



UNIVERSITÀ DEGLI STUDI DI PADOVA

DIPARTIMENTO DI SCIENZE FARMACEUTICHE

SCUOLA DI DOTTORATO DI RICERCA IN SCIENZE MOLECOLARI

INDIRIZZO SCIENZE FARMACEUTICHE

CICLO XXII

TESI DI DOTTORATO

Advance Methodologies in Linear and Nonlinear
Quantitative Structure-Activity Relationships (QSARs):
from Drug Design to *In Silico* Toxicology Applications

Direttore della Scuola: Prof. MAURIZIO CASARIN

Coordinatore d'Indirizzo: Prof.ssa ADRIANA CHILIN

Supervisore: Prof. STEFANO MORO

Dottoranda: LISA MICHIELAN

31 GENNAIO 2010

Abstract

Novel computational strategies are continuously being demanded by the pharmaceutical industry to assist, improve and speed up the drug discovery process. In this scenario chemoinformatics provide reliable mathematical tools to derive quantitative structure-activity relationships (QSARs), able to describe the correlation between molecular descriptors and various experimental profiles of the compounds. In the last years, nonlinear machine learning approaches have demonstrated a noteworthy predictive capability in several QSAR applications, confirming their superiority over the traditional linear methodologies. Particularly the feasibility of the classification approach has been highlighted in solving complex tasks. Moreover, the introduction of the autocorrelation concept in chemistry allows the structural comparison of the molecules by using a vectorial fixed-length representation to serve as effective molecular descriptor.

In the present thesis we have deeply investigated the wide applicability and the potentialities of nonlinear QSAR strategies, especially in combination with autocorrelation molecular electrostatic potential descriptors projected on the molecular surface. Our intent is arranged in six different case studies that focus on crucial problems in pharmacodynamics, pharmacokinetics and toxicity fields.

The first case study considers the estimation of a physicochemical property, the aqueous solvation free energy, that strictly relates to the pharmacokinetic profile and toxicity of chemicals.

Our discussion on pharmacodynamics deals with the prediction of potency and selectivity of human adenosine receptor antagonists (hAR). The adenosine receptor family belongs to GPCR (G protein-coupled receptors) family A, including four different subtypes, referred to as A_1 , A_{2A} , A_{2B} and A_3 , which are widely distributed in the tissues. They differentiate for both pharmacological profile and effector coupling. Intensive explorative synthesis and pharmacological evaluation are aimed at discovering potent and selective ligands for each adenosine receptor subtype. In the present thesis, we have considered several pyrazolo-triazolo-pyrimidine and xanthine derivatives, studied as promising adenosine receptor antagonists. Then, a second case study focuses on the comparison and the parallel applicability of linear and nonlinear models to predict the binding affinity of human adenosine receptor A_{2A} antagonists and find a consensus in the prediction results. The following studies evaluate the prediction of both selectivity and binding affinity to $A_{2A}R$ and A_3R subtypes by combining classification and regression strategies, to finally investigate

the full adenosine receptor potency spectrum and human adenosine receptor subtypes selectivity profile by applying a multilabel classification approach.

In the field of pharmacokinetics, and more specifically in metabolism prediction, the use of multi- and single-label classification strategies is involved to analyze the isoform specificity of cytochrome P450 substrates. The results lead to the identification of the appropriate methodology to interpret the real metabolism information, characterized by xenobiotics potentially transformed by multiple cytochrome P450 isoforms.

As final case study, we present a computational toxicology investigation. The recent regulatory initiatives due to REACH (Registration, Evaluation, Authorization and Restriction of Chemicals) require the ecotoxicological and risk assessment of chemicals for safety. Most of the current evaluation protocols are based on costly animal experiments. So, chemoinformatic tools are heartily recommended to facilitate the toxicity characterization of chemical substances. We describe a novel integrated strategy to predict the acute aquatic toxicity through the combination of both toxicokinetic and toxicodynamic behaviors of chemicals, by using a machine learning classification method. The goal is to assign chemicals to different levels of acute aquatic toxicity, providing an appropriate answer to the new regulatory requirements. As preliminary validation of our approach, two toxicokinetic and toxicodynamic models have been applied in series to inspect both aquatic toxicity hazard and mode of action of a set of chemical substances with unknown or uncertain toxicodynamic information, assessing the potential ecological risk and the toxic mechanism.

Riassunto

Nuove strategie computazionali vengono continuamente richieste dall'industria farmaceutica per assistere, migliorare e velocizzare il processo di scoperta dei farmaci. In questo scenario la chemoinformatica fornisce affidabili strumenti matematici per ottenere relazioni quantitative struttura-attività (QSAR), in grado di descrivere la correlazione tra descrittori molecolari e vari profili sperimentali dei composti. Negli ultimi anni approcci non lineari di *machine learning* hanno dimostrato una notevole capacità predittiva in diverse applicazioni QSAR, confermando la loro superiorità sulle tradizionali metodologie lineari. E' stata evidenziata particolarmente la praticabilità dell'approccio di classificazione nel risolvere compiti complessi. Inoltre, l'introduzione del concetto di autocorrelazione in chimica permette il confronto strutturale delle molecole attraverso l'uso di una rappresentazione vettoriale di lunghezza fissa che serve da efficace descrittore molecolare.

Nella presente tesi abbiamo studiato approfonditamente l'ampia applicabilità e le potenzialità delle strategie QSAR non lineari, soprattutto in combinazione con i descrittori autocorrelati potenziale elettrostatico molecolare proiettato sulla superficie molecolare. Il nostro intento si articola in sei differenti casi studio, che si concentrano su problemi cruciali nei campi della farmacodinamica, farmacocinetica e tossicologia.

Il primo caso studio considera la valutazione di una proprietà fisico-chimica, l'energia libera di solvatazione acquosa, che è strettamente connessa con il profilo farmacocinetico e la tossicità dei composti chimici.

La nostra discussione in farmacodinamica riguarda la predizione di potenza e selettività di antagonisti del recettore adenosinico umano (hAR). La famiglia del recettore adenosinico appartiene alla famiglia A di GPCR (recettori accoppiati a proteine G), che include quattro diversi sottotipi, cui ci si riferisce come A₁, A_{2A}, A_{2B} e A₃, ampiamente distribuiti nei tessuti. Si differenziano sia per profilo farmacologico che per effetto cui sono accoppiati. Le intense sintesi esplorativa e valutazione farmacologica hanno lo scopo di scoprire ligandi potenti e selettivi per ogni sottotipo del recettore adenosinico. Nella presente tesi abbiamo considerato diversi derivati pirazolo-triazolo-pirimidinici e xantini, studiati come promettenti antagonisti del recettore adenosinico. Quindi, un secondo caso studio si focalizza sul confronto e l'applicabilità in parallelo di modelli lineari e non lineari per predire l'affinità di legame di antagonisti del recettore adenosinico A_{2A} umano e trovare un consenso nei risultati di predizione. Gli studi successivi valutano la predizione

sia della selettività che dell'affinità di legame ai sottotipi A_{2A}R e A₃R combinando strategie di classificazione e regressione, per studiare infine il completo spettro di potenza del recettore adenosinico e il profilo di selettività per i sottotipi hAR mediante l'applicazione di un approccio di classificazione *multilabel*.

Nel campo della farmacocinetica, e più specificamente nella predizione del metabolismo, è coinvolto l'uso di strategie di classificazione *multi-* e *single-label* per analizzare la specificità di isoforma di substrati del citocromo P450. I risultati conducono all'identificazione della metodologia appropriata per interpretare la reale informazione sul metabolismo, caratterizzata da xenobiotici potenzialmente trasformati da multiple isoforme del citocromo P450.

Come caso studio finale, presentiamo un'indagine in tossicologia computazionale. Le recenti iniziative regolatorie dovute al REACH (Registration, Evaluation, Authorization and Restriction of Chemicals) richiedono l'accertamento ecotossicologico e del rischio dei composti chimici per la sicurezza. La maggiorparte dei correnti protocolli di valutazione è basata su costosi esperimenti animali. Così, gli strumenti chemoinformatici sono caldamente raccomandati per facilitare la caratterizzazione della tossicità di sostanze chimiche. Noi descriviamo una nuova strategia integrata per predire la tossicità acquatica acuta attraverso la combinazione di entrambi i comportamenti tossicocinetico e tossicodinamico dei composti chimici, utilizzando un metodo di classificazione *machine learning*. L'obiettivo è assegnare i composti chimici a diversi livelli di tossicità acquatica acuta, fornendo un'appropriata risposta alle nuove esigenze regolatorie. Come validazione preliminare del nostro approccio, due modelli tossicocinetico e tossicodinamico sono stati applicati in serie per esaminare sia il rischio di tossicità acquatica che il modo d'azione di un set di sostanze chimiche con informazione tossicodinamica sconosciuta o incerta, valutandone il potenziale rischio ecologico ed il meccanismo tossico.

List of Papers

- I. Michielan, L.; Bacilieri, M.; Kaseda, C.; Moro, S. Prediction of the aqueous solvation free energy of organic compounds by using autocorrelation of molecular electrostatic potential surface properties combined with response surface analysis. *Bioorg. Med. Chem.* **2008**, *16* (10), 5733-5742.
- II. Michielan, L.; Bacilieri, M.; Schiesaro, A.; Bolcato, C.; Pastorin, G.; Spalluto, G.; Cacciari, B.; Klotz, K. N.; Kaseda, C.; Moro, S. Linear and nonlinear 3D-QSAR approaches in tandem with ligand-based homology modeling as a computational strategy to depict the pyrazolo-triazolo-pyrimidine antagonists binding site of the human adenosine A_{2A} receptor. *J. Chem. Inf. Model.* **2008**, *48* (2), 350-363.
- III. Michielan, L.; Terfloth, L.; Gasteiger, J.; Moro S. Comparison of multilabel and single-label classification applied to the prediction of the isoform specificity of cytochrome P450 substrates. *J. Chem. Inf. Model.* **2009**, *49* (11), 2588-2605.
- IV. Michielan, L.; Bolcato, C.; Federico, S.; Cacciari, B.; Bacilieri, M.; Klotz, K. N.; Kachler, S.; Pastorin, G.; Cardin, R.; Sperduti, A.; Spalluto, G.; Moro, S. Combining selectivity and affinity predictions using an integrated support vector machine (SVM) approach: an alternative tool to discriminate between the human adenosine A_{2A} and A₃ receptor pyrazolo-triazolo-pyrimidine antagonists binding sites. *Bioorg. Med. Chem.* **2009**, *17* (14), 5259-5274.
- V. Michielan, L.; Federico, S.; Terfloth, L.; Cacciari, B.; Klotz, K. N.; Spalluto, G.; Gasteiger, J.; Moro, S. Exploring potency and selectivity receptor antagonist profiles using a multilabel classification approach: the human adenosine receptors as a key study. *J. Chem. Inf. Model.* **2009**, *49* (12), 2820-2836.
- VI. Michielan, L.; Pireddu, L.; Floris, M.; Moro, S. Support vector machine (SVM) as alternative tool to assign acute aquatic toxicity warning labels to chemicals. *Mol. Inf.* **2010**, *in press*.

Contents

1	Introduction	1
1.1	Challenges in pharmaceutical research	1
1.2	Predictive toxicology	3
1.3	Motivation	5
2	Methods	7
2.1	QSAR	8
2.2	Molecular structure building	10
2.3	Molecular descriptors	11
2.3.1	Molecular Electrostatic Potential (MEP)	12
2.3.2	Autocorrelation vectors	13
2.3.3	Sterimol parameters	15
2.3.4	Other molecular descriptors	16
2.4	Data autoscaling	19
2.5	Linear strategies	20
2.5.1	Single and multiple regression	20
2.5.2	Principal Component Analysis	21
2.5.3	Projection to Latent Structures by means of Partial Least Square	24
2.6	Nonlinear strategies	28
2.6.1	Response Surface Analysis	28
2.6.2	Support Vector Machine	31
2.6.3	Cross-training with Support Vector Machine	36
2.6.4	Artificial Neural Networks	38
2.7	Validation and statistical evaluation	42
2.8	Software	45
3	Estimation of the aqueous solvation free energy	47
3.1	Introduction	48
3.2	Results and discussion	50
3.3	Final remarks	54

CONTENTS

4	Parallel application of linear and nonlinear QSAR methodologies	55
4.1	Introduction	56
4.2	Human A _{2A} R antagonists dataset	56
4.3	Results and discussion	59
4.3.1	PLS and RSA models	59
4.3.2	External test set prediction	63
4.4	Final remarks	65
5	Isoform specificity of cytochrome P450 substrates	67
5.1	Cytochrome P450 in drug metabolism	68
5.2	Computational approaches to CYP450	69
5.3	Dataset	71
5.3.1	Data set 1	72
5.3.2	Data set 2	72
5.4	Results	73
5.4.1	Multilabel classification with Data set 1	73
5.4.2	Single-label classification with Data set 2	77
5.4.3	Validation of the models with an external test set	79
5.5	Discussion	81
5.5.1	Aspects related to the data set	81
5.5.2	Considerations on the selected descriptors	81
5.5.3	ct-SVM and CPG NN models	82
5.5.4	ct-SVM and SVM models	82
5.5.5	Analysis of some classified compounds	83
5.6	Final remarks	85
6	Classification and regression to investigate selectivity and binding affinity	87
6.1	Introduction	88
6.2	Dataset	89
6.3	Results and discussion	91
6.3.1	SVM classification model	93
6.3.2	SVM regression models	94
6.3.3	Validation of in series SVM _{class} and SVR models	96
6.4	Final remarks	100
7	Exploring potency and selectivity of hAR antagonists	101
7.1	Introduction	102
7.2	Adenosine receptor antagonists	104
7.3	Dataset	105
7.4	Results and discussion	108
7.4.1	<i>auto</i> MEP/ct-SVM classification models	109

CONTENTS

7.4.2	Internal validation	110
7.4.3	External validation	113
7.5	Final remarks	116
8	Prediction of toxicodynamic and toxicokinetic properties	117
8.1	Introduction	118
8.2	Dataset	121
8.3	Results and discussion	123
8.3.1	TOX <i>class</i> model	124
8.3.2	MOA <i>class</i> model	126
8.3.3	Applicability of TOX <i>class</i> and MOA <i>class</i> models . . .	128
8.4	Final remarks	131
9	General conclusions	133
	Bibliography	135
	Supporting Information	154

CONTENTS

List of Abbreviations

ADMET	Absorption, Distribution, Metabolism, Elimination, Toxicity
ANN	Artificial Neural Networks
AR	Adenosine Receptor
<i>auto</i> MEP	Autocorrelation Molecular Electrostatic Potential
CoMFA	Comparative Molecular Field Analysis
CPG NN	Counter-Propagation Neural Network
ct-SVM	Cross-training with Support Vector Machine
CYP450	Cytochrome P450
DOE	Design of Experiment
EPAFHM	Environmental Protection Agency Fathead Minnow Acute Toxicity
GPCR	G Protein-Coupled Receptor
HOMO	Highest Occupied Molecular Orbital
LUMO	Lowest Unoccupied Molecular Orbital
MCC	Matthews Correlation Coefficient
MEP	Molecular Electrostatic Potential
MLR	Multiple Linear Regression
MOA	Mode of Action
OECD	Organization of Economic Cooperation and Development
PEOE	Partial Equalization of Orbital Electronegativity
PC	Principal Component
PCA	Principal Component Analysis
PCR	Principal Component Regression
PLS	Projection to Latent Structures by means of Partial Least Square
QSAR	Quantitative Structure-Activity Relationship
QSPR	Quantitative Structure-Property Relationship
RBF	Radial Basis Function
REACH	Registration, Evaluation, Authorization and Restriction of Chemicals
RSA	Response Surface Analysis
SOM	Self Organizing Map
SVM	Support Vector Machine
SVR	Support Vector Regression

Preface

Chemoinformatics can provide both description and understanding of various pharmacodynamic, pharmacokinetic and toxicity properties of the compounds by quantitative structure-activity relationships (QSARs). In such applications machine learning methods represent valuable mathematical tools able to solve complex tasks, overcoming the potentialities offered by the classical linear QSAR strategies. The six case studies presented in this thesis constitute interesting examples of the use of both regression and classification models in combination with particular 3D autocorrelated molecular descriptors, to predict different experimental properties.

The first chapter comprises an introductory overview on the modern critical aspects of pharmaceutical research. In this context the recent driving forces due to the international regulatory initiatives are also introduced.

The second chapter focuses on the basic concepts of QSAR analysis and includes the techniques applied in the six case studies, that are described in the following chapters. The details of the calculations are also reported.

The third chapter summarizes the results of the application of a nonlinear QSAR strategy, by using Response Surface Analysis, to predict the aqueous solvation free energy of organic compounds (*Paper I*).

The fourth chapter represents an evaluation of the use of linear and nonlinear approaches in parallel, to obtain a consensus in the prediction of the binding affinity of human A_{2A} adenosine receptor antagonists (*Paper II*).

The fifth chapter analyzes the isoform specificity of cytochrome P450 substrates by comparing the multi- with the single-label classification methods, to find the best model able to interpret an important phase of the metabolism (*Paper III*).

The sixth chapter presents an introduction to the prediction of the receptor subtype selectivity task. In more detail, we consider A_{2A} and A₃

PREFACE

adenosine receptor antagonists to derive an integrated strategy based on a first classifier and a second regression model by applying Support Vector Machine. The goal is to simultaneously discriminate A_{2A}R versus A₃R antagonists and to predict the binding affinity to the corresponding receptor subtype for unknown compounds used as test set (*Paper IV*).

The seventh chapter deeply discusses the selectivity of human adenosine receptor antagonists by extending the case study reported in the sixth chapter to the whole adenosine receptor family (A₁, A_{2A}, A_{2B} and A₃ subtypes). More specifically, we present a novel application of the multilabel classification approach. After introducing three classifiers, based on decreasing thresholds of potency, both potency profile and selectivity are predicted by applying the classification models as in series quantitative sieves (*Paper V*).

The eighth chapter focuses on the estimation of ecotoxicological endpoints and investigates the classification approach as alternative tool to predict toxicokinetic and toxicodynamic properties of chemicals. In particular, a first model is derived to assign chemicals to different levels of acute aquatic toxicity; a second classifier provides the prediction of the mode of action (MOA) of toxic compounds (*Paper VI*).

In light of these investigations, we have draft the final conclusions, that emphasize the appreciable performances of nonlinear QSAR techniques to predict several pharmacodynamic, pharmacokinetic and toxicity profiles.

CHAPTER 1

Introduction

1.1 Challenges in pharmaceutical research

Drug discovery process is aimed at bringing to market new therapeutic agents with desirable pharmacodynamic profile and favourable ADMET (Absorption, Distribution, Metabolism, Elimination and Toxicity) properties. The target selectivity is a further crucial requirement for drugs to avoid efficacy problems and limiting side-effects, incurring when the compounds do not differentiate between different receptors. The goal is to design drugs without, or with minimum, side-effects while retaining the desired function.

Nowadays, the pharmaceutical research has to face many obstacles with the result of a very low success rate, regardless the extremely growing employed resources. [1] According to recent Tufts Center for the Study of Drug Development data, drug development, starting from the clinical trials to the final approval, is about 8.5 years long with a cost exceeding \$40 billion, and only 21.5% of clinical success rate. In fact, the pharmaceutical research is actually involved in the study of more complex diseases, while the increasing size and costs of the clinical trials, the candidate high attrition rates and the late occurrence of failures in the clinical studies are emphasized as the main negative contributions to the economic profile of the research in the pharmaceutical companies. Various aspects have been identified and reported as the causes of the high level of attrition undergone by the compounds during the developmental stages. [2] The reasons for attrition have changed over time and in 2000 some problems of efficacy, safety or toxicological effects were recognized as highly responsible for the failures, covering more than 50% of the

causes for abandoning, as shown in Figure 1.1.1. Clearly, most of the efforts are directed to unproductive clinical trials, since most drug candidates are eliminated late in the clinical development without recovering the starting investment. [2]

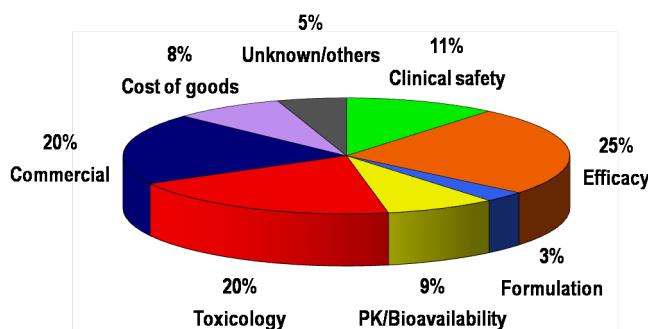


Figure 1.1.1: Reasons for attrition (2000). [2]

In the last decades the potentialities of the combinatorial chemistry have provided new large databases with unknown compounds. Therefore, at the early stage of drug discovery suitable computational approaches are needed to shorten the time and increase the success rate by deriving *in silico* models for the prediction of some corresponding desirable properties. Then, several computational tools have been developed to eliminate lead compounds with undesirable profiles, before they enter the costly late phases of drug development, and to let compounds to proceed in the optimization step. The result would be the reduction of the attrition rate in drug discovery. [3, 4]

However, in the initial stage of drug development the optimization of the properties related to absorption, distribution, metabolism and elimination is expected as well as the study of the pharmacodynamic profile of novel chemical entities. [5, 6] Recent *in silico* methods have focused on the metabolic endpoints and the prediction of drug metabolism directly from structure represents an advanced approach integrated into expert systems. [7, 8] Moreover, the computational tools are suggested to successfully assist the *in vitro* methods for studying the human drug metabolism in order to compensate the limitation of the use of each of these approaches alone. [9]

Quantitative Structure-Activity Relationships (QSARs) or Quantitative Structure-Property Relationships (QSPRs) approaches represent probably the most robust well-known tools to mathematically analyze the correlation between the molecular properties and an experimental endpoint.

Among various algorithms available, novel nonlinear machine learning methods have been applied for the prediction of pharmacodynamic and AD-MET properties. [10-12] In more detail, many QSARs have been attempted to correlate molecular descriptors with druglikeness, activity, selectivity, toxicity and several pharmacokinetic properties, such as aqueous solubility, blood-brain barrier and human intestinal absorption, plasma protein binding, oral bioavailability or steady-state volume of distribution. [12-23]

1.2 Predictive toxicology

The regulatory frame is considered an additional obstacle in the drug discovery process, since a very accurate risk evaluation is required to assess the safety of the drug once on the marketplace. [1] Recently, the poorly efficient risk assessment process and the uncomplete information on hazard properties of chemicals has driven the need for new regulatory dispositions, that have been introduced in European Community on June 1 2007 with the chemical management system REACH (Registration, Evaluation, Authorization and Restriction of Chemicals). [24, 25] The immediate objective of REACH, in a relatively short time period (11 years), is to characterize the toxicological properties of a large group of substances, manufactured or imported in quantities in excess of 1 ton per year. The attempt of this regulation is the increase in the production of useful data for the decisions involving the improvement of the protection of human health and environment, through a better identification and understanding of the chemical properties hazardous to safety. Diverse expensive animal testing experiments are usually expected for *in vivo* toxicological data requirements, as shown in Figure 1.2.1.

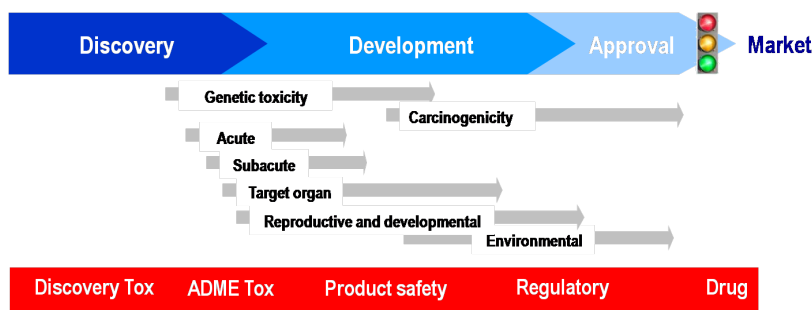


Figure 1.2.1: Classical toxicological testing procedures in the drug discovery process.

The experimental toxicity assessment is relevant for human health. Unfortunately, the huge resource demand deals with the large amount of chemicals needed in the experiments and the cost of animals. Thus, very recently, a paradigm shift has been suggested in toxicology with a specific reference to the computational methods as reliable support in the toxicity assessment. [26] In particular, the predictive toxicology represents an attractive tool to investigate the effects on human health and the potential ecotoxicological risk of chemical substances in the drug discovery process as well as in the environmental hazard assessment. In this context, pharmaceuticals, personal health care products, nutritional ingredients and products of the chemical industries are all potentially dangerous and need to be assessed. Then, the aim of the computational toxicology is to assist their evaluation through *in silico* models, by assigning a priority for the traditional toxicological tests and providing information about the consequences to their exposition, as graphically represented in Figure 1.2.2¹. [27]

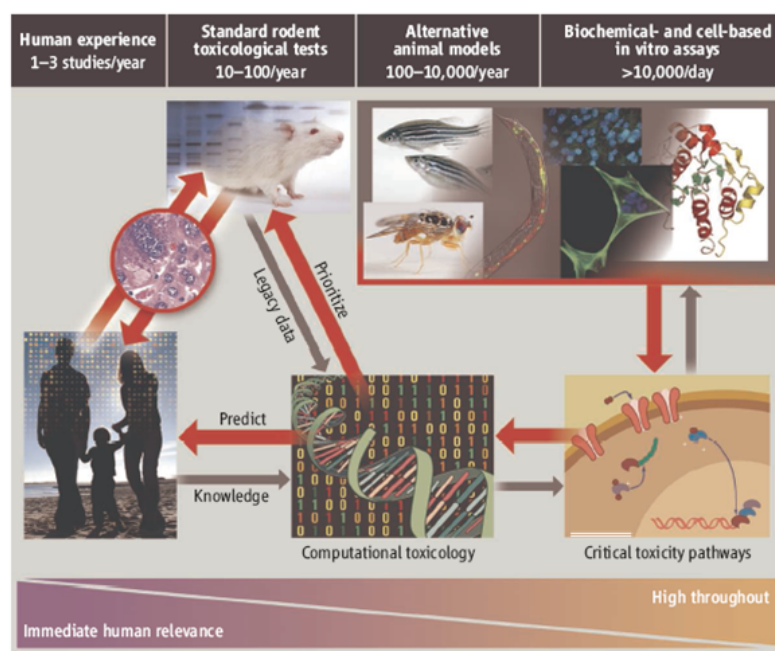


Figure 1.2.2: Roles of computational toxicology, that yields data predictive of results from animal toxicity studies. This discipline will allow prioritization of chemicals for further testing and can assist in prediction of risks to humans.

¹Adapted from Collins, F. S.; Gray, G. M.; Bucher, J. R. *Science* 2009, 319, 906-907.

The same introduction of REACH should speed up the risk assessment process by prioritizing compounds for traditional toxicity testing and providing information on the Exposure Scenarios (ESs) concerning the chemical safety profile. [27] In fact, REACH promotes alternative tools to collect extensive information on hazards of chemicals in order to reduce animal use in toxicology. As a consequence, several Intelligent or Integrated Testing Strategies (ITS) have been proposed as rapid, efficient approaches to obtain exposure and effects data and identify different modes of toxic action. [28, 29] Moreover, *in vitro* or computational methods, optimized *in vivo* studies, chemical categories, read-across analysis and thresholds of toxicological concern (TTCs) are admitted non-testing strategies to replace missing data or endpoints, and profitably reduce costly animal experiments. [27]

So far, powerful computational toxicology prediction systems have been developed for the exposure and hazard assessment to satisfy the new regulatory requests. [30] In drug discovery the *in silico* approaches, and especially machine learning methodologies, for the toxicity prediction of safety-relevant endpoints are precious contributions to early discovery of adverse drug reactions. [31, 32] A brief overview of both tools and models in computational toxicology have been considered. [33] Furthermore, a recent review about the toxicity databases available, *in silico* toxicology tools together with their advantages and limitations has been published. [34]

In toxicology QSARs are widely used approaches to infer the toxicological properties of compounds from their molecular structure. [35] Several studies have focused on the prediction of the environmental toxicity properties of drugs. [36] Aquatic toxicity of chemical substances is lately investigated as basic information in the hazard and environmental risk assessment. [37-41]

1.3 Motivation

Nonlinear strategies offer a useful tool by deriving quantitative structure-activity relationships for the investigation of new molecular structures with the goal to facilitate their evaluation at the early stage of drug discovery process. The present thesis aims to demonstrate the satisfactory predictive capability and the enormous potentialities of several nonlinear QSAR approaches for the prediction of properties ranging from the pharmacodynamics to the ADMET profiles. The chapters 3-8 separately explore six case stud-

ies to evaluate or compare the performances of linear and nonlinear QSAR strategies. We discuss various models by combining different descriptor sets with several algorithms and we have validated our results by introducing new compounds as test set. In more detail, we have predicted the aqueous solvation free energy as physicochemical property (chapter 3). In three case studies the pharmacodynamic property investigated is the binding affinity to the different human adenosine receptor subtypes, by focusing on potency to A_{2A} (chapter 4) or A_{2A}/A_3 subtypes (chapter 6) and on selectivity (chapters 6 and 7) predictions. The chapter 8 relates to the recent debates on the limited toxicological information and propose a novel strategy to predict toxicokinetic and toxicodynamic behaviors of chemicals.

Moreover, we would like to evaluate the efficiency of the molecular descriptors selected in rationalizing the chemical structures to derive robust regression or classification models for the experimental properties in analysis. The autocorrelation seems to represent an efficient strategy to develop QSAR models for structurally different compounds. More specifically, we consider the autocorrelated descriptors encoding for molecular electrostatic potential computed on the molecular surface. Finally, while investigating some of our case studies, we present the results of the further introduction of descriptors (global, topological, quantum chemical, etc.) to interpret cytochrome P450 isoform specificity and toxicity mechanisms.

CHAPTER 2

Methods

In the field of chemoinformatics, quantitative structure-activity relationships (QSAR) have demonstrated to be powerful tools in the prediction of simple chemical-physical properties as well as complex pharmacodynamic, pharmacokinetic and toxicological profiles. Several key steps are involved in any QSAR approach: data collection for the molecular structure building and the calculation of suitable molecular descriptors, data pretreatment, model generation and optimization, and finally, the statistical validation and evaluation of the model. Machine learning represents a well-known family of algorithms based on a solid statistical theory able to handle complex problems, and especially developed as modeling methods. Lately, among non-linear strategies, the machine learning methodologies have been applied as robust alternatives to the traditional linear QSAR techniques, such as the Partial Least Squares (PLS) analysis. To date, they have shown promising potentialities in many scientific studies. Very recently, the current mathematical techniques applied in QSAR approaches have been reviewed. [42]

Artificial neural networks or support vector machines have gained interesting progresses in both classification and quantitative prediction of different endpoints. Based also on the increasing availability of experimental data, these techniques have been further applied in the assessment of several pharmacokinetic properties. [43] Some valid machine learning applications in the prediction of the cytochrome P450 isoform specificity, interactions and inhibition were summarized. [44, 45] In the present thesis, the classification approach is demonstrated to be able to investigate the selectivity problem. However, the model validity has to deal with the recent regulatory context created by REACH law and some reference principles need to be satisfied.

2.1 QSAR

Most drugs attain the therapeutic activity through a specific target recognition process. In the optimization step of drug candidates, whether the information on the target is not available, the *ligand-based* drug design approach might be applied for the evaluation of new compounds.

Chemoinformatics, combining knowledge from different fields, is characterized by the interest on the chemical structures to extract information on the corresponding activity or properties. Among the chemoinformatic methods, the quantitative structure-activity relationships (QSAR) or quantitative structure-property relationships (QSPR) relate molecular descriptors to the quantitative measure of a property. [46] QSAR are based on the general principle that the chemical structures can be mathematically codified as distributions of molecular properties, or *molecular descriptors*. Then, an appropriate statistical modeling method is used to achieve the correlation between the molecular descriptors (X variables) and the defined property (Y variable), such as biological activity, volume of distribution, toxicity, to predict the corresponding property of unknown compounds. [46-49] In QSAR analysis, a training set and a test set are selected from a starting collection of data (Figure 2.1.1).

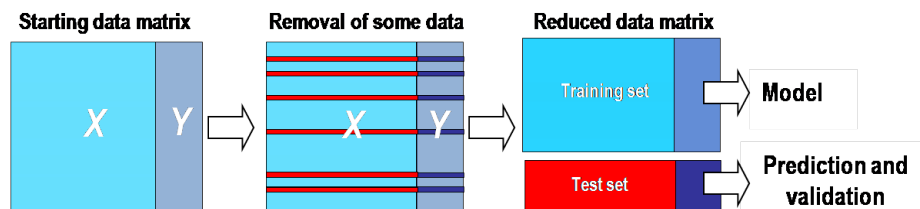


Figure 2.1.1: Training and test set selection procedure. The X matrix contains molecular descriptors, the property data are included in the Y matrix.

The training set is used to generate the model, that is then statistically evaluated on its ability to predict the property values of a test set. Finally, the QSAR model can be applied to the prediction of the property of new chemical structures. The main processes involved in a QSAR analysis are represented in Figure 2.1.2.

Moreover, based on the nature of the relationship between the molecular descriptors and the property in analysis, linear and nonlinear strategies can be distinguished. In particular, multiple regression, principal component

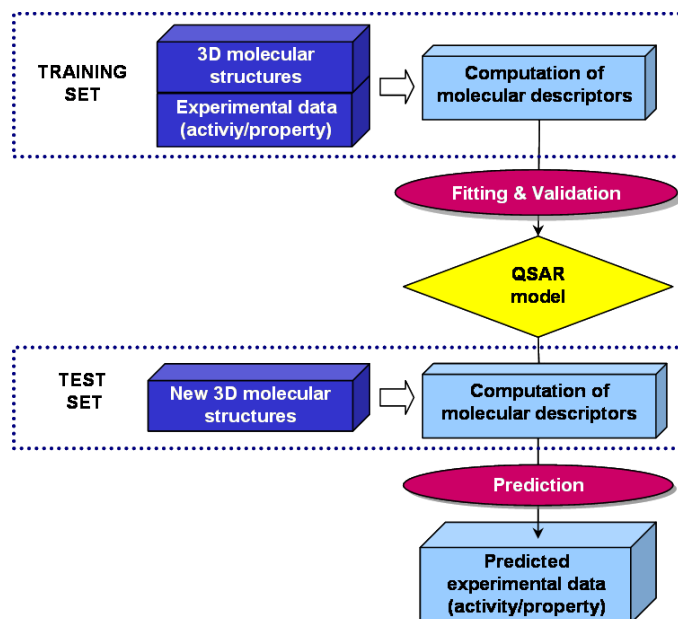


Figure 2.1.2: Crucial steps in QSAR analysis involving the training and the test sets: model generation, validation and prediction.

analysis, projection to latent structures by means of partial least square will be discussed as linear strategies, while we will describe response surface analysis, support vector machine and artificial neural networks as nonlinear techniques.

Considering the input data, the regression and classification approaches can be defined. If the considered property is represented by continuous data, the QSAR model is referred to as regression regardless whether the relationship is linear or not; if binary data are introduced as qualitative measure of the property for the model generation, discrete classification models, or SARs, are derived with the aim to separate the compounds into different classes (Figure 2.1.3). [50]

The classification allows to assign a sample to one class (*single-label*) or to more classes (*multilabel*) to derive a qualitative prediction. Moreover, in the traditional single-label classification the classes are considered mutually exclusive; when the samples belong to multiple classes, the multilabel classification analysis seems to be more appropriate.

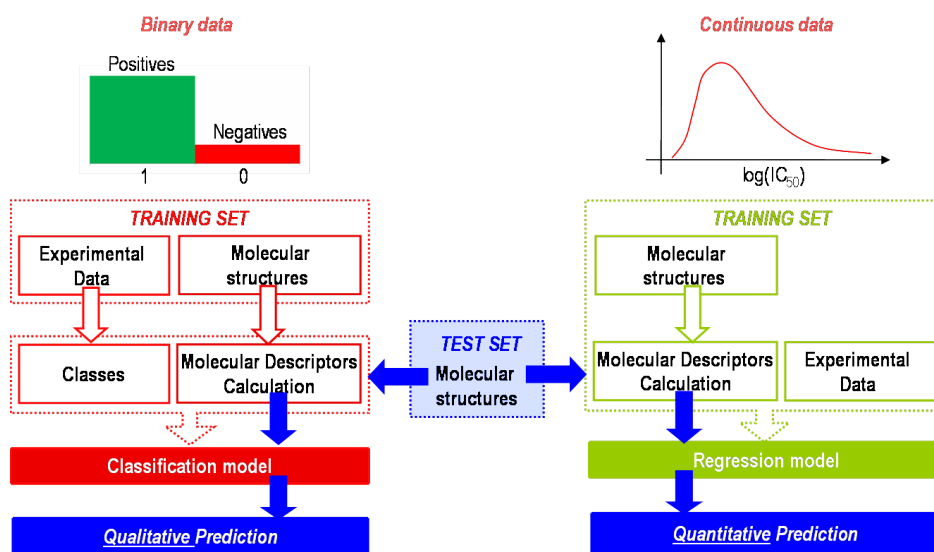


Figure 2.1.3: Classification and regression approaches. In classification analysis the input data are structured into classes, while the regression approach is based on continuous data. Finally, qualitative and quantitative predictions are carried out, respectively.

2.2 Molecular structure building

The database generation step is required to order the selected molecules and visualize their three-dimensional structures. In the present thesis, 3D models of all molecules in the training sets, validation sets, internal and external test sets were obtained using the 3D structure generator Corina, setting parameters to standard values. Corina is an integral part of ADRIANA. Code suite. [51]

Conformer selection is a crucial step in the approaches considering 3D molecular descriptors. If the information about the possible binding mode of the compounds to the corresponding target is limited, we have decided to select the energetically most stable conformers produced by the software conformational analysis. We verified that the conformations derived by Corina are reasonably similar to the poses obtained by docking experiments. Protonation states are selected in agreement with the corresponding pK_a at the physiological pH value (7.4 unit).

The final molecules are globally neutral, so, they can be used for the calculation of the molecular descriptors.

2.3 Molecular descriptors

In a QSAR study we need for each molecule some numerical properties, through a mathematical description of their structure. [46, 49] The molecular descriptors are numerical representations of physicochemical or topological properties. As anticipated, they can be used as independent variables in QSARs and the descriptors selection should be accurate according to the type of experimental data for achieving satisfactory modeling results.

The descriptors derive from experimental measurements, theoretical calculations or mathematical operations, and they may refer to the whole molecule or one molecular fragment. Moreover, they can be represented as scalars or vectors and they are defined according to the number of dimensions they require for the computation. Consequently, their complexity is related to the dimensionality of the molecular representation (Figure 2.3.1)¹.

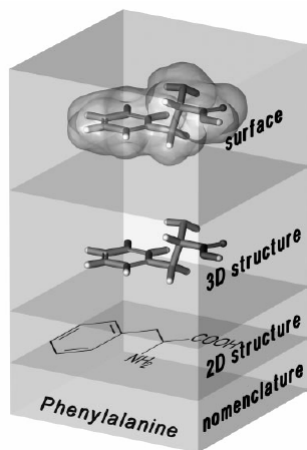


Figure 2.3.1: Dimensional levels of structural information.

Simple 1D descriptors need knowledge of only the code of a molecule and consider the presence of a particular element. More complex global molecular properties or functional-group counts require the connection table to be computed (2D descriptors). The 3D descriptors reflecting molecular shape or the distribution of a property on the molecular surface need the previous computation of the three-dimensional molecular structure.

We have selected different molecular descriptors for our analysis, as described in the following paragraphs.

¹Adapted from Gasteiger, J.; Engel, T. *Cheminformatics*, Wiley-VHC, 2003.

2.3.1 Molecular Electrostatic Potential (MEP)

Autocorrelation molecular electrostatic potential (MEP) vectors have been introduced by Gasteiger and collaborators as molecular descriptors computed on the molecular surface (Figure 2.3.2a). [52] In our models MEPs derive from a classical point charge model: the electrostatic potential for each molecule is obtained by moving a unit positive point charge across the molecular surface, and it is calculated at various points j on this surface by applying the following equation:

$$V_i = \frac{1}{4\pi\epsilon_0} \sum_i^{\text{atoms}} \frac{q_i}{r_{ji}} \quad (2.3.1)$$

where q_i represents the partial charge of each atom i and r_{ji} is the distance between points j and atom i . Starting from the 3D model of a molecule and its partial atomic charges, the electrostatic potential is calculated for points on the molecular surface.

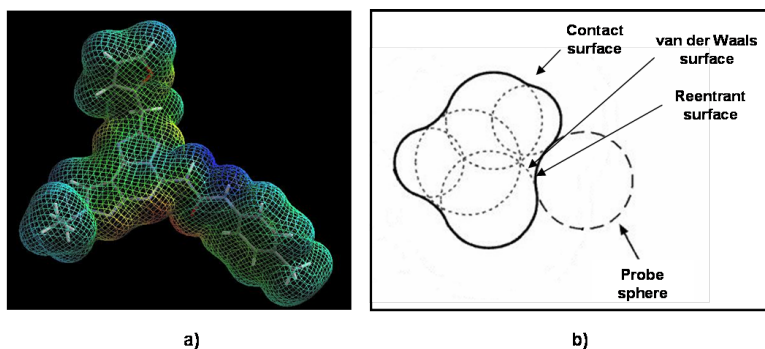


Figure 2.3.2: a) Representation of the molecular electrostatic potential; b) references for the surface calculation.

Partial atomic charges were calculated by the PEOE (*Partial Equalization of Orbital Electronegativity*) method and its extension to conjugated systems implemented in *ADRIANA.Code*. [51, 53, 54] As reference for the surface calculation, we have considered Connolly's solvent accessible surface, obtained by moving a probe sphere on the van der Waals surface, as shown in Figure 2.3.2b². Connolly's solvent accessible surface with a solvent radius of 2.0 Å and van der Waals radius reduction factor³ of 1.00 have been used

²Adapted from Gasteiger, J., Engel, T. *Chemoinformatics*, Wiley-VHC, 2003.

³Factor of reduction, which the van der Waals radius is multiplied for.

to project the corresponding MEP. Once the autocorrelation function has been applied, the autocorrelation vector is derived. [51, 53, 54]

2.3.2 Autocorrelation vectors

The autocorrelation function transforms the constitution of a molecule into a fixed length representation. In fact, as originally computed, the MEP properties depend on the spatial orientation of the molecule and a previous alignment is needed to compare different molecular structures. The mathematical notion of autocorrelation is schematically reported in Figure 2.3.3.

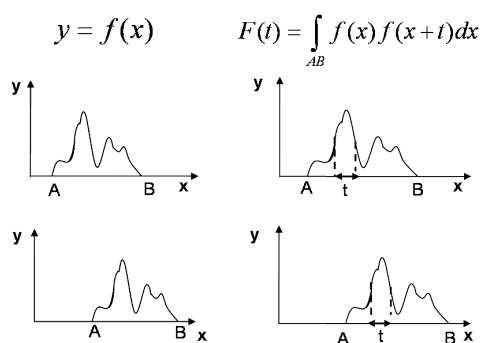


Figure 2.3.3: a) Effect of the translation of $f(x)$ on x axis; b) geometrical meaning of the autocorrelated function $F(t)$.

Given the function $f(x)$ measuring a property in AB domain, if the inner variable t is introduced, the new autocorrelated function $F(t)$ is defined as:

$$F(t) = \int_{AB} f(x)f(x+t)dx \quad (2.3.2)$$

where $F(t)$ is an intrinsic descriptor, not relying on the external reference systems and on the translation of the function $f(x)$ on the x axis. Conversely, the translation changes $f(x)$ values for any $x \in AB$ domain.

Firstly investigated by Moreau and Broto, this concept was introduced in chemistry to analyze the properties of different molecules without molecular superimposition. [55, 56] By formulating a topological definition, they considered that a certain property p of an atom i can be correlated with the corresponding property p of atom j and the products of p values can be summed over all atom pairs having a certain topological distance d .

Each component of the autocorrelation vector is consequently calculated as follows:

$$A(d) = \sum_{i,j}^N p_i p_j \delta(d_{ij}, d) \quad \delta = \begin{cases} 1 & \forall d_{ij} = d \\ 0 & \forall d_{ij} \neq d \end{cases} \quad (2.3.3)$$

where $A(d)$ is the autocorrelation coefficient referring to atom pairs i,j at the topological distance d and p_i is the atomic property. [55, 56]

The molecular recognition processes and the physicochemical phenomena involve interactions between molecular surfaces and, therefore, representations of molecular surfaces should be appropriate to understand the diversity in the binding affinity and chemical behaviors. We are under the restriction of having to represent molecular surfaces of different size, thus the autocorrelation concept has been extended to 3D structures to achieve this goal. [52, 57, 58] Starting from the topological autocorrelation examples of Moreau and Broto, a set of randomly distributed points on the molecular surface has to be generated first. Then, all distances between the surface points are calculated and sorted into the preset intervals d_{lower} - d_{upper} . Finally, the autocorrelation coefficients are computed:

$$A(d_{lower}, d_{upper}) = \frac{1}{L} \sum_{j=i}^N \sum_{i=1}^N p_j p_i \delta(d_{ij}, d_{lower}, d_{upper}) \quad (2.3.4)$$

$$\delta = \begin{cases} 1 & \forall d_{lower} < d_{ij} \leq d_{upper} \\ 0 & \forall d_{ij} \leq d \vee d_{ij} > d_{upper} \end{cases}$$

According to equation 2.3.4, the component of the autocorrelation vector $A(d_{lower}, d_{upper})$, referring to the i,j distance d in the interval d_{lower} - d_{upper} is the sum of all products of the properties p_i and p_j for atoms i and j . We consider N the number of atoms in the molecule and L a parameter representing the total number of distances in the interval d_{lower} - d_{upper} .

The application of this concept made possible the comparison of different molecular properties, as this 3D descriptor represents a compressed expression of the distribution of the property p on the molecular surface, as shown in Figure 2.3.4.

For the calculation of the autocorrelation coefficients we have applied the default values for parameter computation, since no improving in statistical model capability was observed by changing them in various way. Default

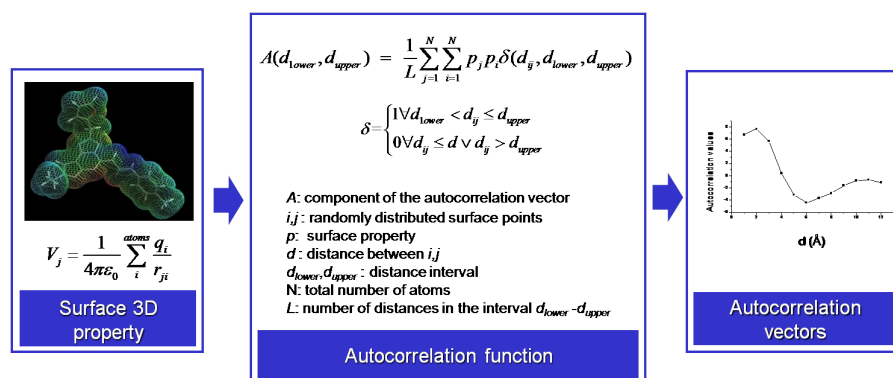


Figure 2.3.4: Calculation of autoMEP vectors, from the 3D molecular structure to a fixed length representation.

parameters values are the following: $d_{lower} = 1 \text{ \AA}$; $d_{upper} = 13 \text{ \AA}$; $L = 12$; *point density* = 10 points/\AA^2 ; *vdW radius reduction factor* = 1.000. Consequently, we have derived 12 autocorrelation vectors per molecule, computed at the 12 (L value) distances in the interval from 1 to 13 \AA with a step width of 1 \AA . By considering the size of the molecules in our datasets, we decided that the step width of 1 \AA was sufficient to describe in an accurate way the distribution of the MEP property on the molecular surface. This transformation produces a molecular descriptor which is a unique fingerprint of each molecule under consideration.

2.3.3 Sterimol parameters

3D topological Sterimol descriptors ($B1$, $B2$, $B3$, $B4$ and L) have been introduced by Verloop to consider the volumes of the molecular substituents with different geometries (Figure 2.3.5). [59]

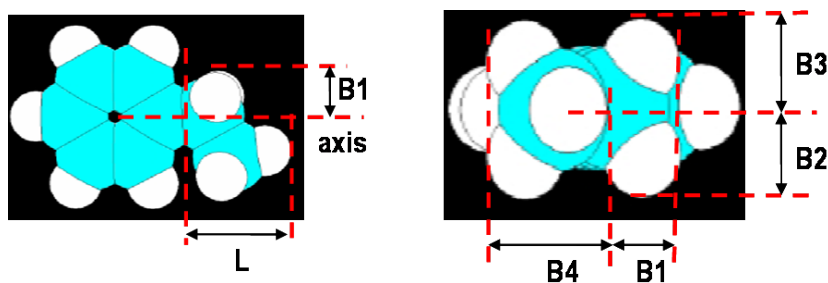


Figure 2.3.5: Geometrical meaning of Sterimol parameters.

The transposition of this concept to the whole molecule produces global

descriptors, that, being intrinsically independent on rotation and translation of the molecule, can be used together with *autoMEP* vectors. [60] In more detail, L is the length of the molecule and refers to the principal axis, $B1$ is perpendicular to x axis and it is the smallest distance from L axis to a side of the parallelepiped containing the molecule, $B2$ - $B4$ have a similar geometrical meaning and they are perpendicular to $B1$.

2.3.4 Other molecular descriptors

Further descriptors calculated in our studies are listed in Table 2.1. These descriptors are, to a large extent, 2D and 3D molecular descriptors and reflect shape and reactivity properties.

The capability for participating in hydrogen bonding is described directly by the number of hydrogen-bonding acceptors/donors or the hydrogen-bond acceptor/donor potential, or indirectly by the number of basic nitrogen atoms and the number of acidic groups.

The highest hydrogen-bonding acceptor potential is defined as the maximum lone-pair electronegativity on an atom considering all N, O, and F atoms in a compound. The highest hydrogen-bonding donor potential is defined as the most positive charge on the hydrogen atom in the functional groups -OH, -NH, and -SH in a compound.

LogP(o/w) , or log of the n -octanol/water partition coefficient, describes the partitioning equilibria. This 2D molecular descriptor was initially used in drug design to quantify lipophilicity. Various methods for the estimation of the experimental logP values have been developed and one of the well-known techniques, published by Nys and Rekker, is based on additive fragment contributions to the total molecular lipophilicity. [46] In our studies, logP(o/w) is calculated from a linear atom type model by considering a pH value of the aqueous phase such that the predominant form of the chemical is un-ionized.

As anticipated, two-dimensional topological autocorrelations derive atom properties from the structure diagram, whereas spatial autocorrelation descriptors are based on the information encoded by the 3D molecular structure. We have to consider atom identities or a property (σ -electronegativity, π -electronegativity, σ -charge or π -charge) for their computation. In both cases, the function of autocorrelation was applied to derive the autocorrelation vectors.

Table 2.1: List of descriptors used in our analysis, arranged by class.

No.	Name	Details	Ref(s).
Global			
1	MW	molecular weight	[61]
2	HAccPot	highest hydrogen-bond acceptor potential	[61]
3	HDonPot	highest hydrogen-bond donor potential	[61]
4	HAcc	number of hydrogen-bonding acceptors derived from the sum of nitrogen and oxygen atoms in the molecule	[61]
5	HDon	number of hydrogen-bonding donors derived from the sum of NH and OH groups in the molecule	[61]
6	TPSA	topological polar surface area	[62]
7	ASA	approximate surface area	[63]
8	α	mean polar polarizability	[64-67]
9	μ	molecular dipole moment	[68]
10	logP(o/w)	log of <i>n</i> -octanol/water partition coefficient	
Topological			
11-12	χ^0, χ^1	connectivity χ indeces	[69]
13-14	κ_1, κ_2	κ shape indeces	[69]
15	W	Wiener path number	[70]
16	χ^R	Randic index	[68]
Size/Shape			
17	D_3	diameter	[71]
18	R_3	radius	[70]
19	I_3	geometric shape coefficient	[71, 72]
20	r_2	radius perpendicular to D_3	
21	r_3	radius perpendicular to D_3 and r_2	
22-24	$\lambda_1, \lambda_2, \lambda_3$	principal moment of inertia	[68]
25	r_{gyr}	radius of gyration	[73, 74]
26	r_{span}	radius of the smallest sphere, centered at the center of the mass which completely encloses all atoms in the molecule	[74]
27	ϵ	molecular eccentricity	[68]
28	Ω	molecular asphericity	[68]
Quantum Chemical			
29	HOMO	energy (eV) of Highest Occupied Molecular Orbital	[75]
30	LUMO	energy (eV) of Lowest Unoccupied Molecular Orbital	[75]
31	LUMO-HOMO	difference between LUMO and HOMO	[75]
Functional-Group Counts			
32	n_{aliph_amino}	number of aliphatic amino groups	
33	n_{aro_amino}	number of aromatic amino groups	

(continued on next page)

No.	Name	Details	Ref(s).
34	n_{prim_amino}	number of primary aliphatic amino groups	
35	n_{sec_amino}	number of secondary aliphatic amino groups	
36	n_{tert_amino}	number of tertiary aliphatic amino groups	
37	$n_{prim_sec_amino}$	$n_{prim_amino} + n_{sec_amino}$	
38	$n_{aro_hydroxy}$	number of aromatic hydroxy groups	
39	$n_{aliph_hydroxy}$	number of aliphatic hydroxy groups	
40	$n_{guanidine}$	number of guanidine groups	
41	$n_{basic_nitrogen}$	number of basic, N-containing functional groups	
42	n_{acidic_groups}	number of acidic functional groups	
43	$n_{acylsulfonamides}$	number of sulfonamide-C=O groups	
44	$n_{enolate_groups}$	number of enolate groups	
Vectorial (topological and spatial)			
45-55	2D-AC χ^{LP}	property: lone-pair electronegativity χ^{LP}	
56-66	2D-AC χ^σ	property: σ -electronegativity χ^σ	
67-77	2D-AC χ^π	property: π -electronegativity χ^π	
78-88	2D-AC q^σ	property: σ -charge q^σ	
89-99	2D-AC q^π	property: π -charge q^π	
100-110	2D-AC q_{tot}	property: total charge q_{tot}	
111-121	2D-AC α	property: polarizability α	
122-249	3D-AC $identity$	property: identity	
250	$\chi_{\sigma_1} = \sum \chi_\sigma^2$	property: σ -electronegativity χ^σ	
251	$\chi_{\pi_1} = \sum \chi_\pi^2$	property: π -electronegativity χ^π	
252	$q_{\sigma_1} = \sum q_\sigma^2$	property: σ -charge q^σ	
253	$q_{\pi_1} = \sum q_\pi^2$	property: π -charge q^π	
254-265	SurfACorr_HBP	property: hydrogen-bonding potential	

The hydrogen atoms were included before the computation of the vectorial descriptors. In more detail, each component $A(d)$ of the autocorrelation vector for the topological distance d is calculated as:

$$A(d) = \frac{1}{L} \sum_{j=i}^N \sum_{i=1}^N p_j p_i \delta(d_{t,ij}, d) \quad \delta = \begin{cases} 1 & \forall d_{t,ij} = d \\ 0 & \forall d_{t,ij} \neq d \end{cases} \quad (2.3.5)$$

where $A(d)$ represents the autocorrelation coefficient referring to atom pairs i, j ; N is the number of atoms in the molecule; p_i and p_j are the properties of atoms i and j , respectively; $d_{t,ij}$ is the i, j topological distance (i.e. the number of bonds corresponding to the shortest path in the structure diagram). A maximum distance $d = 10$ was selected, and 11 2D topological autocorrelation components per molecule were obtained.

The calculation of 3D autocorrelation vectors was performed as described

in the 2.3.2 section, by using $p_i = p_j = 1$ in eq. 2.3.3 for property identity vectors (descriptors 122-249 in Table 2.1). Default parameter values were $d_{lower} = 1 \text{ \AA}$ and $d_{upper} = 13.8 \text{ \AA}$, with a resolution of 0.1 \AA . Then, 128 components for this descriptor were obtained. The sums of the squares of the σ -electronegativity (descriptor 250 in Table 2.1), π -electronegativity (descriptor 251 in Table 2.1), σ -charge (descriptor 252 in Table 2.1), and π -charge (descriptor 253 in Table 2.1) were calculated to reflect the electronegativities and charge distributions in the aliphatic and conjugated systems. They can be obtained by calculating the first component of the autocorrelation vector while setting the distance to zero ($d_{lower} = 0 \text{ \AA}$, $d_{upper} = 0 \text{ \AA}$, and number of intervals = 1). For these descriptors, only one component ($d = 0$) of the autocorrelation coefficients has been considered. The parameters for the calculation of the hydrogen-bonding potential (descriptors 254-265 in Table 2.1) autocorrelation coefficients were as follows: $d_{lower} = 1 \text{ \AA}$, $d_{upper} = 13 \text{ \AA}$, $point\ density = 10 \text{ points/\AA}^2$, and 12 autocorrelation coefficients were obtained.

An extensive presentation of the remaining descriptors in Table 2.1 was reported in a previous work. [76]

2.4 Data autoscaling

Once the molecular descriptors have been computed, the data is ordered in a matrix form suitable to proceed with the statistical analysis. The independent x variables (molecular descriptors) should be distinguished from the dependent y variable (experimental property). However, the descriptor values might cover different intervals and show diverse distribution. Consequently, a variable with high variance might have a stronger influence than the other variables in the model development, but this effect should be avoided. Then, each variable needs to be subjected to both *scaling* and *mean-centering* procedures, in the autoscaling process, in order to return the data in an unique scale and make them homogeneous. (Figure 2.4.1). [47] The scaling of data is achieved by multiplying the elements of each variable by the corresponding standard deviation (*unit variance scaling*), while in the mean-centering the mean value of the corresponding variable data is detracted from each data. Finally, the information is contained in the same interval for all variables.

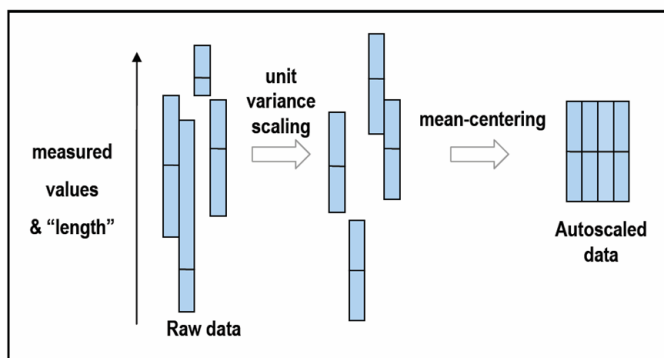


Figure 2.4.1: Scheme of the autoscaling process.

2.5 Linear strategies

2.5.1 Single and multiple regression

The linear regression represents the simplest mathematical technique to derive QSAR models, when the independent variables (molecular descriptors) are correlated with the dependent ones (experimental properties) in a linear way, as reported for example in the following relationship:

$$y = ax + c \quad (2.5.1)$$

with a single dependent y variable described as function of one independent x variable. The best straight line achieved is able to approximate data distribution with the minimum root mean squared error (RMSE) by considering the predicted and experimental y values⁴.

If further independent x variables are introduced, the calculation becomes more complex and in this case we refer to as multiple linear regression (MLR). The data is structured into a two-matrix organization and the variables are differently weighted, according to the angular coefficient of each contribute to the new regression straight line [48]:

$$y = a_1x_1 + a_2x_2 + a_3x_3 + \dots + a_ix_i + c$$

$$y = \sum a_nx_n + c \quad (2.5.2)$$

⁴The root mean square of error is defined as: $RMSE = \sqrt{\frac{\sum_{i=1}^n (y_p - y_e)^2}{n}}$ where $y_p = y_{predicted}$, $y_e = y_{experimental}$ and n is the total number of observations.

The problems related to this approach are due to the same variables, that should be independent from each other as more as possible, without errors, since they influence the model performance. Finally, they should be selected according to the relevance for the considered y variable. Moreover, five samples at least are needed for each selected descriptor in the analysis. Therefore, novel tools have been introduced to overcome these requirements in the regression-based linear QSAR strategies.

2.5.2 Principal Component Analysis

A widely used approach to reduce the dataset dimensionality, i.e. the number of independent variables, is the Principal Component Analysis (PCA). This technique allows to optimize the information contained in the data matrix, by finding out the most contributing independent variables and eliminating intercorrelations between molecular descriptors. [46-49] In fact, if two variables x_1 and x_2 are highly correlated, it would be redundant to consider both of them in the model generation. So, a new single variable is introduced to synthesize the information: the *principal component (PC)* is a linear combination of x_1 and x_2 .

If we consider a multivariate analysis with multiple descriptors, the dependent variable can be described by new variables, the *principal components*, linear combinations of the input independent variables:

$$PC_1 = a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \dots + a_{1n}x_n + b$$

$$PC_2 = a_{21}x_1 + a_{22}x_2 + a_{23}x_3 + \dots + a_{2n}x_n + b$$

$$PC_i = a_{i1}x_1 + a_{i2}x_2 + a_{i3}x_3 + \dots + a_{in}x_n + b = \sum a_{in}x_n + b \quad (2.5.3)$$

Each PC in the X space is represented by the straight line that crosses the origin and best approximates the data distribution, in order to minimize the sum of the squared distances from the straight line, i.e. when $\sum e^2$ tends to assume a value close to zero. (Figure 2.5.1). The first principal component PC_1 is the straight line that maximizes the variance in the data, then, the data shows the best distribution along PC_1 . The variance that has to be further explained, is progressively justified by the remaining principal components. Then, the second component is derived orthogonally to the first one and the following PC s are carried out in a similar way, as far as to equal

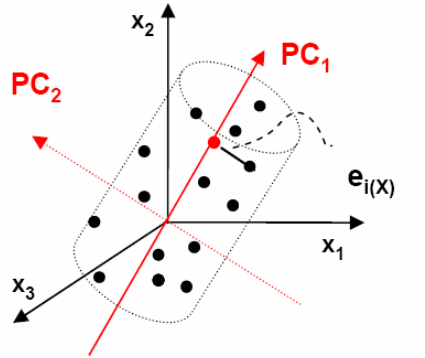


Figure 2.5.1: Data distribution and computation of PCs.

the number of the independent variables themselves. After the analysis, the data will present a large distribution of the descriptor values, with a low degree of correlation between the newly calculated variables.

Several statistical measures with the corresponding graphical representations can be considered in PCA: *variances*, *loadings*, *scores*, *residuals* and *leverage*. In Figure 2.5.2 a typical trend of the explained *variance* by increasing the number of PCs is reported. Each PC enriches the information carried by the previous one in terms of percentage (%) of variance, consequently, the last PCs are less significant than the first ones. Moreover, the decreasing slope highlighted by the graphical representation in Figure 2.5.2, allows the detection of the number of significant principal components, which defines also the model complexity.

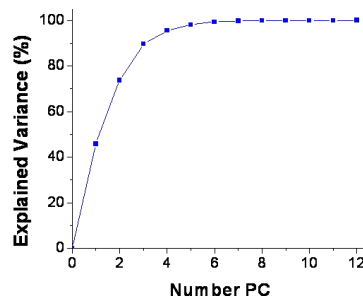


Figure 2.5.2: Graphical representation of the percentage (%) of variance progressively explained by the PCs.

In the *loading* plot the coefficients of the descriptors for each PC are shown. In particular, each descriptor is weighted according to cosine of the

angle that each variable x forms with the considered PC straight line, giving the loading value p . Furthermore, the correlation between the descriptors and PC s might be positive or negative and have a different importance: for a PC , the loading value of x variables can change both its sign and intensity with respect to the zero reference.

After the calculation of the principal components, new coordinates are identified for each sample by the orthogonal projection of each point on the considered PC . The intercepted value represents the *score*, according to the different principal components, which the projection refers to. The new coordinates, indicated as t_n , with n corresponding to the considered PC , reflect the information of the original space in the new PC_i space. The score values can be graphically reported in a 2D plane, with the axis represented by two diverse PC s (Figure 2.5.3).

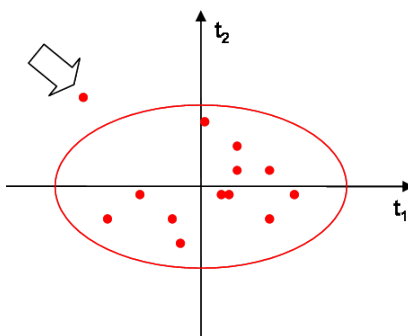


Figure 2.5.3: Score plot PC_1/PC_2 , with the outlier indicated by the arrow.

In this plot, the samples in the same quadrant present similar properties, moreover, according to the sample position in the projected space, we can evaluate which PC s are more descriptive for a particular data. The anomalous position of data in the score plot (external to the ellipse, delimiting the confidence interval, as indicated in Figure 2.5.3) might correspond to outliers⁵.

The *leverage* represents the data influence on the model: a sample with high leverage value tends to carry the analysis in a particular direction. This effect is proportional to the distance of the data from the global center. By evaluating the leverage, we are able to distinguish the presence of strong outliers.

⁵Discordant data if compared with the remaining data. Their chemical structure and molecular descriptors should be investigated before their exclusion from the model.

The detection of moderate outliers is achieved by the analysis of the variance not explained by the PC s, with the meaning of Distance to the Model in the X space or $DModX$: it indicates the *residual* (i.e. data variation non captured by the considered PC s) for each data in the X space. The data can be represented in the plane $DModX$ /number of observations, as shown in the Figure 2.5.4.

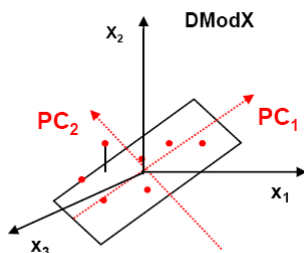


Figure 2.5.4: Geometrical interpretation of $DModX$.

Finally, the principal components are further correlated to the y dependent variable by deriving a new linear regression model, called Principal Component Regression (PCR), represented by the following equation:

$$y = aPC_1 + bPC_2 + \dots + kPC_n + d \quad (2.5.4)$$

2.5.3 Projection to Latent Structures by means of Partial Least Square

In chemometrics Projection to Latent Structure by means of Partial Least Square (PLS) analysis is able to detect liner correlations between x independent variables and y dependent variables by introducing new variables, the *principal components*, defined as latent variables, since they hide the distribution of data in the input XY space. [77] PLS technique is an extension of the PCA methodology, by considering the computation of the principal components both in X and Y spaces.

In Figure 2.5.5 a dataset constituted of X matrix, with K corresponding to the number of descriptors ($K = 3$), and Y matrix, with M referring to the number of experimental data ($M = 3$), is reported as example. The aim is to build a model for the experimental data, to find the relationship between two groups of variables. As previously described for the PCA technique, several principal components can be defined as combinations of the input

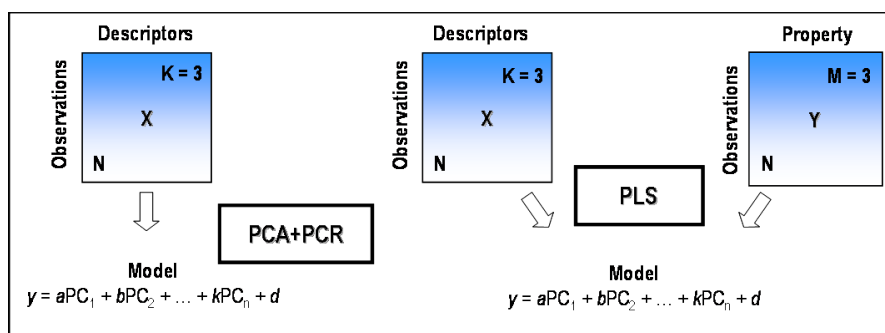


Figure 2.5.5: Meaning of the PLS analysis with respect to PCR and PCA.

variables. *PCs* are straight lines passing by the origin and orthogonal one to each other, able to approximate the data distribution in both X and Y spaces (Figure 2.5.6).

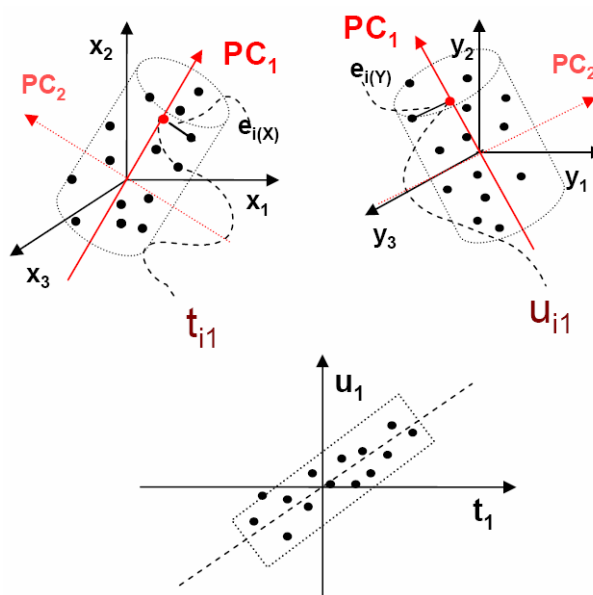


Figure 2.5.6: Calculation of PCs in PLS; t_1 and u_1 scores referred to PC_1 .

Then, for each point one can identify new coordinates, obtained by the projection of each point on the principal components: t_{i1} and u_{i1} scores. The following inner relationship between the projections is defined to correlate the data distribution in the X and Y spaces:

$$u_{i1} = mt_{i1} + h_i \quad (2.5.5)$$

where u and t are the score values of each data in the X and Y matrices, respectively, and h the residuals⁶. In this way, the information is condensed in a smaller space than the input space, by achieving the desired dimensionality reduction. The relationship between the descriptors and the y property requires the obtainment of a straight line in the plot u_1/t_1 (Figure 2.5.6); moreover, the ideal function correlating t with u is the bisector of the first and third quadrants ($m = 1$). To achieve this objective, a further step is represented by the oscillation of the straight line PC_i in the X space in a way that t values result as more as equal to u values, tending to the ideal linear correlation in the plot t/u .

If one dependent variable is analyzed, the data can be represented by using a single dimension in the Y space with u corresponding to y (Figure 2.5.7).

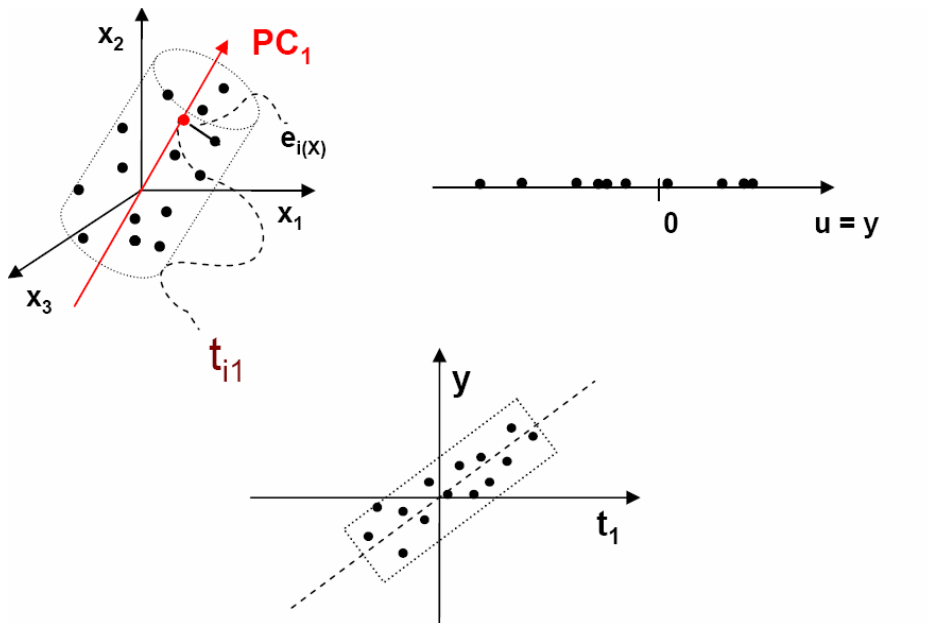


Figure 2.5.7: y and t_1 scores referred to the first PC with three independent and one dependent variables.

By introducing further PC s the percentage (%) of variance explained increases, but the contribute of each one progressively decreases, as previously described for the Principal Component Analysis.

⁶Data variation not explained by the PC s.

The considerations related to the PLS analysis are similar to the statistical definitions in the last section. In more detail, the plots corresponding to the relationship between X and Y spaces, and the data representation in the Y space can be also considered, as for example the X and Y score plots, the X and Y loadings. The score plot u/t (for the different components) reports the observations in the X (T) and Y (U) projected spaces and gives information on the correlation between Y and X spaces.

In the $y_{experimental}$ vs $y_{predicted}$ plot the prediction results are compared with the experimental data (Figure 2.5.8).

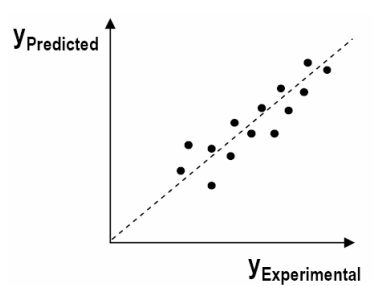


Figure 2.5.8: $y_{experimental}$ vs $y_{predicted}$ plot.

Ideally, the data should lay on a straight line with an angular coefficient equal to one and crossing the origin. We can refer to the *score* plots, the u/t score plot, the leverage, which is reported for both X and Y spaces, and to X - or Y -variance residuals/leverage plots to determine the presence of outliers⁷. Moreover, we can detect the outliers by observing the $y_{experimental}$ vs $y_{predicted}$ plot, if there is data distant from the ideal distribution. In the scatter plot the loading weights are reported, i.e. the weights of each descriptor in both X and Y spaces for the PCs , to quantify the influence of each input x and y variables in the analysis. The VIP plot (Variable Importance in the Projection) shows the coefficients of each original variable in the model for all the PCs .

