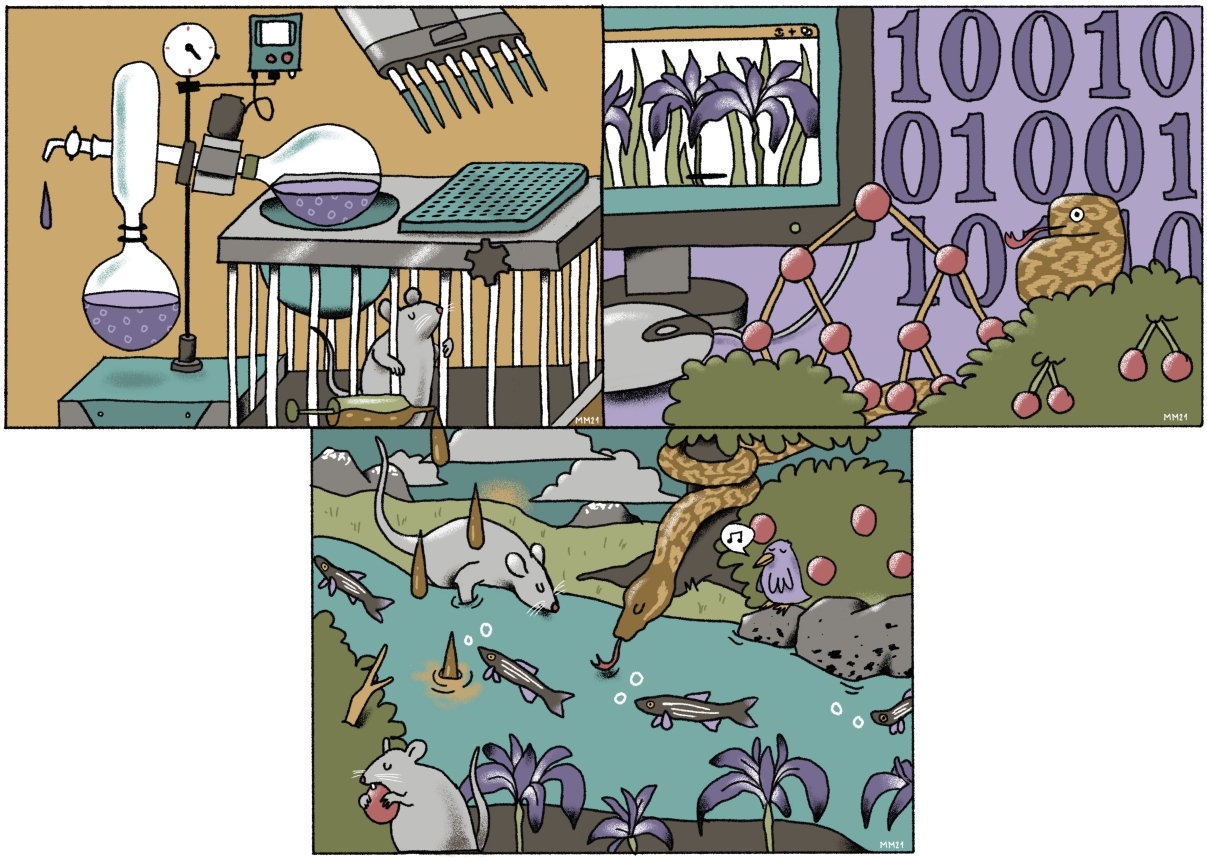
Strategies to enhance the applicability of *in silico* toxicity prediction methods



Inaugural-Dissertation
to obtain the academic degree
Doctor of Philosophy in Natural Science (Ph.D. in Natural Science)

submitted to the Department of Biology, Chemistry, Pharmacy
of Freie Universität Berlin
by

ANDREA LILIAN MORGER

2022

The presented thesis was prepared from August 2017 until March 2022 under the (main) supervision of Prof. Dr Andrea Volkamer at the Institute of Physiology of the Charité Universitätsmedizin Berlin, further supervision by Prof. Dr Gerhard Wolber and mentoring by PD Dr Robert Landsiedel.

1st reviewer: Prof. Dr Andrea Volkamer
2nd reviewer: Prof. Dr Gerhard Wolber

Date of defence: 17.06.2022

Acknowledgements

Firstly, I would like to thank Prof. Dr Andrea Volkamer for making this thesis possible and for putting together such a wonderful group. I would like to thank all colleagues, who contributed to the friendly atmosphere, in particular to Dominique and Talia for our friendship and to Jaime for his enthusiasm in teaching Pythonic Programming. I would also like to thank Prof. Dr Gerhard Wolber and PD Dr Robert Landsiedel for their support as second supervisor and mentor.

Secondly, I would like to thank our external collaboration partners without whom this thesis would not have been possible. I am especially grateful for the experience in computational toxicology and conformal prediction that I could acquire with them. Many thanks to my collaboration partners at BASF SE (in particular Dr Miriam Mathea, Marina Garcia de Lomana, Dr Janosch Achenbach, PD Dr Robert Landsiedel, and Roland Buesen, Ph.D.), to the collaboration partners at BfR (especially Dr Saskia Klutzny and Dr Sebastian Dunst), to Dr Fredrik Svensson, Dr Ulf Norinder, Prof. Dr Ola Spjuth, and Prof. Dr Johannes Kirchmair.

During my PhD, I had the opportunity for a research stay in the Pharmaceutical Bioninformatics group of Prof. Dr Ola Spjuth and I would like to thank Ola and Ulf for making this valuable experience possible. The research stay was financed by FUBright Mobility Allowance, and I would like to further thank the HaVo-Stiftung, BASF SE, and BB3R for funding for this thesis. A thank you also to Dr Vivian Kral for her efforts in leading the BB3R graduate school.

I would like to thank Andrea, Gerhard, Robert, Fredrik, Talia, and Bart for proof-reading (parts of) this thesis. A big thank you also to my brother Moreno (`https://morenomorger.ch/`) for the beautiful illustrations. He nicely illustrated the interplay of the different topics that are coming together in this thesis. Many thanks to my family, and especially to Bart, for all of their support.

Herewith I certify that I have prepared and written my thesis independently and that I have not used any sources and aids other than those indicated by me.

Contents

Abstract

Given the ubiquity of synthetic chemicals in our daily life, it is crucial to assess the hazardous effects of new chemical substances on humans, animals, and the environment. Toxicity assessment has traditionally been based on *in vitro* and *in vivo* studies, but ethical and economic arguments call for reduction and replacement of animal testing. Therefore, computational toxicity prediction has gained momentum to support toxicity studies and to ultimately reduce animal testing. *In silico* approaches are comparably fast and inexpensive, and many of them can be applied prior to synthesis and *in vitro* testing of new chemicals.

Computational methods such as machine learning (ML), similarity search, and structural alerts are already in use during the development of new chemicals. They are often restrained by limitations in data availability for training and by the need for applicability domain determination for predictions on new data. In this thesis, novel *in silico* strategies for guiding *in vivo* and *in vitro* toxicity testing were developed. Means to maximise the gain from limited available data were explored, as well as strategies to improve the applicability of *in silico* toxicity prediction approaches. A special focus was laid on studying the potential of the conformal prediction (CP) framework, which is built on top of an ML model, to allow for confidence estimation. The CP framework utilises an extra calibration set to compare the predicted probabilities of new query compounds to those previously seen. The calibrated probabilities are returned in the form of so-called p-values.

To support toxicity testing of new chemicals, the Python-based KnowTox pipeline was developed. Following a holistic approach, a compound of interest can *in silico* be examined from three perspectives: KnowTox searches for known toxic substructures, returns how similar compounds were tested *in vitro*, and queries pre-trained CP models. The value of complementing the outputs of different *in silico* approaches was demonstrated in a retrospective case study on two triazole molecules from industry. Focusing on the estrogen receptor (ER), an important target for endocrine disruption, we further explored whether *in silico* predictions can help to pre-select compounds for *in vitro* experiments. Starting from nine newly discovered ER active compounds (using the recently-developed E-Morph Screen ER assay), it was prospectively shown that similarity search and CP models can help to increase the hit rate of *in vitro* screens, enabling fast and efficient identification of novel endocrine disruptors.

In the above described studies, CP was used as it outputs valid confidence estimates and guarantees pre-defined error rates (on the exchangeability assumption). Moreover, allowing class-wise calibration, data imbalances are usually well-handled. The potential of CP was, in this thesis, further investigated for the generation of bioactivity descriptors, and to mitigate data drift effects. The ChemBioSim project addressed the challenge of predicting *in vivo* toxicological effects by informing CP models with bioactivity de-

scriptors originating from *in vitro* data. Compared to chemical descriptors, bioactivity descriptors may provide more mechanistic information and help to better capture complex *in vivo* outcomes. To avoid *in vitro* testing of every query molecule, p-values returned by CP models trained on *in vitro* datasets served as bioactivity descriptors. For the investigated MNT and cardiotoxicity endpoints, *in vivo* toxicity prediction could be improved by using bioactivity (instead of chemical) descriptors.

The CP framework is designed to yield valid predictions, provided that training and test set are exchangeable. This assumption is not always fulfilled; data drifts may occur e.g. when the chemical space or assay conditions change. To mitigate effects of data drifts, we have developed a recalibration strategy, suggesting to exchange the calibration set with data closer to the test data. The strategy, developed based on the Tox21 data, was further analysed for temporal data drifts using ChEMBL data and for differences between public and proprietary data. In most cases, recalibration led to restored validity, a prerequisite for model applicability.

Besides applications of computational toxicity prediction methods and CP, this thesis further discusses general aspects of data and applicability domain in the context of *in silico* toxicology. While regulatory agencies still require animal studies, with the computational strategies discussed in this work, we aim to foster the reliability of predictions and the applicability of models, to ultimately reduce animal testing.

Zusammenfassung

Synthetische Chemikalien sind in unserem täglichen Leben allgegenwärtig, was die Untersuchung neuer Chemikalien auf toxische Effekte unerlässlich macht. Toxikologische Untersuchungen werden traditionellerweise anhand von *in vitro* und *in vivo* Studien durchgeführt. Jedoch fordern ethische und wirtschaftliche Argumente die Reduktion und letztendlich den Ersatz von Tierversuchen. Daher hat die computergestützte Toxizitätsvorhersage an Bedeutung gewonnen, um Toxizitätsstudien zu unterstützen und letztlich Tierversuche zu reduzieren. *In silico* Methoden sind vergleichsweise schnell und günstig, und viele von ihnen können vor der Synthese und *in vitro*-Prüfung neuer Chemikalien angewendet werden.

Computergestützte Methoden wie Maschinelles Lernverfahren (ML), Ähnlichkeitssuche und Substruktursuche werden im Entwicklungsprozess neuer Chemikalien bereits angewendet. Sie stossen jedoch oft an ihre Grenzen. Ein Grund dafür ist die limitierte Datenverfügbarkeit, ein anderer die Gewährleistung der Anwendbarkeit der Modelle. Im Zuge dieser Arbeit wurden neuartige *in silico* Strategien zur Steuerung von *in vivo* und *in vitro* Versuchen entwickelt. Es wurden Strategien zur Maximierung des Nutzens aus den begrenzt verfügbaren Daten sowie Strategien zur Verbesserung der Anwendbarkeit von *in silico*-Toxizitätsvorhersageansätzen untersucht. Ein besonderer Schwerpunkt der Arbeit lag auf der Untersuchung des Potenzials des Conformal Prediction (CP) Frameworks, das auf einem ML-Modell aufbaut, um eine Vetrauensabschätzung zu ermöglichen. CP verwendet ein zusätzliches Kalibrierungsset, mithilfe dessen die von ML Modellen vorhergesagten Wahrscheinlichkeiten für neue Moleküle kalibriert werden. Die Kalibrierung erfolgt anhand von Vorhersagen für bereits bekannte Moleküle und die kalibrierten Wahrscheinlichkeiten werden als sogenannte p-Werte zurückzugeben.

Um das Planen von toxikologischen Studien und die Risikobeurteilung von Chemikalien zu unterstützen, wurde die Python-basierte KnowTox Pipeline entwickelt. KnowTox verfolgt einen ganzheitlichen Ansatz, bei dem eine neue Substanz aus drei Perspektiven *in silico* beurteilt wird: KnowTox sucht nach bekannten unerwünschten Substrukturen, ermittelt wie ähnliche Substanzen *in vitro* getestet wurden, und es werden Vorhersagen mit vortrainierten CP Modellen gemacht. In einer retrospektiven Fallstudie mit zwei ehemaligen Entwicklungskandidaten aus der Industrie konnte der Nutzen des Kombinierens verschiedener *in silico* Methoden aufgezeigt werden. Unsere nächste Studie konzentrierte sich auf den Östrogenrezeptor (ER), einen wichtigen Angriffspunkt für hormonaktive Substanzen. Es wurde untersucht, ob *in silico* Vorhersagen auch bei der Vorselektionierung von Testsubstanzen für *in vitro* Versuche nützlich sein können. Anhand von neun Substanzen, die mithilfe des kürzlich entwickelten E-Morph Screen ER Assays als ER-aktiv eingestuft worden sind, konnte prospektiv gezeigt werden, wie Ähnlichkeitssuche und CP-Modelle die Trefferquote von *in vitro* Screeningverfahren erhöhen können, was eine schnellere und effizientere Identifizierung neuartiger Endokriner Disruptoren ermöglicht.

In den oben beschriebenen Studien wurde die CP-Methode gewählt, weil sie valide Vertrauensabschätzungen macht und vordefinierte Fehlerraten garantiert. Zusätzlich kann CP durch klassenweise Kalibrierung gut mit den für toxikologische Datensätze üblichen Ungleichgewichten zwischen der Anzahl aktiver und inaktiver Moleküle umgehen. Desweiteren wurde in dieser Arbeit das Potenzial der CP-Methode für die Generierung von Bioaktivitäts-Deskriptoren und zur Abschwächung von Datendrifteffekten untersucht. Das ChemBioSim Projekt befasste sich mit der Herausforderung, toxikologische *in vivo* Effekte vorherzusagen, indem CP-Modelle mit Bioaktivitäts-Deskriptoren aus *in vitro*-Daten angereichert wurden. Im Vergleich zu chemischen Deskriptoren, könnten Bioaktivitäts-Deskriptoren mehr mechanistische Informationen enthalten und helfen, komplexe *in vivo*-Endpunkte besser zu erfassen. Um zu vermeiden, dass jedes vorhergesagte Molekül auch synthetisiert und *in vitro* getestet werden muss, wurden CP Modelle auf *in vitro* Datensätzen trainiert und die ausgegebenen p-Werte als Bioaktivitäts-Deskriptoren verwendet. Für die untersuchten MNT- und Kardiotoxizitäts-Endpunkte konnte die Vorhersage der *in vivo* Toxizität mithilfe der Bioaktivitätsdeskriptoren, im Vergleich zu chemischen Deskriptoren, verbessert werden.

Das CP Framework wurde so konzipiert, dass die Modelle gültige Vorhersagen liefern, vorausgesetzt, dass Trainings- und Testdatensatz austauschbar sind. Diese Annahme ist jedoch nicht immer erfüllt. Es kann zum Beispiel zu Datendrifts kommen, wenn sich der chemische Raum oder die Assay-Bedingungen ändern. Um die Auswirkungen solcher Datendrifte abzuschwächen, haben wir eine sogenannte 'Rekalibrierungs-Strategie' entwickelt, bei der das Kalibrierungsset durch neue Daten ersetzt wird, die näher am Testdatensatz liegen. Die Strategie wurde anhand der Tox21 Datensätze entwickelt und anschliessend weiter für die Anwendung auf temporale Datendrifts sowie auf Unterschiede zwischen öffentlichen und proprietären Daten untersucht. In den meisten Fällen führte die Rekalibrierung zur Wiederherstellung der Validität, eine Voraussetzung für die Anwendbarkeit des Modells.

Neben den Anwendungen von computergestützten Methoden und CP zur Vorhersage der Toxizität, werden in dieser Arbeit auch allgemeine Aspekte der Daten und der Anwendbarkeit im Kontext der *in silico* Toxikologie diskutiert. Während die Aufsichtsbehörden nach wie vor Tierversuche verlangen, zielen die in dieser Arbeit erörterten Strategien darauf ab, die Zuverlässigkeit der Vorhersagen und die Anwendbarkeit der Modelle zu verbessern, um letztendlich Tierversuche zu reduzieren.

Acronyms

ML machine learning

CP conformal prediction

QSAR quantitative structure-activity relationship

AD applicability domain

AD$_{Hanser}$ applicability domain

RD$_{Hanser}$ reliability domain

DD$_{Hanser}$ decidability domain

RF random forest

SVM support vector machine

KNN k-nearest neighbours

ER estrogen receptor

AR androgen receptor

REACH Registration, Evaluation, Authorisation and Restriction of Chemicals

OECD Organisation for Economic Co-operation and Development

HTS high-throughput screening

nc score nonconformity score

EDC endocrine-disrupting chemical

MNT micro nucleus test

Introduction

1.1 Toxicity assessment of chemical substances

Synthetic chemicals are ubiquitous in our daily life. They appear e.g. in the form of drugs, pesticides, and cosmetic ingredients, but are also present in cleaning agents, or used to soften plastic. Consequently, humans and the environment are inevitably exposed to a variety of chemical substances, which may not only have beneficial, but also toxic effects [1]. The risk that a substance is actually causing harm, depends on its hazard (i.e. toxicity) and the exposure to it [2].

Toxicity is still one of the most important reasons for drug attrition during the development of new medicines [3–5]. According to a study conducted by four large pharmaceutical companies and performed on data for 812 oral development candidates, nominated between 2000 and 2010, 40% of the failed drug candidates dropped out due to non-clinical toxicology, i.e. toxic effects detected before progressing to the clinics. During Phase I and II clinical studies, drug attrition due to clinical safety accounted for 25% of the failed compounds [4]. Already during the development of new drugs and other types of chemicals, it is crucial that potential toxic effects on humans or the environment are assessed. Adverse effects of chemical substances comprise various types of toxicity. For example, organs, such as the heart, liver, or lung might be adversely affected. Exposure to chemicals could also cause reproductive and developmental toxicity, or mutagenicity [4–7].

An important class of toxic chemical substances are endocrine-disrupting chemicals (EDCs). They disturb the hormone system and as a consequence exhibit toxic effects such as reproductive disfunction, hormone-dependent cancers, or disruption of brain or immune system development [8–11]. EDCs can act via various mechanisms of action; some need further investigation [8, 11]. They may mimic or partly mimic natural hormones such as estrogens, androgens, or thyroid hormones. They may also directly interact with hormone receptors or interfere with other players of the endocrine system [8, 9, 12, 13]. Well-known EDCs are, among others, bisphenol-A used in plastic and can manufacturing [10], pesticides such as dichlorodiphenyltrichloroethane or permethrin, and synthetic steroids which are present in contraceptives such as estradiol and estrone [14]. Awareness and concerns about severe effects of EDCs on humans and wildlife have grown since the 1990's. This has led to a variety of (inter-)national actions, such as revised guidelines, government-established inventories of produced or imported

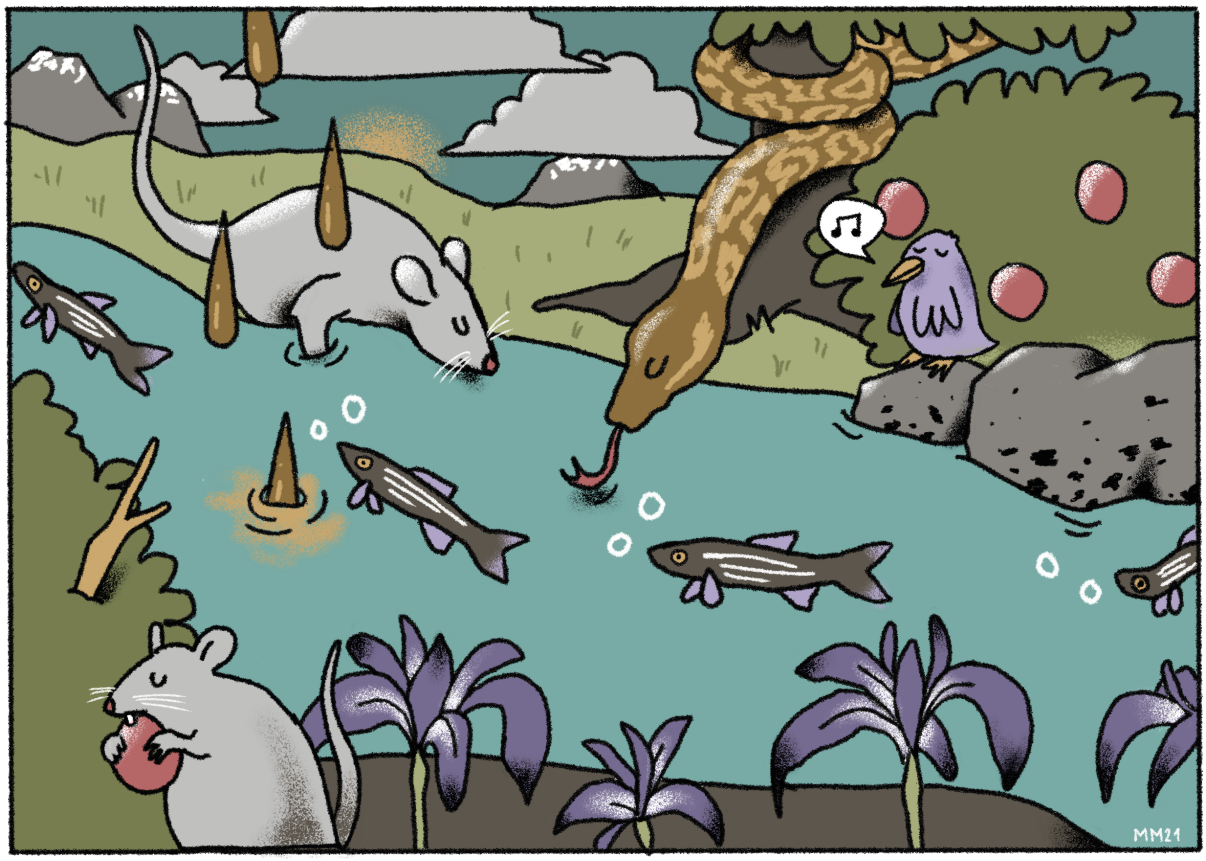


Figure 1.1: The prevalence of chemicals in the environment can lead to unwanted effects on plants and animals. Illustration by Moreno Morger.

chemical substances, or the launch of EDC screening programs, among others by the EU, USA, WHO, and Organisation for Economic Co-operation and Development (OECD) [8, 15–19].

While the awareness of the potential 'risk' of chemicals rose, it also became clear that safety data was missing for a large portion of marketed chemicals [20–23]. Hence, in 2006, the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH) legislation was enforced. The REACH legislation requires hazard assessment for all chemicals available on the European Union's market. The hazard information requirements depend on the amount and how they were produced, imported, and used [21, 24].

1.2 Replace, reduce, and refine animal testing

Traditionally, toxicity studies in animals, such as mice, rats, or monkeys, and subsequent extrapolation to humans have been used for the assessment of toxicological effects of chemical substances [25]. In 2019, almost 475,000 animals were employed in Germany for production, quality control, and toxicological safety assessment [26]. Such *in vivo* studies are expensive, time-consuming, and come with the uncertainty of extrapolation from test animals to humans.

Moreover, the ethical aspects of animal experimentation need to be considered. In 1959, the 3R's principle, i.e. replacement, reduction, and refinement of animal testing, was introduced by William Russell and Rex Burch and published in their book on 'The Principles of Humane Experimental Technique' [27]. They called for substantial effort to research for alternatives to animal testing (replacement), more efficient, reliable, and reproducible experiments (reduction), and to minimise suffering of laboratory animals (refinement). Also the REACH legislation from 2006 promotes non-animal based methods. Based on animal welfare considerations, animal tests should only be used as a last resort [21, 24, 28].

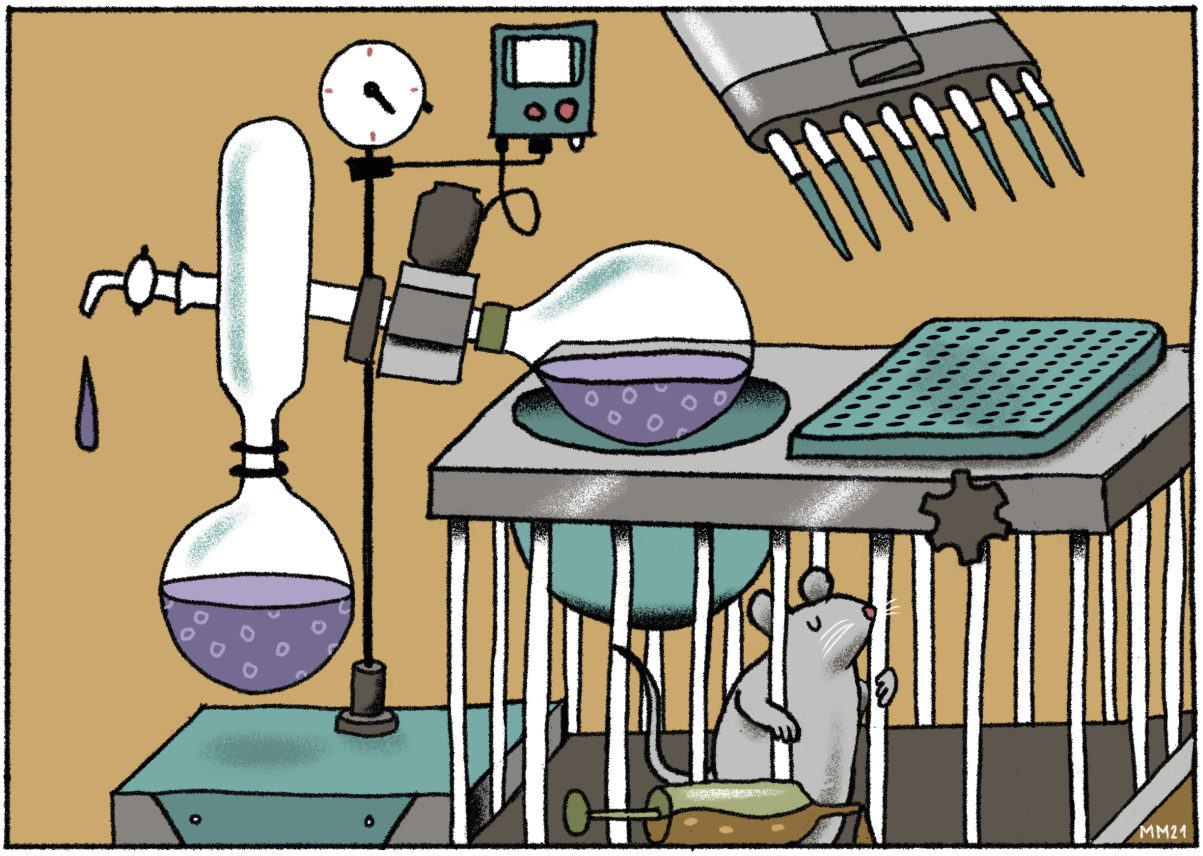


Figure 1.2: Toxicity of new chemicals is traditionally assessed in animal studies. *In vitro* testing is an essential alternative for already synthesised compounds. Illustration by Moreno Morger.

In the following, different alternatives to animal toxicity testing will be introduced, i.e. read-across, *in vitro* testing, and *in silico* predictions. These methods are also referred to as new approach methodologies (NAMs) [29].

Read-across approach

The read-across approach [24, 30] to fill data gaps is based on the assumption that similar compounds may exhibit similar toxic effects. If data is missing on the properties of a query molecule, more may be learnt from reading-across information from sufficiently

similar molecules. Read-across is one of the non-animal-based methods, which is in certain cases accepted by the regulatory agencies [6, 24, 31, 32]. The aim of read-across predictions is to be (more or less) equivalent to and thus able to replace a standard animal study [32]. Therefore, sound and well-justified read-across is required [24]. This entails proper planning, documentation for reproducibility and transparency, and expert judgment. A big challenge is also the assessment of uncertainty, of both the underlying experimental data, and the similarity justification [31, 32]. The state-of-the-art and insights on the read-across studies submitted to REACH were summarised by Ball et al. [6].

In the second report on 'The Use of Alternatives to Testing on Animals for the REACH Regulation', it was reported that up to 75% of the dossiers contained a read-across or category approach for at least one endpoint [24, 33]. Several successful read-across studies have been published since then [31, 32]. For example, van Ravenzwaay et al. provided a retrospective read-across case study with three phenoxy-herbicides. They demonstrated how a 90-day study in rats could have been waived with the use of metabolomics data from blood samples in a 28-day study [34].

Read-across is a so-called expert method and typically performed manually. However, computational support can be useful to mine for appropriate similar molecules in a large database, to rank them, and to collect experimental information about them [19, 35, 36].

In vitro approaches

In vitro methods provide valuable alternatives to animal testing. Nowadays, many biochemistry and cell-based experiments can be performed in the form of high-throughput screening (HTS) assays [25]. While such assays are much faster, cheaper, and more ethical than *in vivo* experiments, the interpretation and extrapolation from *in vitro* to *in vivo* experiments remains challenging [23, 37]. *In vitro* and *in vivo* experiments are often performed with different doses and measurements taken at different time points. Moreover, certain systemic effects can be difficult to detect *in vitro* when biochemical and cell-based *in vitro* systems cannot capture whole-animal systems. Another difficulty is the analysis of the formation and the toxic effects of metabolites which are formed upon bioactivation in the liver or other tissues.

Both *in vitro* and *in vivo* experiments come with experimental errors, i.e. variability between assay conditions, the person conducting the experiment, or animal physiology [38–40].

In silico approaches

In silico methods, which build on the information gained from *in vitro* and *in vivo* studies, provide another level of alternatives to animal testing. Computational methods for toxicity prediction [3, 31, 41, 42] can be used in the early stages of the development of

new chemicals, to guide toxicity testing and hopefully lead to the exclusion of harmful compounds in the early stages. A variety of methods that can help to assess the hazard of compounds, such as similarity search, structural alerts, and machine learning (ML), will be introduced in the next section. Computational predictions are relatively fast and cheap, compared to *in vivo* and *in vitro* experiments. *In silico* tools can even be deployed to not yet synthesised compounds and to compounds not available in sufficient amount for testing. Additionally, raw data, code, models, and environments can be stored, shared, and reused by independent researchers to reproduce the results [40].

The use of *in silico* toxicology in regulatory frameworks is still limited [31]. The first guideline to allow *in vitro* experiments to be replaced by *in silico* predictions was the ICH M7 guideline for mutagenicity assessment. Given certain prerequisites, if the predictions from two *in silico* approaches — typically structural alerts and ML — come to the same (negative) result, an *in vitro* AMES assay for manufacturing impurities assessment may be waived [31, 43].

Also the combination of *in vitro* and *in silico* methods is promising, although there are only a few publications exploring such strategies [44–47]. The full potential is not yet exploited.

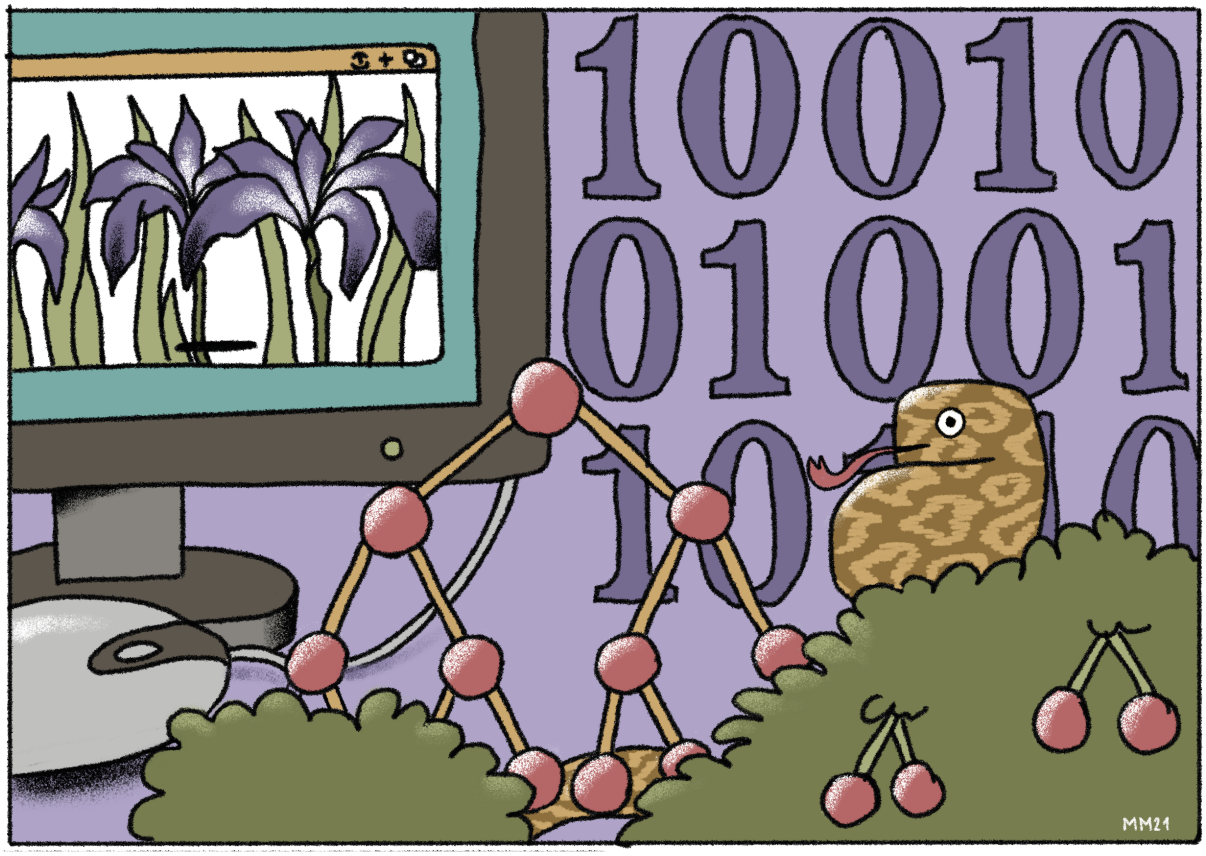


Figure 1.3: Computational methods, such as machine learning models trained on *in vitro* or *in vivo* data, are promising alternatives to animal testing, particularly for not yet synthesised chemicals. Illustration by Moreno Morger.

1.3 *In silico* methods for toxic endpoint prediction

Computational methods used for risk assessment and toxicity prediction of novel compounds are diverse. The main computational methods will be introduced in the following.

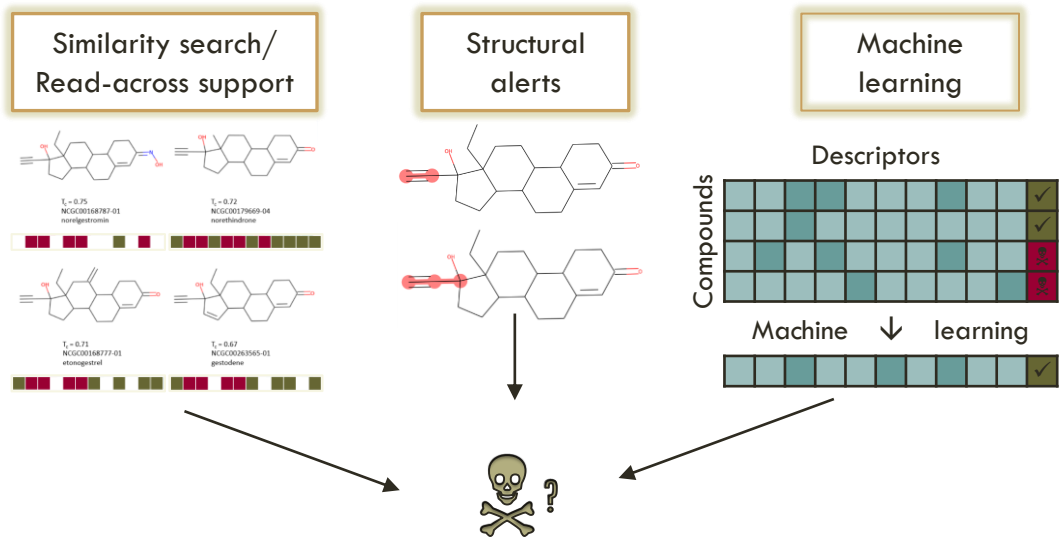


Figure 1.4: An overview of computational toxicity prediction methods.

Similarity-based approaches

Many computational prediction methods follow the similarity hypothesis. The similarity property principle, introduced by Johnson and Maggiora in 1990, states that similar compounds should have similar properties, especially similar biological effects [48, 49]. Similarity can be defined and determined in different forms, such as chemical, molecular, or biological similarity. According to Maggiora et al. [49], chemical similarity focuses on the physicochemical properties of a compound. These are usually values that describe a global property of a molecule, such as molecular weight, logP, or boiling point. Molecular similarity takes structural features of molecules and their representations into account. It can be defined by shared substructures or topologies of the molecules. In some cases, chemically or structurally very similar molecules might still show large differences in potency, so-called activity cliffs [49–52]. Biological similarity, i.e. comparing the outcomes of biological or biochemical assays, is a strategy to tackle activity cliffs [53–55]. While molecular similarity can be directly computed from chemical structures, traditional biological similarity implies that the compounds to compare were previously synthesised and assayed.

In cheminformatics, similarity between molecules is typically assessed by comparing their representations in the form of descriptors. Descriptors are vectors of binary and/or

continuous values, each value corresponding, for example, to a chemical, molecular, or biological feature of a compound (see Section 3.1).

Similarity between chemicals might be easily determined by eye — if two substances are compared. When working with large chemical databases, a computational similarity search can be useful to retrieve resembling molecules for a query compound and rank them by similarity. Such a similarity search (see Section 3.2 for more details) mainly consists of three steps. First, the molecules are encoded in the form of descriptors (see Section 3.1). Second, a similarity metric is needed to calculate the pairwise similarity between the query compound and each molecule in the database. Finally, the molecules in the database are ranked by similarity to the query molecule.

The concept of a similarity search is fairly straightforward and many different molecule encodings and similarity measures are available. However, not all similarity measures are suitable for comparing compounds in large databases due to hardware and software restrictions [56]. Also the interpretation of the outcomes, which are typically values ranging between 0 and 1, is challenging. The similarity values not only depend on the molecule representation and the similarity measure. Even if a combination is selected, properties such as the molecule size and the position of a dissimilar atom or functional group have an influence [49]. For example, the Tanimoto similarity coefficient, which takes into account the number of shared features by two compounds A and B and the number of unique features in the union of A and B, tends to be lower when comparing small molecules, i.e. molecules with a low bit density [57].

If the similarity between chemicals is sufficient, they might qualify for the read-across approach to fill data gaps. Support for read-across is available, for example, in the OECD QSAR toolbox [19, 35] and the eChemPortal[36]. Although these tools are openly available, they cannot easily be integrated into existing applications, e.g. internal platforms and pipelines in industry.

Structural alerts

Structural alerts, also known as toxicophores, are molecular substructures that have been associated with particular unwanted, e.g. toxic, effects [58]. The detection of structural alerts is an expert method, which assumes that chemicals with similar substructures might exhibit similar toxic effects. Such substructures were initially derived from structure-activity relationship studies [59, 60]. Meanwhile, the collection of available structural alerts has been extended by structural alerts extracted through statistical evaluation and computational methods [31, 59–63]. The application of structural alerts appears already in the early stages of new chemicals' development to avoid the design of potentially toxic molecules [59]. They can, for example, be employed when assembling (virtual) screening libraries for drug discovery, as described by Brenk et al. [64]. The authors provide a list of unwanted, e.g. potentially mutagenic or reactive, groups in their publication. Many other such lists of unwanted substructures are available in the literature [58, 65]. The e-MolTox

webserver for *in silico* drug safety analysis, for example, allows to query a molecule for the occurrence of any of such potentially toxic substructures [65, 66]. ToxAlerts is another web-based platform that offers a collection of structural alerts as well as the use of the alerts for virtual screening to flag potentially harmful substances [58, 67].
Considerable advantages of structural alerts are the fast and easy use, as well as their interpretability [62, 63]. Moreover, if more knowledge becomes available, i.e. from data-driven methods or from experts' experience [31, 59], the lists with undesired substructures can easily be extended. While structural alert approaches (if alert, then toxic) are usually highly sensitive, there is also a high chance for the occurrence of false positives, i.e. not all compounds with an undesired structural moiety are actually toxic as the effects on chemical toxicity depend on the structural environment [31, 68, 69]. Furthermore, there is no guarantee for negative predictions (i.e. if no alert, then non-toxic). The method can only reveal existent knowledge. Therefore, Alves et al. suggest that the appearance of a structural alert should rather be seen as a hypothesis about the mechanism of action or the toxic effect [69]. The relevance of the alert should, however, be assessed by experts. It is also noted that some structural moieties might be responsible for both the mode of action and the adverse effect of a compound [31, 70]. It is, therefore, not always desired to remove any compound with such a substructure. For example, triazole substructures are responsible for the interaction of fungicides with both the intended target fungal lanosterol 14α-demethylase and the homologous human off-target aromatase [71, 72]. Another drawback to keep in mind is that structural alerts only focus on one part of a molecule. If effects depend on multiple groups existing in the same molecule, they cannot be analysed with structural alerts [31].

Machine learning

Machine learning is a method to enable computers to learn from data [73, 74]. For supervised ML methods, the model built on labelled training data is then used to make predictions about the label of yet unseen instances. ML models can be used for the prediction of potential toxicological effects of chemical substances. In the context of relating a property, such as toxicity or activity, to a set of molecular descriptors, ML models are typically referred to as quantitative structure-activity relationship (QSAR) models [31, 42, 75, 76]. In this case, the labels correspond to a specific endpoint, such as a specific type of toxicity, or an assay outcome. The labels are binary or categorical in the case of classification, or continuous variables in regression. The molecules can be represented in the form of descriptors (see Section 3.1). Several traditional ML algorithms, such as random forest, support vector machine, and k-nearest neighbours are employed (see Section 3.3)[7, 77].

With the availability of more powerful computers, new algorithms, and more data, ML has become more popular for toxicity prediction in recent years [42]. A famous promoter for ML model building for toxicity prediction was the Tox21 Data Challenge [77]. The

organisers motivated 40 participating groups to build ML models on nuclear receptor and stress response pathway datasets with about 8000 compounds and submit totally 178 models for final evaluation. While the grand challenge winner (DeepTox) [78] used a deep neural network approach, all winning models achieved high AUC-ROC scores between 0.81 and 0.95 [77], Furthermore, CERAPP and CoMPARA [12, 79] are two recent, highly collaborative, studies on estrogen and androgen receptor modelling. In the Collaborative Estrogen Receptor Activity Prediction Project (CERAPP) and the Collaborative Modeling Project for Androgen Receptor Activity (CoMPARA), 17 or 25 international groups built models for estrogen receptor (ER) (CERAPP) or androgen receptor (AR) (CoMPARA) binding, agonism, and antagonism using QSAR and docking approaches. From these models, one consensus model was built per project with the aim to reveal new potential EDCs.

Although similar descriptors as in similarity search and computational read-across support are used, with ML algorithms, more complex, especially non-linear, patterns can be learnt from the molecules and their labels. Well-performing models can be trained automatically and benefit more from computational rather than toxicological expertise [31, 80]. Nevertheless, ML models should not only make accurate predictions, but also be useful, i.e. with respect to the selected endpoint, applicability domain, and interpretation of the outputs. Therefore, in 2007 the OECD issued guidelines for harmonised evaluation of QSAR models [31, 81].

Tropsha et al. published a series of best practices in QSAR papers, among others focusing on the importance of data curation, model validation, and applicability domain determination [38, 82–84]. Another prominent challenge building ML classification models for toxicological effects are the imbalances between classes [55, 76, 85]. Toxicological datasets typically contain more inactive (non-toxic) than active (toxic) compounds (see Section 1.5 for more detail on toxicity data).

While many ML models and tools for toxicity prediction have been published, ML is still a developing field [31, 65, 86–90]. More research is needed, especially to face challenges with regard to data limitations and applicability domain determination.

1.4 Confidence in machine learning predictions

Whenever an ML model is built and used for predictions, it is crucial to know if the model can be confidentially applied to new data. The desired level of confidence depends on the application context. For compound prioritisation purposes, the risk of a false prediction is inferior, thus a lower accuracy can be accepted. The accuracy of a prediction becomes more important, if a compound is more advanced in the (drug) discovery pipeline, if the study becomes more expensive, and if decisions with respect to hazard and risk assessment are made [91, 92]. While global accuracy of a batch of compounds can be determined in

internal and external validation and may be sufficient during compound prioritisation, the confidence in a prediction for an individual compound needs to be assessed differently [92]. The need for a so-called applicability domain (AD) for ML models is even noted down in the OECD guidelines on the validation of (Q)SAR models. The AD is defined as the 'response and chemical structure space in which the model makes predictions with a given reliability' [93]. 'Response' stands for the output (label) of an ML model. For regression, the label is a continuous variable, which can differ from the label space used to train the model. In classification, the labels are categorical and, thus, the AD is typically based on the (molecular) descriptor space only, i.e. the descriptors used to encode the molecules [94].

Applicability, reliability, and decidability domains

Hanser et al. have outlined that AD is an even more complex concept. A confident decision can only be derived from a prediction if it is valid, reliable, and decisive; it is difficult to address all three aspects with one question. Therefore, they suggested to consider three levels of confidence, i.e. the applicability, the reliability, and the decidability domains (see Figure 1.5) [92, 95].

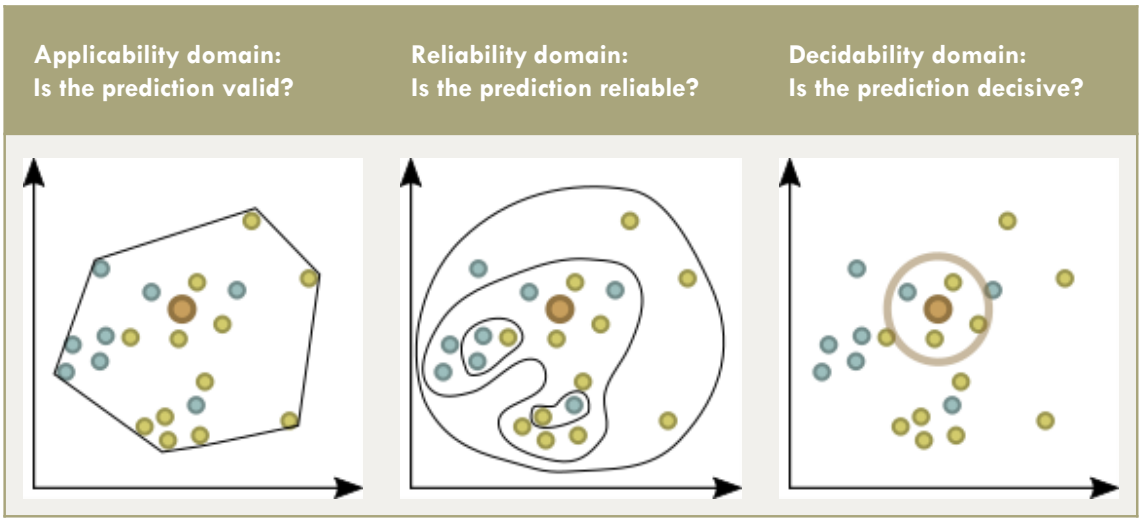


Figure 1.5: The concept of applicability domain can be divided into three levels: applicability, reliability, and decidability domain [92, 95]

The applicability domain (AD_{Hanser}) defines the boundaries, in terms of descriptors, compound classes, and structural features, within which a model is suitable to be applied. These boundaries can, for example, be defined by the range of descriptor values in the training set or by a convex hull, i.e. the minimum convex area, around the first components of a principal component analysis fitted on the training data [92, 94, 95].

The reliability domain (RD_{Hanser}) depends on the quality, quantity, and relevance of the underlying model information, e.g. a prediction of a query compound is expected to be more reliable if more similar compounds are contained in the training set. RD_{Hanser}

can be estimated by distance- and density-based methods, therefore, nearest-neighbour approaches can be helpful. For example, the reliability of a prediction is considered higher if there are more close neighbours, if the distance to the nearest neighbours is smaller, or if the neighbours are more equally distributed. Even within the AD_{Hanser}, a prediction is more reliable if the compound is in an area with higher information density [92, 95].

According to Hanser et al., the AD_{Hanser} and RD_{Hanser} can be estimated before making a prediction, while the decidability domain (DD_{Hanser}) is only defined with the prediction. The DD_{Hanser} estimates if a prediction is decisive or equivocal. It requires probability or likelihood predictions per class label. Such a probability could be derived directly from an ML algorithm, i.e. the number of trees in a random forest that produce a specific label or the label distribution among the nearest neighbours. It is recommended to calibrate the decidability level, so it can reflect the expected accuracy. This can e.g. be achieved within the conformal prediction (CP) framework as explained in Section 3.4 [92, 94, 95].

In a review on chemoinformatic classification methods and their applicability domain, Mathea et al. [94] had already discussed a concept similar to the DD_{Hanser}. They differentiated between AD approaches for novelty detection and confidence estimation. Novelty detection methods only consider the descriptor space and 'mark' compounds which are outside the descriptor space covered by the training set. Confidence estimation, similar to the DD_{Hanser} concept introduced by Hanser et al., takes into account the 'labels' and hence compounds at the decision boundary of the two classes. It was stated that the error rate of a batch of predictions might be reduced more efficiently by rejecting predictions at the overlap region than by excluding remote object predictions — remote objects may still lie on the correct side of the decision boundary. It is, therefore, important to have means to detect compounds at the decision boundary.

Conformal prediction for confidence estimation

Conformal prediction (CP) [87, 96–98] (see Section 3.4) has become a popular method for confidence estimation, with the DD_{Hanser} concept included in the framework. The CP framework is built on top of an ML algorithm and contains an additional calibration step. It is possible to define an accepted error rate, which will not be exceeded given that data is exchangeable, i.e. stems from the distribution [96, 98]. Instead of a single value, the prediction output is a prediction set (classification) or a prediction interval (regression). For binary conformal classification, the prediction set contains either of the classes (single-label prediction), both classes, or it is empty. The interpretation of the three types of prediction sets can be related to the AD_{Hanser} and DD_{Hanser}. An empty prediction set suggests that the compound is outside the AD_{Hanser}. A single-label prediction is returned if the model's confidence is high enough, i.e. if the compound is inside the AD and inside the DD_{Hanser}. If the prediction set contains both classes, the compound is considered inside the AD_{Hanser} but outside the DD_{Hanser} between the two

classes, i.e. the model does not have enough information to assign the compound to one of the classes.

Advantages of the CP framework are the straight-forward interpretation of predictions. Moreover, for a batch of compounds, it can (retrospectively) be checked if the applicability domain assumption is correct, i.e. if training and test data stem from the same distribution. In classification, there is an additional advantage that (Mondrian [99]) CP is usually dealing well with imbalanced data.

The exchangeability assumption is, however, not always fulfilled, i.e. when test instances have drifted from the training data's descriptor space or when transitioning to a new experimental setup for investigating the same biological effects [100, 101]. More research is needed on how to mitigate effects of such data drifts on the performance of CP models.

1.5 Toxicity data as a basis for *in silico* predictions

The aforementioned toxicity prediction approaches all rely on sufficient and, ideally, consistent and standardised data [76]. In particular, the quality of ML models, which are trained on large datasets, relies on well-curated and standardised data.

These desires are not always satisfied, particularly for toxicological datasets, which are often sparse and imbalanced [42, 55, 76, 85]. This can be explained by economical and ethical considerations: if a compound is confirmed toxic for one crucial endpoint, it will most likely be attrited and not tested in any further assay. Moreover, only non-harmful compounds should be submitted to regulatory agencies (and will be published).

Therefore, different actions have been taken to improve the situation. First, initiatives such as the Tox21 and ToxCast programs set their sights on generating more complete publicly-available *in vitro* datasets [77, 102, 103]. The combined Tox21 and ToxCast dataset contains about 8000 compounds with measurements taken in about 1000 consistent HTS assays and are a good basis for the training of ML models. Second, 13 pharmaceutical companies teamed up in the eTOX project with the aim to share pre-clinical toxicological data to be used for read-across, model building, and evaluation [3]. Last but not least, there are large bioactivity databases, such as PubChem and ChEMBL. The PubChem database, hosted by the US National Institutes of Health, contains information on chemical substances and their biological activities — 96.5 million unique chemical structures and 1.25 million biological assays as of August 2018 — which are provided by university labs, governmental agencies, and industries [104, 105]. The ChEMBL database is developed and maintained by the EMBL-EBI in the UK and consists of manually extracted information on binding, functional, and ADMET properties (of drug-like bioactive compounds) from the literature [106–108]. ChEMBL Release 28 contains more than 2 million distinct compounds and activity measurements for more than 14,000 targets [108]. The PubChem and ChEMBL databases report distinct and complementary data types.

Nevertheless, there is some overlap in the databases as PubChem and ChEMBL have developed a mutual exchange mechanism for dose-response assay data [104, 107].

To manage the combination of data produced in different laboratories, standardisation tools or pipelines for chemical structures are utilised [104, 106]. Particularly, ChEMBL has recently put substantial effort into structure curation and published an open source chemical structure curation pipeline [109]. Another popular and publicly-available standardisation tool is the IMI eTox project standardiser tool by Atkinson [110]. Important steps of chemical structure curation are the removal of salts and solvents, charge neutralisation, and normalisation of structures according to defined rules and conventions.

High quality data lay the foundation for well-performing predictive models. And only if the predictions are accurate and reliable, toxicologists and regulatory agencies may be persuaded by the usefulness of *in silico* models.

Aim and Objectives

Ethical and economic aspects of hazard assessment of new chemicals motivate the replacement, reduction, and refinement of animal testing. Computational methods are promising tools to predict potential toxicological effects of new chemicals. While methods such as similarity search, structural alerts, and machine learning are already applied at different stages during the development process, some important challenges remain. The main constraints of computational toxicity prediction methods are the limited data and the need to assess the confidence in predictions. The aim of this thesis is to explore several toxicity prediction methods, with a focus on the CP framework, to answer the following questions:

- Can the effects of limited data be mitigated by combining different computational toxicity prediction methods? Does this strategy lead to complementary information gain or a consensus when in a case study applied to compounds from industry?

- Does the hit rate in toxicity screening improve when combining *in silico* and *in vitro* methods? Is this approach useful in the identification of endocrine-disrupting chemicals?

- Can the need for experimental measurements be bypassed when using conformal predictions to generate predicted bioactivity descriptors? How do models based on such descriptors perform compared to using molecular descriptors?

Although CP can be useful in the above described cases, the challenge of restricted model applicability domains remains. Data drifts between old training and new test data restrain the application of CP models. It would be helpful to find a strategy to mitigate the effects of such data drifts.

- How can CP be used to assess and mitigate the effects of data drifts between the datasets from the Tox21 Data Challenge?

- Is the strategy developed for the Tox21 data transferable to other datasets, i.e. can it be applied to temporal data drifts and differences in chemical or assay space between external and internal data?

The outcomes of the thesis should provide strategies to tackle some of the main challenges of computational toxicology, such as data limitations and effects of data drifts. With this, it aims to promote the application of computational toxicology in the reduction of animal experimentation.

Computational methods

In this section, first, the descriptors used in this thesis are introduced. Then, the methodology of similarity search is explained, followed by an introduction of the ML algorithms. Finally, the CP framework is described.

3.1 Molecular, chemical, and bioactivity descriptors

For similarity searches and the building of ML models, compounds need to be encoded in a computer readable format. Descriptors can consist of binary or count features, also called fingerprints, or contain continuous variables per feature. In this thesis, several types of molecular, chemical, and bioactivity descriptors were employed, which will be described in the following.

Molecular ACCess System keys

Molecular ACCess System (MACCS) keys [111] are SMARTS-pattern based descriptors. MACCS consists of 166 keys that were pre-defined by medicinal chemists. Every key — corresponding to one bit position — checks whether the molecule contains a certain functional group or atom type. Rules for MACCS keys differ slightly between implementations [111]. For this work, the RDKit Python library [112] was used to calculate the MACCS keys; a list of the 166 rules in RDKit can be found in [113]. Since the MACCS descriptor has a pre-defined number of 166 keys, it is quickly calculated, and can, furthermore, be interpreted easily. As a disadvantage, the few keys might be insufficient to distinguish similar molecules. For example, key 134 checks for the appearance of any halogen and does not differentiate Cl from F substituents.

Extended Connectivity Fingerprints

Extended Connectivity Fingerprints (ECFPs) [114] are 2D topological descriptors which were specially developed for structure-activity modelling. In the ECFP, every bit corresponds to an atom of the molecule and its circular environment of a given diameter. For a detailed explanation of the algorithm, the reader is referred to the publication by Rogers and Hahn [114]. For the generation of identifiers, which are comparable between

molecules, every (core) atom is hashed to a 32-bit integer. The 2^{32} long, but sparse fingerprint can be folded into a smaller descriptor, e.g. 1024 bit long. Such shorter fingerprints use less storage and calculation time (e.g. for model building), although some information and interpretability can be lost due to rare bit collisions. In this work, the RDKit [112] implementation of the Morgan fingerprint was used, which follows the same above described algorithm, with the exception that it takes the radius instead of the diameter as input parameter. As a consequence, Morgan (radius 2) is equal to ECFP (diameter 4).

Physicochemical descriptors

Physicochemical descriptors are one-dimensional properties with continuous (non-binary) variables, which can, among others be calculated within RDKit [112]. Examples of such physicochemical descriptors are the molecular weight, logP, or the number of rotatable bonds. A list of available descriptors in RDKit can be found in the documentation [115].

Signature molecular descriptors

The algorithm of the signature molecular descriptor [116, 117] defines fragments of a molecule specified by a 'height' parameter, i.e. the number of atomic bonds and enumerates them. In this work, the count of these fragments was used. Note that signature descriptors are extremely sparse and that the enumeration of fragments is dataset-dependent.

Bioactivity descriptors

For the bioactivity descriptors, used in this work, CP binary classification models were trained on 373 datasets for *in vitro* biological effects that can lead to toxicological endpoints, such as cytotoxicity, genotoxicity, or thyroid hormone homeostasis. These models were used to calculate the predicted p-values (per model and class) for new compounds. The predicted p-values were used as bioactivity descriptors.

3.2 Similarity search

A computational similarity search can be useful to find compounds which are similar to a query molecule within a large database. Such a similarity search requires, first, a way to describe the molecules (see Section 3.1 for more details on descriptors) and, second, a measure to calculate the similarity between two descriptors [49, 56]. A well-known similarity coefficient for binary descriptors is the Tanimoto index [118]. It is calculated by the number of features shared by two compounds A and B divided by the number of unique features in the union of A and B. For count fingerprints or descriptors with continuous variables, another metric such as euclidean distance is required.
After encoding the molecules and calculating pairwise similarity to the query molecule,

the compounds are ranked by similarity. It can be helpful to highlight the substructures that two molecules have in common, e.g. using RDKit [112].

3.3 Machine learning algorithms

The supervised ML algorithms used in this thesis are explained in the following. These are random forest (RF) and support vector machine (SVM) for classification and k-nearest neighbours (KNN) for regression.

Random forest

RF is an ensemble learning method which consists of a group of decorrelated decision trees. Each tree is built on a random subset of the training data and a given number of randomly selected features and is later used to make a prediction. The final RF prediction is based on a majority voting system between all trees. For classification, the class or the probabilities per class can be returned. The main parameter to be tuned is the number of decision trees, which is ideally high for increased classifier performance (but adds computational cost). Moreover, the RF ensemble learning approach typically yields quite robust models with some sacrifice in interpretability [73, 74, 119].

Support vector machine

In an SVM, the classes are separated by a hyperplane; the samples, which lie nearest to the hyperplane are called support vectors. The model is optimised by maximising the margin between the decision boundary and the closest samples. Since many datasets might not be linearly separable, a kernel function can be introduced, i.e. the samples are projected into a higher-dimensional space by creating nonlinear combinations from the original features. The kernel helps to reduce the computational cost, i.e. by transforming high-dimensional distance measures to a similarity measure between 0 and 1. Examples for kernels are linear, polynomial, or the RBF (radial basis function). In the RBF, the value is only dependent on the distance to a fixed point. The most important SVM tuning parameters are the cost function C and the kernel coefficient γ. C stands for the penalty accounted for instances on the wrong side of the hyperplane. γ determines how much influence a single training example has. With a small γ, the influence of an individual data point reduces further [73, 74, 119].

k-nearest neighbours

The KNN algorithm makes a prediction for a test sample based on the label of its nearest neighbours. The algorithm is fundamentally different from RF and SVM as its discriminating power lies in memorising the training dataset [120]. The algorithm basically consists of three steps. First, the number k of nearest neighbours to consider (or the

distance within which neighbours should be considered) is defined. Second, the nearest neighbours around the sample to predict are identified. Third, a prediction is computed using the mean of the sample's neighbours. Note that the KNN algorithm does not involve a training step. This comes with the advantage of easy expandability with additional training data and the drawback of computational complexity growing linearly with the number of training samples [74, 119].

3.4 Conformal prediction

Besides obtaining internally valid ML models, it is crucial to estimate the confidence in the predictions made on an external test set. In this work, CP [87, 96–98] was used for confidence estimation. In the following, the CP framework as well as the evaluation of CP models are explained. Note that the description is referring to the Mondrian setting [99] (i.e. separate handling of active and inactive compounds for calibration) for binary classification, as used in this work.

Originally, CP was designed in an on-line setting, meaning that, after each prediction, the label was determined and used to inform the model before predicting a new sample [121]. In ML applications, the computationally more efficient inductive conformal predictors (ICPs) [122] (see Figure 3.1a) are typically used.

Analogous to ML, a training and a test set are the basis for CP model training and validation. To allow for the additional calibration step, the training set is split into a so-called proper training and a calibration set. An ML model is then trained on the compounds (descriptors and binary labels) of the training set and used to make predictions for the calibration and the test set compounds. The prediction outputs from the ML model are further transformed into so-called nonconformity scores (nc scores), which indicate how unusual a certain prediction is relative to previous predictions [121, 123].

Note that the use of a normaliser model (see Figure 3.1f) is optional in classification. This would be an additional regressor model, e.g. KNN regression, which is trained on the nc scores of the proper training set. The model output is the mean nc score of the nearest neighbours for a query compound and can be used to normalise the nc score obtained from the base model. This normalisation is important for conformal regression [123–125], but has, to our knowledge, not been applied to classification before. In this work, it was shown useful in the application of CP classification models, trained on public data, to internal data from industry.

Following the binary and Mondrian setting [99], the nc scores for the calibration set compounds are sorted into two lists; one with the nc scores for the active class compounds; and one for the inactive class, as illustrated in Figure 3.1b. The nc scores for the test compounds, which are two values per instance (1 active, 1 inactive), are arranged within the lists to calculate the two p-values (see Figure 3.1b-d). A p-value is calculated as

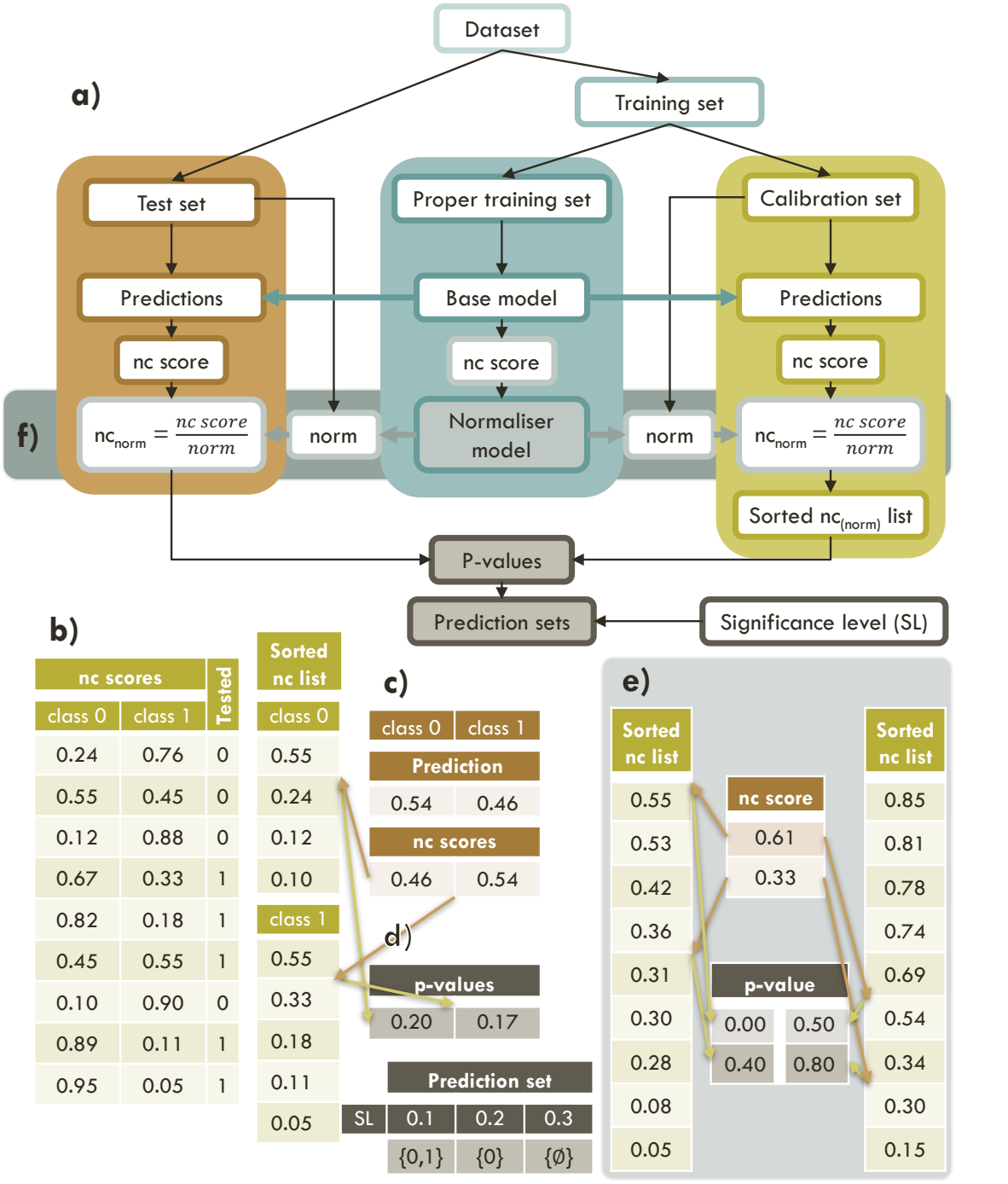


Figure 3.1: Concepts of conformal prediction. a) Overview of an ICP. b) Nonconformity scores (nc scores) for an example calibration set with four inactive (class 0) and five active (class 1) compounds. The nc scores are sorted into two lists. c) Predictions and nonconformity scores (calculated using an inverse probability error function) for one example compound. d) Arranging the nc score for the example compound into the nc score lists from the calibration set, the p-values can be calculated. Depending on the selected significance level, the subsequent prediction sets may differ. e) The calculated p-value depends on the calibration set. f) Optionally, an additional normaliser model can be included in the workflow.

the ratio of the (previously predicted) calibration set compounds belonging to the class, which are more unusual, i.e. having a higher nc score, than the test instance.

The concept of calibration could be compared to the calibration of a measuring device in an experimental setting. Analogously, the device is calibrated based on the conditions (e.g. temperature, air pressure, or moisture), and the CP model needs to be calibrated with compounds from a similar feature or prediction space. A different calibration set leads to different p-values and thus different evaluation values, as illustrated in Figure 3.1e.

While ML classification models return a single-point estimate, i.e. the labels or the probability belonging to a class predicted for a test instance, the CP output is a prediction set. For the prediction sets, a significance level is chosen, which refers to the maximum accepted error rate in the predictions and is commonly set to 0.2 or 20% [126]. The prediction set contains all class labels, for which the p-value is larger than the significance level. Exemplified in Figure 3.1d, for two example p-values of 0.20 (class 0) and 0.17 (class 1) and the significance level 0.2, the prediction set is $\{0\}$. However, with an expected error rate of 0.1, the prediction set would be $\{0, 1\}$ and if more errors, i.e. 0.3, are accepted, the prediction set would be empty ($\{\emptyset\}$).

For the evaluation of the prediction sets it is important to know that in CP, a prediction (set) is defined to be correct if it contains the correct label. This means that a prediction set is always correct if it contains both labels ($\{0,1\}$) and wrong if it is empty ($\{\emptyset\}$), although these sets do not provide much information.

The main CP evaluation measures are validity and efficiency, which are calculated for a chosen significance level. Validity is defined as the ratio of correct predictions, i.e. prediction sets containing the correct label. A typical efficiency measure is the ratio of single-class predictions (i.e. $\{0\}$ and $\{1\}$). Since (informationally) efficient predictions are not necessarily correct, often the additional 'accuracy' evaluation measure is used. Accuracy is calculated as the proportion of correct single-class predictions.
Note that while the non-efficient predictions do not add any information on the compounds' label, they can provide useful information about new compounds to be assayed for model improvement. Screening of more compounds from the empty category can improve the applicability domain while assaying more chemicals from the both category can improve the decidability domain [85].

Results

4.1 KnowTox: pipeline and case study for confident prediction of potential toxic effects of compounds in early phases of development

In silico toxicity prediction methods are promising alternatives to animal testing. Their potential has, however, not yet been fully exploited. The applicability of ML models is, for example, limited by the size of available datasets and the need for prediction confidence estimates. In the following work, we explore if combining toxicity prediction methods and presenting the predictions in the form of a holistic picture can lead to more informative and reliable predictions. To this end, the Python-based KnowTox tool will be developed. Designed for prioritising compounds and guiding toxicity testing, KnowTox combines machine learning models, alerts for toxic substructures and read-across support. To illustrate the value of the holistic approach in the early stages of new chemical's development, KnowTox will retrospectively be applied to a case study with proprietary compounds from industry.

