universität
wien

DISSERTATION / DOCTORAL THESIS

Titel der Dissertation /Title of the Doctoral Thesis

„In silico prediction of in vitro and in vivo toxicity endpoints based on chemical and biological descriptors“

verfasst von / submitted by

Grda. Marina Garcia de Lomana Rodriguez, M.Sc.

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Doktorin der Naturwissenschaften (Dr.rer.nat.)

Wien, 2022 / Vienna 2022

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	UA 796 610 449
Dissertationsgebiet lt. Studienblatt / field of study as it appears on the student record sheet:	Pharmazie
Betreut von / Supervisor:	Assoc.-Prof. Mag. Dr. Johannes Kirchmair
Betreut von / Supervisor:	Univ.-Prof. Mag. Dr. Thierry Langer

Acknowledgements

First and foremost, I would like to thank my supervisors Johannes Kirchmair and Thierry Langer. Thank you, Johannes, for giving me the opportunity to be part of the group, challenging me to develop my scientific thinking, and always supporting me. Your enthusiasm, broad knowledge, and way of working are a great inspiration for me. I also thank you, Thierry, for taking the lead during the first year and making all this possible.

I would also like to express my gratitude to BASF AG for funding my Ph.D. and giving me the wonderful opportunity to participate in a synergistic collaboration between academia and industry. I am especially thankful to my mentors, Miriam Mathea, Janosch Achenbach and Prof. Schleifer. Miriam, I am really grateful to have been able to learn from you and to share many good moments and inspiring conversations with you. It was a great pleasure having you by my side throughout this journey. Janosch, thank you for your guidance and valuable advice during the first year of my Ph.D. I highly appreciate all those lessons you gave me to introduce me to the cheminformatics world. Thank you, Prof. Schleifer, for your constant support and for giving me the great chance of doing my Ph.D. at BASF.

I am also very grateful to the entire Computational Chemistry group for making me part of the team and allowing me to learn from their experience in a cheerful atmosphere. I also feel lucky to belong to the very friendly and supportive COMP3D group at the University of Vienna, with whom I enjoyed exchanging experiences and getting involved in scientific discussions.

My gratitude goes as well to all my collaborators, both internal and external, especially Ulf Norinder and Roland Buesen, for sharing your knowledge about conformal prediction and toxicology with me. I also appreciate all the interesting discussions with Andrea Volkamer, Fredrik Svensson, Robert Landsiedel, Andrea Morger, Andreas Weber, and Barbara Birk from which I learned so much and that highly contributed to this dissertation.

In addition, I am grateful to Jennifer Hemmerich, Anke Wilm, and Conrad Stork for proofreading parts of this dissertation.

Finally, many thanks to my family for their continuous support and life advice. I am extremely lucky to have you and be able to count on each one of you.

Abstract

Given the large number of new small organic molecules that are being developed and commercialized for diverse applications, there is an urgent need for robust risk assessment strategies to ensure the safety of chemicals with respect to health and the environment. Toxicity assessment currently relies on animal assays that imply serious ethical concerns as well as high costs in terms of time and money. Therefore, substantial efforts are set on the development of in vitro assays and computational tools that can reduce or replace these animal assays while ensuring the compounds' safety.

This dissertation aims at the development of computational toxicology tools tackling current difficulties for the advancement of safe and robust alternative methods. The first part focuses on approaching in vivo toxicity prediction from a lower complexity perspective by predicting single molecular initiating events (MIEs) as a starting point for the determination of adverse outcome pathways of in vivo effects. This was studied at the example of endocrine disruption, a specially challenging toxicity endpoint due to the many regulation pathways involved. The following three parts of the dissertation tackle the prediction of in vivo effects directly. These models aim to enhance in vivo toxicity prediction by including complementary biological information (e.g. pharmacokinetics, MIEs or metabolism) already in the model development phase. Moreover, these in vivo models explore the opportunities that conformal prediction, a framework for the mathematical estimation of the confidence of the predictions, offers in the context of computational toxicology.

Zusammenfassung

Angesichts der großen Zahl neuer kleiner organischer Moleküle, die kontinuierlich für verschiedene Anwendungen entwickelt und vermarktet werden, wird der Bedarf an robusten Risikobewertungsstrategien zur Gewährleistung der Sicherheit für unsere Gesundheit und Umwelt verstärkt. Die Toxizitätsbewertung beruht derzeit auf Tierversuchen, die sowohl ethisch bedenklich als auch zeitlich und finanziell sehr kostspielig sind. Daher werden erhebliche Anstrengungen unternommen, um in-vitro-Assays und in-silico-Tools zu entwickeln, die diese Tierversuche reduzieren oder ersetzen können und gleichzeitig die Sicherheit der Substanzen gewährleisten.

Diese Dissertation befasst sich mit der Entwicklung von in-silico-Toxizitätsmodellen, die aktuelle Herausforderungen für die Weiterentwicklung sicherer und robuster alternativer Methoden adressieren. Der erste Teil beschäftigt sich mit der Vorhersage von in-vivo Toxizität aus einer Perspektive geringerer Komplexität, indem einzelne molekulare auslösende Ereignisse (MIEs) als Ausgangspunkt für die Bestimmung adverser Signalwege von toxischen in-vivo-Effekten vorhergesagt werden. Dies wurde am Beispiel der Störung des Hormonsystems untersucht, einem besonders schwierigen Toxizitätsendpunkt aufgrund der vielen beteiligten Regulationswege. Die weiteren drei Teile der Dissertation befassen sich direkt mit der Vorhersage von in-vivo Effekten und insbesondere mit der Entwicklung neuer Methoden, um in-vivo Toxizitätsmodelle zu verbessern. Zu diesem Zweck werden komplementäre biologische Informationen (z. B. Pharmakokinetik, MIE oder Metabolismus) bereits in der Modellentwicklungsphase eingeführt. Darüber hinaus untersuchen diese in-vivo Modelle die Möglichkeiten, die Conformal Prediction, ein Verfahren zur mathematischen Schätzung der Zuverlässigkeit der Vorhersagen, im Kontext der computergestützten Toxikologie bietet.

Contents

1 Introduction

1.1 Safety assessment of newly developed compounds

Humans and animals are constantly exposed to a great variety of chemicals. Some of these chemicals are naturally formed (e.g. by microorganisms) while others have been developed by humans for a variety of purposes. As the number of available drugs, cosmetics, agrochemicals, or household products grows, our organism is inevitably in contact with more and more new compounds. Not only us, but also the environment can suffer from the increasing amount of chemicals, especially through the vast use of pesticides.[1, 2]

All chemicals can potentially induce toxicity on one or more species through on-target or off-target effects. On-target toxicity refers to exaggerated or adverse effects caused by small molecules interacting with their intended target. This is, for example, the case of the drug statin, an inhibitor of β-hydroxy β-methylglutaryl-CoA (HMG-CoA) reductase widely used to reduce cholesterol in blood in people with cardiovascular risks. Since this reductase is also necessary for other vital functions (e.g. generation of coenzyme Q10 and heme-A), its on-target inhibition also causes several adverse effects (like muscles weakness or inflammation).[3] Off-target toxicity corresponds to adverse effects derived from the interaction of substances with unintended (and often promiscuous) targets. For instance, many drugs that have been withdrawn from the market due to their interaction with human ether-à-go-go-related gene (hERG) channels that results in cardiotoxicity issues.[4]

Extensive testing of newly developed compounds on a variety of toxicity endpoints is of high importance to avoid the risk of severe adverse reactions. In this regard, the EU Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH) regulation deals with the approval of newly developed chemicals to ensure their safety for humans and the environment. REACH establishes a series of protocols to determine the properties and hazards of compounds and decide upon their authorization. Since a comprehensive study of all possible toxicological endpoints is not feasible, the required studies are based on the intended use of the chemical and its exposure routes and concentrations. For instance, the ecological safety of a compound (e.g. with regard to bees or fish) is often required for the registration of agrochemicals or household products, while it is usually not a requirement for the approval of pharmaceutical drugs.

Toxicological data on closely related compounds may sometimes be used to fill data gaps for the authorization of substances by the so-called "read-across" approach. Read-across is based on the assumption that similar compounds have similar bioactivities.[5] Following this principle, the activity of a new compound may be inferred from experimental safety data on closely related compounds. Although the definition of molecular similarity is not straightforward or unique, similarity is commonly determined by the Euclidean or Tanimoto distance based on features describing molecular properties.[6] Both the increasing amount of available experimental data and the demanding testing requirements imposed by REACH are boosting the application of read-across cases in the last few years. Moreover, the support of more accurate in silico quantitative structure-activity relationship (QSAR) predictions also facilitates the acceptance of read-across studies. Nevertheless, a thorough documentation of each read-across case is still required and the final decision on its acceptance remains at the judgement of the regulatory body.

Identifying toxic compounds in early stages of the substance discovery pipeline can prevent investing resources on unfeasible substances that will fail on the unavoidable requirements of the safety assessment. Nevertheless, toxicity testing is usually first conducted at advanced stages of the substance development process as it involves a high number of animal tests, as well as high costs and time investment. Despite important differences between animal species leading to mismatches between the outcome of animal assays and clinical trials, in vivo animal assays are still considered the most accurate safety assessment approach and are required for most toxicological endpoints.[7] However, significant efforts are ongoing to develop alternative methods based on in vitro assays and in silico modeling to limit animal suffering to a non-reducible minimum.[8, 9]

Computational methods for predicting the toxicity of compounds represent valuable tools at several stages of the substance discovery pipeline. In early stages, they can help to prioritize promising candidates with less or no predicted toxicity warnings. As candidate substances advance in the development pipeline and require safety evaluation, toxicity prediction models can highlight certain toxicity assays that present the highest risk of a positive result to accelerate the identification of failing candidates. Finally, in silico methods can support read-across cases to avoid extensive testing of all registered compounds.

1.2 In vivo toxicity testing

In vivo assays are so far the most widespread and accepted approaches for determining the toxicity of a compound on a specific endpoint. These assays are carried out on living animals and can take years to complete (e.g. carcinogenicity studies) and involve several generations of animals (e.g. reproductive toxicity and embryotoxicity studies). Therefore, in vivo assays not only entail ethical concerns due to animal suffering, but also an immense resource investment with regard to time and money. Considering the large amount of compounds being developed and released to the environment every year, it is not feasible to conduct in vivo studies for each compound on a high number of toxicity endpoints.

The indisputable advantage of in vivo assays over other approaches lies in the consideration of the complex interactions present in whole organisms. Taking all of these interactions into account is of utmost importance, as these can have a great influence on the observed effects of a compound on the target organisms. The outcome may be for instance determined by the pharmacokinetics of the compound and its availability at a given time and location in the body. The effects defining this complex system are generally summarized as the absorption, distribution, metabolism and excretion (ADME) parameters.

Absorption refers to the ability of a compound to move from the site of administration to the bloodstream. The bloodstream is the most common vehicle for compounds to move inside the organism and is often reached through intestinal absorption or inhalation. Absorption may be hence determined, among other properties, by the compound's solubility or its chemical stability in the stomach. The distribution of compounds to the target organ or tissue through the bloodstream is the next necessary step to elicit a response. Several factors like the blood flow at the target site, the molecular size or the binding to serum proteins can influence the distribution of substances.

Metabolism also plays a major role in the toxic outcome of a substance, as it may chemically modify its molecular structure and, hence, its properties. In principle, metabolism is the primary defense of the body for the detoxification of xenobiotics. Nevertheless, compounds may also be bioactivated through metabolic transformations resulting in more active (as is the case of prodrugs) or toxic compounds. Metabolic transformations usually convert compounds into more hydrophilic metabolites to facilitate their elimination. The excretion of compounds and their metabolites through the kidneys (via urine) or the bile (via feces) avoids their accumulation and possible toxic effects.

Results from animal assays are extrapolated to humans to decide upon the toxicity of a compound prior to clinical trials. However, biological differences between organisms may lead to substantial discrepancies in the observed effects on animals and humans.[10] For instance, off-target proteins responsible for toxic outcomes in humans may not be present in the animal models. Also variations in the metabolic pathways may lead to the formation of different (potentially toxic) metabolites or different metabolic rates.[11] Perel et al.[12] conducted a systematic review comparing the (beneficial or harmful) effects of drugs in humans during clinical trials and the observed effects in animal studies. From the six analyzed endpoints, the animal studies agreed with the observations during clinical trials in only half of the cases. Moreover, some of these agreements are believed to be also biased by the study design or be caused by random error.

Despite the limited extrapolation to humans, in vivo animal assays are nowadays still the most reliable method for assessing the safety of compounds as they consider a whole biological system and therefore all the interactions that the substance may undergo in the body. However, following the 3Rs-Principle (Reduction, Replacement, Refinement) roadmap for a more ethical use of animal testing, safety assessment is experiencing a shift towards the use of alternative methods better aligned with ethical concerns and regulatory requirements. Especially in the case of cosmetic ingredients, in vitro assays and in silico methods are widely applied and set a good precedent for the development of toxicity testing in this direction.[13]

1.3 In vitro toxicity testing

In vitro toxicity assays are carried out on microorganisms or isolated biological targets, cells or tissues. These assays present the advantages of avoiding animal suffering and often supporting high-throughput screening. Therefore, they may also be conducted early in the substance development pipeline to identify toxic hazards straight away.

There are many efforts set by industry and governmental entities on the development of reliable in vitro toxicity assays. Some well-known initiatives in this area are the Tox21 (Toxicology in the 21st Century program) and ToxCast (U.S. Environmental Protection Agency's (EPA) Toxicity Forecaster) projects. The Tox21 program was initiated by several U.S. government agencies with the aim of detecting adverse effects of small molecules on humans based on high-throughput screening assays. The project covers around 70 assays and over 9000 substances that span from commercial chemicals and pesticides to food additives and medical compounds. Within the ToxCast project from the U.S. EPA, 4500 substances have been screened in more than 700 high-throughput assays. Similarly to the Tox21, the ToxCast library of compounds also includes a high number of pesticides, but also pharmaceutical compounds, cosmetic ingredients and food additives.[14]

In parallel, the Organisation for Economic Co-operation and Development (OECD) is focusing on the identification of adverse outcome pathways (AOP). The AOPs describe the sequence of causally linked events at different levels of the biological systems and help to analyze the pathways leading to a toxic effect. As part of this effort, the OECD is hosting the Adverse Outcome Pathway KnowledgeBase (AOP-KB) to enable sharing and discussing AOPs in the scientific community. Among others, AOPs are of highest relevance for the reduction of animal tests for the identification of endocrine disruptors, as these trigger cascades of events difficult to evaluate with isolated in vitro assays.[15]

Although in vitro toxicity testing is in some cases a valid alternative to animal tests, it also has some clear limitations. These assays only account for localized effects on the target protein, cell or tissue and hence do not consider the ADME effects (or only a subset of them). Therefore, there are often inconsistencies observed between in vitro and in vivo assay outcomes. To overcome the limitations of individual in vitro assays, a battery of assays may be combined to reproduce a concrete in vivo toxicity endpoint. This is the case of the strategy for identifying skin sensitizing substances recently approved by the OECD, which includes three in vitro methods and has shown to perform at least as well as the widely accepted murine local lymph node assay (LLNA) animal test.[16] Another strategy to reduce the gap between in vitro and in vivo results is the emulation of some ADME properties in the in vitro assays. This may be achieved by reproducing the organ tissue instead of working with isolated cells or molecular targets, or by including metabolizing enzymes (with liver microsomes, liver S-9 fractions or hepatocytes) in the culture medium. Moreover, the outcome of in vitro assays may also be combined with in silico models based on human or animal data with the aim of further reducing the in vitro-in vivo gap and reliably substitute animal assays.

1.4 In silico prediction of toxicity

In silico tools for predicting the toxicity of new, untested compounds are developed using existing, measured biological data. Such computational tools can be designed for high-throughput profiling, enabling predictions of the toxicity (or any other property) of hundreds of compounds within a matter of seconds or minutes. Moreover, in silico tools bring a paradigm shift into the compound development pipelines as they enable the evaluation and interactive optimization of conceptual molecules not yet synthesized. These methods are hence a powerful tool for the prioritization of molecules for testing or for choosing relevant assays (e.g. by predicting molecular initiating events (MIEs)). Since computational tools can be trained on human data, they also present the advantage of avoiding issues related to interspecies differences.

The robustness and reliability of in silico models is limited by the quantity and quality of the available experimental data. These models are usually not able to extrapolate the results to new compounds that are overly different from those used to train the models. Therefore, before applying the models, it is of utmost importance to determine both their accuracy and their applicability domain (AD). The AD defines the compounds for which a reliable prediction can be made, and its clear definition and communication are of high relevance to ensure the correct usage of the model (see "Applicability domain" section for details).

As the available data and computational power increase, in silico tools for toxicity prediction have evolved as well. The first QSAR models were based on small data sets and linear combinations of physically meaningful descriptors. Higher amounts of data and evolving machine learning (ML) algorithms allowed the development of more accurate models, often at the cost of less interpretable predictions.[17] Nowadays, the term QSAR is commonly used to describe all forms of predictive models, including ML and deep learning models.

1.4.1 Classical quantitative structure-activity relationship analysis

Structure-activity relationship (SAR) analysis aims to relate the substitution of functional groups with the observed changes in the bioactivity of a compound. Building on that, QSAR analysis intends to quantify these differences through mathematical or statistical relationships. These relationships are typically represented by linear models combining one or several structural or physicochemical properties.[18, 19] One example of such a QSAR model is a recently developed model by Naseem et al. for the prediction of human skin permeability of neutral organic chemicals based on a linear combination of the partition coefficients for octanol-water and air-water systems.[20]

In order to relate small chemical changes with the biological activity, traditional QSAR analyses are generally restricted to specific classes of compounds or ligand-target interactions and binding modes. This characteristic is hence limiting the usability of classical QSAR models for the prediction of some toxicity endpoints, in which several physicochemical and biological events are involved (e.g. permeability, bioavailability, interaction with off-targets, etc.). Nevertheless, classical QSAR analysis may still be useful for the examination of isolated toxicological targets in order to understand and minimize the interactions leading to the toxic outcome.[21, 22]

1.4.2 Machine learning

ML models are statistical methods that uncover patterns on new data, based on learned observations on past data. The methods can be classified in two main groups based on the input data and the final goal: unsupervised and supervised ML. In unsupervised ML, the input data is unlabeled (i.e. there is not a predefined correct answer to the problem) and the aim of the model is to find patterns that separate the data in two or more groups. These methods are often used in in silico toxicity applications for visualizing the chemical space covered by the available data. Popular models used for this purpose are principal component analysis[23] (PCA) and uniform manifold approximation and projection[24] (UMAP). These methods conduct a dimension reduction on the input features to allow a two-dimensional representation of the samples. PCA reduces the dimensions by creating new variables that maximize the explained variance (i.e. the information content). In contrast, UMAP is designed to maintain the pairwise distance between samples in the lower dimensional space.

In supervised ML the input data are labeled (e.g. with measured activity values) and the aim is to map the input features describing the samples with the given values. In toxicity prediction, supervised ML models are applied to predict the activity of untested compounds, based on assay outcomes of a set of measured compounds. For training, the models are presented with a set of features representing characteristics of the compounds (see "Molecular encoding" section for details), as well as the measured activity of these compounds. Supervised ML models can be further divided into classification (if the label is categorical; e.g. "active" and "inactive") and regression (if the label is continuous; e.g. LD_{50} value) models. Given the small size of the available data sets in toxicity prediction applications, the development of regression models returning a continuous activity value (e.g. LD_{50}) is often challenging. Regression models usually require samples to cover a wide range of values and to have a similar distribution over all possible values (i.e. no skewed data) in order to be robust. Classification models, on the other hand, are normally trained using binary classes for a given compound (e.g. "active" and "inactive"), hence simplifying the problem. Nevertheless, setting a cut-off on the activity value to label the compounds as "active" or "inactive" is not always straightforward and may strongly influence the relevance, quality and interpretation of the classification model.

Some of the most commonly employed, supervised ML algorithms are linear and logistic regression[25] (LR; for regression and classification applications, respectively), support vector machine[26] (SVM), random forest[27] (RF) and gradient boosted trees[28] (GBT). Linear regression and LR models map a function describing a linear combination of one or more independent features to the given values. SVMs project the features into a hyperplane that maximizes the margin between samples from different classes and which is used as a decision boundary. RF classifiers combine the predictions from an ensemble of decision trees trained on different subsets of data. GBT models are also based on ensembles of decision trees but built in a stage-wise manner where each tree is designed to correct the mistakes made by the previous one. Moreover, deep learning, a subtype of ML based on neural networks (NN),[29] has been strongly developed and applied in recent years. These networks are formed by several hidden layers of interconnected neurons that aim to simulate the functions of the human brain. The model learns by transferring information back and forth along the network with the objective of minimizing the error between the prediction and the correct label.

Even in cases where a relatively large amount of toxicological data are available, these data often contain only small amounts of active compounds. This data imbalance can hinder the recognition of active compounds by the model, as a small variety of active samples may not be enough to derive general characteristics.[30] Moreover, having a small percentage of active compounds may bias the model to predict all compounds as inactive, since ML models are trained to minimize wrong predictions (and predicting a compound as inactive is in most cases correct). To avoid this bias, a method for bias correction, such as class weight balancing, oversampling or undersampling, is often applied. By balancing the class weights, errors committed on samples from the minority class (usually active compounds) are penalized harder than errors on samples from the majority class (usually inactive compounds). The oversampling technique revolves around the increase of the number of samples representing the minority class in the training data, either by duplicating samples or by creating new synthetic samples by inference (e.g. with the SMOTE[31] method). In the undersampling approach, some samples from the majority class are removed from the training data to balance the ratio of samples between classes (effectively causing a loss of information).

It is essential to define the AD of ML models and recognize unreliable predictions, as the false prediction of a toxic compound as non-toxic must be avoided. In general, the AD and quality of the predictions of ML models rise as the amount and diversity of training data increase. Large and diverse data sets may allow the model to generalize better to unseen compounds, as they represent more significant correlations between the input features and the activity than data sets with only few hundreds of compounds. Unfortunately, the amount of available data is often the biggest limitation for developing well performing toxicity prediction models. In the case of in vivo assays, data is usually exceptionally scarce, as the number of animal experiments are small. Occasionally, in vivo data for drugs is also available from reports during clinical trials or from consumers once the compound has been released to the market. Problematic drugs may then be withdrawn from the market or annotated with a "black box warning" and also serve as input data for ML models. Thanks to initiatives like the Tox21 and ToxCast, the amount of publicly available in vitro toxicity data has increased over the last years. The higher amount of data together with the lower complexity of the assays (compared to in vivo endpoints) makes in vitro assays often good candidates for the development of ML models. Several well-performing in silico models for the prediction of in vitro assays have already been developed. To assess the performance of different ML algorithms and workflows on in vitro toxicity endpoints, the Tox21 data challenge compared several approaches for the prediction of the outcome of 12 assays covering nuclear receptor signaling and stress pathway assays using only the chemical structure of the compounds.[32, 33] The best performing models of the challenge were normally consensus models combining the predictions from several underlying models and types of descriptors, proving that different models can learn different, useful information. However, there is a great variety of algorithms and approaches for training a ML model and no universal best method for it. Hence, the model development workflow usually needs to be adapted for each individual endpoint.

1.4.2.1 Molecular encoding

Chemical information describing the compounds needs to be encoded in machine readable features from which ML models can learn. The selection of input descriptors may have an important influence on model performance, as different encodings capture distinct properties and characteristics of the molecules.

General features derived from 0D to 3D properties of the molecules can be used as input descriptors for ML models. 0D descriptors (such as atom and bond counts, or sums of atom properties) require no structural information for calculation. 1D descriptors represent information about molecular substructures, such as the count of functional groups or molecular fragments. 2D descriptors are calculated from the graph representation of a molecule, considering connectivity and adjacency properties. These can describe features such as size, shape or polarity, as well as atom-specific properties like the hybridization state. Finally, 3D descriptors are derived from the molecular conformation (i.e. geometrical representation). These descriptors can capture properties that are particularly relevant to the interactions between compounds and biological targets but are limited by the uncertainties related to the biologically active conformations of compounds. For calculating 3D descriptors, intense conformational sampling is generally conducted to predict the energetically most favorable conformation, which is yet not necessarily the biologically active one. Moreover, quantum chemical features describing e.g. atomic charges, energies of the highest occupied (HOMO) and lowest unoccupied (LUMO) molecular orbitals or orbital electron densities, are also widely applied in QSAR models and are particularly well suited for modeling reactivity and physical properties.[34] There are several cheminformatic tools available for calculating physicochemical properties, like RDKit,[35] alvaDesc[36] or the Molecular Operating Environment (MOE).[37] Besides physicochemical properties, one of the most commonly used chemical descriptors are molecular fingerprints, which encode structural features in the form of a vector indicating the absence, presence or count of each feature in a molecule. MACCS keys[38] are an example of a common and straightforward molecular fingerprint that encodes a set of 166 fixed structural properties (e.g. presence of a sulfur bond or more than three oxygen atoms). Popular and more complex descriptors are the extended-connectivity fingerprints (ECFP)[39] or Morgan fingerprints, a type of circular fingerprints encoding the presence of molecular fragments of different length (depending on the selected radius). The features corresponding to the circular fragments are mapped (i.e. “folded”) into a vector of fixed length (typically 1024 or 2048 bits) to reduce the size and sparsity of the vector containing all possible fragments.

One type of features commonly used in toxicity prediction are the so-called structural alerts. Structural alerts are substructures that commonly appear in compounds exhibiting toxicity with a similar mode of action. They exist for a number of toxicological endpoints and are usually derived from expert knowledge[40] or by statistical evaluation of the appearance of molecular fragments in toxic compounds.[41, 42] For ML applications, a vector encoding the presence or absence of each structural alert can be used as input descriptor. The drawback of structural alerts is their limitation to already known problematic features that prevents the extraction of information or recognition of new modes of action. However, their simplicity and interpretability still make them a useful asset for understanding the underlying mechanism of toxicity and help to interpret the predictions.[43]

In recent years there has been an explosion of methods for deriving task-specific molecular descriptors that are directly learned from the molecular graph. These methods are usually based on graph-convolutional networks (GCN) that automatically extract the best representation of the molecules (i.e. the most relevant features) for the given task. GCNs are a very promising tool for property prediction and have already shown to perform at least as good as predefined molecular fingerprints like ECFPs in a variety of setups.[44, 45] However, this kind of models generally need a high amount of input data to avoid model overfitting and derive descriptors that can generalize well to unseen data.

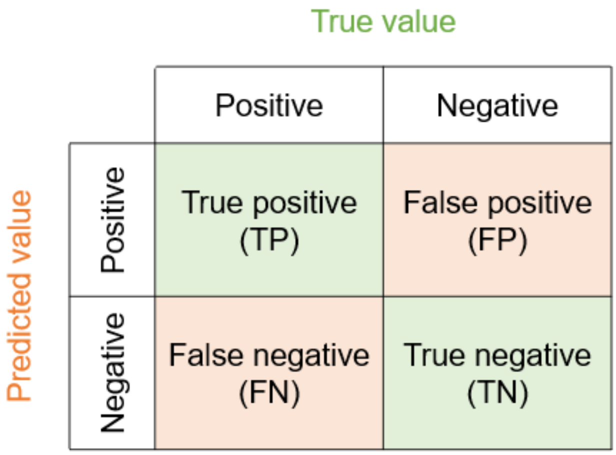
Other research studies have shown that descriptors derived from biological data are also promising for training activity prediction models.[46-48] To this end, the outcomes of a set of compounds on several assays (often high-throughput screening assays) are concatenated and used as input features. This information may improve ML models by describing the behavior of compounds in biological systems and hence complementing the structural information contained in chemical descriptors. The bottleneck of this approach is usually the quality and quantity of the available biological assay data for building these bioactivity descriptors.[47, 49] Generally, to train a ML model with bioactivity descriptors, the outcome of compounds in all the assays comprising the descriptor is needed. Otherwise, the data matrix would contain missing values that must be filled using imputation techniques, which at the same time may add bias and/or noise to the models.

1.4.2.2 Model performance evaluation

In order to estimate the quality and reliability of the predictions, model performance should be evaluated on a test set of samples with known labels not used during model training. One of the most common model evaluation frameworks is cross-validation (CV), which is especially popular in applications where the available data are limited. In a *k*-fold CV workflow, the data are split into *k* fixed groups of samples, and at each fold, *k*-1 groups are used as a training set (on which the model is trained) and the remaining group as a test set (on which the model is evaluated). With this approach, all samples are once contained in the test set, and used for model training in the remaining *k*-1 folds. Generally, the data are randomized prior to the splitting to avoid possible bias due to the ordering of samples in the data set. Furthermore, there are also other variations of CV that allow e.g. stratified splitting (containing the same percentage of samples from each class in each fold), time splitting (separating the training and test set based on a timestamp in the data), or cluster splitting (maintaining complete clusters of similar samples always in the same set to ensure chemical diversity between training and test set).

A holdout test set with samples not considered at any point during model training may also be used to estimate model performance. This approach is usually employed when the amount of data is large and reserving some data for evaluation does not impact the robustness of the model, or in scenarios where new data are generated after model development.

The performance evaluation on the test set is performed by comparing the predictions with the real value. One way to assess the performance of a binary classification model is the confusion matrix, which indicates the number of true positives, true negatives, false positives and false negatives (Figure 1.1). These values can also be summarized in a single metric (e.g. F1 score or Matthews correlation coefficient (MCC)) to facilitate the comparison of models (see Methods for details). In the case of regression models, some of the most common metrics used to assess model performance are the coefficient of determination (R^2) and the root mean squared error (RMSE). The R^2 represents the proportion of the variance that is explained by the model and the RMSE measures the difference between the correct and the predicted value (being an RMSE of zero the perfect prediction).



		True value	
		Positive	Negative
Predicted value	Positive	True positive (TP)	False positive (FP)
	Negative	False negative (FN)	True negative (TN)

Figure 1.1: Confusion matrix for the evaluation of predictions of binary classifiers. There are four types of possible predictions depending on the predicted and the true value (true positive, true negative, false positive, and false negative).

1.4.3 Applicability domain

Model performance is usually estimated on a small subset of compounds (test set) and should reflect how accurate the predictions on a new subset of unlabeled compounds will be. However, the error rate on new compounds not well embedded in the chemical space of the training data may be higher than for those in the test set (which usually have a similar distribution to the training set).[50] To avoid predictions with a higher than expected error rate, a definition of the compounds on which the model should be applied to is needed. This domain of compounds for which the model can make reliable predictions is referred to as the AD of a model. Compounds may fall out of the AD mainly due to two reasons: novelty and anomaly.[51] On the one hand, novel compounds are located in a widely differing descriptor space than the training data and the model may not have enough information to make reliable predictions on them. These compounds can be identified by novelty detection techniques, like the distance to the nearest neighbor in the descriptor space of the training data. On the other hand, anomalous compounds may be well embedded in the descriptor space covered by the model but be outliers with regard to their label. These compounds are usually detected with confidence estimations of the prediction (e.g. built-in class probability estimates). A large benchmark study comparing error reduction using several AD definitions concluded that built-in class probability estimates generally performed better than the alternatives (e.g. distance measures).[52] Since confidence estimation methods consider not only the descriptor space (like novelty detection methods) but also the class labels, these methods may be more reliable.[51]

A popular method for confidence estimation and AD definition of in silico models is conformal prediction (CP). CP models generate predictions at a user-defined error rate as long as the exchangeability assumption between training and test sets holds.[53] These models have the advantage of mathematically defining the AD by just determining the allowed error rate, and without the need of setting more or less arbitrary thresholds to confidence estimates like distance measures. This advantage makes CP an especially powerful tool for toxicity prediction, where the definition of the AD is of utmost importance.[54-56]

To estimate the uncertainty of the predictions in the CP framework, the training set is further divided into a proper training set and a calibration set. The ML model is only trained on the proper training set, while the calibration set is used to estimate the confidence of the predictions made on the test set (Figure 1.2.a.).[57] For that purpose, both the predictions on the calibration and test sets are first transformed into a nonconformity (nc) score (by applying a nc function) that presents low values for predictions close to the true value. The calibrated probabilities (named p-values) are then calculated as the rank of the nc score obtained for the test sample among the nc scores for the calibration set. Based on the p-value and the defined allowed error rate, a set of classes (in the case of classification) or a value range (in the case of regression) are reported (see Methods for details).

There are several variations of the CP framework depending on (a) the splitting of the training data and averaging of the predictions or (b) the separation of nc scores based on the class label (Figure 1.2.). The inductive CP (ICP) is the baseline framework and uses fixed proper training and calibration sets. In aggregated CP[58] (ACP), the splitting in proper training and calibration sets is repeated several times to minimize the possible bias introduced by the random splitting of the training set, as well as the effects of the information loss caused by only using part of the training data for model development. The final p-value is obtained by averaging the resulting p-values from the different models. In the synergy CP[59] (SCP) workflow, the calibration set is kept constant, while several models are trained on different proper training sets. The nc scores for the calibration and test sets are averaged across the predictions made by the different models. This approach has shown to be useful for federated learning, where individual institutions can train a different model on their (confidential) data without needing to pool or disclose the data.[60] Moreover, these three CP types may be combined with the Mondrian CP approach to address the problem of imbalanced data in binary classification.[61] Although CP models should output the defined error rate (as long as the exchangeability assumption holds), the errors may be unevenly distributed between classes. To maintain the expected error rate in both classes, in the Mondrian approach the nc scores obtained for the calibration set are separated in independent nc score lists depending on the class label of the sample. For the test samples, one p-value for each of the two classes is calculated by comparing the predictions with the respective nc score list.

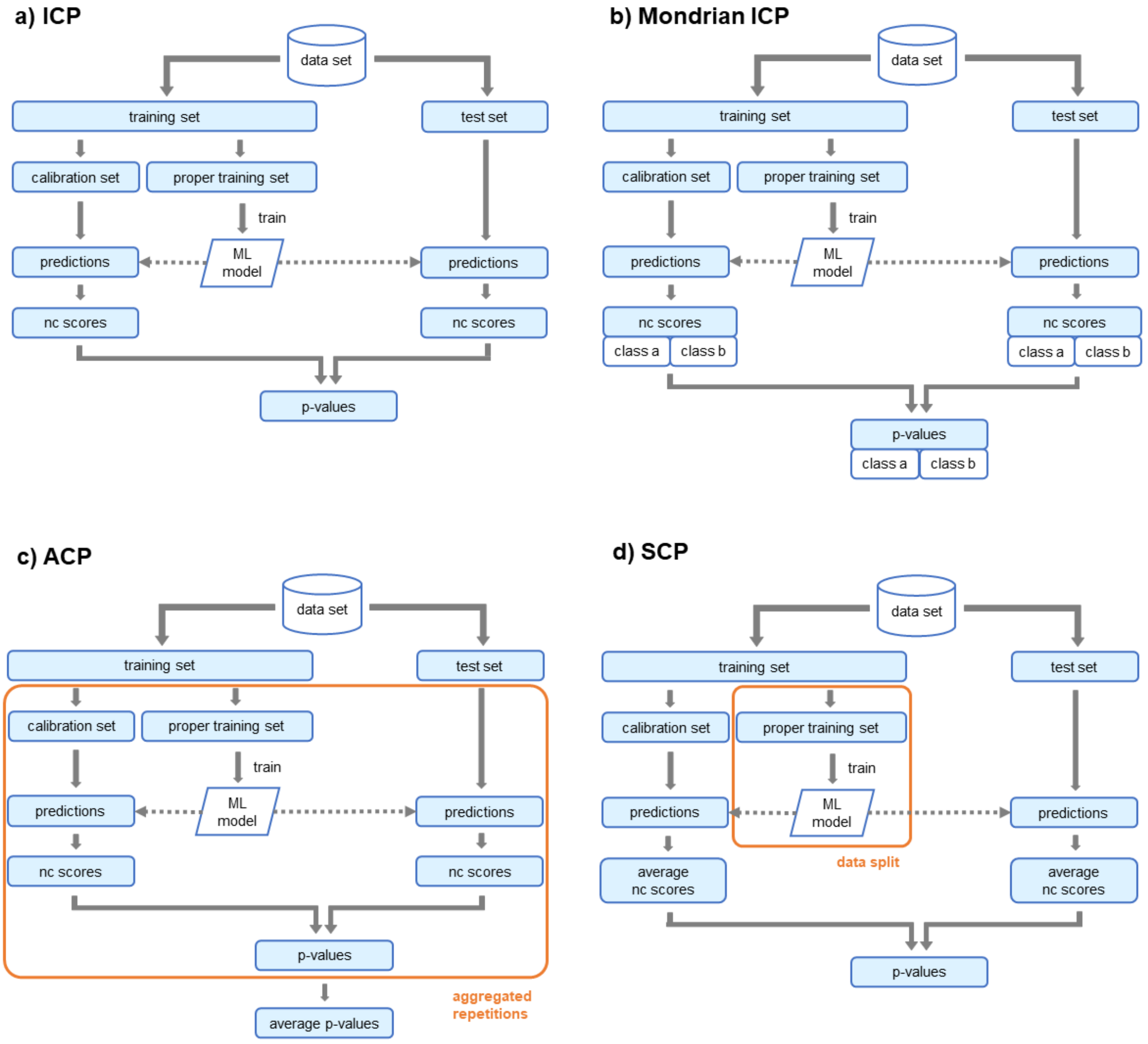


Figure 1.2: Overview of different conformal prediction (CP) workflows. Compared to the (a) baseline inductive CP (ICP) workflow, the main differences between workflows reside in the way the data is split for model training ((c) aggregated CP (ACP) and (d) synergy CP (SCP)) or in the class separation for calculating the nonconformity (nc) scores and p-values ((b) Mondrian ICP).

2 Aims

The safety assessment of chemicals with regard to human health and the environment is an indispensable requirement for the authorization of newly developed compounds, including drugs, cosmetics and agrochemicals. However, toxicity testing entails critical ethical concerns from the use of animals and is highly time consuming and expensive. Therefore, toxicity is usually first evaluated at late stages of the substance development process, causing high attrition rates. Computational methods are a useful tool for predicting the toxicity of compounds in a high-throughput manner early in the development pipeline to help the prioritization of promising candidates. Moreover, they can also support read-across cases as well as the extrapolation from in vitro to in vivo results.

This dissertation focuses on the development of novel in silico toxicity tools approaching in vivo toxicity prediction from different angles and complexity levels. With the presented studies we aim to answer the following questions:

1. **Can in silico models identify endocrine disruptors and determine which MIEs they are triggering?** The identification of endocrine disruptors, a complex toxicity endpoint involving many regulation pathways, was tackled by the development of in silico models for a set of MIEs involved in the perturbation of hormone homeostasis. This approach was elaborated at the example of thyroid hormones, a highly relevant endocrine pathway, for which only few in silico tools have been developed so far. Experimental data for a battery of protein targets involved in thyroid hormone homeostasis was collected and curated. These data were used to develop ML and deep learning models that can predict MIEs of endocrine disruption to determine or confirm the triggered AOPs.
2. **To what extent can in vivo toxicity prediction be enhanced by bioactivity descriptors representing the activity of compounds in biological systems?** The outcome of in vivo assays is often challenging to predict with in silico models due to the high number of parameters that come into play when considering whole organisms. To address the issue of building well-performing ML models for in vivo endpoints, novel predicted bioactivity descriptors (defining the outcome of compounds in over 300 in vitro and pharmacokinetics assays) were developed within a CP framework. By combining these descriptors with state-of-the-art chemical features, we aimed to bridge the in vitro-in vivo gap.

3. **Can we mitigate the effects of data drifts on CP models to make them applicable to samples from different feature distributions?** Data drifts between the training and test data may appear over time or when models trained on public data are applied on proprietary data. This is usually the case if the test data cover a different part of the descriptor space or were derived with differing experimental conditions. We evaluated an approach for recalibrating models (without the need to retrain them) in order to overcome the problem of the described data drift scenarios and make models directly applicable to differing test sets.
4. **Can the incorporation of information about predicted metabolites in toxicity models improve the identification of toxic compounds?** Initially safe compound structures may be bioactivated by metabolism into reactive and toxic metabolites. Considering possible metabolic structures of each parent compound in toxicity prediction models may be a key element to improve the predictions when this bioactivation occurs. To study this possibility, a variety of approaches for including predicted xenobiotic metabolism information into ML models were explored with the aim of improving toxicity predictions.

3 Methods

3.1 Data collection, curation and processing

Most of the data used in this thesis for model development was collected from public domain databases and literature (including ToxCast,[14, 62] eMolTox[63] and eChemPortal).[64] Details about the data used in each study can be found in the respective section. Only in section 4.3. proprietary data from BASF SE for two endpoints (micronucleus test (MNT) and liver toxicity) were used.

In order to standardize the collected molecular structures, data curation and processing workflows were developed using KNIME.[65] Starting with the molecular structure represented as SMILES strings, steps for removing solvents, salts and small fragments, annotating aromaticity, removing stereochemical information, neutralizing charges and mesomerizing structures (returning their canonical resonant form) were included. The canonical SMILES was calculated from the processed structure and further used for deduplication. In cases of duplicate SMILES with conflicting class labels, the structures were removed from the specific data set. This deduplication procedure was also applied when data from different sources was merged to increase the size and coverage of an endpoint-specific data set.

3.2 Machine learning approaches

As input for the ML models developed in this thesis, the molecular structure of the compounds was encoded with Morgan fingerprints, a type of circular fingerprints, and physicochemical descriptors calculated with RDKit (see "Molecular encoding" section for details). Moreover, some of the models (sections 4.2. and 4.3.) also included predicted bioactivity descriptors encoding the outcome of compounds in pharmacokinetics and in vitro assays.

The unsupervised models PCA and UMAP were used in this dissertation to visualize the chemical space covered by the collected data sets (see "Machine learning" section for details). PCA was trained on a set of physically meaningful descriptors to interpret differences in the properties of active and inactive compounds, while UMAP was trained on the whole set of calculated descriptors to analyze the representation of the chemical space used as input for training the predictive ML models.

The supervised ML models developed within this thesis were in all cases classification models for the prediction of discrete class labels ("active" and "inactive" or "toxic" and "non-toxic"). Several classification algorithms, such as LR, SVM, RF and GBT, were explored (see "Machine learning" section for details). Moreover, multilayer perceptron NN (a type of feedforward NN with multiple hidden layers) were trained for single-task and multi-task models. Multi-task models are trained on several endpoints at the same time to try to enable transfer learning and benefit from features learned from complementing data on other endpoints. Keras[66] was used to develop the single- and multi-task NN models. For the PCA, LR, SVM, RF and GBT models the scikit-learn[67] implementation was employed.

3.3 Conformal prediction framework

The CP framework was used in this thesis to mathematically define the AD of the models with a fixed error rate and to explore how the CP characteristics may be exploited in different ML scenarios. More specifically, CP was applied for the derivation of predicted bioactivity descriptors, as well as for the prediction of genotoxicity in vivo and organ toxicity (section 4.2.). Moreover, a method for mitigating the effects of data drifts on CP models was also explored (section 4.3.).

For developing the CP models, the data sets were split into a training and a test set (80% and 20% of the data, respectively; with a random stratified split). The training set was then further divided into a proper training set and a calibration set (70% and 30% of the training data, respectively) using also random stratified splitting. RF models were trained on the proper training set and applied on both the calibration and the test sets to obtain the predicted probabilities. From the predicted probabilities, nc scores were calculated by applying the inverse probability error function (i.e. one minus the predicted probability for the true class).

The nc score of each test sample was then compared with the list of nc scores from the calibration set, and its rank in the list was used to calculate the p-value (from which the predicted class is finally derived as described below; Figure 3.1.). To ensure that the significance level (i.e. error rate) is evenly distributed between classes, the Mondrian CP approach was used (Figure 1.2.b.). Hence, the nc scores of the calibration set were separated in two lists based on the experimental class labels. After comparing the nc scores of a test sample with both lists, a p-value for each class was obtained.

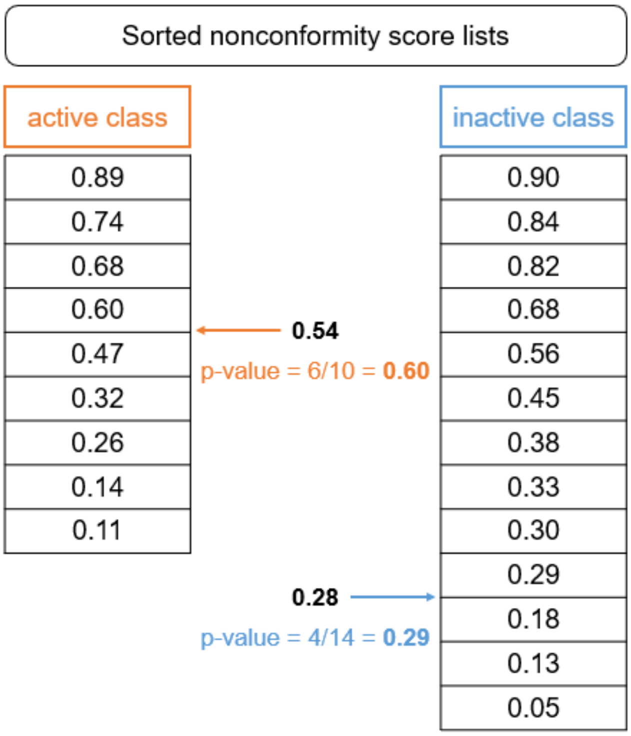


Figure 3.1: Calculation of p-values by comparing the nonconformity scores of the calibration set (separated by class labels) and the nonconformity scores of a test sample. At a significance level of 0.20 the predicted labels would be both “active” and “inactive”, while at a significance level of 0.30 the only predicted label would be “active”.

An aggregated CP approach was conducted by repeatedly splitting the training set into different calibration and proper training sets (Figure 1.2.c.). Within the aggregated CP framework, several models trained on different proper training sets were applied on a variety of calibration sets (and on the unaltered test set). The final p-value for a test sample was calculated as the median p-value among all splits.

The output of CP models is a set of class labels, which is derived from the defined significance level and the p-values. If the p-value for a class is above the significance level, that class is assigned to the sample. Besides the binary outcomes (e.g. “active” and “inactive”), a sample can also be predicted to be “both” (if both p-values are above the significance level) or “none” (if none of the p-values are above the significance level). These two sets of labels indicate that the model does not have enough information to make a single class prediction at that significance level, or that the sample is outside the AD of the model.

3.4 Model performance evaluation

In the context of this thesis, models were evaluated within 5-fold or 10-fold CV using random stratified splitting. Also holdout test sets were employed in section 4.3. to evaluate the performance of CP models after data drifts. These holdout test sets were either derived by a time-splitting approach or collected from additional sources.

Several metrics were applied to evaluate the performance of the predictions on the respective test sets: recall, precision, F1 score, MCC, balanced accuracy and area under the receiver operating curve (AUC). The recall (Eq. 1) measures the ratio of predicted true positives among all real positives, while the precision (Eq. 2) measures the ratio of predicted true positives among all positive predictions. These two metrics may be further summarized into the F1 score (Eq. 3), which is the harmonic mean of the precision and the recall. Another popular metric is the MCC (Eq. 4), as it considers the four classes of predictions (true positive, true negative, false positive, and false negative predictions). The MCC takes values in the range of -1 to +1, being +1 the perfect prediction. Other commonly used metrics are balanced accuracy (Eq. 5), which quantifies the average recall obtained for each class, and the AUC (Eq. 6), which measures the ability of the model to rank the predictions according to their true label. The F1 score, MCC and balanced accuracy have the advantage that they are robust against data imbalance, which is the usual scenario in toxicity prediction applications.

$$Recall = \frac{TP}{TP+FN} \quad (1)$$

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

$$F1\ Score = 2 * \frac{Precision*Recall}{Precision+Recall} \quad (3)$$

$$MCC = \frac{TP*TN-FP*FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (4)$$

$$Balanced\ Accuracy = \frac{(\frac{TP}{P}+\frac{TN}{N})}{2} \quad (5)$$

$$AUC = \int_{x=0}^{1} \frac{TPR}{FPR(x)} dx \quad (6)$$

In the case of CP models, there are two further specific metrics that need to be considered due to the output of a set of labels (instead of single labels): validity and efficiency. The validity measures the ratio of predictions containing the correct label, being “both” predictions always correct and “none” predictions always wrong. The validity should be approximately one minus the significance level (if the calibration and test data are exchangeable), indicating that the model is valid and guarantees the defined error rate. The efficiency measures the ratio of single class predictions (i.e. predictions containing exactly one label) and gives an indication of how good the coverage of the model is on the test set. Moreover, the abovementioned metrics for the general evaluation of ML models can also be calculated on the single class predictions of CP models to evaluate their quality.

4 Results

4.1 Prediction of molecular initiating events of endocrine disruption at the example of thyroid hormones

Endocrine disrupting chemicals are compounds affecting hormone homeostasis in the body by interfering with the synthesis, transport, degradation, or action of hormones. Given the large amount of chemicals we are exposed to and the high relevance of hormone homeostasis for multitude of body functions, regulatory agencies are setting their focus on the detection of these compounds.[68, 69] Computational methods could help not only to detect endocrine disruptors in a high-throughput set up, but also to identify the addressed biological target and to give a mechanistic explanation (which is often missing in in vivo assays). Identifying the addressed off-target can enable the understanding of the disruption mechanism and the redesign of the toxic compound to avoid the interaction with the target.

Several QSAR and ML approaches have already been developed for the identification and characterization of estrogen and androgen disruptors.[70-74] However, only few studies about in silico tools for identifying thyroid hormone disruptors are available so far. Moreover, these studies are limited to two protein targets (thyroid peroxidase (TPO) and thyroid receptor (TR)).[75-77]

In the following study, ML models for a battery of assays on targets related to the dysregulation of thyroid hormone homeostasis (TPO, TR, deiodinases 1, 2 and 3, sodium/iodide symporter, thyrotropin-releasing hormone receptor, and thyroid-stimulating hormone receptor) were developed based on data from the ToxCast database and related literature. After a thorough data curation procedure, predictive ML models were developed by optimizing the combination of the selected algorithm (including RF, LR, SVM, GBT, single-task and multi-task NN) and the data balancing technique (including class weight balancing, oversampling and undersampling). Moreover, a deeper analysis on the predictivity of the models was conducted by evaluating the correlation of the performance with distance metrics and probability estimates.

[P1] In Silico Models to Predict the Perturbation of Molecular Initiating Events Related to Thyroid Hormone Homeostasis

Marina Garcia de Lomana, Andreas Georg Weber, Barbara Birk, Robert Landsiedel, Janosch Achenbach, Klaus-Juergen Schleifer, Miriam Mathea, and Johannes Kirchmair

Chemical Research in Toxicology, 2021

Available at https://doi.org/10.1021/acs.chemrestox.0c00304

Contribution:

M. Garcia de Lomana, M. Mathea and J. Kirchmair conceptualized the research. M. Garcia de Lomana and A. Weber compiled the data sets and analyzed them together with B. Birk and R. Landsiedel. M. Garcia de Lomana developed the machine learning models with contributions of J. Achenbach, K.J. Schleifer and M. Mathea. A. Weber, B. Birk and R. Landsiedel wrote the introduction with contributions of M. Garcia de Lomana and J. Kirchmair. M. Garcia de Lomana wrote the remaining parts of the manuscript, with contributions from J. Achenbach, K.J. Schleifer, M. Mathea and J. Kirchmair. J. Kirchmair and M. Mathea supervised the work.

The following article was reprinted with permission from:

Garcia de Lomana, M.; Weber, A. G.; Birk, B.; Landsiedel, R.; Achenbach, J.; Schleifer, K. J.; Mathea, M. and Kirchmair, J. In Silico Models to Predict the Perturbation of Molecular Initiating Events Related to Thyroid Hormone Homeostasis, *Chem. Res. Toxicol.*, **2021**, 34, 396–411.

Chemical Research in Toxicology

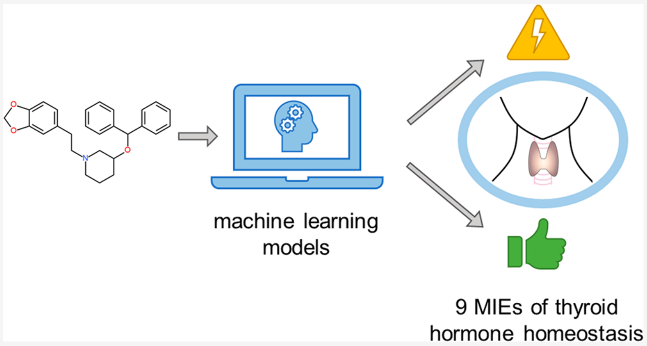
In Silico Models to Predict the Perturbation of Molecular Initiating Events Related to Thyroid Hormone Homeostasis

Marina Garcia de Lomana, Andreas Georg Weber, Barbara Birk, Robert Landsiedel, Janosch Achenbach, Klaus-Juergen Schleifer, Miriam Mathea,* and Johannes Kirchmair*

Cite This: *Chem. Res. Toxicol.* 2021, 34, 396−411

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information



ABSTRACT: Disturbance of the thyroid hormone homeostasis has been associated with adverse health effects such as goiters and impaired mental development in humans and thyroid tumors in rats. In vitro and in silico methods for predicting the effects of small molecules on thyroid hormone homeostasis are currently being explored as alternatives to animal experiments, but are still in an early stage of development. The aim of this work was the development of a battery of in silico models for a set of targets involved in molecular initiating events of thyroid hormone homeostasis: deiodinases 1, 2, and 3, thyroid peroxidase (TPO), thyroid hormone receptor (TR), sodium/iodide symporter, thyrotropin-releasing hormone receptor, and thyroid-stimulating hormone receptor. The training data sets were compiled from the ToxCast database and related scientific literature. Classical statistical approaches as well as several machine learning methods (including random forest, support vector machine, and neural networks) were explored in combination with three data balancing techniques. The models were trained on molecular descriptors and fingerprints and evaluated on holdout data. Furthermore, multi-task neural networks combining several end points were investigated as a possible way to improve the performance of models for which the experimental data available for model training are limited. Classifiers for TPO and TR performed particularly well, with F1 scores of 0.83 and 0.81 on the holdout data set, respectively. Models for the other studied targets yielded F1 scores of up to 0.77. An in-depth analysis of the reliability of predictions was performed for the most relevant models. All data sets used in this work for model development and validation are available in the Supporting Information.

INTRODUCTION

Thyroid hormones regulate physiological processes such as basal metabolism and the growth and development of the pituitary gland, heart, liver, bone, and brain.[1] Disturbances of the thyroid hormone homeostasis have been linked to goiters, hypothyroidism, and impaired mental development in humans[2−5] and thyroid tumor formation in rats.[6−9] Thyroid hormone homeostasis is maintained by a complex system involving thyroid hormone synthesis, distribution via the bloodstream, metabolism, elimination, and a negative feedback loop between the hypothalamic−pituitary−thyroid (HPT) axis. In brief, the hypothalamus secretes the thyrotropin-releasing hormone (TRH), which binds to the thyrotropin-releasing hormone receptor (TRHR) in the anterior pituitary, triggering the production and secretion of the thyroid stimulating hormone (TSH).[10] TSH binds to the TSH receptor (TSHR) of the thyroid gland, initiating thyroid hormone synthesis.[11] As an initial step of the thyroid hormone synthesis, the sodium iodide symporter (NIS), an intrinsic membrane transporter located at the basolateral membrane of thyrocytes, mediates the active transport of iodide into the thyroid gland.[12] Thyroid peroxidase (TPO), a heme containing peroxidase located at the apical membrane of the thyrocytes, catalyzes the iodination as well as the coupling of tyrosine residues to thyroglobulin to form tetraiodothyronine (T4) and, to a lesser extent, the more active form triiodothyronine (T3). Deiodinases (DIO), a group of selenocysteine-containing enzymes, regulate thyroid hormone signaling through the deiodination of thyroid hormones, resulting in the formation of thyroid hormone metabolites with differing activity. DIO1 not only plays an important role in systemic T3 production in the thyroid but also in recycling iodide from thyroid hormone metabolites in excreting organs like the liver and kidney. DIO2 and DIO3 regulate local thyroid hormone signaling in peripheral tissue through

Special Issue: Computational Toxicology

Received: July 29, 2020
Published: November 13, 2020

https://dx.doi.org/10.1021/acs.chemrestox.0c00304

activation (DIO2) and inactivation (DIO3) of thyroid hormones. DIO2-expressing tissues include the pituitary gland, skeletal muscle, bone, brown adipose tissue, and the thyroid, while DIO3 is mainly present in placental tissue and the developing embryo as well as in neurons in the brain.[13] The transcription of thyroid hormone-regulated genes is initiated through the binding of thyroid hormones (T3 in particular) to thyroid hormone receptors (TR). Upon hormone binding, the TR–thyroid hormone complex translocates into the nucleus and interacts with response elements on the DNA, leading to the transcription of thyroid hormone-regulated genes.[14]

Chemicals have been reported to disturb the HPT axis through a variety of mechanisms. In the context of regulations (EU) no. 528/2012 and (EC) no. 1107/2009, the European Food Safety Authority published a guideline for the identification of endocrine disrupting compounds. This guideline defines scientific criteria for the determination of endocrine-disrupting properties of chemicals,[15] leading to an increased need for methods to detect endocrine-mediated effects.

The Organization for Economic Co-operation and Development (OECD) proposes a tiered approach for the evaluation of potential endocrine disruptors using all existing toxicological data. Level 1 of this tiered approach involves physical and chemical property analysis, read-across, quantitative structure–activity relationship (QSAR) analysis, and further in silico methods. Level 2 involves in vitro assays for individual end points, and Levels 3–5 involve in vivo assays providing different layers of information.[16] In vitro models are available for many key events related to the HPT axis,[17] but none of these have been validated and accepted by the OECD yet. In silico and in vitro methods can guide product development and avoid higher-tier regulatory testing, hence reducing the need for in vivo studies in accordance with the 3R principle.[18] Further, in vitro and in silico models can be used to build and confirm adverse outcome pathways (AOPs); multiple HPT-axis-related AOPs are already available at https://aopwiki.org/.[19] AOPs can serve as guidance for integrated testing and assessment strategies and enable the integration of in vivo and in vitro data.

A variety of in vitro methods for the evaluation of end points involved in thyroid hormone homeostasis have been reported in the scientific literature. Moreover, the Endocrine Disruptor Screening Program of the United States Environmental Protection Agency (U.S. EPA) has started high-throughput in vitro assays for key events in the regulation of thyroid hormone homeostasis and has fed their testing results into the Toxicity Forecaster (ToxCast) database.[20] Many of these high-throughput assays show high rates of positive outcomes. These are in part related to nonspecific effects such as cytotoxicity, protein synthesis inhibition, nonspecific enzyme inhibition, and others. For this reason, any compounds reported as active by these assays are generally subjected to testing in orthogonal assays.

In vitro data have been utilized to develop in silico models. For example, Rosenberg et al.[21] have developed QSAR models for predicting the interaction of substances with the TPO based on data obtained within ToxCast phase 1 and phase 2 (consisting of primarily pesticides and chemicals of research and regulatory interest) as well as E1K (such as chemicals of interest to the EPA's Endocrine Disruption Screening Program). Rosenberg et al. first built a model on the 1126 chemicals in the ToxCast phase 1 and 2 data sets and tested it on the ToxCast E1K data set (containing 771 compounds that are not included in the ToxCast phase 1 and 2 data sets), on which it obtained a balanced accuracy of 85%. In addition, the authors generated a classifier on the combined data set. This classifier obtained an averaged balanced accuracy of 83% during a five-time two-fold stratified cross-validation. Several QSAR models for predicting the binding affinity of small molecules to the TR have also been reported.[22−24]

The aim of this study was the development of a battery of machine learning models for the prediction of interactions of small molecules with proteins involved in molecular initiating events (MIEs) of thyroid hormone homeostasis, including the three DIOs (DIO1, DIO2, and DIO3), TPO, TR, NIS, TRHR, and TSHR. In addition to logistic regression (LR), random forest (RF), gradient boosting (XGB), support vector machine (SVM), and neural networks (NN) were explored as well as strategies for the generation of multi-task models.

The in silico approaches presented in this work could provide guidance in the assessment of the safety profiles of small molecules during early development phases. The models could also prove useful in mode of action prediction for endocrine disruptors.

■ MATERIALS AND METHODS

Data Sets. For DIO1, TPO, TR, NIS, TRHR, and TSHR, data sets with measured binary assay outcomes ("active", "inactive") were obtained from the ToxCast database[25] (Figure 1; Table 1). All these data sets have in common that they include at least 50 active compounds after data processing (see below for details on the data processing procedure). Binary activity labels were assigned according to the "hitc" value ("active" if the "hitc" value is one and "inactive" if it is zero; Table 2). The hitc value is calculated by fitting a curve to

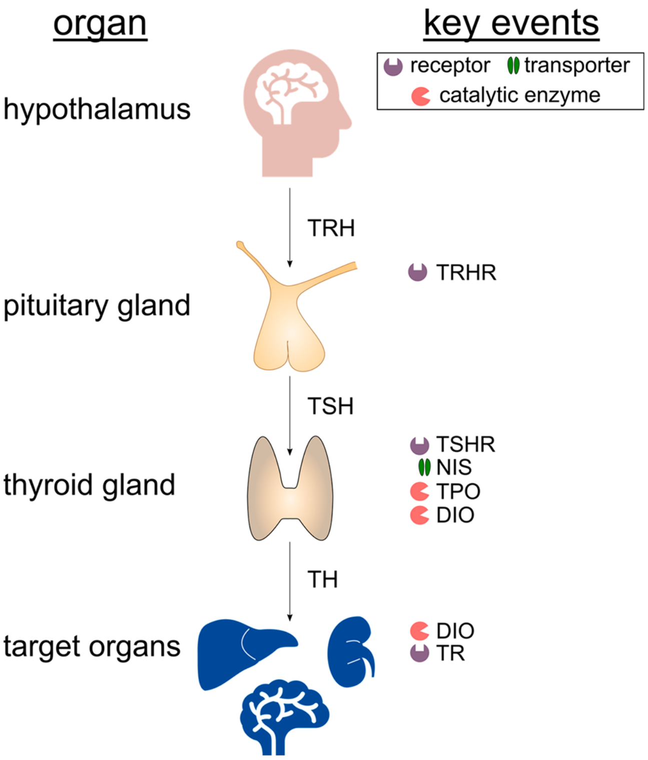


Figure 1. Overview of the protein families involved in MIEs of thyroid hormone homeostasis that are investigated in this work.

Table 1. Overview of the Modeled Assays and the ToxCast Compound Libraries the Assays Were Tested on

end point	assay model	assay description	ToxCast compound library tested on the assay	source of the assay data used in this work
deiodinase activity	recombinant DIO enzyme	measurement of DIO inhibition using an iodide release assay with recombinant DIO enzyme and quantification via Sandell–Kolthoff reaction	ToxCast phase 1_v2, phase 2, and e1k database	DIO1: NHEERL_MED_hDIO1_dn[a] + Olker et al., 2019 DIO2 + 3: Olker et al., 2019
thyroid peroxidase activity	rat thyroid microsomes	quantification of TPO inhibition via the oxidation of Amplex UltraRed in the AUR-TPO assay	ToxCast phase 1_v2, phase 2, and e1k database	NCCT_TPO_AUR_dn[a] + Friedman et al., 2016
thyroid hormone receptor modulation	GH3.TRE-LUC (rat pituitary tumor cell line transfected with the TR regulated luciferase reporter and TH response elements)	antagonistic modulation of TR binding measured via thyroid hormone-dependent luciferase expression	Tox21 compound library	TOX21_TR_LUC_GH3_Antagonist[a]
NIS-mediated iodide uptake	hNIS-HEK293T-EPA (hNIS transfected HEK293T-EPA cells)	quantification of NIS inhibition via radioactive iodide uptake	ToxCast phase 1_v2 and phase 2 database	NIS_RAIU_inhibition[a]
TRH receptor modulation	TRHR-HEK293 cells (TRHR transfected HEK-293 cells)	agonistic and antagonistic activation of the TRHR measured via quantification of intracellular Ca^{2+} concentration using a fluorescent dye	Tox21 compound library	TOX21_TRHR_HEK293_Antagonist[a]
TSH receptor modulation	HEK293-TSHR (TSHR transfected HEK-293 cells)	agonistic and antagonistic modulation of the G-protein coupled TSHR was measured through quantification of cAMP production as a marker for TSHR activation. cAMP production was quantified with Förster resonance energy using a competitive immunoassay to differentiate between naive and labeled cAMP	Tox21 compound library	TOX21_TSHR_Agonist_ratio[a] TOX21_TSHR_Antagonist_ratio[a]

[a] Data taken from the ToxCast database; the identifier indicates the assay name.

Table 2. Overview of the Data Sets Used for In Silico Model Development

		number of		
target abbreviation	assay name	active compounds	inactive compounds	ratio
DIO1	NHEERL_MED_hDIO1_dn	109	1610	1:15
DIO2	DIO2 inhibition	178	1551	1:9
DIO3	DIO3 inhibition	183	1545	1:8
TPO	NCCT_TPO_AUR_dn	256	796	1:3
TR	TOX21_TR_LUC_GH3_Antagonist	1251	5091	1:4
NIS	NIS_RAIU_inhibition	55	747	1:14
TRHR	TOX21_TRHR_HEK293_Antagonist	70	6548	1:94
TSHRAnt	TOX21_TSHR_Antagonist_ratio	116	6591	1:57
TSHRAg	TOX21_TSHR_Agonist_ratio	202	6587	1:33

concentration−response data and determining whether the minimum activity threshold, defined individually for each assay, was reached.[26]

For DIO1, TPO, and NIS, the ToxCast database only includes compounds that were tested in a multiconcentration assay (after they had previously been tested active in a single-concentration assay). Therefore, information on inactive compounds (these are the compounds that were tested negative in the single-concentration assay) was collected from the scientific literature (note that these works originate from the same lab as large parts of the ToxCast database). More specifically, data on 1678 compounds inactive on DIO1 were collected from Olker et al.,[27] data on 746 compounds inactive on TPO were collected from Friedman et al.,[28] and data on 663 compounds inactive on NIS were collected from Wang et al.[29]

For DIO2 and DIO3, all data used in this work were extracted from Olker et al. The data are derived with a colorimetric single-concentration assay measuring the release of iodide from the hormone substrate (at 200 μM concentration). Compounds inhibiting either deiodinase by at least 50% were then tested at multiple concentrations in the same assay setup. For the purpose of this study, binary activity labels were assigned according to the following rules: Any compounds with inhibition rates in the multiconcentration assay of 20% or higher were labeled as "active"; all other compounds, including those showing <50% inhibition in the single-concentration assay, were labeled as "inactive".

The compounds tested in the assays for the three DIOs, TPO, TR, NIS, TRHR, and TSHR are part of the Tox21 (Toxicology in the 21st Century program) and ToxCast (EPA's Toxicity Forecaster) projects. The Tox21 program is a collaboration between United States government agencies to develop high-throughput assays for the determination of adverse effects of small molecules on human health. The Tox21 library contains over 9000 substances, covering commercial chemicals, pesticides, food additives, and medical compounds. The ToxCast project is run by the U.S. EPA and has screened around 4500 substances in more than 700 high-throughput assays so far. The substances screened in the ToxCast project include not only a high number of pesticides but also food additives, pharmaceutical compounds, and cosmetics.[30] The ToxCast compound library has been built incrementally, by adding new subsets of compounds in each phase. For the assays considered in this work, different subsets of the ToxCast data sets or the complete Tox21 compounds library were tested in each assay (Table 1; see also the "Compound library" column in the Supporting Information Excel file).

The ToxCast database provides information (flags; see Table S1 for details) that can help in the identification of potentially false-positive and false-negative assay outcomes. For the seven data sets collected from the ToxCast database, data points tagged with any flag that indicate a potential quality issue were filtered out.

The results of confirmatory assays for TR and TSHR are also included in the ToxCast database and were used for refining the corresponding data sets with the following procedure: For the compounds tested in the confirmatory assay of TR ("TOX21_TR_LUC_GH3_Antagonist_Followup"),[31] the activity labels of the initial data set were corrected with the confirmatory assay information. For the TSHR end point, any compounds tested positive in an assay setup that lacks the TSHR reporter ("TOX21_TSHR_wt_ratio") were removed from the data (as positive results in this assay indicate that a compound's activity is not caused by a specific interaction with the TSHR; Figure 2; Table 3).

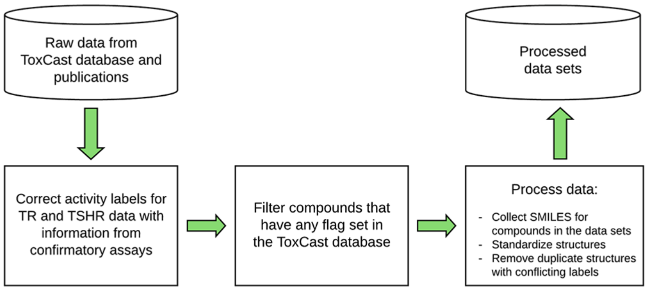

Raw data from ToxCast database and publications
Processed data sets
Correct activity labels for TR and TSHR data with information from confirmatory assays
Filter compounds that have any flag set in the ToxCast database
Process data:
- Collect SMILES for compounds in the data sets
- Standardize structures
- Remove duplicate structures with conflicting labels

Figure 2. Data processing workflow from the raw data to the final processed data sets used for modeling.

Table 3. Number of Compounds for Each Class at Different Steps in the Data Processing Workflow

	number of compounds					
	raw data		after filtering of compounds with any ToxCast flag		processed data sets used for model development	
end point	active	inactive	active	inactive	active	inactive
DIO1	136	1683	119	1683	109	1610
DIO2	194	1625	–[a]	–[a]	178	1551
DIO3	194	1625	–[a]	–[a]	183	1545
TPO	489	830	264	810	256	796
TR	2376	5929	1354	5574	1251	5091
NIS	282	756	55	756	55	747
TRHR	317	7554	81	7161	70	6548
TSHRAnt	336	7535	116	7206	116	6591
TSHRAg	489	7382	222	7192	202	6587

[a] Data not in the ToxCast database - no flag filtering step.

A "global thyroid toxicity" data set was generated by merging the nine data sets (see section Structure Preparation for details). This data set as well as the data source of each data point and the assay setup on which each compound was tested are provided as Supporting Information ("Complete data set" Excel sheet). Two complementary Excel sheets in the Supporting Information file report the data points filtered out due to a ToxCast flag ("Flag filtered compounds" Excel sheet) and the raw and standardized SMILES with the standardization steps applied on each compound ("Raw and standardized SMILES" Excel sheet).

The DrugBank,[32] containing a total of 11,355 approved, experimental, or withdrawn drugs, served as a reference data set to

Table 4. Overview of the Criteria Employed for Filtering Compounds for Cytotoxicity and of Resulting Data Set Compositions

			number of compounds after filtering cytotoxicity	
end point	data source	cytotoxicity filter	active	inactive
DIO1	ToxCast database	*Z*-score > 3	17	1610
DIO2	No data	–	–	–
DIO3	No data	–	–	–
TPO	Friedman et al.[28]	selectivity value > 1	188	796
TR	ToxCast database	TOX21_TR_LUC_GH3_Antagonist_viability hitc value = 1	422	5072
NIS	Wang et al.[29]	Hit2 value = 0	31	747
TRHR	ToxCast database	*Z*-score > 3	5	6552
TSHRAnt	ToxCast database	*Z*-score > 3	1	6593
TSHRAg	ToxCast database	*Z*-score > 3	41	6590

represent the drug-like chemical space. The EU CosIng database,[33] containing 1089 compounds, was utilized for the representation of the chemical space of cosmetic substances. Herbicides, insecticides, and fungicides were represented by all 522 compounds in the ChEMBL database[34] that have a mechanism of action classification assigned according to the Fungicide Resistance Action Committee (FRAC), Herbicide Resistance Action Committee (HRAC), or Insecticide Resistance Action Committee (IRAC) systems.

Data Sets Filtered for Cytotoxicity and Nonspecificity. In an attempt to further increase the quality of the data sets utilized for model development, any compound for which there was any data available suggesting that its measured activity could be related to cytotoxicity, the inhibition of cell growth or multiplication, or nonspecific protein inhibition was removed from the data sets. For the end points, for which these types of interference have been specifically studied and published (i.e., TPO, TR and NIS), the information was collected from the related publications (Table 4). For DIO1, TRHR, TSHRAnt, and TSHRAg, the *Z*-score from the ToxCast database, based on the AC_{50} of the assay of interest and of a cytotoxicity assay, was used for determining cytotoxicity. For DIO2 and DIO3, no information on the cytotoxicity of the compounds tested in these assays was identified. In the case of TPO, the selectivity value calculated by Friedman et al.[28] served as the criterion for identifying cytotoxic compounds and nonspecific inhibitors. Any compounds with a selectivity value below 1.0 were discarded. In the case of TR, cytotoxicity data were collected from the viability assay provided as part of the ToxCast database (assay "TOX21_TR_LUC_GH3_Antagonist_viability"). For NIS, the outcome of a cytotoxicity filter was obtained from the work of Wang et al.[29] In the case of DIO1, TRHR, and two TSHR end points, compounds with a *Z*-score from the ToxCast database lower than 3.0 were removed. With this information, data sets containing only compounds that did not show any interference were compiled for DIO1, TPO, TR, NIS, TRHR, TSHRAnt, and TSHRAg. This data set is also provided as Supporting Information ("Filtered data set" Excel sheet). In the Supporting Information, filtered out compounds in this data set are tagged with the label "filtered out data point", and the data source for the filtering is indicated as well.

Software and Hardware Setup. All calculations were performed on Linux workstations running Red Hat Enterprise 7.8 and equipped with Intel Xeon Gold 6136 processors (3.00 GHz) and 64 GB of main memory.

KNIME[35] was used for the preparation of the structures (with the ChemAxon Standardizer[36] and RDKit Canon SMILES[37] nodes) and descriptor calculation (RDKit Count-Based Fingerprint and RDKit Descriptor calculation nodes). The principal component analysis (PCA) as well as model training and evaluation were performed in Python with the packages scikit-learn[38] and Keras.[39]

Structure Preparation. The molecules tested in one or several of the nine assays (including those assays not included in the ToxCast database) originate from one or more chemical libraries compiled within the ToxCast program (phases 1, 2, and 3). The SMILES strings for these compounds were obtained from the ToxCast database, where available. In the absence of such information, the NCI/CADD Chemical Identifier Resolver[40] was queried with the CAS number instead. Ultimately, for compounds without a match, the "RDKit from IUPAC" node of RDKitin KNIME was used to try to derive a structure from the chemical name.

All structures in the modeling data sets were processed and standardized with the ChemAxon Standardizer node in KNIME. More specifically, the tool was used for removing solvents, stripping salts, detecting and annotating aromaticity, removing stereochemical information, neutralizing charges, mesomerizing structures, and removing small fragments. Canonical SMILES were derived from the standardized molecules with RDKit (with default parameters) and used for deduplication. Duplicate compounds with conflicting activity labels for an assay were removed. The global thyroid toxicity data set, generated by merging the nine end-point-specific data sets based on the previously generated canonical SMILES, consists of 8001 substances.

Descriptor Calculation. Count-based Morgan fingerprints with a radius of 2 bonds and a length of 2048 bits were calculated with the "RDKit Count-Based Fingerprint" node of RDKit in KNIME. In addition, all 119 one-dimensional (1D) and two-dimensional (2D) physicochemical property descriptors implemented in the "RDKit Descriptor Calculation" node were computed, which describe, among other properties, the number of particular types of atoms, the numbers of bonds and rings in a molecule, as well as polarity and solubility. Prior to model building, the 1D and 2D descriptors were subjected to *Z*-score normalization using the "Normalizer" node in KNIME. Descriptors for which no variance was observed for the global thyroid data set were removed.

Chemical Space Analysis. Dimensionality reduction was performed on the global thyroid data set with the PCA implementation of scikit-learn, based on a subset of 23 physically meaningful and interpretable molecular descriptors generated with RDKit (Table S2).

Machine Learning Methods. Five machine learning approaches for classification were explored: LR, RF, XGB, SVM, and NN. LR classification models employ a mathematical function that is a linear combination of one or more independent variables. RF is an ensemble learning method that utilizes a multitude of decision trees for making predictions. The XGB algorithm makes decisions based on an ensemble of decision trees, too, with the special feature that each new tree is designed to correct the mistakes made by the previous one. SVMs project the features into a hyperplane that maximizes the distance to each class point in space and which then acts as the decision boundary. Multilayer perceptron NN are formed by nodes, or so-called "neurons", located in different interconnected layers. Information is transferred back and forth between layers to update the functions in the neurons, with the objective of minimizing the error between the correct class and the prediction.

The NN models were generated with Keras, and all other types of models were implemented with scikit-learn in Python. The optimization of hyperparameters (Table 5) was performed during a grid search within a 10-fold cross-validation framework. The F1 score was used as the optimization criterion.

Table 5. Overview of Hyperparameters Applied for Each Method

method	hyperparameters[a]	values[b]
logistic regression	C	0.7, 0.8, 1
random forest	number of estimators	500, 1000
	min_samples_leaf	1, 2
gradient boosting	estimators	500, 1000
support vector machine	C	0.01, 1, 10
	gamma	scale, auto
neural network	number of layers	3
	neurons	(4000, 1000, 1), (1000, 500, 1)
	dropout rate	0, 0.3
	learning rate	0.001, 0.0001

[a] Hyperparameters for which the default values were preserved are not reported. [b] A grid search was conducted to identify the optimum value for parameters for which more than one value is reported in this table; otherwise, the value was fixed.

Generation and Evaluation of Single-Task Models. To address data imbalance (excess of inactive compounds in this case), weight balancing, undersampling, and oversampling techniques were explored.

For the weight balancing approach, balanced weights for the active and inactive classes were calculated with scikit-learn and employed in combination with the ML methods: RF, LR, SVM, and NN. For XGB, balanced weights were not used, as the method itself is designed to deal with class imbalance by successively constructing training sets with misclassified examples.

An inner 10-fold cross-validation (CV) was applied for hyperparameter selection, and an outer 10-fold CV was applied for performance assessment.

For the undersampling approach, the following workflow was developed, which generates an ensemble of models built on different training sets (Figure 3):

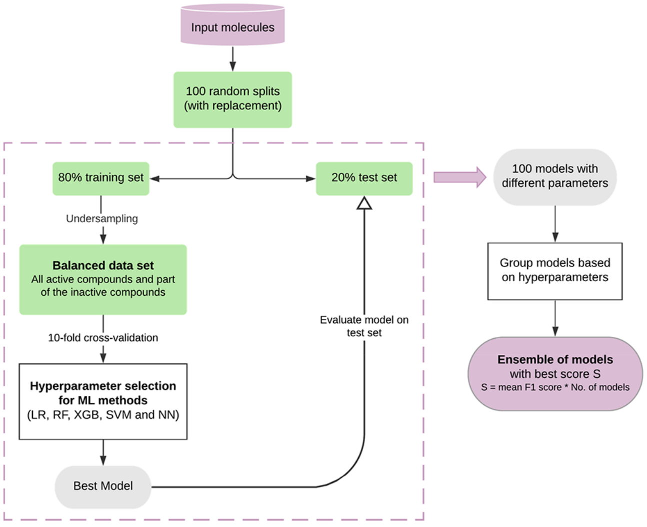

Input molecules
100 random splits (with replacement)
80% training set
20% test set
Undersampling
Balanced data set
All active compounds and part of the inactive compounds
10-fold cross-validation
Hyperparameter selection for ML methods (LR, RF, XGB, SVM and NN)
Best Model
Evaluate model on test set
100 models with different parameters
Group models based on hyperparameters
Ensemble of models with best score S
S = mean F1 score * No. of models

Figure 3. Workflow for generating and testing models based on training sets balanced by undersampling. The hyperparameters of the ML models are optimized during a grid search within a 10-fold CV framework. The performance of the resulting best model is evaluated on the test set. The result of the workflow is an ensemble of models with optimized hyperparameters for each method.

(1) Preparation of the data sets: The data were divided into a training set (80%) and a test set (20%). To evade class imbalance, the number of inactive compounds (majority class) in the training set was reduced by random selection, while all active compounds were retained. For data sets with an active-to-inactive ratio of <1:10, the ratio was changed to 2:3. For data sets with an active-to-inactive ratio ≥1:10, the ratio was changed to 1:2 (Table 6).

Table 6. Composition of the Training Sets after Undersampling

	number of		
end point	active compounds	inactive compounds	ratio of active and inactive compounds
DIO1	87	147	1:2
DIO2	142	213	2:3
DIO3	146	219	2:3
TPO	205	307	2:3
TR	1001	1501	2:3
NIS	44	88	1:2
TRHR	56	112	1:2
TSHRAnt	93	186	1:2
TSHRAg	162	324	1:2

(2) Hyperparameter optimization: Hyperparameter optimization was performed on the resampled data sets within a 10-fold CV framework. The 10 models obtained from the CV were grouped based on the selected hyperparameter values to calculate the mean F1 score for each hyperparameter set. The best model of the group with the highest mean value was selected and subsequently evaluated on the test set.

(3) Generation of the final ensemble of models: By repeating this workflow 100 times, an ensemble of 100 models, trained on different balanced data sets, was obtained for each method. In order to determine the best overall hyperparameters for the end point classification problem and ensure model robustness, the 100 models were grouped according to their hyperparameter values, and the best ensemble of models was chosen as the predictive model. The selection of the best ensemble is based on a score calculated as the mean F1 score plus the number of models in the ensemble.

For the oversampling approach, the SMOTENC[41] method was employed. Molecular fingerprints were defined as categorical features, and the "sampling strategy" parameter, which defines the resulting ratio between the minority and majority class, was set to 0.7. The RF, LR, XGB, SVM, and NN models were trained on these oversampled data sets, with an inner 10-fold CV for hyperparameter selection and an outer 10-fold CV for performance estimation.

Generation and Evaluation of Multi-task Models. A multi-task model was generated based on the global thyroid toxicity data set. Additional three multi-task models were generated from subsets of the global data set that include only a subset of end points. All models were derived with multilayer perceptron NNs with a shared architecture for all tasks. Only the output layer is independent for each learned task. Missing values in the training set (related to the fact that not all compounds have been tested in all assays) were not considered during model training and evaluation by masking (i.e., ignoring) them during the loss and performance calculation. Class imbalance was addressed by balancing the class weights for the loss calculation based on the active-to-inactive ratio in the training set. A workflow similar to the one used for the single-task models (but skipping the undersampling step) was employed to derive an ensemble of models (Figure 3). A grid search for hyperparameter optimization was carried out within a 10-fold CV framework (Table 7), and four combinations of assay end points were evaluated. The combinations covered two to nine end points, starting with TPO and TR, and incrementally adding (i) the three DIOs, (ii) NIS and

Table 7. Overview of Combinations of Hyperparameters Explored

number of layers	parameter	values[a]
4	neurons	(8000, 4000, 1000, X), (4000, 2000, 500, X)
	dropout rate	0, 0.3
	regularizer rate	0.000001
	learning rate	0.0001
5	neurons	(9000, 4000, 1000, 100, X), (5000, 2000, 1000, 100, X)
	regularizer rate	0, 0.0000001
	learning rate	0.0001

[a] "X" in the number of neurons denotes the number of end points employed for each multi-task model (i.e., number of neurons in the output layer).

TRHR, and (iii) both TSHR end points. In the case of the multi-task models, the performance was evaluated and optimized on the mean F1 score among all end points included in the model.

Metrics for Model Performance Evaluation. Six different metrics were employed for the evaluation of model performance:

(1) Precision: measures the proportion of true positive predictions out of all positive predictions (eq 2).
(2) Recall: measures the proportion of correctly identified positive samples (eq 3).
(3) F1 score: is the harmonic mean of precision and recall (eq 4). It is robust against data imbalance.
(4) Matthews correlation coefficient (MCC): considers all four classes of predictions (true positive, true negative, false positive, and false negative predictions; eq 5). MCC values range from −1 to +1, with a value of +1 indicating perfect prediction. The metric is robust against data imbalance.
(5) Balanced accuracy: quantifies the average recall obtained for each class and, therefore, is robust against data imbalance (eq 6).
(6) Area under the receiver operating curve (AUC): is a measure of the ability of a model to distinguish between positive and negative samples. The AUC is calculated as the bidimensional area under the receiver operating curve (eq 7).

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2}$$

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3}$$

$$\text{F1 score} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{4}$$

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}} \tag{5}$$

$$\text{balanced accuracy} = \frac{\left(\frac{\text{TP}}{P} + \frac{\text{TN}}{N}\right)}{2} \tag{6}$$

$$\text{AUC} = \int_{x=0}^{1} \frac{\text{TPR}}{\text{FPR}(x)} dx \tag{7}$$

where FN is false negatives, FP is false positives, TN is true negatives, TP is true positives, FPR is false positive rate, and TPR is true positive rate.

RESULTS AND DISCUSSION

In this study, five machine learning methods (RF, LR, XGB, SVM, and NN) were employed with the aim to develop predictive classifiers for nine end points involved in thyroid hormone homeostasis: DIO1, DIO2, DIO3, TPO, TR, NIS, TRHR, TSHRAnt, and TSHRAg. Because of a lack of active compounds across all training sets (the active class represents only 1 to 32% of the training data), a weight balancing approach, an undersampling method, and an oversampling strategy were explored. In addition, the use of multi-task models was investigated as a possible avenue to obtain better performing and more widely applicable in silico models.

Chemical Space. The chemical space represented by the training data defines the applicability domain of a model. An in-depth analysis of the composition and properties of the ToxCast and Tox21 data sets was conducted by Richard et al.[30] In their work, Richard et al. describe how the chemicals included in the ToxCast data sets were selected (e.g., compounds with available in vivo toxicity results, donated by pharmaceutical companies, or known endocrine disruptors) and how this selection yielded a high chemical structure diversity and a broad chemical property coverage.

Table 8. Percentage of Compounds in the Reference Data Sets Covered by a Compound in the End-Point-Specific Data Sets at the Given Tanimoto Similarity Thresholds

		end point								
	Tanimoto similarity	DIO1	DIO2	DIO3	TPO	TR	NIS	TRHR	TSHRAnt	TSHRAg
% coverage pesticides	1.0	57	57	57	47	56	35	65	65	70
	≥0.8	58	58	58	48	57	36	66	65	71
	≥0.6	67	68	68	58	68	47	76	76	79
	≥0.4	84	84	84	78	85	68	87	89	90
	≥0.2	99	99	99	98	99	98	99	99	99
% coverage cosmetics	1.0	16	16	16	9	37	7	39	40	39
	≥0.8	20	20	20	11	41	9	43	44	43
	≥0.6	34	34	34	19	58	17	59	60	59
	≥0.4	69	70	70	52	98	49	86	86	86
	≥0.2	95	95	95	92	98	92	98	98	98
% coverage drugs	1.0	5	5	5	3	22	2	20	20	20
	≥0.8	5	5	5	3	24	3	22	22	22
	≥0.6	10	10	10	7	37	6	33	34	33
	≥0.4	28	29	29	22	62	20	60	60	60
	≥0.2	95	95	95	93	98	91	98	98	98

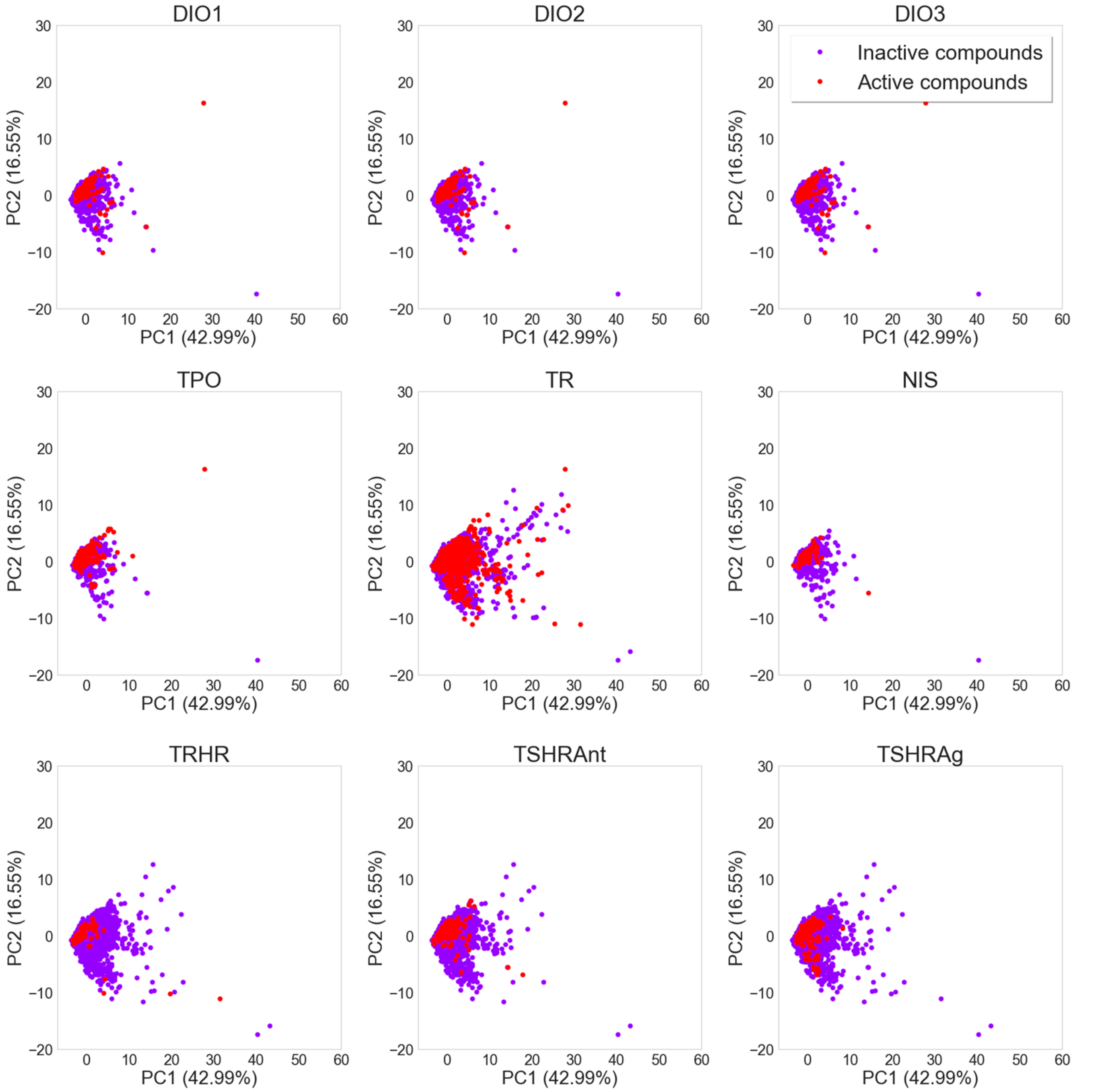


Figure 4. PCA based on a selection of interpretable molecular descriptors generated with the RDKit for the end-point-specific data sets. Active compounds are colored in red and inactive compounds in purple. The shift of the active compounds toward higher values on the *y*-axis is mainly due to a high number of aromatic rings.

