Technische Universität München
Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt

In silico modeling using *in vitro* high throughput screening data for toxicity prediction within REACH

Ahmed Mohamed Abdelaziz Sayed Ahmed

Vollständiger Abdruck der von der Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzender: Prof. Dr. Heiko Briesen

Prüfer der Dissertation:
1. Prof. Dr. Dr. Karl-Werner Schramm
2. Prof. Dr. Hans-Werner Mewes

Die Dissertation wurde am 27.04.2016 bei der Technischen Universität München eingereicht und durch die Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt am 22.09.2016 angenommen.

*"*The sciences do not try to explain, they hardly even try to interpret, they mainly make models. By a model is meant a mathematical construct which, with the addition of certain verbal interpretations, describes observed phenomena. The justification of such a mathematical construct is solely and precisely that it is expected to work.*"*

Johann Von Neumann US (Hungarian-born) computer scientist, mathematician (1903 - 1957)

"Prediction is very difficult, especially about the future."

Niels Henrik David Bohr Danish physicist (1885 – 1962)

Acknowledgements

I would like to express my sincere gratitude to Prof. Dr. Dr. Karl Werner Schramm for supervising my doctoral study. Prof. Schramm's advice allowed me to see the grander picture of how QSAR research can benefit the scientific community and guided me to investigate useful applications for the basic research. He pointed me in directions that would continue to inspire me for future research. I am indebted to his patience, support and trust along the way. His backing and insights helped me in steering the project and writing the thesis.

My Sincere thanks go to Prof. Dr. Hilde Spahn-Langguth for her continued trust, useful advices and for hosting me through my internship in Mainz. I would like to thank Marina Shalaeva, Dr. Veerabahu Shanmugasundaram and Dr. Laurence Philippe for making my stay in Pfizer Inc., Groton global research site such an incredible experience. Prof. Dr. Willie Peijnenburg for guiding me through my internship in Leiden.

I am grateful to Dr. Igor Tetko for selecting me to join the ECO-ITN program and for his help. The discussions we had in his research group helped shape my current understanding of QSAR. I extend my gratitude to many research colleagues for their continuous discussions, advices and support: Yurii Sushko, Anil Pandey, Robert Körner, Sergii Novotarskyi, Stefan Brandmaier, Wolfram Teetz and Eva Schlosser as well as many other scientists who briefly visited our group. Many thanks to Prof. Dr. Hans-Werner Mewes and Prof. Dr. Michael Sattler for hosting us in the institutes of Bioinformatics and Systems Biology (IBIS) and the institute of Structural Biology (STB) at the HelmholtzZentrum München (HMGU). The interactions with excellent scientific community through lectures and seminars broadened the perspective and applications of my research.

Many thanks go to the TUM Graduate School for granting me a partial scholarship, which permitted me to pursue a double degree and complete my MBA and doctoral studies simultaneously.

I would like to thank my parents and lovely sister for their belief in me reaching my goals despite any challenges along the way. They infused me with confidence that I needed to reach the closing.

Special thanks are due to my wife, Amira. Thank you for all you have done and for being always reliable in managing our small family and raising our lovely daughter Laila at times where I was too busy to contribute but a little. I am truthfully grateful.

The research leading to these results has received partial funding from the European Community's Seventh Framework Programme [FP7/2007-2013] under Grant Agreement n. 238701 of Marie Curie ITN Environmental Chemoinformatics (ECO) project.

Ahmed Mohamed Abdelaziz Sayed

December 2015

Zusammenfassung

REACH und prädiktive Risikobewertung:

Quantitative Struktur-Wirkungs-Beziehungen (QSAR: Quantitative Structure-Activity Relationships) haben in der prädiktiven Toxikologie in den letzten Jahren deutlich an Relevanz und an Popularität gewonnen. Insbesondere der Fortschritt bei Algorithmen des ‚Machine Learning' und die abnehmenden Kosten bei Rechenressourcen ermöglichen die Analyse von großen Datenmengen, wie sie z.B. beim High-Throughput-Screening (HTS; i.e., *in vitro*-Testmethoden mit angemessenem prädiktivem Wert für *in vivo*) anfallen. Inzwischen fordert die REACH-Verordnung (im Jahr 2007 in der EU eingeführt) eine gestaffelte Registrierung aller auf dem Markt befindlichen Chemikalien und eine Behebung von Informationslücken. Da die Übergangsfristen im Jahre 2018 ablaufen und dann auch in geringerem Ausmaß produzierte chemische Verbindungen betroffen sind, auf der anderen Seite aus ethischen, wirtschaftlichen und praktischen Gründen für betroffene Chemikalien flächendeckende Tierversuche nicht angemessen sind, wurden multizentrisch prädiktive QSAR-Modelle auf Basis von standardisierten Daten aus HTS-Toxizitätsscreenings evaluiert.

Entwicklung von QSAR-Modellen auf der Basis von HTS-Toxizitätsdaten:

Die vorliegende Arbeit beinhaltet eine Verarbeitung von Daten aus HTS-Assays im prädiktiven QSAR-Kontext, d.h. die Verwendung von großen HTS-Datenmengen als biologisch basierte Deskriptoren für die QSAR-Modellierung und Prädiktion von *in vivo*-Toxizität. Heute stehen HTS-Toxizitätsassays zur Verfügung, mit denen es möglich ist, Toxizitätsmechanismen (basierend auf biochemischen Stoffwechselwegen, der Interaktion mit nuklearen Rezeptoren oder der Bindung an Proteine) für eine große Anzahl von Verbindungen zu evaluieren und demzufolge große Datenmengen verarbeiten bzw. generieren zu können, um Erkenntnisse zu Toxizitätsmechanismen von chemischen Verbindungen zu erhalten und somit die zugrunde liegenden, gestörten biochemischen Prozesse [‚Adverse Outcome Pathways (AOP)] zu verstehen. Die entsprechenden verfügbaren HTS-Datenreihen wurden als „Training sets" für *in silico*-basierte QSAR-Modelle verwendet, die nach den OECD-Grundsätzen für QSAR-Modellerstellung konzipiert und optimiert wurden. Die resultierenden (priorisierenden) Modelle haben einen hohen und validen prädiktiven Wert und ermöglichen daher eine erhebliche Aufwands- und Kostenminimierung in Zusammenhang mit erforderlichen Toxizitätscharakterisierungen von chemischen Verbindungen.

Prädiktiver Wert für das biologische System:

Eine Prädiktion von präklinischen *in vivo*-Toxizitäten auf Basis von *in silico*-Deskriptoren für komplexe Endpunkte erscheint nur sinnvoll, wenn es sich um Substanzbibliotheken mit einheitlichen biochemischen, biologischen bzw. Wirkparametern handelt. Beispielsweise ergab sich in Bezug auf die Acetylcholinesterase-Hemmung durch Organophosphor-Verbindungen eine Vorhersage-Genauigkeit von über 90 %. Für andere Endpunkte (z.B. Reproduktions- und Entwicklungstoxizität bei der Ratte) ergaben sich ca. 70 %. Die Kombination einer *in vitro*-HTS-Profilierung von chemischen Verbindungen mit den entsprechenden *in silico*-Deskriptoren führte bei einigen Endpunkten (z.B. Embryotoxizität bei

Ratten) zu einer - gegenüber der Toxizitätsprädiktion ausschliesslich auf Basis der *in silico*-Deskriptoren - signifikanten Verbesserung der Vorhersagegenauigkeit von QSAR-Modellen (p-Werte <0,05). Durch die mechanistische Klassifizierung und Neukategorisierung von Daten aus *in vitro*-HTS-Assays entsprechend ihrem molekularen Toxizitätsmechanismus konnte zudem die Vorhersage für einige *in vivo*-Toxizitätsendpunkte (z.B. Bildung von chronischen neoplastischen hepatischen Läsionen und generationsübergreifende Lebensfähigkeit bei Ratten *in vivo*) zusätzlich deutlich verbessert werden ($p < 0,05$). Die Analyse *von in vitro*-HTS-Daten erwies sich somit als sinnvoll beim Nachweis der Relevanz molekularer Mechanismen, die mit der *in vivo*-Toxizität korrelieren. Toxizitätsdeterminierende biochemische Prozesse können dadurch identifiziert werden können.

Eine Vorhersage des Toxizitätspotentials über *in silico*-Deskriptoren auf Basis der Daten aus den *in vitro*-Assays zu Substanzinteraktionen mit multiplen nuklearen Rezeptoren sowie biochemischen (Toxizitäts-)Mechanismen aufgrund von im entsprechenden System hervorgerufenen Stressreaktionen (Daten aus den ToxCast- und Tox21-Projekten) zeigten für eine Reihe von inkludierten, als relevant erachteten HTS-Endpunkten und Surrogat-Parametern exzellente Ergebnisse mit Vorhersagegenauigkeiten (‚balanced accuracies') von über 80 %. Hierbei sind insbesondere die AHR-Aktivierung (mit 86 %), die Beeinträchtigung der Funktion der Mitochondrien Membran (mit 88 %) und die Androgen-Rezeptor-Aktivierung (mit 82 %) zu nennen. Bei der Validierung des „bagging" zeichnet sich deutlich die gute Anwendbarkeit der entwickelten Modelle auf externe Validierungsdatensätze ab. Zudem ergab sich im Consensus-Modeling-Prozess eine zusätzliche Verbesserung der Prädiktionsgenauigkeit, die sowohl bei Validierungs- als auch bei Test-Datensätzen evident war.

<u>Anwendungen und Testpriorisierung (‚ToPS):</u>

Um Einsetzbarkeit und Relevanz der entwickelten Verfahren zu prüfen, wurde in zwei Fällen *in silico* eine Umwelt-Risiko-Abschätzung durchgeführt und die Ergebnisse diskutiert.

(1) Im ersten Projekt wurde der umfangreiche Datensatz der im EINECS-Verzeichnis (=Altstoffverzeichnis) enthaltenen Verbindungen im entwickelten Ansatz auf mögliche biochemische Störeffekte untersucht. Die Evaluierungen zeigen, dass mit hoher Wahrscheinlichkeit ein bestimmter Prozentsatz der chemischen Verbindungen (zwischen 4,6 und 12,6 %, je nach biologischem Endpunkt) molekulare Signalwege negativ beeinflusst. In diesem Zusammenhang wurde zudem ein vereinfachendes bzw. zusammenfassendes priorisierendes Punktesystem vorgeschlagen [„toxicity-testing priority score (ToPS)"], auf Basis dessen unter Miteinbeziehung aller unter Verwendung des HTS-Daten für die verschiedenen Toxizitätsmechanismen etablierten Modelle eine Beurteilung des Gesamt-Risikoprofils einer chemischen Verbindung möglich ist.

(2) Die zweite Anwendung untersucht eine Reihe von halogenierten Carbazolen, die sich in Europa und den Vereinigten Staaten als ökotoxikologisch relevant erwiesen haben, interessanterweise (für den Fall Europa) ohne dort produziert oder dorthin importiert zu werden. Bei der Analyse von HTS-Daten, im Rahmen derer die Aktivierung des Aryl-Hydrocarbon-Rezeptors (AHR) evaluiert worden war, wurde eine sehr hohe Korrelation des

Vorkommens des Carbazolyl-Strukturelements mit einer AHR-Aktivierung gefunden (p-Wert: 3 x 10^{-25}), zudem ergab sich bei den experimentellen Untersuchungen ein hoher Anreicherungsfaktor (> 6-fach). Allgemein führt das Vorkommen bestimmter Substituenten am Carbazol-Gerüst (wie z.B. das Vorkommen von aromatischen Aminen) - mit hoher Wahrscheinlichkeit - zu einer AHR-Aktivierung (p-Wert: im Bereich von 10^{-5} bis 10^{-7}). Für alkoholische und phenolische Substituenten wurde eine Tendenz zu AHR-Inaktivität bzw. geringer AHR-Aktivität gefunden (p-Wert: im Bereich von 10^{-5} bis 10^{-6}). Die Ergebnisse, die zu den halogenierten Carbazolen erhalten wurden, zeigen – bei hoher Vorhersage-Präzision – ein hohes Toxizitätspotential in Bezug auf alle im Rahmen der *in vitro*-HTS-Toxizitätsassays untersuchten und in die vorliegende Arbeit miteinbezogenen Endpunkte.

Mögliche Nutzung der Modelle:

Die in dieser Arbeit entwickelten QSAR-Modelle wurden im Rahmen von Wettbewerben, die von den National Institutes of Health (NIH) sowie der Environmental Protection Agency (EPA) organisiert waren, mehrfach ausgezeichnet. Alle verwendeten Datensätze und entwickelten QSAR-Modelle sind öffentlich zugänglich und im wissenschaftlichen und regulatorischen Bereich allgemein nutzbar. Es wurde für die Modellbildung die Plattform *iPrior* eingesetzt, in der Daten aus den ToxCast-, den Tox21- und den e1K-Projekten zu finden sind. Auch die im Rahmen der vorliegenden Arbeit auf Basis der Tox21-Daten entwickelten Modelle sind öffentlich zugänglich unter http://amaziz.com/article/tox21. Dadurch sind sie für andere Wissenschaftler im Rahmen von prospektiven und retrospektiven Untersuchungen uneingeschränkt nutzbar. Zudem sind die Ergebnisse der verschiedenen Analysen und Applikationen in einem GitHub-Verzeichnis zugänglich (https://amaziz.com/dissertation/supplementary).

Es kann davon ausgegangen werden, dass die in der vorliegenden Arbeit entwickelten und von Regulierungs-/Zulassungsbehörden und der Wissenschaft akzeptierten Modelle für Toxizitätsvorhersagen und für eine Reduktion der Anzahl von erforderlichen Tierversuchen eine wichtige Rolle spielen können.

Abstract

Quantitative structure activity relationships (QSARs) have been gaining popularity in predictive toxicology. The advancement in machine learning algorithms and the diminishing costs of computational resources allow the analysis of large datasets resulting from high throughput screening (HTS). On the other hand, the REACH regulations introduced in the European Union calls for a phased registration of all compounds used and filling information gaps related to such chemicals. Ethical, economic and practical reasons edicts that such information gaps must not be filled by animal experiments except as a last resort. Alternative testing approaches, including QSAR, are therefore called to action.

This thesis focuses on the utilization of HTS assays in the QSAR context. The use of large collection of HTS assays as biologically derived descriptors for modeling *in vivo* toxicity outcomes was investigated. Furthermore, HTS assays focusing on a specific biochemical pathway, nuclear receptor or protein binding and applied on a large dataset of compounds can give insights on the mode of adverse action of chemicals and contribute to the understanding of adverse outcome pathways related to chemicals. These HTS assays were used to train *in silico*–based QSAR models according to the OECD principles for QSAR model building. The current approach and the models showed high prediction accuracy and, therefore, can reduce the cost, analysis time and allow the *in silico* screening of larger compound datasets.

Prediction of preclinical *in vivo* animal toxicity using *in silico* descriptors for complex end points was only feasible for restricted compound libraries with the same mode of action. For instance, predictive balanced accuracy for organophosphorus compounds' inhibition of acetylcholine esterase exceeded 90%. Other endpoints such as developmental rat maternal toxicity reached 70%. Combining data derived from HTS *in vitro* profiling of chemicals, with *in silico* descriptors showed a significant improvement in the predictive ability of QSAR models for some endpoints **(p-values <0.05)** (such as rat fetal pathology). Furthermore, the mechanistic classification and regrouping of the HTS *in vitro* assay responses in the form of pathway perturbations significantly improved **(with p <0.05)** the predictivity for other *in vivo* toxicity endpoints such as chronic rat liver neoplastic lesions development and multigenerational rat viability among others. Furthermore, analysis of *in vitro* HTS proved useful in detecting molecular pathways that are most correlated to *in vivo* toxicity outcomes and therefore could assist in understanding the underlying mechanism of toxicity and the essential biochemical pathways involved.

Prediction of *in vitro* assays outcomes of multiple nuclear receptors and stress response pathways relevant to toxicological responses (from ToxCast and Tox21 projects) using *in silico* descriptors showed high success with balanced accuracies reaching up to more than 80% for several endpoints. This includes endpoints such as aryl hydrocarbon receptor activation (86%), mitochondrial membrane disruption (88%) and androgen receptor activation (82%). Bagging validation provided a good indication for the models' predictive ability on external validation sets. Furthermore, Consensus modeling improved the predictive ability of QSAR models as signified by both validation and evaluation set accuracies.

Finally, two specific applications for environmental risk assessment were computed and discussed. The first application screens the large dataset of EINECS compounds for potential pathway perturbations. The predictions show, with high confidence, that a certain percentage of chemicals (between 4.6% and 12.6% depending on the target) are likely to disrupt molecular pathways. Furthermore, a point-based system was suggested: *toxicity-testing priority score (ToPS) to* provide a universal overview of a compound's molecular pathways perturbation considering models' applicability domain and assesses chemicals' overall risk profile.

The second application investigates a set of halogenated carbazole compounds emerging in the European and US ecology without being actively produced or imported. Analyzing HTS data showed that the presence of carbazolyl moiety highly correlates with Aryl Hydrocarbon Receptor (AHR) activation (p-value: 3×10^{-25}). The carbazolyl moiety provides high enrichment factor (> 6-fold) for AHR activation. Certain carbazolyl substitutions (such as aromatic amines) are more likely to lead to AHR activation (p-value: 10^{-5} to 10^{-7}) while alcohols and phenols were more likely to be associated with AHR inactive compounds (p-value: 10^{-5} to 10^{-6}). Prediction of halogenated carbazoles' pathways perturbation shows, with high prediction accuracy, an activity against most pathways.

QSAR models developed in this thesis were recognized by winning multiple awards in challenges organized by the National Institute of Health (NIH) as well as the environmental protection agency (EPA). The outcomes of the dissertation are made available to regulators and the scientific community. The public platform iPrior was deployed and is hosting data from ToxCast, Tox21, and e1K projects. Moreover, the developed models based on the Tox21 study are made publicly available at http://amaziz.com/article/tox21, thus allowing other researchers to use them for prospective and retrospective analyses. Finally, the results of different analyses and applications are made available in an open GitHub repository (https://amaziz.com/dissertation/supplementary). It is hypothesized that those developed and freely accessible models may become accepted by the regulators and the scientific community and therefore play a significant role in predicting *in vivo* toxicity and reduce animal toxicity testing.

Preface - Chemicals and biological systems

Systems biology involves the mathematical and computational modeling of complex biological systems such as molecules, cells, or organisms and up to entire species. As such living systems are dynamic, it may be hard to predict their behavior from the properties of their individual parts. Therefore, systems biology brings interdisciplinary methodologies from fields such as engineering and mathematics to study the complex interactions within biological systems using a holistic approach.

The continuous advancement in high throughput screening (HTS) techniques and mapping of the human genome has promoted the tracking of a biological system's exposure to potential stressors (e.g. chemicals) during its lifetime. The concept of "exposome" –generally– studies how such exposure to a chemical (from diet, environment, occupation, or lifestyle) may impact a biological system (human, animal or an environment). The overall exposure integrates particular properties of a chemical that determine the uptake into an organism as well as its disposition (toxicokinetics) and the toxicity mechanisms.

To better understand how exposures may affect an organism, this current work utilizes HTS data to investigate selected molecular pathways of interest for toxicity estimation. With the aid of machine-learning algorithms, the interaction between small molecules and genetics may be revealed out of *in vitro* data pools. The utilization of state-of-the-art machine learning in combination with HTS *in vitro* assays is an alternative approach for chemicals' risk assessment, it includes biologically relevant HTS data, spares animal studies, and may prioritize - yet toxicologically undefined - chemicals for more thorough investigations. This approach is encouraged by the REACH (**R**egistration, **E**valuation, **A**uthorization and restriction of **CH**emicals) legislations.

Outline

The dissertation is structured in four parts. In the first part, a general introduction about the chemical regulations in Europe, the interaction of xenobiotics with biological systems, and the steps of quantitative structure activity relationship (QSAR) are introduced. The development of the "receptor" history is introduced and relevant targets for toxicity estimations using HTS are reviewed. Moreover, the role of computational toxicology and *in vitro* assays in alternate testing is discussed.

In the second part, different algorithms, related methods and software tools are introduced. Classical and modern machine learning algorithms and molecular descriptors used in building QSAR models are explained. For such models, the techniques used for variable selection, assessment of goodness of fit and prediction, model comparison and applicability domain estimation are elucidated.

The third part of the dissertation shows, how the previously defined methods have been used to build and validate QSAR models, which were constructed to directly predict *in vivo* toxicity endpoints as well as predict selected *in vitro* outcomes related to adverse outcome pathways.

The last part of the dissertation shows some potential applications for using the developed QSAR models in environmental risk assessment, quantifying the potential hazards of EINECS chemicals and prioritizing toxicity testing. The potential toxicity of a particular class of halogenated carbazoles was also examined.

Table of contents

1. Introduction

1.1 Chemicals regulations in the European Union

1.1.1 Chemicals legislations in Europe

The European Union (EU) started controlling chemicals use and marketing since 1976. The process was slow with only 100 substances restricted from use and marketing and another 900 substances restricted from marketing to the general public due to **C**arcinogenicity, **M**utagenicity or toxicity to **R**eproduction (**CMR** substances)[1].

A distinction was made under *Regulation (EEC) No 793/93*[2], between "existing" and "new" chemicals. It was based on a cut-off date of 1981. Chemicals already recorded in the European Inventory of Existing Commercial Chemical Substances (EINECS), between 1st of January 1971 and 18th of September 1981, were labeled as "existing". This accounted for an excess of 100,000 substances. "New" chemicals are those (more than 3800) introduced to the European Community market after 1981[1].

Different regulations governed the two categories. "New" chemicals must be tested before being introduced to the market; no similar obligation exists for "existing" chemicals. "Existing" chemicals were left with insufficient public information on their properties and uses. It was therefore difficult to effectively evaluate and regulate these substances[1].

Furthermore, the responsibilities allocation was not appropriate. Risk assessment of substances was the responsibility of public authorities and not the enterprises that place them into the market through manufacturing, import, or use. The requirements for such assessments were comprehensive rather than use-specific. This placed huge load on the authorities. Only 141 high-volume chemicals have been identified as priority substances for risk assessment from the period 1993-2007. Risk reduction recommendations were only issued for a narrow subset of such chemicals which completed the entire evaluation procedure according to *Regulation (EEC) 793/93*. This can also be related to the fact that only substances' manufacturers and importers were required to provide information. While the regulations ignored downstream industrial users and formulators, unless the substance was classified and a safety data sheet had to be supplemented along the supply chain. This lead to a scarcity in information about substance use and exposure[1].

On the contrary, "new" chemicals, per *Regulation (EEC) 793/93*, were notified and tested starting from very low annual volumes of 10 kg. Due to the bureaucratic process and testing burden, the EU chemical industry was unenthusiastic about research and innovation of new chemicals. The industry heavily favored the development of "existing" substances, for example, using mixtures which present a challenge on its own for regulators.

From the first of June 2007, a new European Community regulation on chemical substances, REACH (Registration, Evaluation, Authorization and Restriction of Chemicals) was introduced and came into effect[3]. REACH aims to protect humans, animals and the environment while enhancing the competitiveness of the chemical industry in the EU. The most hazardous chemicals will be progressively substituted as soon as suitable alternatives are found. Through

increased transparency, the regulations aim to prevent internal market fragmentation. The legislation gradually assesses the potential risks caused by chemical substances in a tiered process. REACH also aims to reduce animal testing through the promotion of alternative risk assessment methods[3].

The Industry, rather that the authorities, now needs to ensure that chemicals it puts on the EU market are not harmful for human health or the environment. For that, knowledge about certain properties of the substances must be known and certain risks should be assessed. Authorities can now better use available resources for making sure those industry players are attaining to their responsibilities and acting on substances of very high concern.

REACH unified all chemicals, "existing" and "new". Substances that were not marketed or produced before the enforcement of REACH regulations are referred to as 'non-phase-in' substances. On the other hand, substances that were listed in EINECS or were manufactured in the EC but were not located on the Community market for the last 15 years or the compounds that listed as 'no longer polymers' of *Directive 67/548* are referred to as 'phase-in' substances[1].

There are big benefits to be expected from the REACH implementation. The occupational and public health influence of REACH is expected to reduce chemical-related diseases such as respiratory and bladder malignancies, mesothelioma, as well as skin, eye and respiratory disorders, among others[1]. The benefits will increase as more information is being gathered to better implement the legislations and controls. The authorization requirements for substances of very high concern and the faster restrictions will also contribute to a better human health and environment.

The extended impact assessment of the Commission calculated the public health benefits, for implementing REACH, based on a World Bank[4] evaluation and a number of careful assumptions. With the assumption that chemically-linked diseases are responsible for around 1% of the total disease burden in the European Union and consequently, that a 10% reduction in such diseases as a result of REACH would lead to a 0.1% decrease in the total drain of disease in the EU. This would be corresponding to about 4,500 cancer-deaths avoided every year. Accounting for a €1 million value of life, the potential health profits of REACH were estimated to be about €50 billion over a period of 30 years.

A study contracted by EC Directorate-General for the Environment[5] investigated the returns of REACH implementation due to the decrease in release of compounds in the environment and the exposure of humans through the environment. The study inspected many cases utilizing different appraisal methods showing that the long-term benefits of REACH is expected to be significant.

REACH is expected to contribute to reduced air, soil, and water pollution and to decrease stress on biodiversity. REACH will also assist in reducing endocrine disrupting chemicals effect. The information required for the safe handling of chemicals would be recorded in the central databank managed by the European Chemicals Agency (ECHA, Helsinki)[6].

1.1.2 The European chemical agency (ECHA)

ECHA is the driver among regulatory authorities in implementing the EU's chemicals legislation. It helps companies to fulfill their responsibilities in the legislation, encourages the safe use of chemicals and offers information on and addresses concerns about chemicals. ECHA also coordinates with the European Commission and the EU member states for safeguarding the wellbeing of human health and the environment.

Through its role, ECHA observes innovation and competitiveness. The organization is independent from external interests and is impartial in its decision-making. This builds credibility among all stakeholders through qualified handling of technical, scientific and administrative aspects of the regulation. In its capacity, ECHA also manages the chemical registration process, evaluates submitted dossiers, takes decisions about suspicious chemicals and runs the databases of available hazard information thus linking consumers and experts.

ECHA identifies needs for regulatory risk management on EU-wide level. ECHA can initiate the identification of substances of very high concern and restrictions and manages the applications for authorization by the industry. ECHA also manages the process for harmonized classification and labelling of substances.

ECHA also facilitates the information exchange between companies that plan to register the same phase-in chemical. Such companies are obliged to join a Substance Information Exchange Forum (SIEF) to share data on the basic properties of the chemical and to avoid duplication of studies. Specifically, companies are required to share all test data on vertebrate animals. SIFE leads to one joint submission for each substance, therefore lowers costs and eliminates unnecessary animal testing.

1.1.3 The regulatory process for chemicals risk assessment in REACH

The process starts with a registration application by manufacturers and/or importers who are required to acquire relevant information on their substances and to use such data for safe handling. All data concerning studies on vertebrate animal testing is mandatorily shared. For other tests, data sharing is required on request. The downstream users and formulators are brought along. They receive information on hazards and risk management through the supply chain.

For a product to be allowed into the EU market, a registration dossier should be submitted to ECHA for each substance manufactured or imported in quantities of one ton or above per year. Figure 1 shows the deadline for registration of substances per their production volume. The tonnage-band also controls the information that needs to be reported for each substance. Figure 2 shows the annexes that describe the information needs relevant to each production volume. Dispersively used substances, in quantities ranging between one to ten tons, that are potentially hazardous to the human health or the environment receive priority. Those hazardous compounds are chemicals listed in CMR categories 1 or 2, as well as persistent, bioaccumulative and toxic (PBT) and very persistent and very bioaccumulative (vPvB) substances.

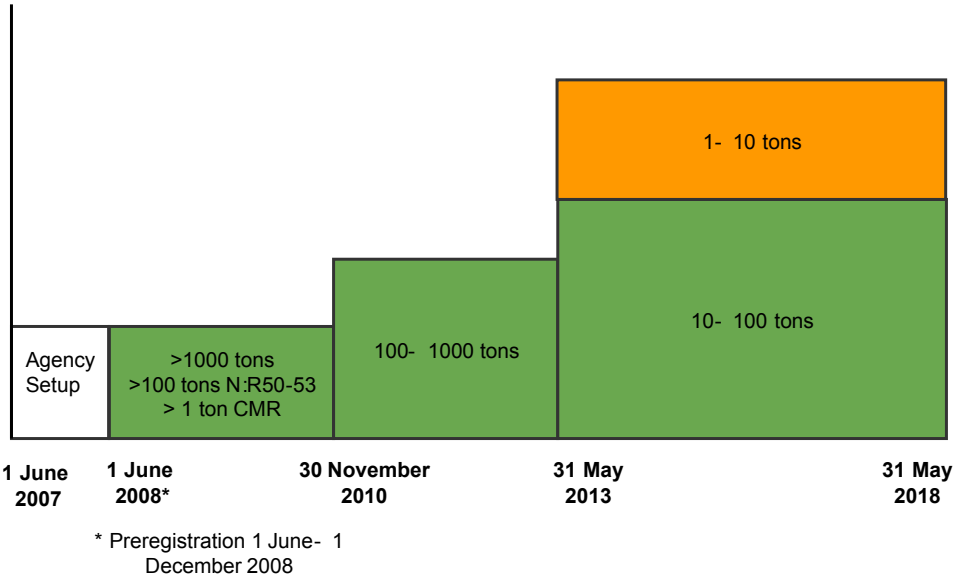


Figure 1. Registration deadline per substance production volume. The European chemical agency was established in June 2007 and started accepting registration dossiers in June 2008. The first band is for chemicals with production volumes above one thousand tons per year, or which may be toxic to the aquatic environment or may cause long term effects (N: R50-53) with production volumes above 100 tons/year or chemicals that are categorized as carcinogenic, mutagenic, or toxic for reproduction (CMR) with an annual production volume above one ton. Such chemicals had to be registered before 30th of November 2010. Chemicals with annual production volumes between 100 and 1000 tons had a registration deadline until the 31st of May 2013 while those chemicals of lower production volumes must be registered until the 31st of May 2018.

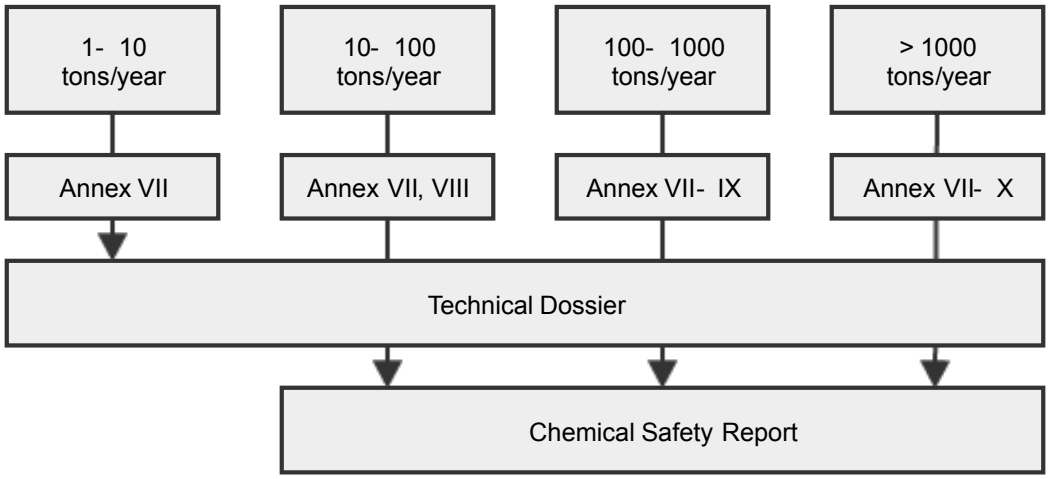


Figure 2. Minimum data requirements for chemicals registration according to REACH. The legislation requires information on the intrinsic properties of chemicals submitted in a technical report. If the chemicals are manufactured or imported with an annual volume above 10 tons/year, a chemical safety report must also be submitted. Such report explains the potential hazards of the substance (e.g., PBT or vPvB) and explains the potential exposure scenarios for the given uses. Information requirements vary based on the tonnage band of the chemicals. These information requirements are listed in annexes VII to X of the REACH legislation. In all cases, registrants are required to collect all available information available to them on the chemicals they are registering regardless on the necessity of the information based on the production volume. This includes any relevant information about physicochemical, toxicological or ecotoxicological endpoints. The registrants must have permission to use the data in order to utilize it for their dossiers. Additional testing may be needed to meet the minimum information requirements. However, per Article 13 of REACH), the use of alternative testing and the exhaustion of other options must be considered.

For filling the information gaps in dossiers, REACH only allows new experimental testing if no alternatives were available. The use of existing information or alternative techniques such as *in vitro* assays, quantitative structure-property relationships (QSPRs), and read-across is encouraged.

The evaluation process is done by ECHA. The agency evaluates proposals for testing by SIFE to check compliance with the registration requirements. It also synchronizes substance assessment by authorities to inspect substances with perceived risks. This evaluation can be used to plan restrictions or authorization proposals.

Substances showing properties of very high concern requires authorisation; ECHA issues a list of these candidate substances. Industry applicants should prove that risks connected with the intended uses of these substances are being sufficiently controlled or that the socioeconomic profits of their use would overshadow such risks. Applicants must also investigate safer appropriate substitute substances or technologies. In case such substituents exist, applicants must present a plan for substitution. In case they do not exist, applicants should, if appropriate, offer evidence for research and development efforts to create such substitutions.

Finally, restriction (**R**EACH) is a necessary control for implementing a framework through which manufacturing, use or market introduction of certain unsafe chemical substances could be prohibited or subjected to special provisions. The labelling and classification of hazardous substances encourages the industry to agree on substances classification. It is also possible (e.g., in case of substances of high concern) that a community-wide classification harmonization is enforced by the authorities.

1.2 Interaction of xenobiotics with biological systems – Quantitative Structure Activity/Property Relationship (QSAR/QSPR/QPPR)

In this section, the history of the QSAR field is introduced, the applications and challenges facing QSAR/QSPR as well as the opportunities that arise due to advances in robotics and "big data" implementation in healthcare where the cost of running large-scale High Throughput Screening (HTS) *in vitro experiments* is reduced. Finally, the author explains the motivation behind this work and briefly highlights what was accomplished in the study.

1.2.1 History

Quantitative structure–activity relationships were slowly developed over a period of over hundred years. Multiple associations between the toxicity and narcotic activities of organic compounds and their lipophilicity, expressed in oil/water partition coefficients[7,8], were noticed. In 1868, Crum Brown and Fraser spotted a significant change in pharmacological activities of some organic bases due to the quaternation of the basic nitrogen atom. They described the dependence of "physiological properties" $\emptyset$ on chemical structures C[9] as:

$$\boldsymbol{\phi} = \boldsymbol{f}(\boldsymbol{C})$$ Equation 1

Or

$$\boldsymbol{\Delta\phi} = \boldsymbol{f}(\boldsymbol{\Delta C})$$ Equation 2

While the first equation describes the mathematical basis of the QSAR field, the second provides the practical base for its implementation. In contrast to biological activities which can be well-defined (e.g., a growth inhibition concentration (IGC50), half maximal inhibitory concentration (IC50), or median lethal dose, i.e., kills 50% of test population, (LD50)) chemical structures are not similarly so. It is impossible to exactly describe a chemical structure. Therefore, the second equation is necessary, where only the changes of biological activities and chemical structures are being correlated. These chemical changes can be quantified either on structural basis (such as indicator or dummy variables or Free-Wilson parameters analysis) or through the change in physicochemical or other properties.

Many years later, in 1893, the inverse correlation between simple organic chemicals' solubility and their cytotoxicity was discovered by Richet[10]. Meanwhile, Meyer and Overton developed independent theories of narcosis as being related to partitioning between olive oil and water phases. Such theories contributed to a better understanding of the relation between lipophilicity and narcotic and toxic activity[11]. This work is often (though not correctly) denoted as the historical beginning for the correlation between physicochemical properties and biological effects. Due to this relation, measured lipophilicity may be used directly to estimate biological properties of compounds[12]. This approach can be referred to as quantitative property-property relationship (QPPR) because the chemical structures were not explicitly needed to construct the relationship.

Ferguson, in 1939, delivered a thermodynamic interpretation for the non-specific lipophilicity-activity relationships. He offered a simplification for the relation between the relative saturation of volatile compounds, in their administration vehicle, and their observed depressant activity. This also rationalized the activities' "cutoff" after a certain lipophilicity optimum[13].

Hammett, derived Equation 3 and Equation 4 describing equilibrium constants (K^e) and rate constants (K^r) for various aromatic reactions using a reaction-dependent constant (ρ) and a substituent parameter (σ) which depends only on the nature of the substituent (X) of the corresponding aromatic compounds, using hydrogen as a reference substituent; (ρ) values are based on the ionization constants of substituted benzoic acid[14]. Hammett published his breakthrough as the Physical Organic Chemistry[15] in 1940 showing that the effects of substituents could be quantified and giving rise to the "sigma-rho" culture. In the subsequent years, different σ scales were needed for various systems leading to the proliferation of substituent constant scales.

$$\log \boldsymbol{K}^{\boldsymbol{e}}_{\boldsymbol{R-X}} - \log \boldsymbol{K}^{\boldsymbol{e}}_{\boldsymbol{R-H}} = \boldsymbol{\rho\sigma}$$ Equation 3

$$\log K^{r}_{R-X} - \log K^{r}_{R-H} = \rho\sigma$$ Equation 4

In 1962, Hansen derived the first Hammett-type relationship between the toxicities of substituted benzoic acids and the electronic σ constants of their substituents, thus for the first time, applying the Hammett approach to a biological property[16]. It was later discovered that this relationship was only a chance-correlation due to the interrelation between Hammett (σ) parameter and the lipophilicity constant (π).

On parallel, "Hansch and Muir discovered the SAR of growth regulators in plants and their reliance on hydrophobicity and Hammett constants[17]. They used regression analysis and descriptors for the hydrophobic, electronic and steric properties of molecules to present the first nonlinear multi-parameter equation (Equation 5) describing biological activity values:

$$\log \frac{1}{C} = -2.14\pi^2 + 4.08\pi + 2.78\sigma + 3.36$$ Equation 5

After such developments, the field of QSAR commenced to get its modern shape through two independent publications. Hansch and Fujita[18] as well as Free and Wilson[19] described two new approaches for quantitative structure–activity relationships. The approaches were later referred to as "Hansch analysis" (linear free energy-related approach, extrathermodynamic approach) and "Free–Wilson analysis". By combining different physicochemical parameters in a linear additive manner, both approaches presented a breakthrough in the QSAR arena.

While Equation 5 represent typical Hansch model, the Free–Wilson model can be expressed by Equation 6, where (a_{ij}) is the group contribution of a substituent (X_i) in the position (j) and (μ) is the measured or calculated biological activity value of a reference compound within the series. All group contributions (a_{ij}) for different substituents (X_i) refer to the substitution of corresponding substituents (usually hydrogen) in the reference compound:

$$\log \frac{1}{C} = \sum a_{ij} + \mu$$ Equation 6

After a while, linear equations became insufficient for cases with extended hydrophobicity ranges. This led to the development of the Hansch parabolic equation for describing nonlinear lipophilicity–activity relationships[20]. An example can be seen in Equation 7.

$$\log \frac{1}{C} = a(logP)^2 + b\, logP + c\,\sigma + \cdots + const.$$ Equation 7

The definition of these models led to the quick development in QSAR analysis and related methods. Kubinyi combined Hansch equations with indicator variables[21] (Equation 8), which may be considered as a mixed Hansch/Free–Wilson model leading to further improvements . He also formulated the bilinear model[21] as a theoretical non-linear model for describing the transport and distribution of drugs in biological systems (Equation 9). This model is a refinement of the parabolic model and has been superior in many cases[22].

$$\log\frac{1}{C} = a(logP)^2 + b\,logP + c\,\sigma + \cdots + \sum a_{ij} + const.$$ Equation 8

$$\log\frac{1}{C} = a(logP) + b.\log(\beta P + 1) + c\,\sigma + d.MR + \cdots + const.$$ Equation 9

Both Hansch and Free-Wilson approaches can be considered two-dimensional QSAR (2D-QSAR) approaches. They do not depend of the three-dimensional (3D) structures of chemicals. Thus, all conformations of a chemical structure have identical predicted activity. In 2D-QSAR, the hypothesis is that compound properties are totally determined by knowledge of its topology (2D-structure).

The 3D-QSAR emerged to extend these approaches. Knowing that conformations of a single compound do not necessarily exhibit equivalent biological behavior, the basic hypothesis of 3D-QSAR is that macroscopic properties (e.g., biological or physicochemical) of substances are determined by the spatial arrangements (conformations) of a given molecular structure. 3D-QSAR models relate computed atom-based properties (e.g., steric and electrostatic potentials) to target macroscopic properties, and assuming that changes in these spatial arrangements (and structures) lead to altered properties.

Many 3D-QSAR approaches have been described. In general, they share many common steps: starting with the determination of the biologically active conformation for the molecules, aligning such molecules, computing some molecular properties and correlating these properties to the activity at hand. Among the most famous approaches are Comparative Molecular Field Analysis (CoMFA), described by Cramer et al. in 1988[23], Comparative Molecular Similarity Indices (CoMSIA) method[24], the Voronoi Field Analysis (VFA) developed by Chuman et al[25] and Comparative Molecular Moment Analysis (CoMMA)[26] among many others[27]. Most 3D-QSAR methods rely on the linear free-energy formalism, like traditional 2D-QSAR approaches.

The number of descriptors generated in 3D-QSAR is much larger than in the classical 2D-approaches. This may result in collinearities developing across these descriptors and thus resulting in chance correlations when multiple linear regression analysis is used. Such collinearities may be dealt with using dimension reduction approaches like PLS (see 3.5 Machine learning algorithms).

In 1997, Hopfinger et al.[28,29] described 3D-QSAR models with the 4D-QSAR analysis formalism. Such formalism uses ensemble averaging to allow both conformational flexibility and freedom of alignment. In this setup, the sampling of conformations ensemble is considered a “fourth” dimension. 4D-QSAR uses a grid to select the binding regions in 3D space. Then, the compounds are partitioned into atom/region types based on seven interaction pharmacophore elements (IPEs). Conformational ensemble profiles (CEPs) are then constructed using molecular dynamics simulations. Next, the Grid Cell Occupancy Descriptors (GCODs) are calculated based on the selected IPEs for each conformation in the CEP. Every conformation of a specific compound is aligned in the reference grid and the frequency of a specific IPE in a particular grid cell is counted. The grid occupancy data are reduced by a Partial

Least Squares (PLS) regression analysis. Finally, a model is constructed between the remaining GCODs and the target biological activity.

Angelo Vedani et al. presented the concept of Quasar to introduce the proxy (atomistic) receptor model and hence the 5D-QSAR[30]. This approach addresses the inherent flexibility of the receptor protein that has often been largely overlooked in previous QSAR approaches. When ligand binds to the receptor it transforms to a lower energy conformation state distorting the receptor structure in the process. The result is the formation of a binding complex as the receptor engulfs the ligand. Two receptor states are described, the unbound (Apo) and the bound (Holo). The receptor surface model gives an idea about a hypothetical binding site. The multiple representations of ligand topology to study conformation, isosteriomer, protonation and orientation are referred to as the new dimensions of 4D- QSAR. By representing multiple induced fits, allowing multiple representations of the topology of the quasi-atomistic receptor surrogate, a fifth dimension is added and hence the 5D-QSAR[30].

Later, an extra dimension was added to account for a solvation function[31] as an extension to the Quasar technology. The new dimension considers simulations from different solvation models[32]. This was termed the 6D-QSAR.

1.3 Development of the receptor theory – ligand-target interaction

The elucidation of the function and structure of drug receptors has been the basis for SAR studies[33]. This has been more prominent with the unparalleled advances in genomics. The generally accepted theory is that exogenous as well as endogenous substances interact with a binding site on a specific macromolecular receptor governed by intermolecular forces. Such interaction is responsible for the pharmacological or toxic responses that these substances initiate depending on its ultimate site of action.

In 1878, Langely pioneered the idea that chemicals interacted with specific biological receptors. While studying the mutually antagonistic action of the alkaloids, he realized that they interacted with some receptive body in the nerve endings of the glands[34]. Paul Ehrlich stated that “Corpora non agunt nisi fixata” suggesting that chemicals interact with specific macromolecules in the body in order to exert their biological action[35]. He defined the receptor as the “binding group of the protoplasmic molecule to which a foreign newly introduced group binds”[36]. In 1905 Langley’s studies on the muscular contraction effects of curare led him to describe the first characteristics of receptors; their capacity for certain ligands and an amplification component that leads to a pharmacological response[37].

Enzyme proteins act as receptors through which chemicals can exert their action. They are typically more favorable in QSAR studies due to their ease of isolation and amplification. On the other hand, membrane receptors are whole proteins entrenched in the phospholipid bilayer of cell membranes. They need a thorough treatment with detergents in order to dissociate them for study. This often leads to loss of structural integrity and thus function. However, some membrane-bound receptors have been isolated and their three-dimensional structures elucidated. However, the membrane separation usually ensures loss of reactivity keeping the receptor chemistry understanding a challenge.

Nucleic acids are also important chemical receptors (aptamers). They interact with a diverse set of small organic molecules. Aptamers were successfully isolated and studied using *in vitro* selection techniques[38]. Also lexitropsins have been subject to QSAR studies as groove-binding ligands for potential drug development[39].

Prior to the revolutionary development of molecular graphics simulations and the high advancements in resolving the atomic coordinates for enzyme-ligand complexes using X-ray crystallography, the study of ligand-receptor interactions using QSAR has focused on enzymes. These recent advancements encouraged the elucidation of the mechanistic foundations of ligand-receptor interactions[8].

Fischer's rigid lock-and-key concept was dispelled through probing different enzymes by various ligands. The lock-and-key concept suggests that the ligand acts as a key that fits exactly to its lock (receptor). In that sense, it was thought that a "negative" imprint stands on the surface of the enzyme leading to the geometric complementarity that forms the ligand-receptor complex. However, this theory does not explain allosteric ligands. This encouraged the development of the induced-fit model. Updated models picturing flexible keys and "deformable" locks were suggested based on NMR and other structural studies[40].

The ligand receptor interactions are attributed to the low energy state of the ligand- receptor complex. This is necessary, because the low concentrations of ligand and receptors in the biological system do not permit the law of mass action to explain the pronounced pharmacological effect of chemicals. Therefore, the biological activity of chemicals is determined by their receptor affinity, which is measured by its K_D, the dissociation constant at equilibrium. A small K_D indicates a high affinity (larger concentration of ligand-receptor complex). This is mostly a result of noncovalent interactions sometimes augmented by a few covalent bonds.

$$\boldsymbol{K_D = \frac{[Ligand][Receptor]}{[Ligand - Receptor\ complex]}} \qquad \text{Equation 10}$$

The spontaneous bond-formation between atoms leads to a decrease in free energy (ΔG is negative). Equation 11 shows that the change in free energy ΔG is related to the equilibrium constant K_{eq}. Therefore, a little change in ΔG^0 can result in a measurable effect on the equilibrium constant.

$$\boldsymbol{\Delta G^0 = RT\ lnK_{eq}} \qquad \text{Equation 11}$$

1.4 Targets relevant for toxicity estimation via HTS

Chemicals that mimic natural hormones can disrupt vital functions of the human and wildlife. Such xenobiotics are referred to as endocrine disrupting chemicals (EDCs). They are capable of exerting unfavorable outcomes through different mechanisms. Therefore, there is a growing interest in studying these diverse compounds. Although more studies are being conducted through *in vitro* and *in vivo* experiments, many SAR and QSAR research programs have developed in the recent years[41]. Experimental testing of chemicals is expensive and time-consuming, and in many cases, impractical for application to the large number of synthetic

chemicals in use. This section deals with important biological targets that are relevant for this work.

1.4.1 The aryl hydrocarbon receptor

The aryl hydrocarbon receptor (AHR) is a member of the basic region helix–loop–helix–PER/ARNT/SIM (bHLH–PAS) family. The ligand-activated receptor has been shown to play a key regulatory role in a variety of endogenous developmental processes[42–44]. A consistent response of activating the AHR is induction of gene expression. The receptor and its ability to specifically bind to 2,3,7,8-tetrachlorodibenzo-p-dioxin (TCDD, dioxin) were discovered in 1976[45]. Since then, identification of its ligands has been of high interest.

Among the most characterized chemical classes that are known to be ligands for AHR are environmental toxins, such as the Halogenated Aromatic Hydrocarbons (HAHs) and Polycyclic aromatic hydrocarbons (PAHs)[46–48].However recently, a large number of natural, synthetic as well as endogenous AHR agonists have also been identified that does not share the same structural scaffolds or physicochemical characteristics of HAHs or PAHs[46,49,50]. A need thus aroused for the investigation of the potential for compounds to be AHR ligands. This is of special interest to the molecular design process in pharmaceutical companies.

After binding to its ligand, AHR changes its conformation. This exposes a nuclear localization sequence (NLS) as well as the dimerization interface for the aryl hydrocarbon receptor nuclear translocator (ARNT protein)[51,52]. The ligand-bound-AHR complex then translocates to the nucleus[53,54]. The AHR is then released from its protein complex following its dimerization with ARNT. Formation of the ligand-bound AHR:ARNT heterodimer converts the AHR complex into a high-affinity DNA binding form[55,56] and binding of the complex to its specific DNA recognition site, the dioxin-responsive element (DRE), upstream of AHR-responsive genes leads to coactivator recruitment, chromatin rearrangement, increased promoter accessibility and increased gene transcription[56–59].

The wide variety of adverse effects of dioxins suggests how harmful AHR activation could be. However, many AhR activators are present in our daily diet (for example: compounds such as indolo-(3,2-b)-carbazole, flavonoids, and sulforaphane). Also, the discovery of endogenous AHR activators such as bilirubin, eicosanoids, tryptophan, cAMP, and indirubin suggests that the receptor activation may be a normal physiological process. It was proposed that high level persistent activation of AHR (such as that cause by Dioxin) is the reason of adverse effects[44].

1.4.2 The estrogen receptor

There are two identified estrogen receptors, alpha (ERα) and beta (ERβ). Both are members of the Class I of the nuclear hormone receptor superfamily and are composed of six functional domains. Separate genes located on different chromosomes control both ERα and ERβ[60]. Both subtypes show different tissue distribution. Whereas ERα mRNA is mostly expressed in the liver, heart, kidney, testis, skeletal muscles, uterus, pituitary and mammary gland, The ERβ mRNA is expressed in the ovary and prostate. Other tissues show equal levels of both mRNA, though with different cellular distribution. These tissues include the adrenals, epididymis, gonad, thyroid, and different brain regions.

Upon binding to a ligand, the ligand-binding domain (LBD) of the ER undergoes a conformational change, which allows the interaction with coactivators[61–63]. The ligand-bound estrogen receptor then undergoes dimerization and binds to the estrogen response element (ERE) thus activating the transcription of target genes. This has been termed the classical mechanism for ER activation. However, ER action can also involve ligand-independent activation (i.e., in absence of estrogen). In this case, the activation can be modulated by a number of signaling pathways including growth factors, protein kinase A, and protein kinase C. ER might also regulate the target genes in the absence of EREs. These different mechanisms mediate and enhance the transcription of ER and allow its activation in low hormone levels[64].

Due to its association with negative reproductive effects[65,66], ER is among the most comprehensively investigated receptors in the context of EDCs[67,68]. Various regulations necessitate the assessment of the estrogenic activity of chemicals[69–71]. There are several *in vitro* and *in vivo* protocols to identify potential endocrine pathway-mediated effects of chemicals, including interactions with hormone receptors[68,72–74]. However, due to their costs and limitations, toxicological data for estrogenic activity is only available for a limited number of chemicals[75–78].

1.4.3 The androgen receptor

The androgen receptor (AR) is another example of the steroid-receptor subfamily of the nuclear receptors. Its gene consists of 8 exons and is located on the long arm of the X chromosome.

After ligand binding, the AR undergoes a conformational change leading to the dissociation of the heat shock proteins (HSP). The AR then undergoes translocation from the cytoplasm into the nucleus. It uses its DNA binding domain to interact as a homodimer to specific DNA sequences termed androgen response elements (AREs)[79]. By dimerizing on the DNA, AR interacts with DNA regions in the nucleus leading to activation of gene expression. The AR is known to repress the expression of a number of genes[80,81].

Testosterone and its metabolite dihydrotestosterone (DHT) exert their effects on gene expression via the activation of AR. The interaction of these natural ligands with the AR in various target tissues regulates the final phases in the cellular cascade of normal male sexual differentiation. Complete insensitivity to androgens leads to a female phenotype[82,83]. The AR plays an important role fetal sexual differentiation[84], puberty and adulthood. The cellular resistance to androgens causes the androgen insensitivity syndrome (AIS)[82,85].

1.4.4 Peroxisome proliferator-activated receptor gamma

Peroxisome proliferator-activated receptor gamma (PPAR-γ) is a member of the nuclear receptor subfamily 1, group C and is encoded by the PPARG gene[86]. The genes are mainly expressed in white adipose tissue (WAT) and brown adipose tissue (BAT). It is considered the major controller of adipogenesis and a potent modulator of whole-body lipid metabolism and insulin sensitivity[87]. PPAR-γ, like other members of the PPARs, forms heterodimers with the retinoid X receptor (RXR)[86].

PPAR-γ has been intensively studied as a drug target because of its link to insulin sensitization and obesity[88]. PPAR-γ activated genes stimulate lipid uptake and adipogenesis by fat cells while PPAR-γ knockout mice fed a high-fat diet fail to produce adipose tissue.

Fatty acids and their derivatives have long been recognized to bind and activate PPARγ. However, specific endogenous ligands proved difficult to clearly define[89,90]. Synthetic thiazolidinediones (TZDs) are known to be potent activators of PPARγ and show robust insulin-sensitizing activities[91]. They were thus used in treatment of resistant Type-2 diabetes. However, meta-analyses of clinical trials have associated the use of rosiglitazone (thiazolidinedione medication marketed as Avandia®) with an increased risk of developing congestive heart failure, myocardial infarction, cardiovascular disease and all-cause mortality[92,93] which prompted the European Medicines Agency (EMA) to recommend its suspension[94].

1.4.5 Pregnane X receptor

The pregnane X receptor (PXR), also known as steroid and xenobiotic sensing receptor (SXR), is a member of the nuclear receptor subfamily 1, group I and is encoded by the NR1I2 gene. Upon activation, PXR forms a heterodimer with the retinoid X receptor. It then binds to hormone response elements of the CYP3A4 promoter on DNA, which provokes gene expression.

PXR plays a role in the metabolism of many xenobiotics through the induction of CYP3A4 oxidative enzyme[95,96]. PXR also plays a role in up-regulating the induction of conjugating enzymes and glutathione S-transferase (phase II metabolism)[97] as well as OATP2 protein responsible for uptake and efflux (phase III)[98] and MDR1[99]. PXR activators include a great variety of chemicals, both endogenous and exogenous. Such ligands include antibiotics, antimycotics, steroids as well as bile acids, dexamethasone and rifampicin[95,100].

Because of the great role that CYP3A4 plays in drug-metabolism, *in vitro* assays studying the transactivation of human PXR can play a significant role in investigating the effect of compounds on CYP3A4. This reduces the need for *in vitro* assays based on human liver tissue and primary hepatocytes, which are expensive and limited by donor availability. This can provide cost-effective means for predicting whether compounds would activate CYP3A4 *in vivo*, a long-standing goal in pharmacology and toxicology for studying drug interactions[96].

1.4.6 The aromatase

Aromatase is a member of the cytochrome P450 (CYP450) enzyme superfamily. It is also referred to as estrogen synthase due to its key role in estrogen biosynthesis. Specifically, the aromatase enzyme is responsible for the aromatization of androgens into estrogens[101]. The enzyme complex is localized and expressed in the endoplasmic reticulum. It consists of two components; a form of the CYP450 (Cytochrome P-450_{AROM}) which binds to the steroidal ligand and catalyzes the aromatization reactions series, and a flavoprotien (NADPH-Cytochrome P-450 reductase) which transfers the reducing equivalents from NADPH to Cytochrome P-450. In humans, the enzyme is encoded by the gene CYP19[101] which has nine exons as well as many non-coding first exons that regulate tissue-specific expression. In addition to the gonads, aromatase is also present in various tissues in both genders including the brain, breast, skin, bone as well as adipose tissue[102].

Mutations in CYP19A1 gene expressions can lead to abnormal estrogenic activity. Excessive expression can result in gynecomastia in boys and precocious puberty and gigantomastia in girls. It has also been related to short stature due to early epiphyseal closure in both genders[103]. On the other hand, reduced aromatase activity can result in Aromatase deficiency syndrome. It is characterized by accumulations of androgens during pregnancy resulting in possible female virilization. Girls will have primary amenorrhea. Both genders can be tall, as lack of estrogen does not bring the epiphyseal lines to closure. These effects have been further studied using *in vivo* animal models on aromatase knockout mice (ArKO) or (AROM+) mice overexpressing human aromatase[102].

1.4.7 HERG

The human ether-a-go-go channel (hERG) is a member of the Kv family of voltage-gated potassium channels. The crystal structure of the channel is yet to be solved. However, basic understanding exists about the 3D topology of the protein structure being similar to other members of the voltage-gated K^+-channel family. The hERG related gene (Kv11.1) encodes for a voltage dependent ion channel, the blocking of which has been associated with the withdrawal of several non-cardiovascular drugs due to potential severe heart arrhythmia[104–109]. Three types of conformational states exist for the hERG channel: closed, open, and inactivated. The binding affinity of the many hERG blockers can be correlated to the conformational states of the channel[110,111].