Contribution:

First author
Conceptual design (30%)
Computational experiments (90%)
Visualization (100%)
Manuscript preparation (80%)

Morger *et al. J Cheminform* (2020) 12:24
https://doi.org/10.1186/s13321-020-00422-x

Journal of Cheminformatics

RESEARCH ARTICLE Open Access

KnowTox: pipeline and case study for confident prediction of potential toxic effects of compounds in early phases of development

Andrea Morger[1], Miriam Mathea[2], Janosch H. Achenbach[2], Antje Wolf[2], Roland Buesen[2], Klaus-Juergen Schleifer[2], Robert Landsiedel[2] and Andrea Volkamer[1*]

Abstract

Risk assessment of newly synthesised chemicals is a prerequisite for regulatory approval. In this context, in silico methods have great potential to reduce time, cost, and ultimately animal testing as they make use of the ever-growing amount of available toxicity data. Here, KnowTox is presented, a novel pipeline that combines three different in silico toxicology approaches to allow for confident prediction of potentially toxic effects of query compounds, i.e. machine learning models for 88 endpoints, alerts for 919 toxic substructures, and computational support for read-across. It is mainly based on the ToxCast dataset, containing after preprocessing a sparse matrix of 7912 compounds tested against 985 endpoints. When applying machine learning models, applicability and reliability of predictions for new chemicals are of utmost importance. Therefore, first, the conformal prediction technique was deployed, comprising an additional calibration step and per definition creating internally valid predictors at a given significance level. Second, to further improve validity and information efficiency, two adaptations are suggested, exemplified at the androgen receptor antagonism endpoint. An absolute increase in validity of 23% on the in-house dataset of 534 compounds could be achieved by introducing KNNRegressor normalisation. This increase in validity comes at the cost of efficiency, which could again be improved by 20% for the initial ToxCast model by balancing the dataset during model training. Finally, the value of the developed pipeline for risk assessment is discussed using two in-house triazole molecules. Compared to a single toxicity prediction method, complementing the outputs of different approaches can have a higher impact on guiding toxicity testing and de-selecting most likely harmful development-candidate compounds early in the development process.

Keywords: Toxicity prediction, ToxCast, Read-across, Random forest, Conformal prediction, Confidence estimation, Applicability domain, Case study, Androgen receptor, Triazoles

Introduction

Before newly developed chemicals can be approved, their potential toxic effects on humans and the environment inevitably need to be assessed. Most regulations such as REACH [1] require animal studies for risk assessment. E.g. more than 540,000 animals were employed in Germany in 2017 for production, quality control, and safety assessment [2].

Given the ever growing amount of available toxicity data, computational toxicity prediction methods have great potential to reduce time, cost, and ultimately

*Correspondence: andrea.volkamer@charite.de
[1] In Silico Toxicology and Structural Bioinformatics, Institute of Physiology, Charité Universitätsmedizin Berlin, Charitéplatz 1, Berlin, Germany
Full list of author information is available at the end of the article

animal testing. Using historical data, they can help to disclose relationships between compounds that would not have been identified manually and, thus, reveal potential risk of compounds in early phases of development. In silico predictions can hint at potentially hazardous interactions or critical structural moieties of new molecules. If the corresponding assays are conducted first, harmful compounds can be filtered out before performing a wide range of additional experiments. Moreover, in silico methods can support product optimisation and reduce long-term animal toxicity studies [3, 4].

In silico strategies for supporting risk assessment range from computational read-across approaches and search for substructural alerts to statistical methods. Especially, quantitative structure-activity relationship (QSAR) techniques such as machine learning (ML) [5] methods require a large precompiled dataset.

The US Environmental Protection Agency (EPA) has provided the ToxCast dataset [6] consisting of roughly 8000 compounds, such as pharmaceuticals, pesticides, and environmental chemicals, that were tested on up to 1000 endpoints, e.g. cell cycle, steroid receptors, and cytotoxicity. ToxCast has since been used: to develop QSAR models [7–9]; to generate biological fingerprints for in vivo endpoint predictions [10]; to decipher adverse outcome pathways [7, 11]; and as a basis for read-across [12–14].

Read-across is a common, often manual, approach in toxicology [12, 15, 16], based on the assumption that similar molecules can evoke similar toxic effects. Missing information on query chemicals' properties may be gathered by reading across information from very similar molecules. Using different molecular encodings and diverse similarity measures, computers can search through large compound databases to identify the most similar compounds and—given a decent similarity—transfer knowledge to a query compound. Prerequisite for successful read-across is a robust and reproducible test system of the underlying experimental data [16], i.e. a standardised assay set-up to ensure comparable read-outs. Another challenge is the determination of the amount of required similarity between two compounds that allows safe and reliable knowledge transfer.

Since often not the complete molecule, but rather a specific functional group or fragment, is responsible for an unwanted effect, identifying such toxic substructures in a query molecule is of high practical value. Several authors published lists of toxic alerts or other undesired substructures which can be used to flag novel compounds [17, 18]. For instance, the OCHEM ToxAlert server allows to browse and query structural alerts for various toxicological endpoints [17, 19].

Often the relationship between molecular structure and toxic effect is not linear, thus, statistical methods such as QSAR models are applied to recognise more complex patterns in datasets. The set-up of high-performing toxicity prediction models has recently been promoted in the Tox21 Data Challenge. Research groups competed in model performance on 12 nuclear receptor and stress response pathways trained on roughly 10,000 compounds [20], including various ML algorithms such as random forest, support vector machine, and deep learning approaches [21–23]. The winning models on all 12 endpoints showed AUC-ROC scores between 0.81 and 0.95 on an external blinded test set [20].

Fuart Gatnik and Worth published an overview on publicly and commercially available software tools, such as the well-known TOPKAT [24] and DEREK [25] methods, for toxicity prediction [26]. Concluding, the authors stated that the availability and quality of the models is endpoint-dependent and they emphasised on the observation that generally more research is needed in terms of assessment of the applicability of the in silico models. Besides pure predictions, for practical applications, knowledge about the applicability domain, i.e. the space of chemicals the model can make reliable predictions for, is of major importance. Hanser et al. [27] suggested to further divide this concept into three domains: applicability, reliability, and decidability. The applicability domain indicates whether a model can be applied to make a prediction for a certain use case. It can be defined, for example, by a convex hull around the main components of a principal component analysis (PCA) fitted on the features of the training data. The reliability domain gives information on whether the obtained prediction is reliable enough for the use case. It can be explored by investigating the average distance to the nearest neighbours. The decidability domain returns if a clear decision can be made, based on the outcome of the prediction. Therefore, the distribution of the nearest neighbour's labels can be analysed [27].

A recently promoted method for confidence estimation, especially regarding reliability and decidability, is conformal prediction (CP) [28, 29]. A conformal predictor returns, whether enough evidence is given to reliably assign the query substance to a certain class. CP models have recently been developed and applied in drug discovery [29–31], and toxicology, e.g. to predict cytoxicity [32], endocrine disruption [33], and skin penetration [34]. Moreover, recently, eMolTox was introduced, a webserver offering 174 CP models [35]. However, to the best of our knowledge, few information about applying such models to real-world use cases has been published.

In this work, KnowTox, a holistic toxicity prediction approach, that integrates refined conformal predictors,

structural alerts, and read-across support based on molecular similarity, is introduced and applied to industrial chemicals. The main source of toxicity information is the publicly available ToxCast dataset. Being aware of the challenge to apply ML models trained on public data to an industrial setting, first, the CP model performance was optimised focusing on the androgen receptor endpoint and validated on an in-house dataset. The focus is on endocrine disruption as a disturbance of steroidal hormone homeostasis can cause severe toxic effects, e.g. leading to male feminisation or reproduction disorders [36, 37]. Thus, screening for agonistic and antagonistic activities on androgen and estrogen receptors is frequently conducted in yeast cells (so-called YES- and YAS assays [38]) and sufficient validation data is available. Finally, CP models were trained using the same CP set-up for another 87 ToxCast endpoints with enough training data available. Moreover, with KnowTox, the refinement of chemical structures is guided by the implementation of warnings about unfavourable structural moieties described in literature [35, 39, 40]. To support read-across, a similarity search is proposed which can automatically point to toxic effects in cells and interactions known for the most similar molecules within ToxCast. In a case study, the potential of KnowTox is exemplified on two in-house triazoles. Multiple components of the KnowTox pipeline indicated liver toxicity and endocrine disruption which is in accordance with literature and retrospective test results.

Data and methods

In the following, first the main datasets and their preparation will be introduced, followed by the individual methods for the KnowTox toxicity prediction tool, including CP, PCA, toxic substructure and similarity search.

Datasets

ToxCast dataset

The source of molecules and assay data for KnowTox is the freely available ToxCast dataset provided by the EPA. It consists of over 8000 compounds tested on up to 1092 different toxic endpoints. The data was downloaded from EPA's National Center for Computational Toxicology [41] (date 23.06.2017). Toxicity values were directly adopted from the hitcalls defined by the EPA. Flags were not considered, but endpoints corresponding to background measurements were excluded. This yielded a sparse matrix of 8390 compounds with respective toxicity value (0,1, NaN) per tested endpoint (985 total). The ToxCast dataset represented the basis for the similarity search as well as for CP.

Table 1 Size and purpose of androgen receptor antagonism datasets used to validate the original conformal prediction model

Dataset	Purpose	Actives	Inactives
ToxCast-AA 762	Train and test model	868	5842
In-house-AA	Validation I	280	254
External-AA	Validation II	160	201

Androgen receptor datasets

To validate and optimise the CP set-up for model application on external data, three datasets for androgen receptor antagonism (AA) were collected (see Table 1).

ToxCast-AA The AA assay from ToxCast (assay endpoint id 762) was selected. The assay originates from the Tox21 platform and was conducted in human kidney cells (HEK293T). It is a reporter gene assay that measures beta lactamase induction upon antagonistic activity regulated by the human androgen receptor. Activity data are available for 6710 chemicals.

In-house-AA The in-house dataset from BASF consists of 534 chemicals tested in YES/YAS assays [38]. They are mainly pesticides, such as fungicides and herbicides, and not part of the ToxCast dataset. These compounds were not launched on the market but failed for different reasons during the development. In the YAS assay, human androgen receptor is expressed in yeast cells. Upon binding of an androgenic compound, the lacZ reporter gene is activated, which is responsible for expression of β-galactosidase. Presence of this enzyme can be detected by a colour change. Anti-androgenic effects can be observed if binding of a known androgenic agent is inhibited and thus the colour change is reduced or does not occur at all. YES assays are conducted similarly, but in yeast cells that express the human estrogen receptor.

External-AA Another external dataset, collected by Jensen et al. [42] and by Vinggaard et al. [43] for QSAR modelling, was downloaded from Norinder et al. [33]. The dataset consists of initially 925 molecules that were especially selected to represent a large chemical space [43]. 361 of these molecules, that are not part of ToxCast, were used in this study. Data originate from an AA assay reporting luminescence response upon inhibition of androgen binding to a synthetic androgen receptor and following gene expression in chinese hamster ovary cells.

Dataset preprocessing

Standardisation

Each molecule was standardised by applying the following workflow: first, duplicates (compounds tested more than once for a specific endpoint) were removed. Only one instance was kept if the assay outcomes agreed—otherwise both instances were discarded. Next, molecules were standardised using the IMI eTox project standardiser tool [44]. This included discarding non-organic compounds, application of certain structure standardisation rules (e.g. handling of tautomers, shifting protons between heteroatoms), neutralisation, and removal of, mainly organic, salts. Due to this standardisation step new duplicates occurred; they were treated as described above. Next, remaining mixtures as well as fragments with less than three heavy atoms were removed yielding a cleaned dataset of 7912 ToxCast molecules tested on up to 985 endpoints (see Fig. 1, top). The resulting total number of active and inactive compounds for the AA datasets are listed in Table 1.

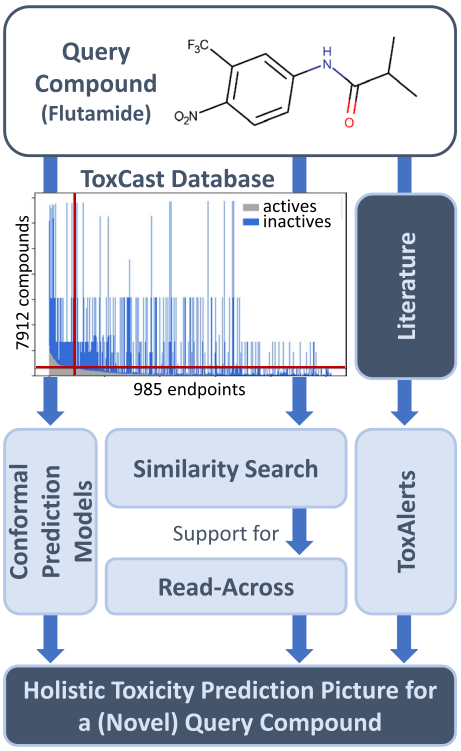


Fig. 1 Overview of KnowTox. Combining toxicity information from different sources, the complementary outputs of the KnowTox tool help to generate a holistic toxicity prediction picture for a novel query compound. ToxCast Database bar plot: Number of actives (grey) and inactives (blue) available per endpoint, sorted by number of actives. Red vertical line: CP models were built for the endpoints on the left side of the threshold line (at least 300 active and inactive compounds tested, red horizontal line)

Descriptor calculation

For similarity search as well as CP, all molecules were encoded by molecular descriptors implemented in RDKit. For similarity search and the original CP model, a combination of the SMARTS-pattern based MACCS keys and the circular-environment based Morgan fingerprint (radius 3, 1024 bits) was chosen. MACCS keys [45] represent the presence or absence of predefined functional groups. Morgan fingerprints [46] are a more abstract representation of a molecule, covering every atom and its circular environment including all atoms and bonds within a defined radius. Concatenation of the two descriptors resulted in a 1191-bit long feature vector representation per molecule. For the normalised and normalised + balanced CP models (see Table 2), the concatenated descriptor (binary values) was reduced to bits with feature variance of equal or higher than 0.01. Additionally, 200 physicochemical descriptors within RDKit [47] (float values) were calculated, normalised and reduced (feature variance threshold 0.001). Finally, these two descriptor sets were concatenated resulting in a feature vector of length 1341. Normalisation of physicochemical parameters and feature reduction were performed based on all standardised ToxCast molecules.

KnowTox pipeline

KnowTox allows input of a query molecule and offers in silico support for risk assessment from various view points, comprising CP, similarity search to support read-across, and search for toxic substructures (see Fig. 1). In the following, the individual methods will be explained.

Table 2 Conformal prediction models built for androgen receptor antagonism

Model name	Descriptors	nc[a]	Balancing
Original	Morgan + MACCS	Default	No
Normalised	Morgan + MACCS + physchem[b]	Normalised	No
Normalised + balanced	Morgan + MACCS + physchem[b]	Normalised	Yes

[a] nc: nonconformity score

[b] physicochemical descriptors

Machine learning and conformal prediction

General CP workflow The CP framework is built on top of ML models and is designed to make valid predictions at a given significance level (SL), assuming exchangeability [30]. An overview of the CP workflow used here (offline-mode, binary classification setting) is shown in Fig. 2. Similar to the standard ML setting, the dataset is stratified and randomly split into a training and a test set. Then, an additional calibration step is introduced, in which training data is further split into a proper training and a calibration set. An underlying ML model, e.g. a random forest, is fitted on the proper training set and used to make a prediction (probability $\hat{p}$) for compounds of the calibration and the test set. The prediction outcome per class is transformed into a so-called nonconformity score (nc score). A nonconformity error function is chosen in the way that more ideal predictions yield lower nc scores; a typical error function for random forest classification models is the inverse probability (Eq. 1):

$$nc\ score = 1 - \hat{p} \tag{1}$$

To improve reliability estimation of predictions, an additional normaliser regression model (e.g. kNN) can be fitted on the descriptors of the proper training set and their nc scores. For a new compound, the normaliser regression model returns a normalised nc score (nc $score_{norm}$), by dividing the nc score of the compound by the average nc score of the compound's k nearest neighbours within the proper training set (see Eq. 2).

$$nc\ score_{norm} = \frac{nc\ score}{norm} \tag{2}$$

Using mondrian classification [48], the CP algorithm generates for each class a sorted list of nc scores or nc $scores_{norm}$ for the calibration set. The ratio of these nonconformity scores higher, and thus more nonconforming, than the nc score predicted for a query compound is called p-value. If a p-value is larger than a given SL ϵ (maximum allowed error rate), that label is assinged to a compound. Thus, for a binary classification problem, the output prediction set per compound contains either one class ({0},{1}), both classes ({0,1}), or an empty prediction set ({}). To obtain more stable predictions, multiple conformal predictors can be trained and the p-values are averaged, so-called aggregated conformal predictors (ACPs) [49] are generated.

CPs are typically evaluated regarding validity, efficiency and accuracy. Validity is defined as the ratio of predictions containing the correct label. A common efficiency measure is the ratio of single class predictions (SCPs). Accuracy of SCPs corresponds to the ratio of correct SCPs divided by all SCPs.

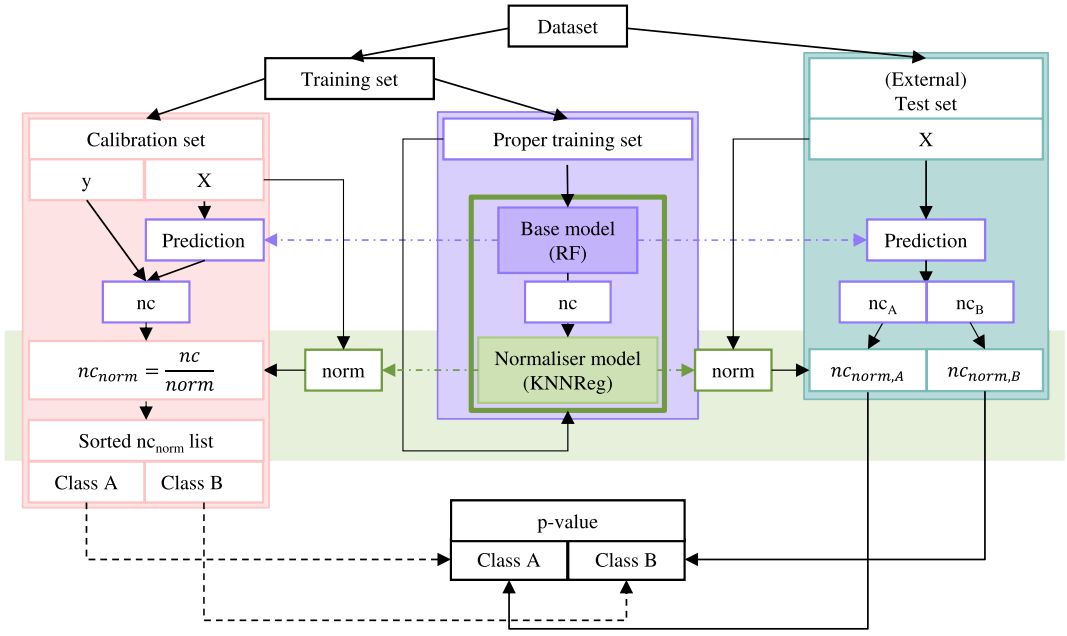


Fig. 2 Schematic description of CP workflow. Data is split into training and test set (blue box). The training set is further divided into calibration (red box) and proper training set (violet box). An ML model is fitted on the proper training set and used to predict compounds of the calibration and test set. Predictions are transformed into nonconformity scores (nc scores). Calibration is conducted by sorting the nc scores of the calibration set (class-wise, mondrian) into two lists. The nc score of a test compound is arranged in the list and thus the p-value calculated. An additional normaliser model (green box) can optionally be fitted on the descriptors and nc scores of the compounds of the proper training set

CP model set-up in this study Three different settings for CP were applied. The corresponding models will further be called 'original', 'normalised' and 'normalised + balanced' model (see Table 2).

For the original model, data was split into 80% training and 20% test data. Within each loop of a fivefold cross-validation, an ACP with 25 loops was generated. In each ACP loop, training data was split into 70% proper training and 30% calibration data (see Carlsson et al. [49]). Random forest models (500 estimators, else default parameters) were trained on the proper training sets and the predictions calibrated using the respective calibration sets (inverse probability error function, mondrian condition). P-values were aggregated by their median as suggested by Linusson et al. [50]. Finally, the mean p-value of the cross-validation was calculated.

For the normalised model, information from the nearest neighbours in the training set was taken into account as described in Eq. 2. The normaliser model was fitted using the KNNRegressor algorithm (scikit-learn, default parameters).

In the normalised + balanced model, per ACP loop, the proper training and calibration data were five times randomly subsampled to equal numbers of actives and inactives.

After evaluation, normalised + balanced models were built for all ToxCast endpoints for which at least 600 compounds were measured—300 active (toxic) and 300 inactive (non-toxic)—yielding 88 CP models (see Fig. 1, ToxCast Database bar plot, vertical red threshold line).

Principal component analysis (PCA) for AA data

For chemical space analysis, a 2-component PCA was fitted on ToxCast AA data. ToxCast-AA, in-house-AA, and external-AA data were projected into the descriptor space. Same descriptors were used as described for the normalised and normalised + balanced CP models.

Structural alerts

To identify potentially toxic or unwanted substructures in the query molecules, known structural alerts, encoded as SMARTS patterns, collected from literature are used. A list of 919 structural alerts incorporated in KnowTox was kindly provided by the authors of eMolTox [35]. Using RDKit, a substructure search for all these patterns in the query molecule is performed. Matching substructures are stored together with information about the associated toxic effect, individually highlighted in the molecule and labelled.

Similarity search and read-across

Computational support for read-across in KnowTox is implemented via a similarity search and subsequent extraction of information from ToxCast. For similarity search, a query compound is compared to all ToxCast compounds using the calculated descriptors. Finally, ToxCast compounds are ranked by Tanimoto similarity to the query compound. The tool returns the most similar compounds together with their respective maximum common substructure (MCS) with the query compound highlighted. Subsequent read-across is supported by extracting experimental activity of these similar molecules from the ToxCast dataset for all 985 endpoints.

Python libraries and versions

Molecules were standardised using the standardiser library [44] version 0.1.9. Descriptor calculation, structural alerts and similarity search were implemented using RDKit [47] version 2018.03.4. For local calculation of feature variances, normalisation of physicochemical parameters, and PCA, scikit-learn [51] version 0.19.2 was used. CP models were trained using nonconformist [52] version 2.1.0 and underlying ML models using scikit-learn (version 0.19.0). Plots were generated using matplotlib version 2.2.3.

Supplementary information on github

A github repository with supplementary information is provided under https://github.com/volkamerlab/knowtox_manuscript_SI. It contains the pre-processed ToxCast and external-AA data, as well as a notebook demonstrating the conformal prediction set-ups used in this work.

Results and discussion

In this section, first, the optimisation of the CP model with respect to applicability to in-house and external data will be discussed, with focus on prediction of AA assay outcome as well as on the complete set of 88 ToxCast endpoints. Finally, the full spectrum of predictions provided by KnowTox will be show cased on two triazoles.

Conformal predictors—validation of AA model

The aim of this study was to generate reliable toxicity prediction CP models which can be applied to in-house industrial chemicals. Data from the freely available and comparably large ToxCast toxicity database is used, which contains experimental data from consistent measurements per assay endpoint. As there is a shift in chemical and descriptor space expected, when applying the models to in-house compounds, it is important to validate the method carefully. Thus, a CP model to predict androgen receptor antagonism (AA) was selected for validation. Here, an in-house dataset with 534 industrial compounds was available, as well as another external dataset with 361 compounds. AA is an important endpoint to examine a compounds' risk for endocrine

disruption disorders such as male feminisation or sexual disruption in fish [36] and other species [37].

By design, conformal predictors are valid at a given SL, assuming data exchangeability [30]. This is also observed when training a standard CP model on ToxCast AA data. Figure 3a shows a calibration plot of the internal validation of the original ToxCast-AA model. Ideally, the error rate is equal to the significance (diagonal in Fig. 3a), thus, the original ToxCast-AA model is valid (orange line in Fig. 3a). Also, high efficiency (ratio of SCPs) of 0.87 is achieved at SL 0.2. Since evaluation at SL 0.2 is commonly used in literature, the values will also be given when describing the further validation process. Furthermore, the performance of the ToxCast-AA model is in

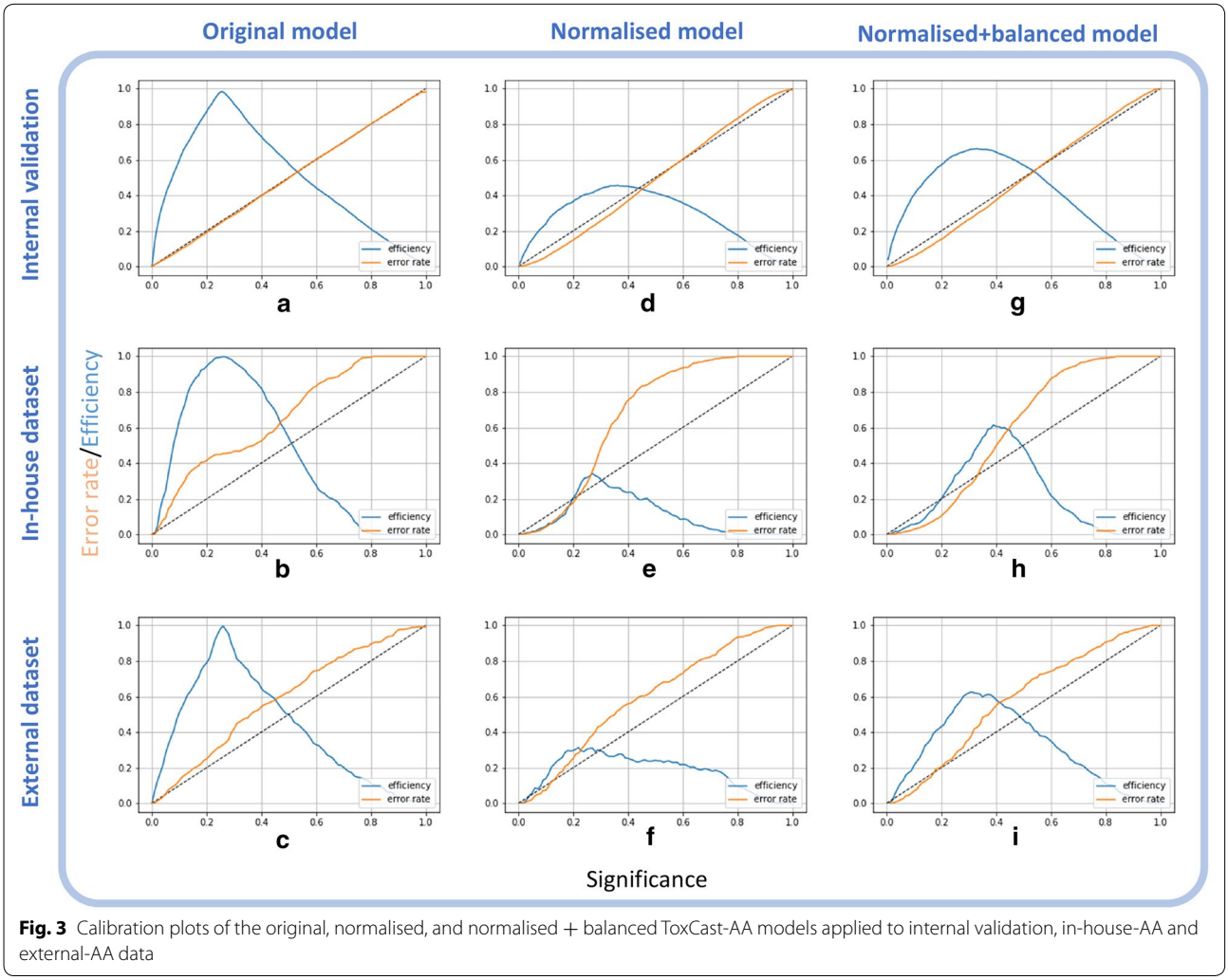


Fig. 3 Calibration plots of the original, normalised, and normalised + balanced ToxCast-AA models applied to internal validation, in-house-AA and external-AA data

Table 3 Comparison of original conformal prediction model for androgen receptor antagonism at 0.2 SL with other studies from literature

Model	Validity			Efficiency	Accuracy	
	All	Class 1[c]	Class 0[c]	All	Class 1[c]	Class 0[c]
KnowTox-AA	0.81	0.82	0.81	0.87	0.80	0.78
eMolTox [35][a]	–	0.76–0.81	0.81–0.82	0.94–0.99	–	–
Norinder et al. [33][b]	0.80–0.81	0.81–0.83	0.79–0.82	–	0.79–0.82	0.78–0.79

[a] Values of models fitted on two different AA datasets.

[b] Three models with different fingerprints trained on one AA dataset

[c] class 1 = actives, class 0 = inactives

Table 4 Information on KnowTox-AA and other CP methods using the random forest ML algorithm to predict androgen receptor antagonism

Method	Data source: actives/inactives	CP aggregation method[a]	Descriptors
KnowTox-AA	ToxCast: 868/5842	ACP	Morgan+MACCS (+physchem)
eMolTox [35][b]	Literature:[d] (1) 532/6207 (2) 406/6256	ACP	Morgan + physchem
Norinder et al. [33][c]	Jensen et al. [42]: 293/637	CCP	(1) Dragon (2) Signatures (3) Physchem

[a] *ACP* aggregated conformal predictor, *CCP* cross-conformal predictor [48]

[b] Two models ((1), (2)) fitted on two different AA datasets

[c] Three models ((1), (2), (3)) with different fingerprints trained on one AA dataset

[d] Data for a total of 174 CP models originated from ChEMBL, Pubchem, Toxnet, eChemPortal databases and literature [35]

line with two other AA models extracted from literature (see Table 3), i.e. the eMolTox webserver [35] and work by Norinder et al. [33]. Validity, efficiency, and accuracy values for all three studies (if reported) at SL 0.2 are in the range of 0.76–0.83, 0.87–0.99, and 0.79–0.82, respectively. Although the above described AA models all use CP, they are only partly directly comparable as underlying data, techniques and/or features differ (see Table 4).

Note that some other QSAR models for AA have been published, based on similar data, using random forest, deep learning [53], and the Case Ultra system [54]. Since set-up and reported performance measures differ from this CP study, they can not directly be compared. Very recently, CoMPARA, an extensive study on androgen receptor modelling, was published by Mansouri et al. [55]. Scientists from 25 research groups have contributed to consensus models for androgen receptor binding, agonism, and antagonism with a predictive accuracy of 78% for the AA evaluation set (which is in the same range as the CP accuracy (SCP) obtained for the original KnowTox-AA model, see Table 3). The individual AA models were trained on 1525 ToxCast chemicals using, amongst others, neural networks as well as tree-based and linear modelling approaches.

When applying the original ToxCast-AA model to the libraries of in-house (Fig. 3b) and external molecules (Fig. 3c), validity at 0.2 SL dropped from 0.81 for the internal validation to 0.59 for the in-house dataset. Furthermore, a high discrepancy was observed between the ratio of correct predictions of the active (0.98) and inactive (0.16) class for the in-house data (see Additional file 1: Table S1). Reasons for lower validity could be lacking exchangeability between the compounds of the datasets (pharmaceuticals vs. industrial chemicals) and data originating from different assays.

Hence, the chemical space was analysed with respect to 1) the most similar compounds and 2) the descriptor space using PCA. First, the average Tanimoto similarity to the ten most similar molecules in ToxCast decreases from 0.51 for intra ToxCast similarity to 0.44 for external data and 0.37 for in-house data. Second, the PCA (Fig. 4) reveals that the in-house data (blue dots) shows the highest density in the lower right corner, which is different from the dense area of the ToxCast data (red dots). The external dataset (grey dots) is more similar to the ToxCast distribution, occupying a dense area in the middle of the plot. Varying distribution and density contribute to poor exchangeability between the different datasets.

To improve reliability of the models, the chemical space was considered by including information about the nearest neighbours to normalise the conformal predictions. While such a normalisation of the nc scores is important for regression models [56, 57], to the best of our knowledge, it has not been applied to classification tasks so far. Including the KNN normalisation clearly improved validity for internal validation and the in-house dataset from 0.81 to 0.85 and from 0.59 to 0.82 at 0.2 SL, respectively (see Additional file 1: Table S2). Figure 3d,e show the lower error rate at a higher confidence area (small SLs), but decreased efficiency. Improved validity comes with the cost that less SCPs are made by the model, i.e. efficiency of 0.37 for ToxCast-AA and 0.21 for in-house-AA at 0.2 SL. From an application point of view, this is acceptable, since it is preferred to make no prediction rather than a wrong assertion.

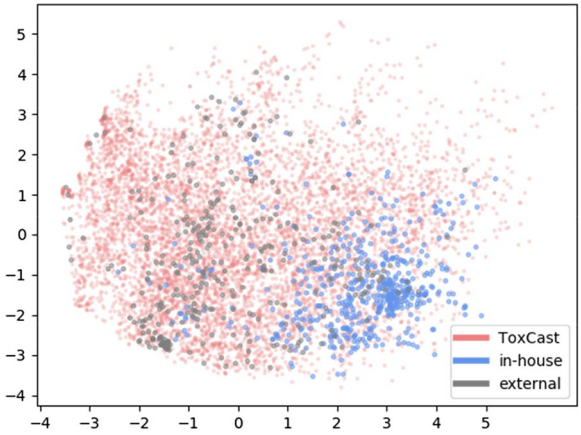


Fig. 4 ToxCast, in-house and external data are projected into the descriptor space of a 2-component PCA trained on ToxCast-AA data

Nevertheless, still, a high discrepancy between the accuracy of the active and inactive classes can be observed, with the highest discrepancy of 0.54 for the prediction of the external-AA data (see Additional file 1: Table S2). This is due to the high imbalance in the training data with a ratio of 1 active to 6.7 inactives in which the KNN algorithm is searching for nearest neighbours. While balancing in a mondrian ACP setting is normally not necessary, in the case of the additional KNN normalisation, random equal size sampling of the proper training and the calibration set, clearly reduced the discrepancy between the two classes for accuracy, as well as efficiency (see Table 5, Fig. 3g–i).

The following factors should be noted regarding model performance: Firstly, the refined, normalised + balanced conformal predictors have been validated for use at low SLs. They are valid on the in-house dataset at SLs below 0.3, on the external dataset below 0.2. Therefore, predictions for the case study compounds are based on SL 0.2. As there is no interest in predictions with high error rates, the low validity at higher SLs can be ignored. Secondly, the three datasets all originate from different assays (i.e. performed in human, yeast, and hamster cells; a human androgen receptor was expressed in both human and yeast cells). Due to a limited amount of available toxicity data, it is inevitable to compare data from different organisms, nevertheless, caution should be exercised.

The knowledge gained from creating the normalised + balanced model was applied to the remaining endpoints of the ToxCast dataset. Using the validated strategy, totally, 88 models were built with overall validity between 0.81 and 0.86, and overall efficiency between 0.32 and 0.68 at SL 0.2 (see Fig. 5, top). The accuracy of single class predictions ranged from 0.65 to 0.95 (see Fig. 5, bottom). Numbers for all 88 models and information about the endpoints can be found in Additional file 1: Tables S4 and S5, respectively.

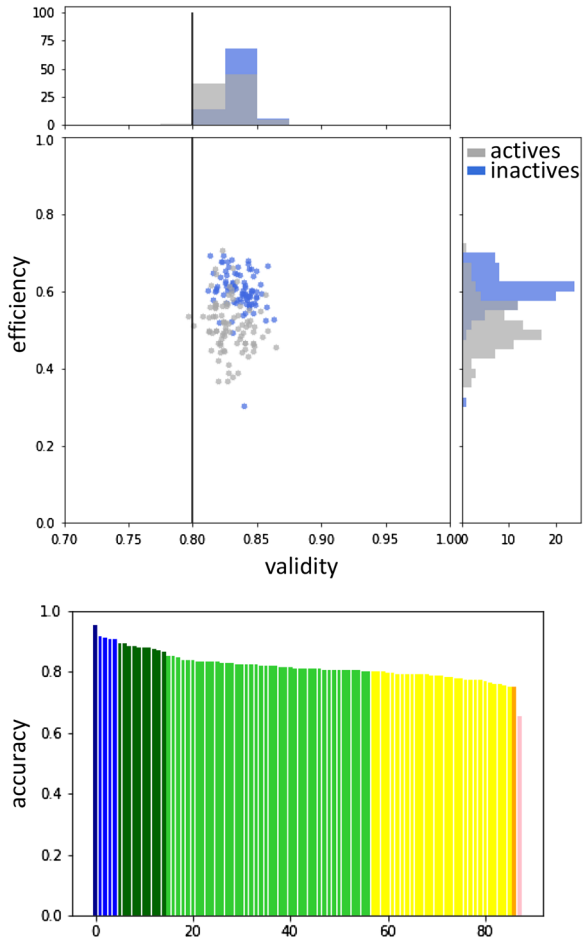


Fig. 5 Evaluation of final 88 CP models. Top: validity vs. efficiency for inactives (blue) and actives (grey). Bottom: sorted, overall accuracy

KnowTox—case study

If KnowTox is queried with a compound of interest, three modules are envoked: conformal prediction (CP) for 88 endpoints, screening for unfavourable structural moieties, and support for read-across from similar compounds (see Fig. 6).

In this section, KnowTox usage is exemplified on two *triazoles* from the in-house dataset. They were designed as potential fungicides, but discontinued for various reasons. Both molecules share an epoxide structure with two halogenated phenyl moieties and a triazole ring with a thioether substitute (see Fig. 6a).

Table 5 Evaluation of normalised + balanced[a] conformal prediction model for androgen receptor antagonism at 0.2 SL

Dataset	Purpose	Validity			Efficiency			Accuracy		
		All	cl.1[b]	cl.0[b]	All	cl.1[b]	cl.0[b]	All	cl.1[b]	cl.0[b]
ToxCast-AA	train model	0.85	0.84	0.85	0.57	0.39	0.60	0.89	0.76	0.91
In-house-AA	validation I	0.90	0.90	0.89	0.20	0.18	0.23	0.75	0.80	0.71
External-AA	validation II	0.80	0.76	0.82	0.43	0.33	0.52	0.74	0.67	0.78

[a] normalised nc score and balancing of calibration and proper training set

[b] cl.: class (class 1 = actives, class 0 = inactives)

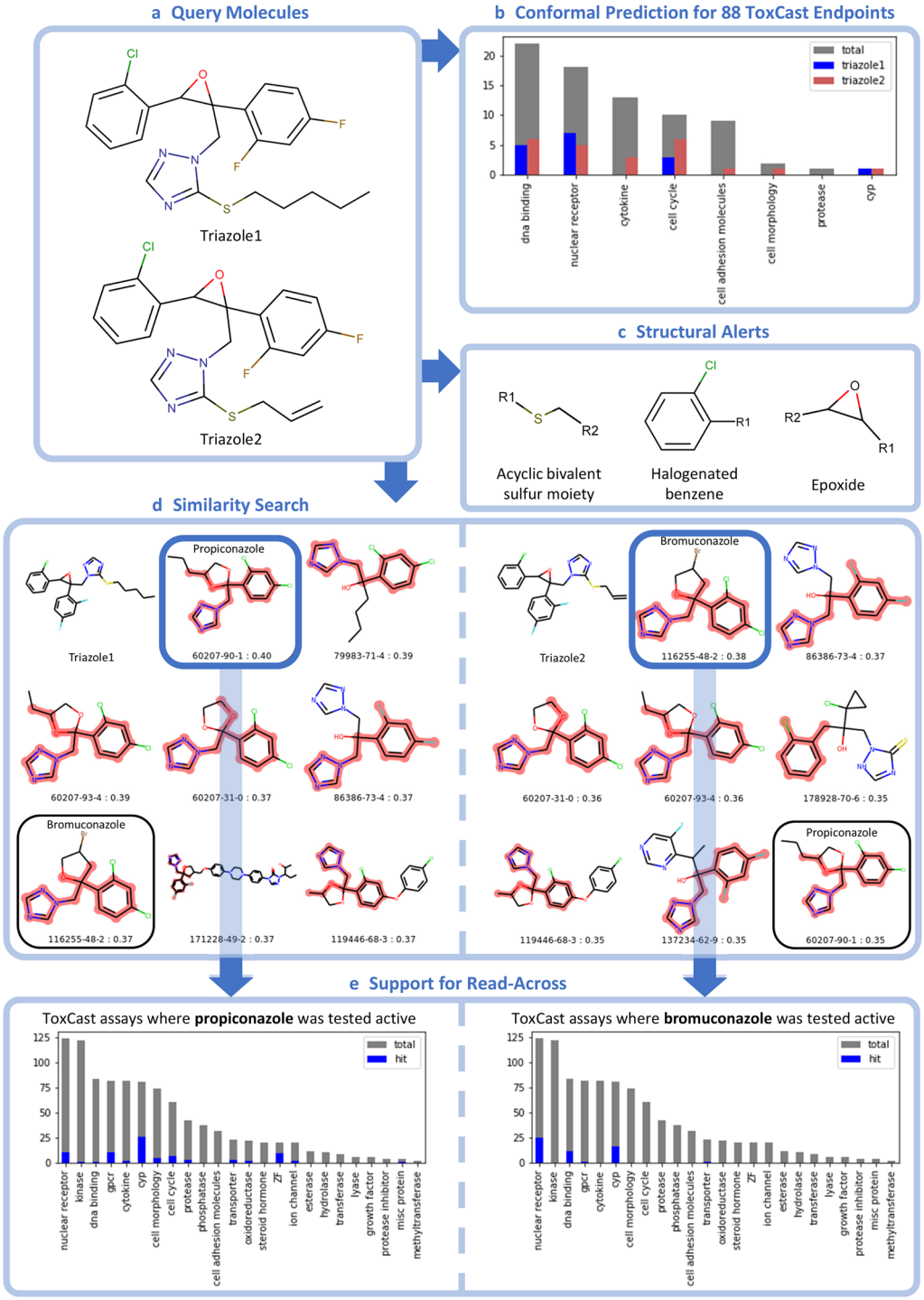


Fig. 6 KnowTox tool applied in a case study. **a** *Triazoles1&2* used as query compounds for the case study. **b** Output of CP. Grey: number of endpoints per family available for CP. Red and blue: number of endpoints where *triazoles1&2* were predicted to be active (SCP) at SL 0.2. **c** Three selected toxic alerts found for *triazoles1&2*. (Note that the potentially critical "triazole" substructure is not considered in this work). **d** *Triazole1* (left) and *triazole2* (right) and their most similar molecules in ToxCast including CAS number and Tanimoto similarity. Red: maximum common substructure. **e** Experimental information from ToxCast for propiconazole (left) and bromuconazole (right). Grey: available assays in ToxCast. Blue: assays where compound was tested active

Firstly, a conformal prediction with every of the above described 88 models is made. Each model returns two p-values, one for the inactive (p0) and one for the active (p1) class. The higher p-value denotes the class the compound is most likely assigned to. For example, the ToxCast-AA model predicts *triazole1* to be active with p-values $p0 = 0.19$ and $p1 = 0.56$. In literature, CPs are often evaluated at a specified maximum accepted error rate (equivalent to SL ϵ). For instance, if no more than 20% errors are accepted (SL = 0.2), the result is a prediction set containing all labels with p-values above 0.2. Thus, *triazole1* is predicted AA ({1}) while *triazole2* ($p0 = 0.21$, $p1 = 0.60$ for AA prediction) is assigned both labels ({0,1}). Therefore, no decision is made for *triazole2*. However, if 25% errors would be allowed (SL = 0.25), *triazole2* would also be predicted to be AA only ({1}).

Alternatively, evaluation can be independent from a predefined SL, i.e. with respect to credibility and confidence [28]. Credibility is defined as the largest p-value, this means the highest SL where a compound is still assigned to the corresponding label. Confidence is defined as 1−second largest p-value; since a high p-value of an alternate class reduces the confidence in the prediction. *Triazole1* is predicted to be AA with credibility = 0.56 and confidence = 0.81.

Referring to the three domains concept by Hanser et al. [27] (applicability, reliability, decidability), mentioned in the introduction, higher p-values, indicate higher reliability of a prediction while a large difference between the two p-values corresponds to increased decidability.

Considering the predictions by all 88 CP models (see Fig. 6b), both triazoles were predicted to be only active (SL 0.2) at a total of 15 endpoints, related to DNA binding, nuclear receptors, cell cycle as well as for aromatase inhibition (CYP19A1). A full list of the p-values for the predictions can be found in Additional file 1: Table S3.

Potential interaction of triazoles with aromatase can be explained through the mode of action of triazole fungicides. They inhibit the biosynthesis of ergosterol—an essential component of fungal cell membranes—changing the composition of the cell membrane. More precisely, the fungal enzyme lanosterol 14α-demethylase (CYP51) is inhibited which is closely related to human CYP15 and CYP19 (aromatase). Homology of fungal CYP51 to human CYP19 suggests likewise effects on steroidogenesis in humans [58]. Aromatase is responsible for catalysing the transformation of androgens into estrogens [59]. Inhibition can have a severe impact on hormone levels, though the actual physiological effects remain unclear [60, 61].

Besides, both *triazoles* were predicted to induce transcription factor activity and, thus, elevate the level of pregnane X receptor (PXR) response element and phenobarbital-responsive enhancer module mRNA. The two response elements are bound by members of the endogeneous human nuclear receptor subfamily 1 (PXR and constitutive androstane receptor (CAR), respectively), and are involved in overlapping pathways of xenobiotic detoxification, mainly occuring in the liver [62]. PXR is responsible for the expression of xenobiotic metabolising enzymes (e.g. cytochromes) in humans and is activated by a wide range of xenobiotics (e.g antibiotics) as well as endobiotics [63]. Activation of PXR has previously been observed by other azole fungicides such as miconazole and propiconazole [64] (see Fig. 7a, b). Moreoever, many conazoles are known to be involved in inhibition and induction of mammalian cytochromes P450 [65]. Generally, metabolism and elimination of foreign substances, such as fungicides, is favourable, it is mainly alarming when it comes to drug-drug interactions [66] (e.g. induction of xenobiotic metabolism by one drug may also affects metabolism and thus plasma levels of another drug). An example is the antimycotic drug ketoconazole (Fig. 7c) which is preferably applied topically rather than orally due to its high drug-drug interaction potential [67].

Triazoles1&2 were both predicted to have antagonistic effects on the thyroid receptor. Indeed, thyroid endocrine effects of triazole fungizides have not yet extensively been

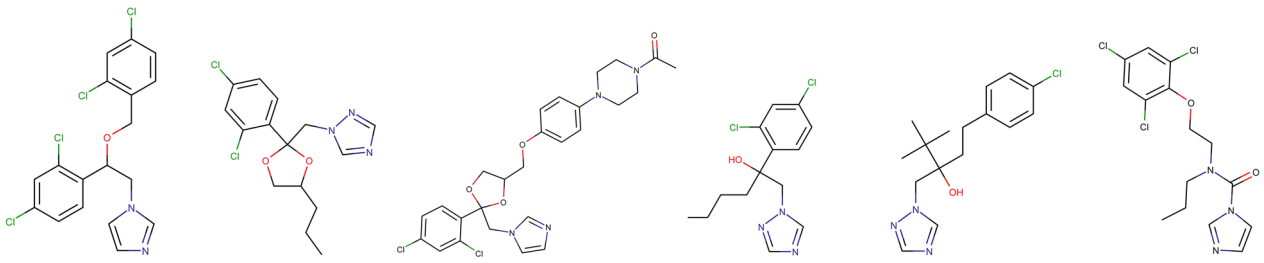


Fig. 7 Chemical structures of triazole and imidazole fungicides referred to in the case study section

studied: There is no indication of an effect in mammalians in vivo and only few reports in vitro and in zebrafish: Thyroid endocrine effects have previously been reported for two triazole fungicides hexaconazole and tebuconazole (see Fig. 7d, e) in zebrafish larvae [68]. Yu et al. suggested that the latter two triazoles can influence both, thyroid hormone levels and gene transcription in the hypothalamic-pituitary-thyroid axis. Changing thyroid hormone levels can affect several important physiological processes, e.g. tissue growth and differentiation, energy homeostasis, and metabolism [69, 70].

Furthermore, *triazoles1&2* were predicted to interfere with the cell cycle, i.e. leading to cytotoxicity. Also, in literature, evidence for cytotoxicity and cell cycle inhibition by triazole fungicides or mixtures containing such is given. For instance, Schwarzbacherova et al. reported cytotoxic and genotoxic effects, such as reduced cell viability, decreased cell proliferation, and apoptosis of bovine lymphocytes induced by fungicides [71]. In another study, they found bovine lymphocytes proliferation inhibited by a mixture of two conazole fungicides [72]. Additionally, Zhou et al. [73] described apoptotic effects of tebuconazole (see Fig. 7e) on human placental trophoblast cells.

Summarising, it could be shown that the CP models make reasonable predictions for potential toxic effects of these compounds, which could be substantiated with evidence in literature.

Secondly, with a search for structural alerts, toxicity prediction is supported with information from literature about substructures that have been previously assigned to specific toxic endpoints. Each query compound is screened against totally 919 available alerts and any critical substructure is highlighted.

Three alerts found for *triazoles1&2* are shown in Fig. 6c. According to Benigni et al. halogenated benzenes are prone to non-genotoxic carcinogenicity via agonistic or antagonistic interaction with the aryl hydrocarbon receptor (AhR) [74]. AhR activation can result in altered gene expression and thus various types of toxicity, e.g. immunotoxicity, liver tumor promotion, and carcinogenicity [74, 75].

The sulfur moiety points to a study by Liu et al. [40], where 23 drugs containing acyclic bivalent sulfur moieties were investigated. Eight out of them are known for liver toxicity, another 14 are possibly hepatotoxic. Since only for one of the investigated drugs, liver toxicity could be excluded certainly, potential liver toxicity should be considered for these moieties. Conversely, this alert must not be an exclusion criterion, as the above drugs were still launched to the market.

Another warning is issued towards the epoxide substructure, a highly reactive group. Presence of the oxygen makes the carbons in the three-membered ring electrophilic. Thus they are typically accessed by nucleophiles, via an S_N2-type mechanism resulting in ring opening and a covalent bond. This may cause mutagenic or carcinogenic effects, as well as skin sensitization and aquatic toxicity [39, 76–78]. While the nucleophile preferentially attacks the less substituted ring carbon, [76] in the case of *triazoles1&2*, access to any ring carbon is sterically hindered due to the three surrounding substituents. Thus, the present epoxides can be considered inert.

Note that the issued warnings are based on the 919 toxic alerts incorporated into KnowTox. If the collection of structural alerts is desired to be even more comprehensive, it can always be extended by literature or in-house knowledge. For example, the triazole substructure, which is also included in the ToxAlerts tool as an "extended functional group" [19, 79], is not considered in this work. As seen in the CP part, this moiety is responsible for both, the antifungal activity, and adverse effects due to aromatase inhibition.

Thirdly, risk assessment is complemented through inclusion of information from experimental ToxCast assay outcomes of similar molecules. For a query compound, the 7912 compounds of the ToxCast dataset are screened to identify the molecules with highest Tanimoto similarity and toxicity information of these most similar molecules is displayed. To simplify the assessment of the grade of similarity, and thus the reliability in the read-across, the Tanimoto index, as well as the MCS between the molecules are indicated. Similarity search and support for read-across can especially be valuable for those endpoints where minority class data was too few to build a CP model.

When querying the triazoles in the similarity search, eight fairly similar molecules are returned (see Fig. 6d). The similarity is mainly reflected in the triazole substructure and halogenated benzenes, mostly connected in three- or four-membered ring-systems. Note that no other molecule with an epoxide substructure is captured within the similarity search.

Assuming that the found molecules are similar enough, known experimental information about them could be used to support read-across. Although ToxCast provides data from 985 assays, the most similar molecules to the two triazoles were only assayed for 32 to 639 endpoints each. The most similar molecule to *triazole1*, propiconazole, was, amongst others, tested active at several nuclear receptor-related endpoints (e.g. PXR*e*, CAR, androgen, thyroid and estrogen receptors), cytochromes P450 (i.e. 19, 1a, 2b, 2c, 2d, 3a), and GPCRs (e.g. opioid receptors, muscarinic cholinergic receptors, and histamine receptor H2). Furthermore, it had effects on several developmental endpoints of zebrafish embryos [80] (see Fig. 6e). Experimentally observed activity for bromuconazole, the most

similar compound to *triazole2*, was mainly restricted to nuclear receptors (e.g. retinoic acid, androgen, and estrogen receptors, PXR), DNA binding (AhR, p53, sterol regulatory element binding protein), and cytochromes P450 (19A1, 2a, 2b, 2c, 3a). It should, moreover, be noted that Br-substituents, as in bromuconazole, are generally more reactive than F- or Cl-substituents [39]. So, certain toxic effects might be more distinct in bromuconazole than in molecules without Br- substituted moieties, such as *triazoles1&2*.

The toxic effects described for *triazoles1&2* above can be related to pathways, such as CAR/RXR and PXR/RXR activation, xenobiotic metabolism signaling, and AhR signaling, which were also investigated in a study by Hester et al. [65] and related to hepatocarcinogenesis.

An association of bromuconazole with xenobiotic metabolism and nuclear receptors (i.e. PXR), as suggested by the similarity-based read-across, is further supported by a recent study by Abdelhadya et al. [81]. They reported, inter alia, that the liver oxidative damage is associated with increased PXR activity and concurrent decrease in expression of the CAR gene.

In conclusion, indications of liver toxicity, liver enzyme induction, and aromatase inhibition were found in rats treated with these two triazoles in in-house studies. Thus, further development of these two triazole candidates was discontinued. Also, according to literature, several conazole fungicides have been associated to potential AA endocrine disruption [82]. For example, AA effects were reported for prochloraz (Fig. 7f) in human prostate cancer cells [83]. Also, propiconazole (Fig. 7b) showed AA activity in vitro, though it could not be asserted in vivo [84]. Moreover, another explanation for triazole-induced liver toxicity was recently provided by Knebel et al. who investigated molecular mechanisms of hepatic steatosis [85]. The triazole fungicides propiconazole and tebuconazole (see Fig. 7b, e) were shown to influence the expression of steatosis-related genes. Especially, the observed additivity of equimolar mixtures suggests a common mode of action.

To conclude, KnowTox was able to predict many interactions, especially with respect to the induction of xenobiotic enzymes, endocrine effects, and liver toxicity. The discussed predictions could be supported by literature findings for other related molecules. Also, the KnowTox tool could reproduce the main in vivo effects of two *triazole* compounds, which have been discontinued as development candidates.

To sum up, such a holistic analysis of the toxic potential of a novel molecule can be of high reward in compound (de-)selection, planning further toxicity testing, and to support read-across. Nevertheless, its benefit can still be increased by incorporation of larger datasets, biological activity fingerprints characterising the compounds, and in vivo endpoint data for model development. Note that KnowTox is based on the ToxCast dataset chosen for its size, scope and accessibility. Used in early stages of new chemical's development, the tool can provide a broad overview on possible interactions with toxicity-related targets. For application in regulatory toxicity testing, it is beneficial to have toxicity data which fundamentally support regulatory required toxicity assays in animals, e.g. reproduction toxicity studies. In case of occurrence of toxic effects, the tool will help to identify a potential mode-of-action. In addition, it will increase certainty if data support the absence of toxic effects. Thus, if, in future, sufficient standard toxicity data will be available for model training, the introduced pipeline has the potential to become even more powerful. Also, information about the compound's bioavailability and in vitro to in vivo translation of the assays would be of high interest [10, 86–88]. According to Grenet et al. [87], it seems to be more challenging to predict long-term in vivo endocrine disruption, compared to predicting short-term in vivo endocrine effects. Furthermore, for a complete risk assessment, the quantitative dose-response needs to be considered. That is beyond the scope of this paper. Information on the type and amount of formed metabolites is highly desirable (see the prominent role of xenobiotic metabolism in the toxic effects of triazole fungicides).

In vitro toxicology has embarked on combining data from different sources to derive more reliable and more relevant information on potential toxic effects of compounds [89, 90]. This concept also applies to in silico toxicology and combinations of the different in vivo, in vitro, in silico methods: combining the input from different, complementary models can provide advantageous information which cannot be obtained from one single source.

Conclusion

In silico methods for toxicity prediction are promising tools assisting in the reduction and replacement of animal testing. In this work, three different approaches were combined in order to support holistic risk assessment for new query molecules.

In praxis, it is not only important to have well performing models, but also to know that they can be confidently applied to novel compounds (applicability domain), that the predictions are reliable (reliability domain) and informative (decidability domain). A popular technique for confidence estimation for machine learning models is conformal prediction, which enables straightforward training of valid and balanced models with little optimisation effort. While this advantage was also witnessed during internal validation, in this work, some challenges emerged during application to an

external dataset where exchangeability was not given. Therefore, the models were refined in two steps: firstly, using k-nearest neighbour normalisation improved validity of both internal and in-house data predictions (reliability domain). Secondly, random equal size sampling of the training set improved informational efficiency of the predictions (decidability domain). This strategy was initially validated on an AA model and subsequently transferred to totally 88 ToxCast endpoint models. Complemented with structural alerts from literature and providing support for read-across, the KnowTox tool generates a risk assessment picture to examine potential toxicity of a novel query compound from different angles as exemplified by the case study on two triazoles.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s13321-020-00422-x.

Additional file 1. Additional Tables S1–S5.

Abbreviations

ML: Machine learning; MCS: Maximum common substructure; QSAR: Quantitative structure-activity relationship; PCA: Principal component analysis; ACP: Aggregated conformal predictor; SL: Significance level; CP: Conformal prediction; nc score: Nonconformity score; AA: Androgen receptor antagonism; SCP: Single class prediction; PXR: Pregnane X receptor; AhR: Aryl hydrocarbon receptor; CAR: Constitutive androstane receptor.

Acknowledgements

All authors thank Peter Geyer, Maike Huisinga, Christian Pilger, Saskia Sperber, and Volker Strauss for collaboration. AM and AV thank Fredrik Svensson, Ulf Norinder, and Andreas Bender for helpful discussions about conformal prediction. Furthermore, the authors thank Changge Ji and Andreas Bender for sharing the list of toxic alerts from their eMolTox webserver. AM and AV thank Jaime Rodríguez-Guerra for reviewing the github repository with the supplementary notebook. AM thanks the HPC Service of ZEDAT, Freie Universität Berlin, for computing time.

Authors' contributions

AM conducted the study and wrote the paper under close supervision of AV and in intense collaboration with all project partners: MM, JHA, RB, AW, KJS and RL. AM, AV, MM and JHA contributed to the set-up of the computational framework. MM and JHA pre-processed the ToxCast as well as the in-house datasets. RB and RL provided the in-house dataset and knowledge about the case study compounds. AW, KJS, RL and AV designed the study. All authors revised the final manuscript. All authors read and approved the final manuscript.

Funding

AM and AV thank BMBF (Grant No. 031A262C) and the HaVo-Stiftung for funding. Furthermore, the authors have received internal BASF funding (creativity project "*KnowTox*"). The authors acknowledge support from the German Research Foundation (DFG) and the Open Access Publication Fund of Charité – Universitätsmedizin Berlin.

Availability of data and materials

ToxCast data was used for model training and is publicly available at https://figshare.com/articles/ToxCast_and_Tox21_Data_Spreadsheet/6062503 [41]. For in depth evaluation and applicability optimization of the ToxCast-AA model the external-AA and in-house-AA datasets were used. External-AA data are available from Norinder et al. [33]. The 534 in-house-AA data are proprietary to BASF SE. The structures of the two BASF SE case study molecules are shown in this manuscript. The pre-processed ToxCast and External-AA data, as well as a notebook demonstrating the process of training and evaluating conformal prediction models, based on this manuscript's methods, is available under https://github.com/volkamerlab/knowtox_manuscript_SI.

Competing interests

The authors declare that they have no competing interests. MM, JHA, AW, KJS, RB and RL are employees of BASF SE, a company developing and marketing fungicides.

Author details

[1] In Silico Toxicology and Structural Bioinformatics, Institute of Physiology, Charité Universitätsmedizin Berlin, Charitéplatz 1, Berlin, Germany. [2] BASF SE, 67056 Ludwigshafen, Germany.