The final mathematical model is able to summarize the variance in the data by introducing a number of latent variables (the new principal components) lower than the input variables (the molecular descriptors), but maintaining the input information.

⁷The outliers are data with high leverage and residual values.

2.6 Nonlinear strategies

2.6.1 Response Surface Analysis

Response Surface Analysis (RSA) is widely applied in the Design of Experiments (DOE) approach, useful to solve optimization problems in research and development. [78] A Design of Experiment is a structured methodology consisting of a sequence of experimental determinations to describe the relationship between parameters x_n involved in a process and the corresponding response variables y_n , as illustrated in Figure 2.6.1.

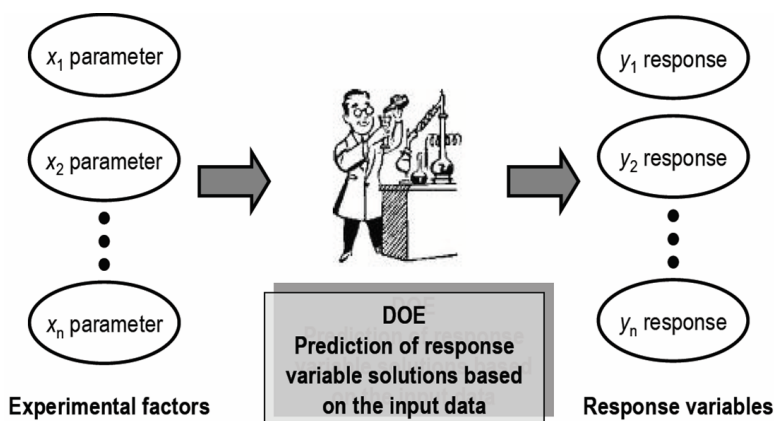


Figure 2.6.1: Simplified graphical representation of Design of Experiments (DOE) approach. x parameters (experimental factors) are related to y variables (observed responses).

The objective of DOE is to design or improve a process/product of interest. During the experiments all relevant parameters are changed systematically, in order to find the ideal conditions. For example, we can consider the study of a formulation, in which different ingredients are mixed together. After the analysis of the results (formulation parameters), the optimal conditions and the role of the input parameters (ingredients factors) to determine the experimental outcome can be identified.

Several steps are involved in the DOE approach: *a)* define the objective to the study; *b)* select the design parameters to be varied during the experiment and their intervals; *c)* identify the response variables that will be measured; *d)* perform the experiments and collect the data. Only a small set of experiments is needed to be performed for studying a process under various conditions, if the computational methods are introduced.

In particular, Response Surface Analysis (RSA) refers to a collection of mathematical and statistical techniques, applied for analyzing the influence of the independent variables (input parameters) on the response(s). [78] Figure 2.6.2 graphically represents how one can derive the generic relationships between the response variables y_1, y_2 , and the input variables x_1, x_2 . The response values, corresponding to different (x_1, x_2) pairs, yield as a surface lying above the X plane. Moreover, we can derive the best x_1, x_2 pairs corresponding to the maximum y values to achieve the optimal compromise solution for both y_1 and y_2 responses.

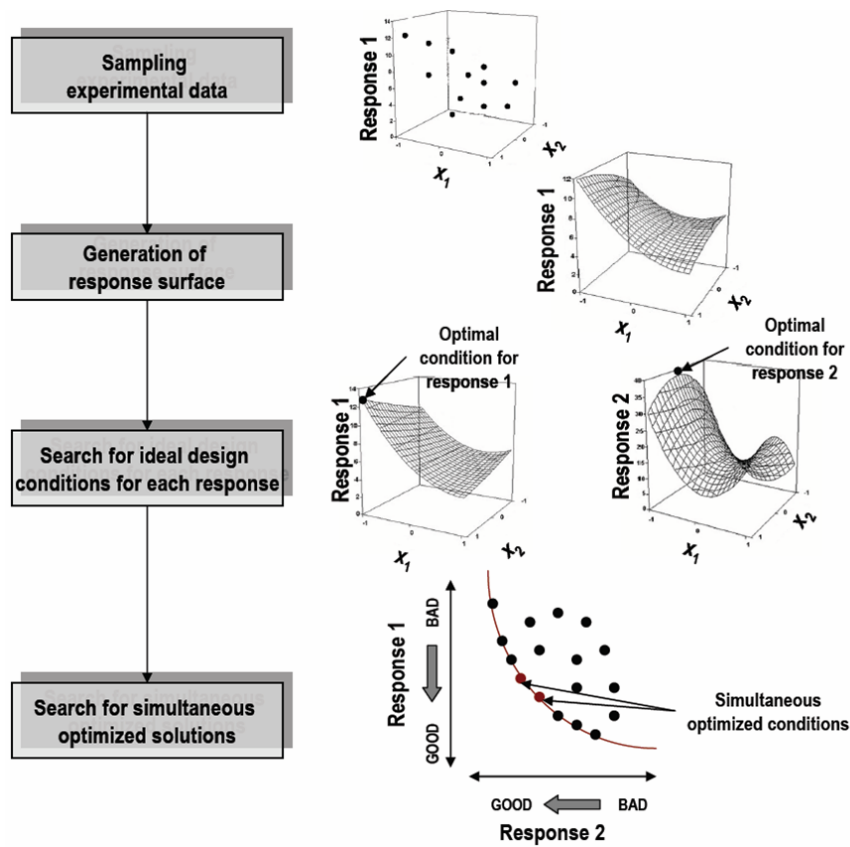


Figure 2.6.2: Flowchart of an optimum design experiment by applying Response Surface Analysis (RSA). Starting from experimental samples, each response surface is then generated for x_1, x_2 pairs. The ideal conditions are separately identified for the responses, to finally search for the best compromise solutions.

In most problems the relationship $f()$ between the response variable and the independent variables is unknown. Thus, RSA technique is aimed at

approximating the function $y = f(x_1, x_2, x_3, \dots, x_n)$, where $f(x)$ is a first-order or a second-order equation. The approximation usually employs a low-order polynomial in some region of the X space. If the response is well-modeled by a linear function of the Y variables, then the approximating function is a first-order model, as the following:

$$y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2. \quad (2.6.1)$$

If a curvature is present in the system or in the region of the optimum, a polynomial of higher degree, corresponding to a nonlinear model, should be used to approximate the response, which is analyzed to locate the optimum, i.e. the set of independent variables such that the partial derivatives of the model response with respect to the individual independent variables is equal to zero. For the second-order equation the addition of a parameter of interaction between the independent variables $x_1 x_2$ is required to introduce a curvature in the response function:

$$y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \alpha_{11} x_1^2 + \alpha_{22} x_2^2 + \alpha_{12} x_1 x_2. \quad (2.6.2)$$

Then, a more general formulation of the second-order response function is:

$$y = \alpha_0 + \sum \alpha_i x_i + \sum \alpha_{ii} x_i^2 + \sum \sum \alpha_{ji} x_j x_i. \quad (2.6.3)$$

The eventual objective of RSA is to determine the optimum operating conditions for the system, or a region which satisfies the operating specifications. Almost all RSA problems utilize one or both of these approximating polynomials. In the present thesis RSA is based on a multivariate thin plate spline algorithm derived by the Green's theorem [79]:

$$y = \sum_{i=1}^n \alpha_i g(d_i) + \sum_{j=1}^p c_j x_j \quad (2.6.4)$$

where α_i and c_j are the weight coefficients, p the number of independent variables x , n the number of data points and $g(d_i)$ the Green's function applied to the Euclidean distances between the data i and any coordinate in x axis. According to this algorithm, the response surface function is the result of an elastic beam displacement in the x_n space, where the elastic beam has to bent to reach the data points in the y space. [79]

The workflow for the model development is reported in Figure 2.6.3. Input values are regarded as points of force actions, while output values as displaced values. The surface response is the result of a smoothing procedure, that reduces the influence of the background noise carried by the input data.

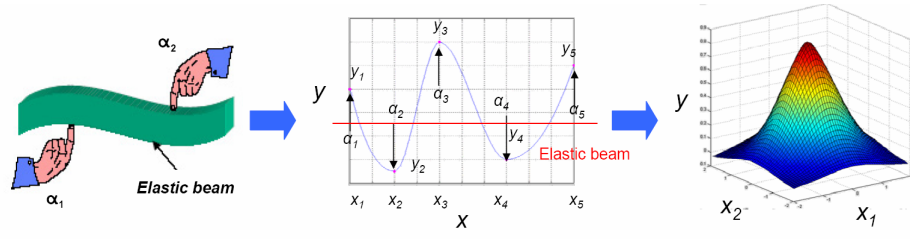


Figure 2.6.3: Schematic procedure of the thin plate spline algorithm applied in Response Surface Analysis.

In RSA technique, the selection of the most informative independent variables is performed to reduce the dimensionality of the final model and improve the model predictivity. Linear stepwise regression and nonlinear cluster analysis have been applied, and we combined the results to select the most statistically relevant independent variables. [79] The stepwise regression combines both forward selection and backward elimination processes, that progressively adds/eliminates the independent variables that are best/worst correlated to the response, respectively. The cluster analysis is a nonlinear analysis able to divide the dataset into groups, according to a similarity criterion applied to the samples. Then, similar samples belong to the same cluster, while different clusters contain diverse samples.

2.6.2 Support Vector Machine

Support Vector Machines (SVMs) are supervised learning systems originated from Statistical Learning Theory, recently developed by Vapnik, and characterized by novel attractive features and optimal generalization performance. [80, 81] The theory of SVM has been described in several books, and here we briefly introduce some principles. [82-87]

A supervised learning problem requires the resolution of a function approximation problem (approximation of an unknown response function), where the available data set (training set) is represented as a set of pairs (examples), $T = \{(x_1, y_1), (x_2, y_2), \dots, (x_i, y_i)\}$, where x_i is an input data and y_i is the corresponding observed response value. Usually, $x_i \in \mathbb{R}^n$,

while if $y_i \in \{-1, +1\}$ the learning problem is a binary classification, if $y_i \in \mathbb{R}$ the learning problem is a regression. In both cases, the aim of the learning system is to select a hypothesis $f(x)$ that approximates the desired response y_i in an optimal fashion, i.e. by minimizing some risk functional R . In particular, we would like the function $f(x)$ to be a reasonable estimate of the functional relation between input-output pairs (prediction or generalization property). R weights the cost of the approximation, while the error of a hypothesis is given by a loss function $L(\cdot)$ that measures the distance between y_i and $f(x)$. A common example of loss is given by the quadratic error function:

$$L(f(x_i), y_i) = (y_i - f(x_i))^2. \quad (2.6.5)$$

The average error over the training set is the *expected risk* R . If we assume that a probability distribution $P(x_i, y_i)$ exists and it is known to govern the data and the underlying function dependences, R can be expressed as

$$R = \int L(f(x_i), y_i) dP(x_i, y_i). \quad (2.6.6)$$

The learning process aim at selecting the hypothesis $f^{pt}(x)$ to minimize R .

The only available information to the learning system is the training set; how the learning system uses the training set in order to minimize R is called the *inductive principle*, that is a general prescription for obtaining the estimate of $f^{pt}(x)$ from the training set. Given an inductive principle, the learning algorithm tells how to use the data to obtain the estimate. One of the most popular inductive principles is Empirical Risk Minimization, which search for $f^{pt}(x)$ by minimizing the empirical error:

$$R_{emp} = \frac{1}{n} \sum_{i=1}^n L(f(x_i), y_i). \quad (2.6.7)$$

The linear SVM is based on: *a*) linear hypotheses corresponding to separating hyperplanes in the \mathbb{R}^n space, i.e. $f(x) = w \cdot x + b = \sum_{i=1}^n w_i x_i + b$ where \cdot is the dot product between vectors; *b*) the solution of a quadratic optimization problem that represents a trade-off between the minimization of the empirical error, i.e. the error over the training set, and the maximization of the smoothness of $f(x)$. [80] Nonlinear versions of SVM can be obtained by the introduction of a kernel. [88]

We will discuss these issues both for classification and regression.

Support Vector Classification

In the last years, several classification problems have been solved by using SVM approach, such as the discrimination between active and non active compounds. [89-99] Binary classification is widely performed to discriminate a set of examples in two classes. Different formulations for SVM are possible according to the loss function $L(f(x_i), y_j)$ used. We adopt the standard formulation derived by using the Hinge loss function and slack variables ξ_i :

$$\min_{w,b} = \frac{1}{2} \|w\|^2 + C \sum_{i=1}^n \xi_i \quad (2.6.8)$$

subject to: $\forall i \in \{1, \dots, n\} \quad y_i(w \cdot x + b) \geq 1 - \xi_i \quad \text{and} \quad \xi_i \geq 0,$

where we recall that $y_i \in \{-1, +1\}$, w and b are the parameters that control the function $f(x)$, and the constraints are satisfied with zero error when it is possible to find a function able to classify any positive example ($y_i = +1$) by returning a positive value that has some margin from zero, i.e. $f(x) \geq 1$, and any negative example ($y_i = -1$) returning a negative value that has some margin from zero, i.e. $f(x) \leq -1$. If such function does not exist, then errors need to be compensated by choosing non zero values for the corresponding slack variables ξ_i . The geometrical interpretation of Support Vector Classification is shown in Figure 2.6.4.

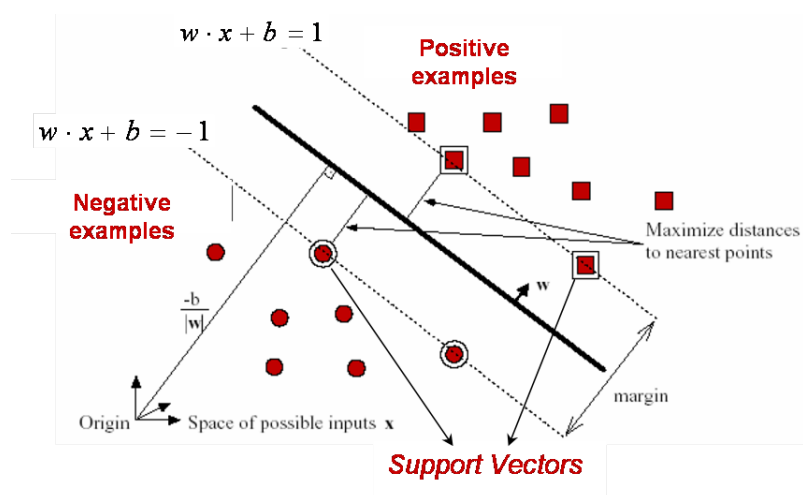


Figure 2.6.4: A binary classification problem. The optimal separating hyperplane is orthogonal to the shortest line connecting the convex hulls of the two classes, and intersects it half-way between the two classes.

The trade-off between the minimization of the norm of the weight vector and the empirical error is given by the constant C . The above quadratic constrained minimization problem (eq. 2.6.8) can be more easily solved by resorting to the corresponding dual problem:

$$\max_{\alpha} = \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{j=1}^n \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) \quad (2.6.9)$$

$$\text{subject to: } \sum_{i=1}^n y_i \alpha_i = 0, \text{ and } \forall i \in \{1, \dots, n\} \quad 0 \leq \alpha_i \leq C.$$

The optimal weight vector w^o of the first formulation is linked to the optimal solution vector α^o of the dual problem (α_i are called dual variables) by the following relation:

$$w^o = \sum_{i=1}^n \alpha_i^o y_i x_i. \quad (2.6.10)$$

The input vectors x_i for which the corresponding dual variables satisfy $\alpha_i^o > 0$ are referred to as support vectors. Finally, the decision rule is given by $\text{sgn}(f(x))$. The characteristic nonlinearity of the boundary separating positive from negative samples is achieved by projecting the input vectors into a higher dimensional feature space, i.e. $x \mapsto \Phi(x)$ (Figure 2.6.5).

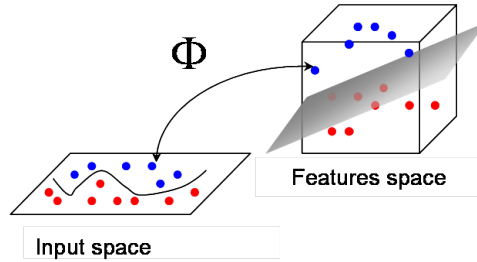


Figure 2.6.5: Transformation operated by the kernel.

In this way the dot product $x_i \cdot x_j$ is replaced by a kernel function $k(x_i, x_j)$ representing the dot product in the transformed space, i.e. $k(x_i, x_j) = \Phi(x_i) \Phi(x_j)$. The decision function takes the final form:

$$f(x) = \text{sgn}\left(\sum_{i=1}^n \alpha_i y_i k(x_i, x) + b\right). \quad (2.6.11)$$

An example of kernel function is the gaussian RBF (radial basis function), $k(x_i, x_j) = e^{-\gamma(x_i - x_j)^2}$, which has demonstrated its good performances by producing a closed decision boundary. Regarding this kernel, the γ parameter should be appropriately selected: very small values of γ correspond to complex models with a high number of support vectors and risk of overfitting, while large values of γ might lead to a separating hyperplane described with few support vectors and too smooth for an accurate classification.

Support Vector Regression

In the last decade Support Vector Regression (SVR) has been used as a non-linear methodology to derive quantitative structure-activity relationships for the prediction of different chemical and biological properties. [100-106]

As anticipated, in a regression problem $y_i \in \mathbb{R}^n$, then the mathematical formulation has to consider the approximation errors. A "reasonable" approximation is defined by introducing the constraint that for each input x_i we should have $|y_i - f(x_i)| \leq \varepsilon$, where ε is a small positive constant representing the tolerance we allow on approximation errors. This requirement can be described by two linear constraints, i.e. $(y_i - w \cdot x_i - b) \leq \varepsilon$ and $(w \cdot x_i + b - y_i) \leq \varepsilon$. Errors above the tolerance are typically linearly penalized by resorting to the linear ε -insensitive loss function, in the following:

$$L^\varepsilon(x, y, f) = |y - f(x)|_\varepsilon = \max(0, |y - f(x)| - \varepsilon) \quad (2.6.12)$$

Based on the above considerations, the standard SVR model is defined as:

$$\min_{w, b, \xi_i, \xi_i^*} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^n (\xi_i + \xi_i^*), \quad (2.6.13)$$

subject to: $\forall i \in \{1, \dots, n\}$

$$(\langle w \cdot x_i \rangle + b) - y_i \leq \varepsilon + \xi_i,$$

$$y_i - (\langle w \cdot x_i \rangle + b) \leq \varepsilon + \xi_i^*,$$

$$\xi_i, \xi_i^* \geq 0,$$

where a set of slack variables ξ_i, ξ_i^* is added to quantify the violation of the imposed constraints: ξ_i accounts for the underestimation of the target values, while ξ_i^* accounts for the overestimation of the target values.

The solution of a linear regression problem results to be a tube with radius ε which approximates the data distribution (Figure 2.6.6). [80, 81, 88, 107] Even for SVR a kernel can be used to introduce nonlinearity. Then, the kernel expansion of the decision function f is:

$$f(x, \alpha_i^*, \alpha_i) = \sum_{i=1}^n (\alpha_i^* - \alpha_i) k(x_i, x) + b. \quad (2.6.14)$$

The final hypothesis regression function is a weighted sum of the kernel function evaluated at the support vectors, defined as the training points located on the border of the regression tube (Figure 2.6.6).

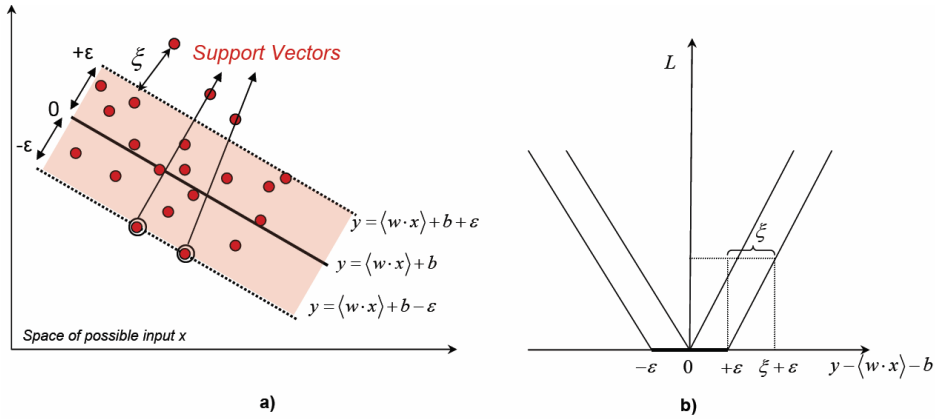


Figure 2.6.6: In Support Vector Regression a tube with radius ε is fitted to the data. The trade-off between model complexity and points lying outside of the tube (with positive slack variables ξ) is determined by minimizing eq. 2.6.13.

2.6.3 Cross-training with Support Vector Machine

The concept of cross-training has been introduced by Boutell and collaborators. [108] They turned from the previously less performing attempted strategies for multilabel data to present a new classification method suitable for multiple and overlapping classes tasks, with samples simultaneously associated to more than one class (Figure 2.6.7).

In the cross-training approach the multilabel data are used more than once when training the classification model. Moreover, each sample is assigned a positive label for each actual class to which it belongs. [108] Cross-training with Support Vector Machine (ct-SVM) represents a novel applica-

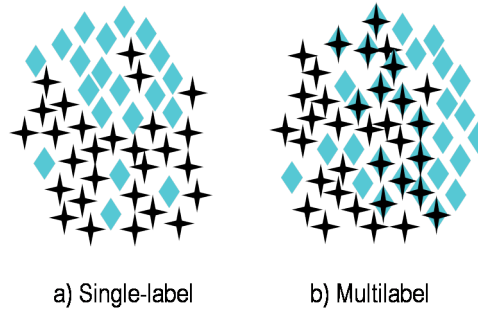


Figure 2.6.7: a) *Single-label classification: the samples that belong to two different classes are often difficult to separate;* b) *multilabel classification: the data marked with both symbols belongs simultaneously to both classes.*

tion of SVM analysis, when classes overlap in the feature space. In ct-SVM technique the output real-valued scores of the trained binary classifiers for each class are transformed into the final labels according to different testing criteria, as previously reported. [109] In more detail, n binary classifiers, n corresponding to the number of the considered target responses, were built using a radial basis function (RBF) kernel. The parameters for each SVM classifier (C and γ) were automatically optimized on the training set during the learning process by using a 10-fold cross-validation and predicting a small validation set. [108] After applying the cross-training approach, the real-valued scores were obtained.

Recently, three different testing criteria (P, T and C) have been proposed. [108] The P-criterion assigns to the samples all labels corresponding to a positive SVM score. If none of the scores is positive, the sample is classified as "unknown". The T-criterion uses the Closed World Assumption (CWA), according to which all samples belong to at least one class: if all SVM scores for a particular sample are negative, the pattern is assigned to the class corresponding to the less negative score. The C-criterion considers SVM scores without any sign and the decision depends on the closeness between the top SVM scores. In our studies a validation set has been used to select the closeness between two scores. Once each binary classifier was optimized by predicting a validation set, the training and the validation sets were merged and the new model was computed by using the previously optimized parameters. The final model was applied in the prediction of a test set in order to evaluate its statistical robustness.

2.6.4 Artificial Neural Networks

Artificial Neural Networks (ANN) have been introduced as a more flexible class of modeling techniques naturally able to deal with complex nonlinear systems both in classification and regression problems. Their architecture is particularly suitable in the studies with a large number of observations. The innovative potentialities and applicability of the neural network methodology in drug discovery have been recently described. [110]

The neural networks algorithm is able to model the functionality of the brain. In the biological neuron, the dendrites are fibers connecting each neuron to the neighboring neurons (Figure 2.6.8a). [46] A neuron receives the new information from the neighboring neurons and converts it to the final single signal in the soma. The signal, if strong enough (higher than a particular threshold value), is transmitted to the axon. Then, the axon carries the information further to the other dendrites and the transmission takes place at the level of the synapse. The aim of the learning process is to build the synapse strength.

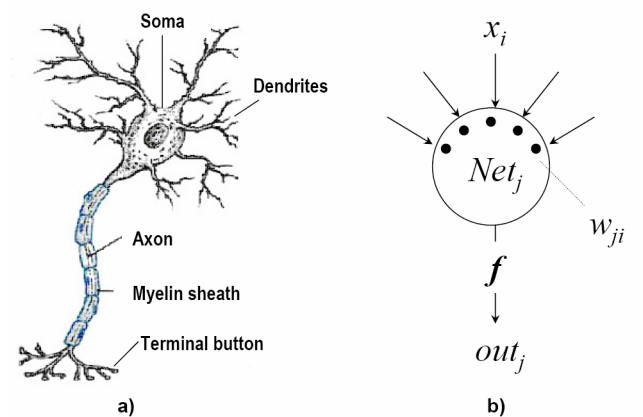


Figure 2.6.8: a) Schematic structure of a biological neuron; b) unit i of an artificial neuron.

Artificial neural network is an interconnected feed-forward network modeling consisting of interrelated neurons, which interchange signals. The feed-forward neural network depends on the input layer, where the units receive the input data from the previous layer, so the information is processed unidirectionally. It is formed by neurons (units) that process the information and cooperate, simulating the connections of biological neurons. [46, 111]

In fact, an artificial neuron (*unit*) behaves similarly to the biological

neuron (Figure 2.6.8b): the neuron j receives the input signals x_i , that are multiplied by w_{ji} (weights) and summed, obtaining the global signal Net_j , as in the following equation:

$$Net_j = \sum_i w_{ji}x_i \quad (2.6.15)$$

where w_{ji} are the weights codified by a vector, and establish the connection strength. The final signal Net_j is filtered and modified by a transfer function, deciding whether the signal can be transmitted as out_j to other neurons. The most common activation functions are linear, such as the sigmoidal transfer function: it can assume zero or one values as indicated by the relation:

$$out_j = \frac{1}{1 + e^{-(\alpha Net_j + \vartheta)}}. \quad (2.6.16)$$

In ANN the units form a network and the network is structured in different interconnected layers: input layer, one or more hidden layers, with not accessible output, and output layer. [46] The neurons in the same layer receive the signal from the layer above and simultaneously produce a unique set of outputs. A typical network is shown in Figure 2.6.9.

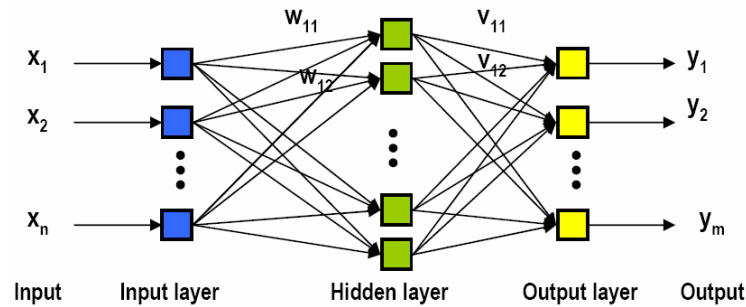


Figure 2.6.9: Artificial neural network architecture, comprising input, hidden and output layers.

Two different learning approaches are available: *unsupervised* and *supervised*. In the *unsupervised* learning the network classifies the input vectors, according to their similarity and the property to be analyzed is not used in the training process. Consequently, the data in the same neuron or in topologically adjacent neurons are similar, since similar data tends to form clusters. An example of unsupervised neural network is the Kohonen net-

work. [46, 111] The *supervised* learning is aimed at assigning the response signal to the input data. In the learning process the input data are sent to the input layer, then to the nodes of the hidden layer, and finally a response is elaborated in the output layer. In more detail, the input vectors are introduced to characterize the objects, and the output vectors correspond to the property of the object to investigate.

During the learning process (training) the neural networks learn from examples, so the connection between neurons are adapted, i.e the weights are adjusted. The input data x_i enters the network system to generate the output out_i , as previously described, and out_i is compared with the target value, yielding the error δ . Then, the weights are corrected to reduce δ during more learning cycles (epochs), in which input data are processed in the network, as summarized in Figure 2.6.10. The supervised learning is applied, for example, in counter-propagation neural networks, where the error is back-propagated from the output layer to the previous hidden layers.

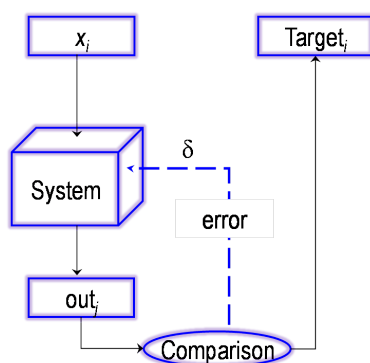


Figure 2.6.10: Schematic representation of the supervised learning.

The Kohonen Network or self organizing map (SOM) is an example of recurrent associative neural network, where each layer feeds the input units back into its own units and the information is transmitted dynamically. It is able to project the samples from a multidimensional space into a two-dimensional plane, retaining the input information (Figure 2.6.11)⁸. [46] In the network architecture each column in this two-dimensional system represents a neuron; each cuboide in a column corresponds to a dimension of the input data (molecular descriptor) and it is also associated to a weight. At the beginning of the training process the weights are random numbers;

⁸Adapted from Gasteiger, J.; Engel, T. *Chemoinformatics*, Wiley-VHC, 2003.

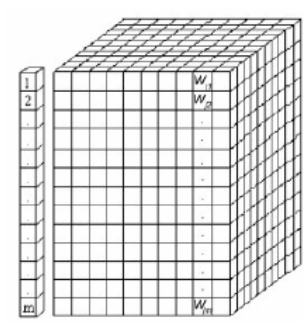


Figure 2.6.11: Structure of the Kohonen network.

when an input vector enters the network, the weights more similar to that neuron are determined according to the Euclidean distance $\sum_{i=1}^m (x_s - w_{ji})^2$ calculated between the input vector x_s and the weights w_j . The winning neuron is associated to the minimum Euclidean distance, according to:

$$\min \|X(t) - W_j\| = \min \left\{ \sum_{i=1}^m [x_i(t) - w_{ji}]^2 \right\} \quad (2.6.17)$$

where the index j refers to a particular neuron, n is the total number of neurons, W_j are the weight vectors and $X(t)$ are the input vectors. Based on their distance from the winning neuron, the weights of the other neurons are adapted. The process is repeated for all remaining input data, until a training epoch is completed.

We have applied the supervised learning in a classification task by using the counter-propagation neural network methodology. A counter-propagation neural network (CPG NN) is a well-known extension of Kohonen self organizing maps analysis, where some output layers (output block), corresponding to the classes (Y variables), are added to the Kohonen input layers, that represent the molecular descriptors (X variables) (Figure 2.6.12)⁹. Figure 2.6.12 refers to a classification problem with four classes; in the output layers, each vector component is one for positive examples or zero for negative examples, according to the assigned classes. During the learning process only the input layers are considered to determine the winning neuron. Then all the weights, including the output layers, are adapted and the trained network can be used to predict unknown property vectors. Different topologies (rectangular and toroidal) and map sizes were used in the CPG NN analysis.

⁹Adapted from Gasteiger, J.; Engel, T. *Chemoinformatics*, Wiley-VHC, 2003.

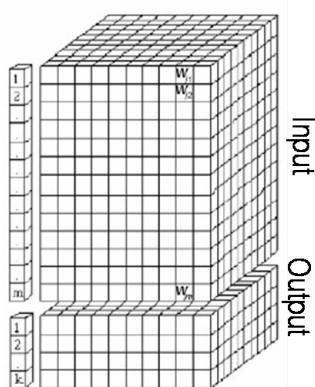


Figure 2.6.12: Counter-propagation neural network; the input and the output layers with the corresponding weights are indicated.

2.7 Validation and statistical evaluation

For regulatory purposes some reference principles, introduced by Organization of Economic Cooperation and Development (OECD), have been recommended in the QSAR model development. [112] The guideline document has underlined the following requirements: *a)* a defined endpoint, *b)* an unambiguous algorithm, *c)* a defined domain of applicability, *d)* appropriate measures of goodness-of-fit, robustness and predictivity and *e)* a mechanistic interpretation, if possible. Some of these principles and their role in the regulatory context have been clarified. [113]

We have used several methods to evaluate the statistical reliability of our models. Their predictive power was verified by performing a validation procedure: the *internal* validation was applied by excluding the samples composing the training set (*cross-validation*), the *external* validation has considered the prediction of the samples not used in the model generation (test set prediction). In particular, LOO (*leave-one-out*), 10-, 5-, 3- and 2-fold cross-validation procedures were performed.

In *n*-fold cross-validation procedure, the data set is divided randomly into *n* subsets with a similar number of samples and class distribution (in the classification approach), according to a stratified methodology. In the first step, one partition *n* is considered as test set, while the others (*n*-1) partitions are used to fit the model, used then to predict the test set. The process is repeated *n* times, until all the partitions are considered as test set.

Concerning the regression models, the correlation coefficient r is calculated to evaluate the quality of the fitting process (model calibration), as follows:

$$r = \frac{\sum_{i=1}^n (x_{Exp} - \bar{x})(y_{Pred} - \bar{y})}{\sqrt{\sum_{i=1}^n (x_{Exp} - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_{Pred} - \bar{y})^2}} \quad (2.7.1)$$

where x_{Exp} represents the experimental data, y_{Pred} the predicted values and \bar{x} , \bar{y} the corresponding averages. The values of the correlation coefficient are included in the interval between zero and one; r indicates the ideal correlation when it is equal to one. The same parameter has been calculated after LOO cross-validation procedure (r_{cv}). The following statistical requirements should be satisfied to achieve a good modeling performance:

$$r_{cv}^2 > 0.3 \text{ and } r^2/r_{cv}^2 \sim 1$$

A further validation procedure in the regression analysis is the y variable randomization, which should give bad performing models.

In the single-label classification approach, following the OECD principles, we have applied an extensive n -fold cross-validation to evaluate the predictivity of our models. The average, standard deviation, minimum and maximum rates were collected for each n -fold cross-validation method. Moreover, the *confusion matrix* is required for the model evaluation (Figure 2.7.1).

		Experimental	
		+	-
Predicted	+	TP	FP
	-	FN	TN

Figure 2.7.1: *Confusion matrix.* The samples classified by the model (rows) and the experimental classes (columns) are reported. TP (true positives) are correct positive predictions, FP (false positives) are incorrect positive predictions, FN (false negatives) are uncorrect negative predictions and TN (true negatives) are correct negative predictions.

Then, the true positive (TP) rate, the false positive (FP) rate, the true negative (TN) rate, the false negative (FN) rate, *accuracy*, *recall*, or *sensitivity*, and *precision*, or *specificity*, were calculated for each binary classifier from the confusion matrix, as summarized in Table 2.2.

Table 2.2: Statistical parameters to evaluate the classification models.

Name	Calculation details
True positive rate	TP/(FN+TP)
False positive rate	FP/(TN+FP)
True negative rate	TN/(TN+FP)
False negative rate	FN/(FN+TP)
Recall	TP rate
Precision	TP/(FP+TP)
% correct predictions (accuracy)	(TP+TN)/Total number of compounds · 100
Matthews correlation coefficient	$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$
	TP = number of true positives; FP = number of false positives;
	TN = number of true negatives; FN = number of false negatives

The values of the rates are included in the interval between zero and one. Low values of FP and FN rates, high values of TP and TN rates, high percentage (%) values of correct predictions, recall and precision correspond to good modeling performances.

Moreover, we calculated the *Matthews correlation coefficient* (MCC), that falls in the range $-1 \leq MCC \leq 1$. A value of $MCC = 1$ indicates perfect agreement between predicted and experimental classes for each binary classifier, whereas a value $MCC = -1$ indicates the worst possible prediction.

The evaluation of a multilabel classification model performance is more complicated in comparison with the statistical quality of a single-label classification model. The confusion matrix was extracted from the predictions of the validation set and the internal test set to assess the robustness of our models. In this case, the accuracy is referred to the overall performance on the tested data set, while the TPs, FPs, TNs, FNs, the recall and precision are base-class measures, calculated for each class after the comparison between actual and predicted labels by our multilabel models. [108, 109] In ct-SVM analysis, the ranking process provides a function to order the labels for each sample and to assign scores to the samples. Several ranking-based performance measures have been mathematically defined. [108, 109] *One-error* represents the ratio of the number of not top-ranked labels to the total number of actual labels. It can take on values between zero and one and values close to zero indicate a good performance. *Coverage* measures how far one needs, on average, to go down the list of labels to cover all actual

labels. The coverage interval is between one and the number of the classes; then, the best performance corresponds to a value of zero. *Average precision*, that refers to the whole system, reflects the effectiveness of the label ranking and indicates the frequency of the top-ranking for the actual labels. The extreme values are zero and one and the best performance is achieved when the average precision is equal to one. [108]

2.8 Software

Most modeling studies were carried out on a 16 CPU (Intel Core™2 Quad CPU 2.40 GHz) linux cluster running under openMosix architecture (*Paper I, II, IV-VI*). [114]

Molecular structure building and *autocorrelation* molecular electrostatic potential (*autoMEP*) descriptors, based on Connolly's solvent accessible surfaces, have been carried out using ADRIANA (version 2.0) [115] and ADRIANA.*Code* suite (version 2.2) (*Papers I-VI*). [51] The number of hydrogen bonding donors and acceptors, Topological Polar Surface Area (TPSA) descriptors have been carried out using ADRIANA.*Code* software (version 2.2) (*Papers III, VI*). [51] Sterimol descriptors, logP(o/w), Approximate Surface Area (ASA), HOMO and LUMO energy descriptors have been calculated using Molecular Operating Environment (MOE, ver. 2008.10) (*Paper VI*). [116] The remaining molecular descriptors in Table 2.1 have been calculated by using ADRIANA.*Code* software (*Paper III*). [51]

Partial Least Square (PLS) analysis has been performed using "The Unscrambler" statistical software (*Paper I*). [117]

Response Surface Analysis (RSA) has been performed using DataFOREST and DataNESIA softwares (*Paper I, II*). [118, 119]

Some SVM classification analysis and Support Vector Regression models have been performed by using SVM^{light} software (*Paper IV*). [120]

Most single-label classification models were built using Weka data mining software (*Papers III, VI*). [121]

Cross-training with SVM (ct-SVM) multilabel classification models were generated with the R software and package e1071 (*Papers III, V*). [122, 123]

The counter-propagation neural network (CPG NN) analysis was performed using SONNIA software (*Paper III*). [124]

Estimation of the aqueous solvation free energy

Several quantitative structure-property relationship (QSPR) approaches have been explored for the prediction of aqueous solubility or aqueous solvation free energy, ΔG_{hyd} , as crucial parameter affecting the pharmacokinetic profile and toxicity of chemical compounds. It is mostly accepted that aqueous solvation free energies can be expressed quantitatively in terms of properties of the molecular surface electrostatic potentials of the solutes. In the present study we have introduced *autocorrelation* molecular electrostatic potential (*autoMEP*) vectors in combination with nonlinear Response Surface Analysis (RSA) as alternative 3D-QSPR strategy to evaluate the aqueous solvation free energy of organic compounds. A robust QSPR model ($r_{cv} = 0.93$) has been obtained by using a collection of 248 organic chemicals. An external test set based on 23 molecules confirmed the good predictivity of the *autoMEP*/RSA model suggesting its further applicability in the *in silico* prediction of water solubility of large organic compounds libraries.

3.1 Introduction

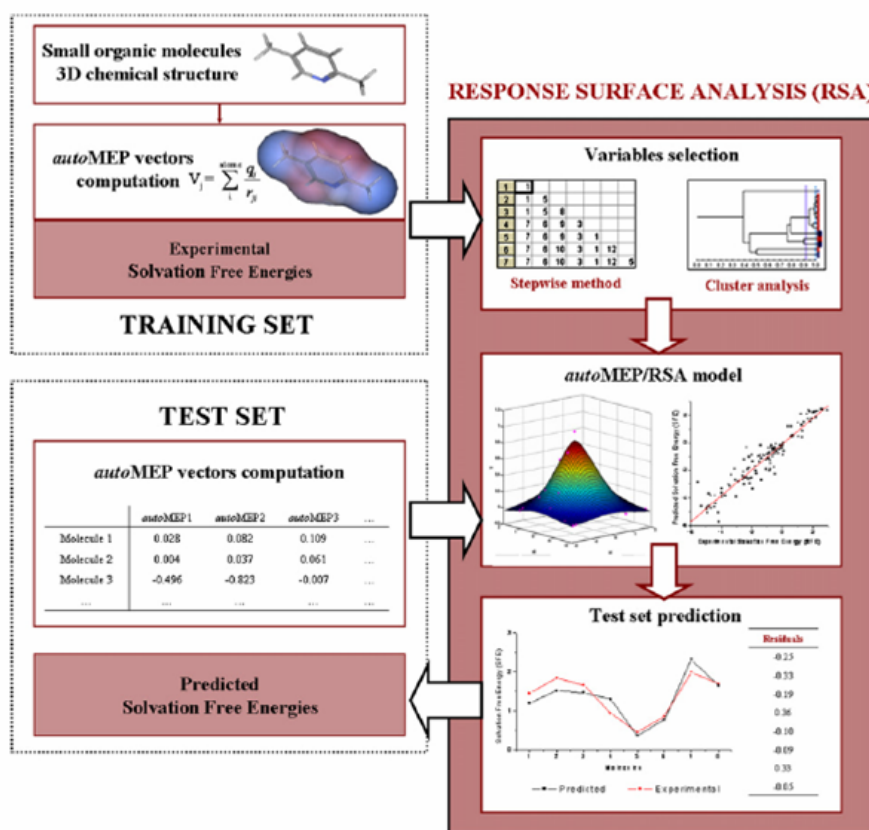
In the last decades a wide interest has been focused on the prediction of aqueous solubility as relevant property affecting the pharmacokinetic profile and toxicity of chemical compounds. [125] Especially in the early phase of drug discovery, molecular solubility represents an important determinant of drug-likeness, since it relates, for example, to drug bioavailability. [61] In fact, solvation effects play one of the major contributes influencing the quantity of free drug for biological processes, and increased solubility might correspond to improved therapeutic effectiveness of potential new drugs.

From a thermodynamic point of view, the solvation free energy describes the effects of the solvent on the solute, when the solute is transferred in solution phase at constant temperature and it is surrounded by the solvent molecules. The solvation process is energetically favoured if the new interactions between the molecules in solution lead to a more stable thermodynamic system with respect to not interacting solvent and solute. Then, an efficient solvation process is due to favourable interactions between the solute and the water (aqueous solvation free energies, ΔG_{hyd}). Moreover, the formation of a receptor-ligand complex requires a trade-off between an unfavourable electrostatic desolvation penalty, occurring when the ligand binds the receptor in aqueous solution, and the generally favourable intermolecular interactions involved in the complex. [126] So, the solvation effects are responsible for the most probable binding mode of a receptor-ligand complex as well as for the binding affinity of organic compounds.

Well performing computer simulations of the solute-solvent system are commonly applied to calculate the free energy of solvation, but their high computational and time demands are not consistent with the study of a large number of compounds. However, a classical quantitative structure-property relationship (QSPR) approach is suitable for the evaluation of any solute-solvent system, as previously reported in several papers. [127-131] Unfortunately, the datasets that were used for the generation of these models are restricted to organic compounds.

Recently, it was reported that autocorrelation Molecular Electrostatic Potential (*auto*MEP) vectors in combination to Partial Least Square (PLS) and/or Response Surface Analysis (RSA) techniques can represent a powerful three-dimensional quantitative structure-activity relationship (3D-QSAR) approach. [60, 132-135] In fact, topological and electrostatic complemen-

tarities are considered two key concepts in molecular recognition processes. Gasteiger and collaborators investigated the MEP on the molecular surface as particularly useful method for rationalizing the interactions between molecules and molecular recognition processes. [52, 57, 58] The electrostatic forces are a fundamental component of the interactions between the solute and the solvent. Moreover, the major contribution to the solvation free energy of the solute is represented by the surface of the solute that is accessible to the solvent, and by the screening effect of the solvent. Therefore, MEP distribution on the molecular surface can be used as parameter to describe aqueous solvation/desolvation processes. To this aim, we have introduced *auto*MEP vectors in combination with Response Surface Analysis (RSA) as alternative 3D-QSPR strategy for the estimation of the aqueous solvation free energy of organic compounds, as shown in Figure 3.1.1.



3.2 Results and discussion

The solvation free energy is a thermodynamic parameter to describe the effects of the solvent. [136] The solvation effect is globally due to intermolecular interactions between the solute and the solvent, as well as a change in the intramolecular interactions of the solute and, a reorganization of the solvent because of the solute. Among these phenomena, the electrostatics represent the main contribution in the abovementioned interactions.

As anticipated, *auto*MEP descriptors encode into autocorrelation vectors the three-dimensional spatial distribution and the intensity of the electrostatic potential projected on the molecular surface. Then, we have used *auto*MEP descriptors in combination with a response surface analysis technique (*auto*MEP/RSA) to predict the solvation free energy of a set of 248 organic chemicals (training set). This training set is a collection of small organic molecules, that belong to different chemical classes (Table 3.1).

Table 3.1: *Frequency of functional groups in the training set.*

No.	Functionalities
14	Alkanes
15	Alkenes
7	Alkines
57	Halogen derivatives
22	Aromatics and cycles
20	Aromatics and N containing compounds
7	Nitro derivatives and nitriles
10	Amines
30	Alcohols
7	Ketones
12	Aldehydes
16	Ethers
26	Esters
5	S containing compounds

The parameters for the calculation of autocorrelation coefficients are the following: $d_{lower} = 0$; $d_{upper} = 5$; $L = 12$; *point density* = 20 points/Å², according to eq. 2.3.4. The preliminary application of the stepwise regression and the cluster analysis on the original twelve *auto*MEP descriptors led to the selection of five independent variables into RSA model: *auto*MEP 1, 7, 8, 10 and 12. The calibration step, performed as described in 2.6.1 section,

has provided a very high correlation coefficient ($r = 0.99$), confirming the good choice of the independent variables, as summarized in Figure 3.2.1 and Table 3.2.

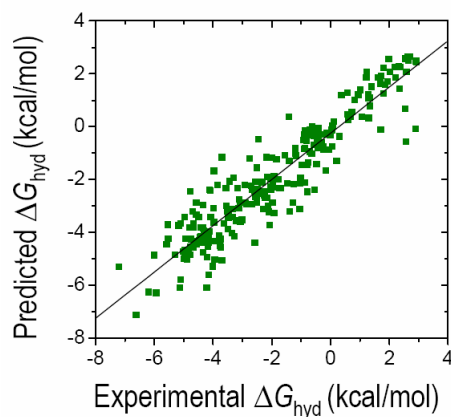


Figure 3.2.1: *AutoMEP/RSA model; experimental ΔG_{hyd} values vs predicted ΔG_{hyd} values after LOO cross-validation on the training set.*

Table 3.2: *Summary of the statistical parameters of autoMEP/RSA model.*

Number of molecules	248
X variables	5
r	0.99
r_{cv}^a	0.93
Slope	0.87
Offset	-0.25
q^b	0.92
RMR ^c	0.069
RSS ^d	1.19

^aCross-validated r after LOO cross-validation procedure: $r_{cv} = [SXY / (SXX)^{\frac{1}{2}} (SYY)^{\frac{1}{2}}]$, $SXY = \sum (X - X_{mean})(Y - Y_{mean})$, $SXX = \sum (X - X_{mean})^2$ and $SYY = \sum (Y - Y_{mean})^2$ with $X = Y_{Experimental}$ and $Y = Y_{Predicted}$; ^b r of the internal test set; ^croot mean square of residuals: RMR; ^dresidual sum of squares: RSS.

A LOO cross-validation technique has been applied for validating the final *autoMEP/RSA* model to statistically confirm its robustness ($r_{cv} = 0.93$). Interestingly, *autoMEPs* 1 and 7 seem to play a major role in describing the complexity of the final response surface. A representation of solvation free energy as function of *autoMEP* 1 and *autoMEP* 7 is shown in *Paper I*.

By analyzing Table 3.1, the predictivity of *autoMEP/RSA* model does not present any particular dependence from the chemical structure of the

considered organic compounds. The residuals of 248 derivatives of the training set overcome 1 kcal/mol, and it happens especially whether chloride and fluorine atoms are present or for some aliphatic and aromatic alcohols and aromatic amines, as reported in *Paper I*. In most cases the solvation free energy of halogen derivatives is overestimated, while alcohols and aromatic amines are generally underestimated, if compared to the respective experimental values. However, it is interesting to note that as all other 3D-QSPR approaches, also *autoMEP/RSA* model is able to discriminate among stereoisomers, improving the limits of some models that have utilized, for example, atomic constants as molecular descriptors (*Paper I*). [128]

A test set of 23 molecules with a different chemical structure and solvation free energy values has been selected to further validate our *autoMEP/RSA* model. The experimental vs predicted solvation free energies values are collected in Figure 3.2.2 and Table 3.3.

The predicted and experimental solvation free energy values are very similar. A very good correlation coefficient calculated on the test set ($q = 0.92$) is an additional evidence about the good predictivity of the *autoMEP/RSA* model, as reported in Table 3.2 (*Paper I*). The predicted solvation free energies result very close to the experimental values, as shown in Figure 3.2.2.

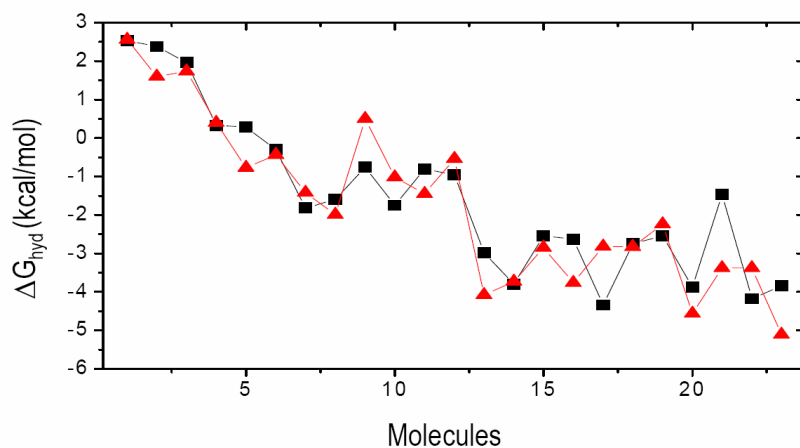


Figure 3.2.2: Comparison between experimental (\blacktriangle) and predicted (\blacksquare) ΔG_{hyd} values by *autoMEP/RSA* model for the test set.

Table 3.3: Experimental and predicted solvation free energies (ΔG_{hyd} in kcal/mol) by our *autoMEP/RSA* model for the test set of 23 molecules.

No.	Molecule name	Exp. ΔG_{hyd} (kcal/mol)	Pred. ΔG_{hyd} (kcal/mol)	Residuals*
1	2-methylpentane	2.56	2.53	-0.03
2	<i>cis</i> -1,2-dimethylcyclohexane	1.60	2.38	0.78
3	1-hexene	1.73	1.96	0.23
4	2,3-dimethyl-1,3-butadiene	0.40	0.33	-0.07
5	toluene	-0.77	0.28	1.05
6	<i>tert</i> -butylbenzene	-0.44	-0.30	0.14
7	dichloromethane	-1.42	-1.82	-0.40
8	1,3-dibromopropane	-1.99	-1.60	0.39
9	chloroethylene	0.50	-0.76	-1.26
10	1,4-dichlorobenzene	-1.02	-1.75	-0.73
11	diethyl sulfide	-1.45	-0.80	0.65
12	diisopropyl ether	-0.54	-0.96	-0.42
13	ethane thiol	-4.08	-2.98	1.10
14	3-hexanol	-3.73	-3.8	-0.07
15	hexanal	-2.85	-2.54	0.31
16	2-butanone	-3.76	-2.63	1.13
17	methylformate	-2.82	-4.34	-1.52
18	ethylpropionate	-2.83	-2.75	0.08
19	isoamylacetate	-2.24	-2.54	-0.30
20	propylamine	-4.56	-3.88	0.68
21	dibutylamine	-3.38	-1.46	1.92
22	1-nitropropane	-3.38	-4.18	-0.80
23	2-isobutylpyrazine	-5.11	-3.84	1.27

*Predicted ΔG_{hyd} (kcal/mol) - Experimental ΔG_{hyd} (kcal/mol).

Indeed, less accurate estimation are again reported for halogen derivatives and amines (see for example, molecules **9**, **21** and **23** in Figure 3.2.2 and Table 3.3). These molecules have a chemical structure in common with the worst predicted compounds in the training set, such as chloroethylene, dibutylamine and 2-isobutylpyrazine, for which *autoMEP/RSA* deviations from the corresponding experimental values are higher than 1 kcal/mol. Overall, we can consider the combination of autocorrelation MEP vectors with a response surface analysis an alternative tool to evaluate the aqueous solvation free energy of organic compounds.

3.3 Final remarks

The solvent environment of molecules plays a very important role in their structure and function. Consequently, it is important to consider solvation effects accurately and efficiently in the prediction and simulation of the molecular properties.

In this work, we present an alternative 3D-QSPR approach combining *auto*MEP molecular descriptors with Response Surface Analysis (RSA) technique to evaluate the aqueous solvation free energy of organic compounds. Considering our results, *auto*MEP vectors can be considered an interesting electrostatic fingerprint able to describe the solvation effects, crucial in both pharmacodynamic and pharmacokinetic processes.

Parallel application of linear and nonlinear QSAR methodologies

The autocorrelated descriptors encoding for Molecular Electrostatic Potential (*autoMEP*) in combination with both linear (Partial Least Square, PLS) and nonlinear (Response Surface Analysis, RSA) strategies was demonstrated to be a reliable tool to quantitatively predict the binding affinity of human adenosine receptor antagonists. In this work, a collection of 127 known human hA_{2A} antagonists has been utilized to generate two 3D-QSAR models (*autoMEP/PLS* and *autoMEP/RSA*). PLS analysis is able to describe linear correlations, whether RSA detects the possible nonlinearity in the relationships between the molecular descriptors and the target property. However, we show that the parallel approach by using both techniques can lead to a more robust consensus in the prediction results. To validate our approach we have used our strategy to predict the binding affinity of five new human hA_{2A} pyrazolo-triazolo-pyrimidine antagonists.

4.1 Introduction

Ligand-based approaches are widely and successfully used to develop quantitative models able to correlate, and predict, the biological activities based on various molecular descriptors, especially when the bioactive conformation of the ligand is unknown, as in the case of some G protein-coupled receptors (GPCRs). The bioactive conformation represents the starting point of all 3D-QSAR strategies such as Comparative Molecular Field Analysis (CoMFA) or 3D-Pharmacophore search. [137, 138]

As anticipated, 3D-QSAR methods require the knowledge of the conformational properties of the molecules in order to calculate their structural or property descriptors. It was demonstrated that the *autoMEP* vectors in combination with Partial Least Square (PLS) analysis can represent an alternative 3D-QSAR tool to CoMFA. [132-134] However, both CoMFA and *autoMEP*/PLS methodologies can be classified as linear QSAR methods considering the mathematical relationship among molecular descriptors and the chemical/biological response space. Very recently, a nonlinear method based on a response surface analysis (RSA) application in tandem with the *autoMEP* descriptors (*autoMEP*/RSA) was also presented as an alternative 3D-QSAR method. [60, 135]

As case study we have considered the prediction of the pharmacodynamic profile of human adenosine A_{2A} receptor antagonists. More specifically, we would like to show how the applicability in parallel of both linear and nonlinear 3D-QSAR methods (*autoMEP*/PLS and *autoMEP*/RSA) can help to predict the binding affinity data of a new set of human adenosine A_{2A} receptor antagonists.

4.2 Human A_{2A}R antagonists dataset

Briefly, the adenosine A_{2A} receptors are classified in the adenosine receptor (AR) family of GPCRs, which includes A₁, A_{2A}, A_{2B} and A₃ different subtypes, abundantly expressed in diverse areas of human body and potentially in the same cellular types¹. [139] Being heptahelical transmembrane GPCRs, they are involved in several signal transduction pathways, as shown in Figure 4.2.1.

¹Further details on the four human AR subtypes are reported in section 7.2.

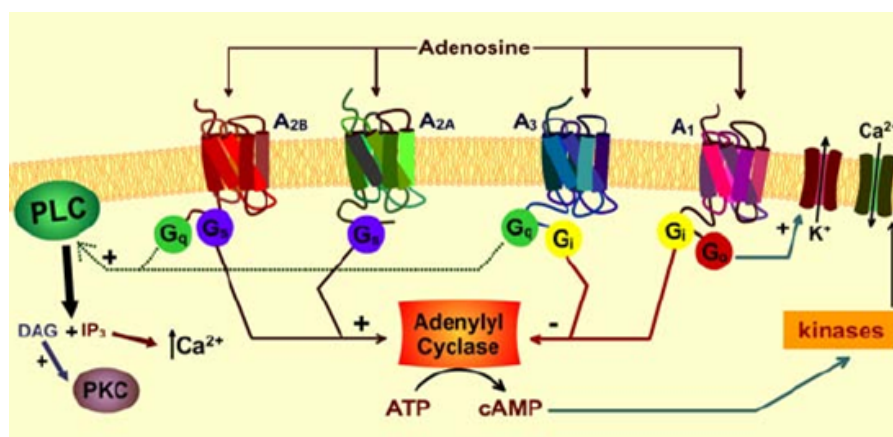


Figure 4.2.1: Signal transduction pathways of adenosine receptors.

Moreover, they are codified by distinct genes and they have been cloned from various mammalian species, where they seem to differentiate for their pharmacological profile. [139] In particular, the human adenosine A_{2A} receptor (hA_{2A}R) has been discovered to be crucial in some neurological disorders, which involve also other neurotransmission systems, above all the dopamine D₂ receptor, antagonistically associated to this adenosine receptor subtype. [140] It has been demonstrated that the activation of the human A_{2A}R causes the inhibition of platelet aggregation, attenuates the inflammatory responses mediated by cytokines, involves the regulation of the immune cells functions, while adenosine A_{2A} receptor antagonists show a neuroprotective activity during ischemic processes. [141] The inhibition of A_{2A} receptor, blocking the effects of adenosine, has been suggested as key strategy for the treatment of diverse pathologies. [142] In more detail, one of the main potential therapeutic applications of A_{2A} receptor antagonists is the promotion of cellular survival and the reduction of neuronal damage in Parkinson's or Huntington's diseases. [142-144]

In the last few years, several different potent and selective human adenosine A_{2A} receptor antagonists have been discovered. [145] In particular, pyrazolo-triazolo-pyrimidine and triazolo-pyridine derivatives were described as promising hA_{2A}R antagonists. Their chemical structures are summarized in Figure 4.2.2. In the present study, a collection of 127 known human A_{2A}R antagonists has been utilized to derive a couple of 3D-QSAR models (*auto*MEP/PLS and *auto*MEP/RSA). The binding affinity of the compounds is expressed as K_i (nM) after displacement experiments by using [3H]-NECA

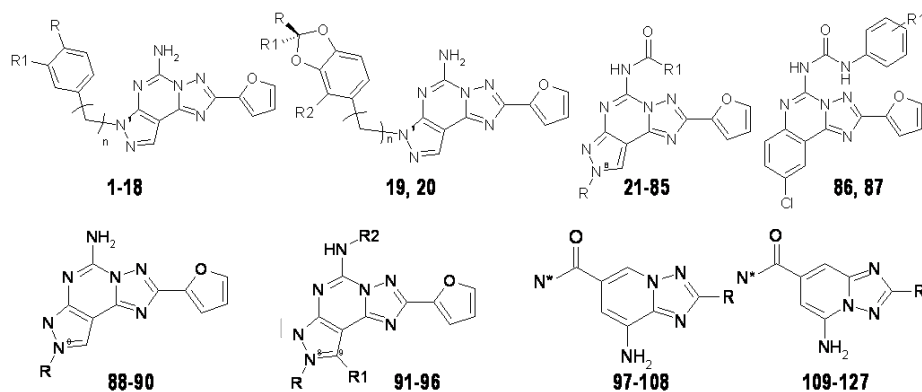


Figure 4.2.2: Chemical structures of known pyrazolo-triazolo-pyrimidine and triazolo-pyridine human A_{2A}R antagonists in the training set.

binding at human A_{2A} receptors expressed in CHO or HEK-293 cells (*Paper II*). To validate our in tandem approach, they have been utilized to predict the binding affinity of five new human A_{2A}R pyrazolo-triazolo-pyrimidine antagonists, following the flowchart shown in Figure 4.2.3.

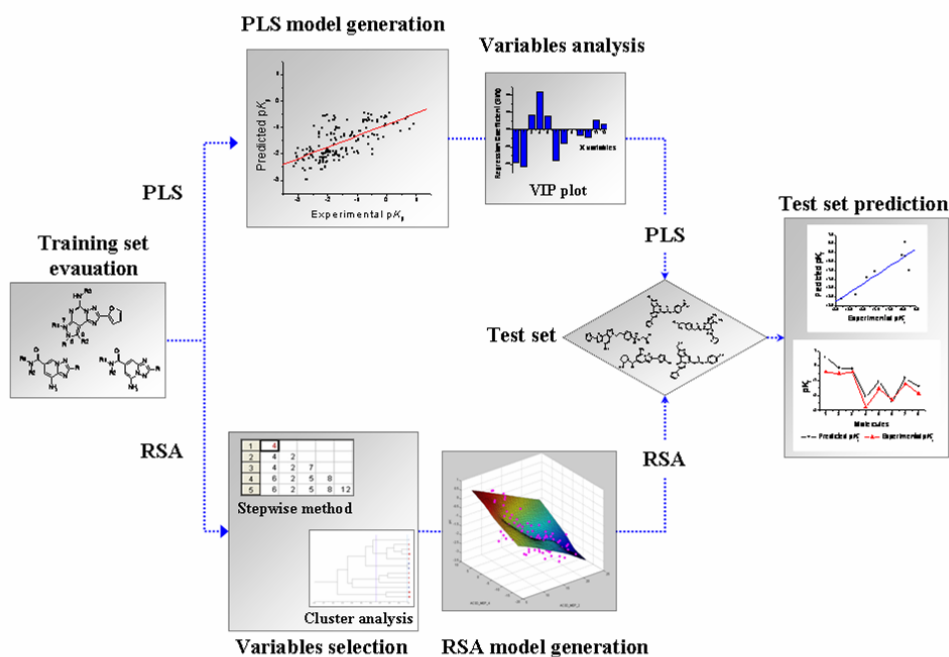


Figure 4.2.3: Partial Least Square (PLS) and Response Surface Analysis (RSA) approaches applied in tandem.

As summarized in Figure 4.2.3, a collection of pyrazolo-triazolo-pyrimidine and triazolo-pyridine analogues (molecules **1-127**) was selected as training set in both linear and nonlinear QSAR models. [146-152] An internal test set of 10 training set analogues (molecules **128-137**) was selected for the validation process of both PLS and RSA models. Finally, five new pyrazolo-triazolo-pyrimidine analogues (molecules **138-142**) has been analyzed as additional validation set².

4.3 Results and discussion

Topological and electrostatic complementarities are considered two key aspects in the molecular recognition processes. Both potentialities and advantages of the use of *autoMEP* descriptors in different 3D-QSAR applications have been previously discussed. [60, 132-135] In this work, using the *autoMEP* vectors, mentioned in chapter 2, we have assessed the possibility to combine in parallel two different linear and nonlinear strategies, PLS and RSA, to find a consensus in the quantitative binding affinity predictions.

4.3.1 PLS and RSA models

Both 3D-QSAR models have been derived by using 96 pyrazolo-triazolo-pyrimidine and 31 triazolo-pyridine derivatives as training set of known A_{2A} receptor antagonists. Moreover, both models have been subjected to a validation process by using a test set of 10 molecules (defined as the internal test set), structurally related to those included into the training set. Twelve *autoMEP* vectors have been used as independent variables in both PLS and RSA analysis (calculated as described in chapter 2); a preliminary variable selection step has been introduced before deriving RSA model (*Paper II*).

Concerning PLS analysis, the resulting model has shown acceptable statistical quality in both calibration and internal validation steps as demonstrated by the r and r_{cv} values of 0.80 and 0.78, respectively, using only three latent variables (Figure 4.3.1 and Table 4.1). The robustness of the PLS model is also supported by the good value of the correlation coefficient calculated on the test set ($q = 0.85$), as reported in Figure 4.3.2, Table 4.1 and Table 4.2.

²Experimental binding affinity data kindly provided by the work coordinated by Prof. G. Spalluto (University of Trieste) for the synthesis and by Prof. K. N. Klotz (University of Würzburg) for the pharmacological characterization.

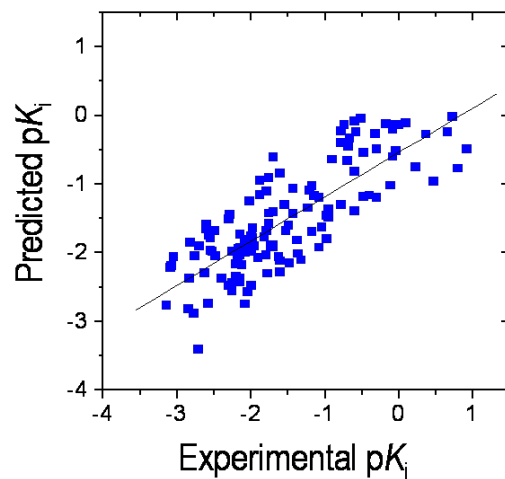


Figure 4.3.1: AutoMEP/PLS model; experimental pK_i data plotted vs predicted pK_i values (after LOO cross-validation) for the training set.

Table 4.1: Summary of the statistical parameters of autoMEP/PLS model.

Number of molecules	127
Latent variables	3
r	0.80
r_{cv}^a	0.78
Slope	0.62
Offset	-0.57
q^b	0.85
RMR ^c	0.043

^aCross-validated r after LOO cross-validation procedure: $r_{cv} = [SXY / (SXX)^{\frac{1}{2}} (SYY)^{\frac{1}{2}}]$, $SXY = \sum (X - X_{mean})(Y - Y_{mean})$, $SXX = \sum (X - X_{mean})^2$ and $SYY = \sum (Y - Y_{mean})^2$ with $X = Y_{Experimental}$ and $Y = Y_{Predicted}$; ^b r of the internal test set; ^croot mean square of residuals: RMR.

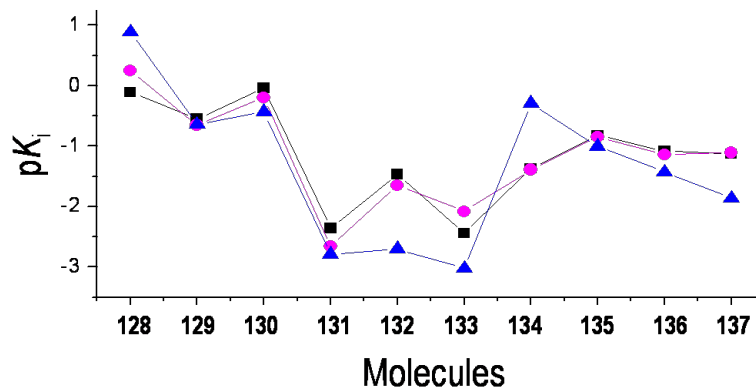
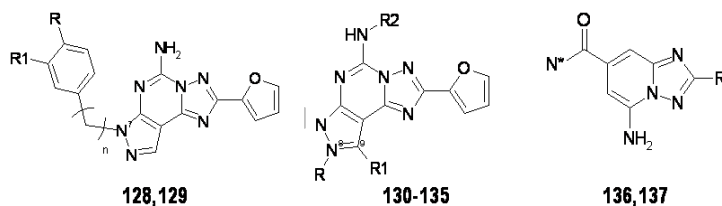


Figure 4.3.2: Comparison of autoMEP/PLS (■) and autoMEP/RSA predicted pK_i (●) with experimental pK_i values (▲) of the internal test set.

Table 4.2: Experimental and predicted pK_i for the internal test set. Differences between predicted and experimental pK_i values for both models are reported; PLS = autoMEP/PLS model; RSA = autoMEP/RSA model.



No.	Exp. pK_i	Pred. pK_i PLS	Pred. pK_i RSA	Δ PLS pK_i^a	Δ RSA pK_i^a
128	0.89	-0.11	0.25	-1.00	-0.64
129	-0.64	-0.55	-0.65	0.09	-0.01
130	-0.43	-0.04	-0.20	0.39	0.23
131	-2.79	-2.36	-2.66	0.43	0.13
132	-2.70	-1.47	-1.65	1.23	1.05
133	-3.02	-2.44	-2.08	0.58	0.94
134	-0.29	-1.37	-1.39	-1.08	-1.10
135	-1.00	-0.82	-0.85	0.18	0.15
136	-1.43	-1.08	-1.14	0.35	0.29
137	-1.86	-1.13	-1.11	0.73	0.75

^aPredicted pK_i - Experimental pK_i .

In parallel, we have delivered a nonlinear RSA model using the same training and test sets. The stepwise regression analysis together with the cluster analysis on the original 12 molecular descriptors led us to select five of them as final combination to utilize as independent variables into the RSA model: *autoMEP* 2, 4, 6, 7 and 11 (*Paper II*). The statistical parameters and the final RSA model are collected in Figure 4.3.3 and Table 4.3.

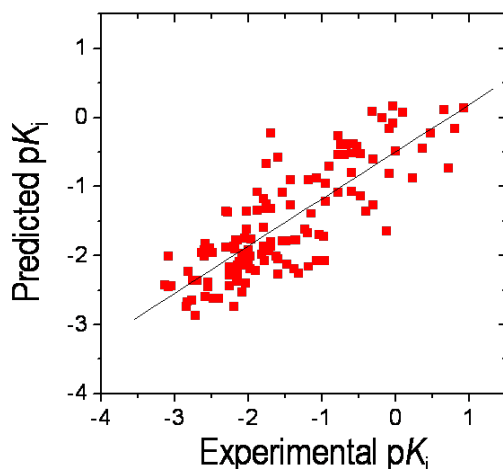


Figure 4.3.3: *AutoMEP/RSA* model; experimental pK_i data plotted vs predicted pK_i values (after LOO cross-validation) for the training set.

Table 4.3: Summary of the statistical parameters of *autoMEP/RSA* model.

Number of molecules	127
X variables	5
r	0.98
r_{cv}^a	0.82
Slope	0.68
Offset	-0.49
q^b	0.87
RMR ^c	0.043

^aCross-validated r after LOO cross-validation procedure: $r_{cv} = [SXY / (SXX)^{\frac{1}{2}} (SYY)^{\frac{1}{2}}]$, $SXY = \sum (X - X_{mean})(Y - Y_{mean})$, $SXX = \sum (X - X_{mean})^2$ and $SYY = \sum (Y - Y_{mean})^2$ with $X = Y_{Experimental}$ and $Y = Y_{Predicted}$; ^b r of the internal test set; ^croot mean square of residuals: RMR.

We can observe a very high correlation coefficient ($r = 0.98$) for the calibration step confirming the good choice of the independent variables.

The correlation coefficients after LOO cross-validation and calculated on the test set are also appreciable ($r_{cv} = 0.82$ and $q = 0.87$, respectively). These results represent additional evidences about the good predictivity of the *autoMEP*/RSA model, as shown in Figure 4.3.2 and Table 4.2.

Even if both methods are statistically acceptable, the *autoMEP*/RSA model presents higher predictivity than *autoMEP*/PLS model (Table 4.2). However, both methodologies are able to coherently discriminate between "more active" and "less active" analogues. This result is very interesting because ensemble, or *consensus*, approaches to classification and regression have been considered as attractive tools. [153] In fact, these methods have been shown to outperform a single predictor usability on a wide range of scientific tasks. [153]

4.3.2 External test set prediction

The application in parallel of different strategies may confirm the predictions achieved by using single models alone. In this study, we used both *autoMEP*/PLS and *autoMEP*/RSA models as an ensemble of binding affinity predictors to prioritize the synthesis of new human A_{2A}R antagonists.

Following these encouraging results, we have tested the real predictive capability of our PLS and RSA models on an external test set, which consisted in five new pyrazolo-triazolo-pyrimidine analogues. This is a preliminary proof of concept of our parallel PLS/RSA approach. As anticipated in the Introduction, we aim at simultaneously performing different 3D-QSAR approaches to create a more even balance between false positive and false negative performance rates than the use of a single method can achieve.

In our laboratories, we are still developing new potent and selective human A_{2A}R antagonists decorating the pyrazolo-triazolo-pyrimidine scaffold using different strategies. [154] In this context, we analyzed a new class of N⁵-substituted derivatives in which a different series of benzyloxy-phenyl-acetyl substituents are present³. Once *autoMEP* vectors have been computed for this new set of ligands, we have applied both PLS and RSA models for their binding affinity predictions (Figure 4.3.4 and Table 4.4).

³Experimental binding affinity data kindly provided by the work coordinated by Prof. G. Spalluto (University of Trieste) for the synthesis and by Prof. K. N. Klotz (University of Würzburg) for the pharmacological characterization.

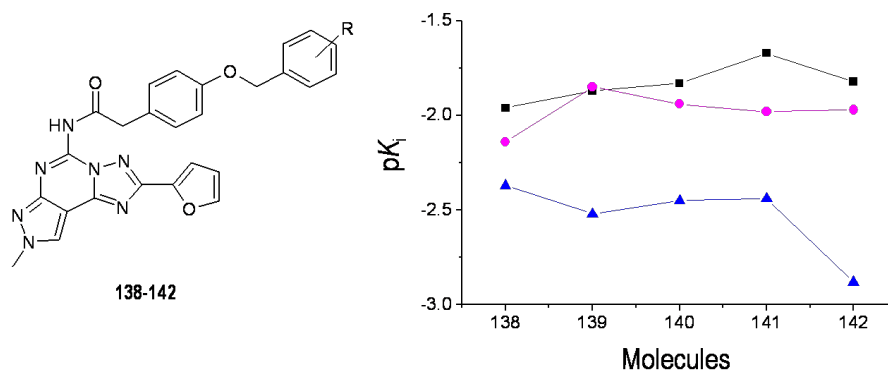


Figure 4.3.4: Experimental pK_i activity data (▲) of the external test set compared to pK_i values predicted by *autoMEP/PLS* (■) and *autoMEP/RSA* (●) models. The relative molecular scaffold is reported on the left.

