UNIVERSITY OF TARTU
Faculty of Science and Technology
Institute of Computer Science
Data Science Curriculum

Ida Rahu

Machine learning for assessing toxicity of chemicals identified with mass spectrometry

Master's Thesis (15 ECTS)

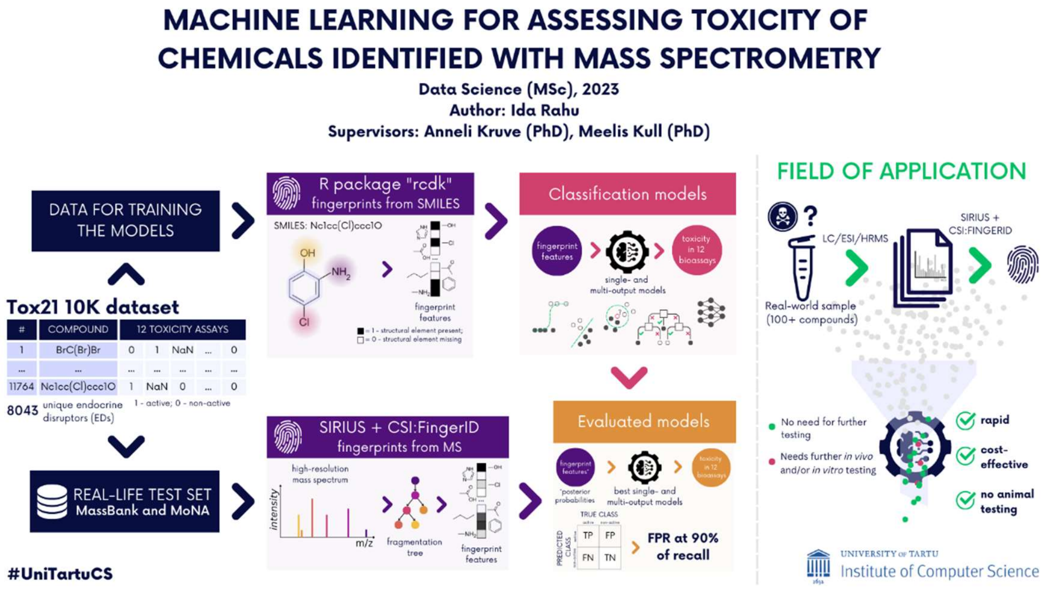
Supervisors: Anneli Kruve, PhD
Meelis Kull, PhD

Tartu 2023

Machine learning for assessing toxicity of chemicals identified with mass spectrometry

Abstract:

Real-world samples can contain hundreds to thousands of chemicals, with endocrine-disrupting chemicals (EDCs) posing a severe threat to human health. Unfortunately, reliable and rapid methods for detecting these compounds from complex mixtures are lacking. One of the potential solutions could be to leverage the capabilities of non-target liquid chromatography high-resolution mass spectrometry (LC/HRMS) combined with machine learning methods. This study aimed to investigate whether the biochemical activity of compounds can be estimated based on chemical fingerprints calculated from HRMS spectra and thereby flag the compounds that require further analysis due to the potential risk they pose to human health. For that, several classification models based on a variety of machine learning algorithms were trained, and their accuracy was evaluated using chemical fingerprints derived from experimental mass spectra. As a result, it was found that the proposed methodology has great potential in the field of *in silico* toxicology.



MACHINE LEARNING FOR ASSESSING TOXICITY OF CHEMICALS IDENTIFIED WITH MASS SPECTROMETRY

Data Science (MSc), 2023
Author: Ida Rahu
Supervisors: Anneli Kruve (PhD), Meelis Kull (PhD)

DATA FOR TRAINING THE MODELS

Tox21 10K dataset

#	COMPOUND	12 TOXICITY ASSAYS				
1	BrC(Br)Br	0	1	NaN	...	0
...	...	...	...	...	...	...
11764	Nc1cc(Cl)ccc1O	1	NaN	0	...	0

8043 unique endocrine disruptors (EDs)

1 - active; 0 - non-active

REAL-LIFE TEST SET MassBank and MoNA

#UniTartuCS

Keywords:

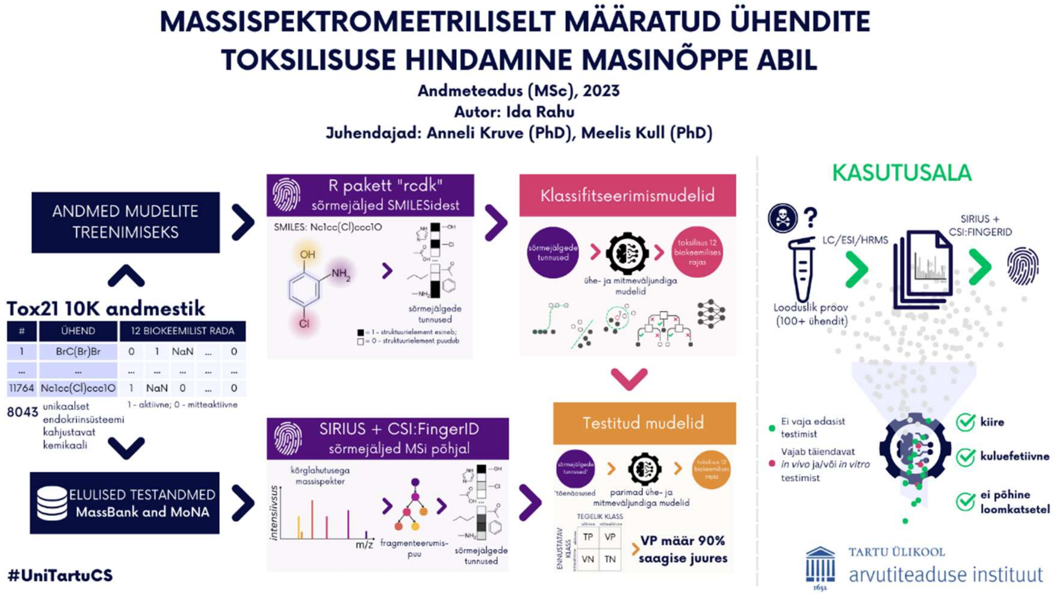
High-resolution mass spectrometry, molecular fingerprints, endocrine disruptors, Tox21, multi-task learning

CERCS: P176 – Artificial intelligence, P300 – Analytical chemistry, P305 – Environmental chemistry, B740 – Pharmacological sciences, pharmacognosy, pharmacy, toxicology

Massispektromeetriliselt määratud ühendite toksilisuse hindamine masinõppe abil

Lühikokkuvõte:

Looduslikud proovid võivad sisaldada tuhandeid ühendeid, millest mitmed on inimestele kahjulikud. Viimase kümnendi jooksul, on leitud, et eriti suurt riski kujutavad endast kemikaalid, mis avaldavad mõju endokriinsüsteemile. Kahjuks aga puuduvad seni usaldusväärsed ja kiired meetodid, mis suudaksid neid keerulistest ainete segudest detekteerida. Üheks lahenduseks võiks olla tundamatute ühendite analüüsimeetodite, mis põhinevad kõrglahutus-massispektromeetrial (HRMS), rakendamine koos masinõppega. Antud töös uuriti, kas HRMS spektri põhjal leitud keemiliste sõrmejälgede alusel on võimalik hinnata proovis sisalduvate ühendite biokeemilist aktiivsust ja seeläbi märgistada ühendeid, mis vajaksid edasist analüüsi, kuna võivad avaldada potentsiaalselt kahjulikku mõju endokriinsüsteemile. Selleks treeniti mitmeid erinevatel algoritmidel põhinevaid klassifitseerimismudelid ning hinnati nende täpsust kasutades eksperimentaalsetest massispektritest arvutatud keemilisi sõrmejälgi. Töö tulemusena leiti, et selline meetod on rakendatav ning omab suurt potentsiaali *in silico* toksikoloogias.



MASSISPEKTROMEETRILISELT MÄÄRATUD ÜHENDITE
TOKSILISUSE HINDAMINE MASINÕPPE ABIL

Andmeteadus (MSc), 2023
Autor: Ida Rahu
Juhendajad: Anneli Kruve (PhD), Meelis Kull (PhD)

Tox21 10K andmestik

#	ÜHEND	12 BIOKEEMILIST RADA				
1	BrC(Br)Br	0	1	NaN	...	0
...	...	...	...	...	...	...
11764	Nc1cc(Cl)ccc1O	1	NaN	0	...	0

1 - aktiivne; 0 - mitteaktiivne

8043 unikaalset endokriinsüsteemi kahjustavat kemikaali

ANDMED MUDELITE TREENIMISEKS
ELULISED TESTANDMED MassBank and MoNA
R pakett "rcdk" sõrmejäljed SMILESidest
SMILES: Nc1cc(Cl)ccc1O
OH
NH2
Cl
sõrmejälgede tunnused
■ = 1 - struktuurielement esineb;
□ = 0 - struktuurielement puudub
SIRIUS + CSI:FingerID sõrmejäljed MSi põhjal
kõrglahutusega massispekter
intensiivsus
m/z
fragmenteerumis-puu
sõrmejälgede tunnused
Klassifitseerimismudelid
sõrmejälgede tunnused
toksilisus 12 biokeemilises rajas
ühe- ja mitmeväljundiga mudelid
Testitud mudelid
sõrmejälgede tunnused'
'tõenäosused
parimad ühe- ja mitmeväljundiga mudelid
toksilisus 12 biokeemilises rajas
TEGELIK KLASS
ENNUSTATAV KLASS
TP
VP
VN
TN
VP määr 90% saagise juures
KASUTUSALA
?
LC/ESI/HRMS
SIRIUS + CSI:FINGERID
Looduslik proov (100+ ühendit)
Ei vaja edasist testimist
Vajab täiendavat *in vivo* ja/või *in vitro* testimist
kiire
kuluefetiivne
ei põhine loomkatsetel
#UniTartuCS
TARTU ÜLIKOOL
arvutiteaduse instituut
1632

Võtmesõnad:

Kõrglahutus-massispektromeetria, keemilised sõrmejäljed, endokriinsüsteemi kahjustavad ained, Tox21, mitme ülesandega õpe

CERCS: P176 – Tehisintellekt, P300 – Analüütiline keemia, P305 – Keskkonnakeemia, B740 – Farmakoloogia, farmakognoosia, farmaatsia, toksikoloogia

Table of Contents

1 Introduction

Over the last decade, the importance of human exposomics has been rising. Exposomics is a study that investigates how all the exposures of individuals during their lifetime affect their health. [1] As a result of these studies, it has become clear that most real-world samples, such as wastewater and food, contain hundreds to thousands of chemicals, with endocrine-disrupting chemicals (EDCs) posing a particularly serious threat to human health [2]. Meanwhile, the advent of non-target liquid chromatography high-resolution mass spectrometry (LC/HRMS) has made it possible to detect thousands of chemicals in real-world mixtures [1,3,4]. However, due to many limitations, gaps still remain in the downstream evaluation of their toxic effects.

Firstly, although thousands of chemicals, more precisely molecular features, are detected during an LC/HRMS analysis, a tiny fraction (up to 2%) of them are typically identified [4]. For example, in the analysis of household dust, around 5000 molecular features (2000 and 3000 in negative and positive ionisation mode, respectively) were detected, but only 33 were distinctly identified [1]. Secondly, the toxicity information about identified chemicals is incomplete or entirely unavailable. For instance, the PubChemLite database [5], which is a "shortlist" of 400,000 chemicals that humans are most likely to be exposed to through food, agriculture, pharmaceuticals, *etc*., only contains comparable toxicity data for around 3000 of these chemicals for mice in the EPA CompTox database [6]. (For other species, the data is even more scarce.)

Alternative approaches that comprise both *in vitro* and *in silico* methods have been developed to address these problems. Proposed techniques offer the possibility to efficiently generate toxicity information for many chemicals simultaneously without the need for animal testing, which is one of the main shortcomings of classical toxicity testing methods. One such approach is the measurement of bioassay endpoints [7], which has been suggested as an *in vitro* method for obtaining experimental data rapidly. The potential of this methodology is illustrated by the EPA ToxCast project [8]. During this project, the endpoint values for around 10,000 chemicals (Tox21 10K library) were observed and are readily available for toxicity evaluation.

The increase in available toxicity data is complemented by the advances in the field of *in silico* toxicology. *In silico* methods, including read-across [9], structural alerts [10], quantitative structure-activity relationship (QSAR) [11], and other machine learning [12,13] models, have made it feasible to screen a large number of compounds efficiently and are therefore valuable tools, for prioritising the chemicals, that need further testing or fill the data gaps for untested chemicals. Unfortunately, these methods are limited by the requirement for the known structure of the compound under investigation as input, which is a crucial drawback while analysing real-world samples. However, there is a promising workaround.

The molecule's toxicity relates to specific structural patterns, so-called toxicophores. For example, a phenolic functional group is found to be associated with oestrogenic and androgenic endocrine-disrupting activity [14]. While analysing the chemical mixtures with LC/HRMS, that kind of structural information about unidentified substances can be obtained by utilising the capabilities of tools such as SIRIUS+CSI:FingerID [15], which maps the structural patterns present in spectra to molecular fingerprint features. Thus, this study aimed to investigate the potential of HRMS data in predicting the toxicities of compounds without the need for their identification by using the data from the Tox21 10K library.

The following hypotheses were formulated:

- The fingerprint features computable by SIRIUS+CSI:FingerID from HRMS are characteristic enough (representing meaningful toxicophores) to predict the toxicity of compounds with different machine learning algorithms.
- The real-life HRMS data (together with SIRIUS+CSI:FingerID) can be employed with sufficient accuracy as an input of the trained models.
- Since biological pathways are often correlated, multitask learning may be beneficial in applications where several toxicity endpoints for the same compound are predicted.

2 Abbreviations

AHR	aryl hydrocarbon receptor
AR	androgen receptor
ARE	antioxidant response element
ATAD5	ATPase family AAA domain-containing protein 5
DNN	deep neural network
ED	endocrine disruptor
EPA	US Environmental Protection Agency
ER	oestrogen receptor
ESI	electrospray ionisation
FDA	US Food and Drug Administration
FPR	false positive rate
HRMS	high-resolution mass spectrometry
HSE	heat shock element
HTS	high-throughput screening
*k*NN	*k*-nearest neighbours
LBD	ligand binding domain
LC	liquid chromatography
MMP	mitochondrial membrane potential
MOA	mode of action
MS	mass spectrometer/spectrometry
NB	naïve Bayes
NCATS	National Center for Advancing Translational Sciences
NTS	non-target screening
p53	tumor protein p53
PBDE	polybrominated diphenyl ether
PPARg	peroxisome proliferator-activated receptor gamma
(Q)SAR	(quantitative) structure-activity relationship
RF	random forest
ROC-AUC	area under the receiver operating characteristic curve
ROSE	random over-sampling
SA	structural alert
SHAP	Shapley Additive exPlanations
SMARTS	SMILES arbitrary target specification

SMILES	simplified molecular-input line-entry system
SMOTE	synthetic minority over-sampling
SR	stress response
SVM	support vector machine
TPR	true positive rate/recall
t-SNE	t-distributed stochastic neighbour embedding

3 Background

3.1 Endocrine disruptors

In recent decades the chemicals called endocrine disruptors (ED) have globally become the focus of risk assessment and management plans [16,17]. Based on the definition phrased by R.T. Zoeller *et al.*, the ED is an exogenous chemical or mixture of chemicals that interferes with any aspect of hormone action [18].

The endocrine system, also called the hormone system, is one of the most important regulatory systems in the body, along with the central nervous system. It controls several physiological processes, including growth and development, metabolism, reproductive functions, immunity, and the capability to deal with light cycles, temperature fluctuations and other stressors by releasing hormones. [19]

Hormones are organic molecules with diverse structures (proteins and peptides, steroids, catecholamines, eicosanoids, *etc*.) that act as chemical "messengers". They are primarily produced in glands (hypothalamus, pituitary gland, pineal gland, thyroid gland, parathyroid gland, thymus gland, adrenal glands, pancreas, ovary, and testis) and act via binding to nuclear receptors and cell membrane receptors. Due to this high-affinity interaction, they can significantly impact the body's processes even at low concentrations. Based on the receptor type, there are two main mechanisms of how hormones exert their effects. Steroid and thyroid hormones that bind to the nuclear receptors regulate the gene expression (slow response), and peptide and amine hormones that bind to cell membrane receptors activate the different signalling pathways, such as cyclic adenosine monophosphate (cAMP), guanosine 3,5-cyclic monophosphate (cGMP), and Ca^{2+} pathways (rapid response). [19,20]

Endocrine disruptors can mimic or block the actions of hormones or interfere with their synthesis, transport, or metabolism through receptor-mediated and non-receptor-mediated mechanisms [21,22]. In the next chapter, the main concepts of these mechanisms are discussed.

3.1.1 Mechanisms of endocrine disruptors

Recently, the ten key characteristics [21] through which endocrine disruptors can modify the normal functioning of the endocrine system. Of the numerous mechanisms, the most well-known and studied are the receptor-mediated mechanisms, where the main mode of action (MOA) involves EDs acting directly as ligands for hormone receptors. In these mechanisms, EDs can either function as agonists or antagonists.

Agonists are substances that bind to receptors and activate them, leading to a response similar to that of natural ligands. For instance, dichlorodiphenyltrichloroethane (DDT), an insecticide, can function as an agonist by activating both the nuclear oestrogen receptor (ER) [23,24] and the cell membrane receptor human follitropin receptor (FSHR) [25]. ER is a transcription factor that has a crucial role in the development of the mammary gland, among other functions. Thus, the binding of ED to ER (α or β subtype) alters the gene expression, which leads to potential adverse effects such as the increased risk of breast cancer. [26] The binding to FSHR, which is a G protein-coupled receptor (GPCR), on the other hand, triggers the signalling cascade, where cAMP serves as a secondary messenger. Upon ED interaction with FSHR, intracellular cAMP concentration rises, stimulating protein kinase A (PKA), which then modifies the activity of target proteins by phosphorylating them. [27]

Antagonists are structural analogues of natural ligands that bind to receptors but do not activate them, thereby blocking or inhibiting the effect of natural ligands. It is important to note that even though EDs can act as antagonists for nuclear and cell membrane receptors, the research focuses mainly on the first category because many of them tend to be promising drugs. For example, flutamide is a nonsteroidal antiandrogen (NSAA) used to treat androgen-dependent severe health conditions, primarily prostate cancer but also other conditions such as acne. As an antagonist to the androgen receptor (AR), after binding, it blocks the activity of natural androgens, *e.g.* testosterone and dihydrotestosterone. However, it has been found that this compound may be hepatotoxic and cause other side effects, like a decrease in libido. [28]

Besides activating or antagonising the hormone receptors, EDs can also alter their expression level. Such up- or downregulation of specific receptors can impact the amount of hormone needed to initiate a response and/or the strength of the response. For instance, exposure to bisphenol A (BPA), which is widely used in plastic production, has been linked to altering ER expression in the heart (decreasing ERα and increasing ERβ), leading to heart dysfunction and cardiovascular diseases [29]. In addition, EDs can also modify signal transduction, a process where extracellular signal triggers a series of intracellular events that ends with the ultimate response. An illustration of this is the study of fungicide tolylfluanid (TF). The study found that TF reduces a concentration of insulin receptor substrate 1 (IRS-1) and, thereby, the protein kinase B (PKB) activation, which is like other protein kinases, is a key element in regulating the properties of proteins. [30]

The non-receptor-mediated mechanism can be divided into two groups: MOAs directly related to the hormones and MOAs where EDs exert their effect via transgenerational epigenetic inheritance. Due to the wide variety of hormones with distinct characteristics, the former group encompasses many different mechanisms that involve the effects of EDs on hormone biosynthesis, transport, and metabolism, all having a commonality of EDs impacting hormone concentration and availability. [21]

The biosynthetic pathways of hormones are determined by their structure (see Appendix I). Because of the multitude of reaction mechanisms employed, EDs have numerous opportunities to interfere with hormone production. In steroidogenesis, several enzymes are required for catalysing the reactions: cytochrome P450 family enzymes (CYPs) for hydroxylations, hydroxysteroid dehydrogenases (HSDs) for dehydrogenation and steroid reductases for reduction reactions. All of these enzymes can be targets for endocrine disruption, which alters the hormonal balance in the body. For instance, research has shown that parabens, commonly used as preservatives in cosmetics, inhibit the 17β-HSD enzyme [31], while agricultural insecticides, neonicotinoids, hinder the aromatase (CYP19) expression and catalytic activity [32]. These examples also illustrate the EDs capability to modify the metabolism of the hormones.

In order to pass on information that they carry, hormones have to reach their target tissue. Based on the way they travel to their receptors, hormones can be classified as endocrine (hormones that travel in the bloodstream to distant tissues; distant-signalling), paracrine (hormones that, after the secretion to extracellular space, diffuse to neighbouring cells; adjacent-signalling) and autocrine (hormones that are synthesised in their target cells; self-signalling) hormones. Solubility is the major factor that defines how hormones circulate in the body or pass through the cell membranes. Hydrophilic hormones that are easily dissolved in the blood travel freely in the bloodstream. However, they are unable to diffuse through the phospholipid bilayer. Lipophilic hormones, on the other hand, can passively move through the cell membranes but need special protein carriers for transportation throughout the body. [27]

This diversity of transportation mechanisms makes it possible for EDs to have a wide range of MOAs to exert their effects. For example, imidazolines, a class of organic compounds derived from imidazoles that hold promising potential in the pharmaceutical field, have been linked to influencing insulin exocytosis [33]. The commonly used plasticisers phthalates can interfere with steroid hormones transportation by interacting with sex hormone-binding globulin (SHBG) that are responsible for carrying them [34].

Finally, research has demonstrated that exposure to EDs can also result in epigenetic changes. These modifications include DNA methylation, regulation of non-coding RNA expression, and histone modifications that ultimately affect gene expression and occur via mechanisms that either target the epigenetic machinery globally (for instance, through changes in the levels or activity of epigenetic regulators [35]) or at specific gene loci [36].