Received: 18 December 2019 Accepted: 9 March 2020
Published online: 14 April 2020

References

1. ECHA (2007) REACH. https://echa.europa.eu/regulations/reach/understanding-reach. Accessed 5 Apr 2019
2. BMEL (2018) Versuchstierdaten 2017. https://www.bmel.de/DE/Tier/Tierschutz/_texte/Versuchstierzahlen2017.html. Accessed 24 Mar 2019
3. Thomford NE, Senthebane DA, Rowe A, Munro D, Seele P, Maroyi A, Dzobo K (2018) Natural products for drug discovery in the 21st century: innovations for novel drug discovery. Int J Mol Sci. https://doi.org/10.3390/ijms19061578
4. Kavlock RJ, Ankley G, Blancato J, Breen M, Conolly R, Dix D, Houck K, Hubal E, Judson R, Rabinowitz J, Richard A, Setzer RW, Shah I, Villeneuve D, Weber E (2008) Reviews: computational toxicology—a state of the science mini review. Toxicol Sci 103(1):14–27. https://doi.org/10.1093/toxsci/kfm297
5. Yang H, Sun L, Li W, Liu G, Tang Y (2018) In silico prediction of chemical toxicity for drug design using machine learning methods and structural alerts. Front Chem 6:30. https://doi.org/10.3389/fchem.2018.00030
6. Richard AM, Judson RS, Houck KA, Grulke CM, Volarath P, Thillainadarajah I, Yang C, Rathman J, Martin MT, Wambaugh JF, Knudsen TB, Kancherla J, Mansouri K, Patlewicz G, Williams AJ, Little SB, Crofton KM, Thomas RS (2016) ToxCast chemical landscape: paving the road to 21st century toxicology. Chem Res Toxicol 29(8):1225–1251. https://doi.org/10.1021/acs.chemrestox.6b00135
7. Gadaleta D, Manganelli S, Roncaglioni A, Toma C, Benfenati E, Mombelli E (2018) QSAR modeling of ToxCast assays relevant to the molecular initiating events of AOPs leading to hepatic steatosis. J Chem Inform Model 58(8):1501–1517. https://doi.org/10.1021/acs.jcim.8b00297
8. Liu J, Mansouri K, Judson RS, Martin MT, Hong H, Chen M, Xu X, Thomas RS, Shah I (2015) Predicting hepatotoxicity using ToxCast in vitro bioactivity and chemical structure. Chem Res Toxicol 28(4):738–751. https://doi.org/10.1021/tx500501h
9. Bhhatarai B, Wilson DM, Price PS, Marty S, Parks AK, Carney E (2016) Evaluation of OASIS QSAR models using Toxcast™ in vitro estrogen and androgen receptor binding data and application in an integrated endocrine screening approach. Environ Health Perspect 124(9):1453–1461. https://doi.org/10.1289/EHP184
10. Liu J, Patlewicz G, Williams AJ, Thomas RS, Shah I (2017) Predicting organ toxicity using in vitro bioactivity data and chemical structure. Chem Res Toxicol 30(11):2046–2059. https://doi.org/10.1021/acs.chemrestox.7b00084
11. Bell SM, Angrish MM, Wood CE, Edwards SW (2016) Integrating publicly available data to generate computationally predicted adverse outcome pathways for fatty liver. Toxicol Sci 150(2):510–520. https://doi.org/10.1093/toxsci/kfw017
12. Zhu H, Bouhifd M, Kleinstreuer N, Kroese ED, Liu Z, Luechtefeld T, Pamies D, Shen J, Strauss V, Wu S, Hartung T (2016) Supporting read-across using biological data. ALTEX 1848(2):3047–3054. https://doi.org/10.1016/j.bbamem.2015.02.010.Cationic

13. Chushak YG, Shows HW, Gearhart JM, Pangburn HA (2018) In silico identification of protein targets for chemical neurotoxins using ToxCast in vitro data and read-across within the QSAR toolbox. Toxicol Res 7(3):423–431. https://doi.org/10.1039/c7tx00268h
14. Shah I, Liu J, Judson RS, Thomas RS, Patlewicz G (2016) Systematically evaluating read-across prediction and performance using a local validity approach characterized by chemical structure and bioactivity information. Regul Toxicol Pharmacol 79:12–24. https://doi.org/10.1016/J.YRTPH.2016.05.008
15. Teubner W, Landsiedel R (2015) Read-across for hazard assessment: the ugly duckling is growing up. Altern Lab Anim 43(6):P67–P71. https://doi.org/10.1177/026119291504300617
16. van Ravenzwaay B, Sperber S, Lemke O, Fabian E, Faulhammer F, Kamp H, Mellert W, Strauss V, Strigun A, Peter E, Spitzer M, Walk T (2016) Metabolomics as read-across tool: a case study with phenoxy herbicides. Regul Toxicol Pharmacol 81:288–304. https://doi.org/10.1016/j.yrtph.2016.09.013
17. Sushko I, Salmina E, Potemkin VA, Poda G, Tetko IV (2012) ToxAlerts: a web server of structural alerts for toxic chemicals and compounds with potential adverse reactions. J Chem Inform Model 52(8):2310–2316. https://doi.org/10.1021/ci300245q
18. Brenk R, Schipani A, James D, Krasowski A, Gilbert IH, Frearson J, Wyatt PG (2008) Lessons learnt from assembling screening libraries for drug discovery for neglected diseases. ChemMedChem 3(3):435–444. https://doi.org/10.1002/cmdc.200700139
19. OCHEM (2012) ToxAlerts. www.ochem.eu/alerts. Accessed 8 Nov 2019
20. Huang R, Xia M, Nguyen DT, Zhao T, Sakamuru S, Zhao J, Shahane SA, Rossoshek A, Simeonov A (2017) 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. https://doi.org/10.3389/978-2-88945-197-5
21. Banerjee P, Siramshetty VB, Drwal MN, Preissner R (2016) Computational methods for prediction of in vitro effects of new chemical structures. J Cheminform 8(1):1–11. https://doi.org/10.1186/s13321-016-0162-2
22. Mayr A, Klambauer G, Unterthiner T, Hochreiter S (2016) DeepTox: toxicity prediction using deep learning. Front Environ Sci 3:80. https://doi.org/10.3389/fenvs.2015.00080
23. Banerjee P, Eckert AO, Schrey AK, Preissner R (2018) ProTox-II: a webserver for the prediction of toxicity of chemicals. Nucleic Acids Res. https://doi.org/10.1093/nar/gky318
24. Accelrys (2015) TOPKAT. https://omictools.com/topkat-tool
25. Greene N, Judson PN, Langowski JJ, Marchant CA (1999) Knowledge-based expert systems for toxicity and metabolism prediction: DEREK, StAR and METEOR. SAR QSAR Environ Res 10(2–3):299–314. https://doi.org/10.1080/10629369908039182
26. Fuart-Gatnik M, Worth AP (2010) Review of software tools for toxicity prediction. JRC Sci Tech Rep. https://doi.org/10.2788/60101
27. Hanser T, Barber C, Marchaland JF, Werner S (2016) Applicability domain: towards a more formal definition. SAR QSAR Environ Res 27(11):865–881. https://doi.org/10.1080/1062936X.2016.1250229
28. Mathea M, Klingspohn W, Baumann K (2016) Chemoinformatic classification methods and their applicability domain. Mol Inform 35(5):160–180. https://doi.org/10.1002/minf.201501019
29. Eklund M, Norinder U, Boyer S, Carlsson L (2015) The application of conformal prediction to the drug discovery process. Ann Math Artif Intell 74(1–2):117–132. https://doi.org/10.1007/s10472-013-9378-2
30. Norinder U, Carlsson L, Boyer S, Eklund M (2014) Introducing conformal prediction in predictive modeling for regulatory purposes. A transparent and flexible alternative to applicability domain determination. Regul Toxicol Pharmacol 71(2):279–284. https://doi.org/10.1016/j.yrtph.2014.12.021
31. Svensson F, Norinder U, Bender A (2017a) Improving screening efficiency through iterative screening using docking and conformal prediction. J Chem Inform Model 57(3):439–444. https://doi.org/10.1021/acs.jcim.6b00532
32. Svensson F, Norinder U, Bender A (2017b) Modelling compound cytotoxicity using conformal prediction and PubChem HTS data. Toxicol Res 6(1):73–80. https://doi.org/10.1039/C6TX00252H
33. Norinder U, Rybacka A, Andersson P (2016) Conformal prediction to define applicability domain—a case study on predicting ER and AR binding. SAR QSAR Environ Res 27(4):303–316. https://doi.org/10.1080/1062936X.2016.1172665
34. Lindh M, Karlén A, Norinder U (2017) Predicting the rate of skin penetration using an aggregated conformal prediction framework. Mol Pharm 14(5):1571–1576. https://doi.org/10.1021/acs.molpharmaceut.7b00007
35. Ji C, Svensson F, Zoufir A, Bender A (2018) eMolTox: prediction of molecular toxicity with confidence. Bioinformatics. https://doi.org/10.1093/bioinformatics/bty135
36. Rostkowski P, Horwood J, Shears JA, Lange A, Oladapo FO, Besselink HT, Tyler CR, Hill EM (2011) Bioassay-directed identification of novel antiandrogenic compounds in bile of fish exposed to wastewater effluents. Environ Sci Technol 45(24):10,660–10,667. https://doi.org/10.1021/es202966c
37. MacLusky NJ, Luine VN, Gerlach JL, Fischette C, Naftolin F, McEwen BS (1988) The role of androgen receptors in sexual differentiation of the brain: effects of the testicular feminization (Tfm) gene on androgen metabolism, binding, and action in the mouse. Psychobiology 16(4):381–397. https://doi.org/10.3758/BF03327335
38. Kolle S, Kamp H, Huener HA, Knickel J, Verlohner A, Woitkowiak C, Landsiedel R, van Ravenzwaay B (2010) In house validation of recombinant yeast estrogen and androgen receptor agonist and antagonist screening assays. Toxicol In Vitro 24(7):2030–2040. https://doi.org/10.1016/J.TIV.2010.08.008
39. Hermens JL (1990) Electrophiles and acute toxicity to fish. Environ Health Perspect 87:219–225. https://doi.org/10.1289/ehp.9087219
40. Liu R, Yu X, Wallqvist A (2015) Data-driven identification of structural alerts for mitigating the risk of drug-induced human liver injuries. J Cheminform 7(1):4. https://doi.org/10.1186/s13321-015-0053-y
41. EPA's National Center for Computational Toxicology (2019) ToxCast and Tox21 Data Spreadsheet. https://figshare.com/articles/ToxCast_and_Tox21_Data_Spreadsheet/6062503
42. Jensen GE, Niemelä JR, Wedebye EB, Nikolov NG (2008) QSAR models for reproductive toxicity and endocrine disruption in regulatory use—a preliminary investigation. SAR QSAR Environ Res 19(7–8):631–641. https://doi.org/10.1080/10629360802550473
43. Vinggaard AM, Niemelä J, Wedebye EB, Jensen GE (2008) Screening of 397 chemicals and development of a quantitative structure-activity relationship model for androgen receptor antagonism. Chem Res Toxicol 21(4):813–823. https://doi.org/10.1021/tx7002382
44. Atkinson FCGEE (2014) Standardiser. https://github.com/flatkinson/standardiser
45. Accelrys (2014) The Keys to Understanding MDL Keyset Technology. http://www.3dsbiovia.com/products/pdf/keys-to-keyset-technology.pdf
46. Rogers D, Hahn M (2010) Extended-connectivity fingerprints. J Chem Inform Model 50(5):742–754. https://doi.org/10.1021/ci100050t
47. Landrum GA (2018) RDKit: Open-source cheminformatics. http://www.rdkit.org
48. Sun J, Carlsson L, Ahlberg E, Norinder U, Engkvist O, Chen H (2017) Applying Mondrian cross-conformal prediction to estimate prediction confidence on large imbalanced bioactivity data sets. J Chem Inform Model 57(7):1591–1598. https://doi.org/10.1021/acs.jcim.7b00159
49. Carlsson L, Eklund M, Norinder U, Carlsson L, Eklund M, Norinder U, Conformal A, Lazaros P, Maglogiannis I, Papadopoulos H, Sioutas S, Ifip CM (2016) Aggregated Conformal Prediction To cite this version : Aggregated conformal prediction. In: IFIP advances in information and communication technology, pp 231–240
50. Linusson H, Norinder U, Boström H, Johansson U, Löfström T (2017) On the Calibration of aggregated conformal predictors. In: Proceedings of the sixth workshop on conformal and probabilistic prediction and applications, vol. 60, pp 154–173
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 (2011) Scikit-learn: machine learning in Python. J Mach Learn Res 12(Oct):2825–2830
52. Linusson H (2015) Nonconformist. http://donlnz.github.io/nonconformist/
53. Idakwo G, Thangapandian S, Luttrell J, Zhou Z, Zhang C, Gong P (2019) Deep learning-based structure-activity relationship modeling for multi-category toxicity classification: a case study of 10K Tox21 chemicals with high-throughput cell-based androgen receptor bioassay data. Front Physiol 10(August):1–13. https://doi.org/10.3389/fphys.2019.01044

54. Jensen GE (2012) QSAR model for androgen receptor antagonism–data from CHO cell reporter gene assays. J Steroids Hormonal Sci. https://doi.org/10.4172/2157-7536.s2-006
55. Mansouri K, Kleinstreuer N, Abdelaziz AM, Alberga D, Alves VM, Andersson PL, Andrade CH, Bai F, Balabin I, Ballabio D, Benfenati E, Bhhatarai B, Boyer S, Chen J, Consonni V, Farag S, Fourches D, García-Sosa AT, Gramatica P, Grisoni F, Grulke CM, Hong H, Horvath D, Hu X, Huang R, Jeliazkova N, Li J, Li X, Liu H, Manganelli S, Mangiatordi GF, Maran U, Marcou G, Martin T, Muratov E, Nguyen DT, Nicolotti O, Nikolov NG, Norinder U, Papa E, Petitjean M, Piir G, Pogodin P, Poroikov V, Qiao X, Richard AM, Roncaglioni A, Ruiz P, Rupakheti C, Sakkiah S, Sangion A, Schramm KW, Selvaraj C, Shah I, Sild S, Sun L, Taboureau O, Tang Y, Tetko IV, Todeschini R, Tong W, Trisciuzzi D, Tropsha A, Van Den Driessche G, Varnek A, Wang Z, Wedebye EB, Williams AJ, Xie H, Zakharov AV, Zheng Z, Judson RS (2020) CoMPARA: collaborative modeling project for androgen receptor activity. Environ Health Perspect 128(2):027,002. https://doi.org/10.1289/EHP5580
56. Papadopoulos H, Vovk V, Gammerman A (2011) Regression conformal prediction with nearest neighbours. J Artificial Intell Res 40:815–840. https://doi.org/10.1613/jair.3198
57. Svensson F, Aniceto N, Norinder U, Cortes I, Spjuth O, Carlsson L, Bender A (2018) Conformal regression for QSAR modelling—quantifying prediction uncertainty. J Chem Inform Model 58:1132–1140. https://doi.org/10.1021/acs.jcim.8b00054
58. Rey Moreno MC, Fussell KC, Gröters S, Schneider S, Strauss V, Stinchcombe S, Fegert I, Veras M, Van Ravenzwaay B (2013) Epoxiconazole-induced degeneration in rat placenta and the effects of estradiol supplementation. Birth Defects Res Part B Dev Reprod Toxicol 98(3):208–221. https://doi.org/10.1002/bdrb.21055
59. Balthazart J, Ball GF (1998) New insights into the regulation and function of brain estrogen synthase (aromatase). Trends Neurosci 21(6):243–249. https://doi.org/10.1016/S0166-2236(97)01221-6
60. Stinchcombe S, Schneider S, Fegert I, Rey Moreno MC, Strauss V, Gröters S, Fabian E, Fussell KC, Pigott GH, Van Ravenzwaay B (2013) Effects of estrogen coadministration on epoxiconazole toxicity in rats. Birth Defects Res Part B Dev Reprod Toxicol 98(3):247–259. https://doi.org/10.1002/bdrb.21059
61. Schneider S, Hofmann T, Stinchcombe S, Moreno MCR, Fegert I, Strauss V, Gröters S, Fabian E, Thiaener J, Fussell KC, Van Ravenzwaay B (2013) Species differences in developmental toxicity of epoxiconazole and its relevance to humans. Birth Defects Res Part B Dev Reprod Toxicol 98(3):230–246. https://doi.org/10.1002/bdrb.21058
62. Wang YM, Ong SS, Chai SC, Chen T (2012) Role of CAR and PXR in xenobiotic sensing and metabolism. Expert Opin Drug Metab Toxicol 8(7):803–817. https://doi.org/10.1517/17425255.2012.685237
63. Ihunnah CA, Jiang M, Xie W (2011) Nuclear receptor PXR, transcriptional circuits and metabolic relevance. Biochim Biophys Acta Mol Basis Dis 1812(8):956–963. https://doi.org/10.1016/j.bbadis.2011.01.014
64. Lange A, Corcoran J, Miyagawa S, Iguchi T, Winter MJ, Tyler CR (2017) Development of a common carp (*Cyprinus carpio*) pregnane X receptor (cPXR) transactivation reporter assay and its activation by azole fungicides and pharmaceutical chemicals. Toxicol In Vitro 41:114–122. https://doi.org/10.1016/j.tiv.2017.02.023
65. Hester S, Moore T, Padgett WT, Murphy L, Wood CE, Nesnow S (2012) The hepatocarcinogenic conazoles: cyproconazole, epoxiconazole, and propiconazole induce a common set of toxicological and transcriptional responses. Toxicol Sci 127(1):54–65. https://doi.org/10.1093/toxsci/kfs086
66. Oladimeji P, Cui H, Zhang C, Chen T (2016) Regulation of PXR and CAR by protein-protein interaction and signaling crosstalk. Expert Opin Drug Metab Toxicol 12(9):997–1010. https://doi.org/10.1080/17425255.2016.1201069
67. Choi FD, Juhasz ML, Atanaskova Mesinkovska N (2019) Topical ketoconazole: a systematic review of current dermatological applications and future developments. J Dermatol Treat. https://doi.org/10.1080/09546634.2019.1573309
68. Yu L, Chen M, Liu Y, Gui W, Zhu G (2013) Thyroid endocrine disruption in zebrafish larvae following exposure to hexaconazole and tebuconazole. Aquatic Toxicol 138–139:35–42. https://doi.org/10.1016/j.aquatox.2013.04.001
69. Jugan ML, Levi Y, Blondeau JP (2009) Endocrine disruptors and thyroid hormone physiology. Biochem Pharmacol 79(7):939–947. https://doi.org/10.1016/j.bcp.2009.11.006
70. Kjærstad MB, Andersen HR, Taxvig C, Hass U, Petersen MA, Metzdorff SB, Vinggaard A (2007) Effects of azole fungicides on the function of sex and thyroid hormones. https://orbit.dtu.dk/en/publications/id(682969e7-48a3-431a-9e45-34ab315cb866).html
71. Schwarzbacherová V, Wnuk M, Lewinska A, Potocki L, Zebrowski J, Koziorowski M, Holečková B, Šiviková K, Dianovský J (2017) Evaluation of cytotoxic and genotoxic activity of fungicide formulation Tango & #x00AE; super in bovine lymphocytes. Environ Pollut 220:255–263. https://doi.org/10.1016/j.envpol.2016.09.057
72. Schwarzbacherová V, Šiviková K, Drážovská M, Dianovský J (2015) Evaluation of DNA damage and cytotoxicity induced by triazole fungicide in cultured bovine lymphocytes. Caryologia 68(3):233–238. https://doi.org/10.1080/00087114.2015.1032613
73. Zhou J, Zhang J, Li F, Liu J (2016) Triazole fungicide tebuconazole disrupts human placental trophoblast cell functions. J Hazard Mater 308:294–302. https://doi.org/10.1016/j.jhazmat.2016.01.055
74. Benigni R, Bossa C, Tcheremenskaia O (2013) Nongenotoxic carcinogenicity of chemicals: mechanisms of action and early recognition through a new set of structural alerts. Chem Rev 113(5):2940–2957. https://doi.org/10.1021/cr300206t
75. Bock KW, Köhle C (2005) Ah receptor- and TCDD-mediated liver tumor promotion: clonal selection and expansion of cells evading growth arrest and apoptosis. Biochem Pharmacol 69(10):1403–1408. https://doi.org/10.1016/J.BCP.2005.02.004
76. Schramm F, Mueller A, Hammer H, Paschke A, Schueuermann G (2011) Epoxide and thiirane toxicity in vitro with the ciliates *Tetrahymena pyriformis*: structural alerts indicating excess toxicity. Environ Sci Technol 45(13):5812–5819. https://doi.org/10.1021/es200081n
77. Niklasson IB, Broo K, Jonsson C, Luthman K, Karlberg AT (2009) Reduced sensitizing capacity of epoxy resin systems: a structure-activity relationship study. Chem Res Toxicol 22(11):1787–1794. https://doi.org/10.1021/tx900193s
78. Fraenkel-Conrat H (1944) The action of 1,2-epoxides on proteins. J Biol Chem 154:227–238
79. Salmina ES, Haider N, Tetko IV (2016) Extended functional groups (EFG): an efficient set for chemical characterization and structure-activity relationship studies of chemical compounds. Molecules 21(1):1–8. https://doi.org/10.3390/molecules21010001
80. Truong L, Reif DM, St Mary L, Geier MC, Truong HD, Tanguay RL (2014) Multidimensional in vivo hazard assessment using zebrafish. Toxicol Sci 137(1):212–33. https://doi.org/10.1093/toxsci/kft235
81. Abdelhadya DH, El-Magd MA, Elbialy ZI, Saleh AA (2017) Bromuconazole-induced hepatotoxicity is accompanied by upregulation of PXR/CYP3A1 and downregulation of CAR/CYP2B1 gene expression. Toxicol Mech Methods 27(7):544–550. https://doi.org/10.1080/15376516.2017.1333555
82. Lv X, Pan L, Wang J, Lu L, Yan W, Zhu Y, Xu Y, Guo M, Zhuang S (2017) Effects of triazole fungicides on androgenic disruption and CYP3A4 enzyme activity. Environ Pollut 222:504–512. https://doi.org/10.1016/j.envpol.2016.11.051
83. Robitaille CN, Rivest P, Sanderson JT (2015) Antiandrogenic mechanisms of pesticides in human LNCaP prostate and H295R adrenocortical carcinoma cells. Toxicol Sci 143(1):126–135. https://doi.org/10.1093/toxsci/kfu212
84. Paul Friedman K, Papineni S, Marty MS, Yi KD, Goetz AK, Rasoulpour RJ, Kwiatkowski P, Wolf DC, Blacker AM, Peffer RC (2016) A predictive data-driven framework for endocrine prioritization: a triazole fungicide case study. Crit Rev Toxicol 46(9):785–833. https://doi.org/10.1080/10408444.2016.1193722
85. Knebel C, Buhrke T, Süssmuth R, Lampen A, Marx-Stoelting P, Braeuning A (2019) Pregnane X receptor mediates steatotic effects of propiconazole and tebuconazole in human liver cell lines. Arch Toxicol 93(5):1311–1322. https://doi.org/10.1007/s00204-019-02445-2
86. Browne P, Judson RS, Casey WM, Kleinstreuer NC, Thomas RS (2015) Screening chemicals for estrogen receptor bioactivity using a computational model. Environ Sci Technol 49(14):8804–8814. https://doi.org/10.1021/acs.est.5b02641
87. Grenet I, Comet JP, Schorsch F, Ryan N, Wichard J, Rouquié D (2019) Chemical in vitro bioactivity profiles are not informative about the long-term in vivo endocrine mediated toxicity. Comput Toxicol 12(June):100,098. https://doi.org/10.1016/j.comtox.2019.100098

88. Thomas RS, Black MB, Li L, Healy E, Chu TMM, Bao W, Andersen ME, Wolfinger RD, Lili L, Healy E, Chu TMM, Bao W, Andersen ME, Wolfinger RD (2012) A comprehensive statistical analysis of predicting in vivo hazard using high-throughput in vitro screening. Toxicol Sci 128(2):398–417. https://doi.org/10.1093/toxsci/kfs159
89. Tollefsen KE, Scholz S, Cronin MT, Edwards SW, de Knecht J, Crofton K, Garcia-Reyero N, Hartung T, Worth A, Patlewicz G (2014) Applying Adverse Outcome Pathways (AOPs) to support Integrated Approaches to Testing and Assessment (IATA). Regul Toxicol Pharmacol 70(3):629–640. https://doi.org/10.1016/J.YRTPH.2014.09.009
90. Gabbert S, Leontaridou M, Landsiedel R (2017) A critical review of adverse outcome pathway-based concepts and tools for integrating information from nonanimal testing methods: the case of skin sensitization. Appl In Vitro Toxicol 3(3):250–264. https://doi.org/10.1089/aivt.2017.0015

Publisher's Note

Additional file

KnowTox: Pipeline and Case Study for Confident Prediction of Potential Toxic Effects of Compounds in Early Phases of Development

Andrea Morger, Miriam Mathea, Janosch H Achenbach, Antje Wolf, Roland Buesen, Klaus-Juergen Schleifer, Robert Landsiedel and Andrea Volkamer

Table S1: Evaluation of original conformal prediction model for androgen receptor antagonism at 0.2 significance level (SL)

		Validity			Efficiency			Accuracy		
Dataset	Purpose	all	cl.1[a]	cl.0[a]	all	cl.1	cl.0	all	cl.1	cl.0
ToxCast-AA	train model	0.81	0.82	0.81	0.87	0.89	0.87	0.78	0.80	0.78
In-house-AA	validation I	0.59	0.98	0.16	0.94	0.98	0.91	0.56	0.97	0.07
External-AA	validation II	0.75	0.77	0.74	0.79	0.77	0.81	0.68	0.70	0.67

[a]cl.: class (class 1 = actives, class 0 = inactives)

Table S2: Evaluation of normalised conformal prediction model (normalised nonconformity score) for androgen receptor antagonism at 0.2 SL

		Validity			Efficiency			Accuracy		
Dataset	Purpose	all	cl.1[a]	cl.0[a]	all	cl.1	cl.0	all	cl.1	cl.0
ToxCast-AA	train model	0.85	0.82	0.85	0.37	0.14	0.40	0.95	0.46	0.98
In-house-AA	validation I	0.82	0.84	0.80	0.21	0.25	0.16	0.85	0.94	0.71
External-AA	validation II	0.75	0.65	0.83	0.29	0.19	0.37	0.77	0.39	0.93

[a]cl.: class (class 1 = actives, class 0 = inactives)

	Triazole1		Triazole2			Triazole1		Triazole2	
Endpoint	**p0**	**p1**	**p0**	**p1**	**Endpoint**	**p0**	**p1**	**p0**	**p1**
1	0.140	0.747	0.123	0.572	241	0.441	0.631	0.276	0.373
45	0.079	0.522	0.086	0.367	243	0.431	0.627	0.300	0.498
63	0.405	0.457	0.466	0.566	249	0.439	0.700	0.240	0.440
64	0.193	0.444	0.169	0.397	251	0.092	0.559	0.071	0.461
66	0.154	0.424	0.142	0.407	253	0.419	0.571	0.274	0.263
69	0.214	0.623	0.241	0.720	257	0.549	0.509	0.302	0.275
74	0.294	0.700	0.226	0.686	267	0.416	0.658	0.261	0.483
75	0.342	0.411	0.220	0.334	277	0.257	0.692	0.176	0.554
82	0.391	0.676	0.381	0.680	287	0.373	0.721	0.212	0.467
84	0.425	0.734	0.364	0.739	291	0.383	0.797	0.271	0.483
91	0.187	0.597	0.134	0.517	297	0.351	0.749	0.268	0.480
97	0.200	0.648	0.186	0.648	299	0.468	0.733	0.322	0.527
98	0.323	0.724	0.289	0.714	301	0.444	0.714	0.308	0.571
100	0.386	0.804	0.319	0.744	303	0.378	0.819	0.166	0.584
101	0.142	0.571	0.178	0.623	305	0.347	0.799	0.192	0.593
102	0.393	0.471	0.268	0.429	307	0.466	0.810	0.262	0.635
103	0.013	0.281	0.024	0.436	309	0.358	0.745	0.185	0.517
104	0.236	0.615	0.221	0.570	315	0.299	0.866	0.165	0.777
106	0.433	0.652	0.230	0.569	317	0.356	0.723	0.177	0.535
107	0.056	0.459	0.108	0.625	762	0.191	0.558	0.214	0.597
113	0.073	0.379	0.082	0.445	765	0.205	0.640	0.213	0.622
114	0.300	0.668	0.248	0.690	767	0.036	0.423	0.086	0.558
117	0.444	0.328	0.270	0.299	785	0.408	0.357	0.410	0.353
134	0.395	0.615	0.334	0.566	786	0.283	0.690	0.244	0.601
135	0.307	0.633	0.202	0.448	788	0.602	0.153	0.648	0.259
142	0.630	0.264	0.493	0.181	789	0.205	0.681	0.232	0.600
145	0.399	0.609	0.273	0.364	793	0.171	0.519	0.227	0.480
147	0.386	0.765	0.285	0.671	794	0.399	0.580	0.337	0.585
153	0.411	0.711	0.221	0.398	797	0.498	0.612	0.389	0.488
157	0.287	0.808	0.218	0.675	804	0.193	0.705	0.190	0.703
159	0.373	0.761	0.256	0.593	806	0.554	0.338	0.593	0.480
165	0.417	0.787	0.284	0.575	1110	0.262	0.692	0.232	0.709
167	0.368	0.470	0.299	0.238	1113	0.067	0.322	0.330	0.662
169	0.357	0.708	0.172	0.369	1116	0.325	0.594	0.269	0.477
171	0.452	0.563	0.442	0.422	1120	0.247	0.767	0.255	0.684
173	0.388	0.774	0.322	0.655	1127	0.289	0.636	0.339	0.590
175	0.478	0.616	0.304	0.414	1317	0.212	0.642	0.174	0.670
177	0.426	0.798	0.253	0.534	1321	0.342	0.593	0.406	0.678
179	0.350	0.807	0.213	0.590	1325	0.242	0.615	0.216	0.498
181	0.391	0.691	0.320	0.557	1329	0.264	0.547	0.222	0.440
185	0.326	0.762	0.303	0.713	1412	0.183	0.558	0.197	0.576
189	0.408	0.662	0.356	0.599	1441	0.297	0.782	0.244	0.695
221	0.329	0.811	0.173	0.599	1496	0.278	0.844	0.277	0.795
235	0.400	0.663	0.249	0.376	1816	0.159	0.436	0.194	0.464

Table S3: P-values obtained from predictions with 88 KnowTox conformal prediction models for *triazoles1&2*.

	Accuracy		Validity		Efficiency			Accuracy		Validity		Efficiency	
Endpoint	**cl0**	**cl1**	**cl0**	**cl1**	**cl0**	**cl1**	**Endpoint**	**cl0**	**cl1**	**cl0**	**cl1**	**cl0**	**cl1**
1	0.82	0.76	0.84	0.82	0.59	0.56	241	0.78	0.74	0.84	0.84	0.54	0.51
45	0.84	0.72	0.85	0.81	0.57	0.56	243	0.83	0.75	0.84	0.83	0.62	0.57
63	0.84	0.78	0.85	0.83	0.62	0.61	249	0.82	0.75	0.84	0.83	0.59	0.52
64	0.83	0.74	0.84	0.83	0.59	0.5	251	0.78	0.81	0.84	0.85	0.59	0.55
66	0.83	0.77	0.83	0.83	0.62	0.47	253	0.81	0.77	0.82	0.82	0.6	0.54
69	0.79	0.72	0.83	0.84	0.61	0.44	257	0.82	0.75	0.86	0.82	0.53	0.48
74	0.81	0.76	0.81	0.82	0.69	0.49	267	0.81	0.76	0.85	0.8	0.56	0.53
75	0.74	0.79	0.83	0.83	0.49	0.63	277	0.77	0.81	0.82	0.86	0.6	0.59
82	0.76	0.7	0.83	0.82	0.57	0.52	287	0.81	0.76	0.85	0.82	0.6	0.49
84	0.81	0.76	0.83	0.83	0.63	0.53	291	0.84	0.77	0.84	0.82	0.57	0.44
91	0.85	0.78	0.84	0.84	0.66	0.52	297	0.76	0.75	0.82	0.85	0.58	0.46
97	0.83	0.76	0.82	0.82	0.64	0.52	299	0.81	0.8	0.83	0.84	0.62	0.48
98	0.84	0.72	0.84	0.82	0.61	0.47	301	0.8	0.75	0.83	0.82	0.61	0.48
100	0.81	0.77	0.83	0.83	0.66	0.54	303	0.78	0.74	0.83	0.84	0.61	0.51
101	0.85	0.79	0.83	0.82	0.65	0.45	305	0.77	0.74	0.83	0.83	0.61	0.52
102	0.85	0.71	0.84	0.83	0.67	0.5	307	0.83	0.8	0.83	0.85	0.61	0.51
103	0.82	0.8	0.82	0.82	0.65	0.66	309	0.8	0.76	0.84	0.86	0.57	0.5
104	0.85	0.77	0.84	0.84	0.64	0.45	315	0.78	0.77	0.82	0.83	0.61	0.59
106	0.83	0.62	0.86	0.83	0.53	0.39	317	0.82	0.8	0.83	0.84	0.61	0.51
107	0.84	0.8	0.82	0.83	0.68	0.58	762	0.91	0.76	0.85	0.84	0.6	0.39
113	0.84	0.81	0.82	0.83	0.68	0.61	765	0.9	0.77	0.84	0.83	0.57	0.38
114	0.83	0.72	0.84	0.84	0.59	0.47	767	0.88	0.77	0.84	0.82	0.6	0.47
117	0.78	0.81	0.81	0.82	0.6	0.71	785	0.88	0.85	0.84	0.81	0.58	0.62
134	0.87	0.77	0.83	0.83	0.65	0.5	786	0.93	0.76	0.85	0.83	0.6	0.37
135	0.85	0.79	0.82	0.82	0.69	0.65	788	0.61	0.85	0.84	0.86	0.3	0.45
142	0.73	0.81	0.82	0.83	0.52	0.66	789	0.92	0.79	0.83	0.83	0.6	0.41
145	0.8	0.71	0.84	0.8	0.56	0.51	793	0.87	0.84	0.83	0.81	0.66	0.54
147	0.81	0.76	0.85	0.84	0.64	0.49	794	0.83	0.84	0.85	0.84	0.57	0.55
153	0.82	0.8	0.83	0.85	0.6	0.47	797	0.97	0.77	0.85	0.82	0.65	0.37
157	0.8	0.77	0.83	0.82	0.62	0.56	804	0.9	0.81	0.83	0.81	0.64	0.53
159	0.82	0.79	0.82	0.82	0.59	0.45	806	0.91	0.85	0.86	0.83	0.67	0.59
165	0.83	0.78	0.84	0.84	0.58	0.43	1110	0.82	0.78	0.83	0.83	0.61	0.57
167	0.77	0.77	0.83	0.84	0.52	0.52	1113	0.79	0.79	0.83	0.82	0.57	0.54
169	0.81	0.78	0.84	0.84	0.56	0.46	1116	0.89	0.78	0.84	0.82	0.59	0.46
171	0.81	0.72	0.84	0.83	0.59	0.5	1120	0.88	0.76	0.83	0.82	0.68	0.5
173	0.84	0.8	0.84	0.81	0.62	0.5	1127	0.89	0.66	0.85	0.82	0.66	0.42
175	0.84	0.79	0.85	0.85	0.59	0.49	1317	0.81	0.78	0.84	0.83	0.59	0.56
177	0.81	0.78	0.85	0.85	0.59	0.46	1321	0.81	0.77	0.84	0.82	0.54	0.53
179	0.8	0.75	0.85	0.82	0.58	0.5	1325	0.89	0.72	0.86	0.81	0.56	0.45
181	0.81	0.75	0.83	0.82	0.61	0.56	1329	0.86	0.76	0.85	0.85	0.54	0.44
185	0.81	0.75	0.83	0.81	0.58	0.49	1412	0.83	0.74	0.83	0.82	0.68	0.57
189	0.83	0.77	0.85	0.85	0.6	0.48	1441	0.79	0.67	0.84	0.81	0.58	0.5
221	0.78	0.76	0.83	0.82	0.57	0.56	1496	0.76	0.73	0.82	0.82	0.61	0.57
235	0.81	0.8	0.83	0.84	0.65	0.56	1816	0.92	0.81	0.84	0.84	0.68	0.48

Table S4: Class-wise evaluation of conformal prediction models.

Endpoint_ID	Description	Endpoint Family	Endpoint Subfamily	# Actives	# Inactives
1	T47D	cell cycle	cytotoxicity	413	1169
45	CellLoss	cell cycle	cytotoxicity	396	533
63	Ahr	dna binding	basic helix-loop-helix protein	397	2670
64	AP	dna binding	basic leucine zipper	557	2501
66	BRE	dna binding	Smad protein	349	2721
69	CRE	dna binding	basic leucine zipper	303	2760
74	EGR	dna binding	zinc finger	441	2624
75	ERE	nuclear receptor	steroidal	794	2254
82	HIF1a	dna binding	basic helix-loop-helix protein	335	2731
84	HSE	dna binding	heat shock protein	369	2696
91	MRE	dna binding	zinc finger	647	2417
97	NRF2	dna binding	basic leucine zipper	1164	1878
98	Oct	dna binding	POU domain protein	481	2578
100	Pax6	dna binding	paired box protein	405	2664
101	PBREM	nuclear receptor	non-steroidal	346	2721
102	PPRE	nuclear receptor	non-steroidal	469	2591
103	PXRE	nuclear receptor	non-steroidal	1494	1549
104	RORE	nuclear receptor	orphan	333	2739
106	Sp1	dna binding	zinc finger	302	2760
107	SREBP	dna binding	basic helix-loop-helix leucine zipper	473	2591
113	VDRE	nuclear receptor	non-steroidal	819	2231
114	Xbp1	dna binding	basic leucine zipper	401	2662
117	ERa	nuclear receptor	steroidal	678	2388
134	PPARg	nuclear receptor	non-steroidal	852	2206
135	PXR	nuclear receptor	non-steroidal	873	2177
142	RXRb	nuclear receptor	non-steroidal	479	2580
145	Eselectin	cell adhesion molecules	selectins	376	947
147	HLADR	cell adhesion molecules	MHC Class II	510	812
153	MCP1	cytokine	chemotactic factor	346	976
157	Proliferation	cell cycle	cytotoxicity	554	767
159	SRB	cell cycle	cytotoxicity	433	886
165	uPAR	cytokine	plasmogen activator	382	940
167	VCAM1	cell adhesion molecules	Immunoglobulin CAM	328	997
169	Vis	cell morphology	cell conformation	396	921
171	Eotaxin3	cytokine	chemotactic factor	403	920
173	MCP1	cytokine	chemotactic factor	358	964
175	Pselectin	cell adhesion molecules	selectins	372	954
177	SRB	cell cycle	cytotoxicity	361	962
179	uPAR	cytokine	plasmogen activator	319	1005
181	VCAM1	cell adhesion molecules	Immunoglobulin CAM	400	923
185	HLADR	cell adhesion molecules	MHC Class II	356	965
189	IP10	cytokine	chemotactic factor	349	974
221	Proliferation	cell cycle	cytotoxicity	401	920
235	CollagenIII	cell adhesion molecules	collagen	403	915
241	IP10	cytokine	chemotactic factor	384	933
243	MCSF	cytokine	colony stimulating factor	388	932
249	PAI1	cytokine	plasmogen activator inhibitor	367	957
251	Proliferation	cell cycle	cytotoxicity	575	746
253	SRB	cell cycle	cytotoxicity	312	1012
257	VCAM1	cell adhesion molecules	Immunoglobulin CAM	383	939
267	MMP9	protease	matrix metalloproteinase	359	965
277	CD40	cytokine	inflammatory factor	397	927
287	MCSF	cytokine	colony stimulating factor	358	962
291	SRB	cell cycle	cytotoxicity	341	981
297	VCAM1	cell adhesion molecules	Immunoglobulin CAM	405	919
299	CD38	cytokine	other cytokine	404	918
301	CD40	cytokine	inflammatory factor	399	922
303	CD69	cytokine	inflammatory factor	370	953
305	Eselectin	cell adhesion molecules	selectins	377	945
307	IL8	cytokine	interleukins	315	1005
309	MCP1	cytokine	chemotactic factor	318	1004
315	Proliferation	cell cycle	cytotoxicity	522	802
317	SRB	cell cycle	cytotoxicity	333	990
762	AR	nuclear receptor	steroidal	868	5845
765	AR	nuclear receptor	steroidal	603	6161
767	Aromatase	cyp	steroidogenesis-related	925	5770
785	ERa	nuclear receptor	steroidal	323	6462
786	ERa	nuclear receptor	steroidal	761	5969
788	ERa	nuclear receptor	steroidal	857	5716
789	ERa	nuclear receptor	steroidal	494	6265
793	GR	nuclear receptor	steroidal	369	6431
794	GR	nuclear receptor	steroidal	333	6471
797	MMP	cell morphology	organelle conformation	705	3810
804	TR	nuclear receptor	non-steroidal	1390	5345
806	AhR	dna binding	basic helix-loop-helix protein	641	6106
1110	ARE	dna binding	basic leucine zipper	1199	4871
1113	HSE	dna binding	heat shock protein	404	5781
1116	p53	dna binding	tumor suppressor	593	6167
1120	FXR	nuclear receptor	non-steroidal	698	5467
1127	PPARg	nuclear receptor	non-steroidal	389	5813
1317	p53	dna binding	tumor suppressor	735	5975
1321	p53	dna binding	tumor suppressor	663	6033
1325	p53	dna binding	tumor suppressor	668	6048
1329	p53	dna binding	tumor suppressor	648	6082
1412	DR4	nuclear receptor	non-steroidal	624	2438
1441	ISRE	dna binding	interferon regulatory factors	405	2655
1496	TCF	dna binding	HMG box protein	384	2670
1816	AR	nuclear receptor	steroidal	916	5286

Table S5: Information about endpoints where conformal prediction models are available. More details about all endpoints can be found under: https://figshare.com/articles/ToxCast_and_Tox21_Data_Spreadsheet/6062503.

4.2 Quantitative high-throughput phenotypic screening for environmental estrogens using the E-Morph Screening Assay in combination with *in silico* predictions

The combination of toxicity prediction methods was shown successful in the KnowTox project (see Section 4.1), yet the extensive prediction output can be challenging to interpret for non-toxicology experts. The following work focuses on the ER and, for facilitated decision making, multiple ER assays from ToxCast will be organised in a consensus approach. Activity on the ER can lead to endocrine disruption, a crucial threat to the environment with consequences such as hormone-dependent cancers or reproductive dysfunction. In the following study, the recently-developed E-Morph Screen ER assay, aimed at the detection of novel estrogenic substances, will be validated. Although conducting the *in vitro* assay may be relatively fast, efforts required for synthesis or purchase of substances to be screened should not be neglected. Therefore, we explore to what extent the addition of *in silico* toxicity prediction methods can make the screening more efficient. It will be prospectively investigated if pre-selecting more-likely active compounds with a similarity search and conformal predictions can increase the hit rate.

Contribution:

Middle author
Conceptual design (20%)
Computational experiments (90%)
In vitro experiments (0%)
Visualization (10%)
Manuscript preparation (20%)

Environment International 158 (2022) 106947

Contents lists available at ScienceDirect

Environment International

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

Quantitative high-throughput phenotypic screening for environmental estrogens using the E-Morph Screening Assay in combination with *in silico* predictions

Saskia Klutzny [a], Marja Kornhuber [a,b], Andrea Morger [c], Gilbert Schönfelder [a,d], Andrea Volkamer [c], Michael Oelgeschläger [a], Sebastian Dunst [a,*]

[a] *Experimental Toxicology and ZEBET, German Federal Institute for Risk Assessment (BfR), German Centre for the Protection of Laboratory Animals (Bf3R), Berlin, Germany*
[b] *Freie Universität Berlin, Berlin, Germany*
[c] *In silico Toxicology and Structural Bioinformatics, Institute of Physiology, Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin, Humboldt-Universität zu Berlin, and Berlin Institute of Health, Berlin, Germany*
[d] *Institute of Clinical Pharmacology and Toxicology, Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin, Humboldt-Universität zu Berlin, and Berlin Institute of Health, Berlin, Germany*

ARTICLE INFO

Handling Editor: Olga Kalantzi

ABSTRACT

Background: Exposure to environmental chemicals that interfere with normal estrogen function can lead to adverse health effects, including cancer. High-throughput screening (HTS) approaches facilitate the efficient identification and characterization of such substances.
Objectives: We recently described the development of the E-Morph Assay, which measures changes at adherens junctions as a clinically-relevant phenotypic readout for estrogen receptor (ER) alpha signaling activity. Here, we describe its further development and application for automated robotic HTS.
Methods: Using the advanced E-Morph Screening Assay, we screened a substance library comprising 430 toxicologically-relevant industrial chemicals, biocides, and plant protection products to identify novel substances with estrogenic activities. Based on the primary screening data and the publicly available ToxCast dataset, we performed an *in silico* similarity search to identify further substances with potential estrogenic activity for follow-up hit expansion screening, and built seven *in silico* ER models using the conformal prediction (CP) framework to evaluate the HTS results.
Results: The primary and hit confirmation screens identified 27 'known' estrogenic substances with potencies correlating very well with the published ToxCast ER Agonist Score ($r = +0.95$). We additionally detected potential 'novel' estrogenic activities for 10 primary hit substances and for another nine out of 20 structurally similar substances from *in silico* predictions and follow-up hit expansion screening. The concordance of the E-Morph Screening Assay with the ToxCast ER reference data and the generated CP ER models was 71% and 73%, respectively, with a high predictivity for ER active substances of up to 87%, which is particularly important for regulatory purposes.
Discussion: These data provide a proof-of-concept for the combination of *in vitro* HTS approaches with *in silico* methods (similarity search, CP models) for efficient analysis of large substance libraries in order to prioritize substances with potential estrogenic activity for subsequent testing against higher tier human endpoints.

1. Introduction

Endocrine-disrupting chemicals (EDCs) are a group of exogenous substances that interfere with the endocrine system, leading to adverse health effects, including cancer (WHO/IPCS, 2002). Global cancer burden has substantially increased over the last decades and incidence

* Corresponding author at: German Federal Institute for Risk Assessment (BfR), German Centre for the Protection of Laboratory Animals (Bf3R), Max-Dohrn-Straße 8-10, 10589 Berlin, Germany
E-mail address: Sebastian.Dunst@bfr.bund.de (S. Dunst).

https://doi.org/10.1016/j.envint.2021.106947
Received 14 July 2021; Received in revised form 14 October 2021; Accepted 18 October 2021
Available online 28 October 2021

rates are projected to further rise in the future (Sung et al. 2021; Wild et al. 2020). In 2020, female breast cancer has been the most commonly diagnosed cancer worldwide (Sung et al. 2021) and estrogens are an important risk factor (Yager and Davidson 2006). Thus, the identification of exogenous substances that mimic estrogen function and subsequent reduction of human exposure to such substances are important measures for the effective prevention of endocrine-related cancers such as breast cancer.

Environmental sources of substances with estrogenic activity are manifold and include consumer products and food packaging materials, preservatives, food additives, and pesticides, but are also naturally found in food (e.g. phytoestrogens) (Paterni et al. 2017). In regulatory toxicology, the detection of potential substance-related adverse health effects still calls for traditional *in vivo* test guideline studies that mainly use rodents to evaluate the potential hazards of single substances (OECD 2001; 2018a; b; c). However, such cost- and time-consuming *in vivo* studies not necessarily mimic human-relevant physiological and disease conditions. In addition, the ethical issues related to animal testing in general and the high numbers of test animals needed further fuel the rising need and interest in human-relevant alternative *in silico*, *in chemico*, and *in vitro* test methods to reduce and eventually replace animal testing according to the 3Rs principle (Russell and Burch 1959).

Following the assumption that structurally similar substances can have similar toxicological effects (Maggiora et al. 2014), risk assessment commonly uses read-across approaches to effectively reduce costs and animal testing (Carrio et al. 2016; Hemmerich and Ecker 2020; Raies and Bajic 2016). Indeed, various established QSAR models, which represent statistical *in silico* models that relate a set of structural descriptors of a substance to its biological activity, can support read-across (Ma et al. 2015; Tropsha 2010). Still, the prediction of toxicity from structural similarity remains challenging for risk assessment because *in silico* predictions alone do not yet sufficiently fulfill information requirements for complex human health endpoints.

In recent years, machine learning approaches gained momentum, which use the increasingly comprehensive *in chemico* and *in vitro* test data in an iterative process with growing certainty to identify combinations of substance features that may lead to a specific toxicological effect (Gayvert et al., 2016; Huang and Xia, 2017; Mayr et al., 2016). A special case of machine learning is conformal prediction (CP), which adds confidence estimation to the model predictions (Alvarsson et al. 2021; Norinder et al. 2014; Vovk et al. 2005). The CP framework is built on top of a machine learning method and uses an additional calibration step based on experimental test results to determine the confidence when making predictions on new data. CP models have recently been intensively and successfully built and applied to toxicological questions (Morger et al. 2020; Morger et al. 2021; Norinder et al. 2016; Svensson et al. 2017b; Zhang et al. 2021).

The increasing quantity and diversity of chemicals that are produced and marketed worldwide stimulate the establishment of high-throughput screening (HTS) research programs that use *in chemico* and *in vitro* assays to efficiently generate comprehensive concentration–response information for a large number of substances. For example, the U.S. EPA Toxicity Forecaster (ToxCast) project generated screening data for over 10,000 environmental chemicals that were tested in hundreds of HTS assays addressing toxicological and endocrine endpoints in order to rank and prioritize substances for subsequent *in vivo* testing (Dix et al. 2007; Judson et al. 2010; Reif et al. 2010; Rotroff et al. 2013). These screening data have further been integrated into an *in silico* ToxCast ER pathway model, which converts results from 18 automated ER screening assays into a relative ER bioactivity score ranging from 0.00 (no activity) to 1.00 (bioactivity of the reference substance 17-alpha-Ethinylestradiol) (Browne et al. 2015; Judson et al. 2015). More recently, Judson *et al.* demonstrated that a reduced set of four out of the originally 18 ER screening assays achieves a comparable performance (Judson et al. 2017). Furthermore, the Collaborative Estrogen Receptor Activity Prediction Project (CERAPP) developed another *in silico* consensus model for prediction of ER binding, agonistic, and antagonistic activities of chemicals (Mansouri et al. 2016). This CERAPP ER consensus model integrates 48 individual computational models using different QSAR and structure-based approaches, which have been trained and optimized using the relative potency information from the ToxCast ER Agonist Model (Browne et al. 2015; Judson et al. 2015).

The *in chemico* and *in vitro* HTS assays that provide the data for the ToxCast ER pathway model each cover single, mechanistic events of estrogen signaling (ER binding, ER dimerization, regulation of gene expression, and cell proliferation) but the immediate relevance of the derived data regarding adverse effects, including human cancer, is still limited. Hence, development and application of novel cell-based test methods that combine HTS capability with more human-relevant, functional endpoints can support a more direct extrapolation of the test results to the complex signaling events and regulatory mechanisms that drive adverse effects including cancer progression and metastasis. We have recently shown that the cell-based E-Morph Assay provides such an endpoint, i.e., the machine learning-based analysis of estrogen-dependent phenotypic changes at adherens junctions (AJ) (Bischoff et al. 2020; Kornhuber et al. 2021). The E-Morph Assay is based on the observation that the inhibition of ER signaling in an MCF-7 breast cancer cell line led to a prominent reorganization of AJs and induced the clustering of the AJ protein E-cadherin (E-Cad), which could be reverted by co-treatment with estrogenic substances (Bischoff et al. 2020). These changes in cell morphology correlated with increased cellular stiffness and decreased cell motility, with deregulation of these two parameters often being associated with breast cancer progression and metastasis (Bischoff et al. 2020). In addition, we could describe comparable changes in E-Cad localization in clinical breast cancer tissue samples supporting the clinical relevance of the assay endpoint.

In the present study, we optimized the E-Morph Assay for automated robotic HTS and used this advanced E-Morph Screening Assay to analyze a substance library comprising 430 toxicologically-relevant industrial chemicals, biocides and plant protection products that are reported to act through various nuclear hormone receptors. Using our HTS data in combination with already publicly available *in chemico* and *in vitro* ToxCast data as well as *in silico* prediction approaches using the CP framework, we could further identify additional, novel substances with potential estrogenic activity.

2. Materials and methods

2.1. Cell line and cell culture conditions

The MCF-7/E-Cad-GFP cell line (de Beco et al. 2009; 2020) that stably expresses a fluorescent E-Cadherin-GFP fusion protein was kindly provided by Sylvie Coscoy (Laboratoire Physico-Chimie Curie, Institut Curie, PSL Research University - Sorbonne Universités, UPMC-CNRS, Paris, France).

Routine cell cultures were maintained at 37 °C with 5% CO_2 in normal-serum medium containing Dulbecco's modified Eagle's medium (DMEM, low glucose, pyruvate, no glutamine, no phenol red) (Gibco/Thermo Fisher Scientific, Waltham, MA, USA), 10% (v/v) Fetal Bovine Serum (FBS, S0615, Estradiol levels: 22.3 pg/ml) (Biochrom/Merck, Darmstadt, Germany), 2 mM stable glutamine (Gibco/Thermo Fisher Scientific), 100 μg/ml streptomycin / 100 U/ml penicillin (Biochrom/Merck), and 0.4 mg/ml geneticin (Gibco/Thermo Fisher Scientific). Cells were sub-cultured over a maximum of 10–12 passages, and regularly tested using the Eurofins Genomics mycoplasma test service (Eurofins Genomics, Ebersberg, Germany).

Experiments were performed in reduced-serum medium as described above but containing only 5% (v/v) FBS to minimize background estrogen levels and potential test chemical binding to serum lipids and proteins in the exposure medium. The final estradiol concentration in reduced-serum medium (4.1 pM) was in the range of physiological

serum levels of postmenopausal women (Rothman et al. 2011). If not otherwise stated, cells were seeded into multi-well plates at suitable concentrations to achieve 80–90% confluency after 24 h. Cells were then exposed to reduced-serum medium containing the respective test chemical in combination with the anti-estrogen Fulvestrant (Fulv, 10 nM) (Sigma-Aldrich/Merck, Darmstadt, Germany) for 48 h, followed by the sample preparation procedure. Experimental controls included the solvent control, the Fulv control containing 10 nM Fulv only, and the co-treatment (reactivity) control containing 10 nM Fulv + 10 μM Estrone (Sigma-Aldrich/Merck) (each in reduced-serum medium). In all experiments, the solvent control corresponds to the respective experimental conditions, excluding Fulv and test chemicals. The DMSO (Sigma-Aldrich/Merck) concentration in the solvent control was always adjusted to the highest DMSO concentration used in the experiment, i.e. in the range of 0.1–0.4% depending on the experimental setting.

2.2. Quantitative PCR

Cells were seeded into 12-well plates at a concentration of $4x10^5$ cells/well in 1 ml reduced-serum medium and exposed to test substances as described above. RNA extraction (RNeasy Kit, Qiagen, Hilden, Germany), cDNA synthesis (High-Capacity cDNA Reverse Transcription Kit, Applied Biosystems/Thermo Fisher Scientific, Waltham, MA, USA), and quantitative PCR (qPCR) (PowerUp SYBR Green Master Mix, Applied Biosystems/Thermo Fisher Scientific) were conducted according to the manufacturers protocols using a QuantStudio 7 Flex Real-Time PCR System (Applied Biosystems/Thermo Fisher Scientific) (40 cycles; denaturation for 15 s at 95 °C; annealing, extension, and fluorescence read for 1 min at 60 °C). RNA concentrations (A260) and purity ratios (A260/A280 and A260/A230) were determined using a NanoDrop 2000 spectrophotometer (Thermo Fisher Scientific). Purity ratios of ~ 2.0 (A260/A280) and 2.0–2.2 (A260/A230) were generally considered acceptable. cDNA synthesis was performed using 1 μg RNA and RT random primers (High-Capacity cDNA Reverse Transcription Kit, Applied Biosystems/Thermo Fisher Scientific). For qPCR, 1 μl of 1:10 diluted (water) cDNA was added to 10 μl master mix containing water, primers and SYBR Green. RNA expression levels (fold change) were calculated according to the $\Delta\Delta C_T$ method (Livak and Schmittgen 2001). Tyrosine 3-Monooxygenase/Tryptophan 5-Monooxygenase Activation Protein, Zeta (YWHAZ) was used as housekeeping gene. If not otherwise stated, each experiment was performed in technical triplicates and in at least three independent repetitions. Primers used (5′-3′ orientation):

BCL2L1 (CAGCTTGGATGGCCACTTAC, TGCTGCATTGTTCCCATAGA);

TFF1 (CATCGACGTCCCTCCAGAAGAG, CTCTGGGACTAATCACCGTGCTG);

PGR (TCAACTACCTGAGGCCGGAT, GCTCCCACAGGTAAGGACAC);

AREG (TGGATTGGACCTCAATGACA, TAGCCAGGTATTTGTGGTTCG);

ESR1 (CCACCAACCAGTGCACCATT, GGTCTTTTCGTATCCCACCTTTC);

GFP (AAGCTGACCCTGAAGTTCATCTGC, CTTGTAGTTGCCGTCGTCCTTGAA);

CDH1 (AGGAGCCAGACACATTTATGGAA, GCTGTGTACGTGCTGTTCTTCAC);

mCdh1 (AACCCAAGCACGTATCAGGG, GAGTGTTGGGGGCATCATCA);

YWHAZ (ACTTTTGGTACATTGTGGCTTCAA, CCGCCAGGACAAACCAGTAT).

2.3. Western blot

Cells were seeded into 6 or 12-well plates at a concentration of $1x10^6$ or $4x10^5$ cells/well in 2 ml or 1 ml reduced-serum medium and exposed to substances as described above. For protein extraction, cells were washed with ice-cold Phosphate-buffered saline (PBS) and scraped in 100–200 μl lysis buffer (50 mM Tris/HCl pH 7.4, 150 mM NaCl, 0.1% (w/v) Na-deoxycholate, 0.1% (w/v) SDS, 1% (v/v) IGEPAL CA-630/NP-40, 5 mM EDTA pH 8.0, 5 mM EGTA, 1X cOmplete Protease Inhibitor Cocktail (Roche, Basel, Switzerland), PhosSTOP Phosphatase Inhibitor Cocktail (Roche) and incubated for 30 min on ice. Lysates were centrifuged at 13,000 g and 4 °C for 10 min, and the supernatant was collected. Total protein concentrations were determined using a Pierce BCA Protein Assay Kit (Thermo Scientific/Thermo Fisher Scientific, Waltham, MA, USA) and a BSA standard (Thermo Fisher Scientific) according to the manufacturer's instructions. Protein lysates were separated by SDS-PAGE using Mini-PROTEAN precast gels (4–15% polyacrylamide) (Bio-Rad Laboratories, Hercules, CA, USA) according to the manufacturer's instructions. Proteins were transferred onto nitrocellulose membranes (Bio-Rad Laboratories) using a semi-dry Trans-Blot Turbo Transfer System (1.3 A per gel, 25 V for 7 min) (Bio-Rad Laboratories). Membranes were blocked with 5% low-fat milk powder for 60 min, rinsed in Tris-buffered saline containing Tween 20 (TBS-T) (TBS, 0.1% Tween 20), and incubated with primary and secondary antibodies in 0.6% low-fat milk powder in TBS-T (TBS, 0.1% Tween 20) over night at 4 °C and for 3 h at room temperature, respectively. Antibodies/dyes used: mouse anti-E-Cad (1:1,000) (Clone 36, BD Biosciences, Franklin Lakes, NJ, USA) and HRP-conjugated goat anti-mouse secondary antibody (1:10,000) (Jackson ImmunoResearch, West Grove, PA, USA). Protein detection was carried out using a Pierce ECL Western Blotting Substrate (Thermo Scientific/Thermo Fisher Scientific) in a Fusion Solo S (VWR, Radnor, PA, USA) imaging system. Coomassie Brilliant Blue (Bio-Rad Laboratories) total protein staining of nitrocellulose membranes was used as loading control (Welinder and Ekblad 2011). Semi-quantitative densitometric analysis of western blot bands was performed using the FIJI software (Schindelin et al. 2012). The band intensities were normalized to the respective Coomassie total protein staining of each lane. The results from each treatment condition were then normalized to the solvent control.

2.4. siRNA knockdown

Cells were seeded into 6 or 12-well plates at a concentration of $1x10^6$ or $4x10^5$ cells/well in 2 ml or 1 ml reduced-serum medium and exposed to substances as described above. Transfections were carried out using the HiPerFect Transfection Reagent (Qiagen) and a mix of four *ESR1* siRNAs (FlexiTube GeneSolution GS2099, Quiagen) (10 nM) with different target sequences (SI02781401; SI03114979; SI03065615; SI00002527). Cells were transfected at the time of cell seeding according to the manufacturer's reverse-transfection protocol.

2.5. ER binding experiments

Cells were seeded into 96-well plates and exposed to substances as described above. Transfections were carried out using the FuGENE HD Transfection Reagent (Promega, Madison, WI, USA), the pBIND-ERα [*hRluc*] vector (50 ng), and the pGL4.35[*luc2P/9XGAL4 UAS/Hygro*] vector (50 ng) (both Promega) according to the manufacturer's protocol. The binding of a test substance with estrogenic activity to a fusion protein containing an estrogen receptor-ligand binding domain (ER-LBD) and a yeast Gal4 DNA-binding domain (Gal4-DBD) (pBIND-ERα [*hRluc*] vector) led to the expression of an UAS-controlled Firefly luciferase reporter protein (pGL4.35[luc2P/9XGAL4UAS/Hygro] vector), which was detected using the Dual-Glo Luciferase Reagent (Promega). The detected Firefly luminescence (pGL4.35[luc2P/9XGAL4UAS/Hygro] vector) was normalized to the Renilla luminescence (pBIND-ERα [*hRluc*] vector) to derive a relative signal intensity.

2.6. E-Morph Screen: Cell seeding and test substance exposure scenarios

Cells were seeded into CellCarrier-96 Ultra Microplates (PerkinElmer, Waltham, MA, USA) at a concentration of $9x10^4$ cells/well in

225 µl reduced-serum medium, grown until 80–90% confluency for 24 h, and then exposed to 250 µl reduced-serum medium containing each test chemical in combination with 10 nM Fulv for 48 h.

All 430 test substances (Sigma-Aldrich/Merck) of the BfR-ChemLibrary were previously dissolved in DMSO (Sigma-Aldrich/Merck) at a stock concentration of 10 mM and stored at the *Compound Management Unit* of the Leibniz Institute of Molecular Pharmacology (FMP, Berlin, Germany). For this project, a copy of the BfR-ChemLibrary was provided by the FMP on five 96-well microplates (Greiner Bio-One, Frickenhausen, Germany) along with an empty column for the assay controls. The preparation of the exposure medium and its application to cells was performed using a JANUS Automated Liquid Handling Workstation (PerkinElmer) and customized treatment protocols written in WinPREP (PerkinElmer).

For the hit selection (primary) screen, 2 µl of the test substance (10 mM) were transferred into an empty 96-well microplate (Greiner Bio-One) and then dissolved (1:100) in 198 µl reduced-serum medium containing 100 nM Fulv. Subsequently, 25 µl of the exposure medium containing the diluted test substance (100 µM) were then transferred from the compound plates to the cell culture assay plates containing 225 µl reduced-serum medium to achieve a final test substance concentration of 10 µM and a final Fulv concentration of 10 nM. Considering that the nominal concentration of a test chemical does not necessarily reflect the concentration at the target site due to potential partitioning of test chemicals to other extracellular compartments in *in vitro* assays (Proença et al., 2021), a starting concentration of 10 µM is often used for hit selection in comparable HC/HT screening projects in order to maximize exposure of cells to the test substance. Sufficiently high exposure levels ensure confidence in negative test results and are particularly important for detection of substances with weak estrogenic activities, such as industrial chemicals, that were, in contrast to pharmaceuticals, not designed to act on the ER pathway.

For the hit confirmation (potency) and the hit expansion screens, 9 µl of the test substance (10 mM) were transferred into an empty 96-well microplate (Greiner Bio-One) and then dissolved (1:33) in 291 µl reduced-serum medium containing 100 nM Fulv. From this start concentration (300 µM), serial dilutions were generated at a 1:3 ratio in reduced-serum medium containing 100 nM Fulv. Subsequently, 25 µl of the exposure medium containing the diluted test substance (300 µM to 10 pM) was then transferred from the compound plates to the cell culture assay plates containing 225 µl reduced-serum medium to achieve a final test substance concentration of 30 µM to 1 pM and a final Fulv concentration of 10 nM.

For all screening approaches, the three solvent control wells of each plate contained 0.2% (primary screen) or 0.4% (hit confirmation screen) DMSO, the three Fulv control wells contained 10 nM Fulv, and the two co-treatment (reactivity) control wells contained 10 nM Fulv + 10 µM Estrone (each in reduced-serum medium).

2.7. E-Morph Screen: Fluorescence microscopy and quantitative image analysis

The preparation of the cells for fluorescence microscopy was performed using a JANUS Automated Liquid Handling Workstation (PerkinElmer) and an ELx405 Select CW Microplate Washer (BioTek Instruments, Winooski, VT, USA). After treatment for 48 h, the cells were stained in PBS containing 1 µM CellTrace Far Red (Molecular Probes/Thermo Fisher Scientific, Waltham, MA, USA) to visualize the cell–cell contact morphology according to (Kornhuber et al. 2021) as an internal quality control and 2 µg/ml Hoechst 33,342 (Molecular Probes/Thermo Fisher Scientific) to label nuclei for 20 min at 37 °C with 5% CO_2, then washed twice with PBS, fixed with 4% formaldehyde solution for 15 min at room temperature, and finally washed again with PBS. During this procedure, the E-Cad-GFP signal was preserved and did not require additional staining.

Cells were subsequently imaged with an Opera Phenix High-Content Screening System (PerkinElmer) using a 20x air objective (NA 0.4) at three standardized positions per well and three or four optical sections with 4 µm or 3 µm spacing per position. Image analysis was performed using the integrated Harmony software (PerkinElmer) and customized image analysis routines (Fig. S1A). First, nuclei were identified using the Hoechst 33,342 channel to define each cell. Nuclei touching the edge of the image were excluded from further analysis. Next, cell outlines were identified using the GFP channel and the E-Cad-GFP signal intensity was measured for each cell. Finally, a mean E-Cad-GFP signal intensity was calculated across all cells per well.

For visualization of concentration-response curves of relative E-Cad-GFP signal intensities ($SI^{E\text{-}Cad\text{-}GFP}$), the mean E-Cad-GFP signal intensity (SI) per well (SI_{well}) was normalized to the average SI (SI_{avr}) of the corresponding three solvent control wells on each plate according to (Malo et al. 2006):

$$SI^{E\text{-}Cad\text{-}GFP} = \frac{SI_{well}^{Substance}}{SI_{avr}^{Solv}} * 100$$

2.8. E-Morph Screen: Automated data evaluation

The process automation software KNIME [v4.1.2] (Berthold et al. 2008) was used to build a customized pipeline (Fig. S1B; File S1) for fast and efficient automated processing, evaluation, and statistical analysis of the quantitative image data obtained from the individual screens. Briefly, this KNIME workflow retrieved all *.txt* files that were exported from the Harmony software (PerkinElmer) into a specified folder and executed the following steps in a loop function: a) import *.txt* files and convert to tables b) adjust table columns and rows (e.g. remove unnecessary columns), c) merge all measurement tables into a global table, and d) join measurement data with the plate assignment metadata (e.g. substance name and concentration).

In order to detect potential cytotoxic substance effects on cell viability (CV), the mean number of nuclei (N) per well (N_{well}) was normalized to the average N (N_{avr}) of the corresponding three Fulv control wells on each plate according to (Malo et al. 2006):

$$CV = \frac{N_{well}^{Substance}}{N_{avr}^{Fulv}} * 100\%$$

Substances leading to a CV $<$ 75% (i.e., representing a $\geq$ 25% reduction of the number of nuclei compared to the 10 nM Fulv control) in at least two out of three runs were assigned to the group of 'Toxic substances'. For substances leading to a CV $\geq$ 75%, the mean E-Cad-GFP signal intensities were further analyzed.

To identify potential estrogenic substances, the mean E-Cad-GFP signal intensity (SI) per well (SI_{well}) was normalized to both the average SI (SI_{avr}) of the corresponding three solvent control wells (SI = 100) AND the three Fulv control wells (SI = 0) on each plate according to (Malo et al. 2006):

$$SI = \frac{SI_{avr}^{Fulv} - SI_{well}^{Substance}}{SI_{avr}^{Fulv} - SI_{avr}^{Solv}} * 100$$

Substances leading to an SI $\geq$ 20 in at least two out of three runs were considered as potential estrogenic substances in the primary screen. The image data of these substances were furthermore visually assessed for ambiguous results and potential imaging artifacts.

For quality assessment, the signal separation (effect size) between the solvent control and Fulv control and the deviation of values within each control group was determined for each plate and run. The Z'-factor (Z') was calculated based on the average SI (SI_{avr}) and the standard deviation (SI_{sd}) of both the three Fulv control and the three solvent control wells on each plate according to (Iversen et al. 2006; Zhang et al. 1999):

$$Z' = \frac{\left(SI_{avr}^{Solvent} - 3SI_{sd}^{Solvent}\right) - \left(SI_{avr}^{Fulv} + 3SI_{sd}^{Fulv}\right)}{SI_{avr}^{Solvent} - SI_{avr}^{Fulv}} = 1 - \frac{3\left(SI_{sd}^{Solvent} + SI_{sd}^{Fulv}\right)}{SI_{avr}^{Solvent} - SI_{avr}^{Fulv}}$$

The acceptance criterion for a valid run was Z' > 0.5.

2.9. Data visualization and statistical analyses

All quantitative data were exported into Excel (Microsoft, Redmond, WA, USA)-readable files. Graphical visualizations and statistical analyses of data were performed using Prism 8 (GraphPad Software, San Diego, CA, USA). Quantitative data were plotted using descriptive statistical indexes, i.e. mean and standard deviation. SI concentration–response curves from the hit confirmation and hit expansion screens were fitted using the non-linear fit algorithm (four parameters, variable hill slope) to calculate half-maximal concentrations (EC50). The Pearson correlation coefficient (r) has been used to measure the strength of association and the direction of the relationship between the determined EC50 values of the substances tested in the E-Morph Assay and the ToxCast ER Agonist Score. Statistical methods used for computational tools are described below in detail. Figures were generated using Illustrator CC 2020 (Adobe, San Jose, CA, USA).

2.10. Computational predictions

The *in silico* toxicity prediction pipeline follows the main steps from the previously published KnowTox project (Morger et al. 2020). In this study, the KnowTox pipeline was slightly adapted for the identification of substances with estrogenic activities and the individual steps are briefly explained in the following. A more detailed description of the underlying concepts is available in the original publication (Morger et al. 2020).

2.10.1. Dataset and preprocessing

2.10.1.1. ToxCast dataset. The publicly available U.S. EPA ToxCast dataset, comprising 8,390 chemicals tested against up to 1,092 endpoints, was downloaded from U.S. EPA's Center for Computational Toxicology and Exposure (U.S. EPA, 2017) and used for training of the *in silico* conformal prediction (CP) models and for the similarity search. In addition to the data preparation steps described in (Morger et al. 2020), canonical SMILES were extracted from the PubChem database using the PubChem PUG REST API (Kim et al. 2015). If no canonical SMILES were available from PubChem, the original ToxCast SMILES were retained.

2.10.1.2. Standardization. First, all instances (molecules and mixtures comprising multiple chemicals) were standardized using the IMI eTOX project standardizer tool (Atkinson 2014) applying the following steps: discard non-organic compounds, neutralize, apply certain structure standardization rules (e.g. handling of tautomers, shifting protons between heteroatoms), neutralize, and remove (mainly organic) salts. Second, standardized molecules and mixture components with less than four heavy atoms, as well as all remaining mixtures, were discarded. This standardized dataset, containing 7,911 molecules, was used as a basis for the similarity search. For subsequent read-across support, measured activities of these molecules in seven ToxCast screening assays covering relevant ER-related endpoints (see Table 4) were considered.

To train CP models on seven estrogen receptor assay datasets, the measured activities (binary) in these assays were assigned to the standardized molecules. Next, each molecule was represented as InChI (International Chemical Identifier), a standardized format, to recognize molecules that appear more than once in the dataset. Chemicals with duplicate InChIs were merged and measurements were aggregated by median (median = 0.5 was discarded). This resulted in a data frame of 7,135 compounds tested on up to seven endpoints.

2.10.1.3. Descriptor calculation. As input for similarity search and CP, descriptors were calculated for all molecules using the RDKit Python library [v.2020.03.1] (Landrum 2006). For similarity search, the circular-environment based Morgan fingerprint (1024 bits, radius 3) and the SMARTS-pattern based MACCS keys (167 bits) were calculated and concatenated. For CP, the above fingerprints were further extended with 200 physicochemical descriptors calculated using the RDKit library. The physicochemical descriptors were normalized based on mean and standard deviation of the physicochemical descriptors of substances for which ToxCast assay data was available. Further molecular descriptors used for the analysis of bisphenols (i.e. MorganCount, MACCS, pharmacophore fingerprints) were also calculated using the RDKit functionalities.

2.10.2. In silico methods

2.10.2.1. Similarity search and read-across support. To support read-across with *in vitro* activity information from similar molecules, a similarity search was implemented using RDKit functionalities. For a query molecule, the Tanimoto similarity to all molecules in the standardized ToxCast data set was calculated, based on the above-described descriptors. The ten most similar compounds were returned together with their Tanimoto similarity and the maximum common substructure with the query molecule.

2.10.2.2. Conformal prediction (CP). CP is a framework on top of a machine learning algorithm, which has the advantage to provide a measure of confidence to the prediction as it includes an additional calibration step (Alvarsson et al. 2021; Norinder et al. 2014; Vovk et al. 2005). Therefore, besides the proper training set, an additional calibration set is needed. By comparing the predictions for a query compound with the predictions already made for the calibration set, the algorithm calculates how well the new prediction conforms to the pre-calculated data points per class (i.e., binary: active and inactive, mondrian classification (Sun et al. 2017)) using calculated p-values.

Given that the training and test data are exchangeable, conformal predictors are designed to conform to a pre-defined maximum error rate. This error rate (significance level) functions as a threshold, so the CP output is a prediction set, which contains all classes for which the p-value is higher than the significance level. For a binary classification problem with classes '0' and '1', the possible prediction outputs are: 'single class' ({0},{1}), 'both class' ({0,1}), or an empty prediction set ({}). A more detailed description of CP, and specifically the use of an additional normalizer model to improve the applicability of the CP models to unseen data, is provided in (Morger et al. 2020).

For each of the ER related endpoints, CP models (nonconformist Python library (Linusson 2015)) were trained and evaluated within a fivefold cross-validation. Thus, the ToxCast data was randomly and stratified split into five parts. In each fold, 80% training and 20% test data were used and an aggregated conformal predictor (ACP) (Carlsson et al. 2014) with 20 loops was initialized. In every ACP loop, the training data was further split into a proper training (70%) and a calibration set (30%) and a random forest model (500 estimators, else default parameters, scikit-learn Python library [v.0.22.2] (Pedregosa et al. 2011)) was trained on the proper training set. The predictions were calibrated using the calibration set, inverse probability error function, and mondrian condition (Sun et al. 2017). Furthermore, the predicted values were normalized using information from the nearest neighbors of the proper set (KNNRegressor [v2.1.0] (Linusson 2015) as described in (Morger et al. 2020)). Median was used to aggregate the p-values from the 20 ACP loops, as recommended by (Linusson et al. 2017). The p-values of the cross-validation were averaged by their mean.

2.10.2.3. CP evaluation. To evaluate the CP ER models and the predictions of the hit expansion compounds, evaluation measures such as

validity, efficiency and accuracy were used. Validity was calculated as the percentage of prediction sets containing the correct class, i.e., the fraction of all 'both class' ({0,1}) and correct 'single class' ({0}, {1}) predictions. Efficiency of the models was calculated as the ratio of prediction sets only containing a single class, i.e., {0} and {1}. Accuracy was determined by the ratio of correct 'single class' classifications compared to all 'single class' predictions.

2.10.2.4. Consensus prediction. To derive a single prediction per compound over all seven CP ER models, the predictions from the individual models were merged into a so-called consensus prediction. Thus, for each prediction, the prediction set was calculated for a maximum accepted error rate of 20%. Only the efficient single class predictions were considered and the mode was calculated following a 'majority vote' principle to obtain one consensus prediction for ER agonism per substance.

2.10.2.5. Comparative docking. Docking is a structure-based modeling technique to predict the preferred orientation of a ligand when it is bound to a protein, which can be mainly divided into a placement and a scoring step (Brooijmans and Kuntz 2003). Docking was performed using the Endocrine Disruptor Monitoring tool (2019 EDMonv3) of the @TOME-2 platform, an inverse screening pipeline that was developed to study interactions between ERα and potentially ER active substances (Pons and Labesse 2009). The EDMonv3/@TOME-2 webserver provides docking of ligands into several endocrine disruption targets (collected in the database "NR_HUMAN_I90_2019M8 (Aug 2019)").

In this study, the binding modes of five bisphenol(-like) ligands (Bisphenol F, 4-Benzylphenol, 4,4''-Dihydroxybiphenyl, 4,4''-Dihydroxybenzophenone, and Bisphenol E) were analyzed when docked to ERα, for which the database contains 221 different protein structures for docking. For the docking experiment, 3D coordinates for each of the five compounds were generated using RDKit, and default parameters were used for the EDMonv3/@TOME-2 screening on ERα (H_NR3A1_ERa).

Out of 221 available structural supports for ERα, the server returned the 20 best complexes per compound. To be able to directly compare the resulting poses for the five compounds, a common target structure that was returned as one of the top 20 complexes for all queries was desired. Twelve crystal structures were returned for all five query compounds consistently. Amongst them, the 3UUA ER agonist PDB structure was chosen, which was already investigated in the analysis of bisphenol-ER interactions by (Delfosse et al. 2012). Docking results, i.e., the crystal structure of 3UUA with the best docked pose of the respective ligand, were downloaded (as *.pdb* file) from the server and further analyzed using LigandScout [v4.4.3] (Wolber and Langer 2005).

2.11. Performance calculations

For each comparison, the overall concordance of active and inactive class predictions, i.e., the proportion of all substances that are correctly classified as active ($N_{True\ Actives}$) or inactive ($N_{True\ Inactives}$) from all tested substances (N), were calculated as follows:

$$\text{Concordance} = \frac{N_{True\ Actives} + N_{True\ Inactives}}{N} * 100$$

The accuracy of active class predictions ($P_{active\ class}$), i.e., the proportion of all substances that are correctly classified as active ($N_{True\ Actives}$) from all substances that are active in the reference method ($N_{True\ Actives} + N_{False\ Inactives}$), were calculated as follows:

$$P_{active\ class} = \frac{N_{True\ Actives}}{N_{True\ Actives} + N_{False\ Inactives}} * 100$$

The accuracy of inactive class predictions ($P_{inactive\ class}$), i.e., the proportion of all substances that are correctly classified as inactive ($N_{True\ Inactives}$) from all substances that are inactive in the reference method ($N_{True\ Inactives} + N_{False\ Actives}$), were calculated as follows:

$$P_{inactive\ class} = \frac{N_{True\ Inactives}}{N_{True\ Inactives} + N_{False\ Actives}} * 100$$

3. Results and Discussion

3.1. E-Cadherin-GFP cell membrane signal intensity as a novel readout to efficiently measure estrogen signaling activity

As described in (Kornhuber et al. 2021), the E-Morph Assay allows the identification and characterization of estrogenic substances based on quantitative changes in the morphology of cell–cell contacts at the level of AJs in the MCF-7/vBOS breast cancer cell line, which occur 24–48 h after exposure to an anti-estrogenic compound such as Fulvestrant (Fulv). In this first description of the assay, the cell–cell contact morphology was visualized using live-cell staining and analyzed by applying a quantitative image analysis pipeline with an integrated classification model (Kornhuber et al. 2021).

