

Department of  
Earth and Environmental Sciences

PhD program in Chemical, Geological and Environmental Sciences, Cycle XXXIV  
Curriculum in Chemical Sciences

## **Advancing the prediction of Nuclear Receptor modulators through machine learning methods**

Valsecchi Cecile

748094

Tutor: Prof. Marco Orlandi

Supervisor: Prof. Davide Ballabio

Co-supervisor: Prof. Laura Bonati, Dr. Viviana Consonni

Coordinator: Prof. Marco Giovanni Malusa'



*“...when the brain is released from the constraints of reality, it can generate any sound, image, or smell in its repertoire, sometimes in complex and “impossible” combinations.”*

Oliver Sacks



# Abstract

Cecile VALSECCHI

*Advancing the prediction of Nuclear Receptor modulators through machine learning methods*

Nuclear receptors are transcription factors involved in processes critical to human health and are a relevant target for toxicological risk assessment and the drug discovery process. Computational models can be a useful tool (i) to prioritize chemicals that can mimic natural hormones and thus be endocrine disruptors and (ii) to identify new possible lead for drug discovery.

Therefore, the main goal of this project is to study potential interactions between chemicals and nuclear receptors, with the dual purpose of developing *in silico* tools to search for new modulators and to identify possible endocrine disrupting chemicals.

After creating an exhaustive collection of nuclear receptor modulators, we applied machine learning methods to fill the data gap and prioritize modulators by building predictive models. In particular, modeling strategies included multi-tasking machine learning algorithms to investigate the complex relationships between chemicals and multiple nuclear receptors.



# Acknowledgements

First of all, I would like to express my sincere gratitude to my supervisor Prof. Davide Ballabio for giving me the opportunity to undertake this PhD project, being always open to discuss the project with me and giving me useful advice.

In addition, I would like to thank Prof. Roberto Todeschini, Dr. Viviana Consonni and Prof. Laura Bonati who supported and reviewed my work and involved me in other interesting activities.

I am also grateful to Dr. Francesca Grisoni especially for her guidance in the world of deep learning and medicinal chemistry and for her inspiring advice, and to Schneider's MODLAB group at ETH for their hospitality and stimulating work environment.

I would like to thank all the other people I have met in the lab over the years, who have contributed on a working and social level: Giacomo, Stefano, Veronica, Magda, Martina C. and Martina B. and my family and friends for all the unconditional support during this very intense academic year.

To conclude, I cannot forget to thank Gabriele for the scientific discussions, emotional support and for believing in me.



# Contents

<b>Abstract</b>	<b>v</b>
<b>Acknowledgements</b>	<b>vii</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Nuclear Receptors . . . . .	1
1.1.1 Nuclear Receptors modulation . . . . .	1
1.1.2 Nuclear Receptors classification . . . . .	3
1.1.3 Nuclear receptors as targets in medicinal chemistry . . . . .	7
1.1.4 Nuclear receptors as targets for endocrine disruptors . . . . .	7
1.2 Machine learning and nuclear receptors . . . . .	9
1.2.1 Machine learning for drug design and virtual screening . . . . .	10
1.2.2 Machine learning for prioritizing modulators . . . . .	11
1.3 Objectives . . . . .	12
<b>2 Tools and methods</b>	<b>13</b>
2.1 Molecular structure encoding . . . . .	13
2.1.1 SMILES . . . . .	13
2.1.2 Molecular descriptors . . . . .	15
Fingerprints . . . . .	18
WHALES . . . . .	20
2.2 Traditional QSAR classification methods . . . . .	25
2.2.1 k-Nearest Neighbours and N-Nearest Neighbours . . . . .	25
2.2.2 Random Forest . . . . .	27
2.2.3 Naïve Bayes . . . . .	28
2.2.4 Applicability domain . . . . .	29
2.3 Artificial neural networks . . . . .	30
2.3.1 Feedforward neural networks . . . . .	31
2.3.2 Graph convolutional networks . . . . .	32
2.3.3 Multi-task learning . . . . .	33
2.3.4 Parameters tuning . . . . .	34
2.4 Model validation . . . . .	36
2.5 Consensus modelling . . . . .	37
2.5.1 Majority Voting . . . . .	38
2.5.2 Bayesian Consensus . . . . .	39
2.6 Metrics for classification performances . . . . .	40
2.7 Software . . . . .	43

<b>3 Data</b>	<b>45</b>
3.1 CoMPARA dataset . . . . .	46
3.2 NURA dataset . . . . .	47
3.2.1 Target selection . . . . .	47
3.2.2 Data collection . . . . .	48
3.2.3 Data curation . . . . .	49
3.2.4 Data analysis . . . . .	51
3.2.5 Data driven insights . . . . .	54
3.3 Summary of results and concluding remarks . . . . .	58
<b>4 Consensus analysis of CoMPARA models</b>	<b>59</b>
4.1 Individual QSAR Models . . . . .	60
4.2 Analysis of Consensus Strategies . . . . .	62
4.2.1 Consensus Based on Subsets of Models . . . . .	68
4.3 Summary of results and concluding remarks . . . . .	70
<b>5 Multi-task modelling to predict nuclear receptor modulators</b>	<b>73</b>
5.1 Data curation . . . . .	75
5.2 Parameters tuning . . . . .	75
5.3 Comparison on individual tasks . . . . .	80
5.4 Multi-task vs single-task modelling . . . . .	81
5.4.1 Comparison on global performance . . . . .	83
5.4.2 Performance on the evaluation set . . . . .	84
5.5 Summary of results and concluding remarks . . . . .	87
<b>6 Pseudo multi-task modelling of ligand-receptor pairs</b>	<b>89</b>
6.1 Data . . . . .	91
6.2 Ligand-receptor pair . . . . .	92
6.2.1 Optimization . . . . .	92
6.2.2 Applicability domain . . . . .	94
6.3 Pseudo multi-task results . . . . .	94
6.4 Summary of results and concluding remarks . . . . .	97
<b>7 Conclusions</b>	<b>99</b>
<b>A Parameters tuning</b>	<b>101</b>
<b>B Publications and Conferences</b>	<b>105</b>
<b>C Deliverables</b>	<b>107</b>
<b>Bibliography</b>	<b>109</b>

# List of Figures

1.1 Example of dose-response curve . . . . .	3
1.2 Types of nuclear receptors . . . . .	5
1.3 Tissue-specific protein expression of nuclear receptors . . . . .	8
1.4 Workflow of a drug approval. . . . .	11
2.1 Examples of SMILES . . . . .	14
2.2 Molecular descriptors . . . . .	17
2.3 ECFP example for Tamoxifen . . . . .	19
2.4 Overview of the WHALES concept . . . . .	23
2.5 Toy example of kNN . . . . .	26
2.6 Toy example of a decision tree classifier . . . . .	27
2.7 Toy example of applicability domain . . . . .	30
2.8 Toy example of a feed forward neural network . . . . .	31
2.9 Comparison of single-task and multi-task data . . . . .	34
2.10 Toy example of optimization methods . . . . .	36
2.11 QSAR development pipeline . . . . .	37
2.12 Example of majority voting assignation . . . . .	39
3.1 Analysis of the individual sources used to develop NURA dataset	53
3.2 Multidimensional scaling of the molecules in the curated NURA dataset . . . . .	54
3.3 Distribution of molecules in NURA dataset . . . . .	55
3.4 Summary of the NURA data-driven analysis . . . . .	57
4.1 Violin plots of CoMPARA models performances . . . . .	61
4.2 Plot of sensitivities ( $S_n$ ) versus specificities ( $S_p$ ) for the individual CoMPARA models . . . . .	63
4.3 MDS for CoMPARA models and consensus strategies. . . . .	66
4.4 NER and coverage as a function of the number of CoMPARA models included in the consensus calculation. . . . .	69
5.1 Schematic representation of a multitask neural network . . . . .	74
5.2 Relative frequency of network parameters in the chromosomes of the final GA population . . . . .	78
5.3 PCA biplot of relative frequencies of network parameters . . . . .	79
5.4 Analysis of classification performance on individual tasks . . . . .	82
5.5 P values of the paired Wilcoxon signed-rank test performed on each pair of classification approaches . . . . .	84
5.6 Multidimensional scaling calculated over the evaluation set coloured according to the prediction errors . . . . .	86

6.1 Workflow of the pseudo-multitask approach . . . . .	90
6.2 Normalized sum of reciprocal ranks for the first fifty consensus combinations . . . . .	95
6.3 Performances of "pseudo" multi-task approaches . . . . .	96
A.1 Non-Error Rate of the best solution for each optimization strategy . . . . .	103
C.1 Screenshot of Zenodo website. . . . .	108

# List of Tables

1.1 Name, abbreviation and NRCN grouping for the 48 human nuclear receptors . . . . .	5
2.1 Summary of PDBbind crystallographic structures . . . . .	24
2.2 Binary classification confusion matrix . . . . .	41
3.1 Summary of CoMPARA evaluation sets . . . . .	46
3.2 Summary of nuclear receptors information . . . . .	48
3.3 Summary of the considered data sources for NURA dataset creation . . . . .	50
4.1 Classification Performance of the Consensus Approaches for Binding, Agonism, and Antagonism Endpoints . . . . .	65
4.2 Summary of the misclassified CoMPARA molecules . . . . .	68
4.3 Classification Performance of the Consensus Approaches for Binding, Agonism, and Antagonism Endpoints considering 5 models . . . . .	70
5.1 Dataset description: number of molecules and class distributions among the tasks for the training and test sets . . . . .	76
5.2 Classification performance ( $NER_t$ ) on the test set for each task . . . . .	80
5.3 Global classification measures on the external evaluation set . . . . .	85
6.1 Pseudo multi-task data . . . . .	91
6.2 Summary of the ten developed models . . . . .	93
A.1 Summary information of the considered multi-task datasets . . . . .	102
A.2 Summary of parameters to be optimized . . . . .	102
A.3 Results in terms of overall Non-error Rate for each optimization strategy . . . . .	103



# List of Abbreviations

<b>AD</b>	Applicability Domain
<b>ANN</b>	Artificial Neural Network
<b>AR</b>	Androgen Receptor
<b>ECFPs</b>	Extended Connectivity FingerPrints
<b>ER</b>	Estrogen Receptor
<b>FFNN</b>	FeedForward Neural Network
<b>FPs</b>	FingerPrints
<b>GCN</b>	Graph Convolutional Network
<b>GR</b>	Glucocorticoid Receptor
<b>kNN</b>	k Nearest Neighbours
<b>MD</b>	Molecular Descriptors
<b>MDS</b>	Multi Dimensional Scaling
<b>MTL</b>	Multi-Task Learning
<b>NB</b>	Naïve Bayes
<b>NRs</b>	Nuclear Receptors
<b>PDB</b>	Protein Data Bank
<b>PPAR</b>	Peroxisome Proliferator-Activated Receptor
<b>PR</b>	Progesterone Receptor
<b>PXR</b>	Pregnane X Receptor
<b>QSAR</b>	Quantitative Structure Activity Relationship
<b>RF</b>	Random Forest
<b>RXR</b>	Retinoid X Receptor
<b>WHALES</b>	Weighted Holistic Atom Localization and Entity Shape



To my dad



# Chapter 1

## Introduction

### 1.1 Nuclear Receptors

Deoxyribonucleic acid (DNA) is a double helix of two polynucleotide chains carrying genetic instructions (i.e., genes) for the development, functioning, growth and reproduction of all known organisms and many viruses. A transcription factor is a protein that regulates the expression of specific genes by binding to a specific DNA sequence. The biggest family of human transcription factors is constituted by nuclear receptors (NRs) whose action depends on ligand binding, in other words to explicate their action a formation of ligand-receptor complex is needed.

#### 1.1.1 Nuclear Receptors modulation

There are several types of modulators of nuclear receptors, including specific ligands and inhibitors of interactions of nuclear receptors with various proteins related to the transcription (Gronemeyer, Gustafsson, and Laudet, 2004; Mangelsdorf et al., 1995). This thesis will focus only on NRs ligand modulators, which can modulate the NRs activities in different ways. Usually they bind directly to the receptor in a so-called binding pocket (orthosteric binding) or on the external surface (allosteric binding) (Christopoulos, 2002). The formation of a ligand-receptor complex (LR) is an equilibrium process (Ariens et al., 1954). Ligand (L) binds to the receptor (R) and dissociates from it according to the following equation. The brackets denote concentrations.

$$[L] + [R] \xrightleftharpoons{K_d} [LR] \quad (1.1)$$

The binding causes conformational changes and, thus, triggers a biological effect (Germain et al., 2006). Depending on the resulting activity a binder can be identified as agonist or antagonist. In the former case the formation of the ligand-receptor complex activates the transcription to produce a biological response (Germain et al., 2006). On the contrary, by binding to the receptor, an antagonist blocks or dampens a biological response. Competitive antagonists bind to receptors at the same binding site as the endogenous ligand or agonist, but without activating the receptor. In this case the inhibition is the result of the competition for the same binding site on the receptor. In other words, once bound, a competitive antagonist will block agonist binding (Germain et al., 2006). A non-competitive antagonist may bind to an

allosteric site of the receptor, or irreversibly bind to the active site of the receptor, inducing a conformational change and, thus, preventing the agonist binding (Burris et al., 2013).

The distinction between agonism and antagonism is multi-faceted. For example the so-called partial agonists can act as a competitive antagonist in the presence of a full agonist by exerting a lower expression occupying the binding site. In addition, some modulators display tissue and/or target gene specificity in terms of their agonist, antagonist, or inverse agonist activity (i.e., selective receptor modulators) (Burris et al., 2013). Some small molecules, can be agonist in a tissue and antagonist in another tissue. Most studied examples include Estrogen receptor agonists/antagonists (ERAs) (Begam, Jubie, and Nanjan, 2017).

Nowadays it is possible to rapidly identify NRs ligands throughout specific experimental tests. In particular, high-throughput screening (HTS) consists in quickly and simultaneously testing of thousands of compounds using a battery of *in vitro* assays. *In vitro* HTS assays are faster and more cost-effective than traditional *in vivo* toxicity testing, and they avoid the ethical concerns associated with animal tests (Rotroff et al., 2013).

HTS assays can be divided broadly into two categories: biochemical assays and cell-based assays. Biochemical assays are direct and specific to the target of interest while cell-based assays assess the efficacy of compounds in a cellular environment. Although the former provides robuster and more reproducible results, the latter takes into account cellular-related requirements such as cellular cofactors, membrane permeability, off-target effects and cytotoxicity (An and Tolliday, 2010). Examples of biochemical assays include fluorescent polarization of a probe linked to a specific ligand, nuclear magnetic resonance or surface plasmon resonance investigate the direct interaction between ligand and receptor and consequent conformational changes (Ishigami-Yuasa and Kagechika, 2020). On the other hand, cell-based assays are affected by the choice of the biological system (primary cell, native, or engineered cell-line, model organism); the type of approach (functional, reporter gene) and the assay readout modality (uniform well readout or high content) (An and Tolliday, 2010).

Depending on the assays different types of resulting experimental readouts exist, the most used are: the half maximal concentration on the dose-response curve for inhibition or effect ( $IC_{50}$  and  $EC_{50}$ ) and the dissociation and inhibition constants ( $K_d$  and  $K_i$ ).  $IC_{50}$  and  $EC_{50}$  are the concentration of ligand causing 50% of displacement or response (Sebaugh, 2011). Dissociation constant  $K_d$  is an equilibrium constant that measures the dissociation of a complex into its components equilibrium as in Equation 1.1. The inhibitory constant  $K_i$ , on the other hand, describes the binding affinity between an inhibitor and its corresponding protein, which essentially also represent a dissociation constant. The difference between  $K_d$  and  $K_i$  is that  $K_d$  is a more general, all-encompassing term, whilst  $K_i$  is more narrowly used to indicate the dissociation equilibrium constant of the protein-inhibitor complex (Berg, Tymoczko, and Stryer, 2002).

A dose-response curve describes the magnitude of a biological response

(Y-axis) as a function of exposure to a certain amount of a chemical (X-axis). Figure 1.1 shows an example of a dose-response curve.

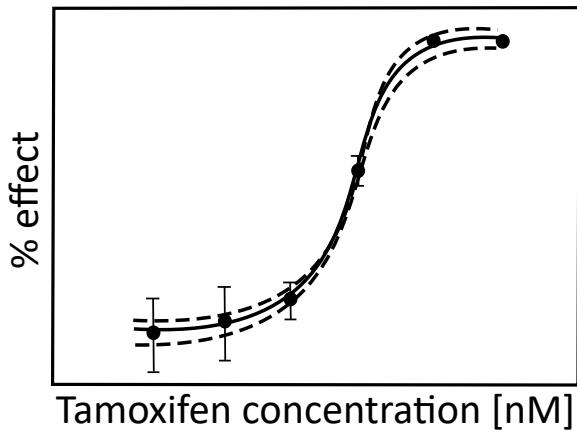


FIGURE 1.1: Example of dose-response curve.

### 1.1.2 Nuclear Receptors classification

The human NRs superfamily comprises 48 members which are evolutionarily and structurally related. In spite of the remarkable structural similarity, these proteins regulate an extremely large number of biological processes. The greatest homology is preserved in the amino acid sequence of the DNA-binding domain (C) and the ligand-binding domain (E) (Figure 1.2c) (Mangelsdorf et al., 1995). These C and E domains are responsible for the association of the transcription factor with specific DNA sequences and the binding of small ligands, usually of lipophilic nature, respectively. A third, non-conserved N-terminal domain named the regulatory domain (A/B) shows variable length and sequence in the different family members and is recognized by coactivators and/or other transcription factors. The ability of the E domain to activate transcription is controlled by the C-terminal helix 12, also termed F (Mangelsdorf et al., 1995).

In the 1950s, it was thought that steroid hormones enter the cells by simple diffusion through the plasma membrane and trigger series of metabolic oxidations and reductions. At the IV International Congress of Biochemistry in Vienna (1958), a new concept of hormone receptor arised when Dr. Elwood Jensen proved that tritiated estrogens bind to a receptor within the cell without chemical changes (Jensen et al., 1968). Then, this hormone-receptor complex must translocate to the cell nucleus and regulates the expression of specific genes (Mazaira et al., 2018).

The estrogen receptor (ER) was indeed the first member of the NR family to be identified biochemically (Jensen, 1962). Since then, study on NRs have increased leading to their identification as a superfamily of transcription factors, and steroid receptors were grouped as a subfamily (Mazaira et al., 2018).

The group of non-steroidal receptors was also added to the family comprising the thyroid hormone receptors (TR), the retinoic-acid receptors (RAR).

A group of receptor whose endogenous ligands were unknown, were grouped as the orphan receptor subfamily.

Nuclear receptor genes are encoded and expressed from the simplest to the most complex organisms of the animal kingdom. More than 900 nuclear receptors genes have been identified in different animals, and it appears that the number of receptors increases with the functional complexity of the organisms, reaching forty-nine members in mammals. However, nuclear receptors are absent in fungi, plants and also in the closest known relatives of metazoans, i.e. eukaryotes of the Choanoflagellatea class. Hence, it is thought that these receptors appeared on the scene of evolution about 635 million years ago with the first metazoans and played a key role during the Cambrian explosion of life forms nearly 540 million years ago (Mazaira et al., 2018).

The nuclear receptor superfamily can be generally divided into the following four groups of unequal size based on their DNA-binding properties and dimerisation preferences (Novac and Heinzel, 2004):

- Type I: ligand-receptor complex formation take place in the cytosol by the dissociation of heat shock proteins. Then the receptor homodimerizes, translocates into the nucleus and binds to specific sequence of DNA (i.e., response elements). Type I nuclear receptors include steroid hormone receptors (Figure 1.2a).
- Type II: receptors located in the nucleus regardless of the ligand binding status, which bind to DNA as heterodimers (usually with RXR)(Figure 1.2b).
- Type III: receptors similar to type I receptors which bind to DNA as homodimers. However, type III nuclear receptors, bind primarily to direct repeat instead of inverted repeat response elements.
- Type IV: receptors which bind either as monomers or dimers, but only a single DNA binding domain of the receptor binds to a single half site response element.

Furthermore, a new phylogeny-based nomenclature approved by the Nuclear Receptor Nomenclature Committee (NRNC) has been proposed for nuclear receptors. This nomenclature system is based on the multiple alignment procedures, phylogenetic tree reconstruction methods and other evolutionary implications, and subdivides the nuclear receptor superfamily into seven subfamilies which are numbered from 0 to 6. The phylogenetically closest members of each subfamily are combined into groups designated by capital letters arranged in the alphabetical order and the individual genes within each group are defined by Arabic numerals as reported in Table 1.1.

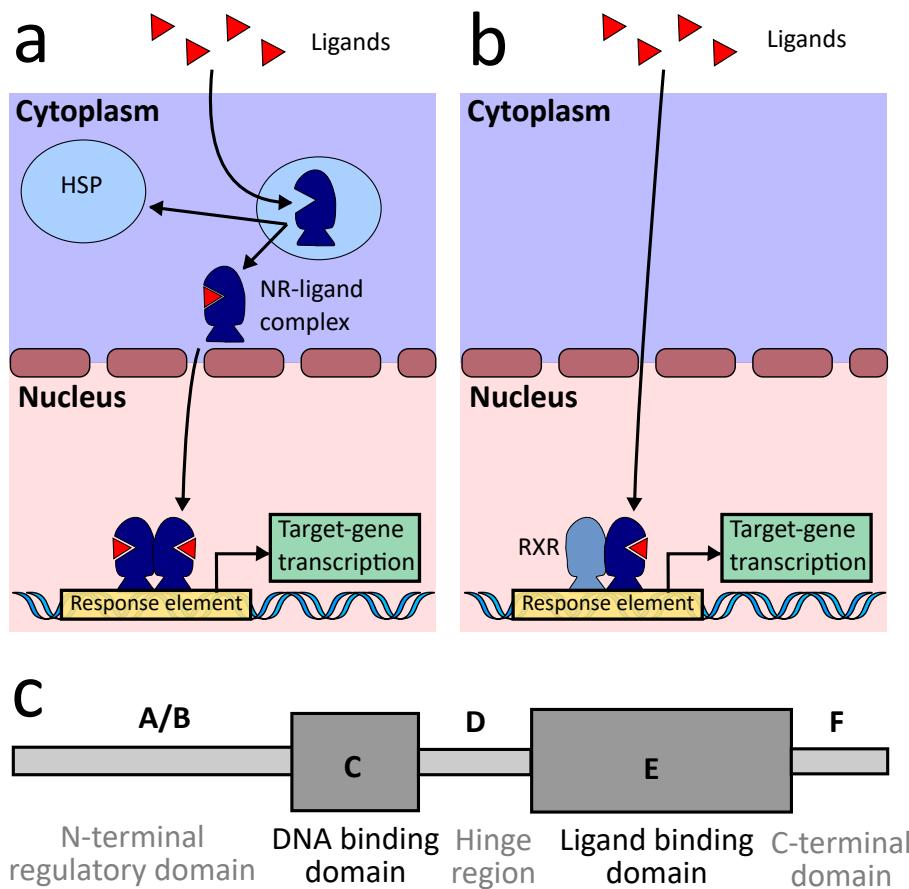


FIGURE 1.2: (a) Type I steroid nuclear receptors are synthesized in inactive forms associated with heat-shock protein (HSP) complexes in the cytoplasm. Direct hormone binding causes a conformational change, dissociation from HSP complexes and translocation into the nucleus. (b) Type II heterodimeric nuclear receptors bind constitutively to DNA with RXRs as obligate partners. Ligand binding causes a conformational change, dissociation of co-repressor complexes and recruitment of co-activators. (c) Members of the nuclear receptor superfamily have a common domain structure consisting of an amino-terminal activation domain (A/B), a DNA-binding domain (C), a hinge region (D), a ligand binding domain (E) and a carboxy-terminal domain (F).

TABLE 1.1: Summary of 48 human nuclear receptors divided according to the Nuclear Receptor Nomenclature Committee (NRNC) in group and subgroup (Sub.).

Name	Acronym	NRNC	Group	Sub.
DAX Short heterodimeric partner	DAX1 SHP	NR0B1 NR0B2	B	0

Thyroid hormone receptor $\alpha$	TR $\alpha$	NR1A1	A	1
Thyroid hormone receptor- $\beta$	TR $\beta$	NR1A2		
Retinoic acid receptor- $\alpha$	RAR $\alpha$	NR1B1	B	
Retinoic acid receptor- $\beta$	RAR $\beta$	NR1B2		
Retinoic acid receptor- $\gamma$	RAR $\gamma$	NR1B3		
PPAR- $\alpha$	PPAR $\alpha$	NR1C1	C	
PPAR- $\delta$	PPAR $\delta$	NR1C2		
PPAR- $\gamma$	PPAR $\gamma$	NR1C3		
Reverse-Erb- $\alpha$	REV-ERB $\alpha$	NR1D1	D	
Reverse-Erb- $\beta$	REV-ERB $\beta$	NR1D2		
Retinoic acid-related orphan- $\alpha$	ROR $\alpha$	NR1F1	F	
Retinoic acid-related orphan- $\beta$	ROR $\beta$	NR1F2		
Retinoic acid-related orphan- $\gamma$	ROR $\gamma$	NR1F3		
Farnesoid X receptor- $\alpha$	FXR $\alpha$	NR1H4	H	
Farnesoid X receptor- $\beta$	FXR $\beta$	NR1H5		
Liver X receptor- $\alpha$	LXR $\alpha$	NR1H3		
Liver X receptor- $\beta$	LXR $\beta$	NR1H2		
Vitamin D receptor	VDR	NR1I1	I	
Pregnane X receptor	PXR	NR1I2		
CAR	NR1I3	NR1I3		
Hepatocyte nuclear Factor-4- $\alpha$	HNF4 $\alpha$	NR2A1	A	2
Hepatocyte nuclear Factor-4- $\gamma$	HNF4 $\gamma$	NR2A2		
Retinoid X receptor- $\alpha$	RXR $\alpha$	NR2B1	B	
Retinoid X receptor- $\beta$	RXR $\beta$	NR2B2		
Retinoid X receptor- $\gamma$	RXR $\gamma$	NR2B3		
Testicular Receptor 2	TR2	NR2C1	C	
Testicular Receptor 4	TR4	NR2C2		
TLX	TLX	NR2E1	E	
PNR	PNR	NR2E2		
COUP-TF $\alpha$	COUP-TF $\alpha$	NR2F1	F	
COUP-TF $\beta$	COUP-TF $\beta$	NR2F2		
COUP-TF $\gamma$	COUP-TF $\gamma$	NR2F6		
Estrogen receptor- $\alpha$	ER $\alpha$	NR3A1	A	3
Estrogen receptor- $\beta$	ER $\beta$	NR3A2		
Estrogen-related receptor- $\alpha$	ERR $\alpha$	NR3B1	B	
Estrogen-related receptor- $\beta$	ERR $\beta$	NR3B2		
Estrogen-related receptor- $\gamma$	ERR $\gamma$	NR3B3		
Androgen receptor	AR	NR3C4	C	
Glucocorticoid receptor	GR	NR3C1		
Mineralocorticoid receptor	MR	NR3C2		
Progesterone receptor	PR	NR3C3		
Nerve growth Factor 1B	NGF1-B	NR4A1	A	4
Nurr-related Factor 1	NURR1	NR4A2		
NOR-1	NOR-1	NR4A3		
Steroidogenic Factor 1	SF-1	NR5A1	A	5
Liver receptor Homolog-1	LRH-1	NR5A2		
Germ cell nuclear factor	GCNF	NR6A1	A	6

### 1.1.3 Nuclear receptors as targets in medicinal chemistry

Since NRs exert a key role in physiological processes such as homeostasis, metabolism and development, they have become very attractive drug target. It is estimated that 15% of drug targets belong to the NRs superfamily (Santos et al., 2017). Drugs that target nuclear receptors are widely used and commercially successful also because of their roles in the development and progression of several diseases, such as carcinogenesis for steroid hormones receptors (Honma, Matsuda, and Mikami, 2021). In particular, the functional status of Androgen Receptor (AR) is an important mediator of prostate cancer progression (Heinlein and Chang, 2004). Estrogen Receptor alpha (ER $\alpha$ ) and Progesterone Receptor (PR), have been shown to play a major role in breast cancer development and progression (Kittler et al., 2013). Indeed, both receptors are used to classify breast cancers and to predict response to specific therapies. To treat breast cancer, ER $\alpha$  is thus one of the main targets, through inhibitors, such as tamoxifen.

In addition, nuclear receptors widely present in human tissues (Figure 1.3) such as Farnesoid X Receptor (FXR), Retinoid X Receptor (RXR), Glucocorticoid Receptor (GR) and Peroxisome proliferator-activated receptors (PPARs) are involved in the development and treatment of several diseases including cancers, cardiovascular and metabolic diseases.

For example, bexarotene and alitretinoin (RXRs), fibrates (PPAR $\alpha$ ), and thiazolidinediones (PPAR $\gamma$ ) are drugs approved for treating cancer, hyperlipidemia, and type 2 diabetes, respectively (Kittler et al., 2013), (Dhiman, Bolt, and White, 2018), (Dixon et al., 2021), (Shao et al., 2021).

Furthermore, Pregnane X Receptor (PXR) is used routinely to screen all new drug candidates for potentially dangerous drug-drug interactions in the pharmaceutical industry.

### 1.1.4 Nuclear receptors as targets for endocrine disruptors

Endocrine disrupting chemicals (EDCs) or endocrine disruptors are chemicals that are suspected to cause adverse effects in the endocrine system by mimicking endogenous hormone activity and, therefore, interfering with their synthesis, transport, degradation or action. Since Nuclear Receptors are involved in several physiological processes, many EDCs interfere directly or indirectly with them causing dysfunctional NRs signaling which often leads to proliferative, reproductive, and metabolic diseases, including hormonal cancers, infertility, obesity, or diabetes.

Human body is exposed to endocrine disruptors through different exogenous sources including dietary lipids and vitamins, pharmaceutical agents, plant-derived compounds and industrial byproducts (Hall and Greco, 2020).

The group of EDCs is highly heterogeneous and comprises compounds that are often distantly related to endogenous ligands in terms of size or chemical structure (Balaguer, Delfosse, and Bourguet, 2019). This is the case of tributyltin which, in spite of being structurally dissimilar from the endogenous ligand 9-cis retinoic acid, can activate RXR at nanomolar concentrations. Tributyltin occupies only a small part of the ligand binding pocket

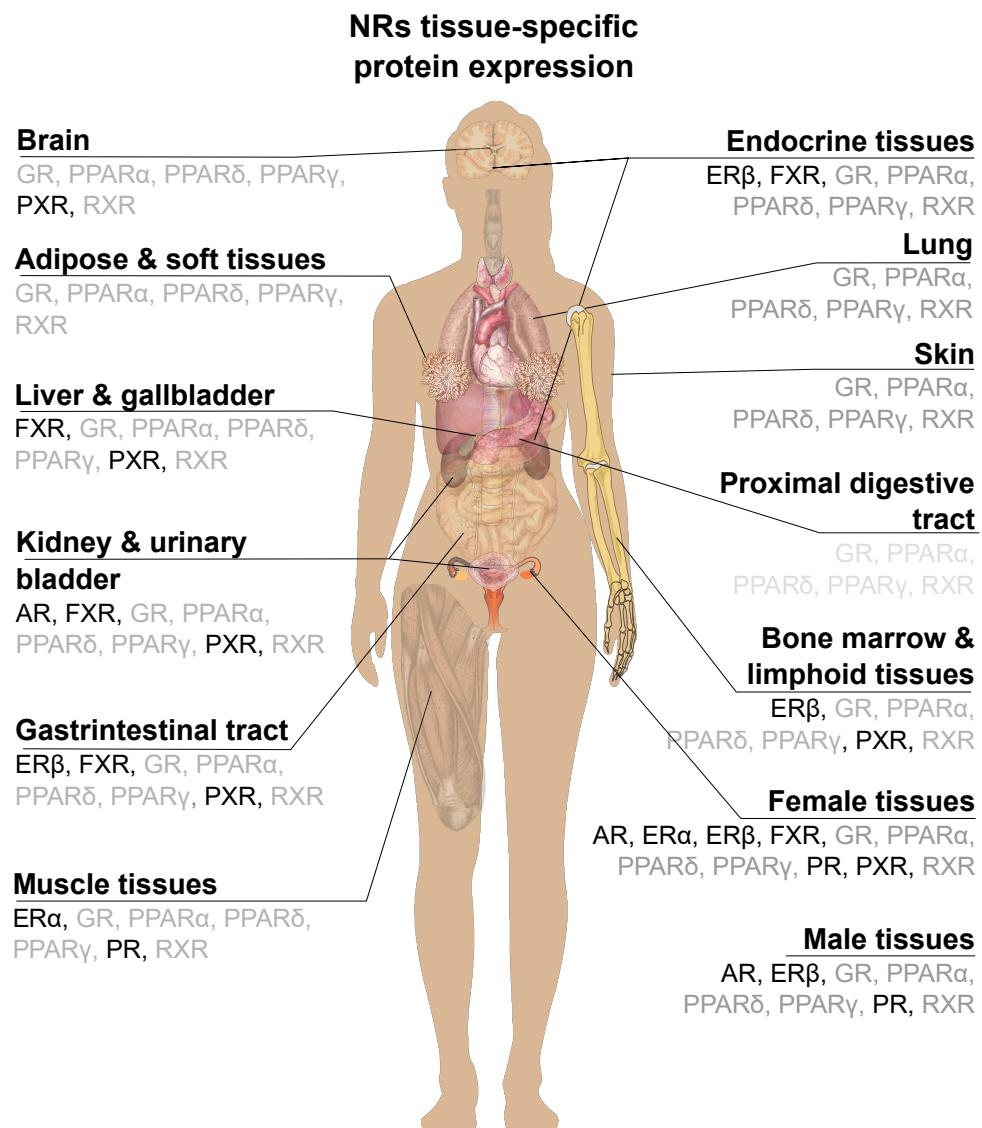


FIGURE 1.3: Maps of tissue with specific protein expression of nuclear receptors. Nuclear receptors present in all the highlighted tissues (i.e., GR, PPAR $\alpha$ , PPAR $\gamma$ , PPAR $\delta$  and RXR) are reported in grey. (Atlas, Accessed: 2021-06-10)

in comparison with 9-cis retinoic acid and is able to form a covalent bond between its tin atom and the sulfur atom of conserved cysteine of the ligand binding pocket (Toporova and Balaguer, 2020).

The most known effects of EDCs are related to steroid hormone receptors (e.g. estrogen and androgen, thyroid hormone receptors). However, there is emerging evidence that interactions of EDCs with other NRs, may coincide with chronic diseases such as obesity and type II diabetes/metabolic syndrome (Hall and Greco, 2020).

The pharmaceutical diethylstilbestrol (DES) provides an other example of endocrine disruption. Prenatal exposure to DES, used in the 1970s to prevent

miscarriage in women with high risk pregnancies, was linked with the development of vaginal cancer and its toxic effects were subsequently attributed to its interaction with estrogen receptors (ERs) (Le Maire, Bourguet, and Balaguier, 2010).

## 1.2 Machine learning and nuclear receptors

As mentioned before, nuclear receptors (NRs) are involved in fundamental human health processes and are a relevant target for both medicinal chemistry and toxicological risk assessment. To help the identification of new possible drug candidate (hits) and the prioritization of chemicals that can be EDCs, computational models can be a useful tool. The final aim of computational tools is to anticipate the properties of a compound with reasonably accuracy before testing it in laboratory (Butler et al., 2018).

As a subfield of artificial intelligence (AI), machine-learning (ML) learn the relationships that underlie a dataset by assessing a portion of that data and building a model to make predictions. Machine-learning algorithms can, thus, be viewed as searching through a large space of candidate models, guided by training experience, to find a model that optimizes the performance metric.

Machine-learning algorithms vary greatly. For example they differ by the representation of the candidate model (e.g., decision trees, mathematical functions, and general programming languages) or by the searching strategy through the space of models (Jordan and Mitchell, 2015). Machine learning can be viewed as a crossroad of computer science, statistics and other disciplines involved in the automatic improvement over time, and inference and decision-making under uncertainty.

The training of a machine-learning model may be supervised, unsupervised or semi-supervised, depending on the learning task and the type and amount of available data (Mitchell, 1997). In supervised learning, the training data consist of sets of input and associated output values. The goal of the algorithm is to derive a function that predicts the output values to an acceptable degree of fidelity. If the available dataset consists of only input values, unsupervised learning can be used in an attempt to identify trends, patterns or clustering in the data (Butler et al., 2018).

Quantitative Structure Activity Relationships (QSARs) are used to predict the behavior of chemicals from their structures, leading to better understanding of the adverse effects of the studied substances in cells and tissues. Therefore QSARs belong to ML specifically applied to chemical structures (Selassie and Verma, 2003). QSARs techniques make use of existing experimental data to predict the activity or property of new or unseen chemicals. The conceptual basis of QSARs is that similar structures are expected to exhibit similar biological behavior. The appropriate theoretical descriptors calculated from structural information are used to train the models and predict the biological activity of the chemicals (Cherkasov et al., 2014).

### 1.2.1 Machine learning for drug design and virtual screening

AI tools are widely used in medicinal chemistry application especially in drug discovery, which represents the very first step in the approval of a new drug before clinical trials (see Figure 1.4) and usually takes from three to six years. It is estimated that for each approved drug, 10'000 molecular candidates are usually selected by AI.

In particular, AI tools are useful to design new potential drug, repurpose existing drug or virtually screen large libraries to find potential drugs.

Generative artificial intelligence consists in designing new drug-like compounds with desired activities from scratch (Button et al., 2019). This design concept comprises three tasks: (i) molecule generation, (ii) molecule scoring, and (iii) molecule optimisation. For example among the ligand-based studies, in a recent work (Merk et al., 2018b) Merk et al. trained a recurrent neural network to capture the constitution of a large set of known bioactive compounds. The general model was fine-tuned on recognizing RXR and PPAR agonists. Four of the five synthesized top-ranking compounds designed by the generative model revealed nanomolar to low-micromolar receptor modulatory activity in cell-based assays.

Drug repurposing is a strategy for identifying new uses for approved or investigational drugs that are outside the scope of the original medical indication. It allows to reduce costs and speed up the development timelines. For example, Raloxifene is a selective estrogen receptor modulator originally developed to treat osteoporosis, which thanks to retrospective clinical analysis, was repurposed also to treat breast cancer and recently proposed to treat SARS-CoV-2 infection (Allegretti et al., 2021). Computational approaches constitute the main route to drug repurposing (Pushpakom et al., 2019).

Virtual screening of wide libraries, such as natural products library (Merk et al., 2018a), allows (1) to identify new scope for existing drugs ad thus speed up the application process, (2) to find new alternative structures throughout scaffold hopping and (3) find potential hazardous compounds.

