



HAL
open science

Evaluation of the use of public toxicological data for chemical hazard prediction through computational methods

Ingrid Grenet

► **To cite this version:**

Ingrid Grenet. Evaluation of the use of public toxicological data for chemical hazard prediction through computational methods. Machine Learning [cs.LG]. COMUE Université Côte d'Azur (2015 - 2019), 2019. English. NNT: 2019AZUR4050 . tel-02612815

HAL Id: tel-02612815

<https://theses.hal.science/tel-02612815v1>

Submitted on 19 May 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

THÈSE DE DOCTORAT

De l'utilisation des données publiques pour la
prédiction de la toxicité des produits chimiques

Ingrid GRENET

Laboratoire d'Informatique, Signaux et Systèmes de Sophia Antipolis (I3S)
UMR7271 UNS CNRS

Présentée en vue de l'obtention du
grade de docteur en Informatique
d'Université Côte d'Azur

Dirigée par :
Jean-Paul COMET

Co-encadrée par :
David ROUQUIE

Soutenue le : 9 juillet 2019

Devant le jury, composé de :

Jean-Paul Comet, PR, Université Côte d'Azur
Mohamed Elati, PR, Université de Lille
Kevin Merlo, Dassault Systèmes
Lysiane Richert, PR, Université de Franche-Comté
David Rouquié, DR, Bayer
Céline Rouveirol, PR, Université Paris 13
Olivier Taboureau, PR, Université Paris Diderot



THÈSE DE DOCTORAT

Evaluation of the use of public toxicological data
for chemical hazard prediction through
computational methods

Ingrid GRENET

Laboratoire d'Informatique, Signaux et Systèmes de Sophia Antipolis (I3S)
UMR7271 UNS CNRS

Présentée en vue de l'obtention du
grade de docteur en Informatique
d'Université Côte d'Azur

Dirigée par :
Jean-Paul COMET

Co-encadrée par :
David ROUQUIE

Soutenue le : 9 juillet 2019

Devant le jury, composé de :

Jean-Paul Comet, PR, Université Côte d'Azur
Mohamed Elati, PR, Université de Lille
Kevin Merlo, Dassault Systèmes
Lysiane Richert, PR, Université de Franche-Comté
David Rouquié, DR, Bayer
Céline Rouveirol, PR, Université Paris 13
Olivier Taboureau, PR, Université Paris Diderot



Evaluation of the use of public toxicological data for chemical hazard prediction through computational methods

Composition du jury :

Président du jury

Céline Rouveirol, Professeure des Universités, Université Paris 13

Rapporteurs

Mohamed Elati, Professeur des Universités, Université de Lille

Olivier Taboureau, Professeur des Universités, Université Paris Diderot

Examineurs

Jean-Paul Comet, Professeur des Universités, Université Côte d'Azur

Lysiane Richert, Professeure des Universités, Université de Franche-Comté

David Rouquié, Directeur de Recherches, Bayer SAS

Céline Rouveirol, Professeure des Universités, Université Paris 13

Invité

Kevin Merlo, Chef de projet Informatique, Dassault Systèmes BIOVIA

Remerciements

Tout d'abord, je souhaite remercier les membres du jury pour avoir assisté à ma soutenance de thèse et en particulier Mohamed Elati et Olivier Taboureau pour avoir accepté de rapporter sur mon manuscrit.

Ensuite je remercie à la fois le laboratoire I3S et l'entreprise Bayer, en particulier le site de Sophia Antipolis, pour m'avoir fait confiance et donné l'opportunité de réaliser cette thèse CIFRE.

Je remercie très particulièrement mes encadrants : Jean-Paul Comet mon directeur de thèse et David Rouquié mon co-encadrant chez Bayer. Merci pour tout ce que vous m'avez appris et apporté durant ces trois années. David je te remercie pour ton soutien et ta bonne humeur au quotidien ainsi que ton enthousiasme et ton optimisme sans faille (ce qui n'a pas toujours été mon cas). Jean-Paul je te remercie également pour ton soutien et ta confiance mais aussi pour ton implication exemplaire, quelque soit le jour ou l'heure (j'ai fait mon maximum pour ne pas te déranger pendant les week-ends et tes vacances car je savais que tu aurais pris le temps de me répondre!). J'ai énormément appris à tes côtés et ça a été un réel plaisir de travailler avec toi (nos réunions me manquent déjà).

Durant cette thèse j'ai eu la chance de pouvoir collaborer avec des équipes de recherche académiques et industrielles ce qui m'a été très bénéfique et je remercie en particulier Erol Gelenbe et Yonghua Yin de l'Imperial College de Londres ainsi que Kevin Merlo de Dassault Systèmes avec qui je garde un super souvenir de nos réunions et appels téléphoniques (je n'ai pas oublié le repas réunionais).

Ensuite mes remerciements vont aux membres des deux structures au sein desquelles j'ai travaillé. D'une part l'ensemble du site de Bayer Sophia, en particulier l'ancienne équipe Tox Recherche dirigée par Rémy Bars et désormais l'équipe Early Tox dirigée par Angela Becker. Plus précisément je souhaite remercier toutes les personnes que j'ai côtoyées chez Bayer ou en dehors : Laurent, Béné, Davy, Caro, Isabelle, Alice, Amandine, Oscar, Marie-France, Daria, Benoit, Mélanie, etc. Je remercie également Frédéric Schorsch pour son aide dans la classification des effets tox. Et enfin un remerciement spécial à Fabien Porée qui a su prendre le relais pendant l'absence de David : nos discussions m'ont beaucoup apportée et fait réfléchir. Toujours chez Bayer mais hors de France : I would like to thank Natalia Ryan (from the US) for everything she did for me. She was always very helpful and interested in what I proposed to do, always available at almost anytime for a skype meeting or a short chat and also accompanying me at the SOT in San Antonio in 2018 : thanks a lot ! I also thank Joerg Wichard (from Berlin) who spent six months with us and who taught me a lot, regarding prediction of toxicity but also of soccer game results during Euro 2016 : "allez les Bleus !"

D'autre part je remercie les membres de l'I3S et notamment les doctorants (anciens, futurs et actuels) qui ont contribué à la bonne ambiance et à la bonne humeur au sein du labo : Ben, John, Emilien, Heytem, Ophélie, Laetitia, Laetitia 2, Assia, Rémy, Samvel, Nico, Eman, Cyrille, Oussama. Merci aussi aux permanents : Gilles, Alexandre, Cinzia, Arnaud, Marie, Jean-Charles, Enrico, etc. Et un grand merci à Magali pour son sourire, sa grande aide et sa disponibilité.

Je souhaite désormais remercier mes amis, en commençant par Ophélie et Estelle sans qui ces 3 années n'auraient pas été les mêmes. Merci pour votre présence, votre soutien et tous les bons moments qu'on a passé (je m'arrête ici pour vous car ça pourrait devenir très long). Merci aussi aux autres Newtoniens : Alex, Emilie et Ben ainsi qu'aux autres copains Niçois : Tevaï, Charlotte, Nico, Matthieu (allez l'OM), Julien. Une pensée également à mes amis Clermontois qui ont dû et su s'adapter à mes emplois du temps chargés lors de mes retours à Clermont !

Enfin, un énorme merci à ma famille qui a toujours cru en moi, m'a toujours soutenue et a été présente même à distance : Maman, Roland, Quentin (merci spécial à mon frère adoré qui m'a aidée dans la réalisation de certaines figures de mon manuscrit), Papy, Mamie, Mamie Ginette, Héléne, Marion, Leslie, Annick, Laura, Sacco, etc. Je vous aime et n'aurai jamais pu aller aussi loin sans vous !

Je termine en remerciant toutes les personnes qui ont pu contribuer de près ou de loin à ces 3 belles années et que je n'ai pas citées ici et je vous remercie encore une fois tous car si c'était à refaire je recommencerais !

Titre : De l'utilisation des données publiques pour la prédiction de la toxicité des produits chimiques

Résumé : L'évaluation de la sécurité des composés chimiques repose principalement sur les résultats des études *in vivo*, réalisées sur des animaux de laboratoire. Cependant, ces études sont coûteuses en terme de temps, d'argent et d'utilisation d'animaux, ce qui les rend inadaptées à l'évaluation de milliers de composés. Afin de prédire rapidement la toxicité potentielle des composés et de les prioriser pour de futures études, des solutions alternatives sont actuellement envisagées telles que les essais *in vitro* et les modèles prédictifs d'apprentissage automatique. L'objectif de cette thèse est d'évaluer comment les données publiques de ToxCast et ToxRefDB peuvent permettre de construire de tels modèles afin de prédire les effets *in vivo* induits par les composés, uniquement à partir de leur structure chimique.

A cette fin, et après pré-traitement des données, nous nous focalisons d'abord sur la prédiction de la bio-activité *in vitro* à partir de la structure chimique puis sur la prédiction des effets *in vivo* à partir des données de bio-activité *in vitro*.

Pour la prédiction de la bio-activité *in vitro*, nous construisons et testons différents modèles de machine learning dont les descripteurs reflètent la structure chimique des composés. Puisque les données d'apprentissage sont fortement déséquilibrées en faveur des composés non toxiques, nous testons une technique d'augmentation de données et montrons qu'elle améliore les performances des modèles. Aussi, par une étude à grande échelle sur des centaines de tests *in vitro* de ToxCast, nous montrons que la méthode ensembliste "stacked generalization" mène à des modèles fiables sur leur domaine d'applicabilité.

Pour la prédiction des effets *in vivo*, nous évaluons le lien entre les résultats des essais *in vitro* ciblant des voies connues pour induire des effets endocriniens et les effets *in vivo* observés dans les organes endocriniens lors d'études long terme. Nous montrons que, de manière inattendue, ces essais ne sont pas prédictifs des effets *in vivo*, ce qui soulève la question essentielle de la pertinence des essais *in vitro*. Nous faisons alors l'hypothèse que le choix d'essais capables de prédire les effets *in vivo* devrait reposer sur l'utilisation d'informations complémentaires comme, en particulier, les données mécanistiques.

Mots clés : Prédiction de la toxicité, données publiques, apprentissage automatique

Title : Evaluation of the use of public toxicological data for chemical hazard prediction through computational methods

Abstract : Currently, chemical safety assessment mostly relies on results obtained in *in vivo* studies performed in laboratory animals. However, these studies are costly in term of time, money and animals used and therefore not adapted for the evaluation of thousands of compounds. In order to rapidly screen compounds for their potential toxicity and prioritize them for further testing, alternative solutions are envisioned such as *in vitro* assays and computational predictive models. The objective of this thesis is to evaluate how the public data from ToxCast and ToxRefDB can allow the construction of this type of models in order to predict *in vivo* effects induced by compounds, only based on their chemical structure. To do so, after data pre-processing, we first focus on the prediction of *in vitro* bioactivity from chemical structure and then on the prediction of *in vivo* effects from *in vitro* bioactivity data.

For the *in vitro* bioactivity prediction, we build and test various models based on compounds' chemical structure descriptors. Since learning data are highly imbalanced in favor of non-toxic compounds, we test a data augmentation technique and show that it improves models' performances. We also perform a large-scale study to predict hundreds of *in vitro* assays from ToxCast and show that the stacked generalization ensemble method leads to reliable models when used on their applicability domain.

For the *in vivo* effects prediction, we evaluate the link between results from *in vitro* assays targeting pathways known to induce endocrine effects and *in vivo* effects observed in endocrine organs during long-term studies. We highlight that, unexpectedly, these assays are not predictive of the *in vivo* effects, which raises the crucial question of the relevance of *in vitro* assays. We thus hypothesize that the selection of assays able to predict *in vivo* effects should be based on complementary information such as, in particular, mechanistic data.

Keywords : Toxicity prediction, public data, machine learning

CONTENTS

Abbreviations	5
Introduction	9
1 Risk assessment for new molecules	15
1.1 The regulatory context	15
1.2 From hazard characterization to risk assessment: general principles	16
1.2.1 Definitions	17
1.2.2 Risk assessment	17
1.2.3 Toxicology studies	20
1.2.4 Classification and labelling	21
1.2.5 Assessing the relevance to human: MoA studies	22
1.2.6 The case of endocrine active chemicals	23
1.3 The need for alternative methods	26
1.3.1 The rationale	26
1.3.2 How to address the need?	27
1.4 Position of the thesis in this context	29
2 Machine learning: generalities	31
2.1 Principle of machine learning	31
2.2 Dataset processing	33
2.3 Learning algorithms	35
2.3.1 Types of learning	35
2.3.2 Methods and types of algorithms for supervised learning	36
2.4 Model evaluation	42
2.4.1 Validation	43
2.4.2 Estimation of performance	44
2.5 Dealing with imbalanced data	47

3	Computational toxicology for toxicity prediction	51
3.1	Data in toxicology	51
3.1.1	Types of data	51
3.2	Existing sources of toxicological data	53
3.2.1	Challenges in using toxicity data for computational purpose	55
3.2.2	Focus on the data used in this thesis	56
3.3	From chemical structure to <i>in vitro</i> activity or <i>in vivo</i> toxicity	59
3.3.1	Non machine learning approaches	59
3.3.2	Quantitative Structure-Activity relationship (QSAR)	60
3.4	From <i>in vitro</i> activity to <i>in vivo</i> toxicity	68
3.4.1	Non machine learning approaches	69
3.4.2	Machine learning approaches	71
3.4.3	Combination of several types of information	73
3.5	Summary	75
4	From structure to activity: a preliminary study	77
4.1	Machine learning on datasets constrained by <i>in vivo</i> data	77
4.1.1	Data used	77
4.1.2	Generation and selection of chemical descriptors	81
4.1.3	Learning procedure	81
4.1.4	Results on imbalanced datasets	82
4.1.5	Results on balanced datasets	86
4.1.6	Analysis of the chemical descriptor space of the compounds	90
4.2	Machine learning on extended datasets and comparison to the <i>in vivo</i> constrained ones	91
4.2.1	Data used	92
4.2.2	Learning procedure and evaluation	92
4.2.3	Learning procedure	92
4.2.4	Results on both <i>in vivo</i> constrained and extended datasets	93
4.3	Conclusion	94
5	From structure to activity: a large scale analysis	95
5.1	Datasets building	95
5.1.1	Data used	95
5.1.2	Generation and selection of chemical descriptors	96
5.1.3	Datasets filtering	97
5.1.4	Datasets are highly imbalanced	97
5.2	Simple classifiers	97
5.2.1	Machine learning algorithms	98
5.2.2	Learning procedure and validation	98

5.2.3	Results	99
5.3	An ensemble method: the stacked generalization	106
5.3.1	General principle	106
5.3.2	Learning procedure and validation	106
5.3.3	Results	107
5.4	Estimation of the applicability domain to assess the quality of predictions	110
5.4.1	Estimation of the applicability domain	111
5.5	Conclusion	113
6	From <i>in vitro</i> activity to <i>in vivo</i> toxicity	115
6.1	Endocrine Disruptor Chemicals: reminders	115
6.2	Data used	116
6.3	Estimation of relation between <i>in vitro</i> assays and <i>in vivo</i> outcomes	118
6.3.1	Methods	118
6.3.2	Results	121
6.4	Machine Learning to predict <i>in vivo</i> outcomes	123
6.4.1	Methods	124
6.4.2	Performance results	127
6.5	Conclusion	133
	Conclusion	135
	Appendices	144
A	Toxicological data resources description	145
A.1	Chemical structures	145
A.2	<i>In vitro</i> data	145
A.3	<i>In vivo</i> data	146
A.4	Genomics data	147
A.5	Mechanistic data	148
B	List of the 37 assays used in Chapter 4	149
C	List of the 42 selected assays in Chapter 6	151
D	List of <i>in vivo</i> effects from ToxRefDB grouped into outcomes categories	153
E	Performance of the ML models built in Chapter 6	155
	Bibliography	155

ABBREVIATIONS

ACToR: Aggregated Computational Toxicology Resource AD: Applicability Domain
ADASYN: Adaptive Synthetic Sampling
ADI: Acceptable Daily Intake
ADME: Absorption, Distribution, Metabolism, Excretion
ANN: Artificial Neural Network
AOP: Adverse Outcome Pathway
AOP-KB: AOP-KnowledgeBase
AR: Androgen Receptor
ARfD: Acute Reference Dose
BA: Balanced Accuracy CAR: Constitutive Androstane Receptor
CART: Classification and Regression Tree
CEBS: Chemical Effects in Biological Systems
CMap: Connectivity Map
CMR: Carcinogen, Mutagen, Reprotoxic
CNN: Convolutional Neural Network
CPDB: Carcinogenic Potency DataBase
CTD: Comparative Toxicogenomics DataBase
DILI: Drug-Induced Liver Injury
DSSTox: Distributed Structure-Searchable Toxicity
EADB: Estrogenic Activity DataBase
EATS: Estrogenic, Androgenic, Thyroidal and Steroidogenic
EBI: European Bioinformatics Institute
ECFP: Extended-Connectivity Fingerprints
ECHA: European Chemicals Agency
ED: Endocrine Disruption
EDC: Endocrine Disrupting Chemical
EDSP: Endocrine Disruptor Screening Program
EFSA: European Food Safety Authority

ABBREVIATIONS

EKDB: Endocrine Disruptors Knowledge DataBase
EMBL: European Molecular Biology Laboratory
EPA: Environmental Protection Agency
ER: Estrogen Receptor
EU: European Union
EURL ECVAM: European Union Reference Laboratory for alternatives to animal testing
FedTEX: Fertility and Developmental Toxicity in Experimental animals
FDA: Food and Drug Administration
GLP: Good Laboratory Practices
GSH: Globally Harmonized System of Classification and Labelling of Chemicals
HESS: Hazard Evaluation Support System
HTS: High-Throughput Screening
IATA: Integrated Approaches to Testing and Assessment
ICATM: International Cooperation on Alternative Test Methods
InChi: IUPAC International Chemical Identifier
IPCS: International Programme on Chemical Safety
IUPAC: International Union of Pure and Applied Chemistry
IVIVE: In vitro - in vivo extrapolation
KE: Key Event
kNN: K-Nearest Neighbor
LD50: Lethal Dose 50%
LDA: Linear Discriminant Analysis
LOAEL: Lowest-Observed-Adverse-Effect-Level
MACCS: Molecular ACCess System
MBEA: Maximum Biclique Enumeration Algorithm
MCC: Matthew's correlation coefficient
MIE: Molecular Initiating Event
ML: Machine Learning
MLP: Multi Layer Perceptron
MLRNN: Multi-Layer Random Neural Network
MOA: Mode Of Action
MRL: Maximum Residue Level
NCATS: National Center for Advancing Translation Sciences
NCBI: National Center for Biotechnology Information
NCCT: National Center for Computational Toxicology
NIEHS: National Institute of Environmental Health Sciences
NIH: National Institute of Health
NITE: National Institute of Technology and Evaluation
NOAEL: No-Observed-Adverse-Effect-Level
NTP: National Toxicology Program

NRC: National Research Council
OECD: Organization for Economic Co-operation and Development
PBTK: Physiologically based toxicokinetic
PCA: Principal Components Analysis
PLP: Pipeline Pilot
PPAR: Peroxisome Proliferator-Activated Receptor
PPPs: Plant Protection Products
PXR: Pregnane X Receptor
QSAR: Quantitative Structure-Activity relationships
RBF: Radial Basis Function
REACH: Registration, Evaluation, Authorization and registration of Chemicals
RF: Random Forest
RfD: Reference Dose
RMSE: Root Mean Squared Error
RecNN: Recurrent Neural Network
RNN: Random Neural Network
ROC: Receiver Operating Characteristic
SAR: Structure-Activity Relationship
SDF: Structure Data File
SEURAT: Safety Evaluation Ultimately Replacing Animal Testing
SOM: Self-Organizing Map
SMILES: Simplified Molecular Input Line Entry Specification
SMOTE: Synthetic-Minority Over-Sampling
SVM: Support Vector Machine
TG-GATEs: Toxicogenomics Project-Genomics Assisted Toxicity Evaluation Systems
TGP: Toxicogenomics Project
ToxRefDB: Toxicity Reference Database
TSH: Thyroid Stimulation Hormone
US: United States
WHO: World Health Organization

INTRODUCTION

Toxicity caused by chemical compounds (whether natural or synthetic) to living organisms and the environment is a growing concern for the entire population. In particular for humans, xenobiotics have been associated with several human health issues such as increased incidences of cancers, reduced fertility, metabolic disorders such as diabetes [208]. Indeed, beside the exposure of all the natural small molecules contained in the food and beverages we are eating and drinking, we are daily exposed to a variety of synthetic chemicals such as food additives, households products, personal care products, drugs or pesticides which can induce various adverse effects above certain dose levels and duration of exposure. Adverse effects include carcinogenicity (induction of tumors), mutagenicity (induction of DNA damages), reproductive toxicity (alteration of sexual functions and fertility) and **endocrine mediated toxicity** (alteration of the endocrine system). Endocrine toxicity has raised a lot of attention, in particular in Europe, as endocrine mediated toxicants are believed to induce adverse effects via non-threshold mechanisms consequently preventing the classical use of exposure-based risk assessment. This led instead to the application of hazard-based regulation for this type of compounds.

The current regulation for the registration and marketing of non genotoxic (worldwide) and non endocrine active (in Europe) compounds is based on a **risk assessment** process that evaluates the toxicity caused by molecules to human health and the environment in order to define their safe conditions of use. This assessment relies on a series of **toxicity studies** performed *in vivo* in several rodent and non-rodent species of laboratory animals, for different durations of exposure (from some days up to the whole life-time of animals) and over critical windows of exposure (gestation and shortly after birth). These studies aim at identifying the potential **hazard** of chemicals to finally characterize their **risk** through an assessment of individuals' **exposure** to the compounds. *In vitro* studies, performed using biochemical assays and cell based assays, can also complete the *in vivo* ones in order to determine the biological **mechanisms** that are involved in the pathways leading to toxic effects. The elucidation of these pathways is important for risk assessment since it allows the regulators to conclude if effects are relevant to humans. However, toxicity studies, in particular the *in vivo* ones, are time, money and laboratory animals

consuming which raises ethical and economical concerns. They are also questioned about their ability to inform about the toxicity caused to humans due to the interspecies extrapolation. In addition, many compounds already on the market or natural compounds have very little to no information about their toxic potential preventing a good evaluation of their risk. Nonetheless, this lack of data cannot be filled using the traditional *in vivo* testing methods since they are not adapted to the rapid evaluation of thousands of chemicals. Therefore, regulations in Europe such as REACH [82] (Registration, Evaluation, Authorization and registration of Chemicals) and authorities such as the US Environmental Protection Agency (EPA) in the United States are asking for new optimized testing methods which are more predictive of toxicity as well as more reliable, faster, and less animals consuming than actual methods.

In recent years, this urgent challenge has been considered by the toxicology community who proposed a new testing paradigm, the Tox21 vision [55], that aims at shifting from the traditional testing approaches towards a predictive toxicology based on new alternative approaches enabling a rapid screening of compounds for their toxic potential in order to prioritize them for further testing. Such alternative approaches are also of great interest for phytopharmaceutical and pharmaceutical companies since they follow essentially the same approach to de-risk early candidate compounds in the hit identification and hit optimization phases. Moreover, these new approaches should also focus on the identification of biological mechanisms that induce toxicity. Among the envisioned alternative methods for predictive toxicology are: (1) the use of human-relevant *in vitro* assays to provide hints about the bioactivity of chemicals, meaning their ability to affect biological processes and, (2) the use of *in silico* methods to perform data analysis and integration and to develop predictive models of *in vivo* toxicity.

Following the new paradigm, initiatives have been launched in order to generate large amount of data such as the ToxCast and Tox21 programs which performed *in vitro* High-Throughput Screening on thousands of chemicals [260, 217]. Moreover, many studies have been completed in order to develop computational models to either understand the mechanisms of toxicity, based for example on systems biology or to predict toxicity, based for example on machine learning. Precisely, systems biology aims at modeling the interactions of the components of complex biological systems, at different levels of organizations, while machine learning aims at predicting a specific property based on existing data that enable a learning algorithm to associate characteristic features to the specific property by finding informative relationships. These types of computational approaches and in particular machine learning, require a large amount of data to generate accurate models in order to go over the maximum of conditions which induce toxicity. Nevertheless, the availability and usability of toxicological data is often challenging for diverse reasons and necessitates an evaluation of their suitability to *in silico* modeling.

Even if a lot of work has been already performed in using computational approaches for toxicity prediction, there are still needs for good and reliable methods. Specifically, machine learning has been widely used for the prediction of diverse types of effects and appeared to result in a broad range of models' performances (from really bad to quite good). Thus, no "ultimate" method has been clearly identified to produce sufficiently accurate models.

Actually, this challenge of finding the best method is in practice unsolvable because of a well known limitation of machine learning, exposed by Wolpert in 1996 in the "No Free Lunch" theorem [279]: no algorithm works better than the others for every problem and an algorithm which is good for one specific problem may not be good for another one. Indeed, since a model is an idealization of the reality, it is made of hypotheses and assumptions that highly depend on the considered problem. Thus, these hypotheses differ for each problem and different problems should be associated with different models. Although the theorem has been stated more than two decades ago, it is still discussed and considered today by machine learning scientists [112]. Specifically, the "No Free Lunch" theorem implies that, if we want to build a good machine learning model, we should try multiple algorithms to find the one that works the best. Consequently, a large part of research in machine learning focuses on finding the type of algorithm that generates good models for a specific type of data; toxicological data in our case.

Objective and approach

This thesis comes from a CIFRE¹ agreement between academic research and the agrochemicals division of Bayer and takes place at the interface between two major disciplines: **computer science** and **toxicology**. According to the toxicological context, our motivation is to develop *in silico* tools, and in particular machine learning (ML) methods, for the early prediction of potential hazard of compounds using the publicly available data. This motivation fits with one of the current objective of Bayer, which intends to use computational approaches to perform compounds' selection in the early phases of the development of new **plant protection products** in order to reduce animal use, cost and development time.

As illustrated in Figure 1, the ideal objective would be to predict effects observed in long-term *in vivo* studies, directly from the chemical structure of compounds. Nonetheless, this long-term prediction seems to be ambitious [259] because of the high level of biological complexity and variability and because toxicity can result from a long chain of causality involving multiple pathways at different biological levels [226]. Indeed, the performance of ML models tends to be better when the complexity of the modeled property is low and conversely. We therefore propose a **two-stage ML approach** where the first stage (1) is to predict bioactivity (*in vitro* assays) based on chemical structures and the second stage (2) is to predict *in vivo* toxicity from *in vitro* data, possibly combined with chemical structure (2'). Models resulting from these two stages should then be chained up to predict *in vivo* toxicity directly from structural data for a new compound for which only the molecular structure is known.

Similar works to our two-stage approach have already been proposed. For example, Martin *et al.* in 2011 [179] developed the Profile-QSAR method which is composed of two steps: first, ML models are built to predict the activity of compounds in hundreds of kinase *in vitro* assays based on chemical structure and second, predictions from the previous models are used as input of another ML model to predict the activity for a new kinase. In 2017, a new version of this

¹In French: CIFRE = Conventions Industrielles de Formation par la REcherche

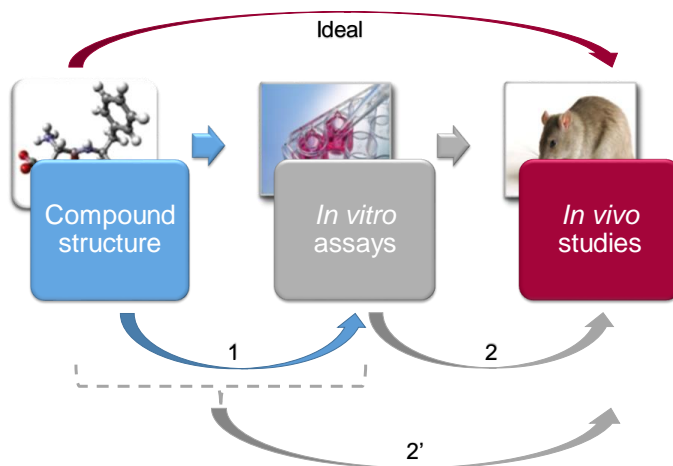


Figure 1: Objective and the proposed 2-stage machine learning approach. The objective is to develop ML models to predict potential hazard of compounds. The ideal would be to directly predict *in vivo* outcomes from chemical structure but it seems too ambitious. We therefore propose a two-stage approach constituted of: (1) the prediction of *in vitro* results based on chemical structure and (2) the prediction of *in vivo* outcomes based on *in vitro* data, possibly combined with chemical structure (2').

method has been published: the Profile-QSAR 2.0 which is based on different ML algorithms than the first version [180]. More recently, in 2018, Guan *et al.* [110] proposed to predict rat carcinogenicity using a ML model whose descriptors are the predictions from four ML models predicting four different *in vitro* bioassays.

These works illustrate that the general idea of chaining two types of models has been considered by several scientists but it is only recently that it has been implemented, and for specific cases. Here we suggest to evaluate how this global idea could be applied in a large scale context, without focusing on a particular set of *in vitro* assays as intermediate data. Indeed, we aim at using all the available *in vitro* assay results to find the combination of assays that is the best predictive of an effect of interest through machine learning. It is only when this set has been identified that we propose to build ML models to predict each of these assays. Obviously, even if we are interested here in *in vitro* assays, other intermediate data could be envisaged.

In order to build ML models such as proposed in our approach, three types of data are required: chemical structure of compounds, results from *in vitro* assays and results of *in vivo* studies. The first step therefore consists in reviewing the publicly available data and choosing the most appropriated ones for our purpose by taking into account the various challenges raised by the different resources. Then, each stage of the two-stage approach can be evaluated independently by building ML models using different types of learning algorithms. Indeed, because of the "No Free Lunch" theorem, all the work presented in this manuscript focuses on the use of several types of algorithms and machine learning methods, as well as the study of criteria that enable the building of the best and most reliable models. If sufficient performance are obtained, the chaining of the two types of ML models will therefore be envisaged. Indeed, since the performance of the global approach is a composition of the performance of each of the two stages, these two performances should be high enough to result in a correct global performance. In particular, the

studies proposed in this manuscript show that our two-stage approach is not generalizable to the prediction of all types of effects due to insufficient performances obtained in each of the two stages.

Reading guide

This manuscript is organized in six chapters. The three first chapters focus on the multidisciplinary context of the work while the three last ones present the studies that have been performed in order to contribute to the overall objective.

Chapter 1 presents the toxicological context of this work and aims at defining the important notions required by the reader to understand the general concepts of toxicology. It first focuses on the current regulation of chemical compounds and then develops the principles that enable their risk assessment, on which the regulation relies. It finally highlights the need for alternative approaches and gives examples of existing initiatives that have been started or achieved in order to address this need.

Chapter 2 focuses on generalities on machine learning methods, because numerous methods are used afterwards, see the "No free lunch" theorem. In particular, it details the principle of the global methodology and the different steps that are required to build ML models. These steps include data processing, learning and evaluation and each of these steps can be performed by considering various techniques which are more or less detailed according to their importance for the understanding of the following work.

Chapter 3 puts the disciplines of toxicology and computer science together in order to inform the reader on what is already available for toxicity prediction using *in silico* tools. It first lists the existing resources for toxicological data as well as the various challenges they raise and puts an accent on the data used in our work. Secondly, it provides a large overview of the state of the art of compounds' bioactivity and toxicity prediction using computational tools. This overview is split into two parts corresponding to the two stages of our approach: the prediction of *in vitro* bioactivity and the prediction of *in vivo* toxicity.

Chapter 4 presents the results of a preliminary study that focuses on the first stage of our two-stage approach: the prediction of *in vitro* assay results from the chemical structure of compounds. Constrained by the amount of data, we build ML models for only 37 *in vitro* assays, using two datasets composed of different number of compounds and we also tests several ML methods as well as data augmentation.

Chapter 5 presents a second study that aims at developing ML models for the prediction of *in vitro* bioactivity from chemical structure of compounds. Contrary to the previous study, this one proposes a large scale analysis by building ML models for more than 500 assays. It enables us to evaluate how to build good ML models with the type of data used. Indeed, after comparing several ML methods, we propose an approach that results in reliable models when combined with the estimation of the applicability domain.

Chapter 6 presents the results of a study related to the second stage of our approach: the pre-

diction of *in vivo* toxicity from *in vitro* data, possibly combined to chemical structure. In this study we specifically focus on *in vivo* effects observed in endocrine organs and try to evaluate their relation with *in vitro* assays supposed to target the biological pathways leading to endocrine effects. This relation is first evaluated through simple statistical analysis and then by ML modeling. The results show that the *in vitro* assays are not sufficiently informative of the *in vivo* effects to enable their prediction., we further discuss the possible reasons of this finding.

We end this manuscript by providing a general discussion and conclusion on the results obtained in the three studies. In particular, we further discuss the possible reasons of the findings of the last study. We finally propose perspectives for further work.

CHAPTER 1

RISK ASSESSMENT FOR NEW MOLECULES: ACTUAL AND FUTURE TRENDS

This chapter introduces the toxicological context and the generalities about the risk assessment for new small molecules, focusing on the regulation of plant protection products (PPPs) (*i.e.* pesticides). We show that current approaches raise various issues and concerns regarding animal usage and extrapolation of the results obtained in laboratory animals to humans and that there is a need for new alternative methods, including computational tools.

1.1 The regulatory context

All the living beings are daily exposed to a variety of chemical compounds, either naturals or synthetics, which can induce **adverse toxic effects** if a combination of dose level and duration of **exposure** sufficiently is achieved. Regarding the synthetic compounds manufactured by the chemical industry, we can distinguish several types of compounds ranging from pharmaceutical ones to **plant protection products (PPPs)** through food additives, household products or cosmetics. Worldwide, every new industrial compound should be demonstrated as safe for the human health and the environment in order to be registered and placed on the market. This is done by performing various toxicity studies allowing the characterization of the potential **hazard** caused by the chemicals. These studies are different according to the type of chemical compounds: the more in depth safety evaluation being for pharmaceutical and plant protection products by far. For example, in the case of pharmaceuticals, pre-clinical studies are first performed using laboratory animals (rodent and non-rodent species) and they are followed by clinical studies directly performed in humans. Regarding cosmetic compounds, animal testing is banned in several countries including the EU (since 2013) and tests are only performed using non animal alternative methods. In this work, we are mainly focusing on PPPs for which the studies conducted for the risk assessment process follow essentially the same approach as the one for pharmaceutical products in the pre-clinical phase. For obvious reasons, PPPs are not tested in humans.

Regarding the registration of chemical compounds, each country has its own legislation and standards which can sometimes be difficult to harmonize [118]. Here we focus on the European Union (EU) legislation.

Since 2007, the **REACH** regulation about the Registration, Evaluation, Authorization and restriction of Chemicals governs the industrial chemicals legislation in the EU, in particular for high production volume chemicals [82]. Indeed, REACH asks all companies, manufacturers, importers and users of more than one ton of chemical substances per year to register the compounds with the **European Chemicals Agency (ECHA)** (European agency specifically created in 2007 for the REACH regulation). To apply for a registration, companies should identify the **risks** of the substance and provide a technical dossier as well as a chemical safety report which demonstrate how the substance can be used safely. Then, ECHA evaluates the registration and assess how the risk can be managed and finally gives its authorization as well as restrictions if necessary.

The REACH regulation applies for any type of chemical compounds, including PPPs in particular for their intermediate of synthesis. In the EU, the legislation for approval of PPPs is principally based on the Regulation (EC) No 1107/2009 [84] which is accompanied by other regulations and directives and is in accordance with the REACH regulation. The **European Food Safety Authority (EFSA)** is the European Commission agency in charge of reviewing all test results provided by the entity asking for the approval of a new substance. In particular, they review the risk assessments (described in Section 1.2.2), provide scientific advice to the European Commission on possible risks related to the substance and set maximum residue levels (MRLs) which are the legal limits for the substance residue in food and animal feed. EFSA is also in charge of the renewal of substance approval. Indeed, an approval is generally valid for 10 years, after which an application for a renewal should be submitted.

In the United States, the **Environmental Protection Agency (EPA)** is in charge of the registration of PPPs.

Whatever the considered agency for registration, a risk assessment should be performed by the requesting entity and its process is described in the next section.

1.2 From hazard characterization to risk assessment: general principles

Before introducing the principle and process of risk assessment for PPPs, some important notions and concepts have to be defined.

1.2.1 Definitions

Adverse effect: An adverse effect is a "change in the morphology, physiology, growth, development, reproduction or life span of an organism, system, or (sub) population that results in an impairment of functional capacity, an impairment of the capacity to compensate for additional stress, or an increase in susceptibility to other influences" [134].

Hazard: The hazard represents the "inherent property of an agent or situation having the potential to cause adverse effects when an organism, system or population is exposed to that agent" [134]. Basically, it is a potential source of harm for a person or the environment.

Exposure: The exposure represents the "concentration or amount of a particular agent that reaches a target organism, system or (sub) population in a specific frequency for a defined duration" [134].

Risk: The risk is the "probability of the adverse effect occurring in an organism, system or (sub) population caused under specified circumstances by exposure to an agent" [134]. More generally, the risk is the likelihood that an organism may be harmed if exposed to the hazard and it therefore depends on both the hazard and the exposure according to the following definition:

$$Risk = Hazard \times Exposure$$

1.2.2 Risk assessment

Chemical **risk assessment** corresponds to the process that enables the evaluation of the potential risk caused by a chemical to organisms, systems and populations in the context of defined exposure. This process has been introduced in 1983 by a famous publication from the US National Research Council called the "Red Book" [54] and further detailed in another report in 1994 [56]. Currently, the overall process consists of four steps which are detailed hereafter and illustrated in Figure 1.1.

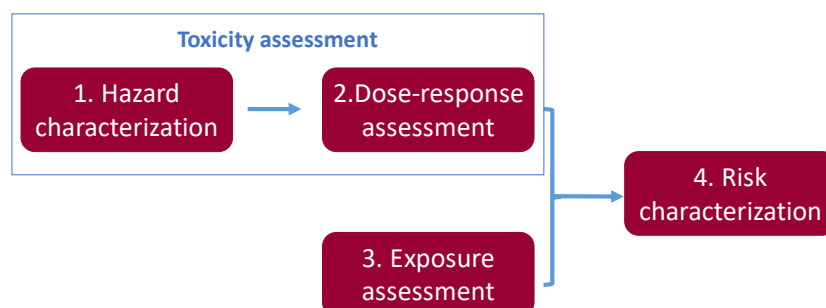


Figure 1.1: The risk assessment process. The process includes 4 steps.

- 1. Hazard characterization:** This step aims at identifying the adverse effects that can be caused by a chemical. In particular, it describes the target organs and the type of toxicity

(neurotoxicity, carcinogenicity, hepatotoxicity, developmental toxicity, *etc.*) as well as the conditions under which it occurs. In the case of PPPs, this information is obtained from *in vivo* studies performed in laboratory animals (rodents and non-rodents) during which several parameters are measured or evaluated (clinical signs, weights, clinical and anatomic pathology, histopathology, *etc.*). More details about these studies are given in Section 1.2.3.

- Dose-response assessment:** This step further evaluates the conditions under which the hazard can occur by quantifying the relationship between the dose of exposure and the severity of the toxic effects. The dose-response concept is based on the famous maxim from Paracelsus who said that "only the dose makes the poison" which enunciates that a substance can induce its adverse effects only if it reaches a certain concentration in an organism. Therefore, the dose-response assessment aims at looking for the minimal dose at which an effect is induced which is called the **Lowest-Observed-Adverse-Effect-Level** (LOAEL). Another commonly used dose is the **No-Observed-Adverse-Effect-Level** (NOAEL) which corresponds to the highest tested dose of exposure for which no effect could be observed compared to a control group in a given repeated dose toxicity study. These doses are obtained after *in vivo* studies performed in laboratory animals and are expressed either in parts per million (ppm) or milligram per kilogram of body weight per day (mg/kg/d).

In order to estimate "safe" levels for humans by extrapolation, the smallest NOAEL obtained in the set of conducted toxicity studies is divided by an uncertainty factor which takes into account the interspecies variability (between laboratory animals and humans) and the intra-species variability (due to human sensitivity differences according to gender, age, health, genetics, *etc.*). In general, a factor of 10 is applied for the two types of variability, resulting in a total uncertainty factor of 100 (10×10). Thus, the resulting "safe" exposure for humans equals $NOAEL/100$ and is called the Reference Dose (RfD) or **Acceptable Daily Intake (ADI)** and corresponds to the amount of substance to which a human can be exposed on a daily basis over lifetime without any undesirable effect. For example, if the smallest NOAEL of 50 mg/kg/d is obtained for a given compound, the resulting ADI for humans will be 0.5 mg/kg/d. A representation of the dose-response curve with the different doses described above is displayed in Figure 1.2

Hazard characterization and dose-response assessment constitute the **toxicity assessment**.

- Exposure assessment:** This step measures (for already marketed compounds) or estimates (for new compounds) the levels at which the organisms are exposed to the considered chemical or its residues. In particular, it involves the identification of the population that could be exposed as well as the route, the magnitude and the duration or frequency of exposure. In particular for PPPs, we distinguish operators who will handle the products from the rest of the population and another limit value is specifically derived for operators: the Acceptable Operator Exposure Level (AOEL). Overall, the exposure assessment step results in an exposure estimate.

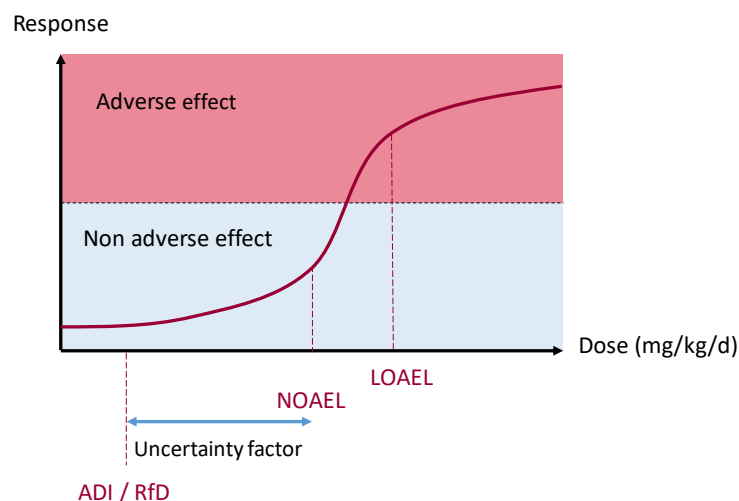


Figure 1.2: Schema of the dose-response curve with the corresponding doses. LOAEL = Lowest-Observed-Adverse-Effect-Level; NOAEL = No-Observed-Adverse-Effect-Level; ADI = Acceptable Daily Intake; RfD = Reference Dose.

4. **Risk characterization:** This step integrates the information obtained from the toxicity assessment and the exposure assessment in order to conclude on the likelihood that the hazard associated with the compound will occur in the population. It defines the nature of the risk but also the uncertainties encountered in the different steps of the risk assessment process. At the end, the risk is either "acceptable" (*i.e.* the expected exposure does not exceed the limit values) or "unacceptable" (*i.e.* the expected exposure does exceed the limit values). In the latter case, the compound cannot be commercialized.

Risk characterization is often followed by the **risk management** which aims at mitigating the risk and evaluating the impacts of regulatory measures on the risk. In the end, risk characterization and risk management lead to registration decisions.

The risk assessment approach as detailed here is "threshold based" since it considers that compounds induce their adverse effects above a certain dose. Nonetheless, other approaches for chemical regulation are considered such as:

- **The non-threshold based approach:** it is followed by North American Authorities and corresponds to the case of carcinogenic chemicals for which dose-response curves are assumed to have in a first place no threshold, meaning that any exposure can lead to a cancer. Moreover, the effect accumulates after each exposure to the chemical resulting in a linear dose-response from which the **slope factor**, also known as Q^* (Q-star), can be computed. It corresponds to the dose level which would induce an incidence of cancer of one individual in a population of 1 million individuals taking into account the incidence of tumors observed in the corresponding bioassay. It is used to estimate the risk over individuals lifetime continuously exposed to the chemical. To have the compound evaluated in a classical risk based approach, the applicant needs to make the demonstration that the

compound is inducing tumor via threshold based mechanisms. The demonstration represents basically the elucidation of the toxic Mode of Action involved in the tumor formation (see Section 1.2.5).

- **The hazard-based approach:** in that case, if some specific hazards are identified to be caused by chemicals, they are directly banned from the market, without considering the exposure. This is in particular the case for genotoxicity worldwide and endocrine active chemicals in EU.

These two approaches are often considered at the same time since it is the characterization of hazard that inform about the non-threshold characteristic.

1.2.3 Toxicology studies

In order to generate data to enable risk assessment, toxicity *in vivo* studies are performed using laboratory animals. They aim at looking at various **endpoints** (or outcomes) which are the results of an interaction between a chemical substance and the biology of an organism. Several types of studies are required by the authorities and should follow the international test guidelines proposed by the **Organization for Economic Co-operation and Development (OECD)** [202]. In particular, these guidelines require that regulatory studies should be performed according to the **Good Laboratory Practice (GLP)** [197] which ensures the generation of high quality and reliable data by providing international standards regarding the process and conditions under which studies should be planned, performed, monitored, recorded and reported. Toxicology studies include [202]:

- **Acute toxicity studies:** they evaluate the mortality induced by a compound after either a single-dose exposure or multiple exposures in a short period of time (< 24 hours). Administration routes can be oral, dermal and inhalation. If mortality is observed, a **LD50** (Lethal Dose 50%) is derived, corresponding to the dose at which 50% of the individuals dies (if the compound is administered by inhalation, it is the LC50 for Lethal Concentration 50%). Acute toxicity studies enable the definition of a limit value used for PPPs: the **Acute Reference Dose (ARfD)** [81].
- **Short-term studies:** they evaluate the adverse effects of repeated exposure to different dose levels during up to 10% of the animal's lifespan. Usually, at least 3 dose levels are used in order to inform about the dose-response relationship and derive the NOAEL. We distinguish **sub-acute** studies from **subchronic** studies according to the time of exposure (respectively 14 to 28 days and 90 days in rodents).
- **Long-term studies:** also called **chronic studies**, they evaluate the adverse effects of repeated exposure during the expected lifespan of the animal (2 years in rodents). They inform about the systemic toxicity and dose-response relationship. They are usually combined to **carcinogenicity studies** which look for the potential of compounds to induce cancer.
- **Reproductive and developmental studies:** they evaluate the potential adverse effects on the sexual function and fertility of both male and female adults and the development

of offsprings since the embryonic and fetus stages. These studies require two generations of animals.

- **Genotoxicity studies:** they evaluate the capacity of compounds to induce DNA damage and chromosomal changes (regarding the number or the structure). More specifically, **mutagenicity studies** look for the genetic changes that are caused by mutations in the genes. A combination of *in vitro* and *in vivo* assays can be used for these studies and no NOAEL can be derived since they are a category of carcinogens with no-threshold effects.
- **Neurotoxicity studies:** they evaluate the capacity of compounds to affect the structure, function and development of the nervous system in both acute and repeated dose effects.
- **Toxicokinetic studies:** they inform about the fate of a compound and its metabolites when it enters the body of an organism, in particular regarding the 4 following processes: absorption, distribution, metabolism and excretion (**ADME**). Usually, a single dose administration (and two dose levels) is sufficient to measure the ADME properties. Information from toxicokinetic studies help for further dose selection for long-term studies and contribute to the **animal-to-human extrapolation** for risk assessment.

1.2.4 Classification and labelling

For most of these studies, a classification of chemicals can be performed according to the **Globally Harmonized System of Classification and Labelling of Chemicals** (GSH) which has been developed by the United Nations in order to standardize and harmonize classification and labelling of chemicals [92]. GHS is an international voluntary system which can be adopted by countries and adapted to their own regulations. In the EU, the Regulation (EC) No 1272/2008 on the **Classification, Labelling and Packaging** of Substances and Mixtures (CLP regulation) [83] constitutes the reference and requires that new substance category, classification and labelling should be notified to ECHA before their marketing. The different classes are related to physical hazards (*e.g.* explosives, flammable, corrosive, *etc.*), health hazards (*e.g.* acute toxicity, mutagenicity, carcinogenicity, *etc.*) or environmental hazards (*e.g.* aquatic toxicity, hazard to the ozone layer). Each hazard class is divided into numbered categories informing about the severity of the hazard with a category of 1 corresponding to the most severe hazard. For example the carcinogenicity class is divided into 3 categories:

- **1A:** Substances known to have carcinogenic potential for humans, based on human evidence
- **1B:** Substances presumed to have carcinogenic potential for humans, based on animal evidence (experimental animal data)
- **2:** Suspected human carcinogens, based on evidence obtained from human and/or animal studies but not sufficiently convincing to place the substance in Categories 1A and 1B.

Then, once a chemical has been assigned to a class and a category, the labeling elements that should appear on its packaging can be determined. They contain signal words, hazard pictograms (among 9), hazard statements and precautionary statements. Moreover, a Safety Data Sheet should be provided to give more advice and safety precaution for the use of the substance.

Among the health hazard classes of the CLP are the carcinogenicity, mutagenicity and toxicity for the reproduction which represent the most hazardous ones and therefore the most highly concerned by regulators. The three classes are divided into the categories 1A, 1B and 2 and compounds inducing at least one of the three adverse effects are classified as **CMRs** (Carcinogens, Mutagens, Reprotoxic). CMR substances benefit of a particular interest in the risk assessment since they should be avoided and replaced as much as possible except if their exposure has been determined as "negligible" (only for categories 1B and 2).

1.2.5 Assessing the relevance to human: MoA studies

The studies described previously are used for descriptive toxicology which aims at producing data to inform about the potential hazard of compounds by identifying adverse outcomes. But the use of **mechanistic studies** is also required in the risk assessment process in order to understand the nature of the hazard and its relevance to humans. Basically, descriptive studies identify "what" happens and mechanistic studies describe "how" it happens. Indeed, mechanistic studies aim at identifying and evaluating the biological processes leading to an adverse effect, called the **Mode Of Action (MOA)**. This term has been originally defined by the International Program on Chemical Safety (IPCS) from the World Health Organization (WHO) in a conceptual framework for evaluating the MOA of chemical carcinogens [244]. Later on, the concept of **Adverse Outcome Pathways (AOP)** has been proposed to describe the existing knowledge regarding the causal links of a sequence of biological events that lead to an adverse effect [8]. These two concepts are highly similar and we use them interchangeably in the following.

The AOP concept is illustrated in Figure 1.3. It starts with a biological event called the **Molecular Initiating Event (MIE)** that corresponds to the chemical interaction between the compound and its biological target and leads to an Adverse Outcomes (AO) through intermediate events called **Key Events (KE)**. An AOP begins at the molecular level and goes up to the individual or the population level through key events at the cellular, organ and organism levels. The key events are required but often not sufficient to induce the adverse outcome in the absence of other key events. Moreover, key event relationships (KER) should describe the causal relationships between the key events. This causal link is demonstrated based on the modified Bradford-Hill criteria and using a weight of evidence approach as proposed by Meek *et al.* [186]. Key criteria are the dose and time concordances as well as reversibility evaluations. The entire elucidation of an AOP (MOA) is performed using a combination of both *in vitro* and *in vivo* assays. It is important to keep in mind that several AOPs or MOAs can lead to a specific adverse outcome through different MIEs and KEs.

In 2012, the OECD launched a program for the development of AOPs in order to make them available in a web-based platform so that all the knowledge from the community can be brought together. This resource is called AOP-KB [267].

Once a MOA has been elucidated for a specific toxic outcome, its relevance to human can be assessed according to the IPCS Human Relevance Framework [27]. This evaluation of the relevance to humans then helps to perform risk assessment, classification and labeling. A compound inducing an adverse effect through a non relevant to human MOA is assumed to be "safe".

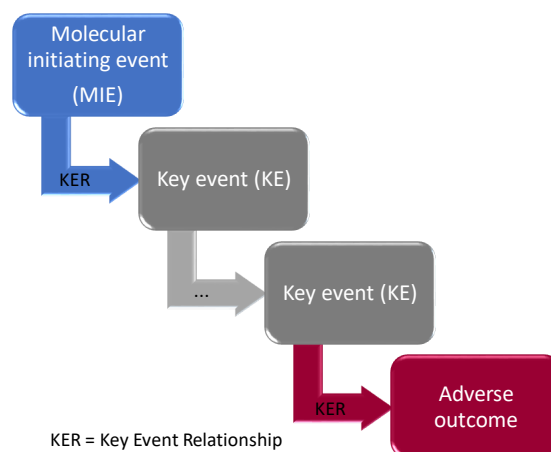


Figure 1.3: Schema of the Adverse Outcome Pathway concept. An AOP is composed of a Molecular Initiating Event (MIE) and several Key Events (KE) leading to an adverse outcome. Causal relationships between the events are described by Key Event Relationships (KER).

Figure 1.4 represents the postulated AOP for liver-mediated thyroid tumors leading to follicular cell adenomas in the thyroid of male mice [223]. This AOP starts with the activation of the Constitutive Androstane Receptor (CAR) or the Pregnane X Receptor (PXR) in the liver (MIE) which induces metabolic enzymes of the phase II in the liver (KE1) and consequently a changes in circulating concentration of thyroid hormones (KE2): increase of the Thyroid Stimulation Hormone (TSH) and decrease of the Thyroxine 4 (T_4). At the cellular level, this is translated by an increased cellular proliferation (KE3) and a follicular cell hyperplasia (KE4) finally leading to the development of tumors. Regarding the relevance to humans, the literature provides evidence that the process involving hepatic activity on thyroid hormones and leading to thyroid tumors in rodents is not likely to induce tumor development in humans [62]. This AOP is therefore probably not relevant to humans.

Beyond providing mechanistic information for the evaluation of relevance to humans in a regulatory context, the AOP concept offers several advantages to the entire toxicology community. In particular, they allow the collection of mechanistic information at different biological levels, the linkage of mechanistic events to adverse outcomes or can be used for the design of new *in vitro* methods and computational models for toxicity prediction [164].

1.2.6 The case of endocrine active chemicals

Since the 1990s, endocrine active chemicals, also called **endocrine disrupting chemicals (EDCs)** have raised a lot of interest and concern regarding their risk for human health and the envi-

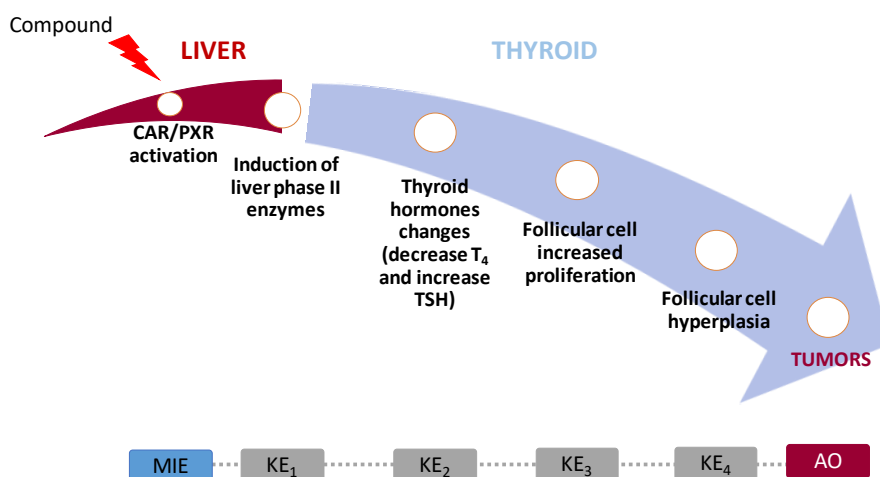


Figure 1.4: Example of Adverse Outcome Pathway: the liver-mediated thyroid tumors AOP leading to adenomas in thyroid of male mice [223]. The MIE corresponds to an activation of CAR/PXR hepatic nuclear receptors leading to the adverse outcome through three key events at the cellular and organ levels. CAR:Constitutive Androstane Receptor, PXR:Pregnane X Receptor, TSH:Thyroid Stimulation Hormone, T_4 : Thyroxine 4.

ronment [143]. An **endocrine disruptor** has been defined by the World Health Organization (WHO) as "an exogenous substance or mixture that alters function(s) of the endocrine system and consequently causes adverse health effects in an intact organism, or its progeny, or (sub)populations" [133]. Specifically, they interfere with metabolic processes, they bind to hormones' receptors and mimic their biological activity but lead to unwanted response and they also bind to transport proteins in blood and alter the hormones circulation. If sufficiently potent, these functional disruptions can lead to diverse adverse outcomes at the whole organism level such as developmental and reproductive effects, neurobehavioral troubles, immune disorders or cancers [232]. There are numerous mechanistic pathways that result in these effects, including activation of nuclear receptors (*e.g.* Estrogen Receptor (ER), Androgen Receptor (AR)), alteration of steroid pathway enzymes and neurotransmitter receptors [64]. The rationale for which EDCs are of high concern is that their mechanisms of action are not threshold dependent: their effect are not necessarily observed above a certain dose and can even appear at low doses, resulting in non-monotonic dose response curves as illustrated in Figure 1.5 [265]. This prevents the use of traditional toxicity studies to characterize their hazard. Therefore, their risk assessment and regulation should be based on mechanistic studies rather than descriptive toxicity.

EDCs represent a broad variety of chemicals, ranging from natural such as mycotoxins and phytoestrogens, to synthetic such as drugs, household products, plastics and PPPs. However, as of today, most of the EDCs are still lacking assessment regarding their potential endocrine activity.

In 1999, the US EPA created the **Endocrine Disruptor Screening Program (EDSP)** in order to screen PPPs and environmental chemicals for their potential to affect the endocrine sys-

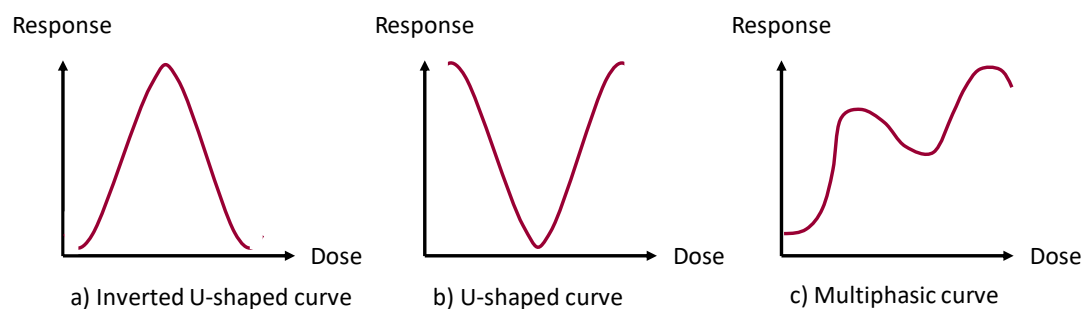


Figure 1.5: Examples of non-monotonic response curves that can be obtained with EDCs. a) Inverted U-shaped curve, b) U-shaped curve, c) Multiphasic curve

tems [88]. The program is based on a two-tiered approach proposing a battery of *in vitro* and *in vivo* assays to detect estrogen and androgen related effects induced by various MOA. On the one hand, the tier 1 is composed of five *in vitro* tests (estrogen and androgen receptor binding, estrogen transcriptional activation, aromatase and steroidogenesis activity) and six *in vivo* assays (rat uterotrophic assay, rat Hershberger assay, rat male and female pubertal assays, amphibian metamorphosis assay and short-term reproduction in fish assay). On the other hand, the tier 2 includes *in vivo* multi-generational studies which have the goal to further characterize the compounds that were identified as active in the Tier 1 assays. The actual testing of compounds started in 2009 with a first list of 67 compounds followed in 2010 by a second list of 109 compounds.

Following the launch of the EDSP program, the OECD developed in 2002 a conceptual framework that provides an approach built on several tiers for testing and assessment of endocrine-disrupting potential of chemicals [196]. The initial level is based on existing information and aim at prioritizing compounds for further testing in the following levels that include more and more biological complexity by starting with *in vitro* assays and going to more complex *in vivo* studies.

In the EU, according to the Regulation 1107/2009, a compound with endocrine disrupting properties cannot be considered and is completely banned from the market. Nonetheless, before 2018, there was no existing criteria to define the potential endocrine effects induced by chemicals. Indeed, since 2009 several rules and guidances for the evaluation of the endocrine disrupting potential of PPPs have been published but there were a lot of uncertainties. In 2013, EFSA proposed a draft criteria to define EDC that is based on 3 requirements [53]:

- The presence of an adverse effect in an intact organism,
- The presence of an endocrine activity,
- A plausible causal relationship between the previous two.

With this definition, endocrine disruption is considered as a mode of action and not a simple adverse effect and it therefore requires mechanistic studies to be demonstrated. Moreover, EFSA proposed that EDC could be divided into two categories analogous to the CMR classification (known or presumed and suspected ED).

Finally, it was only in 2018 that the ED scientific identification criteria for PPPs has been officially

published by the European commission in a new directive with a new EFSA/ECHA guidance document [75]. The novelty compared to other adverse effects is that regulation of EDCs is only based on hazard and does not take into account exposure. Nonetheless, the guidance focuses on ED effects caused by **estrogenic, androgenic, thyroidal and steroidogenic (EATS)** pathways because they are the ones for which there is a good mechanistic understanding and for which OECD *in vivo* and *in vitro* test guidelines exist.

1.3 The need for alternative methods

1.3.1 The rationale

The current risk assessment process and related toxicology studies raise several concerns. First, the approach is expensive and time consuming and requires a high number of laboratory animals which is an important ethical issue. Regarding this last issue, the **Three R (3R)** rule has been proposed in 1959 and presents three principles for a more ethical use of laboratory animals: replacement (find new methods to avoid animal use), reduction (use of methods that enable the use of fewer animals for the same quality and quantity of information) and refinement (use of methods that increase animal welfare and reduce animal pain and stress) [225].

Moreover, the number of compounds that need to be tested is always increasing and there are a lot of chemicals for which little or no toxicity data are available (including PPPs). This increase is partly due to the new directives and initiatives for a better toxicity testing, such as the REACH regulation which asks for more information for an important number of chemicals that are currently lacking of data. Indeed, an estimation of the number of chemicals falling under REACH ranged from 68,000 to 100,000 corresponding to 54 million of laboratory animals and a testing cost of 9.5 billion euros [224]. Unfortunately, the traditional approaches cannot deal with this high need for testing.

Lastly, because data are obtained using laboratory animal studies on which high doses of chemicals are tested, they may not be representative of the actual risk to human health [155]. Indeed, if the hazard is successfully characterized in those studies, the extrapolation of its risk to other species and life stages is not evident which raises the question of concentration relevance.

According to this situation, there is an urgent need for new alternative methods that would help to face all the issues encountered with the traditional approaches. Ideally, these methods should be more predictive, more reliable, faster, cheaper and provide information about the MOA of chemicals. The entire community knows that the development of new methods can take years and that the methods will raise uncertainties which should be identified in order to determine whether the new methods are more reliable than the actual ones. Since the last two decades, efforts are made in that sense and the goal is to move from a risk assessment based on animal hazard data towards human based *in vitro* assays in which compounds are tested at relevant concentrations. This "shift" is further illustrated in the next section.