The early assessment of hERG-related cardiotoxicity has become a common practice in drug discovery. Specifically, the drug-induced long QT Syndrome (LQTS) may cause avoidable sudden cardiac arrest. FDA thus mandates the testing for hERG safety.

Many *in vitro* assays exist for the pre-clinical evaluation of hERG-related cardiotoxicity[112] such as *in vitro* electrophysiology measurements, rubidium-flux assays, fluorescence-based assays, and radioligand binding assays[113]. *In silico* (i.e., computer-based) models have also been proposed to predict the potential hERG blockers in early virtual screening for drug discovery[111,114].

1.4.8 Antioxidant responsive element (Nrf2/ARE)

Nrf2 is a transcription factor is encoded by the NFE2L2 gene[115] in humans. The strong similarity with the ARE consensus sequence[116,117] associate these proteins as candidate factors for regulating the antioxidant responsive element (ARE) response. Reactive oxygen species and electrophiles cause the activation of the transcription factor which leads to the induction of (ARE)-mediated genes responsible for oxidative stress and phase II detoxification[118,119].

Nrf2 is kept in the cytoplasm by a cluster of proteins that degrade it quickly under normal conditions (i.e., in lack of stress conditions). In case of oxidative stress, Nrf2 does not degrade. It travels to the nucleus where it attaches to a DNA promoter. Antioxidative genes are therefore transcribed and the corresponding proteins are expressed.

Oxidative stress has been implicated in the pathogenesis of a variety of diseases ranging from cancer to neurodegeneration. Reactive oxygen species (ROS) and electrophiles may lead to DNA damage and therefore cause malignancies or develop other diseases[120–122]. To defend against these risks, organisms use multiple defense mechanisms[123,124], including the use of

phase II detoxifying enzymes and oxidative stress-induced proteins[125,126]. Model studies for carcinogenesis have demonstrated how such mechanisms might perform its role[127].

Tecfidera is a marketed drug used for multiple sclerosis patients to reduce relapse rates and increased time to disability progression. Tecfidera activates the Nrf2 pathway and was recognized *in vitro* as a nicotinic receptor agonist[128]. The mechanism by which it exerts its action is still unknown. Oltipraz is another NRF2 inducer that inhibits cancer formation in rodent organs[129]. However, it failed to demonstrate efficacy in human trials and has been associated with severe side-effects including neurotoxicity and gastrointestinal toxicity[130].

The activation of NRF2 have been also associated with the formation of de novo cancerous tumors[131] and the raising of plasma and liver cholesterol levels resulting in atherosclerosis[132]. Such adverse effects have been suggested to overshadow potential gains from antioxidant induction by NRF2 activation[132,133].

1.4.9 ATAD5

Human ATAD5 protein is encoded by the Genome Instability Gene 1 (ELG1; human ATAD5). Its levels increase in response to different types of DNA damage. Thus, it has been used as a biomarker for identifying genotoxic compounds. ATAD5 is involved in the RAD9A-related damage checkpoint pathway. Such pathway is crucial in checking whether DNA damage is compatible with cell survival or whether apoptosis sequence should be induced[134].

The need of cancer cells to insistently grow has been targeted by many chemotherapeutic agents. These drugs exploit the sensitivity of cancer cells to the inhibition of DNA replication through DNA damage by genotoxic chemicals. Exposure to such agents leads to DNA lesions, which stops DNA replication, collapses replication forks, and produces DNA double-strand breaks (DSBs), resulting in cell death. However, the genomic damage caused by these agents can also induce mutations that might make cells more resilient to cell-cycle checkpoints and apoptosis.

Ligands that induce ATAD5 have been studied with the aim to identify genotoxic compounds that can kill rapidly dividing cancer cells with minimal adverse effects[135].

1.4.10 Heat shock factor response elements (HSEs)

Heat shock factor response elements (HSEs) are specific DNA sequences located in the heat-shock responsive genes. They bind to the Heat Shock Factor (HSF) in response to an exposure to stress conditions such as heat shock. HSF is a transcription factor, which regulates the heat shock protein expression[136]. In humans, three transcription factors exist (HSF-1, -2, and -4). The heat shock response (HSR) is rapid. The translocation of HSF from cytoplasm, activation and binding to HSE occurs within minutes of exposure to elevated temperatures[137]. The HSF-bound sequence motifs represent only a small fraction of the total HSEs present in the genome[138]. The heat shock proteins ensure that cells are able to cope with stressful condition, which would otherwise cause irreversible cell damage and consequently cell-death[137]. The heat shock proteins (referred to as molecular chaperons) play a significant role in the synthesis, transport and folding of proteins.

Numerous chemicals, ecological and physiological stress conditions can activate the HSR. In addition to overcoming the impact of such conditions, proper HSF function was also found to be crucial for the animal development[139] and the survival of cancer cells[140].

1.4.11 Mitochondrial membrane potential (MMP; Δψm)

Mitochondrial membrane potential (MMP) is one of the parameters used to measure the integrity of mitochondrial functions. Mitochondria generate ATP by making use of the proton electrochemical gradient potential (Δp) acting as the cellular power plant. The electrochemical proton motive force is generated by a series of reduction reactions referred to as the respiratory electron transport chain (ETC). The inner mitochondrial membrane plays an essential role in the process. Its ETC protein complexes I through IV reductively transfer electrons providing enough energy to push protons against their concentration across the membrane and out of the mitochondrial cytoplasm. The accumulation of protons outside the membrane then flows back through ATP-synthase (Complex V) generating ATP in the process and closing the ETC cycle. Δp provides the force behind the ATP production. It is a function of both the MMP and the mitochondrial pH gradient. MMP constituent of Δp delivers the gradient charge necessary for Ca^{2+} sequestration in the mitochondria, and controls the formation of reactive oxygen species (ROS). It is thus considered a fundamental controller of cellular well-being[141,142].

Under stress conditions, MMP may in turn be altered by the dysregulation of intracellular ionic charges [e.g., Ca^{2+} [141–143] or K^{+} [144]]. This would lead to an alteration in Δp and consequently the production of ATP. However, the mitochondrial capacity for overcoming such changes may be exhausted resulting in ionic fluxes and ultimately a collapse of the Δp, MMP, and/or mitochondrial pH gradient may collapse. This, in turn, would result into bioenergetic stress due to the inability to produce ATP[143].

Due to is crucial role in different cellular processes; MMP is considered a key marker of cell health or injury. It was used as a tool to monitor changes in physiological functions of the mitochondria and its capacity to generate ATP through oxidative phosphorylation. *In vitro* fluorescence assays evaluate MMP to assess the potential of chemically induced mitochondrial toxicity; Using lipophilic cationic fluorescent dyes, a decrease in MMP following exposure to chemical stressors can be detected.

1.4.12 Tumor protein p53

Tumor protein p53 (p53) is encoded, in humans, by TP53 gene. It encodes, at least, twelve protein isoforms, which are collectively referred to as P53 isoforms. The diverse isoforms regulates the cell fate in reaction to various stresses though differently regulating gene expression[145]. The isoforms are differentially expressed in numerous human cancer types where they are found to modulate p53 transcriptional activity and tumor-suppressor functions. p53 is a transcription factor. As a tetramer, it directly binds specifically to p53-responsive elements on DNA. This induces or represses gene expression[146,147]. Studies show that an estimated 3,600 target genes are directly regulated by p53[148].

p53 is also referred to as tumor suppressor p53 or “the genome guardian”. This is due to its role in increasing genome stability, reducing mutation and suppressing cancer formation. Generally, P53 performs its function by preventing the proliferation of damaged cells. Such

cells are more likely to have mutations and show abnormal cell growth leading to cancer formation. The TP53 gene is considered the most frequently mutated gene in human cancer (more than 50%)[145].

Upon cellular exposure to stress conditions or DNA damage, P53 expression is activated. Based on the kind of stress and the degree of DNA damage, the activated p53 would trigger either cell-cycle arrest, DNA repair or programmed cell-death (apoptosis). The mechanism behind such choice is still not clearly understood[145]. p53-mediated apoptosis is thought to be the principal source of tumor suppression. Measuring the activation of P53 can thus provide a good indication of cellular insult or DNA damage.

1.5 The general QSAR problem in toxicity

As discussed earlier, the concept of QSAR is founded on the hypothesis that biological activity, like any other property, is a function of molecular structure, consequently it is expected that alterations in the molecular structure is reflected in a change in the biological activity of the compound. However, direct prediction of a compound's properties from its molecular structure (*ab initio*) is typically very limited. Therefore, the discipline of QSARs uses an indirect approach in order to tackle this problem as illustrated in Figure 3.

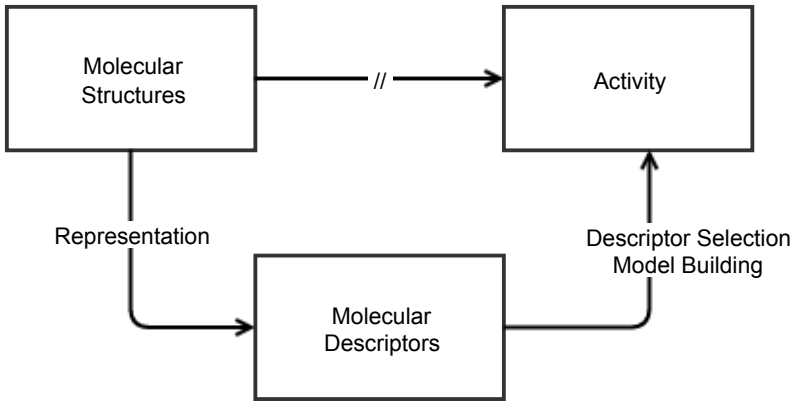


Figure 3. The general QSAR problem. Chemicals are represented in the form of molecular structures which cannot be directly correlated to the activity. Therefore, molecular descriptors are calculated from the given structural representations and correlated to the activity under investigation using a model building process.

Numerical descriptors are used to describe the chemical information encoded in the molecular structure for a set of compounds (termed a training set). Once numerical descriptors are available, the QSAR problem becomes essentially a problem in statistical model building, where statistical methodologies can be used to relate the set of numbers representing the structures to those representing the biological activities. This is an inductive technique that depends on the availability of a compound set for which the activity (or any other property) is already determined experimentally.

Figure 4 depicts the general QSAR model building process. First, the molecular structures of training set compounds are entered and stored. At minimum, these structures provide information on the molecules' topology. Multiple approaches, experimental or theoretical, can be used to determine a reasonable 3D structure for the compounds and therefore allow

molecular descriptors to gain information related to structures' geometries. Second, *in silico* descriptors are calculated. Then, statistical methods are applied to build models that relate the descriptors with the activity or property of interest. Finally, the models are validated (e.g., with an external dataset which has not been used for model building). Therefore, the steps of a QSAR study are generalized as: (1) Structure modeling and the selection of a geometry, (2) Descriptors calculation, (3) Descriptors selection (prefiltering), (4) Model building (fitting), and (5) Model validation. Steps 3-5 can be iterated upon, in combination with an optimization algorithm (e.g., with cross-validation or bootstrap aggregation) allowing the selection of a descriptor subset with maximum predictivity. Furthermore, the fifth step can also be performed on an external set that was not part of any training.

1.5.1 Role of computational toxicology in environmental risk assessment

ECHA described the role of animals in ensuring the safe use of chemical substances as being the last resort. This is one of the key principles for the REACH legislations. It encourages the use of so-called "alternative approaches" to reduce animal testing. QSAR modeling is one of the promoted mechanisms for alternative chemical risk assessment. Guiding documents exist that explain the best-practices and the requirements for accepting QSAR models' predictions[149]. These guidelines are essential for directing the stakeholders on how to utilize QSAR methodologies in a manner that gets accepted by the regulators. Thus, evaluating the human and environmental toxicity risks, complying with the regulatory requirements and reducing the need for animal testing at the same time.

1.5.2 The five OECD principles for QSAR model construction in general

Although the alternative approaches for animal testing are highly encouraged, their proper use must be established. For QSAR model building, five OECD principles were developed to ascertain the validity of good QSAR models for use in regulatory purposes and assessment of chemicals' risks. In 2004, the 37th OECD's Joint Meeting of the Chemicals Committee and the Working Party on Chemicals, Pesticides and Biotechnology established these principles for the validation of QSAR Models for regulatory purposes[150,151]. As this work is intended for the consideration of REACH applications, the OECD principles were taken into consideration during the development of all QSAR models. This section describes the five OECD principles for QSAR model validation.

Principle 1: Defined Endpoint: To ensure the transparency in any physicochemical, biological or environmental effect that a model is trying to assess, such an endpoint needs to be well defined. This includes the experimental conditions and measurement protocols. Ideally, data used in QSAR model development should belong to a single protocol. In practice, this is seldom possible. It is often sensible to combine data produced form different protocols[151].

Principle 2: Unambiguous algorithm: The "algorithm" refers to the form of relationship between the descriptors of chemical structure and the endpoint in the QSAR model. This can be a mathematical/statistical methods or rule-based models defined by experts. Presenting a clear description of the algorithm ensures transparency and allows others to reproduce the model and explain how predictions are generated. Some proprietary models do not use publicly available algorithms; therefore, their results could be reproduced but not explained. The ability to reproducibly select an appropriate QSAR model (considering the third principle),

calculate molecular descriptors and use them to produce an estimate (i.e., prediction) is crucial for the acceptance of QSAR models for regulatory purposes. However, it is not necessary to delve into the mathematical and statistical details of algorithm development in order to offer a transparent description. A regression-based QSAR can be explicitly defined without particular discussion of the regression approach[151].

Principle 3: Defined domain of applicability (AD): QSAR models are expected to give reliable predictions only for chemicals that are similar to the ones used in the model's training process. Therefore, the scope and limitations of the model must be defined by the model developers. This is based on the physicochemical, structural and/or response information of the training set used. An extrapolation outside the model's applicability domain boundary is likely to give unreliable estimates. At bare minimum, a binary response should be provided on whether a certain prediction falls inside or outside the applicability domain of the QSAR model. Quantitative assessment of the model's confidence in prediction can also be expressed in the form of an AD confidence interval. This reports the degree of similarity between the compound to be predicted and the model's training set[152,153].

Principle 4: Appropriate measures of goodness-of-fit, robustness and predictivity: This principle highlights the need for statistical validation of QSAR models in order to judge models' performance. Such performance validation can be either internal or external. Internal validation judges the ability of the model to correlate the structures and activities (i.e., the molecular descriptors with the property of interest) within the training set. This is referred to as "fitting" and is measured through the goodness-of-fit and robustness. To avoid the risk of "over-fitting", external validation is used to check predictivity. In this case the ability of the QSAR model to provide reliable estimates for an external set of compounds (i.e., that was not used in its training) is tested considering the model's applicability domain. Among the techniques discussed by OECD are response randomization test, cross-validation, bootstrapping, training/test splitting as well as external validation test sets.

Principle 5: Mechanistic interpretation, if possible: This principle aims to encourage finding mechanistic basis for the validated QSAR model that adds to the understanding of the statistical validity and the domain of applicability. Describing the relation between chemical structure and activity (or any property thereof) using statistical and machine learning approaches is supposed to complement (and not replace) the existing chemical and toxicological knowledge. Thus, efforts should be taken, during QSAR models validation, to show the consistency of such models to the related known chemical and toxicological processes. Consistency of the model with existing theories and knowledge of biochemical mechanisms justifies and explains how predicted values from the model are generated and therefore increases the transparency of judging the model's performance.

The "if possible" phrase shows that the mechanistic interpretation is not mandatory for model acceptance by regulators. Sometimes, the iterative model building process and the involvement of data-mining techniques increases the complexity of the developed QSAR models through multiple training set refinements rendering the mechanistic interpretation hard to establish.

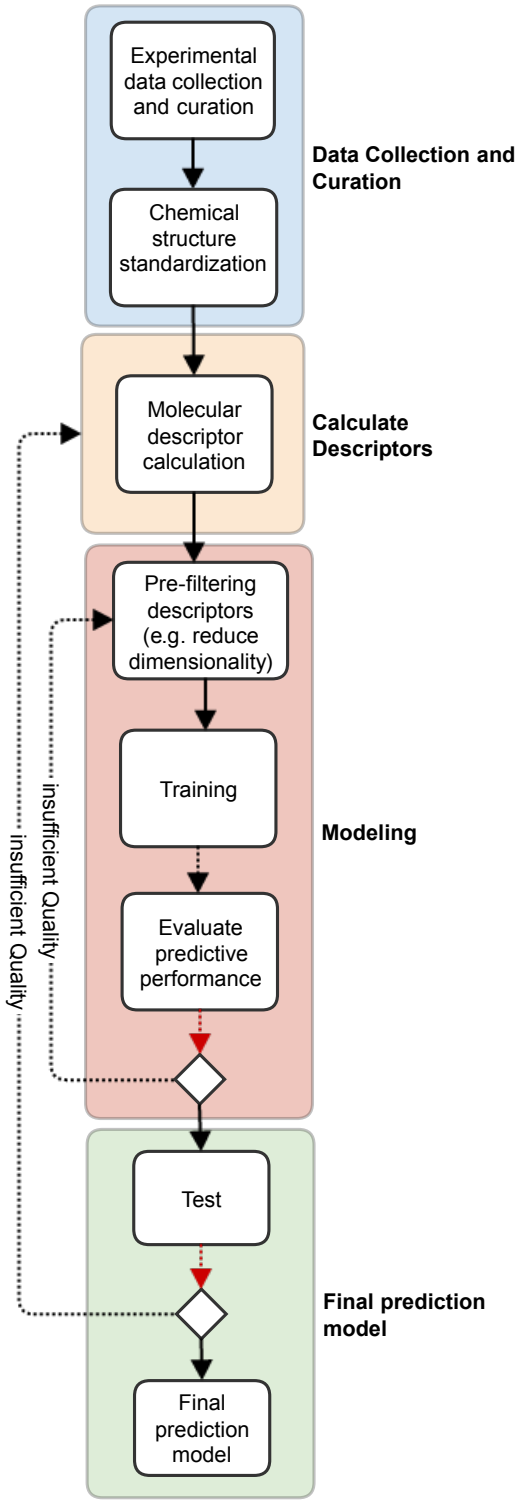


Figure 4. Diagram depicting the general steps in QSAR model building process. The first step is the collection and curation of high quality data including the standardization of the structural representation of chemicals. Then, descriptors are calculated from such representation. Afterwards, QSAR models are trained and validated before potentially being tested on external test sets.

1.6 Role of *in vitro* assays in alternative testing – databases generation and growth (TOXCAST, TOX21 and EDSP)

High-throughput screening (HTS) allows researchers to conduct millions of chemical, genetic, or pharmacological experiments with minimal intervention. Such procedure can quickly distinguish active compounds, antibodies, or genes that control particular biochemical pathways. The results of these assays can guide the research process and thus has become a viable tool for large-scale chemical testing[154–156]. The large amounts of data generated by HTS can be used to correlate chemical structures to their biological activity. QSARs can support the identification of key characteristics in chemical structures responsible for such activity. This knowledge can then be used to provide predictions on the possible activity of chemicals in virtual screening settings for regulatory purposes. The quality of QSAR models based on large chemical libraries from HTS experiments can vary. However, the accuracy is usually high enough to support prioritizing chemicals that are worth subjecting to experimental testing. This satisfies the imminent need to prioritize chemicals for accelerating the chemical registration process and lowering the experimental testing costs[157].

As high throughput technologies advance, more data are being produced from *in vitro* profiling of chemical substances[158]. Numerous chemical and biological databanks have immensely developed in recent years regarding their diversity and size. Such data can contribute as a potential substitute or complementation for *in vivo* animal studies.

The U.S. Environmental Protection Agency (EPA) participated in projects to profile *in vitro* bioactivity of chemical substances. ToxCast[159,160] program was launched in 2007 as a multi-phased project that uses automated HTS assays[161–163] to test the effect of exposing *in vitro* cells or isolated proteins to chemical substances. Afterwards, the treated living cells or proteins are tested for alteration in their biological activity. This could suggest a possibility for toxic effects that may lead to potential adverse effects on human health. Such advanced technologies can rapidly and efficiently screen large number of substances and reduce the need for animal toxicity studies[164].

Tox21[165,166] is another example of a multi-agency effort that uses HTS assays for toxicity modeling and prediction. EPA, The National Institutes of Health (NIH), The National Center for Advancing Translational Sciences (NCATS), The National Institutes of Environmental Health Sciences/National Toxicology Program (NIEHS/NTP) and the Food and Drug Administration (FDA) cooperate in screening chemical substances for some potential toxic effects. The screening data can then be used, with the assistance of *in silico* techniques, for the prediction of toxicity. This has the potential for providing an economical method for toxicity testing prioritization for thousands of still untested compounds[167].

EDSP21 is another example for the usage of pioneering screening techniques for prioritization of toxicity assessment of chemicals; in 2011, EPA published the EDSP21 work plan as a successor of the Endocrine Disruptor Screening Program (EDSP) established in 1998. The work plan explains the foundation and basis upon which the transition will materialize. EPA shares the aim of significantly reducing animal testing, making testing fast and less costly and of providing characterizations of chemicals, chemical mixtures, and toxicity endpoints[168]. The EPA stated that “Using this current process [EDSP] to continue to identify chemicals for

screening, having them screened, and making decisions about more definitive testing, is not sustainable to evaluate the tens of thousands of chemicals that fall within the purview of EPA"[169] .

Figure 5 shows the size of the chemical libraries and number of assays in these projects. Within all mentioned studies ToxCast has the most comprehensive *in vitro* assay panel with about 600 assays. In its first phase, the program screened 309 chemical substances most of which were food pesticides. As such, these chemicals had a wide-range of animal toxicity data available. These chemicals come from various sources (as shown in Figure 6) resulting in wide diversity in structural groups, complexity and physicochemical properties.

ToxCast program aims to construct "bioactivity signatures" that are intended to prioritize chemicals for targeted testing and predict possible adverse outcome pathways for such chemicals[170]. These signatures can be built through acquiring enough information on a collection of chemicals and using them to define distinguishing patterns of toxicities, or phenotypes, detected in existing experimental animal toxicity studies.

Earlier studies examined the possibility of utilizing *in vitro* assays in prediction of *in vivo* endpoints[171–173] and investigated the biochemical pathways behind toxic or adverse effects. While many studies inspected particular *in vivo* toxicity endpoints, a comprehensive investigation of *in vitro* – to – *in vivo* predictive power across multiple toxicity endpoints utilizing the ToxCast HTS assays has been independently conducted by Thomas et al[174]. In their analysis, the authors used different statistical classification methods, in combination with cross-validation, to test the ability of *in vitro* assays to predict the outcomes of 60 *in vivo* toxicity endpoints from ToxCast Phase I screening data. They suggested that the assays and chemicals used in ToxCast phase I can be useful for prioritization of chemical testing but otherwise have restricted applicability in predicting *in vivo* chemical hazards.

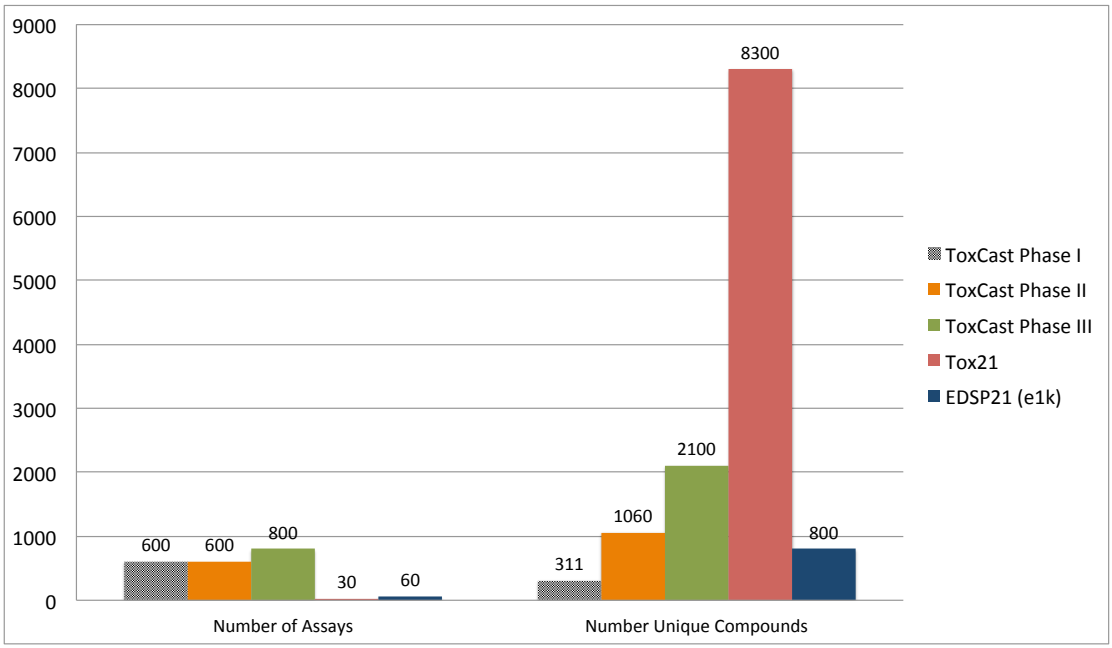


Figure 5. Different screening programs managed by the US EPA and its partners. The ToxCast program has the most comprehensive number of *in vitro* assays while the Tox21 project includes the most diverse set of chemicals (8300). ToxCast phase III will extend the chemical library of ToxCast by 1000 new compounds and an additional 200 assays.

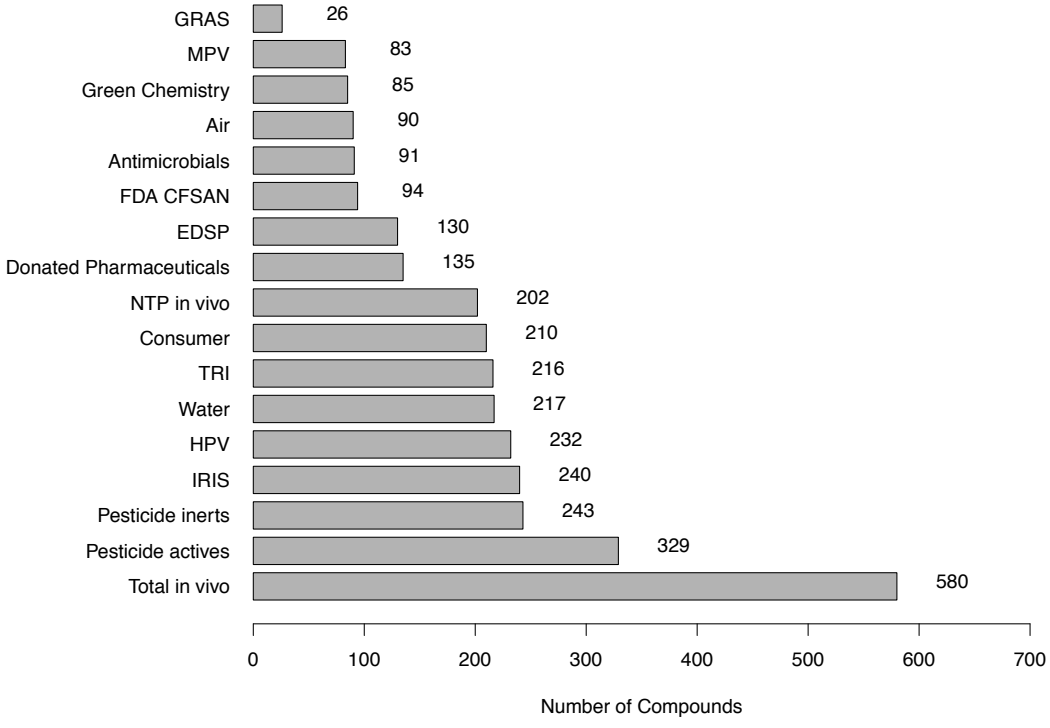


Figure 6. Inventory sources for ToxCast Phase I & II chemicals. Phase I & Phase II covers 1060 chemical compounds, EDSP21 (e1k) adds another 800 compounds (total: 1860). Total 2806 chemicals overlap across 16 diverse inventories. GRAS: Food and Drug Administration (FDA) - Generally Recognized as Safe. MPV: Medium Production Volume, FDA CFSAN: Center for Food Safety and Applied Nutrition, EDSP: Endocrine Disruptor Screening Program, NTP: National Toxicology Program, TRI: Toxics Release Inventory, IRIS: Integrated Risk Information System, HPV: High Production Volume.

2 Motivation and aims

Multiple studies were performed in the area of QSAR prediction of toxicity. These studies, however, were limited in their scope, the number of machine learning algorithms used, and the diversity of the descriptor packages or the integration of applicability domain for estimation of model confidence in predicting new compounds. The aim of the current thesis was to improve the situation and study different aspects of the utilization of high throughput screening in predictive toxicology.

- The work was supposed to expand on the analysis of the HTS *in vitro* assays ability to build a bioactivity signature both alone and in combination with different *in silico* descriptor packages. In particular, more *in silico* descriptor packages and classification algorithms using comprehensive validation protocols (bootstrap aggregation and cross-validation) were to be assessed.
- Another objective of this work was investigating the ability of *in silico* modeling to predict the outcome of HTS screening. For this purpose, a wide range of stress-response elements and nuclear receptors that are associated with toxicity in mammals was to be examined in multiple QSAR studies.

All QSAR studies in this work aimed to be complemented with applicability domain estimation to assess the suitability of models for application within the scope of any given chemical space. Structural information and *in vitro* pathways were also to be used, when possible, to provide mechanistic interpretation for the witnessed toxicity. All studies should be designed to comply with the OECD guidelines for QSAR model building. Models were designed to be publicly accessible to assist regulators and toxicologist in screening their chemical libraries for potential toxicity.

The overall aim of the work was to present an in-depth examination of the successes and limitations of the current HTS initiatives including ToxCast, Tox21 and others within QSAR studies. It was conceptualized to include construction of *in silico* QSAR models for the prediction of *in vitro* assays outcomes, analysis to which extent *in silico* descriptors can capture information in the *in vitro* assays (some *in vitro* assays, such as the activation of several nuclear receptors, have already been associated with potential toxicities), the prediction of such outcome using computational modeling alone that could save both time and cost, assessment of the sources of variability in the datasets, and finally testing the applicability of an online tool deployed for the exploration of *in vitro* assay datasets.

3 Tools and methods

3.1 Experimental data sources

As QSAR aims to correlate chemical structures to their activities, the first step is acquiring reliable and consistent experimental activity data for such compounds. The principle of "defined endpoint" encourages using data from a single source to ensure the uniformity of experimental conditions. However, single experiments rarely hold enough data for a proper QSAR model construction. For practical reasons, data is often collected from multiple sources. Therefore, data curation is essential to safeguard the compatibility of experimental conditions. Thorough examination of literature is therefore necessary to acquire data from reliable sources.

Luckily, reliable data collections exist. Multiple public databases hold a capital of experimental data for compounds covering a wide range of properties (physicochemical, biological, toxicological as well as data on environmental fate). Among these databases are PubChem[175], ChemSpider[176] and ChemExper[177]. The PubChem BioAssay database allows the examination of information in PubChem BioAssay records, including experimental conditions provided by assay depositors as well as chemical annotations from PubChem. Assay information can be retrieved through the assay ID (AID) while information about certain chemicals can generally be accessed through the Chemical Abstract Registration Number (CASRN)[178,179] or a structural representation such as SMILES or a structure hash as the INCHI[180] key. The US EPA also offers a large set of databases through its online warehouse ACToR[181,182] (Aggregated Computational Toxicology Resource). Data is searchable online though a web-interface and can also be downloaded for offline analysis.

Another potential source of data collections comes from QSAR modeling software that integrates experimental databases. OECD QSAR toolbox holds a large set of referenced entries accessible through its user-interface. It allows searching through a multitude of physicochemical properties and environmental endpoints. The EPA software "Estimation Program Interface (EPI) Suite" also publishes its datasets publicly. The Online Chemical Modeling environment (OCHEM)[183] also publishes referenced records of experimental data used in building QSAR models. Other related projects based on OCHEM, such as QSPR Thesaurus[184,185] and iPrior[186,187] follow the same strategy.

While these sources are generally reliable, online datasets as well as those integrated in QSAR modeling software may still hold errors[188]. One of the commonly encountered errors is the presence of duplicates of molecules.

Curation is needed not only for experimental data but also for the chemical structure representations. Some substances may be excluded from a QSAR analysis due to the lack of appropriate cheminformatics techniques for their handling. For example, mixtures, inorganic and organometallic compounds would generally be discarded from the datasets. Salts are generally handled by removal of counter-ion and neutralization of the charges. When data is being collected from multiple sources, certain chemotypes could be represented differently. For example, diverse chemical sketching software could handle Nitro-groups and aromatic rings differently. Thus, a normalization of such chemotypes is important to ensure consistency

in descriptor calculation. More difficult cases such as poly-zwitterions, tautomers[189], or anionic heterocycles need more effort and multiple normalization steps[188]. For example, choosing between keto-enol forms could have a significant effect on the predictivity of QSAR models built with one form or the other. Such decision can be taken with knowledge of the compound's mechanism of action, if it is known for the investigated activity[190].

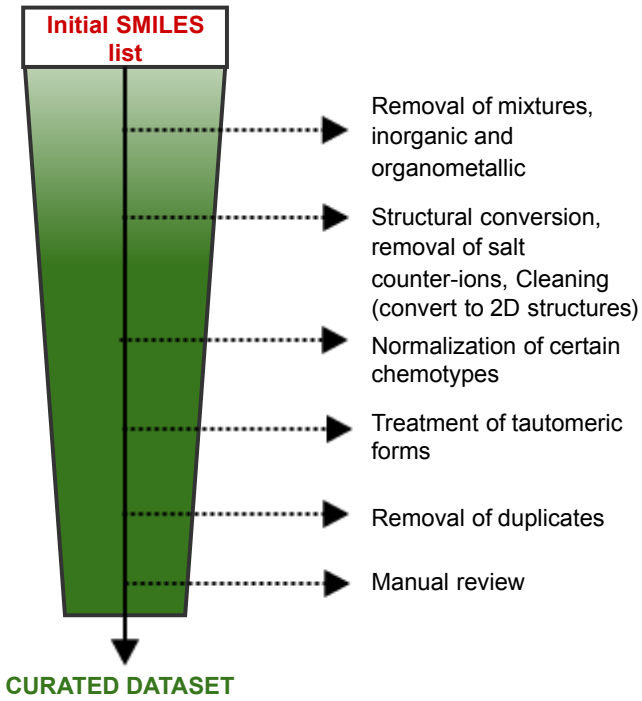


Figure 7. Data curation process in QSAR model building including the removal of structures that cannot represented by descriptors (such as mixtures and inorganics, etc.) and the standardization of the representation of different functional groups and 3D structure generation (when applicable). Finally, whenever possible, a manual expert review may be valuable (e.g., for detecting abnormalities and picking correct tautomer forms).

Finally, duplicate records should be inspected. These duplicates could be identical records that can be easily fixed by removing the redundant ones. More often, duplicates represent exact chemical structures with varying activity response. This could be due to different measurement units, error in data-entry or varying experimental conditions or inter-lab variations. For every QSAR study all values should be converted to an appropriate measurement unit suitable for the study (example: molar units rather than weights to describe activity). Data-entry errors could be fixed by going back to the original references and experimental records. That's why a referenced data collection is important in QSAR modeling. Inter-lab variations in experimental conditions should be manually examined and if needed, such data should be modeled disjointedly. Figure 7 shows the steps involved in the data curation process. Figure 8 shows an example molecular structure undergoing the standardization process.

The manual inspection of the dataset is therefore important for comprehensive assessment. Workflow tools exist that facilitate the curation and inspection process[191]. These tools allow

efforts to be focused on the technically challenging tasks in curation while automating most routine steps. In this work, the main tools used are KNIME and OCHEM.

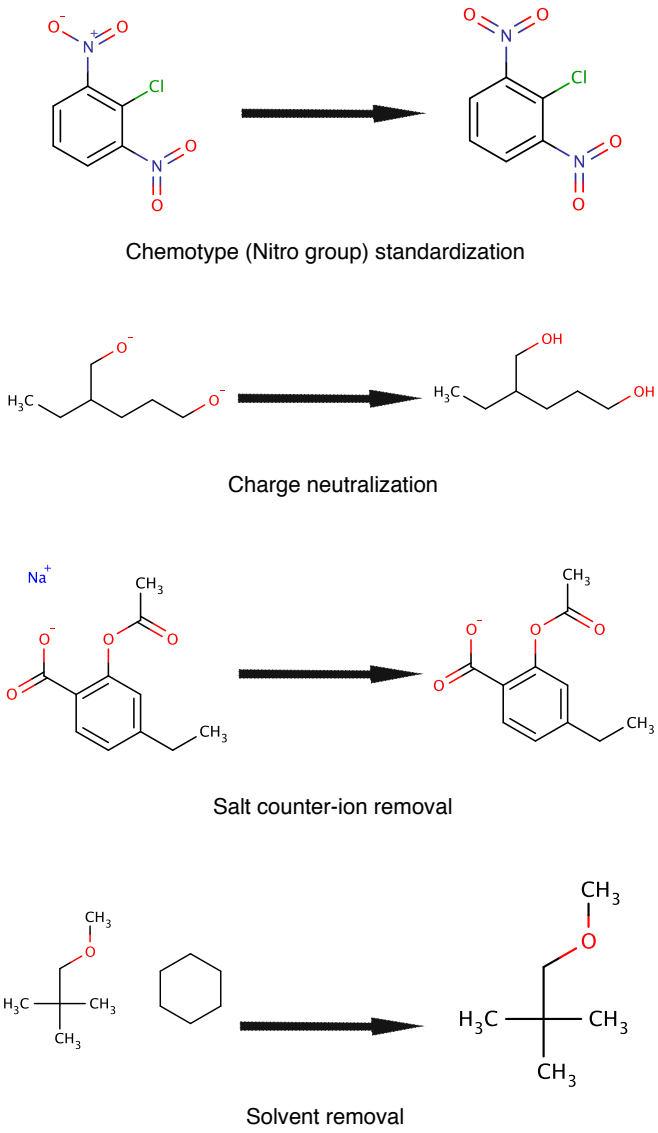


Figure 8. Examples of chemicals' preprocessing steps.

3.2 Workflow tools

3.2.1 OCHEM / iPrior

OCHEM[183] offers an interactive web interface (http://www.ochem.eu) that may be used to explore the data, construct QSAR models and run predictions. It also offers the ability to interpret results using prediction-driven matched molecular pairs[192]. Handling large datasets and thousands of QSAR models is more convenient using workflow systems such as KNIME[193]. For that, OCHEM exposes a number of methods through SOAP web services[194]. These methods allow the user to login, upload data, create properties, create or delete QSAR models, download model statistics, and to run predictions on the constructed models. OCHEM implements an XML format that allows users to configure the QSAR modeling tasks regarding all steps including descriptors calculation, descriptors pre-filtering, and configuring the machine learning algorithms.

iPrior was developed based on OCHEM[183] and QSPR Thesaurus[184] platforms. It offers an interactive web interface (http://iPrior.ochem.eu) (screenshot shown in Figure 9). Like OCHEM, it can be used to explore the data, construct QSAR models and run predictions.

All models are associated with a unique identification number (model id). That id can be used to access the model's profile page (see Figure 10). To access a certain model, users visit: http://iPrior.ochem.eu/model/[modelID] replacing [modelID] with the model identification number. The profile page lists, besides the model name and property predicted, the algorithm and descriptors used, pre-filtering parameters as well as the model's statistics. From this page, users have also access to the applicability domain graphs as shown in Figure 11. These graphs are automatically calculated (whenever applicable) based on the distance-to-model (DM) approach[195]. There are multiple DMs implemented on the platform such as: the standard deviation for an ensemble of models (STDEV) such as in the case of bagging models, the correlation in the models space (CORREL)[196] and the Mahalanobis distance (LEVERAGE).

Model quality can be judged through the statistical parameters presented in the model profile page. Whenever users are satisfied with the quality of the model they can apply it to new compounds. They get the option to draw a chemical structure directly through the web browser, select from a previous dataset or upload their own structure file in SDF format. Modeling results and statistics can also be queried using workflow systems such as KNIME.

Figure 9. Screenshot of the iPrior homepage showing different *in vitro* assays for which data are available as well as an excerpt of the published models available for users to run predictions against.

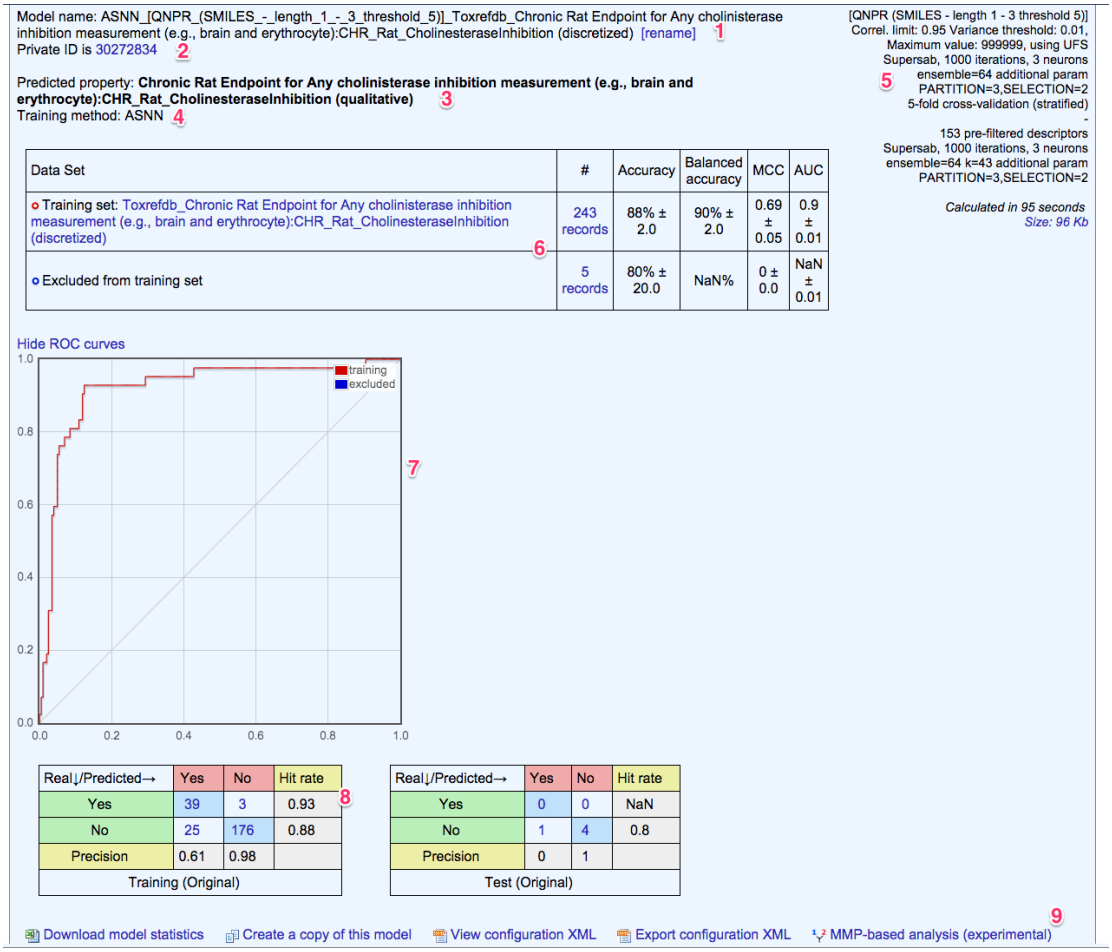


Figure 10. Model profile page for a good performing model showing (1) model name (2) model id (3)the predicted endpoint (4) the machine-learning algorithm used (5) The configuration for the learning algorithm and the pre-filtering parameters (6) The model's accuracy, balanced accuracy, Matthew's Correlation Coefficient and area under the receiver operating characteristic curve (AUROC) (7) The ROC curve (8) model confusion matrix showing hit rate and precision (9)different tools allowing model statistics download, model replication, exporting model configuration or analyzing the data matched molecular pairs.

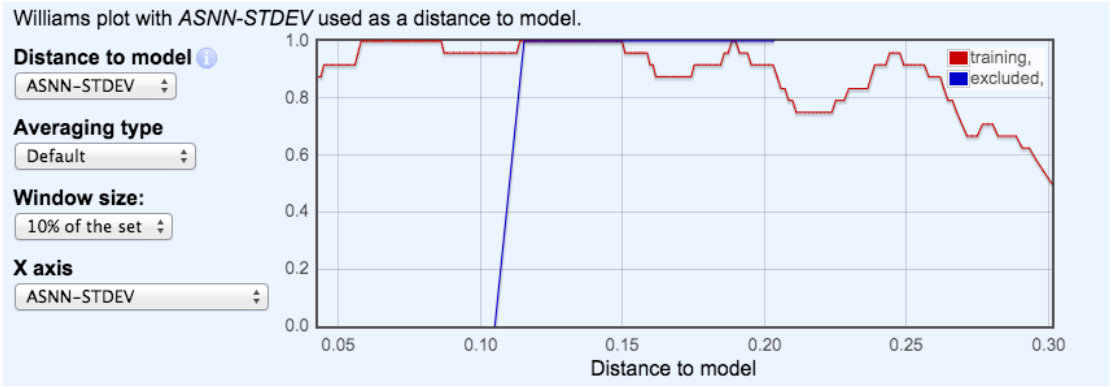


Figure 11. The applicability domain graph for the above model showing distance-to-model (DM) in respect of standard deviation of the ASNN ensemble (x-axis) and model accuracy (y-axis)

3.2.2 KNIME

KNIME[193,197] (Konstanz Information Miner) is a data analysis environment with a modular design. It allows data scientists to graphically program their data inspection routines. It also allows an intuitive approach for analysis. The software platform was originally designed at the chair of Bioinformatics and Information mining at the University of Konstanz. Since then, it became widely adopted by data scientists in different disciplines[198]. It implements a user-friendly graphical workbench based on the Eclipse open-source platform. KNIME workbench can be used throughout the entire data analysis process including initial data access, download, transformation, inspection, modeling, prediction analytics as well as visualization and report generation.

KNIME.com AG now develops multiple tools. The KNIME Desktop is a free and open-source workbench platform that was extensively used in this work. It is licensed under the GNU General Public License (GPL)[199] and includes more than 1000 nodes (modules) developed by the Company or its partners or contributed by the users' community. Many well-known third-party tools are integrated with KNIME and extend its functionalities. They range from statistical analysis and data mining tools such as WEKA[200,201], R[202] and MATLAB[203] to Cheminformatics tools[204] such as CDK, RDKit, Chemaxon, EMBL-EBI tools inter alia. For example, through Erl Wood Informatics, the Lilly group published 30 open source KNIME nodes[205]. These include format conversion nodes, viewers as well as nodes for fingerprint generation, docking, R-group analysis, matched pairs, multi-objective optimization and activity cliffs analysis.

Borrowing from its parent platform, Eclipse, KNIME is organized into a set of windows. This section gives a brief description of the important windows.

- The workflow editor is the main window, in which the visual programming and construction of the workflow takes place. Multiple workflows can be opened simultaneously in different tabs.
- The "Node Repository" window lists all installed and initialized nodes (modules) available for use. Nodes are sorted in an intuitive tree structure based on their function (for the native nodes) or their provider (for third-party nodes). Nodes can be searched based on their name. Every node performs a specific function (for example: filter rows). Scripting nodes are available that allows the execution of custom scripts in Perl, Java, Python, R, MATLAB among many languages.
- "Favorite Nodes" window is similar to the "Node repository" but only listing the latest and most frequently used nodes as well as the user's personal favorite nodes. It provides a useful shortcut to the user saving the effort of searching among the nodes tree.
- The "Node Description" window displays a quick help article describing the currently selected node, its actions and supported input and output ports.
- "Console" window displays the textual messages concerning the workflow execution. This includes warnings and error messages.
- "Outline" window shows a graphical thumbnail of the entire workflow and highlights the section currently viewed in the main window. This window is useful for large workflows where scrolling through the workflow space might be confusing.

For building workflows, nodes are dragged from the Node Repository into the workflow editor to be included in the workflow. Nodes can be connected to each other through input and output ports. The compatibility of such connections is governed by the corresponding node classes' metadata. Branching of the workflow is permitted.

For executing workflows, the nodes are executed in order (left-to-right) according to their connectivity. If the workflow has multiple left edges (possible starting points) or branching points, they would be executed in parallel. Almost all nodes process input data on row-by-row basis. KNIME can also interact with external software tools (through the "External tool" node) or with any external web service (through "Generic Web-service Client" node).

OCHEM also offers several nodes for directly integrating into KNIME[206]. These nodes aim to rendering essential OCHEM features accessible in KNIME workflows. Among these features are executing QSAR prediction as well as importing and exporting data. However, more features can be directly accessed directly through calling OCHEM SOAP web services via KNIME "Generic Web-service Client" node. Throughout this work, this node was used to bridge between both platforms as necessary. This proved particularly useful in uploading sizable sets of *in vitro* data, building a large number of QSAR models and downloading their statistics.

3.3 *In silico* representation of chemicals

While researchers regularly use the International Union of Pure and Applied Chemistry (IUPAC) names or 2D sketches to refer to chemical structures, these forms are typically not suitable for computation purposes. Computer-friendly molecule depictions are usually non-directed graph representations of the molecular structures. Among the most used are: SMILES / Unique SMILES, Molfile / SDF, MOL2 and InChI / InChIKey.

SMILES (Simplified molecular input line entry specification)[207] is an ASCII string representation of the molecular structure that was designed to be short, unambiguous and human-readable. SMILES are not unique (i.e., single chemical structures can have multiple SMILES representations). However, canonicalization algorithms have been developed that allow the generation of the same SMILES string for any given molecule. This unique selection of a specific representation forms the unique (canonical) SMILES[208]. Other set of rules is used to denote chirality, isotopism and configuration about double bonds. Collectively, these rules are referred to as isomeric SMILES[180]. The shortness of the representations comes at the expense of disposal of specific atomic coordinates and thus SMILES is not suitable for distinguishing conformations.

The short SMILES string is generated following a graph-based approach; by printing the symbol nodes encountered in a depth-first tree traversal of the chemical graph. Hydrogen atoms are removed and cycles are broken into a spanning tree. Numeric suffix labels are added to designate connected nodes in places where cycles were broken. Branching points on the tree are marked by parentheses.

InChI (IUPAC International Chemical Identifier)[209] is another textual one-string representation of molecular structures that was originally developed by the IUPAC[210] and the National Institute of Standards and Technology (NIST)[211]. InChI provides a human-readable

representation suitable for electronic chemical information storage[212,213]. The algorithms are non-proprietary and were later expanded through the InChI Trust[214,215].

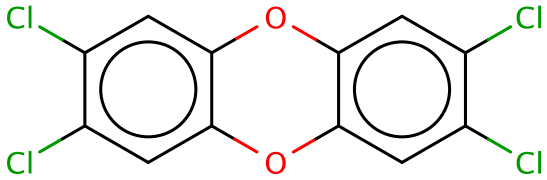
The InChI specifications represent chemical structures in the form of layers. There are four important layers included in the standard InChI format. They describe the atoms and bond connectivity (main layer), tautomerism, stereochemistry, isotopism as well as electronic charge information[214]. Some layers (such as tautomerism) are not mandatory and can be omitted. The advantage of separator-prefix format is the possibility to parse large amounts of molecules in InChI representation by wildcards or regular expressions to filter molecules with specific features. The layers and sub-layers are separated by a "/" delimiter. Layers also begin with a characteristic prefix letter (except for the chemical formula sub-layer of the main layer). The presence of such delimiter and prefixes facilitate the computer sub-search of information in certain layers of information using wildcards and regular expressions.

A hashed version of the InChI representation was developed to facilitate search engine indexing of chemicals on the web. The format is called InChIKey and embodies a 27-character string generated by hashing the standard InChI representation using the SHA-256. The InChIKey comprises 4 parts separated by hyphens. The first 14 characters hash the connectivity information layer in the original InChI, the next 9 characters hash the other layers, then a single character indicates the InChI algorithm version. Finally, a checksum character verifies the consistency of the key record. Due to the nature of the algorithm, a minute chance of hash collision exists (i.e., two different compounds having the exact same hash). This was estimated as 1 in 75 billion unique structures[216]. While the hash cannot give information on the chemical structures, it is used mainly for indexing and searching large chemical databases. Figure 12 displays some examples of SMILES, InChI, InChIKey representations as well as a 2D depiction of dioxin.

Molfile (or MDL) is a chemical-table file format[217]. It was originally developed by MDL Information Systems (MDL)[218], which is now a subsidiary of Dassault Systemes[219]. It consists of two main sections, the header (which holds the substance name, software used to generate the file as well as any other arbitrary information or comments) and a connection table section. The connection table contains, in addition to the counts line, three blocks for atoms, bond and properties. The counts line describes the total number of atom types, atoms and bonds as well as the specifications version. The atoms block details the Cartesian coordinates (in angstroms) for each atom (1 line per atom) as well as its atom type. The bonds block describes the connection between atoms (referring to their order in the atoms block) as well as the bond type. Finally, the properties section describes any complex information or properties available about the substance. While this format is not as concise as the SMILES or InChI, it was designed to support storing detailed information about the chemical structures.

SDF (structure-data file) extended the chemical-table file format by adding the support for storing additional information. The format provides the flexibility in representing any additional metadata through a key-value pair system (termed “tags”). Such information can vary from molecule synonyms to measured experimental data (e.g., logP) or calculated molecular properties (as molecular weight). The format also supports the storage of multiple structures in the same file separated by a four-dollar-sign $$$$ delimiter.

MOL2 is another chemical format that was developed by Tripos. To avoid restrictions of fixed file formats, the format specifications are feely available[220]. Similar to MDL Molfile format, the Mol2 is ASCII-based molecule representation. It can store information on atom coordinates, atom connectivity and bond types, as well as other additional information. It supports more atom types than SDF. For example, it distinguishes aromatic and non- aromatic carbon atoms. It can also store partial atom charges. This is an important feature for calculating charge-sensitive descriptors in QSAR.



IUPAC: 2,3,7,8-tetrachlorodibenzo[b,e][1,4]-dioxin

SMILES: Clc2cc1Oc3c(Oc1cc2Cl)cc(Cl)c(Cl)c3

InChI: 1S/C12H4Cl4O2/c13-5-1-9-10(2-6(5)14)18-12-4-8(16)7(15)3-11(12)17-9/h1-4H

InChIKey: HGUFODBRKLSHSI-UHFFFAOYSA-N

Figure 12. Molecular representation of Dioxin in different formats

3.4 Molecular descriptors

A molecular structure descriptor can be considered a mathematical representation of a molecule, which captures certain structural information. An important criterion for the success of QSAR studies is the degree to which these descriptors capture structural information relevant to the biological activity or property being studied. For example, if a property is sensitive to chirality while the descriptors used are not, one should not expect a good performance of such models.

Descriptors may represent physicochemical parameters (e.g., hydrophobic, steric, or electronic). They may also be structural descriptors (e.g., frequency of occurrence of a substructure), topological (e.g., connectivity indices), electronic (from a molecular orbital calculation), geometric (e.g., from a molecular surface calculation), inter alia. Many descriptors were specifically developed to suit a particular problem or group of problems. The number of descriptor can now be measured in thousands. It shows that there is no end to the ways by which one can represent a chemical structure[221,222].

Chemical compounds are considered complex systems with different ways of representing them[222]. Most chemical properties cannot be directly derived from the summation of the contribution of its individual parts (atoms or fragments).

In silico descriptors convert the structural representations of molecules into a matrix of numbers allowing statistical methods (e.g., machine learning algorithms) to correlate them to the properties of interest. Molecular descriptors can be generally classified according to the information used in their calculation:

- 0D-descriptors: These can be calculated directly from the chemical formula. Specific atom-type counts, constitutional indices and molecular weight are examples for these descriptors.
- 1D-descriptors: Require only fractional knowledge about the chemical structures with regard to fragments and functional groups. It is used in substructure analysis such as structural alerts[223,224] that correlate the presence of certain fragments (or their counts) with pharmacological or toxicological activities.
- 2D-descriptors: are calculated form the molecular topological graph representation based on graph theory. Topological and connectivity indices are examples of such descriptors. SMILES representations (2D) were also used as molecular descriptors in building QSAR[225] models.
- 3D-descriptors: They require knowledge about the molecular geometry. This can be obtained theoretically (*ab initio*), experimentally (e.g., X-ray) or using a molecular mechanics simulation among other approaches. These descriptors are derived from the 3D-representations of a certain conformer of the molecule under investigation. An example of such descriptors is the 3D-polar surface area.
- 4D-descriptors: study multiple conformations of the molecule of interest and thus account for the flexibility of the molecular representation. This notion is used in the Grid-based QSAR techniques such as Comparative Molecular Field Analysis (CoMFA)[226,227] by analyzing the averages and standard deviations of the 3D-descriptors calculated from an ensemble of conformations for the structures of interest.

Other descriptors that depend not only on the substance's structure but also on its interaction with the target protein are called chemogenomic descriptors. They are only relevant in the context of a certain biological inquest at hand. They describe the interaction between the small molecule and the biological target of interest (e.g., receptor binding). They may characterize protein atomic coordinates, the relative position of the protein and the interacting small molecule or any other features of the binding site[228].