Table 4.4: Experimental and predicted pK_i for the new synthesized pyrazolo-triazolo-pyrimidine derivatives 138-142. Differences between predicted and experimental pK_i values for both models are reported; PLS = *autoMEP/PLS* model; RSA = *autoMEP/RSA* model.

No.	R	Exp. K_i^a (nM)	Exp. pK_i	Pred. pK_i PLS	Pred. pK_i RSA	ΔpK_i^b PLS	ΔpK_i^b RSA
138	H	233	-2.37	-1.96	-2.14	0.41	0.23
139	2-CH ₃	333	-2.52	-1.87	-1.85	0.65	0.67
140	2,6-Cl	281	-2.45	-1.83	-1.94	0.62	0.51
141	3-Cl	278	-2.44	-1.67	-1.98	0.77	0.46
142	4-CH ₃	751	-2.88	-1.82	-1.97	1.06	0.91

^a K_i obtained by displacement of specific [3H]-NECA binding at human A_{2A} receptors expressed in CHO cells; ^b Predicted pK_i - Experimental pK_i .

As shown in Table 4.4, both methods predicted all five derivatives active in the low nM range (K_i values in between to 50 and 150 nM), with a better performance of the *autoMEP/RSA* model with respect to *autoMEP/PLS* one. Interestingly, even if both methods overestimated all binding affinities, new synthesized pyrazolo-triazolo-pyrimidine analogues are active in the nM range (K_i values in between to 230 and 750 nM), as theoretically predicted (Figure 4.3.4).

4.4 Final remarks

In light of the consideration that GPCR ligands represent one of the major continuing source of novel potent therapeutic agents, and that 3D structures of GPCRs as determined by experimental techniques are still unavailable, ligand-based drug discovery methods remain the most feasible computational approaches to the analysis of growing data sets of developmental GPCR ligands.

We have proposed the application of a couple of complementary QSAR methodologies to evaluate the binding affinity data of a new series of human A_{2A}R antagonists. Two statistically meaningful models have been generated from a common training set, and their predictivity has been evaluated by using both internal and external test sets. We are continuously analyzing new series of ligands with the aim to improve the robustness and the predictivity of our QSAR models. The purpose is to perform *in silico* screening of real or virtual libraries to research for new potent and selective human A_{2A}R antagonists.

Isoform specificity of cytochrome P450 substrates

Each drug can potentially be metabolized by different cytochrome P450 (CYP450) isoforms. In the development of new drugs, the prediction of the metabolic fate is important to prevent drug-drug interactions. The present study deeply analyzes a collection of 554 CYP450 substrates by applying multi- and single-label classification strategies, after the computation and the selection of suitable molecular descriptors. Cross-training with support vector machine and counter-propagation neural network modeling methods were used in the multilabel approach, which allows one to classify the compounds simultaneously in multiple classes. In the single-label models automatic variable selection was combined with various cross-validation experiments and modeling techniques. Moreover, the reliability of both multi- and single-label models was assessed by the prediction of an external test set. Finally, the predicted results of the best models were compared to show that, even if the models present similar performances, the multilabel approach more coherently reflects the real metabolism information.¹

¹This work has been carried out at Molecular Networks, Erlangen (Germany), with the supervision of Prof. J. Gasteiger and the collaboration of Dr. L. Terfloth.

5.1 Cytochrome P450 in drug metabolism

The metabolic profile of a drug candidate is an important aspect to be considered in the selection of a potential new drug. Several problems related to stability, toxicity of xenobiotics and drug interactions might represent serious adverse effects. In fact, in the case of co-administration of drugs, the pharmacological profile of each drug might be modified by the presence of other drugs in the human body. [155, 156] If two drugs are co-administered and metabolized by the same enzyme, the competition for the binding site can result in the inhibition of the biotransformation of one or both drugs.

Focusing the attention on metabolism in the ADMET process, a crucial role is played by the cytochrome P450 (CYP450) class of hemoprotein enzymes. The CYP450 superfamily of enzymes is abundantly expressed in the liver and remarkably in the small intestine, where it is responsible for the detoxification of xenobiotics. [157, 158] In the Phase I metabolism cytochrome P450 isoforms chemically modify a large variety of substrates mainly through oxidation reactions to make them more water-soluble and to ease their elimination. [159] This detoxification system is highly complex, since it includes many different CYP450 isoforms characterized by multiple binding sites, polymorphism and enzyme induction or modulation phenomena. [160] These aspects are involved in drug-drug interactions, which might lead to unpredictable blood concentrations of one of the xenobiotics with consequent possible toxic effects or loss of activity. CYP450 enzymes are classified in several isoforms according to the similarity of their amino acidic sequences. We have investigated CYP450 1A2, 2C9, 2D6, 2E1 and 3A4 substrates, that cover almost all possible metabolism routes (Figure 5.1.1).

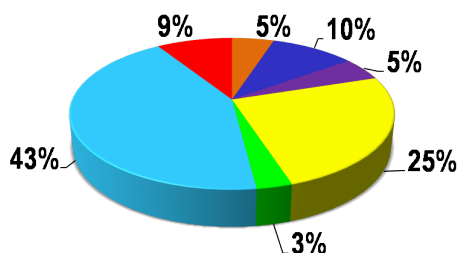


Figure 5.1.1: Importance, as percentage (%), of the different CYP450 isoforms in the metabolic process; CYP1A2 (orange), CYP2C9 (blue), CYP2C19 (purple), CYP2D6 (yellow), CYP2E1 (green), CYP3A4 (light blue), other isoforms (red).

In our analysis, we have excluded CYP2C8 and CYP2C19 isoforms, poorly represented, as they are not much involved in the metabolic process. The same analysis with seven CYP450 isoforms has been performed, as reported in *Paper III*. CYP450 1A2 metabolizes planar molecules characterized by moderate volume and basicity. [157] CYP450 2C9 substrates are acidic or neutral, and lipophilic molecules, in particular sulfonylureas and NSAIDs (non steroidal anti-inflammatory drugs) drug classes. [157] CYP450 2D6 shows polymorphism and about 25% of all drugs are at least partial substrates of this isoform. They show a hydrophilic character and have a basic nitrogen atom. [157] Mostly small and polar molecules, like volatile anesthetics, are substrates of CYP450 2E1 isoform which is involved in many drug interactions. [157] The CYP450 3A4 isoform, ubiquitously found, is responsible for the metabolism of high volume and lipophilic xenobiotics; almost 50% of all drugs are metabolized by this isoform. [157] For this reason its activity is largely affected by chemically different compounds and CYP3A4 represents the most populated class in the dataset investigated in this study. However, different isoforms might be responsible for the detoxification of the same drug.

5.2 Computational approaches to CYP450

The early detection of ADMET properties of drugs under *in vivo* conditions is experimentally time-consuming and expensive, therefore computational methods can profitably speed up the collection of new data. [161-163] A challenging problem in this field is the prediction of CYP450 isoform specificity. Different chemoinformatic tools have already been attempted for the prediction of CYP-related metabolism properties. [164, 165] In more detail, several ligand-based approaches were applied to classify CYP450 substrates according to their route of metabolism. [44, 45, 166-170] Anyway, most solutions consider local models for each CYP450 isoform and they do not approach the problem globally. Among these, the traditional single-label classification deals exclusively with non-overlapping classes. Following this approach, a classification model was developed to predict the isoform specificity for CYP3A4, CYP2D6 and CYP2C9 substrates considering non-overlapping classes, i.e. assuming each compound to be metabolized by a single, predominant CYP450 isoform. [76]

In the present study we would like to extend the abovementioned model to cover other CYP450 isoforms and to find a strategy to predict the substrates which are metabolized by more than one isoform. Such a multilabel classification analysis represents a different approach that can be applied whether our dataset comprises elements assigned simultaneously to more than one class. The prediction of the metabolism profile of CYP450 substrates represents a novel application of this methodology.

This work aims at the prediction of the isoform specificity from information on the substrates metabolism. The substrates were represented by different sets of structural and physicochemical descriptors. Then, various classification techniques - both multilabel and single-label approaches - were used to derive models for the prediction of the isoform(s) responsible for metabolism of CYP450 1A2, 2C9, 2D6, 2E1 and 3A4 substrates. The procedures we followed are separately illustrated in Figure 5.2.1.

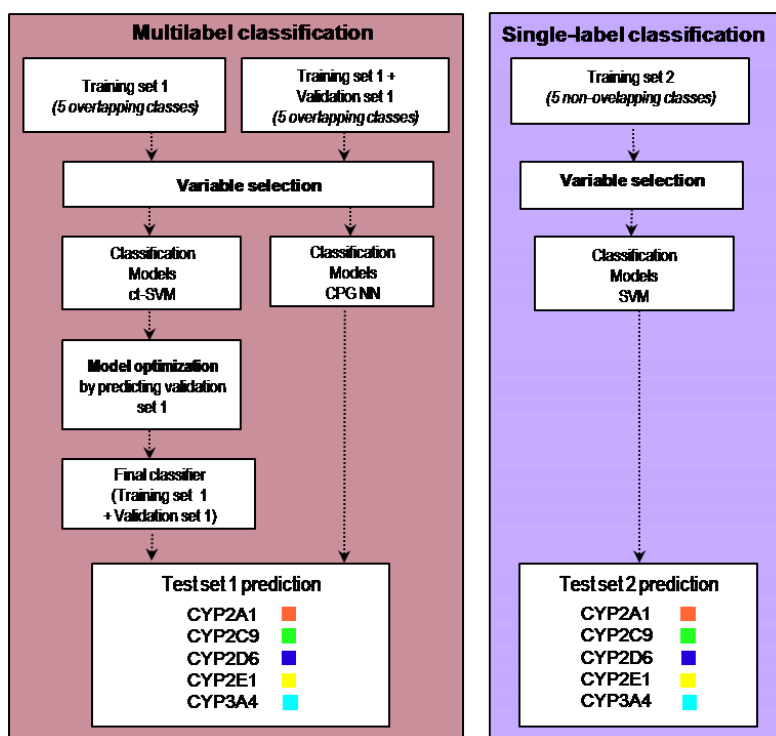


Figure 5.2.1: Flowchart of the multilabel and single-label approaches for the prediction of CYP450 isoform specificity.

We applied a multilabel classification method to distinguish between substrates of five CYP450 isoforms. Then, only single-label compounds, i.e. compounds metabolized by a single isoform, were utilized to build a classifier. Variable selection and optimization of the models were performed in the present study. The single-label classification models were directly generated after an automatic variable selection process. The modeling results and the metabolism profiles of the CYP450 substrates of the test sets predicted by the different approaches were compared to verify the reliability of both multi- and single-label classification methods.

5.3 Dataset

A collection of 554 cytochrome P450 substrates with different chemical structures was used to derive our classification models. In particular, we considered the dataset compiled in the recently published paper by Block *et al.* [170] It includes 253 substrates metabolized by CYP1A2, CYP2C19, CYP2C8, CYP2C9, CYP2D6 and CYP3A4 isoforms. At least one isoform may be responsible for the metabolism of a single substrate, i.e. a compound might be metabolized by several isoforms and then belongs to several classes. Since we considered all possible routes of metabolism, our classification problem is multilabel. Further substrates were extracted from other published papers and publicly available lists. [171-174] Moreover, 267 additional compounds from the Metabolite reaction database were included. [175] When inconsistencies resulted in the comparison between the datasets, we considered more reliable the information about the compounds metabolic fate published by Block *et al.* In the final collection (554 compounds) 484 substrates are metabolized by one CYP450 isoform and the remaining 70 compounds are metabolized by several CYPs. Only 13% of all compounds in the dataset are multilabel. Considering the 484 compounds metabolized by one isoform, 46 are CYP1A2 substrates (9.5%), 50 are CYP2C9 substrates (10.3%), 106 are CYP2D6 substrates (21.9%), 49 are CYP2E1 substrates (10.1%) and 233 are CYP3A4 substrates (48.2%). Not all isoforms have the same relevance in the xenobiotic metabolism, consequently these classes are differently populated and our dataset is quite unbalanced.

For our modeling studies two different data sets were compiled (Data set 1 and Data set 2). Data set 1 was manually split into training, validation

and test set with a similar distribution of the considered classes in the entire data set and the subsets. Most of the compounds from the Metabolite data set were used as test set, even if some substrates were included in the training and the validation set (*Paper III*). Data set 2 was simply split into a training and a validation set, applying the same selection criterion (similar distribution of the substrates in the considered classes).

5.3.1 Data set 1

Data set 1 comprises the initial collection of 554 chemically different substrates, metabolized by five CYP450 isoforms (CYP1A2, CYP2C9, CYP2D6, CYP2E1 and CYP3A4 single- and multilabel substrates). Data set 1 was used to perform a multilabel classification analysis. The distribution of the classes in the training, validation and test sets is reported in Table 5.1.

Table 5.1: *Data set 1 used in this analysis. 554 single- and multilabel CYP450 substrates classified in five isoforms (CYP1A2, CYP2C9, CYP2D6, CYP2E1 and CYP3A4). The distribution of the substrates in training, validation and test sets 1 are listed.^a*

CYP450 Isoform	Training set 1		Validation set 1		Test set 1	
	S	M	S	M	S	M
CYP1A2	26	17	6	5	14	9
CYP2C9	31	13	8	3	11	7
CYP2D6	46	24	7	8	53	11
CYP2E1	30	2	8	0	11	1
CYP3A4	110	32	21	10	102	14
Total (554)	243	40	50	12	191	18

^aSingle-label substrates occur only once in each class; multilabel substrates belong to more than one class. Consequently, the sum of multilabel substrates for all the classes within a column is higher than the number of multilabel substrates for Training set 1, Validation set 1 and Test set 1; S = single-label; M = multilabel.

5.3.2 Data set 2

Only the single-label substrates in Data set 1 were selected to perform a single-label classification analysis. Then, 484 compounds, a subset of Data set 1, was used as Data set 2. In the new training set (Training set 2) exactly the same single-label substrates collected in the Training set 1 and Validation set 1 (totally 293 compounds) were included, while in the Test set 2 only the single-label substrates in the Test set 1 (191 compounds) were considered.

Data set 2 was utilized to develop a single-label classification model. The results of the splitting process for Data set 2 can be inferred by considering the "S" columns for the training, validation and test sets in Table 5.1.

5.4 Results

The prediction of isoform specificity represents a multilabel classification problem, characterized by high complexity of the feature space. In this study, we built a model to simultaneously classify a collection of substrates metabolized by five CYP450 isoforms (CYP1A2, CYP2C9, CYP2D6, CYP2E1 and CYP3A4). Various descriptors and data analysis techniques were combined to predict the isoform specificity using two different data sets, which correspond to (1) multilabel classification models (Data set 1) and (2) single-label classification model (Data set 2). In more detail, cross-training with Support Vector Machine (ct-SVM) and counter-propagation neural networks (CPG NN) have been applied as multilabel classification techniques, while SVM was used to develop single-label classification models. Only the best-performing experiments are reported in this chapter.

5.4.1 Multilabel classification with Data set 1

Since the information about the reaction site of the substrates was not reported, several models were derived combining global, topological, shape and functional group counts descriptors with 2D topological or 3D spatial autocorrelation vector components (descriptors reported in section 2.3 and in Table 2.1). A manual descriptor selection process was combined with ct-SVM multilabel modeling method in the optimization procedure using the Training set 1. The best subset of global, topological, shape and functional groups count descriptors was selected according to the model performance after each single-descriptor addition step. In more detail, a descriptor was included in the best subset if the model predictivity results on the Validation set 1 improved. The best model corresponded to the final set of 27 descriptors, as reported in Table 5.2.

The information encoded by the autocorrelation molecular electrostatic potential descriptors improved the predictivity of the model in comparison to the model computed without these twelve variables (*Paper III*). It supports the concept that the distribution of electrostatic properties on the molecu-

Table 5.2: *Twenty seven descriptors selected for the training set in multilabel classification models with Data set 1.*

No.	Name	Details
1	MW	Molecular weight
1	HAccPot	Hydrogen bond acceptor potential
1	TPSA	Topological polar surface area
1	ASA	Approximate surface area
1	D_3	Diameter
1	R_3	Radius
1	I_3	Geometric shape coefficient
1	r^2	Radius perpendicular to D_3
1	r^3	Radius perpendicular to D_3 and R_2
1	n_{aro_amino}	Number of aromatic amino groups
1	n_{tert_amino}	Number of tertiary aliphatic amino groups
1	$n_{prim_sec_amino}$	$n_{prim_amino} + n_{sec_amino}$
1	$n_{basic_nitrogen}$	Number of basic, N containing functional groups
1	n_{acidic_groups}	Number of acidic functional groups
1	$q_{\pi_1} = \sum q_{\pi}^2$	property: π -charge q^{π}
12	SurfACorr_ESP	Surface autocorrelation; property: molecular electrostatic potential

lar surface represents an important determinant in the prediction of isoform specificity. In the ct-SVM modeling method, the best results were achieved using the T-criterion to transform the real-valued scores assigned by the corresponding classifier into labels. This descriptor set was used to derive a first ct-SVM classification model with Training set 1. As previously described, the model parameters were optimized by predicting the Validation set 1. Then, TP, FP, TN, FN rates and the percentage (%) of correct predictions were calculated from the confusion matrix (Table 5.3).

Table 5.3: *Multi-classification ct-SVM model; the statistical parameters for each class after prediction on the Validation set 1 (62 substrates) are reported.*

Classes	TP rate	FP rate	TN rate	FN rate	Recall	Precision	% correct predictions
CYP1A2	0.64	0.04	0.96	0.36	0.64	0.78	90.3
CYP2C9	0.82	0.05	0.95	0.18	0.82	0.75	95.2
CYP2D6	0.87	0.00	1.00	0.13	0.87	1.00	93.5
CYP2E1	0.88	0.00	1.00	0.12	0.88	1.00	98.4
CYP3A4	0.87	0.25	0.75	0.13	0.87	0.75	87.1

As seen in Table 5.3, a good predictivity is achieved for almost all the

classes if we analyze the values of recall and precision. Training set 1 and the Validation set 1 were merged to derive a new classifier with the optimized parameters. The performance measures of the final ct-SVM model are shown in Table 5.4.

Table 5.4: Multi-classification ct-SVM model; performance measures after prediction on the Validation set 1 (62 substrates) and the Test set 1 (209 compounds) are reported.

Model prediction	Accuracy $_{ML}$	One-error	Coverage	Average precision
Validation set 1	0.84	0.10	1.53	0.93
Test set 1	0.70	0.25	1.52	0.85

Satisfactory results for the multilabel approach were also achieved by the application of CPG NN technique with the set of 27 descriptors in Table 5.2. A graphical representation of our analysis is shown in Figure 5.4.1.

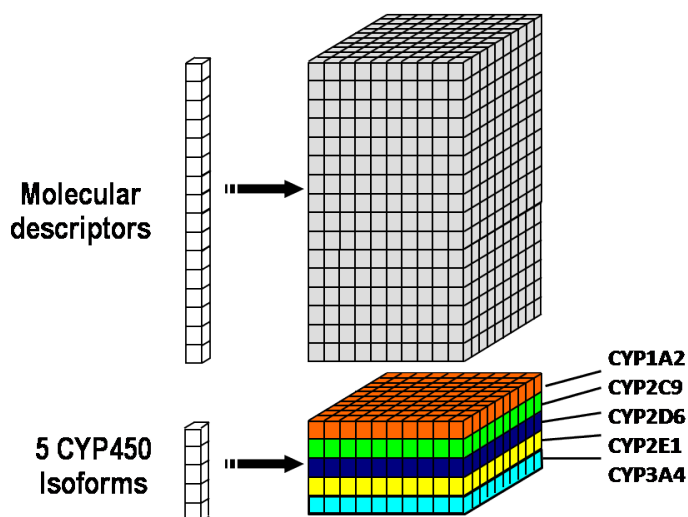


Figure 5.4.1: Flowchart of the CPG NN analysis with five output layers (CYP450 isoforms) applied to our multilabel classification problem.

Different map sizes, topologies and number of epochs were selected for the learning process and the best model with a rectangular topology is discussed here. After LOO and 5-fold cross-validation procedures, the best CPG NN model was selected. The model performances are reported in Table 5.5. Good values of precision and recall were obtained for CYP2C9, CYP2D6, CYP2E1 and CYP3A4 classes, by applying the best CPG NN model. A low predictivity of CPG NN model can be observed for CYP1A2 and CYP2C9

Table 5.5: Multi-classification CPG NN; the statistical parameters for each class after LOO and 5-fold cross-validation (345 substrates) are reported.

Classes	TP rate		FP rate		TN rate		FN rate		Recall		Precision	
	LOO	5-fold	LOO	5-fold	LOO	5-fold	LOO	5-fold	LOO	5-fold	LOO	5-fold
CYP1A2	0.33	0.48	0.10	0.10	0.90	0.90	0.66	0.52	0.33	0.48	0.39	0.47
CYP2C9	0.60	0.58	0.07	0.07	0.93	0.93	0.40	0.42	0.60	0.58	0.62	0.60
CYP2D6	0.79	0.79	0.09	0.06	0.91	0.93	0.21	0.21	0.79	0.79	0.74	0.80
CYP2E1	0.87	0.85	0.01	0.03	0.99	0.98	0.12	0.15	0.87	0.85	0.95	0.77
CYP3A4	0.77	0.77	0.24	0.21	0.76	0.78	0.22	0.23	0.77	0.77	0.76	0.78

classes in terms of lower TP rate or, alternatively, higher FN rate.

In general, a model with a minimum predictivity for one class might not be able to detect substrates metabolized by this particular isoform, whether it was applied to an external test set. However, at least 75% of the compounds for each class were correctly classified (*Paper III*). In Figure 5.4.2 the five output layers of the CPG NN network are shown.

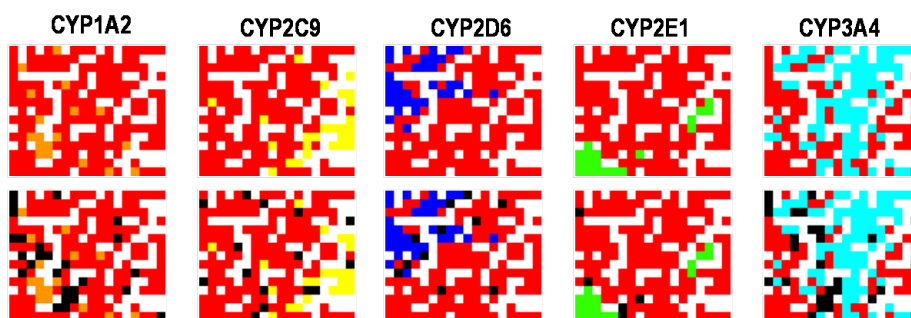


Figure 5.4.2: CPG NN model; projection of the CYP450 substrates into five maps, corresponding to CYP1A2, CYP2C9, CYP2D6, CYP2E1 and CYP3A4 classes. CYP1A2 substrates in orange, CYP2C9 substrates in yellow, CYP2D6 substrates in blue, CYP2E1 substrates in green, CYP3A4 substrates in light blue are indicated (according to the most frequent output). For each map, neurons containing substrates which do not belong to the corresponding class are red. In the maps below conflicting neurons in black are also shown. Black neurons contain compounds of different CYP450 classes. White squares represent empty neurons.

In each layer, the CYP450 substrates tend to cluster. This tendency is not particularly evident for CYP1A2 class, where the occupied neurons are spread out in the corresponding layer. Few conflict neurons are present in the

maps, and most of them are caused by CYP3A4 substrates, conflicting with the substrates metabolized by other classes. In fact, CYP3A4 represents the major and the most chemically heterogeneous class. In some cases, the descriptor set we selected is not able to correctly classify the substrates with similar structural features and different class memberships.

5.4.2 Single-label classification with Data set 2

Data set 2 includes only single-label substrates extracted from Data set 1 and it is therefore suitable for the following data analysis. In the single-label approach, a systematic variable selection procedure was performed, by considering global, topological, shape and functional group counts, 128 spatial autocorrelation vectors (totally 168 descriptors in Table 2.1). Finally, nineteen descriptors were automatically selected by using the Training set 2, as summarized in Table 5.6.

Table 5.6: Nineteen descriptors resulting from automatic variable selection for the training set in single-label classification models with Data set 2.

No.	Name	Details
1	HDonPot	Hydrogen bond donor potential
1	TPSA	Topological polar surface area
1	ASA	Approximate surface area
1	μ	Molecular dipole moment
1	r^2	Radius perpendicular to D_3
1	r^3	Radius perpendicular to D_3 and R_2
1	n_{aliph_amino}	Number of aliphatic amino groups
1	$n_{prim_sec_amino}$	$n_{prim_amino} + n_{sec_amino}$
1	$n_{basic_nitrogen}$	Number of basic, N containing functional groups
1	n_{acidic_groups}	Number of acidic functional groups
1	3D-AC _{identity} [1.2-1.3 Å]	Spatial autocorrelation; property: identity
1	3D-AC _{identity} [1.3-1.4 Å]	Spatial autocorrelation; property: identity
1	3D-AC _{identity} [1.4-1.5 Å]	Spatial autocorrelation; property: identity
1	3D-AC _{identity} [1.7-1.8 Å]	Spatial autocorrelation; property: identity
1	3D-AC _{identity} [2.3-2.4 Å]	Spatial autocorrelation; property: identity
1	3D-AC _{identity} [2.7-2.8 Å]	Spatial autocorrelation; property: identity
1	3D-AC _{identity} [3.1-3.2 Å]	Spatial autocorrelation; property: identity
1	3D-AC _{identity} [4.2-4.3 Å]	Spatial autocorrelation; property: identity
1	3D-AC _{identity} [5.3-5.4 Å]	Spatial autocorrelation; property: identity

The 3D autocorrelation identity descriptors reflect the distribution of the interatomic distances in the 3D molecular structure and complete the information given by the first selected subset. In more detail, the BestFirst automatic criterion implemented in Weka was applied to select the variables. [121] The descriptor space was explored in order to detect the subset that is likely to predict the classes best. The attribute evaluator CfsSubsetEval combined with Best First search method has been applied. The variable selection process was repeated for each fold during the model validation step. The selected nine components of 3D autocorrelation identity vectors correspond to particular atom distances and show an important contribution in the model building process.

We have generated a single-label classification model, by combining the automatic variable selection with Support Vector Machine. The standard parameters suggested in Weka software were selected for the computation. In the SVM model, a polynomial kernel with exponent equal to three was used. The results are reported in Table 5.7.

Table 5.7: *Single-label classification SVM model; the statistical parameters using nineteen descriptors for the Training set 2.*

Partition	CV	No. of runs	% correct predictions			
			Mean	StDev	Min	Max
Training set 2		1	85.7	-	85.7	85.7
Training set 2	LOO	1	75.8	-	75.8	75.8
	10-fold	10	76.3	1.3	74.7	78.2
	5-fold	20	76.0	1.2	73.3	78.2
	3-fold	33	75.3	1.7	71.0	78.1
	2-fold	50	75.1	2.2	69.3	80.2
Test set 2		1	78.0	-	78.0	78.0

The model is quite stable if we analyze the profile of the standard deviation values for each n -fold cross-validation and the predictivity in Table 5.7. In the LOO cross-validation a predictivity of 75.8% was obtained. The percentage of correct predictions is lower for the other n -fold cross-validation procedures, with a difference of 10.6% between the Training set 2 predictivity and the average prediction accuracy in 2-fold cross-validation. Test set 2 is predicted with an accuracy of 78%. Table 5.8 shows the TP, FP, TN and FN rates of the Training set 2 in LOO cross-validation for SVM model. Further details on the models derived are reported in *Paper III*.

Table 5.8: *Single-label classification SVM model; predictivity results after LOO cross-validation for Training set 2.*

Classes	TP rate	FP rate	TN rate	FN rate	Recall	Precision
CYP1A2	0.34	0.01	0.99	0.66	0.34	0.73
CYP2C9	0.72	0.04	0.96	0.28	0.72	0.74
CYP2D6	0.77	0.05	0.95	0.23	0.77	0.77
CYP2E1	0.89	0.06	0.94	0.11	0.89	0.68
CYP3A4	0.82	0.18	0.82	0.18	0.82	0.79

5.4.3 Validation of the models with an external test set

In our analysis different test sets were studied, according to the data set used to build up the classification models, as described in paragraph 5.2. The isoform specificity was predicted for each test set by applying the multilabel (Test set 1) or the single-label classification (Test set 2) models.

Test set 1. 209 substrates (Test set 1) were analyzed by both ct-SVM and CPG NN isoform predictors. The prediction results on Test set 1 by using ct-SVM and CPG NN techniques are summarized in Table 5.9 and Table 5.10, respectively.

Table 5.9: *Multilabel classification ct-SVM: predicted results of the model for the Test set 1. The number of true positives is indicated in the second column for each class.*

Classes	TP rate (TP)	TN rate	Recall	Precision	% correct predictions	Matthews correlation coefficient
CYP1A2	0.65 (15/23)	0.90	0.65	0.44	87.1	0.47
CYP2C9	0.44 (8/18)	0.96	0.44	0.53	91.9	0.44
CYP2D6	0.59 (38/64)	0.92	0.59	0.78	82.3	0.58
CYP2E1	0.75 (9/12)	0.98	0.75	0.69	96.6	0.70
CYP3A4	0.84 (97/116)	0.70	0.84	0.78	77.5	0.56

In the ct-SVM model predictions the FN rate is remarkable for CYP2C9 class. Considering that the classifier is based on five classes, the model prediction capability is quite accurate.

Regarding CPG NN model, only for the CYP2E1 class surprisingly good results were achieved, with a recall of 0.92, as reported in Table 5.10. Statistically acceptable values of TP rate corresponded to CYP2D6 (0.70) and CYP3A4 (0.72) isoforms, while a low predictivity was found for the remaining classes.

Table 5.10: *Multilabel classification CPG NN model; predicted results for the Test set 1.*

Classes	TP rate	TN rate	Recall	Precision	% correct predictions	Matthews correlation coefficient
CYP1A2	0.52	0.85	0.52	0.30	81.3	0.29
CYP2C9	0.61	0.93	0.61	0.44	89.9	0.46
CYP2D6	0.70	0.88	0.70	0.72	82.8	0.58
CYP2E1	0.92	0.97	0.92	0.69	97.1	0.78
CYP3A4	0.72	0.79	0.72	0.81	75.6	0.52

Test set 2. In the single-label model the Test set 2 (191 CYP450 substrates) was used to assess its predictivity. As seen in Table 5.7, the percentage of correct predictions for Test set 2 resulted in 78% for the SVM model. This value is slightly higher than the corresponding value of correct predictions after 2-fold cross-validation. In Table 5.11 the prediction rates of the same single-classification model for each class of the Test set 2 are reported.

Table 5.11: *Single-label classification SVM model; predicted results for the Test set 2. The number of true positives is indicated in the second column for each class.*

Classes	TP rate (TP)	TN rate	Recall	Precision	Matthews correlation coefficient
CYP1A2	0.64 (9/14)	0.97	0.64	0.60	0.59
CYP2C9	0.73 (8/11)	0.98	0.73	0.67	0.68
CYP2D6	0.77 (41/53)	0.91	0.77	0.76	0.67
CYP2E1	1.00 (11/11)	0.97	1.00	0.65	0.79
CYP3A4	0.78 (80/102)	0.85	0.78	0.85	0.64

All CYP2E1 substrates are correctly predicted (the TP rate is equal to one) and the values of TP rate for the remaining classes are included in the interval 60-80%. Moreover, the SVM model was applied to predict the eighteen multilabel substrates in Test set 1, not included in Test set 2. The prediction results were considered in the final comparison with the other multilabel classification methods.

5.5 Discussion

5.5.1 Aspects related to the data set

The classification of multilabel data represents a complex problem. So far, many different strategies were explored to classify drugs metabolized by a single isoform. However, this approach does not reflect the real scenario, in which the route of metabolism might involve several enzymes for the biotransformation. In fact, the same molecular structure can be recognized by different isoforms, on the other hand the same CYP450 isoform might metabolize chemically diverse compounds and this is particularly true for CYP3A4 isoform. Therefore, CYP450 enzymes are not selective and in the metabolic process the CYPs show a different role and relevance. As a consequence, the data set is unbalanced, with few compounds classified as CYP1A2, CP2C19, CYP2C8, CYP2C9, CYP2E1 substrates and more represented by CYP2D6 and CYP3A4 classes.

Moreover, in our analysis we dealt with information coming from different sources and in some cases it was inconsistent or incomplete. Then, the uncertainty of information led us to choose a compromise by considering more reliable the most recent source.

5.5.2 Considerations on the selected descriptors

In the metabolic process a recognition mechanism is responsible for the complementary interaction between the substrates and cytochrome P450 isoforms. Therefore, the chemical nature of the substrates and especially the distribution of particular properties on their surface are involved in the determination of their metabolic fate. The specificity of the interactions is driven by several molecular properties. Moreover, the function of autocorrelation is a useful strategy to overcome the dependence on the spatial rotation and translation of the molecules. In fact, the autocorrelation concept is able to describe the distribution of a particular property on the molecular surface and to represent molecules of different size with a vector of fixed length.

In ct-SVM model the selected descriptors (Table 5.2) have been confirmed relevant by the prediction results for CYP1A2, CYP2C9 and CYP3A4 classes. The CPG NN model is based on the same descriptors set. The clearly distinguishable clusters in the maps corresponding to each layer/class further on support the good choice of the variables. The results demonstrate that

the molecular size and the presence of particular functional groups as well as the distribution of the electrostatic or charge properties positively affect the model predictivity. However, the autocorrelation molecular electrostatic potential descriptors are more understandable than the 3D autocorrelation identity components used in the single-label model, so they can easily substitute the vectorial properties used in the paper by Terfloth *et al.* [76]

Seven out of nineteen descriptors in the single-label classification SVM model are in common with the manually selected variables in the multilabel classification models (TPSA, ASA, r_2 , r_3 , $n_{prim_sec_amino}$, $n_{basic_nitrogen}$ and n_{acidic_groups} descriptors). Various shape and size related descriptors were recognized as important in the single-label analysis, while specific acid-basic properties were selected in all models, confirming that these descriptors are crucial in the prediction of isoform specificity.

5.5.3 ct-SVM and CPG NN models

The prediction results for the Test set 1 were analyzed to compare the performances of the ct-SVM and CPG NN models. The predictivity of ct-SVM for CYP1A2 and CYP3A4 classes is higher than the CPG NN model results. On the other hand, if we consider the remaining isoforms, the values of recall underline the better performances of the CPG NN model. After the comparison of the percentages of correct predictions, the values are similar for the corresponding isoforms.

5.5.4 ct-SVM and SVM models

We compared the prediction results of the multilabel classification ct-SVM model on Test set 1 (209 substrates) and the predictivity of single-label classification SVM model for Test set 2 (191 substrates) to verify whether the multilabel approach might be a valid alternative to the single-label methodology, by applying the same algorithm as modeling method. In this analysis we had to consider that the test sets comprise a different number of compounds, since in Test set 1 eighteen compounds are multilabel.

On first analysis, the recall shows similar performances of the models for CYP1A2 class and an increase of predictivity for CYP3A4 class in the multilabel model. Regarding CYP2C9, CYP2D6 and CYP2E1 classes, the recall drops down in the multilabel classifier. On the other hand, the profile of the precision values reflects better performances of the ct-SVM model if

CYP2D6 and CYP2E1 classes are considered, with the precision values of 0.78 and 0.69, respectively. In the single-label model the values of recall are 0.60 (CYP1A2 class), 0.67 (CYP2C9 class) and 0.85 (CYP3A4 class), higher than the corresponding values in the multilabel classifier. If we analyze the number of TP in the single-label model, 9 out of 14 CYP1A2 substrates, 8 out of 11 CYP2C9 substrates, 41 out of 53 CYP2D6 substrates, 11 out of 11 CYP2E1 substrates and 80 out of 102 CYP3A4 substrates resulted. The number of correctly predicted compounds using the single-label approach (Table 5.11) is very close in comparison to the number of TP in the multilabel results reported in Table 5.9.

It seems clear that the single-label model is not able to give a complete picture of the metabolism information. In fact, the ct-SVM model performances result at least comparable to the SVM model ones. However, in the single-label approach, inevitably, we lose important details about isoform specificity, since each substrate is implicitly supposed to be metabolized by an unique CYP450 isoform.

5.5.5 Analysis of some classified compounds

We compared the prediction results of the ct-SVM, CPG NN and SVM models on the Test set 1, including multi- and single-label substrates. Also the multilabel compounds in the Test set 1 were predicted by the single-label model. The CYP450 substrates in Test set 1 not extracted from the Metabolite database were analyzed. [175] The experimental and predicted classes for these compounds are reported in Table 5.12.

Nine out of 44 compounds (two of them are multilabel) are incorrectly predicted by both multilabel ct-SVM and CPG NN models. The single-label SVM model assigned a wrong class to ten compounds and all the multilabel compounds were correctly assigned to one of the experimental classes by ct-SVM and CPG NN models. Regarding multilabel compounds, most ct-SVM predictions are correct even if partial. Two examples are Clomipramine and Methadone. The drug Clomipramine (**173**) is metabolized by four different CYP450 isoforms and the ct-SVM model correctly predicted three of them (CYP1A2, CYP2D6 and CYP3A4), while both CPG NN and SVM classifiers have only assigned one class. Similarly, Methadone (**189**) is predicted by the multilabel models to be metabolized by CYP2D6 and CYP3A4 isoforms, showing a good correspondence with the experimental metabolic profile.

Table 5.12: Some experimental and predicted isoforms after applying *ct-SVM*, *CPG NN* and *SVM* models are summarized. *Ct-SVM* and *CPG NN* models are multilabel; *SVM* model was carried out by using the single-label approach. In the second column the multilabel substrates are bold; *M* = multi-label; *S* = single-label.; 1A2 = CYP1A2; 2C9 = CYP2C9; 2D6 = CYP2D6; 2E1 = CYP2E1; 3A4 = CYP3A4.

No.	Name	Exp. classes	Pred. ct-SVM (M)	Pred. CPG NN (M)	SVM (S)
166	Acetaminophen	1A2	1A2	1A2	1A2
167	Alpidem	3A4	3A4	3A4	3A4
168	Amiflamine	2D6	1A2	2D6	2D6
169	Aripiprazole	2D6 3A4	3A4	3A4	3A4
170	Azatadine	3A4	3A4	3A4	3A4
171	Bufuralol	2D6	2D6	2D6	2D6
172	Cinnarizine	2D6	2D6	2D6	2D6
173	Clomipramine	1A2 2C9 2D6 3A4	1A2 2D6 3A4	2D6	2D6
174	Clopidogrel	1A2 3A4	2D6	2D6	3A4
175	Deprenyl	2D6	3A4	2D6 3A4	2E1
176	Desipramine	1A2 2D6	2D6	2D6	2D6
177	Dihydrocodeine	2D6 3A4	2D6 3A4	2D6	2D6
178	Ebastine	3A4	3A4	3A4	3A4
179	Enalapril	3A4	3A4	2C9	1A2
180	Fluconazole	3A4	2C9 3A4	3A4	1A2
181	Flunarizine	2D6	2D6	2D6	2D6
182	Formoterol	2C9 2D6	2D6	2D6	2D6
183	Indomethacin	2C9	3A4	2C9	2C9
184	Levonorgestrel	3A4	3A4	3A4	3A4
185	Lidocaine	2D6 3A4	1A2	2D6 3A4	2D6
186	Lisuride	3A4	1A2 3A4	3A4	3A4
187	Lobeline	2D6	2D6 3A4	1A2 2D6 3A4	3A4
188	Lornoxicam	2C9	2C9	2C9	2C9
189	Methadone	1A2 2D6 3A4	2D6 3A4	2D6 3A4	3A4
190	Methoxyphenamine	2D6	2D6	2D6	2D6
191	Mexiletine	1A2 2D6	2D6	2D6	2D6
192	Montelukast	2C9 3A4	3A4	3A4	3A4
193	Omeprazole	2C9 3A4	3A4	1A2	3A4

(continued on next page)

No.	Name	Exp. classes	Pred. ct-SVM (M)	Pred. CPG NN (M)	SVM (S)
194	Ondansetron	1A2 2D6 3A4	1A2	3A4	3A4
195	Phenylbutazone	2C9	1A2	1A2	3A4
196	Quercetin	3A4	1A2	1A2	1A2
197	Ramelteon	1A2 2C9 3A4	3A4	2C9	3A4
198	Remoxipride	2D6	3A4	3A4	3A4
199	Sertindole	3A4	3A4	3A4	3A4
200	Sparteine	2D6	2D6	2D6	2D6
201	Sulfamethizole	2C9	2C9	2C9	3A4
202	Sulfidimidine	3A4	3A4	2C9	3A4
203	Tamsulosin	2D6 3A4	3A4	3A4	3A4
204	Theophylline	1A2 2E1	1A2 3A4	1A2	1A2
205	Tolterodine	2D6 3A4	2D6 3A4	2D6 3A4	2D6
206	Trimethoprim	2C9	3A4	3A4	3A4
207	Valdecoxib	2C9 3A4	2C9	2C9	3A4
208	Zidovudine	3A4	3A4	3A4	1A2
209	Zileuton	1A2 2C9 3A4	3A4	2E1	1A2

These examples confirm that the multilabel approach is able to perform an extensive investigation of the drug metabolism, while the single-label results are limited to the prediction of a single class. A deep analysis of each compound in Table 5.12 is reported in *Paper III*.

5.6 Final remarks

In the present study, we investigated several classification strategies to predict the isoform specificity of known CYP1A2, CYP2C9, CYP2D6, CYP2E1 and CYP3A4 substrates. The multilabel approach was applied to a data set including five classes, by using the ct-SVM and the CPG NN methods. The best model (ct-SVM) was derived after the selection of 27 descriptors and yielded 77.5-96.6% of correct predictions for the five classes of the corresponding test set. Similarly, the CPG NN model achieved 75.6-97.1% of correct predictions. A five-classes data set was used to perform an extensive single-label classification analysis, in combination with automatic variable selection. The highest predictivity on the corresponding test set, achieved by using the SVM technique based on nineteen descriptors, was 78% of cor-

rect predictions. All the presented models show acceptable performances, however the multilabel prediction results reflect more coherently the real metabolic fate of drugs.

In conclusion, our results underline the high complexity of this classification problem and suggest the application of the multilabel approach to predict CYP450 isoform specificity. The advantage of the CPG NN technique is the graphical visualization of the results. Both ct-SVM and the CPG NN strategies might be extended to quantitative data. The multilabel methodology can be used to explore the metabolic profile of new chemical entities and the prediction capability might be improved by collecting other multilabel substrates in the dataset.

Classification and regression to investigate selectivity and binding affinity

The selectivity is an important aspect of drug discovery, and in the development of G protein-coupled receptors (GPCRs) ligands to distinguish between related receptor subtypes is often the key to therapeutic success. Nowadays, the prediction of receptor subtype selectivity represents a very challenging task. In the present study, we present an alternative application of Support Vector Machine (SVM) and Support Vector Regression (SVR) methodologies to simultaneously describe both A_{2A}R versus A₃R subtypes selectivity profile and the corresponding receptor binding affinities. We have implemented an integrated application of SVM-SVR approach, based on the use of the autocorrelated molecular descriptors encoding for the Molecular Electrostatic Potential (*auto*MEP), to simultaneously discriminate A_{2A}R versus A₃R antagonists and to predict the binding affinity to the corresponding receptor subtype of a large dataset of known pyrazolo-triazolo-pyrimidine analogs. To validate our approach, we have synthesized 51 new pyrazolo-triazolo-pyrimidine derivatives anticipating both A_{2A}R/A₃R subtypes selectivity and receptor binding affinity profiles.

6.1 Introduction

G protein-coupled receptors (GPCRs) represent the target for many drugs under development. The efficacy problems and limiting side-effects of some candidates are due to the lack of differentiation between receptor subtypes. There is thus considerable interest in attaining therapeutic selectivity by identifying the single receptor subtype that affects a particular physiology. The goal is to reduce, as more as possible, the side-effects, while retaining the desired function. To date, very few valuable computational tools are available for the prediction of receptor subtype selectivity, which is still considered a complex problem. Conversely, different *in silico* approaches are accessible to estimate the distinct receptor-ligand affinity, in particular QSAR is the commonly used approach in this field. [176, 177]

In the last few years, the possibility to discover new potent and selective adenosine receptors (ARs) antagonists has been intensively explored. Briefly, the adenosine receptor (AR) family belongs to GPCR family A, including four different subtypes, referred to as A_1 , A_{2A} , A_{2B} and A_3 , which are widely but differentially distributed throughout the body¹. [139, 142] In this study we will focus on $A_{2A}R$ and A_3R subtypes and selective ligands to these ARs are becoming increasingly attractive drugs due to their potential role of this receptor in several physiopathological processes. [140, 141, 144, 178, 179] In particular, $A_{2A}R$ antagonists seem to play a role in the reduction of neuronal damage in Parkinson's or Huntington's diseases, while A_3R antagonists have a potential application in the tumor growth inhibition and in the treatment of glaucoma. [142-144]

Consequently, several receptor-based and ligand-based drug design approaches have been carried out with the aim to improve potency and selectivity of different molecular scaffolds and, in particular, the pyrazolo-triazolo-pyrimidine scaffold has been extensively studied. Moreover, it has been demonstrated that proper substitutions at the N^5 and N^8 positions drive the antagonist selectivity to the human A_3R subtype. [149] On the other hand, the substitution at the position N^7 shifts the selectivity profile to the human $A_{2A}R$ subtype. [146] However, this very empirical rule based on experimental evidences does not provide a criterion to assign the correct pharmacological A_{2A} and A_3 receptors profiles of novel pyrazolo-triazolo-pyrimidine derivatives.

¹Further details on the four human AR subtypes are reported in section 7.2.

Here, we describe an alternative application of the Support Vector Machine (SVM) and Support Vector Regression (SVR) methodologies to predict both A_{2A}R versus A₃R subtypes selectivity profile and the corresponding receptor binding affinities. As anticipated, SVM is very utilized to solve both classification and regression problems. [82, 85] In this study, we have implemented an integrated application of SVM-SVR approach, based on the use of *autoMEP* descriptors, to simultaneously discriminate A_{2A}R versus A₃R antagonists and to predict the binding affinity to the corresponding receptor subtype of a large dataset of known pyrazolo-triazolo-pyrimidine analogs. To validate our approach, we have newly synthesized 51 pyrazolo-triazolo-pyrimidine derivatives anticipating both A_{2A}R/A₃R subtypes selectivity and receptor binding affinity profiles.

6.2 Dataset

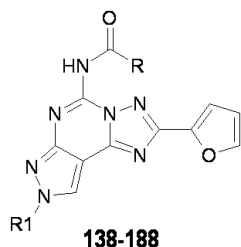
A collection of 104 selective N⁷- and N⁸-substituted pyrazolo-triazolo-pyrimidine analogues (molecules **1-104**) has been selected as training set in the first SVM classification (*SVMclass*) model. [146-151, 180]

In the SVR regression (SVR) model 104 N⁸-substituted pyrazolo-triazolo-pyrimidine derivatives (molecules **1-71**, **105-137**), selective and not selective, have been used as training set of both human A_{2A}R and A₃R nonlinear SVR models. [146, 149, 150, 180, 181]

Finally, a test set of 51 N⁸-substituted pyrazolo-triazolo-pyrimidine analogues (molecules **138-188**) has been selected to validate both *SVMclass* and SVR models² (Table 6.1).

²Experimental binding affinity data kindly provided by the work coordinated by Prof. G. Spalluto (University of Trieste) for the synthesis and by Prof. K. N. Klotz (University of Würzburg) for the pharmacological characterization.

Table 6.1: Biological profile at the $hA_{2A}R$ and hA_3R subtypes of the test set compounds.



No.	R	R1	$hA_{2A}R$ (K_i , nM) ^a	hA_3R (K_i , nM) ^b
138	CH ₂ - α -naphthyl	CH ₂ CH ₂ CH ₂ Ph	305	343
139	CHPh ₂	CH ₃	216	0.25
140	CH ₂ -Ph-Ph	CH ₃	193	11.2
141	CHPh ₂	CH ₂ CH ₂ CH(CH ₃) ₂	159	5.86
142	CH ₂ -Ph-Ph	CH ₂ CH ₂ CH(CH ₃) ₂	9.18	268
143	CHPh ₂	CH ₂ CH ₂ Ph	53.1	6.49
144	CH ₂ -Ph-Ph	CH ₂ CH ₂ Ph	46.9	125
145	CHPh ₂	CH ₂ CH ₂ CH ₂ CH ₃	114	1.20
146	CH ₂ -Ph-Ph	CH ₂ CH ₂ CH ₂ CH ₃	39.6	189
147	CH ₂ -3-Cl-Ph	CH ₃	30.5	1.94
148	CHPh ₂	CH ₂ CH ₂ CH ₃	46.5	0.93
149	CH ₂ -Ph-Ph	CH ₂ CH ₂ CH ₃	52.5	65.4
150	CH ₂ -4-Cl-Ph	CH ₃	156	12.7
151	CH ₂ -Ph-Ph	CH ₂ CH ₃	70.9	14.2
152	CH ₂ -4-OCH ₃ -Ph	CH ₃	62.5	0.95
153	CH ₂ - β -naphthyl	CH ₂ CH ₂ CH(CH ₃) ₂	15.7	409
154	CH ₂ - β -naphthyl	CH ₂ CH ₂ Ph	42.1	180
155	CH ₂ -2-thienyl	CH ₂ CH ₂ Ph	8.8	330
156	CH ₂ -3-thienyl	CH ₂ CH ₂ Ph	12.9	726
157	CH ₂ - α -naphthyl	CH ₂ CH ₃	74.6	3.05
158	CH ₂ - β -naphthyl	CH ₂ CH ₃	18.4	36.3
159	CH ₂ -3-thienyl	CH ₂ CH ₃	23	765
160	CH ₂ -2-thienyl	CH ₂ CH ₃	15.9	196
161	CH ₂ -2-thienyl	CH ₃	56	5.26
162	CH ₂ -3-thienyl	CH ₃	31.3	1.25
163	CH ₂ - β -naphthyl	CH ₃	77.5	14.5
164	CH ₂ - α -naphthyl	CH ₃	80.5	3.47
165	CH ₂ - α -naphthyl	CH ₂ CH ₂ CH ₃	38.7	17.3

^aDisplacement of specific [3H]-NECA binding at human A_{2A} receptors expressed in CHO cells;

^bdisplacement of specific [3H]-NECA binding at human A_3 receptors expressed in CHO cells.

(continued on next page)

No.	R	R1	hA _{2A} (K_i , nM) ^a	hA ₃ (K_i , nM) ^b
166	CH ₂ - β -naphthyl	CH ₂ CH ₂ CH ₃	7.99	95.9
167	CH ₂ - α -naphthyl	CH ₂ CH ₂ CH ₂ CH ₃	11.7	100
168	CH ₂ -4-CF ₃ -Ph	CH ₃	75.9	1.22
169	CH ₂ -2-thienyl	CH ₂ CH ₂ CH ₂ CH ₃	4.15	99.8
170	CH ₂ -3-thienyl	CH ₂ CH ₂ CH ₂ CH ₃	3.13	189
171	CH ₂ -O-Ph-4-Cl	CH ₃	39.3	223
172	CH ₂ -3-Cl-Ph	CH ₂ CH ₂ CH(CH ₃) ₂	1.86	273
173	CH ₂ -4-Cl-Ph	CH ₂ CH ₂ CH(CH ₃) ₂	2.75	56.5
174	CH ₂ -3-Cl-Ph	CH ₂ CH ₂ Ph	5.75	273
175	CH ₂ -4-CF ₃ -Ph	CH ₂ CH ₂ CH(CH ₃) ₂	5.43	266
176	CH ₂ -4-F-Ph	CH ₂ CH ₂ CH(CH ₃) ₂	3.69	116
177	CH ₂ -4-F-Ph	CH ₃	54.1	0.97
178	CH ₂ -2,6-Cl ₂ -Ph	CH ₂ CH ₂ CH(CH ₃) ₂	18.7	207
179	CH ₂ -2,6-Cl ₂ -Ph	CH ₃	45.2	44.4
180	CH ₂ -4-F-Ph	CH ₂ CH ₂ CH ₂ Ph	211	58.5
181	CH-Ph ₂	CH ₂ CH ₂ CH ₂ Ph	326	12.6
182	CH ₂ - β -naphthyl	CH ₂ CH ₂ CH ₂ Ph	73.6	717
183	CH ₂ Ph	CH ₂ CH ₂ CH ₂ Ph	43.9	5.49
184	CH ₂ -3-Cl-Ph	CH ₂ CH ₂ CH ₂ Ph	182	110
185	CH ₂ -4-Cl-Ph	CH ₂ CH ₂ CH ₂ Ph	89.9	30.5
186	CH ₂ -O-Ph-4-Cl	CH ₂ CH ₂ CH ₂ Ph	27.2	400
187	CH ₂ -2,6-Cl ₂ -Ph	CH ₂ CH ₂ CH ₂ Ph	186	601
188	CH ₂ -Ph-Ph	CH ₂ CH ₂ CH ₂ Ph	256	410

^aDisplacement of specific [3H]-NECA binding at human A_{2A} receptors expressed in CHO cells;

^bdisplacement of specific [3H]-NECA binding at human A₃ receptors expressed in CHO cells.

6.3 Results and discussion

Support Vector Machine represents a group of supervised learning techniques, which find now diverse applications in classification and regression problems. SVM has been originally developed for classification, then the introduction of a suitable ε -insensitive loss function together with the advantages of the kernel representation have enabled its application in the regression analysis, as reported in chapter 2.6. Recently, the application of SVM and SVR approaches has helped to solve several classification problems, as for example active and non active compounds discrimination, and to derive QSARs for the prediction of different chemical and biological properties. [89-92, 100, 103] SVR seems to be a promising tool, with good generalization performance and increased robustness compared with the neural networks.

We consider both topological and electrostatic complementarities extremely crucial in describing the receptor subtypes selectivity. Basing on the motivations underlined in section 4.3, we believe that the *autoMEP* vectors can be used as interesting molecular descriptors. We have also reported that pyrazolo-triazolo-pyrimidine is a versatile scaffold to cover a large spectrum of the adenosine receptor selectivity. As anticipated, pyrazolo-triazolo-pyrimidines bearing specific substitutions at the N⁵ and N⁸ positions have been described as highly potent and selective human A₃R antagonists, while the position N⁷ shifts the selectivity profile to the human A_{2A}R subtype. [146, 149] However, the observation of the scaffold and its substitutions is not an unailing strategy to correctly assign the selectivity profile of new pyrazolo-triazolo-pyrimidine antagonists.

In this chapter we present an integrated approach based on the introduction of two distinct support vector machine tools both using as input matrix the *autoMEP* vectors. The first tool is a SVM-driven selectivity classifier and the second one is a couple of SVR-driven receptor-affinity predictors. The workflow of the abovementioned procedure is summarized in Figure 6.3.1.

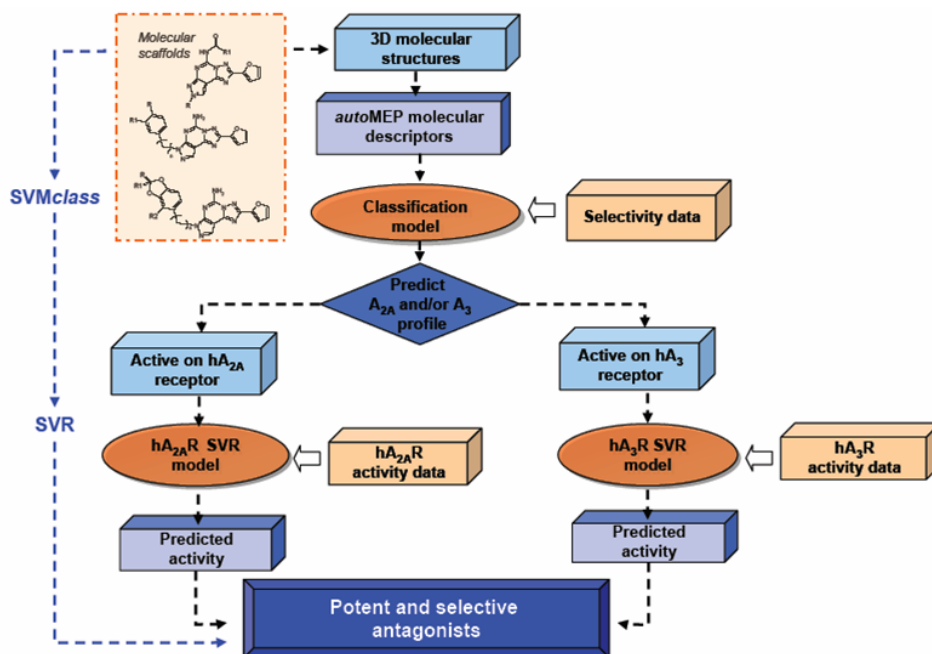


Figure 6.3.1: Flowchart of the in series *autoMEP/SVMclass* and *autoMEP/SVR* approach for the selection of new selective and potent human A_{2A}R and A₃R antagonists.

We have introduced the *autoMEP* of each pyrazolo-triazolo-pyrimidine antagonist to optimize the experimental A_{2A}R/A₃R subtypes selectivity profile using a SVM classifier (*autoMEP/SVMclass*). Then, the two different A_{2A}R and A₃R receptor-affinity predictors can be generated using the *autoMEP* vectors as input values. The application of the *autoMEP/SVMclass* model ahead of the two receptor-affinity predictors can refine the prediction of both A_{2A}R/A₃R subtypes selectivity profile and A_{2A}R/A₃R binding affinity values of new pyrazolo-triazolo-pyrimidine derivatives.

6.3.1 SVM classification model

To build our SVM-driven selectivity classifier, we have selected 104 "selective" pyrazolo-triazolo-pyrimidines derivatives (molecules **1-104**). In more detail, 48.1% of the *SVMclass* model training set include hA_{2A}R antagonists (50 compounds) and the remaining percentage (51.9%) is composed of hA₃R antagonists (54 compounds). The definition of "selectivity" used for our classification approach is based on a simple binary criteria: the selectivity index is set to "+1" if it is referred to a selective hA_{2A}R antagonist, and is "-1" for a selective hA₃R antagonists. Moreover, we have considered as selectivity threshold a difference of at least 2 orders of magnitude between the corresponding K_i values.

Information encoded by twelve *autoMEP* vectors (calculated by selecting default parameters, as described in chapter 2) has been used as input matrix for our *SVMclass* model. The best classifier was obtained by using a Gaussian radial basis function kernel ($C = 150$; $\gamma = 0.01$) and this model has been subjected to an extensive n -fold cross-validation procedure (Table 6.2).

Table 6.2: *AutoMEP/SVMclass* model; the statistical parameters after the cross-validation procedure on the selected classifier are collected.

Partition	CV	% correct predictions				
		No. of runs	Mean	StDev	Min	Max
Training set		1	99.0	-	99.0	99.0
Training set	LOO	1	93.3	-	93.3	93.3
	10-fold	10	91.4	2.5	86.5	95.2
	5-fold	10	91.7	2.1	88.5	94.2
Test set		1	78.4	-	78.4	78.4

The percentages obtained after repeated 10-fold and 5-fold cross-validation processes confirm the statistical reliability of this model. Interestingly, it yielded 93.3% correct predictions after LOO cross-validation. The statistical robustness is also confirmed by the percentage (%) values of sensitivity (92.0%) and specificity (94.4%). Therefore, we decided to select this model as SVM-driven selectivity classifier for the final validation step.

6.3.2 SVM regression models

A different collection of 104 pyrazolo-triazolo-pyrimidine analogs (molecules **1-71**, **105-137**) has been selected as training set of known hA_{2A}R and hA₃R antagonists to derive our *auto*MEP/SVR models. The selection of all training set candidates was not performed according to a selectivity criterion, due to the fact that the principal aim of our regression model is to accurately predict the receptor binding affinity. Indeed we have utilized 17 hA_{2A}R selective antagonists (16.4%), 54 hA₃R selective antagonists (51.9%) and 33 non selective antagonists (31.7%). Also in this step, information encoded by twelve *auto*MEP vectors of all training set antagonists has been used as input matrix. For the generation of both regression models, we have utilized a Gaussian radial basis function kernel.

An acceptable hA_{2A}R SVR model ($C = 200$, $\varepsilon = 0.0005$, $\gamma = 0.0006$) was obtained as indicated by the LOO cross-validated correlation coefficient (r_{cv}) of 0.78 and a root mean square of residuals (RMSR) of 0.050, as reported in Figure 6.3.2 and Table 6.3.

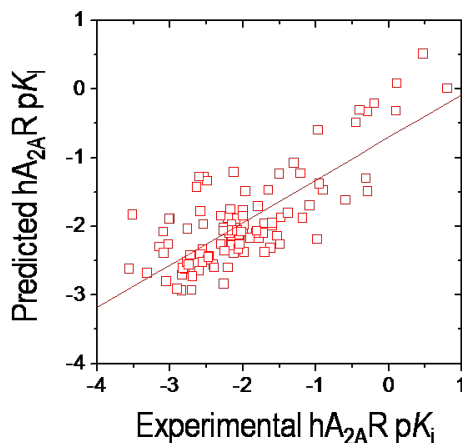


Figure 6.3.2: AutoMEP/SVR $hA_{2A}R$ model; experimental pK_i values vs predicted pK_i values after LOO cross-validation on the training set.

Table 6.3: Statistical parameters of the autoMEP/SVR $hA_{2A}R$ model.

Number of molecules	104
r	0.83
r_{cv}^a	0.78
Slope	0.62
Offset	-0.71
q^b	0.82
RMSR ^c	0.050

^aCross-validated r after LOO cross-validation procedure: $r_{cv} = [SXY / (SXX)^{\frac{1}{2}} (SYY)^{\frac{1}{2}}]$, $SXY = \sum (X - X_{mean})(Y - Y_{mean})$, $SXX = \sum (X - X_{mean})^2$ and $SYY = \sum (Y - Y_{mean})^2$ with $X = Y_{Experimental}$ and $Y = Y_{Predicted}$; ^b r of the internal test set; ^croot mean square of residuals: RMSR.

On the other hand, the best hA_3R SVR model ($C = 150$, $\varepsilon = 0.3$, $\gamma = 0.005$) was derived as described by the LOO cross-validated correlation coefficient (r_{cv}) of 0.85 and a root mean square of residuals (RMSR) of 0.046 (Figure 6.3.3 and Table 6.4). The results of SVR analysis are noteworthy considering that the same training set was used to generate two different robust models. The validation of SVR models is discussed in the following paragraph.

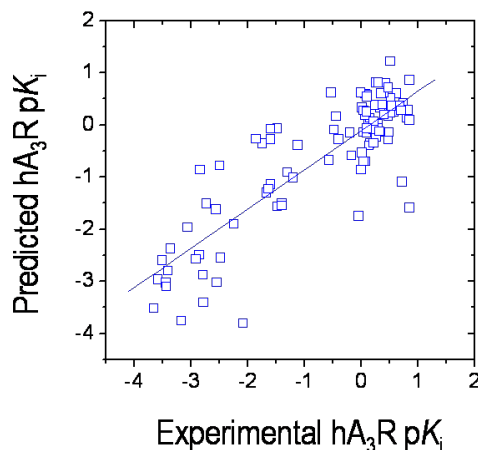


Figure 6.3.3: *AutoMEP/SVR hA₃R model; experimental pK_i values vs predicted pK_i values after LOO cross-validation on the training set.*

Table 6.4: *Statistical parameters of the autoMEP/SVR hA₃R model.*

Number of molecules	104
r	0.95
r_{cv}^a	0.85
Slope	0.75
Offset	-0.11
q^b	0.85
RMSR ^c	0.046

^aCross-validated r after LOO cross-validation procedure: $r_{cv} = [SXY / (SXX)^{\frac{1}{2}} (SYY)^{\frac{1}{2}}]$, $SXY = \sum (X - X_{mean})(Y - Y_{mean})$, $SXX = \sum (X - X_{mean})^2$ and $SYY = \sum (Y - Y_{mean})^2$ with $X = Y_{Experimental}$ and $Y = Y_{Predicted}$; ^b r of the internal test set; ^croot mean square of residuals: RMSR.

6.3.3 Validation of in series SVMclass and SVR models

As anticipated, the principal aim of the present work has been to evaluate the robustness of the tandem *autoMEP/SVMclass* and *autoMEP/SVR* models in the prediction of both A_{2A}R/A₃R subtypes selectivity and receptor binding affinity profiles of new pyrazolo-triazolo-pyrimidine derivatives, and in particular 51 analogs (molecules **138-188**) were considered (Table 6.5).

Table 6.5: Predicted and experimental $hA_{2A}R$ and hA_{3R} both pK_i and K_i for the test set. Differences between predicted and experimental pK_i values for both SVR models ($hA_{2A}R$ and hA_{3R}) are reported. In the middle column $hA_{2A}R$ selective antagonists (red) and hA_{3R} selective antagonists (blue) are highlighted.

Pred. pK_i (nM) $hA_{2A}R$	Exp. pK_i (nM) $hA_{2A}R$	Δ pK_i^*	Pred. K_i (nM) $hA_{2A}R$	Exp. K_i (nM) $hA_{2A}R$	No.	Exp. K_i (nM) hA_{3R}	Pred. K_i (nM) hA_{3R}	Δ pK_i^*	Exp. pK_i (nM) hA_{3R}	Pred. pK_i (nM) hA_{3R}
-2.25	-2.48	0.23	177.83	305	138	343	107.15	0.51	-2.54	-2.03
-1.79	-2.33	0.54	61.66	216	139	0.25	0.47	-0.27	0.60	0.33
-2.04	-2.29	0.25	109.65	193	140	11.2	35.48	-0.50	-1.05	-1.55
-2.07	-2.20	0.13	117.49	159	141	5.86	2.88	0.31	-0.77	-0.46
-1.05	-0.96	-0.09	11.22	9.18	142	268	144.54	0.27	-2.43	-2.16
-1.47	-1.73	0.26	29.51	53.1	143	6.49	1.70	0.58	-0.81	-0.23
-1.18	-1.67	0.49	15.14	46.9	144	125	70.79	0.25	-2.10	-1.85
-1.85	-2.06	0.21	70.79	114	145	1.20	0.56	0.33	-0.08	0.25
-1.21	-1.60	0.39	16.22	39.6	146	189	51.29	0.57	-2.28	-1.71
-1.59	-1.48	-0.11	38.90	30.5	147	1.94	2.24	-0.06	-0.29	-0.35
-1.87	-1.67	-0.20	74.13	46.5	148	0.93	0.60	0.19	0.03	0.22
-1.15	-1.72	0.57	14.13	52.5	149	65.4	44.67	0.17	-1.82	-1.65
-1.67	-2.19	0.52	46.77	156	150	12.7	6.92	0.26	-1.10	-0.84
-1.41	-1.85	0.44	25.70	70.9	151	14.2	72.44	-0.71	-1.15	-1.86
-1.78	-1.80	0.02	60.26	62.5	152	0.95	3.39	-0.55	0.02	-0.53
-1.32	-1.20	-0.12	20.89	15.7	153	409	85.11	0.68	-2.61	-1.93
-1.12	-1.62	0.50	13.18	42.1	154	180	46.77	0.59	-2.26	-1.67
-1.41	-0.94	-0.47	25.70	8.8	155	330	269.15	0.09	-2.52	-2.43
-1.58	-1.11	-0.47	38.02	12.9	156	726	707.95	0.01	-2.86	-2.85
-1.65	-1.87	0.22	44.67	74.6	157	3.05	11.48	-0.58	-0.48	-1.06
-1.52	-1.26	-0.26	33.11	18.4	158	36.3	23.99	0.18	-1.56	-1.38
-1.73	-1.36	-0.37	53.70	23	159	765	489.7	0.19	-2.88	-2.69
-1.52	-1.20	-0.32	33.11	15.9	160	196	186.21	0.02	-2.29	-2.27
-1.64	-1.75	0.11	43.65	56	161	5.26	1.12	0.67	-0.72	-0.05
-1.71	-1.50	-0.21	51.29	31.3	162	1.25	3.55	-0.45	-0.10	-0.55
-1.63	-1.89	0.26	42.66	77.5	163	14.5	3.55	0.61	-1.16	-0.55
-1.79	-1.91	0.12	61.66	80.5	164	3.47	2.51	0.14	-0.54	-0.40
-1.69	-1.59	-0.10	48.98	38.7	165	17.3	67.61	-0.59	-1.24	-1.83
-0.99	-0.90	-0.09	9.77	7.99	166	95.9	57.54	0.22	-1.98	-1.76
-1.48	-1.07	-0.41	30.20	11.7	167	100	61.66	0.21	-2.00	-1.79

*Predicted pK_i - Experimental pK_i .

(continued on next page)

Pred. pK_i (nM)	Exp. pK_i (nM)	Δ pK_i^*	Pred. K_i (nM)	Exp. K_i (nM)	No.	Exp. K_i (nM)	Pred. K_i (nM)	Δ pK_i^*	Exp. pK_i (nM)	Pred. pK_i (nM)
hA _{2A} R	hA _{2A} R		hA _{2A} R	hA _{2A} R		hA ₃ R	hA ₃ R		hA ₃ R	hA ₃ R
-1.76	-1.88	0.12	57.54	75.9	168	1.22	18.20	-1.17	-0.09	-1.26
-1.41	-0.62	-0.79	25.70	4.15	169	99.8	173.78	-0.24	-2.00	-2.24
-0.76	-0.50	-0.26	5.75	3.13	170	189	22.39	0.93	-2.28	-1.35
-1.06	-1.59	0.53	11.48	39.3	171	223	52.48	0.63	-2.35	-1.72
-0.23	-0.27	0.04	1.70	1.86	172	273	97.72	0.45	-2.44	-1.99
0.11	-0.44	0.55	0.78	2.75	173	56.5	125.89	-0.35	-1.75	-2.10
-1.23	-0.76	-0.47	16.98	5.75	174	273	169.82	0.21	-2.44	-2.23
-0.86	-0.73	-0.13	7.24	5.43	175	266	89.13	0.47	-2.42	-1.95
-0.14	-0.57	0.43	1.38	3.69	176	116	173.78	-0.18	-2.06	-2.24
-1.93	-1.73	-0.20	85.11	54.1	177	0.97	14.45	-1.17	0.01	-1.16
-0.85	-1.27	0.42	7.08	18.7	178	207	112.20	0.27	-2.32	-2.05
-1.32	-1.66	0.34	20.89	45.2	179	44.4	8.51	0.72	-1.65	-0.93
-1.96	-2.32	0.36	91.20	211	180	58.5	223.87	-0.58	-1.77	-2.35
-2.34	-2.51	0.17	218.78	326	181	12.6	2.82	0.65	-1.10	-0.45
-1.82	-1.87	0.05	66.07	73.6	182	717	524.81	0.14	-2.86	-2.72
-1.33	-1.64	0.31	21.38	43.9	183	5.49	14.79	-0.43	-0.74	-1.17
-2.07	-2.26	0.19	117.49	182	184	110	245.47	-0.35	-2.04	-2.39
-1.26	-1.95	0.69	18.20	89.9	185	30.5	213.80	-0.85	-1.48	-2.33
-1.29	-1.43	0.14	19.50	27.2	186	400	602.56	-0.18	-2.60	-2.78
-2.48	-2.27	-0.21	302.00	186	187	601	1258.93	-0.32	-2.78	-3.10
-2.18	-2.41	0.23	151.36	256	188	410	48.98	0.92	-2.61	-1.69

*Predicted pK_i - Experimental pK_i .

The experimental A_{2A}R and A₃R binding affinities are collected in Table 6.1. Moreover, 19.6% of the compounds are selective hA_{2A}R (molecules **142**, **153**, **155**, **156**, **159**, **170**, **172**, **174-176** in Table 6.5) and 9.8% are selective hA₃R antagonists (molecules **139**, **141**, **145**, **152**, **181** in Table 6.5) in the test set. Following the workflow reported in Figure 6.3.1, the *autoMEP* vectors of these new 51 antagonists have been used as input matrix for the previously generated *autoMEP/SVMclass* model. Our classification model was able to correctly assign the 78.4% of the compounds in the collected test set to their class (Table 6.2). Almost all selective compounds of our test set are correctly classified and only 11 of them (22%) are erroneously recognized. In *Paper IV* additional information on *autoMEP/SVMclass* model predictions are reported.

After passing the selectivity filtering process, each of the hA_{2A}R and hA₃R antagonists has been analyzed by the corresponding SVR binding affin-

ity predictor. The comparison of all the experimental with the predicted pK_i values by the abovementioned $hA_{2A}R$ and hA_3R SVR models on the test set again support the quality of the predictors, as underlined by the good values of the correlation coefficient ($q = 0.82$ and $q = 0.85$, respectively) (Table 6.3 and Table 6.4).

In Figure 6.3.4 and Figure 6.3.5 only the hA_{2A} classified antagonists predicted by the $hA_{2A}R$ SVR model and only the hA_3R classified antagonists predicted by the hA_3R SVR model, respectively, have been separately considered.

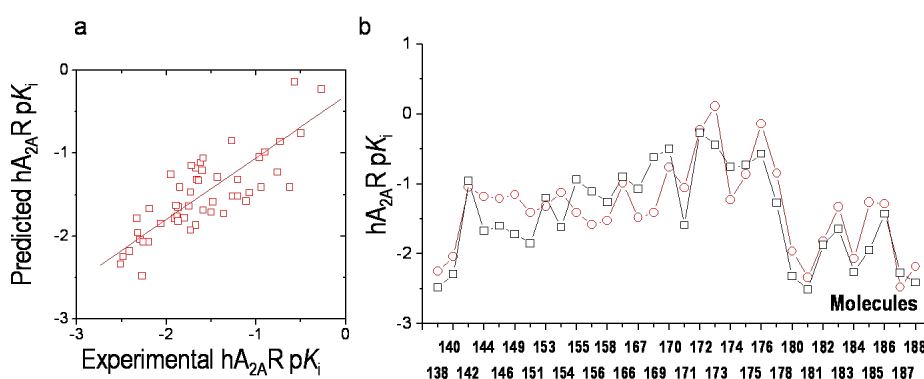


Figure 6.3.4: Test set prediction by autoMEP/SVR $hA_{2A}R$ model. a) Experimental pK_i activity data plotted vs predicted pK_i values; b) experimental pK_i activity data (\square) of the classified selective $hA_{2A}R$ antagonists in the test compared to the pK_i values predicted by autoMEP/SVR $hA_{2A}R$ model (\circ).