1.3.2 How to address the need?

In 2007, the National Research Council (NRC) of the National Academy of Science in the US published a report entitled "Toxicity Testing in the 21st Century: A Vision and a Strategy" [55]. This report proposes a paradigm shift in toxicology by suggesting an increased use of *in vitro* and *in silico* methods for toxicity assessment. In particular, these methods should help to [6]:

- Increase the number of tested chemicals,
- Decrease the time and cost of testing,
- Significantly reduce animal testing,
- Increase the ability to determine the human risks of environmental chemicals by incorporating data on the mode of action and information on tissue doses and human exposure.

In 2008, following the release of the NRC report, the **Toxicology in the 21st century (Tox21)** American consortium has been created. This federal collaboration gathers the US EPA, the National Institute of Health (NIH) (and specifically its National Center for Advancing Translation Sciences, NCATS), the National Toxicology Program (NTP) from the National Institute of Environmental Health Sciences (NIEHS) and the Food and Drug Administration (FDA). The objective of this collaboration is to develop methods to "rapidly and efficiently evaluate the safety of commercial chemicals, pesticides, food additives or contaminants and medical products". Specifically, these methods should help at identifying chemically-induced biological activity, prioritizing compounds for further and deeper testing and developing predictive models of *in vivo* toxicity.

Basically, this Tox21 vision wants to base the risk assessment on the elucidation of toxicity pathways which were defined as "cellular response pathways that, when sufficiently perturbed, would be expected to result in adverse health effects" [55]. Compared to the AOP, the toxicity pathway only considers chemical, macromolecular and cellular levels. In order to elucidate these toxicity pathways, several alternative methods can be envisioned such as *in vitro* High-Throughput Screening (HTS), functional genomics and computational methodologies [6].

HTS: **High-Throughput Screening** techniques allow the testing of several thousands of compounds in several *in vitro* assays in parallel. This can be achieved by the use of robotic platforms that perform standardized protocols of miniaturized biological assays. In the end, HTS enables a rapid screening of chemicals and their prioritization for further testing, at lower cost than animal testing. Assays can be performed cell free (in a biochemical medium containing molecular entities) or can be cell based (in isolated cell lines) using one to several concentrations of compounds and enable measuring the ability of compounds to interact with proteins, genes or cells and therefore affect their functions. More precisely, these assays generally measure the binding of chemicals to receptors, the gene expression levels using reporter genes, the activation or inhibition of enzymatic activity and effects in cells such as cytotoxicity or changes in cell size and

shape. Thus they help in identifying the cellular responses induced by chemicals by evaluating several pathways and triggering MIE and KE of specific AOPs [141]. They potentially also have the advantage of avoiding the animal to human extrapolation since they use human cell lines and concentrations of compounds that are relevant to actual levels of human exposure [222].

Functional genomics: The term "omics" refers to all the fields of biology ending with the suffix *-omics* such as genomics, transcriptomics, proteomics or metabolomics. In "omics" studies, large number of biological entities (genes, transcripts, proteins, metabolites) are measured and analyzed to describe and understand the biological functions and dynamics of organisms. These measures can be performed after either *in vitro* or *in vivo* experiments.

In particular, functional genomics tries to link genes-related data to biological outcomes by analyzing a large panel of genes and proteins and enabling the generation of gene expression profiles. Functional genomics can provide interesting and important information about toxicity pathways and AOPs by looking at the perturbed genes after the exposure to chemicals. Such as *in vitro* assays, genomics can be performed in high-throughput which results in a large amount of data. Therefore, bioinformatics tools are required to perform data analysis and interpretation.

Computational methods: They can be used for two distinct tasks: (1) data analysis and integration or (2) development of computational models for biological simulations and predictions. Regardless of the considered task, computational methods play a significant role since they enable the integration of all the generated data which are then used to build predictive models. This would help to identify toxicity pathways and AOPs and also to prioritize compounds that would need further testing for risk assessment. Among computational tools we can cite **machine learning** (including Quantitative Structure-Activity Relationship, QSAR), grouping and read-across or physiologically based toxicokinetic models (PBTK). These methods are further detailed in Chapter 3.

These alternative methodologies are intended to be used in **Integrated Approaches to Testing and Assessment (IATA)** which are defined as approaches for chemical safety assessment based on the integration of data from various methods and sources [199]. For example they can incorporate *in vitro* and *in vivo* data as well as computational methods for either new data generation or existing data interpretation and integration. In the end, they aim at helping in the development and understanding of AOPs and MOAs.

In line with the Tox21 vision, initiatives have been launched during the last decades in order to seek for alternative methods and several entities have been created:

- The International Cooperation on Alternative Test Methods (ICATM) has been created in 2009 in order to promote the international cooperation for the development and validation of new alternative approaches in order to support the 3R rule. It gathers several govern-

mental organizations from Europe, US, Canada, Japan, Korea, China, Brazil. In Europe, it includes the European Union Reference Laboratory for alternatives to animal testing (EURL ECVAM) which has already validated alternative methods for acute toxicity, eye irritation, genotoxicity, skin sensitization, skin irritation, *etc.*

- In the US, the **Tox21 program** have been launched with the goal to generate HTS data for a large library of compounds. The **ToxCast program** is the US EPA contribution to the Tox21 program [68]. These two programs are further detailed in Chapter 3, Section 3.2.2. The US EPA also launched the **ExpoCast program** in complement to the ToxCast one in order to gather exposure information thanks to the development of high-throughput methods that estimate exposure [131].
- In Europe, REACH promotes the use of alternative methods for the hazard assessment of chemical substances in order to reduce the number of tests on animals. The legislation states that laboratory animal studies should be performed only as a last option and that the use of IATA strategies should be envisioned such as eliminating duplicate studies, promoting data sharing and using existing data [82].
Still in Europe, the **Safety Evaluation Ultimately Replacing Animal Testing (SEURAT-1)** initiative has been launched in 2013 after the banning of cosmetic products containing ingredients tested on animals. It was composed of 7 projects and aimed at looking for new alternative methods to help in the regulatory risk assessment [103]. Following the SEURAT project, the **EU-ToxRisk** initiative has been launched in 2016 and gathers 38 European partners (and one university from the US) with the goal to develop a "new way of risk assessment" [58]. In particular, the objective is to use IATA comprising *in vitro* assays, omics technologies and computational tools in order to get a mechanistic understanding of adverse effects induced by chemicals.
- In Japan, the Toxicogenomics Project had the objective to use omics technologies in order to generate gene expression data and identify mechanisms leading to adverse effects as biomarkers of toxicity and therefore help in risk assessment [263].

1.4 Position of the thesis in this context

According to the current toxicological context, the objective of this thesis is to evaluate how the public data generated for toxicity assessment could be exploited by computational tools to help in the prioritization of compounds and to support the 3R principle. This is in line with the Tox21 vision but it is also responding to the objective of agrochemical companies such as Bayer. The ideal would be to predict *in vivo* outcomes caused by compounds directly from their chemical structure but due to the large number of events happening between the entrance of the chemical in the organism and its effect, it seems quite ambitious and the consideration of other type of information is necessary.

Moreover, since the interest for alternative approaches is quite recent and that we are still in the early stage of the validation of these methods, we think it is important to test the relevance

of *in vitro* derived bioactivity signatures and their ability to predict *in vivo* outcomes through computational models.

We therefore propose to focus on three types of information: chemical structure of compounds, *in vitro* bioactivities generated in HTS and *in vivo* outcomes observed in toxicity studies and we distinguish two stages to predict *in vivo* outcomes from chemical structure. Indeed, in our two-stage approach described in the Introduction, the first stage aims at predicting *in vitro* results based on chemical structure information while the second stage aims at predicting *in vivo* outcomes based on *in vitro* data. To do so, we mostly use machine learning approaches whose generalities are described in the next Chapter.

CHAPTER 2

MACHINE LEARNING: GENERALITIES

This chapter aims at introducing or reminding the generalities and rationales of machine learning.

Artificial intelligence can be defined as the computational science which develops techniques to make the machines able to simulate intelligence and can result for instance, for the general public, in robots or expert systems. Machine learning (ML) is a specific field of artificial intelligence that enables computers to "learn" from existing data on which they are trained. Thanks to learning algorithms that are based on statistical approaches, ML allows the detection of patterns inside the data which are learnt in order to make predictions or decisions on new examples with regard to a specific **task**. We commonly say that algorithms generalize from their experience. The first ML algorithms were developed in the late 1950s and since the 21st century, ML has gained a big interest in many fields. Indeed, due to the emergence and availability of big data on which it is difficult to perform comprehensive analysis using simple statistical methods, computer scientists have focused on ML methods and the development of effective algorithms. Today, a lot of companies are relying on big data and ML to provide competitive products and services. Examples of applications of ML include image processing, face recognition, natural language processing, object detection, financial analysis, medical diagnosis, *etc.*

After briefly having presented the principle of machine learning, we describe in more details the different required steps to generate a model starting with data processing, followed by the learning and finally the evaluation. The last section of this chapter focuses on methods that enable the handling of imbalanced datasets.

2.1 Principle of machine learning

In order to design a ML approach, several elements are required:

- **A task to complete:** it corresponds to the problem that the ML model should solve and we essentially distinguish three main types of task: classification, regression and clustering.

A description of these tasks is provided in Section 2.3.1.

- **Data:** they are essential to build ML models and their quality is directly linked to the performance of the future model. They are used as input to train the algorithm (training set) and test the resulting model (testing set). They are characterized by several input variables also named **descriptors** or **features** and sometimes by a desired output. Most of the time, data are not ready for machine learning and need to be pre-processed [292]. Data can be of various types such as numbers, words, images, sounds, *etc.*
- **A learning algorithm:** this is the mathematical method that learns from the training set. Numerous ML algorithms exist and their choice mostly depends on the type of data used as input as well as the task to complete.
- **Performance metrics:** they are used to evaluate the model's performance and they also differ according to the considered tasks. More details are provided in Section 2.4.2.

The machine learning process can be divided into two phases: the learning and the task completion (see Figure 2.1).

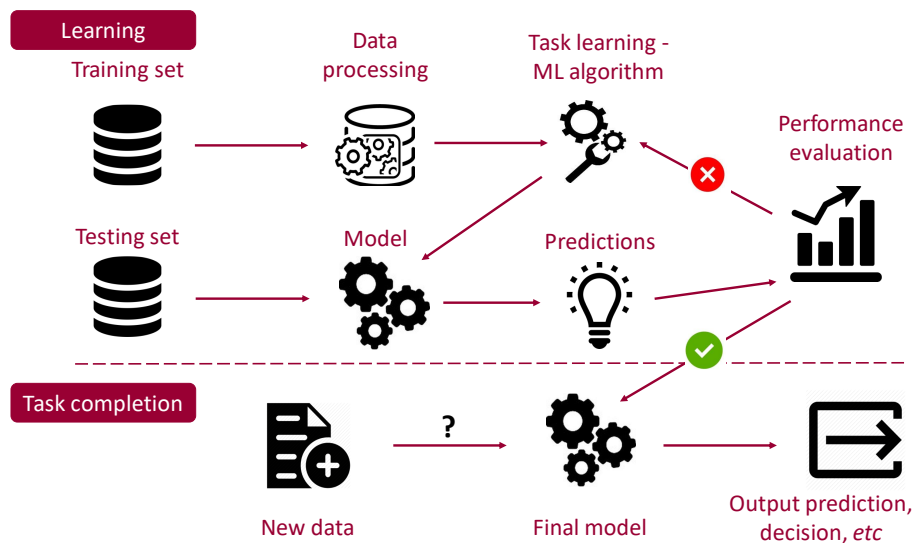


Figure 2.1: Machine learning rationale. **Learning:** based on a training set on which is firstly performed data processing, a ML algorithm tries to learn how to give good answers for a considered task. Performance of the resulting model is evaluated on a testing set and the model is retrained until it reaches acceptable performance. **Task completion:** once the model has sufficient performance, it can be used to complete the specific task for new observations.

1. During the **learning phase**, an algorithm is trained to build a mathematical model that will be able to give answers to complete a considered task. The learning is performed using a **training dataset** composed of input data called **examples** or **observations**. Performance of the model is evaluated during this phase using a **testing set** composed of examples for which we already know the answer to the considered task but that were not included in the training dataset. If performance is sufficient, the model can be used in the next phase, otherwise it is retrained with other hyperparameters (*i.e* parameters controlling the

learning algorithm itself) until it reaches the desired performance; this is called parameter tuning.

2. Once the model has sufficient performance, it can enter the second phase which consists of applying the model on new data in order to complete the considered task.

2.2 Dataset processing

Before entering the learning step, data need to be pre-processed in order to make them handable by the learning algorithms. Indeed, raw data are generally not in a proper format to be read by ML algorithms and they can contain missing, incorrect and noisy values as well as duplicate observations. Moreover, the input data need to be relevant for the task to complete, meaning that the features used should sufficiently describe the problem. Data processing is therefore crucial for the development of good ML models since the quality of data is directly linked to the quality of the resulting model. In general, this step requires the largest amount of time and effort from the scientists [292].

We can distinguish different steps in the data processing:

Merging and formatting the data: Sometimes data are extracted from different sources and need to be aggregated into a unique and consistent format that is readable by the learning algorithm.

Removing of duplicate observations: Depending on the problem to model, having a unique version of each observation can be important to avoid the induction of noise in the model. Indeed, the training data have to be representative of the real world and in some cases, if observations are repeated several times, the model will base its learning in favor of these observations with the risk of giving less consideration on the others.

Removing of outliers: Extreme values are generally not representative of the data and can also affect the performance of the model in the same way that duplicate observations. Here again it depends on the modeling problem and in particular on the information carried by the outliers: if an outlier comes from errors in the data it is safer to remove it but if the outlier is a real one it can have a lot of meaning and may be important to be considered by the model. Several methods exist to automatically detect outliers and therefore proceed to a case by case evaluation of their importance [126].

Dealing with missing values: There are different methods to handle missing values in the input features. Basically, they can be either removed from the dataset or replaced by another value which can be the mean or median of all the other values of the given feature or a computed guess made by more complex algorithms [12].

Scaling of features: When features have different units and various scales, the learning algorithm might give more consideration on the ones with the largest scale. To avoid this and make all the features to contribute equally to the model, feature scaling can be used and the different methods include: standardization (also called Z-score normalization, it rescales the features such that, after treatment, their mean equals 0 and their variance equals 1), min-max normalization (it

rescales the range of the features into the interval $[0,1]$ or $[-1,1]$) and decimal point normalization (it transforms the values of the features by moving the decimal point according to the number of digits of the maximal value; if the maximal value has 3 digits, all the values are divided by 10^3) [236].

Feature selection: It aims at removing irrelevant and redundant features in order to reduce the dimensionality of the feature space and hence avoid the **curse of dimensionality** which states that *the more the features in the dataset, the more observations are required to learn correctly*. In particular, this relation is exponential: in one dimension, if the observations cover 10% of the space, in two dimensions the same data will cover 1% of the space and only 0.1% of a three dimensional space (see Figure 2.2 for an illustration). Several approaches can be used to reduce the number of features: first, features with a low variance can be removed since they do not discriminate the different observations. Next, redundancy is avoided by removing one of two highly correlated features. Then, several algorithms can be used to keep the most relevant features.

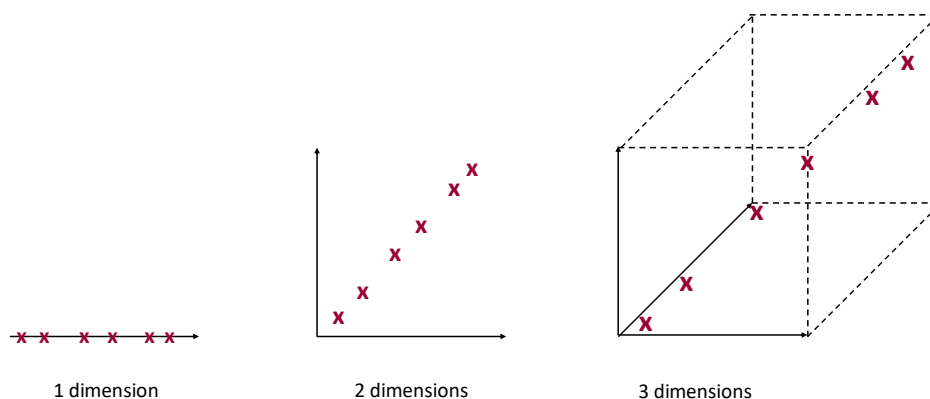


Figure 2.2: The curse of dimensionality. In one dimension, if the observations cover 10% of the space, in two dimensions the same data will cover 1% of the space and only 0.1% of a three dimensional space.

These algorithms can be divided into three categories [111, 254]:

- **Filter methods:** they rank the features according to a statistical test (Chi-squared, correlation coefficient, F-score *etc.*) and keep the best ones [138]. The test is performed individually between each feature and the output, before the learning phase.
- **Wrapper methods:** they are used during the learning phase and aim at selecting the best set of features by testing different combinations during the training [111]. Several methods can be applied to generate the set of features such as forward selection (starts with one feature and add the best significant one at each iteration, stops when performance is not improved anymore), backward elimination (starts with all features and removes the least significant at each iteration, stops when performance is affected) or recursive feature elimination (starts with all features and estimates their importance after learning the model to remove the least important at each turn, stops when performance is affected).

- **Embedded methods:** also used during the learning phase, they combine the advantages of the two previous methods by learning which features are contributing the best to the performance of the model [175]. They are generally preferred to the wrapper methods since they are less computationally costly. Some learning algorithms such as decision trees and Naive Bayes (see Section 2.3.2) already include a built-in mechanism to perform this type of feature selection. Apart from them, the most common algorithms are based on regularization methods, also called penalization methods, that are performing feature weighting. Indeed, by using an objective function (or penalty function) that minimizes the errors of the predictions, they assign a small (or close to zero) coefficient to the features that do not contribute to the model: the features are therefore penalized since their weight is low and they are finally removed from the dataset.

2.3 Learning algorithms

2.3.1 Types of learning

Once the data have been processed and are ready for learning, they can be fed to the learning algorithm [3]. Different types of learning exist according to the input data.

Supervised learning: In that case each training example is labeled with an output (already known) and the algorithm is trained on the input-output pairs of data and build a mathematical model able to predict the unknown output for a new example. In other words, the algorithm tries to map inputs to outputs by inferring a mathematical function [63].

Two tasks can be achieved by supervised learning algorithms: classification and regression. On the one hand, the **classification** is used when the desired output is a discrete variable representing a category or a class (in the case of two classes, we talk about binary classification). On the other hand, the **regression** is used for approximation of continuous output.

Unsupervised learning: In that case the desired output is unknown and the algorithm looks for patterns in the data in order to group them into categories according to similarities and differences in the features. Contrary to the supervised learning which aims at predicting an output for a new observation, the unsupervised model tries to extract general rules that explain relationships between the input variables [3].

The most common tasks of unsupervised learning include **clustering** [200] (*i.e.* grouping of observations into clusters according to their similarity), **dimensionality reduction** [264] (*i.e.* reducing the number of input variables to a few principal components), and **association rules learning** [286] (*i.e.* finding relationships between the descriptive features).

Semi-supervised learning: It is a mix between the two previous types of learning since one part of the input examples is labeled and the other part is not. The model should learn both the structure of the data and the output of labeled data in order to be able to make predictions [41].

Reinforcement learning: Here, the algorithm learns how to behave from its experiences in order to optimize a quantitative reward over time. Each time, it has to make decisions according to its current environment and receives a positive or negative reward. In the end, the goal of the algorithm is to maximize the sum of the reward obtained at each trial meaning that it has learnt the best behavior [250].

Since in this work we only use labeled data in order to predict different types of output, the next section will only focus on supervised learning.

2.3.2 Methods and types of algorithms for supervised learning

Learning algorithms can be classified according to several criteria such as the type of task they perform (classification, regression) or the method on which they are based. Here is a list of the most commonly used methods.

Regression algorithms: They model the relationship between the input variables and the output and try to minimize the predictions error. The model is iteratively refined according to error of predictions. Examples of regression algorithms are the **linear regression** which tries to fit the best linear equation of the form $f(x) = a \times x + b$ between the variables [11] (the linear regression can be simple or multiple according to the number of variables); and the **logistic regression** [70, 201] which is a classification method that learns to fit a sigmoid function of type $f(x) = \frac{1}{1+e^{-x}}$. It returns a value corresponding to the probability of an observation to belong to a class. This predicted probability is then transformed into a predicted binary value according to a threshold (generally 0.5)¹.

Bayesian algorithms: They are mostly used for classification tasks and are based on the Bayes' Theorem [152] which proposes a formula to calculate conditional probabilities. In the case of Bayesian algorithms, the posterior probability of the class c given the observation (input value) x ($P(c|x)$) is computed according to the following equation:

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)}$$

Bayesian classifiers make the strong assumption that all features are independent such that, for each feature x_i of a new observation and each class c_j , the posterior probability $P(c_j|x_i)$ of the class given the feature can be computed according to the previous equation. Then, the probability of the class given all the features $P(c_j|x_1, \dots, x_n)$ is computed as the product of all independent posterior probabilities:

$$P(c_j|x_1, \dots, x_n) = \prod_{i=1}^n P(c_j|x_i)$$

¹Note that despite its name, the logistic regression is a classification method.

The class with the highest probability is finally assigned to the new observation. Bayesian algorithms have the advantage of being fast and highly scalable and the most famous one is the **Naïve Bayes** classifier [189, 154].

Kernel algorithms: They aim at transforming an original data space where data are not linearly separable into a higher dimensional space in order to go back to a linear problem where the data can be discriminated using a linear function. Hence the linear separation in the high dimensional space is equivalent to a non linear separation in the original space. To do so, the algorithms use the "kernel trick" which consists of applying a kernel function to transform and project the input data into the higher dimensional space. Since a computation of the coordinates of the data in the high dimensional space would be too energy consuming, kernel functions have been introduced to avoid computing the scalar product in the high dimensional space and to allow a computation of this product directly in the initial space [231]. Examples of kernel functions include polynomial, Gaussian, sigmoid and hyperbolic tangent functions. In the end, algorithms look for the optimal hyperplane separating the different classes in the higher dimensional space. Two known kernel algorithms are the **Support Vector Machine (SVM)** [192] which tries to find the hyperplane that maximizes the margin between the different classes and the **Linear Discriminant Analysis (LDA)** [166] which assumes that data are normally distributed whereas SVM does not make assumptions. An illustration of the representation of the data before and after applying a kernel function is provided in Figure 2.3.

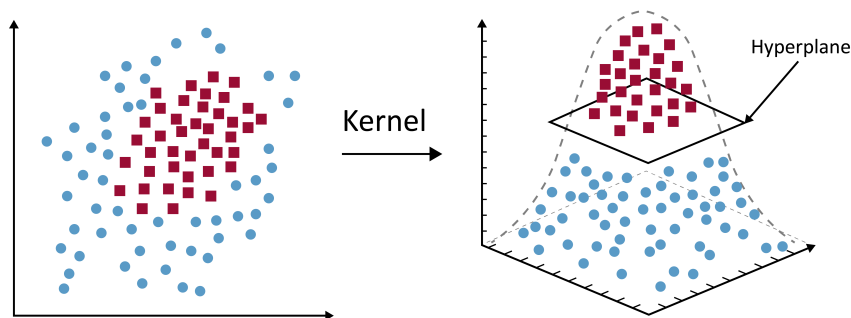


Figure 2.3: Example of the application of a kernel function to linearly separate the data. Left: representation of the data in their original space, dots and squares cannot be linearly discriminated. Right: representation of the data in a higher dimension space when a kernel function has been applied. Dots and squares are separable using a hyperplane.

Instance based algorithms: They compare new data to one or some of the most similar instances of the training set to make their predictions. The most popular instance based algorithm is the **k-Nearest Neighbor (kNN)** which looks for the k-Nearest Neighbors of a new data point by measuring a distance (Euclidean, Manhattan, Jaccard, *etc.*) and assigns their majority vote to the new observation [63]. An illustration of a kNN classifier is given in Figure 2.4.

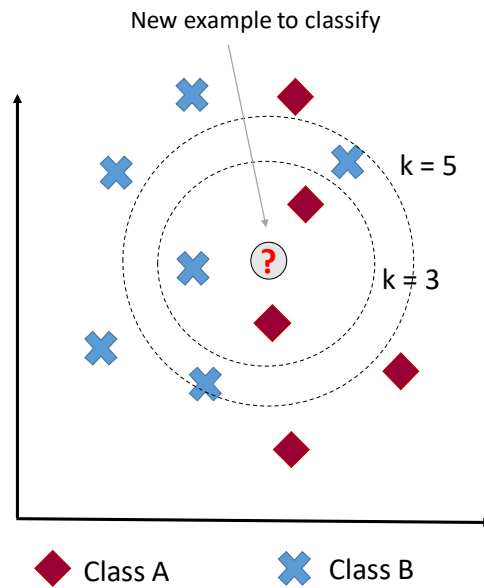


Figure 2.4: *Illustration of a k NN classifier.* When $k=3$, two neighbors of the new example belong to the class A and the other one belongs to the class B: class A is therefore assigned to the observation. When $k=5$, class B is assigned since 3 neighbors are belonging to this class (versus 2 for class A).

Decision tree algorithms: As indicated by their name, they are based on a tree where the nodes correspond to a condition regarding the input variables of the data and the leaves represent the final decision (*i.e.* the prediction). A simple example of decision tree is provided in Figure 2.5. The conditional rules in the nodes are generated based on conditional probabilities computed on the training data. At each node, in order to choose the feature which best splits the observations, algorithms use different criteria that measure the homogeneity of the target output in the possible subsets [63, 154]. Decision trees are fast, accurate and easily interpretable which places them among the favorite methods. Examples of decision tree algorithms are the **Classification and Regression Tree (CART)** and the C5.0 which differ by the criteria they use [171] (Gini impurity for CART and information gain for C5.0).

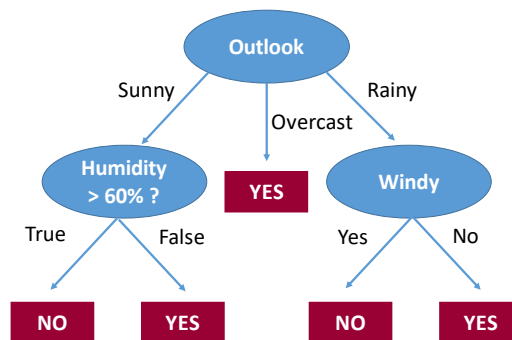


Figure 2.5: *Example of small decision tree that decides if playing tennis is a good idea according to the weather.* Here three descriptors are considered to make the decision: the first one is the outlook and according to it, humidity and wind can also be considered.

Artificial Neural Networks (ANN) algorithms: Inspired from the biological neurons, they are composed of different layers of interconnected nodes (neurons) including an input layer which receives input data, one to several hidden layers and one output layer which returns the predicted output (see Figure 2.6-a) [63]. Each node receives information (values) from the neurons of the previous layer and combine them into a value thanks to a propagation function which is most often a weighted sum of the inputs. Then, an activation function is applied to compute the output of the neuron after comparison to a threshold: below the threshold, the neuron is inactive (its output is generally 0 or -1) and above the threshold, the neuron is active (its output is generally 1). This output is then passed to the neurons of the next layer (*i.e.* the successor neurons). Among the most known activation functions are the linear ones (identity or sigmoid) used for example by the **perceptron** algorithm or the **Radial Basis Function (RBF)** used by the eponymous networks [154]. An illustration of a single neuron (the perceptron) is presented in Figure 2.6-b.

Each connexion between a neuron and its successor carries a weight which is modified and updated during the learning phase which aims at finding the parameters of the network that lead to the correct predictions [189]. To do so, a loss function mapping the predicted values to the actual ones (and therefore calculates the error predictions) has to be minimized. This is done by an algorithm from the gradient descent class in which the gradient of the loss function is computed using a technique called backpropagation. It allows the update of the weights of the network.

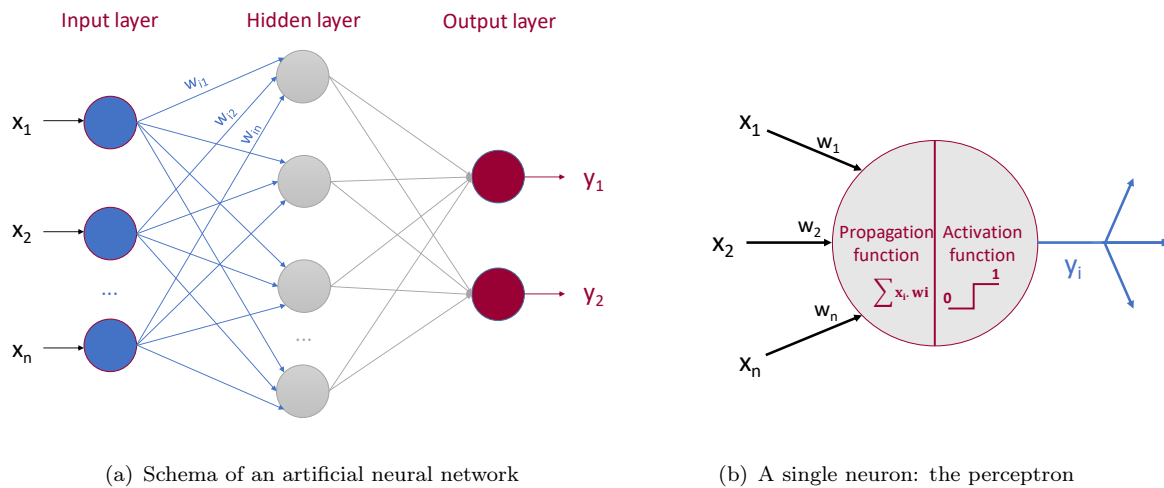


Figure 2.6: a) *Schema of an artificial neural network with the three types of layers: input, hidden and output. The input layer receives the input data and the output layer computes the final predicted output.* b) *Single neuron (perceptron): it receives input values which are combined by a propagation function and passed to an activation function to compute the output that will float to the next layers.*

In general, information in ANN flows from the input to the output without any feedback loop and these networks are called feed-forward networks [11].

Many types of ANN exist and differ according to their structure, their activation function, their

types of learning, *etc.* We can for example mention the family of **Recurrent Neural Networks (RecNN)**² which enables the modeling of sequential data [25, 162]. Indeed, besides computing the output for each time step of the sequence, the RecNN "memorizes" its internal state for this considered time step in order to feed it to the upstream neurons through recurrent connections to compute the output for the next time step. Another example is the class of **Random Neural Networks (RNN)** which are RecNN inspired by the spiking probabilistic behavior of biological neurons with a "product form solution" [98], integrating excitatory and inhibitory spiking signals resulting in respectively increase or decrease of values of the neurons receiving the spikes [99].

Deep learning: It is considered as an entire field of machine learning that uses neural networks with many hidden layers called Deep Neural Networks [162]. For several years, it has become the most popular method of machine learning and have proved its high efficiency in image and natural language processing, computer vision, speech recognition, *etc.* Networks used in deep learning are often referred as a "black box" because of the difficulty to know what is happening within the hidden nodes. Examples of architectures of deep neural networks are the RecNN (same than previously with a many hidden layers) and **Convolutional Neural Networks (CNN)** which include convolutional, pooling and fully connected layers and try to mimic the neurons of the visual cortex [163].

Ensemble algorithms: All the previously described methods allow the building of single models that result in one prediction. Indeed, each type of learning algorithm makes some hypotheses to predict, as well as possible, a particular output and this prediction is sometimes not obvious. The idea of ensemble algorithms is to take several models that have been independently trained to predict the same output in order to combine their hypotheses in a unique model. This combination finally leads to a consensus prediction that is more accurate than the ones of the single models [65].

Several methods combining the single models (also called **weak or base models**) have been developed:

- **Bagging:** in that case, multiple models are trained using different random subsets of the original training data and the prediction for a new observation is either the average of the predictions made by all the models (for regression tasks) or the majority vote of the predictions (for classification tasks) [30]; Figure 2.7-a) illustrates the method. This method is also called **bootstrapped aggregation** and leads to the decrease of the variance of the model [284], *i.e.* how much the predictions for a same observation differ from each other.
- **Random Forests:** it is a special form of bagging that combines multiple decision trees trained on different subsets of the training set. The difference with bagging is that a random selection of the features is performed to build each base model [31]. In other

²Note that Recurrent Neural Networks are commonly abbreviated in "RNN" but here we use the abbreviation RecNN to make the distinction with Random Neural Networks (RNN)

words, the choice of the feature that will split the data at each node is done randomly, from a random subset of the features, which avoids to always consider the same set of features. Figure 2.7-b) illustrates the Random Forest method.

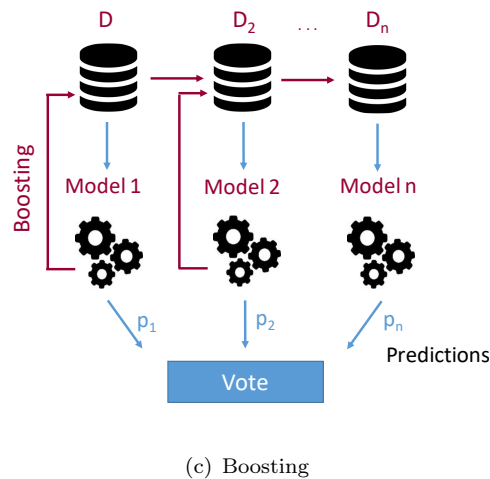
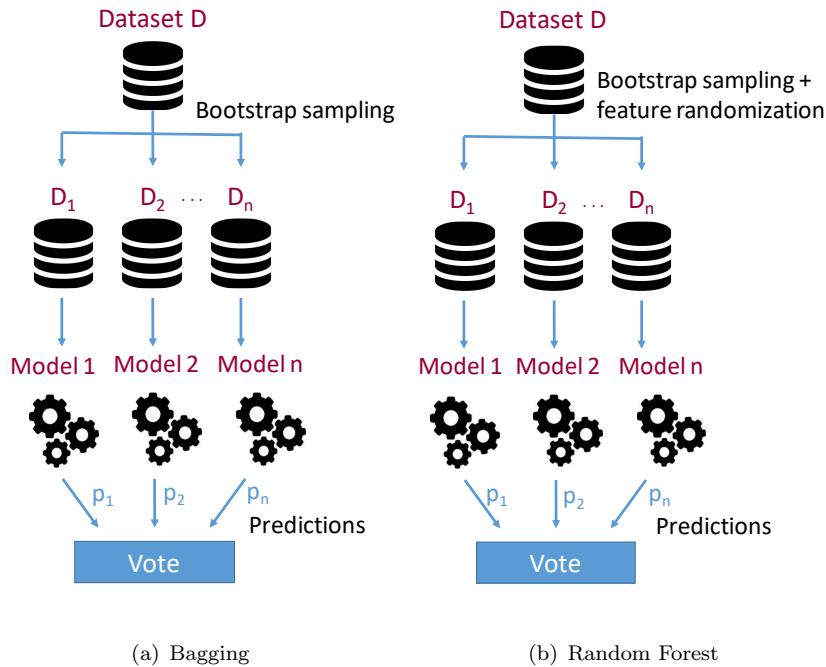


Figure 2.7: Schema of three ensemble methods. a) Bagging, b) Random Forests, c) Boosting

- **Boosting:** these algorithms aim at converting several weak learners into a single strong learner [65, 284]; the method is illustrated in Figure 2.7-c).

Basically, weak models are built sequentially and each new model is trained by considering the performance of predictions of the previous one. For example in the case of AdaBoost, the standard boosting algorithm [289], an important weight is assigned to each observation that has been mispredicted by the previous learner such that the actual one will focus more on difficult examples in order to correctly predict them: we say that these observations are

"boosted". Another well known boosting algorithm is the **gradient boosting** where each weak learner is trained on the remaining error from the previous model [95]. Indeed, this error is computed in a loss function which is minimized thanks to the gradient descent method.

Boosting methods allow the decrease of both bias and variance of the final model (see Section 2.4.2 for details).

- **Stacking**: also called **stacked generalization**, it combines the base models that learnt on a same subset of the training data into a **meta-classifier** which uses the predictions from the base models as input to generate an output [278]. Unlike bagging and boosting, the stacking method has the advantage of allowing the combination of models built using different types of learning algorithms and therefore takes advantage of their various underlying hypotheses. Stacking methods are more described in Chapter 5 since we apply them in the context of toxicological data.

Most of the algorithms described in this section have **hyperparameters** (*e.g.* the number of k-neighbors for the KNN, connection weights for ANN, *etc.*) that can be tuned in order to obtain the best model. This parameter tuning is performed during the validation step which aims at building the model with correct sufficient performance.

2.4 Model evaluation

In order to ensure that a model is good enough to be used for real cases and even choose the best model among several, it is essential to evaluate its performance during a validation step. This step allows the verification that a model generalizes well, meaning that it is able to provide good predictions for unseen data and not only for the data used to train it. Indeed, when a model performs very well on the training data but is not able to make good predictions for new data (*i.e.* the variance is high), we talk about **overfitting**. In other words, the model has learnt so much that it perfectly represents and fits the data it has seen to generalize to others. By contrast, **underfitting** corresponds to the fact that a model is too simple to be able to produce good predictions at all, whether on the training data or on new data. Therefore, the goal is to find a compromise for which the predictions on the training data are good and the generalization to new data is the best. This compromise is known as the **bias-variance trade-off**. Bias and variance are two different sources of error for ML models [101]: on the one hand the **bias** corresponds to the difference between predicted and actual values and reflects underfitting; on the other hand the **variance** is the variability of the model to predict a given observation. A high variance means that changes in the training data on which is trained a model will result in varying predictions for a same observation and this reflects overfitting. In general, models have low bias and high variance or conversely. The objective is to find the best balance where both bias and variance are low, meaning that the model neither overfits nor underfits.

2.4.1 Validation

We distinguish two types of validation: internal and external. In general, when no sufficient data is available, the original dataset is directly split into a training set used to train the model and on which the **internal validation** is performed and a test set which is never used for the learning and will enable to perform the **external validation** [189].

Internal validation: It is performed during the learning phase using the training data and can sometimes allow the tuning of hyperparameters of the algorithm in order to reach the best performance on these training data.

The most used technique is the **k-fold cross-validation** in which the training set is randomly split into k folds of equivalent size [189], as illustrated in Figure 2.8. At each turn, one fold is held out to serve as a test set and the $k-1$ remaining folds are used to train the model which predicts the outputs of the isolated fold. The process is repeated for each fold so that all the observations have been used once in a test set. In the end, the average of the performance obtained on the k -folds can be computed.

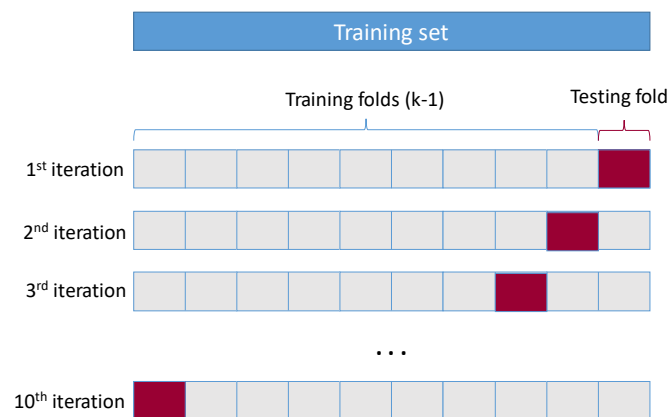


Figure 2.8: Schema of the cross-validation process, example with 10 folds. The entire training set is split into 10 distinct folds and at each step of the cross-validation, 9 folds are used to train the model while the remaining one is used as a test set to compute performance. The process is repeated as many times as the number of folds, here 10 times.

In the case of a classification task, it is recommended to perform **stratification** in order to ensure that the training and the test set have the same proportions of observations from the different classes. For example, considering a training set composed of 1000 observations with 20% belonging to a class A and 80% to a class B, in a 10-fold stratified cross-validation, each "held out" fold will contain 100 observations with 20 from the class A and 80 for the class B. The usual number of folds is 5 or 10 but a specific case of the k -fold cross-validation is the **leave-one-out cross-validation** which leaves out only one example at each turn and the model learns using all the remaining. One big disadvantage of this method is that it highly increases the computation time. Besides, the leave-one-cluster-out cross-validation is a particular case

which requires to first apply clustering: grouping observations into several clusters such that the observations within a same cluster are more similar to each other than to the observations in the other clusters. Then, in the leave-one-cluster-out cross-validation, an entire cluster of observations is used to test the model and evaluate how good the model can predict observations different to the ones in the training set.

Another method is called the **Monte Carlo cross-validation** where training and test data are independently and randomly partitioned in each run so that a same observation can be used in multiple test sets [282].

External validation: It is performed when a model is completely trained and it uses a test set which has not been used during the learning phase. It is the real validation since it mimics a real world scenario (evaluation of performance on unseen data) but it is sometimes not performed because of the unavailability of data.

Another useful method to test a ML model is the **y-scrambling, or y-randomization** [228]. In that case, the output value (y) of the training data are randomly reassigned to the input (x) such that there is no link between inputs and outputs. A model is built using this "random" training set and its performance is computed on the same test set than the real model. The performance of the two models are compared and if they are equivalent, it means that the real model is performing randomly.

2.4.2 Estimation of performance

After validation, performance metrics can be computed to evaluate the model. These metrics are different according to the task that is modeled.

Classification task: In the case of a binary classification problem, we use a confusion matrix as illustrated in Figure 2.9 where:

- **TP (True Positives)** is the number of examples predicted as positives which are actually positives
- **TN (True Negatives)** is the number of examples predicted as negatives which are actually negatives
- **FP (False Positives)** is the number of examples predicted as positives but which are actually negatives
- **FN (False Negatives)** is the number of examples predicted as negatives but which are actually positives

Note that most of the algorithms do not directly compute a binary value but rather a probability of belonging to a class. In order to obtain the binary value, they compare the computed probability to a decision threshold which is usually set to 0.5 by default. Thereby, if the probability is greater than this threshold, the prediction is "positive" and "negative" otherwise.

		Actual	
		1	0
Predicted	1	TP	FP
	0	FN	TN

Figure 2.9: Confusion matrix

Many performance metrics can be derived from this confusion matrix and we describe here the ones that will be used in the following [127]:

- The **sensitivity** (or recall) is the proportion of actual positives that have been predicted correctly:

$$\frac{TP}{TP + FN}$$

- The **specificity** is the proportion of actual negatives that have been predicted correctly:

$$\frac{TN}{TN + FP}$$

- The **Balanced Accuracy (BA)** is the mean between sensitivity and specificity:

$$\frac{\text{Sensitivity} + \text{Specificity}}{2}$$

- The **accuracy** is the proportion of correctly predicted examples among all:

$$\frac{TP + TN}{TP + TN + FP + FN}$$

Another commonly used metric is the **ROC score (or AUC ROC)** and corresponds to the value of the area under the Receiver Operating Characteristic (ROC) curve. This curve plots the *sensitivity* against $(1 - \text{specificity})$ (or False Positive Rate) such as illustrated in Figure 2.10. Basically, a ROC curve is a step curve which is built as follow: for each decision threshold ranging from 0 to 1 with a given step, the corresponding confusion matrix is determined after comparison with the probability computed by the model to the decision threshold³. Therefore, for a threshold of 0, since probabilities are positive, all the binary predicted values will equal to 1 (positive). Conversely, a threshold of 1 will result in only negative predicted values (0). Then, *sensitivity* and *specificity* are computed from the confusion matrix and plotted on the graph representing the *sensitivity* according to $(1 - \text{specificity})$. In the case of a decision threshold of 0, *sensitivity* and *specificity* respectively equal to 1 and 0 and conversely for a threshold of 1. They correspond to the 2 extreme points of the ROC curve. The process is repeated for various decision thresholds between 0 and 1 so that the curve can be drawn. The more numerous decision thresholds are, the more precise the curve is. Note that the decision threshold value is

³Note that only the binary predicted values change according to the decision threshold but not the predicted probabilities.

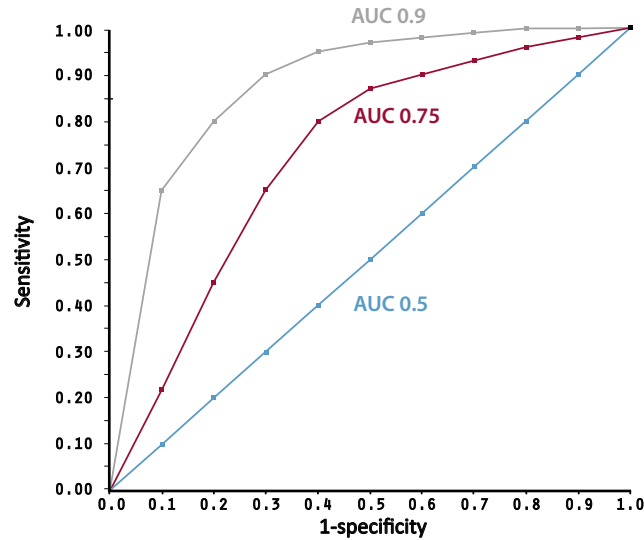


Figure 2.10: Example of ROC curves for three different models. The model resulting in the grey curve is better than the one resulting in the red curve but both are better than a random model. The blue curve is the identity line and corresponds to random predictions.

not visible on the plot. Finally, the AUC (Area Under the Curve) of the ROC is computed and correspond to the ROC score.

The closest this ROC score is to 1, the best the model is and a score of 0.5 means that the model is predicting randomly (*i.e* it corresponds to the identity line). In the example given in Figure 2.10, the blue line is the identity line which highlights randomness, the green and red curves correspond to two different models which are better than random models, with the model corresponding to the green curve being the best with a high ROC score (0.9).

Unlike the previous metrics, the ROC curve has the advantage of considering a model's performance obtained for different decision thresholds and not only one (0.5 most of the time) and therefore allows the identification (and even selection) of the best threshold depending on the desired performance. For example, if we want to get a high sensitivity despite a high specificity, we will probably select a threshold lower than 0.5 which would correspond to a point on the left part of the ROC.

Many other metrics can be computed from the confusion matrix, including the precision (also called Positive Predicted Value), the Negative Predicted Value, the Matthew's correlation coefficient (MCC) and the F-score.

Regression task: In the case of regression models, the goal is to predict continuous values as close as possible to the actual (correct) values: it is an approximation. In order to quantify this, two types of metrics can be computed, based either on the distance or the correlation between predicted and actual values.

Regarding distance based metrics, the most commonly used is the **Root Mean Squared Error (RMSE)** which corresponds to the square root of the average squared sum of distances between

the predicted values (\hat{y}) and the actual ones (y), according to the following formula:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

The other well known metric, easier to interpret, is the **Mean Absolute Error (MAE)** and corresponds to the absolute difference between the predicted and actual values, with the following formula:

$$MAE = \frac{\sum_{i=1}^n |y_i - \hat{y}_i|}{n}$$

With regard to correlation, the **coefficient of determination (noted R^2)** is the common metric that indicates how the predicted values (\hat{y}) are correlated to the actual ones (y). It corresponds to the difference between 1 and the ratio of the residual sum of squares ($SS_{res} = \sum_{i=1}^n (y_i - \hat{y}_i)^2$) and the total sum of squares ($SS_{tot} = \sum_{i=1}^n (y_i - \bar{y})^2$, where \bar{y} is the mean of the actual values). Thus, the R^2 is computed according to the following formula:

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

In the case of linear regression, R^2 is the square of the Pearson correlation coefficient.

2.5 Dealing with imbalanced data

There are many classification problems for which the datasets are imbalanced, meaning that the number of observations in the different classes is unequal, in particular in the case of two classes. In such cases, classifiers tend to favor the majority class and result in good predictions regarding this class versus bad predictions regarding the minority class. In order to limit the bias induced by imbalanced datasets, several techniques can be applied and we describe the most commonly used in the following.

Cost-sensitive learning: Unlike regular learning where all misclassifications are treated equally, this method applies a cost to penalize the misclassifications and this cost is heavier when the minority class is misclassified [79]. Since the goal of the ML algorithm is to minimize the total cost, it will learn by paying more attention to the minority class.

Undersampling: In that case, observations from the majority class are removed from the training set, either randomly (we talk about random undersampling) or based on the representation of the original dataset into the descriptors space (we talk about informed undersampling). For the second approach, the kNN algorithm can be applied to select the observations to remove (from the majority class) according to their distance to the ones of the minority class [298]. Basically, this method can be applied in different ways by, for example, removing observations having the smallest (or largest) average distance to the k closest (or farthest) minority class examples, or keeping only a given number of majority class examples that surround each minority class

example. Cluster-based undersampling is also used and consists in applying a clustering algorithm on the dataset and to select one or some representative observation(s) from each cluster to keep in the training dataset [287]. Note that to apply undersampling, the original dataset should contain a lot of data.

Oversampling: In contrast to undersampling, oversampling aims at increasing the number of observations of the minority class and is usually applied when the amount of data in the original dataset is insufficient. We also talk about **data augmentation**. Here also, the oversampling can be done randomly by duplicating random existing observations or cluster-based by duplicating observations in each cluster obtained after clustering to get an equal number in all clusters [114]. Nonetheless, these two approaches can suffer from overfitting since they use exact replications of the minority samples. The most popular oversampling technique used to face this issue is called **Synthetic-Minority Over-Sampling Technique** [42, 43], or **SMOTE** for short.

The SMOTE method aims at creating new synthetic samples based on linear interpolation of actual data. Basically, for each observation i of the minority class, it randomly selects one of its k -nearest neighbors (k) of the minority class in the descriptors space and generate a random example that is along the line between i and k according to the following formula:

$$x_{new} = x_i + (x_k - x_i) \times \delta$$

where x corresponds to the vector of descriptors (input features) of the different observations and δ is a random number from the interval $[0,1]$.

This process is repeated for all or part of the k -nearest neighbors of each observation from the minority class, according to the desired final number of new samples.

Adaptive algorithms of the SMOTE technique have been proposed including:

- Borderline-SMOTE [116] : unlike the original SMOTE where all observations from the minority class are used, in that case only the borderline observations from this class are over-sampled. These observations correspond to the ones for which the number of neighbors, among its k -nearest neighbors, belonging to the majority class is greater than its number of neighbors from the minority class. The borderline observations are therefore close to the border between the majority and the minority class.
- Adaptive Synthetic Sampling (ADASYN) [123]: unlike the standard SMOTE method where the same number of samples is synthesized for each minority class example, here the number of new samples generated varies. In particular, this number depends on the distribution of the majority examples around the considered minority class observation: the more neighbors belong to the majority class, the more samples are synthesized. This thus forces the algorithm to focus on these examples from the minority class that are more difficult to learn.

In this chapter we provided a broad overview of machine learning. Some of the methods described will be applied to toxicological data in order to build: (1) ML models that predict *in*

vitro bioactivities from chemical structure of compounds and (2) ML models that predict *in vivo* toxic outcomes from *in vitro* data, eventually combined with chemical structures. In order to build such models, we first need to look for the appropriate public data and the next chapter provides a review of the available resources.

CHAPTER 3

COMPUTATIONAL TOXICOLOGY FOR TOXICITY PREDICTION

This chapter focuses on the computational tools that are used to establish the link between the different types of toxicological data. A key component for computational toxicology is the development of large databases gathering well structured and standardized information easily downloadable and handable by users. Therefore, the first part of this chapter focuses on the different types of toxicological data and existing databases, with emphasis on the data used in this thesis. In particular, we focus on three types of data: compound structure, *in vitro* activity and *in vivo* toxicity. Secondly, we propose to present a large but not exhaustive overview of what has been done so far using these data in order to predict potential toxicity of compounds and help in their prioritization and risk assessment. With this overview we demonstrate that, due to the quite recent interest for computational methods applied to toxicology, there is still a lack of uniformity and concordance with respect to the methods to use.

3.1 Data in toxicology

3.1.1 Types of data

Chemical structure: The chemical structure is the most simple type of information regarding a compound. It does not require to perform any experimental test to be determined except if the compound is unknown. Hence, the structure is a cost-free and rapidly accessible information having a lot of interest in toxicity prediction. Indeed, if the prediction could be based only on the structural information, it would save a lot of time, money and laboratory animals.

In order to be read and interpreted by computers, the information provided by the chemical structure should be represented in understandable format. Various types of chemical structure representations are known and we review some of them in Section [i](#)).

***In vitro* data:** As stated in Chapter 1, *in vitro* assays are tests performed in solutions containing biological molecules, in cells or even in "lower" organisms (*e.g.* zebrafish, *c. elegans*). Each test measures a specific bioactivity such as protein binding, gene activation, cell death, *etc.* In general, the activity corresponds to a percentage of activation or inhibition compared to a baseline control. To test the effect of compounds on these activities, assays are performed either using one single concentration of compound or a range of concentrations (in nanoMolar nM , microMolar μM or milliMolar mM). In the first case the result is a binary value (*i.e.* active / inactive) and in the second case it results in a concentration-response curve which enables the computation of several values, see Figure 3.1. The most used values are the EC_{50} (half maximal effective concentration) and IC_{50} (half maximal inhibitory concentration) which respectively refers to the concentration of compound inducing 50% of activity or inhibition. They are determined by considering both the baseline and the maximal response which is reached when the response is stable (*i.e.* the curve has reached a plateau).

These assays can be performed in high-throughput screening (HTS) allowing the generation of a lot of data in a short amount of time thanks to the use of robots. During the last decades, HTS has been performed in the context of several initiatives in order to obtain bioactivity signatures of compounds for prioritization purposes.

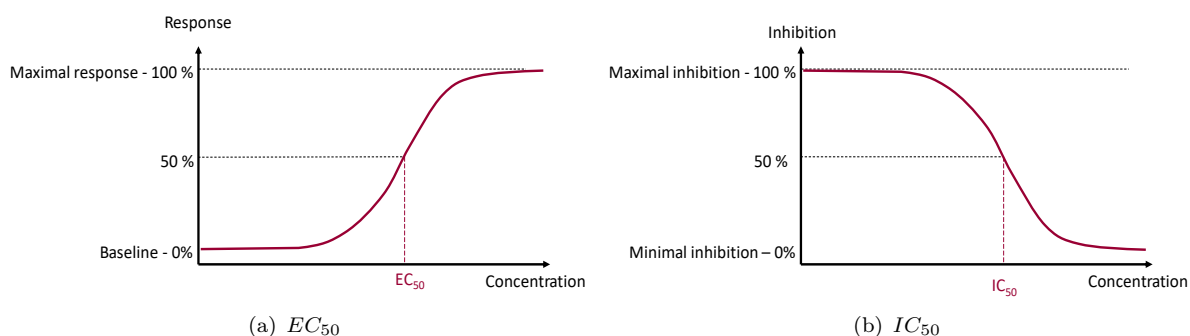


Figure 3.1: Concentration-response curve. a) Concentration-activity curve to determine the EC_{50} value which corresponds to the concentration of compound inducing 50% of activity, b) Concentration-inhibition curve to determine the IC_{50} value which corresponds to the concentration of compound inducing 50% of inhibition.

***In vivo* data:** They are the results of *in vivo* experiments performed in laboratory animals. There are various types of *in vivo* toxicity studies with different durations and outcomes of interest, aiming either at the description of the hazardous profile of compounds (guideline studies) or at a better understanding of adverse effects (MoA studies). The most important ones have been already described in Chapter 1, Section 1.2.3. The results of these studies available in databases can be of different types such as:

- Number of animals with the outcome and total number of animals in the study, for each tested dose and sometimes several time points. These numbers can also be expressed as a ratio.
- Doses such as LD50 (Lethal Dose 50%), NOAEL (No-Observed-Adverse-Effect-Level),

LOAEL (Lowest-Observed-Adverse-Effect-Level), *etc.* (see Chapter 1, Section 1.2.3).

- Binary value informing about the presence or absence of the outcome.

Omics data: As also stated in Chapter 1, omics refers to several disciplines generating data for various biological entities in order to map them to biological outcomes (diseases, adverse effects, phenotypes, *etc.*). In particular, **toxicogenomics** studies aim at analyzing the expression of a large panel of genes and their transcripts after exposure of cells (*in vitro*) or of an entire organism (*in vivo*) with a chemical of interest in order to investigate the molecular mechanisms inducing toxicity. This analysis results in gene expression profiles where each gene activity is quantified using a baseline control. These profiles help in understanding the molecular mechanisms of toxicity and ultimately build links between cellular and physiological mechanism taking place during the formation of an adverse effect. Such as *in vitro* assays, toxicogenomics can be performed in high-throughput studies thanks to the use of, for example, DNA micro-arrays (or DNA chips) or Next Generation Sequencing.

Mechanistic data: They refer to all the information that is useful to characterize mechanisms of toxicity. More specifically, they correspond to MOA and AOPs and should include MIE, KEs as well as weight of evidence to demonstrate the causal link between each KE. See Figures 1.3 and 1.4 of Section 1.2.5 for more details.

3.2 Existing sources of toxicological data

A number of sources (either publicly accessible or with restricted access) provide toxicological related data. Table 3.1 is a synthetic summary of the most important resources, which focuses on publicly available ones. A more detailed description of these resources is available in Appendix A. Note that other data sources specific to pharmaceutical compounds are also available but since we are not specifically interested in this type we decided to not integrate them in this summary table.

There are several sources providing information about chemical structures only but they are generally not used for toxicological purpose. Indeed, most of the databases that contain toxicological data also provide information regarding the structure of compounds. Nonetheless, we can still cite the two following resources for chemical structures: PubChem [268, 269] and the Distributed Structure-Searchable Toxicity database [216], named DSSTox.

Regarding *in vitro* data, we can mention several resources such as, already cited PubChem, ChEMBL [21], Chemical Effects in Biological Systems (CEBS) [271], the Estrogenic Activity Database (EADB) [237] and data generated during the Tox21 [260] and ToxCast [68, 217] programs. Note that some of these resources also include *in vivo* and / or genomics data.

Table 3.1: Summary of existing resources for toxicological data. NA = Non Available. For abbreviations, see main text and list of abbreviations

Name of the database	Type of data	URL				
ACToR	<i>in vitro</i> , <i>in vivo</i> , chemical structures, exposure, EDSP program, etc	https://actor.epa.gov/actor/home.xhtml				
AOP-KB	Mechanistic information	https://aopwiki.org/				
CEBS	<i>in vitro</i> , <i>in vivo</i>	https://manticore.niehs.nih.gov/cebssearch/				
CHEMBL	<i>in vitro</i>	https://www.ebi.ac.uk/chembl/				
COSMOS DB	Chemical structures, <i>in vitro</i>	http://www.cosmostox.eu/what/COSMOSdb/				
CMap	Genomics	https://portals.broadinstitute.org/cmap/				
CPDB	<i>in vivo</i> cancer studies	https://toxnet.nlm.nih.gov/cpdb/				
CTD	Toxicogenomics	http://ctdbase.org/				
DSSTox	Chemical structures	ftp://newftp.epa.gov/COMPTOX/Sustainable_Chemistry_Data/Chemistry_Dashboard				
EADB	<i>in vitro</i> , <i>in vivo</i> estrogenic activity related data	https://www.fda.gov/ScienceResearch/BioinformaticsTools/EstrogenicActivityDatabaseEADB/default.htm				
eTOX	<i>in vitro</i>	http://etoxsys.com/				
FedTex	<i>in vitro</i> developmental and reproductive studies	http://publica.fraunhofer.de/starweb/pub09/en/index.htm				
HESS	<i>in vitro</i>	https://www.nite.go.jp/en/chem/qsar/hess-e.html				
PubChem	Chemical structures, <i>in vitro</i>	https://pubchem.ncbi.nlm.nih.gov/				
RepDose	<i>in vitro</i>	https://repdose.item.fraunhofer.de/index.php				
TG-GATES	Toxicogenomics, <i>in vitro</i>	https://dbarchive.biosciencedbc.jp/en/open-tggates/download.html				
Tox21	<i>in vitro</i>	https://ntp.niehs.nih.gov/results/tox21/tbox/index.html				
ToxCast	<i>in vitro</i>	https://www.epa.gov/chemical-research/exploring-toxcast-data-downloadable-data				
ToxRefDB	<i>in vitro</i>	ftp://newftp.epa.gov/comptox/High_Throughput_Screening_Data/Animal_Tox_Data/				
Name of the database	Provider	Number of compounds	Number of endpoints / studies	Type of compounds	Last release	References
ACToR	US EPA	> 500,000	NA (>500 data sources)	All	Regularly updated	[140]
AOP-KB	OECD – collaborative work	NA	258 AOPs and 1910 KEs	NA	Regularly updated	[275]
CEBS	NTP / NIEHS	> 11,000	> 8,000 studies for 19 data types	All	Regularly updated	[271, 161]
CHEMBL	EMBL / EBI	>1.8 millions	>12,000 targets	Drugs	Regularly updated	[21]
COSMOS DB	Part of the SEURAT European project	> 40,000 structures and > 1600 compounds with <i>in vitro</i> results	12,000 studies	Cosmetics	2013	[283]
CMap	Broad Institute	>1,300	>7,000 genes + L1000 assay	All	2017	[160, 248]
CPDB	University of California, Berkeley	1547	6540 studies	All	2001	[91]
CTD	North Carolina State University	> 15,000	> 45,000 genes	All	Regularly updated	[182, 109]
DSSTox	US EPA	> 700,000	NA	All	Regularly updated	[216]
EADB	US FDA	>8,212	1,284 <i>in vitro</i> assays and <i>in vitro</i> studies	All	2012	[237]
eTOX	Consortium of 30 partners	>8,000	>8,000	Mostly pharmaceuticals	2016	[245]
FedTex	Fraunhofer Institute for Toxicology	300	535 studies	All ?	NA	[23]
HESS	Japanese NITE	>500	>500 studies	Around 500	Regularly updated	[229]
PubChem	NCBI	> 97 millions structures and > 3 millions compounds with <i>in vitro</i> data	> 1 million bioassays	All	Regularly updated	[268, 269]
RepDose	Fraunhofer Institute for Toxicology	830	3,100 studies	All	NA	[24]
TG-GATES	Japanese Toxicogenomics Project	170	> 20,000 genes	Pharmaceuticals	2012	[132, 263]
Tox21	Tox21 consortium	Around 10,000	> 70 assays	All	October 2018	[260]
ToxCast	US EPA	Around 1,800	> 1,000 assays	All	October 2018	[68]
ToxRefDB	US EPA	>400	Around 5,900 studies	All	2019	[214]

Concerning the *in vivo* data, the following resources can be highlighted: the Hazard Evaluation Support System (HESS) [229], the RepDose database [24], the Fertility and Developmental Toxicity in Experimental animals database (FedTEX) [23], the Carcinogenic Potency Database (CPDB) [91], the COSMOS database [283], eTOX [245] and ToxRefDB [214].

Moreover, examples of databases dedicated to genomics data are the Toxicogenomics Project-Genomics Assisted Toxicity Evaluation Systems (TG-GATEs) [132], the Comparative Toxicogenomics Database (CTD) [182, 109] and the Connectivity Map (CMap) [160, 248]. For mechanistic data, we can evoke AOP-KB [275].