Descriptors can also be based on experimental data (i.e., biodescriptors[229]). Even if the exact small molecule→target interaction is unknown, these descriptors can capture relevant information of the interaction between the small molecules and their biological targets. Data from HTS-derived concentration-response curves[230] is an example of such descriptors.

Another way of descriptor classification is to consider those that represent some substituent in the molecule, and those that capture some of the properties of the molecules as a single undivided entity.

Multiple descriptor packages from different providers have been used and their performance was evaluated and compared in the course of this work. Below is a brief summary of each package:

ALOGPS is a software program that calculates lipophilicity (presented as a log of the octanol-water partition coefficient) and water solubility. Both properties are important in QSAR modeling, since they implicitly affect many other physicochemical and biological properties of molecules. Therefore, LogP and LogS are of particular interest as molecular descriptors. Most

often, experimentally measured values for molecules used in modeling are not available and are substituted with their predicted values. The underlying algorithm utilizes the ESTATE descriptors and the associative neural networks[231]. Although many other packages include algorithms for calculating these two properties (e.g., Chemaxon calculators package), the ALOGPS algorithm was specifically selected due to its reported higher performance in many studies[232,233]. Throughout this work, version 3 of the software was used[234].

ADRIANA.Code. This package was developed by Molecular Networks and is now part of the CORINA Symphony package[235]. It includes a collection of geometric and physicochemical descriptors (from single dimensional descriptors to molecular surfaces) that are easy to interpret[236]. When combined with linear models or rule-based trees, it can be very useful for understanding the effect of numerous physicochemical and structural factors on the property of interest. ADRIANA.Code can utilize performs empirical 3D optimization of the chemical structures or utilize a pre-optimized representation.

E-State descriptors. Electrotopological state descriptors combine information about the molecules electronic and topological properties[237]. The descriptors are atoms centered and divided according to the atom / bond type. The descriptors include both E-state indices as well as counts (counting atom / bond type of the respective index).

ISIDA Fragmentor utility is a package that calculates the molecular fragments counts (MFC)[238,239]. The algorithm splits molecules into substructure molecular fragments (SMF) of particular size range. The presence of each fragment is then counted in the entire dataset. Therefore, the number of descriptors generated (i.e., the fragments counted) differs based on the underlying dataset analyzed. Furthermore, for each fragment, three subtypes are defined. These are atom types only (A), bond types only (B), or both atom and bond types (AB). This work uses fragments of size range 2 - 5 atoms with AB descriptors.

Dragon descriptors. Dragon is a famous descriptors package developed by the Milano Chemometrics and QSAR research group of Prof. Todeschini and Talete srl[221,222]. The package includes a large collection of 0D-3D descriptors that have proven successful in capturing chemical information relevant to many QSAR studies[240]. The descriptors range in complexity and therefore they interpretability from simple atom types and functional groups counts to complicated indices that are hard to interpret.

Throughout this work, version 6 of the Dragon software[241] was used as integrated into the OCHEM platform. It calculates 4885 molecular descriptors grouped into 29 different blocks such as topological and constitutional indices, atom-type E-states, geometrical descriptors, functional groups and 2D and 3D atom pairs inter alia.

GSFrag descriptors. They are a group of 2D descriptors that count the number of certain fragments (of size ranging from 2-10 non-hydrogen atoms). Such fragments were shown to provide a unique signature for a wide range of chemicals[242]. An extension of this package (GSFRAG-L) also considers the fragments containing labeled vertex and thus captures the effect of heteroatoms.

Inductive descriptors. These descriptors were developed by Dr. Cherkasov resulting from the Linear Free Energy Relationships(LFER)-based equations[243]. They are calculated according to inductive and steric effect models as well as inductive electronegativity and molecular capacitance. They reflect characteristics of inter- and intramolecular interactions. They were used in many QSAR studies to successfully model chemical and biological properties[244].

MERA and MerSy descriptors. MERA is a non-parametric algorithm that covers many 3D descriptors. They can be broadly divided into 4 categories: Geometric descriptors (covering linear and quadratic geometrical descriptors, molecular volumes, ratios on molecular portions as well as symmetry and chirality), energy-related descriptors (Coulomb energy and Van der Waals forces as well as intermolecular decomposition energies) and physicochemical descriptors (such as entropy, heat capacity, association probabilities and pKa)[245–248].

MerSy (MERA Symmetry) package extends the MERA algorithm by providing quantitative estimations of molecular symmetry with respect to certain symmetry axes and inversion-rotational axis. It also quantitatively assesses the molecular chirality according to the negative chirality criteria (i.e., absence of inversion-rotational axes in the molecular point group).

Spectrophore descriptors. These electrostatic molecular descriptors are calculated using the Electronegativity Equalization Method (EEM)[249]. The calculated descriptors give similar results to those calculated using the Density Functional Theory (B3LYP/6-31G*) calculations with high performance in calculation. The algorithm provides a fast method for calculating quantum mechanical descriptors[250]. The descriptors include atomic charges, Fukui functions, hardness and softness among other related descriptors.

Quantitative name-property relationship (QNPR). These descriptors are concerned with directly converting chemical names into descriptors that can be used for predicting target physicochemical or biological responses. The descriptors use either canonical IUPAC names or SMILES representations as input and dissects them into fragments of preconfigured lengths[251].

In this work SMILES representations were used as input. Only fragments of lengths 1-3 characters were concerned with a minimum fragments-count threshold of 5.

Chemistry Development Kit (CDK) descriptors. CDK is an open-source Java library for structural Cheminformatics and Bioinformatics[252]. It is licensed under the GNU Lesser General Public License (LGPL) agreement[253]. Making it friendly to integration into both academic and commercial packages[254]. Therefore, CDK packages gained wide acceptance in the scientific community[255]. CDK includes a descriptors engine that is capable of performing 2D and 3D molecular descriptor calculations. It supports 204 descriptors divided into 6 blocks: topological, electronic, geometrical, constitutional, and hybrid descriptors. Furthermore, it calculates substructure keys including MACCS, PubChem, and E-state keys, molecular fingerprints of 1024 bits based on the Daylight theory[180,255]. CDK also provides a graphical user interface (GUI) for its descriptors calculations. It is also available as a group of KNIME nodes.

In this work, only the structural descriptors (not the fingerprints) were considered. They were calculated using CDK integrated within OCHEM.

Chemaxon descriptors. These descriptors (also referred to as calculator plugins) are developed by Chemaxon Kft[256]. They include a range of physicochemical and biological properties. The descriptors are divided into 7 different groups: elemental analysis, charge, geometry, partitioning, protonation, isomers, and others. In the course of this work, Chemaxon descriptor plugins were integrated into OCHEM and used for building QSAR models.

3.5 Machine learning algorithms

3.5.1 k-nearest neighbors (kNN)

The kNN approach can be used for both classification and regression. In classification, it functions through assuming that class probabilities are approximately uniform within its neighborhood. Therefore, it predicts the new sample's class based on the majority class of its k neighbors. Such assumption might be invalid with high-dimensional datasets. kNN was found to perform better in classification of such datasets than for regression[257]. The parameters to configure are the distance metric (e.g., Euclidean or Manhattan) and the numbers of neighbors to consider (k). The distance metric is usually defined in the descriptor space. kNN works well only in balanced training sets[258,259]. This can be optimized for each dataset by iterating through different values and comparing the errors in prediction. The computational needs of constructing kNN models are typically minimal. This is due to the fact that the model can be fully described by the descriptors matrix of its training set. Therefore, the majority of calculations are deferred to the prediction of new instances.

In this work, the Euclidean distance was used based on normalized descriptors (with a 0 mean and a standard deviation of 1). As the parameter k influences the decision of class membership, a systematic search was conducted in the range (1, 100) to optimize the number of nearest neighbors that provide the highest classification accuracy.

3.5.2 Artificial neural networks (ANN)

ANNs are inspired by the way biological neurons work. Multilayered perceptrons[260] are among the most commonly used NN in which all output from one layer is fed into the input of the next layer. Thus, it can be represented by a directed graph of multiple layers as shown in Figure 13. The final prediction is the output of a single neuron at the last layer of the network.

$$y(x_1 \dots, x_n) = f\left(\sum_{i=1}^{n} w_j . x_i\right)$$ Equation 12

W = $\{w_{ij}, i = 1..L, j = 1..N_i\}$,

where $(x_1 \dots, x_n)$ represent the input (i.e., descriptors or output of the previous-layer neurons), w_i represent the weights of neurons and f is a non-linear response function, L is the total number of layers and N_i is the number of neurons in the i^{th} layer.

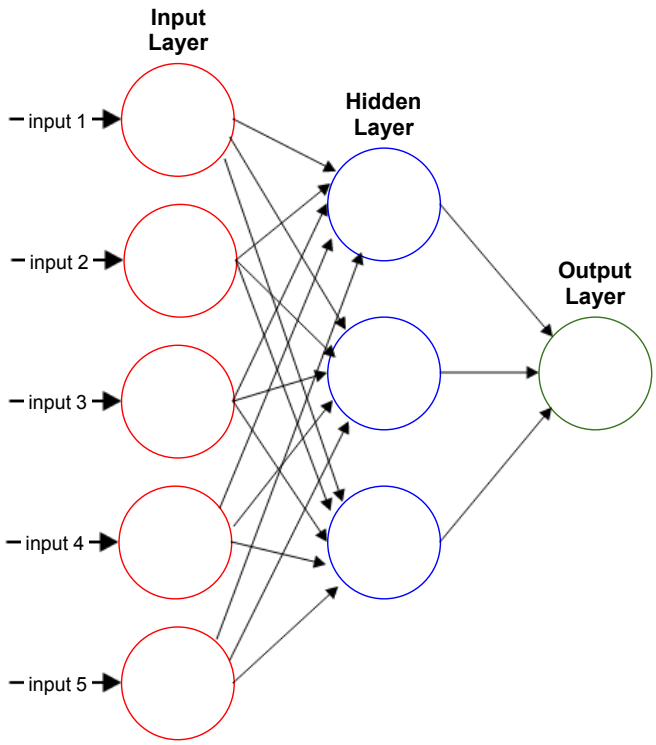


Figure 13. Graph representing a neural network, In its simple form, a neural network consists of 3 layers an input (attaining descriptors), hidden layer (performing operations) and output layer (giving predictions).

As exemplified in Equation 30, the neural network can be completely defined by the set of neural weights applied on the input. Training the neural network (i.e., constructing the predictive model) is the process through which all input weights to the neurons are optimized in order to minimize a predefined cost function. Many training methods exist that trade between the quality of prediction and computational cost such as SuperSAB, Levenberg-Marquardt, momentum, RPROP, QuickProp and differential equation.

In this work, the associative neural networks (ASNN)[261,262] were used. ASNN uses kNN over the space of ensemble predictions. This allows for a local correction for the ensemble of neural networks. The distance is based on the correlation between the vectors of predicted samples by the networks of the ensemble. The configuration of the algorithm was kept to OCHEM/iPrior defaults (i.e., 3 neurons in the hidden layer, 1000 iterations, using model ensemble size of 64, the method for neural network training was SuperSAB[263]).

3.5.3 C4.5 decision tree

C4.5 is a decision tree classifier based on the concept of entropy gain[264]. The tree nodes are optimized to split the molecule sets most effectively between the binary classes. The criterion for this optimization is choosing the descriptor that results into maximum normalized information gain (entropy difference).

As most decision trees, C4.5 trees are built in a top-down manner[265]. At each decision point (node), an attribute is selected that can maximize the separation between instance classes. The tree continues to build until a stop criterion is reached or all instances fall into a single category and therefore creating a leaf. Multiple branches can reutilize the same attribute at

different levels. Decision trees are generally easy to interpret as they can be translated into a rule set. Decision trees can be represented as a connection graph without cycles[266] as shown in Figure 14. The attributes are represented as nodes while the edges represent particular values of the parent attribute.

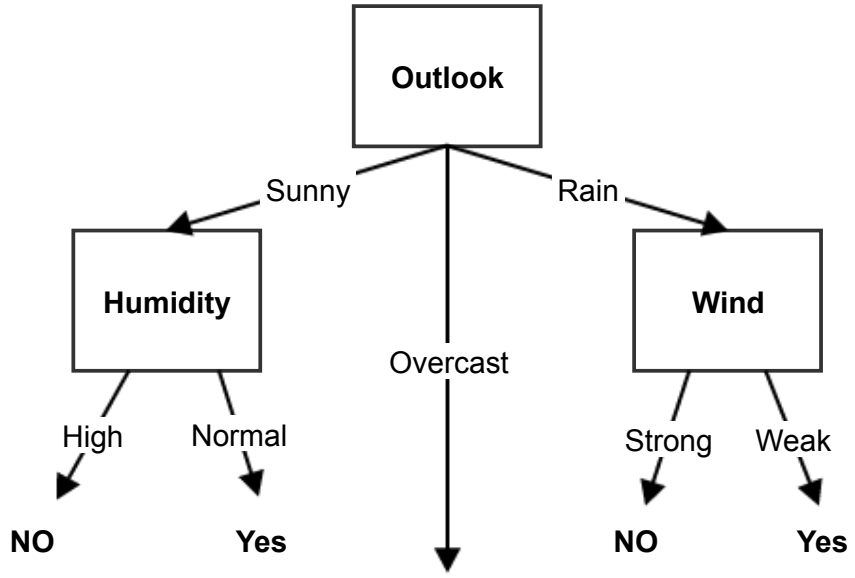


Figure 14. Example of ID3 decision tree on whether to play baseball. Nodes (boxes) perform condition checks while edges (arrows) direct the logic based on the results of such checks.

C4.5 decision trees offer multiple advantages over its predecessor, the ID3 algorithm. It can support tree pruning where the algorithm revises the created tree for unnecessary branches and trims them (i.e., convert them into leaf nodes). C4.5 can also support attributes with different costs. Furthermore, the algorithm can ignore missing attribute values not including them into the entropy calculations. C4.5 can also handle continuous attribute values by setting a threshold splitting the instances into sets based on the value of the attribute being higher or lower than the set threshold (directed discretization).

In this work, a Java implementation of C4.5 decision tree (referred to as J48) in the statistical software WEKA[200,201] was used. The default parameters provided by WEKA were used with no further optimization.

3.5.4 Multiple linear regression analysis (MLRA)

Regression methods detect continuous correlation between the descriptor space and the property to be predicted. It predicts the activity as a function of an optimal linear combination of independent variables (descriptors), which is selected to minimize the training set error as shown in Equation 13

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \varepsilon_i \quad i = 1,2, \ldots, n \qquad \text{Equation 13}$$

where n is the total number of instances, p is the number of independent variables, y_i is the observed response of the ith instance, $x_{i1}, x_{i2}, \ldots, x_{ip}$ are the independent variables of the ith instance and ε_i is the error term.

The purpose of the MLRA algorithm is to estimate the parameters β_0-β_p. MLR is based on the Orthogonal Least Square (OLS) algorithm that minimizes the sum of squares of the error between the predicted $\hat{y}_i$ and the observed values (y_i). The underlying assumption is that the errors are a normally distributed random variable with constant variance. This can generate optimal models when the variables (i.e., descriptors) are unbiased, efficient, and consistent. In which case, according to the law of large numbers[267], the bias and variance approaches zero as the number of instances approach infinity.

Equation 13 is more often represented in the matrix form as shown in Equation 14, which is more convenient in software schemes.

$$\hat{y} = bX \qquad \text{Equation 14}$$

$$b = (X'X)^{-1} X'y$$

Where: b is the vector of estimated parameters
y is the vector of observed responses
$\hat{y}$ is the vector of predicted values
X is the matrix of descriptors

However, studies reported that MLRA is prone to over-fitting[268,269] (misinterpretation due to the use of large number of intercorrelated descriptors). The method is therefore sensitive to collinearity between the independent variables. Thus, prefiltering of correlated descriptors is necessary to remove insignificant coefficients and reduce the risk of multi-collinearity.

In this work, the MLRA method used stepwise variable selection. It eliminated a single descriptor on each step. The descriptor was selected for elimination based on the t-test when its regression coefficient insignificantly differed from zero. The only parameter in this method is the p-value (ALPHA) according to which variables would be preserved for the regression (p=0.05 was used).

3.5.5 Fast stagewise multiple linear regression (FSMLR)[270]

FSMLR constructs linear regression models by using greedy descriptor selection. It is a particular kind of the regression boosting (additive regression) procedure specially intended to utilize a three-set approach. It uses three different subsets for learning: training set, internal tuning/validation set as well as the external test set. The internal set is used for determining the optimal number of descriptors considered in the model. In this framework, an error vector is calculated from the experimental measurements of the training set-compounds. Then, the descriptor with highest correlation to this vector is added to the set of selected descriptors its corresponding regression model is used for recalculating an error vector, which will be used in the next cycle to select the subsequent descriptor, and so on. Such iterative descriptor selection process and the corresponding model formulation continues until the minimal prediction error for an internal test set can be achieved.

3.5.6 Partial least squares (PLS)

PLS is another powerful statistical approach that addresses the problem of intercorrelation of variables and the singularity of X'X due to large number of variables as compared to instances

in MLRA (Equation 14). PLS is also known as principal component analysis (PCA) regression as it combines both PCA and linear regression. While PLS involves orthogonal transformation to the variable space, it differs from PCA in its ability to rank the components not only by their variance in the variable space but also by their correlation to the target property vector. PLS decomposes the variables matrix (X) into orthogonal scores T and loadings matrix P (called outer relationship; Equation 15) while decomposing the observations (Y) into the score matrix (Z) and the loading matrix (Q) as shown in Equation 16. Finally, the two scoring matrices (U and T) are correlated using a transformation matrix (B) (called inner relationship; Equation 17) which can be thought of as a regression model relating both scoring matrices.

$$X = T \cdot P' + E \qquad \text{Equation 15}$$

Where E is an error term

$$Y = U \cdot Q' + F \qquad \text{Equation 16}$$

Where F is an error term

$$U' = B \cdot T' \qquad \text{Equation 17}$$

In PLS, the components are referred to as latent variables (LVs). They are calculated by singular value decomposition (SVD), decomposing the cross product of the variables as shown in Figure 15. PLS has the advantages of being efficient in calculation and preserving the linearity and therefore can support mechanistic interpretations. PLS can also be used to predict multivariate responses[271]. PLS has been widely used in QSAR studies and covered in OECD QSAR validation guidelines.

In this work, a 5-fold cross-validation protocol (on the training set) was used to automatically optimize the number of latent variables.

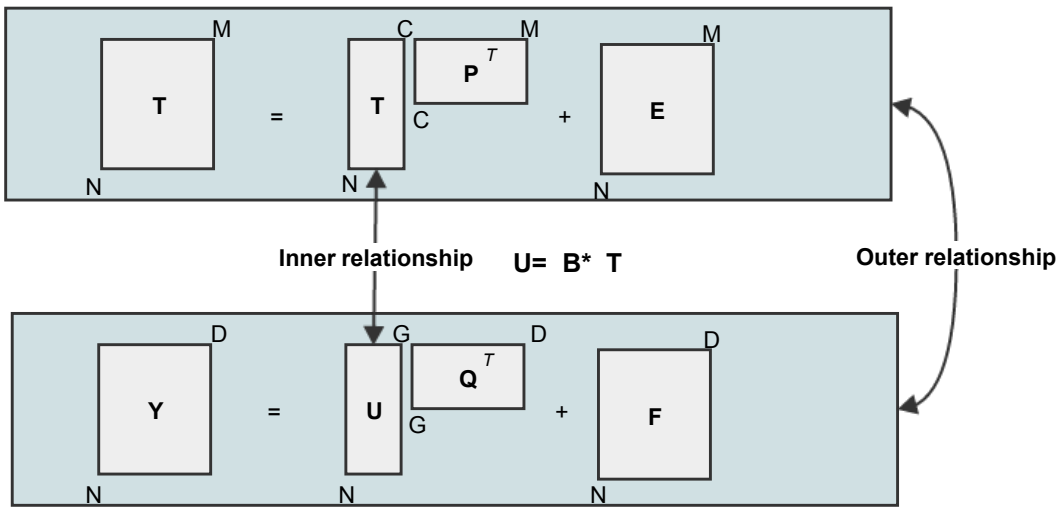


Figure 15. The PLS analysis decomposes the descriptors matrix (X) as well as the target property (Y). The score matrices (T and U) are related in order to keep the orthogonal transformation

3.5.7 Random trees / random forests (RF)

RF[272] is another example of a decision tree algorithm. The random trees are constructed by choosing random m variables at each node, from the entire variables domain M, to form the decision at that node. The tree is based on a randomly selected subset of the entire training set of N instances. The subset is selected through bootstrap sampling (i.e., sampling N times with replacement) and therefore has the same size as the original dataset.

Random trees are often used in combination with a meta-learning method (bootstrap aggregation)[266]. Multiple trees are created and the final class membership results from the consensus voting of individual trees.

In this work, the algorithm used was a Java implementation of random forests in the statistical software WEKA[200,201]. Each RF model was constructed with ten trees As the RF approach was combined with bootstrap aggregation (64 sets), each QSAR model consisted of 640 trees. The default parameters provided by WEKA were used with no further optimization.

3.5.8 Support vector machines (SVM)

SVM[273] was developed by Vapnik to address classification problems[274–276]. It was initially a linear non-probabilistic binary classification method. It locates a hyper-plane that can separate the multi-dimentional data into two-classes. Placing such a hyperplane considers maximizing the distance to the adjacent data points for both classes (such distance is referred to as "functional margin") as shown in Figure 16. When such a hyperplane that can classify all training examples correctly cannot be found, a 'soft margin' approach is proposed. It allows misclassification of some instances while still tries to maximize the margin to the examples correctly classified[277]. This introduces extra parameters to measure and penalize for the misclassification of examples. The approach has also been generalized to multi-class classification by employing multiple hyperplanes.

However, on complex problems, dependence could be frequently non-linear. Therefore, a non-linear transformation was introduced through the means of utilizing a kernel function.

This allows the transformation of original data into a higher dimensional space that can separate the data. Many kernel functions were developed, among the most widely used are linear, sigmoid, polynomial as well as radial basis functions (RBF). The SVM technique was also extended to address regression problems. The linear model in the high-dimensional space can be described per Equation 18.

$$f(X, w) = \sum_{j=1}^{p} w_j g_j(X) + b$$ Equation 18

where g_j ($\mathbf{X}$), j = 1, ... , p represent a set of nonlinear transformations and b is the bias term.

The approach is intuitive and well established. However, many computational challenges could arise. There is also the risk of over-fitting the training sets. To address these difficulties, SVM introduces many parameters, other than the kernel type. The most important parameters, namely the cost constant 'C' and the parameter ϵ, are discussed in this section.

The parameter C, controls the degree of data fitting to the model. Its value is typically set between 0 and 10. The correct value is a question that is only answered relative to the dataset. With a too-high value the model runs into the risk of over-fitting the training dataset. On the other hand, setting its value too low risks unsatisfactory fit and inability to extract relevant information during model training. Therefore, it is necessary to adjust the C parameter relative to the degree of noise in the dataset through implementing appropriate fitting and validation techniques (e.g., using a grid search with cross-validation method).

The parameter ϵ is the trade-off between maximizing the functional margin and minimizing the error rate. It controls the number of support vectors (SVs) with the value of the epsilon parameter being inversely proportional to the number of selected SVs.

In this work, the LibSVM implementation was used. Multiple configuration options are supported for the SVM type (either epsilon-SVR or nu-SVR or one-class SVM) and the kernel type (linear, polynomial, radial basis function, and sigmoid). The Classic epsilon-SVR and the RBF kernel were used. A grid search was performed as specified in the LibSVM manual to optimize the cost “C” and width parameters of the RBF kernel.

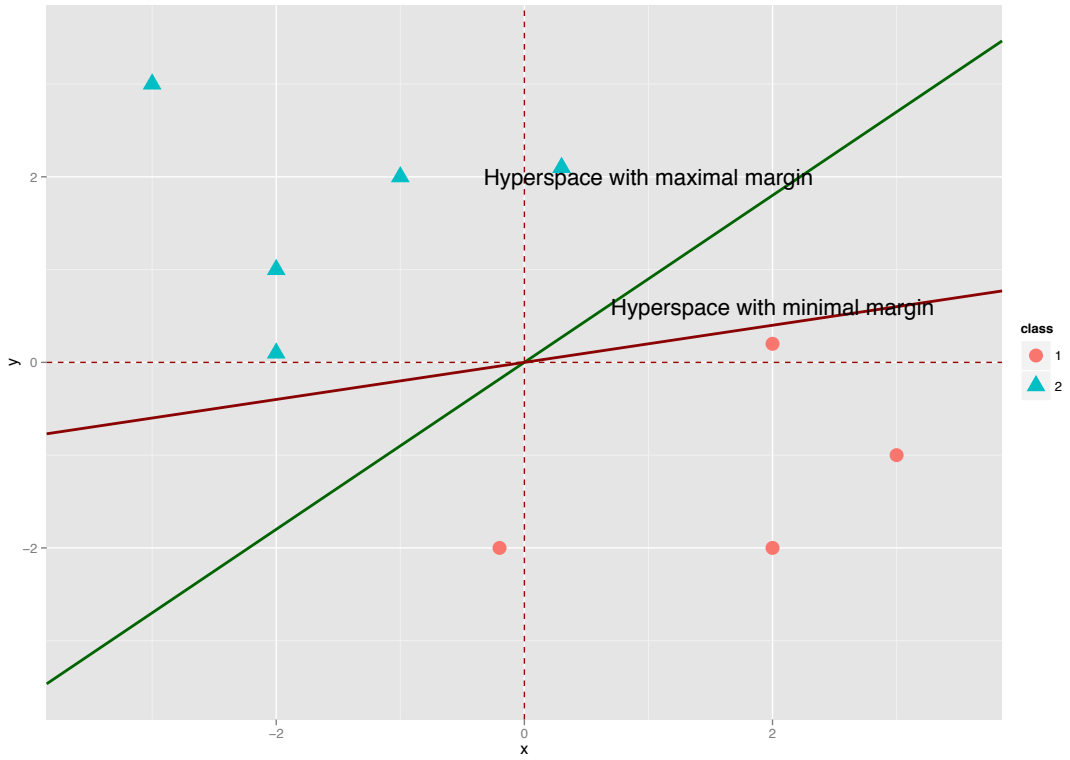


Figure 16. Maximizing the functional margin in SVM hyperplane selection. The hyperspace with maximal margin (green) is preferred for the separation between the two classes.

3.6 Variable selection

As discussed earlier, variables (i.e., descriptors) with low variance can reduce the performance of distance based machine-learning algorithms. Also, linear regression performance suffers as the number of descriptors, when compared to instances, increase. Therefore, in this work, all descriptors were pre-filtered prior to model development. The following pre-filtering criteria were used: first, descriptors that are constant among all compounds, offering no information gain, were removed. Then, normalized descriptors that have variance smaller than < 0.01 were removed. Finally, descriptors were grouped if they showed pair-wise Pearson's correlation coefficient (R) > 0.95. The same pre-filtering steps were applied to all descriptor packages and machine-learning algorithms.

Some algorithms implemented automatic scaling as part of their training protocols, i.e., in ASNN the descriptors were scaled to [0,1] interval, in MLRA and kNN the variables are normalized to zero mean and unit variance. For other algorithms, descriptors were scaled to the range [0-1] prior to the application of the algorithms. The prefiltering step was also applied within the bootstrap aggregation protocol for any of the algorithms used. Thus, the exact numbers of descriptors used could be different for each model.

3.7 Goodness of fit and prediction

The ability of QSAR model to derive predictions on new chemicals not previously measured is very valuable. However, the quality of such predictions must be rigorously checked. This has also been highlighted through the fourth OECD principle regarding the use of QSAR models in hazard assessment of chemicals[278] (see section 1.5.2).

Multiple statistical indices can be used to assess the quality of a QSAR model[222]. When such parameters are used to examine the model's training set, they measure the goodness of fit. Alternatively, when applied on a set of compounds that were not included in the model construction process, these parameters measure the model Predictivity (i.e., the model's ability to predict properties of new chemicals). The most relevant statistical indices to this work are discussed in this section

Classification parameters

Multiple statistical indices were described in literature to measure the quality of a classification model[222,279]. For a 2-class classification problem, the possible outcome of a QSAR model prediction can be described using the confusion matrix as shown in Table 1. The output of a prediction (Positive (class 1) or Negative (class 2)) can either match the observed class (True) or not (False).

Table 1. The confusion matrix for a 2-class classification problem. It shows all possible outcomes of a classification model. The table also lists some statistical parameters that were used for judging the quality of the QSAR models throughout the work.

	Experimental measurement	
Predicted outcome	True positive (TP)	False positive (FP)
	False negative (FN)	True negative (TN)
Sensitivity (SN) = TP / (TP + FN) Specificity (SP) = TN / (TN + FP) Accuracy (ACC) = (TP + TN) / (TP + FP + TN + FN) Balanced accuracy (BA) = (sensitivity + specificity) / 2 Matthews correlation coefficient (MCC) = (TP * TN – FP * FN) / [(TP + FP) (TN + FP) (TP + FN) (TN + FN)]$^{1/2}$		

3.7.1 Sensitivity

Sensitivity measures the model's ability to correctly predict a positive outcome. It is also known as the True Positive Rate (TPR) or recall and can be calculated per Equation 19.

$$Sensitivity = \frac{TP}{TP + FN}$$ Equation 19

3.7.2 Specificity

Specificity measures the model's ability to correctly predict a negative outcome. It is also known as the True Negative Rate (TNR) or precision and can be calculated per Equation 20. A complementary metric is called the False Positive Rate (FPR) or fall-out and can be calculated per Equation 22.

$$Specificity = \frac{TN}{TN + FP}$$ Equation 20

$$FPR = 1 - specificity = \frac{FN}{TN + FP}$$ Equation 21

3.7.3 Total accuracy (ACC)

Total accuracy shows the portion of the instances that were correctly assigned to their respective class. (i.e., an accuracy of 0.7 means that the model could assign 70% of the cases to their correct classification). However, ACC does not take into account the ratio of the classes' size. It can be calculated per Equation 22.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$ Equation 22

3.7.4 Balanced accuracy (BACC)

Total accuracy can be misleading when dealing with unbalanced dataset where the distribution of instances differs widely between both classes. Models with poor performance can be disguised through bias towards the majority class. Therefore, balanced accuracy is more relevant in assessing such datasets. It considers both sensitivity and specificity equally and can be thought of as the average of sensitivity and specificity per Equation 23. It is also known as the non-error rate (NER) and is often expressed in percentage form (NER%). A complementary metric is the Error Rate (ER%), which can be calculated as shown in Equation 24.

Since it is common for QSAR problems to be presented in unbalanced classes (e.g., Small number of active compounds against certain target), the balanced accuracy is often the metric of choice for optimizing predictive models' performance.

$$Balanced\ Accuracy = \frac{Sensitivity + Specificity}{2}$$ Equation 23

$$ER\% = 100 - NER\%$$ Equation 24

3.7.5 Positive predictive value (PPV)

The positive predictive value (PPV) measures the proportion of positive results that are truly positive and can be calculated according to Equation 25. A complementary metric that measures the negative instances that were misclassified as positive is called the false discovery rate (FDR) and can be calculated per Equation 26.

$$PPV = \frac{TP}{TP + FP}$$ Equation 25

$$FDR = 1 - PPV = \frac{FP}{TP + FP}$$ Equation 26

3.7.6 Negative predictive value (NPV)

The negative predictive value (NPV) measures the proportion of negative results that are truly negative and can be calculated according to Equation 29. A complementary metric that measures the negative instances that were misclassified as positive is called the false omission rate (FOR) and can be calculated per Equation 30.

$$NPV = \frac{TN}{TN + FN}$$ Equation 27

$$FOR = 1 - NPV = \frac{FN}{TN + FN}$$ Equation 28

3.7.7 Matthews correlation coefficient (MCC)

Matthews correlation coefficient (MCC) can be considered a correlation coefficient between the observed and predicted outcomes of the classification[280]. Therefore, it takes values between -1 and +1. Whereas a +1 represents a perfect prediction (i.e., all instances were correctly assigned to their classes), a value of zero represents lack of correlation between the observed and predicted values (i.e., equivalent to random predictions). An MCC value of -1 represents a perfect inverse correlation where all instances were assigned to the opposite class. The MCC is another measure that is suitable for use in unbalanced datasets as it considers all the true and false positive and negative predictions. It can be calculated according to Equation 29:

$$MCC = \frac{TP.TN - FP.FN}{\sqrt{(TP + FP).(TP + FN).(TN + FP).(TN + FN)}}$$ Equation 29

3.7.8 Area under the receiver operating characteristic curve

The receiver-operating characteristic (ROC) is a graphical plot that describes the variance in the discrimination power of a binary classification model. The curve is created by plotting the models' sensitivity against its fallout (false positive rate) at various threshold levels Therefore, The ROC curve describes the sensitivity as a function of fall-out as shown in Figure 17.

The area under the ROC curve (AUROC) has long been used in model comparison in machine-learning[281]. It can be calculated by averaging multiple trapezoidal approximations. Assuming that an 'active' compound is ranked higher than a 'negative' one, AUROC can be represented as the probability that a given classifier ranks randomly selected active compound higher than a randomly selected inactive one when using normalized units.

AUROC has the advantage of considering all cases of false and true positive and negative predictions and therefore can be used with unbalanced datasets. However, it is difficult to use a single number to summarize the entire ROC curve. Such a number loses information on the performance pattern of the underlying model that can be gained from the curve itself. Studies have suggested that using AUROC as a classification metric can be noisy and misleading[282,283] especially in datasets with small sample sizes[284].

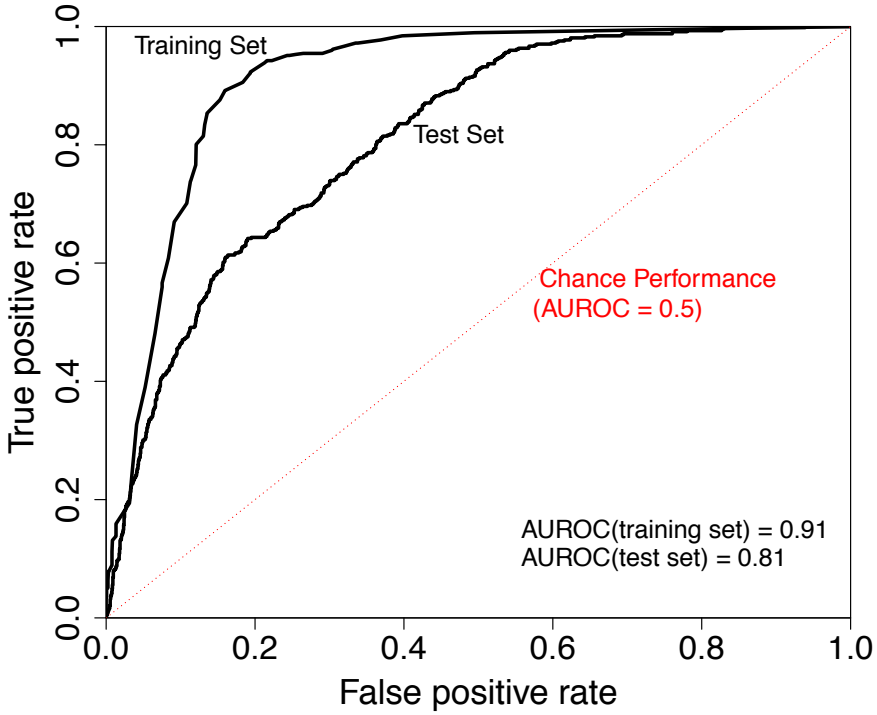


Figure 17. Receiver Operating Curve (ROC) showing a classification's model sensitivity as a function of the fallout (1-specificity). The performance of two hypothetical datasets are shown, a training set in blue and a test set in red.

3.8 Models comparison

As discussed in the previous sections, multiple criteria exist for judging on the quality of QSAR models. Two main aspects need to be considered for comparing multiple models based on these statistical parameters. First, which ones to use (e.g., If Model A has a higher balanced accuracy than model B, this does not necessarily guarantee that it will have a higher area under the ROC) and second, whether the difference in the selected parameter(s) is significant and not due to chance correlation. In this section, the multi-criteria decision making and the statistical significance are discussed.

Multi-criteria decision making (MCDM)

Certain criteria of the dataset or the machine-learning method can limit the suitability of available parameters. Although most feature selection techniques optimize only one parameter (for example: balanced accuracy), many parameters could be simultaneously considered to achieve a compromise between predictive ability and model complexity. Multi-Criteria Decision Making (MCDM) methods were developed to address this challenge. These methods considers multiple statistical parameters by utilizing the concepts of utility indices and desirability to perform a multivariate ranking to optimally compare multiple models[285,286].

The utility is calculated as the arithmetic mean of the parameters under consideration while the desirability represents the geometric mean of such parameters. The utility (U_i) for each non-weighted alternative parameter (i) can be calculated per Equation 30. Where some parameters are thought to be more important than others, weights can be used. Utility is calculated for weighted parameters using Equation 31

$$U_i = \frac{\sum_{j=1}^{p} t_{ij}}{p} \qquad 0 \leq U_i \leq 1$$ Equation 30

$$U_i = \sum_{j=1}^{p} w_j t_{ij} \qquad 0 \leq U_i \leq 1$$ Equation 31

Where: p is the number of criteria t

The desirability (D_i) for each non-weighted alternative parameter (i) can be calculated according to Equation 32 while for the weighted parameters using Equation 33

$$D_i = \sqrt[p]{t_{i1} t_{i2} \ldots\ldots t_{ip}} \qquad 0 \leq U_i \leq 1$$ Equation 32

$$D_i = t_{i1}^{w_1} t_{i2}^{w_2} \ldots\ldots t_{ip}^{w_p} \qquad 0 \leq U_i \leq 1$$ Equation 33

The weight constraint is given by Equation 34:

$$\sum_{j=1}^{p} w_r = 1$$ Equation 34

The weights are calculated using the method of normalized weights for ranked criteria as shown in Equation 35[287,288]:

$$w_j = \frac{Q/r_j^k}{\sum_{j=1}^{p} Q/r_j^k}$$

$$Q = \prod_{j=1}^{p} r_j^k = \exp\left[\sum_{j=1}^{p} k \ln (r_j)\right]$$ Equation 35

Where:
r_j: criterion rank
k : smoothing parameter

Statistical significance

When comparing the performance of two models (e.g., built using different descriptor packages or machine learning algorithms) using the same statistical parameter (e.g., balanced accuracy), the significance of the difference in the values of such parameter should be checked (i.e., Whether 81% is truly higher than 80% on the given validation set or whether that could be due to a random chance).

The classical hypothesis testing can be used for checking this statistical significance (the difference in the model performance). Two hypotheses are defined; the null hypothesis (H_0) is the case that both QSAR models have similar performance. While the alternative hypothesis

claims that the models in questions has a difference in performance that cannot be attributed to chance. Then, the probability of the null hypothesis is calculated (referred to as the p-value) given the dataset at hand. The p-value is compared to pre-defined levels (usually 0.05 is used). Whenever the probability is lower than that level, the null-hypothesis is thought to be improbable and thus is rejected in favor of the alternative hypothesis. The H_0 is accepted when p-value is higher than the predefined threshold and therefore the QSAR models would be thought to have similar performance.

Many tests can be used to measure p-value. They are generally divided into parametric and non-parametric tests. Parametric tests (such as student's t-test) assume certain statistical distribution for the data (usually, the normal distribution) and calculate p-value as a function of the statistics distribution parameters. The non-parametric tests (such as Wilcoxon test or bootstrap test) involve fewer assumptions on the data and thus are more unanimous.

Throughout this work, a p-level of 0.05 was used to indicate a significant difference. When a p-value of 0.001 was used, it was referred to as highly significant. The non-parametric bootstrap test was used to calculate p-values. It involves resampling the original test set with replacement for N times (N = 1000 was used). The models in comparison were applied to all generated bootstrap sets and the statistical parameters of choice (e.g., balanced accuracy or AUROC) were calculated and compared for both models in a pairwise fashion. For a model to be superior than the other in a statistically significant manner, it needs to show a better value for the metric in question in 95% of the comparisons (p=0.05).

3.9 Model validation

3.9.1 External validation

Challenging a QSAR model with a new set of chemical compounds can be the ultimate test for model's predictivity. Prior to constructing a QSAR model, the data pool (with m number of instances) is divided into two separate sets (training and test sets). The training is bigger (typically 75-80%) and is used to develop the QSAR model. The smaller test set (usually 20-25%) is used to evaluate the model's predictive ability. The test set should not be included in any model calibration or descriptors selection processes (i.e., completely neglected during model construction).

The external validation approach is suitable only for big datasets. It has the disadvantage of not fully utilizing the data at hand to maximally feed the QSAR model with all available information. Furthermore, an important consideration is the procedure of data split. While a random split is suitable for a sufficiently large dataset, smaller datasets (e.g., <50 instances) are more prone to selection bias. For instance, in a classification problem, significantly more instances of certain class can fall in a particular set and therefore negatively affect the model's performance.

3.9.2 Cross-validation (CV)

For avoiding any information loss in the form of instances not utilized in model training and cope with smaller datasets, a cross-validation approach can be used. Data is divided into multiple portions checking one another. Two famous types of the cross-validation techniques exist; namely, the n-fold CV and the leave-one-out CV.

A. n-fold cross-validation

Data is partitioned into n folds. The folds are normally chosen to have equal sizes. The number of folds to choose has been investigated in literature and is a trade-off. For a large n, each training sample used in n-fold cross-validation has size m−m/n = m(1−1/n) (illustrated by the right vertical red line in Figure 18), which is close to m, the size of the full sample, but the training samples are much similar. Thus, the method tends to have a small bias but a large variance. In contrast, smaller values of n lead to more diverse training samples but their size (shown by the left vertical red line in Figure 18) is significantly less than m, thus the method tends to have a smaller variance but a larger bias. In most QSAR studies 5 or 10 folds are considered. The QSAR model fitting process is repeated n times. During each iteration, one fold is considered as a validation set and is excluded from the fitting process while the remaining labeled instances are combined and used as a training set. Then, the resulting QSAR model is used to predict the labels of the left-out n-fold and performance statistics are calculated. The process is repeated for n iterations as shown in Figure 18. The final model quality statistics are calculated using the validation folds. Usually, the n folds are portioned using contiguous blocks or venetian blinds methods:

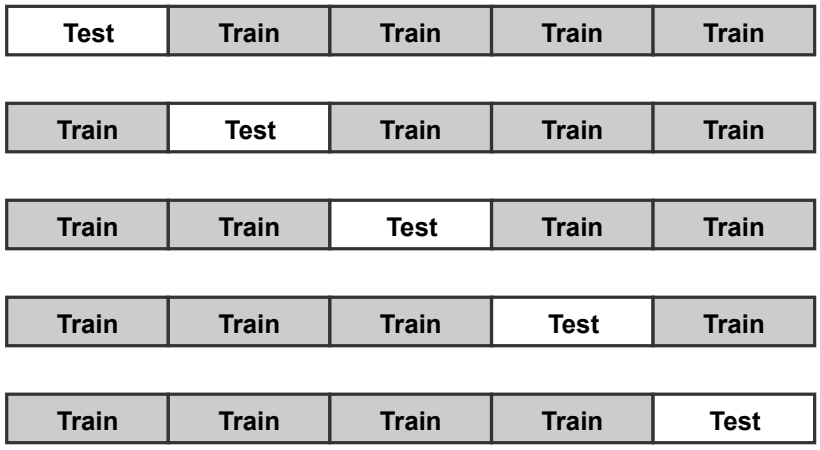
- in venetian blinds technique, each n-th instance of the dataset is selected for the test set, starting at the first sample.

- in contiguous blocks, the test set is constructed by selecting m/n samples from the dataset, starting at the first sample.

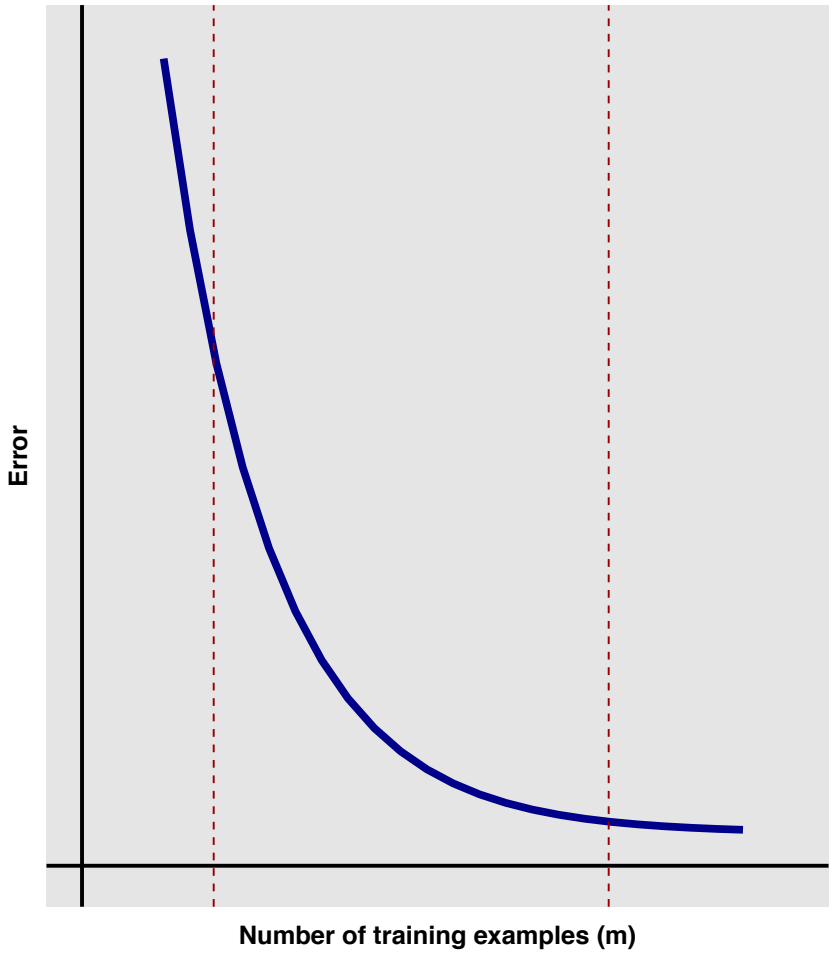
B. Leave-One-Out cross validation (LOO)

LOO can be considered as a special case of the n-fold cross validation where the number of folds (n) is equal to the number of instances (m). In LOO, each molecule is excluded once from the model training, while the remaining molecules are used to fit a QSAR model that predicts the outcome of the excluded molecule. This process is repeated m times. The final model quality statistics are calculated considering only the predicted left-out molecules.

The disadvantage of the LOO approach has been shown to result in overly optimistic prediction statistics as it only omits one compound at a time[289]. A better approach has been proposed which omits many compounds at a time, thus called Leave-Many-OUT (LMO)[290]. Despite its robustness, LMO is computationally expensive. Moreover, due to the random nature of the selection process, its results are irreproducible. LOO is also computationally expensive, as it requires the calculation of m number of QSAR models each with a size m-1. It is usually left for smaller datasets.



(a)



(b)

Figure 18(a) Illustration of the partitioning of the training data into five folds. (b) Typical plot of a classifier's prediction error as a function of the size of the training sample: the error decreases as a function of the number of training points.

C. Stratified cross validation

Stratified cross validation is another special case of the n-fold cross validation technique. It is used in classification problems where the folds are selected to include an equal distribution of the target property classes. It is useful for avoiding the difficulties of unbalanced datasets with machine-learning algorithms. For each of the folds, equal number of positive and negative instances are randomly selected. The number of compounds per fold cannot exceed double the size of the minority class.

3.9.3 Bootstrap aggregation (Bagging)

Bagging[272] is a meta-algorithm that involves the random sampling, with repetition, of many subsets of the original dataset (with 'n' number of instances). A predefined number of iterations are executed. In each iteration, 'n' training examples are selected randomly from the original dataset allowing duplicates (i.e., resampling with replacement) and a QSAR model is fitted. The model is used to predict the outcome of the validation set (i.e., out of bag compounds). Thus, bagging validation creates an ensemble of models for each bagging meta-model constructed. Multiple prediction outcomes exist for each molecule. The final prediction outcome is determined by voting among the ensemble of models that predicted the given molecule. The final meta-model quality statistics are calculated based on the final prediction for the instances.

The number of iterations (bagging folds) is usually selected to be at least 32 to ensure that each example is represented at least once in the validation set. Given the probability of selecting the same instance multiple times, the probability of a given example to fall into the training set is approximately 63.2% (in each iteration) as shown in Equation 36. Subsequently, the probability of not being selected in the training set (i.e., becomes part of the validation set) is approximately 36.8% as shown by Equation 37.

$$probability\ for\ training\ set\ selection = 1 - e^{-1} \approx 63.2\% \qquad \text{Equation 36}$$

$$probability\ for\ validation\ set\ selection = \ e^{-1} \ \approx 36.8\% \qquad \text{Equation 37}$$

Throughout this work, stratified bagging[291] was used as a validation protocol. It also served to handle the unbalance of the training set[292292]. An ensemble size of 64 bagging folds was used.

Finally, regardless of the validation protocol in use, it is essential to remove duplicates during data curation as discussed earlier. The presence of a duplicate instance in both training and test sets simultaneously can affect the statistical parameters of model quality. Leading to the fitting quality of some compounds to be reported as predictions and therefore resulting in over optimistic predictivity measure for the QSAR models.

3.10 Models applicability domain (AD)

The statistical metrics for judging model's performance, as discussed in section 3.7 Goodness of fit and prediction above, are evaluated during the construction of such model. However, they cannot guarantee the suitability of the model for the infinite chemical space. Furthermore, these metrics comprises an average of the model's performance. On the other hand, Chemicals that are more similar to the majority of the training set (interpolation space) are expected to perform better than that average while those that are further from it

(extrapolation space) will perform worse. This has also been highlighted through the third OECD principle (see section 1.5.2) regarding the use of QSAR models in hazard assessment of chemicals[278]. A well-defined AD is essential before a model can be considered validated. The description of the model's AD is therefore important to allow regulators to assess whether the provided prediction falls within the range of reliability described by the model. Many approaches for assessing the AD have been reported in literature[152,293]. A general overview of such approaches is discussed in this section.

Different AD algorithms try to estimate the interpolation space for the model's training set. The efficiency of such algorithms can be estimated based on their ability to maximize the retention of true predictions while rejecting false ones. In classification models, the algorithms try to maximize the allocation of test molecules to their true classes. While in case of regression, the aim is to lower the prediction error.

Algorithms can define the AD in a variety of ways. For instance, it can be defined based on the model's descriptor space. In this case, a certain test compound is said to fall inside the model's applicability domain (i.e., interpolation space) if its descriptor values match certain criteria. Another approach is defining a mechanistic AD; where a test molecule falls into the interpolation space if its mode of action matches that of the training set compounds. A metabolic AD can be defined based on the possibility of chemical substances being metabolized or undergo certain transformation[1,294].

The four major categories of AD definitions in the descriptor space are: range-based methods, geometric methods, distance-based methods and Probability Density Distribution based methods[152,293,295,296]. The range-based methods include the bounding box method, which considers individual descriptors used for model building. The bounding box interpolation space is defined by the minima and maxima of all descriptors' values. This method cannot identify the empty spaces within the interpolation space, which is a major drawback[293,297]. The PCA bounding box is another range-based method that considers the projection of chemical structures in principal component space. The method takes into account the minimum and maximum values of the principal components rather than the individual descriptors. Because descriptors are projected as principal components, this method overcomes the problems of intercorrelated descriptors. It still however is not able to define the empty areas within the interpolation space.

The geometric approaches include the Convex Hull method, which defines the applicability domain based on the smallest convex region enclosing the entire training set. This approach is typically limited to QSAR models with small number of descriptors due to the challenges in implementation on higher complexity datasets (e.g., above three dimensional data)[298].

The distance based methods estimate the distance of the molecule to be predicted to a certain point such as the training set centroid (called centroid-based distance approach) or from the nearest *k* molecules of the training set (referred to as K-nearest neighbors based approaches). Such distance is then compared with a certain threshold, above which the compound is considered outside of the model's interpolation space. The definition of such threshold is user-defined with no obvious rules. This has the disadvantage of not necessarily reflecting the data density[152,293,295,297]. Another approach is to define the distance to model in the property

space (rather than the descriptor space)[196]. This approach uses the standard deviation between the predictions of an ensemble of models (generated through bagging or an ensemble of neural networks). Other than standard deviations (STD), other measures exist, such as the correlation of prediction vectors of ensemble models for a certain compound (CORRELL). In binary classification, the difference between the numeric value predicted by the machine-learning algorithm and the class's value can also be used as a measure of prediction certainty by the model. For example, if 0 represents class A and 1 presents Class B, a prediction value of 0.6 is thought to be less certain than a value of 0.9 although both are classified as Class B prediction. This method is referred to as the rounding effect (CLASS-LAG). A combination of multiple methods is also possible. The prediction based distance to model for applicability domain estimation has the advantage of overcoming problems of activity cliffs where small differences in descriptor values result in large difference in activity.

The probability Density Function methods are based on estimating the Probability Density Function (PDF) of the training set descriptors using parametric or non-parametric methods. It has the advantage of being able to identify empty regions in the interpolation space[293].

Although multiple approaches exist for defining the model's applicability domain, no certain method was officially accepted or recognized by authorities[299]. The increasing awareness of AD is important for increasing the confidence in adopting QSAR and alternative approaches for regulatory purposes. In this work, the distance-to-model in the prediction space was used to assess the applicability domain of models. The STD between the predictions of ensemble models was used as the distance measure, unless otherwise stated.

4 QSAR case studies for risk assessment and computational modeling of datasets

4.1 ToxCast™ phase I

The ToxCast™ project was covered above (see 1.6 Role of *in vitro* assays in alternative testing). The following analysis was conducted using the Phase I data from the project. The primary aim of the study was to examine the possibility to directly predict animal toxicity outcomes using chemical structures, *in vitro* assay response panels or a combination of both. The second aim was to explore whether the outcome of the *in vitro* assay panel can itself be predicted from the chemical structures.

4.1.1 Data setup and curation

A. In vitro toxicity assays datasets

The *in vitro* screening data from ToxCast Phase I "Toxminer v17" MySQL dump was downloaded, reconstructed and integrated into a dedicated instance of OCHEM called iPrior[186], using the KNIME workflow tool.

The downloaded database includes, in addition to the screening data, biochemical pathways and processes correlated to the target toxicological endpoints, together with assay-gene, and gene-pathway mappings. The genes and pathways correlations were gathered from different sources such as Gene Ontology (GO)[300], pathway commons[301], Ingenuity Pathways analysis (IPA, Ingenuity systems Inc.)[302], Kyoto Encyclopedia of Genes and Genomes (KEGG)[303] and the OMIM[304] phenotype databases. It also includes the chemical structure files (in SDF format) of all ToxCast phase I chemicals (309 compounds). The *in vitro* screening data covers 467 assays, some of which include multiple time-points, resulting in a total of 669 assay endpoints. These assays test both direct interactions between chemicals and identified receptors and enzymes, as well as downstream events on receptor gene activity or cellular consequence. They cover nine technologies: cell-free HTS assays; multiplexed transcription reporter; biologically multiplexed activity profiling; high-content cell imaging; multiplexed gene expression; cell-based HTS; phase I and II XME cytotoxicity; real-time cell electronic sensing; and HTS genotoxicity. The response of the chemicals to *in vitro* assays varies across categories. The distribution of the ToxCast phase I chemicals regarding the number of positive *in vitro* assay hits (out of the total 669 measured) and the corresponding number of positive pathway perturbations calculated is shown in Figure 19.

ToxCast™ database included either the calculated half maximum activity concentration (AC_{50}) or lowest effective concentration (LEC) in umol/L for each assay/compound combination but not both. Analysis of the AC50/LEC ration was therefore not possible. This study only considered classification models in an effort to compensate for the limitations due to the small size of the screened chemical set as well as the uncertainties typically associated with HTS experimental accuracy. Successful classification models can encourage the future exploration of the data using regression models.

To construct binary classification models, all screening assay outcomes were discretized to (response/no response) factors. The absence of response per certain assay threshold was

reported with a flag value of 10^6 umol/L in the original database. The same value was used as the discretization threshold "no response" while any other reported value was considered positive "response", i.e., maintaining the original threshold of the ToxCast assays.

Roughly, only 7% of the full assay/chemical interaction matrix presented a positive response. An alternative approach was considered for consolidating assays data. The ability of a chemical perturbing a given pathway, regardless of the exact gene affected in such perturbation was examined. Therefore, 1456 pathways were correlated to chemicals in order to construct chemical-pathway perturbations. The pathways-genes correlations were examined to detect whether a compound would show activity towards any assay associated with these genes. Such compound would thus be considered perturbing the investigated pathway. Subsequently, a chemical/pathway-perturbation matrix was constructed where 14% of the potential interactions were positive. Because single assay might correlate with several pathways and vice-versa, the pathway perturbations represent a different re-grouping of the *in vitro* assay screening data. Such regrouping resulted into less sparse datasets (as shown in Figure 20) and, therefore, may be better suited for applying machine learning algorithms. The study investigates whether such regrouping can improve the predictive models.