In order to determine the relevance of the data employed in this study, we compared the chemical space covered by our global thyroid toxicity data set (containing measured data on the nine modeled thyroid end points for 8001 compounds) as well as the end-point-specific data sets to the chemical space covered by pesticides (all compounds in ChEMBL that are linked with the HRAC, IRAC or FRAC systems), cosmetic substances (from the EU CosIng database), and drugs (from DrugBank). We found that the global thyroid toxicity data set covers pesticides (coverage 78%) better than cosmetic substances (39%) and drugs (25%). Analysis of the end-point-specific data sets shows that at least 47% of all agrochemicals are represented by training set compounds with a Tanimoto coefficient (based on Morgan2 fingerprints) of 0.6 or higher (Table 8). For cosmetics and drugs, this percentage is only 17% and 6%, respectively. Only in the case of TR, TRHR, and both TSHR end points, the coverage of cosmetics and drugs is higher (58% and 33%; at a similarity threshold of 0.6). The higher coverage is related to the fact that the size of the training sets for these end points is much larger and that the compounds tested in these assays include the Tox21 compound library, which has a higher percentage of cosmetics and drugs.

PCA scatter plots derived from the global thyroid toxicity data set using physically meaningful and interpretable molecular descriptors (Figure 4) show a strong overlap of the areas most densely populated by the active and inactive compounds of any of the target-specific subsets. A small number of outliers is observed for any of the data sets. These are mostly macrocyclic molecules or large compounds with a high number of rings. For most end points, the active compounds tend to have high values in the second component of the PCA (*y*-axis), which are primarily a result of high numbers of aromatic rings.

Molecular diversity within the end-point-specific data sets was analyzed with plots of the pairwise similarities (based on atom-pair fingerprints)[42] among (a) all pairs of active compounds, (b) all pairs of inactive compounds, and (c) all pairs consisting of one active and one inactive compound. The distribution of similarities among these three sets of compounds is comparable and shows a tailing toward small similarities (examples for DIO1 and TPO are shown in Figure 5 and are representative of all targets; the figures for all other investigated targets are provided in Figure S1). This analysis confirms the high molecular diversity of the compounds included in the data sets, as it was also concluded by Richard et al.[30] Note that the distribution of pairwise similarities among the active compounds is comparable to the distribution of pairwise similarities between the active and inactive compounds.

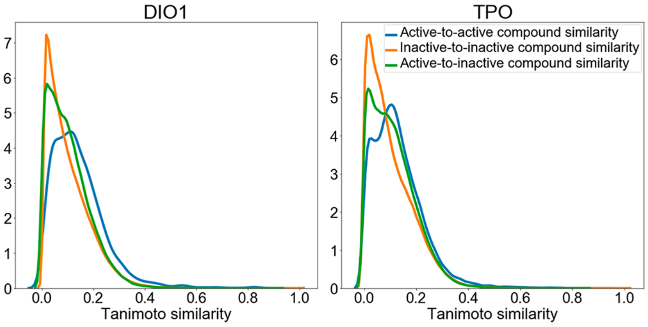


Figure 5. Examples of the distribution of pairwise Tanimoto similarities based on atom-pair fingerprints for three types of compound pairs: (a) active-to-active, (b) inactive-to-inactive, and (c) active-to-inactive. The distributions for all other end-point-specific data sets are provided in Figure S1.

To further analyze the chemical diversity of the data sets, we calculated the number of distinct Murcko scaffolds in each end-point-specific data set and in the global thyroid toxicity data set. Additionally, also the number of compounds without a Murcko scaffold (i.e., without a ring system) and the number of compounds with a unique scaffold (defined as the sum of compounds with a unique Murcko scaffold and compounds without Murcko scaffold) were calculated (Table 9). From this analysis, it can be seen that there is a high number of distinct scaffolds in the data sets (between 330 distinct Murcko scaffolds for NIS and 2327 for the global data set) and that around half of the compounds have a unique scaffold (between 45% for the global data set and 61% for the NIS data set).

The relationship between specific chemical groups and active compounds for the different assays was analyzed by searching the list "SMARTS Patterns for Functional Group Classification"[43] distributed by Open Babel,[44] which contains 309 SMARTS patterns, in the respective inactive and active compounds of each data set. The number of hits per class was analyzed, and a ratio, defined as the number of hits in active compounds divided by the number of hits in inactive compounds, was calculated. Only functional groups with ratios >1.7 were considered. The total number of hits was also taken into account, and only functional groups found in at least six compounds were regarded. Following these criteria, only for the TPO and TR end points, a relationship between some functional groups and active compounds could be established. Compared to inactive compounds, a high proportion of active compounds for TPO have at least one primary aromatic amine, phenol, sulfenic derivative, enol, thiourea, vinylogous acid, and phosphoric acid derivative (Table 10). Among the compounds active on TR organometallic compounds, diarylthioethers and enamine groups are over-represented.

Table 9. Number of Distinct Murcko Scaffolds and Compounds without a Ring System

end point	number of distinct Murcko scaffolds	number of compounds without ring systems	percentage of unique scaffolds[a]
DIO1	554	455	53%
DIO2	557	456	52%
DIO3	557	456	52%
TPO	418	231	55%
TR	1877	1608	48%
NIS	330	202	61%
TRHR	1810	1712	47%
TSHRAnt	1834	1733	47%
TSHRAg	1876	1728	47%
global data set	2327	1871	45%

[a] Unique scaffolds are defined as the sum of compounds with unique Murcko scaffold and compounds without Murcko scaffold.

Single-Task Classification Models. For each of the nine thyroid-related end points, the data obtained from the ToxCast database and relevant publications were employed for training and evaluation of single-task classification models (see Methods for details). The models were developed based on molecular fingerprints and physicochemical descriptors. All possible combinations of the five ML algorithms and three data balancing techniques were explored.

The performance of the models based on any of the five ML algorithms was in general very similar. For example, the maximum difference in the F1 scores observed among ML algorithms in combination with the oversampling approach was no higher than 0.10 (maximum difference observed for the NIS end point, with F1 scores of 0.70 and 0.60 for the LR and RF models, respectively).

The impact of the data balancing approach on model performance was also, in general, small. The largest differences in the mean F1 scores for different balancing approaches among the ML models for the same end point were between 0.02 (for TR) and 0.19 (for TRHR) (see Figure 6 for a comparison of the F1 scores obtained by the RF models; the figures for all other models are provided in Figure S2). However, a tendency for ML models to perform best when trained on oversampled data was observed. The maximum difference in F1 scores between a ML method trained on oversampled data and one trained on undersampled or imbalanced data (using weight balancing) was −0.23 (for the TRHR model with SVM in combination with undersampling). Only in one case, which is the RF model for NIS, the model based on undersampled data performed favorably to the model based on oversampled data (F1 score 0.66 vs 0.60). The biggest differences related to data sampling were observed for the TRHR and the two TSHR end points, for which the undersampling approach yielded up to 0.24 lower mean F1 scores than the other two sampling approaches. The reason for

Table 10. Number of Hits of Functional Groups in the Inactive and Active Compounds of the Data Sets

	SMARTS hits for the functional groups (inactive:active compounds)[a]									
end point	primary aromatic amine	organometallic compounds	phenol	sulfenic derivative	diarylthioether	enol	enamine	thiourea	vinylogous acid	phosphoric acid derivative
TPO	20:40	–	48:95	6:15	–	1:6	–	1:8	10:20	13:27
TR	–	7:27	–	–	9:49	–	7:20	–	–	–

[a] Only values with ratios (active/inactive compounds) >1.7 and with at least 6 hits in the active compounds are shown.

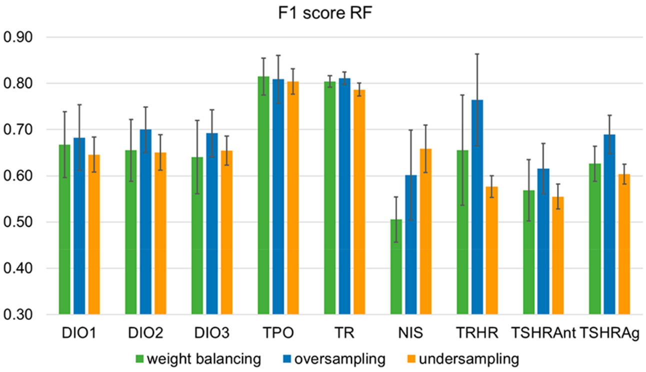


Figure 6. Comparison of the mean F1 score obtained with the RF method in combination with the different data sampling techniques (weight balancing, oversampling, and undersampling) for the nine thyroid end points.

this is likely the extreme imbalance of the training sets for these three end points, with only 1–3% of active compounds. Undersampling in these cases leads to a substantial loss of information on inactive compounds, which is otherwise preserved. However, the gain in performance related to oversampling comes at the cost of an increased standard deviation across models trained on different splits of the data.

Because of the overall favorable performance of models trained on oversampled data, further discussion focuses on these models. Unless stated otherwise, all results refer to mean values obtained by 10-fold cross-validation. Although the main text only discusses the F1 score results, MCC values, balanced accuracies, and AUC values are also provided in Table 11.

The classification models derived for DIO1, DIO2, and DIO3 all showed comparable performance, with mean F1 scores ranging from 0.67 to 0.71, depending on the ML method used (Table 11). Within the individual end points, the largest difference in F1 scores between ML methods was just 0.04. SVM produced the best model for DIO1 (mean F1 score of 0.71) and DIO2 (mean F1 score of 0.71), whereas NN worked best for DIO3 (mean F1 score of 0.71).

The models for the TPO and TR end points yielded mean F1 scores between 0.77 (for TR with LR and SVM) and 0.83 (for TPO with XGB). The best-performing algorithm for TPO was XGB (mean F1 score of 0.83), while RF performed best on the TR data set (mean F1 score of 0.81). For the NIS models, the mean F1 scores ranged from 0.60 (with RF) to 0.70 (with LR). Linear models (LR and SVM) outperformed decision trees (RF and XGB) and NNs on this data set, with up to 0.10 higher F1 scores. The standard deviation of the F1 score among the 10-fold CV models ranges from 0.08 to 0.10 with the different algorithms. The high standard deviation may be related to an overfitting of models as a result of the low number of active compounds in the training set (only 55 active compounds and 747 inactive compounds).

It should also be considered that the data sets for the DIOs, TPO, and NIS include data measured in single- and multiconcentration assays. The class labels for the single-concentration results were set considering the inhibition cutoff of 50%, while for the multiconcentration results, the class labels were derived from a more precise curve fitting on the concentration–response data (for DIO1, TPO, and NIS) or from an inhibition cutoff of 50% (for DIO2 and DIO3; see Materials and Methods for details). The combination of these two types of data may increase the uncertainty of the models and result in lower performance. This difference in the cut-offs for the multiconcentration results may also be the reason why for DIO2 and DIO3, a higher percentage of active compounds does not seem to be beneficial to model performance when compared to DIO1. Although for the latter end point the number of active compounds is lower, they were identified by curve fitting instead of the fixed 50% inhibition threshold applied for DIO2 and DIO3, which may cause a higher number of false positives. In the case of TPO, the better performance of the models could be explained, to some extent, by the fact that the active compounds were also derived from the concentration–response curve and that the percentage of active compounds is higher. Similar causes could explain the performance of the TR models, for which all data was derived from the multiconcentration assays and which has a higher percentage of active compounds.

For the TRHR end point, where the number of active compounds is also small (70 active compounds and 6545 inactive compounds), the standard deviation of the mean F1 score was between 0.08 and 0.12. However, the mean F1 scores were higher than for NIS and ranged from 0.68 (with XGB) to 0.77 (with SVM). The mediocre results and the variability of these models may be caused by the assay design itself. In this assay, the activity of compounds against this receptor is derived from the concentration of intracellular calcium as a marker of GPCR activation (via fluorescence) and is thus prone to interference, for example, by any alteration of intracellular calcium or autofluorescence.

The outcomes of the two TSHR assays were predicted with mean F1 scores ranging from 0.60 (for TSHRAnt with LR) to 0.69 (for TSHRAg with RF). For TSHRAnt, the best results (mean F1 score of 0.65) were obtained with NN, whereas for TSHRAg, the best results (mean F1 score of 0.69) were obtained with RF. An important limitation of the data used for model development is related to assay technology, which employs fluorescent antibodies coupled to a second messenger to derive the activity of the compounds against TSHR.[45] Since this second messenger is nonspecific and may be activated via several pathways, and fluorescence measurements may be positive due to fluorescent compounds and dyes, the false-positive rate in the data may be substantial.

Overall, the presented models could contribute to the first level of the OECD approach for the evaluation of potential endocrine disruptors, by making available models for an initial

Table 11. Mean and Standard Deviation of the Performance of Different Methods for All Modeled End Points

end point	method	F1 score	MCC	balanced accuracy	AUC
DIO1	RF	0.68 (± 0.07)	0.41 (± 0.15)	0.64 (± 0.06)	0.87 (± 0.04)
	LR	0.68 (± 0.06)	0.37 (± 0.12)	0.67 (± 0.07)	0.83 (± 0.08)
	XGB	0.70 (± 0.08)	0.45 (± 0.15)	0.66 (± 0.07)	0.84 (± 0.05)
	SVM	**0.71 (± 0.07)**	**0.44 (± 0.15)**	**0.68 (± 0.06)**	**0.86 (± 0.04)**
	NN	0.70 (± 0.05)	0.43 (± 0.10)	0.67 (± 0.06)	0.86 (± 0.08)
DIO2	RF	0.70 (± 0.05)	0.43 (± 0.11)	0.66 (± 0.04)	0.85 (± 0.06)
	LR	0.67 (± 0.05)	0.35 (± 0.10)	0.67 (± 0.05)	0.81 (± 0.05)
	XGB	0.70 (± 0.06)	0.41 (± 0.11)	0.67 (± 0.05)	0.81 (± 0.06)
	SVM	**0.71 (± 0.04)**	**0.43 (± 0.09)**	**0.68 (± 0.04)**	**0.84 (± 0.04)**
	NN	0.69 (± 0.05)	0.39 (± 0.10)	0.67 (± 0.04)	0.82 (± 0.05)
DIO3	RF	0.69 (± 0.05)	0.41 (± 0.10)	0.66 (± 0.05)	0.85 (± 0.04)
	LR	0.70 (± 0.05)	0.39 (± 0.09)	0.69 (± 0.04)	0.82 (± 0.05)
	XGB	0.69 (± 0.05)	0.40 (± 0.11)	0.67 (± 0.05)	0.82 (± 0.06)
	SVM	0.68 (± 0.04)	0.38 (± 0.08)	0.66 (± 0.04)	0.85 (± 0.04)
	NN	**0.71 (± 0.05)**	**0.42 (± 0.11)**	**0.68 (± 0.05)**	**0.85 (± 0.06)**
TPO	RF	0.81 (± 0.05)	0.63 (± 0.10)	0.79 (± 0.05)	0.91 (± 0.04)
	LR	0.80 (± 0.06)	0.60 (± 0.12)	0.80 (± 0.07)	0.88 (± 0.05)
	XGB	**0.83 (± 0.04)**	**0.67 (± 0.09)**	**0.82 (± 0.05)**	**0.90 (± 0.04)**
	SVM	0.80 (± 0.05)	0.60 (± 0.10)	0.80 (± 0.05)	0.88 (± 0.05)
	NN	0.82 (± 0.04)	0.64 (± 0.08)	0.81 (± 0.04)	0.90 (± 0.04)
TR	**RF**	**0.81 (± 0.01)**	**0.62 (± 0.03)**	**0.80 (± 0.01)**	**0.92 (± 0.01)**
	LR	0.77 (± 0.02)	0.54 (± 0.04)	0.76 (± 0.02)	0.87 (± 0.03)
	XGB	0.80 (± 0.02)	0.61 (± 0.04)	0.79 (± 0.02)	0.91 (± 0.02)
	SVM	0.77 (± 0.04)	0.54 (± 0.09)	0.75 (± 0.04)	0.87 (± 0.05)
	NN	0.79 (± 0.01)	0.59 (± 0.02)	0.77 (± 0.02)	0.89 (± 0.02)
NIS	RF	0.60 (± 0.10)	0.23 (± 0.20)	0.58 (± 0.07)	0.86 (± 0.10)
	LR	**0.70 (± 0.08)**	**0.41 (± 0.16)**	**0.68 (± 0.06)**	**0.86 (± 0.08)**
	XGB	0.66 (± 0.09)	0.32 (± 0.19)	0.63 (± 0.07)	0.82 (± 0.11)
	SVM	0.68 (± 0.08)	0.40 (± 0.15)	0.66 (± 0.08)	0.84 (± 0.10)
	NN	0.66 (± 0.10)	0.32 (± 0.20)	0.64 (± 0.09)	0.81 (± 0.12)
TRHR	RF	0.76 (± 0.10)	0.58 (± 0.17)	0.70 (± 0.09)	0.91 (± 0.05)
	LR	0.72 (± 0.09)	0.46 (± 0.18)	0.69 (± 0.09)	0.86 (± 0.07)
	XGB	0.68 (± 0.08)	0.39 (± 0.15)	0.66 (± 0.10)	0.84 (± 0.14)
	SVM	**0.77 (± 0.11)**	**0.57 (± 0.22)**	**0.73 (± 0.11)**	**0.90 (± 0.03)**
	NN	0.72 (± 0.12)	0.45 (± 0.25)	0.69 (± 0.13)	0.83 (± 0.07)
TSHRAnt	RF	0.62 (± 0.05)	0.30 (± 0.13)	0.58 (± 0.04)	0.87 (± 0.06)
	LR	0.60 (± 0.06)	0.22 (± 0.14)	0.58 (± 0.04)	0.78 (± 0.09)
	XGB	0.63 (± 0.06)	0.28 (± 0.15)	0.60 (± 0.04)	0.82 (± 0.06)
	SVM	0.63 (± 0.06)	0.32 (± 0.15)	0.59 (± 0.05)	0.82 (± 0.07)
	NN	**0.65 (± 0.06)**	**0.32 (± 0.13)**	**0.62 (± 0.05)**	**0.76 (± 0.08)**
TSHRAg	**RF**	**0.69 (± 0.04)**	**0.44 (± 0.08)**	**0.63 (± 0.03)**	**0.89 (± 0.03)**
	LR	0.66 (± 0.06)	0.34 (± 0.13)	0.62 (± 0.06)	0.80 (± 0.06)
	XGB	0.67 (± 0.05)	0.36 (± 0.11)	0.63 (± 0.04)	0.83 (± 0.04)
	SVM	0.66 (± 0.04)	0.38 (± 0.07)	0.62 (± 0.03)	0.82 (± 0.04)
	NN	0.68 (± 0.06)	0.37 (± 0.11)	0.64 (± 0.05)	0.79 (± 0.07)

screen to detect the interaction of small molecules with key targets related to thyroid hormone homeostasis. Moreover, the models could help to build or confirm HPT-axis related AOPs.

However, it is important to highlight the intrinsic nature of the modeled assays. These are high-throughput in vitro assays, which usually show high rates of (false) positive outcomes due to interferences, as shown by Paul-Friedman et al.[31] for the case of TR. Therefore, compounds showing activity in these assays should be tested in orthogonal assays, and the same principle should be applied to the presented models.

In-Depth Analysis of Model Performance and Prediction Reliability. Among all end points investigated, the best models were obtained for TPO and TR. As these well-performing models will be of primary relevance to investigators, we conducted additional analyses with them in order to gain an in-depth understanding of model performance and the reliability of predictions. Since all algorithms showed a similar performance on TPO and TR, the analysis is exemplified for the RF models in combination with oversampling, which obtained a mean F1 score of 0.81 for both end points during 10-fold CV.

First, we investigated how the distance of the prediction probability to the decision boundary relates to the reliability of a prediction. More specifically, we gradually reduced the coverage of the model by removing compounds from the test set which are predicted with probabilities close to the decision threshold, starting with those closest to the boundary (Figure 7). For both the TPO and TR models, the F1 scores increased

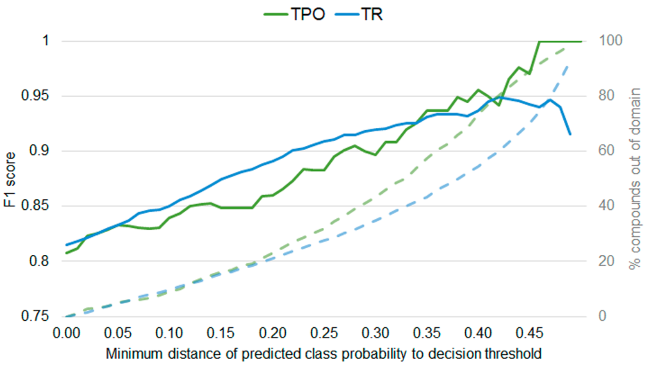


Figure 7. Changes in the F1 score (solid lines) and coverage (dashed lines), as compounds with predicted probabilities close to the decision boundary for the RF model of TPO (green) and TR (blue) were considered out of the applicability domain and removed.

as more compounds close to the decision boundary were removed, indicating that there was a higher rate of wrong predictions among compounds closer to the cutoff. We also investigated the number of compounds that are not covered by the model, as we increase the minimum distance to the decision threshold. When excluding around 20% of the test compounds, the TPO model had an F1 score of 0.86 (+0.05) and the TR model an F1 score of 0.89 (+0.07). Reducing the coverage of the model to those compounds predicted with high confidence could therefore increase the validity of the model.

The similarity of the query compounds to the training data can be decisive for prediction success. To determine how this affects model performance, for each compound in the test set, the (average) Tanimoto similarity of the ECFP fingerprint to the one, three, and five nearest neighbors in the training set was calculated. For both the TPO and TR end points, a linear relationship between the similarity of the compounds and the F1 score was observed, consistent when considering different numbers of nearest neighbors (Figure 8). For the TPO model, the F1 score was 0.21 points higher for compounds that are similar to the training data (Tanimoto similarity higher than 0.8) than for compounds that are not represented by structurally related molecules in the training data (Tanimoto similarity lower than 0.2) when considering one nearest neighbor. For the TR model, this difference was 0.18 points. Determining the similarity of new compounds to those in the training sets can therefore help to estimate the reliability of the predictions.

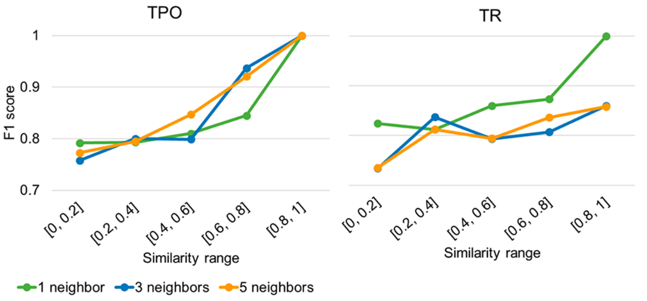


Figure 8. F1 scores as a function of the Tanimoto similarity between the compounds in the test set and in the training set. The similarity was calculated based on the ECFP fingerprint between one, three, or five nearest neighbors.

Single-Task Models Generated from Filtered Data Sets. The data modeled so far originate from high-throughput screening assays and are therefore often error-prone. False positive outcomes may occur if, for instance, a nonspecific interaction between a compound and a protein is measured, or if a compound is falsely perceived as active due to its cytotoxicity. On the other hand, false negative outcomes may be caused by the volatility or low solubility of compounds, which reduces their concentration in the assay sample. In some cases, they may also be caused by the cytotoxicity of compounds, as it impedes the identification of a possible interaction.

Available information about the specificity and cytotoxicity of the assay outcomes was collected from the ToxCast database as well as other publications, and the affected measured data were filtered out from the data sets (see Materials and Methods for details). After this filtering step, sufficient amounts of data for model development (i.e., at least 50 active compounds) remained available only for TPO and TR. Compared to the complete data sets, the filtered data sets for TPO and TR contain 27% and 66% less active compounds (total of 68 and 829 active compounds less), respectively. For TPO, the number of inactive compounds remains the same, and for TR, it is reduced by only 0.3% (16 compounds). Note that filtering does not mitigate the problem of false-negative outcomes related to, for example, compound volatility or solubility issues.

With the filtered data sets for TPO and TR, classification models with the same five ML algorithms in combination with oversampling were developed. For TPO, the models obtained F1 scores of up to 0.81 (with RF in combination with oversampling; Table 12). However, the best F1 score obtained by models trained on the unfiltered data set was marginally higher (0.83). Also for TR, the highest F1 score obtained by the models built on the filtered data set (0.68, obtained with the RF model in combination with oversampling) was 0.13 points lower than the best F1 score obtained by the models trained on the complete data set. The observed lower performance of the models on the filtered data sets may be related to the substantial reduction of active compounds, which leads to a significant loss of information.

Although reducing the number of compounds to only those more specific for the inhibitory or antagonistic activity of the targets does not improve the ability of the model to differentiate between active and inactive compounds, these models may have more biological relevance, as they represent a more specific mechanism. However, the substantial reduction of the data sets severely narrows the coverage of the chemical space by the models and therefore their applicability domain.

Multi-task Classification Models. In a further attempt to maximize the performance and scope of in silico models, we explored the use of multi-task models for toxicity prediction, which present the opportunity to combine information and learn a common representation for the molecules.[46] These models are trained on multiple end points simultaneously and may hence benefit from regularization and transfer learning (Figure 9). This could be particularly beneficial in the case of small or imbalanced training sets, like some of the ones handled in this work. For the implementation of multi-task models, we selected NNs as they are the preferred approach for multi-task models in the literature[47,48] and benefit most from the use of larger data sets.

Table 12. Mean and Standard Deviation of the Performance of Different Methods for the Models Built on the Filtered Data Sets for Nonspecific and Cytotoxic Compounds for TPO and TR

end point	method	F1 score	MCC	balanced accuracy	AUC
TPO	**RF**	**0.81 (± 0.05)**	**0.63 (± 0.10)**	**0.78 (± 0.04)**	**0.91 (±0.03)**
	LR	0.79 (± 0.07)	0.59 (± 0.13)	0.79 (± 0.06)	0.87 (± 0.04)
	XGB	0.80 (± 0.04)	0.60 (± 0.09)	0.79 (± 0.04)	0.89 (± 0.02)
	SVM	0.79 (± 0.03)	0.58 (± 0.07)	0.78 (± 0.04)	0.89 (± 0.03)
	NN	0.79 (± 0.05)	0.58 (± 0.11)	0.77 (± 0.05)	0.88 (± 0.04)
TR	**RF**	**0.68 (± 0.05)**	**0.39 (± 0.10)**	**0.65 (± 0.04)**	**0.88 (± 0.02)**
	LR	0.63 (± 0.03)	0.28 (± 0.05)	0.62 (± 0.02)	0.77 (± 0.05)
	XGB	0.67 (± 0.05)	0.37 (± 0.10)	0.64 (± 0.04)	0.85 (± 0.04)
	SVM	0.66 (± 0.04)	0.34 (± 0.07)	0.66 (± 0.05)	0.82 (± 0.04)
	NN	0.64 (± 0.04)	0.29 (± 0.07)	0.61 (± 0.04)	0.81 (± 0.04)

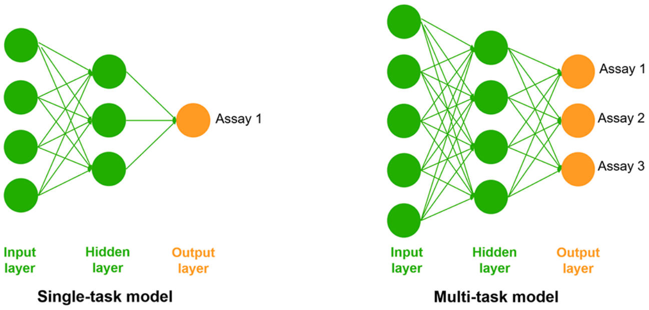


Figure 9. Representation of single-task (left) and multi-task (right) NNs. On the single-task models, only one problem (assay result) is solved at a time, while multi-task models can learn and solve different problems simultaneously.

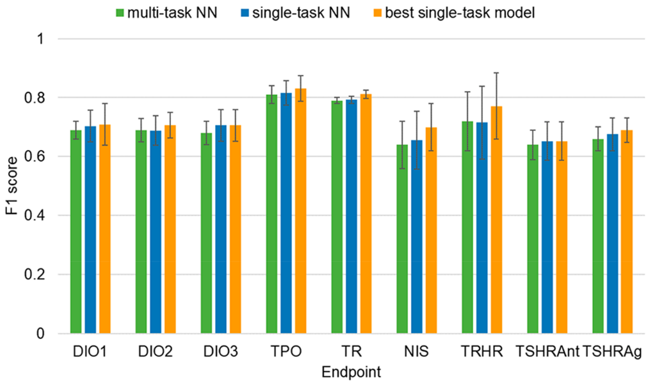


Figure 10. Comparison of single- and multi-task models. Results for single-task methods are divided in (a) best method (orange) and (b) NN method (blue). Performance of multi-task NN is shown in green.

Four multi-task models were built based on different combinations of end points, each covering two to nine end points. As the single-task models for TPO and TR showed good performance (indicating that the training sets for these end points have a high information content), these end points were included in all multi-task models. The other end points were incrementally added to the training data of the multi-task models.

The multi-task models were developed within a workflow that generates 100 models built on different training sets and with optimized hyperparameters (see Materials and Methods for details). The 100 models are grouped based on their hyperparameters, and one group of models with common hyperparameters is selected as the final model. This selection is based on the number of models in the ensemble and its mean F1 score over the respective test sets of the single models. The performance of the multi-task models was evaluated on the mean F1 score of the selected ensemble.

In all cases, the performance of the multi-task NN models was similar to that of the single-task NN model implementing the oversampling approach (Figure 10). The best mean F1 scores obtained among the models with different end point combinations were of 0.81 for TPO (vs 0.82 for the single-task NN model), 0.79 for TR (vs 0.79), 0.69 for DIO1 (vs 0.70), 0.69 for DIO2 (vs 0.69), 0.68 for DIO3 (vs 0.71), 0.64 for NIS (vs 0.66), 0.72 for TRHR (vs 72), 0.64 for TSHRAnt (vs 0.65), and 0.66 for TSHRAg (vs 0.68) (Table 13). The mean F1 score of the multi-task models was also in general comparable to the one obtained by the best single-task model (Figure 10).

Those end points implemented in models with different combinations of end points showed similar performance in all combinations (difference in the mean F1 score up to 0.02 points), suggesting that an increase in the number of end points and data sets represented by a model does not contribute much to the learning process. Although all targets are related to thyroid hormone homeostasis, their structure and functions as well as the assays employed for measuring their function are diverse. The transfer of information between end points is then limited to simple molecular features, without benefiting from common biological features. Since these features are already contained in the descriptors used as input for all the models, there would be no information gain in the combination of these end points, explaining the similar results to the single-task models.

■ CONCLUSIONS

We have compiled a comprehensive set of experimental data on the interference of small molecules with nine targets involved in molecular initiating events of thyroid hormone homeostasis (DIO1, DIO2, DIO3, TPO, TR, NIS, TRHR, and TSHR antagonism and agonism) from the ToxCast database and published studies. Five ML algorithms in combination with three data balancing approaches were explored for the generation of single-task models. In addition, NNs were explored for the development of multi-task models combining several end points.

The classifiers for TPO and TR showed high predictive performance during a 10-fold CV, with mean F1 scores of up to 0.83 and 0.81, respectively. The models for the other end points (DIO1, DIO2, DIO3, NIS, TRHR, TSHRAnt, and TSHRAg), for which the quantity and quality of the available data were more limited, yielded mean F1 scores between 0.65

Table 13. Mean F1 Score and Standard Deviation for the Multi-task Models With Different End Point Combinations

	F1 score			
end point	model 1	model 2	model 3	model 4
DIO1		0.67 (± 0.05)	0.69 (± 0.03)	0.68 (± 0.05)
DIO2		0.68 (± 0.05)	0.68 (± 0.04)	0.69 (± 0.04)
DIO3		0.67 (± 0.04)	0.67 (± 0.04)	0.68 (± 0.04)
TPO	0.81 (± 0.03)	0.81 (± 0.03)	0.80 (± 0.03)	0.80 (± 0.03)
TR	0.79 (± 0.02)	0.79 (± 0.02)	0.79 (± 0.01)	0.78 (± 0.01)
NIS			0.63 (± 0.08)	0.64 (± 0.08)
TRHR			0.72 (± 0.02)	0.72 (± 0.10)
TSHRAnt				0.64 (± 0.05)
TSHRAg				0.66 (± 0.04)

and 0.77. Overall, the impact of the selected ML algorithm and data balancing method on model performance was minor. Larger differences in the performance of the different models were observed for end points for which the amount of data available for model development is very limited (mainly NIS, TRHR, and TSHR). For these end points, models derived in combination with weight balancing and oversampling usually performed better than models derived in combination with undersampling (F1 scores up to 0.24 higher). However, this increase in performance comes with the cost of a higher standard deviation during CV. The performance of the multi-task models was comparable to those of the single-task models, indicating that these models were not able to benefit from a transfer of information. We also showed that the reliability of the predictions is correlated with the similarity of the test compounds and the training instances as well as with the distance of the predicted probability from the decision boundary.

The initial data sets were further filtered with complementary information available on the reliability of assay outcomes (related to cytotoxicity and nonspecific protein inhibition). However, the substantial reduction of training data caused by this refinement procedure resulted in models that did in no case outperform the models trained on unfiltered data. Although the chemical space represented by these models is narrower than the chemical space of those derived from the unfiltered data, these models may be of higher biological relevance as they represent a more specific interaction of the compounds with the target protein.

Overall, the models presented in this work can help in the identification of substances with the potential to disturb the thyroid hormone homeostasis and point out which key events are affected. Thus, they may help to prioritize compounds for further testing in early stages of development and to support read-across. This will ultimately reduce animal testing and increase efficiency of product development and regulatory testing.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.chemrestox.0c00304.

Additional tables and figures: Flags available in the ToxCast database for tagging potential errors in class labeling; list of molecular descriptors used in principal component analysis; distribution of pairwise Tanimoto similarities based on atom-pair fingerprints; comparison of the mean F1 scores obtained for the nine thyroid end points with different machine learning algorithms and data sampling techniques (PDF)

Additional data: The processed data sets, including the "Complete data set" and the "Filtered data set", used in this work for model development and validation as well as complementary information on (i) the filtered out compounds due to ToxCast flags and (ii) the raw SMILES and the SMILES standardization steps (XLSX)

■ AUTHOR INFORMATION

Corresponding Authors

Miriam Mathea – *BASF SE, 67063 Ludwigshafen am Rhein, Germany;* orcid.org/0000-0002-3214-1487; Phone: +49 621 60-29054; Email: miriam.mathea@basf.com

Johannes Kirchmair – *Department of Pharmaceutical Chemistry, Faculty of Life Sciences, University of Vienna, 1090 Vienna, Austria;* orcid.org/0000-0003-2667-5877; Phone: +43 1-4277-55104; Email: johannes.kirchmair@univie.ac.at

Authors

Marina Garcia de Lomana – *BASF SE, 67063 Ludwigshafen am Rhein, Germany; Department of Pharmaceutical Chemistry, Faculty of Life Sciences, University of Vienna, 1090 Vienna, Austria;* orcid.org/0000-0002-9310-7290

Andreas Georg Weber – *BASF SE, 67063 Ludwigshafen am Rhein, Germany;* orcid.org/0000-0001-5545-7583

Barbara Birk – *BASF SE, 67063 Ludwigshafen am Rhein, Germany;* orcid.org/0000-0002-1208-8527

Robert Landsiedel – *BASF SE, 67063 Ludwigshafen am Rhein, Germany*

Janosch Achenbach – *BASF SE, 67063 Ludwigshafen am Rhein, Germany;* orcid.org/0000-0001-9712-1471

Klaus-Juergen Schleifer – *BASF SE, 67063 Ludwigshafen am Rhein, Germany;* orcid.org/0000-0003-3428-1384

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.chemrestox.0c00304

Notes

The authors declare the following competing financial interest(s): M.G.d.L., A.G.W., B.B., R.L., K.-J.S., and M.M. are employed at BASF SE. J.A. was employed at BASF SE during the time he was contributing to this work.

■ ACKNOWLEDGMENTS

We thank Katie Paul-Friedman from the United States Environmental Protection Agency for the information about the ToxCast database assays. We are also grateful for the valuable comments and suggestions received by the two anonymous expert reviewers.

■ ABBREVIATIONS

AOP, adverse outcome pathway; AUC, area under the receiver operating characteristic curve; CV, cross-validation; DIOs, deiodinases; DIO1, deiodinase 1; DIO2, deiodinase 2; DIO3, deiodinase 3; EPA, Environmental Protection Agency; FRAC, Fungicide Resistance Action Committee; HPT, hypothalamic−pituitary−thyroid; HRAC, Herbicide Resistance Action Committee; IRAC, Insecticide Resistance Action Committee; LR, logistic regression; MCC, Matthews correlation coefficient; MIE, molecular initiating event; ML, machine learning; NIS, sodium/iodide symporter; NN, neural networks; PCA, principal component analysis; QSAR, quantitative structure−activity relationship; RF, random forest; SVM, support vector machine; T3, triiodothyronine; T4, tetraiodothyronine; TPO, thyroid peroxidase; TR, thyroid hormone receptor; TRH, thyrotropin-releasing hormone; TRHR, thyrotropin-releasing hormone receptor; TSH, thyroid-stimulating hormone; TSHR, thyroid-stimulating hormone receptor; TSHRAg, thyroid-stimulating hormone receptor agonism; TSHRAnt, thyroid-stimulating hormone receptor antagonism; XGB, gradient boosting

■ REFERENCES

(1) Zoeller, R. T., Tan, S. W., and Tyl, R. W. (2007) General Background on the Hypothalamic-Pituitary-Thyroid (HPT) Axis. *Crit. Rev. Toxicol. 37*, 11−53.

(2) Kim, W. G., and Cheng, S. Y. (2013) Thyroid Hormone Receptors and Cancer. *Biochim. Biophys. Acta, Gen. Subj. 1830*, 3928−3936.

(3) Brent, G. A. (2012) Mechanisms of Thyroid Hormone Action. *J. Clin. Invest. 122*, 3035−3043.

(4) Leemans, M., Couderq, S., Demeneix, B., and Fini, J.-B. (2019) Pesticides With Potential Thyroid Hormone-Disrupting Effects: A Review of Recent Data. *Front. Endocrinol. 10*, 743.

(5) De Cock, M., Maas, Y. G., and Van De Bor, M. (2012) Does Perinatal Exposure to Endocrine Disruptors Induce Autism Spectrum and Attention Deficit Hyperactivity Disorders? Review. *Acta Paediatr. 101*, 811−818.

(6) Hill, R. N., Crisp, T. M., Hurley, P. M., Rosenthal, S. L., and Singh, D. V. (1998) Risk Assessment of Thyroid Follicular Cell Tumors. *Environ. Health Perspect. 106*, 447−457.

(7) Liu, J., Liu, Y., Barter, R. A., and Klaassen, C. D. (1995) Alteration of Thyroid Homeostasis by UDP-Glucuronosyltransferase Inducers in Rats: A Dose-Response Study. *J. Pharmacol. Exp. Ther. 273*, 977−985.

(8) McClain, R. M., Levin, A. A., Posch, R., and Downing, J. C. (1989) The Effect of Phenobarbital on the Metabolism and Excretion of Thyroxine in Rats. *Toxicol. Appl. Pharmacol. 99*, 216−228.

(9) McClain, R. M. (1989) The Significance of Hepatic Microsomal Enzyme Induction and Altered Thyroid Function in Rats: Implications for Thyroid Gland Neoplasia. *Toxicol. Pathol. 17*, 294−306.

(10) Harris, A. R., Christianson, D., Smith, M. S., Fang, S.-L., Braverman, L. E., and Vagenakis, A. G. (1978) The Physiological Role of Thyrotropin-Releasing Hormone in the Regulation of Thyroid-Stimulating Hormone and Prolactin Secretion in the Rat. *J. Clin. Invest. 61*, 441−448.

(11) Vassart, G., and Dumont, J. E. (1992) The Thyrotropin Receptor and the Regulation of Thyrocyte Function and Growth. *Endocr. Rev. 13*, 596−611.

(12) Dohan, O., De la Vieja, A., Paroder, V., Riedel, C., Artani, M., Reed, M., Ginter, C. S., and Carrasco, N. (2003) The Sodium/Iodide Symporter (NIS): Characterization, Regulation, and Medical Significance. *Endocr. Rev. 24*, 48−77.

(13) Köhrle, J. (2000) The Deiodinase Family: Selenoenzymes Regulating Thyroid Hormone Availability and Action. *Cell. Mol. Life Sci. 57*, 1853−1863.

(14) Desvergne, B. (1994) How Do Thyroid Hormone Receptors Bind to Structurally Diverse Response Elements? *Mol. Cell. Endocrinol. 100*, 125−131.

(15) Andersson, N., Arena, M., Auteri, D., Barmaz, S., Grignard, E., Kienzler, A., Lepper, P., Lostia, A. M., Munn, S., et al. (2018) Guidance for the Identification of Endocrine Disruptors in the Context of Regulations (EU) No 528/2012 and (EC) No 1107/2009. *EFSA J. 16*, No. e05311.

(16) (2018) Revised Guidance Document 150 on Standardised Test Guidelines for Evaluating Chemicals for Endocrine Disruption. *OECD Series on Testing and Assessment*, OECD Publishing, Paris.

(17) Murk, A. J., Rijntjes, E., Blaauboer, B. J., Clewell, R., Crofton, K. M., Dingemans, M. M. L., David Furlow, J., Kavlock, R., Köhrle, J., Opitz, R., Traas, T., Visser, T. J., Xia, M., and Gutleb, A. C. (2013) Mechanism-Based Testing Strategy Using in Vitro Approaches for Identification of Thyroid Hormone Disrupting Chemicals. *Toxicol. In Vitro 27*, 1320−1346.

(18) Russell, W. M. S., and Burch, R. L. (1959) *The Principles of Humane Experimental Technique*, Methuen & Co. Limited, London.

(19) Noyes, P. D., Friedman, K. P., Browne, P., Haselman, J. T., Gilbert, M. E., Hornung, M. W., Barone, S., Jr, Crofton, K. M., Laws, S. C., Stoker, T. E., et al. (2019) Evaluating Chemicals for Thyroid Disruption: Opportunities and Challenges with in Vitro Testing and Adverse Outcome Pathway Approaches. *Environ. Health Perspect. 127*, 095001.

(20) (2017) Continuing Development of Alternative High-Throughput Screens to Determine Endocrine Disruption, Focusing on Androgen Receptor, Steroidogenesis, and Thyroid Pathways. *FIFRA Scientific Advisory Panel*, Vol. *30*, U.S. EPA, Washington, DC.

(21) Rosenberg, S. A., Watt, E. D., Judson, R. S., Simmons, S. O., Friedman, K. P., Dybdahl, M., Nikolov, N. G., and Wedebye, E. B. (2017) QSAR Models for Thyroperoxidase Inhibition and Screening of U.S. and EU Chemical Inventories. *Comput. Toxicol. 4*, 11−21.

(22) Politi, R., Rusyn, I., and Tropsha, A. (2014) Prediction of Binding Affinity and Efficacy of Thyroid Hormone Receptor Ligands Using QSAR and Structure Based Modeling Methods. *Toxicol. Appl. Pharmacol. 280*, 177−189.

(23) Liu, H., and Gramatica, P. (2007) QSAR Study of Selective Ligands for the Thyroid Hormone Receptor *β*. *Bioorg. Med. Chem. 15*, 5251−5261.

(24) Azimi, G., Afiuni-Zadeh, S., and Karami, A. (2012) A QSAR Study for Modeling of Thyroid Receptors *β*1 Selective Ligands by Application of Adaptive Neuro-Fuzzy Inference System and Radial Basis Function. *J. Chemom. 26*, 135−142.

(25) (2019) *ToxCast and Tox21 Summary Files for invitroDBv3.2*, U.S. EPA, Washington, DC. (accessed 2020-08-27)

(26) Filer, D. L., Kothiya, P., Setzer, R. W., Judson, R. S., and Martin, M. T. (2016) tcpl: the Tox Cast Pipeline for High-Throughput Screening Data. *Bioinformatics* 33, 618−620.

(27) Olker, J. H., Korte, J. J., Denny, J. S., Hartig, P. C., Cardon, M. C., Knutsen, C. N., Kent, P. M., Christensen, J. P., Degitz, S. J., and Hornung, M. W. (2019) Screening the Tox Cast Phase 1, Phase 2, and e1k for Inhibitors of Iodothyronine. *Toxicol. Sci. 168*, 430−442.

(28) Friedman, K. P., Watt, E. D., Hornung, M. W., Hedge, J. M., Judson, R. S., Crofton, K. M., Houck, K. A., and Simmons, S. O. (2016) Tiered High-Throughput Screening Approach to Identify Thyroperoxidase Inhibitors Within the Tox Cast Phase I and II Chemical Libraries. *Toxicol. Sci. 151*, 160−180.

(29) Wang, J., Hallinger, D. R., Murr, A. S., Buckalew, A. R., Lougee, R. R., Richard, A. M., Laws, S. C., and Stoker, T. E. (2019) High-Throughput Screening and Chemotype-Enrichment Analysis of Tox Cast Phase II Chemicals Evaluated for Human Sodium-Iodide Symporter (NIS) Inhibition. *Environ. Int. 126*, 377−386.

(30) Richard, A. M., Judson, R. S., Houck, K. A., Grulke, C. M., Volarath, P., Thillainadarajah, I., Yang, C., Rathman, J., Martin, M. T., Wambaugh, J. F., Knudsen, T. B., Kancherla, J., Mansouri, K., Patlewicz, G., Williams, A. J., Little, S. B., Crofton, K. M., and Thomas, R. S. (2016) Tox Cast Chemical Landscape: Paving the Road to 21st Century Toxicology. *Chem. Res. Toxicol. 29*, 1225−51.

(31) Paul-Friedman, K., Martin, M., Crofton, K. M., Hsu, C. W., Sakamuru, S., Zhao, J., Xia, M., Huang, R., Stavreva, D. A., Soni, V., Varticovski, L., Raziuddin, R., Hager, G. L., and Houck, K. A. (2019) Limited Chemical Structural Diversity Found to Modulate Thyroid Hormone Receptor in the Tox21 Chemical Library. *Environ. Health Perspect. 127*, 097009.

(32) *DrugBank Version 5.1.5.* https://www.drugbank.ca (accessed 2020-02-14).

(33) *CosIng (Cosmetic Ingredient Database)* - Growth - European Commission. http://ec.europa.eu/growth/tools-databases/cosing/index.cfm?fuseaction=search.simple (accessed 2020-02-14).

(34) Gaulton, A., Bellis, L. J., Bento, A. P., Chambers, J., Davies, M., Hersey, A., Light, Y., McGlinchey, S., Michalovich, D., Al-Lazikani, B., and Overington, J. P. (2012) ChEMBL: A Large-Scale Bioactivity Database for Drug Discovery. *Nucleic Acids Res. 40*, D1100−D1107.

(35) Berthold, M. R., Cebron, N., Dill, F., Gabriel, T. R., Kötter, T., Meinl, T., Ohl, P., Sieb, C., Thiel, K., and Wiswedel, B. (2007) KNIME: The Konstanz Information Miner. In *Studies in Classification, Data Analysis, and Knowledge Organization (GfKL 2007)*, Springer, Berlin.

(36) Standardizer was used for structure canonicalization and transformation. *JChem 3.5.0*, ChemAxon, Budapest (http://www.chemaxon.com).

(37) Landrum, G. (2018) *RDKit: Open-Source Cheminformatics Software*, version 2018.09.1.

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

(39) Chollet, F., and Al, E. (2015) *Keras* (Version 2.2.4.).

(40) *NCI/CADD Chemical Identifier Resolver.* https://cactus.nci.nih.gov/chemical/structure (accessed 2019-10-01).

(41) Chawla, N. V., Bowyer, K., Hall, L. O., and Kegelmeyer, P. O. (2002) *SMOTE: Synthetic Minority Over-Sampling Technique. 16*, 321−357.

(42) Carhart, R. E., Smith, D. H., and Venkataraghavan, R. (1985) *J. Chem. Inf. Model. 25*, 64−73.

(43) Laggner, C. (2005) SMARTS Patterns for Functional Group Classification, *Git Hub repository*, Inte:Ligand Software-Entwicklungs und Consulting GmbH, Austria, https://github.com/openbabel/openbabel/blob/master/data/SMARTS_InteLigand.txt (accessed 2020-09-01).

(44) O'Boyle, N. M., Banck, M., James, C. A., Morley, C., Vandermeersch, T., and Hutchison, G. R. (2011) Open Babel: An open chemical toolbox. *J. Cheminf. 3*, 33.

(45) Friedman, K. P., Zhao, J., Huang, R., Xia, M., Crofton, K., and Houck, K. (2017) Screening the Tox21 10K Library for Thyroid Stimulating Hormone Receptor Agonist and Antagonist Activity. Proceedings from the *Society of Toxicology Annual Meeting*, March 12−16, 2017, Baltimore, MD, Society of Toxicology, Reston, VA.

(46) Caruana, R. (1997) Multi-task Learning. *Mach. Learn. 28*, 41−75.

(47) Li, X., Xu, Y., Lai, L., and Pei, J. (2018) Prediction of Human Cytochrome P450 Inhibition Using a Multi-task Deep Autoencoder Neural Network. *Mol. Pharmaceutics 15*, 4336−4345.

(48) Wenzel, J., Matter, H., and Schmidt, F. (2019) Predictive Multi-task Deep Neural Network Models for ADME-Tox Properties: Learning from Large Data Sets. *J. Chem. Inf. Model. 59*, 1253−1268.

4.2 Enhancement of in vivo toxicity prediction with predicted bioactivity descriptors

In vivo toxicity prediction is still a challenging problem due to the high complexity of the endpoints and the scarce data. Chemical similarity to measured compounds is often not enough to predict the outcome of untested compounds in in vivo assays, as small structural changes may influence any of the ADME parameters determining the in vivo effects. Some studies on read-across applications[78-80] and toxicity prediction[46-48] have already shown that describing the behavior of compounds in biological systems may better capture similarities on a biological level.

A common problem of in silico models including biological information is the sparsity of the available experimental data for building bioactivity descriptors. The following study presents an approach for exploiting the benefits of biological data while overcoming the data sparsity issue by the generation of predicted bioactivity descriptors. For developing these descriptors, data sets for over 300 in vitro and pharmacokinetics assays were collected and used to train ML models. These models were then used to compute predicted bioactivity descriptors that, alone or in combination with chemical descriptors, constituted the input features for the development of in vivo toxicity prediction models. The approach was tested on a genotoxicity in vivo assay (MNT) and two organ toxicity endpoints (drug-induced liver injury (DILI) and cardiological complications (DICC)). All developed models were built within a CP framework, which enabled the direct definition of the AD for both the bioactivity models (used for calculating the bioactivity descriptors) and the in vivo toxicity models. Moreover, an analysis of the most important bioactivity descriptors for the prediction of each endpoint was conducted to understand relevant biological relationships found by the models and define the best strategy for applying the method on new in vivo endpoints.

[P2] ChemBioSim: Enhancing Conformal Prediction of In Vivo Toxicity by Use of Predicted Bioactivities

Marina Garcia de Lomana, Andrea Morger, Ulf Norinder, Roland Buesen, Robert Landsiedel, Andrea Volkamer, Johannes Kirchmair, and Miriam Mathea

Journal of Chemical Information and Modeling, 2021

Available at https://doi.org/10.1021/acs.jcim.1c00451

Contribution:

M. Garcia de Lomana, U. Norinder, M. Mathea and J. Kirchmair conceptualized the research. M. Garcia de Lomana along with U. Norinder, A. Morger, A. Volkamer, M. Mathea and J. Kirchmair designed the experiments. M. Garcia de Lomana compiled the data sets and analyzed them with contributions from R. Landsiedel and R. Buesen. M. Garcia de Lomana developed the machine learning models. M. Garcia de Lomana wrote the manuscript, with contributions from A. Morger, A. Volkamer, U. Norinder, R. Landsiedel, R. Buesen, M. Mathea and J. Kirchmair. J. Kirchmair and M. Mathea supervised the work.

The following article was reprinted with permission from:

Garcia de Lomana, M.; Morger, A.; Norinder, U.; Buesen, R.; Landsiedel, R.; Volkamer, A.; Kirchmair, J. and Mathea, M. ChemBioSim: Enhancing Conformal Prediction of In Vivo Toxicity by Use of Predicted Bioactivities, *J. Chem. Inf. Model.*, **2021**, 61, 3255–3272.

JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

pubs.acs.org/jcim Article

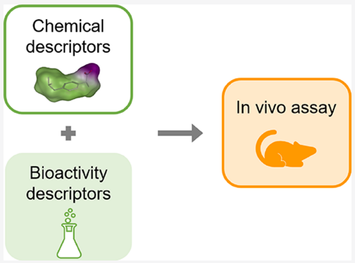
ChemBioSim: Enhancing Conformal Prediction of In Vivo Toxicity by Use of Predicted Bioactivities

Marina Garcia de Lomana, Andrea Morger, Ulf Norinder, Roland Buesen, Robert Landsiedel, Andrea Volkamer, Johannes Kirchmair,* and Miriam Mathea*

Cite This: *J. Chem. Inf. Model.* 2021, 61, 3255−3272

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information



ABSTRACT: Computational methods such as machine learning approaches have a strong track record of success in predicting the outcomes of in vitro assays. In contrast, their ability to predict in vivo endpoints is more limited due to the high number of parameters and processes that may influence the outcome. Recent studies have shown that the combination of chemical and biological data can yield better models for in vivo endpoints. The ChemBioSim approach presented in this work aims to enhance the performance of conformal prediction models for in vivo endpoints by combining chemical information with (predicted) bioactivity assay outcomes. Three in vivo toxicological endpoints, capturing genotoxic (MNT), hepatic (DILI), and cardiological (DICC) issues, were selected for this study due to their high relevance for the registration and authorization of new compounds. Since the sparsity of available biological assay data is challenging for predictive modeling, predicted bioactivity descriptors were introduced instead. Thus, a machine learning model for each of the 373 collected biological assays was trained and applied on the compounds of the in vivo toxicity data sets. Besides the chemical descriptors (molecular fingerprints and physicochemical properties), these predicted bioactivities served as descriptors for the models of the three in vivo endpoints. For this study, a workflow based on a conformal prediction framework (a method for confidence estimation) built on random forest models was developed. Furthermore, the most relevant chemical and bioactivity descriptors for each in vivo endpoint were preselected with lasso models. The incorporation of bioactivity descriptors increased the mean F1 scores of the MNT model from 0.61 to 0.70 and for the DICC model from 0.72 to 0.82 while the mean efficiencies increased by roughly 0.10 for both endpoints. In contrast, for the DILI endpoint, no significant improvement in model performance was observed. Besides pure performance improvements, an analysis of the most important bioactivity features allowed detection of novel and less intuitive relationships between the predicted biological assay outcomes used as descriptors and the in vivo endpoints. This study presents how the prediction of in vivo toxicity endpoints can be improved by the incorporation of biological information—which is not necessarily captured by chemical descriptors—in an automated workflow without the need for adding experimental workload for the generation of bioactivity descriptors as predicted outcomes of bioactivity assays were utilized. All bioactivity CP models for deriving the predicted bioactivities, as well as the in vivo toxicity CP models, can be freely downloaded from https://doi.org/10.5281/zenodo.4761225.

■ INTRODUCTION

Modern toxicity testing heavily relies on animal models, which entails ethical concerns, substantial costs, and difficulties in the extrapolation of results to humans.[1] The increasing amount and diversity of not only drugs but also more generally of chemicals present in the environment and the lack of knowledge about their toxic potential require the development of more efficient toxicity assessment tools.

In recent years, in silico tools for toxicity prediction have evolved into powerful methods that can help to decrease animal testing.[2−4] This is particularly true when applied in tandem with in vitro methods.[5] Machine learning (ML) models trained on data sets of compounds with known activities for an assay can be used as predictive tools for untested compounds.[6] These models are generally trained on chemical and structural features of compounds with measured activity values.[7] However, the outcomes of in vivo toxicological assays depend on a number of biological interactions such as the administration, distribution, metabolism, and excretion (ADME) and the interaction with different cell types.[4] The ability of chemical property descriptors to capture these complex interactions and, consequently, the predictive power of ML models trained on these molecular representations are limited. By the example of classification models for hit expansion[8,9] and toxicity prediction,[10−13] recent studies have shown that the predictive power of in silico models can be improved by the amalgamation of chemical and biological

Received: April 20, 2021
Published: June 21, 2021

Table 1. Overview of Collected Assay Data

database/ endpoint	description	source
ToxCast database	• 222 high-throughput screening assays, including endpoints related to cell cycle and morphology control, steroid hormone homeostasis, DNA-binding proteins, and other protein families (e.g., kinases, cytochromes, and transporters)	ToxCast database version 3.3[24]
eMolTox database	• 136 in vitro assays, including endpoints related to mutagenicity, cytotoxicity, hormone homeostasis, neurotransmitters, and several protein families (e.g., nuclear receptors, cytochromes, and cell surface receptors)	Ji et al.[25]
genotoxicity	• AMES mutagenicity assay • chromosome aberration (CA) assay	AMES assay: eChemPortal,[26] Benigni et al.,[28] Hansen et al.[29]
	• mammalian mutagenicity (MM) assay	CA and MM assays: eChemPortal, Benigni et al.
bioavailability	• human oral bioavailability assay	Falcón-Cano et al.[27]
permeability	• Caco-2 assay	Wang et al.[30]
thyroid hormone homeostasis	• deiodinases 1, 2, and 3 inhibition assays • thyroid peroxidase inhibition assay • sodium iodide symporter inhibition assay • thyroid hormone receptor antagonism assay • thyrotropin-releasing hormone receptor antagonism assay • thyroid stimulating hormone receptor agonism and antagonism assays	Garcia de Lomana et al.[31]
P-glycoprotein inhibition	• P-glycoprotein (ABCB1) inhibition assay	Broccatelli et al.[32]

information. More specifically, it has been shown that bioactivity descriptors could help to infer the activity of new substances by capturing the similarity of compounds in the biological space, i.e., identifying those compounds that behave similarly in biological systems (but may be chemically dissimilar). However, options to integrate biological data into models are limited by the sparsity of the available experimental data. In principle, the use of bioactivity features in ML requires compounds of interest to be tested in all assays conforming the bioactivity descriptor set. Norinder et al.[14] however showed, by the example of conformal prediction (CP) frameworks built on random forest (RF) models, that the use of predicted bioactivity descriptors in combination with chemical descriptors can yield superior cytotoxicity and bioactivity predictions while circumventing the problems of sparsity of data and extensive testing. CP models are a robust type of confidence predictors that generate predictions with a fixed error rate determined by the user.[15] To estimate the confidence of new predictions, the predicted probabilities of a set of compounds with known activity (calibration set) are used to rank the predicted probabilities for new compounds and calculate their so-called p-values (i.e., calibrated probabilities). An additional feature of CP models is their ability to handle data imbalance and predict minority classes more accurately.[16]

The CP approach offers the advantage of a mathematical definition of a model's applicability domain (AD); i.e., chemical space within the model makes predictions with a defined reliability based on the allowed error rate.[17] Other common approaches for defining the applicability domain are based on compound similarity or predicted probability and a more or less arbitrary (user-defined) threshold. However, CP models return a statistically robust class membership probability for each class. Under the exchangeability assumption of the samples (assumption also made for classical ML models), the observed error rate returned by CP models will be equal to (or very close to) the allowed (i.e., user-defined) error rate.

The aim of this study is to determine if, and to what extent, classification models for the prediction of in vivo toxicity endpoints can benefit from integrating chemical representations with data from biological assays. To include the biological assay information in the models, predicted bioactivities were derived from 373 CP models, each representing an individual biological assay. The results obtained for models trained exclusively on chemical descriptors ("CHEM"), trained exclusively on bioactivity ("BIO") descriptors, or trained on the combination of chemical and bioactivity descriptors ("CHEMBIO") were analyzed for three toxicological in vivo endpoints: in vivo genotoxicity (with the in vivo micronucleus test (MNT)), drug-induced liver injury (DILI), and cardiological complications (DICC).

The in vivo MNT assay is used to detect genetic (clastogenic and aneugenic) damage induced by a substance causing the appearance of micronuclei in erythrocytes or reticulocytes of mice or rats.[18] DILI describes the potential hepatotoxicity of a compound. Although there is no consensus method for assessing the DILI potential of a compound, the U.S. Food and Drug Administration (FDA) proposed a systematic classification scheme based on the FDA-approved drug labeling.[19] The DICC endpoint comprises five cardiological complications induced by drugs and annotated in clinical reports: hypertension, arrhythmia, heart block, cardiac failure, and myocardial infarction.

Severe organ toxicity, as observed with DILI and DICC, but also genotoxicity (which can lead to carcinogenesis and teratogenic effects) must be avoided and hence recognized early in the development of industrial chemicals and drugs. Both hepatic and cardiovascular adverse effects are listed as two of the most common safety reasons for drug withdrawals[20] and failures in drug development phases I–III.[21] Moreover, REACH, the chemical control regulation in the European Union, is requiring the in vivo MNT as follow up of a positive result in any genotoxicity test in vitro.[22] The Organisation for Economic Co-operation and Development (OECD) Guideline 474 and the International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH) list the in vivo MNT assay as one of the recommended tests for detecting genotoxicity, as it can account for ADME factors and DNA repair processes.[18,23]

This study introduces an improvement of the in silico prediction of in vivo toxicity endpoints by considering the activity of compounds in multiple biological test systems. We show that predicted bioactivities, which present the benefit of not needing further experimental testing for new compounds, are often enough to achieve ML models with increased performance.

■ MATERIALS AND METHODS

Data Sets. In the following paragraphs, the data from biological assays used for generating descriptors based on predicted bioactivities are introduced followed by the data related to the three in vivo toxicological endpoints (MNT, DILI, and DICC). Finally, the reference data sets used to analyze the chemical space covered by the in vivo endpoints are described.

All information required for the download of any of the data sets used for modeling in this study (including download links, exact json queries, as well as MD5 file checksums) are provided in Table S1 (for the in vivo endpoints) and Table S2 (for the biological assays).

Biological Assays. For the generation of descriptors from predicted bioactivities, a total of 373 data sets (each belonging to a single biological assay) were collected (Table 1): 372 data sets from in vitro assays obtained from the ToxCast,[24] eMolTox,[25] and eChemPortal[26] databases and the literature, and one data set from an in vivo assay (a human oral bioavailability assay) obtained from Falcón-Cano et al.[27] From the ToxCast and eMolTox databases, only endpoints with at least 200 active and 200 inactive compounds listed (after structure preparation and deduplication; see the section Structure Preparation for details) were considered for modeling. Besides the endpoints selected from these two databases, data sets for assays covering genotoxicity, bioavailability, permeability, thyroid hormone homeostasis disruption, and P-glycoprotein inhibition were considered (Table 1). A more detailed description of the data collection and activity labeling of these data sets is provided in Table S2. The numbers of active and inactive compounds in each of the 373 data sets (after the structure preparation and deduplication steps) are reported in Table S3.

In Vivo Endpoints. During the development of this study, a larger number of publicly available in vivo endpoint data sets were investigated for their suitability for modeling. Taking into account the quantity and quality of the data, as well as the regulatory relevance of the toxicological endpoints, three in vivo endpoints were selected for this study: MNT, DILI, and DICC. The collection of the respective data sets is introduced in the following paragraphs.

MNT Data Set. For the MNT assay, data from the European Chemicals Agency (ECHA) available at the eChemPortal were collected. Only experimental data derived according to the OECD Guideline 474 (or equivalent) were considered. All assay outcomes annotated as unreliable or related to compounds that are cytotoxic were discarded. All compounds (identified based on CAS numbers) with conflicting activity data were also removed. Additional data were obtained from the work of Benigni et al.,[28] which includes curated data sets from the European Food Safety Authority (EFSA) data. In addition, data sets for MNT on mouse (1001 compounds) and rat (127 compounds) compiled by Yoo et al.[33] and containing binary activity labels for MNT were obtained. These additional data sets include data, among other sources, from the FDA approval packages, the National Toxicology Program (NTP) studies, the U.S. EPA GENETOX database, the Chemical Carcinogenesis Information System (CCRIS) and the public literature. The mouse and rat data sets did not contain overlapping compounds and an overall MNT result (independent from the species) was derived for the 1128 compounds in the data set. The final data set (after the structure preparation and deduplication steps) contains a total of 1791 compounds (316 active and 1475 inactive compounds; Table 2).

Table 2. Overview of the Data Sets for the in Vivo Endpoints

	number of		
endpoint	active compounds	inactive compounds	ratio
MNT	316	1475	1:5
DILI	445	247	2:1
DICC	988	2268	1:2

DILI Data Set. The data for the DILI endpoint were obtained from the verified DILIrank data set compiled by the FDA.[34] In this data set, drugs are classified as “Most-DILI-concern”, “Less-DILI-concern”, “No-DILI-concern”, and “Ambiguous-DILI-concern”. For the purpose of this study, compounds in the “Most-DILI-concern” and “Less-DILI-concern” classes were labeled as ″active″ and compounds in the “No-DILI-concern” class were labeled as ″inactive″. Compounds of the ″Ambiguous-DILI-concern″ class were removed from the data set. The final binary DILI data set contained 692 compounds (445 active and 247 inactive compounds).

DICC Data Set. For the DICC endpoint, the data set compiled by Cai et al.[35] on different cardiological complications was used. In their work, Cai et al. gathered individual data sets for hypertension, arrhythmia, heart block, cardiac failure, and myocardial infarction from five databases: Comparative Toxicogenomics Database (CTD),[36] SIDER[37] (side effect resource), Offsides[38] (database of drugs effects), MetaADEDB[39] (adverse drug events database), and DrugBank.[40] In this study, a unique DICC data set was built that combines the five data sets of Cai et al. In the DICC data set, compounds were labeled as “active” if they were measured to be active on at least one of the cardiological endpoints (and active, inactive, or “missing” on the remaining endpoints), and as “inactive” otherwise. This resulted in a data set of 3256 compounds after the structure preparation and deduplication steps (988 active and 2268 inactive compounds; see section Structure Preparation for details).

Reference Data Sets. Three reference data sets were obtained to represent the chemical space of pesticide active ingredients, cosmetic ingredients, and drugs in order to analyze the coverage of these types of substances by the in vivo endpoint data sets. The chemical space of pesticides was represented by the 2417 compounds (after structure preparation and deduplication; see the section Structure Preparation for details) collected in the Pesticide Chemical Search database[41] (from the Environmental Protection Agency’s (EPA) Office of Pesticide Programs) and downloaded from the CompTox Dashboard.[42] The chemical space of cosmetic ingredients was represented by the 4503 compounds (after structure preparation and deduplication) included in the COSMOS cosmetics database,[43] created as part

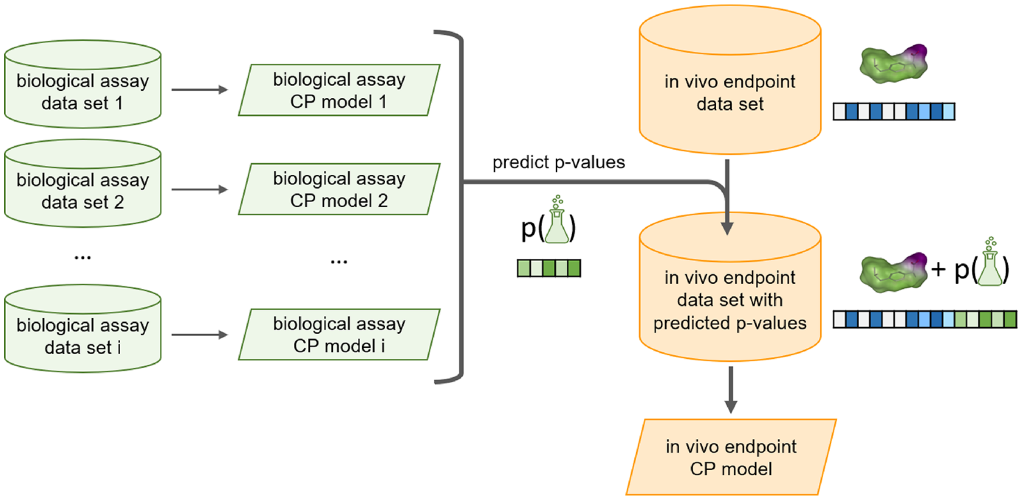

biological assay data set 1
biological assay CP model 1
biological assay data set 2
biological assay CP model 2
...
...
biological assay data set i
biological assay CP model i
predict p-values
p( )
in vivo endpoint data set
in vivo endpoint data set with predicted p-values
+ p( )
in vivo endpoint CP model

Figure 1. Workflow for the derivation of the bioactivity descriptors for the in vivo toxicity CP models. For each biological assay, a conformal prediction model is built and used to predict the *p*-values of the compounds in the three in vivo endpoint data sets. These predicted *p*-values are used as bioactivity descriptors, in combination with chemical descriptors, for training the models of the in vivo endpoints.

of a European Union project for determining the safety of cosmetics in industry without the use of animals, and downloaded from the CompTox Dashboard as well. The chemical space of drugs was represented by the 10087 (after structure preparation and deduplication) approved, experimental, or withdrawn drugs contained in DrugBank.[44]

Structure Preparation. The structures of all molecules were prepared starting from the respective SMILES strings, which are directly available from most data resources. For resources that do not provide SMILES strings (e.g., eChemPortal and the work of Yoo et al.), this information was obtained by querying the PubChem PUG REST interface[45] with the CAS numbers. CAS numbers for which no SMILES was retrieved by this PubChem search were queried with the NCI/CADD Chemical Identifier Resolver.[46] For the 977 compounds that did not produce any match with this procedure either, the "RDKit from IUPAC" node of RDKit[47] in KNIME[48] was used in an attempt to derive a structure from the chemical name. For 131 out of these 977 compounds, the chemical structure was successfully derived with this method. The remaining 846 compounds, without known chemical structures (e.g., including compound mixtures and unspecific formulas), were removed.

All obtained SMILES notations were interpreted, processed, and standardized with the ChemAxon Standardizer[49] node in KNIME. As part of this process, solvents and salts were removed, aromaticity was annotated, charges were neutralized, and structures were mesomerized (taking the canonical resonant form of the molecule). All compounds containing any element other than H, B, C, N, O, F, Si, P, S, Cl, Se, Br, and I were removed from the data set with the "RDKit Substructure Filter" node in KNIME. In the case of multicomponent compounds, the structures of the individual components forming the compound were compared. More specifically, the canonical SMILES of the components were derived with RDKit, and in case the components had identical canonical SMILES, one of them was kept; otherwise, the whole compound was filtered out. Lastly, compounds with fewer than four heavy atoms were discarded.

Canonical SMILES were derived with RDKit from all standardized compounds. For each endpoint data set, duplicate canonical SMILES with conflicting activity labels were removed from the respective endpoint data set.

A KNIME workflow with the specific steps and settings for the preparation of the structures as well as for the calculation of the chemical descriptors (see Descriptor Calculation section) is provided in the Supplementary Information.

Descriptor Calculation. *Chemical Descriptors.* Molecular structures were encoded using count-based Morgan fingerprints with a radius of 2 bonds and a length of 2048 bytes, computed with the ″RDKit Count-Based Fingerprint″ node in KNIME. Morgan fingerprints encode circular environments and capture rather local properties of the molecules. To capture global molecular properties, all 119 1D and 2D physicochemical property descriptors implemented in the "RDKit Descriptor Calculation" node in KNIME were calculated. These descriptors encode properties such as the number of bonds and rings in a molecule, the number of particular types of atoms, or the polarity and solubility of the compound. Two acidic and two basic pK_a values were also calculated per molecule with the "pK_a" KNIME node from ChemAxon.[50] Missing pK_a values (for molecules without two acidic or basic groups) were replaced with the mean value of the data set.

Bioactivity Descriptors. For the calculation of the bioactivity descriptors, first, 373 CP models—one per assay—were fitted on the respective biological assay sets (see the Data Sets section for details). The workflow for the generation of these models is explained in detail in the "Model development" section. With the generated bioactivity CP models, two *p*-values for each compound contained in the three in vivo endpoint data sets were predicted (Figure 1). Both the *p*-values for the active (p1) and for the inactive (p0) classes for each assay were used as bioactivity descriptors, resulting in 746 descriptors.

Chemical Space Analysis. To visualize the chemical space covered by the data sets of the in vivo endpoints, dimensionality reduction was performed on a subset of 23 physically meaningful and interpretable molecular descriptors generated with RDKit (Table S4). For that purpose, the principal component analysis (PCA) implementation of scikit-learn[51] was applied on the merged in vivo endpoint data sets

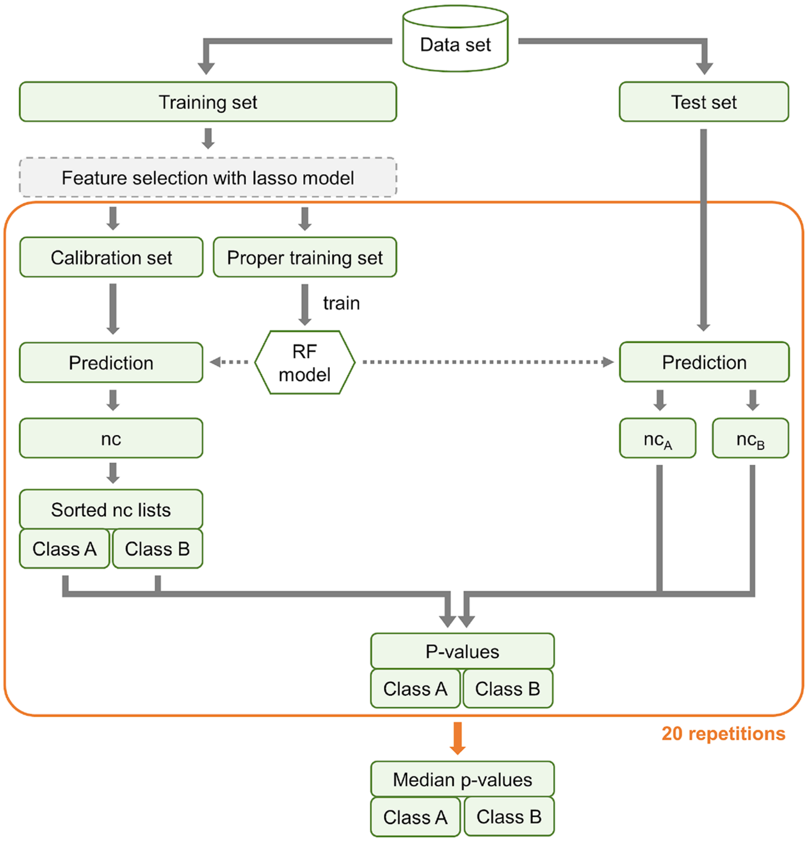


Figure 2. Workflow of the aggregated Mondrian CP set up for the development of the models for the biological assays and the in vivo endpoints. The aggregated CP framework included 20 random splits in calibration and proper training data sets, on which individual RF models were trained, and the resulting *p*-values per test compound were afterward averaged. The feature selection step was implemented with a lasso model and only included in the development of the in vivo toxicity CP models (in vivo toxicity CP models without feature selection were also trained for comparison).

(merged on the canonical SMILES). A further visualization of the chemical space defined by the complete CHEM and CHEMBIO descriptor sets was performed with the Uniform Manifold Approximation and Projection (UMAP).[52] This method conducts a dimension reduction while maintaining the global structure of the data (i.e., the pairwise distance between samples). For each of the three in vivo endpoint data sets, a two-dimensional projection was performed on the CHEM and CHEMBIO descriptor sets, respectively, with 50 nearest neighbors, a minimum distance of 0.2, and use of the “euclidean” metric as the distance measure.

The molecular similarities of the compounds of the in vivo endpoint data sets and the collected pesticides, cosmetics, and drugs reference data sets were quantified with Tanimoto coefficients calculated from Morgan fingerprints with a radius of 2 bonds and a length of 1024 bits (fingerprints computed with the ″RDKit Fingerprint″ node in KNIME).

Model Development for the Biological Assays and In Vivo Toxicity Endpoints. *Workflow for the Development of CP Models.* The same model development workflow was followed to train the CP models used for the calculation of the bioactivity descriptors, as well as to train the final models for the in vivo toxicity endpoints. Note that the structure preparation and chemical descriptor calculation was done in KNIME, but the following workflow was implemented in Python. All hyperparameters of the functions used in the workflow for deriving the CP models are specified in Table S5.

Prior to model development, a variance filter was applied to all features used as input for the in vivo toxicity CP models (including the bioactivity features if present) in order to remove any features with low information content. More specifically, any features with a variance (among the compounds in the respective data set) of less than 0.0015 were removed. Note that, in order to preserve the homogeneity of the input features, this variance filter was not part of the workflow for the biological assay CP model development (used to calculate the bioactivity descriptors). Also, in all cases (including the biological assay CP models), the features were scaled (by subtracting the mean and scaling to unit variance) prior to model development by applying the StandardScaler class of scikit-learn on each endpoint-specific data set.

For CP model development, each endpoint-specific data set was divided into 80% training and 20% test set using the StratifiedShuffleSplit class of scikit-learn (Figure 2). For performance assessment, this splitting of the data was performed within a 5-fold cross-validation (CV) framework. During each CV run, the training set was further divided (stratified) into a proper training set (70% of the training set) and a calibration set (30% of the training set) with the RandomSubSampler class from the nonconformist Python package.[53] An RF model was trained on the proper training set using the scikit-learn implementation (with 500 estimators and default values for the rest of the hyperparameters). The trained RF model was then used to predict the probabilities of the compounds in the calibration set. From these probabilities, the so-called nonconformity score (nc score) was derived by applying a nonconformity error function, which yields low nc scores for predictions close to the true value. Here, the inverse probability error function from the nonconformist package (named “InverseProbabilityErrFunc”) was used to calculate the nc scores. This error function is defined as

$$\text{nc score} = 1 - \hat{P}(y_i|x),$$

with $\hat{P}(y_i \mid x)$ being the probability of predicting the correct class.

By definition, errors produced by CP models do not exceed the significance level ε (i.e., indicated error rate) under the assumption that training and test compounds are independent and belong to the same distribution. However, these errors may be unevenly distributed across classes. To achieve conditional validity with respect to the active and inactive classes, the Mondrian approach was used. Following the Mondrian CP approach, a sorted nc score list with the calculated nc scores of the calibration set was created for each class (active/inactive) independently. After calculating the nc scores (one per class) for the test compounds, their rank (with regard to the calibration set) in the respective list was calculated. The rank of the nc score of each test compound defines the predicted p-value for the respective class.

An aggregated CP approach[54] was conducted by repeating the random splitting of the proper training and calibration sets 20 times. As a result, the p-values for a test set were calculated 20 times and the final p-value was derived from the median value.

CP models output a set of labels, which contain one class (“active” or “inactive”), both classes, or none. If the final p-value for any of the classes was higher than the significance level ε, the compound was assigned to that class (or to both classes if both p-values were higher than ε). Thus, based on the p-values and the significance level, the CP model determines whether a compound is within the applicability domain (AD) of the model.[55] Compounds within the AD of the model are assigned to one or both classes and those outside of the AD are assigned to the empty class (i.e., no class label is assigned).

The predicted p-values obtained by applying the bioactivity CP models on the in vivo endpoint data sets (for the generation of the bioactivity descriptors) were used as is, and no class labeling was performed (i.e., no significance level was assigned). Instead, the p-values for both classes were considered.

In Vivo Toxicity CP Models Including Feature Selection. The workflow for developing the in vivo toxicity CP models that include feature selection is similar to the general workflow described in the previous section but additionally includes a least absolute shrinkage and selection operator (lasso) model.[56] Lasso is a regression method that penalizes the coefficients of the input features for the selection of variables and the regularization of models. Some feature coefficients are shrunk to zero and therefore eliminated from the model.

In our workflow, a lasso model with the LassoCV implementation of scikit-learn was trained on the complete training set (prior to splitting the complete training set into proper training and calibration set; see Figure 2). To optimize the regularization parameter alpha of the lasso model, an inner 5-fold CV is applied. The list of coefficients assigned to each feature is obtained, and those features with a coefficient shrunken to zero are filtered out from the data set. Only the selected features (i.e., with a coefficient higher than zero) are used as input for the aggregated CP workflow described in the previous section.

In order to use the coefficients for ranking the features according to their importance for the analysis of the models, the mean among the absolute values of the coefficients obtained during each outer CV run was calculated.

Since the lasso model discards highly correlated features, considering only the lasso coefficients for the analysis of the most relevant features could lead to an underestimation of the importance of some biological assays. Therefore, this analysis was mainly based on the feature importance values of the RF models without feature preselection with lasso. The feature importance values of RF were extracted, and the mean across CV runs were calculated. Lastly, to better estimate the relative importance of each feature, a min-max normalization with the MinMaxScaler class of scikit-learn (with a range of 0.01 to one) was applied on the mean coefficients higher than zero and on the mean feature importance values of RF.

Performance Evaluation of CP Models. Two important metrics for the evaluation of CP models were calculated based on all predictions of the respective test sets: the validity and the efficiency. CP models are proven to be valid (i.e., guarantee the error rate indicated by the user) if the training and test data are exchangeable.[15] To achieve the indicated validity of the predictions, CP models output a set of class labels that can be empty, contain both labels, or only one of the labels (i.e., single class predictions). The validity is defined as the ratio of predictions containing the correct label (the “both” class set is therefore always correct and the “empty” set is always wrong). The efficiency measures the ratio of single class predictions (i.e., predictions containing only one class label) and, therefore, how predictive a model for a given endpoint is.

Additionally, the F1 score, Matthews correlation coefficient (MCC), specificity, sensitivity, and accuracy (both overall and independently for each class) were calculated (on the single class predictions only), to determine the model quality. The F1 score is the harmonic mean of precision and recall and is robust against data imbalance. The MCC considers all four classes of predictions (true positive, true negative, false positive, and false negative predictions) and takes values in the range of −1 to +1 (a value of +1 indicates perfect prediction). This metric is also robust against data imbalance. The specificity is determined by the proportion of inactive compounds correctly identified, while the sensitivity is determined by the proportion of active compounds correctly identified. The accuracy is defined as the ratio of correct predictions.

The CP models were evaluated at a significance level ε of 0.2, i.e., at a confidence level $(1 - \varepsilon)$ of 0.80. The set of predicted classes at this confidence level will contain the true class label in at least 80% of the cases (for valid models). This significance level was selected because it usually offers an adequate trade-off between efficiency and validity.[57,58]

The difference in performance between models with distinct descriptors was evaluated with the nonparametric Mann−Whitney U test.[59] For each pair of models compared, the distribution of values obtained in the different CV runs for a given performance metric (e.g., efficiency) was given as input in the “mannwhitneyu” function implemented in SciPy.[60]

RESULTS AND DISCUSSION

In this study, we investigated if, and to what extent, the consideration of predicted bioactivities can improve the performance of in silico models for the prediction of the in vivo toxicity endpoints MNT, DILI, and DICC. To this end, we first trained CP models for 373 biological assays and applied them on the in vivo endpoint data sets for deriving the predicted bioactivities. For training the models for the three in vivo endpoints, we embedded three types of RF models in CP frameworks: (a) CHEM models based exclusively on chemical descriptors, (b) BIO models based exclusively on (predicted) bioactivity descriptors, and (c) CHEMBIO models based on the combination of both types of descriptors.

Chemical Space Analysis. In order to develop an understanding of the chemical space represented by the training data from the three in vivo endpoints (MNT, DILI, and DICC), we compared the overlap of the chemical space between the in vivo endpoint data sets and three reference data sets. The overlap between data sets serves as an indication of the relevance of models trained on the in vivo data sets for different chemical domains (pesticides, cosmetics, and drugs). The reference data sets represent pesticides (2417 compounds from the EPA's Office of Pesticide Programs), cosmetics (4503 cosmetics ingredients from the COSMOS database), and drugs (10,087 approved, experimental, or withdrawn drugs from DrugBank).

We found that the MNT data set covers 16% of the pesticides reference set, 10% of the cosmetics reference set, and 8% of the drugs reference set, considering exact matches only (exact matches defined as any pair of compounds with a Tanimoto coefficient of 1.00; Table 3). The DICC data set covers 34% of the drugs reference set but just 7 and 6% of the cosmetics and pesticides reference sets, respectively. The lowest coverage rates were observed for the DILI data set (as it is also the smallest data set), with just 6, 2, and 1% for the drugs, pesticides, and cosmetics reference sets, respectively.

Table 3. Percentage of Compounds in the Reference Data Sets Covered by Compounds in the Three In Vivo Endpoint Data Sets (MNT, DILI, DICC) at Given Similarity Thresholds

		endpoint		
parameter	Tanimoto coefficient threshold[a]	MNT	DILI	DICC
% coverage pesticides	1.0	16	2	6
	≥0.8	17	2	7
	≥0.6	29	3	11
	≥0.4	62	10	36
	≥0.2	99	85	97
% coverage cosmetics	1.0	10	1	7
	≥0.8	14	1	9
	≥0.6	29	3	17
	≥0.4	68	17	58
	≥0.2	99	89	99
% coverage drugs	1.0	8	7	34
	≥0.8	9	8	37
	≥0.6	16	15	51
	≥0.4	40	34	73
	≥0.2	99	96	100

[a]Tanimoto coefficients calculated from binary Morgan fingerprints (1024 bits and radius 2).

For assessing the structural relationships between the active and inactive compounds present in the MNT, DILI, and DICC in vivo data, we referred to PCA. The PCA was performed on selected interpretable molecular descriptors, which describe, e.g., the number of bonds, rings, and particular types of atoms in a molecule, or the polarity and solubility of the compounds (Table S4). The three in vivo toxicity data sets were combined (containing 4987 compounds) and used to perform the PCA.

The PCA plots reported in Figure 3 indicate that the physicochemical properties of the active and inactive compounds of the individual in vivo endpoint data sets are mostly similar, with only a few outliers. Outliers with high values for the first principal component (PC1, *x* axis) are molecules with high molecular weight. Outliers with low values in the second component of the PCA (PC2, *y* axis) are mostly acyclic and polar, while molecules with high values on this axis have a high number of rings. Most outliers are inactive on the three investigated endpoints. The loadings plots (indicating how strongly each descriptor influences a principal component) are provided in Figure S1.

In order to investigate the chemical space with regard to the full set of descriptors used for model training, we utilized UMAP to compare the two-dimensional projections of the CHEM and CHEMBIO descriptor sets. UMAP conducts a dimension reduction of the data while maintaining the pairwise distance structure among all samples. In general, no clear separation of activity classes emerged for any of the three endpoints. Moreover, no significant difference was observed in the projections derived from the two descriptor sets regarding their ability to cluster compounds with different activity labels. The resulting UMAP plots are provided in Figure S2.

The structural diversity within the individual compound sets was determined based on the distribution of pairwise Tanimoto coefficients (based on atom-pair fingerprints)[61] among (a) all pairs of active compounds, (b) all pairs of inactive compounds, and (c) all pairs consisting of one active and one inactive compound (Figure 4). For the three in vivo endpoints, the distribution of pairwise compound similarities shows a tailing toward low similarities for the three sets of compounds (a, b, and c), indicating a high molecular diversity in the data sets. It is also shown that compounds in one class are not more similar to each other than they are to compounds of the other class, since the distribution of similarities of the three subsets is in all cases comparable.

Hence, the classification of compounds in the active and inactive classes based only on their structural similarity is not straightforward and complementary information may be necessary for in silico methods to be able to differentiate between classes.

Performance of CP Models for Deriving the Predicted Bioactivities. With the aim to improve the predictive performance for in vivo toxicity endpoints, we included information about the outcome of the compounds in biological assays (obtained from the ToxCast database, eMolTox, eChemPortal, and other publications) as input for the in vivo toxicity CP models. To avoid increased sparsity of the data due to missing experimental values, a fingerprint based on predicted bioactivities was developed. More specifically, for each of the 373 collected biological assay data sets, a bioactivity CP model was trained on molecular fingerprints and physicochemical property descriptors (see Materials and Methods for details).