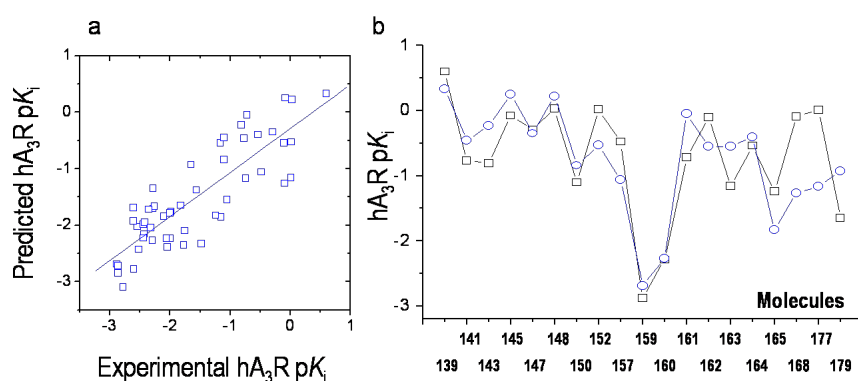


Figure 6.3.5: Test set prediction by autoMEP/SVR hA_3R model. a) Experimental pK_i activity data plotted vs predicted pK_i values; b) experimental pK_i activity data (\square) of the classified selective hA_3R antagonists in the test compared to the pK_i values predicted by autoMEP/SVR hA_3R model (\circ).

The prediction accuracies, as demonstrated by the differences between the experimental and the predicted pK_i values, are statistically satisfactory for both hA_{2A}R and hA₃R SVR models, with few exceptions in particular regarding the hA₃R binding affinity prediction, such as molecules **168**, **170**, **185** and **188** (see Table 6.5, Figure 6.3.4b and Figure 6.3.5b). In *Paper IV* additional considerations on the external test set potency profiles are reported.

6.4 Final remarks

To show the useful application of the machine learning in solving a pharmacodynamic selectivity problem, we have presented a combination of Support Vector Machine tools able to predict both A_{2A}R versus A₃R subtypes selectivity profile and the corresponding receptor binding affinities of a large dataset of known pyrazolo-triazolo-pyrimidine analogs. The preliminary results based on a new set of 51 pyrazolo-triazolo-pyrimidines are very encouraging. To further validate our integrated SVM approach, we are extending the applicability of this method to other classes of hAR antagonists and, at the same time, we are exploring the possibility to describe, using a multi-classifier, the full adenosine receptor selectivity spectrum, as we propose in the following chapter.

Exploring potency and selectivity of hAR antagonists

Nowadays, in medicinal chemistry adenosine receptors (ARs) represent some of the most studied targets, and there is growing interest on the different AR subtypes. The AR subtypes selectivity is highly desired in the development of potent ligands to achieve the therapeutic success. So far, very few ligand-based strategies have been investigated to predict the receptor subtypes selectivity. We have carried out a novel application of the multilabel classification approach by combining the autocorrelated molecular descriptors encoding for the Molecular Electrostatic Potential (*autoMEP*) with Support Vector Machines (SVMs). Three valuable models, based on decreasing thresholds of potency, have been generated as in series quantitative sieves for the simultaneous prediction of the hA₁R, hA_{2A}R, hA_{2B}R and hA₃R subtypes potency profile and selectivity of a large collection, more than 500, of known antagonists such as xanthine and pyrazolo-triazolo-pyrimidine analogs. The robustness and reliability of our multilabel classification models were assessed by predicting an internal test set. Finally, we have applied our strategy to 13 newly synthesized pyrazolo-triazolo-pyrimidine derivatives inferring their full adenosine receptor potency spectrum and hAR subtypes selectivity profile.

7.1 Introduction

Adenosine receptors (ARs) are widely considered interesting and promising therapeutic targets. In the last decade, the growing knowledge about the different adenosine receptor subtypes has inspired the development of potent and selective ligands. [142, 182] During the optimization step of drug discovery process, the general aim is to design drugs more effective in the therapeutic treatment, but with minimum side-effects. If compounds do not differentiate between receptor subtypes, their therapeutic application might be accompanied by efficacy problems or side-effects. Therefore, after identifying the single receptor subtype that is responsible for a particular function, the drug candidates may be sifted out according to criteria of high potency profile and subtype selectivity.

The detection of selective compounds by using *in silico* tools represents a difficult task and to date, not many examples of selectivity prediction have been described in the literature. [183, 184] Only few pioneer studies suggest an integration of both traditional classification and regression analysis as useful filtering strategy to select potent and selective ligands, as previously described in chapter 6.

We would like to demonstrate how a novel application of the multilabel classification approach by combining the autocorrelated molecular descriptors encoding for the Molecular Electrostatic Potential (*auto*MEP) vectors with Support Vector Machine (SVM) analysis can represent a very powerful tool to simultaneously describe the hA₁R, hA_{2A}R, hA_{2B}R and hA₃R potency profiles and identify the possible subtype selectivity for hAR antagonists.

In the previous chapter, we have developed an integrated SVM-SVR method by using the *auto*MEP molecular descriptors to discriminate A_{2A}R versus A₃R antagonists and to predict the binding affinity to the corresponding receptor subtype. However, in the traditional single-label classification classes are considered mutually exclusive. In our classification task some samples belong to multiple classes, since the hAR antagonists may present a good potency profile for more subtypes. The multilabel classification analysis seems to be appropriate, whereas our dataset deals with non-mutually exclusive and overlapping classes. In this field a novel multilabel classification technique, cross-training with Support Vector Machines (ct-SVM), has been recently proposed. [108, 109]

In the present study, the combination of *autoMEP* vectors with ct-SVM analysis (*autoMEP*/ct-SVM) represents a novel strategy for the prediction of the complete hARs potency profiles and infer hAR subtypes selectivity of known xanthine and pyrazolo-triazolo-pyrimidine analogs. Interestingly, our *autoMEP*/ct-SVM approach has been extended to all four hAR subtypes. In more detail, a large collection of hAR antagonists has been utilized to carry out and validate three *autoMEP*/ct-SVM models. They have been applied in series as quantitative sieves, based on decreasing thresholds of potency (500 nM, 250 nM and 100 nM), corresponding to different binding affinity K_i values. For the further validation of our strategy, we have synthesized 13 new pyrazolo-triazolo-pyrimidine derivatives to inspect their A_{1R} , A_{2AR} , A_{2BR} and A_{3R} potency profiles.¹ The workflow applied in our analysis is represented in Figure 7.1.1.

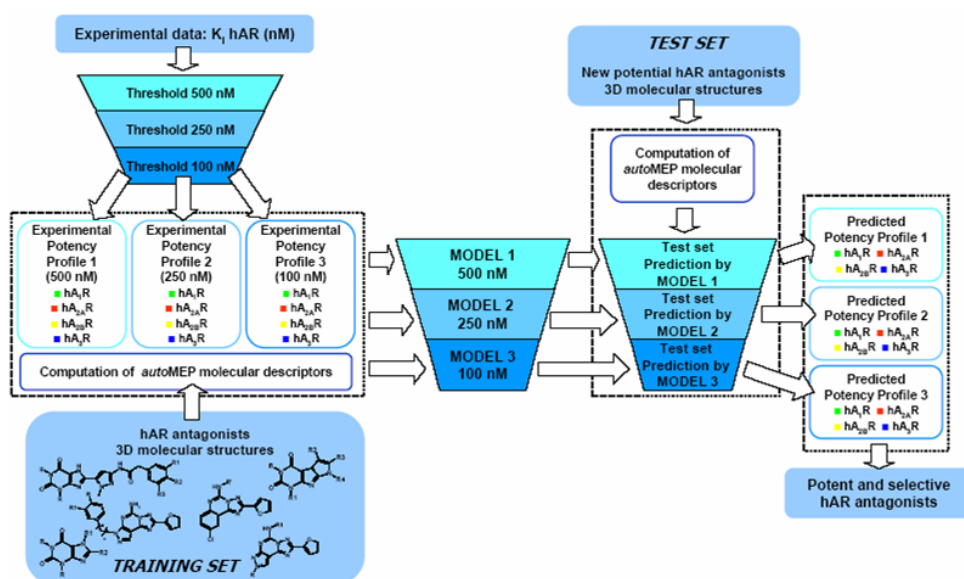


Figure 7.1.1: Workflow for the generation of *autoMEP*/ct-SVM multilabel classification models and the prediction of the hAR antagonists potency profiles.

¹Experimental binding affinity data kindly provided by the work coordinated by Prof. G. Spalluto (University of Trieste) for the synthesis and by Prof. K. N. Klotz (University of Würzburg) for the pharmacological characterization.

7.2 Adenosine receptor antagonists

In the last few years an intensive exploration of the chemical space has been pursued to discover new highly potent and selective adenosine receptors (ARs) antagonists. As anticipated, the adenosine receptor family belongs to GPCR (G protein-coupled receptors) family A, including four different subtypes, referred to as A_1 , A_{2A} , A_{2B} and A_3 , which are widely but differentially distributed in the tissues (Figure 7.2.1)². [142, 182] Moreover, they have been cloned from various mammalian species, where they differentiate for both their pharmacological profile and effector coupling. [139]

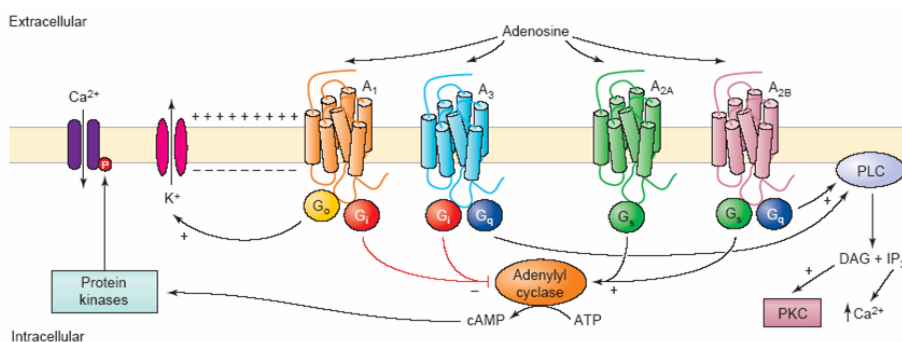


Figure 7.2.1: Signal transduction pathways involved in the activation of adenosine receptors. A_1 and A_3 activation inhibits adenylyl cyclase through G_i family of G proteins, whereas A_{2A} and A_{2B} receptors activate G_s family, that stimulates the adenylyl cyclase activity. Furthermore, A_1R subtype may lead to the activation of G_o family that increases the conductance of K^+ ions (efflux from inside to outside the cell) and influences some protein kinase activities. Finally, A_3 and A_{2B} receptors activation can involve G_q proteins, with the resulting stimulation of phospholipase C (PLC). Diacylglycerol (DAG) and inositol (1,4,5)-trisphosphate (IP_3) are implicated in the regulation of protein kinase C (PKC) activity and the intracellular concentration of Ca^{2+} ions, respectively. The adenosine receptors activation may involve other G proteins, affecting further cellular pathways.

Diverse potent and selective ligands for each subtype have demonstrated the potential therapeutic role of the adenosine receptor in several physiopathological processes. [140, 141, 143, 144, 178, 179, 185-187] In particular, A_1R selective antagonists have shown anxiolytic effects and they have been

²Adapted from Moro, S.; Spalluto, G.; Jacobson, K. A. *Trends Pharmacol. Sci.* 2005, 26, 44-51.

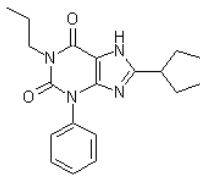
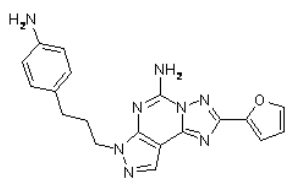
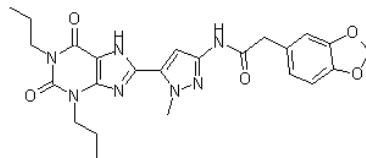
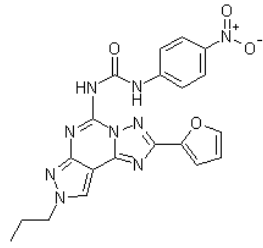
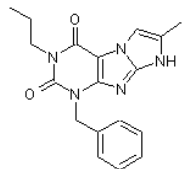
reported as promising candidates for the treatment of cognitive disorders, such as dementia. The antagonism selectivity for A₁R is also the proposed mechanism for some diuretic agents, which are considered effective in congestive heart failure and in oedema. [142, 185] A_{2A}R antagonists have a neuroprotective activity during ischemic processes and seem to play a role in the reduction of neuronal damage in Parkinson’s or Huntington’s diseases. [141-144] A potential therapeutic activity in the asthma disease has been discovered for A_{2B}R selective antagonists or mixed antagonists to A_{2B}Rs and A₃Rs. [142] A_{2B}R antagonists are also studied as hypoglycaemic agents in diabetes, while A₃R antagonists have a potential application in tumor growth inhibition and in the treatment of glaucoma. [142, 178, 179, 187]

Basing on different molecular scaffolds, diverse drug design approaches have been applied for the discovery of more potent and selective human AR (hAR) antagonists. To this aim, the xanthine and pyrazolo-triazolo-pyrimidine scaffolds have been properly modified to introduce novelty in the chemical space of known adenosine receptors antagonists. [146, 149, 188] In particular, improving selectivity for the hA₁R subtype has been obtained by decorating the classical xanthine scaffold with 8-aryl or 8-cycloalkyl substituents. [189] On the other hand, hA_{2B}R selective antagonists have been developed with different substitutions at the N¹, N³ and 8 positions in the same xanthine scaffold. [190-192] Regarding the pyrazolo-triazolo-pyrimidine derivatives, the position N⁷ has been suggested to be crucial for the selectivity to the hA_{2A}R subtype. [146] Conversely, proper substituents at N⁵ and N⁸ positions shift the antagonism towards the hA₃R subtype. [149] Furthermore, an alternative imidazo[2,1-*f*]purinone scaffold has been discovered to improve potency and selectivity for the human A₃R antagonists. [193] This structural information is summarized in Table 7.1.

7.3 Dataset

Unfortunately, only a limited number of known AR antagonists has been synthesized and tested on all four human AR subtypes. In the past, most of the literature partially reported the binding affinity to some hAR subtypes or the dataset was obtained by using ARs cloned from other mammalian species. We have collected 514 hAR antagonists, synthesized and tested on all four hAR subtypes, to derive and validate our three *auto*MEP/*ct*-SVM multilabel

Table 7.1: Examples of potent and selective AR antagonists to the four A_1 , A_{2A} , A_{2B} , and A_3 subtypes.

Structures	Exp. binding affinity	Ref(s).
<i>A₁ adenosine receptor antagonists</i>		
	hA₁R $K_i = 7.1$ nM hA _{2A} R $K_i = 1200$ nM hA _{2B} R $K_i = 625$ nM hA ₃ R $K_i = 395$ nM	[190]
<i>A_{2A} adenosine receptor antagonists</i>		
	hA ₁ R $K_i = 2160$ nM hA_{2A}R $K_i = 0.22$ nM hA _{2B} R $K_i > 10,000$ nM hA ₃ R $K_i > 10,000$ nM	[146]
<i>A_{2B} adenosine receptor antagonists</i>		
	hA ₁ R $K_i = 566$ nM hA _{2A} R $K_i > 1,000$ nM hA_{2B}R $K_i = 18$ nM hA ₃ R $K_i > 1,000$ nM	[146]
<i>A₃ adenosine receptor antagonists</i>		
	hA ₁ R $K_i = 1214$ nM hA _{2A} R $K_i = 1115$ nM hA _{2B} R $K_i = 305$ nM hA₃R $K_i = 0.81$ nM	[192]
	hA ₁ R $K_i > 1,000$ nM hA _{2A} R $K_i > 1,000$ nM hA _{2B} R $K_i > 1,000$ nM hA₃R $K_i = 0.8$ nM	[193]

classification models. [134, 146, 147, 149-151, 180, 181, 188-195] They are xanthine derivatives, N⁷ and N⁸ pyrazolo-triazolo-pyrimidine analogs. This large dataset was split into training set (318 compounds), validation set (65 compounds) and internal test set (131 compounds). In Figure 7.3.1 the potency and selectivity spectrum for each hAR subtype, considering only our collected homogeneous data on human ARs, is summarized.

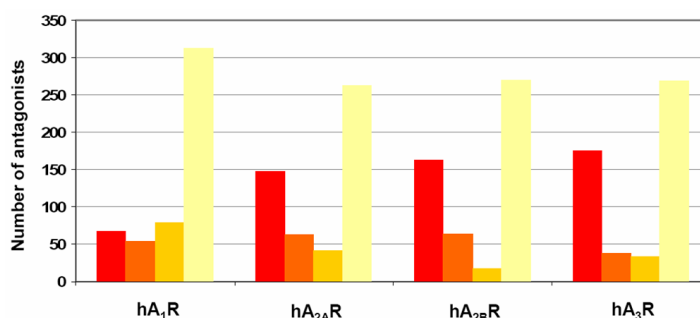


Figure 7.3.1: Distribution representation of the experimental K_i binding affinity data of hAR antagonists in our dataset (514 molecules), including training set, validation set and internal test set. The classes are completely overlapping.

The number of hAR antagonists are reported in different ranges of binding affinity for each subtype: $K_i \leq 100$ nM (●), 100 nM $< K_i \leq 250$ nM (●), 250 nM $< K_i \leq 500$ nM (●), $K_i > 500$ nM (●). A similar distribution of the corresponding potency profiles intervals is present for hA_{2A}R, hA_{2B}R and hA₃R subtypes, with a noteworthy number of potent antagonists having a binding affinity K_i value lower or equal to 100 nM to the distinct subtypes (Figure 7.3.1). Concerning the hA₁R subtype, our collection is short of potent compounds and many hA₁R antagonists have a binding affinity K_i value higher than 250 nM. In the final collection, 209 hAR antagonists (41%) are selective for one hAR subtype, with the corresponding K_i values lower than 100 nM. Considering these 209 compounds, 5 are selective for hA₁R (2%), 31 are selective for hA_{2A}R (15%), 79 are hA_{2B}R selective antagonists (38%) and the remaining 94 are hA₃R selective antagonists (45%). Considering the limited information on hA₁R subtype and on human AR antagonists, it is very difficult to correctly predict the complete hAR potency profiles and infer the selectivity of novel xanthine and pyrazolo-triazolo-pyrimidine derivatives.

Finally, 13 newly synthesized pyrazolo-triazolo-pyrimidine analogues have been selected as external test set.

7.4 Results and discussion

In the last years, the Support Vector Machine method has shown good generalization performance and high accuracy as supervised learning technique in many classification tasks. [82, 85] Unfortunately, the traditional classification models do not give any quantitative information about the biological affinity of the compounds to the corresponding receptor. Moreover, in chemoinformatics the classical single-label classification considers mutually exclusive classes, while in some cases compounds are simultaneously labeled in multiple classes. In the present work, hAR antagonists may present multiple molecular properties of multiple hAR subtypes. Since our dataset deals with non-mutually exclusive and overlapping classes, the alternative ct-SVM multilabel classification provides an appropriate approach. [108, 109]

Both topological and electrostatic complementarities are crucial in describing the receptor subtypes selectivity, and the investigation of the MEP on the molecular surface is a useful strategy for rationalizing the interactions involved in the molecular recognition processes. [52, 57, 58] In this case, the distribution of the MEP on the molecular surface is able to discriminate the selectivity for different receptor subtypes. The introduction of the autocorrelation vector allows then for overcoming the MEP information inconvenience to be reliant on the spatial rotation and translation of the molecule.

As anticipated, the pyrazolo-triazolo-pyrimidine scaffold can be properly modified to obtain hA_{2A}R or hA₃R selectivity (chapter 6). On the other hand, the xanthine scaffold has been investigated to develop highly potent and selective hA₁R or A_{2B}R antagonists. However, the observation of the scaffold and its substitutions is not an unfailing strategy to correctly assign the potency profile covering all hAR subtypes of new human pyrazolo-triazolo-pyrimidine and xanthine AR antagonists.

Our "sieve system" is composed of three in series ct-SVM multilabel classification models using as input matrix our *auto*MEP vectors (Figure 7.1.1). We aim at introducing *auto*MEP descriptors of each antagonist to best approximate the experimental hA₁R, hA_{2A}R, hA_{2B}R and hA₃R subtypes potency profile applying three independent *auto*MEP/ct-SVM classification models. In particular, our *auto*MEP/ct-SVM models have been derived after the selection of different thresholds of binding affinity, corresponding to three diverse K_i values: 500 nM for MODEL 1, 250 nM in MODEL 2 and 100 nM for MODEL 3 (Figure 7.1.1). Interestingly, our models can provide at the

same time a quantitative information about the binding affinity K_i values to all hAR subtypes. In fact, starting from the calculation of *autoMEP* vectors of novel pyrazolo-triazolo-pyrimidine and xanthine analogs, our MODELS 1, 2 and 3 are able to detect potent and selective hAR antagonists.

7.4.1 *autoMEP*/ct-SVM classification models

Each model is characterized by n binary classifiers, with n corresponding to the number of AR subtypes ($n=4$). The score values have been transformed in the predicted classes according to the C criterion. [108] To derive our *autoMEP*/ct-SVM models we selected a collection of 318 pyrazolo-triazolo-pyrimidine and xanthine derivatives (molecules **1-318**), and we have defined them as our training set. Furthermore, we have considered 65 additional pyrazolo-triazolo-pyrimidine and xanthine analogs (molecules **319-383**) as validation set for each model to find the optimal parameters of the four binary classifiers. As previously described, in our *autoMEP*/ct-SVM models the actual labels (experimental classes) are assigned by selecting a different binding affinity K_i value as threshold. In particular, a binary criterion assign "1" if the hAR subtype binding affinity K_i value is lower than the selected threshold (500 nM, 250 nM or 100 nM according to the MODELS 1, 2 or 3, respectively). Conversely, "0" is assigned if the hAR antagonists are less potent than the threshold, i.e. the corresponding hAR subtype binding affinity K_i value is higher than 500 nM, 250 nM or 100 nM in the corresponding MODELS 1, 2 and 3. The selected thresholds act as meshes of our "sieve system", able to filter hAR antagonists with increasing potency.

We have considered as selectivity criterion a difference of at least 2 orders of magnitude between the corresponding hAR subtypes K_i values, with the lower receptor subtype K_i value ≤ 100 nM. To carry out the final *autoMEP*/ct-SVM classification models the corresponding training set and validation set were merged in a collection of 383 compounds for the training of the new classifiers with the ct-SVM optimized parameters³. Finally, an internal test set (molecules **384-514**) has been selected to validate our MODELS 1, 2 and 3.

³MODEL 1 classifiers: hA₁R ($C = 4$, $\gamma = 2$); hA_{2A}R ($C = 4$, $\gamma = 0.5$); hA_{2B}R ($C = 16$, $\gamma = 0.5$); hA₃R ($C = 16$, $\gamma = 0.5$). MODEL 2 classifiers: hA₁R ($C = 4$, $\gamma = 2$); hA_{2A}R ($C = 8$, $\gamma = 0.5$); hA_{2B}R ($C = 4$, $\gamma = 0.5$); hA₃R ($C = 16$, $\gamma = 1$). MODEL 3 classifiers: hA₁R ($C = 4$, $\gamma = 0.5$); hA_{2A}R ($C = 4$, $\gamma = 0.5$); hA_{2B}R ($C = 4$, $\gamma = 0.5$); hA₃R ($C = 4$, $\gamma = 0.5$).

7.4.2 Internal validation

As anticipated, our *autoMEP/ct-SVM* multilabel classification models have been evaluated in their in series applicability as "sieve system" for new xanthine and pyrazolo-triazolo-pyrimidine derivatives. In this validation process each of the hA₁R, hA_{2A}R, hA_{2B}R and hA₃R antagonists has been analyzed by MODELS 1, 2 and 3. In Figure 7.4.1 the potency profiles of five structurally different and correctly predicted hAR antagonists have been considered as examples for the predictions interpretation. Differently colored columns are used for each antagonist to show the predicted subtypes/labels by MODELS 1, 2 and 3. The predicted value "1" by the corresponding hAR binary classifier is represented with a color for each model to indicate the relative percentage (%) of potency profile. In Figure 7.4.1 the colors have been assigned in this way: hA₁R (●), hA_{2A}R (●), hA_{2B}R (●) and hA₃R (●).

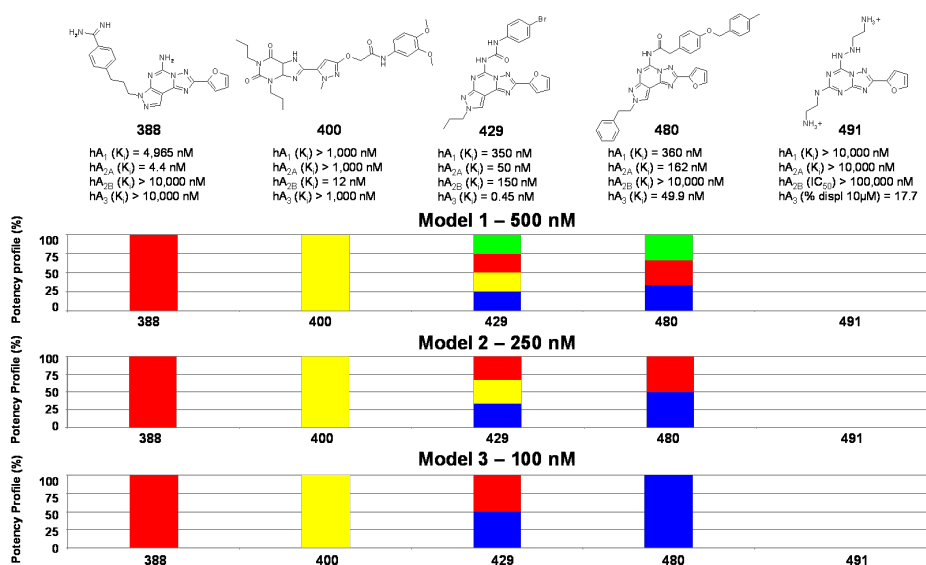


Figure 7.4.1: Flowchart of the correctly predicted profiles of five hAR antagonists by applying our *autoMEP/ct-SVM* multilabel classification models on the internal test set.

Thus, the prediction "1"/"0" refers to the corresponding hAR subtype K_i value, resulting to be lower/higher than the thresholds 500 nM, 250 nM or 100 nM for the MODELS 1, 2 or 3, respectively. A predicted relative percentage (%) of potency profile lower than 100% for at least one class means that the compound is more potent than the selected threshold K_i

for multiple hAR subtypes. Moreover, the selectivity can be inferred by one-color columns in all three models. Then, by observing Figure 7.4.1, the first N⁷-substituted pyrazolo-triazolo-pyrimidine analog (compound **388**) resulted to have 100% hA_{2A}R potency profile for all three models, with totally red columns. This corresponds to a K_i value lower than 100 nM only for the hA_{2A}R subtype and a K_i value higher than 500 nM for the remaining hAR subtypes. Consequently, the compound **388** is hA_{2A}R selective. The compound **400**, a xanthine derivative, is predicted to have a complete hA_{2B}R potency profile by the MODELS 1, 2 and 3. By observing the resulting columns, both compounds **429** and **480** show a mixed potency profile. They are N⁸-substituted pyrazolo-triazolo-pyrimidine analogs and each *auto*MEP/*ct*-SVM model is able to inform about the experimental K_i values for all hAR subtypes. Finally, we have predicted a negative potency profile for all hAR subtypes in the last example (compound **491**), characterized by a different chemical structure. In fact, no colored columns are present in the *auto*MEP/*ct*-SVM models prediction.

In our analysis the internal test set has been sifted out by three quantitative filters. We have compared the experimental with the corresponding predicted relative percentages (%) of potency profile by MODELS 1, 2 and 3 for the internal test set, as illustrated in Figure 7.4.2, 7.4.3 and 7.4.4, respectively. The colored columns indicate the positive potency profile for hA₁R (●), hA_{2A}R (●), hA_{2B}R (●) and hA₃R (●) profiles.

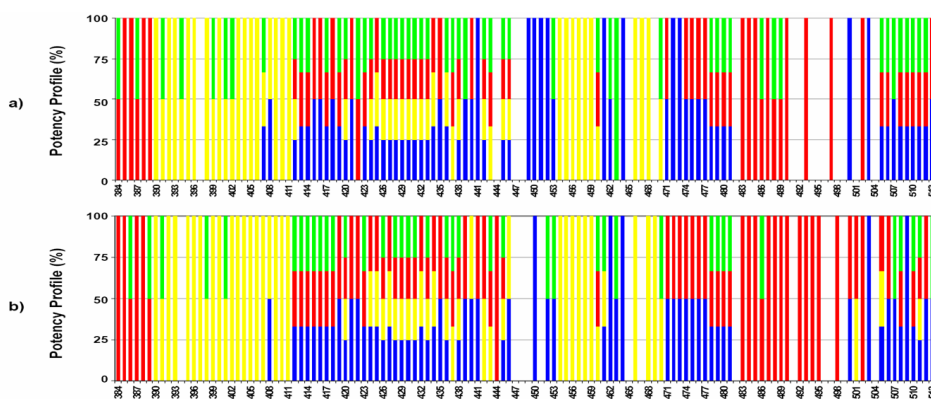


Figure 7.4.2: Graphical representation of the (a) experimental and (b) predicted by MODEL 1 (threshold = 500 nM) potency profiles for the internal test set. The colored columns show whether the K_i values to the relative hAR subtype are lower than 500 nM.

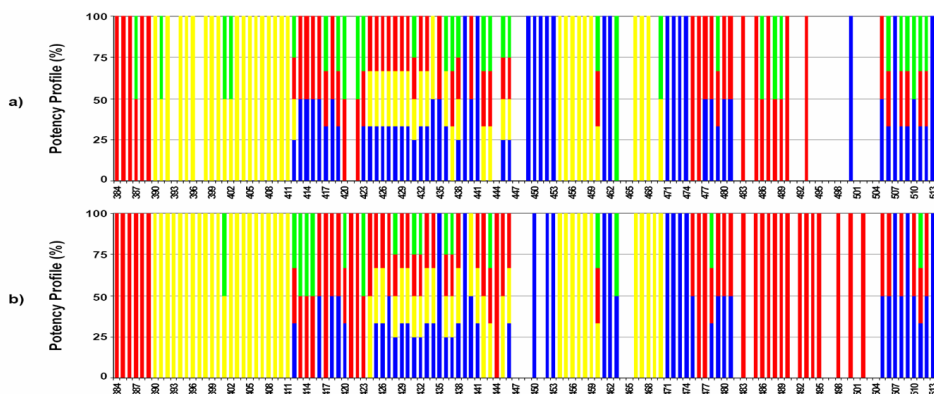


Figure 7.4.3: Graphical representation of the (a) experimental and (b) predicted by MODEL 2 (threshold = 250 nM) potency profiles for the internal test set. The colored columns show whether the K_i values to the relative hAR subtype are lower than 250 nM.

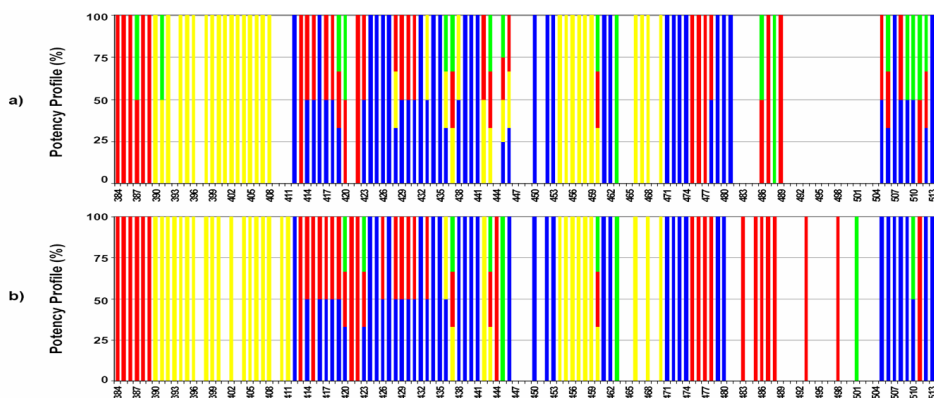


Figure 7.4.4: Graphical representation of the (a) experimental and (b) predicted by MODEL 3 (threshold = 100 nM) potency profiles for the internal test set. The colored columns show whether the K_i values to the relative hAR subtype are lower than 100 nM.

Good values of accuracy were obtained for our *autoMEP*/ct-SVM multi-label classification models in the prediction of the collected internal test set (86%, 81% and 78% for MODELS 1, 2 and 3, respectively). The satisfying prediction accuracies can be also inferred by the graphical comparison of the experimental (Figure 7.4.2a, 7.4.3a and 7.4.4a) with the predicted (Figure 7.4.2b, 7.4.3b and 7.4.4b) potency profiles for each model. The comparison of the experimental with the corresponding predicted graphical representations has shown high similarity in the color distribution.

Moreover, the satisfactory prediction results are highlighted by the good values of the statistical base-class parameters (Table 7.2).

Table 7.2: Statistical parameters of our models after prediction of the internal test set.

Classes	MODEL 1		MODEL 2		MODEL 3	
	Recall	Precision	Recall	Precision	Recall	Precision
hA ₁ R	0.61	0.79	0.34	0.65	0.39	0.70
hA _{2A} R	0.92	0.84	0.95	0.82	0.79	0.75
hA _{2B} R	0.96	0.86	0.96	0.89	0.78	0.91
hA ₃ R	0.94	0.97	0.80	0.92	0.92	0.98

The analysis of the experimental and the predicted hAR potency profiles of the internal test set by the abovementioned models, with the exception of the hA₁R subtype, again support the quality of our filtering strategy. After considering in series our *autoMEP/ct-SVM* models, 26/49 selective hAR antagonists have been perfectly predicted (molecules **385, 388, 390, 392, 395, 396, 400, 404-406, 408, 425, 431, 450, 454-459, 466, 468, 471, 474, 487** and **507**) and we are able to infer hAR subtype selectivity of 10 compounds by analyzing their partially correct predictions (molecules **384, 386, 389, 394, 398, 403, 407, 452, 461** and **513**). However, MODEL 3 has correctly assigned the potency profile of other 10 out of 49 selective hAR antagonists (molecules **412, 424, 427, 432, 434, 440, 441, 463, 472** and **473**), as reported in Figure 7.2.2, 7.2.3 and 7.2.4. Concerning the prediction of potent ($K_i \leq 100$ nM) hAR antagonists, MODEL 3 has detected 39% potent hA₁R antagonists, 79% potent hA_{2A}R antagonists, 78% potent hA_{2B}R antagonists and 96% potent hA₃R antagonists.

7.4.3 External validation

In the optimization step of drug discovery, the principal application of the *autoMEP/ct-SVM* models described in the present work is the prediction of the complete hAR binding affinity profile and hAR subtypes selectivity of new potential antagonists. To evaluate the prediction capability of our *autoMEP/ct-SVM* strategy, we have considered as potential new hAR antagonists 13 novel N⁸-substituted pyrazolo-triazolo-pyrimidine analogs (external test set, compounds **515-527**), which have been synthesized and tested on all four hAR subtypes.

The experimental human A₁R, A_{2A}R, A_{2B}R and A₃R binding affinities are collected in Table 7.3.⁴

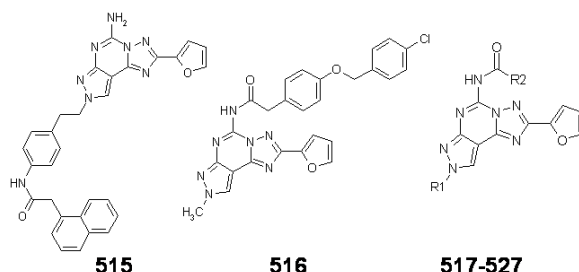


Table 7.3: Biological profile at the four hAR subtypes of the external test set (515-527).

Mol.	R ₁	R ₂	hA ₁ (K _i nM) ^a	hA _{2A} R (K _i nM) ^b	hA _{2B} R (K _i nM) ^c	hA ₃ R (K _i nM) ^d
515	-	-	1,250	87.3	3,700	1,537
516	-	-	1,710	2,520	>10,000	25.5
517	CH ₃	CH ₂ -O-Ph	42	15.1	>10,000	692
518	CH ₂ CH ₃	CHPh ₂	156	131	248	0.98
519	CH ₂ CH ₂ CH ₃	CH ₂ -2-thienyl	228	6.21	2,853	255
520	CH ₂ CH ₂ CH ₃	CH ₂ -3-thienyl	170	4.51	2,820	3.14
521	CH ₂ CH ₂ CH ₂ CH ₃	CH ₂ -β-naphthyl	113	7.94	1,200	174
522	CH ₂ CH ₂ CH(CH ₃) ₂	CH ₂ -4-OCH ₃ -Ph	22.6	1.95	460	359
523	CH ₂ CH ₂ CH(CH ₃) ₂	CH ₂ -2-thienyl	20.5	2.03	2,030	308
524	CH ₂ CH ₂ CH(CH ₃) ₂	CH ₂ -3-thienyl	31.5	3.88	2,487	540
525	CH ₂ CH ₂ CH(CH ₃) ₂	CH ₂ -O-Ph-4-Cl	155	16.1	>10,000	2,306
526	CH ₂ CH ₂ Ph	CH ₂ -O-Ph-4-Cl	199	31.7	>30,000	3,251
527	CH ₂ CH ₂ CH ₂ Ph	CH ₂ -4-OCH ₃ -Ph	44.8	8.93	2,690	120

^aDisplacement of specific [³H]-CCPA binding at human A₁ receptors expressed in CHO cells, (n=3-6);

^bdisplacement of specific [³H]-NECA binding at human A_{2A} receptors expressed in CHO cells; ^cK_i values of the inhibition of NECA-stimulated adenylyl cyclase activity in CHO cells expressing hA_{2B} receptors;

^ddisplacement of specific [³H]-NECA binding at human A₃ receptors expressed in CHO cells. Data are expressed as geometric means, with 95% confidence limits.

These derivatives are the result of various substitution and homologation experiments in two different positions to increase the selectivity of the pyrazolo-triazolo-pyrimidine scaffold to hA_{2A}R and hA₃R subtypes. *autoMEP* vectors of these new 13 hAR antagonists have been used as input matrix for the previously generated *autoMEP/ct-SVM* multilabel classification MODELS 1, 2 and 3.

⁴Experimental binding affinity data kindly provided by the work coordinated by Prof. G. Spalluto (University of Trieste) for the synthesis and by Prof. K. N. Klotz (University of Würzburg) for the pharmacological characterization.

In Figure 7.4.5 we have reported the predicted potency profiles by MODELS 1, 2 and 3, where the relative percentages (%) indicate positive potency profile for hA₁R (●), hA_{2A}R (●), hA_{2B}R (●) and hA₃R (●) subtypes. In most cases our models are able to assign the potency profile with at least the 75% of accuracy for each compound in the collected external test set (Figure 7.4.5).

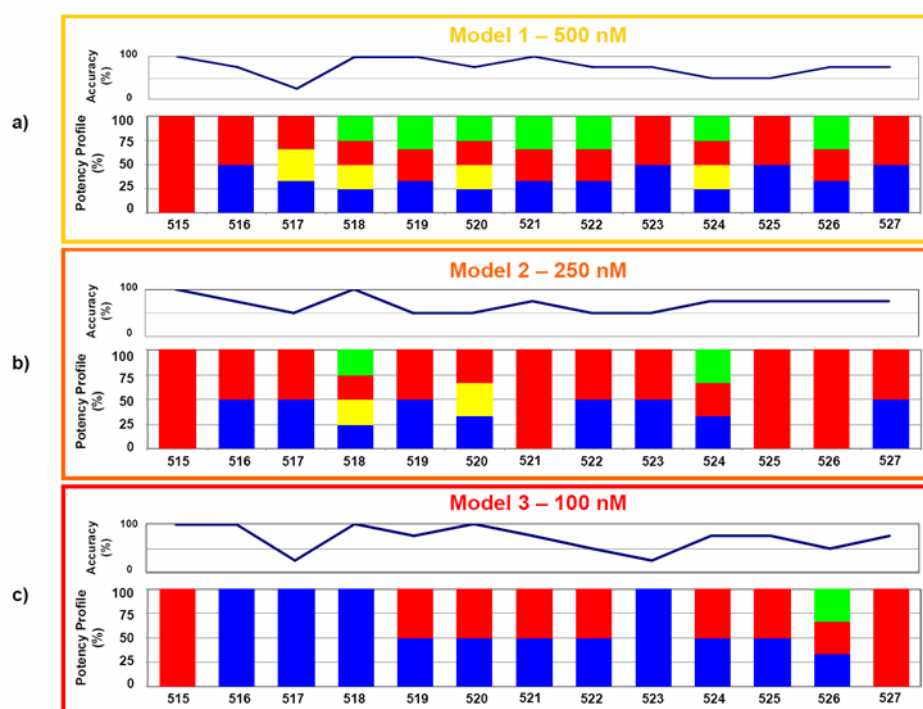


Figure 7.4.5: External test set predictions by a) MODEL 1 - 500 nM, b) MODEL 2 - 250 nM, c) MODEL 3 - 100 nM. The graphical representation indicates the percentage (%) of accuracy for each compound and the predicted relative percentages (%) of the potency profiles for hAR subtypes.

Our methodology has correctly classified most of the selective compounds in the external test set. In more detail, the compounds **515** and **518**, hA_{2A}R and hA₃R selective antagonists, respectively, are perfectly predicted. Considering their similar structure, our *autoMEP/ct-SVM* models are able at least to select almost all potent hA_{2A}R antagonists (molecules **519-527**), without missing out any potent hA₃R antagonist (molecule **520**) in the filtering procedure (see Figure 7.4.5c). Surprisingly, MODEL 3 has not misattributed any compound to hA_{2B}R class.

7.5 Final remarks

A novel multilabel classification approach combining *autoMEP* molecular descriptors with Support Vector Machine (*autoMEP/ct-SVM*) has been presented as powerful tool to predict hA₁R, hA_{2A}R, hA_{2B}R and hA₃R subtypes potency profile and infer the potential selectivity of known xanthine and pyrazolo-triazolo-pyrimidine derivatives. Three statistically meaningful models have been generated from the same training set by using different binding affinity K_i values as thresholds for hAR classifiers and very positive results were achieved in the validation procedure.

The independent application of each of our models can be used to select with high accuracy hAR antagonists having a binding affinity K_i value lower than 500 nM, 250 nM or 100 nM, according to the aim of the filtering process. To further improve the predictivity of our dynamic *autoMEP/ct-SVM* strategy, we aim at integrating new information on hAR antagonists in our dataset, especially regarding the hA₁R subtype.

Prediction of toxicodynamic and toxicokinetic properties

Quantitative structure-activity relationship (QSAR) analysis has been frequently utilized as computational tool for the prediction of several ecotoxicological parameters including the acute aquatic toxicity. In this chapter we describe a novel integrated strategy to predict the acute aquatic toxicity through the combination of both toxicokinetic and toxicodynamic behaviors of chemicals. In particular, a robust classification model (*TOXclass*) has been derived by combining Support Vector Machine (SVM) analysis with three classes of toxicokinetic-like molecular descriptors: the autocorrelation molecular electrostatic potential (*autoMEP*) vectors, Sterimol topological descriptors and $\log P(o/w)$ property values. *TOXclass* model is able to assign chemicals to different levels of acute aquatic toxicity, providing an appropriate answer to the new regulatory requirements. Moreover, we have extended the abovementioned toxicokinetic-like descriptor set with more toxicodynamic-like descriptors, as for example HOMO and LUMO energies, to generate a valuable SVM classifier (*MOAclass*) for the prediction of the mode of action (MOA) of toxic chemicals. As preliminary validation of our approach, the toxicokinetic (*TOXclass*) and the toxicodynamic (*MOAclass*) models have been applied in series to inspect both aquatic toxicity hazard and mode of action of 296 chemical substances with unknown or uncertain toxicodynamic information to assess their potential ecological risk and toxic mechanism.

8.1 Introduction

The need for various toxicological data of chemicals in limited time and animal experiments requires the application of alternative computational solutions. As anticipated in chapter 1, the attempt of REACH regulation is the improvement of the toxicity assessment process, by identifying the most hazardous properties of chemicals. [24, 25] Regardless animal and *in vivo* testing strategies are still reliable methods for the human and environmental toxicology risk assessment, the computational toxicology offers a valuable tool to speed up the costly screening of high numbers of compounds. As a consequence, several Intelligent or Integrated Testing Strategies (ITS) have been proposed as rapid, efficient approaches to obtain exposure and effects data and identify different modes of toxic action. [28, 29] In more detail, *in vitro* or computational methods, optimized *in vivo* studies, chemical categories, read-across analysis and thresholds of toxicological concern (TTCs) are admitted non-testing strategies to replace missing data or endpoints, and profitably reduce costly animal experiments. [27] So far, powerful computational toxicology prediction systems have been developed for the exposure and hazard assessment to satisfy the new regulatory requests. [30] In drug discovery the *in silico* approaches for the toxicity prediction of safety-relevant endpoints are precious contributions to early discovery of adverse drug reactions. [31, 35]

In the last years several "data driven systems" and "expert systems" have become available for the prediction of toxicological endpoints. Data driven based programs generate statistically valuable structure-activity relationships (SARs) by processing large groups of unrelated chemicals, without user bias or prior organization, to find associations based on similar chemical structures, known as structural alerts, that most probably correspond to the same toxicological mechanism. Examples for data driven softwares able to predict toxicity endpoints are TOPKAT, MCASE and Lazar. [196-198] Unfortunately, the ease prediction in these techniques is penalized by the accurate statistical validation needed. For this reason, they are better suggested to detect general alerting properties. On the other hand, the expert systems embody a series of knowledge based rules, considering small classes of similar-acting chemicals or groups of compounds with similar structure to build classes of potential toxicity. Even if their application is more limited in comparison with the data driven systems, the expert systems offer

more easily interpretable results. [31] In particular, various expert systems can provide the prediction of aquatic toxicity endpoints, such as ECOSAR, DEREK, HazardExpert and OASIS. [199-206]

In toxicology quantitative structure-activity relationships are widely used approaches to infer the toxicological properties of compounds from their molecular structure. The traditional linear QSAR methodologies are still the most applied strategies in this field. Several studies have focused on the prediction of the aquatic toxicity of chemical substances as basic information in the hazard and environmental risk assessment for species living in the water. [37-41] In this context, two main QSAR strategies, chemical class-based and toxicological-based, have been carried out. In the class-based approach the aquatic toxicity is modeled on small series of homologous chemical substances, according to the concept that similar compounds should behave with a similar toxic mode. [207-209] A problem in this classification scheme is represented by the difficult treatment of complex molecular structures. Alternatively, the toxicological-based QSAR models are developed for compounds supposed to act with the same toxic mechanism. [210-212] However, the toxicological-based approach is closely related to the toxicodynamic information and considers toxicity dependent on the mode of action (MOA). Moreover, in both methods the class assignment is not unambiguous, if more functional groups are involved in the same compound/mechanism of action.

In a previous work, Colombo has highlighted the advantages of the application of simple constitutional and quantum chemical descriptors for the classification of chemicals into structure-related subsets to derive QSAR local models, independently from the mode of action. [213] As proposed so far, most of the QSAR local models require the chemical grouping into structural or toxicodynamic classes first.

The typical endpoint for the assessment of acute toxicity is the concentration lethal to 50% of the organisms (LC_{50}), produced by chemicals with different mechanisms, and well-defined thresholds for acute toxicity have been established by UNECE in the Globally Harmonized System of Classification and Labeling of Chemicals (GHS). [214] The toxic effect may involve different types of biochemical molecular interaction between the chemicals and the biological target. Concerning aquatic toxicology, the classification scheme published by Verhaar and included in Toxtree is one of the first strategies to assign chemicals to mechanisms of action. [215, 216]

Russom and collaborators have utilized the information on the toxicodynamic profiles in the *fathead minnow* specie to develop an expert system based on substructural fragments for the classification of chemical substances in different modes of action. [217] In the recent years, alternative reliable classification models based on various molecular descriptors have been introduced. [218-222] The classification approach has recently been proposed in the prediction of genotoxicity, a property deeply involved in the toxicological profile of compounds. In such application, a preprocessing of experimental toxicity data by selecting a threshold was required for the classes definition. [223] Very recently, the well-known EPA Fathead Minnow Acute Toxicity (EPAFHM) database, reporting 96-h LC_{50} values for fathead minnow of diverse industrial chemicals, has been selected to derive a local support vector regression model. [224]

In the present work we describe a novel classification approach able to assign a large number of compounds in the publicly available EPAFHM dataset to different classes of acute aquatic toxicity and modes of toxic action. The workflow of our approach is illustrated in Figure 8.1.1.

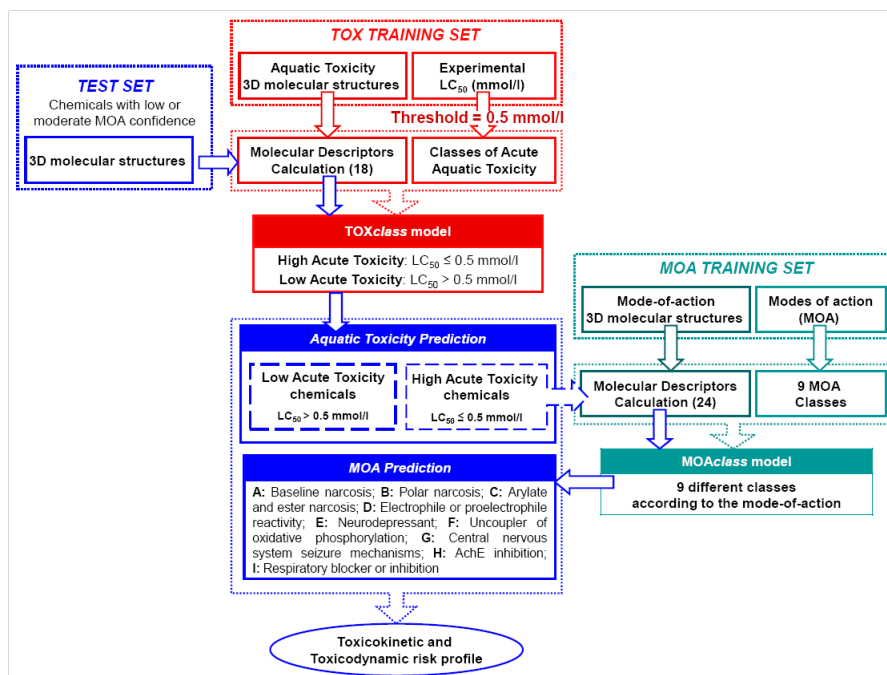


Figure 8.1.1: Flowchart of the in series TOXclass and MOAclass approach for the prediction of the toxicokinetic and toxicodynamic risk profiles of new chemicals.

Our first attempt is to provide an easily interpretable answer to the regulatory requirements by defining two classes of environmental hazard based on a 96-h LC₅₀ threshold value. We have carried out a couple of novel classification models combining Support Vector Machine with two different sets of molecular descriptors (*TOXclass* and *MOAclass* models). In particular, a statistically appreciable classification model (*TOXclass*) has been derived by combining SVM analysis with three classes of molecular descriptors: the autocorrelation molecular electrostatic potential (*autoMEP*) vectors, Sterimol topological descriptors and logP(o/w) property values. Once developed a model to predict the level of acute aquatic toxicity, we have extended the previous descriptor set by introducing further properties, that influence the toxicodynamic profile. Based on this descriptor set, we have generated a robust SVM classifier (*MOAclass*) for the prediction of the multiple MOA of toxic chemicals. The toxicokinetic and the toxicodynamic models have been applied in series to identify both aquatic toxicity classes and modes of action of 296 chemicals with unknown MOA or moderate confidence MOA assignment to assess both potential ecological risk and mechanism of toxicity.

8.2 Dataset

A collection of 559 industrial chemicals in the original EPAFHM database has been selected to derive and validate our *TOXclass* and *MOAclass* models. [217] EPAFHM dataset provides for each compound the chemical structure, 96-h LC₅₀ values in mg/L and mmol/L and specifies the mode of action with the corresponding confidence. The toxicity profiles distribution of the dataset considered in the present work is summarized in Table 8.1 and reported in Figure 8.2.1.

Table 8.1: *The classification of substances in mg/L according to the GHS legislation and the corresponding LC₅₀ (mmol/L) intervals for our dataset are reported; tox = toxicity.*

Minor classes	Values LC ₅₀ (mg/L)	Values LC ₅₀ (mmol/L)
Acute Tox 1 (AT1)	AT1 ≤ 1 mg/L	AT1 ≤ 0.00848 mmol/L
Acute Tox 2 (AT2)	1 mg/L < AT2 ≤ 10 mg/L	0.001 ≤ AT2 ≤ 0.131 mmol/L
Acute Tox 3 (AT3)	10 mg/L < AT3 ≤ 100 mg/L	0.028 ≤ AT3 ≤ 1.285 mmol/L
No Acute Tox (nAT)	nAT > 100 mg/L	nAT ≥ 0.360 mmol/L

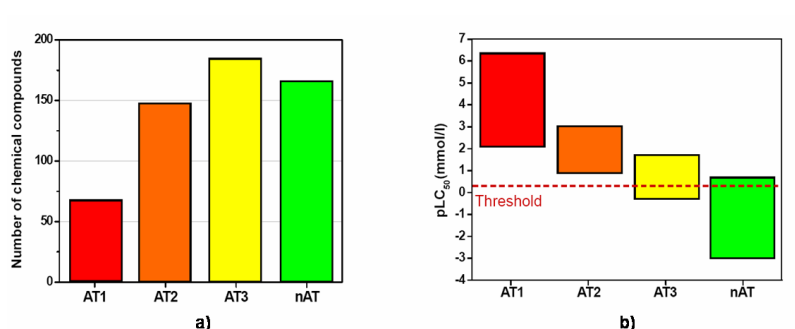


Figure 8.2.1: Graphical representation of 559 chemicals in our dataset. a) Data distribution indicating the acute toxicity classes according to the GHS classification scheme for substances hazardous to the aquatic environment: $LC_{50} \leq 1$ mg/L (red), 1 mg/L $< LC_{50} \leq 10$ mg/L (orange), 10 mg/L $< LC_{50} \leq 100$ mg/L (yellow), $LC_{50} > 100$ mg/L (green); (b) data distribution indicating the intervals of pLC_{50} (mmol/L) values corresponding to the classes reported in (a); the threshold for acute toxicity ($pLC_{50} = 0.3$) is highlighted.

In the present study, a training set of 554 compounds has been used to carry out our *TOXclass* model (*TOXclass* training set). Moreover, 263 chemicals have been selected as training set of our *MOAclass* model (*MOAclass* training set). They show a high or high-moderate level of confidence in the assigned nine modes of action and this collection is partially overlapping to the abovementioned *TOXclass* training set. Finally, a preliminary test set comprising 296 compounds has been considered in both *TOXclass* and *MOAclass* models. These chemicals display a moderate or low level of confidence in the classification by MOA and they are partially included in the *TOXclass* training set. In particular, 187 chemicals in the test set (subset 1) have a low or moderate MOA confidence, while for the remaining 109 compounds (subset 2) the mode of action is unknown. The mechanistic information should be experimentally determined to verify the goodness of our predictions.

Acute toxicity data are expressed as LC_{50} values (in mmol/L) to quantify the concentration lethal to 50% juvenile fathead minnows (*Pimephales promelas*) in 96-flow-through exposure tests. [217] In our analysis LC_{50} values in mmol/L are transformed in classes of aquatic toxicity, as described in section 8.3. The mode of action of chemicals was assigned based on joint toxic action studies, involving the analysis of behavioral responses and dose-response relationships after interpreting 96 h LC_{50} experiments. [217]

8.3 Results and discussion

Computer-based identification of molecular structure properties qualitatively (SAR) or quantitatively (QSAR) related to biological activity represents the main useful application in predictive toxicology. [30, 31, 35] In fact, the unknown toxicological properties of new chemicals can be inferred by considering the available information on toxic molecular structures or fragments.

Even if SAR and QSAR methodologies are more and more rising importance in this field, their applicability requires good quality and homogeneous experimental data. Firstly, LC₅₀ precise value of acute aquatic toxicity is not needed in the case of general evaluation of the environmental risk of a chemical. Just one unit of difference between experimental and predicted pLC₅₀ corresponds to ten times the same difference expressed in LC₅₀ (mmol/L). So the claim for the prediction of exactly the LC₅₀ value by using a classical QSAR local model is very challenging. Moreover, if the mechanism of action is unknown or uncertain, and LC₅₀ values are inaccurate, the possibility to derive reliable QSAR local models for each MOA subset is excluded. So, these methods are better tailored for a posteriori identification of alerting classes from the predicted values. Therefore, a classification approach independent from MOA, by selecting a particular threshold of aquatic toxicity, seems to be more appropriate.

To date, the classical models for acute aquatic toxicity prediction by chemical class or mode of action have approached toxicity as a property depending on the presence of particular scaffolds/functional groups and on the available toxicodynamic information. Some chemicals with different structure might be very toxic by acting with the same mechanism (*Paper VI*). Then, we can hypothesize that local models based on chemical classes are not able to correctly assess the level of toxicity of new chemically different compounds. Moreover, the different molecular intrinsic properties might be somehow responsible for a similar toxicokinetic. In EPAFHM dataset the toxic compounds displaying a different mode of action, but analogous molecular structures, are present (*Paper VI*). Even in this case, a local model by MOA might not recognize as toxic new compounds with similar structure but different toxicodynamic profile.

Our in series TOX*class*/MOA*class* strategy has been developed to overcome these problems in the toxicity prediction. Interestingly, we have considered the general toxicity as a property well described by molecular de-

scriptors, rather than a priori influenced by the chemical structure or toxic mechanism. As synthetically represented in Figure 8.3.1, we have derived two robust *TOXclass* and *MOAclass* models to predict the level of acute aquatic toxicity and the mode of toxic action, respectively. Our procedure has analyzed toxicity as a property independent from the mechanism of action, since differently acting chemicals can be equally toxic. Moreover, MOA information is not needed to predict the potential toxicological risk. Consequently, we suggest to apply *TOXclass* and *MOAclass* models in series for the prediction of the toxicokinetic and toxicodynamic profiles of new chemicals.

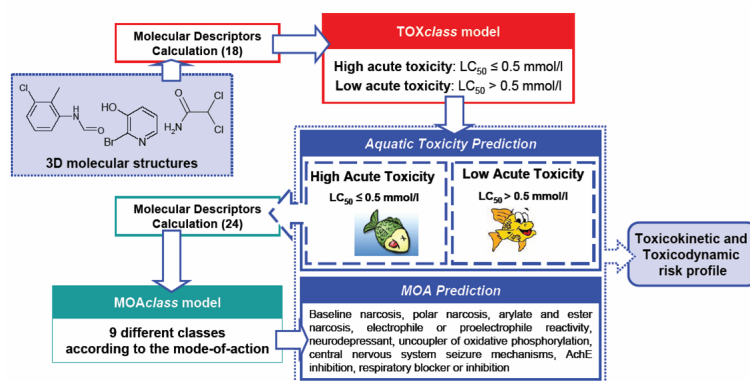


Figure 8.3.1: Synthesis of the *in series* *TOXclass* and *MOAclass* approach.

8.3.1 *TOXclass* model

We have used 554 structurally different chemicals in the same EPAFHM dataset as our training set to carry out a first binary classification model (*TOXclass*). By considering data distribution of LC_{50} values in mmol/L according to GHS legislation, reported in Table 8.1 and Figure 8.2.1, we have selected 0.5 mmol/L as LC_{50} threshold value for high acute toxicity ($pLC_{50}=0.3$). Then, we have divided our training set into two classes: high acute aquatic toxicity ($LC_{50} \leq 0.5$ mmol/L), that comprises 66% of the training set, and low acute aquatic toxicity ($LC_{50} > 0.5$ mmol/L), including the remaining 34%. As anticipated, we have combined three classes of molecular descriptors with SVM analysis: *autoMEP* vectors (calculated by using default parameters indicated in chapter 2), Sterimol topological descriptors and $\log P(o/w)$ property values, as reported in Table 8.2.

Table 8.2: List of the eighteen descriptors selected for the binary classification model to predict the level of acute toxicity.

No.	Name	Details
12	<i>auto</i> MEP vectors	Spatial autocorrelation; property: molecular electrostatic potential
5	Sterimol	Topological
1	logP(o/w)	Log octanol/water partition coefficient

In this study, the autocorrelation allows in particular to easily compare surface properties of structurally unrelated compounds of different size. Sterimol descriptors have been introduced to represent topological properties, as in previous papers. [60, 225] Finally, logP, or log of the *n*-octanol/water partition coefficient, describes the partitioning equilibria and is parameter commonly used as molecular descriptor in the evaluation of pharmacokinetic and pharmacodynamic properties and in modeling of toxicity. [212, 219-222] We have selected these molecular properties to represent properly the chemical influence to the toxicological profile of each compound.

A statistically appreciable *TOXclass* model has been carried out by using a Gaussian radial basis function kernel setting the *C* parameter value to 32 and the γ parameter value to 1. The results are shown in Table 8.3 and Table 8.4.

Table 8.3: Statistical parameters of *TOXclass* model after cross-validation procedure.

Partition	CV	% correct predictions				
		No. of runs	Mean	StDev	Min	Max
Training set		1	88.6	-	88.6	88.6
Training set	LOO	1	84.3	-	84.3	84.3
	10-fold	10	82.4	4.5	70.9	92.9
	5-fold	20	82.0	3.4	70.3	89.2
	3-fold	33	81.3	2.2	76.2	85.9
	2-fold	50	80.5	1.8	76.2	85.2

Table 8.4: Statistical parameters of *TOXclass* model after LOO cross-validation.

Classes	TP (TP rate)	FP (FP rate)	Recall	Precision
<i>high</i> AT	323 (0.88)	43 (0.23)	0.88	0.88
<i>low</i> AT	144 (0.77)	44 (0.12)	0.77	0.77

The percentages (%) of correct predictions obtained after the extensive n -fold cross-validation procedures are higher than 80% (80.5% is the minimum average predictivity for the 2-fold cross-validation), with 84.3% correctly classified chemicals after LOO cross-validation procedure. The robustness of TOXclass model is confirmed by the values of recall and precision for both classes. These results demonstrate the good capability of TOXclass model to infer the toxicokinetic profile of chemicals prior any consideration on the toxicodynamic mechanism. Therefore, we have decided to apply this model as toxicity classifier, as described in 8.3.3 paragraph.

8.3.2 MOAclass model

Regarding the MOA training set, 263 chemicals with high or high-moderate MOA confidence in the EPAFHM database have been assigned to nine different MOA and their distribution is given in Figure 8.3.2.

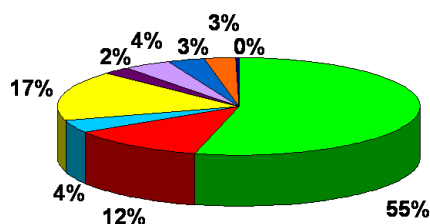


Figure 8.3.2: MOAclass model. Data distribution indicating the percentage (%) of the single MOA classes: baseline narcosis (green), polar narcosis (red), arylate and ester narcosis (light blue), electrophile or proelectrophile phosphorylation (yellow), neurodepressant (purple), uncoupler of oxidative phosphorylation (violet), central nervous system seizure mechanisms (blue), AchE inhibition (orange) and respiratory blocker or inhibition mechanisms (black). The percentages correspond to 142 baseline narcosis, 32 polar narcosis, 11 arylate and ester narcosis, 44 electrophile or proelectrophile phosphorylation, 6 neurodepressant, 11 uncoupler of oxidative phosphorylation, 9 central nervous system seizure mechanisms, 7 AchE inhibition and 1 respiratory blocker or inhibition acting chemicals, respectively.

This training set is characterized by very unbalanced classes, with few chemicals acting with neurodepressant, uncoupler of oxidative phosphorylation, central nervous system seizure, AchE inhibition and respiratory blocker or inhibition mechanisms. However, this distribution reflects the actual well-established knowledge on the toxicodynamic profile of chemicals.

The descriptor set used in the *MOAclass* model is an extension of the family of descriptors reported above for the *TOXclass* model, as listed in Table 8.5.

Table 8.5: List of the twenty four descriptors selected for the *MOAclass* model.

No.	Name	Details
12	<i>auto</i> MEP vectors	Spatial autocorrelation; property: molecular electrostatic potential
5	Sterimol	Topological
1	logP(o/w)	Log octanol/water partition coefficient
1	HDon	Number of hydrogen bonding donors derived from the sum of NH and OH groups in the molecule
1	HAcc	Number of hydrogen bonding acceptors derived from the sum of nitrogen and oxygen atoms in the molecule
1	TPSA	Topological polar surface area
1	ASA	Approximate surface area
1	HOMO	The energy (eV) of the Highest Occupied Molecular Orbital
1	LUMO-HOMO	Difference between the energy (eV) of the Lowest Unoccupied Molecular Orbital and the energy (eV) of the Highest Occupied Molecular Orbital

Besides *auto*MEP vectors, Sterimol and logP(o/w) descriptors, further quantum-physical-chemical properties were introduced to more accurately describe chemistry related the toxicodynamic profile (the number of hydrogen bonding donors and acceptors, the topological polar and approximate surface areas, HOMO and the difference LUMO-HOMO energy descriptors).

The SVM methodology has been utilized in combination with the above-mentioned descriptors to derive a robust *MOAclass* classifier, by using Gaussian radial basis function kernel ($C = 32$, $\gamma = 0.9$). *MOAclass* model statistical parameters are summarized in Table 8.6 and Table 8.7.

An acceptable *MOAclass* model has been obtained as indicated by the percentage (%) of correct predictions after LOO cross-validation procedure and the recall and precision values higher than 70% for the A, B, E, F and H MOA classes. Interestingly, 77% correct predictions, yielded after LOO cross-validation, confirm the reliability of this model. Regardless the high n -fold standard deviations, *MOAclass* analysis gave noteworthy results considering that MOA classes are highly unbalanced. Moreover, the inclusion of new molecular descriptors considerably improved the model predictivity for

Table 8.6: *Statistical parameters of the MOAclass model after cross-validation.*

Partition	CV	% correct predictions			
		No. of runs	Mean	StDev	Min Max
Training set		1	97.3	-	97.3 97.3
Training set	LOO	1	77.2	-	77.2 77.2
	10-fold	10	75.1	8.0	57.7 92.6
	5-fold	20	73.3	5.5	58.5 84.9
	3-fold	33	69.7	4.3	55.7 78.2
	2-fold	50	65.8	4.0	57.2 74.0

Table 8.7: *MOAclass model. The statistical parameters after LOO cross-validation procedure. In the first column baseline narcosis (green), polar narcosis (red), arylate and ester narcosis (light blue), electrophile or proelectrophile phosphorylation (yellow), neurodepressant (purple), uncoupler of oxidative phosphorylation (violet), central nervous system seizure mechanisms (blue), AchE inhibition (orange) and respiratory blocker or inhibition (black) mode-of-action classes are highlighted.*

Classes	TP (TP rate)	FP (FP rate)	Recall	Precision
● (A)	126 (0.89)	30 (0.25)	0.89	0.81
● (B)	23 (0.72)	10 (0.04)	0.72	0.70
● (C)	5 (0.45)	4 (0.02)	0.45	0.56
● (D)	25 (0.57)	11 (0.05)	0.57	0.70
● (E)	5 (0.83)	0 (0.00)	0.83	1.00
● (F)	9 (0.82)	2 (0.01)	0.82	0.82
● (G)	5 (0.56)	3 (0.01)	0.56	0.62
● (H)	5 (0.71)	0 (0.00)	0.71	1.00
● (I)	0 (0.00)	0 (0.00)	0.00	0.00

B, C, D and G classes (polar narcosis, arylate and ester narcosis, electrophile or proelectrophile phosphorylation and central nervous system seizure mechanisms, respectively). This MOAclass predictor has been selected to evaluate the test set, as described in the following paragraph.

8.3.3 Applicability of TOXclass and MOAclass models

As preliminary proof, TOXclass and MOAclass models have been applied in series to predict both toxicokinetic and toxicodynamic profiles of 296 chemicals (test set) with experimentally uncertain mechanisms of toxicity. In particular, our attempt is the evaluation of the potential ecological risk and,

in a second step, the mode of toxic action.

Our classification approach is able to intrinsically predict the toxicological classes instead of the numerical values of LC_{50} , with a low probability of significant errors with respect to classical QSAR regression methods. In more detail, our *TOXclass* model can directly assign chemicals to "high acute aquatic toxicity" or "low acute aquatic toxicity" classes, corresponding to LC_{50} values in mmol/L lower or higher than 0.5, respectively. Following the workflow reported in Figure 8.3.1, the eighteen descriptors reported in Table 8.2 of all chemicals in the test set have been used as input matrix for *TOXclass* model and the prediction results are reported in Table 8.8.

Table 8.8: *Statistical parameters after the test set prediction by TOXclass model.*

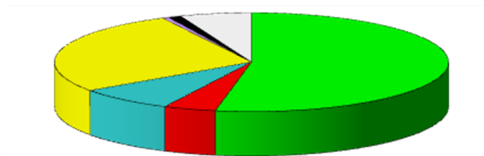
Classes	TP (TP rate)	FP (FP rate)	Recall	Precision
<i>highAT</i>	183 (0.93)	18 (0.18)	0.93	0.91
<i>lowAT</i>	81 (0.82)	14 (0.07)	0.82	0.85

TOXclass model yielded a percentage (%) of 89.2 correct predictions for the test set, with recall and precision higher than 80% for both aquatic toxicity classes. Only 14 high acute toxic aquatic chemicals have been erroneously recognized as low acute toxic.

After predicting the toxicokinetic profile, the ideal procedure should be the exclusive prediction of MOA for compounds with hazard toxicity ($LC_{50} \leq 0.5$ mmol/L). In this work, we have applied *MOAclass* model to the whole test set and the results have been analyzed separately for the subsets 1 and 2. In this case, the descriptor set reported in Table 8.5 was used to represent the compounds in the test set for the prediction by *MOAclass* model. The experimental (low and moderate confidence) and predicted MOA classes for subset 1 are summarized in Table 8.9 and Table 8.10, respectively.

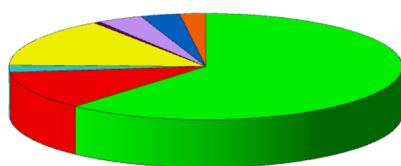
Only 49% predicted classes corresponds to the experimental toxicodynamic profiles. The comparison of the experimental with the predicted MOA classes for the subset 1 supports the debatable quality of the experimental data. However, a good correspondence in the distribution of chemicals within the toxicological classes has been found, as underlined by the comparison of the graphical representations in Table 8.9 and Table 8.10.

Table 8.9: Graphical representation of the experimental toxicodynamic classes for chemicals with low or moderate MOA confidence in the test set (subset 1). The percentages (%) and the corresponding colors indicating the distribution of chemicals within the classes are reported in the table below.



Experimental MOA classes	Percentage (%)
● (A) - Baseline narcosis	53
● (B) - Polar narcosis	4
● (C) - Arylate and ester narcosis	8
● (D) - Electrophile or proelectrophile phosphorylation	27
● (E) - Neurodepressant	0
● (F) - Uncoupler of oxidative phosphorylation	1
● (G) - Central nervous system seizure mechanisms	0
● (H) - AchE inhibition	0
● (I) - Respiratory blocker or inhibition	1
○ (A and B)	6

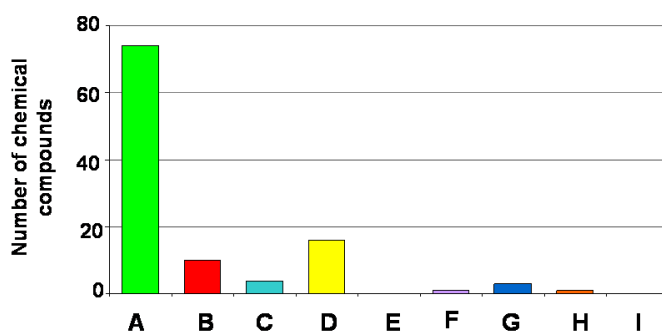
Table 8.10: Graphical representation of the predicted toxicodynamic classes for chemicals with low or moderate MOA confidence in the test set (subset 1). The percentages (%) and the corresponding colors indicating the distribution of chemicals within the classes are reported in the table below.



Predicted MOA classes	Percentage (%)
● (A) - Baseline narcosis	61
● (B) - Polar narcosis	12
● (C) - Arylate and ester narcosis	2
● (D) - Electrophile or proelectrophile phosphorylation	15
● (E) - Neurodepressant	1
● (F) - Uncoupler of oxidative phosphorylation	4
● (G) - Central nervous system seizure mechanisms	3
● (H) - AchE inhibition	2
● (I) - Respiratory blocker or inhibition	0

In Table 8.11 the compounds in subset 2, classified by MOAclass model, have been reported.

Table 8.11: Graphical representation of the predicted toxicodynamic classes by MOAclass model for the chemicals with unknown MOA in the test set (subset 2). The number of compounds for each class are indicated, and the corresponding percentages (%) with the colors showing the distribution of chemicals within the classes are reported in the table below.