Finally, the Aggregated Computational Toxicology Resource (ACToR) is the EPA's on line warehouse that stores data from thousands of public sources [140]. In particular, it includes some of the previously mentioned sources (*i.e.* DSSTox, ToxCast, ToxRefDB) as well as the results from the EDSP and ExpoCast programs (see Chapter 1, Section 1.2.6 and 1.3.2, respectively).

Although a lot of data are publicly available regarding compound toxicity, they are not always appropriate for an efficient use in *in silico* purpose. The next section lists some of the challenges encountered when using these resources.

3.2.1 Challenges in using toxicity data for computational purpose

Toxicity data are characterized by several limitations obliging scientists to face various challenges. First, the data structures are not always in **machine readable format**: either the structure is not provided in the database (*e.g.* in TG-GATEs) and the mapping to other databases or services such as PubChem is required or it is provided but in an inappropriate representation for further use and therefore need to be transformed and curated.

Moreover, one of the biggest issue with toxicological data is the **lack of homogeneity**, either within a database or between several ones. This heterogeneity can be observed for both the compounds' identifiers and the results of toxicity studies. In particular, regarding *in vitro* studies, differences come from the protocol used to perform a same assay in various laboratories (inter-labs differences), the name of assay targets or the types of results that are provided (*e.g.* dose-response curves, EC_{50} or active/inactive). With respect to *in vivo* data, heterogeneity concerns all the parameters of the studies such as the laboratory animal species used, the doses, the time points, the administration routes, *etc.*, but also the name of observed endpoints (discussed later) and the types of results (*e.g.* LOAEL, incidence of the considered endpoints, *etc.*). Consequently, this heterogeneity rises challenges to aggregate data from various databases and such aggregation should be done with caution with a particular attention on redundancy and inconsistency of results.

Last but not least, a big challenge regarding toxicity data is the use of a consistent and universal **ontology** [121]. Indeed, as stated earlier, the name of endpoints is a marker of heterogeneity, whether between or within databases. Most of the databases do not use controlled vocabularies to describe the endpoints which is an important problem since it is what is considered in the final

risk assessment. An ontology is defined as an effective and structured way of describing terms and concepts. The entire community agrees that ontology is required and some initiatives such as ToxML [218] and OpenTox [255, 9] are working in that way. Moreover, the use of standardize format like Standard for Exchange of Nonclinical Data (SEND) [50] is considered as well as the development of universal pathology terminology in the International Harmonization of Nomenclature and Diagnostic (INHAND) project [145].

Finally, another specific characteristic of toxicological data is the **imbalanced** property of the results: most of the compounds for which data are available are "negative". It is in particular the case for *in vitro* activity data where most of the compounds are inactive. This imbalanced characteristic in favor of inactive compounds has to be taken into account in the context of predictive toxicology.

Currently, US EPA is one of the most advanced entity that provides inter-operable data enabling the aggregation of several types of information. Indeed, they provide the three types of data we are focusing on (chemical structure, *in vitro* and *in vivo*) with a common identifier for the compounds. Moreover, these data are almost ready-to-use avoiding an important processing step. Finally, since these data have been generated in the context of the Tox21 vision in order to enable prioritization of compounds for further testing and development of predictive approaches, they are well adapted for a computational purpose. For these reasons, we decided to use EPA's data in further work. We therefore provide a broader description of these data in the next section.

3.2.2 Focus on the data used in this thesis

The US EPA provides chemical structures of compounds in the **DSSTox** database, *in vitro* bioactivity data from both Tox21 and ToxCast program in the **ToxCast** database and *in vivo* data in the **ToxRefDB**. All these data are also available in the ACToR system.

DSSTox: The Distributed Structure-Searchable Toxicity database is a resource [216] that gathers chemical structures of compound, their corresponding physico-chemical properties and toxicity data (*in vitro* bioassays). Today it includes data for over 700,000 compounds and is publicly available through the EPA's website. In the work presented in this manuscript, the release from 2015 was used and contains 9011 structures along with their unique identifier (DSSTox_GSID), name, molecular formula, molecular weight and other structural representations such as SMILES, InChi, InChiKey, *etc.* All these information are provided in a single file, using the Structure Data File (SDF) format (see Section i)).

Tox21: In order to apply the new vision described in Chapter 1, the Tox21 consortium launched the Tox21 program that aimed at using the expertise from the different agencies to generate a lot of data to rapidly screen compounds. In particular, the ultimate objective is to identify *in vitro* signatures that could help in the prediction of *in vivo* toxicity [128]. Between 2005 and 2016, two phases have been conducted in order to generate HTS data for more than 75

assays (corresponding to more than 100 endpoints) originally developed in the context of a collaboration between NCATS and NTP. These assays are performed in 1,536-well plates allowing the rapid testing of a large number of compounds. As of today, a library of approximately 10,000 compounds, the **Tox21 10K library** [130], has been tested in these assays and this library includes environmental chemicals, food additives, drugs, chemical mixtures, *etc.* [260]. Assays can be categorized into phenotypic, target-specific and pathway-based assays and measure the following list of endpoints [128]:

- Cell viability
- Apoptosis (cell death)
- Membrane integrity
- DNA damage / epigenetics
- Mitochondrial toxicity
- Phospholipidosis
- Ion channel and GPCR signaling - G-protein coupled receptor
- Cytokine secretion
- CYP induction - cytochrome P450
- Nuclear Receptor signaling
- Stress response signaling

All the data generated during the Tox21 program are available through different sources including the Tox21 Toolbox¹, CEBS, PubChem and the ToxCast database from EPA.

ToxCast: The **Toxicity Forecasting** (ToxCast) program is one of the EPA's contribution to the Tox21 collaboration. It was originally launched in 2007 by the US EPA in order to deal with the large number of environmental chemicals that require testing for their potential toxicity. The main goal of the program is to develop methods based on HTS, computational approaches and omics technologies to detect and predict potential of compounds for toxicity and prioritize them for further screening and animal testing [68, 217]. By testing chemicals in hundreds of HTS assays, the objective is to obtain **bioactivity profiles** or signatures that are predictive of *in vivo* toxic potential. Currently, ToxCast uses more than 1000 HTS assays that are performed in several independent laboratory platforms and vendors which are coordinated by the National Center for Computational Toxicology (NCCT) in EPA.

Most of the cell based assays are performed in rat and human cell lines. Around 1,800 compounds have been screened in these assays and the assays which turned out to be the most useful have then been considered by the Tox21 program to extend the testing to the 10K library. Results generated within both the ToxCast and the Tox21 program are analyzed by the NCCT and correspond today to 1192 assay endpoints (mostly performed at multiple concentrations) and more than 9000 compounds. Nonetheless, note that not all the compounds have been tested in all the assays.

Depending on the needs of the users, the ToxCast and Tox21 data are available through several

¹<https://ntp.niehs.nih.gov/results/tox21/tbox/index.html>

sources, namely the iCSS ToxCast Dashboard² for a quick and easy access and consultation, the ToxCast data download page³ for the download of entire data files and a MySQL database provided with a R package (*tcpl*) for a full data access [227].

Several versions of the ToxCast and Tox21 data have been released since the beginning due to the acquisition of new data and the evolution of data analysis (in particular quality control) methods and the last release (third version) dates from October 2018. In the work presented here, we used the October 2015 release (second version) still available through the download page⁴. Different types and levels of information are available in the 2015 release such as assays information, concentration-response plots and summary files of various results obtained after data analysis. This analysis has been performed by the NCCT using a pipeline provided by the R package *tcpl* and composed of 6 different levels [90]. Basically, this pipeline aims at processing, normalizing and fitting the acquired concentration-response for each pair of compound and assay to a mathematical regression model (constant, hill or gain-loss) and finally determine activity values such as the AC_{50} which corresponds to the concentration inducing 50% of activity (either activation or inhibition) and usually known as EC_{50} or IC_{50} ⁵. These AC_{50} are reported in a matrix where assays are listed across the columns and compounds down the rows and are expressed in micromolar (μM). When a compound was not tested in an assay, it is reported as "NA" and if a compound was inactive in the assay (*i.e.* no AC_{50} could be determined), a value of 1,000,000 is reported. Another useful matrix is the "hit call" matrix (*hitc*) which indicates if a compound is active (1), inactive (0) or not tested (-1) in the assay. All the compounds are identified using their ToxCast identifier (*chid*) and their DSSTox identifier (*DSSTox_GSID*) which enables the crossing of the different sources of information.

ToxRefDB: The **Toxicity Reference Database** has been developed by the US EPA to gather results of *in vivo* studies from various sources using partially harmonized terminology [214]. **In particular, it is probably one of the best examples of pretty well curated database including quite standardized ontology.** This database was designed to help in the development of models for the prediction of toxicity and for the validation of the ToxCast *in vitro* assays. The database is regularly improved and updated and the current ToxRefDB2.0 version contains data for more than 400 compounds tested in various toxicology studies. These studies were conducted in rat, mouse, rabbit and dog with exposure duration ranging from some days to the animal lifetime. It also gathers multi-generational and reproductive studies. In total, more than 1,000 endpoints are characterized and for each endpoint induced by a compound, the corresponding NOAEL or LOAEL is given. Each study referenced in the database is accompanied by various information such as its quality, its duration, the laboratory animal species used, the number of animals in each group of test, the tested doses, *etc.* The ToxRefDB is available through the ToxCast data

²<https://actor.epa.gov/dashboard/>

³<https://www.epa.gov/chemical-research/exploring-toxcast-data-downloadable-data>

⁴ftp://newftp.epa.gov/COMPTOX/High_Throughput_Screening_Data/Previous_Data/ToxCast_Data_Release_Oct_2015/

⁵Note that this AC_{50} term is specific to ToxCast and refers to both EC_{50} and IC_{50}

download page. In the work presented here, we used the release from October 2014 where results are reported into flat files in which compounds are also identified using their DSSTox_GSID. Table 3.2 provides a summary of the three sources of data used in the work presented here.

Table 3.2: Summary of data used. In this work we focus on data regarding compound structures, *in vitro* bioactivity and *in vivo* studies provided by the US EPA.

Type of data	Compound structures	<i>In vitro</i> assays	<i>In vivo</i> studies
Source	DSSTox	ToxCast / Tox21	ToxRefDB
Number of compounds	9,011 structures	10,000	> 400
Number of assays / endpoints	NA	821 assays	> 1,000 endpoints
Compounds identifier	DSSTox_GSID	DSSTox_GSID	DSSTox_GSID

After having sketched the toxicological databases, we propose an overview that shows how these data can be used for a computational purpose.

3.3 From chemical structure to *in vitro* activity or *in vivo* toxicity

Firstly, we focus on the computational methods that evaluate the link between chemical structure and *in vitro* activity or *in vivo* toxicity of compounds.

3.3.1 Non machine learning approaches

So-called non-testing methods aim at generating data about chemicals' effect only from their structure and are not necessarily based on machine learning. In particular, they include the three following approaches which have been acknowledged by the OECD and REACH as alternative methods [213].

Rule based models: **Structural alerts** are fragments within a molecule which are associated with a specific activity, they are also called **toxicophores**. They enable the definition of rules of the form "*if A is B then T*" where A is a structural alert, B is its value (presence or absence of A) and T is the toxic effect. These rules are derived either from experts knowledge or from probabilities computed using large datasets. Rule-based models are used to design new chemical compounds (drugs and pesticides) and lists of structural alerts are available for specific toxic endpoints such as carcinogenicity, hepatotoxicity, cytotoxicity, *etc.* [212]. Moreover, rule-based models have been implemented and made available in several software called "**expert systems**" including Toxtree [203], Derek Nexus or Meteor Nexus [178] (from Lhasa Limited company).

Chemical category and read-across: Read-across aims at making interpolation of activities from a group of similar compounds for which the considered activity is known [57]. It first requires to perform **chemical grouping** (or category formation) where chemicals are grouped into categories according to their similarity on several characteristics [80]: chemical structure,

mechanism of action, physico-chemical properties, biological interactions, *etc* [212]. In read-across, the hypothesis is that compounds belonging to a same category will be associated with the same activity or effect. Therefore, the predicted activity of a new compound will correspond to the one of the compounds that share the same class (called analogs). There are two types of read across: the **analog approach** when the new compound is compared to only one or a small number of similar compounds and the **category approach** when it is compared to many other compounds of the same group. If category approach is better to increase confidence in the predictions, it is also more difficult since the category has to be well defined and sometimes the data do not allow performing category approach because of bad quality and / or uncertainty.

Structure-activity relationship (SAR): This is a qualitative method which aims at looking for relationships between fragments of the chemical structure (chemical functions, groups) and the presence or absence of an activity. These techniques suppose that the biological effects of new chemical compounds can be deduced (or predicted) from their molecular structure. In particular, they assume that compounds with similar structures may share similar biological and toxicological properties. Hence, SAR methods base their predictions on existing data regarding compounds that are structurally similar to the new compound [280]. In order to look for this similarity, they sometimes apply clustering methods using various similarity measures such as the Tanimoto index or the Euclidean distance [147].

The quantitative version of SAR is the so-called **QSAR** and belongs to the machine learning field. Since most of the work presented in the following chapters involve QSAR modeling, the next section focuses on this approach in further details.

3.3.2 Quantitative Structure-Activity relationship (QSAR)

QSAR is a well known ML approach that aims at predicting compounds activity from their chemical structure. It has been firstly introduced by Hansch in 1962 [120, 119] and intensively studied since then. On the one hand, input features correspond to **molecular descriptors** which are automatically computed from the structure of the compounds. On the other hand, the predicted output is an **activity** which can be a physico-chemical property (*e.g.* logP, solubility), a biological activity measured in *in vitro* assay (*e.g.* molecular binding, gene transcription, mutagenicity), or an effect observed in *in vivo* studies (*e.g.* carcinogenicity, endocrine effect).

i) Molecular descriptors

In order to prepare the datasets for learning, molecular descriptors of the compounds have to be computed by software using the information provided by the chemical structure for which various representations exist. Nonetheless, these representations are sometimes incorrect and therefore need to be cleaned before the computation of the molecular descriptors.

In the following, we first review the most used structural representations of compounds' structure, we then detail how to ensure the quality of these representations thanks to a data curation process

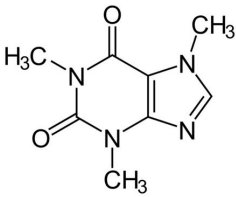

and we finally summarize the different types of molecular descriptors.

Chemical structure representations: In order to be stored in databases, chemical structures must be encoded into standardized formats that are readable by software. These representations must be **unique** and **non-ambiguous** which is not the case of the molecular formula or the International Union of Pure and Applied Chemistry (IUPAC) nomenclature [270].

Chemical structure representations can be either in one dimension (linear) or in two dimensions (connection tables). Here are the most commonly used representations (illustrated in Table 3.3 for the caffeine molecule):

- **SMILES:** it stands for Simplified Molecular Input Line Entry Specification and is a linear string notation encoding the atoms of a molecule (except hydrogens) and their bonds (simple and double bonds, branches, rings) [273]. The biggest limitation of SMILES is that there is no unique algorithm to generate them and therefore one structure can have several versions of SMILES (but one SMILES refers to a unique compound). Nonetheless, standard methods have been recently proposed to generate canonical SMILES [194].
- **InChi:** it stands for IUPAC International Chemical Identifier and is also a linear string notation which corresponds to the digital form of the IUPAC name [125]. It is unambiguous and unique and is composed of a maximum of six layers informing about different characteristics of the structure (formula, connectivity, hydrogens, isotopes, stereochemistry and charge). The **InChiKey** has been derived from InChi in order to reduce the length of the InChi into a condensed identifier with a fixed number of characters (27) making easier their processing in databases [270].
- **Molfile format:** it is a file text format which contains header information followed by a **connection table** which represents the structure of molecules in two dimensions. The connection table is composed of [270]:
 1. A **counts line** informing about the number of atoms and bonds,
 2. An **atom block** where each line corresponds to one atom and inform about its 3D coordinates, symbol, charge, *etc*,
 3. A **bond block** where each line corresponds to a bond between two atoms,
 4. A **properties block** giving information about additional properties such as charges and isotopes.
- **Structure Data File (SDF) format:** it is also a text format which gathers structural information for several compounds, unlike the Molfile [270]. For each compound, it includes the corresponding Molfile with associated data that can be identifiers, physico-chemical properties, biological activity, *etc*. Information regarding each compound are delimited by a separator.

Table 3.3: Examples of representations of the caffeine structure.

Molecule name	Caffeine
2D structure	
Molecular formula	$C_8H_{10}N_4O_2$
IUPAC name	1,3,7-trimethylpurine-2,6-dione
SMILES	<chem>CN1C=NC2=C1C(=O)N(C(=O)N2C)C</chem>
InChi	1S/C8H10N4O2/c1-10-4-9-6-5(10)7(13)12(3)8(14)11(6)2/h4H,1-3H3
InChi Key	RYYVLZVUVIJVGH-UHFFFAOYSA-N
Connection table (MolFile format)	<pre> 24 25 0 0 0 0 0 0999 V2000 3.7320 2.0000 0.0000 O 0 0 0 0 0 0 0 0 0 0 2.0000 -1.0000 0.0000 O 0 0 0 0 0 0 0 0 0 0 3.7320 -1.0000 0.0000 N 0 0 0 0 0 0 0 0 0 0 5.5443 0.8047 0.0000 N 0 0 0 0 0 0 0 0 0 0 4.5981 -0.5000 0.0000 C 0 0 0 0 0 0 0 0 0 0 3.7320 1.0000 0.0000 C 0 0 0 0 0 0 0 0 0 0 [...] 6.4443 1.5626 0.0000 H 0 0 0 0 0 0 0 0 0 0 6.0476 2.3446 0.0000 H 0 0 0 0 0 0 0 0 0 0 5.2656 1.9479 0.0000 H 0 0 0 0 0 0 0 0 0 0 2.3100 1.5369 0.0000 H 0 0 0 0 0 0 0 0 0 0 1.4631 1.3100 0.0000 H 0 0 0 0 0 0 0 0 0 0 1.6900 0.4631 0.0000 H 0 0 0 0 0 0 0 0 0 0 1 9 2 0 0 0 0 2 10 2 0 0 0 0 3 8 1 0 0 0 0 3 10 1 0 0 0 0 3 12 1 0 0 0 0 4 7 1 0 0 0 0 4 11 1 0 0 0 0 [...] 13 20 1 0 0 0 0 13 21 1 0 0 0 0 14 22 1 0 0 0 0 14 23 1 0 0 0 0 14 24 1 0 0 0 0 M CHG..... M ISO... M END </pre> <div style="display: flex; justify-content: space-between; align-items: center;"> <div style="width: 45%;"> <p>Counts line</p> <p>Atom block</p> <p>Bond block</p> <p>Properties block</p> </div> <div style="width: 5%; text-align: center;">  </div> <div style="width: 45%; font-size: small;"> <p>Counts line</p> <p>Atom block</p> <p>Bond block</p> <p>Properties block</p> </div> </div>

Chemical data curation: One other crucial step before the generation of molecular descriptors is, as stated in Chapter 2, to ensure the correctness of the input data and in particular the quality of the structure representation in the case of QSAR modeling. Fourches *et al.* [93] proposed a standardized data curation procedure which they recommend to apply before any QSAR modeling and whose main steps are summarized below.

- 1. Removal of inorganics and mixtures:** most of descriptor-generating software cannot process inorganic compounds and salts and results in many errors during the descriptors' calculation. Therefore, this type of molecules, which is not known to induce biological activity, should be removed from the data.

Additionally, in some cases a single structural representation encodes for several compounds (mixtures) and not only one which is also not well suited to compute molecular descriptors. When mixtures contain several organic compounds of the same size, it is recommended to completely discard it. Otherwise, when there is one major organic compound mixed with

- smaller ones and/or small inorganic molecules, the biggest one can be kept while the others are deleted.
2. **Structural conversion and cleaning:** the conversion of SMILES into 2D structures enables visualizing the molecules and cleaning them from salts, ions and other non desired charges in order to get neutralized compounds. It is also recommended to add explicit hydrogens.
 3. **Normalization of specific chemical function:** sometimes the same functional groups can be differently represented due to an electronic effect within the molecule which makes the electrons moving inside the functional group. In the chemical structure representation, this generally results in double bonds that are represented at different locations of the functional group. Nonetheless, all these representations correspond to the same molecule and it is therefore recommended to normalize them into a standard form.
 4. **Removal of duplicates:** some compounds can appear several times in a same dataset due to the use of different IDs and only one must be kept in order to avoid bias when modeling. The removal can be done by using canonical SMILES.
 5. **Manual checking:** The last recommended step would consist of carefully and manually looking at each curated structure, in particular when the previous steps have been performed automatically. Nonetheless, due to the large number of compounds that are constituting ML datasets, this step would lead to the loss of a considerable amount of time. Furthermore, this is not so obvious that a "human" check would be more efficient than an automated check.

To perform the different steps of the data curation process, except the manual checking, various software are publicly available [262].

Molecular descriptors: All the structure representations described above can be used to compute the molecular descriptors that will constitute the input features of the QSAR model. Many types of descriptors exist with different levels of complexity and chemical structure representations. Indeed, we distinguish 1D, 2D and 3D descriptors which are respectively computed from the molecular formula, the two-dimensional structural formula or the three-dimensional conformation of the molecule [48]. The most popular descriptors are the 1D and 2D and include the following classes [73]:

- **Constitutional descriptors:** they are related to the components of the molecule (*e.g.* number of specific atoms, number of single, double, triple bounds, number of rings, *etc.*).
- **Physico-chemical descriptors:** they are estimation of the physico-chemical properties of the molecule and are sometimes included in the previous class (*e.g.* molecular weight, solubility, partition coefficient such as logP which corresponds to the logarithm of the ratio of a compound's concentrations measured in the two phases of the octanol/water solvent, *etc.*).

- **Geometrical descriptors:** they are related to the spatial arrangements of the components of the molecule (*e.g.* molecular surface area, volume, moments of inertia, *etc.*).
- **Topological descriptors:** they are based on the topology of the molecular graph and lead to various indexes (*e.g.* Wiener index, Randic index, *etc.*).
- **Electrostatic and quantum descriptors:** they are related to the electronic nature of the molecule and to the molecular orbital (*e.g.* partial atomic charge, polarizability, energy, *etc.*).
- **Fragment-based descriptors and fingerprints:** they are binary vectors coding the presence or absence of specific substructures in the molecule [37]. Different types and "dictionaries" of fingerprints exist including Molecular ACCess System (MACCS) keys [74], PubChem fingerprints [117], Extended-Connectivity Fingerprints (ECFP) [220], E-state fingerprints [115], Klekota-Roth fingerprints [221], *etc.* They can be computed by a variety of open source and proprietary software packages [37].

As already described in Chapter 2 Section 2.2, once generated these descriptors have to be selected using different existing methods in order to reduce the dimensionality and keep only the ones that are the most relevant [104]. In particular, a general rule of thumb suggests that at least five times more observations than descriptors should be used for QSAR modeling [233, 261]. Then, learning and evaluation of the QSAR model can be performed as described in Chapter 2. Finally, an important concept specific to QSAR models and whose purpose is to estimate the reliability of the predictions has to be considered: the applicability domain. Because of its importance and since we consider the applicability domain in our work, the next section focus on this notion.

ii) Applicability domain

Requirements for the validation of good QSAR models have been proposed and in particular the OECD defined five principles for toxicological predictive models which are the following [198]:

1. A defined endpoint: the bioactivity, toxic effect that is predicted by the model should be clear and the experimental conditions used to measure it should be identified;
2. An unambiguous algorithm: the learning algorithm as well as the parameters used to learn the model should be precise to ensure reproducibility of the predictions;
3. A defined **domain of applicability**: refers to the structural, physico-chemical space defined by the compounds that constitute the training set of the model and in which the model is applicable to make new predictions for unseen compounds;
4. Appropriate measures of goodness-of-fit, robustness, and predictivity: the first two points refer to the use of an internal validation using a training set and the third one to the use of an external validation using an appropriate test set. The training and test set should be clearly identified as well as the metrics used to measure resulting performance;

5. A mechanistic interpretation, if possible: this refers to the description of the association between the descriptors used in the model and the endpoint that is predicted. Nonetheless, this association is not always obvious.

These five principles clearly highlight the importance of the applicability domain and we provide here a definition of this concept as well as the methods used to estimate it.

Definition: Several definitions of the Applicability Domain (AD) have been proposed and the Setubal Workshop report [136] proposed the following one: “The AD of a (Q)SAR is the physico-chemical, structural, or biological space, knowledge or information on which the training set of the model has been developed, and for which it is applicable to make predictions for new compounds. The AD of a (Q)SAR should be described in terms of the most relevant parameters *i.e.* usually those that are descriptors of the model. Ideally, the (Q)SAR should only be used to make predictions within that domain by interpolation not extrapolation.” Therefore, this tool allows one to evaluate the reliability of predictions by defining when and for which compounds a supposed valid model is applicable.

Methods to define the applicability domain: Basically, the AD enables the estimation of the similarity between training and test compounds [239]. In practice, there are different approaches to define the applicability domain [137, 190]:

- **Range-based methods:** these methods consider the ranges of values of each individual descriptor to define an n -dimensional hyper-rectangle space with n being the number of descriptors (Figure 3.2-a). Nonetheless, limitations are that the hyper-rectangle can include empty space (*i.e.* large regions in the n -dimensional space where no compound is represented due to non-uniformity of data distribution) and that correlation between descriptors is not taken into account. A more advanced method that considers correlations and reduces the empty space is the **Principal Component Analysis (PCA)** [276], see Figure 3.2-b. Indeed, it transforms the axes into Principal Components (PCs) that are aligned with the directions of the greatest variations of the training set. The computed ranges then correspond to the minimum and maximum values of each PC and define a new n -dimensional hyper-rectangle. The number of PCs to keep depends on the amount of the total variance one wishes to explain. There is no particular rule for choosing the number of PCs to keep but several heuristics have been proposed such as choosing the desired amount of total variance to explain and keep the PCs in accordance or selecting the PCs whose eigenvalue is greater than a predefined threshold [207].
- **Distance-based methods:** they compute the distance between a new data point and all the points of the training set using different approaches: the distance to the mean, the average distance between the new point and all the points of the training set or the maximum distance between the new point and all the points of the training set. This distance can then be compared to a threshold in order to decide if the new point is close

or not to the training data. The three most used distance are Euclidean, Mahalanobis [60] and Manhattan distance.

- **Geometric methods:** they aim at computing the convex hull which is the smallest convex space that include the original data, see Figure 3.2-c. The biggest limitation of these methods is that the computation of the convex hull is a geometry problem with a high complexity [44].

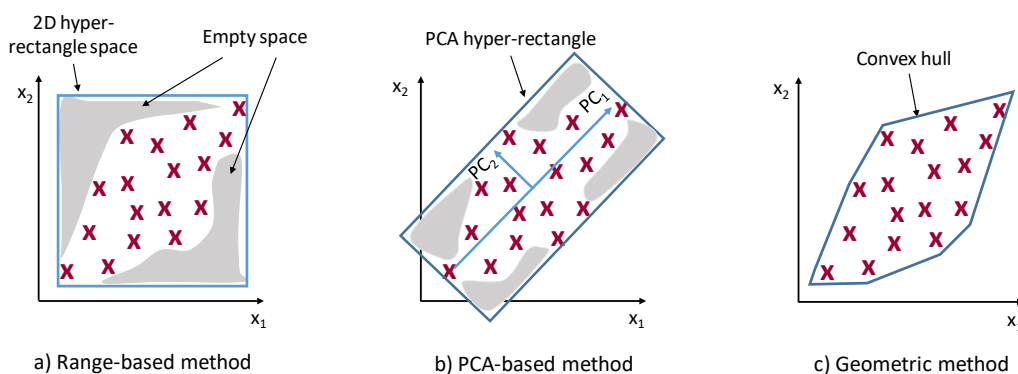


Figure 3.2: Illustration of methods to define the applicability domain in a 2-dimensional space
 a) Range-based method, b) PCA-based method, c) Geometric method.

- **Probability density distribution methods:** these methods aim at estimating the probability density of the data and are either parametric or non-parametric [240]. While the parametric methods assumes that the probability density distribution of the data corresponds to standard distributions (Gaussian or Poisson), non-parametric ones do not make any assumption. Once the probability density has been estimated, the goal is to find the smallest region (in one dimension or more) that comprises a desired fraction of the total probability; this region is called the "highest density region".

Global vs local QSAR: QSAR can be divided into two types of models: global and local ones. **Global QSAR** are built using a large and diverse set of compounds and are therefore characterized by a wide AD while **local QSAR** are usually developed using a specific class of similar compounds resulting in a narrow AD. In term of performance, global QSAR are naturally good for a diverse set of compounds but can result in poor performance for highly similar molecules while local QSAR are good for only similar compounds to the ones used in the training set [87]. The choice between these two strategies when building a new model is not evident as it depends on the nature of available data and the initial objective for building the models. The literature is still unclear on this topic since it appears to be a case-by-case problem [124, 238].

iii) Current use of QSAR in toxicology

QSAR methods can be used to help in risk assessment of chemical compounds, for example by supporting prioritization for further testing, providing information to fill data gaps and complementing existing experimental data, or by directly replacing some *in vitro* assays and *in vivo*

studies [280].

In particular, QSAR and other computational methods are already applied in the regulatory context. For example, computational toxicology approaches are accepted in lieu of *in vitro* testing for pharmaceutical compounds by the International Conference on Harmonisation (ICH) M7 guideline in order to assess the mutagenic potential of impurities [256]: it requires the prediction of an *in vitro* mutagenicity assay through two complementary methodologies, namely QSAR and rule-based experts. Besides, in a guidance for dietary risk assessment, EFSA also asks for at least two independent models (rule based and ML based) to predict genotoxicity of PPPs and their plant residues, such as QSAR and rule-based [77]. Lhasa Limited [178] (Nexus) and Leadscope⁶ are two examples of companies providing QSAR and rule-based models to enable such predictions, in compliance with the proposed guidelines.

Moreover, the REACH legislation in EU promotes the use of QSAR and the OECD has established several reports regarding the regulatory use of QSAR and other alternative methods. It has also developed, in collaboration with ECHA (European Chemicals Agency), an open-access tool aiming at filling data gaps by providing a workflow that integrates chemical grouping and read-across: the **OECD QSAR Toolbox**⁷ [66]. Nonetheless, it does not directly provides QSAR models.

Another example of the desire of having QSAR models for regulatory purposes is the **CAESAR project** which has been funded by the European Commission. Indeed, this project aimed at developing models for the REACH legislation [15]. Basically, five QSAR models have been developed for the five following endpoints: bioconcentration in fish, skin sensitization, mutagenicity, carcinogenicity and developmental toxicity. These models have been validated according to the five OECD principles mentioned earlier [198] and are freely available on the project website⁸. These models can also be found in the VEGA⁹ software along with other robust and validated QSAR models [17].

In the US, the EPA also developed a tool with QSAR models to allow the users to estimate the toxicity of their compounds: the Toxicity Estimation Software Tool (TEST)¹⁰.

TopKat (Dassault Systèmes, Biovia)¹¹ is another software providing this type of models for various toxic endpoints. In particular the models have been reported to the European Commission Joint Research Center.

Finally, an ensemble method using a Bayesian model has been recently proposed to predict the potential of a chemical to be carcinogenic or not. The input features of the model are the predictions of four open-source QSAR tools used for regulatory risk assessment (including OECD ToolBox) [210].

An extensive review of all the tools and software available for toxicity prediction (including QSAR

⁶<http://www.leadscope.com/>

⁷www.qsartoolbox.org

⁸www.caesar-project.eu/

⁹www.vegahub.eu/portfolio-item/vega-qsar/

¹⁰www.epa.gov/chemical-research/toxicity-estimation-software-tool-test#pubs

¹¹www.3dsbiovia.com/products/datasheets/ds_topkat.pdf

and other methods) has been released as a technical report by the European Commission [97].

Apart from these existing available QSAR models which aim to help for risk and hazard assessment, a lot of work has been performed to develop models to predict various *in vitro* activities or toxic effects.

Examples of tentative of predicted outcomes *in vivo* include: acute oral toxicity [295, 159, 167], carcinogenicity [16, 18], drug induced liver injury [291], hepatotoxicity [174, 169], hepatocellular hypertrophy [5], cytotoxicity [251], mutagenicity [18], *etc.*

Regarding the prediction of *in vitro* bioactivity, if we only consider the *in vitro* data from the ToxCast / Tox21 project, we can already mention a number of studies. First, the Tox21 challenge held in 2014 asked the different competitors to build QSAR models to predict 12 *in vitro* assays related to stress response and nuclear receptor signaling pathways, using a dataset of more than 12,000 compounds [129]. This challenge raised a lot of interest and resulted in good models that were based on different types of methods [36, 71, 1] with the best ones using deep learning [183]. Following this challenge, other models have been developed to predict the same assays [10, 215]. More generally, assays measuring nuclear receptor binding are intensively studied and modeled using QSAR. For example, before the Tox21 challenge, QSAR for estrogen and androgen receptor binding had already been proposed [290, 47] and recently improved using new techniques [193]. Then, Ng *et al.* developed decision tree models to also predict estrogen receptor binding using different sources, including ToxCast [191]. Moreover, Gadaleta *et al.* recently tried to build QSAR models to predict assays related to Molecular Initiating Events of AOPs that lead to hepatic steatosis [96]. These assays include 6 receptor binding assays and one nuclear factor activation assay from ToxCast. Others used data from Tox21 to predict aromatase binding from fingerprints [72]. Finally, a really recent study used several ML approaches to detect potential endocrine disrupting chemicals (EDCs) by predicting their activity against six nuclear receptor targets, related to endocrine disruption (ED) [249].

In summary, these studies show that QSAR models are better when they predict direct chemical effects corresponding to molecular initiating events (binding, activation, *etc.*) than events or outcomes occurring further in the AOPs. Therefore, the development of such predictive models, when sufficiently performing, could help for screening and prioritization of compounds for further testing. However, QSAR is not adapted to the prediction of *in vivo* toxicity and we can anticipate that other types of information are required to reach this goal, such as *in vitro* activity.

3.4 From *in vitro* activity to *in vivo* toxicity

We now focus on the prediction of toxicity observed *in vivo* from *in vitro* data using non-ML and ML computational approaches. A summary of the studies reviewed here is provided at the end of this section in Table 3.4.

3.4.1 Non machine learning approaches

Simple statistical analysis: Statistical measures such as correlation coefficient, odds ratio, Student's t-test, Fisher's test, chi-squared, *etc.* aim at explaining the relationship between variables. In particular, **univariate association** describes the relation between pairs of variables and has been applied to measure the association between *in vitro* assays and *in vivo* outcomes of interest. For example, Chandler *et al.* [38] performed univariate associations between results of tests performed in embryonic stem cells and *in vivo* developmental toxicity using Student's and Fisher's tests (they also applied this method between embryonic stem cells tests and ToxCast assays). Kleinstreuer *et al.* [150] used the same approach between ToxCast *in vitro* assays and *in vivo* carcinogenesis observed in several organs of rats and mice. Indeed, they computed the odds ratio between each pair of assay / cancer endpoint to find the most significant pairs. Judson *et al.* [139] also used correlation metrics to demonstrate the utility of ToxCast *in vitro* data. In particular they found association between some *in vitro* assays and rat liver tumors and they highlighted a negative correlation between the number of pathways disturbed by a chemical (according to the related *in vitro* assays) and the lowest dose at which the compound induces *in vivo* toxicity. An extension of the univariate association is the multivariate analysis where several variables are used to explain one variable of interest.

Moreover, a recent study used various types of methods to evaluate how carcinogenicity could be predicted based on *in vitro* ToxCast assays related to "key carcinogen characteristics" [13]. They attempted to explain the difference between carcinogenic and non carcinogenic compounds based on their activity in *in vitro* assays. To do so, they used descriptive statistics measures including the Negative Predicted Value, the Sensitivity, the Specificity, the p.values of statistical tests and similarity metrics such as the Jaccard coefficient.

Linear additive models: These models simply aggregate the results of several *in vitro* assays into one model by assuming an equal contribution from each assay. This method has been used by US EPA researchers for the development of estrogen receptor activity [142, 33] and androgen receptor activity [149] models. Basically, respectively 18 and 12 ToxCast assays target key events along the ER and AR pathways and chemicals were tested in these assays at different concentrations. For each compound tested in these assays, and each concentration, the models compute a predicted value for the activity of the compound in the global pathway by linearly integrating the results of each of the 18 (resp. 12) assays. Finally, a concentration-response curve was obtained for the entire pathway using predicted activities for all tested concentrations. The AUC of the curve has been computed and corresponds to the score of the model, ranging from 0 to 1. This score was then compared to a threshold in order to determine if the compound was agonist, antagonist or inactive for the corresponding pathway. The results were compared to *in vitro* and short term *in vivo* public data regarding ER and AR activity (including the EDSP results, see Chapter 1, Section 1.2.6) which enabled the researchers to validate their models.

Connectivity mapping: A connectivity map aims at establishing connections between gene expression signatures of compounds and biological processes leading for example to toxic effects or diseases. The method is based on the comparison of a given gene expression profile to several reference profiles using a specific statistical test (Kolmogorov-Smirnov test). This comparison enables assigning a "connectivity score" ranging from -1 (negative connectivity) to 1 (positive connectivity) for each reference profile. Profiles are then ranked according to their score [160]. Connectivity mapping has been applied for toxicity prediction and its ability to detect potential toxicity of compounds with an indication about the involved mechanisms has been shown [243]. Moreover, Caiment *et al.* have demonstrated the usefulness of the method to predict compounds hepatocellular carcinogenicity using the TG-Gates database as a reference dataset [35].

Physiologically-Based Toxicokinetics modeling (PBTK): *In vitro to in vivo extrapolation (IVIVE)* is an approach that aims at transposing *in vitro* experimental data to predict *in vivo* physiological phenomena such as toxicokinetics and toxicodynamics. Toxicokinetics (TK) describes the fate of compounds in living organisms, including Absorption, Distribution, Metabolism and Excretion (ADME). Toxicodynamics (TD) refers to the toxic effects induced by the compounds highlighted by its dynamic interactions with the target molecules [14]. IVIVE analyses are based on quantitative models composed of mathematical equations that numerically simulate the *in vivo* systems using parameters measured in *in vitro* assays.

PBTK (known as PBPK for pharmacology) is a commonly used method for IVIVE to describe the ADME properties of compounds. The method consists of mathematically modeling the physiological processes occurring in several tissue compartments of the living organisms (*e.g.* liver, kidney, brain, skin, *etc.*) when a compound is administered. On the one hand, tissue compartments are characterized by multiple parameters such as volume, blood flows, transport, metabolism, protein abundance, *etc.* and are connected by the blood [14]. On the other hand, compounds are characterized by: (1) physico-chemical properties such as molecular weight, lipophilicity and pKa; and (2) biological properties such as unbound fraction, partition coefficients, permeability, solubility, clearance, *etc.* [157]. In the end, when both organs parameters and compounds properties are known by the model, associated with a given dose and route of administration, it allows the simulation of the concentration of the compound of interest in different tissues as well as its metabolites, at various time points. Note that these models are applied to different species and can also account for inter and intra population variability. There are several commercial software available for PBTK modeling such as PK-Sim (Bayer) [78] and Simcyp (Certara) [135]. PBTK **reverse dosimetry** refers to the use of PBTK models in the reverse order in such a way that they calculate the dose of compound required to obtain a desired concentration in a specific tissue [172]. This approach has been used a lot, for example to predict neurotoxicity, developmental, liver or kidney toxicity [172] or to compare *in vitro* and *in vivo* estrogen receptor activity [40]. More recently, reverse dosimetry has been used to investigate if LOAEL obtained in *in vivo* studies measuring ED endpoints could be extrapolated from observed-effect concentrations measured in *in vitro* assays looking for potential endocrine disrupting compounds [85].

3.4.2 Machine learning approaches

Unsupervised machine learning: Shah *et al.* [235] performed an unsupervised multivariate analysis to compare compounds nuclear receptor (NR) activity measured in *in vitro* assays with several stages of liver cancer lesions observed in rat and mouse *in vivo* studies. Basically, they performed two independent hierarchical clustering to partition the compounds into: (1) 7 nuclear receptors groups (for the *in vitro* data) and (2) 8 cancer lesions progression groups (for the *in vivo* data). For each pair of NR group and lesion group, they computed the ratio of the number of compounds belonging to both groups of the pair and the total number of compounds in the considered lesion group. They could conclude that compounds having a high NR activity were associated with hepatocarcinogens while compounds with low NR activity did induce mild or no lesions at all. Moreover, Sipes *et al.* [241] clustered 976 ToxCast compounds according to their structure and bioactivity results from 331 assays and were able to identify potential targets of compounds as well as possible modes of action.

Another example of unsupervised learning is the use of self-organizing map (SOM) that enables dimensionality reduction and therefore representative visualization of high-dimensional data [153]. SOM has been applied by Huang *et al.* to predict 72 *in vivo* toxicity endpoints from the results of 30 ToxCast *in vitro* assays for around 10,000 compounds (*i.e.* the Tox21 10K library) [130]. Basically, the compounds were initially clustered using a SOM algorithm according to their activity profiles and using the Euclidean distance to measure similarity. Then, these clusters were used to predict the 72 *in vivo* endpoints: for each cluster and endpoint, a toxicity score was computed using a Fisher’s test to represent the toxic potential of the compounds from the cluster. This score considers the proportion of compounds in the cluster that are positive for the endpoint according to the proportion in the entire dataset. These scores were finally used to predict the potential toxicity of a new compound after its assignment to a cluster. Note that this study has also been performed using structural fingerprints instead of *in vitro* activity and the resulting models showed better performances.

Supervised machine learning: Martin *et al.* [181] and Sipes *et al.* [242] proposed supervised ML models that respectively predict rat reproductive toxicity and developmental toxicity based on ToxCast *in vitro* assays results, using LDA. They previously selected *in vitro* assays (among more than 500) that were associated with the endpoints of interest by applying univariate feature selection (*i.e.* they measured the association between each assay and the *in vivo* outcomes using Pearson correlation, Student’s t-test and chi-squared test). The selected assays were then aggregated into groups of assays sharing a same target and the average of AC50 values measured for each assay of the group was computed and assigned to the group. This allowed the reduction of the number of total features and avoided the use of redundant assays. Finally, these aggregated values (one per group of aggregated assays) were used as input features of ML models that predict several reproductive or developmental toxicity endpoints using a LDA algorithm. After cross-validation, the resulting BA were greater than 70%.

Nonetheless, the approaches used by Martin and Sipes have been reconsidered by Thomas *et al.* [259] in a large analysis which compared various scenarios and ML methods to predict 60 *in vivo* endpoints from ToxRefDB, resulting in 84 ML models in total. Among the scenarios was the aggregation of the results of several *in vitro* assays sharing a same target into only one descriptor for ML models and they showed that it did not lead to better performance than the use of unaggregated data. Moreover, they suggested that pre-filtering assays according to a univariate selection prior the cross-validation (*i.e.* only once for the whole learning) indeed inflates the results but they suggested that this was due to an induction of bias into the models. They therefore recommended to use this method during the cross-validation: for each loop of the cross-validation, the prefiltering is performed based on the actual examples of the training set which are different for each loop.

This recommendation has been followed by Liu *et al.* in two studies, in which they developed ML models to predict either three specific endpoints regarding hepatotoxicity [169] or 35 outcomes regarding chronic toxicity of 19 other organs [170], observed in rat *in vivo* studies collected from ToxRefDB. In both studies, they evaluated the performances of different types of ML algorithms and used two to three types of descriptors: molecular descriptors, structural fragments, and results of *in vitro* ToxCast assays (note that structural fragments were used only in the second study). For each endpoint, they built several models with a varying number of descriptors used as input (from 5 to 60 depending on the studies) that were selected inside the cross-validation loop according to the measure of the univariate association between each descriptor and the endpoint (t-test or ANOVA F-value). Finally, for each descriptors type, they looked at the descriptors that were the most frequently used (*i.e.* selected after univariate association) among all the models.

Moreover, in the previously described study that evaluated how carcinogenicity could be predicted based on *in vitro* ToxCast assays, ML models have also been built using Logistic Regression, CARTs, Bayesian methods and Random Forests [13]. All the methods used in this study, whether based on ML or not, showed that it was not possible to predict carcinogenicity from the results of the selected *in vitro* assays.

Finally, other ML models developed for toxicity prediction use genomic profiles as input features. For example, expression profiles of 27 genes of the liver were used in Random Forest algorithms to predict hepatocarcinogenicity [89]. Moreover, as part of a challenge, data from the Tox21 1000 genomes project [4] were used to predict human toxic responses to environmental compounds [76]. These data include the genotype profiling of more than 800 cell lines from various populations, gene expression data for some of these cell lines and cytotoxic assay results performed in these cell lines for more than 100 compounds. Indeed, the 1000 genomes project provides genomics data for around 1000 humans from different populations (European, African, Asia, *etc.*).

Table 3.4: Examples of methods used to link *in vitro* bioactivity and other types of information to *in vivo* outcomes. Methods are grouped into non machine learning and machine learning ones (both unsupervised and supervised). ER: Estrogen Receptor, AR: Androgen Receptor, NR: Nuclear Receptor

Method	Input data	Predicted outcome	Reference
Non machine learning methods			
Univariate association	Embryonic stem cells assays	Developmental toxicity	[38]
	<i>In vitro</i> assays	Rodent carcinogenesis	[150]
Descriptive statistics	<i>In vitro</i> assays	<i>In vivo</i> toxicity in general	[139]
	<i>In vitro</i> "key carcinogenic" assays	Carcinogenicity	[13]
Linear additive models	<i>In vitro</i> assays (ER related)	Estrogen receptor activity	[142, 33]
	<i>In vitro</i> assays (AR related)	Androgen receptor activity	[149]
Connectivity mapping	Gene expression signatures	Hepatocarcinogenicity	[35]
PBTK	<i>In vitro</i> assays	Neurotoxicity, developmental toxicity, liver toxicity, kidney toxicity	[172]
	<i>In vitro</i> assays (ER related)	Estrogen receptor activity	[40]
	<i>In vitro</i> assays (AR/ER related)	Endocrine effects	[85]
Machine learning methods			
Clustering	<i>In vitro</i> assays (NR related)	Hepatocarcinogenicity	[235]
	Chemical structure, <i>in vitro</i> assays	Modes of action	[241]
	<i>In vitro</i> cytotoxicity assays	Rat acute toxicity	[295]
	Chemical structure, bioactivity profiles	72 <i>in vivo</i> adverse outcomes	[130]
Supervised machine learning	<i>In vitro</i> assays	Reproductive toxicity	[181]
	<i>In vitro</i> assays	Developmental toxicity	[242]
	Chemical structure, <i>in vitro</i> assays	60 <i>in vivo</i> adverse outcomes	[259]
	Chemical structure, <i>in vitro</i> assays	Hepatotoxicity	[169]
	Chemical structure, structural fragments, <i>in vitro</i> assays	35 <i>in vivo</i> adverse outcomes	[170]
	<i>In vitro</i> "key carcinogenic" assays	Carcinogenicity	[13]
	Genomic profiles	Hepatocarcinogenicity	[89]
	Genomic profiles	Human general cytotoxicity	[76]
	Chemical structure + <i>in vitro</i> concentration - response	Acute toxicity	[234]
	Chemical structure + <i>in vitro</i> concentration - response + target affinity	Acute toxicity	[2]
	Structural fingerprints + HTS data	General compounds' activity	[219]
	Chemical structure + transcriptomics	Hepatotoxicity	[174]
	Chemical structure + imaging assays	Hepatotoxicity	[297]
Chemical structure + MOA information	Drug induced liver injury	[281]	

3.4.3 Combination of several types of information

As already mentioned previously, some studies tried to combine several types of information to improve *in vivo* toxicity predictions [173]. For example, Huang *et al.* used SOM to cluster compounds according to either their bioactivity profiles or their chemical structure. They also combined the two types of clusters to generate "consensus clusters" and used them as input of ML models for 67 endpoints. They concluded that performances of the models based on combined data were higher than the ones of the structure-based models and the bioactivity-based models [130]. Other works also proposed to use clustering methods to integrate several information, for example for rat acute toxicity prediction [295].

The most common way of using several types of information is to perform data pooling, meaning

merging the data into "hybrid" sets of descriptors for further modeling [173]. For instance, Sedykh *et al.* pooled concentration-response data obtained in cell based assays available in PubChem and chemical structure information to predict acute toxicity and showed that the models had better performance than the basic QSAR ones [234]. This study has been extended by adding a third type of descriptors: the protein target affinity information, which again resulted in improved accuracy of predictions [2]. Then, Riniker *et al.* transformed HTS data from PubChem assays into so called binary or float HTS-fingerprints and showed that models built with both structural fingerprints and HTS-fingerprints were leading to better or similar performance than QSAR models based on structural fingerprints only [219].

However, this improvement is not clear in every study. Indeed, in their large analysis for the prediction of 60 *in vivo* endpoints, Thomas *et al.* [259] also compared the performance of models based on either *in vitro* assays alone, molecular descriptors alone, or a combination of both. They were able to conclude that ML models based on *in vitro* assays do not perform better than those based on structural descriptors and that the combination of both types of descriptors does not improve the performance.

These conclusions have also been obtained when combining structural information to either transcriptomics [174] or imaging assays [297] results to predict drug hepatotoxicity.

Moreover, in their two studies [169, 170], Liu *et al.* also built ML models based on a combination of different types of descriptors. For hepatotoxicity, the combination of structural descriptors and *in vitro* bioactivity data did not lead to significantly higher performances than the use of structural descriptors alone. Nonetheless, the second study demonstrated that a combination of bioactivity data with chemical descriptors or structural fragments is slightly more predictive than the use of chemical descriptors or structural fragments alone.

Finally, MOA information obtained in *in vitro* assays have been combined to chemical structure information to predict drug-induced liver injury (DILI) and resulted in a small improvement of performance compared to simple QSAR models [281]. In particular, 4 assays have been identified to be more predictive of DILI.

Apart from data pooling which has been exemplified here, other approaches have been proposed to integrate chemical and biological data. These methods include [173]:

- **Model pooling:** it is based on ensemble modeling. For example, the results from chemical-based models and biological-based models are pooled in order to make a consensus vote to obtain the final predictions.
- **Network modeling:** known associations between chemicals, biological targets, bioactivities and toxicity endpoints are modeled into graphs where edges represent the associations between the different entities (represented by nodes). These networks may consider several types of associations such as direct interactions or statistical correlations and aim at inferring new associations from the existing ones. This type of networks has been mainly proposed in pharmacology in order to identify drug-target interactions [26].

- Use of chemical structure to **estimate biological parameters**: this method is mostly used for PBTK when some required parameters for modeling are unknown. In such cases, QSAR is applied to estimate these parameters.

3.5 Summary

In this chapter we first introduced the existing data in toxicology and then proposed a large overview of the state of the work related to computational methods applied to these data.

We showed that: (1) there is a high variety and variability of toxicological data as well as a lack of standardization in data sources and (2) there is an important diversity of the types of methods and learning algorithms used to establish the links between the different types of data.

Consequently, it is difficult to define universal criteria to select the most appropriated computational approaches for toxicity modeling and prediction. This can be partly due to the "No Free Lunch" theorem but also to the variability of toxicological data (in term of quality, quantity, diversity of acquisition methods, *etc.*) and above all, to the long and complex chain of causality between a Molecular Initiating Event (MIE) and a long-term outcome. Moreover, we can also point out the key importance of the physiological relevance of *in vitro* assays and the lack of quantitative translation from *in vitro* to *in vivo* (from μM to mg/kg/day). Therefore, the community encounters difficulties to take out some recommendations from the experience of the last 15 years that could be applied for a large diversity of problems. This is particularly illustrated by the numerous studies that are permanently published regarding the development and evaluation of new alternative methods.

CHAPTER 4

FROM STRUCTURE TO ACTIVITY: A PRELIMINARY STUDY GUIDED BY THE TWO-STAGE APPROACH

In this chapter we focus on the first stage of the two-stage approach whose goal is to build machine learning models that predict the activity measured in *in vitro* assays from the chemical structure of compounds. As already described in Chapter 3, this type of relation is known as Quantitative Structure-Activity Relationship (QSAR). Here we focus on the results of 37 *in vitro* assays from the ToxCast database and we build models to classify the compounds into actives or inactives for each of these assays. We use different types of machine learning methods and we apply data augmentation to balance the original datasets.

Part of this chapter has been published in the international journal *Sensors* [107] and has also been presented at the *International Conference on Artificial Neural Networks* [108].

4.1 Machine learning on datasets constrained by *in vivo* data

4.1.1 Data used

Since the objective of the two-stage approach is to predict *in vivo* outcomes, the datasets are restricted to compounds for which both *in vitro* and *in vivo* data are available. We therefore use the 3 sources of data provided by the US EPA to build the datasets:

- the DSSTox database which provides more than 9,000 unique chemical structures as SDF files (October 2015 release);
- the ToxCast database which gathers AC50 measured for about 10,000 compounds tested in more than 1,000 *in vitro* assays (October 2015 release);
- the ToxRef database which collects the results of different types of *in vivo* studies performed on several hundreds of compounds (October 2014 release).

As already mentioned previously, it is important to use correct data to build relevant models. Therefore, we proceed to a data curation of the chemical structures by applying an automated workflow implemented in Knime [22] which performs the following tasks: removal of inorganics and salts, conversion into 2D structures, standardization of chemical functions, conversion into canonical SMILES and removal of duplicates. In the end, 8325 unique and cleaned chemical structures are kept.

Since we aim to build classifiers, we transform the AC50 continuous values from ToxCast into binary ones. As stated in Section 3.2.2, when no activity had been measured for a compound, its reported value for the considered assay is 1,000,000. We then set 1 if the AC50 is comprised into the interval $[0; 1,000,000[$ and 0 if it is equal to 1,000,000.

Based on the three databases, the selection of the *in vivo* constrained datasets is obtained according to the three following steps, illustrated in Figure 4.17(b).

i) Looking for the overlap between structural, *in vitro* and *in vivo* data

First, we look for the overlap of compounds which are common to the three databases and for which *in vivo* results of studies performed in rats during two years are available in ToxRefDB. This results in a matrix composed of 418 compounds and 915 available ToxCast assays. Nonetheless, since not all the compounds have been tested in all the assays, this matrix is incomplete and contains 16% of missing values. We therefore need to look for a complete matrix in order to avoid any noise that can be induced by missing values.

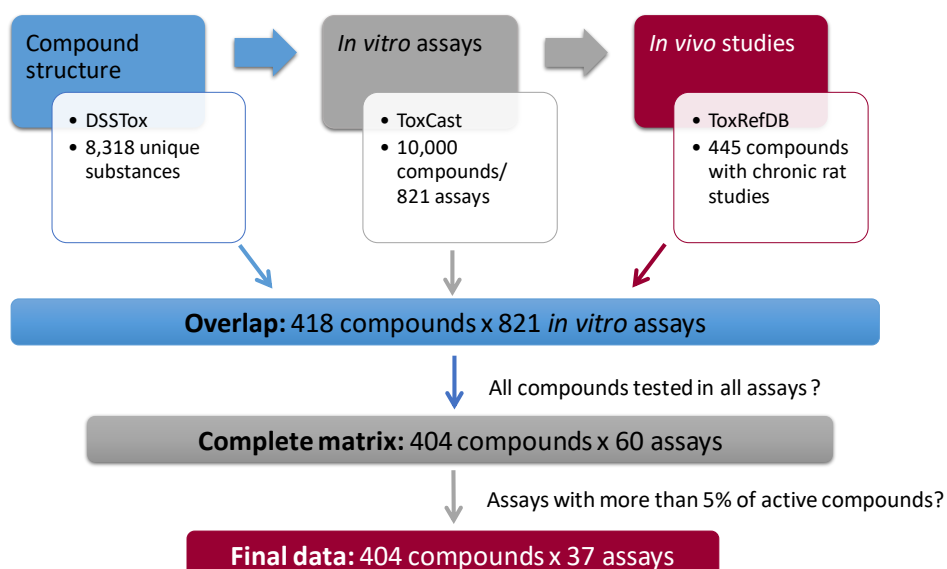


Figure 4.1: Schema of the process to obtain the final data for the *in vivo* constrained datasets. Starting with the available compound structures, *in vitro* assay results and *in vivo* studies, we first look for the overlap of compounds, then for the corresponding complete matrix and we finally remove assays containing less than 5% of active compounds. The final matrix corresponds to 404 compounds and 37 assays.

ii) Looking for the maximal biclique: a NP-complete problem

The matrix can be considered as a **bipartite graph** which is commonly used to represent relationships between two different types of data constituting two disjoint sets of vertices [293]. In our case, the two sets of vertices correspond to the 418 compounds and the 915 assays and edges between these two sets inform about an existing relationship between one assay and one compound (Figure 4.2). Basically, when a compound has been tested in an assay, an edge is present between these two vertices; it is absent otherwise.

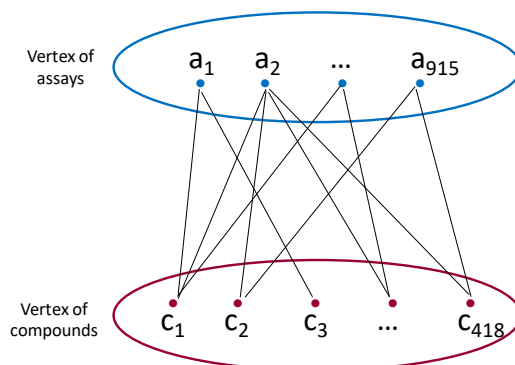


Figure 4.2: Schema of a bipartite graph. One vertex corresponds to the 915 assays and the other one to the 418 compounds.

Looking for a complete matrix of n compounds and m assays where all compounds have been tested in all assays is equivalent to finding a subgraph of the bipartite graph where all edges exist between the two subsets of vertices. This complete subgraph is called a **biclique** [206]. Since we want to maximize the number of compounds and assays in our matrix, we should look for the maximum biclique which is the largest one of the bipartite graph.

The maximum biclique problem has been largely studied since more than 50 years [177] and can be broken down into two problems:

- The **vertex maximum** biclique that looks for a biclique with the maximal number of vertices,
- The **edge maximum** biclique that looks for a biclique with the maximal number of edges.

If the first problem has a polynomial complexity, the second one has been proved to be **NP-complete** by reduction to the clique problem [206] which itself has been shown NP-complete by reduction to the 3SAT problem [144]. Moreover, the number of maximal bicliques can be exponential in the graph size. Nonetheless, several algorithms have been proposed to face the edge maximum biclique problem and the last one is called "**Maximum Biclique Enumeration Algorithm**" (MBEA) and is able to find all maximal bicliques of a bipartite graph [293]. The MBEA was inspired by the well known Bron-Kerbosh algorithm which has been designed to find all cliques of undirected graphs [32]. Basically, it combines branch and bound technique and pruning to efficiently eliminate paths of the tree search that do not include maximal bicliques. We applied the MBEA on our graph composed of 418 compounds and 915 assays (418×915) with

a total number of 319,700 edges (*i.e.* number of existing pairs of assay and tested compound). The algorithm found a total of 3302 bicliques, a vertex maximal biclique of size 214×905 and an edge maximal biclique of size 324×745 . However, since ML performs better with larger number of examples in the datasets, we decide to use the biclique composed of the largest number of compounds. Among the 3302 bicliques found by the MBEA algorithm, the one that corresponds to this constraint is of size 404×60 . It has a number of assays lower than those of the maximal bicliques (*i.e.* 905 and 745) but it is sufficient for the purpose of this study.

iii) Assays filtering

Now that we have a complete matrix with the largest possible number of compounds (404), we finally proceed to a filtering of assays and remove all the ones with less than 5% of active (or inactive) compounds so that we work with datasets with a reasonable minimum number of observations in the two classes. This corresponds to 23 assays and we finally end up with a matrix of 404 compounds and **37 assays**.

Figure 4.3 shows the percentage of positive compounds in the datasets corresponding to these 37 *in vitro* assays. We observe that the datasets are highly imbalanced in favor of negative compounds, the ratio between positive and negative ranging from 5% (assay number 11) to 30% (assay number 17), with a mean around 12%.

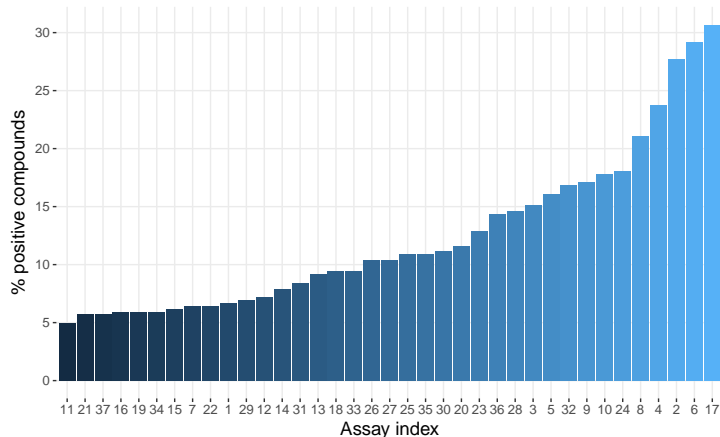


Figure 4.3: Percentage of positive compounds (*Y-axis*) in the 37 *in vivo* constrained datasets corresponding to the 37 *in vitro* assays (*X-axis*).

All these 37 *in vitro* ToxCast assays have been performed by the NIH as part of the Tox21 project and they mostly target nuclear receptors (Androgen Receptor (AR), Estrogen Receptor (ER), Glucocorticoid Receptor (GR), Peroxisome Proliferator-Activated Receptor (PPAR), Tyrosine Receptor (TR)) and the transcription factor p53. The full list of assays with their corresponding target is provided in Appendix B.

4.1.2 Generation and selection of chemical descriptors

Two types of chemical descriptors are computed from the SDF of the 404 compounds:

1. **74 physico-chemical** properties such as the molecular weight, the number of bonds, the solubility, *etc.*, computed using the RDKit Open-Source software¹;
2. **4870 fingerprints** computed using the Python's *pybel* package [195] and the PaDEL software [285].

Then, the physico-chemical properties are normalized into the interval $[0, 1]$ using the min-max normalization [204] and fingerprints being present in less than 5% of compounds are removed, leading to a final number of **731** fingerprints.