Although screening outcomes for all 669 *in vitro* assay endpoints were available for all ToxCast Phase I compounds, most of the assays showed only few active hits. *in vitro* assays with less than 35 hits among tested compounds were filtered out for having too few data for a statistical QSAR approach. Only 144 assays showed sufficient active hits and thus were modeled using QSAR. The list of selected *in vitro* assays is provided in the supplementary materials (Supplementary 1: List of *in vivo* endpoints from ToxCast / ToxRefDB, their respective total number of hits and whether it was selected for modeling.). The description of the *in vitro* assays methodology is available from EPA[182].

B. In vivo animal studies dataset

The Toxicity Reference Database (ToxRefDB), part of the ACToR system, contains summary outcomes from primary toxicological studies presented to the EPA for pesticides' active ingredients[305]. These data were gathered from EPA Office of Pesticide Programs (OPP) evaluations of studies, based on harmonized test guidelines from EPA Office of Prevention, Pesticides and Toxic Substances (OPPTS). Thousands of studies were characterized in ToxRefDB using standardized vocabulary, with consistent structure across multiple study types, and a high level of internal and external quality control (QC) for the abstraction of endpoints valuable in constructing predictive models[306]. Toxicity studies were performed on mice, rats and rabbits (single species per study).

A subset of the (ToxRefDB)[307] which is related to ToxCast substances was incorporated in the ToxMiner v17. It reported results from 461 animal toxicity endpoints. For each toxicity endpoint, *in vivo* toxicity data for chemicals were discretized to a binary outcome (toxic / non-toxic). The same flag value reported with the original database (10^6) for lack of toxicity was used as the cutoff between active and inactive responses. Between 234-251 compounds were tested per toxicity endpoint. Only 61 *in vivo* toxicity endpoints (out of 461) revealed toxic outcome for 35 or more compounds, a tentative threshold that was used to filter out

endpoints with insufficient data for statistical modeling QSAR. A list of *in vivo* toxicity endpoints and count of their associated toxic compounds is provided in the supplementary materials (Supplementary 2: List of *in vitro* assay endpoints, their respective total number of hits and whether it was selected for modeling.). Full details of the data collected can be found in literature[306].

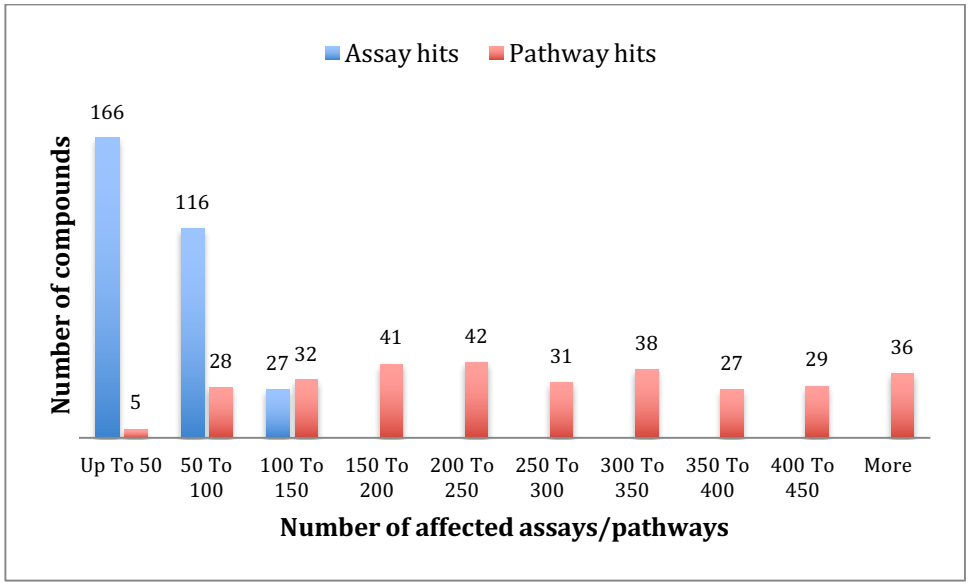


Figure 19. Histogram showing count of chemicals showing positive assay and pathway hits for 309 compounds of ToxCast Phase I. The assay data (blue bars) is very sparse - most chemicals affect only a few assays. Regrouping assays into affected pathways (red bars) allowed to retrieve a dataset that is less sparse and, therefore, more informative to machine learning algorithms.

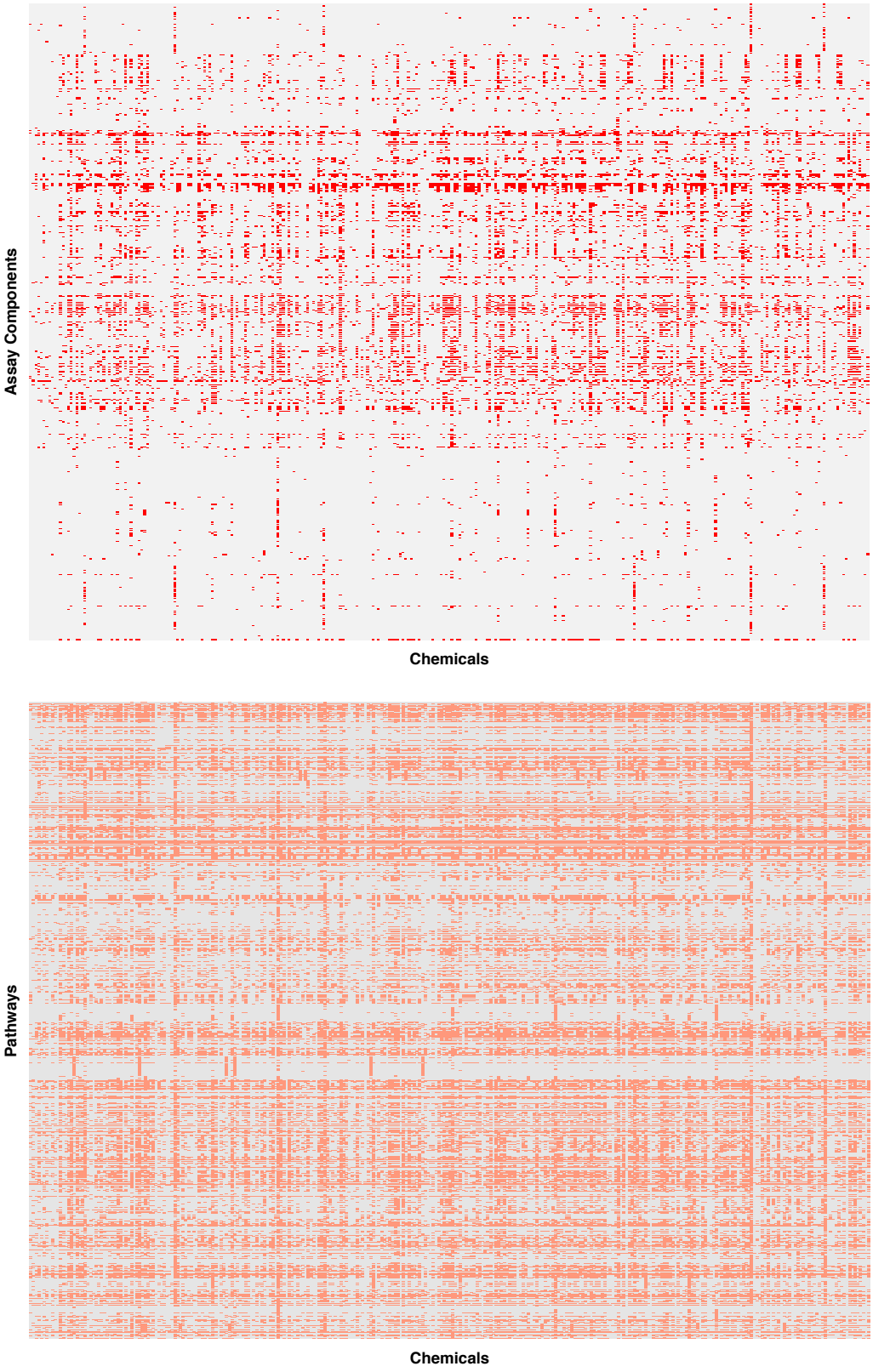


Figure 20. Heatmap of the assay-chemical activity matrix with 7% of all possible interactions resulting in positive hits (top) and pathway-chemical perturbation matrix with 14% positive hits (bottom). The regrouping of assay results into pathways perturbations resulted into less sparse matrix

4.1.2 Methods

A. Interacting with iPrior

Throughout this study, different KNIME workflows were used to upload the data, initialize the QSAR modeling on iPrior (as shown in Figure 9) and download the modeling results (as shown in Figure 22) (see 3.2.1 OCHEM / iPrior). All QSAR models construction was done on iPrior.

B. In silico Descriptors calculation

The preprocessing of chemical compounds was conducted using Chemaxon Standardizer, integrated within iPrior workflow. The standardization workflow consisted of removal of salt counter-ions, neutralization of charges, as well as standardizing the representations of aromatic rings and nitro groups. For the 3D descriptors, structures were optimized using CORINA.[308] iPrior web platform[186] was used to calculate 11 in silico descriptor packages gathered from multiple academic and commercial partners (see Table 2).

Descriptors calculation failed for ten chemicals, such as mixtures, inorganics, large macrocyclic compounds or organometallics. These structures were excluded (see Supplementary 3: List of ToxCast Phase I chemicals excluded from modeling due to failed descriptors calculation.). The remaining 299 compounds (out of 309) were used throughout this study for conducting the analysis.

C. Prefiltering criteria:

Descriptors with low variance can reduce the performance of distance based machine-learning algorithms. Thus, all descriptors were pre-filtered before model development. The following pre-filtering criteria were used: first, descriptors that are constant among all compounds, offering no information gain, were removed. Then, normalized descriptors that have variance smaller than < 0.01 were removed. Finally, descriptors were grouped if they showed pair-wise Pearson's correlation coefficient (R) > 0.95. The same pre-filtering steps were also applied to biologically derived descriptors (assay results and pathways perturbations) for modeling *in vivo* toxicity endpoints. Some algorithms implemented automatic scaling as part of their training protocols, i.e., in ASNN the descriptors were scaled to [0,1] interval, in MLRA and $k$nn the variables are normalised to zero mean and unit variance. For other algorithms, descriptors were scaled to the range [0-1] prior to the application of the algorithms. The prefiltering step was also applied within the bootstrap aggregation protocol for any of the algorithms used. Thus, the exact numbers of descriptors used could be different for each model.

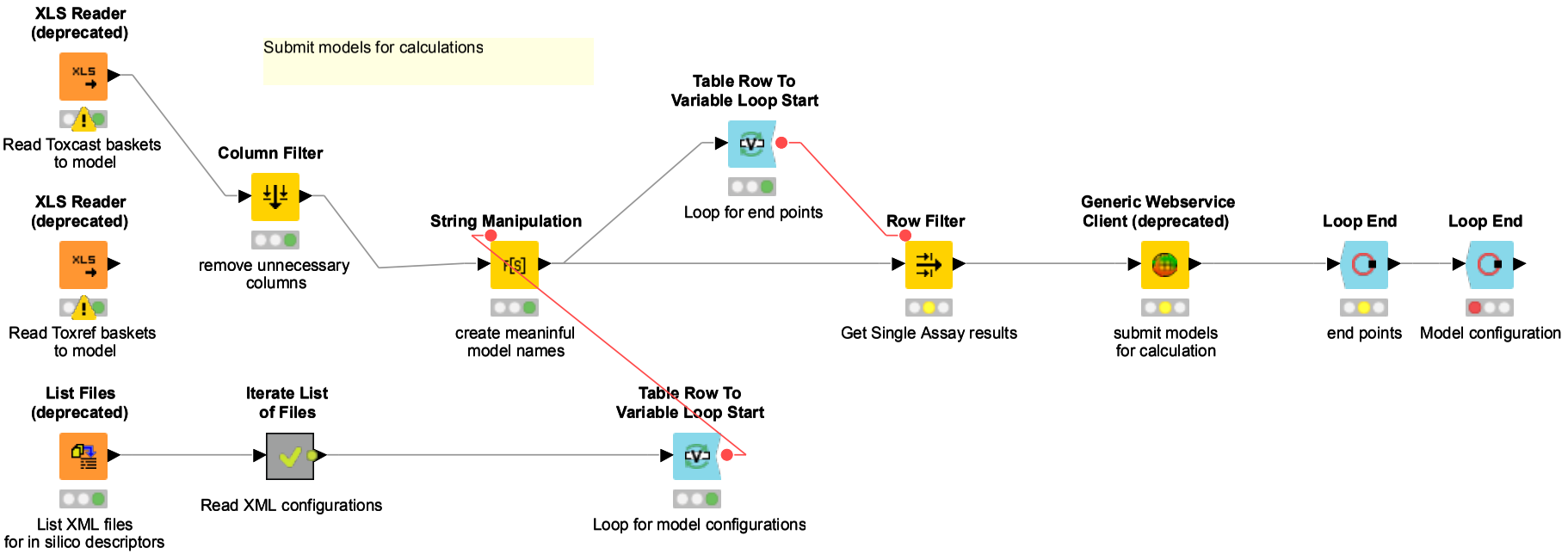


Figure 21. KNIME workflow showing the QSAR model-building process on iPrior. Different loops iterate over the model configuration XML files and endpoints to model. Overall 20968 QSAR models were constructed.

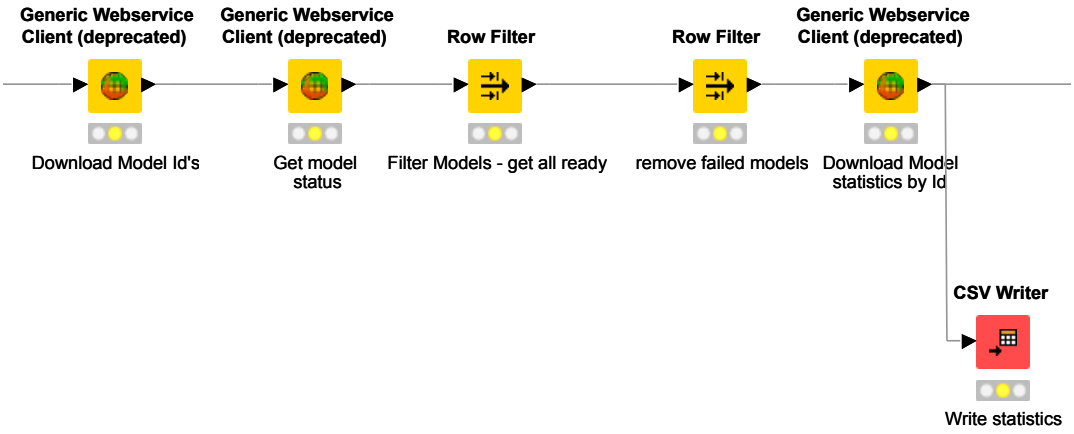


Figure 22. Partial KNIME workflow showing steps to collect QSAR modeling results from iPrior. The process starts with querying iPrior for QSAR models' IDs, then the workflow requests model status, filters by ready models that were successfully completed and commences with downloading their statistics

Table 2 List of *in silico* and biological descriptor packages used in the study. The number of descriptors within the package is shown.

Descriptors package name	Type	Descriptors count
ALOGPS[309,310]+ OEstate indices[311,312]	*in silico*	2+ 222
Chemaxon descriptors[313]	*in silico* (3D)	465
GSFragments[314]	*in silico*	588
ISIDA fragments[315]	*in silico*	1487
CDK[255]	*in silico* (3D)	204
Dragon 6[222]	*in silico* (3D)	3127
inductive descriptors[316]	*in silico* (3D)	40
MERA + MerSy[246–248]	*in silico* (3D)	529 + 42
QNPR[251]	*in silico*	
Spectrophores[249,250]	*in silico* (3D)	144
Adriana.Code[235]	*in silico* (3D)	183
ToxCast *in vitro* assays	Biological	407
ToxCast *in vitro* assays + CDK	Biological + *in silico*	407 + 204
pathways perturbation	Biological	1178
pathways perturbation + CDK	Biological + *in silico*	1178 + 204
ToxCast *in vitro* assays + pathways perturbation	Biological	407 + 1178
ToxCast *in vitro* assays + pathways perturbation + CDK	Biological + *in silico*	407 + 1178 + 204

D. Machine learning methods

Eight machine-learning methods were applied. These are k-Nearest Neighbors (kNN), Associative neural networks (ASNN), C4.5 decision tree (J48), Multiple Linear Regression Analysis (MLRA), Fast Stagewise Multiple Linear Regression (FSMLR), Partial Least Squares (PLS), Random Forests (RF) and Support Vector Machine (SVM). The description and configurations of each method is described in section 3.5 Machine learning algorithms above.

E. Performance measures and validation protocol

Different measures for accuracy estimation of models were calculated. These measures include: sensitivity, specificity, balanced accuracy (BAC), total accuracy (ACC), positive predictive value (PPV) and Matthews correlation coefficient (MCC). Section 3.7 Goodness of fit and prediction above shows the equations for these measures. Due to the unbalanced nature of the datasets, balanced accuracy was used throughout the study as the primary measure for comparing models. Stratified bagging was used as the validation protocol as described in section 3.9.3 Bootstrap aggregation (Bagging).

F. Modeling in vivo animal toxicity

Different feature combinations were used for modeling as listed in Table 2. These were the 11 in silico descriptor packages as well as 6 biological-derived features. The biological features were: The ToxCast in vitro assay, the pathway perturbations as described before, the combination of both and finally combining CDK descriptors (as an example of a widely used in silico descriptors package) with each of the three features.

G. Modeling in vitro assays

An important addition to previously conducted studies[174] is to explore the extent by which *in silico* descriptors could predict the outcome of *in vitro* assays. For that, QSAR models were constructed using the 11 *in silico* packages listed in Table 2.

4.1.3 Results and discussion

A. Modeling in vivo animal toxicity

In total 8 machine-learning algorithms were applied to 17 feature combinations (see Table 2) to model the 61 *in vivo* toxicity endpoints resulting in 8296 QSAR models constructed with 64-bagging-validation.

Figure 23 summarizes the balanced accuracies for all 8296 models as grouped by their respective endpoints. Each endpoint is represented by one vertical line, therefore 61 vertical lines in total. The upper tip of the line represents the maximum achievable balanced accuracy among all 136 combinations (17 feature combinations * 8 learning algorithms) that were used to model that endpoint. Likewise, the bottom tip of the line represents the lowest balanced accuracy. The triangle shows the median balanced accuracy among the models. More statistical parameters such as: specificity, sensitivity, Matthews's correlation coefficient (MCC) and overall accuracy were calculated and deposited in an open GitHub repository[317].

The low median balanced accuracies among models confirm the difficulties of modeling animal toxicity as previously reported[174]. However, it is worth investigating the top-performing models. Table 3 lists the balanced accuracy of the top-five predicted *in vivo* toxicity endpoints. Ranking was based on the maximum balanced accuracy achieved across the 136 models generated for each respective endpoint. The machine learning algorithms that contributed to the best models widely differed between the cases. In some cases, it was better to have a linear algorithm while in others the non-linear algorithms predominated. This could be related to the complexity of the endpoint and the descriptors involved. Naturally *in vitro* assays carry some errors that could contribute to the modeling difficulties

Table 3. The five best predicted *in vitro* assays based on the maximum balanced accuracy of the respective models. TPR: true positive rate (i.e., sensitivity), TNR: true negative rate (i.e., specificity), BACC: balanced accuracy, MCC: Matthews correlation coefficient, ASNN: associative neural networks, ISIDA: framgementor descriptors. [1]Model identification; models can be accessed from https://amaziz.com/iprior/model/id replacing "id" with the model identification number. [2]68% confidence intervals of the mean balanced accuracy are also shown. [3]Algorithms described in section 3.5 Machine learning algorithms. [4]Descriptor packages described in section 3.4 Molecular descriptors

Id[1]	Predicted endpoint	TPR	TNR	BACC[2]	MCC	Algorithm[3]	Descriptor package[4]
63	Chronic Rat Endpoint for Any cholinesterase inhibition measurement (e.g., brain and erythrocyte)	0.93	0.95	0.94 ±0.02	0.83	ASNN	ISIDA
62	Developmental Rat Maternal (Systemic)	0.75	0.7	0.72 ±0.04	0.33	FSMLR	CDK
66	Developmental Rat Maternal	0.77	0.71	0.74 ±0.04	0.35	J48	CDK + ToxCast Pathways
64	Developmental rat Maternal (General Maternal)	0.75	0.7	0.72 ±0.04	0.33	FSMLR	CDK
65	Chronic Mouse Endpoint for All effect related to apoptosis and necrosis	0.67	0.77	0.72 ±0.04	0.35	ASNN	ALOGPS, OEstate

B. Understanding significant features using ToxAlerts

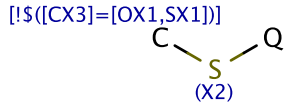
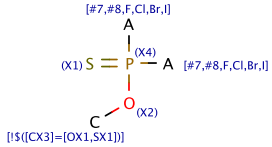
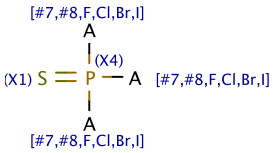
Acetylcholinesterase inhibition was one of the most predictable endpoints, with the balanced prediction accuracy reaching more than 90%. The analysis below concerns this endpoint. It was used as an example to investigate the success with modeling some endpoints (as shown in Table 3). To analyze the most significant features contributing to toxicity, the "Set Compare" and "ToxAlerts" tools, previously developed within OCHEM platform[224,291,318], were used. ToxAlerts is a web platform that is freely available for the storage of structural alerts. It holds a collection of alerts that have been collected from published literature in the form of SMARTS[319] patterns. The current collection includes more than 2000 alerts from more than 25 publications. It also allows *in silico* screening of chemical libraries for the detection of potential toxic or adverse effects. Among the stored alerts are patterns for potential mutagenicity, carcinogenicity, acute aquatic toxicity, skin sensitization, and possible idiosyncratic drug toxicity. It has previously been used to identify AlphaScreen frequent hitters in small-molecule HTS as well[320].

Upon comparing the 2 sub-sets of the rat acetylcholinesterase inhibition chemicals (toxic and non-toxic) using ToxAlerts[224], many significant toxic groups were detected. The three most significant alerts are shown in Table 4 together with their respective p-value. Indeed, organophosphorus insecticides functions through acetylcholinesterase inhibition as its primary mechanism of action. In the dataset analyzed, only one phosphorus derivative was reported as non-toxic. This may be a simple scaffold for descriptor packages accounting for atom counts or fragments to detect such toxicity. However, it could be more challenging for *in vitro* assays to indirectly capture the presence of such scaffold. The p-values and enrichment factors for the significance of each SMARTS pattern detected is shown in Table 4.

C. Detecting significant in vitro assays using SetCompare

To detect which *in vitro* assays or pathways perturbations can better act as a biological descriptor for building a bioactivity signature, the "Set Compare" tool was used. The Acetylcholinesterase inhibition was selected as an example endpoint for the analysis. The toxic vs. non-toxic sets of compounds were compared. Table 5 shows the most significant *in vitro* assays together with their respective p-values. Indeed, Acetylcholinesterase (AChE)-related *in vitro* assays in both rat and human were the most significant ones. While this sounds logical *a posteriori*, the analysis did not assume any prior knowledge of the underlying *in vivo* toxicity pathway. This confirms the potential of using HTS *in vitro* screening to understand mechanisms of toxicity not previously known.

Table 4. Most common toxicity alerts for toxic acetylcholinesterase inhibitors identifying organophosphorus compounds. [1]SMARTS pattern[319] describing the alert.

Toxicity Alert	# Toxic set (42)	# non-toxic set (206)	Enrichment factor	p-value
 [#6&!$([CX3]=[OX1,SX1])][Sv2X2][!#1!#6][1]	14 (33.30%)	1 (0.50%)	68.7	10^{-11}
 [SX1]=[Pv5X4]([OX2][#6&!$([CX3]=[OX1,SX1])])([#7,#8,F,Cl,Br,I])[#7,#8,F,Cl,Br,I]	10 (23.80%)	1 (0.50%)	49	10^{-8}
 [SX1]=[Pv5X4]([#7,#8,F,Cl,Br,I])([#7,#8,F,Cl,Br,I])[#7,#8,F,Cl,Br,I]	10 (23.80%)	1 (0.50%)	49	10^{-8}

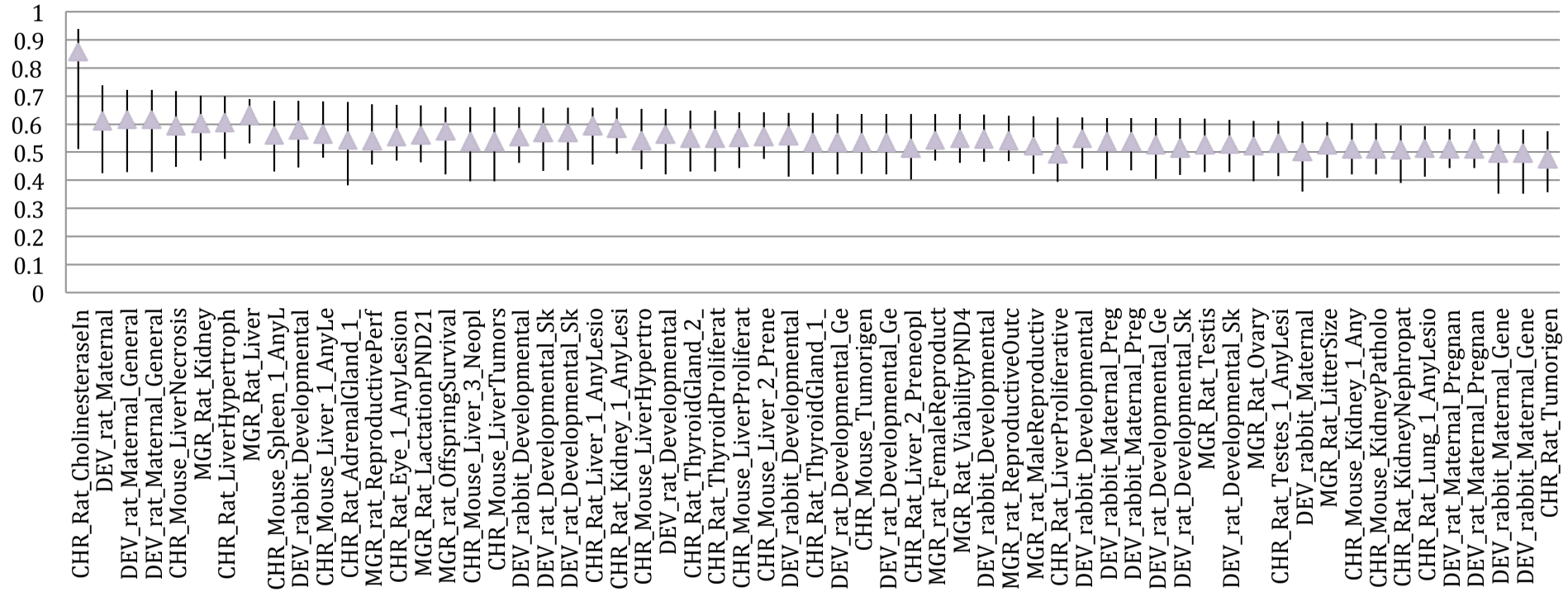


Figure 23. Plot showing the difference in the balanced accuracy for the 8296 models constructed using 136 algorithm/features combinations for each of the 61 *in vivo* toxicological endpoints from the ToxRefDB. The lower and upper boundaries of the line represents the maximum and minimum balanced accuracy achieved; respectively. Endpoints are sorted by the maximum achievable balanced accuracy. The triangle shows the median balanced accuracy among the models irrespective of the *in silico* or biological descriptors or the machine-learning algorithm used. More statistical parameters of each model are provided in the supplementary materials. The x-axis shows the endpoints names, according to ToxRefDB format: study type_species_organ_effect_category. The full list of endpoints and their description is available from EPA website[182] . Study type: DV, developmental; CHR, chronic; MGR, multigenerational. Species: Rt, rat; Rb, rabbit; Ms, mouse. Effect and category: Mat, maternal; GL-Mt, general maternal; Dev, developmental; PregRel, pregnancy related; PregLoss, pregnancy loss; AnyLes, any lesion; Skl, skeletal; PreneoplastLes, preneoplastic lesion; GenFetal, general fetal; Prolif- eratLes, proliferative lesion; WghtReg, weight reduction; NeoplastLes, neoplastic lesion; Reproduct, reproductive; ThyroidGlnd, thyroid gland; ReproductTract, reproductive tract; Perform, performance; Cholinester, cholinesterase; Inhibit, inhibition.

Table 5. Most significant *in vitro* assays for toxic acetylcholinesterase inhibitors showing the association of acetylcholinesterase pathway

Toxicity Alert	# Toxic set (42)	# non-toxic set (206)	Enrichment factor	p-value
ToxCast assay: Novascreen Rat AChE	13 (31.0%)	1 (0.5%)	63.8	10^{-10}
ToxCast_Pathway: 512 Glycerophospholipid metabolism	13 (31.0%)	1 (0.5%)	63.8	10^{-10}
ToxCast_Pathway: 801 Process: response to wounding GO id:0009611	10 (23.8%)	1 (0.5%)	49	10^{-8}
ToxCast_Pathway: 796 -- Component: basal lamina Description: Component: basal lamina GO id:0005605	10 (23.8%)	1 (0.5%)	49	10^{-8}
ToxCast_Pathway: 793 Function: acetylcholinesterase activity GO id: 0003990	10 (23.8%)	1 (0.5%)	49	10^{-8}
Novascreen Human AChE	10 (23.8%)	1 (0.5%)	49	10^{-8}

D. Comparing performance across algorithms and descriptor packages

For each toxicity endpoint, the individual model that showed the highest balanced accuracy was selected. Then, the different algorithms and descriptor packages were ranked according to the number of times they contributed to such models. Table 7 shows the ranking of different descriptor packages in their success to achieve the best predictive model (considering the highest balanced accuracy). It also shows the number of toxicity endpoints for which the descriptor package contributed to its best model.

Generally, different *in silico* descriptor packages exhibited similar performance as shown in Table 7 while *biological descriptors* performed worse. However, it is worth noticing that biological descriptors outperformed '*in silico* descriptors-only models' for the prediction of 9 toxicity endpoints. In 8 cases, the pathways, either alone, in combination with the assays or in combination with the *in silico* descriptors significantly improved the model prediction ability (Table 6) with p-values <0.05. This suggests that the re-arrangement of *in vitro* assay outcomes in the form of pathways perturbation provided extra information for modeling these toxicity endpoints. Also, biological descriptors in combination with *in silico* descriptors contributed to achieving the highest balanced accuracy for 5 endpoints. For these cases, it might be because each kind of descriptors encoded for different information related to the chemical structure or the *in vivo* target.

Error! Reference source not found. shows the endpoints where use of biological descriptors outperformed the *in silico-only* descriptors. It also presents the balanced accuracies of the biological descriptors alone, *in silico* descriptors alone and the combination of both.

Table 8 compares different machine learning on their performance. In general, both simple methods, such as the FSMLR, MLRA and kNN and non-linear high-resolution methods, such as random forests and neural networks showed comparable results.

E. Modeling in vitro assays

The same *in silico* descriptors and machine learning algorithms were applied as previously explained for the *in vivo* toxicity endpoint. Eight machine-learning algorithms in combination with 11 descriptor packages were used to model 144 endpoints resulting in a total of 12672 QSAR models. All models were constructed with 64-bagging validation.

Regarding the best performing descriptors, unlike the case with *in vivo* toxicity endpoints, Dragon 6[222] and ALOGPS[309,310]+ OEstate indices[311,312] performed better than other descriptors as shown in Table 7**.** Regarding machine-learning algorithms there was no distinctive difference in performance as shown in Table 8**.**

Figure 24 summarizes the balanced accuracies for all 12672 models as grouped by their respective *in vitro* endpoints. Each endpoint is represented by one vertical line, therefore 144 vertical lines in total. The upper tip of the line represents the maximum achievable balanced accuracy among all 88 combinations (11 *in silico* descriptors * 8 learning algorithms) that were used to model that endpoint. Likewise, the bottom tip of the line represents the lowest balanced accuracy. The triangle shows the median balanced accuracy among the models. The figure shows generally higher balanced accuracies achieved across the *in vitro* endpoints compared to the prediction of *in vivo* endpoints. Table 9 lists the statistics for the five best-predicted *in vitro* assays based on the maximum-achievable balanced accuracy among the 88 models built per endpoint. Multiple *in vitro* assays (79 out of 144) presented a balanced accuracy of > 0.7. More statistical parameters were calculated and deposited in an open GitHub repository[317]. *In vitro* assays measuring the expression of different CYP450 isoforms were among the most successful to be modeled. This agrees with earlier QSAR studies reporting similar success[321–323].

F. Modelability of the datasets

Aside from the modeling process itself, many characteristics of the training set can affect the predictive power of QSAR models including its diversity, size, presence of activity cliffs and activity distribution[324,325]. Previous study investigated ToxCast Phase I data with regard to its modelability showing that too many activity cliffs in comparison with the dataset size makes it not suitable for QSAR modeling[326]. In this study, the median balanced accuracy of all models built for each individual endpoint was used to compare its relative ease of modeling.

For *in vivo* endpoints, only 7 endpoints had a median balanced accuracy of above 0.6. The chronic rat acetylcholinesterase inhibition stands as a clear exception for an endpoint that is easy to model where the median balanced accuracy for all models exceeds 0.85. For *in vitro* endpoints, the provided statistics reveal that 21 endpoints have median balanced accuracy above 0.70. Comparatively, *in vitro* assays where easier to model than *in vivo* toxicity endpoints.

Table 6. Toxicity endpoints where the biological descriptors contributed to the best predictive QSAR model (with the underlined balanced accuracy). Balanced accuracies for models developed using CDK (as an example for *in silico* descriptors) as well as different biological descriptors are shown.

Toxicity endpoint	Best model (out of 136 models per endpoint)			Balanced accuracies						
	id	BACC	Algorithm	CDK	CDK+ Pathway[1]	CDK+ Assays	CDK+Assays+ Pathway	Pathway	Assays	Assays + Pathways
Chronic rat liver preneoplastic lesion	79[3]	0.64 ±0.04[4]	PLS	0.49	0.54	0.53	0.60	0.54	0.56	<u>0.64</u>
Chronic rat testes any lesion	70	0.61 ±0.04	FSMLR	0.54	0.56	0.53	<u>0.61</u>	0.50	0.52	0.51
Chronic rat endpoint for all neoplastic and non-neoplastic proliferative liver lesions	71	0.62 ±0.03	J48	0.50	0.55	0.58	0.61	0.60	0.58	<u>0.62</u>
Developmental rabbit maternal (mLEL_rabbit)	72	0.61 ±0.06	FSMLR	0.50	0.44	0.51	<u>0.61</u>	0.36	0.46	0.44
Developmental rat maternal (mLEL_rat)	66	0.74 ±0.04	J48	0.63	<u>0.74</u>	0.55	0.55	0.52	0.59	0.51
Developmental rat general fetal pathology	73	0.62 ±0.05	FSMLR	0.47	0.51	0.47	0.49	0.56	<u>0.62</u>	0.54
Developmental Rat Skeletal Appendicular	74	0.62 ±0.04	ASNN	0.53	0.58	0.55	0.55	0.61	0.59	<u>0.62</u>
Multigenerational rat reproductive performance	75	0.67 ±0.04	PLS	0.50	<u>0.67</u>	0.52	0.57	0.63	0.59	0.53
Multigenerational rat endpoint for viability Index	76	0.64 ±0.04	ASNN	0.49	0.60	0.60	0.58	<u>0.64</u>	0.60	0.57

[1]Assays: ToxCast Assays, [2]Pathways: ToxCast pathways perturbation. [1]Model identification; models can be accessed from https://amaziz.com/iprior/model/id replacing "id" with the model identification number. [4]68% confidence intervals.

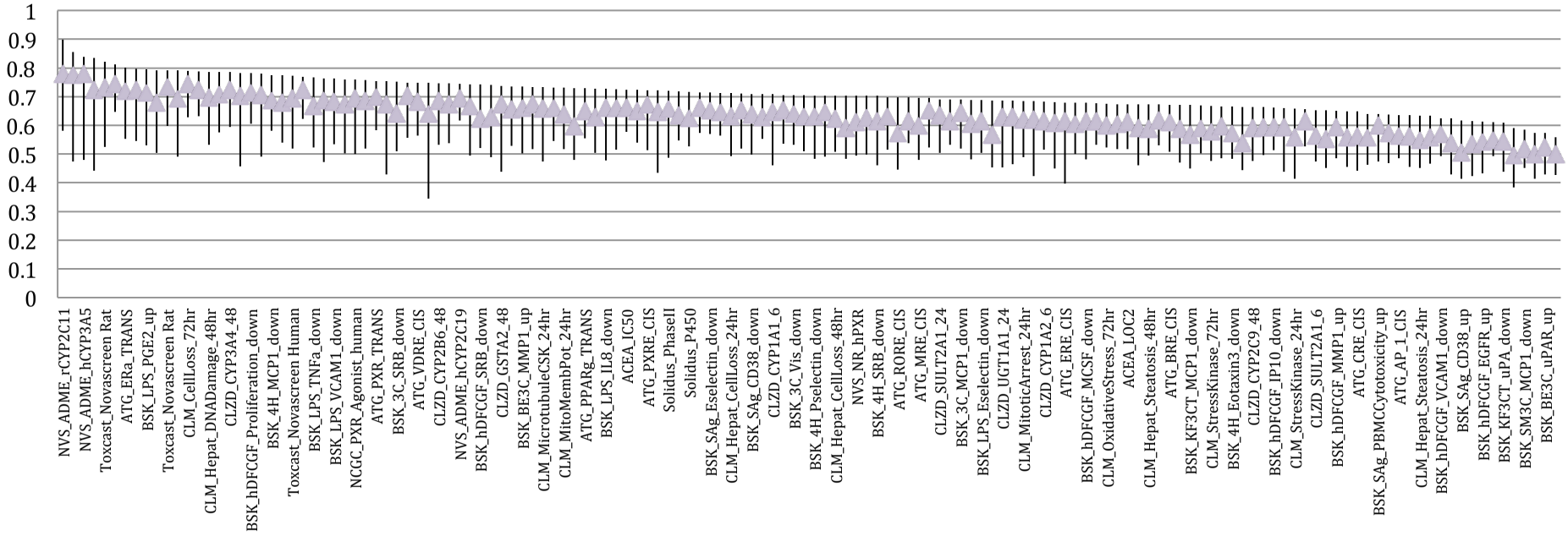


Figure 24. Plot showing the difference in the balanced accuracies for the 12672 models constructed using 88 algorithm/*in silico* descriptors combinations for each of the 144 *in vitro* assay endpoints from the ToxCast database. The lower and upper boundaries of the line represents the maximum and minimum balanced accuracy achieved; respectively. Endpoints are sorted by the maximum achievable balanced accuracy. The triangle shows the median balanced accuracy among the models irrespective of the *in silico* descriptor package or the machine learning algorithm used. More statistical parameters of each model are provided in the supplementary materials. ACEA: ACEA - Real-time Cell Electronic Sensing; ATG: Attagene - Transcription factor assays; BSK: BioSeek - Cell-based protein level assays; CLM: Cellumen - Cell imaging assays; CLZD: CellzDirect - Transcription assays; NCGC: NCGC - nuclear receptor assays; NVS: Novascreen / Caliper - receptor binding and enzyme inhibition assays; Solidus: Solidus - P450 vs. cytotoxicity assays

Table 7. Comparing the performance of different descriptor packages in constructing QSAR models for *in vivo* toxicity and *in vitro* assays. The number of *in vivo* toxicity endpoints / *in vitro* assays where the descriptor package contributed to the model with highest balanced accuracy is shown.

Descriptors	*In vivo* rank	# *in vivo* endpoints	*In vitro* rank	# *in vitro* endpoints
ISIDA	1	9	3	16
CDK	2	7	7	8
ALOGPS + OEstate	3	6	2	26
GSFrag	4	6	8	8
Dragon6	5	5	1	35
Spectrophores	6	5	11	2
Adriana	7	3	6	12
Inductive descriptors	8	3	10	3
QNPR	9	3	9	6
ToxCast assays + ToxCast Pathways	10	3		
Chemaxon Descriptors	11	2	4	15
MERA + MerSy	12	2	5	14
CDK + ToxCast Pathways	13	2		
CDK, ToxCast assays + ToxCast Pathways	14	2		
CDK + ToxCast assays	15	1		
ToxCast Pathways	16	1		
ToxCast assays	17	1		
Total		61		145[1]

[1]One *in vitro* assay endpoint showed an exact tie between 2 models

Table 8. Comparing the performance of the machine-learning algorithms in constructing QSAR models for *in vivo* toxicity and *in vitro* assays. The number of *in vivo* toxicity endpoints and *in vitro* assays where the algorithm contributed to the model with highest balanced accuracy is shown.

Algorithm	*In vivo* rank	# *in vivo* endpoints	*In vitro* rank	# *in vitro* endpoints
FSMLR	1	16	3	21
ASNN	2	13	4	18
WEKA-RF	3	11	1	33
PLS	4	7	2	26
MLRA	5	5	8	6
*k*NN	6	3	7	10
LibSVM	7	3	5	18
WEKA-J48	8	3	6	13
Total		61		145[1]

[1]One *in vitro* assay endpoint showed an exact tie between 2 models

Table 9. The five best-predicted *in vitro* assays based on the maximum achievable balanced accuracy for the endpoints.

Model id	property	Sensitivity	Specificity	Balanced accuracy	MCC [1]	Algorithm	Descriptor package
67[3]	Novascreen Human CYP2B6	0.81	0.91	0.86 ±0.03[2]	0.6	ASNN	ALOGPS, OEstate
77	Novascreen Human CYP2C18	0.74	0.93	0.84 ±0.04	0.62	LibSVM	ALOGPS, OEstate
68	Novascreen Rat CYP2C6	0.83	0.81	0.82 ±0.03	0.5	WEKA-RF	ALOGPS, OEstate
69	Novascreen Rat CYP2C11	0.93	0.86	0.9 ±0.02	0.65	FSMLR	Dragon6
78	Novascreen Human CYP3A5	0.81	0.87	0.83 ±0.03	0.58	LibSVM	Dragon6

[1]MCC: Matthews correlation coefficient. [2]68% confidence intervals of the mean balanced accuracy are also shown. [3]Models can be accessed from https://amaziz.com/iprior/model/id replacing "id" with the model identification number.

4.1.4 Summary of ToxCast™ phase I analysis aspects

Several *in vivo* endpoints with a promising predictive balanced accuracy exceeding 0.75 were identified (examples listed in Table 3). In some cases, the biological descriptors derived from the *in vitro* profiling of chemicals significantly improved (p-values <0.05) models' predictive ability compared to the use of *in silico* descriptors alone (Table 6). Also, the regrouping of the *in vitro* assay responses in the form of pathway perturbations significantly improved (p-values <0.05) the predictivity for some toxicity endpoints

However, analysis of ToxCast Phase I compounds remains challenging for most *in vivo* endpoints as shown by the median performance of constructed QSAR models. It remains difficult to replace animal toxicity testing using predictive QSAR models, with a possible exception for the acetylcholinesterase inhibition. However, the comprehensive modeling with multiple machine learning algorithms and descriptors shows relative success for selected endpoints (Table 3). Thomas et al[174] presented similar findings and advised the combination of QSAR and *in vitro* profiling of chemicals as means for prioritization, rather than substitution, of animal toxicity testing. The "Set Compare" utility proved successful for detecting the most significant *in vitro* assays correlated to toxicity endpoints (Table 5). This shows that *in vitro* assays could assist in understanding the underlying mechanism of toxicity.

Multiple *in vitro* assays showed a high balanced accuracy (>0.8) (Table 9) when modeled by *in silico* descriptors. This represents a different methodology towards toxicity modeling where *in silico* descriptors can be used to model *in vitro* assay outcomes known to be related to *in vivo* effect. Tox21 project explores this possibility by profiling large number of chemicals using *in vitro* assays as an investigation and exploratory tool (see Tox21 project).

Many challenges remain in place: first, QSAR modeling, as a statistical approach, necessitates a significant amount of data. The low number of chemicals (as training instances) restricts the

modeling process. This constrain would gradually diminish as more data becomes available in future stages of ToxCast and other programs. The applicability domain and predictive power of models is more likely to increase. Secondly, the *in vitro* representation could be too simple to address the complexity of the interactions *in vivo*. Bioavailability and biotransformation can play a significant role in inducing or diminishing toxic effects for chemicals. The importance of absorption, distribution, metabolism, elimination (ADME) for both drug discovery as well as environmental risk assessment cannot be overestimated[327–329]. Thirdly and finally, the assays conducted might not be sufficient for capturing biochemical events on the molecular level or depict the pathways responsible for toxicity. With that taken into consideration, ToxCast Phase I still provided useful overview of the chemical initiating events. Some assays may be redeemed unnecessary in future tests, as they were focused on promiscuous endpoints or, vice versa, were not sensitive enough. As more data is being gathered from chemical providers through programs like REACH, ToxCast and Tox21, QSAR modeling will play more significant role.

For this study, the public platform iPrior[186] (Figure 9) was deployed and is currently hosting data from ToxCast, Tox21, e1K projects. iPrior is a public online tool. It allows users to reproduce any of the QSAR models created in this study as well as run predictions on new chemicals using these models. The configurations (machine learning algorithm and descriptors) that provided the best balanced-accuracy for each endpoint is provided in the supplementary materials (Supplementary 4: Statistical parameters for the models with best balanced-accuracy for each of the 144 *in vitro* assay endpoints from the ToxCast database.) as well as (Supplementary 5: Statistical parameters for the models with best balanced-accuracy for each of the 61 *in vivo* toxicological endpoints from the Toxicity reference database.) Users are encouraged to investigate the model profiles, applicability domains and run predictions using their own chemical structures. It is open to researchers to upload more data or contribute their descriptor packages. iPrior supports the full cycle of QSAR research online. The platform is freely accessible for the non-commercial use of the academic community. The required workflows and modeling infrastructure is in place to assist scientists in developing predictive bioactivity signatures. This infrastructure remains open for the investigation of upcoming data releases. Models that become accepted by the community and regulators can play a role in predicting *in vivo* toxicity and reduce animal testing.

4.2 Lowest effect level prediction

This study was made as part of the ToxCast challenge 2014 organized by EPA and the Topcoder competition platform. The aim of this study is to develop QSAR models to predict Lowest effect level (LEL) concentration based on *in vitro* measurements and calculated *in silico* descriptors. LEL is defined as "the lowest dose that shows adverse effects in these animal toxicity tests." The LEL values are used by regulators to put limits on exposure to chemicals to ensure that they are tolerated by majority of the population.

The author participated in the challenge under the name (AMAZIZ) and achieved the fifth rank among 47 participants. EPA published a summary of the challenge that can be accessed from the web archives[330]. This study expands on the work done during the competition and provides further analysis.

4.2.1 Data acquisition and curation

The total dataset used during the challenge included 1,854 molecules divided into three groups. The first group (Group A) covered 483 compounds for which their LEL values were disclosed. These chemicals were intended to be used as a training set. The second group (Group B) included 143 chemicals, for which the LEL values was not revealed. A randomly selected subset consisting of 63 compounds from Group B was used as a provisional test set. Contestants are allowed to submit predictions against the provisional test set and receive a score, during the competition, in order to optimize their algorithms. Finally, the remaining 80 compounds from Group B were considered as the final test set that were used to determine the final ranking for the competition.

The EPA did not reveal which compounds belong to which group. Users were always asked to predict the outcome of all 1,854 compounds including 1228 compounds of group C, which were part of neither training nor test sets.

Besides the chemical structures, the dataset included measurements from more than 700 *in vitro* assays from different biochemical and cell-based assays. As described earlier (see 4.1 ToxCast™ phase I - In vitro toxicity assays dataset), they cover a wide range of proteins, pathways, and cellular processes against which chemicals may interact.

Data was downloaded from the EPA website. The data was divided into multiple files that contain experimental measurements, structure representations (in SMILES format) as well as few *in silico* descriptors. A KNIME workflow (shown in Figure 25) was used to examine the dataset and join the chemical structures to their corresponding LEL values for the training set (483 compounds).

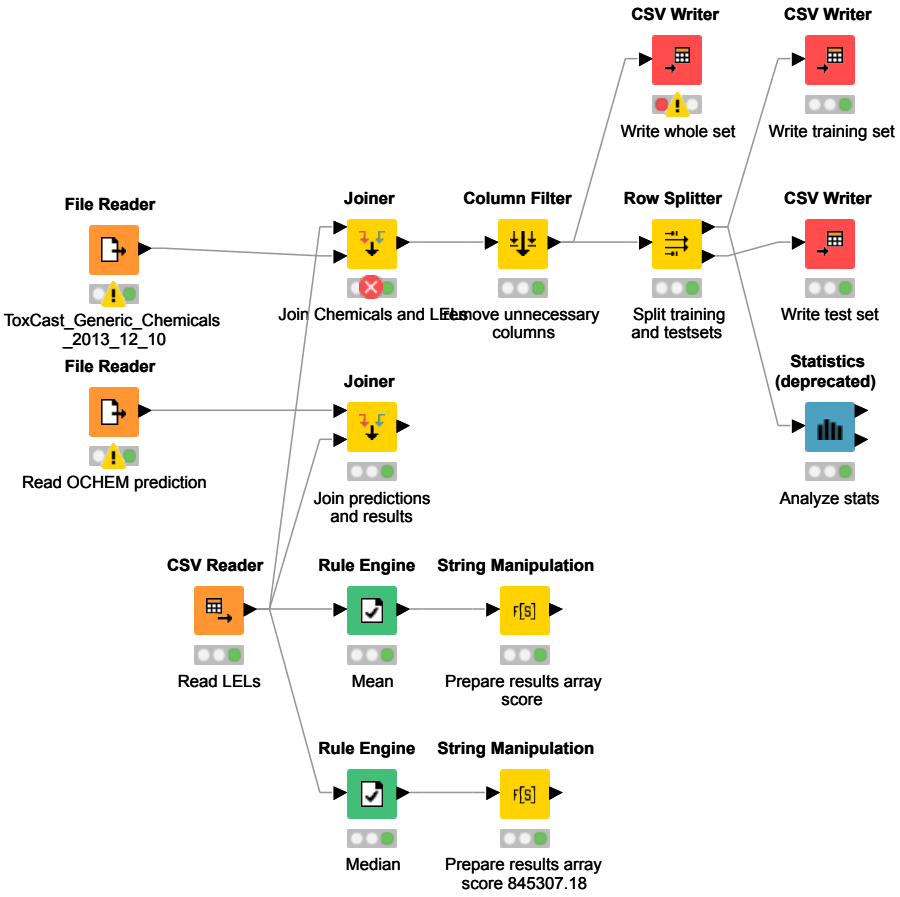


Figure 25. KNIME workflow used to analyze the LEL data and prepare the submission files

4.2.2 Methods

Two modeling approaches were used to build QSAR models for the prediction of LEL values. These are Feature nets and consensus modeling.

A. Feature nets

The first approach used feature nets (FN) to simultaneously build models for LEL as well as Octanol/Water partition coefficient and water solubility. The model used OEstate indices[311,312] as *in silico* descriptors. The *in vitro* assays were ignored, because it did not improve the prediction accuracy. FN is an example of inductive knowledge transfer approaches where, unlike the conventional single-task learning (STL) modeling focused only on a single target property without any relations to other properties, in the framework of inductive transfer approach, the individual models are viewed as nodes in the network of interrelated models built sequentially[331]. The rationale behind selecting lipophilicity and water solubility, for such inductive knowledge transfer, is that they have been known to correlate well with many biological and toxicological activities[7,8]. The dataset used for training the FN consists, in addition to the LEL measurements discussed above, of 8072 experimental solubility measurements and 16823 lipophilicity measurements. It is the same dataset behind the ALOGPS model[234], which has been shown high predictive power in multiple studies[332,333].

The predictions by the final QSAR model developed in this approach was used as the author's participation in the challenge (team: AMAZIZ)

B. Consensus modeling

The second approach used ten *in silico* descriptor packages implemented in OCHEM to build 10 independent QSAR models. The ten descriptor packages are: OEstate indices[311,312], Chemaxon descriptors[313], GSFragments[314], ISIDA fragments[315], CDK[255], Dragon 6[222], inductive descriptors[316], MERA + MerSy[246–248], QNPR[251] and Adriana.Code[235] descriptors. The 3D structure representation was generated using Corina[17]. Finally, a consensus model was constructed to average the prediction outcomes from all ten models. This approach is similar to the approach used in study 4.1 ToxCast™ phase I above. It was used by contestant, Novserj[334], and is included for comparison.

Descriptors calculation failed for 37 molecules for different packages (11 of which are from the training set). These compounds presented chemotypes that were unsupported by some descriptor packages. For instance, CDK descriptors package fails for chemicals containing [Sn], [Hg], [B] or [As] atoms. Other compounds were too large to be calculated with the current descriptors implementation. These compounds included rifampicin, alpha-cyclodextrin, milbemectin, emamectin that failed due to calculation time-out or structure conversion errors. To compensate for unavailable predictions for failed molecules, the median LEL value for all training set compounds (logLEL = - 3.201 log(M)) was used in the first approach while the mean LEL value (logLEL = -3.2602 log(M)) was used in the second.

For comparison, two more models were built. The first is based only on *in vitro* assay outcomes while the second was based on two simple descriptors: molecular weight and number of carbon atoms using linear regression, ASNN and LibSVM. The purpose of these models is to compare *in silico* and *in vitro*-based descriptor performances as well as judge whether, and to what extent, complex machine learning and descriptor packages improve prediction accuracy.

Associative neural networks (ASNN) was the selected machine-learning algorithm for building QSAR models in both approaches. The algorithm was used as described in 3.5.2 Artificial neural networks (ANN). The same descriptor selection workflow was used for both approaches as described in section 3.6 Variable selection above.

To avoid over-fitting, Bootstrap aggregation (bagging) was used for estimating models performance in both approaches as described in section 3.9.3 Bootstrap aggregation (Bagging) above with 64 models bag. Bagging was also used to validate the consensus model built in the second approach. As the provisional test set was much smaller (N = 63) than the training set (N = 483). Therefore, optimizing the prediction algorithm against such provisional set is likely to result in overfitting. For this reason, a conscious decision was made in both approaches to neglect the provisional test set and rely solely on the bagging standard deviation on the training set as a measure for models' confidence intervals.

The challenge organizers used a scoring function (Equation 38) based on Root Mean Square Error (RMSE) as the statistical metric to compare models' performance. Therefore, models with lower RMSE will receive a higher score and therefore be judged as more superior (i.e.,

ranked higher). Furthermore, the organizers provided Pearson correlation coefficient and AUROC for some winning submissions in a summary report released after the challenge[330].

$$Score = 1000000 \times (2 - RMSE)$$ Equation 38

4.2.3 Results and discussion

Table 10 summarizes the statistical results of the competition as published by the organizers on the challenge website and the summary report[330] complemented with results from the investigations in this study. The RMSE of the two-descriptor model with linear regression on the training set was 1.0 ± 0.04 log unit. While LibSVM with the same two simple descriptors decreased RMSE to 0.97 ± 0.04 log unit. This error was significantly higher than that obtained by either the consensus modeling approach or the feature net approaches. On the other hand, it was exactly equal to the performance of the QSAR model built using the *in vitro* assays measurements (RMSE = 0.97±0.04).

The first approach resulted in RMSE of 0.92±0.04 on the training set and R^2 of 0.19 ± 0.02. The final predictions for all 1854 compounds is deposited in an open GitHub repository[317] for reference.

In the second approach, models developed with different *in silico* descriptor sets resulted in similar performance as shown in Table 11. Interestingly, *in vitro* assays measurements provided the lowest accuracy (RMSE: 0.97±0.04) compared to other descriptor packages. Consensus modeling achieved an improved performance regarding both RMSE (0.88±0.04) and R^2 (0.27±0.04) measures.

It is worth noticing that there have been large swings in ranking between provisional and final test sets. For example, the author's submission (AMAZIZ) was ranked 20th in the provisional submission but achieved the fifth place in the final prediction. Likewise, the fourth winner in the final test set was ranked 27th in the provisional set. On the other hand, the first provisional rank was only able to score ninth in the final test set. As expected, the provisional ranking was not a good indicator of the final ranking. This may be due to contestants optimizing their submissions for the provisional set.

This can be explained by investigating the confidence intervals as a function of dataset size. The consensus model RMSE was 0.88 ± 0.04 for N=472 training set molecules. The confidence interval of the provisional set was estimated by random sampling of N = 63 molecules from the training set, for each of which the intervals were calculated. The confidence interval for a set of such size was (± 0.08) and therefore twice as large as the training set. This means that a selection of a model based on its performance about the provisional test set would be about twice uncertain as the selection based on the training set. It is therefore advisable to rely on the estimated accuracy on the training set, rather than the provisional test set for model selection.

It is also worth noticing that the confidence interval for the final test set (N = 80 molecules) is about the same as for N = 63 molecules. The wide confidence intervals for both provisional and final test contributed to the fluctuations of ranks of challenge models for both sets. Provisional and final model ranks were correlated only with R=0.76.

It is worth noting that the RMSE of the top eight models were in range 1.12 to 1.16 and thus were within the confidence intervals of the winning model. Therefore, from a statistical perspective, these models had the same performance and their scoring differences are not more significant than random chance.