In order to improve the HTS capability of the E-Morph Assay and to streamline the visualization and analysis procedures, we now selected a MCF-7 cell line that stably expressed an *E-Cad-GFP* transgene encoding for the mouse E-Cad fused to GFP, which has been shown to actively engage in the AJ assembly, maintenance, and dissociation process (de Beco et al. 2009; 2020). Treatment of this MCF-7/E-Cad-GFP cell line with the anti-estrogen Fulv for 48 h resulted in an AJ phenotype similar to the one observed in the MCF-7/vBOS cell line (Bischoff et al. 2020; Kornhuber et al. 2021) (Fig. 1A), and, in addition, influenced the E-Cad-GFP signal intensity (SI) (Fig. 1A). The SI increased in a concentration-dependent manner upon Fulv treatment (Fig. 1B, light grey) with a mean EC50 of 0.95 nM (Fig. 1C, light grey). In turn, co-treatment with increasing concentrations of 17β-Estradiol (E2) reduced the effect of Fulv on the SI again (Fig. 1B, grey) with a mean EC50 of 32.1 pM (Fig. 1C, grey).

The dependence of the SI on Fulv- or E2-mediated changes of estrogen signaling was verified by gene expression analyses of the estrogen receptor alpha (ERα) target genes *BCL2L1*, *TFF1*, *PGR*, and *AREG*. In line with results from the MCF-7/vBOS cell line (Bischoff et al. 2020), these mRNA expression levels were downregulated or upregulated in MCF-7/E-Cad-GFP cells under anti-estrogenic (Fulv) or estrogenic (Fulv + E2) conditions, whereas *ESR1* (encoding for ERα) itself was not affected (Fig. 1D). Interestingly, the expression of the *E-Cad-GFP* transgene itself clearly increased upon Fulv-treatment on the mRNA and protein level, whereas the expression level of the endogenous human E-Cad was hardly affected (Fig. 1D-F, Fig. S2A-B). The latter observation is again in line with our results from the MCF-7/vBOS cell line (Bischoff et al. 2020), whereas the interesting effect of estrogen signaling on the expression of the *E-Cad-GFP* transgene has not been described before and will be subject of future analyses. As shown by (Bischoff et al. 2020), the described changes in ERα target gene expression levels and the resulting AJ reorganization process occur at different times and differ in their kinetics. Although the authors identified several relevant cellular components that are involved in the phenotype formation process, the precise mechanisms connecting estrogen signaling activity and AJ reorganization is not yet fully understood.

Importantly, these Fulv-induced effects were indeed directly caused by inhibition of ERα activity, since specific depletion of ERα by small interfering RNAs (siRNAs) targeting *ESR1* sufficed for the formation of the AJ phenotype, the increased SI, and the elevated *E-Cad-GFP* expression levels (Fig. 1G-H; Fig. S2B). ERα activity was effectively depleted by si*ESR1* as indicated by the reduction of *AREG* and the increase of *BCL2L1* ERα target gene expression levels (Fig. S2B-C). Moreover, the application of Fulv to si*ESR1* knockdown cells only slightly further increased the SI and the *E-Cad-GFP* expression level (Fig. 1G-H). Likewise, addition of E2 could also not rescue these si*ESR1*-mediated effects (Fig. 1G-H).

Together, these data support the conclusion that the MCF-7/E-Cad-

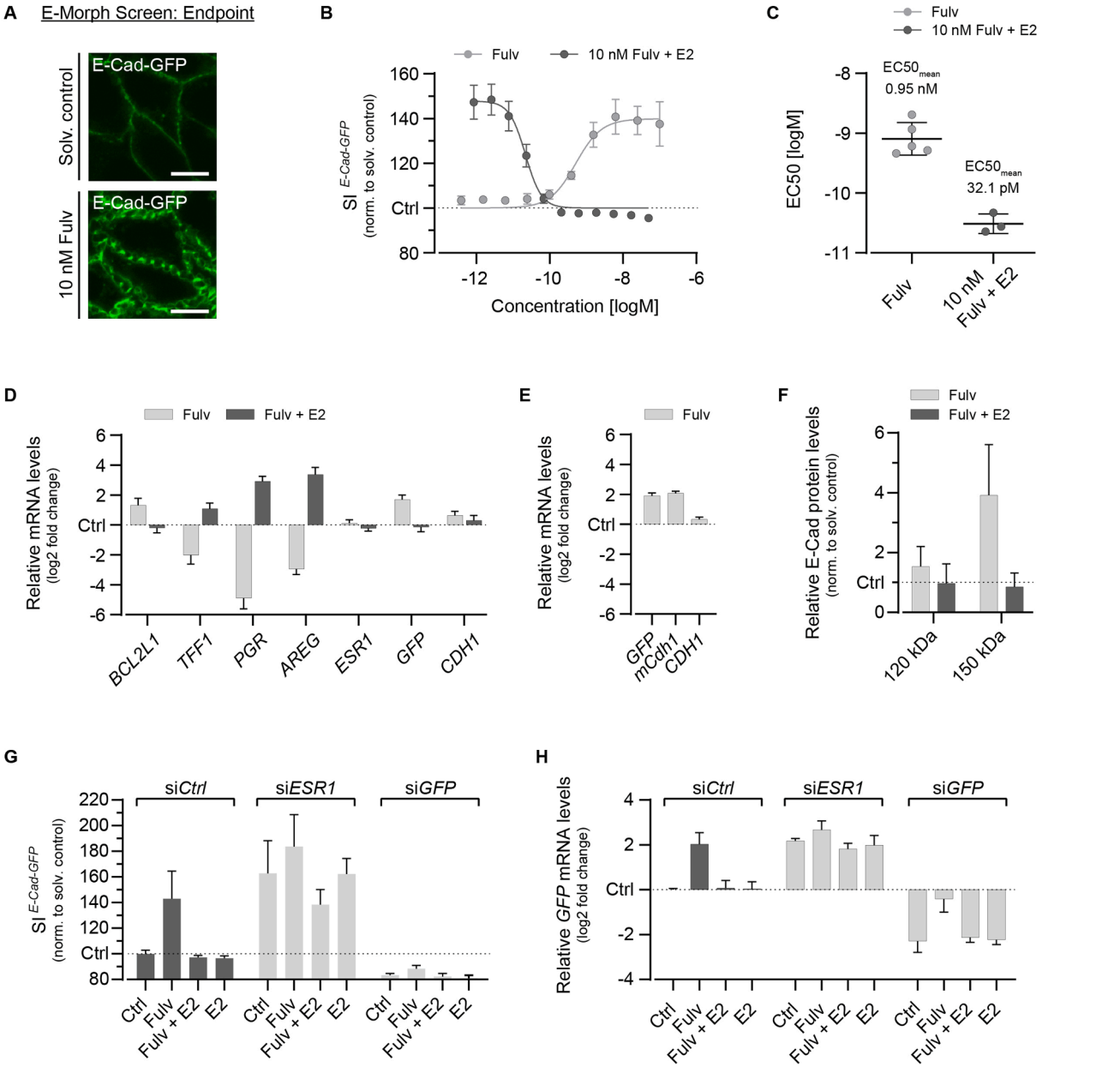


Fig. 1. Visualization and quantification of E-Cad-GFP signal intensity (SI) along the cell membrane as an endpoint for estrogenic activity in the E-Morph Screening Assay. (A) Fluorescence images showing E-Cad distribution and intercellular spacing in MCF-7/E-Cad-GFP cells upon Fulvestrant (10 nM Fulv) treatment for 48 h as compared to the solvent control. Scale bars, 10 μM. (B) Representative concentration–response curves from quantitative image analysis of E-Cad-GFP expressing cells under anti-estrogenic (Fulv titration) and estrogenic (10 nM Fulv + 17β-estradiol (E2) titration) conditions (treatment for 48 h). The plot depicts the relative E-Cad-GFP signal intensity, which increases under anti-estrogenic conditions and decreases under estrogenic conditions. Signal intensities are normalized to the solvent control (Ctrl). Non-linear fit (four parameters, variable hill slope, bottom constrained to Ctrl). Biological replicates, n = 1. Error bars, mean +/- SD from ≥ 3 technical replicate experiments. (C) Mean half-maximal concentrations ($EC50_{mean}$) derived from dose-response curves as described in Fig. 1B. Non-linear fit (four parameters, variable hill slope, bottom constrained to the solvent control). Biological replicates, n = 5 (Fulv titration, normalized to the solvent control) and n = 3 (10 nM Fulv + E2 titration, normalized to the solvent control and the 10 nM Fulv control). (D) Quantitative PCR measurement of mRNA expression levels of typical ERα target genes (*BCL2L1*, *TFF1*, *PGR*, *AREG*), *ESR1*, *GFP*, and *CDH1* under anti-estrogenic (10 nM Fulv) and estrogenic (10 nM Fulv + 10 nM E2) conditions (treatment for 48 h). Relative mRNA expression levels for each treatment condition are normalized to the solvent control (Ctrl). Biological replicates, n ≥ 3. Error bars, mean + SD. (E) Quantitative PCR measurement of mRNA expression levels of *GFP*, murine *Cdh1* and human *CDH1* under anti-estrogenic (10 nM Fulv) conditions (treatment for 48 h). Relative mRNA expression levels for each treatment condition are normalized to the solvent control (Ctrl). Biological replicates, n = 3. Error bars, mean + SD. (F) Quantification of protein expression levels of endogenous E-Cad (120 kDa) and transgenic E-Cad-GFP (150 kDa) bands from chemiluminescence western blots shown in Fig. S2A. Relative protein expression levels under anti-estrogenic (10 nM Fulv) and estrogenic (10 nM Fulv + 10 nM E2) conditions (treatment for 48 h) are normalized to the solvent control (Ctrl). Biological replicates, n = 3. Error bars, mean + SD. Loading control, Coomassie total protein staining. G) Quantification of E-Cad-GFP signal intensities from cells transfected with *ESR1* siRNA or *GFP* siRNA compared to cells transfected with scrambled control siRNA (si*Ctrl*) for 72 h. Cells were additionally grown under anti-estrogenic (10 nM Fulv) and estrogenic (10 nM Fulv + 10 nM E2 or 10 nM E2 alone) conditions (treatment for 48 h). Relative E-Cad-GFP signal intensities are normalized to cells treated with scrambled control siRNA and the solvent control (Ctrl). Biological replicates, n = 2. Error bars, mean + SD. (H) Quantitative PCR measurement of *GFP* mRNA expression levels of data shown in Fig. 1G. Relative mRNA expression levels are normalized to cells treated with scrambled control siRNA (si*Ctrl*) and the solvent control (Ctrl). Biological replicates, n = 2. Error bars, mean + SD.

GFP cell line can adequately replace the original MCF-7/vBOS cell line in the E-Morph Screening Assay and that the SI represents a novel and reliable readout for estrogenic activity. This readout further simplifies quantitative image analysis pipelines because it does not require training of a supervised machine learning algorithm for image classification such as in the original E-Morph Assay readout.

3.2. Automated high-throughput screening for estrogenic substances

Besides implementing the novel primary readout of the E-Morph Screening Assay, we adapted and optimized the cell treatment and staining procedure for automated handling of 96- and 384-well-plates using a robotic platform. To speed up the image data analysis and evaluation procedure, we further improved the automated imaging and quantitative image analysis pipeline and built an automated image data evaluation pipeline using the KNIME software (Fig. S1A-B, File S1). In this KNIME workflow, substance-related cell death or altered cell proliferation were automatically detected by counting the number of nuclei as a readout for the number of cells. In a next step, the measured SI was normalized to both the solvent control (SI = 100) and the 10 nM Fulv control (SI = 0) for cut-off-based classification of substances with potential estrogenic activity (ER activity: SI ≥ 20, i.e., representing a ≥ 20% reduction of the Fulv-mediated increase in E-Cad-GFP membrane signal intensity).

We then applied this automated HTS pipeline to screen a substance library comprising 430 toxicologically-relevant industrial chemicals, biocides, and plant protection products, as well as reference substances with already known, specific activities on different nuclear receptor

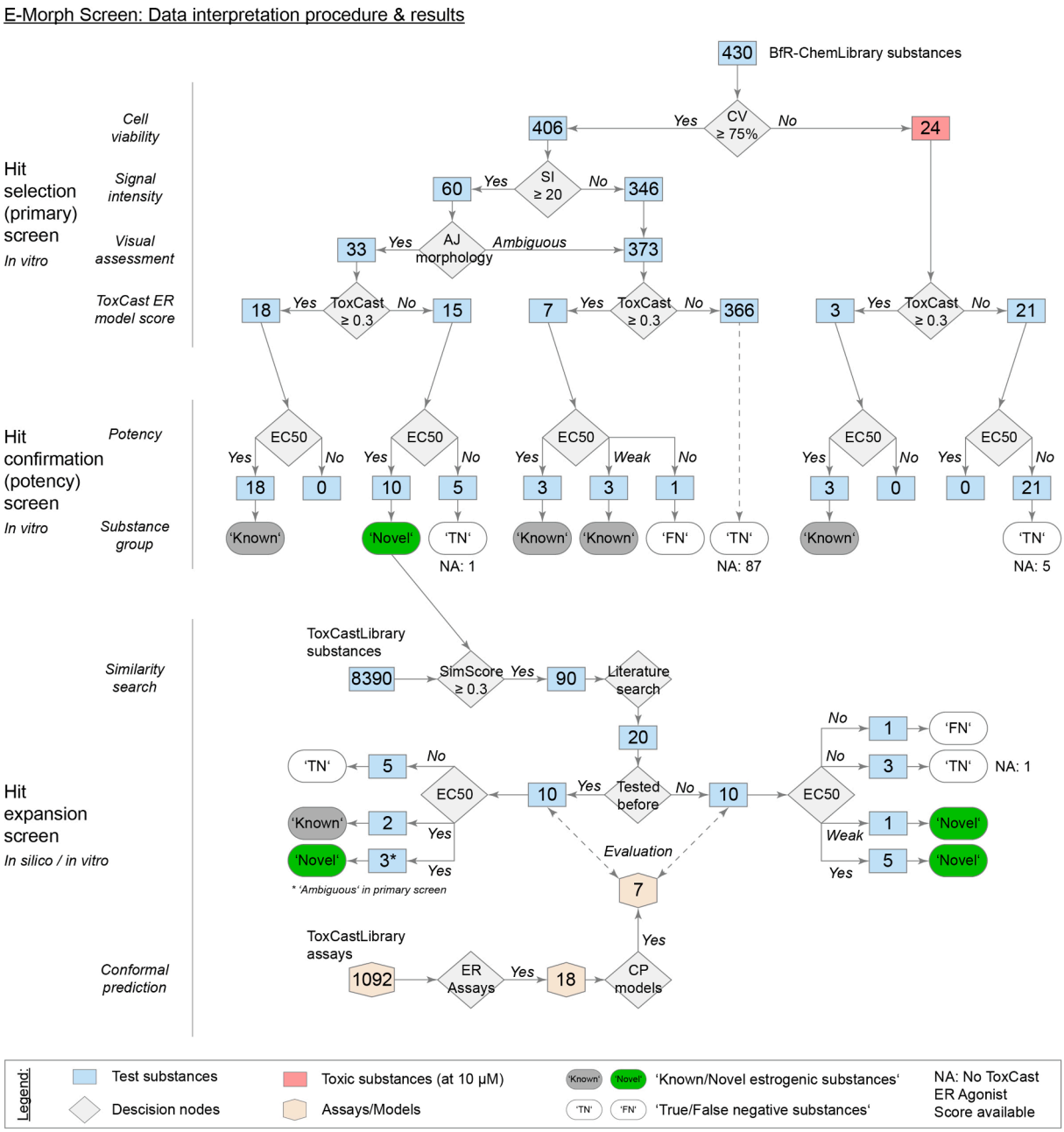


Fig. 2. E-Morph screen workflow and results. Decision tree describing the data interpretation procedure and the results of the three consecutive E-Morph screens involving *in vitro* (E-Morph Screening Assay) and *in silico* (similarity search, conformal prediction) methods to identify substances with estrogenic activity. See main text for details. Rectangular boxes, numbers of substances. Hexagonal boxes, numbers of assays/models. Oval boxes, substance groups based on comparison to the published ToxCast ER Agonist Score (see Fig. 3).

signaling pathways (Filer et al., 2014; Wetzel et al., 2017; EFSA, 2015; OECD, 2017; U.S. NIEHS, 2019). A complete list of these BfR-ChemLibrary substances and the corresponding screening results are available in Supplementary Table S1. The data interpretation procedure for identification of estrogenic substances and the screening results are summarized in Fig. 2. The quality and performance of the screens was furthermore evaluated in the KNIME workflow based on a commonly used statistical parameter, i.e., the Z'-factor (Z'). Each run achieved a Z'-factor > 0.5, which indicates a very robust HTS assay according to (Iversen et al. 2006; Zhang et al. 1999) and demonstrates the applicability of the E-Morph Screening Assay for HTS purpose. Notably, (Kornhuber et al. 2021) already compared the robustness of the selected 48 h time point with a shorter treatment period of 30 h and concluded that the effect size of the assay declined when the test chemical exposure time was reduced.

3.3. Primary screen and hit selection

In the primary screen (Fig. 2), we tested the 430 substances at a single concentration of 10 μM in the presence of 10 nM Fulv for 48 h in three independent runs. Of those, 24 substances led to a significantly reduced cell viability (CV) < 75% (Figs. 2 and 3, red) and were therefore subsequently re-tested at lower concentrations (1 pM - 30 μM) in the hit confirmation screen as described below. We identified 60 potential hit substances (SI ≥ 20), of which 33 substances clearly influenced the characteristic estrogen-dependent AJ morphology in a similar way as compared to the Estrone reactivity control (visual assessment), corresponding to an overall hit-rate of 7.7% (Fig. 2). The other 27 substances did either not clearly influence the AJ phenotype or caused rather unrelated changes in fluorescence intensity and were therefore first considered 'inactive' but flagged as 'ambiguous' in Table S1. Comparing the results for the 33 clear hit substances with the published ToxCast ER Agonist score (Browne et al. 2015; Judson et al. 2015) identified 18 substances with a ToxCast ER Agonist Score ≥ 0.3 that were considered verified actives in the primary screen (Figs. 2 and 3, grey). The ToxCast ER Agonist Score of the remaining 15 hit substances was < 0.3 or 'not available' (NA) indicating potential yet undescribed estrogenic activity (Figs. 2 and 3, green). Notably, seven of the 373 substances that were first classified as 'inactive' had a ToxCast ER Agonist Score ≥ 0.3 and were therefore included in the subsequent hit confirmation screen to be tested at higher concentrations.

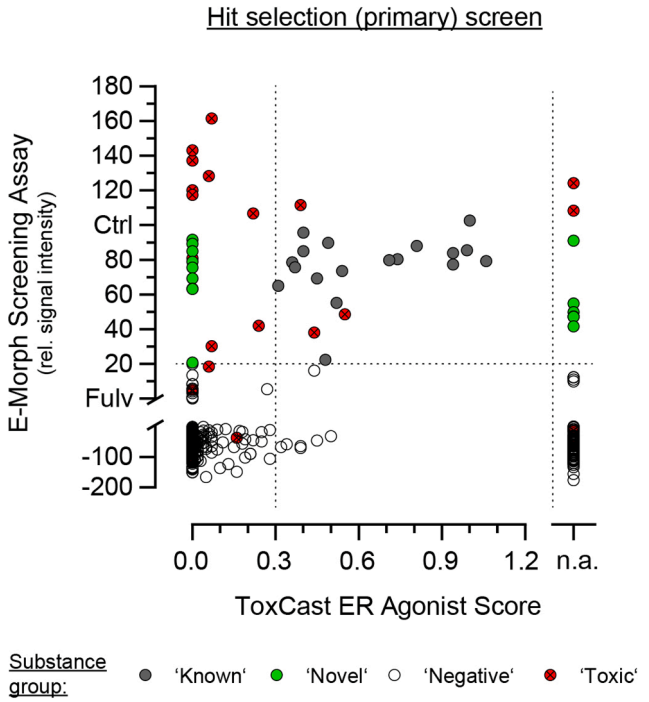


Fig. 3. Primary screening and hit selection. Relative E-Cad-GFP signal intensities of 430 test substances that were measured in the hit selection (primary) screen as compared to the published ToxCast ER Agonist Score. Cells were exposed to 10 nM Fulv + 10 μM test substance for 48 h. Each data point represents the mean relative signal intensity obtained from three independent runs. Relative E-Cad-GFP signal intensities are normalized to the solvent control (Ctrl, SI = 100) and the 10 nM Fulv control (Fulv, SI = 0). Substances that induce an increase of the relative signal intensity above the assay threshold (horizontal dashed line, SI ≥ 20) are considered as primary hit substances. Hit substances with a ToxCast ER Agonist Score ≥ 0.3 (vertical dashed line) are depicted in grey color and assigned to the group of 'Known estrogenic substances'. Hit substances with a ToxCast ER Agonist Score < 0.3 or 'not available' (NA) (vertical dashed line) are highlighted in green color and assigned to the group of potential 'Novel estrogenic substances'. Test substances leading to a cell viability (CV) < 75% are indicated in red color and assigned to the group of 'Toxic substances'. Data from substances that were excluded after visual assessment of images are not displayed. Biological replicates, n = 3. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

3.4. Hit confirmation screen and potency determination

In the hit confirmation screen (Fig. 2), we re-tested the 33 primary hit substances (SI ≥ 20), the 24 substances displaying cytotoxicity at 10 μM, and the seven 'inactive' substances with a ToxCast ER agonist score ≥ 0.3 (in total 64 substances) at multiple concentrations ranging from 1 pM to 30 μM in the presence of 10 nM Fulv for 48 h in multiple independent runs. A clear concentration-dependent estrogenic activity was detected for 28 out of the 33 primary hit substances (SI ≥ 20), three out of the 24 'cytotoxic' substances (CV < 75%) when tested at concentrations < 10 μM, and six out of the seven 'inactive' substances (ToxCast ER agonist score ≥ 0.3) when tested at concentrations > 10 μM. The remaining substances did not show a clear activity in the tested concentration range. These data show that the results of the primary screen and the hit confirmation screen were concordant to a large extent and highlight the need of testing substances in a wide concentration range to increase the hit rate. Based on the respective ToxCast ER Agonist Score, we grouped the active substances as 'Known estrogenic substances' (≥0.3; 27 substances) (Fig. 2, grey; Table 1) or potential 'Novel estrogenic substances' (<0.3 or not available ('NA'); 10 substances) (Fig. 2, green; Table 2). Accordingly, the inactive substances were grouped as 'True negative substances' ('TN') (<0.3; 299 substances), 'False negative substances' ('FN') (≥0.3; 1 substance), or 'NA' if no ToxCast ER Agonist Score was available (93 substances) (Fig. 2, white; Table S1).

Using the concentration–response data that was collected in the hit confirmation screen, we further determined the potencies (EC50) and relative ER bioactivities (logEC50 normalized to 17-alpha-Ethinylestradiol) for 24 out of the 27 'Known estrogenic substances' (Fig. 2, {'Known','Yes'}; Fig. 4A; Table 1), which correlated well with the respective ToxCast ER Agonist Scores (r = +0.95) (Fig. 4A-B; Table 1). No EC50 values could be determined for the remaining three substances, including Tamoxifen, which showed only weak estrogenic activity in the hit confirmation screen (Fig. 2, {'Known','Weak'}; Table 1). The weak estrogenic activity of Tamoxifen might reflect its partial agonistic function (Jordan 1977), which is further supported by its estrogenic activity in the uterotrophic bioassay in rodents (Kleinstreuer et al. 2016). Overall, the measured activities were also concordant with the CERAPP ER consensus model predictions (Table 1). Among the substances with a ToxCast ER Agonist Score ≥ 0.3, only 2,4′-DDT may have been a potential false negative substance in the E-Morph Screen (Fig. 2, 'FN'; Table 1). This particular chemical appears to be difficult to detect in the tested concentration range as it also showed weak or no activity in other ER testing systems according to the Integrated Chemical Environment database (Bell et al. 2020; Bell et al. 2017) of the U.S. National Toxicology Program.

Table 1
Screening results for the group of 'Known estrogenic substances' compared to published *in silico* ER model data from the U.S. EPA.

Chemical name	CAS No.	U.S. EPA *in silico* ER models		E-Morph Screening Assay	Hit confirmation screen				
		ToxCast ER Agonist Score[a)]	CERAPP ER Agonist Model[b)]	Substance group	Potency [M]			ER Bioactivity [rel. LogEC50]	Comment
					EC50	SD	n		
Diethylstilbestrol	56-53-1	0.94	active	Known	7.97E-10	6.99E-10	3	1.03	active
Beta-Estradiol	50-28-2	0.94	active	Known	9.00E-10	1.30E-10	3	1.03	active
Hexestrol	84-16-2	0.99	active	Known	1.05E-09	2.59E-10	4	1.02	active
Ethinyl Estradiol	57-63-6	1.00	active	Known	1.56E-09	3.04E-10	3	1.00	active
Mestranol	72-33-3	0.74	active	Known	2.26E-09	7.13E-10	4	0.98	active
Estrone	53-16-7	0.81	active	Known	3.18E-09	2.83E-09	4	0.96	active
Alpha-Estradiol	57-91-0	1.06	active	Known	5.50E-09	1.74E-09	4	0.94	active
Zearalenone	17924-92-4	0.71	active	Known	1.66E-07	6.01E-08	4	0.77	active
Bisphenol AF	1478-61-1	0.55	active	Known	2.96E-07	1.47E-07	3	0.74	active
Trenbolone-Dea	10161-33-8	0.48	active	Known	4.30E-07	1.70E-07	3	0.72	active
Genistein	446-72-0	0.54	active	Known	4.35E-07	4.00E-07	4	0.72	active
Norethisterone	68-22-4	0.52	active	Known	7.71E-07	4.22E-07	4	0.69	active
Bisphenol B	77-40-7	0.49	active	Known	1.07E-06	5.16E-07	4	0.68	active
5alpha-Androstan-17beta-OL-3ON	521-18-6	0.40	active	Known	1.28E-06	1.59E-06	4	0.67	active
Nonylphenol techn Gemisch	84852-15-3	0.44	active	Known	1.39E-06	1.77E-07	2	0.67	active
Dehydroisoandrosterone	53-43-0	0.37	active	Known	1.54E-06	6.70E-07	4	0.66	active
4-tert.-Octylphenol	140-66-9	0.39	active	Known	1.80E-06	5.59E-07	2	0.65	active
Bisphenol A	80-05-7	0.45	active	Known	2.23E-06	4.10E-07	3	0.64	active
Biochanin A	491-80-5	0.36	active	Known	3.09E-06	1.05E-06	4	0.63	active
Daidzein	486-66-8	0.44	active	Known	3.10E-06	9.26E-07	2	0.63	active
2,2',4,4'-Tetrahydroxybenzophenon	131-55-5	0.40	active	Known	4.26E-06	2.32E-06	4	0.61	active
17alpha-Hydroxyprogesterone	68-96-2	0.34	active	Known	8.22E-06	4.64E-06	2	0.58	active
Levonorgestrel	797-63-7	0.39	active	Known	8.79E-06	2.99E-06	2	0.57	active
Apigenin	520-36-5	0.31	active	Known	2.16E-05	2.16E-05	4	0.53	active
1,1,1-Tris(4-hydroxyphenyl) ethane	27955-94-8	0.32	active	Known	NA	NA	1	NA	weakly active at 30 µM (22,3 % rescue)
2,4'-DDT Lösung	789-02-6	0.39	active	FN	NA	NA	1	NA	negative
17a-Methyltestosterone	58-18-4	0.50	active	Known	NA	NA	1	NA	weakly active at 10 µM (21,3 % rescue)
Tamoxifen	10540-29-1	0.45	inactive	Known	NA	NA	1	NA	weakly active at 10 µM (26,6 % rescue)

Overall classifications and potencies of 28 substances with a ToxCast ER Agonist Score ≥ 0.3. EC50, mean potency from multiple independent runs (n). SD, standard deviation. ER Bioactivity, potency (logEC50) normalized to 17-alpha-Ethinylestradiol (1.00). NA, not available/not applicable. FN, false negative substance.
[a)]Screening Chemicals for Estrogen Receptor Bioactivity Using a Computational Model (Browne et al. 2015)
[b)]CERAPP: Collaborative Estrogen Receptor Activity Prediction Project (Mansouri et al. 2016)

Potencies (EC50) and relative ER bioactivities were also determined for all 10 substances that were assigned to the group of 'Novel estrogenic substances' (Fig. 2, 'Novel'; Fig. 4C; Table 2). The calcium chelating agent EDTA was considered a false positive substance because of its known interference with the Ca + -dependent E-Cad activity and therefore excluded from further analysis. The two most potent substances in this hit class were the pesticide Hexythiazox (insecticide) with an EC50 of 10 nM, for which no estrogenic activity has been described before, and the progestin Norethisterone acetate (NETA) with an EC50 of 625 nM. The remaining substances showed weaker activities in the µM range. The estrogenic activities of NETA, Nandrolone (androgen and anabolic steroid (AAS)), Phloretin (flavonoid), and Bisphenol F (industrial chemical), for which no ToxCast ER Agonist Score was available, were consistent with previous studies (Branham et al. 2002; Chwalisz

Table 2
Screening results for the group of 'Novel estrogenic substances' compared to published *in silico* ER model data from the U.S. EPA.

Chemical name	CAS No.	U.S. EPA *in silico* ER models		E-Morph Screening Assay	Hit confirmation screen				
		ToxCast ER Agonist Score [a)]	CERAPP ER Agonist Model [b)]	Substance group	Potency [M]			ER Bioactivity [rel. LogEC50]	Comment
					EC50	SD	n		
Hexythiazox	78587–05-0	0.00	inactive	Novel	1.01E-08	3.94E-09	4	0.91	active
Norethindrone acetate (NETA)	51–98-9	NA	active	Novel	6.25E-07	4.62E-07	6	0.70	active
EDTA iron(III) sodium salt	15708–41-5	0.00	inactive	Novel	1.10E-06	7.34E-07	3	0.68	active
Nandrolone	434–22-0	NA	active	Novel	2.04E-06	1.38E-06	6	0.65	active
Phloretin	60–82-2	NA	active	Novel	3.36E-06	1.41E-06	5	0.62	active
2,4,6-Tri-*tert*-butylphenol (TTBP)	732–26-3	0.00	inactive	Novel	4.73E-06	3.12E-06	4	0.60	active
Bisphenol F	620–92-8	NA	active	Novel	4.79E-06	1.41E-06	5	0.60	active
Diuron	330–54-1	0.00	inactive	Novel	6.04E-06	2.91E-06	3	0.59	active
Azoxystrobin	131860–33-8	0.00	inactive	Novel	6.34E-06	3.68E-06	6	0.59	active
Zineb	12122–67-7	NA	inactive	Novel	8.71E-05	8.48E-05	2	0.46	active

Overall classifications and potencies of 10 substances that were active in the E-Morph Screening Assay with a ToxCast ER Agonist Score = 0.00 or 'not available' (NA). EC50, mean potency from multiple independent runs (n). SD, standard deviation. ER Bioactivity, potency (logEC50) normalized to 17-alpha-Ethinylestradiol (1.00). NA, not available/not applicable.
[a)]Screening Chemicals for Estrogen Receptor Bioactivity Using a Computational Model (Browne et al. 2015)
[b)]CERAPP: Collaborative Estrogen Receptor Activity Prediction Project (Mansouri et al. 2016)

et al. 2012; Rochester and Bolden 2015; Sirianni et al. 2012), results from ER transactivation screening assays (Table S4), as well as the CERAPP ER consensus model (Table 2). Nandrolone (19-nortestosterone) was further shown to be active in the STTA and VM7Luc ER transactivation assays of OECD TG 455 (OECD 2016). For 2,4,6-TTBP (industrial chemical), Azoxystrobin (fungicide), Hexythiazox, and Diuron (herbicide), the ToxCast ER Agonist Score was 0.00 and they were also inactive in the relevant ER screening assays (Table S4) as well as the CERAPP ER consensus model (Table 2). Furthermore, these substances were also classified as 'non-binder' in the FW and CERI ER binding assays of OECD TG 493 (OECD 2015). In addition, for the fungicide Zineb (Zink-ethylen-1,2-bis-dithiocarbamat) no conclusive data demonstrating estrogenic activity was available, yet (Table S4).

These partially discordant results between the E-Morph Screening Assay, the ToxCast ER HTS data, and the *in silico* ER models (ToxCast and CERAPP) can have various reasons (including false positive results), but may also reflect that, in contrast to the simplifying ER HTS assays conducted in the ToxCast project, the functional E-Morph Screening Assay integrates multiple interacting cellular pathways. On the one hand, it therefore provides a more complete picture of relevant cellular mechanisms mediating estrogen-dependent effects. On the other hand, integration of multiple mechanistic events or cellular signaling pathways in a single assay increases the degrees of freedom for possible modes of action of test substances and necessitates running secondary assays to confirm substance-specific effects on distinct signaling pathways.

3.5. Verification of 'Novel estrogenic substances'

In order to verify the nine (excluding EDTA) potential 'Novel estrogenic substances' (Table 2), we first determined the mRNA expression levels of the ERα target genes *BCL2L1*, *TFF1*, *PGR*, and *AREG* along with *ESR1* and *CDH1*. Cells were exposed to each test substance at a concentration of 10 μM in the presence of 10 nM Fulv for 48 h and the effects were compared to the mRNA expression profiles under anti-estrogenic (Fulv) and estrogenic (Fulv + E2) conditions (Fig. 5A; Fig. S2D). Hexythiazox and NETA showed the most similar expression profiles when compared to the E2 reference substance, which was in line with the high potency that was measured in the E-Morph Screening Assay. Nandrolone, Phloretin, and Diuron also showed an estrogenic expression profile, albeit to a weaker extent. Bisphenol F slightly inhibited the Fulv effect, particularly for *TFF1* expression. Importantly, these effects on gene expression profiles could be confirmed when cells were only exposed to these test substances without Fulv co-treatment (Fig. 5B; Fig. S2E). Interestingly, Bisphenol F showed the strongest effect in this case. The expression profiles of the phenol 2,4,6-TTBP and the pesticide Zineb did not support an estrogenic effect neither in the competitive treatment nor in the single treatment scenario (Fig. 5A-B; Fig. S2D-E). The gene expression pattern of Azoxystrobin rather showed some surprising anti-estrogenic effect, particularly for *TFF1* and *PGR*, when applied to cells without Fulv (Fig. 5B; Fig. S2E).

To characterize the underlying mechanism of action of the potential 'Novel estrogenic substances' (Table 2), we next performed a 'pBIND-ERα vector assay', which allows the identification of substances that directly bind to the ERα ligand binding domain. In this assay, MCF-7/E-Cad-GFP cells were transiently transfected with both the reporter and control vector plasmids and subsequently treated with each test substance at 10 μM for 48 h (Fig. 5C). The results were very similar to the detected gene expression profiles. Hexythiazox and NETA showed the highest activity in this assay, whereas Nandrolone, Phloretin, Diuron, and Bisphenol F showed weaker effects. Again, an estrogenic activity at the level of ERα binding could not be identified for 2,4,6-TTBP, Zineb, and Azoxystrobin.

Together, these secondary assay data support the detected estrogenic activity of Hexythiazox, NETA, Nandrolone, Phloretin, Diuron, and Bisphenol F from the group of potential 'Novel estrogenic substances' that were identified by the E-Morph Screening Assay (Fig. 2; Table 2). These data further underline the importance of running secondary assays to identify potential false positive substances, i.e., 2,4,6-TTBP, Zineb, and Azoxystrobin. These substances might act on E-cadherin or AJs in an estrogen-independent manner that will be addressed in future analyses.

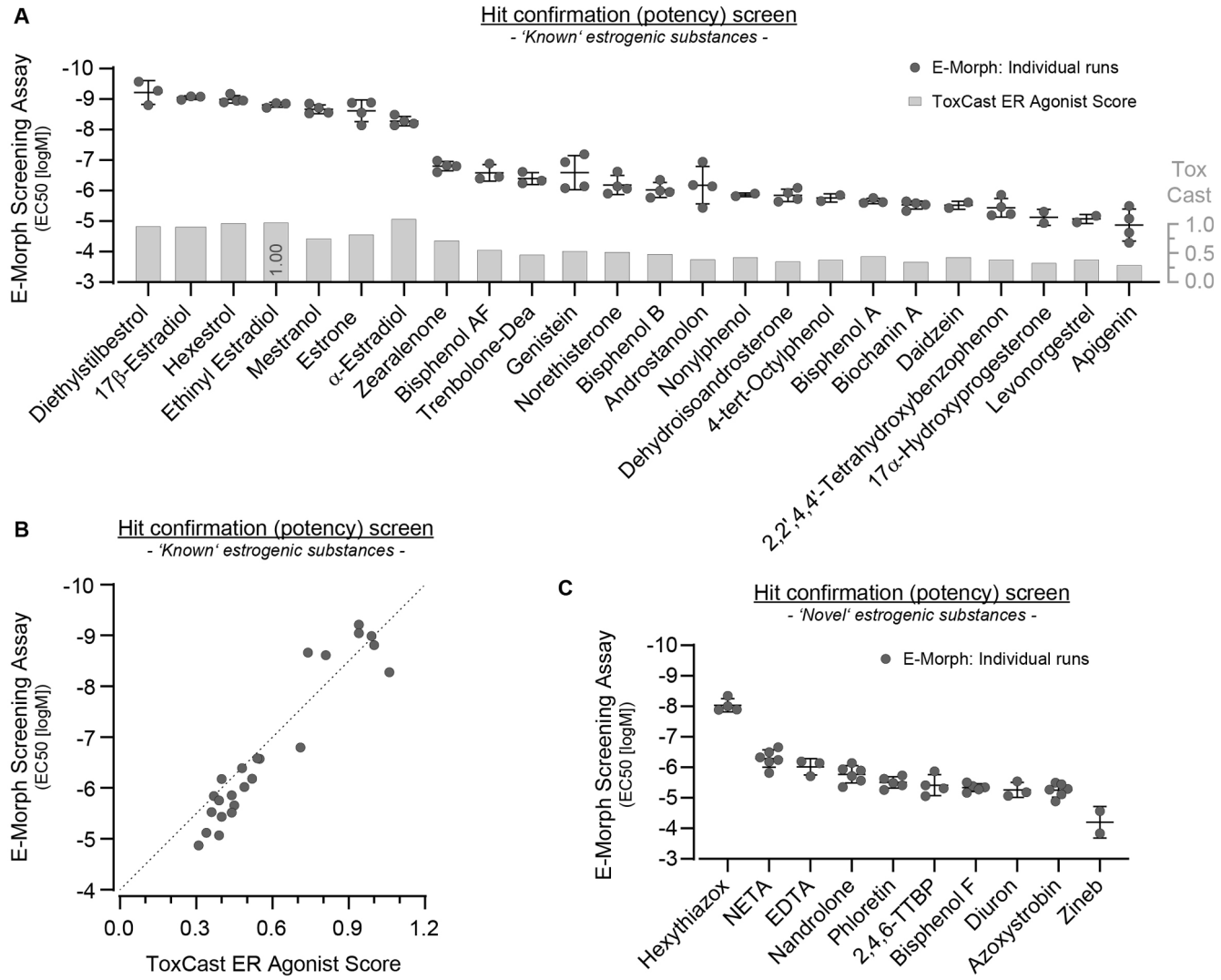


Fig. 4. Hit confirmation screening and potency determination. A) Potencies (half-maximal concentrations, EC50) of 24 'Known' hit substances that were active in the hit confirmation (potency) screen with a ToxCast ER Agonist Score ≥ 0.3. Each data point represents an individual run. The three 'Known' hit substances 1,1,1-Tris-(4-hydroxyphenyl)-ethan, Methyltestosterone, and Tamoxifen showed only weak activity (no EC50 value could be determined) and are therefore not shown. Biological replicates, n ≥ 3. Error bars, mean +/- SD. (B) Correlation between the potencies (half-maximal concentrations, EC50) of 24 'Known' hit substances obtained from the E-Morph Assay and the published ToxCast ER Agonist Score. Each data point represents the mean of the relative bioactivities obtained from individual runs shown in Fig. 4A. The contour line indicates full correlation. Pearson r = +0.95. (C) Potencies (half-maximal concentrations, EC50) of 10 'Novel' hit substances that were active in the hit confirmation (potency) screen with a ToxCast ER Agonist Score = 0.00 or 'not available' (NA). Each data point represents an individual run. Biological replicates, n ≥ 3. Error bars, mean +/- SD.

3.6. Identification of structurally similar substances and hit expansion screening

Based on the assumption that structurally similar substances can interact with similar targets (Bender and Glen 2004), we performed an *in silico* similarity search against the substances for which ToxCast assay data was available (Fig. 2). The chemical structures of the nine (excluding EDTA) potential 'Novel estrogenic substances' (Table 2) were used as input for the identification of other structurally similar substances (Table S2; File S2). Based on the resulting similarity scores (Tanimoto index) and literature search, we selected a final set of 20 similar substances and measured their potential estrogenic activity and potency in the E-Morph Screening Assay (Fig. 6; Table 3). Notably, we set a relatively low global Tanimoto cut-off (>0.3) for inclusion of similar substances into the hit expansion screening in order to account for the diverse structural complexities of the input substances. Furthermore, the number of selected similar substances per input substance varied because of the composition of the ToxCast substance library. For example, the ToxCast database contains assay data for many different bisphenols with relatively high similarity scores (>0.5) to Bisphenol F, but no data for substances that are similar to the pesticide Hexythiazox with a score above 0.4 (Table 3; Table S2; File S2).

Of the 20 similar substances, 10 substances had already been tested in the primary or hit confirmation screens (Fig. 2; Table 3) and their re-testing in the hit expansion screening (5 active, 5 inactive) provided concordant results (Table S1). Interestingly, three of these substances (4,4′-Dihydroxybiphenyl (92–88-6), 4,4′-Dihydroxybenzophenone (611–99-4), Triclocarban (101–20-2)) were initially considered as 'ambiguous' by visual inspection in the primary screen at 10 μM (Fig. 2; Table S1). However, based on the hit expansion screening of a wider concentration range, these substances could now be re-assignment to the group of 'Novel estrogenic substances' (Table 3; Table S1). Hence, a visual inspection of the AJ phenotype at a single concentration also bears the risk of misinterpretation of substance effects, particularly at concentrations near the cytotoxic range. Of the remaining 10 newly tested substances, six substances were active and four substances were

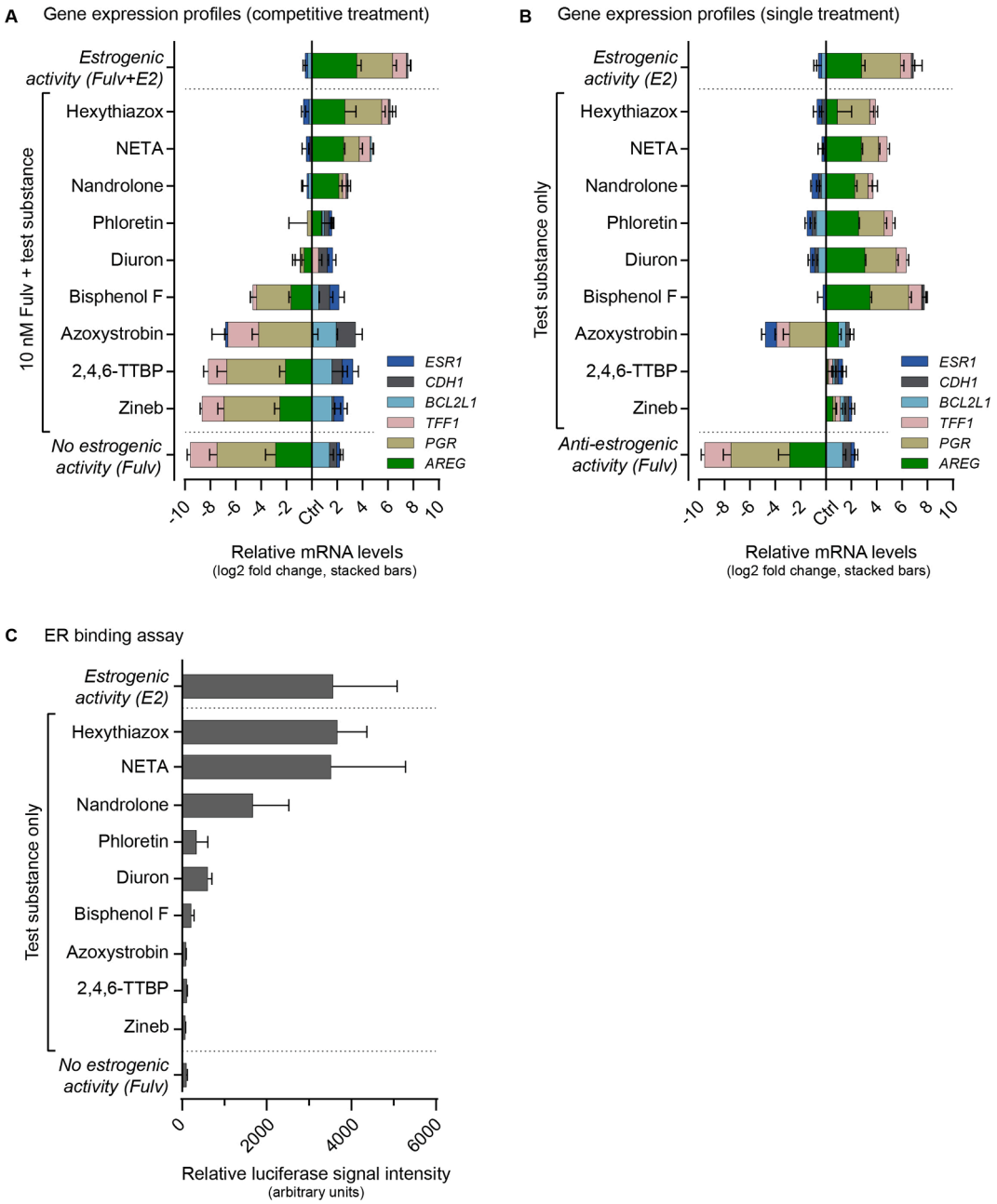


Fig. 5. Verification of 'Novel' hit substances. (A) Gene expression profiles from quantitative PCR measurements of *ESR1*, *CDH1*, and typical ERα target genes (*BCL2L1*, *TFF1*, *PGR*, *AREG*). Cells were exposed to 10 nM Fulv + 10 μM test substance (competitive treatment) for 48 h. Measurements from estrogenic (10 nM Fulv + 10 nM E2, top) and anti-estrogenic (10 nM Fulv, bottom) conditions provide reference gene expression profiles. Relative mRNA expression levels are normalized to cells treated with the solvent control (Ctrl). Biological replicates, n = 3. (B) Gene expression profiles from quantitative PCR measurements of *ESR1*, *CDH1*, and typical ERα target genes (*BCL2L1*, *TFF1*, *PGR*, *AREG*). Cells were exposed to 10 μM test substance only (single treatment) for 48 h. Measurements from estrogenic (10 nM E2, top) and anti-estrogenic (10 nM Fulv, bottom) conditions provide reference gene expression profiles. Relative mRNA expression levels are normalized to cells treated with the solvent control (Ctrl). Biological replicates, n = 3. C) Relative luciferase signal intensities obtained from an ERα reporter gene assay. Cells were co-transfected for 72 h with a pBIND-ERα expression vector (Gal4-DBD fused to ERα-LBD, Renilla luciferase) and a target vector expressing a UAS-controlled Firefly luciferase. Cells were exposed to 10 μM test substance only for 48 h. The detected Firefly luminescence was normalized to Renilla luminescence. Relative signal intensities from estrogenic (10 nM E2, top) and anti-estrogenic (10 nM Fulv, bottom) conditions provide reference measurements. Biological replicates, n = 3. Error bars, mean + SD.

inactive (Fig. 2; Table 3, Table S1). Interestingly, Norgestrel, a racemate of D-Norgestrel and L-Norgestrel/Levonorgestrel enantiomers (Kuhl 2005), was inactive in the hit expansion screen despite a ToxCast ER Agonist Score of 0.39 (Table 3, 'FN'), whereas the known active enantiomer Levonorgestrel by itself was active in our assay (Table 3, 'Known'). Thus, selecting candidate substances based on structural similarities to our primary hits, resulted in the identification of in total nine additional 'Novel estrogenic substances'. The measured estrogenic activities for most of these substances were supported by the CERAPP ER consensus model (Table 3).

Together, these screening data provide strong support for an estrogenic activity of NETA, Nandrolone, Phloretin, and Bisphenol F and demonstrate that the combination of *in silico* similarity search and *in vitro* testing supports the identification of estrogenic activities.

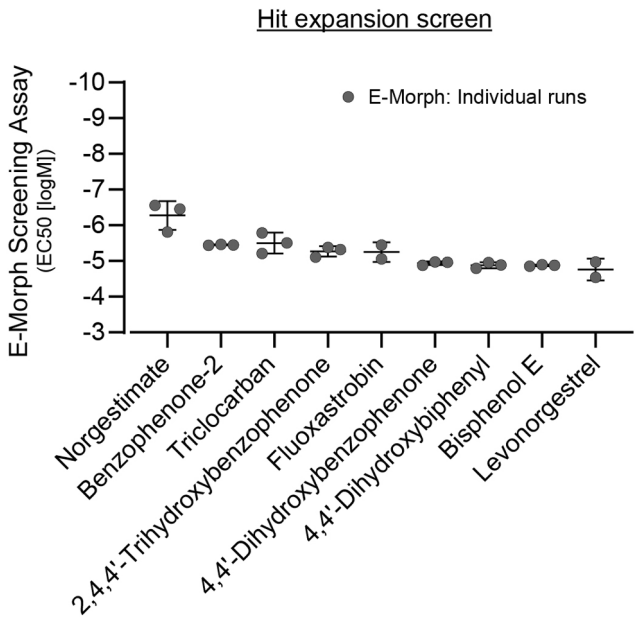


Fig. 6. Hit expansion screening and potency determination. Potencies (half-maximal concentrations, EC50) of nine additional hit substances that were active in the hit expansion screen. Each data point represents an individual run. Biological replicates, n = 3. Error bars, mean +/- SD.

Although, potential false positive results, like 2,4,6-TTBP, Zineb, or Azoxystrobin, are inevitable, the identification of Hexythiazox and Diuron as potential estrogenic substances seems to be relevant and warrant further investigation.

3.7. Docking of bisphenols into ERα

Even though the primary hit substance Bisphenol F shares a very similar chemical structure with 4-Benzylphenol (Fig. 7, File S2), the latter was inactive in the E-Morph Screening Assay (Table 3; Table S1). This was particularly surprising since this substance had the highest Tanimoto similarity score of 0.78 as compared to three additional substances with similarity scores between 0.49 and 0.56 that were active in the E-Morph Assay. Here, it needs to be considered that terminal functional group differences in molecules often have less influence on the calculated similarity than central atom changes (because of the nature of the calculation of the circular environments, such as in the Morgan fingerprint). The different activities could neither be explained using diverse other molecular fingerprints (MorganCount, MACCS, pharmacophore fingerprints), which were very similar for these bisphenols (Table S3). In scientific literature, such phenomena are often described as activity cliffs (Maggiora et al. 2014) in which, e.g., changes of one atom or functional group in otherwise very similar molecules can lead to a notable difference in activity. To better understand this effect, we performed a docking analysis (Brooijmans and Kuntz 2003) and investigated potential variations in interaction patterns of the bisphenols with ERα using the @TOME-2 web-server according to (Delfosse et al. 2012). The docking results showed an apparent difference in the binding of 4-Benzylphenol to ERα (Fig. 7, red arrow) as compared to the other four tested bisphenols. The latter four substances are all capable of forming undirected interactions with the hydrophobic pocket core, while being anchored on both sides through hydrogen bonds (see pharmacophoric interactions in Fig. 7). Hence, the inactivity of 4-Benzylphenol in the E-Morph Screening Assay could be explained by the absence of an important hydrogen bond to histidine H524 (because of a missing hydroxy (–OH) group), which may significantly reduce the binding capacity of 4-Benzylphenol to ERα.

3.8. Generation of in silico models for prediction of estrogenic activities in HTS approaches

Ideally, these kinds of structural analyses will eventually lead to the development of highly predictive (Q)SAR tools to identify estrogenic activities in environmental chemicals based on structural fingerprints. As a proof-of-concept for the application of *in silico* prediction models in HTS approaches, we used ToxCast ER screening data to build seven *in silico* prediction models based on the conformal prediction (CP) framework (Morger et al. 2020; Norinder et al. 2014; Vovk et al. 2005) for the relevant mechanistic events of estrogen signaling, i.e., ER binding, ER dimerization, regulation of gene expression, and cell proliferation (Table 4), that are also included in the ToxCast ER Agonist Model (Browne et al. 2015). We trained these CP ER models on all binary readouts (active/inactive) of the corresponding *in chemico* and *in vitro* ER screening assays that were conducted in the ToxCast project. A fivefold cross-validation was employed to assess the performance of each model (Table 4). The benefit of the CP method over simpler similarity search is thereby two-fold. First, machine learning (ML) models are statistical models that relate a set of structural descriptors of a chemical compound to its biological activity. Thus, the CP model learns which features in the molecule contribute more (or less) to the outcome, whereas a simpler similarity search treats all features the same. In other words, while similarity search looks for more obvious similarities between molecules, CP may detect more hidden, and also non-linear, relationships. Second, CP is built on top of a ML framework and adds a calibration step. This allows for monitoring the reliability of the predictions more closely, thus providing a measure of confidence in the prediction per molecule.

All CP ER models were valid at the 0.2 significance level (validity $\geq$ 0.8), i.e., making<20% prediction errors when considering 'single class' (active or inactive) and 'both class' (active and inactive) predictions (Table 4). This high mean validity of 0.85 $\pm$ 0.01 indicates that the models were well calibrated and can therefore be reliably applied to new data. The mean efficiency, i.e., the fraction of single class predictions made by the models, was 0.39 $\pm$ 0.12 (Table 4) and notably lower than compared to other CP ER models described before (Ji et al. 2018; Norinder et al. 2016). This can be a consequence of the use of the additional normalizer regression model and prior equal size sampling of the proper training and calibration set, which was shown to improve the prediction performance on external data (Morger et al. 2020). The CP ER models had a mean accuracy, i.e., the fraction of correct single class predictions made by the models, of 0.71 $\pm$ 0.10 (Table 4). Regarding the class-wise evaluation, the mean accuracy for prediction of the active class was rather high with an average of 0.83 $\pm$ 0.03, whereas the mean accuracy for prediction of the inactive class was slightly lower with an average of 0.67 $\pm$ 0.13 (Table 4). The reduced mean accuracy for inactive class predictions was mainly caused by two rather weakly performing endpoint models (aeid_788 and aeid_2) (Table 4, italics). Nevertheless, the overall results show that the individual CP ER models can reliably predict agonistic ER activity, especially since the focus of this study is to detect active substances.

We then applied the seven CP ER models to classify the nine (excluding EDTA) potential 'Novel estrogenic substances' (Table 2) from primary and hit confirmation screening as well as the selected 20 structurally similar substances from the hit expansion screening (Table 3). The respective p-values, which describe the certainty of the active/inactive predictions of the individual models, are summarized in Table S4. To facilitate direct comparison of the CP ER model classifications with the E-Morph Screening Assay results, we converted the seven individual CP ER model predictions into an overall 'consensus prediction' by applying a 'majority rule' principle (Table 5; Table S4). The *in chemico* and *in vitro* test results of the seven corresponding ER screening assays included in the ToxCast project were converted in the same way to obtain an overall 'consensus test result' for each individual substance (Table 5; Table S4). Notably, for some of the 29 test

Table 3
Screening results for the hit expansion substances compared to published *in silico* ER model data from the U.S. EPA.

Chemical name	CAS No.	US EPA *in silico* ER models		Similarity search	E-Morph Screening Assay	Hit expansion screen					
		ToxCast ER Agonist Score [a)]	CERAPP ER Agonist Model [b)]	Similarity score	Substance group	Potency [M] EC50	SD	n	ER Bioactivity [rel. LogEC50]	Comment	Tested in primary screen
Hexythiazox	***78587–05-0***	***0.00***	***inactive***	***1.00***	***Novel***	***1.01E-08***	***3.94E-09***	***4***	***0.91***	***active***	***Y***
Iprodion	36734–19-7	0.00	inactive	0.36	TN	NA	NA		NA	inactive	Y
Norethindrone acetate (NETA)	***51–98-9***	***NA***	***active***	***1.00***	***Novel***	***6.25E-07***	***4.62E-07***	***6***	***0.70***	***active***	***Y***
Ethynodiol diacetate	297–76-7	NA	active	0.73	Novel	< 1.37E-08	< 1.37E-08	3	NA	active	–
Norgestrel	6533–00-2	0.39	active	0.55	FN	NA	NA		NA	inactive	–
Levonorgestrel	797–63-7	0.39	active	0.55	Known	1.97E-05	1.27E-05	2	0.53	active	Y
Norgestimate	35189–28-7	NA	active	0.51	Novel	7.27E-07	7.14E-07	3	0.70	active	–
Nandrolone	***434–22-0***	***NA***	***active***	***1.00***	***Novel***	***2.04E-06***	***1.38E-06***	***6***	***0.65***	***active***	***Y***
Norgestrel	6533–00-2	0.39	active	0.60	FN	NA	NA		NA	inactive	–
Levonorgestrel	797–63-7	0.39	active	0.60	Known	1.97E-05	1.27E-05	2	0.53	active	Y
Phloretin	***60–82-2***	***NA***	***active***	***1.00***	***Novel***	***3.36E-06***	***1.41E-06***	***5***	***0.62***	***active***	***Y***
Benzophenone-2	131–55-5	0.40	active	0.43	Known	3.55E-06	1.00E-07	3	0.62	active	Y
2,4,4′-Trihydroxybenzophenone	1470–79-7	NA	active	0.42	Novel	5.60E-06	1.98E-06	3	0.60	active	–
Diphenolic acid	126–00-1	0.17	active	0.38	Novel	NA	NA		NA	active at ≥ 30 μM	–
2,4,6-Tri-tert-butylphenol (TTBP)	***732–26-3***	***0.00***	***inactive***	***1.00***	***Novel***	***4.73E-06***	***3.12E-06***	***4***	***0.60***	***active***	***Y***
Butylhydroxytoluene	128–37-0	0.00	inactive	0.72	TN	NA	NA		NA	inactive	Y
2,5-Di-*tert*-butylhydroquinone	88–58-4	0.00	inactive	0.60	TN	NA	NA		NA	inactive	–
Bisphenol F	***620–92-8***	***NA***	***active***	***1.00***	***Novel***	***4.79E-06***	***1.41E-06***	***5***	***0.60***	***active***	***Y***
4-Benzylphenol	101–53-1	NA	active	0.78	NA	NA	NA		NA	inactive	Y
4,4′-Dihydroxybiphenyl	92–88-6	NA	active	0.56	Novel	1.34E-05	2.59E-06	3	0.55	active	Y
4,4′-Dihydroxybenzophenone	611–99-4	NA	active	0.49	Novel	1.17E-05	1.42E-06	3	0.56	active	Y
Bisphenol E	2081–08-5	NA	active	0.49	Novel	1.34E-05	6.56E-07	3	0.55	active	–
Diuron	***330–54-1***	***0.00***	***inactive***	***1.00***	***Novel***	***6.04E-06***	***2.91E-06***	***3***	***0.59***	***active***	***Y***
Linuron	330–55-2	0.00	inactive	0.67	TN	NA	NA		NA	inactive	Y
Swep	1918–18-9	NA	inactive	0.59	NA	NA	NA		NA	inactive	–
Troclocarban	101–20-2	0.00	inactive	0.52	Novel	3.65E-06	2.33E-06	3	0.62	active	Y
Azoxystrobin	***131860–33-8***	***0.00***	***inactive***	***1.00***	***Novel***	***6.34E-06***	***3.68E-06***	***6***	***0.59***	***active***	***Y***
Picoxystrobin	117428–22-5	0.00	inactive	0.45	TN	NA	NA		NA	inactive	–
Fluoxastrobin	361377–29-9	0.00	inactive	0.33	Novel	6.23E-06	3.72E-06	2	0.59	active	–
Zineb	***12122–67-7***	***NA***	***inactive***	***1.00***	***Novel***	***8.71E-05***	***8.48E-05***	***2***	***0.46***	***active***	***Y***
Maneb	12427–38-2	0.00	inactive	1.00	TN	NA	NA		NA	inactive	Y

Overall classifications and potencies of 20 hit expansion substances as compared to the ToxCast ER Agonist Score. The similar substances for the nine (excluding EDTA) potential 'Novel estrogenic substances' (bold italic) were selected based on their structural similarity and their activity in different ToxCast steroidal nuclear receptor assays. EC50, mean potency from multiple independent runs (n). SD, standard deviation. ER Bioactivity, potency (logEC50) normalized to 17-alpha-Ethinylestradiol (1.00). NA, not available/not applicable. FN, false negative substances. TN, true negative substance.

[a)]Screening Chemicals for Estrogen Receptor Bioactivity Using a Computational Model (Browne et al. 2015)

[b)]CERAPP: Collaborative Estrogen Receptor Activity Prediction Project (Mansouri et al. 2016)

Table 4
Development and evaluation (cross-validation) of CP ER models.

U.S. EPA *in chemico/in vitro* ER screening assays			Conformal prediction ER models										
Assay name	Assay ID (aeid)	Biological mechanism	Validity			Efficiency			Accuracy			no. of compounds	
			all	inactive	active	all	inactive	active	all	inactive	active	inactive	active
NVS_NR_hER	714	receptor binding	0.85	0.85	0.85	0.40	0.36	0.57	0.73	0.69	0.81	819	204
OT_ER_ERaERb_1440	745	receptor dimerization	0.86	0.85	0.88	0.32	0.29	0.53	0.68	0.62	0.88	1333	180
ATG_ERE_CIS_up	75	gene expression, mRNA	0.85	0.85	0.85	0.43	0.40	0.51	0.74	0.71	0.81	2257	792
ATG_ERa_TRANS_up	117	gene expression, mRNA	0.84	0.84	0.84	0.50	0.47	0.61	0.77	0.75	0.83	2385	681
TOX21_ERa_BLA_Agonist_ratio	785	gene expression, protein	0.86	0.86	0.84	0.57	0.57	0.57	0.87	0.87	0.84	6465	320
TOX21_ERa_LUC_VM7_Agonist	788	gene expression, protein	0.85	0.85	0.85	0.25	0.22	0.41	0.58	*0.5*	0.84	5719	858
ACEA_T47D_80hr_Positive	2	cell proliferation	0.86	0.87	0.84	0.24	0.20	0.41	0.63	*0.54*	0.81	1307	266
		Mean	**0.85**			**0.39**			**0.71**	**0.67**	**0.83**		
		SD	**0.01**			**0.12**			**0.10**	**0.13**	**0.03**		

Information on the seven CP models built for seven ER screening assay endpoints conducted in the ToxCast project. Accuracies highlighted in italics indicate reduced performance of two CP ER models for prediction of inactive substances.

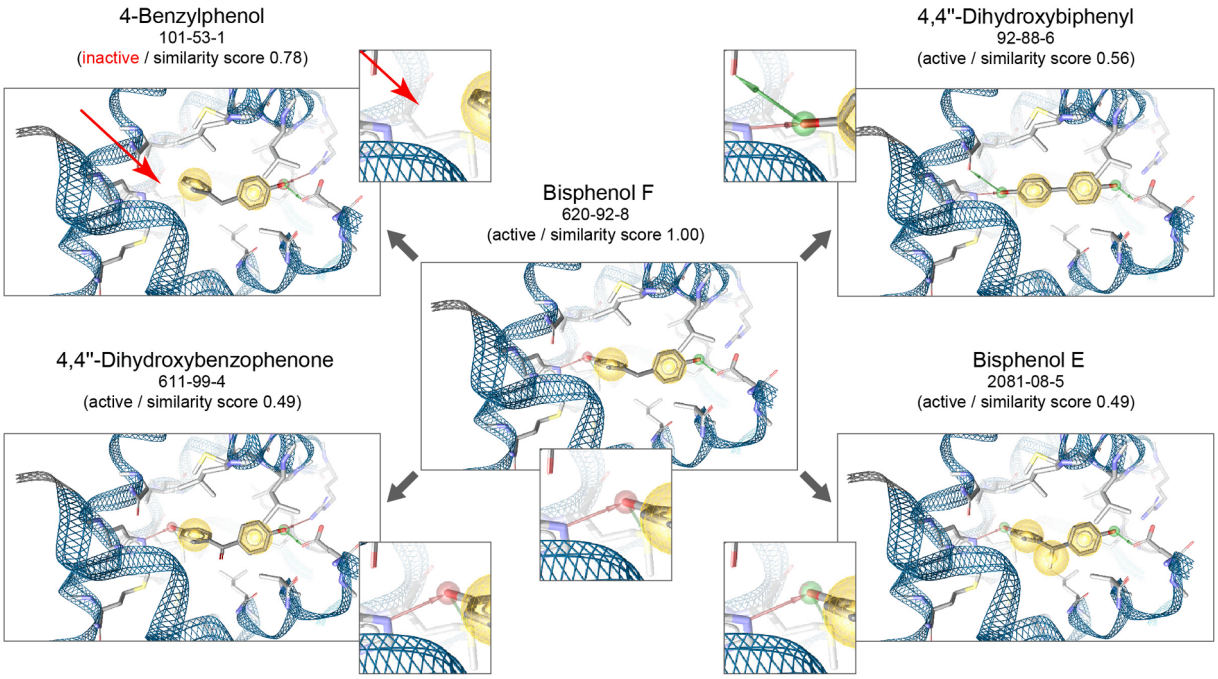


Fig. 7. Docking of bisphenols into ERα. Pharmacophoric interactions of different bisphenols with ERα. For each substance, the outcome of the E-Morph Screening Assay (active/inactive) and the Tanimoto similarity to Bisphenol F are indicated. The red arrow points to a clear difference in the binding of 4-Benzylphenol to ERα (missing hydrogen bond) as compared to the other four tested bisphenols. Visualizations using LigandScout after docking with the @TOME-2 webserver according to (Delfosse et al. 2012). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

substances, no conclusive (Table S4, 'NC') CP ER model classifications were achieved, i.e., 'both class' predictions were returned by the CP framework (no decisions could be made), and not all these substances were tested in every of the seven ER screening assays (Table S4, 'NA').

For this small subset of 29 substances, the overall concordance of the CP ER consensus model, the E-Morph Screening Assay, and the consensus of the ER screening assay test results was in the range of 71–76% (Table 6). Hence, the performance of the CP ER consensus model was comparable to the mean accuracy of the seven individual CP ER models (see Table 4) supporting the consensus model approach, which integrates multiple relevant mechanistic events of estrogen signaling. Importantly, the high predictivity of the CP ER consensus model (100%) and the E-Morph Screening Assay (87%) for active class substances promotes their future use in HTS frameworks. With regard to the evaluation of the E-Morph Assay screening results, the CP ER consensus model supported 89% of the active class assignments for the subset of 29 substances. The reduced predictivity of the E-Morph Screening Assay (54%) for inactive class substances was mainly caused by a higher frequency of non-concordant 'false positive' results, which, however, represent substances with potential estrogenic activity ('Novel estrogenic substances' group, Table S1) that are of particular interest for prioritized follow-up testing.

Taken together, these data suggest that the generated CP ER models are applicable for fast and efficient browsing of large substance libraries to prioritize substances with potential estrogenic activity for subsequent *in vitro* testing in HTS approaches, such as E-Morph. Benchmarking the CP prediction models and the E-Morph Screening Assay against a larger set of reference substances will provide further insights into their

Table 5
Comparison of results from the E-Morph Screen and CP ER models with published *in vitro* and *in silico* ER data from the U.S. EPA.