4.1.3 Learning procedure

For each of the 37 *in vitro* assays, we build 8 different models using 8 learning algorithms based on three types of methods: neural networks, tree ensemble and SVM (see Section 2.3.2). The details of the chosen algorithms are the following:

- **Radial Basis Function Network (RBF)**: it is a one-hidden layer ANN where each neuron uses a radial basis activation function in the form of the Gaussian function to compute the output.
- **Random Neural Network (RNN)**: it is a model of the spiking probabilistic behavior of biological neural systems where each single cell (neuron) receives spikes from the cells from the previous layer [99]. The structure and weights of the RNN are determined using the cross-validation approach developed in [294].
- **Multi-Layer RNN (MLRNN)**: it is a deep learning extension of the RNN which proposes a multi-layer architecture composed of several hidden layers of single cells [288]. The structure is fixed as having 20 inputs and 100 intermediate nodes and a cross-validation approach is used to determine the optimal structure and weights of the network; 20 trials are conducted to average the results.
- **Dense RNN**: it is a more complex extension of the RNN where cells are grouped into clusters such that they all receive the same spikes from the cells of the previous layer [100]. In each cluster, cells communicate with each other in a RecNN using direct interactions called "soma-to-soma" interactions. This is also inspired by the human brain where some areas are composed of densely grouped neurons which directly communicate together. Similar to the MLRNN, 20 trials are conducted to average the results and the structure of the Dense RNN used is composed of 20 inputs and 100 intermediate nodes.
- **Convolutional Neural Network (CNN)**: it is one of the well-known types of deep learning networks based on weight-sharing and composed of convolutional, pooling and fully connected layers [163]. Here, we use the following structure of layers: input - convolutional - convolutional - pooling - fully connected - output, organized in sequence [49].

¹RDKit: Open-source cheminformatics; <http://www.rdkit.org>

- **Multi Layer Perceptron (MLP):** it is another commonly used method in deep learning, which is generally a multi-layer fully-connected neural network [165]. It can use different types of activation functions and here we use the following: $\frac{e^{ax}}{1+e^{ax}}$ with $a \geq 0$.
- **XGBoost:** it is a tree ensemble method that uses gradient boosting. The open-source software library XGBoost [46] provides an easy-to-use tool for implementing it.
- **SVM:** a support vector machine that uses a Radial Basis Function kernel. Here we use the source code described in [39] and available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.

In order to evaluate the performance of the classifiers, each *in vivo* constrained dataset is randomly and equally separated into a training set (50%) and a testing set (50%). The process is repeated 50 times and the performance metrics are computed as the average of the 50 experiments. We use the four following metrics to evaluate the performance: sensitivity, specificity, balanced accuracy (BA) and ROC score.

4.1.4 Results on imbalanced datasets

Results of the 8 algorithms for the 37 *in vitro* assays

Figures 4.4 and 4.5 summarize the mean-value of the four metrics obtained respectively on the training sets and the testing sets for the classifiers that predict all 37 assays using the 8 methods described previously. Figures 4.6 and 4.7 represent the standard-deviations computed for each metric over the 50 runs of learning.

From the BA mean values, as shown in Figures 4.4(a) and 4.5(a), the MLP, CNN, DenseRNN, MLRNN, XGBoost and RBFN classify the training and testing datasets well while the RNN and SVM fail to do so. However, MLP, CNN, XGBoost and RBF reach really high performance on the training datasets and when comparing to the performance of testing set, this denotes overfitting.

Concerning the testing results, none of the method is able to reach good sensitivity, whatever the *in vitro* assay, but the specificity is high meaning that the classifiers tend to predict everything as negative. This is explained by the imbalanced property of the datasets since sensitivity is higher when datasets are more balanced. As an example, the assay 17 is the most balanced one with around 30% of positive compounds and the corresponding classifiers reach higher sensitivity (except for the SVM).

Regarding the ROC AUC means, all methods, including the RNN and SVM, obtain acceptable training and testing values. The reason could be that the classification thresholds for the methods trained with unbalanced training datasets may need to be adjusted carefully and accordingly rather than using the standard one (*e.g.*, 0.5). Indeed, the thresholds could be chosen based on the proportion of positive instances. Because we work on 37 distinct datasets and assays and since the best decision threshold is unique to each model, we do not investigate the effect of the thresholds here. Nonetheless, it would have been worth to do it if we had wished to develop one

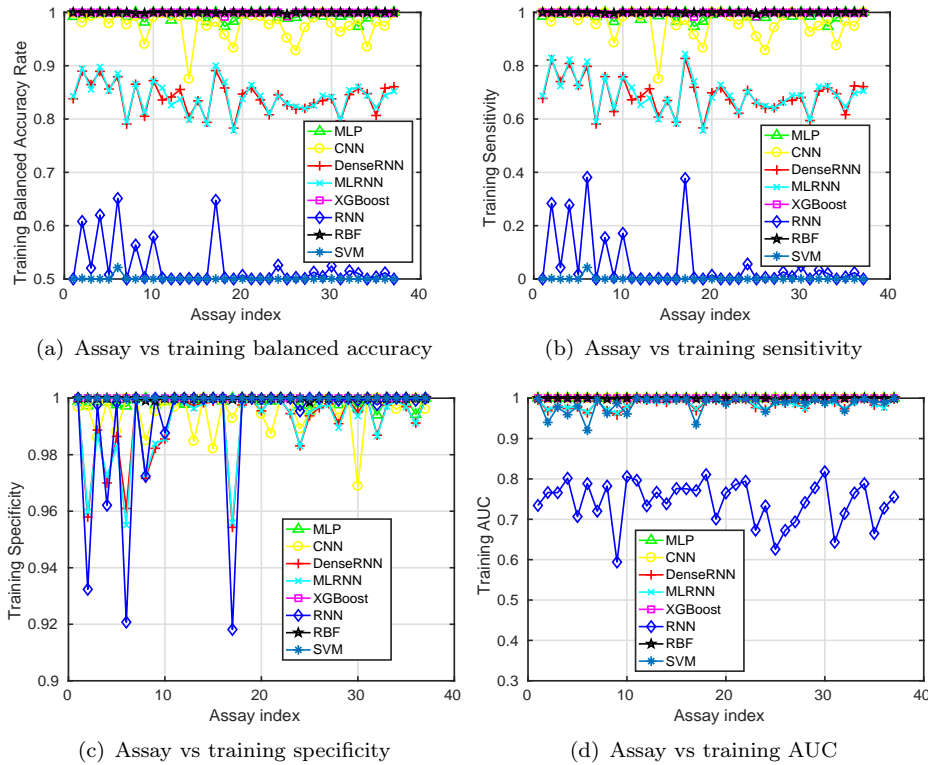


Figure 4.4: Training mean-value results (Y-axis) versus different assays (X-axis) when the MLP, CNN, DenseRNN, MLRNN, XGBoost, RNN, RBFN and SVM are used to classify the in vivo constrained datasets.

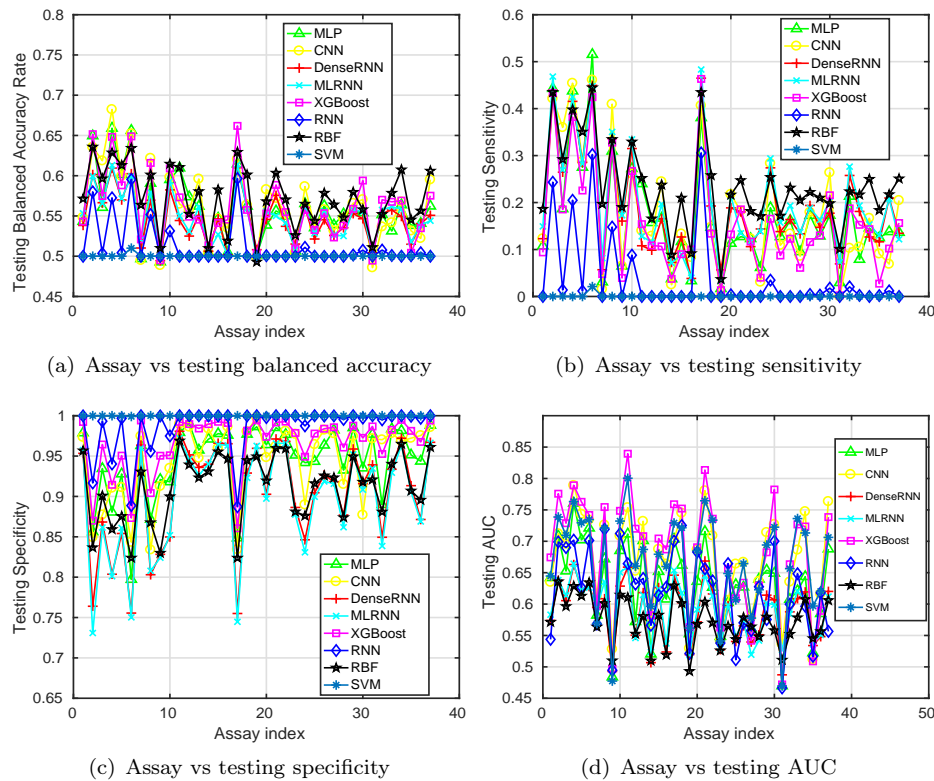


Figure 4.5: Testing mean-value results (Y-axis) versus different assays (X-axis) when the MLP, CNN, DenseRNN, MLRNN, XGBoost, RNN, RBFN and SVM are used to classify the in vivo constrained datasets.

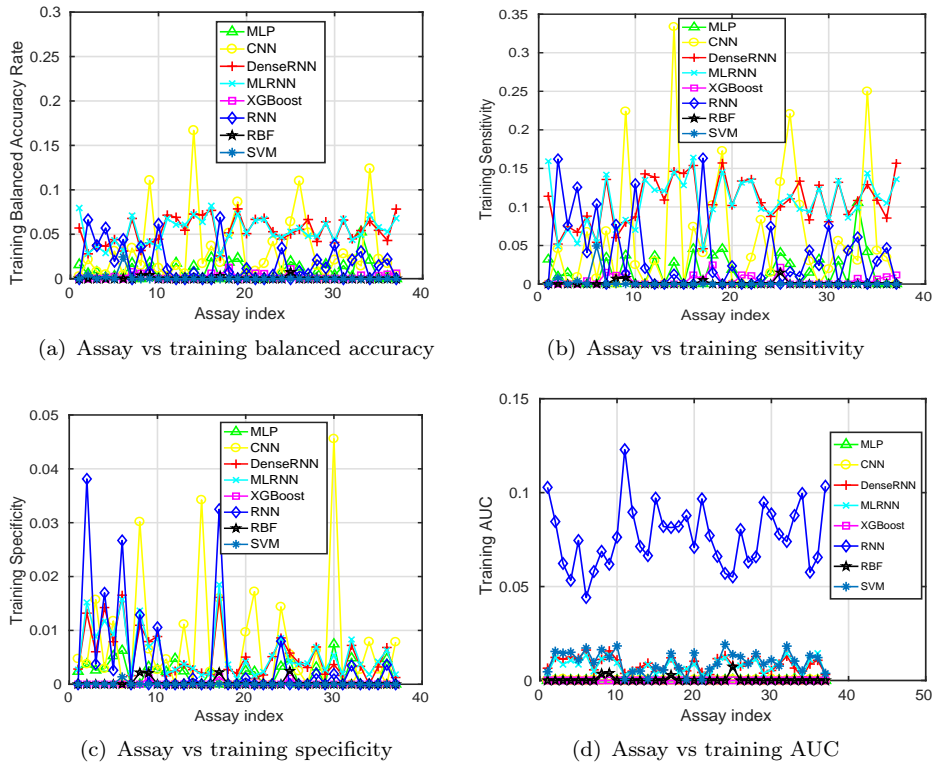


Figure 4.6: Training standard-deviation results (Y-axis) versus different assays (X-axis) when the MLP, CNN, DenseRNN, MLRNN, XGBoost, RNN, RBFN and SVM are used to classify the *in vivo* constrained datasets.

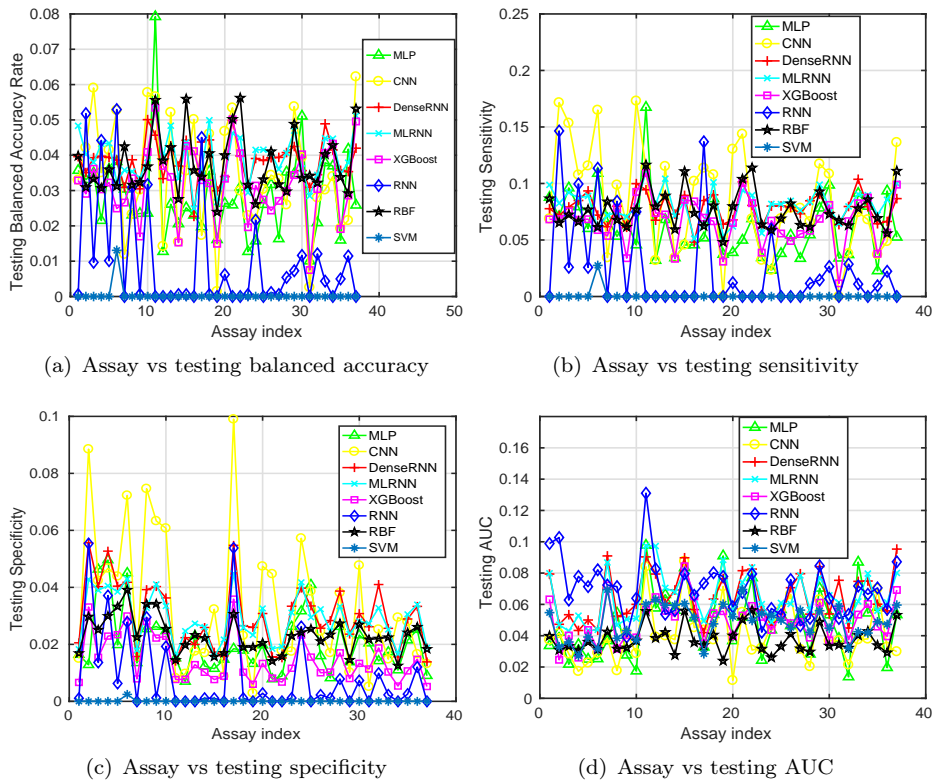


Figure 4.7: Testing standard-deviation results (Y-axis) versus different assays (X-axis) when the MLP, CNN, DenseRNN, MLRNN, XGBoost, RNN, RBFN and SVM are used to classify the *in vivo* constrained datasets.

or two particular ML models, as Chen *et al.* did for three datasets related to cancer using four ML methods [45].

Finally, the standard deviations values summarized in Figures 4.6 and 4.7 are in concordance with the mean values. Indeed, they equal 0 in the case of RNN and SVM for BA, sensitivity and specificity since these algorithms are not able to learn well and therefore predict everything as negative. For the other methods, standard deviations are varying in the same way that their corresponding mean values and they are higher for sensitivity compared to BA and specificity since true positives are the most difficult to predict due to imbalanced data.

Overall, these results show that the algorithms tend to lead to comparable performance when applied to this type of data. More particularly, the sensitivity is quite low and the specificity is high but the results vary between assays and depend on their corresponding datasets. In order to demonstrate this relation between performance and the number of positive compounds in the datasets, we then look at their correlation.

Impact of the number of positive compounds on the performance

To investigate the performance differences among the assays, we rank the testing BA and ROC AUCs of the 37 assays using the **average-rank** ranking method [29], where a higher rank represents a better performance. Basically, for each of the 8 algorithms, the 37 datasets are ordered according to their BA (resp. ROC AUCs) mean values or standard deviations computed after the 50 runs of learning. Then, the average rank of each of the 37 datasets is computed over the rankings obtained for the 8 algorithms. For better comparisons, values of all average ranks are linearly normalized into the interval [0;1]. The method is illustrated in Table 4.1 for an example considering 4 assays, 3 algorithms and arbitrary BA mean values.

Table 4.1: Example of the average-rank ranking method. The example is for 4 assays and 3 algorithms: first, for each algorithm, a rank is given to the 4 assays according to their average BA mean values computed during the 50 runs of learning. Then, the average rank of each assay over the 3 algorithms is computed and normalized into the interval [0;1].

	Algo 1		Algo 2		Algo 3		Average rank	Normalized rank
	Average BA mean	Assay rank	Average BA mean	Assay rank	Average BA mean	Assay rank		
Assay 1	0.7	1	0.8	1	0.7	2	1.33	1
Assay 2	0.6	4	0.74	4	0.68	3	3.66	0
Assay 3	0.65	3	0.76	2	0.71	1	2	0.71
Assay 4	0.67	2	0.75	3	0.66	4	3	0.28

Figures 4.8(a) and 4.8(b) respectively show the normalized average ranks of mean values and standard deviations of testing BA and ROC AUCs for the 37 assays. The proportion of positive instances in the 37 datasets (ranging from 5 to 30%) is also displayed on the two figures and assays are arranged in ascending order of percentage of positives along the x axis.

For the results of mean values, the correlation coefficient between the proportion of positive in-

stances and the BA (resp. the AUC ranks) equals 0.57 (resp. 0.38) with p-value of 2.410×10^{-4} (resp. 2.0×10^{-2}). This positive correlation can be seen in Figure 4.8(a), in particular when looking beyond the assay 19. Nonetheless, this correlation is not clear at each point of the plot suggesting that there are some assays for which the ML methods used are not suitable. Generally, assays with sufficient positive instances (*e.g.*, Assays 2, 6 and 17) have a high average rank meaning that they tend to have a relatively high testing BA and ROC AUC whatever the algorithm used to build the classifiers. On the contrary, for the assays with few positive instances (*e.g.*, Assays 14 and 19), the testing performance tend to be low.

The results of Figure 4.8(b) clearly highlight a negative correlation between the standard deviations of BA and AUC ranks and the proportion of positive instances in the datasets and coefficient correlation values are respectively -0.72 and -0.91 with p-values of 1.9×10^{-7} and 2.3×10^{-15} . Therefore, the more the datasets are balanced, the more the ML methods behave similarly: if the average rank over all algorithms is close to 1, this denotes that they are all performing well on the considered dataset. However, when the number of positive instances in a dataset is low, performance tend to vary much more according to the methods.

Overall, these results show that the number of instances from the minority class (*i.e.* the positive compounds here) is correlated with the performance of the models. We now wonder if an equivalent number of positive and negative compounds would increase the performance and we use a **data augmentation** technique to generate balanced datasets.

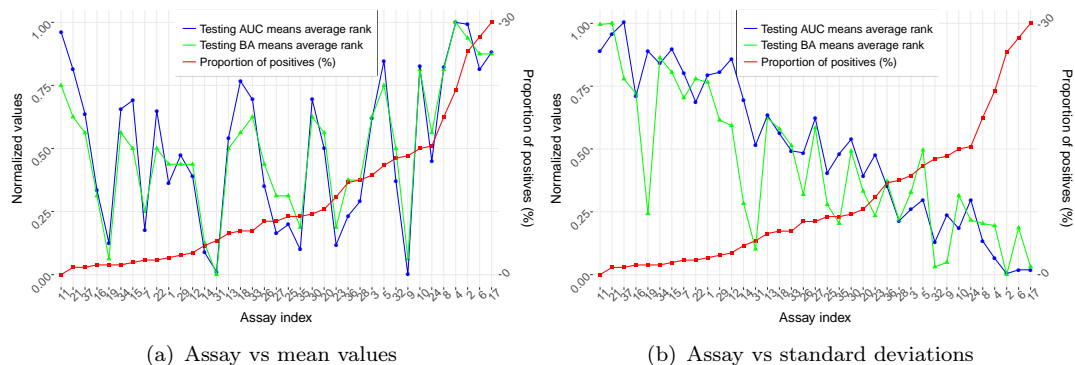


Figure 4.8: Average ranks of testing balanced accuracies and testing AUCs means (a) or standard deviations (b) (left Y-axis) and the proportion of positive instances (right Y-axis) versus the 37 assays arranged in ascending order of percentage of positives (X-axis) when the MLP, CNN, DenseRNN, MLRNN, XGBoost, RNN, RBFN and SVM are used to classify the *in vivo* constrained datasets.

4.1.5 Results on balanced datasets

In order to create balanced datasets, we apply the SMOTE [43] data augmentation technique on the training sets while the corresponding testing sets remain unchanged (see Section 2.5).

As previously, Figures 4.9, 4.10, 4.11 and 4.12 summarize the mean-value and standard deviation results of the four metrics obtained on the training sets and on the testing sets using the 8 ML methods for the 37 assays.

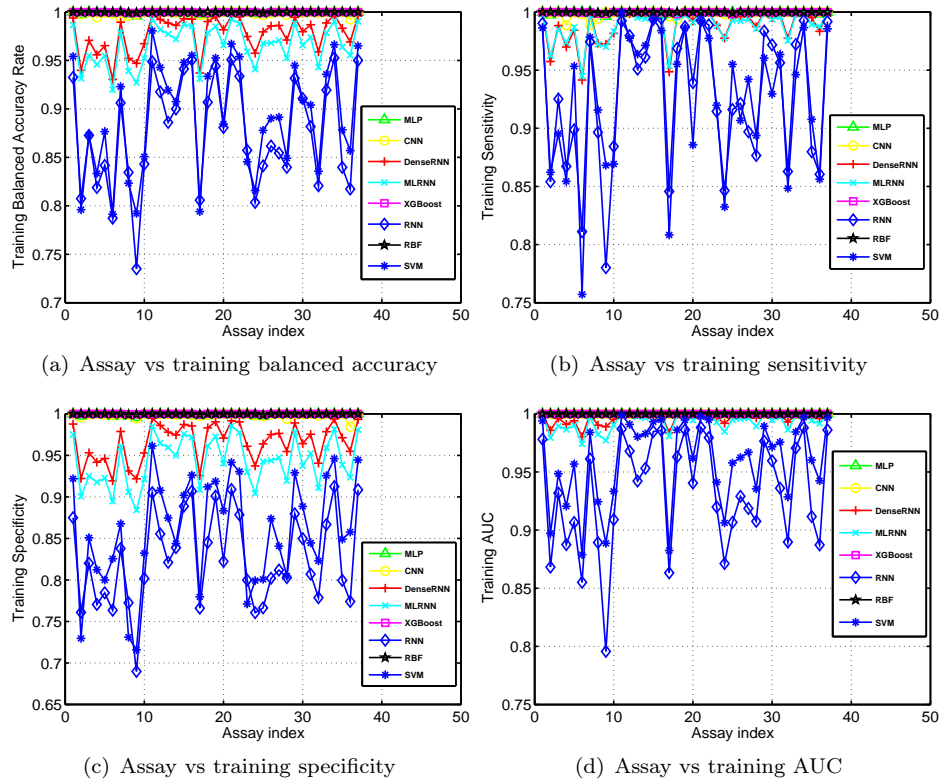


Figure 4.9: Training mean-value results (Y-axis) versus different assays (X-axis) when the 8 ML methods are used to classify the *in vivo* constrained datasets after data augmentation.

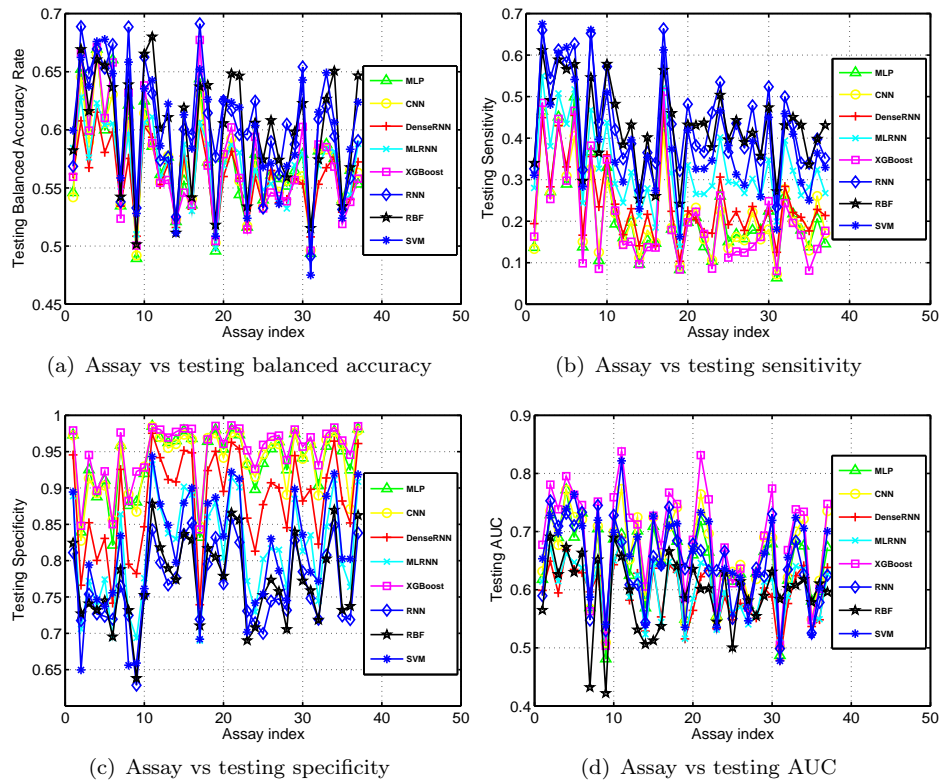


Figure 4.10: Testing mean-value results (Y-axis) versus different assays (X-axis) when the 8 ML methods are used to classify the *in vivo* constrained datasets after data augmentation.

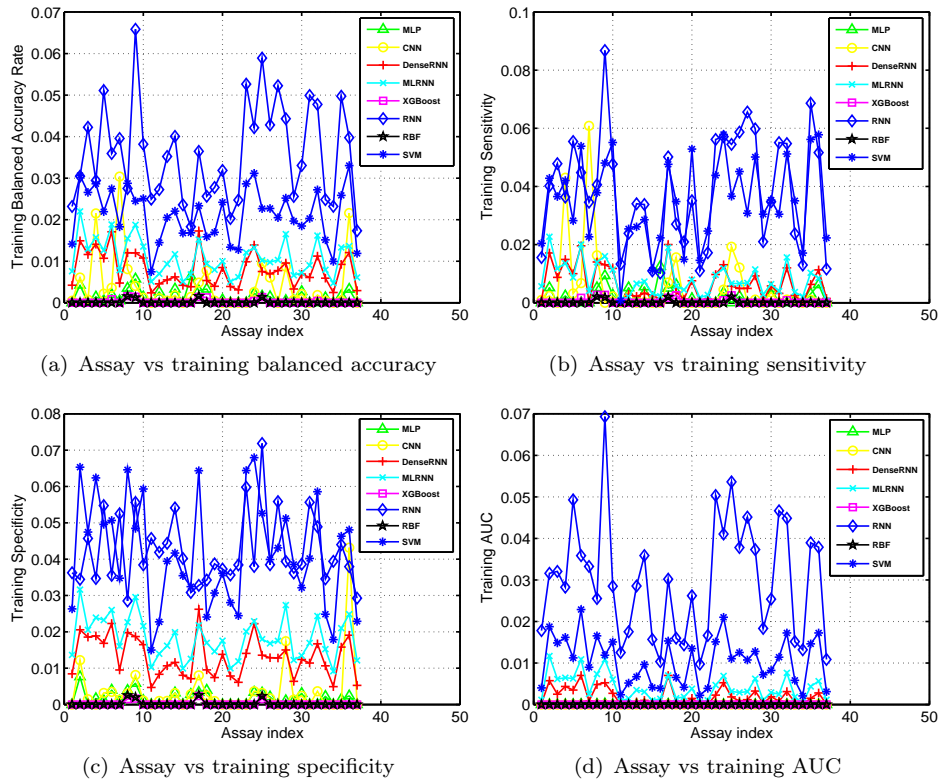


Figure 4.11: Training standard-deviation results (Y-axis) versus different assays (X-axis) when the 8 ML methods are used to classify the *in vivo* constrained datasets after data augmentation.

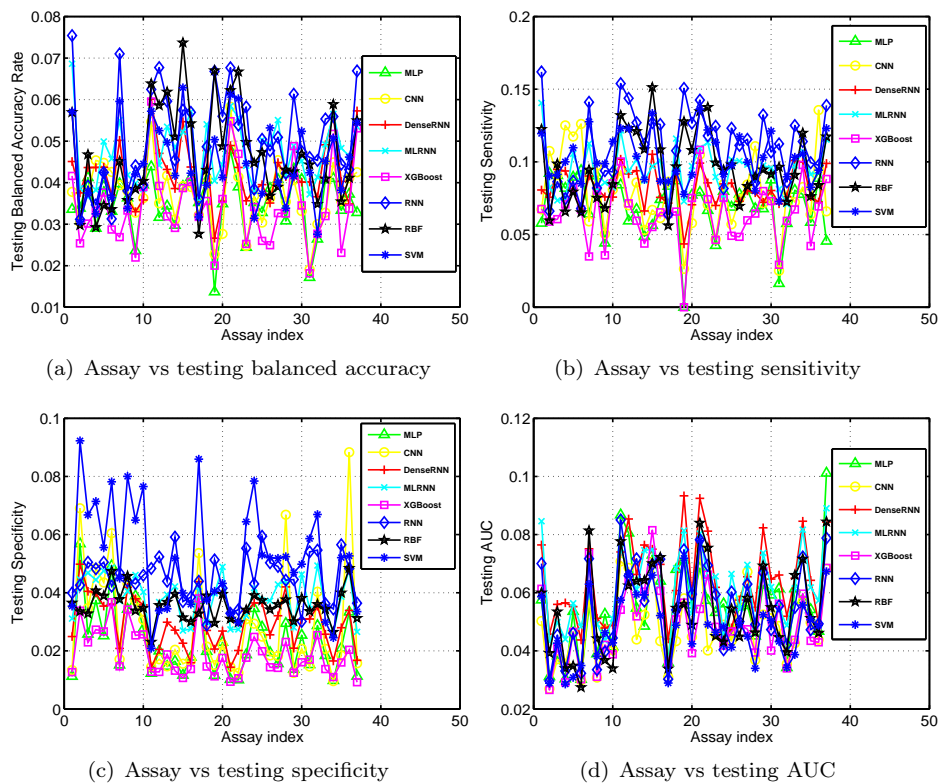


Figure 4.12: Testing standard-deviation results (Y-axis) versus different assays (X-axis) when the 8 ML methods are used to classify the *in vivo* constrained datasets after data augmentation.

According to Figure 4.9, we see that training BAs are always above 70%, even using the RNN and SVM methods which are still the ones that perform worse. Moreover, we see that the training sensitivity is largely increased compared to Figure 4.4(b) with the lowest values around 75%. Nonetheless, we see that RBF, XGBoost, CNN and MLP result in training values close or equal to 1 which typically suggests overfitting since this trends is not observed anymore for testing results where all the algorithms have varying performance according to the assays. In particular, comparing Figures 4.10(a) and 4.10(b) (based on balanced training datasets) with Figure 4.5(a) and 4.5(b) (based on imbalanced training datasets), we see that BAs and sensitivities obtained on the testing datasets for some assays are increased with data augmentation. Finally, as already observed for imbalanced datasets, standard deviations values (Figures 4.11 and 4.12) are consistent with their corresponding mean values. Indeed, methods with the lowest training performance have larger training standard deviations (RNN and SVM) and these values equal 0 for methods that suffer from overfitting. After testing, similar to the mean values, the standard deviations vary according to the assays for all algorithms. Furthermore, they are in the same range than the ones obtained on imbalanced testing datasets (Figure 4.7).

Figure 4.13 presents a comparison between the results based on unbalanced and balanced training datasets. For each assay, we report the highest testing BA, sensitivity, specificity and ROC AUC achieved on unbalanced and balanced dataset and that we previously observed in Figures 4.5 and 4.10. In consequence, the algorithm that led to these values can differ between unbalanced and balanced results as well as between assays.

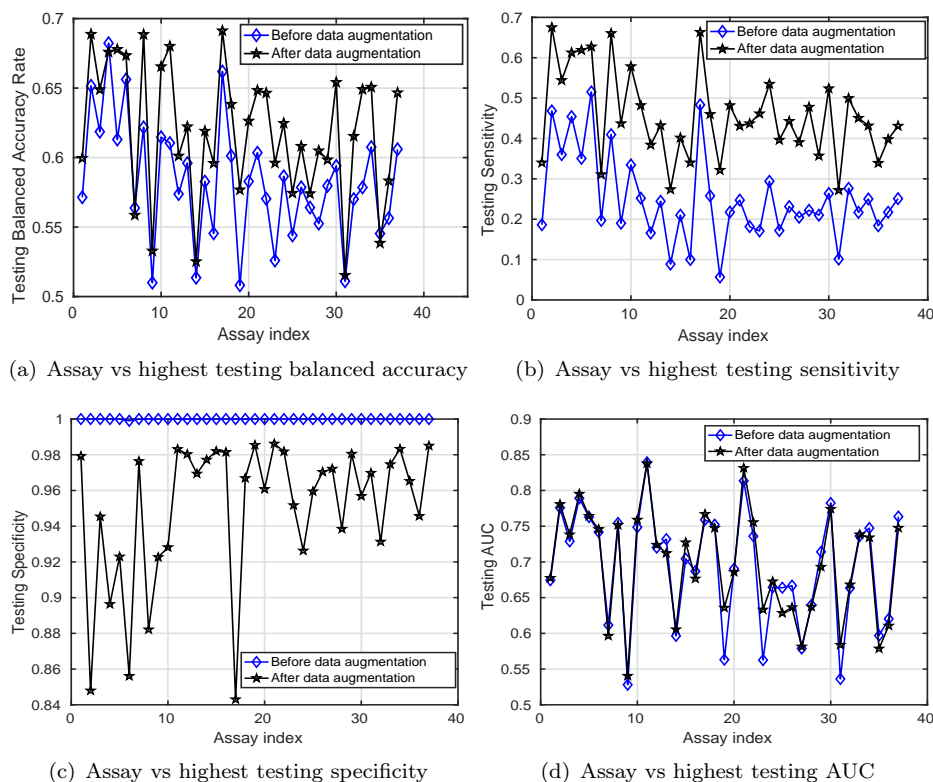


Figure 4.13: Highest testing results (Y-axis) versus different assay index (X-axis) of the 8 ML methods to classify the *in vivo* constrained datasets before and after data augmentation.

We observe that the highest testing BAs and testing sensitivities for most assays are increased after data augmentation. This suggests that, when training datasets have been artificially balanced, the classifiers are better in detecting positive compounds and do not predict almost all compounds as negative. This finally leads to an increase of the sensitivity and a decrease of the specificity to more acceptable values.

All these results show that data augmentation is able to increase performance of models but for most of the assays the testing performance is still not sufficient.

4.1.6 Analysis of the chemical descriptor space of the compounds

In order to explore the distribution of the compounds in the chemical space, we apply a **Principal Components Analysis (PCA)** on the compounds of the training set to reduce their 805 chemical descriptors to two principal components. We then project the compounds of both the training set and the test set on this two-dimensional space. This analysis has been performed for the 37 datasets and led to the same results for all of them: Figures 4.14(a) and 4.14(b) illustrate the results for assays 14 and 17, respectively. These figures display the plots obtained for the datasets of the two assays which respectively contain 8 and 30.7% of positive compounds and for which the performance of the classification task were the lowest and the highest.

From the figures of the PCA, we can not see any pattern in the plots that separates positive and negative compounds but since the two first PCs explain only 25% of the variance, we could think that a better separation would be observed in a higher dimension space. However, this is probably not the case because the percentage of variance explained by the third PC and the following is already low (*e.g.* PC 3 and 4 respectively explain 7% and 4% of variance for assay 17).

Most of the compounds of the test set (crosses) are included in the domain delimited by the compounds of the training set meaning that they are well represented by the training set. Nevertheless, for the few test compounds that are not within this domain but not far from it, we can wonder if they are well predicted by the classifiers. This brings out the interest of defining an **applicability domain** to assess the reliability of a machine learning model and its predictions. Next, to look at the effect of the data augmentation, we project the observations generated by the SMOTE method on the same PCs than previously (*i.e.* the ones obtained with the training set compounds before SMOTE). Figures 4.14(c) and 4.14(d) are the resulting plots for assays 14 and 17 and clearly show that the new "synthetic" compounds are inside the space delimited by the compounds of the original training set. This observation is not surprising since SMOTE is an interpolation method. Therefore, this data augmentation technique cannot help to improve the predictions for compounds that are not included in this space.

According to the classification task results, performances of some assays are low (in particular the sensitivity) even after applying data augmentation. Furthermore, the PCA analysis of the chemical space pointed out that, whatever the ratio of positive compounds, no pattern

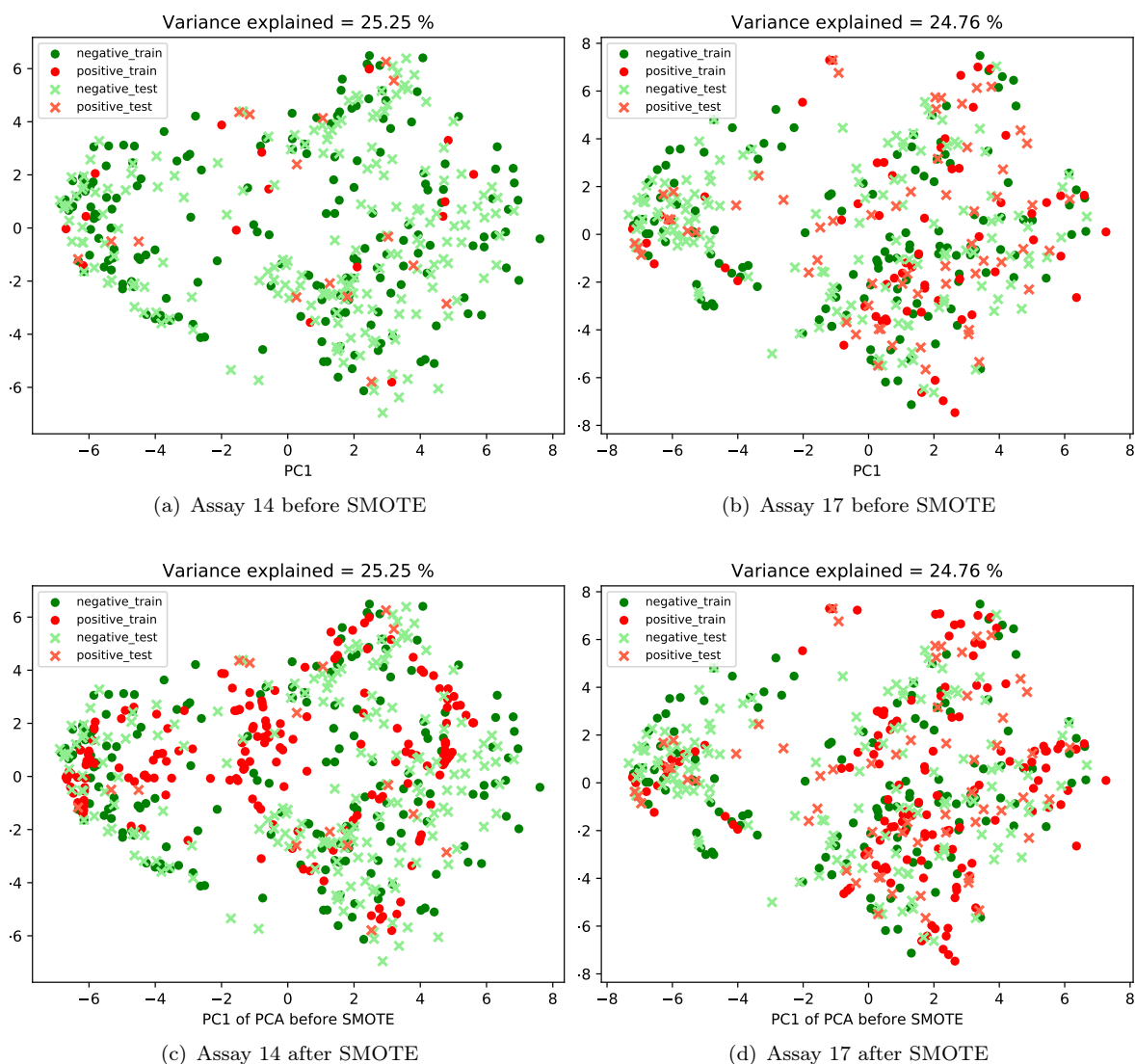


Figure 4.14: Principal Component Analysis. (a),(b): PCA of the compounds of the training sets for assays 14 and 17 before data augmentation and plots of the compounds of the test sets. (c),(d): Plots of the generated observations by SMOTE on the PCA "before" data augmentation.

discriminating the two classes of compounds could be observed in a two-dimension space. This suggests that the chemical descriptors cannot explain the classification performance variability among assays and we now wonder if this could be due to a too small number of compounds. To test this hypothesis, we use larger datasets.

4.2 Machine learning on extended datasets and comparison to the *in vivo* constrained ones

In this section we use bigger datasets in order to evaluate the impact of using more observations on the performance of the models. To get a first idea, we build new models using both the previous *in vivo* constrained datasets and bigger ones in order to compare the performance on the same 37 *in vitro* assays.

4.2.1 Data used

To obtain larger datasets, we extend the *in vivo* constrained datasets composed of 404 compounds to the entire ToxCast database which leads to a total number of 7691 compounds. As illustrated in Figure 4.15, the 37 resulting datasets are still highly imbalanced in favor of negative compounds with a range of percentage of positive compounds going from 2.5% for assay number 1 to 22% for assay number 17.

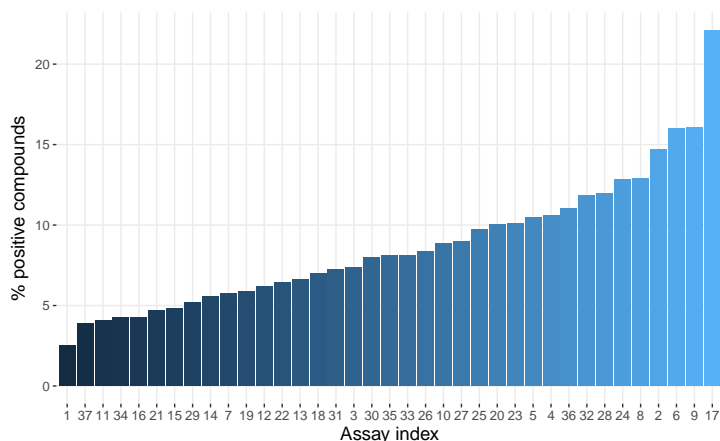


Figure 4.15: Percentage of positive compounds (Y-axis) in the 37 extended datasets corresponding to the 37 *in vitro* assays (X-axis).

4.2.2 Learning procedure and evaluation

The same descriptors than previously are computed for the 7691 compounds (74 physico-chemical properties and 4870 fingerprints) but we use a different method to select them for learning. Indeed, the descriptors selection is composed of the two following steps:

1. Removal of descriptors having a variance close to 0: in such case, the descriptor is not sufficiently informative to discriminate the compounds;
2. Removal of highly correlated descriptors: for each of the 37 *in vitro* assays, we perform a Fisher test (F-test) between each descriptor and the binary value of the assay and we select the 20% descriptors that have the lowest p-value (they correspond to the most associated ones to the considered assay).

Thereby, the finally selected descriptors are different for each dataset and depend on the correlation with the corresponding assay outputs.

This descriptor selection is applied to the two types of datasets (*in vivo* constrained and extended ones) for a valid comparison.

4.2.3 Learning procedure

Here we choose a classical and usual learning algorithm to build the new models based on the two types of datasets: we use the **Random Forest** (RF) [31] classifier from Python *Scikit-learn* toolbox [205].

We perform an internal 10-fold cross-validation and repeat the process 10 times to finally compute the average BA, sensitivity and specificity and evaluate the models' performance.

4.2.4 Results on both *in vivo* constrained and extended datasets

Imbalanced datasets

Figure 4.16 presents the results obtained on both the *in vivo* constrained datasets (404 compounds) and the extended ones (7691 compounds). We see that the performance obtained with the extended datasets are higher than the ones obtained on the *in vivo* constrained datasets. Nonetheless, for both types of datasets, the sensitivity is low ($< 50\%$) and the specificity is high ($> 90\%$), suggesting that the RF algorithm is not able to detect true positives and simply predicts almost everything as negative. As already observed using the *in vivo* constrained datasets, performance vary a lot between assays and here again, we want to evaluate the effect of balancing the datasets.

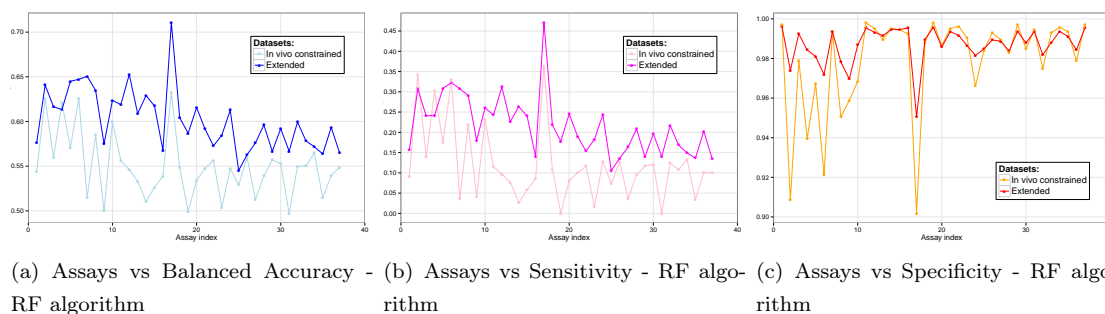


Figure 4.16: Performance results (Y-axis) versus different assays (X-axis) when the RF is used to classify both the *in vivo* constrained datasets and extended datasets.

Balanced datasets

We use the same data augmentation technique than previously (SMOTE) to generate balanced training sets for all the models. The obtained BAs, sensitivities and specificities are presented in Figure 4.17. We observe that the results are improved after data augmentation (compared to Figure 4.16) for both types of datasets. In particular, the sensitivities and BAs are respectively increased by 8% and 3% in average. Nonetheless, sensitivity stays low compared to specificity. Even after data augmentation, performance is higher for the extended datasets than for the *in vivo* constrained ones confirming that a larger number of observations helps the learning and leads to better classifiers. Since performances are still not sufficient for most of the assays when using extended and balanced datasets, it suggests that they do not only depend on the methods used but mainly on the suitability of the data to the considered problem.

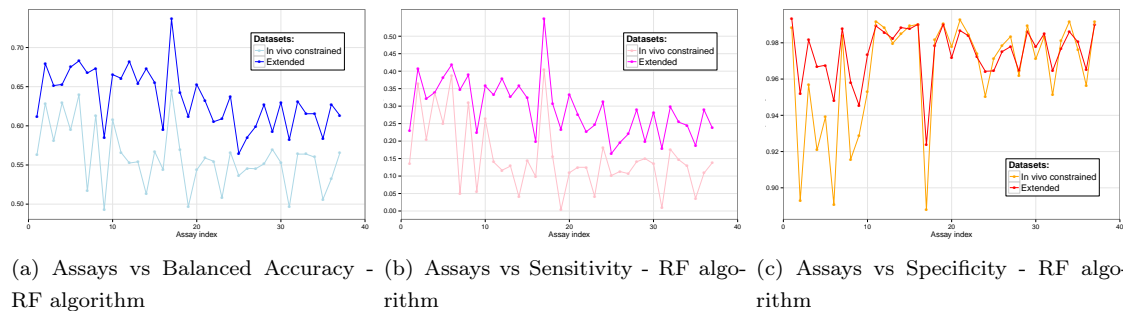


Figure 4.17: Performance results (*Y*-axis) versus different assays (*X*-axis) when the RF is used to classify both the *in vivo* constrained datasets and extended datasets with data augmentation.

4.3 Conclusion

In this chapter we built ML models to predict the results observed in *in vitro* assays based on the structure of the compounds. Since the two-stage approach requires *in vivo* data, the data used were constrained by the available data in both ToxCast and ToxRefDB and we were able to build classifiers for 37 *in vitro* assays from ToxCast.

We first used 404 compounds and 8 ML algorithms to build the 37 models and we then applied the SMOTE data augmentation technique to balance the datasets which were originally highly imbalanced in favor of negative compounds. We showed that for some assays, ML models were able to correctly predict *in vitro* activity, in particular after balancing the datasets. Nonetheless, we highlighted that classification performance depend on the assays since assays having the most balanced datasets were easier to predict than the ones with few instances of positive compounds. We then used extended datasets composed of 7691 compounds to evaluate the importance of the datasets size on the prediction of the same 37 *in vitro* assays. The results showed that the performance increases with the size of the datasets but are still low for most of the assays, in particular the sensitivity, due to the imbalanced property of data. In that case again, data augmentation improved performance.

Nevertheless, this study is not sufficient to demonstrate that good models can be obtained to constitute the first stage of the two-stage approach, since performances are insufficient for most of the assays. Hence, a larger analysis on more assays would be interesting in order to explore if we could build more relevant models for some specific assays and which types of methods are suitable for the data available.

CHAPTER 5

FROM STRUCTURE TO ACTIVITY: A LARGE SCALE ANALYSIS

This chapter focuses on a large scale analysis using all the available data in ToxCast in order to evaluate how to build good ML models with this type of data, relaxing the constraint due to the two-stage approach. We therefore build ML classifiers for each of the ToxCast *in vitro* assay, first using simple algorithms and then using an ensemble method. Also, the applicability domain is used in order to assess the relevance of the predictions. Part of this chapter has been published in the international *Journal of Information and Chemical Modelling* [106].

5.1 Datasets building

5.1.1 Data used

In this work we use the entire ToxCast database released in October 2015. Precisely, we use the "hitc matrix" which reports the values of 9076 compounds tested in 1192 *in vitro* assays. Among all the compounds, 8599 correspond to unique structures according to their SMILES identifiers [273] available in the DSSTox database.

As described in Chapter 3, the hitc matrix results are provided as categorical values:

- 0 means inactive,
- 1 means active,
- -1 means undetermined,
- NA means Non Assigned.

Since our goal here is to build binary classifiers that predict if a compound is active or inactive in the assays, compounds with "undetermined" or "Non Assigned" values are not used. Hence, assays that contain only -1 and/or NA values are removed from the matrix, resulting in a total of 1092 assays.

For each of these 1092 assays, we build a dataset composed of the number of compounds tested

in the assay (*i.e.* compounds with either 1 or 0 values), which implies that the number of compounds in the datasets varies according to the assays. For all the compounds, we generate chemical descriptors and Figure 5.1 summarizes the workflow that we use to compute and select these descriptors.

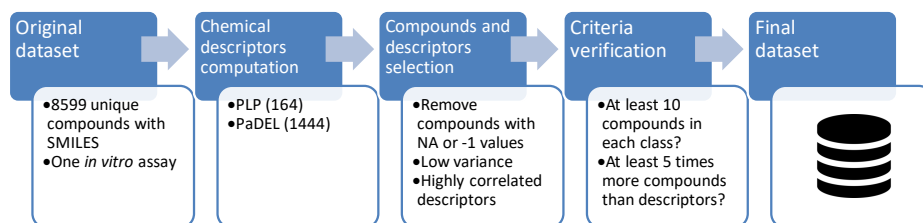


Figure 5.1: *Workflow of the data processing for one in vitro assay.* For each in vitro assay, the original dataset contains the list of tested compounds with their structure from which chemical descriptors are computed. The compounds for which the reported value for the assay is NA or -1 are removed and the molecular descriptors having a low variance or being highly correlated with other descriptors are then removed (see methods in Section 5.1.2). Only datasets containing at least 10 compounds of the two classes (positive and negative) and at least 5 times more compounds than descriptors are kept. They correspond to the final datasets.

5.1.2 Generation and selection of chemical descriptors

i) Descriptors computing

Two software are used to compute several 1D and 2D chemical descriptors using the SMILES representation of the compounds: PaDEL-Descriptor [285], an open source software, and Pipeline Pilot (PLP) [59], developed by Dassault Systèmes BIOVIA. Respectively 1444 and 164 descriptors are computed with PaDEL and PLP and are mostly continuous values except for constitutional descriptors which refer to the number of various components in the compounds. Each type of descriptors will be used independently for machine learning, meaning that there are two datasets for each *in vitro* assay: one PLP dataset and one PaDEL dataset. Regarding PaDEL, there are compounds for which the software is not able to compute all the descriptors, probably due to errors in the SMILES representation. Indeed, as described in Chapter 3, some molecular descriptors cannot be computed with improper structures. Even if we performed a data curation, we could have omitted some errors during the final manual check due to the important number of compounds (> 8000). To avoid the use of incorrect data, we choose to remove from the datasets the compounds for which at least one descriptor cannot be computed. After this step, **2184 datasets** are available: 1092 with 164 PLP descriptors and 1092 with 1444 PaDEL descriptors.

ii) Selection of descriptors

In order to reduce the number of descriptors in the datasets, we proceed to a selection depending on the nature of their values (categorical or continuous).

First, for the categorical descriptors, we remove the ones that have a low variance because they do not enough discriminate the compounds. To do so, we use the *nearZeroVar* function from

the R *caret* package [158] which looks at two characteristics to remove a descriptor: (1) the ratio of the frequency of the most common value to the frequency of the second most common value and (2) the number of unique values out of the total number of compounds in the datasets. If (1) is greater than 4 and (2) is lower than 0.3, the descriptor is removed.

Then, to limit redundancy of descriptors with continuous values, we discard one of two highly correlated descriptors using the *findCorrelation* function from the R *caret* package. Basically, this function computes the absolute value of pair-wise correlations and compares it to a defined threshold (here we choose 0.8). If two descriptors have a correlation greater than this threshold, the function looks at the average absolute correlations of the two descriptors with all the other descriptors and removes the one with the highest average.

The set of descriptors finally kept differs among the datasets since it depends on the compounds that have been tested in the corresponding *in vitro* assays.

5.1.3 Datasets filtering

Finally, in order to keep datasets with enough representative of each class, we remove the ones with less than 10 members of active or inactive compounds. We also remove the datasets for which the number of compounds is lower than 5 times the number of descriptors. We finally end up with 515 datasets with PLP descriptors and 414 with PaDEL ones.

5.1.4 Datasets are highly imbalanced

Figures 5.2-(a) and 5.2-(c) show the total number of compounds in the 515 PLP and 414 PaDEL datasets, respectively. For PLP, this number ranges between 115 and 7810 with average and median values of 3054 and 3362, respectively. For PaDEL, the size ranges from 1391 to 7516 with average and median values of 3546 and 3259, respectively.

More interestingly, Figures 5.2-(b) and 5.2-(d) represent the percentage of active compounds in the datasets. Regarding PLP, the average percentage is about 12% with the highest percentage being 83%. However, 58% of the datasets (300/515) contain less than 10% of actives. For PaDEL, there is no dataset with more than 50% of active compounds and 66% of the datasets (275/414) have less than 10% active compounds. The average percentage is about 9%.

Overall, these numbers highlight that, based on a binary classification of the assays results, the datasets are **highly imbalanced** in favor of inactive compounds.

5.2 Simple classifiers

For each of the PLP and PaDEL datasets, we first build simple classifiers using the following algorithms and learning procedure. Unlike the previous chapter, here we use only five basic algorithms since it is a large scale study with an important number of datasets which prevents us from reviewing various algorithms.

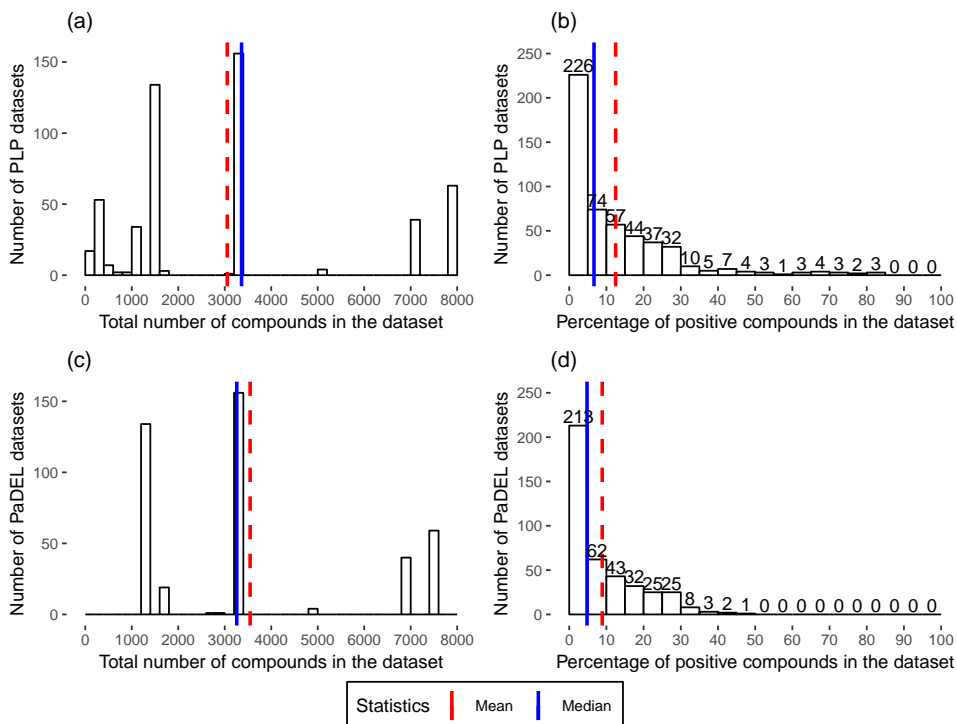


Figure 5.2: Distribution of datasets according to the total number of compounds or the percentage of positive compounds. (a), (b) Distribution of the 515 PLP datasets according to the total number of compounds in the datasets and the percentage of active compounds in the datasets. (c), (d) Distribution of the 414 PaDEL datasets according to total the number of compounds in the datasets and the percentage of active compounds in the datasets.

5.2.1 Machine learning algorithms

We choose five commonly used algorithms that cover the different types of existing classification methods (see Section 2.3.2 for details):

- A Linear Discriminant Analysis (LDA): we use the implementation from the R package *MASS* [266];
- An Artificial Neural Network (ANN): we use the implementation from the R package *NNET* [211], the network is composed of a single hidden layer;
- A Support Vector Machine (SVM): we use the implementation from the R package *e1071* [188];
- A Naïve Bayes (Bayesian): implemented in Pipeline Pilot ;
- A Random Forest (RF): implemented in Pipeline Pilot.

5.2.2 Learning procedure and validation

The learning procedure applied to each dataset and for each algorithm is described in Figure 5.3. First, we split each dataset into a training set (80%) and a test set that will constitute the external test set (20%).

We then use the training set to learn the models and perform a stratified 5-fold internal cross-validation to estimate internal performance. To guarantee that each fold contains at least one member of each class of molecules, active and inactive compounds are first separated and ran-

domly split into 5 sub-folds. Then, the five folds of the cross-validation are built by merging one sub-fold from each class. The cross-validation is repeated 5 times and the average performance is computed.

Then we perform external validation using the model built on the entire training set to predict the values of the test set previously set aside and we compute external performance.

The entire process (from the splitting of the dataset to the external validation) is repeated 5 times and the average of the performance metrics is computed.

The following four metrics are used to evaluate the performance: ROC score, balanced accuracy, sensitivity and specificity.

As already mentioned in Chapter 2, in the case of two-class predictions, the predicted continuous number that is returned by the algorithms is automatically transformed into a binary one according to a threshold. There are several ways to determine this threshold and it has been shown that the traditional default method (threshold = 0.5) was unreliable for most of the datasets [94]. One of the best approach is to maximize the percentage of correctly classified observations (*i.e.* the accuracy or BA). We chose an equivalent approach based on the opposite, that is to say the minimization of the balanced error rate which corresponds to the average of the errors on each class, and equals to $(1 - BA)$, or:

$$\frac{1}{2} \left(\frac{FN}{TP + FN} + \frac{FP}{TN + FP} \right)$$

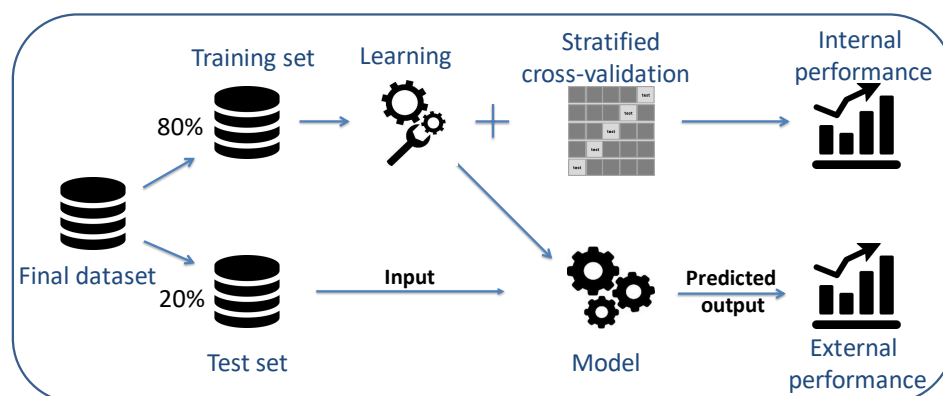


Figure 5.3: Workflow of the learning procedure for one *in vitro* assay. The dataset obtained in 5.1 is split into training set (80%) and test set (20%). The training set is used to learn the model using one of 5 different algorithms and a stratified 5-fold cross-validation is performed to get the internal performance of the model. The test set is then used to compute external performance of the model.

5.2.3 Results

Internal cross-validation results on all datasets

For the five ML algorithms, we compute the mean and standard deviation of each performance metric obtained after internal cross-validation over the 515 PLP models (Figure 5.4-(a)) or the 414 PaDEL models (Figure 5.4-(b)). For the two types of descriptors and the five algorithms, the

four metric means are comprised between 0.6 and 0.73 (except ANN-PaDEL sensitivity which is equal to 0.52) and standard deviations are large, ranging from 0.06 to 0.19. According to these results, none of the algorithms is able to reach high performance and we are not able to rank them because of too large and overlapping standard deviations.

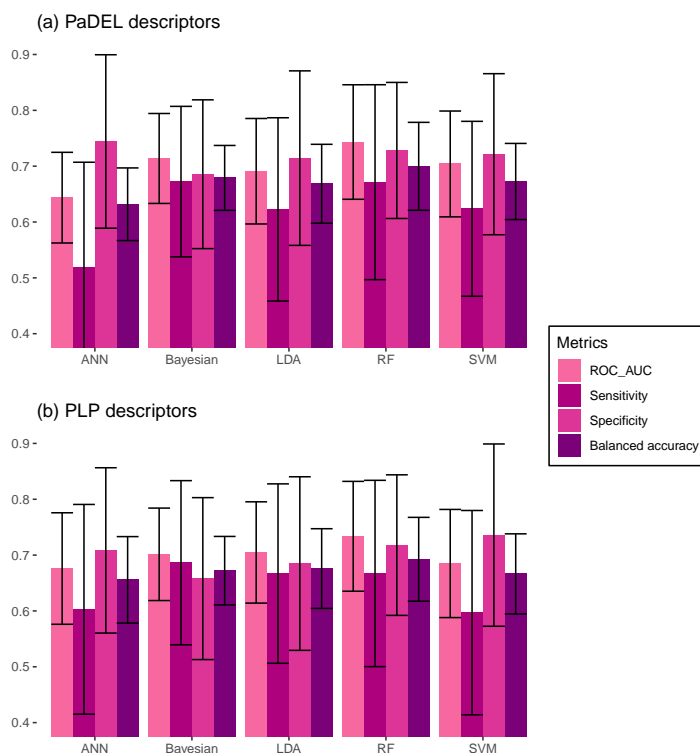


Figure 5.4: Comparison of performance metrics (ROC score, sensitivity, specificity and BA) after internal cross-validation for the 5 ML algorithms. (b) Models using PaDEL descriptors and all 414 datasets. (a) Models using PLP descriptors and all 515 datasets. None of the algorithms is able to reach high performance.