Predicted MOA classes	Percentage (%)
● (A) - Baseline narcosis	67
● (B) - Polar narcosis	9
● (C) - Arylate and ester narcosis	4
● (D) - Electrophile or proelectrophile phosphorylation	15
● (E) - Neurodepressant	0
● (F) - Uncoupler of oxidative phosphorylation	1
● (G) - Central nervous system seizure mechanisms	3
● (H) - AchE inhibition	1
● (I) - Respiratory blocker or inhibition	0

Most chemicals have been predicted as baseline narcosis (74 compounds) and electrophile or proelectrophile phosphorylation (16 compounds) acting (Table 8.11). The following step should be to experimentally verify our subset 2 predictions and extend the applicability of our approach to new chemicals.

8.4 Final remarks

The toxicokinetic and toxicodynamic aspects of aquatic toxicity have been investigated to provide useful tools in agreement with the recent regulatory system for the evaluation of the environmental hazards. We have applied

novel powerful classification methods to discriminate chemicals into classes of toxicity and mode of toxic action. The novelty of the strategy is represented by a first approach to toxicity not related to toxicodynamic and chemical information. Moreover, our models have provided an easily interpretable answer to the regulatory requirements by defining two classes of acute aquatic toxicity. In particular, we have presented a couple of robust SVM classification models in combination with two families of molecular descriptors. The first *TOXclass* classifier has been applied to a test set to predict the level of aquatic toxicity, while the second *MOAclass* model has been used to infer the toxic mechanism.

The experimental evaluation of the acute aquatic toxicity and the mode of toxic action would represent an effective validation of our *TOXclass*/*MOAclass* approach. Finally, we aim at incorporating new compounds in the training sets to extend the *TOXclass*/*MOAclass* strategy to other chemicals for the aquatic toxicity prediction.

General conclusions

The present thesis has focused on the development of nonlinear QSAR models, mainly by using machine learning methods, as an attractive and helpful strategy in drug discovery. Our first intent is to demonstrate the wide applicability of nonlinear methodologies in pharmaceutical research tasks. The promising prediction results underlined by the six reported case studies in the field of pharmacodynamic, pharmacokinetics and toxicology have been obtained by combining several families of molecular descriptors with nonlinear techniques, to properly describe the relationship between the molecular properties and the desired endpoint.

Response Surface Analysis in combination with *auto*MEP vectors, encoding for 3D molecular structures, resulted as a robust methodology in the evaluation of both aqueous solvation free energy of organic compounds and binding affinity K_i values of human A_{2A}R antagonists, as described in chapters 3 and 4. Moreover, the case study discussed in chapter 4 has emphasized the potential parallel use of nonlinear methods to support the predictivity of linear strategies.

Especially Support Vector Machine has been suggested for its good predictive power and the generalization capability in a large number of regression and classification studies. The classification-based approaches aim at predicting classes of activity (high and low) or mechanisms of activity, while the regression-based methods offer the precise prediction of activity or property data. In chapter 5 we have reported interesting results about the application of nonlinear classification approaches to predict the cytochrome P450 metabolism undergone by xenobiotics in humans. In the pharmacodynamic

field, novel powerful SVM methodologies have been proposed to predict potency and selectivity of human AR antagonists, first focusing on A_{2A}R and A₃R subtypes (chapter 6) to finally extend our task to all hAR subtypes (chapter 7). In more detail, we have generated single-label and multilabel classification models, respectively, trying to develop a filtering strategy to detect potent and selective hAR antagonists. Further algorithms and screening approaches are being developing to improve the results in the selection process of drug candidates.

Regarding the prediction of the toxicological profile of chemicals, *in silico* approaches are more and more applied as alternative methods to animal testing in order to refine and reduce the experiments. Consequently, in computational toxicology investigations the QSAR models should be in agreement with the recent regulatory system for the evaluation of the environmental hazards. In chapter 8 we have discussed a novel classification strategy to evaluate the toxicokinetic and toxicodynamic aspects of aquatic toxicity. Interestingly, we introduced a new approach for studying toxicity, which is not strictly dependent on the toxicodynamic profile of chemicals. Then, the efforts should be directed to the interpretation of other toxicological endpoints, in order to built a complete system supporting costly animal testing protocols.

Bibliography

- [1] Kaitin, K. I. Obstacles and opportunities in new drug development. *Nature Clin. Pharm. Ther.* **2008**, *83*, 210-212.
- [2] Kola, I.; Landis, J. Can the pharmaceutical industry reduce the attrition rates? *Nature Rev. Drug Discov.* **2004**, *3*, 711-715.
- [3] Chadwick, A.; Hajek, M. Learning to improve the decision-making process in research. *Drug Discov. Today* **2004**, *9*, 251-257.
- [4] Li, H.; Yap, C. W.; Xue, Y.; Li, Z. R.; Ung, C. Y.; Han, L. Y.; Chen, Y. Z. Statistical learning approach for predicting specific pharmacodynamic, pharmacokinetic, or toxicological properties of pharmaceutical agents. *Drug Dev. Res.* **2006**, *66*, 245-259.
- [5] Tang, W.; Lu, A. Y. H. Drug metabolism and pharmacokinetics in support of drug design. *Curr. Pharm. Des.* **2009**, *15*, 2170-2183.
- [6] Korfmacher, W. A. Advances in the integration of drug metabolism into the lead optimization paradigm. *Mini-Rev. Med. Chem.* **2009**, *9*, 703-716.
- [7] Czodrowski, P.; Kriegl, J. M.; Scheuerer, S.; Fox, T. Computational approaches to predict drug metabolism. *Exp. Opin. Drug Met.* **2009**, *5*, 15-27.
- [8] Madden, J. C.; Cronin, M. T. D. Structure-based methods for the prediction of drug metabolism. *Exp. Opin. Drug Met.* **2006**, *2*, 545-557.
- [9] Jolivet, L. J.; Ekins, S. Methods for predicting human drug metabolism. *Adv. Clin. Chem.* **2007**, *43*, 131-176.

BIBLIOGRAPHY

- [10] Duch, W.; Swaminathan, K.; Meller, J. Artificial intelligence approaches for rational drug design and discovery. *Curr. Pharm. Des.* **2007**, *13*, 1497-1508.
- [11] Yap, C. W.; Li, H.; Ji, Z. L.; Chen, Y. Z. Regression methods for developing QSAR and QSPR models to predict compounds of specific pharmacodynamic, pharmacokinetic and toxicological properties. *Mini-Rev. Med. Chem.* **2007**, *7*, 1097-1107.
- [12] Norinder, U.; Bergström, C. A. S. Prediction of ADMET properties. *Chem. Med. Chem.* **2006**, *1*, 920-937.
- [13] Mager, D. E. Quantitative structure-pharmacokinetic/pharmacodynamic relationships. *Adv. Drug Deliv. Rev.* **2006**, *58*, 1326-1356.
- [14] Khan, M. T.; Sylte, I. Predictive QSAR modeling for the successful predictions of the ADMET properties of candidate drug molecules. *Curr. Drug Discov. Technol.* **2007**, *4*, 141-149.
- [15] Yap, C. W.; Xue, Y.; Li, H.; Li, Z. R.; Ung, C. Y.; Han, L. Y.; Zheng, C. J.; Cao, Z. W.; Chen, Y. Z. Prediction of compounds with specific pharmacodynamic, pharmacokinetic or toxicological property by statistical learning methods. *Mini-Rev. Med. Chem.* **2006**, *6*, 449-459.
- [16] Chohan, K. K.; Paine, S. W.; Waters, N. J. Quantitative structure activity relationships in drug metabolism. *Curr. Top. Med. Chem.* **2006**, *6*, 1569-1578.
- [17] Ruiz-Gracia, A.; Bermejo, M.; Moss, A.; Casabo, V. G. Pharmacokinetics in drug discovery. *J. Pharm. Sci.* **2008**, *97*, 654-690.
- [18] Wang, J.; Hou, T. Recent advances on in silico ADME Modeling. *Ann. Rep. Comput. Chem.* **2009**, *5*, 101-127.
- [19] Mehdipour, A. R.; Hamidi, M. Brain drug targeting: a computational approach for overcoming blood-brain barrier. *Drug Discov. Today* **2009**, *14*, 1030-1036.
- [20] Hou, T.; Wang, J.; Zhang, W.; Xu, X. ADME evaluation in drug discovery. 7. Prediction of oral absorption by correlation and classification. *J. Chem. Inf. Model.* **2007**, *47*, 208-218.
- [21] Hou, T. ADME evaluation in drug discovery. 8. The prediction of human intestinal absorption by a support vector machine. *J. Chem. Inf. Model.* **2007**, *47*, 2408-2415.
- [22] Sui, X.; Sun, J.; Wu, X.; Li, H.; Liu, J.; He, Z. Predicting the volume of distribution of drugs in humans. *Curr. Drug Metab.* **2008**, *9*, 574-580.

-
- [23] Paixão, P.; Gouveia, L. F.; Morais, J. A. G. Prediction of drug distribution within blood. *Eur. J. Pharm. Sci.* **2009**, *36*, 544-554.
- [24] European Union. Corrigendum to Regulation (EC) No 1907/2006 of the European Parliament and of the Council of 18 December 2006 concerning the Registration, Evaluation, Authorization and Restriction of Chemicals (REACH), establishing a European Chemicals Agency, amending Directive 1999/45/EC and repealing Council Regulation (EEC) No 793/93 and Commission Regulation (EC) No 1488/94 as well as Council Directive 76/769/EEC and Commission Directives 91/155/EEC, 93/67/EEC, 93/105/EC and 2000/21/EC (OJ L 396, 30.12.2006). *Off. J. Eur. Union*, L 136, 50, **2007**.
- [25] <http://ecb.jrc.it/reach/reach-legislation/>: REACH in brief.
- [26] Collins, F. S.; Gray, G. M.; Bucher, J. R. Transforming environmental health protection. *Science Tox.* **2009**, *319*, 906-907.
- [27] Schaafsma, G.; Kroese, E. D.; Tielemans, E. L.; Van de Sandt, J. J.; Van Leeuwen, C. J. REACH, non-testing approaches and the urgent need for a change in mind set. *Regul. Toxicol. Pharm.* **2009**, *53*, 70-80.
- [28] Bradbury, S. P.; Feijtel, T. C.; Van Leeuwen, C. J. Meeting the scientific needs of ecological risk assessment in a regulatory context. *Environ. Sci. Technol.* **2004**, *38*, 463-470.
- [29] Nendza, M.; Wenzel, A. Discriminating toxicant classes by mode of action. 1.(Eco)toxicity profiles. *Environ. Sci. Pollut. Res.* **2006**, *13*, 192-203.
- [30] Benfenati, E. Predicting toxicity through computers: a changing world. *Chem. Cent. J.* **2007**, *32*, 1-7.
- [31] Muster, W.; Breidenbach, A.; Fischer, H.; Kirchner, S.; Müller, L.; Pähler, A. Computational toxicology in drug development. *Drug Disc. Today* **2008**, *13*, 303-310.
- [32] Ma, X. H.; Wang, R.; Xue, Y.; Li, Z. R.; Yang, S. Y.; Wei, Y. Q.; Chen, Y. Z. Advances in machine learning prediction of toxicological properties and adverse drug reactions of pharmaceutical agents. *Curr. Drug Saf.* **2008**, *3*, 100-114.
- [33] Nigsch, F.; Macaluso, N. J.; Mitchell, J. B.; Zmuidinavicius, D. Computational toxicology: an overview of the sources of data and of modelling methods. *Exp. Opin. Drug Metab. Toxicol.* **2009**, *5*, 1-14.

- [34] Valerio, L. G. Jr. In silico toxicology for the pharmaceutical sciences. *Toxicol. Appl. Pharm.* **2009**, *241*, 356-370.
- [35] Helma, C. In silico predictive toxicology: the state-of-the-art strategies to predict human health effects. *Curr. Opin. Drug Disc.* **2005**, *8*, 27-31.
- [36] Fent, K.; Weston, A. A.; Caminada, D. Ecotoxicology of human pharmaceuticals. *Aquatic Tox.* **2006**, *76*, 122-159.
- [37] Papa, E.; Villa, F.; Gramatica, P. Statistically validated QSARs, based on theoretical descriptors, for modeling aquatic toxicity of organic chemicals in *Pimephales promelas* (Fathead Minnow). *J. Chem. Inf. Model.* **2005**, *45*, 1256-1266.
- [38] Mazzatorta, P.; Smiesko, M.; Lo Piparo, E.; Benfenati, E. QSAR models for predicting pesticide aquatic toxicity. *J. Chem. Inf. Model.* **2005**, *45*, 1767-1774.
- [39] Netzeva, T. I.; Pavan, M.; Worth, A. P. Review of (quantitative) structure-activity relationships for acute aquatic toxicity. *QSAR Comb. Sci.* **2008**, *27*, 77-90.
- [40] Castillo-Garit, J. A.; Marrero-Ponce, Y.; Escobar, J.; Torrens, F.; Rotondo, R. A novel approach to predict aquatic toxicity from molecular structure. *Chemosphere*, **2008**, *73*, 415-427.
- [41] Zhu, H.; Tropsha, A.; Fourches, D.; Varnek, A.; Papa, E.; Gramatica, P.; Oberg, T.; Dao, P.; Cherkasov, A.; Tetko, I. V. Combinatorial QSAR modeling of chemical toxicants tested against *Tetrahymena pyriformis*. *J. Chem. Inf. Model.* **2008**, *48*, 766-784.
- [42] Liu, P.; Long, W. Current mathematical methods used in QSAR/QSPR studies. *Int. J. Mol. Sci.* **2009**, *10*, 1978-1998.
- [43] Fox, T.; Kriegl, J. M. Machine learning techniques for in silico modeling of drug metabolism. *Curr. Top. Med. Chem.* **2006**, *6*, 1579-1591.
- [44] Arimoto, R. Computational models for predicting interactions with cytochrome P450 enzyme. *Curr. Top. Med. Chem.* **2006**, *6*, 1609-1618.
- [45] Yap, C. W.; Xue, Y.; Chen, Y. Z. Application of support vector machines to in silico prediction of cytochrome P450 enzyme substrates and inhibitors. *Curr. Top. Med. Chem.* **2006**, *6*, 1593-1607.
- [46] Gasteiger, J.; Engel, T. *Chemoinformatics*; Wiley-VHC, 2003.
- [47] Esbensen, K. H. *Multivariate data analysis - in practice*; CAMO, 2001.

- [48] Eriksson, L.; Johansson, E.; Kettaneh-Wold, N.; Wold, S. *Multi- and megavariate data analysis. Principles and applications*; Umetrics Academy, 2001.
- [49] Leach, A. R.; Gillet, V. J. *An introduction to chemoinformatics*; Kluwer Academic Publishers, 2003.
- [50] Burbidge, R.; Trotter, M.; Buxton, B.; Holden, S. Drug design by machine learning: support vector machines for pharmaceutical data analysis. *Comput. Chem.* **2001**, *26*, 5-14.
- [51] *ADRIANA.Code*; version 2.2; Molecular Networks GmbH: Erlangen, Germany, 2008.
- [52] Gasteiger, J.; Li, X.; Rudolph, C.; Sadowski, J.; Zupan, J. Representation of molecular electrostatic potentials by topological feature maps. *J. Am. Chem. Soc.* **1994**, *116*, 4608-4620.
- [53] Gasteiger, J.; Marsili, M. Iterative partial equalization of orbital electronegativity - a rapid access to atomic charges. *Tetrahedron* **1980**, *36*, 3219-3228.
- [54] Gasteiger, J.; Saller, H. Calculation of the charge distribution in conjugated systems by a quantification of the resonance concept. *Angew. Chem. Int. Ed. Engl.* **1985**, *24*, 687-689.
- [55] Moreau, G.; Broto, P. The autocorrelation of a topological structure: a new molecular descriptor. *Nouv. J. Chim.* **1980**, *4*, 359-360.
- [56] Moreau, G.; Broto, P. Autocorrelation of molecular structures, application to SAR studies. *Nouv. J. Chim.* **1980**, *4*, 757-764.
- [57] Wagener, M.; Sadowski, J.; Gasteiger, J. Autocorrelation of molecular surface properties for modeling corticosteroid binding globulin and cytosolic Ah receptor activity by neural networks. *J. Am. Chem. Soc.* **1995**, *117*, 7769-7778.
- [58] Bauknecht, H.; Zell, A.; Bayer, H.; Levi, P.; Wagener, M.; Sadowski, J.; Gasteiger, J. Locating biologically active compounds in medium-sized heterogeneous datasets by topological autocorrelation vectors: dopamine and benzodiazepine agonists. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 1205-1213.
- [59] Verloop, A. *The sterimol approach to drug design*; Marcel Dekker: New York, 1987.
- [60] Bacilieri, M.; Varano, F.; Deflorian, F.; Marini, M.; Catarzi, D.; Colotta, V.; Filacchioni, G.; Galli, A.; Costagli, C.; Kaseda, C.; Moro, S. Tandem 3D-QSARs approach as valuable tool to predict binding affinity data: design of

- new Gly/NMDA receptor antagonists as a key study. *J. Chem. Inf. Model.* **2007**, *47* (5), 1913-1922.
- [61] Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug. Deliver. Rev.* **1997**, *23*, 3-25.
- [62] Ertl, P.; Rohde, B.; Selzer, P. Fast calculation of molecular polar surface area as a sum of fragment-based contributions and its application to the prediction of drug transport properties. *J. Med. Chem.* **2000**, *43*, 3714-3717.
- [63] Labute, P. A widely applicable set of descriptors. *J. Mol. Graphics Modell.* **2000**, *18*, 464-477.
- [64] Gasteiger, J.; Hutchings, M. G. New empirical models of substituent polarisability and their application to stabilisation effects in positively charged species. *Tetrahedron Lett.* **1983**, *24*, 2537-2540.
- [65] Gasteiger, J.; Hutchings, M. G. Quantitative models of gas-phase proton-transfer reaction involving alcohols, ethers, and their thio analogues. Correlation analysis based on residual electronegativity and effective polarisability. *J. Am. Chem. Soc.* **1984**, *106*, 6489-6495.
- [66] Kang, K. K.; Jhon, M. S. Additivity of atomic polarisabilities and dispersion coefficients. *Theor. Chim. Acta* **1982**, *61*, 41-48.
- [67] Miller, K. J. Additivity methods in molecular polarisability. *J. Am. Chem. Soc.* **1990**, *112*, 8533-8542.
- [68] Todeschini, R.; Consonni, V. *Handbook of molecular descriptors*; Wiley-VCH: Weinheim, 2000; Vol. 11, pp 1-667.
- [69] Hall, L. H.; Kier, L. B. *The molecular connectivity chi indexes and kappa shape indexes in structure-property modeling*; Wiley-VHC: New York, 1991; Vol. 2, pp 367-422.
- [70] Wiener, H. Structural determination of paraffin boiling points. *J. Am. Chem. Soc.* **1947**, *69*, 17-20.
- [71] Petitjean, M. Applications of the radius-diameter diagram to the classification of topological and geometrical shapes of chemical compounds. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 331-337.
- [72] Bath, P. A.; Poirette, A. R.; Willet, P.; Allen, F. H. The extent of the relationship between the graph-theoretical and the geometrical shape coefficients of chemical compounds. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 714-716.

- [73] Tanford, C. *Physical chemistry of macromolecules*; Wiley: New York, 1961.
- [74] Volkenstein, M. V. *Configurational statistics of polymeric chains*; Wiley: New York, 1963; pp 1-562.
- [75] Stewart, J. J. P. *MOPAC manual*; seventh edition, 1993. <http://www.cdm.tu-dresden.de/edv/mopac7/mirror/mopac.html>.
- [76] Terfloth, L.; Bienfait, B.; Gasteiger, J. Ligand-based models for the isoform specificity of cytochrome P450 3A4, 2D6 and 2C9 substrates. *J. Chem. Inf. Model.* **2007**, *47*, 1688-1701.
- [77] Wold, H. *Research papers in statistics*; Wiley: New York, 1966.
- [78] Myers, R.; Montgomery, D. C. *Response methodology surface*; Wiley Interscience: New York, 1995.
- [79] Kaseda, C. *Response surface methodology using a spline algorithm*; Nashboro Press: Fujisawa-shi Kanagawa, Japan, 2004.
- [80] Vapnik, V. *The nature of statistical learning theory*; Springer: New York, 1995.
- [81] Vapnik, V. *Statistical learning theory*; Wiley: New York, 1998.
- [82] Cristianini, N.; Shawe-Taylor, J. *An introduction to support vector machines*; Cambridge University Press, 2000.
- [83] Boser, B. E., Guyon, I. M.; Vapnik, V. *Proceedings of the 5th annual ACM workshop on computational learning theory*; ACM: Pittsburgh, 1992, pp 144-152.
- [84] Hastie, T.; Tibshirani, R.; Friedman, J. *The elements of statistical learning; data mining, inference, and prediction*; Springer: New York, 2001.
- [85] Burges, C. J. C. *Data mining and knowledge discovery*; Springer: New York, 1998, Vol. 2, pp 121-167.
- [86] Lodhi, H.; Saunders, C., Shawe-Taylor, J.; Cristianini, N.; Watkins, C. *Journal of machine learning research*; MIT Press, 2002, Vol. 2, pp 419-444.
- [87] Ralaivola, L.; Swamidass, S. J.; Saigo, H.; Baldi, P. *Neural networks*; 2005, Vol. 18, pp 1093-1110.
- [88] Smola, A. J., Scholkopf, B. *Learning with kernels. Support vector machines, regularization, optimization, and beyond*; MIT Press, 2002.

BIBLIOGRAPHY

- [89] Czermiński, R.; Yasri, A.; Hartsough, D. Use of support vector machine in pattern classification: application to QSAR studies. *Quant. Struct. Act. Relat.* **2001**, *20*, 227-240.
- [90] Warmuth, M. K.; Liao, J.; Rätsch, G.; Mathieson, M.; Putta, S.; Lemmen, C. Active learning with support vector machine in the drug discovery process. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 667-673.
- [91] Jorissen, R. N.; Gilson, M. K. Virtual screening of molecular databases using a support vector machine. *J. Chem. Inf. Model.* **2005**, *45*, 549-561.
- [92] Bruce, C. L.; Melville, J. L.; Pickett, S. D.; Hirst, J. D. Contemporary QSAR classifiers compared. *J. Chem. Inf. Model.* **2007**, *47*, 219-227.
- [93] Li, J.; Liu, H.; Yao, X.; Liu, M.; Hu, Z.; Fan, B. Structure-activity relationship study of oxindole-based inhibitors of cyclin-dependent kinases based on least-squares support vector machines. *Anal. Chim. Acta* **2007**, *581*, 333-342.
- [94] Coi, A.; Bianucci, A. M.; Calderone, V.; Testai, L.; Digiacomio, M.; Rapposelli, S.; Balsamo, A. Predictive models, based on classification algorithms, for compounds potentially active as mitochondrial ATP-sensitive potassium channel openers. *Bioorg. Med. Chem.* **2009**, *17*, 5565-5571.
- [95] Kortagere, S.; Chekmarev, D.; Welsh, W. J.; Ekins, S. Hybrid scoring and classification approaches to predict human pregnane X receptor activators. *Pharm. Res.* **2009**, *26*, 1001-1011.
- [96] Luan, F.; Liu, H. T.; Ma, W. P.; Fan, B. T. Classification of estrogen receptor- β ligands on the basis of their binding affinities using support vector machine and linear discriminant analysis. *Eur. J. Med. Chem.* **2008**, *43*, 43-52.
- [97] Zhang, H.; Xiang, M.-L.; Zhao, Y.-L.; Wei, Y.-Q.; Yang, S.-Y. Support vector machine and pharmacophore-based prediction models of multidrug-resistance protein 2 (MRP2) inhibitors. *Eur. J. Pharm. Sci.* **2009**, *36*, 451-457.
- [98] Liew, C. Y.; Ma, X. H.; Liu, X.; Yap, C. W. SVM model for virtual screening of Lck inhibitors. *J. Chem. Inf. Model.* **2009**, *49*, 877-885.
- [99] Liu, X. H.; Ma, X. H.; Tan, C. Y.; Jiang, Y. Y.; Go, M. L.; Low, B. C.; Chen, Y. Z. Virtual screening of Abl inhibitors from large compound libraries by support vector machines. *J. Chem. Inf. Model.* **2009**, *49*, 2101-2110.
- [100] Liu, H. X.; Zhang, R. S.; Yao, X. J.; Liu, M. C.; Hu, Z. D. QSAR study of ethyl 2-[(3-methyl-2,5-dioxo(3-pyrrolinyl)amino]-4-(trifluoromethyl)pyrimi-

- dine-5-carboxylate: an inhibitor of AP-1 and NF- κ B mediated gene expression based on support vector machines. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1288-1296.
- [101] Norinder, U. Support vector machine models in drug design: applications to drug transport processes and QSAR using simplex optimisations and variable selection. *Neurocomputing* **2003**, *55*, 337-346.
- [102] Xue, C. X.; Zhang, R. S.; Liu, H. X.; Yao, X. J.; Liu, M. C.; Hu, Z. D.; Fan, B. T. QSAR models for the prediction of binding affinities to human serum albumin using the heuristic method and a support vector machine. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1693-1700.
- [103] Oloff, S.; Mailman, R. B.; Tropsha, A. Application of validated QSAR models of D₁ dopaminergic antagonists for database mining. *J. Med. Chem.* **2005**, *48*, 7322-7332.
- [104] Xia, B.; Ma, W.; Zheng, B.; Zhang, X.; Fan, B. Quantitative structure-activity relationship studies of a series of non-benzodiazepine structural ligands binding to benzodiazepine receptor. *Eur. J. Med. Chem.* **2008**, *43*, 1489-1498.
- [105] Chen, H.-F. Computational study of histamine H₃-receptor antagonist with support vector machines and three dimension quantitative structure activity relationship methods. *Anal. Chim. Acta* **2008**, *624*, 203-209.
- [106] Yuan, Y.; Zhang, R.; Hu, R.; Ruan, X. Prediction of CCR5 receptor binding affinity of substituted 1-(3,3-diphenylpropyl)-piperidiny amides and ureas based on the heuristic method, support vector machine and projection pursuit regression. *Eur. J. Med. Chem.* **2009**, *44*, 25-34.
- [107] Smola, A. J.; Scholkopf, B. *A tutorial on support vector regression*; NeuroCOL Technical report series, NCR2-TR-1998-030, 1998.
- [108] Boutell, M. R.; Luo, J.; Shen, X.; Brown, C. M. C. Learning multi-label scene classification. *Pattern Recogn.* **2004**, *37*, 1757-1771.
- [109] Hristozov, D.; Gasteiger, J.; Da Costa, F. B. Multilabeled classification approach to find a plant source for terpenoids. *J. Chem. Inf. Model.* **2008**, *48*, 56-67.
- [110] Winkler, D. A. Neural networks as robust tools in drug lead discovery and development. *Mol. Biotech.* **2004**, *27*, 139-167.
- [111] Zupan, J.; Gasteiger, J. *Neural networks in chemistry and drug design*, second edition; Wiley-VHC, Weinheim, 1999.

BIBLIOGRAPHY

- [112] Organization for Economic Cooperation and Development. *Principles for the validation, for regulatory purposes, of (Quantitative) Structure-Activity Relationship Models*; 2004, available online at <http://www.oecd.org/dataoecd/33/37/37849783.pdf>.
- [113] Gramatica, P. Principles of QSAR models validation: internal and external. *QSAR Comb. Sci.* **2007**, *26* (5), 694-701.
- [114] *OpenMolix*, version 2.4.26; Moshe Bar, Tel Aviv University: Israel, 2004.
- [115] *Adriana*, version 2.0; Molecular Networks GmbH: Erlangen, Germany, 2003.
- [116] *Molecular Operating Environment*, version 2008.10; Chemical Computing Group: Montreal, Canada, 2008.
- [117] *The Unscrambler*, version 9.6; CAMO Process AS: Oslo, Norway, 2006.
- [118] *DataFOREST*, version 9; Yamatake Corporation: Fujisawa-shi Kanagawa, Japan, 2007.
- [119] *DataNESIA*, version 3.2; Yamatake Corporation: Fujisawa-shi Kanagawa, Japan, 2007.
- [120] Joachims, T. *SVM^{light}*; version 6.01; Support Vector Machine. <http://svmlight.joachims.org>, 2004.
- [121] *Weka: Waikato Environment for Knowledge Analysis*; University of Waikato, New Zealand. <http://www.cs.waikato.ac.nz/ml/weka/> (accessed May 27, 2008).
- [122] R Development Core Team. *R: a Language and Environment for Statistical Computing*, version 2.2.1; 2005. URL: <http://www.r-project.org> (accessed June 2006).
- [123] Dimitriadou, E.; Hornik, K.; Leisch, F.; Meyer, D.; Weingessel, A. *e1071: Misc functions of the Department of Statistics (e1071)*, TU Wien, 2005.
- [124] *SONNIA*, Molecular Networks GmbH, Erlangen, Germany. <http://www.molecular-networks.com> (accessed April, 2008).
- [125] Johnson, S. R.; Zheng, W. Recent progress in the computational prediction of aqueous solubility and absorption. *AAPS J.* **2006**, *8*, 27-40.
- [126] Wang, H.; Ben-Naim, A. A possible involvement of solvent-induced interactions in drug design. *J. Med. Chem.* **1996**, *39*, 1531-1539.

- [127] Duffy, E. M.; Jorgensen, W. L. Prediction of properties from simulations: free energies of solvation in hexadecane, octanol, and water. *J. Am. Chem. Soc.* **2000**, *122*, 2878-2888.
- [128] Viswanadhan, V. N.; Ghose, A. K.; Singh, U. C.; Wendoloski, J. J. Prediction of solvation free energies of small organic molecules: additive-constitutive models based on molecular fingerprints and atomic constants. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 405-412.
- [129] Katritzky, A. R.; Oliferenko, A. A.; Oliferenko, P. V.; Petrukhin, R.; Tatham, D. B.; Maran, U.; Lomaka, A.; Acree, W. E. Jr. A general treatment of solubility. 1. The QSPR correlation of solvation free energies of single solutes in series of solvents. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1794-1805.
- [130] Mansson, R. A.; Frey, J. G.; Essex, J. W.; Welsh, A. H. Prediction of properties from simulations: a re-examination with modern statistical methods. *J. Chem. Inf. Model.* **2005**, *45*, 1791-1803.
- [131] Kang, H.; Choi, H.; Park, H. Prediction of molecular solvation free energy based on the optimization of atomic solvation parameters with genetic algorithm. *J. Chem. Inf. Model.* **2007**, *47*, 509-514.
- [132] Moro, S.; Bacilieri, M.; Ferrari, C.; Spalluto, G. Autocorrelation of molecular electrostatic potential surface properties combined with partial least squares analysis as alternative attractive tool to generate ligand-based 3D-QSARs. *Curr. Drug Discov. Technol.* **2005**, *2*, 13-21.
- [133] Moro, S.; Bacilieri, M.; Cacciari, B.; Spalluto, G. Autocorrelation of molecular electrostatic potential surface properties combined with partial least squares analysis as new strategy for the prediction of the activity of human A₃ adenosine receptor antagonists. *J. Med. Chem.* **2005**, *48*, 5698-5704.
- [134] Moro, S.; Bacilieri, M.; Cacciari, B.; Bolcato, C.; Cusan, C.; Pastorin, G.; Klotz, K. N.; Spalluto, G. The application of a 3D-QSAR (*autoMEP/PLS*) approach as an efficient pharmacodynamic-driven filtering method for small-sized virtual library: application to a lead optimization of a human A₃ adenosine receptor antagonist. *Bioorg. Med. Chem.* **2006**, *14*, 4923-4932.
- [135] Bacilieri, M.; Kaseda, C.; Spalluto, G.; Moro, S. Response surface analysis as alternative 3D-QSAR tool: human A₃ adenosine receptor antagonists as a key study. *Lett. Drug. Des. Discov.* **2007**, *4*, 122-127.
- [136] Lorentz H. A., *Theory of Electrons*; Dover, NY, 1952.

BIBLIOGRAPHY

- [137] Moro, S.; Bacilieri, M.; Defflorian, F. Combining ligand-based and structure-based drug design in the virtual screening arena. *Exp. Opin. Drug Discov.* **2007**, *2*, 37-49.
- [138] Cramer, R. D. I.; Patterson, D. E.; Bunce, J. D. Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5969-5967.
- [139] Fredholm, B. B.; Arslan, G.; Halldner, L.; Kull, B.; Schulte, G.; Wasserman, W. Structure and function of adenosine receptors and their genes. *Naunyn-Schmiedeberg's Arch. Pharmacol.* **2000**, *362*, 364-374.
- [140] Ferré, S.; Von Euler, G.; Johansson, B.; Fredholm, B. B.; Fuxe, K. Stimulation of high-affinity adenosine A₂ receptors decreases the affinity of dopamine D₂ receptors in rat striatal membranes. *Proc. Natl. Acad. Sci. USA* **1991**, *88*, 7238-7241.
- [141] Ribeiro, J. A.; Sebastião, A. M.; de Mendonça, A. Adenosine receptors in the nervous system: pathophysiological implications. *Prog. Neurobiol.* **2002**, *68*, 377-392.
- [142] Jacobson, K. A.; Gao, Z. G. Adenosine receptors as therapeutic targets. *Nat. Rev. Drug Discov.* **2006**, *5*, 247-264.
- [143] Xu, K.; Bastia, E.; Schwarzschild, M. Therapeutic potential of adenosine A_{2A} receptor antagonists in Parkinson's disease. *Pharmacol. Ther.* **2005**, *105*, 267-310.
- [144] Johnston, T. H.; Brotchie, J. M. Drugs in development for Parkinson's disease: an update. *Curr. Opin. Investig. Drugs* **2004**, *5*, 720-726.
- [145] Cristalli, G.; Cacciari, B.; Dal Ben, D.; Lambertucci, C.; Moro, S.; Spalluto, G.; Volpini, R. Highlights on the development of A_{2A} adenosine receptor agonists and antagonists. *Chem. Med. Chem.* **2007**, *2*, 260-281.
- [146] Baraldi, P. G.; Tabrizi, M. A.; Bovero, A.; Avitabile, B.; Preti, D.; Fruttarolo, F.; Romagnoli, R.; Varani, K.; Borea, P. A. Recent developments in the field of A_{2A} and A₃ adenosine receptor antagonists. *Eur. J. Med. Chem.* **2003**, *38*, 367-382.
- [147] Baraldi, P. G.; Cacciari, B.; Romagnoli, R.; Spalluto, G.; Monopoli, A.; Ongini, E.; Varani, K.; Borea, P. A. 7-Substituted 5-amino-2-(2-furyl) pyrazolo[4,3-*e*]-1,2,4-triazolo[1,5-*c*]pyrimidines as A_{2A} adenosine receptor antagonists: a study on the importance of modifications at the side chain on the activity and solubility. *J. Med. Chem.* **2002**, *45*, 115-126.

- [148] Baraldi, P. G.; Cacciari, B.; Spalluto, G.; Bergonzoni, M.; Dionisotti, S.; Ongini, E.; Varani, K.; Borea, P. A. Design, synthesis, and biological evaluation of a second generation of pyrazolo[4,3-*e*]-1,2,4-triazolo[1,5-*c*]pyrimidines as potent and selective A_{2A} adenosine receptor antagonists. *J. Med. Chem.* **1998**, *41*, 2126-2133.
- [149] Baraldi, P. G.; Cacciari, B.; Moro, S.; Spalluto, G.; Pastorin, G.; Da Ros, T.; Klotz, K. N.; Varani, K.; Gessi, S.; Borea, P. A. Synthesis, biological activity, and molecular modeling investigation of new pyrazolo[4,3-*e*]-1,2,4-triazolo[1,5-*c*]pyrimidine derivatives as human A₃ adenosine receptor antagonists. *J. Med. Chem.* **2002**, *45*, 770-780.
- [150] Baraldi, P. G.; Cacciari, B.; Romagnoli, R.; Spalluto, G.; Varani, K.; Gessi, S.; Merighi, S.; Borea, P. A. Pyrazolo[4,3-*e*]1,2,4-triazolo[1,5-*c*]pyrimidine derivatives: a new pharmacological tool for the characterization of the human A₃ adenosine receptor. *Drug. Dev. Res.* **2001**, *52*, 406-415.
- [151] Baraldi, P. G.; Fruttarolo, F.; Tabrizi, M. A.; Preti, D.; Romagnoli, R.; El-Kashef, H.; Moorman, A.; Varani, K.; Gessi, S.; Merighi, S.; Borea P. A. Design, synthesis, and biological evaluation of C⁹- and C²-substituted pyrazolo[4,3-*e*]-1,2,4-triazolo[1,5-*c*]pyrimidines as new A_{2A} and A₃ adenosine receptors antagonists. *J. Med. Chem.* **2003**, *46*, 1229-1241.
- [152] Guba, W.; Nettekoven, M.; Püllmann, B.; Riemer, C.; Schmitt, S. Comparison of inhibitory activity of isomeric triazolopyridine derivatives towards adenosine receptor subtypes or do similar structures reveal similar bioactivities? *Bioorg. Med. Chem. Lett.* **2004**, *14*, 3307-3312.
- [153] Brown, G.; Wyatt, J.; Harris, R.; Yao, X. Diversity creation methods: a survey and categorisation. *J. Inf. Fusion.* **2005**, *6*, 1-28.
- [154] Cacciari, B.; Bolcato, C.; Spalluto, G.; Klotz, K.-N.; Bacilieri, M.; Deflorian, F.; Moro S. Pyrazolo-triazolo-pyrimidines as adenosine receptor antagonists: a complete structure-activity profile. *Purinerg. Signal.* **2007**, *3*, 183-193.
- [155] Henry, J. Drug-associated disease: cytochrome P450 interactions. *Crit. Care Clin.* **2006**, *22*, 329-345.
- [156] Lynch, T.; Price, A. The effect of cytochrome P450 metabolism on drug response, interactions, and adverse effects. *Am. Fam. Physician* **2007**, *76*, 391-396.
- [157] Anzenbacher, P.; Anzenbacherová, E. Cytochrome P450 and metabolism of xenobiotics. *Cell. Mol. Life Sci.* **2001**, *58*, 737-747.

BIBLIOGRAPHY

- [158] Kalra, B. S. Cytochrome P450 enzyme isoforms and their therapeutic implications: an update. *Indian J. Med. Sci.* **2007**, *61*, 102-116.
- [159] Brown, C. M.; Reisfeld, B.; Mayeno, A. N. Cytochrome P450: a structure-based summary of biotransformations using representative substrates. *Drug Metab. Rev.* **2008**, *40*, 1-100.
- [160] Ingelman-Sundberg, M. Human drug metabolising cytochrome P450 enzymes: properties and polymorphisms. *Biomed. Life Sci.* **2003**, *369*, 89-104.
- [161] Lewis, D. F. V.; Modi, S. Dickins, M. Structure-activity relationship for human cytochrome P450 substrates and inhibitors. *Drug Metab. Rev.* **2002**, *34*, 69-82.
- [162] de Groot, M. J.; Kirton, S. B.; Sutcliffe, M. J. In silico methods for predicting ligand binding determinants of cytochromes P450. *Curr. Top. Med. Chem.* **2004**, *4*, 1803-1824.
- [163] de Groot, M. J. Designing better drugs: predicting cytochrome P450 metabolism. *Drug Discov. Today* **2006**, *11*, 601-606.
- [164] Crivori, P.; Poggesi, I. Computational approaches for predicting CYP-related metabolism properties in the screening of new drugs. *Eur. J. Med. Chem.* **2006**, *41*, 795-808.
- [165] Li, H.; Sun, J.; Fan, X.; Sui, X.; Zhang, L.; Wang, Y.; He, Z. Considerations and recent advances in QSAR models for cytochrome P450-mediated drug metabolism prediction. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 843-855.
- [166] Vermeulen, N. P. Prediction of drug metabolism: the case of cytochrome P450 2D6. *Curr. Top. Med. Chem.* **2003**, *3*, 1227-1239.
- [167] de Graaf, C.; Vermeulen, N. P. E.; Feenstra, K. A. Cytochrome P450 in silico: an integrative modeling approach. *J. Med. Chem.* **2005**, *48*, 2725-2755.
- [168] Yap, C. W.; Chen, Y. Z. Prediction of cytochrome P450 3A4, 2D6 and 2C9 inhibitors and substrates by using support vector machines. *J. Chem. Inf. Model.* **2005**, *45*, 982-992.
- [169] Yamashita, F.; Hara, H.; Ito, T.; Hashida, M. Novel hierarchical classification and visualization method for multiobjective optimization of drug properties: application to structure-activity relationship analysis of cytochrome P450 metabolism. *J. Chem. Inf. Model.* **2008**, *48*, 364-369.
- [170] Block, J. H.; Henry, D. R. Evaluation of descriptors and classification schemes to predict cytochrome substrates in terms of chemical information. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 385-392.

- [171] Bonnabry, P.; Sievering, J.; Leemann, T.; Dayer, P. Quantitative drug interactions prediction system (Q-DIPS): a dynamic computer-based method to assist in the choice of clinically relevant *in vivo* studies. *Clin. Pharmacokinet.* **2001**, *40*, 631-640.
- [172] Pharmacist's Letter, 2006. <http://www.pharmacistsletter.com>.
- [173] Flockhart. Cytochrome P450 drug-interaction table. <http://medicine.iupui.edu/flockhart/table.htm>.
- [174] Manga, N.; Duffy, J. C.; Rowe, P. H.; Cronin, M. T. Structure-based methods for the prediction of the dominant P450 enzyme in human drug biotransformation: consideration of CYP3A4, CYP2C9, CYP2D6. *QSAR Environ. Res.* **2005**, *16*, 43-61.
- [175] Metabolite database; MDL Inc. <http://www.mdl.com/products/predictive/metabolite/index.jsp>.
- [176] Moro, S.; Bacilieri, M.; Deflorian, F.; Spalluto, G. G protein-coupled receptors as challenging druggable targets: insights from *in silico* studies. *New J. Chem.* **2006**, *30*, 301-308.
- [177] Moro, S.; Deflorian, F.; Bacilieri, M.; Spalluto, G. Ligand-based homology modeling as attractive tool to inspect GPCR structural plasticity. *Curr. Pharm. Des.* **2006**, *12*, 2175-2185.
- [178] Müller, C. E. Medicinal chemistry of adenosine A₃ receptor ligands. *Curr. Top. Med. Chem.* **2003**, *3*, 445-462.
- [179] Jacobson, K. A. *A₃ adenosine receptors*. In Annual Reports in Medicinal Chemistry; Elsevier: San Diego, CA, 2003.
- [180] Baraldi, P. G.; Cacciari, B.; Romagnoli, R.; Spalluto, G.; Moro, S.; Klotz, K. N.; Leung, E.; Varani, K.; Gessi, S.; Merighi, S.; Borea, P. A. Pyrazolo[4,3-*e*]1,2,4-triazolo[1,5-*c*]pyrimidine derivatives as highly potent and selective human A₃ adenosine receptor antagonists: influence of the chain at the N⁸ pyrazole nitrogen. *J. Med. Chem.* **2000**, *43*, 4768-4780.
- [181] Baraldi, P. G.; Cacciari, B.; Romagnoli, R.; Klotz, K. N.; Spalluto, G.; Varani, K.; Gessi, S.; Merighi, S.; Borea, P. A. Pyrazolo[4,3-*e*]1,2,4-triazolo[1,5-*c*]pyrimidine derivatives as adenosine receptor ligands: a starting point for searching A_{2B} adenosine receptor antagonists. *Drug Dev. Res.* **2001**, *53*, 225-235.
- [182] Moro, S.; Gao, Z. G.; Jacobson, K. A.; Spalluto, G. Progress in the pursuit of therapeutic adenosine receptor antagonists. *Med. Res. Rev.* **2006**, *26*, 131-159.

BIBLIOGRAPHY

- [183] Kadam, R.; Chavan, A.; Monga, V.; Kaur, N.; Jain, R.; Roy, N. Selectivity-based QSAR approach for screening and evaluation of TRH analogs for TRH-R1 and TRH-R2 receptors subtypes. *J. Mol. Graph. Model.* **2008**, *27*, 309-320.
- [184] Giorgi, I.; Leonardi, M.; Pietra, D.; Biagi, G.; Borghini, A.; Massarelli, I.; Ciampi, O.; Bianucci, A. M. Synthesis, biological assays and QSAR studies of N-(9-benzyl-2-phenyl-8-azapurin-6-yl)-amides as ligands for A₁ adenosine receptors. *Bioorg. Med. Chem.* **2009**, *17*, 1817-1830.
- [185] Maemoto, T.; Tada, M.; Mihara, T.; Ueyama, N.; Matsuoka, H.; Harada, K.; Yamaji, T.; Shirakawa, K.; Kuroda, S.; Akahane, A.; Iwashita, A.; Matsuoka, N.; Mutoh, S. Pharmacological characterization of FR194921, a new potent, selective, and orally active antagonist for central adenosine A₁ receptors. *J. Pharmacol. Sci.* **2004**, *96*, 42-52.
- [186] Holgate, S. The identification of the adenosine A_{2B} receptor as a novel therapeutic target in asthma. *Br. J. Pharmacol.* **2005**, *145*, 1009-1015.
- [187] Okamura, T.; Kurogi, Y.; Hashimoto, K.; Sato, S.; Nishikawa, H.; Kiryu, K.; Nagao, Y. Structure-activity relationships of adenosine A₃ receptor ligands: new potential therapy for the treatment of glaucoma. *Bioorg. Med. Chem. Lett.* **2004**, *14*, 3775-3779.
- [188] Tabrizi, M. A.; Baraldi, P. G.; Preti, D.; Romagnoli, R.; Saponaro, G.; Baraldi, S.; Moorman, A. R.; Zaid, A. N.; Varani, K.; Borea, P. A. 1,3-Dipropyl-8-(1-phenylacetamide-1H-pyrazol-3-yl)-xanthine derivatives as highly potent and selective human A_{2B} adenosine receptor antagonists. *Bioorg. Med. Chem.* **2008**, *16*, 2419-2430.
- [189] Weyler, S.; Fülle, F.; Diekmann, M.; Schumacher, B.; Hinz, S.; Klotz, K. N., Müller, C. E. Improving potency, selectivity, and water solubility of adenosine A₁ receptor antagonists: xanthines modified at position 3 and related pyrimido[1,2,3-cd]purinediones. *Chem. Med. Chem.* **2006**, *1*, 891-902.
- [190] Elzein, E.; Kalla, R.; Li, X.; Perry, T.; Parkhill, E.; Palle, V.; Varkhedkar, V.; Gimbel, A.; Zeng, D.; Lustig, D.; Leung, K.; Zablocki, J. Novel 1,3-dipropyl-8-(1-heteroaryl-methyl-1H-pyrazol-4-yl)-xanthine derivatives as high affinity and selective A_{2B} adenosine receptor antagonists. *Bioorg. Med. Chem. Lett.* **2006**, *16*, 302-306.
- [191] Kalla, R. V.; Elzein, E.; Perry, T.; Li, X.; Gimbel, A.; Yang, M.; Zeng, D.; Zablocki, J. Selective, high affinity A_{2B} adenosine receptor antagonists: N-1 monosubstituted 8-(pyrazol-4-yl)xanthines. *Bioorg. Med. Chem. Lett.* **2008**, *18*, 1397-1401.

- [192] Baraldi, P. G.; Tabrizi, M. A.; Preti, D.; Bovero, A.; Romagnoli, R.; Fruttarolo, F.; Zaid, N. A.; Moorman, A. R.; Varani, K.; Gessi, S.; Merighi, S.; Borea, P. A. Design, synthesis, and biological evaluation of new 8-heterocyclic xanthine derivatives as highly potent and selective human A_{2B} adenosine receptor antagonists. *J. Med. Chem.* **2004**, *47*, 1434-1477.
- [193] Baraldi, P. G., Preti, D.; Tabrizi, M. A.; Romagnoli, R.; Saponaro, G.; Baraldi, S.; Botta, M.; Bernardini, C.; Tafi, A.; Tuccinardi, T.; Martinelli, A.; Varani, K.; Borea, P. A. Structure-activity relationship studies of a new series of imidazo[2,1-*f*]purinones as potent and selective A_3 adenosine receptor antagonists. *Bioorg. Med. Chem.* **2008**, *16*, 10281-10294.
- [194] Michielan, L.; Bolcato, C.; Stephanie, F.; Cacciari, B.; Bacilieri, M.; Klotz, K. N.; Kachler, S.; Pastorin, G.; Cardin, R.; Sperduti, A.; Moro, S. Combining selectivity and affinity predictions using an integrated support vector machine (SVM) approach: an alternative tool to discriminate between the human adenosine A_{2A} and A_3 receptor pyrazolo-triazolo-pyrimidine antagonists binding sites. *Bioorg. Med. Chem.* **2009**, *17*, 5259-5274.
- [195] Michielan, L.; Bacilieri, M.; Schiesaro, A.; Bolcato, C.; Pastorin, G.; Spalluto, G.; Cacciari, B.; Klotz, K. N.; Kaseda, C.; Moro, S. Linear and nonlinear 3D-QSAR approaches in tandem with ligand-based homology modeling as a computational strategy to depict the pyrazolo-triazolo-pyrimidine antagonists binding site of the human adenosine A_{2A} receptor. *J. Comput. Inf. Model.* **2008**, *48*, 350-363.
- [196] <http://www.accelrys.com/products/topkat/>.
- [197] <http://www.multicase.com/product/prod09.htm>.
- [198] <http://lazar.in-silico.de/>.
- [199] <http://www.epa.gov/oppt/exposure/docs/episuitedl.htm>.
- [200] <http://www.lhasalimited.org/>.
- [201] <http://www.compudrug.com> (commercial).
- [202] <http://www.oasis-lmc.org/software.php>.
- [203] de Roode, D.; Hoekzema, C.; de Vries-Buitenweg, S.; Van de Waart, B.; Van der Hoeven, J. QSARs in ecotoxicological risk assessment. *Regul. Toxicol. Pharm.* **2006**, *45*, 24-35.
- [204] Madden, J. C.; Enoch, S. J.; Hewitt, M.; Cronin, M. T. D. Pharmaceuticals in the environment: good practice in predicting acute ecotoxicological effects. *Toxicol. Lett.* **2009**, *185*, 85-101.

BIBLIOGRAPHY

- [205] Reuschenbach, P.; Silvani, M.; Dammann, M.; Warnecke, D.; Knacker, T. ECOSAR model performance with a large test set of industrial chemical. *Chemosphere*, **2008**, *71*, 1986-1995.
- [206] Sanderson, H.; Thomsen, M. Comparative analysis of pharmaceuticals versus industrial chemicals acute aquatic toxicity classification according to the United Nations classification system for chemicals. Assessment of the (Q)SAR predictability of pharmaceuticals acute aquatic toxicity and their predominant acute toxic mode-of-action. *Toxicol. Lett.* **2009**, *187*, 84-93.
- [207] Papa, E.; Battaini, F.; Gramatica, P. Ranking of aquatic toxicity of esters modelled by QSAR. *Chemosphere*, **2005**, *58*, 559-570.
- [208] Lo Piparo, E.; Fratev, F.; Lemke, F.; Mazzatorta, P.; Smiesko, M.; Fritz, J. I.; Benfenati, E. QSAR models for *Daphnia magna* toxicity prediction of benzoxazinone allelochemicals and their transformation products. *J. Agr. Food Chem.* **2006**, *54*, 1111-1115.
- [209] Hodges, G.; Roberts, D. W.; Marshall, S. J.; Dearden, J. C. The aquatic toxicity of anionic surfactants to *Daphnia magna* - A comparative QSAR study of linear alkylbenzene sulphonates and ester sulphonates. *Chemosphere* **2006**, *63*, 1443-1450.
- [210] Casalegno, M.; Benfenati, E.; Sello, G. An automated group contribution method in predicting aquatic toxicity: the diatomic fragment approach. *Chem. Res. Toxicol.* **2005**, *18*, 740-746.
- [211] Casalegno, M.; Sello, G. Quantitative aquatic toxicity prediction: using group contribution and classification methods on polar and non-polar narcotics. *J. Mol. Struc.-Teochem.* **2005**, *727*, 71-80.
- [212] Yuan, H.; Wang, Y.-Y.; Cheng, Y.-Y. Mode of action-based local QSAR modeling for the prediction of acute toxicity in the fathead minnow. *J. Mol. Graph. Model.* **2007**, *26*, 327-335.
- [213] Colombo, A.; Benfenati, E.; Karelson, M.; Maran, U. The proposal of architecture for chemical splitting to optimize QSAR models for aquatic toxicity. *Chemosphere* **2008**, *72*, 772-780.
- [214] http://www.unece.org/trans/danger/publi/ghs/ghs_rev02/02files_e.html (Part 4 - Environmental hazards).
- [215] Verhaar, H. J. M.; Van Leeuwen, C. J.; Hermens, J. L. M. Classifying environmental pollutants. 1: Structure-activity relationships for prediction of aquatic toxicity. *Chemosphere*, **1992**, *25*, 471-491.

- [216] <http://ecb.jrc.it/qsar/qsar-tools/>.
- [217] Russom, C. L.; Bradbury, S. P.; Broderius, S. J.; Hammermeister, D. E.; Drummond, R. A. Predicting modes of toxic action from chemical structure: acute toxicity in the fathead minnow (*Pimephales promelas*). *Environ. Toxicol Chem.* **1997**, *16*, 948-967. Dataset available on line at http://www.epa.gov/ncct/dsstox/sdf_epafhm.html.
- [218] Enoch, S. J.; Hewitt, M.; Cronin, M. T. D.; Azam, S.; Madden, J. C. Classification of chemicals according to mechanism of aquatic toxicity: an evaluation of the implementation of the Verhaar scheme in Toxtree. *Chemosphere* **2008**, *73*, 243-248.
- [219] Spycher, S.; Nendza, M.; Gasteiger, J. Comparison of different classification methods applied to a mode of toxic action data set. *QSAR Comb. Sci.* **2004**, *23*, 779-791.
- [220] Yao, X. J.; Panaye, A.; Doucet, J. P.; Chen, H. F.; Zhang, R. S.; Fan, B. T.; Liu, M. C.; Hu, Z. D. Comparative classification study of toxicity mechanisms using support vector machines and radial basis function neural networks. *Anal. Chim. Acta* **2005**, *535*, 259-273.
- [221] Norinder, U.; Lidén, P.; Boström, H. Discrimination between modes of toxic action of phenols using rule based methods. *Mol. Divers.* **2006**, *10*, 207-212.
- [222] Niu, B.; Jin, Y.; Lu, W.; Li, G. Predicting toxic action mechanisms of phenols using AdaBoost learner. *Chemometr. Intell. Lab.* **2009**, *96*, 43-48.
- [223] Du, H.; Wang, J.; Watzl, J.; Zhang, X.; Hu, Z. Classification structure-activity relationship (CSAR) studies for prediction of genotoxicity of thiophene derivatives. *Toxicol. Lett.* **2008**, *177*, 10-19.
- [224] Maunz, A.; Helma, C. Prediction of chemical toxicity with local support vector regression and activity-specific kernels. *SAR QSAR Environ. Res.* **2008**, *19*, 413-431.
- [225] Leonard, J. T.; Roy, K. Comparative QSAR modeling of CCR5 receptor binding affinity of substituted 1-(3,3-diphenylpropyl)-piperidinyl amides and ureas. *Bioorg. Med. Chem. Lett.* **2006**, *16*, 4467-4474.

Supporting Information

In the electronic version of this thesis you find attached files containing the supplementary information material.

Appendix A includes the training set, the test set with the relative references used to develop the *autoMEP*/RSA model for aqueous solvation free energy prediction, as described in chapter 3 (*Paper I*).

Appendix B reports the binding affinity data for the training set, both internal and external test sets, with the corresponding references, utilized for the generation of linear *autoMEP*/PLS and nonlinear *autoMEP*/RSA models derived in chapter 4 (*Paper II*).

Appendix C collects the whole dataset of cytochrome P450 substrates, with the relative references, as described in chapter 5 (*Paper III*). Additional tables reporting the results of a multi-classification model by considering 7 classes, together with further parameters corresponding to the models derived by using 5 classes are included.

Appendix D reports both training sets used in the *SVMclass* and *SVR* models, as summarized in chapter 6 (*Paper IV*). Moreover, the binding affinity data for the test set is reported.

Appendix E includes the training set, validation set and internal test set used in the multi-classification models to predict potency and selectivity of adenosine receptor antagonists, as described in chapter 7 (*Paper V*), together with further tables reporting additional information on the validation process and the predictions on the external test set.

Appendix F reports both training sets utilized in chapter 8 for the generation of *TOXclass* and *MOAclass* classifiers, with the predicted toxicodynamic and toxicokinetic information about the test set (*Paper VI*).

APPENDIX A

Paper I

Supplementary Information

A.1 Training set.

A.2 Test set.

A.3 References.

A.1 Training set

1-248: [1]

Mol Id	Molecule name	Exp. ΔG_{hyd} (kcal/mol)	Pred. ΔG_{hyd} (kcal/mol)
1	methane	1.98	2.55
2	ethane	1.81	2.22
3	propane	2.02	2.08
4	<i>n</i> -butane	2.18	2.26
5	2-methylpropane	2.32	2.29
6	<i>n</i> -pentane	2.36	2.25
7	2,2-dimethylpropane	2.69	2.65
8	cyclopentane	1.22	1.87
9	<i>n</i> -hexane	2.58	2.07
10	3-methylpentane	2.54	2.57
11	cyclohexane	1.24	2.08
12	methylcyclopentane	1.62	1.87
13	2,2-dimethylbutane	2.63	2.63
14	<i>n</i> -heptane	2.65	2.49
15	2,4-dimethylpentane	2.92	2.48
16	methylcyclohexane	1.73	1.92
17	<i>n</i> -octane	2.93	2.43
18	2,2,4-trimethylpentane	2.89	2.47
19	ethylene	1.30	0.32
20	propylene	1.28	1.23
21	1-butene	1.40	1.51
22	2-methylpropene	1.30	1.06
23	1-pentene	1.69	1.80
24	<i>trans</i> -2-pentene	1.35	1.09
25	cyclopentene	0.57	1.29
26	2-methyl-2-butene	1.33	0.78
27	3-methyl-1-butene	1.85	1.58
28	cyclohexene	0.37	1.20
29	4-methyl-1-pentene	1.93	1.72
30	<i>trans</i> -2-heptene	1.69	1.61
31	1-methylcyclohexene	0.68	0.38
32	1-octene	2.20	1.85
33	1,3-butadiene	0.57	0.46
34	1,4-pentadiene	0.95	1.21
35	2-methyl-1,3-butadiene	0.69	0.44
36	1,5-hexadiene	1.02	1.39
37	propyne	-0.48	-1.42
38	butyne	-0.17	-1.67
39	1-pentyne	0.01	-0.32
40	1-hexine	0.29	-0.37

Training set: continued

Mol Id	Molecule name	Exp. ΔG_{hyd} (kcal/mol)	Pred. ΔG_{hyd} (kcal/mol)
41	1-heptyne	0.61	0.25
42	1-octyne	0.72	0.55
43	1-nonyne	1.06	0.37
44	benzene	-0.90	-0.30
45	ethylbenzene	-0.62	0.01
46	<i>o</i> -xylene	-0.91	-0.50
47	<i>m</i> -xylene	-0.82	-0.41
48	<i>p</i> -xylene	-0.82	-0.53
49	propylbenzene	-0.54	-0.10
50	2-propylbenzene	-0.30	-0.61
51	1,2,4-trimethylbenzene	-0.87	-0.15
52	butylbenzene	-0.40	-0.07
53	2-butylbenzene	-0.46	-0.29
54	<i>tert</i> -amylbenzene	-0.18	-0.30
55	naphthalene	-2.45	-0.48
56	acenaphthene	-3.44	-2.07
57	anthracene	-4.34	-2.97
58	phenanthrene	-4.12	-3.98
59	chloromethane	-0.54	-2.39
60	trichloromethane	-1.04	-2.14
61	tetrachloromethane	0.10	0.20
62	bromomethane	-0.80	-1.45
63	dibromomethane	-1.99	-1.17
64	tribromomethane	-2.16	-1.91
65	iodomethane	-0.90	-0.85
66	chlorofluoromethane	-0.79	-2.47
67	chlorotrifluoromethane	2.56	-0.56
68	dichlorodifluoromethane	1.71	-0.11
69	chloroethane	-0.64	-1.03
70	bromoethane	-0.70	-0.12
71	iodoethane	-0.73	-0.22
72	1,1-dichloroethane	-0.86	-2.12
73	1,2-dichloroethane	-1.75	-2.34
74	1,2-dibromoethane	-2.13	-2.99
75	1-chloro-2-bromoethane	-1.98	-2.88
76	1,1,1-trichloroethane	-0.25	-0.76
77	1,1,2-trichloroethane	-1.98	-2.31
78	pentachloroethane	-1.38	-3.21
79	hexachloroethane	-1.42	0.38
80	chloropentafluoroethane	2.90	-0.08
81	1,1,2,2-tetrachloro-difluoroethane	0.83	1.02
82	1,1,2-trichloro-trifluoroethane	1.80	1.25
83	1,1-dichloro-tetrafluoroethane	2.54	0.68

Training set: continued

Mol Id	Molecule name	Exp. ΔG_{hyd} (kcal/mol)	Pred. ΔG_{hyd} (kcal/mol)
84	1,2-dichloro-tetrafluoroethane	2.35	1.44
85	1-chloropropane	-0.36	-0.71
86	2-chloropropane	-0.25	-0.87
87	1-bromopropane	-0.57	-0.35
88	2-bromopropane	-0.48	-0.29
89	1-iodopropane	-0.62	-0.32
90	2-iodopropane	-0.47	-0.49
91	1,2-dichloropropane	-1.27	-2.88
92	1,3-dichloropropane	-1.92	-3.28
93	1,2-dibromopropane	-1.96	-2.10
94	1-chlorobutane	-0.14	-0.24
95	1-bromobutane	-0.41	-0.35
96	1-bromo-2-methylpropane	-0.03	-0.33
97	1-iodobutane	-0.26	-0.29
98	1,1-dichlorobutane	-0.70	-0.95
99	1-chloropentane	-0.07	-0.87
100	2-chloropentane	0.07	0.04
101	3-chloropentane	0.04	-0.13
102	1-bromo-3-methylbutane	0.21	-1.30
103	<i>cis</i> -1,2-dichloroethylene	-1.19	-1.37
104	<i>trans</i> -1,2-dichloroethylene	-0.77	-0.82
105	1,2,3-trichloroethylene	-0.44	-0.82
106	tetrachloroethylene	0.06	-0.76
107	3-chloropropene	-0.58	-0.70
108	chlorobenzene	-1.02	-1.93
109	bromobenzene	-1.48	-2.15
110	1,2-dichlorobenzene	-1.38	-1.30
111	1,3-dichlorobenzene	-0.99	-1.46
112	1,4-dibromobenzene	-2.33	-2.46
113	<i>p</i> -bromotoluene	-1.41	-1.29
114	1-bromo-2-ethylbenzene	-1.20	-1.02
115	<i>o</i> -bromocumene	-0.86	-0.80
116	dimethyl ether	-1.92	-1.88
117	dimethyl sulfide	-1.56	-1.09
118	1,3-dioxolane	-4.14	-5.60
119	diethyl ether	-1.77	-2.00
120	methylpropyl ether	-1.69	-2.34
121	methyl isopropyl ether	-2.03	-2.00
122	tetrahydrofuran	-3.51	-3.20
123	dioxane	-5.11	-3.67
124	ethylpropyl ether	-1.84	-1.23
125	methyl <i>tert</i> -butyl ether	-2.24	-2.69
126	2-methyltetrahydrofuran	-3.34	-3.11

Training set: continued

Mol Id	Molecule name	Exp. ΔG_{hyd} (kcal/mol)	Pred. ΔG_{hyd} (kcal/mol)
127	tetrahydropyran	-3.16	-2.32
128	dipropyl ether	-1.17	-1.10
129	1,2-dietoxyethane	-3.30	-1.94
130	1,1-dietoxyethane	-3.32	-3.43
131	di- <i>n</i> -butyl ether	-0.84	-0.38
132	anisole	-1.05	-3.11
133	thioanisole	-2.76	-1.21
134	2,2'-dichlorodietyl sulfide	-3.97	-1.68
135	methanol	-5.14	-6.09
136	methan thiol	-1.26	-2.17
137	ethanol	-4.96	-4.66
138	2,2,2-trifluoroethanol	-4.35	-4.27
139	1-propanol	-4.92	-4.24
140	2-propanol	-4.81	-3.81
141	allyl alcohol	-5.10	-5.78
142	1,1,1-trifluoro-2-propanol	-4.21	-6.09
143	2,2,3,3-tetrafluoropropanol	-4.96	-4.77
144	2,2,3,3,3-pentafluoropropanol	-4.20	-4.95
145	1-butanol	-4.78	-4.85
146	2-butanol	-4.67	-3.80
147	<i>tert</i> -butyl alcohol	-4.57	-2.86
148	2-methyl-1-propanol	-4.57	-3.34
149	1-pentanol	-4.55	-4.46
150	2-pentanol	-4.45	-4.18
151	2-methyl-1-butanol	-4.48	-4.31
152	2-methyl-2-butanol	-4.49	-3.42
153	1-hexanol	-4.42	-4.15
154	cyclohexanol	-5.02	-4.74
155	2,3-dimethylbutanol	-3.97	-3.62
156	2-methyl-3-pentanol	-3.94	-4.33
157	4-methyl-2-pentanol	-3.79	-4.11
158	2-methyl-2-pentanol	-3.98	-3.09
159	1-heptanol	-4.31	-3.94
160	1-octanol	-4.16	-4.39
161	phenol	-6.62	-7.13
162	4-bromophenol	-7.20	-5.30
163	thiophenol	-2.58	-4.39
164	2-cresol	-5.94	-6.29
165	4-nitrophenol	-6.20	-6.26
166	4- <i>tert</i> -butylphenol	-6.00	-4.86
167	acetaldehyde	-3.55	-3.00
168	propanal	-3.48	-3.30
169	butanal	-3.22	-3.56

Training set: continued

Mol Id	Molecule name	Exp. ΔG_{hyd} (kcal/mol)	Pred. ΔG_{hyd} (kcal/mol)
170	pentanal	-3.07	-2.86
171	heptanal	-2.71	-2.24
172	octanal	-2.32	-2.45
173	nonanal	-2.10	-3.23
174	<i>trans</i> -2-butenal	-4.28	-4.03
175	<i>trans</i> -2-hexenal	-3.73	-4.26
176	<i>trans</i> -2-octenal	-3.48	-3.44
177	<i>trans-trans</i> -2,4-hexadienal	-4.70	-3.21
178	benzaldehyde	-4.08	-3.43
179	acetone	-3.85	-2.22
180	2-pentanone	-3.56	-2.33
181	2-heptanone	-3.11	-2.75
182	2-octanone	-2.92	-2.22
183	2-nonanone	-2.51	-1.56
184	2-undecanone	-2.18	-1.16
185	acetophenone	-4.64	-4.32
186	ethylformate	-2.68	-2.39
187	methylacetate	-3.36	-4.03
188	propylformate	-2.51	-2.71
189	isopropylformate	-2.04	-2.28
190	ethylacetate	-3.12	-3.08
191	methylpropionate	-3.01	-3.53
192	isobutylformate	-2.25	-2.44
193	propylacetate	-2.89	-2.86
194	isopropylacetate	-2.68	-2.95
195	methylbutyrate	-2.87	-2.93
196	isoamylformate	-2.16	-2.70
197	butylacetate	-2.58	-2.45
198	isobutylacetate	-2.39	-2.48
199	propylpropionate	-2.49	-2.38
200	isopropylpropionate	-2.25	-2.61
201	ethylbutyrate	-2.53	-2.47
202	methylpentanoate	-2.57	-2.66
203	amylacetate	-2.49	-2.53
204	propylbutyrate	-2.31	-2.51
205	ethylpentanoate	-2.56	-2.09
206	methylhexanoate	-2.51	-2.87
207	hexylacetate	-2.29	-3.13
208	amylpropionate	-2.02	-2.38
209	methyloctanoate	-2.07	-3.73
210	ethylheptanoate	-2.33	-2.10
211	methylbenzoate	-4.34	-4.76
212	ethylamine	-4.67	-3.76

Training set: continued

Mol Id	Molecule name	Exp. ΔG_{hyd} (kcal/mol)	Pred. ΔG_{hyd} (kcal/mol)
213	butylamine	-4.43	-3.81
214	pentylamine	-4.14	-4.15
215	hexylamine	-4.09	-2.43
216	dimethylamine	-4.34	-3.83
217	diethylamine	-4.12	-2.39
218	pyrrolidine	-5.54	-3.71
219	piperidine	-5.17	-3.74
220	dipropylamine	-3.70	-1.16
221	hexamethyleneimine	-4.97	-3.42
222	trimethylamine	-3.27	-4.15
223	triethylamine	-3.07	-3.61
224	<i>n</i> -methylpyrrolidine	-4.02	-5.29
225	<i>n</i> -methylpiperidine	-3.94	-5.04
226	propionitrile	-3.90	-4.34
227	butyronitrile	-3.69	-5.07
228	nitroethane	-3.76	-4.51
229	2-nitropropane	-3.18	-3.58
230	nitrobenzene	-4.17	-4.54
231	2-nitrotoluene	-3.63	-3.67
232	3-nitrotoluene	-3.50	-3.87
233	pyridine	-4.75	-4.26
234	2-methylpyridine	-4.68	-4.58
235	3-methylpyridine	-4.84	-4.70
236	4-methylpyridine	-4.99	-5.01
237	2-ethylpyridine	-4.38	-4.38
238	4-ethylpyridine	-4.78	-2.45
239	2,3-dimethylpyridine	-4.88	-5.02
240	2,4-dimethylpyridine	-4.92	-4.97
241	2,5-dimethylpyridine	-4.77	-4.55
242	2,6-dimethylpyridine	-4.66	-3.66
243	3,4-dimethylpyridine	-5.28	-4.85
244	3,5-dimethylpyridine	-4.90	-4.76
245	2-methylpyrazine	-5.58	-4.45
246	2-ethylpyrazine	-5.53	-4.23
247	2-ethyl-3-methoxypyrazine	-4.45	-4.19
248	2-isobutyl-3-methoxypyrazine	-3.73	-4.56

A.2 Test set

1-23: [1]

Mol Id	Molecule name	Exp. ΔG_{hyd} (kcal/mol)	Pred. ΔG_{hyd} (kcal/mol)
1	2-methylpentane	2.56	2.53
2	<i>cis</i> -1,2-dimethylcyclohexane	1.60	2.38
3	1-hexene	1.73	1.96
4	2,3-dimethyl-1,3-butadiene	0.40	0.33
5	toluene	-0.77	0.28
6	<i>tert</i> -butylbenzene	-0.44	-0.3
7	dichloromethane	-1.42	-1.82
8	1,3-dibromopropane	-1.99	-1.6
9	chloroethylene	0.50	-0.76
10	1,4-dichlorobenzene	-1.02	-1.75
11	diethyl sulfide	-1.45	-0.8
12	diisopropyl ether	-0.54	-0.96
13	ethane thiol	-4.08	-2.98
14	3-hexanol	-3.73	-3.8
15	hexanal	-2.85	-2.54
16	2-butanone	-3.76	-2.63
17	methylformate	-2.82	-4.34
18	ethylpropionate	-2.83	-2.75
19	isoamylacetate	-2.24	-2.54
20	propylamine	-4.56	-3.88
21	dibutylamine	-3.38	-1.46
22	1-nitropropane	-3.38	-4.18
23	2-isobutylpyrazine	-5.11	-3.84

A.3 References

- [1] Viswanadhan, V. N.; Ghose, A. K.; Singh, U. C.; Wendoloski, J. J. Prediction of solvation free energies of small organic molecules: additive-constitutive models based on molecular fingerprints and atomic constants. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 405-412.

APPENDIX B

Paper II

Supplementary Information

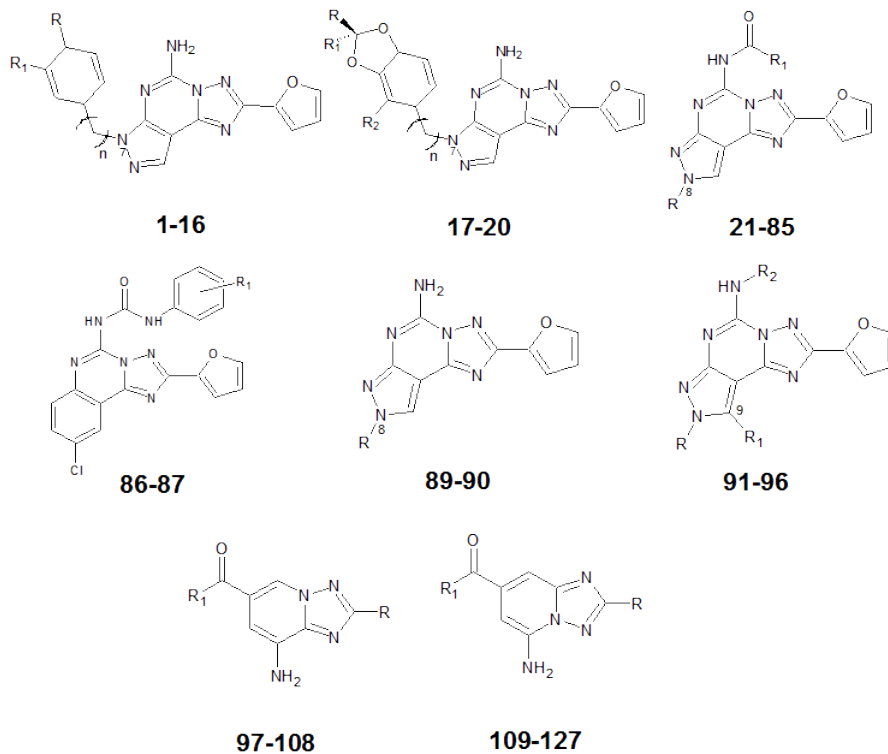
B.1 Training set.