Chemical name	CAS No.	U.S. EPA *in silico* ER models		U.S. EPA *in chemico/in vitro* ER screening assays		E-Morph Screening Assay		Conformal prediction ER models	
		ToxCast	CERAPP	Consensus **test results**		Substance group	Potency [M]	Consensus **predictions**	
		ER Agonist Score[a)]	ER Agonist Model[b)]	active	inactive		EC50	active	inactive
Hexythiazox	***78587–05-0***	***0.00***	***inactive***	***1 (14%)***	***6 (86%)***	***Novel***	***1.01E-08***	***0***	***0***
Norethindrone acetate (NETA)	***51–98-9***	***NA***	***active***	***2 (100%)***	***0***	***Novel***	***6.25E-07***	***6 (100%)***	***0***
Norgestimate	35189–28-7	NA	active	2 (100%)	0	Novel	7.27E-07	4 (100%)	0
Nandrolone	***434–22-0***	***NA***	***active***	***2 (100%)***	***0***	***Novel***	***2.04E-06***	***7 (100%)***	***0***
Phloretin	***60–82-2***	***NA***	***active***	***2 (100%)***	***0***	***Novel***	***3.36E-06***	***7 (100%)***	***0***
Benzophenone-2	131–55-5	0.40	active	7 (100%)	0	Known	3.55E-06	7 (100%)	0
Troclocarban	101–20-2	0.00	inactive	0	6 (100%)	Novel	3.65E-06	1 (100%)	0
2,4,6-Tri-tert-butylphenol (TTBP)	***732–26-3***	***0.00***	***inactive***	***1 (20%)***	***4 (80%)***	***Novel***	***4.73E-06***	***4 (80%)***	***1 (20%)***
Bisphenol F	***620–92-8***	***NA***	***active***	***4 (100%)***	***0***	***Novel***	***4.79E-06***	***7 (100%)***	***0***
2,4,4′-Trihydroxybenzophenone	1470–79-7	NA	active	5 (100%)	0	Novel	5.60E-06	7 (100%)	0
Diuron	***330–54-1***	***0.00***	***inactive***	***0***	***6 (100%)***	***Novel***	***6.04E-06***	***0***	***5 (100%)***
Fluoxastrobin	361377–29-9	0.00	inactive	1 (20%)	4 (80%)	Novel	6.23E-06	0	0
Azoxystrobin	***131860–33-8***	***0.00***	***inactive***	***1 (17%)***	***5 (83%)***	***Novel***	***6.34E-06***	***1 (100%)***	***0***
4,4′-Dihydroxybenzophenone	611–99-4	NA	active	2 (100%)	0	Novel	1.17E-05	7 (100%)	0
Bisphenol E	2081–08-5	NA	active	4 (100%)	0	Novel	1.34E-05	7 (100%)	0
4,4′-Dihydroxybiphenyl	92–88-6	NA	active	2 (100%)	0	Novel	1.34E-05	2 (100%)	0
Levonorgestrel	797–63-7	0.39	active	6 (86%)	1 (14%)	Known	1.97E-05	6 (100%)	0
Zineb	***12122–67-7***	***NA***	***inactive***	***0***	***0***	***Novel***	***8.71E-05***	***0***	***2 (100%)***
Ethynodiol diacetate	297–76-7	NA	active	2 (100%)	0	Novel	< 1.37E-08	5 (100%)	0
Diphenolic acid	126–00-1	0.17	active	6 (86%)	1 (14%)	Novel	pos ≥ 30 μM	5 (100%)	0
Norgestrel	6533–00-2	0.39	active	5 (100%)	0	FN	inactive	6 (100%)	0
4-Benzylphenol	101–53-1	NA	active	2 (100%)	0	NA	inactive	4 (100%)	0
Picoxystrobin	117428–22-5	0.00	inactive	1 (17%)	5 (83%)	TN	inactive	1 (100%)	0
Butylhydroxytoluene	128–37-0	0.00	inactive	1 (17%)	5 (83%)	TN	inactive	2 (100%)	0
2,5-Di-*tert*-butylhydroquinone	88–58-4	0.00	inactive	1 (14%)	6 (86%)	TN	inactive	4 (100%)	0
Swep	1918–18-9	NA	inactive	0 (0%)	4 (100%)	NA	inactive	0	5 (100%)
Linuron	330–55-2	0.00	inactive	0 (0%)	7 (100%)	TN	inactive	0	1 (100%)
Iprodion	36734–19-7	0.00	inactive	1 (17%)	5 (83%)	TN	inactive	0	0
Maneb	12427–38-2	0.00	inactive	0 (0%)	6 (100%)	TN	inactive	0	2 (100%)

Overall classifications and potencies for the nine (excluding EDTA) potential 'Novel estrogenic substances' (bold italic) and the 20 hit expansion chemicals that were tested in the E-Morph Screening Assay as compared to the ToxCast ER Agonist Score and the available ToxCast ER agonist assay screening data. Consensus predictions considering modes of all single class predictions from the individual ER models or ER screening assays. EC50, mean potency from multiple independent runs. NA, not available/not applicable. FN, false negative substances. TN, true negative substances.
[a)]Screening Chemicals for Estrogen Receptor Bioactivity Using a Computational Model (Browne et al. 2015).
[b)]CERAPP: Collaborative Estrogen Receptor Activity Prediction Project (Mansouri et al. 2016).

predictive capacities in future studies. In iterative *in silico - in vitro* screening cycles, the newly generated data could then also be used to update and improve the CP ER models. Particularly, additional training data that cover the chemical space for which the current models make poor predictions could prompt more efficient and accurate predictions, which in turn supports the identification of additional active substances by HTS (Svensson et al. 2017a).

4. Conclusion

Over the years, many organizations worldwide have published candidate lists of suspected EDCs, which include hundreds of substances that may pose a potential threat to human health and the environment (WHO/UNEP, 2012). The identification and regulatory restriction of such EDCs is a central goal of chemicals management frameworks and policies worldwide. Commitments such as the 'European Green Deal' even pursue a zero-pollution ambition towards a fully 'toxic-free environment' in the next decades. This intention is reflected in the recently adopted 'Chemicals Strategy for Sustainability' (EC 2020), which also promotes the 'safe-by-design' approach, i.e. the use of substances that pose less or no harm to humans and the environment. For now, adverse health effects of EDCs are mainly investigated in animal experiments (OECD, 2001, 2018a, 2018b, 2018c), although animal data are not necessarily directly translatable to (patho-)physiological processes in humans (Holen et al. 2017). Moreover, these *in vivo* assays are not necessarily specific for individual endocrine mechanisms or suitable for the analysis of 'real-life' co-exposure scenarios. The projected doubling of the global chemical sales by 2030 (WHO, UNEP, 2019) and, in parallel, the intended phasing out of animal experimentation in toxicological testing (Grimm 2019) emphasize the need for novel human-relevant HTS methods and computational approaches to ensure protection of human health and the environment.

Table 6
Predictivity of the E-Morph Screening Assay and CP ER models.

			U.S. EPA *in chemico/in vitro* ER screening assays	E-Morph Screening Assay
			Consensus test results	Test results
Conformal prediction ER models	Consensus predictions	N	25	26
		$N_{True\ Actives}$	15	16
		$N_{False\ Inactives}$	0	2
		$N_{True\ Inactives}$	4	3
		$N_{False\ Actives}$	6	5
		Concordance	**76%**	**73%**
		$P_{active\ class}$	**100%**	**89%**
		$P_{inactive\ class}$	**40%**	**38%**
			U.S. EPA *in chemico/in vitro*ER screening assays	Conformal prediction ER models
			Consensus test results	Consensus predictions
E-Morph Screening Assay	Test results	N	28	26
		$N_{True\ Actives}$	13	16
		$N_{False\ Inactives}$	2	5
		$N_{True\ Inactives}$	7	3
		$N_{False\ Actives}$	6	2
		Concordance	**71%**	**73%**
		$P_{active\ class}$	**87%**	**76%**
		$P_{inactive\ class}$	**54%**	**60%**
			E-Morph Screening Assay	Conformal prediction ER models
			Test results	Consensus test results
U.S. EPA *in chemico/in vitro* ER screening assays	Consensus predictions	N	28	25
		$N_{True\ Actives}$	13	15
		$N_{False\ Inactives}$	6	6
		$N_{True\ Inactives}$	7	4
		$N_{False\ Actives}$	2	0
		Concordance	**71%**	**76%**
		$P_{active\ class}$	**68%**	**71%**
		$P_{inactive\ class}$	**78%**	**100%**

Concordance between the E-Morph Screening Assay, the CP ER models, and the available ToxCast ER agonist assay screening data was calculated based on the results of the nine (excluding EDTA) 'Novel estrogenic substances' and the 20 hit expansion chemicals. Note that the total numbers (n) of substances differ because for some of the 29 test substances, no conclusive CP ER model classifications were achieved and not all of the 29 substances were tested in every of the seven ToxCast ER screening assays. Consensus predictions considering modes of all single class predictions from the individual CP ER models or ToxCast ER screening assays. n, total number of substances. N, number of true/false active/inactive substances for each type of comparison. P, predictivity for active/inactive substances.

4.1. The E-Morph screening Assay provides a reliable and robust human-relevant readout to determine ER signaling activity by phenotypic HTS

The E-Morph Assay (Kornhuber et al. 2021) addresses a human-relevant functional endpoint of adversity, i.e., the perturbation of cell–cell adhesion leading to breast cancer progression and metastasis (Bischoff et al. 2020). In the present study, we further developed the applicability of the original E-Morph Assay for automated HTS using local changes in E-Cad-GFP signal intensity (SI) as a novel, simple and reliable HTS-compatible phenotypic readout for estrogenic activity (see Fig. 1). The SI readout was very robust, with each valid run in the primary screen achieving a Z'-factor above 0.5 (Iversen et al. 2006; Zhang et al. 1999). The determined EC50 values under anti-estrogenic (Fulv treatment) and estrogenic (Fulv + E2 treatment) conditions were directly comparable to the results of the original E-Morph Assay (Kornhuber et al. 2021) with the advantage that the adapted assay avoids both live-cell staining and extensive quantitative image analysis procedures. Based on an intact, complete, and interconnected endogenous estrogen signaling pathway, the E-Morph Screening Assay therefore facilitates the efficient identification of substances with estrogenic activities and the determination of their potencies from concentration-response curves. It could therefore help to accelerate the identification of new substances of concern and support the comprehensive assessment of potential mixture effects of EDCs (Schlotz et al. 2017; Yu et al. 2019).

4.2. E-Morph phenotypic screening correctly identified 27 'known' estrogenic substances and 10 'novel' substances with potential estrogenic activity

We used the E-Morph Screening Assay to analyze a novel substance library (BfR-ChemLibrary) comprising 430 toxicologically-relevant industrial chemicals, biocides, and plant protection products (see Figs. 2 and 3). We identified 27 estrogenic substances of which the potencies of 24 substances correlated very well with the ToxCast ER pathway model (Browne et al. 2015; Judson et al. 2015) (see Fig. 4, Table 1). We further identified 10 additional potential estrogenic substances that have not been described as such in ToxCast before (see Fig. 4, Table 2). According to a recently proposed human-relevant potency threshold (HRPT) for ERα agonism, the minimum relative activity of a test substance must be at least 0.01% of strong estrogens (E2 or 17α-Ethinylestradiol) to exert adverse effects in humans via an ERα-mediated mechanism (Borgert et al. 2018). In the E-Morph Screening Assay, the potencies of the active substances (Fig. 4; Table 1 and 2) were in the range of 1 nM (strong activity, e.g., E2) to 10 μM (weak activity, e.g., Apigenin) and, thus, fitted well into the HRPT with only Zineb displaying a potency slightly below. Subsequent hit verification studies, including gene expression profiling and ERα binding supported the detected estrogenic activity of Hexythiazox, NETA, Nandrolone, Phloretin, Diuron, and Bisphenol F but not of 2,4,6-TTBP, Zineb, and Azoxystrobin (see Fig. 5).

4.3. Use of in silico tools increased the hit-rate and supported the hit evaluation

The E-Morph screening results for the group of 'novel' substances (Table 2) were further substantiated by additional testing of 20 structurally similar substances that were selected based on an *in silico* similarity search for subsequent hit expansion screening, which identified in total another nine ER active substances (see Fig. 6 and Table 3). While being structurally very similar to Bisphenol F, 4-Benzylphenol was inactive in the E-Morph Screening Assay. Additional docking studies detected a difference in the binding mode (i.e., a missing hydrogen bond) to ERα, which could explain the difference in activity of otherwise very similar molecules (see Fig. 7). Hence, computational docking analyses can significantly support the interpretation of *in vitro* screening results, particularly regarding the capability of substances to bind to nuclear hormone receptors. In addition, we built seven *in silico* ER models using the CP framework and the publicly available ToxCast assay data to predict further substances with potential estrogenic activity and to support the E-Morph screening results. The high predictivity for active substances (see Table 5 and 6), which is particularly important from a regulatory point of view to protect human health and the environment, support that the E-Morph Screening Assay and the CP ER models are fit-for-purpose to be applied to new data in an automated manner.

4.4. Future applications of the E-Morph screening Assay and the CP ER models

Provided that a future validation study demonstrates transferability and inter-laboratory reproducibility, the E-Morph Screening Assay appears to be generally suitable for inclusion in existing HTS projects, where it can be used to identify both substances with estrogenic and anti-estrogenic activities using the same phenotypic readout. In ER testing batteries, the E-Morph Screening Assay could be used for efficient analysis of comprehensive substance libraries in order to prioritize substances for subsequent testing against higher tier endpoints, thereby avoiding unnecessary animal testing. ER testing strategies could additionally benefit from further development and implementation of *in silico* tools, including similarity search approaches and CP models, in the evaluation of screening results and targeted selection of candidate substances for follow-up *in vitro* analysis. Well-trained CP ER models may ultimately even replace existing ER HTS assays that resemble the complex (patho-)physiological processes in humans to a rather limited extent. The combination of human-relevant HTS assays and CP models in testing and assessment strategies can ultimately help to increase confidence in *in vitro* results for the regulatory decisions making and thus make an important contribution to achieve the goal of a next generation risk assessment framework that does no longer depend on animal experimentation.

CRediT authorship contribution statement

Saskia Klutzny: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Marja Kornhuber:** Validation, Investigation, Data curation, Writing – review & editing. **Andrea Morger:** Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing. **Gilbert Schönfelder:** Resources, Writing – review & editing, Supervision, Project administration, Funding acquisition. **Andrea Volkamer:** Conceptualization, Methodology, Resources, Writing – original draft, Writing – review & editing, Supervision, Project administration, Funding acquisition. **Michael Oelgeschläger:** Conceptualization, Resources, Writing – review & editing, Supervision, Project administration. **Sebastian Dunst:** Conceptualization, Methodology, Validation, Data curation, Writing – original draft, Writing – review & editing, Visualization, Supervision, Project administration.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper: [The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: A patent application (published under EP3517967, WO/2019/145517, US20210055311) for the endpoint and conceptual design of the E-Morph Assay to screen substances for estrogenic or anti-estrogenic activity has been filed at the European Patent Office by the employer (German Federal Institute for Risk Assessment (BfR)) of the authors. The German Federal Institute for Risk Assessment (BfR) is a scientifically independent institution within the portfolio of the Federal Ministry of Food and Agriculture (BMEL) in Germany. The authors' freedom to design, conduct, interpret, and publish research is explicitly not compromised.].

Acknowledgements

We thank Sylvie Coscoy (Laboratoire Physico-Chimie Curie, Institut Curie, PSL Research University - Sorbonne Universités, UPMC-CNRS, Paris, France) for providing the MCF-7/E-Cad-GFP cell line. We are particularly grateful to Verena Fetz (BfR, Berlin, Germany) for supporting the set-up of automated HTS workflows and Edgar Specker and Marc Nazaré (Compound Management Unit and Medicinal Chemistry Group, Leibniz Institute of Molecular Pharmacology, Berlin, Germany) for generation and management of the BfR-ChemLibrary compound plates. We further gratefully acknowledge our colleagues from BfR/Bf3R for scientific input and comments on the manuscript. This work was supported by an internal BfR research funding program (Sonderforschungsprojekt 1322-683), the HaVo-Stiftung for A.M., and a BMBF grant (031A262C) for A.V.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.envint.2021.106947.

References

Alvarsson, J., Arvidsson McShane, S., Norinder, U., Spjuth, O., 2021. Predicting With Confidence: Using Conformal Prediction in Drug Discovery. J. Pharm. Sci. 110 (1), 42–49.

Atkinson, F. Standardiser. https://github.com/flatkinson/standardiser; 2014.

Bell, S., Abedini, J., Ceger, P., Chang, X., Cook, B., Karmaus, A.L., Lea, I., Mansouri, K., Phillips, J., McAfee, E., Rai, R., Rooney, J., Sprankle, C., Tandon, A., Allen, D., Casey, W., Kleinstreuer, N., 2020. An integrated chemical environment with tools for chemical safety testing. Toxicol. In Vitro 67, 104916. https://doi.org/10.1016/j.tiv.2020.104916.

Bell, S.M., Phillips, J., Sedykh, A., Tandon, A., Sprankle, C., Morefield, S.Q., Shapiro, A., Allen, D., Shah, R., Maull, E.A., Casey, W.M., Kleinstreuer, N.C., 2017. An Integrated Chemical Environment to Support 21st-Century Toxicology. Environ. Health Perspect 125, 054501.

Bender, A., Glen, R.C., 2004. Molecular similarity: a key technique in molecular informatics. Org. Biomol. Chem. 2 (22), 3204. https://doi.org/10.1039/b409813g.

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. Data Analysis, Machine Learning and Applications; 2008.

Bischoff, P.; Kornhuber, M.; Dunst, S.; Zell, J.; Fauler, B.; Mielke, T.; Taubenberger, A.V.; Guck, J.; Oelgeschlager, M.; Schonfelder, G. Estrogens Determine Adherens Junction Organization and E-Cadherin Clustering in Breast Cancer Cells via Amphiregulin. iScience 2020;23:101683.

Borgert, C.J., Matthews, J.C., Baker, S.P., 2018. Human-relevant potency threshold (HRPT) for ERalpha agonism. Arch Toxicol 92, 1685–1702.

Branham, W.S.; Dial, S.L.; Moland, C.L.; Hass, B.S.; Blair, R.M.; Fang, H.; Shi, L.; Tong, W.; Perkins, R.G.; Sheehan, D.M. Phytoestrogens and mycoestrogens bind to the rat uterine estrogen receptor. J Nutr 2002;132:658–664.

Brooijmans, N., Kuntz, I.D., 2003. Molecular recognition and docking algorithms. Annu. Rev. Biophys. Biomol. Struct. 32 (1), 335–373.

Browne, P., Judson, R.S., Casey, W.M., Kleinstreuer, N.C., Thomas, R.S., 2015. Screening Chemicals for Estrogen Receptor Bioactivity Using a Computational Model. Environ. Sci. Technol. 49 (14), 8804–8814.

Carlsson, L.; Eklund, M.; Norinder, U. Aggregated Conformal Prediction. Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications; 2014.

Carrió, P., Sanz, F., Pastor, M., 2016. Toward a unifying strategy for the structure-based prediction of toxicological endpoints. Arch. Toxicol. 90 (10), 2445–2460.

Chwalisz, K., Surrey, E., Stanczyk, F.Z., 2012. The hormonal profile of norethindrone acetate: rationale for add-back therapy with gonadotropin-releasing hormone agonists in women with endometriosis. Reprod. Sci. 19 (6), 563–571.

de Beco, S., Gueudry, C., Amblard, F., Coscoy, S., 2009. Endocytosis is required for E-cadherin redistribution at mature adherens junctions. Proc. Natl. Acad. Sci. U S A 106 (17), 7010–7015.

de Beco, S.; Gueudry, C.; Amblard, F.; Coscoy, S. Correction for de Beco et al., Endocytosis is required for E-cadherin redistribution at mature adherens junctions. Proc Natl Acad Sci U S A 2020;117:23191.

Delfosse, V., Grimaldi, M., Pons, J.-L., Boulahtouf, A., le Maire, A., Cavailles, V., Labesse, G., Bourguet, W., Balaguer, P., 2012. Structural and mechanistic insights into bisphenols action provide guidelines for risk assessment and discovery of bisphenol A substitutes. Proc. Natl. Acad. Sci. U S A 109 (37), 14930–14935.

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.

EC. European Commission, Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions - Chemicals Strategy for Sustainability - Towards a Toxic-Free Environment ed˚eds. Brussels; 2020.

EFSA. The 2015 European Union report on pesticide residues in food. EFSA J 2017;15:e04791.

Filer, D., Patisaul, H.B., Schug, T., Reif, D., Thayer, K., 2014. Test driving ToxCast: endocrine profiling for 1858 chemicals included in phase II. Curr Opin Pharmacol 19, 145–152.

Gayvert, K.M.; Madhukar, N.S.; Elemento, O. A Data-Driven Approach to Predicting Successes and Failures of Clinical Trials. Cell Chem Biol 2016;23:1294–1301.

Grimm, D., 2019. EPA to eliminate all mammal testing by 2035. Science. https://doi.org/10.1126/science:aaz4593.
Hemmerich, J., Ecker, G.F., 2020. In silico toxicology: From structure-activity relationships towards deep learning and adverse outcome pathways. Wires Comput. Mol. Sci. 10 (4) https://doi.org/10.1002/wcms.v10.410.1002/wcms.1475.
Holen, I., Speirs, V., Morrissey, B., Blyth, K., 2017. In vivo models in breast cancer research: progress, challenges and future directions. Dis. Model. Mech. 10, 359–371.
Huang, R.L.; Xia, M.H. Editorial: Tox21 Challenge to Build Predictive Models of Nuclear Receptor and Stress Response Pathways As Mediated by Exposure to Environmental Toxicants and Drugs. Front Env Sci-Switz 2017;5.
Iversen, P.W., Eastwood, B.J., Sittampalam, G.S., Cox, K.L., 2006. A comparison of assay performance measures in screening assays: signal window, Z' factor, and assay variability ratio. J. Biomol. Screen 11 (3), 247–252.
Ji, C.; Svensson, F.; Zoufir, A.; Bender, A. eMolTox: prediction of molecular toxicity with confidence. Bioinformatics 2018;34:2508–2509.
Jordan, V.C., 1977. Effects of tamoxifen in relation to breast cancer. Br Med J 1 (6075), 1534–1535.
Judson, R.S., Houck, K.A., Kavlock, R.J., Knudsen, T.B., Martin, M.T., Mortensen, H.M., Reif, D.M., Rotroff, D.M., Shah, I., Richard, A.M., Dix, D.J., 2010. In vitro screening of environmental chemicals for targeted testing prioritization: the ToxCast project. Environ. Health Perspect. 118 (4), 485–492.
Judson, R.S., Houck, K.A., Watt, E.D., Thomas, R.S., 2017. On selecting a minimal set of in vitro assays to reliably determine estrogen agonist activity. Regul. Toxicol. Pharmacol. 91, 39–49.
Judson, R.S., Magpantay, F.M., Chickarmane, V., Haskell, C., Tania, N., Taylor, J., Xia, M., Huang, R., Rotroff, D.M., Filer, D.L., Houck, K.A., Martin, M.T., Sipes, N., Richard, A.M., Mansouri, K., Setzer, R.W., Knudsen, T.B., Crofton, K.M., Thomas, R. S., 2015. Integrated Model of Chemical Perturbations of a Biological Pathway Using 18 In Vitro High-Throughput Screening Assays for the Estrogen Receptor. Toxicol. Sci. 148 (1), 137–154.
Kim, S., Thiessen, P.A., Bolton, E.E., Bryant, S.H., 2015. PUG-SOAP and PUG-REST: web services for programmatic access to chemical information in PubChem. Available Nucleic Acids Res 43, W605–W611. https://pubchemdocs.ncbi.nlm.nih.gov/pug-rest.
Kleinstreuer, N.C., Ceger, P.C., Allen, D.G., Strickland, J., Chang, X., Hamm, J.T., Casey, W.M., 2016. A Curated Database of Rodent Uterotrophic Bioactivity. Environ Health Perspect 124 (5), 556–562.
Kornhuber, M., Dunst, S., Schönfelder, G., Oelgeschläger, M., 2021. The E-Morph Assay: Identification and characterization of environmental chemicals with estrogenic activity based on quantitative changes in cell-cell contact organization of breast cancer cells. Environ. Int. 149, 106411. https://doi.org/10.1016/j.envint.2021.106411.
Kuhl, H., 2005. Pharmacology of estrogens and progestogens: influence of different routes of administration. Climacteric 8 (sup1), 3–63.
Landrum, G.A. RDKit: Open-source cheminformatics. http://www.rdkit.org; 2006.
Linusson, H. Nonconformist. http://donlnz.github.io/nonconformist/; 2015.
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. Proceedings of Machine Learning Research: PMLR; 201.
Livak, K.J., Schmittgen, T.D., 2001. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. Methods 25, 402–408.
Ma, J., Sheridan, R.P., Liaw, A., Dahl, G.E., Svetnik, V., 2015. Deep neural nets as a method for quantitative structure-activity relationships. J. Chem. Inf. Model. 55 (2), 263–274.
Maggiora, G., Vogt, M., Stumpfe, D., Bajorath, Jürgen, 2014. Molecular similarity in medicinal chemistry. J. Med. Chem. 57 (8), 3186–3204.
Malo, N., Hanley, J.A., Cerquozzi, S., Pelletier, J., Nadon, R., 2006. Statistical practice in high-throughput screening data analysis. Nat. Biotechnol. 24 (2), 167–175.
Mansouri, K., Abdelaziz, A., Rybacka, A., Roncaglioni, A., Tropsha, A., Varnek, A., Zakharov, A., Worth, A., Richard, A.M., Grulke, C.M., Trisciuzzi, D., Fourches, D., Horvath, D., Benfenati, E., Muratov, E., Wedebye, E.B., Grisoni, F., Mangiatordi, G. F., Incisivo, G.M., Hong, H., Ng, H.W., Tetko, I.V., Balabin, I., Kancherla, J., Shen, J., Burton, J., Nicklaus, M., Cassotti, M., Nikolov, N.G., Nicolotti, O., Andersson, P.L., Zang, Q., Politi, R., Beger, R.D., Todeschini, R., Huang, R., Farag, S., Rosenberg, S.A., Slavov, S., Hu, X., Judson, R.S., 2016. CERAPP: Collaborative Estrogen Receptor Activity Prediction Project. Environ. Health Perspect. 124 (7), 1023–1033.
Mayr, A.; Klambauer, G.; Unterthiner, T.; Hochreiter, S. DeepTox: Toxicity Prediction using Deep Learning. Front. Env. Sci. Switz 2016;3.
Morger, A., Mathea, M., Achenbach, J.H., Wolf, A., Buesen, R., Schleifer, K.J., Landsiedel, R., Volkamer, A., 2020. KnowTox: pipeline and case study for confident prediction of potential toxic effects of compounds in early phases of development. J. Cheminform. 12, 24.
Morger, A., Svensson, F., McShane, S.A., Gauraha, N., Norinder, U., Spjuth, O., Volkamer, A., 2021. Assessing the Calibration in Toxicological in Vitro Models with Conformal Prediction. J. Cheminformatics.
Norinder, U., Carlsson, L., Boyer, S., Eklund, M., 2014. Introducing conformal prediction in predictive modeling. A transparent and flexible alternative to applicability domain determination. J. Chem. Inf. Model. 54 (6), 1596–1603.
Norinder, U., Rybacka, A., Andersson, P.L., 2016. Conformal prediction to define applicability domain - A case study on predicting ER and AR binding. SAR QSAR Environ. Res. 27 (4), 303–316.
OECD. Test No. 416: Two-Generation Reproduction Toxicity edˆeds: OECD Publishing; 2001.
OECD. Test No. 493: Performance-Based Test Guideline for Human Recombinant Estrogen Receptor (hrER) In Vitro Assays to Detect Chemicals with ER Binding Affinity edˆeds: OECD Publishing; 2015.
OECD. Test No. 455: Performance-Based Test Guideline for Stably Transfected Transactivation In Vitro Assays to Detect Estrogen Receptor Agonists and Antagonists edˆeds: OECD Publishing; 2016.
OECD. New Scoping Document on in vitro and ex vivo Assays for the Identification of Modulators of Thyroid Hormone Signalling edˆeds; 2017.
OECD. Test No. 443: Extended One-Generation Reproductive Toxicity Study edˆeds: OECD Publishing; 2018a.
OECD. Test No. 451: Carcinogenicity Studies edˆeds: OECD Publishing; 2018b.
OECD. Test No. 452: Chronic Toxicity Studies edˆeds: OECD Publishing; 2018c.
Paterni, I., Granchi, C., Minutolo, F., 2017. Risks and benefits related to alimentary exposure to xenoestrogens. Crit. Rev. Food Sci. Nutr. 57 (16), 3384–3404.
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., 2011. Scikit-learn: machine learning in Python. J. Mach. Learn. Res. 12(Oct):2825–2830.
Pons, J.-L., Labesse, G., 2009. @TOME-2: a new pipeline for comparative modeling of protein-ligand complexes. Nucleic Acids Res 37 (Web Server), W485–W491.
Proença, S., Escher, B.I., Fischer, F.C., Fisher, C., Grégoire, Sébastien, Hewitt, N.J., Nicol, B., Paini, A., Kramer, N.I., 2021. Effective exposure of chemicals in in vitro cell systems: A review of chemical distribution models. Toxicol. In Vitro 73, 105133. https://doi.org/10.1016/j.tiv.2021.105133.
Raies, A.B., Bajic, V.B., 2016. In silico toxicology: computational methods for the prediction of chemical toxicity. Wiley Interdiscip Rev. Comput. Mol. Sci. 6 (2), 147–172.
Reif, D.M., Martin, M.T., Tan, S.W., Houck, K.A., Judson, R.S., Richard, A.M., Knudsen, T.B., Dix, D.J., Kavlock, R.J., 2010. Endocrine profiling and prioritization of environmental chemicals using ToxCast data. Environ. Health Perspect 118 (12), 1714–1720.
Rochester, J.R., Bolden, A.L., 2015. Bisphenol S and F: A Systematic Review and Comparison of the Hormonal Activity of Bisphenol A Substitutes. Environ. Health Perspect 123 (7), 643–650.
Rothman, M.S., Carlson, N.E., Xu, M., Wang, C., Swerdloff, R., Lee, P., Goh, V.H.H., Ridgway, E.C., Wierman, M.E., 2011. Reexamination of testosterone, dihydrotestosterone, estradiol and estrone levels across the menstrual cycle and in postmenopausal women measured by liquid chromatography-tandem mass spectrometry. Steroids 76 (1-2), 177–182.
Rotroff, D.M., Dix, D.J., Houck, K.A., Knudsen, T.B., Martin, M.T., McLaurin, K.W., Reif, D.M., Crofton, K.M., Singh, A.V., Xia, M., Huang, R., Judson, R.S., 2013. Using in vitro high throughput screening assays to identify potential endocrine-disrupting chemicals. Environ. Health Perspect 121 (1), 7–14.
Russell, W.M.S., Burch, R.L., 1959. Principles of humane experimental technique edˆeds. Methuen, London.
Schindelin, J., Arganda-Carreras, I., Frise, E., Kaynig, V., Longair, M., Pietzsch, T., Preibisch, S., Rueden, C., Saalfeld, S., Schmid, B., Tinevez, J.-Y., White, D.J., Hartenstein, V., Eliceiri, K., Tomancak, P., Cardona, A., 2012. Fiji: an open-source platform for biological-image analysis. Nat. Methods 9 (7), 676–682.
Schlotz, N., Kim, G.-J., Jäger, S., Günther, S., Lamy, E., 2017. In vitro observations and in silico predictions of xenoestrogen mixture effects in T47D-based receptor transactivation and proliferation assays. Toxicol. In Vitro 45, 146–157.
Sirianni, R., Capparelli, C., Chimento, A., Panza, S., Catalano, S., Lanzino, M., Pezzi, V., Andò, S., 2012. Nandrolone and stanozolol upregulate aromatase expression and further increase IGF-I-dependent effects on MCF-7 breast cancer cell proliferation. Mol. Cell Endocrinol. 363 (1-2), 100–110.
Sun, J., Carlsson, L., Ahlberg, E., Norinder, U., Engkvist, O., Chen, H., 2017. Applying Mondrian Cross-Conformal Prediction To Estimate Prediction Confidence on Large Imbalanced Bioactivity Data Sets. J. Chem. Inf. Model. 57 (7), 1591–1598.
Sung, H., Ferlay, J., Siegel, R.L., Laversanne, M., Soerjomataram, I., Jemal, A., Bray, F., 2021. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J. Clin. 71 (3), 209–249.
Svensson, F., Norinder, U., Bender, A., 2017a. Improving Screening Efficiency through Iterative Screening Using Docking and Conformal Prediction. J. Chem. Inf. Model. 57 (3), 439–444.
Svensson, F., Norinder, U., Bender, A., 2017b. Modelling compound cytotoxicity using conformal prediction and PubChem HTS data. Toxicol Res (Camb) 6 (1), 73–80.
Tropsha, A., 2010. Best Practices for QSAR Model Development, Validation, and Exploitation. Mol. Inform. 29 (6-7), 476–488.
U.S. EPA CCTE. Center for Computational Toxicology and Exposure, ToxCast and Tox21 Data Spreadsheet. [Online]. Available: https://figshare.com/articles/dataset/ToxCast_and_Tox21_Data_Spreadsheet/6062503. Accessed: 2017-06-23.
U.S. NIEHS. Reference Chemical Lists for Test Method Development and Evaluation [Online]. Available: https://ntp.niehs.nih.gov/whatwestudy/niceatm/resources-for-test-method-developers/refchem/index.html. Accessed: 2019-07-17.
Vovk, V., Gammerman, A., Shafer, G., 2005. Algorithmic Learning in a Random World edˆeds. Springer, Boston, MA.
Welinder, C., Ekblad, L., 2011. Coomassie staining as loading control in Western blot analysis. J Proteome Res 10 (3), 1416–1419.
Wetzel, C., Pifferi, S., Picci, C., Gök, C., Hoffmann, D., Bali, K.K., Lampe, A., Lapatsina, L., Fleischer, R., Smith, E.S.J., Bégay, V., Moroni, M., Estebanez, L., Kühnemund, J., Walcher, J., Specker, E., Neuenschwander, M., von Kries, J.P., Haucke, V., Kuner, R., Poulet, J.F.A., Schmoranzer, J., Poole, K., Lewin, G.R., 2017. Small-molecule inhibition of STOML3 oligomerization reverses pathological mechanical hypersensitivity. Nat. Neurosci. 20 (2), 209–218.

WHO/IPCS. IPCS global assessment of the state-of-the-science of endocrine disruptors. WHO/PCS/EDC/022 2002:35-50.
WHO/UNEP. State of the science of endocrine disrupting chemicals - 2012 edˆeds; 2013.
WHO/UNEP. Global Chemicals Outlook II - From Legacies to Innovative Solutions: Implementing the 2030 Agenda for Sustainable Development - Synthesis Report. 2019.
Wild, C.P.; Weiderpass, E.; Stewart, B.W.; editors. World Cancer Report: Cancer Research for Cancer Prevention edˆeds. Lyon, France: International Agency for Research on Cancer. Available from: http://publications.iarc.fr/586. Licence: CC BY-NC-ND 3.0 IGO; 2020.
Wolber, G., Langer, T., 2005. LigandScout: 3-D pharmacophores derived from protein-bound ligands and their use as virtual screening filters. Available J. Chem. Inf. Model. 45, 160–169. https://www.inteligand.com/ligandscout/.
Yager, J.D., Davidson, N.E., 2006. Estrogen carcinogenesis in breast cancer. N. Engl. J. Med. 354 (3), 270–282.
Yu, H., Caldwell, D.J., Suri, R.P., 2019. In vitro estrogenic activity of representative endocrine disrupting chemicals mixtures at environmentally relevant concentrations. Chemosphere 215, 396–403.
Zhang, J., Norinder, U., Svensson, F., 2021. Deep Learning-Based Conformal Prediction of Toxicity. J. Chem. Inf. Model. 61 (6), 2648–2657.
Zhang, J.H.; Chung, T.D.; Oldenburg, K.R. A Simple Statistical Parameter for Use in Evaluation and Validation of High Throughput Screening Assays. J Biomol Screen 1999;4:67-73.

4.3 ChemBioSim: enhancing conformal prediction of *in vivo* toxicity by use of predicted bioactivities

In the KnowTox (see Section 4.1) and E-Morph Screen (see Section 4.2) studies it was explored how *in silico* toxicity prediction can guide *in vitro* testing. Since individual *in vitro* assays cannot completely represent complex *in vivo* systems, it is desirable to create *in silico* models for the prediction of *in vivo* toxicity. Besides direct interactions with off-targets, *in vivo* toxicological effects of a chemical are also affected by its exposure, as well as by complex cellular mechanisms and downstream pathways. Therefore, predicting *in vivo* toxicity solely from the molecular structure is challenging. In the following study, we will investigate if including information from *in vitro* assay outcomes in the form of bioactivity descriptors can improve the performance of CP models built for the genotoxicity, liver toxicity, and cardiotoxicity endpoints. Using standard bioactivity descriptors would require that each query molecule be synthesised and tested in all the descriptor-specific assays before a prediction can be made. Therefore, one aim of this work will be to explore if such bioactivity descriptors can be computed from CP models. CP models will be built on *in vitro* datasets and the predicted p-values for query compounds will be used as bioactivity descriptors. The performance of CP models based on such bioactivity descriptors will be compared to such using chemical (i.e. molecular and physicochemical) descriptors only or a combination of both.

Contribution:

Middle author
Conceptual design (20%)
Computational experiments (20%)
Visualization (0%)
Manuscript preparation (20%)

JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

pubs.acs.org/jcim Article

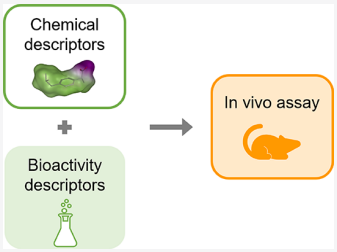
ChemBioSim: Enhancing Conformal Prediction of In Vivo Toxicity by Use of Predicted Bioactivities

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	AMES assay: eChemPortal,[26] Benigni et al.,[28] Hansen et al.[29]
	• chromosome aberration (CA) assay	
	• 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	Garcia de Lomana et al.[31]
	• 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	
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]

https://doi.org/10.1021/acs.jcim.1c00451
J. Chem. Inf. Model. 2021, 61, 3255−3272

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

https://doi.org/10.1021/acs.jcim.1c00451
J. Chem. Inf. Model. 2021, 61, 3255−3272

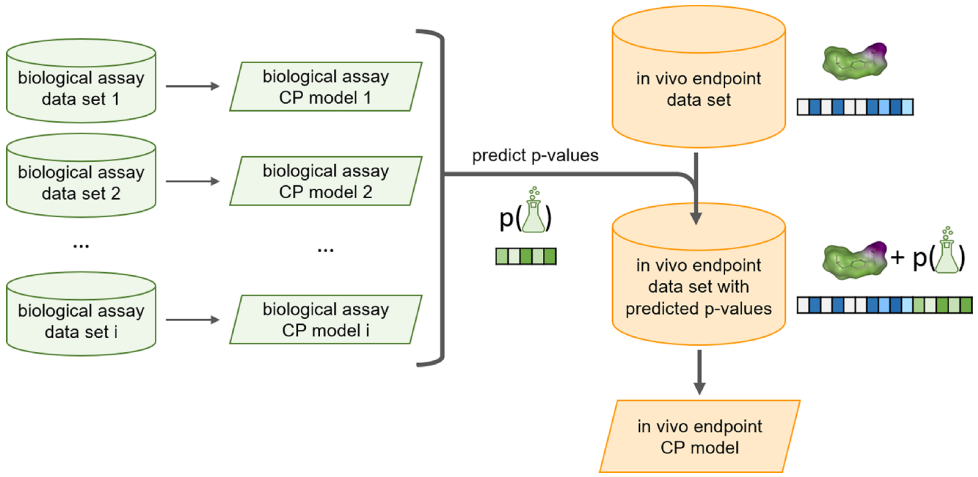


Figure 1. Workflow for the derivation of the bioactivity descriptors for the in vivo toxicity CP models. For each biological assay, a conformal prediction model is built and used to predict the *p*-values of the compounds in the three in vivo endpoint data sets. These predicted *p*-values are used as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

https://doi.org/10.1021/acs.jcim.1c00451
J. Chem. Inf. Model. 2021, 61, 3255−3272

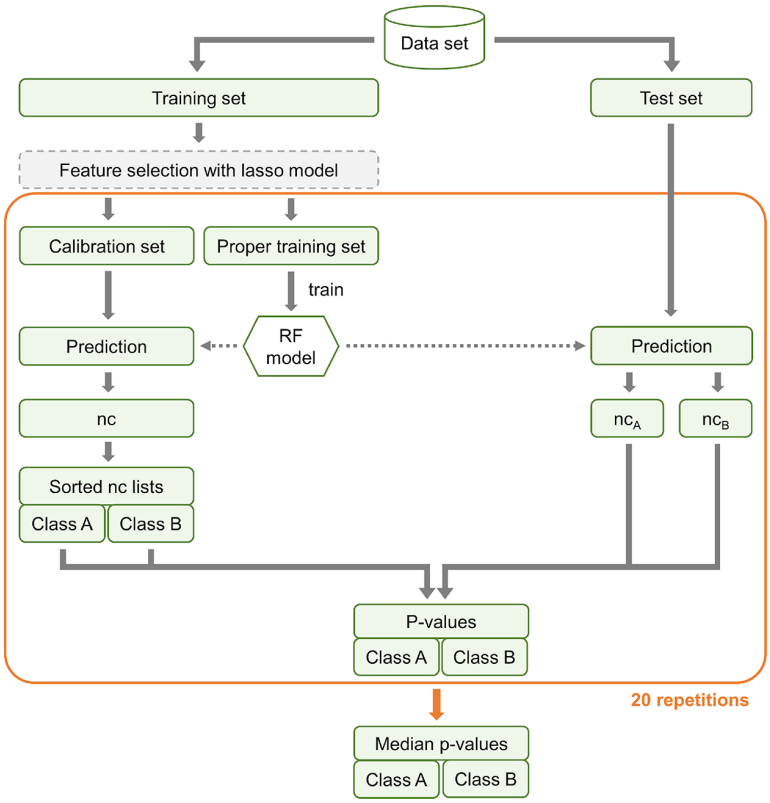


Figure 2. Workflow of the aggregated Mondrian CP set up for the development of the models for the biological assays and the in vivo endpoints. The aggregated CP framework included 20 random splits in calibration and proper training data sets, on which individual RF models were trained, and the resulting *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

https://doi.org/10.1021/acs.jcim.1c00451
J. Chem. Inf. Model. 2021, 61, 3255−3272

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]

https://doi.org/10.1021/acs.jcim.1c00451
J. Chem. Inf. Model. 2021, 61, 3255−3272

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).

https://doi.org/10.1021/acs.jcim.1c00451
J. Chem. Inf. Model. 2021, 61, 3255−3272

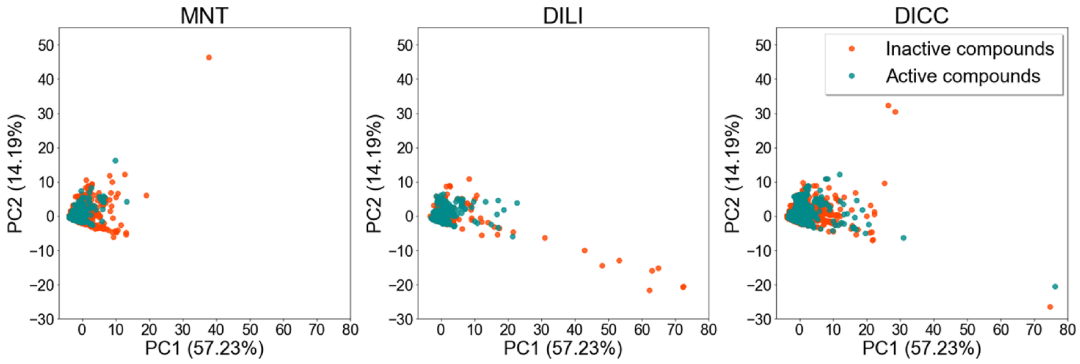


Figure 3. Principal component analysis based on a selection of interpretable molecular descriptors generated with RDKit on the merged in vivo toxicity data sets. Inactive compounds are colored in red and active compounds in green. The variance explained by the first two principal components is indicated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

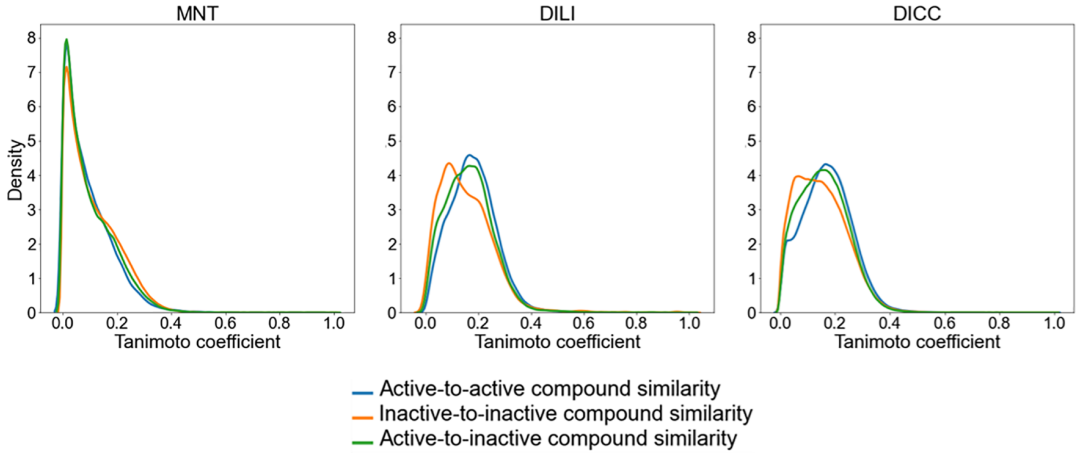


Figure 4. Distribution of pairwise Tanimoto coefficients based on atom-pair fingerprints for three types of compound pairs: (a) active-to-active (blue), (b) inactive-to-inactive (orange), and (c) active-to-inactive (green).

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

https://doi.org/10.1021/acs.jcim.1c00451
J. Chem. Inf. Model. 2021, 61, 3255−3272

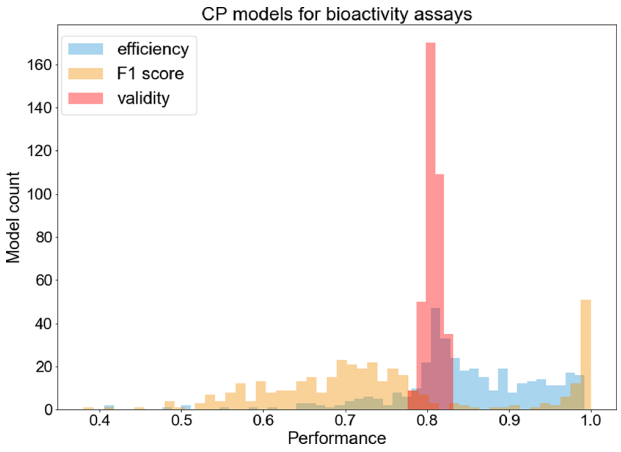


Figure 5. Histogram of the performance distribution of the CP models for the biological assays. All models were valid but their efficiencies and F1 scores showed a high degree of variability.

"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 "CHEM-BIO" 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

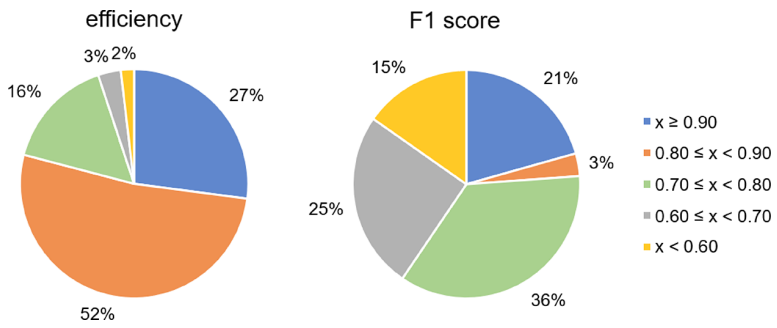


Figure 6. Percentage of the 373 bioactivity CP models showing mean efficiencies and mean F1 scores in the four given ranges.

https://doi.org/10.1021/acs.jcim.1c00451
J. Chem. Inf. Model. 2021, 61, 3255−3272

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).

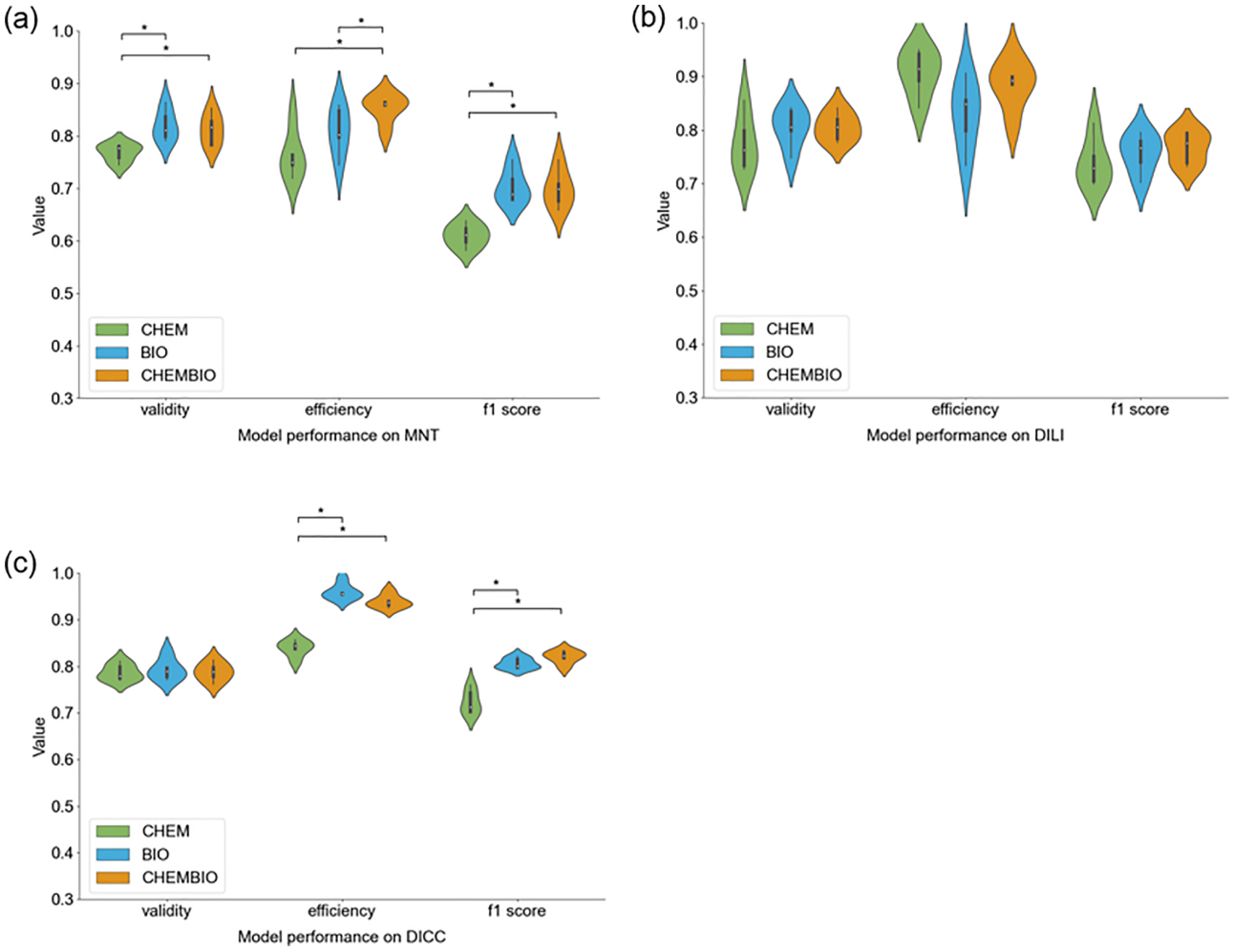


Figure 7. Distribution of the validity, efficiency, and F1 score values obtained within the 5-fold CV framework for the (a) MNT, (b) DILI, and (c) DICC CP models built on the different descriptor sets after feature selection. The CHEM descriptor set includes the molecular fingerprint and physicochemical 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.

https://doi.org/10.1021/acs.jcim.1c00451
J. Chem. Inf. Model. 2021, 61, 3255−3272

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

https://doi.org/10.1021/acs.jcim.1c00451
J. Chem. Inf. Model. 2021, 61, 3255−3272

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

https://doi.org/10.1021/acs.jcim.1c00451
J. Chem. Inf. Model. 2021, 61, 3255−3272

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

https://doi.org/10.1021/acs.jcim.1c00451
J. Chem. Inf. Model. 2021, 61, 3255−3272

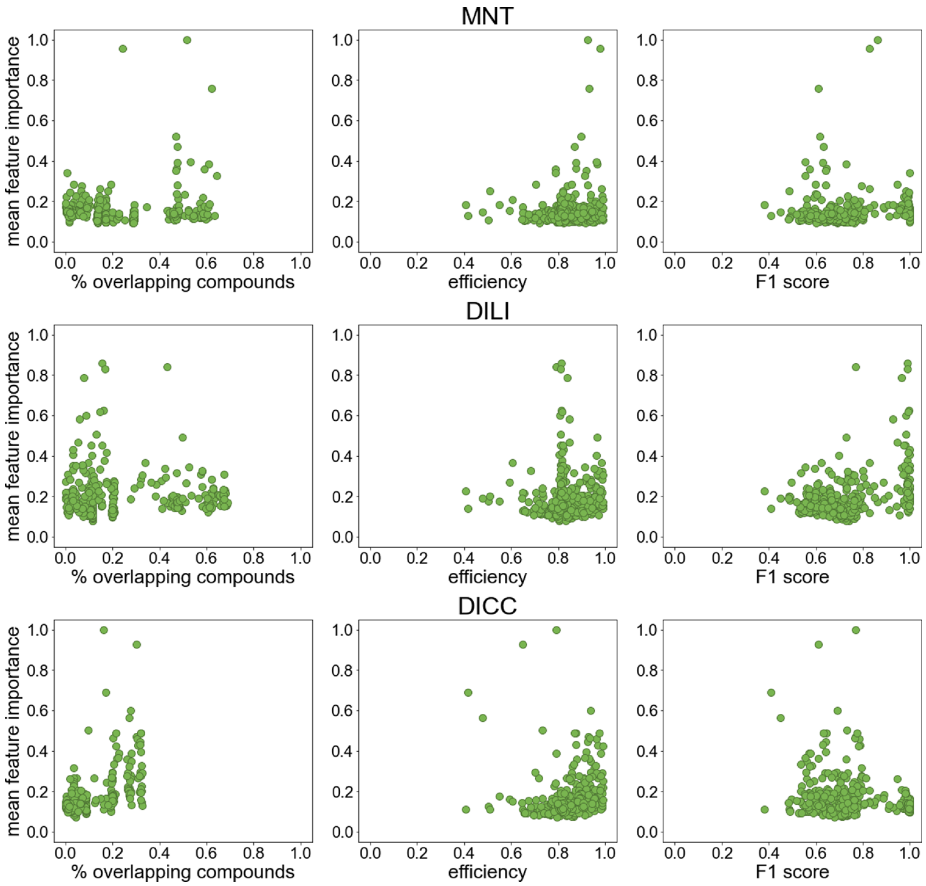


Figure 8. Mean feature importance reported by the RF model for the bioactivity descriptors in relationship with the percentage of overlapping compounds (of the in vivo data set), the efficiency and F1 score of the models for each biological assay. For each of the 373 biological assays, the highest mean 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

https://doi.org/10.1021/acs.jcim.1c00451
J. Chem. Inf. Model. 2021, 61, 3255−3272

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.;

https://doi.org/10.1021/acs.jcim.1c00451
J. Chem. Inf. Model. 2021, 61, 3255−3272

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.

SUPPORTING INFORMATION

ChemBioSim: Enhancing Conformal Prediction of in vivo Toxicity by Use of Predicted Bioactivities

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, 10117 Berlin, Germany

[4] MTM Research Centre, School of Science and Technology, Örebro University, SE-70182 Örebro, Sweden

* johannes.kirchmair@univie.ac.at;

miriam.mathea@basf.com

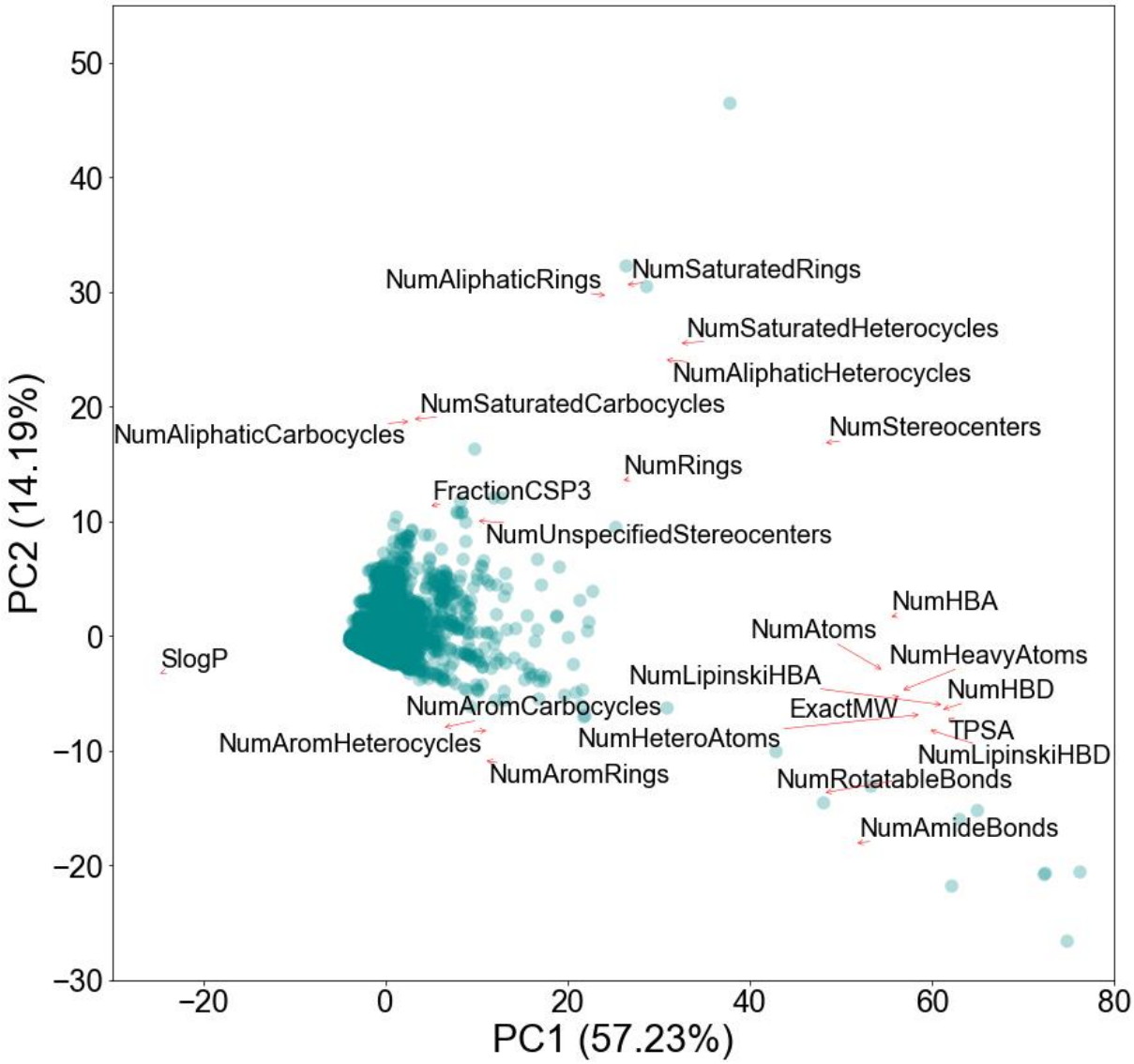


Figure S1. Loadings plot of the PCA based on a selection of interpretable molecular descriptors generated with RDKit on the global in vivo toxicity data set. The loadings plot shows how strongly each feature influences a principal component.

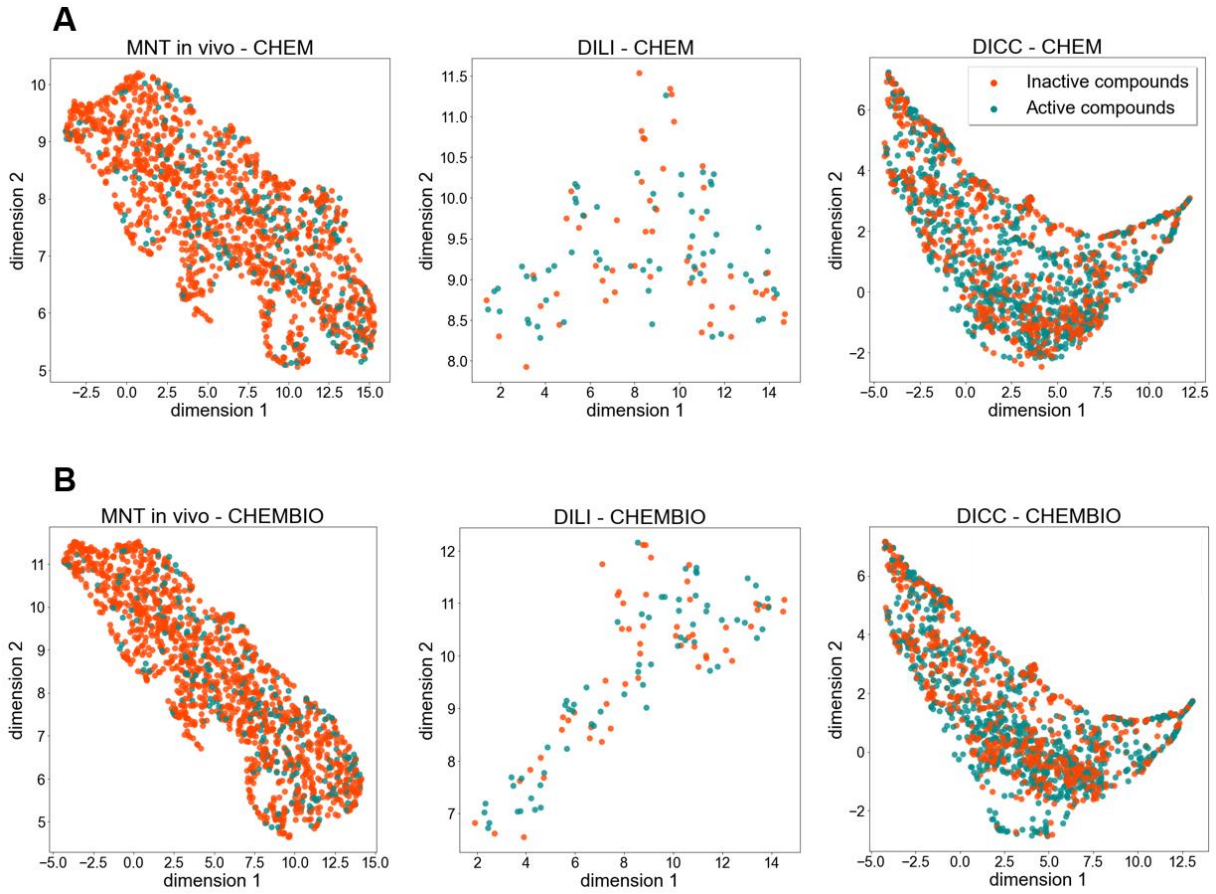


Figure S2. UMAP projections for the three in vivo endpoints (MNT in vivo, DILI and DICC) on (A) the CHEM descriptor set and (B) the CHEMBIO descriptor set.

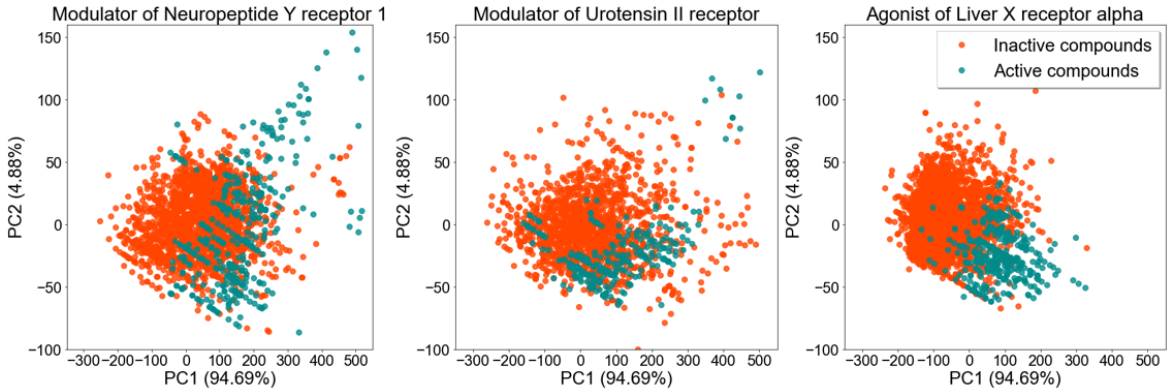
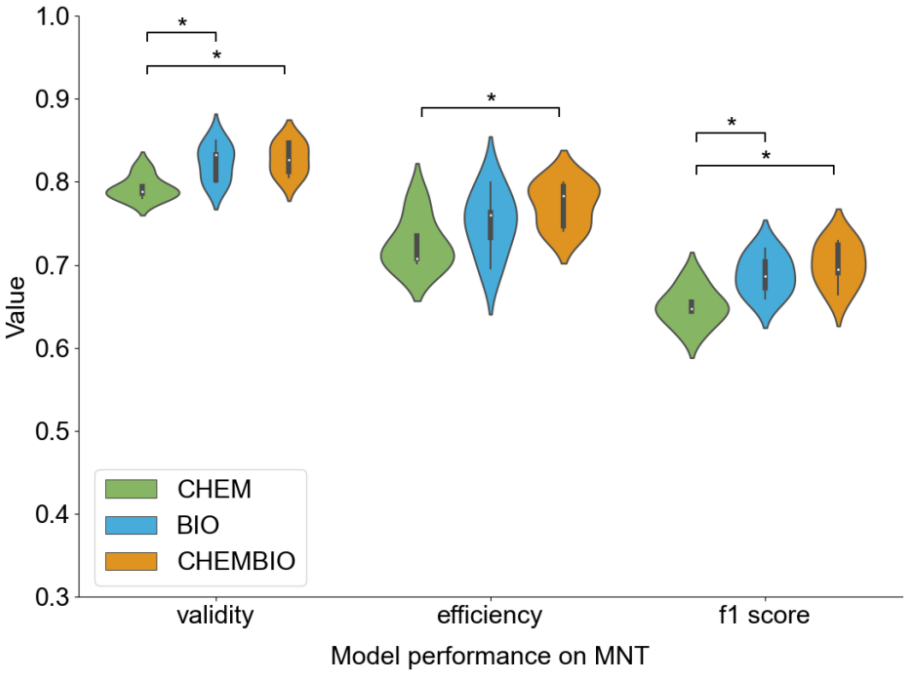
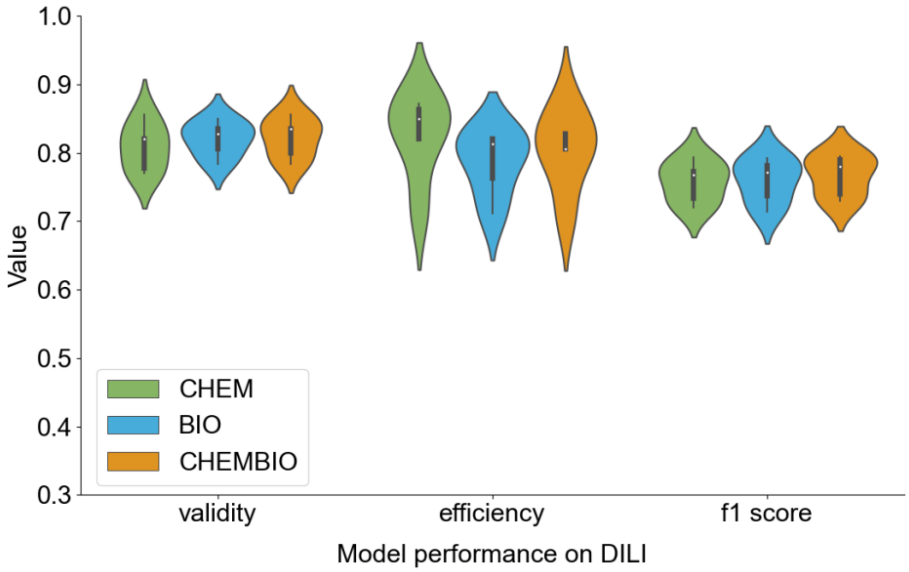


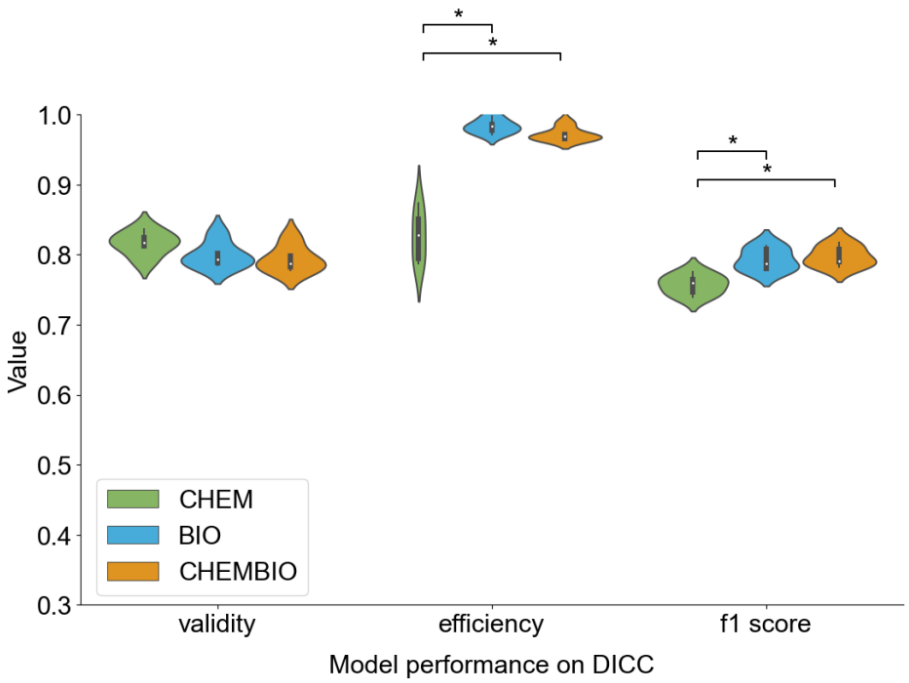
Figure S3. Principal component analysis based on a selection of interpretable molecular descriptors generated with RDKit. The PCA was derived from the merged data set of three eMolTox assays ("Modulator of Neuropeptide Y receptor type 1", "Modulator of Urotensin II receptor" and "Agonist of Liver X receptor 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.


1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
Value
*
*
*
*
*
CHEM
BIO
CHEMBIO
validity
efficiency
f1 score
Model performance on MNT

(a)


1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
Value
CHEM
BIO
CHEMBIO
validity
efficiency
f1 score
Model performance on DILI

(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.