As we already showed that the data are imbalanced, we now consider if this characteristic can explain the previous results. For all the 5 algorithms, Figure 5.5 presents the plots of the balanced accuracy (BA) obtained for each dataset according to the percentage of active compounds. These plots display a “funnel” shape with BA variability decreasing when datasets are more balanced. In particular, we observe lower BA variability for datasets containing at least 10% of compounds in the minority class (*i.e.* positive compounds). For datasets with low percentage of positives, most of the BA variability is due to the variability of the sensitivity, which depends on the number of positive compounds in the datasets: for a same ratio of positive compounds, the larger the number of positives, the higher the sensitivity. Note that this funnel shape is also observed for the plots of the 4 other metrics and that we obtain similar results with PaDEL datasets except that the percentage of positive compounds stops at 50% (data not shown).

Figure 5.6 presents the variance of ROC score obtained with RF algorithm trained on PLP datasets according to the percentage of positive compounds in the datasets. For each range of percentages, we compute the variance of ROC score over all the datasets having a percentage of

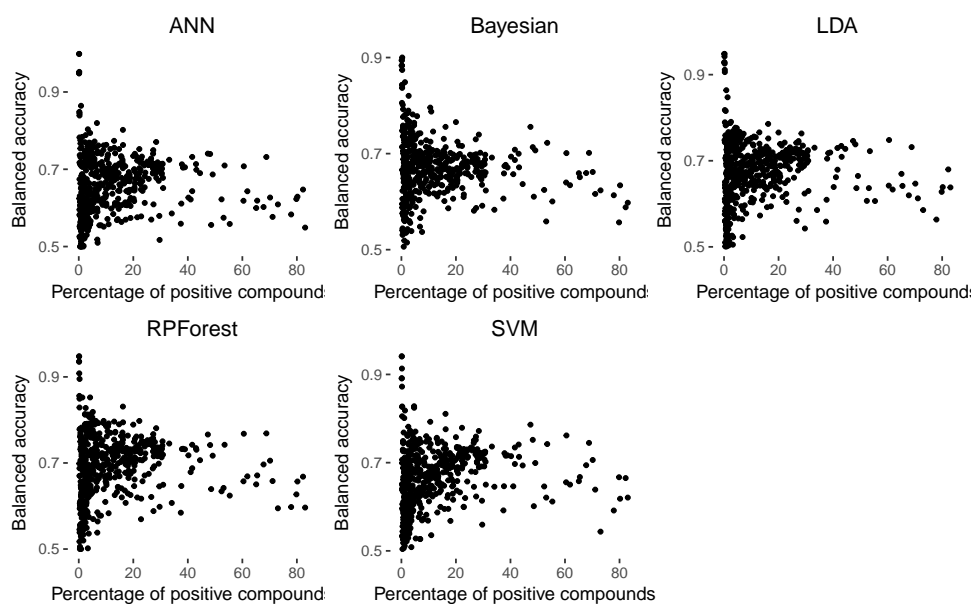


Figure 5.5: *Balanced accuracy according to the percentage of positive compounds in PLP datasets for the 5 ML algorithms. BA gets stable when the percentage of positive compounds in datasets increases. For datasets with a low percentage of positives, most of the BA variability is due to the variability of the sensitivity, which depends on the number of positive compounds in the datasets: the larger the number of positives, the higher the sensitivity.*

positive compounds in that range. The figure shows that the variance of the ROC score tends to decrease when datasets are more balanced. We can determine a cut-off: when datasets contain at least 10% of positive compounds, the associated variance is always below 0.0065. Similar results are obtained with PaDEL datasets and other algorithms (data not shown).

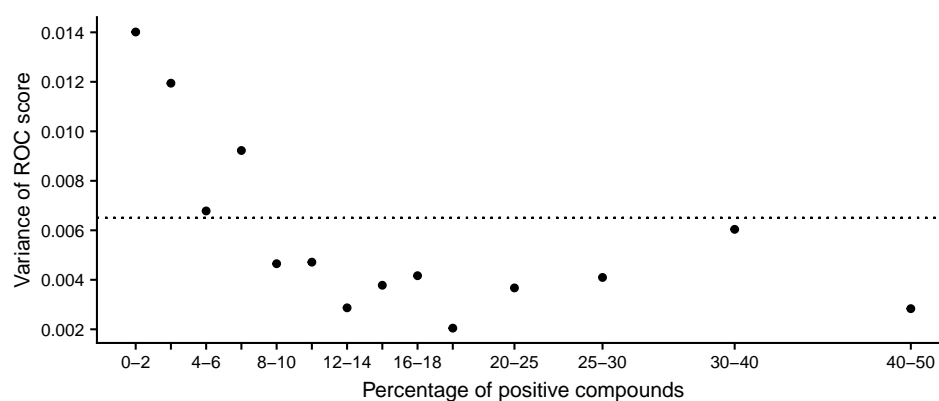


Figure 5.6: *Variance of ROC score according to the percentage of positive compounds after internal cross-validation for Random Forest models based on PLP descriptors. Variance is lower than 0.0065 when percentage of positive compounds is greater than 10%.*

Together, these results suggest that the imbalanced nature of datasets has a negative impact on models' performances and that these classical learning methods are not suitable for highly imbalanced datasets. This finding is in agreement with previous work from different domains which highlighted the need for techniques to face imbalanced issues, for example for detection of oil spills [156] or fraudulent telephone calls [86]. Indeed, the use of common algorithms does not

cope with the problems related to the imbalanced property of data [43].

Moreover, the models with very few positive compounds will be characterized by a limited applicability domain regarding this class of compounds. Consequently, a new positive compound will have a low chance to be within the applicability domain and its associated prediction will be of low confidence.

Based on these considerations, we decide to try if we could obtain better performance when considering datasets which contain at least 10% of compounds belonging to the minority class. This leads us to focus on 139 PaDEL and 215 PLP remaining datasets.

Internal cross-validation results on a subset of more balanced datasets

Figure 5.7 shows the means of each performance metric over the 139 and 215 models obtained on the PaDEL and PLP datasets, respectively. The means of all metrics are between 0.63 and 0.75 and here again the results are in the same range for the 5 ML algorithms (BA around 0.68). Interestingly, standard deviations are lower than previously (when using all the datasets): BA standard deviation decreases in average from 0.07 to 0.04, sensitivity ones from 0.16 to 0.10 and specificity ones from 0.14 to 0.09. However, despite this improvement, it is still difficult to rank the algorithms and conclude on the most appropriated to the data, no matter the type of descriptors used.

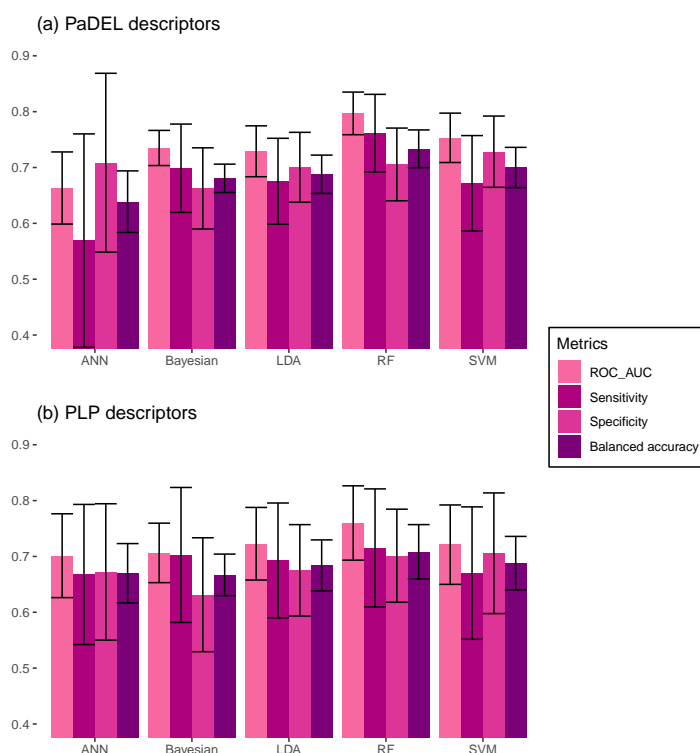


Figure 5.7: Comparison of performance metrics (ROC score, sensitivity, specificity and BA) after internal cross-validation for the 5 ML algorithms. (a) Models using PaDEL descriptors and the 139 datasets with at least 10% of compounds in the minority class. (b) Models using PLP descriptors and the 215 datasets with at least 10% of compounds in the minority class. Performance of the 5 algorithms are on the same range and standard deviations are smaller compared to the previous results based on all the datasets.

In order to quantify the advantage of using more balanced datasets, we perform a Student’s t-test: Table 5.1 shows the p-values of the t-test that compare the mean of the 4 metrics between the datasets containing strictly less than 10% of compounds of the minority class and the ones containing at least 10% of compounds, for the 5 algorithms and the 2 types of descriptors used. Note that here we assume that datasets are independent, so that there means are independent too and therefore comparable using the t-test. According to this test, most of the metrics means are significantly different, meaning that the use of more balanced datasets has an impact on the models’ performances. In particular, the sensitivity is always significantly increased suggesting that the use of more balanced datasets helps in the detection of true positives. We also remark that for ANN_PADEL, Bayesian_PADEL and Bayesian_PLP, if sensitivity and specificity are significantly different between the two types of datasets considered for the test, this difference is not sufficient to obtain a significant difference for the BA.

Overall, the same tendencies are observed with both types of descriptors but PLP datasets contain fewer descriptors which confers several advantages. In particular, they are generally easier to understand as they are related to well-known physico-chemical properties such as molecular weight or solubility. Also, other advantages of building ML models with fewer descriptors are to decrease model complexity, to reduce chances of overfitting, and to decrease computational time [104, 111], as already mentioned in Chapter 2. Therefore, we focus only on PLP datasets for the rest of the study.

Table 5.1: p-values of Student’s t-test performed on the 4 metrics for the 5 algorithms and 2 types of descriptors, between the datasets that contain strictly less than 10% of active compounds and the datasets that contain at least 10% of active compounds. The p-values lower than 0.05 are in bold. Most of the metrics means are significantly different meaning that the use of more balanced datasets has an impact on the models’ performances.

Method	Descriptor type	ROC AUC	Balanced Accuracy	Sensitivity	Specificity
ANN	PADEL	1.10×10^{-4}	9.06×10^{-2}	1.58×10^{-4}	1.15×10^{-3}
Bayesian	PADEL	6.32×10^{-7}	5.96×10^{-1}	5.95×10^{-4}	2.00×10^{-3}
LDA	PADEL	1.29×10^{-13}	4.85×10^{-7}	2.98×10^{-9}	9.03×10^{-2}
RF	PADEL	2.40×10^{-23}	1.04×10^{-15}	1.70×10^{-22}	8.48×10^{-4}
SVM	PADEL	8.76×10^{-22}	3.55×10^{-13}	3.85×10^{-8}	3.73×10^{-1}
ANN	PLP	1.66×10^{-7}	1.20×10^{-4}	3.30×10^{-13}	7.59×10^{-7}
Bayesian	PLP	2.11×10^{-1}	7.82×10^{-2}	2.55×10^{-2}	1.47×10^{-4}
LDA	PLP	3.79×10^{-5}	1.69×10^{-2}	6.80×10^{-4}	1.79×10^{-1}
RF	PLP	2.15×10^{-8}	9.39×10^{-6}	8.98×10^{-10}	5.69×10^{-3}
SVM	PLP	6.71×10^{-15}	2.31×10^{-10}	1.46×10^{-17}	1.09×10^{-4}

External validation results

We then perform an external validation on the 215 PLP datasets and we choose to present in Figure 5.8 the average BA, sensitivity and specificity obtained for the 5 ML algorithms.

Except for the Bayesian algorithm, sensitivity is very low (under 0.4) and specificity is high (greater than 0.8) which leads to a BA around 0.6. Moreover, sensitivity and specificity standard deviations for ANN, LDA, RF and SVM models are large (between 0.09 and 0.27), which once

again prevents us from drawing conclusions on the usefulness of these methods. According to these results, ANN, LDA, RF and SVM do not seem to be able to build good models, even when only focusing on datasets with at least 10% of compounds in the minority class. We hypothesize that they are too sensitive to imbalanced datasets as Mazurowski *et al.* [184] already showed for neural networks. Indeed, they performed a large scale analysis on simulated imbalanced data and studied the impact of this characteristic on two neural network methods. He concluded that even a small imbalance in the training set led to a deterioration of the performance.

On the contrary, the Bayesian algorithm seems to be more suited to imbalanced datasets since BA, sensitivity and specificity are greater than 0.6 with smaller standard deviations. However, in order to ensure confidence in the models, we consider that models are not good enough when their performance are lower than 0.7 for all metrics. Thus, the different types of methods used here seem to not be suited to the data and we suggest that using more than one simple algorithm could improve the models' performances. In particular, we propose to use ensemble techniques.

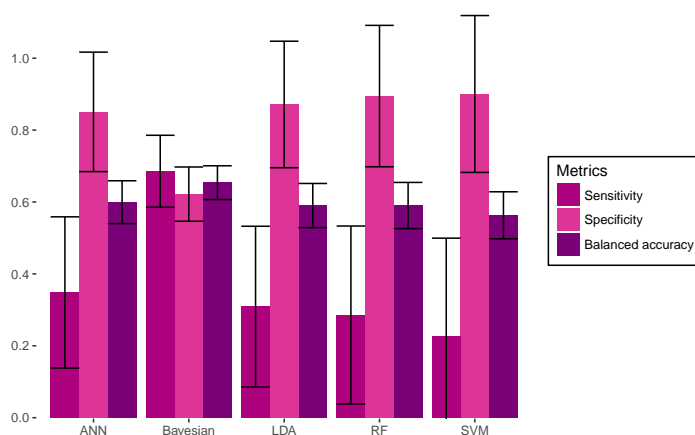


Figure 5.8: Comparison of performance metrics (sensitivity, specificity and balanced accuracy) after external validation for the 5 algorithms on PLP datasets with at least 10% of compounds in the minority class (215 datasets). ANN, LDA, RF and SVM are not able to lead to good performance and do not seem suitable for the data while Bayesian seems to be more appropriate.

Descriptor importance analysis

When interpreting a QSAR model, we often want to understand how each descriptor contributes to the final decision. In order to assess descriptor relevance, different methods exist including the "permutation importance" [247] which, for each descriptor in turn, compares the performances between a model built on the original dataset and a model built on a dataset where the considered descriptor has been randomly permuted. The permutation of a descriptor refers to the random reassignment of all its values to the observations of the dataset, such that there is no more relation between the observations and the descriptor. In the end, if the permuted descriptor has a strong effect on the output prediction, the model trained on the permuted data will result in lower performance than the original one.

Here we use the permutation importance approach in order to quantify the relative importance

of each descriptor in the datasets. For each model built with a randomly permuted descriptor, we compute the absolute difference between its ROC score and the ROC score of the original model (non permuted). Then, we are able to rank the descriptors according to this difference: the larger the difference, the greater the presumed importance of the descriptor.

We perform the permutation importance on the 10 RF models (based on PLP datasets) that results in the best performance, *i.e.* the ones that reached BA, sensitivity and specificity greater than 0.7 after external validation. We finally compute the average rank of each descriptor over this 10 models.

Table 5.2 shows the top three descriptors that obtained the highest rank:

1. **The Ghose and Crippen octanol-water partition coefficient AlogP [102]**: it is the logarithmic ratio of the concentration of a compound in a two-phase solvent, here octanol and water. It indicates hydrophobicity and hydrophilicity of a compound.
2. **The molecular solubility [257]**: it informs about the ability of a compound to enter the systemic circulation and directly affects its bioavailability [230].
3. **The molecular weight**: like AlogP, it is one of the 4 components taken into account by the Lipinski’s rule of five [168] which considers 4 molecular criteria important for chemical discovery.

Descriptor	Average rank
AlogP	1.27
Molecular Solubility	2.27
Molecular Weight	5.54

Table 5.2: Top three descriptors obtained by computing the average rank of each descriptor after applying the "permutation importance" method on the ten datasets corresponding to the ten best models. AlogP, molecular solubility and molecular weight are the three most important descriptors.

Our analysis also highlights other important descriptors such as the sum of carbons with two single bonds or the sum of carbons with two single bonds and one double bond. This type of descriptors might be helpful if we would like to go further and investigate the modes of action that explain the activity of the molecules.

In this partial analysis, the most important descriptors are found to be the ones that are the easiest to interpret biologically and chemically. Moreover, when looking at the 10 assays corresponding to the 10 best models used¹, we see that most of them measure receptor binding which suggests that these three descriptors are important for this specific type of assays. These results therefore need to be extended to more assays by performing a large scale analysis on more datasets and using other algorithms.

¹List of the 10 assays:

ATG_PXR_TRANS_up; ATG_SREBP_CIS_up; TOX21_AR_BLA_Agonist_ratio; TOX21_GR_BLA_Agonist_ch1; TOX21_GR_BLA_Agonist_ratio; TOX21_GR_BLA_Antagonist_ch2; TOX21_GR_BLA_Antagonist_viability; TOX21_MMP_ratio_down; TOX21_MMP_viability; TOX21_p53_BLA_p1_viability.

5.3 An ensemble method: the stacked generalization

As the previous results showed that one algorithm alone is not able to lead to good models on our data, we now test if combining several models in an ensemble method (see Chapter 2) could improve the results. Since we want to combine different types of methods, we decide to use the Stacked Generalization technique [278, 113] which, unlike bagging and boosting, allows the use of heterogeneous models. Furthermore, this method has been recently shown to be more appropriate to imbalanced data in a study performed on 5 imbalanced datasets from different domain [176]. They showed that simple methods such as Bayesian, Decision trees and Logistic Regression are not sufficient to reach high performance and that each of these methods has its own pros and cons. They therefore proposed a theoretical analysis of the effectiveness of ensemble methods to learn from imbalanced data and concluded in a favor for the boosting and stacking techniques. Regarding the methods to combine, we first choose to keep the Bayesian method because it seems to be the most suitable for our data according to the previous results. Moreover, we choose the RF method because it is based on a really different representation of the data than the Bayesian one (trees versus instances).

5.3.1 General principle

The **Stacked generalization** ensemble technique (also called stacking) involves the training of a learning algorithm called "meta learner" that combines the predictions of several "base learners" in order to obtain a higher level learner. Basically, the training set is split into two disjoint sets where the first part is used to train the base learners and the second part to test them and generate predictions. These predictions are finally used as the inputs of the meta learner with the corresponding outputs being the correct responses.

5.3.2 Learning procedure and validation

The procedure used to learn the stacked models is illustrated in Figure 5.9 and detailed hereafter. First, we randomly split each dataset into one training set (Train 1) and two test sets (Test 1 and Test 2) in the following proportions: 60%, 24% and 16% (see ovals in Figure 5.9). On Train 1, we train both a Bayesian and an RF algorithms to build the base models B1 and RF1 and we test them on Test 1 to obtain predictions P1-b and P1-rf (see the blue workflow in Figure 5.9). We then use the predictions P1-b and P1-rf, respectively produced by models B1 and RF1, as input descriptors of the meta-learner to train a so-called stacked model using a naive Bayesian algorithm. The outputs to learn are the actual outputs of Test 1 (orange workflow of Figure 5.9). We choose the Bayesian learner as meta learner due to its ease of implementation and fast computing.

Then, a stratified 5-fold cross-validation is performed to find the best threshold for the stacked model: the training set containing P1-b and P1-rf as input features and actual output values of Test 1 is split into 5 folds which are all used once as a test set while the 4 remaining folds are

used to learn the model. In some cases, the number of positive compounds in Test 1 is lower than 5 and the 5-folds are therefore impossible to compute (since all the folds should have the same ratio of positives to negatives). In those cases, the entire dataset is removed from the study. In the other cases, the 5-fold cross-validation is repeated three times. Finally, we evaluate external predictive performance of the stacked model on Test 2 (orange workflow of Figure 5.9).

In order to compare the performance of the stacked models with the simple classifiers on the same test set, we merge Train 1 and Test 1 to train simple Bayesian and Random Forest learners and build models B2 and RF2 (see red workflow in Figure 5.9). We then compute their external predictive performances on Test 2.

Note that we also built stacked models using 3 to 5 base models (by iteratively adding ANN, SVM and LDA to RF and Bayesian) on a subset of datasets (data not shown). Since the performance did not change significantly according to the number of models used, we decided to keep only two base models for a matter of computing time.

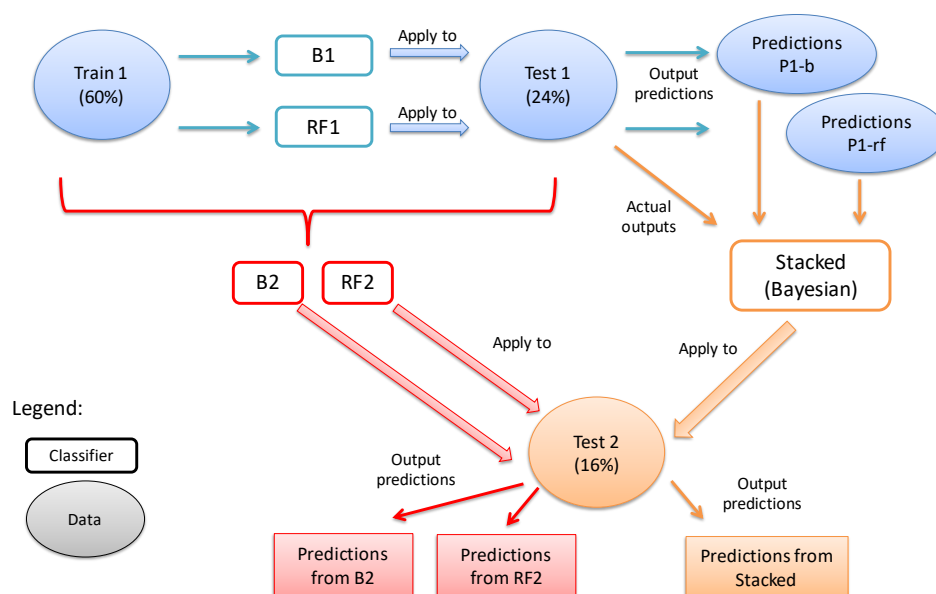


Figure 5.9: Principle of the Stacked generalization method.

Blue Workflow: A Bayesian B1 and a RF RF1 models are built using Train 1 (60% of the dataset). The models are then used to make predictions P1-b and P1-rf on Test 1.

Orange workflow: the predictions P1-b and P1-rf are used to train a Stacked Bayesian model (Stacked).

Red workflow: By merging Train 1 and Test 1 to train a simple Bayesian model B2 and simple RF model RF2, we are able to compare performances of B2, RF2 and the Stacked model on Test 2.

5.3.3 Results

Comparison to simple classifiers

The stacked generalization is applied to all the PLP datasets, including those with less than 10% of positive compounds. Over the 515 datasets, 32 are removed because of the failure to compute the 5-fold cross-validation as described above. We therefore obtain 483 Stacked models with the

483 corresponding B2 and RF2 models.

We compare the predictive performances of the 483 simple Bayesian and RF models (B2 and RF2) with the Stacked ones. Table 5.3 shows, for the three types of methods, the number of models that reach a certain value of ROC score. First, less RF2 models are able to reach ROC score greater than 0.6 compared to the B2 and Stacked ones and only 30 among the 483 have ROC score above 0.8. Furthermore, if an equivalent number of B2 and Stacked models reach 0.6 and 0.7 values of ROC score, when looking at higher and better performance (above 0.75 and 0.8), the Stacked method becomes clearly better than the Bayesian one. Finally, when we look at the method that leads to the highest ROC score for each model, the Stacked is the winner for 61% of models (294/483), followed by the Bayesian one with 30% of models (147/483) and the RF with only 9% (42/483) (data not shown). These results confirm that the Stacked generalization method is able to build more models with good performances than simple algorithms.

Table 5.3: Comparison of simple Bayesian B2 and Random Forest RF2 models with Stacked generalization models on the 483 PLP datasets. The comparison of the number of models that reach a certain value of ROC score shows that the Stacked generalization method is able to build more models with higher ROC score than simple methods.

Method	ROC \geq 0.60	ROC \geq 0.70	ROC \geq 0.75	ROC \geq 0.80
Stacked	417	319	253	144
Bayesian	416	321	223	90
RF	356	205	89	30

Figure 5.10 shows the ROC curves of B2 and stacked models for one particular assay (TOX21_ ERa_BLA_Antagonist_ratio) measuring the expression of the Estrogen Receptor gene². The associated dataset is composed of 7810 molecules, 13% of which are active in the *in vitro* assay. We observe that the ROC curve of the Stacked model is always above the one of the B2 model and the ROC scores are equal to 0.84 and 0.77 for the two models, respectively. Since the ROC curve displays the sensitivities and their corresponding specificities obtained for all threshold values between 0 and 1, we can choose a specific threshold according to a desired sensitivity or specificity. As an example, a sensitivity of 85% corresponds to a specificity of about 73% with the Stacked model and only of 52% with the Bayesian one (see dotted lines in Figure 5.10). This again illustrates the difference of performances between the two models and the ability of the Stacked model to detect more inactive compounds than the Bayesian one, for the same number of actives detected. Naturally, one can move this threshold depending on the desired stringency of the model output. The same analysis on other assays lead to the same observations and conclusions (data not shown).

²see <https://actor.epa.gov/dashboard>

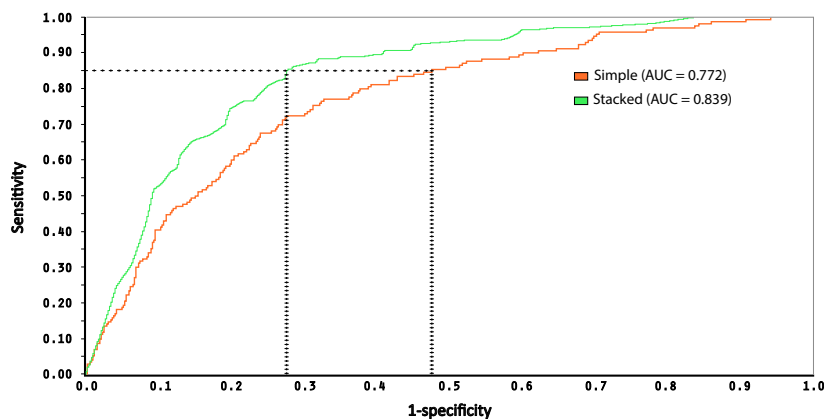


Figure 5.10: ROC curves obtained for models trained on the dataset of the assay *TOX21_ERa_BLA_Antagonist_ratio* with the two types of methods (Stacked generalization and simple Bayesian). ROC curve of the Stacked model is always above the one of simple Bayesian model. For a given Sensitivity of 85%, the Bayesian model detects 52% of the inactive molecules whereas the Stacked model detects 73% of them.

Focus on a subset of *in vitro* assays highlighted as being correlated to *in vivo* toxicity

Since the bioactivity assays can be seen as an intermediate step towards the evaluation of the *in vivo* toxicity, several works have relevantly focused on the link between ToxCast *in vitro* assays and toxicity outcomes observed *in vivo*. In particular, as already mentioned in Chapter 3, in 2015 Liu and co-workers [169] built ML models that predict *in vivo* chronic toxicity observed in liver based on either chemical descriptors, bioactivity descriptors (*i.e.* ToxCast *in vitro* assays) or a combination of both. This study has been extended in 2017 to 19 other organs [170]. They extracted in both studies the 36 (resp. 50) *in vitro* assays most frequently used in their models and which were supposed to be the most correlated with *in vivo* liver (resp. 19 other organs) toxicity.

Since we built models for the majority of the ToxCast *in vitro* assays, we propose here to focus on the ones that predict these assays. More precisely, among the 36 (resp. 50) *in vitro* assays highlighted by Liu, we were able to build classifiers for 25 (resp. 38) of them. Because 11 assays are common to both sets (25 and 38), we finally have models for 52 assays, using the simple algorithms and the stacked method. Table 5.4 summarizes the best ROC score we obtained for the models of these 52 assays and the corresponding method used (simple Bayesian, RF or Stacked generalization). For 71% of the assays (37/52), the Stacked generalization is the method leading to the best ROC score. Also, for 62% of the assays (32/52) the ROC score is greater than 0.75 meaning that we are able to build good classifiers to predict some of the *in vitro* assays featured by Liu. Altogether, these results show that the Stacked generalization method allows one to build classifier models that predict *in vitro* assays which have been previously shown to be associated to *in vivo* toxicity outcomes. This suggest that we could think about replacing all or part of these *in vitro* assays by *in silico* predictions and use these predictions as input of Liu’s ML models, as we already suggest in our two-stage approach.

Table 5.4: Most frequently used assays in Liu’s models and the best ROC scores we obtained with either the Stacked generalization or simple methods (Bayesian or RF). Assays are sorted by decreasing ROC score. Stacked generalization method has the best ROC score for 71% of the assays and this score is greater than 0.75 for 62%.

Assay name	Liver	Others	ROC score	Bayesian(B)/ Stacked(S)/ RF (RP)
BSK_KF3CT_SRB_down		×	0,855	S
TOX21_TR_LUC_GH3_Antagonist	×		0,842	S
APR_HepG2_CellLoss_24h_dn		×	0,842	B
TOX21_ERa_BLA_Antagonist_ratio	×		0,839	S
ATG_VDRE_CIS_up	×	×	0,836	S
ATG_SREBP_CIS_up		×	0,834	S
ATG_PBREM_CIS_up		×	0,834	S
ATG_MRE_CIS_up		×	0,834	S
ATG_PXR_TRANS_up		×	0,833	S
APR_HepG2_MitoticArrest_72h_up	×	×	0,830	S
TOX21_ERa_LUC_BG1_Antagonist		×	0,828	S
TOX21_PPARd_BLA_agonist_ratio		×	0,827	S
TOX21_Aromatase_Inhibition	×		0,821	S
ATG_TGFb_CIS_up		×	0,819	S
ATG_RARa_TRANS_up	×		0,815	S
APR_HepG2_StressKinase_1h_up		×	0,815	B
ATG_RORE_CIS_up		×	0,812	B
BSK_3C_Vis_down		×	0,811	S
ATG_PXRE_CIS_up	×	×	0,810	S
BSK_BE3C_SRB_down		×	0,809	S
ATG_NRF2_ARE_CIS_up	×		0,793	S
ATG_PPRE_CIS_up	×	×	0,791	S
NVS_GPCR_hOpiate_mu	×		0,791	S
ATG_RXRb_TRANS_up	×		0,787	S
ATG_C_EBP_CIS_up		×	0,785	S
ATG_LXRb_TRANS_up		×	0,780	B
ATG_BRE_CIS_up	×	×	0,780	S
TOX21_PPARg_BLA_antagonist_ratio		×	0,775	S
ATG_Oct_MLP_CIS_up	×		0,771	S
APR_HepG2_MicrotubuleCSK_72h_up		×	0,760	B
APR_HepG2_MitoMembPot_1h_dn		×	0,759	B
ATG_NFI_CIS_up		×	0,754	B
ATG_NF_kB_CIS_up		×	0,749	B
NVS_MP_rPBR	×	×	0,749	RF
NVS_ADME_hCYP2C19	×	×	0,733	S
APR_HepG2_CellCycleArrest_24h_up		×	0,733	B
NVS_NR_hAR	×		0,727	B
ATG_ERE_CIS_up		×	0,717	S
NVS_ADME_hCYP1A2	×		0,717	B
NVS_NR_mERa	×		0,708	S
ATG_IR1_CIS_up		×	0,708	S
APR_HepG2_CellCycleArrest_72h_dn		×	0,705	S
NVS_NR_hPXR	×		0,691	B
NVS_NR_hCAR_Antagonist		×	0,689	S
NVS_MP_hPBR	×	×	0,682	S
TOX21_ERa_LUC_BG1_Agonist	×	×	0,676	S
NVS_TR_hNET	×		0,668	S
APR_HepG2_NuclearSize_72h_up	×		0,650	B
NVS_NR_hER	×	×	0,649	S
TOX21_TR_LUC_GH3_Agonist		×	0,646	S
APR_HepG2_MitoMass_24h_up		×	0,599	B
APR_HepG2_MitoMass_72h_up	×	×	0,537	S

5.4 Estimation of the applicability domain to assess the quality of predictions

The previous results showed that we were able to build good models to predict *in vitro* activity based on the structure of compounds, in particular using the stacked generalization method. In

order to evaluate if these models could be used for further investigation, we propose to estimate their applicability domain (AD) and assess the relevance of the predictions.

5.4.1 Estimation of the applicability domain

Here we use two different approaches to estimate the applicability domain of the previously built models: a range-based method using Principal Component Analysis (PCA) and a distance-based method (see Sectionii)).

PCA-based approach: In the first approach, we apply a PCA directly on the chemical descriptors of the compounds used in the training set in order to reduce the space to only a few principal components (PCs). Basically, the PCs are computed to explain a minimum of 80% of the variance or a minimum of 10 components if 80% of the variance is explained with fewer components. Thus, the number of principal components is different for each dataset. Then, for each PC descriptor, a range of acceptable values is defined by the minimal and maximal PCs values observed in the training set. For an unseen compound, if the value of at least one of its PC descriptors is out of the previously computed range, the compound is flagged "out of domain". For each model, we compute the average of the four performance metrics obtained on an external test set using either all the compounds or only the compounds that are in the AD (note that when the test set has only one active compound within the AD, the dataset is removed from the analysis).

Table 5.5 shows the percentage of assays for which performance metrics are higher when considering the test set with only the compounds belonging to the AD compared to the entire test set (compounds in and out of the AD). For more than 50% of the assays, ROC score and BA are higher when "out of AD" compounds are excluded. Specificity is higher for 88% of the assays but only 16% show higher sensitivity when using only "in AD" compounds. This can be explained by the low number of active compounds in the test sets which results in equivalent performance regardless of the compounds taken into consideration. Moreover, the number of compounds "out of AD" is comprised between 5 and 10 for all the test sets meaning that most of the compounds are included in the AD and that the compared results are based on highly similar test sets (both "in AD test set" and "entire test set" differ from only 5 to 10 compounds). These results are therefore not really highlighting the usefulness of the estimation of the AD but show that our models do represent well the compounds of the test set.

Table 5.5: Percentage of models with higher performance when using only "in AD" compounds than when using all compounds.

Metric	% of assays
ROC score	53.2
BA	53.0
Sensitivity	16.0
Specificity	88.3

Distance based approach: In the second approach, we compute the average Euclidean distance from each molecule of the test set to its three closest neighbors from the training set [239]. After sorting the compounds in ascending order of the average distance, we first cut the test set into disjoint blocks of 50 compounds. We then count the number of good predictions in each of these blocks.

Figure 5.11 displays the results of this approach applied to the previously built simple Bayesian³ and Stacked models of the assay TOX21_ERa_BLA_Antagonist_ratio which was reported by Liu to be linked to chronic liver toxicity [169].

The ROC scores of the two models were 0.77 and 0.84 for the simple Bayesian and the Stacked, respectively. The test set that enabled to build these models was composed of 1251 compounds, corresponding to 24 disjoint blocks of 50 compounds and one block of 51.

In Figure 5.11, we observe that the percentage of good predictions decreases (from 100% to less than 50%) when the average distance to the three closest compounds in the training set increases. This highlights that we can be more confident in the predictions when the compounds are closer to the ones of the training set. Moreover, the percentage of good predictions with the Stacked model is greater than with the Bayesian one (except for blocks 20 to 23).

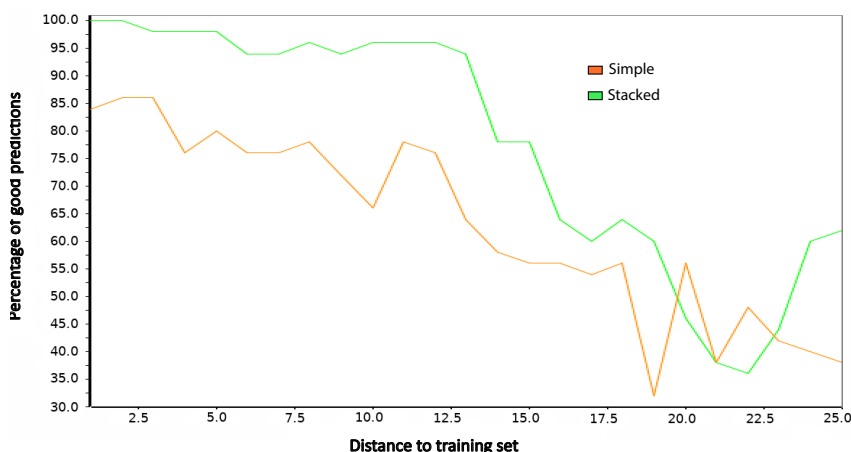


Figure 5.11: Percentage of good predictions made by the Stacked model and the simple Bayesian in all blocks of 50 predictions according to the average distance between the molecules of the test set and their three closest compounds of the training set. The dataset used corresponds to the assay TOX21_ERa_BLA_Antagonist_ratio. Percentage of good predictions decreases when the average Euclidean distance to the training set increases. Percentage of good predictions of Stacked model is almost always greater than that of the Bayesian model.

Together, these two approaches that estimate the AD demonstrate that it is an important parameter to take into account for further new predictions. Indeed, it enables one to evaluate the reliability of predictions and can be used as an important parameter for the selection of compounds for further testing.

³Note that we do not compare to the RF model since it has lower performance than the Bayesian one.

5.5 Conclusion

In this chapter, we performed a large scale analysis in order to build classifiers for all the assays from the ToxCast database, based on the compounds' chemical structures. We first used classical learning methods and two different sets of descriptors (PLP and PaDEL). The results showed that all classifiers performed similarly and led to insufficient performance, even after removing the most imbalanced datasets.

We then built ensemble classifiers using the stacked generalization method and obtained higher performances. In particular, good performances were reached for *in vitro* assays that have previously been shown to be related to specific *in vivo* toxicity outcomes.

Finally, we proposed to estimate the applicability domain of the models in order to evaluate the reliability of the predictions. We showed, using two approaches, that the AD is an important parameter to take into account for further use of the models.

Overall, we demonstrated that we could develop good and reliable machine learning models to predict some of the *in vitro* ToxCast assays by using the stacked generalization method combined to an estimation of the AD.

Now that we have explored the link between chemical structure information and *in vitro* bioactivity, we decide to focus on the second interesting relationship for toxicity prediction: between *in vitro* bioactivity data and *in vivo* observed adverse outcomes. Indeed, in the following chapter we evaluate this link using simple statistical methods and machine learning.

CHAPTER 6

FROM IN VITRO ACTIVITY TO IN VIVO TOXICITY

This chapter is devoted to the evaluation of the link between *in vitro* bioactivity data from ToxCast and *in vivo* outcomes observed in rat long-term studies available in ToxRefDB. Since endocrine disruptor chemicals (EDCs) are currently of high interest, we focus on endocrine toxicity by considering: (1) *in vitro* assays related to Estrogen Receptor (ER), Androgen Receptor (AR) and steroidogenesis pathways and (2) *in vivo* effects observed in three endocrine organs (*i.e.* adrenal gland, testis, ovary) and two sex accessory organs (*i.e.* prostate and uterus). In a first part, we look at the relationship between each *in vitro* assay and the *in vivo* outcomes and compare it with the relationship between the published ER (resp. AR) EPA's computational model results and these same outcomes. Then, we build ML models to predict the *in vivo* outcomes, either based on *in vitro* assays, on chemical structure or on both types of descriptors¹.

6.1 Endocrine Disruptor Chemicals: reminders

As already stated in Chapter 1, Endocrine Disruptor Chemicals (EDCs) are of high priority since they can lead to adverse outcomes, and are considered by regulatory authorities in EU as being able to induce adverse outcomes via non monotonic dose-response phenomena. Thus in Europe, a chemical that have been shown to induce endocrine adverse effects is banned from the market. Therefore, it is important for the agrochemicals industry to be able to assess, as early as possible, the potential endocrine mediated adverse effects of compounds under development. Moreover, since the majority of marketed compounds lacks of evaluation concerning these type of effects, it is also important for the regulatory authorities to rapidly screen for potential EDCs and prioritize them for further testing. Several short term *in vivo* assays have been proposed by the OECD to evaluate endocrine mediated toxicity, including the uterotrophic assay in rodents (screening for estrogenic properties), a 28-day oral toxicity study in rodents, a 21-day fish assay and a reproduction test performed in *Daphnia Magna* (a small planktonic crustacean) [19]. These tests

¹Part of the work presented in this chapter has been submitted to an international journal and is currently in the reviewing process [105].

were originally included in the Tier 1 of the US EPA Endocrine Disruptor Screening Program (EDSP) (see Chapter 1) which is composed of six *in vivo* tests and five *in vitro* assays related to androgen, estrogen and steroidogenesis pathways. These tests have been performed for a short list of compounds and results are publicly available in the ACToR system (see Chapter 3).

In line with the Tox21 vision, alternative methods are envisioned to assess endocrine potential of compounds and finally to replace the *in vivo* tests after validation [19]. In particular, this validation is possible only if the new methods perform equivalently or better than the existing approved ones. Furthermore, as evoked in Chapter 1, the risk assessment of EDCs should be based on mechanistic studies rather than descriptive toxicology since endocrine disruption is considered as a mode of action. Therefore, alternative methods should be based on the mechanistic evaluation of EDCs. In particular, among the well known pathways leading to endocrine effects are the estrogenic, androgenic, thyroidal and steroidogenic pathways (known as EATS). In this work we focus on the **estrogenic** (E), **androgenic** (A) and **steroidogenic** (S) pathways. Basically, the estrogenic and androgenic pathways are the ones induced by the activation of the estrogen receptor (ER) and the androgen receptor (AR), respectively. The steroidogenesis pathway is the one leading to the synthesis of all the steroid hormones such as the cortisol (synthesized in the adrenal glands), the estradiol (sexual hormone synthesized in ovaries) or the testosterone (sexual hormone synthesized in the testes and ovaries). These three pathways are involved in various biological functions in the whole organism and their perturbation can lead to several adverse effects. In particular, these effects are observed in female reproductive organs (such as ovaries and uterus), in male reproductive organs (such as testes and prostate) and in adrenal glands when respectively estrogenic pathway, androgenic pathway and steroidogenesis pathway are altered. In the ToxCast program, several HTS *in vitro* assays are targeting these pathways. Some of the results have already been used in order to assess their relevance and to develop predictive models and part of these studies has been reviewed in Chapter 3. In particular, the results from 18 assays related to the estrogenic pathway and 12 assays related to the androgenic pathway have been integrated in computational linear additive models in order to predict agonist and antagonist compounds of the two pathways, as already described in Chapter 3 [33, 142, 149]. A high correlation between the predictions based on *in vitro* ToxCast assays and short-term *in vivo* effects observed in estrogen and androgen dependent tissues could be measured on a limited number of compounds (50 to 100). Since we use the results from these models in the work presented here, more details about the methods to compute them are provided later on.

6.2 Data used

Again, the data used in this study are the ones released by the US EPA.

Chemical structures: Chemical structure information is obtained from the DSSTox database, released in October 2015 and providing a SDF file containing 9011 unique substances (see Chapter 3 for details).

***In vivo* toxicity data:** *In vivo* data are obtained from the ToxRefDB v1.0 released in October 2014 (see Chapter 3 for details).

Here we focus on outcomes observed in rat long-term studies, referred to as “CHR” *i.e.* studies in ToxRefDB. Of the CHR studies, 80% are 2-year rat carcinogenicity studies and 20% are 13-week to 31-month studies. We only use studies that have been referred to as “acceptable guideline” in the database (*i.e.* studies are complete and meet official guideline requirements), therefore retaining studies for 445 compounds. We focus on outcomes observed in five endocrine related organs: adrenal glands, ovary, testis, prostate and uterus. For each organ, various specific effects are listed in ToxRefDB and often describe a more general outcome. We therefore classify and group these specific effects into categories related to a particular adverse outcome and the full list of effects included into each category is provided in Appendix D. We finally obtain 9 outcomes distributed as follows among the 5 organs:

- 3 outcomes for **Adrenal glands**: steroidogenesis effects, stimulation and injury,
- 2 outcomes for **Ovary**: effects on germinal cells, effects on interstitial cells,
- 1 outcome for **Uterus**: effect in uterus,
- 2 outcomes for **Testis**: effects on spermatogenesis, effects on interstitial cells,
- 1 outcome for **Prostate**: effect in prostate.

For each category independently, we use these *in vivo* results as binary data by assigning 1 to compounds that induce an effect, whatever the corresponding LOAEL or NOAEL, and 0 otherwise.

***In vitro* bioactivity data:** *In vitro* data are extracted from the ToxCast database, released in October 2015. Precisely, we use the AC_{50} matrix which reports AC_{50} measured for each compound-assay pair (see also Chapter 3 for details).

Of the 445 compounds selected from ToxRefDB, 418 are also in the ToxCast database. As we focus on the E, A and S endocrine pathways, we manually select *in vitro* assays related to these pathways, based on literature, expertise and knowledge:

- For E: we select the 18 assays used by the published EPA’s ER computational models [33, 142],
- For A: we select the 12 assays used by the published EPA’s AR computational models [149],
- For S: we select 11 assays that measure the activity concentration of hormones synthesized during the steroidogenesis and 2 assays measuring the activity of one enzyme involved in this synthesis (the aromatase, which in particular converts testosterone into estradiol),
- Others: we also select 8 assays related to receptors known to be present in the 5 considered organs and involved in endocrine effects.

In an effort to ensure a robust dataset with enough representatives of active versus inactive compounds, we apply a filter to keep only assays that have at least 5% of active compounds (among the 418 compounds overlapping between ToxCast and ToxRefDB) and we end up with 42 assays.

The final list of assays is available in Table 6.1 and Appendix C provides the same list with the

associated pathway and type of each assay. In summary we use 12 assays related to ER (E), 9 related to AR (A), 2 related to aromatase (S), 11 related to steroidogenesis hormones (S) and 8 related to other receptors (O).

6.3 Estimation of relation between *in vitro* assays and *in vivo* outcomes

Firstly, we estimate the relation between each of the 42 *in vitro* assays and each of the 9 *in vivo* effect categories. For this we use a simple statistical method described below.

6.3.1 Methods

Statistical measure

In vitro assay results (AC_{50}) are turned into binary values: 1 if the ToxCast data analysis pipeline determined that the chemical was active in the assays and so an AC_{50} was reported, and 0 otherwise.

For each pair of *in vitro* assay and *in vivo* outcome, we compute the three following metrics:

- Sensitivity = $TP/(TP + FN)$;
- Specificity = $TN/(TN + FP)$;
- Balanced Accuracy (BA) = $\frac{(Sensitivity+Specificity)}{2}$;

where TP (respectively TN) is the number of True Positive (respectively True Negative) compounds, *i.e.* compounds which are positive (respectively negative) for both *in vitro* assay and *in vivo* outcome; and FP (respectively FN) is the number of False Positive (respectively False Negative) compounds, *i.e.* compounds which are positive (respectively negative) *in vitro* but negative (respectively positive) *in vivo*.

Comparison to ER and AR computational models

In order to not only look at the relation of the *in vivo* outcomes with each *in vitro* assay considered independently, we also evaluate the relation of these outcomes with a combination of assays. To do so, we refer to the published work of EPA's researchers for the prediction of ER [142, 33] and AR activity of compounds [149], see Section 3.4.1. Indeed, they proposed computational models that aggregate the results of several *in vitro* assays related to ER and AR pathways. Since we use the results of these models in our work, the end of this section technically presents the method developed by the authors to combine the assays and compute a predicted activity of compounds.

As briefly stated in Section 3.4.1, for each tested compound, these models integrate the concentration - response curves obtained for each of the 18 (resp. 12) assays related to ER (resp. AR) pathway. In particular, the models take into account the non specific assay interference

Table 6.1: Summary of in vitro data. Number of positive and negative compounds in each of the 42 in vitro assays selected and in the EPA's models for ER and AR over the 418 compounds.

Assay name	Pathway	# of tested compounds	# of inactive compounds	# of active compounds	% of active compounds
ACEA_T47D_80hr_Positive	E	367	319	48	13.08
ATG_ERE_CIS_up	E	397	277	120	30.23
ATG_ERa_TRANS_up	E	397	297	100	25.19
OT_ER_EraERb_0480	E	368	328	40	10.87
OT_ER_EraERb_1440	E	368	341	27	7.34
OT_ER_ErbERb_0480	E	368	328	40	10.87
OT_ER_ErbERb_1440	E	368	345	23	6.25
OT_Era_EREFGFP_0120	E	368	340	28	7.61
OT_Era_EREFGFP_0480	E	368	345	23	6.25
TOX21_Era_BLA_Antagonist_ratio	E	404	319	85	21.04
TOX21_Era_LUC_BG1_Agonist	E	404	335	69	17.08
TOX21_Era_LUC_BG1_Antagonist	E	404	332	72	17.82
NVS_NR_cAR	A	373	329	44	11.80
NVS_NR_hAR	A	388	346	42	10.82
NVS_NR_rAR	A	397	377	20	5.04
OT_AR_ARELUC_AG_1440	A	368	343	25	6.79
OT_AR_ARSRC1_0480	A	368	336	32	8.70
OT_AR_ARSRC1_0960	A	368	307	61	16.58
TOX21_AR_BLA_Antagonist_ratio	A	404	292	112	27.72
TOX21_AR_LUC_MDAKB2_Antagonist	A	404	308	96	23.76
TOX21_AR_LUC_MDAKB2_Antagonist2	A	402	278	124	30.85
CEETOX_H295R_11DCORT_dn	S	349	301	48	13.75
CEETOX_H295R_ANDR_dn	S	349	307	42	12.03
CEETOX_H295R_CORTISOL_dn	S	349	314	35	10.03
CEETOX_H295R_DOC_dn	S	349	319	30	8.60
CEETOX_H295R ESTRADIOL_up	S	349	328	21	6.02
CEETOX_H295R ESTRONE_dn	S	349	331	18	5.16
CEETOX_H295R ESTRONE_up	S	349	324	25	7.16
CEETOX_H295R OHPROG_dn	S	349	312	37	10.60
CEETOX_H295R OHPROG_up	S	349	324	25	7.16
CEETOX_H295R PROG_up	S	349	322	27	7.74
CEETOX_H295R TESTO_dn	S	349	314	35	10.03
NVS_ADME_hCYP19A1	S	384	360	24	6.25
TOX21_Aromatase_Inhibition	S	404	286	118	29.21
ATG_Sp1_CIS_up	O	397	344	53	13.35
ATG_GRE_CIS_dn	O	397	360	37	9.32
ATG_SREBP_CIS_up	O	397	290	107	26.95
NVS_NR_bPR	O	384	355	29	7.55
NVS_NR_hGR	O	393	340	53	13.49
NVS_NR_hPR	O	393	371	22	5.60
TOX21_GR_BLA_Agonist_ratio	O	404	375	29	7.18
TOX21_GR_BLA_Antagonist_ratio	O	404	372	32	7.92
ER EPA model	E	361	356	5	1.39
AR EPA model	A	361	306	55	15.24

of compounds (*i.e.* compounds that are active in the assay but not due to a specific activity towards the biological target): for each concentration, the models assume that the global activity of the compound in the entire pathway is a non-weighted linear sum of its activities in all the assays triggering the pathways but also of non specific activities. Also, by using a non-weighted sum, the models assume that each of the 18 (resp. 12) assay equally contributes to the global ER (resp. AR) pathway activity.

Finally, for each compound and each concentration, the models enable the computation of a global predicted activity for the entire pathway. The activities for all the concentrations are then transposed into a concentration-curve for which the AUC can be computed and which corresponds to the predicted final (ER) (resp. AR) score of the ER (resp. AR) pathway model, ranging from 0 to 1. More details about the mathematical functions used to compute these models can be found in Judson *et al.* [142] (ER model) and Kleinstreuer *et al.* [149] (AR model). For both models, these predictions were translated as follows:

- $score \geq 0.1$: compound is active in the pathway;
- $0 < score < 0.1$: inconclusive;
- $score = 0$: compound is inactive in the pathway.

Moreover, the models also consider assay-interference and cytotoxicity by computing several quality criteria and putting some flags to the final scores.

Regarding the ER model, two scores are computed for each pair of compound and assay:

- a **Z-score** that flags non selective assay activity due to cytotoxicity by using the results of 35 cytotoxicity assays available in ToxCast. In particular this score assumes that, if a compound shows cytotoxicity at a concentration lower than the concentration at which it is active in an ER-related assay, then this activity is not relevant and the compound is a false positive for the considered ER assay. For each assay and compound, the Z-score is computed by measuring the difference between the AC_{50} obtained for the compound in the assay and the median AC_{50} obtained in the cytotoxicity assays, with regards to global cytotoxicity across all chemicals. This could be seen to an equivalent to a standardization which rescales the data by centering and reducing according to the cytotoxicity results. A low Z-score (lower than 3) indicates that the activity measured in the considered assay could be due to cytotoxicity and not to a target-selective mechanism;
- a **T-score** that corresponds to the maximum activity measured (*i.e.* the highest point of the concentration-response curve). Indeed, since concentration-response curves are normalized compared to a control or baseline, the maximal activity is a relative percentage not necessarily equals to 100%. The T-score corresponds to the highest value of this relative percentage.

For each compound, a median of all their Z-scores and T-scores obtained in all assays are computed and referred to respectively as **med.Z** and **med.T**.

Regarding the AR model, for each pair of compound and assay, a confidence score is provided

and takes into account the score of the model, the same Z-score as for the ER model and the results of a supplemental assay which can confirm the antagonist activity of chemicals and therefore that it is a True Positive. Basically, one of the assay was run twice with two different initial experimental conditions which allowed the distinction between specific and non specific compounds for this assay.

Among the 418 compounds from our study, 361 have a score available for the ER and AR models. To discriminate between positive and negative compounds for these models, we choose the following thresholds for the different values available:

- Positive for ER model if model score > 0.1 and med.T $> 50\%$ and med.Z > 3 ; negative otherwise,
- Positive for AR model if model score > 0.1 and confidence score for antagonist activity > 0 ; negative otherwise.

In the end, 5 compounds are positive for the ER model and 55 for the AR model.

We perform the analysis of the relation between the ER and AR models and the *in vivo* effects as previously described for the 42 assays independently. Note that from the 18 assays used by Judson *et al.* [142] in their model for ER, 7 are excluded from our study because their hit rate is below our cutoff of 5%. This is also the case for 3 assays of the 12 used by Kleinstreuer *et al.* [149] for the AR model.

6.3.2 Results

From the 418 compounds that have both *in vivo* data for chronic rat studies in ToxRefDB and *in vitro* results in ToxCast, 349 have been tested in all the 42 selected *in vitro* assays. Table 6.1 provides for each assay and the two EPA's model the total number of compounds tested (among the 418) with the number of positive and negative ones as well as the corresponding percentage of actives.

The results show a range of percentage of actives only between 5 and 30% (mean 12.7%) indicative of highly imbalanced data in favor of negative compounds.

Regarding the computational models for ER and AR, scores are available for 361 compounds among the 418. After applying the previously described filters to discriminate between positives and negatives, only 5 compounds are positive among the 361 for the ER model (1.4%) and 55 for the AR model (15%).

Table 6.2 summarizes the *in vivo* data used with the number of positive and negative compounds among the 418 for each of the 9 effect categories for the 5 organs. Here again the data are highly imbalanced in favor of negative compounds with 4 to 16% of positive compounds depending on the category.

For each *in vivo* outcome, we plot the sensitivity, specificity and balanced accuracy in order to evaluate the relation of the outcome with the 42 selected *in vitro* assays and the AR and ER

Table 6.2: Summary of *in vivo* data. Number of positive and negative compounds in each of the 9 endocrine *in vivo* outcomes over the 418 compounds.

<i>In vivo</i> endpoint	# negative compounds	# positive compounds	# tested compounds	% positive compounds	Figure
Steroidogenesis adrenal glands	360	58	418	13.88	Figure 6.1
Stimulation adrenal glands	350	68	418	16.27	Figure 6.1
Injury adrenal glands	363	55	418	13.16	Figure 6.1
Germinal cells ovary	387	31	418	7.42	Figure 6.2
Interstitial cells effect ovary	382	36	418	8.61	Figure 6.2
Uterus effect	375	43	418	10.29	Figure 6.2
Spermatogenesis testis	375	43	418	10.29	Figure 6.3
Interstitial cells testis	351	67	418	16.03	Figure 6.3
Prostate effect	401	17	418	4.07	Figure 6.3

models. Figures 6.1, 6.2 and 6.3 respectively correspond to the effects in adrenals, ovaries and uterus, and testes and prostate.

Relation with the 42 *in vitro* assays: We observe that for all pairs of assay and *in vivo* outcome, the specificity is high (between 0.85 and 0.95) and the sensitivity is very low (lower than 0.3) leading to an overall BA around 0.5. Of note, ER pathway related assays (E)² do not show a higher BA for ovary and uterus outcomes compared to the other assays (Figure 6.2). The same is observed for the steroidogenesis (S) and AR (A)³ pathways related assays and adrenal gland (Figure 6.1) and testis and prostate outcomes (Figure 6.3), respectively.

Relation with the ER and AR computational models:

- For the ER model, we do observe an increase of BA and sensitivity for 4 outcomes (effects on germinal cells ovary, on uterus, on spermatogenesis testis and on prostate) but since there is only 1.4% of positive compounds for this model in our datasets, we cannot consider this finding as relevant without another study extended to more compounds.
- For the AR model, when specifically looking at the relation with effects in prostate and testis, we do not see any difference compared to the individual AR pathway related assays. This result show that even the aggregation of several assays related to the AR pathway does not improve the link with the selected *in vivo* outcomes that are examined here.

Overall, this simple statistical analysis demonstrates that, globally, there is no mutual linear correlation between the 42 *in vitro* assays and any of the selected *in vivo* outcomes when we use the results of each assay independently. Somewhat surprisingly, this observation is also made for the *in vitro* assays with targets physiologically related to specific *in vivo* adverse outcomes (*e.g.* *in vitro* AR assays are not correlated with effects observed in prostate or testis known to result

²Here we do not consider the ER EPA's model since we only look at each assay independently.

³Here we do not consider the AR EPA's model since we only look at each assay independently.

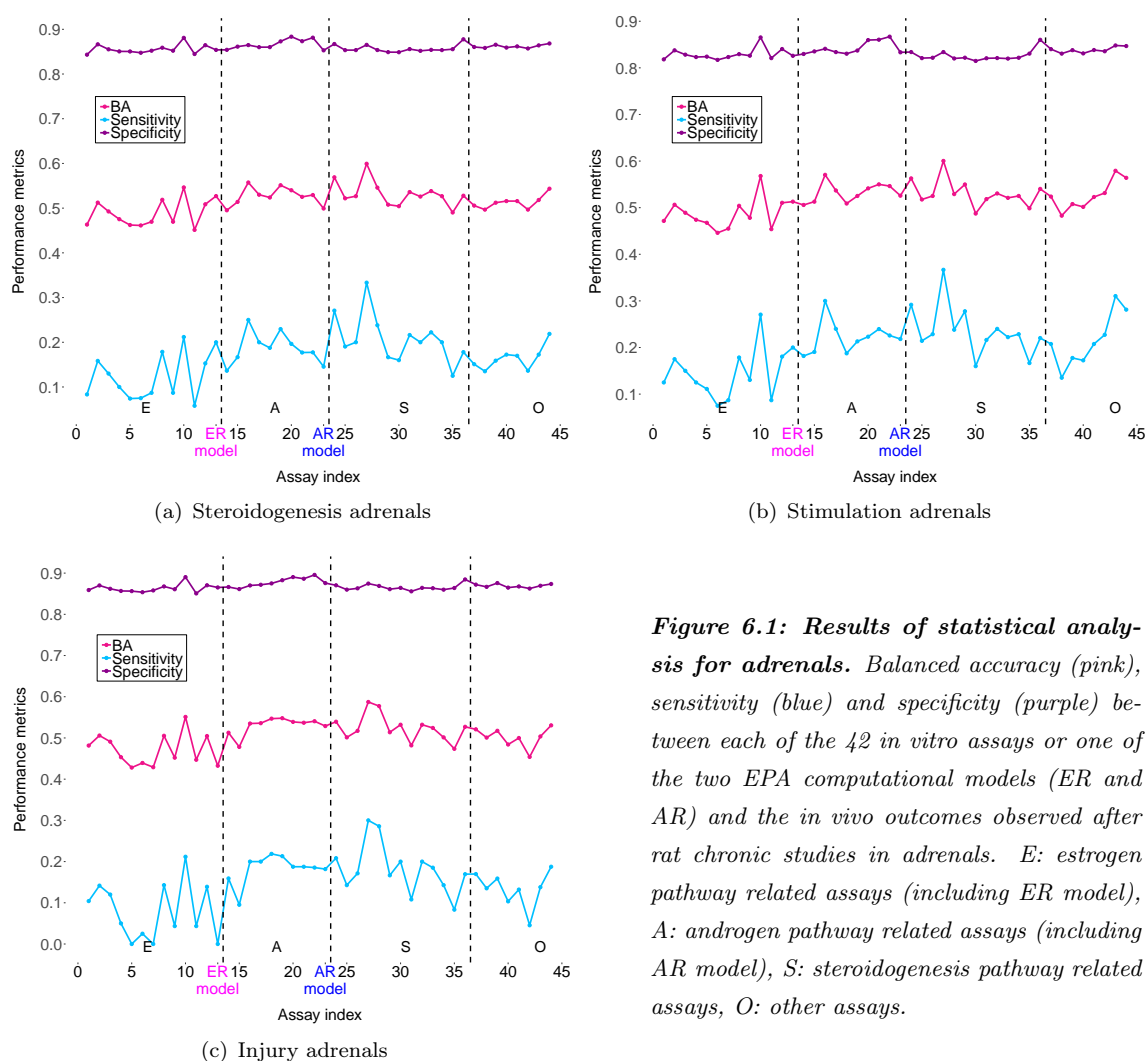


Figure 6.1: Results of statistical analysis for adrenals. Balanced accuracy (pink), sensitivity (blue) and specificity (purple) between each of the 42 *in vitro* assays or one of the two EPA computational models (ER and AR) and the *in vivo* outcomes observed after rat chronic studies in adrenals. E: estrogen pathway related assays (including ER model), A: androgen pathway related assays (including AR model), S: steroidogenesis pathway related assays, O: other assays.

from a perturbation of the AR pathway).

We could already anticipate that a single assay alone would not be sufficient to inform about a long-term *in vivo* outcome. That is why we also evaluated the relation between the *in vivo* outcomes and the results of the EPA’s computational models that consider a linear combination of several *in vitro* assays. Nonetheless, still no relation could be concluded. Since these computational models only perform a linear combination, we propose to investigate non linear combinations through ML methods and therefore investigate if *in vivo* outcomes could be predicted by a combination of several of the 42 *in vitro* assays using machine learning.