B.2 Internal test set.

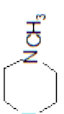
B.3 External test set.

B.4 References.

B.1 Training set



1-15: [1]; 16: [2]; 17-18: [3]; 19-20: [1]; 21-81: [4]; 82-84: [5]; 85: [1]; 86-87: [4]; 88-90: [5]; 91-96: [6]; 97-108: [7]; 109-127: [7]

Mol. Id	n	R	R ₁ /Amine*	R ₂	Pred. pK _i (nM) (PLS)	Pred. pK _i (nM) (RSA)	Exp. pK _i (nM) hA _{2A} R	Exp. K ₁ (nM) hA _{2A} R
1	2	N(CH ₂ CH ₂ OH) ₂	H	-	-0.49	0.14	0.92	0.12
2	3	NH ₂	H	-	-0.24	0.11	0.66	0.22
3	3	NO ₂	H	-	-0.14	-0.49	0.00	100
4	2	NHCOCH ₃	H	-	-0.45	-0.38	-0.68	4.80
5	3	N(CH ₂ CH ₂ OH) ₂	H	-	-0.52	0.16	-0.04	1.10
6	3	COOCH ₂ CH ₃	H	-	-0.08	-0.52	-0.60	4.00
7	2	OCH ₂ COOCH ₂ CH ₃	H	-	-0.27	-0.45	0.37	0.43
8	3	C(NOH)NH ₂	H	-	-0.40	-0.54	-0.78	6.00
9	3	COOH	H	-	-0.34	-0.40	-0.67	4.63
10	2	SO ₂ NH ₂	H	-	-1.02	-1.65	-0.12	1.31
11	2	SO ₂ N(CH ₂ CH ₂ OH) ₂	H	-	-0.11	0.07	0.10	0.80
12	2	SO ₂ N 	H	-	-0.24	-0.38	-0.58	3.80
13	2	SO ₂ N(CH ₂ CH ₂ Cl) ₂	H	-	-0.75	-0.88	0.23	0.59
14	2	SO ₂ NHCH ₂ COOH	H	-	-0.61	-0.23	-1.70	50.0
15	3	H	H	-	-0.20	-0.16	-0.08	1.20
16	2	H	H	-	-0.14	-0.08	-0.04	1.10
17	3	OCH ₂ O	H	-	-0.04	-0.42	0.52	3.30
18	3	OH	H	-	-0.12	-0.01	-0.18	1.50
19	3	CH ₂ OH	CH ₂ OH	H	-0.01	-0.73	0.72	0.19
20	3	COOCH ₂ CH ₃	H	H	-0.14	-0.39	-0.74	5.48
21	-	CH ₃	3,4-Cl ₂ -Ph-NH	-	-2.42	-2.38	-2.16	143

Training set: continued

Mol Id	n	R	R ₁ /Amine*	R ₂	Pred. p <i>K</i> _i (nM) (PLS)	Pred. p <i>K</i> _i (nM) (RSA)	Exp. p <i>K</i> _i (nM) hA _{2A} R	Exp. <i>K</i> _i (nM) hA _{2A} R
22	-	CH ₃	3,4-OCH ₂ O-Ph-NH	-	-2.38	-2.67	-2.83	680
23	-	CH ₃	4-NO ₂ -Ph-NH	-	-2.81	-2.74	-2.84	695
24	-	CH ₃	4-CH ₃ -Ph-NH	-	-2.57	-2.40	-2.04	110
25	-	CH ₃	4-Br-Ph-NH	-	-2.47	-2.13	-2.00	100
26	-	CH ₃	4-F-Ph-NH	-	-2.75	-2.53	-2.08	120
27	-	CH ₃	4-CF ₃ -Ph-NH	-	-2.37	-2.14	-2.15	140
28	-	CH ₃	2-OCH ₃ -Ph-NH	-	-2.44	-2.28	-2.26	180
29	-	CH ₃	3-OCH ₃ -Ph-NH	-	-2.35	-2.74	-2.20	160
30	-	CH ₃	2-Cl-Ph-NH	-	-2.48	-1.88	-2.30	200
31	-	CH ₃	4-Cl-Ph-NH	-	-2.56	-2.16	-2.26	180
32	-	C ₂ H ₅	3,4-Cl ₂ -Ph-NH	-	-1.79	-2.39	-2.55	352
33	-	C ₂ H ₅	3,4-OCH ₂ O-Ph-NH	-	-2.05	-2.36	-2.76	576
34	-	C ₂ H ₅	4-CH ₃ -Ph-NH	-	-2.15	-2.13	-1.48	30
35	-	C ₂ H ₅	4-Br-Ph-NH	-	-2.12	-2.05	-1.60	40
36	-	C ₂ H ₅	4-F-Ph-NH	-	-2.30	-2.08	-1.78	60
37	-	C ₂ H ₅	4-CF ₃ -Ph-NH	-	-1.90	-1.93	-1.72	53
38	-	C ₂ H ₅	2-OCH ₃ -Ph-NH	-	-2.18	-2.09	-2.12	133
39	-	C ₂ H ₅	3-OCH ₃ -Ph-NH	-	-2.03	-2.26	-2.15	140
40	-	C ₂ H ₅	2-Cl-Ph-NH	-	-2.07	-1.91	-2.18	150
41	-	C ₂ H ₅	4-Cl-Ph-NH	-	-2.16	-1.86	-2.20	160
42	-	<i>n</i> -C ₃ H ₇	3,4-Cl ₂ -Ph-NH	-	-1.59	-2.01	-2.60	401





Training set: continued

Mol. Id	n	R	R ₁ / Amine*	R ₂	Pred. pK _i (nM) (PLS)	Pred. pK _i (nM) (RSA)	Exp. pK _i (nM) hA _{2A} R	Exp. K _i (nM) hA _{2A} R
43	-	<i>n</i> -C ₃ H ₇	3,4-OCH ₂ -Ph-NH	-	-1.85	-2.23	-2.82	667
44	-	<i>n</i> -C ₃ H ₇	4-NO ₂ -Ph-NH	-	-2.06	-2.44	-3.05	1115
45	-	<i>n</i> -C ₃ H ₇	4-CH ₃ -Ph-NH	-	-1.92	-2.08	-1.08	12
46	-	<i>n</i> -C ₃ H ₇	4-Br-Ph-NH	-	-1.88	-1.87	-1.70	50
47	-	<i>n</i> -C ₃ H ₇	4-F-Ph-NH	-	-2.06	-2.01	-1.62	42
48	-	<i>n</i> -C ₃ H ₇	4-CF ₃ -Ph-NH	-	-1.68	-1.79	-1.53	34
49	-	<i>n</i> -C ₃ H ₇	2-OCH ₃ -Ph-NH	-	-1.99	-1.92	-2.00	100
50	-	<i>n</i> -C ₃ H ₇	3-OCH ₃ -Ph-NH	-	-1.89	-2.14	-2.05	113
51	-	<i>n</i> -C ₃ H ₇	2-Cl-Ph-NH	-	-1.83	-1.97	-2.08	121
52	-	<i>n</i> -C ₃ H ₇	4-Cl-Ph-NH	-	-1.93	-1.77	-2.15	140
53	-	<i>n</i> -C ₄ H ₉	3,4-OCH ₂ O-Ph-NH	-	-1.75	-1.83	-2.58	376
54	-	<i>n</i> -C ₄ H ₉	4-NO ₂ -Ph-NH	-	-1.90	-2.36	-2.70	503
55	-	<i>n</i> -C ₄ H ₉	4-CH ₃ -Ph-NH	-	-1.82	-2.19	-1.38	24
56	-	<i>n</i> -C ₄ H ₉	4-Br-Ph-NH	-	-1.77	-1.99	-1.82	66
57	-	<i>n</i> -C ₄ H ₉	4-F-Ph-NH	-	-1.91	-1.80	-1.70	50
58	-	<i>n</i> -C ₄ H ₉	4-CF ₃ -Ph-NH	-	-1.60	-1.79	-1.49	31
59	-	<i>n</i> -C ₄ H ₉	2-OCH ₃ -Ph-NH	-	-1.90	-1.76	-1.96	91
60	-	<i>n</i> -C ₄ H ₉	3-OCH ₃ -Ph-NH	-	-1.78	-2.19	-1.98	95
61	-	<i>n</i> -C ₄ H ₉	2-Cl-Ph-NH	-	-1.76	-2.01	-2.00	100
62	-	<i>n</i> -C ₄ H ₉	4-Cl-Ph-NH	-	-1.80	-1.75	-2.05	111
63	-	CH ₃	Ph-CH ₂	-	-2.30	-1.96	-2.63	423


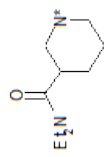
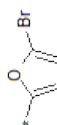




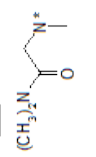

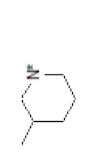

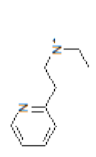
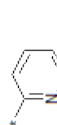

Training set: continued

Mol Id	n	R	R ₁ /Amine*	R ₂	Pred. pK _i (nM) (PLS)	Pred. pK _i (nM) (RSA)	Exp. pK _i (nM) hA _{2A} R	Exp. K _i (nM) hA _{2A} R
64	-	C ₂ H ₅	Ph-CH ₂	-	-1.98	-1.89	-2.53	335
65	-	<i>n</i> -C ₃ H ₇	Ph-CH ₂	-	-1.68	-1.96	-2.49	306
66	-	<i>n</i> -C ₄ H ₉	Ph-CH ₂	-	-1.65	-1.92	-2.60	400
67	-	CH ₃	Ph-NH	-	-2.74	-2.60	-2.58	381
68	-	C ₂ H ₅	Ph-NH	-	-2.28	-2.27	-1.60	40
69	-	<i>n</i> -C ₃ H ₇	Ph-NH	-	-2.04	-1.86	-1.79	62
70	-	<i>n</i> -C ₄ H ₉	Ph-NH	-	-1.94	-1.78	-2.17	148
71	-	CH ₃	4-SO ₃ -Ph-NH	-	-2.88	-2.65	-2.77	594
72	-	C ₂ H ₅	4-SO ₃ -Ph-NH	-	-2.37	-2.62	-2.40	249
73	-	<i>n</i> -C ₃ H ₇	4-SO ₃ -Ph-NH	-	-2.05	-2.62	-2.48	305
74	-	<i>n</i> -C ₄ H ₉	4-SO ₃ -Ph-NH	-	-1.97	-2.46	-2.55	352
75	-	CH ₃	4-OCH ₃ -Ph-NH	-	-2.77	-2.43	-3.14	1390
76	-	CH ₃	3-Cl-Ph-NH	-	-2.19	-2.45	-3.08	1195
77	-	C ₂ H ₅	3-Cl-Ph-NH	-	-1.98	-2.44	-2.26	180
78	-	<i>n</i> -C ₃ H ₇	4-OCH ₃ -Ph-NH	-	-2.21	-2.01	-2.15	140
79	-	<i>n</i> -C ₃ H ₇	3-Cl-Ph-NH	-	-1.73	-2.27	-3.09	1220
80	-	<i>n</i> -C ₄ H ₉	4-OCH ₃ -Ph-NH	-	-2.07	-2.22	-1.90	80
81	-	<i>n</i> -C ₄ H ₉	3-Cl-Ph-NH	-	-1.64	-2.01	-1.98	95
82	-	Ph-CH ₂ CH ₂	4-OCH ₃ -Ph-NH	-	-1.99	-2.02	-2.08	120
83	-	Ph-CH ₂ CH ₂	3-Cl-Ph-NH	-	-1.84	-1.62	-2.02	105



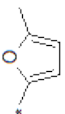

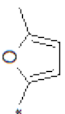



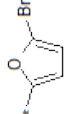






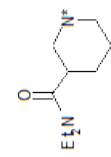
Training set: continued

Mol. Id	n	R	R ₁ / Amine*	R ₂	Pred. pK _i (nM) (PLS)	Pred. pK _i (nM) (RSA)	Exp. pK _i (nM) hA _{2A} R	Exp. K _i (nM) hA _{2A} R
84	-	Ph-CH ₂ CH ₂ CH ₂ CH ₂	3-Cl-Ph-NH	-	-1.51	-1.36	-2.30	200
85	-	H	4-OCH ₃ -Ph-NH	-	-3.41	-2.87	-2.72	520
86	-	-	4-OCH ₃ -Ph-NH	-	-1.80	-2.07	-0.97	9.40
87	-	-	3-Cl-Ph-NH	-	-1.46	-1.72	-0.98	9.50
88	-	<i>n</i> -C ₃ H ₇	-	-	-1.16	-1.36	-0.40	2.51
89	-	Ph-CH ₂ CH ₂	-	-	-0.96	-0.22	0.47	0.34
90	-	Ph-CH ₂ CH ₂ CH ₂ CH ₂	-	-	-0.77	-0.16	0.80	0.16
91	-	CH ₃	SCH ₃	H	-0.60	-0.81	-0.08	1.2
92	-	CH ₃	S(CH ₂) ₂ CH ₃	H	-0.26	0.08	-0.32	2.1
93	-	CH ₃	NHCH ₂ CH ₃	CO-NH-Ph-4-OCH ₃	-2.11	-2.26	-1.32	21
94	-	CH ₃	SCH ₃	COCH ₂ Ph-4-OCH ₃	-1.70	-2.16	-1.18	15
95	-	CH ₃	SCH ₃	COCH ₂ Ph-4-isobutyl	-1.40	-1.32	-1.70	50
96	-	CH ₃	NHCH ₂ CH ₃	COCH ₂ Ph-3,4-medioxy	-1.68	-1.59	1.79	61
97	-			-	-1.31	-1.09	-1.54	35
98	-			-	-1.06	-0.90	-1.43	27







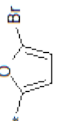







Training set: continued

Mol Id	n	R	R ₁ /Amine*	R ₂	Pred. p <i>K</i> _i (nM) (PLS)	Pred. p <i>K</i> _i (nM) (RSA)	Exp. p <i>K</i> _i (nM) hA _{2A} R	Exp. <i>K</i> _i (nM) hA _{2A} R
99	—			—	-0.84	-0.58	-1.60	40
100	—			—	-1.09	-1.67	-1.20	16
101	—			—	-1.10	-1.18	-1.79	61
102	—			—	-2.01	-1.77	-1.36	23
103	—			—	-1.02	-0.91	-1.18	15
104	—			—	-0.95	-1.08	-1.88	76
105	—			—	-1.44	-1.37	-2.28	192

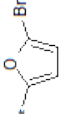
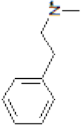








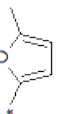

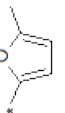

Training set: continued

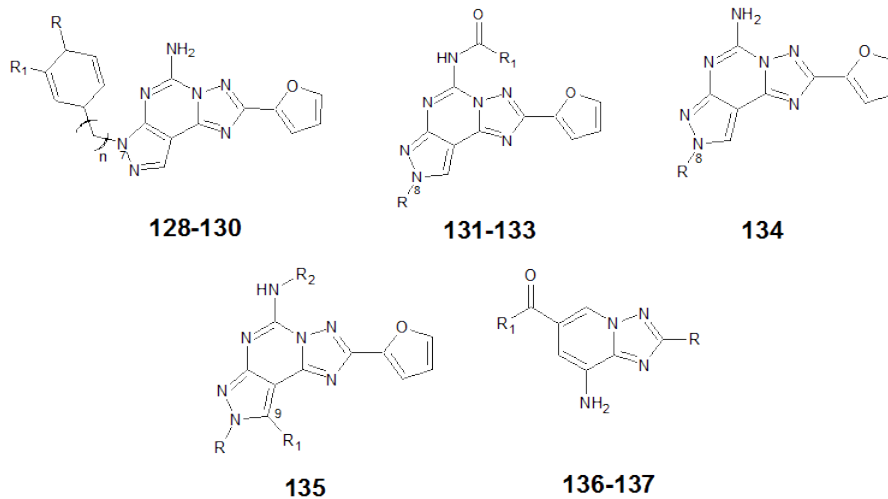
Mol. Id	n	R	R ₁ / Amine*	R ₂	Pred. pK _i (nM) (PLS)	Pred. pK _i (nM) (RSA)	Exp. pK _i (nM) hA _{2A} R	Exp. K _i (nM) hA _{2A} R
106	—			—	-1.42	-1.25	-1.76	57
107	—			—	-1.16	-1.35	-1.88	76
108	—			—	-0.91	-0.67	-1.76	57
109	—			—	-0.55	-0.53	-0.48	3
110	—			—	-1.49	-1.21	-0.95	9
111	—			—	-1.39	-1.07	-0.60	4
112	—			—	-1.30	-1.09	-0.78	6
113	—			—	-0.64	-0.71	-0.90	8

Training set: continued

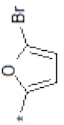

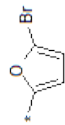

Mol Id	n	R	R ₁ /Amine*	R ₂	Pred. pK _i (nM) (PLS)	Pred. pK _i (nM) (RSA)	Exp. pK _i (nM) hA _{2A} R	Exp. K _i (nM) hA _{2A} R
114	—			—	-0.49	-0.60	-0.30	2
115	—			—	-1.20	-0.88	-1.08	12
116	—			—	-1.20	-1.27	-0.30	2
117	—			—	-1.18	-1.14	-0.48	3
118	—			—	-0.23	-0.27	-0.78	6
119	—			—	-0.81	-0.80	-0.60	4
120	—			—	-1.36	-0.96	-0.95	9

Training set: continued

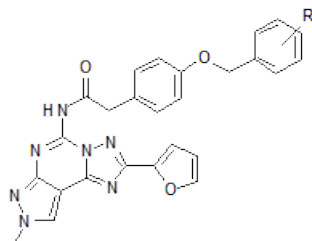
Mol. Id	n	R	R ₁ / Amine*	R ₂	Pred. pK _i (nM) (PLS)	Pred. pK _i (nM) (RSA)	Exp. pK _i (nM) hA _{2A} R	Exp. K _i (nM) hA _{2A} R
121	—			—	-0.66	-0.54	-0.70	5
122	—			—	-1.58	-1.32	-1.76	57
123	—			—	-1.63	-1.70	-1.04	11
124	—			—	-1.43	-1.27	-1.43	27
125	—			—	-1.25	-1.36	-2.02	104
126	—			—	-1.35	-1.62	-1.23	17
127	—			—	-1.17	-1.39	-1.15	14

B.2 Internal test set

128-129: [1]; 130: [3]; 131-133: [4]; 134: [5]; 135: [6]; 136-137: [7]

Mol Id	n	R	R ₁ /Amine*	R ₂	Pred. pK _i (nM) (PLS)	Pred. pK _i (nM) (RSA)	Exp. pK _i (nM) hA _{2A} R	Exp. K _i (nM) hA _{2A} R
128	3	CH ₂ NH ₂	H	—	-0.11	0.25	0.89	0.13
129	3	C(NH)NH ₂	H	—	-0.55	-0.65	-0.64	4.40
130	3	OCH ₃	OCH ₃	—	-0.04	-0.20	-0.43	2.70
131	—	C ₂ H ₅	4-NO ₂ -Ph-NH	—	-2.36	-2.66	-2.79	614
132	—	<i>m</i> -C ₄ H ₉	3,4-Cl-Ph-NH	—	-1.47	-1.65	-2.69	495
133	—	C ₂ H ₅	4-OCH ₃ -Ph-NH	—	-2.44	-2.08	-3.02	1040
134	—	C ₂ H ₅	—	—	-1.37	-1.39	-0.29	1.95
135	—	CH ₃	NCH ₂ CH ₃	H	-0.82	-0.85	-1.00	10
136	—			—	-1.08	-1.14	-1.43	27
137	—			—	-1.13	-1.11	-1.86	73

B.3 External test set

**138-142**

138-142: [8]

Mol Id	R	Predicted pK_i (nM) (PLS)	Predicted pK_i (nM) (RSA)	Experimental pK_i (nM) hA _{2A} R	Experimental K_i (nM) hA _{2A} R
138	H	-1.96	-2.14	-2.37	233
139	2-CH ₃	-1.87	-1.85	-2.52	333
140	2,6-Cl	-1.83	-1.94	-2.45	281
141	3-Cl	-1.67	-1.98	-2.44	278
142	4-CH ₃	-1.82	-1.97	-2.88	751

B.4 References

- [1] Baraldi, P. G.; Tabrizi, M. A.; Bovero, A.; Avitabile, B.; Preti, D.; Fruttarolo, F.; Romagnoli, R.; Varani, K.; Borea, P. A. Recent developments in the field of A_{2A} and A_3 adenosine receptor antagonists. *Eur. J. Med. Chem.* **2003**, *38*, 367-382.
- [2] Baraldi, P. G.; Cacciari, B.; Romagnoli, R.; Spalluto, G.; Monopoli, A.; Ongini, E.; Varani, K.; Borea, P. A. 7-Substituted 5-amino-2-(2-furyl)pyrazolo[4,3-e]-1,2,4-triazolo[1,5-c]pyrimidines as A_{2A} adenosine receptor antagonists: a study on the importance of modifications at the side chain on the activity and solubility. *J. Med. Chem.*, **2002**, *45*, 115-126.
- [3] Baraldi, P. G.; Cacciari, B.; Spalluto, G.; Bergonzoni, M.; Dionisotti, S.; Ongini, E.; Varani, K.; Borea, P. A. Design, synthesis, and biological evaluation of a second generation of pyrazolo[4,3-e]-1,2,4-triazolo[1,5-c]pyrimidines as potent and selective A_{2A} adenosine receptor antagonists. *J. Med. Chem.* **1998**, *41*, 2126-2133.
- [4] Baraldi, P. G.; Cacciari, B.; Moro, S.; Spalluto, G.; Pastorin, G.; Da Ros, T.; Klotz, K. N.; Varani, K.; Gessi, S.; Borea, P. A. Synthesis, biological activity, and molecular modeling investigation of new pyrazolo[4,3-e]-1,2,4-triazolo[1,5-c]pyrimidine derivatives as human A_3 adenosine receptor antagonists. *J. Med. Chem.* **2002**, *45*, 770-780.
- [5] Baraldi, P. G.; Cacciari, B.; Romagnoli, R.; Spalluto, G.; Varani, K.; Gessi, S.; Merighi, S.; Borea, P. A. Pyrazolo[4,3-e]-1,2,4-triazolo[1,5-c]pyrimidine derivatives: a new pharmacological tool for the characterization of the human A_3 adenosine receptor. *Drug Dev. Res.* **2001**, *52*, 406-415.
- [6] Baraldi, P. G.; Fruttarolo, F.; Tabrizi, M. A.; Preti, D.; Romagnoli, R.; El-Kashef, H.; Moorman, A.; Varani, K.; Gessi, S.; Merighi, S.; Borea, P.

- A. Design, synthesis, and biological evaluation of C9- and C2-substituted pyrazolo[4,3-e]-1,2,4-triazolo[1,5-c]pyrimidines as new A_{2A} and A_3 adenosine receptors antagonists. *J. Med. Chem.* **2003**, *46*, 1229-1241.
- [7] Guba, W.; Nettekoven, M.; Pullmann, B.; Riemer, C.; Schmitt, S. Comparison of inhibitory activity of isomeric triazolopyridine derivatives towards adenosine receptor subtypes or do similar structures reveal similar bioactivities? *Biorg. Med. Chem. Lett.* **2004**, *14*, 3307-3312.
- [8] Michielan, L.; Bacilieri, M.; Schiesaro, A.; Bolcato, C.; Pastorin, G.; Spalluto, G.; Cacciari, B.; Klotz, K. N.; Moro, S. Linear and nonlinear 3D-QSAR approaches in tandem with ligand-based homology modeling as computational strategy to depict the pyrazolo-triazolo-pyrimidine antagonists binding site of the human adenosine A_{2A} receptor. *J. Chem. Inf. Model.* **2008**, *48* (2) 350-363.

APPENDIX C

Paper III

Supplementary Information

C.1 Dataset.

C.2 References.

Table C.1 Descriptors selected for the training set in Model 3 in the previous publication to classify CYP450 substrates in three classes.¹ The numbers in the first column refer to the publication by Terfloth *et al.*

Table C.2 Multi-classification model by applying the ct-SVM technique to Data set 1 after the computation of the twelve selected descriptors in Model 3 published in Terfloth *et al.* paper¹: the statistical parameters for each class after prediction on the Validation set 1 (67 substrates) are reported.

Table C.3 Multi-classification model by applying the ct-SVM technique to Data set 1 after the computation of the twelve selected descriptors in Model 3 published in Terfloth *et al.* paper¹: performance measures after prediction on the Validation set 1 (67 substrates) and on the Test set (217 substrates) are summarized.

Table C.4 Multi-label classification ct-SVM/*7classes* model using the same descriptors of Model 3 in ISOCYP paper¹: predicted results of the model for the Test set 1.

Table C.5 ct-SVM/*7classes* model parameters.

¹Terfloth, L.; Bienfait, B.; Gasteiger, J. Ligand-based models for the isoform specificity of cytochrome P450 3A4, 2D6 and 2C9 substrates. *J. Chem. Inf. Model.* 2007, *47*, 1688-1701.

Table C.6 ct-SVM/*5classes* model parameters.

Table C.7 CPG-NN/*5classes* model parameters.

Table C.8 SVM/*5classes* model parameters.

Table C.9 CPG-NN/*5classes* percentage (%) of correct predictions after LOO validation procedure.

C.1 Dataset

Mol Id	Molecule name	CYP1A2	CYP2C19	CYP2C8	CYP2C9	CYP2D6	CYP2E1	CYP3A4	Multilabel	References
1	322	0	0	0	0	0	0	1	No	[1]
2	859	0	0	0	0	1	0	0	No	[1]
3	864	0	0	0	0	0	0	1	No	[1]
4	868	0	0	0	0	0	0	1	No	[1]
5	2689	0	0	0	1	0	0	0	No	[1]
6	3088	0	0	0	1	0	0	0	No	[1]
7	3179	0	0	0	1	0	0	0	No	[1]
8	3326	0	0	0	0	1	0	0	No	[1]
9	3676	0	0	0	0	1	0	0	No	[1]
10	3908	0	0	0	0	0	0	1	No	[1]
11	5018	0	0	0	0	0	0	1	No	[1]
12	5354	0	0	0	0	0	0	1	No	[1]
13	5476	0	0	0	0	1	0	0	No	[1]
14	5491	0	0	0	0	0	0	1	No	[1]
15	5516	0	0	0	0	1	0	0	No	[1]
15	5712	0	0	0	0	1	0	0	No	[1]
17	5726	0	0	0	0	1	0	0	No	[1]
18	6095	0	0	0	0	1	0	0	No	[1]
19	6228	1	0	0	0	0	0	0	No	[1]
20	6389	0	0	0	0	1	0	0	No	[1]
21	6609	0	0	0	0	1	0	0	No	[1]
22	6610	0	0	0	0	1	0	0	No	[1]

Dataset: continued

Mol Id	Molecule name	CYP1A2	CYP2C19	CYP2C8	CYP2C9	CYP2D6	CYP2E1	CYP3A4	Multilabel	References
23	6634	0	0	0	0	1	0	0	No	[1]
24	6635	0	0	0	0	1	0	0	No	[1]
25	8606	0	0	0	0	1	0	0	No	[1]
26	8616	0	0	0	0	1	0	0	No	[1]
27	8641	0	1	0	0	0	0	0	No	[1]
28	9093	0	0	0	0	0	0	1	No	[1]
29	9348	0	0	0	0	1	0	0	No	[1]
30	9862	0	0	0	0	1	0	0	No	[1]
31	9948	0	0	1	0	0	0	0	No	[1]
32	10011	1	0	0	0	0	0	0	No	[1]
33	10989	0	0	0	0	0	0	1	No	[1]
34	11120	0	0	0	1	0	0	0	No	[1]
35	11124	0	0	0	1	0	0	0	No	[1]
36	11295	0	0	0	1	0	0	0	No	[1]
37	11300	0	0	0	0	0	0	1	No	[1]
38	11441	0	0	0	0	1	0	0	No	[1]
39	11443	0	1	0	0	0	0	0	No	[1]
40	11447	0	1	0	0	0	0	0	No	[1]
41	11449	0	0	0	0	1	0	0	No	[1]
42	11750	1	0	0	0	0	0	0	No	[1]
43	12345	0	0	0	0	1	0	0	No	[1]
44	12451	0	0	0	0	1	0	0	No	[1]

Dataset: continued

Mol Id	Molecule name	CYP1A2	CYP2C19	CYP2C8	CYP2C9	CYP2D6	CYP2E1	CYP3A4	Multilabel	References
45	13212	0	0	0	0	1	0	0	No	[H]
46	13414	0	0	0	0	1	0	0	No	[H]
47	13499	0	0	0	0	1	0	0	No	[H]
48	14334	0	0	0	1	0	0	0	No	[H]
49	14343	0	0	0	1	0	0	0	No	[H]
50	14467	0	1	0	0	0	0	0	No	[H]
51	14628	0	0	0	0	1	0	0	No	[H]
52	15061	0	0	0	0	1	0	0	No	[H]
53	15508	0	0	0	0	0	1	0	No	[H]
54	15852	0	0	0	0	0	1	0	No	[H]
55	16165	0	0	0	0	0	1	0	No	[H]
56	16396	0	0	0	0	0	1	0	No	[H]
57	16439	1	0	0	0	0	0	0	No	[H]
58	16457	0	0	0	0	0	1	0	No	[H]
59	16460	0	0	0	0	0	1	0	No	[H]
60	16462	0	0	0	0	0	1	0	No	[H]
61	16608	0	0	0	0	0	1	0	No	[H]
62	16611	0	0	0	0	0	1	0	No	[H]
63	16781	1	0	0	0	0	0	0	No	[H]
64	17178	0	0	0	0	0	0	1	No	[H]
65	17304	0	0	0	0	1	0	0	No	[H]
66	17308	0	0	0	0	1	0	0	No	[H]

Dataset: continued

Mol Id	Molecule name	CYP1A2	CYP2C19	CYP2C8	CYP2C9	CYP2D6	CYP2E1	CYP3A4	Multilabel	References
67	18423	0	0	0	0	1	0	0	No	[1]
68	19636	0	0	0	0	0	0	1	No	[1]
69	19843	1	0	0	0	0	0	0	No	[1]
70	19860	1	0	0	0	0	0	0	No	[1]
71	20159	0	0	0	0	1	0	0	No	[1]
72	20440	0	0	0	0	1	0	0	No	[1]
73	21115	0	0	1	0	0	0	0	No	[1]
74	21125	0	0	0	1	0	0	0	No	[1]
75	21138	0	0	0	1	0	0	0	No	[1]
76	21342	0	0	0	0	0	0	1	No	[1]
77	21344	0	0	0	0	0	0	1	No	[1]
78	21491	0	0	0	0	1	0	0	No	[1]
79	22459	0	0	0	0	0	1	0	No	[1]
80	22636	1	0	0	0	0	0	0	No	[1]
81	23276	0	0	0	0	1	0	0	No	[1]
82	23281	0	0	0	0	1	0	0	No	[1]
83	23669	1	0	0	0	0	0	0	No	[1]
84	23677	1	0	0	0	0	0	0	No	[1]
85	24195	0	0	0	0	0	1	0	No	[1]
86	25165	0	0	0	0	1	0	0	No	[1]
87	25307	0	0	0	0	0	1	0	No	[1]
88	25318	0	0	0	0	0	0	1	No	[1]

Dataset: continued

Mol Id	Molecule name	CYP1A2	CYP2C19	CYP2C8	CYP2C9	CYP2D6	CYP2E1	CYP3A4	Multilabel	References
89	25319	0	0	0	0	0	0	1	No	[H]
90	25414	0	0	0	0	0	0	1	No	[H]
91	25647	0	0	0	0	0	1	0	No	[H]
92	26299	1	0	0	0	0	0	0	No	[H]
93	27253	0	0	0	0	0	0	1	No	[H]
94	27906	0	0	0	0	0	0	1	No	[H]
95	27907	0	0	0	0	0	0	1	No	[H]
96	27909	0	0	0	0	0	0	1	No	[H]
97	29127	1	0	0	0	0	0	0	No	[H]
98	29870	0	0	0	0	1	0	0	No	[H]
99	29889	0	0	0	0	1	0	0	No	[H]
100	30029	0	0	1	0	0	0	0	No	[H]
101	30031	0	0	0	0	0	0	1	No	[H]
102	30034	0	0	0	0	0	0	1	No	[H]
103	30085	0	0	0	0	1	0	0	No	[H]
104	30095	0	0	0	0	0	0	1	No	[H]
105	30529	0	0	0	0	0	0	1	No	[H]
106	30531	0	0	0	0	0	0	1	No	[H]
107	31064	0	0	0	0	0	1	0	No	[H]
108	31537	1	0	0	0	0	0	0	No	[H]
109	31543	1	0	0	0	0	0	0	No	[H]
110	31893	0	0	0	0	0	0	1	No	[H]

Dataset: continued

Mol Id	Molecule name	CYP1A2	CYP2C19	CYP2C8	CYP2C9	CYP2D6	CYP2E1	CYP3A4	Multilabel	References
111	34120	0	0	0	0	0	0	1	No	[1]
112	34454	1	0	0	0	0	0	0	No	[1]
113	34455	1	0	0	0	0	0	0	No	[1]
114	34457	1	0	0	0	0	0	0	No	[1]
115	34458	1	0	0	0	0	0	0	No	[1]
116	34475	0	0	0	0	0	0	1	No	[1]
117	34579	0	0	0	0	0	1	0	No	[1]
118	34690	0	0	0	0	0	1	0	No	[1]
119	35597	0	0	0	0	0	1	0	No	[1]
120	35600	0	0	0	0	0	1	0	No	[1]
121	36005	0	0	0	0	0	0	1	No	[1]
122	36731	0	0	0	0	0	1	0	No	[1]
123	37213	0	0	0	0	0	0	1	No	[1]
124	37474	0	0	0	0	0	0	1	No	[1]
125	37599	0	0	0	0	0	1	0	No	[1]
126	38395	1	0	0	0	0	0	0	No	[1]
127	38398	1	0	0	0	0	0	0	No	[1]
128	38402	0	1	0	0	0	0	0	No	[1]
129	38404	0	1	0	0	0	0	0	No	[1]
130	38407	0	1	0	0	0	0	0	No	[1]
131	38416	1	0	0	0	0	0	0	No	[1]
132	38528	1	0	0	0	0	0	0	No	[1]

Dataset: continued

Mol Id	Molecule name	CYP1A2	CYP2C19	CYP2C8	CYP2C9	CYP2D6	CYP2E1	CYP3A4	Multilabel	References
133	38680	1	0	0	0	0	0	0	No	[1]
134	38744	0	0	0	0	1	0	0	No	[1]
135	39279	0	0	0	0	0	0	1	No	[1]
136	40334	0	0	0	0	0	0	1	No	[1]
137	40472	0	0	0	0	0	0	1	No	[1]
138	40931	0	1	0	0	0	0	0	No	[1]
139	40994	0	0	0	1	0	0	0	No	[1]
140	41354	0	0	0	0	0	0	1	No	[1]
141	41405	0	0	0	0	0	0	1	No	[1]
142	43025	0	0	0	0	0	0	1	No	[1]
143	43253	0	0	0	0	1	0	0	No	[1]
144	43851	0	0	0	0	0	0	1	No	[1]
145	44100	0	0	0	0	1	0	0	No	[1]
146	44150	1	0	0	0	0	0	0	No	[1]
148	44972	0	0	0	0	0	1	0	No	[1]
149	44978	0	0	0	0	1	0	0	No	[1]
150	45411	0	0	0	0	0	1	0	No	[1]
151	46098	1	0	0	0	0	0	0	No	[1]
152	46231	0	0	0	0	0	0	1	No	[1]
153	46232	0	0	0	0	0	0	1	No	[1]
154	46346	1	0	0	0	0	0	0	No	[1]
155	46350	1	0	0	0	0	0	0	No	[1]

Dataset: continued

Mol Id	Molecule name	CYP1A2	CYP2C19	CYP2C8	CYP2C9	CYP2D6	CYP2E1	CYP3A4	Multilabel	References
156	47019	0	0	0	0	1	0	0	No	[1]
157	47020	1	0	0	0	0	0	0	No	[1]
158	47041	0	0	0	0	1	0	0	No	[1]
159	47755	0	0	0	0	0	0	1	No	[1]
160	48285	0	0	0	0	0	0	1	No	[1]
161	48336	0	0	0	0	1	0	0	No	[1]
162	48337	0	0	0	0	1	0	0	No	[1]
163	48566	1	0	0	0	0	0	0	No	[1]
164	48657	0	0	0	0	0	0	1	No	[1]
166	48895	0	0	0	0	0	1	0	No	[1]
167	49260	0	0	0	0	0	1	0	No	[1]
168	50413	0	0	0	1	0	0	0	No	[1]
169	50588	0	0	0	1	0	0	0	No	[1]
170	50826	0	0	0	0	0	0	1	No	[1]
171	51201	0	0	0	0	0	0	1	No	[1]
172	51488	0	0	0	0	0	0	1	No	[1]
173	1727	0	0	0	0	0	0	1	No	[1]
174	51728	0	0	0	0	0	0	1	No	[1]
175	51830	0	0	0	0	0	0	1	No	[1]
176	51831	0	0	0	0	0	0	1	No	[1]
177	52799	0	0	0	0	0	1	0	No	[1]
178	52801	0	0	0	0	0	1	0	No	[1]

Dataset: continued

Mol Id	Molecule name	CYP1A2	CYP2C19	CYP2C8	CYP2C9	CYP2D6	CYP2E1	CYP3A4	Multilabel	References
179	53365	0	0	0	0	0	0	1	No	[1]
180	53367	0	0	0	1	0	0	0	No	[1]
181	55175	0	0	0	0	0	0	1	No	[1]
182	55955	0	0	0	1	0	0	0	No	[1]
183	57002	0	0	0	0	0	0	1	No	[1]
184	57004	0	0	0	0	0	0	1	No	[1]
185	57197	0	0	0	0	0	1	0	No	[1]
186	58162	0	0	0	0	0	0	1	No	[1]
187	58164	0	0	0	0	0	0	1	No	[1]
188	58173	0	0	0	0	0	0	1	No	[1]
189	58176	0	0	0	0	0	0	1	No	[1]
190	59883	0	0	0	0	0	1	0	No	[1]
191	59887	0	0	0	0	0	1	0	No	[1]
192	59889	0	0	0	0	0	1	0	No	[1]
193	60153	0	0	0	0	0	0	1	No	[1]
194	60586	0	0	0	0	0	0	1	No	[1]
195	60616	0	0	0	0	1	0	0	No	[1]
196	60821	0	0	0	0	1	0	0	No	[1]
197	60822	0	0	0	0	1	0	0	No	[1]
198	60823	0	0	0	0	1	0	0	No	[1]
199	60824	0	0	0	1	0	0	0	No	[1]
200	61327	0	0	0	0	1	0	0	No	[1]

Dataset: continued

Mol Id	Molecule name	CYP1A2	CYP2C19	CYP2C8	CYP2C9	CYP2D6	CYP2E1	CYP3A4	Multilabel	References
201	61469	1	0	0	0	0	0	0	No	[1]
202	61470	1	0	0	0	0	0	0	No	[1]
203	61472	1	0	0	0	0	0	0	No	[1]
204	61650	1	0	0	0	0	0	0	No	[1]
205	61815	0	0	0	0	1	0	0	No	[1]
206	61816	0	0	0	0	1	0	0	No	[1]
207	62027	0	0	0	0	0	0	1	No	[1]
208	62126	0	0	0	0	0	0	1	No	[1]
209	62636	0	0	0	0	1	0	0	No	[1]
210	62744	0	0	0	0	1	0	0	No	[1]
211	63097	0	0	0	0	0	1	0	No	[1]
212	63098	0	0	0	0	0	1	0	No	[1]
213	63099	0	0	0	0	0	1	0	No	[1]
214	63100	0	0	0	0	0	1	0	No	[1]
215	63101	0	0	0	0	0	1	0	No	[1]
216	63102	0	0	0	0	0	1	0	No	[1]
217	63104	0	0	0	0	0	1	0	No	[1]
218	63115	0	0	0	0	0	0	1	No	[1]
219	63353	0	0	0	0	0	0	1	No	[1]
220	63824	0	0	0	0	0	0	1	No	[1]
221	63829	0	0	0	0	0	0	1	No	[1]
222	64618	0	0	0	0	0	0	1	No	[1]

Dataset: continued

Mol Id	Molecule name	CYP1A2	CYP2C19	CYP2C8	CYP2C9	CYP2D6	CYP2E1	CYP3A4	Multilabel	References
223	64627	1	0	0	0	0	0	0	No	[1]
224	64819	0	0	0	0	0	0	1	No	[1]
225	64859	0	0	0	0	0	1	0	No	[1]
226	64929	0	0	0	0	0	0	1	No	[1]
227	65245	0	0	0	0	0	0	1	No	[1]
228	66235	0	0	0	0	0	0	1	No	[1]
229	66260	0	0	0	0	0	0	1	No	[1]
230	66337	0	0	0	0	0	0	1	No	[1]
231	66427	0	0	0	0	0	1	0	No	[1]
232	66478	0	0	0	0	0	0	1	No	[1]
233	67107	0	0	0	0	0	0	1	No	[1]
234	67111	0	0	0	0	0	0	1	No	[1]
235	67517	0	0	0	0	0	0	1	No	[1]
236	67596	0	0	0	0	0	0	1	No	[1]
237	68665	0	0	0	0	0	0	1	No	[1]
238	68790	0	0	1	0	0	0	0	No	[1]
239	68791	0	0	1	0	0	0	0	No	[1]
240	69572	0	0	0	1	0	0	0	No	[1]
241	69573	0	0	0	1	0	0	0	No	[1]
242	69575	0	0	0	1	0	0	0	No	[1]
243	69578	0	0	0	1	0	0	0	No	[1]

Dataset: continued

MoI Id	Molecule name	CYP1A2	CYP2C19	CYP2C8	CYP2C9	CYP2D6	CYP2E1	CYP3A4	Multilabel	References
244	69579	0	0	0	1	0	0	0	No	[1]
245	69585	0	0	0	1	0	0	0	No	[1]
246	69734	0	0	0	0	0	0	1	No	[1]
247	69807	0	0	1	0	0	0	0	No	[1]
248	70401	0	0	0	0	1	0	0	No	[1]
249	70678	0	0	0	0	0	0	1	No	[1]
250	70714	0	0	0	0	1	0	0	No	[1]
251	71062	0	0	0	0	1	0	0	No	[1]
252	71072	0	0	0	0	1	0	0	No	[1]
253	71098	1	0	0	0	0	0	0	No	[1]
254	71099	1	0	0	0	0	0	0	No	[1]
255	71410	0	0	0	0	0	0	1	No	[1]
256	72142	0	0	0	0	0	0	1	No	[1]
257	72588	0	0	0	1	0	0	0	No	[1]
258	72922	0	0	0	0	1	0	0	No	[1]
259	73006	1	0	0	0	0	0	0	No	[1]
260	73066	0	0	0	0	0	0	1	No	[1]
261	73076	0	0	0	0	1	0	0	No	[1]
262	73184	0	0	0	0	0	0	1	No	[1]
263	73186	0	0	0	0	0	0	1	No	[1]
264	73201	0	0	0	0	0	0	1	No	[1]
265	73211	0	0	0	0	0	0	1	No	[1]

Dataset: continued

Mol Id	Molecule name	CYP1A2	CYP2C19	CYP2C8	CYP2C9	CYP2D6	CYP2E1	CYP3A4	Multilabel	References
266	73212	0	0	0	0	0	0	1	No	[1]
267	73334	0	0	0	0	0	0	1	No	[1]
268	Acetofenac	0	0	0	1	0	0	0	No	[2]
269	Acenocoumarol	0	0	0	1	0	0	1	Yes	[2]
270	Acetaminophen	1	0	0	0	0	0	0	No	[3]
271	Alfentanil	0	0	0	0	0	0	1	No	[3]
272	Alfuzosin	0	0	0	0	0	0	1	No	[3]
273	Almotriptan	0	0	0	0	0	0	1	No	[3]
274	Alpidem	0	0	0	0	0	0	1	No	[4]
275	Alprazolam	0	0	0	0	0	0	1	No	[3]
276	Alprenolol	0	0	0	0	1	0	0	No	[5]
277	Amiflamine	0	0	0	0	1	0	0	No	[4]
278	Amiodarone	0	0	1	0	0	0	1	Yes	[3]
279	Amitriptyline	1	1	0	1	1	0	1	Yes	[3]
280	Amlodipine	0	0	0	0	0	0	1	No	[3]
281	Amphetamine	0	0	0	0	1	0	0	No	[3]
282	Amprenavir	0	0	0	0	0	0	1	No	[3]
283	Aprepitant	0	0	0	0	0	0	1	No	[3]
284	Aripiprazole	0	0	0	0	1	0	1	Yes	[5]
285	Astemizole	0	0	0	0	0	0	1	No	[3]
286	Atazanavir	0	0	0	0	0	0	1	No	[3]
287	Atomoxetine	0	0	0	0	1	0	0	No	[3]

Dataset: continued

Mol Id	Molecule name	CYP1A2	CYP2C19	CYP2C8	CYP2C9	CYP2D6	CYP2E1	CYP3A4	Multilabel	References
288	Atorvastatin	0	0	0	0	0	0	1	No	[3]
289	Azatadin	0	0	0	0	0	0	1	No	[4]
290	Beclomethasone	0	0	0	0	0	0	1	No	[3]
291	Benzphetamine	0	0	1	0	0	0	0	No	[3]
292	Bepiridil	0	0	0	0	0	0	1	No	[3]
293	Bexarotene	0	0	0	0	0	0	1	No	[3]
294	Bisoprolol	0	0	0	0	1	0	0	No	[3]
295	Bromocriptine	0	0	0	0	0	0	1	No	[3]
296	Budesonide	0	0	0	0	0	0	1	No	[3]
297	Bufuralol	0	0	0	0	1	0	0	No	[4]
298	Buspirone	0	0	0	0	0	0	1	No	[3]
299	Busulfan	0	0	0	0	0	0	1	No	[3]
300	Caffeine	1	0	0	0	0	0	1	Yes	[3]
301	Carbamazepine	0	0	1	0	0	0	1	Yes	[3]
302	Carisoprodol	0	1	0	0	0	0	0	No	[3]
303	Carmustine	0	0	0	0	0	0	1	No	[2]
304	Carvedilol	0	0	0	1	1	0	0	Yes	[3]
305	Celecoxib	0	0	0	1	0	0	0	No	[3]
306	Cerivastatin	0	0	1	0	0	0	1	Yes	[5]
307	Cevimeline	0	0	0	0	1	0	1	Yes	[3]
308	Chlordiazepoxide	1	0	0	0	0	0	0	No	[3]
309	Chlorpheniramine	0	0	0	0	0	0	1	No	[3]

Dataset: continued

Mol Id	Molecule name	CYP1A2	CYP2C19	CYP2C8	CYP2C9	CYP2D6	CYP2E1	CYP3A4	Multilabel	References
310	Chlorpromazine	0	0	0	0	1	0	0	No	[3]
311	Chlorpropramide	0	0	0	0	1	0	0	No	[3]
312	Chlorzoxazone	0	0	0	0	0	1	0	No	[2]
313	Cilostazol	0	1	0	0	0	0	1	Yes	[3]
314	Cinacalcet	1	0	0	0	1	0	1	Yes	[3]
315	Cinnarizine	0	0	0	0	1	0	0	No	[4]
316	Cisapride	0	0	0	0	0	0	1	No	[3]
317	Citalopram	0	1	0	0	0	0	1	Yes	[3]
318	Clarithromycin	0	0	0	0	0	0	1	No	[3]
319	Clindamycin	0	0	0	0	0	0	1	No	[3]
320	Clomipramine	1	1	0	1	1	0	1	Yes	[3]
321	Clonazepam	0	0	0	0	0	0	1	No	[3]
322	Clopidogrel	1	0	0	0	0	0	1	Yes	[3]
323	Clozapine	1	0	0	0	1	0	1	Yes	[3]
324	Cocaine	0	0	0	0	0	0	1	No	[3]
325	Codeine	0	0	0	0	1	0	0	No	[3]
326	Colchicine	0	0	0	0	0	0	1	No	[3]
327	Cortisol	0	0	0	0	0	0	1	No	[3]
328	Cyclobenzaprine	1	0	0	0	1	0	1	Yes	[3]
329	Cyclophosphamide	0	1	0	0	0	0	1	Yes	[3]
330	Cyclosporin	0	0	0	0	0	0	1	No	[2]
331	Dapsone	0	0	0	0	0	1	1	Yes	[3]

Dataset: continued

Mol Id	Molecule name	CYP1A2	CYP2C19	CYP2C8	CYP2C9	CYP2D6	CYP2E1	CYP3A4	Multilabel	References
332	Darifenacin	0	0	0	0	1	0	1	Yes	[3]
333	Debrisoquine	0	0	0	0	1	0	0	No	[5]
334	Delavirdine	0	0	0	0	0	0	1	No	[3]
335	Deprenyl	0	0	0	0	1	0	0	No	[4]
336	Desipramine	1	1	0	0	1	0	0	Yes	[3]
337	Desogestrel	0	0	0	1	0	0	1	Yes	[3]
338	Dexamethasone	0	0	0	0	0	0	1	No	3]
339	Dexfenfluramine	0	0	0	0	1	0	0	No	[3]
340	Dextromethorphan	0	0	0	0	1	0	1	Yes	[3]
341	Diazepam	1	1	0	1	0	0	1	Yes	[3]
342	Diclofenac	0	0	0	1	0	0	0	No	[3]
343	Dihydrocodeine	0	0	0	0	1	0	1	Yes	[2]
344	Dihydroergotamine	0	0	0	0	0	0	1	No	[3]
345	Diltiazem	0	0	0	0	0	0	1	No	[3]
346	Disopyramide	0	0	0	0	0	0	1	No	[3]
347	Docetaxel	0	0	1	0	0	0	1	Yes	[3]
348	Dofetilide	0	0	0	0	0	0	1	No	[3]
349	Dolasetron	0	0	0	0	1	0	1	Yes	[3]
350	Domperidone	0	0	0	0	0	0	1	No	[5]
351	Donepezil	0	0	0	0	1	0	1	Yes	[3]
352	Doxepin	0	0	0	0	1	0	0	No	[3]
353	Doxorubicin	0	0	0	0	0	0	1	No	[3]

Dataset: continued

Mol Id	Molecule name	CYP1A2	CYP2C19	CYP2C8	CYP2C9	CYP2D6	CYP2E1	CYP3A4	Multilabel	References
354	Dronabinol	0	0	0	1	0	0	1	Yes	[3]
355	Duloxetine	1	0	0	0	1	0	0	Yes	[3]
356	Dutasteride	0	0	0	0	0	0	1	No	[3]
357	Ebastine	0	0	0	0	0	0	1	No	[4]
358	Efavirenz	0	0	0	0	0	0	1	No	[3]
359	Enalapril	0	0	0	0	0	0	1	No	[4]
360	Encainide	0	0	0	0	1	0	0	No	[3]
361	Enflurane	0	0	0	0	0	1	0	No	[2]
362	Eplerenone	0	0	0	0	0	0	1	No	[3]
363	Ergotamine	0	0	0	0	0	0	1	No	[3]
364	Erlotinib	1	0	0	0	0	0	1	Yes	[3]
365	Erythromycin	0	0	0	0	0	0	1	No	[3]
366	Estradiol	1	0	0	0	0	0	0	No	[3]
367	Eszopiclone	0	0	0	0	0	1	1	Yes	[3]
368	Ethanol	0	0	0	0	0	1	0	No	[2]
369	Ethinylestradiol	0	0	0	0	0	0	1	No	[3]
370	Ethosuximide	0	0	0	0	0	0	1	No	[3]
371	Ethylmorphine	0	0	0	0	1	0	1	Yes	[2]
372	Etonogestrel	0	0	0	0	0	0	1	No	[3]
373	Etoposide	0	0	0	0	0	0	1	No	[3]
374	Exemestane	0	0	0	0	0	0	1	No	[3]
375	Felodipine	0	0	0	0	0	0	1	No	[3]

Dataset: continued

Mol Id	Molecule name	CYP1A2	CYP2C19	CYP2C8	CYP2C9	CYP2D6	CYP2E1	CYP3A4	Multilabel	References
376	Fenfuramine	0	0	0	0	1	0	0	No	[3]
377	Fentanyl	0	0	0	0	1	0	1	Yes	[3]
378	Fexofenadine	0	0	0	0	0	0	1	No	[3]
379	Finasteride	0	0	0	0	0	0	1	No	[3]
380	Flecainide	0	0	0	0	1	0	0	No	[3]
381	Fluconazole	0	0	0	0	0	0	1	No	[4]
382	Flunarizine	0	0	0	0	1	0	0	No	[4]
383	Flunitrazepam	0	1	0	0	0	0	1	Yes	[2]
384	Fluoxetine	0	0	0	1	1	0	0	Yes	[3]
385	Fluphenazine	0	0	0	0	1	0	0	No	[3]
386	Flurbiprofen	0	0	0	1	0	0	0	No	[3]
387	Flutamide	1	0	0	0	0	0	1	Yes	[3]
388	Fluticasone	0	0	0	0	0	0	1	No	[3]
389	Fluvastatin	0	0	1	1	0	0	0	Yes	[3]
390	Fluvoxamine	1	0	0	0	1	0	0	Yes	[3]
391	Formoterol	0	1	0	1	1	0	0	Yes	[3]
392	Fulvestrant	0	0	0	0	0	0	1	No	[3]
393	Galantamine	0	0	0	0	1	0	1	Yes	[3]
394	Glimepiride	0	0	0	1	0	0	0	No	[3]
395	Glipizide	0	0	0	1	0	0	0	No	[3]
396	Glyburide	0	0	0	1	0	0	0	No	[3]
397	Granisetron	0	0	0	0	0	0	1	No	[2]

Dataset: continued

Mol Id	Molecule name	CYP1A2	CYP2C19	CYP2C8	CYP2C9	CYP2D6	CYP2E1	CYP3A4	Multilabel	References
398	Haloperidol	1	0	0	0	1	0	1	Yes	[3]
399	Halothane	0	0	0	0	0	1	0	No	[2]
400	Hexobarbital	0	1	0	0	0	1	0	Yes	[3]
401	Hydrocodone	0	0	0	0	1	0	1	Yes	[3]
402	Ibuprofen	0	0	0	1	0	0	0	No	[3]
403	Ifosfamide	0	0	0	0	0	0	1	No	[3]
404	Imatinib	0	0	0	0	0	0	1	No	[3]
405	Imipramine	1	1	0	1	1	0	1	Yes	[3]
406	Indinavir	0	0	0	0	0	0	1	No	[3]
407	Indomethacin	0	1	0	1	0	0	0	Yes	[3]
408	Irbesartan	0	0	0	1	0	0	0	No	[3]
409	Irinotecan	0	0	0	0	0	0	1	No	[5]
410	Isoflurane	0	0	0	0	0	1	0	No	[2]
411	Isotretinoin	0	0	1	0	0	0	0	No	[3]
412	Isradipine	0	0	0	0	0	0	1	No	[3]
413	Itraconazole	0	0	0	0	0	0	1	No	[3]
414	Ketoconazole	0	0	0	0	0	0	1	No	[3]
415	Lansoprazole	0	1	0	0	0	0	1	Yes	[3]
416	Lercanidipine	0	0	0	0	0	0	1	No	[5]
417	Letrozole	0	0	0	0	0	0	1	No	[3]
418	Levo bupivacaine	1	0	0	0	0	0	1	Yes	[3]
419	Levonorgestrel	0	0	0	0	0	0	1	No	[4]

Dataset: continued

Mol Id	Molecule name	CYP1A2	CYP2C19	CYP2C8	CYP2C9	CYP2D6	CYP2E1	CYP3A4	Multilabel	References
420	Lidocaine	0	0	0	0	1	0	1	Yes	[3]
421	Lisuride	0	0	0	0	0	0	1	No	[4]
422	Lobeline	0	0	0	0	1	0	0	No	[4]
423	Lopinavir	0	0	0	0	0	0	1	No	[3]
424	Loratadine	0	0	0	0	0	0	1	No	[3]
425	Lornoxicam	0	0	0	1	0	0	0	No	[4]
426	Losartan	0	0	0	1	0	0	1	Yes	[3]
427	Lovastatin	0	0	0	0	0	0	1	No	[3]
428	Maprotiline	0	0	0	0	1	0	0	No	[3]
429	Maraviroc	0	0	0	0	0	0	1	No	[3]
430	Medroxyprogesterone	0	0	0	0	0	0	1	No	[3]
431	Mefenamic acid	0	0	0	1	0	0	0	No	[3]
432	Meloxicam	0	0	0	1	0	0	0	No	[3]
433	Meperidine	0	0	0	0	1	0	0	No	[3]
434	Mephobarbital	0	1	0	0	0	0	0	No	[3]
435	Methadone	1	0	0	0	1	0	1	Yes	[3]
436	Methamphetamine	0	0	0	0	1	0	0	No	[3]
437	Methoxyamphetamine	0	0	0	0	1	0	0	No	[3]
438	Methoxyflurane	0	0	0	0	0	1	0	No	[2]
439	Methoxyphenamine	0	0	0	0	1	0	0	No	[4]
440	Methylprednisolone	0	0	0	0	0	0	1	No	[3]
441	Metoclopramide	0	0	0	0	1	0	0	No	[5]

Dataset: continued

Mol Id	Molecule name	CYP1A2	CYP2C19	CYP2C8	CYP2C9	CYP2D6	CYP2E1	CYP3A4	Multilabel	References
442	Metoprolol	0	0	0	0	1	0	0	No	[3]
443	Mexiletine	1	0	0	0	1	0	0	Yes	[3]
444	Mianserin	1	0	0	0	1	0	1	Yes	[2]
445	Miconazole	0	0	0	0	0	0	1	No	[3]
446	Midazolam	0	0	0	0	0	0	1	No	[3]
447	Mifepristone	0	0	0	0	0	0	1	No	[3]
448	Minaprine	0	0	0	0	1	0	0	No	[5]
449	Mirtazapine	1	0	0	0	1	0	1	Yes	[3]
450	Moclobemide	0	1	0	0	0	0	0	No	[3]
451	Modafinil	0	0	0	0	0	0	1	No	[3]
452	Mometasone	0	0	0	0	0	0	1	No	[3]
453	Montelukast	0	0	0	1	0	0	1	Yes	[3]
454	Morphine	0	0	0	0	1	0	0	No	[3]
455	Naproxen	1	0	0	1	0	0	0	Yes	[3]
456	Nateglinide	0	0	0	1	0	0	1	Yes	[3]
457	Nefazodone	0	0	0	0	0	0	1	No	[3]
458	Nelfinavir	0	1	0	0	0	0	1	Yes	[3]
459	Nevirapine	0	0	0	0	0	0	1	No	[3]
460	Nicardipine	0	0	0	0	0	0	1	No	[3]
461	Nifedipine	0	0	0	0	0	0	1	No	[3]
462	Nilutamide	0	1	0	0	0	0	0	No	[3]
463	Nimodipine	0	0	0	0	0	0	1	No	[3]

Dataset: continued

Mol Id	Molecule name	CYP1A2	CYP2C19	CYP2C8	CYP2C9	CYP2D6	CYP2E1	CYP3A4	Multilabel	References
464	Nisoldipine	0	0	0	0	0	0	1	No	[3]
465	Nitrendipine	0	0	0	0	0	0	1	No	[3]
466	Norethisterone	0	0	0	0	0	0	1	No	[3]
467	Nortriptyline	1	0	0	0	1	0	0	Yes	[3]
468	Olanzapine	1	0	0	0	1	0	0	Yes	[3]
469	Omeprazole	0	1	0	1	0	0	1	Yes	[3]
470	Ondansetron	1	0	0	0	1	0	1	Yes	[3]
471	Oxybutynin	0	0	0	0	0	0	1	No	[3]
472	Oxycodone	0	0	0	0	1	0	0	No	[3]
473	Paclitaxel	0	0	1	0	0	0	1	Yes	[3]
474	Pantoprazole	0	1	0	0	0	0	1	Yes	[3]
475	Paroxetine	0	0	0	0	1	0	0	No	[3]
476	Pentamidine	0	1	0	0	0	0	0	No	[3]
477	Perhexiline	0	0	0	0	1	0	0	No	[5]
478	Perphenazine	0	0	0	0	1	0	0	No	[3]
479	Phenformin	0	0	0	0	1	0	0	No	[5]
480	Phenylbutazone	0	0	0	1	0	0	0	No	[4]
481	Phenytolol	0	1	1	1	0	0	0	Yes	[3]
482	Pimozide	0	0	0	0	0	0	1	No	[3]
483	Pindolol	0	0	0	0	1	0	0	No	[3]
484	Piroxicam	0	0	0	1	0	0	0	No	[3]

Dataset: continued

Mol Id	Molecule name	CYP1A2	CYP2C19	CYP2C8	CYP2C9	CYP2D6	CYP2E1	CYP3A4	Multilabel	References
486	Prednisolone	0	0	0	0	0	0	1	No	[3]
487	Prednisone	0	0	0	0	0	0	1	No	[3]
488	Progesterone	0	1	0	0	0	0	1	Yes	[3]
489	Proguanil	0	1	0	0	0	0	0	No	[3]
490	Promethazine	0	0	0	0	1	0	0	No	[3]
491	Propafenone	1	0	0	0	1	0	0	Yes	[3]
492	Propoxyphene	0	0	0	0	1	0	0	No	[3]
493	Propranolol	1	1	0	0	1	0	0	Yes	[3]
494	Quercetin	0	0	0	0	0	0	1	No	[4]
495	Quetiapine	0	0	0	0	1	0	1	Yes	[3]
496	Quinidine	0	0	0	0	0	0	1	No	[3]
497	Quinine	0	0	0	0	0	0	1	No	[3]
498	Rabeprazole	0	1	0	0	0	0	1	Yes	[3]
499	Ramelteon	1	0	0	1	0	0	1	Yes	[3]
500	Ranolazine	0	0	0	0	1	0	1	Yes	[3]
501	Remoxipride	0	0	0	0	1	0	0	No	[4]
502	Repaglinide	0	0	1	0	0	0	1	Yes	[3]
503	Retinoic acid	0	0	1	0	0	0	0	No	[3]
504	Retinol	0	0	1	0	0	0	0	No	[3]
505	Rifabutin	0	0	0	0	0	0	1	No	[3]
506	Rifampicin	0	0	0	0	0	0	1	No	[3]
507	Riluzole	1	0	0	0	0	0	0	No	[3]

Dataset: continued

Mol Id	Molecule name	CYP1A2	CYP2C19	CYP2C8	CYP2C9	CYP2D6	CYP2E1	CYP3A4	Multilabel	References
508	Risperidone	0	0	0	0	1	0	0	No	[3]
509	Ritonavir	0	0	0	0	1	0	1	Yes	[3]
510	Ropinirole	1	0	0	0	0	0	0	No	[3]
511	Ropivacaine	1	0	0	0	0	0	0	No	[3]
512	Rosiglitazone	0	0	1	1	0	0	0	Yes	[3]
513	Salmeterol	0	0	0	0	0	0	1	No	[3]
514	Saquinavir	0	0	0	0	0	0	1	No	[3]
515	Sertindole	0	0	0	0	0	0	1	No	[4]
516	Sertraline	0	0	0	0	0	0	1	No	[3]
517	Sevoflurane	0	0	0	0	0	1	0	No	[2]
518	Sibutramine	0	0	0	0	0	0	1	No	[3]
519	Sildenafil	0	0	0	1	0	0	1	Yes	[3]
520	Simvastatin	0	0	0	0	0	0	1	No	[3]
521	Sirolimus	0	0	0	0	0	0	1	No	[3]
522	Solifenacin	0	0	0	0	0	0	1	No	[3]
523	Sorafemib	0	0	0	0	0	0	1	No	[3]
524	Sparteine	0	0	0	0	1	0	0	No	[4]
525	Sufentanil	0	0	0	0	0	0	1	No	[2]
526	Sulfamethizole	0	0	0	1	0	0	0	No	[4]
527	Sulfamethoxazole	0	0	0	1	0	0	0	No	[3]
528	Sulfidimidine	0	0	0	0	0	0	1	No	[4]
529	Sunitinib	0	0	0	0	0	0	1	No	[3]

Dataset: *continued*

Mol Id	Molecule name	CYP1A2	CYP2C19	CYP2C8	CYP2C9	CYP2D6	CYP2E1	CYP3A4	Multilabel	References
530	Suprofen	0	0	0	1	0	0	0	No	[5]
531	Tacrine	1	0	0	0	0	0	0	No	[3]
532	Tacrolimus	0	0	0	0	0	0	1	No	[3]
533	Tadalafil	0	0	0	0	0	0	1	No	[3]
534	Tamoxifen	0	0	0	1	1	0	1	Yes	[3]
535	Tamsulosin	0	0	0	0	1	0	1	Yes	[2]
536	Telithromycin	0	0	0	0	0	0	1	No	[3]
537	Temazepam	0	0	0	0	0	0	1	No	[3]
538	Teniposide	0	1	0	0	0	0	0	No	[3]
539	Tenoxicam	0	0	0	1	0	0	0	No	[2]
540	Terfenadine	0	0	0	0	0	0	1	No	[3]
541	Testosterone	0	0	0	0	0	0	1	No	[3]
542	Theophylline	1	0	0	0	0	1	0	Yes	[3]
543	Thioridazine	0	1	0	0	1	0	0	Yes	[3]
544	Timolol	0	0	0	0	1	0	0	No	[3]
545	Tinidazole	0	0	0	0	0	0	1	No	[3]
546	Tipranavir	0	0	0	0	0	0	1	No	[3]
547	Tizanidine	1	0	0	0	0	0	0	No	[3]
548	Tolbutamide	0	1	1	1	0	0	0	Yes	[3]
549	Tolterodine	0	0	0	0	1	0	1	Yes	[3]
550	Torasemide	0	0	0	1	0	0	0	No	[3]
551	Toremifene	0	0	0	0	0	0	1	No	[3]

Dataset: continued

Mol Id	Molecule name	CYP1A2	CYP2C19	CYP2C8	CYP2C9	CYP2D6	CYP2E1	CYP3A4	Multilabel	References
552	Tramadol	0	0	0	0	1	0	1	Yes	[3]
553	Trazodone	0	0	0	0	1	0	1	Yes	[3]
554	Triazolam	0	0	0	0	0	0	1	No	[3]
555	Trimethoprim	0	0	0	1	0	0	0	No	[4]
556	Trimetrexate	0	0	0	0	0	0	1	No	[3]
557	Trimipramine	0	0	0	0	1	0	0	No	[2]
558	Tropisetron	0	0	0	0	1	0	1	Yes	[2]
559	Valdecoxib	0	0	0	1	0	0	1	Yes	[3]
560	Valsartan	0	0	0	1	0	0	0	No	[3]
561	Vardenafil	0	0	0	1	0	0	1	Yes	[3]
562	Venlafaxine	0	0	0	0	0	0	0	No	[3]
563	Verapamil	1	0	1	0	0	0	1	Yes	[3]
564	Vinblastine	0	0	0	0	0	0	1	No	[3]
565	Vincristine	0	0	0	0	0	0	1	No	[3]
566	Vindesine	0	0	0	0	0	0	1	No	[2]
567	Vinorelbine	0	0	0	0	0	0	1	No	[3]
568	Voriconazole	0	1	0	1	0	0	1	Yes	[3]
569	Warfarin-(R)	1	1	1	0	0	0	1	Yes	[3]
570	Warfarin-(S)	0	0	1	1	0	0	0	Yes	[3]
571	Zafirlukast	0	0	0	1	0	0	0	No	[3]
572	Zaleplon	0	0	0	0	0	0	1	No	[3]
573	Zidovudine	0	0	0	0	0	0	1	No	[4]

Dataset: continued

Mol Id	Molecule name	CYP1A2	CYP2C19	CYP2C8	CYP2C9	CYP2D6	CYP2E1	CYP3A4	Multilabel	References
574	Zileuton	1	0	0	1	0	0	1	Yes	[3]
575	Ziprasidone	0	0	0	0	0	0	1	No	[3]
576	Zolmitriptan	1	0	0	0	0	0	0	No	[5]
577	Zolpidem	0	0	0	0	0	0	1	No	[3]
578	Zonisamide	0	0	0	0	0	0	1	No	[3]
579	Zopiclone	0	0	1	0	0	0	0	No	[3]
580	Zuclopenthixol	0	0	0	0	1	0	0	No	[5]

C.2 References

- [1] MDL Inc. <http://www.mdl.com/products/predictive/metabolite/index.jsp> (accessed May 3, 2008).
- [2] Bonnabry, P.; J., S.; Leemann, T.; P., D. Quantitative drug interactions prediction system (Q-DIPS). A dynamic computer-based method to assist in the choice of clinically relevant in vivo studies. *Clin. Pharmacokinet.* **2001**, *40*, 631-640.
- [3] Block, J. H.; Henry, D. R. Evaluation of descriptors and classification schemes to predict cytochrome substrates in terms of chemical information. *J. Comput. Aided Mol. Des.* **2008**, *22*, 385-392.
- [4] Manga, N.; Duffy, J. C.; Rowe, P. H.; Cronin, M. T. Structure-based methods for the prediction of the dominant P450 enzyme in human drug biotransformation: consideration of CYP3A4, CYP2C9, CYP2D6. *QSAR Environ. Res.* **2005**, *16*, 43-61.
- [5] Flockhart. Cytochrome P450 drug-interaction table. <http://medicine.iupui.edu/clinpharm/ddis/table.asp> (accessed February 10, 2008).

Table C.1

No.	Name	Details
94	2D-AC _{identity} (5)	Topological autocorrelation
145	2D-AC _{qπ} (3)	Topological autocorrelation
148	2D-AC _{qπ} (6)	Topological autocorrelation
126	2D-AC _{χπ} (5)	Topological autocorrelation
133	2D-AC _{qσ} (1)	Topological autocorrelation
134	2D-AC _{qσ} (2)	Topological autocorrelation
116	2D-AC _{χσ} (6)	Topological autocorrelation
223	3D-AC _{identity} ([5.8-5.9]Å)	Spatial autocorrelation
33	<i>n_{acidic_groups}</i>	Number of acidic functional groups
27	<i>n_{aliph_amino}</i>	Number of aliphatic amino groups
32	<i>n_{basic_nitrogen}</i>	Number of basic, nitrogen containing functional groups
26	<i>r₃</i>	Radius perpendicular to D_3 and R_2

Table C.2

Classes	TP rate	FP rate	TN rate	FN rate	Recall	Precision	% correct predictions
CYP1A2	0.09	0.02	0.98	0.91	0.09	0.50	83.6
CYP2C19	0.25	0.02	0.98	0.75	0.25	0.50	94.0
CYP2C8	0.25	0.02	0.98	0.75	0.25	0.50	94.0
CYP2C9	0.73	0.04	0.96	0.27	0.73	0.80	92.5
CYP2D6	0.93	0.08	0.92	0.07	0.93	0.78	92.5
CYP2E1	1.00	0.02	0.98	0.00	1.00	0.89	98.5
CYP3A4	0.68	0.17	0.83	0.32	0.68	0.78	76.1

Table C.3

Model prediction	Accuracy $ML(\alpha=1)$	One-error	Coverage	Average precision
Validation set 1	0.70	0.19	2.04	0.85
Test set 1	0.71	0.25	1.94	0.82

Table C.4

Classes	TP rate	FP rate	TN rate	FN rate	Recall	Precision	% correct predictions
CYP1A2	0.48	0.05	0.95	0.52	0.48	0.52	89.9
CYP2C19	0.10	0.01	0.99	0.90	0.10	0.25	94.5
CYP2C8	0.33	0.02	0.98	0.77	0.33	0.17	94.5
CYP2C9	0.44	0.03	0.97	0.56	0.44	0.61	93.1
CYP2D6	0.86	0.15	0.85	0.14	0.86	0.70	85.2
CYP2E1	0.92	0.03	0.97	0.08	0.92	0.61	96.3
CYP3A4	0.78	0.22	0.78	0.22	0.78	0.80	78.3

Table C.5

		Classes CYP						
		1A2	2C19	2C8	2C9	2D6	2E1	3A4
Parameters	C	4	4	4	4	4	4	8
	γ	0.5	2	0.5	0.5	0.5	0.5	0.5

Table C.6

		Classes CYP				
		1A2	2C9	2D6	2E1	3A4
Parameters	C	4	4	8	4	4
	γ	0.5	0.5	0.5	0.5	0.5

Table C.7

Parameter set	
Net dimensions	18x15
Initial span (x, y)	9, 7.5
Span step (x, y)	0.9, 0.75
Number of cycles	3450 (10 epochs)
Topology	rectangular

Table C.8

Parameter set	
Kernel	polynomial, degree=3
C	1

Table C.9

Classes	% correct predictions
CYP1A2	81.4
CYP2C9	87.8
CYP2D6	87.8
CYP2E1	98.0
CYP3A4	76.5

APPENDIX D

Paper IV

Supplementary Information

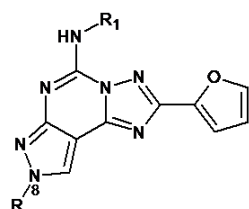
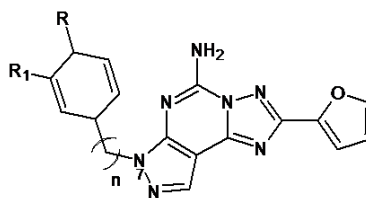
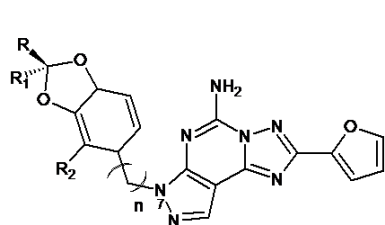
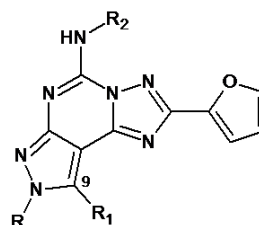
D.1 SVM classification model, training set.