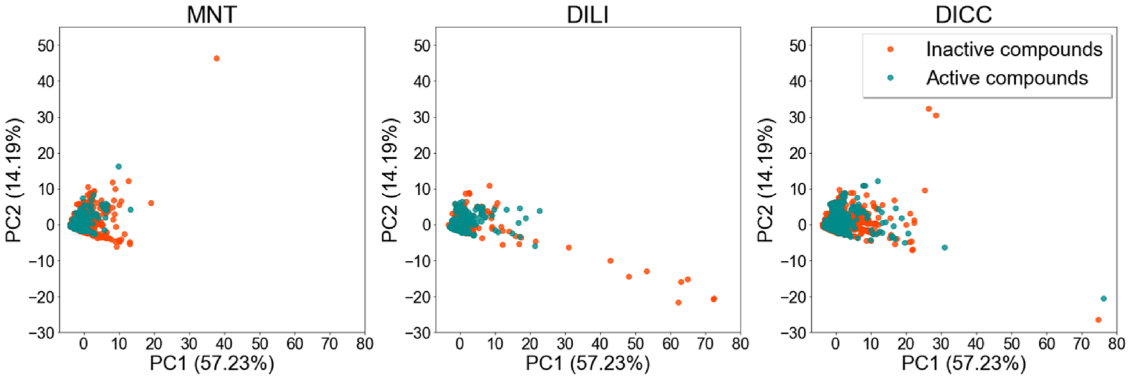


Figure 3. Principal component analysis based on a selection of interpretable molecular descriptors generated with RDKit on the merged in vivo toxicity data sets. Inactive compounds are colored in red and active compounds in green. The variance explained by the first two principal components is indicated in the axes.

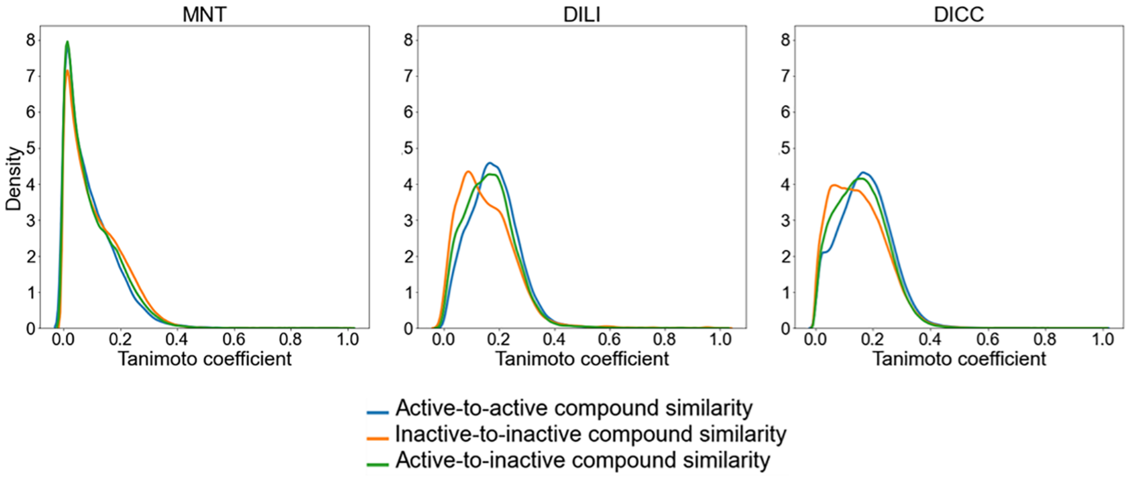


Figure 4. Distribution of pairwise Tanimoto coefficients based on atom-pair fingerprints for three types of compound pairs: (a) active-to-active (blue), (b) inactive-to-inactive (orange), and (c) active-to-inactive (green).

CP models are a type of confidence predictor that use the predictions made by the model on a set of compounds with known activities (calibration set) to rank and estimate the certainty of the predictions for new compounds[57] (see Materials and Methods section for details). These models output a set of labels (instead of only one label), which can contain one class (active or inactive), both classes, or none of them. Therefore, two important metrics for the evaluation of CP models are the validity, which measures the ratio of prediction sets containing the correct label (i.e., the "both" class is always correct), and the efficiency, which measures the ratio of single class predictions. Furthermore, the quality of the single class predictions (covered by the AD of the model) can be evaluated with common metrics like the F1 score or the MCC. The performance of models developed in this work was evaluated on the validity, efficiency, and F1 score results referring to mean values obtained by 5-fold CV at a significance level ε of 0.2 (Table S6). The MCC, specificity, sensitivity, and overall and class-wise mean accuracies of the single class predictions are also provided in Table S6.

The AD of ML models defines the region in chemical space where the model makes predictions with a given reliability. Depending on the focus of the study, there are different ways to define the AD. For example, unusual compounds or unreliable predictions can be flagged, assuming that they are likely outside the aforementioned region. In our case, error rate reduction is the focus of defining an AD; hence, it is mandatory to use confidence measures to identify objects close to the decision boundary and reject their predictions. A large benchmark study from Klingspohn et al. concluded that built-in class probability estimates performed constantly better than the alternatives (e.g., distance measures) in terms of error reduction.[62,63] In the current study, we are using the RF prediction score (best confidence measure for RF) as nonconformity measure for the CP. Hence, it is expected that no other nonconformity measure (or method) will outperform the prediction score to estimate the confidence of the predictions.

All 373 bioactivity CP models showed adequate mean validities for the given significance level (for which the expected validity is 0.80) that ranged from 0.78 to 0.83 (Figure 5) and thus obtained the defined error rate. The mean efficiency values and F1 scores spread over a wider range. There were 19 CP models (5%) with mean efficiencies lower than 0.70 (Figure 6). The lowest mean efficiency (0.41) was obtained for the ToxCast assay "ATG Ahr CIS dn". On the other hand, mean efficiencies higher than 0.90 were achieved for 101 CP models (27%), where the highest mean efficiency of 0.99 was obtained for the two eMolTox assays "Substrates of cytochrome P450 2C19" and "Differential cytotoxicity (isogenic chicken DT40 cell lines)", and the two ToxCast assays "TOX21 ERa LUC VM7 antagonist 0.1nM E2" and

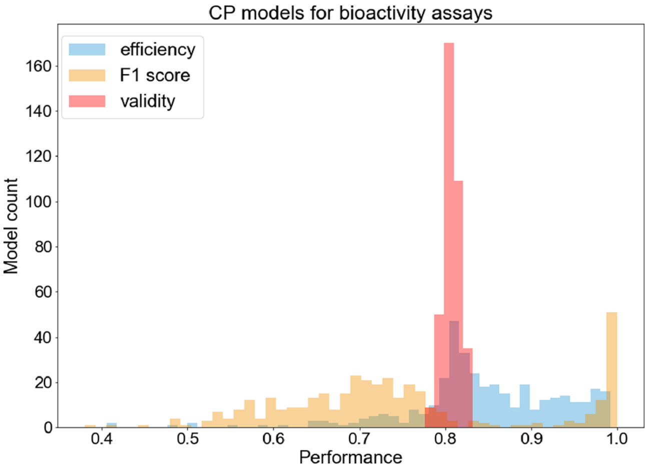


Figure 5. Histogram of the performance distribution of the CP models for the biological assays. All models were valid but their efficiencies and F1 scores showed a high degree of variability.

"TOX21 SBE BLA antagonist ratio". Hence, the ratio of single class predictions obtained by the bioactivity CP models was relatively high and only in a few cases the models showed poor efficiencies. In general, the models with the lowest mean efficiency had highly imbalanced classes and a low number of active compounds, while the contrary was observed for the models showing the highest mean efficiencies.

Seventy-seven models (21%) obtained F1 scores higher than 0.90, indicating a very good performance of these models on the single class predictions. There were 149 CP models (40%) with mean F1 scores lower than 0.70. Only for 15% of all models, the mean F1 scores were lower than 0.60, indicating poor performance. The worst-performing model was that for the ToxCast assay "ATG Ahr CIS dn" (mean F1 score of 0.38) and the best-performing ones for the eMolTox assays "Modulator of Neuropeptide Y receptor type 1", "Modulator of Urotensin II receptor", and "Agonist of Liver X receptor alpha" (F1 score of 1.00). One explanation for the good predictivity could be the fact that the chemical space of the active and inactive compounds is well differentiated (PCA plots of the chemical space of these data sets are shown in Figure S3). The classification of these compounds might therefore be easier than for data sets with more similar compounds between classes.

The performance of all CP models for the biological assays can be found in the Supplementary Information (Table S6).

In Vivo Toxicity CP Model Performance. The in vivo toxicity CP models were trained on three sets of descriptors: (i) the chemical descriptor set ("CHEM") comprising physicochemical features and the molecular fingerprint; (ii) the bioactivity descriptor set ("BIO") containing the predicted *p*-values for the biological endpoints; and (iii) the "CHEMBIO" descriptor set, which contains all features from both the CHEM and the BIO descriptor sets.

The number of features in the CHEM descriptor set (2171 features) is almost three times higher than the number of features of the BIO descriptor set (746 features), and together, they add up to 2917 features. The underrepresentation of bioactivity features in the CHEMBIO descriptor set and, more generally, the high number of total features could lead to a dilution of relevant information in the high-dimensional feature space. Moreover, since no prefiltering has been applied to the BIO descriptor set, some features may be redundant or less relevant for the specific in vivo endpoints. In order to test whether a reduction of the feature space could increase the performance of the in vivo toxicity CP models, we introduced a feature selection procedure based on a lasso model (which assigns coefficients, i.e., weights, to all features) that we applied prior to model training (see Materials and Methods for details).

With each of the CHEM, BIO, and CHEMBIO descriptor sets, two types of models were trained: (i) baseline models based on all features of the respective descriptor set (only filtering out those features with low variance; see Materials and Methods for details) and (ii) models based on a subset of features selected with a lasso model (built on the feature subset after the variance filter). For the model training, only those features with coefficients higher than zero in the lasso model were selected (see Materials and Methods for details).

The models based on the preselected set of features (based on (ii) lasso procedure) generally performed better (details will be discussed together with the individual in vivo endpoint performances below) and also present the computational advantage that only the *p*-values for the selected biological assays need to be computed to build the bioactivity descriptor for new compounds. Therefore, in the following paragraphs, only the results of these models will be further discussed. The results from the baseline models without feature selection with lasso (as described in (i)) are presented in Figure S3 and Table S7. All models were evaluated on the mean validity, efficiency, and F1 score (on the single class predictions) over 5-fold CV at a significance level ε of 0.2. The MCC is presented in Table 4 (see discussion in the next paragraph); specificity, sensitivity, and overall and per class accuracy data are provided in Table S8. The differences in the performance among models with

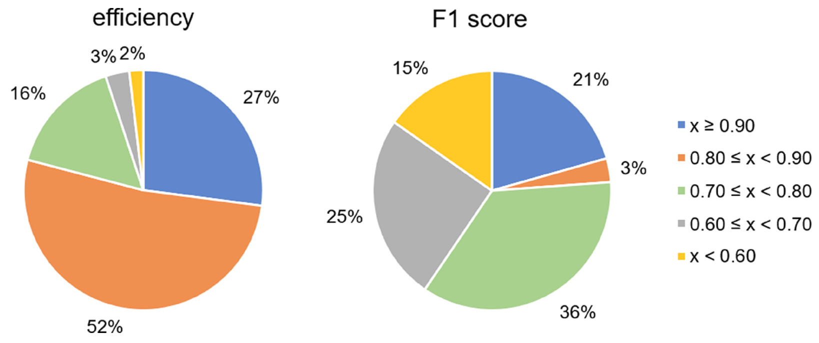


Figure 6. Percentage of the 373 bioactivity CP models showing mean efficiencies and mean F1 scores in the four given ranges.

Table 4. Average Performance of the CP Models Generated from a Selected Set of Features[a]

endpoint	descriptor	validity	STD validity	efficiency	STD efficiency	F1 score	STD F1 score	MCC	STD MCC
MNT	CHEM	0.77	0.02	0.76	0.05	0.61	0.02	0.28	0.05
	BIO	**0.82**	0.03	0.81	0.05	**0.70**	0.03	**0.46**	0.06
	CHEMBIO	0.81	0.03	**0.85**	0.03	**0.70**	0.03	0.44	0.07
DILI	CHEM	0.78	0.05	**0.91**	0.04	0.74	0.05	0.49	0.09
	BIO	**0.81**	0.04	0.83	0.07	0.76	0.04	0.53	0.07
	CHEMBIO	**0.81**	0.03	0.88	0.04	**0.77**	0.03	**0.55**	0.06
DICC	CHEM	0.79	0.02	0.84	0.02	0.72	0.03	0.46	0.05
	BIO	0.79	0.02	**0.96**	0.02	0.81	0.01	0.63	0.02
	CHEMBIO	0.79	0.02	0.94	0.01	**0.82**	0.01	**0.65**	0.03

[a] Mean and standard deviation (STD) calculated over a 5-fold CV. The highest mean per metric and endpoint is highlighted (bold).

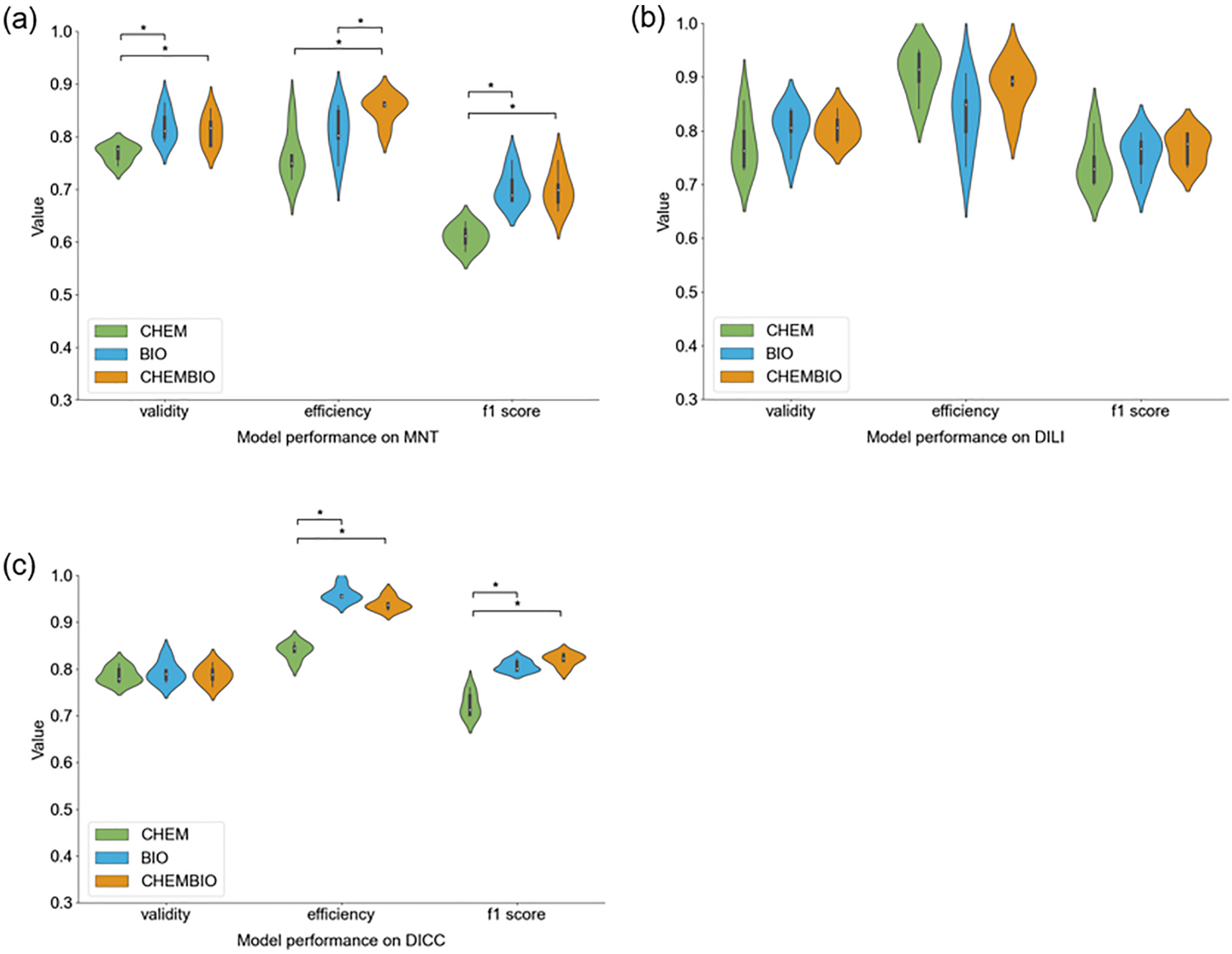


Figure 7. Distribution of the validity, efficiency, and F1 score values obtained within the 5-fold CV framework for the (a) MNT, (b) DILI, and (c) DICC CP models built on the different descriptor sets after feature selection. The CHEM descriptor set includes the molecular fingerprint and physicochemical descriptors; the BIO descriptor set includes the predicted *p*-values for a set of biological endpoints (bioactivity descriptor); the CHEMBIO descriptor set includes the previous two descriptor sets. Significant differences in the distribution (*p*-value <0.05) are denoted by a star.

different descriptors are evaluated with a Mann−Whitney U test at a *p*-value <0.05.

It is important to consider the inherent noise and errors in experimental data, which sets the upper limit for the models' performance, as a model can only be as good as the data it is trained on.[64] Hence, models trained on chemical descriptors only, which already achieve high performance rates, may not benefit from the addition of bioactivity fingerprints, as the noise in the data may be the bottleneck in these cases. Unfortunately, there is no information available on the noise in the data sets under investigation. Since studies such as that by Zhao et al.[65] have shown that low levels of noise are often tolerated by models while the removal of suspicious data points often decreases model performances and causes overfitting issues, we decided to not attempt to identify and remove noise in the data.

Table 5. Summary of Model Performances of the ChemBioSim Models and Existing Methods

endpoint	model	mean sensitivity	mean specificity	evaluation	modeling approach	comments
MNT	Yoo et al.	0.54–0.74	0.77–0.93	5% leave-many-out	Leadscope Enterprise and CASE Ultra software	variations related to different modeling approaches
	our method	0.78	0.76	5-fold CV	CP built on RF models	CHEMBIO model with feature selection
DILI	Ancuceanu et al.	0.83	0.66	nested CV	meta-model with a naïve Bayes model trained on output probabilities of 50 ML models	
	our method	0.78	0.78	5-fold CV	CP built on RF models	CHEMBIO model with feature selection
DICC	Cai et al.	0.69–0.75	0.72–0.81	5-fold CV	combined classifier using neural networks based on four single classifiers	results refer to five cardiological complications endpoints evaluated independently
	our method	0.83	0.86	5-fold CV	CP built on RF models	CHEMBIO model with feature selection

To evaluate the influence of the predicted bioactivities on model performance, the results of the in vivo toxicity CP models (including feature selection with lasso) based on the CHEM, BIO, and CHEMBIO descriptor sets were analyzed for each of the three in vivo endpoints.

For the MNT endpoint, the mean validities obtained by the two models including the BIO descriptor set (0.82 (±0.03) with the BIO and 0.81 (±0.03) with the CHEMBIO descriptor sets) were significantly higher than the validity of the model trained on the CHEM descriptor set alone (mean validity of 0.77 (±0.02); Figure 7, Table 4). While the validity of the model based on the CHEM descriptor set (0.77 ± 0.02) was lower than the expected validity at a significance level of 0.2 (i.e., expected validity of 0.80), the validity could be restored by adding the bioactivity descriptors (in the BIO and CHEMBIO descriptor sets). The mean efficiency obtained with the CHEMBIO descriptor set (0.85 ± 0.03) was significantly higher than the one obtained with the CHEM descriptor set alone (0.76 ± 0.05) but also higher than with the BIO descriptor set (0.81 ± 0.05) only. The two models including the BIO descriptor set significantly increased the predictive performance of the single class predictions, as reflected by the F1 score. More specifically, the model based on the CHEM descriptor set yielded a mean F1 score of 0.61 (±0.02), while the models based on the BIO and CHEMBIO descriptor sets both obtained a mean F1 score of 0.70 (±0.03). Thus, the model based on the CHEMBIO descriptor set not only increased the number of single class predictions but also the accuracy of these predictions.

The analysis of the number and type of the features selected with lasso for the models based on the CHEMBIO descriptor set showed that a total of 157 features were selected, 30 of which were bioactivity features (19%). Of the 15 features with the highest lasso coefficients, seven were bioactivity features and eight are chemical features (Table S10). Compared to the models without feature selection, the efficiency of the CHEMBIO MNT model including feature selection was significantly higher (0.07 higher mean efficiency). Otherwise, the difference in the performance between models with and without feature selection (only comparing models with the same descriptor set) was not significant.

The DILI models obtained mean validities between 0.78 (±0.05; with the CHEM descriptor set) and 0.81 (±0.04 with the BIO and ±0.03 with the CHEMBIO descriptor sets). The distribution of efficiencies within the CV from models trained on the different descriptor sets was not significantly different. However, the mean efficiencies ranged from 0.83 (±0.07; with the BIO descriptor set) to 0.91 (±0.04; with the CHEM descriptor set; Figure 7). The mean F1 score based on the single class predictions was also comparable among the three models and was between 0.74 (±0.05) with the CHEM descriptor set and 0.77 (±0.03) with the CHEMBIO descriptor set. Although there is no model for DILI that outperforms the others, the models including biological features (CHEMBIO and BIO) have a slightly higher mean validity and F1 score (while a lower number of single class predictions is obtained compared to the model trained on the CHEM descriptor set). Thus, both the BIO and CHEM descriptor sets may contain relevant—but not complementing—information for the prediction of the DILI endpoint. In the model based on the CHEMBIO descriptor set, 648 features were selected by the lasso model, 59 of which were bioactivity features (9%). The smaller percentage of bioactivity features (compared to the number of features in the MNT model) among the selected features also reflects the fact that including the bioactivity descriptor set did not improve the performance of the models significantly for this endpoint. Nevertheless, among the 15 features with the highest lasso coefficients, nine were bioactivity features and six were chemical features (Table S10). Compared to the models without feature selection by lasso, the efficiencies of the BIO and CHEMBIO models were significantly increased (up to 0.08 higher mean efficiency).

In the case of the DICC endpoint, the models based on each of the three different descriptor sets yielded mean validities of 0.79 (±0.02). The models trained on the BIO and CHEMBIO descriptor sets showed significantly higher efficiencies (0.96 ± 0.02 and 0.94 ± 0.01, respectively) than the model trained on the CHEM descriptor set (0.84 ± 0.02, Figure 7). Not only the ratio of single class predictions (i.e., efficiency) was improved in the models including the BIO descriptor set but also the quality of these predictions. The two models including the BIO descriptor set obtained significantly higher F1 scores (mean F1 score of 0.81 (±0.01) with the BIO and 0.82 (±0.01) with the CHEMBIO descriptor sets) than the model based on the CHEM descriptor set (mean F1 score of 0.72 (±0.03)). The significantly better performance of the DICC models making use of the BIO descriptor set over the DICC models based solely on CHEM descriptors is also reflected in the nature of the features selected by lasso from the CHEMBIO descriptor set: among the 666 features selected, 101 are bioactivity features (15%). Furthermore, the bioactivity features were assigned high coefficients by the lasso model, and from the top 50 features (ranked after the mean coefficient), 34 belong to the bioactivity descriptor set (15 out of the top 15 features are bioactivity features; Table S10). Compared to the models without feature selection, the efficiencies of the two models including the BIO descriptor set decreased when the feature selection was included (up to 0.03 lower mean efficiency). Also, the mean F1 score of the model trained on the CHEM descriptor set decreased by 0.04 when including the feature selection procedure. One possible explanation for the decrease in performance is the potential overfitting of the models without feature selection to the training data due to the high number of features.

In summary, it was shown that the addition of bioactivity descriptors in the form of predicted *p*-values for a set of biological assay outcomes can improve the predictive ability of CP models with regard to the number of single class predictions as well as to the quality of these predictions. However, this effect and its magnitude were endpoint-dependent and not achieved in all cases. It was also shown that including feature selection before training, the models can help to discard irrelevant features favoring those more relevant for the specific endpoint.

Comparison with Existing Models. Several in silico models for MNT, DILI, and DICC are described in the literature (Table 5). However, to our knowledge, no CP models have been previously developed for these endpoints. Note that the studies cannot be directly compared given differences in underlying data and techniques. Also, the evaluation of the models differs since the quality of the predictions of CP models is in general evaluated on single class predictions only. However, considering existing models can help to put the results of this study into context.

Yoo et al.[33] recently collected data sets for MNT in mice and rats, containing 1001 and 127 compounds, respectively. They developed statistical-based models with the Leadscope and CASE Ultra software combined with different balancing techniques for the mouse data set based on chemical features and structural alerts (functional groups or substructures frequently found in molecules eliciting a determined biological effect). Their best model with regard to specificity (i.e., the proportion of inactive compounds correctly identified) on a 5% leave-many-out framework yielded a mean specificity of 0.93 but a mean sensitivity (i.e., the proportion of active compounds correctly identified) of only 0.54. The model with the highest sensitivity (and also with the most balanced sensitivity-to-specificity ratio) obtained a mean specificity of 0.77 and a mean sensitivity of 0.74. To train our MNT CP models, we combined the mouse and rat data sets from Yoo et al. and added further data sources (see Materials and Methods section) to obtain a data set with 1791 compounds. For comparison, the specificity and sensitivity values obtained by our models trained on the CHEMBIO descriptor set including feature selection with lasso were also calculated (Table 5). The CHEMBIO model for the MNT endpoint yielded a mean specificity of 0.76 and a mean sensitivity of 0.78. Thus, compared to the most balanced model of Yoo et al., our model showed a slightly higher sensitivity and comparable specificity on a significantly larger data set (790 additional compounds).

Several in silico models with adequate predictive performance have already been reported for the DILI endpoint.[66−68] In a recent study based on the same data set as our models, Ancuceanu et al.[68] built 267 different models combining feature selection techniques with ML algorithms. Meta-models using the output of 50 ML models as input for a final model were developed. Their meta-model with the highest balanced accuracy (0.75) evaluated in a nested CV was built training a naïve Bayes model on output probabilities of 50 ML models. This model yielded a mean specificity of 0.66 and a mean sensitivity of 0.83. In comparison, our CHEMBIO DILI model yielded a much more balanced sensitivity-to-specificity ratio. The mean specificity and sensitivity obtained by our model were both 0.78.

Although in silico models for cardiological complications are more scarce, Cai et al.[35] compiled data sets for five different cardiological complications (hypertension, arrhythmia, heart block, cardiac failure, and myocardial infarction), on which our DICC data set is based, and developed a combined classifier for each of the five endpoints. These classifiers yielded mean specificities between 0.72 and 0.81 and sensitivities between 0.69 and 0.75 (depending on the endpoint). Our CHEMBIO model for the DICC endpoint yielded a mean specificity of 0.86 and a mean sensitivity of 0.83, thus increasing the performance observed for the previous models (especially with regard to the sensitivity).

Overall, our models yielded a high balanced sensitivity-to-specificity ratio and often generally good performance. It should be considered that the existing models used for comparison were built on complicated and highly optimized model architectures for the studied endpoint, while in this study, we used simple RF models without hyperparameter optimization embedded in a CP framework for the predictions with the aim of comparing the different descriptors.

Analysis of Feature Importance to Discover Biological Relationships. Understanding which bioactivity features are most important for the prediction can help to identify the most relevant assays for an endpoint and to discover unknown biological relationships. From the complete CHEMBIO descriptor set (i.e., the descriptor set without feature selection with lasso), we analyzed the 15 descriptors that were assigned the highest feature importance values by the RF model. The reason for using the complete set of CHEMBIO descriptors instead of the subset of features selected by the lasso method (which generally yields better performing models) is that the lasso model discards highly correlated features during the feature selection. Therefore, feature importance analysis involving a descriptor preselection with lasso may lead to an underestimation of the importance of some of the features.

The RF model for the MNT endpoint ranked the features from (i) the AMES assay, (ii) the eMolTox assay for mutagenicity, and (iii) the eMolTox assay for agonism on the p53 signaling pathway as the most important features (Table S9). These three in vitro assays are known to be biologically related to the MNT endpoint: the AMES and mutagenicity assays evaluate the genotoxic potential of compounds in vitro by measuring the capability of substances to induce mutations in bacterial strains. DNA damage leading to these gene mutations could also cause the chromosome aberrations observed in the MNT.[69] The tumor suppressor p53 has the capacity of preventing the proliferation of cells with a damaged genome and is also referred to as “the Guardian of the Genome”.[70] The p53 signaling pathway is activated i.a. when DNA damage accumulates in a cell. As a result, a mechanism of cell cycle arrest, cellular senescence or apoptosis is initiated. Since genotoxic damage is one of the primary triggers of the activation of the p53 signaling pathway, the detection of agonism of the p53 pathway could be an indication of the genotoxic activity of a compound, which could also lead to micronuclei formation in vivo.[71] The contribution of the p53 signaling pathway for the prediction of MNT in vivo is highlighted by the high feature importance assigned to features corresponding to further assays related to this endpoint (ToxCast assays "TOX21 p53 BLA p3 ratio," "TOX21 p53 BLA p5 ratio," and "TOX21 p53 BLA p2 ratio" (each measuring the ratio of two measurements with the inducible beta lactamase (BLA) reporter); Table S9). Also the biological function of the constitutive androstane receptor (CAR) and aryl hydrocarbon receptor (AhR) could explain the high importance assigned by the model to the ToxCast assay "TOX21 CAR antagonist" and the eMolTox assay "Activator the aryl hydrocarbon receptor (AhR) signaling pathway." The AhR and the CAR are ligand-activated transcription factors functioning as sensors of xenobiotic compounds. Upon activation of these receptors, i.a. the expression of enzymes involved in the metabolism of xenobiotic compounds, is upregulated.[72,73] The downregulation of enzymes detoxifying compounds (or their metabolites) mediated by CAR antagonists, as well as the AhR-mediated upregulation of enzymes activating compounds to form genotoxic metabolites seem to contribute to the observed effects in the MNT. The remaining features among the 15 most important features for MNT are related to the eMolTox assay "Antagonist of the farnesoid-X-receptor (FXR) signaling pathway." The FXR, also called bile acid receptor, is a nuclear receptor that regulates, among other things, bile acid and hepatic triglyceride levels.[74] Its possible biological relationship with genotoxicity has not been reported so far (to the best of our knowledge). Comparing the features with the highest feature importance values with RF to the features with the highest lasso coefficients during feature selection (Table S9 and Table S10), an overlap of the assays for AMES, the p53 signaling pathway, and the CAR antagonism was observed, highlighting the relevance of these biological endpoints for the prediction of MNT.

Although in the case of DILI the performance of the RF models making use of bioactivity descriptors was not superior (see Table 4) over that of the models trained on chemical descriptors only, 14 out of the 15 top-ranked features were bioactivity features. The highest feature importance was obtained for a chemical descriptor (smr VSA10) that captures polarizability properties of compounds. The bioactivity features ranked at positions 3 and 4 are the two p-values (of the active and inactive classes) for human oral bioavailability, respectively. Since any compound must be absorbed and distributed in order to be able to elicit any kind of biological response, bioavailability is essential to induce liver injury. Moreover, orally administered substances undergo a hepatic first pass before they become systemically available. Other than that, several features related to modulators of G protein-coupled receptors were of high importance (see Table S9). Despite the lack of a clear biological relationship between liver injury and opioid receptors (kappa, mu and delta) or muscarinic acetylcholine receptors (M2, M3, M4 and M5), the activity of compounds against these receptors showed high predictivity for DILI. Between the features with the highest feature importance values for RF and the features with the highest lasso coefficients (Table S10) we found an overlap of descriptors for the bioavailability, mu opioid receptor, and muscarinic acetylcholine receptor assays.

Consistent with the DILI model, also the DICC model assigned high ranks (rank 1 and rank 4) to the two features related to human oral bioavailability (i.e., *p*-values for the active and inactive classes). The importance of these features is plausible, as substances first need to be absorbed in order to be able to elicit any response. We also found the ToxCast assay "TOX21 ERa LUC VM7 agonist", an assay for detecting agonists of the estrogen receptor alpha, to have a high relevance value assigned by the DICC RF model. There is evidence about the important correlation between estrogen levels and cardiovascular diseases.[75] The cardioprotective effects shown by estrogen derive from the increase in angiogenesis and vasodilation as well as the decrease in oxidative stress and fibrosis. Another feature that was assigned a high importance is agonism on the retinoid X receptor (RXR; eMolTox assay "Agonist of the RXR signaling pathway" and ToxCast assay "TOX21 RXR BLA agonist"). Following its activation, RXR forms homo- or heterodimers with other nuclear receptors (e.g., thyroid hormone receptor), regulating the transcription of several genes and therefore playing a role in diverse body functions. It has been shown that the functionality of RXR influences, for example, the composition of the cardiac myosin heavy chain, thus affecting the correct functionality of the heart.[76] The induction of phospholipidosis, a phospholipid storage disorder in the lysosomes, was also assigned a high importance value by the DICC RF model. There is still controversy whether phospholipidosis is a toxic or an adaptive response, as it does not necessarily result in target organ toxicity.[77] However, a high percentage of compounds inducing phospholipidosis has been found to also inhibit the human ether-à-go-go-related gene (hERG),[78,79] an ion channel that contributes to the electrical activity of the heart. Inhibitors of hERG can lead to fatal irregularities in the heartbeat (ventricular tachyarrhythmia).[80] Another bioactivity that was of high importance for the prediction of cardiological complications is the agonism of the p53 signaling pathway (ToxCast assays "TOX21 p53 BLA p2 ratio" and "TOX21 p53 BLA p3 ratio"). As already mentioned, the p53 transcription factor is related to tumor suppressor mechanisms of the cell, but it also inhibits the hypoxia-inducible factor-1 (Hif-1) in the heart. Inhibition of Hif-1 hinders cardiac angiogenesis (i.e., the formation of new blood vessels). This hindrance presents a problem in cases of cardiac hypertrophy (an adaptive response to increased cardiac workload), as blood pressure overload can lead to heart failure.[81,82] Recently, heart failure has also been related to DNA damage. Higo et al.[83] showed that single-stranded DNA damage is accumulated in cardiomyocytes of failing hearts and that mice lacking DNA repair mechanisms are more prone to heart failure. This relationship between DNA damage and heart failure could also explain the high relevance assigned by the DICC RF model to the three features related to genotoxicity in cells lacking DNA damage response pathways (from the eMolTox assay "Differential cytotoxicity against isogenic chicken DT40 cell lines with known DNA damage response pathways - Rad54Ku70 mutant cell line" and the ToxCast assay "TOX21 DT40 657"). The comparison of the most important features for RF with the features assigned the highest coefficients by lasso showed an overlap of the descriptors for the bioavailability and estrogen agonism assays. Furthermore, other assays related to genotoxicity (and correlated with the ones with a high feature importance shown in Table S9) were also assigned high coefficients.

Apart from biological relationships, there are other factors that may influence the importance values assigned to the respective bioactivity features. One should keep in mind that predicted *p*-values are used for the representation of biological properties, not measured bioactivity values. This means that

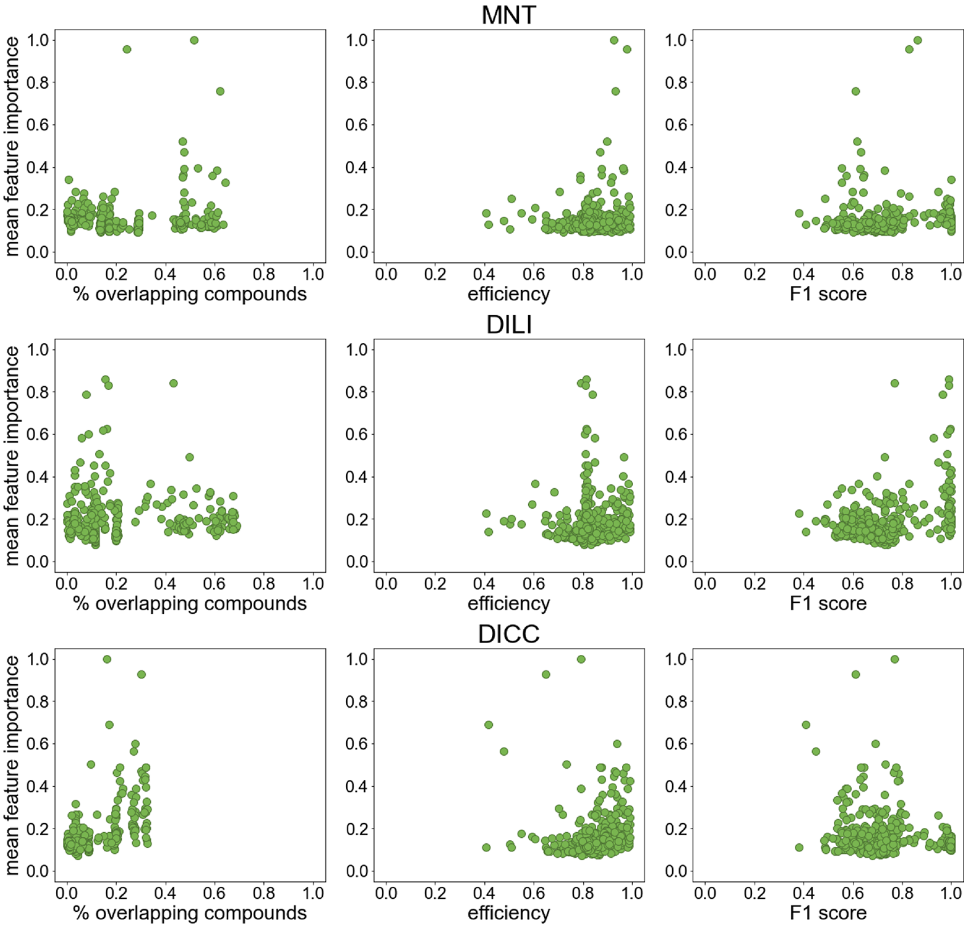


Figure 8. Mean feature importance reported by the RF model for the bioactivity descriptors in relationship with the percentage of overlapping compounds (of the in vivo data set), the efficiency and F1 score of the models for each biological assay. For each of the 373 biological assays, the highest mean feature importance of the two *p*-values used as descriptors (for the active and inactive classes of each assay) was taken. The feature importance values were normalized with a min-max normalization (from 0.01 to 1; see Materials and Methods section) for easier comparison.

feature importance values are likely affected by the performance and applicability of the individual models used for predicting the *p*-values. For example, bioactivity features based on biological assay data sets with a strong overlap with the in vivo endpoint data sets could be favored by a model, as the predicted *p* -values for structurally similar compounds are likely more accurate (as they were also used to train the bioactivity model itself).

Therefore, the overlap between the in vivo endpoint data set and the data sets of the selected biological assays, as well as the performance of the biological assay models, was analyzed to test possible correlations with the assigned feature coefficients. Overall, we observed no strong correlation between the extent of overlaps in the data and the assigned feature importance values. Also, no pronounced correlation between the performance of the bioactivity CP models and the feature importance values was observed (Figure 8), but bioactivity descriptors predicted with models showing lower efficiencies also often resulted in less important features.

The comparison between the data set overlap and model performance with the coefficients obtained during feature selection with the lasso model showed similar effects and correlations to the feature importance of the RF models discussed here (Figure S5).

In general, it was observed that the most predictive biological assays have a clear biological relationship with the corresponding in vivo endpoint. However, not all biological assays with a clear biological connection were assigned a high feature importance. Moreover, biological assays with a less obvious biological relationship were sometimes given a high relevance, as they may describe a more general behavior of the compounds in biological systems. These less obvious relationships could also reflect yet unknown effects and point to further lines of investigation.

■ CONCLUSIONS

In this work, we have explored the potential of incorporating predicted bioactivities to improve the in silico prediction of in vivo endpoints beyond the level of accuracy reached by established molecular descriptors. More specifically, in the first part of this work, we collected 373 compound data sets with biological assay outcomes from the literature for modeling, and in the second part, we developed an elaborate conformal prediction framework in combination with the random forest algorithm, with the aim to identify the scope and limitations of the developed bioactivity descriptors for in vivo toxicity prediction on three selected in vivo endpoints (MNT, DILI, and DICC).

Overall, valid in vivo toxicity CP models could be produced with the different descriptors for all endpoints. For the MNT and DICC endpoints, the incorporation of predicted bioactivities was highly beneficial for the performance of the models. Compared to the models based only on chemical descriptors, the mean efficiencies of the models for MNT and DICC including bioactivity descriptors increased by 0.09 (from 0.76 to 0.85) and 0.12 (from 0.84 to 0.96), respectively. The mean F1 scores also increased by 0.09 (from 0.61 to 0.70) and 0.10 (from 0.72 to 0.82), respectively. The performance of the model for the DILI endpoint did not significantly improve by the integration of bioactivity descriptors, but a slight increase in the mean F1 score was also observed. The chemical and bioactivity descriptors may not complement each other for the prediction of DILI, which could explain the lower influence of the selected descriptor set on the performance. The prediction of the DILI endpoint may be especially challenging due to the nature of the data set, which has a reduced number of compounds and combines substances producing major and less severe effects in the active class. Further investigations are needed to determine how to improve the learning power of ML models for this endpoint.

In general, applying a feature selection procedure with a lasso model prior to model training with RF increased the mean efficiency of the models (up to 0.08 for the MNT and DILI endpoints). Feature selection proved especially beneficial in the models including the bioactivity descriptor set, as some biological assays may be redundant or not related to the in vivo endpoints.

The analysis of the most important features of the models based on the CHEMBIO descriptor set for each in vivo endpoint showed that generally these features had an explainable relationship with the biological mechanism eliciting the toxicity in vivo. For instance, some of the most important features for the MNT, an in vivo genotoxicity assay, are measuring genotoxicity in vitro or are involved in tumor suppressor mechanisms of the cells. In the case of the DILI and DICC endpoints, human oral bioavailability was ranked as one of the most important features, as bioavailability is an unavoidable requirement to elicit organ toxicity. Furthermore, the high feature importance assigned to assays with a less clear biological relationship could hint to unknown interactions that might help to better understand the toxic mechanisms.

The determination of which features will make the largest impact on the in vivo models prior to model development remains a difficult task since there are many factors influencing the relevance of the bioactivity features. However, using biological assays with known biological relevance for the in vivo endpoints is a well-suited approach. Also, for which in vivo endpoints the bioactivity descriptor will enhance the results cannot be predicted beforehand and may require evaluation case-by-case.

Overall, the approach presented in this work shows how the prediction of in vivo endpoints, which entail a high complexity due to all interactions taking place in biological systems, can be improved by the incorporation of bioactivity fingerprints. Moreover, the CP framework supporting the developed models also presents the advantage of intrinsically defining the applicability domain of these models and ensuring a defined error rate. Our approach also showed that bioactivity information can be included in the form of predicted probabilities, opening the possibility to apply these models directly on new compounds, without the need to fill their bioactivity profile experimentally. The bioactivity CP models for deriving the predicted bioactivities as well as the in vivo toxicity CP models trained on the different descriptor sets (and including feature selection with lasso) are freely available for download (https://doi.org/10.5281/zenodo.4761225).[84]

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.1c00451.

Loading plot of the PCA; UMAP projections for the three in vivo endpoints on the CHEM and the CHEMBIO descriptor sets; PCA of the biological assays with a mean F1 score of 1.0; distribution of the performance over 5-fold CV for the models for the three in vivo endpoints without feature preselection with lasso; scatter plots of lasso coefficients vs data set overlap and model performance of the models for the biological assays (PDF)

Download links, queries, and MD5 file checksum of the in vivo endpoint data sets; download links, queries, and MD5 file checksum of the biological assay data sets; data set information for the biological assays used to build the bioactivity descriptors; list of molecular descriptors used in principal component analysis; average performance of the CP models built on the biological assay data sets average performance of the CP for the three in vivo endpoints without feature preselection with lasso; top 15 features with the highest feature importance values for the three in vivo endpoints; top 15 features with the highest lasso coefficients for the three in vivo endpoints (ZIP)

KNIME workflow for the preparation of the molecular structures and calculation of the CHEM descriptors (ZIP)

■ AUTHOR INFORMATION

Corresponding Authors

Johannes Kirchmair − *Department of Pharmaceutical Sciences, Faculty of Life Sciences, University of Vienna, Vienna 1090, Austria;* orcid.org/0000-0003-2667-5877; Phone: +43 1-4277-55104; Email: johannes.kirchmair@univie.ac.at

Miriam Mathea − *BASF SE, Ludwigshafen am Rhein 67063, Germany;* orcid.org/0000-0002-3214-1487; Phone: +49 621 60-29054; Email: miriam.mathea@basf.com

Authors

Marina Garcia de Lomana − *BASF SE, Ludwigshafen am Rhein 67063, Germany; Department of Pharmaceutical Sciences, Faculty of Life Sciences, University of Vienna, Vienna 1090, Austria;* orcid.org/0000-0002-9310-7290

Andrea Morger − *In Silico Toxicology and Structural Bioinformatics, Institute of Physiology, Berlin 10117, Germany;* orcid.org/0000-0003-4774-6291

Ulf Norinder − *MTM Research Centre, School of Science and Technology, Örebro University, Örebro SE-70182, Sweden;* orcid.org/0000-0003-3107-331X

Roland Buesen − *BASF SE, Ludwigshafen am Rhein 67063, Germany;* orcid.org/0000-0002-6531-1200

Robert Landsiedel − *BASF SE, Ludwigshafen am Rhein 67063, Germany;* orcid.org/0000-0003-3756-1904

Andrea Volkamer − *In Silico Toxicology and Structural Bioinformatics, Institute of Physiology, Berlin 10117, Germany;* orcid.org/0000-0002-3760-580X

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.1c00451

Funding

AM and AV thank BMBF (grant No. 031A262C), the HaVo-Stiftung as well as BASF for funding.

Notes

The authors declare no competing financial interest.

All data sets used in this study are publicly available and the original sources, as well as the processing workflow, are described in detail in the Materials and Methods section and Tables S1 and S2. Note that due to licensing issues of the original data, the data used in this study cannot explicitly be added as supporting information. The KNIME workflow used for preprocessing the structures and calculating the chemical descriptors is provided in the Supplementary Information. The workflow and parameters used for developing the models and necessary for reproducing the results are described in the Materials and Methods section. All bioactivity CP models (used for deriving the bioactivity descriptor) and in vivo toxicity CP models can be freely downloaded from https://doi.org/10.5281/zenodo.4761225. We believe that the scientific value of this study resides in researching the improvement of models for in vivo toxicity prediction with a fully automated workflow using predicted bioactivity fingerprints. Thus, this finding can be extrapolated to further endpoints and using other ML methods (and parameters).

M.G.d.L., R.B., R.L., and M.M. are employed at BASF SE. U.N. performed research and served as a consultant for BASF SE.

■ ABBREVIATIONS

AD, applicability domain; ADME, administration, distribution, metabolism, and excretion; AhR, aryl hydrocarbon receptor; BLA, beta lactamase; CA, chromosome aberration; CAR, constitutive androstane receptor; CCRIS, Chemical Carcinogenesis Information System; CP, conformal prediction; CTD, Comparative Toxicogenomics Database; CV, cross-validation; DICC, drug-induced cardiological complications; DILI, drug-induced liver injury; ECHA, European Chemicals Agency; EFSA, European Food Safety Authority; EPA, Environmental Protection Agency; FDA, U.S. Food and Drug Administration; FXR, farnesoid-X-receptor; hERG, human ether-à-go-go-related gene; Hif-1, hypoxia-inducible factor-1; ICH, International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use; MCC, Matthews correlation coefficient; ML, machine learning; MM, mammalian mutagenicity; MNT, micronucleus test; nc, nonconformity; NTP, National Toxicology Program; Papp, apparent permeability coefficient; PCA, principal component analysis; OECD, Organisation for Economic Co-operation and Development; RF, random forest; RXR, retinoid X receptor; STD, standard deviation; TSHR, thyroid stimulating hormone receptor

■ REFERENCES

(1) Akhtar, A. The Flaws and Human Harms of Animal Experimentation. *Camb. Q. Healthc. Ethics* **2015**, *24*, 407−419.

(2) Van Norman, G. A. Limitations of Animal Studies for Predicting Toxicity in Clinical Trials: Is it Time to Rethink our Current Approach? *JACC Basic Transl. Sci.* **2019**, *4*, 845−854.

(3) Van Norman, G. A. Limitations of Animal Studies for Predicting Toxicity in Clinical Trials: Part 2: Potential Alternatives to the Use of Animals in Preclinical Trials. *JACC Basic Transl. Sci.* **2020**, *5*, 387−397.

(4) Gleeson, M. P.; Modi, S.; Bender, A.; Robinson, R. L. M.; Kirchmair, J.; Promkatkaew, M.; Hannongbua, S.; Glen, R. C. The Challenges Involved in Modeling Toxicity Data In Silico: A Review. *Curr. Pharm. Des.* **2012**, *18*, 1266−1291.

(5) Doke, S. K.; Dhawale, S. C. Alternatives to Animal Testing: A Review. *Saudi Pharm. J.* **2015**, *23*, 223−229.

(6) Hand, D. J.; Mannila, H.; Smyth, P., *Principles of Data Mining*; Bradford Book: 2001.

(7) Hansch, C.; Fujita, T. p-σ-π Analysis. A Method for the Correlation of Biological Activity and Chemical Structure. *J. Am. Chem. Soc.* **1964**, *86*, 1616−1626.

(8) Helal, K. Y.; Maciejewski, M.; Gregori-Puigjané, E.; Glick, M.; Wassermann, A. M. Public Domain HTS Fingerprints: Design and Evaluation of Compound Bioactivity Profiles from PubChem's Bioassay Repository. *J. Chem. Inf. Model.* **2016**, *56*, 390−398.

(9) Riniker, S.; Wang, Y.; Jenkins, J. L.; Landrum, G. A. Using Information from Historical High-Throughput Screens to Predict Active Compounds. *J. Chem. Inf. Model.* **2014**, *54*, 1880−1891.

(10) Guo, Y.; Zhao, L.; Zhang, X.; Zhu, H. Using a hybrid read-across method to evaluate chemical toxicity based on chemical structure and biological data. *Ecotoxicol. Environ. Saf.* **2019**, 178−187.

(11) Xu, T.; Ngan, D. K.; Ye, L.; Xia, M.; Xie, H. Q.; Zhao, B.; Simeonov, A.; Huang, R. Predictive Models for Human Organ Toxicity Based on *In Vitro* Bioactivity Data and Chemical Structure. *Chem. Res. Toxicol.* **2020**, *33*, 731−741.

(12) Liu, J.; Patlewicz, G.; Williams, A. J.; Thomas, R. S.; Shah, I. Predicting Organ Toxicity Using in Vitro Bioactivity Data and Chemical Structure. *Chem. Res. Toxicol.* **2017**, *30*, 2046−2059.

(13) Su, R.; Wu, H.; Liu, X.; Wei, L. Predicting Drug-Induced Hepatotoxicity Based on Biological Feature Maps and Diverse Classification Strategies. *Brief. Bioinform.* **2021**, *22*, 428−437.

(14) Norinder, U.; Spjuth, O.; Svensson, F. Using Predicted Bioactivity Profiles to Improve Predictive Modeling. *J. Chem. Inf. Model.* **2020**, *60*, 2830−2837.

(15) Vovk, V., Gammerman, A., Shafer, G., *Algorithmic Learning in a Random World*; Springer US: 2005.

(16) Norinder, U.; Boyer, S. Binary Classification of Imbalanced Datasets Using Conformal Prediction. *J. Mol. Graphics Modell.* **2017**, *72*, 256−265.

(17) Vovk, V. Conditional Validity of Inductive Conformal Predictors. *Mach. Learn.* **2013**, *92*, 349−376.

(18) OECD, *Test No. 474: Mammalian Erythrocyte Micronucleus Test*; 2016.

(19) Chen, M.; Vijay, V.; Shi, Q.; Liu, Z.; Fang, H.; Tong, W. FDA-Approved Drug Labeling for the Study of Drug-Induced Liver Injury. *Drug Discovery Today* **2011**, *16*, 697−703.

(20) Fung, M.; Thornton, A.; Mybeck, K.; Wu, J. H.-H.; Hornbuckle, K.; Muniz, E. Evaluation of the Characteristics of Safety Withdrawal of Prescription Drugs from Worldwide Pharmaceutical Markets-1960 to 1999. *Drug Inf. J.* **2001**, *35*, 293−317.

(21) Watkins, P. B. Drug Safety Sciences and the Bottleneck in Drug Development. *Clin. Pharmacol. Ther.* **2011**, *89*, 788−790.

(22) ECHA *Guidance on Information Requirements and Chemical Safety Assessment. Chapter R.7a: Endpoint specific guidance.*; 2017.

(23) ICHS2(R1) Guidance on Genotoxicity Testing and Data Interpretation for Pharmaceuticals Intended for Human Use; *International Conference on Harmonization of Technical Requirements for Registration of Pharmaceuticals for Human Use*; ICH Expert Working Group 2011.

(24) EPA, U. S., *ToxCast & Tox21 Data Spreadsheet from invitrodb_v3.3*. Retrieved from https://www.epa.gov/chemical-research/toxicity-forecaster-toxcasttm-data on September 7, 2020. Data released September 2020.

(25) Ji, C.; Svensson, F.; Zoufir, A.; Bender, A. eMolTox: Prediction of Molecular Toxicity With Confidence. *Bioinformatics (Oxford, England)* **2018**, *34*, 2508−2509.

(26) *eChemPortal*. https://www.echemportal.org/echemportal/(accessed August 6, 2020).

(27) Falcón-Cano, G.; Molina, C.; Cabrera-Pérez, M. A. ADME Prediction With KNIME: Development and Validation of a Publicly Available Workflow for the Prediction of Human Oral Bioavailability. *J. Chem. Inf. Model.* **2020**, *60*, 2660−2667.

(28) Benigni, R.; Laura Battistelli, C.; Bossa, C.; Giuliani, A.; Fioravanzo, E.; Bassan, A.; Fuart Gatnik, M.; Rathman, J.; Yang, C.; Tcheremenskaia, O. *Evaluation of the Applicability of Existing (Q)SAR Models for Predicting the Genotoxicity of Pesticides and Similarity Analysis Related With Genotoxicity of Pesticides for Facilitating of Grouping and Read Across*; EFSA Support. Publ.: 2019, *1598E*.

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

(30) Wang, N.-N.; Dong, J.; Deng, Y.-H.; Zhu, M.-F.; Wen, M.; Yao, Z.-J.; Lu, A.-P.; Wang, J.-B.; Cao, D.-S. ADME Properties Evaluation in Drug Discovery: Prediction of Caco-2 Cell Permeability Using a Combination of NSGA-II and Boosting. *J. Chem. Inf. Model.* **2016**, *56*, 763−773.

(31) Garcia de Lomana, M.; Weber, A. G.; Birk, B.; Landsiedel, R.; Achenbach, J.; Schleifer, K.-J.; Mathea, M.; Kirchmair, J. In Silico Models to Predict the Perturbation of Molecular Initiating Events Related to Thyroid Hormone Homeostasis. *Chem. Res. Toxicol.* **2021**, *34*, 396−411.

(32) Broccatelli, F.; Carosati, E.; Neri, A.; Frosini, M.; Goracci, L.; Oprea, T. I.; Cruciani, G. A Novel Approach for Predicting P-Glycoprotein (ABCB1) Inhibition Using Molecular Interaction Fields. *J. Med. Chem.* **2011**, *54*, 1740−1751.

(33) Yoo, J. W.; Kruhlak, N. L.; Landry, C.; Cross, K. P.; Sedykh, A.; Stavitskaya, L. Development of Improved QSAR Models for Predicting the Outcome of the in Vivo Micronucleus Genetic Toxicity Assay. *Regul. Toxicol. Pharmacol.* **2020**, *113*, 104620.

(34) Chen, M.; Suzuki, A.; Thakkar, S.; Yu, K.; Hu, C.; Tong, W. DILIrank: The Largest Reference Drug List Ranked by the Risk for Developing Drug-Induced Liver Injury in Humans. *Drug Discovery Today* **2016**, *21*, 648−653.

(35) Cai, C.; Fang, J.; Guo, P.; Wang, Q.; Hong, H.; Moslehi, J.; Cheng, F. In Silico Pharmacoepidemiologic Evaluation of Drug-Induced Cardiovascular Complications Using Combined Classifiers. *J. Chem. Inf. Model.* **2018**, *58*, 943−956.

(36) Mattingly, C. J.; Rosenstein, M. C.; Colby, G. T.; Forrest, J. N., Jr.; Boyer, J. L. The Comparative Toxicogenomics Database (CTD): a Resource for Comparative Toxicological Studies. *J. Exp. Zool. A Comp. Exp. Biol.* **2006**, *305A*, 689−692.

(37) Kuhn, M.; Letunic, I.; Jensen, L. J.; Bork, P. The SIDER Database of Drugs and Side Effects. *Nucleic Acids Res.* **2016**, *44*, D1075−D1079.

(38) Tatonetti, N. P.; Ye, P. P.; Daneshjou, R.; Altman, R. B. Data-Driven Prediction of Drug Effects and Interactions. *Sci. Transl. Med.* **2012**, *4*, 125ra31.

(39) Cheng, F.; Li, W.; Wang, X.; Zhou, Y.; Wu, Z.; Shen, J.; Tang, Y. Adverse Drug Events: Database Construction and in Silico Prediction. *J. Chem. Inf. Model.* **2013**, *53*, 744−752.

(40) Wishart, D. S.; Knox, C.; Guo, A. C.; Cheng, D.; Shrivastava, S.; Tzur, D.; Gautam, B.; Hassanali, M. DrugBank: A Knowledgebase for Drugs, Drug Actions and Drug Targets. *Nucleic Acids Res.* **2008**, *36*, D901−D906.

(41) *Pesticide Chemical Search*; EPA: https://iaspub.epa.gov/apex/pesticides/f?p=chemicalsearch:1 (accessed February 1, 2021).

(42) Williams, A. J.; Grulke, C. M.; Edwards, J.; McEachran, A. D.; Mansouri, K.; Baker, N. C.; Patlewicz, G.; Shah, I.; Wambaugh, J. F.; Judson, R. S.; Richard, A. M. The CompTox Chemistry Dashboard: A Community Data Resource for Environmental Chemistry. *Aust. J. Chem.* **2017**, *9*, 61. (accessed EFebruary 1, 2021)

(43) COSMOS *cosmetics database*. http://www.cosmostox.eu/home/welcome/(accessed February 1, 2021).

(44) *DrugBank* Version 5.1.5. https://www.drugbank.ca (accessed February 14, 2020).

(45) Kim, S.; Thiessen, P. A.; Cheng, T.; Yu, B.; Bolton, E. E. An Update on PUG-REST: RESTful Interface for Programmatic Access to PubChem. *Nucleic Acids Res.* **2018**, *46*, W563−w570.

(46) *NCI/CADD Chemical Identifier Resolver*. https://cactus.nci.nih.gov/chemical/structure(accessed October 2019).

(47) Landrum, G., *RDKit: Open-Source Cheminformatics Software*. 2016.

(48) Berthold, M. R.; Cebron, N.; Dill, F.; Gabriel, T. R.; Kötter, T.; Meinl, T.; Ohl, P.; Sieb, C.; Thiel, K.; Wiswedel, B., KNIME: The Konstanz Information Miner. In *Studies in Classification, Data Analysis, and Knowledge Organization (GfKL 2007)*, Springer: 2007; p (Version 4.1.1.).

(49) *Standardizer was used for structure canonicalization and transformation, JChem 3.5.0, ChemAxon* (http://www.chemaxon.com), JChem 3.5.0.

(50) *The pKa Plugin was used for the calculation of the pKa constant value of molecules, JChem 3.5.0, ChemAxon* (http://www.chemaxon.com), JChem 3.5.0.

(51) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, É. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825−2830. (Version 0.22.1)

(52) McInnes, L.; Healy, J.; Melville, J., Umap: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv preprint* **2018**, arXiv:1802.03426.

(53) Linusson, H., *Nonconformist*. 2015 (Version 2.1.0).

(54) Carlsson, L.; Eklund, M.; Norinder, U. In Aggregated Conformal Prediction, *Artificial Intelligence Applications and Innovations. AIAI 2014*. IFIP Advances in Information and Communication Technology, v., Ed. Springer, Berlin, Heidelberg: *2014*; pp. 231−240.

(55) Norinder, U.; Carlsson, L.; Boyer, S.; Eklund, M. Introducing Conformal Prediction in Predictive Modeling. A Transparent and Flexible Alternative to Applicability Domain Determination. *J. Chem. Inf. Model.* **2014**, *54*, 1596−1603.

(56) Tibshirani, R. Regression Shrinkage and Selection via the Lasso. *J. R. Statist. Soc. B* **2014**, *58*, 267−288.

(57) Cortés-Ciriano, I.; Bender, A., Concepts and Applications of Conformal Prediction in Computational Drug Discovery. *arXiv preprint* **2019**, *abs/1908.03569*.

(58) Svensson, F.; Aniceto, N.; Norinder, U.; Cortes-Ciriano, I.; Spjuth, O.; Carlsson, L.; Bender, A. Conformal Regression for Quantitative Structure−Activity Relationship Modeling—Quantifying Prediction Uncertainty. *J. Chem. Inf. Model.* **2018**, *58*, 1132−1140.

(59) Mann, H. B.; Whitney, D. R. On a Test of Whether one of Two Random Variables is Stochastically Larger Than the Other. *Ann. Math. Statist.* **1947**, *18*, 50−60.

(60) Virtanen, P.; Gommers, R.; Oliphant, T. E.; Haberland, M.; Reddy, T.; Cournapeau, D.; Burovski, E.; Peterson, P.; Weckesser, W.; Bright, J.; van der Walt, S. J.; Brett, M.; Wilson, J.; Millman, K. J.; Mayorov, N.; Nelson, A. R. J.; Jones, E.; Kern, R.; Larson, E.; Carey, C. J.; Polat, İ.; Feng, Y.; Moore, E. W.; VanderPlas, J.; Laxalde, D.; Perktold, J.; Cimrman, R.; Henriksen, I.; Quintero, E. A.; Harris, C. R.; Archibald, A. M.; Ribeiro, A. H.; Pedregosa, F.; van Mulbregt, P.; Vijaykumar, A.; Bardelli, A. P.; Rothberg, A.; Hilboll, A.; Kloeckner, A.; Scopatz, A.; Lee, A.; Rokem, A.; Woods, C. N.; Fulton, C.; Masson, C.; Häggström, C.; Fitzgerald, C.; Nicholson, D. A.; Hagen, D. R.; Pasechnik, D. V.; Olivetti, E.; Martin, E.; Wieser, E.; Silva, F.; Lenders, F.; Wilhelm, F.; Young, G.; Price, G. A.; Ingold, G.-L.; Allen, G. E.; Lee, G. R.; Audren, H.; Probst, I.; Dietrich, J. P.; Silterra, J.; Webber, J. T.; Slavič, J.; Nothman, J.; Buchner, J.; Kulick, J.; Schönberger, J. L.; de Miranda Cardoso, J. V.; Reimer, J.; Harrington, J.; Rodríguez, J. L. C.; Nunez-Iglesias, J.; Kuczynski, J.; Tritz, K.; Thoma, M.; Newville, M.; Kümmerer, M.; Bolingbroke, M.; Tartre, M.; Pak, M.; Smith, N. J.; Nowaczyk, N.; Shebanov, N.; Pavlyk, O.; Brodtkorb, P. A.; Lee, P.; McGibbon, R. T.; Feldbauer, R.; Lewis, S.; Tygier, S.; Sievert, S.; Vigna, S.; Peterson, S.; More, S.; Pudlik, T.;

Oshima, T.; Pingel, T. J.; Robitaille, T. P.; Spura, T.; Jones, T. R.; Cera, T.; Leslie, T.; Zito, T.; Krauss, T.; Upadhyay, U.; Halchenko, Y. O.; Vázquez-Baeza, Y.; SciPy, C. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nat. Methods* **2020**, *17*, 261−272. (Version 1.4.1.).

(61) Carhart, R. E.; Smith, D. H.; Venkataraghavan, R. Atom Pairs as Molecular Features in Structure-Activity Studies: Definition and Applications. 1985, *25*, 64−73.

(62) Klingspohn, W.; Mathea, M.; ter Laak, A.; Heinrich, N.; Baumann, K. Efficiency of Different Measures for Defining the Applicability Domain of Classification Models. *Aust. J. Chem.* **2017**, *9*, 44.

(63) Sushko, I.; Novotarskyi, S.; Körner, R.; Pandey, A. K.; Cherkasov, A.; Li, J.; Gramatica, P.; Hansen, K.; Schroeter, T.; Müller, K.-R.; Xi, L.; Liu, H.; Yao, X.; Öberg, T.; Hormozdiari, F.; Dao, P.; Sahinalp, C.; Todeschini, R.; Polishchuk, P.; Artemenko, A.; Kuz'min, V.; Martin, T. M.; Young, D. M.; Fourches, D.; Muratov, E.; Tropsha, A.; Baskin, I.; Horvath, D.; Marcou, G.; Muller, C.; Varnek, A.; Prokopenko, V. V.; Tetko, I. V. Applicability Domains for Classification Problems: Benchmarking of Distance to Models for Ames Mutagenicity Set. *J. Chem. Inf. Model.* **2010**, *50*, 2094−2111.

(64) Kramer, C.; Kalliokoski, T.; Gedeck, P.; Vulpetti, A. The Experimental Uncertainty of Heterogeneous Public K(i) Data. *J. Med. Chem.* **2012**, *55*, 5165−5173.

(65) Zhao, L.; Wang, W.; Sedykh, A.; Zhu, H. Experimental Errors in QSAR Modeling Sets: What We Can Do and What We Cannot Do. *ACS Omega* **2017**, *2*, 2805−2812.

(66) He, S.; Ye, T.; Wang, R.; Zhang, C.; Zhang, X.; Sun, G.; Sun, X. An In Silico Model for Predicting Drug-Induced Hepatotoxicity. *Int. J. Mol. Sci.* **2019**, *20*, 1897.

(67) Wang, Y.; Xiao, Q.; Chen, P.; Wang, B. In Silico Prediction of Drug-Induced Liver Injury Based on Ensemble Classifier Method. *Int. J. Mol. Sci.* **2019**, *20*, 4106.

(68) Ancuceanu, R.; Hovanet, M. V.; Anghel, A. I.; Furtunescu, F.; Neagu, M.; Constantin, C.; Dinu, M. Computational Models Using Multiple Machine Learning Algorithms for Predicting Drug Hepatotoxicity With the DILIrank Dataset. *Int. J. Mol. Sci.* **2020**, *21*, 2114.

(69) Kirkland, D.; Zeiger, E.; Madia, F.; Corvi, R. Can in Vitro Mammalian Cell Genotoxicity Test Results be Used to Complement Positive Results in the Ames Test and Help Predict Carcinogenic or in Vivo Genotoxic Activity? II. Construction and Analysis of a Consolidated Database. *Mutat. Res., Genet. Toxicol. Environ. Mutagen.* **2014**, *775-776*, 69−80.

(70) Toufektchan, E.; Toledo, F. The Guardian of the Genome Revisited: p53 Downregulates Genes Required for Telomere Maintenance, DNA Repair, and Centromere Structure. *Cancers* **2018**, *10*, 135.

(71) Kumari, R.; Kohli, S.; Das, S. p53 Regulation Upon Genotoxic Stress: Intricacies and Complexities. *Mol. Cell. Oncol.* **2014**, *1*, No. e969653.

(72) Yang, H.; Wang, H. Signaling Control of the Constitutive Androstane Receptor (CAR). *Protein Cell* **2014**, *5*, 113−123.

(73) Brauze, D.; Rawłuszko, A. A. The Effect of Aryl Hydrocarbon Receptor Ligands on the Expression of Polymerase (DNA Directed) Kappa (Polκ), Polymerase RNA II (DNA Directed) Polypeptide A (PolR2a), CYP1B1 and CYP1A1 Genes in Rat Liver. *Environ. Toxicol. Pharmacol.* **2012**, *34*, 819−825.

(74) Rizzo, G.; Renga, B.; Mencarelli, A.; Pellicciari, R.; Fiorucci, S. Role of FXR in Regulating Bile Acid Homeostasis and Relevance for Human Diseases. *Curr. Drug Targets Immune Endocr. Metabol. Disord.* **2005**, *5*, 289−303.

(75) Iorga, A.; Cunningham, C. M.; Moazeni, S.; Ruffenach, G.; Umar, S.; Eghbali, M. The Protective Role of Estrogen and Estrogen Receptors in Cardiovascular Disease and the Controversial Use of Estrogen Therapy. *Biol. Sex Differ.* **2017**, *8*, 33.

(76) Long, X.; Boluyt, M. O.; O'Neill, L.; Zheng, J.-S.; Wu, G.; Nitta, Y. K.; Crow, M. T.; Lakatta, E. G. Myocardial Retinoid X Receptor, Thyroid Hormone Receptor, and Myosin Heavy Chain Gene Expression in the Rat During Adult Aging. *J. Gerontol. A* **1999**, *54*, B23−B27.

(77) Reasor, M. J.; Hastings, K. L.; Ulrich, R. G. Drug-Induced Phospholipidosis: Issues and Future Directions. *Expert Opin. Drug Saf.* **2006**, *5*, 567−583.

(78) Sun, H.; Xia, M.; Shahane, S. A.; Jadhav, A.; Austin, C. P.; Huang, R. Are hERG Channel Blockers Also Phospholipidosis Inducers? *Bioorg. Med. Chem. Lett.* **2013**, *23*, 4587−4590.

(79) Slavov, S.; Stoyanova-Slavova, I.; Li, S.; Zhao, J.; Huang, R.; Xia, M.; Beger, R. Why are Most Phospholipidosis Inducers Also hERG Blockers? *Arch. Toxicol.* **2017**, *91*, 3885−3895.

(80) Calderone, V.; Testai, L.; Martinotti, E.; Del Tacca, M.; Breschi, M. C. Drug-Induced Block of Cardiac hERG Potassium Channels and Development of Torsade de Pointes Arrhythmias: the Case of Antipsychotics. *J. Pharm. Pharmacol.* **2005**, *57*, 151−161.

(81) Sano, M.; Minamino, T.; Toko, H.; Miyauchi, H.; Orimo, M.; Qin, Y.; Akazawa, H.; Tateno, K.; Kayama, Y.; Harada, M.; Shimizu, I.; Asahara, T.; Hamada, H.; Tomita, S.; Molkentin, J. D.; Zou, Y.; Komuro, I. p53-Induced Inhibition of Hif-1 Causes Cardiac Dysfunction During Pressure Overload. *Nature* **2007**, *446*, 444−448.

(82) Mak, T. W.; Hauck, L.; Grothe, D.; Billia, F. p53 Regulates the Cardiac Transcriptome. *Proc. Natl. Acad. Sci.* **2017**, *114*, 2331.

(83) Higo, T.; Naito, A. T.; Sumida, T.; Shibamoto, M.; Okada, K.; Nomura, S.; Nakagawa, A.; Yamaguchi, T.; Sakai, T.; Hashimoto, A.; Kuramoto, Y.; Ito, M.; Hikoso, S.; Akazawa, H.; Lee, J.-K.; Shiojima, I.; McKinnon, P. J.; Sakata, Y.; Komuro, I. DNA Single-Strand Break-Induced DNA Damage Response Causes Heart Failure. *Nat. Commun.* **2017**, *8*, 15104.

(84) Garcia de Lomana, M.; Morger, A.; Norinder, U.; Buesen, R.; Landsiedel, R.; Volkammer, A.; Kirchmair, J.; Mathea, M. ChemBioSim: Biological Assay and in Vivo Toxicity Models. *Zenodo.* **2021**, DOI: 10.5281/zenodo.4761226.

4.3 Mitigation of data drift effects on conformal prediction models

In order to get predictions with the expected error rate, ML models should only be applied if test and training data have the same distribution. This requirement limits the applicability of predictive ML models and makes them unsuitable as they get outdated over time or need to be applied to slightly diverging data. However, the amount of available data similar to the new test data is often not enough to train robust ML models for the specific application.[50, 81]

In the following study, a strategy for mitigating the effects of data drifts on CP models, without the need of completely retraining them, was investigated on two common scenarios: (i) changes in the descriptor space over time (simulated with a time-split on data from 12 ChEMBL[82] endpoints) and (ii) the application of models trained on public domain data to predict proprietary data (analyzed on two BASF SE inhouse data sets). In these two scenarios, test data deviates from the training data as the assay set up or conditions change, or as the descriptor space covered by the different data sets diverges. Data drifts can be recognized by a decrease in the expected validity (i.e. 1 - significance level) of CP models, which indicates that the exchangeability assumption between calibration and test set is not hold. With the aim of restoring the validity of the models on the new test data, the calibration set was replaced by samples from the same (or more similar) distribution as the test data, while keeping the trained ML model unchanged. The influence of this recalibration strategy was investigated on the balanced validity, efficiency and accuracy obtained for models trained on data from 14 toxicity-related endpoints and based on CHEMBIO descriptors (developed as part of this thesis; section 4.2.).

[P3] Studying and Mitigating the Effects of Data Drifts on ML Model Performance at the Example of Chemical Toxicity Data

Andrea Morger^{+}, Marina Garcia de Lomana^{+}, Ulf Norinder, Fredrik Svensson, Johannes Kirchmair, Miriam Mathea and Andrea Volkamer

$^{+}$ these authors contributed equally to this work

Scientific Reports, 2022

Available at https://doi.org/10.1038/s41598-022-09309-3

Contribution:

A. Morger, M. Garcia de Lomana, M. Mathea and A. Volkamer conceptualized the research. A. Morger, M. Garcia de Lomana, U. Norinder, F. Svensson, J. Kirchmair, M. Mathea and A. Volkamer designed the experiments. A. Morger and M. Garcia de Lomana compiled the data sets. A. Morger developed the machine learning models. A. Morger and M. Garcia de Lomana evaluated the results and wrote the manuscript, with contributions from U. Norinder, F. Svensson, J. Kirchmair, M. Mathea and A. Volkamer. M. Mathea and A. Volkamer supervised the work.

The following article was reprinted with permission from:

Morger, A.; Garcia de Lomana, M.; Norinder, U.; Svensson, F.; Kirchmair, J.; Mathea, M. and Volkamer, A. Studying and Mitigating the Effects of Data Drifts on ML Model Performance at the Example of Chemical Toxicity Data, *Sci. Rep.*, **2022**, 12, 7244.

scientific reports

OPEN

Studying and mitigating the effects of data drifts on ML model performance at the example of chemical toxicity data

Andrea Morger[1,8], Marina Garcia de Lomana[2,3,8], Ulf Norinder[4,5,6], Fredrik Svensson[7], Johannes Kirchmair[3], Miriam Mathea[2]✉ & Andrea Volkamer[1]✉

Machine learning models are widely applied to predict molecular properties or the biological activity of small molecules on a specific protein. Models can be integrated in a conformal prediction (CP) framework which adds a calibration step to estimate the confidence of the predictions. CP models present the advantage of ensuring a predefined error rate under the assumption that test and calibration set are exchangeable. In cases where the test data have drifted away from the descriptor space of the training data, or where assay setups have changed, this assumption might not be fulfilled and the models are not guaranteed to be valid. In this study, the performance of internally valid CP models when applied to either newer time-split data or to external data was evaluated. In detail, temporal data drifts were analysed based on twelve datasets from the ChEMBL database. In addition, discrepancies between models trained on publicly-available data and applied to proprietary data for the liver toxicity and MNT in vivo endpoints were investigated. In most cases, a drastic decrease in the validity of the models was observed when applied to the time-split or external (holdout) test sets. To overcome the decrease in model validity, a strategy for updating the calibration set with data more similar to the holdout set was investigated. Updating the calibration set generally improved the validity, restoring it completely to its expected value in many cases. The restored validity is the first requisite for applying the CP models with confidence. However, the increased validity comes at the cost of a decrease in model efficiency, as more predictions are identified as inconclusive. This study presents a strategy to recalibrate CP models to mitigate the effects of data drifts. Updating the calibration sets without having to retrain the model has proven to be a useful approach to restore the validity of most models.

Machine learning (ML) models are usually trained—and evaluated—on available historical data, and then used to make predictions on prospective data. This strategy is often applied in the context of toxicological data to predict potential toxic effects of novel compounds[1–6]. Internal cross-validation (CV) is a common practice for assessing the performance of ML models. When applying the model to new data, it is advisable to observe the applicability domain (AD) of an ML model[7,8]. The AD determines the compound space and the response value (label) range in which the model makes reliable predictions[9]. Investigating classification models, Mathea et al.[8] distinguished AD methods that rely on novelty from those relying on confidence estimation. Novelty detection methods focus on the fit of the query samples to the given descriptor space. Confidence estimation methods determine the reliability of the predictions by taking into account that samples may be well-embedded in the descriptor space but be unusual in terms of their class membership.

[1]In Silico Toxicology and Structural Bioinformatics, Institute of Physiology, Charité Universitätsmedizin Berlin, Berlin 10117, Germany. [2]BASF SE, 67056 Ludwigshafen, Germany. [3]Division of Pharmaceutical Chemistry, Department of Pharmaceutical Sciences, University of Vienna, Vienna 1090, Austria. [4]Department of Pharmaceutical Biosciences, Uppsala University, Uppsala 751 24, Sweden. [5]Dept Computer and Systems Sciences, Stockholm University, Kista 164 07, Sweden. [6]MTM Research Centre, School of Science and Technology, 701 82 Örebro, Sweden. [7]Alzheimer's Research UK UCL Drug Discovery Institute, London WC1E 6BT, UK. [8]These authors contributed equally: Andrea Morger and Marina Garcia de Lomana. ✉email: miriam.mathea@basf.com; andrea.volkamer@charite.de

A popular method for confidence estimation is conformal prediction (CP)[10,11]. The framework of an inductive conformal predictor uses three types of datasets: proper training, calibration, and test set. The proper training set is used to train an underlying ML model. With this model, predictions are made for the calibration and test set. According to the rank that is obtained for the prediction outcome of the test compound as compared to the calibration set, so-called p-values are calculated to give an estimate of the likelihood of a compound to belong to a certain class. If a significance level, i.e. an expected error rate, is defined, the compounds are assigned labels for those classes where the p-value is larger than the significance level. For binary classification, the possible prediction sets are 'empty' ({Ø}), 'single class' ({0}, {1}), and 'both' ({0,1}). Single class predictions indicate a confident prediction for a certain class. Additionally, the CP framework recognises compounds for which it cannot make a reliable prediction ({Ø}) and compounds at the decision boundary, for which the predictions are reliable but indecisive ({0,1}). Provided that the calibration and test data are exchangeable, the framework of the conformal predictor is mathematically proven to yield valid predictions at a given significance level[10,11].

The performance and AD of a model are determined by the quality and quantity of the data it has been trained on. One prerequisite for building good models is the availability of large, well-distributed and consistent datasets. To assemble large datasets, modellers often need to collect data from different sources, e.g. data which were produced in different assays or laboratories or over longer periods of time[12–14]. However, data from different sources and data taken at different time points may have distinct property distributions, reflecting, for example, the evolution of research interests or changes in assay technologies and protocols[15,16]. Since the predictivity of ML models is constrained by their AD, data drifts pose a challenge to modelling tasks, including toxicity or bioactivity prediction.

When ML models are validated using CV, the data is usually randomly split into training and test data. The resulting sets intrinsically stem from the same distribution and, typically, high model performance on the test set is observed. Nevertheless, it has been shown that model performance can be substantially lower for datasets obtained by time split or datasets from other sources[5,17–19]. This may be an indicator that the distribution of the data has changed. Hence, it is essential to confirm that ML models can be applied to a specific dataset and to determine the confidence in the predictions.

The data drifts, which challenge the underlying ML models, do also affect conformal predictors when the trained and calibrated models are applied to a new dataset. In previous work[17], a new strategy was introduced to mitigate the effects related to data drifts by exchanging the calibration set with data closer to the holdout set. The study built on the Tox21 data challenge[2], which was invented to support and compare ML models for twelve toxicity endpoints and included three subsequently released datasets. We showed that internally valid CP models resulted in poor performance when predicting the holdout data. The observed effects were associated to data drifts between datasets and could be mitigated by exchanging the calibration set with the intermediate set—without the need to retrain the models.

Here, we aim to expand and challenge our previous analysis on the recalibration strategy by a wider variety of datasets, beyond Tox21. Furthermore, we utilise enhanced compound encodings which combine molecular fingerprints with predicted bioactivity descriptors, specifically designed for toxicity prediction[12,20].

First, temporal data drifts are studied using twelve toxicity-related endpoint datasets extracted from the ChEMBL database[21,22]. The ChEMBL database is a manually-curated data collection containing quantitative and qualitative measurements for more than two million compounds tested in up to more than 1.3 million assays. The large size of the database makes it a primary data resource for ML, in particular in the context of activity prediction[23–25] and target prediction[26,27]. Moreover, it is one of only a few publicly-available bioactivity databases that provides temporal information on bioactivity measurements in the form of the publication date.

In the second part of this study, the impact on model validity from using data with differences in assay setups and source laboratories is investigated. Therefore, models were trained on public datasets for two in vivo endpoints, i.e., 'liver toxicity' and 'in vivo micro nucleus test (MNT)', and applied to predict proprietary data. Both, liver toxicity and MNT are in vivo endpoints with high relevance for the registration and authorisation of new chemical compounds[28–30].

Data and methods

In this section, first, the used datasets are described, including chemical structure standardisation, data splitting and compound encoding. Second, the CP setup together with the individual modelling strategies is explained. Finally, further data analysis and visualisation methods are outlined.

Data assembly.

Dataset description, collection and filtration. Large toxicity-related ChEMBL datasets. To investigate temporal data drifts, the ChEMBL database[21,22] version 26 was queried following the protocol described by Škuta et al.[31]. In short, the presented 29 target datasets each containing more than 1000 compounds were downloaded with measured pIC50 values and publication year. Next, the datasets were cleaned to handle molecules contained more than once in a target dataset, called duplicates (see Supplementary Material Section A1.1). Then, compounds were standardised (see Section "Data assembly") and the datasets temporally split (see Section "Data assembly"). Activity was assigned based on the target family and following the activity cutoff suggestions by the *Illuminating the Druggable Genome* Consortium[32]. Only datasets with more than 50 active and 50 inactive compounds in the holdout set were retained for the study. From the resulting 20 target datasets, only twelve targets that are linked to toxicity[33,34] (see Supplementary Material Section A1.1 and Table 1) were selected for this study.

Public and inhouse datasets for liver toxicity and MNT. To assess drifts between data originating from different sources, public and proprietary datasets for two in vivo endpoints (drug-induced liver injury (DILI) and MNT)

ChEMBL ID	Name	Active compounds	Inactive compounds
CHEMBL220	Acetylcholinesterase (human)	1334	1339
CHEMBL4078	Acetylcholinesterase (fish)	2056	1755
CHEMBL5763	Cholinesterase	1871	884
CHEMBL203	EGFR erbB1	2955	1104
CHEMBL206	Estrogen receptor alpha	826	590
CHEMBL279	VEGFR 2	3782	1392
CHEMBL230	Cyclooxygenase-2	1148	872
CHEMBL340	Cytochrome P450 3A4	2501	815
CHEMBL240	hERG	1601	3375
CHEMBL2039	Monoamine oxidase B	1413	1121
CHEMBL222	Norepinephrine transporter	406	1160
CHEMBL228	Serotonin transporter	449	1662

Table 1. ChEMBL target datasets used to investigate data drifts including the target name and the number of active and inactive compounds.

were collected. For CP model training, the same public datasets for DILI and MNT were used as compiled and described by Garcia de Lomana et al.[12]. After data pre-processing and deduplication the respective DILI dataset consists of 445 active and 247 inactive compounds; the MNT dataset of 316 active and 1475 inactive compounds (see Supplementary Material Section A1.2 for more details). Note that we will from here on refer to the DILI endpoint as 'liver toxicity'.

Two proprietary BASF SE inhouse datasets for liver toxicity and MNT in vivo were used as independent test and update sets. In short, liver toxicity was measured in rats according to the OECD Guidelines 407, 408 and 422[35–37]. MNT was determined in mice following the OECD Guideline 474[29], or in (non-GLP) screening assays. The liver toxicity dataset contains 63 active and 77 inactive compounds and the MNT dataset contains 194 active and 172 inactive compounds, after data pre-processing and deduplication (see Supplementary Material Section A1.3).

Chemical structure standardisation. Standardisation of chemical structures was conducted as described by Garcia de Lomana et al.[12]. Briefly, the SMILES of each of the compounds were standardised with the ChemAxon Standardizer[38] node in KNIME[39,40] to remove solvents and salts, annotate aromaticity, neutralise charges and mesomerise structures (i.e. taking the canonical resonant form of the molecules). Multi-component compounds as well as compounds containing any unwanted element were removed from the dataset. Canonical SMILES were derived for the standardised compounds and used for removing duplicates. In cases where duplicate SMILES had conflicting labels, the compounds were removed from the dataset.

Compound encoding. To encode the molecules for training the CP models, the 'CHEMBIO' descriptors developed by Garcia de Lomana et al.[12] were used. These descriptors combine chemical with predicted bioactivity descriptors to describe the compounds. The chemical descriptor comprises a 2048-byte Morgan count fingerprint (with a radius of 2 bonds)[41] and a 119-byte physicochemical property descriptor from RDKit[42] (calculated with KNIME[39,40]).

For deriving the bioactivity descriptors, Garcia de Lomana et al.[12] first built binary classification CP models for 373 in vitro toxicological endpoints, such as cytotoxicity, genotoxicity and thyroid hormone homeostasis (including datasets from ToxCast[33], eMolTox[43] and literature). These models were used to calculate the p-values (see Section "Conformal prediction") per target endpoint model and class, thus, resulting in a 746-byte predicted bioactivity fingerprint. For use in CP-based toxicity prediction model studies, the individual features were scaled prior to model training. The combination of chemical and bioactivity descriptors into the 2913-byte 'CHEMBIO' descriptor has shown superior performance in the CP study by Garcia de Lomana et al.[12] and was therefore used in this study.

Data splitting. After standardising the compounds (see Section "Data assembly"), the target datasets derived from the ChEMBL database were temporally split based on the publication year. This resulted in four subsets, i.e. train, update1, update2, and holdout set, see Table 2. Thus, compounds were ordered by publication year (old to new).

Aiming for the typically used ratio of 80% training (further divided in 70% proper training and 30% calibration set) and 20% test set[5,6,44], year thresholds were set to assign at least 50% of the total compound number to the proper training set, and at least 12% to each calibration set. The remaining compounds were used as holdout data (see Supplementary Material Section A1.4 for more details).

For the computational experiments with the liver toxicity and MNT data, the standardised public datasets were used for training. The standardised proprietary data were time-split into update and holdout set based on the internal measurement date (see Supplementary Material Section A1.4 for details). Due to the small number of available inhouse compounds, only one update set was deducted, containing at least 50% of the total available inhouse dataset, see Table 2.

	Training set			Update1 set			Update2 set			Holdout set		
Target (ID)	Thresh*	Inactive	Active	Thresh*	Inactive	Active	Thresh*	Inactive	Active	Thresh*	Inactive	Active
CHEMBL220	2014	802	840	2016	211	248	2017	217	138	2020	104	113
CHEMBL4078	2014	1031	1008	2015	259	275	2016	267	202	2020	499	270
CHEMBL5763	2015	1125	600	2016	302	75	2017	307	95	2020	137	114
CHEMBL203	2012	1660	433	2014	526	213	2016	428	291	2020	341	167
CHEMBL206	2006	437	325	2012	117	63	2016	114	97	2020	158	105
CHEMBL279	2010	1955	649	2013	523	307	2014	618	137	2020	686	299
CHEMBL230	2010	475	542	2013	218	78	2015	237	80	2020	218	172
CHEMBL340	2012	1272	496	2014	439	153	2015	341	59	2020	449	107
CHEMBL240	2012	797	1938	2014	301	413	2016	265	526	2020	238	498
CHEMBL2039	2014	710	645	2015	189	192	2017	380	212	2020	134	72
CHEMBL222	2009	231	673	2011	61	227	2015	40	206	2020	74	54
CHEMBL228	2009	242	858	2011	97	373	2014	31	235	2020	79	196
Micro nucleus test	-	1475	316	2005	70	134	–	–	–	2020	98	50
Liver toxicity	-	247	445	2011	42	48	–	–	–	2020	35	15

Table 2. Number of active and inactive compounds and year threshold used for the time split. ChEMBL data were temporally split into training, update1, update2 and holdout set based on the publication year. Models for the micro nucleus test and liver toxicity endpoint were trained on public data while the inhouse data were split into update and holdout set based on the internal measurement date. *Thresh: Data points published (ChEMBL) or measured (micro nucleus test, liver toxicity) until this year threshold are included in the corresponding subset.

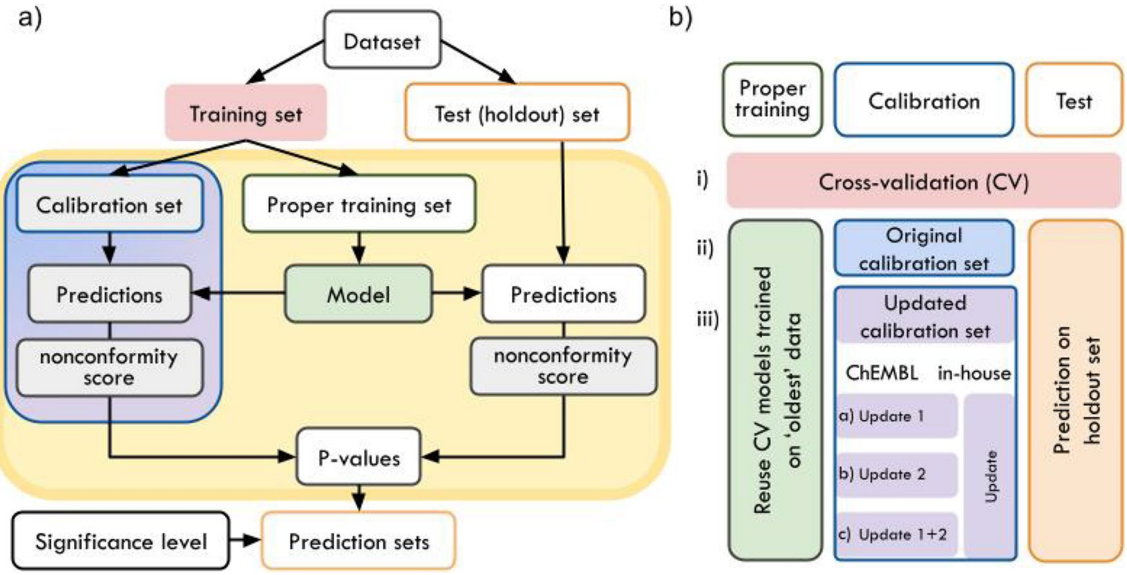


Figure 1. (**a**) Framework of an inductive conformal predictor. An ML model is fitted on the compounds of the proper training set to make predictions for the calibration and test (holdout) set instances. The predictions are transformed into nonconformity scores. By comparing the outcome of the test compound to the outcomes of the calibration set, p-values are calculated, which give an estimate on the likelihood of the compound to belong to a certain class. If a significance level is selected, prediction sets are calculated. *Blue-purple box* In the 'update calibration set' strategy, the calibration set is updated. *Yellow box* If multiple conformal predictors are aggregated, the part highlighted in the yellow box is repeated n times. (**b**) Overview of CP experiment setup: Experiments (i) CV, and prediction of holdout set using (ii) original calibration set, (iii) updated calibration sets to investigate temporal data drifts and drifts between data from different origin, i.e., ChEMBL and inhouse data.