6.4 Machine Learning to predict *in vivo* outcomes

We use ML methods to predict the *in vivo* outcomes observed in the five considered organs from either the structure of compounds and / or their *in vitro* bioactivity. In particular, we build three classifiers for each of the 9 *in vivo* effect categories for the five organs: one based on biological descriptors (the 42 *in vitro* assays), one based on the chemical structure (physicochemical properties and fingerprints) and one based on a combination of both types of descriptors.

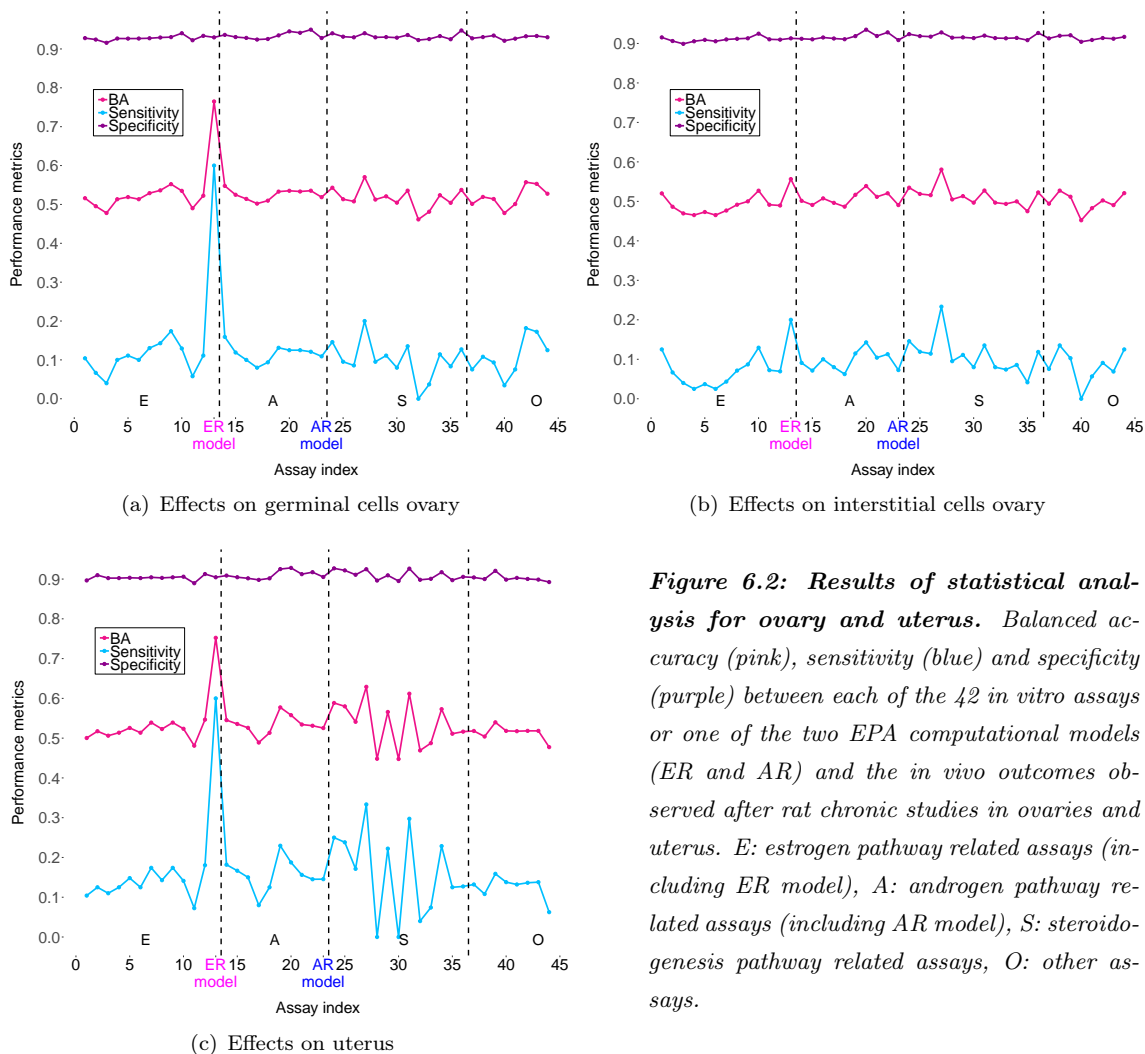


Figure 6.2: Results of statistical analysis for ovary and uterus. Balanced accuracy (pink), sensitivity (blue) and specificity (purple) between each of the 42 *in vitro* assays or one of the two EPA computational models (ER and AR) and the *in vivo* outcomes observed after rat chronic studies in ovaries and uterus. E: estrogen pathway related assays (including ER model), A: androgen pathway related assays (including AR model), S: steroidogenesis pathway related assays, O: other assays.

6.4.1 Methods

The ML approach used here is inspired from the work of Liu *et al.* who built ML models based on the same types of descriptors to predict *in vivo* effects observed in the liver [169], see Section 3.4.2.

Datasets

Since the ML methods used in this approach are not good at handling missing data, we identify a complete matrix between all *in vitro* and *in vivo* available data. For this we only include compounds that have been tested in all the 42 assays previously selected. From the 418 compounds available in both ToxCast and ToxRefDB, 341 compounds meet this criteria which is eight less than previously because these 8 compounds are not included in ToxRefDB.

Table 6.3 summarizes the number of positive and negative compounds in the datasets used for each *in vivo* effect category for the 5 organs. As Liu *et al.*, we choose to be quite stringent and call a compound negative (assigned a value of 0) for a specific organ only if it is negative for all the effect categories related to this organ. For example, if a compound does not induce any of

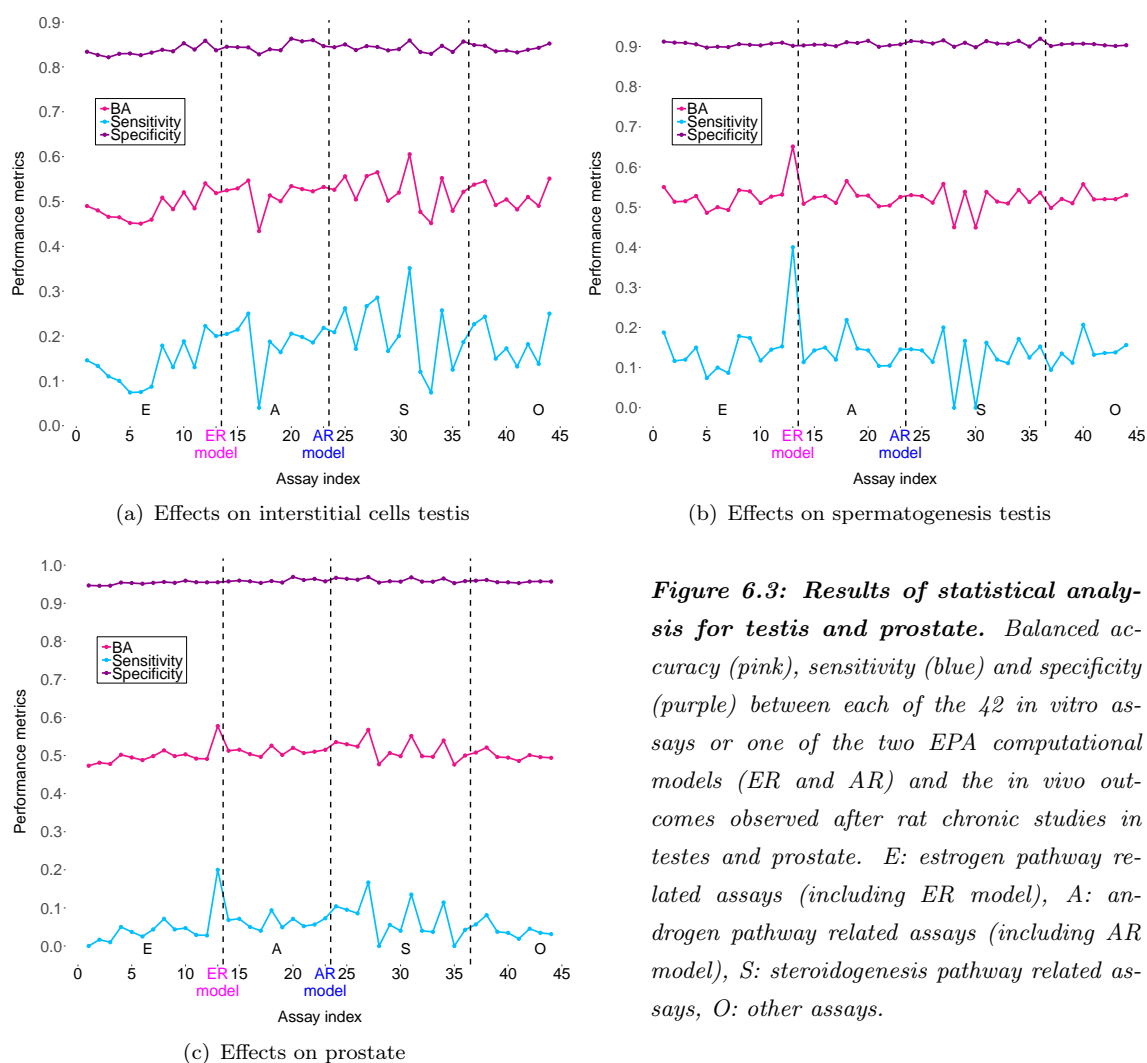


Figure 6.3: Results of statistical analysis for testis and prostate. Balanced accuracy (pink), sensitivity (blue) and specificity (purple) between each of the 42 *in vitro* assays or one of the two EPA computational models (ER and AR) and the *in vivo* outcomes observed after rat chronic studies in testes and prostate. E: estrogen pathway related assays (including ER model), A: androgen pathway related assays (including AR model), S: steroidogenesis pathway related assays, O: other assays.

the 3 category effects in the adrenal glands (Steroidogenesis effects, Stimulation or Injury), it is considered as “negative”. However, if a compound induces one of these 3 category effects (*e.g.*, Stimulation), it is considered positive in the “Stimulation” dataset but discarded from the two other adrenal glands datasets (Steroidogenesis effects and Injury). Therefore, not all the 341 compounds are represented in each dataset.

The table illustrates that for all the endpoints considered, the datasets for ML are ones again highly imbalanced in favor of negative compounds. The lowest percentage of actives is for the effects in the prostate (4.5%) and the highest is for stimulation in adrenal glands (19%).

Chemical descriptors

The SDF for the 341 compounds is an extract of the SDF from the DSSTox database containing 9011 structures.

After cleaning the structures of the compounds (removing salts and inorganic elements, neutralizing and checking for duplicates), we compute two types of molecular descriptors:

- **74 physico-chemical properties** using the RDKit⁴ tool available in Knime [22]. Continuous

⁴<http://www.rdkit.org/>

Table 6.3: Datasets for machine learning. Number of positive and negative compounds for each dataset to predict the 9 *in vivo* outcomes corresponding to 5 endocrine organs. For adrenal glands, testis and ovary, compounds are negatives for the organ if they are negative for all the organ’s categories.

Organ name	Endpoint	# positive compounds	% positive compounds	# negative compounds
Adrenal glands	Steroidogenesis effects	51	16	264
	Stimulation	62	19	
	Injury	47	15	
Ovary	Effect on germinal cells	25	7.5	307
	Effect on interstitial cells	29	8.6	
Uterus	Uterus effect	33	9.7	308
Testis	Effect on spermatogenesis	33	11	270
	Effect on interstitial cells	56	17	
Prostate	Prostate effect	15	4.5	326

values are normalized between 0 and 1 using the min-max normalization [204];

- 731 fingerprints using the *pybel* package in Python [195] and PaDEL software [285]

Bioactivity descriptors

The 42 *in vitro* assays selected are used as individual descriptors of the bioactivity of compounds. As in Liu *et al.* [169], we set AC_{50} values of inactive compounds to $1.10^6 \mu M$ and transformed all the AC_{50} according to the following formula:

$$AC'_{50} = 6 - \log_{10}(AC_{50})$$

This formula gives inactive compounds a value of 0 and represents active ones on a continuous ascending scale and reflects their potential activity. Then the values are normalized between 0 and 1 using the min-max normalization [204].

Learning procedure and evaluation

The learning procedure is also inspired from Liu *et al.* and is implemented in Python2.7.

In total, we use the 8 following classification algorithms (see Chapter 2 for details):

- Linear Discriminant Analysis (LDA)
- Naïve Bayes (NB)
- Support Vector Machines (SVM) with two different kernels: linear (SVCL) and radial basis function (SVCR)
- Classification and regression trees (CART)
- K-nearest neighbors (KNN)
- Random Forest (RF)

- Ensemble technique (ENSMB) for which the prediction corresponds to the majority vote of the six previous classifiers

All these algorithms are used with their default parameters except for RF for which we use 100 trees.

Compared to Liu, we add the Random Forest algorithm because it is an ensemble method that performs better in terms of generalization than a single regression tree which lowers risk of overfitting and that can handle many input features [252].

For each model, a 10-fold cross-validation testing is performed and repeated 20 times. For each step in the cross-validation loop, the descriptors are ranked by computing their importance score using the Random Forest attribute *feature_importance* which is based on the Gini index [187] (measure that provides a relative ranking of the features). Note that this is different from Liu *et al.* who computed the univariate association between each pair of descriptor and *in vivo* outcome. Indeed, we prefer the *feature_importance* because, unlike the univariate association, it considers all the features together at the same time to compute the rank, and not one by one. This ranking reflects which of the descriptors are the most important to build the trees that best split the data into subsets corresponding to the two classes. Then, classifiers are built using the 10 best descriptors and iteratively adding one descriptor at each step. This iteration stops at 42 descriptors when only *in vitro* assays are used and at 70 (arbitrary value inspired from Liu *et al.*) when molecular descriptors are used (either alone or combined to *in vitro* ones). For each outcome and classification algorithm, the process finally leads to 33 models (10 to 42 descriptors) based on *in vitro* assays, 61 models (10 to 70 descriptors) based on molecular descriptors and also 61 models based on a combination of both. Performance of all classifiers are evaluated using the three following metrics: sensitivity, specificity, and balanced accuracy (BA). Finally, for each triplet of outcome, algorithm and descriptor type, we look for the model that reaches the highest BA among all and report its corresponding sensitivity and specificity as well as the number of descriptors used in the model.

Data augmentation

Since the datasets in this study are imbalanced (more inactive compounds than active ones) we use a data augmentation technique to rebalance the data and evaluate whether it affects the performance of the classifiers. We utilize the Synthetic minority over-sampling technique [43], SMOTE for short, see Section 2.5 for details. We use this technique in each step of the cross-validation loop in order to increase the number of compounds of the minority class of each training set.

6.4.2 Performance results

Figures 6.4, 6.5 and 6.6 show the performances of the models obtained for the 9 effects using the Random Forest (RF) algorithm, before and after performing data augmentation using SMOTE.

The results obtained with all the methods are available in Appendix E.

Performance of ML models based on *in vitro* assays

Regarding ML models that are based on *in vitro* assays alone, whatever the *in vivo* outcomes, all models have low sensitivity (between 0.05 and 0.09) and high specificity (between 0.95 and 0.99), and BA between 0.50 and 0.53.

In order to evaluate the impact of the SMOTE technique, we perform a paired t-test that compares the average performance over the 200 (20×10 -fold cross-validation) values obtained on the original data (imbalanced) and the average performance over the 200 values obtained after applying SMOTE (balanced data). Table 6.4 reports the p-values obtained for the t-test for the 9 *in vivo* outcomes and the 3 performance metrics. This table shows that, considering a cutoff of 0.05 for significance, the SMOTE method significantly affects sensitivity and specificity. When looking at Figures 6.4, 6.5 and 6.6, the sensitivity is increased by a factor of at least 2 for all outcomes (for example in Figure 6.1, (a) and (b), sensitivity increases from 0.08 to 0.48 for the outcome "steroidogenesis adrenals"). On the contrary, the specificity is always decreased (from 0.96 to 0.89 for the same outcome, see Figure 6.1, (a) and (b)). Consequently, these opposite variations result in a BA still around 0.50, not significantly different from the results obtained on imbalanced datasets, except for two outcomes. Similar results are obtained with the other algorithms (see Appendix E).

Table 6.4: Effect of the SMOTE technique on ML models based on *in vitro* assays. The table presents the p-values of the t-test which compares the average performance of ML models before and after applying the SMOTE technique. p-values lower than 0.05 are in bold. bio = *in vitro* assays

<i>In vivo</i> outcome	Figure	Bio		
		BA	Sensitivity	Specificity
Steroidogenesis adrenal glands	6.4(a) & (b)	9.12×10^{-1}	8.58×10^{-30}	9.29×10^{-36}
Stimulation adrenal glands	6.4(c) & (d)	5.49×10^{-1}	1.72×10^{-13}	2.95×10^{-18}
Injury adrenal glands	6.4(e) & (f)	8.21×10^{-4}	8.60×10^{-30}	2.55×10^{-32}
Germinal cells ovary	6.5(a) & (b)	2.07×10^{-1}	1.44×10^{-2}	2.05×10^{-20}
Interstitial cells effect ovary	6.5(c) & (d)	1.07×10^{-2}	2.94×10^{-7}	1.06×10^{-18}
Uterus effect	6.5(e) & (f)	1.63×10^{-1}	7.19×10^{-8}	6.63×10^{-14}
Spermatogenesis testis	6.6(a) & (b)	1.86×10^{-1}	1.47×10^{-3}	3.68×10^{-7}
Interstitial cells effect testis	6.6(c) & (d)	8.28×10^{-1}	2.00×10^{-27}	1.35×10^{-33}
Prostate effect	6.6(e) & (f)	3.04×10^{-1}	2.32×10^{-2}	1.80×10^{-15}

These results show that ML models that predict *in vivo* effects observed in endocrine organs from the selected *in vitro* assays do not perform better than chance (BA around 0.5) and that data augmentation does not help to increase their performance. This highlights that a combination (linear or not) of different *in vitro* assays is also not correlated to the selected long-term *in vivo* effects and cannot help to predict them.

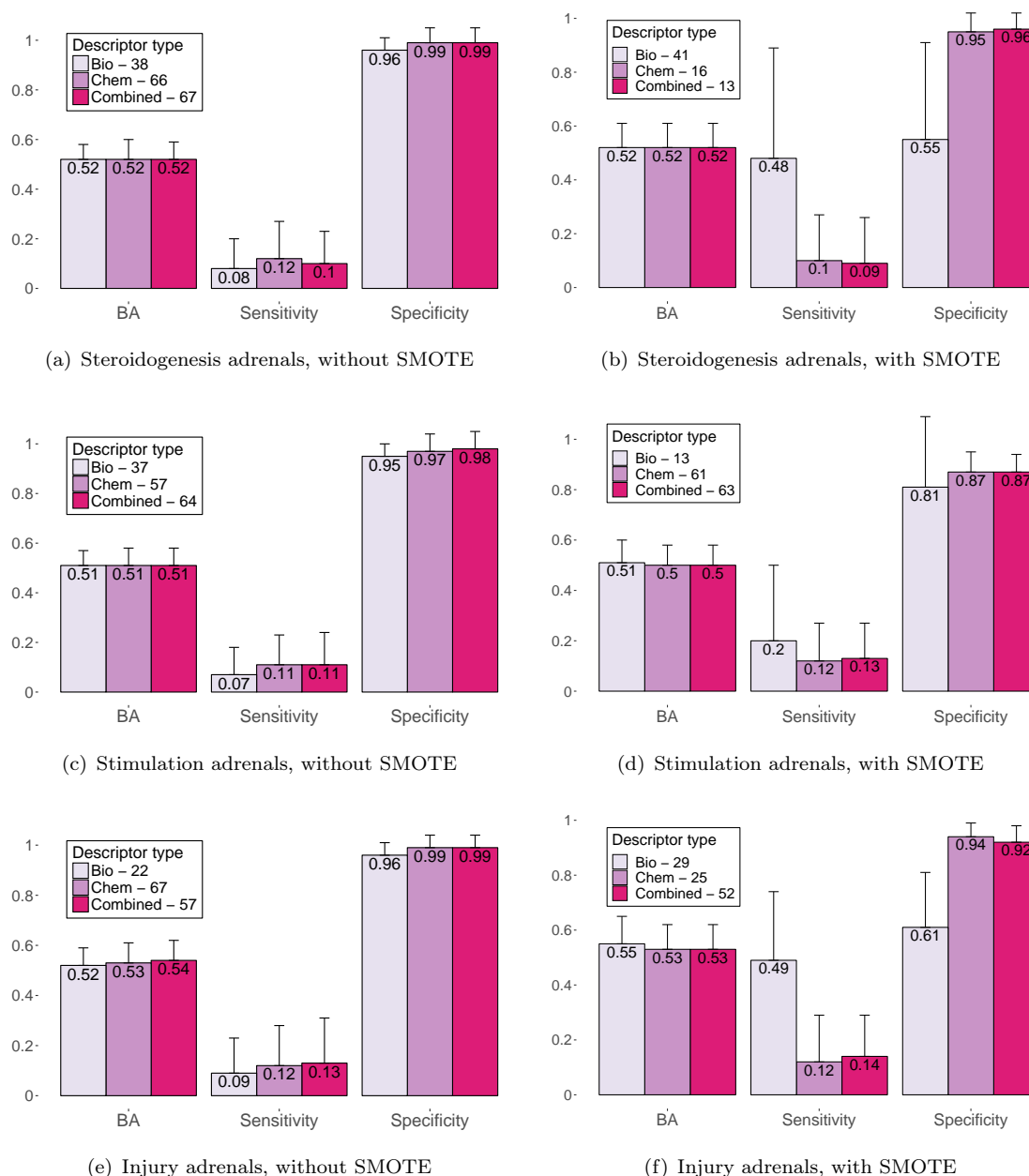


Figure 6.4: Performances of ML models that predict adrenals outcomes using RF algorithm. a,b)- Steroidogenesis, c,d)- Stimulation, e,f)- Injury. Left panel: without SMOTE methods, right panel: with SMOTE method. The reported results correspond to the models that reaches the highest BA among all the 33 or 61 RF models with its corresponding sensitivity, specificity as well as the number of descriptors used (numbers in the legend). The different colors represent the types of descriptors used in the models (bio: *in vitro* assays, chem: molecular descriptors, combined: combination of both).

Performance of ML models based on molecular descriptors

We now look at the performance of ML models based on only molecular descriptors or combined with *in vitro* assays in order to see if chemical structure information helps in the predictions. Without data augmentation, we observe that the performance of the models is similar to the ones of the models built with *in vitro* assays alone. These observations are confirmed by a t-test

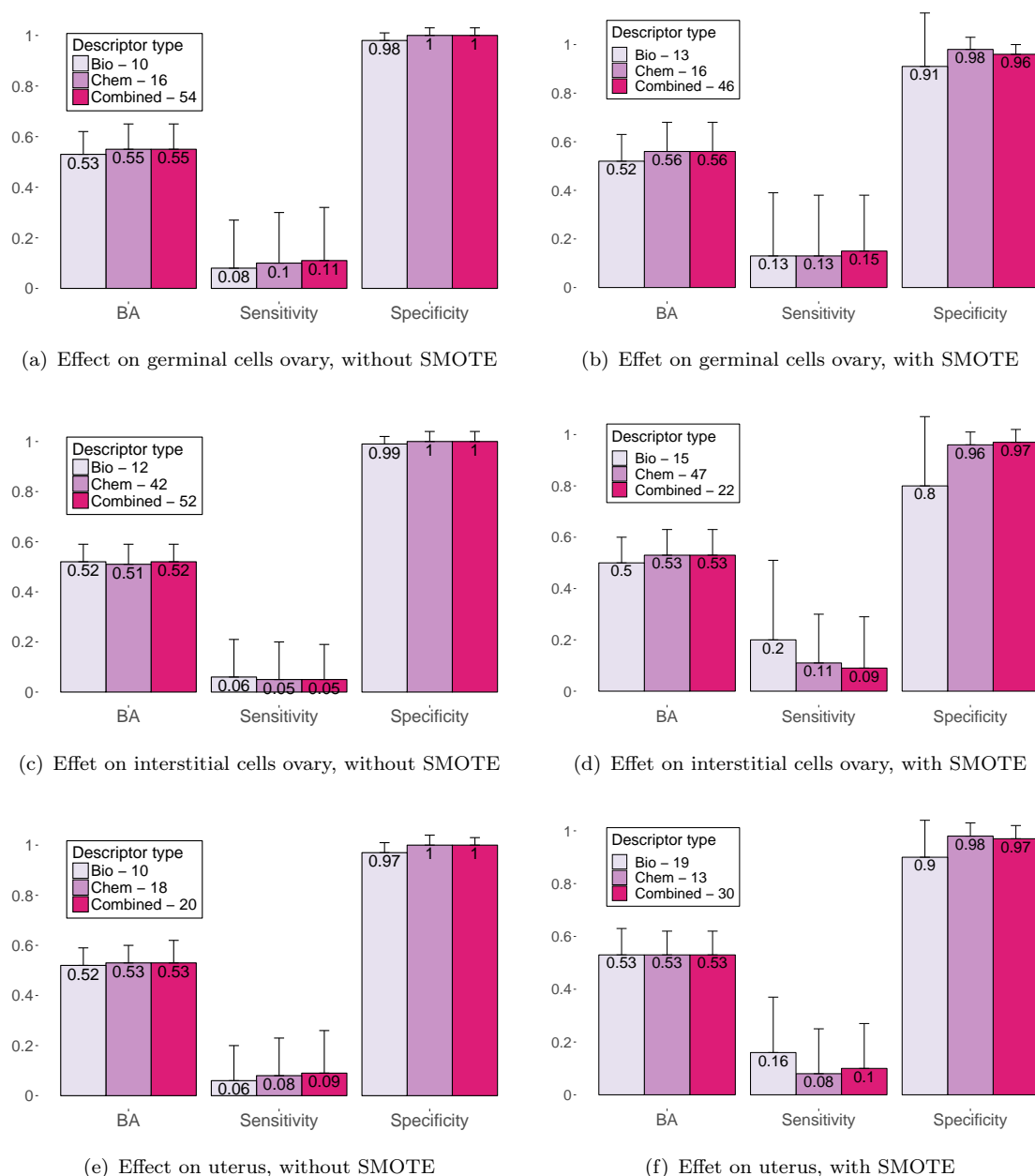


Figure 6.5: Performances of ML models that predict ovaries and uterus outcomes using RF algorithm. a,b)- Effects on germinal cells, c,d)- Effects on interstitial cells, e,f)- Effects on uterus. Left panel: without SMOTE methods, right panel: with SMOTE method. The reported results correspond to the models that reaches the highest BA among all the 33 or 61 RF models with its corresponding sensitivity, specificity as well as the number of descriptors used (numbers in the legend). The different colors represent the types of descriptors used in the models (bio: *in vitro* assays, chem: molecular descriptors, combined: combination of both).

(performed in the same way than previously) comparing the results of the ML models based on *in vitro* assays and either molecular descriptors only or molecular descriptors combined with *in vitro* assays: no significant difference of the average BA between each type of descriptors is observed, using a p.value cutoff of 0.05 (data not shown).

Regarding the impact of the SMOTE technique, the p-values of the t-test comparing results before and after applying the SMOTE method are presented in Table 6.5. From Figures 6.4, 6.5, 6.6

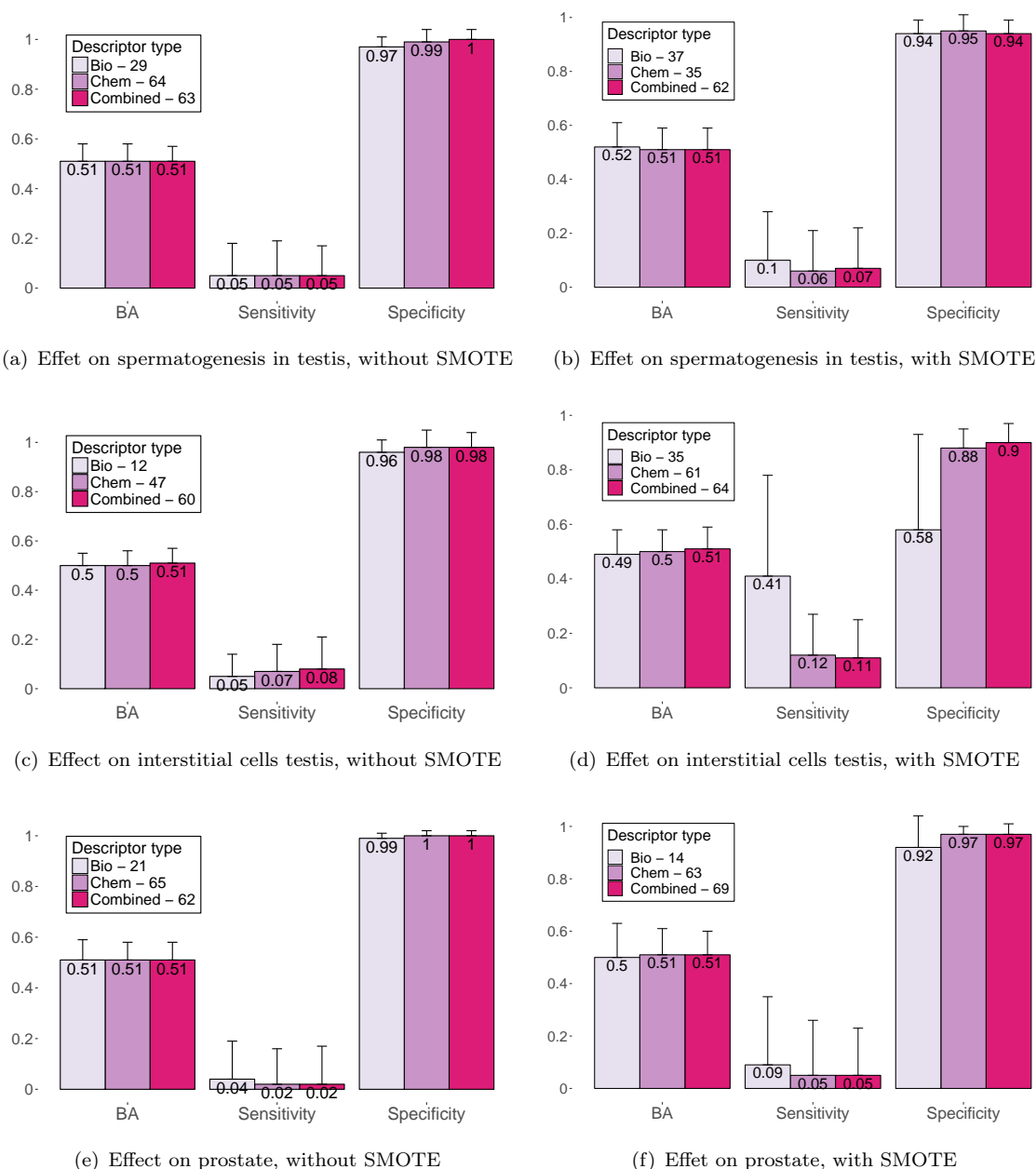


Figure 6.6: Performances of ML models that predict testis and prostate outcomes using RF algorithm. a,b)- Effects on spermatogenesis, c,d)- Effects on interstitial cells, e,f)- Effects on prostate. Left panel: without SMOTE methods, right panel: with SMOTE method. The reported results correspond to the models that reaches the highest BA among all the 33 or 61 RF models with its corresponding sensitivity, specificity as well as the number of descriptors used (numbers in the legend). The different colors represent the types of descriptors used in the models (bio: in vitro assays, chem: molecular descriptors, combined: combination of both).

and Table 6.5, we observe that, whether for the models based on molecular descriptors alone (chem) or the ones based on the two types of descriptors (combined), the specificity is always significantly decreased. Nonetheless, the sensitivity is generally not significantly increased resulting in an overall BA not significantly impacted.

Table 6.5: Effect of the SMOTE technique on ML models based on molecular descriptors. The table presents the *p*-values of the *t*-test which compares the average performance of ML models before and after applying the SMOTE technique. *p*-values lower than 0.05 are in bold. chem = molecular descriptors

<i>In vivo</i> outcome	Figure	Chem			Combined		
		BA	Sensitivity	Specificity	BA	Sensitivity	Specificity
Steroidogenesis adrenal glands	6.4(a) & (b)	7.10×10^{-1}	3.06×10^{-1}	3.36×10^{-4}	6.03×10^{-1}	4.49×10^{-1}	1.99×10^{-4}
Stimulation adrenal glands	6.4(c) & (d)	3.48×10^{-2}	9,84E-02	2.31×10^{-13}	7.21×10^{-2}	7.61×10^{-2}	3.64×10^{-12}
Injury adrenal glands	6.4(e) & (e)	8.53×10^{-1}	5.78×10^{-1}	2.32×10^{-2}	1.23×10^{-1}	4.82×10^{-1}	1.67×10^{-12}
Germinal cells ovary	6.5(a) & (b)	3.07×10^{-1}	9.30×10^{-2}	5.94×10^{-12}	3.86×10^{-1}	6.56×10^{-2}	2.45×10^{-11}
Interstitial cells effect ovary	6.5(a) & (d)	7.97×10^{-3}	3.28×10^{-6}	1.91×10^{-17}	3.29×10^{-2}	7.38×10^{-3}	1.98×10^{-2}
Uterus effect	6.5(e) & (e)	8.69×10^{-1}	4.39×10^{-1}	4.15×10^{-14}	9.93×10^{-1}	9.29×10^{-2}	3.98×10^{-22}
Spermatogenesis testis	6.6(a) & (b)	7.68×10^{-1}	3.72×10^{-1}	1.38×10^{-4}	7.89×10^{-1}	1.60×10^{-2}	9.49×10^{-15}
Interstitial cells effect testis	6.6(c) & (d)	9.68×10^{-1}	9.87×10^{-7}	1.29×10^{-20}	8.37×10^{-1}	4.28×10^{-2}	3.29×10^{-7}
Prostate effect	6.6(e) & (e)	3.93×10^{-1}	7.41×10^{-2}	4.96×10^{-11}	8.61×10^{-1}	1.80×10^{-1}	7.38×10^{-15}

Table 6.6: Comparison of models' performance according to the types of descriptors they use and when SMOTE is applied. The table presents the *p*-values of the *t*-test which compares the average performance of ML models based on *in vitro* assays alone and the ones based either on chemical descriptors or on a combination of *in vitro* assays and chemical descriptors (combined). *p*-values lower than 0.05 are in bold. bio = *in vitro* assays

<i>In vivo</i> outcome	Figure	BA		Sensitivity		Specificity	
		Bio VS chemical	Bio VS combined	Bio VS chemical	Bio VS combined	Bio VS chemical	Bio VS combined
Steroidogenesis adrenal glands	6.4(b)	5.44×10^{-1}	7.19×10^{-1}	1.51×10^{-25}	1.56×10^{-28}	2.54×10^{-35}	1.19×10^{-36}
Stimulation adrenal glands	6.4(d)	1.08×10^{-1}	3.32×10^{-1}	1.03×10^{-5}	3.34×10^{-5}	5.63×10^{-5}	1.30×10^{-4}
Injury adrenal glands	6.4(f)	2.75×10^{-2}	1.88×10^{-2}	1.78×10^{-24}	1.44×10^{-22}	3.02×10^{-30}	2.76×10^{-27}
Germinal cells ovary	6.5(b)	4.28×10^{-4}	5.05×10^{-4}	9.04×10^{-1}	2.53×10^{-1}	1.17×10^{-19}	1.66×10^{-12}
Interstitial cells effect ovary	6.5(b)	2.32×10^{-4}	8.52×10^{-4}	9.40×10^{-4}	7.12×10^{-5}	1.76×10^{-14}	7.63×10^{-17}
Uterus effect	6.5(f)	5.26×10^{-1}	3.82×10^{-1}	9.44×10^{-5}	2.20×10^{-3}	2.00×10^{-17}	3.96×10^{-12}
Spermatogenesis testis	6.6(b)	1.59×10^{-1}	7.83×10^{-2}	3.50×10^{-2}	9.77×10^{-2}	5.53×10^{-2}	8.23×10^{-1}
Interstitial cells effect testis	6.6(d)	3.70×10^{-1}	1.92×10^{-1}	1.37×10^{-18}	1.02×10^{-19}	1.02×10^{-26}	7.90×10^{-27}
Prostate effect	6.6(f)	2.64×10^{-1}	7.33×10^{-1}	5.86×10^{-2}	6.82×10^{-2}	5.70×10^{-11}	3.01×10^{-8}

Table 6.6 presents the results of the *t*-test comparing the performance of ML models based on *in vitro* assay alone and either molecular descriptors only or molecular descriptors combined

with *in vitro* assays, when the SMOTE technique is applied. From Figures 6.4, 6.5, 6.6 and Table 6.6, we can see that, except for one outcome (spermatogenesis testis), the specificity is always significantly higher for models based on molecular descriptors, whether alone or combined with *in vitro* assays, than the one of models based on *in vitro* assays only. Conversely, sensitivity is generally significantly lower when molecular descriptors are used. According to what we previously noticed, this decrease in sensitivity is probably due to the fact that SMOTE significantly increases sensitivity of *in vitro* assay-based models but not the one of models based on molecular descriptors. This could be explained by the nature of the descriptor values which are continuous in the first case and mostly binary in the second case. Indeed, the original version of SMOTE has been implemented for continuous features and may perform worse when applied to a mix between continuous and binary features. In most cases, the models results in a BA not significantly different between the three types of models (it is significantly different for injury in adrenals and the two outcomes of ovary). Therefore, it seems that neither the type of descriptors used nor the SMOTE data augmentation technique contributes to an improvement of the models' performance. However, since the Figure and Tables presented here provide the performance only for the *model which reaches the best BA* but not the best sensitivity, we may observe higher sensitivity for other models but the BA would be lower (*e.g.* the best sensitivity obtained for steroidogenesis in adrenal glands is 0.15 but the corresponding BA equals 0.51). Similar results are obtained with the other algorithms (see Appendix E).

Finally, these results show that using chemical structure to predict the long-term *in vivo* effects is neither better nor worse than using the results of the selected *in vitro* assays.

6.5 Conclusion

In this chapter we evaluated the ability of 42 *in vitro* assays related to endocrine pathways to predict *in vivo* outcomes observed in rat long-term studies in three endocrine organs and two sex accessory organs, using data from ToxCast and ToxRefDB. Despite the small number of compounds (418) for which both *in vitro* and *in vivo* data are available, we were able to draw some conclusions.

First, we showed that there is no relationship between the 42 *in vitro* assays and the *in vivo* outcomes, even for assays that are specific for relevant pathways known to occur in the target organs. To determine if better results could be obtained by considering results from multiple assays, we also performed the analysis using the results from the published ER and AR computational models from US EPA. For the AR model, the use of a linear additive approach that considers several assays targeting the same pathway did not show more important relation with long-term *in vivo* outcomes. Regarding the ER model, we were not able to conclude because of the small number of positive compounds.

Second, using several ML algorithms, we demonstrated that the combination (not necessarily linear) of the 42 assays was not able to predict the *in vivo* outcomes. After applying a data augmentation technique to face the highly imbalance nature of the training datasets, performances

of the models were not improved.

Finally, the predictions from ML models built on the selected *in vitro* assays are not better than those derived from molecular descriptors alone. Moreover, a combination of both types of descriptors also did not improve performances.

In conclusion, this study highlights that the 42 selected *in vitro* assays do not provide information about the *in vivo* outcomes observed in endocrine and associated organs in rat long-term studies and raises the question of the utility of these *in vitro* assays for compounds' prioritization, in particular in the case of endocrine mediated effects. This kind of study should be extended to other *in vitro* assays and *in vivo* outcomes in order to explore if the same trends come out.

CONCLUSION

Outcomes of the work

The overall objective of this thesis was to evaluate how the public data generated for toxicity assessment could be exploited by machine learning to help in the prioritization of compounds, in line with the Tox21 vision initially proposed by the US NRC in 2007. This objective also fits with that of Bayer, which supported this work as part of a CIFRE agreement. Indeed, Bayer is developing computational approaches to perform compounds' selection in the early phases of the development of Plant Protection Products. Here, we focused on information provided by chemical structure of compounds, on *in vitro* bioactivity results and on *in vivo* studies performed in laboratory animals. Since the direct prediction of *in vivo* effects from the chemical structure was ambitious, we preferred to focus on two sub-tasks: (1) the prediction of *in vitro* bioactivity from the chemical structure and (2) the prediction of *in vivo* effects from the *in vitro* results. Then, if the two types of ML models had performed well, we would have proposed to chain them in a two-stage approach.

Available toxicological data

Before performing any computational study, data should be collected from various sources, curated and standardized to make them consistent for the purpose of the study. They also need to be harmonized for a relevant aggregation of several data sources. When considering the publicly available data, we highlighted various challenges regarding their use by computational methods such as uneasy access, heterogeneity within and between databases (regarding the format of results, the representation of compounds, *etc*) and a lack of harmonized ontology. Moreover, with regard to toxicological data, we could observe variability in term of quality and results, a high imbalanced property (a lot more negative compounds compared to the positive ones) and in general a small volume of data. All these challenges are known by the community [121] and initiatives have emerged to work on it such as ToxML [218] or OpenTox [255, 9]. Various organizations are also making progress in developing terminology and ontologies. For example, the US FDA

utilizes a standardized format called Standard for Exchange of Nonclinical Data (SEND) [50] and the Society of Toxicology Pathology and the European Society of Toxicology are leading the International Harmonization of Nomenclature and Diagnostic (INHAND) project [145], that aims at defining criteria for observed effects in rats and mice. In the pharmaceutical domain, such initiatives are more advanced and we can cite the Open PHACTS [122] and the eTOX [245] projects.

In our work, we decided to use the data released by the US EPA (DSSTox, ToxCast and ToxRefDB), in particular because they provide the three types of information we wanted to use and because it seems the least challenging and the most suitable resource for our purpose. Indeed, the three databases share the same compounds' identifiers facilitating their aggregation and provide almost "ready-to-use" data avoiding a lot of pre-processing to make them machine readable.

Moreover, by proposing an overview of the studies that have been performed so far regarding computational tools applied to toxicological data, with a specific interest in ML methods, we showed that there is no universal recommendation for the development of such models, probably due to the important variability of data (number of data points, number of positive compounds, data quality) and the high complexity of biological phenomena characterized in the different *in vitro* assays.

From chemical structure to *in vitro* bioactivity

Regarding the first task which aimed at predicting *in vitro* bioactivity based on chemical structure, we demonstrated in two studies that models' performances were highly depending on the assays but that it was still possible to build models reaching almost 70% of Balanced Accuracy (BA) for some assays. In particular, we showed that model's performance could be improved by the use of data augmentation techniques and by an increase of the number of observations used in the training set. Moreover, in a large scale study we developed models to predict risk associated with hundreds of ToxCast assays and we highlighted that the stacked generalization ensemble method was appropriate for the data used and could lead to models reaching an AUC ROC of more than 0.75. We could therefore think about using these models for the filtering of chemical structures. We also showed the importance of taking into account the applicability domain to estimate the reliability of predictions.

Since we built models for several assays without focusing on specific ones, observations and conclusions raised by the two studies are general. However, if ones want to build ML models for one or some specific assay(s), even if these observations can be used as a starting point, further work would be required to construct better models. For example, we could think of using other molecular descriptors, fine tuning hyper-parameters, testing other ensemble methods and data augmentation techniques, *etc.* In particular, we refer to the numerous published papers about the building of good QSAR models [61, 262, 48] which result from a long expertise in this domain. In any, case, it is important to have in mind that models' performance highly depends on the

quality of the input data.

More generally, we should highlight the growing interest of regulatory authorities in the use of QSAR models for toxicity prediction. Indeed, some QSAR models are already requested by authorities for risk assessment such as the ones that predict mutagenicity (in particular the Ames test) of pharmaceutical compounds [256] (ICH M7 guideline) and genotoxicity of PPPs [77] (EFSA guidance on dietary risk assessment). Moreover, EFSA has recently evaluated the applicability of existing *in silico* models that predict genotoxicity and gave recommendations for a Weight-of-Evidence approach that integrates *in silico* predictions, experimental results and expert knowledge to perform risk assessment [20]. In particular, they concluded that all QSAR models for the Ames test (mutagenicity) were resulting in significant predictions but not the models predicting other genotoxicity endpoints (*e.g.* chromosomal aberrations), which therefore should be improved. This specific case demonstrates that regulatory authorities are willing to move towards the proposed paradigm for a risk assessment based on alternative approaches and are encouraging the scientific community to develop reliable *in silico* methods. Another example is the case of acute toxicity: in 2018 a workshop has been organized by the ICCVAM acute toxicity workgroup in order to discuss the possibility of using acute toxicity predictive models instead of *in vivo* studies in regulatory purpose [151]. Indeed, the combined predictions of several *in silico* models have been shown to reach really good performance when compared to animal data. The workshop members therefore proposed recommendations and next steps for their integration into regulatory use.

From *in vitro* bioactivity to *in vivo* outcomes

With respect to the second task whose goal was to predict *in vivo* outcomes from *in vitro* assays, we highlighted that no relation could be observed between *in vitro* assays targeting pathways known to induce endocrine effects and *in vivo* effects observed in endocrine organs after rat long-term studies. This observation was first made for *in vitro* assays considered alone (one assay versus one effect). However, the use of ML methods that combine the results of several assays into predictive models of the *in vivo* effects also did not lead to good performance, since BA was around 0.5. These results can be explained by several reasons.

First, the *in vitro* assays that were used do not give information about the ADME properties of compounds and therefore their results do not reflect the dose dependencies that can be observed in *in vivo* context. However, these properties are critical to enable reliable *in vitro* - *in vivo* extrapolation (IVIVE) and therefore to obtain accurate *in vivo* predictions. This aspect was recently highlighted by Thomas [258] and Klaren *et al.* [148].

Besides, these assays do not represent the set of all possible biological pathways leading to adverse endocrine effects but only target some of them. Indeed, each assay only targets one specific biological event and it is clearly acknowledged that one assay alone will not help to predict toxicity since it will be only partially related to an *in vivo* effect [296]. Also, *in vitro* assays only mimic a small spectrum of the cellular and physiological processes taking place in complex whole

organisms such as mammals since they do not capture intercellular and inter-organ communications that actually happen in the whole organism [34]. This prevents from detecting *in vivo* responses that require multi-tissue interactions [7], which is specifically the case in endocrine mediated toxicity with the important role of the pituitary gland in the regulation of the endocrine function.

Also, assays selected in the ToxCast project were not originally designed to be predictive of specific long-term *in vivo* outcomes or toxicological modes of action and are probably not the most suitable for ML modeling [185]. Indeed, these assays were mostly selected based on their technical and economical feasibility and not according to their biological importance, or toxicity relevance [68].

Finally, we can also discuss the quality and relevance of *in vivo* data from ToxRefDB. Indeed, challenges in using this database in its initial release have already been discussed by Plunkett *et al.* [209]. Nonetheless, we should state that major improvements have been performed since the first release such that the ToxRefDB 2.0 version is one of the best examples of pretty well curated databases including standardized ontology [272].

Even if our observations are specific to endocrine toxicity and that the evaluation of the link between *in vitro* assays and *in vivo* effects should be expanded to other types of adverse outcomes, it is known that the previously mentioned limitations of *in vitro* assays are valid for most of HTS assays and need to be considered. The Tox21 consortium recently released a new strategic and operational plan that takes into account these key challenges in order to gain a scientific confidence in the *in vitro* assays [258]. In particular, they intend to address the lack of metabolic competence of *in vitro* assays, to develop alternative tests predictive of human dose-response and to deploy new methods and computational approaches to perform IVIVE and focus on assays targeting molecular events in high priority AOPs.

According to the results obtained for the two types of predictions, we can already anticipate that the two-stage approach proposed in the Introduction of this manuscript is not relevant. Indeed, if we chain two ML models for which BA performances are around 0.7 and 0.5, we do not expect a final good BA since prediction errors will propagate from one model to the other. More work is therefore required to improve the intermediate steps and generate good models before considering the entire approach.

Perspectives

The results presented in this manuscript along with the numerous drawbacks of *in vitro* assays highlighted the big challenge to predict *in vivo* outcomes based on *in vitro* data alone. Therefore, the consideration and the integration of other types of information in computational tools seems to be essential to face this challenge. In particular, as pointed in the Tox21 vision, the focus on mechanistic knowledge is crucial since it enables the understanding and characterization of

pathways leading to adverse outcomes in a systemic approach rather than the simple identification and the characterization of hazards caused by compounds. In a series of three papers recently published, Wolf, Doe, Cohen *et al.* [277, 69, 52] considered the specific case of carcinogenicity and suggested that the current evaluation of carcinogenic potential of compounds that is based on the identification of hazard alone (after long-term rodent studies) should be replaced because of its insufficient biological relevance. Indeed, they stated that the carcinogenic potency of a compound is the result of a multi-stage probabilistic process that depends on the dose level, on the duration of exposure and on toxicokinetic properties. Therefore, they proposed that the entire risk potential should be assessed using a rationale stepwise process which takes into account the AOP knowledge.

Here we propose two approaches that consider mechanistic knowledge. On the one hand, existing mechanistic knowledge could be potentially used to build good and relevant ML models to predict *in vivo* toxicity. On the other hand, ML can help in the discovery and elucidation of possible AOPs based on available data.

Towards an integration of mechanistic knowledge: a mechanistic-driven approach

In general, since the results of our work showed that there is no evident link between the 42 *in vitro* assays selected and *in vivo* endocrine mediated effects observed in rat long term studies, we would suggest being cautious when interpreting the meaning and relevance of *in vitro* assay results in general. In particular, we believe that compounds' prioritization should be based on the evaluation of the risk of toxicity pathways alterations that are causally associated with *in vivo* adverse outcomes rather than proposing a pure *in vivo* adverse outcomes prediction. Ideally, a **mechanistic-driven approach** should be conducted to help in the selection of appropriate *in vitro* assays specifically addressing the alteration of a given adverse outcome pathway [222]. Indeed, since one or few assays do not represent the complexity of the entire organism, several assays should be considered and carefully selected.

Our proposed mechanistic-driven approach is illustrated in Figure 1. Basically, considering an adverse outcome (AO) of interest for which an AOP or mode of action has been elucidated, we suggest to base the predictive ML models on the assays triggering the MIE and KEs of the AOP (**AO model**). Indeed, we first make the hypothesis that if a compound is active in all or part of the n assays targeting the AOP's events, it could induce the AO when the compound is administered to the laboratory animal. Thus, we think that a ML model that uses the results of the n *in vitro* assays as descriptors to predict the AO of interest would correctly inform on the potential of compounds to alter the AOP leading to this AO. Of course, these *in vitro* descriptors can also be combined with structural features if it can improve the performance of the final AO model. Nonetheless, this implies that the physiological relevance of the *in vitro* assays selected has been demonstrated. In general, assays that enable the measure of MIEs and KEs have been defined for validated AOPs and information are publicly available, for example in the AOP-Wiki. On the contrary, relevant assays would have to be developed and validated. It is for example

the case of developmental toxicity for which limited knowledge exists in term of mode of action, preventing the development of corresponding assays.

Moreover, the *in vitro* data can be used to build ML models for each of the *in vitro* assays targeting the MIE and KEs (respectively **MIE models** and **KE models**) based on compounds' structure. Thus, if a compound has not been tested in all the assays required by the ML model predicting the considered AO, the missing values could be predicted from its structure (with a error risk) rather than by performing the experimental test.

We think that the most important challenge for the construction of this type of ML model (whether AO model or MIE and KE models) will be the lack of data regarding results for the considered *in vitro* assays as well as the adverse outcome. Therefore, we suggest that HTS will have to be performed for the maximum number of chemicals, including all the ones for which *in vivo* data are available. Furthermore, one important limitation for the *in vivo* prediction concerns the lack of description of the bioavailability properties of compounds, mainly obtained from few *in vitro* assays and physico-chemical properties. In order to face this limitation and to also take into account the dose-effect relationship, we believe that **toxicokinetics** information (ADME properties) of compounds are necessary to complete the evaluation of the pathways altered by the compounds in *in vitro* assays. If not already available, this type of information can be obtained either by the use of *in vitro* assays or estimated by PBTK models, possibly coupled with a reverse dosimetry approach, or even by ML methods such as, recently proposed, multi-task deep learning [274]. Moreover, **toxicogenomics** data regarding genes that are involved in the considered AOP can also be integrated in this approach.

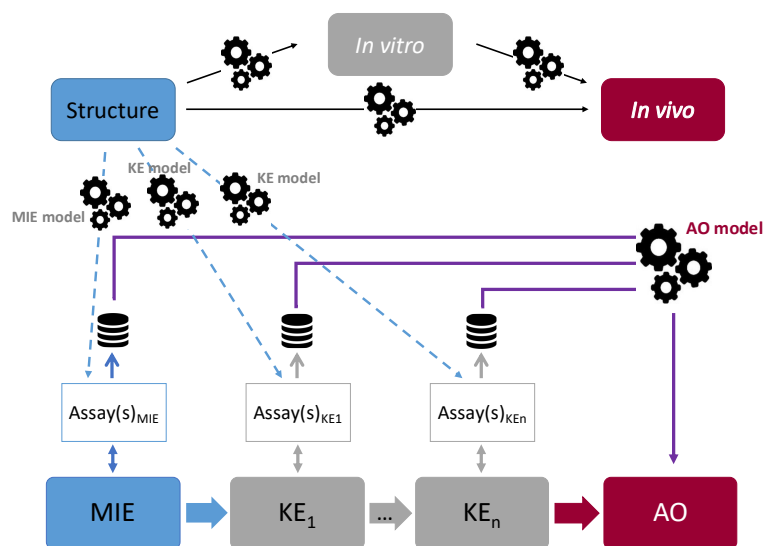


Figure 1: A mechanistic-driven approach. The upper part of the figure summarizes the ML models that we intended to build in the presented work. The lower part describes the proposed mechanistic-driven approach: starting with a known AOP (bottom line), *in vitro* assays targeting the MIE and KEs are identified. Results from these assays are used as descriptors of a ML model to predict the AO of interest (AO model). ML models can also be built to predict the *in vitro* assays constituting the AO model, based on the structure of compounds. MIE: Molecular Initiating Event, KE: Key event, AO: Adverse Outcome.

Finally, we can consider two use-cases to apply these ML models for the detection of the potential of a new compound to induce the AO of interest through an alteration of the considered AOP. These use-cases depend on the type of data available:

1. If the new compound has been tested in all the n *in vitro* assays used in the model: in that case, the AO model predicting the AO of interest can be directly used with the *in vitro* results as input descriptors, possibly combined with structural features;
2. If the new compound has not been tested at all or has been partially tested in some of the *in vitro* assays used in the model: in that case, missing bioactivities can be either measured by directly testing the compound in the corresponding *in vitro* assays or predicted if reliable MIE and KE models have been developed for the corresponding *in vitro* assays.

In the end, if sufficient data is available, we hope that this approach would result in *in silico* models with high predictive performance and meaningful mechanistic interpretation.

Machine Learning to help in the development of new AOPs: a data-driven approach

In the previous approach, we suggested to use mechanistic knowledge to help in the development of good ML models. On the contrary, the **data-driven approach** proposes to use existing data and ML methods to increase mechanistic knowledge and help in the elucidation of new AOPs. Indeed, as already done with genomics data for the reconstruction of gene regulation networks, we propose here to use *in vitro* data for the reconstruction of mechanistic networks.

The approach is illustrated in Figure 2. Basically, starting with a pool of *in vitro* assays for which enough data are available and considering an AO of interest, the idea is to build several ML models to predict the AO (step 1 of the figure). At the beginning, the input training set for modeling is composed of bioactivity descriptors corresponding to the results of all available *in vitro* assays. Contrary to what we did for the prediction of endocrine outcomes in Chapter 6, no "knowledge-based" pre-selection of the assays is performed. Then various models are built using several methods of feature selection (including filter, wrapper and embedded methods detailed in Chapter 2) in order to accurately look for the most descriptive *in vitro* assays. The construction of such ML models should be performed using an automatic and systematic workflow such that a large number of models can be built but only the ones reaching the desired performance are kept.

We then propose to interpret the *in vitro* assays that are used to build these good models by identifying their biological targets (step 2 of the figure). This interpretation should be based on a biological knowledge that enables a mapping between assays and biological events.

Finally, the idea is to find relationships between identified events (still based on biological knowledge) in order to extract mechanistic information and deduce new possible AOPs (step 3 of the figure). This step should be combined to systems biology approaches such as the development of toxicological networks to finally identify previously unknown compound-target interactions and toxicity mechanisms. Obviously, the resulting new hypothesis should be assessed and confirmed

or rejected thanks to biological experiments performed in "wet laboratory".

Here also we can think about integrating other types of information such as **ADME properties** and **toxicogenomics**. This is in particular what US EPA, in collaboration with other partners, intends to do within the Human Toxome project that aims at generating omics data to deduce toxicological modes of action [28].

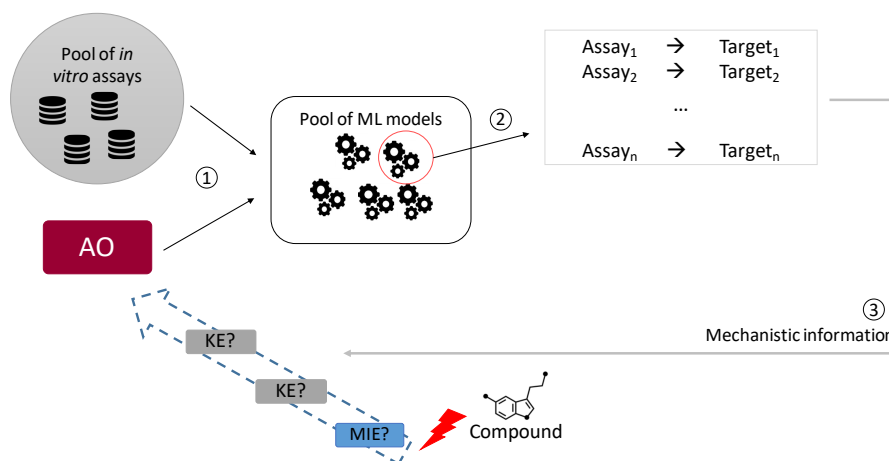


Figure 2: A data-driven approach to elucidate potential new AOP of a given adverse outcome. 1) ML models are built using data from all *in vitro* assay results available. 2) Assays that are used to build good models are interpreted and their biological targets are identified. 3) Mechanistic information are extracted to develop new AOPs.

This kind of approach that aims at identifying new mechanisms has already been proposed and some of them based on clustering methods have been reviewed in Chapter 3 [130, 241]. Moreover, Kim *et al.* [146] developed a workflow allowing the identification of assays relevant for the prediction of hepatotoxicity and finally providing new insights into the possible mechanisms and events that lead to this AO. Besides, as part of systems biology, toxicological networks are developed to identify interactions between chemicals, biological entities and outcomes based on available data [253].

In the end, we hope that this kind of data-driven approaches will enable the discovery of previously unknown mechanisms and here again, the most important limitation is the availability of sufficient data, both *in vitro* to broadly evaluate toxicity pathways and *in vivo*, necessary to build the physiological and phenotypic anchoring.

To conclude, in a purpose of early toxicity prediction, computational tools can be very useful to integrate the relevant *in vitro* biological activities induced by compounds. Nonetheless, the currently available technologies and *in vitro* assays that describe bioactivities of compounds associated with *in vivo* adverse outcomes only enable us to characterize a risk of altering a toxicity pathway or an AOP instead of a true *in vivo* adverse outcome. Moreover, *in silico* approaches cannot work alone and an important biological knowledge should be considered, regarding var-

ious types of information (bioactivity, toxicokinetics, mechanistic information), integrated by systems approaches.

We propose here two types of approaches that integrate these information and a recent paper reinforces our proposal since it suggests to consider the same kind of approaches (either mechanistic-driven or data-driven) [51]. Since our two approaches should be used for specific AO of interest, we would suggest to first apply them to well known AOs such as hepatotoxicity in order to provide a first proof of concept. Then, if the mechanistic-driven approach turns out to result in good ML models and the data-driven approach retrieves known AOPs and mechanistic events, they could be extended and generalized to other AOs in order to cover the largest mechanistic knowledge.

Last but not least, we hope that this kind of integrative approach could help in making the new vision of Tox21 a reality !

APPENDICES

APPENDIX A

TOXICOLOGICAL DATA RESOURCES DESCRIPTION

A.1 Chemical structures

PubChem: This database contains the structure of more than 97 million of compounds along with their physico-chemical properties and other various information. Since PubChem also contains *in vitro* data, it is further described below.

DSSTox: The Distributed Structure-Searchable Toxicity database is a US EPA resource [216] that maps chemical structures of compounds with their corresponding physico-chemical property and toxicity data (*in vitro* bioassays) into SDF files (see Section i)). Today it includes data for over 700,000 compounds. It is also included in the ACToR system (see below).

A.2 *In vitro* data

Here are listed resources that mostly provide data obtained in *in vitro* experiments but some of them also include *in vivo* and / or genomics data.

PubChem: Released in 2004, this public database provides *in vitro* activity results as well as chemical structure of compounds [268, 269]. It is maintained by the National Center for Biotechnology Information (NCBI) from the NIH and is freely available. Today, the database contains more than 97 million of compound structure entries from all categories of small chemicals and includes results for more than 1 million of biological assays (not necessarily related to toxicity). These results concern around 3 million of compounds corresponding to almost 240 million of bioactivities.

ChEMBL: This public database is maintained by the European Bioinformatics Institute (EBI) since 2009 and gathers curated literature data for bioassay results regarding drug-like compounds [21]. Today, more than 15 million of bioactivities are available for almost 2 million of distinct compounds. Some of PubChem bioassays are included in ChEMBL.

CEBS: It stands for Chemical Effects in Biological Systems and is a public database developed by the NIEHS to gather data generated by the NTP [271, 161] but also from other depositors from academic, industrial and governmental laboratories. It also includes *in vivo* data from various types of studies as well as toxicogenomics data.

EADB: The Estrogenic Activity Database has been developed by the US Food and Drug Administration to gather estrogenic activity *in vitro* and *in vivo* data from the public domain [237]. On the one hand, it stores results of more than 1,200 *in vitro* assays mostly performed in human and rat cell lines for more than 8,000 compounds. On the other hand, it contains *in vivo* data from two short term assays that measure the effect of compounds on uterine weight and called uterotrophic assays. Nonetheless, the EADB has been populated only when it was created in 2012 and has not been updated afterwards. The results of these assays are also available in the FDA’s Endocrine Disruptors Knowledge Database (EKDB) [67].