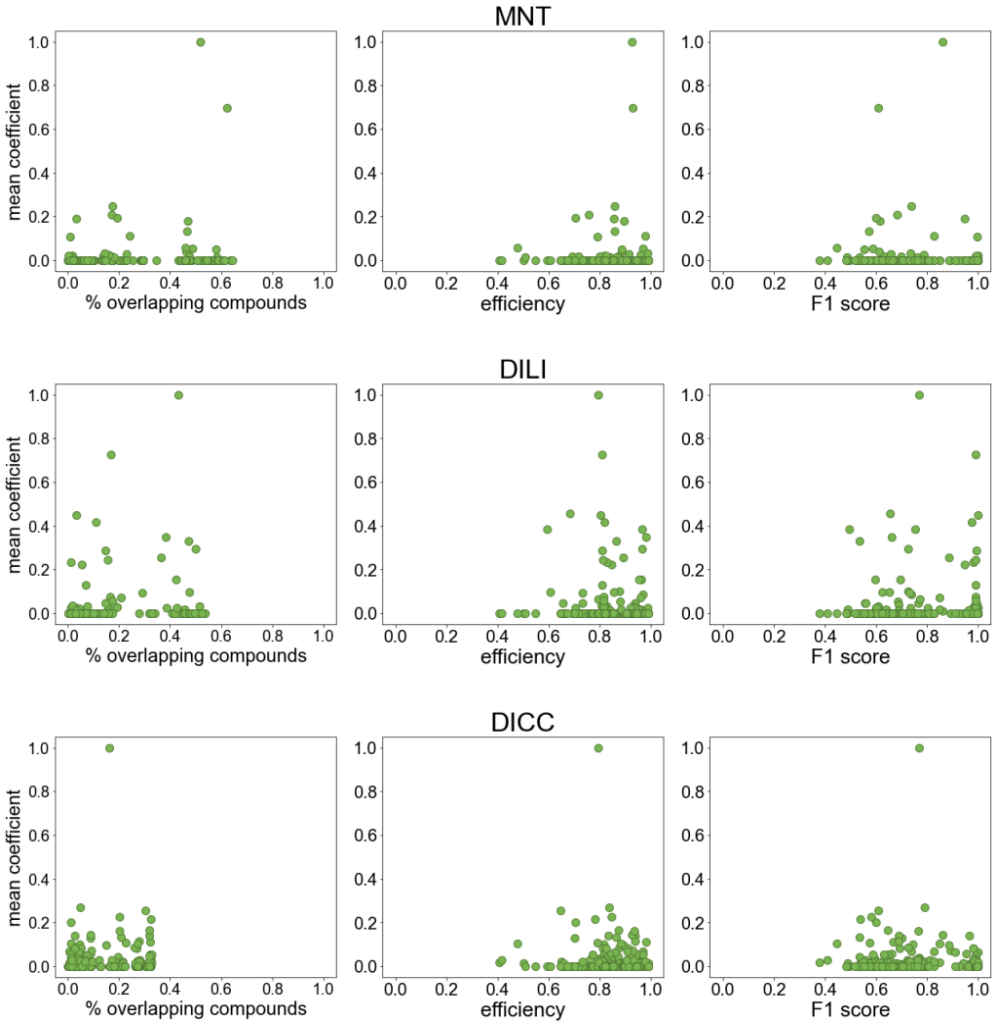


Figure S5. Mean coefficient reported by the lasso model for the bioactivity descriptors in relationship with the percentage of overlapping compounds (of the in vivo data set), the efficiency and F1 score of the models for each biological assay. For each of the 373 biological assays, the highest mean 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.

4.4 Assessing the calibration in toxicological *in vitro* models with conformal prediction

In our previous studies, CP was used for confidence estimation (see Sections 4.1 and 4.2) and to generate bioactivity descriptors (see Section 4.3). In the following project, a third potential application of CP will be explored, i.e. the use of CP for the mitigation of data drift effects. While the CP framework is designed to yield valid models given that training and test data are exchangeable, this exchangeability assumption is not always fulfilled. CP models are typically well-calibrated within a cross-validation, but validity may drop when the models are applied to a batch of new query compounds. Violations of the exchangeability assumption may not be uncommon. They occur, for example, when a new chemical space is explored, or when data is produced in different laboratories. In the following work, data drift effects will be studied with the example of the Tox21 datasets, a collection of toxicological *in vitro* datasets, that were originally produced for a data challenge and subsequently released in three subsets. Multiple potential reasons for poor model calibration will be discussed and a strategy to mitigate effects of data drifts exploiting the CP framework will be introduced.

Contribution:

First author
Conceptual design (40%)
Computational experiments (100%)
Visualization (100%)
Manuscript preparation (80%)

Morger *et al. J Cheminform* (2021) 13:35
https://doi.org/10.1186/s13321-021-00511-5

Journal of Cheminformatics

RESEARCH ARTICLE Open Access

Assessing the calibration in toxicological in vitro models with conformal prediction

Andrea Morger[1], Fredrik Svensson[2], Staffan Arvidsson McShane[3], Niharika Gauraha[3,4], Ulf Norinder[3,5,6], Ola Spjuth[3†] and Andrea Volkamer[1*†]

Abstract

Machine learning methods are widely used in drug discovery and toxicity prediction. While showing overall good performance in cross-validation studies, their predictive power (often) drops in cases where the query samples have drifted from the training data's descriptor space. Thus, the assumption for applying machine learning algorithms, that training and test data stem from the same distribution, might not always be fulfilled. In this work, conformal prediction is used to assess the calibration of the models. Deviations from the expected error may indicate that training and test data originate from different distributions. Exemplified on the Tox21 datasets, composed of chronologically released Tox21Train, Tox21Test and Tox21Score subsets, we observed that while internally valid models could be trained using cross-validation on Tox21Train, predictions on the external Tox21Score data resulted in higher error rates than expected. To improve the prediction on the external sets, a strategy exchanging the calibration set with more recent data, such as Tox21Test, has successfully been introduced. We conclude that conformal prediction can be used to diagnose data drifts and other issues related to model calibration. The proposed improvement strategy—exchanging the calibration data only—is convenient as it does not require retraining of the underlying model.

Keywords: Toxicity prediction, Conformal prediction, Data drifts, Applicability domain, Calibration plots, Tox21 datasets

Introduction

Machine learning (ML) methods are ubiquitous in drug discovery and toxicity prediction [1, 2]. In silico toxicity prediction is typically used to guide toxicity testing in early phases of drug design [3]. With more high-quality standardised data available, the (potential) impact of ML methods in regulatory toxicology is growing [4]. The collection of available toxicity data is increasing, thanks in part to high-throughput screening programs such as ToxCast [5] and Tox21 [6, 7], but also with public-private partnerships such as the eTOX and eTRANSAFE projects, which focus on the sharing of (confidential) toxicity data and ML models across companies [8, 9]. In any case, no matter which underlying data and ML method is used, it is essential to know or assess if the ML model can be reliably used to make predictions on a new dataset.

Hence, validation of ML models is crucial to assess their predictivity. Several groups investigated random vs. rational selection of optimal test/training sets, e.g. using cluster- or activity-based splits, with the goal of better reflecting the true predictive power of established models [10–14]. Martin et al. [11] showed that rational selection of training and test sets—compared to random splits—generated better statistical results on the (internal) test sets. However, the performance of both types of regression models on the—artificially created—external evaluation set was comparable.

Thus, further metrics to define the applicability domain (AD), the domain in which an ML classifier can reliably

*Correspondence: andrea.volkamer@charite.de
†Ola Spjuth and Andrea Volkamer—Shared senior authorship
[1] In Silico Toxicology and Structural Bioinformatics, Institute of Physiology, Charité Universitätsmedizin, Berlin, Germany
Full list of author information is available at the end of the article

be applied [15–21], are needed. Besides traditional metrics accounting for chemical space coverage, Sheridan [20] discussed uncertainty prediction regression models, fitted with the activity prediction errors as labels and diverse AD metrics as descriptors (e.g. accounting for variation among RF tree predictions, predicted activity ranges with different confidence, or similarity to nearest neighbours). Since in classification models, the response/activity is a categorical value, only the chemical space remains to define the AD. Mathea et al. [15] categorised the available methods into novelty and confidence estimation techniques. The former consider the fit into the underlying chemical descriptor space as a whole, whereas the latter focus on the reliability of predictions, i.e. data points may be well embedded in the descriptor space but abnormal regarding their class label.

A popular method for confidence estimation is conformal prediction (CP), which has in recent years been widely applied in the drug discovery and toxicity prediction context [15, 22]. In CP, ML models are trained, and with the help of an additional calibration set (inductive conformal prediction [23]), the predictions are calibrated, i.e. ranked based on previously seen observations, resulting in so-called conformal p-values or simply p-values (not to be confused with statistical p-values from hypothesis testing). The design of the CP statistical framework guarantees that the error rate of the predictions will not exceed a user-specified significance level. The control of this significance level makes CP advantageous compared to traditional confidence estimation methods, such as distance from the decision boundary, or ensemble models [15].

ML algorithms rely on the assumption that the probability distribution of the training data and test data are *I.I.D.* (independent and identically distributed). For conformal prediction, a slightly weaker assumption in the form of exchangeability is assumed for producing well-calibrated models [24]. This assumption is nevertheless not always fulfilled, especially when training and test data come from different sources. For example, data drifts were observed between training and test data of the USPS (handwritten digits) and the Statlog Satellite (satellite image) datasets [25]. Similar observations were made in the toxicity prediction context when applying androgen receptor agonism CP models trained on publicly available data to an industrial dataset [26]. Some efforts to look at data exchangeability include studies using martingales to uncover exchangeability issues in an online setting [25].

In this work, we explored how the above described concepts of conformal prediction can be used to assess the quality of the model calibration when trained and applied on various toxicological in vitro datasets or subsets. For this purpose, the freely available Tox21 datasets [27], initially prepared for a data challenge to encourage model building and benchmarking toxicity prediction, were used. We show that conformal prediction allows us to identify data drifts between the Tox21 datasets, and we also propose strategies to mitigate this.

Data and methods

In this section, first the used Tox21 datasets are introduced. Second, the general conformal prediction framework along with aspects such as aggregation, evaluation and strategies to improve the calibration are described. Finally, the set-up and the individual computational experiments of this work are explained, including a reference to code and data availability.

Data collection, preprocessing and encoding

Tox21 datasets

The investigations in this work were performed on the freely available Tox21 datasets [27]. They consist of approximately 10,000 chemicals, which were tested on up to 12 endpoints of the nuclear receptor (NR) and stress response (SR) pathways. As the dataset was released in a challenge setting, the three subsets were chronologically published to the Tox21 Data Challenge participants: Tox21Train for training the models, Tox21Test as an intermediate set for the leaderboard to check the performance (and for participants to improve their models), and Tox21Score as the final dataset to determine the best performing models. The respective datasets were downloaded from the US National Center for Advancing Translational Sciences [28] on January 29th, 2019. Each compound was provided in sdf-format together with a binary value (0/non-toxic, or 1/toxic) for each of the 12 endpoints (*X* if no assay outcome was available for the compound). Note that throughout this manuscript, the Tox21 datasets are, consistently, referred to as Tox21Train, Tox21Test and Tox21Score, this should not be confused with additional training and test set splits necessary for the ML/CP model set-ups.

Data preprocessing

The datasets were standardised as described in Morger et al. [26]. Briefly, the IMI eTox standardiser tool was applied to discard non-organic compounds, to exert certain structure standardisation rules, to neutralise, and to remove salts [29]. Before and after applying the standardisation protocol, compounds with duplicate InChIs (IUPAC International Chemical Identifiers [30]) but disagreeing labels were discarded. Furthermore, remaining mixtures and fragments with less than four heavy atoms were removed. The numbers of data points available per dataset and endpoint after standardisation are presented

in Table 1. The corresponding numbers before standardisation can be found in the Additional file 1: Table S1.

Compound encoding

Converting molecules into numerical data was performed using the signature molecular descriptor [32, 33], using the program CPSign [34] version 0.7.14. The signature descriptor has been used extensively in previous QSAR studies [35–37]. In brief, the signature molecular descriptor enumerates all fragments of a molecule using a specified number of atomic bonds, often referred to as height, here using height 1 to 3 (e.g., height 1 creates fragments containing a center atom and all its one-bond connected atoms). This descriptor is often extremely sparse as there is a large number of fragments in a dataset and each molecule contains only a small set of these fragments. Herein, the count of each fragment was used; it is also possible to use a bit-type vector, where 0/1 indicates whether the fragment is present or not. The composition of the training set and hence the number of descriptors is different per endpoint. On average 36,721 ($\pm$ 2363 std) fragments were defined per endpoint in the Tox21Train set, whereas the signatures for Tox21Test and Tox21Score are based on the fragments in Tox21Train.

Modelling

Conformal prediction

Conformal prediction (CP) is a statistical framework, which provides means for confidence estimation [15, 38]. The baseline conformal predictor is the computationally efficient inductive conformal predictor (ICP) [23] (indicated in purple in Fig. 1a). An ICP operates on the output from an underlying model. To allow calibration of the outputs, the training set is divided into a proper training set and a calibration set. An underlying model, most often a machine learning model, is fitted on the proper training set, predictions are made for both the test and the calibration set compounds, and transformed into so-called nonconformity scores. In a binary Mondrian setting [38, 39], for each test compound two p-values are calculated, one per class, by comparing the outcome of each instance with the outcomes of the corresponding calibration set compounds. Given the two p-values and a predefined significance level $\epsilon = 1 -$ confidence level, a prediction set is calculated. The prediction set contains all class labels, for which the p-value is larger than the significance level. For more information on conformal prediction, see Alvarsson et al. [40] and Norinder et al. [41].

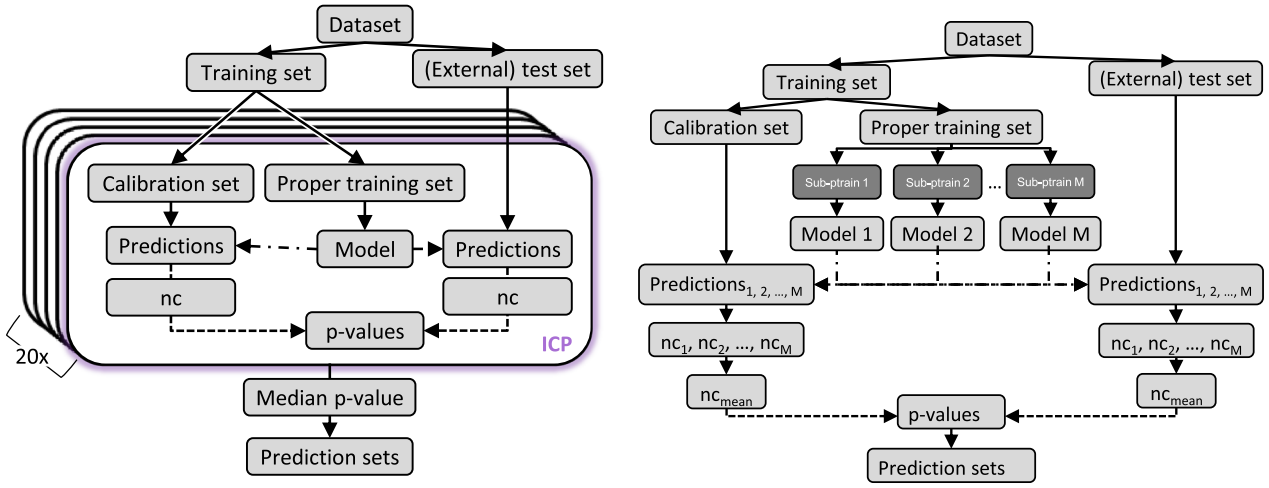
Aggregated conformal prediction methods

To reduce the variance in efficiency of ICPs, multiple conformal predictors can typically be aggregated [42, 43] (see Fig. 1). In the commonly used aggregated conformal prediction (ACP) [43] aggregation method, the training set is randomly split n times into a proper training set and a calibration set, with which n ICPs are trained and calibrated (Fig. 1a). The p-values resulting from the different ICPs are then averaged. While the consolidation of multiple models stabilises the predictions, a uniform distribution of the p-values is not necessarily observed after their averaging [42].

The influence of ACPs on the calibration can be analysed by additionally incorporating the recently developed synergy conformal prediction (SCP) method (Fig. 1b) [44]. In the SCP, one fixed calibration set is

Table 1 Number of compounds (separated as actives and inactives) available per Tox21 dataset and endpoint after standardisation. The full names for the endpoints are adopted from Huang et al. [31]

Endpoint	Tox21Train		Tox21Test		Tox21Score	
	Actives	Inactives	Actives	Inactives	Actives	Inactives
Aryl hydrocarbon receptor (NR_AhR)	933	6687	29	236	71	506
Androgen receptor, full length (NR _AR)	373	8370	3	282	11	549
Androgen receptor, ligand binding domain (NR_AR_LBD)	295	7742	4	242	8	543
Aromatase (NR_Aromatase)	338	6362	18	192	36	466
Estrogen receptor, full length (NR_ER)	901	6290	27	231	49	441
Estrogen receptor, ligand binding domain (NR_ER_LBD)	419	7763	10	270	20	548
Peroxisome proliferator-activated receptor gamma (NR_PPAR)	204	7414	15	245	31	543
Nuclear factor (erythroid-derived 2)-like 2/antioxidant responsive element (SR_ARE)	1032	5653	47	181	89	433
ATAD5 (SR_ATAD5)	322	8179	25	240	36	554
Heat shock factor response element (SR_HSE)	386	7233	10	250	19	558
Mitochondrial membrane potential (SR_MMP)	1094	5719	38	195	56	457
p53 (SR_p53)	515	7542	28	234	40	543



a Aggregated conformal predictor (ACP) **b** Synergy conformal predictor (SCP)

Fig. 1 Inductive conformal predictor (ICP) and the aggregated conformal prediction methods used in this study. **a** Aggregated conformal prediction (ACP) and ICP (box with purple edge): The dataset is split into a training set and a test set. The training set is further split into a proper training set to train the model and a calibration set. The predictions made for the test set compounds are used to calculate nonconformity scores (nc) and compared to nonconformity scores in the calibration set to calculate p-values and generate prediction sets. In ACP, multiple models are trained and calibrated with randomly selected proper training and calibration sets, and p-values from these are averaged. **b** Synergy conformal prediction (SCP): In order to ensure a uniform distribution of p-values, SCP averages the nonconformity scores instead. Multiple models are trained on (subsets of) the proper training set and with each model predictions are made for the test set and for a fixed calibration set

randomly selected, and the proper training set is split into n subsets to train multiple sub-models. Note that the analysis of other options to build an SCP, e.g. training several models using different ML algorithms on the same (sub)set, is out of scope for this work. The predictions made with every sub-model are aggregated before calculating the p-values and prediction sets. A fixed calibration set reduces the number of available training compounds, but the needlessness of averaging p-values ensures a uniform distribution of the latter and hence leads to theoretically valid models [44].

Model evaluation

CP models are typically evaluated by their validity and efficiency [15]. Validity, for a given significance level, is defined as the ratio of prediction sets that contain the true label. The efficiency of a model is a way to measure the information content of the model, and we herein use the most widely used efficiency metric: ratio of single label sets at a given significance level. In binary CP, the possible prediction sets are {∅}, {0}, {1} and {0,1}, where only the {0} and {1} (i.e. single label sets) are informative, and 'empty' and 'both' sets are uninformative in a sense. Thus, the fraction of single label sets should be maximised for best efficiency.

Model calibration

When evaluating the predictive performance on a test set, deviations from the underlying assumption that all data come from the same distribution will lead to predictions that are invalid and hence the results might be misleading. In this work, we use calibration plots to identify deviations from acceptable levels of calibration, and also discuss potential mitigation strategies.

Assessing model calibration

In a conformal prediction setting, the observed error rate of predictions is theoretically proven to not be larger than the specified significance level. In return, any deviations between these values may indicate data drifts (or other causes for the deviations, such as a too small test set). The level of calibration can be visualised in a so-called calibration plot, where the observed error rate (y-axis) is plotted versus the significance level (desired error rate, x-axis). For valid (well-calibrated) models the values should lie on the diagonal line. Deviations from this behaviour signals deviations from perfect calibration. We also include efficiency in the plot, calculated as the fraction of single-class predictions. These plots, from hereon called calibration and efficiency plots (CEPs), were used in this work to assess the model calibration and efficiency (see Fig. 2). As a measure of the level of calibration, we

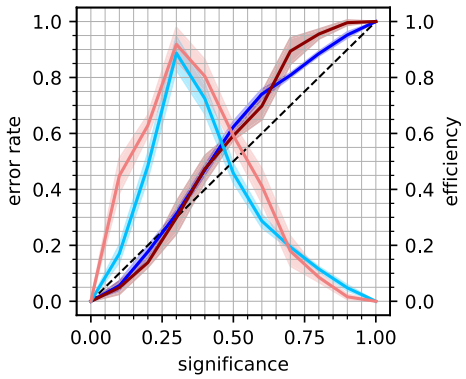


Fig. 2 Calibration and efficiency plot. The dark lines show the mean error rate for the active (dark red) and inactive (dark blue) compounds. For a well-calibrated model, the error rate ideally follows the dashed diagonal line. The light coloured lines illustrate the mean efficiencies expressed as ratio of single label sets for the active (light red) and inactive (light blue) compounds. The shaded areas indicate the respective standard deviations within the fivefold CV. Class 0: inactive compounds, class 1: active compounds

use the root-mean-square deviation (RMSD) between the specified significance and the observed error rate.

Model update strategies

In a setting where a model has been trained but new data on the same or a similar endpoint is made available, it is interesting to consider how the new data should be utilised in order to improve primarily the level of calibration but also the efficiency. We investigated two update strategies, see Fig. 3. The first strategy included updating the whole training set with new data followed by subsequent retraining of the model (see Fig. 3a). In the second strategy, the proper training set was kept and only the calibration set was exchanged with more recent data (see Fig. 3b).

Study design

In this work, six different CP experiments were explored as illustrated in Fig. 4 and Table 2. The first experiment consisted of a cross-validation (CV) using ACP on the Tox21Train dataset (*1-internal_CV*), the second comprised predictions with the CV-models from experiment 1 on the Tox21Score dataset (*2-pred_score*). In the third experiment, the influence of ACP on the calibration was assessed by training an SCP model on Tox21Train and predicting Tox21Score (*3-pred_score_SCP*). Finally, in the last three experiments, the model update strategies to improve the calibration were evaluated (see Fig. 3). Thus, in experiment 4 the training set was updated (*4-train_update*) and the model retrained, while in experiment 5 and 6 only the calibration set was updated (*5-cal_update* and *6-cal_update_2*).

The individual experiments were conceptualised in a way that the proper training sets were consistent across all experiments (where applicable). A fivefold CV was implemented, not only for internal validation (*1-internal_CV*), but conserved for all experiments. Hence, the selected data per CV loop of a fivefold CV were retained for all trained models (i.e. in the *1-internal_CV*, *4-train_update* and *3-pred_score_SCP* experiments). Specifically, the indices of the training compounds were saved, so that the same training sets could be used for the subsequent experiments. This ensures that the results from the different experiments can be directly compared. For the ACP model, 20 aggregated ICPs were used with 30% (of the training set) set aside as a calibration set and 70% as a

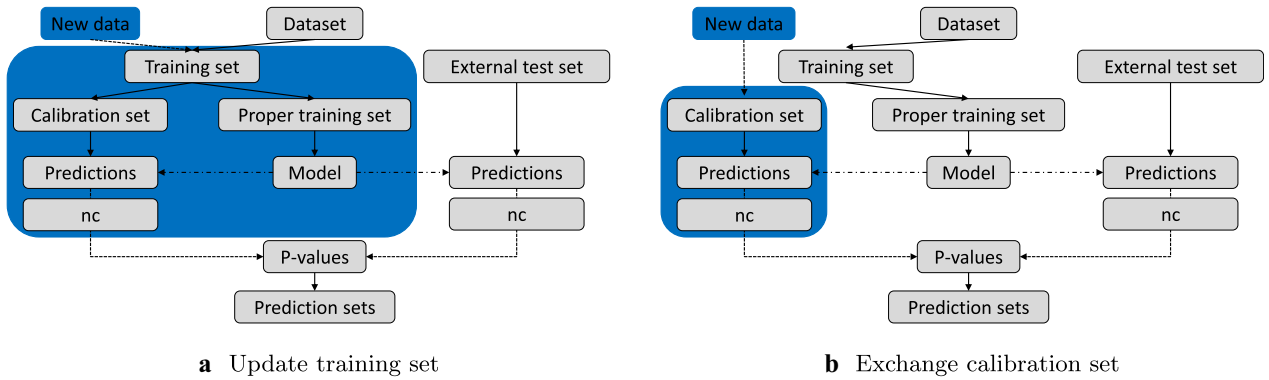


Fig. 3 Model update strategies analysed to improve calibration. **a** Update training set: The whole training set is updated with new data. This involves retraining a new model. **b** Exchange calibration set: Only the calibration set is updated with new data. Models can hereby be re-calibrated without training a new model

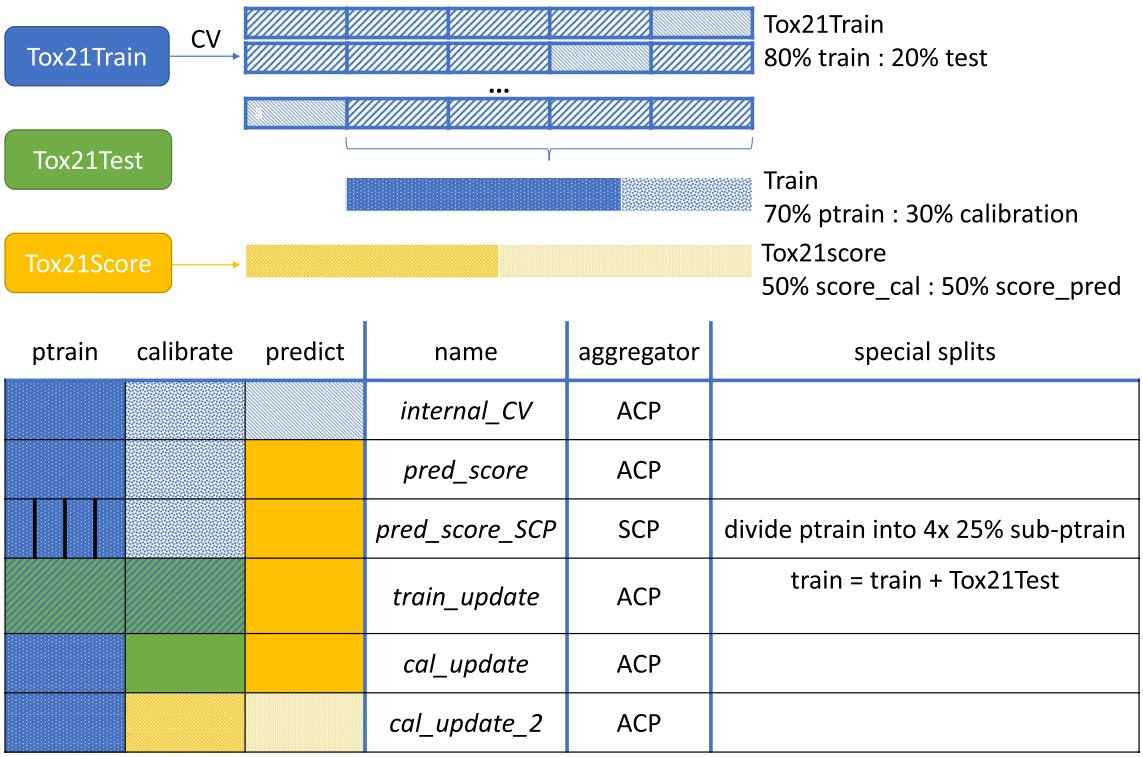


Fig. 4 Overview of the experiments discussed in this work. Top: Splitting of Tox21 data into (proper) training, calibration and test set. Bottom: Data for training, calibration, and prediction as well as aggregator used in the specific experiments.

Table 2 Overview of the experiments discussed in this work. Note that all splits were performed randomly stratified

Nr.	Name	Explanation
1	*internal_CV*	A fivefold CV, training one ACP per fold, is performed on the Tox21Train dataset and internally evaluated on the respective hold out data.
2	*pred_score*	Using the CV-models trained within the above described CV, the Tox21Score data are predicted.
3	*pred_score_SCP*	The same CV splits are applied as described above. The training set is then split into a fixed calibration set and four proportionate sub-proper training sets. For each of the four corresponding sub-proper training sets, an ML model is trained. Predictions are made for Tox21Score (and the calibration set compounds) with every model; the four nonconformity scores (ncs) are averaged before calculating the p-values.
4	*train_update*	The training set from the CV is combined with the Tox21Test set. This updated training set is then split into proper training and calibration set to train new ACP models for the CV set-up. Tox21Score data are predicted with the new models.
5	*cal_update*	The CV-models from experiment 1 are used, but the calibration is updated with the Tox21Test data to predict Tox21Score.
6	*cal_update_2*	The CV-models from experiment 1 are used, but the calibration is updated with 50% of Tox21Score data. The other 50% of Tox21Score are predicted. In every fold of the CV, Tox21Score is split in two equal subsets.

proper training set. For the *3-pred_score_SCP* experiment (using SCP, see Fig. 1a), the training set was split into a fixed 30% calibration set and the proper training set divided into four equally sized partitions. For the *4-train_update* experiment, the training set was first updated with the Tox21Test dataset and then split into calibration and proper training set using the above described ratios. For the two experiments updating the calibration set, the same trained CV-model from *1-internal_CV* was calibrated with only the Tox21Test dataset (*5-cal_update*) and in the last experiment (*6-cal_update_2)* replacing the calibration examples with 50% randomly stratified split Tox21Score data.

SVM models were trained using the Scikit-learn Python library [45] version 0.23.2 with an RFB kernel, $C = 50$, $\gamma = 0.002$) [37]. For conformal prediction, the nonconformist Python library [46] was used with margin error function, Mondrian condition [38, 39] version 2.1.0.

For ACP, p-values were aggregated by median (see [42]), for SCP the nonconformity scores were averaged before calculating p-values.

Code and data availability

A GitHub repository associated with this work is available at https://github.com/volkamerlab/cptox21_manuscript_SI. It contains the signature fingerprints for all pre-processed datasets as well as example code to demonstrate how the different ACP experiments were performed. The repository also provides the result files containing the respective measures for all experiments, from which the CEPs and boxplots can be generated. The SCP code is available from the original SCP repository by Gauraha et al. [44, 47].

Results and discussion

The aim of this study was to assess the level of calibration between the initial release of the Tox21Train data and the subsequently released Tox21Score data using conformal prediction (experiments 1–3). In follow-up experiments, we also investigated two model update strategies for incorporating the Tox21Test data (experiments 4–6). An overview of the error rates and efficiencies at significance level 0.2 for all experiments is provided in the Additional file 1: Table S2.

Experiment 1: Cross-validation on the Tox21Train datasets

Before applying a model to external data, it needs to be validated by ensuring that the model is internally well calibrated. Hence, in a first experiment (*1-internal_CV*), models were built in a fivefold CV scenario on the Tox21Train datasets. The models for the 12 Tox21 endpoints were internally valid with a mean error rate of 0.17 (± 0.01 std) at significance level 0.2, as well as a high mean efficiency of 0.77 (± 0.13 std).

The error rates and efficiencies over all significance levels (mean and std of the five CV folds per model) are illustrated in CEPs (Fig. 5a) for three example endpoints (namely SR ARE, NR_Aromatase and NR_AR; the

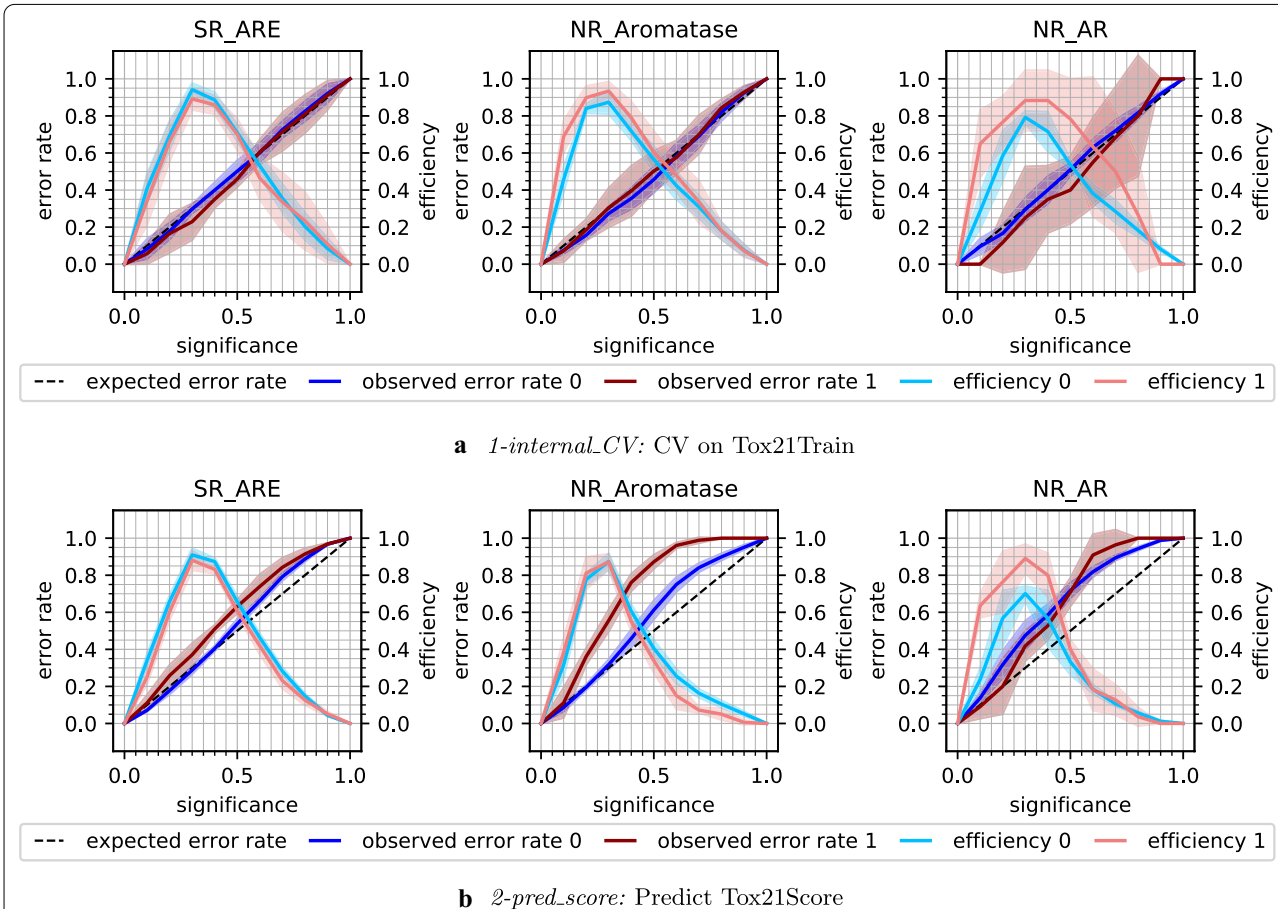


Fig. 5 CEPs for models trained on Tox21Train and subsequent internal cross-validation (**a**) and predictions on Tox21Score (**b**). CEPs for a selection of three example endpoints (SR_ARE, NR_Aromatase, NR_AR). Class 0: inactive compounds, class 1: active compounds. For a detailed explanation of all the components in the CEP, see Fig. 2

remaining CEPs are shown in the Additional file 1: Figure S1). While the models are overall well calibrated, i.e. the observed error rates follow the diagonal line in the CEPs, and the standard deviations between the individual runs are low, there are a few outliers. The high variance (see shaded areas in the CEPs) for the active compounds and the low efficiency for NR_AR reflect the observations in the Tox21 data challenge that NR_AR was one of the most difficult targets to model and has, with 387 active and 9201 inactive compounds, the lowest active compound rate after NR_PPARγ and NR_AR_LBD [31]. The well-calibrated models were ready to be applied to external data which stem from the same distribution as the training data.

Experiment 2: Model performance on the Tox21Score datasets

To investigate how well the CP models from the cross-validation perform on an external dataset, predictions were made for the Tox21Score data (*2-pred_score*). A mean error rate at significance level 0.2 of 0.31 ($\pm$ 0.12 std) was achieved. The efficiency dropped only slightly to 0.72 ($\pm$ 0.14 std). The deviations from the diagonal line in the CEPs (Fig. 5b, Additional file 1: Figure S2) for most of the endpoints indicate that the calibration of the models was poor when predicting Tox21Score.

Note that predictions were also made for the Tox21Test compounds (shown in the Additional file 1: Figure S3 only, referred to as *pred_test*). This set-up was similar to the intermediate setting in the Tox21 challenge, where predictions on Tox21Test were decisive for the leaderboard. The mean error rate at significance level 0.2 over all endpoints was higher than expected (0.26 $\pm$ 0.11 std). So, the models were not well-calibrated for predictions on Tox21Test. The mean efficiency was 0.70 ($\pm$ 0.15 std), i.e. similar to 2-*pred_score results*. The poor calibration for the predictions on both (external) datasets is an indication that the Tox21Score and the Tox21Test data might come from a different distribution than the Tox21Train data.

Experiment 3: Influence of aggregation method on the calibration

Reasons for poor calibration can be the difference between the distribution of two datasets, but also the data set size (discussed later) or the aggregation strategy for the conformal predictor (here ACP). From a theoretical perspective, the use of ACP can affect the calibration, as ACPs have not been proven to be always valid [42]. In ACP, the p-values from all ICPs are aggregated, which in theory could result in a non-uniform distribution. To rule out that the use of ACP is the (main) reason for the poor calibration, the recently developed SCP aggregation method was applied. In the SCP framework (see Fig. 1b), nonconformity scores are averaged before calculating the p-values, which are the basis for the calibration. This aggregation method has been shown to be theoretically valid [44].

Applying SCP improved the calibration on the Tox21Score dataset (*3-pred_score_SCP*), the mean error rate decreased to 0.27 ($\pm$ 0.12 std) and the mean efficiency at significance level 0.2 was 0.73 ($\pm$ 0.13 std). The error rates and efficiencies over all significance levels are shown in the CEPs in Fig. 6 for the SR_ARE, NR_Aromatase, and NR_AR endpoints and in the Additional file 1: Figure S4 for all 12 endpoints. It is especially noticeable that the calibration curves in the CEPs became less sigmoidal for many endpoints—such sigmoidal curves have typically been observed for ACPs [42, 44]. The sigmoidal shape is unfavourable from a theoretical perspective as it means

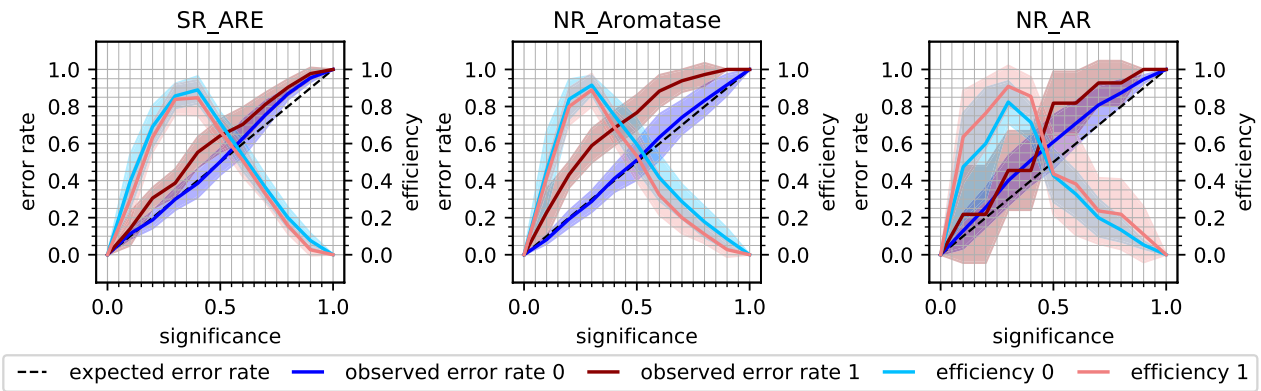


Fig. 6 Results from experiment *3-pred_score_scp:* SCP models were trained on Tox21Train and predictions made for Tox21Score. CEPs are shown for a selection of three example endpoints (SR_ARE, NR_Aromatase, NR_AR). Class 0: inactive compounds, class 1: active compounds. For a detailed explanation of all the components in the CEP, see Fig. 2

that the model is poorly calibrated at low and high significance levels, but may be less problematic in an application context since the error rate is typically over-conservative at lower (i.e. relevant) significance levels. One drawback of SCP is the fixed calibration set, which means that part of the training set information is never used for training. Together with the smaller proper training set partitions, this can lead to less efficient predictions. This can be seen in the relatively large standard deviations of the error and efficiency rates in the CEPs (Fig. 6 and Additional file 1: Figure S4). For this reason, and since ACP is commonly used in literature, which makes the outcomes more comparable with work by other scientists, ACP was used for the subsequent experiments.

Summarising the results from experiment 1–3, it was concluded that the Tox21Test and Tox21Score data may originate from slightly different distributions than the Tox21Train data. This could be explained by knowing that the three Tox21 datasets were created (screening of compounds) at different stages. For the Tox21Train set, the actual "Tox21 10K dataset" [31] was used, for which data had been available at the start of the challenge. The Tox21Test dataset is part of the LOPAC^{1280} (Library of Pharmacologically Active Compounds) dataset, which was used to validate the Tox21 assays [31, 48]. The Tox21Score data were separately provided by the EPA and only screened during the challenge [31]. So-called data or assay drifts typically occur over time or when moving towards a different chemical space [49].

Experiment 4: Effects on calibration by updating the training set

When the model is not well calibrated for the predictive task and newer data are available, one would intuitively combine these additional data (i.e. Tox21Test) with the previous training data (i.e. from *1-internal_CV*), train a new model, and use it to predict Tox21Score (*4-train_update*). Following this strategy, the mean error rate over the 12 endpoints dropped to 0.23 (± 0.06 std) compared to the predictions with the model built on the Tox21Train data (*2-pred_score*, 0.31 ± 0.12 std). The mean efficiency at significance level 0.2 (0.71 ± 0.15 std) was in a similar range as with the original training set (0.72 ± 0.14 std). Thus, the updating of the training set and retraining the model led to a small improvement in calibration (see CEPs in Additional file 1: Figure S5). One reason why we observed only a minor improvement of the calibration could be the sizes of the two datasets. The update set (254 ± 22 compounds) is small compared to the original training set (7647 ± 692 compounds) and has thus a lower influence on the new model. Furthermore, this strategy involves additional computational resources and the data of the previous model needs to be available for retraining.

Effects on calibration by updating the calibration set

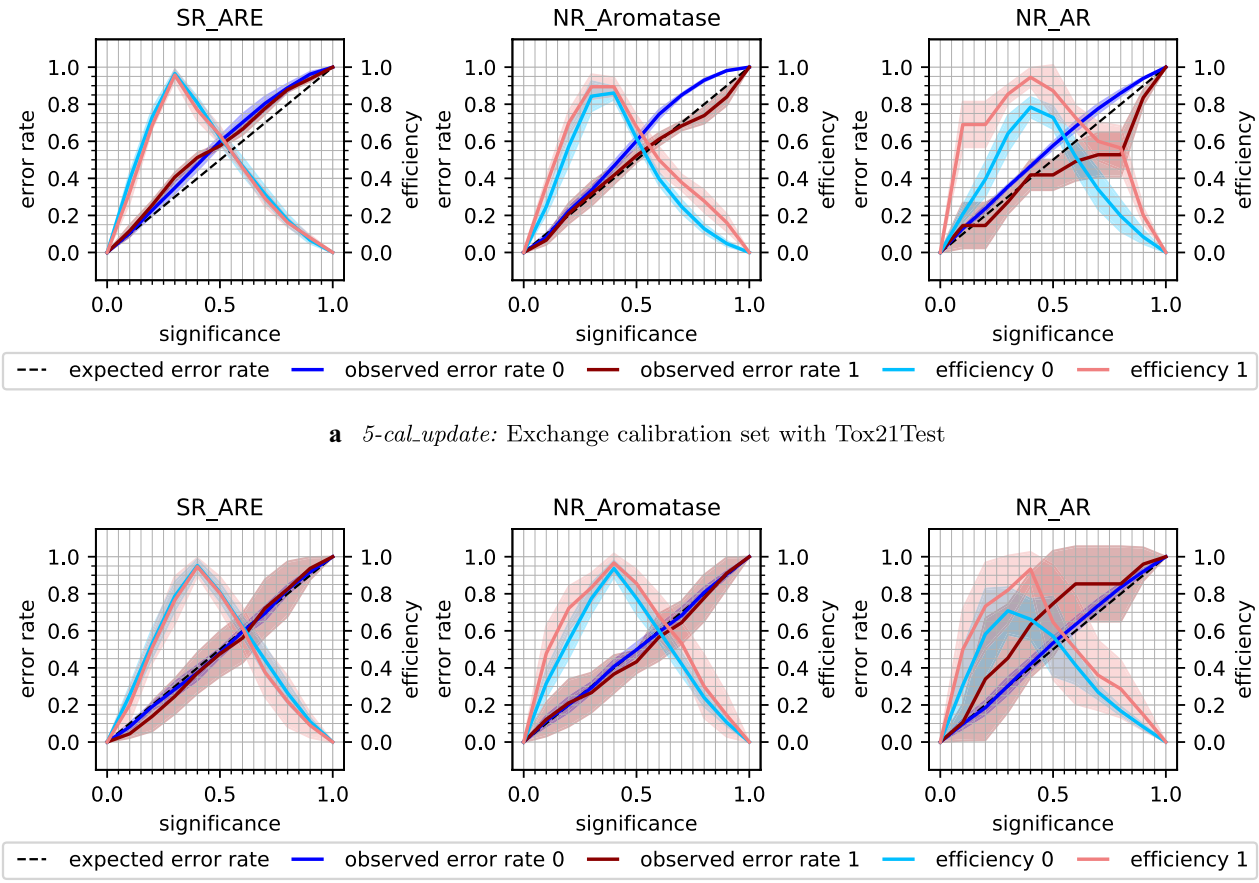
Experiment 5: Replace the calibration set with observations from Tox21Test

An alternative to updating the whole training set is to replace only the calibration set with the more recent data. This comes with the additional advantage that the calibration set can be renewed even if the training data are unavailable.

Updating the calibration set did result in a lower mean error rate of 0.21 (± 0.05 std) for the predictions on Tox21Score (*5-cal_update*). The mean efficiency at significance level 0.2 dropped to 0.51 (± 0.18 std). The loss in efficiency at low significance levels can be observed in the CEPs (Fig. 7a and Additional file 1: Figure S6), where the peak in efficiency is shifted towards higher significance levels. In the same CEPs, the improved calibration can be seen in the lower error rates. For six endpoints, when considering inactive compounds, or 11 endpoints, with regard to active compounds, even overconservative validity, i.e. a lower than expected error rate was achieved.

Experiment 6: Exchange the calibration set with half of Tox21Score

The chronological order of how the experimental data were produced is given by the Tox21 challenge organisers [31]. However, it is not clear if the compounds contained in Tox21Score (and Tox21Test) were really developed later than those in Tox21Train. For a 'perfect' calibration, it is required that the calibration and the test set stem from the same distribution. To simulate this, a second updating experiment, i.e. *6-cal_update_2*, was implemented. While still using the same proper training set as for the former experiments, the updated calibration set was created from Tox21Score. In every of the five (original) CV folds, 50% of Tox21Score was (randomly stratified) selected to constitute the calibration set while the other 50% of Tox21Score was used as test set. With this set-up, calibration and test set originate from the same distribution. This was also reflected in the mean error rate of 0.18 (± 0.01 std) at significance level 0.2, which was in a similar range as for the *1-internal_CV* with the original calibration set (0.17 ± 0.01 std). Similar to the previous updating experiment *5-cal_update*, the efficiency decreased to 0.50 (± 0.17 std) at significance level 0.2. Note that also the size of the calibration set was similar to the former *5-cal_update* experiment, as the Tox21Score set contains roughly twice as many compounds (551 ± 35) as Tox21Test (254 ± 22). On the other hand, by using half of Tox21Score for calibration, only the other half of the compounds was available for use as test set. This could lead to higher variations, e.g. in the error rate, especially for datasets with few test compounds. Such an example is shown for the NR_AR endpoint, for which



a *5-cal_update:* Exchange calibration set with Tox21Test

b *6-cal_update_2:* Exchange calibration set with half of Tox21Score

Fig. 7 Updating the calibration set with more recent data from Tox21Test (**a**) or Tox21Score (**b**). CEPs for a selection of three example endpoints (SR_ARE, NR_Aromatase, NR_AR). Class 0: inactive compounds, class 1: active compounds. For a detailed explanation of all the components in the CEP, see Fig. 2

Tox21Score only contains 11 actives. The standard deviation (shaded area) for the error rate and efficiency of the active compounds (red) increased compared to the *5-cal_update* experiment (Fig. 7). For the other two example endpoints in Fig. 7b (SR_ARE and NR_Aromatase), the calibration improved considerably. Summarising, the CEPs in the Additional file 1: Figure S7 illustrate how the calibration improved after exchanging the calibration set with data from the same distribution as the test set, but also how the efficiency dropped compared to the *4-train_update* strategy (Additional file 1: Figure S5). The decrease in efficiency in the 'cal_update' experiments is undesired but can be an acceptable trade-off in cases where validity could be restored. However, it has to be noted that the *6-cal_update_2* scenario is not often practically applicable as the updated calibration data needs to be available before making predictions.

Ultimately, updating the calibration set has no impact on the applicability domain of the underlying model. Improved calibration level and lower efficiency rather indicate that more compounds outside the applicability domain might be detected and classified as 'both' prediction sets. Thus, applying the *5-cal_update* over the *4-train_update* strategy is mainly promising in a situation as described in this work where the number of available new compounds is limited.

Quantification of the calibration for all experiments

The error rates (discussed above) depend on the desired significance level. In the calibration plot, the error rates are plotted over a range of significance levels. However, if the model will only be applied at a certain significance level, obtaining a good level of calibration at that significance level might be enough. But, if the calibration of the model is assessed from a theoretical perspective, all significance levels must be considered. This was illustrated for the individual experiments with the help of CEPs as

discussed above. To have a comparable metric, the root-mean-square deviation (RMSD) over all significance levels (step-width 0.1) was calculated.

Boxplots illustrating the RMSDs between observed and expected error rates over all endpoints are available in Fig. 8a for the active compounds and Fig. 8b for the inactive compounds, and show how the error rate deviations behave between the individual experiments. The mean RMSD values (overall, actives and inactives) for all experiments are provided in the Additional file 1: Table S3.

Clearly, the RMSD for the actives and inactives is low in the internal CV with Tox21Train (*1-internal_CV*) for most of the endpoints (mean overall RMSD: 0.022), while the deviations increased for the predictions on Tox21Score (*2-pred_score*, mean overall RMSD: 0.150). When using the SCP aggregation method (*3-pred_score_SCP*), the RMSD decreased for eight endpoints, albeit, only by a small amount (Fig. 8, mean overall RMSD: 0.121). Updating the training set (*4-train_update*, using ACP) led only to a small improvement of the mean RMSD of the active compounds (mean RMSD, actives: 0.135, Fig. 8a), while the improvement was more distinct for the inactive compounds (mean RMSD, inactives: 0.089, see Fig. 8b). When exchanging the calibration set with Tox21Test (*5-cal_update*), the RMSD decreased for 11 endpoints (except for SR_ARE, for which the calibration was already very good (overall RMSD SR_ARE, *1-pred_score*: 0.055) with the original calibration set). The mean overall RMSD (0.054) was, however, still not at the same level as for *1-internal_CV*. This can be attributed to overconservative validity, especially for the active compounds (see Additional file 1: Figure S6) which led to an increased RMSD for several endpoints. The overconservative validity almost disappeared when the calibration set was exchanged with data which are inherently exchangeable with the test set (*6-cal_update_2*). The mean RMSD (0.018) value of the inactive compounds is at a similar level as for the internal CV on Tox21Train (*1-internal_CV*) as shown in Fig. 8b. The RMSD values of the active compounds vary more between the different endpoints. This may be explained by the small number of active compounds available in the calibration and test sets for some endpoints. To summarise, the CP models trained on Tox21Train were internally well calibrated (*1-internal_CV*) but showed poorer calibration for the prediction of Tox21Score (*2-pred_score*). Applying SCP (*3-pred_score_SCP*) or updating the training set with Tox21Test (*4-train_update*) did not improve the calibration to the same extent as when exchanging the calibration set only (*5-cal_update, 6-cal_update_2*).

Impact of data size on the calibration

Importantly, the proofs on CP validity are made assuming an asymptotic number of test examples (i.e. requiring an infinite number of test examples) [24]. Hence, the poor calibration is not necessarily only due to exchangeability issues (or the use of ACP, for which there are no validity guarantees). The calibration could also be affected by the

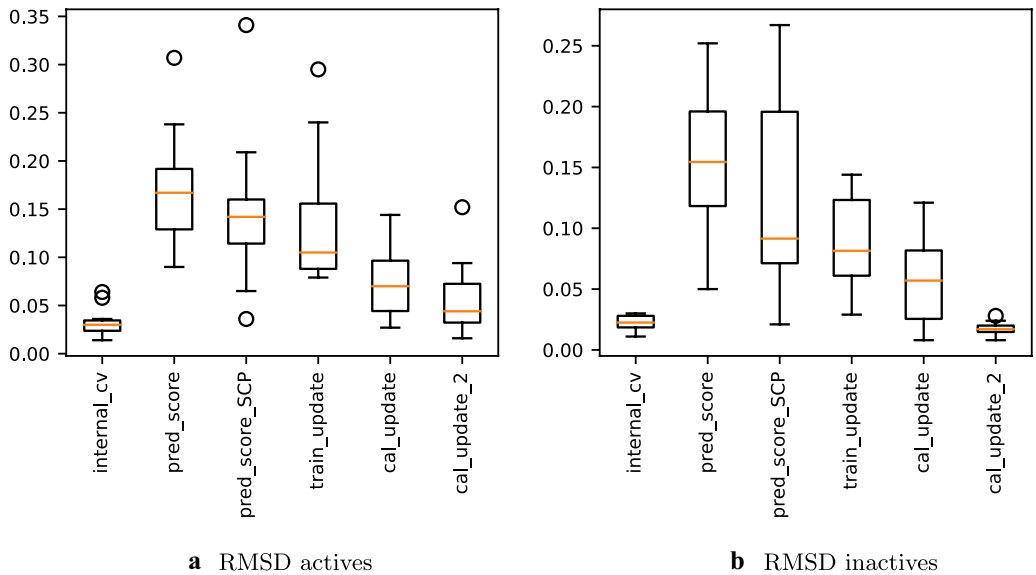


Fig. 8 Box plots for the root-mean-square deviation (RMSD) between the expected and observed error rates for all 12 Tox21 endpoints compared amongst the different experiments are shown. On the left results for the active compounds (**a**), on the right for the inactive compounds (**b**) are plotted. Note that the y-axis ranges differ

statistical variation due to finite test sets in all computational experiments. In the broadest sense, also the overconservative validity could be due to the finite number of test examples.

Looking at the outliers in the RMSD (Fig. 8), they mainly arise from endpoints NR_AR_LBD, NR_AR and SR_ATAD5, which are, besides NR_PPARγ, the endpoints with the smallest overall number of actives (in all three Tox21 datasets combined). For the NR_AR and NR_AR_LBD datasets, the predictive performance (both in validity and efficiency) is expected to be less good for the active compounds, as the number of available active compounds is very small (i.e. 3 and 4 in Tox21Test and 11 and 8 in Tox21Score, respectively). If we have only eight compounds in the calibration set, this means that only nine different p-values can be obtained for a new active compound. This low resolution obviously makes it impossible to obtain perfect calibration. Since it is difficult to define a minimum required number of actives, and since the resolution for the p-values of the inactive compounds is much higher, results for all endpoints were included in the evaluation. The calibration might generally improve if the experiments were repeated on larger and/or more balanced datasets.

Although, the composition of the three Tox21 datasets may not conform with all model assumptions, this may more closely resemble many real-life scenarios where data is generated at different time points and older data is often used to predict new outcomes. All the more, it is therefore important to have strategies to improve the calibration and thus the application of CP models on new data.

Conclusions

In this work, the potential of CP to diagnose data drifts in toxicity datasets was investigated on the Tox21 data. Deviations between observed and expected error rates was monitored using calibration plots and quantified using the RMSD from the expected calibration level. Poor calibration was observed for models trained on Tox21Train and predictions made on Tox21Score, indicating the presence of drifts between the two datasets. The distribution of the data may not be the only reason for error rate deviations in the calibration plot. In additional experiments using the newly introduced SCP framework, it was ruled out for 10 endpoints that the employed CP aggregation method (ACP) has a major impact. A second influencing factor on the calibration can be the small data set size. It was discussed that the calibration may be improved to some extent by having larger datasets, especially containing more active compounds, for model training, calibration and testing. Overall, it was concluded that the three Tox21 datasets likely do not originate from the same distribution and may be challenging for ML methods. Nonetheless, these datasets do reflect outcomes that may occur in experimental screening scenarios.

Two different model update strategies using the intermediate Tox21Test data were investigated with the aim to improve the poor calibration. The calibration of predictions on Tox21Score could be slightly enhanced by updating the training set with more recent data (Tox21Test) and retraining the models—the more natural behaviour if new data has been obtained. However, exchanging only the calibration set with newer data (Tox21Test) led to a slightly smaller error rate, albeit often with a reduction in efficiency. As an additional advantage of the *5-cal_update* strategy, retraining of the model is not required.

Abbreviations

ML: Machine learning; SVM: Support vector machine; EP: Endpoint; NR_AhR: Aryl hydrocarbon receptor; NR_AR: Androgen receptor, full length; NR_ER: Estrogen receptor, full length; NR_AR_LBD: Androgen receptor, ligand-binding domain; NR_ER_LBD: Estrogen receptor, ligand-binding domain; NR_PPAR_γ: Peroxisome proliferator-activated receptor gamma; SR_ARE: Nuclear factor (erythroid-derived 2)-like 2/antioxidant responsive element; SR_HSE: Heat shock factor response element; SR_MMP: Mitochondrial membrane potential; SR_ATAD5: ATAD5; NR_Aromatase: Aromatase; SR_p53: p53; CP: Conformal prediction; ICP: Inductive conformal predictor; ACP: Aggregated conformal prediction; SCP: Synergy conformal prediction; nc: Nonconformity score; RMSD: Root-mean-square deviation; Tox21Train: Training data for the Tox21 Data Challenge; Tox21Test: Test data for the Tox21 Data Challenge; Tox21Score: Final scoring data for the Tox21 Data Challenge; CEP: Calibration and efficiency plot; CV: Cross-validation; AD: Applicability domain.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13321-021-00511-5.

Additional file 1: Table S1. Number of compounds available per Tox21 dataset and endpoint before standardisation. **Table S2**. Mean ± standard deviation values over all twelve endpoints for observed error rate and efficiency at SL 0.2 for all experiments. **Figure S1.** *1-internal_CV*: ACP models were trained and calibrated on Tox21Train and internally validated. **Figure S2.** *2-pred_score*: ACP models were trained and calibrated on Tox21Train and predictions were made for Tox21Score. **Figure S3.** *pred_test*: ACP models were trained and calibrated on Tox21Train and predictions were made for Tox21Test. **Figure S4.** *3-pred_score_SCP*: SCP models were trained on Tox21Train and predictions made for Tox21Score. **Figure S5.** *4-train_update*: The training set from Tox21Train was updated with Tox21Test. **Figure S6.** *5-cal_update*: ACP models were trained on Tox21Train and calibrated on Tox21Test. **Figure S7.** *6-cal_update_2*: ACP models were trained on Tox21Train and calibrated on 50% of Tox21Score. **Table S3.** Mean RMSD values over all 12 endpoints, calculated for all compounds, as well as for active and inactive compounds, separately.

Acknowledgements

Computational resources were provided by SNIC-UPPMAX under project [SNIC 2019/8-149]. AM and AV would additionally like to thank the HPC Service of ZEDAT, Freie Universität Berlin, for computing time.

Authors' contributions

AM conducted the study under close supervision of UN, OS and AV, and in intense collaboration with all other project partners: FS, SAMS and NG. SAMS supported the signature generation with CPSign and NG provided support

with the SCP framework. AM, FS, UN, OS and AV designed the study. AM wrote the first draft of the manuscript and all authors contributed actively to the text. All authors revised the final manuscript. All authors read and approved the final manuscript.

Funding
Open Access funding enabled and organised by Projekt DEAL. AM acknowledges support from FUBright Mobility Allowances and the HaVo-Stiftung. AV thanks BMBF (grant 031A262C) for funding. ARUK UCL DDI is funded by Alzheimer's Research UK (Grant No. 560832). OS is supported by Swedish FORMAS (Grant 2018-00924) and Swedish Research Council (Grants 2020-03731 and 2020-01865). NG and OS were supported by Swedish Foundation for Strategic Research (Grant BD150008).

Availability of data and materials
A GitHub repository with supplementary information is available under https://github.com/volkamerlab/cptox21_manuscript_SI. In the repository, the signature fingerprints for all pre-processed datasets are provided, as well as the output evaluation files from all experiments, which contain the underlying data for the CEPs and boxplots. The repository, also contains example code to demonstrate how the different ACP experiments were performed. For the SCP code, the reader is referred to the original SCP repo by Gauraha [44, 47].

Declarations

Competing interests
OS declares ownership of Aros Bio AB, a company developing the CPSign software. SAMS declares contributions to the CPSign codebase.

Author details
[1]In Silico Toxicology and Structural Bioinformatics, Institute of Physiology, Charité Universitätsmedizin, Berlin, Germany. [2]Alzheimer's Research UK UCL Drug Discovery Institute, London WC1E 6BT, UK. [3]Department of Pharmaceutical Biosciences and Science for Life Laboratory, Uppsala University, 751 24 Uppsala, Sweden. [4]Division of Computational Science and Technology, KTH, 100 44 Stockholm, Sweden. [5]Dept. Computer and Systems Sciences, Stockholm University, Box 7003, 164 07 Kista, Sweden. [6]MTM Research Centre, School of Science and Technology, Örebro University, 70 182 Örebro, Sweden.

Received: 7 February 2021 Accepted: 10 April 2021
Published online: 29 April 2021

References

1. Yang H, Sun L, Li W, Liu G, Tang Y (2018) in silico prediction of chemical toxicity for drug design using machine learning methods and structural alerts. Front Chem 6:30. https://doi.org/10.3389/fchem.2018.00030
2. Klambauer G, Hochreiter S, Rarey M (2019) Machine learning in drug discovery. J Chem Inf Model 59(3):945–946. https://doi.org/10.1021/acs.jcim.9b00136
3. Vo AH, Van Vleet TR, Gupta RR, Liguori MJ, Rao MS (2020) An overview of machine learning and big data for drug toxicity evaluation. Chem Res Toxicol 33(1):20–37. https://doi.org/10.1021/acs.chemrestox.9b00227
4. Steger-Hartmann T, Boyer S (2020) Computer-based prediction models in regulatory toxicology. Regulatory toxicology. Springer, Berlin, pp 123–131. https://doi.org/10.1007/978-3-642-35374-1_36
5. Richard AM, Judson RS, Houck KA, Grulke CM, Volarath P, Thillainadarajah I, Yang C, Rathman J, Martin MT, Wambaugh JF, Knudsen TB, Kancherla J, Mansouri K, Patlewicz G, Williams AJ, Little SB, Crofton KM, Thomas RS (2016) ToxCast chemical landscape: paving the road to 21st century toxicology. Chem Res Toxicol 29(8):1225–1251. https://doi.org/10.1021/acs.chemrestox.6b00135
6. Huang R, Sakamuru S, Martin MT, Reif DM, Judson RS, Houck KA, Casey W, Hsieh JH, Shockley KR, Ceger P, Fostel J, Witt KL, Tong W, Rotroff DM, Zhao T, Shinn P, Simeonov A, Dix DJ, Austin CP, Kavlock RJ, Tice RR, Xia M (2014) Profiling of the Tox21 10K compound library for agonists and antagonists of the estrogen receptor alpha signaling pathway. Sci Rep 4:1–9. https://doi.org/10.1038/srep05664
7. Richard AM, Huang R, Waidyanatha S, Shinn P, Collins B, Thillainadarajah I, Grulke CM, Williams AJ, Lougee RR, Judson RS, Houck KA, Shobair M, Yang C, Rathman JF, Yasgar A, Fitzpatrick SC, Simeonov A, Thomas RS, Crofton KM, Paules RS, Bucher JR, Austin CP, Kavlock RJ, Tice RR (2020) The Tox21 10K compound library: collaborative chemistry advancing toxicology. Chem Res Toxicol. https://doi.org/10.1021/acs.chemrestox.0c00264
8. Cases M, Briggs K, Steger-Hartmann T, Pognan F, Marc P, Kleinöder T, Schwab CH, Pastor M, Wichard J, Sanz F (2014) The eTOX data-sharing project to advance in Silico drug-induced toxicity prediction. Int J Mol Sci 15(11):21,136-21,154. https://doi.org/10.3390/ijms151121136
9. Pastor M, Quintana J, Sanz F (2018) Development of an infrastructure for the prediction of biological endpoints in industrial environments. Lessons learned at the eTOX project. Front Pharmacol 9:1–8. https://doi.org/10.3389/fphar.2018.01147
10. Golbraikh A, Shen M, Xiao Z, Xiao YD, Lee KH, Tropsha A (2003) Rational selection of training and test sets for the development of validated QSAR models. J Comput Aided Mol Des 2003(17):241–53. https://doi.org/10.1023/A:1025386326946
11. Martin TM, Harten P, Young DM, Muratov EN, Golbraikh A, Zhu H, Tropsha A (2012) Does rational selection of training and test sets improve the outcome of QSAR modeling? J Chem Inf Model 52(10):2570–2578. https://doi.org/10.1021/ci300338w
12. Leonard JT, Roy K (2006) On selection of training and test sets for the development of predictive QSAR models. QSAR Combinat Sci 25(3):235–251. https://doi.org/10.1002/qsar.200510161
13. Andrada MF, Vega-Hissi EG, Estrada MR, Garro Martinez JC (2017) Impact assessment of the rational selection of training and test sets on the predictive ability of QSAR models. SAR QSAR Environ Res 28(12):1011–1023. https://doi.org/10.1080/1062936X.2017.1397056
14. Klimenko K, Rosenberg SA, Dybdahl M, Wedebye EB, Nikolov NG (2019) QSAR modelling of a large imbalanced aryl hydrocarbon activation dataset by rational and random sampling and screening of 80,086 REACH pre-registered and/or registered substances. PLoS ONE 14(3):1–21. https://doi.org/10.1371/journal.pone.0213848
15. Mathea M, Klingspohn W, Baumann K (2016) Chemoinformatic classification methods and their applicability domain. Mol Inf 35(5):160–180. https://doi.org/10.1002/minf.201501019
16. Hanser T, Barber C, Marchaland JF, Werner S (2016) Applicability domain: towards a more formal definition. SAR QSAR Environ Res 27(11):865–881. https://doi.org/10.1080/1062936X.2016.1250229
17. Bosnić Z, Kononenko I (2008) Comparison of approaches for estimating reliability of individual regression predictions. Data Knowl Eng 67(3):504–516. https://doi.org/10.1016/j.datak.2008.08.001
18. Aniceto N, Freitas AA, Bender A, Ghafourian T (2016) A novel applicability domain technique for mapping predictive reliability across the chemical space of a QSAR: Reliability-density neighbourhood. J Cheminf 8(1):1–20. https://doi.org/10.1186/s13321-016-0182-y
19. Dragos H, Gilles M, Alexandre V (2009) Predicting the predictability: a unified approach to the applicability domain problem of QSAR models. J Chem Inf Model 49(7):1762–1776. https://doi.org/10.1021/ci9000579
20. Sheridan RP (2015) The relative importance of domain applicability metrics for estimating prediction errors in QSAR varies with training set diversity. J Chem Inf Model 55(6):1098–1107. https://doi.org/10.1021/acs.jcim.5b00110
21. Alves VM, Muratov EN, Zakharov A, Muratov NN, Andrade CH, Tropsha A (2018) Chemical toxicity prediction for major classes of industrial chemicals: Is it possible to develop universal models covering cosmetics, drugs, and pesticides? Food Chem Toxicol 112:526–534. https://doi.org/10.1016/j.fct.2017.04.008
22. Cortés-Ciriano I, Škuta C, Bender A, Svozil D (2020) QSAR-derived affinity fingerprints (part 2): modeling performance for potency prediction. J Cheminf 12(1):1–17. https://doi.org/10.1186/s13321-020-00444-5
23. Vovk V (2013) Conditional validity of inductive conformal predictors. Mach Learn 92(2–3):349–376. https://doi.org/10.1007/s10994-013-5355-6
24. Vovk V, Gammerman A, Shafer G (2005) Algorithmic learning in a random world. Springer Science & Business Media, Berlin
25. Fedorova V, Gammerman A, Nouretdinov I, Vovk V (2012) Plug-in martingales for testing exchangeability on-line. In: Proceedings of the 29th international conference on machine learning, ICML 2012 2. pp 1639–1646

26. Morger A, Mathea M, Achenbach JH, Wolf A, Buesen R, Schleifer KJ, Landsiedel R, Volkamer A (2020) KnowTox: pipeline and case study for confident prediction of potential toxic effects of compounds in early phases of development. J Cheminf 12(1):1–17. https://doi.org/10.1186/s13321-020-00422-x
27. Huang Ruili, Xia Menghang (2017) Editorial: Tox21 challenge to build predictive models of nuclear receptor and stress response pathways as mediated by exposure to environmental toxicants and drugs. Front Environ Sci 5(3):5. https://doi.org/10.1038/ncomms
28. NIH—National Center for Advancing Translational Sciences (2014) Tox21 data challenge. https://tripod.nih.gov/tox21/challenge/data.jsp
29. Atkinson FCGEE (2014) Standardiser. https://github.com/flatkinson/standardiser
30. Heller SR, McNaught A, Pletnev I, Stein S, Tchekhovskoi D (2015) InChI, the IUPAC international chemical identifier. J Cheminf. https://doi.org/10.1186/s13321-015-0068-4
31. Huang R, Xia M, Nguyen DT, Zhao T, Sakamuru S, Zhao J, Shahane SA, Rossoshek A, Simeonov A (2016) Tox21 challenge to build predictive models of nuclear receptor and stress response pathways as mediated by exposure to environmental chemicals and drugs. Front Environ Sci 3:1–9. https://doi.org/10.3389/fenvs.2015.00085
32. Faulon JL, Visco DP, Pophale RS (2003) The signature molecular descriptor. 1. Using extended valence sequences in QSAR and QSPR studies. J Chem Inf Comput Sci 43(3):707–720. https://doi.org/10.1021/ci020345w
33. Faulon JL, Collins MJ, Carr RD (2004) The signature molecular descriptor. 4. Canonizing molecules using extended valence sequences. J Chem Inf Comput Sci 44(2):427–436. https://doi.org/10.1021/ci0341823
34. Aros Bio (2020) CPSign. https://arosbio.com/cpsign/
35. Kensert A, Alvarsson J, Norinder U, Spjuth O (2018) Evaluating parameters for ligand-based modeling with random forest on sparse data sets. J Cheminf 10(1):1–10. https://doi.org/10.1186/s13321-018-0304-9
36. Alvarsson J, Eklund M, Engkvist O, Spjuth O, Carlsson L, Wikberg JE, Noeske T (2014a) Ligand-based target prediction with signature fingerprints. J Chem Inf Model 54(10):2647–2653. https://doi.org/10.1021/ci500361u
37. Alvarsson J, Eklund M, Andersson C, Carlsson L, Spjuth O, Wikberg JE (2014b) Benchmarking study of parameter variation when using signature fingerprints together with support vector machines. J Chem Inf Model 54(11):3211–3217. https://doi.org/10.1021/ci500344v
38. Sun J, Carlsson L, Ahlberg E, Norinder U, Engkvist O, Chen H (2017) Applying Mondrian cross-conformal prediction to estimate prediction confidence on large imbalanced bioactivity data sets. J Chem Inf Model 57(7):1591–1598. https://doi.org/10.1021/acs.jcim.7b00159
39. Toccaceli P, Gammerman A (2019) Combination of inductive Mondrian conformal predictors. Mach Learn 108(3):489–510. https://doi.org/10.1007/s10994-018-5754-9
40. Alvarsson J, Arvidsson McShane S, Norinder U, Spjuth O (2021) Predicting with confidence: using conformal prediction in drug discovery. J Pharm Sci 110(1):42–49. https://doi.org/10.1016/j.xphs.2020.09.055
41. Norinder U, Carlsson L, Boyer S, Eklund M (2014) Introducing conformal prediction in predictive modeling. A transparent and flexible alternative to applicability domain determination. J Chem Inf Model 54(6):1596–1603. https://doi.org/10.1021/ci5001168
42. Linusson H, Norinder U, Boström H, Johansson U, Löfström T (2017) On the calibration of aggregated conformal predictors. In: Proceedings of the sixth workshop on conformal and probabilistic prediction and applications 60. pp 154–173
43. Carlsson L, Eklund M, Norinder U, Carlsson L, Eklund M, Norinder U, Conformal A, Lazaros P, Maglogiannis I, Papadopoulos H, Sioutas S, Ifip CM (2014) Aggregated conformal prediction. In: IFIP advances in information and communication technology. pp 231–240
44. Gauraha N, Spjuth O (2018) Synergy conformal prediction. DiVA preprint 360504. http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-360504
45. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E (2011) Scikit-learn: machine learning in python. J Mach Learn Res 12:2825–2830
46. Linusson H (2015) Nonconformist. http://donlnz.github.io/nonconformist/
47. Gauraha N (2018) Synergy conformal prediction code. https://github.com/niha21/scp_code
48. Attene-Ramos MS, Miller N, Huang R, Michael S, Itkin M, Kavlock RJ, Austin CP, Shinn P, Simeonov A, Tice RR, Xia M (2013) The Tox21 robotic platform for the assessment of environmental chemicals—from vision to reality. Drug Discov Today 18(15–16):716–723. https://doi.org/10.1016/j.drudis.2013.05.015.
49. Azadeh M, Sondag P, Wang Y, Raines M, Sailstad J (2019) Quality controls in ligand binding assays: recommendations and best practices for preparation, qualification, maintenance of lot to lot consistency, and prevention of assay drift. AAPS J. https://doi.org/10.1208/s12248-019-0354-6

Publisher's Note

Additional file

Assessing the Calibration in Toxicological in Vitro Models with Conformal Prediction

Andrea Morger, Fredrik Svensson, Staffan Arvidsson McShane, Niharika Gauraha, Ulf Norinder, Ola Spjuth and Andrea Volkamer

Table S1: Number of compounds available per Tox21 dataset and endpoint before standardisation.