Conformal prediction.

Inductive and aggregated conformal predictor. The framework of an inductive conformal predictor (ICP) (see Fig. 1a) uses three types of datasets: proper training set, calibration set, and test set[45]. On the proper training set, an underlying ML model is fitted to make predictions for the calibration and test set instances. The outcomes, i.e. the probabilities for a compound to be assigned to class 0 or 1 in binary classification, are converted into so-called nonconformity (nc scores) by using a nonconformity function.

Here, the inverse probability error function, which is typically used together with random forest (RF) models, is applied[20,46–48].

For each test data point, the calibrated model outputs two so-called p-values in the binary setup. Therefore, the nc scores of the calibration set are sorted into two lists, one per class. The ratio of nc scores of the calibration set, which are larger than the nc scores for a test sample, results in a p-value. If a significance level, i.e. an expected error rate, is selected, prediction sets can be derived. They contain the class labels for which the p-value is larger than the significance level. For binary classification, the possible prediction sets are {Ø}, {0}, {1}, {0,1}. Given that calibration and test data are exchangeable, the CP framework ensures that the observed error rate does not exceed the significance level[10,11].

In an ICP, only part of the information available in the training set is used for calibration as the other part is required to fit the underlying ML model. To improve the informational efficiency, multiple ICPs are typically aggregated in an aggregated conformal predictor (ACP)[47], as in this study. Therefore, the training and prediction part (see yellow box in Fig. 1) is repeated n times (here $n = 20$). In fact, the training set was 20 times split into calibration and proper training set, 20 models were built on the proper training set and calibrated with the corresponding calibration set. Each compound was predicted 20 times and the calculated p-values were aggregated taking the median value[49].

Evaluation of conformal predictors. Conformal predictors are generally evaluated with respect to their validity, efficiency and accuracy of single class predictions. Validity is defined as the ratio of prediction sets containing the correct class label. As predictions are considered correct when they contain the correct label, 'both' predictions ({0,1}) are always correct. Empty prediction sets ({Ø}) count as erroneous. Efficiency of the predictions can be assessed by the ratio of prediction sets containing a single class label, i.e. {0} and {1}. The ratio of these single class predictions containing the correct label is often calculated as the single class accuracy. In the case of unbalanced datasets, class-wise metrics, i.e. separate metrics for the compounds belonging to the active and inactive class, can also be calculated. Balanced metrics (e.g. balanced validity, balanced efficiency and balanced accuracy), are then calculated as the arithmetic mean of the class-wise metrics.

CP setup and experiments. In this work, it was further explored how effects of data drifts can be mitigated by recalibrating a CP model. In the 'update calibration set' strategy, the original calibration set (Fig. 1a, blue-purple box) is exchanged with data assumed to be closer to the holdout set (Fig. 1b). Three main experiments were performed and compared. First, an internal fivefold CV experiment was performed (Fig. 1b.i). Hence, the training set was five times randomly stratified split into 80% training and 20% test set. Within each CV fold, an ACP consisting of 20 ICPs (inverse probability error function, Mondrian condition, nonconformist Python library, version 2.1.0[46]) using an underlying RF classifier (500 estimators, else default parameters, scikit-learn Python library, version 0.22.2[50]) was implemented. Each model was trained on 70% (proper training set) and calibrated on 30% (original calibration set) of the selected training data. The test sets from the CV-splits were predicted with the CV-models calibrated with the original training set. Second, the same calibrated CV-models were used to predict the holdout set, i.e. the 'newest' data from the ChEMBL datasets or the inhouse DILI and MNT test sets (Fig. 1b.ii). Third, the same models were recalibrated using the update sets, which were determined as described in Section "Data assembly". For the experiments with the ChEMBL data, two update sets (update1 and update2) were used each, as well as a combination of update1+update2. For the inhouse data, only one update set was investigated. The recalibrated models were used to make predictions on the same holdout sets (Fig. 1b.iii) All models were evaluated at a significance level of 0.2, as it has been shown that this level offers a good trade-off between efficiency and validity[51,52].

Visualisation and further data analysis.

Visualisation. Data visualisations were created using matplotlib version 3.2.1[53].

UMAP. For descriptor space analysis, UMAPs were generated on the CHEMBIO fingerprints using the umap-learn Python library, version 0.4.6[54]. The parameters were set to $n\_neighbors = 100$, $min\_distances = 0.8$ and $distance\_metric = “euclidean''$, meaning that a range of 100 nearest neighbours was considered to learn the manifold data structure. The distance between two points plotted in the UMAP is at least 0.8 and the distance between two data points is calculated using the euclidean distance.

Compound clustering. To analyse commonalities between compounds per set, compounds were clustered, using the "Hierarchical Clustering" node in KNIME. The clusters were annotated based on the Tanimoto coefficients of Morgan fingerprints (1024 bits, radius 2) between all compound pairs. A distance threshold of 0.5 was chosen, i.e., clusters were split so that all compounds within a cluster have a smallest distance below the threshold. Since the analysis focused on detecting clusters that spread over more than one set (training/test/update/holdout), clusters with less than two compounds, i.e. singletons, were not considered. Clustering and fingerprint calculation was performed in KNIME.

Results and discussion

When using (ML) algorithms, it is assumed that the training data and test data are independent and identically distributed (*I.I.D.*). Similarly, CP models are designed to be valid if training and test data originate from the same distribution, i.e., are exchangeable[10]. This prerequisite, however, is not always fulfilled, especially when new compound spaces or different assay sources are explored. Hence, given comprehensive training data and modelling

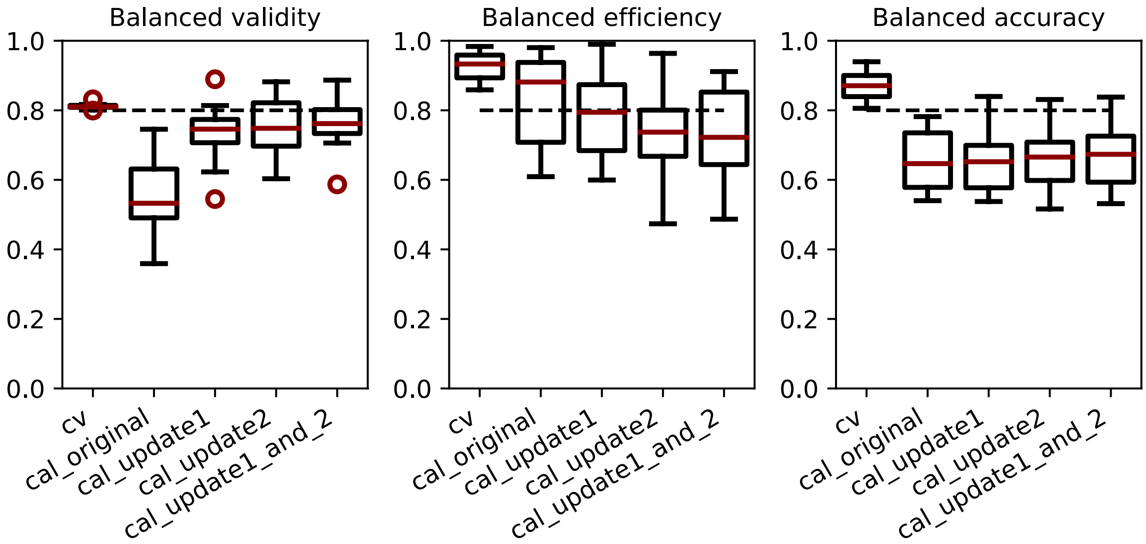


Figure 2. Time split evaluation (balanced validity, balanced efficiency, balanced accuracy) of CV experiments and predictions for the holdout set using the original (cal_original), update1 (cal_update1), update2 (cal_update2) and combined update1_and_2 (cal_update1_and_2) calibration sets for twelve ChEMBL datasets.

		Predict holdout set			
	CV	Cal_original	Cal_update1	Cal_update2	Cal_update1_and_2
Validity	0.81 ± 0.01	0.57 ± 0.14	0.75 ± 0.07	0.77 ± 0.09	0.78 ± 0.07
Efficiency	0.93 ± 0.04	0.82 ± 0.14	0.78 ± 0.12	0.74 ± 0.13	0.73 ± 0.15
Accuracy	0.87 ± 0.04	0.68 ± 0.10	0.68 ± 0.08	0.70 ± 0.10	0.70 ± 0.09
Balanced validity	0.81 ± 0.01	0.56 ± 0.11	0.73 ± 0.09	0.76 ± 0.08	0.77 ± 0.08
Balanced efficiency	0.93 ± 0.04	0.83 ± 0.14	0.79 ± 0.12	0.74 ± 0.13	0.73 ± 0.15
Balanced accuracy	0.87 ± 0.04	0.65 ± 0.09	0.65 ± 0.09	0.66 ± 0.10	0.67 ± 0.09
Validity inactive class	0.81 ± 0.01	0.62 ± 0.26	0.76 ± 0.22	0.78 ± 0.22	0.78 ± 0.20
Efficiency inactive class	0.93 ± 0.04	0.84 ± 0.14	0.79 ± 0.14	0.72 ± 0.14	0.73 ± 0.16
Accuracy inactive class	0.87 ± 0.05	0.72 ± 0.26	0.69 ± 0.26	0.68 ± 0.29	0.70 ± 0.24
Validity active class	0.81 ± 0.01	0.50 ± 0.22	0.71 ± 0.19	0.74 ± 0.18	0.75 ± 0.14
Efficiency active class	0.93 ± 0.05	0.81 ± 0.14	0.78 ± 0.13	0.75 ± 0.10	0.73 ± 0.16
Accuracy active class	0.87 ± 0.04	0.59 ± 0.20	0.61 ± 0.26	0.64 ± 0.23	0.64 ± 0.20

Table 3. Overall, balanced and class-wise evaluation of time-split experiments with ChEMBL data.

tasks, valid CP models can often be generated in a random-split k-fold CV setup. However, when predictions on external test data are performed, model performance has been shown to drop[55]. Here, we analysed the effects of data drifts on the validity of CP models. Thereby, we assessed the impact of recalibrating a CP model with updated data to restore the validity and positively affect performance. Note that this strategy has been introduced in the previous study, exemplified on the Tox21 challenge data[17], and is further investigated here for different datasets, molecular encodings and study settings.

In the first part of this study, temporal data drifts were analysed on twelve toxicity-related datasets from the ChEMBL database. In the second part, the applicability of models trained on public data to proprietary toxicity datasets was investigated.

Time-split experiments with twelve ChEMBL datasets.

To analyse the impact of temporal data drifts on CP model performance, ChEMBL datasets for twelve endpoints were prepared. The selected endpoints are toxicologically-relevant targets, known for off-target effects, drug-drug interactions or as ecotoxicological endpoints, which need to be considered during the development of new chemicals[33,34] (see Supplementary Table S1). The collected datasets were temporally split into training, update1, update2 and holdout subsets based on their publication date (see Section "Data and methods" and Table 2).

Experiments i and ii: CV and predictions using original calibration set. Fivefold CV on the training data produced valid (mean balanced validity: 0.81), efficient (mean balanced efficiency: 0.93), and accurate (mean balanced accuracy: 0.87) models at significance level of 0.2 (see experiment *cv_original* in Table 3 and Fig. 2). However, predictions with the same CV-models on the holdout data, i.e., newest data w.r.t. publication year, resulted in non-valid models with a higher-than-expected error rate (mean balanced validity of 0.56) as well as lower mean efficiency and accuracy (see experiment *cal_original* in Table 3 and Fig. 2). Class-wise evaluations for all experiments are provided in Supplementary Fig. S1.

The poor calibration of the model, i.e., a mean absolute loss in balanced validity of 0.25, for predictions on the holdout set may be an indicator for data drifts over time. Changes in the descriptor space or assay conditions (also due to diverse groups investigating the same target class) over the years may be responsible for such data drifts. Note that the data points in the holdout set were published at least five to ten years later than the training set instances (depending on the endpoint, see Table 2). Thus, it was investigated if the effects of these drifts can be mitigated by updating the calibration set with intermediately published data, i.e. update1 or update2 sets.

Experiment iii: update calibration set. To investigate whether valid models can be obtained with a small amount of new data, the calibration set was updated with more recent data while the trained CV-models were left unchanged[17]. For the ChEMBL experiments, the new calibration sets consist of the update1, update2 set, or a combination of both update sets.

Measured over all twelve endpoints, updating the calibration set with update1 or update2 led to an improvement of the mean balanced validity by up to 0.20 compared to the models with the original calibration set, reaching values of 0.73 and 0.76 with update1 and update2, respectively (see experiments *cal_update1* and *cal_update2* in Table 3 and Fig. 2). However, a slight decrease in the mean balanced efficiency by up to 0.09 was also observed (reaching values of 0.79 and 0.74 for update1 and update2, respectively).

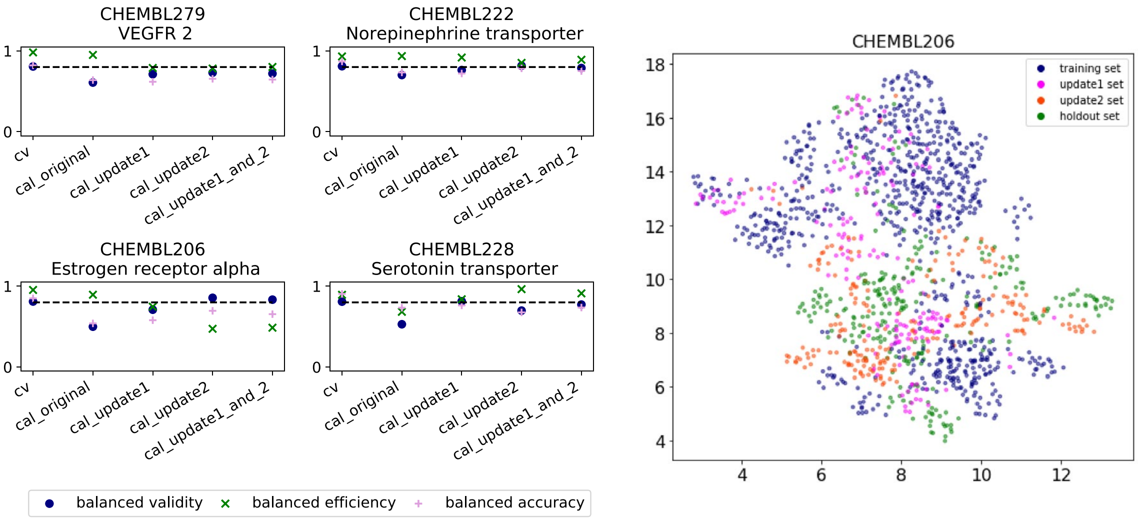
It should be noted that restoring the validity is a prerequisite for applying CP models with confidence[7,17]. In the absence of validity, the confidence of the predictions is not guaranteed and the efficiency becomes an irrelevant metric (CP model would not offer any advantage and could be exchanged by the base model (e.g. random forest) to obtain an efficiency of one). With validity being a prerequisite for the application of CP models, restoring it by recalibration is an improvement. The concurrent loss in efficiency is undesired but also expected, since many instances in the holdout set may fall outside the AD of the underlying model. Lower efficiency along with improved validity indicates that the model recognises more compounds, for which it does not have enough information to classify them into a single class. Hence they are predicted as 'both'. To avoid the loss in efficiency, the underlying model could be retrained with more up-to-date data. For example, compound representatives classified as empty or both sets by the current model could be experimentally screened to include their outcomes in an updated training set, feeding the model the necessary information to increase its efficiency. However, to achieve an improvement in the efficiency by retraining, a high amount of new data is usually required. Other studies[56–58] have explored the use of CP-based active learning approaches to select data points that provide the most information to the model if experimentally evaluated. By using these approaches, a small number of additional data points can greatly extend the AD of the model.

While no overall improvement—or impairment—was observed in terms of accuracy (see Table 3 and Fig. 2), restored validity allows predictions with an associated confidence.

To analyse the impact of the size of the calibration set on the model performance, the two update sets were combined and used as a new calibration set (update1 + 2). In summary, all evaluation values remained at a similar level as for the update1 and update2 experiments. Mean balanced validity of 0.77, mean balanced efficiency of 0.73 and mean balanced accuracy of 0.67 were achieved (see experiment *cal_update1_and_2*) in Table 3 and Fig. 2). This indicates that the variation in size of the different calibration sets (from around 500 compounds in the original, update1, and update2 calibration sets to around 1000 compounds in the update1 + 2 set) in the 'update calibration set' strategy does not have a major influence on model performance in this study. Previous studies have shown that the size of the calibration set, nevertheless, has an influence on the resolution of the p-values, i.e. if more data points are available for calibration, the calculation of the p-values becomes more precise/distinct[6,17]. For instance, a calibration set with only 4 active compounds can only produce five different p-values, while a larger calibration set will be more precise in the p-value assignment.

ChEMBL data composition analysis. It is concluded that the validity of predictions for the holdout set can be restored when using more recent data to calibrate the CP models.

This could be attributed to the fact that the distribution of calibration and holdout sets are more similar compared to the training data. The efficiency of the models is slightly affected by this strategy, as the model still lacks information to make single class predictions. Nevertheless, the characteristics of the time-split within the ChEMBL data based on the publication year should be considered with care. In theory, a cluster CV (where by design compounds belonging to the same cluster are always in the same splits) should present a more challenging task than a temporal CV (where series of compounds could be further developed after the splitting date)[26]. However, this situation could be different for time splits on public domain data. Yang et al.[19] showed on a benchmark study that time-split CV is a much harder task on public domain data (PDBbind[59–61] in this case) than in industry setups. Using ChEMBL data, we observe that one publication may contain a whole chemical series, which was developed over a longer period of time, but is labelled in ChEMBL with the same publication date. Moreover, the fact that public data in ChEMBL arise from different sources reduces the chances that a compound series is further developed over time (and is therefore present in several splits). This might increase the chemical diversity between time-splits within openly collected data compared to data from a single institution. Analysing the molecular clusters of the ChEMBL data used in this study and their distribution among time-splits, we observed that only few clusters are scattered over different splits. Only between 7% and 16% of the compounds in a single cluster (with distance threshold of 0.5 and only considering clusters with at least two compounds) were spread over more than one split (see Supplementary. Fig. S5). This result indicates that, in this case, the prediction of



(a)

(b)

Figure 3. Analysis of individual endpoints (**a**) Balanced evaluation of time-split experiments for four selected ChEMBL endpoints. Each plot represents CV results (cv) and predictions for the holdout set using the original (cal_original), update1 (cal_update1), update2 (cal_update2) and combined update1_and_2 (cal_update1_and_2) calibration sets. The doted line at 0.8 denotes the expected validity for the chosen significance level of 0.2. (**b**) UMAP showing the descriptor space covered by the compounds in the different time-split sets for ChEMBL206 endpoint.

the holdout set may be even more challenging than in an industrial (time-split) scenario, where early developed compounds of a compound series may be included in the consecutive training/update/holdout sets.

Individual endpoint performance analysis. The above discussed performance values referred to average values over models built for twelve endpoints. This led to the conclusion that updating the calibration set on average improves the validity at the cost of a small loss in efficiency. Considering the endpoints individually, the influence of updating the calibration set on the performance of the models varied. On average there was no substantial difference between updating the calibration set with update1 or update2 data. However, looking at individual models (Fig. 3a, Supplementary Fig. S4), e.g. endpoint ChEMBL228, the continuous calibration worked better in restoring the validity with update1 than update2 sets. In contrast, recalibrating with the update2 sets led to better performance for endpoints ChEMBL206, ChEMBL222, and ChEMBL279 (see also Supplementary Figs. S2 and S3).

The observations that the effects of recalibration for each endpoint are dependent on the update set might be explained by the descriptor space covered by the respective holdout, update and training sets. Our hypothesis is that updating the calibration set might be more beneficial if the update set compounds cover a descriptor space more similar to the holdout compounds than the original calibration set.

To investigate the influence of the descriptor space, the compounds' 'CHEMBIO' descriptors of the training, update1, update2, and holdout set were transformed into a two-dimensional space using UMAP (Fig. 3b). For endpoint ChEMBL206, for which the update2 strategy worked clearly better, a large part of the update1 set overlaps with the training set, indicating that less improvement can be expected when recalibrating with it. Contrary, there is more overlap between the holdout and update2 sets. This might explain the particularly positive effects of recalibrating with update2 on the validity and accuracy for predicting the ChEMBL206 holdout set.

To quantify these differences in a rational manner, the Tanimoto coefficient based on Morgan fingerprints of each holdout compound to its nearest neighbour in the training and update sets, respectively, was calculated. Exemplified for endpoint ChEMBL206, the median coefficient of the holdout compounds to their nearest neighbour in the respective sets confirmed that the the holdout set is on average more similar to the update2 set (median coefficient of 0.42) than to the update1 or training sets (median coefficients of 0.29 and 0.33, respectively; distribution of distances to nearest neighbours provided in Supplementary Fig. S6).

Update calibration strategy on inhouse datasets.

When insufficient internal data are available to build ML models (or, in general, to extent the descriptor space coverage of the models), public data can be used in industrial setups for model training. Exemplified by MNT in vivo and liver toxicity CP models, we explored whether the applicability and validity of predictions on internal data could be improved by recalibrating models trained on public data with part of the internal data.

CP models were fitted on publicly-available data for MNT in vivo and liver toxicity, previously collected and used for model building by Garcia de Lomana et al.[12]. Liver toxicity induced by chemicals is a growing cause of acute liver failure[62]. MNT in vivo is an assay to assess mutagenicity[29]. Both endpoints are highly relevant for registration and authorisation of new chemicals[28–30]. The internal data were temporally split into update (older

	Liver toxicity			Micro nucleus test		
		Predict holdout set			Predict holdout set	
	CV	Cal_original	Cal_update	CV	Cal_original	Cal_update
Balanced validity	0.81	0.47	0.82	0.82	0.50	0.74
Balanced efficiency	0.81	0.89	0.38	0.79	0.94	0.40
Balanced accuracy	0.77	0.43	0.49	0.77	0.49	0.39
Validity inactive class	0.81	0.75	0.84	0.80	0.99	0.61
Efficiency inactive class	0.84	0.84	0.45	0.79	0.89	0.54
Accuracy inactive class	0.77	0.70	0.63	0.75	0.99	0.29
Validity active class	0.82	0.20	0.80	0.83	0.00	0.88
Efficiency active class	0.78	0.95	0.31	0.79	1.00	0.26
Accuracy active class	0.77	0.16	0.35	0.78	0.00	0.50
Validity	0.82	0.58	0.84	0.81	0.66	0.70
Efficiency	0.80	0.87	0.40	0.79	0.93	0.45
Accuracy	0.77	0.52	0.57	0.76	0.63	0.33

Table 4. Evaluation of experiments to investigate drifts between internal and external data.

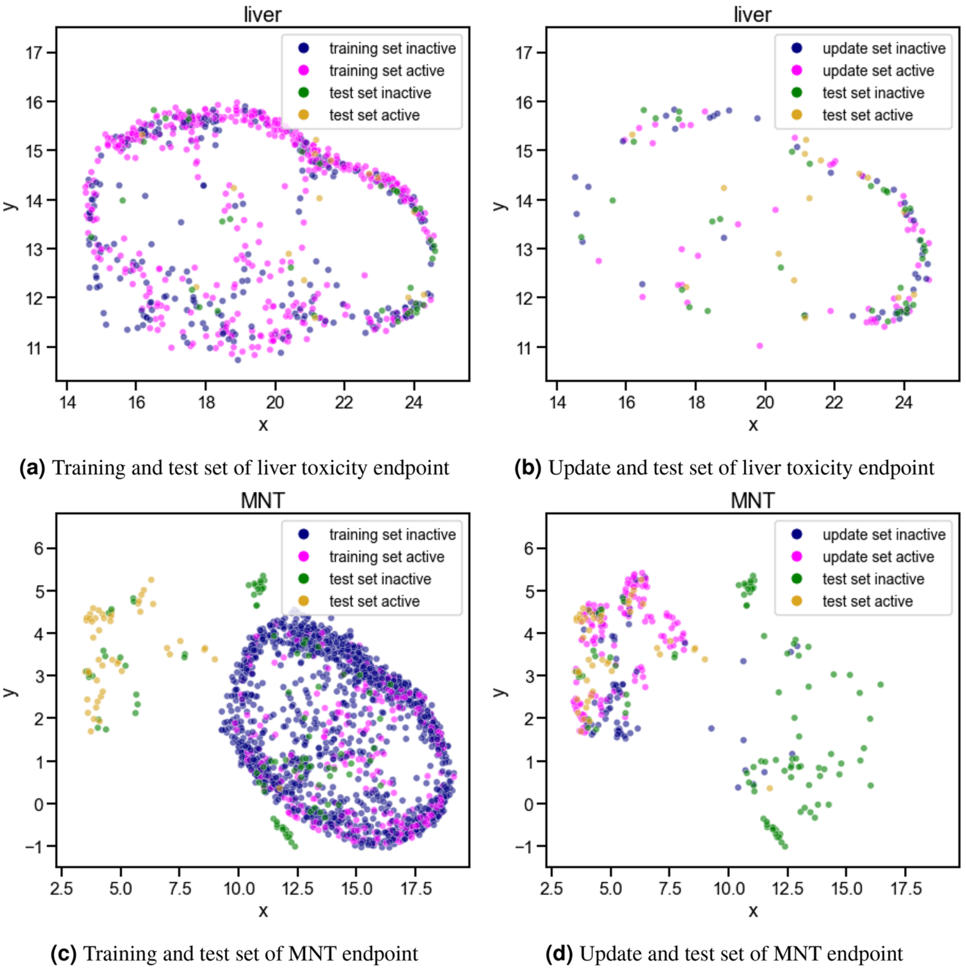
data) and holdout (more recent data) sets. Note that due to the limited data size only one update set was created (see Table 2).

Experiments i and ii: CV and predictions using original calibration set. The CP models were built on the publicly-available training data and validated within a fivefold CV. The predictions for the liver toxicity and the MNT endpoints resulted in a balanced validity of 0.81 and 0.82, a balanced efficiency of 0.81 and 0.79 and a balanced accuracy of 0.77 and 0.77, respectively (see Table 4). Thus, valid models with high efficiency and accuracy were obtained when evaluated within CV experiments.

Applying these models to the holdout set containing internal data, the balanced validity dropped drastically by up to 0.34 points (liver toxicity: 0.47, MNT: 0.50). The balanced accuracy of the models also decreased strongly (liver toxicity: 0.43, MNT: 0.49), while the balanced efficiency increased (liver toxicity: 0.89, MNT: 0.94). The latter indicates that mostly single class predictions were made. The class-wise evaluation of the MNT model predictions discloses that almost all internal compounds were predicted to be inactive (accuracy inactive compounds: 0.99, accuracy active compounds: 0, see Table 4 and Supplementary Fig. S7). For the liver endpoint, a similar trend was observed (accuracy inactive compounds: 0.7, accuracy active compounds: 0.16). These observations indicate that the distributions of the holdout and calibration data, i.e. of internal and external data, are highly different. Summarising, applying the models trained on public data to the internal data resulted in non-valid models that mainly predict all internal compounds as inactive.

Experiment iii: update calibration sets. For the liver toxicity endpoint, exchanging the calibration set with the earliest developed internal data (years 2005-2019, containing at least 50% of all internal data) could restore the validity for both compound classes (inactive: 0.84, active: 0.80). The balanced efficiency decreased largely from 0.89 to 0.38 (inactive compounds: 0.45, active compounds: 0.31) as many single class predictions were now identified as inconclusive and shifted to the 'both' class. The balanced accuracy increased only slightly from 0.43 to 0.49. Nevertheless, the accuracy became more balanced (inactive: 0.63, active compounds: 0.35), as now more active compounds were correctly identified as such. The observations for the liver toxicity endpoint are similar to those for the ChEMBL endpoints. It is promising that the validity could be restored, although the balanced efficiency dropped. The improved balanced accuracy of 0.49 still leaves room for further improvements. To visualise the differences in the descriptor space covered by the public and internal data, UMAPs were derived (see Fig. 4a,b). Both datasets seem to cover a similar area of the descriptor space calculated with UMAP. The low accuracy obtained by applying the model on internal data could thus be better explained by the differences in the endpoint definition, as public and internal data were derived from different assays and species. These differences could lead to inconsistencies in the class labelling of a compound (i.e. one compound having different outcomes in each assay). Although the validity of the models could be restored by recalibration, these inconsistencies could be one explanation for the poor performance in terms of accuracy.

For MNT, updating the calibration set led to an improved balanced validity from 0.50 to 0.74 (inactive compounds: 0.61, active compounds: 0.88) and a strongly reduced balanced efficiency from 0.94 to 0.40 (inactive compounds: 0.54, active compounds: 0.26). The fact that the validity for the active class is high while the efficiency of this class remains low, indicates a high number of both predictions for the active compounds. Thus, the model is lacking information about active compounds to make single class predictions. A reduction in the balanced accuracy to 0.39 was observed, while the values are again more balanced between classes (inactive compounds: 0.29, active compounds: 0.50). Concluding, in the case of MNT, the balanced validity could be improved when recalibrating the models, but for the inactive compounds, it could not be restored to the expected level of 0.8. Analysing the descriptor space of the different datasets and their class labels (see UMAPs in Fig. 4c,d), it can be observed that almost all holdout compounds overlapping with the training set are inactive, while most of the holdout compounds overlapping with the update set are active. After updating the calibration set, the validity



(a) Training and test set of liver toxicity endpoint

(b) Update and test set of liver toxicity endpoint

(c) Training and test set of MNT endpoint

(d) Update and test set of MNT endpoint

Figure 4. Descriptor space analysis of the liver toxicity (**a**, **b**) and MNT datasets (**c**, **d**) derived by UMAP. The descriptor space covered by the active and inactive compounds of the test sets is compared to the space covered by the training (**a**, **c**) and update sets (**b**, **d**), respectively.

of the active class increased and could be restored, as this class is now better represented in the calibration set. However, the contrary is observed for the inactive class. Moreover, the efficiency drops as the analysed compounds are very different from the training set and the models are missing information about this area of the descriptor space to make single class predictions.

Although exchanging the calibration set with data from the same origin as the holdout set, i.e. with inhouse data, did help to increase the validity, these results show that the descriptor space of the holdout set still needs to be better represented by the training set to obtain efficient and accurate—and therefore useful—models.

Conclusion

CP models, or generally ML models, are widely used for molecular property predictions, including activity and toxicity[5,6,63]. Notably, the CP framework is based on the assumption that test and calibration data stem from the same distribution[10,11]. If this prerequisite is not given, the models are not guaranteed to be valid (i.e. return the expected error rate). The goal of this study was twofold. Firstly, the performance of internally valid CP models, when applied to either newer time-split or (true) external data, was assessed. Second, the impact of model updating strategies exchanging the CP calibration set with data closer to the prediction set was evalutated. Building on previous work performed on the Tox21 datasets[17], we investigated here two scenarios with data subsets that may stem from different distributions. First, temporal data drifts were analysed at the example of twelve toxicity-related datasets collected from the ChEMBL bioactivity database. Second, discrepancies between performance of models trained on publicly-available data vs. models recalibrated on inhouse data was evaluated on holdout inhouse data for the liver toxicity and MNT in vivo endpoints.

Due to changes in descriptor space and assays, over time or between laboratories, data drifts occur and were observed through the performed experiments (i and ii) on both the twelve ChEMBL as well as the liver toxicity and MNT datasets. Overall, valid CP models within CV were built for all endpoint datasets at a significance level of 0.2. In contrast, validity dropped below the expected error rate of 0.8, when applied to the holdout sets. Resulting mean balanced validities were 0.56 ± 0.11 over all twelve ChEMBL datasets, 0.47 for liver toxicity and 0.50 for MNT.

To address the poor validity on the holdout set, CP updating strategies were implemented (experiment iii), in which the calibration sets were exchanged by part of the newer or proprietary data, with the aim of restoring the validity. For most of the ChEMBL endpoints, the validity (at 0.2 significance level) could be mostly restored (mean balanced validity: 0.77 ± 0.08). The same holds for predictions on the proprietary liver toxicity endpoint data (balanced validity: 0.82). For the MNT data, the calibration was also improved, but to a lower extent (balanced validity: 0.74). Note that the improved validity comes at the cost of reduced efficiency for ten of the ChEMBL endpoints (average absolute loss between 0.04 and 0.10, depending on the update set used), which is more prominent for the liver toxicity and the MNT endpoints (absolute loss up to 0.55). A drop in efficiency is, however, more acceptable than non-valid models, which cannot be confidently applied. Too low efficiency may indicate that the model lacks information, e.g., chemical and biological descriptor space coverage, for classifying the new compounds.

With regard to the accuracy of the single class predictions, no change was observed on average for the ChEMBL endpoints when updating the calibration set. However, for the liver toxicity and MNT endpoints a more balanced accuracy between classes was observed after the update, as more compounds were identified as active.

In principle it is not possible to define an overall update/calibration criteria for all applications, but more research is needed to derive a generic approach on how to define it within the specific use-cases. In future studies it should be investigated how the degree of deviation of the calibration set from the training and holdout sets influences the models validity, efficiency and accuracy. This trade-off between the similarity of the calibration data to each set and the amount of available update data will probably determine in which scenarios the recalibration strategy is a good approach to overcome data drifts, and when a complete model retraining is necessary.

It is in the nature of the field of compound toxicity prediction or drug design that ML models are applied to completely new compounds that are potentially quite different from the training set. This work showed the necessity of considering data drifts when applying CP or ML models to new and external data and the need of developing strategies to mitigate the impact on the performance.

Data availability

The input data for the twelve ChEMBL endpoint models can be retrieved from https://doi.org/10.5281/zenodo.5167636. The public data for the liver toxicity and in vivo MNT endpoints are freely available as described in Garcia de Lomana et al.[12]. The in house data for liver toxicity and in vivo MNT are proprietary to BASF SE.

Code availability

Code is available on GitHub at https://github.com/volkamerlab/CPrecalibration_manuscript_SI. The GitHub repository contains example notebooks on how to perform the recalibration experiments on a selected endpoint as well as on all twelve ChEMBL endpoints together. The code can be adapted and used for other datasets.

Received: 28 September 2021; Accepted: 17 March 2022
Published online: 04 May 2022

References

1. Zhang, L. *et al.* Applications of machine learning methods in drug toxicity prediction. *Curr. Top. Med. Chem.* **18**, 987–997. https://doi.org/10.2174/1568026618666180727152557 (2018).
2. Huang, R. *et al.* Tox21 challenge to build predictive models of nuclear receptor and stress response pathways as mediated by exposure to environmental toxicants and drugs. *Front. Environ. Sci.* **3**, 85. https://doi.org/10.3389/978-2-88945-197-5 (2016).
3. Mansouri, K. *et al.* CoMPARA: Collaborative modeling project for androgen receptor activity. *Environ. Health Perspect.* **128**, 027002. https://doi.org/10.1289/EHP5580 (2020).
4. Idakwo, G. *et al.* A review on machine learning methods for in silico toxicity prediction. *J. Environ. Sci. Health C* **36**, 169–191. https://doi.org/10.1080/10590501.2018.1537118 (2018).
5. Morger, A. *et al.* KnowTox: Pipeline and case study for confident prediction of potential toxic effects of compounds in early phases of development. *J. Cheminform.* **12**, 1–17. https://doi.org/10.1186/s13321-020-00422-x (2020).
6. Svensson, F., Norinder, U. & Bender, A. Modelling compound cytotoxicity using conformal prediction and PubChem HTS data. *Toxicol. Res.* **6**, 73–80. https://doi.org/10.1039/C6TX00252H (2017).
7. Hanser, T., Barber, C., Guesne, S., Marchaland, J. F. & Werner, S. Applicability domain: Towards a more formal framework to express the applicability of a model and the confidence in individual predictions. In *Advances in Computational Toxicology* (ed. Hong, H.) 215–232 (Springer, Cham, 2019).
8. Mathea, M., Klingspohn, W. & Baumann, K. Chemoinformatic classification methods and their applicability domain. *Mol. Inform.* **35**, 160–180. https://doi.org/10.1002/minf.201501019 (2016).
9. OECD. *Guidance Document on the Validation of (Quantitative) Structure-Activity Relationship [(Q)SAR] Models* (OECD Publishing, 2014).
10. Vovk, V., Gammerman, A. & Shafer, G. *Algorithmic Learning in a Random World* (Springer, 2005).
11. Alvarsson, J., Arvidsson McShane, S., Norinder, U. & Spjuth, O. Predicting with confidence using conformal prediction in drug discovery. *J. Pharm. Sci.* **110**, 42–49. https://doi.org/10.1016/j.xphs.2020.09.055 (2021).
12. Garcia de Lomana, M. *et al.* ChemBioSim: Enhancing conformal prediction of in vivo toxicity by use of predicted bioactivities. *J. Chem. Inf. Model.* https://doi.org/10.1021/acs.jcim.1c00451 *(2021).*
13. Chen, Y., Stork, C., Hirte, S. & Kirchmair, J. NP-scout: Machine learning approach for the quantification and visualization of the natural product-likeness of small molecules. *Biomolecules.* https://doi.org/10.3390/biom9020043 *(2019).*
14. Stepanov, D., Canipa, S. & Wolber, G. HuskinDB, a database for skin permeation of xenobiotics. *Sci. Data* **7**, 1–8. https://doi.org/10.1038/s41597-020-00764-z (2020).
15. Fourches, D., Muratov, E. & Tropsha, A. Trust but verify: On the importance of chemical structure curation in chemoinformatics and QSAR modeling research. *J. Chem. Inf. Model.* **50**, 1189–1204 (2010).
16. Arvidsson McShane, S., Ahlberg, E., Noeske, T. & Spjuth, O. Machine learning strategies when transitioning between biological assays. *J. Chem. Inf. Model.* https://doi.org/10.1021/acs.jcim.1c00293 *(2021).*
17. Morger, A. *et al.* Assessing the calibration in toxicological in vitro models with conformal prediction. *J. Cheminform.* **1**, 1–14. https://doi.org/10.1186/s13321-021-00511-5 (2021).

18. Kosugi, Y. & Hosea, N. Prediction of oral pharmacokinetics using a combination of in silico descriptors and in vitro ADME properties. *Mol. Pharm.* https://doi.org/10.1021/acs.molpharmaceut.0c01009 *(2021).*
19. Yang, K. *et al.* Analyzing learned molecular representations for property prediction. *J. Chem. Inf. Model.* **59**, 3370–3388. https://doi.org/10.1021/acs.jcim.9b00237 (2019).
20. Norinder, U., Spjuth, O. & Svensson, F. Using predicted bioactivity profiles to improve predictive modeling. *J. Chem. Inf. Model.* **60**, 2830–2837. https://doi.org/10.1021/acs.jcim.0c00250 (2020).
21. Mendez, D. *et al.* ChEMBL: Towards direct deposition of bioassay data. *Nucleic Acids Res.* **47**, D930–D940. https://doi.org/10.1093/nar/gky1075 (2019).
22. Davies, M. *et al.* ChEMBL web services: Streamlining access to drug discovery data and utilities. *Nucleic Acids Res.* **43**, W612–W620. https://doi.org/10.1093/nar/gkv352 (2015).
23. Cortés-Ciriano, I., Škuta, C., Bender, A. & Svozil, D. QSAR-derived affinity fingerprints (part 2): Modeling performance for potency prediction. *J. Cheminform.* **12**, 1–17. https://doi.org/10.1186/s13321-020-00444-5 (2020).
24. Bosc, N. *et al.* Large scale comparison of QSAR and conformal prediction methods and their applications in drug discovery. *J. Cheminform.* **11**, 1–16. https://doi.org/10.1186/s13321-018-0325-4 (2019).
25. Sakai, M. *et al.* Prediction of pharmacological activities from chemical structures with graph convolutional neural networks. *Sci. Rep.* **11**, 1–14. https://doi.org/10.1038/s41598-020-80113-7 (2021).
26. Mayr, A. *et al.* Large-scale comparison of machine learning methods for drug target prediction on ChEMBL. *Chem. Sci.* https://doi.org/10.1039/C8SC00148K *(2018).*
27. Mathai, N. & Kirchmair, J. Similarity-based methods and machine learning approaches for target prediction in early drug discovery: Performance and scope. *Int. J. Mol. Sci.* https://doi.org/10.3390/ijms21103585 *(2020).*
28. Watkins, P. B. Drug safety sciences and the bottleneck in drug development. *Clin. Pharmacol. Ther.* **89**, 788–790. https://doi.org/10.1038/clpt.2011.63 (2011).
29. OECD. *Test No. 474: Mammalian Erythrocyte Micronucleus Test* (OECD Publishing, 2016).
30. ICHS2(R1). Guidance on Genotoxicity Testing and Data Interpretation for Pharmaceuticals Intended for Human Use. *International Conference on Harmonization of Technical Requirements for Registration of Pharmaceuticals for Human Use* (2011).
31. Škuta, C. *et al.* QSAR-derived affinity fingerprints (part 1): Fingerprint construction and modeling performance for similarity searching, bioactivity classification and scaffold hopping. *J. Cheminform.* **12**, 1–16. https://doi.org/10.1186/s13321-020-00443-6 (2020).
32. IDG. *Illuminating the Druggable Genome: Target Development Levels* (2022).
33. Richard, A. M. *et al.* ToxCast chemical landscape: Paving the road to 21st century toxicology. *Chem. Res. Toxicol.* **29**, 1225–1251. https://doi.org/10.1021/acs.chemrestox.6b00135 (2016).
34. Bowes, J. *et al.* Reducing safety-related drug attrition: The use of in vitro pharmacological profiling. *Chem. Res. Toxicol.* https://doi.org/10.1038/nrd3845 (2012).
35. OECD. *Test No. 407: Repeated Dose 28-day Oral Toxicity Study in Rodents* (OECD Publishing, 2008).
36. OECD. *Test No. 408: Repeated Dose 90-Day Oral Toxicity Study in Rodents* (OECD Publishing, 2018).
37. OECD. *Test No. 422: Combined Repeated Dose Toxicity Study with the Reproduction/Developmental Toxicity Screening Test* (OECD Publishing, 1996).
38. ChemAxon.
39. Berthold, M. R. *et al.* KNIME: The Konstanz information miner. *ACM SIGKDD Explor. Newsl.* **11**, 26. https://doi.org/10.1145/1656274.1656280 (2009).
40. Fillbrunn, A. *et al.* KNIME for reproducible cross-domain analysis of life science data. *J. Biotechnol.* **261**, 149–156. https://doi.org/10.1016/j.jbiotec.2017.07.028 (2017).
41. Rogers, D. & Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **50**, 742–754. https://doi.org/10.1021/ci100050t (2010).
42. Landrum, G. A. RDKit: Open-source cheminformatics. http://www.rdkit.org (2018).
43. Ji, C., Svensson, F., Zoufir, A. & Bender, A. eMolTox: Prediction of molecular toxicity with confidence. *Bioinformatics.* https://doi.org/10.1093/bioinformatics/bty135 *(2018).*
44. Norinder, U., Carlsson, L., Boyer, S. & Eklund, M. Introducing conformal prediction in predictive modeling: A transparent and flexible alternative to applicability domain determination. *J. Chem. Inf. Model.* https://doi.org/10.1021/ci5001168 *(2014).*
45. Vovk, V. Conditional validity of inductive conformal predictors. *Mach. Learn.* **92**, 349–376. https://doi.org/10.1007/s10994-013-5355-6 (2013).
46. Linusson, H. Nonconformist. http://donlnz.github.io/nonconformist/ (2015).
47. Carlsson, L., Eklund, M. & Norinder, U. Aggregated conformal prediction. *IFIP Adv. Inf. Commun. Technol.* **1**, 231–240 (2014).
48. Shen, Y. *Loss functions for binary classification and class probability estimation*. Ph.D. thesis, University of Pennsylvania (2005).
49. Linusson, H., Norinder, U., Boström, H., Johansson, U. & Löfström, T. On the Calibration of Aggregated Conformal Predictors. *Proceedings of the Sixth Workshop on Conformal and Probabilistic Prediction and Applications* **60**, 154–173 (2017).
50. Pedregosa, F. *et al.* Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
51. Cortés-Ciriano, I. & Bender, A. *Concepts and Applications of Conformal Prediction in Computational Drug Discovery. ArXiv* 1–40 (2019).
52. Svensson, F. *et al.* Conformal regression for QSAR modelling: Quantifying prediction uncertainty. *J. Chem. Inf. Model.* **58**, 1132–1140. https://doi.org/10.1021/acs.jcim.8b00054 (2018).
53. Hunter, J. D. Matplotlib: A 2D graphics environment. *Comput. Sci. Eng.* **9**, 90–95. https://doi.org/10.1109/MCSE.2007.55 (2007).
54. McInnes, L., Healy, J. & Melville, J. *Uniform Manifold Approximation and Projection for Dimension Reduction* (UMAP, 2018).
55. Vovk, V. Cross-conformal predictors. *Ann. Math. Artif. Intell.* **74**, 9–28. https://doi.org/10.1007/s10472-013-9368-4 (2015).
56. Makili, L. E., VegaSanchez, J. A. & Dormido-Canto, S. Active learning using conformal predictors: Application to image classification. *Fusion Sci. Technol.* **62**, 347–355 (2012).
57. Corrigan, A. M. *et al.* Batch mode active learning for mitotic phenotypes using conformal prediction. *Proc. Mach. Learn. Res.* **128**, 1–15 (2020).
58. Svensson, F., Norinder, U. & Bender, A. Improving screening efficiency through iterative screening using docking and conformal prediction. *J. Chem. Inf. Model.* **57**, 439–444. https://doi.org/10.1021/acs.jcim.6b00532 (2017).
59. Wang, R., Fang, X., Lu, Y. & Wang, S. The PDBbind database: Collection of binding affinities for protein-ligand complexes with known three-dimensional structures. *J. Med. Chem.* **47**, 2977–2980. https://doi.org/10.1021/jm030580l (2004).
60. Wang, R., Fang, X., Lu, Y., Yang, C. Y. & Wang, S. The PDBbind database: Methodologies and updates. *J. Med. Chem.* **48**, 4111–4119. https://doi.org/10.1021/jm048957q (2005).
61. Wu, Z. *et al.* MoleculeNet: A benchmark for molecular machine learning. *Chem. Sci.* **9**, 513–530. https://doi.org/10.1039/c7sc02664a (2018).
62. Norman, B. H. Drug induced liver injury (DILI). Mechanisms and medicinal chemistry avoidance/mitigation strategies. *J. Med. Chem.* **63**, 11397–11419. https://doi.org/10.1021/acs.jmedchem.0c00524 (2020).
63. Wang, Y. *et al.* Discrimination of different species of dendrobium with an electronic nose using aggregated conformal predictor. *Sensors.* https://doi.org/10.3390/s19040964 *(2019).*

Acknowledgements

A.V. thanks BMBF (Grant 031A262C) for funding. A.M. acknowledges support from the HaVo-Stiftung. Furthermore, the authors have received internal BASF SE funding. The authors thank Roland Buesen and Robert Landsiedel, BASF SE, for providing the inhouse data to study data drifts between freely available and inhouse data. Additionally, A.M. and A.V. would like to thank the HPC Service of ZEDAT, Freie Universität Berlin, for computing time.

Author contributions

A.M., M.M. and A.V. conceived the study, A.M. and M.G.L. conducted the computational experiments, A.M., M.G.L., J.K., U.N., M.M. and A.V. analysed the results. J.K., M.M. and A.V. supervised the study, consulted by U.N. and F.S.. A.M., M.G.L., M.M. and A.V. wrote the manuscript draft. All authors reviewed the manuscript. All authors agreed to the submitted version of the manuscript.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Competing interests

M.G.L. and M.M. are employed at BASF SE. U.N. performed research and served as a consultant for BASF SE. Other authors do not have conflict of interest.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-022-09309-3.

Correspondence and requests for materials should be addressed to M.M. or A.V.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

4.4 Consideration of xenobiotic metabolism information in toxicity prediction models

One of the most relevant parameters influencing the discrepancies between in vitro and in vivo assay results is xenobiotic metabolism. In vivo, the parent compound may undergo several biotransformations resulting in a variety of metabolite structures. Although metabolism is intended to detoxify xenobiotics and facilitate their excretion, compounds may also be activated into more reactive or toxic metabolites. The relevance of metabolism for the toxic outcome has already been described in multiple studies[83-85] and the formation of toxic metabolites has forced the withdrawal of several drugs from the market (e.g. the antidepressant nefazodone).[86] Nevertheless, studies considering metabolism information in toxicity prediction models are very scarce and often only limited to data sets with measured metabolites[87, 88] or focused on a specific endpoint[89-92] from which no general conclusions can be drawn.

The following publication analyses different approaches for including metabolism information in toxicity prediction models based on five endpoints: two genotoxicity assays (AMES in vitro with metabolic activation and MNT in vivo), two organ toxicity endpoints (DILI and DICC) and a skin sensitization assay (LLNA). The developed approaches could be divided in two main workflows, (i) the generation of new input features for a parent compound based on chemical properties of its metabolites or on the metabolic transformations the compound may undergo, and (ii) the combination of the predictions made for a parent compound and its metabolites. In order to expand the coverage of the models and make them directly applicable to new, untested compounds, predicted metabolites instead of measured ones were employed in this study. Moreover, filters to discard unlikely or highly soluble metabolites were also implemented to further clean the predicted metabolism data and improve the toxicity predictions.

[P4] Consideration of Predicted Small-Molecule Metabolites in Computational Toxicology

Marina Garcia de Lomana, Fredrik Svensson, Andrea Volkamer, Miriam Mathea and Johannes Kirchmair

Digital Discovery, 2022

Available at https://doi.org/10.1039/D1DD00018G

Contribution:

M. Garcia de Lomana, M. Mathea and J. Kirchmair conceptualized the research. M. Garcia de Lomana along with F. Svensson, A. Volkamer, M. Mathea and J. Kirchmair designed the experiments. M. Garcia de Lomana compiled the data sets and developed the machine learning models. M. Garcia de Lomana wrote the manuscript, with contributions from F. Svenson, A. Volkamer, M. Mathea and J. Kirchmair. M. Mathea and J. Kirchmair supervised the work.

The following article was reprinted with permission from:

Garcia de Lomana, M.; Svensson, F.; Volkamer, A.; Mathea, M. and Kirchmair, J. Consideration of Predicted Small-Molecule Metabolites in Computational Toxicology, *Digital Discov.*, **2022**, 1, 158-172.

Digital Discovery

PAPER

View Article Online
View Journal | View Issue

Check for updates

Cite this: *Digital Discovery*, 2022, **1**, 158

Consideration of predicted small-molecule metabolites in computational toxicology†

Marina Garcia de Lomana,[ab] Fredrik Svensson,[c] Andrea Volkamer,[d] Miriam Mathea *[a] and Johannes Kirchmair *[b]

Xenobiotic metabolism has evolved as a key protective system of organisms against potentially harmful chemicals or compounds typically not present in a particular organism. The system's primary purpose is to chemically transform xenobiotics into metabolites that can be excreted *via* renal or biliary routes. However, in a minority of cases, the metabolites formed are toxic, sometimes even more toxic than the parent compound. Therefore, the consideration of xenobiotic metabolism clearly is of importance to the understanding of the toxicity of a compound. Nevertheless, most of the existing computational approaches for toxicity prediction do not explicitly take metabolism into account and it is currently not known to what extent the consideration of (predicted) metabolites could lead to an improvement of toxicity prediction. In order to study how predictive metabolism could help to enhance toxicity prediction, we explored a number of different strategies to integrate predictions from a state-of-the-art metabolite structure predictor and from modern machine learning approaches for toxicity prediction. We tested the integrated models on five toxicological endpoints and assays, including *in vitro* and *in vivo* genotoxicity assays (AMES and MNT), two organ toxicity endpoints (DILI and DICC) and a skin sensitization assay (LLNA). Overall, the improvements in model performance achieved by including metabolism data were minor (up to +0.04 in the F1 scores and up to +0.06 in MCCs). In general, the best performance was obtained by averaging the probability of toxicity predicted for the parent compound and the maximum probability of toxicity predicted for any metabolite. Moreover, including metabolite structures as further input molecules for model training slightly improved the toxicity predictions obtained by this averaging approach. However, the high complexity of the metabolic system and associated uncertainty about the likely metabolites apparently limits the benefit of considering predicted metabolites in toxicity prediction.

Received 5th October 2021
Accepted 10th February 2022

DOI: 10.1039/d1dd00018g

rsc.li/digitaldiscovery

Introduction

The metabolic system has evolved as the primary defense system against xenobiotic, potentially toxic substances. Its protective function is based on the biotransformation of xenobiotics into more hydrophilic and, hence, more rapidly excretable compounds (metabolites). However, a minority of metabolites produced by the metabolic system are more active than their parent compound (which is exploited by the prodrug concept) or even toxic.[1]

The important role of metabolism in the toxicity of small organic molecules highlights the need for the consideration of metabolic pathways also in the computational prediction of toxicity. However, so far only a few *in silico* models for toxicity prediction have integrated metabolism information. For example, Dmitriev *et al.*[2] built linear models for the prediction of rat acute toxicity using self-consistent regression, thereby considering parent compounds and measured metabolites. More specifically, they trained a model on about 3000 parent compounds and used it to predict the LD_{50} value of 37 test parent compounds and their measured metabolites (around 200 known metabolites). To calculate the final LD_{50} value, different strategies for averaging the LD_{50} values predicted for the parent compounds and their metabolites were investigated. However, only minor improvements in the overall performance of the model were achieved compared to using only the predicted probability of the parent compounds (R^2 increased from 0.78 to 0.81 and RMSE remained at 0.49). In a more recent study

[a] *BASF SE, 67063 Ludwigshafen am Rhein, Germany. E-mail: miriam.mathea@basf.com; Tel: +49-621-60-29054*

[b] *Department of Pharmaceutical Sciences, Faculty of Life Sciences, University of Vienna, 1090 Vienna, Austria. E-mail: johannes.kirchmair@univie.ac.at; Tel: +43-1-4277-55104*

[c] *Alzheimer's Research UK UCL Drug Discovery Institute, University College London, London WC1E 6BT, UK*

[d] *In Silico Toxicology and Structural Bioinformatics, Institute of Physiology, Charité Universitätsmedizin Berlin, 10117 Berlin, Germany*

† Electronic supplementary information (ESI) available. See DOI: 10.1039/d1dd00018g

from the same research group,[3] classification models based on a Bayesian approach were trained on parent compounds with annotated bioactivity data for a variety of endpoints. The bioactivity of a compound was then calculated as the maximum probability predicted among the parent compound and its measured metabolites. For the 28 endpoints in the "toxic and adverse effects" category (with data sets ranging from 15 to 5583 toxic and non-toxic compounds), an increase of up to 0.14 in the precision and 0.16 in the recall during leave-one-out cross-validation (CV) was obtained on average (compared to taking the predicted probability of the parent compound only). These results show that the consideration of metabolism in prediction models can substantially improve the identification of potentially toxic compounds.

Data on measured metabolites can be valuable for estimating the toxicity of compounds but such approaches rely on the availability of experimental data. For this reason, *in silico* approaches to predict the likely metabolites of substances based on molecular structures are in high demand. Several predictors of this kind are available today, including BioTransformer,[4] CyProduct,[5] GloryX,[6] Meteor Nexus,[7] SyGMA,[8] TIssue MEtabolic Simulator (TIMES)[9] and XenoSite.[10]

In previous works, researchers from the Laboratory of Mathematical Chemistry (LMC) have combined *in silico* models for toxicity prediction with their TIMES metabolite predictor. The first model from LMC taking into account the parent compound and its metabolites (predicted with the S-9 metabolism simulator of TIMES) was developed for the prediction of *in vitro* mutagenicity (*i.e.* outcomes of the AMES assay).[11,12] This AMES model was based on decision trees trained on the reactivity profile of compounds and labeled a compound as toxic if any of its predicted metabolites were predicted as toxic. The evaluation of the model on the training data showed that the metabolism-aware approach resulted in lower sensitivity (0.77) and specificity (0.74) compared to the performance of the model considering only the parent compound (sensitivity 0.82; specificity 0.94). The lower sensitivity obtained by this approach may be related to the fact that compounds without any predicted metabolites were automatically classified as inactive. Another drawback of this approach is the decrease in specificity due to false positive predictions derived from non-mutagenic parents with metabolites predicted as mutagenic. In addition to the training data, the model was evaluated on a test set of 36 mutagenic compounds, obtaining a sensitivity of 0.58 (corresponding to 21 correctly classified compounds). Despite the overall drop in performance, the metabolism-aware approach correctly identified compounds of which their mutagenicity is related to the metabolites formed.

Two further decision tree models from LMC targeting skin and respiratory sensitization, respectively,[13,14] also included the evaluation of several properties of predicted metabolites (*e.g.* reactivity profile or ability to cross-link proteins) to classify the parent compounds as non-sensitizers or sensitizers (further distinguished between strong or weak sensitizers in the case of the skin). The evaluation of this skin sensitization model on the training data yielded 80% correct predictions for strong sensitizers, 34% for weak sensitizers and 72% for non-sensitizers, while the respiratory sensitization model obtained a sensitivity of 0.89 and a specificity of 0.52.

A further model of this kind from LMC was reported for the *in vivo* micronucleus test (MNT).[15] By comparing the assay outcomes of the (*in vitro*) AMES assay with a liver genotoxicity and an MNT *in vivo* assays, bioactivated compounds and "bio-exhausted" compounds (*i.e.* highly reactive compounds interacting with off-targets before reaching the target) were analyzed to establish *in vitro*–*in vivo* relationships. Based on this analysis, an *in vivo* rat liver metabolism predictor reproducing phase II conjugation reactions and detoxification pathways was developed. The toxicity prediction model of MNT applied on the predicted metabolites (derived with the *in vivo* metabolite predictor) reached a sensitivity of 0.82 and a specificity of 0.61 on the training data.

The performance of this MNT model, as well as the skin and respiratory sensitization models, was not compared to the performance of models not considering predicted metabolites. Therefore it is not possible to conclude on the benefits or drawbacks of these metabolism-aware models compared to models considering only parent compounds.

Overall, these recent reports on efforts to enhance toxicity prediction of small organic molecules by the consideration of their biotransformation provide valuable insights and starting points for the further development of methods for computational toxicology. Although metabolism is key to understanding the pharmacokinetics and toxicity of compounds, the inherent uncertainty of the complex metabolic data could also hinder the improvement of models integrating this information. So far, the existing works on this topic are either based on only a few parent compounds and their measured metabolites, or focused on a single endpoint, making it therefore difficult to derive more general conclusions.

With this work, we aim to provide a systematic study on how, and to what extent, the consideration of metabolism can help the *in silico* prediction of toxicity. In order to cover a wide chemical space and make models applicable to new, untested compounds, we referred to the use of predicted metabolites. Five relevant toxicological endpoints and assays were selected for investigation: the *in vitro* AMES assay (considering metabolic activation with S-9 liver extract), the *in vivo* micronucleus test (MNT), a skin sensitization assay (the murine local lymph node assay, LLNA), and the drug-induced liver injury (DILI) and cardiological complications (DICC) endpoints.[16–18] All selected endpoints and assays have in common that their outcome is known to be related, to some extent, to the biological activity of metabolites. Positive outcomes of the genotoxicity assays (AMES and MNT) and the skin sensitization assay (LLNA) can be produced by reactive metabolites that bind to DNA or skin proteins. The *in vitro* AMES assay (considering metabolic activation) was specifically chosen to evaluate the impact of adding metabolism information to a less complex endpoint (that is less dependent on pharmacokinetic variables than other *in vivo* endpoints). Moreover, reactive metabolites are also known to be a recurrent trigger of idiosyncratic adverse effects of drugs (*i.e.* unpredictable and infrequent adverse reactions often unrelated to dose).[16] The role of metabolites in the two organ toxicity endpoints (DILI and DICC), often triggered by idiosyncratic adverse reactions, was hence also investigated.[17,18]

Materials and methods

Data sets

AMES. AMES assay data were collected from the Chemical Carcinogenesis Research Information System (CCRIS),[19] the Genetic Toxicology Data Bank (GENE-TOX)[20] and the U.S. National Toxicology Program (NTP; Table S1†).[21] These data sources were selected because they provide information about the consideration of metabolic activation in the assay setup. Since the influence of the metabolites on the toxic effect was investigated in this study, only results obtained from the AMES assay accounting for metabolic activation were considered.

More specifically, the CCRIS database (stored in XML file format) was queried for mutagenicity studies based on the AMES assay, resulting in 67 907 study results (*i.e.* experimental assay outcomes on a set of compounds). For extracting these studies, the word "ames" was queried in the test system field ("mstu/tsstm") of the XML file. The retrieved AMES data were further filtered for experiments that test for metabolic activation, by querying the data for the words "liver", "hepatocytes", "s9" and "s-9" in the "matvm" field. The resulting data (38 267 study results) were further curated by removing any inconclusive or potentially ambiguous results. This was achieved by removing studies with results labeled as "weak" or as both "positive" and "negative" (*e.g.* "positive (retest was negative)"). Also inconclusive results caused by precipitating compounds were removed from the data set by querying the labels "negative" and "precipitation" (*e.g.* "negative, precipitation at 3 highest doses.").

The remaining data (38 200 study results) were labeled as "toxic" if the results field matched the word "positive", or "non-toxic" if the results field matched the word "negative". To obtain only one result per compound, the data were deduplicated based on the CAS number and any compounds with conflicting class labels were removed from the data set. This resulted in 4721 compounds with AMES data.

The GENE-TOX database was obtained from PubChem.[22] The different genotoxicity study types contained in this database were queried to select only those studies belonging to the AMES assay (*i.e.* matching the "Histidine reverse gene mutation, Ames assay" assay type). From the 1057 compounds with AMES data only the 238 results considering metabolic activation (*i.e.* matching "with metabolic activation" in the "activation" field) were conserved. The activity labels were used as is.

The NTP AMES data set contains 64 246 study results. Results from assay setups without S-9 activation and from assays with microsome-activating conditions of less than 5% were removed from the data set. Results without an activity label reported in the study conclusion and results labeled as "equivocal" were removed from the data set. These filtering steps resulted in 40 859 study results. Study outcomes with a "positive" or "weakly positive" study conclusion label were annotated as "toxic", and study outcomes with the "negative" conclusion label as "non-toxic". Compounds were deduplicated based on the CAS number, and duplicate compounds with conflicting labels were removed from the data set. In contrast to the above data sets, the NTP set did not include SMILES strings for the tested compounds. The SMILES strings were obtained by querying PubChem *via* the PUG REST interface[23] using the CAS numbers provided with the NTP data set. This resulted in 1959 compounds annotated with AMES results.

The data from the three databases were merged based on the canonical SMILES (see section Structure preparation for details). Compounds with identical canonical SMILES but differing AMES activity labels (72 compounds) were removed from the data set. This resulted in a total of 5061 compounds (1908 toxic and 3153 non-toxic compounds; Table 1).

Micronucleus test. MNT data was collected, as described by Garcia de Lomana *et al.*,[24] from (i) the European Chemicals Agency (ECHA; available at the eChemPortal),[25] (ii) the European Food Safety Authority (EFSA), curated by Benigni *et al.*,[26] and (iii) the work of Yoo *et al.*[27] The final, processed and deduplicated MNT data set consists of a total of 1775 compounds (315 toxic and 1460 non-toxic compounds; Table 1).

Drug-induced liver injury. The data set for the DILI endpoint was obtained from the verified DILIrank data (*i.e.* the revised version of their original DILIrank data set) of the U.S. Food and Drug Administration (FDA).[28] These data were derived from the observed hepatotoxicity of FDA-approved drugs described in drug labeling documents as well as evidence in literature. The drugs in this data set are classified as "most-DILI-concern", "less-DILI-concern", "no-DILI-concern" and "ambiguous-DILI-concern". For this study, binary class labels were assigned: 182 "most-DILI-concern" and 271 "less-DILI-concern" compounds were labeled as "toxic", 268 "no-DILI-concern" compounds as "non-toxic", and 239 "ambiguous-DILI-concern" compounds were removed from the data set. The final, processed and deduplicated DILI data set consists of a total of 661 compounds (435 toxic and 226 non-toxic compounds; Table 1).

Drug-induced cardiological complications. The data set for DICC was compiled, as described by Garcia de Lomana *et al.*,[24] from the work of Cai *et al.*[29] The DICC data set covers five cardiological complications: hypertension, arrhythmia, heart block, cardiac failure and myocardial infarction. Compounds were labeled as "toxic" if they were active in at least one of the five cardiological endpoints and labeled as "non-toxic" otherwise. The final, processed and deduplicated DICC data set

Table 1 Sizes of the data sets used in this work

	Number of		
Endpoint	Toxic compounds	Non-toxic compounds	Ratio
AMES	1908	3153	1 : 2
MNT	315	1460	1 : 5
DILI	435	226	2 : 1
DICC	965	2243	1 : 2
LLNA	521	749	1 : 1

contains a total of 3208 compounds (965 toxic and 2243 non-toxic compounds; Table 1).

Murine local lymph node assay. The data set for the LLNA was obtained from the work of Wilm *et al.*[30] The binary activity labels from this data set were used as is, resulting, after processing and deduplication in a total of 1270 compounds (521 toxic and 749 non-toxic compounds; Table 1).

Structure preparation

The standardization of the molecular structures followed the same procedure as described by Garcia de Lomana *et al.*[24] (with one exception, indicated below). Briefly, the SMILES strings were standardized with the ChemAxon Standardizer[31] node in KNIME[32] to remove solvents and salts, annotate aromaticity, neutralize charges and mesomerize structures (*i.e.* returning the canonical resonant form of the molecule). Moreover, compounds containing any element other than H, B, C, N, O, F, Si, P, S, Cl, Se, Br and I as well as multi-component compounds were removed from the data set. Lastly, compounds with fewer than four heavy atoms or with molecular weight greater than 1000 Da (this criterion has been introduced for the current work only) were filtered out from the respective data set.

For the remaining standardized structures, canonical SMILES were derived with RDKit[33] in KNIME. These canonical SMILES were used for the deduplication of compounds in each data set. Compounds with identical canonical SMILES but conflicting labels for an endpoint were removed from the respective endpoint data set.

Descriptor calculation

Molecular structures were encoded with count-based Morgan fingerprints with a radius of 2 bonds and a length of 2048 bytes (computed with the "RDKit Count-Based Fingerprint" node in KNIME) plus 119 1D and 2D physicochemical property descriptors (computed with the "RDKit descriptor calculation" node in KNIME). These RDKit physicochemical property descriptors capture properties such as the number of occurrences of a specific atom type, bond or ring, as well as global molecular properties such as polarity and solubility. Moreover, up to two acidic and two basic pK_a values were calculated for each molecule with the "pK_a" KNIME node from ChemAxon.[34] For molecules with fewer than two acidic or basic groups, the remaining pK_a feature values were filled with the mean value of the respective data set.

Model development and evaluation

Prior to model development, a variance filter was applied to all input features to remove those with a variance of less than 0.001. The remaining features were scaled with the StandardScaler class of scikit-learn[35] by subtracting the mean and scaling to unit variance. Both variance filtering and scaling were performed individually for each data set.

The models were evaluated within a 5-fold cross-validation (CV) framework by splitting the data into 80% training and 20% test set with the StratifiedShuffleSplit class of scikit-learn. To account for data imbalance, oversampling with SMOTENC (an extension of SMOTE that handles categorical features)[36] was performed on the training set (with a ratio of samples in the minority class with respect to the majority class of 0.8). All molecular fingerprints and discrete RDKit descriptor features (*e.g.* number of hydrogen bond donors or ring count) were specified as categorical features in SMOTENC.

For each training set, random forest (RF) models were trained with the RandomForestClassifier of scikit-learn, with default parameters, except for num_trees = 1000, min_samples_leaf = 3 and class_weights = "balanced".

For evaluating the performance of the models, the precision, recall, F1 score and Matthews Correlation Coefficient (MCC) were calculated on the respective test set of the CV. The precision measures the proportion of true positive predictions out of all positive predictions, while the recall measures the proportion of correctly predicted positive samples. The F1 score is the harmonic mean of precision and recall. The MCC takes into consideration all four classes of predictions (true positive, true negative, false positive and false negative predictions) and ranges between −1 and +1 (being +1 the perfect prediction). Both the F1 score and the MCC are robust against data imbalance.

Differences in the performance between models were evaluated with the nonparametric Mann–Whitney *U* test.[37] For comparing a pair of models, the values for a given performance metric obtained in the different CV runs were used as input for the "mannwhitneyu" function implemented in SciPy.[38] The *p*-value threshold of 0.05 was applied to consider a difference as significant. Due to the negligible number of significant results, a correction of the *p*-value accounting for the number of comparisons performed was deemed to be not necessary.

Metabolite prediction with Meteor Nexus

The metabolites were predicted with Meteor Nexus,[7,39] a leading software package for metabolism prediction that is widely applied in the industries. Meteor Nexus covers a broad range of approximately 500 manually curated biotransformations gathered from several public sources and proprietary data sets from member organizations of Lhasa Limited.

In this study, starting from the prepared molecular structures (canonical SMILES), four generations of metabolites were predicted and subsequently scored with the "Site of Metabolism (SOM) Scoring" method,[40] which is the default scoring method of Meteor Nexus. Other processing options were retained at their default setting. The score given to each metabolite is based on experimental data for compounds that are chemically related to the query compound around the site of metabolism. The molecular structures of the predicted metabolites were prepared and standardized following the same procedure described for the parent compounds (starting from the SMILES string output by Meteor Nexus).

Predicted metabolite information as input descriptors for parent compounds

Two different approaches for including metabolite information as input features in machine learning were explored (Fig. 1A). In the first approach, the above-mentioned molecular fingerprints

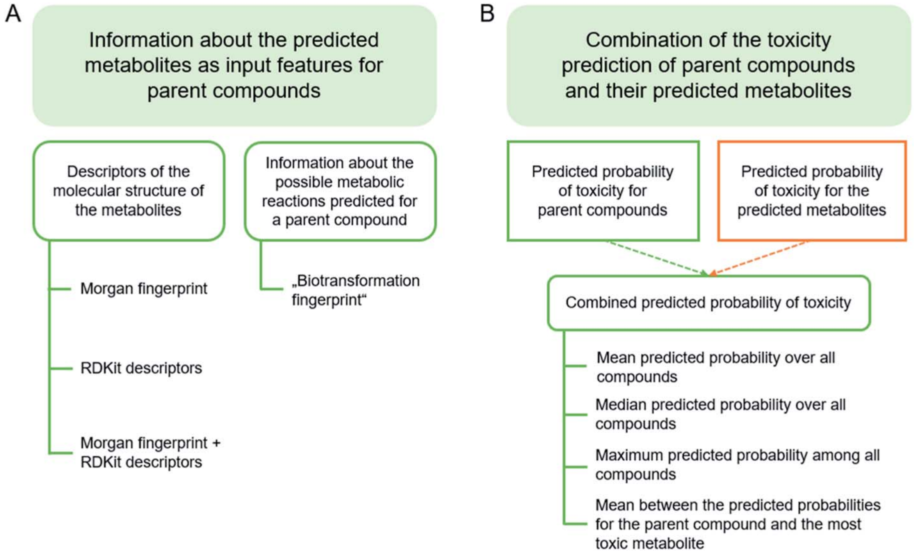

A
Information about the predicted metabolites as input features for parent compounds
Descriptors of the molecular structure of the metabolites
Information about the possible metabolic reactions predicted for a parent compound
Morgan fingerprint
RDKit descriptors
Morgan fingerprint + RDKit descriptors
„Biotransformation fingerprint“
B
Combination of the toxicity prediction of parent compounds and their predicted metabolites
Predicted probability of toxicity for parent compounds
Predicted probability of toxicity for the predicted metabolites
Combined predicted probability of toxicity
Mean predicted probability over all compounds
Median predicted probability over all compounds
Maximum predicted probability among all compounds
Mean between the predicted probabilities for the parent compound and the most toxic metabolite

Fig. 1 Overview of the different strategies explored to integrate predicted metabolite information into the *in silico* models.

and physicochemical properties for each parent compound were concatenated with chemical descriptors calculated for the top-5 predicted metabolites of that parent compound (if available; metabolite scoring with Meteor). The chemical descriptors of the metabolites comprise count-based Morgan fingerprints (radius of 2; length of 1024 bytes) and all of the 200 physicochemical property descriptors of RDKit listed under "rdkit.Chem.Descriptors._descLis". For parent compounds with fewer than five predicted metabolites, the empty values of the Morgan fingerprint vectors from the remaining metabolites were filled with zeros (indicating the absence of the structural feature) and the features corresponding to RDKit descriptors were filled with the mean value of the whole data set for that feature. Models were trained combining the molecular descriptors of the parent compounds with (a) Morgan fingerprints of the metabolites, (b) RDKit physicochemical property descriptors of the metabolites or (c) a combination of both.

In the second approach, the above-mentioned fingerprints and physicochemical properties for each parent compound were concatenated with a count-based "biotransformation fingerprint". The biotransformation fingerprint encodes the number of occurrences of a particular biotransformation (as labeled by Meteor Nexus) in the predicted metabolic tree. For each endpoint data set only those biotransformation predicted for at least one parent compound were included in the fingerprint. The feature length of the fingerprint ranges from 238 for the LLNA data set to 330 for the AMES data set. In addition to models based on the complete descriptor vector, models were also built on subsets of features selected prior to model building (in an attempt to reduce noise related to the sparsity of the biotransformation fingerprints). The feature selection was conducted on all descriptors (including fingerprints and physicochemical descriptors) and using the LassoCV implementation from scikit-learn within a 5-fold CV. Any feature with an output coefficient of zero was removed from the data prior to the training of the RF models.

Combination of the probabilities of toxicity predicted for a parent compound and its predicted metabolites

Overall predicted probability of a compound's toxicity. An overall probability for the parent compounds' toxicity was calculated by combining the predicted probabilities for the parent compounds and their predicted metabolites.

Two types of models were used for predicting the probability of toxicity:

(i) Baseline model: without the consideration of metabolites (*i.e.* trained only on the parent compounds).

(ii) Metabolism-aware model: with the consideration of metabolites (*i.e.* trained on the parent compounds and labeled metabolites).

The molecular descriptors defined in the "Descriptor calculation" section were used as input features for the parent compounds and metabolites in both types of models. For the metabolism-aware model the labels of the metabolites were assigned according to the workflow described in "Assignment of toxicity labels to metabolites". The predicted probabilities for the parent compounds (with the baseline model) were used as a baseline result to analyze whether model performance improves when considering metabolites for the prediction of toxicity.

In an attempt to obtain the most accurate predicted probability for the parent compounds and metabolites, two approaches combining the baseline model and metabolism-aware model were investigated:

(a) Baseline-approach: baseline model for the prediction of both parent compounds and metabolites.

(b) Hybrid-approach: baseline model for the prediction of parent compounds plus metabolism-aware model for the prediction of metabolites.

To obtain the overall probability of toxicity of a compound (*i.e.* with the consideration of its metabolites), the selected model was applied to calculate the probability of toxicity of the parent compound and that of the predicted metabolites (up to four generations; Fig. 2). In addition, a number of different strategies for filtering predicted metabolites according to their relevance to toxicity were explored by a grid search. These filters are based on calculated log *P*, the Meteor score and/or predicted phase II metabolism, and are intended to remove any non-toxic (since readily excretable or unlikely) metabolites. The investigated threshold values, below which metabolites were removed, are 0 and 3 for log *P*, and 100, 200 and 300 for the Meteor score. When the phase II metabolism filter was applied, metabolites formed by phase II reactions, as well as those metabolites further transformed by phase II reactions, were filtered out. A grid search over the 23 possible combinations of filters (always including the possibility of not filtering for one or more properties) was performed.

The predicted probabilities of toxicity calculated for the selected metabolites were then combined with the predicted probability for the respective parent compound. For the combination of the predicted probabilities of toxicity, four strategies were explored (Fig. 1B):

(1) Strategy 1: mean predicted probability over all compounds (*i.e.* the parent compound and all predicted metabolites).

(2) Strategy 2: median predicted probability over all compounds (*i.e.* the parent compound and all predicted metabolites).

(3) Strategy 3: maximum predicted probability among the parent compound and its predicted metabolites.

(4) Strategy 4: mean between the predicted probability of the parent compound and the maximum probability among all predicted metabolites.

If the overall probability was above 0.5, the compound was predicted as toxic and otherwise as non-toxic.

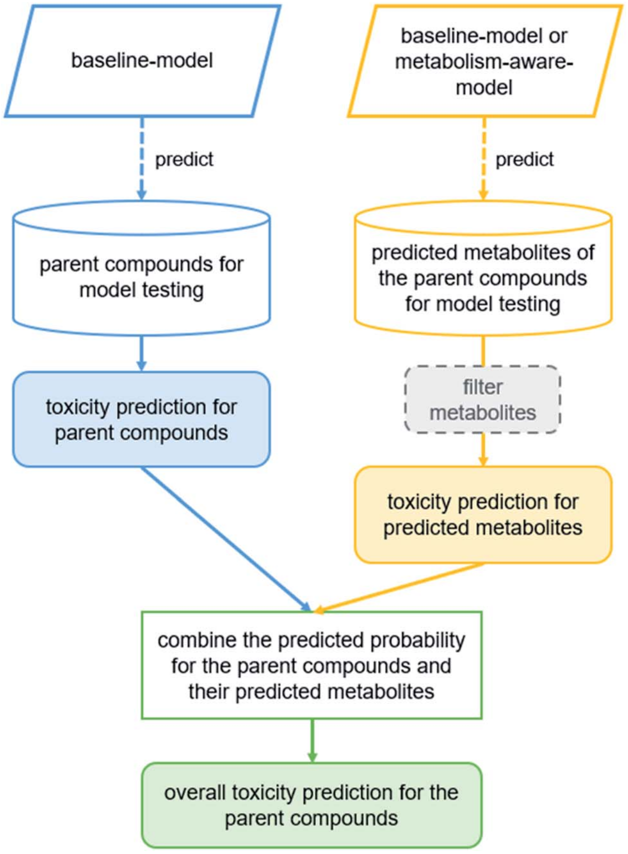

baseline-model
baseline-model or metabolism-aware-model
predict
predict
parent compounds for model testing
predicted metabolites of the parent compounds for model testing
filter metabolites
toxicity prediction for parent compounds
toxicity prediction for predicted metabolites
combine the predicted probability for the parent compounds and their predicted metabolites
overall toxicity prediction for the parent compounds

Fig. 2 Workflow for calculating the overall probability of toxicity. The baseline model or the metabolism-aware model are used to predict the probability of toxicity of parent compounds and predicted metabolites independently. The predictions for a compound and its predicted metabolites are then combined into an overall probability to obtain the toxicity label.

Assignment of toxicity labels to metabolites. In preparation of the use of the predicted metabolites for the generation of the metabolism-aware models, the metabolites were assigned toxicity labels according to the following procedure, individually for each endpoint data set:

(1) All metabolites with identical canonical SMILES as a parent compound were assigned the toxicity label of the parent compound.

(2) All metabolites not covered by step 1 and originating from non-toxic parent compounds were labeled as "non-toxic".

(3) All metabolites not covered by step 1 and originating from toxic parent compounds were compared with the already labeled metabolites. If an identical metabolite (based on the canonical SMILES) was labeled in one of the previous steps (as toxic or non-toxic), the same label was assigned.

(4) The remaining unlabeled metabolites from toxic parent compounds were labeled as "toxic" (Table 2).

Data splitting. All models were trained within a 5-fold CV framework. In order to ensure comparability between the baseline models and the metabolism-aware models, the same splits (with regard to parent compounds) were used in both cases.

To ensure that no data leak occurred in the metabolism-aware model due to the presence of identical metabolites in the training and test sets, the following procedure was conducted on each split:

(1) Stratified shuffle split was applied on the parent compounds (see Model development for details).

(2) The metabolites from the parent compounds in the test and training set were collected independently.

(3) The metabolites in the training set, which were also present in the test set (as parent or metabolite), were removed from the training set.

(4) The compounds of the training set were deduplicated based on the canonical SMILES (duplicates may appear due to repeated metabolites or metabolites identical to parent compounds).

Machine learning methods for further modeling optimization

RF, gradient boosted trees and *k*-nearest neighbors models with optimized hyperparameters were also trained in the hybrid-approach. The scikit-learn implementations 'GradientBoostingClassifier' and 'KNeighborsClassifier' were used for training the gradient boosted trees and *k*-nearest neighbor models, respectively. The hyperparameter optimization was conducted on the training set within a grid search evaluated on an inner 5-fold CV over the hyperparameters shown in Table 3.

A further set of molecular descriptors, the Continuous and Data-Driven molecular Descriptors (CDDD),[41] was employed as input for RF models. These descriptors are derived from a neural network trained to translate between two syntactically

Table 2 Overview of the metabolites labeled in each step of the labeling workflow

		Percentage of metabolites			
Endpoint	Number of metabolites	With the same molecular structure as a parent compound (step 1) (%)	Originating from non-toxic parent compounds (step 2) (%)	Originating from toxic parent compounds already labeled as toxic (step 3) (%)	Labeled as toxic as part of step 4 (%)
AMES	86 629	5.19	59.03	3.43	32.34
MNT	27 105	2.11	81.53	2.22	14.14
DILI	10 730	0.40	32.25	4.60	62.75
DICC	46 881	2.21	67.43	4.82	25.54
LLNA	16 842	3.46	51.62	5.66	39.26

Table 3 Grid of hyperparameters applied for each method

Method	Hyperparameter	Values
Random forest	n_estimators	400, 700, 1000
	Min_samples_leaf	1, 2, 3
	Class_weight	'Balanced'
Gradient boosted trees	n_estimators	200, 400, 600
	Min_samples_leaf	1, 2, 3
	Learning_rate	0.1, 0.01
K-nearest neighbors	n_neighbors	3, 5, 8
	Weights	'Uniform', 'distance'

different molecular representations. In order to make the translation, the model first learns to compress meaningful information for the representation of molecules into a vector. This vector can hence be used as a data-driven molecular descriptor, offering a conceptually different method to represent molecules, compared to the fixed Morgan fingerprints and RDKit physicochemical descriptors.

Results and discussion

Analysis of the chemical space of the parent compounds and their predicted metabolites

To understand the nature and composition of the metabolites predicted for the parent compounds in each data set, several characteristics of the predicted metabolites were analyzed.

The predicted metabolites result from phase I or phase II reactions (considering up to four generations of metabolites). The number of unique metabolites for the individual parent compounds (after removing duplicate metabolites from the respective metabolic tree) varied greatly (from 0 to 828). However, the median number of predicted metabolites among all parent compounds of an endpoint-specific data set was between 8 and 12 in all cases (Table 4).

By comparing the molecular properties of the parent compounds and their predicted metabolites (Fig. 3 reports on the AMES and MNT data sets; the graphs for the other endpoints are provided in Fig. S1†) we found the latter to have, averaged over all endpoints, a higher molecular weight (+43.9 Da) as well as a larger polar surface area (+44.4 Å^2). The predicted metabolites also tended to have a lower log *P* value than the parent compounds (−1.5; averaged over all endpoints). These shifts are primarily a result of the addition of polar groups to the parent compounds, which make them more water soluble and therefore easier to excrete. This observation is in concordance with the higher number of hydrogen bond donors and acceptors observed in metabolites compared to parent compounds (1.8 more hydrogen bond donors and acceptors on average; Fig. 3). Overall, the shifts in the physicochemical property space between the parent compounds and the predicted metabolites are consistent with those observed for parent compounds and experimentally detected metabolites,[42] a fact that supports the relevance of the predicted metabolites.

Analysis of metabolites originating from toxic and non-toxic parent compounds

The toxicity observed for a compound may be a direct result of the parent compound or of one or several of its metabolites.

Table 4 Overview of the number of predicted metabolites for the parent compounds in each endpoint data set

Endpoint	Mean number of metabolites per compound	Median number of metabolites per compound	Percentage of parent compounds without any predicted metabolite	Percentage of parent compounds with fewer than five predicted metabolites
AMES	17.34	10	1.28	19.67
MNT	15.52	9	1.66	20.90
DILI	16.28	12	0.30	11.53
DICC	14.74	10	0.88	15.94
LLNA	13.38	8	0.87	23.75

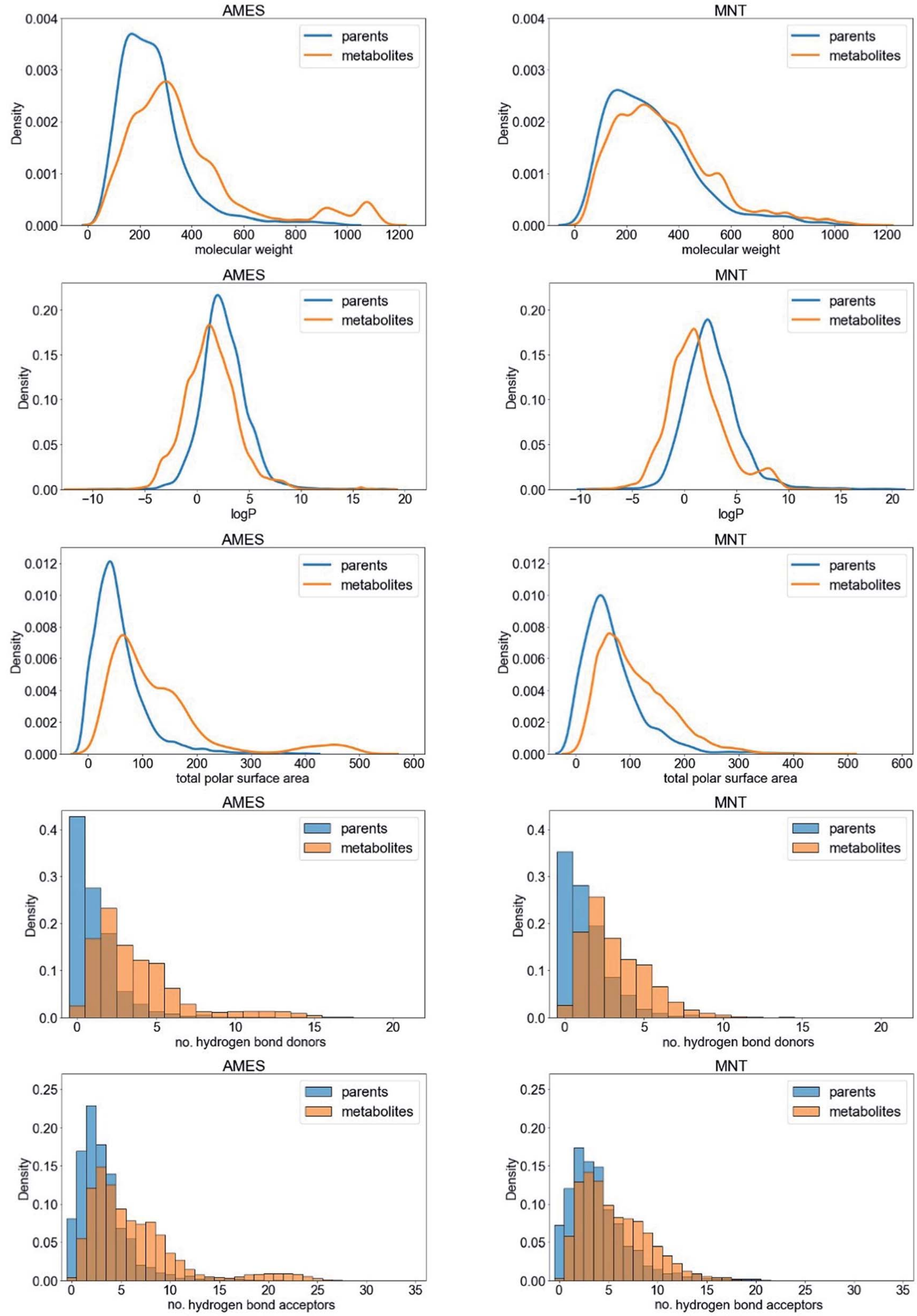


Fig. 3 Comparison of the physicochemical properties of the parent compounds (blue) and predicted metabolites (orange) represented in the AMES and MNT data sets.

Understanding the differences in the metabolites formed by toxic and non-toxic compounds may therefore help in their discrimination. However, when comparing the physicochemical properties of the (predicted) metabolites originating from toxic and from non-toxic parent compounds, we did not detect any substantial, systematic differences. This is not surprising because toxic effects may be related to a single metabolite, which is difficult to detect.

Most notable was a minor shift in the log *P* distribution (see Fig. S2† for an example of the log *P* distributions of AMES and MNT): the log *P* of metabolites originating from non-toxic compounds was generally lower (log *P* of 0.8; averaged over all metabolites of all endpoints) than for metabolites from toxic compounds (log *P* of 1.2; averaged over all metabolites of all endpoints). The higher log *P* of metabolites originating from toxic parent compounds could be related to the observed toxicity, as these metabolites are more likely to evade excretion and to cross membranes.