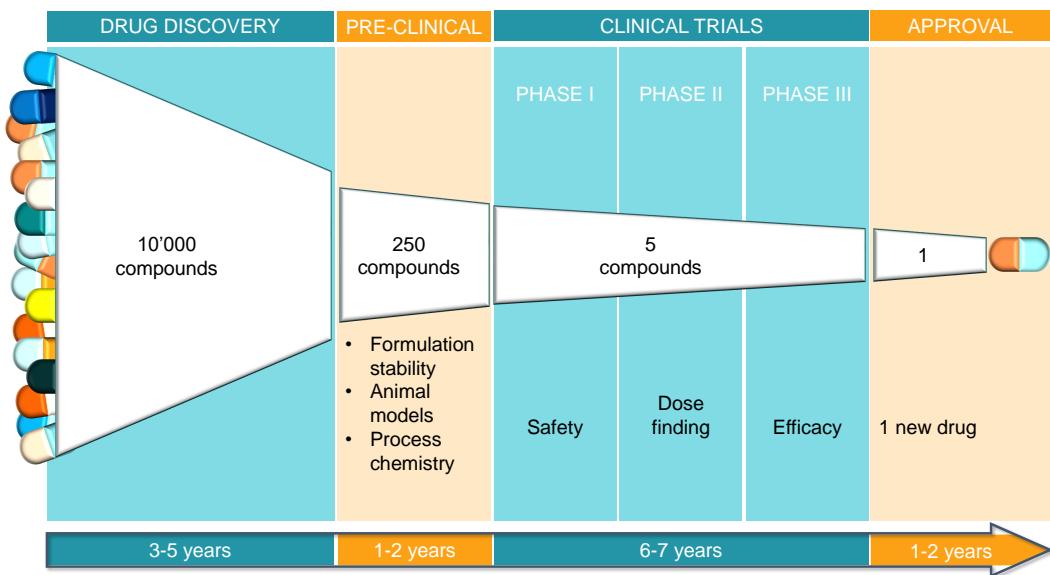


FIGURE 1.4: Workflow of a drug approval.

### 1.2.2 Machine learning for prioritizing modulators

*In-silico* modelling based on machine learning can be used to fill the data-gap in order to prioritize compounds, reduce animal testing and protect human health.

In nuclear receptor frameworks, the use of computational methods to screen and prioritize chemicals for endocrine activity has been already initiated at the EPA's National Center for Computational Toxicology (NCCT), the U.S. Environmental Protection Agency (EPA) and the NTP Interagency Center for Evaluation of Alternative Toxicological Methods (NICEATM), with a special focus on ER and AR (Mansouri et al., 2020; Mansouri et al., 2016).

Starting with ER, a total of 18 *in vitro* assays targeting the main estrogen-signaling steps (three cell-free radioligand binding assays; six dimerization assays using both ER $\alpha$  and ER $\beta$ ; two DNA binding assays; two RNA transcription assays; two agonist-mode protein expression assays; two antagonist-mode protein expression assays; and one cell proliferation assay) were developed. The collected bioactivity data were then used in the Collaborative Estrogen Receptor Activity Prediction Project (CERAPP) to develop a total of 48 QSAR and docking predictive models, which were evaluated using an external set from the literature and then combined into consensus models. The consensus models were then used to virtually screen a library of 32,464 unique chemical structures compiled from different lists of interest to the EPA, which identified approximately 4,000 chemicals with evidence of ER activity (Mansouri et al., 2016). A recent study applied explainable tools on a deep learning architecture aimed to identify potential endocrine disruptors starting from CERAPP data (Mukherjee, Su, and Rajan, 2021).

CERAPP workflow was applied to Androgen receptor in the later Collaborative Modeling Project for Androgen Receptor (CoMPARA) where 11

assays covering the androgen signaling pathway were used to collect bioactivity data. Collaborators from 25 international research groups contributed a total of 91 qualitative and quantitative predictive QSAR models for binding, agonist, and antagonist AR activities (Mansouri et al., 2016).

Other studies were focused on PPAR isoforms. For example, Al Sharif and coworkers (Al Sharif et al., 2017) presented a workflow to screen chemicals based on their potential ability to bind and activate PPAR $\delta$  and thus identify chemicals with hepatotoxic potential; while Oshida et al. developed methods useful for screening of environmental chemicals for PPAR  $\alpha$  bioactivity (Oshida et al., 2015).

Another recent work aimed to supply models that could be used to screen compounds with potential endocrine disrupting characteristics considering six targets, including AR, ER, GR, aromatase, TR, and PPAR $\gamma$  (Sun et al., 2019).

### 1.3 Objectives

The main aim of this work is the advancing of nuclear receptors modulators prediction through machine learning methods.

Nuclear receptors are a superfamily of transcription factors that play a key role in several physiological processes such as cell growth control, development, homeostasis, and metabolism. Because of their biological relevance, NR receptors have been a privileged target for computational applications based on machine learning models in order to i) prioritize the testing of potential harmful compounds such as endocrine disruptors and ii) find new leads for possible selective or promiscuous drug candidates.

Machine learning models are "data hungry" and are strongly influenced by the quality of the input data according to the "garbage in, garbage out" paradigm. Therefore, curation of a comprehensive dataset on nuclear receptor modulators, which can overcome the problem of "data fragmentation" among medicinal chemistry and toxicology sources, is highly beneficial for the scientific community and especially for the development of machine learning models.

In particular, multitask neural networks can exploit all possible information and model underrepresented tasks or, in our case, modulation for nuclear receptors less represented in the experimental data.

We therefore wanted to create a reliable dataset containing information on nuclear receptor modulators and, thus, evaluate the advantages and limitations of multitask neural networks over traditional single-task approaches, in identifying modulators.

Additional objectives included evaluating (i) consensus methods to reduce the effects of conflicting information by averaging model predictions and (ii) different approaches for tuning the hyperparameters of the multitask neural network.

## Chapter 2

# Tools and methods

This chapter describes the methods applied during the project in order (i) to codify the input, i.e., the molecules, (ii) to model the response, i.e., the bioactivity prediction and finally (iii) to assess the model's reliability.

## 2.1 Molecular structure encoding

For storing and modelling purposes it is necessary to encode the molecular structures as string (SMILES) or vectors of molecular descriptors.

In the following sections SMILES notation and the molecular descriptions used in this work will be briefly explained.

### 2.1.1 SMILES

The majority of molecular activity databases encoded molecules in SMILES (Simplified Molecular Input Line Entry System) notation (Weininger, 1988). In this way, the molecular structures can be stored as ASCII string, with advantages in storing space and readability. SMILES line notation (a typographical method using printable characters) consists of a series of characters containing no spaces. Hydrogen atoms may be omitted (hydrogen-suppressed graphs) or included (hydrogen-complete graphs) and aromatic structures may be specified directly or in a Kekulé form.

There are five generic SMILES encoding rules, corresponding to specification of atoms, bonds, branches, ring closures, and disconnections (Figure 2.1 shows some examples).

1. Atoms are represented by their atomic symbols: this is the only required use of letters in SMILES. Each non-hydrogen atom is specified independently by its atomic symbol enclosed in square brackets, [ ]. The second letter of two-character symbols must be entered in lower case. Elements in the 'organic subset' B, C, N, O, P, S, F, Cl, Br, and I may be written without brackets if the number of attached hydrogens conforms to the lowest normal valence consistent with explicit bonds.
2. Single, double, triple, and aromatic bonds are represented by the symbols -, =, #, and :, respectively. Adjacent atoms are assumed to be connected to each other by a single or aromatic bond (single and aromatic bonds may always be omitted).

3. Branches are specified by enclosing them in parentheses, and can be nested or stacked.
4. Cyclic structures are represented by breaking one bond in each ring. The bonds are numbered in any order, designating ring opening (or ring closure) bonds by a digit immediately following the atomic symbol at each ring closure. There are usually many different, but equally valid descriptions of the same structure.
5. Disconnected compounds are written as individual structures, which are separated by a '.' (period). The order in which ions or ligands are listed is arbitrary.

SMILES	2D structure	Name
CC	C–C	ethane
O=C=O	O=C=O	carbon dioxide
C#N	C≡N	hydrogen cyanide
CC(=O)O		acetic acid
c1ccccc1		benzene
C1CCCCC1		cyclohexane
[OH3+]	OH <sub>3</sub> <sup>+</sup>	hydronium ion

FIGURE 2.1: Some examples of molecules written in SMILES format associated with their name and 2D-structure.

In the SMILES language, there are two fundamental types of symbols: atoms and bonds. Using these SMILES symbols, it is possible to specify a molecule's graph (its 'nodes' and 'edges') and assign 'labels' to the components of the graph (that is, say what type of atom each node represents, and what type of bond each edge represents).

SMILES can be used as input by software to compute quantitative representations of chemical structures (molecular descriptors). Some text-based modelling techniques such as recurrent neural networks have proved to be

able to work directly on SMILES representation for new structures generation (Gupta et al., 2018; Grisoni et al., 2020).

Although recently alternatives to SMILES have been proposed (e.g. SELFIES in (Krenn et al., 2019)), SMILES still is the most popular chemical representation format.

### 2.1.2 Molecular descriptors

QSAR applications rely on the principle that the biological properties of any chemical are the effects of its structural characteristics (Hansch and Fujita, 1964). Analogously, compounds with similar molecular structures will be likely to show the same biological and physicochemical profile. Following this assumption, it is possible to numerically encode molecular properties in different forms called molecular descriptors.

Because of their numeric nature, molecular descriptors allow to link the theoretical information arising from the molecular structure (e.g., geometric, steric, and electronic properties) (Todeschini and Consonni, 2008) to some experimental evidence on the molecule (e.g., acute/chronic toxicity, receptor binding). Thus, molecular descriptors have become the input to many computational toxicology applications (Grisoni et al., 2018a). The information encoded by molecular descriptors ranges from simple bulk properties to complex three-dimensional definitions. In particular, different levels of complexity (or dimensionality) can be used to represent any given molecule (Figure 2.2), as follows (Grisoni et al., 2018a):

- 0-Dimensional (0D). The chemical formula is the simplest molecular representation since it specifies the chemical elements and their occurrence in a molecule. For instance, the chemical formula of tamoxifen (a selective estrogen receptor modulator used to treat breast cancer) is  $C_{26}H_{29}NO$ , which indicates the presence of 26 Carbon, 29 Hydrogen, 1 Nitrogen, and 1 Oxygen atoms. This representation is independent of any knowledge about atom connectivity and bond types. Hence, molecular descriptors obtained from the chemical formula are referred to as 0D descriptors and capture bulk properties. 0D descriptors are very simple to compute and interpret, but show a low information content and a high degeneration degree, that is, they may have equal values for different molecules. Some examples of 0D descriptors are atom counts (e.g., number of carbon atoms), molecular weight, and sum or average of atomic properties (e.g., atomic van der Waals volumes).
- 1-Dimensional (1D). According to this representation, molecules are perceived as a set of substructures, such as functional groups or atom-centered fragments. This representation does not require the complete knowledge of molecular structures. The 1D molecular representation is reflected in the derived descriptors, which usually are binary (encoding for the presence/absence of given substructures) or occurrence frequencies.

- 2-Dimensional (2D). This representation adds an additional information level to the 1D representation, by also considering how the atoms are connected, in terms of both presence and nature of chemical bonds. Usually, the molecule is represented as a graph, whose vertexes are the atoms and edges are the bonds. From a graph representation, several numerical quantifiers of molecular topology are mathematically derived in a direct and unambiguous manner. They are commonly known as topological indices (TIs). TIs encode topological properties (e.g., adjacency, connectivity) and are usually sensitive to structural features such as size, shape, symmetry, branching, and cyclicity. Often, also specific chemical properties of atoms are considered, e.g., mass and polarizability, or the presence of hydrogen bond donors/acceptors. Thus, topological indices can be logically divided into two categories: (1) topostructural indices, which encode only information about adjacency and through-bond distances between atoms, and (2) topochemical indices, which quantify information about topology but also specific chemical properties of atoms, such as their chemical identity and hybridization state.
- 3-Dimensional (3D). An additional level of complexity may be added by perceiving the molecule not only in terms of atom type, connectivity, and adjacency but also by viewing it as a geometrical object in space, characterized by the spatial configuration of the atoms. In other words, the molecule is defined in terms of atom types and their x-y-z coordinates. Descriptors deriving from 3D representation have a high information content and can be particularly useful for modeling pharmaceutical and biological properties. When dealing with the 3D representation, users have to keep in mind several issues connected to the geometric optimization of molecules, such as (1) the influence of the optimization method on the coordinate values; (2) the presence of more than one similar minimum energy conformer for highly flexible molecules; and (3) the difference between the bioactive geometry and the optimized geometry, the degree of deformation depending upon the number of freely rotatable bonds in the molecule. For these reasons, the cost/benefit of using 3D descriptors is case-dependent and has to be carefully evaluated.
- 4-Dimensional (4D). In addition to the molecular geometry, a 'fourth dimension' can be introduced, usually aiming to identify and characterize quantitatively the interactions between the molecule(s) and the active site(s) of a biological receptor. For instance, a grid-based representation can be obtained by placing molecules in a 3D grid of several thousands of evenly spaced grid points and by using a probe (steric, electrostatic, hydrophilic, etc.) to map the surface of the molecule. The molecule can be then described through its molecular interactions with

the probe (e.g., see Comparative Molecular Field Analysis and Comparative Molecular Similarity Indices Analysis descriptors). 4D representations may also be ‘ensemble-based’, that is, they can include conformational flexibility and freedom of alignment, through an ensemble of the spatial features of different members of a training set, or by representing each ligand by an ensemble of conformations, protonation states, and/or orientations.

Descriptors can be chosen based on an a priori knowledge and/or on their performance for the problem under analysis. Molecular descriptors can be grouped according to the rationale underlying their design, which influences their applicability to computational problems and the required modeling steps. In particular, molecular descriptors can be divided into classical molecular descriptors and binary fingerprints.

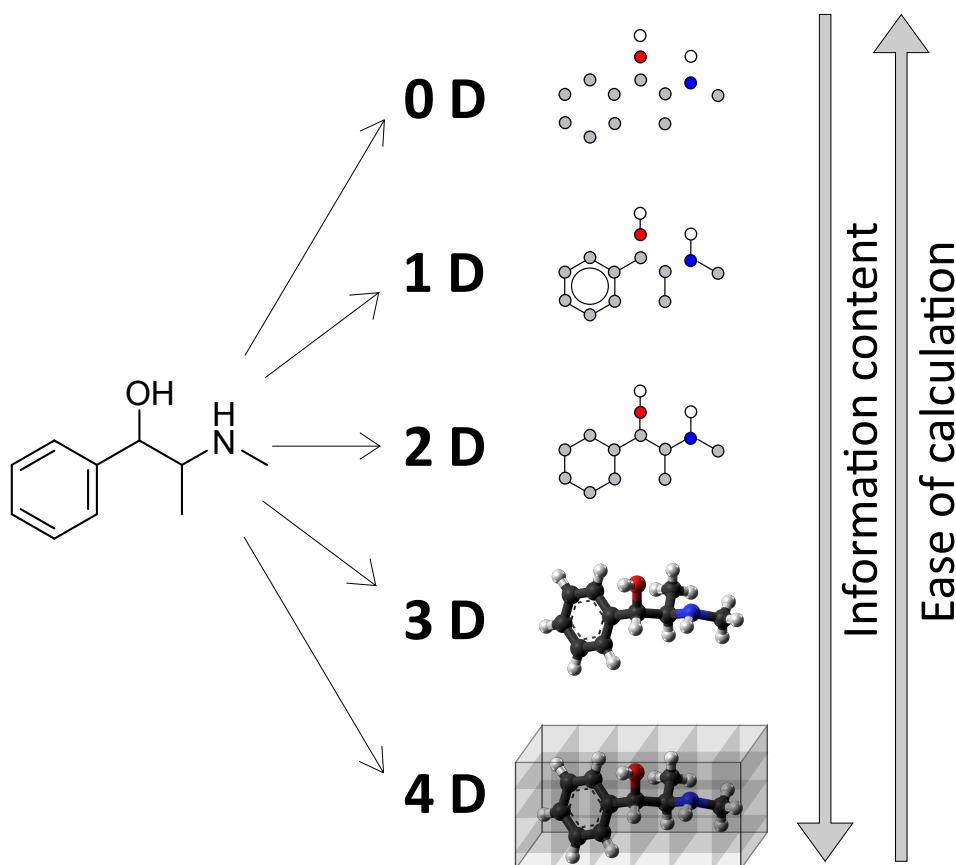


FIGURE 2.2: Graphical example of different molecular representations of the same structure (vanillin, here depicted as a 2D structure).

Classical molecular descriptors (MDs) are designed to encode a precise structural/chemical feature (or a set of features of different complexity) into one, single number. Thus, each descriptor can be used alone or in combination with other descriptors. Classical descriptors can have different measurement scales: they can be integers (e.g., number of double bonds and counts

of atom types), binary (e.g., presence/absence of a given substituent) or can have continuous values (e.g., molecular weight). The majority of classical molecular descriptors are usually interpretable to a certain extent, and, in some cases, they can be mapped back onto sets of structural features (i.e., reversible decoding).

## Fingerprints

Binary fingerprints (FPs) provide a complete representation of all the structural fragments of a molecule in a binary form (Willett, 2006; Rogers and Hahn, 2010). Unlike classical descriptors, fingerprints encode the 2D structural information in a series of binary digits or bits that represent the presence or absence of specific substructures in the molecule and are meaningful only when used as a whole. Usually, a set of patterns (e.g., branched/linear fragments or substructures) are generated from a given molecule, the presence and absence of a pattern are encoded within a string of a given length and represented as 1 or 0, respectively (Shemetulskis et al., 1996).

Hash function is a mathematical algorithm that maps some input of variable length to a fixed length value. In the case of FPs, a hash function is used to map a substructure to a binary vector of fixed length. This procedure is applied to all the considered substructures, using the logical OR operator, to produce the fingerprint of the whole molecule. This means that the hashing procedure allows the mapping of each substructure to the final FP by a certain number of bits that are set to one.

A hash function is deterministic, that is, a certain substructure will be always mapped to the same set of bits (when the same approach and parameters are used), although it usually does not allow reversible-decoding, that is, it is not possible, starting from a given set of bits in the FP, to recreate the original substructure that led to the observed configuration unlike structural keys. In particular, hashing algorithms often lead to a 'collision' of multiple features in the same bit(s) and to the loss of the one-to-one correspondence with molecular features.

Fingerprints allow performing quick calculations for molecule similarity/diversity problems and the frequency of the molecular fragments encoded into FPs can be used to interpret the structural features underlying the observed bioactivity patterns (Todeschini et al., 2012).

FPs are, thus, fixed-size binary vectors that encode the structural information of a molecule by subdividing its structure in all the possible substructure patterns (following a given set of rules), and then processing these patterns by hashing algorithm. A pattern means, for example, a path of predefined length characterized by the nature of atoms and bonds along the path or a circular substructure rooted at a specific atom (Shemetulskis et al., 1996).

Extended Connectivity FingerPrints (ECFPs) are a particular group of FPs that considers the molecular substructures as atom-centred fragments using a variant of the Morgan's extended connectivity algorithm (Rogers and Hahn, 2010). Atom-centred fragments are circular substructures, i.e., fully

explored labelled trees of a particular length, rooted at a particular vertex in the molecular graph.

The user-definable 'pattern length' parameter identifies the maximum atom-centred fragment radius to be explored, that is, the number of bonds along a path starting from the central atom (that is 2 by default). Considering a length equal to 0, substructures are just represented by individual atoms. An iterative process is used to explore the atom environment in order to generate larger substructural fragments. At length 1, the information of all the atoms directly bonded to each atom is taken into account. At length 2, the information of all the atoms within a diameter of 4 chemical bonds is accounted for and so on. The process is complete when the chosen neighborhood size is reached.

Figure 2.3 shows an ECFP generated with default settings for the FP parameters and selecting the maximum length of 1. The active bits (i.e., 1) are associated to the corresponding circular fragments of radius equal to 0 or 1 in the processed molecule. In yellow, the active bits with fragment collision (i.e., bits encoding more than one fragment-type) are highlighted.

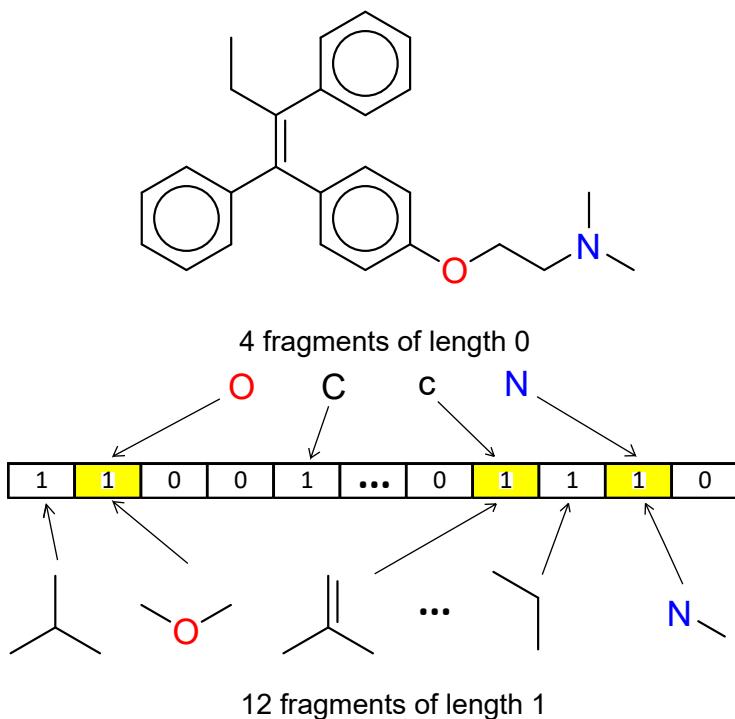


FIGURE 2.3: Extended Connectivity FingerPrints (ECFPs) with a maximum length of 1 for Tamoxifen.

## WHALES

The Weighted Holistic Atom Localization and Entity Shape (WHALES) descriptors (Grisoni et al., 2018b) encode geometric inter-atomic distances, molecular shape, and atomic properties in a holistic way. For each 3D atom position, weighted covariance matrix is computed capturing the spatial distribution of the surrounding molecular atoms. This covariance matrix is then used to normalize the inter-atomic distances and account for local feature distributions. These obtained inter-atomic distances are proportional to the remoteness of each atom from the center of local atomic distributions, measured in variance units. Additionally, the contribution of each atom to the atom-centered covariance matrix is weighted by atomistic partial charges in order to consider potential ligand-receptor interaction patterns or 'pharmacophore' features. The procedure to calculate WHALES descriptors is briefly explained below and in Figure 2.4A, for further details see (Grisoni et al., 2018b).

- Step 1 Atom-centered covariance matrix calculation. Let  $X$  be the matrix of the atom coordinates, made of as many rows as there are non-hydrogen atoms ( $n$ ) and three columns corresponding to the 3D coordinates of each non-hydrogen atom. The distribution of atoms and their partial charges around any  $j$ -th atom is captured using an atom-centered weighted covariance matrix ( $S_{w(j)}$ ),

$$\mathbf{S}_{w(j)} = \frac{\sum_{i=1}^n |\delta_i| \cdot (x_i - x_j) (x_i - x_j)^T}{\sum_{i=1}^n |\delta_i|} \quad (2.1)$$

where  $(x_i - x_j)$  are the differences between the 3D coordinates of the  $j$ -th atomic center and those of any  $i$ -th atom, while  $|\delta_i|$  is the absolute value of the partial charge of the  $i$ -th atom. The atom-centered covariance is computed for any non-hydrogen atom of the molecule. The weighted covariance matrix is influenced by the density and partial charges of atoms surrounding  $j$ .

- Step 2 Atom-centered Mahalanobis distance calculation. From  $S_{w(j)}$ , the atom-centered Mahalanobis (ACM) distance from the center  $j$  to any  $i$ -th atom is calculated as follows:

$$\text{ACM}(i,j) = (x_i - x_j)^T \cdot S_{w(j)}^{-1} \cdot (x_i - x_j) \quad (2.2)$$

All of the pairwise normalized interatomic distances calculated according to Eq. 2.2 are collected in the ACM matrix.

- Step 3 Calculation of atomic indices. From the ACM matrix, three indices are calculated for each atom:

1. Remoteness (Rem), which is the ACM matrix row-average, calculated as follows:

$$Rem(j) = \frac{\sum_{i=1}^n ACM(j, i)}{n - 1} \quad (2.3)$$

2. Isolation degree (Isol), which is the ACM matrix column minimum (excluding the atomic center):

$$Isol(j) = \min_i (ACM(i, j)) \quad (2.4)$$

3. Isolation-Remoteness ratio, calculated as:

$$IR(j) = \frac{Isol(j)}{Rem(j)} \quad (2.5)$$

The remoteness, isolation degree values and their ratio calculated for negatively charged atoms are assigned a negative sign, as follows:

$$\text{if } \delta_j < 0 \begin{cases} Isol(j) = -Isol(j) \\ Rem(j) = -Rem(j) \\ IR(j) = -IR(j) \end{cases} \quad (2.6)$$

This procedure allows to distinguish atoms having the same values of isolation degree and remoteness but charged differently.

- Step 4 WHALES descriptors calculation. A binning procedure is applied to obtain a fixed-length representation, enabling the straightforward comparison of molecules with different numbers of atoms. The WHALES descriptors are, thus, usually calculated as deciles plus minimum and maximum of (i) Isol, (ii) Rem, and (iii) IR. Thus, each molecule is characterized by the same number of descriptors (i.e., 11 values for each atomic index, for a total of 33 descriptors), regardless of the number of atoms considered.

For WHALES calculations usually the Gasteiger-Marsili (Gasteiger and Marsili, 1980) partial charges and MMFF94 (Halgren, 1996) energy-minimized structures are used.

An adaptation of the traditional WHALES was carried out by computing of Euclidean distance instead of Mahalanobis distance without accounting for the covariance matrices and thus replacing steps 1 and step 2 with a partial charge weighted Euclidean distance calculation.

WHALES descriptors can be able to embed features related to the formation of the ligand-receptor complex, therefore were chosen in this study also to describe protein features (Figure 2.4B). In this case the 3D geometry is given by the atom coordinates listed in Protein Data Bank (PDB) files. The primary information stored in a PDB file, indeed, consists of coordinate information. A typical PDB formatted file includes a large 'header' section of text that summarizes the protein, citation information, and the details of the structure solution, followed by the sequence and a long list of the atoms and

their coordinates. For this analysis we used the crystallographic structures of ligand-nuclear receptor complexes from the PDBbind database (PDB, Accessed: 2021-07-21; Berman et al., 2000) summarized in Table 2.1.

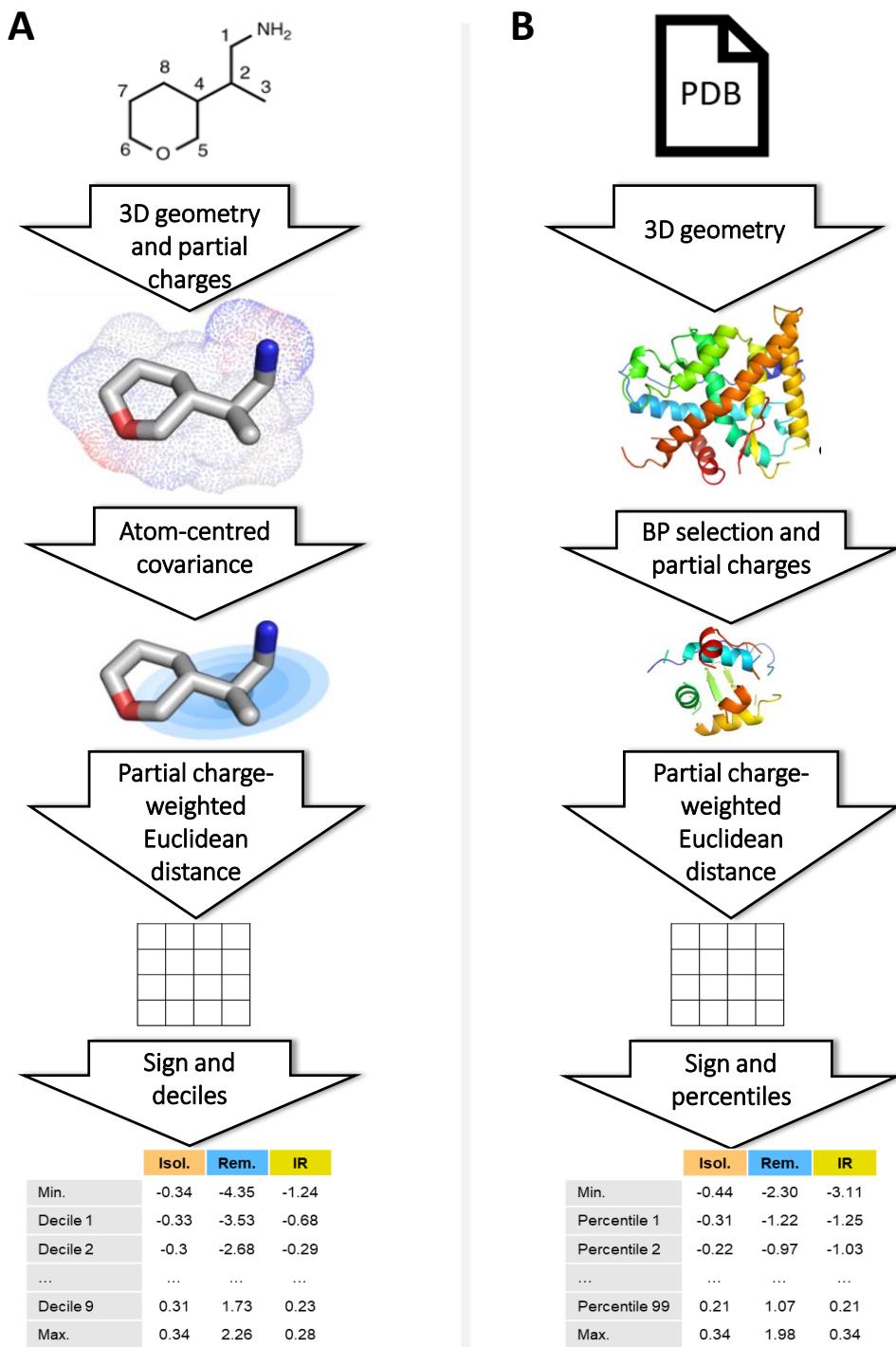


FIGURE 2.4: (A) Starting from a given molecular graph, 3D energy minimization and partial charge calculation are performed. The coordinates and the computed partial charges are used to calculate the atom-centered weighted covariance and then the ACM distance matrix. From the ACM, WHALES descriptors are calculated as the deciles, the minimum and the maximum of isolation degree (Isol.), remoteness (Rem.) and isolation-remoteness ratio (IR). (B) Adaptation of WHALES pipeline to the binding pocket (BP) of a protein structures extracted from a PDB file. In this case percentiles are computed.

TABLE 2.1: Summary of PDBbind crystallographic structures.

Target	No. ligands	PDB structures
AR	22	1E3G, 1Z95, 2AM9, 2AMA, 2AX6, 2AX9, 2HVC, 2IHQ, 2NW4, 2OZ7, 3B5R, 3B65, 3B66, 3B67, 3B68, 3G0W, 3V49, 4QL8, 5CJ6, 5T8E, 5T8J, 5V8Q
ER $\alpha$	13	1X7E, 1X7R, 2I0J, 3ERD, 3ERT, 5FQP, 5FQT, 5FQV
ER $\beta$	36	1NDE, 1QKN, 1U3Q, 1U3R, 1U3S, 1U9E, 1X76, 1X78, 1X7B, 1YY4, 1YYE, 1ZAF, 2GIU, 2I0G, 2J7X, 2JJ3, 2NV7, 2Q TU, 2Z4B 3DCT, 3OKH, 3OKI, 3OLF, 3OMM, 3OOF, 3OOK, 4OIV, 5Q0I, 5Q0J, 5Q0L, 5Q0M, 5Q0N, 5Q0O, 5Q0P, 5Q0Q, 5Q0R, 5Q0S,
FXR	68	5Q0T, 5Q0U, 5Q0V, 5Q0W, 5Q0X, 5Q0Y, 5Q10, 5Q11, 5Q12, 5Q13, 5Q14, 5Q15, 5Q16, 5Q17, 5Q18, 5Q19, 5Q1A, 5Q1B, 5Q1C, 5Q1D, 5Q1F, 5Q1G, 5Q1I
GR	18	1NHZ, 1P93, 3K22, 3K23, 4CSJ, 4P6W, 4P6X
PPAR $\alpha$	11	1I7G, 1KKQ, 3FEI, 3G8I, 3KDT, 3KDU
PPAR $\delta$	8	3DY6, 3GWX, 3GZ9, 3PEQ, 3TKM 1FM9, 1I7I, 1NYX, 1ZEO, 2ATH, 2F4B, 2G0G, 2G0H, 2GTK, 2HFP, 2I4J, 2I4Z, 2P4Y, 2Q8S, 2YFE, 3B1M, 3FEJ, 3FUR,
PPAR $\gamma$	59	3G9E, 3H0A, 3IA6, 3LMP, 3OSI, 3OSW, 3R5N, 3R8I, 3SZ1, 3T03, 3TY0, 3U9Q, 4A4V, 4A4W, 4JAZ, 4PRG, 4R06, 4XTA, 4XUH, 4XUM, 4Y29, 5F9B, 5LSG, 5TWO, 5U5L
PR	16	1A28, 1SQN, 1SR7, 1ZUC, 2W8Y, 3G8O, 3HQ5, 3KBA, 4OAR
PXR	4	1ILH, 1M13, 2O9I
RXR	20	1RDT, 3FAL, 3NSQ, 3OZJ, 3PCU, 3R2A, 3R5M, 4K4J, 4K6I, 4M8E, 4M8H, 4POH, 4POJ, 4PP3, 4PP5, 4ZSH, 5MKJ

After removing the structures which include an allosteric ligand, we aligned all the remaining structures and, in each of them, we selected only residues within 5 Å from any of the crystallographic ligands. This procedure returns the overlapping portion of each crystallographic receptor in correspondence of the binding site, compared to all the receptors, that was used for the calculation of the WHALES. To take into account the bigger dimensions of the binding pockets compared to ligands and thus to better capture the pocket conformation, percentiles were computed instead of deciles.

## 2.2 Traditional QSAR classification methods

Classification (or, more formally, supervised classification) is the specific area of machine learning that aims to assign objects to one of several predefined classes. The objects, that represent the input of a classification task, are called records, instances or observations and in QSARs problems are typically represented by molecules. Each of these records is characterized by a tuple  $(\mathbf{x}, y)$ , where  $\mathbf{x}$  represents the set of attributes or features of the object and  $y$  its class label(s). The attributes can be both numerical (i.e., continuous values) or categorical (i.e., discrete, cardinal values), while the class labels must be categorical. Classification is the task of learning a target function  $f$  that maps each attribute set  $\mathbf{x}$  to one of the predefined class labels  $y$  (Tan, Steinbach, and Kumar, 2005).

An example of a classification problem in the domain of QSAR could be the one of predicting the activity of a molecule for a given biological target. The set of class labels could be  $\{\text{active}, \text{inactive}\}$ . In this case we are facing a binary classification problem. The set of features instead could be represented by molecular descriptors (e.g.  $\{\text{molecular weight}, \text{hydrophobicity}, \text{presence of specific functional groups}\}$ ). Ideally, all the descriptors provided during the training phase to the classification algorithms contribute to learning a function that maps the attribute set to one of the two class labels. The final goal of this classification model is to be able to learn a generalized function such that, given a new set of molecules (described by the predefined set of features) the model will predict their activity towards the selected target.

Given these preliminaries related to the supervised classification problem, several classification algorithms (i.e., the procedures responsible to learn from the data the mapping function  $f$ ) were developed throughout the years. In the following paragraphs we will present some instances of the classification algorithms used in this project.

### 2.2.1 k-Nearest Neighbours and N-Nearest Neighbours

$k$ -Nearest Neighbours ( $k$ NN) is a local classification method that assigns a target molecule to the most represented class among the  $k$  most structurally similar records of the training set according to a predefined distance computed among features (Wilkinson et al., 1983; Keller, Gray, and Givens, 1985). The class membership of a new object will thus be defined as a majority vote of its neighbors, with the new object being assigned to the class most common among its  $k$  nearest neighbors.

The training objects are vectors in a multidimensional feature space, each paired to a class label. The training phase of the algorithm consists of storing the feature vectors and class labels of the training objects; while in the classification phase, all the distances (using a user-defined metric) between the unlabeled vector of features of the new object and the training objects are computed. The new object is classified by assigning the label which is most frequent among the  $k$  training objects nearest to that query point (Figure 2.5).

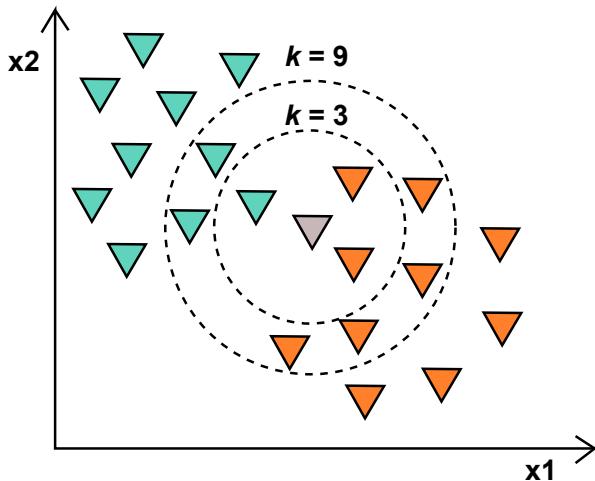


FIGURE 2.5: Toy example of kNN with basic majority voting. Objects are described by two features ( $x_1$  and  $x_2$ ) and are divided into 2 classes (green and orange triangles). A new object is represented by a grey triangle. The two circles highlight the areas including 3 and 9 nearest neighbours. In both cases the new object will be assigned to the orange class.

A drawback of the basic majority voting classification occurs when the class distribution is unbalanced. That is, the most frequent class tends to dominate the prediction of the new object, because the objects belonging to the most frequent class tend to be widely spread among the neighbors due to their large number. One way to overcome this problem is to weight the object contribution, considering the distance from the test object to each of its  $k$  nearest neighbors. The class (or value, in regression problems) of each of the  $k$  nearest objects is multiplied by a weight proportional to the inverse of the distance from that point to the test object.