In conclusion, EDs, commonly present in many everyday items, have a complex and far-reaching impact on biological systems. The discussed mechanisms highlight how diverse the impact of EDs can be on the endocrine system and the wide range of physiological processes they can therefore disrupt, leading to various negative outcomes. Thus, it is essential to continue researching and addressing the consequences of EDs in order to mitigate their effects on human health and the environment.

3.2 Toxicity testing

In order to determine whether a compound possesses any harmful effects on living organisms, a relevant toxicity testing methodology is needed. The first approved methods were based on animal testing (*in vivo* testing) and utilised lethal dose/concentration for 50% of the tested population (LD_{50}/LC_{50}). However, increasing concerns about the ethics and reliability of using animals in research have led to the development of more cost-effective and less time-consuming *in vitro* and *in silico* techniques. [37,38]

In vitro methods use isolated biological matter, such as cells, tissues, and organs, as model systems to assess chemicals' toxicity and shed light on their MOAs. Even though *in vitro* methods are much faster and cheaper compared to *in vivo* testing, it is still not feasible to measure the properties of all the chemicals under every set of conditions experimentally. Thus, to fill the gaps in the data, the use of computational methods has proliferated. [38]

In silico approaches are mainly applied for the preliminary screening of chemicals, for instance, in drug development or in time-critical tasks. They aim to help identify the compounds that may possess any risk and thereby prioritise the substances that need further testing. The fundamental principle of *in silico* toxicology is that the biological activity of the compound is the function of its chemical structure and, therefore, its properties. [38]

In silico methods can be divided into two main categories: expert systems that use predefined rules based on human reasoning to make predictions and learning systems, where predictions are made automatically using conventional statistical analysis or machine learning techniques. These two subclasses include several types of computational approaches, such as structural alerts (SAs), (quantitative) structure-activity relationship ((Q)SAR) modelling and read-across analysis. [39]

The structural alerts technique relies on searching the structural patterns, so-called toxicophores, which are known to be associated with specific types of toxic effects, assuming that the presence and absence of SAs can explain the compound's overall toxicity [39]. For instance, the (poly)brominated diphenyl ether ((P)BDE) group, commonly present in many flame retardant structures, is linked to their antagonising properties in oestrogenic and androgenic receptor-binding assays and can be used as SA in the evaluation of endocrine-disrupting activity [10,40].

Like in other approaches, the central assumption in the read-across analysis is that compounds with similar structures have similar biological activity. In this technique, the endpoint values for the target compounds are estimated by leveraging the relevant endpoint data of their closest analogue(s). [39]

QSAR methods are mathematical models that use structural information expressed as different molecular descriptors to output the compound's activity in a particular biochemical assay. Molecular descriptors, both experimental (*e.g.* partition coefficient) and theoretical (*e.g.* molecular formula), are often categorised based on their dimensionality. For example, 0D descriptors are molecular weights, counts of atoms and bonds. The structural fragment counts and molecular fingerprints are representatives of 1D descriptors. Molecular fingerprints are key components in most cheminformatics applications. Even though several types of fingerprints exist, most of them can be described as binary vectors, where a value of "1" indicates the presence of a specific structural feature and a "0" absence of it (see Figure 1). The examples of 2D and 3D descriptors are different graph representations of molecules and weighted holistic invariant molecular (WHIM) descriptors, respectively. Depending on the field of study, descriptors with higher dimensionality are also used.[41]

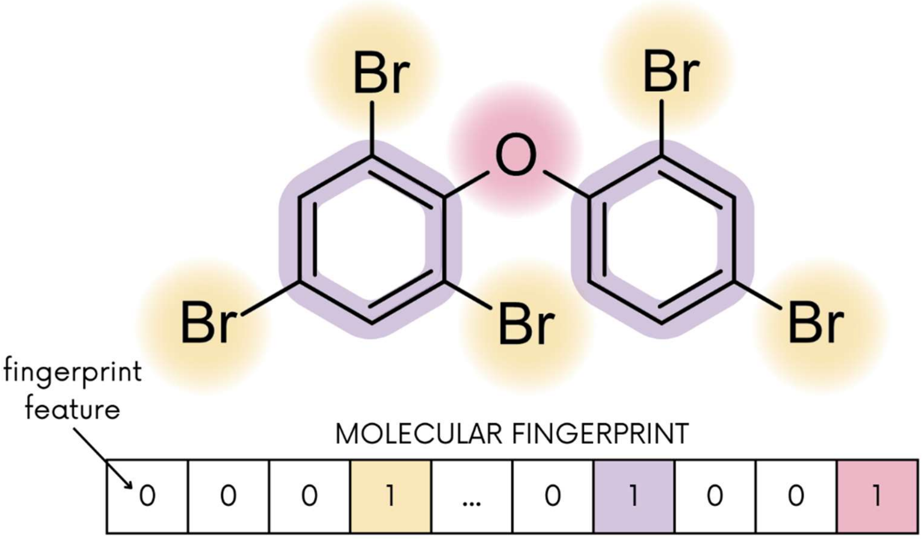


Figure 1 Example of theoretical molecular fingerprint of 2,2',4,4',6-pentabromodiphenyl ether (PBDE-100) that is shown to have one of the highest estrogenic potency among PBDEs [42]

In QSAR modelling, a wide range of algorithms has been employed. Alongside the historically used simple and easily interpretable multiple linear regression (MLP), several other more sophisticated machine learning methods are utilised nowadays, such as support vector machines (SVMs), *k*-nearest neighbours (*k*NN), partial least squares (PLS), random forest (RF), boosting and artificial neural networks (ANNs). [43]

It is important to note that even though the field of *in silico* toxicology is gaining popularity and applying its tools is seen as common practice in future risk assessment procedures, some limitations still inhibit their development. One of the most crucial hindrances is the lack of comprehensive and reliable data to train and validate the models. Different programs and initiatives have been established to promote generating and sharing high-quality toxicity data to overcome this problem. Among these initiatives, by far the most significant and well-known is Tox21 [44].

3.3 Toxicology in the 21st Century and Tox21 Data Challenge 2014

Toxicology in the 21st Century, also known as Tox21, is a US federal research collaboration between the National Institute of Environmental Health Sciences (NIEHS)/the National Toxicology Program (NTP), the National Center for Advancing Translational Sciences (NCATS), the US Food and Drug Administration (FDA), and the US Environmental Protection Agency (EPA) (the National Center for Computational Toxicology (CCTE)) launched in 2008. It is an innovative program developed to leverage the latest advances in high-throughput screening (HTS) technologies to assess the potential risk of chemicals to human health and the environment. [37,44]

Tox21 is revolutionising the field of toxicity testing by developing novel *in vitro* and *in silico* methods to gain a deeper understanding of the mechanisms by which different chemicals affect the organisms and the biological responses they may elicit, as well as to help to select a feasible number of compounds that should go through comprehensive testing. These methods address major issues associated with traditional toxicity testing, including ethical concerns with animal testing, limited resources (including money and time), and the complications associated with clinical trials. [37,44]

In 2014, the NCATS announced the Tox21 Data Challenge[1], soliciting the support of scientists from around the globe to achieve their goals. They invited the researchers to build computational models that could predict the chemical activity of compounds in 12 different bioassays, including seven nuclear receptor (NR) and five stress response (SR) panel pathways (see Table 1). For training the models, the organisers provided a dataset of ~10,000 licensed drugs and environmental chemicals (Tox21 10K library), which endocrine-disrupting activity had been tested using a quantitative HTS setup. Every compound in this dataset had information about its activity in each assay, expressed with three categories: active, inactive or inconclusive, and its structure in either the simplified molecular-input line-entry system (SMILES) or structure-data file (SDF) format. [45]

In machine learning approaches, SMILES [46] notation is the most commonly used representation of chemicals' structures. It utilises ASCII (American Standard Code for Information Interchange) characters to encode the compound's structural information (atoms and bonds) as a single-line string. For example, the SMILES of PBDE-100, shown in Figure 1, is "C1=CC(=C(C=C1Br)Br)OC2=C(C=C(C=C2Br)Br)Br".

The 40 teams from 18 countries employed an extensive array of strategies and tools to construct the most accurate predictive models, incorporating a diverse selection of molecular fingerprints along with other chemical descriptors as inputs, as well as utilising various machine-learning algorithms, such as random forest, deep neural networks (DNN), support vector machines, and *k*-nearest neighbours. The winning teams commonly used multiple descriptor types, applied feature selection to select the most relevant descriptors, employed multiple modelling algorithms, and applied consensus models to make the final predictions. Additionally, the Grand Challenge winner [13] (the best model predicting all the 12 assays) used external data from literature and public databases like PubChem[2] and ChEMBL[3] to improve their predictions. Depending on the biochemical assay, the balanced accuracy of the winning models ranged from 0.550 ("nr.er.lbd") to 0.904 ("sr.mmp") and

[1] https://tripod.nih.gov/tox21/challenge/

[2] https://pubchem.ncbi.nlm.nih.gov/

[3] https://www.ebi.ac.uk/chembl/

the area under the receiver operating characteristic curve (ROC-AUC) from 0.810 ("nr.er") to 0.950 ("sr.mmp").[45]

Table 1 Twelve different toxicity assays used in the Tox21 dataset

Panel	Dataset	Toxicity pathway	Abbreviation
nuclear receptor	activators of aryl hydrocarbon receptor	aryl hydrocarbon receptor (AHR) (full receptor) agonism in HepG2 cells	nr.ahr
	activators of androgen receptor	androgen receptor (AR) (full receptor) agonism in MDA-kb-2 cells	nr.ar
	activators of androgen receptor ligand binding domain	AR (partial receptor) agonism in Hek293 cells	nr.ar.lbd
	aromatase inhibitors	inhibition of aromatase in MCF-7 cells	nr.aromatase
	oestrogen receptor activators	oestrogen receptor (ER) alpha (full receptor) agonism in BG1 cells	nr.er
	activators of oestrogen receptor ligand binding domain	ER alpha (partial receptor) agonism in Hek293 cells	nr.er.lbd
	activators of peroxisome proliferator-activated receptor gamma	peroxisome proliferator-activated receptor gamma (PPARg) (partial receptor) agonism in Hek293 cells	nr.ppar.gamma
stress response	activators of antioxidant response element	antioxidant response element (ARE) agonism in HepG2 cells	sr.are
	activators of heat shock response signalling pathway	heat shock response (HSR) signalling pathway activation in HSE-bla (beta-lactamase reporter gene under the control of heat shock response elements) HeLa cells	sr.hse
	ATPase family AAA domain-containing protein 5	induced stabilisation of the ATAD5 protein in Hek293 cells	sr.atad5
	disruptors of mitochondrial membrane potential	mitochondria membrane potential in HepG2 cells	sr.mmp
	activators of p53 signaling pathway	induced stabilisation in HCT-116 cells	sr.p53

3.4 Liquid-chromatography high-resolution mass spectrometry

Real-world samples (wastewater, blood, food, *etc.*) contain hundreds to thousands of chemicals. Therefore, to evaluate the toxicity of each component, it is necessary first to separate the individual chemicals from the complex mixture and then identify and quantify them accurately. In recent years, non-target liquid chromatography high-resolution mass spectrometry (LC/HRMS) has become a widely used method for this purpose [1,3,4].

LC is a separation technique where the mixture's components are separated based on their polarity. Commonly, LC is coupled to MS to detect the separated chemicals. In MS, the chemicals from the LC are converted to gas-phase ions, and for this, electrospray ionisation (ESI) in both positive and negative ionisation modes is widely used. [47]

After ionisation, the resulting ions are directed into a mass analyser, where they are separated based on their mass-to-charge ratio (*m*/*z*) in an electric and/or magnetic field. In applications where MS is utilised for identifying unknown compounds, so-called non-target screening (NTS), instruments with high-resolution mass analysers, such as time-of-flight (ToF) and orbitrap, are usually used. In the final stage of MS, the number of ions with specific *m*/*z* values is recorded to generate a mass spectrum. [47]

Some applications, like analysing complex mixtures, require more information than the *m*/*z* of the detected chemical. Therefore, a special two-step technique, called tandem mass spectrometry or MS^2, is developed that combines multiple mass analysers. In the first stage of MS^2, ions generated during ionisation are separated based on their *m*/*z* values. After that, the interesting ions with predetermined *m*/*z* are isolated from the rest of the ions and are fragmented further (see Figure 2). The weaker bonds in the ions are broken during the fragmentation, and characteristic fragments are produced. The higher the collision energy, the more different fragments are generated. The resulting fragments are separated based on their *m*/*z*, and their detection produces the MS^2 spectrum. [48]

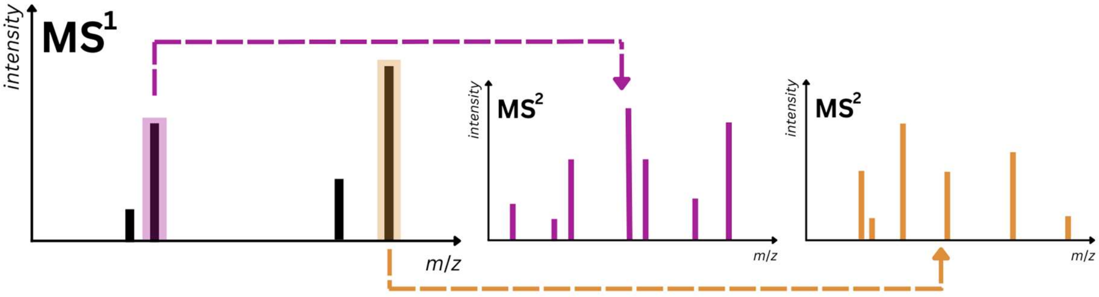


Figure 2 Schematic representation of tandem mass spectrometry. In the first stage, all the ions generated during the ionisation are separated based on their *m*/*z* values (MS^1 spectrum is generated). Based on that, interesting precursor ions of specific *m*/*z* are selected (highlighted with purple and yellow) and fragmented further to produce the MS^2 spectra.

In order to utilise the data obtained from the LC/HRMS analysis for identifying unknown chemicals, a multi-step analysis procedure needs to be conducted. In the first step of this identification procedure, the appropriate molecular formulas are allocated to unknown compounds based on the registered *m*/*z* values (in the MS^1 spectrum). One widely-used algorithm [49] utilises seven heuristic rules, commonly called the "seven golden rules", which are applied to calculate scores for each possible molecular formula. Next, formulas with the highest scores are searched in the chemistry databases for potential structures and mass spectra. The unknown compounds' fragmentation spectra (MS^2) are then measured and compared with literature spectral data. If the patterns of the two spectra match, the compound's identity can be confirmed. However, due to the fact that compounds may have several structural isomers (compounds with the same molecular formula but different structures), additional information (*e.g.* retention time, collision cross-section, and preliminary knowledge about the sample composition) is often required. [50]

Besides the need for complementary data, the main limiting factor of this process is the lack of fragmentation spectra in the literature, which often makes it impossible to achieve a high confidence level in identification. Schymanski's five-level system [4] is frequently employed to describe different confidence levels. In this system, level 1 is the ideal situation, where the structure of the compound under investigation is confirmed, and level 5 is the case where *m*/*z* is known, but due to lack of information, it is impossible even to assign a molecular formula to the compound. In terms of non-target LC/HRMS analysis, this means that while the technique can detect thousands of chemicals simultaneously, only a small fraction (up to 2%) of them can be unambiguously identified. [4]

Tools such as SIRIUS CSI:FingerID are used to complement non-target LC/HRMS analysis to improve identification confidence. SIRIUS is a software that utilises the MS^2 spectra to compute the fragmentation trees (*i*.*e*. annotate the fragmentation spectra with chemical formulas) by employing combinatorial optimisation. Fragmentation trees are graph-based data structures that shed light on the fragmentation pathways of the compounds under investigation. Every node in these trees represents the fragment's chemical formula, and edges describe the losses between the precursor ions (larger fragments) and product ions (smaller fragments). In order to increase the accuracy of this process, the MS^1 spectra are additionally required (for example, to flag the presence of halogens). [15,51]

The CSI:FingerID uses the constructed fragmentation trees to predict molecular fingerprints. It employs the linear support vector machine algorithm to calculate the probability of the presence or absence of each molecular property (structural pattern) for an unknown compound. The calculated fingerprints include CDK Substructe, ECFP6, Klekota-Roth, MACCS, Open Babel FP3, PubChem CACTVS, ring systems, and custom-made SMARTS (SMILES arbitrary target specification) fingerprints. However, it is important to emphasise that not entire fingerprints are outputted, but only those molecular properties, so-called fingerprint features, that were found to have reasonable prediction quality when the SVMs were trained. If the HRMS measured in the positive ionisation mode is given as an input, the process yields 3878 fingerprint features, and when the negative mode data is provided, the number of produced features is 4072. Between those fingerprint feature subsets, there are 3494 overlapping molecular properties. [15,51] CSI:FingerID offers a valuable feature in the case where the entire structure of the compound is unknown because it allows for predicting its characteristics and MOAs, and therefore, it may help to overcome the problem of the lack of structural information that impedes EDs discovery.

4 Data and methods

4.1 Data and its preprocessing

4.1.1 Toxicity data

In order to train machine learning models that can classify compounds as toxic and non-toxic based on structural information, toxicity data is needed. The present study utilised the dataset that was provided as a training set in the Tox21 Data Challenge. The raw dataset was downloaded as twelve files in ".smiles" format, each containing information about one specific toxicity assay, including compounds' structures in SMILES format, their NCATS identifiers and activity data in the respective assay obtained *in vitro*. Combining these datasets resulted in a single dataset with 11764 instances. Among these, 5090 instances represented compounds whose SMILES occurred only once, and the remaining 6674 instances corresponded to compounds with multiple occurrences in the datasets. However, it is important to note that some experimental results about the same compound were inconsistent.

Therefore, the following rules were applied (for reasoning, see Chapter 5.4.1)

1) if any of the duplicate rows have a value of "1" for a particular assay, the compound is classified as active in that assay in the combined dataset;
2) if at least one of the duplicated rows has a value of "0" for a particular assay while the others have missing values, the compound is classified as inactive in that assay in the combined dataset;
3) if all the duplicated rows have a value of "NA" for a particular assay, the compound is marked as inconclusive for that assay in the combined dataset.

After processing the whole combined dataset by following these principles, a total of 8043 unique compounds remained.

Another important data-cleaning step in the Tox21 data preprocessing pipeline was removing the compounds unsuitable for LC/ESI/HRMS analysis or, if possible, modifying their structures to make them compatible. The Tox21 dataset contained 1600 compounds that had disconnected structures (SMILES notation contains the character "."; see Figure 3): containing ionic bonds (*i.e.* were classified as salts) or coordinating bonds (*i.e.* coordination complexes). For these chemicals, the non-toxic cations and anions (such as Na^+, K^+, Ca^{2+}, nitrate and acetate ions) and solvent molecules (*e.g.* H_2O, ethanol) were removed if possible. Furthermore, the remaining ions were neutralised by taking into account the valence of the atoms (the number of bonds they can form). In both of these tasks, the functions from open-source cheminformatics and machine learning software RDKit [52] were utilised in Python. The ions and solvent molecules were removed by using module "*rdkit.Chem.SaltRemover*", and the remaining ions were neutralised with the function "*neutralize_atoms()*", which algorithm was written by Noel O'Boyle [53] and adapted to RDKit by Vincent Scalfani. In the current work, it was assumed that the overall biochemical activity of the compound would not be affected by the elimination of the non-toxic ion. Therefore, if it was unclear which ion was potentially toxic, the compound was discarded.

All the compounds were also evaluated individually to exclude those unsuitable for mass spectrometric analysis (*e.g.* indium arsenide). As a part of the deduplication, all the SMILES of the compounds were standardised via the function "*neutralize_atoms()*", and the obtained "*Molecule class*" objects were converted back to SMILES. Finally, 7483 unique chemicals (in this work called as original dataset) remained for training and testing the models (see Figure 5**;** the exact proportions of the active, inactive and inconclusive compounds per bioassay are given in Appendix II).

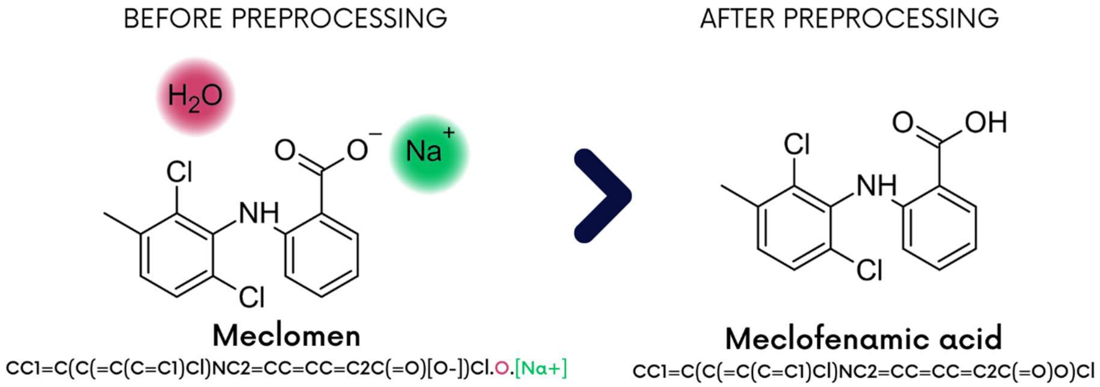

BEFORE PREPROCESSING
AFTER PREPROCESSING
H2O
O
O−
Na+
Cl
NH
Cl
Meclomen
CC1=C(C(=C(C=C1)Cl)NC2=CC=CC=C2C(=O)[O-])Cl.O.[Na+]
O
OH
Cl
NH
Cl
Meclofenamic acid
CC1=C(C(=C(C=C1)Cl)NC2=CC=CC=C2C(=O)O)Cl

Figure 3 Example of modifying the SMILES notation of the compound that has a disconnected structure. The original structure and SMILES are shown on the left, and on the right, the structure and its SMILES representation obtained after removing non-toxic ions and solvent molecules are given.

4.1.2 Mass spectrometric data

Since this study aimed to build machine learning models that use LC/ESI/HRMS data to predict toxicity, the experimental high-resolution mass spectra of compounds were needed to evaluate the final models. Two databases widely acknowledged in the field of analytical chemistry, MassBank (version 2022.06a) [54] and MassBank of North America (MoNA) [55], were used to derive the spectrometric data.