Table 10. Summary of the performance of the top-ranked models in EPA ToxCast challenge

Model	training set[a]		test sets				
			provisional		final		
	RMSE	R^2	RMSE	rank	RMSE	R^2	Rank
novserj	0.88±0.04	0.27±0.04	1.03±0.08[b]	8	1.12±0.08[b]	0.31	1
NobuMiu			1.03	9	1.131	0.30	2
a9108tc			1.05	16	1.134	0.29	3
klo86min			1.09	27	1.139	0.29	4
amaziz	0.92±0.04	0.19 ± 0.02	1.06	20	1.145	0.29	5
in vitro assays[c]	0.97±0.04	0.11±0.03					
MW + NC[d]	0.97±0.04	0.11±0.03					

[a]The accuracy of predictions for the validation "out-of-the-bag" samples. [b]Confidence intervals were estimated using the sets, which were sampled from the training set and had each the same size as the respective test set[318]. [c]The best model based on the *in vitro assays* descriptors developed using LibSVM method. [d]The model based on molecular weight (MW) and number of carbon atoms (NC) developed using LibSVM method.

Table 11. Performance of QSAR models based on *in silico* descriptors for the prediction of LEL

Descriptor packages	RMSE	R^2
OEstate	0.95	0.18
CDK	0.92	0.23
Dragon6	0.92	0.23
ISIDA Fragmentor	0.95	0.2
GSFragments	0.96	0.16
MERA + MerSy	0.93	0.21
Chemaxon Descriptors	0.92	0.23
Inductive Descriptors	0.94	0.18
Adriana	0.93	0.21
QNPR	0.97	0.17

4.2.4 Summary of LEL prediction aspects

QSAR models for the prediction of LEL were developed. Two different approaches were discussed, both of which received a top ranking in the EPA ToxCast challenge, which was organized by the Topcoder platform[335]. The performance of *in vitro* assays and *in silico* descriptors was compared. *In vitro* descriptors alone performed in par with *in silico*

descriptors. The winning approaches described in this study were able to achieve such ranking[334] despite not including *in vitro* descriptors.

The exclusion of the model based on *in vitro* descriptors did not change the accuracy of the ASNN consensus model. Using a model based on the combination of both *in silico* and *in vitro*-based descriptors requires the availability of both descriptors. This hinders its application to chemicals for which *in vitro* measurements are available. Performing such experiments has a higher cost and is more time consuming than the calculation of *in silico* descriptors alone. Therefore, LEL models based on *in silico* descriptors only are recommended as they are more feasible and does not compromise on prediction accuracy.

4.3 Tox21 project

4.3.1 Introduction and data source

This study was made as part of a challenge organized by National Institute of Health (NIH) / National Center for Advancing Translational Sciences (NCATS). Through this challenge, HTS assay data from 12 targets were made available to contestants to predict the potential of activation of such targets using different *in silico* approaches. Targets were divided into 2 panels; a nuclear receptor-signaling panel as well as a stress response panel. For each target, the datasets were given in 2 portions; an initial training set and a leaderboard set. The logic behind the leaderboard set was to allow competing teams to understand their standing as compared to other competitors using a unified test set. The ground truth (labels) for the evaluation sets was released towards the end of the competition to allow all contestants to maximize the learning for their model[336]. The final ranking was done using an external validation set of 222 compounds, for which all contestants were asked to predict their possible response against all targets. The ground truth of these compounds was also made available after the teams' ranking was released. Table 12 shows the number of records for each target.

The Tox21 Data challenge follows the open-innovation principles[337] to crowdsource scientists' efforts in analyzing HTS data generated through the Tox21 project. It aspires to predict the pathways' interference of chemicals using only their chemical structures. Such predictions can therefore guide regulators and participating governmental agencies in identifying the chemicals (either drugs or industrial) that carry the highest concern for human and environmental risks. The aim of this study is to describe the methodologies used by the winning corresponding author during the challenge (team: AMAZIZ)[338] and to extend the analysis on the chemical libraries beyond what was possible during the limited duration of the challenge. The study investigates a comprehensive approach on consensus modeling and analyzes multiple descriptor packages.

The pathway endpoints investigated were:

A. Estrogen receptor (ER)

Tox21 compounds library was screened for potentially acting as agonist at the estrogen receptor alpha. Such activators could lead to reproductive dysfunction(Aop:30). Two different cell lines were used:

- ER-alpha-UAS-bla GripTiteTM cell line (AID: 743077[340]): This cell line is developed by Invitrogen, Carlsbad, CA, USA. Cells contain a beta-lactamase reporter gene controlled by an Upstream Activator Sequence (UAS) stably integrated into HEK293 cells. Throughout this work, this dataset is referred to as (ER-LBD).
- BG1-Luc-4E2 cell line (AID: 743079[341]): Dr. Michael Denison from University of California provided the cell line. Cells endogenously express the full-length ER-alpha and are stably transfected with a plasmid containing four estrogen responsive elements (ERE) under the control of an upstream luciferase reporter gene. Throughout this work, this dataset is referred to as (ER-full).

B. Androgen receptor (AR)

The ability of chemical compounds to agonist the estrogen receptor alpha was measured in 2 different cell lines that were used to screen the Tox21 compound library.

- GeneBLAzer AR-UAS-bla-GripTite cell line (AID: 743053[342]): This cell line is provided by Invitrogen, Carlsbad, CA, USA. Cells contain a beta-lactamase reporter gene controlled by an upstream activator sequence (UAS) stably integrated into HEK293 cells. To assess the possibility for false positive or false negative results, the chemicals were also tested for auto fluorescence, which could interfere with the biological target readout. Throughout this work, this dataset is referred to as (AR-LBD).
- MDA-kb2 AR-luc cell line (AID: 743040[343]): This cell line was deposited by Wilson et al. It is human breast carcinoma cell line that was stably transfected with a luciferase reporter gene under control of the MMTV promoter containing response elements for both androgen receptor (AR) and glucocorticoid receptor (GR). Throughout this work, this dataset is referred to as (AR-full).

C. Aryl hydrocarbon receptor (AHR) (AID: 743122[344])

A cell based HepG2-AhR-luc assay, developed by Dr. Michael S. Denison (University of California at Davis), was used to assess the activation of AhR for Tox21 compounds. The human hepatocellular carcinoma (HepG2) Cells were stably transfected with an Ah receptor-responsive firefly luciferase reporter gene plasmid carrying 20 dioxin responsive elements and luciferase reporter gene. AhR activation leads to an increase in luciferase activity and therefore ligands can be detected. Cell viability was measured using CellTiter-Fluor assay in the same wells to detect chemical cytotoxicity against the HepG2-AhR-luc cell line.

D. Peroxisome proliferator-activated receptor gamma (PPAR-gamma) (AID: 743140[345])

GeneBLAzer PPAR gamma UAS-bla HEK293H cell line was used in this assay. This cell line is provided by Invitrogen, Carlsbad, CA, USA. Cells contain a beta-lactamase reporter gene controlled by an upstream activator sequence (UAS) stably integrated into HEK293H cells. To assess the possibility for false positive or false negative results, the chemicals were also tested for auto fluorescence, which could interfere with the biological target readout.

E. Nuclear factor (erythroid-derived 2)-like 2/antioxidant responsive element (Nrf2/ARE) (AID: 743219[346])

The CellSensor ARE-bla Hep-G2 assay was used to assess the activation of the report gene and thus identify chemicals that stimulate oxidative stress. The cells contain a beta-lactamase reporter gene controlled by the Antioxidant Response Element (ARE) stably integrated into HepG2 cells. Fluorescence intensity was measured to assess the activation of the responsive element. Cell viability was measured, using CellTiter-Glo assay (Promega, Madison, WI), in the same wells to detect chemical cytotoxicity against the ARE-bla cell line. Furthermore, compounds were tested for auto fluorescence to identify the possibility for false target readout.

F. Aromatase enzyme inhibitors (AID: 743139[347])

The MCF-7 aro ERE cell line (human breast carcinoma), as provided by Dr. Shiuan Chen (Beckman Research Institute of the City of Hope), was used to identify aromatase inhibitors. Cells were stably transfected with a promoter plasmid, pGL3-Luc, encompassing three repeats

of estrogen responsive element (ERE). Cell viability was measured using CellTiter-Fluor assay (Promega, Madison, WI) in the same wells to detect chemical cytotoxicity against the MCF-7 aro ERE cell line.

G. ATAD5 receptor (ATAD5) (AID: 720516[348])

A cell-based assay using embryonic kidney cells (HEK293T) was used to screen the Tox21 compounds library. The assay was developed by Kyungjae Myung (NHGRI, NIH) to detect any enhanced Level of Genome Instability Gene 1 (ELG1; human ATAD5) protein, which increase in response to different kinds of DNA damage. The assay uses a luciferase reporter-gene tagged with ATAD5 to measure the induction of ELG1. Therefore, an increase in luciferase activity marks a chemically induced genetic stress. Cytotoxicity was also assessed through measuring protease activity within live cells.

H. Heat shock response element (HSE) (AID: 743228[349])

HSE-bla HeLa cell line was utilized in this HTS assay. This cell line is provided by Invitrogen, Carlsbad, CA, USA. Cells contain a beta-lactamase reporter gene controlled by the heat shock response elements. Cell viability was measured, using CellTiter-Glo assay (Promega, Madison, WI), in the same wells to detect chemical cytotoxicity against the HSE-bla cell line. Furthermore, to assess the possibility for false positive or false negative results, the chemicals were also tested for auto fluorescence, which could interfere with the biological target readout.

I. Disruptors of the mitochondrial membrane potential (MMP) (AID: 720637[350])

An assay based on a homogenous cell-based assay with a water-soluble mitochondrial membrane potential sensor (m-MPI, Codex Biosolutions, MD) was applied to the Tox21 compounds to identify those that can induce mitochondrial toxicity. In healthy cells, the water-soluble dye accumulates in the mitochondria as aggregates, causing red fluorescence. In case of a decrease in MMP, the dye cannot accumulate in the mitochondria and thus remains in the cytoplasm as monomers causing green fluorescence. Cytotoxicity was also assessed in the same wells to detect chemical cytotoxicity through the quantitation of ATP present.

J. Agonists of the p53 signaling pathway (P53) (AID: 720552[351])

Using CellSensor p53RE-bla HCT-116 cell line, the Tox21 compounds were screened. This cell line is provided by Invitrogen, Carlsbad, CA, USA. Cells contain a stably integrated beta-lactamase (BLA) reporter gene controlled by the p53 response elements. Fluorescence intensity was measured to assess the activation of the responsive element. Cell viability was measured by measuring the intra cellular ATP content in the same wells to detect chemical cytotoxicity against the p53 RE-bla HCT-116 cell line.

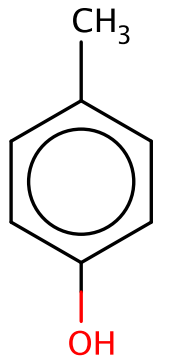
4.3.2 Data acquisition and curation

Data were downloaded from the Tox21 challenge website[336] in both SDF and SMILES formats. The files contained the molecular representation (SDF or SMILES), a molecule name as well as the target response. In addition, SDF files contained few extra tags for the DSSTox compound ID (DSSTox_CID), the chemical formula and the average mass (FW). Both file formats were compared to examine consistency. KNIME[193] was used to compare the structures and responses in both file formats. The data covered 12 pathway endpoints covering the 'Nuclear

Receptor Signaling Panel' (7 assays) and the 'Stress Response Panel' (5 assays). All assay endpoints are listed in Table 12.

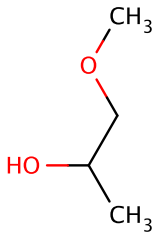
For each molecular pathway endpoint, both training and leaderboard test sets were combined to form a whole training set. Some molecules were presented multiple times (i.e., exact SMILES representation despite different molecule names). The basis for such duplicated records may be the result of intentional repetitive testing for quality control purpose. The Online CHEmical database and Modeling environment platform (OCHEM)[183] was used to check records duplication. It calculates the INCHI[180] key structure hash to compare structures. Some records showed different experimental responses despite exhibiting the same molecular structures. Figure 26 shows an example of such duplicates with conflicting experimental measurements. Table 12 shows the number of records per dataset as well as the number of unique molecules.

SDF molecular representations included no 3D coordinates. The files showed the signature of Marvin tool for compiling the SDF files. A single molecular representation in the final evaluation dataset (ID: NCGC00357026-01) held an ambiguous aromaticity. The Marvin tool was used to adopt a corrected aromatic diazole ring structure as shown in Figure 27.



Name: p-Kresol
SMILES: CC1=CC=C(O)C=C1

NCGC00013272-01	Active
NCGC00091519-04	Inactive
NCGC00257956-01	Inactive
NCGC00253980-01	Inactive
NCGC00258667-01	Inactive



Name: 1-methoxypropan-2-ol
SMILES: COCC(C)O

NCGC00256978-01	Active
NCGC00259352-01	Inactive

Figure 26. Example of conflicting training data. The examples shown were obtained from the estrogen nuclear receptor subset. In some cases, such as p-Kresol, it could be reasonable to assume that the compound would be inactive (4 records shows inactive against only one active record). In other cases, such as methoxypropan-2-ol, it is not possible tell whether the compound was truly activating the estrogen nuclear receptor (with one record in each class). Compounds are compared using their calculated INCHI keys generated from the SDF representation. All twelve targets showed similar cases.

Figure 27. To the left, Compound NCGC00357026-01 provided structure from the smiles and SDF files as depicted by Marvin Sketch. On the right, the corrected aromatic diazole ring adopted.

Table 12. Number of records and unique molecules in each dataset. Nuclear receptor (nr) assay panel contained 7 assays while the stress response (sr) assay panel covered 5 assays

Molecular pathway endpoint	Training set records [unique molecules]	Test set records	Complete training set records [unique molecules]
Nuclear Receptor Signaling Panel			
Aryl hydrocarbon receptor (nr-ahr)	8169 [6716]	272	8441 [6988]
Androgen receptor MDA-kb2 AR-luc cell line (nr-ar)	9362 [7468]	292	9654 [7760]
Androgen receptor GeneBLAzer AR-UAS-bla-GripTite cell line (nr-ar-lbd)	8599 [6927]	253	8852 [7180]
Aromatase enzyme (nr-aromatase)	7226 [5966]	214	7440 [6180]
Estrogen receptor alpha BG1-Luc-4E2 cell line (nr-er)	7697 [6334]	265	7962 [6599]
Estrogen receptor alpha ER-alpha-UAS-bla GripTiteTM cell line (nr-er-lbd)	8753 [7138]	287	9040 [7425]
Peroxisome proliferator-activated receptor gamma (nr-ppar-gamma)	8184 [6607]	267	8451 [6874]
Stress Response Panel			
Nuclear factor (erythroid-derived 2)-like 2/antioxidant responsive element (Nrf2/ARE) (sr-are)	7167 [5959]	234	7401 [6193]
ATAD5 receptor (sr-atad5)	9091 [7256]	272	9363 [7528]
Heat shock factor response element (sr-hse)	8150 [6617]	267	8417 [6884]
Mitochondrial membrane potential (sr-mmp)	7320 [5941]	238	7558 [6179]
p53 signaling pathway (sr-p53)	8634 [6931]	269	8903 [7200]

4.3.3 Methods

A. Software tools

Throughout this study, different KNIME[193] workflows were used to explore the data, initialize the QSAR model building process on OCHEM (Figure 28) and download the models' predictions (Figure 29). All QSAR models were built using OCHEM. CRAN R[202] was used to build consensus models and analyze models' performance.

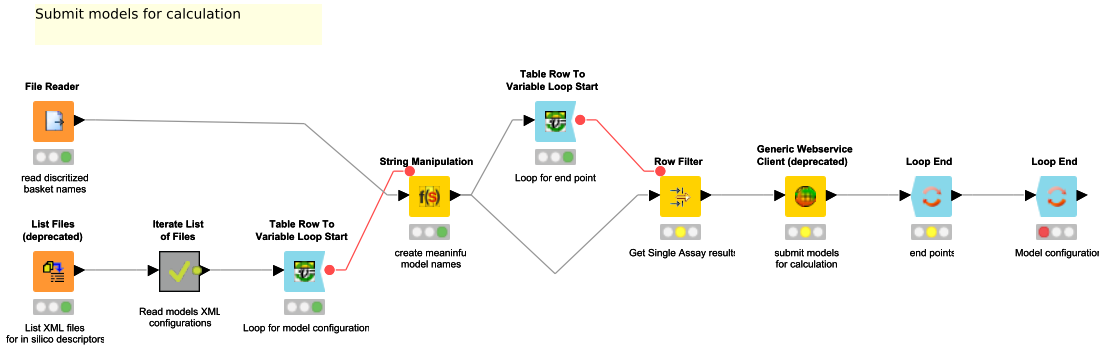


Figure 28. KNIME workflow used to submit models for calculation on OCHEM. The workflow submits XML configuration with the specific instructions for the machine learning algorithm, descriptor packages as well as the descriptors prefiltering and chemical structure standardization instruction. The workflow utilizes a previously prepared set of chemicals uploaded to OCHEM (chemical baskets) that contain the training set for building the models.

B. In silico descriptors calculation

Ten descriptor packages were selected from OCHEM to be used for constructing QSAR models. These packages were compiled from multiple academic and commercial sources. The selected packages are: GSFrag[314], ISIDA fragments (length 2 - 4)[315], Chemaxon descriptors[313], Estate indices[311,312] & ALOGPS[309,352], CDK (using all constitutional, topological, geometrical, electronic and hybrid descriptors)[255], Inductive descriptors[316], Dragon 6[222], Adriana.Code[235], MERA & MerSy[246–248], QNPR (using SMILES representations - length 1 - 3 and a threshold of 5)[251]. Further details on these packages and their integration within OCHEM can be found in section 3.4 Molecular descriptors.

The same structure-preprocessing protocol was used prior to the calculation of any descriptor package utilizing Chemaxon Standardizer that is integrated within OCHEM workflow. The standardization workflow consisted of salt counter-ion removal, charge neutralization and the standardizing of certain chemotype representations; such as nitro groups and aromatic rings. For 3D descriptors, structural coordinates were optimized using CORINA[308] starting from a clean SMILES representation. Descriptors calculation failed for some chemicals, the number of failed molecules depends on the nature of the descriptor package. Reasons for calculation failure may be a large molecules size or undefined chemotypes. The count of failed molecules for each constructed model is available, together with the detailed modeling results, deposited in an open GitHub repository[317].

C. Machine learning

The associative neural networks (ASNN)[261,262] algorithm was used to construct all models as described in section 3.5 Machine learning algorithms.

D. Performance measures and validation protocol

Due to the unbalanced nature of the datasets, balanced accuracy was used throughout the study, as well as during the challenge, as the primary measure for comparing models' performance. It is important to notice that the challenge did not only account for the balanced accuracy but also the Area Under the Receiver Operating Characteristic (AUROC) curve[281].

Bagging[272] was used to validate the accuracy of the training set. Bagging is a meta-algorithm that involves the aggregation of many models, each of which is based on its own training set

("bag"). Bagging utilized the random sampling, with repetition, of many subsets of the training set. In each bagging meta-model constructed, an ensemble of 64 models was developed. For each model in the ensemble the training examples were selected randomly from the original training set allowing duplicates (i.e., resampling with replacement). The prediction of each classification was determined by majority voting among the ensemble members. Stratified bagging[291] was used as the validation protocol. It also served to handle the unbalance of the training set[292]. In the current implementation, for each of the 64 models in an ensemble, equal numbers of active and inactive compounds were randomly selected. Thus, the size of the training set was always double the size of the minority class.

The calculation of statistical measures was done only using the validation set (out of bag compounds). For molecules with conflicting experimental measurements (see Figure 26), the class with more experimental measurements (majority vote) was selected. Molecules that showed equal number of active and inactive experimental measurements were excluded.

E. Consensus modeling

For each endpoint, consensus models were built using all possible combinations of the underlying ten models (each built using different *in silico* descriptor package), i.e., $\sum_{i=1}^{10} C_i^{10}$. In total, 12276 models (1023 x 12 endpoints) were constructed. Simple averaging of the predictions was used for building each of the consensus models.

Two approaches for consensus model selection were investigated in this study. The first approach considers consensus models that show the highest validated balanced accuracy on the training set. The second approach considers consensus models which combine models built with all ten descriptor packages regardless of the resulting validation balanced accuracy. Both approaches performed comparatively well with no significant difference in most cases.

F. Applicability domain

In this study, a distance-based method was used to estimate the applicability domain for all models. The distance to model is defined in the property space (rather than the descriptor space)[196]. This approach uses the standard deviation between the predictions of an ensemble of models (generated through bagging) as a measure of distance.

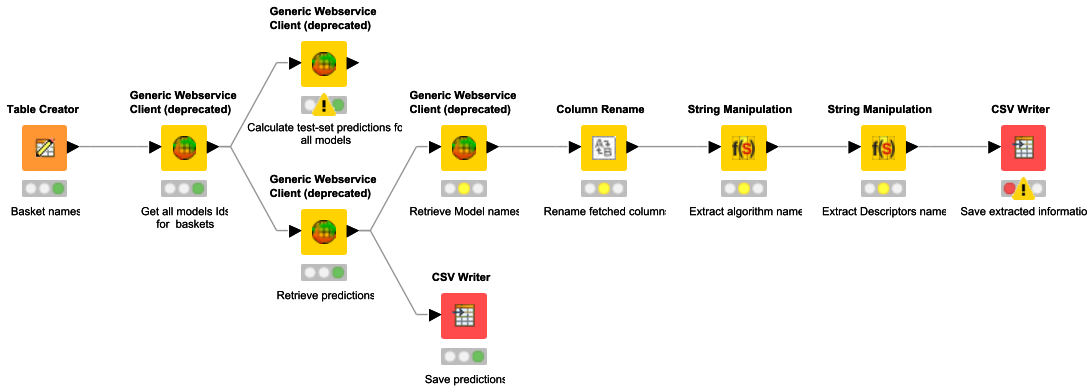


Figure 29. KNIME workflow used to retrieve QSAR model IDs from OCHEM. The model predictions on the training set are retrieved for analysis of models' performance. Information on the model name are also retrieved and used to store meta-information on the models' algorithms and descriptors. Finally, KNIME sends instructions to OCHEM to calculate predictions for the test set compounds.

4.3.4 Results and discussion

A. Individual models

In total 10 descriptor packages were used to model twelve *in vitro* assay endpoints resulting in 120 QSAR models constructed with 64-bagging-validation. Different endpoints showed varying success. All models are published online and may be examined through http://www.ochem.eu/mode/[model-id] replacing [model-id] with the respective model identification number available in the results tables. Users can see a model's summary with performance statistics, applicability domain graphs as well as apply the model to new compounds. Figure 30 shows the balanced accuracy of all 120 models as grouped by their respective targets. Other statistical parameters such as specificity, sensitivity, Matthews's correlation coefficient (MCC) and overall accuracy are deposited in an open GitHub repository[317] where the summary statistics of all models are publicly available.

To compare descriptor packages success, each package was given a score from one to ten according to its rank (a score of 10 was given to the descriptor package contributing to the model with the highest balanced accuracy and 1 for the lowest). The scores were summed for all endpoints. The final rank of descriptors can be seen in Table 13. Dragon and CDK descriptor packages shared the top positions in both training and evaluation sets.

As shown in Figure 30, a direct correlation exists between the validated training and the evaluation set balanced accuracies except for the nr-ar-lbd endpoint. This can also be seen by directly plotting the training set against the evaluation set balanced accuracies as shown in Figure 31.

Table 14 lists the performance of the single descriptor package models with the highest balanced accuracy for each pathway endpoint together with their corresponding performance on the final evaluation set. The highest balanced accuracy achieved by any team (measured on the evaluation set) during the challenge was reported online[353]. It is also shown in Table 3 (referred to as "winning balanced accuracy") for reference.

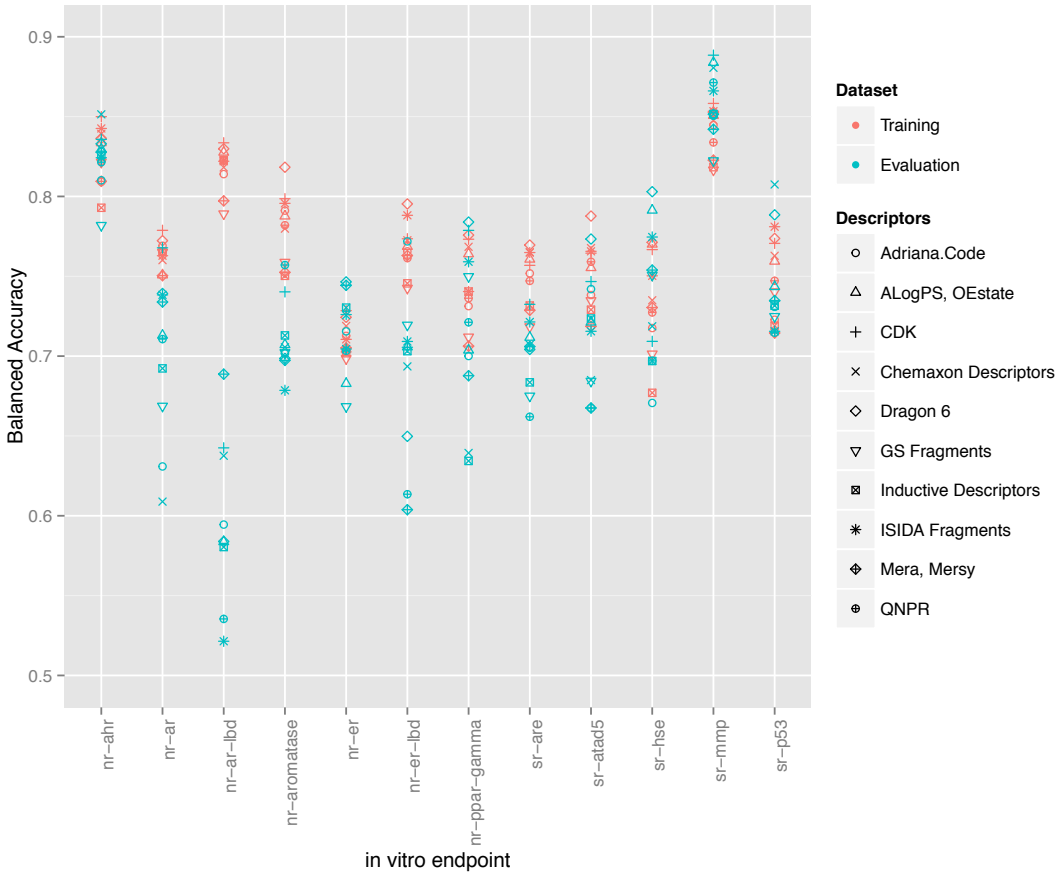


Figure 30. Training set balanced accuracies for all 120 models as grouped by their respective endpoints. Red points represent the validated (through bagging) balanced accuracies calculated on the training set. Blue points represent the balanced accuracy on the evaluation set.

Table 13. Comparison of the performance of different descriptor packages in constructing QSAR models for *in vitro* pathway disruption prediction

Descriptors package	Training total score	Training set rank	Evaluation total score	Evaluation set rank
Dragon 6	111	1	86	2
CDK	105	2	98	1
ISIDA Fragments	88	3	65	5
Chemaxon Descriptors	79	4	71	4
ALOGPS, OEstate	73	5	79	3
Adriana.Code	55	6.5	55	8
QNPR	55	6.5	45	9
Inductive Descriptors	36	8	57	7
MERA, MerSy	30	9	62	6
GS Fragments	28	10	42	10

Table 14. Performance of the single-descriptor-package models with the highest training set balanced accuracy for each pathway endpoint. The balanced accuracies of winning models in the data challenge[353] are shown for reference. Cases where models perform better than wining balanced accuracy are underlined. Three significant digits are shown for comparison. However, the difference in the balanced accuracy in many cases is not significant to justify some models as being more superior than others. The upper and lower boundaries for balanced accuracies as well as p-values are available, together with detailed QSAR results, from an open GitHub repository[317].

Molecular pathway endpoint	Descriptors package	Training balanced accuracy	Evaluation balanced accuracy	Wining balanced accuracy (evaluation set)
nr-ahr	CDK	0.850	0.836	0.853
nr-ar	CDK	0.779	0.768	0.736
nr-ar-lbd	CDK	0.834	0.643	0.650
nr-aromatase	Dragon 6	0.818	0.699	0.737
nr-er	CDK	0.728	0.726	0.749
nr-er-lbd	Dragon 6	0.795	0.650	0.715
nr-ppar-gamma	Dragon 6	0.776	0.784	0.785
sr-are	Dragon 6	0.770	0.704	0.729
sr-atad5	Dragon 6	0.788	0.773	0.741
sr-hse	Dragon 6	0.771	0.803	0.799
sr-mmp	CDK	0.858	0.888	0.904
sr-p53	ISIDA Fragments	0.781	0.716	0.765

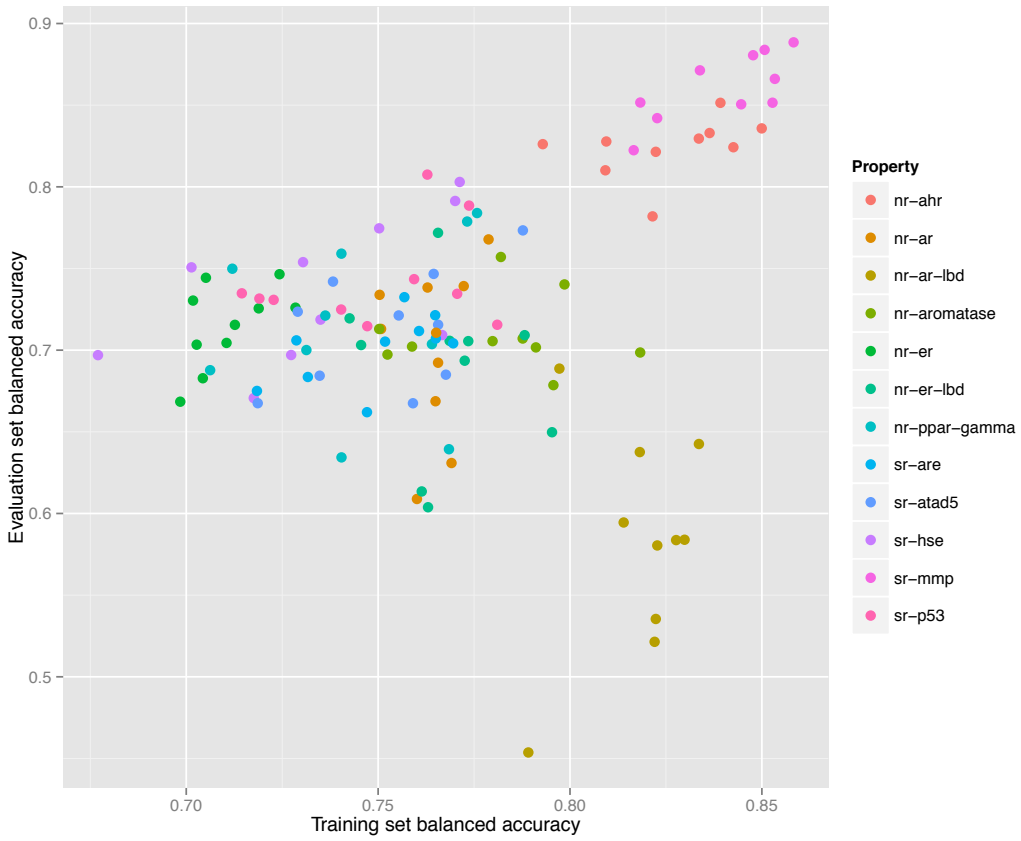


Figure 31. Correlation between training and validation set balanced accuracies for 120 models constructed for 12 endpoints using 10 individual descriptor packages for each endpoint.

B. Consensus modeling

For comparison, Table 16 shows the performance of the consensus models involving all ten underlying descriptor packages for each pathway endpoint. In seven endpoints, the predictive ability of these models on the evaluation set slightly exceeded those of the highest validated balanced accuracy.

Descriptor packages differed in their success in representing the chemical structures. Some descriptor packages failed during the calculation phase for some of the molecules (e.g., reporting a chemical structure being too large for calculation). Therefore, models based on them would be deprived from any information gain from those failed molecules (i.e., will have a smaller training set size). A QSAR model built on such descriptors may show good statistics on the smaller training set but fail to perform similarly for an external evaluation set.

The second approach has the advantage of covering the largest number of molecules by compensating for the failure of some packages in descriptors calculation. It can also compensate for some packages bias by offering a wider range of molecular representations. However, it might suffer from the disadvantage of picking noise from descriptor packages with particularly bad performance. It also involves the highest computational expense, as applying such models to new molecules would require calculation of all descriptors from ten packages. On the other hand, the first approach has the advantage of picking fewer descriptor packages with the highest performance.

In Table 15, the consensus models with highest validated balanced accuracy based on the training set for each endpoint are listed as well as their respective performance on the evaluation set. For all endpoints, consensus modeling could improve the performance on the training set. In six endpoints, the consensus models' predictive ability on the evaluation set would have been better than the winning balanced accuracy. The developed consensus models can be accessed at http://amaziz.com/article/tox21.

For comparison, Table 16 shows the performance of the consensus models involving all ten underlying descriptor packages for each pathway endpoint. In seven endpoints, the predictive ability of these models on the evaluation set slightly exceeded those of the models showing highest validated balanced accuracy (Equation 15)

Descriptor packages differed in their success in representing the chemical structures. Some descriptor packages failed during the calculation phase for some of the molecules (e.g., reporting a chemical structure being too large for calculation). Therefore, models based on them would be deprived from any information gain from those failed molecules (i.e., will have a smaller training set size). A QSAR model built on such descriptors may show good statistics on the smaller training set but fail to perform similarly for an external evaluation set.

The second approach has the advantage of covering the largest number of molecules by compensating for the failure of some packages in descriptors calculation. It can also compensate for some packages bias by offering a wider range of molecular representations. However, it might suffer from the disadvantage of picking noise from descriptor packages with particularly bad performance. It also involves the highest computational expense, as applying such models to new molecules would require calculation of all descriptors from ten packages. On the other hand, the first approach has the advantage of picking fewer descriptor packages with the highest performance.

Table 15. Performance of the consensus models with the highest training set balanced accuracy for each pathway endpoint. The balanced accuracies of winning models in the data challenge[353] are shown for reference. Cases where models perform better than wining balanced accuracy are underlined. Three significant digits are shown for comparison. However, the difference in the balanced accuracy in many cases is not significant to justify some models as being more superior than others. The upper and lower boundaries for balanced accuracies as well as p-values are available, together with detailed QSAR results, from an open GitHub repository[317].

Molecular pathway endpoint	Training set balanced accuracy	Evaluation set balanced accuracy	Wining balanced accuracy (evaluation set)	Ids for models used in building consensus
nr-ahr	0.865	0.859	0.853	512
nr-ar	0.785	0.752	0.736	515
nr-ar-lbd	0.838	0.592	0.650	516
nr-aromatase	0.824	0.715	0.737	513
nr-er	0.736	0.756	0.749	517
nr-er-lbd	0.810	0.726	0.715	518
nr-ppar-gamma	0.802	0.741	0.785	514
sr-are	0.799	0.730	0.729	534
sr-atad5	0.809	0.734	0.741	519
sr-hse	0.794	0.767	0.799	520
sr-mmp	0.882	0.900	0.904	521
sr-p53	0.795	0.783	0.765	522

Table 16. Performance of the consensus models involving all 10 descriptor packages for each pathway endpoint. The balanced accuracies of winning models in the data challenge[353] are shown for reference. Cases where models perform better than wining balanced accuracy are underlined. Three significant digits are shown for comparison. However, the difference in the balanced accuracy in many cases is not significant to justify some models as being more superior than others. More detailed models statistics were deposited to an open GitHub repository[317].

Molecular pathway endpoint	Training set balanced accuracy	Evaluation set balanced accuracy	Wining balanced accuracy (evaluation set)
nr-ahr	0.850	0.858	0.853
nr-ar	0.770	0.754	0.736
nr-ar-lbd	0.824	0.599	0.650
nr-aromatase	0.811	0.760	0.737
nr-er	0.730	0.744	0.749
nr-er-lbd	0.794	0.756	0.715
nr-ppar-gamma	0.779	0.759	0.785
sr-are	0.789	0.707	0.729
sr-atad5	0.786	0.727	0.741
sr-hse	0.766	0.773	0.799
sr-mmp	0.875	0.903	0.904
sr-p53	0.784	0.759	0.765

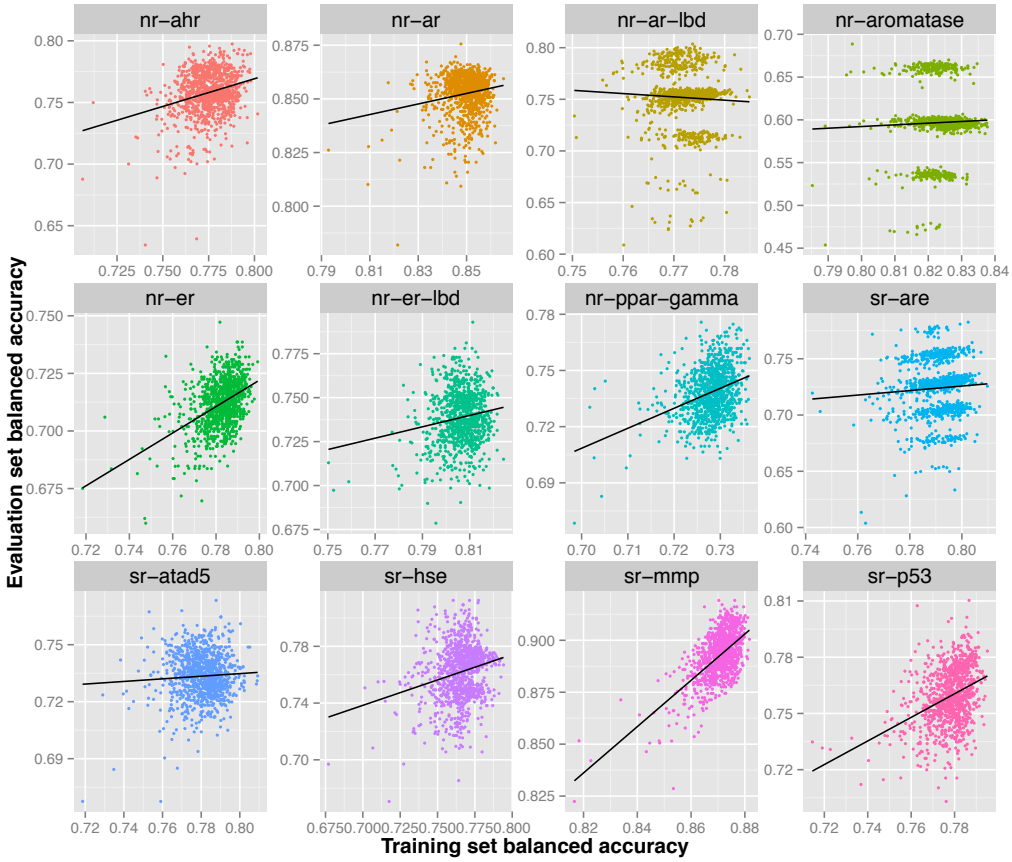


Figure 32. Each sub-figure shows the performance of 1023 consensus models constructed for a single endpoint with x-axis representing the validated balanced accuracy on the training set and y-axis shows the balanced accuracy on the evaluation set. A positive trend line can be noticed with all endpoints except nr-ar-lbd.

Table 17. Models used for the final submission by team AMAZIZ during the Tox21 challenge. Consensus models involving all 10 descriptor packages (sr-are and sr-mmp) failed for the calculation of 23 molecules of the evaluation set and were replaced by simpler models, based on the consensus of 3 models only, predicting these molecules.

Molecular pathway endpoint	Ids for models used in building consensus
nr-ahr	523
nr-ar	524
nr-ar-lbd	525
nr-aromatase	351
nr-er	526
nr-er-lbd	527
nr-ppar-gamma	528
sr-are	533
sr-atad5	529
sr-hse	530
sr-mmp	531
sr-p53	532

4.3.5 Summary of Tox21 analysis aspects

Using QSAR for modeling *in vitro* assays representing molecular pathways showed promising success with balanced accuracies reaching up to more than 85% for some endpoints as shown in Table 15. The relatively high balanced accuracies among models confirm the possibility of modeling HTS *in vitro* assays using *in silico* descriptors as reported in the study 4.1 ToxCast™ phase I[187].

Bagging validation provides a good indication for the models' predictive ability on external validation sets (Figure 31). Stratified bagging may counter the unbalanced nature of the training set and reduce bias towards the majority class. However, it is important to realize that due to the variance in data, selecting the highest validated accuracy does not guarantee the highest predictive ability for any particular evaluation set.

The Androgen receptor GeneBLAzer AR-UAS-bla-GripTite cell line endpoint showed exceptional difficulty in modeling. Big discrimination exists between validated performance on the training set and the prediction ability on the evaluation set. Indeed, the endpoint had the lowest success in modeling in the challenge with the wining model being able to achieve a balanced accuracy of only 65% (the lowest among all endpoints).

Further investigation of the models constructed for this endpoint shows multiple models that would have been able to achieve a higher predictive ability on the evaluation set (0.75-0.80) as shown in Figure 32. However, such models did not show the highest validated balanced accuracy and were thus not selected. The lack of direct correlation between validated balanced accuracy and predictive ability on the evaluation set (Figure 31) may suggest that the split of the whole cluster of chemicals into training and evaluation sets may not have been random.

Consensus modeling improves the predictive ability of models as signified by both validation and evaluation set accuracies. This can be due to the complementarity between descriptor packages, therefore capturing more aspects of the molecular structures. Presence of more packages may also compensate for each other's failure to represent certain chemical scaffolds and thus covering the entire the training set.

Due to the time constraint during the challenge, the consensus models' selection for the author (team AMAZIZ) was based on expert knowledge including the criteria discussed in this study, namely the performance of the models with regard to their balanced accuracy and to a lesser extent the AUROC, preference to descriptor packages, which show more success in representing a larger size of the training set and the simplicity of the underlying descriptor packages (e.g., 2D descriptors are simpler in calculation than 3D descriptors, as they lack the need for 3D optimization). Table 17 shows the models that were used for the final submission of team AMAZIZ in the challenge. All models can be accessed through their identification numbers for further analysis and to run predictions on new compounds. This study represents a systemic approach to consensus models' selection as well as a deeper analysis beyond the challenge.

The combination of the workflow tool (KNIME), the QSAR modeling platform (OCHEM) and the statistical package (CRAN R) allowed the creation and analysis of thousands of models

with high efficiency. Finally, the use of HTS *in vitro* assays to construct QSAR models that can predict certain molecular pathways' perturbation paves the way towards a better understanding for the mode of chemical toxicity and allows for prioritization of testing efforts. This is in line with the vision of EPA and ECHA for replacing unnecessary animal toxicity testing, rapidly filing information gaps, and achieving higher outcomes with available efforts and resources.

4.4 Pregnane X receptor activators (PXR)

A library of 1889 compounds was screened against human PXR activation using the DPX-2 cell line, provided by Puracyp Inc. The HepG2 cells were co-transfected with a PXR response element and a luciferase construct containing CYP3A4 promoter. Therefore, an increase in luciferase activity marks compounds that activate the PXR pathway. The original assay is reported on PubChem BioAssay AID: 720659[354]

4.4.1 Data acquisition and curation

Data was downloaded from PubChem BioAssay AID: 720659[354]. Records with activity being reported as inconclusive were neglected (433 records). Furthermore, 21 molecules were reported as being both active and inactive (duplicate records with mismatching results), these molecules were excluded from the analysis. The final training data set included 1889 unique compounds of which 205 were active.

4.4.2 Methods

All QSAR models were built using OCHEM. Eleven descriptor packages were used to represent the molecular features using nine machine-learning algorithms to build the QSAR models. The models were validated through 5-fold stratified cross validation and stratified bootstrap aggregation with 64 bags. The machine learning algorithms used are k-Nearest Neighbors (kNN), Associative neural networks (ASNN)[261,262], C4.5 decision tree (J48)[200,201], multiple linear regression analysis (MLRA), fast stagewise multiple linear regression (FSMLR), partial least squares (PLS), random forests (RF), LADTree and support vector machine (SVM). Section 3.5 Machine learning algorithms cover the algorithms configuration in details.

The descriptor packages evaluated were Adriana.Code (3D)[235,236], CDK (3D)[255], Chemaxon calculators (3D)[313], Dragon 6 (3D)[221,222], Estate[311,355,356], ISIDA Fragments[315], GSFrag[242,314], Inductive descriptors (3D)[244,316,357], MERA (3D)[358], Spectrophore fingerprints (3D)[359], QNPR(Quantitative Name Property Relationship)[251]. Section 3.4 Molecular descriptors contains details about each descriptor package.

For each model, sensitivity, specificity, accuracy, balanced accuracy, MCC, and AUROC were calculated. As the dataset is unbalanced, the balanced accuracy (BAC) was used as the primary metric for judging models' performance. The applicability domain based on a distance-to-model approach was estimated for all models as described in section 3.10 Models applicability domain (AD).

4.4.3 Results and discussion

Table 18 shows the balanced accuracies for all 108 QSAR models built using different machine learning algorithms (columns) and descriptor packages (rows) for the prediction of PXR activation. All models were validated using 5-fold cross validation. Similarly, Table 19 shows the balanced accuracies for the 108 QSAR models built using the same machine learning algorithms and descriptor packages, which were validated using stratified bootstrap aggregation. In general, Bagging validation resulted in a slightly better performance for the same machine learning algorithms and descriptor combinations.

The highest validated balanced accuracy for a single model was based on the dragon descriptor package and the associative neural networks (BAC: 83.6% ± 0.9). The final model is

published on OCHEM and can be accessed through the link: http://amaziz.com/model/6816904. The area under the ROC curve for this model is 0.91 ± 0.01.

The applicability domain of the model was estimated using the distance-to-model approach. Figure 33 shows a plot of the applicability domain of the aforementioned model as a function of its bagging standard deviation. The lower the deviation is, the higher the model's accuracy. Therefore, the bagging STD can be used to estimate the confidence in prediction for new molecules. When considering the training set, Figure 34 shows that more than 50% of the compounds were predicted with an accuracy higher then 90%.

Table 18. Balanced accuracies for 108 QSAR models built using different machine learning algorithms (columns) and descriptor packages (rows) for the prediction of PXR activation. All models were validated using 5-fold cross validation

	ASNN	kNN	SVM	FSMLR	MLRA	PLS	J48	LADTree	RF
CDK	82%	82%	80%	77%	77%	74%	76%	81%	78%
Dragon6	84%	78%	73%	77%	75%	81%	77%	80%	80%
ALOGPS, OEstate	78%	72%	82%	74%	78%	79%	81%	81%	80%
ISIDA	75%	61%	80%	68%	72%	73%	65%	38%	71%
GSFrag	77%	73%	78%	77%	75%	67%	73%	76%	78%
MERA, MerSy	79%	74%	29%	70%	76%	70%	71%	78%	75%
Chemaxon Descriptors	79%	75%	46%	70%	79%	63%	82%	82%	83%
Inductive Descriptors	79%	72%	71%	41%	75%	74%	70%	74%	77%
Adriana	81%	78%	50%	81%	76%	75%	77%	81%	77%
Spectrophores	75%	69%	46%	65%	68%	63%	69%	68%	71%
QNPR	73%	59%	77%	72%	67%	74%	72%	69%	71%

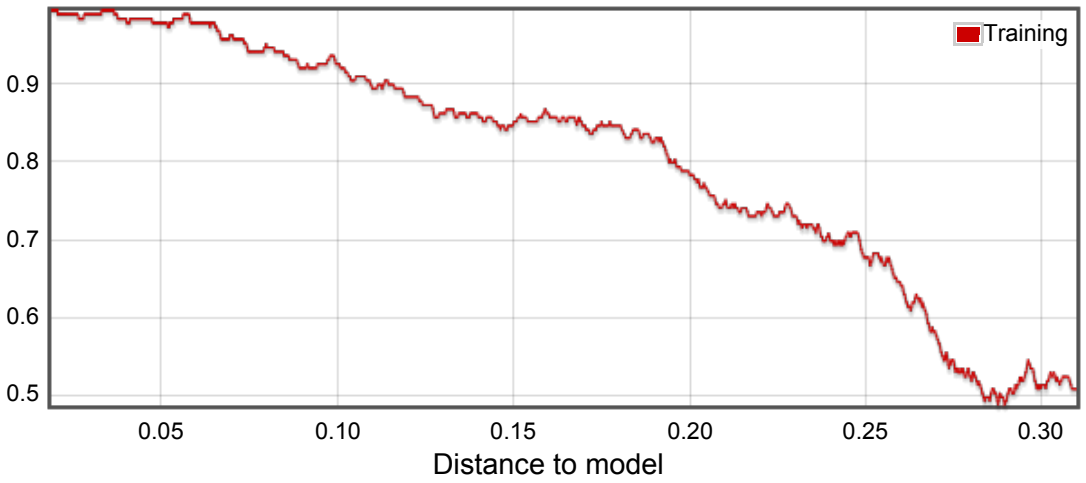


Figure 33. Williams plot showing the applicability domain of the aforementioned model as a function of the bagging standard deviation. The lower the deviation is, the higher the model's accuracy. The bagging STD can thus be used to estimate the error in prediction for new molecules.

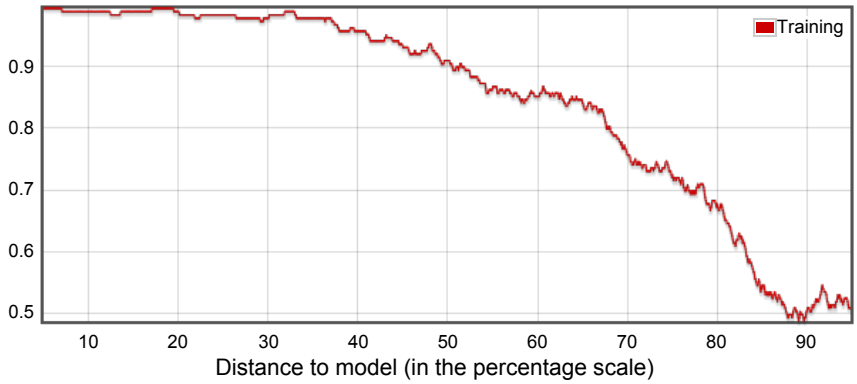


Figure 34. Williams plot showing the performance of the model within the dataset. 50% of the dataset is predicted with >90% balanced accuracy.

Table 19. Balanced accuracy for 108 QSAR models built using different machine learning algorithms (columns) and descriptor packages (rows) for the prediction of PXR activation. All models were validated using bootstrap aggregation (64-stratified bagging)

	ASNN	kNN	SVM	FSMLR	MLRA	PLS	J48	LADTree	RF
CDK	82%	81%	80%	80%	79%	81%	80%	81%	82%
Dragon6	84%	81%	77%	80%	81%	83%	80%	82%	81%
ALOGPS, OEstate	82%	72%	84%	81%	78%	80%	80%	83%	83%
ISIDA	77%	62%	81%	78%	77%	79%	67%	77%	75%
GSFrag	81%	77%	78%	80%	78%	78%	79%	80%	79%
MERA, MerSy	81%	77%	50%	79%	80%	72%	80%	81%	79%
Chemaxon Descriptors	81%	79%	53%	80%	81%	79%	83%	84%	85%
Inductive Descriptors	82%	77%	74%	61.20 %	78%	77%	75%	80%	80%
Adriana	83%	80%	50%	81%	81%	78%	81%	83%	84%
Spectrophores	77%	70%	50%	69%	70%	69%	74%	77%	76%
QNPR	76%	60%	81%	76%	77%	76%	72%	79%	74%

The highest validated balanced accuracy for a single model was based on the dragon descriptor package and the associative neural networks (BAC: 83.6% ± 0.9). The final model is published on OCHEM and can be accessed through the link: http://amaziz.com/model/6816904. The area under the ROC curve for this model is 0.91 ± 0.01.

The applicability domain of the model was estimated using the distance-to-model approach. Figure 33 shows a plot of the applicability domain of the aforementioned model as a function of its bagging standard deviation. The lower the deviation is, the higher the model's accuracy. Therefore, the bagging STD can be used to estimate the confidence in prediction for new molecules. When considering the training set, Figure 34 shows that more than 50% of the compounds were predicted with an accuracy higher then 90%.

4.4.4 Consensus modeling

A consensus model [model id: 7103264] was built using the four best performing models regardless of their underlying algorithm or descriptor packages. Another model [model id: 29021089] was constructed using the five best ASNN models (with different descriptor packages).

Both consensus models resulted in better model statistics. The consensus between the best 4 models resulted in BAC = 86% ± 1.0 and AUROC = 0.924 ± 0.01. The consensus between the best neural network models had BAC = 85% ± 1.0 and AUROC = 0.92 ± 0.01. Both models can be accessed on OCHEM platform for prediction of new compounds.

4.4.5 Summary of PXR activators prediction aspects

Analysis showed that data has good potential for modeling than the best single model. The performance for bootstrap aggregation models was slightly better than that of the 5-fold cross validation. On average the associative neural networks (ASNN) showed better performance that other machine learning methods. CDK descriptor package was generally better than others. Consensus modeling slightly improved the balanced accuracy reaching 86% ± 1.0.

4.5 Aryl hydrocarbon receptor activation – extended study

The Scripps Research Institute Molecular Screening Center (SRIMSC) conducted an HTS assay to identify compounds that act as agonists of the human AHR.

The cell-based assay measured the ability of chemical compounds to activate AHR signaling. The assay uses human hepatoma (HepG2) cells transfected with the AHR-dependent pGudLuc6.1-DRE plasmid (HG2L6.1c3 cell line), which expresses the firefly luciferase reporter gene under control of a minimal promoter containing a synthetic DRE[360,361362]. The experimental protocol was described in details in the PubChem Bioassay repository[363].

Each compound was tested in a single final nominal concentration of 5.0 UM. Cells were incubated with test compounds for 24 hours. Cell lysis was then performed and well luminescence detected using commercially available luciferase reagent. The concept behind this assays is that chemicals that act as agonists for the AHR will increase its activity and nuclear translocation. This will thus raise the activity of the DRE and increase the transcription of the luciferase transporter gene leading finally to higher luminescence. DMSO was used as a low control and Indirubin as a high control.

The criteria to judge a certain compound as being active or not depended on the amount of luminescence that was detected as compared to the two control compounds.

A % activation value was then calculated for every compound per Equation 40

$$\%\,Activation = 100\times \frac{\text{Luminescence}_{\text{Tested Compound}} - \text{Median Luminescence}_{\text{DMSO}}}{\text{Median Luminescence}_{\text{Indirubin}} - \text{Median Luminescence}_{\text{DMSO}}} \quad \text{Equation 39}$$

Compounds were then classified as either activators or non-activators. The cutoff for compounds to be considered as activators was to have %activation more than three standard deviations above the average for all compounds.

Compounds were also ranked according to their observed activation with the highest activity given a score of 100 and negative activities a score of zero. The non-activators had a score range 0-15 while activators had a score of 15-100

The assay was run on a total library of 324858 substances. According to INCHI calculations performed though OCHEM, these substances represented 324744 different compounds (324751 distinct compound identification numbers assigned by PubChem). Of the total library, only 7988 compounds were active according to the criteria above. Therefore, the ratio of activators to non-activators was 1:40.

4.5.1 Data acquisition and curation

The HTS assay data for AHR activators/non-activators were collected from the PubChem Bioassay database (AID: 2796)[363]. Data were downloaded in two files; An SDF file format for the chemical structures) and a CSV file for the assay results. The files were linked through a key field representing the PubChem Substance Id (PUBCHEM_SID) that is present in both files. The correlation was conducted through the software package KNIME[364] with no errors reported.

All chemical structures were standardized using Chemaxon Standardizer[365] integrated into OCHEM[366183], salt counter ions were strapped, ions were neutralized.

4.5.2 Methods

Eight descriptor packages were used to represent the molecular features. Six machine-learning algorithms were employed to build classification QSAR models. All models were built using OCHEM. All models were validated through stratified bootstrap aggregation with 64 bags. The machine learning algorithms used are k-Nearest Neighbors (kNN), Associative neural networks (ASNN)[261,262], multiple linear regression analysis (MLRA), fast stagewise multiple linear regression (FSMLR), random forests (RF) and support vector machine (SVM).

The descriptor packages evaluated were Adriana.Code (3D)[235,236], CDK (3D)[255], Chemaxon calculators (3D)[313], Dragon 6 (3D)[221,222], Estate[311,355,356], ISIDA Fragments[315], GSFrag[242,314], Inductive descriptors (3D)[244,316,357], MERA (3D)[358], Shape signatures (3D)[367], Spectrophore fingerprints (3D)[359].

For each model, sensitivity, specificity, accuracy, balanced accuracy, Matthew's correlation coefficient, and Area under the ROC curve were calculated. The positive predictive values were also an important parameter to consider as aim of the users who are likely to use the model would be to reduce the number of hits from the under-represented class of activators since AHR activators represent a potential toxicity hazard. As the dataset is unbalanced, the balanced accuracy (BAC) was used as the primary metric for judging model performance. Applicability domain based on a distance-to-model approach was estimated for all models as described in section 3.10 Models applicability domain (AD).

4.5.3 Results and discussion

Table 20 shows the balanced accuracy for all 108 QSAR models built using different machine learning algorithms (columns) and descriptor packages (rows) for the prediction of AHR activation.