For a kNN model, it is necessary to optimize the number of  $k$  neighbours as well as the distance metric. The best choice of  $k$  is data-dependent; generally, higher values of  $k$  reduces effect of the noise, but make boundaries between classes less distinct. On the contrary, a model with  $k = 1$  will be more affected by noise.

In binary (two class) classification problems, it is helpful to choose  $k$  to be an odd number as this avoids tied votes.

The best choice of distance metric depends on the features type. A commonly used distance metric for continuous features is Euclidean distance, while for binary features (e.g., fingerprints) Jaccard-Tanimoto distance is preferred.

N-Nearest Neighbours (N3) (Todeschini et al., 2015) method, as kNN, considers only local information to perform the classification of each object. Unlike kNN, N3 method takes into account all the  $n - 1$  objects to classify the  $i$ -th new object. The  $n - 1$  neighbours are sorted from the most similar to the least similar to the new object and the corresponding similarity rank

vector is obtained; then, the neighbour contributions to class assignment exponentially decrease as the similarities diminish, since they are weighted by the rank, whose role is modulated by an  $\alpha$  exponent to be optimized.

### 2.2.2 Random Forest

Decision tree classification models are a family of predictive models used for supervised classification, represented by a tree structure composed by branches and leaves (Rokach and Maimon, 2007). A decision tree is used as a predictive model to go from observations of the features  $x$  of an object (represented in the branches) to conclusions about the object's target class  $y$  (represented by the leaves). In these tree structures, leaves are associated with class labels and branches represent conjunctions of features that lead to those class labels. An illustrative example of a decision tree model is presented in Figure 2.6.

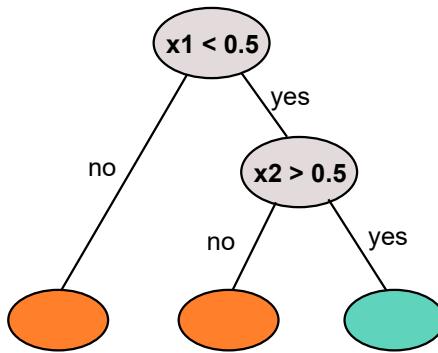


FIGURE 2.6: Toy example of a decision tree classifier. Leaf (i.e., terminal) nodes are coloured according to the class assignation (green and orange) while in the two decision nodes include the splitting criteria of the branches.

A decision tree can be learned by splitting the data into subsets based on attribute value tests (i.e., logic statements on attribute values). This process is repeated on each derived subset in a recursive manner, called recursive partitioning. The recursion is completed when the whole subset belongs to the same target class, or when splitting no longer adds value to the classification performance. Many specific learning algorithms were proposed in the literature throughout the years, and the most relevant are: ID3, C4.5, and CART (Rokach and Maimon, 2007). Algorithms for constructing decision trees usually work top-down, by choosing a variable at each step that best splits the set of available examples. Different algorithms use different metrics for measuring the best split. These, generally, measure the homogeneity of the target variable within the subsets. These metrics are applied to each candidate subset, and the resulting values are combined (e.g. averaged) to provide a measure of the quality of the split. The most used metrics to determine the quality of a split are *Gini impurity* and *information gain*.

A peculiar type of classification procedure that gained increasing success in the machine learning community is represented by ensemble learning. The general idea is to boost classification performances by averaging across the predictions of an ensemble of classifiers instead of relying on a single one. In the context of decision trees *Random forest* (Breiman, 2001) classifiers were introduced.

Random forest is an ensemble learning method for classification, that operates by constructing a multitude of randomized and uncorrelated decision trees at training time and outputting the class that is the mode of the classes of the individual trees. The main advantage of random forests over decision trees is the capability of avoiding overfitting to the training set, thanks to the *bagging* (or bootstrap aggregating) procedure.

Several parameters can be optimized in a Random Forest model, such as the number of trees to be considered and the prune and split criteria.

### 2.2.3 Naïve Bayes

Naïve Bayes classifiers are a family of probabilistic classifiers, based on the Bayes theorem and on the strong (naïve) assumption of conditional independence between features  $\mathbf{x}$  given the class  $y$  (Tan, Steinbach, and Kumar, 2005). In this setting we consider the classification problem from a probabilistic perspective. The class variable is assumed to have a non-deterministic relationship with the features. Thus, we treat the set of features as a set of random variables  $\mathbf{X}$  and the class as random variable  $Y$  and capture their relationship stochastically with the conditional probability  $P(Y|\mathbf{X})$ . This conditional probability is also known as the posterior probability of  $Y$ , as opposed to its prior probability,  $P(Y)$ . During the training phase, we need to learn the conditional probability  $P(\mathbf{X}|Y)$ , i.e., the likelihood for every combination of  $\mathbf{X}$  and  $Y$  based on information gathered from the training data.

Accurately estimating the posterior probabilities for every possible combination of class label and feature values is a difficult problem because it requires a very large training set, even for a moderate number of features. The well-known Bayes theorem becomes useful, because it allows to express the *posterior probability* in terms of the prior probability  $P(Y)$ , the *class-conditional probability* (or *likelihood*)  $P(\mathbf{X}|Y)$ , and the prior probability of the evidence *evidence*  $P(\mathbf{X})$ :

$$P(Y|\mathbf{X}) = \frac{P(\mathbf{X}|Y)P(Y)}{P(\mathbf{X})} \quad (2.7)$$

When comparing the posterior probabilities for different values of  $Y$ , the denominator term  $P(\mathbf{X})$  is always constant, and thus, can be ignored. The prior probability can be easily estimated by computing the class proportion for the objects in the training set. To estimate the class-conditional probabilities  $P(\mathbf{X}|Y)$  the naïve assumption comes in handy. A Naïve Bayes classifier estimates the class-conditional probability by assuming that the features, i.e., the components of  $\mathbf{X}$  are conditionally independent, given the class label  $y$ . Thus the class-conditional probability can be estimated for each feature  $X_i$  in the attribute set  $\mathbf{X} = \{X_1, X_2, \dots, X_d\}$  independently. The likelihood  $P(\mathbf{X}|Y)$  can

be formally computed as:

$$P(\mathbf{X}|Y) = \prod_{i=1}^d P(X_i|Y) \quad (2.8)$$

With the conditional independence assumption, instead of computing the class-conditional probability for every combination of  $\mathbf{X}$ , we only have to estimate the conditional probability of each  $X_i$  given  $Y$ .

This approach is more practical because it does not require a very large training set to obtain a good estimate and its optimization is straightforward since no parameters has to be preliminary defined.

Finally, different assumptions can be drawn on the nature of the class-conditional distributions, corresponding to variants of the naïve bayes classifier. Whereas *multinomial* and *Bernoulli* naïve Bayes classifiers are used for discrete/categorical set of attributes, *gaussian* naïve Bayes is used to deal with continuous values.

## 2.2.4 Applicability domain

'The applicability domain of a (Q)SAR model is the response and chemical structure space in which the model makes predictions with a given reliability' (Netzeva et al., 2005) (see Figure 2.7). In other word, the applicability domain is the chemical space where predictions are obtained by model interpolation and thus are associated with higher confidence (Sahigara et al., 2013).

In the QSAR field, the Applicability Domain (AD) is widely understood to express the scope and limitations of a model, i.e., the range of chemical structures for which the model is considered to be applicable. Therefore, reliable predictions are limited to the chemicals that are structurally similar to the ones used to build the model.

The most commonly adopted approach consists in defining the AD of the model by means of the features distance from training objects.

In this work, a previously published approach (Sahigara et al., 2013) which is based on a set of local thresholds corresponding to the training data points and defining the width of their neighbourhood, was applied. For each  $i$ -th training molecule, the associated threshold  $t_i$  was calculated as the average Jaccard-Tanimoto distance on ECFPs to the first  $k_i$  neighbours, the number  $k_i$  being variable and depending on the object density in the chemical space.

If a test molecule exceeds the threshold of all the training molecules, then it is considered as outside the AD, and its prediction is considered as unreliable. On the contrary, if the molecule falls inside the neighbourhood of at least one training molecule, it will be considered inside the domain of applicability and associated with a reliable prediction. Therefore, given the training set TR, for each test molecule  $j$ , the AD decision rule is

$$j \in AD \iff \exists i \in TR : D_{ij} \leq t_i \quad (2.9)$$

where  $D_{ij}$  is the binary Jaccard-Tanimoto distance between the  $j$ -th test and the  $i$ -th training molecule.

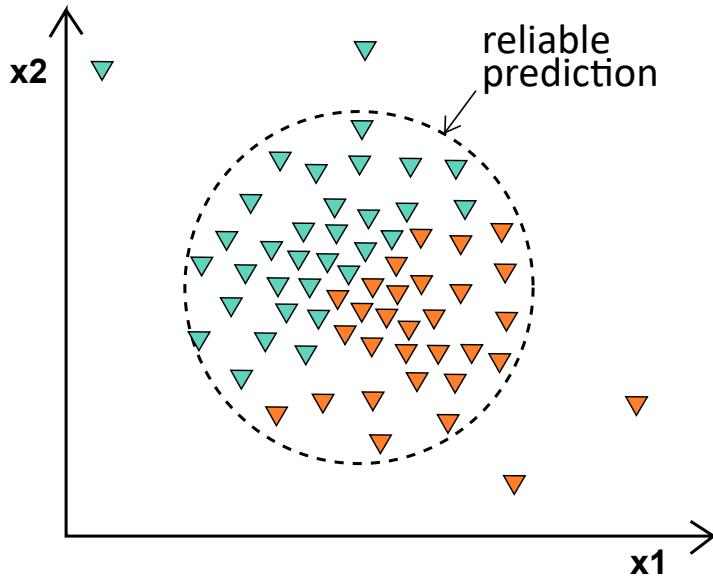


FIGURE 2.7: Toy example of applicability domain as a a region with reliable predictions.

## 2.3 Artificial neural networks

The study of artificial neural networks (ANN) was inspired by attempts to simulate biological neural systems. Analogous to human brain structure, an ANN is composed of an inter-connected assembly of nodes and directed links or connections between the nodes (Tan, Steinbach, and Kumar, 2005). The simplest ANN model is called *Perceptron* (Minsky and Papert, 1988). Perceptrons were developed in the 1950s by Frank Rosenblatt, considered one of the fathers of deep learning. Nowadays other models of artificial neurons are widely used, but to understand neural networks, it's useful to first understand perceptrons (Nielsen, 2015).

The perceptron consists of two types of nodes: *input nodes*, which are associated with the input features  $\mathbf{x}$ , and an *output node(s)*, which represents the model output  $y$  (i.e., the class). The nodes in a neural network's architecture are commonly known as neurons. In a perceptron, each input node is connected via a weighted connection to the output node. The weighted connection emulates the strength of synaptic links between neurons. As in biological neural systems, training a perceptron model means to adapt the weights of the connections until they fit the input-output relationships of the underlying data.

A perceptron computes its output value  $\hat{y}$ , by performing a weighted sum on its inputs  $x_i \in \mathbf{x}$ , subtracting a bias factor  $b$  from the sum, and then applying a non-linear transformation  $g(\cdot)$ , called *activation function*. More specifically, the output of a perceptron model can be expressed as follows:

$$\hat{y} = g\left(\sum_{i=1}^d w_i x_i - b\right) \quad (2.10)$$

Where  $d$  is the number of input neurons, which equals the number of features in the data.  $w_i$  is the learned weight for the  $i$ -th neurons and  $g(.)$  is a non-linear activation function (e.g. sigmoid, hyperbolic tangent, rectified linear units), that projects the weighted sum onto a specific numerical interval (e.g.  $[0, 1]$ ).

### 2.3.1 Feedforward neural networks

The most used ANN supervised classifier is the *Multi-layer perceptron* (MLP) or *Feedforward neural network* (FNN), that introduces a more complex structure to the simple perceptron described so far. The network may contain several intermediary layers between its input and output layers (an example is reported in Figure 2.8). Such intermediary layers are called hidden layers and the nodes embedded in these layers are called hidden neurons. These additional complexities allow FNN to model more complex relationships between the input and output variables.

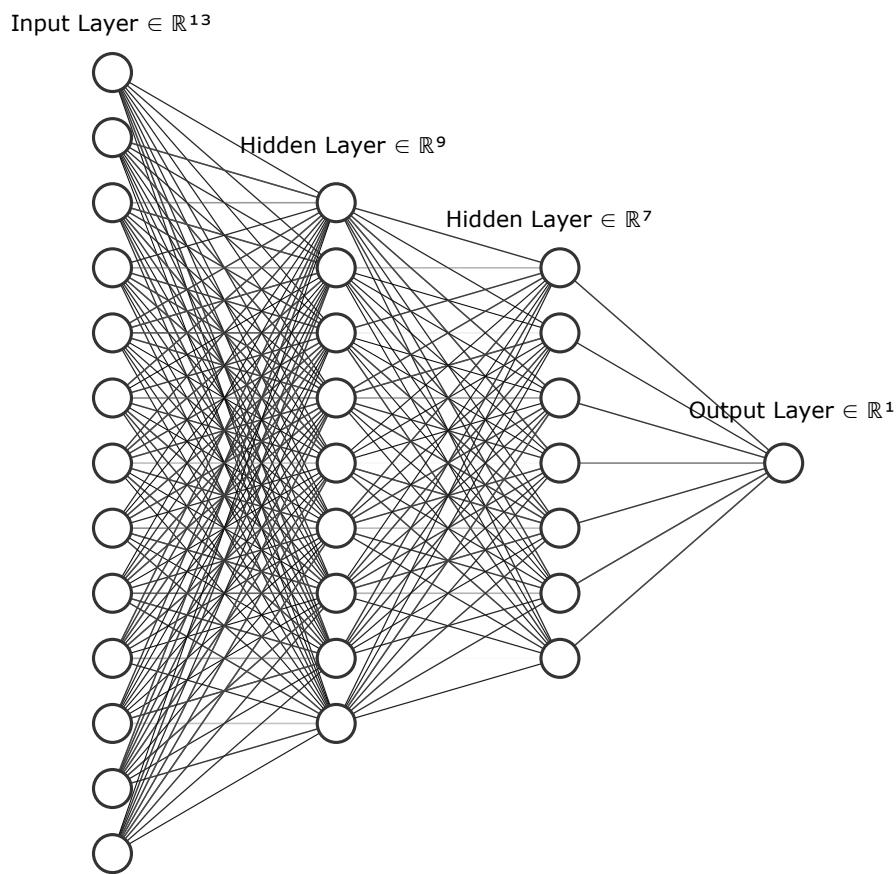


FIGURE 2.8: Toy example of a feedforward neural network with an input feature vector of length 13, two hidden layers of 9 and 7 neurons, respectively and one output.

The process of learning the weights of a FNN from data is a non-trivial task. The goal of the learning algorithm is to determine the set of weights  $\mathbf{w}$

that minimize a defined loss function (e.g. sum of squared errors or cross-entropy) over the predicted classes  $\hat{y}$  and the real classes  $y$ . Note that the loss function depends on  $w$  because the predicted class  $\hat{y}$  is a function of the weights assigned to the hidden and output nodes. Greedy algorithms, such as those based on the gradient descent method, have been developed to efficiently solve the optimization problem (i.e., find the global minimum of the loss function). The general update formula is based on the gradient of the computed error/loss. It states to update the weights in the direction that reduces the overall error term. In order to update and learn the weights of the output and hidden nodes of a neural network an efficient algorithm based on gradient descent has been developed. It is called *Backpropagation* and described in details in (Hecht-Nielsen, 1992).

Each connection represents a weight, whereas each node represents a learning function  $f$  that, in the feedforward phase, processes the information of the previous layer to be fed into the subsequent layer. In the backpropagation phase, each weight is adjusted according to the loss function and the optimization algorithm. Different types of learning or activation functions exist in literature; the most known functions are sigmoid ( $\sigma$ ), Rectified Linear Unit (ReLU), hyperbolic tangent (tanh) and leaky ReLU (Nair and Hinton, 2010; Maas, Hannun, Ng, et al., 2013; Agostinelli et al., 2014). To iteratively adjust the weights, a loss function is computed considering the experimental and the predicted response. Neural network tuning implies setting a learning rate that determines the update of the weights in each iteration with respect to the gradient of the loss function; this parameter can be fixed or changed during the learning (e.g., exponential decay).

For computational and learning efficiency, in each training step (i.e., iteration) a subset of training objects called batch is used. When all the objects are seen by the model, i.e., after a number of iterations equal to the training set size divided by the batch size, a training epoch is completed.

Furthermore, several strategies called regularization techniques can improve the network's generalizing ability and reduce overfitting. This is the case of dropout and weight decay (L1 or L2 regularization). In this work, we used the cross-entropy as loss function () which is the standard loss function for classification problems; it can handle multiple outputs also in the case of missing data.

The threshold of assignment for the output nodes for each neural network was optimized on the basis of ROC curves (Fawcett, 2006), that is, if the output of the neural network ensemble node is equal or lower than the best task-specific threshold selected using the ROC curve, the compound is predicted inactive, otherwise active.

### 2.3.2 Graph convolutional networks

Graphs naturally arise in many real-world applications, including molecular representations. By representing a molecule as a graph, the structural information can be encoded to model the relations among atoms, and furnish more insights underlying the data (Kipf and Welling, 2016).

Currently, the most used graph neural network models are the Graph Convolutional Networks (GCNs), where filter parameters are typically shared over all locations in the graph.

For GCN models, the goal is to learn a function of features on a graph  $G = (\nu, \varepsilon)$  which takes as input:

- A feature description  $x_i$  for every node  $i$ ; summarized in a  $N \times D$  feature matrix  $X$ , where  $N$  is number of nodes and  $D$  the number of input features.
- A representative description of the graph structure in matrix form; typically in the form of an adjacency matrix  $A$ .

Then GCN can produce either a node-level output  $Z$  (an  $N \times F$  feature matrix, where  $F$  is the number of output features per node) or a graph-level outputs by introducing some form of pooling operation (Duvenaud et al., 2015). Since in the latter case, the task is to predict a single class for the entire graph instead of for every node, it is necessary to aggregate the representations of all the nodes and potentially the edges to form a graph-level representation. Such process is more commonly referred as a readout. A simple choice is to average the node features of a graph (Wang et al., 2019).

In our case, since the objective is the assignation of a class label to an entire molecule (i.e., graph) we considered graph-level outputs.

### 2.3.3 Multi-task learning

Multi-task learning (MTL) works as an inductive transfer mechanism aimed to improve generalization performance by leveraging the domain-specific information contained in the training features of related tasks (Caruana, 1997).

In classification problems, MTL models are able to assign objects to one of several predefined classes for multiple tasks without the condition of mutual exclusivity as in the case of multiclass learning. In this case  $y$  is a vector of labels, one for each task. The tasks have to be related, but values of  $y$  are not mutually exclusive (see Figure 2.9).

Multitask learning usually rely on fully connected neural network layers trained on joint tasks, where the output is shared among all learning tasks and then fed into individual classifiers, one for each task (Ramsundar et al., 2015; Dahl, Jaitly, and Salakhutdinov, 2014; Proschak, Stark, and Merk, 2018).

When some dependence relationships exist among the tasks, the model should learn a joint representation and, thus, benefit from an information boost (Ruder, 2017; Xu et al., 2017; Caruana, 1997).

As in single-task, in multi-task feedforward neural networks, input vectors are mapped to output vectors with repeated compositions of layers and the output layer consists of as many nodes as tasks.

We used multitask networks with and without a bypass net, which is an independent additional layer that ‘bypass’ shared layers to directly connect inputs with outputs leading to more robust results (Ramsundar et al., 2015).

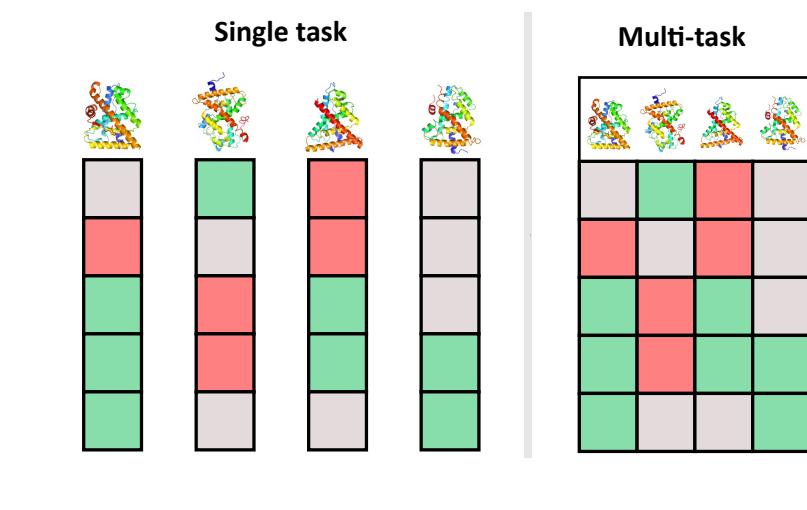


FIGURE 2.9: Comparison of single-task and multi-task data. In the former case one model for each task is needed, while in the latter one model is able to predict simultaneously all the tasks

Several mechanisms might help MTL models to generalize better, for example (Caruana, 1997):

- Statistical Data Amplification, which is an increase in object size due to extra information in the training data of related tasks.
- Attribute Selection, which, as a consequence of data amplification, is an increase ability of the model to distinguish relevant input features.
- Eavesdropping, which is the ability of a MTL model to eavesdrop, i.e., learn of, important features for one tasks and transfer this knowledge to other tasks.

Recent attempt made to demystifying the role of MTL (Xu et al., 2017) in QSAR filed states that: 'when tasks contain molecules in the training set with structures similar to those in other tasks and the activities (task labels) between these similar molecules are correlated (either positively or negatively), building a multitask DNN can boost the predictive performance; otherwise, if the activities between these similar molecules are uncorrelated, using multitask models can degrade the predictive performance'. Moreover, it was proved that when tasks do not share structural similarities between molecules, multitask model will show comparable predictive performance to single task models, regardless of whether the activities in the tasks are correlated (Xu et al., 2017).

### 2.3.4 Parameters tuning

The choice of a modelling approach usually requires the choosing of related learning parameters which can vary greatly in number from one approach to another. For example kNN models depend on the distance metrics selection

as well as on the definition of  $k$ . The number of parameters to tune increases when considering neural networks. Hence, the following parameters are example of user-definable parameters and thus prone to tuning:

- number of layers; in case of one or two hidden layer the network is called shallow, while a network with more than two hidden layers is called deep;
- number of neuron per layers;
- learning rate which is a number determining the update of the weights in each iteration with respect to the gradient of the loss function;
- batch size which indicates the size of the subset of training examples to use in each iteration;
- dropout, which indicates a percentage of randomly selected neurons to turn off;
- type of norm penalty (none, L1 or L2 function);
- type of activation function, the most used alternatives are ReLU, sigmoid and hyperbolic tangent;
- type of optimization algorithm which includes stochastic gradient descend, RMSProp, Adam and Adamax.

Greater complexity comes from the fact that many of these parameters are interdependent, such as the number of layers and the number of neurons per layer in defining the network architecture.

Different strategies to tune the network parameters exists, the most known are grid search, random search and genetic algorithm (Liashchynskyi and Liashchynskyi, 2019) (Figure 2.10).

- Grid search (GS) strategy, as a brute-force strategy, is the most computationally expensive and time consuming strategy, but allows to explore all the possible combinations of parameters at the selected levels in the search space. Grid search suffers from high dimensional spaces, but often can easily be parallelized.
- Random search (RS) strategy avoids the complete selection of all combinations by a random selection of combinations and thus is more efficient than GS (Bergstra and Bengio, 2012). We chose to randomly select a subset of GS combinations and thus to limit the exploration. The number of random combinations to test is user-defined, usually on the basis of a trade-off between available computational time/power and satisfying performance.
- Genetic algorithm (GA) is an heuristic stochastic evolutionary search algorithm based on sequential selections, combinations, and parameter

mutations simulating biological evolution. In other words, combinations of parameters leading to higher performances or fitness survive and have higher chances to reproduce and generate children with a pre-defined mutation probability. Each generation consists of a population of chromosomes representing points in search space. Each individual is represented as a binary vector. In our case, the genetic algorithm begins with a randomly generated population of  $n$  chromosomes and the computation of their fitness as in  $k$ -fold cross validation. Then, the selection and recombination process, based on each chromosome's fitness, lead to the generation of two children with a mutation probability for each bit usually equal to 10%. After inserting the children into the population, the two worst-performing individuals (i.e., those with the lowest fitness) were discarded. This process is iterated till reaching a stopping criterion. In addition, after some generations a cataclysm can be simulated by replacing the worst performing half of the population with new randomly generated chromosomes.

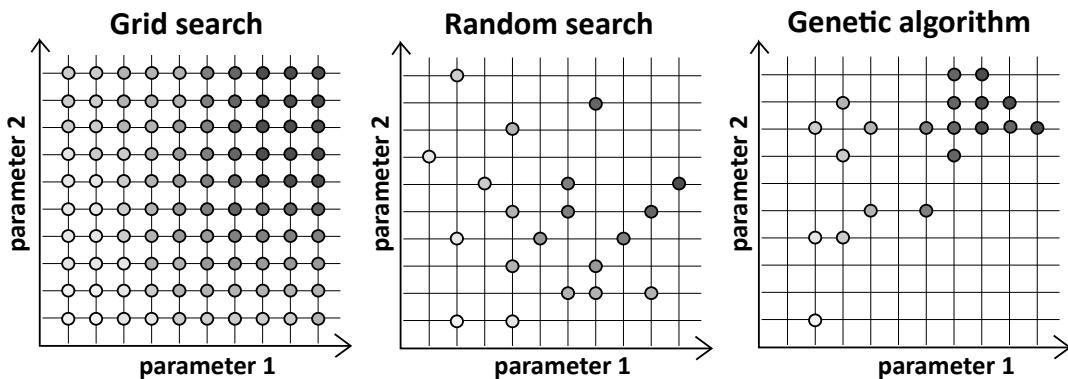


FIGURE 2.10: Toy example of optimization methods for 2 parameters with 10 levels each. In this case, random search and genetic algorithm explore 20% of the possible combinations.

In Appendix A a study carried out to compare these three different optimization algorithms is illustrated.

## 2.4 Model validation

The validation of a model consists in testing its predictive ability, which should provide similar statistical parameters both for the objects used to train the model and new (in AD) objects in order to be considered stable. Basically, the validation techniques are used to calibrate the model parameters and to verify that the model avoids overfitting.

Usually a part of the dataset is used to build a reduced model (training set), which subsequently is used to predict the remaining part (evaluation or test set).

There are several validation techniques. The most common technique is cross validation. In  $k$ -fold cross-validation the original dataset is partitioned into  $k$  equal-sized disjoint folds. During each run, one of the partitions is chosen for testing, while the rest of them are used for training. This procedure is repeated  $k$  times such that each partition is used for testing exactly once. The performances of the model are measured by summing up the errors of the model across the  $k$  runs. This method allows to abstract the results to the specific partition performed once on the data. Furthermore, cross-validation provides a more consistent estimation of the error, since all the instances in the dataset are used for testing.

Figure 2.11 summarizes the main step in a QSAR development pipeline. Starting from a set of chemicals paired to experimental properties or bioactivities and after a pretreatment of the chemical structures and molecular descriptors calculation, the data are partitioned into a training and a test set. The former will be further partitioned and undergo to a validation procedure in order to calibrate the model's parameters and test the robustness. The test set will be used to mimic the final use of the model which should provide satisfying performance also on unseen molecules falling inside the applicability domain.

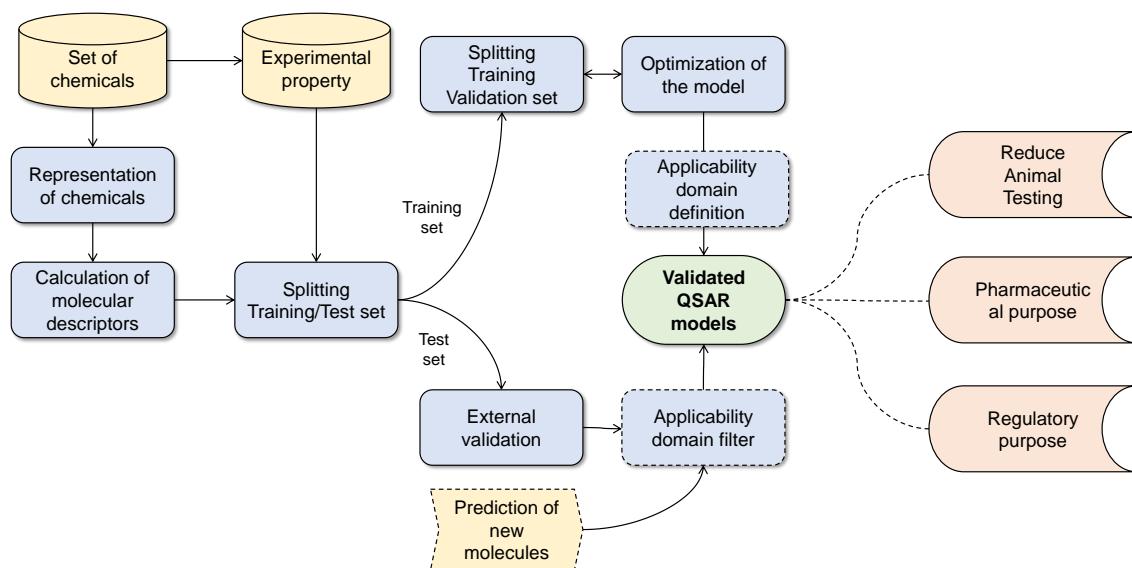


FIGURE 2.11: Pipeline of QSAR development: from experimental data and molecular representation through molecular descriptor computation and data splitting to model optimization and validation.

## 2.5 Consensus modelling

Consensus approaches, also known as high-level data fusion or ensemble approaches, are mathematical and statistical techniques aimed to combine and

integrate information derived from different sources to increase the outcome reliability and overcome limitations of single approaches (Hewitt et al., 2007).

In the framework of QSARs, consensus approaches are generally recognized to reduce the effects of underestimating uncertainties in the prediction of biological activities (Neumann and Gujer, 2008; Weber, VanBriesen, and Small, 2006). The main underlying assumption is that individual models, consider only partial structure-activity information, as encoded by molecular descriptors and adopted algorithms. Thus, the combination of multiple QSAR predictions may provide a wider knowledge and increase the reliability associated with the predictions compared to individual models (Hewitt et al., 2007; Jaworska and Hoffmann, 2010).

Reducing the effects of contradictory information by averaging the predictions of models (Hewitt et al., 2007; Grisoni et al., 2015; Votano et al., 2004) is one of the main advantages of the consensus methods, although this is not always reflected in an improvement of the predictive ability compared to single models (Hewitt et al., 2007; Grisoni et al., 2015). Another advantage of consensus methods is the broadening of the domain of applicability compared to single models.

For these reasons, consensus methods have been extensively applied in QSAR studies (Ballabio et al., 2017; Pradeep et al., 2016; Chauhan and Kumar, 2018; Ruiz et al., 2017; Mansouri et al., 2020; Mansouri et al., 2016). In particular, recent studies on the improvement achieved with large-scale consensus approaches for quantitative models can be found in the literature (Zakharov et al., 2019; Ambure et al., 2019).

In this thesis, two consensus strategies were applied to integrate the predictions provided by individual models: majority voting and the Bayes consensus with discrete probability distributions. These methods are briefly described below.

### 2.5.1 Majority Voting

The family of voting strategies combines the predictions given by independent models with different frequency-based approaches, such as averaging and scoring (Ruiz et al., 2017; Abdelaziz et al., 2016; Marzo et al., 2016).

The most trivial and intuitive voting strategy is the majority voting (MV) rule, which assigns a chemical to the most frequently predicted class among the pool of considered models (Ballabio, Todeschini, and Consonni, 2019; Mansouri et al., 2013). A protective version of voting strategies can be obtained by considering only those predictions with a sufficiently high concordance, based on a user-defined threshold, among the pool of models. In this thesis, we considered three different thresholds and, thus, three different majority voting strategies: (i) majority voting loose (MVL), (ii) majority voting intermediate (MVI), and (iii) majority voting strict (MVS). The 'loose' approach classifies molecules according to the most recurrent class assignment, i.e., with a two-class case, this corresponds to the class predicted with a frequency higher than 50%. The 'intermediate' and 'strict' criteria (MVI and MVS, respectively) are protective approaches. MVS assigns the compound

only if the prediction agreement is higher than or equal to 75%. The MVS approach provides a prediction for a given molecule only if all of the individual models predict the same class (100% agreement). Figure 2.12 provides three examples of assignment.

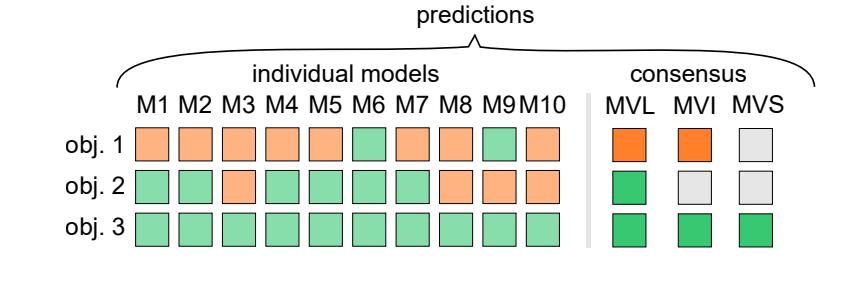


FIGURE 2.12: Example of majority voting strict (MVS), intermediate (MVI) and loose (MVL) applied to predictions provided by ten individual models (M1, M2,..., M10) on three objects (rows). Green and orange colours denote the class belonging.

## 2.5.2 Bayesian Consensus

Probabilistic method, such as Bayesian consensus, offer an alternative to the majority voting approach. The Bayesian rule (Pradeep et al., 2016; Ballabio et al., 2019; Fernández et al., 2012) estimates the prior probability for a molecule to belong to a specific class for each information source and then combines this information to provide a joint probability (Borràs et al., 2015). In particular, the Bayesian consensus with discrete probability distributions (Fernández et al., 2012) initially takes into account the first evidence,  $e$ , which is in this case the class (active or inactive) predicted by the first model. Then, the posterior probabilities  $p(h_g|e)$  that hypothesis  $h_g$  is true given evidence  $e$  are calculated for any class  $g$ , as follows (similarly to 2.7):

$$p(h_g|e) = \frac{p(e|h_g) \cdot p(h_g)}{\sum_g p(e|h_g) \cdot p(h_g)} \quad (2.11)$$

where  $p(e|h_g)$  is the likelihood probability that evidence  $e$  is observed given that hypothesis  $h_g$  is true and  $p(h_g)$  is the prior probability that hypothesis  $h_g$  is true in the absence of any specific evidence. With two hypotheses (i.e., class equal to 'active' or 'inactive'), the prior equal (non-informative) probability is estimated as  $p(h_{ACTIVE}) = p(h_{INACTIVE}) = 0.50$ . The prior proportional (informative) probability for each hypothesis  $h_g$  would be  $p(h_g) = \frac{n_g}{n}$ , where  $n_g$  is the number of molecules belonging to the  $g$ -th experimental class within the  $n$  total molecules.

Likelihood probabilities for each model can be estimated from its confusion matrix, where the numbers of correct and incorrect classifications are collected (Fernández et al., 2012). Once posterior probabilities for the first model have been calculated, the Bayes consensus proceeds with the following iterative procedure. Posterior probabilities of the first model are used as new prior probabilities for the second step, where the class predicted by the

second model is the new evidence  $e$  on the basis of which the posterior probabilities are calculated. These posterior probabilities become the new prior probabilities in the third iteration and so on, until predictions of all models have been used in the consensus process. At the end of the iterations, the posterior probabilities corresponding to the combination of all of the information sources are obtained.

Therefore, the Bayes consensus assigns a probability value to each class, which is then used for prediction, by choosing the class with the maximum posterior probability. As for the majority voting strategies, the Bayes consensus can be used in a protective manner by setting a posterior probability threshold (in this study, 95%) that has to be fulfilled to predict the class (Fernández et al., 2012). When proportional prior probabilities are used with models calibrated on data with unbalanced class distributions.

## 2.6 Metrics for classification performances

The evaluation of the models performances need to be evaluated on the basis of the adopted model strategies as well as on the expected outcome. Since this study was focused on classification problems (i.e., tasks) (Ballabio, Grisoni, and Todeschini, 2018), the assessment of the performance is based on the analysis of the so-called confusion matrix, which encodes the number of both correct and incorrect predictions for each class. From this matrix, several well-established class indices can be derived, such as sensitivity, specificity and precision. These measures describe the classification results achieved on each single class. Other measures are defined for the global estimation of classification performances; accuracy being the most known and used.

Indices of global classification performance are useful to assess the model's quality with a single numerical value. This is particularly needed to assess the performance of multi-task models.

Given a dataset,  $n$  denotes the number of objects and  $G$  the number of experimental classes;  $n_g$  represents the number of objects belonging to the  $g - th$  class, while  $n'_g$  the number of objects predicted in the  $g - th$  class. The classification results can be represented in the confusion matrix, also known as contingency table. It is a square matrix ( $G \times G$ ), whose rows and columns represent experimental and predicted classes, respectively. Each entry  $c_{gk}$  represents the number of objects belonging to class  $g$  and predicted as belonging to class  $k$ . Consequently, the diagonal elements  $c_{gg}$  represent the number of correctly classified objects, while off-diagonal elements represent the numbers of classification errors.

The confusion matrix contains all the information related to the distribution of objects within the classes and to the classification performance. For instance, the number of objects in the dataset ( $n$ ) is equal to the sum of all elements of the confusion matrix:

$$n = \sum_{g=1}^G \sum_{k=1}^G c_{gk} \quad (2.12)$$

TABLE 2.2: Confusion matrix for binary classification problem.

		Predicted class	
		Active	Inactive
Experimental class	Active	TP	FN
	Inactive	FP	TN

Moreover, the number of objects belonging to the  $g - th$  class ( $n_g$ ) corresponds to the sum of the  $g - th$  row elements:

$$n_g = \sum_{k=1}^G c_{gk} \quad (2.13)$$

The number of objects predicted in the  $g - th$  class ( $n'_g$ ) corresponds to the sum of the  $g - th$  column elements:

$$n'_g = \sum_{k=1}^G c_{kg} \quad (2.14)$$

The confusion matrix is usually asymmetric, since the number of objects belonging to class  $g$  and assigned to class  $k$  ( $c_{gk}$ ) may not be equal to the number of objects belonging to  $k$  and assigned to  $g$  ( $c_{kg}$ ). The information on the outcome of the classification modelling contained in the confusion matrix is generally encoded into one or more classification measures (Ballabio, Grisoni, and Todeschini, 2018).