D.2 SVM regression model, training set.

D.3 Test set.

D.4 References.

D.1 SVM classification model, training set

**1-71****72-96****97-101****102-104**

1-42: [1]; 43: [2]; 44-56: [3]; 57-62: [4]; 63-71: [5]; 72-91: [2]; 92-93: [6]; 94-96: [7]; 97-101: [2]; 102-104: [8]

Mol Id	n	R	R ₁	R ₂	Class in SVMclass model	Exp. K _i (nM) hA _{2,4} R	Exp. K _i (nM) hA ₃ R
1	-	CH ₃	3,4-Cl ₂ -Ph-NH-CO	-	-1	143	3.40
2	-	CH ₃	3,4-OCH ₂ O-Ph-NH-CO	-	-1	680	0.24
3	-	CH ₃	4-NO ₂ -Ph-NH-CO	-	-1	695	0.43
4	-	CH ₃	4-CH ₃ -Ph-NH-CO	-	-1	110	0.31
5	-	CH ₃	4-Br-Ph-NH-CO	-	-1	100	0.46
6	-	CH ₃	4-F-Ph-NH-CO	-	-1	120	0.34
7	-	CH ₃	4-CF ₃ -Ph-NH-CO	-	-1	140	0.74
8	-	CH ₃	2-OCH ₃ -Ph-NH-CO	-	-1	180	0.70
9	-	CH ₃	3-OCH ₃ -Ph-NH-CO	-	-1	160	0.80
10	-	CH ₃	2-Cl-Ph-NH-CO	-	-1	200	0.91
11	-	CH ₃	4-Cl-Ph-NH-CO	-	-1	180	0.29
12	-	C ₂ H ₅	3,4-Cl ₂ -Ph-NH-CO	-	-1	352	3.00
13	-	C ₂ H ₅	3,4-OCH ₂ O-Ph-NH-CO	-	-1	576	0.27
14	-	C ₂ H ₅	4-NO ₂ -Ph-NH-CO	-	-1	614	0.65
15	-	C ₂ H ₅	2-OCH ₃ -Ph-NH-CO	-	-1	133	0.56
16	-	C ₂ H ₅	3-OCH ₃ -Ph-NH-CO	-	-1	140	0.86
17	-	C ₂ H ₅	2-Cl-Ph-NH-CO	-	-1	150	0.30
18	-	C ₂ H ₅	4-Cl-Ph-NH-CO	-	-1	160	0.20
19	-	<i>m</i> -C ₃ H ₇	3,4-Cl ₂ -Ph-NH-CO	-	-1	401	2.50
20	-	<i>m</i> -C ₃ H ₇	3,4-OCH ₂ O-Ph-NH-CO	-	-1	667	0.30
21	-	<i>m</i> -C ₃ H ₇	4-NO ₂ -Ph-NH-CO	-	-1	1115	0.81
22	-	<i>m</i> -C ₃ H ₇	2-OCH ₃ -Ph-NH-CO	-	-1	100	0.34
23	-	<i>m</i> -C ₃ H ₇	3-OCH ₃ -Ph-NH-CO	-	-1	113	0.40

SVM classification model, training set: continued

Mol Id	n	R	R ₁	R ₂	Class in SVMclass model	Exp. K _i (nM) hA _{2,4} R	Exp. K _i (nM) hA ₃ R
24	-	<i>n</i> -C ₃ H ₇	2-Cl-Ph-NH-CO	-	-1	121	0.71
25	-	<i>n</i> -C ₃ H ₇	4-Cl-Ph-NH-CO	-	-1	140	0.34
26	-	<i>n</i> -C ₄ H ₉	3,4-Cl ₂ -Ph-NH-CO	-	-1	495	3.70
27	-	<i>n</i> -C ₄ H ₉	3,4-OCH ₂ O-Ph-NH-CO	-	-1	376	0.50
28	-	<i>n</i> -C ₄ H ₉	4-NO ₂ -Ph-NH-CO	-	-1	503	0.55
29	-	<i>n</i> -C ₄ H ₉	2-Cl-Ph-NH-CO	-	-1	100	0.86
30	-	<i>n</i> -C ₄ H ₉	4-Cl-Ph-NH-CO	-	-1	111	0.43
31	-	CH ₃	Ph-CH ₂ -CO	-	-1	423	0.81
32	-	C ₂ H ₅	Ph-CH ₂ -CO	-	-1	335	1.03
33	-	<i>n</i> -C ₃ H ₇	Ph-CH ₂ -CO	-	-1	306	1.01
34	-	<i>n</i> -C ₄ H ₉	Ph-CH ₂ -CO	-	-1	400	1.11
35	-	CH ₃	Ph-NH-CO	-	-1	381	0.16
36	-	<i>n</i> -C ₄ H ₉	Ph-NH-CO	-	-1	148	0.21
37	-	CH ₃	4-OCH ₃ -Ph-NH-CO	-	-1	1390	0.20
38	-	CH ₃	3-Cl-Ph-NH-CO	-	-1	1195	0.40
39	-	C ₂ H ₅	4-OCH ₃ -Ph-NH-CO	-	-1	1040	0.60
40	-	C ₂ H ₅	3-Cl-Ph-NH-CO	-	-1	180	1.60
41	-	<i>n</i> -C ₃ H ₇	4-OCH ₃ -Ph-NH-CO	-	-1	140	0.80
42	-	<i>n</i> -C ₃ H ₇	3-Cl-Ph-NH-CO	-	-1	1220	0.91
43	-	H	4-OCH ₃ -Ph-NH-CO	-	-1	520	0.14
44	-	H	3-Cl-Ph-NH-CO	-	-1	248	0.50


SVM classification model, training set: continued

Mol Id	n	R	R ₁	R ₂	Class in SVMclass model	Exp. K _i (nM) hA _{2A} R	K _i	Exp. K _i (nM) hA ₃ R
45	-	CH ₃	H	-	1	2.80		300
46	-	CH ₂ -CH=CH ₂	4-OCH ₃ -Ph-NH-CO	-	-1	176		0.48
47	-	n-C ₄ H ₉	H	-	1	1.60		600
48	-	t-C ₄ H ₉	H	-	1	45		1149
49	-	t-C ₄ H ₉	4-OCH ₃ -Ph-NH-CO	-	-1	545		0.80
50	-	t-C ₄ H ₉	3-Cl-Ph-NH-CO	-	-1	796		2.78
51	-	(CH ₃) ₂ CH-CH ₂ -CH ₂	H	-	1	0.78		700
52	-	(CH ₃) ₂ C=CH ₂ -CH ₂	H	-	1	0.80		811
53	-	(CH ₃) ₂ C=CH ₂ -CH ₂	4-OCH ₃ -Ph-NH-CO	-	-1	1025		40
54	-	2,4,5-Br ₃ -Ph-CH ₂ -CH ₂	4-OCH ₃ -Ph-NH-CO	-	-1	2030		25
55	-	2,4,5-Br ₃ -Ph-CH ₂ -CH ₂	3-Cl-Ph-NH-CO	-	-1	3600		71
56	-	2-(α -naphthyl)ethyl	4-OCH ₃ -Ph-NH-CO	-	-1	3260		16
57	-	C ₂ H ₅	H	-	1	1.95		3759
58	-	n-C ₃ H ₇	H	-	1	2.51		613
59	-	Ph-CH ₂ -CH ₂	H	-	1	0.34		2785
60	-	Ph-CH ₂ -CH ₂ -CH ₂	H	-	1	0.16		2666
61	-	Ph-CH ₂ -CH ₂	4-OCH ₃ -Ph-NH-CO	-	-1	120		1.47
62	-	Ph-CH ₂ -CH ₂ -CH ₂	4-OCH ₃ -Ph-NH-CO	-	-1	1010		19.81
63	-	CH ₂ -CH ₃	CO-CHPh ₂	-	-1	131		0.98
64	-	CH ₂ -CH ₂ -CH(CH ₃) ₂	CO-CH ₂ -2-thienyl	-	1	2.03		308

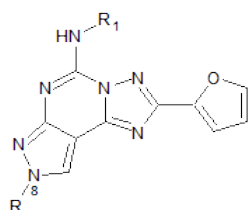
SVM classification model, training set: continued

Mol Id	n	R	R ₁	R ₂	Class in SVMclass model	Exp. K _i (nM) hA _{2A} R	Exp. K _i (nM) hA ₃ R
65	-	CH ₂ -CH ₂ -CH(CH ₃) ₂	CO-CH ₂ -3-thienyl	-	1	3.88	540
66	-	CH ₂ -CH ₂ -CH ₂ -CH ₃	CO-CH ₂ -β-naphthyl	-	1	7.94	174
67	-	CH ₃	CO-CH ₂ -O-Ph	-	1	15.1	692
68	-	CH ₂ -CH ₂ -CH(CH ₃) ₂	CO-CH ₂ -4-OCH ₃ -Ph	-	1	1.95	359
69	-	CH ₂ -CH ₂ -CH(CH ₃) ₂	CO-CH ₂ -O-Ph-4-Cl	-	1	16.1	2306
70	-	CH ₂ -CH ₂ -Ph	CO-CH ₂ -O-Ph-4-Cl	-	1	31.7	3251
71	-	CH ₂ -CH ₂ -CH ₂ -Ph	CO-CH ₂ -4-OCH ₃ -Ph	-	1	8.93	120
72	2	N(CH ₂ CH ₂ OH) ₂	H	-	1	0.12	>10,000
73	3	NH ₂	H	-	1	0.22	>10,000
74	3	CH ₂ NH ₂	H	-	1	0.13	>10,000
75	2	SO ₃ H	H	-	1	100	>10,000
76	3	NO ₂	H	-	1	1.00	>10,000
77	2	NHCOCH ₃	H	-	1	4.80	>10,000
78	3	N(CH ₂ CH ₂ OH) ₂	H	-	1	1.10	>10,000
79	3	CN	H	-	1	86	>10,000
80	3	COOEt	H	-	1	4.00	>10,000
81	2	OCH ₂ COOEt	H	-	1	0.43	>10,000
82	3	C(NO ₂)NH ₂	H	-	1	6.00	>10,000
83	2	NH ₂	H	-	1	55	>10,000
84	3	C(NH)NH ₂	H	-	1	4.4	>10,000
85	3	COOH	H	-	1	4.63	>10,000

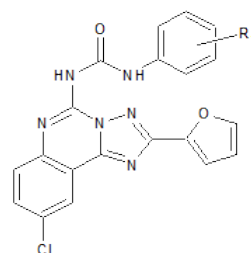
SVM classification model, training set: continued

Mol Id	n	R	R ₁	R ₂	Class in SVMclass model	Exp. K _i (nM) hA _{2A} R	Exp. K _i (nM) hA ₃ R
86	2	SO ₂ NH ₂	H	-	1	1.31	>10,000
87	2	SO ₂ N(CH ₂ CH ₂ OH) ₂	H	-	1	0.80	>10,000
88	2	SO ₂ N  CH ₃	H	-	1	3.80	>10,000
89	2	SO ₂ N(CH ₂ CH ₂ Cl) ₂	H	-	1	0.59	>10,000
90	2	SO ₂ NHCH ₂ COOH	H	-	1	50.00	>10,000
91	3	SO ₃ H	H	-	1	140	>10,000
92	3	H	H	-	1	1.20	>10,000
93	2	H	H	-	1	1.10	>10,000
94	3	OCH ₂ O	OCH ₂ O	-	1	3.30	>10,000
95	3	OCH ₃	OCH ₃	-	1	2.7	>10,000
96	3	OH	H	-	1	1.50	>10,000
97	3	CH ₂ OH	CH ₂ OH	H	1	0.19	>10,000
98	3	H	H	SO ₃ H	1	75	>10,000
99	3	COOEt	H	H	1	5.48	>10,000
100	3	COOH	COOH	H	1	120	>10,000
101	3	COOH	H	H	1	59	>10,000
102	-	CH ₃	S(CH ₂) ₂ CH ₃	H	1	2.1	224
103	-	CH ₃	SCH ₃	COCH ₂ Ph-4-OCH ₃	1	15	>1,000
104	-	CH ₃	SCH ₃	COCH ₂ Ph-4-iso butyl	1	50	>1,000

D.2 SVM regression model, training set



1-71, 105-126, 129-137



127-128

1-42: [1]; 43: [2]; 44-56: [3]; 57-62: [4]; 63-71: [5]; 105-128: [1]; 129-133: [3]; 134-135: [4]; 136-137: [9]

Mol Id	R	R ₁	Pred. (nM) (SVM model)	pK _i hA _{2A} R (nM) (SVM model)	Pred. (nM) (SVM model)	pK _i hA ₃ R (nM) (SVM model)	Exp. pK _i hA _{2A} R (nM)	Exp. pK _i hA ₃ R (nM)	Exp. K _i hA _{2A} R (nM)	Exp. K _i hA ₃ R (nM)
1^a	CH ₃	3,4-Cl ₂ -Ph-NH-CO	-2.25		0.62		-2.16	-0.53	143	3.40
2^a	CH ₃	3,4-OCH ₂ O-Ph-NH-CO	-2.71		0.60		-2.83	0.62	680	0.24
3^a	CH ₃	4-NO ₂ -Ph-NH-CO	-2.94		0.35		-2.84	0.37	695	0.43
4^a	CH ₃	4-CH ₃ -Ph-NH-CO	-2.14		1.22		-2.04	0.51	110	0.31
5^a	CH ₃	4-Br-Ph-NH-CO	-2.04		0.33		-2.00	0.34	100	0.46
6^a	CH ₃	4-F-Ph-NH-CO	-2.31		0.16		-2.08	0.47	120	0.34
7^a	CH ₃	4-CF ₃ -Ph-NH-CO	-2.14		0.04		-2.15	0.13	140	0.74
8^a	CH ₃	2-OCH ₃ -Ph-NH-CO	-2.84		0.14		-2.26	0.15	180	0.70
9^a	CH ₃	3-OCH ₃ -Ph-NH-CO	-2.60		0.52		-2.20	0.10	160	0.80
10^a	CH ₃	2-Cl-Ph-NH-CO	-2.26		-0.15		-2.30	0.04	200	0.91
11^a	CH ₃	4-Cl-Ph-NH-CO	-2.02		0.24		-2.26	0.54	180	0.29
12^a	C ₂ H ₅	3,4-Cl ₂ -Ph-NH-CO	-2.34		-0.10		-2.55	-0.48	352	3.00
13^a	C ₂ H ₅	3,4-OCH ₂ O-Ph-NH-CO	-2.45		0.74		-2.76	0.57	576	0.27
14^a	C ₂ H ₅	4-NO ₂ -Ph-NH-CO	-2.63		-0.34		-2.79	0.19	614	0.65
15^a	C ₂ H ₅	2-OCH ₃ -Ph-NH-CO	-2.39		0.81		-2.12	0.25	133	0.56
16^a	C ₂ H ₅	3-OCH ₃ -Ph-NH-CO	-2.28		0.59		-2.15	0.07	140	0.86
17^a	C ₂ H ₅	2-Cl-Ph-NH-CO	-2.10		0.23		-2.18	0.52	150	0.30
18^a	C ₂ H ₅	4-Cl-Ph-NH-CO	-1.88		0.36		-2.20	0.70	160	0.20
19^a	<i>n</i> -C ₃ H ₇	3,4-Cl ₂ -Ph-NH-CO	-2.65		-0.28		-2.60	-0.40	401	2.50
20^a	<i>n</i> -C ₃ H ₇	3,4-OCH ₂ O-Ph-NH-CO	-2.61		0.51		-2.82	0.52	667	0.30

^a Molecules already introduced in the previous SVM classification model

SVM regression model, training set: continued

Mol Id	R	R ₁	Pred. (nM) (SVM model)	pK _i hA _{2A} R (nM) (SVM model)	Pred. (nM) (SVM model)	pK _i hA ₃ R (nM) (SVM model)	Exp. pK _i hA _{2A} R (nM)	Exp. pK _i hA ₃ R (nM)	Exp. K _i hA _{2A} R (nM)	Exp. K _i hA ₃ R (nM)
21^a	<i>n</i> -C ₃ H ₇		-2.80		-0.15		-3.05	0.09	1115	0.81
22^a	<i>n</i> -C ₃ H ₇	4-NO ₂ -Ph-NH-CO	-1.77		-0.28		-2.00	0.47	100	0.34
23^a	<i>n</i> -C ₃ H ₇	2-OCH ₃ -Ph-NH-CO	-2.10		-0.14		-2.05	0.40	113	0.40
24^a	<i>n</i> -C ₃ H ₇	3-OCH ₃ -Ph-NH-CO	-2.93		-0.23		-2.70	0.26	503	0.55
25^a	<i>n</i> -C ₃ H ₇	2-Cl-Ph-NH-CO	-1.85		-0.71		-2.00	0.07	100	0.86
26^a	<i>n</i> -C ₄ H ₉	4-Cl-Ph-NH-CO	-2.13		0.39		-2.05	0.37	111	0.43
27^a	<i>n</i> -C ₄ H ₉	3,4-Cl ₂ -Ph-NH-CO	-1.43		0.25		-2.63	0.09	423	0.81
28^a	<i>n</i> -C ₄ H ₉	3,4-OCH ₂ O-Ph-NH-CO	-1.28		0.62		-2.53	-0.01	335	1.03
29^a	<i>n</i> -C ₄ H ₉	4-NO ₂ -Ph-NH-CO	-1.34		-0.86		-2.49	0.00	306	1.01
30^a	<i>n</i> -C ₄ H ₉	2-Cl-Ph-NH-CO	-1.28		-1.75		-2.60	-0.05	400	1.11
31^a	CH ₃	4-Cl-Ph-NH-CO	-1.78		0.13		-2.58	0.80	381	0.16
32^a	C ₂ H ₅	Ph-CH ₂ -CO	-2.93		-0.23		-2.70	0.26	503	0.55
33^a	<i>n</i> -C ₃ H ₇	Ph-CH ₂ -CO	-1.85		-0.71		-2.00	0.07	100	0.86
34^a	<i>n</i> -C ₄ H ₉	Ph-CH ₂ -CO	-2.13		0.39		-2.05	0.37	111	0.43
35^a	CH ₃	Ph-NH-CO	-1.43		0.25		-2.63	0.09	423	0.81
36^a	<i>n</i> -C ₄ H ₉	Ph-NH-CO	-1.75		0.37		-2.17	0.68	148	0.21
37^a	CH ₃	4-OCH ₃ -Ph-NH-CO	-2.30		0.28		-3.14	0.70	1390	0.20
38^a	CH ₃	3-Cl-Ph-NH-CO	-2.39		0.20		-3.08	0.40	1195	0.40

^a Molecules already introduced in the previous SVM classification model

SVM regression model, training set: continued

Mol Id	R	R ₁	Pred. (nM) (SVM model)	pK _i hA _{2A} R (nM) hA _{2A} R (SVM model)	Pred. (nM) (SVM model)	pK _i hA _{3R} hA _{3R}	Exp. pK _i (nM) hA _{2A} R	Exp. pK _i (nM) hA _{3R}	Exp. K _i (nM) hA _{2A} R	Exp. K _i (nM) hA _{3R}
39 ^a	C ₂ H ₅	4-OCH ₃ -Ph-NH-CO	-2.26		0.07		-3.02	0.22	1040	0.60
40 ^a	C ₂ H ₅	3-Cl-Ph-NH-CO	-2.07		-0.15		-2.26	-0.20	180	1.60
41 ^a	<i>n</i> -C ₃ H ₇	4-OCH ₃ -Ph-NH-CO	-2.06		-0.07		-2.15	0.10	140	0.80
42 ^a	<i>n</i> -C ₃ H ₇	3-Cl-Ph-NH-CO	-2.08		-0.70		-3.09	0.04	1220	0.91
43 ^a	H	4-OCH ₃ -Ph-NH-CO	-2.42		0.09		-2.72	0.85	520	0.14
44 ^a	H	3-Cl-Ph-NH-CO	-2.60		0.81		-2.39	0.30	248	0.50
45 ^a	CH ₃	H	-0.50		-2.55		-0.45	-2.48	2.80	300
46 ^a	CH ₂ -CH=CH ₂	4-OCH ₃ -Ph-NH-CO	-2.36		-0.12		-2.25	-3.31	176	0.48
47 ^a	<i>n</i> -C ₄ H ₉	H	-0.21		-3.41		-0.20	-2.78	1.60	600
48 ^a	<i>t</i> -C ₄ H ₉	H	-1.47		-1.97		-1.65	-3.06	45	1149
49 ^a	<i>t</i> -C ₄ H ₉	4-OCH ₃ -Ph-NH-CO	-2.56		0.54		-2.74	0.10	545	0.80
50 ^a	<i>t</i> -C ₄ H ₉	3-Cl-Ph-NH-CO	-2.91		0.16		-2.90	-0.44	796	2.78
51 ^a	(CH ₃) ₂ CH-CH ₂ -CH ₂	H	0.08		-2.49		0.11	-2.85	0.78	700
52 ^a	(CH ₃) ₂ C=CH ₂ -CH ₂	H	-0.32		-2.57		0.10	-2.91	0.80	811
53 ^a	(CH ₃) ₂ C=CH ₂ -CH ₂	4-OCH ₃ -Ph-NH-CO	-1.89		-0.08		-3.01	-1.60	1025	40
54 ^a	2,4,5-Br ₃ -Ph-CH ₂ -CH ₂	4-OCH ₃ -Ph-NH-CO	-2.68		-1.56		-3.31	-1.40	2030	25
55 ^a	2,4,5-Br ₃ -Ph-CH ₂ -CH ₂	3-Cl-Ph-NH-CO	-2.62		-0.27		-3.56	-1.85	3600	71
56 ^a	2-(<i>o</i> -naphthyl)ethyl	4-OCH ₃ -Ph-NH-CO	-1.83		-1.01		-3.51	-1.20	3260	16

^a Molecules already introduced in the previous SVM classification model

SVM regression model, training set: continued

Mol Id	R	R ₁	Pred.		pK _i		Exp.		Exp.		Exp.	
			(nM) (SVM model)	hA _{2A} R (nM) (SVM model)	hA ₃ R (nM) (SVM model)	hA _{2A} R (nM) (SVM model)	hA ₃ R (nM) (SVM model)	hA _{2A} R (nM) (SVM model)	hA ₃ R (nM) (SVM model)	hA _{2A} R (nM) (SVM model)	hA ₃ R (nM) (SVM model)	hA _{2A} R (nM) (SVM model)
57^a	C ₂ H ₅	H	-0.33	-2.97	-0.29	-3.58	1.95	3759				
58^a	n-C ₃ H ₇	H	-0.31	-2.88	-0.40	-2.79	2.51	613				
59^a	Ph-CH ₂ -CH ₂	H	0.51	-3.03	0.47	-3.44	0.34	2785				
60^a	Ph-CH ₂ -CH ₂ -CH ₂	H	0.01	-3.10	0.80	-3.43	0.16	2666				
61^a	Ph-CH ₂ -CH ₂	4-OCH ₃ -Ph-NH-CO	-2.06	-0.59	-2.08	-0.17	120	1.47				
62^a	Ph-CH ₂ -CH ₂ -CH ₂	4-OCH ₃ -Ph-NH-CO	-1.89	-0.91	-3.00	-1.30	1010	19.81				
63^a	CH ₂ -CH ₃	CO-CHPh ₂	-1.21	-0.54	-2.12	0.01	131	0.98				
64^a	CH ₂ -CH ₂ -CH(CH ₃) ₂	CO-CH ₂ -2-thienyl	-1.30	-0.78	-0.31	-2.49	2.03	308				
65^a	CH ₂ -CH ₂ -CH(CH ₃) ₂	CO-CH ₂ -3-thienyl	-1.62	-1.51	-0.59	-2.73	3.88	540				
66^a	CH ₂ -CH ₂ -CH ₂ -CH ₃	CO-CH ₂ -β-naphthyl	-1.47	-1.90	-0.90	-2.24	7.94	174				
67^a	CH ₃	CO-CH ₂ -O-Ph	-1.88	-0.86	-1.18	-2.84	15.1	692				
68^a	CH ₂ -CH ₂ -CH(CH ₃) ₂	CO-CH ₂ -4-OCH ₃ -Ph	-1.49	-1.75	-0.29	-2.56	1.95	359				
69^a	CH ₂ -CH ₂ -CH(CH ₃) ₂	CO-CH ₂ -O-Ph-4-Cl	-1.23	-2.37	-1.21	-3.36	16.1	2306				
70^a	CH ₂ -CH ₂ -Ph	CO-CH ₂ -O-Ph-4-Cl	-1.24	-2.61	-1.50	-3.51	31.7	3251				
71^a	CH ₂ -CH ₂ -CH ₂ -Ph	CO-CH ₂ -4-OCH ₃ -Ph	-1.38	-3.80	-0.95	-2.08	8.93	120				

^a Molecules already introduced in the previous SVM classification model

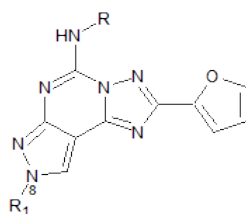
SVM regression model, training set: continued

Mol Id	R	R ₁	Pred. (nM) (SVM model)	pK _i hA _{2A} R (nM) (SVM model)	Pred. (nM) (SVM model)	pK _i hA ₃ R (nM) (SVM model)	Exp. pK _i hA _{2A} R (nM)	Exp. pK _i hA ₃ R (nM)	Exp. K _i hA _{2A} R (nM)	Exp. K _i hA ₃ R (nM)
105	C ₂ H ₅	4-CH ₃ -Ph-NH-CO	-1.87		0.86		-1.48	0.85	30	0.14
106	C ₂ H ₅	4-Br-Ph-NH-CO	-1.99		0.80		-1.60	0.43	40	0.37
107	C ₂ H ₅	4-F-Ph-NH-CO	-2.17		0.18		-1.78	0.07	60	0.86
108	C ₂ H ₅	4-CF ₃ -Ph-NH-CO	-2.10		0.33		-1.72	0.01	53	0.97
109	n-C ₃ H ₇	4-CH ₃ -Ph-NH-CO	-1.70		0.61		-1.08	0.40	12	0.40
110	n-C ₃ H ₇	4-Br-Ph-NH-CO	-1.96		0.60		-1.70	0.35	50	0.45
111	n-C ₃ H ₇	4-F-Ph-NH-CO	-2.32		0.36		-1.62	0.54	42	0.29
112	n-C ₃ H ₇	4-CF ₃ -Ph-NH-CO	-2.14		0.01		-1.53	0.29	34	0.51
113	n-C ₄ H ₉	4-CH ₃ -Ph-NH-CO	-1.81		0.35		-1.38	0.68	24	0.21
114	n-C ₄ H ₉	4-Br-Ph-NH-CO	-2.10		0.27		-1.82	0.04	66	0.91
115	n-C ₄ H ₉	4-F-Ph-NH-CO	-2.38		0.58		-1.70	0.10	50	0.80
116	n-C ₄ H ₉	4-CF ₃ -Ph-NH-CO	-2.26		-0.40		-1.49	0.14	31	0.72
117	n-C ₄ H ₉	2-OCH ₃ -Ph-NH-CO	-1.49		0.38		-1.96	0.24	91	0.57
118	n-C ₄ H ₉	3-OCH ₃ -Ph-NH-CO	-1.98		-0.10		-1.98	0.22	95	0.60
119	C ₂ H ₅	Ph-NH-CO	-1.95		0.42		-1.60	0.74	40	0.18
120	n-C ₃ H ₇	Ph-NH-CO	-1.71		0.28		-1.79	0.82	62	0.15
121	CH ₃	4-SO ₃ -Ph-NH-CO	-2.52		-1.50		-2.77	-1.40	594	25
122	C ₂ H ₅	4-SO ₃ -Ph-NH-CO	-2.56		-1.14		-2.40	-1.60	249	40
123	n-C ₃ H ₇	4-SO ₃ -Ph-NH-CO	-2.49		-1.57		-2.48	-1.48	305	30

SVM regression model, training set: continued

Mol Id	R	R ₁	Pred.		pK _i		Exp.		Exp. K _i (nM) hA ₃ R
			(nM) (SVM model)	hA _{2A} R (nM) (SVM model)	hA ₃ R (nM) (SVM model)	hA _{2A} R (nM) (SVM model)	hA ₃ R (nM) (SVM model)	hA _{2A} R (nM) (SVM model)	
124	<i>n</i> -C ₄ H ₉	4-SO ₃ -Ph-NH-CO	-2.52	-1.30	-2.55	-1.67	352	47	
125	<i>n</i> -C ₄ H ₉	4-OCH ₃ -Ph-NH-CO	-2.17	-0.14	-1.90	0.49	80	0.32	
126	<i>n</i> -C ₄ H ₉	3-Cl-Ph-NH-CO	-2.38	-0.35	-1.98	0.22	95	0.60	
127	-	4-OCH ₃ -Ph-NH-CO	-0.60	-1.59	-0.97	0.85	9.40	0.14	
128	-	3-Cl-Ph-NH-CO	-2.19	-1.09	-0.98	0.72	9.50	0.19	
129	H	H	-1.08	-3.02	-1.30	-2.54	20	348	
130	(CH ₃) ₂ CH-CH ₂ -CH ₂	4-OCH ₃ -Ph-NH-CO	-2.07	-0.07	-1.81	-1.48	65	30	
131	(CH ₃) ₂ CH-CH ₂ -CH ₂	3-Cl-Ph-NH-CO	-2.32	-0.28	-2.06	-1.60	115	40	
132	(CH ₃) ₂ C=CH ₂ -CH ₂	3-Cl-Ph-NH-CO	-2.26	-0.36	-2.04	-1.74	110	55	
133	2,4,5-Br ₃ -Ph-CH ₂ -CH ₂	H	-1.97	-3.52	-2.54	-3.65	348	4481	
134	Ph-CH ₂ -CH ₂	3-Cl-Ph-NH-CO	-2.08	-0.39	-2.02	-1.12	105	13.28	
135	Ph-CH ₂ -CH ₂ -CH ₂	3-Cl-Ph-NH-CO	-1.85	-1.23	-2.30	-1.63	200	42.65	
136	Ph-CH ₂ -CH ₂	(CH ₃) ₃ COCONH(CH ₂) ₃ -CO	-2.43	-3.77	-2.46	-3.17	288	1480	
137	(CH ₃) ₂ CH-CH ₂ -CH ₂	(CH ₃) ₃ COCONH(CH ₂) ₃ -CO	-2.45	-2.80	-2.47	-3.40	297	2488	

D.3 Test set

**138-188**

138-188: [5]

Mol Id	R	R ₁	Exp. class in SVMclass model	Pred. class by SVMclass model	Pred. pK _i (nM) (SVM model)	Pred. pK _i (nM) (SVM model)	Exp. pK _i (nM) hA _{2A} R	Exp. pK _i (nM) hA ₃ R	Exp. K _i (nM) hA _{2A} R	Exp. K _i (nM) hA ₃ R
138	CH ₂ - α -naphthyl	CH ₂ -CH ₂ -CH ₂ -Ph	1	1	-1.63	-2.76	-2.48	-2.54	305	343
139	CHPh ₂	CH ₃	-1	-1	-1.08	-0.74	-2.33	0.60	216	0.25
140	CH ₂ -Ph-Ph	CH ₃	-1	1	-1.75	-1.46	-2.29	-1.05	193	11.2
141	CHPh ₂	CH ₂ -CH ₂ -CH(CH ₃) ₂	-1	-1	-1.20	0.02	-2.20	-0.77	159	5.86
142	CH ₂ -Ph-Ph	CH ₂ -CH ₂ -CH(CH ₃) ₂	1	1	-1.05	-1.60	-0.96	-2.43	9.18	268
143	CHPh ₂	CH ₂ -CH ₂ -Ph	-1	-1	-1.57	-0.75	-1.73	-0.81	53.1	6.49
144	CH ₂ -Ph-Ph	CH ₂ -CH ₂ -Ph	1	1	-1.59	-2.45	-1.67	-2.10	46.9	125
145	CHPh ₂	CH ₂ -CH ₂ -CH ₂ -CH ₃	-1	-1	-1.19	-0.16	-2.06	-0.08	114	1.20
146	CH ₂ -Ph-Ph	CH ₂ -CH ₂ -CH ₂ -CH ₃	1	1	-1.59	-1.47	-1.60	-2.28	39.6	189
147	CH ₂ -3-Cl-Ph	CH ₃	-1	-1	-1.79	-0.31	-1.48	-0.29	30.5	1.94
148	CHPh ₂	CH ₂ -CH ₂ -CH ₃	-1	-1	-1.23	-0.04	-1.67	0.03	46.5	0.93
149	CH ₂ -Ph-Ph	CH ₂ -CH ₂ -CH ₃	1	1	-1.49	-1.37	-1.72	-1.82	52.5	65.4
150	CH ₂ -4-Cl-Ph	CH ₃	-1	-1	-1.36	-3.43	-2.19	-1.10	156	12.7
151	CH ₂ -Ph-Ph	CH ₂ -CH ₃	-1	1	-1.84	-1.41	-1.85	-1.15	70.9	14.2
152	CH ₂ -4-OCH ₃ -Ph	CH ₃	-1	-1	-2.06	-0.45	-1.80	0.02	62.5	0.95
153	CH ₂ - β -naphthyl	CH ₂ -CH ₂ -CH(CH ₃) ₂	1	1	-1.59	-1.52	-1.20	-2.61	15.7	409
154	CH ₂ - β -naphthyl	CH ₂ -CH ₂ -Ph	1	1	-1.61	-2.18	-1.62	-2.26	42.1	180
155	CH ₂ -2-thienyl	CH ₂ -CH ₂ -Ph	1	1	-1.69	-1.45	-0.94	-2.52	8.8	330
156	CH ₂ -3-thienyl	CH ₂ -CH ₂ -Ph	1	1	-1.73	-1.30	-1.11	-2.86	12.9	726
157	CH ₂ - α -naphthyl	CH ₂ -CH ₃	-1	-1	-2.21	-0.43	-1.87	-0.48	74.6	3.05
158	CH ₂ - β -naphthyl	CH ₂ -CH ₃	1	1	-1.42	-2.40	-1.26	-1.56	18.4	36.3

Test set: continued

Mol Id	R	R ₁	Exp. class in SVMclass model	Pred. class by SVMclass model	Pred. (nM) (SVM model)	pK _i hA _{2A} R hA _{2A} R model	Pred. (nM) hA _{3R} hA _{3R} model	Exp. pK _i (nM) hA _{2A} R hA _{3R}	Exp. pK _i (nM) hA _{3R}	Exp. K _i (nM) hA _{2A} R	Exp. K _i (nM) hA _{3R}
159	CH ₂ -3-thienyl	CH ₂ -CH ₃	1	-1	-2.01		0.11	-1.36	-2.88	23	765
160	CH ₂ -2-thienyl	CH ₂ -CH ₃	1	-1	-1.80		0.41	-1.20	-2.29	15.9	196
161	CH ₂ -2-thienyl	CH ₃	-1	-1	-1.79		0.24	-1.75	-0.72	56	5.26
162	CH ₂ -3-thienyl	CH ₃	-1	-1	-1.82		-0.13	-1.50	-0.10	31.3	1.25
163	CH ₂ -β-naphthyl	CH ₃	-1	-1	-1.91		-0.55	-1.89	-1.16	77.5	14.5
164	CH ₂ -α-naphthyl	CH ₃	-1	-1	-2.21		-0.13	-1.91	-0.54	80.5	3.47
165	CH ₂ -α-naphthyl	CH ₂ -CH ₂ -CH ₃	-1	-1	-1.70		-1.32	-1.59	-1.24	38.7	17.3
166	CH ₂ -β-naphthyl	CH ₂ -CH ₂ -CH ₃	1	1	-1.83		-1.07	-0.90	-1.98	7.99	95.9
167	CH ₂ -α-naphthyl	CH ₂ -CH ₂ -CH ₂ -CH ₃	1	1	-1.81		-0.96	-1.07	-2.00	11.7	100
168	CH ₂ -4-CF ₃ -Ph	CH ₃	-1	-1	-1.68		-1.81	-1.88	-0.09	75.9	1.22
169	CH ₂ -2-thienyl	CH ₂ -CH ₂ -CH ₂ -CH ₃	1	1	-1.58		-0.71	-0.62	-2.00	4.15	99.8
170	CH ₂ -3-thienyl	CH ₂ -CH ₂ -CH ₂ -CH ₃	1	1	-1.73		-0.77	-0.50	-2.28	3.13	189
171	CH ₂ -O-Ph-4-Cl	CH ₃	1	1	-1.65		-3.50	-1.59	-2.35	39.3	223
172	CH ₂ -3-Cl-Ph	CH ₂ -CH ₂ -CH(CH ₃) ₂	1	1	-1.58		-1.43	-0.27	-2.44	1.86	273
173	CH ₂ -4-Cl-Ph	CH ₂ -CH ₂ -CH(CH ₃) ₂	1	1	-1.58		-1.60	-0.44	-1.75	2.75	56.5
174	CH ₂ -3-Cl-Ph	CH ₂ -CH ₂ -Ph	1	1	-1.71		-1.92	-0.76	-2.44	5.75	273
175	CH ₂ -4-CF ₃ -Ph	CH ₂ -CH ₂ -CH(CH ₃) ₂	1	1	-2.14		-2.27	-0.73	-2.42	5.43	266
176	CH ₂ -4-F-Ph	CH ₂ -CH ₂ -CH(CH ₃) ₂	1	1	-1.83		-1.63	-0.57	-2.06	3.69	116
177	CH ₂ -4-F-Ph	CH ₃	-1	-1	-2.06		-0.64	-1.73	0.01	54.1	0.97
178	CH ₂ -2,6-Cl ₂ -Ph	CH ₂ -CH ₂ -CH(CH ₃) ₂	1	1	-1.12		-1.03	-1.27	-2.32	18.7	207

Test set: continued

Mol Id	R	R ₁	Exp. class SVMmodel	in SVMmodel	Pred. class SVMmodel	by SVMmodel	Pred. (nM) (SVM model)	pK _i hA _{2A} R	Pred. (nM) (SVM model)	pK _i hA ₃ R	Exp. pK _i hA _{2A} R (nM)	Exp. pK _i hA ₃ R (nM)	Exp. K _i hA _{2A} R (nM)	Exp. K _i hA ₃ R (nM)
179	CH ₂ -2,6-Cl ₂ -Ph	CH ₃	-1	-1	-1	-1.05	-2.47	-1.66	-1.65	-1.66	-1.65	44.4	45.2	44.4
180	CH ₂ -4-F-Ph	CH ₂ -CH ₂ -CH ₂ -Ph	-1	1	1	-0.98	-3.24	-2.32	-1.77	-2.32	-1.77	211	211	58.5
181	CHPh ₂	CH ₂ -CH ₂ -CH ₂ -Ph	-1	1	1	-1.29	-2.74	-2.51	-1.10	-2.51	-1.10	326	326	12.6
182	CH ₂ -β-naphthyl	CH ₂ -CH ₂ -CH ₂ -Ph	1	1	1	-1.38	-2.50	-1.87	-2.86	-1.87	-2.86	717	73.6	717
183	CH ₂ -Ph	CH ₂ -CH ₂ -Ph	-1	1	1	-1.70	-1.64	-1.64	-0.74	-1.64	-0.74	43.9	43.9	5.49
184	CH ₂ -3-Cl-Ph	CH ₂ -CH ₂ -CH ₂ -Ph	-1	1	1	-1.54	-2.31	-2.26	-2.04	-2.26	-2.04	182	182	110
185	CH ₂ -4-Cl-Ph	CH ₂ -CH ₂ -CH ₂ -Ph	-1	1	1	-1.72	-1.97	-1.95	-1.48	-1.95	-1.48	89.9	89.9	30.5
186	CH ₂ -O-Ph-4-Cl	CH ₂ -CH ₂ -CH ₂ -Ph	1	1	1	-1.32	-2.45	-1.43	-2.60	-1.43	-2.60	27.2	27.2	400
187	CH ₂ -2,6-Cl ₂ -Ph	CH ₂ -CH ₂ -CH ₂ -Ph	1	-1	-1	-0.88	-2.14	-2.27	-2.78	-2.27	-2.78	186	186	601
188	CH ₂ -Ph-Ph	CH ₂ -CH ₂ -CH ₂ -Ph	1	-1	-1	-0.75	-2.96	-2.41	-2.61	-2.41	-2.61	256	256	410

D.4 References

- [1] Baraldi, P. G.; Cacciari, B.; Moro, S.; Spalluto, G.; Pastorin, G.; Da Ros, T.; Klotz, K. N.; Varani, K.; Gessi, S.; Borea, P. A. Synthesis, biological activity, and molecular modeling investigation of new pyrazolo[4,3-e]-1,2,4-triazolo[1,5-c]pyrimidine derivatives as human A₃ adenosine receptor antagonists. *J. Med. Chem.* **2002**, *45*, 770-780.
- [2] Baraldi, P. G.; Tabrizi, M. A.; Bovero, A.; Avitabile, B.; Preti, D.; Fruttarolo, F.; Romagnoli, R.; Varani, K.; Borea, P. A. Recent developments in the field of A_{2A} and A₃ adenosine receptor antagonists. *Eur. J. Med. Chem.* **2003**, *38*, 367-382.
- [3] Baraldi, P. G.; Cacciari, B.; Romagnoli, R.; Spalluto, G.; Moro, S.; Klotz, K. N.; Leung, E.; Varani, K.; Gessi, S.; Merighi, S.; Borea, P. A. Pyrazolo[4,3-e]1,2,4-triazolo[1,5-c]pyrimidine derivatives as highly potent and selective human A₃ adenosine receptor antagonists: influence of the chain at the N⁸ pyrazole nitrogen. *J. Med. Chem.* **2000**, *43*, 4768-4780.
- [4] Baraldi, P. G.; Cacciari, B.; Romagnoli, R.; Spalluto, G.; Varani, K.; Gessi, S.; Merighi, S.; Borea, P. A. Pyrazolo[4,3-e]1,2,4-triazolo[1,5-c]pyrimidine derivatives: a new pharmacological tool for the characterization of the human A₃ adenosine receptor. *Drug Dev. Res.* **2001**, *52*, 406-415.
- [5] Michielan, L.; Bolcato, C.; Federico, S.; Cacciari, B.; Bacilieri, M.; Klotz, K. N.; Kachler, S.; Pastorin, G.; Cardin, R.; Sperduti, A.; Spalluto, G.; Moro, S. Combining selectivity and affinity predictions using an integrated Support Vector Machine (SVM) approach: an alternative tool to discriminate between the human adenosine A_{2A} and A₃ receptor pyrazolo-triazolo-pyrimidine antagonists binding sites. *Bioorg. Med. Chem.* **2009**, *17*, 5259-5274.

- [6] Baraldi, P. G.; Cacciari, B.; Romagnoli, R.; Spalluto, G.; Monopoli, A.; Ongini, E.; Varani, K.; Borea, P. A. 7-Substituted 5-amino-2-(2-furyl)pyrazolo[4,3-e]-1,2,4-triazolo[1,5-c]pyrimidines as A_{2A} adenosine receptor antagonists: a study on the importance of modifications at the side chain on the activity and solubility. *J. Med. Chem.* **2002**, *45*, 115-126.
- [7] Baraldi, P. G.; Cacciari, B.; Spalluto, G.; Bergonzoni, M.; Dionisotti, S.; Ongini, E.; Varani, K.; Borea, P. A. Design, synthesis, and biological evaluation of a second generation of pyrazolo[4,3-e]-1,2,4-triazolo[1,5-c]pyrimidines as potent and selective A_{2A} adenosine receptor antagonists. *J. Med. Chem.* **1998**, *41*, 2126-2133.
- [8] Baraldi, P. G.; Fruttarolo, F.; Tabrizi, M. A.; Preti, D.; Romagnoli, R.; El-Kashef, H.; Moorman, A.; Varani, K.; Gessi, S.; Merighi, S.; Borea, P. A. Design, synthesis, and biological evaluation of C^9 - and C^2 -substituted pyrazolo[4,3-e]-1,2,4-triazolo[1,5-c]pyrimidines as new A_{2A} and A_3 adenosine receptors antagonists. *J. Med. Chem.* **2003**, *46*, 1229-1241.
- [9] Baraldi, P. G.; Cacciari, B.; Romagnoli, R.; Klotz, K. N.; Spalluto, G.; Varani, K.; Gessi, S.; Merighi, S.; Borea, P. A. Pyrazolo[4,3-e]-1,2,4-triazolo[1,5-c]pyrimidine derivatives as adenosine receptor ligands: a starting point for searching A_{2B} adenosine receptor antagonists. *Drug Dev. Res.* **2001**, *53*, 225-235.

APPENDIX E

Paper V

Supplementary Information

E.1 MODELS 1, 2, 3, training set.

E.2 MODELS 1, 2, 3, validation set.

E.3 MODELS 1, 2, 3, internal test set.

E.4 References.

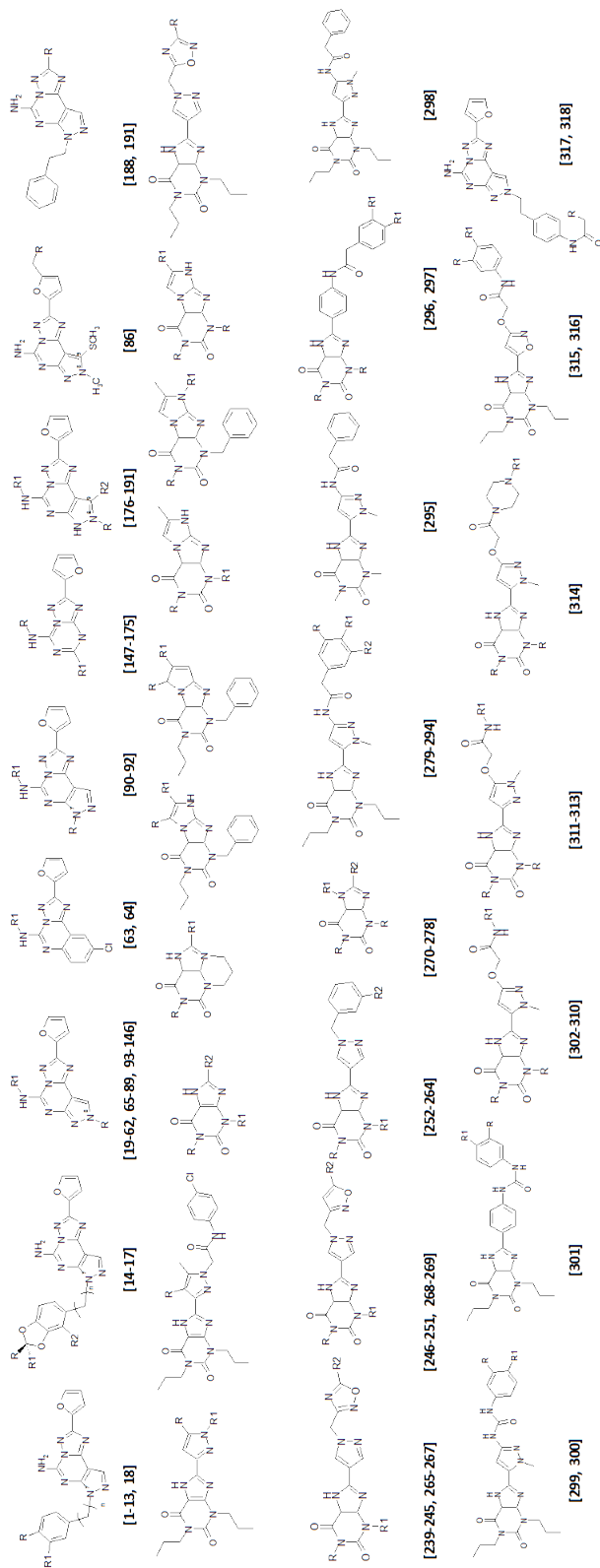
Table E.1 Prediction of the validation set (MODEL 1).

Table E.2 Prediction of the validation set (MODEL 2).


Table E.3 Prediction of the validation set (MODEL 3).

Table E.4 MODELS 1, 2, 3: prediction of the external test set.

E.1 MODELS 1, 2, 3, training set.



1-17, 19: [1]; 18: [2]; 20-64: [3]; 65-68: [4]; 69-81: [4]; 82-94: [5]; 95-120: [6]; 127-131: [7]; 176-191: [8]; 192-202: [9]; 203-206: [10]; 207-228:
 [11]; 229-251: [12]; 252-269: [13]; 270-316: [14]

Mol Id	n	R	R ₁	R ₂	Exp. K_i (nM) hA ₁ R (% displ. 10 μ M)	Exp. K_i (nM) hA ₂ A (nM)	Exp. K_i (nM) hA ₂ B (IC ₅₀ nM)	Exp. K_i (nM) hA ₃ R (% displ. 10 μ M)
1	2	N(CH ₂ CH ₂ OH) ₂	H	-	123	0.12	> 10,000	> 10,000
2	3	NH ₂	H	-	2,160	0.22	> 10,000	> 10,000
3	3	CH ₂ NH ₂	H	-	576	0.13	> 10,000	> 10,000
4	3	NO ₂	H	-	1,026	1.00	> 10,000	> 10,000
5	2	NHCOCH ₃	H	-	419	4.80	> 10,000	> 10,000
6	3	N(CH ₂ CH ₂ OH) ₂	H	-	558	1.10	> 10,000	> 10,000
7	2	OCH ₂ COOEt	H	-	4,197	0.43	> 10,000	> 10,000
8	2	NH ₂	H	-	75	55	> 10,000	> 10,000
9	3	COOH	H	-	4,927	4.63	> 10,000	> 10,000
10	2	SO ₂ N(CH ₂ CH ₂ OH) ₂	H	-	2,495	0.80	> 10,000	> 10,000
11	2	SO ₂ N 	H	-	369	3.80	> 10,000	> 10,000
12	2	SO ₂ N(CH ₂ CH ₂ Cl) ₂	H	-	5,297	0.59	> 10,000	> 10,000
13	3	SO ₃ H	H	-	139	140	> 10,000	> 10,000
14	3	H	H	SO ₃ H	6,392	75	> 10,000	> 10,000
15	3	COOEt	H	H	1,793	5.48	> 10,000	> 10,000
16	3	COOH	COOH	H	9,399	120	> 10,000	> 10,000
17	3	COOH	H	H	3,599	59	> 10,000	> 10,000
18	3	H	H	-	350	1.20	> 10,000	> 10,000
19	-	H	4-OCH ₃ -Ph-NH-CO	-	448	520	350	0.14

MODELS 1, 2, 3, training set: continued

Mol Id	n	R	R ₁	R ₂	Exp. K_i (nM) hA ₁ R or (% displ. 10 μ M)	Exp. K_i (nM) hA ₂ A R	Exp. K_i (nM) hA ₂ B R or (IC ₅₀ nM)	Exp. K_i (nM) hA ₃ R or (% displ. 10 μ M)
20	-	CH ₃	3,4-OCH ₂ O-Ph-NH-CO	-	1,015	680	142	0.24
21	-	CH ₃	4-NO ₂ -Ph-NH-CO	-	1,115	695	180	0.43
22	-	CH ₃	4-CH ₃ -Ph-NH-CO	-	731	110	302	0.31
23	-	CH ₃	4-Br-Ph-NH-CO	-	600	100	181	0.46
24	-	CH ₃	2-OCH ₃ -Ph-NH-CO	-	450	180	223	0.70
25	-	CH ₃	3-OCH ₃ -Ph-NH-CO	-	500	160	251	0.80
26	-	CH ₃	2-Cl-Ph-NH-CO	-	400	200	101	0.91
27	-	C ₂ H ₅	3,4-Cl ₂ -Ph-NH-CO	-	526	352	40	3.00
28	-	C ₂ H ₅	3,4-OCH ₂ O-Ph-NH-CO	-	802	576	48	0.27
29	-	C ₂ H ₅	4-Br-Ph-NH-CO	-	500	40	296	0.37
30	-	C ₂ H ₅	4-F-Ph-NH-CO	-	602	60	246	0.86
31	-	C ₂ H ₅	4-CF ₃ -Ph-NH-CO	-	450	53	235	0.97
32	-	C ₂ H ₅	2-Cl-Ph-NH-CO	-	348	150	97	0.30
33	-	C ₂ H ₅	4-Cl-Ph-NH-CO	-	400	160	85	0.20
34	-	n-C ₃ H ₇	3,4-Cl ₂ -Ph-NH-CO	-	611	401	31	2.50
35	-	n-C ₃ H ₇	3,4-OCH ₂ O-Ph-NH-CO	-	842	667	53	0.30
36	-	n-C ₃ H ₇	4-NO ₂ -Ph-NH-CO	-	1,214	1,115	305	0.81
37	-	n-C ₃ H ₇	2-Cl-Ph-NH-CO	-	300	121	61	0.71
38	-	n-C ₃ H ₇	4-Cl-Ph-NH-CO	-	325	140	76	0.34

MODELS 1, 2, 3, training set: continued

Mol Id	n	R	R ₁	R ₂	Exp. K_i (nM) hA ₁ R or (% displ. 10 μ M)	Exp. K_i (nM) hA ₂ A R	Exp. K_i (nM) hA ₂ B R or (IC ₅₀ nM)	Exp. K_i (nM) hA ₃ R or (% displ. 10 μ M)
39	-	<i>n</i> -C ₄ H ₉	3,4-Cl ₂ -Ph-NH-CO	-	708	495	34	3.70
40	-	<i>n</i> -C ₄ H ₉	3,4-OCH ₂ O-Ph-NH-CO	-	410	376	78	0.50
41	-	<i>n</i> -C ₄ H ₉	4-NO ₂ -Ph-NH-CO	-	584	503	52	0.55
42	-	<i>n</i> -C ₄ H ₉	4-CH ₃ -Ph-NH-CO	-	281	24	161	0.21
43	-	<i>n</i> -C ₄ H ₉	4-Br-Ph-NH-CO	-	300	66	121	0.91
44	-	<i>n</i> -C ₄ H ₉	4-F-Ph-NH-CO	-	318	50	136	0.80
45	-	<i>n</i> -C ₄ H ₉	4-CF ₃ -Ph-NH-CO	-	350	31	130	0.72
46	-	<i>n</i> -C ₄ H ₉	2-OCH ₃ -Ph-NH-CO	-	200	91	125	0.57
47	-	<i>n</i> -C ₄ H ₉	3-OCH ₃ -Ph-NH-CO	-	248	95	150	0.6
48	-	<i>n</i> -C ₄ H ₉	2-Cl-Ph-NH-CO	-	280	100	78	0.86
49	-	CH ₃	Ph-CH ₂ -CO	-	702	423	165	0.81
50	-	C ₂ H ₅	Ph-CH ₂ -CO	-	714	335	161	1.03
51	-	<i>n</i> -C ₃ H ₇	Ph-CH ₂ -CO	-	351	306	143	1.01
52	-	<i>n</i> -C ₃ H ₇	Ph-NH-CO	-	176	62	241	0.15
53	-	<i>n</i> -C ₄ H ₉	Ph-NH-CO	-	409	148	50	0.21
54	-	CH ₃	4-SO ₃ H-Ph-NH-CO	-	>10,000	594	>10,000	25
55	-	<i>n</i> -C ₃ H ₇	4-SO ₃ H-Ph-NH-CO	-	>10,000	305	>10,000	30
56	-	<i>n</i> -C ₄ H ₉	4-SO ₃ H-Ph-NH-CO	-	>10,000	352	>10,000	47
57	-	CH ₃	4-OCH ₃ -Ph-NH-CO	-	1,097	1,390	261	0.20

MODELS 1, 2, 3, training set: continued

Mol Id	n	R	R ₁	R ₂	Exp. K_i (nM) hA ₁ R or (% displ. 10 μ M)	Exp. K_i (nM) hA ₂ A R	Exp. K_i (nM) hA ₂ B R or (IC ₅₀ nM)	Exp. K_i (nM) hA ₃ R or (% displ. 10 μ M)
58	-	CH ₃	3-Cl-Ph-NH-CO	-	350	1,195	205	0.40
59	-	C ₂ H ₅	3-Cl-Ph-NH-CO	-	249	180	150	1.60
60	-	<i>m</i> -C ₃ H ₇	4-OCH ₃ -Ph-NH-CO	-	1,197	140	2,056	0.80
61	-	<i>m</i> -C ₄ H ₉	4-OCH ₃ -Ph-NH-CO	-	296	80	303	0.32
62	-	<i>n</i> -C ₄ H ₉	3-Cl-Ph-NH-CO	-	245	95	70	0.60
63	-	-	4-OCH ₃ -Ph-NH-CO	-	7.60	9.40	22	0.14
64	-	-	3-Cl-Ph-NH-CO	-	8.10	9.50	30	0.19
65	-	Ph-CH ₂ -CH ₂	H	-	1	0.34	5.1	2,785
66	-	Ph-CH ₂ -CH ₂ -CH ₂	H	-	3	0.16	49	2,666
67	-	Ph-CH ₂ -CH ₂	3-Cl-Ph-NH-CO	-	251	105	90	13.28
68	-	Ph-CH ₂ -CH ₂ -CH ₂	3-Cl-Ph-NH-CO	-	1,500	200	226	42.65
69	-	(CH ₃) ₂ CH-CH ₂ -CH ₂	H	-	2.00	0.78	9.1	700
70	-	2,4,5-Br ₃ -Ph-CH ₂ -CH ₂	H	-	247	348	1,113	4,481
71	-	H	H	-	140	20	51	348
72	-	<i>t</i> -C ₄ H ₉	3-Cl-Ph-NH-CO	-	1,015	796	450	2.78
73	-	(CH ₃) ₂ CH-CH ₂ -CH ₂	3-Cl-Ph-NH-CO	-	100	115	61	40
75	-	(CH ₃) ₂ C=CH-CH ₂	4-OCH ₃ -Ph-NH-CO	-	300	1,025	58	40
76	-	(CH ₃) ₂ C=CH-CH ₂	3-Cl-Ph-NH-CO	-	352	110	81	55
77	-	2,4,5-Br ₃ -Ph-CH ₂ -CH ₂	4-OCH ₃ -Ph-NH-CO	-	255	2,030	>10,000	25

MODELS 1, 2, 3, training set: continued

Mol Id	n	R	R ₁	R ₂	Exp. K_i (nM) hA ₁ R (% displ. 10 μ M)	K_i or displ. (nM) hA ₂ A R	Exp. K_i (nM) hA ₂ B R (IC ₅₀ nM)	K_i or displ. (nM) hA ₃ R (% displ. 10 μ M)
78	-	2,4,5-Br ₃ -Ph-CH ₂ -CH ₂	3-Cl-Ph-NH-CO	-	389	3,600	1,816	71
79	-	2-(α -naphthyl)ethyl	4-OCH ₃ -Ph-NH-CO	-	2,466	3,260	>10,000	16
80	-	2-(α -naphthyl)ethyl	3-Cl-Ph-NH-CO	-	3,164	>10,000	>10,000	51
81	-	<i>n</i> -C ₄ H ₉	H	-	14	1,60	53	600
82	-	Ph-CH ₂ -CH ₂	CO-NH-CH-(CH ₃) ₂	-	60	>10,000	>10,000	9.0
83	-	Ph-CH ₂ -CH ₂	CO-NH-C(CH ₃) ₃	-	35	>10,000	>10,000	4.9
84	-	(CH ₃) ₂ CH-CH ₂ -CH ₂	CO-NH-CH-(CH ₃) ₂	-	389	>10,000	>10,000	65
85	-	Ph-CH ₂ -CH ₂ -CH ₂	CO-NH-CH-(CH ₃) ₂	-	849	>10,000	>10,000	55
86	-	Ph-CH ₂ -CH ₂	CO(CH ₂) ₃ NH ₃ +Cl-	-	1.6	54.5	27.6	65
87	-	Ph-CH ₂ -CH ₂	CO(CH ₂) ₃ NHCOOC(CH ₃) ₃	-	11.0	288	>10,000	1,480
88	-	(CH ₃) ₂ CH-CH ₂ -CH ₂	CO(CH ₂) ₃ NH ₃ +Cl-	-	9.9	241	52.8	90
89	-	(CH ₃) ₂ CH-CH ₂ -CH ₂	CO(CH ₂) ₃ NHCOOC(CH ₃) ₃	-	99	297	>10,000	2,488
90	-	Ph-CH ₂ -CH ₂	CH ₂ -CH ₂ -OH	-	5,485	84.9	>10,000	>10,000
91	-	Ph-CH ₂	CH ₂ -CH ₂ -OH	-	59.4	750	>10,000	>10,000
92	-	cC ₆ H ₁₁ -CH ₂ -CH ₂	CH ₂ -CH ₂ -OH	-	714	51.3	>10,000	>10,000
93	-	Ph-CH ₂ -CH ₂	CO-CH ₂ -NH ₃ +Cl-	-	4.5	182	70	163
94	-	Ph-CH ₂ -CH ₂	CO(CH ₂) ₂ NH ₃ +Cl-	-	2.5	50.0	37.6	80
95	-	CH ₃	Ph ₂ -CH-CO	-	139	216	116	0.25
96	-	CH ₃	Ph-Ph-CH ₂ -CO	-	1,460	193	(>10,000)	11.2

MODELS 1, 2, 3, training set: continued

Mol Id	n	R	R ₁	R ₂	Exp. K_i (nM) hA ₁ R (% displ. 10 μ M)	Exp. K_i (nM) hA ₂ A R	Exp. K_i (nM) hA ₂ B R (1C ₅₀ nM)	Exp. K_i (nM) hA ₃ R (% displ. 10 μ M)
97	-	CH ₂ -CH ₂ -CH(CH ₃) ₂	Ph ₂ -CH-CO	-	441	159	(>10,000)	5.86
98	-	CH ₂ -CH ₂ -CH(CH ₃) ₂	Ph-Ph-CH ₂ -CO	-	84.3	9.18	(8,960)	268
99	-	Ph-CH ₂ -CH ₂	Ph ₂ -CH-CO	-	115	53.1	(5,840)	6.49
100	-	Ph-CH ₂ -CH ₂	Ph-Ph-CH ₂ -CO	-	84.3	46.9	(11,100)	125
101	-	C ₂ H ₅	Ph-Ph-CH ₂ -CO	-	557	70.9	(>50,000)	14.2
102	-	CH ₃	4-OCH ₃ -Ph-CH ₂ -CO	-	133	62.5	(>10,000)	0.95
103	-	CH ₂ -CH ₂ -CH(CH ₃) ₂	β -naphthyl-CH ₂	-	140	15.7	(>10,000)	409
104	-	Ph-CH ₂ -CH ₂	β -naphthyl-CH ₂	-	57.5	42.1	(>50,000)	180
105	-	Ph-CH ₂ -CH ₂	2-thienyl-CH ₂	-	18.5	8.8	(>10,000)	330
106	-	Ph-CH ₂ -CH ₂	3-thienyl-CH ₂	-	22.8	12.9	(>10,000)	726
107	-	C ₂ H ₅	3-thienyl-CH ₂	-	154	23	(9,370)	765
108	-	C ₂ H ₅	2-thienyl-CH ₂	-	148	15.9	(15,000)	196
109	-	CH ₃	2-thienyl-CH ₂	-	444	56	(>10,000)	5.26
110	-	CH ₃	3-thienyl-CH ₂	-	461	31.3	(13,700)	1.25
111	-	CH ₃	β -naphthyl-CH ₂	-	190	77.5	(>10,000)	14.5
112	-	CH ₃	α -naphthyl-CH ₂	-	872	80.5	(>10,000)	3.47
113	-	<i>n</i> -C ₃ H ₇	β -naphthyl-CH ₂	-	160	7.99	(5,370)	95.9
114	-	<i>n</i> -C ₄ H ₉	3-thienyl-CH ₂	-	25.5	3.13	290	189
115	-	CH ₃	4-Cl-Ph-OCH ₂ -CO	-	963	39.3	(>10,000)	223








MODELS 1, 2, 3, training set: continued

Mol Id	n	R	R ₁	R ₂	Exp. K_i (nM) hA ₁ R (% displ. 10 μ M)	Exp. K_i (nM) hA ₂ A R	Exp. K_i (nM) hA ₂ B R (1C ₅₀ nM)	Exp. K_i (nM) hA ₃ R (% displ. 10 μ M)
116	-	CH ₂ -CH ₂ -CH(CH ₃) ₂	4-Cl-Ph-CH ₂ -CO	-	22.6	2.75	391	56.5
117	-	CH ₂ -CH ₂ -CH(CH ₃) ₂	4-CF ₃ -Ph-CH ₂ -CO	-	48.2	5.43	(5,480)	266
118	-	CH ₂ -CH ₂ -CH(CH ₃) ₂	4-F-Ph-CH ₂ -CO	-	51.8	3.69	(5,060)	116
119	-	CH ₂ -CH ₂ -CH(CH ₃) ₂	2,6-Cl ₂ -Ph-CH ₂ -CO	-	121	18.7	461	207
120	-	CH ₃	2,6-Cl ₂ -Ph-CH ₂ -CO	-	908	45.2	(6,450)	44.4
121	-	Ph-CH ₂ -CH ₂ -CH ₂	4-F-Ph-CH ₂ -CO	-	726	211	(>50,000)	58.5
122	-	Ph-CH ₂ -CH ₂ -CH ₂	Ph ₂ -CH-CO	-	307	326	(>10,000)	12.6
123	-	Ph-CH ₂ -CH ₂ -CH ₂	3-Cl-Ph-CH ₂ -CO	-	605	182	(50,000)	110
124	-	Ph-CH ₂ -CH ₂ -CH ₂	4-Cl-Ph-CH ₂ -CO	-	256	89.9	(>10,000)	30.5
125	-	Ph-CH ₂ -CH ₂ -CH ₂	2,6-Cl ₂ -Ph-CH ₂ -CO	-	786	186	(>10,000)	601
126	-	Ph-CH ₂ -CH ₂ -CH ₂	Ph-Ph-CH ₂ -CO	-	1,170	256	(>100,000)	410
127	-	CH ₃	Ph-CH ₂ O-Ph-CH ₂ -CO	-	2,640	233	(12,900)	29.4
128	-	CH ₃	2-CH ₃ -Ph-CH ₂ O-Ph-CH ₂ -CO	-	2,940	333	(>30,000)	6.80
129	-	CH ₃	2,6-Cl-Ph-CH ₂ O-Ph-CH ₂ -CO	-	1,680	281	(17,800)	35.3
130	-	CH ₃	3-Cl-Ph-CH ₂ O-Ph-CH ₂ -CO	-	1,980	278	(10,400)	9.93
131	-	CH ₃	4-CH ₃ -Ph-CH ₂ O-Ph-CH ₂ -CO	-	6,760	751	(>30,000)	24.4
132	-	CH ₃	2,5-CH ₃ -Ph-CH ₂ O-Ph-CH ₂ -CO	-	5,730	609	>30,000	11.7
133	-	CH ₃	3,4-CH ₃ -Ph-CH ₂ O-Ph-CH ₂ -CO	-	5,250	397	>30,000	10.8
134	-	CH ₃	3,5-CH ₃ -Ph-CH ₂ O-Ph-CH ₂ -CO	-	7,260	357	>30,000	11.9

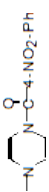
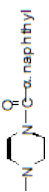
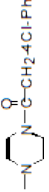
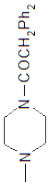

MODELS 1, 2, 3, training set: continued

Mol Id	n	R	R ₁	R ₂	Exp. K_i (nM) hA ₁ R (% displ. 10 μ M)	Exp. K_i (nM) hA ₂ A R	Exp. K_i (nM) hA ₂ B R (IC ₅₀ nM)	Exp. K_i (nM) hA ₃ R (% displ. 10 μ M)
135	-	CH ₂ -CH ₂ -CH(CH ₃) ₂	3,4-CH ₃ -Ph-CH ₂ O-Ph-CH ₂ -CO	-	396	37.2	>30,000	135
136	-	CH ₂ -CH ₂ -CH(CH ₃) ₂	3,5-CH ₃ -Ph-CH ₂ O-Ph-CH ₂ -CO	-	7,620	954	>30,000	936
137	-	CH ₂ -CH ₂ -CH(CH ₃) ₂	2,4,6-CH ₃ -Ph-CH ₂ O-Ph-CH ₂ -CO	-	975	83.7	>10,000	197
138	-	Ph-CH ₂ -CH ₂	2,5-CH ₃ -Ph-CH ₂ O-Ph-CH ₂ -CO	-	148	52.9	>10,000	61.3
139	-	Ph-CH ₂ -CH ₂	3,5-CH ₃ -Ph-CH ₂ O-Ph-CH ₂ -CO	-	348	156	>10,000	46.3
140	-	Ph-CH ₂ -CH ₂	2,4,6-CH ₃ -Ph-CH ₂ O-Ph-CH ₂ -CO	-	1,180	185	>10,000	79.5
141	-	Ph-CH ₂ -CH ₂ -CH ₂	2-CH ₃ -Ph-CH ₂ O-Ph-CH ₂ -CO	-	4,590	432	>10,000	278
142	-	Ph-CH ₂ -CH ₂ -CH ₂	4-CH ₃ -Ph-CH ₂ O-Ph-CH ₂ -CO	-	1,680	222	>10,000	88.8
143	-	Ph-CH ₂ -CH ₂ -CH ₂	3,4-CH ₃ -Ph-CH ₂ O-Ph-CH ₂ -CO	-	1,550	351	>10,000	279
144	-	Ph-CH ₂ -CH ₂ -CH ₂	2,5-CH ₃ -Ph-CH ₂ O-Ph-CH ₂ -CO	-	3,610	469	>10,000	271
145	-	Ph-CH ₂ -CH ₂ -CH ₂	3,4-CH ₃ -Ph-CH ₂ O-Ph-CH ₂ -CO	-	63.3	8.41	1,920	150
146	-	Ph-CH ₂ -CH ₂ -CH ₂	2,4,6-CH ₃ -Ph-CH ₂ O-Ph-CH ₂ -CO	-	2,190	759	>10,000	587
147	-	H	NH(CH ₂) ₂ Ph-OH	-	774	1.6	75	743
148	-	H	OPh	-	2,720	18.3	(3,420)	489
149	-	H	SCH ₃	-	1,730	96.5	(14,900)	2,576
150	-	Ph-CH ₂ -CO	SCH ₃	-	1,420	429	(>100,000)	4,200
151	-	4-CH ₃ -Ph-NH-CO	SCH ₃	-	3,440	742	(16,800)	2,200
152	-	CH ₃ (CH ₂) ₄ CO	OPh	-	2,900	189	(>100,000)	2,200
153	-	4-CH ₃ -Ph-NH-CO	OPh	-	35% inhib	214	(20,000)	750

MODELS 1, 2, 3, training set: continued

Mol Id	n	R	R ₁	R ₂	Exp. K_i (nM) hA ₁ R or (% displ. 10 μ M)	Exp. K_i (nM) hA _{2A} R	Exp. K_i (nM) hA _{2B} R or (IC ₅₀ nM)	Exp. K_i (nM) hA ₃ R or (% displ. 10 μ M)
154	-	Ph ₂ -CH-CO	N(CH ₃) ₂	-	951	4,090	(>100,000)	473
155	-	Ph-CH ₂ -CO	N(CH ₃) ₂	-	6,080	6,700	(>100,000)	1,050
156	-	CH ₃ (CH ₂) ₄ CO	N(CH ₃) ₂	-	1,570	10,300	(>100,000)	2,200
157	-	Ph-NH-CO	N(CH ₃) ₂	-	3,150	580	(>100,000)	311
158	-	4-CH ₃ -Ph-NH-CO	N(CH ₃) ₂	-	2,720	2,720	(>100,000)	9,600
159	-	H		-	>10,000	423	(>10,000)	(56.5)
160	-	NH(CH ₂) ₂ NH-Boc	CH ₂ -CH ₂ -NH-Boc	-	1,148	3,180	(>100,000)	788
161	-	H	CH ₂ -CH ₂ -NH-Boc	-	1,410	463	(30,000)	696
162	-	H	NH(CH ₂) ₂ -CH ₂ -NH-Boc	-	408	65	(20,600)	(77.3)
163	-	H		-	141	23	(>30,000)	(71.8)
164	-	H		-	5,630	20	(>30,000)	(52.7)
165	-	H		-	787	174	(>10,000)	(71.2)
166	-	H		-	255	9.5	(30,000)	(47.6)
167	-	H		-	303	252	(>100,000)	(49.6)
168	-	H		-	120	439	(>100,000)	(33.5)

MODELS 1, 2, 3, training set: continued

Mol Id	n	R	R ₁	R ₂	Exp. K_i (nM) hA ₁ R (% displ. 10 μ M)	Exp. K_i (nM) hA ₂ A (% displ. 10 μ M)	Exp. K_i (nM) hA ₂ B (IC ₅₀ nM)	Exp. K_i (nM) hA ₃ R (% displ. 10 μ M)
169	-	H		-	415	1,670	(>100,000)	(50.8)
170	-	H		-	>10,000	3,780	(>100,000)	(31.1)
171	-	H		-	626	83	(>100,000)	1,060
172	-	H		-	>10,000	578	(>100,000)	(40.6)
173	-	H		-	>10,000	404	(>100,000)	(32.2)
174	-	PhCO	OPh	-	>10,000	87	>10,000	(32.3)
175	-	PhCO	H	-	3,830	32	7,060	(67.1)
176	-	CH ₃	H	NHCH ₂ CH ₃	50	10	81	225
177	-	CH ₃	H	4-OCH ₃ -Ph-NH	260	>1,000	>1,000	>1,000
178	-	CH ₃	H	CH ₃ -N-piperazine	30	156	35	>1,000
179	-	CH ₃	H	S(CH ₂) ₂ CH ₃	9	2.1	69	224
180	-	CH ₃	4-OCH ₃ -Ph-NH-CO	CH ₃ -N-piperazine	316	>1,000	26	>1,000
181	-	SCH ₃	4-OCH ₃ -Ph-NH-CO	SCH ₃	70	3.1	24	212
182	-	SCH ₃	4-OCH ₃ -Ph-CH ₂ -CO	SCH ₃	80	15	45	>1,000
183	-	SCH ₃	4-isobutyl-Ph-CH ₂ -CO	SCH ₃	780	50	180	>1,000