The highest validated balanced accuracy for a single model was based on the CDK descriptor package and the associative neural networks (BAC: 81.9% ± 0.2). The final model is published on OCHEM and can be accessed through the link: http://amaziz.com/model/222449. The area under the ROC curve for this model is 0.894 ± 0.01.

The applicability domain of the model was estimated using the distance-to-model approach. Figure 35 shows a plot of the applicability domain of the aforementioned model as a function of its bagging standard deviation. The lower the deviation is, the higher the model's accuracy. Therefore, the bagging STD can be used to estimate the confidence in prediction for new molecules. When considering the training set, Figure 36 shows that more than 45% of the compounds were predicted with an accuracy higher than 80%.

4.5.4 Summary of the extended AhR study

Analysis showed that data has good potential for modeling. On average the associative neural networks (ASNN) showed better performance that other machine learning methods. CDK descriptor package was generally better than others. Consensus modeling slightly improved the balanced accuracy reaching 81.9% ±0.2

Table 20. Balanced accuracy for 48 QSAR models built using different machine learning algorithms (columns) and descriptor packages (rows) for the prediction of AhR activation. All models were validated using bootstrap aggregation (64-stratified bagging).

	ASNN	LibSVM	kNN	RF	FSMLR	MLRA
EState, ALOGPS	81.30%	80.10%	76.43%	79.60%	77.40%	77.70%
CDK	81.92%	80.70%	78.20%	78.10%	75.90%	78.60%
Chemaxon Descriptors	80.60%	78.60%	76.58%	77%	75.40%	75.30%
Adriana	81%	79.70%	75.90%	76.90%	69.10%	74.90%
Spectrophores	70.60%	69.10%	66.40%	67%	57.06%	67.20%
GSFrag	80.90%	78.10%	73.80%	77.50%	73.50%	75.80%
MERA	79.10%	78%	74.20%	75.90%	55.77%	77.40%
Inductive Descriptors	76.50%	75%	73%	75.20%	56.93%	73.70%

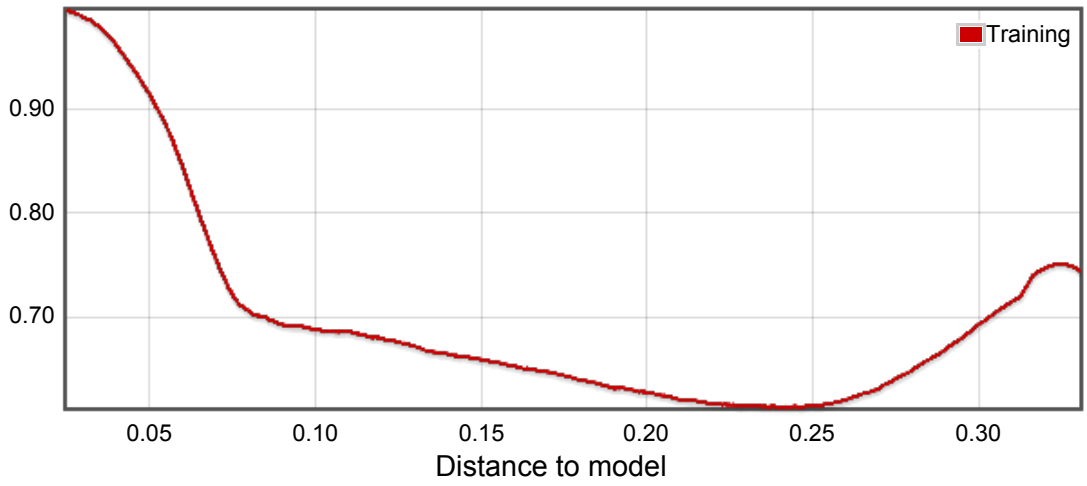


Figure 35. Williams plot showing the applicability domain of the best performing classification model (based on ASNN and CDK descriptors) as a function of the bagging standard deviation. The lower the deviation is, the higher the model's accuracy. The bagging STD can thus be used to estimate the error in prediction for new molecules.

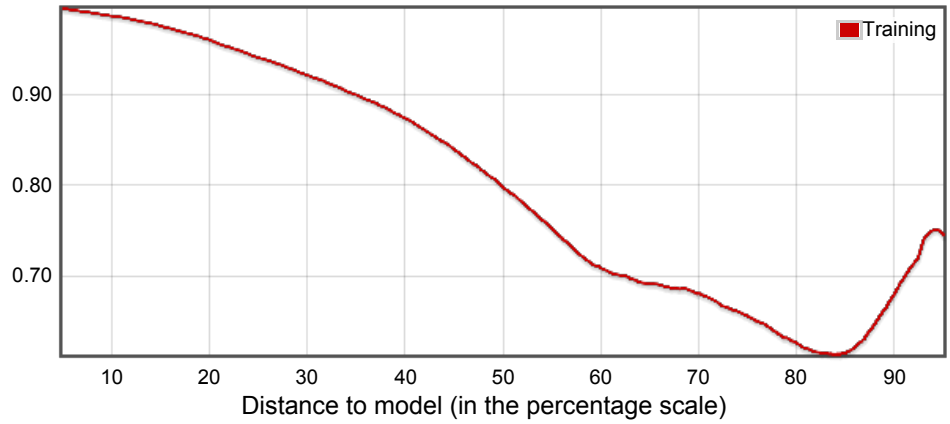


Figure 36. Williams plot showing the performance of the AHR activation model within the dataset. 45% of the dataset is predicted with >80% balanced accuracy.

5 Applications of the developed and validated computational methodologies

This chapter discusses two practical applications for employing QSAR studies in environmental risk assessment. The first application screens the large dataset of EINECS compounds (see 1.1 Chemicals regulations in the European Union) for potential pathway perturbations. The second application investigates a set of halogenated carbazole compounds emerging in the European and US ecology without being actively produced or imported.

Twelve QSAR models developed in the previous chapter (Table 15) will be used to assess the potential hazards of the chemicals on specific nuclear receptors as well as their ability to induce a stress response through selected biological pathways.

5.1 Toxicity-testing priority Score (ToPS) for EINECS

5.1.1 Introduction

As discussed earlier, EINECS compounds were considered as already "existing" in the European market between 1 January 1971 and 18 September 1981. They were left with insufficient data regarding their properties or effects making them difficult to regulate.

Therefore, EINECS compounds demonstrate a high need for filling information gaps and a good case of application for alternative testing approaches. The developed QSAR models can be directly used to provide such information.

5.1.2 Methods

A. Dataset

EINECS compounds are available on the OCHEM platform as a compound tag and can be used for screening and analysis. The dataset consists of 68779 unique compounds. A preliminary analysis of the compounds basic properties was conducted using CDK molecular descriptors. Table 21 shows the mean, median and standard deviation of some basic descriptors for EINECS compounds.

Table 21. Mean, median and standard deviation of some basic descriptors for EINECS compounds

Descriptor	Mean	Std. deviation	Skewness	Median
Hydrogen Bond Acceptors	3.79	4.05	3.03	3.00
Hydrogen Bond Donors	1.21	1.77	3.72	1.00
Topological Polar Surface Area	75.84	77.49	2.94	54.37
Molecular Weight	318.40	207.16	2.34	256.07
XLogP	3.33	4.18	2.40	2.71

B. QSAR models

The QSAR models with the highest balanced accuracies developed in 4.3 Tox21 project, listed in Table 15 were applied to the EINECS compounds. The applicability domain for the models was also assessed using the standard-deviation (STD)-based distance-to-models DMs. It has been shown to provide the best separation between accurate and inaccurate predictions in

multiple studies[368]. To calculate STD values, the consensus standard deviation was used as the distance measure. The DM calculated in such a way is referred to as CONSENSUS-STD. This procedure was applied to all QSAR models, providing 12 CONSENSUS-STD DMs.

C. Toxicity-testing Priority Score (ToPS)

A point-based scoring system was applied to all EINECS compounds. A compound would gain a full point if it is predicted to be toxic to the specified target with the respective QSAR model and where the compound falls completely within the applicability domain of the model (i.e., the model has 100% estimated accuracy for the compound's activity prediction). In practice, no given compound is likely to receive such perfect estimated prediction accuracy. As this analysis covered 12 molecular pathways, the maximum points a compound might collect is 12 and the minimum is 0.

Equation 40 describes the suggested point-based scoring system. For each model (m), the prediction (P_m) is multiplied by the applicability domain (AD_m) and a weight (w_m). The prediction can be binary where 0 represents a prediction of an inactive compound (i.e., with no perturbation to a given molecular pathway) and 1 for an active compound. It can also be any fraction in between. This is useful for models that provide a quantitative estimation of the binary classification.

Likewise, the applicability domain can be, in its simplest form, a binary classification where 0 means a compound is out of applicability domain and 1 means a compound is within the applicability domain of a given model. For more advanced quantitative approaches for applicability domain estimation (such as DM) a fraction between 0 and 1 is expected. This fraction describes how near, to the model's applicability, a certain prediction is, with 1 being the nearest.

To allow for flexibility, a weighting scheme can be included. Such scheme can be used to assign a higher priority to certain targets that might be deemed more important for certain investigation. In that case, the model predicting such target should be adjusted by multiplication by a desired weight (w_m). The weight can take values between 0 and 1. Equation 40 gives the general formula for the score.

$$\boldsymbol{PTS} = \sum_{\boldsymbol{m=1}}^{\boldsymbol{m}} \boldsymbol{w_n} * \boldsymbol{AD_n} * \boldsymbol{P_m} \qquad \text{Equation 40}$$

Where

w_m: is a weight given to the model associated with specific pathway endpoint based on its relative importance

AD_m: is the model's applicability domain for a given prediction

P_m: is the model's prediction

The models used in this study provide a quantitative estimation of the likelihood of classification (where, 0 - <0.5 represents the inactive class and 0.5 – 1 for the active class). Such value was used as a quantitative prediction score (P_m). The CONSENSUS-STD was used as

a quantitative measure for the distance to each model (AD_m). A uniform weight was used (i.e., w = 1 for all models).

5.1.3 Results

A. EINECS prediction

Predictions for twelve pathways perturbations for 68779 compounds resulted in 824605 predictions (743 predictions failed to compute due to descriptors calculation errors; only 10 compounds failed for all endpoints predictions). All predictions also included the associated applicability domain estimates. The entire prediction set was deposited in an open GitHub repository[317] and is made available for the scientific community and regulators to use.

Figure 37 shows the correlation between consensus standard deviation (CONSENSUS-STD) and the estimated predicted accuracy for twelve pathway endpoints. The applicability of the models to the EINECS compounds differed between targets. As shown in Table 22, The androgen receptor MDA-kb2 AR-luc cell line (nr-ar-lbd) was predicted with at least 85% accuracy for 92% of the compounds. On the other hand, The Nuclear factor (erythroid-derived 2)-like 2/antioxidant responsive element was only predicted for 35% of the compounds with such high accuracy. The histograms in Figure 38 show the distribution of the estimated prediction accuracy for the twelve pathway endpoints among EINECS compounds.

A tradeoff always exists between the coverage of compound set and the minimum accuracy considered. As the purpose of this study is to draw attention to the compounds that are most likely to be harmful and therefore may receive a priority for testing, a high accuracy cutoff of 85% was used. Under such cutoff, the percentage of compounds predicted to disrupt different molecular pathways ranged from 4.6% to 12.6% of the entire EINECS compounds as shown in Table 22. The heatmap in Figure 39 shows the Predicted chemical/pathway perturbation matrix of the EINECS compounds for all twelve pathway endpoints with high accuracy (>85%). The active compounds are shown in red while inactive compounds are shown in green. Compounds with prediction accuracy <0.85 are omitted (Grey).

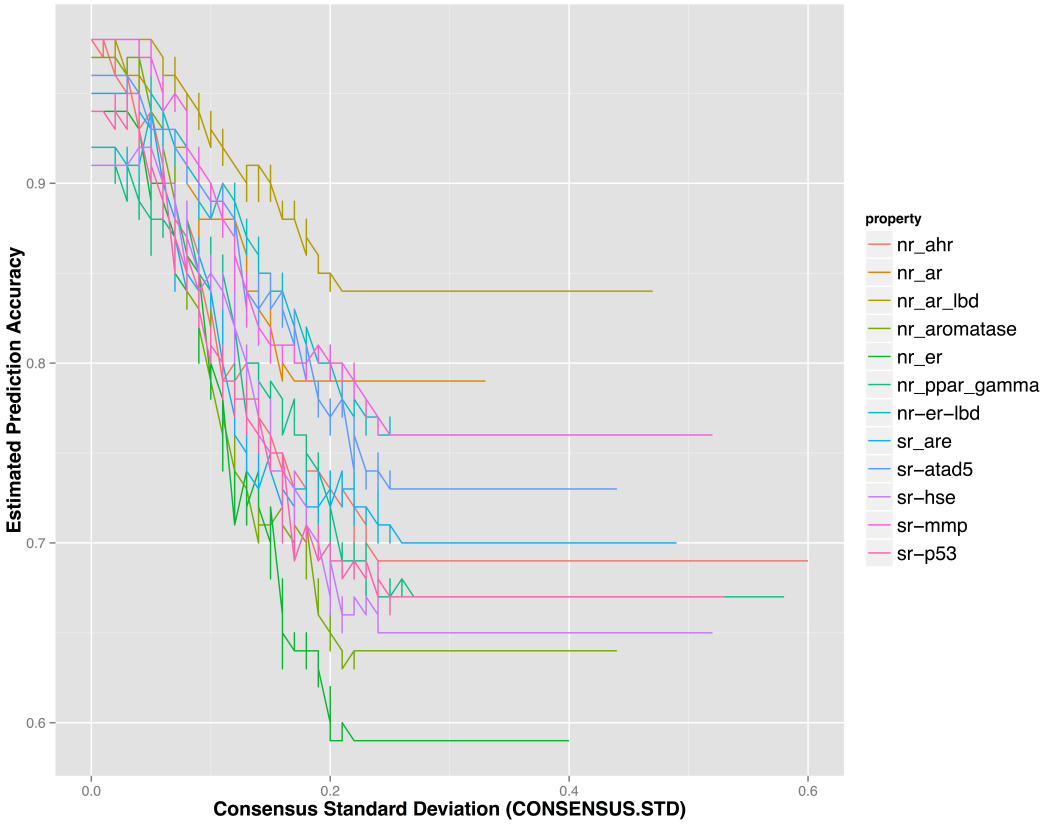


Figure 37. The correlation between consensus standard deviation (CONSENSUS-STD) and the estimated predicted accuracy for twelve pathway endpoints.

Table 22. Percentage of EINECS compounds with high prediction accuracy (>85%) and the percentage of active compounds (i.e., disrupting the molecular pathways) for twelve endpoints.

Pathway endpoints	Percentage of accurate predictions (>=85% estimated accuracy)	Percentage of active compounds among accurate predictions (and among total)
nr-ahr	66.6%	17.1% (11.4%)
nr-ar	77.3%	7.2% (5.5%)
nr-ar-lbd	92.0%	6.5% (5.9%)
nr-aromatase	39.7%	11.7% (4.6%)
nr-er	39.4%	12.6% (5.0%)
nr-ppar-gamma	41.8%	16.9% (7.1%)
nr-er-lbd	73.1%	12.8% (9.4%)
sr-are	34.5%	24.0% (8.3%)
sr-atad5	56.8%	12.2% (6.9%)
sr-hse	52.3%	18.0% (9.4%)
sr-mmp	61.0%	20.7% (12.6%)
sr-p53	47.5%	13.0% (6.2%)

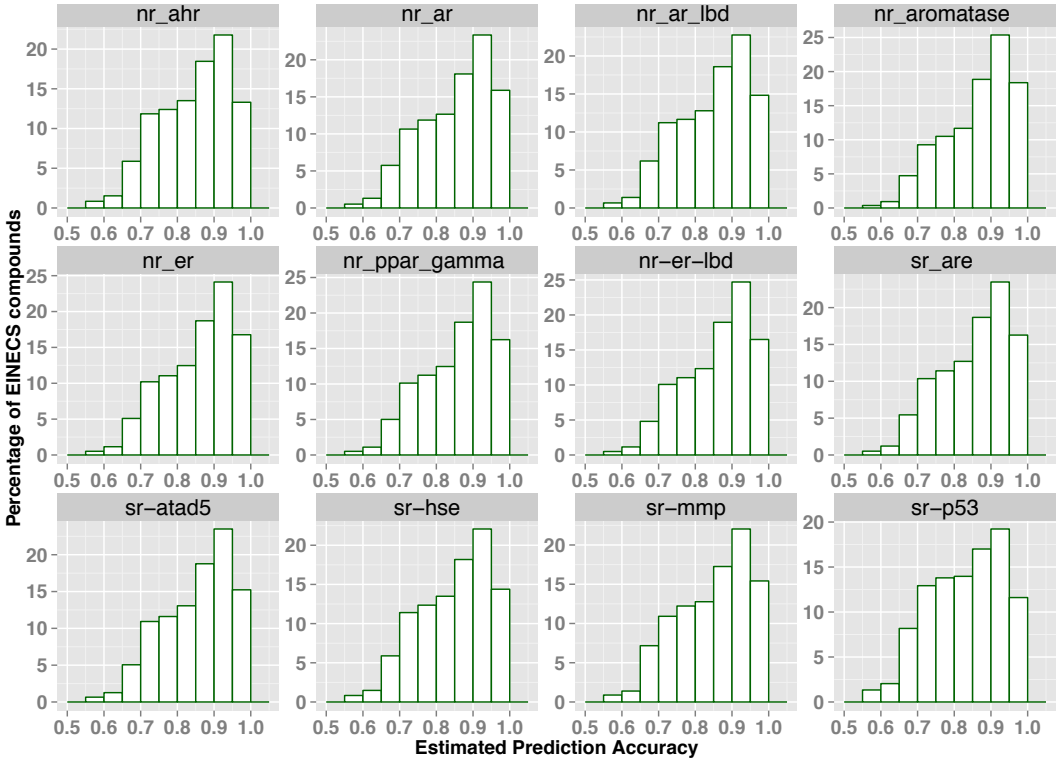


Figure 38. Distribution of the estimated prediction accuracy for twelve pathway endpoints among EINECS compounds

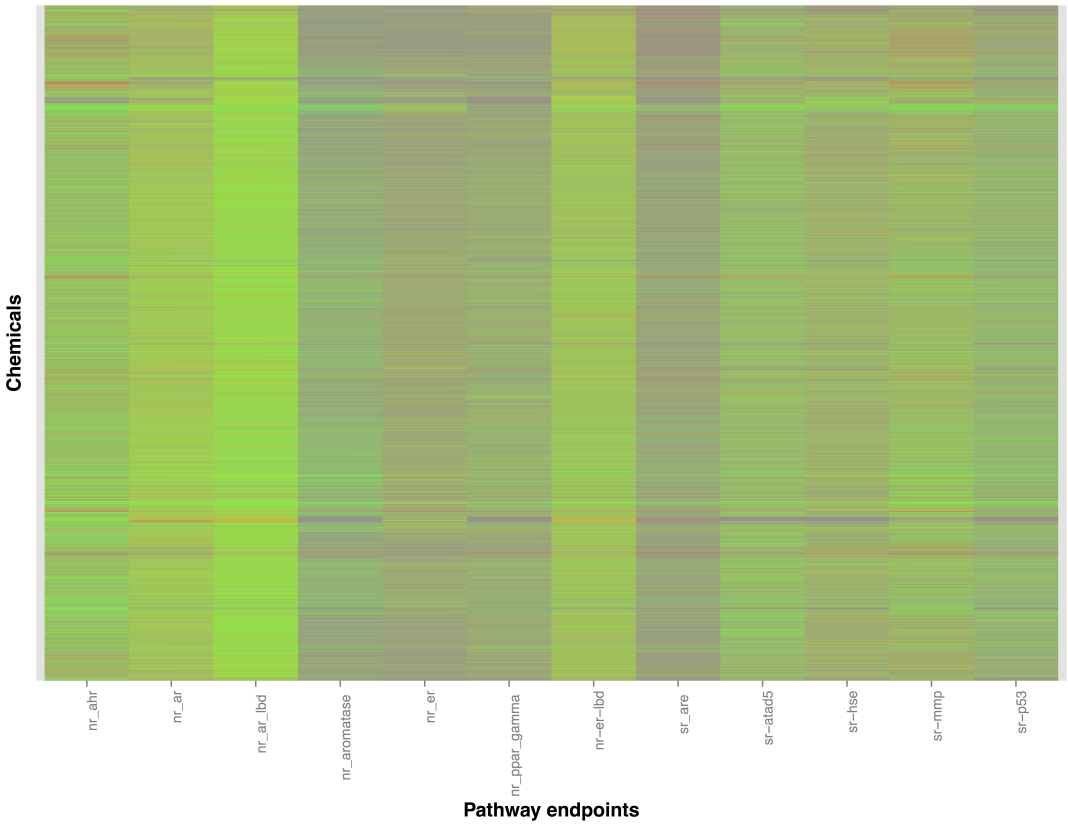


Figure 39. Predicted chemical/pathway perturbation matrix of EINECS compounds for 12 pathway endpoints with high accuracy (>85%). (Red: Active perturbation, Green: no perturbation, Grey: estimated prediction accuracy < 0.85)

B. Toxicity-testing Prioritization Score (ToPS)

The toxicity-testing priority score (ToPS) was calculated for all EINECS compounds. The scoring function takes into account the estimated accuracy based on the distance to model applicability domain through the AD_m term. Therefore, all predictions were considered without cutoff. The ToPS scores for all compounds is deposited in an open GitHub repository[317] for public scrutiny.

Most compounds showed ToPS less than 5. Figure 40 shows a histogram for the distribution ToPS scores among EINECS compounds. Examining the compounds with highest ToPS scores show that they are highly conjugated fused-ring systems suggesting potential reactivity. Figure 42 shows the six compounds with highest ToPS scores.

5.1.4 Summary

The QSAR models developed in earlier studies were used to predict EINECS compounds. These chemicals are used in the EU and were left with insufficient data regarding their toxicity risk. The predictions show, with high confidence, that a certain percentage of chemicals (between 4.6% and 12.6% depending on the target) are likely to disrupt molecular pathways and are worth of further investigations.

The cross-reactivity against multiple nuclear receptors has been reported earlier in multiple studies. Crosstalk between estrogen- and aryl hydrocarbon receptors leads to inhibition of estrogenic signaling in experimental animals as well as *in vitro*. Studies suggest that ARNT is a coactivator of ER[369]. Crosstalk has also been reported between AhR, PXR, and the constitutive androstane receptor (CAR)[370].

Furthermore, a toxicity-testing prioritization score (ToPS) is suggested that can give a holistic overview of the compound's molecular pathways perturbation and assess its overall risk profile. Rather than judging compounds priority arbitrarily, ToPS offers a systematic rationale and a prioritization scheme. The score is calculated as a factor of the predicted toxicity, applicability domain (as a distance to model) and a weight for the different endpoints.

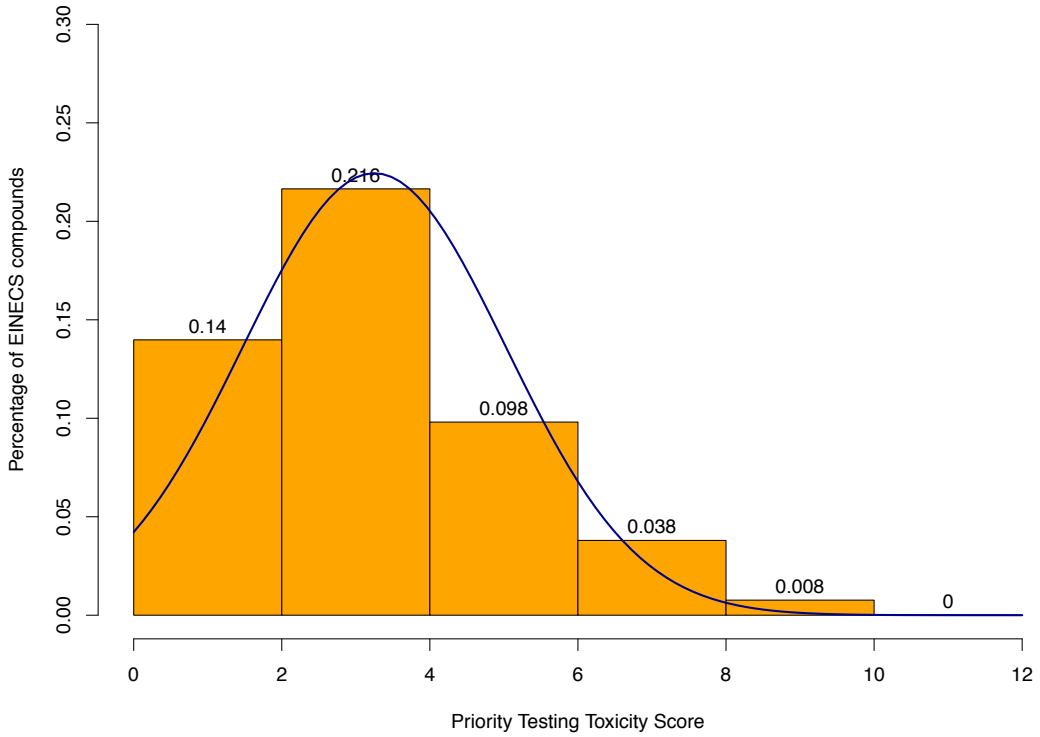


Figure 40. Histogram of the distribution of ToPS scores among EINECS compounds

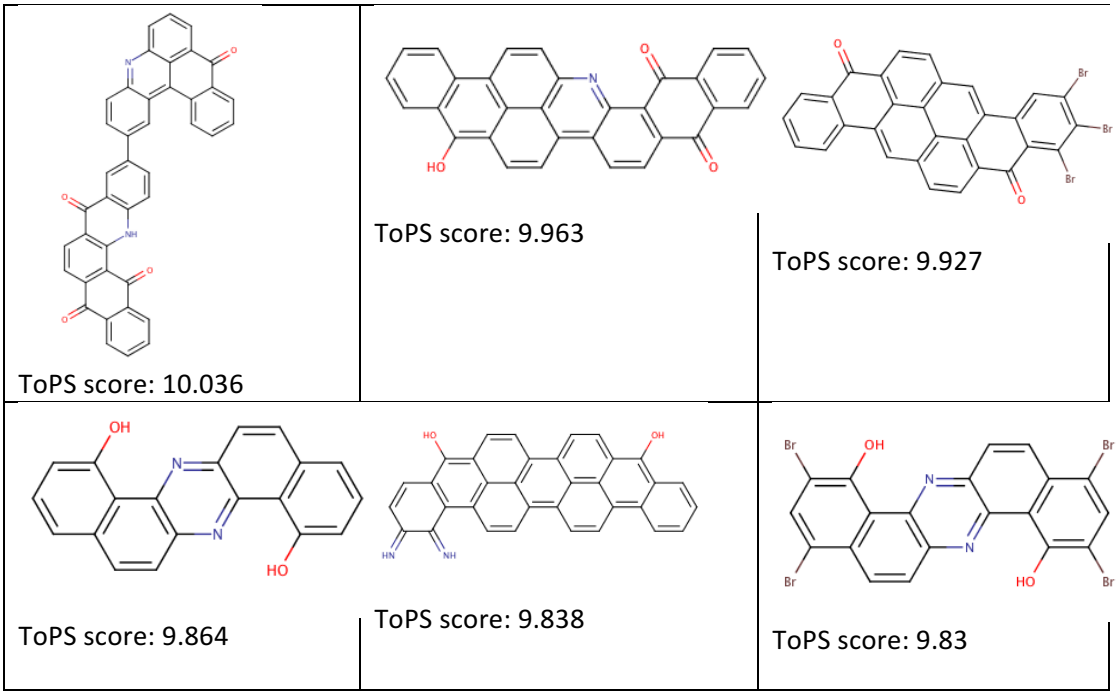


Figure 41. EINECS compounds with highest ToPS scores suggesting high disturbance of molecular pathways

5.2 Prediction of potential toxicity of halogenated carbazoles

5.2.1 Introduction

Halogenated carbazoles belong to the heterocyclic aromatic hydrocarbons group. Limited

information is available on the properties of halogenated carbazoles. They have recently been reported in lakes[371], sediments[372], soil[373,374] and sea[372,375] in Europe and the United States. Carbazole as well as chloro- and bromocarbazoles have been previously reported to exhibit dioxin-like toxicity by EROD induction[376] and have been recently regarded by the European Commission as potentially persistent, bioaccumulative, and toxic (PBT) substances.

Due to the emergence of, previously unknown, halogenated carbazoles in soil and water samples[377], many studies are looking into their sources[378] and potential effects[376]. Although the compounds are not directly synthesized, their detection in the environment raises concerns. Besides their toxicological effects, studies have shown their persistence in soil[376].

REACH regulations transferred the risk assessment responsibility to the producers and importers of chemicals. For compounds that are not actively produced or imported, ambiguity stands on the industrial responsibility to assess the potential toxicity of such compounds giving the high cost for conducting toxicity studies. QSAR can play a role in prioritizing the testing and guiding regulators on potential toxicity pathways that could be affected by these compounds.

The aim of this application is to investigate the ability of carbazole derivatives to activate the AHR and whether such activation can be detected using QSAR modeling.

5.2.2 Methods

A. Analysis of AHR activation by carbazoles in HTS in vitro assays

Two HTS *in vitro* AHR activation assay datasets were searched for compounds that show a carbazole substructure scaffold. The HTS assays came from the PubChem bioassay database (AID: 2796)[363] (described in 4.5.1 Data acquisition and curation above) and the Tox21 assay dataset (described in C Aryl hydrocarbon receptor (AHR) (AID: 743122[344]) above). The carbazole derivatives detected in these datasets were examined for being AHR activators and the results were statistically compared to the average probability of the presence of AHR activators in the respective dataset using hypergeometric distribution.

Furthermore, the first dataset (AID: 2796) was large enough to allow the examination of structural features that may contribute to the activation of the AHR receptor. Therefore, active and inactive carbazolyl-bearing compounds from the first dataset were compared using the "Set Compare" utility and the ToxAlerts structural alerts[224]. ToxAlerts is a collection of 2310 SMARTS template patterns collected from literature. These structural patterns act as alerts for endpoints related to different adverse outcome such as skin sensitization, carcinogenicity, metabolic activation, mutagenicity, and compounds that may form reactive metabolites with potential adverse reactions. The alerts are available online as part of the OCHEM platform. The "Set Compare" utility compares two sets of chemicals (in this case, AHR active and inactive compounds) for the presence of certain features (in this case the ToxAlerts SMARTS patterns), counts the number of occurrences of each pattern in both sets and quantifies the p-value for the probability of such occurrences by chance according to a geometric distribution.

B. Carbazoles as drugs

The DrugBank database[379] version 4.3 was searched for the presence of carbazole derivatives. Carbazolyl drugs detected in the database were examined for AHR activation using both HTS *in vitro* AHR dataset (AIDs: 2796 and 743122340) mentioned above to determine whether these drugs are AHR activators. DrugBank hosts comprehensive drug information including drug protein targets. It contains more than 8206 drug entries. These are 207 FDA-approved biotech products (proteins and peptides), 1991 FDA approved small molecule drugs, and more than 6000 experimental drugs. The database allows full chemical structure search (including substructure queries) as well as text and sequence searches.

C. Prediction of AHR activity for halogenated carbazoles

Earlier studies suggested that bromo- and iodocarbazoles could be more persistent, bioaccumulative, and toxic than the parent carbazoles[375]. Therefore, the toxicity of mono-, di-, tri- and tetra- halogenated carbazoles was predicted using QSAR. Figure 42 shows the markush representation of the carbazoles. Overall, 250 unique chemical structures resulted from enumerating the markush representation using Marvin Sketch. Structures were uploaded to OCHEM in SDF format. The impact of these chemicals on twelve molecular targets was predicted using the QSAR models developed in 4.3 Tox21 project and listed in Table 15. Finally, predictions were downloaded using KNIME together with the prediction's distance-to-model as a measure for applicability domain.

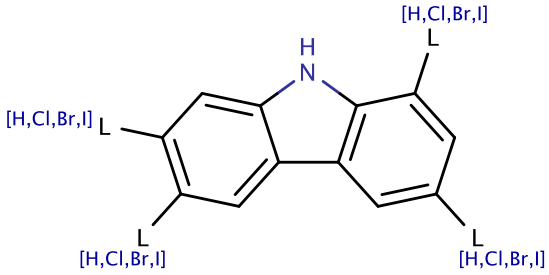


Figure 42. Markush representation of halogenated carbazoles investigated using QSAR models for 12 toxicity-related targets

5.2.3 Results

A. Analysis of carbazoles activation in HTS in vitro assays

Tox21 dataset comprised 6988 unique compounds, of which 817 compounds were AHR activators. The dataset included 6 carbazole derivatives of which 5 were AHR activators (83.3%).

The PubChem dataset (AID: 2796) tested 324744 unique compounds. Among them, 291 held the carbazolyl scaffold. Two compounds were reported in both active and inactive datasets and were excluded from the analysis. Among the remaining 289 compounds, 46 of them (15.91%) being activators for the AHR. This represents more than 6-fold enrichment from the average presence AHR activators within the set (7987 substances ≈2.46%). This represents a highly significant increase (p-value: 3×10^{-25}) in the AHR activation for carbazole derivatives.

Analyzing the PubChem dataset by comparing both active (46 carbazole derivatives) and inactive (243 carbazole derivatives) sets using the ToxAlerts SMARTS templates suggested some structural patterns to be highly correlated to the AHR activation of carbazoles. These structural patterns are listed in Table 23. The shown scaffolds suggest that aromatic amines are highly correlated to the activation of AHR while alcohols and phenols are highly represented among AHR inactive compounds.

B. Carbazoles as drugs

Searching the DrugBank database for carbazole derivatives resulted in the identification of two FDA-approved drugs (carvedilol and carprofen) shown in Figure 43 as well as 22 other investigational or experimental compounds.

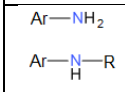
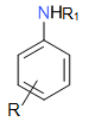
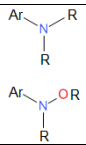
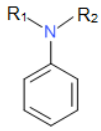
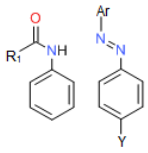
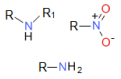
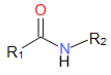
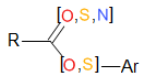
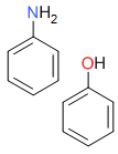
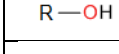
Carvedilol is both a competitive beta-adrenoceptor antagonist[380–382] and an arterial vasodilator[383]. At higher concentrations, carvedilol is also a calcium channel antagonist[380,381,383]. The vasodilatory actions result primarily from alpha-adrenoceptor blockade[383] although in certain vascular beds, calcium channel blockade may also contribute to vasodilation[381]. In addition to these actions on the heart and vasculature, carvedilol has also been shown to possess significant antioxidant and antiproliferative actions.

Carprofen is a non-steroidal anti-inflammatory drug (NSAID)[384] that is used in veterinary medicine to treat geriatric dogs with arthritic symptoms. The drug was initially used as a human drug (1985-1995). The compound was highly tolerated with intestinal ulcers as the only side effect that was reported after high dose exposure in animals[384]. The drug was pulled on voluntary basis by Pfizer for commercial reasons[385]. Other mild adverse effects were similar to those reported by aspirin and other NSAIDs including gastrointestinal pain and nausea.

Both compounds were tested in both HTS AHR activation *in vitro* assays (AIDs: 2796 and 743122340). Carprofen was reported as inactive in both assays while carvedilol result was inconclusive. It was reported as active in the Tox21 dataset only (AID: 743122340).

Among the 22 experimental/investigational compounds, experimental AHR activation data was available for 2 compounds only (Staurosporine and (S)-Wiskostatin) shown in Figure 43. Both compounds were reported as inactive.

Table 23 Structural patterns with high significance to the AHR activation for the carbazoles within the *in vitro* assay screening dataset (AID: 2796). P-values are calculated through a hypergeometric distribution

Structural pattern	Pattern name	Occurances in AHR active compounds	Occurances in AHR inactive compounds	Enrichment factor	p-value
	Aromatic amines	29 (63.0%)	56 (23.0%)	2.7	2.32E-7
	Aromatic primary and secondary amines	29 (63.0%)	56 (23.0%)	2.7	2.32E-7
	Aromatic amines	31 (67.4%)	68 (27.9%)	2.4	-10^{-7}
	Anilines, anilides	29 (63.0%)	62 (25.4%)	2.5	-10^{-6}
	Aromatic amines precursors	25 (54.3%)	50 (20.5%)	2.7	-10^{-6}
	Aromatic N-Groups alcohols or phenols	27 (58.7%)	60 (24.6%)	2.4	-10^{-5}
	Carboxylic acid secondary amides	26 (56.5%)	57 (23.4%)	2.4	-10^{-5}
	Esters of aromatic alcohols and their thio and aza analogues	25 (54.3%)	53 (21.7%)	2.5	-10^{-5}
	Simple anilines and phenols	29 (63.0%)	71 (29.1%)	2.2	-10^{-5}
Patterns more frequent in inactive compounds					
	Alcohols	3 (6.5%)	83 (34.0%)	5.2	-10^{-5}
	Hydroxyl compounds: alcohols or phenols	4 (8.7%)	99 (40.6%)	4.7	-10^{-6}

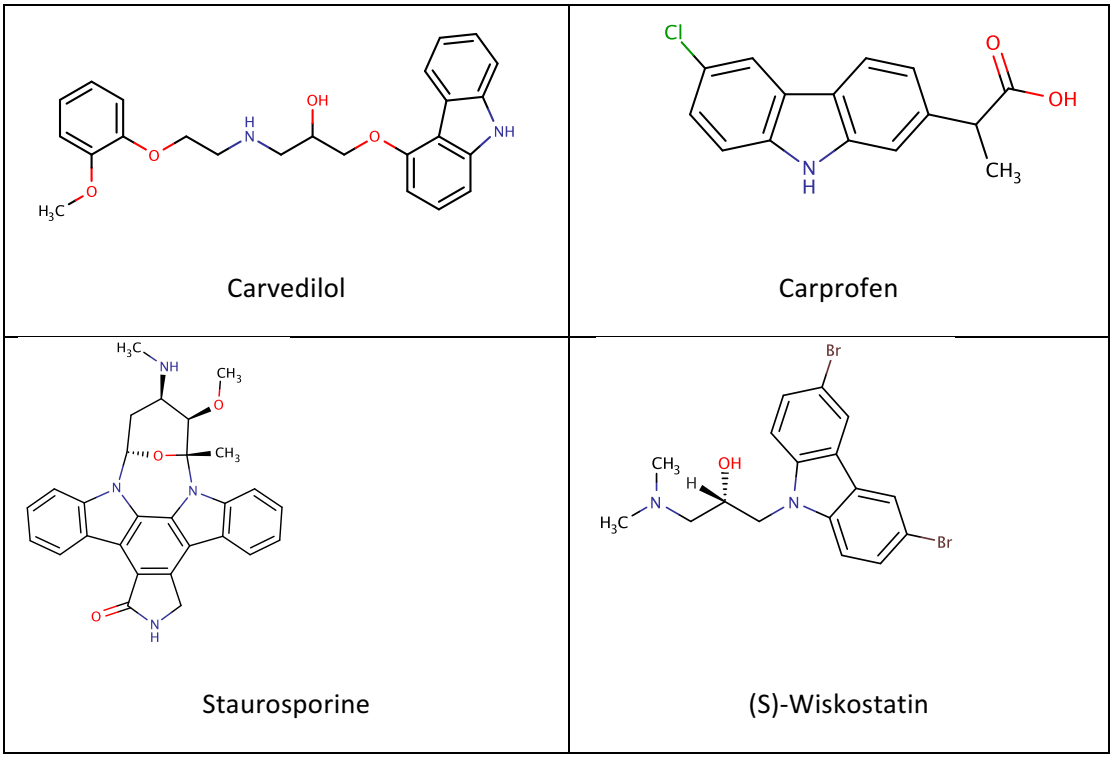


Figure 43. The top row shows the chemical structures of the only 2 FDA-approved drugs that show a carbazole substructure according to the DrugBank database; Carvedilol (top-left) and Carprofen (top-right). The bottom row shows the chemical structures of 2 experimental drugs for which AHR activation HTS assay data were available; Staurosporine (bottom-left) and (S)-Wiskostatin (bottom-right)

C. Prediction of AHR activity for halogenated carbazoles

Applying all twelve models (Table 15) to 250 unique chemicals results in 3000 predictions with no calculation errors. Most predictions are estimated to have high prediction accuracy as shown in Figure 44.

Halogenated carbazoles are predicted to cause perturbation to all investigated pathway endpoints, except the androgen receptor MDA-kb2 AR-luc cell line (nr-ar) as shown in Table 24. The heatmap in Figure 45 shows the predicted chemical/pathway perturbation matrix of the halogenated carbazoles for all twelve pathway endpoints with high accuracy (>85%).

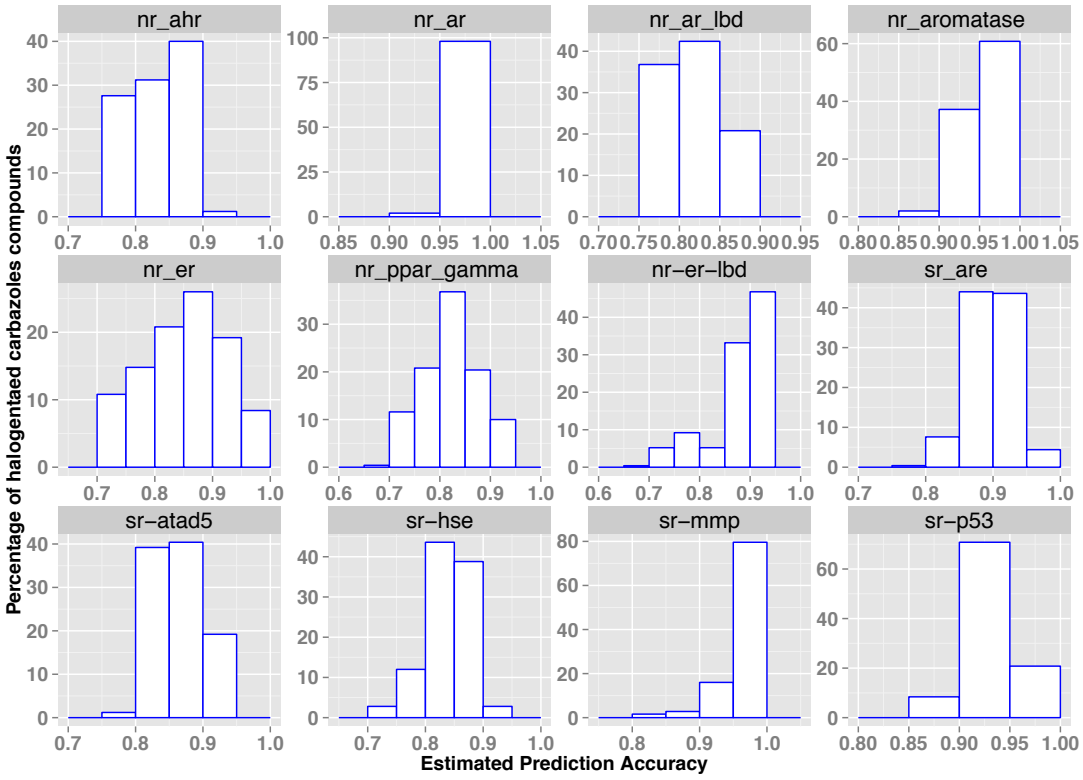


Figure 44. Distribution of the estimated prediction accuracy for twelve pathway endpoints among halogenated carbazoles compounds

Table 24. Percentage of halogenated carbazole compounds with high prediction accuracy (>85%) and the percentage of active compounds (i.e., disrupting the molecular pathways) for twelve endpoints.

Pathway endpoints	Percentage of accurate predictions (>=85% estimated accuracy)	Percentage of active compounds among accurate predictions (and among total)
nr-ahr	100%	100% (100%)
nr-ar	21%	0% (0%)
nr-ar-lbd	100%	98% (98%)
nr-aromatase	30%	100% (30%)
nr-er	80%	100% (80%)
nr-ppar-gamma	41%	99% (41%)
nr-er-lbd	92%	100% (92%)
sr-are	54%	100% (54%)
sr-atad5	60%	100% (60%)
sr-hse	42%	100% (42%)
sr-mmp	98%	100% (98%)
sr-p53	100%	100% (100%)

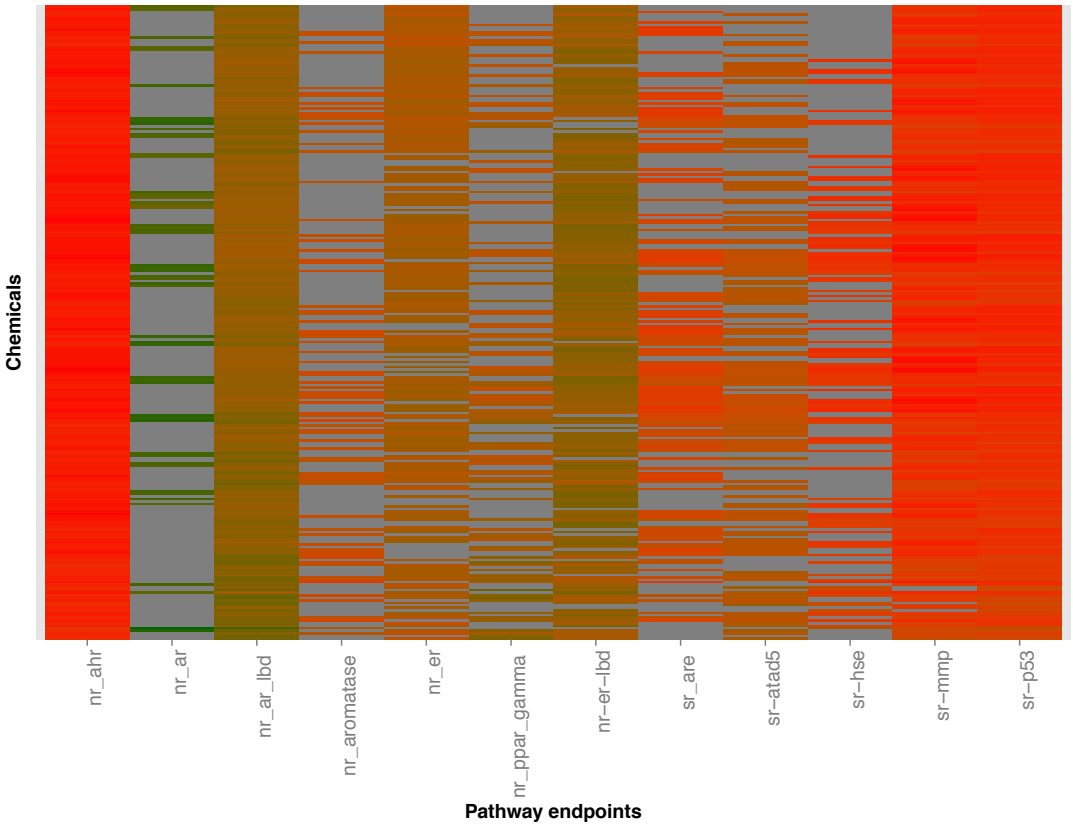


Figure 45. Predicted chemical/pathway perturbation matrix of halogenated carbazoles compounds for twelve pathway endpoints with high accuracy (>85%). Color gradient indicates the prediction distance from class limits (green for inactive compounds and red for active perturbation)

5.2.4 Discussion

The interesting structural features and auspicious pharmacological behaviors of carbazoles resulted in an immense growth in the carbazole chemistry. Alkaloid derivatives of carbazoles are famous for many pharmacological activities, such as antifungal, anti-bacterial, anti-cancer and anti-HIV activities. Some carbazole derivatives (e.g., poly(vinylcarbazole)) are being used in industrial applications as optoelectronic materials[386]. The presence of halogen atoms allows these compounds to be used as substrates in coupling reactions such as the Suzuki-Miyaura cross-coupling.

Investigating carbazole derivatives among HTS *in vitro* assays for AHR activation revealed that the presence of carbazolyl moiety highly correlates with AHR activation (p-value: 10^{-25}) and that such moiety provides high enrichment factor (> 6-fold). Such correlation adds to the weight of evidence that link carbazoles to dioxin-like side effects. It may suggest that AHR activation is a key event in this adverse outcome pathway. This is in accordance with earlier studies[376,387] which suggest that carbazoles act through the AHR activation. Further investigation of relative effect potencies of different carbazoles is needed to shed light on potential toxicity from such activation.

Further analysis of carbazole substitutions to determine which substituents are more likely to lead to AHR activation showed that aromatic amines were highly likely to activate AHR (p-

value: 10^{-5} to 10^{-7}) while alcohols and phenols were more likely to be associated with AHR inactive compounds (p-value: 10^{-5} to 10^{-6}).

Searching for FDA-approved carbazolyl drugs showed only 2 candidates (carvedilol and carprofen) in the DrugBank. The results of these drugs in HTS AHR activation *in vitro* assays were either inactive (carprofen) or contradictory (carvedilol). On the other hand, 22 experimental compounds were reported in the DrugBank with carbazole substructure. Experimental data available for 2 compounds showed them being AHR inactive.

The prediction of pathway perturbations for 250 halogenated carbazoles against 12 assays (representing ten biochemical pathways) show high confidence in prediction for most targets. All halogenated carbazoles show activity against all pathways except the androgen receptor MDA-kb2 AR-luc cell line (nr-ar). However, it is important to notice that this particular target had the lowest applicability domain coverage among all targets (only 21% of the compounds had an estimated accuracy > 85%).

Finally, activation of AHR in itself does not necessarily mandate toxicity. AHR has multiple reported endogenous activators. It was proposed that the high level persistent stimulation of AHR by ligands is the cause of toxic effects[44]. Future studies are needed to quantify the duration and magnitude of receptor activation revealing more about the pharmacodynamics of such AHR activators. However, this study shows that halogenated carbazoles represent a class of persistent organic pollutants exhibiting dioxin-like toxicity.

6 Summarizing discussion

6.1 Outcome of the studies and conclusions

The aim of this work was to investigate the ability of QSAR modeling to predict potential systems toxicity of chemicals. The overall work was based on the 5 OECD principles for QSAR models construction. It included available in-vitro databases, which were provided through the - also REACH-related - framework of development and generation of HTS (*in vitro*) profiling methods, and included development and optimization of predictive computational models. For the purpose of modeling and prediction, multiple approaches were used including the direct correlation of chemical structures to *in vivo* animal toxicity, the combination of *in vitro* HTS and *in silico* descriptors to predict *in vivo* outcomes as well as the prediction of specific pathways perturbations by correlating *in silico* descriptors to results from HTS of such targets.

Generally, the direct prediction of *in vivo* animal toxicity using *in silico* descriptors for complex end points yielded limited success. Prediction was only feasible for restricted compound libraries with the same mode of action (e.g., organophosphorus compounds' toxicity inhibition of acetylcholine esterase). Some *in vivo* endpoints, with a promising predictive balanced accuracy exceeding 70% were identified in the course of the studies. These include multiple rat maternal toxicity endpoints and chronic apoptosis and necrosis in mice.

Data, derived from HTS *in vitro* profiling of chemicals, were combined with *in silico* descriptors to build "biological descriptors". This approach showed a significant improvement in the predictive ability of QSAR models for some endpoints (p-values <0.05) compared to the use of *in silico* descriptors alone (such as rat fetal pathology). Furthermore, the mechanistic classification and regrouping of the HTS *in vitro* assay responses in the form of pathway perturbations significantly improved (with p <0.05) the predictivity for some toxicity endpoints. These includes chronic rat liver neoplastic lesions and multigenerational rat viability among others.

Overall, the prediction of final *in vivo* toxicity remains challenging. This was confirmed by the low median performance of QSAR models predicting the final *in vivo* toxicity by analysis of ToxCast Phase I compounds. Therefore, as to be expected from the complexity of a living organism as opposed to *in vitro* isolated targets or *in vitro* biological systems, it remains difficult to directly replace animal toxicity testing using predictive QSAR models, with a possible exception of acetylcholinesterase inhibition.

In vitro HTS is also useful in detecting molecular pathways that are most correlated to *in vivo* toxicity outcomes (by using the "Set Compare" utility). This indicates that *in vitro* assays could assist in understanding the underlying mechanism of toxicity and the essential biochemical pathways involved. Furthermore, fragment-analysis techniques used to support the investigation of potential modes of action were also promising.

As opposed to handling or generating data from *in vivo* systems, prediction of *in vitro* assays outcomes using *in silico* descriptors showed high success. This was confirmed by the high balanced accuracy for predicting ToxCast Phase I assay results. The concept represents a different approach towards toxicity prediction where *in silico* descriptors can be used to

model *in vitro* assay outcomes that are known to be related to specific *in vivo* effects. The Tox21 project explores this possibility by profiling a large number of chemicals using *in vitro* assays as an investigational and exploratory tool.

Using QSAR for modeling the outcome of Tox21 *in vitro* assays (representing different molecular pathways) showed promising success with balanced accuracies reaching up to more than 80% for several endpoints, such as aryl hydrocarbon receptor activation (86%), mitochondrial membrane disruption (88%) and androgen receptor activation (82%). The relatively high balanced accuracies among models confirmed the possibility of modeling HTS results from *in vitro* assays using *in silico* descriptors as reported in earlier studies[187].

Bagging validation provided a good indication for the models' predictive ability on external validation sets. *Stratified bagging* addressed the unbalanced nature of the training set and reduced bias towards the majority class. The stratified bagging contributed models, which were optimized towards the balanced accuracy. Models developed in the Tox21 study calculated the best balanced accuracy across all twelve analyzed targets. Furthermore, the used strategy allowed to calculate the highest AUROC scores for two targets. It is also important to realize that, due to the model prediction variances, selecting a model with the highest validated accuracy does not guarantee the highest predictive ability for an evaluation set.

Consensus modeling improved the predictive ability of models as signified by both validation and evaluation set accuracies. To a high degree this result was achieved thanks to the diversity of descriptor packages, which captured different aspects of the molecular structures. Use of different descriptors also compensated for failure of some descriptors to represent certain structures, thus covering the entire training set. This methodology achieved the highest balanced accuracy for all 12 targets of the Tox21 Data Challenge organized by the NIH. The same approach was used to build QSAR models for the activation of pregnane X receptor as well as QSAR models on an extended dataset for the aryl hydrocarbon receptor activation. Both studies showed high prediction accuracy.

Finally, two specific applications were computed and discussed. These applications put the developed *in silico* to *in vitro* QSAR studies on the twelve molecular pathway endpoints of the Tox21 project in practical solicitation for environmental risk assessment.

(1) The first application screens the large dataset of EINECS compounds for potential pathway perturbations. The predictions show, with high confidence, that a certain percentage of chemicals (between 4.6% and 12.6% depending on the target) are likely to disrupt molecular pathways. Furthermore, in conclusion, a point-based system was suggested: *toxicity-testing priority score (ToPS).* This score provides a universal overview of a compound's molecular pathways perturbation and assesses its overall risk profile. ToPS offers a systematic rationale for a compound-prioritization scheme. The score represents a factor of the predicted biochemical pathways perturbation, applicability domain and allows weighing different targets according to the investigated application or potential exposure scenario in the environment.

(2) The second application investigates a set of halogenated carbazole compounds emerging in the European and US ecology without being actively produced or imported. All halogenated carbazoles show, with high prediction accuracy, an activity against all pathways except the androgen receptor MDA-kb2 AR-luc cell line (nr-ar). This particular target had the lowest applicability domain coverage (21%) - among all targets - for high accuracy predictions (>85%). Analyzing HTS data showed that the presence of carbazolyl moiety highly correlates with Aryl Hydrocarbon Receptor (AHR) activation (p-value: 3×10^{-25}). The carbazolyl moiety provides high enrichment factor (> 6-fold) for AHR activation. Certain carbazolyl substitutions (such as aromatic amines) are more likely to lead to AHR activation (p-value: 10^{-5} to 10^{-7}) while alcohols and phenols were more likely to be associated with AHR inactive compounds (p-value: 10^{-5} to 10^{-6}).

QSAR models developed in this thesis were recognized by winning multiple awards in challenges organized by the National Institute of Health (NIH) as well as the environmental protection agency (EPA). The outcomes of the dissertation are available to regulators and the scientific community. The public platform iPrior was deployed and is hosting data from ToxCast, Tox21, and e1K projects. It is open for researchers to apply the developed models on new compounds, upload more data, or contribute their descriptor packages. Moreover, the developed models based on the Tox21 study are made publicly available at http://amaziz.com/article/tox21, thus allowing their use for prospective and retrospective analyses. Finally, the results of different studies and applications are made available in an open GitHub repository. It is hypothesized that those freely accessible models may become accepted by the regulators and the scientific community and therefore play a significant role in predicting *in vivo* toxicity and reduce animal testing.

6.2 Outlook and recommendations

The ultimate goal of computational toxicology would be to achieve a precise prediction of human, animal and environmental risk of chemicals and minimize the need for conducting animal studies. However, a number of open questions and underlying assumptions must always be kept in mind:

1. The *in-vitro* to *in-vivo* extrapolation of toxic effects remains a challenge that still needs to be further investigated and validated. The extrapolation suffers from multiple limitations. The contribution of bioavailability (i.e., entrance or uptake of a chemical into the biological system) in activation or elimination of toxicity-relevant chemicals remains noteworthy for future investigation. *In vitro studies with* cell line settings do not account for metabolic first-pass effect in the gastrointestinal tract (in cases of oral ingestion this may be relevant).
2. *in vitro* cell-lines (being frequently based on carcinoma cells) express different patterns of metabolizing enzymes, not comparable with healthy human cells. *In vitro* assays may lack the bioactivation pathways that are exerted *in vivo*. The assumption of similarity of gene expression between carcinoma cells and normal human cells should always be questioned.