- Primary measures related to single classes. Sensitivity and specificity are two well-known class-based measures that can be used to estimate the classification performance achieved on each class separately. For example, the sensitivity of the  $g - th$  class ( $Sn_g$ ) represents the ability of the given classifier to correctly identify the objects of the  $g - th$  class and is calculated as:

$$Sn_g = \frac{c_{gg}}{n_g} \quad (2.15)$$

The specificity of the  $g - th$  class ( $Sp_g$ ) represents the ability of the classifier to reject objects of other classes, and it is calculated as the ratio of objects not belonging to the  $g - th$  class which were not classified in the  $g - th$  class over the total number of objects not belonging to the  $g - th$  class ( $n_g$ ). Both sensitivity and specificity have values between 0 (no class discrimination) and 1 (perfect class discrimination). When dealing with binary classification, these measures are defined in terms of true/false positive/negative values and objects are usually labelled as positive or negative and the confusion matrix is reduced to a  $2 \times 2$  numerical table with the following structure:

where TP (true positives) is the number of positive objects correctly predicted as positive, TN (true negatives) is the number of negative objects correctly predicted as negative, FP (false positives) is the number of negative objects predicted as positive, and FN (false negatives) is the

number of positive objects predicted as negative. Furthermore, In a binary classification, as in our case, usually the term sensitivity denotes the sensitivity of the active class, while specificity denotes the sensitivity of the inactive class. Hence, sensitivity ( $Sn$ ) is defined as the ratio between TP and the total number of positive objects:

$$Sn = \frac{TP}{TP + FN} \quad (2.16)$$

Specificity ( $Sp$ ) can be defined as the ratio between TN and the total number of negative objects:

$$Sp = \frac{TN}{TN + FP} \quad (2.17)$$

- Global indices derived from primary class measures. Class measures can be aggregated in different ways to calculate global measures of classification performances. The average sensitivity (Non Error Rate, NER) and average precision is calculated as arithmetic mean of sensitivity values of the  $G$  classes.

$$NER = \frac{\sum_{g=1}^G Sn_g}{G} \quad (2.18)$$

In case of binary classification NER can also be expressed as:

$$NER = \frac{Sn + Sp}{2} \quad (2.19)$$

- Global multi-task indices. The model performance on each binary  $t - th$  task was quantified using sensitivity ( $Sn_t$ ), specificity ( $Sp_t$ ) and non-error rate ( $NER_t$ ), defined as follows:

$$Sn_t = \frac{TP_t}{TP_t + FN_t} \quad Sp_t = \frac{TN_t}{TN_t + FP_t} \quad NER_t = \frac{Sn_t + Sp_t}{2} \quad (2.20)$$

where  $TP_t$ ,  $TN_t$ ,  $FP_t$  and  $FN_t$  are the number of true positive, true negative, false positive and false negative molecules for the  $t - th$  task. To compare the overall performance of models, ‘global’ sensitivity, specificity and non error rate measures ( $Sn_T$ ,  $Sp_T$ ,  $NER_T$ ) were computed as follows:

$$Sn_T = \frac{\sum_{t=1}^T TP_t}{\sum_{t=1}^T TP_t + \sum_{t=1}^T FN_t} \quad Sp_T = \frac{\sum_{t=1}^T TN_t}{\sum_{t=1}^T TN_t + \sum_{t=1}^T FP_t} \quad (2.21)$$

$$NER_T = \frac{Sn_T + Sp_T}{2} \quad (2.22)$$

where  $t$  runs over each task, and  $T$  is the total number of tasks.  $S_{n_T}$  and  $S_{p_T}$  represent the fraction or percentage (if multiplied by 100) of active and inactive molecules correctly predicted over all tasks, respectively.

## 2.7 Software

Data curation and integration was performed in KNIME 4.0.1 (Berthold et al., 2009). SMILES were canonicalized using KNIME 4.0.1 ('RDKit Canon SMILES' node). Pocket overlap scores were computed using PocketMatch (Yeturu and Chandra, 2008) in Python v3.6.

Extended connectivity fingerprints (ECFPs) (Rogers and Hahn, 2010) were calculated with Dragon 7 (KodeSrl, 2017) with the following settings: 'Bits per pattern' = 2; 'Count fragments': True; 'Atom Options': [Atom type, Aromaticity, Connectivity total, Charge, Bond order].

WHALES were calculated in Python v3.6 (Van Rossum and Drake Jr, 1995) using freely available script as described in (Grisoni et al., 2018b).

MultiDimensional Scaling and Principal Component Analysis were computed using MATLAB v2018b (The Mathworks Inc) were performed using in-house MATLAB 2018b or Python v3.6 code.

Structural alignment between crystallographic receptors and residue selection was performed using Pymol ('The PyMOL Molecular Graphics System, 2018') and in-house codes.

Single-task models were calculated and optimized in MATLAB 2019b by means of in-house scripts. Published and freely accessible MATLAB code for PCA, N3, and NB, KNN and RF was used, as available on Milano Chemometrics website.

Multitask neural networks were built and optimized by means of the keras (Chollet et al., 2015) module with TensorFlow (Abadi et al., 2015) backend in Python v3.6.

Graph convolutional networks were built and optimized by means of the dgl (Wang et al., 2019) module with PyTorch (Paszke et al., 2017) backend in Python v3.6.

Wilcoxon signed-ranked test and t-test were performed in Python v3.6 using the SciPy library (Virtanen et al., 2020).



## Chapter 3

# Data

Due to their biological relevance, NRs have became the target of numerous computational projects for both toxicological (Khandelwal et al., 2008; Klein-streuer et al., 2017; Mansouri et al., 2016; Mansouri et al., 2020) and medicinal chemistry applications (Grisoni et al., 2018b; Heitel et al., 2019; Motta et al., 2018; Merk et al., 2018a; Park, Kufareva, and Abagyan, 2010; Rupp et al., 2010). These computational projects are often based on machine learning approaches, which are “data-hungry” and require as many training data as possible to reach satisfying levels of predictivity and generalization ability (Halevy, Norvig, and Pereira, 2009).

In this framework, the creation of datasets comprising as many experimental data as possible becomes fundamental. Furthermore, since public and commercial databases can contain up to 10% errors in structural and/or experimental annotations (Fourches, Muratov, and Tropsha, 2010; Young et al., 2008), data curation becomes a key step in order to avoid potential inconsistencies and ensure reliable molecular modeling research (Fourches, Muratov, and Tropsha, 2010; Young et al., 2008).

Nowadays, several freely accessible databases containing information on nuclear receptor modulation for small molecules exist (Gaulton et al., 2017; Reau et al., 2018; Gilson et al., 2016; Tice et al., 2013). However, the type and amount of annotated chemical structures and scaffolds, and the distribution of biologically active or inactive molecules depend on the database focus (Wassermann and Bajorath, 2011). Indeed, each of the available repositories might contain different sets of compounds and investigated targets, and may exhibit a different proportion of modulators and non-modulators (Wassermann and Bajorath, 2011).

This chapter begins with a description of the CoMPARA project data (Mansouri et al., 2020) for androgen receptor, then we will discuss the creation of a new dataset called NURA (Valsecchi et al., 2020b) that includes curated data for eight NRs since it is generally accepted that curated data are a valuable resource for quantitative modeling of the structure-activity relationship and corresponding decision making in medicinal chemistry, toxicology, and related fields (Cronin and Schultz, 2003; Griffen et al., 2018; Tropsha, 2010; Vangala et al., 2011).

### 3.1 CoMPARA dataset

As mentioned before, data for NRs modulators are available from several different sources. Some of them are specialized on specific nuclear receptor (e.g. CERAPP (Mansouri et al., 2016) and CoMPARA (Mansouri et al., 2020) dataset for estrogen and androgen receptor, respectively). Data coming from medicinal chemistry related sources are more focused on very active molecules (activity values lower than 1 uM) and on similar structures, while toxicology related sources are interested in weak activity and collateral effects given for example by EDs.

In this section CoMPARA dataset will be described and characterized as a source of data coming from a toxicology-related project.

CoMPARA is a collaborative project (Collaborative Modeling Project of Androgen Receptor Activity), coordinated by the National Center of Computational Toxicology (U.S.Environmental Protection Agency) (Mansouri et al., 2020). CoMPARA aimed to develop *in silico* approaches to identify potential androgen receptor (AR) modulators. This project involved 25 research groups worldwide, which were provided with a calibration set consisting of 1689 chemicals and the corresponding experimental annotations on binding, agonism, and antagonism activities (in the form of qualitative labels, active/inactive), as determined by a battery of 11 *in vitro* assays coming from Tox21 results (Tice et al., 2013). The research groups were then asked to predict another 55'450 chemicals for one or more endpoints (binding, agonism, and antagonism) using their own developed QSAR models. Finally, these predictions were merged through *ad hoc* consensus approaches, which are currently being used by the CoMPARA coordinators to prioritize experimental tests for potential endocrine-disrupting chemicals.

TABLE 3.1: Number of Chemicals (Total, Actives, and Inactives) included in the CoMPARA Binding, Antagonism, and Agonism Evaluation Sets and Number of Models Developed within the CoMPARA Project for Each Endpoint

	<b>binding</b>	<b>agonism</b>	<b>antagonism</b>
<b>No. chemicals</b>	3540	3667	4408
active	411 (11.6 %)	314 (8.6 %)	164 (3.7 %)
inactive	3129 (88.4%)	3353 (91.4 %)	4244 (96.3 %)
<b>No.models</b>	34	22	21

The predictive ability of individual QSAR models was assessed by the project coordinators on the basis of three specific evaluation sets, which were embedded within the large prediction set of 55'450 chemicals, to carry out a blinded verification. These sets were created from literature data extracted from different sources and curated for quality, by considering target, modality, hit call, and concordance among the annotated values.

As can be noted from Table 3.1, the three data sets are characterized by a prevalence of inactive compounds (more than 88% for all the endpoints). In this case agonists and antagonists can not be always identified as binders,

thus, the number of binders is lower than the sum of agonists and antagonists.

## 3.2 NURA dataset

### 3.2.1 Target selection

Following the retrieval pipeline of CoMPARA data collection, we aimed to create an exhaustive dataset comprising *in vitro* bioactivity data on eight nuclear receptors, selected based on their biological relevance and data availability in public databases. The considered NRs are:

- Androgen receptor (AR), which plays a key role in many sexual, somatic and behavioral functions critical to lifelong health, as well as in the development of several diseases such as prostate cancer and cardiovascular diseases (Davey and Grossmann, 2016).
- Estrogen receptor (ER), which is the main mediator of estrogen action in development and reproductive system as well as in brain function, bone maintenance, cardiovascular system and adipose tissue. Several diseases are associated with this receptor, including osteoporosis, obesity and Alzheimer's disease (Mueller and Korach, 2001).
- Progesterone receptor (PR), which mainly affects the female sexual development end pregnancy and it is a promising target for the treatment of breast cancer, cardiovascular disease, and central nervous system disorders (Huang, Chandra, and Rastinejad, 2010; Schug et al., 2011).
- Glucocorticoid receptor (GR), which plays multiple roles in physiology, e.g., immune mediation, inflammation, glucose balance, the stress response, fat distribution, and normal growth and is involved in the development of several disorders, such as diabetes mellitus, hypertension and cardiovascular diseases (Huang, Chandra, and Rastinejad, 2010).
- Peroxisome proliferator-activated receptor (PPAR), which controls lipid homeostasis with isoform-specific lipid regulation, insulin action and cell proliferation and is linked to obesity, dyslipidemia and atherosclerosis risk (Berger and Moller, 2002; Schug et al., 2011).
- Pregnane X receptor (PXR), which regulates the detoxification and clearance of xenobiotic substances, exerting a protective function (Ekins et al., 2009; Francis et al., 2003). PXR has been associated to cancer, and to inflammatory and metabolic diseases (Banerjee, Robbins, and Chen, 2015).
- Retinoid X receptor (RXR), which regulates metabolic homeostasis and forms heterodimers with numerous other nuclear receptors. Drugs that target RXR heterodimers are used to treat cancer, dermatologic diseases, endocrine disorders, and the metabolic syndrome (Penvose et al., 2019; Shulman and Mangelsdorf, 2005).

TABLE 3.2: Summary of names, ligands, main functions and related diseases of the eight nuclear receptors considered in this thesis.

Acronym	Ligand	Main function	Related disease
AR	androgens	Sexual maturation	Prostate cancer
ER $\alpha$	estrogens	Sexual maturation	Breast cancer
ER $\beta$		and gestation	obesity
FXR	bile acids	Homeostasis	Liver cancer, renal cancer
GR	glucocorticoids	Inflammatory responses, cellular proliferation and differentiation	Metabolic disease cancer
PPAR $\alpha$ , PPAR $\delta$ , PPAR $\gamma$	fatty acids	Lipid metabolism	Obesity, diabetes, atherosclerosis
PR	progesterogens	Regulation of gene expression and cellular proliferation and differentiation	Breast cancer
PXR	steroids	Metabolism and secretion of xenobiotics	Colorectal cancer, liver cancer
RXR	9-cis retinoic acid	Dimerization	Renal cancer, pancreatic cancer

- Farnesoid X receptor (FXR), or bile-acid activated transcription factor, which contributes to the liver physiology and can be targeted to treat metabolic and hepatic disorders (Francis et al., 2003).

Table 3.2 summarizes the main physiological roles and related diseases for the eight most studied nuclear receptors with two and three isoforms for ER and PPAR, respectively.

### 3.2.2 Data collection

In order to merge data from medicinal chemistry and computational toxicology related database, we considered four different sources for data collection, namely:

- ChEMBL25 (Gaulton et al., 2017), which is a large-scale, open database containing drug-like bioactive molecules with *in vitro* bioactivity annotations. For the chosen NRs, we filtered bioactivity data referred to single proteins (Table 3.3) according to the BioAssay Ontology (BAO) signature (Visser et al., 2011) BAO0000190, BAO0000188, BAO0000192, BAO0000034, BAO0000186, BAO0000199, BAO0002583, BAO0002809. As an additional filter, we used ChEMBL25 confidence score, which is based on the assessed record quality and ranges from 0 (non-curated

data entries), to 9 (high-quality data), to retain compounds with confidence score greater than 8 (Gaulton et al., 2017). Records annotated as “potential transcription error” were removed (7 records). Records with exhaustive assay type information were retained.

- BindingDB (Gilson et al., 2016), which is a public database of measured binding affinities focusing on small, drug-like ligands; bioactivity data referred to nuclear receptors.
- NR-DBIND (Nuclear Receptors - DataBase Including Negative Data) (Reau et al., 2018), which is a repository dedicated to drug-like nuclear receptor ligands. All the data referred to the selected NRs were collected.
- Tox21 (NIH). The Tox21 (Toxicology in the 21st Century) program (Kleinsteuer et al., 2017) adopts high-throughput screening (HTS) *in vitro* techniques to test large numbers of chemicals that could be toxic *in vivo*. For this purpose, Tox21 established a library of 10 K chemicals - composed of environmental chemicals and approved drugs - which has been screened against different cell-based assays. Some of these assays (Table 3.3) focused on nuclear receptor modulation were considered in our study. In particular, we collected the NR-related data of Tox21 from PubChem BioAssay Repository (Kim et al., 2019). Molecules labelled as antagonists in agonism assays (or as agonists in antagonism assays) were removed. Records with inconclusive readouts were removed.

For ER and PPAR, for which more than 1000 isoform specific annotations were retrieved, isoform related bioactivity data were collected separately (i.e, alpha and beta isoforms for ER; and alpha, delta and gamma for PPAR), obtaining a total of 11 macromolecular targets (AR, ER $\alpha$ , ER $\beta$ , PR, GR, PPAR $\alpha$ , PPAR $\delta$ , PPAR $\gamma$ , PXR, RXR, FXR). For the selected targets, we collected *in vitro* data referred to binding, agonistic and antagonistic effects (referred to as “endpoints”), obtaining a total of 33 endpoints. We retained the database entries corresponding to the following two types of experimental readouts: (i) half maximal concentration on the dose-response curve for inhibition or effect ( $IC_{50}$  and  $EC_{50}$ , respectively) and (ii) the dissociation and inhibition constants ( $K_d$  and  $K_i$ ), which describe the affinity between a ligand and a protein (with  $K_d$  measuring the equilibrium between the ligand-protein complex and the dissociated components, while  $K_i$  being specific for inhibitors). For Tox21, the activity concentration at half-maximal response ( $AC_{50}$ ) as determined by a panel of *in-vitro* assays (Tice et al., 2013) was considered.

### 3.2.3 Data curation

Data from different sources were collected and arranged in a record with the following format: (i) ligand molecular structure (expressed as SMILES strings (Weininger, 1988)), (ii) experimental readout (including the unit of measure and the experimental response value), (iii) effect type, if available (agonism, antagonism, binding), (iv) target organism and (v) target nuclear

TABLE 3.3: Summary of the considered data sources. Receptor acronym, number of *in vitro* records, PubChem Assay ID and ChEMBL ID are reported. For targets having more than 1000 isoform-specific records (i.e., ER and PPAR), the experimental data referred to each isoform was collected separately.

Target	No. bioactivity records				PubChem Assay ID	ChEMBL ID
	Tox21	ChEMBL	BindingDB	NRDB.		
AR	10486	8095	4591	1,513	743053, 743063, 743054	CHEMBL1871
ER $\alpha$	10486	11148	1308	2,054	743053, 743078, 743091	CHEMBL206
ER $\beta$	10486	7749	9151	1,826	1259394, 1259396	CHEMBL242
FXR	9305	3769	2552	136	743239, 743240	CHEMBL2047
GR	10486	11934	217	1,935	720719, 720725	CHEMBL2034
PPAR $\alpha$	0	7108	3929	1,018	n.a.	CHEMBL239
PPAR $\delta$	10486	4941	2118	525	743227, 743226	CHEMBL3979
PPAR $\gamma$	10486	11362	2118	1,454	743140, 743199	CHEMBL235
PXR	9667	1964	659	1	1347033	CHEMBL3401
PR	9667	5239	3324	1,403	1347036, 1347031	CHEMBL208
RXR	9667	3564	5406	340	1159531	CHEMBL2061, 1870, 2004

receptor (among the 11 selected, isoforms for PPAR and ER included). On each of these records, the data curation procedure was carried out with the following sequential steps:

1. Only records referred to *Homo sapiens* were retained;
2. Records with the experimental readout expressed as  $EC_{50}$ ,  $IC_{50}$ ,  $AC_{50}$ ,  $K_i$  or  $K_d$  were retained;
3. All records referring to disconnected structures, salts, mixtures, inorganic compounds and compounds containing elements different from H, C, N, O, F, Br, I, Cl, P or S were removed. All the structures were converted into canonical SMILES strings (O’Boyle, 2012);
4. Each record was assigned a discrete bioactivity label, according to its experimental readout, as follows: (i) “active”, for experimental bioactivities equal to or lower than 10,000 nM; (ii) “weakly active”, for activity values between 10,000 and 100,000 nM; (iii) “inactive”, for entries with activity values exceeding 100,000 nM. Records containing a range of potency (specified as ‘<’ or ‘>’) were retained only if the specified range was either lower than 10,000 nM or higher than 100,000 nM (and subsequently assigned to the “active” or “inactive” classes, respectively);
5. For each target, records referred to the same molecule (as identified by the canonical SMILES string) were merged. The information obtained from such multiple records was used to assess the reliability of the assigned label(s) for a given molecule on a given target. If all the records for a molecule showed the same bioactivity label on a target, the molecule-target pair was retained in the dataset. Molecules having conflicting labels in the corresponding records for a given target and a given endpoint (e.g., presence of both “active” and “inactive” labels) were assigned the label “inconclusive”, to highlight the lack of a final bioactivity assessment. Whenever a molecule was retained for at least one of the macromolecular targets, the lack of collected bioactivity information for other targets was identified with the label “missing”.

### 3.2.4 Data analysis

The contribution of the individual data sources to the final dataset, in terms of novel and shared molecules and molecular scaffold diversity and the distribution of bioactivity labels for each selected target was analysed. Finally, a ligand- and structure-based analysis allows to obtain some additional data-driven insights into the captured structure-activity landscapes. Each database source provided a different contribution to the final dataset.

The final dataset contained (Figure 3.1):

- 6504 molecules from Tox21, with 150,571 activity labels in total (1.73% active, 3.76% weakly active and 94.51% inactive), no activity labels for RXR antagonism were retained;

- 3951 molecules from ChEMBL, with 12,159 activity labels in total (82.29% active, 6.28% weakly active and 11.43% inactive);
- 5491 molecules from NR-DBIND, with 13,711 activity labels (90.74% active, 3.80% weakly active and 5.46% inactive), no activity labels for PPAR $\alpha$  were retained;
- 1125 molecules from BindingDB, with 1570 activity labels (81.21% active, 6.18% weakly active and 12.61% inactive), no activity labels for RXR and PXR antagonism and for PPAR $\alpha$  were retained

As already seen for CoMPARA data, Tox21 data contains a larger number of inactive compounds (94.51%), mostly due to its focus on toxicological evaluation of man-made chemicals. On the contrary, medicinal chemistry databases focus mostly on bioactive compounds (82.29%, 90.74% and 82.29% for ChEMBL, NR-DBIND and BindingDB, respectively) (Reau et al., 2018).

We extracted the most frequent atomic molecular scaffolds (Bemis and Murcko, 1996) for each source to investigate the structural similarity between the molecules annotated in the considered data sources. 1713 molecules out of 15,247 (11%), corresponding to 701 unique scaffolds out of 4334 (16%), are shared among two or more sources (Figure 3.1B). Most of the common scaffold (576, corresponding to 13% of the total) are shared between the sources aiming at medicinal chemistry applications, i.e., ChEMBL, BindingDB and NR-DBIND. This reflects a certain similarity of the chemical space covered by these sources. At the same time, each source contributes with unique atomic scaffolds i.e., 1680 novel scaffolds contained in Tox21 (39%), 803 in ChEMBL (19%), 1043 in NR-DBIND (24%) and 107 in BindingDB (2%). These aspects highlight the benefit of merging different sources to expand the atomic scaffolds covered in the curated dataset.

Additionally, Tox21 covers a significantly different property space (Figure 3.1C i.e., molecular weight, lipophilicity, number of aromatic rings and rotatable bonds;  $p < 0.05$ , t-test) than the other databases.

In order to further investigate the overlap of the considered chemical sources, we represented the chemical space of NURA dataset by means of a multidimensional scaling (MDS), which compresses the information on molecular similarity in a two-dimensional plot (Figure 3.2A). In this representation, regions mainly characterized by molecules labelled as active can be identified (Figure 3.2A). These regions correspond in particular to the overlap between ChEMBL, BindingDB and NR-DBIND molecules (Figure 3.2B). The main region of overlap (Figure 3.2B, i) contains drug-like compounds with heteroaromatic rings and alicyclic compounds with alkyne bonds and hydroxyl functional group, while a smaller region of overlap contains alicyclic compound with hormone like scaffolds (Figure 3.2B ii). On the contrary, Tox21 molecules occupy different regions, mostly characterized by inactive molecules, which also contain inactive linear aliphatic compounds (Figure 3.2B, iv). Finally, BindingDB contributes a unique set of molecules binding to RXR (Figure 3.2B, iii), containing triazole and pyridine heterocycles with fluorine substituents.

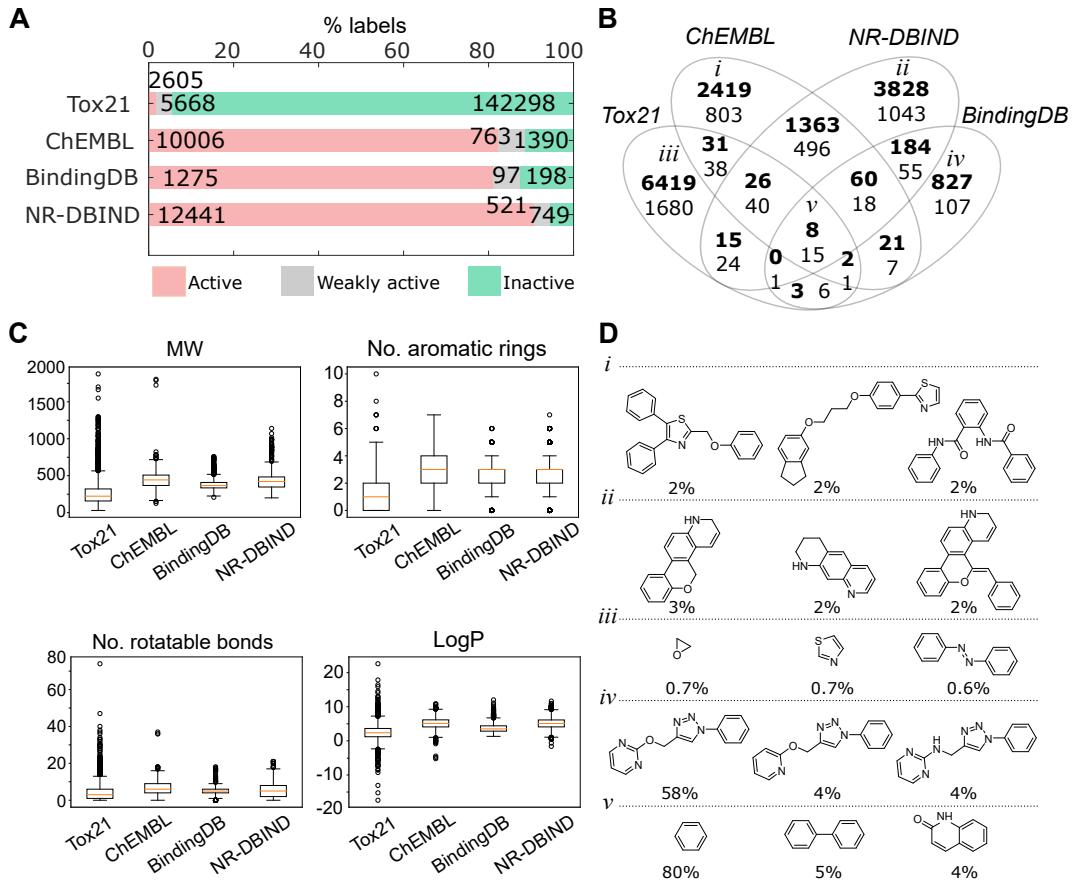


FIGURE 3.1: Analysis of the individual sources used to develop NURA dataset. (A) Percentage of records labelled as active (activity lower than 10,000 nM), weakly active (activity between 10,000 nM and 100,000 nM) and inactive (activity higher than 100,000 nM) grouped by data source. (B) Venn diagram of the data collected from Tox21, ChEMBL, NR-DBIND and Binding-DB. The numbers of shared and not shared molecules (in bold) and scaffolds are reported; (C) distribution of molecular weights (MW), number of aromatic rings, rotatable bonds and octanol-water partition coefficients (LogP) per data source. Tox21 molecules have statistically significant ( $p < 0.05$ , t-test) values in the computed properties compared to the other data sources. (D) Three most frequently occurring scaffolds present in only one source and in all sources (the frequency reported as percentage). Roman numerals correspond to the set the scaffolds belong to, as specified in (B).

The aggregation and curation of data from the selected sources led to a dataset containing 15,247 molecules with activity annotation for 33 endpoints, i.e., 11 NRs with the respective labels for three activity modulations (“binding”, “agonism” or “antagonism”, Figure 3.3). Each endpoint contains on average 4.5 k molecules with annotated activity. The endpoints with the highest number of annotations are the PPAR $\gamma$  binding (7362 molecules), GR binding (7128 molecules) and ER $\beta$  binding (6779 molecules). The endpoints

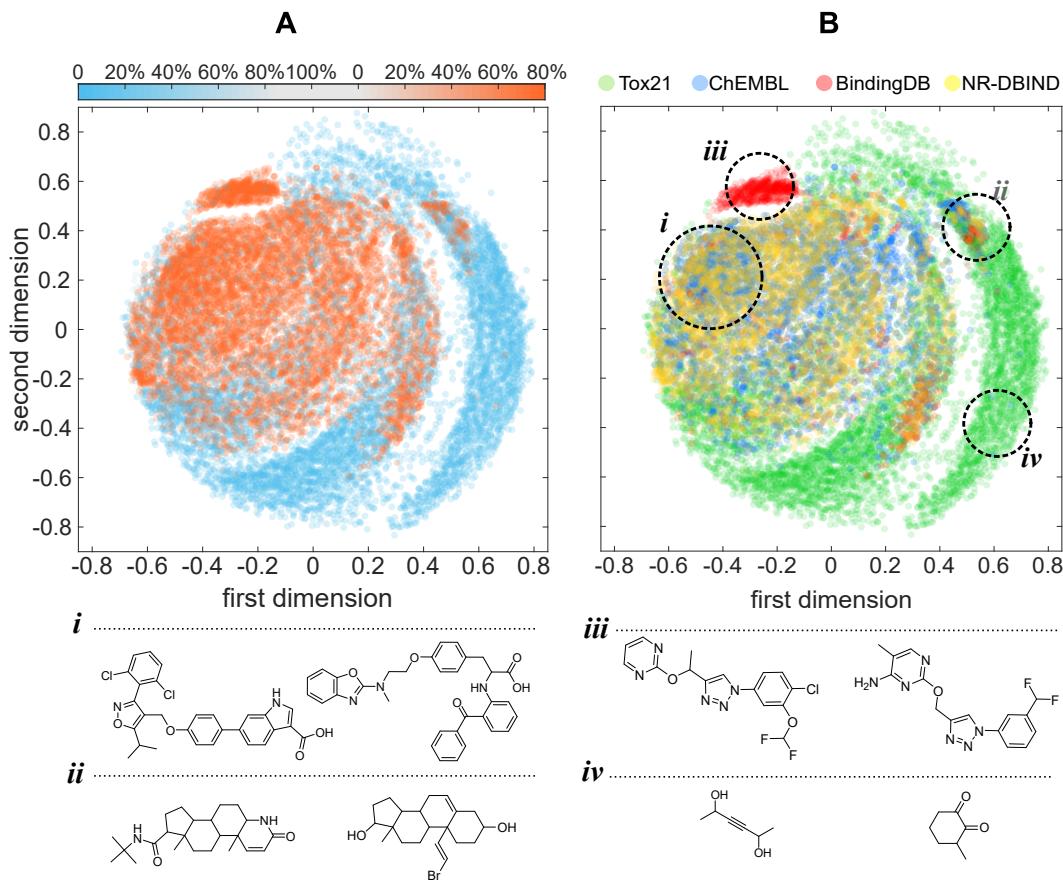


FIGURE 3.2: Multidimensional scaling of the molecules in the curated NURA dataset, as obtained on ECFPs (stress error = 0.356); (A) molecules are colored based on their percentage of “active” labels over their total number of annotated endpoints; (B) molecules are colored based on the original source, with some representative structures highlighted.

relative to antagonism on RXR (119 molecules), PPAR $\alpha$  (19 molecules) and PXR (10 molecules) contain the lowest number of annotations. The dataset contains a different balance between active and inactive chemicals depending on the endpoint considered. For instance, antagonism on RXR, binding and agonism on PPAR $\alpha$  show the largest percentage of molecules labelled as active (96.7%, 89.2%, 90.9%, respectively), while the endpoints relative to PPAR $\delta$  antagonism, PPAR $\gamma$  antagonism, and FXR antagonism mainly comprise inactive molecules (99.1%, 95.6% and 94.8%, respectively). 87% of the molecules have an activity label for at least two endpoints, with an average of 11 annotations (over the 33 endpoints) per molecule.

### 3.2.5 Data driven insights

Data analysis is a key step in every machine learning pipeline. Therefore, to gain more information on the data and verify if the NURA dataset is suitable

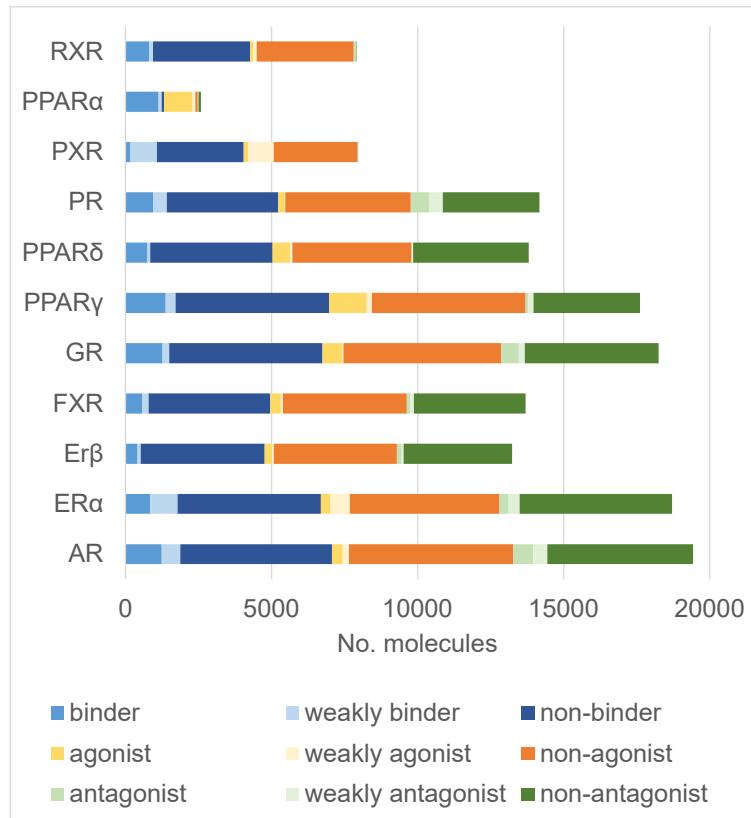


FIGURE 3.3: Distribution of molecules per considered nuclear receptors in the curated NURA dataset, divided into active (activity lower than 10,000 nM), weakly active (activity ranging from 10,000 nM to 100,000 nM) and inactive (activity larger than 100,000 nM).

for multi-task learning (Xu et al., 2017) two types of *post-hoc* analysis were performed:

- a “ligand-centric” analysis, which is aimed to identify active ligands shared among different endpoint. We considered the number of active molecules shared between pairs of endpoints as a measure of overlap. For any given pair of endpoints ( $i$  and  $j$ ), the overlap in their activity annotations ( $S_{ij}$ ) was calculated using the following index:

$$S_{ij} = \frac{a}{a + b} \quad (3.1)$$

where  $a$  is the number of molecules active for both endpoints  $i$  and  $j$ ;  $b$  is the number of molecules with different activity labels for  $i$  and  $j$ . Therefore,  $S_{ij}$  gives the fraction of molecules annotated as actives in both endpoints, without considering the presence of shared inactive molecules. Note that weakly active molecules were not considered in this analysis.

- a “pocket-centric” analysis, which is aimed to identify correlation patterns of the binding pockets. The selected NRs were evaluated for the overlap of their binding pockets, using the PocketMatch algorithm (Yeturu and Chandra, 2008). PocketMatch compares the binding site in a frame-invariant manner, by calculating 90 lists of sorted distances capturing the shape and the chemical nature of the site. The algorithms provide a score (PMscore) ranging from 0 to 100 for any considered pair of binding sites; the greater the score, the higher the overlap. For this analysis we used the crystallographic structures of ligand-nuclear receptor complexes from the PDBbind database (“PDBbind 2018”) summarized in Table 2.1.

As expected, binding-agonism and binding-antagonism pairs for the same nuclear receptor are characterized by high overlap of active molecules (fraction of common active molecules higher than 0.66 and 0.85, respectively Figure 3.4), while little to no overlap is present for agonism-antagonism pairs (lower than 24% of shared actives for all targets). The only exception is RXR, where only one molecule is shared (5-Fluorinated trienoic acid), which behaves as both agonist and antagonist (Gernert et al., 2003). AR, GR and PR, as well as PPAR and ER isoforms show an high fraction of common active molecules, i.e. 0.85, 0.91, and 0.94 between AR-GR, AR-PR and GR-PR for binding, respectively; 0.95 between ER $\alpha$ -ER $\beta$  for binding; 0.99, 0.99 and 0.88 between PPAR $\alpha$ -PPAR $\delta$ , PPAR $\alpha$ -PPAR $\gamma$  and PPAR $\delta$ -PPAR $\delta$ , respectively.

To analyze the physico-chemical, volumetric and geometrical diversity of the binding pockets of the chosen receptors, we calculated the median of the PMscores for each pair of targets (Figure 3.4B). The diagonal values are computed using different crystallographic structures of the same receptor, and, thus, they represent both the experimental uncertainty in the crystallographic structure and the pocket flexibility. The lowest diagonal scores are those of ER $\alpha$  and PPAR $\alpha$ . Despite the studied receptors belong to the same superfamily, some diversity in the pocket features can be observed, with the lowest PMscore being equal to 62. This highlights a good coverage of the dataset in terms of included receptors, which might possess relatively different binding pockets. This binding pocket analysis might be an additional support to complement structure-activity investigations in the field of polypharmacology and/or selectivity optimization for nuclear receptors.

To further compare the ligand-centric and the pocket-centric analysis, we applied hierarchical clustering to both ligand-based overlap scores (Figure 3.4C) and structure-based scores (PMscores, Figure 3.4D).