	Tox21Train		Tox21Test		Tox21Score	
endpoint	actives	inactives	actives	inactives	actives	inactives
NR_AhR	950	7214	30	241	73	537
NR_AR	380	8977	3	288	12	574
NR_AR_LBD	303	8291	4	248	8	574
NR_Aromatase	360	6861	18	196	39	489
NR_ER	937	6756	27	237	51	465
NR_ER_LBD	446	8302	10	276	20	580
NR_PPAR_γ	222	7957	15	251	31	574
SR_ARE	1097	6067	47	186	93	462
SR_ATAD5	338	8748	25	246	38	584
SR_HSE	428	7718	10	256	22	588
SR_MMP	1142	6174	38	199	60	483
SR_p53	537	8092	28	240	41	575

Table S2: Mean ± standard deviation values over all twelve endpoints for observed error rate and efficiency at SL 0.2 for all experiments.

nr.	name	error rate at SL 0.2	efficiency at SL 0.2
1	*internal_CV*	0.17 ± 0.01	0.77 ± 0.13
2	*pred_score*	0.31 ± 0.12	0.72 ± 0.14
–	*pred_test**	0.26 ± 0.11	0.70 ± 0.15
3	*pred_score_SCP*	0.27 ± 0.12	0.73 ± 0.13
4	*train_update*	0.23 ±0.06	0.71 ±0.15
5	*cal_update*	0.21 ± 0.05	0.51 ± 0.18
6	*cal_update_2*	0.18 ± 0.01	0.50 ± 0.17

**pred_test*: the CV-models from *internal_cv* were used to make predictions on Tox21Test.

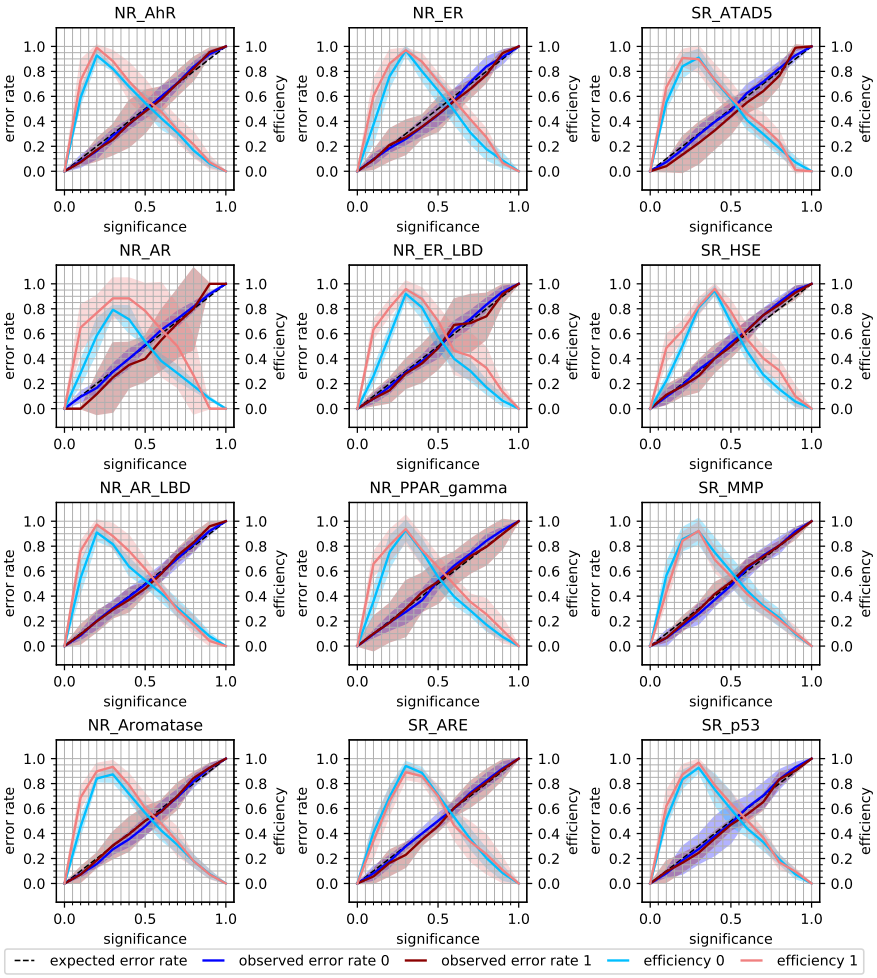


Figure S1: *1-internal_CV:* ACP models were trained and calibrated on Tox21Train and internally validated. CEPs for all twelve Tox21 endpoints are shown. Class 0: inactive compounds, class 1: active compounds. For a detailed explanation of all the components in the CEP, see Figure 2.

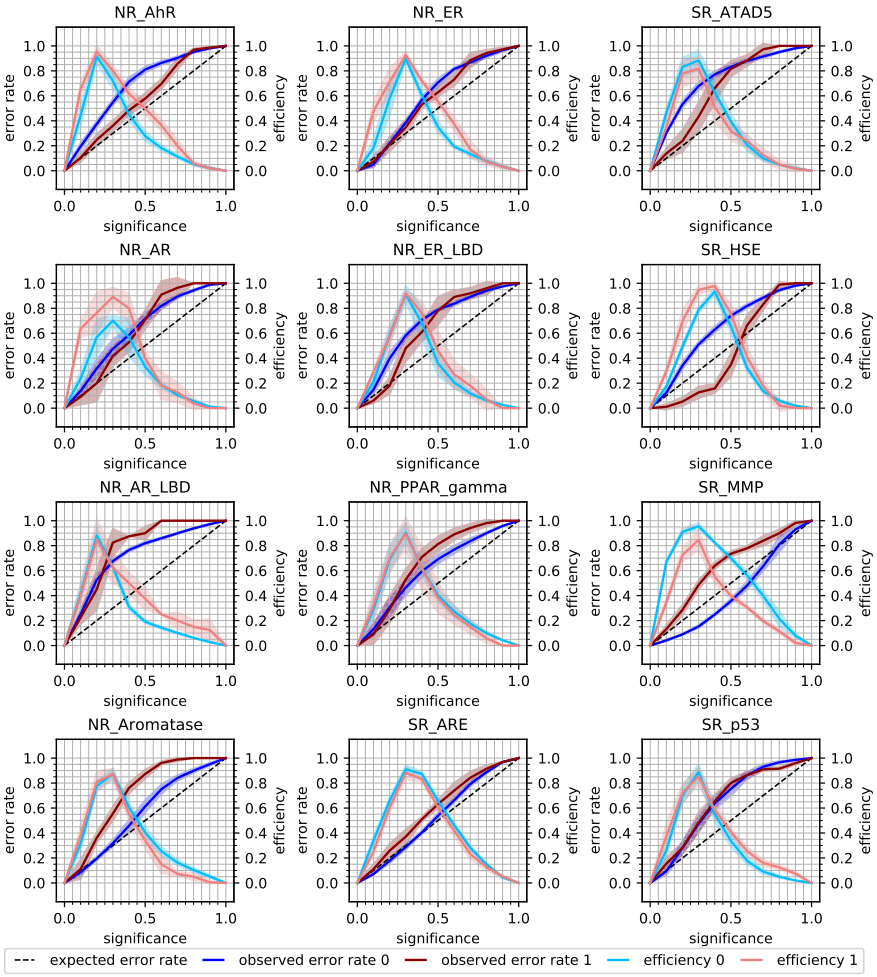


Figure S2: *2-pred_score:* ACP models were trained and calibrated on Tox21Train and predictions were made for Tox21Score. CEPs for all twelve Tox21 endpoints are shown. Class 0: inactive compounds, class 1: active compounds. For a detailed explanation of all the components in the CEP, see Figure 2.

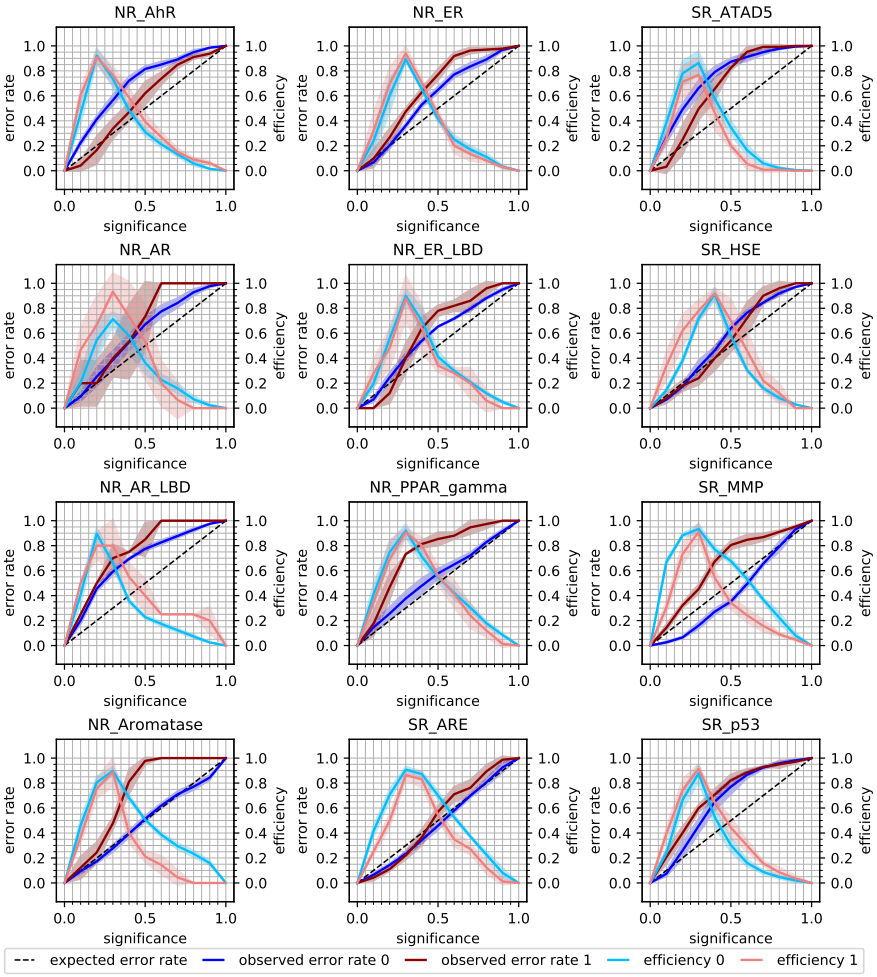


Figure S3: *pred_test:* ACP models were trained and calibrated on Tox21Train and predictions were made for Tox21Test. CEPs for all twelve Tox21 endpoints are shown. Class 0: inactive compounds, class 1: active compounds. For a detailed explanation of all the components in the CEP, see Figure 2.

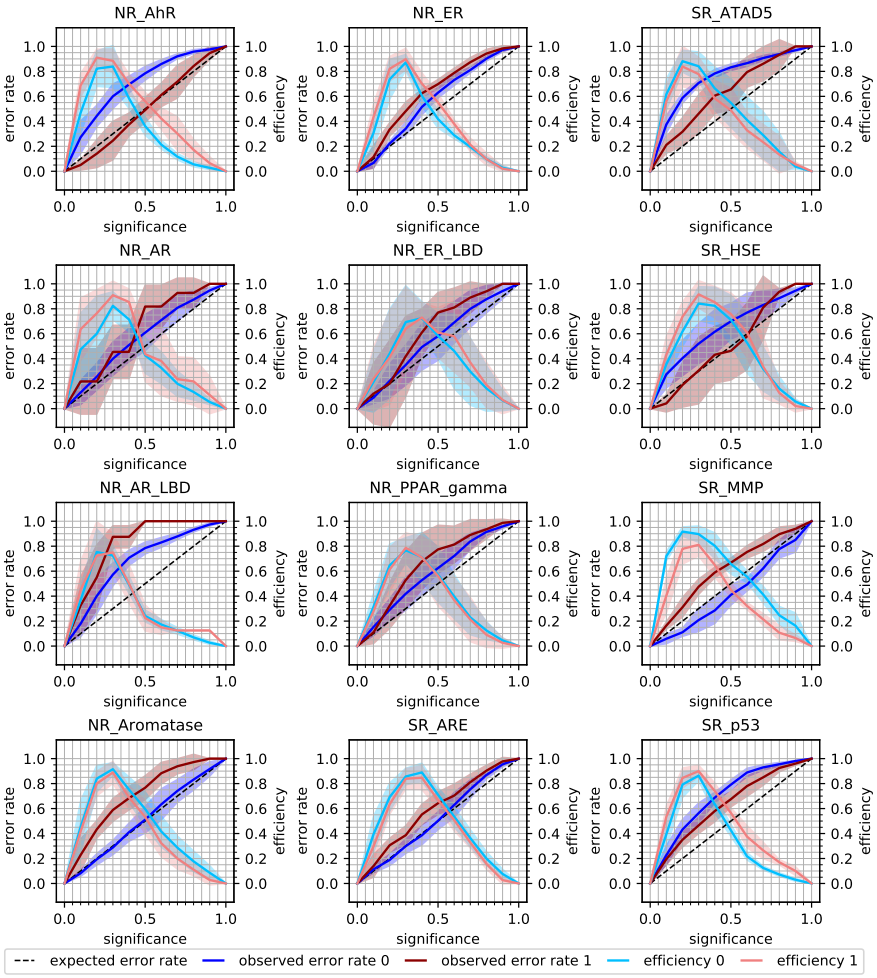


Figure S4: *3-pred_score_SCP:* SCP models were trained on Tox21Train and predictions made for Tox21Score. CEPs for all twelve Tox21 endpoints are shown. Class 0: inactive compounds, class 1: active compounds. For a detailed explanation of all the components in the CEP, see Figure 2.

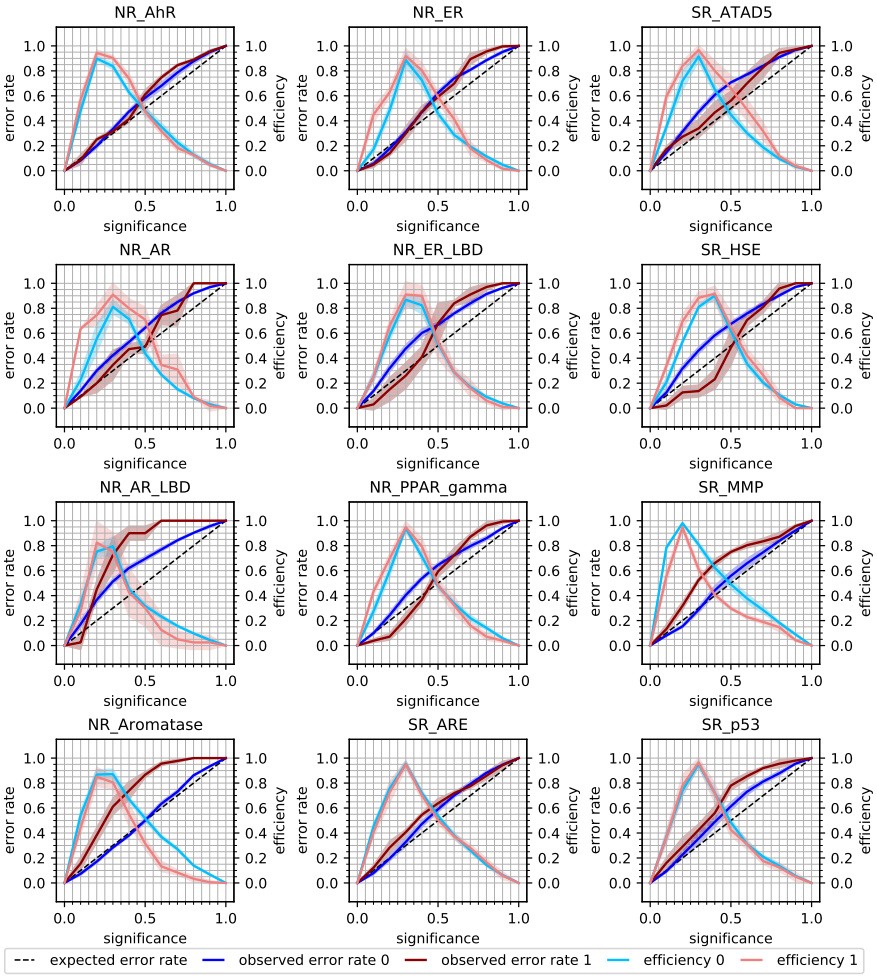


Figure S5: *4-train_update:* The training set from Tox21Train was updated with Tox21Test. An ACP model was retrained and predictions were made for Tox21Score. CEPs for all twelve Tox21 endpoints are shown. Class 0: inactive compounds, class 1: active compounds. For a detailed explanation of all the components in the CEP, see Figure 2.

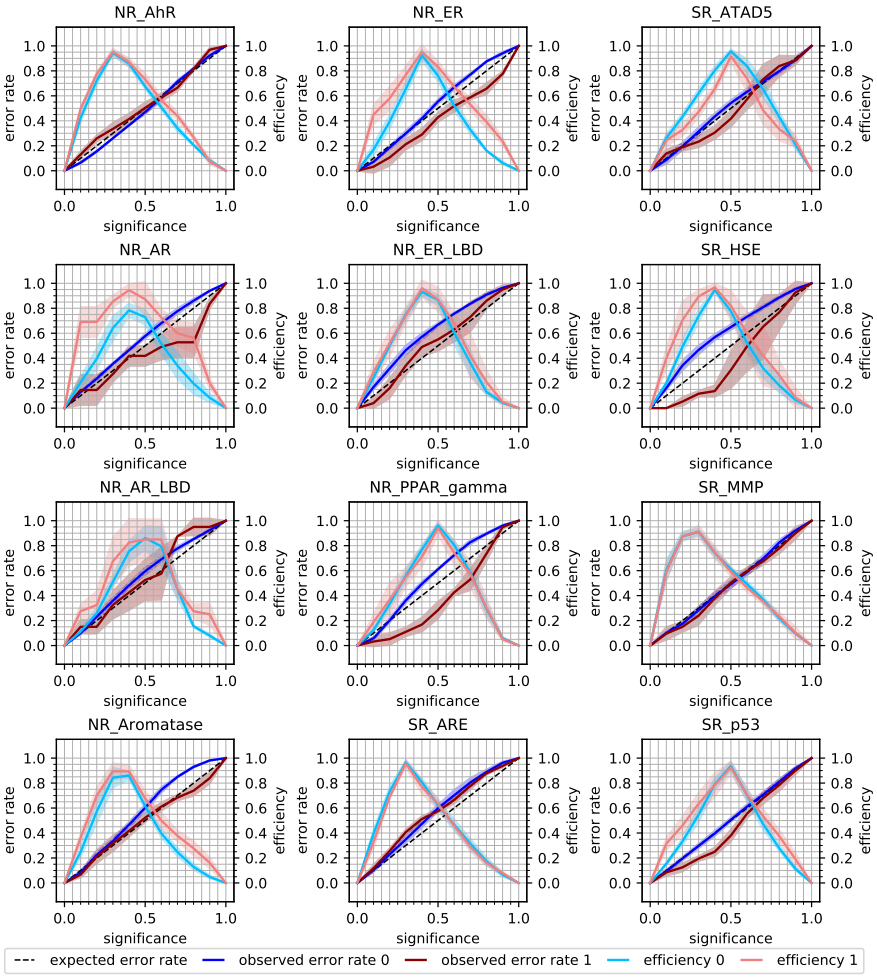


Figure S6: *5-cal_update:* ACP models were trained on Tox21Train and calibrated on Tox21Test. Predictions were made for Tox21Score. CEPs for all twelve Tox21 endpoints are shown. Class 0: inactive compounds, class 1: active compounds. For a detailed explanation of all the components in the CEP, see Figure 2.

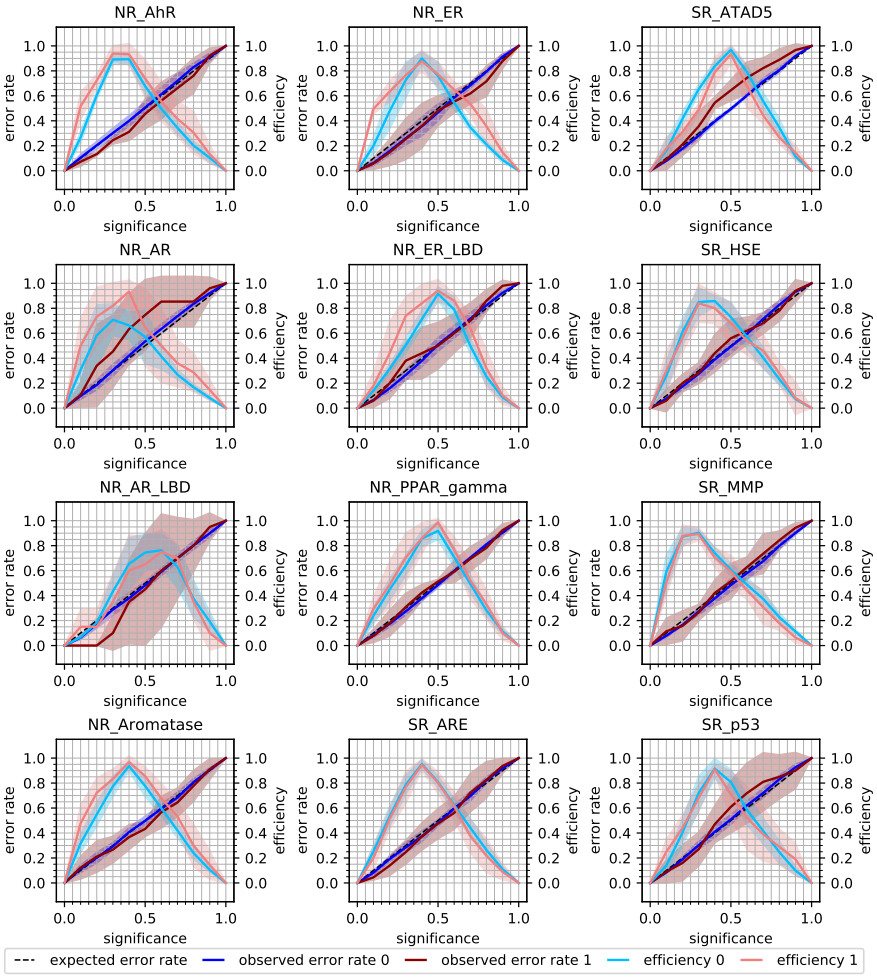


Figure S7: *6-cal_update_2:* ACP models were trained on Tox21Train and calibrated on 50% of Tox21Score. Predictions were made for the other 50% of Tox21Score. CEPs for all twelve Tox21 endpoints are shown. Class 0: inactive compounds, class 1: active compounds. For a detailed explanation of all the components in the CEP, see Figure 2.

Table S3: Mean RMSD values over all 12 endpoints, calculated for all compounds, as well as for active and inactive compounds, separately.

nr.	name	all	actives	inactives
1	*internal_CV*	0.022	0.032	0.022
2	*pred_score*	0.150	0.167	0.154
–	*pred_test**	0.116	0.180	0.119
3	*pred_score_SCP*	0.121	0.147	0.124
4	*train_update*	0.090	0.135	0.089
5	*cal_update*	0.054	0.073	0.058
6	*cal_update_2*	0.018	0.057	0.018

**pred_test*: the CV-models from *internal_cv* were used to make predictions on Tox21Test.

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

The previous study (see Section 4.4) could show the success of recalibrating conformal prediction models with the example of the Tox21 datasets. It would be useful, if this strategy could also be applied to mitigate temporal data drifts, or in the case of training and test data originating from different sources. Therefore, the recalibration strategy will be further explored in two real-life scenarios. First, mitigating effects of temporal data drifts will be investigated with the example of twelve toxicity-related ChEMBL datasets. ChEMBL is one of only a few bioactivity databases, which contains temporal information — it allows splitting the data based on the publication date. Second, conformal prediction models will be trained on publicly-available liver toxicity and genotoxicity data and the recalibration strategy will be explored to calibrate the models for application on proprietary data from industry.

Contribution:

Co-first author
Conceptual design (70%)
Computational experiments (80%)
Visualization (70%)
Manuscript preparation (50%)

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

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

[1] *In Silico* Toxicology and Structural Bioinformatics, Institute of Physiology, Charité Universitätsmedizin, Berlin, 10117, Germany
[2] BASF SE, Ludwigshafen, 67056, 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, Örebro, 701 82, Sweden
[7] Alzheimer's Research UK UCL Drug Discovery Institute, London, WC1E 6BT, United Kingdom
[*] andrea.volkamer@charite.de, miriam.mathea@basf.com
[+] these authors contributed equally to this work

ABSTRACT

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 re-train the model has proven to be a useful approach to restore the validity of most models.

1 Introduction

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 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.

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' ($\{\emptyset\}$), '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 ($\{\emptyset\}$) 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 applicability domain 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 cross-validation (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 machine learning, 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].

2 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.

2.1 Data assembly

2.1.1 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 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 2.1.2) and the datasets temporally split (see Section 2.1.4). 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.

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

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

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 micro nucleus test (MNT)) 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, or in (non-GLP) screening assays[29]. 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).

2.1.2 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.

2.1.3 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 2.2.1) 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.

2.1.4 Data splitting

After standardising the compounds (see Section 2.1.2), 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.

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.

target (ID)	training set			update1 set			update2 set			holdout set		
	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

*thresh: Data points published (ChEMBL) or measured (micro nucleus test, liver toxicity) until this year threshold are included in the corresponding subset.

2.2 Conformal Prediction

2.2.1 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 scores (nc scores) by using a nonconformity function. Here, the inverse probability error function, which is typical 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 score 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 $\{\emptyset\}$, $\{0\}$, $\{1\}$, $\{0,1\}$. 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 (here $n = 20$) times. 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].

2.2.2 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 ($\{\emptyset\}$) 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.

2.2.3 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 2.1.4. 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 used. 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].

2.3 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.

2.4 Availability of data and code

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.

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.

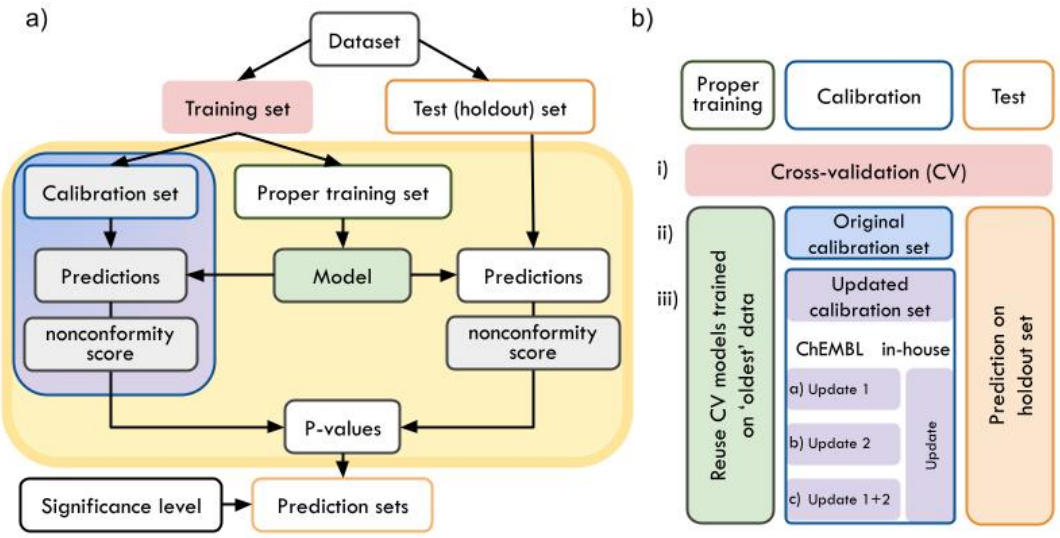


Figure 1. a) Framework of an inductive conformal predictor (ICP). An ML model is fitted on the compounds of the proper training set to make predictions for the calibration and test (holdout) set instances. The predictions are transformed into nonconformity scores (nc scores). By comparing the outcome 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) update calibration sets to investigate temporal data drifts and drifts between data from different origin, i.e., ChEMBL and inhouse data.

3 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, conformal prediction (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 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 performances have 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.

3.1 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 2 and Table 2).

Experiments i and ii: CV and predictions using original calibration set Five-fold 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 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.

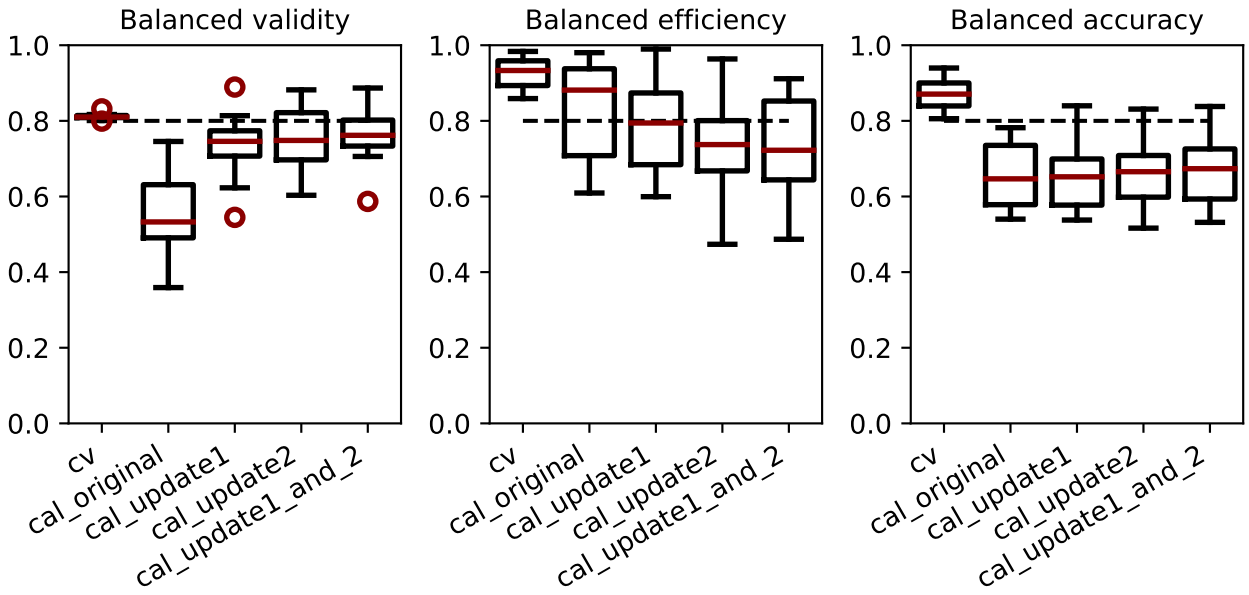


Figure 2. Time split evaluation (balanced validity, balanced efficiency, balanced accuracy) of CV experiments and predictions for the holdout set using the original (cal_original), update1 (cal_update1), update2 (cal_update2) and combined update1_and_2 (cal_update1_and_2) calibration set for twelve ChEMBL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

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 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 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 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 test set may fall outside the applicability domain 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 conformal prediction 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.

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

		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

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 cross-validation (where by design compounds belonging to the same cluster are always in the same splits) should present a more challenging task than a temporal time-split (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 cross-validation 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 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.

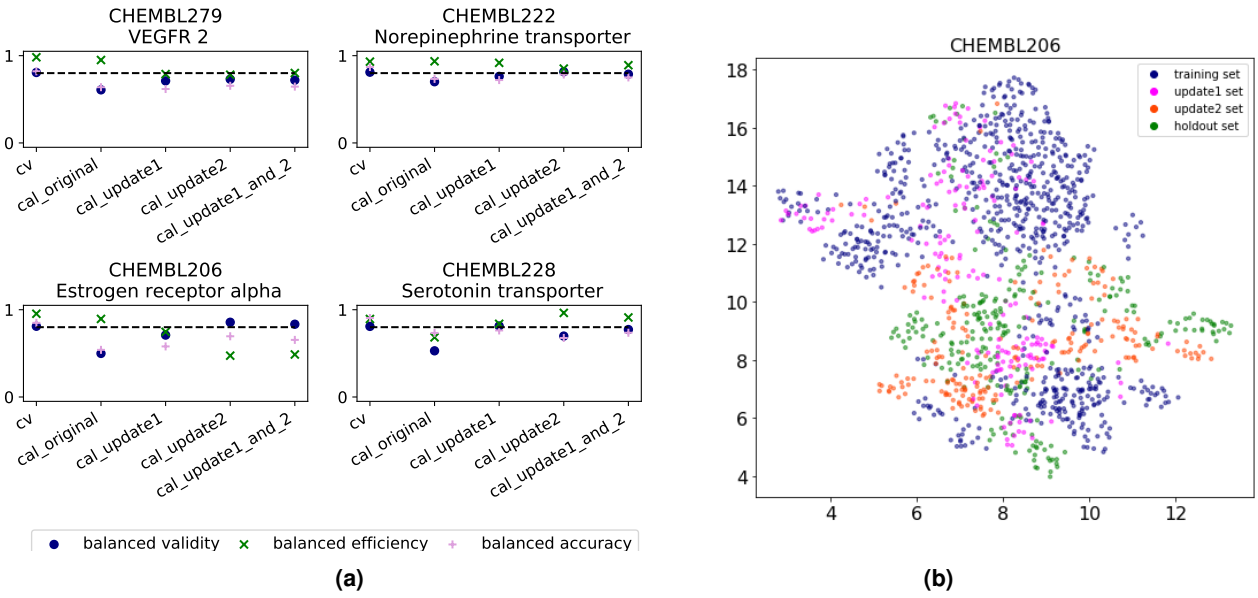


Figure 3. Analysis of individual endpoints **(a)** Balanced evaluation of time-split experiments for four selected ChEMBL endpoints. Each plot represents CV results (cv) and predictions for the holdout set using the original (cal_original), update1 (cal_update1), update2 (cal_update2) and combined update1_and_2 (cal_update1_and_2) calibration set. 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

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 Fig. 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 fewer 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 neighbour (NN) provided in Supplementary Fig. S6).

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

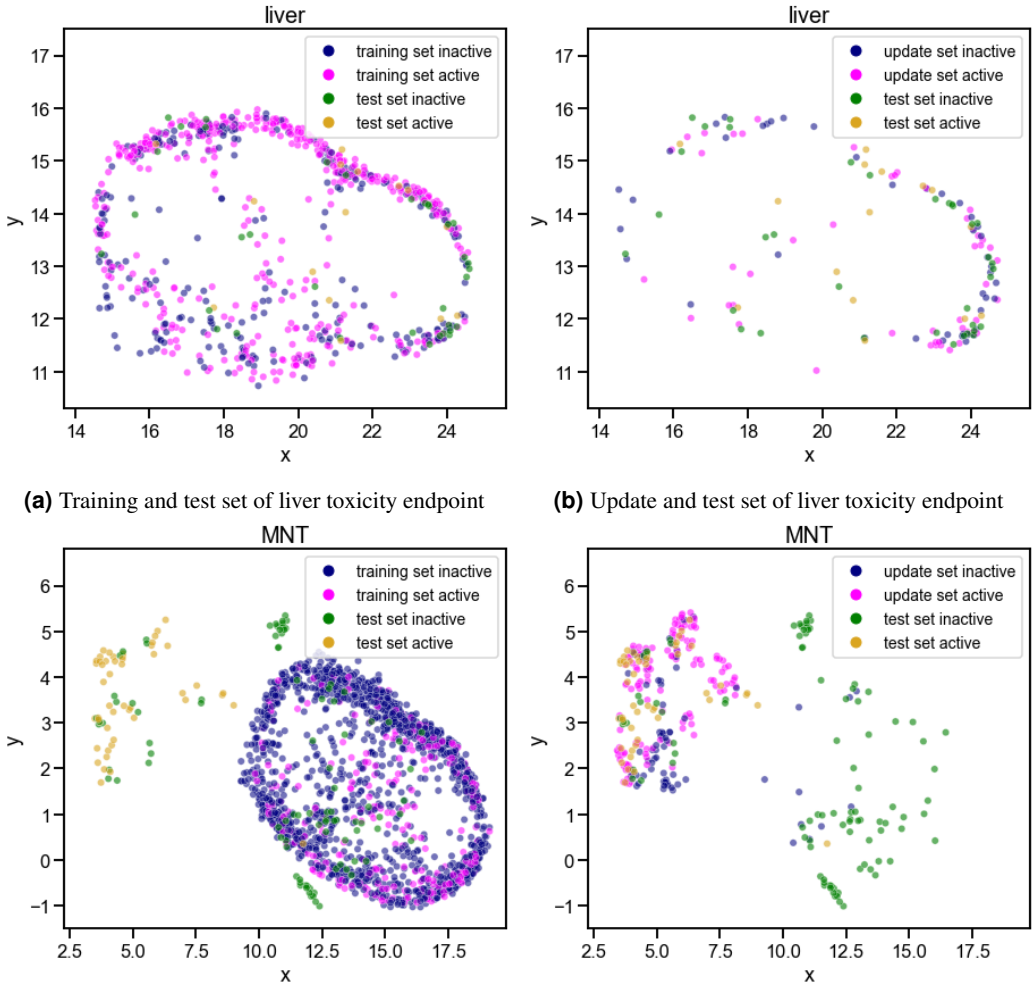
Applying these models to the holdout set containing internal data, the balanced validity dropped drastically by up to 0.34 points (liver toxicity: 0.47, MNT: 0.50). The balanced accuracy of the models also decreased strongly (liver toxicity: 0.43, MNT: 0.49), while the balanced efficiency increased (liver 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.

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

	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

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 and 4b). 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



(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 and b) and MNT datasets (c and 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 and c) and update sets (b and d), respectively.

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 and 4d), 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 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.

4 Conclusion

CP models, or generally ML models, are widely used for molecular property predictions, including activity or 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.

References

1. Zhang, L. *et al.* Applications of Machine Learning Methods in Drug Toxicity Prediction. *Curr. Top. Medicinal Chem.* **18**, 987–997, DOI: 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, DOI: 10.3389/978-2-88945-197-5 (2016).
3. Mansouri, K. *et al.* CoMPARA: Collaborative Modeling Project for Androgen Receptor Activity. *Environ. Heal. Perspectives* **128**, 027002, DOI: 10.1289/EHP5580 (2020).
4. Idakwo, G. *et al.* A review on machine learning methods for in silico toxicity prediction. *J. Environ. Sci. Heal. - Part C Environ. Carcinog. Ecotoxicol. Rev.* **36**, 169–191, DOI: 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. Cheminformatics* **12**, 1–17, DOI: 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, DOI: 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 Hong, H. (ed.) *Advances in Computational Toxicology*, chap. 11, 215–232 (Springer, Cham, 2019).

8. Mathea, M., Klingspohn, W. & Baumann, K. Chemoinformatic Classification Methods and their Applicability Domain. *Mol. Informatics* **35**, 160–180, DOI: 10.1002/minf.201501019 (2016).

9. OECD. *Guidance Document on the Validation of (Quantitative) Structure-Activity Relationship [(Q)SAR] Models* (OECD Publishing, Paris, 2014).

10. Vovk, V., Gammerman, A. & Shafer, G. *Algorithmic Learning in a Random World* (Springer Science & Business Media, 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, DOI: 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.* DOI: 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* **9**, DOI: 10.3390/biom9020043 (2019).

14. Stepanov, D., Canipa, S. & Wolber, G. HuskinDB, a database for skin permeation of xenobiotics. *Sci. Data* **7**, 1–8, DOI: 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.* DOI: 10.1021/acs.jcim.1c00293 (2021).

17. Morger, A. *et al.* Assessing the Calibration in Toxicological in Vitro Models with Conformal Prediction. *J. Cheminformatics* 1–14, DOI: 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.* DOI: 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, DOI: 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, DOI: 10.1021/acs.jcim.0c00250 (2020).

21. Mendez, D. *et al.* ChEMBL: Towards direct deposition of bioassay data. *Nucleic Acids Res.* **47**, D930–D940, DOI: 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, DOI: 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. Cheminformatics* **12**, 1–17, DOI: 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. Cheminformatics* **11**, 1–16, DOI: 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. Reports* **11**, 1–14, DOI: 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.* DOI: 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.* **21**, DOI: 10.3390/ijms21103585 (2020).

28. Watkins, P. B. Drug safety sciences and the bottleneck in drug development. *Clin. Pharmacol. Ther.* **89**, 788–790, DOI: 10.1038/clpt.2011.63 (2011).

29. OECD. *Test No. 474: Mammalian Erythrocyte Micronucleus Test* (OECD Publishing, Paris, 2016).

30. ICHS2(R1). Guidance on Genotoxicity Testing and Data Interpretation for Pharmaceuticals Intended for Human Use. *Int. Conf. on Harmon. Tech. Requir. for Regist. Pharm. for Hum. 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. Cheminformatics* **12**, 1–16, DOI: 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, DOI: 10.1021/acs.chemrestox.6b00135 (2016).

34. Bowes, J. *et al.* Reducing safety-related drug attrition: The use of in vitro pharmacological profiling, DOI: 10.1038/nrd3845 (2012).

35. OECD. *Test No. 407: Repeated Dose 28-day Oral Toxicity Study in Rodents* (OECD Publishing, Paris, 2008).

36. OECD. *Test No. 408: Repeated Dose 90-Day Oral Toxicity Study in Rodents* (OECD Publishing, Paris, 2018).

37. OECD. *Test No. 422: Combined Repeated Dose Toxicity Study with the Reproduction/Developmental Toxicity Screening Test* (OECD Publishing, Paris, 1996).

38. ChemAxon.

39. Berthold, M. R. *et al.* KNIME - the Konstanz information miner. *ACM SIGKDD Explor. Newsl.* **11**, 26, DOI: 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, DOI: 10.1016/j.jbiotec.2017.07.028 (2017).

41. Rogers, D. & Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **50**, 742–754, DOI: 10.1021/ci100050t (2010).

42. Landrum, G. A. RDKit: Open-source cheminformatics (2018).

43. Ji, C., Svensson, F., Zoufir, A. & Bender, A. eMolTox: prediction of molecular toxicity with confidence. *Bioinformatics* 1–2, DOI: 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.* DOI: 10.1021/ci5001168 (2014).

45. Vovk, V. Conditional validity of inductive conformal predictors. *Mach. Learn.* **92**, 349–376, DOI: 10.1007/s10994-013-5355-6 (2013).

46. Linusson, H. Nonconformist (2015).

47. Carlsson, L., Eklund, M. & Norinder, U. Aggregated Conformal Prediction. *IFIP Adv. Inf. Commun. Technol.* 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. *Proc. Sixth Work. on Conformal Probabilistic Predict. Appl.* **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, DOI: 10.1021/acs.jcim.8b00054 (2018).

53. Hunter, J. D. Matplotlib: A 2D graphics environment. *Comput. Sci. Eng.* **9**, 90–95, DOI: 10.1109/MCSE.2007.55 (2007).

54. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction (2018).

55. Vovk, V. Cross-conformal predictors. *Annals Math. Artif. Intell.* **74**, 9–28, DOI: 10.1007/s10472-013-9368-4 (2015).

56. Makili, L. E., Vega Sanchez, J. A. & Dormido-Canto, S. Active learning using conformal predictors: Application to image classification. *Fusion Sci. Technol.* **62**, 347–355, DOI: 10.13182/FST12-A14626 (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, DOI: 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. Medicinal Chem.* **47**, 2977–2980, DOI: 10.1021/jm030580l (2004).

60. Wang, R., Fang, X., Lu, Y., Yang, C. Y. & Wang, S. The PDBbind database: Methodologies and updates. *J. Medicinal Chem.* **48**, 4111–4119, DOI: 10.1021/jm048957q (2005).

61. Wu, Z. *et al.* MoleculeNet: A benchmark for molecular machine learning. *Chem. Sci.* **9**, 513–530, DOI: 10.1039/c7sc02664a (2018).

62. Norman, B. H. Drug Induced Liver Injury (DILI). Mechanisms and Medicinal Chemistry Avoidance/Mitigation Strategies. *J. Medicinal Chem.* **63**, 11397–11419, DOI: 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 (Basel)* **19**, DOI: 10.3390/s19040964 (2019).

64. Chen, M. *et al.* DILIrank: The largest reference drug list ranked by the risk for developing drug-induced liver injury in humans. *Drug Discov. Today* **21**, 648–653, DOI: 10.1016/j.drudis.2016.02.015 (2016).

65. OECD. eChemPortal (2020).

66. Benigni, R. *et al.* 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.* **16**, DOI: 10.2903/sp.efsa.2019.en-1598 (2019).

67. Yoo, J. W. *et al.* Development of improved QSAR models for predicting the outcome of the in vivo micronucleus genetic toxicity assay. *Regul. Toxicol. Pharmacol.* **113**, 104620, DOI: 10.1016/j.yrtph.2020.104620 (2020).

68. Moretto, A. Experimental and clinical toxicology of anticholinesterase agents. *Toxicol. Lett.* **102-103**, 509–513, DOI: 10.1016/S0378-4274(98)00245-8 (1998).

69. Hogan, D. B. Long-term efficacy and toxicity of cholinesterase inhibitors in the treatment of Alzheimer disease. *Can. J. Psychiatry* **59**, 618–623, DOI: 10.1177/070674371405901202 (2014).

70. Bianchini, D., Jayanth, A., Yu, J. C. & Cunningham, D. Epidermal growth factor receptor inhibitor-related skin toxicity: Mechanisms, treatment, and its potential role as a predictive marker. *Clin. Color. Cancer* **7**, 33–43, DOI: 10.3816/CCC.2008.n.005 (2008).

71. Hervent, A. S. & De Keulenaer, G. W. Molecular mechanisms of cardiotoxicity induced by ErbB receptor inhibitor cancer therapeutics. *Int. J. Mol. Sci.* **13**, 12268–12286, DOI: 10.3390/ijms131012268 (2012).

72. Buluş, A. D. *et al.* The evaluation of possible role of endocrine disruptors in central and peripheral precocious puberty. *Toxicol. Mech. Methods* **26**, 493–500, DOI: 10.3109/15376516.2016.1158894 (2016).

73. La Merrill, M. A. *et al.* Consensus on the key characteristics of endocrine-disrupting chemicals as a basis for hazard identification. *Nat. Rev. Endocrinol.* **16**, 45–57, DOI: 10.1038/s41574-019-0273-8 (2020).

74. Eskens, F. A. & Verweij, J. The clinical toxicity profile of vascular endothelial growth factor (VEGF) and vascular endothelial growth factor receptor (VEGFR) targeting angiogenesis inhibitors; A review. *Eur. J. Cancer* **42**, 3127–3139, DOI: 10.1016/j.ejca.2006.09.015 (2006).

75. Grosser, T., Fries, S. & FitzGerald, G. A. Biological basis for the cardiovascular consequences of COX-2 inhibition: Therapeutic challenges and opportunities. *J. Clin. Investig.* **116**, 4–15, DOI: 10.1172/JCI27291 (2006).

76. Guengerich, F. P. Common and uncommon cytochrome P450 reactions related to metabolism and chemical toxicity. *Chem. Res. Toxicol.* **14**, 611–650, DOI: 10.1021/tx0002583 (2001).

77. Vandenberg, J. I., Walker, B. D. & Campbell, T. J. HERG K+ channels: Friend and foe. *Trends Pharmacol. Sci.* **22**, 240–246, DOI: 10.1016/S0165-6147(00)01662-X (2001).

78. Nicotra, A. & Parvez, S. H. Cell death induced by MPTP, a substrate for monoamine oxidase B. *Toxicology* **153**, 157–166, DOI: 10.1016/S0300-483X(00)00311-5 (2000).

79. Mayer, A. F. *et al.* Influences of norepinephrine transporter function on the distribution of sympathetic activity in humans. *Hypertension* **48**, 120–126, DOI: 10.1161/01.HYP.0000225424.13138.5d (2006).

80. Stahl, S. M. Mechanism of action of serotonin selective reuptake inhibitors. Serotonin receptors and pathways mediate therapeutic effects and side effects. *J. Affect. Disord.* **51**, 215–235, DOI: 10.1016/S0165-0327(98)00221-3 (1998).

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 statement

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.

Additional information

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. F.S. served as a consultant for BASF SE. Other authors do not have conflict of interest.

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, 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.

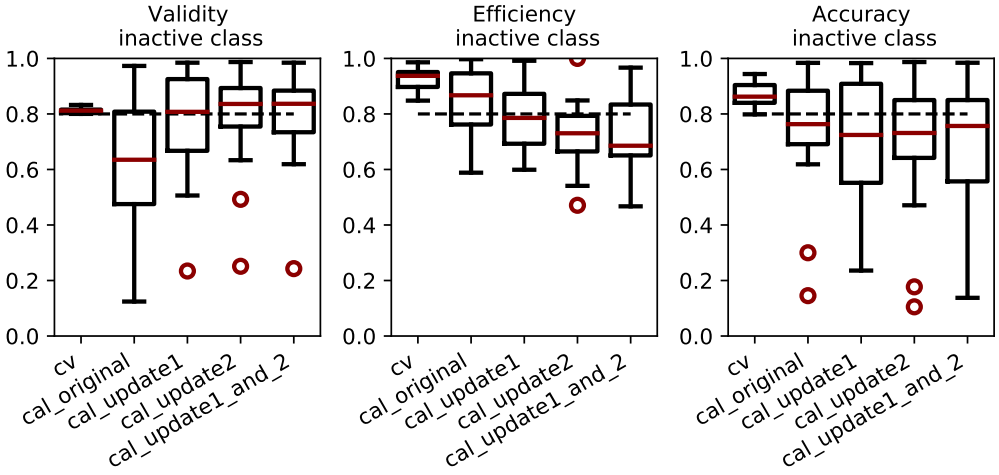
ChEMBL data After standardising the compounds (see 2.1.2), the ChEMBL data were time-split into four datasets, i.e. train, update1, update2, and holdout set based on the publication year. A minimum number of compounds per dataset was defined based on a predefined ratio, i.e. the training set must contain 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.

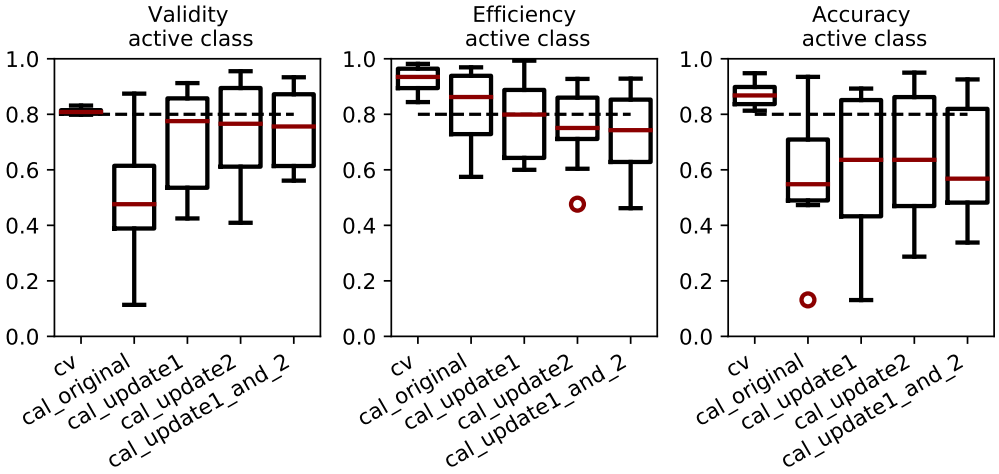
Table S1. ChEMBL datasets and their biological relevance. A selection of possible toxicological or adverse effects due to agonism (or activation) or antagonism (or inhibition) with the targets is provided.

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.

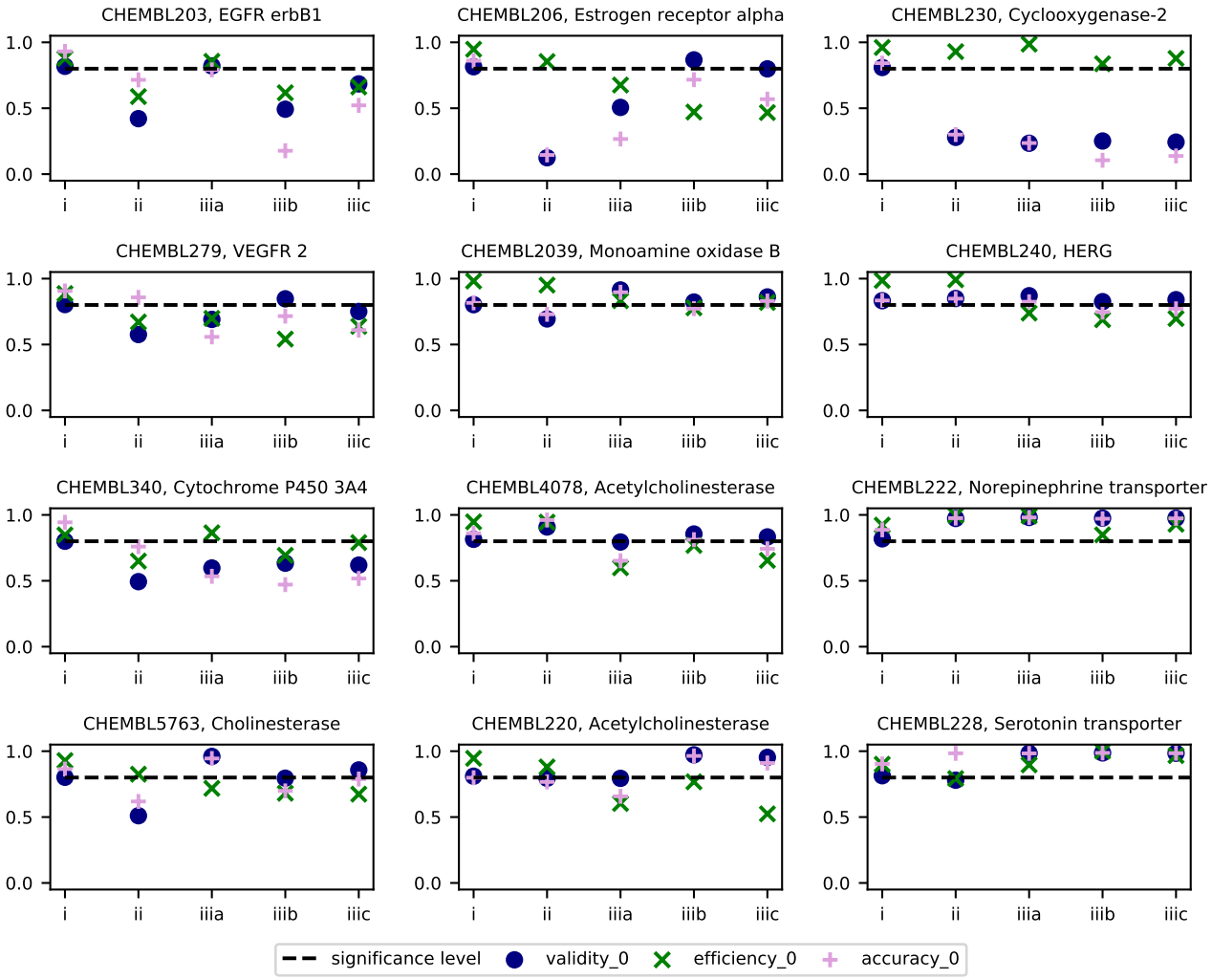


Figure S2. Inactive compounds evaluation of time-split experiments for individual ChEMBL endpoints. i) cross-validation on training data, predict holdout data using ii) original calibration set iiia) update1, iiib) update2, iiic) combined update1+2 calibration sets.

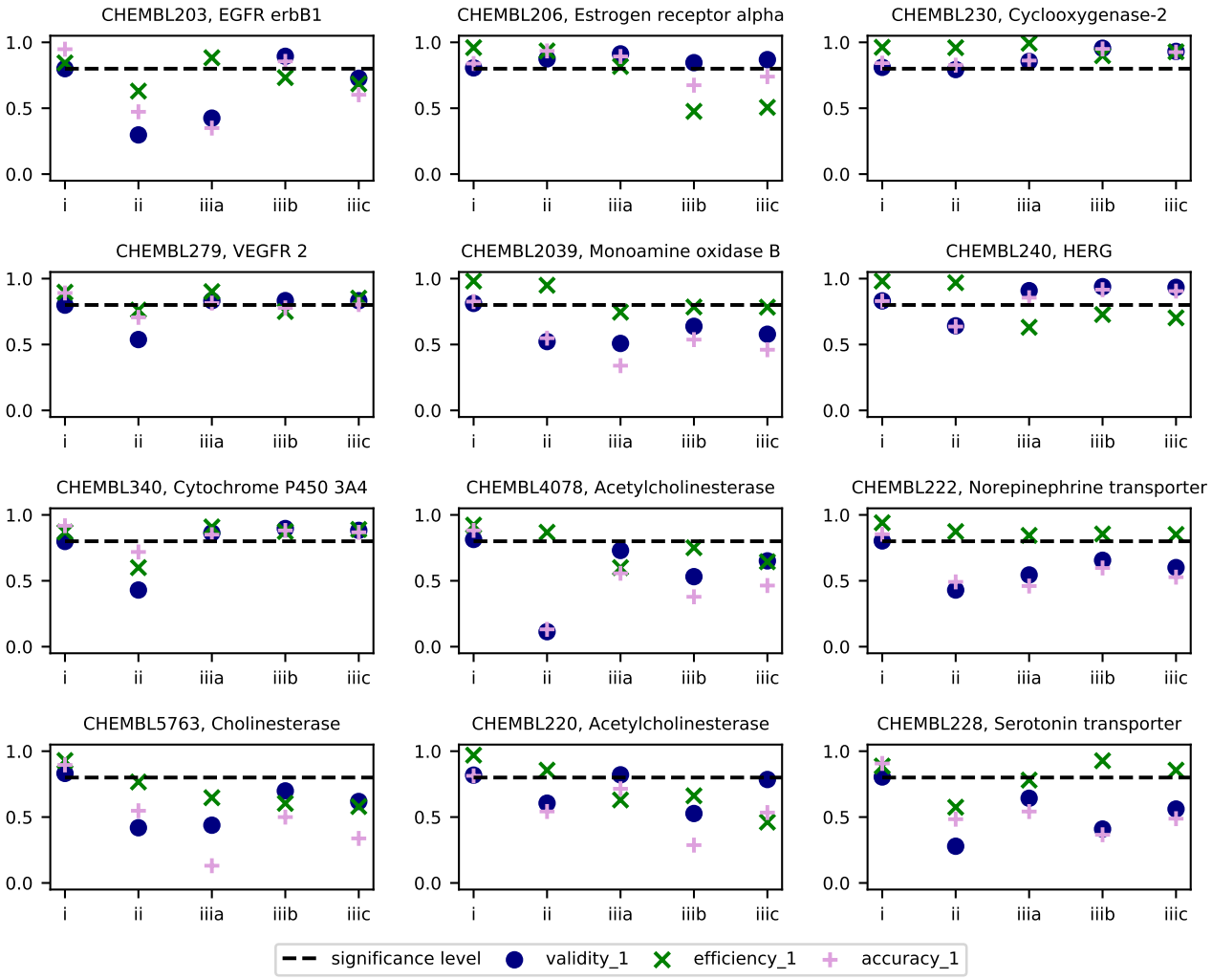


Figure S3. Active compounds evaluation of time-split experiments for individual ChEMBL endpoints. i) cross-validation on training data, predict holdout data using ii) original calibration set iiia) update1, iiib) update2, iiic) combined update1+2 calibration sets.

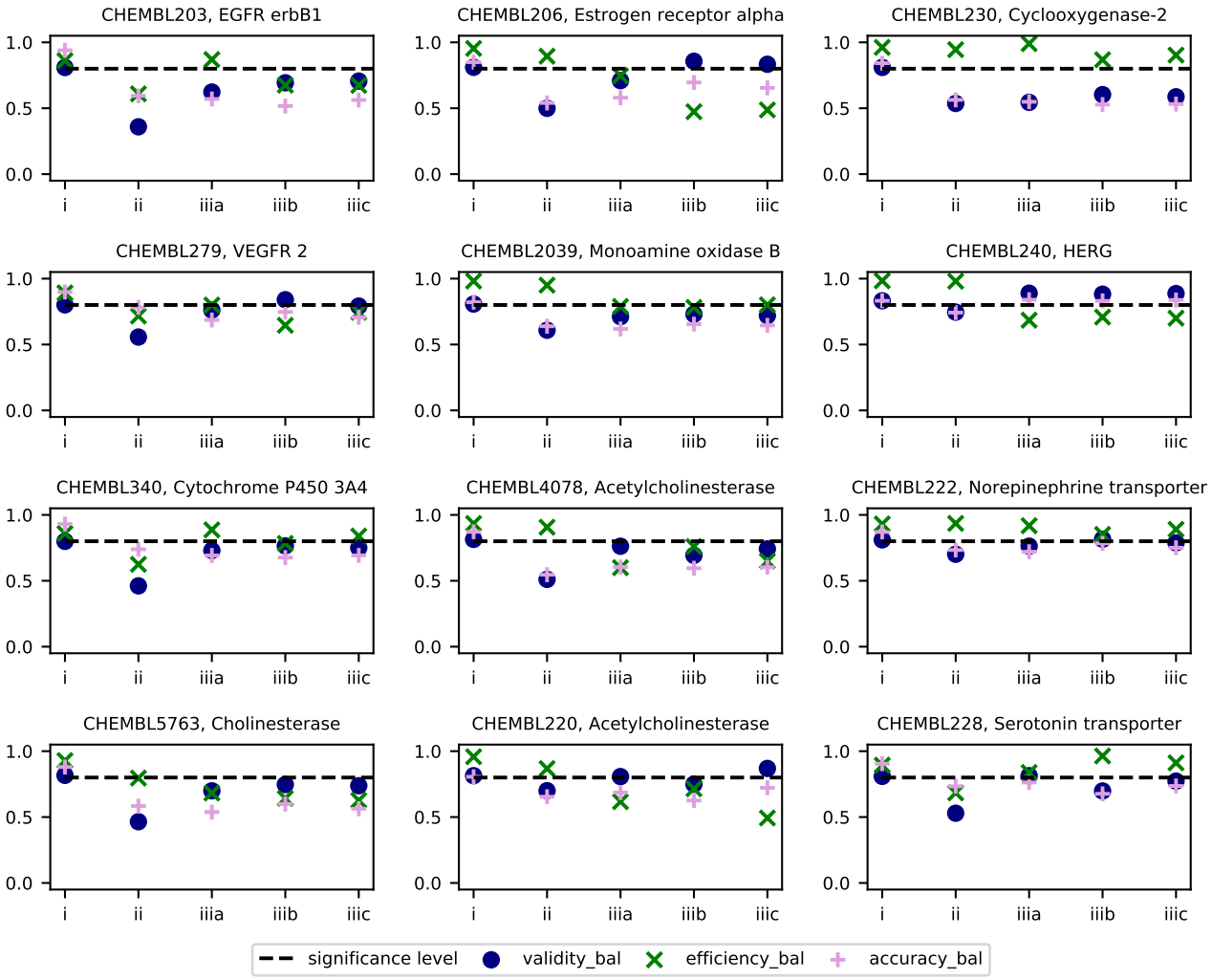


Figure S4. Balanced evaluation of time-split experiments for individual ChEMBL endpoints. i) cross-validation on training data, predict holdout data using ii) original calibration set, iii) updated calibration set, a) update1, b) update2, c) combined update1+2 sets. The doted line at 0.8 denotes the expected validity for the chosen significance level (of 0.2).

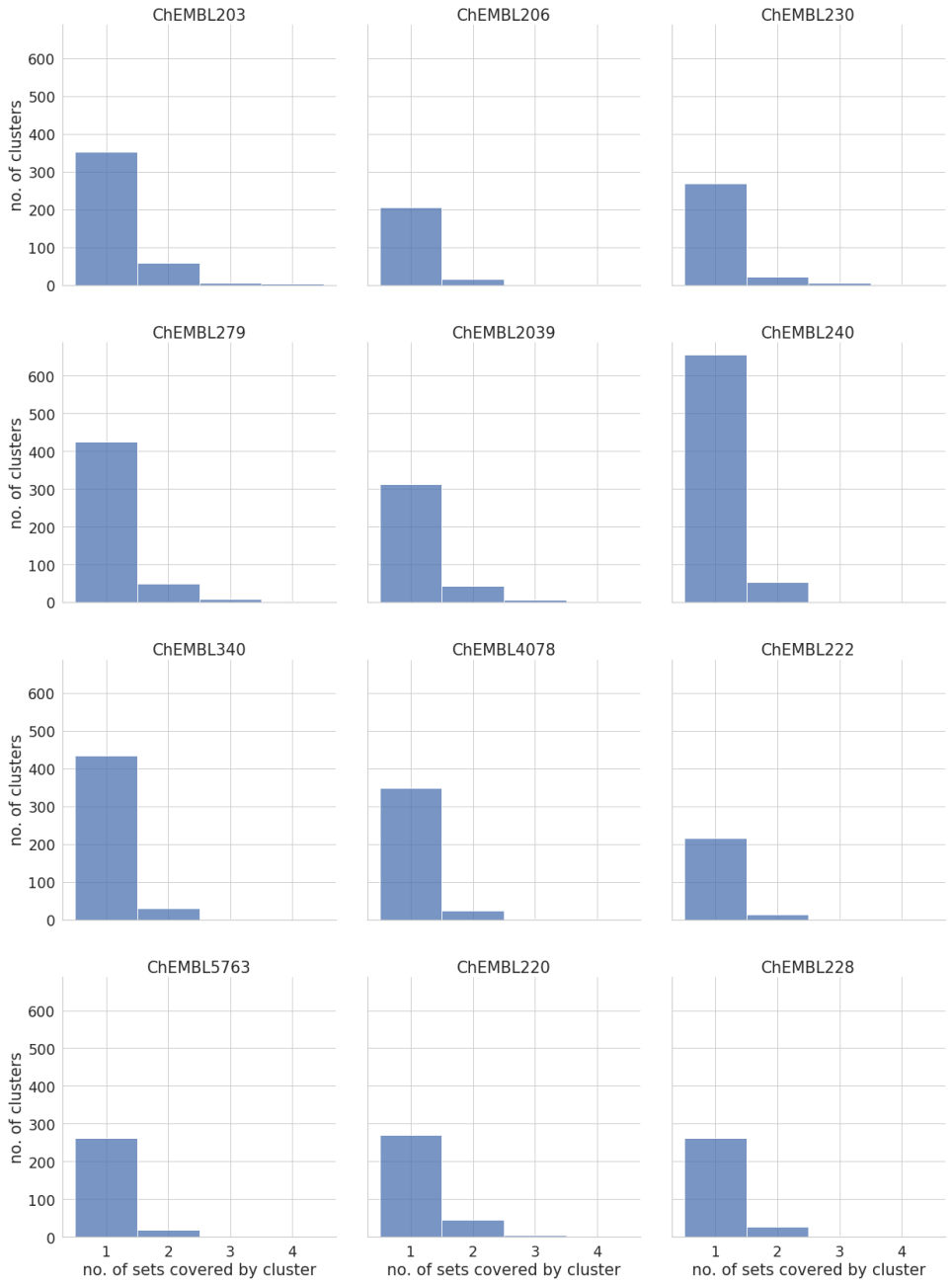
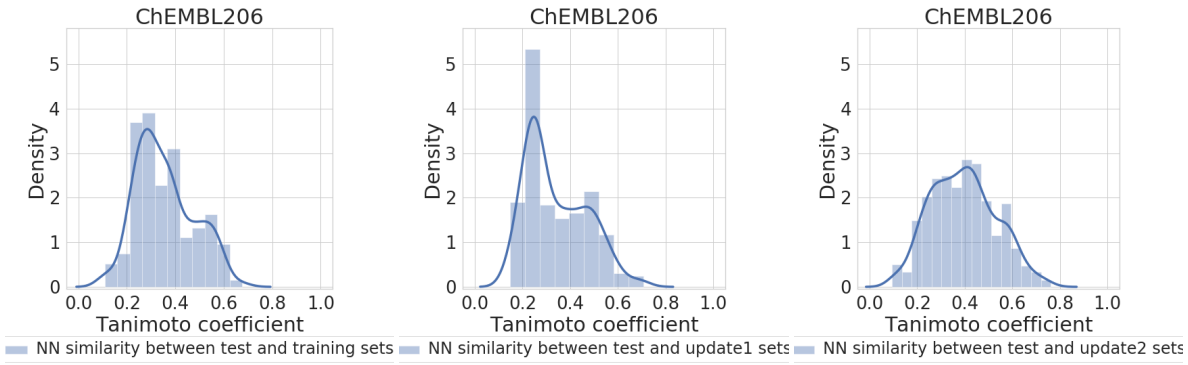


Figure S5. Spreading of clusters amongst the data subsets (i.e. splits) for the ChEMBL datasets. Most of the clusters (with at least two compounds) do not spread over more than one subset (i.e. training, update1, update2 or holdout set).