Another aspect that could differ from toxic to non-toxic compounds are the types of biotransformations that they are undergoing. Testa *et al.*[43] observed that some reactions are more prone to generate reactive or toxic metabolites than others. They showed that toxic metabolites are mainly formed by redox reactions, followed by conjugation reactions and, lastly, hydrolysis. Hence, the type of biotransformation that a compound undergoes may be an indicator of the compound's toxicity. To investigate whether the types of biotransformations in the metabolic trees of toxic and non-toxic compounds differ, the percentage of parent compounds of each toxicity class undergoing each biotransformation (as labeled by Meteor Nexus) was calculated for all endpoints.

We observed that some biotransformations occur more frequently in toxic parent compounds than in non-toxic ones (and *vice versa*). However, there was no single biotransformation observed to be related to the same toxicity class for all endpoints (see Fig. S3† for the examples of AMES and MNT). For instance, "aromatic reductive dehalogenation" is predicted more frequently for toxic compounds in the MNT assay (than for non-toxic compounds in this assay) while it is more often observed for non-toxic compounds in the AMES assay (than for toxic compounds in this assay).

In an analogous way, the enzymes catalyzing biotransformations in the metabolic tree of toxic and non-toxic compounds were also investigated. Similar results as for the biotransformations were observed, but, in this case, the differences between classes were smaller (*i.e.* there were few enzymes metabolizing a higher percentage of toxic or non-toxic compounds).

Baseline performance of the models

To enable the (later) quantification of the added value of metabolism prediction in toxicity prediction we generated baseline models trained exclusively on physicochemical properties of the parent compounds (encoded by count-based Morgan fingerprints and RDKit physicochemical property descriptors; see Materials and methods section for details).

The mean F1 score obtained by the baseline models within 5-fold CV ranged from 0.64 (for MNT) to 0.82 (for AMES; Table 5). The superior performance of the AMES baseline model (F1 score at least 0.09 higher than for any other baseline model) is attributed to the larger size of the data set (it is the biggest data set considered in this study with at least 1853 compounds more than any other data set) as well as the nature of the endpoint: the AMES test is an *in vitro* assay carried out on bacteria, hence representing a more simple problem than the *in vivo* endpoints based on living mammals and considered in this work. Among the *in vivo* endpoints, the model for the LLNA assay, a skin sensitization assay measuring cellular proliferation in the draining lymph nodes of mice, obtained the highest mean F1 score (0.73). The lowest F1 score (0.64) was obtained by the MNT baseline model. The precision and recall yielded by each endpoint-specific model were on a similar level in all cases, indicating a balanced ratio of false positive and false negative predictions.

Metabolite information as input descriptors for parent compounds

Molecular descriptors for metabolites. One or several chemical features present in the metabolites could be associated with the toxic effect observed for a parent compound. In an attempt to include this information in the model, molecular descriptors of the five best-scored predicted metabolites were included as further input features for model building. These molecular descriptors include (a) count-based Morgan fingerprints, (b) RDKit physicochemical property descriptors and (c) a combination thereof (see Materials and methods for details). In cases where fewer than five metabolites were predicted for a parent compound (between 12% and 24% of the compounds; Table 4), the remaining features were filled with zeros (in the case of the Morgan fingerprints) or with the mean value of the feature (in the case of the RDKit property descriptors). The trained models were evaluated by comparing the predicted label for each test parent compound with their experimental toxicity label within 5-fold CV.

When comparing the performance of these models containing metabolite information with that of the baseline models, no improvements of performance were observed (Table S2†). The few minor gains in performance did not exceed a value of +0.04 among all evaluated metrics and were not significant (at a *p*-value of 0.05; Table S3†). In several cases the addition of descriptors for the predicted metabolites led to small decreases in performance (up to a value of −0.09 among all metrics).

Biotransformation fingerprint. Our analysis of the types of biotransformations recorded for toxic and non-toxic compounds (see "Analysis of metabolites originating from toxic and non-toxic parent compounds") found indications that this information could be utilized to enhance toxicity prediction. Therefore, we derived a biotransformation fingerprint which encodes the number of occurrences of each biotransformation in the predicted metabolic tree of a compound. In combination with the molecular descriptors calculated for the parent compounds, this biotransformation fingerprint was used for the training of machine learning models (see the Materials and methods section for details).

Within the 5-fold CV framework, the performance of these models was comparable to the baseline performance of each endpoint. For all evaluated metrics the difference from the baseline performance did not exceed ±0.01 (Tables S4 and

Table 5 Performance of the baseline models within 5-fold cross-validation[a]

Endpoint	F1 score	MCC	Precision	Recall
AMES	0.82 (±0.01)	0.65 (±0.03)	0.83 (±0.01)	0.82 (±0.01)
MNT	0.64 (±0.03)	0.29 (±0.05)	0.67 (±0.02)	0.62 (±0.03)
DILI	0.68 (±0.04)	0.37 (±0.08)	0.69 (±0.04)	0.68 (±0.04)
DICC	0.69 (±0.02)	0.39 (±0.04)	0.71 (±0.02)	0.69 (±0.03)
LLNA	0.73 (±0.02)	0.47 (±0.04)	0.74 (±0.02)	0.73 (±0.02)

[a] Numbers reported in parentheses are the standard deviations.

S5†). The lack of an improvement in performance may be related to the sparsity of the biotransformation fingerprint: most of the biotransformations were not predicted to take place on more than 10% of the compounds. This low coverage of compounds may not be sufficient to enhance toxicity prediction. In order to remove possible noise caused by the sparse fingerprints, feature selection with a lasso model was applied to all input features (in order to discard irrelevant features prior to the training of the RF model). However, no relevant improvement in the performance compared to the baseline models was observed when feature selection was included prior to model training (F1 score deviations ranged from −0.05 to +0.01 among all endpoints).

Combination of predicted probabilities for parent compounds and metabolites

Another approach for considering metabolite information in toxicity prediction is the calculation of an "Overall predicted probability of toxicity" by combining the probabilities predicted for the parent compounds and their metabolites. A related approach (although based on distinct modeling methods and utilizing measured metabolites; explored for different endpoints) was applied, with some success, by Dmitriev *et al.*[2] and Filimonov *et al.*[3] (see the Introduction section for details).

In this work, we explored four strategies to combine prediction probabilities:

Strategy 1: mean of the probabilities of the parent compound and all predicted metabolites.

Strategy 2: median probability of the parent compound and all predicted metabolites.

Strategy 3: maximum probability among the parent compound and all predicted metabolites.

Strategy 4: mean between the predicted parent compound probability and the maximum probability among all metabolites (*i.e.* the probability of the metabolite that the model deems most likely to be toxic, among all predicted metabolites).

To evaluate model performance, the obtained "Overall probability of toxicity" (derived by the different strategies) was compared to the experimental toxicity label of each parent

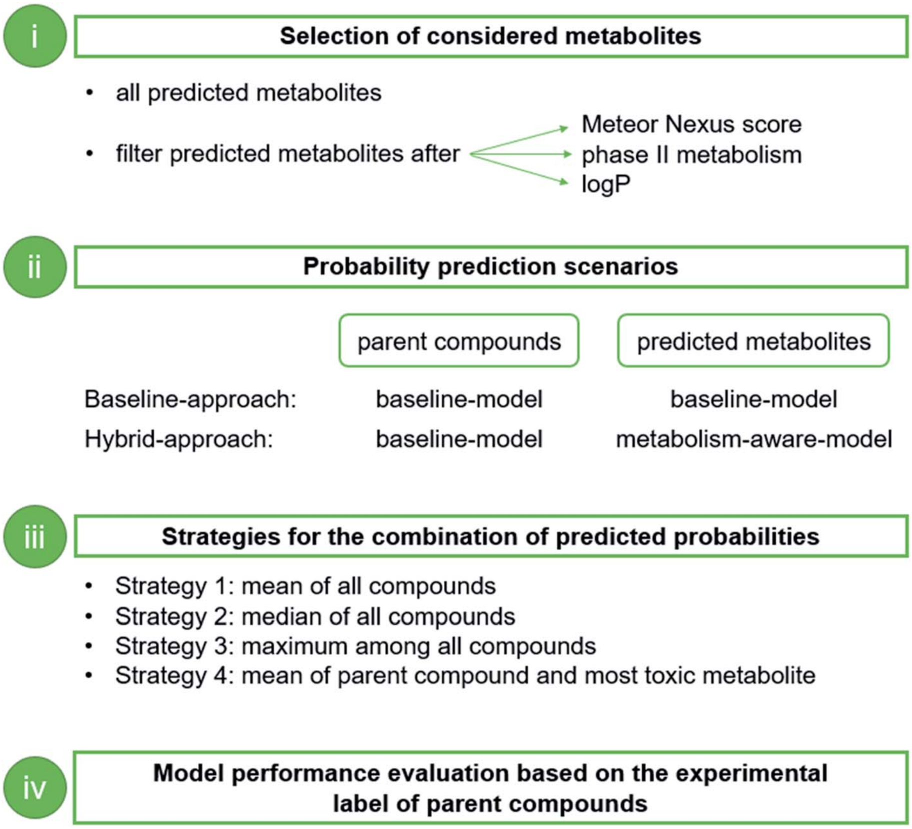

i
Selection of considered metabolites
• all predicted metabolites
• filter predicted metabolites after
Meteor Nexus score
phase II metabolism
logP
ii
Probability prediction scenarios
parent compounds
predicted metabolites
Baseline-approach:
baseline-model
baseline-model
Hybrid-approach:
baseline-model
metabolism-aware-model
iii
Strategies for the combination of predicted probabilities
• Strategy 1: mean of all compounds
• Strategy 2: median of all compounds
• Strategy 3: maximum among all compounds
• Strategy 4: mean of parent compound and most toxic metabolite
iv
Model performance evaluation based on the experimental label of parent compounds

Fig. 4 Overview of the steps (i–iv) of the workflow for combining the predicted probability of parent compounds and predicted metabolites, showing the variations investigated at each stage. A grid search among all combinations of parameters at the different stages was conducted to identify the optimum solution.

compound (within 5-fold CV; see the "Data splitting" section for details). Note that all predicted metabolites (not only the five best-scored metabolites) were considered here.

The four strategies were applied to two approaches that differ in the underlying models used for calculating the predicted probabilities (Fig. 4ii). In the baseline-approach, we applied the baseline models on the test parent compounds and their metabolites and combined them with each of the above-mentioned strategies. With strategy 1, strategy 2 and strategy 3, a drop in F1 score and MCC was observed for all investigated endpoints. Strategies 1 and 2 especially showed a decrease in recall (up to −0.17), which was sometimes compensated, to some extent, by an increased precision (up to +0.04), while the opposite effect was observed for strategy 3 (Table S6†).

Out of the four strategies, the best classification performance was obtained, in general, with strategy 4. However, the gain in F1 scores compared to the respective baseline models was 0.02 or less (and hence not significant, according to the Mann–Whitney *U* test; see Table S7† for details). Compared to strategies 1 and 2, strategy 4 may provide a well-balanced compromise between an improved capacity to detect toxicity related to metabolism and noise introduced by the predicted metabolites. A similar result was also observed in the study by Dmitriev *et al.*,[2] where several strategies to combine the predicted LD_{50} value (for acute rat toxicity) for parent compounds and their measured metabolites were investigated (mean of the predicted LD_{50} of all metabolites; mean of the predicted LD_{50} of the parent compound and all metabolites; maximum predicted LD_{50} among all metabolites; mean of the predicted LD_{50} of the parent compound and the most toxic metabolite). In agreement with our observations, Dimitriev *et al.* obtained their best results for the prediction of acute rat toxicity when taking the mean of the predicted LD_{50} for the parent compound and that of the most toxic metabolite. Also the increase in model performance (compared to taking the prediction of the parent compound only) in their case was minor (+0.03 in R^2 and no differences in RMSE).

In the hybrid-approach, the predicted probabilities of the metabolites to be toxic were calculated with a dedicated model. We addressed the possibility that the absence of relevant improvement by the four above-mentioned strategies was due to a deficient coverage of the chemical space of the metabolites by the baseline model. The differences observed in the chemical space of parent compounds and metabolites (see the section "Analysis of the chemical space of the parent compounds and predicted metabolites" for details) could indicate that some metabolites fall outside the applicability domain of the models trained only on parent compounds (baseline models).

To expand the chemical space coverage of the models and try to improve the toxicity predictions for the metabolites, models including metabolites as input data (*i.e.* with their molecular descriptors as input features and the assigned toxicity as class label) were also developed (metabolism-aware models). The toxicity label of the metabolites for these models was assigned following the workflow described in the section "Assignment of toxicity labels to metabolites". Instead of applying this straightforward labeling approach, the toxicity labels of the metabolites could have also been predicted with the baseline model. However, we did not investigate this option further as it would increase the complexity of the workflow and does not fit the purpose of this study. By labeling the metabolites we pretend to analyze whether the reason for the small model performance improvement is due to poor quality of the predicted probabilities of toxicity of the metabolites. Hence, predicting the toxicity label of the metabolites would suffer from the same limitation. We acknowledge that any manual or automatic metabolite labeling approach is a limitation of this study. The only way to overcome this limitation is the use of a large dataset of metabolites with measured toxicities. However, to our best knowledge no such dataset is in existence in the public domain.

With the hybrid-approach we aim to obtain the best predictions for each compound by predicting the probability of the parent compound to be toxic with the baseline model, and the probability to be toxic of the individual metabolites with the metabolism-aware model. Note that we also investigated the possibility to predict both the toxicity of the parent compound and the metabolites with the metabolism-aware model, but we did not see a relevant improvement compared to the baseline- or hybrid-approaches in this case and therefore did not further investigate this direction.

Compared to the baseline-approach, the hybrid-approach yielded better results in toxicity prediction. However, with improvements in the F1 scores and MCCs not exceeding 0.03 and 0.05, respectively, these results are not significantly better (based on the Mann–Whitney *U* test) than those obtained with the baseline model (Table 6). Few significant improvements were recorded for precision or recall for the MNT and DICC models (Table S8†).

The decrease in performance with strategies 1 and 2 (considering the predictions of all metabolites) in combination with the hybrid-approach was in general not as drastic as with the baseline-approach. This may indicate that the predicted probabilities for the metabolites were more accurate and did not include as much noise in the overall prediction. Again in this case, the best performance was observed with strategy 4 (averaging the probability of the parent compound and the most toxic metabolite), with only minor improvements in the F1 score of up to +0.03. Only for the DILI endpoint the F1 score decreased (by −0.02) with this strategy.

In addition, we analyzed whether the improvements in model performance may be limited due to the consideration of metabolites that are irrelevant to the observed toxic effect. In order to reduce the noise in the prediction caused by these metabolites, we applied several metabolite filters removing predicted metabolites that (a) have a low Meteor Nexus prediction score, (b) have a low calculated log *P*, or (c) are predicted to be further metabolized by conjugating enzymes (Fig. 4i).

Metabolites predicted with a low score by Meteor Nexus may be less likely to be observed *in vivo* and hence irrelevant to toxicity prediction. Metabolic reactions often lead to compounds with low log *P* values, making them more water

Table 6 Average performance within 5-fold cross-validation for the different combinations of predicted probabilities with the hybrid-approach

Endpoint	Combination[a]	F1 score	MCC	Precision	Recall
AMES	Baseline performance	0.82 (±0.01)	0.65 (±0.03)	0.83 (±0.01)	0.82 (±0.01)
	Strategy 1	0.80 (±0.01)	0.61 (±0.02)	0.82 (±0.01)	0.80 (±0.01)
	Strategy 2	0.80 (±0.01)	0.61 (±0.02)	0.81 (±0.01)	0.79 (±0.01)
	Strategy 3	0.79 (±0.02)	0.60 (±0.03)	0.79 (±0.01)	0.81 (±0.01)
	Strategy 4	0.83 (±0.02)	0.65 (±0.03)	0.82 (±0.02)	0.83 (±0.02)
MNT	Baseline performance	0.64 (±0.03)	0.29 (±0.05)	0.67 (±0.02)	0.62 (±0.03)
	Strategy 1	0.61 (±0.03)	0.31 (±0.05)	0.75 (±0.02)	0.59 (±0.03)
	Strategy 2	0.61 (±0.04)	0.29 (±0.06)	0.74 (±0.03)	0.59 (±0.03)
	Strategy 3	0.65 (±0.02)	0.31 (±0.03)	0.64 (±0.02)	0.67 (±0.02)
	Strategy 4	0.66 (±0.03)	0.33 (±0.06)	0.69 (±0.04)	0.65 (±0.03)
DILI	Baseline performance	0.68 (±0.04)	0.37 (±0.08)	0.69 (±0.04)	0.68 (±0.04)
	Strategy 1	0.66 (±0.03)	0.33 (±0.06)	0.67 (±0.03)	0.66 (±0.03)
	Strategy 2	0.66 (±0.03)	0.32 (±0.06)	0.67 (±0.03)	0.65 (±0.03)
	Strategy 3	0.59 (±0.05)	0.31 (±0.07)	0.73 (±0.03)	0.60 (±0.03)
	Strategy 4	0.66 (±0.03)	0.37 (±0.05)	0.73 (±0.02)	0.65 (±0.03)
DICC	Baseline performance	0.69 (±0.02)	0.39 (±0.04)	0.71 (±0.02)	0.69 (±0.03)
	Strategy 1	0.68 (±0.02)	0.40 (±0.03)	0.75 (±0.02)	0.66 (±0.01)
	Strategy 2	0.68 (±0.02)	0.39 (±0.04)	0.73 (±0.02)	0.66 (±0.02)
	Strategy 3	0.68 (±0.01)	0.38 (±0.01)	0.67 (±0.00)	0.70 (±0.00)
	Strategy 4	0.72 (±0.02)	0.44 (±0.03)	0.72 (±0.01)	0.72 (±0.02)
LLNA	Baseline performance	0.73 (±0.02)	0.47 (±0.04)	0.74 (±0.02)	0.73 (±0.02)
	Strategy 1	0.70 (±0.02)	0.42 (±0.04)	0.73 (±0.02)	0.70 (±0.02)
	Strategy 2	0.71 (±0.03)	0.44 (±0.05)	0.73 (±0.02)	0.71 (±0.03)
	Strategy 3	0.69 (±0.01)	0.42 (±0.03)	0.71 (±0.02)	0.71 (±0.01)
	Strategy 4	0.74 (±0.02)	0.48 (±0.05)	0.74 (±0.02)	0.74 (±0.03)

[a] The baseline performance corresponds to models considering only parent compounds. Strategies 1, 2 and 3 correspond to taking the mean, median and maximum predicted probability among the parent compound and its metabolites, respectively. Strategy 4 corresponds to the mean between the predicted probability for the parent compound and the highest probability predicted for any of its metabolites.

soluble and therefore easier to excrete. These metabolites are also unlikely to cross membranes and they are less likely to induce toxic effects. Along the same lines, phase II metabolism facilitates the conjugation of compounds with polar moieties, making them more water soluble. It has already been observed that only few conjugation reactions lead to toxic metabolites.[43] Following this reasoning, several thresholds for removing metabolites based on their Meteor Nexus score as well as calculated log *P* values were investigated. Also strategies to remove metabolites formed by phase II reactions, or remove metabolites which are further transformed by phase II reactions were explored. A grid search over all filtering possibilities (and all above-mentioned approaches and strategies) was conducted on each data set to obtain the most favorable combinations.

In most cases, reducing the number of metabolites considered for the prediction based on these parameters did not yield better models. Among the five top-ranked models (based on the F1 score) of the grid search, only in a few cases, minor improvements of up to +0.06 among all metrics and endpoints were observed (Table S9†). However, these performance improvements were not significant for any endpoint compared to the baseline performance (Table S10†).

Exploration of further modeling approaches with the hybrid-approach

To evaluate whether the predictions may be improved by optimizing the modeling approach, different machine learning modeling methods with optimized hyperparameters (within a grid search; see Materials and methods section for details) and a further, distinct set of descriptors (CDDD descriptors)[41] were investigated at the example of the best performing approach, namely the hybrid-approach.

The F1 score obtained for the following machine learning setup combinations is shown in Table S11:† RF, gradient boosted trees and k-nearest neighbors, each with and without the use of oversampling with SMOTENC (based on Morgan fingerprint and RDKit physicochemical descriptors as input descriptors). Moreover, the performance of RF models trained on CDDD descriptors (including oversampling with SMOTE) are also provided.

The results obtained with these new models do not deviate from those obtained with the models generated with the initial modeling setup (*i.e.* RF with fixed hyperparameters; combination of Morgan fingerprints and RDKit physicochemical descriptors; oversampling with SMOTENC; results reported in Table 6): the largest observed improvement in F1 scores yielded by the new models was of just +0.01. The conclusions derived in the 'Combination of predicted probabilities for parent compounds and metabolites' section remain consistent with the new results. The explicit incorporation of predicted metabolite information in toxicity prediction models did not significantly improve the toxicity predictions of these models either. Although there was often no benefit compared to the baseline models (or the benefit was small), the best strategy for combining the predicted probabilities of parent compounds and metabolites was, also in this case, strategy 4 (taking the mean between the predicted probability of the parent compound and the maximum probability among all predicted metabolites).

Conclusions

In this work we systematically investigated a variety of strategies to enhance toxicity prediction by taking into account xenobiotic metabolism. Our results show that none of these strategies produces models that consistently outperform others. The best results were obtained by averaging the probability of toxicity predicted for the parent compound and the maximum probability of toxicity predicted for any metabolite. This approach yielded models with F1 scores up to +0.03 higher than the baseline models disregarding metabolism.

We observed that models trained exclusively on the parent compounds often produce poor predictions for the metabolites as their chemistry often differs. Including labeled metabolites in the training set of the models slightly improved the predictions of toxicity for the metabolites and hence the overall result of averaging the probabilities of toxicity for parent compounds and their metabolites. In some cases, discarding unlikely or water-soluble metabolites slightly improved the predictions (F1 score up to +0.04 higher than for the baseline models).

While metabolites can be key to detecting and understanding toxicity, they also add a new layer of complexity. The metabolites formed, their concentrations in the organism, and their excretion kinetics are often unknown. Therefore, including metabolism data in toxicity prediction poses veritable challenges. The fragile balance between added signal and added noise, when working with predicted metabolites in machine learning, may explain the small differences in performance of the models including metabolism information for toxicity prediction compared to the baseline models. It is clear from these results that there is still a long way to go in the development of sufficiently accurate models for metabolism prediction which, in turn, can boost toxicity prediction.

Abbreviations

CCRIS	Chemical carcinogenesis research information system
CV	Cross-validation
DICC	Drug-induced cardiological complications
DILI	Drug-induced liver injury
ECHA	European chemicals agency
EFSA	European food safety authority
FDA	U.S. food and drug administration
GENE-TOX	Genetic toxicology data bank
LLNA	Murine local lymph node assay
LMC	Laboratory of mathematical chemistry
MCC	Matthews correlation coefficient
MNT	Micronucleus test
NTP	National toxicology program
RF	Random forest
TIMES	Tissue metabolic simulator

Data availability

All data sets used in this study are publicly available. Due to licensing reasons, the original data and the predicted metabolites cannot be provided with this publication. However, a detailed protocol for the reproducible collection and preprocessing of the data utilized in this work is provided in the Materials and methods section. Moreover, Table S1† contains links for downloading the original data and complementary information about the data sets. Also detailed KNIME workflows used for preprocessing each data set and calculating the chemical descriptors of the parent compounds are provided in the ESI.† The workflows and parameters used for developing the models and necessary for reproducing the results are described in detail in the Materials and methods section. The code used for model training and evaluation can be accessed at https://github.com/marinaglr/metabio.

Conflicts of interest

MGL and MM are employed at BASF SE. AV served as consultant for BASF SE.

Acknowledgements

We thank Robert Landsiedel and Roland Buesen from BASF SE for their valuable comments and suggestions based on their expertise in toxicology. A. V. is supported by the Federal Ministry of Education and Research (BMBF), Germany (grant 031A262C).

References

1 M. Pirmohamed, N. R. Kitteringham and B. Kevin Park, The Role of Active Metabolites in Drug Toxicity, *Drug Saf.*, 1994, **11**, 114–144.

2 A. Dmitriev, A. Rudik, D. Filimonov, A. Lagunin, P. Pogodin, V. Dubovskaja, V. Bezhentsev, S. Ivanov, D. Druzhilovsky, O. Tarasova and V. Poroikov, Integral Estimation of Xenobiotics' Toxicity With Regard to Their Metabolism in Human Organism, *Pure Appl. Chem.*, 2017, **89**, 1449–1458.

3 D. A. Filimonov, A. V. Rudik, A. V. Dmitriev and V. V. Poroikov, Computer-Aided Estimation of Biological Activity Profiles of Drug-Like Compounds Taking into Account Their Metabolism in Human Body, *Int. J. Mol. Sci.*, 2020, **21**, 7492.

4 Y. Djoumbou-Feunang, J. Fiamoncini, A. Gil-de-la-Fuente, R. Greiner, C. Manach and D. S. Wishart, BioTransformer: A Comprehensive Computational Tool for Small Molecule Metabolism Prediction and Metabolite Identification, *J. Cheminf.*, 2019, **11**, 2.

5 S. Tian, X. Cao, R. Greiner, C. Li, A. Guo and D. S. Wishart, CyProduct: A Software Tool for Accurately Predicting the Byproducts of Human Cytochrome P450 Metabolism, *J. Chem. Inf. Model.*, 2021, **61**, 3128–3140.

6 C. de Bruyn Kops, M. Šícho, A. Mazzolari and J. Kirchmair, GLORYx: Prediction of the Metabolites Resulting from

Phase 1 and Phase 2 Biotransformations of Xenobiotics, *Chem. Res. Toxicol.*, 2021, **34**, 286–299.

7 C. A. Marchant, K. A. Briggs and A. Long, *In Silico* Tools for Sharing Data and Knowledge on Toxicity and Metabolism: Derek for Windows, Meteor, and Vitic, *Toxicol. Mech. Methods*, 2008, **18**, 177–187.

8 L. Ridder and M. Wagener, SyGMa: Combining Expert Knowledge and Empirical Scoring in the Prediction of Metabolites, *ChemMedChem*, 2008, **3**, 821–832.

9 O. Mekenyan, S. Dimitrov, T. Pavlov, G. Dimitrova, M. Todorov, P. Petkov and S. Kotov, Simulation of Chemical Metabolism for Fate and Hazard Assessment. V. Mammalian Hazard Assessment, *SAR QSAR Environ. Res.*, 2012, **23**, 553–606.

10 J. Zaretzki, M. Matlock and S. J. Swamidass, XenoSite: Accurately Predicting CYP-Mediated Sites of Metabolism with Neural Networks, *J. Chem. Inf. Model.*, 2013, **53**, 3373–3383.

11 O. Mekenyan, S. Dimitrov, R. Serafimova, E. Thompson, S. Kotov, N. Dimitrova and J. D. Walker, Identification of the Structural Requirements for Mutagenicity by Incorporating Molecular Flexibility and Metabolic Activation of Chemicals I: TA100 Model, *Chem. Res. Toxicol.*, 2004, **17**, 753–766.

12 G. M. Ovanes, D. D. Sabcho, S. P. Todor and D. V. Gilman, A Systematic Approach to Simulating Metabolism in Computational Toxicology. I. The TIMES Heuristic Modelling Framework, *Curr. Pharm. Des.*, 2004, **10**, 1273–1293.

13 S. D. Dimitrov, L. K. Low, G. Y. Patlewicz, P. S. Kern, G. D. Dimitrova, M. H. I. Comber, R. D. Phillips, J. Niemela, P. T. Bailey and O. G. Mekenyan, Skin Sensitization: Modeling Based on Skin Metabolism Simulation and Formation of Protein Conjugates, *Int. J. Toxicol.*, 2005, **24**, 189–204.

14 O. Mekenyan, G. Patlewicz, C. Kuseva, I. Popova, A. Mehmed, S. Kotov, T. Zhechev, T. Pavlov, S. Temelkov and D. W. Roberts, A Mechanistic Approach to Modeling Respiratory Sensitization, *Chem. Res. Toxicol.*, 2014, **27**, 219–239.

15 O. G. Mekenyan, P. I. Petkov, S. V. Kotov, S. Stoeva, V. B. Kamenska, S. D. Dimitrov, M. Honma, M. Hayashi, R. Benigni, E. M. Donner and G. Patlewicz, Investigating the Relationship between *in Vitro–in Vivo* Genotoxicity: Derivation of Mechanistic QSAR Models for *in Vivo* Liver Genotoxicity and *in Vivo* Bone Marrow Micronucleus Formation Which Encompass Metabolism, *Chem. Res. Toxicol.*, 2012, **25**, 277–296.

16 T. Cho and J. Uetrecht, How Reactive Metabolites Induce an Immune Response That Sometimes Leads to an Idiosyncratic Drug Reaction, *Chem. Res. Toxicol.*, 2017, **30**, 295–314.

17 N. P. Chalasani, P. H. Hayashi, H. L. Bonkovsky, V. J. Navarro, W. M. Lee and R. J. Fontana, ACG Clinical Guideline: The Diagnosis and Management of Idiosyncratic Drug-Induced Liver Injury, *Am. J. Gastroenterol.*, 2014, **109**, 950–966; quiz 967.

18 I. Hopper, Cardiac Effects of Non-Cardiac Drugs, *Aust. Prescr.*, 2011, **34**, 52–54.

19 National Institutes of Health, Chemical Carcinogenesis Research Information System (CCRIS), accessed February 19, 2021, https://ftp.nlm.nih.gov/projects/ccrislease/.

20 National Institutes of Health, GENE-TOX, accessed February 19, 2021, https://www.nlm.nih.gov/databases/download/genetox.html.

21 U.S. Department of Health and Human Services, National Toxicology Program, accessed February 19, 2021, https://cebs.niehs.nih.gov/datasets/search/ames.

22 NCBI, PubChem Bioassay Record for AID 1259408, GENE-TOX Mutagenicity Studies, Source: Genetic Toxicology Data Bank (GENE-TOX), accessed February 19, 2021, https://pubchem.ncbi.nlm.nih.gov/bioassay/1259408.

23 S. Kim, P. A. Thiessen, T. Cheng, B. Yu and E. E. Bolton, An Update on PUG-REST: RESTful Interface for Programmatic Access to PubChem., *Nucleic Acids Res.*, 2018, **46**, W563–w570.

24 M. Garcia de Lomana, A. Morger, U. Norinder, R. Buesen, R. Landsiedel, A. Volkamer, J. Kirchmair and M. Mathea, ChemBioSim: Enhancing Conformal Prediction of *In Vivo* Toxicity by Use of Predicted Bioactivities, *J. Chem. Inf. Model.*, 2021, **61**, 3255–3272.

25 eChemPortal, accessed August 6, 2020, https://www.echemportal.org/echemportal/.

26 R. Benigni, C. Laura Battistelli, C. Bossa, A. Giuliani, E. Fioravanzo, A. Bassan, M. Fuart Gatnik, J. Rathman, C. Yang and O. Tcheremenskaia, Evaluation of the Applicability of Existing (Q)SAR Models for Predicting the Genotoxicity of Pesticides and Similarity Analysis Related With Genotoxicity of Pesticides for Facilitating of Grouping and Read Across, *EFSA Support. Publ.*, 2019, 1598E.

27 J. W. Yoo, N. L. Kruhlak, C. Landry, K. P. Cross, A. Sedykh and L. Stavitskaya, Development of Improved QSAR Models for Predicting the Outcome of the *in Vivo* Micronucleus Genetic Toxicity Assay, *Regul. Toxicol. Pharmacol.*, 2020, **113**, 104620.

28 M. Chen, A. Suzuki, S. Thakkar, K. Yu, C. Hu and W. Tong, DILIrank: The Largest Reference Drug List Ranked by the Risk for Developing Drug-Induced Liver Injury in Humans, *Drug Discovery Today*, 2016, **21**, 648–653.

29 C. Cai, J. Fang, P. Guo, Q. Wang, H. Hong, J. Moslehi and F. Cheng, *In Silico* Pharmacoepidemiologic Evaluation of Drug-Induced Cardiovascular Complications Using Combined Classifiers, *J. Chem. Inf. Model.*, 2018, **58**, 943–956.

30 A. Wilm, U. Norinder, M. I. Agea, C. de Bruyn Kops, C. Stork, J. Kühnl and J. Kirchmair, Skin Doctor CP: Conformal Prediction of the Skin Sensitization Potential of Small Organic Molecules, *Chem. Res. Toxicol.*, 2021, **34**, 330–344.

31 Standardizer was used for structure canonicalization and transformation, *JChem 3.5.0*, ChemAxon, http://www.chemaxon.com.

32 M. R. Berthold, N. Cebron, F. Dill, T. R. Gabriel, T. Kötter, T. Meinl, P. Ohl, C. Sieb, K. Thiel and B. Wiswedel, KNIME: The Konstanz Information Miner, in *Studies in*

© 2022 The Author(s). Published by the Royal Society of Chemistry

Classification, Data Analysis, and Knowledge Organization (GfKL 2007), Version 4.3.3., Springer, 2007.

33 G. Landrum, *RDKit: Open-Source Cheminformatics Software, Version 4.2.0.*, 2016.

34 The pKa Plugin was used for the calculation of the pKa constant value of molecules, *JChem 3.5.0*, ChemAxon, http://www.chemaxon.com.

35 F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and É. Duchesnay, *Scikit-learn: Machine Learning in Python, version 0.22.1*, 2011, vol. 12, pp. 2825–2830.

36 N. V. Chawla, K. Bowyer, L. O. Hall and P. O. Kegelmeyer, *SMOTE: Synthetic Minority Over-Sampling Technique*, 2002, **16**, 321–357.

37 H. B. Mann and D. R. Whitney, On a Test of Whether one of Two Random Variables is Stochastically Larger Than the Other, *Ann. Math. Stat.*, 1947, **18**, 50–60.

38 P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, A. Vijaykumar, A. P. Bardelli, A. Rothberg, A. Hilboll, A. Kloeckner, A. Scopatz, A. Lee, A. Rokem, C. N. Woods, C. Fulton, C. Masson, C. Häggström, C. Fitzgerald, D. A. Nicholson, D. R. Hagen, D. V. Pasechnik, E. Olivetti, E. Martin, E. Wieser, F. Silva, F. Lenders, F. Wilhelm, G. Young, G. A. Price, G.-L. Ingold, G. E. Allen, G. R. Lee, H. Audren, I. Probst, J. P. Dietrich, J. Silterra, J. T. Webber, J. Slavič, J. Nothman, J. Buchner, J. Kulick, J. L. Schönberger, J. V. de Miranda Cardoso, J. Reimer, J. Harrington, J. L. C. Rodríguez, J. Nunez-Iglesias, J. Kuczynski, K. Tritz, M. Thoma, M. Newville, M. Kümmerer, M. Bolingbroke, M. Tartre, M. Pak, N. J. Smith, N. Nowaczyk, N. Shebanov, O. Pavlyk, P. A. Brodtkorb, P. Lee, R. T. McGibbon, R. Feldbauer, S. Lewis, S. Tygier, S. Sievert, S. Vigna, S. Peterson, S. More, T. Pudlik, T. Oshima, T. J. Pingel, T. P. Robitaille, T. Spura, T. R. Jones, T. Cera, T. Leslie, T. Zito, T. Krauss, U. Upadhyay, Y. O. Halchenko, Y. Vázquez-Baeza and C. SciPy, SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python (Version 1.4.1.), *Nat. Methods*, 2020, **17**, 261–272.

39 *Meteor Nexus v3.1.0*, Lhasa Limited.

40 C. A. Marchant, E. M. Rosser and J. D. Vessey, A k-Nearest Neighbours Approach Using Metabolism-related Fingerprints to Improve *In Silico* Metabolite Ranking, *Mol. Inf.*, 2017, **36**, 1600105.

41 R. Winter, F. Montanari, F. Noé and D. A. Clevert, Learning Continuous and Data-Driven Molecular Descriptors by Translating Equivalent Chemical Representations, *Chem. Sci.*, 2019, **10**, 1692–1701.

42 J. Kirchmair, A. Howlett, J. E. Peironcely, D. S. Murrell, M. J. Williamson, S. E. Adams, T. Hankemeier, L. van Buren, G. Duchateau, W. Klaffke and R. C. Glen, How Do Metabolites Differ from Their Parent Molecules and How Are They Excreted?, *J. Chem. Inf. Model.*, 2013, **53**, 354–367.

43 B. Testa, A. Pedretti and G. Vistoli, Reactions and Enzymes in the Metabolism of Drugs and Other Xenobiotics. Drug Discov, *Today*, 2012, **17**, 549–560.

5 Conclusions and future directions

Risk assessment is a requirement for the registration of any developed chemical, including pharmaceutical drugs, agrochemicals and cosmetics. Nowadays, toxicity testing still relies heavily on animal assays. Not only do these tests involve ethical concerns but they also are demanding with respect to costs, expertise and time. As the number and amounts of chemicals released to the environment are growing and more detailed studies on the safety of chemicals are required by the authorities for their registration, the urge to develop alternative methods and reduce animal assays is increasing. Such alternative methods rely on in vitro, in chemico or in silico methods that aim to reproduce in vivo effects. However, the substitution of in vivo animal assays by alternative methods is still a challenging task for most toxicity endpoints due to the highly complex processes taking place in the body that cause poor in silico predictivity and discrepancies between the in vitro and in vivo observations. For instance, compounds active on an off-target in vitro may not show such activity in vivo if the off-target effect is compensated by other routes or regulation pathways (e.g. thyroid hormone homeostasis can be regulated, among others, by the conversion to more or less active forms of the thyroid hormone or by changes in the synthesis rate of the hormones). The contrary relationship between in vitro and in vivo observations is also possible, as compounds that are inactive in vitro may cause toxic effects in vivo if, for example, compounds are metabolized in the living organism into toxic metabolites. Therefore, integrative approaches including in vitro and in silico models for different targets and pharmacokinetic processes are necessary for a successful substitution of animal assays and improvement of the current risk assessment strategies. Alternative models, designed directly on human tissues or effects, may even be more accurate than in vivo models, as they avoid the interspecies discrepancies observed when interpolating results from animal models to humans. To benefit from all current tools in regulatory contexts as well, the Integrated Approaches to Testing and Assessment[93] (IATA) aim to integrate existing information and methods (including QSAR, in vitro and in vivo models) to make decisions about the safety of compounds.

This thesis aims to contribute to the establishment of alternative methods by the development of computational toxicology tools that help to reproduce in vivo effects and bridge the in vitro-in vivo gap from several perspectives. In the first study, we tackled the prediction of endocrine disruptors by setting the focus on a set of single, more approachable biological targets involved in the endocrine regulation. Since hormone homeostasis is regulated by highly complex pathways, the in silico prediction of endocrine disruption effects in vivo is particularly challenging. Reducing the complexity of the problem to the prediction of MIEs allows the development of useful tools for determining the AOPs and guiding the detection of endocrine disruptors. This approach was investigated at the example of thyroid hormones, one of the least studied endocrine systems in computational toxicology. For this study, data sets for nine assays related to MIEs of thyroid hormone homeostasis disruption were collected (from the ToxCast database and related literature) and thoroughly curated (section 4.1.). This data served for the development of ML models, which were optimized with regard to the ML algorithm (among five possible algorithms) and the data balancing technique (among three techniques). The models for the TPO and TR endpoints achieved high predictive performance (0.83 and 0.81 F1 score within a 10-fold CV), while the F1 score for the remaining endpoints ranged from 0.65 to 0.77. The reduced performance of these endpoints may be related to the limited amount of data available for model development (especially active compounds) as well as the quality of this data (e.g. due to assay set ups prone to false positives). Multi-task NN models, trained on all endpoints (or a subset of them) at the same time, were also investigated. In contrast to our expectation, these models did not benefit from a transfer of information between endpoints and showed similar performance to the single-task models. The lack of benefit of multi-task models may be due to the small complementarity of the data sets (since all data sets contain compounds from the ToxCast library) and the limited biological and structural relationship of the biological targets. In order to do a deeper analysis of the reliability of the models, we evaluated the confidence of the predictions with regard to (i) their distance to the decision threshold and (ii) the similarity of the test compounds to the compounds in the training set. We observed that these two metrics correlate with the confidence of the predictions, and may be used to define thresholds for the AD of the models.

The models developed within this study can help to identify the mechanisms of action of endocrine disruptors and to elucidate or confirm their related AOPs. The establishment of high quality AOPs can serve for the generation of reliable assessment strategies based on in vitro and in silico data, and hence reduce the frequency of animal assays to its minimum. As more compounds are tested on the studied assays in the future, and orthogonal assays are carried out to discard false positive readouts, it would be interesting to update the models to ensure the best coverage of the chemical space and highest quality of the models.

In the second study of this thesis, an integrative approach was developed for enhancing the performance of in silico models targeted directly to the prediction of in vivo effects. For this purpose, the outcome of compounds in in vitro and pharmacokinetics assays were included in in vivo toxicity prediction models in the form of bioactivity descriptors (section 4.2.). We investigated whether these bioactivity descriptors can better represent the behavior of compounds in biological systems and enhance the prediction of in vivo toxicity endpoints compared to using only chemical descriptors. Exclusively public domain data was used in this study, which covered over 300 biological assays (employed to build the bioactivity descriptors) and three in vivo endpoints (used as final modeling target): a genotoxicity assay (MNT) and two organ toxicity assays (DILI and DICC). Due to the sparsity of the biological assay data, predicted outcomes for the in vitro and pharmacokinetics assays were introduced instead of measured ones. For calculating the predicted bioactivity descriptors, a ML model was trained on each biological assay and applied on the in vivo toxicity data sets to compute the predicted outcomes. All the developed models in this study were embedded in a CP framework, which enabled the mathematical definition of their AD. The results showed a significant increase in the performance of models including bioactivity descriptors for two of the three analyzed in vivo endpoints (MNT and DICC) compared to models trained exclusively on chemical descriptors. Both an increase in the efficiency (up to 0.12), measuring the ratio of single class predictions, and in the F1 score (up to 0.10), measuring the quality of the single class predictions, were obtained with the developed bioactivity descriptors for these endpoints.

The positive outcome of this study is not only relevant for the development of better performing ML models for in vivo toxicity prediction. The predicted bioactivity descriptors are also promising for improving the similarity assessment of compounds for read-across applications. After confirming the suitability of the described approach, future work should be centered on the application of the predicted bioactivity descriptors on defined endpoints following the OECD principles[94] (i.e. with homogeneous data sets derived by the exact same experimental protocol and conditions), in order to make the models suitable for regulatory applications as well.

Before applying ML models on new data for regulatory purposes or research applications, it has to be ensured that the model can be applied on the test data with confidence. A common issue in computational toxicology is the appearance of data drifts over time or between data sets from different sources, which hinder the reliable application of ML models. When data drifts appear and the training and test sets have different distributions, CP models can no longer ensure the defined error rate (i.e. they are not valid). This was also the observation made when applying the developed CP models for MNT and DILI to BASF SE inhouse data that covered different parts of the descriptor space than the training data or were derived with differing assay conditions. To preserve the information learned from the public data and avoid a complete model retraining (for which the amount of inhouse data was also too scarce), we evaluated a strategy to adapt the already trained models to the new data. This strategy consists of exchanging only the calibration set of the CP models by data from the same distribution as the new test data. Simulating a real-case scenario, the oldest BASF data was used for recalibration and the model performance was evaluated on the newest inhouse data (section 4.3.). A similar strategy was also tested on ChEMBL time-split data to analyze the value of the approach for mitigating temporal data drifts. It was observed that exchanging the calibration set helped to restore the validity to its expected value (i.e. 0.8 at a significance level of 0.2) in most cases, making the models applicable to the new test sets. The models with restored validity were hence able to identify samples for which the model lacks information to make a reliable prediction. Since the new test data was in many cases only poorly covered by the data used to train the model (which remained unchanged), a high number of unreliable predictions was observed. In contrast to the original models, the recalibrated models identified these unreliable predictions and therefore showed decreased efficiencies and increased validities. Compared to the original models, the updated models yielded up to 0.10 lower efficiencies for the ChEMBL endpoints, and up to 0.55 lower efficiencies for the two inhouse endpoints. This efficiency drop may be considered acceptable as long as the validity is restored, which is the prerequisite for applying the CP models with confidence. In cases where the efficiency is too low, completely retraining the model may be required to have useful models in practice. With regard to the quality of the predictions, the balanced accuracy remained stable for the ChEMBL endpoints after the recalibration. In the case of the two inhouse endpoints, a more balanced accuracy between active and inactive predictions was obtained compared to the original models, which predicted almost all compounds to be inactive. In future work, it should be investigated further how the recalibration strategy compares to the results of retraining the ML models as the quantity of new data increases, or as the new data show different degrees of similarity to the original training data. Although defining an overall best approach may be unfeasible, such an analysis could deliver a thorough best practice guide for future applications.

With the previously developed bioactivity descriptors we achieved to account for some of the ADME parameters that contribute to the discrepancies between in vitro and in vivo outcomes. In the last study of this thesis, we tried to go a step further and include information about metabolism in the predictive models for five toxicity endpoints (covering two genotoxicity, two organ toxicity and one skin sensitization endpoints; section 4.4.). Xenobiotic metabolism is one of the key parameters influencing the outcome of in vivo assays but is, nevertheless, usually disregarded in in silico toxicology. In this study, predicted metabolites instead of measured ones were employed to expand the coverage of the chemical space and make the developed models easily applicable to new data. Several strategies for incorporating metabolism information in ML models were evaluated, including (i) the computation of a "biotransformation fingerprint" (indicating the occurrence of specific metabolic transformations on a parent compound), (ii) the concatenation of physicochemical descriptors of the predicted metabolites as further input features for model development, or (iii) the combination of the predicted probability of toxicity for each parent compound and its predicted metabolites. Among all the strategies, the best results were obtained by taking the mean of the probability of toxicity for the parent compound and the maximum probability of toxicity for any of its metabolites. However, compared to the baseline models not including metabolism information, this approach yielded only up to 0.03 higher F1 scores among all evaluated endpoints. A further data cleaning step using several filters to remove unlikely or highly soluble metabolites was also implemented but did not result in relevant model performance improvements either. Although metabolism plays a major role in the toxicity outcome of a compound, it was shown that including predicted metabolism information in the models may incorporate high levels of uncertainty that counteract the valuable added information. As more metabolism data are available and metabolism predictors get better at prioritizing relevant structures, these approaches could be revised to try to improve the results.

In recent years there has been a great increase in the number of ML algorithms and computational power that allows the development of highly accurate models. However, the performance of toxicity prediction models remains mainly limited by the quality and availability of the data. Experimental data contain inherent noise and errors that lower the upper limit for the predictive performance, as models can only get as good as the underlying data.[95] Moreover, sufficiently large data sets for model training may often only be achieved by merging data from several sources. This approach may introduce noise in the data, as experimental results obtained in different laboratories or with different protocols often present inconsistencies. Also merging data sets from publications derived with different data curation workflows may reduce the consistency and hence the quality of the data.

In order to improve the acceptance of in silico toxicity models in early discovery stages and regulatory applications, both the data quality and the chemical space covered by the models need to be expanded to allow for a wider AD and more confident predictions. For that matter, governmental initiatives like the ToxCast project need to be further developed to support the increase of high-quality toxicity data. Special emphasis should be set on screening putative active compounds to increase the representation of toxic substances in the data sets, as these are often the limiting factor for training well-generalizing and sensitive predictive models. Besides these governmental projects, a lot of effort is put on the development of federated learning methods that allow training ML models on confidential data from several companies without the need to pool or reveal this information.[96, 97] Such approaches enable the exploitation of data otherwise only accessible to the individual companies, as well as the increase of the overall amount and coverage of available training data. To this end, projects like MELLODDY,[98] a consortium of several industrial and public organizations, may be key for pushing the development of ML models with greater generalization capacity that boost the utility of in silico approaches in substance development pipelines. However, one of the biggest challenges that federated learning approaches are still facing is guaranteeing the safety of the intellectual property of industrial partners, while still having access to useful data for the development and interpretation of the models.[99]

As the amount of toxicity data increases over time, e.g. via more published data or federated approaches, more complex, big-data-oriented ML algorithms like deep learning methods are becoming more powerful tools for enhancing the performance of computational toxicology models. However, the increased performance of these models comes at the cost of a low interpretability of the predictions.[17] Not having a mechanistic understanding behind the prediction not only limits the acceptance of the models by the users, but also their usefulness for guiding the redesign of compounds in the desired direction. As deep learning algorithms become more frequent in many research areas, the field of explainable artificial intelligence has also gained great relevance.[100] Explainable artificial intelligence allows the understanding of how the different input parameters are influencing the prediction and can hence help to guide research in the desired direction.[101, 102] This information may also be valuable to confirm that models are learning meaningful relationships in the data and not just noise or bias. The implementation of these explainability methods in toxicity prediction models can also promote their acceptance in regulatory applications and may be mandatory in the future, as the OECD principles for QSAR models already require mechanistic interpretation whenever possible.[94]

When enough data is still not available for training robust deep learning models, transfer learning methods offer the opportunity to exploit the benefits of these complex architectures and at the same time reuse information learned on related topics with abundant data.[103] A common transfer learning approach consists of first training a model on a task and using it as a starting point (as a whole or parts of it) for training and fine-tuning a model on a second task. Also multi-task models, where several tasks are learned at the same time (as described in section 4.1.), are commonly applied in chemical property prediction applications to exploit the benefits of transfer learning.[104, 105] Besides the information exchange between related tasks, transfer learning approaches also present further advantages, such as shorter training times (especially important for big data applications) and better generalization by reducing overfitting.

All these advancements in data sciences and the increase of available toxicological data pose in silico toxicology as a powerful tool for the development of robust IATA in the following years. The determination of in vivo effects is a highly challenging problem that requires safety assessment strategies approaching it from different perspectives to achieve robust and reliable results. Here, we aim to cover some of these aspects with the development of computational tools that, for instance, predict MIEs for the establishment of AOPs, or improve the molecular similarity definition for read-across applications using bioactivity descriptors.

Bibliography

1. Martyniuk, C.; Mehinto, A. C.; Denslow, N., Organochlorine Pesticides: Agrochemicals With Potent Endocrine-Disrupting Properties in Fish. *Mol. Cell. Endocrinol.* **2020,** *507*, 110764.
2. Santarossa, M.; Coleone de Carvalho, A.; Paganoti de Mello, N.; Ignácio, N.; Machado, A.; Silva, J.; Velini, E.; Machado-Neto, J., Contamination of Fee-Fishing Ponds With Agrochemicals Used in Sugarcane Crops. *SN Appl. Sci.* **2020,** *2*, 1498.
3. Golomb, B. A.; Evans, M. A., Statin Adverse Effects: A Review of the Literature and Evidence for a Mitochondrial Mechanism. *Am. J. Cardiovasc. Drugs* **2008,** *8*, 373-418.
4. Kalyaanamoorthy, S.; Barakat, K. H., Development of Safe Drugs: The hERG Challenge. *Med. Res. Rev.* **2018,** *38*, 525-555.
5. Maggiora, G.; Vogt, M.; Stumpfe, D.; Bajorath, J., Molecular Similarity in Medicinal Chemistry. *J. Med. Chem.* **2014,** *57*, 3186-3204.
6. Mellor, C. L.; Marchese Robinson, R. L.; Benigni, R.; Ebbrell, D.; Enoch, S. J.; Firman, J. W.; Madden, J. C.; Pawar, G.; Yang, C.; Cronin, M. T. D., Molecular Fingerprint-Derived Similarity Measures for Toxicological Read-Across: Recommendations for Optimal Use. *Regul. Toxicol. Pharmacol.* **2019,** *101*, 121-134.
7. Van Norman, G. A., Limitations of Animal Studies for Predicting Toxicity in Clinical Trials: Is it Time to Rethink our Current Approach? *JACC Basic Transl. Sci.* **2019,** *4*, 845-854.
8. Van Norman, G. A., Limitations of Animal Studies for Predicting Toxicity in Clinical Trials: Part 2: Potential Alternatives to the Use of Animals in Preclinical Trials. *JACC Basic Transl. Sci.* **2020,** *5*, 387-397.
9. Vinken, M.; Benfenati, E.; Busquet, F.; Castell, J.; Clevert, D.-A.; de Kok, T. M.; Dirven, H.; Fritsche, E.; Geris, L.; Gozalbes, R.; Hartung, T.; Jennen, D.; Jover, R.; Kandarova, H.; Kramer, N.; Krul, C.; Luechtefeld, T.; Masereeuw, R.; Roggen, E.; Schaller, S.; Vanhaecke, T.; Yang, C.; Piersma, A. H., Safer Chemicals Using Less Animals: Kick-off of the European ONTOX Project. *Toxicology* **2021,** *458*, 152846.
10. Toutain, P.-L.; Ferran, A.; Bousquet-Mélou, A., Species Differences in Pharmacokinetics and Pharmacodynamics. In *Comparative and Veterinary Pharmacology*, Cunningham, F.; Elliott, J.; Lees, P., Eds. Springer Berlin Heidelberg: Berlin, Heidelberg, 2010; pp 19-48.

11. Martignoni, M.; Groothuis, G. M. M.; de Kanter, R., Species Differences Between Mouse, Rat, Dog, Monkey and Human CYP-Mediated Drug Metabolism, Inhibition and Induction. *Expert Opin. Drug Metab. Toxicol.* **2006,** *2*, 875-894.
12. Perel, P.; Roberts, I.; Sena, E.; Wheble, P.; Briscoe, C.; Sandercock, P.; Macleod, M.; Mignini, L. E.; Jayaram, P.; Khan, K. S., Comparison of Treatment Effects Between Animal Experiments and Clinical Trials: Systematic Review. *BMJ* **2007,** *334*, 197.
13. Dent, M. P.; Vaillancourt, E.; Thomas, R. S.; Carmichael, P. L.; Ouedraogo, G.; Kojima, H.; Barroso, J.; Ansell, J.; Barton-Maclaren, T. S.; Bennekou, S. H.; Boekelheide, K.; Ezendam, J.; Field, J.; Fitzpatrick, S.; Hatao, M.; Kreiling, R.; Lorencini, M.; Mahony, C.; Montemayor, B.; Mazaro-Costa, R.; Oliveira, J.; Rogiers, V.; Smegal, D.; Taalman, R.; Tokura, Y.; Verma, R.; Willett, C.; Yang, C., Paving the Way for Application of Next Generation Risk Assessment to Safety Decision-Making for Cosmetic Ingredients. *Regul. Toxicol. Pharmacol.* **2021,** *125*, 105026.
14. Richard, A. M.; Judson, R. S.; Houck, K. A.; Grulke, C. M.; Volarath, P.; Thillainadarajah, I.; Yang, C.; Rathman, J.; Martin, M. T.; Wambaugh, J. F.; Knudsen, T. B.; Kancherla, J.; Mansouri, K.; Patlewicz, G.; Williams, A. J.; Little, S. B.; Crofton, K. M.; Thomas, R. S., ToxCast Chemical Landscape: Paving the Road to 21st Century Toxicology. *Chem. Res. Toxicol.* **2016,** *29*, 1225-51.
15. *New Scoping Document on in Vitro and ex Vivo Assays for the Identification of Modulators of Thyroid Hormone Signalling*. Paris: OECD Publishing: 2017.
16. OECD, *Guideline No. 497: Defined Approaches on Skin Sensitisation*. 2021.
17. Polishchuk, P., Interpretation of Quantitative Structure–Activity Relationship Models: Past, Present, and Future. *J. Chem. Inf. Model.* **2017,** *57*, 2618-2639.
18. Hansch, C.; Maloney, P. P.; Fujita, T.; Muir, R. M., Correlation of Biological Activity of Phenoxyacetic Acids with Hammett Substituent Constants and Partition Coefficients. *Nature* **1962,** *194*, 178-180.
19. Roy, K.; Das, R. N.; Popelier, P. L. A., Predictive QSAR Modelling of Algal Toxicity of Ionic Liquids and its Interspecies Correlation with Daphnia Toxicity. *Environ. Sci. Pollut. Res.* **2015,** *22*, 6634-6641.
20. Naseem, S.; Zushi, Y.; Nabi, D., Development and Evaluation of Two-Parameter Linear Free Energy Models for the Prediction of Human Skin Permeability Coefficient of Neutral Organic Chemicals. *J. Cheminformatics* **2021,** *13*, 25.

21. Fengxian, C.; Reti, H., Analysis of Positions and Substituents on Genotoxicity of Fluoroquinolones With Quantitative Structure-Activity Relationship and 3D Pharmacophore Model. *Ecotoxicol. Environ. Saf.* **2017,** *136*, 111-118.
22. Tong, L.; Guo, L.; Lv, X.; Li, Y., Modification of Polychlorinated Phenols and Evaluation of Their Toxicity, Biodegradation and Bioconcentration Using Three-Dimensional Quantitative Structure–Activity Relationship Models. *J. Mol. Graph. Model.* **2017,** *71*, 1-12.
23. Pearson, K., LIII. On Lines and Planes of Closest Fit to Systems of Points in Space. *Lond. Edinb. Dublin Philos. Mag. J. Sci.* **1901,** *2*, 559-572.
24. McInnes, L.; Healy, J.; Melville, J., Umap: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv preprint* **2018**, arXiv:1802.03426.
25. Harrell Jr., F. E., *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*. Springer: 2001.
26. Cristianini, N.; Ricci, E., Support Vector Machines. In *Encyclopedia of Algorithms*, Kao, M.-Y., Ed. Springer US: Boston, MA, 2008; pp 928-932.
27. Ho, T. K., Random Decision Forests. In *Proceedings of the Third International Conference on Document Analysis and Recognition (Volume 1) - Volume 1*, IEEE Computer Society: 1995; p 278.
28. Friedman, J. H., Greedy Function Approximation: A Gradient Boosting Machine. *Ann. Stat.* **2001,** *29*, 1189-1232.
29. Aggarwal, C. C., *Neural Networks and Deep Learning.* Springer: 2018.
30. He, H.; Garcia, E. A., Learning from Imbalanced Data. *IEEE Trans. Knowl. Data Eng.* **2009,** *21*, 1263-1284.
31. Chawla, N. V.; Bowyer, K.; Hall, L. O.; Kegelmeyer, P. O., SMOTE: Synthetic Minority Over-Sampling Technique. **2002,** *16*, 321-357.
32. Huang, R.; Xia, M., Editorial: Tox21 Challenge to Build Predictive Models of Nuclear Receptor and Stress Response Pathways as Mediated by Exposure to Environmental Toxicants and Drugs. *Front. Environ. Sci.* **2017,** *5*.
33. Huang, R.; Xia, M.; Nguyen, D.-T.; Zhao, T.; Sakamuru, S.; Zhao, J.; Shahane, S. A.; Rossoshek, A.; Simeonov, A., Tox21Challenge to Build Predictive Models of Nuclear Receptor and Stress Response Pathways as Mediated by Exposure to Environmental Chemicals and Drugs. *Front. Environ. Sci.* **2016,** *3*.
34. Karelson, M.; Lobanov, V. S.; Katritzky, A. R., Quantum-Chemical Descriptors in QSAR/QSPR Studies. *Chem. Rev.* **1996,** *96*, 1027-1044.

35. Landrum, G., RDKit: Open-Source Cheminformatics Software. **2016**.

36. Mauri, A., alvaDesc: A Tool to Calculate and Analyze Molecular Descriptors and Fingerprints. In *Ecotoxicological QSARs*, Roy, K., Ed. Springer US: New York, NY, 2020; pp 801-820.

37. *Molecular Operating Environment (MOE)*, 2019.01; Chemical Computing Group ULC: 1010 Sherbooke St. West, Suite #910, Montreal, QC, Canada, H3A 2R7, 2021.

38. Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G., Reoptimization of MDL Keys for Use in Drug Discovery. *J. Chem. Inf. Comput. Sci.* **2002,** *42*, 1273-1280.

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

40. Ashby, J.; Tennant, R. W., Chemical Structure, Salmonella mutagenicity and Extent of Carcinogenicity as Indicators of Genotoxic Carcinogenesis Among 222 Chemicals Tested in Rodents by the U.S. NCI/NTP. *Mutat. Res.* **1988,** *204*, 17-115.

41. Ferrari, T.; Cattaneo, D.; Gini, G.; Golbamaki Bakhtyari, N.; Manganaro, A.; Benfenati, E., Automatic Knowledge Extraction From Chemical Structures: The Case of Mutagenicity Prediction. *SAR QSAR Environ. Res.* **2013,** *24*, 365-83.

42. Ahlberg, E.; Carlsson, L.; Boyer, S., Computational Derivation of Structural Alerts From Large Toxicology Data Sets. *J. Chem. Inf. Model.* **2014,** *54*, 2945-52.

43. Alves, V. M.; Muratov, E. N.; Capuzzi, S. J.; Politi, R.; Low, Y.; Braga, R. C.; Zakharov, A. V.; Sedykh, A.; Mokshyna, E.; Farag, S.; Andrade, C. H.; Kuz'min, V. E.; Fourches, D.; Tropsha, A., Alarms About Structural Alerts. *Green Chem.* **2016,** *18*, 4348-4360.

44. Kearnes, S.; McCloskey, K.; Berndl, M.; Pande, V.; Riley, P., Molecular Graph Convolutions: Moving Beyond Fingerprints. *J. Comput. Aided Mol. Des.* **2016,** *30*, 595-608.

45. Coley, C. W.; Barzilay, R.; Green, W. H.; Jaakkola, T. S.; Jensen, K. F., Convolutional Embedding of Attributed Molecular Graphs for Physical Property Prediction. *J. Chem. Inf. Model.* **2017,** *57*, 1757-1772.

46. Sedykh, A.; Zhu, H.; Tang, H.; Zhang, L.; Richard, A.; Rusyn, I.; Tropsha, A., Use of in Vitro HTS-Derived Concentration-Response Data as Biological Descriptors Improves the Accuracy of QSAR Models of in Vivo Toxicity. *Environ. Health Perspect.* **2011,** *119*, 364-370.

47. Cortés-Ciriano, I.; Škuta, C.; Bender, A.; Svozil, D., QSAR-Derived Affinity Fingerprints (Part 2): Modeling Performance for Potency Prediction. *J. Cheminform.* **2020,** *12*, 41.

48. Sturm, N.; Sun, J.; Vandriessche, Y.; Mayr, A.; Klambauer, G.; Carlsson, L.; Engkvist, O.; Chen, H., Application of Bioactivity Profile-Based Fingerprints for Building Machine Learning Models. *J. Chem. Inf. Model.* **2019,** *59*, 962-972.
49. Allen, C. H. G.; Mervin, L. H.; Mahmoud, S. Y.; Bender, A., Leveraging Heterogeneous Data From GHS Toxicity Annotations, Molecular and Protein Target Descriptors and Tox21 Assay Readouts to Predict and Rationalise Acute Toxicity. *J. Cheminform.* **2019,** *11*, 36.
50. Morger, A.; Svensson, F.; Arvidsson McShane, S.; Gauraha, N.; Norinder, U.; Spjuth, O.; Volkamer, A., Assessing the Calibration in Toxicological in Vitro Models With Conformal Prediction. *J. Cheminform.* **2021,** *13*, 35.
51. Mathea, M.; Klingspohn, W.; Baumann, K., Chemoinformatic Classification Methods and Their Applicability Domain. *Mol. Inf.* **2016,** *35*, 160-180.
52. Klingspohn, W.; Mathea, M.; ter Laak, A.; Heinrich, N.; Baumann, K., Efficiency of Different Measures for Defining the Applicability Domain of Classification Models. *J. Cheminform.* **2017,** *9*, 44.
53. Vovk, V., Conditional Validity of Inductive Conformal Predictors. *Mach. Learn.* **2013,** *92*, 349-376.
54. Wilm, A.; Norinder, U.; Agea, M. I.; de Bruyn Kops, C.; Stork, C.; Kühnl, J.; Kirchmair, J., Skin Doctor CP: Conformal Prediction of the Skin Sensitization Potential of Small Organic Molecules. *Chem. Res. Toxicol.* **2021,** *34*, 330-344.
55. Morger, A.; Mathea, M.; Achenbach, J. H.; Wolf, A.; Buesen, R.; Schleifer, K.-J.; Landsiedel, R.; Volkamer, A., KnowTox: Pipeline and Case Study for Confident Prediction of Potential Toxic Effects of Compounds in Early Phases of Development. *J. Cheminform.* **2020,** *12*, 24.
56. Zhang, J.; Norinder, U.; Svensson, F., Deep Learning-Based Conformal Prediction of Toxicity. *J. Chem. Inf. Model.* **2021,** *61*, 2648-2657.
57. Norinder, U.; Carlsson, L.; Boyer, S.; Eklund, M., Introducing Conformal Prediction in Predictive Modeling. A Transparent and Flexible Alternative to Applicability Domain Determination. *J. Chem. Inf. Model.* **2014,** *54*, 1596-1603.
58. Carlsson, L.; Eklund, M.; Norinder, U. In *Aggregated Conformal Prediction*, Artificial Intelligence Applications and Innovations. AIAI 2014. IFIP Advances in Information and Communication Technology, v., Ed. Springer, Berlin, Heidelberg: 2014; pp 231-240.
59. Gauraha, N.; Spjuth, O., Synergy Conformal Prediction. In *Proceedings of the Tenth Symposium on Conformal and Probabilistic Prediction and Applications*, Lars, C.;

Zhiyuan, L.; Giovanni, C.; Khuong An, N., Eds. PMLR: Proceedings of Machine Learning Research, 2021; Vol. 152, pp 91--110.

60. Norinder, U.; Spjuth, O.; Svensson, F., Synergy Conformal Prediction Applied to Large-Scale Bioactivity Datasets and in Federated Learning. *J. Cheminform.* **2021,** *13*, 77.
61. Norinder, U.; Boyer, S., Binary Classification of Imbalanced Datasets Using Conformal Prediction. *J. Mol. Graph. Model.* **2017,** *72*, 256-265.
62. Dix, D. J.; Houck, K. A.; Martin, M. T.; Richard, A. M.; Setzer, R. W.; Kavlock, R. J., The ToxCast Program for Prioritizing Toxicity Testing of Environmental Chemicals. *Toxicol. Sci.* **2007,** *95*, 5-12.
63. Ji, C.; Svensson, F.; Zoufir, A.; Bender, A., eMolTox: Prediction of Molecular Toxicity With Confidence. *Bioinformatics (Oxford, England)* **2018,** *34*, 2508-2509.
64. eChemPortal. https://www.echemportal.org/echemportal/ (accessed August 6, 2020).
65. Berthold, M. R.; Cebron, N.; Dill, F.; Gabriel, T. R.; Kötter, T.; Meinl, T.; Ohl, P.; Sieb, C.; Thiel, K.; Wiswedel, B., KNIME: The Konstanz Information Miner. In *Studies in Classification, Data Analysis, and Knowledge Organization (GfKL 2007)*, Springer: 2007, (Version 4.3.3.).
66. Chollet, F. Keras. https://keras.io.
67. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, É., Scikit-learn: Machine Learning in Python. **2011,** *12*, 2825-2830.
68. Andersson, N.; Arena, M.; Auteri, D.; Barmaz, S.; Grignard, E.; Kienzler, A.; Lepper, P.; Lostia, A. M.; Munn, S., Guidance for the Identification of Endocrine Disruptors in the Context of Regulations (EU) No 528/2012 and (EC) No 1107/2009. *EFSA J.* **2018,** *16*, e05311.
69. OECD, Revised Guidance Document 150 on Standardised Test Guidelines for Evaluating Chemicals for Endocrine Disruption. OECD Publishing, Paris: OECD Series on Testing and Assessment, No. 150, 2018.
70. Schneider, M.; Pons, J.-L.; Labesse, G.; Bourguet, W., In Silico Predictions of Endocrine Disruptors Properties. *Endocrinology* **2019,** *160*, 2709-2716.
71. Balabin, I. A.; Judson, R. S., Exploring Non-Linear Distance Metrics in the Structure–Activity Space: QSAR Models for Human Estrogen Receptor. *J. Cheminform.* **2018,** *10*, 47.

72. Zhang, Q.; Yan, L.; Wu, Y.; Ji, L.; Chen, Y.; Zhao, M.; Dong, X., A Ternary Classification Using Machine Learning Methods of Distinct Estrogen Receptor Activities Within a Large Collection of Environmental Chemicals. *Sci. Total Environ.* **2017,** *580*, 1268-1275.

73. Grisoni, F.; Consonni, V.; Ballabio, D., Machine Learning Consensus to Predict the Binding to the Androgen Receptor within the CoMPARA Project. *J. Chem. Inf. Model.* **2019,** *59*, 1839-1848.

74. Yang, X.; Liu, H.; Yang, Q.; Liu, J.; Chen, J.; Shi, L., Predicting Anti-Androgenic Activity of Bisphenols Using Molecular Docking and Quantitative Structure-Activity Relationships. *Chemosphere* **2016,** *163*, 373-381.

75. Rosenberg, S. A.; Watt, E. D.; Judson, R. S.; Simmons, S. O.; Friedman, K. P.; Dybdahl, M.; Nikolov, N. G.; Wedebye, E. B., QSAR Models for Thyroperoxidase Inhibition and Screening of U.S. and EU Chemical Inventories. *Comput. Toxicol.* **2017,** *4*, 11-21.

76. Politi, R.; Rusyn, I.; Tropsha, A., Prediction of Binding Affinity and Efficacy of Thyroid Hormone Receptor Ligands Using QSAR and Structure Based Modeling Methods. *Toxicol. Appl. Pharmacol.* **2014,** *280*, 177-189.

77. Azimi, G.; Afiuni-Zadeh, S.; Karami, A., A QSAR Study for Modeling of Thyroid Receptors β1 Selective Ligands by Application of Adaptive Neuro-Fuzzy Inference System and Radial Basis Function. *J. Chemometrics* **2012,** *26*, 135-142.

78. Hartung, T., Making Big Sense From Big Data in Toxicology by Read-Across. *Altex* **2016,** *33*, 83-93.

79. Zhu, H.; Bouhifd, M.; Donley, E.; Egnash, L.; Kleinstreuer, N.; Kroese, E. D.; Liu, Z.; Luechtefeld, T.; Palmer, J.; Pamies, D.; Shen, J.; Strauss, V.; Wu, S.; Hartung, T., Supporting Read-Across Using Biological Data. *Altex* **2016,** *33*, 167-82.

80. Guo, Y.; Zhao, L.; Zhang, X.; Zhu, H., Using a Hybrid Read-Across Method to Evaluate Chemical Toxicity Based on Chemical Structure and Biological Data. *Ecotoxicol. Environ. Saf.* **2019,** *178*, 178-187.

81. Arvidsson McShane, S.; Ahlberg, E.; Noeske, T.; Spjuth, O., Machine Learning Strategies When Transitioning between Biological Assays. *J. Chem. Inf. Model.* **2021,** *61*, 3722-3733.

82. Gaulton, A.; Hersey, A.; Nowotka, M.; Bento, A. P.; Chambers, J.; Mendez, D.; Mutowo, P.; Atkinson, F.; Bellis, L. J.; Cibrián-Uhalte, E.; Davies, M.; Dedman, N.; Karlsson, A.; Magariños, M. P.; Overington, J. P.; Papadatos, G.; Smit, I.; Leach, A. R., The ChEMBL Database in 2017. *Nucleic Acids Res.* **2016,** *45*, D945-D954.

83. Amacher, D. E., Reactive Intermediates and the Pathogenesis of Adverse Drug Reactions: The Toxicology Perspective. *Curr. Drug Metab.* **2006,** *7*, 219-29.

84. Williams, P. D.; Kitteringham, R. N.; Naisbitt, J. D.; Pirmohamed, M.; Smith, A. D.; Park, K. B., Are Chemically Reactive Metabolites Responsible for Adverse Reactions to Drugs? *Curr. Drug Metab.* **2002,** *3*, 351-366.

85. Leung, L.; Kalgutkar, A. S.; Obach, R. S., Metabolic Activation in Drug-Induced Liver Injury. *Drug Metab. Rev.* **2012,** *44*, 18-33.

86. Kalgutkar, A. S.; Vaz, A. D. N.; Lame, M. E.; Henne, K. R.; Soglia, J.; Zhao, S. X.; Abramov, Y. A.; Lombardo, F.; Collin, C.; Hendsch, Z. S.; Hop, C. E. C. A., Bioactivation of the Nontricyclic Antidepressant Nefazodone to a Reactive Quinone-Imine Species in Human Liver Microsomes and Recombinant Cytochrome P450 3A4. *Drug Metab. Dispos.* **2005,** *33*, 243.

87. Dmitriev, A.; Rudik, A.; Filimonov, D.; Lagunin, A.; Pogodin, P.; Dubovskaja, V.; Bezhentsev, V.; Ivanov, S.; Druzhilovsky, D.; Tarasova, O.; Poroikov, V., Integral Estimation of Xenobiotics' Toxicity With Regard to Their Metabolism in Human Organism. *Pure Appl. Chem.* **2017,** *89*, 1449-1458.

88. Filimonov, D. A.; Rudik, A. V.; Dmitriev, A. V.; Poroikov, V. V., Computer-Aided Estimation of Biological Activity Profiles of Drug-Like Compounds Taking Into Account Their Metabolism in Human Body. *Int. J. Mol. Sci.* **2020,** *21*, 7492.

89. Mekenyan, O.; Dimitrov, S.; Serafimova, R.; Thompson, E.; Kotov, S.; Dimitrova, N.; Walker, J. D., Identification of the Structural Requirements for Mutagenicity by Incorporating Molecular Flexibility and Metabolic Activation of Chemicals I: TA100 Model. *Chem. Res. Toxicol.* **2004,** *17*, 753-766.

90. Mekenyan, O. G.; Petkov, P. I.; Kotov, S. V.; Stoeva, S.; Kamenska, V. B.; Dimitrov, S. D.; Honma, M.; Hayashi, M.; Benigni, R.; Donner, E. M.; Patlewicz, G., Investigating the Relationship between in Vitro–in Vivo Genotoxicity: Derivation of Mechanistic QSAR Models for in Vivo Liver Genotoxicity and in Vivo Bone Marrow Micronucleus Formation Which Encompass Metabolism. *Chem. Res. Toxicol.* **2012,** *25*, 277-296.

91. Dimitrov, S. D.; Low, L. K.; Patlewicz, G. Y.; Kern, P. S.; Dimitrova, G. D.; Comber, M. H. I.; Phillips, R. D.; Niemela, J.; Bailey, P. T.; Mekenyan, O. G., Skin Sensitization: Modeling Based on Skin Metabolism Simulation and Formation of Protein Conjugates. *Int. J. Toxicol.* **2005,** *24*, 189-204.

92. Mekenyan, O.; Patlewicz, G.; Kuseva, C.; Popova, I.; Mehmed, A.; Kotov, S.; Zhechev, T.; Pavlov, T.; Temelkov, S.; Roberts, D. W., A Mechanistic Approach to Modeling Respiratory Sensitization. *Chem. Res. Toxicol.* **2014,** *27*, 219-239.

93. OECD Integrated Approaches to Testing and Assessment (IATA). https://www.oecd.org/chemicalsafety/risk-assessment/iata-integrated-approaches-to-testing-and-assessment.htm (accessed November 23, 2021).

94. OECD OECD Principles for the Validation, for Regulatory Purposes, of (Quantitative) Structure-Activity Relationship Models. https://www.oecd.org/chemicalsafety/risk-assessment/37849783.pdf (accessed November 23, 2021).

95. Kramer, C.; Kalliokoski, T.; Gedeck, P.; Vulpetti, A., The Experimental Uncertainty of Heterogeneous Public K(i) Data. *J. Med. Chem.* **2012,** *55*, 5165-73.

96. Gedeck, P.; Skolnik, S.; Rodde, S., Developing Collaborative QSAR Models Without Sharing Structures. *J. Chem. Inf. Model.* **2017,** *57*, 1847-1858.

97. Martin, E. J.; Zhu, X.-W., Collaborative Profile-QSAR: A Natural Platform for Building Collaborative Models among Competing Companies. *J. Chem. Inf. Model.* **2021,** *61*, 1603-1616.

98. Machine Learning Ledger Orchestration for Drug Discovery (MELLODDY). https://www.melloddy.eu (accessed November 23, 2021).

99. Pejó, B.; Biczók, G., Quality Inference in Federated Learning With Secure Aggregation. *arXiv preprint arXiv:2007.06236* **2020**.

100. Angelov, P. P.; Soares, E. A.; Jiang, R.; Arnold, N. I.; Atkinson, P. M., Explainable Artificial Intelligence: An Analytical Review. *WIREs Data Min. and Knowl.* **2021,** *11*, e1424.

101. Wenzel, J.; Matter, H.; Schmidt, F., Predictive Multitask Deep Neural Network Models for ADME-Tox Properties: Learning From Large Data Sets. *J. Chem. Inf. Model.* **2019,** *59*, 1253-1268.

102. Henderson, R.; Clevert, D.-A.; Montanari, F., Improving Molecular Graph Neural Network Explainability With Orthonormalization and Induced Sparsity. *arXiv preprint arXiv:2105.04854* **2021**.

103. Weiss, K.; Khoshgoftaar, T. M.; Wang, D., A Survey of Transfer Learning. *J. Big Data* **2016,** *3*, 9.

104. Sadawi, N.; Olier, I.; Vanschoren, J.; van Rijn, J. N.; Besnard, J.; Bickerton, R.; Grosan, C.; Soldatova, L.; King, R. D., Multi-Task Learning With a Natural Metric for Quantitative Structure Activity Relationship Learning. *J. Cheminform.* **2019,** *11*, 68.

105. Sosnin, S.; Vashurina, M.; Withnall, M.; Karpov, P.; Fedorov, M.; Tetko, I. V., A Survey of Multi-task Learning Methods in Chemoinformatics. *Mol. Inf.* **2019,** *38*, 1800108.

Bibliography of this dissertation's publications

[P1] Garcia de Lomana, M.; Weber, A. G.; Birk, B.; Landsiedel, R.; Achenbach, J.; Schleifer, K. J.; Mathea, M. and Kirchmair, J. In Silico Models to Predict the Perturbation of Molecular Initiating Events Related to Thyroid Hormone Homeostasis, *Chem. Res. Toxicol.*, **2021**, 34, 396–411.

[P2] Garcia de Lomana, M.; Morger, A.; Norinder, U.; Buesen, R.; Landsiedel, R.; Volkamer, A.; Kirchmair, J. and Mathea, M. ChemBioSim: Enhancing Conformal Prediction of In Vivo Toxicity by Use of Predicted Bioactivities, *J. Chem. Inf. Model.*, **2021**, 61, 3255–3272.

[P3] Morger, A.; Garcia de Lomana, M.; Norinder, U.; Svensson, F.; Kirchmair, J.; Mathea, M. and Volkamer, A. Studying and Mitigating the Effects of Data Drifts on ML Model Performance at the Example of Chemical Toxicity Data, *Sci. Rep.*, **2022**, 12, 7244.

[P4] Garcia de Lomana, M.; Svensson, F.; Volkamer, A.; Mathea, M. and Kirchmair, J. Consideration of Predicted Small-Molecule Metabolites in Computational Toxicology, *Digital Discov.*, **2022**, 1, 158-172.

List of abbreviations

ACP	Aggregated conformal prediction
AD	Applicability domain
ADME	Absorption, distribution, metabolism and excretion
AOP	Adverse outcome pathway
AOP-KB	Adverse outcome pathway knowledge base
AUC	Area under the receiver operating curve
CP	Conformal prediction
CV	Cross-validation
DICC	Drug-induced cardiological complications
DILI	Drug-induced liver injury
ECFP	Extended-connectivity fingerprints
EPA	U.S. Environmental Protection Agency
GBT	Gradient boosted trees
GCN	Graph-convolutional networks
hERG	Human ether-a-go-go-related gene
HMG-CoA	β-Hydroxy β-methylglutaryl-CoA
HOMO	Highest occupied molecular orbital
IATA	Integrated approaches to testing and assessment
ICP	Inductive conformal prediction
LLNA	Murine local lymph node assay
LR	Logistic regression
LUMO	Lowest unoccupied molecular orbital
MCC	Matthews correlation coefficient
ML	Machine learning
MNT	Micronucleus test
MOE	Molecular operating environment
NN	Neural networks

OECD	Organisation for Economic Co-operation and Development
PCA	Principal component analysis
QSAR	Quantitative structure-activity relationship
R^2	Coefficient of determination
RF	Random forest
REACH	EU Registration, Evaluation, Authorisation and Restriction of Chemicals
RMSE	Root mean squared error
SCP	Synergy conformal prediction
SVM	Support vector machine
TPO	Thyroid peroxidase
TR	Thyroid hormone receptor
UMAP	Uniform manifold approximation and projection

Appendices

Supporting Information for [P1]

This appendix contains the supporting information for the publication:

Garcia de Lomana, M.; Weber, A. G.; Birk, B.; Landsiedel, R.; Achenbach, J.; Schleifer, K. J.; Mathea, M. and Kirchmair, J. In Silico Models to Predict the Perturbation of Molecular Initiating Events Related to Thyroid Hormone Homeostasis, *Chem. Res. Toxicol.*, **2021**, 34, 396–411.

SUPPORTING INFORMATION

In Silico Models to Predict the Perturbation of Molecular Initiating Events Related to Thyroid Hormone Homeostasis

Marina Garcia de Lomana,[1,2] *Andreas Georg Weber,*[1] *Barbara Birk,*[1] *Robert Landsiedel,*[1] *Janosch Achenbach,*[1⊥] *Klaus-Juergen Schleifer,*[1] *Miriam Mathea*[1]* *and Johannes Kirchmair*[2]*

[1] BASF SE, 67063 Ludwigshafen am Rhein, Germany

[2] Department of Pharmaceutical Chemistry, Faculty of Life Sciences, University of Vienna, 1090 Vienna, Austria

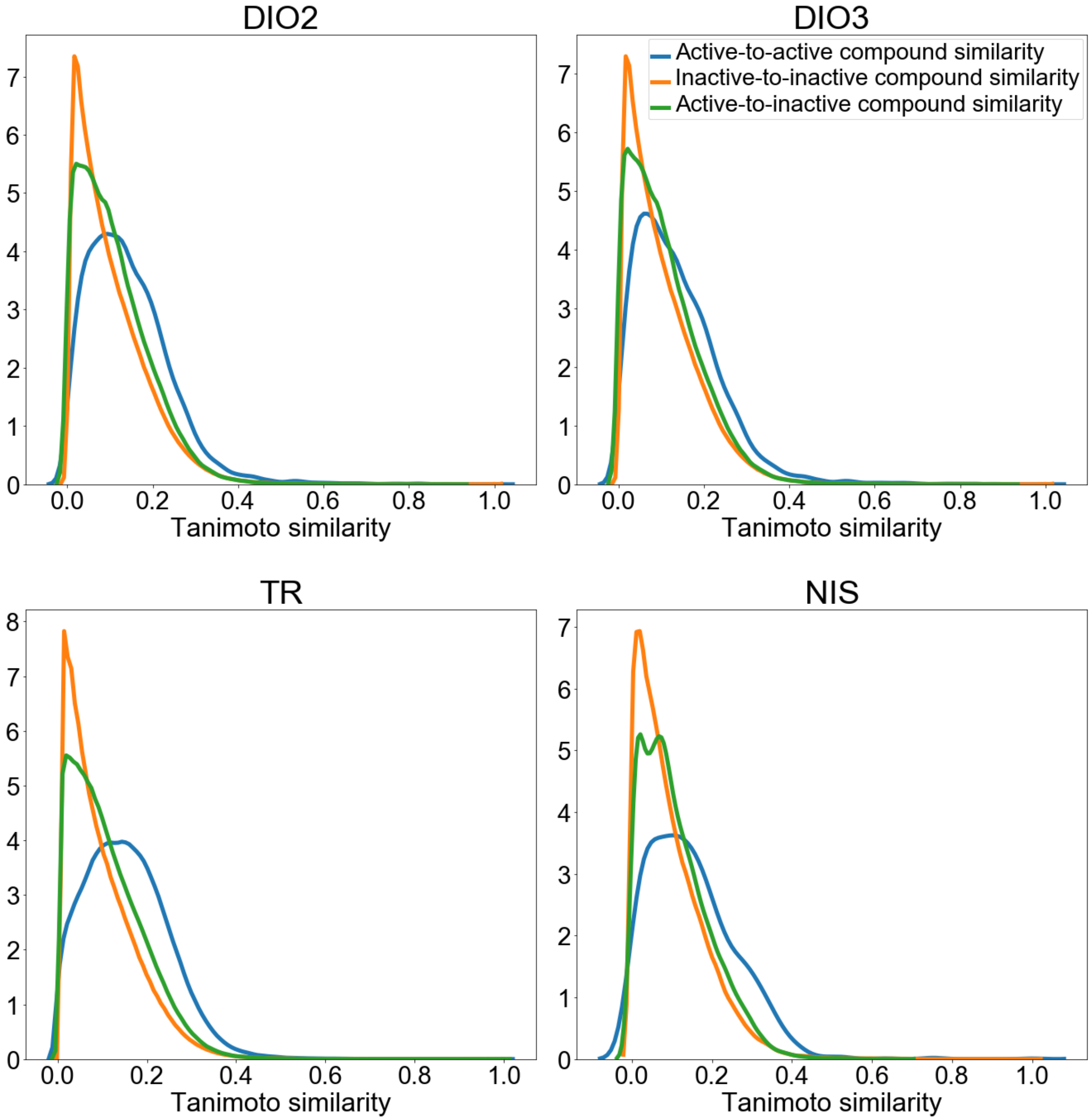
* miriam.mathea@basf.com; Tel.: +49 621 60-29054;
johannes.kirchmair@univie.ac.at; Tel.: +43 1-4277-55104

Table S1. Flags Available in the ToxCast Database for Tagging Potential Errors in Class Labeling.

ToxCast flags
Only one concentration above baseline (active)
Multiple points above baseline (inactive)
Noisy data
Borderline active
Borderline inactive
Gain AC_{50} lower than the lowest concentration and loss AC_{50} lower than the mean concentration
Gain AC_{50} lower than the lowest concentration and loss AC_{50} lower than the mean concentration
Hit-call potentially confounded by overfitting
Biochemical assay with less than 50% efficacy

Table S2. List of Molecular Descriptors Used in Principal Component Analysis.