Despite molecules annotated as binders might not necessarily bind in the orthosteric site, a good overlap between the ligand-based and structure-based hierarchical clustering can be observed (Figure 3.4). The ligand- and structure-based dendograms reproduce some of the known evolutionary relationships, i.e., among ER $\alpha$  and ER $\beta$ , PPAR subtypes, or among the steroid hormone receptors GR, PR, and AR (Mangelsdorf et al., 1995; Edman et al., 2015). The good correspondence between the ligand- and structure-based information indicates a good coverage of the obtained dataset, in terms of

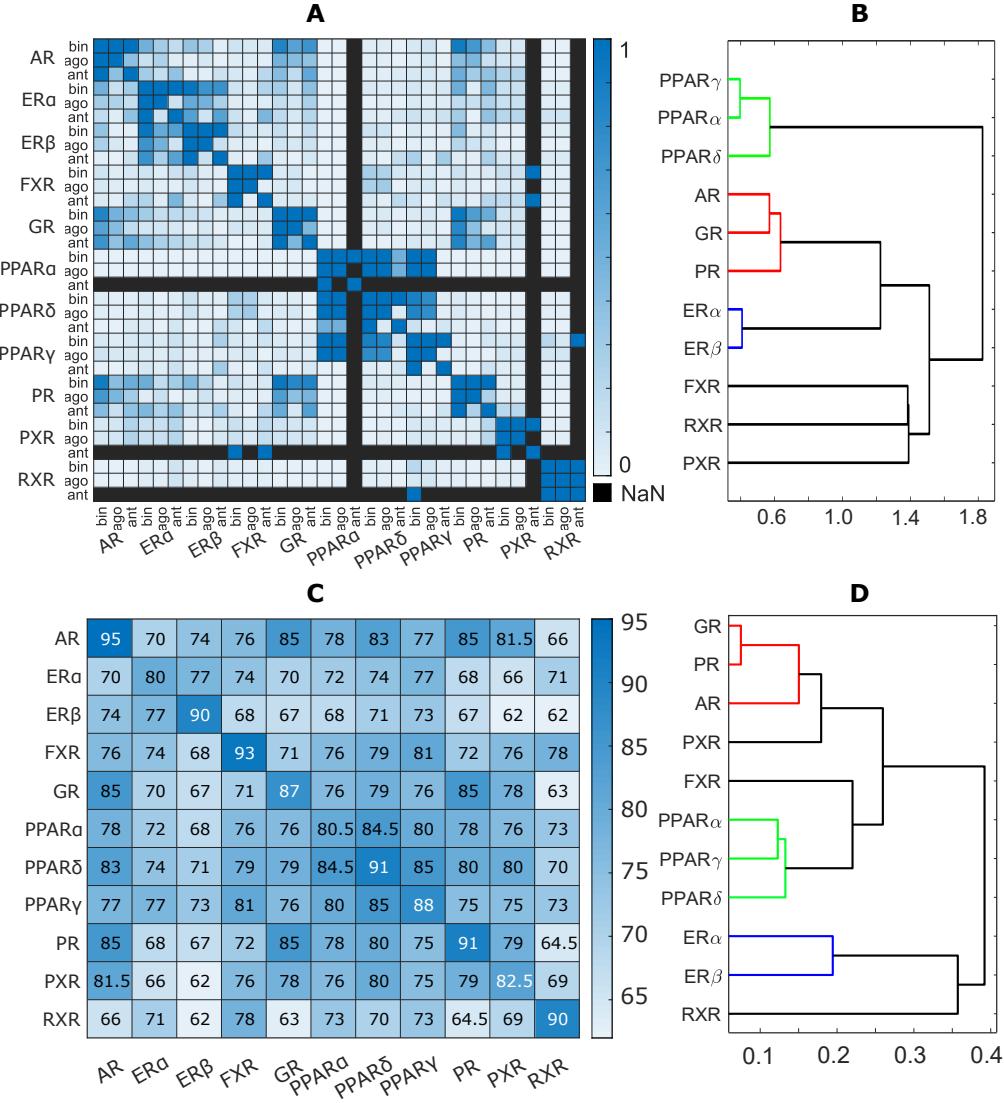


FIGURE 3.4: Summary of the data-driven analysis. (A) Heatmap of the degree of shared active molecules calculated on pair of endpoints considering only active and inactive labels. The colors indicate the degree of shared active molecules from blue (high) to white (low) scores. Black color highlights endpoints that do not have any ligand in common (“NaN”). (B) Dendrogram derived from the degree of shared active molecules calculated on pair of endpoints and targets considering only active and inactive labels. (C) Heatmap of the median of the PMscores for each pair of targets, the darker, the higher the PMscore. (D) Dendrogram derived from the median of the PMscores for each pair of targets, the darker the blue the higher the PMscore.

structure-activity relationships represented. The observed correspondence is, in fact, not always obtained by considering each source separately.

The overlap between the analyzed endpoints (both in terms of ligands and pockets) makes NURA suitable for the application in multi-task learning approaches. Indeed, the target correlation play a fundamental role when dealing with neural networks multi-task learning (Caruana, 1997; Ramsundar et al., 2015; Sadawi et al., 2019; Xu et al., 2017).

### 3.3 Summary of results and concluding remarks

Studying data collected in the CoMPARA project (Mansouri et al., 2020), a need of an integration between medicinal-chemistry and toxicology-related database was underlined. Hence, aiming to build a comprehensive dataset on nuclear receptor bioactivity, we integrated and curated information on binding, agonism and antagonism for 11 selected nuclear receptors, using four well-known chemical databases. The resulting dataset, NURA, includes 15,247 molecules with binding, agonism and/or antagonism annotations for 11 NRs, with 11 endpoints annotated on average per molecule. The data curation and aggregation pipeline successful allowed to bridge the gap between toxicology-related databases (containing information mostly on inactive molecules) and medicinal-chemistry-related databases (mostly focusing on the chemical space of bioactive compounds). Our results show that NURA dataset is enriched in terms of number of molecules, structural diversity and covered atomic scaffolds compared to the single sources.

NURA dataset is the most exhaustive collection of small molecules annotated for their modulation of the chosen nuclear receptors. NURA can serve as a basis to develop machine learning methods for toxicological and/or medicinal chemistry applications, e.g., to predict the modulation of a panel of receptors or the selectivity among the selected NRs. In fact, the increased coverage of the chemical and bioactivity space and of atomic scaffolds offers the opportunity to develop models with an increased applicability domain and improved robustness compared to those developed on the single sources of data. Moreover, for most of the receptors, the data aggregation improved the balance between active and inactive molecules. NURA dataset can be downloaded for free via [Zenodo](#).

In the next chapter, CoMPARA evaluation set is considered to investigate the benefit of consensus approaches, while NURA dataset is the data basis to enhance the prediction of NRs modulators as described in the following chapters.

## Chapter 4

# Consensus analysis of CoMPARA models

As mentioned in Chapter 2, consensus approaches aim to combine and integrate information derived from different sources to increase the outcome reliability and overcome limitations of single approaches. In the framework of QSARs, they are generally recognized as valuable tools to reduce the effects of underestimating uncertainties in the prediction of biological activities (Neumann and Gujer, 2008; Weber, VanBriesen, and Small, 2006).

The main underlying assumption of consensus modeling in QSAR is that individual models, due to their reductionist nature, consider only partial structure-activity information, as encoded by the considered molecular descriptors and adopted algorithms. Thus, the combination of multiple QSAR predictions may provide a wider knowledge and increase the reliability associated with the predictions compared to individual models (Hewitt et al., 2007; Jaworska and Hoffmann, 2010).

Although recent studies on the improvement achieved with large scale consensus approaches for regression models can be found in the literature (Zakharov et al., 2019; Ambure et al., 2019), to the best of our knowledge, no thorough evaluation of the consensus versus single classification models performance has been carried out to date, since only a few QSAR models are usually available for the same endpoint (Baurin et al., 2004; Hanser et al., 2016; Mansouri et al., 2016).

The present study was based on the evaluation sets of CoMPARA (Mansouri et al., 2020) project, which comprise experimental values on androgen receptor (AR) modulation and corresponding QSAR predictions, namely, (i) binding to AR (34 QSAR models), (ii) AR antagonism (22 QSAR models), and (iii) AR agonism (21 QSAR models). CoMPARA was chosen as a test system due to the large availability of diverse QSAR-based predictions.

In the framework of CoMPARA project, two *ad hoc* consensus approaches were applied by combining predictions with a weighting score based on the goodness-of-fit, predictivity, and robustness of models (Mansouri et al., 2020).

However, the aim of the reported study was not a comparison with these former consensus approaches, which were specifically targeted to screen and prioritize chemicals for endocrine activity, but the systematic investigation of the advantages of further consensus strategies compared to single QSAR

models. To this end, approaches with varying levels of complexity (majority voting and Bayesian methods, in both protective and non-protective versions) were considered.

Furthermore, we investigated the influence of the exclusion of the worst-performing models on the consensus outcome, in terms of chemical space coverage and predictive performance (Chauhan and Kumar, 2018; Asturiol, Casati, and Worth, 2016).

Finally, a structural similarity analysis was carried out to identify specific chemical regions where individual QSAR models, and the respective consensus outcome, fail in their predictions.

The three CoMPARA evaluation sets included 3540 chemicals annotated with binding activities, 4408 with agonism, and 3667 with antagonism. All evaluation sets are characterized by unbalanced sample distribution toward inactivity with 88.4, 91.4, and 96.3% of inactive chemicals for binding, antagonism, and agonism, respectively as reported in Table 3.1.

Although the project coordinators also provided quantitative binding, agonism, and antagonism activities, the participants developed only a few regression models (five, five, and three for binding, agonist, and antagonist, respectively). For this reason and since this thesis is focused on classification problems, we considered, only classification models for consensus approaches to allow for a comprehensive and systematic analysis.

CoMPARA consortium members trained QSAR models to classify chemicals for their potential of AR binding (34 models), agonism (21 models), and antagonism (22 models). Models were mainly developed on the same calibration set of 1689 chemicals, using different modeling strategies (e.g., artificial neural networks, k-nearest neighbors, support-vector machines, partial least squares discriminant analysis, classification trees) and molecular descriptors (e.g., binary fingerprints and nonbinary descriptors) (Mansouri et al., 2020).

Each submitted prediction was associated with the applicability domain (AD) assessment. The percentage of reliably predicted chemicals (coverage, Cvg, i.e. percentage of molecules in AD) was used as an additional criterion to assess the model performances.

## 4.1 Individual QSAR Models

Figure 4.1 summarizes the distribution of the classification estimators of the individual CoMPARA models for the three modeled endpoints.

All models have a good predictive performance, with the median NER ranging from 71.0% (antagonism) to 83.8% (agonism). All models identify better inactive than active compounds and, thus, specificity values (Sp) are always higher than sensitivities (Sn). Except for the agonism endpoint, sensitivity is associated with an higher variability than specificity, with values ranging from ~ 20 to ~ 80% on both binding (relative standard deviation equal to ~ 28%) and antagonism (relative standard deviation equal to ~ 29%) endpoints. The main reasons of this general behavior can be identified

in the data class unbalance for binding and antagonism endpoints, which are strongly skewed toward inactivity (88.4 and 91.4% of inactive molecules for binding and antagonism data sets, respectively; Table 3.1), and in the differences in the ranges of testing between training and evaluation sources, as reported in the literature (Mansouri et al., 2020).

The models for agonism show the best trade-off between sensitivity (Sn) and specificity (Sp), with most models characterized by sensitivity values in the range of ~ 70 to ~ 84% and specificity in the range of ~ 76 to ~ 100%.

In particular, agonism models have the highest median sensitivity (76.2%), specificity (96.3%), and NER (83.8%), although the agonism data set includes only 3.7% of actives and is thus the most unbalanced among the three evaluation sets (see Table 3.1).

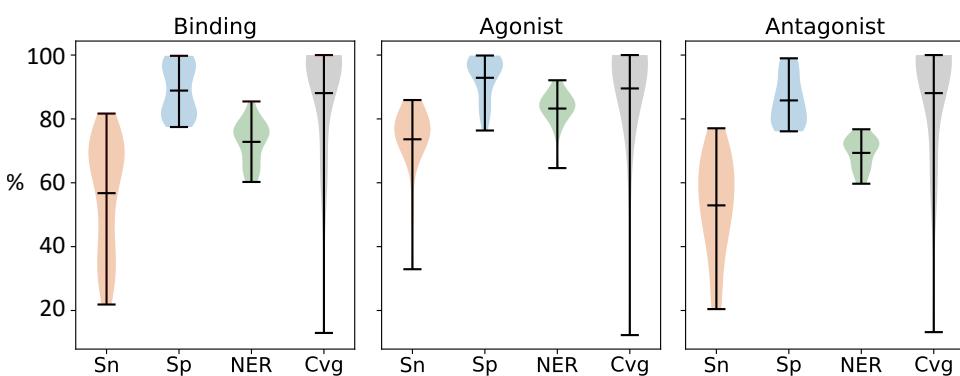


FIGURE 4.1: Violin plots of sensitivity (Sn), specificity (Sp), non-error rate (NER), and coverage (Cvg) for the individual CoMPARA models on the binding, antagonism, and agonism evaluation sets. Thin black lines indicate the first and fourth quartiles and the mean values. Shapes indicate the underlying data distribution.

Models for binding and antagonism have similar median NERs (74.8 and 71%, respectively), moderately low median sensitivities (64.1 and 55.9%), and high median specificities (88.3 and 85.5%).

The majority of individual models are characterized by a high percentage of reliably predicted chemicals (coverage values equal to 88.1, 88.1, and 89.5% on average for binding, antagonism, and agonism, respectively).

The models with the lowest coverage are associated with the highest classification performance, thus confirming that high classification performance is more likely on a narrow applicability domain.

In fact, the four best models to predict the binding activity (NER higher than 80%) were characterized by a limited percentage of chemicals in their applicability domain (coverage values equal to 13, 43.7, 60.7, and 69%), suggesting that these single models have limited applications for prioritization purposes.

## 4.2 Analysis of Consensus Strategies

The selected five consensus strategies (i.e., Bayes [B], protective Bayes [Bp], majority voting loose [MVL], majority voting intermediate [MVI], and majority voting strict [MVS]) were used to integrate the predictions of the individual QSAR models for binding, antagonism, and agonism endpoints.

When applying protective consensus strategies, the outcome predictions were rejected if related to potential uncertainty, that is, (i) prediction agreement lower than 75 and 100% for MVI and MVL, respectively, and (ii) posterior probability lower than 95% for protective Bayes.

For majority voting loose (MVL), no prediction was provided in the case of equal frequency for the two classes (50%).

In analogy with the individual models, the consensus approaches were evaluated for their classification performance, in terms of sensitivity ( $Sn$ ), specificity ( $Sp$ ), non-error rate (NER), and coverage (Cvg) (Table 4.1).

A graphical comparison with individual models is reported in Figure 4.2 with plots of sensitivity versus specificity values. Moreover, since sensitivity, specificity, and coverage have the same unit scale and optimality direction (i.e., ranging from 0 to 100%; the closer to 100%, the better), an overall performance index was calculated as their arithmetic average, denoted as “Utility” in the framework of ranking analysis and multicriteria decision making (Sailaukhanuly et al., 2013; Keeney, Raiffa, and Meyer, 1993; Hendriks et al., 1992).

Both consensus and individual QSARs were ranked for decreasing values of Utility (last row in Table 4.1). On average, consensus strategies have better NERs than individual QSAR models, without a substantial losses in terms of coverage compared to individual models; additionally, consensus models are always ranked among the top 10 positions (Table 4.1). The exception is MVS, which provides a remarkably lower coverage (lower than 52% for all of the endpoints), due to the required 100% agreement among multiple predictions (up to 34 predictions).

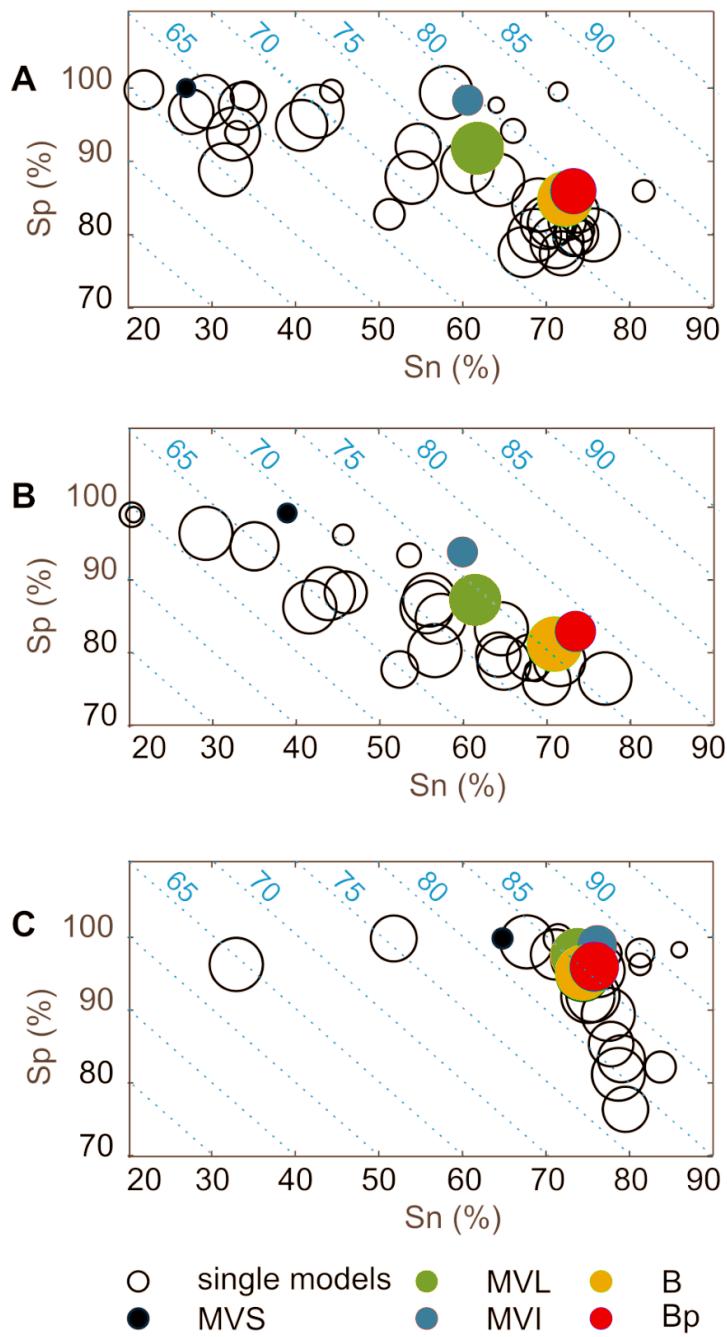


FIGURE 4.2: Plot of sensitivities (Sn) versus specificities (Sp) for the individual models (black empty circles) and for the consensus approaches (filled, colored circles) for each endpoint: (A) binding, (B) antagonism, and (C) agonism. Orange, red, blue, green, and black circles indicate Bayes (B), Bayes protective (Bp), majority voting intermediate (MVI), loose (MVL), and strict (MVS) consensus, respectively. The size of the circles is proportional to the coverage (Cvg); the smaller a circle, the lower the coverage. Isolines represent NER variations (5% steps).

The low coverage of MVS, however, was not counterbalanced by a better performance compared to the other consensus approaches. MVS, in fact, showed the lowest NER and Cvg values among all of the tested consensus strategies. For these reasons, MVS was not analyzed further in this framework. Except for MVS, the other consensus strategies have generally shown a better trade-off between classification performance and chemical space coverage than individual QSARs. For instance, the two single models for binding endpoint falling in the upper-right region of the sensitivity versus specificity space (Figure 4.2A) provided the best predictive performance for binding, with NERs equal to 85.5 and 83.8%, respectively, but they cover only a limited portion of the chemical space, i.e. they have a small coverage (43.7 and 60.7%, respectively). On the other hand, the protective Bayes (Bp) reached a slightly lower NER (79.6%) but a significantly higher coverage (96.1%).

The models on binding and antagonism (Figure 4.2A,B) endpoints are characterized by the unbalanced specificity and sensitivity values, with several models showing high specificity ( $Sp > 90\%$ ) and low sensitivity ( $Sn < 50\%$ ). For these endpoints, consensus methods achieved more balanced values of sensitivity and specificity, due to the compensation in the integration of diverse sources of information. This is particularly evident in the case of the Bayes approaches (Table 4.1), ranked as the best overall approach for both binding and antagonism, and confirms that the uncertainty can be reduced by the integration of conflicting sources. The difference in the performance between consensus and individual QSARs is less pronounced when considering agonism (Figure 4.2C), since the individual models have more homogeneous NERs and balanced Sn and Sp values compared to the other case studies.

Therefore, consensus methods converged to similar performances. Majority voting approaches derive the high specificity values of individual models for both binding and antagonism endpoints, while the Bayes consensus led to a higher sensitivity. This trend could be caused by the low false positive rates of individual models (Figure 4.1) and the way this information is weighted and integrated into the Bayes calculation (eq 2.11). Thus, in this framework, if a compound is predicted with an equal frequency as active and inactive by the individual models, it will be more likely assigned to the active class by the Bayes consensus. Protective approaches (MVI and Bp) yielded slightly better results in terms of the classification performance (NER) compared to their non-protective counterparts, but with a relatively larger loss in coverage (up to 18.7% loss), especially when dealing with majority voting schemes. This explains the worse position within the ranking of protective approaches with respect to non-protective ones (Table 4.1). As an example, the MVL approach on the binding endpoint led to an NER of 76.8% and a coverage of 99.3% (rank 4), while the protective MVI led to a slightly higher NER (79.5%) but considerably lower coverage (80.6% and a worse rank (8).

To evaluate the existence of potential associations between misclassifications and structural chemical features, molecular structures were encoded

TABLE 4.1: Classification Performance of the Consensus Approaches for Binding, Agonism, and Antagonism Endpoints. For each consensus approach, sensitivity (Sn), specificity (Sp), non-error rate (NER), coverage (Cvg), and total ranking are reported. B, Bayes; Bp, protective Bayes; MVL, majority voting loose; MVI, majority voting intermediate; MVS, majority voting strict.

Endpoint	Performance	consensus approach				
		MVL	MVI	MVS	B	Bp
<b>Binding (34 models)</b>	<i>Sn (%)</i>	61.8	60.6	26.9	72.3	73.3
	<i>Sp (%)</i>	91.8	98.3	100	84.9	85.9
	<i>NER (%)</i>	76.8	79.5	63.5	78.6	79.6
	<i>Cvg (%)</i>	99.3	80.6	37.5	100	96.1
	<i>rank</i>	4	8	39	1	7
<b>Antagonism (22 models)</b>	<i>Sn (%)</i>	61.5	60	39	71	73.5
	<i>Sp (%)</i>	87.3	93.8	99.2	81.2	82.9
	<i>NER (%)</i>	74.4	76.9	69.1	76.1	78.2
	<i>Cvg (%)</i>	98.9	80.1	42.4	100	92.9
	<i>rank</i>	3	4	25	1	2
<b>Agonism (21 models)</b>	<i>Sn (%)</i>	73.8	76.1	64.8	74.4	75.8
	<i>Sp (%)</i>	97.5	99	99.9	95.1	95.9
	<i>NER (%)</i>	85.7	87.5	82.3	84.7	85.9
	<i>Cvg (%)</i>	99.7	91.5	51.4	100	97.7
	<i>rank</i>	2	6	17	3	4

by extended connectivity fingerprints (ECFPs). We then performed a multi-dimensional scaling (MDS) on the computed ECFPs to visualize the Jaccard-Tanimoto similarity coefficients in a bidimensional plot.

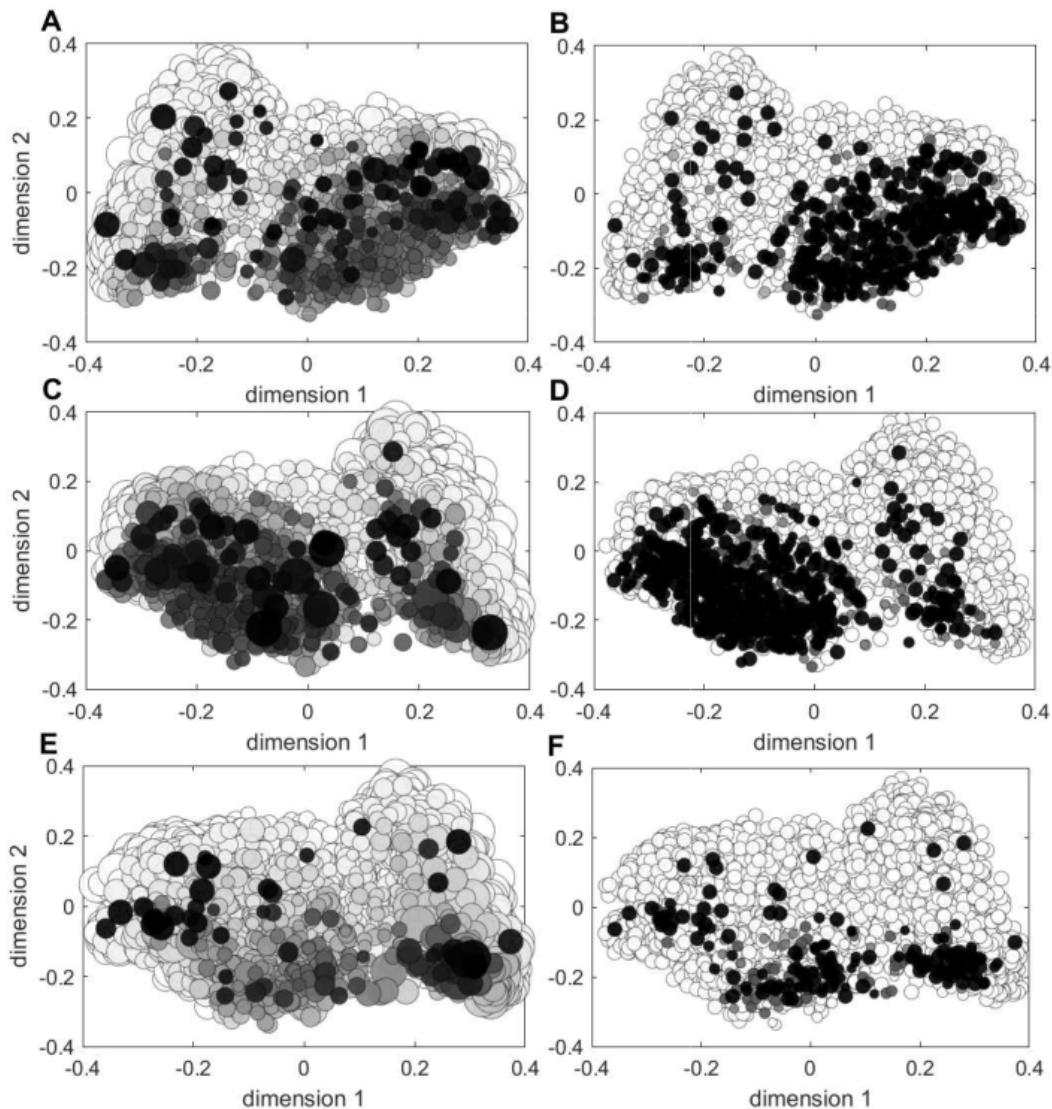


FIGURE 4.3: Plot of the first and second dimensions of the MDS (ECFPs, Jaccard-Tanimoto similarity). Each point represents a chemical, colored based on the number of misclassifications of individual QSAR models for binding, antagonist and agonist endpoint (A, C and E, respectively) and consensus strategies (B, D and F, respectively); the darker the point, the higher the number of misclassifications. The size of each point is proportional to the percentage of models or consensus strategies that provided a prediction for the chemical.

This allowed us to analyze the relationship between such a structural representation and the number of models (individual or consensus), providing reliable predictions.

In the obtained MDS representation for the binding endpoint (Figure 4.3A-B), chemicals are arranged in two clusters. The cluster characterized by negative scores on the first dimension for the binding endpoint (Figure 4.3A,B) is mainly composed of aliphatic molecules with long alkyl chains, as well as cyclic aliphatic compounds, mostly with sp<sup>3</sup>-hybridized carbon atoms. In this cluster, the most frequent functional groups are carbonyls, hydroxyls, ethers, and esters, while conjugated structures or p-systems are almost absent.

The second cluster, located in the positive score region on the first dimension, is mainly composed of conjugated structures, primarily aromatic rings with many electron acceptor substituents (e.g., -NO<sub>2</sub>, -PO<sub>3</sub>, -SO<sub>3</sub>, -F, -Cl, and -CO) and a few donating groups (e.g., -NH<sub>2</sub> and -OH). Most of the misclassified molecules cluster in specific regions of the chemical space. Aliphatic chemicals (characterized by negative scores on the first dimension for the binding endpoint) are in general well-predicted; on the other hand, misclassifications seem to be mainly grouped in the aromatic cluster (positive scores on the first dimension). Besides incorrect predictions, this region is also associated with lower coverage of the individual models (Figure 4.3A). Similarly, the intermediate region between the two clusters is characterized by low coverage, reflecting regions of model uncertainty.

Similar distributions were obtained for agonism and antagonism data sets (Figure 4.3C-F).

These observations seem to confirm the existence of a relationship between chemical features (as encoded in ECFPs) and model performance, since misclassifications are primarily found in limited portions of chemical space, where molecules are often outside the domain of model applicability.

The following chemicals were misclassified by all individual QSAR models despite falling in their domain of applicability: 19 molecules for binding (all false negatives), 28 for agonism (25 false negatives and 3 false positives), and 37 for antagonism (25 false negatives and 12 false positives). We identified some recurrent problems that could explain the observed misclassifications:

- **Borderline Compounds.** Several active molecules that were consistently predicted as inactive are labeled as having experimental weak or very weak potency (Table 4.2), as quantified by the half-maximal activity ( $AC_{50}$ , the molar concentration that produces 50% of the maximum possible activity). The molecules were thus labeled as active, but they actually are borderline between activity and inactivity. Additionally, different activity values due to differences among experimental protocols have been already reported on this set of chemicals (Mansouri et al., 2020). In such cases, models and experimental data can be regarded as belonging to the same level of assessment (Vighi et al., 2019) and QSAR models might provide an indication of the potential inactivity of these consistently misclassified compounds.

- Differences between Charged and Neutralized Forms. Another reason could be related to the different activities of charged compounds toward their neutralized counterparts. In fact, traditional QSAR pipelines do not consider annotated counterions and rely on the neutralized form for descriptor calculations. Nine false negatives (two, one, and six for binding, antagonism, and agonism sets, respectively) showed a different activity in their neutralized form and with an annotated counterion (Table 4.2). For example, 1-butyl-4-methylpyridinium hexafluorophosphate (DTXSID4049296, CASRN 401788-99-6) is a moderate antagonist ( $AC_{50} = 1.94\mu M$ ), but its neutralized form is identical to the neutralized forms of 1-butyl-4-methylpyridinium bromide (DTXSID2049345, CASRN65350-59-6) and 1-butyl-4-methylpyridinium trifluoro methanesulfonate (DTXSID5049368, CASRN 882172-79-4), which are inactive. This highlights the need for considering the effect of charge and counterions on the final biological activity.

Although consensus methods reduced the uncertainty (Figure 4.3B-D-F), misclassifications and unclassified chemicals are still mainly located in the critical regions (e.g. positive scores of the MDS space for binding endpoint), thus following the same pattern as individual models. This corroborates that consensus approaches can reduce uncertainty but cannot eliminate it altogether, as the integration of incorrect information still leads to poor predictions. The performance of consensus models could improve by considering the structural characteristics of chemicals and the performance of individual models in chemical space.

TABLE 4.2: Summary of the molecules which were considered outside the applicability domain or misclassified by all the individual QSAR models. FPs and FNs stand for false positive and false negatives, respectively. Number of FN molecules with concentration of half maximal activity ( $AC_{50}$ ) above  $20\mu M$  (borderline compounds) and with at least a correspondent inactive neutralized form are listed.

	Borderline compounds	FNs	FPs		
		Differences between charged and neutralized forms	others	Total	Total
<b>Binding</b>	12	2	5	19	0
<b>Antagonism</b>	14	1	10	25	12
<b>Agonism</b>	3	6	16	25	3

### 4.2.1 Consensus Based on Subsets of Models

When integrating multiple sources of information, one might decide to select only the most reliable ones with the goal of neglecting misleading information and potentially improving prediction performance. To this end, we

studied the performance of consensus strategies as a function of the number of individual QSAR models considered, sorted by decreasing predictive performance.

For each endpoint, subsets of models were selected as inputs for consensus approaches using the following strategy: (i) individual QSAR models were ranked according to their NER; (ii) consensus approaches were calculated by iteratively adding one model at a time, starting from an initial subset that included the best five (Figure 4.4).

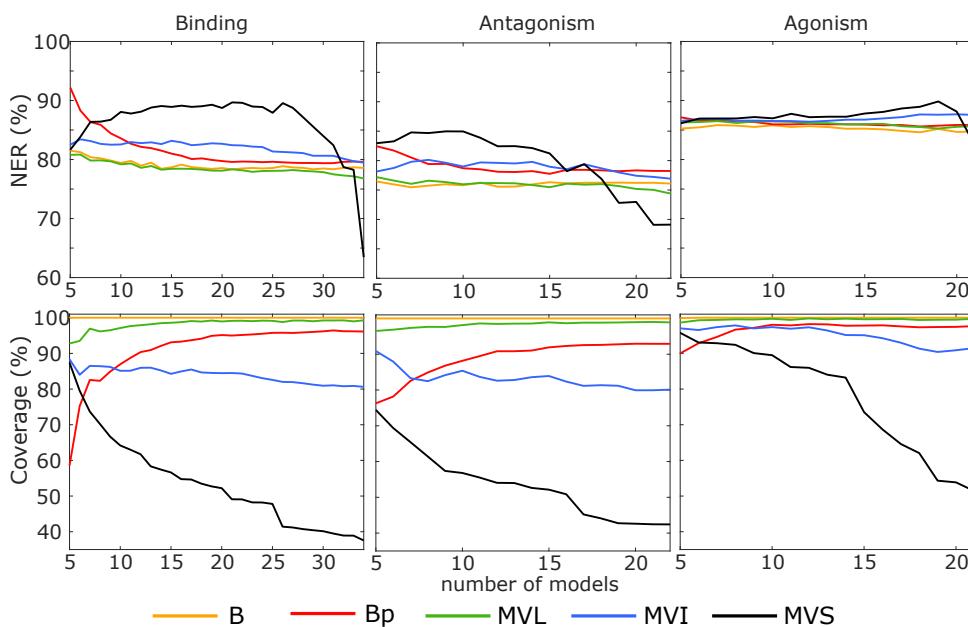


FIGURE 4.4: Plot of NER and coverage as a function of the number of models included in the consensus calculation. B, Bayes; Bp, protective Bayes; MVL, majority voting loose; MVI, majority voting intermediate; MVS, majority voting strict.

The NERs of B, MVI, and MVL are only slightly influenced by the number of included models. It can be concluded that these methods are not sensitive to the integration of poor sources of information in the consensus process. On the contrary, the protective Bayes approach (Bp) is characterized by better performances when a few good models are included, at the expense of the coverage, which shows a considerable decrease. Therefore, when the maximization of the prediction reliability is the only priority, only the most reliable sources of information shall be used in the consensus.

When the final goal is to screen a large set of chemicals for testing prioritization the inclusion of all of the available sources of information can considerably enhance the coverage without a significant loss of performance.

MVS is the consensus approach showing the highest dependence on the number of included models; in particular, as soon as spurious information sources enter in the consensus process, the coverage significantly decreases. Table 4.3 collects the classification performance of consensus approaches calculated on the top five models (chosen based on NER), which is on average

TABLE 4.3: Classification Performance of the Consensus Approaches Estimated on the Binding, Antagonism, and Agonism Sets Considering the Best Five Models Only (Selected Based on NER). For each consensus approach, sensitivity (Sn), specificity (Sp), non-error rate (NER), coverage (Cvg), and total ranking are reported. B, Bayes; Bp, protective Bayes; MVL, majority voting loose; MVI, majority voting intermediate; MVS, majority voting strict.

Endpoint	Performance	consensus approach				
		MVL	MVI	MVS	B	Bp
<b>Binding (5 models)</b>	<i>Sn (%)</i>	63.9	65.7	63.8	72	88.3
	<i>Sp (%)</i>	97.7	99.3	99.5	91	96.2
	<i>NER (%)</i>	80.8	82.5	81.6	81.5	92.2
	<i>Cvg (%)</i>	92.8	88.4	87.4	100	58.6
	<i>rank</i>	3	4	6	1	7
<b>Antagonism (5 models)</b>	<i>Sn (%)</i>	71.6	71.9	78.3	73.2	79.4
	<i>Sp (%)</i>	82.8	84.4	87.6	79.7	85.5
	<i>NER (%)</i>	77.2	78.1	83	76.5	82.4
	<i>Cvg (%)</i>	96.5	90.9	74.4	100	76.3
	<i>rank</i>	2	3	5	1	4
<b>Agonism (5 models)</b>	<i>Sn (%)</i>	73.8	74.1	73.1	74.4	76.1
	<i>Sp (%)</i>	98.8	99	99.2	96.1	98.2
	<i>NER (%)</i>	86.3	86.5	86.1	85.2	87.1
	<i>Cvg (%)</i>	98.6	97	95.8	99.9	90
	<i>rank</i>	2	4	6	3	7

better than that of individual models, with consensus strategies occupying the first seven ranking positions for all of the three considered case studies.

The protective consensus (Bp, MVI, and MVS) obtained on this reduced pool of models provided higher sensitivities than those based on the integration of all available models (Table 4.1), especially for binding and antagonism. However, protective approaches are always ranked worse than the non-protective counterparts.

Finally, the performance of MVS improves, since it is easier to reach a 100% prediction agreement with a few input models compared to using the whole set. For example, for binding endpoints, the NER increased from 63.5 to 81.6% and the coverage increased from 37.5 to 87.4%, respectively.

### 4.3 Summary of results and concluding remarks

In this chapter, we evaluated the extent to which consensus modeling can outperform individual QSARs, by leveraging a large set of QSAR model predictions on androgen receptor binding, agonism, and antagonism.

The protective and non-protective majority voting and Bayes consensus methods were evaluated for their capability to reduce the prediction uncertainty, increase the classification performance, and overcome limitations of individual QSAR models.

The applied consensus strategies provided a better trade-off between the classification performance and the number of reliably predicted chemicals compared to single QSARs. In fact, consensus methods could correctly weigh in and integrate diverse sources of information, leading to balanced values of sensitivity and specificity, as well as to increased coverage compared to the average of individual QSARs. Only a few models could perform better than consensus in terms of classification indices, but they included a limited percentage of chemicals in their applicability domain.

Protective consensus approaches were found to be suitable to incorporate information of less reliable predictions into the final assessment, thereby providing a slightly better classification performance, at the expense of the coverage. However, consensus strategies were not able to perform well in those critical regions of the chemical space where most of the individual models failed, since the integration of erroneous information leads, by definition, to poor predictions.

Implementation of a structure-driven model selection could help overcome these limitations of consensus approaches.

The performance of consensus strategies was finally evaluated as a function of the number of models included in the integration approach. The difference in terms of the classification performance between non-protective consensus strategies applied to all of the available models and to the subset of the five most reliable ones is on average around 1% of the non-error rate (balanced accuracy). Therefore, the performance of non-protective strategies was not significantly influenced by the presence of poorly predictive individual models, thus again demonstrating the ability of these methods to weigh in and integrate conflicting information. On the contrary, protective approaches benefit from the selection of the most predictive models.

As a general recommendation, it is advisable to choose consensus approaches based on the intended application of the model. For prioritization purposes, where one might want to predict as many compounds as possible, we recommend using non-protective approaches. In this case, since MV and Bayes consensus lead to comparable performance, MV may be the method of choice because of the easier implementation and interpretation of results. When the goal is, on the other hand, to obtain the most accurate estimate possible, at the expense of the chemical space covered, protective methods should be applied on a selected, best performing subset of models.



## Chapter 5

# Multi-task modelling to predict nuclear receptor modulators

Multitask modelling implies the simultaneous learning of several related responses or tasks (Caruana, 1997).

Approaches based on deep neural networks are often associated with this type of modelling, where deep learning refers to machine learning strategies based on neural networks with multiple layers of nonlinear processing (Le-Cun, Bengio, and Hinton, 2015).