First, the high-resolution tandem mass spectra (MS^2 data) were extracted from MassBank and matched with a cleaned Tox21 dataset based on standardised SMILES, yielding a subset of over 1000 chemicals with MS^2 and toxicity data. 748 (10% of all the compounds) compounds from this subset were selected to form the so-called real-life test set that would represent the overall chemical space of the dataset well. In this study, using random sampling for that purpose was impossible because data was highly imbalanced in each bioassay and contained up to 26% missing values depending on the assay. (Figure 4 shows the number of non-missing values per compound across all the bioassays.) Instead, the fractions of active, inactive and inconclusive compounds in each bioassay in the Tox21 dataset were calculated, and the dataset was repeatedly sampled until the obtained subset resembled these proportions.

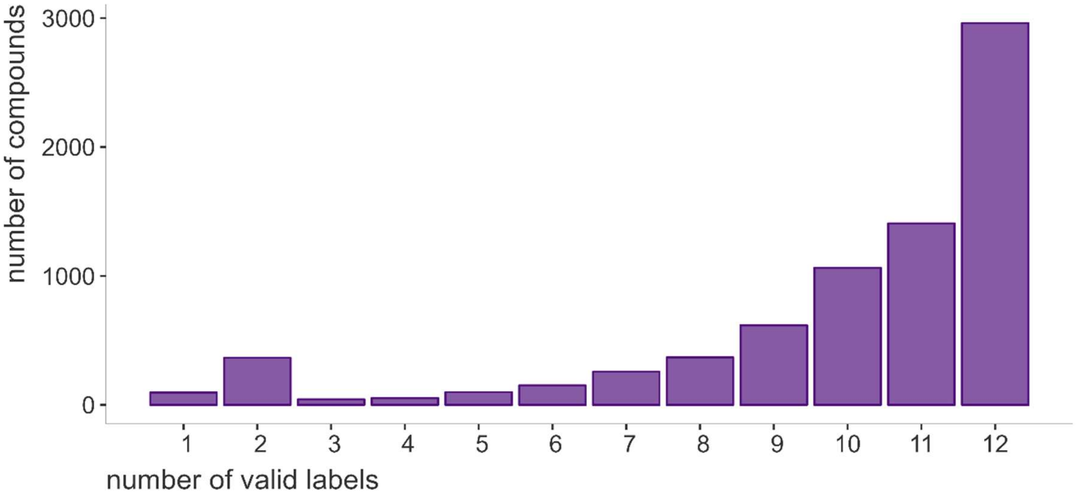


Figure 4 Number of valid labels ("active" or "inactive") per compound in the original dataset. 2961 compounds (around 40%) had labels for all the bioassays, *i.e.* around 60% of the compounds had at least one missing label ("inconclusive").

The real-life test set was used only for the final evaluation of the chosen models, while the remaining 6735 compounds were employed for training the models and for intermediate testing (train/test set).

4.1.3 Data preparation for models training and intermediate testing

A sufficiently large and representative training set consisting of high-quality data is the fundamental building block for developing precise machine learning models. Due to the lack of experimental high-resolution mass spectra of all compounds in Tox21, the usage of probabilistic molecular fingerprints calculated with SIRIUS+CSI:FingerID (see Chapter 3.4) is hindered. Thus, for each compound in the previously defined train/test set (in a total of 6735 compounds), the exact fingerprint features were calculated from SMILES using the R package "*rcdk*"[56] and the full set of SMARTS (structural patterns) that corresponded to the molecular properties computed by SIRIUS+CSI:FingerID.

In order to build classification models that are applicable regardless of which ionisation mode (positive or negative) was used during the HRMS analysis, after the computation process, only the overlapping features (3494 fingerprint features; see Chapter 3.4) were kept for training the models. Furthermore, the fingerprint features with zero and near-zero variance were discarded by using the function "*nearZeroVar()*" with default parameters (*freqCut*" = 95/5, and "*uniqueCut*" = 10). To remove the highly correlated fingerprint features "*findCorrelation()*" with three different cutoff values (0.7, 0.8 and 0.9) from R package "*caret*" [57] was utilised.

The function "*nearZeroVar()*" helps to identify the features that have very little or no variance by flagging the ones that have either only one unique value (zero variance) or whose characteristics follow the values of two parameters: "*freqCut*" (a cutoff value representing the ratio of the most common value to the second most common value within the feature) and "*uniqueCut*" (a cutoff value representing the percentage of distinct values out of the total number of samples). The latter means that all the features where one value prevails over the others (the ratio of the frequencies of the most common and second most common value is relatively large) and that have very few unique values relative to the number of samples are marked as near-zero variance features.

The function "*findCorrelation()*" detects the features that should be removed in order to ensure that the maximum absolute pairwise correlation between them is less than the cutoff value. First, it determines the two features with the highest absolute pairwise correlation based on the correlation matrix given as input. Then it calculates the average correlations between the selected and all the other features and removes the one with the highest average correlation. The procedure is repeated until no absolute correlations are above the cutoff value.

After performing these preprocessing steps, three datasets were obtained for the same set of compounds, with varying numbers of fingerprint features, namely 247, 340, and 476, depending on the selected cutoff value used to eliminate the highly correlated features.

Prior to training the models, the dataset should be split into the training and test set (in this work, the latter is also called the intermediate test set). However, due to imbalance of the data and missing values (Chapter 4.1.2), the random sampling was inapplicable. Thus, to overcome the limitations and divide the data so that the test set would represent the underlying distribution of the toxicity data, the anticlustering algorithm [58] was used.

This algorithm aims to partition the input data into *K* anticlusters, *i.e.* heterogenous groups that are as similar as possible to each other. It is achieved by maximising a clustering objective function rather than minimising it. The anticlustering algorithm is implemented as the function "*anticlustering()*" in the R package "*anticlust*". To obtain the 80/20 training and test sets, the function parameter *K* was set to 5, meaning that all the compounds were assigned to five anticlusters and one of them was randomly allocated as a test set.

Figure 5 shows the proportions of the active, inactive and inconclusive compounds per toxicity assay in training (a total of 5388 compounds) and intermediate test set (a total of 1347 compounds) (for more details, see Appendix II).

4.1.4 Data preparation for final evaluation of models

In order to use SIRIUS software for calculating the fingerprint features from HRMS data, the spectral data were converted to ".ms" format. It is SIRIUS specific file format, where in addition to mass spectra given as a simple peak list, the meta information, such as ionisation mode, formula, parentmass *etc*., necessary for computations, is also provided. Furthermore, the ".ms" format enables the combination of HRMS data of the same compound, measured using identical experimental conditions but different collision energies. Combining the fragmentation spectra obtained in multiple collision voltages allows SIRIUS to build deeper fragmentation trees since more characteristic fragments are captured, leading to more accurate fingerprint predictions.

For each compound in the real-life test set (748 chemicals, see Chapter 4.1.2), the ".ms" files were generated based on MassBank data. If multiple MS^2 spectra, measured using identical experimental parameters (same ionisation mode and instrument type) but different collision energies, were present for a compound, the information was gathered into one file so that data in each HRMS file (peaks given as *m*/*z* and corresponding intensities) were listed under specific collision voltages. In order to provide the MS^1 data, the isotope patterns were calculated from the chemical formula with the R package "*enviPat*" and function "*isopattern()*" [59] because of the lack of experimental data in mass spectrometry databases.

These files were further used as input in SIRIUS+CSI:FingerID (version 5.6.3) to compute the fingerprint features (given as posterior probability that a specific structural pattern is present in the compound under investigation) explained previously. For 97 compounds, fragmentation spectra appeared to provide insufficient information for generating the characteristic fragmentation trees. Therefore, supplementary spectral data was needed. For that objective, the data from MoNA was used: the HRMS data was queried for each compound with insufficient information, and the ".ms" files were improved. Thanks to this excessive work, the finalised real-life test set, ready for evaluating the trained models, contained information about 734 compounds (for more details, see Appendix II).

The following figure illustrates the proportions of the active, inactive and inconclusive compounds per toxicity assay in the original dataset, training dataset, intermediate test set and real-life test set. Based on this, it can be concluded that the datasets used for training and testing the models represent well the overall chemical space of the original dataset.

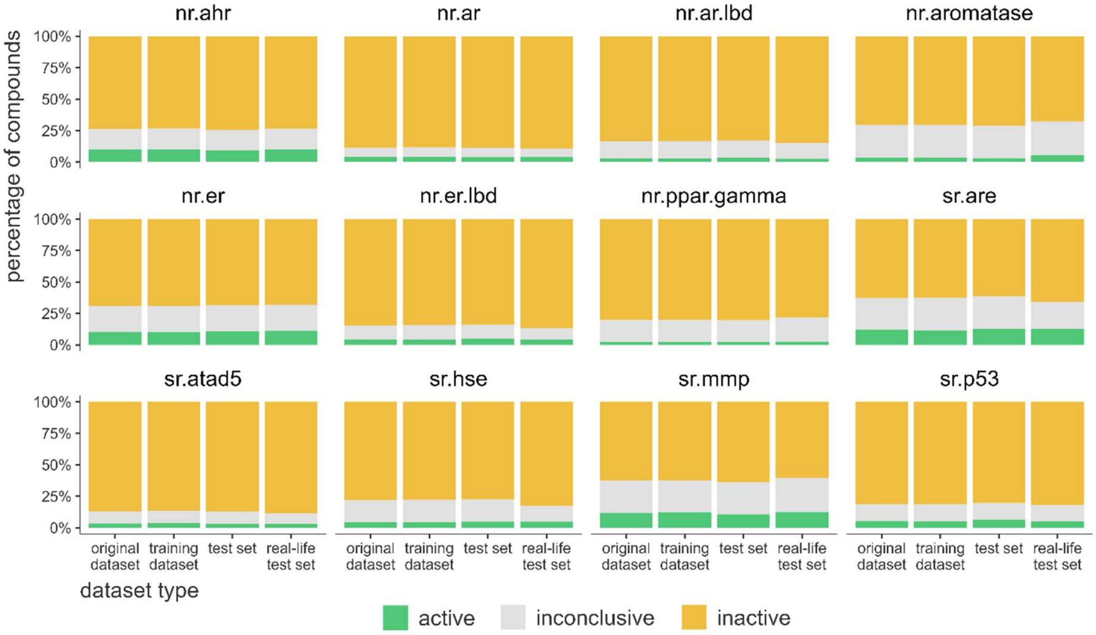


Figure 5 Proportions of the active, inactive and inconclusive compounds per toxicity assay in the original dataset (dataset obtained after deduplication and unsuitable compounds removal from Tox21 data; 7483 compounds), training dataset (used for training the models; 5388 compounds), intermediate test set (used for testing the models and selecting the final models; 1347 compounds), and real-life test set (used for final evaluation of models; 734 compounds). Each subplot illustrates one bioassay.

4.2 Models

In the current study, the machine learning models that use the compounds' binary molecular fingerprint features to predict their endocrine-disrupting activity ("active" or "inactive") in 12 different bioassays were developed. Two concurrent approaches were employed for modelling. The first strategy defined the task as 12 distinct binary classification problems; therefore, the independent classification model was trained for each of the 12 bioassays. In the second approach, however, the task was framed as a multi-label classification problem, and thus, a single model that predicts compound activity in all the assays simultaneously was trained.

The training set, consisting of 5388 chemicals (as described in Chapter 4.1.3), was employed to train the models in both approaches. In the first strategy, the models were trained using three different fingerprint feature sets containing 247, 340, or 476 features, respectively, obtained after removing the highly correlated fingerprint features (cutoff values:

0.7, 0.8, 0.9). While training the multi-output models, only the feature set with 476 molecular properties (cutoff value 0.9) was utilised.

4.2.1 Training the binary classifiers

The trained models comprise a wide variety of machine learning algorithms that are broadly used for classification tasks. Specifically, these models include those based on linear discriminant analysis (LDA), logistic regression (LR), naïve Bayes (NB), *k*-nearest neighbours, support vector machines, and decision tree (DT) algorithms. Since it has been found that aggregating the predictions of a group of models, so-called ensembles, will often improve the accuracy of the prediction compared to the best individual model, the ensemble methods, such as random forest, bagging, and boosting (adaptive boosting, boosted logistic regression, gradient boosting, including stochastic gradient boosting and extreme gradient boosting), were also employed. Additionally, neural network models were trained.

Several implementations of the same algorithm were utilised. For example, besides the classical *k*NN model, an extended version, known as weighted *k*NN [60], was employed. In this realisation of the algorithm, the kernel functions are used to weight the neighbours according to their distances. All the trained models are given in Appendix III.

The R package "*caret*" was used to train the binary classifiers. A grid search approach with 10-fold cross-validation (to build robust models that generalise well on unseen data) was employed to select the optimal set of hyperparameters. The anticlustering algorithm (see Chapter 4.1.3) was utilised to split the data into training and validation sets when the original imbalanced training dataset was used without employing any sampling techniques. The models were assessed using the ROC-AUC, sensitivity, and specificity metrics through the "*twoClassSummary()*" function. However, the ROC-AUC metric was used to select the optimal model.

Sampling methods were employed during the training process to address the imbalanced data issue: depending on the assay, the proportion of active compounds in the training set varied from 2.3% to 12.1%. Four different techniques were considered:

- down-sampling
  (function "*downSample()*" from package "*caret*")
  In the down-sampling approach, a subset of the majority class data points is selected such that the resulting data set has frequencies of the minority and majority classes in close proximity to each other;
- up-sampling
  (function "*upSample()*" from package "*caret*")
  In up-sampling, the minority class data points are randomly sampled with replacement until their number is identical to the amount of majority class data points;
- synthetic minority over-sampling (SMOTE)
  (function "*smote()*" from package "*performanceEstimation*")
  The SMOTE method generates new minority class instances by randomly selecting a minority class data point and synthesising new samples by interpolating the feature values of the selected data point with its *k*-nearest neighbours in the feature space [61];
- random over-sampling (ROSE)
  (function "*ROSE()*" from package "*ROSE*")

The ROSE approach generates new synthetic data points by following the smoothed bootstrap technique [62].

Therefore, each binary classifier was trained for each bioassay on 15 different datasets: three different feature sets (obtained while removing the highly correlated features with different cutoff values) and the original imbalanced dataset together with four balanced datasets (extra datasets obtained by applying the sampling methods).

4.2.2 Training the multi-label classifier

In the current study, deep neural networks (DNNs), which have demonstrated high performance on similar problems [13], were employed for the multi-label classification task. The proposed architectures comprise the DNNs with up to four hidden layers with the rectified linear unit (ReLU) function as an activation function. The number of units in each hidden layer ranged from 512 to 8192. The output layer consisted of 12 sigmoid units: one per task (12 toxicity assays). In order to reduce the issue of vanishing/exploding gradients, the batch normalisation technique was utilised. Simultaneously, dropout, L2 regularisation (Ridge Regression) and early stopping methods were implemented to prevent the DNNs from overfitting.

Due to the missing labels, a regular cross-entropy loss function, broadly used for learning in multi-label classification tasks, was not applicable in the present study. Therefore, its slightly modified version was employed, where the data points with missing labels were discarded while calculating the loss, *i.e.* their loss was fixed to zero.

For optimisation, the Adam (adaptive moment estimation) optimiser was utilised. The hyperparameter-tuning was done using a grid search approach and 10-fold cross-validation (folds were generated utilising an anticlustering algorithm). The optimal set of hyperparameters was selected using the ROC-AUC. Additionally, the learning rate reduction technique was implemented, which reduced the learning rate when the metric did not improve during the number of given epochs. For training the models, "*Keras*" library via TensorFlow for R was used [63]. All the hyperparameters and architectures considered are shown in Table 2. A total of 1080 different settings were tried in multi-output model training.

Table 2 Proposed architectures of DNNs and considered hyperparameters

Considered hyperparameters and architectures	Tried values
number of hidden layers	2, 3, 4
number of hidden units per layer	512, 1024, 2048, 4096, 8192
learning rate	0.01, 0.05, 0.1
learning rate reducing factor	0, 0.1
dropout probability	0, 0.3, 0.5
L2 regularisation penalty	0, 10^{-6}, 10^{-5}, 10^{-4}

4.3 Model selection and final evaluation

The selection of an appropriate metric to evaluate the model relies on the intended application of the model in the future. The current master's thesis aimed to develop machine learning models that are able to pinpoint chemicals that probably have an endocrine-disrupting activity and, therefore, need further examination. In this application, the models should clearly demonstrate a high true positive rate (TPR, also known as recall), meaning they can find as many as possible of the EDs in the sample analysed. On the other hand, due to the problems of classical toxicity testing methods, the number of compounds falsely classified as toxic must be minimal, *i.e.* good models should have a low false positive rate (FPR). Thus, this work uses a false positive rate at 90% of recall as a metric to evaluate models and as a selection criterion (see Figure 6).

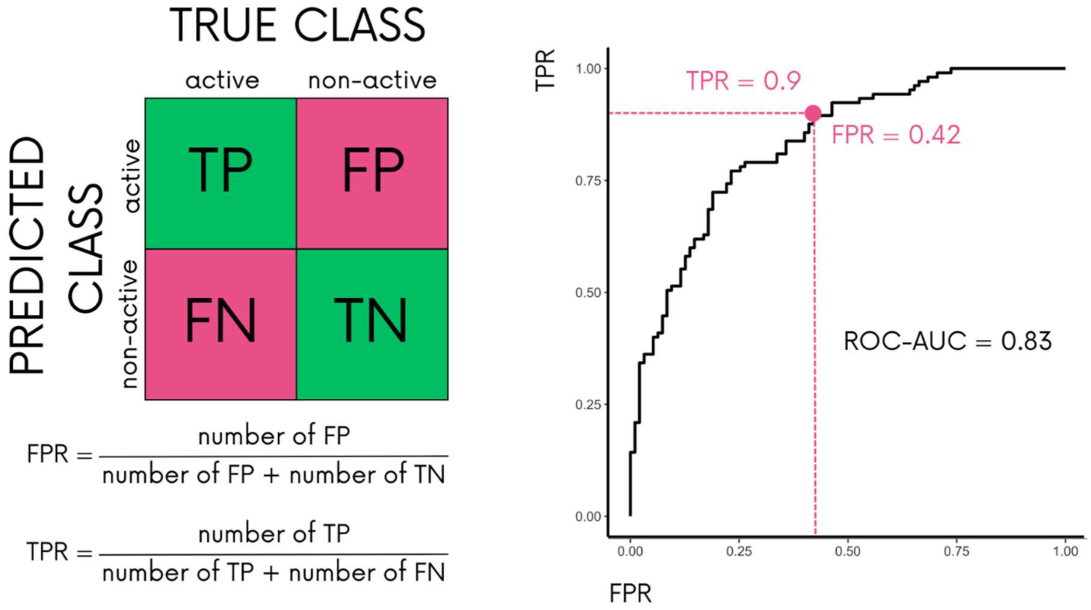


Figure 6 Metrics used for the final evaluation of the models. The left side of the figure shows how the values obtained based on the confusion matrix (TP – true positive, FP – false positive, FN – false negative, TN - true negative) are used for calculating the false positive rate (FPR) and true positive rate (TPR). The ROC graph with an example of determining the FPR at 90% of recall is shown on the right side of the figure. The *x*-axis on this figure represents the FPR and the *y*-axis TPR. Additionally, the value of the area under the curve (ROC-AUC) is given.

After completing the training of the models, each model was evaluated on the intermediate test set. Based on their FPR at 90% of recall, the best models were selected for further evaluation on the real-life test set. Additionally, the balanced accuracy and ROC-AUC values were outputted to enable one to compare the trends in the results with those observed in the Tox21 Data Challenge.

The fingerprint features are given as posterior probabilities instead of binary values in the real-life test set (see Chapter 4.1.4). Thus, an additional preprocessing step is necessary to use them for toxicity predictions with the trained models, during which the probabilities are converted to binary values. In the naive approach, a threshold value could be used for that. For instance, applying a simple threshold value of 0.5: if the predicted probability is equal to or greater than 0.5, the corresponding fingerprint feature would be marked as present ("1"), and if it is less than 0.5, it would be marked as absent ("0"). However, finding a suitable threshold value is a very complex problem. Hence, a more sophisticated strategy was employed. For each compound, every fingerprint feature that is used as input in trained models was sampled 10,000 times using the SIRUS+CSI:FingerID outputted probability (p) that the specific feature should have value "1" and the probability of $1-p$ that it should have a value "0". All 10,000 obtained datasets were utilised to test the models, and their results were averaged to obtain the final prediction. The performance of the models was evaluated by calculating the FPR at 90% of recall.

4.4 Additionally utilised methods and resources

The t-distributed stochastic neighbour embedding (t-SNE) [64] analysis was conducted to explore the potential patterns among the fingerprint features used for training the models. It is a statistical method which is widely used for visualising high-dimensional data in a lower-dimensional (2D or 3D) space. For analysis, the Python module "*scikit-learn*" and the function "*TSNE()*"[65] were utilised.

The SHapley Additive exPlanations (SHAP) [66] technique was utilised to provide insight into a machine learning model's predictions. This commonly used method assists in identifying the features that hold the most significance in the model's prediction process. The analysis was performed by using the R package "*SHAPforxgboost*" [67].

The Nextflow [68] framework was utilised to develop automated workflows that facilitate the training and testing of multiple models concurrently.

During the thesis writing process, Grammarly and ChatGPT were used to improve the quality of the text by helping to rephrase the hardly understandable sentences, suggesting alternative word choices, and correcting the grammar.

The code used in the study can be found at the link: https://github.com/idarahu/MSc_thesis

5 Results and discussions

5.1 Models performance on the intermediate test set

This study aimed to develop machine learning models capable of predicting the endocrine-disrupting activity of compounds using their structural information obtained from HRMS analysis. More precisely, it was hypothesised that the molecular properties derived from SIRIUS+CSI:FingerID are characteristic enough to define compounds' activities ("active" or "inactive") in 12 bioassays related to EDs. Two parallel approaches were employed to test the hypothesis: splitting the task into 12 separate binary classification tasks (*i.e.* training a single-output model for each biochemistry endpoint) or combining all the bioassays into a multi-output classification problem (*i.e.* training a model that simultaneously predicts the values for all the endpoints).