In conclusion, three major concerns with respect to this first pragmatic approach and successful model development are evident: (1) The integration of QSAR bioavailability models

together with toxicokinetic simulations may advance the presented work forward yielding better extensibility towards *in vivo* chemical toxicity prediction. The additional "biological descriptors" (such as biopharmaceutical data) might aid and improve the overall correlation. (2) The *in vitro* assays examined might not be sufficient for capturing biochemical events on the molecular level or depict the pathways responsible for toxicity. (3) QSAR modeling, as a statistical approach, necessitates a significant amount of data. Relatively low numbers of chemicals as training instances restrict the modeling process as shown with the ToxCast study. This constraint would gradually diminish as more data becomes available in future stages of ToxCast and other programs. The applicability domain and predictive power of models is very likely to increase.

6.3 Final remarks

The continuous development and extension of the presented QSAR models is recommended and may be regarded as "natural" development following the growth in knowledge and experience. As machine-learning algorithms get more capable, computational power becomes cheaper and descriptors allow higher resolution representations about chemicals, the QSAR models could be further improved towards a continuously increasing prediction ability. Additionally, more biological targets should be considered widening the spectrum and increasing the probability to identify more toxicologically relevant structural elements in chemicals. The methodology for QSAR models building was performed and presented in a way, which allows its extension to other targets of interest.

In the core of the REACH vision is the belief that QSAR studies should extend from description and providing reliable prediction into guiding the chemical design process. To deduce useful guidance from QSAR models, they should offer more insights into the reasons behind the given predictions, which requires detailed concept- and data analysis, particularly with respect to potential reasons for predicted outcomes. The mechanistic interpretation is, unfortunately, getting harder with the advance of complex statistical approaches and non-linear models. The use of prediction-driven Matched Molecular Pairs (MMP) analysis[192] may be a next step. Such an approach may analyze the space of available compounds for minor chemical modifications that could lead to inverted outcome (e.g., loss of toxicity). This, combined with the applicability domain estimation, would enlighten chemical manufacturers and pharmaceutical companies into the reasons behind developed toxicities and allow the design of safer chemicals while reducing research time and cost.

The goal of QSAR models in predictive toxicology, ordinarily, is to forecast an adverse outcome beyond protein binding or nuclear receptor activation. In this sense, QSAR prediction of molecular pathways' perturbation is, in itself, an attempt to mechanistically understand toxicological risks. In the context of adverse outcome pathways (AOP), such perturbations are considered as molecular initiating events (MIE) or key events (KE) leading to certain adverse outcome. Such KEs are connected through key event relationships (KERs) to form the network of multiple AOPs which form the functional prediction component for real-life circumstances. The investigated molecular pathways have been suggested to play a relevant and significant role in many adverse outcomes.

Finally, it is assumed that the presented work will assist in improving the regulation of chemicals by better deployment of highly potent testing capacities. The more accurate and sophisticated and accessible computational approaches get, the higher the potential for reducing environmental toxicity risks, saving animals, and speeding the – cost-intensive – discovery processes for new developments.

List of abbreviations

2D	Two-Dimensional
3D	Three-Dimensional
ACC	Total Accuracy
ACToR	Aggregated Computational Toxicology Resource
AD	Applicability Domain
ADME	Absorption, Distribution, Metabolism, Excretion
ADMET	Absorption, Distribution, Metabolism, Elimination and Toxicity
AHR	Aryl hydrocarbon receptor
AID	Assay Identification number
AIS	Androgen Insensitivity Syndrome
ANN	Artificial Neural Networks
AR	Androgen Receptor
ar-lbd	androgen receptor MDA-kb2 AR-luc cell line
ARE	Antioxidant Responsive Element
AREs	Androgen Response Elements
ArKO	Aromatase Knockout Mice
ARNT	Aryl Hydrocarbon Receptor Nuclear Translocator
AROM+	Mice Overexpressing Human Aromatase
ASNN	Associative Neural Networks
ATAD5	Human ATAD5 protein
AUC	Area Under the Curve
AUROC	Area Under the ROC Curve
BACC	Balanced Accuracy
Bagging	Bootstrap Aggregation
BAT	Brown Adipose Tissue
CASRN	Chemical Abstract Registration Number
CDK	Chemistry Development Kit
CEPs	Conformational Ensemble Profiles
CMR	Carcinogenic, Mutagenic or Toxic to Reproduction
CMR	Carcinogenicity, Mutagenicity or Toxicity to Reproduction
CoMFA	Comparative Molecular Field Analysis
CONS	Consensus Model
CSR	Chemical Safety Report
CV	Cross-Validation
DHT	Dihydrotestosterone
DM	Distance to Model
DRE	Dioxin-Responsive Element
EC	European Commission
ECHA	European Chemicals Agency
EDCs	Endocrine Disrupting Chemicals
EDSP	Endocrine Disruptor Screening Program
EEM	Electronegativity Equalization Method

EINECS	European Inventory of Existing Commercial Chemical Substances
EMA	European Medicines Agency
EPA	Environmental Protection Agency
EPI	Estimation Program Interface
ER	Estrogen Receptor
er-lbd	ER-alpha-UAS-bla GripTiteTM cell line
ERE	Estrogen Response Element
ETC	Electron Transport Chain
EU	European Union
FDA	Food and Drug Administration
FDR	False Discovery Rate
FOR	False Omission Rate
FPR	False Positive Rate
FSMLR	Fast Stagewise Multiple Linear Regression
GCODs	Grid Cell Occupancy Descriptors
GPL	General Public License
GUI	Graphical User Interface
HAHs	Halogenated Aromatic Hydrocarbons
HERG	Human Ether-A-Go-Go Channel
HSEs	Heat Shock Factor Response Elements
HSF	Heat Shock Factor
HSP	Heat Shock Proteins
HSR	Heat Shock Response
HTS	High Throughput Screening
IC50	Half Maximal Inhibitory Concentration
IGC	Growth Inhibition Concentration
InChI	International Chemical Identifier
IPEs	Interaction Pharmacophore Elements
IUPAC	International Union of Pure and Applied Chemistry
kNN	K-Nearest Neighbors
KRR	Kernel Ridge Regression
LBD	Ligand-Binding Domain
LD50	Median Lethal Dose
LGPL	Lesser General Public License
LMO	Leave-Many-Out
LOO	Leave One Out
LQTS	Long QT Syndrome
LVs	Latent Variables
MAE	Mean Absolute Error
MCDM	Multi-criteria Decision Making
MFC	Molecular Fragments Count
MGD	Multi-Gaussian Distribution
MLR	Multivariate Linear Regression

MLRA	Multiple Linear Regression Analysis
MMP	Mitochondrial Membrane Potential
NCATS	The National Center for Advancing Translational Sciences
NER	Non-Error Rate
NIEHS/NTP	The National Institute of Environmental Health Sciences/ National Toxicology Program
NIH	The National Institutes of Health
NIST	National Institute of Standards and Technology
NLS	Nuclear Localization Sequence
NPV	Negative Predictive Value
nr	nuclear receptor
OCHEM	Online Chemical Modeling Environment
OECD	Organization for Economic Co-Operation and Development
OPP	Office of Pesticide Programs
OPPTS	Office of Prevention, Pesticides and Toxic Substances
P53	Tumor protein p53
PAHs	Polycyclic Aromatic Hydrocarbons
PBT	Persistent, Bioaccumulative And Toxic
PCA	Principal Component Analysis
PDF	Probability Density Function
PLS	Partial Least Squares
PPAR-γ	Peroxisome Proliferator-Activated Receptor Gamma
PPV	Positive Predictive Value
PXR	Pregnane X Receptor
QC	Quality Control
QSAR	Quantitative Structure Activity Relationship
QSAR/QSPR	Quantitative Structure Activity/Property Relationship
QSPR	Quantitative Structure Property Relationship
RBF	Radial Basis Functions
REACH	Registration, Evaluation, Authorization and Restriction of Chemicals
RF	Random Forests
RMSE	Root Mean Square Error
ROC	Receiver-Operating Characteristic
ROS	Reactive Oxygen Species
RXR	Retinoid X Receptor
SDF	Structure Data File
SDF	Structure-Data File
SGD	Single-Gaussian Distribution
SIEF	Substance Information Exchange Forum
SMF	Substructure Molecular Fragments
SMILES	Simplified Molecular Input Line Entry Specification
SMILES	Simplified Molecular Input Line Entry Specification
sr	stress response

STD	Standard Deviation
STD-PROB	Standard Deviation and Probability Based DM
SVD	Singular Value Decomposition
SVM	Support Vector Machine
SVs	Support Vectors
TNR	True Negative Rate
ToxRefDB	Toxicity Reference Database
TPR	True Positive Rate
TZDs	Thiazolidinediones
US EPA	United States Environmental Protection Agency
vPvB	very Persistent and very Bioaccumulative
WAT	White Adipose Tissue

List of figures

List of tables

References

(1) European Commission Environment Directorate General. *REACH in Brief*; 2007.

(2) Eec. *Council Regulation 793/93/EEC on the Evaluation and Control of the Risks of Existing Substances*; 1993; Vol. L 84, 5.4.

(3) REACH - Registration, Evaluation, Authorisation and Restriction of Chemicals http://ec.europa.eu/enterprise/sectors/chemicals/reach/index_en.htm (accessed Sep 8, 2014).

(4) World Bank Group http://www.worldbank.org/ (accessed Jun 19, 2015).

(5) Environment Directorate-General - Environment - European Commission http://ec.europa.eu/dgs/environment/index_en.htm (accessed Jun 19, 2015).

(6) Mission - ECHA http://echa.europa.eu/web/guest/about-us/who-we-are/mission (accessed Jan 1, 2015).

(7) Purcell, W. P.; Bass, G. E.; Clayton, J. M. *Strategy of Drug Design: A Guide to Biological Activity*; John Wiley & Sons, 1973.

(8) Rekker, R. F. The History of Drug Research: From Overton to Hansch. *Quant. Struct. Relationships* **1992**, *11*, 195–199.

(9) Brown, A. C.; Fraser, T. R. On the Connection between Chemical Constitution and Physiological Action; with Special Reference to the Physiological Action of the Salts of the Ammonium Bases Derived from Strychnia, Brucia, Thebaia, Codeia, Morphia, and Nicotia. *J. Anat. Physiol.* **1868**, *2*, 224.

(10) Richet, C. On the Relationship between the Toxicity and the Physical Properties of Substances. *Compt Rendus Seances Soc Biol* **1893**, *9*, 775–776.

(11) Lifnick, R. L. Hans Horst Meyer and the Lipoid Theory of Narcosis. *Trends Pharmacol. Sci.* **1989**, *10*, 265–269.

(12) DeJongh, J.; Verhaar, H. J. M.; Hermens, J. L. M. A Quantitative Property-Property Relationship (QPPR) Approach to Estimate in Vitro Tissue-Blood Partition Coefficients of Organic Chemicals in Rats and Humans. *Arch. Toxicol.* **1997**, *72*, 17–25.

(13) Ferguson, J. The Use of Chemical Potentials as Indices of Toxicity. *Proc. R. Soc. London. Ser. B, Biol. Sci.* **1939**, 387–404.

(14) Hammett, L. P. The Effect of Structure upon the Reactions of Organic Compounds. Benzene Derivatives. *J. Am. Chem. Soc.* **1937**, *59*, 96–103.

(15) Hammett, L. P.; others. Physical Organic Chemistry. **1940**.

(16) Hansen, O. R. Hammett Series with Biological Activity. *Acta Chem Scand* **1962**, *16*, 1593–1600.

(17) Hansch, C.; Maloney, P. P.; Fujita, T.; Muir, R. M. Correlation of Biological

Activity of Phenoxyacetic Acids with Hammett Substituent Constants and Partition Coefficients. **1962**.

(18) 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.

(19) Free, S. M.; Wilson, J. W. A Mathematical Contribution to Structure-Activity Studies. *J. Med. Chem.* **1964**, *7*, 395–399.

(20) Hansch, C.; Clayton, J. M. Lipophilic Character and Biological Activity of Drugs II: The Parabolic Case. *J. Pharm. Sci.* **1973**, *62*, 1–21.

(21) Kubinyi, H. Quantitative Structure-Activity Relations. 7. The Bilinear Model, a New Model for Nonlinear Dependence of Biological Activity on Hydrophobic Character. *J. Med. Chem.* **1977**, *20*, 625–629.

(22) Kubinyi, H. Quantitative Structure-Activity Relationships. IV. Non-Linear Dependence of Biological Activity on Hydrophobic Character: A New Model. *Arzneimittelforschung.* **1975**, *26*, 1991–1997.

(23) Cramer, R. D.; Patterson, D. E.; Bunce, J. D. Comparative Molecular Field Analysis (CoMFA). 1. Effect of Shape on Binding of Steroids to Carrier Proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959–5967.

(24) Klebe, G.; Abraham, U.; Mietzner, T. Molecular Similarity Indices in a Comparative Analysis (CoMSIA) of Drug Molecules to Correlate and Predict Their Biological Activity. *J. Med. Chem.* **1994**, *37*, 4130–4146.

(25) Chuman, H.; Karasawa, M.; Fujita, T. A Novel Three-Dimensional QSAR Procedure: Voronoi Field Analysis. *Quant. Struct. Relationships* **1998**, *17*, 313–326.

(26) Silverman, B. D.; Platt, D. E.; Pitman, M.; Rigoutsos, I. Comparative Molecular Moment Analysis (CoMMA). *Perspect. drug Discov. Des.* **1998**, *12*, 183–196.

(27) G Damale, M.; N Harke, S.; A Kalam Khan, F.; B Shinde, D.; N Sangshetti, J. Recent Advances in Multidimensional QSAR (4D-6D): A Critical Review. *Mini Rev. Med. Chem.* **2014**, *14*, 35–55.

(28) Hopfinger, A. J.; Reaka, A.; Venkatarangan, P.; Duca, J. S.; Wang, S. Construction of a Virtual High Throughput Screen by 4D-QSAR Analysis: Application to a Combinatorial Library of Glucose Inhibitors of Glycogen Phosphorylase B. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1151–1160.

(29) Hopfinger, A. J.; Wang, S.; Tokarski, J. S.; Jin, B.; Albuquerque, M.; Madhav, P. J.; Duraiswami, C. Construction of 3D-QSAR Models Using the 4D-QSAR Analysis Formalism. *J. Am. Chem. Soc.* **1997**, *119*, 10509–10524.

(30) Vedani, A.; Dobler, M. 5D-QSAR: The Key for Simulating Induced Fit? *J. Med. Chem.* **2002**, *45*, 2139–2149.

(31) Vedani, A.; Descloux, A.-V.; Spreafico, M.; Ernst, B. Predicting the Toxic

Potential of Drugs and Chemicals in Silico: A Model for the Peroxisome Proliferator-Activated Receptor γ (PPAR γ). *Toxicol. Lett.* **2007**, *173*, 17–23.

(32) Polanski, J. Receptor Dependent Multidimensional QSAR for Modeling Drug-Receptor Interactions. *Curr. Med. Chem.* **2009**, *16*, 3243–3257.

(33) Selassie, C.; Verma, R. P. *History of Quantitative Structure-Activity Relationships*; 2003; Vol. 1.

(34) Langley, J. N. On the Physiology of the Salivary Secretion. *J. Physiol.* **1878**, *1*, 339–369.

(35) Ehrlich, P. Chemotherapeutics: Scientific Principles, Methods and Results. *Lancet* **1913**, *2*, 353–359.

(36) Ehrlich, P. Werbemessung Des Diphtherieheilserums Und Deren Theoretische Grundlagen. [The Assay of the Activity of Diphtheria-Curative Serum and Its Theoreticalbasis]. *Klin Jahr* **1897**, *6*, 299–326.

(37) Langley, J. N. On the Reaction of Cells and of Nerve-Endings to Certain Poisons, Chiefly as Regards the Reaction of Striated Muscle to Nicotine and to Curari. *J. Physiol.* **1905**, *33*, 374–413.

(38) Famulok, M. Oligonucleotide Aptamers That Recognize Small Molecules. *Curr. Opin. Struct. Biol.* **1999**, *9*, 324–329.

(39) Neidle, S. *Molecular Aspects of Anticancer Drug DNA Interactions*; CRC Press, 1994; Vol. 2.

(40) Roberts, G. C. K. Flexible Keys and Deformable Locks: Ligand Binding to Dihydrofolate Reductase. *Pharmacochem Libr* **1983**, *6*, 91–98.

(41) Devillers, J.; Marchand-Geneste, N.; Carpy, A.; Porcher, J.-M. SAR and QSAR Modeling of Endocrine Disruptors. *SAR QSAR Environ. Res.* **2006**, *17*, 393–412.

(42) McIntosh, B. E.; Hogenesch, J. B.; Bradfield, C. A. Mammalian Per-Arnt-Sim Proteins in Environmental Adaptation. *Annu. Rev. Physiol.* **2010**, *72*, 625–645.

(43) Furness, S. G. B.; Whelan, F. The Pleiotropy of Dioxin Toxicity--Xenobiotic Misappropriation of the Aryl Hydrocarbon Receptor's Alternative Physiological Roles. *Pharmacol. Ther.* **2009**, *124*, 336–353.

(44) Bradshaw, T. D.; Bell, D. R. Relevance of the Aryl Hydrocarbon Receptor (AhR) for Clinical Toxicology. *Clin. Toxicol. (Phila).* **2009**, *47*, 632–642.

(45) Poland, A.; Glover, E.; Kende, A. S. Stereospecific, High Affinity Binding of 2,3,7,8-Tetrachlorodibenzo-P-Dioxin by Hepatic Cytosol. Evidence That the Binding Species Is Receptor for Induction of Aryl Hydrocarbon Hydroxylase. *J. Biol. Chem.* **1976**, *251*, 4936–4946.

(46) Denison, M. S.; Seidel, S. D.; Rogers, W. J.; Ziccardi, M.; Winter, G. M.; Heath-Pagliuso, S. Natural and Synthetic Ligands for the Ah Receptor. *Mol. Biol.*

Approaches to Toxicol. **1998**, 393–410.

(47) Safe, S. Polychlorinated Biphenyls (PCBs), Dibenzo-P-Dioxins (PCDDs), Dibenzofurans (PCDFs), and Related Compounds: Environmental and Mechanistic Considerations Which Support the Development of Toxic Equivalency Factors (TEFs). *CRC Crit. Rev. Toxicol.* **1990**, *21*, 51–88.

(48) Poland, A.; Knutson, J. C. 2, 3, 7, 8-Tetrachlorodibenzo-Thorn-Dioxin and Related Halogenated Aromatic Hydrocarbons: Examination of the Mechanism of Toxicity. *Annu. Rev. Pharmacol. Toxicol.* **1982**, *22*, 517–554.

(49) Denison, M. S.; Nagy, S. R. Activation of the Aryl Hydrocarbon Receptor by Structurally Diverse Exogenous and Endogenous Chemicals. *Annu. Rev. Pharmacol. Toxicol.* **2003**, *43*, 309–334.

(50) Nguyen, L. P.; Bradfield, C. A. The Search for Endogenous Activators of the Aryl Hydrocarbon Receptor. *Chem. Res. Toxicol.* **2007**, *21*, 102–116.

(51) Soshilov, A.; Denison, M. S. Role of the Per/Arnt/Sim Domains in Ligand-Dependent Transformation of the Aryl Hydrocarbon Receptor. *J. Biol. Chem.* **2008**, *283*, 32995–33005.

(52) Ikuta, T.; Eguchi, H.; Tachibana, T.; Yoneda, Y.; Kawajiri, K. Nuclear Localization and Export Signals of the Human Aryl Hydrocarbon Receptor. *J. Biol. Chem.* **1998**, *273*, 2895–2904.

(53) Pollenz, R. S.; Sattler, C. A.; Poland, A. The Aryl Hydrocarbon Receptor and Aryl Hydrocarbon Receptor Nuclear Translocator Protein Show Distinct Subcellular Localizations in Hepa 1c1c7 Cells by Immunofluorescence Microscopy. *Mol. Pharmacol.* **1994**, *45*, 428–438.

(54) Hord, N. G.; Perdew, G. H. Physicochemical and Immunocytochemical Analysis of the Aryl Hydrocarbon Receptor Nuclear Translocator: Characterization of Two Monoclonal Antibodies to the Aryl Hydrocarbon Receptor Nuclear Translocator. *Mol. Pharmacol.* **1994**, *46*, 618–626.

(55) Hankinson, O. The Aryl Hydrocarbon Receptor Complex. *Annu. Rev. Pharmacol. Toxicol.* **1995**, *35*, 307–340.

(56) Denison, M. S.; Fisher, J. M.; Whitlock, J. P. The DNA Recognition Site for the Dioxin-Ah Receptor Complex. Nucleotide Sequence and Functional Analysis. *J. Biol. Chem.* **1988**, *263*, 17221–17224.

(57) Hankinson, O. Role of Coactivators in Transcriptional Activation by the Aryl Hydrocarbon Receptor. *Arch. Biochem. Biophys.* **2005**, *433*, 379–386.

(58) Carlson, D. B.; Perdew, G. H. A Dynamic Role for the Ah Receptor in Cell Signaling? Insights from a Diverse Group of Ah Receptor Interacting Proteins. *J. Biochem. Mol. Toxicol.* **2002**, *16*, 317–325.

(59) Beischlag, T. V; Morales, J. L.; Hollingshead, B. D.; Perdew, G. H. The Aryl

Hydrocarbon Receptor Complex and the Control of Gene Expression. *Crit. Rev. Eukaryot. Gene Expr.* **2008**, *18*.

(60) Walker, V. R.; Korach, K. S. Estrogen Receptor Knockout Mice as a Model for Endocrine Research. *ILAR J.* **2004**, *45*, 455–461.

(61) Bocchinfuso, W. P.; Lindzey, J. K.; Hewitt, S. C.; Clark, J. A.; Myers, P. H.; Cooper, R.; Korach, K. S. Induction of Mammary Gland Development in Estrogen Receptor-α Knockout Mice. *Endocrinology* **2000**, *141*, 2982–2994.

(62) Moggs, J. G.; Orphanides, G. Estrogen Receptors: Orchestrators of Pleiotropic Cellular Responses. *EMBO Rep.* **2001**, *2*, 775–781.

(63) Korach, K. S.; Emmen, J. M. A.; Walker, V. R.; Hewitt, S. C.; Yates, M.; Hall, J. M.; Swope, D. L.; Harrell, J. C.; Couse, J. F. Update on Animal Models Developed for Analyses of Estrogen Receptor Biological Activity. *J. Steroid Biochem. Mol. Biol.* **2003**, *86*, 387–391.

(64) Swope, D. L.; Korach, K. S. Estrogen Receptor Biology and Lessons from Knockout Mice. *Encylopedia Horm.* **2003**, *1*, 608–614.

(65) Hileman, B. Environmental Estrogens Linked to Reproductive Abnormalities, Cancer. *Chem. Eng. News* **1994**, *January 31*, 19–23.

(66) Kavlock, R. J.; Daston, G. P.; DeRosa, C.; Fenner-Crisp, P.; Gray, L. E.; Kaattari, S.; Lucier, G.; Luster, M.; Mac, M. J.; Maczka, C.; *et al.* Research Needs for the Risk Assessment of Health and Environmental Effects of Endocrine Disruptors: A Report of the US EPA-Sponsored Workshop. *Environ. Health Perspect.* **1996**, *104*, 715.

(67) Mueller, S. O.; Korach, K. S. Estrogen Receptors and Endocrine Diseases: Lessons from Estrogen Receptor Knockout Mice. *Curr. Opin. Pharmacol.* **2001**, *1*, 613–619.

(68) Shanle, E. K.; Xu, W. Endocrine Disrupting Chemicals Targeting Estrogen Receptor Signaling: Identification and Mechanisms of Action. *Chem. Res. Toxicol.* **2010**, *24*, 6–19.

(69) Adler, S.; Basketter, D.; Creton, S.; Pelkonen, O.; Van Benthem, J.; Zuang, V.; Andersen, K. E.; Angers-Loustau, A.; Aptula, A.; Bal-Price, A.; *et al.* Alternative (Non-Animal) Methods for Cosmetics Testing: Current Status and Future prospects—2010. *Arch. Toxicol.* **2011**, *85*, 367–485.

(70) US EPA, O. The Safe Drinking Water Act Amendments of 1996.

(71) Compilation of laws enforced by the U.S. Food and Drug Administration and related statutes - UW-Madison Libraries https://search.library.wisc.edu/catalog/9910115343502121 (accessed Jun 29, 2015).

(72) Sung, E.; Turan, N.; Ho, P.-L.; Ho, S.-L.; Jarratt, P. D. B.; Waring, R. H.; Ramsden,

D. B. Detection of Endocrine Disruptors--from Simple Assays to Whole Genome Scanning. *Int. J. Androl.* **2012**, *35*, 407–414.

(73) Jacobs, M. N.; Janssens, W.; Bernauer, U.; Brandon, E.; Coecke, S.; Combes, R.; Edwards, P.; Freidig, A.; Freyberger, A.; Kolanczyk, R.; *et al.* The Use of Metabolising Systems for in Vitro Testing of Endocrine Disruptors. *Curr. Drug Metab.* **2008**, *9*, 796–826.

(74) 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.; *et al.* Using in Vitro High Throughput Screening Assays to Identify Potential Endocrine-Disrupting Chemicals. *Env. Heal. Perspect* **2013**, *121*, 7–14.

(75) Cohen Hubal, E. A.; Richard, A.; Aylward, L.; Edwards, S.; Gallagher, J.; Goldsmith, M.-R.; Isukapalli, S.; Tornero-Velez, R.; Weber, E.; Kavlock, R. Advancing Exposure Characterization for Chemical Evaluation and Risk Assessment. *J. Toxicol. Environ. Heal. Part B* **2010**, *13*, 299–313.

(76) Egeghy, P. P.; Judson, R.; Gangwal, S.; Mosher, S.; Smith, D.; Vail, J.; Hubal, E. A. C. The Exposure Data Landscape for Manufactured Chemicals. *Sci. Total Environ.* **2012**, *414*, 159–166.

(77) Judson, R.; Richard, A.; Dix, D. J.; Houck, K.; Martin, M.; Kavlock, R.; Dellarco, V.; Henry, T.; Holderman, T.; Sayre, P.; *et al.* The Toxicity Data Landscape for Environmental Chemicals. *Env. Heal. Perspect* **2009**, *117*, 685–695.

(78) Knudsen, T. B.; Houck, K. A.; Sipes, N. S.; Singh, A. V; Judson, R. S.; Martin, M. T.; Weissman, A.; Kleinstreuer, N. C.; Mortensen, H. M.; Reif, D. M.; *et al.* Activity Profiles of 309 ToxCast™ Chemicals Evaluated across 292 Biochemical Targets. *Toxicology* **2011**, *282*, 1–15.

(79) Rennie, P. S.; Bruchovsky, N.; Leco, K. J.; Sheppard, P. C.; McQueen, S. A.; Cheng, H.; Snoek, R.; Hamel, A.; Bock, M. E.; MacDonald, B. S. Characterization of Two Cis-Acting DNA Elements Involved in the Androgen Regulation of the Probasin Gene. *Mol. Endocrinol.* **1993**, *7*, 23–36.

(80) Persson, H.; Ayer-Le Lievre, C.; Soder, O.; Villar, M. J.; Metsis, M.; others. Expression of Beta-Nerve Growth Factor Receptor mRNA in Sertoli Cells Downregulated by Testosterone. *Science (80-.).* **1990**, *247*, 704.

(81) Léger, J. G.; Montpetit, M. L.; Tenniswood, M. P. Characterization and Cloning of Androgen-Repressed mRNAs from Rat Ventral Prostate. *Biochem. Biophys. Res. Commun.* **1987**, *147*, 196–203.

(82) Quigley, C. A.; Bellis, A. De; Marschke, K. B.; El-Awady, M. K.; Wilson, E. M.; French, F. S. Androgen Receptor Defects: Historical, Clinical, and Molecular Perspectives*. *Endocr. Rev.* **1995**, *16*, 271–321.

(83) Meschede, D.; Behre, H. M.; Nieschlag, E. Disorders of Androgen Target Organs. In *Andrology*; Springer, 2001; pp. 223–240.

(84) Holterhus, P.-M.; Hiort, O.; Demeter, J.; Brown, P. O.; Brooks, J. D. Differential Gene-Expression Patterns in Genital Fibroblasts of Normal Males and 46,XY Females with Androgen Insensitivity Syndrome: Evidence for Early Programming Involving the Androgen Receptor. *Genome Biol.* **2003**, *4*, R37.

(85) Hiort, O.; Sinnecker, G. H. G.; Holterhus, P.-M.; Nitsche, E. M.; Kruse, K. The Clinical and Molecular Spectrum of Androgen Insensitivity Syndromes. *Am. J. Med. Genet.* **1996**, *63*, 218–222.

(86) Chandra, V.; Huang, P.; Hamuro, Y.; Raghuram, S.; Wang, Y.; Burris, T. P.; Rastinejad, F. Structure of the Intact PPAR-[Ggr]-RXR-[Agr] Nuclear Receptor Complex on DNA. *Nature* **2008**, 350–356.

(87) Tontonoz, P.; Spiegelman, B. M. Fat and beyond: The Diverse Biology of PPARγ. *Annu. Rev. Biochem.* **2008**, *77*, 289–312.

(88) Evans, R. M. PPARs and the Complex Journey to Obesity. *Keio J. Med.* **2004**, *53*, 53–58.

(89) Forman, B. M.; Tontonoz, P.; Chen, J.; Brun, R. P.; Spiegelman, B. M.; Evans, R. M. 15-Deoxy-Delta 12, 14-Prostaglandin J2 Is a Ligand for the Adipocyte Determination Factor PPAR Gamma. *Cell* **1995**, *83*, 803–812.

(90) FORMAN, B.; Chen, J.; Evans, R. M. The Peroxisome Proliferator-Activated Receptors: Ligands and Activatorsa. *Ann. N. Y. Acad. Sci.* **1996**, *804*, 266–275.

(91) Kung, J.; Henry, R. R. Thiazolidinedione Safety. *Expert Opin. Drug Saf.* **2012**, *11*, 565–579.

(92) Graham, D. J.; Ouellet-Hellstrom, R.; MaCurdy, T. E.; Ali, F.; Sholley, C.; Worrall, C.; Kelman, J. A. Risk of Acute Myocardial Infarction, Stroke, Heart Failure, and Death in Elderly Medicare Patients Treated with Rosiglitazone or Pioglitazone. *Jama* **2010**, *304*, 411–418.

(93) Nissen, S. E.; Wolski, K. Effect of Rosiglitazone on the Risk of Myocardial Infarction and Death from Cardiovascular Causes. *N. Engl. J. Med.* **2007**, *356*, 2457–2471.

(94) European Medicines Agency recommends suspension of Avandia, Avandamet and Avaglim http://www.ema.europa.eu/ema/index.jsp?curl=pages/news_and_events/news/2010/09/news_detail_001119.jsp (accessed Jun 30, 2015).

(95) Bertilsson, G.; Heidrich, J.; Svensson, K.; Asman, M.; Jendeberg, L.; Sydow-Bäckman, M.; Ohlsson, R.; Postlind, H.; Blomquist, P.; Berkenstam, a. Identification of a Human Nuclear Receptor Defines a New Signaling Pathway for CYP3A Induction. *Proc. Natl. Acad. Sci. U. S. A.* **1998**, *95*, 12208–12213.

(96) Lehmann, J. M.; McKee, D. D.; Watson, M. a.; Willson, T. M.; Moore, J. T.; Kliewer, S. a. The Human Orphan Nuclear Receptor PXR Is Activated by Compounds That Regulate CYP3A4 Gene Expression and Cause Drug

Interactions. *J. Clin. Invest.* **1998**, *102*, 1016–1023.

(97) Falkner, K. C.; Pinaire, J. A.; Xiao, G.-H.; Geoghegan, T. E.; Prough, R. A. Regulation of the Rat Glutathione S-Transferase A2 Gene by Glucocorticoids: Involvement of Both the Glucocorticoid and Pregnane X Receptors. *Mol. Pharmacol.* **2001**, *60*, 611–619.

(98) Staudinger, J. L.; Goodwin, B.; Jones, S. A.; Hawkins-Brown, D.; MacKenzie, K. I.; LaTour, A.; Liu, Y.; Klaassen, C. D.; Brown, K. K.; Reinhard, J.; *et al.* The Nuclear Receptor PXR Is a Lithocholic Acid Sensor That Protects against Liver Toxicity. *Proc. Natl. Acad. Sci.* **2001**, *98*, 3369–3374.

(99) Synold, T. W.; Dussault, I.; Forman, B. M. The Orphan Nuclear Receptor SXR Coordinately Regulates Drug Metabolism and Efflux. *Nat. Med.* **2001**, *7*, 584–590.

(100) Geick, A.; Eichelbaum, M.; Burk, O. Nuclear Receptor Response Elements Mediate Induction of Intestinal MDR1 by Rifampin. *J. Biol. Chem.* **2001**, *276*, 14581–14587.

(101) Graham-Lorence, S.; Khalil, M. W.; Lorence, M. C.; Mendelson, C. R.; Simpson, E. R. Structure-Function Relationships of Human Aromatase Cytochrome P-450 Using Molecular Modeling and Site-Directed Mutagenesis. *J. Biol. Chem.* **1991**, *266*, 11939–11946.

(102) Czajka-Oraniec, I.; Simpson, E. R. Aromatase Research and Its Clinical Significance. *Endokrynol. Pol.* **2010**, *61*, 126–134.

(103) Fukami, M.; Shozu, M.; Ogata, T. Molecular Bases and Phenotypic Determinants of Aromatase Excess Syndrome. *Int. J. Endocrinol.* **2012**, *2012*.

(104) Brown, A. M. Drugs, hERG and Sudden Death. *Cell Calcium* **2004**, *35*, 543–547.

(105) Choe, H.; Nah, K. H.; Lee, S. N.; Lee, H. S.; Lee, H. S.; Jo, S. H.; Leem, C. H.; Jang, Y. J. A Novel Hypothesis for the Binding Mode of HERG Channel Blockers. *Biochem. Biophys. Res. Commun.* **2006**, *344*, 72–78.

(106) Raschi, E.; Ceccarini, L.; De Ponti, F.; Recanatini, M. hERG-Related Drug Toxicity and Models for Predicting hERG Liability and QT Prolongation. **2009**.

(107) Redfern, W. S.; Carlsson, L.; Davis, A. S.; Lynch, W. G.; MacKenzie, I.; Palethorpe, S.; Siegl, P. K. S.; Strang, I.; Sullivan, A. T.; Wallis, R.; *et al.* Relationships between Preclinical Cardiac Electrophysiology, Clinical QT Interval Prolongation and Torsade de Pointes for a Broad Range of Drugs: Evidence for a Provisional Safety Margin in Drug Development. *Cardiovasc. Res.* **2003**, *58*, 32–45.

(108) De Ponti, F.; Poluzzi, E.; Montanaro, N. QT-Interval Prolongation by Non-Cardiac Drugs: Lessons to Be Learned from Recent Experience. *Eur. J. Clin. Pharmacol.* **2000**, *56*, 1–18.

(109) Meyer, T.; Boven, K.-H.; Günther, E.; Fejtl, M. Micro-Electrode Arrays in Cardiac

Safety Pharmacology. *Drug Saf.* **2004**, *27*, 763–772.

(110) Perry, M.; Stansfeld, P. J.; Leaney, J.; Wood, C.; de Groot, M. J.; Leishman, D.; Sutcliffe, M. J.; Mitcheson, J. S. Drug Binding Interactions in the Inner Cavity of HERG Channels: Molecular Insights from Structure-Activity Relationships of Clofilium and Ibutilide Analogs. *Mol. Pharmacol.* **2006**, *69*, 509–519.

(111) Wang, S.; Li, Y.; Xu, L.; Li, D.; Hou, T. Recent Developments in Computational Prediction of HERG Blockage. *Curr. Top. Med. Chem.* **2013**, *13*, 1317–1326.

(112) Mitcheson, J. S. hERG Potassium Channels and the Structural Basis of Drug-Induced Arrhythmias. *Chem. Res. Toxicol.* **2008**, *21*, 1005–1010.

(113) Polak, S.; Wiśniowska, B.; Brandys, J. Collation, Assessment and Analysis of Literature in Vitro Data on hERG Receptor Blocking Potency for Subsequent Modeling of Drugs' Cardiotoxic Properties. *J. Appl. Toxicol.* **2009**, *29*, 183–206.

(114) Chavan, S.; Abdelaziz, A.; Wiklander, J. G.; Nicholls, I. A. A K-Nearest Neighbor Classification of hERG K+ Channel Blockers. *J. Comput. Aided. Mol. Des.* **2016**, *30*, 229–236.

(115) Moi, P.; Chan, K.; Asunis, I.; Cao, A.; Kan, Y. W. Isolation of NF-E2-Related Factor 2 (Nrf2), a NF-E2-like Basic Leucine Zipper Transcriptional Activator That Binds to the Tandem NF-E2/AP1 Repeat of the Beta-Globin Locus Control Region. *Proc. Natl. Acad. Sci.* **1994**, *91*, 9926–9930.

(116) Igarashi, K.; Kataokat, K.; Itoh, K.; Hayashi, N.; Nishizawa, M.; Yamamoto, M. Regulation of Transcription by Dimerization of Erythroid Factor NF-E2 p45 with Small Maf Proteins. **1994**.

(117) Motohashi, H.; Shavit, J. A.; Igarashi, K.; Yamamoto, M.; Engel, J. D. The World according to Maf. *Nucleic Acids Res.* **1997**, *25*, 2953–2959.

(118) Gold, R.; Kappos, L.; Arnold, D. L.; Bar-Or, A.; Giovannoni, G.; Selmaj, K.; Tornatore, C.; Sweetser, M. T.; Yang, M.; Sheikh, S. I.; *et al.* Placebo-Controlled Phase 3 Study of Oral BG-12 for Relapsing Multiple Sclerosis. *N. Engl. J. Med.* **2012**, *367*, 1098–1107.

(119) Itoh, K.; Wakabayashi, N.; Katoh, Y.; Ishii, T.; Igarashi, K.; Engel, J. D.; Yamamoto, M. Keap1 Represses Nuclear Activation of Antioxidant Responsive Elements by Nrf2 through Binding to the Amino-Terminal Neh2 Domain. **1999**, 76–86.

(120) Miller, J. A. Carcinogenesis by Chemicals: An overview—GHA Clowes Memorial Lecture. *Cancer Res.* **1970**, *30*, 559–576.

(121) Sims, P.; Grover, P. L.; Swaisland, A.; Pal, K.; Hewer, A. Metabolic Activation of Benzo (a) Pyrene Proceeds by a Diol-Epoxide. *Nature* **1974**, *252*, 326–328.

(122) Ames, B. N. Dietary Carcinogens and Anticarcinogens Oxygen Radicals and Degenerative Diseases. *Science (80-.).* **1983**, *221*, 1256–1264.

(123) Bannai, S. Induction of Cystine and Glutamate Transport Activity in Human

Fibroblasts by Diethyl Maleate and Other Electrophilic Agents. *J. Biol. Chem.* **1984**, *259*, 2435–2440.

(124) Primiano, T.; Sutter, T. R.; Kensler, T. W. Antioxidant-Inducible Genes. *Adv. Pharmacol.* **1996**, *38*, 293–328.

(125) Hayes, J. D.; Pulford, D. J. The Glutathione S-Transferase Supergene Family: Regulation of GST and the Contribution of the Lsoenzymes to Cancer Chemoprotection and Drug Resistance Part II. *Crit. Rev. Biochem. Mol. Biol.* **1995**, *30*, 521–600.

(126) Buetler, T. M.; Gallagher, E. P.; Wang, C. H.; Stahl, D. L.; Hayes, J. D.; Eaton, D. L. Induction of Phase I and Phase II Drug-Metabolizing Enzyme mRNA, Protein, and Activity by BHA, Ethoxyquin, and Oltipraz. *Toxicol. Appl. Pharmacol.* **1995**, *135*, 45–57.

(127) Cerutti, P.; Ghosh, R.; Oya, Y.; Amstad, P. The Role of the Cellular Antioxidant Defense in Oxidant Carcinogenesis. *Environ. Health Perspect.* **1994**, *102 Suppl*, 123–129.

(128) Biogen Idec. TECFIDERA HIGHLIGHTS OF PRESCRIBING INFORMATION. **2014**.

(129) Prince, M.; Li, Y.; Childers, A.; Itoh, K.; Yamamoto, M.; Kleiner, H. E. Comparison of Citrus Coumarins on Carcinogen-Detoxifying Enzymes in Nrf2 Knockout Mice. *Toxicol. Lett.* **2009**, *185*, 180–186.

(130) Zhang, Y.; Gordon, G. B. A Strategy for Cancer Prevention: Stimulation of the Nrf2-ARE Signaling Pathway. *Mol. Cancer Ther.* **2004**, *3*, 885–893.

(131) DeNicola, G. M.; Karreth, F. A.; Humpton, T. J.; Gopinathan, A.; Wei, C.; Frese, K.; Mangal, D.; Kenneth, H. Y.; Yeo, C. J.; Calhoun, E. S.; *et al.* Oncogene-Induced Nrf2 Transcription Promotes ROS Detoxification and Tumorigenesis. *Nature* **2011**, *475*, 106–109.

(132) Barajas, B.; Che, N.; Yin, F.; Rowshanrad, A.; Orozco, L. D.; Gong, K. W.; Wang, X.; Castellani, L. W.; Reue, K.; Lusis, A. J.; *et al.* NF-E2--Related Factor 2 Promotes Atherosclerosis by Effects on Plasma Lipoproteins and Cholesterol Transport That Overshadow Antioxidant Protection. *Arterioscler. Thromb. Vasc. Biol.* **2011**, *31*, 58–66.

(133) Araujo, J. A. Nrf2 and the Promotion of Atherosclerosis: Lessons to Be Learned. *Clin. Lipidol.* **2012**, *7*, 123–126.

(134) Sikdar, N.; Banerjee, S.; Lee, K.; Wincovitch, S.; Pak, E.; Nakanishi, K.; Jasin, M.; Dutra, A.; Myung, K. DNA Damage Responses by Human ELG1 in S Phase Are Important to Maintain Genomic Integrity. *Cell Cycle* **2009**, *8*, 3199–3207.

(135) Fox, J. T.; Sakamuru, S.; Huang, R.; Teneva, N.; Simmons, S. O.; Xia, M.; Tice, R. R.; Austin, C. P.; Myung, K. High-Throughput Genotoxicity Assay Identifies Antioxidants as Inducers of DNA Damage Response and Cell Death. *Proc. Natl. Acad. Sci.* **2012**, *109*, 5423–5428.

(136) Sorger, P. K. Heat Shock Factor and the Heat Shock Response. *Cell* **1991**, *65*, 363–366.

(137) Morimoto, R. I.; others. Cells in Stress: Transcriptional Activation of Heat Shock Genes. *Sci. YORK THEN WASHINGTON-* **1993**, *259*, 1409.

(138) Guertin, M. J.; Lis, J. T. Chromatin Landscape Dictates HSF Binding to Target DNA Elements. *PLoS Genet.* **2010**, *6*.

(139) Salamanca, H. H.; Fuda, N.; Shi, H.; Lis, J. T. An RNA Aptamer Perturbs Heat Shock Transcription Factor Activity in Drosophila Melanogaster. *Nucleic Acids Res.* **2011**, *39*, 6729–6740.

(140) Salamanca, H. H.; Antonyak, M. a.; Cerione, R. a.; Shi, H.; Lis, J. T. Inhibiting Heat Shock Factor 1 in Human Cancer Cells with a Potent RNA Aptamer. *PLoS One* **2014**, *9*.

(141) Nicholls, D. G.; Ward, M. W. Mitochondrial Membrane Potential and Neuronal Glutamate Excitotoxicity: Mortality and Millivolts. *Trends Neurosci.* **2000**, *23*, 166–174.

(142) Szabadkai, G.; Duchen, M. R. Mitochondria: The Hub of Cellular Ca2+ Signaling. *Physiology* **2008**, *23*, 84–94.

(143) Nicholls, D. G.; Budd, S. L. Mitochondria and Neuronal Survival. *Physiol. Rev.* **2000**, *80*, 315–360.

(144) Szewczyk, A.; Jarmuszkiewicz, W.; Kunz, W. S. Mitochondrial Potassium Channels. *IUBMB Life* **2009**, *61*, 134–143.

(145) Surget, S.; Khoury, M. P.; Bourdon, J. C. Uncovering the Role of p53 Splice Variants in Human Malignancy: A Clinical Perspective. *Onco. Targets. Ther.* **2013**, *7*, 57–67.

(146) Zemojtel, T.; Vingron, M. P53 Binding Sites in Transposons. *Front. Genet.* **2012**, *3*, 3389.

(147) Funk, W. D.; Pak, D. T.; Karas, R. H.; Wright, W. E.; Shay, J. W. A Transcriptionally Active DNA-Binding Site for Human p53 Protein Complexes. *Mol. Cell. Biol.* **1992**, *12*, 2866–2871.

(148) Li, M.; He, Y.; Dubois, W.; Wu, X.; Shi, J.; Huang, J. Distinct Regulatory Mechanisms and Functions for p53-Activated and p53-Repressed DNA Damage Response Genes in Embryonic Stem Cells. *Mol. Cell* **2012**, *46*, 30–42.

(149) Worth, A. P.; Bassan, A.; Gallegos, A.; Netzeva, T. I.; Patlewicz, G.; Pavan, M.; Tsakovska, I.; Vračko, M. *The Characterisation of (Quantitative) Structure-Activity Relationships: Preliminary Guidance*; Institute for Health and Consumer Protection, Toxicology and Chemical Substances Unit, European Chemical Bureau, 2005.

(150) OECD Quantitative Structure-Activity Relationships Project [(Q)SARs]

http://www.oecd.org/chemicalsafety/testing/oecdquantitativestructure-activityrelationshipsprojectqsars.htm (accessed Jun 23, 2015).

(151) Directorate, E.; Meeting, J.; The, O. F.; Committee, C.; Working, T. H. E.; On, P. OECD Environment Health and Safety Publications Series on Testing and Assessment No . 69 GUIDANCE DOCUMENT ON THE VALIDATION OF (QUANTITATIVE) STRUCTURE-ACTIVITY RELATIONSHIP [(Q) SAR] MODELS Environment Directorate. **2007**.

(152) Sahigara, F.; Mansouri, K.; Ballabio, D.; Mauri, A.; Consonni, V.; Todeschini, R. Comparison of Different Approaches to Define the Applicability Domain of QSAR Models. *Molecules* **2012**, *17*, 4791–4810.

(153) Sushko, I. Applicability Domain of QSAR Models, Technical University of Munich, 2011.

(154) Kavlock, R.; Dix, D. Computational Toxicology as Implemented by the U.S. EPA: Providing High Throughput Decision Support Tools for Screening and Assessing Chemical Exposure, Hazard and Risk. *J. Toxicol. Environ. Health. B. Crit. Rev.* **2010**, *13*, 197–217.

(155) Wetmore, B. A.; Wambaugh, J. F.; Ferguson, S. S.; Sochaski, M. A.; Rotroff, D. M.; Freeman, K.; Clewell, H. J.; Dix, D. J.; Andersen, M. E.; Houck, K. A.; *et al.* Integration of Dosimetry, Exposure, and High-Throughput Screening Data in Chemical Toxicity Assessment. *Toxicol. Sci. An Off. J. Soc. Toxicol.* **2012**, *125*, 157–174.

(156) Judson, R. S.; Kavlock, R. J.; Setzer, R. W.; Hubal, E. A. C.; Martin, M. T.; Knudsen, T. B.; Houck, K. A.; Thomas, R. S.; Wetmore, B. A.; Dix, D. J. Estimating Toxicity-Related Biological Pathway Altering Doses for High-Throughput Chemical Risk Assessment. *Chem. Res. Toxicol.* **2011**, *24*, 451–462.

(157) US EPA. Office of Pollution Prevention and Toxics Homepage.

(158) Oprea, T. I.; Tropsha, A. Target, Chemical and Bioactivity Databases - Integration Is Key. *Drug Discov. Today Technol.* **2006**, *3*, 357–365.

(159) 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.; *et al.* In Vitro Screening of Environmental Chemicals for Targeted Testing Prioritization: The ToxCast Project. *Environ. Health Perspect.* **2010**, *118*, 485–492.

(160) 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.

(161) Martin, M. T.; Dix, D. J.; Judson, R. S.; Kavlock, R. J.; Reif, D. M.; Richard, A. M.; Rotroff, D. M.; Romanov, S.; Medvedev, A.; Poltoratskaya, N.; *et al.* Impact of Environmental Chemicals on Key Transcription Regulators and Correlation to Toxicity End Points within EPA's ToxCast Program. *Chem. Res. Toxicol.* **2010**, *23*, 578–590.

(162) 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.; *et al.* In Vitro Screening of Environmental Chemicals for Targeted Testing Prioritization: The ToxCast Project. *Environ. Health Perspect.* **2010**, *118*, 485–492.

(163) 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. An Off. J. Soc. Toxicol.* **2007**, *95*, 5–12.

(164) Toxicology, U. E.-N. C. for C. ToxCast http://epa.gov/ncct/toxcast/data.html (accessed Sep 3, 2014).

(165) Tice, R.; Kavlock, R.; Christopher Austin. The U.S. "Tox21 Community" and the Future of Toxicology http://www.epa.gov/ncct/bosc_review/2009/posters/1-08_Tice_CompTox_BOSC09.pdf (accessed Jan 15, 2014).

(166) Betts, K. S. Tox21 to Date: Steps toward Modernizing Human Hazard Characterization. *Environ. Health Perspect.* **2013**, *121*, A228.

(167) Toxicology, U. E.-N. C. for C. Tox21 http://www.epa.gov/ncct/Tox21/ (accessed Jan 1, 2015).

(168) EPA. Overview of National Research Council Toxicity Testing Strategy http://www.epa.gov/pesticides/science/nrc-toxtesting.html (accessed Sep 8, 2014).

(169) EPA. *Framework for an EPA Chemical Safety for Sustainability Research Program*; 2011.

(170) Robert J. Kavlock Keith A. Houck, Richard S. Judson, D. J. D.; Richard, M. T. M. and A. M. ToxCast: Developing Predictive Signatures for Chemical Toxicity. In *6th World Congress on Alternatives & Animal Use in the Life Sciences*; Japanese Society for Alternatives to Animal Experiments (JSAAE): Tokyo, Japan, 2007.

(171) Martin, M. T.; Knudsen, T. B.; Reif, D. M.; Houck, K. A.; Judson, R. S.; Kavlock, R. J.; Dix, D. J. Predictive Model of Rat Reproductive Toxicity from ToxCast High Throughput Screening. *Biol. Reprod.* **2011**, *85*, 327–339.

(172) Shah, I.; Houck, K.; Judson, R. S.; Kavlock, R. J.; Martin, M. T.; Reif, D. M.; Wambaugh, J.; Dix, D. J. Using Nuclear Receptor Activity to Stratify Hepatocarcinogens. *PLoS One* **2011**, *6*, e14584.

(173) Kleinstreuer, N. C.; Judson, R. S.; Reif, D. M.; Sipes, N. S.; Singh, A. V; Chandler, K. J.; Dewoskin, R.; Dix, D. J.; Kavlock, R. J.; Knudsen, T. B. Environmental Impact on Vascular Development Predicted by High-Throughput Screening. *Environ. Health Perspect.* **2011**, *119*, 1596–1603.

(174) Thomas, R. S.; Black, M. B.; Li, L.; Healy, E.; Chu, T.-M.; Bao, W.; Andersen, M. E.; Wolfinger, R. D. A Comprehensive Statistical Analysis of Predicting in Vivo Hazard Using High-Throughput in Vitro Screening. *Toxicol. Sci.* **2012**, *128*, 398–417.

(175) Bolton, E. E.; Wang, Y.; Thiessen, P. A.; Bryant, S. H. PubChem: Integrated Platform of Small Molecules and Biological Activities. *Annu. Rep. Comput. Chem.* **2008**, *4*, 217–241.

(176) ChemSpider | Search and share chemistry http://www.chemspider.com/ (accessed Jun 27, 2015).

(177) ChemExper https://www.chemexper.com/ (accessed Jun 27, 2015).

(178) Dittmar, P. G.; Stobaugh, R. E.; Watson, C. E. The Chemical Abstracts Service Chemical Registry System. I. General Design. *J. Chem. Inf. Comput. Sci.* **1976**, *16*, 111–121.

(179) CAS, Chemical Abstracts Service Home Page http://www.cas.org/ (accessed Jun 27, 2015).

(180) James, C. A.; Weininger, D.; Delany, J. Daylight Theory Manual. Daylight Chemical Information Systems. *Inc., Irvine, CA* **1995**.

(181) Judson, R.; Richard, A.; Dix, D.; Houck, K.; Elloumi, F.; Martin, M.; Cathey, T.; Transue, T. R.; Spencer, R.; Wolf, M. ACToR--Aggregated Computational Toxicology Resource. *Toxicol Appl Pharmacol* **2008**, *233*, 7–13.

(182) US EPA, O. of W. C. and O. of E. I. EPA ACTOR Downloads http://actor.epa.gov/actor/faces/Download.jsp (accessed Sep 21, 2013).

(183) Sushko, I.; Novotarskyi, S.; Korner, R.; Pandey, A. K.; Rupp, M.; Teetz, W.; Brandmaier, S.; Abdelaziz, A.; Prokopenko, V. V; Tanchuk, V. Y.; *et al.* Online Chemical Modeling Environment (OCHEM): Web Platform for Data Storage, Model Development and Publishing of Chemical Information. *J. Comput. Aided. Mol. Des.* **2011**, *25*, 533–554.

(184) Brandmaier, S.; Peijnenburg, W.; Durjava, M. K.; Kolar, B.; Gramatica, P.; Papa, E.; Bhhatarai, B.; Kovarich, S.; Cassani, S.; Roy, P. P.; *et al.* The QSPR-THESAURUS: The Online Platform of the CADASTER Project. *Altern. Lab. Anim.* **2014**, *42*, 13–24.

(185) CADASTER QSPR-THESAURUS http://www.qspr-thesaurus.eu/home/show.do (accessed Jun 27, 2015).

(186) iPrior - Prioritization and estimation of toxicity of chemcial compounds http://iprior.ochem.eu/home/show.do (accessed Jan 9, 2015).

(187) Abdelaziz, A.; Sushko, Y.; Novotarskyi, S.; Korner, R.; Brandmaier, S.; V Tetko, I. Using Online Tool (iPrior) for Modeling ToxCast™ Assays Towards Prioritization of Animal Toxicity Testing. *Comb. Chem. High Throughput Screen.* **2015**, *18*, 420–438.

(188) Fourches, D.; Muratov, E.; Tropsha, A. Trust, but Verify: On the Importance of Chemical Structure Curation in Cheminformatics and QSAR Modeling Research. *J. Chem. Inf. Model.* **2010**, *50*, 1189–1204.

(189) Martin, Y. C. Let's Not Forget Tautomers. *J. Comput. Aided. Mol. Des.* **2009**, *23*, 693–704.

(190) Young, D.; Martin, T.; Venkatapathy, R.; Harten, P. Are the Chemical Structures in Your QSAR Correct? *QSAR Comb. Sci.* **2008**, *27*, 1337–1345.

(191) Warr, W. a. Scientific Workflow Systems: Pipeline Pilot and KNIME. *J. Comput. Aided. Mol. Des.* **2012**, *26*, 801–804.

(192) Sushko, Y.; Novotarskyi, S.; Körner, R.; Vogt, J.; Abdelaziz, A.; Tetko, I. V. Prediction-Driven Matched Molecular Pairs to Interpret QSARs and Aid the Molecular Optimization Process. *J. Cheminform.* **2014**, *6*, 48.

(193) Berthold, M. R.; Cebron, N.; Dill, F.; Gabriel, T. R.; Kötter, T.; Meinl, T.; Ohl, P.; Sieb, C.; Thiel, K.; Wiswedel, B. {KNIME}: The {K}onstanz {I}nformation {M}iner. In *Studies in Classification, Data Analysis, and Knowledge Organization (GfKL 2007)*; Preisach, C.; Burkhardt, P. D. H.; Schmidt-Thieme, P. D. D. L.; Decker, P. D. R., Eds.; Springer: Freiburg, Germany, 2007; pp. 319–326.

(194) Using SOAP web-services - OCHEM user's manual - eADMET docs http://docs.ochem.eu/display/MAN/Using+SOAP+web-services (accessed Jan 5, 2015).

(195) Tetko, I. V; Sushko, I.; Pandey, A. K.; Zhu, H.; Tropsha, A.; Papa, E.; Oberg, T.; Todeschini, R.; Fourches, D.; Varnek, A. Critical Assessment of QSAR Models of Environmental Toxicity against Tetrahymena Pyriformis: Focusing on Applicability Domain and Overfitting by Variable Selection. *J. Chem. Inf. Model.* **2008**, *48*, 1733–1746.

(196) Tetko, I. V; Bruneau, P.; Mewes, H.-W.; Rohrer, D. C.; Poda, G. I. Can We Estimate the Accuracy of ADME-Tox Predictions? *Drug Discov. Today* **2006**, *11*, 700–707.