Several recent QSAR studies have shown that deep learning approaches often outperform traditional machine learning approaches both in regression and classification. In particular, deep neural networks have proved to be a valuable tool in drug design and virtual screening (Unterthiner et al., 2014; Imrie et al., 2018; Korotcov et al., 2017; Ma et al., 2015; Lenselink et al., 2017; Gini et al., 2019; Xu et al., 2017).

The interest in simultaneously modelling more than one biological properties (referred to ‘tasks’) has been increasing in the QSAR field (Ramsundar et al., 2015; Sosnin et al., 2019; Dahl, Jaityl, and Salakhutdinov, 2014). However, the issue deriving from predicting multiple responses implies the use of advanced algorithms (Wold, Sjöström, and Eriksson, 2001).

The standard approach in machine learning is to learn one task at a time (i.e., single-task modelling), while multi-task learning assumes that training a unique model simultaneously on multiple related tasks allows to process together all the available information, which can help to learn also very difficult tasks (Wenzel, Matter, and Schmidt, 2019). In particular, multi-task deep neural networks allow to obtain information not only from the multiple hidden layers, but also from a shared internal representation deriving from the multiple related tasks (Sosnin et al., 2019; Caruana, 1997).

Despite the increasing use of deep neural networks in several scientific fields, their performance improvements over other methods are not from universal nor substantial (Xu et al., 2017).

On the one hand, some scientists recommend simpler models for specific applications (e.g., estrogen receptor binding and acute toxicity prediction) (Russo et al., 2018; Liu et al., 2018), while on the other hand, some studies have reported a statistically significant gain in performance over classical approaches (albeit smaller in absolute terms) (Mayr et al., 2018; Rodriguez-Perez and Bajorath, 2019).

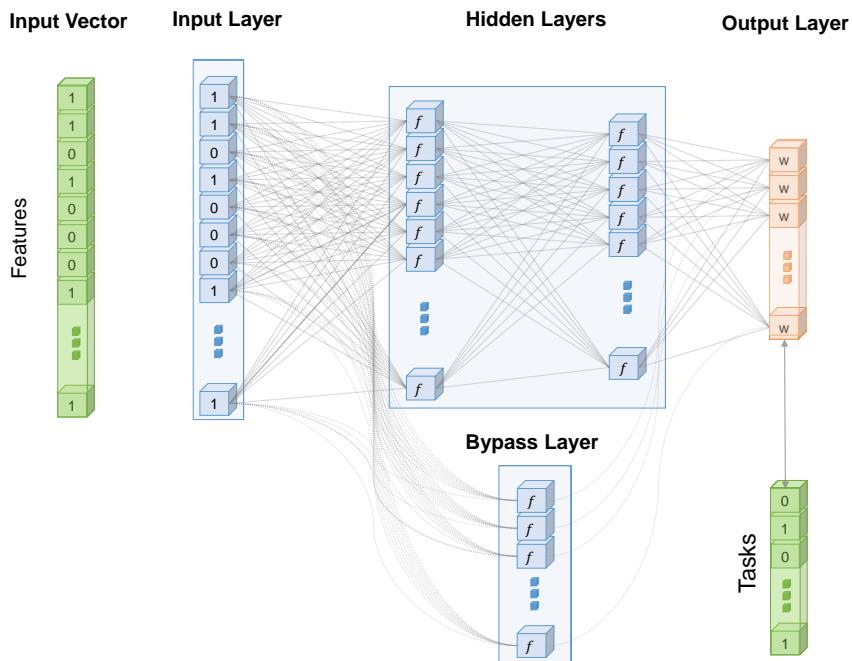


FIGURE 5.1: Schematic representation of the considered multi-task feedforward neural network with bypass layer; input vector is mapped to output layer with repeated compositions of hidden layers, a bypass layer connects directly the input with each task-specific sigmoid neuron in the output layer.

In particular, multitask neural networks have been shown in several QSAR studies to outperform single-task models (Ramsundar et al., 2015; Mayr et al., 2018; Ciallella et al., 2021). The undoubted advantage of the multitask approach, which computes a single model for multiple tasks, is that it is cheaper in terms of computational requirements than traditional single-task QSAR modeling, which involves computing as many models as tasks. Moreover, in multitask modeling, underrepresented tasks can benefit from implicit data augmentation and, thus, achieve higher performance (Caruana, 1997). However, the main drawback is the requirement of a relationship between (Xu et al., 2017) tasks.

In this chapter, we evaluated advantages and limitations of multitask neural networks (both deep and "shallow" neural network-based) in comparison to four single-task reference approaches: random forest (RF)(Breiman, 2001), k-nearest neighbors (kNN),(Wilkinson et al., 1983) N-nearest neighbors (N3)(Todeschini et al., 2015) and Naïve Bayes (NB) (Townsend, Glen, and Mussa, 2012) for predicting bioactivities of NRs.

Comparison was performed on a subset of the NURA dataset on at least one of 30 binary tasks representing agonism, antagonism, or binding (in the form 'active'/'inactive') toward 11 nuclear receptors (Valsecchi et al., 2020b; Valsecchi et al., 2020c). Molecules were randomly divided into training and

test sets. We optimized each model separately in cross-validation through grid search-based protocols, while genetic algorithms were used to tune the parameters of the multitask neural networks (as suggested by the study reported in Appendix A). All approaches were finally evaluated on the test set and on an additional evaluation set of 304 novel chemicals, considering both classification measures on each task and on the entire chemical set.

## 5.1 Data curation

In this work, we used NURA dataset (Chapter 3) and each type of bioactivity for a given receptor (i.e., binding, agonism or antagonism) was considered as a task (e.g., binding activity for androgen receptor), resulting in a total of 33 tasks. Only active and inactive annotations were considered, and tasks containing such annotations for fewer than 200 molecules were discarded (i.e., antagonism for PPAR $\alpha$ , PXR and RXR). The dataset considered thus consists of a total of 14,963 chemicals annotated (as active or inactive) for at least one of the 30 selected tasks (Table 5.1).

Molecules were randomly split into training set (11,970 molecules, 80%) and test set (2,993 molecules, 20%), preserving the proportion between the two classes (actives/inactives) for each task (stratified splitting). The number of molecules for each task and the activity distributions among the tasks are shown in Table 5.1.

For each molecule, we computed ECFPs (Rogers and Hahn, 2010) as input variables.

To further assess the predictivity of the model, we collected an additional set of chemicals, hereafter referred to as the evaluation set. Chemicals were retrieved from the latest available release of ChEMBL database (26, released on 3 March 2020), such that (i) they were not included in the training or test set, (ii) they had an experimental annotation on at least one of the tasks of interest. The retrieved molecules were curated following the same pipeline as the training set and test chemicals, and were labeled for their bioactivity as in the NURA dataset (Chapter 3 and (Valsecchi et al., 2020b)). Because 97.8% of the chemicals were active, the inactivity data (10 molecules) were considered as not numerous enough to have a reliable estimate of classification accuracy and were therefore excluded. The evaluation set consisted of 304 molecules with 435 'active' labels for one task, as reported in the last column of the Table 5.1.

## 5.2 Parameters tuning

The tuning of the neural network parameters was performed by means of genetic algorithm (GA) as suggested by the study reported in Appendix A. The GA resulted in 1515 different parameter combinations (i.e., number of layers and number of neurons per layer, learning rate, optimization algorithm, activation function, regularization and bypass layers). This set contains all the solutions found by the GA approach. The obtained chromosome population

TABLE 5.1: Dataset description: number of molecules and class distributions among the tasks for the training, test and external (ext.) sets.

Task information		Training set		Test set		Ext. set
Receptor	Task Label	# mol.	Act. (%)	# mol.	Act (%)	Act.
Androgen	AR bind	5221	21.3	1328	23.0	5
	AR ago	4861	8.4	1230	8.5	5
	AR ant	4566	13.5	1152	14.0	22
Estrogen ( $\alpha$ )	ER $\alpha$ bind	4927	20.9	1221	21.0	32
	ER $\alpha$ ago	4420	8.4	1116	9.5	7
	ER $\alpha$ ant	4405	6.2	1117	7.8	39
Estrogen ( $\beta$ )	ER $\beta$ bind	5370	17.3	1343	17.3	33
	ER $\beta$ ago	4814	4.7	1216	4.9	24
	ER $\beta$ ant	4291	4.2	1066	4.2	17
Farnesoid X	FXR bind	4627	9.3	1195	9.9	-
	FXR ago	4551	6.4	1170	6.8	40
	FXR ant	3939	2.4	1014	2.8	31
Glucocorticoid	GR bind	5644	25.9	1399	25.3	-
	GR ago	4879	12.0	1242	12.2	2
	GR ant	4173	12.4	1061	13.1	44
PPAR( $\alpha$ )	PPAR $\alpha$ bind	991	98.8	244	98.8	-
	PPAR $\alpha$ ago	808	98.5	204	99.0	-
PPAR( $\gamma$ )	PPAR $\gamma$ bind	5719	23.8	1438	23.4	-
	PPAR $\gamma$ ago	5276	20.7	1299	19.9	6
	PPAR $\gamma$ ant	4261	1.5	1076	2.0	4
PPAR( $\delta$ )	PPAR $\delta$ bind	5165	11.2	1307	11.6	17
	PPAR $\delta$ ago	5005	9.7	1274	10.3	1
	PPAR $\delta$ ant	4463	0.5	1126	0.6	1
Progesterone	PR bind	5029	20.0	1262	19.3	89
	PR ago	4799	6.1	1220	4.8	4
	PR ant	4099	14.3	1042	14.9	-
Pregnane X	PXR bind	3264	5.9	835	5.0	-
	PXR ago	3260	5.7	834	4.9	12
Retinoid X	RXR bind	4352	16.2	1078	14.7	-
	RXR ago	3738	2.8	941	2.8	-

was used to evaluate the influence of the parameters on the classification performance of feedforward neural networks. For this purpose, the 1515 chromosomes were classified based on their fitness function (i.e.,  $NER_T$  in 3-fold cross-validation) and divided into 10 intervals based on the deciles of  $NER_T$ . The relative frequency of each parameter in each decile was calculated (Figure 5.2). A total of 88.5% of the chromosomes included in the highest decile of  $NER_T$  (D10) have learning rate of 0.001, while 5.8% have learning rates of 0.01 and 0.0005, and none have learning rate of 0.025. In addition to the observed optimal learning rate (0.001), other settings are frequent among the best chromosomes, such as (1) Adam optimization (frequency greater than 90% in the best five deciles and equal to 100% in D8, D9 and D10) and (2) no exponential decay (always absent in the models belonging to the best 6 deciles).

To get further insight into the relationship between parameters and classification performance, we performed a principal component analysis (PCA) on the relative frequencies depicted in Figure 5.2. We normalized the values by dividing each relative frequency by the maximum relative frequency of each parameter. Then, we carried out a PCA on the transposed matrix, using the 10 deciles as rows and the 28 parameter values as columns. The first two principal components (Figure 5.3) capture 57% of the data variance, thus providing a good overview on the relationship between network parameters and model performance.

Indeed, the first component (PC1) captures the variation of  $NER_T$  across deciles and, thus, higher PC1 scores correspond to better average classification performance. PC1 confirms the previous considerations from the numerical analysis of the relative frequencies in Figure 5.2: learning rate of 0.001 (LR01), Adam optimization algorithm (OptA) and no exponential decay (ED0) correlate with the highest deciles (D8, D9 and D10), that is, these settings appear frequently in the best models. In contrast, exponential decay (ED1) and gradient descent optimization (OptGD), with the lowest loadings on the first component, are mainly related to the worst deciles (D1 and D2, where  $NER_T$  decreases down to 50%). As for what is captured by PCA, weight decay (WD) and dropout (Drp) seem not to affect  $NER_T$ , having PC1 loadings between -0.1 and 0.1.

Parameter	Value	Relative frequency for $NER_T$ -based decile intervals (%)									
		D1	D2	D3	D4	D5	D6	D7	D8	D9	D10
$NER_T$ (%)	mean	72.1 (50.0-81.8)	83.6 (81.8-85.4)	86.7 (85.4-87.5)	88.1 (87.5-88.7)	88.9 (88.7-89.1)	89.3 (89.1-89.4)	89.5 (89.4-89.6)	89.6 (89.6-89.7)	89.8 (89.7-89.9)	90.1 (89.9-90.6)
min-max											
<b>Architecture</b>	100	Arch1	9.6	2.0	5.9	15.7	7.8	11.8	5.9	21.6	19.6
1000, 100, 10	Arch2	19.2	19.6	11.8	17.6	9.8	3.9	5.9	3.9	7.8	17.3
1000, 100	Arch3	5.8	7.8	7.8	0.0	5.9	11.8	9.8	9.8	13.5	13.5
100, 100	Arch4	21.2	17.6	21.6	15.7	21.6	29.4	13.7	17.6	34.6	34.6
1000	Arch5	7.7	15.7	13.7	21.6	7.8	9.8	15.7	15.7	29.4	7.7
10	Arch6	5.8	9.8	23.5	11.8	15.7	13.7	9.8	13.7	7.8	3.8
10, 5	Arch7	15.4	15.7	9.8	23.5	27.5	11.8	7.8	5.9	5.9	3.8
10, 10	Arch8	15.4	11.8	5.9	7.8	7.8	5.9	9.8	13.7	2.0	1.9
<b>Learning rate</b>	0.01	LR1	11.5	23.5	17.6	39.2	37.3	19.6	13.7	9.8	9.8
0.001	LR01	57.7	25.5	35.3	23.5	45.1	60.8	76.5	86.3	90.2	88.5
0.0005	LR005	19.2	17.6	9.8	5.9	17.6	19.6	9.8	3.9	0.0	5.8
0.025	LR25	11.5	33.3	37.3	31.4	0.0	0.0	0.0	0.0	0.0	0.0
<b>Weight decay</b>	0	WDO	32.7	39.2	49.0	49.0	45.1	35.3	29.4	39.2	37.3
0.01	WD1	67.3	60.8	51.0	51.0	54.9	64.7	70.6	60.8	62.7	50.0
<b>Weight decay type*</b>	L2	48.1	47.1	60.8	70.6	58.8	64.7	58.8	51.0	66.7	67.3
L1	L1	51.9	52.9	39.2	29.4	41.2	35.3	41.2	49.0	33.3	32.7
<b>Dropout</b>	0	Drp0	55.8	54.9	74.5	72.5	64.7	51.0	56.9	60.8	64.7
0.5	Drp1	44.2	45.1	25.5	27.5	35.3	49.0	43.1	39.2	35.3	11.5
<b>Activation function</b>	ReLU	AFr	26.9	41.2	41.2	25.5	31.4	56.9	31.4	47.1	41.2
Leaky ReLU	AFI	15.4	9.8	13.7	35.3	11.8	5.9	25.5	5.9	17.6	88.5
Tanh	AFl	25.0	39.2	31.4	19.6	25.5	29.4	37.3	43.1	39.2	30.8
Sigmoid	AFs	32.7	9.8	13.7	19.6	31.4	7.8	5.9	3.9	2.0	0.0
<b>Bypass layer</b>	1	Bp1	48.1	51.0	47.1	56.9	47.1	51.0	72.5	70.6	43.1
0	Bp0	51.9	49.0	52.9	43.1	52.9	49.0	27.5	29.4	56.9	38.5
<b>Optimization algorithm</b>	Adam	optA	46.2	66.7	49.0	78.4	92.2	100.0	96.1	100.0	100.0
Gradient Descent	optGD	53.8	33.3	51.0	21.6	7.8	0.0	3.9	0.0	0.0	0.0
Exponential decay	ED0	25.0	49.0	96.1	100.0	100.0	100.0	100.0	100.0	100.0	100.0
0	ED1	75.0	51.0	3.9	7.8	0.0	0.0	0.0	0.0	0.0	0.0

\* Relative frequency calculated considering only chromosomes with WD1

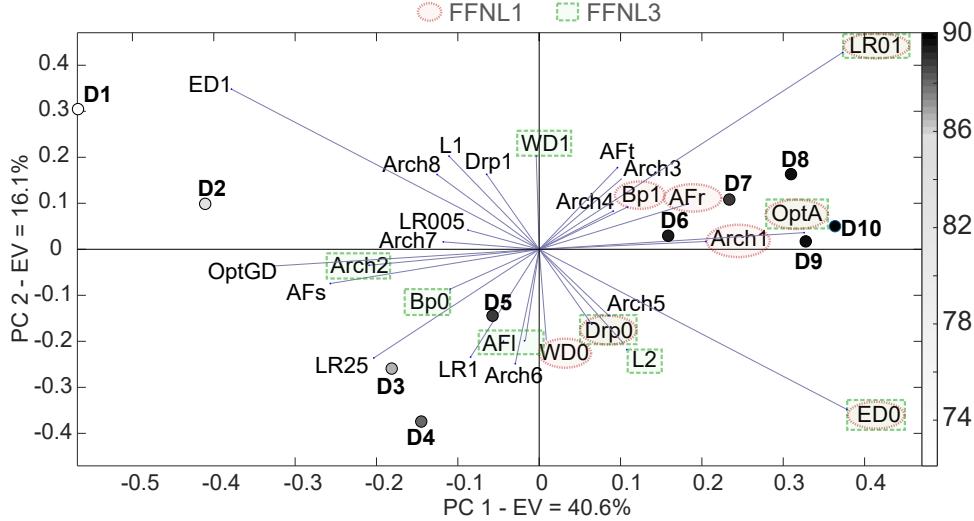


FIGURE 5.3: PCA biplot of relative frequencies of network parameters. Scores (the deciles: D1, D2, ..., D10) are coloured according to the average  $NER_T$ . FFNL1 and FFNL3 stand for the two multitask feedforward neural networks whose selected parameter combinations are highlighted in red and green, respectively

The type of activation function appears to have a moderate influence on classification performance, with ReLU (AFr) being the one most associated with high values of  $NER_T$

In order to select the best performing parameter combination, we then selected the settings associated with the highest PC1 loadings for each parameter, i.e., a multitask neural network (FFNL1) consisting of a hidden layer of 100 neurons (Arch1) with a ReLU (AFr) activation function, a bypass network (Bp1), an Adam optimization algorithm (OptA), a learning rate of 0.001 (LR01) and no regularization (no weight decay, WD0, no dropout, Drp0, and no exponential decay, ED0). This solution is associated with a  $NER_T$  in cross-validation of 90.6%.

FFNL1 can be considered as a ‘shallow’ neural network, since it has only one hidden layer. To compare the performance of shallow and deep multitask neural networks, we also considered the best architecture selected by GA with three layers (Arch2, the deepest). We set the three most relevant parameters to their optimal values as determined by PCA (LR01, OptA, ED0), and searched for the best combination of the remaining parameters.

The best result obtained ( $NER_T$  in cross-validation equal to 90.4%) corresponds to three hidden layers (1000, 100, 10), leaky ReLu as the activation function (AFl), no bypass net (Bp0), Adam optimization algorithm (OptA), learning rate of 0.001 (LR01), no dropout (Drp0), weight decay type L2 of 0.001 (WD1) and no exponential decay (ED0). This model will hereafter be referred to as model FFNL3.

TABLE 5.2: Classification performance ( $NER_T$ ) on the test set for each task. The last row collects the average  $NER_T$  of each modelling method. For each task, the best  $NER_T$  is highlighted in bold while the worst  $NER_T$  is underlined.

task	FFNL1	FFNL3	NB	N3	kNN	RF
AR bind	97.4	96.3	89.6	97.7	97.8	97.9
AR ago	95.3	93.2	89.8	93.0	93.5	94.4
AR ant	92.3	92.0	84.2	92.6	92.5	91.7
ER $\alpha$ bind	95.3	95.9	86.8	95.5	95.0	95.1
ER $\alpha$ ago	87.8	87.7	80.9	87.9	86.1	85.7
ER $\alpha$ ant	88.3	89.8	85.9	87.4	85.0	84.3
ER $\beta$ bind	97.6	97.7	87.0	97.8	98.0	97.0
ER $\beta$ ago	94.2	93.3	85.6	95.4	89.8	91.3
ER $\beta$ ant	89.9	89.5	84.7	89.6	88.1	88.5
FXR bind	96.2	95.4	91.9	97.0	95.2	95.4
FXR ago	97.4	98.1	94.3	99.5	97.1	97.3
FXR ant	87.0	84.4	82.9	85.6	78.3	76.7
GR bind	96.6	95.9	91.0	97.9	98.1	97.9
GR ago	97.0	96.6	93.9	98.1	97.9	98.4
GR ant	92.4	93.2	88.2	93.5	94.0	92.8
PPAR $\alpha$ bind	65.8	65.8	64.2	65.8	66.5	66.0
PPAR $\alpha$ ago	74.0	49.0	72.8	74.0	74.8	74.3
PPAR $\delta$ bind	99.3	99.4	96.0	98.7	99.0	99.2
PPAR $\delta$ ago	99.0	99.5	96.1	99.1	99.0	98.7
PPAR $\delta$ ant	74.9	75.0	75.0	75.9	78.6	71.4
PPAR $\gamma$ bind	95.2	94.2	90.8	95.6	96.0	95.5
PPAR $\gamma$ ago	95.1	95.3	92.0	96.9	97.1	95.9
PPAR $\gamma$ ant	62.5	68.8	74.6	77.1	56.1	56.5
PR bind	98.4	98.1	90.6	98.9	99.0	98.6
PR ago	98.9	98.0	93.9	98.3	98.7	98.8
PR ant	94.6	94.4	87.8	94.0	93.3	91.5
PXR bind	85.9	82.9	75.9	77.5	61.7	71.4
PXR ago	85.8	83.8	75.6	78.3	60.7	73.3
RXR bind	94.0	94.4	91.1	95.2	95.2	95.5
RXRago	74.6	78.9	76.8	77.9	76.8	74.8
<b>average</b>	90.1	89.2	85.7	90.4	88.0	88.2

### 5.3 Comparison on individual tasks

Single and multitask optimized models were used to predict the bioactivity of test set compounds. Only test set compounds within the AD (2970 out of 2993, corresponding to 99.2% of the total) were predicted. The table 5.2 collects the classification performance of all the models on the test set for each task (expressed as  $NER_T$ ).

The last row of Table 5.2 collects the average  $NER_T$  achieved by each modelling method. All methods were able, on average, to correctly classify

most of the test chemicals: the average  $NER_T$  is above 85% for all methods, with the lowest average  $NER_T$  being 85.7% (NB), and the highest being 90.4% (N3), 90.1% (FFNL1) and 89.2% (FFNL3). Comparing the  $NER_T$  obtained on the training set and test set (Table 5.2), the average difference is 9% and 10% for FFNL1 and FFNL3, respectively. Considering these slight differences between fitting and prediction performance, the potential presence of overfitting can be ruled out.

## 5.4 Multi-task vs single-task modelling

To provide a graphical representation of the model's performance, we performed a PCA considering the six classification approaches as samples (rows) and their  $NER_T$  for the 30 tasks as variables. To facilitate the comparison, we added two theoretical benchmarks, consisting of the maximum ('B', best) and minimum ('W', worst)  $NER_T$  values achieved on each task, respectively (Figure 5.4A). The first component obtained (PC1) is related to the overall predictive ability of the classification methods, because the artificial points 'B' and 'W' have the lowest and highest PC1 scores, respectively.

Deep (FFNL3) and shallow (FFNL1) multitask neural networks, along with N3, kNN and RF appear clustered and close to the best point ('B'), indicating their tendency to provide good overall classification. Naive Bayes (NB) shows the worst average performance, as its PC1 score is the highest. These results resemble the average performance shown in the last row of Table 5.2.

The second component (PC2) explains the different behaviour of kNN and RF compared the other best performing methods (N3, FFNL1, FFNL3), which mainly depends on the low  $NER_T$  on the six tasks with the lowest negative loadings on PC2 (RXR agonism, ER antagonism, PXR agonism, PXR binding, FXR antagonism and PPAR $\gamma$  antagonism). These tasks have a remarkably low number of active chemicals (lower than 6%, Figure 5.4B) and kNN and RF provide a suboptimal performance, as can be seen from their low sensitivities ( $S_{nt}$  equal to 72% for ER $\alpha$  antagonism and lower than 58% on the other five tasks, Figure 5.4C). In contrast, multitask feedforward models (FFNL1 and FFNL3) provided the highest  $NER_T$  in five out of six cases (RXR agonism, ER antagonism, PXR agonism, PXR binding and FXR antagonism). Notably, the sensitivity values achieved by FFNL1 and FFNL3, together with N3, on the PXR binding (82.5%, 75.0% and 76.2%, respectively) are significantly higher than those of most of other approaches (kNN and RF showed sensitivity lower than 58% for binding). These tasks share a substantial number of active chemicals with other tasks (75%, 75%, 23% and 23% for ER $\alpha$  antagonism, FXR antagonism, PXR binding and agonism, respectively) (Valsecchi et al., 2020b), potentially suggesting the benefit of multitask models, where simultaneous learning can help lesser-represented tasks to be better modelled by exploiting available data from the other tasks.

Considering individual tasks, no approach is clearly better than the others, in fact all methods show similar performance especially for tasks that are

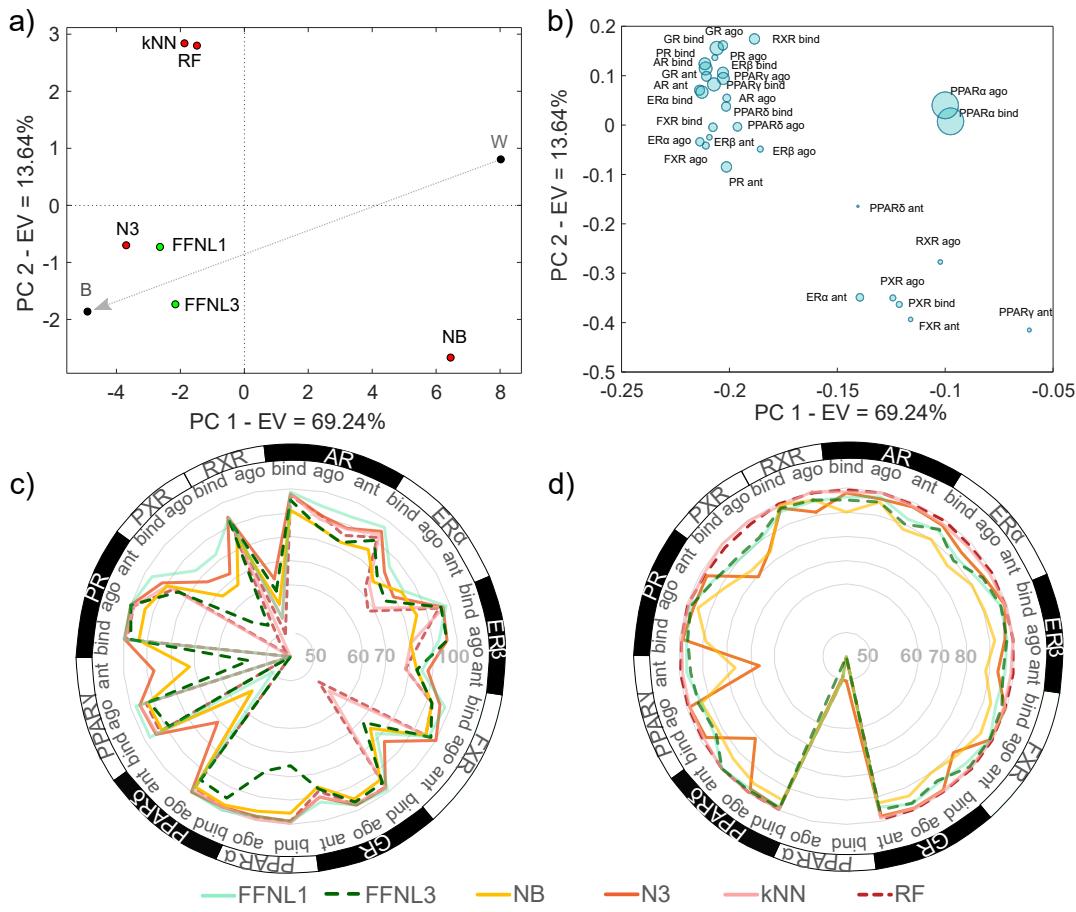


FIGURE 5.4: Analysis of classification performance on individual tasks. (A) PCA score plot on the task-specific classification performance, expressed as  $NER_T$ , of all the models considered; B and W represent the best and worst theoretical performance, respectively; multitask and single-task models are represented by green and red points, respectively. (B) PCA loading plot; each circle represents a task and its size is proportional to the percentage of active chemicals. Radar plots of (C) sensitivity ( $Sn_t$ ) and (D) specificity ( $Sp_t$ ) obtained on test molecules for each task.

easy to model. For example, for binding and agonism on PPAR $\delta$ , all classification approaches provide  $NER_T$  higher than 95%. The same consideration also applies to discrimination of active or inactive chemicals; for example, all approaches correctly classify more than 91% of active chemicals for FXR binding and more than 94% for FXR agonism. In contrast, tasks associated with lower values of  $NER_T$  show greater variation in results depending on the modeling approach considered. For example, N3 and NB achieved significantly higher sensitivity values (86.4% and 72.7% respectively, Figure 5.4C) and higher overall classification performances for PPAR $\gamma$  antagonism ( $NER_T$  equal to 77.1% and 74.6%, respectively) than the other methods ( $NER_T$  equal or lower than 68.8%). One possible explanation for the poor performance of

all models in classifying inactive chemicals for PPAR $\alpha$  agonism ( $S_{pt}$  less than 50% for all methods, Figure 5.4D) may be due to the fact that this is the only task, along with PPAR $\alpha$  binding, with more than 98% active molecules. For all other tasks, methods tend to classify inactive chemicals better, with specificity (Figure 5.4D) comparable to or higher than sensitivity (Figure 5.4C), as expected due to the generally higher number of inactive compounds.

For at least three tasks (PPAR $\delta$ , PPAR $\gamma$  and FXR antagonism), N3 shows higher sensitivity at the expense of specificity, which decreases slightly compared to other models. This is apparent, for example, when modelling PPAR $\delta$  antagonism, for which only N3 provides an acceptable sensitivity value (71.4%). The tendency of N3 to favour the less numerous classes has been already observed (Grisoni, Consonni, and Ballabio, 2019), due to the algorithm normalization on the number of neighbors used belonging to a given class for the calculation of the prediction (Todeschini et al., 2015).

Since all models show similar mean performance (with a difference of about 5% between the highest and lowest average  $NER_T$ , see the last row of Table 5.2), we tested the statistical significance of the observed differences with a Wilcoxon signed-rank test by considering all possible pairs of models and considering  $Sn_t$ ,  $S_{pt}$  and  $NER_T$  separately. The test returned a decision (and an associated  $p$ -value) for the null hypothesis that the median difference in rank between models' performance on all tasks is zero, that is, for each pair of models and for each classification measure, we tested whether there were no significant differences ( $p$ -value > 0.05). Figure 5.5 shows the results of the Wilcoxon signed-rank test, expressed as  $p$  values. Considering the models forming a cluster in the PCA score plot (Figure 5.4A), we can conclude that no statistically significant differences were found between N3 and FFNL1, FFNL1 and FFNL3, N3 and FFNL3. Comparing FFNL1 and FFNL3 with kNN, no statistically significant difference was found on the  $NER_T$  values, but significant differences were found in the specificity and sensitivity results, thus indicating a different behavior of these methods in predicting active or inactive compounds.

Finally, both deep and shallow networks demonstrated to have significantly better classification performance than NB in terms of  $NER_T$ , sensitivity and specificity.

#### 5.4.1 Comparison on global performance

The multitask and single-task approaches were also compared based on the overall classification performance on the test set, in terms of  $Sn_T$ ,  $S_{pt}$  and  $NER_T$  (Table 5.3). Unlike the task-specific indices, the global measures give an idea of the overall classification capability of the model, regardless of the performance on each individual task. In fact, these metrics represent the percentage of active ( $Sn_T$ ) and inactive ( $S_{pt}$ ) chemicals correctly predicted over the entire dataset. Thus, the overall performance is inherently affected by the cardinality of the task (i.e., the number of molecules with annotation in a given task). The higher the number of molecules annotated as active or inactive for a given task, the greater the influence of the task on the overall

p-value	FFNL1				
<i>Snt</i>	<b>0.391</b>				
<b>FFNL3</b>	<b>0.891</b>	<b>FFNL3</b>			
<i>Spt</i>					
<i>NERt</i>	<b>0.222</b>				
<i>Snt</i>	0.000	0.000			
<b>NB</b>	<b>0.000</b>	0.000	<b>NB</b>		
<i>Spt</i>					
<i>NERt</i>	0.000	0.000			
<i>Snt</i>	<b>0.367</b>	<b>0.923</b>	0.000		
<b>N3</b>	<b>0.600</b>	<b>0.417</b>	0.003	<b>N3</b>	
<i>Spt</i>					
<i>NERt</i>	<b>0.214</b>	<b>0.057</b>	0.000		
<i>Snt</i>	0.000	0.000	<b>0.459</b>	0.000	
<b>kNN</b>	<b>0.000</b>	0.000	0.000	0.000	<b>kNN</b>
<i>Spt</i>					
<i>NERt</i>	<b>0.428</b>	<b>0.558</b>	0.008	<b>0.082</b>	
<i>Snt</i>	0.000	0.000	<b>0.909</b>	0.000	<b>0.052</b>
<b>RF</b>	<b>0.000</b>	0.000	0.000	0.000	0.016
<i>Spt</i>					
<i>NERt</i>	0.030	<b>0.098</b>	0.001	0.001	<b>0.229</b>

FIGURE 5.5: P-values of the paired Wilcoxon signed-rank test performed on each pair of classification approaches and for each performance measure ( $S_{nt}$ ,  $S_{pt}$ , and  $NER_T$ ). P-values greater than 0.05, which support the evidence that the null hypothesis is true (i.e., no statistically significant median difference between model performance on all the tasks) are highlighted in bold. Model pairs with P-values consistently greater than 0.05 are highlighted with a gray background.

calculated metrics. FFNL1 and kNN provided the highest classification ability with  $NER_T$  equal to 95.3%; however, FFNL3, RF and N3 achieved very similar performance (95.2%, 95.2% and 94.2%,).

All approaches were able to discriminate active and inactive molecules well (high sensitivity and specificity). Inactive chemicals were predicted better than active ones, as seen by the specificity values, which were generally higher than the sensitivity values, with the only exception for N3, NB, FFNL1, and FFNL3. For these methods, a similar ability to classify active and inactive molecules was observed. kNN and RF achieved the highest  $S_{pt}$  (around 99%), indicating an excellent ability to correctly predict inactive compounds. In contrast, FFNL1 and FFNL3 achieved the highest sensitivity ( $S_{nt}$  around 95%), indicating a good ability to classify active compounds.

#### 5.4.2 Performance on the evaluation set

Both single-task and multitask approaches were further tested on the evaluation set. All models were retrained on the entire set of training and test

TABLE 5.3: Global classification measures expressed as  $NER_T$ ,  $Sn_T$  and  $Sp_T$  (as defined in Chapter 2 Equations 2.21–2.22) achieved on the test set and global sensitivity  $Sn_T$  on the external evaluation set.

Model	Test set			$Sn_T$	Evaluation set
	$NER_T$	$Sp_T$	$Sn_T$		
<b>FFNL1</b>	95.3	95.3	95.4	93.0	
<b>FFNL3</b>	95.2	95.4	95.1	82.6	
<b>NB</b>	89.7	90.3	89.1	89.1	
<b>N3</b>	94.2	94.2	94.3	81.7	
<b>kNN</b>	95.3	98.9	91.7	68.9	
<b>RF</b>	95.2	99.2	91.2	74.9	

molecules (with previously optimized parameters) and used to predict the 304 active molecules in the external evaluation set. Only four chemicals out of 304 were found to be outside the applicability domain and were excluded from the evaluation. Because the evaluation set contains only active molecules for 21 tasks, with several tasks represented by only a few molecules (9 tasks with less than 10 molecules, Table 5.1), only the overall sensitivity  $Sn_T$  was considered as a measure for performance comparison.

When looking at the predictions on these compounds (Table 5.3), RF and kNN underperformed the other models, having  $Sn_T$  ranging from 68.9% (kNN) to 74.9% (RF). FFNL1 has the highest sensitivity ( $Sn_T = 93.0\%$ ). The shallow multitask neural networks (FFNL1), together with Naive Bayes (NB), have comparable  $Sn_T$  values on the test and evaluation sets (Table 5.3), with a difference lower than 3%. On the contrary, several discrepancies can be noticed especially for similarity-based approaches, whose  $Sn_T$  decreased from 91.7% to 68.9% for kNN and from 94.3% to 81.7% for N3. This could be due to the AD approach considered, which is based on the structural similarity of a target molecule to all chemicals in the dataset and does not take into account only for the chemicals with annotated response for an individual task. This AD method was chosen to have a unique approach regardless of tasks and modelling algorithms, in order to improve comparability between single and multitask models. This issue could mainly affect similarity-based approaches and could explain the observed discrepancies in the sensitivities on the test and evaluation set. To visualize the differences between the misclassifications of each model, we performed a non-classical multidimensional scaling (MDS)(Krzanowski, 2000) on the chemicals of the evaluation set (Figure 5.6).

The MDS was calculated on the distance matrix that collects the Jaccard-Tanimoto distances between all possible pairs of 300 molecules, numerically described as ECFPs fingerprints.

Figure 5.6 shows the obtained two-dimensional MDS, where each chemical is coloured according to the fraction of correct predictions among all annotated tasks. White points indicate molecules of the evaluation set associated only with correct predictions, while black points represent chemicals with all wrong predictions among available tasks.