An intermediate test set was utilised to select the final single- and multi-output models based on their false positive rate at 90% of recall. Depending on the biochemical assay, the lowest achieved values of this metric ranged from 0.196 ("nr.ahr") to 0.670 ("nr.er"). The performance of all the trained models is displayed in Figure 7. The single-output models chosen for final evaluation are highlighted in yellow, and the multi-output model in green, and the parameters of these models are given in Appendix IV.

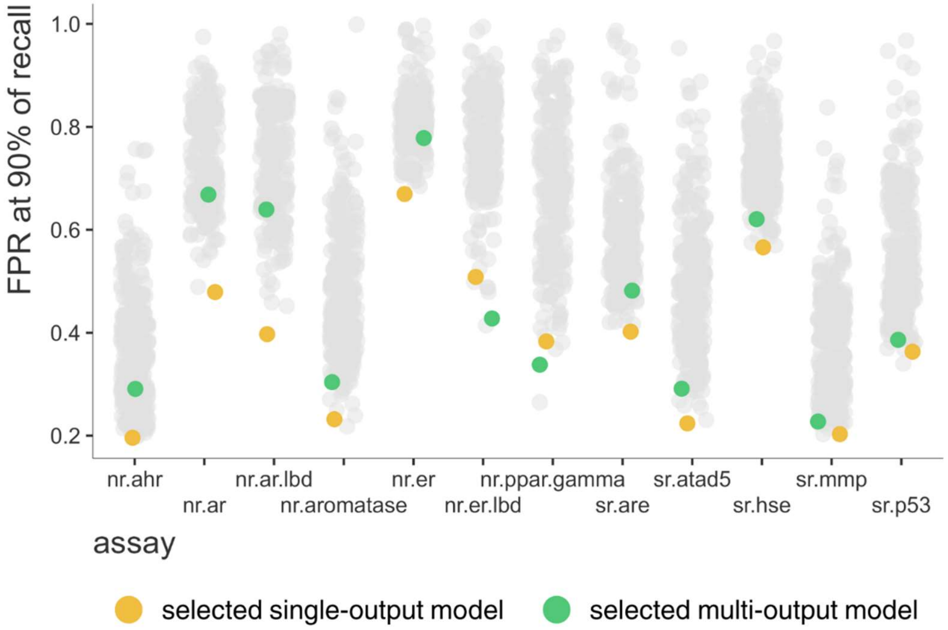


Figure 7 Models' performance on the intermediate test set expressed as FPR at 90% of recall (metric employed to select the models for final evaluation). (Trained models with a metric value of 1 were excluded from the comparison.) The data points highlighted in yellow represent the single-output models selected for final evaluation on the real-life test set, while the data points highlighted in green represent the multi-output model chosen for the same purpose.

As one can see, in most of the bioassays, the selected single-output models had lower FPR at 90% of recall than the selected multi-output model. One of the reasons for that is the difference in selection strategy. Single-output models were selected by identifying the models with the lowest FPR at 90% of recall for specific endocrine-disrupting activity endpoints. In contrast, for choosing the multi-output model, the average FPR at 90% of recall values across all the bioassays were compared, and the model with the lowest value (0.455) was selected. Among all the algorithms used for training the single-output models, only the ones that use ensemble methods, such as random forest and boosting, were represented in the final selection. However, it is important to emphasise that it does not mean that the selected models were statistically better than the others.

Based on the results presented in Figure 7, it becomes apparent that the method proposed in this master's thesis, which utilises the molecular properties that can be computed from HRMS data by SIRIUS+CSI:FingerID, can be employed for modelling compounds' endocrine-disrupting activities with sufficient accuracy. However, some aspects that are also reflected in the figure should be considered beforehand.

5.2 Underlying patterns that affect the performance of the models

Although the measured metric value varied considerably across the trained models, some trends are observable between the bioassays. For example, the models generally performed better on bioassays such as "nr.ahr" and "sr.mmp", where the lowest FPR at 90% of recall among all the models were 0.196 and 0.203, respectively. On the other hand, the "nr.er" endpoint modelling was more challenging, with the best FPR at 90% of recall being 0.670. There can be several reasons for these fluctuations. However, in terms of the applicability of the proposed methodology, it is crucial to determine whether these are solely due to the experimental design. For instance, it could be possible that the fingerprint features used are not representative toxicophores or that valuable information was removed during the data cleaning step. To address this, the trends observed in the current work were compared to those observed for the models' submitted to the Tox21 Data Challenge.

In Tox21 Data Challenge, the balanced accuracy and ROC-AUC values were used to compare the models. Therefore, utilising the intermediate test set, the same metrics were calculated for all the trained models (see Appendix V). However, it is essential to note that a direct numerical comparison is not possible as the test set used here and in Tox21 Data Challenge do not completely overlap.

In Tox21 Data Challenge, the bioassays "nr.ahr" and "sr.mmp" received the models with the best performance, where the highest ROC-AUC scores were greater than 0.9 and average scores above 0.8. Conversely, the lowest average ROC-AUC scores, around 0.7, were reported for "nr.ar" and "nr.ar.lbd" bioassays and the lowest ROC-AUC among the winning models was 0.810, achieved for "nr.er" endpoint prediction. The highest balanced accuracies ranged from 0.650 ("nr.ar.lbd") to 0.904 ("sr.mmp"), with the lowest balanced accuracy across the winning models being 0.550 for "nr.er.lbd" bioassay. [45] These patterns across toxicity assays are in accordance with the trends observed in this study.

In the current work, the highest average ROC-AUC scores were achieved for "nr.ahr" and "sr.mmp" bioassays (0.856 and 0.870, respectively), while the bioassay with the lowest maximum ROC-AUC score (0.768) was "nr.er". The highest balance accuracies ranged from 0.709 ("nr.er") to 0.848 ("nr.ahr"). The most significant difference between the results of the current work and the Tox21 Data Challenge was observed in the "sr.hse" bioassay, which had relatively higher predictive performance in the competition (in the current study, this bioassay had the lowest average ROC-AUC score (0.704)).

Based on this survey, it can be concluded that even though the experimental design definitely impacts the outcome of the models, some assay-specific aspects also play an important role. The Tox21 Data challenge organisers acknowledged that the models developed for assays with higher levels of active compounds tended to perform better. Additionally, they emphasised that although several computational approaches can be employed to handle the data imbalance, their efficacy is restricted by the limited real structural information that can be extracted. [45]

In the case of bioassays, *e.g*. "nr.ahr" and "sr.mmp", where all the models tend to perform generally better, indeed, have a relatively high active compound rate (Figure 8). On the other hand, the difficulty in modelling the "nr.er" endpoint indicates the impact of other aspects.

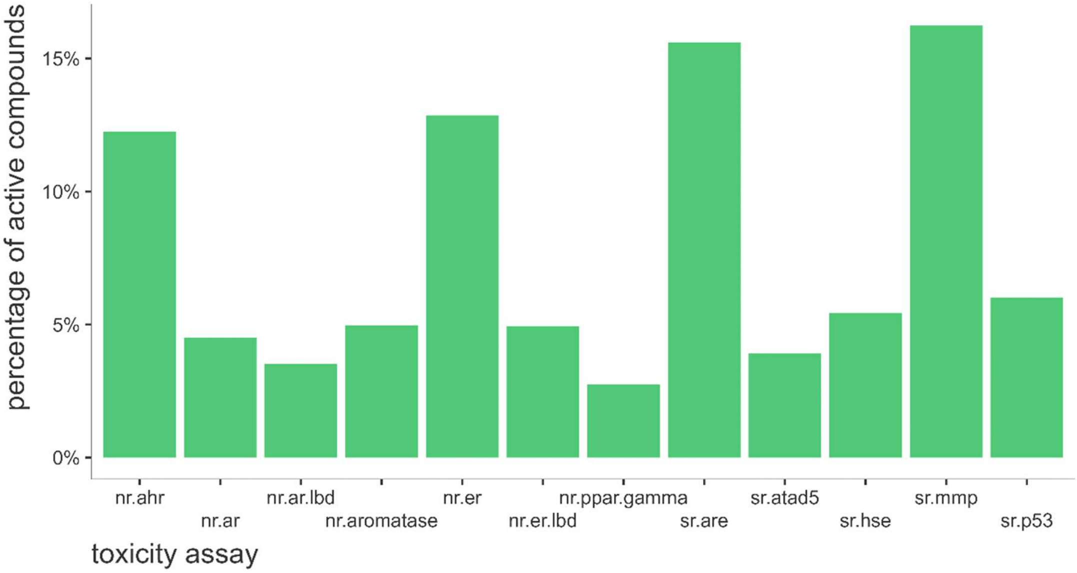


Figure 8 Proportions of the active compounds per bioassay. In the calculations, only the active and inactive compounds were considered (*i.e.* inconclusive compounds were discarded)

One of the key elements that affect the analysis results in the field of *in silico* toxicology is the complexity of the endpoint under investigation. The authors of the Tox21 10K library have brought out that compounds that belong to the same structure classes may have contradicting effects [69]. For instance, flavonoids, which are widely present in plant-based foods, can act as either oestrogen receptor agonists or antagonists [69,70]. This example vividly illustrates the limitations of *in silico* approaches, which rely on the theory that a chemical's structure defines its biochemical properties: if compounds with similar structures have different biological activities, finding the characteristic toxicophores that could explain this variation is very difficult. In order to explore potential patterns among the fingerprint features used for training the models, that could shed light on the quality of the toxicophores captured by these molecular properties, the t-distributed stochastic neighbour embedding (t-SNE) analysis was conducted (see Figure 9).

Although t-SNE has limitations, it provided valuable insights into the analysed data. For example, the active and inactive compounds in the "nr.ahr" assay were predominantly separated on the t-SNE plot, suggesting that the fingerprint features could capture some discriminatory information regarding this assay. In contrast, the compounds in the "nr.er" bioassay were more widely dispersed, indicating a higher degree of complexity and possibly a lower predictive power of the fingerprint features for this endpoint. This may be one of the reasons why models built for the "nr.er" assay showed lower performance metrics compared to those for the "nr.ahr" assay.

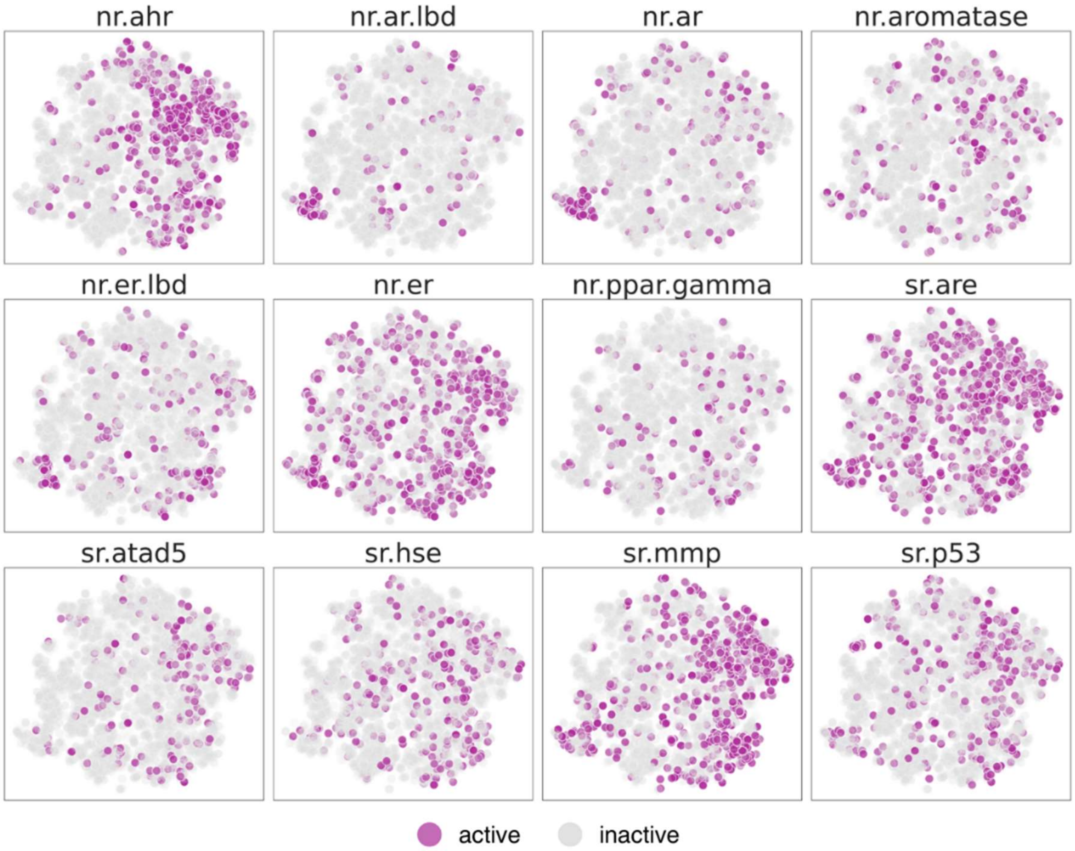


Figure 9 Results of the t-SNE analysis. In every assay, only the active (purple) and inactive (grey) compounds are shown, and each plot uses the same coordination system

In order to determine whether the single-output model selected for final evaluation, which was designed to predict the endocrine-disrupting activity of compounds in the "nr.ahr" assay and had the lowest FPR at 90% recall among all models on the intermediate test set, captured meaningful toxicophores, a Shapley Additive exPlanations (SHAP) analysis was performed. The ten most important variables and their contribution to the models' predictions are shown in Figure 10.

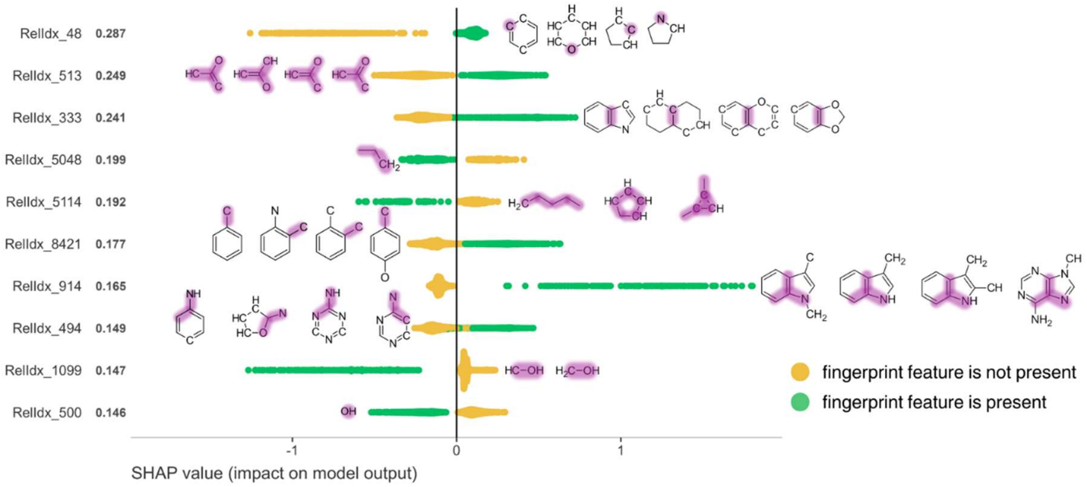


Figure 10 Variable importance analysis for "nr.ahr" model that basis on the extreme gradient boosting algorithm. According to SHAP analysis, the results of the ten most important variables are shown together with their graphical descriptions from SIRIUS+CSI:FingerID (the purple highlight indicates the exact structural patterns that correspond to the feature under investigation). The *x*-axis of the plot represents the directionality of the effect of variables. The colour of the points indicates the absence ("0") or presence ("1") of the respective structural fragment. Fingerprint naming "RelIdx" refers to the absolute index numbering system in SIRIUS+CSI:FingerID.

The fingerprint feature (RelIdx_48) with the highest importance score and whose presence indicates that the compound under investigation is more probably active corresponds to the different ring structures. Although the aromatic ring in the structures of the compounds that are able to bind to aryl hydrocarbon receptors (AHR) is one of the key elements, this fingerprint feature is not very informative since it covers a broad spectrum of chemicals. On the other hand, the molecular property representing the indole moiety (RelIdx_914) is highly relevant because indole-derived chemicals have been shown to act as AHR ligands [71]. The same applies to the fingerprint feature (RelIdx_333) representing annelated rings since polycyclic aromatic hydrocarbons (PAHs) are widely recognized as AHR agonists [72]. Based on these results, it can be concluded that fingerprint features used for training the models contain characteristic toxicophores, at least for some of the bioassays, and machine learning models are able to learn them. This discovery is significant because it provides an opportunity to delve deeper into compounds' toxicological effects and explore how various structural patterns contribute to predicting their endocrine-disrupting activity. Also, it confirms once more that the methodology proposed in this study, which employs machine learning models trained on molecular fingerprints computable from HRMS data to predict compounds' activity in bioassays, is a valuable and promising approach for *in silico* toxicology.

Finally, the correlations between the bioassays were examined to explain the observed trends in the models' performance comparison. The t-SNE plots effectively demonstrated the correlations between certain bioassays, including "nr.ar.lbd" and "nr.ar". Latter is expected since both endpoints are related to the androgen receptor signalling pathway. Additionally, a pairwise correlation matrix was calculated to obtain a more comprehensive understanding of the underlying relationships within the dataset (see Figure 11). This analysis also confirms that bioassays that are designed to measure the compounds' activity towards the same biological targets, such as "nr.ar" and "nr.er", together with their ligand binding domain counterparts, are closely related. However, based on the results, it is also evident that the nature of the relationships could be more complex. For instance, the compounds that act as agonists of antioxidant-responsive element (ARE) also tend to disrupt the mitochondrial membrane potential (MMP). This observation is consistent with the literature [73] demonstrating the direct link between antioxidative stress and mitochondrial membrane potential. The information about the correlations between the assays is highly valuable and helps to understand why the trained models, in the case of some assays, perform similarly and use the analogous fingerprint features for decision-making. Additionally, it illustrates why multi-output models could be more beneficial than single-output models in the field of *in silico* toxicology. By considering multiple bioassays simultaneously, multi-output models can leverage the correlations between the assays and extract more information from the data, potentially leading to more robust and accurate predictions.

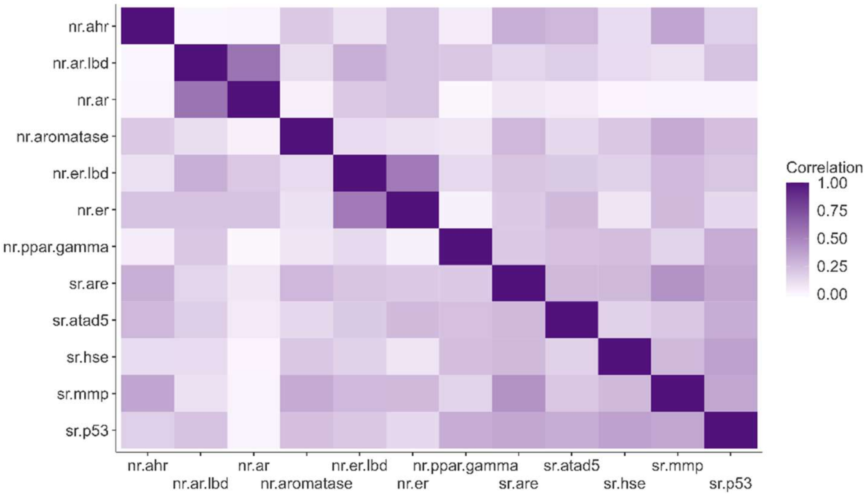


Figure 11 Pairwise correlation matrix for all the biochemical assays

5.3 Models evaluation on the real-life test set

After thoroughly investigating the underlying patterns in the data that may affect the outcome of the models, the selected classifiers were tested on the real-life test set (a separate dataset where all the fingerprint features are calculated from HRMS by SIRIUS+CSI:FingerID). To overcome the limitations related to the fact that SIRIUS+CSI:FingerID outputs the molecular properties as posterior probabilities (p), a sampling strategy that utilises the p to convert the probabilities into binary fingerprint features was used (see Chapter 4.3). This technique basis on the idea that after repeating the procedure where probabilities are converted to binary values, which are used for predicting the toxicity of the compounds enough times, the average of all the predictions becomes constant and reflects the underlying distributions better.

Several sampling iterations were tried during the experiments to determine the optimal value that would result in the models' predictions converging. Based on these experiments, it was found that a sampling strategy involving 10,000 samplings was sufficient for all of the models, as the average predictions remained constant for all compounds even as the number of iterations increased. Figure 12 (left panel) illustrates an example of this technique, where each line represents the predictions made for one compound (a total of 50 compounds are plotted) and shows how the cumulative prediction of the single-output model changes with the increase in iterations. The pink lines in the figure highlight the importance of adequate sampling iterations. For these particular compounds, applying a threshold value of 0.5 to convert the model's prediction into endocrine-disrupting activity would have produced a different result after the first iteration as compared to the 10,000th iteration. For this bioassay ("nr.ahr"), a total of 61 compounds out of 614 would have exhibited similar behaviour as the chemicals shown with pink lines. The same figures for all the models and the table that combines the information about the number of compounds whose activity would be different depending on the sampling iteration are given in Appendix VI.

Additionally, the utilised strategy was compared to a naive approach, in which the fingerprint features outputted by SIRIUS+CSI:FingerID would have been converted to binary values using a fixed cutoff value of 0.5. This approach would mark all probabilities greater than or equal to 0.5 as "1" (fingerprint feature present) and "0" (fingerprint feature missing) for those lower than 0.5. For all the models, the percentage of compounds per each assay was calculated to determine how many compounds predicted to be active or inactive using the utilised strategy would have been predicted to belong to the opposite class if the naive approach had been used. The results of this analysis are displayed in the right panel of Figure 12.