(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 .

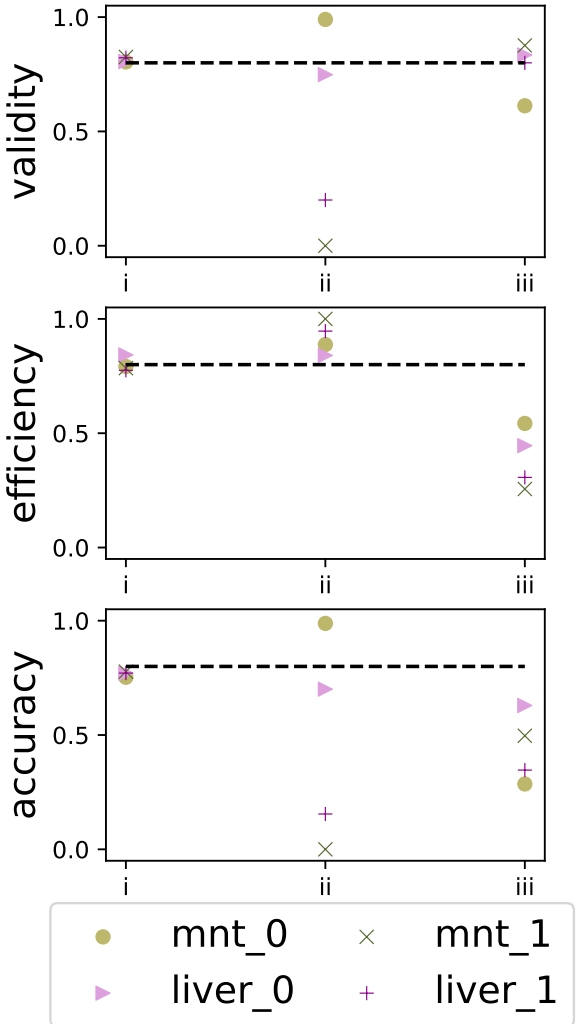


Figure S7. Time split evaluation (validity, efficiency, accuracy) of experiments i) CV, predictions using ii) original calibration set, iii) update calibration set for the liver toxicity and MNT inhouse datasets.

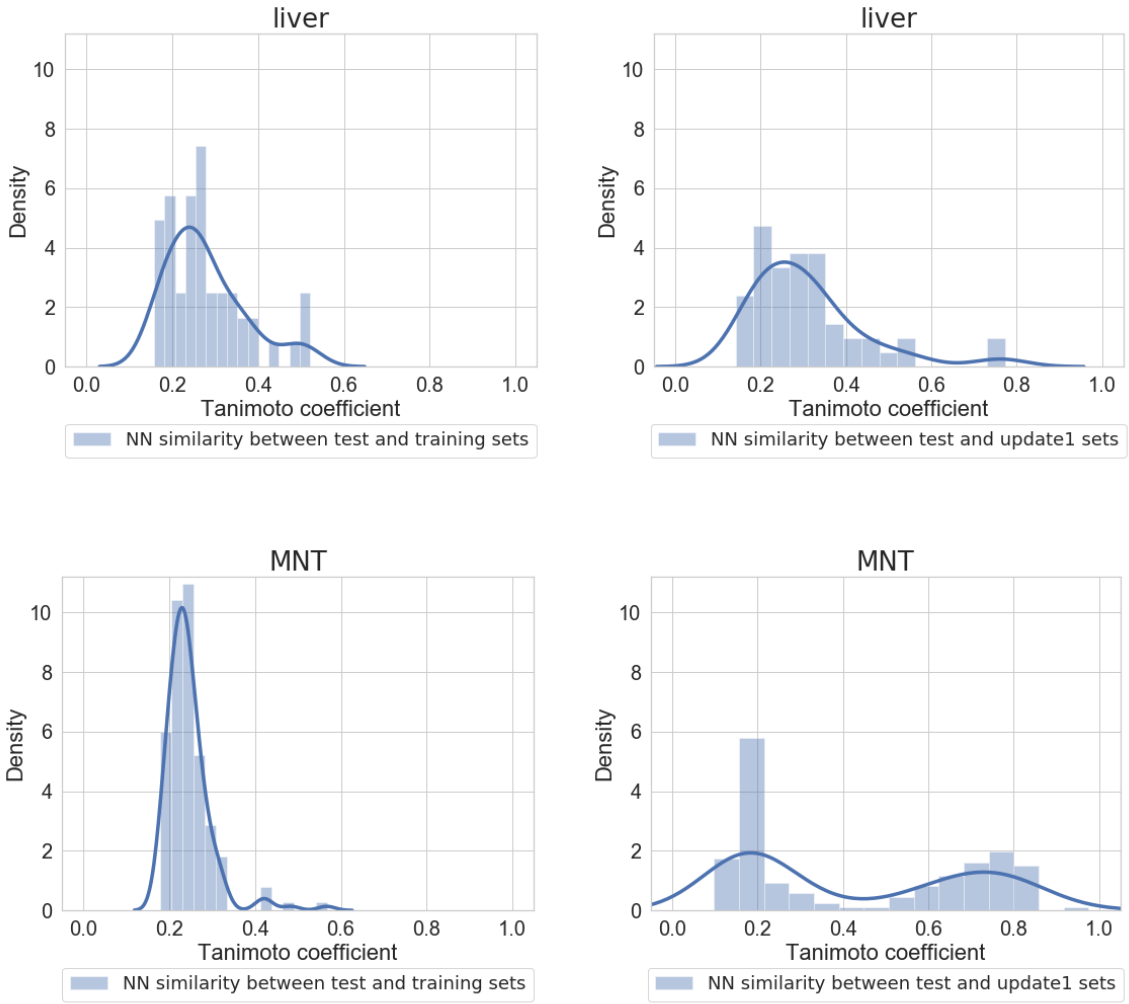


Figure S8. Distribution of Tanimoto coefficients between each holdout compound to its nearest neighbour in the training (left) and update (right) set for the liver (top) and MNT (bottom) endpoints.

Discussion

Drugs and other chemical compounds may not only be beneficial, but can also cause undesired effects to humans and the environment. Therefore, hazard assessment for new chemicals is crucial. Toxic effects of chemicals on humans are still mainly estimated with the help of *in vivo* studies, but animal studies are linked to ethical, economic, and technical concerns [40]. Following the 3R's principle by Russell and Burch [27], there is great need and interest to replace, reduce, and refine animal testing. *In silico* toxicity prediction methods can have an increasing impact on the reduction and the replacement of animal experiments. In this chapter, first, applications of multiple computational toxicology methods are reviewed. Second, two main challenges in computational toxicology, i.e. the need for confidence estimation and data limitations, are discussed. Finally, an outlook on computational toxicology as an alternative to animal experimentation is presented.

5.1 *In silico* methods for toxic endpoint prediction

KnowTox tool for holistic toxicity prediction

The KnowTox tool [127] was designed to support toxicologists in planing toxicological studies and to support the toxicological assessment of chemical substances. It integrates three methods that can help to assess the hazard of chemical substances: KnowTox comprises CP models for 88 toxicity-related biological effects, alerts for 919 toxic substructures, and read-across support based on 7912 molecules which have been experimentally tested in up to 985 *in vitro* assays.

The potential of combining multiple computational toxicity prediction methods into one platform has also been recognised by other groups around the time when KnowTox was developed. To name a few, the Protox II webserver offers toxicity predictions based on molecular similarity and ML models [128, 129]. Another example is the VEGA-HUB platform, which incorporates multiple tools, mainly read-across based, but also ML models, available for regulatory purposes for the exploration and analysis of chemical properties with a focus on ecotoxicology [130, 131]. Probably most similar to KnowTox is the e-MolTox webserver [65, 66]. E-MolTox provides CP models and toxic substructure analysis, also returning the most similar molecule within a training dataset. Another important source for estimating the (eco-)toxicological potential of new chemicals

is the OECD QSAR Toolbox, which was developed by the Laboratory of Mathematical Chemistry in close collaboration with the European Chemicals Agency (ECHA) [19, 35]. Nevertheless, there are advantages of developing an in-house tool, such as having access to the source code and being able to modify and extend it to individual needs. KnowTox consists of individual building blocks, which can be exchanged or extended based on the needs and data availability. Moreover, KnowTox outputs a customisable report, and the CP strategy was validated on new molecules which were candidates for product development at BASF.

The holistic approach from KnowTox was shown beneficial when applied in a retrospective case study on two former triazole fungicide candidates, which were attrited during toxicity studies [127]. The main reasons for their attrition, i.e. growth of liver and aromatase inhibition could retrospectively be explained with the KnowTox model outputs. This indicates that, in the case of the two triazoles, two critical types of toxicity could have been discovered during development, and thus that computational predictions can successfully guide project decisions and inform on the choice of critical toxicological screening assays.

The underlying hypothesis of KnowTox was that the combination of different computational toxicology methods, which integrate data from different sources, can lead to predictions with improved reliability estimates. This hypothesis was to some extent corroborated in the provided case study. For example, both a structural alert as well as ML predictions warned against potential liver toxicity of the triazole query molecules. A concordant result among multiple predictive methods implies more confidence, while contradicting predictions may reduce the reliability and call for experimental investigations. Certain reliability estimates can already be extracted from the integrated methods. To estimate the confidence in ML predictions in particular, the CP framework (see Section 5.2) was employed. Also, the similarity search provides support for reliability estimation, both by calculating the Tanimoto similarity to the query compound and by highlighting the maximum common substructure. Incorporating information from different data sources was further demonstrated to be a strategy to tackle data limitations. If no ML model was trained on a given dataset due to data available for fewer than 300 active or 300 inactive chemicals (e.g. certain cytochromes in the triazole case study), ML predictions may be substituted with a similarity search and subsequent extraction of existing experimental data for sufficiently similar molecules.

The KnowTox prediction outputs for a novel query compound — depending on the data situation — can yield a large amount of data to interpret. CP confidence estimates and Tanimoto similarity help estimating the reliability of individual predictions. The relevance of a model output is yet to be judged by toxicology experts, who subsequently decide, if a compound with predicted hazard should be excluded, or if any and which (confirmatory) assays should be conducted.

To facilitate the interpretation and encourage the application of the holistic tool, it

could be helpful to focus on so-called 'cut-off' criteria in regulatory toxicology (e.g. mutagenicity, carcinogenicity, reproductive toxicity). Therefore, a grouping of models, biological assays for read-across, and alerts per endpoint or toxicity type should be explored in future studies. Further challenges of the KnowTox tool and the underlying methods and data, respectively, are discussed in Sections 5.2 and 5.3.

E-Morph Screen Assay combined with *in silico* methods to increase hit rates

During the validation of the E-Morph Screen Assay [132], the benefits of a combination of *in silico* and *in vitro* methods were investigated in the example of estrogen receptor activity. The E-Morph Screen assay measures changes in the organisation of adherens junctions and can be used to identify substances that interact with the estrogen system [133]. For the assay validation, ToxCast assays relevant for ER activity were collected and a consensus CP model was built on the corresponding seven ER agonism datasets. Starting with nine novel ER active compounds identified with the E-Morph Screen, a similarity search within ToxCast, as well as support from the ER consensus CP model led to the detection of six additional novel estrogenic substances, which were experimentally validated. The study showed that the hit rate of the E-Morph Screen measurements was clearly increased when combining it with computational similarity search and CP. This means that ER active compounds can be found in a faster and cheaper manner when combining experimental and *in silico* methods.

The usefulness of CP to discover active compounds at a higher rate has also been described by Svensson et al. [47]. They showed how iterative rounds of CP predictions, docking, and experimental screening can lead to an increased hit rate in a retrospective study on 41 targets from the Directory of Useful Decoys, Enhanced (DUD-E) dataset. Notably, in our study, we could demonstrate the hit rate increase prospectively.

Our study also underlined that similarity is relative. Depending on the selected descriptors and similarity measures, two molecules may be perceived as more or less similar. This, together with the impact that the molecule size and the position of differing functional groups may have [49, 57], makes it challenging to define a similarity threshold in a screening workflow and to compare outcomes from different 'strategies'. Instead of a threshold, our study focused on the 10 most similar compounds and final screening candidates were selected after subsequent literature search. Another challenge occurred in the form of activity cliffs. Among a subset of compounds from ToxCast including the four most similar molecules to Bisphenol F, the most similar (Tanimoto index, various molecular fingerprints), and only ER-inactive molecule, was 4-Benzylphenol. The only difference to Bisphenol F is one missing hydroxy group. With similarity search — several types of molecular descriptors were investigated — or CP, this difference was not reckoned enough. Both ligand-based methods do not take into account the properties of the target protein.

Ligand-protein interactions could, however, be analysed with a docking approach. 4-

Benzylphenol was virtually placed in the binding pocket of estrogen receptor alpha protein structures. We could illustrate that a hydrogen bond, present in the interactions with the bisphenols, could not be formed due to the lack of the second hydroxy group. In fact, other groups demonstrated that docking can be supportive in screening cascades. For example, Svensson et al. [47] used docking to pre-select an initial set of compounds for screening. Ballante et al. [134] summarised prospective studies using molecular docking to screen for G protein-coupled receptor ligands. While docking was successful to identify new ligands in several cases, systematic integration of docking was out of scope for this thesis for several reasons. One of the biggest challenges of docking is the relative prediction of binding affinities [134, 135]. Moreover, the usefulness of docking heavily relies on the availability of protein structures for the target or at least a close analogue together with the amino acid sequence. Success of docking may also rely on availability of the correct ligand stereochemistry [135, 136].

In our work [132], a proof-of-concept for the combination of *in vitro* and *in silico* methods to increase the hit rate in *in vitro* screens was demonstrated. It is important to note that an increase in the hit rate does not guarantee that all predictions are correct. Rather, a larger proportion of *in silico* predictions are correct, which still need to be confirmed *in vitro*. Hence having the *in vitro* E-Morph Screen assay in place is a big advantage. The *in vitro*-*in silico* combination requires less testing, i.e. saving time and costs, while still giving experimental results for the more promising compounds. In a next step, the proposed combination strategy could be applied to screen larger databases for novel ER-active substances. Furthermore, the activity cliffs observed with several types of molecular descriptors might be tackled by more mechanism-focused, e.g. bioactivity or CHEMBIO [137], descriptors.

CHEMBIO descriptors for improved prediction of *in vivo* endpoints

The aim of the ChemBioSim project [137] was to build accurate CP models for the prediction of *in vivo* toxicity endpoints such as genotoxicity, hepatotoxicity, and cardiotoxicity. We wanted to find out if results from *in vitro* assays that are related to the *in vivo* endpoint, can inform and hence improve the models. To avoid dependency on experimental data for the training and test compounds, predicted bioactivity descriptors were generated by building CP classification models on 373 *in vitro* datasets.

It should be highlighted that the prediction of *in vivo* endpoints is an important challenge in itself. Thomas et al. [138] and Liu et al. [55] have tackled this with the example of the ToxRef database. While well-performing models could be generated in the latter study, no specific strategy was shown to yield overall best-performing models for every endpoint.

The application of *in vitro* assay data in the form of bioactivity descriptors was developed by Petrone et al. [53], who were looking for a new method to interrogate large-scale chemical biology data. Since the underlying data originates from HTS assays, the descrip-

tors were named HTS fingerprints. In several follow-up studies [54, 139–141], these HTS fingerprints were used e.g. for virtual screening, hit expansion, or target prediction. One major limitation of HTS fingerprints is the need for large amounts of available experimental data [141]. An elegant way to circumvent the often missing or sparse experimental data is the building of ML or CP models, and using the prediction outputs as bioactivity descriptors, as recently demonstrated by Cortes, Škuta, Norinder et al. [126, 142, 143].

In the ChemBioSim project [137], binary conformal predictors with underlying random forest models were built on *in vitro* datasets and both predicted p-values were used as bioactivity descriptors. The performance of bioactivity, chemical (physicochemical and molecular), and CHEMBIO (combination of the bioactivity and chemical) descriptors was compared when training CP models for the micro nucleus test (MNT), cardiac, and hepatic toxicity *in vivo* endpoints. In the prediction of MNT and cardiotoxicity, CHEMBIO or bioactivity descriptors were shown superior compared to chemical descriptors only. For the hepatotoxicity endpoint, no significant performance change was observed. These performance differences indicate that a conclusion drawn for a selection of *in vivo* endpoints may not be directly transferable to all other *in vivo* toxicological endpoints, for different reasons. First, it has to be noted that when using chemical descriptors alone the liver toxicity model was already more efficient compared to the models for MNT and cardiotoxicity, leaving limited room for improvement. Especially, since experimental measurements come with inherent noise and errors [141, 144], including predicted bioactivity descriptors may introduce more noise than benefit for already well-performing models. Secondly, it was discussed that the information gain from bioactivity descriptors may be reduced with high correlation of their features with the chemical descriptors. Finally, it is important that the encoded information is relevant for the *in vivo* endpoint to be predicted [143]. One could consider pre-selecting *in vitro* datasets for model building with known relation of the biological effects to the *in vivo* endpoint. In the ChemBioSim project [137], predicted bioactivity descriptors from all available models were used and, subsequently, feature importance was investigated. Such feature importance studies afford opportunities to find new relationships between query compound and bioactivity, and can help to explore the mode of action of novel compounds.

5.2 Confidence in machine learning predictions

Applicability, reliability, and decidability domains for conformal predictions

In order to confidentially apply ML models, it is crucial to determine their AD. In this work, CP was used as an AD approach for confidence estimation, i.e., taking into account the decision boundary between two classes. The CP framework includes several parts of the AD concept proposed by Hanser et al. [95], who suggested to split the AD definition into three steps: applicability domain (AD_{Hanser}), reliability domain (RD_{Hanser}), and

decidability domain (DD_{Hanser}). See Section 1.4 for a description of the three domains.

In principle, all three 'Hanser' domains are covered by the CP framework, although in a different manner. The DD_{Hanser} is clearly reflected in the CP framework returning prediction sets, and not being restricted to single-point estimates. single-label predictions, i.e. prediction sets that contain exactly one class, can be considered inside the DD_{Hanser}, whereas empty prediction sets and those containing both classes are regarded as outside the DD_{Hanser}. The content of the prediction set does not only depend on a single-point estimation, but on the calibrations based on two class-wise sorted lists. The calibrated p-values may be interpreted as RD_{Hanser}, i.e. the higher the p-value (and the lower the p-value for another class [96]) the more reliable the assignment of a class label to the compound. A difference to RD_{Hanser} is that the p-values in CP are only available after making the prediction, while RD_{Hanser} can e.g. already be determined based on the distance to and density of the nearest neighbours in the predictor space [95]. At first glance, one may be tempted to treat the AD_{Hanser} equal to whether a prediction set contains any class label or is empty. There are, however, three arguments indicating deviations from the concepts introduced by Hanser et al. Firstly, by definition, the AD_{Hanser} is supposed to be determined before making a prediction, and no prediction should be made if a compound is outside [95]. Yet, the CP framework does by design allow a certain fraction of false predictions [97], which may be empty prediction sets. Secondly, whether a prediction set is empty does not only depend on the instance to predict, but also on the chosen significance level [97]. Thirdly, in novelty detection, the applicability domain is typically based on the descriptor space [94], while in CP (at a given significance level), the prediction sets are derived from the p-value space. Nevertheless, with the CP concepts, the AD_{Hanser} can partly be investigated.

The limitations of CP are reached when AD_{Hanser} estimates for new, and especially individual, compounds are desired. Therefore, we still rely on traditional AD definitions, e.g. based on the descriptor space, the distance to, or the local density of the nearest neighbours [94, 95].

Moreover, it is desirable that the applicability domain of a CP model could be tailored to the set of instances to which it is applied. The, for this purpose, developed strategy to mitigate effects of data drifts is discussed in the next section.

Mitigation of data drifts with conformal prediction (recalibration strategy)

In our work on the Tox21 datasets [145], a recalibration strategy was introduced and shown useful to mitigate effects of data drifts on the model applicability domain. The recalibration strategy suggests to exchange the calibration set with new data if data drifts are observed. Data drifts can lead to poorly calibrated models. Poor calibration can, for example, be detected in the form of deviations from the diagonal line when plotting expected versus observed error rate in a so-called calibration plot [87, 98, 101, 127, 145].

CP underlies the exchangeability assumption: Provided that training and test data

stem from the same distribution, the CP framework is designed to yield valid models [96, 98]. Note that exchangeability is slightly weaker than the assumption of data being Independent and Identically Distributed ($I.I.D.$), which is generally assumed for all ML methods [98]. Nevertheless, in many cheminformatics use-cases, data drifts between training and test data are observed [87, 101, 127, 145, 146]. This may be due to entering a new chemical space (e.g. developing a new chemical series), further developing an assay for a certain biological effect, or when experiments have been conducted in different labs, by different people, or under different experimental conditions. Such data drifts typically become visible in CP when valid models are trained in a cross-validation, but validity is not given anymore when the same model is applied to a new dataset.

Similarly to when a measuring probe needs to be calibrated for field work in a new environment, CP models should be well-calibrated for the data to predict. If the model is not well-calibrated, the predictions are not necessarily reliable [95, 101].

Hence, our recalibration strategy was developed [145]. We propose to exchange the calibration set with available new data that is more similar to the test set than the original calibration set. The recalibration strategy was first developed with the example of the Tox21 data [145]. In a follow-up study, the strategy was then further investigated for the use on time-split ChEMBL data and to address differences between external and internal (BASF) data [147]. Our studies could show that validity can be restored, if calibration and test data stem from the same distribution, and if the size of the datasets is large enough. Note that, in practice, a calibration set stemming from the exact same distribution as the test data is not often available as the calibration set is needed prior to making predictions [145].

If applied as in the recalibration study, restored validity often comes with a loss in efficiency. Given a well-calibrated model, more compounds outside the AD (or more specifically outside the DD_{Hanser}) are recognised and classified as both prediction sets, which is represented in the low efficiency [145]. While, following Hanser et al. [92], validity is considered first priority for model applicability, efficiency (i.e. a high proportion of single-label predictions) is, nevertheless, important for the usefulness of a model. More research is needed to study which levels of efficiency are still acceptable. Similarly to the error rate, such an efficiency threshold may, of course, be context dependent. An improvement in efficiency might be achieved by retraining a model with an updated training set, though retraining requires access to training data and stronger computational resources. The efficiency, which is related to RD_{Hanser} and DD_{Hanser}, could e.g. be improved by screening compounds from the both and empty prediction sets and including this data in the training set [85]. A recent conformal regression study on assay transition showed that the efficiency of valid models could be improved when using data from old and new assays for training, but only using data from the current assay for calibration [101]. It would be interesting to investigate if these findings are transferable to the use-case of time-split data and data from different sources, as well as to the classification context.

As a drawback, this approach will not spare the retraining of a new ML model.

5.3 Toxicity data as a basis for *in silico* predictions

The quality of the ML and other modelling approaches highly depends on the quality and quantity of available data. In fact, other researchers have observed indications that, for the performance of ML models, data may matter more than the algorithm [73, 148, 149]. To be used for QSAR modelling, toxicological datasets are ideally large and contain well-curated structures and consistent measurements from standardised assays. Models should be built for endpoints with high relevance for the intended use-case and ideally training data and instances to be predicted are $I.I.D.$. Note that $I.I.D.$ is an assumption and can typically not be guaranteed before making a prediction, especially when predicting individual instances.

Comparing sources of toxicological data

For the work performed within this thesis, several popular freely-available sources of toxicological data were used, among others from the ToxCast, Tox21, and ChEMBL databases, as well as proprietary data from industry.

ToxCast and Tox21 The ToxCast database comprises HTS data for about 8000 chemical substances, such as pharmaceuticals, pesticides, or cosmetic ingredients, which have been tested for up to about 1000 biological effects. The largest subset origins from the Tox21 platform, which was combined with ToxCast [77, 102, 103].

ToxCast and Tox21 are large publicly-available toxicity databases with measurements taken from consistent assays [102], which makes them useful and popular for ML model building in the toxicology area [7, 12, 77–79, 128].

Nevertheless, 8000 is still a small number, especially since experimental measurements are noticeably not available for all 1000 biological effects. According to the REACH regulation, about 32,000 registered chemical substances were on the market in 2018 [150]. Moreover, in 2014, more than 15,000 molecules were registered in the Chemical Abstracts Service (CAS) in a single day [151]. Also drug libraries of the pharmaceutical industry can contain between tens of thousands and millions of chemicals [103]. Moreover, the predictive power of this HTS data for *in vivo* endpoints, and hence their impact on regulatory toxicology, has been discussed by several authors [138, 152–154]. On the one hand, they outlined limited predictivity of ToxCast phase I assays and chemicals to estimate *in vivo* hazards [138] and carcinogenicity [153]. On the other hand, authors described the potential of ToxCast for chemical prioritisation with respect to endocrine disruption [152] and *in vivo* hazards [138]. Punt et al. [154] investigated the potential of ToxCast to assess the safety of food chemicals. They suggested that it could have an

impact on regulatory risk assessment, especially by elucidating mechanistic information and for read-across.

In the KnowTox project [127], the ToxCast data was helpful to show the concept of holistic toxicity prediction and for the development of a retrospective case study. If available, it would, however, be desired to build the tool around more relevant and larger datasets. For example, the integrated CP models could be complemented or replaced by models for toxicity-related biological effects that are more easily transferable to *in vivo* or clinical data. Likewise, outcomes of similarity search and subsequent read-across support might be more informative if performed on more complete and larger datasets.

ChEMBL ChEMBL is a manually curated database, the ChEMBL Release 28 contains more than 2 million distinct drug-like molecules and activity measurements on more than 14,000 targets [106, 108].

Hence, ChEMBL provides a much larger source of data, which is also freely available, yet it has to be noted that not all targets are related to toxicity. With the measurements, ChEMBL provides temporal information in the form of the publication year [108]. This additional information is useful for splitting ML datasets during model validation [155]. Although ChEMBL is one of the few databases that comes with such temporal information, the order of the publication date does not necessarily coincide with the order of synthesising a molecule or testing it. Since authors may collect data for a series of compounds for one publication, we observed [147] that such time-splits of the ChEMBL data may less resemble such of datasets from industry, but rather a scaffold split, which is considered more challenging than a time split [156]. Similarly, Yang et al. had demonstrated that time-split cross-validation on public domain data may present a more challenging task than in an industrial setting [157].

Finally, the ChEMBL data originating from different sources and the recording of different units and prefixes require additional data curation steps [142, 158].

Proprietary data from industry The datasets from industry used in two projects [127, 147] were provided by BASF. We had access to internal data for androgen receptor antagonism (YAS assay), liver toxicity, and genotoxicity (MNT).

Such proprietary data present precious real-world data. As opposed to ToxCast data which was specifically generated for modelling, the main purpose of proprietary toxicity data is safety assessment. Hence, if a molecule is assessed to be toxic at one critical endpoint, typically no further measurements will be taken. Since large datasets are required for ML, available models are often restricted to the most frequently conducted assays and studies. For example, the YAS assay is among the more frequently conducted assays and the dataset has a comparatively balanced ratio between positive and negative outcomes. Therefore, sufficient AR antagonism data was available for model validation in the KnowTox project [127]. If data was available, the normalisation strategy would,

ideally, have been externally validated for each of the CP models contained in KnowTox.

Limitations of toxicity data and how to overcome them

Toxicity testing is an exclusion criterion for a compound's further development. If confirmed toxic at one crucial endpoint, a substance will most likely not be tested in any further assay. Moreover, only compounds, which are considered safe, will be submitted to regulatory agencies. This, together with the fact that most of the data is proprietary, can lead to the typical and challenging sparse and imbalanced toxicological datasets.

Sparse data lead to smaller training datasets and may decrease the predictivity of ML models. One strategy to optimally utilise the limited available data, as well as to better estimate the reliability of the predictions, is to combine multiple models or modelling techniques in a consensus approach. In the KnowTox study [127], it was demonstrated that more information can be obtained by combining prediction outputs from CP, structural alerts, and read-across support. In the E-Morph Screen Assay validation project [132], trust in ER activity predictions was increased by consolidating conformal predictions from seven ER CP models in one single prediction [132]. A consensus approach, based on different ML models was also praised and seen effective in the literature, for example in the CERAPP and CoMPARA studies [12, 79]. Earlier, Reif et al. had developed a prioritisation scheme to determine a compound's potential endocrine activity based on information from *in vitro* assays, chemical descriptors, and biological pathways [152].

Nevertheless, it is desired to enhance available toxicity datasets with more data, for example by mutual data sharing between companies. Such important opportunities are already being embraced by members from industry and academia in consortia such as eTOX [3] and eTRANSAFE [159]. The aim of eTOX is the extraction and sharing of preclinical study data from 13 pharmaceutical companies which could subsequently be used for read-across and training of prediction models [3, 90, 160]. The eTRANSAFE project builds on the achievements of eTOX and focuses on translation between pre-clinical and clinical safety data. While substantial effort was required to safeguard intellectual property in the data [3, 159], an interesting outcome of the eTRANSAFE project is the concept of virtual control groups that are expected to reduce the number of animals required for *in vivo* studies [41]. If resources are available and ethical considerations allow (e.g. for *in vitro* data), one could strategically measure more compounds to enlarge available datasets. Such a testing could be guided by CP: Analysing the compounds predicted as empty or both prediction sets and updating the model could specifically increase the model applicability domain [85].

Small datasets are also a limitation for modern AI methods such as deep learning (DL), which is typically successful with big data. DL models behind google translate were trained on millions of data points [161] and the ImageNet Large Scale Visual Recognition Challenge was based on millions of images to train DL models for image recognition [162, 163]. This might explain why the winning DL models (DeepTox) of the Tox21 data

challenge outperformed traditional ML models only slightly. Both compared to their own models based on non-DL methods and compared to the models of other challenge participants, the difference in AUC on the final test set was below 0.1 [78].

Ignoring class-imbalances in datasets may lead to overestimation of global evaluation measures. For example, if a dataset contains less than 3% active compounds, as the case for the PPAR-γ training set from the Tox21 data challenge, a model could reach 97% accuracy but sensitivity of 0% by predicting all compounds as inactive. Hence, data imbalances need to be considered during training and evaluation. There are data-balancing techniques in ML such as under- or oversampling [164, 165]. While undersampling reduces the amount of data for modelling, oversampling techniques may create artificial data, i.e. data which was not initially present in the data and created synthetically. Such is the case of Smote [166]. CP typically provides a simple way to handle data imbalances when using the Mondrian condition [99]. The resolution of the p-values may be lower, but classes are handled separately [145]. Esposito et al. [167] recently suggested an alternative approach for handling imbalanced data by adjusting the decision threshold in ML. The method especially showed an advantage for classifying compounds ending up in the both prediction sets with CP.

Finally, the investigated modelling techniques (CP models, ToxAlerts, and read-across) were applied in the form of binary classifiers. Outputs can serve as indicators to detect potential hazardous effects of new molecules and thereby guide toxicity testing. It has to be noted that a single activity threshold per dataset is used to label compounds as toxic and non-toxic, respectively. In reality, different compounds may exert toxic effects at different exposure [153, 168, 169]. This concept is especially relevant for drugs and synthetic chemicals that may enter the human (or animal) body. Based on its pharmacokinetic profile, a chemical stays in the body for a certain amount of time and reaches a certain plasma concentration. The amount and frequency of how a drug is dosed further depends on its potency at a target. Hence, some information may be lost in binary classification. The toxicity threshold of classification models may be adapted or one may consider regression models, which output an activity value (or range — in conformal regression), which can be used for further analysis.

Last but not least, the data used in this work originates mainly from HTS assays. To estimate toxic effects *in vivo*, subsequent *in vitro* to *in vivo* extrapolation is required. In this work [137, 147], *in vitro* data in the form of bioactivity descriptors was used to predict *in vivo* toxicological outcomes. Provided that *in vitro* data is available, another approach for *in vitro* to *in vivo* extrapolation would be to use physiologically-based toxicokinetic modelling (PBTK) [37, 170, 171]. PBTK uses compound-specific (*in vitro* or *in silico*) and physiological input parameters, which are extrapolated to (*in vivo*) pharmacokinetic parameters. Predicted maximal plasma concentration levels of the compound are compared to lowest observed effect concentrations determined *in vitro* to identify at which dose a chemical could exhibit toxic effects [170]. Importantly, accurate modelling

of *in vivo* endpoints relies on the understanding of mechanisms and pathways, which may not yet be sufficiently elucidated for complex toxicological endpoints [31, 126, 172].

5.4 Replace, reduce, and refine animal testing

Ethical and economic reasons call for the replacement, reduction, and refinement of animal testing and computational toxicology provides a potential for the development of alternative methods [40, 173]. The investigated strategies such as combination of modelling methods, conformal prediction for confidence estimation, and recalibration to mitigate data drifts effects can potentially contribute to make ML models more widely applicable in the field of predictive toxicology. With the integrated confidence estimates and consensus approaches we aim at fostering trust in predictions and enhancing the acceptance of computational methods in the toxicology field.

Early attrition of hazardous molecules can save many animals and money and shorten discovery and development timelines [88, 173–175]. Hence, it is crucial to embed such models and methods for predictive toxicology into the discovery and early development pipeline for new chemicals while keeping in mind the applicability of the models and the confidence in predictions [83, 94, 95]. While the available methods can now already be used to guide the toxicity testing, for example to hint at most critical endpoints to be investigated, or to immediately exclude likely hazardous compounds [31, 173], it may still be a long way ahead until computational toxicology approaches are more widely accepted by regulatory agencies [31, 40, 43, 88]. An important reason for this matter of fact are the particularly high requirements on the accuracy, especially sensitivity, of such regulatory-accepted models. A wrong decision may lead to harmful effects on humans or the environment when a hazardous substance reaches the market.

Restricted data availability has been described as one of the main reasons for limited performance and trust in computational toxicity prediction. Furthermore, especially recently evolving techniques such as deep learning and reinforcement learning [176, 177] are based on big data. Therefore, it will most likely not be possible to completely replace animal experiments with *in silico* predictions in the near future [25, 174].

Nevertheless, the impact that predictive toxicology may have on the reduction of animal testing should not be neglected. Strategies suggested in this work, such as complementary and consensus approaches for more informative and reliable predictions, or the recalibration strategy to mitigate effects of data drifts aim to support the creation of safer chemicals while reducing the need for animal testing.

Conclusion

To stop harmful chemicals from entering the market, it is crucial to assess their toxic potential. Toxicological effects of chemical substances on humans, animals, and the environment have traditionally been assessed based on *in vivo* and *in vitro* experiments [40]. Computational methods, although only marginally accepted by regulatory agencies, have a large potential to reduce animal testing, especially in early phases of chemical development.

Computational toxicity prediction methods are often limited by the small amount of available data and confidence estimates are desired to estimate the reliability of the predictions. The aim of this work was to investigate if the combination of multiple *in silico* toxicology methods, such as substructure and similarity search, and CP, can improve the confidence and efficiency of predictions. A special focus was set on the CP framework. It is built upon ML models and the predicted probabilities are calibrated with the help of a calibration set. The power of the CP framework was explored for various applications: as a means for confidence estimation [127, 132], to calculate bioactivity descriptors [137, 147], and to mitigate effects of data drifts [145, 147].

With KnowTox [127], a toxicity prediction tool was developed that comprises CP models, structural alerts, and read-across support. In a retrospective case study, it was shown that gathering information originating from different methods is a promising way to gain more insight into the toxicity profile of a new chemical. Using KnowTox, it was possible to uncover the key toxic effects, i.e. liver toxicity and aromatase inhibition, of two former fungicide candidates from BASF that led to their attrition.

During the validation of the E-Morph Screen [132], a recently developed estrogen receptor assay for the detection of potential endocrine disrupting chemicals, we illustrated the prospective use of *in silico* toxicology methods for the prioritisation of compounds to be tested in *in vitro* assays. When pre-selecting potentially active compounds with the help of similarity search, a hit rate increase for detecting novel estrogenic substances in the ToxCast library could be achieved. In future studies, this *in-silico-in-vitro*-combination could be used to screen large datasets of yet uncharacterised chemical substances for endocrine disruptive agents.

In the ChemBioSim study [137], the use of *in vitro* assay information for the prediction of *in vivo* endpoints was investigated. To avoid synthesis and testing of every query molecule, CP models were trained on a collection of toxicological *in vitro* datasets

and the generated p-values, i.e. calibrated predicted probabilities, were used as so-called bioactivity descriptors. For MNT and cardiotoxicity, it could be demonstrated that the bioactivity descriptors, or their combination with chemical (molecular and physicochemical) descriptors, can yield superior predictions compared to using chemical descriptors alone.

Given exchangeability, CP is designed to yield automatically valid models at a predefined maximum error rate. The exchangeability assumption, which is also made for ML, is not always fulfilled. Typically, valid models are obtained in a cross-validation, however, when applied to the test data, validity may drop. Impaired validity on test data may be explained by data drifts, e.g. when exploring a new chemical space over time or when predicting data from different sources. Such data drifts can retrospectively be detected in poorly calibrated CP models. When comparing observed versus expected validity in so-called calibration plots, deviations from the diagonal line may be observed.

As an approach to mitigate data drift effects, the recalibration strategy was developed with the example of the Tox21 datasets [145]. Subsequently, application of the strategy to temporal data drifts in time-split ChEMBL datasets and to discrepancies between training and test data from different sources [147] was investigated. The recalibration strategy involves exchanging the calibration set with data assumed to be closer to the test set. It was demonstrated that validity can be restored or improved when the calibration data origins from an exact same or a similar distribution as the test set, respectively. While more research is needed to minimise the drop in efficiency when restoring validity, the recalibration strategy comes with the advantage that no retraining of the model is needed. Hence, a comparatively small dataset of updated data may be sufficient to retrieve valid models.

Considering the need for more and consistent data, as well as the complexity of certain toxicological endpoints, computational toxicology methods will most likely not be able to completely replace animal testing in the near future. Nevertheless, this work suggests ways to improve confidence in and applicability of the methods. The results illustrate how *in silico* toxicology can help to guide toxicity testing, hence contributing towards the development of safer chemicals while reducing animal experimentation.

Bibliography

1. World Health Organization. *The Public Health Impact of Chemicals: Knowns and Unknowns* 2016. `http://apps.who.int/iris/bitstream/handle/10665/206553/WHO_FWC_PHE_EPE_16.01_eng.pdf`.
2. Duffus, J. in *Fundamental toxicology for chemists* (eds Duffus, J. *et al.*) chap. 1 (The Royal Society of Chemistry, 1996).
3. Cases, M. *et al.* The eTOX Data-Sharing Project to Advance in Silico Drug-Induced Toxicity Prediction. *International Journal of Molecular Sciences* **15,** 21136–21154 (2014).
4. Waring, M. J. *et al.* An analysis of the attrition of drug candidates from four major pharmaceutical companies. *Nature Reviews Drug Discovery* **14,** 475–486 (2015).
5. Sacks, L. V. *et al.* Scientific and regulatory reasons for delay and denial of FDA approval of initial applications for new drugs, 2000-2012. *JAMA* **311,** 378–384 (2014).
6. Ball, N. *et al.* Toward good read-across practice (GRAP) guidance. *ALTEX* **33,** 149–166 (2016).
7. Yang, H. *et al.* In Silico Prediction of Chemical Toxicity for Drug Design Using Machine Learning Methods and Structural Alerts. *Frontiers in chemistry* **6,** 30 (2018).
8. Yang, X. *et al.* in *Advances in Computational Toxicology* (ed Hong, H.) 315–335 (Springer, 2019).
9. Buluş, A. D. *et al.* The evaluation of possible role of endocrine disruptors in central and peripheral precocious puberty. *Toxicology Mechanisms and Methods* **26,** 493–500 (2016).
10. Maffini, M. V. *et al.* Endocrine disruptors and reproductive health: The case of bisphenol-A. *Molecular and Cellular Endocrinology* **254-255,** 179–186 (2006).
11. Monneret, C. What is an endocrine disruptor? *Comptes Rendus - Biologies* **340,** 403–405 (2017).
12. Mansouri, K. *et al.* CoMPARA: Collaborative Modeling Project for Androgen Receptor Activity. *Environmental Health Perspectives* **128,** 027002 (2020).

13. Diamanti-Kandarakis, E. *et al.* Endocrine-disrupting chemicals: An Endocrine Society scientific statement. *Endocrine Reviews* **30,** 293–342 (2009).

14. Ain, Q. U. *et al.* in *Endocrine Disrupting Chemicals-induced Metabolic Disorders and Treatment Strategies* (eds Akash, M. S. H. *et al.*) 113–123 (Springer, Cham, 2021).

15. European Commission. Community Strategy for Endocrine Disrupters - a range of substances suspected of interfering with the hormone systems of humans and wildlife. *Comunication from the Commission to the Council and the European Parliament* (1999).

16. National Research Council. *Hormonally active agents in the environment* (National Academies Press (US), 1999).

17. OECD. *Revised Guidance Document 150 on Standardised Test Guidelines for Evaluating Chemicals for Endocrine Disruption* (OECD Publishing, Paris, 2018).

18. World Health Organization. WHO — Global assessment of the state-of-the-science of endocrine disruptors. *WHO* (2013).

19. Dimitrov, S. D. *et al.* QSAR Toolbox – workflow and major functionalities. *SAR and QSAR in Environmental Research* **27,** 203–219 (2016).

20. Rovida, C. *et al.* Integrated testing strategies (ITS) for safety assessment. *ALTEX* **32,** 25–40 (2015).

21. Anon. Regulation (EC) No 1907/2006 of the European Parliament and of the Council of 18 December 2006 concerning the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH), establishing a European Chemicals Agency, amending Directive 1999/4. *Official Journal of the European Union* (2006).

22. Hartung, T. *et al.* Chemical regulators have overreached. *Nature* **460,** 1080–1081 (2009).

23. Zaunbrecher, V. *et al.* Has toxicity testing moved into the 21st century? A survey and analysis of perceptions in the field of toxicology. *Environmental Health Perspectives* **125,** 1–10 (2017).

24. Teubner, W. *et al.* Read-across for hazard assessment: the ugly duckling is growing up. *Alternatives to laboratory animals : ATLA* **43,** P67–P71 (2015).

25. Huang, R. in *Advances in Computational Toxicology* (ed Hong, H.) 279–297 (Springer, Cham, 2019).

26. Bundesministerium für Ernährung und Landwirtschaft. *Verwendung von Versuchstieren im Jahr 2019* 2020.

27. Russell, W. *et al. The principles of humane experimental technique* 1959.

28. Hartung, T. *et al.* Food for thought ... on in silico methods in toxicology. *ALTEX* **26,** 155–166 (2009).

29. Kavlock, R. J. *et al.* Accelerating the Pace of Chemical Risk Assessment. *Chemical Research in Toxicology* **31,** 287–290 (2018).

30. Zhu, H. *et al.* Supporting Read-Across Using Biological Data. *ALTEX* **1848,** 3047–3054 (2016).

31. Hemmerich, J. *et al.* In silico toxicology: From structure–activity relationships towards deep learning and adverse outcome pathways. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **10,** 1–23 (2020).

32. Schultz, T. W. *et al.* Lessons learned from read-across case studies for repeated-dose toxicity. *Regulatory Toxicology and Pharmacology* **88,** 185–191 (2017).

33. ECHA. *The Use of Alternatives to Testing on Animals for the REACH Regulation* **3** (2014).

34. Van Ravenzwaay, B. *et al.* Metabolomics as read-across tool: A case study with phenoxy herbicides. *Regulatory Toxicology and Pharmacology* **81,** 288–304 (2016).

35. OECD. *OECD QSAR toolbox* 2020. `https://www.oecd.org/chemicalsafety/risk-assessment/oecd-qsar-toolbox.htm`.

36. OECD. *eChemPortal* 2020. `https://www.echemportal.org/echemportal/`.

37. Zhang, Q. *et al.* Bridging the Data Gap From in vitro Toxicity Testing to Chemical Safety Assessment Through Computational Modeling. *Frontiers in Public Health* **6** (2018).

38. Fourches, D. *et al.* Trust but verify: on the importance of chemical structure curation in chemoinformatics and QSAR modeling research. *Journal of Chemical Information and Modeling* **50,** 1189–1204 (2010).

39. Pham, L. L. *et al.* Variability in in vivo studies: Defining the upper limit of performance for predictions of systemic effect levels. *Computational Toxicology* **15** (2020).

40. Kusko, R. *et al.* in *Advances in Computational Toxicology* (ed Hong, H.) 1–11 (Springer, Cham, 2019).

41. Steger-Hartmann, T. *et al.* Computer-based prediction models in regulatory toxicology. *Regulatory Toxicology,* 123–131 (2020).

42. Vo, A. H. *et al.* An Overview of Machine Learning and Big Data for Drug Toxicity Evaluation. *Chemical Research in Toxicology* **33,** 20–37 (2020).

43. European Medicines Agency. *ICH guideline M7(R1) on assessment and control of DNA reactive (mutagenic) impurities in pharmaceuticals to limit potential carcinogenic risk* 2015.

44. Hsieh, J. H. *et al.* Harnessing in Silico, in Vitro, and in Vivo Data to Understand the Toxicity Landscape of Polycyclic Aromatic Compounds (PACs). *Chemical Research in Toxicology* **34,** 268–285 (2021).

45. Hoover, G. *et al.* In vitro and in silico modeling of perfluoroalkyl substances mixture toxicity in an amphibian fibroblast cell line. *Chemosphere* **233,** 25–33 (2019).

46. Hamid, N. *et al.* Prioritizing phthalate esters (PAEs) using experimental in vitro/ vivo toxicity assays and computational in silico approaches. *Journal of Hazardous Materials* **398,** 122851 (2020).

47. Svensson, F. *et al.* Improving Screening Efficiency through Iterative Screening Using Docking and Conformal Prediction. *Journal of Chemical Information and Modeling* **57,** 439–444 (2017).

48. Johnson, M. A. *et al. Concepts and applications of molecular similarity* (Wiley, 1991).

49. Maggiora, G. *et al.* Molecular similarity in medicinal chemistry. *Journal of Medicinal Chemistry* **57,** 3186–3204 (2014).

50. Stumpfe, D. *et al.* Evolving Concept of Activity Cliffs. *ACS Omega* (2019).

51. Maggiora, G. M. On outliers and activity cliffs - Why QSAR often disappoints. *Journal of Chemical Information and Modeling* **46,** 1535 (2006).

52. Stumpfe, D. *et al.* Recent progress in understanding activity cliffs and their utility in medicinal chemistry. *Journal of Medicinal Chemistry* **57,** 18–28 (2014).

53. Petrone, P. M. *et al.* Rethinking molecular similarity: Comparing compounds on the basis of biological activity. *ACS Chemical Biology* **7,** 1399–1409 (2012).

54. Riniker, S. *et al.* Using information from historical high-throughput screens to predict active compounds. *Journal of Chemical Information and Modeling* **54,** 1880–1891 (2014).

55. Liu, J. *et al.* Predicting Organ Toxicity Using in Vitro Bioactivity Data and Chemical Structure. *Chemical Research in Toxicology* **30,** 2046–2059 (2017).

56. Willett, P. *et al.* Chemical similarity searching. *Journal of Chemical Information and Computer Sciences* **38,** 983–996 (1998).

57. Holliday, J. *et al.* Analysis and display of the size dependence of chemical similarity coefficients. *Journal of Chemical Information and Computer Sciences,* 819–828 (2003).

58. Sushko, I. *et al.* ToxAlerts: A web server of structural alerts for toxic chemicals and compounds with potential adverse reactions. *Journal of Chemical Information and Modeling* **52,** 2310–2316 (2012).

59. Yang, H. *et al.* Evaluation of Different Methods for Identification of Structural Alerts Using Chemical Ames Mutagenicity Data Set as a Benchmark. *Chemical Research in Toxicology* **30,** 1355–1364 (2017).

60. Benigni, R. *et al.* Structural Alerts of Mutagens and Carcinogens. *Current Computer Aided-Drug Design* **2,** 169–176 (2006).

61. Liu, R. *et al.* Data-driven identification of structural alerts for mitigating the risk of drug-induced human liver injuries. *Journal of Cheminformatics* **7,** 4 (2015).

62. Allen, T. E. *et al.* Using 2D Structural Alerts to Define Chemical Categories for Molecular Initiating Events. *Toxicological Sciences* **165,** 213–223 (2018).

63. Wedlake, A. J. *et al.* Structural Alerts and Random Forest Models in a Consensus Approach for Receptor Binding Molecular Initiating Events. *Chemical Research in Toxicology* **33,** 388–401 (2020).

64. Brenk, R. *et al.* Lessons learnt from assembling screening libraries for drug discovery for neglected diseases. *ChemMedChem* **3,** 435–444 (2008).

65. Ji, C. *et al.* eMolTox: prediction of molecular toxicity with confidence. *Bioinformatics,* 1–2 (2018).

66. Ji, C. *eMolTox* 2018. `http://xundrug.cn/moltox`.

67. OCHEM. *ToxAlerts* 2012. `www.ochem.eu/alerts`.

68. Myden, A. *et al.* Utility of published DNA reactivity alerts. *Regulatory Toxicology and Pharmacology* **88,** 77–86 (2017).

69. Alves, V. M. *et al.* Alarms about structural alerts. *Green Chemistry* **18,** 4348–4360 (2016).

70. Nepali, K. *et al.* Nitro-Group-Containing Drugs. *Journal of Medicinal Chemistry* **62,** 2851–2893 (2019).

71. Rey Moreno, M. C. *et al.* Epoxiconazole-induced degeneration in rat placenta and the effects of estradiol supplementation. *Birth Defects Research Part B - Developmental and Reproductive Toxicology* **98,** 208–221 (2013).

72. Van Der Ven, L. T. *et al.* A Case Study with Triazole Fungicides to Explore Practical Application of Next-Generation Hazard Assessment Methods for Human Health. *Chemical Research in Toxicology* **33,** 834–848 (2020).

73. Aurélien Géron. *Hands-on Machine Learning with Scikit-Learn and TensorFlow* (O'Reilly Media, 2017).

74. Raschka, S. *Python Machine Learning* (Packt Publishing Ltd., 2015).

75. Raies, A. B. *et al.* In silico toxicology: computational methods for the prediction of chemical toxicity. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **6,** 147–172 (2016).

76. Pawar, G. *et al.* In silico toxicology data resources to support read-across and (Q)SAR. *Frontiers in Pharmacology* **10,** 1–26 (2019).

77. 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* (eds Huang, R. *et al.*) **2002,** 27582769 (Frontiers in Environmental Science, 2017).

78. Mayr, A. *et al.* DeepTox: Toxicity Prediction using Deep Learning. *Frontiers in Environmental Science* **3,** 80 (2016).

79. Mansouri, K. *et al.* CERAPP: Collaborative estrogen receptor activity prediction project. *Environmental Health Perspectives* **124,** 1023–1033 (2016).

80. Toivonen, H. *et al.* Statistical evaluation of the predictive toxicology challenge 2000-2001. *Bioinformatics* **19,** 1183–1193 (2003).

81. OECD. *Guidance on grouping of chemicals, Second Edition* 1–141 (OECD Publishing, Paris, 2017).

82. Cherkasov, A. *et al.* QSAR Modeling: Where Have You Been? Where Are You Going To? *Journal of Medicinal Chemistry* **57,** 4977–5010 (2014).

83. Tropsha, A. Best practices for QSAR model development, validation, and exploitation. *Molecular Informatics* **29,** 476–488 (2010).

84. Golbraikh, A. *et al.* in *Cheminformatics: Basic Concepts and Methods* (eds Engel, T. *et al.*) 465–495 (John Wiley & Sons, 2018).

85. Svensson, F. *et al.* Modelling compound cytotoxicity using conformal prediction and PubChem HTS data. *Toxicology Research* **6,** 73–80 (2017).

86. Badillo, S. *et al.* An Introduction to Machine Learning. *Clinical Pharmacology and Therapeutics* **107,** 871–885 (2020).

87. Cortés-Ciriano, I. *et al.* Concepts and Applications of Conformal Prediction in Computational Drug Discovery. *arXiv preprint arXiv:1908.03569* (2019).

88. Luechtefeld, T. *et al.* Big-data and machine learning to revamp computational toxicology and its use in risk assessment. *Toxicology Research* **7,** 732–744 (2018).

89. Göller, A. H. *et al.* Bayer's in silico ADMET platform: a journey of machine learning over the past two decades. *Drug Discovery Today* **25,** 1702–1709 (2020).

90. Sanz, F. *et al.* Integrative Modeling Strategies for Predicting Drug Toxicities at the eTOX Project. *Molecular Informatics* **34,** 477–484 (2015).

91. Tollefsen, K. E. *et al.* Applying Adverse Outcome Pathways (AOPs) to support Integrated Approaches to Testing and Assessment (IATA). *Regulatory Toxicology and Pharmacology* **70,** 629–640 (2014).

92. Hanser, T. *et al.* in *Advances in Computational Toxicology* (ed Hong, H.) 215–232 (Springer, Cham, 2019).

93. OECD. *Guidance Document on the Validation of (Quantitative) Structure-Activity Relationship [(Q)SAR] Models* (OECD Publishing, Paris, 2014).

94. Mathea, M. *et al.* Chemoinformatic Classification Methods and their Applicability Domain. *Molecular Informatics* **35,** 160–180 (2016).

95. Hanser, T. *et al.* Applicability domain: towards a more formal definition. *SAR and QSAR in Environmental Research* **27,** 865–881 (2016).

96. Vovk, V. *et al. Algorithmic learning in a random world* (Springer Science & Business Media, 2005).

97. Norinder, U. *et al.* Introducing conformal prediction in predictive modeling. A transparent and flexible alternative to applicability domain determination. *Journal of Chemical Information and Modeling* **54,** 1596–1603 (2014).

98. Alvarsson, J. *et al.* Predicting With Confidence: Using Conformal Prediction in Drug Discovery. *Journal of Pharmaceutical Sciences* **110,** 42–49 (2021).

99. Sun, J. *et al.* Applying Mondrian Cross-Conformal Prediction to Estimate Prediction Confidence on Large Imbalanced Bioactivity Data Sets. *Journal of Chemical Information and Modeling* **57,** 1591–1598 (2017).

100. Fedorova, V. *et al.* Plug-in martingales for testing exchangeability on-line. *Proceedings of the 29th International Conference on Machine Learning, ICML 2012* **2,** 1639–1646 (2012).

101. Arvidsson McShane, S. *et al.* Machine Learning Strategies When Transitioning between Biological Assays. *Journal of Chemical Information and Modeling* (2021).

102. Richard, A. M. *et al.* ToxCast Chemical Landscape: Paving the Road to 21st Century Toxicology. *Chemical Research in Toxicology* **29,** 1225–1251 (2016).

103. Richard, A. M. *et al.* The Tox21 10K Compound Library: Collaborative Chemistry Advancing Toxicology. *Chemical Research in Toxicology* **34,** 189–216 (2021).

104. Kim, S. *et al.* PubChem substance and compound databases. *Nucleic Acids Research* **44,** D1202–D1213 (2016).

105. Kim, S. *et al.* PubChem 2019 update: Improved access to chemical data. *Nucleic Acids Research* **47,** D1102–D1109 (2019).

106. Mendez, D. *et al.* ChEMBL: Towards direct deposition of bioassay data. *Nucleic Acids Research* **47,** D930–D940 (2019).

107. Gaulton, A. *et al.* ChEMBL: A large-scale bioactivity database for drug discovery. *Nucleic Acids Research* **40,** 1100–1107 (2012).

108. EMBL-EBI. *ChEMBL* 2021. `https://www.ebi.ac.uk/chembl/`.

109. Bento, A. P. *et al.* An open source chemical structure curation pipeline using RDKit. *Journal of Cheminformatics* **12,** 1–16 (2020).

110. Atkinson, F. *Standardiser* 2014. `https://github.com/flatkinson/standardiser`.

111. Dalke, A. *MACCS key 44* 2014. `http://www.dalkescientific.com/writings/diary/archive/2014/10/17/maccs_key_44.html`.

112. Landrum, G. A. *RDKit: Open-source cheminformatics* 2018. `http://www.rdkit.org`.

113. *RDKit MACCS source code* 2020. `https://github.com/rdkit/rdkit/blob/Release_2020_09_1/rdkit/Chem/MACCSkeys.py#L38`.

114. Rogers, D. *et al.* Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling* **50,** 742–754 (2010).

115. *RDKit descriptors* 2020. `https://www.rdkit.org/docs/GettingStartedInPython.html#list-of-available-descriptors`.

116. Faulon, J.-L. *et al.* The Signature Molecular Descriptor. 1. Using Extended Valence Sequences in QSAR and QSPR Studies. *Journal of Chemical Information and Computer Sciences* **43,** 707–720 (2003).

117. Faulon, J.-L. *et al.* The signature molecular descriptor. 4. Canonizing molecules using extended valence sequences. *Journal of Chemical Information and Computer Sciences* **44,** 427–436 (2004).

118. Holliday, J. *et al.* Grouping of Coefficients for the Calculation of Inter-Molecular Similarity and Dissimilarity using 2D Fragment Bit-Strings. *Combinatorial Chemistry & High Throughput Screening* **5,** 155–166 (2002).

119. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12,** 2825–2830 (2011).

120. Hastie, T. *et al. The Elements of Statistical Learning* (Springer, 2009).

121. Shafer, G. *et al.* A tutorial on conformal prediction. *Journal of Machine Learning Research* **9,** 371–421 (2008).

122. Vovk, V. Conditional validity of inductive conformal predictors. *Machine Learning* **92,** 349–376 (2013).

123. Linusson, H. *Nonconformist* 2015. `http://donlnz.github.io/nonconformist/`.

124. Svensson, F. *et al.* Conformal Regression for QSAR Modelling – Quantifying Prediction Uncertainty. *Journal of Chemical Information and Modeling* **58,** 1132–1140 (2018).

125. Papadopoulos, H. *et al.* Regression Conformal Prediction with Nearest Neighbours. *Journal of Artificial Intelligence Research* **40,** 815–840 (2011).

126. Cortés-Ciriano, I. *et al.* QSAR-derived affinity fingerprints (part 2): Modeling performance for potency prediction. *Journal of Cheminformatics* **12,** 1–17 (2020).

127. Morger, A. *et al.* KnowTox: Pipeline and case study for confident prediction of potential toxic effects of compounds in early phases of development. *Journal of Cheminformatics* **12,** 1–17 (2020).

128. Banerjee, P. *et al.* ProTox-II: a webserver for the prediction of toxicity of chemicals. *Nucleic Acids Research,* 1–7 (2018).

129. Preissner, R. *et al. ProTox-II* 2018. `https://tox-new.charite.de/protox_II/`.

130. Benfenati, E. *et al. VEGA HUB* `https://www.vegahub.eu/about-vegahub/`.

131. Benfenati, E. *et al.* in *Advances in Computational Toxicology* (ed Hong, H.) 365–381 (Springer, Cham, 2019).

132. Klutzny, S. *et al.* Quantitative high-throughput phenotypic screening for environmental estrogens using the E-Morph Screening Assay in combination with in silico predictions. *Environment International* **158** (2022).

133. Kornhuber, M. *et al.* The E-Morph Assay: Identification and characterization of environmental chemicals with estrogenic activity based on quantitative changes in cell-cell contact organization of breast cancer cells. *Environment International* **149** (2021).

134. Ballante, F. *et al.* Structure-Based Virtual Screening for Ligands of G Protein-Coupled Receptors: What Can Molecular Docking Do for You? *Pharmacological Reviews* (2021).

135. Torres, P. H. *et al.* Key topics in molecular docking for drug design. *International Journal of Molecular Sciences* **20,** 1–29 (2019).

136. Corbeil, C. R. *et al.* Variability in docking success rates due to dataset preparation. *Journal of Computer-Aided Molecular Design* **26,** 775–786 (2012).

137. Garcia de Lomana, M. *et al.* ChemBioSim: Enhancing Conformal Prediction of In Vivo Toxicity by Use of Predicted Bioactivities. *Journal of Chemical Information and Modeling* **61,** 3255–3272 (2021).

138. Thomas, R. S. *et al.* A comprehensive statistical analysis of predicting in vivo hazard using high-throughput in vitro screening. *Toxicological Sciences* **128,** 398–417 (2012).

139. Wassermann, A. M. *et al.* A screening pattern recognition method finds new and divergent targets for drugs and natural products. *ACS Chemical Biology* **9,** 1622–1631 (2014).

140. Helal, K. Y. *et al.* Public Domain HTS Fingerprints: Design and Evaluation of Compound Bioactivity Profiles from PubChem's Bioassay Repository. *Journal of Chemical Information and Modeling* **56,** 390–398 (2016).

141. Cortes Cabrera, A. *et al.* Optimal HTS Fingerprint Definitions by Using a Desirability Function and a Genetic Algorithm. *Journal of Chemical Information and Modeling* **58,** 641–646 (2018).

142. Škuta, C. *et al.* QSAR-derived affinity fingerprints (part 1): Fingerprint construction and modeling performance for similarity searching, bioactivity classification and scaffold hopping. *Journal of Cheminformatics* **12,** 1–16 (2020).

143. Norinder, U. *et al.* Using Predicted Bioactivity Profiles to Improve Predictive Modeling. *Journal of Chemical Information and Modeling* **60,** 2830–2837 (2020).

144. Kramer, C. *et al.* The experimental uncertainty of heterogeneous public Ki data. *Journal of Medicinal Chemistry* **55,** 5165–5173 (2012).

145. Morger, A. *et al.* Assessing the Calibration in Toxicological in Vitro Models with Conformal Prediction. *Journal of Cheminformatics* **13,** 1–14 (2021).

146. Cortés-Ciriano, I. *et al.* Discovering Highly Potent Molecules from an Initial Set of Inactives Using Iterative Screening. *Journal of Chemical Information and Modeling* **58,** 2000–2014 (2018).

147. Morger, A. *et al.* Studying and Mitigating the Effects of Data Drifts on ML Model Performance at the Example of Chemical Toxicity Data. *manuscript submitted.*

148. Banko, M. *et al.* Mitigating the paucity-of-data problem: exploring the effect of training corpus size on classifier performance for natural language processing. *The first international conference,* 1–5 (2001).

149. Halevy, A. *et al.* The unreasonable effectiveness of data. *IEEE Intelligent Systems* **24,** 8–12 (2009).

150. ECHA. *REACH 2018 registration statistics* 2018. `https://echa.europa.eu/documents/10162/750652/reach_2018_deadline_statistics_en.pdf/ecfe225f-caf0-5bad-7c7f-ce57d2c8938f`.

151. Wang, Zhongyu *et al.* in *Advances in Computational Toxicology* (ed Hong, H.) 15–36 (Springer, Cham, 2019).

152. Reif, D. M. *et al.* Endocrine Profling and Prioritization of Environmental Chemicals Using ToxCast Data. *Environmental Health Perspectives* **118,** 1714–1720 (2010).

153. Benigni, R. Evaluation of the Toxicity Forecasting Capability of EPA's ToxCast Phase I Data: Can ToxCast In Vitro Assays Predict Carcinogenicity? *Journal of Environmental Science and Health, Part C* **31,** 201–212 (2013).

154. Punt, A. *et al.* Potential of ToxCast Data in the Safety Assessment of Food Chemicals. *Toxicological Sciences* **174,** 326–340 (2020).

155. Lenselink, E. B. *et al.* Beyond the hype: Deep neural networks outperform established methods using a ChEMBL bioactivity benchmark set. *Journal of Cheminformatics* **9,** 1–14 (2017).

156. Mayr, A. *et al.* Large-scale comparison of machine learning methods for drug target prediction on ChEMBL. *Chemical Science* (2018).

157. Yang, K. *et al.* Analyzing Learned Molecular Representations for Property Prediction. *Journal of Chemical Information and Modeling* **59,** 3370–3388 (2019).

158. Sydow, D. *et al.* TeachOpenCADD: A teaching platform for computer-aided drug design using open source packages and data. *Journal of Cheminformatics* **11,** 1–7 (2019).

159. Pognan, F. *et al.* The eTRANSAFE Project on Translational Safety Assessment through Integrative Knowledge Management: Achievements and Perspectives. *Pharmaceuticals* **14,** 1–18 (2021).

160. Pastor, M. *et al.* Development of an infrastructure for the prediction of biological endpoints in industrial environments. Lessons learned at the eTOX project. *Frontiers in Pharmacology* **9,** 1–8 (2018).

161. Wu, Y. *et al.* Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *arXiv preprint arXiv:1609.08144,* 1–23 (2016).

162. Russakovsky, O. *et al.* ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision* **115,** 211–252 (2015).

163. Chen, H. *et al.* The rise of deep learning in drug discovery. *Drug Discovery Today* **23,** 1241–1250 (2018).

164. Idakwo, G. *et al.* A review on machine learning methods for in silico toxicity prediction. *Journal of Environmental Science and Health, Part C* **36,** 169–191 (2018).

165. Batista, G. *et al.* A study of the Behavior of Several Methods for Balancing Machine Learning Training Data. *ACM SIGKDD Explorations Newsletter* **6,** 20–29 (2004).

166. Chawla, N. V. *et al.* SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research* **16,** 321–357 (2002).

167. Esposito, C. *et al.* GHOST: Adjusting the Decision Threshold to Handle Imbalanced Data in Machine Learning. *Journal of Chemical Information and Modeling* **61,** 2623–2640 (2021).

168. Miljković, F. *et al.* Machine Learning Models for Human in Vivo Pharmacokinetic Parameters with In-House Validation. *Molecular Pharmaceutics* **18,** 4520–4530 (2021).

169. Thomas, R. S. *et al.* The Next Generation Blueprint of Computational Toxicology at the U.S. Environmental Protection Agency. *Toxicological Sciences* **169,** 317–332 (2019).

170. Fabian, E. *et al.* In vitro-to-in vivo extrapolation (IVIVE) by PBTK modeling for animal-free risk assessment approaches of potential endocrine-disrupting compounds. *Archives of Toxicology* **93,** 401–416 (2019).

171. Sarigiannis, D. A. *et al.* Physiology-based toxicokinetic modelling in the frame of the European Human Biomonitoring Initiative. *Environmental Research* **172,** 216–230 (2019).

172. Madden, J. C. *et al.* A Review of In Silico Tools as Alternatives to Animal Testing: Principles, Resources and Applications. *Alternatives to laboratory animals : ATLA* **48,** 146–172 (2020).

173. Muster, W. *et al.* Computational toxicology in drug development. *Drug Discovery Today* **13,** 303–310 (2008).

174. Stefaniak, F. Prediction of compounds activity in nuclear receptor signaling and stress pathway assays using machine learning algorithms and low-dimensional molecular descriptors. *Frontiers in Environmental Science* **3,** 1–7 (2015).

175. Carrió, P. *et al.* Toward a unifying strategy for the structure-based prediction of toxicological endpoints. *Archives of Toxicology* **90,** 2445–2460 (2016).

176. Popova, M. *et al.* Deep reinforcement learning for de novo drug design. *Science Advances* **4,** 1–15 (2018).

177. Zhou, Z. *et al.* Optimization of Molecules via Deep Reinforcement Learning. *Scientific Reports* **9,** 1–10 (2019).

List of Publications

1. Klutzny, S. *et al.* Quantitative high-throughput phenotypic screening for environmental estrogens using the E-Morph Screening Assay in combination with in silico predictions. *Environment International* **158.** https://doi.org/10.1016/j.envint.2021.106947 (2022).
2. Garcia de Lomana, M. *et al.* ChemBioSim: Enhancing Conformal Prediction of In Vivo Toxicity by Use of Predicted Bioactivities. *Journal of Chemical Information and Modeling* **61,** 3255–3272. https://doi.org/10.1021/acs.jcim.1c00451 (2021).
3. Morger, A. *et al.* Assessing the Calibration in Toxicological in Vitro Models with Conformal Prediction. *Journal of Cheminformatics* **13,** 1–14. https://doi.org/10.1186/s13321-021-00511-5 (2021).
4. Morger, A. *et al.* KnowTox: Pipeline and case study for confident prediction of potential toxic effects of compounds in early phases of development. *Journal of Cheminformatics* **12,** 1–17. https://doi.org/10.1186/s13321-020-00422-x (2020).
5. Sydow, D. *et al.* TeachOpenCADD: A teaching platform for computer-aided drug design using open source packages and data. *Journal of Cheminformatics* **11,** 1–7. https://doi.org/10.1186/s13321-019-0351-x (2019).