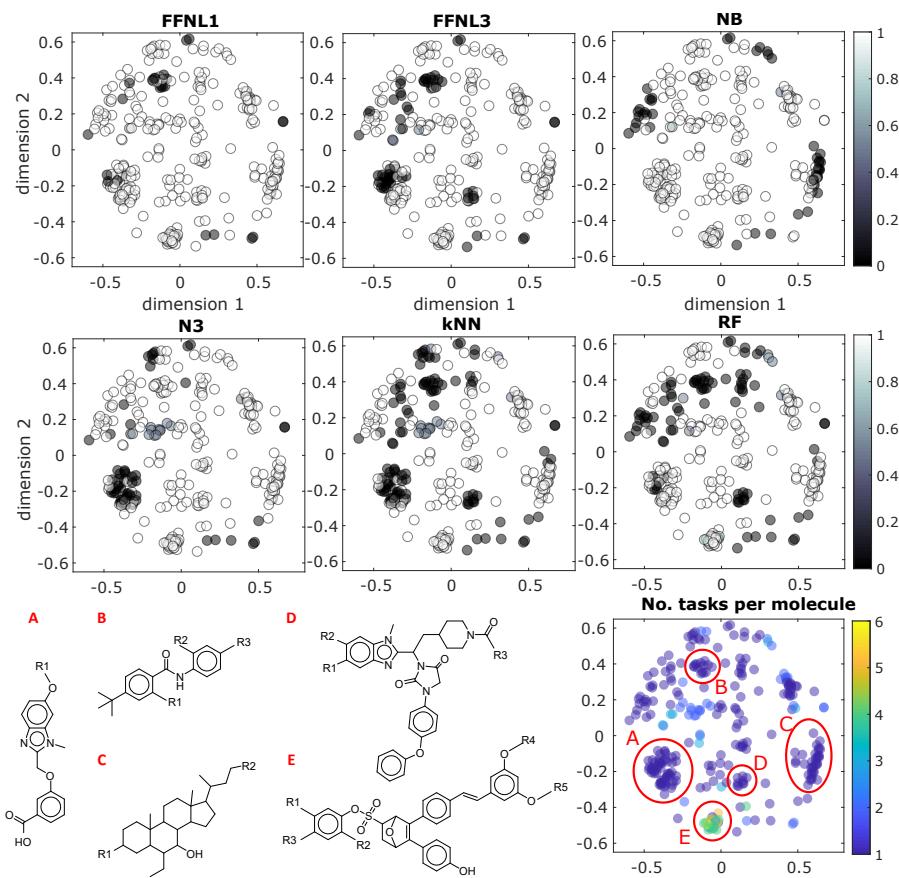


FIGURE 5.6: Multidimensional scaling calculated over the fingerprints (ECFPs) of the evaluation set (stress = 0.30). Each circle represents a molecule and its brightness is proportional to the fraction of correct predictions among all annotated tasks. The last score plot shows the number of tasks with known activities per molecule in terms of the circle's colour. Representative scaffolds of the chemical structures belonging to cluster (A), (B), (C), (D) and (E) are reported below, where  $R_n$  indicates every possible residue.

From Figure 5.6 we see that the incorrect predictions are clustered in chemical space, thus indicating that each modelling method may fail on specific chemical families (darker regions). For example, Imidazo [4,5-c] pyridine derivatives, which were primarily collected from the same scientific study on PPAR $\gamma$ , 68 are grouped in cluster A (Figure 5.6). Most of these chemicals are wrongly predicted by N3, kNN and FFNL3, while FFNL1 and NB provide much better results. Similar considerations can be drawn for the clusters B and D, for which only a few models (e.g., NB and N3) give satisfactory results.

NB, although it provides good overall performance over the entire chemical space, it is the only model to provide inaccurate predictions for molecules belonging to cluster C. These model-specific limitations could be mitigated by averaging model predictions with the application of consensus strategies.

However, groups of chemicals correctly predicted by all models are visible. For example, molecules in cluster E (Figure 5.6), which includes compounds annotated for up to six tasks (Ning et al., 2018), were correctly predicted by all models. This highlights a certain convergence of the structure-activity relationships captured by all the models analyzed when ECFPs are used to describe the molecules.

## 5.5 Summary of results and concluding remarks

In this chapter, we addressed the comparison of the classification performance of deep and shallow multitask neural networks with that of classical single-task classification approaches from a QSAR perspective.

The comparison was performed on a subset of the NURA dataset, i.e., 14,963 chemicals, annotated with agonism, antagonism, and binding activity for 11 nuclear receptors (i.e., 30 tasks), which were divided into training and test sets. Moreover, an additional evaluation set that included 304 chemicals was collected and used to further evaluate the predictive ability of the models.

All models were optimized and, in particular, we performed tuning of the multitask neural networks by means of an *ad hoc* approach based on genetic algorithms and frequency-based selection. This analysis showed that the type of optimization algorithm, the learning rate and its exponential decay are the network parameters that most affected the overall classification performance.

All approaches achieved good classification performance on both test and external molecules. For the data considered, when comparing classical single-task and advanced multitask networks, the results are comparable in terms of average predictive performance, despite some task-dependent exceptions.

Deep and shallow feedforward neural networks achieved the highest classification performance on average, which, however, was often only slightly better than the other methods and not always significantly better. Based on the results of this study, no method was found to clearly outperform all others. The single-task approaches considered in this work have the advantage of avoiding the optimization of several parameters, so they are less computationally demanding than feedforward neural networks. Therefore, we recommend using traditional single-task QSAR approaches when only a few molecular properties need to be predicted.

However, when many tasks need to be modeled simultaneously (such as the 30 tasks modeled in this study), multitask approaches could offer several advantages, such as (1) leveraging information about related tasks and (2) modeling less represented tasks.

Ideally, based on these considerations, multitask models could be a solution to identify selective compounds on desired and less represented biological targets. In addition, having a unique comprehensive model may facilitate several desirable aspects of machine learning in chemistry, such as

(1) defining the applicability domain and (b) developing a "joint" model interpretation for the problem under analysis, thus allowing for better mechanistic understanding.

## Chapter 6

# Pseudo multi-task modelling of ligand-receptor pairs

In pharmaceutical drug research and development ligand and receptor-based *in silico* methods constitute complementary approaches for hit and lead identification. Most of the common modelling approaches in cheminformatics imply a focus on either the ligand features or the protein ones, i.e. ligand-based or structure-based approaches.

Combining ligand- and protein-based descriptors implies the availability of both sources of information and the implementation of a merging strategy. However a combined approach allows to take advantage of a broader information and to predict the activity also for new similar proteins for which only few experimental data are accessible (Tanrikulu et al., 2009).

Nuclear receptors stand as the perfect case study since they are a superfamily of proteins with similar conformation. Although their ligand-binding domains are highly conserved, nuclear receptors have shown to be promiscuous targets, as demonstrated for instance by the endocrine interference exerted by several man-made compounds (Germain et al., 2006).

In this framework, our object under study is ligand-receptor pair which we classified as active if a binding event takes place, as inactive otherwise. Hence, in this chapter only binding activity is taken into account.

The advantage is to be able to simultaneously model different receptors, thus the problem can be seen as a “pseudo-multitask” problem (Caruana, 1997). We dealt with the binding activity of 11 nuclear receptors simultaneously but being the receptor information encoded in the binding-pocket descriptors, the receptor class is implicit.

The data were retrieved from the NURA dataset (Valsecchi et al., 2020b). We described each ligand-receptor pair by a combination of ligand and receptor descriptors. For ligands we computed either Weighted Holistic Atom Localization and Entity Shape (WHALES) descriptors (Grisoni et al., 2018b) or ECFPs (Rogers and Hahn, 2010). As receptor descriptors, we started from an adapted version of WHALES descriptors to handle the dimension of the protein (pWHALES), and we took into account different strategies to uniform the dimensions as depicted in Figure 2.4.

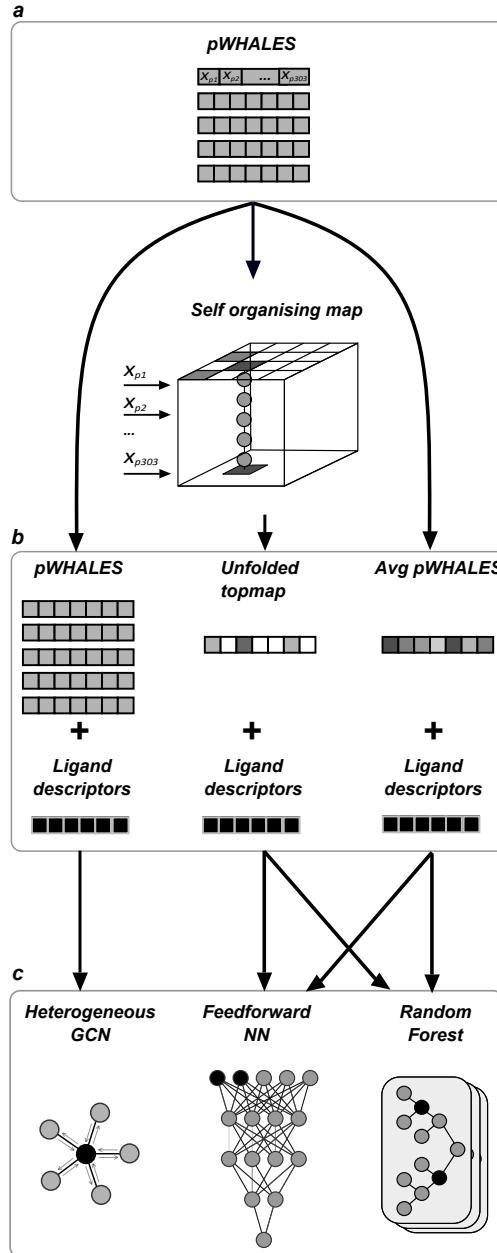


FIGURE 6.1: Workflow of the developed approach: (a) calculation of pWHALES for each structure retrieved for a nuclear receptor, (b) processing pWHALES (three strategies: mean for each pWHALES value, unfolded SOM topmap and unchanged) and merging them with ligand descriptors (either WHALES or ECFPs) and (c) modelling approaches (Random Forest, Feedforward Neural Network and Heterogeneous Graph Convolutional Network).

We applied three modelling algorithms (feedforward neural networks, graph convolutional networks and random forest) using different combinations of descriptors obtaining ten models. After the definition of the applicability domain, we developed a consensus strategy able to successfully predict

binding-event also for nuclear receptors not included in the training set.

We evaluate the performance on three sets. Since the main aim of the work was to determine the generalization capability of the developed approach in predicting binding activity for “unseen” nuclear receptors, we created a customized internal eleven-fold cross-validation by excluding in a step wise manner the data referred to one receptor (i.e. leave one receptor out approach or LORO),

Then, we evaluate the performance on the test set, which contains ligand-receptor pairs for the same 11 nuclear receptors present in the training set.

Finally we evaluated the developed strategy on an external set of ligand-receptor pairs composed by “unseen” nuclear receptors.

Figure 6.1 illustrates the pipeline for the pseudo multi-task modelling approach.

## 6.1 Data

We used again a subset of NURA dataset (Valsecchi et al., 2020b) keeping only binding annotations for a total of 14836 molecules and 11 NRs, i.e. 57387 ligand-receptor pairs (19.5% actives, i.e. binding events).

The selection of the training set followed the criterion of maximizing variability especially for nuclear receptors, i.e., inclusion in the training set of ligands with known activity on multiple receptors and binding to at least one of them was favored. The majority of the ligand-receptor pairs (43725, 76%) represents ligands annotated only for non-binding events.

TABLE 6.1: Summary of the size of the training, test and external set.

Set	No. ligand-receptor pairs	No. ligands	No. receptors
<i>Training</i>	39696	6891	11
<i>Test</i>	17691	7945	11
<i>External</i>	10294	6203	7

In order to evaluate the generalization capability of the workflow to predict binding activity for unseen nuclear receptors, we collected modulation data for additional eight nuclear receptors (LXR $\alpha$ , LXR $\beta$ , ROR $\gamma$ , ROR $\beta$ , RAR $\alpha$ , RAR $\gamma$ , CAR, TR $\beta$ ). We followed the same pipeline explained in Chapter 3 and (Valsecchi et al., 2020b). Each record was assigned a discrete bioactivity label, according to its experimental readout, as follows: ‘active’, for experimental bioactivities equal to or lower than 10,000 nM and ‘inactive’ for entries with activity values exceeding 100,000 nM.

6203 molecules were collected for a total of 10294 ligand-receptor pairs (with a 18.9% of binding events), as reported in Table 6.1.

## 6.2 Ligand-receptor pair

We tested the following three modelling approaches: Random Forest (RF) FeedForward Neural Networks (FFNN) and Graph Convolutional Networks (GCN). RF and FFNN were applied to each combination of ligand-receptor descriptors as input (i.e. mean pWHALES and WHALES, unfolded topmap and WHALES, mean pWHALES and ECFPs and unfolded topmap and ECFPs), while GCN was applied to the unprocessed pWHALES with WHALES or ECFPs, for a total of 10 models as reported in Table 6.2.

Kohonen Maps are self-organising neural networks applied to the unsupervised problems. In Kohonen maps similar input samples are linked to the topological close neurons in the network. Basically, the neurons have as many weights as the number of responses in the target vectors and learn to identify the location in the ANN that is most similar to the input vectors; the weights of the net are updated on the basis of the input sample, i.e. the network is modified each time a sample is introduced and all the samples are introduced for a certain number of times (epochs).

The Kohonen map is usually characterized by being a squared (or hexagonal) toroidal space, that consists of a grid of  $N^2$  neurons, where  $N$  is the number of neurons for each side of the squared space. Each neuron contains as many elements (weights) as the number of input variables. The weights of each neuron are randomly initialised between 0 and 1 and updated on the basis of the input vectors (i.e. samples), for a certain number of times (called training epochs). Both the number of neurons and epochs to be used to train the map must be defined by the user. Kohonen maps can be trained by means of sequential or batch training algorithms. When the sequential training is adopted, in each training step samples are presented to the network, one at a time and weights are updated on the basis of the winner neuron. In each training step, samples are presented to the network, one at a time.

At the end of the network training, samples are placed in the most similar neurons of the Kohonen map; in this way data structure can be visualised and the role of the experimental variables in defining the data structure can be elucidated by looking at the Kohonen weights (Ballabio, Consonni, and Todeschini, 2009) or top map.

We trained a  $50 \times 50$  Kohonen map (or Self Organizing Map - SOM) on the matrix constituted by pWHALES for each crystallographic structures of the selected nuclear receptors (see Table 2.1). Then, we summed up the top maps for each nuclear receptor, obtaining a nuclear receptor specific top map. Finally, we used the unfolded (i.e. vectorized) top map as a descriptors for each receptor.

### 6.2.1 Optimization

Two types of optimization of the model's parameters were taken into account: random search and grid search. The former was performed in order to sift through a wider range of values without costly calculations, while the latter allowed us to search for optimal values in a more limited space.

TABLE 6.2: Summary of the ten developed models characterized by a different combination of modelling approach, ligand descriptors and receptor descriptors and comparison for each model of the overall  $NER_T$  obtained with the applicability domain limitation.

ID	Modelling approach	Ligand descriptors	Receptor descriptors	LORO SET		TEST SET $NER_T$	EXT.SET $NER_T$
				$NER_T$	$NER_T$		
<b>RFwm</b>	Random Forest	WHALES	Mean pWHALES	77.1	89.4	64.1	
<b>RFwt</b>	Random Forest	WHALES	Unfolded topmap	80.4	89.7	86.6	
<b>RFfm</b>	Random Forest	ECFPs	Mean pWHALES	79.5	96.2	57.3	
<b>RFft</b>	Random Forest	ECFPs	Unfolded topmap	69.9	93.6	77.4	
<b>NNwm</b>	Feedforward Neural Network	WHALES	Mean pWHALES	78.6	88.9	77.0	
<b>NNwt</b>	Feedforward Neural Network	WHALES	Unfolded topmap	75.3	82.1	80.4	
<b>NNfm</b>	Feedforward Neural Network	ECFPs	Mean pWHALES	90.5	95.5	83.1	
<b>NNft</b>	Feedforward Neural Network	ECFPs	Unfolded topmap	89.4	96.3	83.0	
<b>GCNw</b>	Graph Convolutional Network	WHALES	Unprocessed pWHALES	80.4	86.8	83.1	
<b>GCNf</b>	Graph Convolutional Network	ECFPs	Unprocessed pWHALES	80.7	93.8	82.0	

Both the optimizations were performed in 3-fold cross validation by means of the scikit-learn library (in particular of the ‘RandomSearchCV’ and ‘GridSearchCV’ classes) (Pedregosa et al., 2011). For GCN and FFNN only a random search of the optimal parameters was performed.

We optimized the overall Non Error Rate ( $NER_T$ ), that is the average between the binding and the non-binding events correctly predicted without distinguishing different nuclear receptors. We also computed an average NER by averaging the  $NER_t$  values computed for each nuclear receptor separately. Since the overall NER is intrinsically influenced by the nuclear receptor’s abundance in the dataset, the average NER provides an indication on the performance giving equal weight to underrepresented receptors (Valsecchi et al., 2020c).

The three modelling approaches (GCN, FFNN and RF) are affected by chance (in the weight initialization or in the random selection of features). Therefore to guarantee the robustness of the performance, each approach was repeated 10 times. As output of each approach the result of a majority voting among replicas was computed. In case of ties (i.e. 5 replicas predict active while the other 5 replicas predict inactive) the ligand-receptor pair will not be predicted.

## 6.2.2 Applicability domain

Reliable model’s predictions are limited generally to the samples that have features similar to the ones used to build that model.

In this case we applied a restrictive approach, in fact to be in AD a new ligand-receptor pair needs to fulfil three rules, i.e. the distance has to be below the threshold for all the considered representations: (i) the WHALES (95th percentile), (ii) ECFPs (95th percentile) and (iii) pWHALES (75th percentile). The thresholds were calculated on training samples.

## 6.3 Pseudo multi-task results

Five receptors (CAR, RAR $\alpha$ , ROR $\gamma$ , PXR, ER $\beta$ ) fall out of the applicability domain and thus are not considered in the further analysis.

The ligand-receptor pairs out of AD are 37.1%, 18.6% and 29.8% for LORO, test and external set respectively.

Despite the intrinsic randomness, all models show robustness, i.e. the standard deviation among the  $NER_T$  (see error bars in Figure 6.3a).

Comparing overall  $NER_T$  and average NER (Figure 6.3a and b) it can be noted that especially for the external set (blue markers), underrepresented receptors contribute to a lowering of average NER compared to overall  $NER_T$ .

The ten models provided  $NER_T$  results ranging from 69.9 to 90.5%, from 82.1 to 96.3% and from 64.1 to 86.6% for LORO, test and external set, respectively. No modelling approach consistently outperformed the others in all the three sets. NNfm, i.e. feedforward neural network with ECFPs and mean pWHALES as descriptors for ligand and protein, respectively, provided the

most consistent results in validation and prediction ( $NER_T$  equal to 90.5%, 95.51%, 83.1% for LORO, test and external set, respectively). However, the models providing the best  $NER_T$  on test set are NNft and RFfm, while RFwt showed the highest  $NER_T$  (86.6%) on the external set.

Since the ten models carry different information, to select the best combination of modelling approaches, all the combinations of the ten approaches were considered starting from a minimum of 3 approaches to a maximum of 10 (i.e. all approaches). The output for each combination was computed using Bayesian protective approach with a threshold ranging between 0.90 and 0.99 (Fernández et al., 2012), for a total of 2190 combinations.

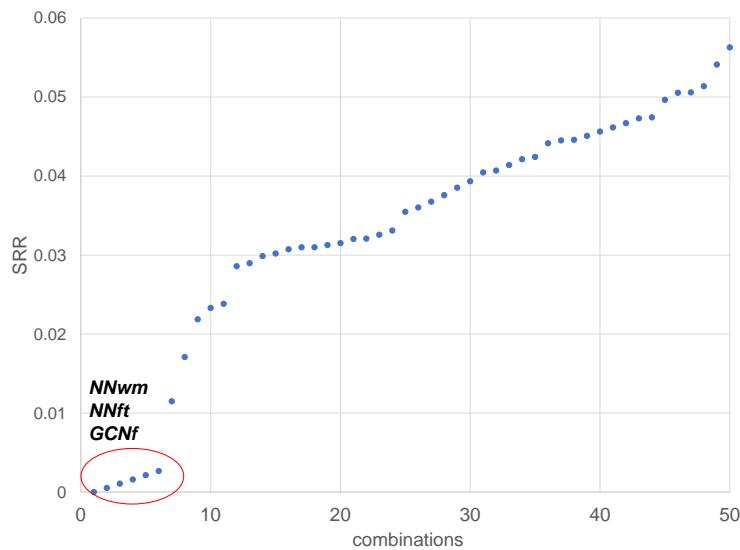


FIGURE 6.2: Normalized sum of reciprocal ranks for the first fifty consensus combinations.

The best combination was chosen according to the lowest normalized sum of reciprocal ranks (SRR) between overall  $NER_T$  and average  $NER$  on the LORO set as shown in Figure 6.2. The six combinations with the lowest SRR are all referred to a three-models consensus including NNwm, NNft and GCNf with different bayesian protective thresholds ranging from 0.93 to 0.98. Hence, the three-models (NNwm, NNft and GCNf) consensus approach with a protective threshold of 0.93 was chosen and its performance are shown in Figure 6.3.

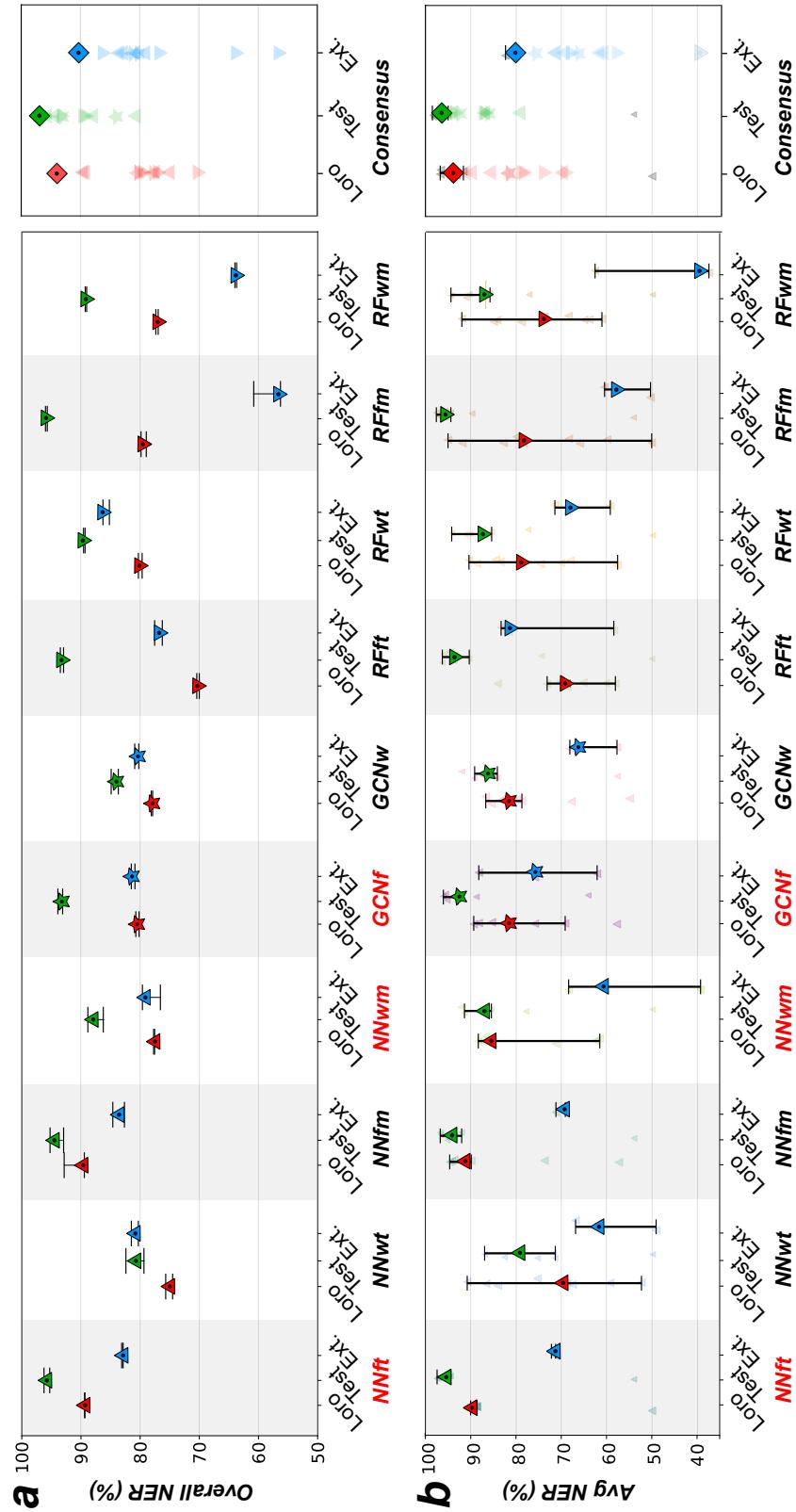


FIGURE 6.3: Performances for single models and consensus approach in terms of overall  $NER_T$  (a) and average  $NER$  (b). Red labels highlight models used for consensus.

The consensus approach improves the overall performance compared to individual models and the overall  $NER_T$  is greater than 90%, while the average  $NER$  is greater than 80% for all the three sets.

## 6.4 Summary of results and concluding remarks

Considering ligand-receptor pairs allowed to exploit all the information available, i.e., ligand and binding pocket descriptors. In addition, with this approach it is possible to predict the binding event for nuclear receptors not considered in the training phase, if they are inside the applicability domain.

The combination of three different modelling approaches in a consensus manner showed to increase both the average  $NER$  and the consistency. In particular, the results obtained especially on the external test set constituted of never seen nuclear receptors are satisfying, with an overall average  $NER$  of 90% and 80%, respectively.

These results are especially promising when applied to virtual screening of large libraries of compounds to find new candidates as selective or promiscuous modulators for the desired nuclear receptor, under the limitations of being within the applicability domain and having crystallographic structures of the receptors from which to calculate descriptors.



# Chapter 7

## Conclusions

Because nuclear receptors are involved in several physiological processes, they are of great interest in areas as diverse as drug development or toxicological evaluation. In this context, computational approaches such as QSAR-based methods offer powerful tools for early detection of new drug candidates or potential harmful molecules.

Analysis of existing nuclear receptor databases revealed purpose-related differences in the type of chemical structures annotated (and chemical scaffolds), and in the proportion of biologically active or inactive molecules. In particular, databases related to computational toxicology show a predominance of inactive compounds, whereas databases with a medicinal chemistry focus contain mainly information on bioactive compounds.

The analysis of the performance of CoMPARA models (Chapter 4) contributed to highlight the need of a curated and heterogeneous set of bioactivity data and the advantage of consensus strategies (Valsecchi et al., 2020a). Therefore, NURA dataset was developed (Chapter 3) as a comprehensive dataset on nuclear receptor bioactivity (Valsecchi et al., 2020b), which collects integrated and curated information on binding, agonism and antagonism for eleven selected nuclear receptors, using well-known chemical databases. In addition, the data curation and aggregation pipeline allowed to bridge the gap between toxicology-related and medicinal-chemistry-related databases.

Our results show that NURA dataset is enriched in terms of number of molecules, structural diversity and covered atomic scaffolds compared to the single sources.

The increased coverage of the chemical and bioactivity space offered by the NURA dataset results in a broader applicability domain and greater robustness of the developed computational models.

Despite being promiscuous targets, the binding domains of nuclear receptors are highly conserved. Therefore, simultaneous modeling of different nuclear receptor bioactivities, i.e., multi-task learning, can lead to increased performance especially for nuclear receptors with few annotated experimental data.

The NURA dataset served as a basis to develop machine learning methods to predict simultaneously the activity for a panel of receptors. To exploit all the information available, we developed multi-task neural networks, which are able to predict together multiple tasks or receptors in our case. The performance of multi-task learning is compared to single task learning

(Valsecchi et al., 2020c) in Chapter 5. Since neural networks require the tuning of several parameters, particular attention was given to model's optimization, carried out by means of genetic algorithms which proved to be the best trade-off between performance and computational requirement and time (see Appendix A).

Compared to single-task models, multi-task neural networks have shown to offer the advantage of exploiting information about related tasks, facilitating the definition of the applicability domain and the development of a "joint" model interpretation for the problem under analysis.

However, the modelling approaches provided satisfying results in predicting both active and inactive molecules, with an average NER provided by the best model for each task equal to 90%. Although human nuclear receptor super family include 48 members, the NURA dataset includes bioactivity data only for the eight most studied nuclear receptors (eleven including ER and PPAR isoforms). In order to provide a modelling approach able to predict the binding event also for nuclear receptors not included in NURA dataset, for which only few experimental data are accessible (Tanrikulu et al., 2009), we combined ligand- and protein-based descriptors in a "pseudo" multi-task approach (Chapter 6) that considers both ligand and binding-pocket descriptors. In this framework, our objects under study are ligand-receptor pairs which we classified as active if a binding event takes place, as inactive otherwise. The advantage is to be able to simultaneously model different receptors. We dealt with the binding activity of 11 nuclear receptors simultaneously but being the receptor information encoded in the binding-pocket descriptors, the receptor class is implicit. We developed different models which we later combined by means of consensus strategies.

In conclusion, the main outcomes of this project include i) the distribution of a curated and freely available dataset on NRs modulators and ii) insights into the application of multi-task neural networks to predict correlated responses in a toxicology and medicinal chemistry context. In particular, the developed approaches can be applied for virtual screening of large libraries of compounds to find new possible modulator candidates of NRs, also in a drug repurposing framework. Candidates can be chosen for their selective modulation for the NR under study. In addition, screening can help the prioritisation of possible endocrine disrupting chemicals.

## Appendix A

### Parameters tuning

Neural networks are increasingly used in chemoinformatics, but they require the tuning of several parameters whose slight fluctuation can strongly affect the result. However, neural network tuning is time consuming and not straightforward. Commonly used optimization techniques include: random search, which is fast, but doesn't guarantee to reach satisfying result (Bergstra and Bengio, 2012), grid search which guarantees to reach the best result but requires a large amount of time and resources (Liashchynskyi and Liashchynskyi, 2019) and genetic algorithm, which can lead to optimal result sparing time and resources.

We compared different strategies to optimize a multi-task neural networks for 3 datasets with special attention on time/computational resources sparing (Valsecchi et al., submitted).

We used the following freely available multi-task datasets whose main characteristics are summarized in Table A.1:

- a subset of the Nuclear Receptor Activity dataset (NURA) (Valsecchi et al., 2020b) composed of a total of 14,963 chemicals annotated (as active or inactive) for at least one of the selected 30 tasks.
- Tox21 dataset which contains qualitative toxicity measurements on 12 biological targets or tasks, including nuclear receptors and stress response pathways. This dataset is curated by MoleculeNet as a benchmark for molecular machine learning (Wu et al., 2018).

The 7831 compounds were pruned for disconnected structures obtaining 7586 molecules.

- ClinTox dataset which contains qualitative data of drugs approved by the FDA and those that have failed clinical trials for toxicity reasons, i.e. 2 tasks. Also this dataset is curated by MoleculeNet (Wu et al., 2018). The 1478 compounds were pruned for disconnected structures obtaining a final dataset of 1472 molecules.

Molecules of each dataset were randomly split into training set (80%) and external test set (20%), trying to preserve the proportion between the two classes (actives/inactives) for each task (stratified splitting).

For each molecule, we computed ECFPs as input variables.

The most used multi-task networks in literature are constituted by fully connected neural network layers trained on multiple tasks, where the output

TABLE A.1: Summary information of the considered multi-task datasets.

Dataset	No. tasks	No. samples	Ref.
<b>NURA</b>	30	14'963	(Valsecchi et al., 2020b)
<b>ClinTox</b>	2	1'472	(Wu et al., 2018)
<b>Tox21</b>	12	7'586	(Wu et al., 2018)

TABLE A.2: Table collecting parameters to be optimized and the levels considered.

Parameters	Levels	Values			
		SGD	Adam	Adamax	RMSProp
<i>Optimization algorithm</i>	4	Sigmoid	eLU	ReLU	Tanh
<i>Activation function</i>	4	L1		L2	
<i>Penalty type</i>	2				
<i>Dropout</i>	2	0		0.5	
<i>Learning Rate</i>	4	0.00001	0.0001	0.001	0.01
<i>Batch size</i>	4	5	10	20	50
<i>Epochs</i>	4	5	50	100	500
<i>No. neurons Layer 1</i>	3	10		100	1000
<i>No. neurons Layer 2</i>	4	0	10	100	1000
<i>No. neurons Layer 3</i>	4	0	10	100	1000

is shared among all learning tasks and then fed into individual classifiers. In this work, we used the binary cross-entropy as loss function; it can handle multiple outputs also in the case of some missing data. We considered both ‘shallow’ (i.e., only one hidden layer) and deep architectures up to three hidden layers, and neurons per layer varying between 0 and 1000. The output layer consists of as many nodes as tasks. The threshold of assignment for the output nodes was optimized on the basis of ROC curves for each task, that is, if the output of the neural network ensemble node is equal or lower than the threshold the compound is predicted inactive, otherwise active. We initialized the network weights randomly according to a truncated normal function with epochs varying from 5 to 500. Table A.2 summarizes the ten parameters as well as their value levels we chose to tune in this work after preliminary experiments. Considering the levels for each parameter, the Cartesian product ( $4 \times 7 \times 3 \times 2 \times 2$ ) gives a total number of possible combinations equal to 196'608, which represent the possible points in our features space which were exhaustively tested by grid search.

TABLE A.3: Results in terms of overall Non-error Rate considering Grid Search (GS) Genetic Algorithm (GA) and Random Search (RS) as optimization strategies in 3-fold cross validation. The computational time for 3-fold cross validation is also reported in hours (h). Mean and confidence interval among 10 replicas are reported for GA and RS results.

	NURA		ClinTox		Tox21	
	Best NERT	time (h)	Best NERT	time (h)	Best NERT	time (h)
GS	95.1	4905.3	93.7	1205.2	76.4	2858.7
GA	<b>94.1 ± 0.2</b>	3.3	<b>91.0 ± 1.1</b>	1.0	<b>74.6 ± 0.5</b>	2.3
RS	<b>93.4 ± 0.4</b>	3.2	<b>88.5 ± 1.4</b>	0.8	<b>73.0 ± 1.0</b>	2.2

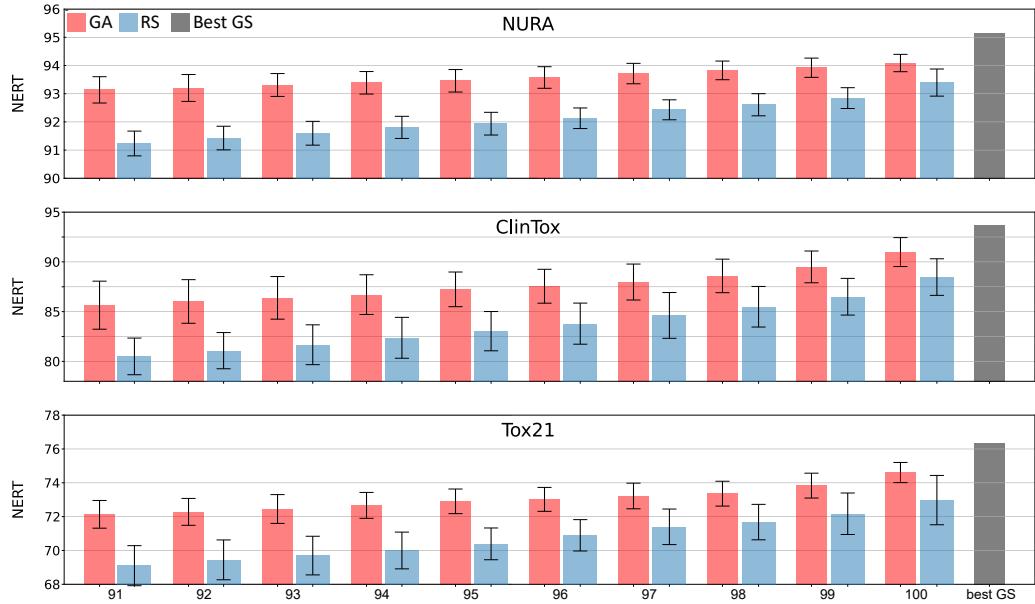


FIGURE A.1: Overall Non-Error Rate of the best 10 solutions for each dataset and optimization method (Genetic Algorithm, GA, Random Search and Grid Search in red, blue and grey, respectively), considering. Error bars are calculated considering 10 replicas.

Performing grid search (GS) is time consuming, unparalleled it takes 1205 hours for the smaller dataset (ClinTox) and 4905 hours for the biggest one (NURA) with processors: 2 x 24-cores Intel Xeon 8160 CPU at 2.10 GHz, Cores: 48 cores/node RAM: 192 GB/node of DDR4. Genetic algorithms (Figure A.1) have shown to be able to converge after a few generations (42) starting from an initial population of 10 chromosomes. Comparing the performance obtained by GS and GA optimization (Figure A.1 and Table A.3) it can be noted that with only 100 total experiments GA is able to converge to near-optimal solutions with a significant reduction of computational times. The difference between the mean  $NER_T$  for GA and the best GS solution is between one (for NURA) to three (for ClinTox) percentage points.



## Appendix B

# Publications and Conferences

Publications:

1. Khan, K., Khan, P. M., Lavado, G., **Valsecchi, C.**, Pasqualini, J., Baderna, D., Marzo, M., Lombardo, A., Roy, K., Benfenati, E. (2019). QSAR modeling of *Daphnia magna* and fish toxicities of biocides using 2D descriptors. *Chemosphere*, 229, 8–17.
2. **Valsecchi, C.**, Grisoni, F., Consonni, V., Ballabio, D. (2019). Structural alerts for the identification of bioaccumulative compounds. *Integrated Environmental Assessment and Management*, 15(1), 19–28.
3. **Valsecchi, C.**, Ballabio, D., Consonni, V., Todeschini, R. (2020). Deep ranking analysis by power eigenvectors (Drape): A polypharmacology case study. *Chemometrics and Intelligent Laboratory Systems*, 203, 104001.
4. **Valsecchi, C.**, Collarile, M., Grisoni, F., Todeschini, R., Ballabio, D., Consonni, V. (2020). Predicting molecular activity on nuclear receptors by multitask neural networks. *Journal of Chemometrics*, e3325.
5. **Valsecchi, C.**, Grisoni, F., Consonni, V., Ballabio, D. (2020). Consensus versus individual qsars in classification: Comparison on a large-scale case study. *Journal of Chemical Information and Modeling*, 60(3), 1215–1223.
6. **Valsecchi, C.**, Grisoni, F., Motta, S., Bonati, L., Ballabio, D. (2020). NURA: A curated dataset of nuclear receptor modulators. *Toxicology and Applied Pharmacology*, 407, 115244.
7. **Valsecchi, C.**, Todeschini, R. (2020). Similarity/diversity indices on incidence matrices containing missing values. *MATCH Communications*, 83(2), 239–260.
8. **Valsecchi, C.**, Todeschini, R. (2021). Deep Ranking Analysis by Power Eigenvectors (Drape): A study on the human, environmental and economic wellbeing of 154 countries. In *Measuring and Understanding complex Phenomena* (pagg. 267–315). Springer.