Descriptors
SlogP
TPSA
ExactMW
NumLipinskiHBA
NumLipinskiHBD
NumRotatableBonds
NumHBD
NumHBA
NumAmideBonds
NumHeteroAtoms
NumHeavyAtoms
NumAtoms
NumStereocenters
NumUnspecifiedStereocenters
NumRings
NumAromaticRings
NumSaturatedRings
NumAliphaticRings
NumAromaticHeterocycles
NumSaturatedHeterocycles
NumAliphaticHeterocycles
NumAromaticCarbocycles
NumSaturatedCarbocycles
NumAliphaticCarbocycles
FractionCSP3


DIO2
DIO3
TR
NIS
Active-to-active compound similarity
Inactive-to-inactive compound similarity
Active-to-inactive compound similarity
Tanimoto similarity
0.0
0.2
0.4
0.6
0.8
1.0
0
1
2
3
4
5
6
7
8

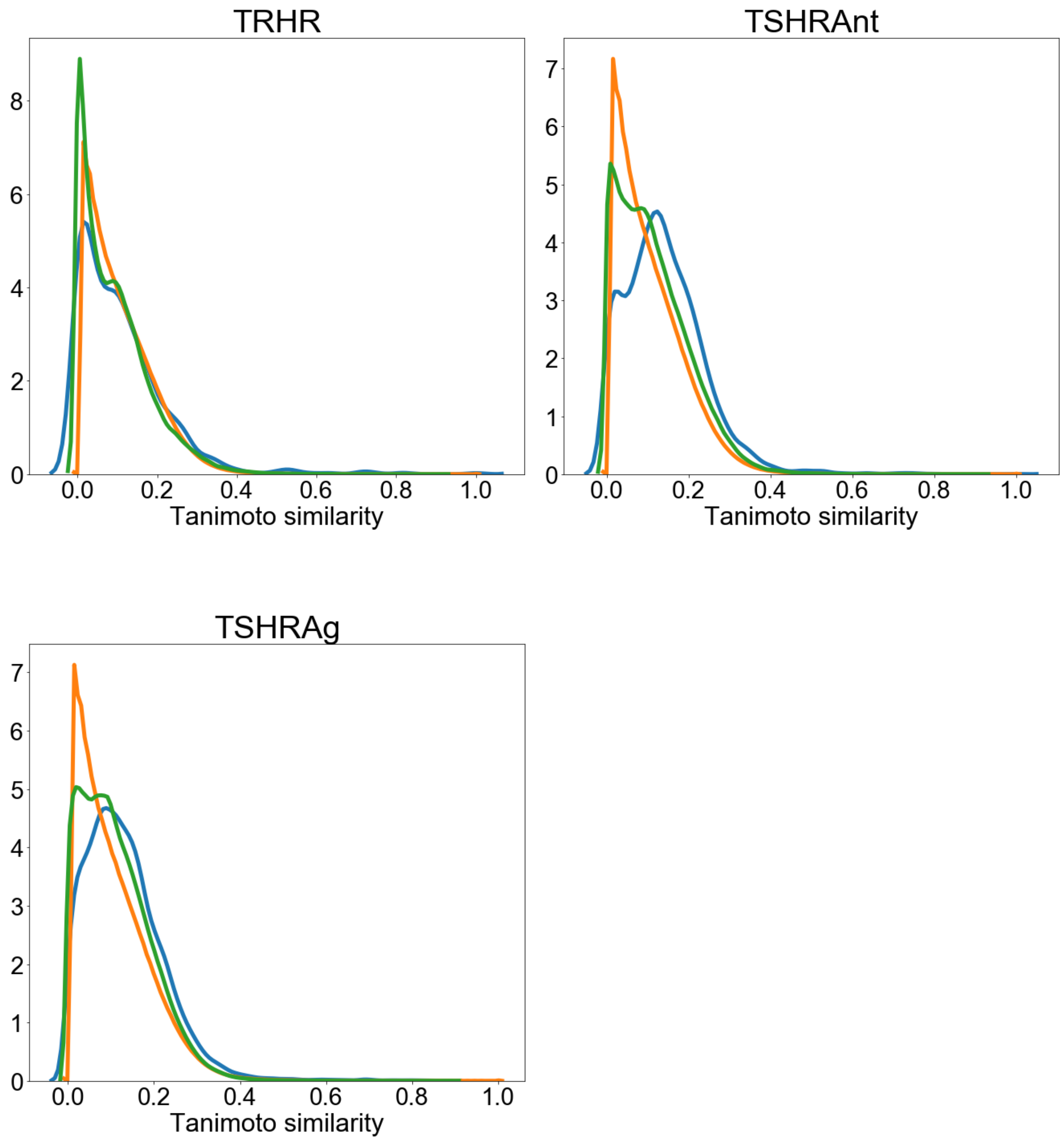
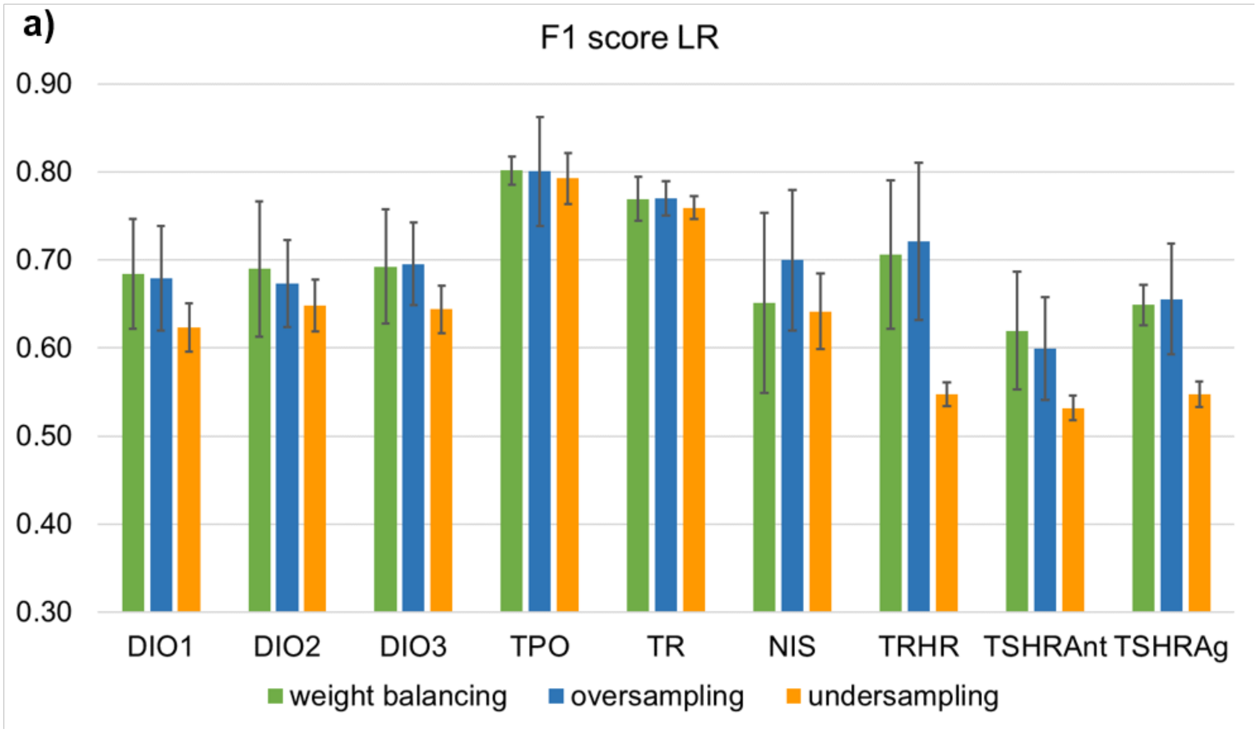
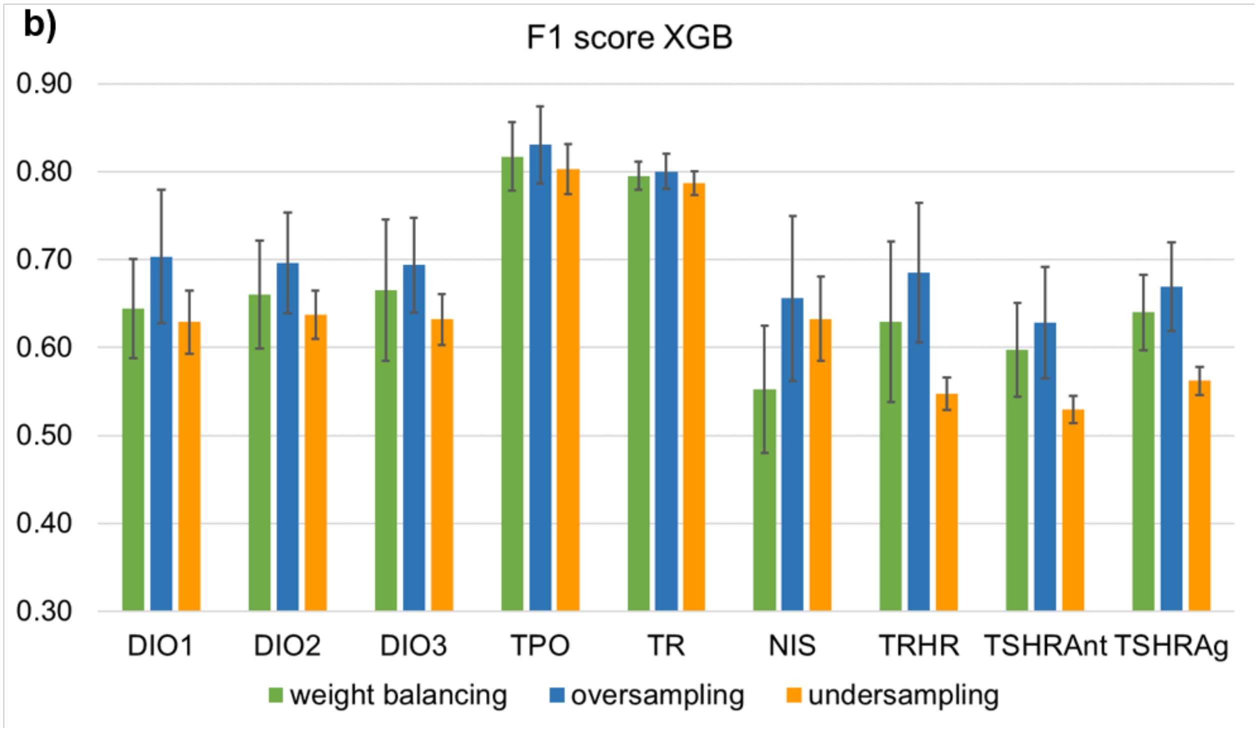


Figure S1. Distribution of pairwise Tanimoto similarities based on atom-pair fingerprints, for DIO2, DIO3, TR, NIS, TRHR, TSHRAg and TSHRAnt and three types of compound pairs: a) active-to-active, b) inactive-to-inactive and c) active-to-inactive.


a)
F1 score LR
0.90
0.80
0.70
0.60
0.50
0.40
0.30
DIO1
DIO2
DIO3
TPO
TR
NIS
TRHR
TSHRAnt
TSHRAg
🟩 weight balancing
🟦 oversampling
🟧 undersampling


b)
F1 score XGB
0.90
0.80
0.70
0.60
0.50
0.40
0.30
DIO1
DIO2
DIO3
TPO
TR
NIS
TRHR
TSHRAnt
TSHRAg
🟩 weight balancing
🟦 oversampling
🟧 undersampling

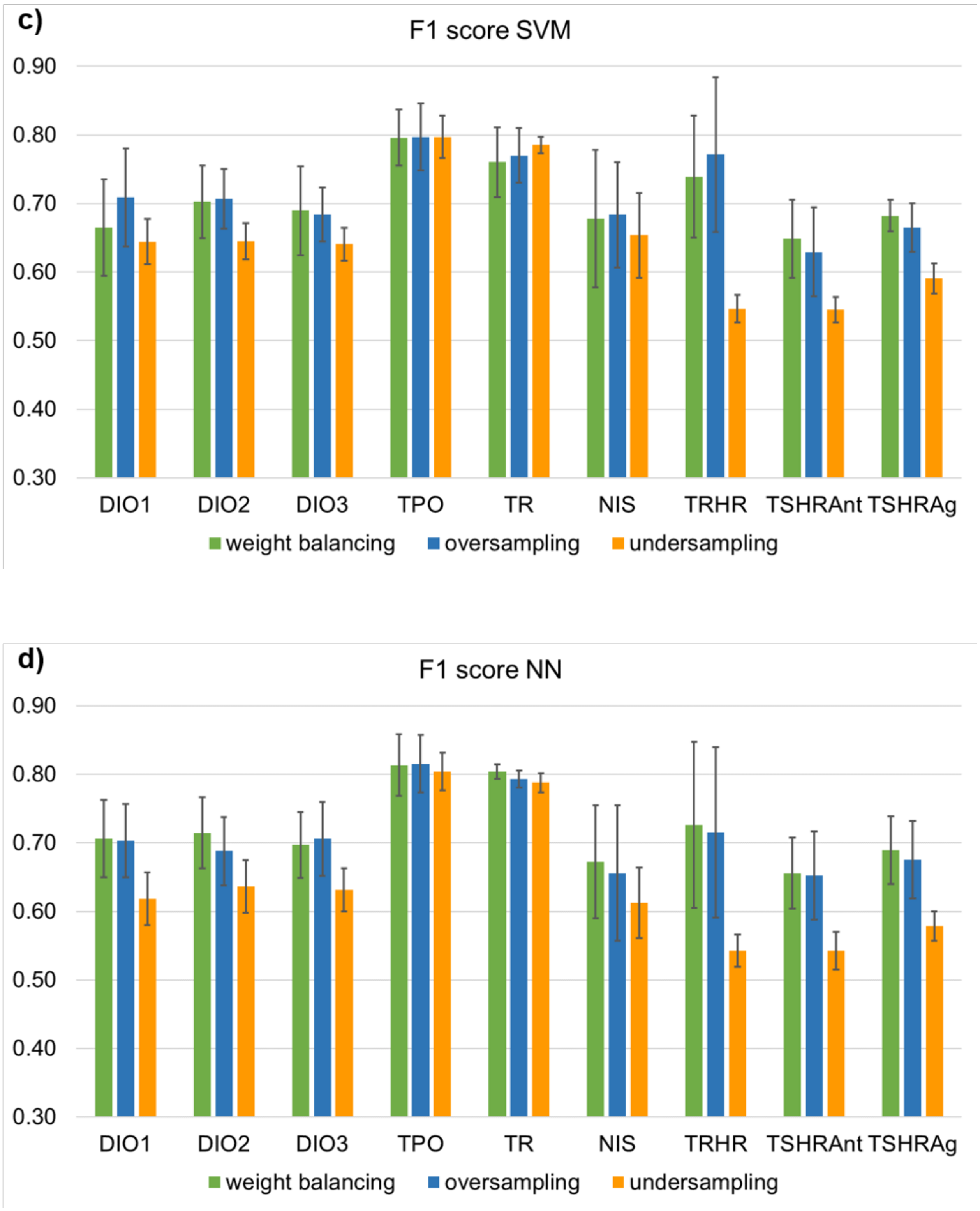


Figure S2. Comparison of the mean F1 score obtained for the nine thyroid end points with (a) LR, (b) XGB, (c) SVM and (d) NN in combination with the three data sampling techniques.

Supporting Information for [P2]

This appendix contains the supporting information for the publication:

Garcia de Lomana, M.; Morger, A.; Norinder, U.; Buesen, R.; Landsiedel, R.; Volkamer, A.; Kirchmair, J. and Mathea, M. ChemBioSim: Enhancing Conformal Prediction of In Vivo Toxicity by Use of Predicted Bioactivities, *J. Chem. Inf. Model.*, **2021**, 61, 3255–3272.

SUPPORTING INFORMATION

ChemBioSim: Enhancing conformal in vivo toxicity predictions by bioactivity descriptors

Marina Garcia de Lomana[1,2], Andrea Morger[3], Ulf Norinder[4], Roland Buesen[1], Robert Landsiedel[1], Andrea Volkamer[3], Johannes Kirchmair[2] and Miriam Mathea[1]**

[1] BASF SE, 67063 Ludwigshafen am Rhein, Germany

[2] Department of Pharmaceutical Sciences, Faculty of Life Sciences, University of Vienna, 1090 Vienna, Austria

[3] In Silico Toxicology and Structural Bioinformatics, Institute of Physiology, Charité Universitätsmedizin Berlin, Charitéplatz 1, Berlin, Germany

[4] MTM Research Centre, School of Science and Technology, Örebro University, SE-70182 Örebro, Sweden

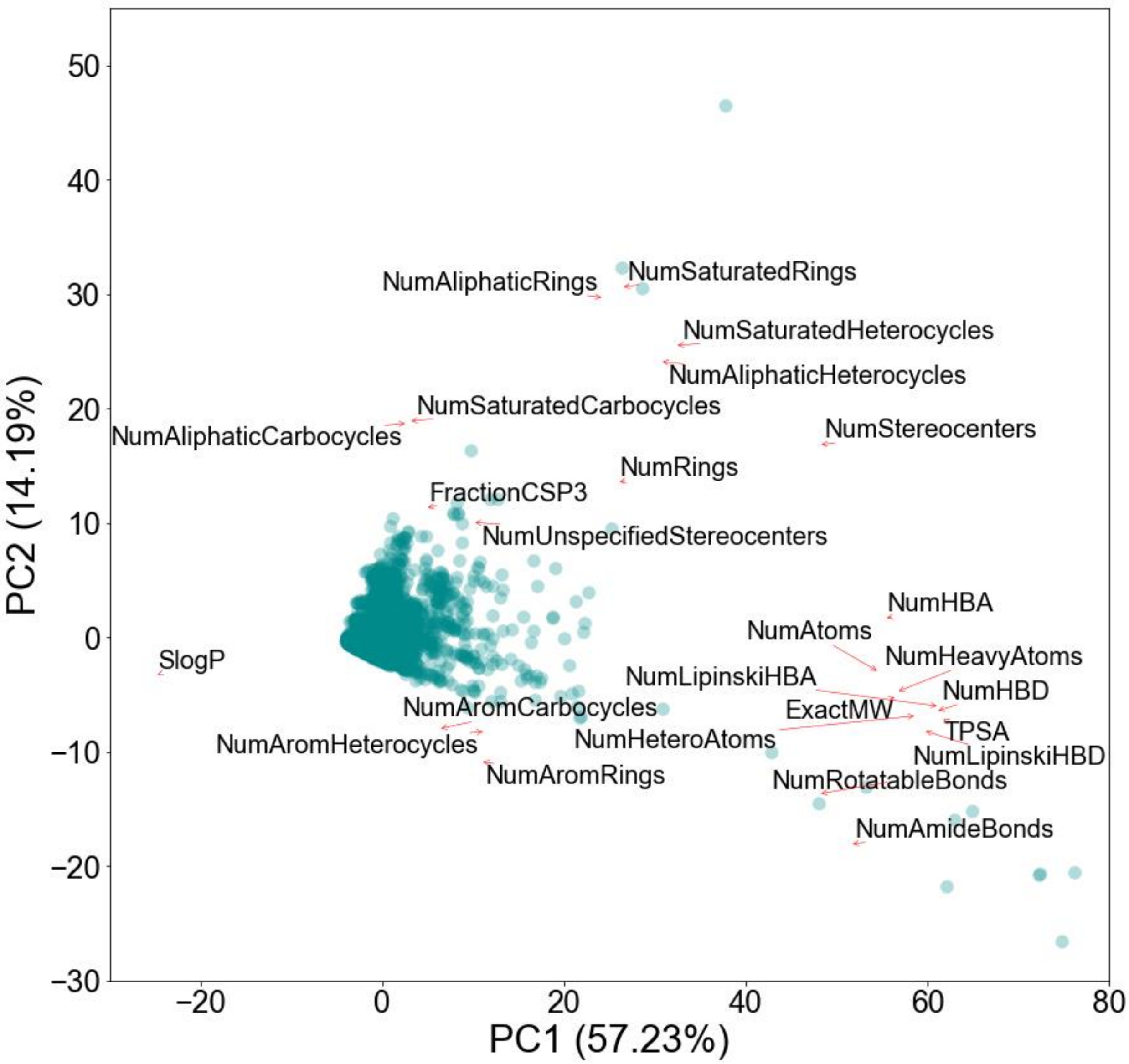


Figure S1. Loadings plot of the PCA based on a selection of interpretable molecular descriptors generated with RDKit on the global in vivo toxicity data set. The loadings plot shows how strongly each feature influences a principal component.

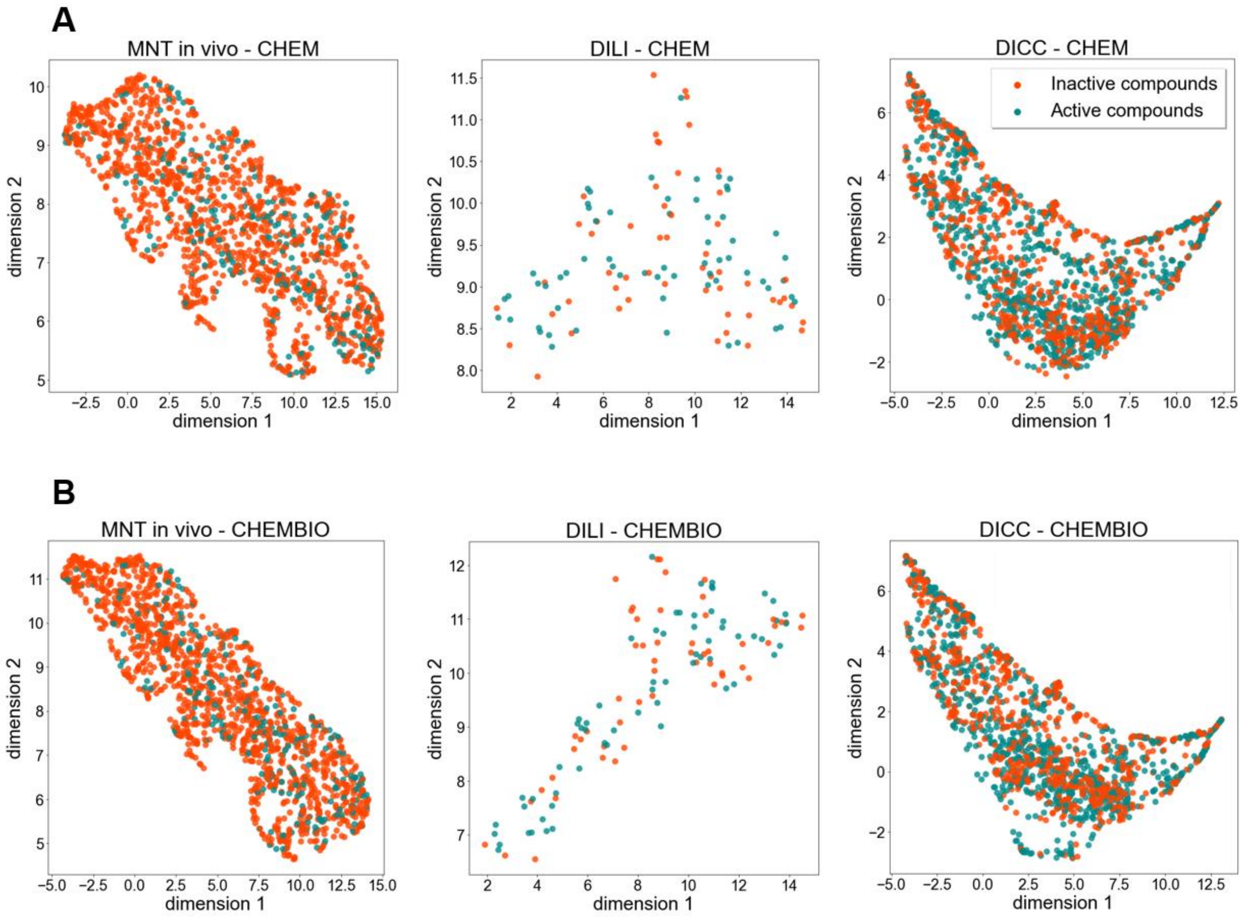


Figure S2. UMAP projections for the three in vivo endpoints (MNT in vivo, DILI and DICC) on (A) the CHEM descriptor set and (B) the CHEMBIO descriptor set.

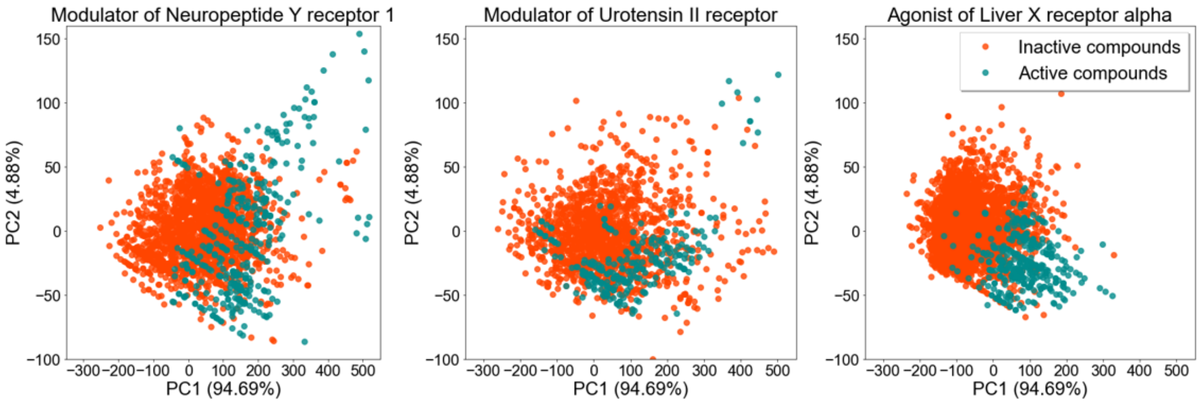
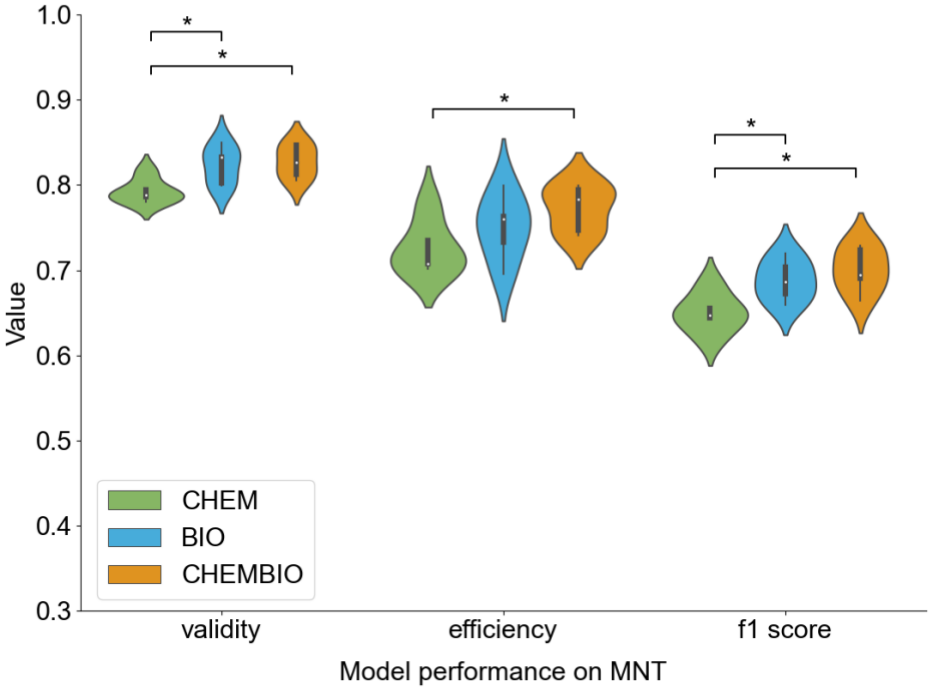
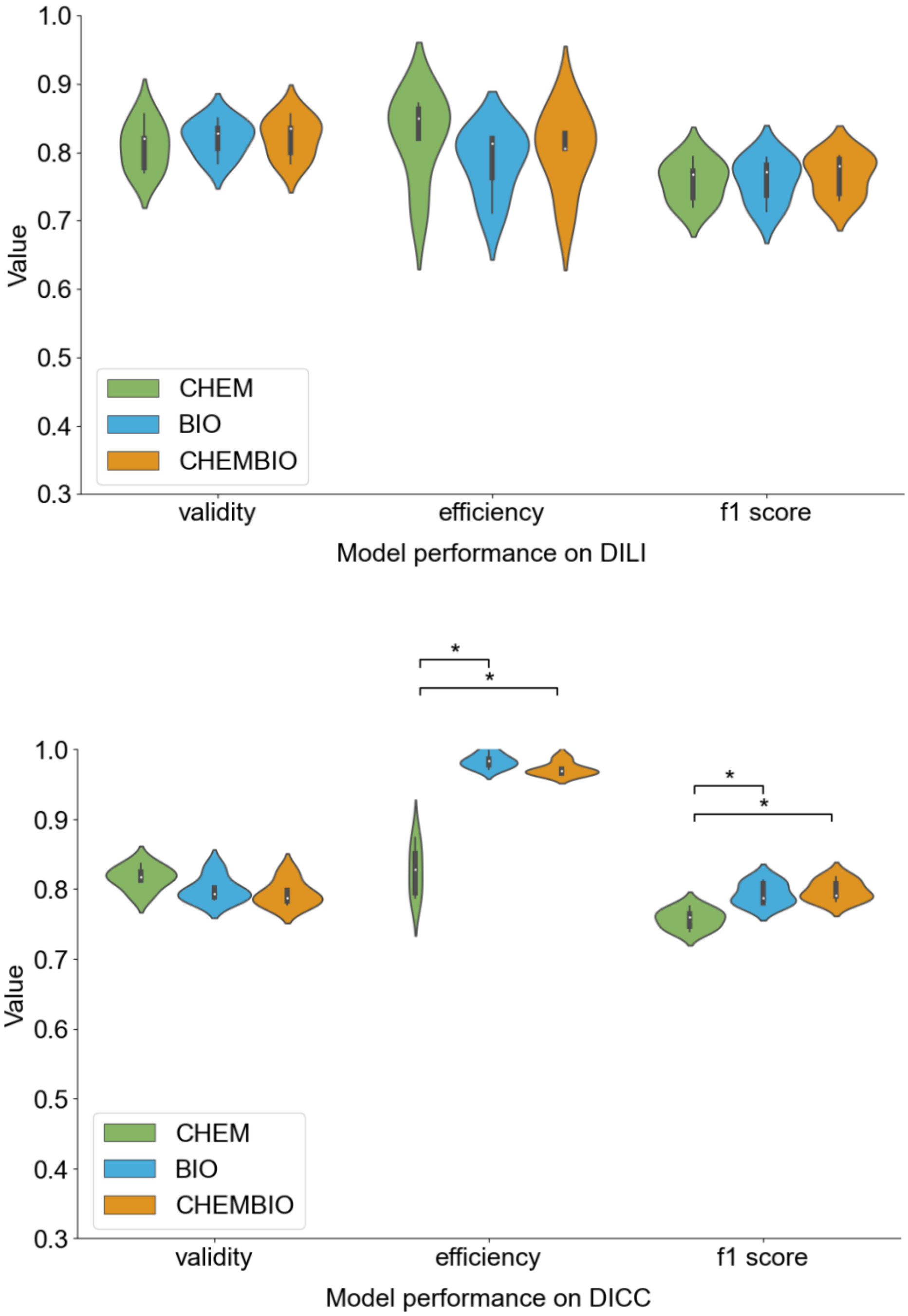


Figure S3. Principal component analysis based on a selection of interpretable molecular descriptors generated with RDKit. The PCA was derived from the merged data set of three eMolTox assays ("Modulator of Neuropeptide Y receptor type 1", "Modulator of Urotensin II receptor" and "Agonist of Liver X receptor alpha") for which the CP models yielded mean F1 scores on the single class predictions of 1.0. The active and inactive compounds of these data sets are located in differentiated parts of the chemical space, facilitating their classification.



(a)



(b)

(c)

Figure S4. Distribution of the validity, efficiency and F1 score values obtained within the 5-fold CV framework for the (a) MNT, (b) DILI and (c) DICC CP models built on the different descriptor sets without feature selection. The CHEM descriptor set includes the molecular fingerprint and physicochemical descriptors; the BIO descriptor set includes the predicted p-values for a set of biological assays (bioactivity descriptor); the CHEMBIO descriptor set includes the previous two descriptor sets. Significant differences in the distribution (p-value < 0.05) are denoted by a star.

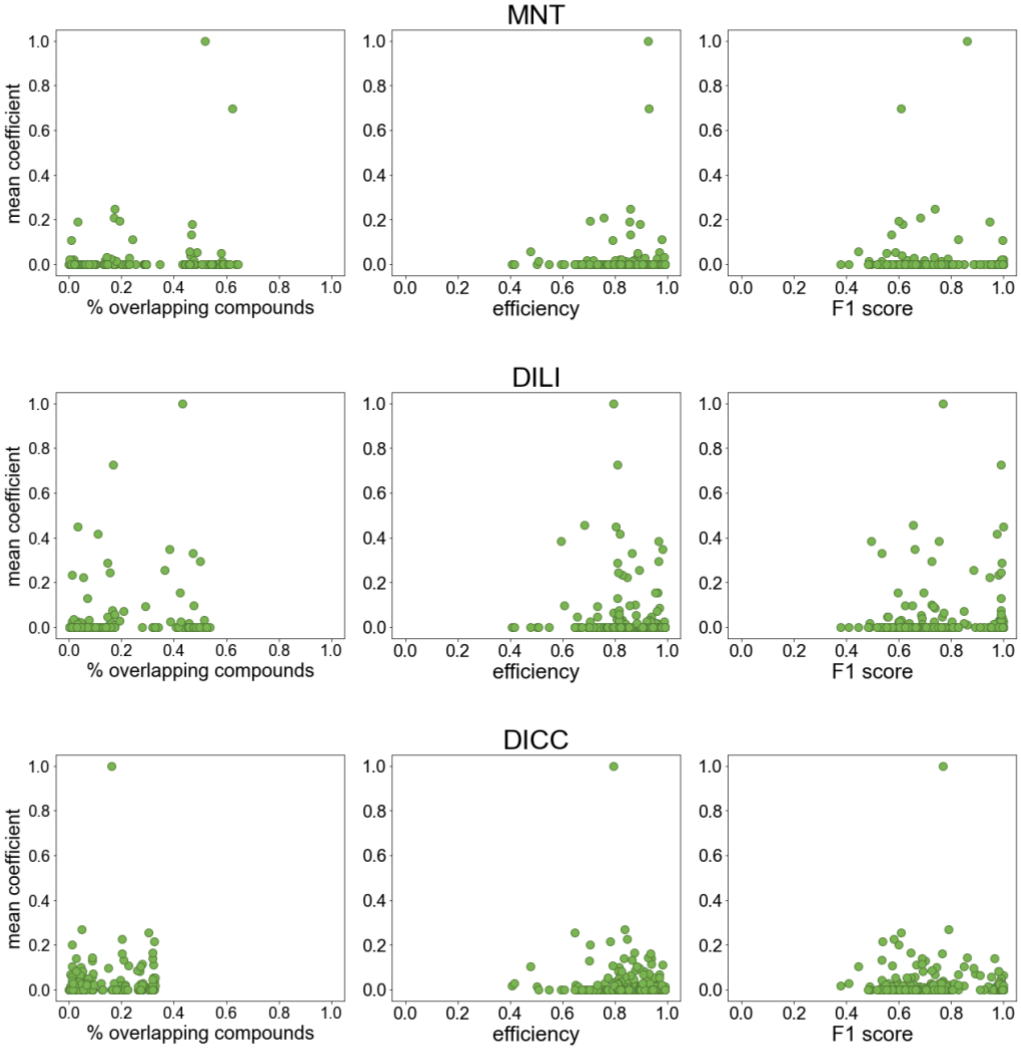


Figure S5. Mean coefficient reported by the lasso model for the bioactivity descriptors in relationship with the percentage of overlapping compounds (of the in vivo data set), the efficiency and F1 score of the models for each biological assay. For each of the 373 biological assays, the highest mean coefficient of the two p-values used as descriptors (for the active and inactive classes of each assay) was taken. The coefficients higher than 0 were normalized with a min-max normalization (from 0.01 to 1; see Materials and Methods section) for easier comparison.

Table S1. Data Sources and Download Links for the Original in Vivo Toxicity Data.

Endpoint	Data sources	Download link[1]	Query (json format)	MD5 file checksum	Checksum input file
MNT	10.1016/j.yrtph.2020.104620	-	-	6174327EB2B69D4326B36E5D610ACDE7	Supplementary .xlsx file
	eChemPortal (active)	https://www.echemportal.org/echemportal/property-search	{"blocks":[{"level":0,"type":"property","id":"dis3i1p7tjkdijou2p","label":"Genetic toxicity in vivo","endpointKind":"GeneticToxicityVivo"}],"endpoints": {"dis3i1p7tjkdijou2p":{"ENDPOINT_STUDY_RECORD.GeneticToxicityVivo.AdministrativeData.StudyResultType":{"1342":"","phrase":["1895"]},"ENDPOINT_STUDY_RECORD.GeneticToxicityVivo.AdministrativeData.Reliability":{"1342":"","phrase":["16","18","24","1342"]},"ENDPOINT_STUDY_RECORD.GeneticToxicityVivo.MaterialsAndMethods.Guideline.Qualifier":{"phrase":["1680","1880"]},"ENDPOINT_STUDY_RECORD.GeneticToxicityVivo.MaterialsAndMethods.Guideline.Guideline":{"1342":"","phrase":["1290"]},"ENDPOINT_STUDY_RECORD.GeneticToxicityVivo.ResultsAndDiscussion.TestRs.Genotoxicity":{"1342":"","phrase":["2276"]},"ENDPOINT_STUDY_RECORD.GeneticToxicityVivo.ResultsAndDiscussion.TestRs.Toxicity":{"phrase":["2170","2197","2207"]},"endpointKind":"GeneticToxicityVivo"}}}	6D1771634AE4FBDFDC9C517A0F2594FC	.csv file resulting from eChemPortal query
	eChemPortal (inactive)	https://www.echemportal.org/echemportal/property-search	{"blocks":[{"level":0,"type":"property","id":"dis3i1p7tjkdijou2p","label":"Genetic toxicity in vivo","endpointKind":"GeneticToxicityVivo"}],"endpoints": {"dis3i1p7tjkdijou2p":{"ENDPOINT_STUDY_RECORD.GeneticToxicityVivo.AdministrativeData.StudyResultType":{"1342":"","phrase":["1895"]},"ENDPOINT_STUDY_RECORD.GeneticToxicityVivo.AdministrativeData.Reliability":{"1342":"","phrase":["16","18","24","1342"]},"ENDPOINT_STUDY_RECORD.GeneticToxicityVivo.MaterialsAndMethods.Guideline.Qualifier":{"phrase":["1680","1880"]},"ENDPOINT_STUDY_RECORD.GeneticToxicityVivo.MaterialsAndMethods.Guideline.Guideline":{"1342":"","phrase":["1290"]},"ENDPOINT_STUDY_RECORD.GeneticToxicityVivo.ResultsAndDiscussion.TestRs.Genotoxicity":{"1342":null,"phrase":["2148"]},"ENDPOINT_STUDY_RECORD.GeneticToxicityVivo.ResultsAndDiscussion.TestRs.Toxicity":{"phrase":["2170","2197","2207"]},"endpointKind":"GeneticToxicityVivo"}}}	10BC1CB5B9C4F0FEDDAEF41600E13937	.csv file resulting from eChemPortal query
DILI	10.1016/j.drudis.2016.02.015	-	-	4EA88A5523A6717B9118AF7C4DAA9442	Supplementary .xlsx file
DICC	10.1021/acs.jcim.7b00641	-	-	80B6A4048F31A9EC74DD84522B1F861D	Supplementary .xlsx file

[1] Missing download links correspond to data sets available as supplementary material of the publication indicated as data source.

Table S2. Description of in Vitro Assays Endpoints and Databases and Corresponding Download Links.

Database/Endpoint	Description
ToxCast database	The ToxCast project is run by the U.S. Environmental Protection Agency (EPA) and has screened thousands of compounds in more than 700 high-throughput assays so far. In this work, the ToxCast database serves as the primary source of in vitro assay data. Binary activity labels for the individual endpoints ("active" and "inactive") were assigned based on the binary "hit-call" values ("0" and "1") obtained from the ToxCast database version 3.3. The hit-call values themselves were derived from activity curves. They indicate whether a minimum activity threshold has been reached. In this study, only endpoints with at least 200 active and 200 inactive compounds listed (after structure preparation and deduplication; see the section Structure Preparation for details) were considered for modeling. Furthermore, endpoints corresponding to background measurements (i.e. assays including “ch1”, “ch2” or “viability” in their name) were discarded. This procedure resulted in a total of 222 ToxCast endpoints (Table S3).
eMolTox database	The eMolTox web server contains models for the prediction of the activity of compounds in 174 in vitro and in vivo toxicity-related assays. The experimental data used for training these models is also available and was collected for this study. The activity labels provided with the eMolTox database are binary and were directly used. In analogy to the protocol followed for the ToxCast database, only endpoints with at least 200 active and 200 inactive compounds (after the structure preparation and deduplication steps; see section Structure Preparation for details) were considered for modeling. Any data on in vivo toxicity endpoints were discarded. This procedure resulted in a total of 136 in vitro assays (Table S3).
AMES mutagenicity assay	For the AMES assay, data from the European Chemicals Agency (ECHA) available at the eChemPortal were collected. Only experimental data derived according to the OECD Guidelines referring to this genotoxicity assay (OECD Guideline 471, 473 or 476; or equivalent) were considered. All assay outcomes annotated as unreliable or related to compounds that are cytotoxic were discarded. All compounds (identified based on CAS numbers) with conflicting activity data were also removed. Additional data on the AMES, chromosome aberration and mammalian cell gene mutation endpoints were obtained from the work of Benigni et al.[28], which includes curated data sets from the European Food Safety Authority (EFSA) data. In addition, the benchmark data set for the AMES test compiled by Hansen et al.[29] was incorporated.
Chromosome aberration assay	For the chromosome aberration assay, data from the European Chemicals Agency (ECHA) available at the eChemPortal were collected. Only experimental data derived according to the OECD Guideline referring to this genotoxicity assay (OECD Guidelines 473 or equivalent) were considered. All assay outcomes annotated as unreliable or related to compounds that are cytotoxic were discarded. All compounds (identified based on CAS numbers) with conflicting activity data were also removed. Additional data on the AMES, chromosome aberration and mammalian cell gene mutation endpoints were obtained from the work of Benigni et al.[28], which includes curated data sets from the European Food Safety Authority (EFSA) data.
Mammalian mutagenicity assay	For the mammalian cell gene mutation assay, data from the European Chemicals Agency (ECHA) available at the eChemPortal were collected. Only experimental data derived according to the OECD Guideline referring to this genotoxicity assay (OECD Guidelines 476 or equivalent) were considered. All assay outcomes annotated as unreliable or related to compounds that are cytotoxic were discarded. All compounds (identified based on CAS numbers) with conflicting activity data were also removed. Additional data on the AMES, chromosome aberration and mammalian cell gene mutation endpoints were obtained from the work of Benigni et al.[28], which includes curated data sets from the European Food Safety Authority (EFSA) data.
Bioavailability	A data set describing human oral bioavailability was collected from Falcón-Cano et al.[27] To derive a binary label for the bioavailability data indicated in the data set, a cut-off of 50% bioavailability, as proposed by Falcón-Cano et al., was applied (compounds were labeled "inactive" if the bioavailability percentage is lower than 50%; otherwise they were labeled "active").
Permeability	Permeability data were obtained from the work of Wang et al.[30] on Caco-2 cells, where the permeability of a compound is indicated by its apparent permeability coefficient (Papp). To derive binary labels, a cut-off of 20 × 10–6 cm/s was applied on the Papp values, as proposed by Wang et al. (compounds were labeled "inactive" if the Papp value is lower than the threshold; otherwise, they were labeled "active").
Thyroid hormone homeostasis	Nine data sets for molecular initiating events related to thyroid hormone homeostasis were collected from Garcia de Lomana et al.[31] These nine assays describe inhibitors of deiodinases 1, 2 and 3, thyroid peroxidase and sodium iodide symporter; antagonists of the thyroid hormone receptor, thyrotropin-releasing hormone receptor and thyroid stimulating hormone receptor (TSHR); as well as agonists of TSHR. The data sets contain binary activity labels primarily obtained from the ToxCast hit-call values and related literature and including data curation steps for removing possible false positive and false negative results.
P-Glycoprotein inhibition	A data set on P-Glycoprotein (ABCB1) inhibition by small molecules was obtained from Broccatelli et al.[32] The binary activity labels provided with this data set are based on IC50 and percent inhibition values of the compounds and were used as is.

Database/Endpoint	**Data sources**	**Download link[1]**	**Query (json format)**	**MD5 file checksum**	**Checksum input file**
ToxCast database	ToxCast version 3.3	ftp://newftp.epa.gov/COMPTOX/High_Throughput_Screening_Data/InVitroDB_V3.3/ToxCast_Data_July_2020/	-	962EFEE512FEE51DFAFB2FECD210ACB5	"INVITRODB_V3_3_LEVEL5.zip" file
eMolTox database	10.1093/bioinformatics/bty135	http://xundrug.cn/moltox/about	-	3401033F5FC1CB907AB295419199301D	"train_data" .tar.gz file
AMES mutagenicity assay	10.1021/ci900161g	-	-	AA408BCD82C3F98D086C9D63C38AA488	Supplementary .smi file
	10.2903/sp.efsa.2019.EN-1598	https://efsa.onlinelibrary.wiley.com/doi/10.2903/sp.efsa.2019.EN-1598	-	2DE6AFD2C3F46FE5238E48E25350ED62	Supplementary .xlsx file
	eChemPortal (active)	https://www.echemportal.org/echemportal/property-search	{"blocks":[{"level":0,"type":"property","id":"ifmhvogurmskdigbmf4","label":"Genetic toxicity in vitro","endpointKind":"GeneticToxicityVitro"}],"endpoints": {"ifmhvogurmskdigbmf4":{"ENDPOINT_STUDY_RECORD.GeneticToxicityVitro.AdministrativeData.StudyResultType":{"1342":"","phrase":["1895"]},"ENDPOINT_STUDY_RECORD.GeneticToxicityVitro.AdministrativeData.Reliability":{"1342":"","phrase":["16","18","24","1342"]},"ENDPOINT_STUDY_RECORD.GeneticToxicityVitro.MaterialsAndMethods.Guideline.Qualifier":{"phrase":["1680","1880"]},"ENDPOINT_STUDY_RECORD.GeneticToxicityVitro.MaterialsAndMethods.Guideline.Guideline":{"1342":"","phrase":["1287"]},"ENDPOINT_STUDY_RECORD.GeneticToxicityVitro.MaterialsAndMethods.GLPComplianceStatement":{"phrase":null},"ENDPOINT_STUDY_RECORD.GeneticToxicityVitro.ResultsAndDiscussion.TestRs.Genotoxicity":{"1342":"","phrase":["2276"]},"ENDPOINT_STUDY_RECORD.GeneticToxicityVitro.ResultsAndDiscussion.TestRs.Cytotoxicity":{"1342":"","phrase":["64852","4120","4121","2196","2207","1342"]},"endpointKind":"GeneticToxicityVitro"}}}	9E450D7E6F3C9FC0D833993187E7D4CE	.csv file resulting from eChemPortal query
	eChemPortal (inactive)	https://www.echemportal.org/echemportal/property-search	{"blocks":[{"level":0,"type":"property","id":"ifmhvogurmskdigbmf4","label":"Genetic toxicity in vitro","endpointKind":"GeneticToxicityVitro"}],"endpoints": {"ifmhvogurmskdigbmf4":{"ENDPOINT_STUDY_RECORD.GeneticToxicityVitro.AdministrativeData.StudyResultType":{"1342":"","phrase":["1895"]},"ENDPOINT_STUDY_RECORD.GeneticToxicityVitro.AdministrativeData.Reliability":{"1342":"","phrase":["16","18","24","1342"]},"ENDPOINT_STUDY_RECORD.GeneticToxicityVitro.MaterialsAndMethods.Guideline.Qualifier":{"phrase":["1680","1880"]},"ENDPOINT_STUDY_RECORD.GeneticToxicityVitro.MaterialsAndMethods.Guideline.Guideline":{"1342":"","phrase":["1287"]},"ENDPOINT_STUDY_RECORD.GeneticToxicityVitro.MaterialsAndMethods.GLPComplianceStatement":{"phrase":null},"ENDPOINT_STUDY_RECORD.GeneticToxicityVitro.ResultsAndDiscussion.TestRs.Genotoxicity":{"1342":null,"phrase":["2148"]},"ENDPOINT_STUDY_RECORD.GeneticToxicityVitro.ResultsAndDiscussion.TestRs.Cytotoxicity":{"1342":"","phrase":["64852","4120","4121","2196","2207","1342"]},"endpointKind":"GeneticToxicityVitro"}}}	7126FB1A36E9CAB84303669F68CD14A6	.csv file resulting from eChemPortal query

Chromosome aberration assay	10.2903/sp.efsa.2019.EN-1598	https://efsa.onlinelibrary.wiley.com/doi/10.2903/sp.efsa.2019.EN-1598	-	2DE6AFD2C3F46FE5238E48E25350ED62	Supplementary .xlsx file
	eChemPortal (active)	https://www.echemportal.org/echemportal/property-search	{"blocks":[{"level":0,"type":"property","id":"ifmhvogurmskdigbmf4","label":"Genetic toxicity in vitro","endpointKind":"GeneticToxicityVitro"}],"endpoints":{"ifmhvogurmskdigbmf4":{"ENDPOINT_STUDY_RECORD.GeneticToxicityVitro.AdministrativeData.StudyResultType":{"1342":"","phrase":["1895"]},"ENDPOINT_STUDY_RECORD.GeneticToxicityVitro.AdministrativeData.Reliability":{"1342":"","phrase":["16","18","24","1342"]},"ENDPOINT_STUDY_RECORD.GeneticToxicityVitro.MaterialsAndMethods.Guideline.Qualifier":{"phrase":["1680","1880"]},"ENDPOINT_STUDY_RECORD.GeneticToxicityVitro.MaterialsAndMethods.Guideline.Guideline":{"1342":null,"phrase":["64850","64851"]},"ENDPOINT_STUDY_RECORD.GeneticToxicityVitro.MaterialsAndMethods.GLPComplianceStatement":{"phrase":null},"ENDPOINT_STUDY_RECORD.GeneticToxicityVitro.ResultsAndDiscussion.TestRs.Genotoxicity":{"1342":null,"phrase":["2276"]},"ENDPOINT_STUDY_RECORD.GeneticToxicityVitro.ResultsAndDiscussion.TestRs.Cytotoxicity":{"1342":"","phrase":["64852","4120","4121","2196","2207","1342"]},"endpointKind":"GeneticToxicityVitro"}}}	EEF5BCA9FF1A0FD128026410D4B83345	.csv file resulting from eChemPortal query
	eChemPortal (inactive)	https://www.echemportal.org/echemportal/property-search	{"blocks":[{"level":0,"type":"property","id":"ifmhvogurmskdigbmf4","label":"Genetic toxicity in vitro","endpointKind":"GeneticToxicityVitro"}],"endpoints":{"ifmhvogurmskdigbmf4":{"ENDPOINT_STUDY_RECORD.GeneticToxicityVitro.AdministrativeData.StudyResultType":{"1342":"","phrase":["1895"]},"ENDPOINT_STUDY_RECORD.GeneticToxicityVitro.AdministrativeData.Reliability":{"1342":"","phrase":["16","18","24","1342"]},"ENDPOINT_STUDY_RECORD.GeneticToxicityVitro.MaterialsAndMethods.Guideline.Qualifier":{"phrase":["1680","1880"]},"ENDPOINT_STUDY_RECORD.GeneticToxicityVitro.MaterialsAndMethods.Guideline.Guideline":{"1342":null,"phrase":["64850","64851"]},"ENDPOINT_STUDY_RECORD.GeneticToxicityVitro.MaterialsAndMethods.GLPComplianceStatement":{"phrase":null},"ENDPOINT_STUDY_RECORD.GeneticToxicityVitro.ResultsAndDiscussion.TestRs.Genotoxicity":{"1342":null,"phrase":["2148"]},"ENDPOINT_STUDY_RECORD.GeneticToxicityVitro.ResultsAndDiscussion.TestRs.Cytotoxicity":{"1342":"","phrase":["64852","4120","4121","2196","2207","1342"]},"endpointKind":"GeneticToxicityVitro"}}}	4408C9F92CAA548888BC21A72D5B4CEE	.csv file resulting from eChemPortal query
Mammalian mutagenicity assay	10.2903/sp.efsa.2019.EN-1598	https://efsa.onlinelibrary.wiley.com/doi/10.2903/sp.efsa.2019.EN-1598	-	2DE6AFD2C3F46FE5238E48E25350ED62	Supplementary .xlsx file

	eChemPortal (active)	https://www.echemportal.org/echemportal/property-search	{"blocks":[{"level":0,"type":"property","id":"ifmhvogurmskdigbmf4","label":"Genetic toxicity in vitro","endpointKind":"GeneticToxicityVitro"}],"endpoints": {"ifmhvogurmskdigbmf4":{"ENDPOINT_STUDY_RECORD.GeneticToxicityVitro.AdministrativeData.StudyResultType":{"1342":"","phrase":["1895"]},"ENDPOINT_STUDY_RECORD.GeneticToxicityVitro.AdministrativeData.Reliability":{"1342":"","phrase":["16","18","24","1342"]},"ENDPOINT_STUDY_RECORD.GeneticToxicityVitro.MaterialsAndMethods.Guideline.Qualifier":{"phrase":["1680","1880"]},"ENDPOINT_STUDY_RECORD.GeneticToxicityVitro.MaterialsAndMethods.Guideline.Guideline":{"1342":null,"phrase":["64855"]},"ENDPOINT_STUDY_RECORD.GeneticToxicityVitro.MaterialsAndMethods.GLPComplianceStatement":{"phrase":null},"ENDPOINT_STUDY_RECORD.GeneticToxicityVitro.ResultsAndDiscussion.TestRs.Genotoxicity":{"1342":null,"phrase":["2276"]},"ENDPOINT_STUDY_RECORD.GeneticToxicityVitro.ResultsAndDiscussion.TestRs.Cytotoxicity":{"1342":"","phrase":["64852","4120","4121","2196","2207","1342"]},"endpointKind":"GeneticToxicityVitro"}}}	3E83FF9EBC8E392FE5D0BEA55934D817	.csv file resulting from eChemPortal query
	eChemPortal (inactive)	https://www.echemportal.org/echemportal/property-search	{"blocks":[{"level":0,"type":"property","id":"ifmhvogurmskdigbmf4","label":"Genetic toxicity in vitro","endpointKind":"GeneticToxicityVitro"}],"endpoints": {"ifmhvogurmskdigbmf4":{"ENDPOINT_STUDY_RECORD.GeneticToxicityVitro.AdministrativeData.StudyResultType":{"1342":"","phrase":["1895"]},"ENDPOINT_STUDY_RECORD.GeneticToxicityVitro.AdministrativeData.Reliability":{"1342":"","phrase":["16","18","24","1342"]},"ENDPOINT_STUDY_RECORD.GeneticToxicityVitro.MaterialsAndMethods.Guideline.Qualifier":{"phrase":["1680","1880"]},"ENDPOINT_STUDY_RECORD.GeneticToxicityVitro.MaterialsAndMethods.Guideline.Guideline":{"1342":null,"phrase":["64855"]},"ENDPOINT_STUDY_RECORD.GeneticToxicityVitro.MaterialsAndMethods.GLPComplianceStatement":{"phrase":null},"ENDPOINT_STUDY_RECORD.GeneticToxicityVitro.ResultsAndDiscussion.TestRs.Genotoxicity":{"1342":null,"phrase":["2148"]},"ENDPOINT_STUDY_RECORD.GeneticToxicityVitro.ResultsAndDiscussion.TestRs.Cytotoxicity":{"1342":"","phrase":["64852","4120","4121","2196","2207","1342"]},"endpointKind":"GeneticToxicityVitro"}}}	96602EA7965095B4245A96C9015450A1	.csv file resulting from eChemPortal query
Bioavailability	doi.org/10.1021/acs.jcim.0c00019	-	-	8DE686BF3664E5CC327420CDA7EB8480	Supplementary .xlsx file
Permeability	10.1021/acs.jcim.5b00642	-	-	879BA561143296B2DD99C334F9A66A9E	Supplementary .xlsx file
Thyroid hormone homeostasis	10.1021/acs.chemrestox.0c00304	-	-	6415103CF4F96B9D4F8845765F30475B	Supplementary .xlsx file
P-Glycoprotein inhibition	10.1021/jm101421d	-	-	7FFACE7954090CBD2847BC41ADE7D42A	Supplementary .xls file

[1] Missing download links correspond to data sets directly available as supplementary material of the publication indicated as data source

Table S3. List of Biological Assays With the Number of Active and Inactive Compounds in Their Data Sets.

Endpoint	**Number of**		**Total**
	active compounds	**inactive compounds**	
AMES	3530	5136	8666
Mammalian cell gene mutation	56	789	845
Chromosome aberration	143	813	956
Human oral bioavailability	797	678	1475
PGP inhibition	651	567	1218
Caco-2	328	718	1046
DIO1 inhibition	108	1563	1671
DIO2 inhibition	175	1504	1679
DIO3 inhibition	180	1498	1678
TPO inhibition	252	780	1032
TR antagonism	1209	4912	6121
NIS inhibition	49	736	785
TRHR antagonism	52	6340	6392
TSHR antagonism	102	6380	6482
TSHR agonism	196	6379	6575
eMolTox - Modulator of Muscarinic acetylcholine receptor M5	287	1625	1912
eMolTox - Antagonist of the thyroid receptor (TR) signaling pathway	356	5102	5458
eMolTox - Modulator of Glucocorticoid receptor	1778	9373	11151
eMolTox - Mutagenicity	3299	2758	6057
eMolTox - Modulator of GABA-A receptor alpha-5beta-3gamma-2	601	1896	2497
eMolTox - Modulator of Muscarinic acetylcholine receptor M2	1075	5730	6805

eMolTox - Modulator of Neuropeptide Y receptor type 1	381	1990	2371
eMolTox - Modulator of Cannabinoid CB2 receptor	3527	11052	14579
eMolTox - Modulator of Androgen Receptor	1269	6890	8159
eMolTox - Modulator of Monoamine oxidase A	435	2250	2685
eMolTox - Agonist of the thyroid stimulating hormone receptor (TSHR) signaling pathway	310	6241	6551
eMolTox - Modulator of Cholecystokinin A receptor	247	1325	1572
eMolTox - Modulator of Neurokinin 1 receptor	2008	9501	11509
eMolTox - Modulator of Serotonin 2c (5-HT2c) receptor	2136	11622	13758
eMolTox - Modulator of GABA-A receptor alpha-2beta-3gamma-2	485	1897	2382
eMolTox - Modulator of Acetylcholinesterase	1717	9250	10967
eMolTox - Modulator of Cannabinoid CB1 receptor	2714	11627	14341
eMolTox - Modulator of Neurokinin 2 receptor	666	3640	4306
eMolTox - Modulator of Histamine H1 receptor	858	4550	5408
eMolTox - Modulator of Alpha-1a adrenergic receptor	1265	7183	8448
eMolTox - Antagonist of the androgen receptor (AR) signaling pathway	507	5640	6147
eMolTox - Modulator of Neuronal acetylcholine receptor protein alpha-7 subunit	344	1920	2264
eMolTox - Activator Alzheimers amyloid precursor	1982	19855	21837
eMolTox - Differential cytotoxicity (isogenic chicken DT40 Rev3 mutant cell line)	1923	4367	6290
eMolTox - Antagonist of the estrogen receptor alpha (ER-alpha) signaling pathway dup	348	5936	6284
eMolTox - Modulator of HERG	1568	8027	9595
eMolTox - Antagonist of the retinoic acid receptor (RAR) signaling pathway	540	4672	5212
eMolTox - Modulator of Serotonin 2b (5-HT2b) receptor	1053	5552	6605

eMolTox - Modulator of Platelet-derived growth factor receptor beta	712	3600	4312
eMolTox - Antagonist of the retinoid-related orphan receptor gamma (ROR-gamma) signaling pathway	483	4600	5083
eMolTox - Modulator of Serotonin 1a (5-HT1a) receptor	3000	11721	14721
eMolTox - Modulator of Neuronal acetylcholine receptor alpha4beta2	525	2960	3485
eMolTox - Agonist of H2AX	379	6241	6620
eMolTox - Modulator of GABA-A receptor alpha-3beta-3gamma-2	610	1898	2508
eMolTox - Induce genotoxicity in human embryonic kidney cells	274	6772	7046
eMolTox - Modulator of Glutamate NMDA receptor	254	1335	1589
eMolTox - Modulator of Norepinephrine transporter	1953	11164	13117
eMolTox - Modulator of Serotonin transporter	3123	11233	14356
eMolTox - Modulator of Sodium channel protein type IX alpha subunit	2116	8545	10661
eMolTox - Agonist of the AP-1 signaling pathway	552	5878	6430
eMolTox - Agonist of the p53 signaling pathway	494	6307	6801
eMolTox - Modulator of Serotonin 3a (5-HT3a) receptor	329	1825	2154
eMolTox - Modulator of Serotonin 1b (5-HT1b) receptor	863	4535	5398
eMolTox - Modulator of Delta opioid receptor	2144	11186	13330
eMolTox - Modulator of Vascular endothelial growth factor receptor 1	1030	5204	6234
eMolTox - Modulator of Beta-2 adrenergic receptor	1026	5471	6497
eMolTox - Modulator of Muscarinic acetylcholine receptor M4	407	2240	2647
eMolTox - Modulator of TNF-alpha	340	1745	2085
eMolTox - Modulator of Calcitonin gene-related peptide type 1 receptor	527	2451	2978

eMolTox - Modulator of Alpha-2a adrenergic receptor	577	3101	3678
eMolTox - Modulator of Sigma opioid receptor	1724	9309	11033
eMolTox - Differential cytotoxicity against isogenic chicken DT40 cell lines with known DNA damage response pathways - Rad54Ku70 mutant cell line	1931	4632	6563
eMolTox - Modulator of GABA-A receptor alpha-1beta-3gamma-2	569	1919	2488
eMolTox - Modulator of Dopamine transporter	1535	8975	10510
eMolTox - Modulator of GABA-A receptor anion channel	927	2748	3675
eMolTox - Differential cytotoxicity (isogenic chicken DT40 cell lines)	1804	4615	6419
eMolTox - Cytotoxicity in HEK293 cells - 16 hour	307	6386	6693
eMolTox - Aromatase inhibitors	295	5515	5810
eMolTox - Modulators of myocardial damage	2264	22689	24953
eMolTox - Antagonist of the androgen receptor (AR) signaling pathway dup	378	5703	6081
eMolTox - Modulator of Adenosine A2a receptor	2924	12881	15805
eMolTox - Modulator of Serotonin 2a (5-HT2a) receptor	2896	13019	15915
eMolTox - Modulator of Beta-3 adrenergic receptor	1273	6617	7890
eMolTox - Agonist of the estrogen receptor alpha (ER-alpha) signaling pathway	402	6575	6977
eMolTox - Modulator of Kappa opioid receptor	2536	13272	15808
eMolTox - Modulator of Adenosine A3 receptor	2634	8893	11527
eMolTox - Modulator of Serotonin 7 (5-HT7) receptor	1378	7083	8461
eMolTox - Modulator of Beta-1 adrenergic receptor	847	4412	5259
eMolTox - Antagonist of the estrogen receptor alpha (ER-alpha) signaling pathway	349	5787	6136
eMolTox - Modulator of Adenosine A2b receptor	1415	5004	6419
eMolTox - Modulator of Vascular endothelial growth factor receptor 3	356	1820	2176

eMolTox - Modulator of Vascular endothelial growth factor receptor 2	4569	9429	13998
eMolTox - Modulator of Receptor protein-tyrosine kinase erbB-2	987	5175	6162
eMolTox - Modulator of Alpha-2b adrenergic receptor	327	1767	2094
eMolTox - Modulator of Serotonin 4 (5-HT4) receptor	433	2225	2658
eMolTox - Modulator of Platelet-derived growth factor receptor alpha	323	1655	1978
eMolTox - Modulator of Melatonin receptor 1B	705	1540	2245
eMolTox - Antagonist of the glucocorticoid receptor (GR) signaling pathway	363	5614	5977
eMolTox - Agonist of the androgen receptor (AR) signaling pathway	240	6496	6736
eMolTox - Cytotoxicity in HEK293 cells - 32 hour	531	6022	6553
eMolTox - Modulator of Muscarinic acetylcholine receptor M1	1112	5935	7047
eMolTox - Modulator of Dopamine D2 receptor	4352	11910	16262
eMolTox - Modulator of Angiotensin II type 2 (AT-2) receptor	283	1224	1507
eMolTox - Modulator of Urotensin II receptor	272	1460	1732
eMolTox - Modulator of P2X purinoceptor 7	1790	4075	5865
eMolTox - Agonist of the antioxidant response element (ARE) signaling pathway	908	4690	5598
eMolTox - Block Bile Salt Export Pump	350	301	651
eMolTox - Cytotoxicity in HepG2 cells - 8 hour	305	6432	6737
eMolTox - Agonist of the farnesoid-X-receptor (FXR) signaling pathway	402	5366	5768
eMolTox - Antagonist of the peroxisome proliferator-activated receptor gamma (PPARg) signaling pathway	368	5388	5756
eMolTox - Cytotoxicity in HepG2 cells - 24 hour	592	5941	6533
eMolTox - Modulator of Adenosine A1 receptor	2520	12934	15454

eMolTox - Agonist of Liver X receptor alpha	396	3407	3803
eMolTox - Modulator of Dopamine D1 receptor	712	3825	4537
eMolTox - Modulator of Peroxisome proliferator-activated receptor gamma	1932	10491	12423
eMolTox - Agonist of the peroxisome proliferator-activated receptor gamma (PPARg) signaling pathway	206	6266	6472
eMolTox - Cytotoxicity in HEK293 cells - 40 hour	641	5878	6519
eMolTox - Modulator of Endothelin receptor ET-B	501	1077	1578
eMolTox - Induce Phospholipidosis	220	520	740
eMolTox - Inhibitors of Hepatocyte nuclear factor 4 (HNF4) dimerization	1711	509	2220
eMolTox - Cytotoxicity in HEK293 cells - 8 hour	257	6458	6715
eMolTox - Modulator of Bradykinin B2 receptor	377	1930	2307
eMolTox - Inhibit CYP2C19 Activity	20287	20283	40570
eMolTox - Disruptors of the mitochondrial membrane potential	935	4918	5853
eMolTox - Modulator of Endothelin receptor ET-A	1073	1007	2080
eMolTox - Modulator of Mu opioid receptor	2532	12213	14745
eMolTox - Agonist of the RXR signaling pathway	208	5374	5582
eMolTox - Modulator of Muscarinic acetylcholine receptor M3	1261	6815	8076
eMolTox - Antagonist of the vitamin D receptor (VDR) signaling pathway	326	5578	5904
eMolTox - Modulator of P2X purinoceptor 3	827	564	1391
eMolTox - Inhibit CYP1A2 Activity	4064	3363	7427
eMolTox - Cytotoxicity in HepG2 cells - 40 hour	786	5685	6471
eMolTox - Cytotoxicity in HEK293 cells - 24 hour	432	6187	6619
eMolTox - Substrates of Cytochrome P450 2C19	1846	5896	7742
eMolTox - Modulator of Alpha-1b adrenergic receptor	960	5417	6377
eMolTox - Antagonist of the farnesoid-X-receptor (FXR) signaling pathway	274	5966	6240

eMolTox - Cytotoxicity in HepG2 cells - 16 hour	445	6203	6648
eMolTox - Activators of the human pregnane X receptor (PXR) signaling pathway	228	1660	1888
eMolTox - Activators of Cytochrome P450 2A9	1368	13676	15044
eMolTox - Modulator of Platelet activating factor receptor	285	688	973
eMolTox - Modulator of Vasopressin V1a receptor	497	2665	3162
eMolTox - Modulator of Type-1 angiotensin II receptor	589	2367	2956
eMolTox - Inhibitors and Substrates of Cytochrome P450 3A4	3271	6192	9463
eMolTox - Modulator of Cyclooxygenase-1	292	1476	1768
eMolTox - Inhibit CYP2C9 Activity	18725	18719	37444
eMolTox - Agonist of Liver X receptor beta	451	3407	3858
eMolTox - Antagonist of the farnesoid-X-receptor (FXR) signaling pathway dup	214	5679	5893
eMolTox - Modulator of Cyclooxygenase-2	1506	7793	9299
eMolTox - Antagonist of the constitutive androstane receptor (CAR) signaling pathway	418	6267	6685
eMolTox - Activator the aryl hydrocarbon receptor (AhR) signaling pathway	838	5721	6559
eMolTox - Inhibitors and Substrates of Cytochrome P450 2D6	1559	5851	7410
eMolTox - Modulator of Angiotensin-converting enzyme	321	1879	2200
eMolTox - Agonist of the constitutive androstane receptor (CAR) signaling pathway	867	5550	6417
eMolTox - Cytotoxicity in HepG2 cells - 32 hour	695	5792	6487
eMolTox - Activators of the heat shock response signaling pathway	389	5842	6231
eMolTox - Inhibitors and Substrates of Cytochrome P450 2C9	1226	6427	7653
ToxCast - ACEA AR antagonist 80hr	676	991	1667
ToxCast - ACEA ER 80hr	398	2376	2774

ToxCast - APR HepG2 CellCycleArrest 72h dn	225	778	1003
ToxCast - APR HepG2 CellLoss 24h dn	284	722	1006
ToxCast - APR HepG2 CellLoss 72h dn	424	581	1005
ToxCast - APR HepG2 MitoticArrest 72h up	266	722	988
ToxCast - APR HepG2 OxidativeStress 24h up	224	779	1003
ToxCast - APR HepG2 OxidativeStress 72h up	235	768	1003
ToxCast - ATG AP 1 CIS up	581	2716	3297
ToxCast - ATG Ahr CIS dn	203	3105	3308
ToxCast - ATG Ahr CIS up	422	2899	3321
ToxCast - ATG BRE CIS up	355	2952	3307
ToxCast - ATG CMV CIS up	502	2794	3296
ToxCast - ATG CRE CIS up	328	2972	3300
ToxCast - ATG C EBP CIS up	223	3088	3311
ToxCast - ATG DR4 LXR CIS dn	647	2649	3296
ToxCast - ATG DR5 CIS up	294	3016	3310
ToxCast - ATG EGR CIS up	472	2829	3301
ToxCast - ATG ERE CIS up	818	2492	3310
ToxCast - ATG ERa TRANS up	735	2583	3318
ToxCast - ATG E Box CIS dn	296	3005	3301
ToxCast - ATG Ets CIS dn	211	3105	3316
ToxCast - ATG HIF1a CIS up	354	2947	3301
ToxCast - ATG HSE CIS up	387	2920	3307
ToxCast - ATG IR1 CIS dn	323	2980	3303
ToxCast - ATG ISRE CIS dn	430	2865	3295
ToxCast - ATG MRE CIS up	685	2615	3300
ToxCast - ATG NFI CIS up	253	3061	3314
ToxCast - ATG NF kB CIS dn	268	3055	3323
ToxCast - ATG NRF2 ARE CIS up	1245	2040	3285
ToxCast - ATG Oct MLP CIS up	526	2762	3288

ToxCast - ATG PBREM CIS up	373	2929	3302
ToxCast - ATG PPARa TRANS up	229	3100	3329
ToxCast - ATG PPARg TRANS up	923	2383	3306
ToxCast - ATG PPRE CIS up	492	2813	3305
ToxCast - ATG PXRE CIS dn	303	3019	3322
ToxCast - ATG PXRE CIS up	1619	1679	3298
ToxCast - ATG PXR TRANS up	934	2347	3281
ToxCast - ATG Pax6 CIS up	419	2882	3301
ToxCast - ATG RARa TRANS dn	217	3093	3310
ToxCast - ATG RORE CIS up	369	2942	3311
ToxCast - ATG RXRb TRANS up	498	2806	3304
ToxCast - ATG SREBP CIS up	490	2814	3304
ToxCast - ATG Sox CIS up	214	3107	3321
ToxCast - ATG Sp1 CIS up	318	2987	3305
ToxCast - ATG TA CIS up	209	3108	3317
ToxCast - ATG TCF b cat CIS dn	405	2882	3287
ToxCast - ATG VDRE CIS up	882	2395	3277
ToxCast - ATG Xbp1 CIS up	422	2879	3301
ToxCast - ATG p53 CIS dn	307	2999	3306
ToxCast - BSK 3C Eselectin down	374	1028	1402
ToxCast - BSK 3C HLADR down	529	872	1401
ToxCast - BSK 3C IL8 down	315	1087	1402
ToxCast - BSK 3C MCP1 down	360	1042	1402
ToxCast - BSK 3C Proliferation down	558	841	1399
ToxCast - BSK 3C SRB down	447	953	1400
ToxCast - BSK 3C TissueFactor down	292	1110	1402
ToxCast - BSK 3C VCAM1 down	331	1071	1402
ToxCast - BSK 3C Vis down	425	971	1396
ToxCast - BSK 3C uPAR down	391	1011	1402

ToxCast - BSK 4H Eotaxin3 down	402	998	1400
ToxCast - BSK 4H MCP1 down	356	1039	1395
ToxCast - BSK 4H Pselectin down	373	1029	1402
ToxCast - BSK 4H SRB down	369	1032	1401
ToxCast - BSK 4H VCAM1 down	391	1010	1401
ToxCast - BSK 4H uPAR down	319	1081	1400
ToxCast - BSK BE3C HLADR down	363	1035	1398
ToxCast - BSK BE3C IL1a down	247	1155	1402
ToxCast - BSK BE3C IP10 down	347	1053	1400
ToxCast - BSK BE3C PAI1 down	257	1139	1396
ToxCast - BSK BE3C tPA down	210	1187	1397
ToxCast - BSK BE3C uPA down	213	1186	1399
ToxCast - BSK CASM3C HLADR down	259	1141	1400
ToxCast - BSK CASM3C MCP1 down	202	1200	1402
ToxCast - BSK CASM3C MCSF down	241	1159	1400
ToxCast - BSK CASM3C Proliferation down	429	970	1399
ToxCast - BSK CASM3C SRB down	235	1168	1403
ToxCast - BSK CASM3C Thrombomodulin up	225	1174	1399
ToxCast - BSK CASM3C VCAM1 down	259	1141	1400
ToxCast - BSK CASM3C uPAR down	232	1167	1399
ToxCast - BSK KF3CT ICAM1 down	205	1196	1401
ToxCast - BSK KF3CT IL1a down	269	1131	1400
ToxCast - BSK KF3CT IP10 down	295	1103	1398
ToxCast - BSK KF3CT MCP1 down	230	1168	1398
ToxCast - BSK KF3CT MMP9 down	359	1040	1399
ToxCast - BSK KF3CT SRB down	217	1182	1399
ToxCast - BSK KF3CT TGFb1 down	258	1138	1396
ToxCast - BSK KF3CT TIMP2 down	286	1111	1397
ToxCast - BSK KF3CT uPA down	241	1155	1396

ToxCast - BSK LPS CD40 down	384	1016	1400
ToxCast - BSK LPS Eselectin down	290	1112	1402
ToxCast - BSK LPS IL1a down	287	1114	1401
ToxCast - BSK LPS IL8 down	293	1102	1395
ToxCast - BSK LPS MCP1 down	310	1090	1400
ToxCast - BSK LPS MCSF down	373	1027	1400
ToxCast - BSK LPS PGE2 down	270	1128	1398
ToxCast - BSK LPS SRB down	346	1058	1404
ToxCast - BSK LPS TNFa down	275	1125	1400
ToxCast - BSK LPS TissueFactor down	211	1189	1400
ToxCast - BSK LPS VCAM1 down	401	1000	1401
ToxCast - BSK SAg CD38 down	418	981	1399
ToxCast - BSK SAg CD40 down	401	996	1397
ToxCast - BSK SAg CD69 down	372	1030	1402
ToxCast - BSK SAg Eselectin down	384	1015	1399
ToxCast - BSK SAg IL8 down	324	1076	1400
ToxCast - BSK SAg MCP1 down	328	1073	1401
ToxCast - BSK SAg MIG down	217	1185	1402
ToxCast - BSK SAg PBMCCytotoxicity down	302	1102	1404
ToxCast - BSK SAg Proliferation down	530	872	1402
ToxCast - BSK SAg SRB down	346	1055	1401
ToxCast - BSK hDFCGF CollagenIII down	398	998	1396
ToxCast - BSK hDFCGF IP10 down	389	1009	1398
ToxCast - BSK hDFCGF MCSF down	381	1017	1398
ToxCast - BSK hDFCGF MIG down	273	1129	1402
ToxCast - BSK hDFCGF MMP1 down	247	1154	1401
ToxCast - BSK hDFCGF PAI1 down	362	1037	1399
ToxCast - BSK hDFCGF Proliferation down	590	809	1399
ToxCast - BSK hDFCGF SRB down	308	1095	1403

ToxCast - BSK hDFCGF TIMP1 down	240	1160	1400
ToxCast - BSK hDFCGF VCAM1 down	377	1022	1399
ToxCast - LTEA HepaRG ABCB11 dn	326	677	1003
ToxCast - LTEA HepaRG ABCB1 up	245	757	1002
ToxCast - LTEA HepaRG ABCG2 up	246	757	1003
ToxCast - LTEA HepaRG ACOX1 dn	208	791	999
ToxCast - LTEA HepaRG AFP dn	351	651	1002
ToxCast - LTEA HepaRG ALPP dn	234	769	1003
ToxCast - LTEA HepaRG APOA5 dn	233	768	1001
ToxCast - LTEA HepaRG CAT dn	235	762	997
ToxCast - LTEA HepaRG CYP1A1 up	793	210	1003
ToxCast - LTEA HepaRG CYP1A2 up	478	523	1001
ToxCast - LTEA HepaRG CYP2B6 up	535	468	1003
ToxCast - LTEA HepaRG CYP2C19 up	327	675	1002
ToxCast - LTEA HepaRG CYP2C9 dn	206	793	999
ToxCast - LTEA HepaRG CYP2E1 dn	489	514	1003
ToxCast - LTEA HepaRG CYP3A4 up	329	673	1002
ToxCast - LTEA HepaRG CYP3A7 up	247	757	1004
ToxCast - LTEA HepaRG CYP4A11 dn	311	691	1002
ToxCast - LTEA HepaRG CYP4A22 dn	305	696	1001
ToxCast - LTEA HepaRG CYP7A1 dn	357	645	1002
ToxCast - LTEA HepaRG DDIT3 up	215	787	1002
ToxCast - LTEA HepaRG FABP1 dn	340	662	1002
ToxCast - LTEA HepaRG FASN dn	239	763	1002
ToxCast - LTEA HepaRG FMO3 dn	260	739	999
ToxCast - LTEA HepaRG GSTA2 dn	232	769	1001
ToxCast - LTEA HepaRG HMGCS2 dn	298	704	1002
ToxCast - LTEA HepaRG IGF1 dn	326	676	1002
ToxCast - LTEA HepaRG IGFBP1 up	238	764	1002

ToxCast - LTEA HepaRG KRT19 dn	261	740	1001
ToxCast - LTEA HepaRG LIPC dn	248	751	999
ToxCast - LTEA HepaRG MYC up	221	779	1000
ToxCast - LTEA HepaRG PEG10 dn	336	668	1004
ToxCast - LTEA HepaRG SLC22A1 dn	314	688	1002
ToxCast - LTEA HepaRG SLCO1B1 dn	203	797	1000
ToxCast - LTEA HepaRG THRSP dn	225	777	1002
ToxCast - LTEA HepaRG UGT1A1 up	366	638	1004
ToxCast - NHEERL ZF 144hpf TERATOSCORE up	471	210	681
ToxCast - OT AR ARSRC1 0480	261	1425	1686
ToxCast - OT AR ARSRC1 0960	377	1310	1687
ToxCast - OT ER ERaERb 0480	226	1460	1686
ToxCast - OT ER ERaERb 1440	300	1381	1681
ToxCast - OT ER ERbERb 0480	213	1472	1685
ToxCast - OT ER ERbERb 1440	229	1457	1686
ToxCast - OT FXR FXRSRC1 0480	366	1302	1668
ToxCast - OT FXR FXRSRC1 1440	325	1343	1668
ToxCast - TOX21 AP1 BLA Agonist ratio	879	5912	6791
ToxCast - TOX21 ARE BLA agonist ratio	1364	4974	6338
ToxCast - TOX21 AR BLA Agonist ratio	414	6736	7150
ToxCast - TOX21 AR BLA Antagonist ratio	1294	5697	6991
ToxCast - TOX21 AR LUC MDAKB2 Agonist	303	6860	7163
ToxCast - TOX21 AR LUC MDAKB2 Antagonist 0.5nM R1881	1314	5490	6804
ToxCast - TOX21 AR LUC MDAKB2 Antagonist 10nM R1881	810	6318	7128
ToxCast - TOX21 AhR LUC Agonist	632	6485	7117
ToxCast - TOX21 Aromatase Inhibition	897	6164	7061
ToxCast - TOX21 CAR Agonist	723	6110	6833

ToxCast - TOX21 CAR Antagonist	572	6243	6815
ToxCast - TOX21 DT40	2419	4462	6881
ToxCast - TOX21 DT40 100	2590	4299	6889
ToxCast - TOX21 DT40 657	2291	4527	6818
ToxCast - TOX21 ERR Agonist	254	6641	6895
ToxCast - TOX21 ERR Antagonist	1439	5303	6742
ToxCast - TOX21 ERa BLA Agonist ratio	331	6822	7153
ToxCast - TOX21 ERa BLA Antagonist ratio	886	6161	7047
ToxCast - TOX21 ERa LUC VM7 Agonist	903	6010	6913
ToxCast - TOX21 ERa LUC VM7 Antagonist 0.1nM E2	857	5923	6780
ToxCast - TOX21 ERa LUC VM7 Antagonist 0.5nM E2	738	6400	7138
ToxCast - TOX21 ERb BLA Antagonist ratio	1295	5471	6766
ToxCast - TOX21 ESRE BLA ratio	205	6361	6566
ToxCast - TOX21 FXR BLA antagonist ratio	901	5545	6446
ToxCast - TOX21 GR BLA Agonist ratio	362	6826	7188
ToxCast - TOX21 GR BLA Antagonist ratio	637	6512	7149
ToxCast - TOX21 H2AX HTRF CHO Agonist ratio	438	6400	6838
ToxCast - TOX21 HDAC Inhibition	461	6425	6886
ToxCast - TOX21 HRE BLA Agonist ratio	274	6599	6873
ToxCast - TOX21 HSE BLA agonist ratio	426	6078	6504
ToxCast - TOX21 MMP fitc	404	6757	7161
ToxCast - TOX21 MMP ratio down	1131	5975	7106
ToxCast - TOX21 MMP ratio up	281	6933	7214
ToxCast - TOX21 MMP rhodamine	952	6167	7119
ToxCast - TOX21 PGC ERR Agonist	279	6607	6886
ToxCast - TOX21 PGC ERR Antagonist	910	5863	6773
ToxCast - TOX21 PPARd BLA antagonist ratio	471	6041	6512
ToxCast - TOX21 PPARg BLA antagonist ratio	659	5833	6492

ToxCast - TOX21 PR BLA Antagonist ratio	1780	4983	6763
ToxCast - TOX21 RAR LUC Agonist	239	6327	6566
ToxCast - TOX21 RAR LUC Antagonist	720	6086	6806
ToxCast - TOX21 RORg LUC CHO Antagonist	695	6111	6806
ToxCast - TOX21 RXR BLA Agonist ratio	276	6502	6778
ToxCast - TOX21 SBE BLA Antagonist ratio	949	5875	6824
ToxCast - TOX21 SSH 3T3 GLI3 Antagonist	1286	5399	6685
ToxCast - TOX21 TR LUC GH3 Antagonist	1899	5150	7049
ToxCast - TOX21 TSHR Agonist ratio	354	6516	6870
ToxCast - TOX21 TSHR Antagonist ratio	242	6651	6893
ToxCast - TOX21 VDR BLA antagonist ratio	388	6145	6533
ToxCast - TOX21 p53 BLA p1 ratio	574	6553	7127
ToxCast - TOX21 p53 BLA p2 ratio	705	6398	7103
ToxCast - TOX21 p53 BLA p3 ratio	611	6511	7122
ToxCast - TOX21 p53 BLA p4 ratio	670	6445	7115
ToxCast - TOX21 p53 BLA p5 ratio	638	6481	7119
ToxCast - UPITT HCI U2OS AR TIF2 Nucleoli Agonist	208	1580	1788
ToxCast - UPITT HCI U2OS AR TIF2 Nucleoli Antagonist	443	1344	1787
ToxCast - UPITT HCI U2OS AR TIF2 Nucleoli Cytoplasm Ratio Antagonist	224	1567	1791
ToxCast - NCCT HEK293T CellTiterGLO	285	246	531

Table S4. List of Molecular Descriptors Used in Principal Component Analysis.

Descriptors
SlogP
TPSA
ExactMW
NumLipinskiHBA
NumLipinskiHBD
NumRotatableBonds
NumHBD
NumHBA
NumAmideBonds
NumHeteroAtoms
NumHeavyAtoms
NumAtoms
NumStereocenters
NumUnspecifiedStereocenters
NumRings
NumAromaticRings
NumSaturatedRings
NumAliphaticRings
NumAromaticHeterocycles
NumSaturatedHeterocycles
NumAliphaticHeterocycles
NumAromaticCarbocycles
NumSaturatedCarbocycles
NumAliphaticCarbocycles
FractionCSP3

Table S5. Hyperparameters Used for Deriving the CP Models.

Step	Implementation	Function	Hyperparameter[1]
variance filter	scikit-learn	VarianceThreshold	threshold=0.0015
CV split	scikit-learn	StratifiedShuffleSplit	n_splits=5, test_size=0.2, random_state=2020
CP framework	nonconformist python package	NcFactory	err_func= InverseProbabilityErrFunc() , normalizer_model=None
		IcpClassifier	condition=(lambda instance: instance[1])
		RandomSubSampler	calibration_portion=0.3
		AggregatedCp	n_models=20, sampler=RandomSubSampler(), aggregation_func=(lambda x: np.median(x,axis=2))
random forest model	scikit-learn	RandomForestClassifier	n_estimators=500, random_state=2020
lasso model	scikit-learn	LassoCV	cv=5, random_state=2020, max_iter=1000000

[1] Hyperparameters not specified here were kept with the default values.