ACToR: The Aggregated Computational Toxicology Resource is the EPA’s warehouse that includes data from various types and sources (more than a thousand) and over 500,000 compounds [140]. In particular it includes HTS data (ToxCast), *in vitro* and *in vivo* data from the EDSP program [88], *in vivo* data from toxicity studies (ToxRefDB), exposure data (ExpoCast), *etc.*

A.3 *In vivo* data

Some of the sources mentioned in the previous section contain *in vivo* data along with *in vitro* ones but here is a broader list of sources providing only *in vivo* data.

HESS DB: HESS stands for Hazard Evaluation Support System and is an integrated platform of the Japanese National Institute of Technology and Evaluation (NITE) that provides two databases, released in 2012 [229]. The first gathers information on *in vivo* toxicity studies and toxicity mechanisms (the repeated dose toxicity database) and the second one is a metabolism knowledge database containing information on ADME properties of compounds measured in rat and humans. More than 500 *in vivo* studies are available in the database for which around 500 effects are evaluated.

RepDose: This database provides NOAELs and LOAELs from *in vivo* repeated dose toxicity studies of several duration (short term to chronic) performed in rat, mouse and dog [24]. It is maintained by the Fraunhofer Institute for Toxicology and Experimental Medicine (Germany) and contains about 3,100 studies corresponding to around 930 compounds.

FedTEX: The Fertility and Developmental Toxicity in Experimental animals database has also been developed by the Fraunhofer Institute to gather reproductive and developmental studies performed in rodent and rabbit for almost 300 compounds [23].

CPDB: The Carcinogenic Potency Database has been developed in the University of California, Berkeley and gathers the data from chronic and long-term cancer *in vivo* studies from the literature and the NTP [91]. A total of 1,547 compounds have data but the database has not been updated after 2001. Among the provided data is the TD_{50} which corresponds to the dose inducing tumors in half of the tested animals. The database is currently available through TOXNET¹ (Toxicology Data Network) which is an online collection of toxicology databases held by the NIH.

COSMOS DB: This database has been developed in the context of the SEURAT European initiative [103]. The goal of the COSMOS project was to develop *in silico* models to predict human repeated dose toxicity of cosmetic compounds [283]. To do so, a database has been released containing more than 40,000 compound structures and more than 12,000 *in vivo* toxicity studies for around 1,600 cosmetic compounds.

eTOX: eTOX is a consortium started in 2010 and funded by the European Innovative Medicines Initiative [246]. It gathers partners from academia, pharmaceutical industry and biotechnology companies in order to develop *in silico* tools to predict the toxicity of pharmaceutical compounds in the early stage of their development. To do so, they established a database where *in vivo* toxicological data obtained in the pharmaceutical companies (including Bayer) are shared. The database also includes public data from the RepDose database. The project ended in 2016 and the database contains more than 1,900 compounds associated to around 8,000 studies which are accessible to the consortium partners only [245].

A.4 Genomics data

TG-GATEs: The Open Toxicogenomics Project-Genomics Assisted Toxicity Evaluation Systems (TG-GATEs) is a database storing toxicogenomics data obtained during the Japanese Toxicogenomics Project (TGP) [132, 263]. The database counts data for 170 pharmaceutical compounds that have been tested either in human and rat hepatocytes (*in vitro*) or in rat (*in vivo*). *In vivo* data from the traditional toxicity studies are also available in the database.

CTD: The Comparative Toxicogenomics Database has been launched in 2004 to provide data describing relationships between chemicals, genes, proteins and diseases in order to enable the understanding about chemical exposures and human health [182]. It is regularly updated with data from the literature and currently contains curated data for more than 15,000 compounds [109]. This resource also provides tools for data analysis and is available through TOXNET.

CMap: The Connectivity Map has been created by the Broad Institute of MIT and Harvard and collects more than 7,000 gene expression profiles obtained from human cells treated with

¹<https://toxnet.nlm.nih.gov/>

more than 1,300 molecules [160]. The most recent version, dating from 2017, provides results for the L1000 assay that measures transcripts of 978 "landmark" human genes and corresponding to more than 1 million profiles [248]. The Broad Institute also provides data analysis tools accessible in a web platform².

A.5 Mechanistic data

AOP-KB: The AOP wiki³ is a collaborative database which gathers information regarding AOPs. It is part of the AOP-Knowledgebase (AOP-KB) repository launched by the OECD to enable the scientific community to share knowledge related to AOPs [275].

²<https://clue.io>

³<https://aopwiki.org>

APPENDIX B

LIST OF THE 37 ASSAYS USED IN CHAPTER 4

Table B.1: List of the 37 assays used in Chapter 4 with the repartition of positive and negative compounds in the corresponding *in vivo* constrained and extended datasets

Assay index	Assay name	Molecular target	In-vivo constrained datasets (404 compounds)			Extended datasets (7691 compounds)				
			Negative	Positive	% positives	Total	Negative	Positive	% positives	Total
1	TOX21_AR_BLA_Antagonist_ch2	Androgen Receptor	377	27	6.68	404	7499	192	2.5	7691
2	TOX21_AR_BLA_Antagonist_ratio	Androgen Receptor	292	112	27.72	404	6560	1131	14.71	7691
3	TOX21_AR_BLA_Antagonist_viability	Androgen Receptor	343	61	15.1	404	7125	566	7.36	7691
4	TOX21_AR_LUC_MDAKB2_Antagonist	Androgen Receptor	308	96	23.76	404	6878	813	10.57	7691
5	TOX21_AhR_LUC_Agonist	Aryl hydrocarbon Receptor	339	65	16.09	404	6888	803	10.44	7691
6	TOX21_Aromatase_Inhibition	Aromatase enzyme	286	118	29.21	404	6463	1228	15.97	7691
7	TOX21_ERa_BLA_Agonist_ch2	Estrogen Receptor	378	26	6.44	404	7248	443	5.76	7691
8	TOX21_ERa_BLA_Antagonist_ratio	Estrogen Receptor	319	85	21.04	404	6701	990	12.87	7691
9	TOX21_ERa_LUC_BG1_Agonist	Estrogen Receptor	335	69	17.08	404	6455	1236	16.07	7691
10	TOX21_ERa_LUC_BG1_Antagonist	Estrogen Receptor	332	72	17.82	404	7012	679	8.83	7691
11	TOX21_GR_BLA_Agonist_ch1	Glucocorticoid Receptor	384	20	4.95	404	7380	311	4.04	7691
12	TOX21_GR_BLA_Agonist_ratio	Glucocorticoid Receptor	375	29	7.18	404	7215	476	6.19	7691
13	TOX21_GR_BLA_Antagonist_ch2	Glucocorticoid Receptor	367	37	9.16	404	7184	507	6.59	7691
14	TOX21_GR_BLA_Antagonist_ratio	Glucocorticoid Receptor	372	32	7.92	404	7266	425	5.53	7691
15	TOX21_GR_BLA_Antagonist_viability	Glucocorticoid Receptor	379	25	6.19	404	7319	372	4.84	7691
16	TOX21_PPARg_BLA_Agonist_ratio	Peroxisome Proliferator-Activated Receptor Gamma	380	24	5.94	404	7364	327	4.25	7691
17	TOX21_TR_LUC_GH3_Antagonist	Tyrosine Receptor	280	124	30.69	404	5990	1701	22.12	7691
18	TOX21_p53_BLA_p1_ch1	p53 transcription factor	366	38	9.41	404	7155	536	6.97	7691
19	TOX21_p53_BLA_p1_ch2	p53 transcription factor	380	24	5.94	404	7237	454	5.9	7691
20	TOX21_p53_BLA_p1_ratio	p53 transcription factor	357	47	11.63	404	6917	774	10.06	7691
21	TOX21_p53_BLA_p1_viability	p53 transcription factor	381	23	5.69	404	7330	361	4.69	7691
22	TOX21_p53_BLA_p2_ch1	p53 transcription factor	378	26	6.44	404	7198	493	6.41	7691
23	TOX21_p53_BLA_p2_ch2	p53 transcription factor	352	52	12.87	404	6915	776	10.09	7691
24	TOX21_p53_BLA_p2_ratio	p53 transcription factor	331	73	18.07	404	6706	985	12.81	7691
25	TOX21_p53_BLA_p2_viability	p53 transcription factor	360	44	10.89	404	6941	750	9.75	7691
26	TOX21_p53_BLA_p3_ch1	p53 transcription factor	362	42	10.4	404	7049	642	8.35	7691
27	TOX21_p53_BLA_p3_ch2	p53 transcription factor	362	42	10.4	404	7002	689	8.96	7691
28	TOX21_p53_BLA_p3_ratio	p53 transcription factor	345	59	14.6	404	6772	919	11.95	7691
29	TOX21_p53_BLA_p3_viability	p53 transcription factor	376	28	6.93	404	7294	397	5.16	7691
30	TOX21_p53_BLA_p4_ch1	p53 transcription factor	359	45	11.14	404	7075	616	8.01	7691
31	TOX21_p53_BLA_p4_ch2	p53 transcription factor	370	34	8.42	404	7135	556	7.23	7691
32	TOX21_p53_BLA_p4_ratio	p53 transcription factor	336	68	16.83	404	6783	908	11.81	7691
33	TOX21_p53_BLA_p4_viability	p53 transcription factor	366	38	9.41	404	7065	626	8.14	7691
34	TOX21_p53_BLA_p5_ch1	p53 transcription factor	380	24	5.94	404	7365	326	4.24	7691
35	TOX21_p53_BLA_p5_ch2	p53 transcription factor	360	44	10.89	404	7067	624	8.11	7691
36	TOX21_p53_BLA_p5_ratio	p53 transcription factor	346	58	14.36	404	6845	846	11	7691
37	TOX21_p53_BLA_p5_viability	p53 transcription factor	381	23	5.69	404	7392	299	3.89	7691

APPENDIX C

LIST OF THE 42 SELECTED ASSAYS IN CHAPTER 6

Table C.1: List of the selected 42 assays that are related to E, A and S endocrine pathways with the pathway they are linked to and their type. E = estrogen, A = androgen, S = steroidogenesis, O = others

Assay name in ToxCast	Pathway	Type of assay
ACEA_T47D_80hr_Positive	E	Cell proliferation
ATG_ERE_CIS_up	E	mRNA induction
ATG_ERa_TRANS_up	E	mRNA induction
OT_ER_EraERb_0480	E	Protein complementation
OT_ER_EraERb_1440	E	Protein complementation
OT_ER_ErbERb_0480	E	Protein complementation
OT_ER_ErbERb_1440	E	Protein complementation
OT_Era_EREgFP_0120	E	Reporter gene
OT_Era_EREgFP_0480	E	Reporter gene
TOX21_Era_BLA_Antagonist_ratio	E	Reporter gene
TOX21_Era_LUC_BG1_Agonist	E	Reporter gene
TOX21_Era_LUC_BG1_Antagonist	E	Reporter gene
NVS_NR_cAR	A	Receptor binding
NVS_NR_hAR	A	Receptor binding
NVS_NR_rAR	A	Receptor binding
OT_AR_ARELUC_AG_1440	A	Reporter gene
OT_AR_ARSRC1_0480	A	Coregulator recruitment
OT_AR_ARSRC1_0960	A	Coregulator recruitment
TOX21_AR_BLA_Antagonist_ratio	A	Reporter gene
TOX21_AR_LUC_MDAKB2_Antagonist	A	Reporter gene
TOX21_AR_LUC_MDAKB2_Antagonist2	A	Reporter gene
CEETOX_H295R_11DCORT_dn	S	Hormone measurement
CEETOX_H295R_ANDR_dn	S	Hormone measurement
CEETOX_H295R_CORTISOL_dn	S	Hormone measurement
CEETOX_H295R_DOC_dn	S	Hormone measurement
CEETOX_H295R ESTRADIOL_up	S	Hormone measurement
CEETOX_H295R ESTRONE_dn	S	Hormone measurement
CEETOX_H295R ESTRONE_up	S	Hormone measurement
CEETOX_H295R_OHPROG_dn	S	Hormone measurement
CEETOX_H295R_OHPROG_up	S	Hormone measurement
CEETOX_H295R_PROG_up	S	Hormone measurement
CEETOX_H295R_TESTO_dn	S	Hormone measurement
NVS_ADME_hCYP19A1	S	Enzyme activity
TOX21_Aromatase_Inhibition	S	Enzyme inhibition
ATG_Sp1_CIS_up	O	mRNA induction
ATG_GRE_CIS_dn	O	mRNA induction
ATG_SREBP_CIS_up	O	mRNA induction
NVS_NR_bPR	O	Receptor binding
NVS_NR_hGR	O	Receptor binding
NVS_NR_hPR	O	Receptor binding
TOX21_GR_BLA_Agonist_ratio	O	Reporter gene
TOX21_GR_BLA_Antagonist_ratio	O	Reporter gene

APPENDIX D

LIST OF IN VIVO EFFECTS FROM TOXREFDB
GROUPED INTO OUTCOMES CATEGORIES

Table D.1: List of observed effects from ToxRefDB that are grouped in each of the 9 category outcomes.

Outcome category	Effects
Adrenals – Steroidogenesis effects	Organ weight, Vacuolization, Vacuolization cytoplasmic, Fatty change, Lipidosis
Adrenals – Stimulation	Organ weight, Hyperplasia, Hypertrophy, Atrophy, Adenoma, Adenoma / carcinoma combined
Adrenals – Injury	Organ weight, Angiectasis, Cyst, Cytologic alterations, Degeneration, Hemorrhage, Hyperemia, Inflammation, Pigmentation, Necrosis
Ovary – Effects on germinal cells	Organ weight, Atrophy, Follicle count, Cyst (follicle), Hypertrophy
Ovary - Effects on interstitial cells	Organ weight, Hyperplasia, Pigmentation, Lipidosis, Stromal hyperplasia, Vacuolization
Uterus	Organ weight, Adenocarcinoma, Atrophy, Carcinoma, Dilatation, Horn distended, Endometriosis, Hydrometra, Hyperplasia, Hypertrophy, Mass, Metaplasia, Nodule(s), Polyp (all sorts),
Testis – Spermatogenesis	Organ weight, Aspermatogenesis, Atrophy, Degeneration, Delayed spermiation, Retained spermatids, Necrosis, Reduced spermatogenesis, Spermatogenic arrest
Testis - Effects on interstitial cells	Organ weight, Adenoma, Interstitial cell tumor benign, Interstitial cell tumor NOS, Hyperplasia, Hypertrophy
Prostate	Organ weight, Adenoma, Atrophy, Hyperplasia

APPENDIX E

PERFORMANCE OF THE ML MODELS BUILT IN CHAPTER 6

Tables E.1 to E.6 summarize the balanced accuracy, sensitivity and specificity obtained with the 8 types of algorithms and the three types of descriptors (bio = *in vitro* assays, chem = molecular structure, cb = combination of both) to predict *in vivo* outcomes observed after rat long-term studies in adrenal glands, ovaries, uterus, testes and prostate, using either imbalanced datasets (without SMOTE method) or balanced datasets (with SMOTE method). For each triplet of outcome, algorithm and descriptor type, the reported results correspond to the model that reaches the highest BA among all with its corresponding sensitivity and specificity as well as the number of descriptors used. Metrics values correspond to the mean of the 20 runs and values in parenthesis to the standard deviations.

Table E.1: Performance of ML models that predict in vivo outcomes in adrenal glands on imbalanced datasets (without SMOTE)

In vivo outcome	ML algorithm	Number of descriptors			Balanced Accuracy			Sensitivity			Specificity		
		bio	chem	cb	bio	chem	cb	bio	chem	cb	bio	chem	cb
Steroidogenesis adrenal glands	CART	20.0	41.0	15.0	0.52 (0.07)	0.55 (0.11)	0.54 (0.10)	0.08 (0.13)	0.23 (0.22)	0.24 (0.20)	0.96 (0.06)	0.88 (0.09)	0.88 (0.09)
	ENSMB	18.0	31.0	20.0	0.50 (0.02)	0.51 (0.05)	0.51 (0.05)	0.01 (0.04)	0.05 (0.10)	0.04 (0.10)	1.00 (0.02)	0.99 (0.03)	0.99 (0.04)
	KNN1	12.0	58.0	35.0	0.51 (0.05)	0.51 (0.07)	0.52 (0.08)	0.05 (0.10)	0.09 (0.13)	0.10 (0.15)	0.98 (0.09)	0.96 (0.08)	0.96 (0.10)
	LDA	11.0	20.0	16.0	0.50 (0.02)	0.53 (0.09)	0.52 (0.09)	0.01 (0.05)	0.16 (0.17)	0.18 (0.17)	1.00 (0.02)	0.98 (0.07)	0.98 (0.08)
	NB	12.0	68.0	43.0	0.50 (0.07)	0.52 (0.13)	0.53 (0.13)	0.06 (0.14)	0.47 (0.30)	0.46 (0.30)	0.99 (0.07)	0.85 (0.30)	0.85 (0.30)
	RF	38.0	66.0	67.0	0.52 (0.06)	0.52 (0.08)	0.52 (0.07)	0.08 (0.12)	0.12 (0.15)	0.10 (0.13)	0.96 (0.05)	0.99 (0.06)	0.99 (0.06)
	SVCL	10.0	13.0	12.0	0.50 (0.00)	0.51 (0.05)	0.51 (0.06)	0.00 (0.00)	0.05 (0.10)	0.06 (0.10)	1.00 (0.00)	1.00 (0.05)	1.00 (0.05)
	SVCR	19.0	64.0	40.0	0.50 (0.02)	0.52 (0.05)	0.51 (0.05)	0.01 (0.04)	0.05 (0.10)	0.05 (0.10)	1.00 (0.02)	1.00 (0.04)	1.00 (0.03)
Stimulation adrenal glands	CART	37.0	47.0	14.0	0.52 (0.06)	0.53 (0.10)	0.54 (0.10)	0.09 (0.11)	0.22 (0.19)	0.24 (0.19)	0.96 (0.05)	0.87 (0.09)	0.87 (0.09)
	ENSMB	12.0	11.0	69.0	0.50 (0.02)	0.50 (0.04)	0.51 (0.04)	0.01 (0.04)	0.03 (0.07)	0.03 (0.07)	1.00 (0.02)	0.99 (0.04)	0.99 (0.04)
	KNN1	17.0	67.0	64.0	0.50 (0.05)	0.50 (0.07)	0.51 (0.07)	0.05 (0.11)	0.10 (0.13)	0.14 (0.15)	0.97 (0.10)	0.94 (0.07)	0.93 (0.10)
	LDA	13.0	29.0	16.0	0.50 (0.02)	0.51 (0.08)	0.51 (0.08)	0.01 (0.04)	0.14 (0.14)	0.15 (0.14)	1.00 (0.02)	0.98 (0.07)	0.98 (0.08)
	NB	17.0	44.0	48.0	0.51 (0.07)	0.52 (0.10)	0.52 (0.11)	0.09 (0.13)	0.38 (0.27)	0.38 (0.27)	0.98 (0.11)	0.85 (0.24)	0.86 (0.24)
	RF	37.0	57.0	64.0	0.51 (0.06)	0.51 (0.07)	0.51 (0.07)	0.07 (0.11)	0.11 (0.12)	0.11 (0.13)	0.95 (0.05)	0.97 (0.07)	0.98 (0.07)
	SVCL	10.0	11.0	25.0	0.50 (0.00)	0.50 (0.05)	0.50 (0.05)	0.00 (0.00)	0.05 (0.09)	0.06 (0.10)	1.00 (0.00)	1.00 (0.04)	1.00 (0.05)
	SVCR	10.0	64.0	69.0	0.50 (0.02)	0.50 (0.04)	0.51 (0.04)	0.01 (0.04)	0.03 (0.06)	0.04 (0.08)	1.00 (0.02)	1.00 (0.04)	1.00 (0.04)
Injury adrenal glands	CART	26.0	26.0	37.0	0.52 (0.07)	0.56 (0.12)	0.56 (0.12)	0.09 (0.14)	0.26 (0.22)	0.25 (0.22)	0.97 (0.06)	0.89 (0.08)	0.89 (0.08)
	ENSMB	33.0	10.0	10.0	0.50 (0.03)	0.51 (0.05)	0.52 (0.06)	0.01 (0.05)	0.05 (0.10)	0.07 (0.11)	1.00 (0.02)	0.99 (0.03)	0.99 (0.03)
	KNN1	26.0	57.0	38.0	0.51 (0.06)	0.51 (0.08)	0.52 (0.08)	0.07 (0.14)	0.10 (0.17)	0.10 (0.17)	0.99 (0.14)	0.95 (0.07)	0.96 (0.09)
	LDA	33.0	14.0	14.0	0.50 (0.02)	0.53 (0.09)	0.53 (0.09)	0.00 (0.04)	0.17 (0.18)	0.16 (0.18)	1.00 (0.02)	0.99 (0.07)	0.99 (0.06)
	NB	20.0	52.0	32.0	0.52 (0.07)	0.53 (0.13)	0.53 (0.12)	0.51 (0.49)	0.50 (0.34)	0.57 (0.36)	0.99 (0.46)	0.87 (0.34)	0.81 (0.31)
	RF	22.0	67.0	57.0	0.52 (0.07)	0.53 (0.08)	0.54 (0.08)	0.09 (0.14)	0.12 (0.16)	0.13 (0.18)	0.96 (0.05)	0.99 (0.05)	0.99 (0.05)
	SVCL	10.0	10.0	11.0	0.50 (0.00)	0.51 (0.06)	0.52 (0.06)	0.00 (0.00)	0.06 (0.11)	0.07 (0.12)	1.00 (0.00)	1.00 (0.05)	1.00 (0.05)
	SVCR	20.0	14.0	13.0	0.50 (0.01)	0.52 (0.05)	0.52 (0.05)	0.00 (0.03)	0.05 (0.10)	0.05 (0.09)	1.00 (0.02)	1.00 (0.03)	1.00 (0.03)

Table E.2: Performance of ML models that predict in vivo outcomes in adrenal glands on balanced datasets (with SMOTE)

In vivo outcome	ML algorithm	Number of descriptors			Balanced Accuracy			Sensitivity			Specificity		
		bio	chem	cb	bio	chem	cb	bio	chem	cb	bio	chem	cb
Steroidogenesis adrenal glands	CART	10.0	61.0	18.0	0.52 (0.09)	0.52 (0.10)	0.53 (0.11)	0.17 (0.32)	0.19 (0.19)	0.25 (0.21)	0.87 (0.26)	0.85 (0.08)	0.82 (0.08)
	ENSMB	17.0	65.0	64.0	0.51 (0.11)	0.52 (0.11)	0.52 (0.11)	0.44 (0.30)	0.26 (0.21)	0.25 (0.19)	0.57 (0.26)	0.78 (0.09)	0.80 (0.09)
	KNN1	38.0	52.0	69.0	0.50 (0.11)	0.52 (0.13)	0.52 (0.13)	0.37 (0.29)	0.35 (0.23)	0.33 (0.23)	0.63 (0.30)	0.68 (0.10)	0.72 (0.10)
	LDA	17.0	62.0	64.0	0.51 (0.11)	0.52 (0.12)	0.53 (0.13)	0.60 (0.31)	0.42 (0.23)	0.43 (0.24)	0.41 (0.26)	0.63 (0.11)	0.64 (0.10)
	NB	25.0	49.0	64.0	0.52 (0.11)	0.49 (0.13)	0.50 (0.13)	0.75 (0.24)	0.58 (0.25)	0.57 (0.25)	0.28 (0.18)	0.41 (0.15)	0.43 (0.16)
	RF	41.0	16.0	13.0	0.52 (0.09)	0.52 (0.09)	0.52 (0.09)	0.48 (0.41)	0.10 (0.17)	0.09 (0.17)	0.55 (0.36)	0.95 (0.07)	0.96 (0.06)
	SVCL	10.0	61.0	64.0	0.52 (0.11)	0.52 (0.13)	0.53 (0.14)	0.54 (0.34)	0.41 (0.23)	0.42 (0.25)	0.51 (0.31)	0.63 (0.11)	0.64 (0.10)
	SVCR	35.0	66.0	61.0	0.51 (0.10)	0.53 (0.11)	0.52 (0.10)	0.48 (0.26)	0.30 (0.21)	0.24 (0.23)	0.54 (0.21)	0.75 (0.11)	0.80 (0.11)
Stimulation adrenal glands	CART	14.0	47.0	63.0	0.50 (0.09)	0.51 (0.10)	0.51 (0.10)	0.25 (0.19)	0.21 (0.18)	0.18 (0.18)	0.76 (0.19)	0.81 (0.09)	0.84 (0.09)
	ENSMB	11.0	64.0	63.0	0.50 (0.10)	0.49 (0.10)	0.50 (0.10)	0.42 (0.28)	0.23 (0.18)	0.23 (0.18)	0.58 (0.28)	0.76 (0.10)	0.77 (0.09)
	KNN1	33.0	48.0	57.0	0.50 (0.09)	0.50 (0.12)	0.50 (0.11)	0.45 (0.34)	0.33 (0.21)	0.35 (0.21)	0.54 (0.32)	0.66 (0.11)	0.65 (0.10)
	LDA	11.0	57.0	63.0	0.50 (0.10)	0.50 (0.12)	0.51 (0.12)	0.54 (0.32)	0.40 (0.22)	0.41 (0.21)	0.46 (0.29)	0.60 (0.10)	0.62 (0.11)
	NB	14.0	54.0	63.0	0.50 (0.10)	0.47 (0.12)	0.48 (0.12)	0.70 (0.25)	0.51 (0.23)	0.51 (0.23)	0.30 (0.21)	0.43 (0.15)	0.45 (0.15)
	RF	13.0	61.0	63.0	0.51 (0.09)	0.50 (0.08)	0.50 (0.08)	0.20 (0.30)	0.12 (0.15)	0.13 (0.14)	0.81 (0.28)	0.87 (0.08)	0.87 (0.07)
	SVCL	11.0	60.0	68.0	0.51 (0.10)	0.51 (0.12)	0.51 (0.12)	0.55 (0.35)	0.41 (0.21)	0.39 (0.21)	0.46 (0.32)	0.60 (0.13)	0.63 (0.12)
	SVCR	14.0	58.0	63.0	0.50 (0.09)	0.51 (0.11)	0.50 (0.11)	0.34 (0.30)	0.31 (0.20)	0.25 (0.22)	0.65 (0.29)	0.70 (0.12)	0.75 (0.14)
Injury adrenal glands	CART	29.0	30.0	23.0	0.54 (0.10)	0.55 (0.11)	0.54 (0.11)	0.47 (0.25)	0.25 (0.21)	0.26 (0.20)	0.62 (0.21)	0.85 (0.08)	0.83 (0.08)
	ENSMB	16.0	31.0	19.0	0.54 (0.11)	0.54 (0.11)	0.54 (0.11)	0.53 (0.32)	0.26 (0.21)	0.24 (0.21)	0.54 (0.29)	0.82 (0.09)	0.83 (0.08)
	KNN1	16.0	27.0	19.0	0.50 (0.11)	0.54 (0.13)	0.54 (0.13)	0.35 (0.32)	0.41 (0.23)	0.42 (0.25)	0.66 (0.29)	0.67 (0.11)	0.66 (0.10)
	LDA	12.0	68.0	23.0	0.54 (0.11)	0.54 (0.13)	0.54 (0.13)	0.63 (0.35)	0.44 (0.25)	0.40 (0.24)	0.44 (0.27)	0.64 (0.10)	0.67 (0.10)
	NB	41.0	68.0	65.0	0.53 (0.10)	0.50 (0.14)	0.50 (0.13)	0.89 (0.25)	0.57 (0.29)	0.57 (0.28)	0.17 (0.19)	0.43 (0.17)	0.43 (0.17)
	RF	29.0	25.0	52.0	0.55 (0.10)	0.53 (0.09)	0.53 (0.09)	0.49 (0.25)	0.12 (0.17)	0.14 (0.15)	0.61 (0.20)	0.94 (0.05)	0.92 (0.06)
	SVCL	16.0	68.0	26.0	0.54 (0.10)	0.54 (0.13)	0.54 (0.13)	0.70 (0.35)	0.43 (0.26)	0.44 (0.23)	0.39 (0.30)	0.65 (0.10)	0.65 (0.12)
	SVCR	16.0	69.0	52.0	0.54 (0.10)	0.54 (0.12)	0.53 (0.11)	0.42 (0.36)	0.32 (0.25)	0.19 (0.24)	0.66 (0.29)	0.76 (0.14)	0.87 (0.12)

Table E.3: Performance of ML models that predict *in vivo* outcomes in ovaries and uterus on imbalanced datasets (without SMOTE)

<i>In vivo</i> outcome	ML algorithm	Number of descriptors			Balanced Accuracy			Sensitivity			Specificity		
		bio	chem	cb	bio	chem	cb	bio	chem	cb	bio	chem	cb
Germinal cells ovary	CART	15.0	22.0	13.0	0.53 (0.10)	0.57 (0.14)	0.57 (0.14)	0.10 (0.20)	0.21 (0.27)	0.21 (0.28)	0.98 (0.04)	0.95 (0.06)	0.96 (0.05)
	ENSMB	16.0	12.0	12.0	0.52 (0.07)	0.55 (0.10)	0.55 (0.10)	0.05 (0.15)	0.11 (0.20)	0.10 (0.19)	1.00 (0.01)	1.00 (0.02)	1.00 (0.02)
	KNN1	21.0	19.0	14.0	0.53 (0.08)	0.56 (0.11)	0.55 (0.11)	0.07 (0.18)	0.13 (0.21)	0.12 (0.21)	1.00 (0.08)	0.99 (0.07)	0.99 (0.09)
	LDA	10.0	12.0	10.0	0.52 (0.06)	0.57 (0.13)	0.57 (0.13)	0.04 (0.12)	0.20 (0.27)	0.22 (0.27)	1.00 (0.02)	0.99 (0.05)	0.99 (0.05)
	NB	11.0	19.0	27.0	0.54 (0.11)	0.58 (0.17)	0.56 (0.17)	0.26 (0.37)	0.88 (0.44)	0.88 (0.43)	0.97 (0.34)	0.49 (0.40)	0.46 (0.39)
	RF	10.0	16.0	54.0	0.53 (0.09)	0.55 (0.10)	0.55 (0.10)	0.08 (0.19)	0.10 (0.20)	0.11 (0.21)	0.98 (0.03)	1.00 (0.03)	1.00 (0.03)
	SVCL	10.0	10.0	10.0	0.50 (0.00)	0.53 (0.10)	0.54 (0.11)	0.00 (0.00)	0.10 (0.19)	0.12 (0.21)	1.00 (0.00)	1.00 (0.04)	1.00 (0.03)
	SVCR	10.0	31.0	49.0	0.52 (0.06)	0.53 (0.08)	0.53 (0.09)	0.04 (0.13)	0.07 (0.17)	0.07 (0.17)	1.00 (0.01)	1.00 (0.02)	1.00 (0.02)
Interstitial cells ovary	CART	10.0	35.0	30.0	0.52 (0.08)	0.55 (0.13)	0.55 (0.13)	0.06 (0.15)	0.18 (0.25)	0.19 (0.26)	0.99 (0.04)	0.94 (0.06)	0.95 (0.06)
	ENSMB	27.0	35.0	13.0	0.51 (0.03)	0.52 (0.06)	0.53 (0.07)	0.01 (0.07)	0.04 (0.12)	0.06 (0.14)	1.00 (0.02)	1.00 (0.03)	1.00 (0.02)
	KNN1	12.0	62.0	61.0	0.51 (0.06)	0.52 (0.08)	0.52 (0.09)	0.04 (0.13)	0.08 (0.18)	0.11 (0.18)	1.00 (0.13)	0.98 (0.09)	0.98 (0.09)
	LDA	20.0	19.0	11.0	0.50 (0.02)	0.56 (0.12)	0.56 (0.13)	0.00 (0.03)	0.18 (0.24)	0.19 (0.25)	1.00 (0.02)	0.98 (0.05)	0.98 (0.05)
	NB	16.0	46.0	41.0	0.51 (0.08)	0.55 (0.13)	0.53 (0.14)	0.33 (0.44)	0.85 (0.44)	0.83 (0.43)	0.95 (0.43)	0.54 (0.40)	0.53 (0.40)
	RF	12.0	42.0	52.0	0.52 (0.07)	0.51 (0.08)	0.52 (0.07)	0.06 (0.15)	0.05 (0.15)	0.05 (0.14)	0.99 (0.03)	1.00 (0.04)	1.00 (0.04)
	SVCL	10.0	10.0	11.0	0.50 (0.00)	0.52 (0.08)	0.52 (0.08)	0.00 (0.00)	0.07 (0.15)	0.07 (0.15)	1.00 (0.00)	1.00 (0.04)	1.00 (0.03)
	SVCR	21.0	61.0	64.0	0.50 (0.00)	0.51 (0.04)	0.51 (0.05)	0.00 (0.00)	0.02 (0.09)	0.03 (0.10)	1.00 (0.01)	1.00 (0.03)	1.00 (0.02)
Uterus effect	CART	11.0	38.0	69.0	0.52 (0.09)	0.54 (0.11)	0.54 (0.12)	0.10 (0.18)	0.16 (0.22)	0.17 (0.23)	0.98 (0.04)	0.93 (0.07)	0.94 (0.06)
	ENSMB	10.0	12.0	20.0	0.51 (0.06)	0.53 (0.06)	0.53 (0.07)	0.03 (0.12)	0.07 (0.13)	0.07 (0.14)	1.00 (0.02)	1.00 (0.02)	1.00 (0.02)
	KNN1	11.0	12.0	17.0	0.51 (0.07)	0.54 (0.09)	0.53 (0.08)	0.05 (0.15)	0.11 (0.18)	0.11 (0.18)	0.99 (0.12)	0.98 (0.06)	0.99 (0.10)
	LDA	10.0	20.0	20.0	0.52 (0.07)	0.51 (0.08)	0.54 (0.10)	0.05 (0.14)	0.08 (0.16)	0.12 (0.19)	1.00 (0.04)	1.00 (0.05)	0.98 (0.05)
	NB	12.0	16.0	14.0	0.55 (0.11)	0.56 (0.13)	0.56 (0.14)	0.58 (0.48)	0.80 (0.42)	0.79 (0.42)	0.96 (0.45)	0.73 (0.38)	0.72 (0.39)
	RF	10.0	18.0	20.0	0.52 (0.07)	0.53 (0.07)	0.53 (0.09)	0.06 (0.14)	0.08 (0.15)	0.09 (0.17)	0.97 (0.04)	1.00 (0.04)	1.00 (0.03)
	SVCL	10.0	14.0	10.0	0.50 (0.00)	0.50 (0.04)	0.52 (0.07)	0.00 (0.00)	0.02 (0.08)	0.06 (0.14)	1.00 (0.00)	1.00 (0.03)	1.00 (0.04)
	SVCR	11.0	44.0	67.0	0.51 (0.04)	0.52 (0.07)	0.52 (0.07)	0.02 (0.08)	0.06 (0.13)	0.05 (0.14)	1.00 (0.01)	1.00 (0.02)	1.00 (0.02)

Table E.4: Performance of ML models that predict *in vivo* outcomes in ovaries and uterus on balanced datasets (with SMOTE)

<i>In vivo</i> outcome	ML algorithm	Number of descriptors			Balanced Accuracy			Sensitivity			Specificity		
		bio	chem	cb	bio	chem	cb	bio	chem	cb	bio	chem	cb
Germinal cells ovary	CART	17.0	35.0	34.0	0.52 (0.11)	0.57 (0.15)	0.57 (0.14)	0.14 (0.22)	0.23 (0.30)	0.23 (0.28)	0.90 (0.13)	0.91 (0.05)	0.92 (0.05)
	ENSMB	13.0	16.0	28.0	0.53 (0.13)	0.59 (0.16)	0.59 (0.15)	0.19 (0.28)	0.29 (0.31)	0.30 (0.30)	0.87 (0.19)	0.89 (0.07)	0.87 (0.07)
	KNN1	13.0	13.0	45.0	0.54 (0.11)	0.60 (0.17)	0.58 (0.16)	0.15 (0.22)	0.44 (0.34)	0.35 (0.31)	0.93 (0.05)	0.75 (0.09)	0.81 (0.08)
	LDA	23.0	25.0	21.0	0.53 (0.15)	0.60 (0.18)	0.61 (0.18)	0.63 (0.36)	0.47 (0.35)	0.44 (0.34)	0.43 (0.27)	0.74 (0.11)	0.78 (0.10)
	NB	23.0	12.0	11.0	0.52 (0.14)	0.55 (0.16)	0.56 (0.17)	0.58 (0.38)	0.74 (0.30)	0.74 (0.32)	0.47 (0.32)	0.37 (0.14)	0.38 (0.13)
	RF	13.0	16.0	46.0	0.52 (0.11)	0.56 (0.12)	0.56 (0.12)	0.13 (0.26)	0.13 (0.25)	0.15 (0.23)	0.91 (0.22)	0.98 (0.05)	0.96 (0.04)
	SVCL	23.0	43.0	52.0	0.53 (0.14)	0.61 (0.17)	0.62 (0.18)	0.60 (0.35)	0.52 (0.34)	0.53 (0.35)	0.46 (0.30)	0.71 (0.10)	0.70 (0.09)
	SVCR	15.0	64.0	63.0	0.53 (0.13)	0.55 (0.14)	0.55 (0.13)	0.18 (0.29)	0.20 (0.27)	0.18 (0.26)	0.87 (0.20)	0.90 (0.06)	0.92 (0.06)
Interstitial cells ovary	CART	15.0	19.0	12.0	0.50 (0.10)	0.55 (0.13)	0.55 (0.13)	0.20 (0.29)	0.21 (0.25)	0.20 (0.24)	0.81 (0.26)	0.89 (0.07)	0.89 (0.06)
	ENSMB	37.0	56.0	41.0	0.52 (0.12)	0.56 (0.14)	0.58 (0.15)	0.48 (0.36)	0.27 (0.28)	0.29 (0.30)	0.56 (0.28)	0.86 (0.08)	0.86 (0.07)
	KNN1	17.0	35.0	24.0	0.51 (0.08)	0.57 (0.16)	0.57 (0.16)	0.07 (0.25)	0.37 (0.30)	0.38 (0.30)	0.94 (0.19)	0.77 (0.09)	0.75 (0.09)
	LDA	36.0	44.0	42.0	0.54 (0.14)	0.59 (0.16)	0.59 (0.16)	0.80 (0.28)	0.48 (0.31)	0.48 (0.32)	0.28 (0.16)	0.71 (0.09)	0.70 (0.10)
	NB	36.0	49.0	65.0	0.54 (0.13)	0.53 (0.14)	0.54 (0.15)	0.82 (0.26)	0.74 (0.31)	0.74 (0.31)	0.26 (0.13)	0.31 (0.22)	0.33 (0.11)
	RF	15.0	47.0	22.0	0.50 (0.10)	0.53 (0.10)	0.53 (0.10)	0.20 (0.31)	0.11 (0.19)	0.09 (0.20)	0.80 (0.27)	0.96 (0.05)	0.97 (0.05)
	SVCL	36.0	67.0	62.0	0.53 (0.13)	0.59 (0.16)	0.59 (0.16)	0.80 (0.30)	0.50 (0.30)	0.49 (0.31)	0.26 (0.21)	0.68 (0.14)	0.69 (0.10)
	SVCR	37.0	68.0	67.0	0.52 (0.11)	0.53 (0.12)	0.53 (0.11)	0.48 (0.37)	0.19 (0.23)	0.17 (0.22)	0.57 (0.31)	0.87 (0.08)	0.89 (0.07)
Uterus effect	CART	18.0	43.0	40.0	0.52 (0.11)	0.53 (0.12)	0.54 (0.11)	0.14 (0.20)	0.18 (0.24)	0.19 (0.22)	0.91 (0.15)	0.89 (0.06)	0.89 (0.07)
	ENSMB	21.0	61.0	37.0	0.54 (0.13)	0.54 (0.13)	0.54 (0.13)	0.26 (0.25)	0.25 (0.26)	0.24 (0.26)	0.81 (0.17)	0.83 (0.07)	0.85 (0.08)
	KNN1	11.0	11.0	53.0	0.55 (0.11)	0.55 (0.15)	0.56 (0.15)	0.18 (0.20)	0.39 (0.28)	0.32 (0.29)	0.91 (0.11)	0.71 (0.08)	0.79 (0.08)
	LDA	11.0	58.0	45.0	0.55 (0.14)	0.54 (0.15)	0.55 (0.15)	0.34 (0.28)	0.46 (0.30)	0.43 (0.28)	0.76 (0.19)	0.62 (0.12)	0.66 (0.10)
	NB	21.0	10.0	11.0	0.53 (0.14)	0.54 (0.15)	0.54 (0.15)	0.52 (0.38)	0.65 (0.32)	0.64 (0.28)	0.54 (0.34)	0.43 (0.19)	0.44 (0.13)
	RF	19.0	13.0	30.0	0.53 (0.10)	0.53 (0.09)	0.53 (0.09)	0.16 (0.21)	0.08 (0.17)	0.10 (0.17)	0.90 (0.14)	0.98 (0.05)	0.97 (0.05)
	SVCL	21.0	57.0	45.0	0.55 (0.14)	0.55 (0.15)	0.56 (0.15)	0.37 (0.29)	0.51 (0.31)	0.47 (0.31)	0.73 (0.24)	0.59 (0.12)	0.64 (0.11)
	SVCR	31.0	61.0	66.0	0.53 (0.13)	0.53 (0.11)	0.53 (0.11)	0.25 (0.26)	0.18 (0.23)	0.19 (0.23)	0.82 (0.20)	0.88 (0.10)	0.87 (0.09)

Table E.5: Performance of ML models that predict in vivo outcomes in testes and prostate on imbalanced datasets (with SMOTE)

In vivo outcome	ML algorithm	Number of descriptors			Balanced Accuracy			Sensitivity			Specificity		
		bio	chem	cb	bio	chem	cb	bio	chem	cb	bio	chem	cb
Spermatogenesis testis	CART	33.0	23.0	63.0	0.51 (0.08)	0.53 (0.11)	0.52 (0.11)	0.06 (0.15)	0.16 (0.20)	0.14 (0.21)	0.97 (0.05)	0.92 (0.07)	0.92 (0.07)
	ENSMB	18.0	49.0	26.0	0.50 (0.02)	0.50 (0.03)	0.50 (0.03)	0.00 (0.03)	0.02 (0.07)	0.01 (0.06)	1.00 (0.02)	0.99 (0.03)	1.00 (0.03)
	KNN1	37.0	45.0	46.0	0.50 (0.05)	0.50 (0.06)	0.50 (0.07)	0.02 (0.11)	0.04 (0.12)	0.05 (0.15)	0.98 (0.10)	0.97 (0.06)	0.98 (0.07)
	LDA	26.0	13.0	13.0	0.50 (0.03)	0.51 (0.09)	0.51 (0.10)	0.01 (0.06)	0.09 (0.16)	0.12 (0.18)	1.00 (0.03)	0.99 (0.06)	0.99 (0.06)
	NB	13.0	66.0	15.0	0.53 (0.10)	0.52 (0.13)	0.53 (0.13)	0.53 (0.48)	0.81 (0.46)	0.84 (0.45)	0.95 (0.45)	0.58 (0.41)	0.58 (0.41)
	RF	29.0	64.0	63.0	0.51 (0.07)	0.51 (0.07)	0.51 (0.06)	0.05 (0.13)	0.05 (0.14)	0.05 (0.12)	0.97 (0.04)	0.99 (0.05)	1.00 (0.04)
	SVCL	10.0	16.0	10.0	0.50 (0.00)	0.50 (0.04)	0.50 (0.05)	0.00 (0.00)	0.02 (0.07)	0.03 (0.09)	1.00 (0.00)	1.00 (0.03)	1.00 (0.03)
	SVCR	16.0	49.0	63.0	0.50 (0.02)	0.51 (0.04)	0.51 (0.05)	0.01 (0.04)	0.02 (0.08)	0.03 (0.09)	1.00 (0.01)	1.00 (0.03)	1.00 (0.03)
Interstitial cells testis	CART	20.0	13.0	61.0	0.51 (0.05)	0.52 (0.09)	0.52 (0.10)	0.06 (0.11)	0.20 (0.18)	0.18 (0.18)	0.97 (0.05)	0.89 (0.09)	0.89 (0.09)
	ENSMB	28.0	43.0	51.0	0.50 (0.03)	0.50 (0.03)	0.50 (0.04)	0.01 (0.05)	0.02 (0.07)	0.02 (0.07)	1.00 (0.03)	0.99 (0.03)	0.99 (0.04)
	KNN1	10.0	58.0	64.0	0.50 (0.05)	0.51 (0.07)	0.51 (0.07)	0.09 (0.19)	0.11 (0.15)	0.12 (0.15)	0.98 (0.16)	0.94 (0.09)	0.94 (0.10)
	LDA	33.0	19.0	11.0	0.50 (0.01)	0.52 (0.08)	0.52 (0.08)	0.01 (0.03)	0.14 (0.15)	0.16 (0.16)	1.00 (0.02)	0.99 (0.07)	0.99 (0.07)
	NB	22.0	68.0	69.0	0.51 (0.06)	0.51 (0.11)	0.52 (0.11)	0.10 (0.19)	0.40 (0.32)	0.44 (0.32)	0.99 (0.19)	0.86 (0.27)	0.84 (0.29)
	RF	12.0	47.0	60.0	0.50 (0.05)	0.50 (0.06)	0.51 (0.06)	0.05 (0.09)	0.07 (0.11)	0.08 (0.13)	0.96 (0.05)	0.98 (0.07)	0.98 (0.06)
	SVCL	10.0	19.0	14.0	0.50 (0.00)	0.50 (0.05)	0.50 (0.05)	0.00 (0.00)	0.04 (0.09)	0.05 (0.10)	1.00 (0.00)	1.00 (0.04)	1.00 (0.05)
	SVCR	23.0	47.0	64.0	0.50 (0.03)	0.50 (0.04)	0.51 (0.04)	0.01 (0.05)	0.03 (0.07)	0.04 (0.09)	1.00 (0.01)	0.99 (0.04)	0.99 (0.04)
Prostate effect	CART	16.0	67.0	19.0	0.51 (0.08)	0.52 (0.12)	0.53 (0.15)	0.04 (0.16)	0.08 (0.25)	0.10 (0.29)	0.99 (0.03)	0.97 (0.05)	0.97 (0.05)
	ENSMB	21.0	44.0	56.0	0.51 (0.06)	0.51 (0.05)	0.50 (0.05)	0.02 (0.12)	0.02 (0.11)	0.01 (0.10)	1.00 (0.01)	1.00 (0.01)	1.00 (0.02)
	KNN1	21.0	49.0	37.0	0.51 (0.07)	0.51 (0.08)	0.51 (0.09)	0.03 (0.15)	0.03 (0.16)	0.04 (0.18)	1.00 (0.14)	1.00 (0.05)	1.00 (0.07)
	LDA	18.0	14.0	18.0	0.51 (0.06)	0.52 (0.12)	0.53 (0.13)	0.02 (0.11)	0.08 (0.24)	0.10 (0.26)	1.00 (0.02)	1.00 (0.04)	0.99 (0.04)
	NB	14.0	38.0	62.0	0.52 (0.16)	0.52 (0.22)	0.54 (0.21)	0.43 (0.47)	0.85 (0.46)	0.84 (0.45)	0.88 (0.43)	0.53 (0.32)	0.56 (0.29)
	RF	21.0	65.0	62.0	0.51 (0.08)	0.51 (0.07)	0.51 (0.07)	0.04 (0.15)	0.02 (0.14)	0.02 (0.15)	0.99 (0.02)	1.00 (0.02)	1.00 (0.02)
	SVCL	10.0	12.0	18.0	0.50 (0.00)	0.50 (0.05)	0.50 (0.06)	0.00 (0.00)	0.01 (0.10)	0.02 (0.11)	1.00 (0.00)	1.00 (0.02)	1.00 (0.02)
	SVCR	11.0	44.0	56.0	0.50 (0.00)	0.51 (0.06)	0.51 (0.06)	0.00 (0.00)	0.02 (0.12)	0.02 (0.11)	1.00 (0.01)	1.00 (0.01)	1.00 (0.01)

Table E.6: Performance of ML models that predict in vivo outcomes in testes and prostate on balanced datasets (with SMOTE)

In vivo outcome	ML algorithm	Number of descriptors			Balanced Accuracy			Sensitivity			Specificity		
		bio	chem	cb	bio	chem	cb	bio	chem	cb	bio	chem	cb
Spermatogenesis testis	CART	33.0	68.0	37.0	0.52 (0.09)	0.51 (0.11)	0.52 (0.11)	0.09 (0.17)	0.13 (0.20)	0.16 (0.21)	0.94 (0.06)	0.89 (0.08)	0.88 (0.08)
	ENSMB	32.0	24.0	11.0	0.51 (0.10)	0.51 (0.12)	0.51 (0.12)	0.11 (0.20)	0.19 (0.23)	0.16 (0.23)	0.92 (0.09)	0.82 (0.07)	0.86 (0.09)
	KNN1	33.0	68.0	58.0	0.52 (0.09)	0.52 (0.14)	0.51 (0.14)	0.10 (0.19)	0.28 (0.27)	0.26 (0.28)	0.95 (0.12)	0.76 (0.09)	0.75 (0.10)
	LDA	14.0	24.0	18.0	0.51 (0.13)	0.51 (0.15)	0.51 (0.15)	0.26 (0.24)	0.35 (0.29)	0.33 (0.29)	0.76 (0.20)	0.67 (0.10)	0.69 (0.09)
	NB	31.0	59.0	63.0	0.48 (0.12)	0.50 (0.14)	0.50 (0.14)	0.46 (0.39)	0.68 (0.30)	0.69 (0.30)	0.50 (0.36)	0.31 (0.13)	0.30 (0.14)
	RF	37.0	35.0	62.0	0.52 (0.09)	0.51 (0.08)	0.51 (0.08)	0.10 (0.18)	0.06 (0.15)	0.07 (0.15)	0.94 (0.05)	0.95 (0.06)	0.94 (0.05)
	SVCL	29.0	68.0	11.0	0.50 (0.12)	0.50 (0.15)	0.50 (0.15)	0.29 (0.34)	0.45 (0.29)	0.27 (0.30)	0.72 (0.30)	0.56 (0.11)	0.74 (0.10)
	SVCR	32.0	62.0	48.0	0.50 (0.08)	0.51 (0.10)	0.51 (0.11)	0.06 (0.16)	0.17 (0.20)	0.11 (0.21)	0.94 (0.10)	0.85 (0.10)	0.91 (0.09)
Interstitial cells testis	CART	20.0	13.0	61.0	0.51 (0.05)	0.52 (0.09)	0.52 (0.10)	0.06 (0.11)	0.20 (0.18)	0.18 (0.18)	0.97 (0.05)	0.89 (0.09)	0.89 (0.09)
	ENSMB	28.0	43.0	51.0	0.50 (0.03)	0.50 (0.03)	0.50 (0.04)	0.01 (0.05)	0.02 (0.07)	0.02 (0.07)	1.00 (0.03)	0.99 (0.03)	0.99 (0.04)
	KNN1	10.0	58.0	64.0	0.50 (0.05)	0.51 (0.07)	0.51 (0.07)	0.09 (0.19)	0.11 (0.15)	0.12 (0.15)	0.98 (0.16)	0.94 (0.09)	0.94 (0.10)
	LDA	33.0	19.0	11.0	0.50 (0.01)	0.52 (0.08)	0.52 (0.08)	0.01 (0.03)	0.14 (0.15)	0.16 (0.16)	1.00 (0.02)	0.99 (0.07)	0.99 (0.07)
	NB	22.0	68.0	69.0	0.51 (0.06)	0.51 (0.11)	0.52 (0.11)	0.10 (0.19)	0.40 (0.32)	0.44 (0.32)	0.99 (0.19)	0.86 (0.27)	0.84 (0.29)
	RF	12.0	47.0	60.0	0.50 (0.05)	0.50 (0.06)	0.51 (0.06)	0.05 (0.09)	0.07 (0.11)	0.08 (0.13)	0.96 (0.05)	0.98 (0.07)	0.98 (0.06)
	SVCL	10.0	19.0	14.0	0.50 (0.00)	0.50 (0.05)	0.50 (0.05)	0.00 (0.00)	0.04 (0.09)	0.05 (0.10)	1.00 (0.00)	1.00 (0.04)	1.00 (0.05)
	SVCR	23.0	47.0	64.0	0.50 (0.03)	0.50 (0.04)	0.51 (0.04)	0.01 (0.05)	0.03 (0.07)	0.04 (0.09)	1.00 (0.01)	0.99 (0.04)	0.99 (0.04)
Prostate effect	CART	10.0	59.0	60.0	0.50 (0.11)	0.53 (0.15)	0.52 (0.15)	0.07 (0.24)	0.12 (0.29)	0.10 (0.29)	0.93 (0.12)	0.94 (0.05)	0.94 (0.05)
	ENSMB	22.0	55.0	69.0	0.50 (0.15)	0.54 (0.18)	0.54 (0.17)	0.14 (0.29)	0.20 (0.35)	0.20 (0.35)	0.87 (0.16)	0.89 (0.06)	0.88 (0.06)
	KNN1	30.0	63.0	69.0	0.50 (0.12)	0.54 (0.19)	0.53 (0.19)	0.07 (0.27)	0.20 (0.38)	0.19 (0.38)	0.93 (0.16)	0.88 (0.07)	0.87 (0.07)
	LDA	12.0	68.0	56.0	0.50 (0.20)	0.58 (0.22)	0.57 (0.22)	0.39 (0.44)	0.45 (0.43)	0.42 (0.43)	0.62 (0.31)	0.72 (0.09)	0.73 (0.08)
	NB	26.0	66.0	53.0	0.51 (0.19)	0.53 (0.22)	0.51 (0.22)	0.64 (0.45)	0.77 (0.45)	0.77 (0.45)	0.39 (0.27)	0.28 (0.12)	0.25 (0.12)
	RF	14.0	63.0	69.0	0.50 (0.13)	0.51 (0.10)	0.51 (0.09)	0.09 (0.26)	0.05 (0.21)	0.05 (0.18)	0.92 (0.12)	0.97 (0.03)	0.97 (0.04)
	SVCL	26.0	67.0	69.0	0.50 (0.20)	0.59 (0.22)	0.58 (0.22)	0.47 (0.44)	0.46 (0.43)	0.44 (0.43)	0.54 (0.32)	0.72 (0.10)	0.72 (0.09)
	SVCR	10.0	55.0	69.0	0.51 (0.14)	0.52 (0.14)	0.51 (0.13)	0.12 (0.28)	0.10 (0.28)	0.11 (0.26)	0.90 (0.11)	0.94 (0.05)	0.92 (0.05)

BIBLIOGRAPHY

- [1] A. Abdelaziz, H. Spahn-Langguth, K. Schramm, and I.V. Tetko. Consensus Modeling for HTS Assays Using *In silico* Descriptors Calculates the Best Balanced Accuracy in Tox21 Challenge. *Frontiers in Environmental Science*, 4:2, 2016.
- [2] C.H.G. Allen, A. Koutsoukas, I. Cortés-Ciriano, D.S. Murrell, T.E. Malliavin, R.C. Glen, and A. Bender. Improving the prediction of organism-level toxicity through integration of chemical, protein target and cytotoxicity qHTS data. *Toxicology Research*, 5(3):883–894, 2016.
- [3] E. Alpaydin. *Introduction to machine learning*. MIT press, 2009.
- [4] D. Altshuler, R.M. Durbin, G.r. Abecasis, D.R. Bentley, A. Chakravarti, A.G Clark, F.S Collins, F. De La Vega, P. Donnelly, M. Egholm, P. Flicek, S.B Gabriel, R. Gibbs, B. Knoppers, E.S. Lander, H. Lehrach, E. Mardis, G.A. McVean, D.A. Nickerson, and R.A. Cartwright. A map of human genome variation from population-scale sequencing. *Nature*, 467:1061–1073, 2010.
- [5] K. Ambe, K. Ishihara, T. Ochibe, K. Ohya, S. Tamura, K. Inoue, M. Yoshida, and M. Tohkin. *In Silico* Prediction of Chemical-Induced Hepatocellular Hypertrophy Using Molecular Descriptors. *Toxicological Sciences*, 162(2):667–675, 2018.
- [6] M.E. Andersen and D. Krewski. Toxicity Testing in the 21st Century: Bringing the Vision to Life. *Toxicological Sciences*, 107(2):324–330, 2009.
- [7] M.E. Andersen and D. Krewski. The Vision of Toxicity Testing in the 21st Century: Moving from Discussion to Action. *Toxicological Sciences*, 117(1):17–24, 2010.
- [8] G.T. Ankley, R.S. Bennett, R.J. Erickson, D.J. Hoff, M.W. Hornung, R.D. Johnson, D.R. Mount, J.W. Nichols, C.L. Russom, P.K. Schmieder, J.A. Serrano, J.E. Tietge, and D.L. Villeneuve. Adverse outcome pathways: A conceptual framework to support ecotoxicology research and risk assessment. *Environmental Toxicology and Chemistry*, 29(3):730–741, 2010.

- [9] N. Baker, A. Boobis, L. Burgoon, E. Carney, R. Currie, E. Fritsche, T. Knudsen, M. Laffont, A. Piersma, A. Poole, S. Schneider, and G. Daston. Building a developmental toxicity ontology. *Birth Defects Research*, 110:502–518, 2018.
- [10] P. Banerjee, V.B. Siramshetty, M.N. Drwal, and R. Preissner. Computational methods for prediction of *in vitro* effects of new chemical structures. *Journal of Cheminformatics*, 8(1):51, 2016.
- [11] I.I. Baskin. *Machine Learning Methods in Computational Toxicology*, pages 119–139. Springer New York, 2018.
- [12] G.E. Batista and M.C. Monard. An analysis of four missing data treatment methods for supervised learning. *Applied Artificial Intelligence*, 17(5-6):519–533, 2003.
- [13] R.A. Becker, D.A. Dreier, M.K. Manibusan, L.A. Cox, T.W. Simon, and J.S; Bus. How well can carcinogenicity be predicted by high throughput “characteristics of carcinogens” mechanistic data? *Regulatory Toxicology and Pharmacology*, 90:185–196, 2017.
- [14] S.M. Bell, X. Chang, J.F. Wambaugh, D.G. Allen, M. Bartels, K.L.R. Brouwer, W.M. Casey, N. Choksi, S.S. Ferguson, and G. Fraczkiewicz *et al.* *In vitro* to *in vivo* extrapolation for high throughput prioritization and decision making. *Toxicology in Vitro*, 47:213 – 227, 2018.
- [15] E. Benfenati. The CAESAR project for *in silico* models for the REACH legislation. *Chemistry Central journal*, 4 Suppl 1:I1, 2010.
- [16] E. Benfenati, R. Benigni, D.M. Demarini, C. Helma, D. Kirkland, T.M. Martin, P. Mazzatorta, G. Ouédraogo-arras, A.M. Richard, B. Schilter, W.G.E.J. Schoonen, R. D. Snyder, and C. Yang. Predictive Models for Carcinogenicity and Mutagenicity: Frameworks, State-of-the-Art, and Perspectives. *Journal of Environmental Science and Health*, 27(2):57–90, 2009.
- [17] E. Benfenati, A. Manganaro, and G. Gini. VEGA-QSAR: AI inside a platform for predictive toxicology. *CEUR Workshop Proceedings*, 1107:21–28, 01 2013.
- [18] R. Benigni. Structure-Activity Relationship Studies of Chemical Mutagens and Carcinogens: Mechanistic Investigations and Prediction Approaches. *Chemical Reviews*, 105(5):1767–1800, 2005.
- [19] R. Benigni, C.L. Battistelli, C. Bossa, A. Giuliani, and O. Tcheremenskaia. Endocrine Disruptors: Data-based survey of *in vivo* tests, predictive models and the Adverse Outcome Pathway. *Regulatory Toxicology and Pharmacology*, 86:18–24, 2017.
- [20] R. Benigni, Chiara Laura B., C. Bossa, A. Giuliani, E. Fioravanzo, A. Bassan, M. Fuart Gatnik, J. Rathman, C. Yang, and O. Tcheremenskaia. Evaluation of the applicability of existing (Q)SAR models for predicting the genotoxicity of pesticides and similarity

-
- analysis related with genotoxicity of pesticides for facilitating of grouping and read across. *EFSA Supporting Publications*, 16(3):1598E, 2019.
- [21] A.P. Bento, A. Gaulton, A. Hersey, B. Al-Lazikani, D. Michalovich, J. Chambers, L.J. Bellis, M. Davies, S. McGlinchey, Y. Light, and J.P. Overington. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Research*, 40(D1):D1100–D1107, 2011.
- [22] M.R. Berthold, N. Cebron, F. Dill, T.R. Gabriel, T. Kötter, T. Meinl, P. Ohl, C. Sieb, Ki. Thiel, and Be. Wiswedel. KNIME: The Konstanz Information Miner. In *Studies in Classification, Data Analysis, and Knowledge Organization*. Springer, 2007.
- [23] A. Bitsch, S. Escher, G. Lewin, C. Melber, N. Simetska, and I. Mangelsdorf. RepDose and FeDTex: Two databases focusing on systemic toxicity: First examples from analyses of repeated dose toxicity and reprotoxicity studies. *Toxicology Letters*, 180:202–210, 2008.
- [24] A. Bitsch, S. Jacobi, C. Melber, U. Wahnschaffe, N. Simetska, and I. Mangelsdorf. REP-DOSE: A database on repeated dose toxicity studies of commercial chemicals—A multi-functional tool. *Regulatory Toxicology and Pharmacology*, 46(3):202 – 210, 2006.
- [25] M. Bodén. A Guide to Recurrent Neural Networks and Backpropagation. In *The Dallas project*, 2002.
- [26] B. Boezio, K. Audouze, P. Ducrot, and O. Taboureau. Network-based Approaches in Pharmacology. *Molecular Informatics*, 36(10):1700048, 2017.
- [27] A.R. Boobis, J.E. Doe, B. Heinrich-Hirsch, M.E. Meek, S. Munn, M. Ruchirawat, J. Schlatter, J. Seed, and C. Vickers. IPCS framework for analyzing the relevance of a noncancer mode of action for humans. *Critical reviews in toxicology*, 38(2):87–96, 2008.
- [28] M. Bouhifd, M.E. Andersen, C. Baghdikian, K. Boekelheide, K.M. Crofton, A.J. Fornace, A. Kleensang, H. Li, C. Livi, A. Maertens, P.D. McMullen, M. Rosenberg, R. Thomas, M. Vantangoli, J.D Yager, L. Zhao, and T. Hartung. The Human Toxome Project. *Alternatives to animal experimentation*, 32(2):112–124, 2015.
- [29] P.B. Brazdil and C. Soares. A comparison of ranking methods for classification algorithm selection. In *European Conference on Machine Learning*, pages 63–75, 2000.
- [30] L. Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.
- [31] L. Breiman. Random Forests. *Machine Learning*, 45(1):5–32, 2001.
- [32] C. Bron and J. Kerbosch. Algorithm 457: finding all cliques of an undirected graph. *Communications of the ACM*, 16(9):575–577, 1973.