9. **Todeschini, R.**, Valsecchi, C. (2022). Evaluation of classification performances of Minimum Spanning Trees by 13 different metrics. *MATCH Communications in Mathematical and in Computer Chemistry*, 87, 273-298.
10. **Valsecchi, C.**, Consonni, V., Todeschini, R., Orlandi, M., Gosetti, F., Ballabio, D (submitted). Parsimonious optimization of multitask neural network hyperparameters. *submitted to Molecules, Special Issue "Data and Low-Data Tools for Artificial Intelligence in Medicinal Chemistry*.
11. **Piazza, G.**, Valsecchi, C., Sottocornola, G. (submitted). Deep learning applied to SEM imagery could support the classification of marine coralline algae. *submitted to Diversity, Special Issue "Machine Learning Methods Applied in Diversity Studies*.

Attended conferences and workshop chronologically ordered:

- **Chemometrics Workshop** (Bergamo, February 25-27, 2019). Oral presentation entitled:  
*"Chemoinformatic approach to search for relevant structural alerts using SARpy software"*.
- **X Colloquium Chemometricum Mediterraneum** (Es Castell, June 11-14, 2019). Poster and flash communication entitled:  
*"Similarity/diversity indices on incidence matrices containing missing values"*.
- **QSAR2021** (Online, June 7-9, 2021). Poster presentation entitled:  
*"Predicting molecular activity on nuclear receptors with deep and machine learning"*.
- **XXVII Congresso Nazionale della Società Chimica Italiana** (Online, September 14-23, 2021). Oral presentation entitled:  
*"Enhanced LC-MS/MS spectra matching through multi-task neural networks and molecular fingerprints"*.
- **III Convegno Annuale Centro 3R** (Online, September 30, 2021). Oral presentation entitled:  
*"Nuclear receptor modulators: catching information by machine learning"*.

Attended summer schools:

- Summer School **DeepLearn2019** (Warsaw, July 22-26, 2019).
- Summer School **DeepLearn2021** (Online, July 26-30, 2021).

# Appendix C

## Deliverables

During my PhD I contributed to the realization of the following deliverables:

**NURA dataset** collects curated information on small molecules that modulate nuclear receptors to be intended for both pharmacological and toxicological applications. NURA contains bioactivity annotations for 15,206 molecules and 11 selected NRs as reported in (Valsecchi et al., 2020b). NURA is accessible free of charge at the following URL on the Zenodo website (see Figure C.1).:

[zenodo.org/record/3991562#.YUl53rgzaUk](https://zenodo.org/record/3991562#.YUl53rgzaUk)

**NURA curation pipeline** used to prune and standardize records downloaded from ChEMBL, NR-DBIND, BindingDB and Tox21 as described in (Valsecchi et al., 2020b). It was developed in KNIME and is accessible free of charge at the following URL:

[michem.unimib.it/download/data/nura/](https://michem.unimib.it/download/data/nura/)

**Bayes and majority voting consensus (for MATLAB)** code and data to reproduce the consensus (high level data fusion) described in (Valsecchi et al., 2020a) and in Chapter 4 is freely available at:

[michem.unimib.it/download/data/bayes-and-majority-voting-consensus-for-matlab/](https://michem.unimib.it/download/data/bayes-and-majority-voting-consensus-for-matlab/)

**MST toolbox for MATLAB** is aimed to visualize Minimum Spanning Trees for small sized datasets using 13 different distance metrics. The toolbox provides both node-based and a link-based strategies to semi-supervised classification as described in [to be added]. In addition, the toolbox allows the classification of a new object using MST-based measures and its graphical representation in a MST. It is available at the following URL:

[michem.unimib.it/download/matlab-toolboxes/mst-viewer-for-matlab/](https://michem.unimib.it/download/matlab-toolboxes/mst-viewer-for-matlab/)

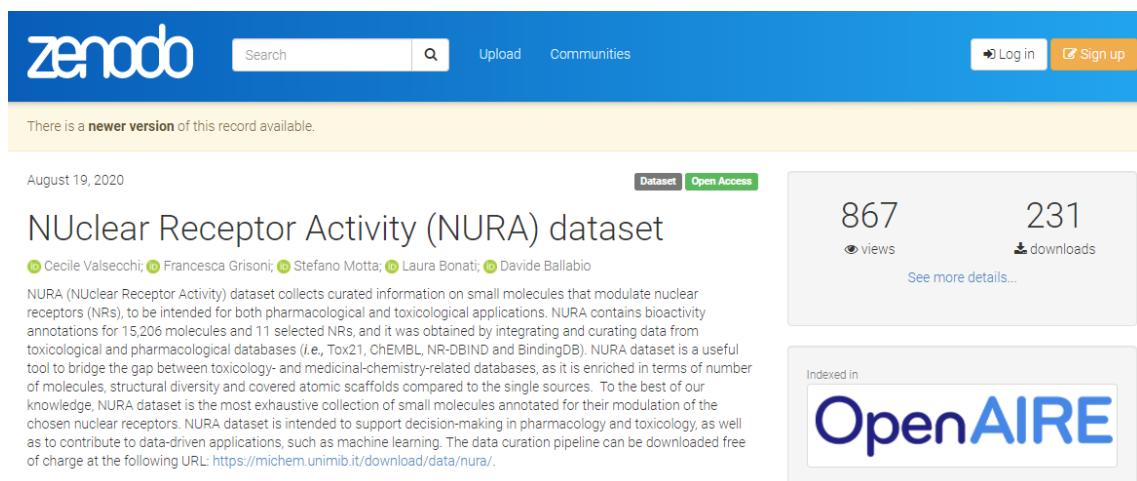


FIGURE C.1: Screenshot of the NURA page on the Zenodo website (accessed on 09/20/2021).

# Bibliography

- Abadi, Martín et al. (2015). *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. Software available from tensorflow.org. URL: <https://www.tensorflow.org/>.
- Abdelaziz, Ahmed et al. (2016). "Consensus modeling for HTS assays using in silico descriptors calculates the best balanced accuracy in Tox21 challenge". In: *Frontiers in Environmental Science* 4, p. 2.
- Agostinelli, Forest et al. (2014). "Learning activation functions to improve deep neural networks". In: *arXiv preprint arXiv:1412.6830*.
- Al Sharif, Merlin et al. (2017). "The application of molecular modelling in the safety assessment of chemicals: A case study on ligand-dependent PPAR $\gamma$  dysregulation". In: *Toxicology* 392, pp. 140–154.
- Allegretti, Marcello et al. (2021). "Repurposing the estrogen receptor modulator raloxifene to treat SARS-CoV-2 infection". In: *Cell Death & Differentiation*, pp. 1–11.
- Ambure, Pravin et al. (2019). "New workflow for QSAR model development from small data sets: small dataset curator and small dataset modeler. integration of data curation, exhaustive double cross-validation, and a set of optimal model selection techniques". In: *Journal of chemical information and modeling* 59.10, pp. 4070–4076.
- An, W Frank and Nicola Tolliday (2010). "Cell-based assays for high-throughput screening". In: *Molecular biotechnology* 45.2, pp. 180–186.
- Ariens, Everhardus Jacobus et al. (1954). "Affinity and intrinsic activity in the theory of competitive inhibition. 1. Problems and theory." In: *Archives internationales de pharmacodynamie et de thérapie* 99, pp. 32–49.
- Asturiol, D, S Casati, and A Worth (2016). "Consensus of classification trees for skin sensitisation hazard prediction". In: *Toxicology in Vitro* 36, pp. 197–209.
- Atlas, The Human Protein (Accessed: 2021-06-10). In: URL: <https://www.proteinatlas.org/>
- Balaguer, Patrick, Vanessa Delfosse, and William Bourguet (2019). "Mechanisms of endocrine disruption through nuclear receptors and related pathways". In: *Current Opinion in Endocrine and Metabolic Research* 7, pp. 1–8.
- Ballabio, D, R Todeschini, and V Consonni (2019). "Recent advances in high-level fusion methods to classify multiple analytical chemical data". In: *Data Handling in Science and Technology* 31, pp. 129–155.
- Ballabio, Davide, Viviana Consonni, and Roberto Todeschini (2009). "The Kohonen and CP-ANN toolbox: a collection of MATLAB modules for self organizing maps and counterpropagation artificial neural networks". In: *Chemometrics and Intelligent Laboratory Systems* 98.2, pp. 115–122.

- Ballabio, Davide, Francesca Grisoni, and Roberto Todeschini (2018). "Multivariate comparison of classification performance measures". In: *Chemosometrics and Intelligent Laboratory Systems* 174, pp. 33–44.
- Ballabio, Davide et al. (2017). "Qualitative consensus of QSAR ready biodegradability predictions". In: *Toxicological & Environmental Chemistry* 99.7-8, pp. 1193–1216.
- Ballabio, Davide et al. (2019). "Integrated QSAR models to predict acute oral systemic toxicity". In: *Molecular informatics* 38.8-9, p. 1800124.
- Banerjee, Monimoy, Delira Robbins, and Taosheng Chen (2015). "Targeting xenobiotic receptors PXR and CAR in human diseases". In: *Drug discovery today* 20.5, pp. 618–628.
- Baurin, Nicolas et al. (2004). "2D QSAR consensus prediction for high-throughput virtual screening. An application to COX-2 inhibition modeling and screening of the NCI database". In: *Journal of chemical information and computer sciences* 44.1, pp. 276–285.
- Begam, A Jameera, S Jubie, and MJ Nanjan (2017). "Estrogen receptor agonists/antagonists in breast cancer therapy: A critical review". In: *Bioorganic chemistry* 71, pp. 257–274.
- Bemis, Guy W and Mark A Murcko (1996). "The properties of known drugs. 1. Molecular frameworks". In: *Journal of medicinal chemistry* 39.15, pp. 2887–2893.
- Berg, J.M., J.L. Tymoczko, and L. Stryer (2002). *Biochemistry, Fifth Edition*. W.H. Freeman. ISBN: 9780716730514. URL: <https://books.google.it/books?id=uDFqAAAAMAAJ>.
- Berger, Joel and David E Moller (2002). "The mechanisms of action of PPARs". In: *Annual review of medicine* 53.1, pp. 409–435.
- Bergstra, James and Yoshua Bengio (2012). "Random search for hyper-parameter optimization." In: *Journal of machine learning research* 13.2.
- Berman, Helen M et al. (2000). "The protein data bank". In: *Nucleic acids research* 28.1, pp. 235–242.
- Berthold, Michael R et al. (2009). "KNIME-the Konstanz information miner: version 2.0 and beyond". In: *AcM SIGKDD explorations Newsletter* 11.1, pp. 26–31.
- Borràs, Eva et al. (2015). "Data fusion methodologies for food and beverage authentication and quality assessment–A review". In: *Analytica Chimica Acta* 891, pp. 1–14.
- Breiman, Leo (Oct. 2001). "Random Forests". In: *Mach. Learn.* 45.1, pp. 5–32. ISSN: 0885-6125. DOI: [10.1023/A:1010933404324](https://doi.org/10.1023/A:1010933404324), URL: <https://doi.org/10.1023/A:1010933404324>.
- Burris, Thomas P et al. (2013). "Nuclear receptors and their selective pharmacologic modulators". In: *Pharmacological reviews* 65.2, pp. 710–778.
- Butler, Keith T et al. (2018). "Machine learning for molecular and materials science". In: *Nature* 559.7715, pp. 547–555.
- Button, Alexander et al. (2019). "Automated de novo molecular design by hybrid machine intelligence and rule-driven chemical synthesis". In: *Nature machine intelligence* 1.7, pp. 307–315.

- Caruana, Rich (1997). "Multitask learning". In: *Machine learning* 28.1, pp. 41–75.
- Chauhan, S and A Kumar (2018). "Consensus QSAR modelling of SIRT1 activators using simplex representation of molecular structure". In: *SAR and QSAR in Environmental Research* 29.4, pp. 277–294.
- Cherkasov, Artem et al. (2014). "QSAR modeling: where have you been? Where are you going to?" In: *Journal of medicinal chemistry* 57.12, pp. 4977–5010.
- Chollet, François et al. (2015). *Keras*. <https://keras.io>.
- Christopoulos, Arthur (2002). "Allosteric binding sites on cell-surface receptors: novel targets for drug discovery". In: *Nature reviews Drug discovery* 1.3, pp. 198–210.
- Ciallella, Heather L et al. (2021). "Predictive modeling of estrogen receptor agonism, antagonism, and binding activities using machine-and deep-learning approaches". In: *Laboratory Investigation* 101.4, pp. 490–502.
- Cronin, Mark TD and T Wayne Schultz (2003). "Pitfalls in QSAR". In: *Journal of Molecular Structure: THEOCHEM* 622.1-2, pp. 39–51.
- Dahl, George E, Navdeep Jaitly, and Ruslan Salakhutdinov (2014). "Multi-task neural networks for QSAR predictions". In: *arXiv preprint arXiv:1406.1231*.
- Davey, Rachel A and Mathis Grossmann (2016). "Androgen receptor structure, function and biology: from bench to bedside". In: *The Clinical Biochemist Reviews* 37.1, p. 3.
- Dhiman, Vineet K, Michael J Bolt, and Kevin P White (2018). "Nuclear receptors in cancer—uncovering new and evolving roles through genomic analysis". In: *Nature Reviews Genetics* 19.3, p. 160.
- Dixon, Emmanuel D et al. (2021). "The Role of Lipid Sensing Nuclear Receptors (PPARs and LXR) and Metabolic Lipases in Obesity, Diabetes and NAFLD". In: *Genes* 12.5, p. 645.
- Duvenaud, David et al. (2015). "Convolutional networks on graphs for learning molecular fingerprints". In: *arXiv preprint arXiv:1509.09292*.
- Edman, Karl et al. (2015). "Ligand binding mechanism in steroid receptors: From conserved plasticity to differential evolutionary constraints". In: *Structure* 23.12, pp. 2280–2290.
- Ekins, Sean et al. (2009). "Challenges predicting ligand-receptor interactions of promiscuous proteins: the nuclear receptor PXR". In: *PLoS Comput Biol* 5.12, e1000594.
- Fawcett, Tom (2006). "An introduction to ROC analysis". In: *Pattern recognition letters* 27.8, pp. 861–874.
- Fernández, Alberto et al. (2012). "Quantitative consensus of bioaccumulation models for integrated testing strategies". In: *Environment international* 45, pp. 51–58.
- Fourches, Denis, Eugene Muratov, and Alexander Tropsha (2010). "Trust, but verify: on the importance of chemical structure curation in cheminformatics and QSAR modeling research". In: *Journal of chemical information and modeling* 50.7, p. 1189.
- Francis, Gordon A et al. (2003). "Nuclear receptors and the control of metabolism". In: *Annual review of physiology* 65.1, pp. 261–311.

- Gasteiger, Johann and Mario Marsili (1980). "Iterative partial equalization of orbital electronegativity—a rapid access to atomic charges". In: *Tetrahedron* 36.22, pp. 3219–3228.
- Gaulton, Anna et al. (2017). "The ChEMBL database in 2017". In: *Nucleic acids research* 45.D1, pp. D945–D954.
- Germain, Pierre et al. (2006). "Overview of nomenclature of nuclear receptors". In: *Pharmacological reviews* 58.4, pp. 685–704.
- Gernert, DL et al. (2003). "Design and synthesis of fluorinated RXR modulators". In: *Bioorganic & medicinal chemistry letters* 13.19, pp. 3191–3195.
- Gilson, Michael K et al. (2016). "BindingDB in 2015: a public database for medicinal chemistry, computational chemistry and systems pharmacology". In: *Nucleic acids research* 44.D1, pp. D1045–D1053.
- Gini, G et al. (2019). "Could deep learning in neural networks improve the QSAR models?" In: *SAR and QSAR in Environmental Research* 30.9, pp. 617–642.
- Griffen, Edward J et al. (2018). "Can we accelerate medicinal chemistry by augmenting the chemist with Big Data and artificial intelligence?" In: *Drug discovery today*.
- Grisoni, Francesca, Viviana Consonni, and Davide Ballabio (2019). "Machine learning consensus to predict the binding to the androgen receptor within the CoMPARA project". In: *Journal of chemical information and modeling* 59.5, pp. 1839–1848.
- Grisoni, Francesca et al. (2015). "QSAR models for bioconcentration: is the increase in the complexity justified by more accurate predictions?" In: *Chemosphere* 127, pp. 171–179.
- Grisoni, Francesca et al. (2018a). "Molecular descriptors for structure–activity applications: a hands-on approach". In: *Computational Toxicology*. Springer, pp. 3–53.
- Grisoni, Francesca et al. (2018b). "Scaffold-Hopping from synthetic drugs by holistic molecular representation". In: *Scientific reports* 8.1, pp. 1–12.
- Grisoni, Francesca et al. (2020). "Bidirectional molecule generation with recurrent neural networks". In: *Journal of chemical information and modeling* 60.3, pp. 1175–1183.
- Gronemeyer, Hinrich, Jan-Åke Gustafsson, and Vincent Laudet (2004). "Principles for modulation of the nuclear receptor superfamily". In: *Nature reviews Drug discovery* 3.11, pp. 950–964.
- Gupta, Anvita et al. (2018). "Generative recurrent networks for de novo drug design". In: *Molecular informatics* 37.1-2, p. 1700111.
- Halevy, Alon, Peter Norvig, and Fernando Pereira (2009). "The unreasonable effectiveness of data". In: *IEEE Intelligent Systems* 24.2, pp. 8–12.
- Halgren, Thomas A (1996). "Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94". In: *Journal of computational chemistry* 17.5-6, pp. 490–519.
- Hall, Julie M and Callie W Greco (2020). "Perturbation of nuclear hormone receptors by endocrine disrupting chemicals: Mechanisms and pathological consequences of exposure". In: *Cells* 9.1, p. 13.

- Hansch, Corwin and Toshio Fujita (1964). "p- $\sigma$ - $\pi$  Analysis. A method for the correlation of biological activity and chemical structure". In: *Journal of the American Chemical Society* 86.8, pp. 1616–1626.
- Hanser, Thierry et al. (2016). "Applicability domain: towards a more formal definition". In: *SAR and QSAR in Environmental Research* 27.11, pp. 865–881.
- Hecht-Nielsen, Robert (1992). "Theory of the backpropagation neural network". In: *Neural networks for perception*. Elsevier, pp. 65–93.
- Heinlein, Cynthia A and Chawnshang Chang (2004). "Androgen receptor in prostate cancer". In: *Endocrine reviews* 25.2, pp. 276–308.
- Heitel, Pascal et al. (2019). "Computer-assisted discovery and structural optimization of a novel retinoid X receptor agonist chemotype". In: *ACS medicinal chemistry letters* 10.2, pp. 203–208.
- Hendriks, Margriet MWB et al. (1992). "Multicriteria decision making". In: *Chemometrics and Intelligent Laboratory Systems* 16.3, pp. 175–191.
- Hewitt, Mark et al. (2007). "Consensus QSAR models: do the benefits outweigh the complexity?" In: *Journal of chemical information and modeling* 47.4, pp. 1460–1468.
- Honma, Naoko, Yoko Matsuda, and Tetuo Mikami (2021). "Carcinogenesis of Triple-Negative Breast Cancer and Sex Steroid Hormones". In: *Cancers* 13.11, p. 2588.
- Huang, Pengxiang, Vikas Chandra, and Fraydoon Rastinejad (2010). "Structural overview of the nuclear receptor superfamily: insights into physiology and therapeutics". In: *Annual review of physiology* 72, pp. 247–272.
- Imrie, Fergus et al. (2018). "Protein family-specific models using deep neural networks and transfer learning improve virtual screening and highlight the need for more data". In: *Journal of chemical information and modeling* 58.11, pp. 2319–2330.
- Ishigami-Yuasa, Mari and Hiroyuki Kagechika (2020). "Chemical Screening of Nuclear Receptor Modulators". In: *International Journal of Molecular Sciences* 21.15, p. 5512.
- Jaworska, Joanna and Sebastian Hoffmann (2010). "Integrated testing Strategy (ItS)–Opportunities to better use existing data and guide future testing in toxicology". In: *ALTEX-Alternatives to animal experimentation* 27.4, pp. 231–242.
- Jensen, Elwood V (1962). "On the mechanism of estrogen action". In: *Perspectives in biology and medicine* 6.1, pp. 47–60.
- Jensen, EV et al. (1968). "A two-step mechanism for the interaction of estradiol with rat uterus." In: *Proceedings of the National Academy of Sciences of the United States of America* 59.2, p. 632.
- Jordan, Michael I and Tom M Mitchell (2015). "Machine learning: Trends, perspectives, and prospects". In: *Science* 349.6245, pp. 255–260.
- Keeney, Ralph L, Howard Raiffa, and Richard F Meyer (1993). *Decisions with multiple objectives: preferences and value trade-offs*. Cambridge university press.
- Keller, James M, Michael R Gray, and James A Givens (1985). "A fuzzy k-nearest neighbor algorithm". In: *IEEE transactions on systems, man, and cybernetics* 4, pp. 580–585.

- Khandelwal, Akash et al. (2008). "Machine learning methods and docking for predicting human pregnane X receptor activation". In: *Chemical research in toxicology* 21.7, pp. 1457–1467.
- Kim, Sunghwan et al. (2019). "PubChem 2019 update: improved access to chemical data". In: *Nucleic acids research* 47.D1, pp. D1102–D1109.
- Kipf, Thomas N and Max Welling (2016). "Semi-Supervised Classification with Graph Convolutional Networks". In: *arXiv preprint arXiv:1609.02907*.
- Kittler, Ralf et al. (2013). "A comprehensive nuclear receptor network for breast cancer cells". In: *Cell reports* 3.2, pp. 538–551.
- Kleinstreuer, Nicole C et al. (2017). "Development and validation of a computational model for androgen receptor activity". In: *Chemical research in toxicology* 30.4, pp. 946–964.
- Korotcov, Alexandru et al. (2017). "Comparison of deep learning with multiple machine learning methods and metrics using diverse drug discovery data sets". In: *Molecular pharmaceutics* 14.12, pp. 4462–4475.
- Krenn, Mario et al. (2019). "SELFIES: a robust representation of semantically constrained graphs with an example application in chemistry". In: *arXiv preprint arXiv:1905.13741*.
- Krzanowski, Wojtek (2000). *Principles of multivariate analysis*. Vol. 23. OUP Oxford.
- Le Maire, Albane, William Bourguet, and Patrick Balaguer (2010). "A structural view of nuclear hormone receptor: endocrine disruptor interactions". In: *Cellular and molecular life sciences* 67.8, pp. 1219–1237.
- LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton (2015). "Deep learning". In: *nature* 521.7553, pp. 436–444.
- Lenselink, Eelke B et al. (2017). "Beyond the hype: deep neural networks outperform established methods using a ChEMBL bioactivity benchmark set". In: *Journal of cheminformatics* 9.1, pp. 1–14.
- Liashchynskyi, Petro and Pavlo Liashchynskyi (2019). "Grid search, random search, genetic algorithm: a big comparison for NAS". In: *arXiv preprint arXiv:1912.06059*.
- Liu, Ruifeng et al. (2018). "Dissecting machine-learning prediction of molecular activity: is an applicability domain needed for quantitative structure–activity relationship models based on deep neural networks?" In: *Journal of chemical information and modeling* 59.1, pp. 117–126.
- Ma, Junshui et al. (2015). "Deep neural nets as a method for quantitative structure–activity relationships". In: *Journal of chemical information and modeling* 55.2, pp. 263–274.
- Maas, Andrew L, Awni Y Hannun, Andrew Y Ng, et al. (2013). "Rectifier nonlinearities improve neural network acoustic models". In: *Proc. icml*. Vol. 30. 1. Citeseer, p. 3.
- Mangelsdorf, David J et al. (1995). "The nuclear receptor superfamily: the second decade". In: *Cell* 83.6, p. 835.
- Mansouri, Kamel et al. (2013). "Quantitative structure–activity relationship models for ready biodegradability of chemicals". In: *Journal of chemical information and modeling* 53.4, pp. 867–878.

- Mansouri, Kamel et al. (2016). "CERAPP: collaborative estrogen receptor activity prediction project". In: *Environmental health perspectives* 124.7, pp. 1023–1033.
- Mansouri, Kamel et al. (2020). "CoMPARA: Collaborative modeling project for androgen receptor activity". In: *Environmental health perspectives* 128.2, p. 027002.
- Marzo, Marco et al. (2016). "Integrating in silico models to enhance predictivity for developmental toxicity". In: *Toxicology* 370, pp. 127–137.
- Mayr, Andreas et al. (2018). "Large-scale comparison of machine learning methods for drug target prediction on ChEMBL". In: *Chemical science* 9.24, pp. 5441–5451.
- Mazaira, Gisela I et al. (2018). "The nuclear receptor field: a historical overview and future challenges". In: *Nuclear receptor research* 5.
- Merk, Daniel et al. (2018a). "Computer-assisted discovery of retinoid X receptor modulating natural products and isofunctional mimetics". In: *Journal of medicinal chemistry* 61.12, pp. 5442–5447.
- Merk, Daniel et al. (2018b). "De novo design of bioactive small molecules by artificial intelligence". In: *Molecular informatics* 37.1-2, p. 1700153.
- Minsky, Marvin L and Seymour A Papert (1988). *Perceptrons: expanded edition*.
- Mitchell, Tom M. (1997). *Machine Learning*. New York: McGraw-Hill.
- Motta, Stefano et al. (2018). "Exploring the PXR ligand binding mechanism with advanced molecular dynamics methods". In: *Scientific reports* 8.1, pp. 1–12.
- Mueller, Stefan O and Kenneth S Korach (2001). "Estrogen receptors and endocrine diseases: lessons from estrogen receptor knockout mice". In: *Current opinion in pharmacology* 1.6, pp. 613–619.
- Mukherjee, Arpan, An Su, and Krishna Rajan (2021). "Deep Learning Model for Identifying Critical Structural Motifs in Potential Endocrine Disruptors". In: *Journal of Chemical Information and Modeling* 61.5, pp. 2187–2197.
- Nair, Vinod and Geoffrey E Hinton (2010). "Rectified linear units improve restricted boltzmann machines". In: *Icml*.
- Netzeva, Tatiana I et al. (2005). "Current status of methods for defining the applicability domain of (quantitative) structure-activity relationships: The report and recommendations of ecvam workshop 52". In: *Alternatives to Laboratory Animals* 33.2, pp. 155–173.
- Neumann, Marc B and Willi Gujer (2008). "Underestimation of uncertainty in statistical regression of environmental models: influence of model structure uncertainty". In: *Environmental science & technology* 42.11, pp. 4037–4043.
- Nielsen, Michael A (2015). *Neural networks and deep learning*. Vol. 25. Determination press San Francisco, CA.
- Ning, Wentao et al. (2018). "Novel hybrid conjugates with dual suppression of estrogenic and inflammatory activities display significantly improved potency against breast cancer". In: *Journal of medicinal chemistry* 61.18, pp. 8155–8173.

- Novac, Natalia and Thorsten Heinzel (2004). "Nuclear receptors: overview and classification". In: *Current Drug Targets-Inflammation & Allergy* 3.4, pp. 335–346.
- Oshida, Keiyu et al. (2015). "Identification of modulators of the nuclear receptor peroxisome proliferator-activated receptor  $\alpha$  (PPAR $\alpha$ ) in a mouse liver gene expression compendium". In: *PloS one* 10.2, e0112655.
- O'Boyle, Noel M (2012). "Towards a Universal SMILES representation-A standard method to generate canonical SMILES based on the InChI". In: *Journal of cheminformatics* 4.1, pp. 1–14.
- Park, So-Jung, Irina Kufareva, and Ruben Abagyan (2010). "Improved docking, screening and selectivity prediction for small molecule nuclear receptor modulators using conformational ensembles". In: *Journal of computer-aided molecular design* 24.5, pp. 459–471.
- Paszke, Adam et al. (2017). "Automatic differentiation in pytorch". In: PDB, RCSB (Accessed: 2021-07-21). In: URL: <https://www.rcsb.org/>.
- Pedregosa, Fabian et al. (2011). "Scikit-learn: Machine learning in Python". In: *the Journal of machine Learning research* 12, pp. 2825–2830.
- Penvose, Ashley et al. (2019). "Comprehensive study of nuclear receptor DNA binding provides a revised framework for understanding receptor specificity". In: *Nature communications* 10.1, pp. 1–15.
- Pradeep, Prachi et al. (2016). "An ensemble model of QSAR tools for regulatory risk assessment". In: *Journal of cheminformatics* 8.1, pp. 1–9.
- Proschak, Ewgenij, Holger Stark, and Daniel Merk (2018). "Polypharmacology by design: a medicinal chemist's perspective on multitargeting compounds". In: *Journal of medicinal chemistry* 62.2, pp. 420–444.
- Pushpakom, Sudeep et al. (2019). "Drug repurposing: progress, challenges and recommendations". In: *Nature reviews Drug discovery* 18.1, pp. 41–58.
- Ramsundar, Bharath et al. (2015). "Massively multitask networks for drug discovery". In: *arXiv preprint arXiv:1502.02072*.
- Reau, Manon et al. (2018). "Nuclear Receptors Database Including Negative Data (NR-DBIND): A Database Dedicated to Nuclear Receptors Binding Data Including Negative Data and Pharmacological Profile: Miniperspective". In: *Journal of medicinal chemistry* 62.6, pp. 2894–2904.
- Rodriguez-Perez, Raquel and Jurgen Bajorath (2019). "Multitask machine learning for classifying highly and weakly potent kinase inhibitors". In: *Acs Omega* 4.2, pp. 4367–4375.
- Rogers, David and Mathew Hahn (2010). "Extended-connectivity fingerprints". In: *Journal of chemical information and modeling* 50.5, pp. 742–754.
- Rokach, Lior and Oded Z Maimon (2007). *Data mining with decision trees: theory and applications*. Vol. 69. World scientific.
- Rotroff, Daniel M et al. (2013). "Using in vitro high throughput screening assays to identify potential endocrine-disrupting chemicals". In: *Environmental health perspectives* 121.1, pp. 7–14.
- Ruder, Sebastian (2017). "An overview of multi-task learning in deep neural networks". In: *arXiv preprint arXiv:1706.05098*.

- Ruiz, P et al. (2017). "Integration of in silico methods and computational systems biology to explore endocrine-disrupting chemical binding with nuclear hormone receptors". In: *Chemosphere* 178, pp. 99–109.
- Rupp, Matthias et al. (2010). "From machine learning to natural product derivatives that selectively activate transcription factor PPAR $\gamma$ ". In: *ChemMedChem: Chemistry Enabling Drug Discovery* 5.2, pp. 191–194.
- Russo, Daniel P et al. (2018). "Comparing multiple machine learning algorithms and metrics for estrogen receptor binding prediction". In: *Molecular pharmaceutics* 15.10, pp. 4361–4370.
- Sadawi, Noureddin et al. (2019). "Multi-task learning with a natural metric for quantitative structure activity relationship learning". In: *Journal of Cheminformatics* 11.1, pp. 1–13.
- Sahigara, Faizan et al. (2013). "Defining a novel k-nearest neighbours approach to assess the applicability domain of a QSAR model for reliable predictions". In: *Journal of cheminformatics* 5.1, pp. 1–9.
- Sailaukhanuly, Yerbolat et al. (2013). "On the ranking of chemicals based on their PBT characteristics: Comparison of different ranking methodologies using selected POPs as an illustrative example". In: *Chemosphere* 90.1, pp. 112–117.
- Santos, Rita et al. (2017). "A comprehensive map of molecular drug targets". In: *Nature reviews Drug discovery* 16.1, pp. 19–34.
- Schug, Thaddeus T et al. (2011). "Endocrine disrupting chemicals and disease susceptibility". In: *The Journal of steroid biochemistry and molecular biology* 127.3–5, pp. 204–215.
- Sebaugh, JL (2011). "Guidelines for accurate EC50/IC50 estimation". In: *Pharmaceutical statistics* 10.2, pp. 128–134.
- Selassie, CD and Rajeshwar P Verma (2003). "History of quantitative structure-activity relationships". In: *Burger's medicinal chemistry and drug discovery* 1, pp. 1–48.
- Shao, Mingyan et al. (2021). "The multi-faceted role of retinoid X receptor in cardiovascular diseases". In: *Biomedicine & Pharmacotherapy* 137, p. 111264.
- Shemetulskis, Norah E et al. (1996). "Stigmata: an algorithm to determine structural commonalities in diverse datasets". In: *Journal of chemical information and computer sciences* 36.4, pp. 862–871.
- Shulman, Andrew I and David J Mangelsdorf (2005). "Retinoid x receptor heterodimers in the metabolic syndrome". In: *New England Journal of Medicine* 353.6, pp. 604–615.
- Sosnin, Sergey et al. (2019). "A survey of multi-task learning methods in chemoinformatics". In: *Molecular informatics* 38.4, p. 1800108.
- Sun, Lixia et al. (2019). "In silico prediction of endocrine disrupting chemicals using single-label and multilabel models". In: *Journal of chemical information and modeling* 59.3, pp. 973–982.
- Tan, Pang-Ning, Michael Steinbach, and Vipin Kumar (2005). *Intro to Data Mining*. Pearson Australia Pty Limited.
- Tanrikulu, Yusuf et al. (2009). "Homology model adjustment and ligand screening with a pseudoreceptor of the human histamine H4 receptor". In: *ChemMedChem: Chemistry Enabling Drug Discovery* 4.5, pp. 820–827.

- Tice, Raymond R et al. (2013). "Improving the human hazard characterization of chemicals: a Tox21 update". In: *Environmental health perspectives* 121.7, pp. 756–765.
- Todeschini, Roberto and Viviana Consonni (2008). *Handbook of molecular descriptors*. Vol. 11. John Wiley & Sons.
- Todeschini, Roberto et al. (2012). "Similarity coefficients for binary chemoinformatics data: overview and extended comparison using simulated and real data sets". In: *Journal of chemical information and modeling* 52.11, pp. 2884–2901.
- Todeschini, Roberto et al. (2015). "N3 and BNN: two new similarity based classification methods in comparison with other classifiers". In: *Journal of chemical information and modeling* 55.11, pp. 2365–2374.
- Toporova, Lucia and Patrick Balaguer (2020). "Nuclear receptors are the major targets of endocrine disrupting chemicals". In: *Molecular and Cellular Endocrinology* 502, p. 110665.
- Townsend, Joe A, Robert C Glen, and Hamse Y Mussa (2012). "Note on naive Bayes based on binary descriptors in Cheminformatics". In: *Journal of chemical information and modeling* 52.10, pp. 2494–2500.
- Tropsha, Alexander (2010). "Best practices for QSAR model development, validation, and exploitation". In: *Molecular informatics* 29.6-7, pp. 476–488.
- Unterthiner, Thomas et al. (2014). "Deep learning as an opportunity in virtual screening". In: *Proceedings of the deep learning workshop at NIPS*. Vol. 27, pp. 1–9.
- Valsecchi, Cecile et al. (2020a). "Consensus versus individual QSARs in classification: Comparison on a large-scale case study". In: *Journal of chemical information and modeling* 60.3, pp. 1215–1223.
- Valsecchi, Cecile et al. (2020b). "NURA: A curated dataset of nuclear receptor modulators". In: *Toxicology and Applied Pharmacology* 407, p. 115244.
- Valsecchi, Cecile et al. (2020c). "Predicting molecular activity on nuclear receptors by multitask neural networks". In: *Journal of Chemometrics*, e3325.
- Valsecchi, Cecile et al. (submitted). "Parsimonious optimization of multitask neural network hyperparameters". In: *Molecules*.
- Van Rossum, Guido and Fred L Drake Jr (1995). *Python reference manual*. Centrum voor Wiskunde en Informatica Amsterdam.
- Vangala, Subrahmanyam et al. (2011). "Translational drug discovery research: Integration of medicinal chemistry, computational modeling, pharmacology, ADME, and toxicology". In: *Encyclopedia of Drug Metabolism and Interactions*, pp. 1–54.
- Vighi, Marco et al. (2019). "Predictive models in ecotoxicology: Bridging the gap between scientific progress and regulatory applicability—Remarks and research needs". In: *Integrated environmental assessment and management* 15.3, pp. 345–351.
- Virtanen, Pauli et al. (2020). "SciPy 1.0: fundamental algorithms for scientific computing in Python". In: *Nature methods* 17.3, pp. 261–272.
- Visser, Ubbo et al. (2011). "BioAssay Ontology (BAO): a semantic description of bioassays and high-throughput screening results". In: *BMC bioinformatics* 12.1, pp. 1–16.

- Votano, Joseph R et al. (2004). "Three new consensus QSAR models for the prediction of Ames genotoxicity". In: *Mutagenesis* 19.5, pp. 365–377.
- Wang, Minjie et al. (2019). "Deep Graph Library: Towards Efficient and Scalable Deep Learning on Graphs." In:
- Wassermann, Anne Mai and Jürgen Bajorath (2011). "BindingDB and ChEMBL: online compound databases for drug discovery". In: *Expert opinion on drug discovery* 6.7, pp. 683–687.
- Weber, Christopher L, Jeanne M VanBriesen, and Mitchell S Small (2006). "A stochastic regression approach to analyzing thermodynamic uncertainty in chemical speciation modeling". In: *Environmental science & technology* 40.12, pp. 3872–3878.
- Weininger, David (1988). "SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules". In: *Journal of chemical information and computer sciences* 28.1, pp. 31–36.
- Wenzel, Jan, Hans Matter, and Friedemann Schmidt (2019). "Predictive multitask deep neural network models for ADME-Tox properties: learning from large data sets". In: *Journal of chemical information and modeling* 59.3, pp. 1253–1268.
- Wilkinson, GN et al. (1983). "Nearest neighbour (NN) analysis of field experiments". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 45.2, pp. 151–178.
- Willett, Peter (2006). "Similarity-based virtual screening using 2D fingerprints". In: *Drug discovery today* 11.23-24, pp. 1046–1053.
- Wold, Svante, Michael Sjöström, and Lennart Eriksson (2001). "PLS-regression: a basic tool of chemometrics". In: *Chemometrics and intelligent laboratory systems* 58.2, pp. 109–130.
- Wu, Zhenqin et al. (2018). "MoleculeNet: a benchmark for molecular machine learning". In: *Chemical science* 9.2, pp. 513–530.
- Xu, Yuting et al. (2017). "Demystifying multitask deep neural networks for quantitative structure–activity relationships". In: *Journal of chemical information and modeling* 57.10, pp. 2490–2504.
- Yeturu, Kalidas and Nagasuma Chandra (2008). "PocketMatch: a new algorithm to compare binding sites in protein structures". In: *BMC bioinformatics* 9.1, pp. 1–17.
- Young, Douglas et al. (2008). "Are the chemical structures in your QSAR correct?" In: *QSAR & combinatorial science* 27.11-12, pp. 1337–1345.
- Zakharov, Alexey V et al. (2019). "Novel consensus architecture to improve performance of large-scale multitask deep learning QSAR models". In: *Journal of chemical information and modeling* 59.11, pp. 4613–4624.