Table S6. Average Performance of the CP Models for the Biological Assays.[1]

Endpoint	Validity	Efficiency	Overall accuracy	Accuracy active class	Accuracy inactive class	F1 score
AMES	0.80 (+/-0.01)	0.93 (+/-0.01)	0.87 (+/-0.01)	0.87 (+/-0.01)	0.87 (+/-0.01)	0.86 (+/-0.01)
Mammalian cell gene mutation	0.82 (+/-0.03)	0.51 (+/-0.04)	0.65 (+/-0.07)	0.67 (+/-0.23)	0.64 (+/-0.08)	0.49 (+/-0.03)
Chromosome aberration	0.79 (+/-0.03)	0.70 (+/-0.02)	0.70 (+/-0.04)	0.72 (+/-0.06)	0.69 (+/-0.05)	0.60 (+/-0.03)
Human oral bioavailability	0.82 (+/-0.04)	0.79 (+/-0.03)	0.77 (+/-0.04)	0.79 (+/-0.07)	0.75 (+/-0.05)	0.77 (+/-0.04)
PGP inhibition	0.81 (+/-0.02)	0.86 (+/-0.02)	0.95 (+/-0.01)	0.94 (+/-0.02)	0.95 (+/-0.01)	0.95 (+/-0.01)
Caco-2	0.83 (+/-0.02)	0.96 (+/-0.01)	0.86 (+/-0.02)	0.87 (+/-0.04)	0.86 (+/-0.03)	0.85 (+/-0.01)
DIO1 inhibition	0.79 (+/-0.03)	0.93 (+/-0.03)	0.78 (+/-0.03)	0.77 (+/-0.14)	0.78 (+/-0.03)	0.59 (+/-0.04)
DIO2 inhibition	0.81 (+/-0.01)	0.91 (+/-0.01)	0.79 (+/-0.02)	0.79 (+/-0.06)	0.80 (+/-0.02)	0.66 (+/-0.02)
DIO3 inhibition	0.83 (+/-0.02)	0.89 (+/-0.01)	0.81 (+/-0.02)	0.77 (+/-0.06)	0.81 (+/-0.02)	0.67 (+/-0.03)
TPO inhibition	0.79 (+/-0.03)	0.96 (+/-0.04)	0.82 (+/-0.03)	0.82 (+/-0.04)	0.82 (+/-0.04)	0.79 (+/-0.03)
TR antagonism	0.80 (+/-0.01)	0.94 (+/-0.01)	0.85 (+/-0.01)	0.87 (+/-0.01)	0.85 (+/-0.01)	0.80 (+/-0.01)
NIS inhibition	0.79 (+/-0.04)	0.88 (+/-0.07)	0.76 (+/-0.03)	0.78 (+/-0.19)	0.76 (+/-0.04)	0.57 (+/-0.04)
TRHR antagonism	0.82 (+/-0.01)	0.95 (+/-0.02)	0.83 (+/-0.02)	0.80 (+/-0.12)	0.83 (+/-0.02)	0.49 (+/-0.01)
TSHR antagonism	0.81 (+/-0.01)	0.86 (+/-0.03)	0.78 (+/-0.01)	0.85 (+/-0.08)	0.78 (+/-0.01)	0.49 (+/-0.01)
TSHR agonism	0.80 (+/-0.02)	0.92 (+/-0.03)	0.79 (+/-0.02)	0.84 (+/-0.03)	0.79 (+/-0.02)	0.53 (+/-0.01)

eMolTox - Modulator of Muscarinic acetylcholine receptor M5	0.81 (+/-0.01)	0.85 (+/-0.01)	0.96 (+/-0.01)	0.94 (+/-0.01)	0.96 (+/-0.01)	0.93 (+/-0.02)
eMolTox - Antagonist of the thyroid receptor (TR) signaling pathway	0.80 (+/-0.01)	0.94 (+/-0.02)	0.79 (+/-0.00)	0.78 (+/-0.05)	0.79 (+/-0.01)	0.60 (+/-0.01)
eMolTox - Modulator of Glucocorticoid receptor	0.81 (+/-0.01)	0.81 (+/-0.01)	1.00 (+/-0.00)	1.00 (+/-0.00)	1.00 (+/-0.00)	1.00 (+/-0.00)
eMolTox - Mutagenicity	0.81 (+/-0.01)	0.98 (+/-0.01)	0.83 (+/-0.01)	0.83 (+/-0.00)	0.83 (+/-0.01)	0.83 (+/-0.01)
eMolTox - Modulator of GABA-A receptor alpha-5beta-3gamma-2	0.82 (+/-0.02)	0.82 (+/-0.02)	1.00 (+/-0.00)	1.00 (+/-0.00)	1.00 (+/-0.00)	1.00 (+/-0.00)
eMolTox - Modulator of Muscarinic acetylcholine receptor M2	0.80 (+/-0.02)	0.81 (+/-0.02)	0.99 (+/-0.00)	0.97 (+/-0.01)	0.99 (+/-0.00)	0.98 (+/-0.01)
eMolTox - Modulator of Neuropeptide Y receptor type 1	0.82 (+/-0.02)	0.82 (+/-0.02)	1.00 (+/-0.00)	1.00 (+/-0.00)	1.00 (+/-0.00)	1.00 (+/-0.00)
eMolTox - Modulator of Cannabinoid CB2 receptor	0.81 (+/-0.01)	0.82 (+/-0.01)	0.99 (+/-0.00)	1.00 (+/-0.00)	0.99 (+/-0.00)	0.99 (+/-0.00)
eMolTox - Modulator of Androgen Receptor	0.81 (+/-0.01)	0.82 (+/-0.01)	1.00 (+/-0.00)	0.99 (+/-0.01)	1.00 (+/-0.00)	1.00 (+/-0.00)
eMolTox - Modulator of Monoamine oxidase A	0.81 (+/-0.01)	0.82 (+/-0.01)	0.99 (+/-0.00)	0.97 (+/-0.02)	0.99 (+/-0.01)	0.97 (+/-0.01)
eMolTox - Agonist of the thyroid stimulating hormone receptor (TSHR) signaling pathway	0.80 (+/-0.01)	0.89 (+/-0.03)	0.77 (+/-0.01)	0.78 (+/-0.04)	0.77 (+/-0.01)	0.55 (+/-0.01)
eMolTox - Modulator of Cholecystokinin A receptor	0.80 (+/-0.04)	0.80 (+/-0.04)	0.99 (+/-0.01)	0.97 (+/-0.04)	1.00 (+/-0.00)	0.99 (+/-0.01)

eMolTox - Modulator of Neurokinin 1 receptor	0.81 (+/- 0.01)	0.81 (+/- 0.01)	1.00 (+/- 0.00)	1.00 (+/- 0.00)	1.00 (+/- 0.00)	1.00 (+/- 0.00)
eMolTox - Modulator of Serotonin 2c (5-HT2c) receptor	0.81 (+/- 0.01)	0.81 (+/- 0.01)	1.00 (+/- 0.00)	0.99 (+/- 0.01)	1.00 (+/- 0.00)	0.99 (+/- 0.00)
eMolTox - Modulator of GABA-A receptor alpha-2beta-3gamma-2	0.80 (+/- 0.02)	0.80 (+/- 0.02)	1.00 (+/- 0.00)	1.00 (+/- 0.01)	1.00 (+/- 0.00)	1.00 (+/- 0.00)
eMolTox - Modulator of Acetylcholinesterase	0.81 (+/- 0.01)	0.82 (+/- 0.01)	0.99 (+/- 0.00)	0.98 (+/- 0.01)	0.99 (+/- 0.00)	0.98 (+/- 0.00)
eMolTox - Modulator of Cannabinoid CB1 receptor	0.81 (+/- 0.01)	0.82 (+/- 0.01)	1.00 (+/- 0.00)	0.99 (+/- 0.00)	1.00 (+/- 0.00)	0.99 (+/- 0.00)
eMolTox - Modulator of Neurokinin 2 receptor	0.81 (+/- 0.01)	0.81 (+/- 0.01)	1.00 (+/- 0.00)	0.99 (+/- 0.01)	1.00 (+/- 0.00)	1.00 (+/- 0.00)
eMolTox - Modulator of Histamine H1 receptor	0.80 (+/- 0.02)	0.81 (+/- 0.01)	0.99 (+/- 0.00)	0.98 (+/- 0.01)	0.99 (+/- 0.00)	0.99 (+/- 0.01)
eMolTox - Modulator of Alpha-1a adrenergic receptor	0.80 (+/- 0.01)	0.80 (+/- 0.01)	1.00 (+/- 0.00)	0.99 (+/- 0.01)	1.00 (+/- 0.00)	0.99 (+/- 0.00)
eMolTox - Antagonist of the androgen receptor (AR) signaling pathway	0.80 (+/- 0.02)	0.99 (+/- 0.00)	0.80 (+/- 0.01)	0.81 (+/- 0.01)	0.80 (+/- 0.01)	0.64 (+/- 0.01)
eMolTox - Modulator of Neuronal acetylcholine receptor protein alpha-7 subunit	0.82 (+/- 0.02)	0.84 (+/- 0.01)	0.98 (+/- 0.00)	0.98 (+/- 0.02)	0.98 (+/- 0.00)	0.96 (+/- 0.01)
eMolTox - Activator Alzheimers amyloid precursor	0.81 (+/- 0.01)	0.90 (+/- 0.01)	0.89 (+/- 0.01)	0.92 (+/- 0.01)	0.89 (+/- 0.01)	0.78 (+/- 0.01)
eMolTox - Differential cytotoxicity (isogenic chicken DT40 Rev3 mutant cell line)	0.80 (+/- 0.01)	0.94 (+/- 0.01)	0.79 (+/- 0.01)	0.79 (+/- 0.03)	0.78 (+/- 0.01)	0.76 (+/- 0.01)

eMolTox - Antagonist of the estrogen receptor alpha (ER-alpha) signaling pathway dup	0.80 (+/-0.02)	0.96 (+/-0.01)	0.79 (+/-0.02)	0.82 (+/-0.03)	0.79 (+/-0.02)	0.59 (+/-0.01)
eMolTox - Modulator of HERG	0.81 (+/-0.00)	0.86 (+/-0.01)	0.95 (+/-0.01)	0.92 (+/-0.01)	0.95 (+/-0.01)	0.91 (+/-0.01)
eMolTox - Antagonist of the retinoic acid receptor (RAR) signaling pathway	0.81 (+/-0.01)	0.90 (+/-0.01)	0.79 (+/-0.01)	0.83 (+/-0.02)	0.79 (+/-0.01)	0.66 (+/-0.01)
eMolTox - Modulator of Serotonin 2b (5-HT2b) receptor	0.81 (+/-0.01)	0.82 (+/-0.02)	0.99 (+/-0.00)	0.96 (+/-0.01)	1.00 (+/-0.00)	0.98 (+/-0.00)
eMolTox - Modulator of Platelet-derived growth factor receptor beta	0.81 (+/-0.01)	0.84 (+/-0.01)	0.97 (+/-0.01)	0.94 (+/-0.03)	0.98 (+/-0.00)	0.95 (+/-0.01)
eMolTox - Antagonist of the retinoid-related orphan receptor gamma (ROR-gamma) signaling pathway	0.81 (+/-0.01)	0.98 (+/-0.01)	0.80 (+/-0.01)	0.81 (+/-0.02)	0.80 (+/-0.01)	0.66 (+/-0.01)
eMolTox - Modulator of Serotonin 1a (5-HT1a) receptor	0.81 (+/-0.01)	0.82 (+/-0.01)	0.99 (+/-0.00)	0.99 (+/-0.01)	0.99 (+/-0.00)	0.99 (+/-0.00)
eMolTox - Modulator of Neuronal acetylcholine receptor alpha4beta2	0.81 (+/-0.03)	0.82 (+/-0.03)	0.99 (+/-0.01)	0.99 (+/-0.01)	0.99 (+/-0.01)	0.98 (+/-0.01)
eMolTox - Agonist of H2AX	0.82 (+/-0.01)	0.86 (+/-0.03)	0.79 (+/-0.01)	0.78 (+/-0.06)	0.79 (+/-0.01)	0.59 (+/-0.01)
eMolTox - Modulator of GABA-A receptor alpha-3beta-3gamma-2	0.83 (+/-0.01)	0.83 (+/-0.01)	1.00 (+/-0.00)	1.00 (+/-0.00)	1.00 (+/-0.00)	1.00 (+/-0.00)
eMolTox - Induce genotoxicity in human embryonic kidney cells	0.80 (+/-0.00)	0.91 (+/-0.03)	0.78 (+/-0.01)	0.84 (+/-0.05)	0.78 (+/-0.01)	0.55 (+/-0.01)

eMolTox - Modulator of Glutamate NMDA receptor	0.80 (+/- 0.03)	0.81 (+/- 0.03)	0.98 (+/- 0.00)	0.95 (+/- 0.02)	0.99 (+/- 0.00)	0.97 (+/- 0.00)
eMolTox - Modulator of Norepinephrine transporter	0.81 (+/- 0.00)	0.81 (+/- 0.00)	1.00 (+/- 0.00)	0.99 (+/- 0.00)	1.00 (+/- 0.00)	0.99 (+/- 0.00)
eMolTox - Modulator of Serotonin transporter	0.81 (+/- 0.01)	0.82 (+/- 0.01)	0.99 (+/- 0.00)	0.99 (+/- 0.00)	0.99 (+/- 0.00)	0.99 (+/- 0.00)
eMolTox - Modulator of Sodium channel protein type IX alpha subunit	0.82 (+/- 0.00)	0.82 (+/- 0.00)	1.00 (+/- 0.00)	1.00 (+/- 0.00)	1.00 (+/- 0.00)	1.00 (+/- 0.00)
eMolTox - Agonist of the AP-1 signaling pathway	0.79 (+/- 0.02)	0.90 (+/- 0.02)	0.77 (+/- 0.02)	0.79 (+/- 0.03)	0.77 (+/- 0.02)	0.62 (+/- 0.02)
eMolTox - Agonist of the p53 signaling pathway	0.80 (+/- 0.01)	0.93 (+/- 0.01)	0.78 (+/- 0.01)	0.79 (+/- 0.05)	0.78 (+/- 0.01)	0.61 (+/- 0.01)
eMolTox - Modulator of Serotonin 3a (5-HT3a) receptor	0.83 (+/- 0.02)	0.84 (+/- 0.02)	0.99 (+/- 0.00)	0.96 (+/- 0.02)	1.00 (+/- 0.00)	0.99 (+/- 0.01)
eMolTox - Modulator of Serotonin 1b (5-HT1b) receptor	0.82 (+/- 0.01)	0.82 (+/- 0.01)	1.00 (+/- 0.00)	0.98 (+/- 0.01)	1.00 (+/- 0.00)	0.99 (+/- 0.00)
eMolTox - Modulator of Delta opioid receptor	0.81 (+/- 0.01)	0.81 (+/- 0.01)	1.00 (+/- 0.00)	1.00 (+/- 0.00)	1.00 (+/- 0.00)	1.00 (+/- 0.00)
eMolTox - Modulator of Vascular endothelial growth factor receptor 1	0.81 (+/- 0.01)	0.85 (+/- 0.01)	0.95 (+/- 0.01)	0.90 (+/- 0.02)	0.96 (+/- 0.01)	0.91 (+/- 0.01)
eMolTox - Modulator of Beta-2 adrenergic receptor	0.82 (+/- 0.01)	0.83 (+/- 0.01)	0.99 (+/- 0.00)	0.99 (+/- 0.01)	0.99 (+/- 0.00)	0.98 (+/- 0.01)
eMolTox - Modulator of Muscarinic acetylcholine receptor M4	0.82 (+/- 0.02)	0.84 (+/- 0.02)	0.98 (+/- 0.01)	0.96 (+/- 0.02)	0.99 (+/- 0.01)	0.97 (+/- 0.02)
eMolTox - Modulator of TNF-alpha	0.82 (+/- 0.01)	0.83 (+/- 0.01)	1.00 (+/- 0.00)	0.97 (+/- 0.02)	1.00 (+/- 0.00)	0.99 (+/- 0.00)

eMolTox - Modulator of Calcitonin gene-related peptide type 1 receptor	0.81 (+/-0.01)	0.81 (+/-0.01)	1.00 (+/-0.00)	1.00 (+/-0.00)	1.00 (+/-0.00)	1.00 (+/-0.00)
eMolTox - Modulator of Alpha-2a adrenergic receptor	0.82 (+/-0.02)	0.83 (+/-0.02)	0.98 (+/-0.00)	0.96 (+/-0.03)	0.99 (+/-0.01)	0.96 (+/-0.01)
eMolTox - Modulator of Sigma opioid receptor	0.80 (+/-0.01)	0.81 (+/-0.02)	0.99 (+/-0.00)	0.98 (+/-0.01)	0.99 (+/-0.00)	0.98 (+/-0.00)
eMolTox - Differential cytotoxicity against isogenic chicken DT40 cell lines with known DNA damage response pathways - Rad54Ku70 mutant cell line	0.80 (+/-0.01)	0.98 (+/-0.01)	0.80 (+/-0.01)	0.81 (+/-0.02)	0.79 (+/-0.01)	0.77 (+/-0.01)
eMolTox - Modulator of GABA-A receptor alpha-1beta-3gamma-2	0.81 (+/-0.01)	0.81 (+/-0.01)	1.00 (+/-0.00)	1.00 (+/-0.00)	1.00 (+/-0.00)	1.00 (+/-0.00)
eMolTox - Modulator of Dopamine transporter	0.80 (+/-0.02)	0.81 (+/-0.02)	0.99 (+/-0.00)	0.98 (+/-0.01)	0.99 (+/-0.00)	0.98 (+/-0.00)
eMolTox - Modulator of GABA-A receptor anion channel	0.81 (+/-0.01)	0.81 (+/-0.01)	1.00 (+/-0.00)	0.99 (+/-0.01)	1.00 (+/-0.00)	0.99 (+/-0.00)
eMolTox - Differential cytotoxicity (isogenic chicken DT40 cell lines)	0.81 (+/-0.01)	0.99 (+/-0.00)	0.81 (+/-0.00)	0.81 (+/-0.03)	0.81 (+/-0.01)	0.78 (+/-0.01)
eMolTox - Cytotoxicity in HEK293 cells - 16 hour	0.80 (+/-0.01)	0.77 (+/-0.02)	0.74 (+/-0.01)	0.77 (+/-0.06)	0.74 (+/-0.01)	0.53 (+/-0.01)
eMolTox - Aromatase inhibitors	0.80 (+/-0.01)	0.90 (+/-0.02)	0.78 (+/-0.01)	0.79 (+/-0.09)	0.78 (+/-0.01)	0.57 (+/-0.01)
eMolTox - Modulators of myocardial damage	0.80 (+/-0.01)	0.91 (+/-0.01)	0.89 (+/-0.00)	0.88 (+/-0.01)	0.89 (+/-0.01)	0.76 (+/-0.01)

eMolTox - Antagonist of the androgen receptor (AR) signaling pathway dup	0.81 (+/- 0.02)	0.89 (+/- 0.02)	0.78 (+/- 0.01)	0.80 (+/- 0.03)	0.78 (+/- 0.01)	0.59 (+/- 0.02)
eMolTox - Modulator of Adenosine A2a receptor	0.81 (+/- 0.01)	0.81 (+/- 0.01)	1.00 (+/- 0.00)	0.99 (+/- 0.01)	1.00 (+/- 0.00)	0.99 (+/- 0.00)
eMolTox - Modulator of Serotonin 2a (5-HT2a) receptor	0.81 (+/- 0.01)	0.82 (+/- 0.01)	0.99 (+/- 0.00)	0.99 (+/- 0.00)	1.00 (+/- 0.00)	0.99 (+/- 0.00)
eMolTox - Modulator of Beta-3 adrenergic receptor	0.82 (+/- 0.01)	0.82 (+/- 0.01)	1.00 (+/- 0.00)	1.00 (+/- 0.00)	1.00 (+/- 0.00)	1.00 (+/- 0.00)
eMolTox - Agonist of the estrogen receptor alpha (ER-alpha) signaling pathway	0.81 (+/- 0.01)	0.79 (+/- 0.02)	0.75 (+/- 0.01)	0.80 (+/- 0.03)	0.75 (+/- 0.01)	0.58 (+/- 0.01)
eMolTox - Modulator of Kappa opioid receptor	0.81 (+/- 0.01)	0.81 (+/- 0.01)	0.99 (+/- 0.00)	0.99 (+/- 0.01)	1.00 (+/- 0.00)	0.99 (+/- 0.00)
eMolTox - Modulator of Adenosine A3 receptor	0.81 (+/- 0.00)	0.82 (+/- 0.01)	0.99 (+/- 0.00)	0.98 (+/- 0.01)	0.99 (+/- 0.00)	0.98 (+/- 0.00)
eMolTox - Modulator of Serotonin 7 (5-HT7) receptor	0.81 (+/- 0.01)	0.81 (+/- 0.01)	0.99 (+/- 0.00)	0.99 (+/- 0.00)	1.00 (+/- 0.00)	0.99 (+/- 0.00)
eMolTox - Modulator of Beta-1 adrenergic receptor	0.81 (+/- 0.02)	0.81 (+/- 0.02)	1.00 (+/- 0.00)	0.99 (+/- 0.01)	1.00 (+/- 0.00)	0.99 (+/- 0.00)
eMolTox - Antagonist of the estrogen receptor alpha (ER-alpha) signaling pathway	0.80 (+/- 0.01)	0.97 (+/- 0.01)	0.80 (+/- 0.01)	0.77 (+/- 0.05)	0.80 (+/- 0.01)	0.59 (+/- 0.01)
eMolTox - Modulator of Adenosine A2b receptor	0.81 (+/- 0.01)	0.81 (+/- 0.01)	1.00 (+/- 0.00)	0.99 (+/- 0.00)	1.00 (+/- 0.00)	1.00 (+/- 0.00)
eMolTox - Modulator of Vascular endothelial growth factor receptor 3	0.82 (+/- 0.02)	0.94 (+/- 0.01)	0.87 (+/- 0.01)	0.87 (+/- 0.02)	0.87 (+/- 0.02)	0.81 (+/- 0.01)

eMolTox - Modulator of Vascular endothelial growth factor receptor 2	0.81 (+/- 0.00)	0.83 (+/- 0.00)	0.98 (+/- 0.00)	0.97 (+/- 0.00)	0.99 (+/- 0.00)	0.98 (+/- 0.00)
eMolTox - Modulator of Receptor protein-tyrosine kinase erbB-2	0.80 (+/- 0.01)	0.81 (+/- 0.01)	0.99 (+/- 0.00)	0.98 (+/- 0.02)	1.00 (+/- 0.00)	0.99 (+/- 0.01)
eMolTox - Modulator of Alpha-2b adrenergic receptor	0.82 (+/- 0.01)	0.85 (+/- 0.01)	0.97 (+/- 0.01)	0.95 (+/- 0.03)	0.98 (+/- 0.01)	0.95 (+/- 0.01)
eMolTox - Modulator of Serotonin 4 (5-HT4) receptor	0.83 (+/- 0.02)	0.83 (+/- 0.02)	1.00 (+/- 0.00)	1.00 (+/- 0.01)	1.00 (+/- 0.00)	1.00 (+/- 0.00)
eMolTox - Modulator of Platelet-derived growth factor receptor alpha	0.80 (+/- 0.02)	0.91 (+/- 0.01)	0.89 (+/- 0.01)	0.89 (+/- 0.03)	0.88 (+/- 0.02)	0.83 (+/- 0.02)
eMolTox - Modulator of Melatonin receptor 1B	0.80 (+/- 0.01)	0.81 (+/- 0.01)	1.00 (+/- 0.00)	1.00 (+/- 0.00)	1.00 (+/- 0.00)	1.00 (+/- 0.00)
eMolTox - Antagonist of the glucocorticoid receptor (GR) signaling pathway	0.82 (+/- 0.01)	0.97 (+/- 0.02)	0.81 (+/- 0.01)	0.83 (+/- 0.04)	0.81 (+/- 0.01)	0.62 (+/- 0.01)
eMolTox - Agonist of the androgen receptor (AR) signaling pathway	0.81 (+/- 0.02)	0.90 (+/- 0.05)	0.79 (+/- 0.03)	0.81 (+/- 0.05)	0.79 (+/- 0.03)	0.55 (+/- 0.02)
eMolTox - Cytotoxicity in HEK293 cells - 32 hour	0.80 (+/- 0.01)	0.85 (+/- 0.02)	0.76 (+/- 0.02)	0.78 (+/- 0.06)	0.76 (+/- 0.02)	0.60 (+/- 0.01)
eMolTox - Modulator of Muscarinic acetylcholine receptor M1	0.81 (+/- 0.01)	0.82 (+/- 0.01)	0.99 (+/- 0.00)	0.98 (+/- 0.02)	0.99 (+/- 0.01)	0.98 (+/- 0.01)
eMolTox - Modulator of Dopamine D2 receptor	0.81 (+/- 0.01)	0.82 (+/- 0.01)	0.99 (+/- 0.00)	0.99 (+/- 0.00)	0.99 (+/- 0.00)	0.99 (+/- 0.00)
eMolTox - Modulator of Angiotensin II type 2 (AT-2) receptor	0.82 (+/- 0.01)	0.82 (+/- 0.01)	1.00 (+/- 0.00)	1.00 (+/- 0.01)	1.00 (+/- 0.00)	1.00 (+/- 0.00)

eMolTox - Modulator of Urotensin II receptor	0.82 (+/-0.02)	0.82 (+/-0.02)	1.00 (+/-0.00)	1.00 (+/-0.00)	1.00 (+/-0.00)	1.00 (+/-0.00)
eMolTox - Modulator of P2X purinoceptor 7	0.81 (+/-0.01)	0.81 (+/-0.01)	1.00 (+/-0.00)	1.00 (+/-0.00)	1.00 (+/-0.00)	1.00 (+/-0.00)
eMolTox - Agonist of the antioxidant response element (ARE) signaling pathway	0.80 (+/-0.01)	0.91 (+/-0.02)	0.78 (+/-0.00)	0.80 (+/-0.04)	0.78 (+/-0.00)	0.70 (+/-0.01)
eMolTox - Block Bile Salt Export Pump	0.83 (+/-0.04)	0.93 (+/-0.02)	0.89 (+/-0.02)	0.90 (+/-0.03)	0.89 (+/-0.04)	0.89 (+/-0.02)
eMolTox - Cytotoxicity in HepG2 cells - 8 hour	0.81 (+/-0.01)	0.67 (+/-0.05)	0.71 (+/-0.03)	0.75 (+/-0.10)	0.71 (+/-0.03)	0.52 (+/-0.01)
eMolTox - Agonist of the farnesoid-X-receptor (FXR) signaling pathway	0.81 (+/-0.02)	0.98 (+/-0.01)	0.81 (+/-0.02)	0.79 (+/-0.04)	0.81 (+/-0.02)	0.63 (+/-0.02)
eMolTox - Antagonist of the peroxisome proliferator-activated receptor gamma (PPARg) signaling pathway	0.80 (+/-0.02)	0.95 (+/-0.03)	0.79 (+/-0.02)	0.78 (+/-0.07)	0.79 (+/-0.02)	0.60 (+/-0.01)
eMolTox - Cytotoxicity in HepG2 cells - 24 hour	0.80 (+/-0.02)	0.85 (+/-0.03)	0.77 (+/-0.02)	0.81 (+/-0.03)	0.76 (+/-0.02)	0.63 (+/-0.01)
eMolTox - Modulator of Adenosine A1 receptor	0.81 (+/-0.01)	0.82 (+/-0.01)	0.99 (+/-0.00)	0.99 (+/-0.00)	0.99 (+/-0.00)	0.99 (+/-0.00)
eMolTox - Agonist of Liver X receptor alpha	0.80 (+/-0.01)	0.80 (+/-0.01)	1.00 (+/-0.00)	1.00 (+/-0.00)	1.00 (+/-0.00)	1.00 (+/-0.00)
eMolTox - Modulator of Dopamine D1 receptor	0.81 (+/-0.01)	0.81 (+/-0.01)	1.00 (+/-0.00)	0.99 (+/-0.01)	1.00 (+/-0.00)	0.99 (+/-0.00)
eMolTox - Modulator of Peroxisome proliferator-activated receptor gamma	0.81 (+/-0.01)	0.81 (+/-0.01)	1.00 (+/-0.00)	0.99 (+/-0.00)	0.99 (+/-0.00)	0.99 (+/-0.00)

eMolTox - Agonist of the peroxisome proliferator-activated receptor gamma (PPARg) signaling pathway	0.81 (+/- 0.02)	0.87 (+/- 0.02)	0.78 (+/- 0.02)	0.81 (+/- 0.02)	0.78 (+/- 0.02)	0.54 (+/- 0.02)
eMolTox - Cytotoxicity in HEK293 cells - 40 hour	0.80 (+/- 0.02)	0.85 (+/- 0.01)	0.77 (+/- 0.02)	0.80 (+/- 0.04)	0.76 (+/- 0.02)	0.64 (+/- 0.02)
eMolTox - Modulator of Endothelin receptor ET-B	0.80 (+/- 0.01)	0.81 (+/- 0.01)	0.99 (+/- 0.00)	0.99 (+/- 0.01)	0.99 (+/- 0.01)	0.99 (+/- 0.01)
eMolTox - Induce Phospholipidosis	0.82 (+/- 0.04)	0.73 (+/- 0.03)	0.75 (+/- 0.05)	0.78 (+/- 0.06)	0.74 (+/- 0.06)	0.73 (+/- 0.05)
eMolTox - Inhibitors of Hepatocyte nuclear factor 4 (HNF4) dimerization	0.81 (+/- 0.03)	0.94 (+/- 0.02)	0.80 (+/- 0.03)	0.80 (+/- 0.04)	0.80 (+/- 0.05)	0.75 (+/- 0.03)
eMolTox - Cytotoxicity in HEK293 cells - 8 hour	0.80 (+/- 0.02)	0.65 (+/- 0.04)	0.69 (+/- 0.02)	0.74 (+/- 0.11)	0.69 (+/- 0.02)	0.48 (+/- 0.01)
eMolTox - Modulator of Bradykinin B2 receptor	0.82 (+/- 0.02)	0.82 (+/- 0.02)	1.00 (+/- 0.00)	1.00 (+/- 0.00)	1.00 (+/- 0.00)	1.00 (+/- 0.00)
eMolTox - Inhibit CYP2C19 Activity	0.81 (+/- 0.00)	0.80 (+/- 0.00)	0.76 (+/- 0.00)	0.76 (+/- 0.01)	0.76 (+/- 0.01)	0.76 (+/- 0.00)
eMolTox - Disruptors of the mitochondrial membrane potential	0.80 (+/- 0.01)	0.92 (+/- 0.01)	0.87 (+/- 0.01)	0.86 (+/- 0.03)	0.87 (+/- 0.01)	0.80 (+/- 0.01)
eMolTox - Modulator of Endothelin receptor ET-A	0.82 (+/- 0.02)	0.84 (+/- 0.01)	0.97 (+/- 0.01)	0.98 (+/- 0.01)	0.97 (+/- 0.00)	0.97 (+/- 0.01)
eMolTox - Modulator of Mu opioid receptor	0.81 (+/- 0.00)	0.81 (+/- 0.00)	1.00 (+/- 0.00)	0.99 (+/- 0.00)	1.00 (+/- 0.00)	0.99 (+/- 0.00)
eMolTox - Agonist of the RXR signaling pathway	0.81 (+/- 0.01)	0.42 (+/- 0.02)	0.54 (+/- 0.04)	0.85 (+/- 0.06)	0.53 (+/- 0.04)	0.41 (+/- 0.02)

eMolTox - Modulator of Muscarinic acetylcholine receptor M3	0.81 (+/- 0.01)	0.82 (+/- 0.01)	1.00 (+/- 0.00)	0.98 (+/- 0.01)	1.00 (+/- 0.00)	0.99 (+/- 0.00)
eMolTox - Antagonist of the vitamin D receptor (VDR) signaling pathway	0.82 (+/- 0.01)	0.99 (+/- 0.01)	0.82 (+/- 0.01)	0.80 (+/- 0.04)	0.82 (+/- 0.02)	0.61 (+/- 0.01)
eMolTox - Modulator of P2X purinoceptor 3	0.79 (+/- 0.02)	0.79 (+/- 0.02)	1.00 (+/- 0.00)	1.00 (+/- 0.00)	1.00 (+/- 0.00)	1.00 (+/- 0.00)
eMolTox - Inhibit CYP1A2 Activity	0.82 (+/- 0.01)	0.98 (+/- 0.01)	0.83 (+/- 0.00)	0.83 (+/- 0.02)	0.83 (+/- 0.02)	0.83 (+/- 0.00)
eMolTox - Cytotoxicity in HepG2 cells - 40 hour	0.81 (+/- 0.01)	0.86 (+/- 0.02)	0.78 (+/- 0.01)	0.78 (+/- 0.04)	0.78 (+/- 0.01)	0.67 (+/- 0.02)
eMolTox - Cytotoxicity in HEK293 cells - 24 hour	0.80 (+/- 0.01)	0.85 (+/- 0.01)	0.76 (+/- 0.01)	0.79 (+/- 0.03)	0.76 (+/- 0.01)	0.58 (+/- 0.01)
eMolTox - Substrates of Cytochrome P450 2C19	0.80 (+/- 0.01)	0.99 (+/- 0.00)	0.80 (+/- 0.01)	0.81 (+/- 0.02)	0.80 (+/- 0.01)	0.76 (+/- 0.01)
eMolTox - Modulator of Alpha-1b adrenergic receptor	0.81 (+/- 0.01)	0.81 (+/- 0.01)	0.99 (+/- 0.00)	0.99 (+/- 0.01)	1.00 (+/- 0.00)	0.99 (+/- 0.00)
eMolTox - Antagonist of the farnesoid-X-receptor (FXR) signaling pathway	0.80 (+/- 0.01)	0.96 (+/- 0.02)	0.79 (+/- 0.01)	0.79 (+/- 0.05)	0.79 (+/- 0.01)	0.57 (+/- 0.01)
eMolTox - Cytotoxicity in HepG2 cells - 16 hour	0.80 (+/- 0.01)	0.81 (+/- 0.02)	0.76 (+/- 0.01)	0.75 (+/- 0.04)	0.76 (+/- 0.01)	0.58 (+/- 0.01)
eMolTox - Activators of the human pregnane X receptor (PXR) signaling pathway	0.81 (+/- 0.03)	0.97 (+/- 0.02)	0.83 (+/- 0.02)	0.85 (+/- 0.08)	0.83 (+/- 0.02)	0.72 (+/- 0.03)
eMolTox - Activators of Cytochrome P450 2A9	0.81 (+/- 0.01)	0.99 (+/- 0.00)	0.81 (+/- 0.01)	0.82 (+/- 0.02)	0.81 (+/- 0.01)	0.66 (+/- 0.01)

eMolTox - Modulator of Platelet activating factor receptor	0.81 (+/- 0.03)	0.82 (+/- 0.03)	0.99 (+/- 0.01)	0.98 (+/- 0.02)	1.00 (+/- 0.00)	0.99 (+/- 0.01)
eMolTox - Modulator of Vasopressin V1a receptor	0.82 (+/- 0.02)	0.82 (+/- 0.02)	1.00 (+/- 0.00)	1.00 (+/- 0.01)	1.00 (+/- 0.00)	1.00 (+/- 0.00)
eMolTox - Modulator of Type-1 angiotensin II receptor	0.80 (+/- 0.02)	0.80 (+/- 0.02)	1.00 (+/- 0.00)	1.00 (+/- 0.00)	1.00 (+/- 0.00)	1.00 (+/- 0.00)
eMolTox - Inhibitors and Substrates of Cytochrome P450 3A4	0.80 (+/- 0.01)	0.89 (+/- 0.01)	0.90 (+/- 0.01)	0.89 (+/- 0.01)	0.90 (+/- 0.01)	0.89 (+/- 0.01)
eMolTox - Modulator of Cyclooxygenase-1	0.80 (+/- 0.01)	0.88 (+/- 0.01)	0.90 (+/- 0.01)	0.93 (+/- 0.01)	0.90 (+/- 0.01)	0.85 (+/- 0.01)
eMolTox - Inhibit CYP2C9 Activity	0.80 (+/- 0.00)	0.82 (+/- 0.01)	0.76 (+/- 0.00)	0.76 (+/- 0.01)	0.76 (+/- 0.01)	0.76 (+/- 0.00)
eMolTox - Agonist of Liver X receptor beta	0.81 (+/- 0.02)	0.81 (+/- 0.02)	1.00 (+/- 0.00)	0.99 (+/- 0.01)	1.00 (+/- 0.00)	1.00 (+/- 0.00)
eMolTox - Antagonist of the farnesoid-X-receptor (FXR) signaling pathway dup	0.80 (+/- 0.02)	0.96 (+/- 0.03)	0.80 (+/- 0.02)	0.80 (+/- 0.10)	0.80 (+/- 0.02)	0.55 (+/- 0.02)
eMolTox - Modulator of Cyclooxygenase-2	0.81 (+/- 0.01)	0.84 (+/- 0.01)	0.97 (+/- 0.01)	0.97 (+/- 0.01)	0.97 (+/- 0.01)	0.94 (+/- 0.01)
eMolTox - Antagonist of the constitutive androstane receptor (CAR) signaling pathway	0.81 (+/- 0.01)	0.79 (+/- 0.02)	0.75 (+/- 0.02)	0.78 (+/- 0.02)	0.75 (+/- 0.02)	0.57 (+/- 0.01)
eMolTox - Activator the aryl hydrocarbon receptor (AhR) signaling pathway	0.80 (+/- 0.01)	0.97 (+/- 0.01)	0.83 (+/- 0.01)	0.86 (+/- 0.01)	0.82 (+/- 0.01)	0.73 (+/- 0.01)
eMolTox - Inhibitors and Substrates of Cytochrome P450 2D6	0.80 (+/- 0.01)	0.93 (+/- 0.01)	0.86 (+/- 0.01)	0.85 (+/- 0.02)	0.86 (+/- 0.01)	0.81 (+/- 0.01)
eMolTox - Modulator of Angiotensin-converting enzyme	0.82 (+/- 0.02)	0.83 (+/- 0.02)	0.99 (+/- 0.00)	0.98 (+/- 0.01)	0.99 (+/- 0.01)	0.98 (+/- 0.01)

eMolTox - Agonist of the constitutive androstane receptor (CAR) signaling pathway	0.81 (+/-0.01)	0.96 (+/-0.01)	0.84 (+/-0.01)	0.87 (+/-0.01)	0.84 (+/-0.01)	0.75 (+/-0.00)
eMolTox - Cytotoxicity in HepG2 cells - 32 hour	0.81 (+/-0.02)	0.86 (+/-0.03)	0.78 (+/-0.02)	0.80 (+/-0.05)	0.77 (+/-0.01)	0.65 (+/-0.02)
eMolTox - Activators of the heat shock response signaling pathway	0.81 (+/-0.01)	0.70 (+/-0.04)	0.73 (+/-0.01)	0.76 (+/-0.05)	0.72 (+/-0.02)	0.56 (+/-0.01)
eMolTox - Inhibitors and Substrates of Cytochrome P450 2C9	0.81 (+/-0.02)	0.99 (+/-0.00)	0.82 (+/-0.02)	0.81 (+/-0.02)	0.82 (+/-0.02)	0.74 (+/-0.02)
ToxCast - ACEA AR antagonist 80hr	0.83 (+/-0.02)	0.84 (+/-0.02)	0.80 (+/-0.02)	0.78 (+/-0.04)	0.81 (+/-0.02)	0.79 (+/-0.02)
ToxCast - ACEA ER 80hr	0.81 (+/-0.01)	0.50 (+/-0.04)	0.62 (+/-0.02)	0.74 (+/-0.09)	0.60 (+/-0.04)	0.56 (+/-0.01)
ToxCast - APR HepG2 CellCycleArrest 72h dn	0.82 (+/-0.03)	0.65 (+/-0.05)	0.72 (+/-0.02)	0.75 (+/-0.03)	0.71 (+/-0.03)	0.66 (+/-0.02)
ToxCast - APR HepG2 CellLoss 24h dn	0.78 (+/-0.03)	0.82 (+/-0.01)	0.74 (+/-0.04)	0.72 (+/-0.03)	0.74 (+/-0.07)	0.70 (+/-0.03)
ToxCast - APR HepG2 CellLoss 72h dn	0.78 (+/-0.03)	0.85 (+/-0.03)	0.74 (+/-0.03)	0.73 (+/-0.03)	0.75 (+/-0.06)	0.74 (+/-0.03)
ToxCast - APR HepG2 MitoticArrest 72h up	0.80 (+/-0.02)	0.78 (+/-0.03)	0.74 (+/-0.02)	0.76 (+/-0.05)	0.73 (+/-0.03)	0.70 (+/-0.02)
ToxCast - APR HepG2 OxidativeStress 24h up	0.81 (+/-0.04)	0.73 (+/-0.04)	0.74 (+/-0.04)	0.78 (+/-0.04)	0.73 (+/-0.06)	0.69 (+/-0.03)
ToxCast - APR HepG2 OxidativeStress 72h up	0.83 (+/-0.03)	0.71 (+/-0.03)	0.75 (+/-0.03)	0.75 (+/-0.06)	0.76 (+/-0.03)	0.70 (+/-0.04)

ToxCast - ATG AP 1 CIS up	0.81 (+/- 0.01)	0.82 (+/- 0.02)	0.77 (+/- 0.01)	0.79 (+/- 0.03)	0.76 (+/- 0.02)	0.69 (+/- 0.01)
ToxCast - ATG Ahr CIS dn	0.79 (+/- 0.02)	0.41 (+/- 0.02)	0.48 (+/- 0.02)	0.60 (+/- 0.10)	0.48 (+/- 0.03)	0.38 (+/- 0.01)
ToxCast - ATG Ahr CIS up	0.80 (+/- 0.02)	0.86 (+/- 0.02)	0.77 (+/- 0.02)	0.75 (+/- 0.09)	0.77 (+/- 0.02)	0.66 (+/- 0.02)
ToxCast - ATG BRE CIS up	0.80 (+/- 0.01)	0.84 (+/- 0.01)	0.76 (+/- 0.01)	0.77 (+/- 0.07)	0.76 (+/- 0.02)	0.62 (+/- 0.01)
ToxCast - ATG CMV CIS up	0.79 (+/- 0.02)	0.90 (+/- 0.02)	0.77 (+/- 0.02)	0.79 (+/- 0.02)	0.77 (+/- 0.02)	0.68 (+/- 0.02)
ToxCast - ATG CRE CIS up	0.81 (+/- 0.02)	0.80 (+/- 0.02)	0.77 (+/- 0.02)	0.76 (+/- 0.03)	0.77 (+/- 0.02)	0.62 (+/- 0.01)
ToxCast - ATG C EBP CIS up	0.80 (+/- 0.01)	0.72 (+/- 0.03)	0.72 (+/- 0.01)	0.74 (+/- 0.08)	0.72 (+/- 0.01)	0.53 (+/- 0.00)
ToxCast - ATG DR4 LXR CIS dn	0.81 (+/- 0.02)	0.89 (+/- 0.01)	0.78 (+/- 0.01)	0.80 (+/- 0.04)	0.78 (+/- 0.01)	0.72 (+/- 0.02)
ToxCast - ATG DR5 CIS up	0.81 (+/- 0.02)	0.74 (+/- 0.04)	0.75 (+/- 0.01)	0.81 (+/- 0.08)	0.74 (+/- 0.01)	0.61 (+/- 0.02)
ToxCast - ATG EGR CIS up	0.80 (+/- 0.01)	0.89 (+/- 0.02)	0.78 (+/- 0.01)	0.77 (+/- 0.03)	0.78 (+/- 0.02)	0.68 (+/- 0.01)
ToxCast - ATG ERE CIS up	0.81 (+/- 0.02)	0.70 (+/- 0.02)	0.73 (+/- 0.02)	0.76 (+/- 0.03)	0.72 (+/- 0.02)	0.70 (+/- 0.02)
ToxCast - ATG ERa TRANS up	0.81 (+/- 0.01)	0.83 (+/- 0.02)	0.77 (+/- 0.01)	0.80 (+/- 0.04)	0.76 (+/- 0.02)	0.73 (+/- 0.01)
ToxCast - ATG E Box CIS dn	0.80 (+/- 0.01)	0.74 (+/- 0.02)	0.73 (+/- 0.02)	0.74 (+/- 0.03)	0.73 (+/- 0.02)	0.57 (+/- 0.02)
ToxCast - ATG Ets CIS dn	0.80 (+/- 0.03)	0.65 (+/- 0.05)	0.70 (+/- 0.03)	0.79 (+/- 0.11)	0.69 (+/- 0.03)	0.53 (+/- 0.02)
ToxCast - ATG HIF1a CIS up	0.79 (+/- 0.01)	0.77 (+/- 0.03)	0.74 (+/- 0.02)	0.78 (+/- 0.04)	0.73 (+/- 0.02)	0.60 (+/- 0.02)
ToxCast - ATG HSE CIS up	0.81 (+/- 0.01)	0.84 (+/- 0.02)	0.77 (+/- 0.01)	0.80 (+/- 0.05)	0.77 (+/- 0.01)	0.65 (+/- 0.01)
ToxCast - ATG IR1 CIS dn	0.80 (+/- 0.02)	0.79 (+/- 0.02)	0.75 (+/- 0.02)	0.79 (+/- 0.02)	0.74 (+/- 0.02)	0.61 (+/- 0.02)
ToxCast - ATG ISRE CIS dn	0.81 (+/- 0.01)	0.75 (+/- 0.02)	0.75 (+/- 0.02)	0.82 (+/- 0.03)	0.74 (+/- 0.02)	0.65 (+/- 0.03)

ToxCast - ATG MRE CIS up	0.80 (+/- 0.01)	0.89 (+/- 0.02)	0.78 (+/- 0.01)	0.80 (+/- 0.06)	0.77 (+/- 0.02)	0.72 (+/- 0.01)
ToxCast - ATG NFI CIS up	0.79 (+/- 0.02)	0.72 (+/- 0.02)	0.71 (+/- 0.02)	0.72 (+/- 0.12)	0.71 (+/- 0.01)	0.55 (+/- 0.02)
ToxCast - ATG NF kB CIS dn	0.81 (+/- 0.02)	0.66 (+/- 0.04)	0.70 (+/- 0.02)	0.73 (+/- 0.10)	0.70 (+/- 0.02)	0.56 (+/- 0.03)
ToxCast - ATG NRF2 ARE CIS up	0.81 (+/- 0.02)	0.87 (+/- 0.01)	0.78 (+/- 0.02)	0.77 (+/- 0.02)	0.78 (+/- 0.02)	0.77 (+/- 0.02)
ToxCast - ATG Oct MLP CIS up	0.81 (+/- 0.02)	0.87 (+/- 0.02)	0.78 (+/- 0.02)	0.77 (+/- 0.02)	0.78 (+/- 0.03)	0.68 (+/- 0.02)
ToxCast - ATG PBREM CIS up	0.79 (+/- 0.01)	0.91 (+/- 0.02)	0.77 (+/- 0.01)	0.76 (+/- 0.05)	0.77 (+/- 0.01)	0.65 (+/- 0.01)
ToxCast - ATG PPARa TRANS up	0.79 (+/- 0.02)	0.78 (+/- 0.01)	0.73 (+/- 0.03)	0.78 (+/- 0.04)	0.73 (+/- 0.03)	0.57 (+/- 0.02)
ToxCast - ATG PPARg TRANS up	0.81 (+/- 0.01)	0.92 (+/- 0.02)	0.80 (+/- 0.02)	0.82 (+/- 0.03)	0.79 (+/- 0.02)	0.77 (+/- 0.01)
ToxCast - ATG PPRE CIS up	0.81 (+/- 0.02)	0.88 (+/- 0.01)	0.78 (+/- 0.03)	0.79 (+/- 0.02)	0.78 (+/- 0.03)	0.69 (+/- 0.03)
ToxCast - ATG PXRE CIS dn	0.79 (+/- 0.01)	0.59 (+/- 0.03)	0.64 (+/- 0.01)	0.70 (+/- 0.08)	0.63 (+/- 0.02)	0.50 (+/- 0.02)
ToxCast - ATG PXRE CIS up	0.80 (+/- 0.01)	0.89 (+/- 0.01)	0.78 (+/- 0.01)	0.78 (+/- 0.03)	0.78 (+/- 0.03)	0.78 (+/- 0.01)
ToxCast - ATG PXR TRANS up	0.80 (+/- 0.02)	0.94 (+/- 0.01)	0.79 (+/- 0.02)	0.81 (+/- 0.03)	0.78 (+/- 0.03)	0.77 (+/- 0.02)
ToxCast - ATG Pax6 CIS up	0.80 (+/- 0.01)	0.87 (+/- 0.02)	0.77 (+/- 0.01)	0.78 (+/- 0.03)	0.77 (+/- 0.01)	0.65 (+/- 0.01)
ToxCast - ATG RARa TRANS dn	0.81 (+/- 0.02)	0.67 (+/- 0.01)	0.71 (+/- 0.03)	0.70 (+/- 0.05)	0.71 (+/- 0.03)	0.53 (+/- 0.02)
ToxCast - ATG RORE CIS up	0.80 (+/- 0.02)	0.91 (+/- 0.02)	0.78 (+/- 0.02)	0.73 (+/- 0.10)	0.78 (+/- 0.01)	0.64 (+/- 0.03)
ToxCast - ATG RXRb TRANS up	0.79 (+/- 0.01)	0.82 (+/- 0.02)	0.75 (+/- 0.01)	0.77 (+/- 0.04)	0.74 (+/- 0.01)	0.66 (+/- 0.01)
ToxCast - ATG SREBP CIS up	0.80 (+/- 0.02)	0.90 (+/- 0.01)	0.78 (+/- 0.02)	0.78 (+/- 0.03)	0.78 (+/- 0.02)	0.68 (+/- 0.02)
ToxCast - ATG Sox CIS up	0.81 (+/- 0.01)	0.66 (+/- 0.04)	0.71 (+/- 0.02)	0.74 (+/- 0.10)	0.71 (+/- 0.03)	0.54 (+/- 0.02)

ToxCast - ATG Sp1 CIS up	0.80 (+/- 0.02)	0.73 (+/- 0.04)	0.73 (+/- 0.02)	0.78 (+/- 0.13)	0.73 (+/- 0.02)	0.59 (+/- 0.02)
ToxCast - ATG TA CIS up	0.80 (+/- 0.02)	0.87 (+/- 0.03)	0.77 (+/- 0.02)	0.81 (+/- 0.07)	0.76 (+/- 0.02)	0.58 (+/- 0.02)
ToxCast - ATG TCF b cat CIS dn	0.80 (+/- 0.01)	0.75 (+/- 0.01)	0.74 (+/- 0.01)	0.77 (+/- 0.04)	0.73 (+/- 0.02)	0.63 (+/- 0.01)
ToxCast - ATG VDRE CIS up	0.79 (+/- 0.01)	0.92 (+/- 0.01)	0.78 (+/- 0.01)	0.81 (+/- 0.02)	0.77 (+/- 0.01)	0.75 (+/- 0.01)
ToxCast - ATG Xbp1 CIS up	0.81 (+/- 0.01)	0.82 (+/- 0.01)	0.76 (+/- 0.01)	0.81 (+/- 0.07)	0.76 (+/- 0.02)	0.65 (+/- 0.02)
ToxCast - ATG p53 CIS dn	0.81 (+/- 0.02)	0.72 (+/- 0.05)	0.74 (+/- 0.03)	0.78 (+/- 0.08)	0.73 (+/- 0.03)	0.60 (+/- 0.02)
ToxCast - BSK 3C Eselectin down	0.79 (+/- 0.02)	0.77 (+/- 0.02)	0.73 (+/- 0.03)	0.70 (+/- 0.07)	0.74 (+/- 0.03)	0.69 (+/- 0.03)
ToxCast - BSK 3C HLADR down	0.79 (+/- 0.04)	0.81 (+/- 0.05)	0.75 (+/- 0.03)	0.73 (+/- 0.02)	0.76 (+/- 0.04)	0.74 (+/- 0.03)
ToxCast - BSK 3C IL8 down	0.80 (+/- 0.03)	0.89 (+/- 0.02)	0.77 (+/- 0.03)	0.82 (+/- 0.07)	0.76 (+/- 0.03)	0.73 (+/- 0.04)
ToxCast - BSK 3C MCP1 down	0.79 (+/- 0.02)	0.84 (+/- 0.03)	0.75 (+/- 0.03)	0.72 (+/- 0.05)	0.76 (+/- 0.04)	0.71 (+/- 0.03)
ToxCast - BSK 3C Proliferation down	0.80 (+/- 0.03)	0.87 (+/- 0.03)	0.77 (+/- 0.03)	0.77 (+/- 0.05)	0.77 (+/- 0.04)	0.76 (+/- 0.03)
ToxCast - BSK 3C SRB down	0.81 (+/- 0.01)	0.86 (+/- 0.01)	0.78 (+/- 0.01)	0.77 (+/- 0.04)	0.78 (+/- 0.01)	0.75 (+/- 0.01)
ToxCast - BSK 3C TissueFactor down	0.82 (+/- 0.02)	0.80 (+/- 0.03)	0.78 (+/- 0.02)	0.80 (+/- 0.05)	0.77 (+/- 0.03)	0.72 (+/- 0.02)
ToxCast - BSK 3C VCAM1 down	0.83 (+/- 0.04)	0.79 (+/- 0.03)	0.78 (+/- 0.04)	0.73 (+/- 0.07)	0.80 (+/- 0.04)	0.73 (+/- 0.04)
ToxCast - BSK 3C Vis down	0.78 (+/- 0.04)	0.83 (+/- 0.02)	0.74 (+/- 0.04)	0.73 (+/- 0.06)	0.75 (+/- 0.05)	0.71 (+/- 0.04)
ToxCast - BSK 3C uPAR down	0.81 (+/- 0.02)	0.85 (+/- 0.01)	0.77 (+/- 0.02)	0.81 (+/- 0.07)	0.76 (+/- 0.04)	0.75 (+/- 0.03)
ToxCast - BSK 4H Eotaxin3 down	0.78 (+/- 0.04)	0.81 (+/- 0.03)	0.73 (+/- 0.05)	0.73 (+/- 0.05)	0.73 (+/- 0.06)	0.70 (+/- 0.04)
ToxCast - BSK 4H MCP1 down	0.80 (+/- 0.02)	0.86 (+/- 0.02)	0.77 (+/- 0.02)	0.78 (+/- 0.07)	0.77 (+/- 0.01)	0.73 (+/- 0.03)

ToxCast - BSK 4H Pselectin down	0.80 (+/- 0.01)	0.85 (+/- 0.02)	0.76 (+/- 0.02)	0.79 (+/- 0.06)	0.75 (+/- 0.03)	0.73 (+/- 0.02)
ToxCast - BSK 4H SRB down	0.81 (+/- 0.02)	0.84 (+/- 0.02)	0.77 (+/- 0.01)	0.77 (+/- 0.02)	0.77 (+/- 0.02)	0.73 (+/- 0.01)
ToxCast - BSK 4H VCAM1 down	0.80 (+/- 0.04)	0.84 (+/- 0.05)	0.76 (+/- 0.03)	0.75 (+/- 0.06)	0.77 (+/- 0.03)	0.73 (+/- 0.04)
ToxCast - BSK 4H uPAR down	0.81 (+/- 0.03)	0.82 (+/- 0.02)	0.77 (+/- 0.03)	0.78 (+/- 0.07)	0.76 (+/- 0.03)	0.71 (+/- 0.04)
ToxCast - BSK BE3C HLADR down	0.80 (+/- 0.03)	0.82 (+/- 0.03)	0.75 (+/- 0.03)	0.77 (+/- 0.03)	0.75 (+/- 0.03)	0.72 (+/- 0.03)
ToxCast - BSK BE3C IL1a down	0.81 (+/- 0.01)	0.81 (+/- 0.02)	0.76 (+/- 0.02)	0.76 (+/- 0.03)	0.76 (+/- 0.02)	0.68 (+/- 0.02)
ToxCast - BSK BE3C IP10 down	0.81 (+/- 0.02)	0.86 (+/- 0.03)	0.78 (+/- 0.01)	0.79 (+/- 0.09)	0.77 (+/- 0.02)	0.74 (+/- 0.02)
ToxCast - BSK BE3C PAI1 down	0.81 (+/- 0.04)	0.87 (+/- 0.02)	0.79 (+/- 0.04)	0.76 (+/- 0.04)	0.79 (+/- 0.04)	0.71 (+/- 0.04)
ToxCast - BSK BE3C tPA down	0.81 (+/- 0.03)	0.83 (+/- 0.03)	0.77 (+/- 0.03)	0.77 (+/- 0.04)	0.77 (+/- 0.04)	0.67 (+/- 0.03)
ToxCast - BSK BE3C uPA down	0.80 (+/- 0.02)	0.73 (+/- 0.03)	0.73 (+/- 0.01)	0.67 (+/- 0.15)	0.74 (+/- 0.01)	0.63 (+/- 0.04)
ToxCast - BSK CASM3C HLADR down	0.80 (+/- 0.02)	0.72 (+/- 0.03)	0.72 (+/- 0.02)	0.78 (+/- 0.06)	0.71 (+/- 0.02)	0.65 (+/- 0.03)
ToxCast - BSK CASM3C MCP1 down	0.81 (+/- 0.01)	0.74 (+/- 0.03)	0.75 (+/- 0.01)	0.70 (+/- 0.11)	0.75 (+/- 0.02)	0.64 (+/- 0.03)
ToxCast - BSK CASM3C MCSF down	0.82 (+/- 0.03)	0.70 (+/- 0.06)	0.74 (+/- 0.03)	0.71 (+/- 0.12)	0.75 (+/- 0.03)	0.66 (+/- 0.04)
ToxCast - BSK CASM3C Proliferation down	0.79 (+/- 0.04)	0.80 (+/- 0.05)	0.74 (+/- 0.03)	0.74 (+/- 0.08)	0.73 (+/- 0.02)	0.71 (+/- 0.04)
ToxCast - BSK CASM3C SRB down	0.82 (+/- 0.01)	0.82 (+/- 0.04)	0.78 (+/- 0.02)	0.77 (+/- 0.11)	0.78 (+/- 0.04)	0.69 (+/- 0.02)
ToxCast - BSK CASM3C Thrombomodulin up	0.83 (+/- 0.03)	0.73 (+/- 0.04)	0.76 (+/- 0.03)	0.81 (+/- 0.03)	0.75 (+/- 0.03)	0.68 (+/- 0.02)

ToxCast - BSK CASM3C VCAM1 down	0.81 (+/- 0.01)	0.74 (+/- 0.02)	0.74 (+/- 0.02)	0.69 (+/- 0.07)	0.75 (+/- 0.02)	0.65 (+/- 0.03)
ToxCast - BSK CASM3C uPAR down	0.80 (+/- 0.04)	0.69 (+/- 0.07)	0.71 (+/- 0.04)	0.74 (+/- 0.06)	0.71 (+/- 0.04)	0.63 (+/- 0.04)
ToxCast - BSK KF3CT ICAM1 down	0.80 (+/- 0.03)	0.83 (+/- 0.04)	0.77 (+/- 0.04)	0.78 (+/- 0.05)	0.76 (+/- 0.04)	0.67 (+/- 0.04)
ToxCast - BSK KF3CT IL1a down	0.82 (+/- 0.02)	0.77 (+/- 0.04)	0.77 (+/- 0.02)	0.79 (+/- 0.04)	0.76 (+/- 0.03)	0.70 (+/- 0.01)
ToxCast - BSK KF3CT IP10 down	0.80 (+/- 0.03)	0.80 (+/- 0.01)	0.75 (+/- 0.03)	0.82 (+/- 0.05)	0.73 (+/- 0.05)	0.70 (+/- 0.03)
ToxCast - BSK KF3CT MCP1 down	0.81 (+/- 0.01)	0.76 (+/- 0.07)	0.75 (+/- 0.02)	0.80 (+/- 0.10)	0.74 (+/- 0.04)	0.68 (+/- 0.01)
ToxCast - BSK KF3CT MMP9 down	0.82 (+/- 0.02)	0.81 (+/- 0.03)	0.78 (+/- 0.02)	0.81 (+/- 0.06)	0.77 (+/- 0.02)	0.74 (+/- 0.03)
ToxCast - BSK KF3CT SRB down	0.82 (+/- 0.03)	0.86 (+/- 0.05)	0.80 (+/- 0.03)	0.80 (+/- 0.07)	0.80 (+/- 0.03)	0.71 (+/- 0.03)
ToxCast - BSK KF3CT TGFb1 down	0.82 (+/- 0.01)	0.76 (+/- 0.05)	0.77 (+/- 0.01)	0.82 (+/- 0.08)	0.75 (+/- 0.02)	0.70 (+/- 0.02)
ToxCast - BSK KF3CT TIMP2 down	0.82 (+/- 0.05)	0.81 (+/- 0.05)	0.78 (+/- 0.05)	0.79 (+/- 0.06)	0.77 (+/- 0.06)	0.71 (+/- 0.06)
ToxCast - BSK KF3CT uPA down	0.82 (+/- 0.01)	0.80 (+/- 0.04)	0.78 (+/- 0.01)	0.79 (+/- 0.07)	0.78 (+/- 0.01)	0.70 (+/- 0.02)
ToxCast - BSK LPS CD40 down	0.82 (+/- 0.02)	0.80 (+/- 0.02)	0.77 (+/- 0.03)	0.79 (+/- 0.04)	0.76 (+/- 0.03)	0.74 (+/- 0.03)
ToxCast - BSK LPS Eselectin down	0.81 (+/- 0.02)	0.77 (+/- 0.03)	0.75 (+/- 0.02)	0.79 (+/- 0.08)	0.74 (+/- 0.03)	0.69 (+/- 0.02)
ToxCast - BSK LPS IL1a down	0.80 (+/- 0.02)	0.81 (+/- 0.01)	0.76 (+/- 0.03)	0.73 (+/- 0.01)	0.77 (+/- 0.04)	0.69 (+/- 0.03)
ToxCast - BSK LPS IL8 down	0.81 (+/- 0.02)	0.83 (+/- 0.03)	0.77 (+/- 0.02)	0.76 (+/- 0.04)	0.77 (+/- 0.01)	0.71 (+/- 0.03)
ToxCast - BSK LPS MCP1 down	0.79 (+/- 0.04)	0.81 (+/- 0.02)	0.74 (+/- 0.04)	0.78 (+/- 0.03)	0.73 (+/- 0.05)	0.69 (+/- 0.04)
ToxCast - BSK LPS MCSF down	0.82 (+/- 0.02)	0.82 (+/- 0.02)	0.77 (+/- 0.02)	0.77 (+/- 0.05)	0.78 (+/- 0.04)	0.74 (+/- 0.02)
ToxCast - BSK LPS PGE2 down	0.80 (+/- 0.02)	0.80 (+/- 0.02)	0.75 (+/- 0.02)	0.79 (+/- 0.08)	0.74 (+/- 0.02)	0.70 (+/- 0.03)

ToxCast - BSK LPS SRB down	0.82 (+/-0.03)	0.83 (+/-0.04)	0.78 (+/-0.03)	0.76 (+/-0.08)	0.78 (+/-0.04)	0.73 (+/-0.03)
ToxCast - BSK LPS TNFa down	0.80 (+/-0.02)	0.80 (+/-0.02)	0.75 (+/-0.02)	0.78 (+/-0.04)	0.74 (+/-0.03)	0.68 (+/-0.01)
ToxCast - BSK LPS TissueFactor down	0.82 (+/-0.02)	0.74 (+/-0.07)	0.75 (+/-0.03)	0.73 (+/-0.06)	0.75 (+/-0.04)	0.64 (+/-0.02)
ToxCast - BSK LPS VCAM1 down	0.82 (+/-0.02)	0.81 (+/-0.03)	0.78 (+/-0.03)	0.75 (+/-0.03)	0.79 (+/-0.03)	0.74 (+/-0.03)
ToxCast - BSK SAg CD38 down	0.80 (+/-0.03)	0.86 (+/-0.03)	0.77 (+/-0.03)	0.80 (+/-0.04)	0.76 (+/-0.03)	0.75 (+/-0.03)
ToxCast - BSK SAg CD40 down	0.82 (+/-0.03)	0.83 (+/-0.04)	0.79 (+/-0.03)	0.83 (+/-0.07)	0.77 (+/-0.04)	0.76 (+/-0.03)
ToxCast - BSK SAg CD69 down	0.83 (+/-0.04)	0.80 (+/-0.06)	0.79 (+/-0.03)	0.82 (+/-0.10)	0.78 (+/-0.01)	0.76 (+/-0.04)
ToxCast - BSK SAg Eselectin down	0.80 (+/-0.04)	0.79 (+/-0.04)	0.75 (+/-0.04)	0.81 (+/-0.06)	0.72 (+/-0.05)	0.73 (+/-0.03)
ToxCast - BSK SAg IL8 down	0.81 (+/-0.02)	0.86 (+/-0.02)	0.77 (+/-0.02)	0.79 (+/-0.09)	0.77 (+/-0.03)	0.73 (+/-0.03)
ToxCast - BSK SAg MCP1 down	0.81 (+/-0.04)	0.79 (+/-0.04)	0.76 (+/-0.04)	0.78 (+/-0.05)	0.76 (+/-0.05)	0.72 (+/-0.04)
ToxCast - BSK SAg MIG down	0.80 (+/-0.02)	0.87 (+/-0.01)	0.77 (+/-0.02)	0.78 (+/-0.08)	0.77 (+/-0.02)	0.67 (+/-0.02)
ToxCast - BSK SAg PBMCCytotoxicity down	0.83 (+/-0.04)	0.85 (+/-0.02)	0.80 (+/-0.04)	0.80 (+/-0.10)	0.80 (+/-0.04)	0.75 (+/-0.05)
ToxCast - BSK SAg Proliferation down	0.81 (+/-0.03)	0.82 (+/-0.03)	0.77 (+/-0.02)	0.73 (+/-0.05)	0.79 (+/-0.02)	0.75 (+/-0.03)
ToxCast - BSK SAg SRB down	0.80 (+/-0.02)	0.86 (+/-0.02)	0.77 (+/-0.02)	0.80 (+/-0.05)	0.76 (+/-0.02)	0.73 (+/-0.02)
ToxCast - BSK hDFCGF CollagenIII down	0.83 (+/-0.02)	0.85 (+/-0.02)	0.80 (+/-0.02)	0.81 (+/-0.06)	0.80 (+/-0.01)	0.77 (+/-0.02)
ToxCast - BSK hDFCGF IP10 down	0.81 (+/-0.02)	0.76 (+/-0.03)	0.75 (+/-0.02)	0.77 (+/-0.02)	0.75 (+/-0.04)	0.72 (+/-0.02)
ToxCast - BSK hDFCGF MCSF down	0.83 (+/-0.03)	0.80 (+/-0.03)	0.79 (+/-0.03)	0.83 (+/-0.03)	0.77 (+/-0.04)	0.76 (+/-0.03)

ToxCast - BSK hDFCGF MIG down	0.81 (+/- 0.01)	0.84 (+/- 0.02)	0.77 (+/- 0.02)	0.78 (+/- 0.12)	0.76 (+/- 0.05)	0.70 (+/- 0.02)
ToxCast - BSK hDFCGF MMP1 down	0.82 (+/- 0.03)	0.83 (+/- 0.03)	0.78 (+/- 0.03)	0.76 (+/- 0.04)	0.78 (+/- 0.04)	0.70 (+/- 0.03)
ToxCast - BSK hDFCGF PAI1 down	0.79 (+/- 0.02)	0.84 (+/- 0.02)	0.75 (+/- 0.02)	0.75 (+/- 0.05)	0.75 (+/- 0.04)	0.71 (+/- 0.02)
ToxCast - BSK hDFCGF Proliferation down	0.80 (+/- 0.02)	0.81 (+/- 0.02)	0.75 (+/- 0.02)	0.77 (+/- 0.05)	0.74 (+/- 0.03)	0.75 (+/- 0.02)
ToxCast - BSK hDFCGF SRB down	0.81 (+/- 0.02)	0.87 (+/- 0.03)	0.79 (+/- 0.02)	0.81 (+/- 0.07)	0.78 (+/- 0.01)	0.74 (+/- 0.03)
ToxCast - BSK hDFCGF TIMP1 down	0.82 (+/- 0.02)	0.79 (+/- 0.02)	0.77 (+/- 0.01)	0.73 (+/- 0.03)	0.78 (+/- 0.01)	0.68 (+/- 0.02)
ToxCast - BSK hDFCGF VCAM1 down	0.81 (+/- 0.03)	0.76 (+/- 0.03)	0.75 (+/- 0.04)	0.77 (+/- 0.10)	0.74 (+/- 0.02)	0.72 (+/- 0.04)
ToxCast - LTEA HepaRG ABCB11 dn	0.78 (+/- 0.01)	0.86 (+/- 0.02)	0.75 (+/- 0.01)	0.76 (+/- 0.03)	0.74 (+/- 0.02)	0.72 (+/- 0.02)
ToxCast - LTEA HepaRG ABCB1 up	0.81 (+/- 0.01)	0.91 (+/- 0.03)	0.79 (+/- 0.01)	0.79 (+/- 0.11)	0.79 (+/- 0.04)	0.74 (+/- 0.01)
ToxCast - LTEA HepaRG ABCG2 up	0.83 (+/- 0.03)	0.84 (+/- 0.03)	0.80 (+/- 0.03)	0.80 (+/- 0.02)	0.80 (+/- 0.03)	0.76 (+/- 0.03)
ToxCast - LTEA HepaRG ACOX1 dn	0.81 (+/- 0.04)	0.79 (+/- 0.03)	0.76 (+/- 0.04)	0.80 (+/- 0.09)	0.75 (+/- 0.04)	0.71 (+/- 0.05)
ToxCast - LTEA HepaRG AFP dn	0.80 (+/- 0.05)	0.94 (+/- 0.03)	0.78 (+/- 0.04)	0.78 (+/- 0.05)	0.78 (+/- 0.06)	0.77 (+/- 0.04)
ToxCast - LTEA HepaRG ALPP dn	0.79 (+/- 0.02)	0.89 (+/- 0.03)	0.76 (+/- 0.02)	0.82 (+/- 0.05)	0.75 (+/- 0.04)	0.72 (+/- 0.01)
ToxCast - LTEA HepaRG APOA5 dn	0.82 (+/- 0.03)	0.77 (+/- 0.04)	0.77 (+/- 0.03)	0.81 (+/- 0.09)	0.75 (+/- 0.01)	0.73 (+/- 0.04)
ToxCast - LTEA HepaRG CAT dn	0.80 (+/- 0.05)	0.80 (+/- 0.03)	0.75 (+/- 0.06)	0.78 (+/- 0.11)	0.73 (+/- 0.07)	0.70 (+/- 0.06)
ToxCast - LTEA HepaRG CYP1A1 up	0.80 (+/- 0.04)	0.61 (+/- 0.06)	0.67 (+/- 0.04)	0.65 (+/- 0.05)	0.74 (+/- 0.10)	0.62 (+/- 0.05)
ToxCast - LTEA HepaRG CYP1A2 up	0.81 (+/- 0.05)	0.55 (+/- 0.04)	0.66 (+/- 0.07)	0.66 (+/- 0.05)	0.67 (+/- 0.10)	0.66 (+/- 0.07)

ToxCast - LTEA HepaRG CYP2B6 up	0.83 (+/- 0.03)	0.73 (+/- 0.05)	0.77 (+/- 0.02)	0.75 (+/- 0.03)	0.78 (+/- 0.04)	0.77 (+/- 0.02)
ToxCast - LTEA HepaRG CYP2C19 up	0.82 (+/- 0.02)	0.84 (+/- 0.03)	0.78 (+/- 0.02)	0.84 (+/- 0.02)	0.75 (+/- 0.04)	0.76 (+/- 0.02)
ToxCast - LTEA HepaRG CYP2C9 dn	0.80 (+/- 0.02)	0.84 (+/- 0.02)	0.76 (+/- 0.02)	0.78 (+/- 0.09)	0.75 (+/- 0.04)	0.70 (+/- 0.03)
ToxCast - LTEA HepaRG CYP2E1 dn	0.80 (+/- 0.01)	0.95 (+/- 0.01)	0.79 (+/- 0.01)	0.80 (+/- 0.02)	0.79 (+/- 0.02)	0.79 (+/- 0.01)
ToxCast - LTEA HepaRG CYP3A4 up	0.83 (+/- 0.03)	0.89 (+/- 0.02)	0.81 (+/- 0.04)	0.79 (+/- 0.06)	0.82 (+/- 0.05)	0.79 (+/- 0.04)
ToxCast - LTEA HepaRG CYP3A7 up	0.82 (+/- 0.03)	0.72 (+/- 0.03)	0.75 (+/- 0.02)	0.75 (+/- 0.05)	0.74 (+/- 0.02)	0.70 (+/- 0.03)
ToxCast - LTEA HepaRG CYP4A11 dn	0.80 (+/- 0.03)	0.92 (+/- 0.02)	0.78 (+/- 0.03)	0.81 (+/- 0.04)	0.77 (+/- 0.05)	0.76 (+/- 0.03)
ToxCast - LTEA HepaRG CYP4A22 dn	0.80 (+/- 0.03)	0.92 (+/- 0.02)	0.79 (+/- 0.03)	0.79 (+/- 0.07)	0.79 (+/- 0.05)	0.77 (+/- 0.03)
ToxCast - LTEA HepaRG CYP7A1 dn	0.80 (+/- 0.02)	0.83 (+/- 0.02)	0.76 (+/- 0.02)	0.74 (+/- 0.06)	0.77 (+/- 0.02)	0.74 (+/- 0.02)
ToxCast - LTEA HepaRG DDIT3 up	0.80 (+/- 0.04)	0.95 (+/- 0.04)	0.79 (+/- 0.04)	0.80 (+/- 0.12)	0.79 (+/- 0.03)	0.74 (+/- 0.05)
ToxCast - LTEA HepaRG FABP1 dn	0.80 (+/- 0.03)	0.87 (+/- 0.02)	0.77 (+/- 0.03)	0.78 (+/- 0.09)	0.77 (+/- 0.03)	0.76 (+/- 0.03)
ToxCast - LTEA HepaRG FASN dn	0.81 (+/- 0.04)	0.77 (+/- 0.07)	0.75 (+/- 0.03)	0.76 (+/- 0.08)	0.75 (+/- 0.04)	0.71 (+/- 0.03)
ToxCast - LTEA HepaRG FMO3 dn	0.78 (+/- 0.02)	0.93 (+/- 0.02)	0.76 (+/- 0.02)	0.78 (+/- 0.05)	0.75 (+/- 0.02)	0.72 (+/- 0.02)
ToxCast - LTEA HepaRG GSTA2 dn	0.79 (+/- 0.03)	0.88 (+/- 0.03)	0.76 (+/- 0.03)	0.79 (+/- 0.01)	0.75 (+/- 0.04)	0.71 (+/- 0.03)
ToxCast - LTEA HepaRG HMGCS2 dn	0.82 (+/- 0.02)	0.87 (+/- 0.02)	0.79 (+/- 0.02)	0.80 (+/- 0.05)	0.78 (+/- 0.02)	0.77 (+/- 0.03)
ToxCast - LTEA HepaRG IGF1 dn	0.81 (+/- 0.03)	0.94 (+/- 0.02)	0.80 (+/- 0.03)	0.82 (+/- 0.07)	0.79 (+/- 0.04)	0.79 (+/- 0.03)
ToxCast - LTEA HepaRG IGFBP1 up	0.81 (+/- 0.03)	0.87 (+/- 0.02)	0.79 (+/- 0.03)	0.83 (+/- 0.05)	0.77 (+/- 0.03)	0.75 (+/- 0.03)
ToxCast - LTEA HepaRG KRT19 dn	0.78 (+/- 0.05)	0.89 (+/- 0.02)	0.76 (+/- 0.05)	0.77 (+/- 0.06)	0.75 (+/- 0.06)	0.72 (+/- 0.05)

ToxCast - LTEA HepaRG LIPC dn	0.82 (+/- 0.04)	0.82 (+/- 0.04)	0.78 (+/- 0.04)	0.84 (+/- 0.07)	0.76 (+/- 0.03)	0.75 (+/- 0.04)
ToxCast - LTEA HepaRG MYC up	0.82 (+/- 0.03)	0.86 (+/- 0.04)	0.79 (+/- 0.02)	0.83 (+/- 0.04)	0.78 (+/- 0.03)	0.74 (+/- 0.02)
ToxCast - LTEA HepaRG PEG10 dn	0.82 (+/- 0.03)	0.89 (+/- 0.02)	0.79 (+/- 0.03)	0.79 (+/- 0.07)	0.80 (+/- 0.04)	0.78 (+/- 0.03)
ToxCast - LTEA HepaRG SLC22A1 dn	0.82 (+/- 0.04)	0.87 (+/- 0.03)	0.79 (+/- 0.04)	0.80 (+/- 0.08)	0.79 (+/- 0.04)	0.77 (+/- 0.05)
ToxCast - LTEA HepaRG SLCO1B1 dn	0.81 (+/- 0.05)	0.79 (+/- 0.04)	0.76 (+/- 0.06)	0.78 (+/- 0.10)	0.75 (+/- 0.05)	0.70 (+/- 0.07)
ToxCast - LTEA HepaRG THRSP dn	0.81 (+/- 0.02)	0.79 (+/- 0.03)	0.75 (+/- 0.02)	0.81 (+/- 0.03)	0.74 (+/- 0.03)	0.70 (+/- 0.02)
ToxCast - LTEA HepaRG UGT1A1 up	0.83 (+/- 0.04)	0.80 (+/- 0.04)	0.78 (+/- 0.04)	0.79 (+/- 0.08)	0.78 (+/- 0.04)	0.77 (+/- 0.04)
ToxCast - NHEERL ZF 144hpf TERATOSCORE up	0.82 (+/- 0.04)	0.84 (+/- 0.03)	0.79 (+/- 0.05)	0.77 (+/- 0.08)	0.82 (+/- 0.10)	0.77 (+/- 0.05)
ToxCast - OT AR ARSRC1 0480	0.81 (+/- 0.02)	0.93 (+/- 0.03)	0.80 (+/- 0.01)	0.85 (+/- 0.04)	0.79 (+/- 0.01)	0.72 (+/- 0.02)
ToxCast - OT AR ARSRC1 0960	0.81 (+/- 0.02)	0.95 (+/- 0.02)	0.80 (+/- 0.01)	0.82 (+/- 0.04)	0.79 (+/- 0.01)	0.75 (+/- 0.02)
ToxCast - OT ER ERaERb 0480	0.81 (+/- 0.02)	0.97 (+/- 0.02)	0.83 (+/- 0.02)	0.85 (+/- 0.05)	0.83 (+/- 0.02)	0.74 (+/- 0.02)
ToxCast - OT ER ERaERb 1440	0.82 (+/- 0.01)	0.83 (+/- 0.04)	0.78 (+/- 0.01)	0.81 (+/- 0.07)	0.78 (+/- 0.02)	0.71 (+/- 0.01)
ToxCast - OT ER ERbERb 0480	0.81 (+/- 0.01)	0.96 (+/- 0.03)	0.81 (+/- 0.01)	0.80 (+/- 0.06)	0.81 (+/- 0.02)	0.70 (+/- 0.02)
ToxCast - OT ER ERbERb 1440	0.82 (+/- 0.02)	0.68 (+/- 0.03)	0.73 (+/- 0.02)	0.80 (+/- 0.06)	0.72 (+/- 0.04)	0.65 (+/- 0.02)
ToxCast - OT FXR FXRSRC1 0480	0.82 (+/- 0.03)	0.96 (+/- 0.02)	0.81 (+/- 0.03)	0.83 (+/- 0.04)	0.80 (+/- 0.03)	0.76 (+/- 0.04)
ToxCast - OT FXR FXRSRC1 1440	0.81 (+/- 0.01)	0.85 (+/- 0.02)	0.78 (+/- 0.01)	0.80 (+/- 0.05)	0.77 (+/- 0.02)	0.71 (+/- 0.01)
ToxCast - TOX21 AP1 BLA Agonist ratio	0.81 (+/- 0.01)	0.91 (+/- 0.01)	0.79 (+/- 0.01)	0.84 (+/- 0.03)	0.79 (+/- 0.01)	0.69 (+/- 0.01)

ToxCast - TOX21 ARE BLA agonist ratio	0.81 (+/- 0.02)	0.88 (+/- 0.01)	0.78 (+/- 0.02)	0.79 (+/- 0.05)	0.78 (+/- 0.02)	0.73 (+/- 0.02)
ToxCast - TOX21 AR BLA Agonist ratio	0.81 (+/- 0.01)	0.92 (+/- 0.02)	0.80 (+/- 0.01)	0.77 (+/- 0.09)	0.80 (+/- 0.01)	0.60 (+/- 0.01)
ToxCast - TOX21 AR BLA Antagonist ratio	0.80 (+/- 0.01)	0.97 (+/- 0.00)	0.80 (+/- 0.01)	0.79 (+/- 0.02)	0.80 (+/- 0.02)	0.73 (+/- 0.01)
ToxCast - TOX21 AR LUC MDAKB2 Agonist	0.80 (+/- 0.01)	0.78 (+/- 0.03)	0.75 (+/- 0.01)	0.82 (+/- 0.06)	0.74 (+/- 0.02)	0.54 (+/- 0.00)
ToxCast - TOX21 AR LUC MDAKB2 Antagonist 0.5nM R1881	0.81 (+/- 0.01)	0.97 (+/- 0.01)	0.84 (+/- 0.01)	0.87 (+/- 0.01)	0.83 (+/- 0.01)	0.78 (+/- 0.01)
ToxCast - TOX21 AR LUC MDAKB2 Antagonist 10nM R1881	0.81 (+/- 0.01)	0.97 (+/- 0.01)	0.81 (+/- 0.01)	0.82 (+/- 0.02)	0.81 (+/- 0.00)	0.69 (+/- 0.01)
ToxCast - TOX21 AhR LUC Agonist	0.80 (+/- 0.02)	0.98 (+/- 0.01)	0.81 (+/- 0.02)	0.86 (+/- 0.02)	0.81 (+/- 0.02)	0.67 (+/- 0.02)
ToxCast - TOX21 Aromatase Inhibition	0.81 (+/- 0.02)	0.94 (+/- 0.01)	0.80 (+/- 0.02)	0.81 (+/- 0.04)	0.80 (+/- 0.02)	0.69 (+/- 0.03)
ToxCast - TOX21 CAR Agonist	0.81 (+/- 0.01)	0.97 (+/- 0.01)	0.83 (+/- 0.01)	0.85 (+/- 0.02)	0.83 (+/- 0.01)	0.71 (+/- 0.01)
ToxCast - TOX21 CAR Antagonist	0.80 (+/- 0.01)	0.90 (+/- 0.02)	0.78 (+/- 0.01)	0.76 (+/- 0.05)	0.78 (+/- 0.01)	0.62 (+/- 0.02)
ToxCast - TOX21 DT40	0.80 (+/- 0.01)	0.96 (+/- 0.01)	0.80 (+/- 0.01)	0.79 (+/- 0.01)	0.80 (+/- 0.01)	0.78 (+/- 0.01)
ToxCast - TOX21 DT40 100	0.80 (+/- 0.01)	0.97 (+/- 0.01)	0.80 (+/- 0.01)	0.80 (+/- 0.03)	0.79 (+/- 0.01)	0.79 (+/- 0.01)
ToxCast - TOX21 DT40 657	0.80 (+/- 0.01)	0.93 (+/- 0.01)	0.78 (+/- 0.01)	0.79 (+/- 0.02)	0.78 (+/- 0.01)	0.77 (+/- 0.01)
ToxCast - TOX21 ERR Agonist	0.81 (+/- 0.01)	0.94 (+/- 0.03)	0.80 (+/- 0.01)	0.84 (+/- 0.04)	0.80 (+/- 0.01)	0.56 (+/- 0.01)
ToxCast - TOX21 ERR Antagonist	0.81 (+/- 0.02)	0.97 (+/- 0.01)	0.80 (+/- 0.02)	0.82 (+/- 0.04)	0.80 (+/- 0.02)	0.75 (+/- 0.02)
ToxCast - TOX21 ERa BLA Agonist ratio	0.81 (+/- 0.02)	0.96 (+/- 0.03)	0.80 (+/- 0.02)	0.83 (+/- 0.05)	0.80 (+/- 0.02)	0.58 (+/- 0.01)

ToxCast - TOX21 ERa BLA Antagonist ratio	0.80 (+/- 0.02)	0.98 (+/- 0.01)	0.80 (+/- 0.02)	0.81 (+/- 0.05)	0.80 (+/- 0.02)	0.69 (+/- 0.02)
ToxCast - TOX21 ERa LUC VM7 Agonist	0.81 (+/- 0.01)	0.65 (+/- 0.02)	0.70 (+/- 0.02)	0.76 (+/- 0.02)	0.69 (+/- 0.02)	0.61 (+/- 0.01)
ToxCast - TOX21 ERa LUC VM7 Antagonist 0.1nM E2	0.80 (+/- 0.01)	0.99 (+/- 0.00)	0.80 (+/- 0.01)	0.82 (+/- 0.04)	0.80 (+/- 0.01)	0.69 (+/- 0.01)
ToxCast - TOX21 ERa LUC VM7 Antagonist 0.5nM E2	0.81 (+/- 0.01)	0.98 (+/- 0.01)	0.81 (+/- 0.01)	0.83 (+/- 0.05)	0.80 (+/- 0.01)	0.68 (+/- 0.01)
ToxCast - TOX21 ERb BLA Antagonist ratio	0.80 (+/- 0.02)	0.99 (+/- 0.01)	0.80 (+/- 0.02)	0.81 (+/- 0.02)	0.79 (+/- 0.02)	0.74 (+/- 0.02)
ToxCast - TOX21 ESRE BLA ratio	0.81 (+/- 0.01)	0.94 (+/- 0.02)	0.79 (+/- 0.01)	0.78 (+/- 0.06)	0.80 (+/- 0.01)	0.54 (+/- 0.01)
ToxCast - TOX21 FXR BLA antagonist ratio	0.81 (+/- 0.02)	0.98 (+/- 0.01)	0.80 (+/- 0.01)	0.82 (+/- 0.02)	0.80 (+/- 0.02)	0.71 (+/- 0.02)
ToxCast - TOX21 GR BLA Agonist ratio	0.80 (+/- 0.02)	0.98 (+/- 0.01)	0.81 (+/- 0.01)	0.85 (+/- 0.06)	0.81 (+/- 0.01)	0.60 (+/- 0.01)
ToxCast - TOX21 GR BLA Antagonist ratio	0.81 (+/- 0.01)	0.95 (+/- 0.02)	0.80 (+/- 0.01)	0.80 (+/- 0.05)	0.80 (+/- 0.01)	0.65 (+/- 0.01)
ToxCast - TOX21 H2AX HTRF CHO Agonist ratio	0.81 (+/- 0.02)	0.87 (+/- 0.02)	0.78 (+/- 0.02)	0.82 (+/- 0.06)	0.78 (+/- 0.02)	0.60 (+/- 0.01)
ToxCast - TOX21 HDAC Inhibition	0.81 (+/- 0.01)	0.94 (+/- 0.00)	0.86 (+/- 0.01)	0.88 (+/- 0.03)	0.86 (+/- 0.02)	0.69 (+/- 0.01)
ToxCast - TOX21 HRE BLA Agonist ratio	0.81 (+/- 0.01)	0.98 (+/- 0.01)	0.80 (+/- 0.01)	0.82 (+/- 0.05)	0.80 (+/- 0.01)	0.57 (+/- 0.01)
ToxCast - TOX21 HSE BLA agonist ratio	0.79 (+/- 0.01)	0.86 (+/- 0.03)	0.75 (+/- 0.01)	0.76 (+/- 0.03)	0.75 (+/- 0.01)	0.57 (+/- 0.01)
ToxCast - TOX21 MMP fitc	0.80 (+/- 0.01)	0.96 (+/- 0.02)	0.84 (+/- 0.02)	0.84 (+/- 0.03)	0.84 (+/- 0.02)	0.64 (+/- 0.02)
ToxCast - TOX21 MMP ratio down	0.81 (+/- 0.01)	0.95 (+/- 0.01)	0.85 (+/- 0.01)	0.87 (+/- 0.02)	0.85 (+/- 0.01)	0.78 (+/- 0.01)

ToxCast - TOX21 MMP ratio up	0.80 (+/- 0.01)	0.97 (+/- 0.02)	0.82 (+/- 0.02)	0.85 (+/- 0.06)	0.82 (+/- 0.02)	0.59 (+/- 0.01)
ToxCast - TOX21 MMP rhodamine	0.79 (+/- 0.01)	0.93 (+/- 0.00)	0.85 (+/- 0.01)	0.85 (+/- 0.02)	0.85 (+/- 0.01)	0.76 (+/- 0.01)
ToxCast - TOX21 PGC ERR Agonist	0.81 (+/- 0.00)	0.91 (+/- 0.03)	0.80 (+/- 0.01)	0.77 (+/- 0.09)	0.80 (+/- 0.01)	0.56 (+/- 0.01)
ToxCast - TOX21 PGC ERR Antagonist	0.81 (+/- 0.03)	0.92 (+/- 0.02)	0.80 (+/- 0.02)	0.81 (+/- 0.04)	0.79 (+/- 0.03)	0.69 (+/- 0.03)
ToxCast - TOX21 PPARd BLA antagonist ratio	0.81 (+/- 0.01)	0.89 (+/- 0.01)	0.78 (+/- 0.01)	0.81 (+/- 0.01)	0.78 (+/- 0.01)	0.61 (+/- 0.01)
ToxCast - TOX21 PPARg BLA antagonist ratio	0.80 (+/- 0.02)	0.97 (+/- 0.01)	0.79 (+/- 0.02)	0.79 (+/- 0.05)	0.79 (+/- 0.02)	0.65 (+/- 0.02)
ToxCast - TOX21 PR BLA Antagonist ratio	0.81 (+/- 0.01)	0.97 (+/- 0.01)	0.83 (+/- 0.01)	0.83 (+/- 0.01)	0.83 (+/- 0.01)	0.80 (+/- 0.01)
ToxCast - TOX21 RAR LUC Agonist	0.80 (+/- 0.02)	0.98 (+/- 0.01)	0.81 (+/- 0.02)	0.85 (+/- 0.05)	0.81 (+/- 0.02)	0.57 (+/- 0.01)
ToxCast - TOX21 RAR LUC Antagonist	0.80 (+/- 0.01)	0.94 (+/- 0.02)	0.79 (+/- 0.01)	0.79 (+/- 0.03)	0.79 (+/- 0.01)	0.66 (+/- 0.02)
ToxCast - TOX21 RORg LUC CHO Antagonist	0.81 (+/- 0.01)	0.98 (+/- 0.01)	0.81 (+/- 0.01)	0.82 (+/- 0.03)	0.81 (+/- 0.02)	0.68 (+/- 0.01)
ToxCast - TOX21 RXR BLA Agonist ratio	0.81 (+/- 0.01)	0.48 (+/- 0.03)	0.61 (+/- 0.01)	0.70 (+/- 0.03)	0.61 (+/- 0.01)	0.45 (+/- 0.01)
ToxCast - TOX21 SBE BLA Antagonist ratio	0.80 (+/- 0.02)	0.99 (+/- 0.00)	0.80 (+/- 0.02)	0.80 (+/- 0.01)	0.80 (+/- 0.02)	0.70 (+/- 0.02)
ToxCast - TOX21 SSH 3T3 GLI3 Antagonist	0.81 (+/- 0.01)	0.89 (+/- 0.02)	0.78 (+/- 0.01)	0.79 (+/- 0.04)	0.78 (+/- 0.02)	0.72 (+/- 0.01)
ToxCast - TOX21 TR LUC GH3 Antagonist	0.81 (+/- 0.01)	0.98 (+/- 0.00)	0.82 (+/- 0.01)	0.84 (+/- 0.01)	0.82 (+/- 0.01)	0.79 (+/- 0.01)
ToxCast - TOX21 TSHR Agonist ratio	0.80 (+/- 0.01)	0.90 (+/- 0.01)	0.78 (+/- 0.01)	0.83 (+/- 0.04)	0.77 (+/- 0.01)	0.57 (+/- 0.01)
ToxCast - TOX21 TSHR Antagonist ratio	0.81 (+/- 0.01)	0.91 (+/- 0.03)	0.79 (+/- 0.00)	0.83 (+/- 0.07)	0.78 (+/- 0.01)	0.54 (+/- 0.01)

ToxCast - TOX21 VDR BLA antagonist ratio	0.81 (+/- 0.02)	0.93 (+/- 0.01)	0.79 (+/- 0.02)	0.82 (+/- 0.05)	0.79 (+/- 0.02)	0.60 (+/- 0.01)
ToxCast - TOX21 p53 BLA p1 ratio	0.81 (+/- 0.02)	0.92 (+/- 0.01)	0.80 (+/- 0.02)	0.86 (+/- 0.06)	0.79 (+/- 0.02)	0.64 (+/- 0.02)
ToxCast - TOX21 p53 BLA p2 ratio	0.81 (+/- 0.01)	0.88 (+/- 0.01)	0.78 (+/- 0.01)	0.81 (+/- 0.03)	0.78 (+/- 0.01)	0.65 (+/- 0.01)
ToxCast - TOX21 p53 BLA p3 ratio	0.81 (+/- 0.01)	0.87 (+/- 0.01)	0.78 (+/- 0.01)	0.83 (+/- 0.02)	0.78 (+/- 0.01)	0.63 (+/- 0.01)
ToxCast - TOX21 p53 BLA p4 ratio	0.80 (+/- 0.01)	0.92 (+/- 0.02)	0.79 (+/- 0.01)	0.79 (+/- 0.01)	0.79 (+/- 0.01)	0.64 (+/- 0.01)
ToxCast - TOX21 p53 BLA p5 ratio	0.81 (+/- 0.01)	0.88 (+/- 0.01)	0.78 (+/- 0.01)	0.78 (+/- 0.02)	0.78 (+/- 0.01)	0.63 (+/- 0.01)
ToxCast - UPITT HCI U2OS AR TIF2 Nucleoli Agonist	0.80 (+/- 0.03)	0.77 (+/- 0.03)	0.74 (+/- 0.04)	0.77 (+/- 0.09)	0.73 (+/- 0.05)	0.63 (+/- 0.04)
ToxCast - UPITT HCI U2OS AR TIF2 Nucleoli Antagonist	0.82 (+/- 0.03)	0.81 (+/- 0.03)	0.78 (+/- 0.03)	0.74 (+/- 0.03)	0.79 (+/- 0.04)	0.73 (+/- 0.03)
ToxCast - UPITT HCI U2OS AR TIF2 Nucleoli Cytoplasm Ratio Antagonist	0.82 (+/- 0.03)	0.78 (+/- 0.01)	0.77 (+/- 0.04)	0.83 (+/- 0.06)	0.76 (+/- 0.04)	0.66 (+/- 0.04)
ToxCast - NCCT HEK293T CellTiterGLO	0.83 (+/- 0.03)	0.84 (+/- 0.03)	0.80 (+/- 0.04)	0.83 (+/- 0.03)	0.76 (+/- 0.06)	0.79 (+/- 0.04)

[1] Numbers in parentheses indicate the standard deviation.

Table S7. Average Performance of the CP Models Generated Based on the Complete Set of Features.[1]

Endpoint	Descriptor	Validity	Efficiency	Overall accuracy	Accuracy active	Accuracy inactive	F1 score	MCC	Sensitivity	Specificity
MNT	CHEM	0,79 (+/- 0.01)	0,73 (+/- 0.03)	0,71 (+/- 0.02)	0,78 (+/- 0.06)	0,70 (+/- 0.03)	0,65 (+/- 0.02)	0,38 (+/- 0.04)	0,70 (+/- 0.03)	0,78 (+/- 0.06)
	BIO	0,82 (+/- 0.02)	0,75 (+/- 0.04)	0,76 (+/- 0.02)	0,78 (+/- 0.09)	0,76 (+/- 0.04)	0,69 (+/- 0.02)	0,44 (+/- 0.05)	0,76 (+/- 0.04)	0,78 (+/- 0.09)
	CHEMBIO	0,83 (+/- 0.02)	0,77 (+/- 0.03)	0,78 (+/- 0.03)	0,79 (+/- 0.07)	0,78 (+/- 0.03)	0,70 (+/- 0.03)	0,46 (+/- 0.05)	0,77 (+/- 0.03)	0,79 (+/- 0.07)
DILI	CHEM	0,81 (+/- 0.04)	0,82 (+/- 0.06)	0,77 (+/- 0.03)	0,77 (+/- 0.05)	0,77 (+/- 0.08)	0,76 (+/- 0.03)	0,53 (+/- 0.06)	0,77 (+/- 0.08)	0,77 (+/- 0.05)
	BIO	0,82 (+/- 0.03)	0,79 (+/- 0.05)	0,77 (+/- 0.03)	0,77 (+/- 0.04)	0,77 (+/- 0.08)	0,76 (+/- 0.03)	0,53 (+/- 0.07)	0,77 (+/- 0.08)	0,77 (+/- 0.04)
	CHEMBIO	0,82 (+/- 0.03)	0,80 (+/- 0.06)	0,78 (+/- 0.03)	0,78 (+/- 0.03)	0,78 (+/- 0.07)	0,77 (+/- 0.03)	0,54 (+/- 0.06)	0,78 (+/- 0.07)	0,78 (+/- 0.03)
DICC	CHEM	0,82 (+/- 0.02)	0,83 (+/- 0.04)	0,78 (+/- 0.01)	0,79 (+/- 0.02)	0,77 (+/- 0.02)	0,76 (+/- 0.01)	0,53 (+/- 0.03)	0,77 (+/- 0.02)	0,79 (+/- 0.02)
	BIO	0,80 (+/- 0.02)	0,98 (+/- 0.01)	0,81 (+/- 0.01)	0,84 (+/- 0.04)	0,80 (+/- 0.02)	0,79 (+/- 0.02)	0,60 (+/- 0.03)	0,80 (+/- 0.02)	0,84 (+/- 0.04)
	CHEMBIO	0,79 (+/- 0.02)	0,97 (+/- 0.01)	0,82 (+/- 0.01)	0,85 (+/- 0.03)	0,80 (+/- 0.02)	0,80 (+/- 0.01)	0,61 (+/- 0.03)	0,80 (+/- 0.02)	0,85 (+/- 0.03)

[1] Numbers in parentheses indicate the standard deviation.