- [33] P. Browne, R.S. Judson, W.M. Casey, N.C. Kleinstreuer, and R.S. Thomas. Screening Chemicals for Estrogen Receptor Bioactivity Using a Computational Model. *Environmental Science & Technology*, 49(14):8804–8814, 2015.
- [34] J.S. Bus and R.A. Becker. Toxicity Testing in the 21st Century: A View from the Chemical Industry. *Toxicological Sciences*, 112(2):297–302, 2009.
- [35] F. Caiment, M. Tsamou, D. Jennen, and J. Kleinjans. Assessing compound carcinogenicity *in vitro* using connectivity mapping. *Carcinogenesis*, 35(1):201–207, 2014.
- [36] S.J. Capuzzi, R. Politi, O. Isayev, S. Farag, and A. Tropsha. QSAR Modeling of Tox21 Challenge Stress Response and Nuclear Receptor Signaling Toxicity Assays. *Frontiers in Environmental Science*, 4(43):3389–3, 2016.
- [37] A. Cereto-Massagué, M.J. Ojeda, C. Valls, M. Mulero, S. Garcia-Vallvé, and G. Pujadas. Molecular fingerprint similarity search in virtual screening. *Methods*, 71:58–63, 2015.
- [38] K.J. Chandler, M. Barrier, S. Jeffay, H.P. Nichols, N.C. Kleinstreuer, A.V. Singh, D.M. Reif, N.S. Sipes, R.S. Judson, D.J. Dix, R. Kavlock, E.S. Hunter, and T.B. Knudsen. Evaluation of 309 Environmental Chemicals Using a Mouse Embryonic Stem Cell Adherent Cell Differentiation and Cytotoxicity Assay. *PLoS ONE*, 6(6):e18540, 2011.
- [39] C.-C. Chang and C.-J. Lin. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27, 2011.
- [40] X. Chang, N. Kleinstreuer, P. Ceger, J. Hsieh, D. Allen, and W. Casey. Application of Reverse Dosimetry to Compare *In Vitro* and *In Vivo* Estrogen Receptor Activity. *Applied In Vitro Toxicology*, 1(1):33–44, 2015.
- [41] O. Chapelle, B. Scholkopf, and A. Zien. Semi-supervised learning. *IEEE Transactions on Neural Networks*, 20(3):542–542, 2009.
- [42] N. Chawla, K. Bowyer, L.O. Hall, and P.W. Kegelmeyer. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002.
- [43] N.V. Chawla, N. Japkowicz, and A. Kotcz. Special issue on learning from imbalanced data sets. *ACM Sigkdd Explorations Newsletter*, 6(1):1–6, 2004.
- [44] B. Chazelle. An optimal convex hull algorithm in any fixed dimension. *Discrete & Computational Geometry*, 10(4):377–409, 1993.
- [45] J.J. Chen, C.-A. Tsai, H. Moon, H. Ahn, J.J. Young, and C.-H. Chen. Decision threshold adjustment in class prediction. *SAR and QSAR in environmental research*, 17:337–52, 2006.

-
- [46] T. Chen and C. Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 785–794, 2016.
- [47] Y. Chen, F. Cheng, L. Sun, W. Li, G. Liu, and Y. Tang. Computational models to predict endocrine-disrupting chemical binding with androgen or oestrogen receptors. *Ecotoxicology and Environmental Safety*, 110:280 – 287, 2014.
- [48] A. Cherkasov, E.N. Muratov, D. Fourches, A. Varnek, I.I. Baskin, M. Cronin, J. Dearden, P. Gramatica, Y. C. Martin, R. Todeschini, V. Consonni, V.E. Kuz'min, R. Cramer, R. Benigni, C. Yang, J. Rathman, L. Terfloth, J. Gasteiger, A. Richard, and A. Tropsha. QSAR Modeling: Where Have You Been? Where Are You Going To? *Journal of Medicinal Chemistry*, 57(12):4977–5010, 2014.
- [49] F. Chollet et al. Keras. <https://github.com/fchollet/keras>, 2015.
- [50] S. Choudhary, A. Walker, K. Funk, C. Keenan, I. Khan, and K. Maratea. The standard for the exchange of nonclinical data (SEND): Challenges and promises. *Toxicologic Pathology*, 46(8):1006–1012, 2018.
- [51] H. Ciallella and H. Zhu. Advancing Computational Toxicology in the Big Data Era by Artificial Intelligence: Data-Driven and Mechanism-Driven Modeling for Chemical Toxicity. *Chemical Research in Toxicology*, 2019.
- [52] S.M. Cohen, A.R. Boobis, V.L. Dellarco, J.E. Doe, P.A. Fenner-Crisp, A. Moretto, T.P. Pastoor, R.S. Schoeny, J.G. Seed, and D.C. Wolf. Chemical carcinogenicity revisited 3: Risk assessment of carcinogenic potential based on the current state of knowledge of carcinogenesis in humans. *Regulatory Toxicology and Pharmacology*, 103:100 – 105, 2019.
- [53] EFSA Scientific Committee. Scientific opinion on the hazard assessment of endocrine disruptors: scientific criteria for identification of endocrine disruptors and appropriateness of existing test methods for assessing effects mediated by these substances on human health and the environment. *EFSA Journal*, 11(3):3132, 2013.
- [54] National Research Council. *Risk Assessment in the Federal Government: Managing the Process*. The National Academies Press, Washington, DC, 1983.
- [55] National Research Council. *Toxicity Testing in the 21st Century: A Vision and a Strategy*. The National Academies Press, Washington, DC, 2007.
- [56] National Research Council et al. *Science and judgment in risk assessment*. National Academies Press, Washington, DC, 1994.
- [57] M.T.D Cronin. Chapter 1 an introduction to chemical grouping, categories and read-across to predict toxicity. In *Chemical Toxicity Prediction: Category Formation and Read-Across*, pages 1–29. The Royal Society of Chemistry, 2013.

- [58] M. Daneshian, H. Kamp, J. Hengstler, M. Leist, and B. van de Water. Highlight report: Launch of a large integrated European *in vitro* toxicology project: EU-ToxRisk. *Archives of Toxicology*, 90(5):1021–1024, 2016.
- [59] Dassault Systèmes BIOVIA. *Pipeline Pilot v. 17.2.0.1361*. 2012.
- [60] R. De Maesschalck, D. Jouan-Rimbaud, and D.L. Massart. The mahalanobis distance. *Chemometrics and intelligent laboratory systems*, 50(1):1–18, 2000.
- [61] J.C. Dearden, M.T.D. Cronin, and K.L.E. Kaiser. How not to develop a quantitative structure–activity or structure–property relationship (QSAR/QSPR). *SAR and QSAR in Environmental Research*, 20(3-4):241–266, 2009.
- [62] V.L. Dellarco, D. McGregor, S.Co. Berry, S. M. Cohen, and A.R. Boobis. Thiazopyr and Thyroid Disruption: Case Study Within the Context of the 2006 IPCS Human Relevance Framework for Analysis of a Cancer Mode of Action. *Critical Reviews in Toxicology*, 36(10):793–801, 2006.
- [63] A. Dey. Machine learning algorithms: a review. *International Journal of Computer Science and Information Technologies*, 7(3):1174–1179, 2016.
- [64] E. Diamanti-Kandarakis, J.P. Bourguignon, L.C. Giudice, R. Hauser, G.S. Prins, A.M. Soto, R.T. Zoeller, and A.C. Gore. Endocrine-Disrupting Chemicals: An Endocrine Society Scientific Statement. *Endocrine Reviews*, 30(4):293–342, 2009.
- [65] T.G. Dietterich. Ensemble methods in machine learning. In *Multiple Classifier Systems*, pages 1–15, 2000.
- [66] 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, and O.G. Mekenyan. QSAR Toolbox – workflow and major functionalities. *SAR and QSAR in Environmental Research*, 27(3):203–219, 2016.
- [67] D. Ding, L. Xu, H. Fang, H. Hong, R. Perkins, S. Harris, E.D. Bearden, L. Shi, and W. Tong. The EDKB: an established knowledge base for endocrine disrupting chemicals. *BMC Bioinformatics*, 11(6):S5, 2010.
- [68] D.J. Dix, K.A. Houck, M.T. Martin, A.M. Richard, R.W. Setzer, and R.J. Kavlock. The ToxCast Program for Prioritizing Toxicity Testing of Environmental Chemicals. *Toxicological Sciences*, 95(1):5–12, 2007.
- [69] J.E. Doe, A.R. Boobis, V. Dellarco, P.A. Fenner-Crisp, A. Moretto, T.P. Pastoor, R.S. Schoeny, J.G. Seed, and D.C. Wolf. Chemical carcinogenicity revisited 2: Current knowledge of carcinogenesis shows that categorization as a carcinogen or non-carcinogen is not scientifically credible. *Regulatory Toxicology and Pharmacology*, 103:124 – 129, 2019.

-
- [70] N.R Draper and H. Smith. *Applied regression analysis*, volume 326. John Wiley & Sons, 2014.
- [71] M. Drwal, V. Siramshetty, P. Banerjee, A. Goede, R. Preissner, and Ma. Dunkel. Molecular similarity-based predictions of the Tox21 screening outcome. *Frontiers in Environmental Science*, 3:54, 2015.
- [72] H. Du, Y. Cai, H. Yang, H. Zhang, X. Yuhan, G. Liu, Y. Tang, and W. Li. *In Silico* Prediction of Chemicals Binding to Aromatase with Machine Learning Methods. *Chemical Research in Toxicology*, 30:1209–1218, 2017.
- [73] A.Z Dudek, T. Arodz, and J. Gálvez. Computational Methods in Developing Quantitative Structure-Activity Relationships (QSAR): A Review. *Combinatorial Chemistry & High Throughput Screening*, 9:213–228, 2006.
- [74] J.L. Durant, B.A. Leland, D.R. Henry, and J.G. Nourse. Reoptimization of MDL Keys for Use in Drug Discovery. *Journal of Chemical Information and Computer Sciences*, 42(6):1273–1280, 2002.
- [75] European Chemical Agency (ECHA) and European Food Safety Authority (EFSA) with the technical support of the Joint Research Centre (JRC). Guidance for the identification of endocrine disruptors in the context of Regulations (EU) No 528/2012 and (EC) No 1107/2009. *EFSA Journal*, 16.
- [76] F.a Eduati, L.M. Mangravite, T. Wang, H.o Tang, J.C. Bare, R. Huang, T. Norman, M. Kellen, M.P. Menden, J. Yang, and *et al.* Prediction of human population responses to toxic compounds by a collaborative competition. *Nature Biotechnology*, 33(9):933–940, 2015.
- [77] Panel on Plant Protection Products EFSA and their Residues. Guidance on the establishment of the residue definition for dietary risk assessment. *EFSA Journal*, 14, 2016.
- [78] T. Eissing. A Computational Systems Biology Software Platform for Multiscale Modeling and Simulation: Integrating Whole-Body Physiology, Disease Biology, and Molecular Reaction Networks. *Frontiers in Physiology*, 2:4, 2011.
- [79] C. Elkan. The Foundations of Cost-Sensitive Learning. In *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence*, pages 973–978, 2001.
- [80] S. Enoch. *Chemical Category Formation and Read-Across for the Prediction of Toxicity*, pages 209–219. Springer Netherlands, Dordrecht, 2010.
- [81] European Commission. *Guidance for the setting of an acute reference dose (ARfD)*. 2001.
- [82] European Commission. *Regulation (EC) No 1907/2006 of the European Parliament and of the Council of 18 December 2006 concerning the Registration, Evaluation, Authorization, and Restriction of Chemicals (REACH), establishing a European Chemicals Agency*,

- amending Directive 1999/45/EC and repealing Council Regulation (EEC) No 793/93 and Commission Regulation (EC) No 1488/94 as well as Council Directive 76/769/EEC and Commission Directives 91/155/EC, 93/67/EEC, 93/105/EC and 2001/21/EC. 2006.
- [83] European Commission. *No 1272/2008 of the European Parliament and of the Council of 16 December 2008 on classification, labelling and packaging of substances and mixtures, amending and repealing Directives 67/548/EEC and 1999/45/EC, and amending Regulation (EC) No 1907/2006*. 2008.
- [84] European Commission. *Regulation (EC) No 1107/2009 of the European Parliament and of the Council of 21 October 2009 concerning the placing of plant protection products on the market and repealing Council Directives 79/117/EEC and 91/414/EEC*. 2009.
- [85] E. Fabian, C. Gomes, B. Birk, T. Williford, T.R. Hernandez, C. Haase, R. Zbranek, B. van Ravenzwaay, and R. Landsiedel. *In vitro-to-in vivo* extrapolation (IVIVE) by PBTK modeling for animal-free risk assessment approaches of potential endocrine-disrupting compounds. *Archives of Toxicology*, 93(2):401–416, 2019.
- [86] T. Fawcett and F. Provost. Adaptive fraud detection. *Data Mining Knowledge Discovery*, 1(3):291–316, 1997.
- [87] M. Feher and T. Ewing. Global or local QSAR: Is there a way out? *QSAR & Combinatorial Science*, 28:850 – 855, 2009.
- [88] P.A. Fenner-Crisp, A.F. Maciorowski, and G.E. Timm. The endocrine disruptor screening program developed by the us environmental protection agency. *Ecotoxicology*, 9(1-2):85–91, 2000.
- [89] M.R. Fielden, A. Adai, R.T. Dunn, A. Olaharski, G. Searfoss, J. Sina, J. Aubrecht, E. Boitier, P. Nioi, S. Auerbach, D. Jacobson-kram, N. Raghavan, Y. Yang, A. Kincaid, J. Sherlock, S.J. Chen, and B. Car. Development and evaluation of a genomic signature for the prediction and mechanistic assessment of nongenotoxic hepatocarcinogens in the rat. *Toxicological Sciences*, 2011.
- [90] D.L. Filer, P. Kothiya, R. Setzer, R. Judson, and M. Martin. *tcpl*: The ToxCast Pipeline for High-Throughput Screening Data. *Bioinformatics*, 33:618–320, 2016.
- [91] R.B. Fitzpatrick. CPDB: Carcinogenic Potency Database. *Medical Reference Services Quarterly*, 27(3):303–311, 2008.
- [92] United Nations Economic Commission for European Secretariat. *Globally Harmonized System of Classification and Labelling of Chemicals (GHS)*. United Nations Publications, 2009.

-
- [93] D. Fourches, E. Muratov, and A. Tropsha. Trust, but verify: On the importance of chemical structure curation in cheminformatics and QSAR modeling research. *50(7)*:1189–1204, 2010.
- [94] E.A. Freeman and G.G. Moisen. A comparison of the performance of threshold criteria for binary classification in terms of predicted prevalence and kappa. *Ecological Modelling*, 217(1-2):48–58, 2008.
- [95] J.H. Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.
- [96] D. Gadaleta, S. Manganelli, A. Roncaglioni, C. Toma, E. Benfenati, and E. Mombelli. QSAR Modeling of ToxCast Assays Relevant to the Molecular Initiating Events of AOPs Leading to Hepatic Steatosis. *Journal of Chemical Information and Modeling*, 58(8):1501–1517, 2018.
- [97] F.M Gatnik and A. Worth. Review of Software Tools for Toxicity Prediction. Technical report, European Commission JRC, 2010.
- [98] E. Gelenbe. Stability of the random neural network model. *Neural Computation*, 2(2):239–247, 1990.
- [99] E. Gelenbe. Learning in the recurrent random neural network. *Neural Computation*, 5(1):154–164, 1993.
- [100] E. Gelenbe and S. Timotheou. Random neural networks with synchronized interactions. *Neural Computation*, 20:2308–2324, 2008.
- [101] P. Geurts. *Bias vs Variance Decomposition for Regression and Classification*, pages 733–746. Springer US, Boston, MA, 2010.
- [102] A.K. Ghose, V.N. Viswanadhan, and J.J. Wendoloski. Prediction of Hydrophobic (Lipophilic) Properties of Small Organic Molecules Using Fragmental Methods: An Analysis of ALOGP and CLOGP Methods. *The Journal of Physical Chemistry A*, 102(21):3762–3772, 1998.
- [103] T. Gocht, E. Berggren, H. Ahr, I. Cotgreave, M. Cronin, G. Daston, B. Hardy, E. Heinzle, J. Hescheler, D. Knight, C. Mahony, M. Peschanski, M. Schwarz, R. Thomas, C. Verfaillie, A. White, and M. Whelan. The SEURAT-1 approach towards animal free human safety assessment. *Alternatives to animal experimentation*, 32(1):9–24, 2015.
- [104] M. Goodarzi, B. Dejaeger, and Y.V. Heyden. Feature Selection Methods in QSAR Studies. *Journal of AOAC International*, 95(3):636–651, 2012.
- [105] I. Grenet, J.-P. Comet, F. Schorsch, N. Ryan, J. Wicharg, and D. Rouquie. Chemical *in vitro* bioactivity profiles are not informative about the long-term *in vivo* endocrine mediated toxicity. 2019. Submitted.

- [106] I. Grenet, K. Merlo, J.P. Comet, R. Tertiaux, D. Rouquié, and F. Dayan. Stacked Generalization with Applicability Domain Outperforms Simple QSAR on *in Vitro* Toxicological Data. *Journal of Chemical Information and Modeling*, 2019.
- [107] I. Grenet, Y. Yin, and J.P. Comet. G-Networks to Predict the Outcome of Sensing of Toxicity. *Sensors*, 18(10):3483, 2018.
- [108] I. Grenet, Y. Yin, J.P. Comet, and E. Gelenbe. Machine Learning to Predict Toxicity of Compounds. In *The 27th International Conference on Artificial Neural Networks (ICANN)*, pages 335–345. 2018.
- [109] C.J. Grondin, D. Sciaky, J. Wieggers, R.J. Johnson, T.C. Wieggers, A.P. Davis, C.J. Mattingly, and R. McMorran. The Comparative Toxicogenomics Database: update 2019. *Nucleic Acids Research*, 47(D1):D948–D954, 2018.
- [110] D. Guan, K. Fan, I. Spence, and S. Matthews. Combining machine learning models of *in vitro* and *in vivo* bioassays improves rat carcinogenicity prediction. *Regulatory Toxicology and Pharmacology*, 94:8 – 15, 2018.
- [111] I. Guyon and A. Elisseeff. An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.
- [112] D. Gómez and A. Rojas. An Empirical Overview of the No Free Lunch Theorem and Its Effect on Real-World Machine Learning Classification. *Neural Computation*, 28(1):216–228, 2016.
- [113] F. Güneş, R. Wolfinger, and P. Tan. Stacked ensemble models for improved prediction accuracy. In *SAS Global Forum Proceedings*, 2017.
- [114] H. He and E.A. Garcia. Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284, 2009.
- [115] L.H. Hall and L.B. Kier. Electrotological State Indices for Atom Types: A Novel Combination of Electronic, Topological, and Valence State Information. *Journal of Chemical Information and Computer Sciences*, 35(6):1039–1045, 1995.
- [116] H. Han, W.-Y. Wang, and B.-H. Mao. Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning. In *Advances in Intelligent Computing*, pages 878–887. Springer Berlin Heidelberg, 2005.
- [117] L. Han, Y. Wang, and S.H. Bryant. Developing and validating predictive decision tree models from mining chemical structural fingerprints and high-throughput screening data in PubChem. *BMC Bioinformatics*, 9(1):401, 2008.
- [118] C.E. Handford, C.T. Elliott, and K. Campbell. A review of the global pesticide legislation and the scale of challenge in reaching the global harmonization of food safety standards. *Integrated Environmental Assessment and Management*, 11(4):525–536, 2015.

-
- [119] C. Hansch. Quantitative structure-activity relationships and the unnamed science. *Accounts of Chemical Research*, 26(4):147–153, 1993.
- [120] C. Hansch, P.P Maloney, T. Fujita, and R.M Muir. Correlation of Biological Activity of Phenoxyacetic Acids with Hammett Substituent Constants and Partition Coefficients. *Nature*, 194:178–180, 1962.
- [121] B. Hardy, G. Apic, P. Carthew, D. Clark, D. Cook, I. Dix, S. Escher, J. Hastings, D..J Heard, N. Jeliaskova, P. Judson, S. matis Mitchell, D. Mitic Potkrajac, G. Myatt, I. Shah, O. Spjuth, O. Tcheremenskaia, L. Toldo, D. Watson, and C. Yang. Toxicology Ontology Perspectives. *Alternatives to animal experimentation*, pages 139–156, 2012.
- [122] L. Harland. Open PHACTS: A Semantic Knowledge Infrastructure for Public and Commercial Drug Discovery Research. In *Knowledge Engineering and Knowledge Management*, pages 1–7. Springer Berlin Heidelberg, 2012.
- [123] H. He., Y. Bai, E. A. Garcia, and S. Li. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In *IEEE International Joint Conference on Neural Networks*, pages 1322–1328, 2008.
- [124] E.A Helgee, L. Carlsson, S. Boyer, and U. Norinder. Evaluation of Quantitative Structure Activity Relationship Modeling Strategies: Local and Global Models. *Journal of Chemical Information and Modeling*, 50(4):677–689, 2010.
- [125] S. Heller, A. McNaught, S. Stein, D. Tchekhovskoi, and I. Pletnev. InChI - the worldwide chemical structure identifier standard. *Journal of Cheminformatics*, 5(1):7, 2013.
- [126] V.J. Hodge and J. Austin. A survey of outlier detection methodologies. *Artificial Intelligence Review*, 22(2):85–126, 2004.
- [127] M. Hossin and M.N. Sulaiman. A review on evaluation metrics for data classification evaluations. *International Journal of Data Mining & Knowledge Management Process*, 5(2):1, 2015.
- [128] C.W. Hsu, R. Huang, M.S.A. Attene-Ramos, C.P Austin, A. Simeonov, and M. Xia. Advances in high-throughput screening technology for toxicology. *International Journal of Risk Assessment and Management*, 20(1/2/3):109–135, 2017.
- [129] R. Huang and M. Xia. Tox21 challenge to build predictive models of nuclear receptor and stress response pathways as mediated by exposure to environmental toxicants and drugs. *Frontiers in Environmental Science*, 5:85, 2017.
- [130] R. Huang, M. Xia, S. Sakamuru, J.H. Zhao, S. Shahane, M.S. Attene-Ramos, T. Zhao, C.P. Austin, and A. Simeonov. Modelling the Tox21 10K chemical profiles for *in vivo* toxicity prediction and mechanism characterization. In *Nature communications*, 2016.

- [131] E.A. Cohen Hubal, A. Richard, L. Aylward, S. Edwards, J. Gallagher, M.-R. Goldsmith, S. Isukapalli, R. Tornero-Velez, E. Weber, and R. Kavlock. Advancing Exposure Characterization for Chemical Evaluation and Risk Assessment. *Journal of Toxicology and Environmental Health, Part B*, 13(2-4):299–313, 2010.
- [132] Y. Igarashi, N. Nakatsu, T. Yamashita, A. Ono, Y. Ohno, T. Urushidani, and H. Yamada. Open TG-GATEs: A large-scale toxicogenomics database. *Nucleic acids research*, 43:D921–D927, 2014.
- [133] IPCS. *Global Assessment of the State-of-Science of Endocrine Disruptors*. Geneva: World Health Organization, 2002.
- [134] IPCS & OECD. *IPCS risk assessment terminology*. Geneva: World Health Organization, 2004.
- [135] M. Jamei, S. Marciniak, K. Feng, A. Barnett, G. Tucker, and A. Rostami-Hodjegan. The Simcyp Population-based ADME Simulator. *Expert Opinion on Drug Metabolism & Toxicology*, 5(2):211–223, 2009.
- [136] J. Jaworska, M. Comber, C. Auer, and K. Van Leeuwen. Summary of a Workshop on Regulatory Acceptance of (Q)SARs for Human Health and Environmental Endpoints. *Environmental health perspectives*, 111:1358–1360, 2003.
- [137] J. Jaworska, N. Nikolova-Jeliazkova, and T. Aldenberg. QSAR applicability domain estimation by projection of the training set descriptor space: a review. *Alternatives to laboratory animals*, 33(5):445–459, 2005.
- [138] A. Jović, K. Brkić, and N. Bogunović. A review of feature selection methods with applications. In *38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pages 1200–1205, 2015.
- [139] R. Judson, K. Houck, R. Kavlock, T. Knudsen, M. Martin, H. Mortensen, D. Reif, D. Rotroff, I. Shah, A. Richard, and D.J. Dix. *In Vitro* Screening of Environmental Chemicals for Targeted Testing Prioritization: The ToxCast Project. *Environmental Health Perspectives*, 118(4):485–492, 2010.
- [140] R. Judson, A. Richard, D. Dix, K. Houck, F. Elloumi, M. Martin, T. Cathey, T.R. Transue, R. Spencer, and M. Wolf. ACToR — Aggregated Computational Toxicology Resource. *Toxicology and Applied Pharmacology*, 233(1):7 – 13, 2008.
- [141] R.S. Judson, R.J. Kavlock, M. T. Martin, D.M. Reif, K.A. Houck, T. Knudsen, A. Richard, R.R. Tice, M. Whelan, M. Xia, R. Huang, C.P. Austin, G.P. Daston, T. Hartung, J.R. Fowle, W. Wooge, W. Tong, and D. Dix. Perspectives on validation of high-throughput assays supporting 21st century toxicity testing. *Alternatives to animal experimentation*, 30 1:51–6, 2013.

-
- [142] R.S. Judson, F.M. Magpantay, V. Chickarmane, C. Haskell, N. Tania, J. Taylor, M. Xia, R. Huang, D.M. Rotroff, D.L. Filer, K.A. Houck, M.T. Martin, N. Sipes, A.M. Richard, K. Mansouri, R.W. Setzer, T.B. Knudsen, K.M. Crofton, and R.S. Thomas. Integrated Model of Chemical Perturbations of a Biological Pathway Using 18 *In Vitro* High-Throughput Screening Assays for the Estrogen Receptor. *Toxicological sciences : an official journal of the Society of Toxicology*, 148(1):137–54, 2015.
- [143] E.R. Kabir, M.S. Rahman, and I. Rahman. A review on endocrine disruptors and their possible impacts on human health. *Environmental Toxicology and Pharmacology*, 40(1):241–258, 2015.
- [144] R.M. Karp. *Reducibility among Combinatorial Problems*, pages 85–103. Springer US, Boston, MA, 1972.
- [145] C.M. Keenan, J.F. Baker, A.E. Bradley, D.G. Goodman, T. Harada, R. Herbert, W. Kaufmann, R. Kellner, B. Mahler, E. Meseck, T. Nolte, S. Rittinghausen, J. Vahle, and K. Yoshizawa. International Harmonization of Nomenclature and Diagnostic Criteria (IN-HAND) progress to date and future plans. *Journal of Toxicologic Pathology*, 28(1):51–53, 2015.
- [146] M. Kim, R. Huang, A. Sedykh, W. Wang, M. Xia, and H. Zhu. Mechanism Profiling of Hepatotoxicity Caused by Oxidative Stress Using the Antioxidant Response Element Reporter Gene Assay Models and Big Data. *Environmental Health Perspectives*, 124:634–641, 2015.
- [147] S. Kim, L. Han, B. Yu, V.D. Hähnke, E.E. Bolton, and S.H. Bryant. PubChem structure-activity relationship (SAR) clusters. *Journal of Cheminformatics*, 2015.
- [148] W.D. Klaren, C. Ring, J.E. Rager, C.M. Thompson, M.A. Harris, S. Borghoff, N.S. Sipes, S.S. Auerbach, and J.-H. Hsieh. Identifying Attributes That Influence *In Vitro*-to-*In Vivo* Concordance by Comparing *In Vitro* Tox21 Bioactivity Versus *In Vivo* DrugMatrix Transcriptomic Responses Across 130 Chemicals. *Toxicological Sciences*, 167(1):157–171, 2018.
- [149] N.C. Kleinstreuer, P. Ceger, E.D. Watt, M. Martin, K. Houck, P. Browne, R.S. Thomas, W.M. Casey, D.J. Dix, D. Allen, S. Sakamuru, M. Xia, R. Huang, and R. Judson. Development and Validation of a Computational Model for Androgen Receptor Activity. *Chemical Research in Toxicology*, 30(4):946–964, 2017.
- [150] N.C. Kleinstreuer, D.J. Dix, K.A. Houck, R.J. Kavlock, T.B. Knudsen, M.T. Martin, K.B. Paul, D.M. Reif, K.M. Crofton, K. Hamilton, R. Hunter, I. Shah, and R.S. Judson. *In Vitro* Perturbations of Targets in Cancer Hallmark Processes Predict Rodent Chemical Carcinogenesis. *Toxicological Sciences*, 131(1):40–55, 2013.

- [151] N.C. Kleinstreuer, A.L. Karmaus, K. Mansouri, D.G. Allen, J.M. Fitzpatrick, and G. Patlewicz. Predictive models for acute oral systemic toxicity: A workshop to bridge the gap from research to regulation. *Computational Toxicology*, 8:21 – 24, 2018.
- [152] K. Koch. Bayes’ theorem. In *Bayesian Inference with Geodetic Applications*, pages 4–8. Springer, 1990.
- [153] T. Kohonen. The self-organizing map. *Proceedings of the IEEE*, 78(9):1464–1480, 1990.
- [154] S.B. Kotsiantis. Supervised Machine Learning: A Review of Classification Techniques. In *Proceedings of the 2007 Conference on Emerging Artificial Intelligence Applications in Computer Engineering*, pages 3–24, Amsterdam, The Netherlands, The Netherlands, 2007.
- [155] D. Krewski, D. Acosta, M. Andersen, H. Anderson, J. Bailar, K. Boekelheide, R. Brent, G. Charnley, V. G Cheung, S. Green, K.T. Kelsey, N. Kerkvliet, A.A. Li, L. Mccray, O. Meyer, W. Patterson, R.D .and Pennie, R.A. Scala, G. Solomon, and L. Zeise. Toxicity Testing in the 21st Century: A Vision and A Strategy. *Journal of toxicology and environmental health. Part B, Critical reviews*, 13:51–138, 2010.
- [156] M. Kubat, R.C. Holte, and S. Matwin. Machine learning for the detection of oil spills in satellite radar images. *Machine learning*, 30(2-3):195–215, 1998.
- [157] L. Kuepfer, C. Niederal, T. Wendl, J-F. Schlender, S. Willmann, J. Lippert, M. Block, T. Eissing, and D. Teutonico. Applied Concepts in PBPK Modeling: How to Build a PBPK/PD Model. *CPT: Pharmacometrics & Systems Pharmacology*, 5(10):516–531.
- [158] M. Kuhn. The caret Package, 2009.
- [159] A. Lagunin, A. Zakharov, D. Filimonov, and V. Poroikov. QSAR Modelling of Rat Acute Toxicity on the Basis of PASS Prediction. *Molecular Informatics*, 30(2-3):241–250, 2011.
- [160] J. Lamb, E.D. Crawford, D. Peck, J.W. Modell, I.C. Blat, M.J. Wrobel, J. Lerner, J.-P. Brunet, A. Subramanian, K.N. Ross, M. Reich, H. Hieronymus, G. Wei, S.A. Armstrong, S.J. Haggarty, P.A. Clemons, R. Wei, S.A. Carr, E.S. Lander, and T.R. Golub. The Connectivity Map: Using Gene-Expression Signatures to Connect Small Molecules, Genes, and Disease. *Science*, 313(5795):1929–1935, 2006.
- [161] I.A. Lea, H. Gong, A. Paleja, A. Rashid, and J. Fostel. CEBS: A comprehensive annotated database of toxicological data. *Nucleic acids research*, 45:D964–D971, 2016.
- [162] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436, 2015.
- [163] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. In *Proceedings of the IEEE*, pages 2278–2324, 1998.

-
- [164] M. Leist, A. Ghallab, R. Graepel, R. Marchan, R. Hassan, S. H. Bennekou, A. Limonciel, M. Vinken, S. Schildknecht, and *et al.* Adverse outcome pathways: opportunities, limitations and open questions. *Archives of Toxicology*, 91(11):3477–3505, 2017.
- [165] M. Leshno, V.Y Lin, A. Pinkus, and S. Schocken. Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural networks*, 6(6):861–867, 1993.
- [166] T. Li, S. Zhu, and M. Ogihara. Using discriminant analysis for multi-class classification: an experimental investigation. *Knowledge and Information Systems*, 10(4):453–472, 2006.
- [167] X. Li, L. Chen, G. Cheng, F. Liu, X. Shen, and Y. Tang. *In Silico* Prediction of Chemical Acute Oral Toxicity Using Multi-Classification Methods. *Journal of Chemical Information and Modeling*, 54(4):1061–1069, 2014.
- [168] C. Lipinski and A. Hopkins. Navigating chemical space for biology and medicine. *Nature*, 432(7019):855–861, 2004.
- [169] J. Liu, K. Mansouri, R.S. Judson, M.T. Martin, H. Hong, M. Chen, X. Xu, R.S. Thomas, and I. Shah. Predicting Hepatotoxicity Using ToxCast *in Vitro* Bioactivity and Chemical Structure. *Chemical Research Toxicology*, 28(4):738–751, 2015.
- [170] J. Liu, G. Patlewicz, A.J. Williams, R.S Thomas, and I. Shah. Predicting Organ Toxicity Using *in Vitro* Bioactivity Data and Chemical Structure. *Chemical Research Toxicology*, 30(11):2046–2059, 2017.
- [171] W. Loh. Classification and regression tree methods. *Wiley StatsRef: Statistics Reference Online*, 2008.
- [172] J. Lousse, K. Beekmann, and I.M.C.M. Rietjens. Use of Physiologically Based Kinetic Modeling-Based Reverse Dosimetry to Predict *in Vivo* Toxicity from *in Vitro* Data. *Chemical Research in Toxicology*, 30(1):114–125, 2017.
- [173] Y. Low, A. Sedykh, I. Rusyn, and A. Tropsha. Integrative Approaches for Predicting *In Vivo* Effects of Chemicals from their Structural Descriptors and the Results of Short-Term Biological Assays. *Current Topics in Medicinal Chemistry*, 14(11):1356–1364, 2014.
- [174] Y. Low, T. Uehara, Y. Minowa, H. Yamada, Y. Ohno, T. Urushidani, A. Sedykh, E. Muratov, V. Kuzmin, D. Fourches, H. Zhu, I. Rusyn, and A. Tropsha. Predicting drug-induced hepatotoxicity using QSAR and toxicogenomics approaches. *Chemical Research in Toxicology*, 2011.
- [175] S. Ma and J. Huang. Penalized feature selection and classification in bioinformatics. *Briefings in bioinformatics*, 9:392–403, 2008.

- [176] K. Madasamy and M. Ramaswami. Data Imbalance and Classifiers: Impact and Solutions from a Big Data Perspective. *IJCIR*, 13(9):2267–2281, 2017.
- [177] Y. Malgrange. Recherche des sous-matrices premières d’une matrice à coefficients binaires. applications à certains problèmes de graphe. In *Proceedings of the Deuxième Congrès de l’AFCALTI*, pages 231–242, 1962.
- [178] C.A. Marchant, K. Briggs, and A. Long. *In Silico* Tools for Sharing Data and Knowledge on Toxicity and Metabolism: Derek for Windows, Meteor, and Vitic. *Toxicology mechanisms and methods*, 18:177–87, 2008.
- [179] E. Martin, P. Mukherjee, D. Sullivan, and J. Jansen. Profile-QSAR: A Novel meta-QSAR Method that Combines Activities across the Kinase Family To Accurately Predict Affinity, Selectivity, and Cellular Activity. *Journal of Chemical Information and Modeling*, 51(8):1942–1956, 2011.
- [180] E.J. Martin, V.R. Polyakov, L. Tian, and R.C. Perez. Profile-QSAR 2.0: Kinase Virtual Screening Accuracy Comparable to Four-Concentration IC50s for Realistically Novel Compounds. *Journal of Chemical Information and Modeling*, 57(8):2077–2088, 2017.
- [181] M.T. Martin, T.B. Knudsen, D.M. Reif, K.A. Houck, R.S. Judson, R.J. Kavlock, and D.J. Dix. Predictive model of rat reproductive toxicity from ToxCast high throughput screening. *Biology of reproduction*, 85(2):327–39, 2011.
- [182] C.J. Mattingly, G.T. Colby, J.N. Forrest, and J.L. Boyer. The Comparative Toxicogenomics Database (CTD). *Environmental Health Perspectives*, 111(6):793–795, 2003.
- [183] A. Mayr, G. Klambauer, T. Unterthiner, and S. Hochreiter. DeepTox: Toxicity Prediction using Deep Learning. *Frontiers in Environmental Science*, 3(80), 2016.
- [184] M.A. Mazurowski, P.A. Habas, J.M. Zurada, J.Y. Lo, J.A. Baker, and G.D. Tourassi. Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance. *Neural Networks*, 21(2-3):427–436, 2008.
- [185] B. Meek and J. Doull. Pragmatic Challenges for the Vision of Toxicity Testing in the 21st Century in a Regulatory Context: Another Ames Test? ... or a New Edition of “the Red Book”? *Toxicological Sciences*, 108(1):19–21, 2009.
- [186] B. Meek, C.M. Palermo, A.N. Bachman, C.M. North, and R. Lewis. Mode of action human relevance (species concordance) framework: Evolution of the Bradford Hill considerations and comparative analysis of weight of evidence. *Journal of applied toxicology*, 34:595–606, 2014.
- [187] B.H. Menze, B.M. Kelm, R. Masuch, U. Himmelreich, P. Bachert, W. Petrich, and F.A. Hamprecht. A comparison of random forest and its gini importance with standard chemo-

-
- metric methods for the feature selection and classification of spectral data. *BMC Bioinformatics*, 10(1):213, 2009.
- [188] D. Meyer. Support Vector Machines. The Interface to libsvm in package e1071., 2001.
- [189] J.B.O. Mitchell. Machine learning methods in chemoinformatics. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 4(5):468–481, 2014.
- [190] T.I Netzeva, A.P Worth, T. Aldenberg, R. Benigni, M.T.D Cronin, P. Gramatica, J.S Jaworska, S. Kahn, G. Klopman, C.A Marchant, et al. Current status of methods for defining the applicability domain of (quantitative) structure-activity relationships. *ATLA*, 33:155–173, 2005.
- [191] H.W. Ng, S.W Doughty, H. Luo, H. Ye, W. Ge, W. Tong, and H. Hong. Development and Validation of Decision Forest Model for Estrogen Receptor Binding Prediction of Chemicals Using Large Data Sets. *Chemical Research in Toxicology*, 28(12):2343–2351, 2015.
- [192] W.S. Noble. What is a Support Vector Machine? *Nature biotechnology*, 24:1565–7, 2007.
- [193] U. Norinder and S. Boyer. Conformal prediction classification of a large data set of environmental chemicals from toxcast and tox21 estrogen receptor assays. *Chemical research in toxicology*, 29(6):1003–1010, 2016.
- [194] N.M. O’Boyle. Towards a Universal SMILES representation - A standard method to generate canonical SMILES based on the InChI. *Journal of cheminformatics*, 4:22, 2012.
- [195] N.M. O’Boyle, C. Morley, and G.R. Hutchison. Pybel: a Python wrapper for the OpenBabel cheminformatics toolkit. *Chemistry Central Journal*, 2(1):5, 2008.
- [196] OECD. Conceptual Framework for Testing and Assessment of Endocrine Disrupters. *Official Journal of the European Union*, 2002.
- [197] OECD. Directive 2004/10/EC of the European Parliament and of the Council of 11 February 2004. The OECD Principles of Good Laboratory Practice (GLP). *Official Journal of the European Union*, 2004.
- [198] OECD. Guidance Document on the Validation of (Quantitative) Structure-Activity Relationship [(Q)SAR] Models. page 154, 2014.
- [199] OECD. Guidance Document for the use of Adverse Outcome Pathways in developing Integrated Approaches to Testing and Assessment (IATA). *Series on Testing and Assessment*, 260, 2017.
- [200] M. Omran, A. Engelbrecht, and A. Salman. An overview of clustering methods. *Intelligent Data Analysis*, 11:583–605, 2007.

- [201] S. Palei and S. Kumar Das. Logistic regression model for prediction of roof fall risks in bord and pillar workings in coal mines: An approach. *Safety Science - SAF SCI*, 47:88–96, 2009.
- [202] S. Parasuraman. Toxicological screening. *Journal of Pharmacology & Pharmacotherapeutics*, 2:74–79, 2011.
- [203] G. Patlewicz, N. Jeliaskova, R.J. Safford, A. Worth, and B. Aleksiev. An evaluation of the implementation of the Cramer classification scheme in the Toxtree software. *SAR and QSAR in environmental research*, 19:495–524, 2008.
- [204] S. Patro and K.K. Sahu. Normalization: A preprocessing stage. *arXiv preprint arXiv:1503.06462*, 2015.
- [205] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
- [206] R. Peeters. The maximum edge biclique problem is NP-complete. *Discrete Applied Mathematics*, 131(3):651–654, 2003.
- [207] P. Peres-Neto, D.A. Jackson, and K. Somers. How Many Principal Components? Stopping Rules for Determining the Number of Non-Trivial Axes Revisited. *Computational Statistics and Data Analysis*, 49:974–997, 2005.
- [208] D. Petrakis, L. Vassilopoulou, C. Mamoulakis, C. Psycharakis, A. Anifantaki, S. Sifakis, A.O. Docea, J. Tsiacoussis, A. Makrigiannakis, and A.M. Tsatsakis. Endocrine Disruptors Leading to Obesity and Related Diseases. *International Journal of Environmental Research and Public Health*, 14(10):1282, 2017.
- [209] L.M. Plunkett, M.A. Kaplan, and R.A. Becker. Challenges in using the ToxRefDB as a resource for toxicity prediction modeling. *Regulatory Toxicology and Pharmacology*, 72(3):610–614, 2015.
- [210] P. Pradeep, R.J. Povinelli, S. White, and S.J. Merrill. An ensemble model of QSAR tools for regulatory risk assessment. *Journal of Cheminformatics*, 8(1):48, 2016.
- [211] R Core Team. *R: A Language and Environment for Statistical Computing*, 2013.
- [212] A.B. Raies and V.B. Bajic. *In silico* toxicology: computational methods for the prediction of chemical toxicity. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 6:147–172, 2016.
- [213] H. Raunio. *In Silico* toxicology - non-testing methods. *Frontiers in Pharmacology*, 2:33, 2011.

-
- [214] D.M. Reif, M.T. Martin, R.J. Kavlock, R.S. Judson, and D.J. Dix. Profiling Chemicals Based on Chronic Toxicity Results from the U.S. EPA ToxRef Database. *Environmental Health Perspectives*, 117(3):392–399, 2008.
- [215] K. Ribay, Ma.T. Kim, W. Wang, D. Pinolini, and H. Zhu. Predictive Modeling of Estrogen Receptor Binding Agents Using Advanced Cheminformatics Tools and Massive Public Data. *Frontiers in Environmental Science*, 4:12, 2016.
- [216] A. Richard. DSSTox web site launch: Improving public access to databases for building structure-toxicity prediction models. *Preclinica*, 2:103–108, 2004.
- [217] A.M. Richard, R.S. Judson, K.A. Houck, C.M. Grulke, P. Volarath, I. Thillainadarajah, C. Yang, J. Rathman, M.T. Martin, J.F. Wambaugh, T.B. Knudsen, J. Kancherla, K. Mansouri, G. Patlewicz, A.J. Williams, S.B. Little, K.M. Crofton, and R.S. Thomas. ToxCast Chemical Landscape: Paving the Road to 21st Century Toxicology. *Chemical Research in Toxicology*, 29(8):1225–1251, 2016.
- [218] A.M. Richard, C. Yang, and R.S. Judson. Toxicity Data Informatics: Supporting a New Paradigm for Toxicity Prediction. *Toxicology Mechanisms and Methods*, 18(2-3):103–118, 2008.
- [219] S. Riniker, Y. Wang, J.L. Jenkins, and G.A. Landrum. Using Information from Historical High-Throughput Screens to Predict Active Compounds. *Journal of Chemical Information and Modeling*, 54(7):1880–1891, 2014.
- [220] D. Rogers and M. Hahn. Extended-Connectivity Fingerprints. *Journal of Chemical Information and Modeling*, 50(5):742–754, 2010.
- [221] F.P. Roth and J. Klekota. Chemical substructures that enrich for biological activity. *Bioinformatics*, 24(21):2518–2525, 2008.
- [222] D. Rouquie, M. Heneweer, J. Botham, H. Ketelslegers, L.K. Markell, T. Pfister, W. Steiling, V. Strauss, and C. Hennes. Contribution of New Technologies to Characterisation and Prediction of Adverse Effects. *Critical Reviews in Toxicology*, 45:1–12, 2015.
- [223] D. Rouquié, H. Tinwell, O. Blanck, F. Schorsch, D. Geter, S. Wason, and R. Bars. Thyroid tumor formation in the male mouse induced by fluopyram is mediated by activation of hepatic CAR/PXR nuclear receptors. *Regulatory Toxicology and Pharmacology*, 70(3):673–680, 2014.
- [224] C. Rovida and T. Hartung. Re-evaluation of animal numbers and costs for *in vivo* tests to accomplish REACH legislation requirements for chemicals - a report by the transatlantic think tank for toxicology. *Alternatives to animal experimentation*, 26 3:187–208, 2009.
- [225] W.M.S. Russell, R.L. Burch, and C.W. Hume. *The principles of humane experimental technique*, volume 238. Methuen London, 1959.

- [226] I. Rusyn, A. Sedykh, Y. Low, K.Z. Guyton, and A. Tropsha. Predictive Modeling of Chemical Hazard by Integrating Numerical Descriptors of Chemical Structures and Short-term Toxicity Assay Data. *Toxicological Sciences*, 127(1):1–9, 2012.
- [227] N. Ryan. A user’s guide for accessing and interpreting toxcast data. Retrieved from <https://lri.americanchemistry.com/Users-Guide-for-Accessing-and-Interpreting-ToxCast-Data.pdf>. 2017.
- [228] C. Rücker, G. Rücker, and M. Meringer. γ -Randomization and Its Variants in QSPR/QSAR. *Journal of Chemical Information and Modeling*, 47(6):2345–2357, 2007.
- [229] Y. Sakuratani, H.Q. Zhang, S. Nishikawa, K. Yamazaki, T. Yamada, J. Yamada, K. Gerova, G. Chankov, O. Mekenyan, and M. Hayashi. Hazard Evaluation Support System (HESS) for predicting repeated dose toxicity using toxicological categories, journal = SAR and QSAR in Environmental Research. 24(5):351–363, 2013.
- [230] K.T. Savjani, A.K. Gajjar, and J.K. Savjani. Drug Solubility: Importance and Enhancement Techniques. *ISRN Pharmaceutics*, 2012:1–10, 2012.
- [231] B. Scholkopf and A.J Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2001.
- [232] T.T. Schug, A.F. Johnson, L.S. Birnbaum, T. Colborn, L.J. Guillette, D.P. Crews, T. Collins, A.M. Soto, F.S. vom Saal, J.A. McLachlan, C. Sonnenschein, and J.J. Heindel. Minireview: Endocrine Disruptors: Past Lessons and Future Directions. *Molecular Endocrinology*, 30(8):833–847, 2016.
- [233] T.W. Schultz, T.I. Netzeva, and M.T.D. Cronin. Selection of data sets for qsars: Analyses of tetrahymena toxicity from aromatic compounds. *SAR and QSAR in Environmental Research*, 14(1):59–81, 2003.
- [234] A. Sedykh, H. Zhu, H. Tang, L. Zhang, A. Richard, I. Rusyn, and A. Tropsha. Use of *in Vitro* HTS-Derived Concentration–Response Data as Biological Descriptors Improves the Accuracy of QSAR Models of *in Vivo* Toxicity. *Environmental Health Perspectives*, 119(3):364–370, 2011.
- [235] I. Shah, K. Houck, R.S. Judson, R.J. Kavlock, M.T. Martin, D.M. Reif, J. Wambaugh, and D.J. Dix. Using Nuclear Receptor Activity to Stratify Hepatocarcinogens. *PLoS ONE*, 6(2):e14584, 2011.
- [236] L.A. Shalabi and Z. Shaaban. Normalization as a Preprocessing Engine for Data Mining and the Approach of Preference Matrix. In *2006 International Conference on Dependability of Computer Systems*, pages 207–214, 2006.
- [237] J. Shen, L. Xu, H. Fang, A.M. Richard, J.D. Bray, R.S. Judson, G. Zhou, T.J. Colatsky, J.L. Aungst, C. Teng, S.C. Harris, W. Ge, S.Y. Dai, Z. Su, A.C. Jacobs, W. Harrouk,

-
- R. Perkins, W. Tong, and H. Hong. EADB: An estrogenic activity database for assessing potential endocrine activity. *Toxicological Sciences*, pages 277–291, 2013.
- [238] R.P. Sheridan. Global Quantitative Structure–Activity Relationship Models vs Selected Local Models as Predictors of Off-Target Activities for Project Compounds. *Journal of Chemical Information and Modeling*, 54(4):1083–1092, 2014.
- [239] R.P. Sheridan, B.P. Feuston, V.N. Maiorov, and S.K. Kearsley. Similarity to molecules in the training set is a good discriminator for prediction accuracy in QSAR. *J. Chem. Inf. Comput. Sci.*, 44(6):1912–1928, 2004.
- [240] B.W Silverman. *Density estimation for statistics and data analysis*. Routledge, 2018.
- [241] N.S. Sipes, M.T. Martin, P. Kothiya, D.M. Reif, R.S. Judson, A.M. Richard, K.A. Houck, D.J. Dix, R.J. Kavlock, and T.B. Knudsen. Profiling 976 ToxCast Chemicals across 331 Enzymatic and Receptor Signaling Assays. *Chemical Research in Toxicology*, 26(6):878–895, 2013.
- [242] N.S. Sipes, M.T. Martin, D.M. Reif, N.C. Kleinstreuer, R.S. Judson, A.V. Singh, K.J. Chandler, D.J. Dix, R.J. Kavlock, and T.B. Knudsen. Predictive Models of Prenatal Developmental Toxicity from ToxCast High-Throughput Screening Data. *Toxicological Sciences*, 124(1):109–127, 2011.
- [243] J. Smalley, T. Gant, and S. Zhang. Application of connectivity mapping in predictive toxicology based on gene-expression similarity. *Toxicology*, 268:143–146, 2009.
- [244] C. Sonich-Mullin, R. Fielder, J. Wiltse, K. Baetcke, J. Dempsey, P. Fenner-Crisp, D. Grant, M. Hartley, A. Knaap, D. Kroese, I. Mangelsdorf, E. Meek, J.M. Rice, and M. Younes. IPCS Conceptual Framework for Evaluating a Mode of Action for Chemical Carcinogenesis. *Regulatory Toxicology and Pharmacology*, 34(2):146 – 152, 2001.
- [245] T. Steger-Hartmann and F. Pognan. The eTOX Consortium: To Improve the Safety Assessment of New Drug Candidates. *Pharmazeutische Medizin*, 1:3–13, 2017.
- [246] T. Steger-Hartmann, F. Pognan, F. Sanz, C. Diaz, A. Sutter, and M. Pastor. *In silico* prediction of *in vivo* toxicity - the first steps of the eTox consortium. *Toxicology Letters*, 196:S250–S251, 2010.
- [247] C. Strobl, A.-L. Boulesteix, T. Kneib, T. Augustin, and A. Zeileis. Conditional Variable Importance for Random Forests. *BMC Bioinformatics*, 9(1):307, 2008.
- [248] A. Subramanian, R. Narayan, S.M. Corsello, D.D. Peck, T.E. Natoli, X. Lu, J. Gould, J.F. Davis, A.A. Tubelli, and J.K. *et al.* A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. *Cell*, 171(6):1437 – 1452.e17, 2017.

- [249] L. Sun, H. Yang, Y. Cai, W. Li, G. Liu, and Y. Tang. *In Silico* Prediction of Endocrine Disrupting Chemicals Using Single-Label and Multilabel Models. *Journal of Chemical Information and Modeling*, 59(3):973–982, 2019.
- [250] R.S Sutton and A.G Barto. *Introduction to reinforcement learning*, volume 135. MIT press Cambridge, 1998.
- [251] F. Svensson, U. Norinder, and A. Bender. Modelling compound cytotoxicity using conformational prediction and PubChem HTS data. *Toxicology Research*, 6(1):73–80, 2017.
- [252] V. Svetnik, A. Liaw, C. Tong, J.C. Culberson, R.P. Sheridan, and B.P. Feuston. Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. *Journal of Chemical Information and Computer Sciences*, 43(6):1947–1958, 2003.
- [253] O. Taboureau and K. Audouze. Human environmental disease network: A computational model to assess toxicology of contaminants. *Alternatives to animal experimentation*, 34(2):289–300, 2017.
- [254] J. Tang, S. Alelyani, and H. Liu. Feature selection for classification: A review. *Data Classification: Algorithms and Applications*, pages 37–64, 01 2014.
- [255] O. Tcheremenskaia, R. Benigni, I. Nikolova, N. Jeliaskova, S. Escher, M. Batke, T. Baier, V. Poroikov, A. Lagunin, M. Rautenberg, and B. Hardy. OpenTox predictive toxicology framework: Toxicological ontology and semantic media wiki-based OpenToxipedia. *Journal of biomedical semantics*, 3 Suppl 1:S7, 2012.
- [256] A. Teasdale. *ICH M7*, chapter 24, pages 667–699. John Wiley & Sons, Ltd, 2017.
- [257] I.V. Tetko, V.Y. Tanchuk, T.N. Kasheva, and A.E.P. Villa. Estimation of Aqueous Solubility of Chemical Compounds Using E-State Indices. *Journal for Chemical Information and Computer Scientists*, 41(6):1488–1493, 2001.
- [258] R. Thomas. The US Federal Tox21 Program: A Strategic and Operational Plan for Continued Leadership. *Alternatives to animal experimentation*, 35:163–168, 2018.
- [259] R.S. Thomas, M.B. Black, L. Li, E. Healy, T. Chu, W. Bao, M.E. Andersen, and R.D. Wolfinger. A Comprehensive Statistical Analysis of Predicting *In Vivo* Hazard Using High-Throughput *In Vitro* Screening. *Toxicological Sciences*, 128(2):398–417, 2012.
- [260] R.R. Tice, C.P. Austin, R.J. Kavlock, and J.R. Bucher. Improving the Human Hazard Characterization of Chemicals: A Tox21 Update. *Environmental Health Perspectives*, 121(7):756–765, 2013.
- [261] J.G. Topliss and R.J. Costello. Chance correlations in structure-activity studies using multiple regression analysis. *Journal of Medicinal Chemistry*, 15(10):1066–1068, 1972.

-
- [262] A. Tropsha. Best Practices for QSAR Model Development, Validation, and Exploitation. *Molecular Informatics*, 29(6-7):476–488, 2010.
- [263] T. Uehara, A. Ono, T. Maruyama, I. Kato, H. Yamada, Y. Ohno, and T. Urushidani. The Japanese toxicogenomics project: Application of toxicogenomics. *Molecular Nutrition & Food Research*, 54(2):218–227.
- [264] L. Van Der Maaten, E. Postma, and J. Van den Herik. Dimensionality reduction: a comparative. *Journal of Machine Learning Research*, 10:66–71, 2009.
- [265] L.N. Vandenberg, R.T. Zoeller, W.V. Welshons, J.P. Myers, T. Colborn, T.B. Hayes, J.J. Heindel, D.R.Jr. Jacobs, D.-H. Lee, A.M. Shioda, T. and Soto, and F.S. vom Saal. Hormones and Endocrine-Disrupting Chemicals: Low-Dose Effects and Nonmonotonic Dose Responses. *Endocrine Reviews*, 33(3):378–455, 2012.
- [266] W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Springer, New York, fourth edition, 2002.
- [267] D.L. Villeneuve, D. Crump, N. Garcia-Reyero, M. Hecker, T.H. Hutchinson, C.A. LaLone, B. Landesmann, T. Lettieri, S. Munn, M. Nepelska, M.A. Ottinger, L. Vergauwen, and M. Whelan. Adverse Outcome Pathway (AOP) Development I: Strategies and Principles. *Toxicological Sciences*, 142(2):312–320, 2014.
- [268] Y. Wang, S.H. Bryant, T. Cheng, J. Wang, A. Gindulyte, B.A. Shoemaker, P.A. Thiessen, S. He, and J. Zhang. PubChem BioAssay: 2017 update. *Nucleic Acids Research*, 45(D1):D955–D963, 2017.
- [269] Y. Wang, J. Xiao, T.O. Suzek, J. Zhang, J. Wang, and S. H. Bryant. PubChem: a public information system for analyzing bioactivities of small molecules. *Nucleic Acids Research*, pages D623–D633, 2009.
- [270] W.A Warr. Representation of chemical structures. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 1:557 – 579, 2011.
- [271] M. Waters, S. Stasiewicz, B. Alex Merrick, K. Tomer, P. Bushel, R. Paules, N. Stegman, G. Nehls, K.J. Yost, Johnson, and *et al.* CEBS Chemical Effects in Biological Systems: a public data repository integrating study design and toxicity data with microarray and proteomics data. *Nucleic Acids Research*, 36(Database):892–900, 2007.
- [272] S. Watford, A. Adrian, J. Wignall, J. Brown, and M. Martin. ToxRefDB 2.0: Improvements in Capturing Qualitative and Quantitative Data from *in vivo* Toxicity Studies, 2017.
- [273] D. Weininger. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences*, 28(1):31–36, 1988.

- [274] J. Wenzel, H. Matter, and F. Schmidt. Predictive Multitask Deep Neural Network Models for ADME-Tox Properties: Learning from Large Data Sets. *Journal of Chemical Information and Modeling*, 59(3):1253–1268, 2019.
- [275] C. Wittwehr, S. Munn, B. Landesmann, and M. Whelan. Adverse Outcome Pathways Knowledge Base (AOP-KB). *Toxicology Letters*, 238:S309, 2015.
- [276] S. Wold, K. Esbensen, and P. Geladi. Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 2(1):37 – 52, 1987.
- [277] D.C. Wolf, .M. Cohen, A.R. Boobis, V.L. Dellarco, P.A. Fenner-Crisp, A. Moretto, T.P. Pastoor, R.S. Schoeny, J.G. Seed, and J.E. Doe. Chemical carcinogenicity revisited 1: A unified theory of carcinogenicity based on contemporary knowledge. *Regulatory Toxicology and Pharmacology*, 103:86 – 92, 2019.
- [278] D.H. Wolpert. Stacked generalization. *Neural Networks*, 5(2):241–259, 1992.
- [279] D.H. Wolpert. The Lack of A Priori Distinctions Between Learning Algorithms. *Neural Computation*, 8(7):1341–1390, 1996.
- [280] A.P. Worth. *The Role of QSAR Methodology in the Regulatory Assessment of Chemicals*, pages 367–382. Springer Netherlands, Dordrecht, 2010.
- [281] L. Wu, Z. Liu, S. Auerbach, R. Huang, M. Chen, K. McEuen, J. Xu, H. Fang, and W. Tong. Integrating Drug’s Mode of Action into Quantitative Structure–Activity Relationships for Improved Prediction of Drug-Induced Liver Injury. *Journal of Chemical Information and Modeling*, 57(4):1000–1006, 2017.
- [282] Q. Xu and Y. Liang. Monte Carlo Cross Validation. *Chemometrics and Intelligent Laboratory Systems*, 56:1–11, 2001.
- [283] L. Yang, J. Rathman, C. Yang, K. Arvidson, M. Cronin, S. Enoch, D. Hristozov, Y. Lan, J. Madden, D. Neagu, A. Richarz, M. Ridley, O. Sacher, and C. Schwab. Development of a COSMOS DB to support *in silico* modelling for cosmetics ingredients and related chemicals. *The Toxicologist – A Supplement to Toxicological Sciences*, pages 132–185, 2013.
- [284] P. Yang, Y. Hwa Yang, B.B Zhou, and A.Y Zomaya. A Review of Ensemble Methods in Bioinformatics. *Current Bioinformatics*, 5(4):296–308, 2010.
- [285] C.W. Yap. PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. *Journal of Computational Chemistry*, 32(7):1466–1474, 2011.
- [286] P. Yazgana and A. Kusakci. A Literature Survey on Association Rule Mining Algorithms. *Southeast Europe Journal of Soft Computing*, 5:1859, 2016.
- [287] S. Yen and Y. Lee. Cluster-based Under-sampling Approaches for Imbalanced Data Distributions. *Expert Systems with Applications*, 36:5718–5727, 2006.

-
- [288] Y. Yin and E. Gelenbe. Single-cell based random neural network for deep learning. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 86–93, 2017.
- [289] C. Ying, M. Qi-Guang, L. Jia-Chen, and G. Lin. Advance and prospects of AdaBoost algorithm. *Acta Automatica Sinica*, 39(6):745–758, 2013.
- [290] Q. Zang, D.M. Rotroff, and R.S. Judson. Binary Classification of a Large Collection of Environmental Chemicals from Estrogen Receptor Assays by Quantitative Structure–Activity Relationship and Machine Learning Methods. *Journal of Chemical Information and Modeling*, 53(12):3244–3261, 2013.
- [291] C. Zhang, F. Cheng, W. Li, G. Liu, P.W. Lee, and Y. Tang. *In silico* Prediction of Drug Induced Liver Toxicity Using Substructure Pattern Recognition Method. *Molecular Informatics*, 35(3-4):136–144, 2016.
- [292] S. Zhang, C. Zhang, and Q. Yang. Data Preparation for Data Mining. *Applied Artificial Intelligence*, 17:375–381, 2003.
- [293] Y. Zhang, C.A. Phillips, G.L. Rogers, E.J. Baker, E.J. Chesler, and M.A. Langston. On finding bicliques in bipartite graphs: a novel algorithm and its application to the integration of diverse biological data types. *BMC Bioinformatics*, 15(1):110, 2014.
- [294] Y. Zhang, Y. Yin, D. Guo, X. Yu, and L. Xiao. Cross-validation based weights and structure determination of Chebyshev-polynomial neural networks for pattern classification. *Pattern Recognition*, 47(10):3414 – 3428, 2014.
- [295] H. Zhu, L. Ye, A. Richard, A. Golbraikh, F.A Wright, I. Rusyn, and A. Tropsha. A Novel Two-Step Hierarchical Quantitative Structure–Activity Relationship Modeling Work Flow for Predicting Acute Toxicity of Chemicals in Rodents. *Environmental health perspectives*, 117:1257–1264, 2009.
- [296] H. Zhu, J. Zhang, M.T. Kim, A. Boison, A. Sedykh, and K. Moran. Big Data in Chemical Toxicity Research: The Use of High-Throughput Screening Assays To Identify Potential Toxicants. *Chemical Research in Toxicology*, 27(10):1643–1651, 2014.
- [297] X. Zhu, A. Sedykh, and S. Liu. Hybrid *in silico* models for drug-induced liver injury using chemical descriptors and *in vitro* cell-imaging information. *Journal of applied toxicology*, 34:281–288, 2014.
- [298] J. Ziang. KNN approach to unbalanced data distributions: a case study involving information extraction. *Proceedings of the International Conference on Machine Learning*, 126, 2003.