(197) Berthold, M. R.; Cebron, N.; Dill, F.; Gabriel, T. R.; Kötter, T.; Meinl, T.; Ohl, P.; Thiel, K.; Wiswedel, B. KNIME-the Konstanz Information Miner: Version 2.0 and beyond. *AcM SIGKDD Explor. Newsl.* **2009**, *11*, 26–31.

(198) Arabie, P.; Baier, N. D.; Critchley, C. F.; Keynes, M. Studies in Classification, Data Analysis, and Knowledge Organization. **2006**.

(199) GNU General Public License http://www.gnu.org/licenses/gpl-3.0.en.html (accessed Jun 28, 2015).

(200) Holmes, G.; Donkin, A.; Witten, I. H. WEKA: A Machine Learning Workbench. In; 1994; pp. 357–361.

(201) Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; Witten, I. H. The WEKA Data Mining Software: An Update. *ACM SIGKDD Explor. Newsl.* **2009**, *11*, 10–18.

(202) R Core Team. R: A Language and Environment for Statistical Computing, 2015.

(203) MATLAB and Statistics Toolbox.

(204) KNIME | Cheminformatics Extensions https://tech.knime.org/cheminformatics-extensions (accessed Jun 28, 2015).

(205) Patel, H.; Bodkin, M. J.; Chen, B.; Gillet, V. J. Knowledge-Based Approach to de Novo Design Using Reaction Vectors. *J. Chem. Inf. Model.* **2009**, *49*, 1163–1184.

(206) KNIME | OCHEM Nodes https://tech.knime.org/book/ochem-nodes (accessed Jun 28, 2015).

(207) Anderson, E.; Veith, G. D.; Weininger, D. *SMILES, a Line Notation and Computerized Interpreter for Chemical Structures*; US Environmental Protection Agency, Environmental Research Laboratory, 1987.

(208) O'Boyle, N. M. Towards a Universal SMILES Representation - A Standard Method to Generate Canonical SMILES Based on the InChI. *J. Cheminform.* **2012**, *4*, 22.

(209) McNaught, A. The Iupac International Chemical Identifier. *Chem. Int.* **2006**, 12–14.

(210) IUPAC - International Union of Pure and Applied Chemistry: The IUPAC International Chemical Identifier (InChI) http://www.iupac.org/home/publications/e-resources/inchi.html (accessed Aug 27, 2014).

(211) US Department of Commerce, N. National Institute of Standards and Technology http://www.nist.gov/ (accessed Jun 27, 2015).

(212) Coles, S. J.; Day, N. E.; Murray-Rust, P.; Rzepa, H. S.; Zhang, Y. Enhancement of the Chemical Semantic Web through the Use of InChI Identifiers. *Org. Biomol. Chem.* **2005**, *3*, 1832–1834.

(213) Prasanna, M. D.; Vondrasek, J.; Wlodawer, A.; Bhat, T. N. Application of InChI to Curate, Index, and Query 3-D Structures. *PROTEINS Struct. Funct. Bioinforma.* **2005**, *60*, 1–4.

(214) Heller, S. R.; McNaught, A.; Pletnev, I.; Stein, S.; Tchekhovskoi, D. *InChI, the IUPAC International Chemical Identifier*; 2015; Vol. 7.

(215) Heller, S.; McNaught, A.; Stein, S.; Tchekhovskoi, D.; Pletnev, I. InChI - the Worldwide Chemical Structure Identifier Standard. *J. Cheminform.* **2013**, *5*, 7.

(216) Pletnev, I.; Erin, A.; McNaught, A.; Blinov, K.; Tchekhovskoi, D.; Heller, S. InChIKey Collision Resistance: An Experimental Testing. *J. Cheminform.* **2012**, *4*, 39.

(217) CTfile Formats http://accelrys.com/products/informatics/cheminformatics/ctfile-formats/no-fee.php (accessed Jun 28, 2015).

(218) Dalby, A.; Nourse, J. G.; Hounshell, W. D.; Gushurst, A. K. I.; Grier, D. L.; Leland, B. A.; Laufer, J. Description of Several Chemical Structure File Formats Used by Computer Programs Developed at Molecular Design Limited. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 244–255.

(219) BIOVIA - Dassault Systèmes http://www.3ds.com/products-services/biovia/ (accessed Jun 28, 2015).

(220) Restrictions, F. Tripos Mol2 File Format http://www.tripos.com/data/support/mol2.pdf (accessed Jun 28, 2015).

(221) Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; John Wiley & Sons, 2008.

(222) Todeschini, R.; Consonni, V. *Molecular Descriptors for Chemoinformatics*; 2nd, Revis ed.; John Wiley & Sons: Milano, Italy, 2009.

(223) Benigni, R.; Bossa, C. Structural Alerts of Mutagens and Carcinogens. *Curr. Comput. - Aided Drug Des.* **2006**, *2*, 169–176.

(224) Sushko, I.; Salmina, E.; Potemkin, V. A.; Poda, G.; Tetko, I. V. ToxAlerts: A Web Server of Structural Alerts for Toxic Chemicals and Compounds with Potential Adverse Reactions. *J. Chem. Inf. Model.* **2012**, *52*, 2310–2316.

(225) Vidal, D.; Thormann, M.; Pons, M. LINGO, an Efficient Holographic Text Based Method to Calculate Biophysical Properties and Intermolecular Similarities. *J. Chem. Inf. Model.* **2005**, *45*, 386–393.

(226) Kubinyi, H. Comparative Molecular Field Analysis (CoMFA). *Handb. Chemoinformatics From Data to Knowl. 4 Vol.* **2008**, 1555–1574.

(227) Cramer 3rd, R. D.; Patterson, D. E.; Bunce, J. D. Recent Advances in Comparative Molecular Field Analysis (CoMFA). *Prog. Clin. Biol. Res.* **1989**, *291*, 161.

(228) Jacob, L.; Hoffmann, B.; Stoven, V.; Vert, J.-P. Virtual Screening of GPCRs: An in Silico Chemogenomics Approach. *BMC Bioinformatics* **2008**, *9*, 363.

(229) Hawkins, D. M.; Basak, S. C.; Kraker, J.; Geiss, K. T.; Witzmann, F. A. Combining Chemodescriptors and Biodescriptors in Quantitative Structure-Activity Relationship Modeling. *J. Chem. Inf. Model.* **2006**, *46*, 9–16.

(230) Sedykh, A.; Zhu, H.; Tang, H.; Zhang, L.; Richard, A.; Rusyn, I.; Tropsha, A. Use of in Vitro HTS-Derived Concentration-Response Data as Biological Descriptors Improves the Accuracy of QSAR Models of in Vivo Toxicity. *Environ. Health Perspect.* **2011**, *119*, 364–370.

(231) Tetko, I. V; Tanchuk, V. Y. Application of Associative Neural Networks for Prediction of Lipophilicity in ALOGPS 2.1 Program. *J. Chem. Inf. Comput. Sci. 42*, 1136–1145.

(232) Tetko, I. V; Poda, G. I.; Ostermann, C.; Mannhold, R. Large-Scale Evaluation of Log P Predictors: Local Corrections May Compensate Insufficient Accuracy and

Need of Experimentally Testing Every Other Compound. *Chem. Biodivers.* **2009**, *6*, 1837–1844.

(233) Du-Cuny, L.; others. Aqueous Solubility of Drug-like Compounds, Universit{ä}ts- und Landesbibliothek Bonn, 2006.

(234) ALOGPS 3.0 - Online Chemical Modeling Environment https://ochem.eu/model/profile.do?id=190384 (accessed Jul 5, 2015).

(235) ADRIANA.Code - Calculation of Molecular Descriptors | Inspiring Chemical Discovery http://www.molecular-networks.com/products/adrianacode (accessed Sep 28, 2013).

(236) Gasteiger, J. Of Molecules and Humans. *J. Med. Chem.* **2006**, *49*, 6429–6434.

(237) Kier, L. B.; Hall, L. H. An Atom-Centered Index for Drug QSAR Models. *Advances in Drug Design*, 1992, *22*, 1–38.

(238) Solov'ev, V. P.; Varnek, A.; Wipff, G. Modeling of Ion Complexation and Extraction Using Substructural Molecular Fragments. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 847–858.

(239) Varnek, A.; Fourches, D.; Horvath, D.; Klimchuk, O.; Gaudin, C.; Vayer, P.; Solov'ev, V.; Hoonakker, F.; Tetko, I. V.; Marcou, G. ISIDA - Platform for Virtual Screening Based on Fragment and Pharmacophoric Descriptors. *Curr. Comput. - Aided Drug Des.* **2008**, *4*, 191–198.

(240) Sawada, R.; Kotera, M.; Yamanishi, Y. Benchmarking a Wide Range of Chemical Descriptors for Drug-Target Interaction Prediction Using a Chemogenomic Approach. *Mol. Inform.* **2014**, *33*, 719–731.

(241) Molecular descriptors calculation - Dragon - Talete srl http://www.talete.mi.it/products/dragon_description.htm (accessed Jul 5, 2015).

(242) Skvortsova, M. I.; Baskin, I. I.; Skvortsov, L. A.; Palyulin, V. A.; Zefirov, N. S.; Stankevich, I. V. Chemical Graphs and Their Basis Invariants. *J. Mol. Struct. THEOCHEM* **1999**, *466*, 211–217.

(243) Cherkasov, A. Inductive QSAR Descriptors. Distinguishing Compounds with Antibacterial Activity by Artificial Neural Networks. *Int. J. Mol. Sci.* **2005**, *6*, 63–86.

(244) Cherkasov, A. Inductive Descriptors: 10 Successful Years in QSAR. *Curr. Comput. Aided. Drug Des.* **2005**, *1*, 21–42.

(245) Bartashevich, E. V; Potemkin, V. A.; Grishina, M. A.; Belik, A. V. A Method for Multiconformational Modeling of the Three-Dimensional Shape of a Molecule. *J. Struct. Chem.* **2002**, *43*, 1033–1039.

(246) Potemkin, V. A.; Grishina, M. A. A New Paradigm for Pattern Recognition of Drugs. *J. Comput. Aided. Mol. Des.* **2008**, *22*, 489–505.

(247) Grishina, M. A.; Bartashevich, E. V; Potemkin, V. A.; Belik, A. V. Genetic Algorithm for Predicting Structures and Properties of Molecular Aggregates in Organic Substances. *J. Struct. Chem.* **2002**, *43*, 1040–1044.

(248) Potemkin, V. A.; Pogrebnoy, A. A.; Grishina, M. A. Technique for Energy Decomposition in the Study of "Receptor-Ligand" Complexes. *J. Chem. Inf. Model.* **2009**, *49*, 1389–1406.

(249) Bultinck, P.; Langenaeker, W.; Lahorte, P.; De Proft, F.; Geerlings, P.; Van Alsenoy, C.; Tollenaere, J. P. The Electronegativity Equalization Method II: Applicability of Different Atomic Charge Schemes. *J. Phys. Chem. A* **2002**, *106*, 7895–7901.

(250) Bultinck, P.; Langenaeker, W.; Carbó-Dorca, R.; Tollenaere, J. P. Fast Calculation of Quantum Chemical Molecular Descriptors from the Electronegativity Equalization Method. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 422–428.

(251) Thormann, M.; Vidal, D.; Almstetter, M.; Pons, M. Nomen Est Omen: Quantitative Prediction of Molecular Properties Directly from IUPAC Names. *Open Appl. Informatics J.* **2007**, *1*, 28–32.

(252) Steinbeck, C.; Hoppe, C.; Kuhn, S.; Floris, M.; Guha, R.; Willighagen, E. L. Recent Developments of the Chemistry Development Kit (CDK)-an Open-Source Java Library for Chemo-and Bioinformatics. *Curr. Pharm. Des.* **2006**, *12*, 2111–2120.

(253) GNU Lesser General Public License https://www.gnu.org/licenses/lgpl-3.0.en.html (accessed Jul 5, 2015).

(254) The Chemistry Development Kit download | SourceForge.net http://sourceforge.net/projects/cdk/ (accessed Jul 5, 2015).

(255) Steinbeck, C.; Han, Y.; Kuhn, S.; Horlacher, O.; Luttmann, E.; Willighagen, E. The Chemistry Development Kit (CDK): An Open-Source Java Library for Chemo-and Bioinformatics. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 493–500.

(256) Calculator Plugins – property prediction & calculation tools « ChemAxon – cheminformatics platforms and desktop applications https://www.chemaxon.com/products/calculator-plugins/?gclid=CjwKEAjwq-OsBRDd95ayprR9wQSJACQnU3Ge_HH6zdiA8vnql8C9zLgsgSZ5f5fSgMqiyFJM-JlghoC-Hzw_wcB (accessed Jul 5, 2015).

(257) Cover, T. M.; Hart, P. E. Nearest Neighbor Pattern Classification. *Inf. Theory, IEEE Trans.* **1967**, *13*, 21–27.

(258) Beyer, K.; Goldstein, J.; Ramakrishnan, R.; Shaft, U. When Is "nearest Neighbor" Meaningful? In *Database Theory—ICDT'99*; Springer, 1999; pp. 217–235.

(259) Aha, D. W.; Kibler, D.; Albert, M. K. Instance-Based Learning Algorithms. *Mach. Learn.* **1991**, *6*, 37–66.

(260) Rosenblatt, F. *The Perceptron, a Perceiving and Recognizing Automaton Project*

Para; First edit.; Cornell Aeronautical Laboratory: New York, USA, 1957.

(261) Tetko, I. V. Associative Neural Network. *Neural Process. Lett.* **2002**, *16*, 187–199.

(262) Tetko, I. V. Neural Network Studies. 4. Introduction to Associative Neural Networks. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 717–728.

(263) Tollenaere, T. SuperSAB: Fast Adaptive Back Propagation with Good Scaling Properties. *Neural Networks* **1990**, *3*, 561–573.

(264) Quinlan, J. R. *C4. 5: Programs for Machine Learning*; Morgan kaufmann, 1993; Vol. 1.

(265) Breiman, L. (University of C. Random Forests. *Mach. Learn.* **1999**, *45*, 1–35.

(266) Breiman, L.; Friedman, J.; Stone, C. J.; Olshen, R. A. *Classification and Regression Trees*; CRC press, 1984.

(267) Hazewinkel, M. Law of Large Numbers. *Encycl. Math. Springer, Berlin* **2001**.

(268) Wold, S.; Geladi, P.; Esbensen, K.; Öhman, J. Multi-Way Principal Components-and PLS-Analysis. *J. Chemom.* **1987**, *1*, 41–56.

(269) Eriksson, L.; Johansson, E. Multivariate Design and Modeling in QSAR. *Chemom. Intell. Lab. Syst.* **1996**, *34*, 1–19.

(270) Zhokhova, N. I.; Baskin, I. I.; Palyulin, V. A.; Zefirov, A. N.; Zefirov, N. S. Fragmental Descriptors with Labeled Atoms and Their Application in QSAR/QSPR Studies. In *Doklady Chemistry*; 2007; Vol. 417, pp. 282–284.

(271) De Jong, S. SIMPLS: An Alternative Approach to Partial Least Squares Regression. *Chemom. Intell. Lab. Syst.* **1993**, *18*, 251–263.

(272) Breiman, L. Bagging Predictors. *Mach. Learn.* **1996**, *24*, 123–140.

(273) Chang, C.-C.; Lin, C.-J. LIBSVM: A Library for Support Vector Machines. *ACM Trans. Intell. Syst. Technol.* **2011**, *2*, 27.

(274) Vapnik, V. N.; Vapnik, V. *Statistical Learning Theory*; Wiley New York, 1998; Vol. 1.

(275) Cristianini, N.; Shawe-Taylor, J. *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*; Cambridge university press, 2000.

(276) Schölkopf, B.; Smola, A. J. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and beyond*; MIT press, 2002.

(277) Cortes, C.; Vapnik, V. Support-Vector Networks. *Mach. Learn.* **1995**, *20*, 273–297.

(278) Organisation for Economic Cooperation and Development. *OECD Guidance Document on the Validation and International Acceptance of New or Updated*

Test Methods for Hazard Assessment; Paris, France, 2005; Vol. 33.

(279) Frank, I. E.; Todeschini, R. *The Data Analysis Handbook*; Elsevier, 1994.

(280) Matthews, B. W. Comparison of the Predicted and Observed Secondary Structure of T4 Phage Lysozyme. *Biochim. Biophys. Acta (BBA)-Protein Struct.* **1975**, *405*, 442–451.

(281) Hanley, J. a; McNeil, B. J. A Method of Comparing the Areas under Receiver Operating Characteristic Curves Derived from the Same Cases. *Radiology* **1983**, *148*, 839–843.

(282) Lobo, J. M.; Jiménez-Valverde, A.; Real, R. AUC: A Misleading Measure of the Performance of Predictive Distribution Models. *Glob. Ecol. Biogeogr.* **2008**, *17*, 145–151.

(283) Hand, D. J. Measuring Classifier Performance: A Coherent Alternative to the Area under the ROC Curve. *Mach. Learn.* **2009**, *77*, 103–123.

(284) Hanczar, B.; Hua, J.; Sima, C.; Weinstein, J.; Bittner, M.; Dougherty, E. R. Small-Sample Precision of ROC-Related Estimates. *Bioinformatics* **2010**, *26*, 822–830.

(285) Keller, H. R.; Massart, D. L.; Brans, J. P. Multicriteria Decision Making: A Case Study. *Chemom. Intell. Lab. Syst.* **1991**, *11*, 175–189.

(286) Hendriks, M. M. W. B.; de Boer, J. H.; Smilde, A. K.; Doornbos, D. A. Multicriteria Decision Making. *Chemom. Intell. Lab. Syst.* **1992**, *16*, 175–191.

(287) Pavan, M.; Todeschini, R. Multicriteria Decision Making Methods. **2009**.

(288) Pavan, M.; Todeschini, R. *Scientific Data Ranking Methods: Theory and Applications*; Elsevier, 2008; Vol. 27.

(289) Alexander, G.; Alexander, T. Beware of Q2. *J Mol Graph Model* **2002**, *20*, 269–276.

(290) Burden, F. R.; Brereton, R. G.; Walsh, P. T. Cross-Validatory Selection of Test and Validation Sets in Multivariate Calibration and Neural Networks as Applied to Spectroscopy. *Analyst* **1997**, *122*, 1015–1022.

(291) Tetko, I. V; Novotarskyi, S.; Sushko, I.; Ivanov, V.; Petrenko, A. E.; Dieden, R.; Lebon, F.; Mathieu, B. Development of Dimethyl Sulfoxide Solubility Models Using 163,000 Molecules: Using a Domain Applicability Metric to Select More Reliable Predictions. *J. Chem. Inf. Model.* **2013**, *53*, 1990–2000.

(292) Kotsiantis, S.; Kanellopoulos, D.; Pintelas, P.; others. Handling Imbalanced Datasets: A Review. *GESTS Int. Trans. Comput. Sci. Eng.* **2006**, *30*, 25–36.

(293) Netzeva, T. I.; Worth, A. P.; Aldenberg, T.; Benigni, R.; Cronin, M. T. D.; Gramatica, P.; Jaworska, J. S.; Kahn, S.; Klopman, G.; Marchant, C. a.; *et al.* Current Status of Methods for Defining the Applicability Domain of (Quantitative) Structure-Activity Relationships. *ATLA Altern. to Lab. Anim.* **2005**,

33, 155–173.

(294) European Chemicals Agency. Guidance on Information Requirements and Chemical Safety Assessment Chapter R . 6 : QSARs and Grouping of Chemicals. **2008**, 134.

(295) Dimitrov, S.; Dimitrova, G.; Pavlov, T.; Dimitrova, N.; Patlewicz, G.; Niemela, J.; Mekenyan, O. A Stepwise Approach for Defining the Applicability Domain of SAR and QSAR Models. *J. Chem. Inf. Model.* **2005**, *45*, 839–849.

(296) Worth, A.; Bassan, a; Gallegos, a; Netzeva, T.; Patlewicz, G.; Pavan, M.; Tsakovska, I.; Vracko, M. The Characterisation of (Quantitative) Structure-Activity Relationships : Preliminary Guidance. *ECB Rep. EUR 21866 Eur. COmmision, Jt. Res. Cent.* **2005**, 95.

(297) Jaworska, J.; Nikolova-Jeliazkova, N.; Aldenberg, T. QSAR Applicability Domain Estimation by Projection of the Training Set in Descriptor Space: A Review. *ATLA Altern. to Lab. Anim.* **2005**, *33*, 445–459.

(298) Preparata, F. P.; Shamos, M. I. Convex Hulls: Basic Algorithms. In *Computational geometry*; Springer, 1985; pp. 95–149.

(299) Sahigara, F.; Ballabio, D.; Todeschini, R.; Consonni, V. Defining a Novel K-Nearest Neighbours Approach to Assess the Applicability Domain of a QSAR Model for Reliable Predictions. *J. Cheminform.* **2013**, *5*, 27.

(300) Gene Ontology Documentation http://www.geneontology.org/GO.contents.doc.shtml (accessed Jan 9, 2015).

(301) Pathway Commons http://www.pathwaycommons.org/about/ (accessed Jan 9, 2015).

(302) Ingenuity IPA - Integrate and understand complex 'omics data http://www.ingenuity.com/products/ipa (accessed Jan 9, 2015).

(303) KEGG: Kyoto Encyclopedia of Genes and Genomes http://www.genome.jp/kegg/ (accessed Jan 9, 2015).

(304) Boyadjiev, S. A.; Jabs, E. W. Online Mendelian Inheritance in Man (OMIM) as a Knowledgebase for Human Developmental Disorders. *Clin. Genet.* **2000**, *57*, 253–266.

(305) Martin, M. T.; Houck, K. A.; McLaurin, K.; Richard, A. M.; Dix, D. J. Linking Regulatory Toxicological Information on Environmental Chemicals with High-Throughput Screening (HTS) and Genomic Data. *Toxicol. CD-An Off. J. Soc. Toxicol.* **2007**, *96*, 219–220.

(306) Martin, M. T.; Judson, R. S.; Reif, D. M.; Kavlock, R. J.; Dix, D. J. Profiling Chemicals Based on Chronic Toxicity Results from the US EPA ToxRef Database. *Environ. Health Perspect.* **2009**, *117*, 392.

(307) US EPA, O. of W. C. and O. of E. I. ToxRefDB (Toxicity Reference Database).

(308) Sadowski, J.; Gasteiger, J.; Klebe, G. Comparison of Automatic Three-Dimensional Model Builders Using 639 X-Ray Structures. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1000–1008.

(309) Tetko, I. V; Tanchuk, V. Y.; Villa, A. E. P. Prediction of N-Octanol/water Partition Coefficients from PHYSPROP Database Using Artificial Neural Networks and E-State Indices. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1407–1421.

(310) Tetko, I. V; Tanchuk, V. Y.; Kasheva, T. N.; Villa, A. E. Estimation of Aqueous Solubility of Chemical Compounds Using E-State Indices. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1488–1493.

(311) Hall, L. H.; Kier, L. B.; Brown, B. B. Molecular Similarity Based on Novel Atom-Type Electrotopological State Indices. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 1074–1080.

(312) Huuskonen, J.; Livingstone, D.; Tetko, I. Neural Network Modeling for Estimation of Partition Coefficient Based on Atom-Type Electrotopological State Indices. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 947–955.

(313) Introduction to Calculator Plugins - Calculator Plugins - ChemAxon - DOCS https://docs.chemaxon.com/display/CALCPLUGS/Introduction+to+Calculator+Plugins (accessed Jan 9, 2015).

(314) Aires-de-Sousa, J.; Gasteiger, J. New Description of Molecular Chirality and Its Application to the Prediction of the Preferred Enantiomer in Stereoselective Reactions. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 369–375.

(315) Varnek, A.; Fourches, D.; Horvath, D.; Klimchuk, O.; Gaudin, C.; Vayer, P.; Solovev, V.; Hoonakker, F.; Tetko, I. V; Marcou, G. ISIDA-Platform for Virtual Screening Based on Fragment and Pharmacophoric Descriptors. *Curr. Comput. Aided. Drug Des.* **2008**, *4*, 191–198.

(316) Cherkasov, A.; Ban, F.; Santos-Filho, O.; Thorsteinson, N.; Fallahi, M.; Hammond, G. L. An Updated Steroid Benchmark Set and Its Application in the Discovery of Novel Nanomolar Ligands of Sex Hormone-Binding Globulin. *J. Med. Chem.* **2008**, *51*, 2047–2056.

(317) Sayed, A. A. GitHub repository - Dissertation https://github.com/amaziz/Dissertation.git (accessed Dec 23, 2015).

(318) Vorberg, S.; Tetko, I. V. Modeling the Biodegradability of Chemical Compounds Using the Online CHEmical Modeling Environment (OCHEM). *Mol. Inform.* **2014**, *33*, 73–85.

(319) Daylight Theory: SMARTS - A Language for Describing Molecular Patterns http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html (accessed Jan 18, 2015).

(320) Schorpp, K.; Rothenaigner, I.; Salmina, E.; Reinshagen, J.; Low, T.; Brenke, J. K.; Gopalakrishnan, J.; Tetko, I. V; Gul, S.; Hadian, K. Identification of Small-

Molecule Frequent Hitters from AlphaScreen High-Throughput Screens. *J. Biomol. Screen.* **2014**, *19*, 715–726.

(321) Roy, K.; Roy, P. P. QSAR of Cytochrome Inhibitors. *Expert Opin. Drug Metab. Toxicol.* **2009**, *5*, 1245–1266.

(322) Lewis, D. F. V; Modi, S.; Dickins, M. Structure-Activity Relationship for Human Cytochrome P450 Substrates and Inhibitors. *Drug Metab. Rev.* **2002**, *34*, 69–82.

(323) Novotarskyi, S.; Sushko, I.; Korner, R.; Pandey, A. K.; Tetko, I. V. A Comparison of Different QSAR Approaches to Modeling CYP450 1A2 Inhibition. *J. Chem. Inf. Model.* **2011**, *51*, 1271–1280.

(324) Fourches, D.; Tropsha, A. Using Graph Indices for the Analysis and Comparison of Chemical Datasets. *Mol. Inform.* **2013**, *32*, 827–842.

(325) Tropsha, A. Best Practices for QSAR Model Development, Validation, and Exploitation. *Mol. Inform.* **2010**, *29*, 476–488.

(326) Golbraikh, A.; Muratov, E.; Fourches, D.; Tropsha, A. Data Set Modelability by QSAR. *J. Chem. Inf. Model.* **2014**, *54*, 1–4.

(327) van de Waterbeemd, H.; Gifford, E. ADMET in Silico Modelling: Towards Prediction Paradise? *Nat. Rev. Drug Discov.* **2003**, *2*, 192–204.

(328) Hou, T.; Wang, J.; Zhang, W.; Wang, W.; Xu, X. Recent Advances in Computational Prediction of Drug Absorption and Permeability in Drug Discovery. *Curr. Med. Chem.* **2006**, *13*, 2653–2667.

(329) Hou, T.; Wang, J. Structure-ADME Relationship: Still a Long Way to Go? **2008**.

(330) EPA ToxCast LELPredictor Marathon Match Results Summary https://web.archive.org/web/20150416015853/http://www.epa.gov/ncct/download_files/ToxCastMMResultSummary.pdf (accessed Dec 24, 2015).

(331) Varnek, A.; Gaudin, C.; Marcou, G.; Baskin, I.; Pandey, A. K.; Tetko, I. V. Inductive Transfer of Knowledge: Application of Multi-Task Learning and Feature Net Approaches to Model Tissue-Air Partition Coefficients. *J. Chem. Inf. Model.* **2009**, *49*, 133–144.

(332) Tetko, I. V; Poda, G. I.; Ostermann, C.; Mannhold, R. Large-Scale Evaluation of Log P Predictors: Local Corrections May Compensate Insufficient Accuracy and Need of Experimentally Testing Every Other Compound. *Chem. Biodivers.* **2009**, *6*, 1837–1844.

(333) Mannhold, R.; Poda, G. I.; Ostermann, C.; Tetko, I. V. Calculation of Molecular Lipophilicity: State-of-the-Art and Comparison of Log P Methods on More than 96,000 Compounds. *J. Pharm. Sci.* **2009**, *98*, 861–893.

(334) Novotarskyi, S.; Abdelaziz, A.; Sushko, Y.; Körner, R.; Vogt, J.; Tetko, I. V. ToxCast EPA in Vitro to in Vivo Challenge: Insight into the Rank-I Model. *Chem. Res. Toxicol.* **2016**, *29*, 768–775.

(335) TopCoder Contest: EPA ToxCast LELPredictor Challenge https://community.topcoder.com/longcontest/stats/?module=ViewOverview&rd=15955 (accessed Dec 24, 2015).

(336) NIH. Tox21 Data Challenge 2014 https://tripod.nih.gov/tox21/challenge/about.jsp.

(337) Chesbrough, H. W. *Open Innovation: The New Imperative for Creating and Profiting from Technology*; Harvard Business Press: Boston, Massachusetts, 2006.

(338) Abdelaziz, A.; Spahn-Langguth, H.; Schramm, K.-W.; Tetko, I. V. Consensus Modeling for HTS Assays Using In Silico Descriptors Calculates the Best Balanced Accuracy in Tox21 Challenge. *Front. Environ. Sci.* **2016**, *4*, 1–12.

(339) Aop:30 - Estrogen receptor antagonism leading to reproductive dysfunction-aopwiki https://aopkb.org/aopwiki/index.php/Aop:30 (accessed Dec 15, 2015).

(340) AID 743077 - qHTS assay to identify small molecule agonists of the estrogen receptor alpha (ER-alpha) signaling pathway: Summary - PubChem BioAssay Summary https://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?aid=743077 (accessed Jul 10, 2015).

(341) AID 743079 - qHTS assay to identify small molecule agonists of the estrogen receptor alpha (ER-alpha) signaling pathway using the BG1 cell line - PubChem BioAssay Summary https://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?aid=743079 (accessed Jul 10, 2015).

(342) AID 743053 - qHTS assay to identify small molecule agonists of the androgen receptor (AR) signaling pathway: Summary - PubChem BioAssay Summary https://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?aid=743053 (accessed Jul 10, 2015).

(343) AID 743040 - qHTS assay to identify small molecule agonists of the androgen receptor (AR) signaling pathway using the MDA cell line - PubChem BioAssay Summary https://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?aid=743040#aDescription (accessed Jul 10, 2015).

(344) AID 743122 - qHTS assay to identify small molecule that activate the aryl hydrocarbon receptor (AhR) signaling pathway: Summary - PubChem BioAssay Summary https://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?aid=743122 (accessed Jul 10, 2015).

(345) AID 743140 - qHTS assay to identify small molecule agonists of the peroxisome proliferator-activated receptor gamma (PPARg) signaling pathway: Summary - PubChem BioAssay Summary https://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?aid=743140 (accessed Jul 10, 2015).

(346) AID 743219 - qHTS assay for small molecule agonists of the antioxidant response element (ARE) signaling pathway: Summary - PubChem BioAssay Summary https://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?aid=743219 (accessed Jul 10, 2015).

(347) AID 743139 - qHTS assay to identify aromatase inhibitors: Summary - PubChem BioAssay Summary https://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?aid=743139 (accessed Jul 10, 2015).

(348) AID 720516 - qHTS assay for small molecules that induce genotoxicity in human embryonic kidney cells expressing luciferase-tagged ATAD5: Summary - PubChem BioAssay Summary https://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?aid=720516 (accessed Jul 10, 2015).

(349) AID 743228 - qHTS assay for small molecule activators of the heat shock response signaling pathway: Summary - PubChem BioAssay Summary https://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?aid=743228 (accessed Jul 10, 2015).

(350) AID 720637 - qHTS assay for small molecule disruptors of the mitochondrial membrane potential: Summary - PubChem BioAssay Summary https://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?aid=720637 (accessed Jul 10, 2015).

(351) AID 720552 - qHTS assay for small molecule agonists of the p53 signaling pathway: Summary - PubChem BioAssay Summary https://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?aid=720552 (accessed Jul 10, 2015).

(352) Tetko, I. V; Tanchuk, V. Y.; Kasheva, T. N.; Villa, A. E. P. Estimation of Aqueous Solubility of Chemical Compounds Using E-State Indices. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1488–1493.

(353) Tox21 Data Challenge 2014 - Final Leaderboard https://tripod.nih.gov/tox21/challenge/leaderboard.jsp (accessed Jun 18, 2015).

(354) AID 720659 - qHTS assay for small molecule activators of the human pregnane X receptor (PXR) signaling pathway - PubChem BioAssay Summary https://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?aid=720659 (accessed Jul 10, 2015).

(355) Hall, L. H.; Kier, L. B. Electrotopological State Indices for Atom Types: A Novel Combination of Electronic, Topological, and Valence State Information. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 1039–1045.

(356) Kier, L. B.; Hall, L. H. Molecular Structure Description. **1999**.

(357) Cherkasov, A.; Jonsson, M. Substituent Effects on Thermochemical Properties

of Free Radicals. New Substituent Scales for C-Centered Radicals. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 1151–1156.

(358) Potemkin, V. A.; Grishina, M. A.; Bartashevich, E. V. Modeling of Drug Molecule Orientation within a Receptor Cavity in the BiS Algorithm Framework. *J. Struct. Chem.* **2007**, *48*, 155–160.

(359) Thijs, G.; Langenaeker, W.; De Winter, H. Application of Spectrophores™ to Map Vendor Chemical Space Using Self-Organising Maps. *J. Cheminform.* **2011**, *3*, 1.

(360) Garrison, P. M.; Tullis, K.; Aarts, J. M.; Brouwer, A.; Giesy, J. P.; Denison, M. S. Species-Specific Recombinant Cell Lines as Bioassay Systems for the Detection of 2,3,7,8-Tetrachlorodibenzo-P-Dioxin-like Chemicals. *Fundam. Appl. Toxicol.* **1996**, *30*, 194–203.

(361) Han, D.; Nagy, S. R.; Denison, M. S. Comparison of Recombinant Cell Bioassays for the Detection of Ah Receptor Agonists. *Biofactors* **2004**, *20*, 11–22.

(362) Zhao, B.; Baston, D. S.; Hammock, B.; Denison, M. S. Interaction of Diuron and Related Substituted Phenylureas with the Ah Receptor Pathway. *J. Biochem. Mol. Toxicol.* **2006**, *20*, 103–113.

(363) AID 2796 - PubChem BioAssay Summary http://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?aid=2796 (accessed Mar 26, 2014).

(364) Takada, N.; Ohmori, N.; Okada, T. Mining Basic Active Structures from a Large-Scale Database. *J. Cheminform.* **2013**, *5*, 15.

(365) Chemaxon Kft. Standardizer User's Guide « ChemAxon – cheminformatics platforms and desktop applications https://www.chemaxon.com/jchem/doc/user/Standardizer.html (accessed Mar 27, 2014).

(366) eADMET GmbH. Online Chemical Modeling Environment https://www.ochem.eu/home/show.do (accessed Mar 27, 2014).

(367) Zauhar, R. J.; Moyna, G.; Tian, L.; Li, Z.; Welsh, W. J. Shape Signatures: A New Approach to Computer-Aided Ligand-and Receptor-Based Drug Design. *J. Med. Chem.* **2003**, *46*, 5674–5690.

(368) Iurii Suschko. Applicability Domain of QSAR Models, Technischer Universität München (TUM), 2012.

(369) Brunnberg, S.; Pettersson, K.; Rydin, E.; Matthews, J.; Hanberg, A.; Pongratz, I. The Basic Helix--Loop--Helix--PAS Protein ARNT Functions as a Potent Coactivator of Estrogen Receptor-Dependent Transcription. *Proc. Natl. Acad. Sci.* **2003**, *100*, 6517–6522.

(370) Pascussi, J.-M.; Gerbal-Chaloin, S.; Duret, C.; Daujat-Chavanieu, M.; Vilarem,

M.-J.; Maurel, P. The Tangle of Nuclear Receptors That Controls Xenobiotic Metabolism and Transport: Crosstalk and Consequences. *Annu. Rev. Pharmacol. Toxicol.* **2008**, *48*, 1–32.

(371) Zhu, L.; Hites, R. A. Identification of Brominated Carbazoles in Sediment Cores from Lake Michigan. *Environ. Sci. Technol.* **2005**, *39*, 9446–9451.

(372) Grigoriadou, A.; Schwarzbauer, J. Non-Target Screening of Organic Contaminants in Sediments from the Industrial Coastal Area of Kavala City (NE Greece). *Water, Air, Soil Pollut.* **2011**, *214*, 623–643.

(373) Reischl, A.; Joneck, M.; Dumler-Gradl, R. Chlorcarbazole in B{ö}den. *Umweltwissenschaften und Schadstoff-forsch.* **2005**, *17*, 197–200.

(374) Tröbs, L.; Henkelmann, B.; Lenoir, D.; Reischl, A.; Schramm, K.-W. Degradative Fate of 3-Chlorocarbazole and 3, 6-Dichlorocarbazole in Soil. *Environ. Sci. Pollut. Res.* **2011**, *18*, 547–555.

(375) Chen, W.-L.; Xie, Z.; Wolschke, H.; Gandrass, J.; Kötke, D.; Winkelmann, M.; Ebinghaus, R. Quantitative Determination of Ultra-Trace Carbazoles in Sediments in the Coastal Environment. *Chemosphere* **2016**, *150*, 586–595.

(376) Mumbo, J.; Henkelmann, B.; Abdelaziz, A.; Pfister, G.; Nguyen, N.; Schroll, R.; Munch, J. C.; Schramm, K.-W. Persistence and Dioxin-like Toxicity of Carbazole and Chlorocarbazoles in Soil. *Environ. Sci. Pollut. Res. Int.* **2014**.

(377) Guo, J.; Chen, D.; Potter, D.; Rockne, K. J.; Sturchio, N. C.; Giesy, J. P.; Li, A. Polyhalogenated Carbazoles in Sediments of Lake Michigan: A New Discovery. *Environ. Sci. Technol.* **2014**, *48*, 12807–12815.

(378) Parette, R.; McCrindle, R.; McMahon, K. S.; Pena-Abaurrea, M.; Reiner, E.; Chittim, B.; Riddell, N.; Voss, G.; Dorman, F. L.; Pearson, W. N. Halogenated Indigo Dyes: A Likely Source of 1,3,6,8-Tetrabromocarbazole and Some Other Halogenated Carbazoles in the Environment. *Chemosphere* **2015**, *127*, 18–26.

(379) Wishart, D. S.; Knox, C.; Guo, A. C.; Shrivastava, S.; Hassanali, M.; Stothard, P.; Chang, Z.; Woolsey, J. DrugBank: A Comprehensive Resource for in Silico Drug Discovery and Exploration. *Nucleic Acids Res.* **2006**, *34*, D668-72.

(380) Ruffolo Jr, R. R.; Boyle, D. A.; Venuti, R. P.; Lukas, M. A. Carvedilol (Kredex): A Novel Multiple Action Cardiovascular Agent. *Drugs Today* **1991**, *27*, 465–492.

(381) Ruffolo, R. R.; Gellai, M.; Hieble, J. P.; Willette, R. N.; Nichols, A. J. The Pharmacology of Carvedilol. *Eur. J. Clin. Pharmacol.* **1990**, *38*, 82–88.

(382) Nichols, A. J.; Sulpizio, A. C.; Ashton, D. J.; Hieble, J. P.; Ruffolo Jr., R. R. In Vitro Pharmacologic Profile of the Novel Beta-Adrenoceptor Antagonist and Vasodilator, Carvedilol. *Pharmacology* **1989**, *39*, 327–336.

(383) Nichols, A. J.; Gellai, M.; Ruffolo Jr, R. R. Studies on the Mechanism of Arterial Vasodilation Produced by the Novel Antihypertensive Agent, Carvedilol.

Fundam Clin Pharmacol **1991**, *5*, 25–38.

(384) Strub, K. M.; Aeppli, L.; Müller, R. K. Pharmacological Properties of Carprofen. *Eur. J. Rheumatol. Inflamm.* **1981**, *5*, 478–487.

(385) The European Agency for the Evaluation of Medicinal Products. COMMITTEE FOR VETERINARY MEDICINAL PRODUCTS - CARPROFEN - SUMMARY REPORT http://www.ema.europa.eu/docs/en_GB/document_library/Maximum_Residue_Limits_-_Report/2009/11/WC500011412.pdf (accessed Apr 3, 2016).

(386) Carbazoles - Halogenated Heterocycles | Sigma-Aldrich http://www.sigmaaldrich.com/chemistry/chemistry-products.html?TablePage=16269059 (accessed Dec 23, 2015).

(387) Riddell, N.; Jin, U. H.; Safe, S.; Cheng, Y.; Chittim, B.; Konstantinov, A.; Parette, R.; Pena-Abaurrea, M.; Reiner, E. J.; Poirier, D.; *et al.* Characterization and Biological Potency of Mono- to Tetra-Halogenated Carbazoles. *Environ. Sci. Technol.* **2015**, *49*, 10658–10666.

7 List of supplementary materials

Supplementary 1: List of *in vivo* endpoints from ToxCast / ToxRefDB, their respective total number of hits and whether it was selected for modeling.

Supplementary 2: List of ToxCast Phase I chemicals excluded from modeling due to failed descriptors calculation.

Supplementary 3: List of *in vitro* assay endpoints, their respective total number of hits and whether it was selected for modeling.

Supplementary 4: Statistical parameters for the models with best balanced-accuracy for each of the 144 *in vitro* assay endpoints from the ToxCast database.

Supplementary 5: Statistical parameters for the models with best balanced-accuracy for each of the 61 *in vivo* toxicological endpoints from the Toxicity reference database.

7.1.1 Supplementary 1: List of *in vivo* endpoints from ToxCast / ToxRefDB, their respective total number of hits and whether it was selected for modeling.

https://amaziz.com/dissertation/supplementary

7.1.2 Supplementary 2: List of *in vitro* assay endpoints, their respective total number of hits and whether it was selected for modeling.

https://amaziz.com/dissertation/supplementary

7.1.3 Supplementary 3: List of ToxCast Phase I chemicals excluded from modeling due to failed descriptors calculation.

https://amaziz.com/dissertation/supplementary

7.1.4 Supplementary 4: Statistical parameters for the models with best balanced-accuracy for each of the 144 *in vitro* assay endpoints from the ToxCast database.

https://amaziz.com/dissertation/supplementary

7.1.5 Supplementary 5: Statistical parameters for the models with best balanced-accuracy for each of the 61 *in vivo* toxicological endpoints from the Toxicity reference database.

https://amaziz.com/dissertation/supplementary

List of cited publications

A. Abdelaziz, A.; Sushko, Y.; Novotarskyi, S.; Korner, R.; Brandmaier, S.; V Tetko, I. Using Online Tool (iPrior) for Modeling ToxCast™ Assays Towards Prioritization of Animal Toxicity Testing. Comb. Chem. High Throughput Screen. 2015, 18, 420–438.
B. Abdelaziz, A.; Spahn-Langguth, H.; Schramm, K.-W.; Tetko, I. V. Consensus Modeling for HTS Assays Using *in silico* Descriptors calculates the Best Balanced Accuracy in Tox21 Challenge. Front. Environ. Sci. 2016, 4, 1–12.

Hiermit erkläre ich an Eides statt, dass ich alleiniger, federführender Hauptautor der zwei oben genannten Publikationen und Studien bin, die in dieser Arbeit wörtlich zitiert wurden. Die betreffenden Passagen wurden ausschließlich von mir verfasst.

Ort, den ....

Unterschrift

Software used

Figures 1, 2, 3, 4, 7, 13, 14, 15, 18(a) were partially generated with **Gliffy** (http://www.gliffy.com/)

Figures 5, 19, 23, 24 were partially generated with **Microsoft Excel,** Microsoft Corporation (https://products.office.com/en/excel)

Figures 6, 16, 17, 18(b), 20, 30, 31, 32, 37, 38, 39, 40, 44, 45 were partially generated with **R**, The R Foundation for Statistical Computing (http://www.R-project.org)

Figure 10 was partially generated with **Skitch**, Evernote Corporation (https://evernote.com/skitch/)

Figures 33, 34, 35, 36 were partially generated with **OCHEM** (https://www.ochem.eu)

Figures 21, 22, 25, 28, 29 were partially generated with **KNIME,** KNIME.COM AG (http://knime.org)

Figures 26, 27, 41, 42, 43 were partially generated with **Marvin Sketch,** Chemaxon Kft. (https://www.chemaxon.com)

Publication record

Work from this thesis were presented and/or discussed through the following scientific avenues:

Peer reviewed articles

1. Mansouri, K.; Abdelaziz, A.; Rybacka, A.; Roncaglioni, A.; Tropsha, A.; Varnek, A.; Zakharov, A.; Worth, A.; Richard, A. M.; Grulke, C. M.; others. CERAPP: Collaborative Estrogen Receptor Activity Prediction Project. J. Environ. Heal. Perspect. 2016.
2. Abdelaziz, A.; Spahn-Langguth, H.; Schramm, K.-W.; Tetko, I. V. Consensus Modeling for HTS Assays Using In Silico Descriptors Calculates the Best Balanced Accuracy in Tox21 Challenge. Front. Environ. Sci. 2016, 4, 1–12.
3. Novotarskyi, S.; Abdelaziz, A.; Sushko, Y.; Körner, R.; Vogt, J.; Tetko, I. V. ToxCast EPA in Vitro to in Vivo Challenge: Insight into the Rank-I Model. Chem. Res. Toxicol. 2016, 29, 768–775.
4. Chavan, S.; Abdelaziz, A.; Wiklander, J. G.; Nicholls, I. A. A K-Nearest Neighbor Classification of hERG K+ Channel Blockers. J. Comput. Aided. Mol. Des. 2016, 30, 229–236.
5. Abdelaziz, A.; Sushko, Y.; Novotarskyi, S.; Korner, R.; Brandmaier, S.; V Tetko, I. Using Online Tool (iPrior) for Modeling ToxCastTM Assays Towards Prioritization of Animal Toxicity Testing. Comb. Chem. High Throughput Screen. 2015, 18, 420–438.
6. Sushko Y, Novotarskyi S, Körner R, Vogt J, Abdelaziz A, Tetko I V: Prediction-driven matched molecular pairs to interpret QSARs and aid the molecular optimization process. J Cheminform 2014, 6:48.
7. Brandmaier, S.; Peijnenburg, W.; Durjava, M. K.; Kolar, B.; Gramatica, P.; Papa, E.; Bhhatarai, B.; Kovarich, S.; Cassani, S.; Roy, P. P.; Rahmberg, M.; Öberg, T.; Jeliazkova, N.; Golsteijn, L.; Comber, M.; Charochkina, L.; Novotarskyi, S.; Sushko, I.; Abdelaziz, A.; D'Onofrio, E.; Kunwar, P.; Ruggiu, F.; Tetko, I. V. The QSPR-THESAURUS: The Online Platform of the CADASTER Project. Altern. Lab. Anim. 2014, 42, 13–24.
8. Mumbo, J.; Henkelmann, B.; Abdelaziz, A.; Pfister, G.; Nguyen, N.; Schroll, R.; Munch, J. C.; Schramm, K.-W. Persistence and Dioxin-like Toxicity of Carbazole and Chlorocarbazoles in Soil. Environ. Sci. Pollut. Res. 2015, 22, 1344–1356.
9. Sushko I, Novotarskyi S, Korner R, Pandey AK, Rupp M, Teetz W, Brandmaier S, Abdelaziz A, Prokopenko V V, Tanchuk VY, Todeschini R, Varnek A, Marcou G, Ertl P, Potemkin V, Grishina M, Gasteiger J, Schwab C, Baskin II, Palyulin VA, Radchenko E V, Welsh WJ, Kholodovych V, Chekmarev D, Cherkasov A, Aires-de-Sousa J, Zhang QY, Bender A, Nigsch F, Patiny L, et al.: Online chemical modeling environment (OCHEM): web platform for data storage, model development and publishing of chemical information. J Comput Aided Mol Des 2011, 25:533–554.

Posters

1. "Development of classification models for the Aryl Hydrocarbon Receptor (AHR) activators using HTS big dataset: Utilizing applicability domain and addressing unbalance in the dataset" (248th ACS meeting, San Francisco, August 2014)
2. "OCHEM: Online public platform for human and environmental toxicity modeling" (248th ACS meeting, San Francisco, August 2014)

3. “Matched Molecular Pairs analysis using OCHEM” Iurii Sushko, Ahmed Abdelaziz, …. Igor V. Tetko (50th International Conference on Medicinal Chemistry, July 2014)
4. “On-line Chemical modeling environment – database and models for physico-chemical properties” Tetko, I.V.; Sushko, I.; Novotarskyi, S.; Körner, R.; Abdelaziz, A. 3rd World Conference on Physico Chemical Methods in Drug Discovery and Development. September 20 -26, 2013, Dubrovnik, Croatia, p. 11.
5. “Challenges with development of a melting point model using public data: not all errors are not the same!” Tetko, I.V.; Patiny, L.; Charochkina, L.; Sushko, I.; Abdelaziz, A.; Asiri, A.M. 3rd World Conference on Physico Chemical Methods in Drug Discovery and Development. September 20 -26, 2013, Dubrovnik, Croatia, p. 40.
6. “On-line chemical modeling environment: Public user contributed tools for drug discovery.” Tetko, I.V.; Sushko, I.; Novotarskyi, S.; Körner, R.; Abdelaziz, A. In International Conference on Medicinal Chemistry: RICT2013. July 3-5 2013, Nice France, p. 366.
7. "Combining HTS *in vitro* assays with *in silico* descriptors for Liver toxicity modeling" Abdelaziz A.; Tetko, I., 244th American Chemical Society meeting (Philadelphia, US, August, 2012)
8. "Using ToxCast™ HTS assays as biologically derived descriptors in QSAR" Abdelaziz A.; Tetko, I., 3rd Strasbourg summer school on chemoinformatics, Strasbourg, France, June 2012
9. “QSAR modeling for the evaluation of Aryl Hydrocarbon receptor activators” Abdelaziz A.; Alexander Safanayev; Tetko, I., ADME and predictive Toxicology Europe & Munich Interact, Munich, Germany, March 2012
10. "QSAR modeling for the evaluation of Aryl Hydrocarbon receptor activators" Abdelaziz A.; Alexander Safanayev; Tetko, I., 12th symposium on ePhyschem, Budapest, Hungary, March 2012
11. “QSAR modeling for *In vitro* assays: linking ToxCast™ database to the integrated modeling framework, OCHEM” Ahmed Abdelaziz, Iurii Sushko, Wolfram Teetz, Robert Körner, Sergii Novotarskyi, Igor V. Tetko, German Conference on Chemoinformatics, Goslar, Germany, 6-8 November 2011
12. “QSAR modeling for *In vitro* assays: linking ToxCast™ database to the integrated modeling framework-OCHEM” Abdelaziz A.; Körner R.; Novotarskyi S.; Teetz W.; Sushko I.; Tetko, I., German Conference on Bioinformatics, Weihenstephan, Germany, 7-9 September 2011
13. "Active and Reactive Metabolites Formed During Hepatic First-Pass: Simulations Featuring Their Contribution to the Overall Effect in Altered Liver Clearance and Drug-Drug Interactions” OpenTox 2011 InterAction Meeting Program, Munich, Germany, 9-12 August 2011
14. “Public QSAR framework with integrated measurements database” Abdelaziz A.; Körner R.; Novotarskyi S.; Teetz W; Pandey A.; Sushko I.; Rupp M.; Tetko, I:OCHEM: Chemaxon eUGM 2011, Budapest, Hungary, 15-20 May 2011
15. “Stepwise D-Optimal design based on latent variables” Brandmaier, S.; Abdelaziz, A.; Sahlin, U.; Oberg, T.; Tetko, I., interact 2011 Munich, Munich, Germany, April 7, 2011

16. “Prediction of kinetic characteristics of drug metabolites in-silico: The distribution characteristics of beta-adrenoceptor antagonists” Abdelaziz A.; Tetko, I.; Spahn-Langguth H., ADMET Europe 2011, Munich, Germany, 28-29 March 2011
17. "OCHEM: public QSAR framework for modeling PK/PD parameters" Abdelaziz A.; Körner R.; Novotarskyi S.; Teetz W; Pandey A.; Sushko I.; Rupp M.; Tetko, I., ADMET Europe 2011, Munich, Germany, 28-29 March 2011
18. Active Metabolites Formed During Hepatic First-pass: Modelling Serum Concentration-Time Profiles [15th Scientific Symposium of the Austrian Pharmacological Society (APHAR)(November 2009 in Graz, Austria)] and [Jahrestagung der DphG 2008 (October 2008 in Bonn, Germany)] and [69th FIP congress (September 2009 in Istanbul, Turkey)]
19. Active metabolites formed during hepatic first-pass: Simulations featuring their contribution to the overall effect in altered liver clearance. [Jahrestagung der DGPT 2009 (March 2009 in Mainz, Germany)]

Talks/ invited lectures

1. Machine-learning applications in clinical research, (Technical University of Munich, GRK Seminar, Freising, Germany, May 2016)
2. Tox21 Challenge to Build Predictive Models of Nuclear Receptor and Stress-Response Pathways as Mediated by Exposure to Environmental Toxicants and Drugs, (Society of Toxicology annual meeting, New Orleans, US, March 2016)
3. The AOP community outreach – AOP-XML: A format standard between AOP-KB modules and interested third parties, (OpenTox Basel, Switzerland, March 2016)
4. “Analyzing ToxCast Phase I HTS assays as a potential descriptors” (ECO closing conference, Chiemsee, Germany, September 2013)
5. “QSAR modeling for the evaluation of Aryl Hydrocarbon receptor activators” (ECO closing conference, Chiemsee, Germany, September 2013)
6. “QSAR in the cloud; OCHEM, a free online platform for modeling and interpretation” (EPFL Workshop on chemical information, Lausanne, Switzerland, August 2013)
7. “Combining HTS *in vitro* assays with *in silico* descriptors for Liver toxicity modeling" 244th American Chemical Society meeting (Philadelphia, US, August 2012)
8. "QSAR modeling for the evaluation of Aryl Hydrocarbon receptor activators" Abdelaziz A.; Alexander Safanayev; Tetko, I., 244th American Chemical Society meeting (Philadelphia, US, August, 2012)
9. "Kinetics of active metabolites: Compartmental approach and in-silico predictions accounting for first-pass metabolism” (Karl-Franzens-Universität Graz, Inst. Fur Pharmazeutische Wissenschaften Bereich Pharmazeutische Chemie, Graz, Austria, June 17, 2010)
10. Drug design summer school – 2007 & 2008 Tübingen, Germany

Internship

1. Pfizer Global Research Site (Groton, US, March-April 2013)
2. Prof. dr. ir. W.J.G.M. (Willie) Peijnenburg (Institute of Environmental Sciences (CML), Faculty of Science, University of Leiden, The Netherlands, October-December 2012)
3. Prof. Dr. Hilde Spahn-Langguth (Mainz, July-August 2012)

Curriculum Vitae

Ahmed Abdelaziz Sayed

Telefon: (+49) 1577 688 7277
Email: contact@amaziz.com
www.amaziz.com

Akademische Ausbildung

- Doktorand in Chemoinformatik; Technische Universität München, Titel:: "*In silico* modeling using *in vitro* High Throughput Screening data for toxicity prediction within REACH". Doktorvater: Prof. Karl-Werner Schramm
- Executive MBA student in Innovation and Business Creation - TU München.
- M.Sc. in der pharmazeutischen Chemie; Fakultät für Pharmazie und Biotechnologie, Deutsche Universität in Kairo, Titel: "Kinetics of active metabolites: In-silico predictions and compartmental approach accounting for first-pass metabolism" (2010; A+)
- B.Sc. Pharmazeutische Wissenschaften; Fakultät für Pharmazie, Ain-Shams-Universität, Kairo, Ägypten. (2005; Very Good with honors)
- Data Science Specialization by Johns Hopkins University

Berufserfahrung

05/2015– gegenwärtig	Independent Berater - Rosettastein Consulting
03/2011 – 01/2015	Chief Commercial Officer, Managing Director - eADMET GmbH
02/2010 – 04/2013	FP7 Marie-Curie fellow, EU researcher - Helmholtz-Zentrum München
09/2006 – 01/2010	Teaching & Research Assistant - German University in Cairo (GUC)

Fähigkeiten

- **Programmiersprachen**: C#, VB.NET, ASP.NET, Java, PHP
- **Datenbanken**: SQL server, MySQL/MariaDB, MongoDB
- **Betriebssysteme**: Linux, MacOSX, Windows
- **Daten Wissenschaft Software**: R, WEKA, Orange, qtiPlot, SPSS, PSPP
- **PK/PD Software**: GastroPlus, ADMET Predictor, WinNonLin, SimCYP
- **Sprachen**: English (fluent), German (business proficiency), Arabic (native)

Auszeichnungen

- Best balanced accuracy for machine-learning predictive models in the NIT/NCATS TOX21 data challenge 2014
- TUM Graduate school scholarship for the executive MBA studies in innovation and business creation.
- FP7 Marie-Curie fellowship for the environmental Chemoinformatics project (2010-2013)
- Best Industrial contribution for 2009 by FIP (International pharmaceutical Federation)
- Semi-finalist in California Berkeley technology entrepreneurship competition 2008. (Intel+ UC Berkeley Technology entrepreneurship challenge)
- First place in the "3rd Arab Technology Business plan competition" by ASTF (Arab science and Technology Foundation) (Morocco, October 2008)
- IEEE Egypt Gold award for (Made in Egypt) competition