Table S8. Average Performance of the CP Models Generated from a Selected Set of Features.[1]

Endpoint	Descriptor	Validity	Efficiency	Overall accuracy	Accuracy active	Accuracy inactive	F1 score	MCC	Sensitivity	Specificity
MNT	CHEM	0,77 (+/- 0.02)	0,76 (+/- 0.05)	0,70 (+/- 0.02)	0,65 (+/- 0.07)	0,71 (+/- 0.03)	0,61 (+/- 0.02)	0,28 (+/- 0.05)	0,71 (+/- 0.03)	0,65 (+/- 0.07)
	BIO	0,82 (+/- 0.03)	0,81 (+/- 0.05)	0,78 (+/- 0.03)	0,78 (+/- 0.07)	0,78 (+/- 0.04)	0,70 (+/- 0.03)	0,46 (+/- 0.06)	0,78 (+/- 0.04)	0,78 (+/- 0.07)
	CHEMBIO	0,81 (+/- 0.03)	0,85 (+/- 0.03)	0,76 (+/- 0.03)	0,76 (+/- 0.06)	0,78 (+/- 0.03)	0,70 (+/- 0.03)	0,44 (+/- 0.07)	0,78 (+/- 0.03)	0,76 (+/- 0.06)
DILI	CHEM	0,78 (+/- 0.05)	0,91 (+/- 0.04)	0,77 (+/- 0.05)	0,77 (+/- 0.08)	0,72 (+/- 0.04)	0,74 (+/- 0.05)	0,49 (+/- 0.09)	0,72 (+/- 0.04)	0,77 (+/- 0.08)
	BIO	0,81 (+/- 0.04)	0,83 (+/- 0.07)	0,75 (+/- 0.03)	0,75 (+/- 0.05)	0,79 (+/- 0.08)	0,76 (+/- 0.04)	0,53 (+/- 0.07)	0,79 (+/- 0.08)	0,75 (+/- 0.05)
	CHEMBIO	0,81 (+/- 0.03)	0,88 (+/- 0.04)	0,78 (+/- 0.03)	0,78 (+/- 0.04)	0,78 (+/- 0.09)	0,77 (+/- 0.03)	0,55 (+/- 0.06)	0,78 (+/- 0.09)	0,78 (+/- 0.04)
DICC	CHEM	0,79 (+/- 0.02)	0,84 (+/- 0.02)	0,74 (+/- 0.02)	0,74 (+/- 0.05)	0,75 (+/- 0.03)	0,72 (+/- 0.03)	0,46 (+/- 0.05)	0,75 (+/- 0.03)	0,74 (+/- 0.05)
	BIO	0,79 (+/- 0.02)	0,96 (+/- 0.02)	0,86 (+/- 0.01)	0,86 (+/- 0.04)	0,81 (+/- 0.02)	0,81 (+/- 0.01)	0,63 (+/- 0.02)	0,81 (+/- 0.02)	0,86 (+/- 0.04)
	CHEMBIO	0,79 (+/- 0.02)	0,94 (+/- 0.01)	0,86 (+/- 0.01)	0,86 (+/- 0.03)	0,83 (+/- 0.02)	0,82 (+/- 0.01)	0,65 (+/- 0.03)	0,83 (+/- 0.02)	0,86 (+/- 0.03)

[1] Numbers in parentheses indicate the standard deviation.

Table S9. Top Fifteen Most Important Features in the RF Models based on the CHEMBIO descriptor set without feature selection.[1]

Endpoint	Feature type	Feature[1]	Feature importance[2]
MNT	BIO	p1 AMES	1.00
	BIO	p0 AMES	1.00
	BIO	p1 eMolTox - Mutagenicity	0.96
	BIO	p0 eMolTox - Mutagenicity	0.90
	BIO	p1 eMolTox - Agonist of the p53 signaling pathway	0.76
	BIO	p0 eMolTox - Agonist of the p53 signaling pathway	0.61
	BIO	p1 ToxCast - TOX21 CAR Antagonist	0.52
	BIO	p0 ToxCast - TOX21 CAR Antagonist	0.52
	BIO	p1 ToxCast - TOX21 p53 BLA p3 ratio	0.47
	BIO	p1 eMolTox - Antagonist of the farnesoid-X-receptor (FXR) signaling pathway	0.40
	BIO	p1 ToxCast - TOX21 p53 BLA p5 ratio	0.39
	BIO	p0 eMolTox - Activator the aryl hydrocarbon receptor (AhR) signaling pathway	0.38
	BIO	p0 eMolTox - Antagonist of the farnesoid-X-receptor (FXR) signaling pathway	0.38
	BIO	p0 ToxCast - TOX21 p53 BLA p5 ratio	0.36
	BIO	p1 ToxCast - TOX21 p53 BLA p2 ratio	0.36
DILI	CHEM	Physicochemical descriptor (smr VSA10)	1.00
	BIO	p0 eMolTox - Modulator of Kappa opioid receptor[3]	0.86
	BIO	p1 eMolTox - Modulator of Kappa opioid receptor[3]	0.84
	BIO	p0 Bioavailability	0.84
	BIO	p1 Bioavailability	0.83
	BIO	p1 eMolTox - Modulator of Mu opioid receptor[3]	0.83
	BIO	p0 eMolTox - Modulator of Mu opioid receptor[3]	0.80
	BIO	p0 eMolTox - Modulator of Muscarinic acetylcholine receptor M4[3]	0.79

	BIO	p1 eMolTox - Modulator of Muscarinic acetylcholine receptor M4[3]	0.65
	BIO	p1 eMolTox - Modulator of Delta opioid receptor[4]	0.63
	BIO	p0 eMolTox - Modulator of Muscarinic acetylcholine receptor M3[3]	0.62
	BIO	p0 eMolTox - Modulator of Delta opioid receptor[4]	0.61
	BIO	p0 eMolTox - Modulator of Muscarinic acetylcholine receptor M2[3]	0.60
	BIO	p1 eMolTox - Modulator of Muscarinic acetylcholine receptor M3[3]	0.59
	BIO	p0 eMolTox - Modulator of Muscarinic acetylcholine receptor M5[3]	0.58
DICC	BIO	p1 Bioavailability	1.00
	BIO	p0 ToxCast - TOX21 ERa LUC VM7 agonist	0.93
	BIO	p1 ToxCast - TOX21 ERa LUC VM7 agonist	0.83
	BIO	p0 Bioavailability	0.81
	BIO	p0 eMolTox - Agonist of the RXR signaling pathway	0.69
	BIO	p0 ToxCast - TOX21 HDAC Inhibition	0.60
	BIO	p0 ToxCast - TOX21 RXR BLA Agonist ratio	0.56
	BIO	p1 eMolTox - Agonist of the RXR signaling pathway	0.52
	BIO	p1 eMolTox - Induce Phospholipidosis	0.50
	BIO	p1 ToxCast - TOX21 p53 BLA p2 ratio	0.49
	BIO	p0 ToxCast - TOX21 p53 BLA p3 ratio	0.49
	BIO	p1 eMolTox - Differential cytotoxicity against isogenic chicken DT40 cell lines with known DNA damage response pathways - Rad54Ku70 mutant cell line	0.49
	BIO	p1 ToxCast - TOX21 p53 BLA p3 ratio	0.47
	BIO	p0 ToxCast - TOX21 DT40 657	0.47
	BIO	p1 ToxCast - TOX21 DT40 657	0.47

[1] p0 and p1 denote the predicted p-values for the inactive (0) and active (1) classes.

[2] Normalized mean feature importance over 5-fold CV.

[3] Features related to modulators of G protein-coupled receptors.

Table S10. Top Fifteen Features Obtained With the Lasso Models Based on the CHEMBIO Descriptor Set.

Endpoint	Feature type	Feature[1]	Mean coefficient[2]
MNT	BIO	p1 AMES	1.00
	BIO	p1 eMolTox - Agonist of the p53 signaling pathway	0.70
	CHEM	Fingerprint (byte vector 1994)	0.28
	CHEM	Fingerprint (byte vector 1730)	0.25
	BIO	p0 ToxCast - BSK BE3C IP10 down	0.25
	CHEM	Fingerprint (byte vector 1809)	0.23
	BIO	p0 ToxCast - BSK KF3CT MCP1 down	0.21
	BIO	p0 Chromosome aberration	0.19
	BIO	p0 PGP inhibition	0.19
	BIO	p1 ToxCast - TOX21 CAR Antagonist	0.18
	CHEM	Fingerprint (byte vector 302)	0.17
	CHEM	Fingerprint (byte vector 1375)	0.16
	CHEM	Fingerprint (byte vector 343)	0.16
	CHEM	Fingerprint (byte vector 853)	0.15
	CHEM	Fingerprint (byte vector 1181)	0.15
DILI	BIO	p1 Bioavailability	1.00
	BIO	p1 eMolTox - Modulator of Mu opioid receptor	0.73
	CHEM	Physicochemical descriptor (peoe VSA3)	0.66
	CHEM	Fingerprint (byte 845)	0.65
	BIO	p1 ToxCast - OT ER ERbERb 1440	0.45
	BIO	p0 eMolTox - Modulator of GABA-A receptor alpha-2beta-3gamma-2	0.45
	CHEM	Physicochemical descriptor (MQN34)	0.43
	BIO	p0 eMolTox - Modulator of Muscarinic acetylcholine receptor M1	0.42
	BIO	p0 ToxCast - ATG PXRE CIS dn	0.39

	BIO	p0 ToxCast - TOX21 ERR Antagonist	0.38
	CHEM	Physicochemical descriptor (MQN25)	0.38
	CHEM	Fingerprint (byte vector1036)	0.35
	BIO	p1 eMolTox – Antagonist of the retinoid-related orphan receptor gamma (ROR-gamma) signaling pathway	0.35
	CHEM	Physicochemical descriptor (fraction of CSP3)	0.34
	BIO	p0 eMolTox – Agonist of the peroxisome proliferator-activated receptor gamma (PPARg) signaling pathway	0.33
DICC	BIO	p0 Bioavailability	1.00
	BIO	p1 Bioavailability	0.96
	BIO	p0 ToxCast - ACEA AR antagonist 80hr	0.27
	BIO	p0 ToxCast - TOX21 ERa LUC VM7 agonist	0.26
	BIO	p0 eMolTox - Cytotoxicity in HEK293 cells - 24 hour	0.23
	BIO	p0 ToxCast - TOX21 AR LUC MDAKB2 agonist	0.22
	BIO	p0 Chromosome aberration	0.20
	BIO	p0 ToxCast - TOX21 p53 BLA p2 ratio	0.17
	BIO	p1 eMolTox - Differential cytotoxicity (isogenic chicken DT40 Rev3 mutant cell line)	0.16
	BIO	p1 AMES	0.14
	BIO	p0 ToxCast - TOX21 Aromatase Inhibition	0.14
	BIO	p1 eMolTox - Modulator_of_Alpha-2a_adrenergic_receptor	0.14
	BIO	p0 eMolTox – Agonist of the peroxisome proliferator-activated receptor gamma (PPARg) signaling pathway	0.13
	BIO	p1 ToxCast – ATG ERE CIS up	0.13
	BIO	p0 ToxCast - TOX21 HDAC Inhibition	0.11

[1] p0 and p1 denote the predicted p-values for the inactive (0) and active (1) classes.

[2] Normalized mean lasso coefficients over 5-fold CV.

Supporting Information for [P3]

This appendix contains the supporting information for the publication:

Morger, A.; Garcia de Lomana, M.; Norinder, U.; Svensson, F.; Kirchmair, J.; Mathea, M. and Volkamer, A. Studying and mitigating the effects of data drifts on ML model performance at the example of chemical toxicity data, *Sci. Rep.*, **2022**, 12, 7244.

Additional file

Studying and mitigating the effects of data drifts on ML model performance at the example of chemical toxicity data

Andrea Morger, Marina Garcia de Lomana Rodriguez, Ulf Norinder, Fredrik Svensson, Johannes Kirchmair, Miriam Mathea and Andrea Volkamer

A1 Additional information on data and methods

A1.1 Target selection for the ChEMBL datasets

Target datasets were selected following a collection of 1360 ligand sets provided by Škuta et al.[31] for similarity searching, bioactivity classification and scaffold hopping. First, the 29 target datasets, for which Škuta et al. found $\geq$ 1000 compounds with reported pIC50 values, were downloaded, including pIC50 values and publication year. The following cleaning procedure was applied to each target dataset: If there were multiple measurements per compound and endpoint, the mean and standard deviation were calculated. Only the mean measurement of those duplicates was kept if the standard deviation was lower or equal than 0.5, otherwise they were discarded. The oldest publication year (i.e. lowest number) was kept for aggregated data points. The compounds were standardised as described in the main manuscript (section 2.1.2) and temporally split into training, update1, update2, and holdout set as explained in 2.1.4. If fewer than 50 active and 50 inactive compounds were left in the holdout set after the time-split, the target dataset was excluded from the study. Finally, 20 targets remained which match the filtering criteria. Of these, a total of twelve targets were selected that are linked to toxicity. A target was defined to be associated to toxicity if it was either assayed in ToxCast[33], or part of the list of targets that are recommended to early assess the potential hazard of a compound[34].

A1.2 Public datasets for liver toxicity and MNT

To assess drifts between data originating from different sources, public and proprietary datasets for liver toxicity and micro nucleus test (MNT) were collected. For CP model training, the same public datasets for liver toxicity (more specifically here drug-induced liver injury (DILI)) and MNT in vivo were used as described by Garcia de Lomana et al.[12]. Data for the DILI endpoint were gathered from the U.S. Food and Drug Administration (FDA)[64] and for the MNT in vivo endpoint from three sources (eChemPortal[65], the work of Benigni et al.[66] and Yoo et al.[67]). The respective datasets contain 692 (445 active and 247 inactive compounds) and 1791 compounds (316 active and 1475 inactive compounds) after the data pre-processing and deduplication steps conducted by Garcia de Lomana et al.[12].

A1.3 Inhouse datasets for liver toxicity and MNT

Two inhouse datasets for liver toxicity and MNT in vivo, with data generated by BASF SE, were used as holdout and update set to investigate data drifts between data with different origin. Liver toxicity was measured in oral assays on rats (including OECD Guidelines 407, 408 and 422, as well as range finding oral studies). Compounds showing adverse or adaptive effects in the liver in any of these studies were labelled as active. MNT in vivo was determined in mice in an assay following the OECD Guideline 474 or in (non-GLP) screening assays (with 18 animals). The liver toxicity dataset contains 140 (63 active and 77 inactive) compounds and the MNT in vivo dataset contains 366 (194 active and 172 inactive) compounds after the data pre-processing and deduplication steps (following the same procedure as Garcia de Lomana et al.[12], see "Chemical structure standardisation").

A1.4 Time-splitting procedure

Note that all compounds published (ChEMBL data) or assayed (inhouse data) in the same year were assigned to the same split.

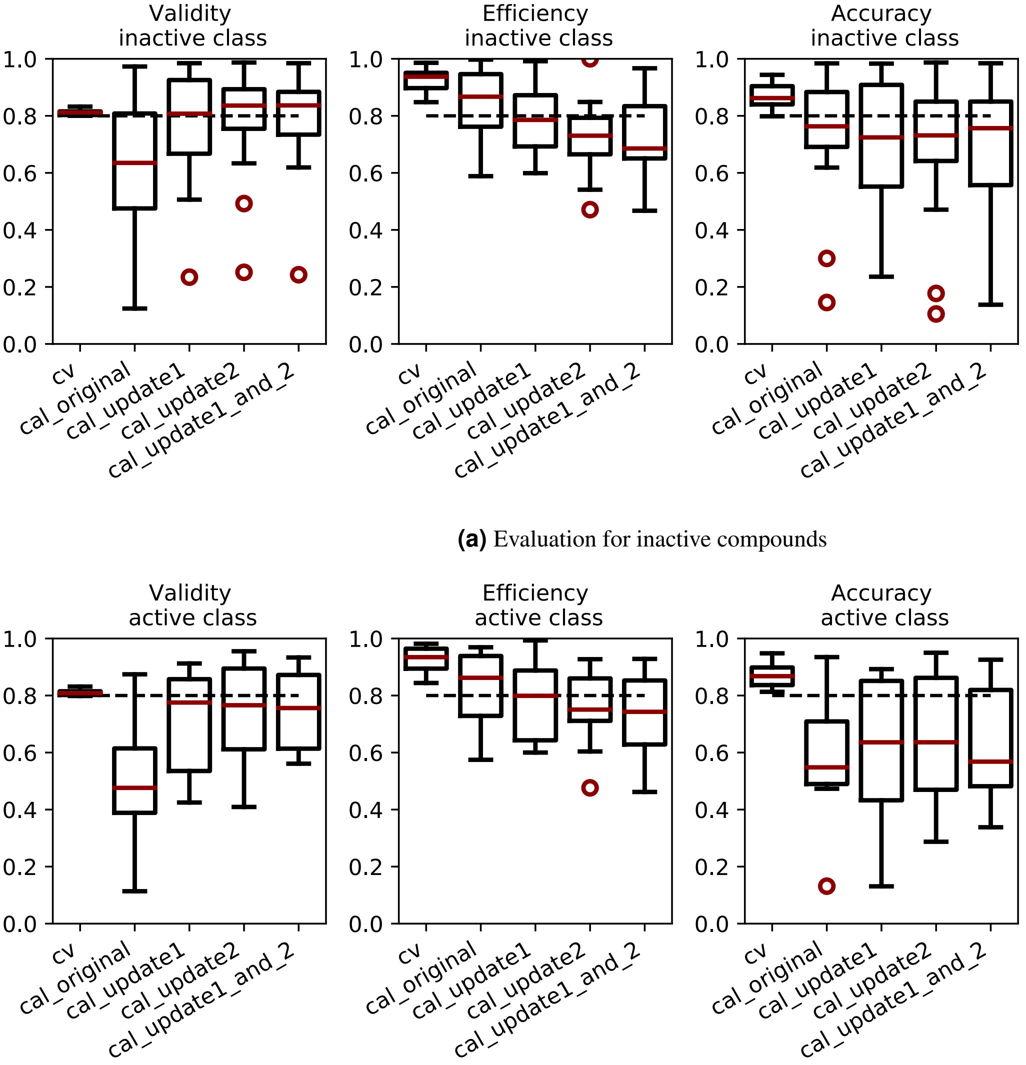
ChEMBL data After standardising the compounds (see 2.1.2), the ChEMBL data were time-split into four datasets, i.e. train, update1, update2, and holdout set based on the publication year. A minimum number of compounds per dataset was defined based on a predefined ratio, i.e. the training set must contain at least 50% of the total number of compounds, the update1 and update2 sets must contain at least 12% each. Starting from the earliest year, all compounds published in that year were assigned to the training set and the number of training compounds was assessed. Same for the next year(s) until the training set contained at least the minimum number of training compounds defined. Then, all compounds published in the following year(s) were assigned to the update1 set until the respective threshold was reached. With the same procedure, the compounds published in the subsequent year(s) were allocated to the update2 set. All remaining compounds belong to the holdout set. The number of active and inactive compounds available per subset of the twelve holdout ChEMBL target datasets, as well as the corresponding time thresholds for splitting, are provided in Table 2.

Liver toxicity and MNT data To investigate the occurrence of discrepancies between external and internal data (see A1.2), the liver toxicity and MNT datasets were investigated. The external data were used for model building as well as for the original calibration set. The internal data were time-split into update and holdout set based on the date they were measured internally. Due to the small number of available inhouse compounds, only one update set was deducted. The data was selected by year as described for the ChEMBL data until at least 50% of the compounds were assigned to the update set. The number of training, update and holdout compounds available for the liver toxicity and MNT endpoints are shown in Table 2.

Table S1. ChEMBL datasets and their biological relevance. A selection of possible toxicological or adverse effects due to agonism (or activation) or antagonism (or inhibition) with the targets is provided.

ChEMBL ID	name	toxicological or adverse effects
CHEMBL220	Acetylcholinesterase (human)	decreased blood pressure or heart rate, increased GI motility[34,68]
CHEMBL4078	Acetylcholinesterase (fish)	decreased blood pressure or heart rate, increased GI motility[34,68]
CHEMBL5763	Cholinesterase	decreased heart rate, QT interval prolongation[69]
CHEMBL203	EGFR erbB1	skin toxicity, cardiotoxicity[70,71]
CHEMBL206	Estrogen receptor alpha	antiandrogenic effects, hormone-dependent cancers[72,73]
CHEMBL279	VEGFR 2	hypertension, disturbed wound healing, GI and skin toxicity[74]
CHEMBL230	Cyclooxygenase-2	myocardial infarction, increased blood pressure, ischaemic stroke, atherothrombosi[34,75]
CHEMBL340	Cytochrome P450 3A4	drug-drug interactions, detoxification by metabolism, activation of toxic metabolites[76]
CHEMBL240	HERG	QT interval prolongation[77]
CHEMBL2039	Monoamine oxidase B	cell death[78]
CHEMBL222	Norepinephrine transporter	increased heart rate or blood pressure, constipation[34,79]
CHEMBL228	Serotonin transporter	increased GI motility, insomnia, anxiety, sexual dysfunction[34,80]

A2 Additional information on results



(a) Evaluation for inactive compounds

(b) Evaluation for active compounds

Figure S1. Class-wise time split evaluation (validity, efficiency, accuracy) of CV experiments and predictions for the holdout set using the original (cal_original), update1 (cal_update1), update2 (cal_update2) and combined update1_and_2 (cal_update1_and_2) calibration sets for twelve ChEMBL datasets.

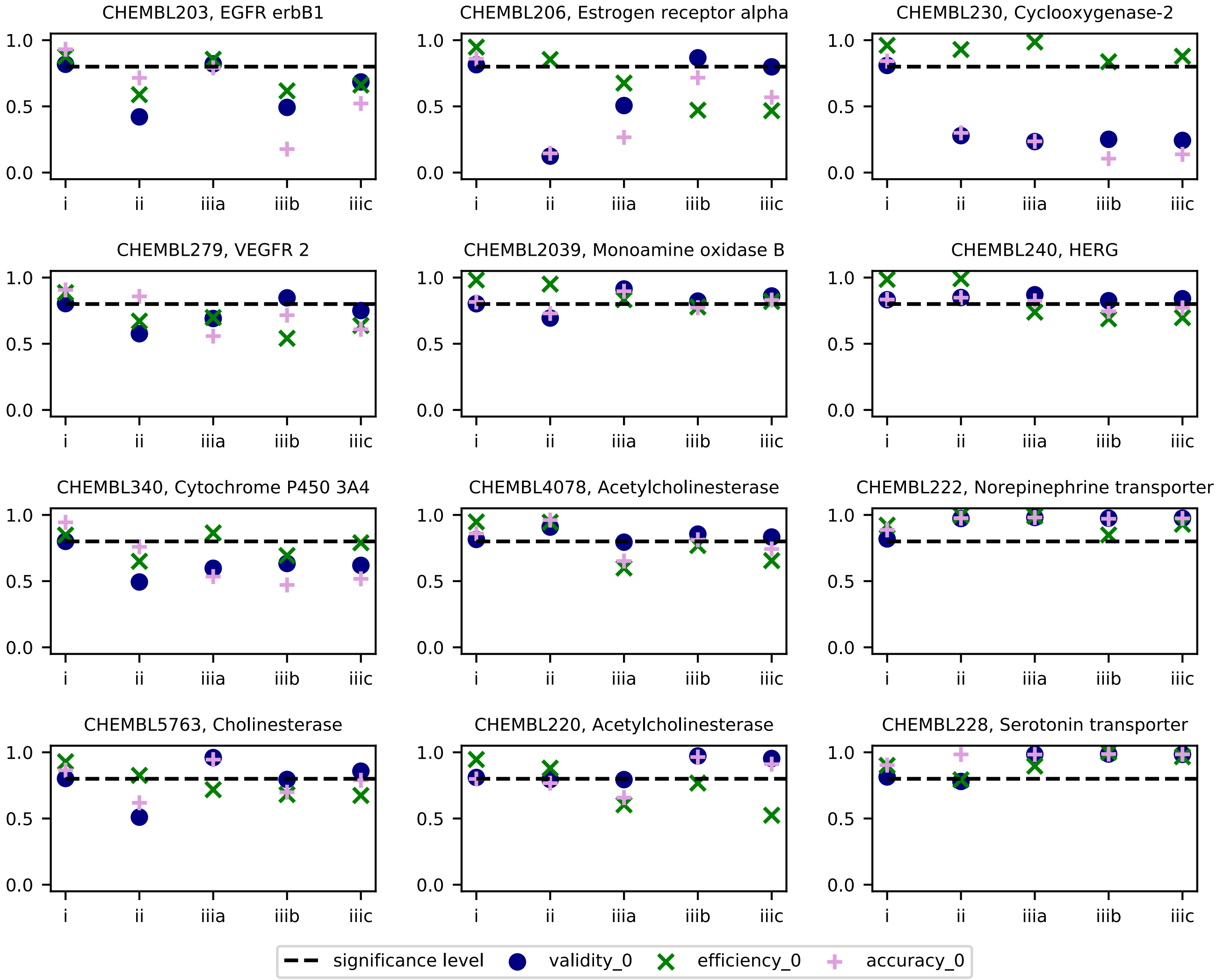


Figure S2. Inactive compounds evaluation of time-split experiments for individual ChEMBL endpoints. i) cross-validation on training data, predict holdout data using ii) original calibration set iiia) update1, iiib) update2, iiic) combined update1+2 calibration sets.

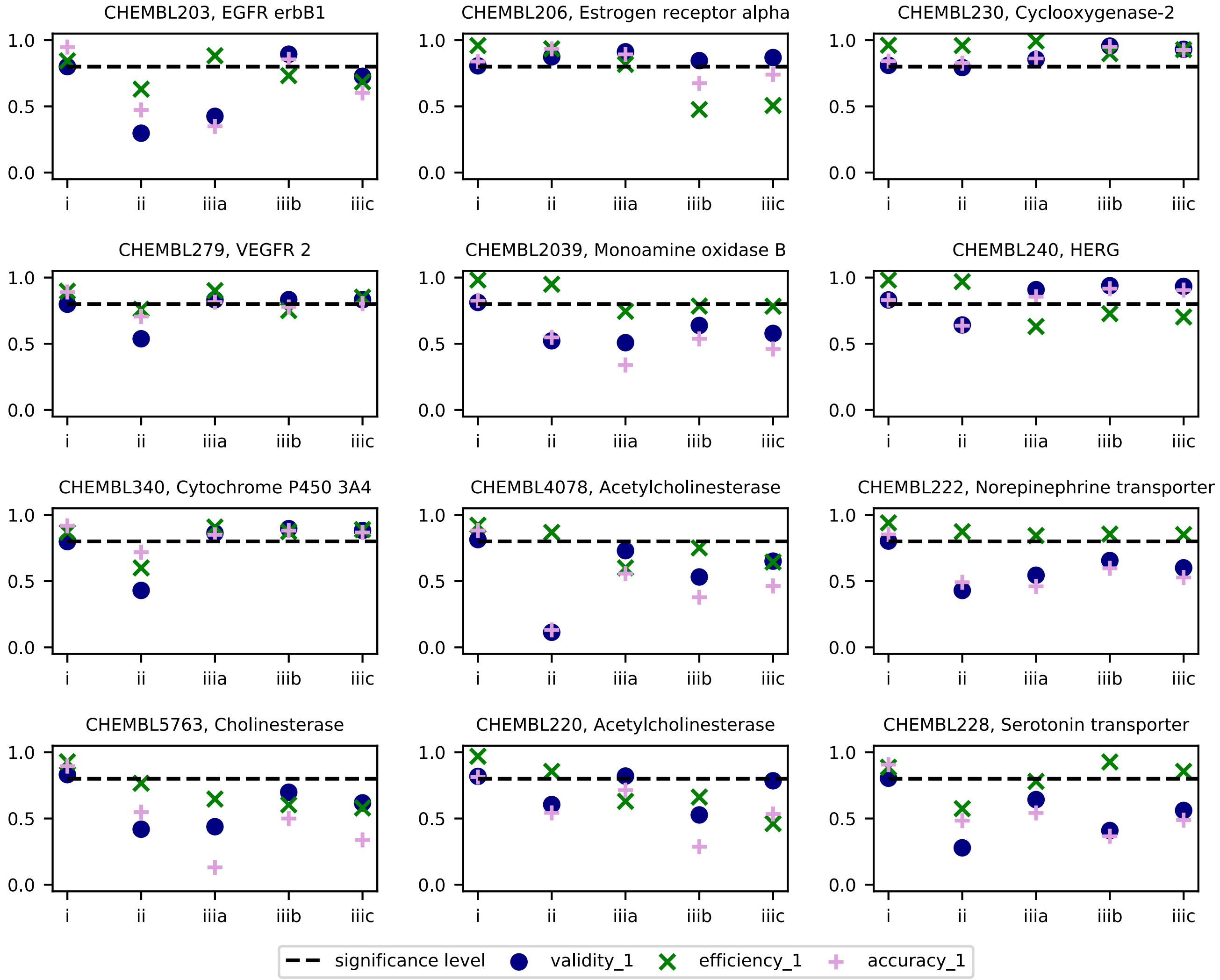


Figure S3. Active compounds evaluation of time-split experiments for individual ChEMBL endpoints. i) cross-validation on training data, predict holdout data using ii) original calibration set iiia) update1, iiib) update2, iiic) combined update1+2 calibration sets.

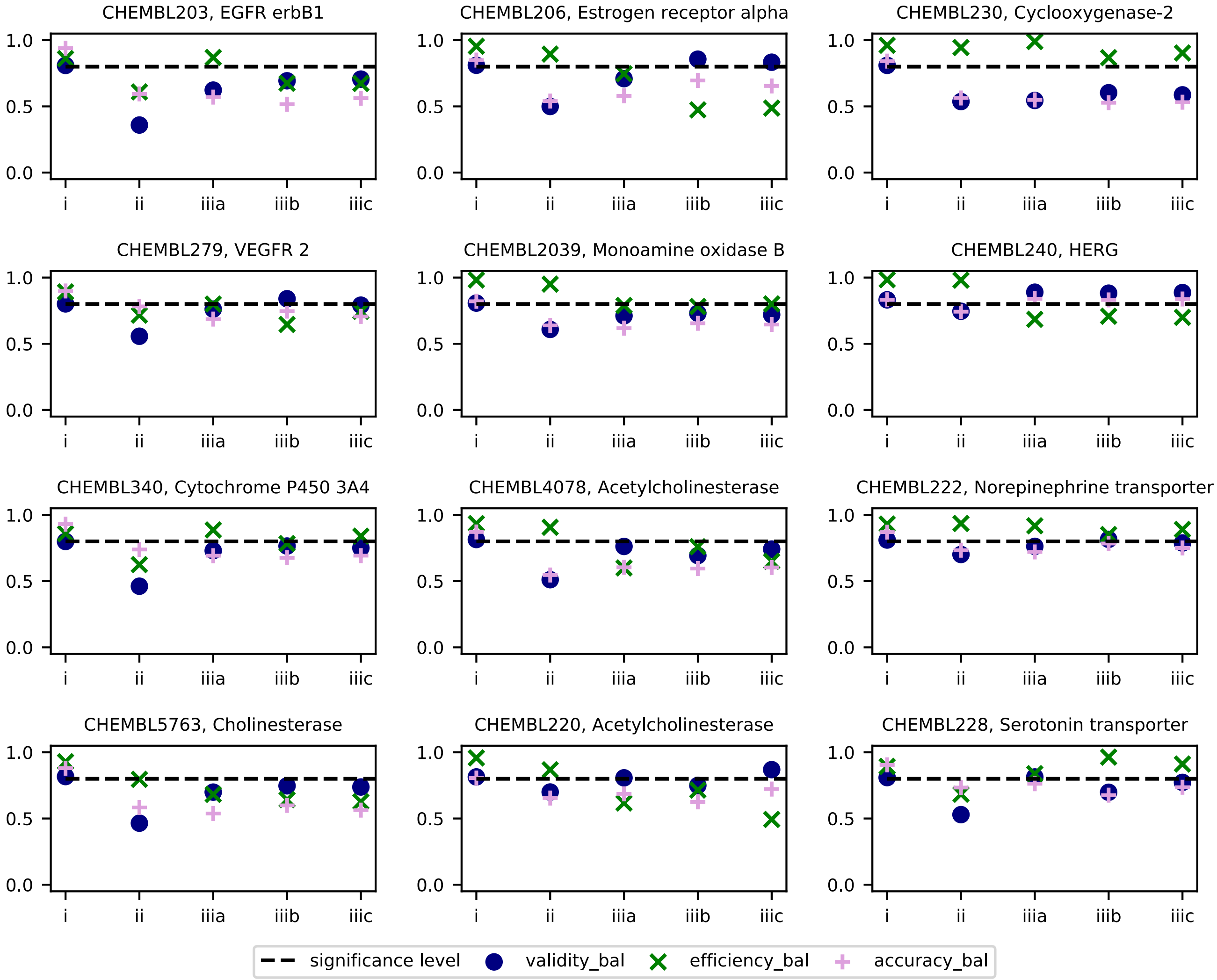


Figure S4. Balanced evaluation of time-split experiments for individual ChEMBL endpoints. i) cross-validation on training data, predict holdout data using ii) original calibration set, iii) updated calibration set, a) update1, b) update2, c) combined update1+2 sets. The doted line at 0.8 denotes the expected validity for the chosen significance level (of 0.2).

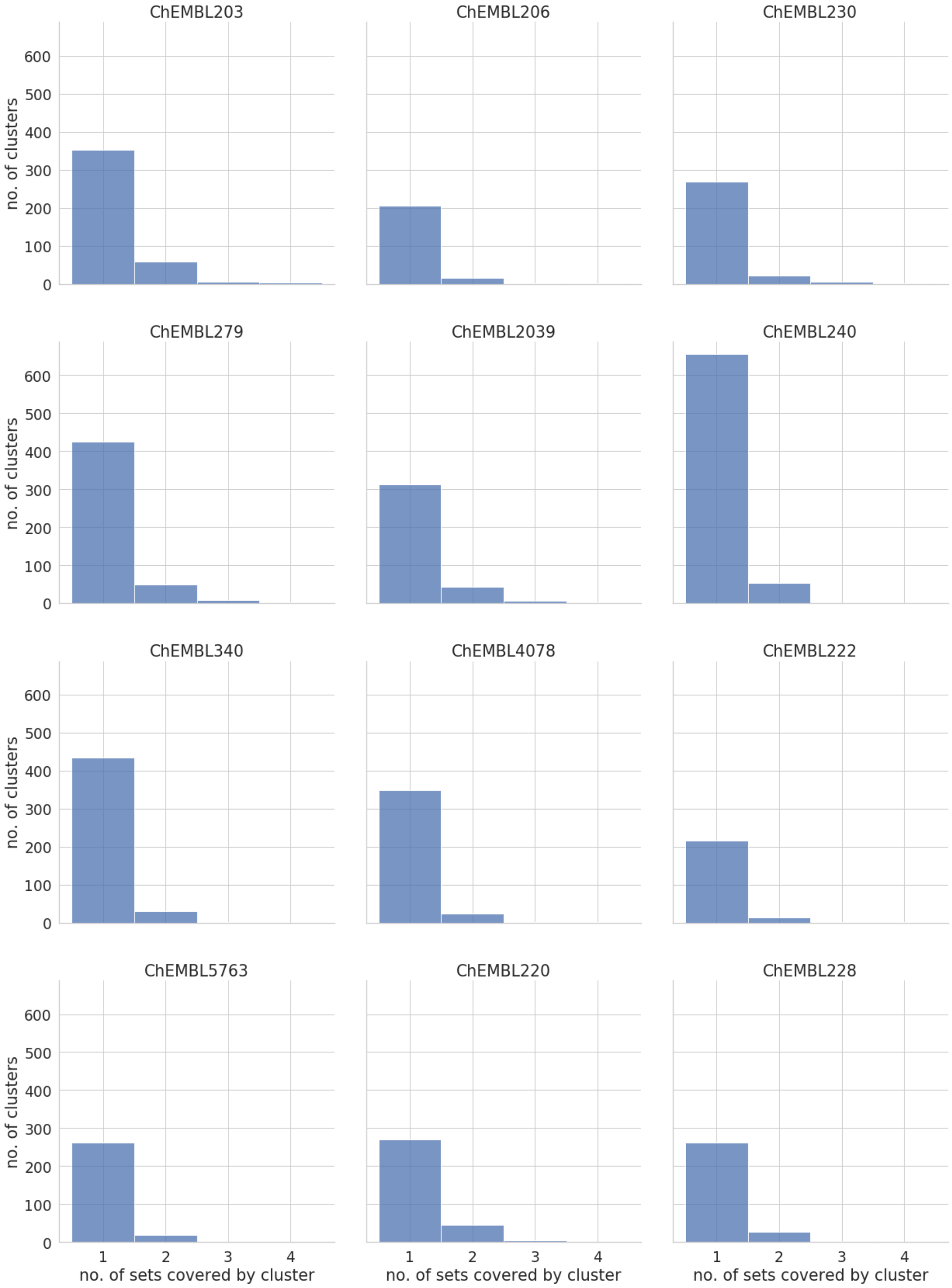
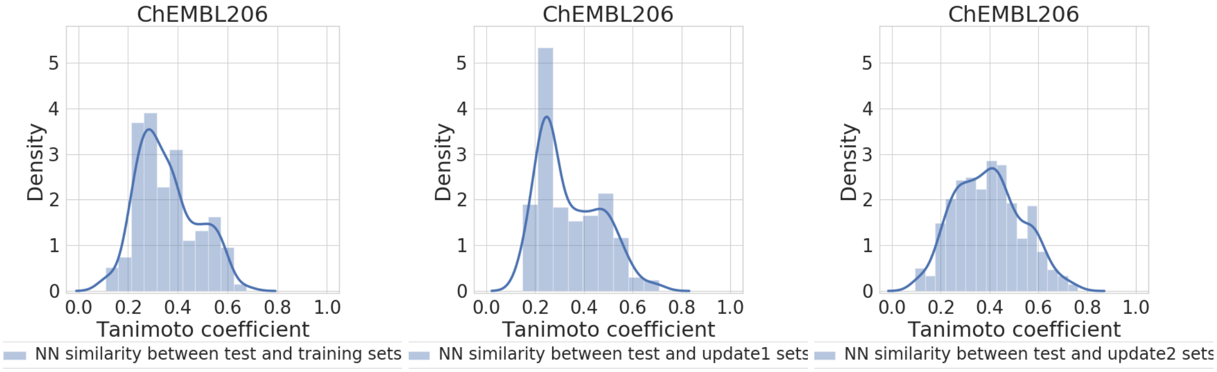


Figure S5. Spreading of clusters amongst the data subsets (i.e. splits) for the ChEMBL datasets. Most of the clusters (with at least two compounds) do not spread over more than one subset (i.e. training, update1, update2 or holdout set).



(a) training set **(b)** Update1 set **(c)** Update2 set

Figure S6. Distribution of Tanimoto coefficients between each holdout compound to its nearest neighbour in the corresponding subset (training, update1 and update2) for ChEMBL206 endpoint .

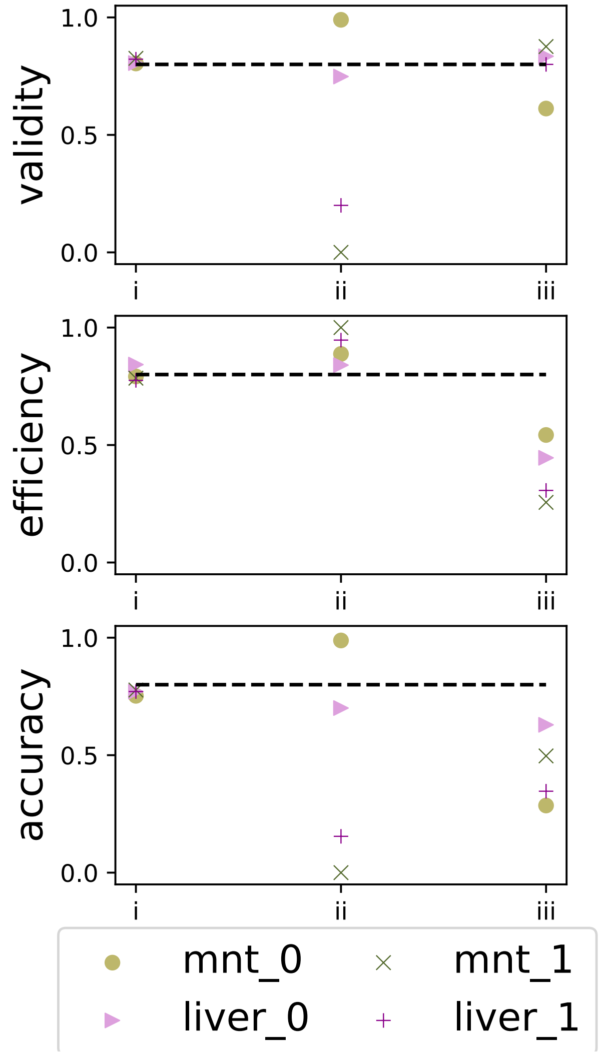


Figure S7. Time split evaluation (validity, efficiency, accuracy) of experiments i) CV, predictions using ii) original calibration set, iii) update calibration set for the liver toxicity and MNT inhouse datasets.

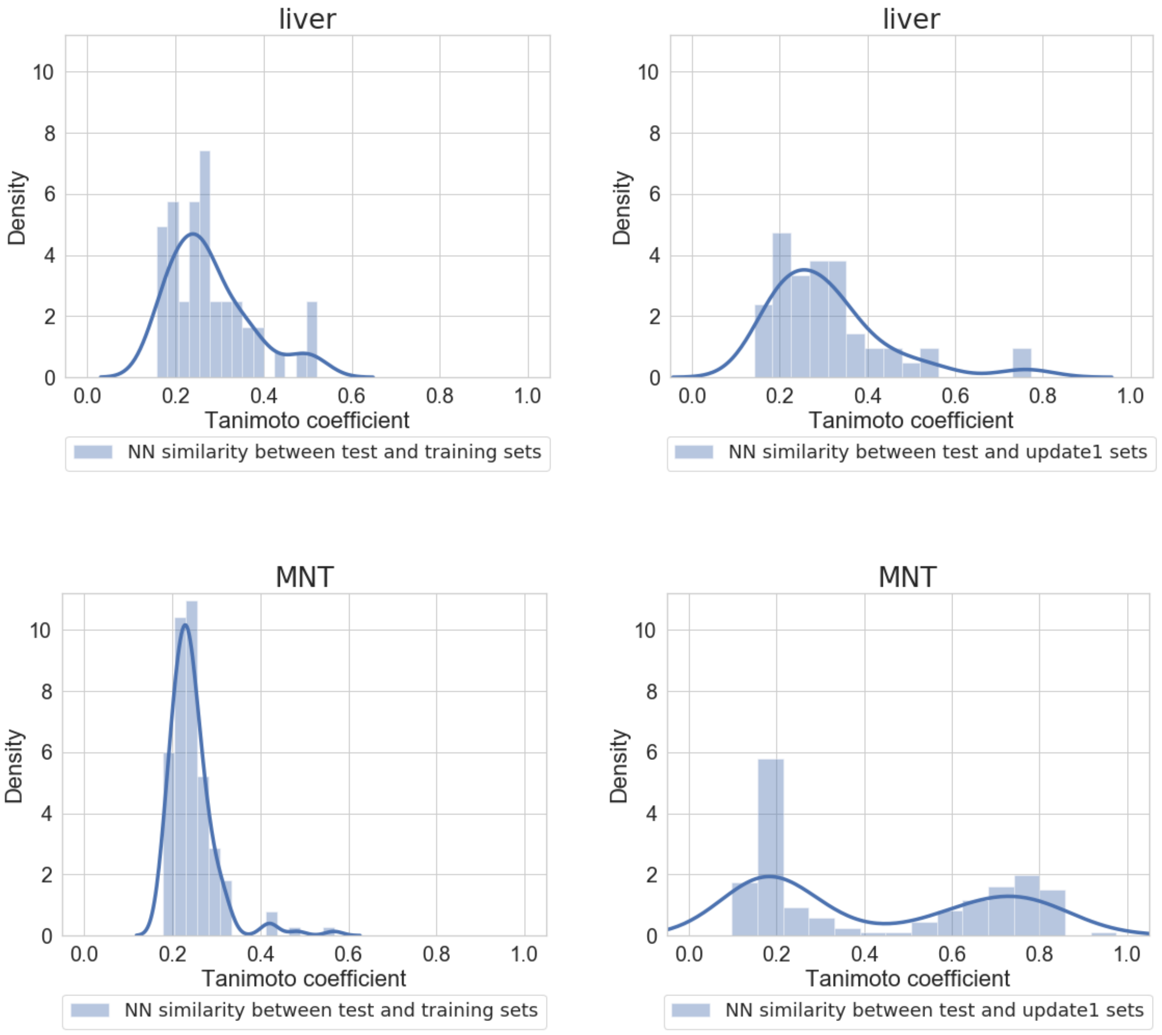


Figure S8. Distribution of Tanimoto coefficients between each holdout compound to its nearest neighbour in the training (left) and update (right) set for the liver (top) and MNT (bottom) endpoints.

Supporting Information for [P4]

This appendix contains the supporting information for the publication:

Garcia de Lomana, M.; Svensson, F.; Volkamer, A.; Mathea, M. and Kirchmair, J. Consideration of predicted small-molecule metabolites in computational toxicology, *Digital Discov.*, **2022**, 1, 158-172.

SUPPORTING INFORMATION

Consideration of predicted small-molecule metabolites in computational toxicology

Marina Garcia de Lomana[1,2]*, Fredrik Svensson*[3]*, Andrea Volkamer*[4]*, Miriam Mathea*[1*] *and Johannes Kirchmair*[2*]

[1] BASF SE, 67063 Ludwigshafen am Rhein, Germany

[2] Department of Pharmaceutical Sciences, Faculty of Life Sciences, University of Vienna, 1090 Vienna, Austria

[3] Alzheimer's Research UK UCL Drug Discovery Institute, University College London, London WC1E 6BT, U.K

[4] In Silico Toxicology and Structural Bioinformatics, Institute of Physiology, Charité Universitätsmedizin Berlin, 10117 Berlin, Germany

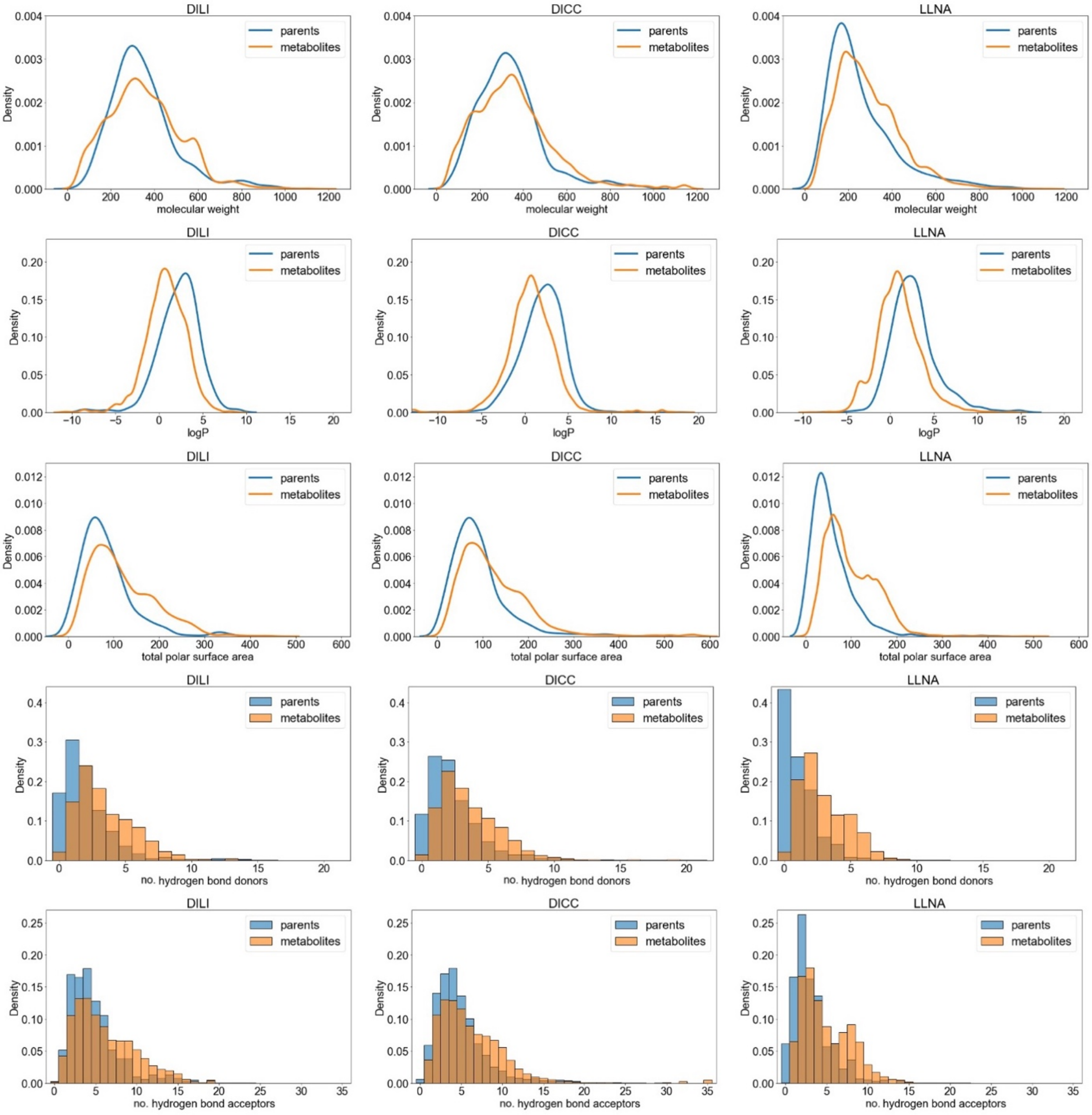


Figure S1. Comparison of the physicochemical properties of the parent compounds (blue) and predicted metabolites (orange) represented in the DILI, DICC and LLNA data sets.

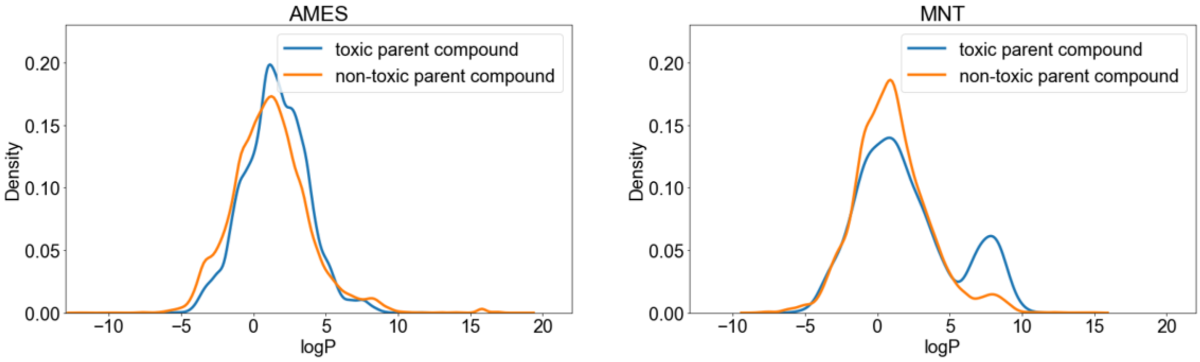


Figure S2. Distribution of the logP values for the metabolites from toxic (blue) and non-toxic (orange) parent compounds in the AMES and MNT data sets.

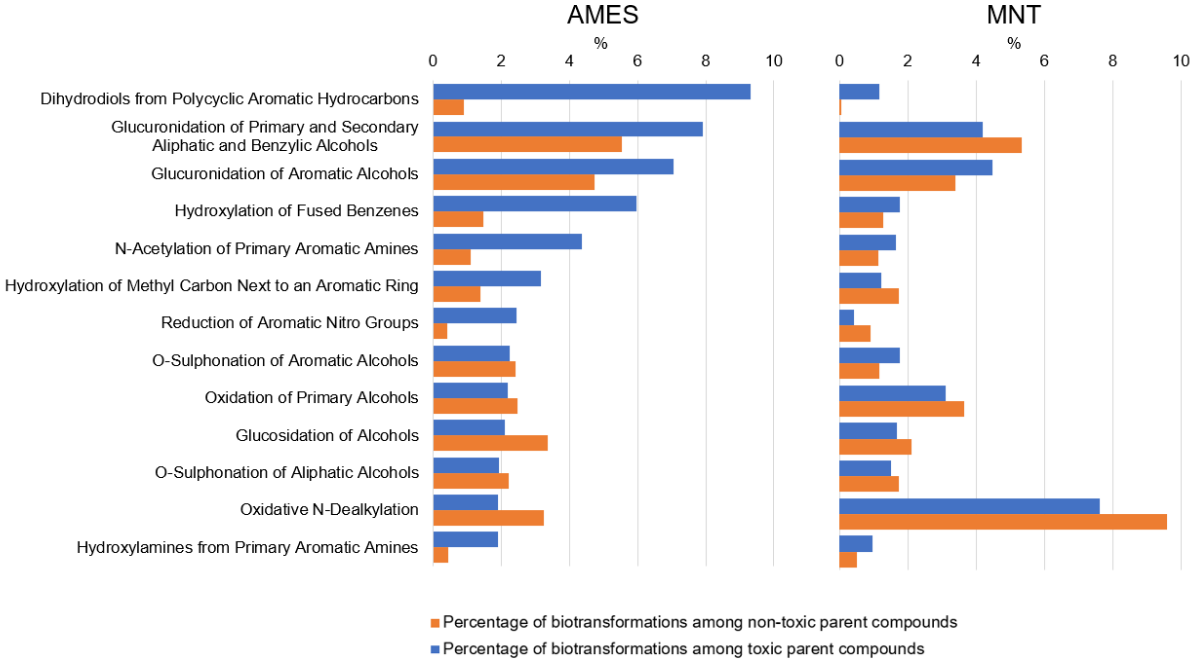


Figure S3. Percentage of occurrence of a subset of biotransformations between toxic (blue) and non-toxic (orange) compounds. The selected subset are the 15 biotransformations most often observed for toxic compounds in AMES. Although some biotransformations appear more often in one of the activity classes, these ratios are different between endpoints.

Table S1. Data Sources and Download Links for the Original Data Sets.

Endpoint	Data sources	Download link[1]	Access date	Query (json format)	MD5 file checksum	Checksum input file
AMES	CCRIS (https://www.nlm.nih.gov/databases/download/ccris.html)	https://ftp.nlm.nih.gov/projects/ccrislease/	19.02.21	-	B411532AD80846CF1D5FDD9B08B79F93	XML file in ccris.xml.20110828.zip
	GENE-TOX (https://www.nlm.nih.gov/databases/download/genetox.html)	https://pubchem.ncbi.nlm.nih.gov/bioassay/1259408	19.02.21	-	1A9706F5C08DF8A24F6DB2BF3DADDAB0	.csv file
	NTP (https://cebs.niehs.nih.gov/datasets/search/ames)	ftp://anonftp.niehs.nih.gov/ntp-cebs/datatype/NTP_Data_Collections/Ames_Conclusions_DataCollection_2020-02-19.xlsx	19.02.21	-	F442B6687587F17B42C9BA1D4D82D4BB	.xlsx file
MNT	10.1016/j.yrtph.2020.104620			-	6174327EB2B69D4326B36E5D610ACDE7	Supplementary .xlsx file

eChemPortal (active)	https://www.echemportal.org/echemportal/property-search	06.08.20	{"blocks":[{"level":0,"type":"property","id":"dis3i1p7tjkdijou2p","label":"Genetic toxicity in vivo","endpointKind":"GeneticToxicityVivo"}],"endpoints":{"dis3i1p7tjkdijou2p":{"ENDPOINT_STUDY_RECORD.GeneticToxicityVivo.AdministrativeData.StudyResultType":{"1342":"","phrase":["1895"]},"ENDPOINT_STUDY_RECORD.GeneticToxicityVivo.AdministrativeData.Reliability":{"1342":"","phrase":["16","18","24","1342"]},"ENDPOINT_STUDY_RECORD.GeneticToxicityVivo.MaterialsAndMethods.Guideline.Qualifier":{"phrase":["1680","1880"]},"ENDPOINT_STUDY_RECORD.GeneticToxicityVivo.MaterialsAndMethods.Guideline.Guideline":{"1342":"","phrase":["1290"]},"ENDPOINT_STUDY_RECORD.GeneticToxicityVivo.ResultsAndDiscussion.TestRs.Genotoxicity":{"1342":"","phrase":["2276"]},"ENDPOINT_STUDY_RECORD.GeneticToxicityVivo.ResultsAndDiscussion.TestRs.Toxicity":{"phrase":["2170","2197","2207"]},"endpointKind":"GeneticToxicityVivo"}}}	6D1771634AE4FBDFDC9C517A0F2594FC	.csv file resulting from eChemPortal query
eChemPortal (inactive)	https://www.echemportal.org/echemportal/property-search	06.08.20	{"blocks":[{"level":0,"type":"property","id":"dis3i1p7tjkdijou2p","label":"Genetic toxicity in vivo","endpointKind":"GeneticToxicityVivo"}],"endpoints":{"dis3i1p7tjkdijou2p":{"ENDPOINT_STUDY_RECORD.GeneticToxicityVivo.AdministrativeData.StudyResultType":{"1342":"","phrase":["1895"]},"ENDPOINT_STUDY_RECORD.GeneticToxicityVivo.AdministrativeData.Reliability":{"1342":"","phrase":["16","18","24","1342"]},"ENDPOINT_STUDY_RECORD.GeneticToxicityVivo.MaterialsAndMethods.Guideline.Qualifier":{"phrase":["1680","1880"]},"ENDPOINT_STUDY_RECORD.GeneticToxicityVivo.MaterialsAndMethods.Guideline.Guideline":{"1342":"","phrase":["1290"]},"ENDPOINT_STUDY_RECORD.GeneticToxicityVivo.ResultsAndDiscussion.TestRs.Genotoxicity":{"1342":null,"phrase":["2148"]},"ENDPOINT_STUDY_RECORD.GeneticToxicityVivo.ResultsAndDiscussion.TestRs.Toxicity":{"phrase":["2170","2197","2207"]},"endpointKind":"GeneticToxicityVivo"}}}	10BC1CB5B9C4F0FEDDAEF41600E13937	.csv file resulting from eChemPortal query

DILI	10.1016/j.drudis.2016.02.015	-	-	4EA88A5523A6717B9118AF7C4DAA9442	Supplementary .xlsx file
DICC	10.1021/acs.jcim.7b00641	-	-	80B6A4048F31A9EC74DD84522B1F861D	Supplementary .xlsx file (Table S1)
LLNA	10.1021/acs.chemrestox.0c00253	-	-	E3F1E3FBCC21AF714295BCFEAA5B4DE2	Supplementary .xlsx file

[1] Missing download links correspond to data sets available as supplementary material of the publication indicated as data source.

Table S2. Performance of Models Including Molecular Descriptors for Metabolites.

Endpoint	Model[1]	F1 score	MCC	Precision	Recall
AMES	baseline performance	0.82 (+/- 0.01)	0.65 (+/- 0.03)	0.83 (+/- 0.01)	0.82 (+/- 0.01)
	Morgan	0.82 (+/- 0.02)	0.65 (+/- 0.03)	0.83 (+/- 0.02)	0.82 (+/- 0.02)
	RDKit	0.81 (+/- 0.01)	0.62 (+/- 0.02)	0.82 (+/- 0.01)	0.81 (+/- 0.01)
	Morgan+RDKit	0.81 (+/- 0.01)	0.62 (+/- 0.03)	0.82 (+/- 0.01)	0.80 (+/- 0.01)
MNT	baseline performance	0.64 (+/- 0.03)	0.29 (+/- 0.05)	0.67 (+/- 0.02)	0.62 (+/- 0.03)
	Morgan	0.61 (+/- 0.03)	0.24 (+/- 0.06)	0.66 (+/- 0.05)	0.59 (+/- 0.02)
	RDKit	0.60 (+/- 0.03)	0.25 (+/- 0.06)	0.69 (+/- 0.05)	0.59 (+/- 0.02)
	Morgan+RDKit	0.59 (+/- 0.04)	0.23 (+/- 0.08)	0.66 (+/- 0.06)	0.58 (+/- 0.03)
DILI	baseline performance	0.68 (+/- 0.04)	0.37 (+/- 0.08)	0.69 (+/- 0.04)	0.68 (+/- 0.03)
	Morgan	0.69 (+/- 0.04)	0.38 (+/- 0.09)	0.70 (+/- 0.05)	0.68 (+/- 0.04)
	RDKit	0.67 (+/- 0.04)	0.36 (+/- 0.07)	0.69 (+/- 0.03)	0.66 (+/- 0.04)
	Morgan+RDKit	0.66 (+/- 0.03)	0.34 (+/- 0.06)	0.68 (+/- 0.04)	0.66 (+/- 0.03)
DICC	baseline performance	0.69 (+/- 0.02)	0.39 (+/- 0.04)	0.71 (+/- 0.02)	0.69 (+/- 0.03)
	Morgan	0.66 (+/- 0.02)	0.33 (+/- 0.05)	0.68 (+/- 0.02)	0.65 (+/- 0.02)
	RDKit	0.64 (+/- 0.03)	0.30 (+/- 0.06)	0.67 (+/- 0.03)	0.64 (+/- 0.03)
	Morgan+RDKit	0.64 (+/- 0.03)	0.30 (+/- 0.06)	0.67 (+/- 0.03)	0.63 (+/- 0.03)
LLNA	baseline performance	0.73 (+/- 0.02)	0.47 (+/- 0.04)	0.74 (+/- 0.02)	0.69 (+/- 0.03)
	Morgan	0.73 (+/- 0.02)	0.46 (+/- 0.04)	0.73 (+/- 0.02)	0.73 (+/- 0.03)
	RDKit	0.69 (+/- 0.02)	0.39 (+/- 0.05)	0.70 (+/- 0.02)	0.69 (+/- 0.02)
	Morgan+RDKit	0.70 (+/- 0.02)	0.40 (+/- 0.04)	0.71 (+/- 0.02)	0.70 (+/- 0.02)

[1] The baseline model does not include any metabolite descriptor. “Morgan” refers to the count-based Morgan fingerprint and “RDKit” to the RDKit physicochemical property descriptors of the metabolites.

Table S3. P-values From the Mann-Whitney U Test Between the Baseline Performance and the Models Including Molecular Descriptors for Metabolites.

Endpoint	Model[1]	F1 score	MCC	Precision	Recall
AMES	Morgan	1,00	1,00	0,92	1,00
	RDKit	0,14	0,14	0,35	0,07
	Morgan+RDKit	0,12	0,14	0,21	0,06
MNT	Morgan	0,21	0,21	0,83	0,14
	RDKit	0,09	0,3	0,4	0,07
	Morgan+RDKit	0,09	0,53	0,92	0,09
DILI	Morgan	1,00	1,00	0,83	0,92
	RDKit	0,53	0,68	1,00	0,60
	Morgan+RDKit	0,30	0,40	0,68	0,35
DICC	Morgan	0,09	0,09	0,09	0,09
	RDKit	0,04	0,04	0,14	0,04
	Morgan+RDKit	0,02	0,04	0,06	0,02
LLNA	Morgan	1,00	1,00	1,00	1,00
	RDKit	0,04	0,04	0,04	0,04
	Morgan+RDKit	0,06	0,06	0,04	0,06

[1] The baseline model does not include any metabolite descriptor. “Morgan” refers to the count-based Morgan fingerprint and “RDKit” to the RDKit physicochemical property descriptors of the metabolites.

Table S4. Performance of Models Including the Biotransformation Fingerprint.

Endpoint	F1 score	MCC	Precision	Recall
AMES	0.82 (+/- 0.02)	0.65 (+/- 0.04)	0.83 (+/- 0.02)	0.82 (+/- 0.02)
MNT	0.63 (+/- 0.03)	0.28 (+/- 0.05)	0.67 (+/- 0.02)	0.62 (+/- 0.03)
DILI	0.69 (+/- 0.04)	0.38 (+/- 0.08)	0.70 (+/- 0.04)	0.68 (+/- 0.04)
DICC	0.69 (+/- 0.02)	0.39 (+/- 0.04)	0.71 (+/- 0.02)	0.69 (+/- 0.02)
LLNA	0.74 (+/- 0.03)	0.48 (+/- 0.05)	0.74 (+/- 0.02)	0.74 (+/- 0.03)

Table S5. P-values From the Mann-Whitney U Test Between the Baseline Performance and the Models Including the Biotransformation Fingerprint.

Endpoint	F1 score	MCC	Precision	Recall
AMES	0,92	0,92	0,83	0,83
MNT	0,68	0,92	0,83	0,53
DILI	1,00	1,00	0,83	0,53
DICC	0,92	1,00	1,00	0,92
LLNA	0,60	0,60	0,53	0,75

Table S6. Average Performance within 5-fold Cross-Validation for the Different Combinations of Predicted Probabilities with the Baseline-Approach.

Endpoint	Combination [1]	F1 score	MCC	Precision	Recall
AMES	baseline performance	0.82 (+/- 0.01)	0.65 (+/- 0.03)	0.83 (+/- 0.01)	0.82 (+/- 0.01)
	Strategy 1	0.73 (+/- 0.00)	0.51 (+/- 0.01)	0.80 (+/- 0.01)	0.71 (+/- 0.00)
	Strategy 2	0.73 (+/- 0.01)	0.50 (+/- 0.03)	0.79 (+/- 0.02)	0.72 (+/- 0.01)
	Strategy 3	0.79 (+/- 0.01)	0.59 (+/- 0.03)	0.79 (+/- 0.01)	0.80 (+/- 0.01)
	Strategy 4	0.82 (+/- 0.01)	0.64 (+/- 0.03)	0.82 (+/- 0.01)	0.82 (+/- 0.01)
MNT	baseline performance	0.64 (+/- 0.03)	0.29 (+/- 0.05)	0.67 (+/- 0.02)	0.62 (+/- 0.03)
	Strategy 1	0.59 (+/- 0.03)	0.25 (+/- 0.05)	0.71 (+/- 0.04)	0.57 (+/- 0.02)
	Strategy 2	0.59 (+/- 0.03)	0.24 (+/- 0.06)	0.70 (+/- 0.04)	0.57 (+/- 0.02)
	Strategy 3	0.61 (+/- 0.03)	0.27 (+/- 0.04)	0.61 (+/- 0.02)	0.67 (+/- 0.02)
	Strategy 4	0.66 (+/- 0.02)	0.33 (+/- 0.04)	0.67 (+/- 0.03)	0.66 (+/- 0.02)
DILI	baseline performance	0.68 (+/- 0.04)	0.37 (+/- 0.08)	0.69 (+/- 0.04)	0.68 (+/- 0.04)
	Strategy 1	0.67 (+/- 0.03)	0.34 (+/- 0.06)	0.67 (+/- 0.02)	0.67 (+/- 0.04)
	Strategy 2	0.66 (+/- 0.02)	0.33 (+/- 0.05)	0.67 (+/- 0.02)	0.66 (+/- 0.03)
	Strategy 3	0.55 (+/- 0.02)	0.22 (+/- 0.06)	0.68 (+/- 0.05)	0.57 (+/- 0.02)
	Strategy 4	0.65 (+/- 0.02)	0.35 (+/- 0.04)	0.71 (+/- 0.02)	0.64 (+/- 0.02)
DICC	baseline performance	0.69 (+/- 0.02)	0.39 (+/- 0.04)	0.71 (+/- 0.02)	0.69 (+/- 0.03)
	Strategy 1	0.58 (+/- 0.04)	0.25 (+/- 0.06)	0.69 (+/- 0.04)	0.59 (+/- 0.03)
	Strategy 2	0.59 (+/- 0.04)	0.25 (+/- 0.07)	0.67 (+/- 0.04)	0.59 (+/- 0.03)
	Strategy 3	0.66 (+/- 0.00)	0.33 (+/- 0.01)	0.66 (+/- 0.01)	0.68 (+/- 0.01)
	Strategy 4	0.70 (+/- 0.01)	0.39 (+/- 0.02)	0.70 (+/- 0.01)	0.70 (+/- 0.02)
LLNA	baseline performance	0.73 (+/- 0.02)	0.47 (+/- 0.04)	0.74 (+/- 0.02)	0.73 (+/- 0.02)

	Strategy 1	0.52 (+/- 0.02)	0.22 (+/- 0.04)	0.68 (+/- 0.05)	0.57 (+/- 0.01)
	Strategy 2	0.51 (+/- 0.02)	0.18 (+/- 0.02)	0.64 (+/- 0.02)	0.56 (+/- 0.01)
	Strategy 3	0.70 (+/- 0.01)	0.41 (+/- 0.02)	0.70 (+/- 0.01)	0.71 (+/- 0.01)
	Strategy 4	0.71 (+/- 0.01)	0.43 (+/- 0.02)	0.71 (+/- 0.01)	0.71 (+/- 0.01)

[1] The baseline performance corresponds to models considering only parent compounds. Strategies 1, 2 and 3 correspond to taking the mean, median and maximum predicted probability among the parent compound and its metabolites, respectively. Strategy 4 corresponds to the mean between the predicted probability for the parent compound and the highest probability predicted for any of its metabolites.

Table S7. P-values From the Mann-Whitney U Test Between the Baseline Performance and the Baseline-Approach Strategies.

Endpoint	**Combination**[1]	**P-value of**[2]			
		F1 score	**MCC**	**Precision**	**Recall**
AMES	Strategy 1	0,01	0,01	0,01	0,01
	Strategy 2	0,01	0,01	0,01	0,01
	Strategy 3	0,02	0,04	0,01	0,17
	Strategy 4	0,83	0,83	0,40	0,92
MNT	Strategy 1	0,06	0,30	0,09	0,03
	Strategy 2	0,06	0,30	0,21	0,05
	Strategy 3	0,30	0,53	0,01	0,04
	Strategy 4	0,14	0,21	0,75	0,09
DILI	Strategy 1	0,68	0,83	0,53	0,83
	Strategy 2	0,53	0,68	0,40	0,53
	Strategy 3	0,01	0,02	0,40	0,01
	Strategy 4	0,21	0,53	0,83	0,21
DICC	Strategy 1	0,01	0,02	0,53	0,01
	Strategy 2	0,01	0,02	0,14	0,01
	Strategy 3	0,03	0,07	0,01	0,29
	Strategy 4	0,83	0,83	0,17	0,53
LLNA	Strategy 1	0,01	0,01	0,06	0,01
	Strategy 2	0,01	0,01	0,01	0,01

	Strategy 3	0,02	0,05	0,01	0,06
	Strategy 4	0,06	0,06	0,07	0,12

[1] Strategies 1, 2 and 3 correspond to taking the mean, median and maximum predicted probability among the parent compound and its metabolites, respectively. Strategy 4 corresponds to the mean between the predicted probability for the parent compound and the highest probability predicted for any of its metabolites.

[2] P-values lower than 0.05 that indicate an improvement compared to the baseline performance are highlighted in bold.

Table S8. P-values From the Mann-Whitney U Test Between the Baseline Performance and the Hybrid-Approach Strategies.

Endpoint	**Combination**[1]	**P-value of**[2]			
		F1 score	**MCC**	**Precision**	**Recall**
AMES	Strategy 1	0,04	0,04	0,53	0,01
	Strategy 2	0,04	0,06	0,17	0,02
	Strategy 3	0,04	0,06	0,02	0,14
	Strategy 4	0,75	0,68	0,92	0,46
MNT	Strategy 1	0,30	0,40	0,01	0,21
	Strategy 2	0,21	0,68	0,01	0,14
	Strategy 3	0,53	0,53	0,06	0,02
	Strategy 4	0,21	0,30	0,30	0,21
DILI	Strategy 1	0,40	0,40	0,30	0,46
	Strategy 2	0,30	0,21	0,21	0,29
	Strategy 3	0,04	0,21	0,53	0,04
	Strategy 4	0,40	1,00	0,40	0,30
DICC	Strategy 1	0,35	1,00	0,01	0,17
	Strategy 2	0,21	0,92	0,05	0,21
	Strategy 3	0,14	0,21	0,03	0,21
	Strategy 4	0,14	0,21	0,21	0,06
LLNA	Strategy 1	0,06	0,14	0,46	0,06
	Strategy 2	0,53	0,53	1,00	0,53
	Strategy 3	0,02	0,09	0,07	0,17
	Strategy 4	0,83	0,83	1,00	0,68

[1] Strategies 1, 2 and 3 correspond to taking the mean, median and maximum predicted probability among the parent compound and its metabolites, respectively. Strategy 4 corresponds to the mean between the predicted probability for the parent compound and the highest probability predicteds for any of its metabolites.

[2] P-values lower than 0.05 that indicate an improvement compared to the baseline performance are highlighted in bold.

Table S9. Best Five Combinations of Model Type, Probability Combination and Metabolite Filters for Each Endpoint.

Endpoint	Scenario	Combination[1]	Minimum Meteor score	Minimum logP	Phase II detoxification	F1 score	MCC	Precision	Recall
AMES	baseline performance					0.82 (+/- 0.01)	0.65 (+/- 0.03)	0.83 (+/- 0.01)	0.82 (+/- 0.01)
	hybrid-approach	Strategy 4	200	-	No	0.83 (+/- 0.02)	0.66 (+/- 0.03)	0.83 (+/- 0.02)	0.83 (+/- 0.02)
	hybrid-approach	Strategy 4	200	0	No	0.83 (+/- 0.01)	0.66 (+/- 0.03)	0.83 (+/- 0.01)	0.83 (+/- 0.02)
	hybrid-approach	Strategy 4	300	0	No	0.83 (+/- 0.02)	0.66 (+/- 0.03)	0.83 (+/- 0.02)	0.83 (+/- 0.02)
	hybrid-approach	Strategy 4	300	-	No	0.83 (+/- 0.02)	0.66 (+/- 0.04)	0.83 (+/- 0.02)	0.83 (+/- 0.02)
	hybrid-approach	Strategy 1	200	3	No	0.83 (+/- 0.01)	0.65 (+/- 0.02)	0.83 (+/- 0.01)	0.82 (+/- 0.01)
MNT	baseline performance					0.64 (+/- 0.03)	0.29 (+/- 0.05)	0.67 (+/- 0.02)	0.62 (+/- 0.03)
	baseline-approach	Strategy 4	-	-	No	0.66 (+/- 0.02)	0.33 (+/- 0.04)	0.67 (+/- 0.03)	0.66 (+/- 0.02)
	baseline-approach	Strategy 4	100	-	No	0.66 (+/- 0.02)	0.33 (+/- 0.04)	0.67 (+/- 0.03)	0.66 (+/- 0.02)
	hybrid-approach	Strategy 4	-	-	Yes	0.66 (+/- 0.04)	0.34 (+/- 0.07)	0.70 (+/- 0.04)	0.64 (+/- 0.03)
	hybrid-approach	Strategy 4	-	-	No	0.66 (+/- 0.03)	0.33 (+/- 0.06)	0.69 (+/- 0.04)	0.65 (+/- 0.03)

	baseline-approach	Strategy 4	200	-	No	0.66 (+/-0.02)	0.32 (+/-0.04)	0.67 (+/-0.03)	0.65 (+/-0.02)
DILI	baseline performance					0.68 (+/-0.04)	0.37 (+/-0.08)	0.69 (+/-0.04)	0.68 (+/-0.04)
	hybrid-approach	Strategy 4	100	3	Yes	0.69 (+/-0.03)	0.39 (+/-0.06)	0.71 (+/-0.03)	0.69 (+/-0.03)
	hybrid-approach	Strategy 4	200	3	Yes	0.69 (+/-0.03)	0.39 (+/-0.06)	0.71 (+/-0.03)	0.69 (+/-0.03)
	hybrid-approach	Strategy 4	-	3	Yes	0.69 (+/-0.03)	0.39 (+/-0.06)	0.71 (+/-0.03)	0.69 (+/-0.03)
	baseline-approach	Strategy 2	300	0	Yes	0.69 (+/-0.04)	0.39 (+/-0.08)	0.70 (+/-0.04)	0.69 (+/-0.04)
	hybrid-approach	Strategy 4	0	3	No	0.69 (+/-0.03)	0.39 (+/-0.06)	0.71 (+/-0.03)	0.68 (+/-0.03)
DICC	baseline performance					0.69 (+/-0.02)	0.39 (+/-0.04)	0.71 (+/-0.02)	0.69 (+/-0.03)
	hybrid-approach	Strategy 4	100	-	Yes	0.73 (+/-0.02)	0.45 (+/-0.04)	0.73 (+/-0.02)	0.72 (+/-0.02)
	hybrid-approach	Strategy 4	-	-	Yes	0.73 (+/-0.02)	0.45 (+/-0.04)	0.73 (+/-0.02)	0.72 (+/-0.02)
	hybrid-approach	Strategy 4	200	-	Yes	0.73 (+/-0.02)	0.45 (+/-0.05)	0.73 (+/-0.02)	0.72 (+/-0.03)
	hybrid-approach	Strategy 4	300	-	No	0.72 (+/-0.02)	0.44 (+/-0.05)	0.73 (+/-0.02)	0.72 (+/-0.03)
	hybrid-approach	Strategy 4	200	-	No	0.72 (+/-0.02)	0.44 (+/-0.04)	0.72 (+/-0.02)	0.72 (+/-0.02)

LLNA	baseline performance					0.73 (+/- 0.02)	0.47 (+/- 0.04)	0.74 (+/- 0.02)	0.73 (+/- 0.02)
	hybrid-approach	Strategy 1	300	0	Yes	0.74 (+/- 0.02)	0.49 (+/- 0.04)	0.75 (+/- 0.02)	0.74 (+/- 0.02)
	hybrid-approach	Strategy 4	-	-	No	0.74 (+/- 0.02)	0.48 (+/- 0.05)	0.74 (+/- 0.02)	0.74 (+/- 0.03)
	hybrid-approach	Strategy 2	300	0	Yes	0.74 (+/- 0.02)	0.48 (+/- 0.04)	0.74 (+/- 0.02)	0.73 (+/- 0.02)
	hybrid-approach	Strategy 4	100	0	No	0.73 (+/- 0.02)	0.47 (+/- 0.04)	0.74 (+/- 0.02)	0.74 (+/- 0.02)
	hybrid-approach	Strategy 4	-	0	No	0.73 (+/- 0.02)	0.47 (+/- 0.04)	0.74 (+/- 0.02)	0.74 (+/- 0.02)

[1] Strategies 1, 2 and 3 correspond to taking the mean, median and maximum predicted probability among the parent compound and its metabolites, respectively. Strategy 4 corresponds to the mean between the predicted probability for the parent compound and the highest probability predicted for any of its metabolites.

Table S10. P-values From the Mann-Whitney U Test for the Best Five Model Combinations for Each Endpoint Compared to the Baseline Performance.

Endpoint	Scenario	Combination[1]	Minimum Meteor score	Minimum logP	Phase II detoxification	F1 score	MCC	Precision	Recall
AMES	baseline performance					0,53	0,53	0,68	0,30
	hybrid-approach	Strategy 4	200	-	No	0,68	0,68	0,84	0,40
	hybrid-approach	Strategy 4	200	0	No	0,84	0,84	1,00	0,60
	hybrid-approach	Strategy 4	300	0	No	0,68	0,68	1,00	0,60
	hybrid-approach	Strategy 4	300	-	No	0,68	0,68	0,68	0,68
	hybrid-approach	Strategy 1	200	3	No	0,14	0,21	0,75	0,10
MNT	baseline performance					0,12	0,12	0,92	0,08
	baseline-approach	Strategy 4	-	-	No	0,40	0,30	0,21	0,40
	baseline-approach	Strategy 4	100	-	No	0,21	0,30	0,30	0,21
	hybrid-approach	Strategy 4	-	-	Yes	0,30	0,30	0,75	0,14

	hybrid-approach	Strategy 4	-	-	No	0,75	0,53	0,40	0,84
	baseline-approach	Strategy 4	200	-	No	0,75	0,53	0,40	0,84
DILI	baseline performance					0,75	0,53	0,40	0,84
	hybrid-approach	Strategy 4	100	3	Yes	1,00	1,00	1,00	0,83
	hybrid-approach	Strategy 4	200	3	Yes	0,84	0,68	0,60	1,00
	hybrid-approach	Strategy 4	-	3	Yes	0,06	0,09	0,10	0,06
	baseline-approach	Strategy 2	300	0	Yes	0,06	0,08	0,10	0,05
	hybrid-approach	Strategy 4	0	3	No	0,10	0,10	0,10	0,10
DICC	baseline performance					0,14	0,14	0,10	0,14
	hybrid-approach	Strategy 4	100	-	Yes	0,10	0,14	0,14	0,06
	hybrid-approach	Strategy 4	-	-	Yes	0,53	0,30	0,30	0,60
	hybrid-approach	Strategy 4	200	-	Yes	0,84	0,84	1,00	0,68
	hybrid-approach	Strategy 4	300	-	No	1,00	0,68	0,4	0,92

	hybrid-approach	Strategy 4	200	-	No	0,92	1,00	0,92	0,84
LLNA	baseline performance					0,92	1,00	0,92	0,84
	hybrid-approach	Strategy 1	300	0	Yes	0,53	0,53	0,68	0,30
	hybrid-approach	Strategy 4	-	-	No	0,68	0,68	0,84	0,40
	hybrid-approach	Strategy 2	300	0	Yes	0,84	0,84	1,00	0,60
	hybrid-approach	Strategy 4	100	0	No	0,68	0,68	1,00	0,60
	hybrid-approach	Strategy 4	-	0	No	0,68	0,68	0,68	0,68

[1] Strategies 1, 2 and 3 correspond to taking the mean, median and maximum predicted probability among the parent compound and its metabolites, respectively. Strategy 4 corresponds to the mean between the predicted probability for the parent compound and the highest probability predicted for any of its metabolites.

Table S11. Mean F1 Score Obtained with Different Machine Learning Methods, Input Descriptors and Oversampling Setups on the Hybrid-Approach.1

		Random forest			**Gradient boosted trees**		**K-Nearest neighbors**	
Endpoint	**Combination**	**CDDD descriptors with oversampling**	**Without oversampling**	**With oversampling**	**Without oversampling**	**With oversampling**	**Without oversampling**	**With oversampling**
AMES	baseline performance[2]	0.81 (+/-0.01)	**0.83 (+/-0.01)**	0.82 (+/-0.01)	0.81 (+/-0.01)	0.82 (+/-0.01)	0.66 (+/-0.02)	0.66 (+/-0.02)
	Strategy 1	0.78 (+/-0.01)	0.80 (+/-0.01)	0.80 (+/-0.01)	0.78 (+/-0.02)	0.79 (+/-0.02)	0.69 (+/-0.02)	0.69 (+/-0.02)
	Strategy 2	0.78 (+/-0.01)	0.80 (+/-0.01)	0.80 (+/-0.01)	0.78 (+/-0.02)	0.79 (+/-0.02)	0.69 (+/-0.03)	0.69 (+/-0.03)
	Strategy 3	0.80 (+/-0.02)	0.80 (+/-0.02)	0.79 (+/-0.02)	0.79 (+/-0.02)	0.78 (+/-0.02)	0.47 (+/-0.03)	0.47 (+/-0.03)
	Strategy 4	0.81 (+/-0.01)	**0.83 (+/-0.02)**	**0.83 (+/-0.02)**	0.82 (+/-0.02)	0.82 (+/-0.02)	0.56 (+/-0.01)	0.56 (+/-0.01)
MNT	baseline performance[2]	0.58 (+/-0.04)	0.60 (+/-0.03)	0.64 (+/-0.03)	0.58 (+/-0.01)	0.62 (+/-0.02)	0.57 (+/-0.05)	0.39 (+/-0.02)
	Strategy 1	0.56 (+/-0.03)	0.59 (+/-0.05)	0.61 (+/-0.02)	0.54 (+/-0.04)	0.61 (+/-0.03)	0.57 (+/-0.01)	0.48 (+/-0.02)
	Strategy 2	0.56 (+/-0.04)	0.59 (+/-0.04)	0.61 (+/-0.02)	0.54 (+/-0.03)	0.61 (+/-0.03)	0.58 (+/-0.02)	0.48 (+/-0.02)
	Strategy 3	0.62 (+/-0.00)	0.64 (+/-0.02)	**0.65 (+/-0.03)**	0.63 (+/-0.02)	0.62 (+/-0.03)	0.52 (+/-0.02)	0.25 (+/-0.01)
	Strategy 4	0.62 (+/-0.01)	0.63 (+/-0.03)	**0.65 (+/-0.04)**	0.61 (+/-0.02)	**0.65 (+/-0.03)**	0.59 (+/-0.02)	0.31 (+/-0.02)
DILI	baseline performance[2]	0.66 (+/-0.02)	0.68 (+/-0.03)	0.68 (+/-0.04)	**0.69 (+/-0.05)**	**0.69 (+/-0.04)**	0.56 (+/-0.02)	0.38 (+/-0.02)
	Strategy 1	0.65 (+/-0.03)	0.67 (+/-0.04)	0.67 (+/-0.04)	0.66 (+/-0.05)	0.68 (+/-0.03)	0.65 (+/-0.03)	0.51 (+/-0.02)
	Strategy 2	0.65 (+/-0.03)	0.65 (+/-0.04)	0.67 (+/-0.03)	0.66 (+/-0.04)	0.67 (+/-0.03)	0.65 (+/-0.05)	0.51 (+/-0.02)
	Strategy 3	0.56 (+/-0.04)	0.57 (+/-0.04)	0.60 (+/-0.04)	0.56 (+/-0.04)	0.57 (+/-0.03)	0.56 (+/-0.04)	0.60 (+/-0.03)
	Strategy 4	0.62 (+/-0.03)	0.64 (+/-0.03)	0.66 (+/-0.04)	0.64 (+/-0.04)	**0.69 (+/-0.05)**	0.63 (+/-0.02)	0.58 (+/-0.02)

DICC	baseline performance[2]	0.65 (+/-0.02)	0.70 (+/-0.02)	0.70 (+/-0.02)	0.68 (+/-0.01)	0.70 (+/-0.02)	0.62 (+/-0.02)	0.44 (+/-0.01)
	Strategy 1	0.61 (+/-0.04)	0.62 (+/-0.02)	0.68 (+/-0.02)	0.61 (+/-0.02)	0.66 (+/-0.02)	0.64 (+/-0.01)	0.53 (+/-0.01)
	Strategy 2	0.62 (+/-0.03)	0.64 (+/-0.01)	0.68 (+/-0.02)	0.62 (+/-0.01)	0.66 (+/-0.01)	0.64 (+/-0.02)	0.53 (+/-0.01)
	Strategy 3	0.68 (+/-0.01)	0.69 (+/-0.02)	0.68 (+/-0.02)	0.68 (+/-0.03)	0.66 (+/-0.02)	0.56 (+/-0.03)	0.31 (+/-0.00)
	Strategy 4	0.68 (+/-0.01)	0.71 (+/-0.02)	**0.72 (+/-0.02)**	0.69 (+/-0.01)	0.71 (+/-0.01)	0.62 (+/-0.02)	0.36 (+/-0.01)
LLNA	baseline performance[2]	**0.74 (+/-0.03)**	**0.74 (+/-0.03)**	**0.74 (+/-0.03)**	0.72 (+/-0.02)	0.73 (+/-0.03)	0.57 (+/-0.03)	0.52 (+/-0.03)
	Strategy 1	0.70 (+/-0.01)	0.71 (+/-0.03)	0.71 (+/-0.02)	0.70 (+/-0.02)	0.71 (+/-0.03)	0.62 (+/-0.02)	0.62 (+/-0.03)
	Strategy 2	0.73 (+/-0.01)	0.71 (+/-0.03)	0.71 (+/-0.03)	0.70 (+/-0.03)	0.70 (+/-0.03)	0.61 (+/-0.01)	0.62 (+/-0.02)
	Strategy 3	0.67 (+/-0.01)	0.68 (+/-0.01)	0.68 (+/-0.0)	0.66 (+/-0.03)	0.66 (+/-0.03)	0.52 (+/-0.03)	0.42 (+/-0.03)
	Strategy 4	**0.74 (+/-0.02)**	0.73 (+/-0.01)	0.73 (+/-0.02)	0.73 (+/-0.01)	**0.74 (+/-0.02)**	0.61 (+/-0.03)	0.53 (+/-0.03)

[1] Unless stated otherwise, Morgan fingerprint and RDKit physicochemical descriptors were used as input for model training. In all models (excepting the baseline models) hyperparameters were optimized within a 5-fold cross-validated grid search.

[2] RF model trained only on parent compounds (including oversampling and without hyperparameter optimization).