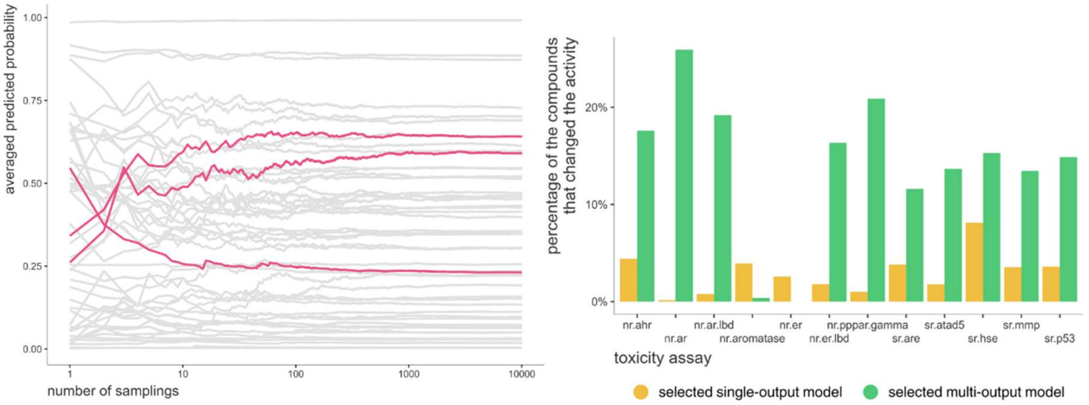


Figure 12 Comparison of the sampling strategy and naive approach for endocrine-disrupting activity prediction using single- and multi-output models. The left panel shows the effect of different sampling iterations on the predictions of the single-output model for 50 compounds in the "nr.ahr" bioassay. Each line represents the predictions made for one compound, and the pink lines indicate the importance of adequate sampling iterations. The right panel displays the percentage of compounds whose endocrine-disrupting activity prediction would have been different if the naive approach had been used instead of the optimal sampling strategy for each bioassay and model.

As one can see, in the case of single-output models, the predictions would have been different for less than 10% of the compounds. However, for the multi-output model, there were significant differences between the two approaches: in "nr.ar" assay, more than 20% of the compounds' endocrine-disrupting activity would have been predicted differently. Nevertheless, this is not surprising while considering the basis of multi-output models. These models are designed to predict multiple properties of a given compound simultaneously, and thus, their predictions depend on the correlation between these properties, *i.e.* they have the ability to learn the hierarchical representations of the data, which can be shared across multiple tasks. Therefore, the input data change can concurrently affect multiple predictions' outcomes. This was one of the reasons why in the present work, the hypothesis was made that using models that are able to predict endpoints of 12 bioassays simultaneously would enable pinpointing new fingerprint features that undergo unnoticed in training individual models due to data sparsity. From all the obtained results (especially in the case of the multi-output model), it can be inferred that the employed sampling strategy for converting the SIRIUS+CSI:FingerID fingerprint features into a usable form for models is effective in achieving consistent and reliable predictions.

Table 3 displays the models' performance on a real-life test set, expressed as FPR at 90% of recall. Upon analysing these results, it can be observed that similar trends to those obtained previously are present. One of the most challenging biochemical endpoints to model is "nr.er" while simultaneously, both the single- and multi-output model tend to perform well on predicting the activity of the compounds in "sr.mmp" bioassay. However, some noteworthy differences exist between the results obtained using the intermediate and real-life test sets. For instance, the selected single-output models tended to have lower FPR at 90% recall than the multi-output model on the intermediate test set (only in the case of two bioassays the multi-output model had a lower metric value). On the real-life test set, both the single- and multi-output models had lower FPR at 90% of recall in half of the bioassays.

Table 3 Models' performance on the real-life test set. The lower FPR at 90% of recall value per bioassay is highlighted in bold.

	FPR at 90% of recall	
Bioassay	**single-output models**	**multi-output model**
nr.ahr	**0.408**	0.430
nr.ar.lbd	**0.688**	0.789
nr.ar	**0.824**	0.904
nr.aromatase	0.520	**0.379**
nr.er.lbd	0.844	**0.576**
nr.er	**0.850**	0.882
nr.pppar.gamma	0.570	**0.537**
sr.are	**0.690**	0.700
sr.atad5	0.856	**0.422**
sr.hse	0.795	**0.754**
sr.mmp	**0.251**	0.339
sr.p53	0.461	**0.254**

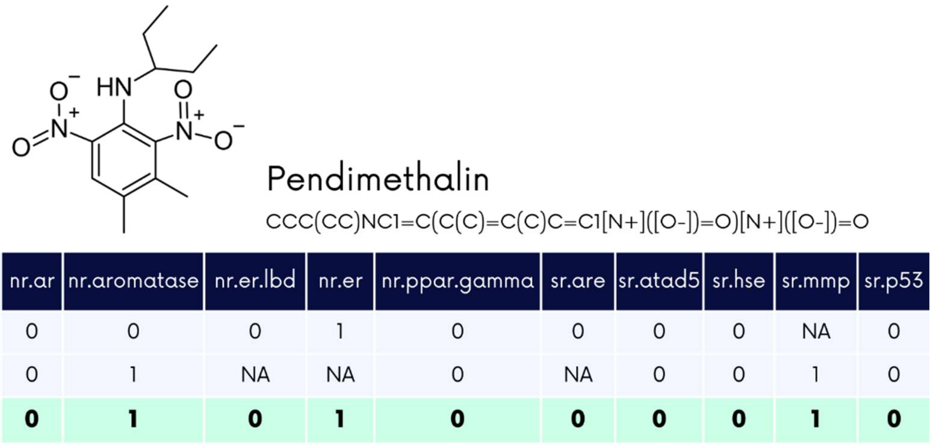
Based on these results, it can be confirmed that it is possible to use the fingerprints calculated from HRMS data by SIRIUS+CSI:FingerID for predicting the compounds' endocrine-disrupting activity as was hypothesised in the present master's thesis. The findings suggest that multi-task learning may be advantageous for such tasks, but more experiments are needed. Further analysis is also necessary to fully understand the impact of model selection and data balancing strategies on overall performance. The following chapter discusses additional considerations and limitations of the proposed methodology.

5.4 Limitations and further considerations of the proposed approach

5.4.1 Quality of the toxicity data and preprocessing strategy

In machine learning, the quantity and quality of the data are crucial: data has to be representative and contain relevant features [72]. The present study utilised the dataset from Tox21 Data Challenge. This widely used dataset in the field of *in silico* toxicology contains qualitative toxicity endpoint measurements in 12 bioassays related to EDs for 8043 unique compounds. Although one of the most comprehensive datasets available, it still possesses many limitations, such as data duplication, missing labels, data imbalance, *etc.*, that should be considered while building the models and interpreting the results.

The original Tox21 dataset included 2953 compounds that were found to occur more than once. This duplication may be caused by the fact that different laboratories tested the same compound or the same compound was repeatedly tested in the same laboratory. However, as noted in Chapter 4.1.1, experimental results regarding the same compound were not always consistent. Figure 13 illustrates one example, where herbicide pendimethalin has two entries (highlighted with light blue) in the Tox21 dataset. Even though these experimental results align with each other in most assays, there is a conflict in the aromatase assay: one experiment suggests that pendimethalin does not exhibit activity in this assay, but another study indicates that the compound is capable of modifying aromatase activity.



nr.ahr	nr.ar.lbd	nr.ar	nr.aromatase	nr.er.lbd	nr.er	nr.ppar.gamma	sr.are	sr.atad5	sr.hse	sr.mmp	sr.p53
0	0	0	0	0	1	0	0	0	0	NA	0
0	0	0	1	NA	NA	0	NA	0	0	1	0
0	**0**	**0**	**1**	**0**	**1**	**0**	**0**	**0**	**0**	**1**	**0**

Figure 13 Example of the duplicated rows in the Tox21 dataset. The endocrine-disrupting activity of the herbicide pendimethalin has been measured twice experimentally (rows that are highlighted with light blue). In most of the bioassays, the measured endpoints are in accordance with each other, except in the case of the aromatase assay. In order to overcome such inconsistencies in the dataset, the deduplication strategy that basis on the three rules (see Chapter 4.1.1) was applied. The endpoint values of pendimethalin obtained while using this approach are highlighted in light green in the third row.

The principles of toxicology (Appendix VII) state that the compound's concentration (dose) plays a key role in determining its toxicity. Therefore, the differences in the experimental data may be caused by the fact that the compounds have been tested on borderline concentrations where small changes in concentrations can lead to different outcomes. Thus, in the present work, in contrast to the usual deduplication strategy, that basis on the "majority vote" idea, the three rules were applied to address the contradicting data points. According to these rules, all the compounds found to have an endocrine-disrupting activity in a specific assay, at least in one replica, were marked to be active in this bioassay in the final dataset. In Figure 13, the values in the third row (highlighted with light green) express the result of employing this approach for deduplicating the data of pendimethalin. As one can see, in the final dataset, the molecule is marked as active in the aromatase assay, which is consistent with the findings reported in the literature [74].

This deduplication technique also served the aim of the master's thesis: developing machine learning models that are able to flag compounds that should go under more thorough *in vitro* or/and *in vivo* testing. In the case of contradicting experimental results, the compound should be tested further to determine its toxic effects unambiguously.

Another important consideration while preprocessing the data for model training is whether it accurately represents the unseen data the models will encounter in practice. In this study, the final models utilise the data of HRMS analysis as input; hence the training data should contain only compounds that are detectable in LC/ESI/HRMS setup. Therefore, all the structures of chemicals were evaluated for suitability and modified or discarded if necessary (see Chapter 4.1.1).

The main part of this preprocessing step was handling the compounds with disconnected structures, such as salts. Salts are dissociated into negatively charged anions and positively charged cations in solutions. Thus, in LC/ESI/HRMS, the complete structures of salts are never registered. For that reason, it was essential to modify their SMILES notations to correspond to real-life scenarios. As mentioned in Chapter 4.1.1, the simplified assumption that the toxicity of salts does not depend on non-toxic ions was made. From the practical point of view, this assumption was vital because it allowed one to use the data that constituted one-fifth of the entire dataset, which would have otherwise been discarded. However, it is necessary to emphasise that since not all of the MOAs of EDs are fully understood, and the toxicodynamics and -kinetics of ions and their neutral counterparts can differ, there is a theoretical possibility that this premise may lead to errors in some cases.

5.4.2 Effect of the usage of SIRIUS+CSI:FingerID on the applicability of models

This study hypothesised that structural information obtained from high-resolution mass spectra could be used to identify unknown compounds requiring additional testing due to potential toxicity concerns. SIRIUS and its integrated tool CSI:FingerID, which maps HRMS data to molecular fingerprint features, are central to this theory. Even though this methodology has some clear advantages, such as allowing one to retrieve information about chemicals' endocrine-disrupting activity without fully identifying them, due to the usage of SIRIUS+CSI:FingerID, it still possesses many limitations.

The first constraints are related to the measured HRMS spectra. The data-rich fragmentation spectra with high mass accuracy are required to build the fragmentation trees, *i.e.* spectra with too few peaks (that correspond to meaningful fragments) cannot be used as an input of SIRIUS+CSI:FingerID. The issue, as observed in the present study when utilising data from MassBank to compute fingerprint features for a real-life test set (see Chapter 4.1.4), can considerably limit the scope of this method since it demands substantial experimental resources to tackle it. Besides the compound's molecular structure, the number of meaningful fragments in MS^2 spectra depends on instrumental parameters, such as collision energy, mass resolution, scanning range and sensitivity. Therefore, the performance in real experiments will depend on the mass spectrometric method used and might differ from the metrics reported here.

A key element of building effective machine learning models is identifying and selecting suitable features for their training. In the field of *in silico* toxicology, determining a sufficient set of structural patterns, or structural alerts, characteristic of toxic compounds is the main challenge. This is especially true when the mechanisms of compounds are very complex and not entirely clear, like in case of EDs (see Chapter 3.1.1)) which strongly affects the applicability of developed methods. In the current master's thesis, all the usable features were predefined by SIRIUS+CSI:FingerID. Thus, the assumption was made that the fingerprint features outputted by this tool are comprehensive enough to describe the toxicity of compounds. However, it should be noted that this premise may not hold for all compounds, as some characteristic toxicophores could be overlooked, which may lead to biased predictions.

Another crucial factor affecting the overall accuracy of the proposed methodology is that the fingerprint features obtained from HRMS data cannot be taken as ground truth. As mentioned, SIRIUS+CSI:FingerID uses the SVM algorithm to predict the molecular properties and provides them as posterior probabilities. Therefore, it is essential to emphasise that even if the posterior probability that the chemical contains the specific structural pattern is 0.99, it does not necessarily mean that this molecular property is truly present. Even when assuming that the accuracy of SIRIUS+CSI:FingerID is as high as 99% for all the fingerprint features, around 34 to 35 out of 3494 overlapping fingerprint features that were used in this work are still expected to be incorrect (Figure 14).

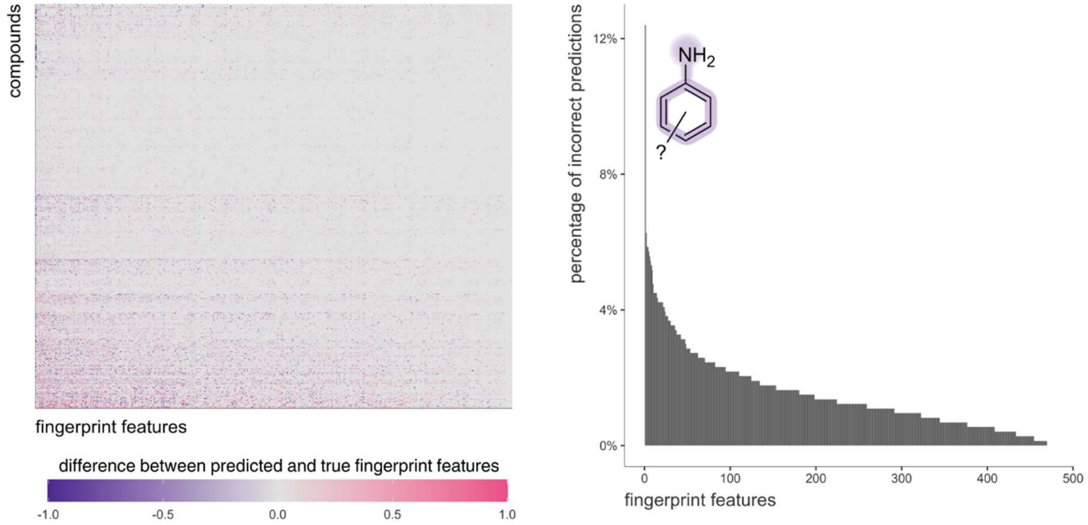


Figure 14 Difference between the predicted fingerprint features, calculated based on HRMS data using SIRIUS+CSI:FingerID ($fpf_{\text{predicted}}$; posterior probabilities) and the true fingerprint features derived from SMILES (fpf_{true}). The figure includes data on 476 molecular properties selected for model training (see Chapter 4.1.3), which are arranged in the same order on both panels. The left panel depicts the difference between $fpf_{\text{predicted}}$ and fpf_{true} for each compound in the real-life test set, calculated as $difference = fpf_{\text{predicted}} - fpf_{\text{true}}$. The right panel illustrates the proportions of compounds for which the predicted and true fingerprint features disagree. For that, a naive approach was used: setting a feature value to "1" if $fpf_{\text{predicted}} \geq 0.5$ and "0" otherwise, and then comparing the results with true values.

As one can see, in most cases, the molecular properties from SIRIUS+CSI:FingerID are in good accordance with the ones calculated from the SMILES notations. The simplified analysis, which considered posterior probabilities greater than or equal to 0.5 as evidence for specific structural patterns, showed that most of the features were mispredicted for less than 4% of the compounds (see Figure 14, right panel). However, the molecular property that represents the benzeneamine group (aniline and its derivatives) demonstrates that concordance is not always as good: for more than 12% of the compounds, its predicted value did not match the true value. This result is especially valuable in the context of the current study, as it highlights the limitations of the developed method. Aniline and its derivatives are reported in the literature as compounds that may influence steroidogenesis [75], making the benzeneamine group a potential characteristic toxicophore. Having said that, it is clear that if SIRIUS+CSI:FingerID is unable to predict its presence with high accuracy, the developed models may misclassify the compounds. This is an important factor that should be considered before employing this approach.

Furthermore, the predicted fingerprint features for certain chemicals deviated from the true ones more frequently (Figure 14). It could result from insufficient meaningful fragments in their MS^2 spectra, inhibiting the building of deep fragmentation trees and thereby affecting the prediction of molecular properties, which again stresses the importance of the quality of experimental data.

6 Conclusions

The current master's thesis aimed to investigate the possibilities of using the information obtained from non-target LC/HRMS analysis to predict the compounds endocrine-disrupting activity in 12 different bioassays without the need for their full identification. It was proposed that leveraging the capabilities of SIRIUS+CSI:FingerID, which is able to translate the compound's structural information that is present in its data-rich HRMS spectra to molecular fingerprint features, it becomes feasible to use machine learning methods to flag the unknown chemicals present in the complex real-world samples that should be going under further *in vitro* and/or *in vivo* testing due to their potential toxicity concerns.

In order to fulfil this objective, the toxicity data from the Tox21 10K library was utilised, and a wide variety of machine learning algorithms were employed to build classification models, including both single- and multi-output models, capable of predicting the experimental toxicity endpoints contained within the dataset.

Based on the results obtained from these experiments, it was found that fingerprint features generated by SIRIUS+CSI:FingerID can be used to determine compounds' bioactivity. Moreover, it was shown that this methodology could enable pinpointing the structural patterns highly characteristic to toxic compounds and thereby help to understand how the compounds under investigation exert their adverse effects on biological systems, shedding light on their mechanisms of action.

Furthermore, the proposed approach was tested under near real-world conditions by utilising the fingerprint features extracted from the experimental HRMS data using SIRIUS+CSI:FingerID as input for trained models. The results of these experiments demonstrate that employing this approach with sufficient accuracy is possible, and thus, it holds great promise for practical applications in the field of *in silico* toxicology. It is also important to notice that one of the key elements of achieving consistent and reliable predictions in this approach relies on the technique developed in the present work for converting the fingerprint features outputted by SIRIUS+CSI:FingerID into true binary fingerprint features.

Finally, the findings of this study suggest that multi-task learning may be advantageous in the field of *in silico* toxicology for predicting values of multiple endpoints for the same compound due to the many underlying correlations between the biochemical pathways; however, further experiments are required to validate this theory. Additionally, a more thorough analysis is required to gain a complete understanding of how the model selection and data balancing strategies employed in the master's thesis affect the overall performance of the proposed methodology.

To conclude, the present study represents a significant step forward in identifying toxic compounds in real-world mixtures without the need for their chemical identification. It provides a strong foundation for further research and the development of new approaches to address the gaps present in the downstream evaluation of their toxic effects.

7 References

[1] J.E. Rager, M.J. Strynar, S. Liang, R.L. McMahen, A.M. Richard, C.M. Grulke, J.F. Wambaugh, K.K. Isaacs, R. Judson, A.J. Williams, J.R. Sobus, Linking high resolution mass spectrometry data with exposure and toxicity forecasts to advance high-throughput environmental monitoring, Environment International. 88 (2016) 269–280. https://doi.org/10.1016/j.envint.2015.12.008.

[2] N. Caporale, M. Leemans, L. Birgersson, P.-L. Germain, C. Cheroni, G. Borbély, E. Engdahl, C. Lindh, R.B. Bressan, F. Cavallo, N.E. Chorev, G.A. D'Agostino, S.M. Pollard, M.T. Rigoli, E. Tenderini, A.L. Tobon, S. Trattaro, F. Troglio, M. Zanella, Å. Bergman, P. Damdimopoulou, M. Jönsson, W. Kiess, E. Kitraki, H. Kiviranta, E. Nånberg, M. Öberg, P. Rantakokko, C. Rudén, O. Söder, C.-G. Bornehag, B. Demeneix, J.-B. Fini, C. Gennings, J. Rüegg, J. Sturve, G. Testa, From cohorts to molecules: Adverse impacts of endocrine disrupting mixtures, Science. 375 (2022) eabe8244. https://doi.org/10.1126/science.abe8244.

[3] B.I. Escher, H.M. Stapleton, E.L. Schymanski, Tracking complex mixtures of chemicals in our changing environment, Science. 367 (2020) 388–392. https://doi.org/10.1126/science.aay6636.

[4] E.L. Schymanski, J. Jeon, R. Gulde, K. Fenner, M. Ruff, H.P. Singer, J. Hollender, Identifying Small Molecules via High Resolution Mass Spectrometry: Communicating Confidence, Environ. Sci. Technol. 48 (2014) 2097–2098. https://doi.org/10.1021/es5002105.

[5] E. Bolton, E. Schymanski, T. Kondic, P. Thiessen, J. Zhang, PubChemLite for Exposomics, (2020). https://doi.org/10.5281/zenodo.4183801.

[6] O. US EPA, CompTox Chemicals Dashboard, (2016). https://www.epa.gov/chemical-research/comptox-chemicals-dashboard (accessed December 20, 2022).

[7] M.T. Martin, D.J. Dix, R.S. Judson, R.J. Kavlock, D.M. Reif, A.M. Richard, D.M. Rotroff, S. Romanov, A. Medvedev, N. Poltoratskaya, M. Gambarian, M. Moeser, S.S. Makarov, K.A. Houck, Impact of environmental chemicals on key transcription regulators and correlation to toxicity end points within EPA's ToxCast program, Chem Res Toxicol. 23 (2010) 578–590. https://doi.org/10.1021/tx900325g.

[8] D.J. Dix, K.A. Houck, M.T. Martin, A.M. Richard, R.W. Setzer, R.J. Kavlock, The ToxCast program for prioritizing toxicity testing of environmental chemicals, Toxicol Sci. 95 (2007) 5–12. https://doi.org/10.1093/toxsci/kfl103.

[9] T.W. Schultz, P. Amcoff, E. Berggren, F. Gautier, M. Klaric, D.J. Knight, C. Mahony, M. Schwarz, A. White, M.T.D. Cronin, A strategy for structuring and reporting a read-across prediction of toxicity, Regul Toxicol Pharmacol. 72 (2015) 586–601. https://doi.org/10.1016/j.yrtph.2015.05.016.

[10] M. Nendza, A. Wenzel, M. Müller, G. Lewin, N. Simetska, F. Stock, J. Arning, Screening for potential endocrine disruptors in fish: evidence from structural alerts and in vitro and in vivo toxicological assays, Environmental Sciences Europe. 28 (2016) 26. https://doi.org/10.1186/s12302-016-0094-5.

[11] S.D. Dimitrov, R. Diderich, T. Sobanski, T.S. Pavlov, G.V. Chankov, A.S. Chapkanov, Y.H. Karakolev, S.G. Temelkov, R.A. Vasilev, K.D. Gerova, C.D. Kuseva, N.D. Todorova, A.M. Mehmed, M. Rasenberg, O.G. Mekenyan, QSAR Toolbox - workflow and major functionalities, SAR QSAR Environ Res. 27 (2016) 203–219. https://doi.org/10.1080/1062936X.2015.1136680.

[12] J. Zhang, U. Norinder, F. Svensson, Deep Learning-Based Conformal Prediction of Toxicity, J. Chem. Inf. Model. 61 (2021) 2648–2657. https://doi.org/10.1021/acs.jcim.1c00208.

[13] A. Mayr, G. Klambauer, T. Unterthiner, S. Hochreiter, DeepTox: Toxicity Prediction using Deep Learning, Frontiers in Environmental Science. 3 (2016). https://www.frontiersin.org/articles/10.3389/fenvs.2015.00080 (accessed October 30, 2022).

[14] S. Terasaka, A. Inoue, M. Tanji, R. Kiyama, Expression profiling of estrogen-responsive genes in breast cancer cells treated with alkylphenols, chlorinated phenols, parabens, or bis- and benzoylphenols for evaluation of estrogenic activity, Toxicol Lett. 163 (2006) 130–141. https://doi.org/10.1016/j.toxlet.2005.10.005.

[15] K. Dührkop, M. Fleischauer, M. Ludwig, A.A. Aksenov, A.V. Melnik, M. Meusel, P.C. Dorrestein, J. Rousu, S. Böcker, SIRIUS 4: a rapid tool for turning tandem mass spectra into metabolite structure information, Nat Methods. 16 (2019) 299–302. https://doi.org/10.1038/s41592-019-0344-8.

[16] COMMUNICATION FROM THE COMMISSION TO THE EUROPEAN PARLIAMENT, THE COUNCIL, THE EUROPEAN ECONOMIC AND SOCIAL COMMITTEE AND THE COMMITTEE OF THE REGIONS Towards a comprehensive European Union framework on endocrine disruptors, 2018. https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1553707706598&uri=CELEX:52018DC0734 (accessed May 8, 2023).

[17] A. Beronius, L.N. Vandenberg, Using systematic reviews for hazard and risk assessment of endocrine disrupting chemicals, Rev Endocr Metab Disord. 16 (2015) 273–287. https://doi.org/10.1007/s11154-016-9334-7.

[18] R.T. Zoeller, T.R. Brown, L.L. Doan, A.C. Gore, N.E. Skakkebaek, A.M. Soto, T.J. Woodruff, F.S. Vom Saal, Endocrine-disrupting chemicals and public health protection: a statement of principles from The Endocrine Society, Endocrinology. 153 (2012) 4097–4110. https://doi.org/10.1210/en.2012-1422.

[19] Introduction; Receptors and Hormone Action, in: The Endocrine System: Systems of the Body Series, 2nd edition, Churchill Livingstone, Edinburgh ; New York, 2010: pp. 1–14; 15–26.

[20] Hormonal Regulation and Integration of Mammalian Metabolism, in: Lehninger Principles of Biochemistry, 8th edition, W.H. Freeman, Austin, 2021: pp. 2940–3086.

[21] M.A. La Merrill, L.N. Vandenberg, M.T. Smith, W. Goodson, P. Browne, H.B. Patisaul, K.Z. Guyton, A. Kortenkamp, V.J. Cogliano, T.J. Woodruff, L. Rieswijk, H. Sone, K.S. Korach, A.C. Gore, L. Zeise, R.T. Zoeller, Consensus on the key characteristics of endocrine-disrupting chemicals as a basis for hazard identification, Nat Rev Endocrinol. 16 (2020) 45–57. https://doi.org/10.1038/s41574-019-0273-8.

[22] Y. Combarnous, T.M.D. Nguyen, Comparative Overview of the Mechanisms of Action of Hormones and Endocrine Disruptor Compounds, Toxics. 7 (2019) 5. https://doi.org/10.3390/toxics7010005.

[23] J. Legler, L.M. Zeinstra, F. Schuitemaker, P.H. Lanser, J. Bogerd, A. Brouwer, A.D. Vethaak, P. De Voogt, A.J. Murk, B. Van der Burg, Comparison of in vivo and in vitro reporter gene assays for short-term screening of estrogenic activity, Environ Sci Technol. 36 (2002) 4410–4415. https://doi.org/10.1021/es010323a.

[24] D.E. Frigo, M.E. Burow, K.A. Mitchell, T.-C. Chiang, J.A. McLachlan, DDT and its metabolites alter gene expression in human uterine cell lines through estrogen receptor-independent mechanisms., Environ Health Perspect. 110 (2002) 1239–1245.

[25] M. Munier, J. Grouleff, L. Gourdin, M. Fauchard, V. Chantreau, D. Henrion, R. Coutant, B. Schiøtt, M. Chabbert, P. Rodien, In Vitro Effects of the Endocrine Disruptor p,p'-DDT on Human Follitropin Receptor, Environ Health Perspect. 124 (2016) 991–999. https://doi.org/10.1289/ehp.1510006.

[26] J.S. Carroll, Mechanisms of oestrogen receptor (ER) gene regulation in breast cancer, Eur J Endocrinol. 175 (2016) R41–R49. https://doi.org/10.1530/EJE-16-0124.

[27] D.L. Nelson, Lehninger Principles of Biochemistry, 8th edition, W.H. Freeman, Austin, 2021.

[28] Zahra Salimi, F. Moradpour, F. Zarei, Z. Rashidi, M.R. Khazaei, S.M. Ahmadi, The Effect of Blockade of Androgen Receptors by Flutamide on Learning and Memory, Synaptic Plasticity and Behavioral Performances: A Review Study, Cell Tiss. Biol. 15 (2021) 337–346. https://doi.org/10.1134/S1990519X21040088.

[29] K.A. Bruno, J.E. Mathews, A.L. Yang, J.A. Frisancho, A.J. Scott, H.D. Greyner, F.A. Molina, M.S. Greenaway, G.M. Cooper, A. Bucek, A.C. Morales-Lara, A.R. Hill, A.A. Mease, D.N. Di Florio, J.M. Sousou, A.C. Coronado, A.R. Stafford, D. Fairweather, BPA Alters Estrogen Receptor Expression in the Heart After Viral Infection Activating Cardiac Mast Cells and T Cells Leading to Perimyocarditis and Fibrosis, Front Endocrinol (Lausanne). 10 (2019) 598. https://doi.org/10.3389/fendo.2019.00598.

[30] R.M. Sargis, B.A. Neel, C.O. Brock, Y. Lin, A.T. Hickey, D.A. Carlton, M.J. Brady, The novel endocrine disruptor tolylfluanid impairs insulin signaling in primary rodent and human adipocytes through a reduction in insulin receptor substrate-1 levels, Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease. 1822 (2012) 952–960. https://doi.org/10.1016/j.bbadis.2012.02.015.

[31] R.T. Engeli, S.R. Rohrer, A. Vuorinen, S. Herdlinger, T. Kaserer, S. Leugger, D. Schuster, A. Odermatt, Interference of Paraben Compounds with Estrogen Metabolism by Inhibition of 17β-Hydroxysteroid Dehydrogenases, International Journal of Molecular Sciences. 18 (2017) 2007. https://doi.org/10.3390/ijms18092007.

[32] -Beaudoin Élyse Caron, R. Viau, J.T. Sanderson, Effects of Neonicotinoid Pesticides on Promoter-Specific Aromatase (CYP19) Expression in Hs578t Breast Cancer Cells and the Role of the VEGF Pathway, Environmental Health Perspectives. 126 (n.d.) 047014. https://doi.org/10.1289/EHP2698.

[33] P. Jakobsen, P. Madsen, H. Andersen, Imidazolines as efficacious glucose-dependent stimulators of insulin secretion, European Journal of Medicinal Chemistry. 38 (2003) 357–362. https://doi.org/10.1016/S0223-5234(03)00041-2.

[34] I.A. Sheikh, R.F. Turki, A.M. Abuzenadah, G.A. Damanhouri, M.A. Beg, Endocrine Disruption: Computational Perspectives on Human Sex Hormone-Binding Globulin and Phthalate Plasticizers, PLOS ONE. 11 (2016) e0151444. https://doi.org/10.1371/journal.pone.0151444.

[35] J. Liu, L. Zhang, L.C. Winterroth, M. Garcia, S. Weiman, J.W. Wong, J.B. Sunwoo, K.C. Nadeau, Epigenetically mediated pathogenic effects of phenanthrene on regulatory T cells, J Toxicol. 2013 (2013) 967029. https://doi.org/10.1155/2013/967029.

[36] M.W. Gordon, F. Yan, X. Zhong, P.B. Mazumder, Z.Y. Xu-Monette, D. Zou, K.H. Young, K.S. Ramos, Y. Li, Regulation of p53-targeting microRNAs by polycyclic aromatic hydrocarbons: Implications in the etiology of multiple myeloma, Mol Carcinog. 54 (2015) 1060–1069. https://doi.org/10.1002/mc.22175.

[37] P. Wexler, ed., QSAR; Risk Assessment, Human Health; ‘Toxic’ and ‘Nontoxic’: Confirming Critical Terminology Concepts and Context for Clear Communication; Toxicity Testing, Alternatives; Toxicity Testing in the 21st Century: Approaches to Implementation; Toxicity Testing, ‘Read Across Analysis’; Toxicology, Toxicology, The History of, in: Encyclopedia of Toxicology, 3rd edition, Academic Press, Amsterdam ; Boston, 2014: pp. 1–9, 158–164, 610–616, 634–637, 673–675, 680–681, 718–720, 731–745.

[38] P. Wexler, ed., Dose–Response Relationship; Endocrine System, Environmental Hormone Disruptors; Ethics: Ethical Issues in Toxicology; High Throughput Screening; In Silico Methods; In Vitro Tests; In Vivo Tests, in: Encyclopedia of Toxicology, 3rd edition, Academic Press, Amsterdam ; Boston, 2014: pp. 224–226, 332–340, 378–380, 498–500, 916–917, 1026–1029, 1101–1102, 1103–1104.

[39] J. Hemmerich, G.F. Ecker, In silico toxicology: From structure–activity relationships towards deep learning and adverse outcome pathways, WIREs Computational Molecular Science. 10 (2020) e1475. https://doi.org/10.1002/wcms.1475.

[40] T. Hamers, J.H. Kamstra, E. Sonneveld, A.J. Murk, M.H.A. Kester, P.L. Andersson, J. Legler, A. Brouwer, In Vitro Profiling of the Endocrine-Disrupting Potency of Brominated Flame Retardants, Toxicological Sciences. 92 (2006) 157–173. https://doi.org/10.1093/toxsci/kfj187.

[41] Background of QSAR and Historical Developments; Chemical Information and Descriptors; Classical QSAR, in: Understanding the Basics of QSAR for Applications in Pharmaceutical Sciences and Risk Assessment, 1st edition, Academic Press, Amsterdam, 2015: pp. 1–46; 47–80; 81–102.

[42] M.A. Siddiqi, R.H. Laessig, K.D. Reed, Polybrominated Diphenyl Ethers (PBDEs): New Pollutants-Old Diseases, Clin Med Res. 1 (2003) 281–290.

[43] Z. Wu, M. Zhu, Y. Kang, E.L.-H. Leung, T. Lei, C. Shen, D. Jiang, Z. Wang, D. Cao, T. Hou, Do we need different machine learning algorithms for QSAR modeling? A comprehensive assessment of 16 machine learning algorithms on 14 QSAR data sets, Briefings in Bioinformatics. 22 (2021) bbaa321. https://doi.org/10.1093/bib/bbaa321.

[44] D. Krewski, D. Acosta, M. Andersen, H. Anderson, J.C. Bailar, K. Boekelheide, R. Brent, G. Charnley, V.G. Cheung, S. Green, K.T. Kelsey, N.I. Kerkvliet, A.A. Li, L. McCray, O. Meyer, R.D. Patterson, W. Pennie, R.A. Scala, G.M. Solomon, M. Stephens, J. Yager, L. Zeise, TOXICITY TESTING IN THE 21ST CENTURY: A VISION AND A STRATEGY, J Toxicol Environ Health B Crit Rev. 13 (2010) 51–138. https://doi.org/10.1080/10937404.2010.483176.

[45] R. Huang, M. Xia, D.-T. Nguyen, T. Zhao, S. Sakamuru, J. Zhao, S.A. Shahane, A. Rossoshek, A. Simeonov, Tox21Challenge to Build Predictive Models of Nuclear Receptor and Stress Response Pathways as Mediated by Exposure to Environmental Chemicals and Drugs, Frontiers in Environmental Science. 3 (2016). https://www.frontiersin.org/articles/10.3389/fenvs.2015.00085 (accessed May 7, 2023).

[46] D. Weininger, SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules, J. Chem. Inf. Comput. Sci. 28 (1988) 31–36. https://doi.org/10.1021/ci00057a005.

[47] Mass Spectrometry; High-Performance Liquid Chromatography, in: Fundamentals of Analytical Chemistry, 10th edition, Cengage Learning, Australia Brazil, 2021: pp. 774–790, 886–909.

[48] E. De Hoffmann, Tandem mass spectrometry: A primer, J. Mass Spectrom. 31 (1996) 129–137. https://doi.org/10.1002/(SICI)1096-9888(199602)31:2<129::AID-JMS305>3.0.CO;2-T.

[49] T. Kind, O. Fiehn, Seven Golden Rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry, BMC Bioinformatics. 8 (2007) 105. https://doi.org/10.1186/1471-2105-8-105.

[50] J. Hollender, E.L. Schymanski, H.P. Singer, P.L. Ferguson, Nontarget Screening with High Resolution Mass Spectrometry in the Environment: Ready to Go?, Environ. Sci. Technol. 51 (2017) 11505–11512. https://doi.org/10.1021/acs.est.7b02184.

[51] SIRIUS | Lehrstuhl Bioinformatik Jena, (n.d.). https://bio.informatik.uni-jena.de/software/sirius/ (accessed May 9, 2023).

[52] G. Landrum, P. Tosco, B. Kelley, Ric, sriniker, gedeck, D. Cosgrove, R. Vianello, NadineSchneider, E. Kawashima, D. N, A. Dalke, G. Jones, B. Cole, M. Swain, S. Turk, AlexanderSavelyev, A. Vaucher, M. Wójcikowski, I. Take, D. Probst, V.F. Scalfani, K. Ujihara, guillaume godin, A. Pahl, F. Berenger, JLVarjo, jasondbiggs, strets123, JP, rdkit/rdkit: 2022_09_5 (Q3 2022) Release, (2023). https://doi.org/10.5281/zenodo.7671152.

[53] Noel O'Blog: No charge - A simple approach to neutralising charged molecules, Noel O'Blog. (2019). https://baoilleach.blogspot.com/2019/12/no-charge-simple-approach-to.html (accessed May 1, 2023).

[54] M. and its contributors, MassBank/MassBank-data: Release version 2022.06, (2022). https://doi.org/10.5281/zenodo.7148841.

[55] MassBank of North America, (n.d.). https://mona.fiehnlab.ucdavis.edu/ (accessed May 1, 2023).

[56] R. Guha, Z. Charlop-Powers, E. Schymanski, rcdk: Interface to the "CDK" Libraries, (2022). https://cran.r-project.org/web/packages/rcdk/index.html (accessed May 10, 2023).

[57] M. Kuhn, Building Predictive Models in R Using the caret Package, Journal of Statistical Software. 28 (2008) 1–26. https://doi.org/10.18637/jss.v028.i05.

[58] M. Papenberg, G.W. Klau, Using anticlustering to partition data sets into equivalent parts, Psychological Methods. 26 (2021) 161–174. https://doi.org/10.1037/met0000301.

[59] M. Loos, C. Gerber, F. Corona, J. Hollender, H. Singer, Accelerated Isotope Fine Structure Calculation Using Pruned Transition Trees, ACS Publications. (2015). https://doi.org/10.1021/acs.analchem.5b00941.

[60] K. Hechenbichler, K. Schliep, Weighted k-Nearest-Neighbor Techniques and Ordinal Classification, in: Universitätsbibliothek der Ludwig-Maximilians-Universität München, 2004. https://doi.org/10.5282/UBM/EPUB.1769.

[61] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, SMOTE: Synthetic Minority Over-sampling Technique, Journal of Artificial Intelligence Research. 16 (2002) 321–357. https://doi.org/10.1613/jair.953.

[62] N. Lunardon, G. Menardi, N. Torelli, ROSE: a Package for Binary Imbalanced Learning, The R Journal. 6 (2014) 79. https://doi.org/10.32614/RJ-2014-008.

[63] TensorFlow for R - Guide to Keras Basics, (n.d.). https://tensorflow.rstudio.com/guides/keras/basics (accessed May 9, 2023).

[64] L. van der Maaten, G. Hinton, Visualizing Data using t-SNE, Journal of Machine Learning Research. 9 (2008) 2579–2605.

[65] sklearn.manifold.TSNE, Scikit-Learn. (n.d.). https://scikit-learn/stable/modules/generated/sklearn.manifold.TSNE.html (accessed May 9, 2023).

[66] S.M. Lundberg, S.-I. Lee, A Unified Approach to Interpreting Model Predictions, in: Advances in Neural Information Processing Systems, Curran Associates, Inc., 2017. https://papers.nips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html (accessed May 9, 2023).

[67] Y. Liu [aut, cre, A. Just [aut, ctb, M. Mayer, SHAPforxgboost: SHAP Plots for "XGBoost," (2021). https://cran.r-project.org/web/packages/SHAPforxgboost/index.html (accessed May 9, 2023).

[68] A DSL for parallel and scalable computational pipelines | Nextflow, (n.d.). https://www.nextflow.io/ (accessed May 10, 2023).

[69] R. Huang, S. Sakamuru, M.T. Martin, D.M. Reif, R.S. Judson, K.A. Houck, W. Casey, J.-H. Hsieh, K.R. Shockley, P. Ceger, J. Fostel, K.L. Witt, W. Tong, D.M. Rotroff, T. Zhao, P. Shinn, A. Simeonov, D.J. Dix, C.P. Austin, R.J. Kavlock, R.R. Tice, M. Xia, Profiling of the Tox21 10K compound library for agonists and antagonists of the estrogen receptor alpha signaling pathway, Sci Rep. 4 (2014) 5664. https://doi.org/10.1038/srep05664.

[70] R. Kiyama, Estrogenic flavonoids and their molecular mechanisms of action, The Journal of Nutritional Biochemistry. 114 (2023) 109250. https://doi.org/10.1016/j.jnutbio.2022.109250.

[71] Z. Dvořák, K. Poulíková, S. Mani, Indole scaffolds as a promising class of the aryl hydrocarbon receptor ligands, Eur J Med Chem. 215 (2021) 113231. https://doi.org/10.1016/j.ejmech.2021.113231.

[72] M.G. Barron, R. Heintz, S.D. Rice, Relative potency of PAHs and heterocycles as aryl hydrocarbon receptor agonists in fish, Marine Environmental Research. 58 (2004) 95–100. https://doi.org/10.1016/j.marenvres.2004.03.001.

[73] T. Satoh, Y. Enokido, H. Aoshima, Y. Uchiyama, H. Hatanaka, Changes in mitochondrial membrane potential during oxidative stress-induced apoptosis in PC12 cells, J Neurosci Res. 50 (1997) 413–420. https://doi.org/10.1002/(SICI)1097-4547(19971101)50:3<413::AID-JNR7>3.0.CO;2-L.

[74] P. Gupta, S.K. Verma, Impacts of herbicide pendimethalin on sex steroid level, plasma vitellogenin concentration and aromatase activity in teleost Clarias batrachus (Linnaeus), Environ Toxicol Pharmacol. 75 (2020) 103324. https://doi.org/10.1016/j.etap.2020.103324.

[75] M.N. Huda Bhuiyan, H. Kang, J.H. Kim, S. Kim, Y. Kho, K. Choi, Endocrine disruption by several aniline derivatives and related mechanisms in a human adrenal H295R cell line and adult male zebrafish, Ecotoxicology and Environmental Safety. 180 (2019) 326–332. https://doi.org/10.1016/j.ecoenv.2019.05.003.

Appendix

I. Biosynthesis of catecholamines and steroid hormones

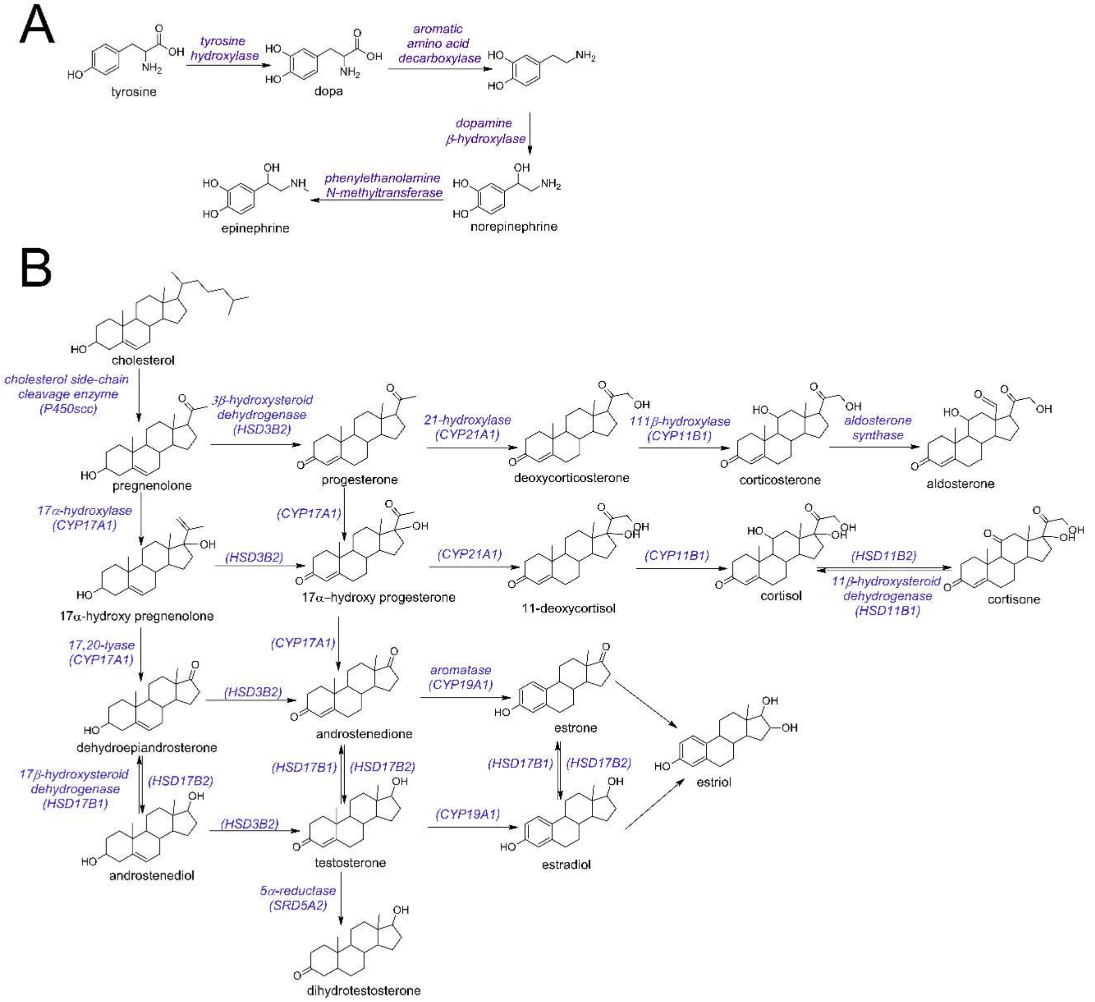


Figure 15 (A) Synthesis of catecholamines (dopamine, epinephrine and norepinephrine synthesis from amino acid tyrosine). (B) Steroidogenesis (steroid hormone synthesis from cholesterol). The names written in black represent the compounds whose structures are shown in the figure; the compounds' names given above and under the arrows (written in purple) represent the enzymes that catalyse the corresponding reactions.

II. Proportions of the compounds in the used datasets

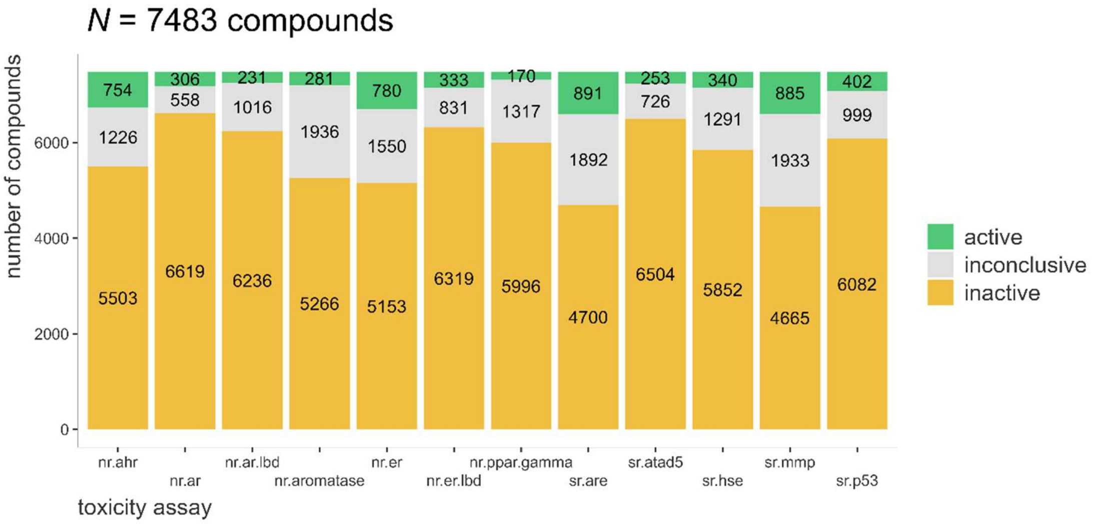


Figure 16 Proportions of the active, inactive and inconclusive compounds per toxicity assay in the original dataset (dataset obtained after deduplication and unsuitable compounds removal from Tox21 data)

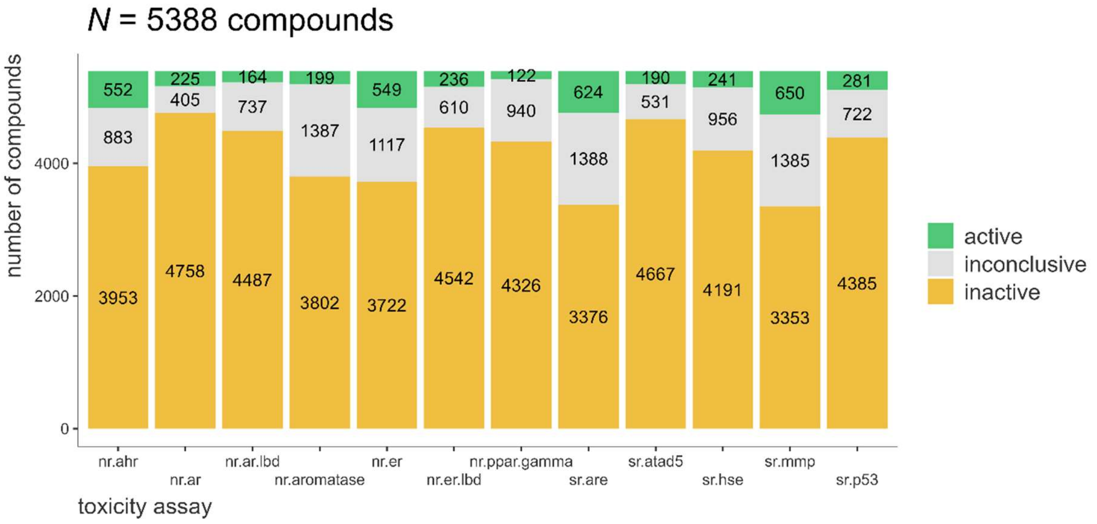


Figure 17 Proportions of the active, inactive and inconclusive compounds per toxicity assay in the training dataset

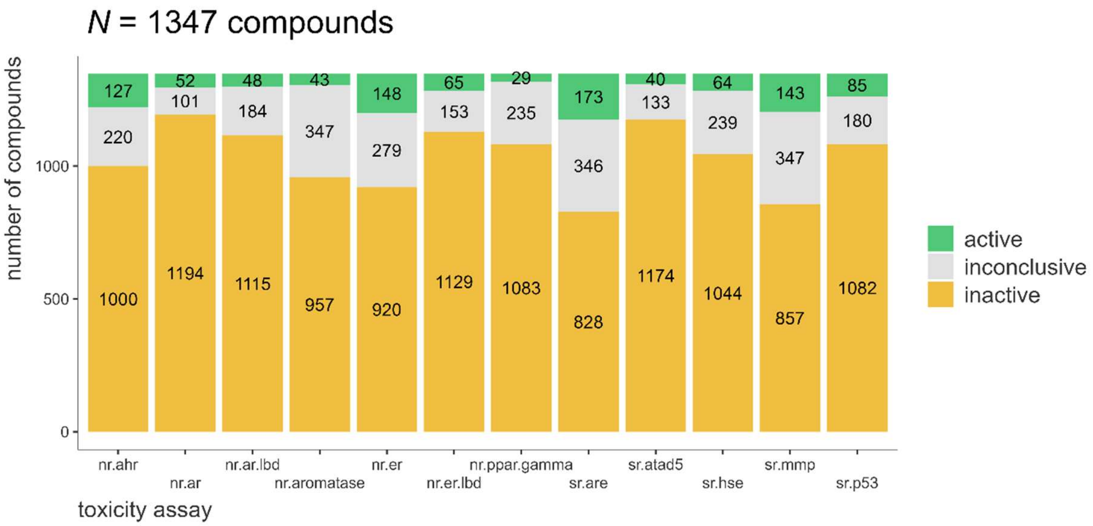


Figure 18 Proportions of the active, inactive and inconclusive compounds per toxicity assay in the intermediate test set

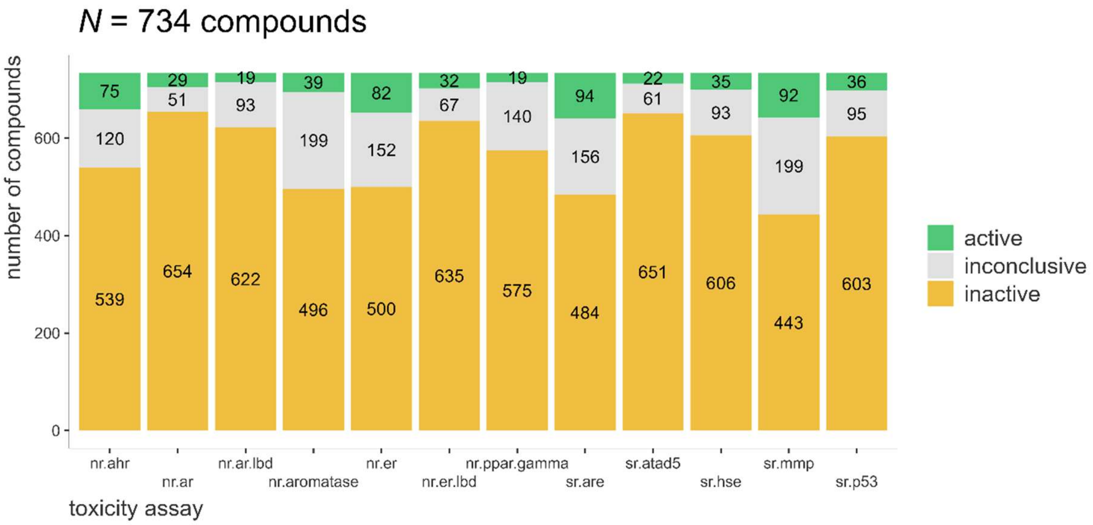


Figure 19 Proportions of the active, inactive and inconclusive compounds per toxicity assay in the real-life test set

III. Trained single-output models

Table 4 Information about all the trained single-output models[4]

Model	Method name	Libraries used	Tuning parameters
Boosted Classification Trees	ada	ada, plyr	iter, maxdepth, nu
Bagged AdaBoost	AdaBag	adabag, plyr	mfinal, maxdepth
Stochastic Gradient Boosting	gbm	gbm, plyr	n.trees, interaction.depth, shrinkage, n.minobsinnode
k-Nearest Neighbors	kknn	kknn	kmax, distance, kernel
k-Nearest Neighbors	knn		k
Linear Discriminant Analysis	lda	MASS	None
Boosted Logistic Regression	LogitBoost	caTools	nIter
Naive Bayes	naive_bayes	naivebayes	laplace, usekernel, adjust
Random Forest	ranger	e1071, ranger, dplyr	mtry, splitrule, min.node.size
Random Forest	Rborist	Rborist	predFixed, minNode
Random Forest	rf	randomForest	mtry
eXtreme Gradient Boosting	xgbDART	xgboost, plyr	nrounds, max_depth, eta, gamma, subsample, colsample_bytree, rate_drop, skip_drop, min_child_weight
Bagged CART	treebag	ipred, plyr, e1071	None
C5.0	C5.0	C50, plyr	trials, model, winnow
Regularised Logistic Regression	regLogistic	LiblineaR	cost, loss, epsilon
eXtreme Gradient Boosting	xgbTree	xgboost, plyr	nrounds, max_depth, eta, gamma, colsample_bytree, min_child_weight, subsample
Linear Support Vector Machines with Class Weights	svmLinearWeights	e1071	cost, weight
Neural Network	nnet	nnet	size, decay
Neural Networks with Feature Extraction	pcaNNet	nnet	size, decay

[4] https://topepo.github.io/caret/available-models.html

IV. Parameters of the selected models

Table 5 Parameters of the selected single-output models

Bioassay	Model	Parameters	Balancing strategy	Cutoff value to remove highly correlated features
nr.ahr	xgbTree	• nrounds = 100 • max_depth = 9 • eta = 0.3 • gamma = 0 • colsample_bytree = 0.6 • min_child_weight = 1 • subsample = 1	down-sampling	0.7
nr.ar.lbd	Rborist	• predFixed = 247 • minNode = 2	SMOTE	0.7
nr.ar	gbm	• n.trees = 50 • interaction.depth =2 • shrinkage = 0.1 • n.minobsinnode = 10	None	0.8
nr.aromatase	gbm	• n.trees = 200 • interaction.depth = 3 • shrinkage = 0.1 • n.minobsinnode = 10	up-sampling	0.7
nr.er.lbd	rf	• mtry = 476	None	0.9
nr.er	rf	• mtry = 239	up-sampling	0.9
nr.pppar.gamma	ranger	• mtry =124 • splitrule = 'extratrees' • min.node.size =1	SMOTE	0.7
sr.are	xgbDART	• nrounds = 150 • max_depth = 9 • eta = 0.3 • gamma =0 • subsample = 1 • colsample_bytree = 0.6 • rate_drop = 0.01 • skip_drop = 0.05 • min_child_weight = 1	down-sampling	0.8
sr.atad5	gbm	• n.trees = 150 • interaction.depth = 3 • shrinkage = 0.1 • n.minobsinnode = 10	up-sampling	0.8
sr.hse	gbm	• n.trees = 150 • interaction.depth = 3 • shrinkage = 0.1 • n.minobsinnode = 10	ROSE	0.7
sr.mmp	rf	• mtry = 171	up-sampling	0.8
sr.p53	gbm	• n.trees = 200 • interaction.depth = 3 • shrinkage = 0.1 • n.minobsinnode = 10	down-sampling	0.8

The architecture and hyperparameters of the selected multi-output model:

- layers = 3
- units in the first layer = 4096
- units in the second layer = 2048
- units in the third layer = 1024
- learning rate = 0.05
- learning rate reducing factor = 0.1
- dropout probability (for each layer) = 0.5
- L2 regularisation penalty = 10^{-6}

V. Performance of the trained models on the intermediate test set expressed as ROC-AUC and balanced accuracy

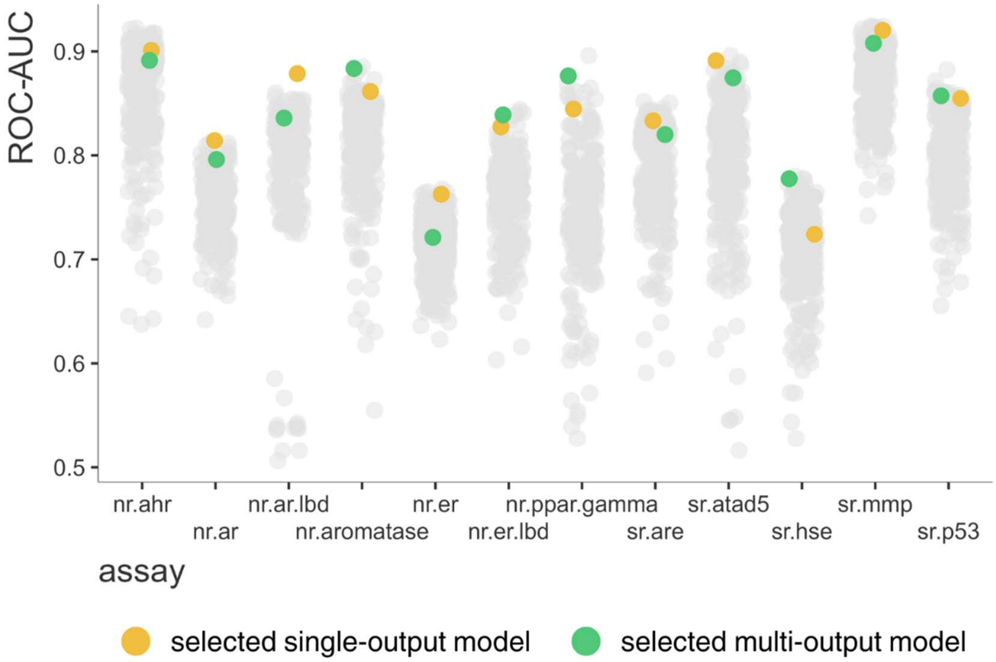


Figure 20 Models' performance on the intermediate test set expressed as ROC-AUC. The data points highlighted in yellow represent the single-output models selected for final evaluation on the real-life test set, while the data points highlighted in green represent the multi-output model chosen for the same purpose.

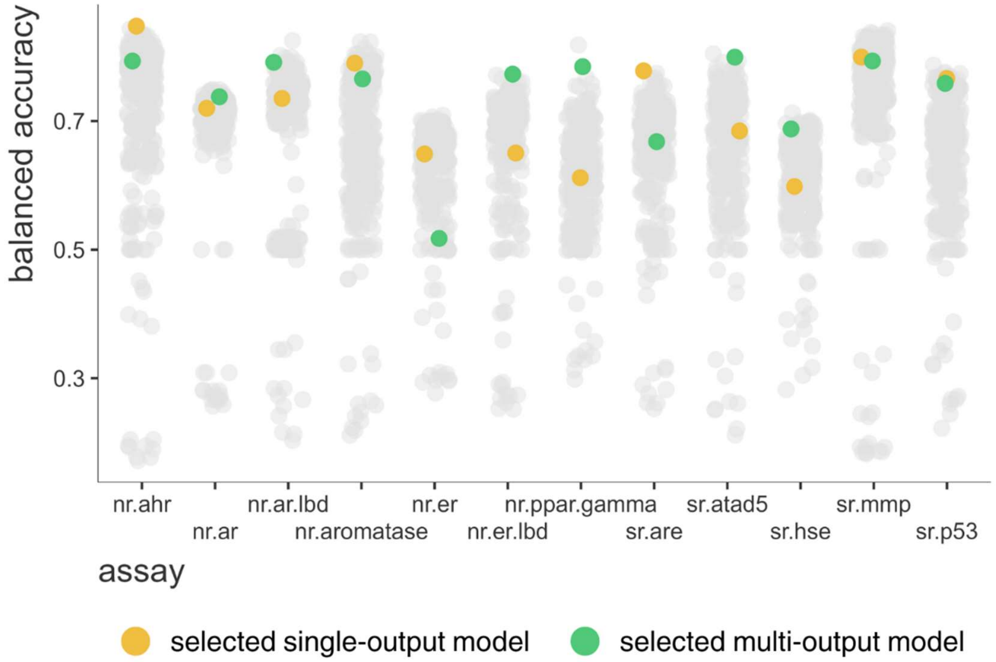


Figure 21 Models' performance on the intermediate test set expressed as balanced accuracy. The data points highlighted in yellow represent the single-output models selected for final evaluation on the real-life test set, while the data points highlighted in green represent the multi-output model chosen for the same purpose.

VI. Effect of sampling strategy

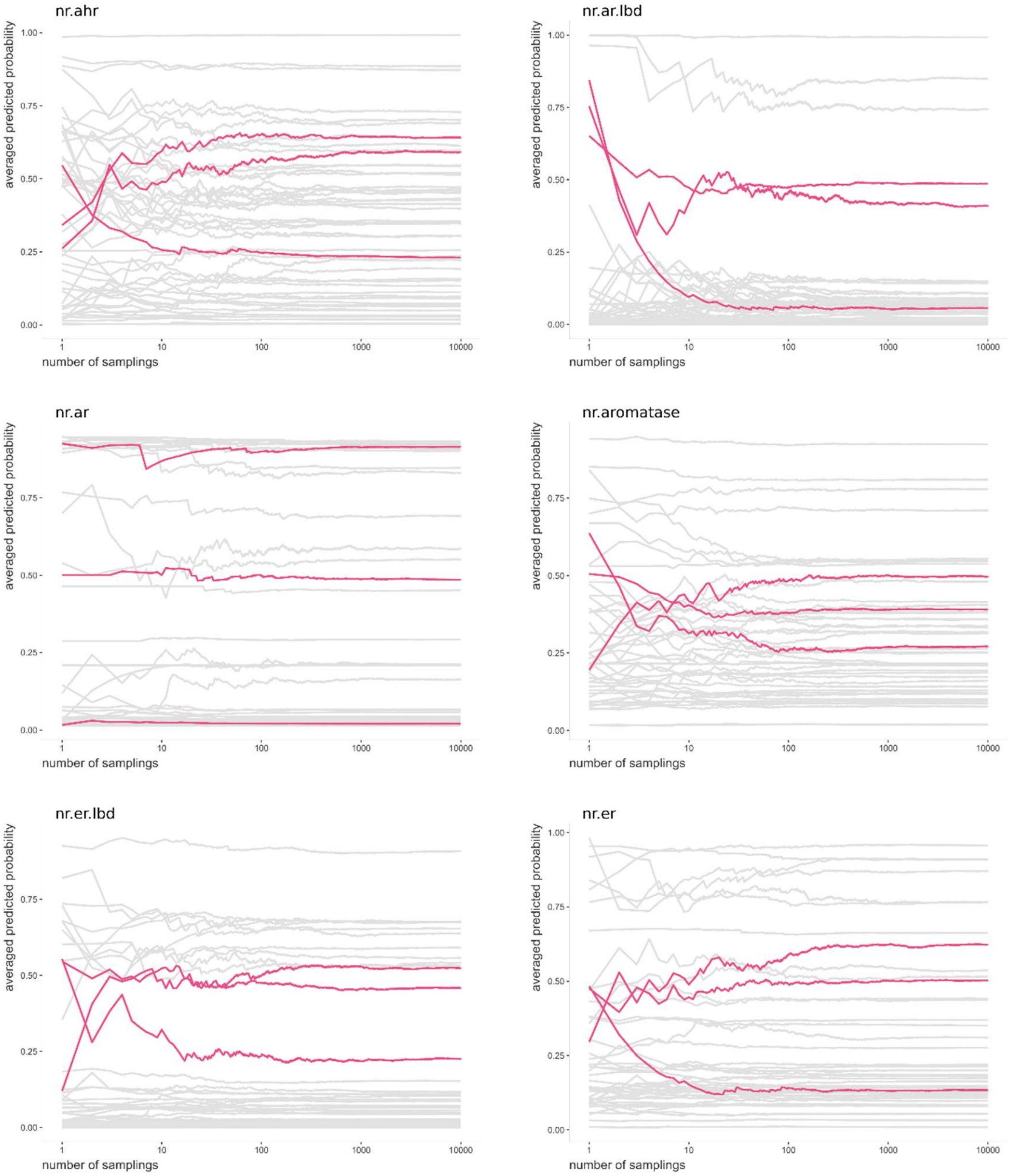


Figure 22 The effect of different sampling iterations (used in the conversion of SIRIUS+CSI:FingerID outputted fingerprint features to the true binary fingerprint features) on the predictions of the single-output model for 50 compounds in the six bioassays. Each line represents the predictions made for one compound, and the pink lines illustrate the importance of adequate sampling iterations

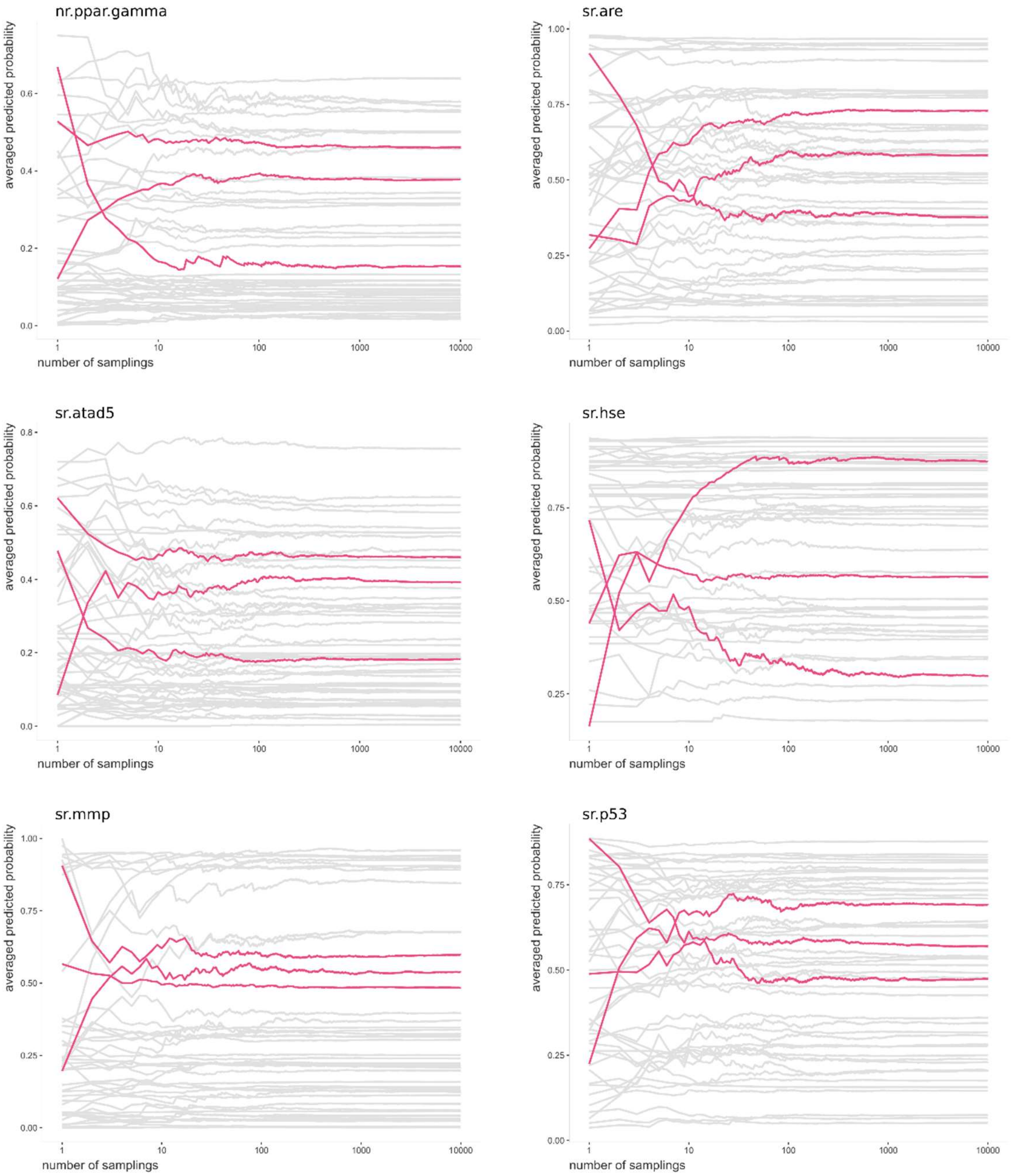


Figure 23 The effect of different sampling iterations (used in the conversion of SIRIUS+CSI:FingerID outputted fingerprint features to the true binary fingerprint features) on the predictions of the single-output model for 50 compounds in the six bioassays. Each line represents the predictions made for one compound, and the pink lines illustrate the importance of adequate sampling iterations

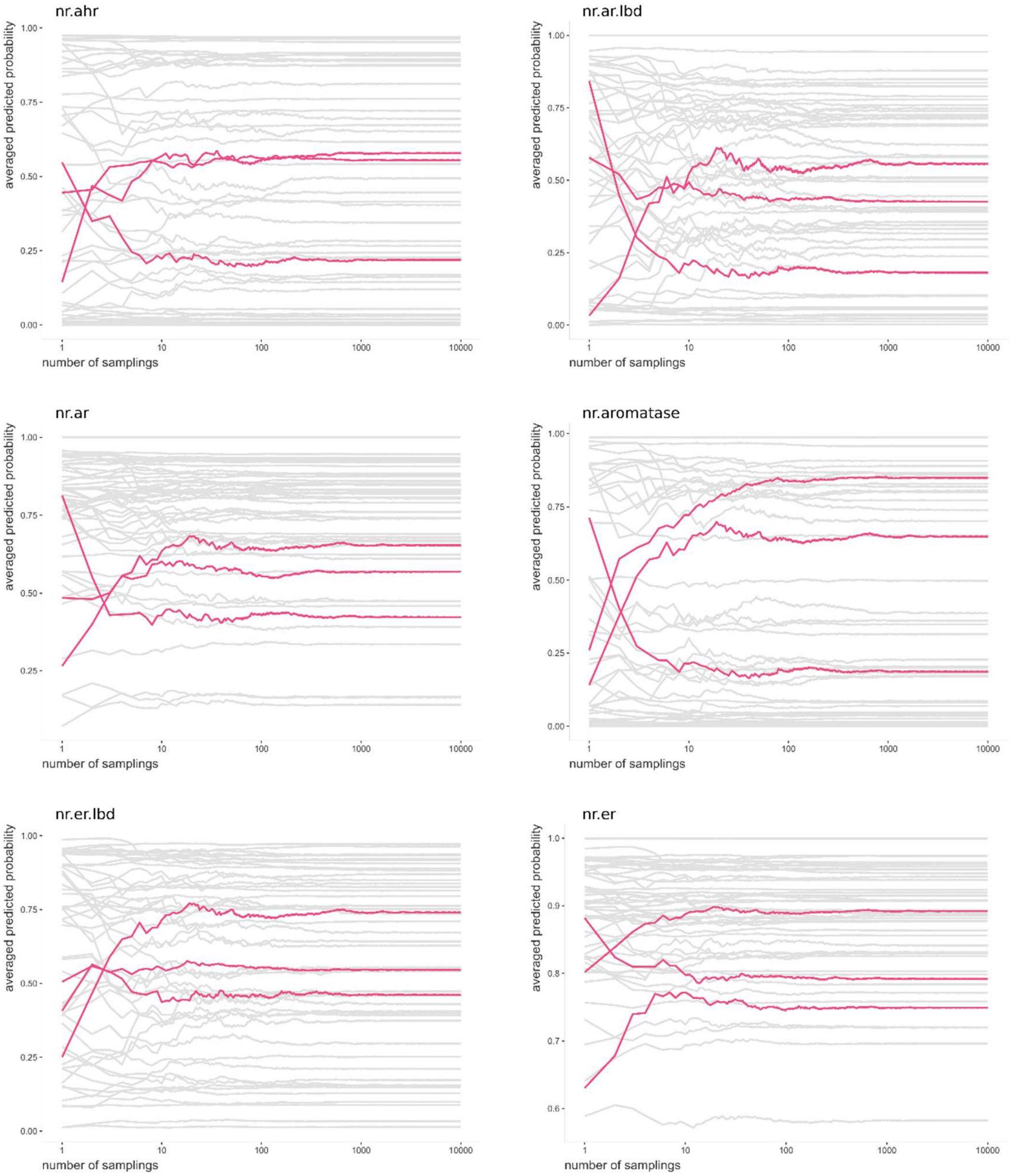


Figure 24 The effect of different sampling iterations (used in converting SIRIUS+CSI:FingerID outputted fingerprint features to the true binary fingerprint features) on the predictions of the multi-output model for 50 compounds in the six bioassays. Each line represents the predictions made for one compound, and the pink lines illustrate the importance of adequate sampling iterations

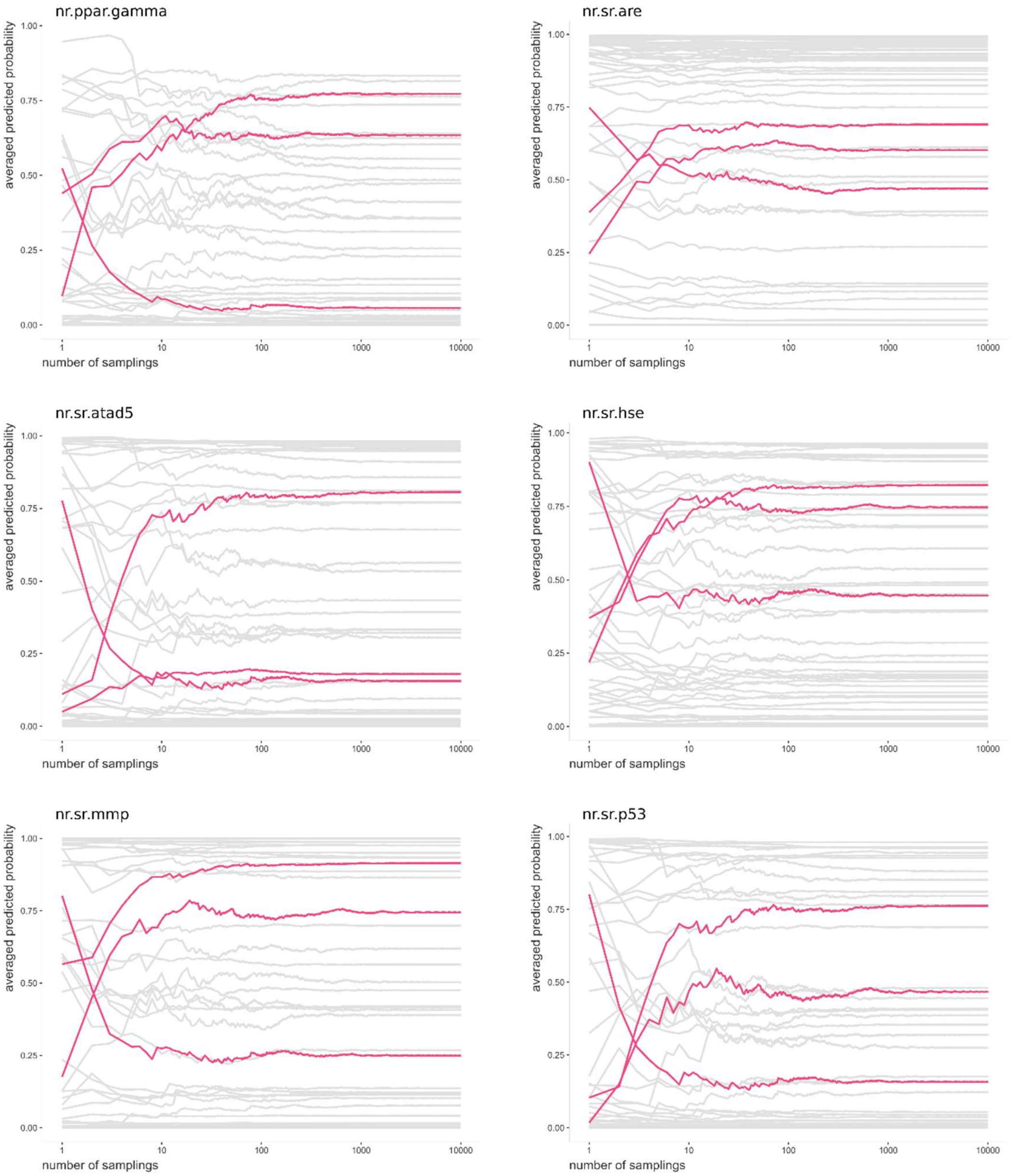


Figure 25 The effect of different sampling iterations (used in converting SIRIUS+CSI:FingerID outputted fingerprint features to the true binary fingerprint features) on the predictions of the multi-output model for 50 compounds in the six bioassays. Each line represents the predictions made for one compound, and the pink lines illustrate the importance of adequate sampling iterations

Table 6 Number of compounds per each bioassay, which predicted endocrine-disrupting activity would have produced a different result after the first iteration as compared to the 10,000th iteration

Bioassay	**Number of compounds**	
	single-output model	**multi-output model**
nr.ahr	61	35
nr.ar.lbd	12	60
nr.ar	1	40
nr.aromatase	30	35
nr.er.lbd	7	59
nr.er	14	0
nr.pppar.gamma	6	31
sr.are	59	13
sr.atad5	17	62
sr.hse	47	55
sr.mmp	17	41
sr.p53	52	59

VII. Toxicology and its principles

Toxicology is a scientific discipline investigating the harmful effects of chemicals (e.g. pesticides, solvents, food additives, drugs) and physical agents (e.g. radiation, coal dust) on living organisms under certain exposure conditions. The roots of toxicology can be traced back to ancient civilisations, where the study of poisons and their effects on the human body was first recorded. The term "*toxicology*" also originates from the ancient Greek word "*toxikón*," representing the poisons used to treat arrowheads before hunting and warfare. Nowadays, this simple "*study of poisons*" has evolved into an interdisciplinary field encompassing pharmacology, biochemistry, environmental science, and epidemiology; and in this integrative approach, toxicologists try to find answers on how toxic substances, so-called toxicants, interact with biological systems and which are the mechanisms behind their adverse effects.[5,6]

Although all the fields are strongly interrelated, toxicology can be divided into three main subareas: descriptive, mechanistic and regulatory/applied toxicology. Descriptive toxicology relates to toxicants' toxicity, epidemiology, and biological quantification through tests such as bioassays and structure-activity studies. It focuses on describing the toxic effects of a substance on an organism, including the dose-response relationship and the symptoms produced by exposure to the substance, without attempting to explain the underlying mechanisms of toxicity. Identifying and understanding these cellular, biochemical and molecular mechanisms by which substances cause toxic effects is the focal point of mechanistic toxicology. The results of studies of these two subfields are used as input in applied toxicology, which evaluates this information on behalf of the government or international organisations, aiming to protect the health of workers, consumers, populations, and the environment.[1]

Even though every mentioned subarea has several branches and subdisciplines, all of them are closely linked and contribute to the general risk assessment process. Latter can be described as a four-step procedure involving hazard identification, dose-response assessment, exposure assessment and risk characterisation, and it basis on the fundamental toxicology concepts. These principles describe how toxic effects are related to the dose of a toxicant, the route of exposure, and the duration of exposure, and take into account various factors, like the target organ and the susceptibility of different populations, such as age, sex, and species.[1]

The dose-response relationship is widely regarded as the most crucial foundation in toxicology. It describes the relationship between the amount of an agent, or dose, and the magnitude of the resulting biological effect or response. Since the term is also used in pharmacology, agricultural sciences, biochemistry *etc*., it is important to note that the response could be wanted or unwanted based on the field of application. The interpretation of dosage depends on the response/endpoint being measured. Historically, in toxicology, mortality is considered an observable response; therefore, the dosages are frequently given as lethal doses (LD) or concentrations (LC). In addition, toxic doses (TD) and sentinel doses (SD) are also used, while seriously harmful or minor adverse effects (*e.g.* headache, fatigue) are reckoned as a response.[1]

[5] In Principles of Toxicology: Environmental and Industrial Applications (eds. Williams, P. L. et al.) 3–34, 232–235 (Wiley–Blackwell, 2000)

[6] In Encyclopedia of Toxicology (ed. Wexler, P.) 4, 1–9, 158–164, 590–594, 610–616, 634–637, 673–675, 680–681, 718–720, 731–745 (Academic Press, 2014)

The dose-response relationship can be characterised in two ways: describing the response of an individual organism to varying doses of a toxicant (individual/graded dose-response relationship) or describing the doses of a substance that produce a given effect in a population (quantal dose-response relationship). However, the latter is more often employed because the unique characteristics of individuals largely influence the impact of a toxicant, and therefore, it is almost impossible to make generalisations based on individual dose-response analysis.[1]

The second principle of toxicology states that the duration and frequency of exposure can affect the toxicity of the substance. Based on the length of the period, the exposure is usually categorised as acute, subacute, subchronic or chronic. The concept is vividly illustrated by the fact that smoking one cigarette does not result in severe complications, but smoking regularly over a long time range can end in lung cancer [37]. Also, chemicals that produce unfavourable effects with a single dose may have no effect if the same total dose is given during a prolonged time interval as multiple smaller portions. Thanks to many coordinated processes that take place in the body, organisms have the ability to eliminate toxicants. If the elimination rate is higher than the administering rate, the toxic concentration of this substance may never be reached. The efficiency of the elimination process is highly related to the organism's characteristics, which leads to the next concept of toxicology.[1]

The susceptibility to chemicals varies between species and among individuals within a species[1]. For example, the botulinum toxin is highly toxic (the most potent known toxin) to humans[7], but vultures[8] have been found to be resistant to it. The interspecies differences in susceptibility and sensitivity can be explained by many factors, such as age, gender, health, and genetics. The young and elderly are usually more affected than adults due to their decreased ability to eliminate toxicants. Also, individuals whose status of well-being is dropped are more prone to the effects of toxic agents. The difference in hormones and physiological processes between males and females may result in nonidentical responses to exposure to the same agent, *e.g.* several studies have shown that women experience more adverse effects of drugs. And last but not least, genetic variability, which makes all organisms unique, can simultaneously give an advantage to one individual and a disadvantage to another in coping with the toxicity of substances.[1] Acatalasia is an autosomal recessive peroxisomal disorder caused by significantly decreased levels of the enzyme catalase. The root of the disorder lies in the genetic mutation in the CAT gene that encodes the particular enzyme. Organisms with acatalasia have much slower rates of removal of hydrogen peroxide; therefore, this chemical may have adverse effects on them.[1,9]

Another major aspect that should be considered while working in the field of toxicology is the routes of exposure. Based on this basic principle, the pathway by which a chemical enters the body, such as inhalation (lung exposure), ingestion (oral/gastrointestinal exposure), skin contact (skin/dermal exposure), or injection (intravenous, intraperitoneal, subcutaneous, intramuscular), can affect its toxicity. The route determines the amount of toxicant passed by and which organs are exposed to the highest concentration. Latter is very important because the absorption site can dictate the elimination rate and alter the observed toxicity. For example, if the overall metabolism is detoxifying, oral or peritoneal administering can be less harmful than other exposure forms because it ensures that the

[7] Dhaked, R. K. et al. Indian J Med Res 132, 489–503 (2010)

[8] In Field Manual of Wildlife Diseases - General Field Procedures and Diseases of Birds 271–281 (CreateSpace Independent Publishing Platform, 2012)

[9] Wang, D. H. et al. Arch Toxicol 70, 189–194 (1996)

toxicant passes the liver, which has a high capacity to break down chemical agents, before damaging the other organs. However, the opposite scenario can occur if toxic byproducts are created during the elimination process.[1]

The last concepts of toxicology are closely associated with the toxicants themselves. In the real world, organisms are concurrently exposed to a mixture of chemicals. However, the toxic effects of these mixtures can significantly differ from the simple sum of toxicities of each component because of the chemical interactions. There are several mechanisms for how chemical interactions take place. These mechanisms involve modifications in toxicokinetics and/or toxicodynamics and thus affect chemical absorption, distribution, metabolism, and excretion, alter binding to a target site like a receptor, or interfere with tissue repair processes. As a result of a broad spectrum of mechanisms, the toxicity of the mixture can be higher (in the case of synergy and potentiation) or lower (in the case of antagonism) compared to the effects of individual compounds. [1]

Finally, the structure of chemicals greatly impacts their toxicity, as it dictates their physicochemical properties, such as solubility, lipophilicity, redox potential, dissociation constant, hydrogen bonding ability, and complex forming ability, to name some of them. Thus, the structure is a crucial factor in determining the reactivity of the compound and, thereby, the mechanisms of its actions. [1]

VIII. License

Non-exclusive licence to reproduce the thesis and make the thesis public

I, **Ida Rahu**,

1. grant the University of Tartu a free permit (non-exclusive licence) to:

reproduce, for the purpose of preservation, including for adding to the DSpace digital archives until the expiry of the term of copyright, my thesis

Machine learning for assessing toxicity of chemicals identified with mass spectrometry,

supervised by Anneli Kruve and Meelis Kull.

2. I grant the University of Tartu a permit to make the thesis specified in point 1 available to the public via the web environment of the University of Tartu, including via the DSpace digital archives, under the Creative Commons licence CC BY NC ND 4.0, which allows, by giving appropriate credit to the author, to reproduce, distribute the work and communicate it to the public, and prohibits the creation of derivative works and any commercial use of the work until the expiry of the term of copyright.

3. I am aware of the fact that the author retains the rights specified in points 1 and 2.

4. I confirm that granting the non-exclusive licence does not infringe other persons' intellectual property rights or rights arising from the personal data protection legislation.

Tartu, **09.05.2023**