MODELS 1, 2, 3, training set: continued

Mol Id	n	R	R ₁	R ₂	Exp. K_i (nM) hA ₁ R or (% displ. 10 μ M)	Exp. K_i (nM) hA ₂ A R	Exp. K_i (nM) hA ₂ B R or (IC ₅₀ nM)	Exp. K_i (nM) hA ₃ R or (% displ. 10 μ M)
184	-	SCH ₃	3,4-medioxy-Ph-CH ₂ -CO	SCH ₃	70	4.1	30	110
185	-	CH ₃	3,4-medioxy-Ph-CH ₂ -CO	NHCH ₂ CH ₃	136	61	65	183
186	-	N-methylpiperazine	-	-	>1,000	>1,000	>1,000	>1,000
187	-	CH ₃	H	4-OH-Ph-NH	666	>1,000	>1,000	308
188	-	2-OEt	-	-	>1,000	>1,000	>1,000	348
189	-	H	-	-	235	>1,000	>1,000	>1,000
190	-	4-OCH ₂ CO ₂ Et	-	-	>1,000	>1,000	>1,000	>1,000
191	-	4-OCH ₂ CO ₂ H	-	-	>1,000	>1,000	>1,000	>1,000
192	-	CH ₃	4-F-Ph-NH-CO-CH ₂	-	500	>1,000	14	>1,000
193	-	CH ₃	4-Cl-Ph-NH-CO-CH ₂	-	>1,000	>1,000	7	>1,000
194	-	CH ₃	4-Br-Ph-NH-CO-CH ₂	-	>1,000	>1,000	35	>1,000
195	-	CH ₃	4-I-Ph-NH-CO-CH ₂	-	1,000	>1,000	28	>1,000
196	-	CH ₃	4- <i>sec</i> -butyl-Ph-NH-CO-CH ₂	-	>1,000	>1,000	>1,000	>1,000
197	-	CH ₃	3-OCH ₃ -Ph-NH-CO-CH ₂	-	480	>1,000	65	>1,000
198	-	CH ₃	3,4-dimethoxy-Ph-NH-CO-CH ₂	-	>1,000	>1,000	60	>1,000
199	-	CH ₃	1-naphthyl-NH-CO-CH ₂	-	>1,000	>1,000	>1,000	>1,000
200	-	Cl	-	-	>1,000	>1,000	222	>1,000
201	-	Br	-	-	>1,000	>1,000	250	>1,000
202	-	I	-	-	>1,000	>1,000	>1000	>1,000

MODELS 1, 2, 3, training set: continued

Mol Id	n	R	R ₁	R ₂	Exp. K_i (nM) hA ₁ R or (% displ. 10 μ M)	Exp. K_i (nM) hA ₂ A R	Exp. K_i (nM) hA ₂ B R or (IC ₅₀ nM)	Exp. K_i (nM) hA ₃ R or (% displ. 10 μ M)
203	-	propyl	propyl	cyclopentyl	3	60	51	243
204	-	propyl	(S)-1-phenylethyl	biphenyl	>10,000	>10,000	>10,000	455
205	-	butyl	3-hydroxypropyl	3-noradamantyl	0.7	980	187	2,300
206	-	propyl	3-noradamantyl	-	13.8	25,000	22,300	188
207	-	H	CH ₃	-	>1,000	>1,000	>1,000	0.8
208	-	H	CH ₂ CH ₃	-	>1,000	>1,000	>1,000	15
209	-	H	CH(CH ₃) ₂	-	460	>1,000	>1,000	31
210	-	H	cyclohexyl	-	>1,000	>1,000	>1,000	555
211	-	CH ₃	CH ₂ CH ₃	-	>1,000	>1,000	>1,000	60
212	-	H	4-OCH ₃ -Ph	-	>1,000	>1,000	>1,000	55
213	-	H	4-Ph-Ph	-	>1,000	>1,000	>1,000	>1,000
214	-	H	4-F-Ph	-	>1,000	>1,000	>1,000	22
215	-	H	CH ₃	-	>1,000	>1,000	400	8
216	-	H	CH ₂ CH ₃	-	>1,000	>1,000	>1,000	3.45
217	-	CH ₃	CH ₃	-	>1,000	>1,000	>1,000	80
218	-	H	CH ₂ -Ph	-	>1,000	>1,000	>1,000	148
219	-	CH ₂ -CH ₂ -OH	CH ₂ -Ph	-	>1,000	>1,000	>1,000	38
220	-	CH ₂ -CH=CH ₂	CH ₂ -Ph	-	>1,000	>1,000	>1,000	5.13

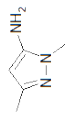
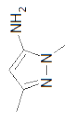
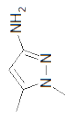
MODELS 1, 2, 3, training set: continued

Mol Id	n	R	R ₁	R ₂	Exp. K_i (nM) hA ₁ R (% displ. 10 μ M)	Exp. K_i (nM) hA ₂ A (nM)	Exp. K_i (nM) hA ₂ B (IC ₅₀ nM)	Exp. K_i (nM) hA ₃ R (% displ. 10 μ M)
221	-	CH ₂ -C≡CH	CH ₂ -Ph	-	>1,000	>1,000	>1,000	15
222	-	CH ₂ -CH ₂ -N(CH ₃) ₂	CH ₂ -Ph	-	>1,000	>1,000	>1,000	98
223	-	H	1-hydroxy-propyl-2-yl	-	>1,000	>1,000	>1,000	>1,000
224	-	H	CH ₂ CH ₃	-	>1,000	>1,000	>1,000	>1,000
225	-	CH ₂ -CH ₂ -CH ₃	Ph	-	373	>1,000	>1,000	99
226	-	CH ₂ -CH(CH ₃) ₂	Ph	-	476	>1,000	>1,000	144
227	-	CH ₃	Ph	-	>1,000	>1,000	>1,000	>1,000
228	-	CH ₃	CH ₃	-	>1,000	>1,000	>1,000	>1,000
229	-	Ph	-	-	102	1,500	22	1,200
230	-	2-Cl-Ph	-	-	2,500	2,500	19	490
231	-	4-CN-Ph	-	-	1,300	>5,000	18	950
233	-	3-Cl-Ph	-	-	652	>5,000	32	490
234	-	3-CF ₃ -Ph	-	-	3,500	>5,000	42	960
235	-	2-CF ₃ -Ph	-	-	530	>5,000	74	340
236	-	2-OCH ₃ -Ph	-	-	1,900	1,600	82	410
237	-	3-OCH ₃ -Ph	-	-	2,200	>5,000	82	600
238	-	4-CH ₃ -Ph	-	-	2,600	3,000	54	8,700
239	-	CH ₂ -CH ₂ -CH ₃	CH ₂ -CH ₂ -CH ₃	Ph	1,000	1,800	21	630

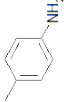
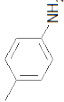
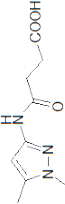
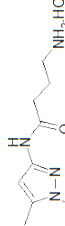
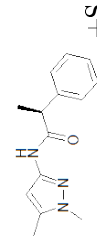
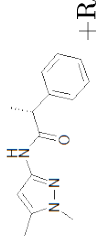
MODELS 1, 2, 3, training set: continued

Mol Id	n	R	R ₁	R ₂	Exp. K_i (nM) hA ₁ R (% displ. 10 μ M)	Exp. K_i (nM) hA ₂ A R	Exp. K_i (nM) hA ₂ B R (IC ₅₀ nM)	Exp. K_i (nM) hA ₃ R (% displ. 10 μ M)
240	-	CH ₂ -CH ₂ -CH ₃	CH ₂ -CH ₂ -CH ₃	4-Cl-Ph	2,900	>5,000	39	>8,300
241	-	CH ₂ -CH ₂ -CH ₃	CH ₂ -CH ₂ -CH ₃	4-CF ₃ -Ph	>6,000	>5,000	21	1,300
242	-	CH ₂ -CH ₂ -CH ₃	CH ₂ -CH ₂ -CH ₃	3-Cl-Ph	3,000	>5,000	134	3,000
243	-	CH ₂ -CH ₂ -CH ₃	CH ₂ -CH ₂ -CH ₃	3-CF ₃ -Ph	>6,000	>5,000	64	1,800
244	-	CH ₂ -CH ₂ -CH ₃	CH ₂ -CH ₂ -CH ₃	4-OCH ₃ -Ph	3,900	>5,000	123	2,300
245	-	CH ₂ -CH ₂ -CH ₃	CH ₂ -CH ₂ -CH ₃	2-OCH ₃ -Ph	930	3,400	83	>8,300
246	-	CH ₂ -CH ₂ -CH ₃	CH ₂ -CH ₂ -CH ₃	Ph	1,500	420	14	1,800
247	-	CH ₂ -CH ₂ -CH ₃	CH ₂ -CH ₂ -CH ₃	4-Cl-Ph	540	1,200	82	3,300
248	-	CH ₂ -CH ₂ -CH ₃	CH ₂ -CH ₂ -CH ₃	4-CF ₃ -Ph	2,300	1,700	48	1,200
249	-	CH ₂ -CH ₂ -CH ₃	CH ₂ -CH ₂ -CH ₃	3-OCH ₃ -Ph	3,500	>5,000	131	>5,000
250	-	CH ₂ -CH ₂ -CH ₃	CH ₂ -CH ₂ -CH ₃	4-OCH ₃ -Ph	2,000	>5,000	63	3,000
251	-	CH ₂ -CH ₂ -CH ₃	CH ₂ -CH ₂ -CH ₃	2-OCH ₃ -Ph	3,300	>5,000	195	4,900
252	-	CH ₂ -CH ₂ -CH ₃	CH ₂ -CH ₂ -CH ₃	H	76	290	11	170
253	-	CH ₂ -CH ₂ -CH ₃	CH ₂ -CH ₂ -CH ₃	F	160	400	14	140
254	-	CH ₂ -CH ₃	H	H	5,600	3,800	37	1,500
255	-	<i>n</i> -C ₄ H ₉	H	H	5,100	3,400	34	110
256	-	<i>i</i> -C ₄ H ₉	H	H	>6,000	>5,000	13	1,100
257	-	cyclopropyl methyl	H	H	3,290	2,760	6	180
258	-	CH ₂ -CH ₃	H	F	>6,000	>5,000	73	2,117

MODELS 1, 2, 3, training set: continued

Mol Id	n	R	R ₁	R ₂	Exp. K_i (nM) hA ₁ R or (% displ. 10 μ M)	Exp. K_i (nM) hA ₂ A R	Exp. K_i (nM) hA ₂ B R or (IC ₅₀ nM)	Exp. K_i (nM) hA ₃ R or (% displ. 10 μ M)
259	-	CH ₂ -CH ₃	H	CF ₃	6,000	>5,000	6	9,000
260	-	CH ₂ -CH ₂ -CH ₃	H	F	1,800	1,730	11	160
261	-	CH ₂ -CH ₂ -CH ₃	H	CF ₃	>6,000	>5,000	8	700
262	-	<i>n</i> -C ₄ H ₉	H	F	4,200	>5,000	18	270
263	-	<i>n</i> -C ₄ H ₉	H	CF ₃	>6,000	>5,000	30	>9,000
264	-	<i>i</i> -C ₄ H ₉	H	F	>6,000	>5,000	24	1,600
265	-	CH ₂ -CH ₂ -CH ₃	H	4-Cl-Ph	>6,000	>5,000	7	>9,000
266	-	CH ₂ -CH ₂ -CH ₃	H	4-CF ₃ -Ph	1,400	>5,000	15	>9,000
267	-	cyclopropyl methyl	H	4-Cl-Ph	>6,000	>5,000	48	>9,000
268	-	<i>i</i> -C ₄ H ₉	H	Ph	>6,000	>5,000	12	>9,000
269	-	cyclopropyl methyl	H	4-CF ₃ -Ph	>6,000	>5,000	15	>9,000
270	-	CH ₃	H		>1,000	>1,000	175	>1,000
271	-	<i>n</i> -C ₃ H ₇	H		140	>1,000	58	>1,000
272	-	<i>n</i> -C ₃ H ₇	H		201	>1,000	235	>1,000

MODELS 1, 2, 3, training set: continued

Mol Id	n	R	R ₁	R ₂	Exp. K _i (nM) hA ₁ R or (% displ. 10 μM)	Exp. K _i (nM) hA ₂ A R	Exp. K _i (nM) hA ₂ B R or (IC ₅₀ nM)	Exp. K _i (nM) hA ₃ R or (% displ. 10 μM)
273	-	<i>n</i> -C ₃ H ₇	H		65	>1,000	9	>1,000
274	-	<i>n</i> -C ₃ H ₇	CH ₃		>1,000	>1,000	516	>1,000
275	-	<i>n</i> -C ₃ H ₇	H		>1,000	>1,000	1,000	>1,000
276	-	<i>n</i> -C ₃ H ₇	H		548	>1,000	2,065	>1,000
277	-	<i>n</i> -C ₃ H ₇	H		>1,000	>1,000	1,000	>1,000
278	-	<i>n</i> -C ₃ H ₇	H		>1,000	>1,000	1,000	>1,000
279	-	H	H	H	900	>1,000	35	>1,000
280	-	OCH ₃	H	H	>1,000	>1,000	96	>1,000

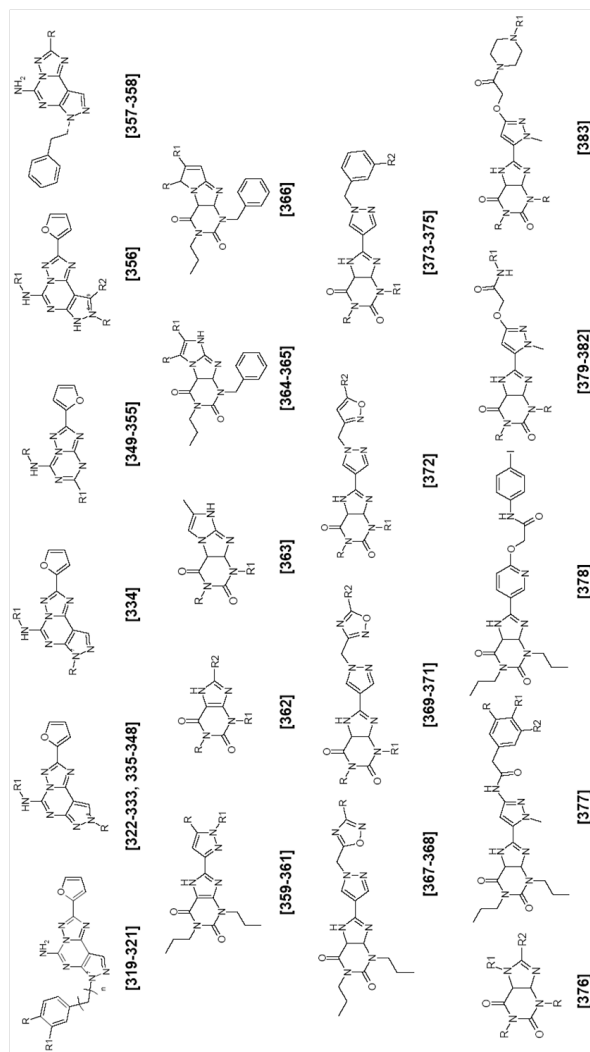
MODELS 1, 2, 3, training set: continued

Mol Id	n	R	R ₁	R ₂	Exp. K_i (nM) hA ₁ R (% displ. 10 μ M)	Exp. K_i (nM) hA ₂ A (nM)	Exp. K_i (nM) hA ₂ B (IC ₅₀ nM)	Exp. K_i (nM) hA ₃ R (% displ. 10 μ M)
281	-	H	CH ₂ CH(CH ₃) ₂	H	>1,000	>1,000	>1,000	>1,000
282	-	H	NO ₂	H	>1,000	>1,000	78	>1,000
283	-	H	OH	H	>1,000	>1,000	103	>1,000
284	-	H	F	H	200	>1,000	88	>1,000
285	-	H	OCH ₃	H	850	>1,000	100	>1,000
286	-	Cl	H	H	4,481	>1,000	160	>1,000
287	-	F	H	H	3,227	>1,000	50	>1,000
288	-	H	N(CH ₃) ₂	H	>1,000	>1,000	1,628	>1,000
289	-	H	Cl	H	520	>1,000	28	>1,000
290	-	H	OCH ₂ - <i>m</i> -CF ₃ -C ₆ H ₅	H	100	>1,000	90	>1,000
291	-	H	OCH ₂ - <i>p</i> -NO ₂ -C ₆ H ₅	H	163	>1,000	111	>1,000
292	-	H	CF ₃	H	746	>1,000	130	>1,000
293	-	OCH ₃	OCH ₃	OCH ₃	>1,000	>1,000	>1,000	>1,000
294	-		OCH ₂ O	H	566	>1,000	18	>1,000
295	-	-	-	-	>1,000	>1,000	569	>1,000
296	-	<i>n</i> -C ₃ H ₇	H	-	>1,000	>1,000	649	>1,000
297	-	<i>n</i> -C ₃ H ₇	OCH ₃	-	1,725	>1,000	95	>1,000
298	-	-	-	-	55	>1,000	34	>1,000
299	-	H	N(CH ₃) ₂	-	2,410	>1,000	59	>1,000

MODELS 1, 2, 3, training set: continued

Mol Id	n	R	R ₁	R ₂	Exp. K_i (nM) hA ₁ R (% displ. 10 μ M)	Exp. K_i (nM) hA ₂ A R	Exp. K_i (nM) hA ₂ B R (IC ₅₀ nM)	Exp. K_i (nM) hA ₃ R (% displ. 10 μ M)
300	-	OCH ₃	H	-	1,993	>1,000	90	>1,000
301	-	Cl	H	-	1,361	>1,000	110	>1,000
302	-	<i>n</i> -C ₃ H ₇	4-F-Ph	-	65	>1,000	12	>1,000
303	-	<i>n</i> -C ₃ H ₇	4-Br-Ph	-	150	>1,000	20	>1,000
304	-	<i>n</i> -C ₃ H ₇	3,4-(methylenedioxy)phenyl	-	200	>1,000	5.5	>1,000
305	-	<i>n</i> -C ₃ H ₇	5-methylpyridin-2-yl	-	>1,000	>1,000	1,012	>1,000
306	-	<i>n</i> -C ₃ H ₇	4- <i>s</i> -Bu-phenyl	-	1,005	>1,000	74	>1,000
307	-	<i>n</i> -C ₃ H ₇	4-tolyl	-	79	>1,000	19	>1,000
308	-	<i>n</i> -C ₃ H ₇	pyridin-4-yl	-	955	>1,000	41	>1,000
309	-	<i>i</i> -C ₄ H ₉	4-F-Ph	-	467	>1,000	303	>1,000
310	-	<i>i</i> -C ₄ H ₉	4-Br-Ph	-	2,427	>1,000	132	>1,000
311	-	<i>n</i> -C ₃ H ₇	4-F-Ph	-	181	>1,000	185	>1,000
312	-	<i>i</i> -C ₄ H ₉	4-Br-Ph	-	49	>1,000	66	>1,000
313	-	<i>i</i> -C ₄ H ₉	4-F-Ph	-	72	>1,000	207	>1,000
314	-	<i>n</i> -C ₃ H ₇	CH ₃	-	>1,000	>1,000	122	>1,000
315	-	H	F	-	>1,000	>1,000	70	>1,000
316	-	H	OCH ₃	-	>1,000	>1,000	53	>1,000
317	-	4-OCH ₃ -Ph	-	-	1,221	>1,000	>10,000	1,760
318	-	4-Cl-Ph	-	-	187	67.3	>10,000	1,490


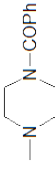
E.2 MODELS 1, 2, 3, validation set.




319-321: [1]; 322-329: [3]; 330: [4]; 331-333: [5]; 334-336: [6]; 337-343: [7]; 356-358: [9]; 359-361: [10]; 362: [11]; 363-366: [12]; 367-372:
 [13]; 373-375: [14]; 376-383: [15]

Mol Id	n	R	R ₁	R ₂	Exp. K_i (nM) hA ₁ R or (% displ. 10 μ M)	Exp. K_i (nM) hA _{2A} R	Exp. K_i (nM) hA _{2B} R (IC ₅₀ nM) or	Exp. K_i (nM) hA ₃ R or (% displ. 10 μ M)
319	2	SO ₃ H	H	-	190	100	>10,000	>10,000
320	2	SO ₂ NH ₂	H	-	2,630	1.31	>10,000	>10,000
321	2	SO ₂ NHCH ₂ COOH	H	-	9,330	50.0	>10,000	>10,000
322	-	CH ₃	4-CF ₃ -Ph-NH-CO	-	750	140	286	0.74
323	-	CH ₃	4-Cl-Ph-NH-CO	-	430	180	128	0.29
324	-	C ₂ H ₅	4-NO ₂ -Ph-NH-CO	-	1,134	614	226	0.65
325	-	C ₂ H ₅	4-CH ₃ -Ph-NH-CO	-	262	30	132	0.14
326	-	<i>n</i> -C ₃ H ₇	4-CF ₃ -Ph-NH-CO	-	400	34	136	0.51
327	-	<i>n</i> -C ₄ H ₉	Ph-CH ₂ -CO	-	602	400	101	1.11
328	-	C ₂ H ₅	Ph-NH-CO	-	326	40	88	0.18
329	-	C ₂ H ₅	4-OCH ₃ -Ph-NH-CO	-	1,026	1,040	245	0.60
330	-	C ₂ H ₅	H	-	5.00	1.95	65	3,759
331	-	2-(α -naphthyl)ethyl	H	-	348	896	4,294	3,416
332	-	<i>t</i> -C ₄ H ₉	H	-	175	45	400	1,149
333	-	H	3-Cl-Ph-NH-CO	-	238	248	70	0.50
334	-	cC ₆ H ₁₁ -CH ₂ -CH ₂	H	-	51.0	0.58	>10,000	>10,000
335	-	CH ₃ -CH ₂ -CH ₂	CO-NH-C(CH ₃) ₃	-	500	>10,000	>10,000	15
336	-	Ph-CH ₂ -CH ₂ -CH ₂	CO-NH-C(CH ₃) ₃	-	876	>10,000	>10,000	65
337	-	CH ₂ -CH ₂ -CH ₃	Ph ₂ -CH-CO	-	80.2	46.5	157	0.93
338	-	CH ₂ -CH ₃	α -naphthyl-CH ₂	-	596	74.6	(>50,000)	3.05
339	-	CH ₂ -CH ₃	β -naphthyl-CH ₂	-	185	18.4	(10,370)	36.3

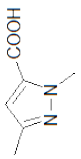
MODELS 1, 2, 3, validation set: continued

Mol Id	n	R	R ₁	R ₂	Exp. K_i (nM) hA ₁ R or (% displ. 10 μ M)	Exp. K_i (nM) hA _{2A} R	Exp. K_i (nM) hA _{2B} R or (IC ₅₀ nM)	Exp. K_i (nM) hA ₃ R or (% displ. 10 μ M)
340	-	CH ₂ -CH ₂ -CH ₃	α -naphthyl-CH ₂	-	388	38.7	(12,400)	17.3
341	-	Ph-CH ₂ -CH ₂	3-Cl-Ph-CH ₂ -CO	-	13.2	5.75	(9,100)	273
342	-	CH ₃	4-F-Ph-CH ₂ -CO	-	741	54.1	(13,000)	0.97
343	-	Ph-CH ₂ -CH ₂ -CH ₂	4-Cl-Ph-CH ₂ -CO	-	71.1	27.2	(>10,000)	400
344	-	CH ₂ -CH ₂ -CH(CH ₃) ₂	2-CH ₃ -Ph-CH ₂ O-Ph-CH ₂ -CO	-	906	52.3	>30,000	358
345	-	CH ₂ -CH ₂ -CH(CH ₃) ₂	3-CH ₃ -Ph-CH ₂ O-Ph-CH ₂ -CO	-	564	29.9	>30,000	184
346	-	Ph-CH ₂ -CH ₂	3,4-CH ₃ -Ph-CH ₂ O-Ph-CH ₂ -CO	-	260	109	>10,000	57.8
347	-	Ph-CH ₂ -CH ₂ -CH ₂	3-CH ₃ -Ph-CH ₂ O-Ph-CH ₂ -CO	-	582	97.3	>10,000	148
348	-	Ph-CH ₂ -CH ₂ -CH ₂	3,5-CH ₃ -Ph-CH ₂ O-Ph-CH ₂ -CO	-	2,190	248	>10,000	79.3
349	-	CO-CH-Ph ₂	OPh	-	<10% inhib	1,880	(>100,000)	4,400
350	-	CO-CH ₂ -Ph	OPh	-	7,260	136	(>100,000)	1,020
351	-	4-OCH ₃ -Ph-CH ₂ -CO	OPh	-	<12% inhib	892	(>100,000)	5,200
352	-	CO-NH-Ph	OPh	-	<17% inhib	38.9	(8,870)	633
353	-	H		-	>10,000	107	(100,000)	(74.9)
354	-	H		-	>10,000	2,140	(>100,000)	(22.5)

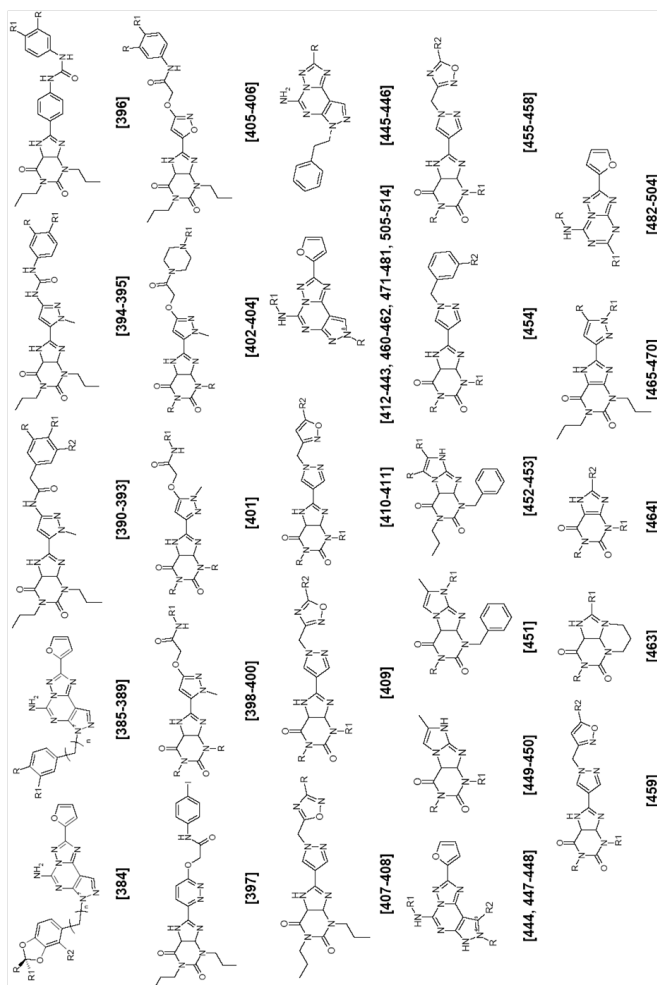
MODELS 1, 2, 3, validation set: continued

Mol Id	n	R	R ₁	R ₂	Exp. K_i (nM) hA ₁ R or (% displ. 10 μ M)	Exp. K_i (nM) hA ₂ A R	Exp. K_i (nM) hA ₂ B R (IC ₅₀ nM)	Exp. K_i (nM) hA ₃ R or (% displ. 10 μ M)
355	-	H		-	> 10,000	192	(> 100,000)	(27.1)
356	-	Ph-CH ₂ -CH ₂ -CH ₂	H	SCH ₃	175	22	31	> 1,000
357	-	N-methylpiperazine	-	-	> 1,000	> 1,000	> 1,000	> 1,000
358	-	morpholine	-	-	> 1,000	> 1,000	> 1,000	> 1,000
359	-	CH ₃	Ph-NH-CO-CH ₂	-	350	> 1,000	15	> 1,000
360	-	CH ₃	4-tolyl-NH-CO-CH ₂	-	> 1,000	> 1,000	56	> 1,000
361	-	CH ₃	4-OCH ₃ -Ph-NH-CO-CH ₂	-	> 1,000	> 1,000	12	> 1,000
362	-	propyl	Ph	cyclopentyl	7.1	1,200	625	395
363	-	CH ₂ -CH ₂ -CH ₂ -OH	CH ₂ -Ph	-	> 1,000	> 1,000	> 1,000	110
364	-	CH ₃	CH ₃	-	> 1,000	> 1,000	> 1,000	36
365	-	H	Ph	-	> 1,000	> 1,000	> 1,000	115
366	-	H	Ph	-	> 1,000	> 1,000	> 1,000	200
367	-	4-OCH ₃ -Ph	-	-	3,000	> 5,000	106	1,500
368	-	3-CH ₃ -Ph	-	-	1,400	> 5,000	128	970
369	-	CH ₂ -CH ₂ -CH ₃	CH ₂ -CH ₂ -CH ₃	2-Cl-Ph	630	2,500	38	330
370	-	CH ₂ -CH ₂ -CH ₃	CH ₂ -CH ₂ -CH ₃	4-CN-Ph	570	> 5,000	14	640
371	-	CH ₂ -CH ₂ -CH ₃	CH ₂ -CH ₂ -CH ₃	2-CF ₃ -Ph	1,900	> 5,000	100	620
372	-	CH ₂ -CH ₂ -CH ₃	CH ₂ -CH ₂ -CH ₃	3-Cl-Ph	4,600	> 5,000	163	4,400

MODELS 1, 2, 3, validation set: continued

Mol Id	n	R	R ₁	R ₂	Exp. K_i (nM) hA ₁ R or (% displ. 10 μ M)	Exp. K_i (nM) hA ₂ A R	Exp. K_i (nM) hA ₂ B R or (IC ₅₀ nM)	Exp. K_i (nM) hA ₃ R or (% displ. 10 μ M)
373	-	CH ₂ -CH ₂ -CH ₃	H	H	1,600	>5,000	13	120
374	-	<i>i</i> -C ₄ H ₉	H	CF ₃	>6,000	>5,000	28	>9,000
375	-	CH ₂ -CH ₂ -CH ₃	CH ₂ -CH ₂ -CH ₃	F	170	230	14	58
376	-	<i>n</i> -C ₃ H ₇	H		>1,000	>1,000	>1,000	>1,000
377	-	H	OCH ₂ -Ph	H	>1,000	>1,000	56	>1,000
378	-	-	-	-	>1,000	>1,000	108	>1,000
379	-	<i>n</i> -C ₃ H ₇	4-acetylphenyl	-	>1,000	>1,000	86	>1,000
380	-	<i>n</i> -C ₃ H ₇	4-N-morpholinylphenyl	-	>1,000	>1,000	86	>1,000
381	-	<i>n</i> -C ₃ H ₇	4-carboxyphenyl	-	>1,000	>1,000	36	>1,000
382	-	<i>n</i> -C ₃ H ₇	3,4-dimethoxyphenyl	-	700	>1,000	10	>1,000
383	-	<i>n</i> -C ₃ H ₇	benzyl	-	810	>1,000	85	>1,000

E.3 MODELS 1, 2, 3, internal test set.



384-388: [1]; 389: [2]; 390-406: [15]; 407-411: [13]; 412-423: [7]; 424-436: [3]; 437-439: [4]; 440-443: [5]; 444-448: [9]; 449-453: [12]; 454-459: [14]; 460-462: [6]; 463-464: [11]; 465-470: [10]; 505-514: [16]

Mol Id	n	R	R ₁	R ₂	Exp. K_i (nM) hA ₁ R or (% displ. 10 μ M)	Exp. K_i (nM) hA ₂ A R	Exp. K_i (nM) hA ₂ B R (IC ₅₀ nM) or	Exp. K_i (nM) hA ₃ R or (% displ. 10 μ M)
384	3	CH ₂ OH	CH ₂ OH	H	432	0.19	>10,000	>10,000
385	3	CN	H	-	6,496	86	>10,000	>10,000
386	3	COOEt	H	-	4,494	4.0	>10,000	>10,000
387	3	C(NOH)NH ₂	H	-	60	6.0	>10,000	>10,000
388	3	(NH)NH ₂	H	-	4,965	4.4	>10,000	>10,000
389	2	H	H	-	549	1.1	>10,000	>10,000
390	-	OCH ₃	OCH ₃	H	>1,000	>1,000	38	>1,000
391	-	H	OCH ₂ - <i>o</i> -CF ₃ -Ph	H	56	>1,000	13	>1,000
392	-	F	F	H	1,898	>1,000	78	>1,000
393	-	OCH ₃	OH	H	>1,000	>1,000	342	>1,000
394	-	Cl	H	-	448	>1,000	39	>1,000
395	-	H	OCH ₃	-	1,440	>1,000	81	>1,000
396	-	OCH ₃	H	-	1,175	>1,000	58	>1,000
397	-	-	-	-	>1,000	>1,000	>1,000	>1,000
398	-	<i>n</i> -C ₃ H ₇	4-(ethoxycarbonyl)phenyl	-	>1,000	>1,000	32	>1,000
399	-	<i>n</i> -C ₃ H ₇	3,4-Cl ₂ -Ph	-	300	>1,000	16	>1,000
400	-	<i>n</i> -C ₃ H ₇	3,4-OCH ₃ -Ph	-	>1,000	>1,000	12	>1,000
401	-	<i>n</i> -C ₃ H ₇	4-Br-Ph	-	168	>1,000	93	>1,000
402	-	<i>n</i> -C ₃ H ₇	Ph	-	250	>1,000	15	>1,000
403	-	<i>n</i> -C ₃ H ₇	4-F-Ph	-	>1,000	>1,000	55	>1,000
404	-	CH ₂ CH=CH ₂	Ph	-	>1,000	>1,000	24	>1,000

MODELS 1, 2, 3, internal test set: continued

Mol Id	n	R	R ₁	R ₂	Exp. K_i (nM) hA ₁ R or (% displ. 10 μ M)	Exp. K_i (nM) hA ₂ A R	Exp. K_i (nM) hA ₂ B R or (IC ₅₀ nM)	Exp. K_i (nM) hA ₃ R or (% displ. 10 μ M)
405	-		OCH ₂ O	-	>1,000	>1,000	47	>1,000
406	-	OCH ₃	OCH ₃	-	>1,000	>1,000	51	>1,000
407	-	-	-	Ph	370	1,100	1	400
408	-	-	-	4-Cl-Ph	740	1,800	21	400
409	-	<i>n</i> -C ₃ H ₇	<i>n</i> -C ₃ H ₇	3-OCH ₃ -Ph	810	>5,000	136	2,000
410	-	<i>n</i> -C ₃ H ₇	<i>n</i> -C ₃ H ₇	2-Cl-Ph	>6,000	2,700	111	2,300
411	-	<i>n</i> -C ₃ H ₇	<i>n</i> -C ₃ H ₇	2-CF ₃ -Ph	3,700	5,000	121	4,200
412	-	CH ₂ -CH ₂ -CH ₂ -CH ₃	Ph ₂ -CH-CO	-	129	114	250	1.2
413	-	CH ₂ -CH ₂ -CH ₂ -CH ₃	Ph-Ph-CH ₂ -CO	-	381	39.6	(8,630)	189
414	-	CH ₃	3-Cl-Ph-CH ₂ -CO	-	410	30.5	(8,960)	1.94
415	-	CH ₂ -CH ₂ -CH ₃	Ph-Ph-CH ₂ -CO	-	622	52.5	(>10,000)	65.4
416	-	CH ₃	4-Cl-Ph-CH ₂ -CO	-	1,850	156	(>50,000)	12.7
417	-	CH ₂ -CH ₂ -CH ₂ -CH ₃	α -naphthyl-CH ₂ -CO	-	170	11.7	(7,370)	100
418	-	CH ₃	4-CF ₃ -Ph-CH ₂ -CO	-	768	75.9	(>10,000)	1.22
419	-	CH ₂ -CH ₂ -CH ₂ -CH ₃	2-thienyl-CH ₂ -CO	-	70.1	4.15	580	99.8
420	-	CH ₂ -CH ₂ -CH(CH ₃) ₂	3-Cl-Ph-CH ₂ -CO	-	13.7	1.86	444	273
421	-	Ph-CH ₂ -CH ₂ -CH ₂	α -naphthyl-CH ₂ -CO	-	241	73.6	(>50,000)	717
423	-	Ph-CH ₂	Ph-CH ₂ -CH ₂ -CO	-	128	43.9	(>3,000)	5.49
424	-	CH ₃	3,4-Cl ₂ -Ph-NH-CO	-	392	143	116	3.4

MODELS 1, 2, 3, internal test set: continued

Mol Id	n	R	R ₁	R ₂	Exp. K _i (nM) hA ₁ R or (% displ. 10 μM)	Exp. K _i (nM) hA ₂ A R	Exp. K _i (nM) hA ₂ B R or (IC ₅₀ nM)	Exp. K _i (nM) hA ₃ R or (% displ. 10 μM)
425	-	CH ₃	4-F-Ph-NH-CO	-	700	120	226	0.34
426	-	C ₂ H ₅	2-OCH ₃ -Ph-NH-CO	-	300	133	196	0.56
427	-	C ₂ H ₅	3-OCH ₃ -Ph-NH-CO	-	400	140	215	0.86
428	-	<i>n</i> -C ₃ H ₇	4-CH ₃ -Ph-NH-CO	-	302	12	46	0.40
429	-	<i>n</i> -C ₃ H ₇	4-Br-Ph-NH-CO	-	350	50	150	0.45
430	-	<i>n</i> -C ₃ H ₇	4-F-Ph-NH-CO	-	376	42	161	0.29
431	-	<i>n</i> -C ₃ H ₇	2-OCH ₃ -Ph-NH-CO	-	250	100	165	0.34
432	-	<i>n</i> -C ₃ H ₇	3-OCH ₃ -Ph-NH-CO	-	275	113	175	0.40
433	-	<i>n</i> -C ₄ H ₉	4-Cl-Ph-NH-CO	-	300	111	87	0.43
434	-	CH ₃	Ph-CH ₂ -CO	-	594	381	222	0.16
435	-	C ₂ H ₅	4-SO ₃ H-Ph-NH-CO	-	>10,000	249	>10,000	40
436	-	<i>n</i> -C ₃ H ₇	3-Cl-Ph-NH-CO	-	30	1,220	57	0.91
437	-	<i>n</i> -C ₃ H ₇	H	-	10	2.51	39	613
438	-	Ph-CH ₂ -CH ₂	4-OCH ₃ -Ph-NH-CO	-	201	120	81	1.47
439	-	Ph-CH ₂ -CH ₂ -CH ₂	4-OCH ₃ -Ph-NH-CO	-	251	1,010	802	19.81
440	-	CH ₂ -CH=CH ₂	4-OCH ₃ -Ph-NH-CO	-	1,531	176	2,030	0.48
441	-	<i>t</i> -C ₄ H ₉	4-OCH ₃ -Ph-NH-CO	-	796	545	1,715	0.8
442	-	CH ₃	H	-	101	2.8	90	300
443	-	(CH ₃) ₂ C=CH-CH ₂	H	-	5.42	0.8	12	811

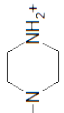
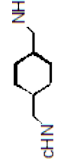

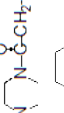

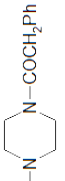
MODELS 1, 2, 3, internal test set: continued

Mol Id	n	R	R ₁	R ₂	Exp. K _i (nM) hA ₁ R or (% displ. 10 μM)	Exp. K _i (nM) hA ₂ A R	Exp. K _i (nM) hA ₂ B R (IC ₅₀ nM)	Exp. K _i (nM) hA ₃ R or (% displ. 10 μM)
444	-	4-OCH ₂ CONHPH-4'-I-Ph	-	-	>1,000	>1,000	>1,000	>1,000
445	-	CH ₃	H	SCH ₃	8.4	1.2	10.3	35
446	-	CH ₃	4-OCH ₃ -Ph-NH-CO	NH-CH ₂ -CH ₃	150	21	37	14
447	-	4-OH-Ph	-	-	>1,000	>1,000	>1,000	>1,000
448	-	4-Cl-Ph	-	-	>1,000	>1,000	>1,000	>1,000
449	-	CH ₂ -CO-CH ₃	CH ₂ -Ph	-	>1,000	>1,000	>1,000	182
450	-	CH ₂ -CO ₂ -Et	CH ₂ -Ph	-	>1,000	>1,000	>1,000	20
451	-	CH ₂ -CH ₂ -CH ₃	CH ₂ CH ₃	-	>1,000	>1,000	>1,000	200
452	-	H	C(CH ₃) ₃	-	>1,000	>1,000	>1,000	99
453	-	H	cyclopropyl	-	350	>1,000	>1,000	23
454	-	cyclopropyl methyl	H	CF ₃	>6,000	>5,000	3	1,000
455	-	propyl	H	Ph	3,500	>5,000	15	>9,000
456	-	<i>n</i> -C ₄ H ₉	H	4-Cl-Ph	>6,000	>5,000	23	>9,000
457	-	<i>i</i> -C ₄ H ₉	H	4-CF ₃ -Ph	>6,000	>5,000	14	>9,000
458	-	cyclopropyl methyl	H	4-CF ₃ -Ph	>6,000	>5,000	13	>9,000
459	-	propyl	H	Ph	>6,000	>5,000	22	>9,000
460	-	cC ₆ H ₁₁ -CH ₂ -CH ₂	H	-	0.70	20.2	55.8	>10,000
461	-	(CH ₃) ₂ CH-CH ₂ -CH ₂	CO-NH-C(CH ₃) ₃	-	523	>10,000	>10,000	39

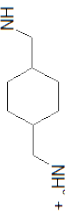
MODELS 1, 2, 3, internal test set: continued

Mol Id	n	R	R ₁	R ₂	Exp. K _i (nM) hA ₁ R or (% displ. 10 μM)	Exp. K _i (nM) hA _{2A} R	Exp. K _i (nM) hA _{2B} R or (IC ₅₀ nM)	Exp. K _i (nM) hA ₃ R or (% displ. 10 μM)
462	-	CH ₃ -CH ₂ -CH ₂	CO-NH-CH-(CH ₃) ₂	-	357	>10,000	>10,000	21
463	-	propyl	3-noradamantyl	-	90.6	34,500	3,190	>10,000
464	-	propyl	(R)-1-phenylethyl	biphenyl	>10,000	>10,000	>10,000	456
465	-	CH ₃	4-N(CH ₃) ₂ -Ph-NH-CO-CH ₂	-	>1,000	>1,000	>1,000	>1,000
466	-	CH ₃	3-Cl-Ph-NH-CO-CH ₂	-	675	>1,000	28	>1,000
467	-	CH ₃	3,4-Cl ₂ -Ph-NH-CO-CH ₂	-	>1,000	>1,000	40	>1,000
468	-	CH ₃	3,4-CH ₃ -Ph-NH-CO-CH ₂	-	>1,000	>1,000	48	>1,000
469	-	CH ₂ -CH ₃	4-Cl-Ph-NH-CO-CH ₂	-	>1,000	>1,000	>1,000	>1,000
470	-	H	4-Cl-Ph-NH-CO-CH ₂	-	140	>1,000	20	>1,000
471	-	CH ₃	2-CH ₃ -Ph-CH ₂ O-Ph-CH ₂ -CO	-	2,940	333	>30,000	6.8
472	-	CH ₃	3-CH ₃ -Ph-CH ₂ O-Ph-CH ₂ -CO	-	6,870	730	>30,000	6.04
473	-	CH ₃	4-CH ₃ -Ph-CH ₂ O-Ph-CH ₂ -CO	-	6,760	751	>30,000	24.4
474	-	CH ₃	2,4-CH ₃ -Ph-CH ₂ O-Ph-CH ₂ -CO	-	4,360	481	>30,000	14.5
475	-	CH ₂ -CH ₂ -CH(CH ₃) ₂	4-CH ₃ -Ph-CH ₂ O-Ph-CH ₂ -CO	-	501	47.3	>30,000	271
476	-	CH ₂ -CH ₂ -CH(CH ₃) ₂	2,4-CH ₃ -Ph-CH ₂ O-Ph-CH ₂ -CO	-	814	69.0	>30,000	315
477	-	CH ₂ -CH ₂ -CH(CH ₃) ₂	2,5-CH ₃ -Ph-CH ₂ O-Ph-CH ₂ -CO	-	882	75.3	>30,000	193
478	-	Ph-CH ₂ -CH ₂	2-CH ₃ -Ph-CH ₂ O-Ph-CH ₂ -CO	-	261	83.4	>10,000	70.1
479	-	Ph-CH ₂ -CH ₂	3-CH ₃ -Ph-CH ₂ O-Ph-CH ₂ -CO	-	240	112	>10,000	48.9
480	-	Ph-CH ₂ -CH ₂	4-CH ₃ -Ph-CH ₂ O-Ph-CH ₂ -CO	-	360	162	>10,000	49.9

MODELS 1, 2, 3, internal test set: continued

Mol Id	n	R	R ₁	R ₂	Exp. K _i (nM) hA ₁ R or (% displ. 10 μM)	Exp. K _i (nM) hA ₂ A R	Exp. K _i (nM) hA ₂ B R or (IC ₅₀ nM)	Exp. K _i (nM) hA ₃ R or (% displ. 10 μM)
481	-	Ph-CH ₂ -CH ₂	2,4-CH ₃ -Ph-CH ₂ O-Ph-CH ₂ -CO	-	278	155	>10,000	98.0
482	-	H		-	1,754	532	(100,000)	(29.8)
483	-	H	BocHN- 	-	895	213	(7,140)	(75.5)
484	-	H	NH-(CH ₂) ₄ CH ₂ -NHBOC	-	3,559	447	(>10,000)	(68.6)
485	-	H	NH-(CH ₂ CH ₂ O) ₂ CH ₂ -CH ₂ -NHBOC	-	696	184	(30,000)	(74.0)
486	-	H		-	53	66	(10,000)	(49.4)
487	-	H	NH-(CH ₂ CH ₂ O) ₂ CH ₂ -CH ₂ -NH ₃ ⁺	-	>10,000	64	(>30,000)	(59.2)
488	-	H		-	87	173	(>30,000)	1,240
489	-	H		-	212	71	(>100,000)	(56.2)
490	-	H		-	2,497	2,000	(>100,000)	(64.5)
491	-	NH-(CH ₂) ₂ -NH ₃ ⁺	NH-(CH ₂) ₂ -NH ₃ ⁺	-	>10,000	>10,000	(>100,000)	(17.7)

MODELS 1, 2, 3, internal test set: continued

Mol Id	n	R	R ₁	R ₂	Exp. K_i (nM) hA ₁ R or (% displ. 10 μ M)	Exp. K_i (nM) hA _{2A} R	Exp. K_i (nM) hA _{2B} R or (IC ₅₀ nM)	Exp. K_i (nM) hA ₃ R or (% displ. 10 μ M)
492	-	H	NH-(CH ₂) ₂ -NH ₃ +	-	>10,000	2,200	(>100,000)	(30.8)
493	-	H	 + ³ NH	-	551	217	(10,700)	(72.8)
494	-	H	NH-(CH ₂) ₃ -NH ₃ +	-	583	2,081	(>30,000)	(40.5)
495	-	H	NH-(CH ₂) ₅ -NH ₃ +	-	>10,000	983	(>10,000)	(48.5)
496	-	H	OH	-	>10,000	709	(>10,000)	(43.9)
497	-	H	NH ₂	-	>10,000	420	(11,100)	(36.3)
498	-	Ph-CO	OH	-	>10,000	5,320	(>100,000)	(75.9)
499	-	N(CH ₃) ₂	H	-	31,400	3800	(33,700)	7,270
500	-	SCH ₃	Ph ₂ -CH-CO	-	557	>10,000	(>100,000)	170
501	-	SCH ₃	4-OCH ₃ -Ph-CH ₂ -CO	-	5,000	1,570	(42,400)	2,100
502	-	SCH ₃	CH ₃ -(CH ₂) ₄ -CO	-	1,520	427	(>100,000)	(24)
503	-	SCH ₃	Ph-NH-CO	-	10,000	1,660	(17,100)	414
504	-	N(CH ₃) ₂	4-OCH ₃ -Ph-CH ₂ -CO	-	8,740	10,300	(>100,000)	2,700
505	-	CH ₃	Ph-NH-CO	-	310	27.7	(3,440)	1.80
506	-	cC ₆ H ₁₁ -CH ₂	CH ₃ -CO-CH ₂ -Ph-NH-CO	-	71.8	6.80	(7,020)	42.3
507	-	CH ₂ -CH ₃	CH ₃ CH ₂ CH(CH ₃)-NH-CO	-	376	559	(>30,000)	0.53
508	-	CH ₂ -CH ₂ -CH ₂ - CH(CH ₃) ₂	CH ₃ -CO-CH ₂ -Ph-NH-CO	-	137	24.8	(13,000)	68.8

MODELS 1, 2, 3, internal test set: continued

Mol Id	n	R	R ₁	R ₂	Exp. K_i (nM) hA ₁ R or (% displ. 10 μ M)	Exp. K_i (nM) hA ₂ A R	Exp. K_i (nM) hA ₂ B R (IC ₅₀ nM) or	Exp. K_i (nM) hA ₃ R or (% displ. 10 μ M)
509	-	CH ₂ -CH ₂ -CH ₃	CH ₃ -CH ₂ -NH-CO	-	77.2	223	(2,810)	0.60
510	-	CH ₂ -CH ₂ -CH ₃	cC ₆ H ₁₁ -CH ₂ -NH-CO	-	13.8	258	(3,440)	3.53
511	-	cC ₆ H ₁₁ -CH ₂	CH ₃ O-CO-CH ₂ -cC ₆ H ₁₁ -NH-CO	-	52.5	5.74	(8,940)	164
512	-	CH ₂ -CH ₂ -CH ₂ -CH ₂ -CH ₃	CH ₃ -CH ₂ O-CO-CH ₂ -Ph-NH-CO	-	54.7	4.77	(9,740)	18
513	-	CH ₂ -CH ₂ -CH ₂ -CH ₂ -CH ₃	CH ₃ CH ₂ CH(CH ₃)-NH-CO	-	576	300	(>100,000)	5.33
514	-	CH ₂ -CH ₂ -CH ₂ -CH ₂ -CH ₃	4-CH ₃ -Ph-NH-CO	-	34.7	3.72	(2,910)	44.4

E.4 References

- [1] Baraldi, P. G.; Tabrizi, M. A.; Bovero, A.; Avitabile, B.; Preti, D.; Fruttarolo, F.; Romagnoli, R.; Varani, K.; Borea, P. A. Recent developments in the field of A_{2A} and A_3 adenosine receptor antagonists. *Eur. J. Med. Chem.* **2003**, *38*, 367-382.
- [2] Baraldi, P. G.; Cacciari, B.; Romagnoli, R.; Spalluto, G.; Monopoli, A.; Ongini, E.; Varani, K.; Borea, P. A. 7-Substituted 5-amino-2-(2-furyl)pyrazolo[4,3-*e*]-1,2,4-triazolo[1,5-*c*]pyrimidines as A_{2A} adenosine receptor antagonists: a study on the importance of modifications at the side chain on the activity and solubility. *J. Med. Chem.*, **2002**, *45*, 115-126.
- [3] Baraldi, P. G.; Cacciari, B.; Moro, S.; Spalluto, G.; Pastorin, G.; Da Ros, T.; Klotz, K. N.; Varani, K.; Gessi, S.; Borea, P. A. Synthesis, biological activity, and molecular modeling investigation of new pyrazolo[4,3-*e*]-1,2,4-triazolo[1,5-*c*]pyrimidine derivatives as human A_3 adenosine receptor antagonists. *J. Med. Chem.* **2002**, *45*, 770-780.
- [4] Baraldi, P. G.; Cacciari, B.; Romagnoli, R.; Spalluto, G.; Varani, K.; Gessi, S.; Merighi, S.; Borea, P. A. Pyrazolo[4,3-*e*]1,2,4-triazolo[1,5-*c*]pyrimidine derivatives: a new pharmacological tool for the characterization of the human A_3 adenosine receptor. *Drug. Dev. Res.* **2001**, *52*, 406-415.
- [5] Baraldi, P. G.; Cacciari, B.; Romagnoli, R.; Spalluto, G.; Moro, S.; Klotz, K. N.; Leung, E.; Varani, K.; Gessi, S.; Merighi, S.; Borea, P. A. Pyrazolo[4,3-*e*]1,2,4-triazolo[1,5-*c*]pyrimidine derivatives as highly potent and selective human A_3 adenosine receptor antagonists: influence of the chain at the N8 pyrazole nitrogen. *J. Med. Chem.* **2000**, *43*, 4768-4780.

- [6] Baraldi, P. G.; Cacciari, B.; Romagnoli, R.; Klotz, K. N.; Spalluto, G.; Varani, K.; Gessi, S.; Merighi, S.; Borea, P. A. Pyrazolo[4,3-*e*]1,2,4-triazolo[1,5-*c*]pyrimidine derivatives as adenosine receptor ligands: a starting point for searching A_{2B} adenosine receptor antagonists. *Drug Dev. Res.* **2001**, *53*, 225-235.
- [7] Michielan, L.; Bolcato, C.; Stephanie, F.; Cacciari, B.; Bacilieri, M.; Klotz, K. N.; Kachler, S.; Pastorin, G.; Cardin, R.; Sperduti, A.; Spalluto, G.; Moro, S. Combining selectivity and affinity predictions using an integrated support vector machine (SVM) approach: a novel tool to discriminate between the human adenosine A_{2A} and A_3 receptor pyrazolo-triazolo-pyrimidine antagonists binding sites. *Bioorg. Med. Chem.* **2009**, *17* (14), 5259-5274.
- [8] Michielan, L.; Bacilieri, M.; Schiesaro, A.; Bolcato, C.; Pastorin, G.; Spalluto, G.; Cacciari, B.; Klotz, K. N.; Moro, S. Linear and nonlinear 3D-QSAR approaches in tandem with ligand-based homology modeling as a computational strategy to depict the pyrazolo-triazolo-pyrimidine antagonists binding site of the human adenosine A_{2A} receptor. *J. Chem. Inf. Model.* **2008**, *48* (2) 350-363.
- [9] Baraldi, P. G.; Fruttarolo, F.; Tabrizi, M. A.; Preti, D.; Romagnoli, R.; El-Kashef, H.; Moorman, A.; Varani, K.; Gessi, S.; Merighi, S.; Borea, P. A. Design, synthesis, and biological evaluation of C9- and C2-substituted pyrazolo[4,3-*e*]1,2,4-triazolo[1,5-*c*]pyrimidines as new A_{2A} and A_3 adenosine receptors antagonists. *J. Med. Chem.* **2003**, *46*, 1229-1241.
- [10] Tabrizi, M. A.; Baraldi, P. G.; Preti, D.; Romagnoli, R.; Saponaro, G.; Baraldi, S.; Moorman, A. R.; Zaid, A. N.; Varani, K.; Borea, P. A. 1,3-dipropyl-8-(1-phenylacetamide-1H-pyrazol-3-yl)-xanthine derivatives as highly potent and selective human A_{2B} adenosine receptor antagonists. *Bioorg.*

Med. Chem. **2008**, *16*, 2419-2430.

- [11] Weyler, S.; Füller, F.; Diekmann, M.; Schumacher, B.; Hinz, S.; Klotz, K. N.; Müller, C. E. Improving potency, selectivity, and water solubility of adenosine A₁ receptor antagonists: xanthine modified at position 3 and related pyrimido[1,2,3-*cd*]purinediones. *Chem. Med. Chem.* **2006**, *1*, 891-902.
- [12] Baraldi, P. G.; Preti, D.; Tabrizi, M. A.; Romagnoli, R.; Saponaro, G.; Baraldi, S.; Botta, M.; Bernardini, C.; Tafi, A.; Tuccinardi, T.; Martinelli, A.; Varani, K.; Borea, P. A. Structure-activity relationship studies of a new series of imidazo[2,1-*f*]purinones as potent and selective A₃ adenosine receptor antagonists. *Bioorg. Med. Chem.* **2008**, *16*, 10281-10294.
- [13] Elzein, E.; Kalla, R.; Li, X.; Perry, T.; Parkhill, E.; Palle, V.; Varkhedkar, V.; Gimbel, A.; Zeng, D.; Lustig, D.; Leung, K.; Zablocki, J. Novel 1,3-dipropyl-8-(1-heteroarylmethyl-1H-pyrazol-4-yl)-xanthine derivatives as high affinity and selective A_{2B} adenosine receptor antagonists. *Bioorg. Med. Chem. Lett.* **2006**, *16*, 302-306.
- [14] Kalla, R. V.; Elzein, E.; Perry, T.; Li, X.; Gimbel, A.; Yang, M.; Zeng, D.; Zablocki, J. Selective, high affinity A_{2B} adenosine receptor antagonists: N-1 monosubstituted 8-(pyrazol-4-yl)xanthines. *Bioorg. Med. Chem.* **2008**, *18*, 1397-1401.
- [15] Baraldi, P. G.; Tabrizi, M. A.; Preti, D.; Bovero, A.; Romagnoli, R.; Fruttarolo, F.; Zaid, N. A.; Moorman, A. R.; Varani, K.; Gessi, S.; Merighi, S.; Borea, P. A. Design, synthesis, and biological evaluation of new 8-heterocyclic xanthine derivatives as highly potent and selective human A_{2B} adenosine receptor antagonists. *J. Med. Chem.* **2004**, *47*, 1434-1477.

- [16] Moro, S.; Bacilieri, M.; Cacciari, B.; Bolcato, C.; Cusan, C.; Pastorin, G.; Klotz, K. N.; Spalluto, G. The application of a 3D-QSAR (*autoMEP/PLS*) approach as an efficient pharmacodynamic-driven filtering method for small-sized virtual library: application to a lead optimization of a human A₃ adenosine receptor antagonists. *Bioorg. Med. Chem.*, **2006**, *14*, 4923-4932.

Table E.1

Classes	TP	FP	TN	FN	Recall	Precision
hA ₁ R	11	12	32	10	0.52	0.48
hA _{2A} R	26	7	27	5	0.84	0.79
hA _{2B} R	28	6	27	4	0.87	0.82
hA ₃ R	29	7	29	2	0.93	0.81

Table E.2

Classes	TP	FP	TN	FN	Recall	Precision
hA ₁ R	2	5	48	10	0.17	0.29
hA _{2A} R	22	6	29	8	0.73	0.79
hA _{2B} R	23	6	29	7	0.77	0.79
hA ₃ R	22	5	34	4	0.85	0.81

Table E.3

Classes	TP	FP	TN	FN	Recall	Precision
hA ₁ R	2	1	58	4	0.33	0.67
hA _{2A} R	14	5	39	7	0.67	0.74
hA _{2B} R	12	4	42	7	0.63	0.75
hA ₃ R	17	5	40	3	0.85	0.77

Table E.4

Classes	MODEL 1		MODEL 2		MODEL 3	
	Recall	Precision	Recall	Precision	Recall	Precision
hA ₁ R	0.54	1.00	0.18	1.00	0.00	0.00
hA _{2A} R	1.00	0.92	1.00	0.92	0.82	1.00
hA _{2B} R	0.50	0.25	1.00	0.50	-	-
hA ₃ R	1.00	0.67	1.00	0.50	1.00	0.28
Accuracy _{ML} ($\alpha=1$)	0.77		0.69		0.57	

APPENDIX F

Paper VI

Supplementary Information

F.1 *TOXclass* model, training set (no. refers to the reference database, Experimental and predicted classes: *highAT*, high acute toxicity; *lowAT*, low acute toxicity).

F.2 *MOAclass* model, training set (no. refers to the reference database. Experimental and predicted classes: A, baseline narcosis; B, polar narcosis; C, arylate and ester narcosis; D, electrophile or proelectrophile phosphorylation; E, neurodepressant; F, uncoupler of oxidative phosphorylation; G, central nervous system seizure mechanisms; H, AchE inhibition; I, respiratory blocker or inhibition).

F.3 Test set (no. refers to the reference database. Experimental and predicted classes by *TOXclass* model: *highAT*, high acute toxicity; *lowAT*, low acute toxicity; experimental and predicted classes by *MOAclass* model: A, baseline narcosis; B, polar narcosis; C, arylate and ester narcosis; D, electrophile or proelectrophile phosphorylation; E, neurodepressant; F, uncoupler of oxidative phosphorylation; G, central nervous system seizure mechanisms; H, AchE inhibition; I, respiratory blocker or inhibition).

F.4 References.

F.1 TOX class model: Training set

No.	Substance chemical name	Substance CASRN	LC ₅₀ (mg/l)	LC ₅₀ (mmol/l)	Exp. class	Tox class	Pred. Tox class
1	4-(Hexyloxy)- <i>m</i> -anisaldehyde	61096-84-2	2.67	0.0113	<i>high</i> AT	<i>high</i> AT	<i>high</i> AT
2	5-Bromo-2-nitrovanillin	98434-34-5	73.30	0.2660	<i>high</i> AT	<i>high</i> AT	<i>high</i> AT
4	<i>p</i> -Chlorophenyl- <i>o</i> -nitrophenyl ether	39145-47-6	1.92	0.0077	<i>high</i> AT	<i>high</i> AT	<i>high</i> AT
5	3'-Chloro- <i>o</i> -formotoluidide	71862-02-7	46.60	0.2750	<i>high</i> AT	<i>high</i> AT	<i>high</i> AT
6	Di- <i>n</i> -butylisophthlate	3126-90-7	0.90	0.0032	<i>high</i> AT	<i>high</i> AT	<i>high</i> AT
7	1,1-Diphenyl-2-propyn-1-ol	3923-52-2	11.10	0.0533	<i>high</i> AT	<i>high</i> AT	<i>high</i> AT
8	4,7-Dithiadecane	22037-97-4	7.52	0.0422	<i>high</i> AT	<i>high</i> AT	<i>high</i> AT
11	4,9-Dithiadodecane	56348-39-1	2.99	0.0145	<i>high</i> AT	<i>high</i> AT	<i>high</i> AT
14	2-chloroethyl-N-cyclohexyl carbamate	31502-57-5	35	0.1700	<i>high</i> AT	<i>high</i> AT	<i>high</i> AT
15	Phenobarbital	50-06-6	484	2.0800	<i>low</i> AT	<i>low</i> AT	<i>low</i> AT
16	2,4-Dinitrophenol	51-28-5	10.90	0.0592	<i>high</i> AT	<i>high</i> AT	<i>high</i> AT
17	Urethane	51-79-6	5240	58.8000	<i>low</i> AT	<i>low</i> AT	<i>low</i> AT
18	Salicylic acid Na+	54-21-7	1720	12.5000	<i>low</i> AT	<i>low</i> AT	<i>low</i> AT
19	Benzamide	55-21-0	661	5.4600	<i>low</i> AT	<i>low</i> AT	<i>low</i> AT
21	1,1-Dimethyl hydrazine	57-14-7	7.85	0.1310	<i>high</i> AT	<i>high</i> AT	<i>high</i> AT
22	Pentobarbital sodium	57-33-0	49.50	0.1990	<i>high</i> AT	<i>high</i> AT	<i>high</i> AT
23	Amobarbital	57-43-2	85.40	0.3770	<i>high</i> AT	<i>high</i> AT	<i>high</i> AT
24	Caffeine	58-08-2	151	0.7780	<i>low</i> AT	<i>low</i> AT	<i>low</i> AT
25	2-Methyl-1,4-naphthoquinone	58-27-5	0.11	0.0006	<i>high</i> AT	<i>high</i> AT	<i>high</i> AT
26	2,3,4,6-Tetrachlorophenol	58-90-2	1.03	0.0044	<i>high</i> AT	<i>high</i> AT	<i>high</i> AT
27	4-Chloro-3-methyl phenol	59-50-7	5.47	0.0384	<i>high</i> AT	<i>high</i> AT	<i>high</i> AT

TOX class model, training set: continued

No.	Substance chemical name	Substance CASRN	LC ₅₀ (mg/l)	LC ₅₀ (mmol/l)	Exp. class	Tox class	Pred. class	Tox class
28	Tolazoline hydrochloride	59-97-2	354	1.8000	lowAT	lowAT	highAT	highAT
29	Amphetamine sulfate	60-13-9	28.80	0.0782	highAT	highAT	highAT	highAT
30	Diethyl ether	60-29-7	2560	34.5000	lowAT	lowAT	lowAT	lowAT
31	Strychnine hemisulphate salt	60-41-3	1.11	0.0011	highAT	highAT	highAT	highAT
32	Aniline	62-53-3	105	1.1300	lowAT	lowAT	lowAT	lowAT
33	Carbaryl (sevin)	63-25-2	8.75	0.0435	highAT	highAT	highAT	highAT
34	Ethanol	64-17-5	14700	319.0000	lowAT	lowAT	lowAT	lowAT
35	Nicotine sulfate	65-30-5	13.80	0.0530	highAT	highAT	lowAT	lowAT
36	2-Hydroxybenzamide	65-45-2	101	0.7360	lowAT	lowAT	lowAT	lowAT
38	Hexanal	66-25-1	17.50	0.1750	highAT	highAT	highAT	highAT
39	Dicumarol	66-76-2	5.11	0.0152	highAT	highAT	highAT	highAT
40	<i>p</i> -Phenoxybenzaldehyde	67-36-7	4.60	0.0232	highAT	highAT	highAT	highAT
41	Methanol	67-56-1	29400	917.0000	lowAT	lowAT	lowAT	lowAT
42	2-Propanol	67-63-0	8680	144.0000	lowAT	lowAT	lowAT	lowAT
43	Acetone	67-64-1	7160	123.0000	lowAT	lowAT	lowAT	lowAT
44	Chloroform	67-66-3	70.70	0.5920	lowAT	lowAT	lowAT	lowAT
45	Methyl sulfoxide	67-68-5	34000	435.0000	lowAT	lowAT	lowAT	lowAT
46	Hexachloroethane	67-72-1	1.42	0.0060	highAT	highAT	highAT	highAT
47	2,2'-Methylene bis(3,4,6-trichlorophenol)	70-30-4	0.02	0.0001	highAT	highAT	highAT	highAT
48	4'-Aminopropiophenone	70-69-9	146	0.9790	lowAT	lowAT	highAT	highAT
49	1-Propanol	71-23-8	4550	75.7000	lowAT	lowAT	lowAT	lowAT
50	<i>n</i> -Butanol	71-36-3	1730	23.3400	lowAT	lowAT	lowAT	lowAT

TOX class model, training set: continued

No.	Substance chemical name	Substance CASRN	LC ₅₀ (mg/l)	LC ₅₀ (mmol/l)	Exp. Tox class	Pred. Tox class
51	1-Pentanol	71-41-0	472	5.3550	lowAT	lowAT
52	Benzene	71-43-2	17.60	0.2250	highAT	highAT
53	1,1,1-Trichloroethane	71-55-6	47.30	0.3550	highAT	highAT
54	Thiopental, sodium salt	71-73-8	26.20	0.0991	highAT	highAT
55	Acetonitrile	75-05-8	1644	40.0500	lowAT	lowAT
56	Ethanal	75-07-0	33.80	0.7670	lowAT	lowAT
57	Dichloromethane	75-09-2	330	3.8850	lowAT	lowAT
58	Iodoform	75-47-8	2.92	0.0074	highAT	highAT
60	2-Methyl-2-propanol	75-65-0	6410	86.4800	lowAT	lowAT
61	2,2,2-Trifluoroethanol	75-89-8	119	1.1900	lowAT	lowAT
62	3,3-Dimethyl-2-butanone	75-97-8	87	0.8690	lowAT	lowAT
63	Pentachloroethane	76-01-7	7.53	0.0372	highAT	highAT
64	5,5-Dimethylhydantoin	77-71-4	16460	128.5000	lowAT	lowAT
65	3-Methyl-3-pentanol	77-74-7	672	6.5770	lowAT	lowAT
66	3-Methyl-1-pentyn-3-ol	77-75-8	1220	12.4300	lowAT	lowAT
67	1-Ethynyl-cyclohexanol	78-27-3	256	2.0610	lowAT	lowAT
68	Tris(2-butoxyethyl) phosphate	78-51-3	11.20	0.0281	highAT	highAT
69	2-Methyl-1-propanol	78-83-1	1430	19.2900	lowAT	lowAT
70	1,2-Dichloropropane	78-87-5	127	1.1240	lowAT	highAT
71	1,2-Diaminopropane	78-90-0	1010	13.6300	lowAT	lowAT
72	2-Butanol	78-92-2	3670	49.5100	lowAT	lowAT
73	2-Butanone	78-93-3	3220	44.6600	lowAT	lowAT

TOX class model, training set: continued

No.	Substance chemical name	Substance CASRN	LC ₅₀ (mg/l)	LC ₅₀ (mmol/l)	Exp. class	Tox class	Pred. class	Tox class
74	1-Amino-2-propanol	78-96-6	2520	33.5500	lowAT	lowAT	lowAT	lowAT
75	1,1,2-Trichloroethane	79-00-5	81.60	0.6120	lowAT	lowAT	lowAT	lowAT
76	Trichloroethylene	79-01-6	44.10	0.3360	highAT	highAT	highAT	highAT
77	Methyl acetate	79-20-9	357	4.8190	lowAT	lowAT	lowAT	lowAT
78	1,1,2,2-Tetrachloroethane	79-34-5	20.30	0.1210	highAT	highAT	highAT	highAT
79	beta-Ionone	79-77-6	5.09	0.0265	highAT	highAT	highAT	highAT
80	4,4'-Isopropylidenebis(2,6-dichlorophenol)	79-95-8	1.33	0.0036	highAT	highAT	highAT	highAT
81	p-tert-Pentylphenol	80-46-6	2.59	0.0158	highAT	highAT	highAT	highAT
82	1,8-Diamino-p-menthane	80-52-4	65.30	0.3830	highAT	highAT	highAT	highAT
83	alpha,alpha-2,6-Tetrachlorotoluene	81-19-6	0.97	0.0042	highAT	highAT	highAT	highAT
84	Acenaphthene	83-32-9	1.73	0.0112	highAT	highAT	highAT	highAT
85	3-Methylindole	83-34-1	8.84	0.0674	highAT	highAT	highAT	highAT
86	Rotenone	83-79-4	0.0052	0.0000	highAT	highAT	highAT	highAT
88	Diphenyl phthalate	84-62-8	0.08	0.0003	highAT	highAT	highAT	highAT
90	Diethyl phthalate	84-66-2	31.80	0.1430	highAT	highAT	highAT	highAT
91	Di-n-butylorthophthalate	84-74-2	1.00	0.0036	highAT	highAT	highAT	highAT
94	Salicylamide	87-17-2	3.95	0.0185	highAT	highAT	highAT	highAT
95	Hexachloro-1,3-butadiene	87-68-3	0.09	0.0003	highAT	highAT	highAT	highAT
96	Pentachlorophenol	87-86-5	0.2450	0.0009	highAT	highAT	highAT	highAT
97	2,4,6-Trichlorophenol	88-06-2	4.89	0.0248	highAT	highAT	highAT	highAT
98	3-Trifluoromethyl-4-nitrophenol	88-30-2	9.14	0.0441	highAT	highAT	highAT	highAT
99	Antbranilamide	88-68-6	395	2.9010	lowAT	lowAT	lowAT	lowAT

TOX class model, training set: continued

No.	Substance chemical name	Substance CASRN	LC ₅₀ (mg/l)	LC ₅₀ (mmol/l)	Exp. Tox class	Pred. Tox class
100	2-Nitrophenol	88-75-5	160	1.1500	lowAT	lowAT
101	2-sec-Butyl-4,6-dinitrophenol (dinoseb)	88-85-7	0.5350	0.0022	highAT	highAT
102	Salicylaldehyde	90-02-8	2.30	0.0188	highAT	highAT
103	1-Naphthol	90-15-3	4.63	0.0321	highAT	highAT
104	2-Phenylphenol	90-43-7	6.15	0.0361	highAT	highAT
106	3,5-Dibromosalicylaldehyde	90-59-5	0.85	0.0030	highAT	highAT
107	Naphthalene	91-20-3	6.14	0.0479	highAT	highAT
108	Quinoline	91-22-5	77.80	0.6020	lowAT	highAT
109	N,N-Diethylcyclohexylamine	91-65-6	21.40	0.1380	highAT	highAT
110	N,N-Diethylamine	91-66-7	16.40	0.1100	highAT	highAT
111	2-(N-Ethyl-m-toluidino)ethanol	91-88-3	52.90	0.2950	highAT	highAT
113	1-Benzoylacetone	93-91-4	1.10	0.0068	highAT	highAT
114	Ethyl-p-amino benzoate	94-09-7	35.70	0.2160	highAT	highAT
115	Piperine (aliphatic)	94-62-2	7.84	0.0275	highAT	highAT
116	2,4-Dihydroxybenzaldehyde	95-01-2	13.10	0.0950	highAT	lowAT
117	o-Xylene	95-47-6	16.40	0.1540	highAT	highAT
118	o-Cresol	95-48-7	14.00	0.1290	highAT	highAT
119	1,2-Dichlorobenzene	95-50-1	9.47	0.0640	highAT	highAT
120	2-Chloroaniline	95-51-2	5.74	0.0450	highAT	highAT
121	2-Fluorotoluene	95-52-3	19.40	0.1760	highAT	highAT
122	2-Chlorophenol	95-57-8	11.40	0.0887	highAT	highAT
123	1,2,4-Trimethylbenzene	95-63-6	7.72	0.0642	highAT	highAT

TOX class model, training set: continued

No.	Substance chemical name	Substance CASRN	LC ₅₀ (mg/l)	LC ₅₀ (mmol/l)	Exp. class	Tox class	Pred. class
124	3,4-Dichlorotoluene	95-75-0	2.91	0.0181	highAT	highAT	highAT
125	3,4-Dichloroaniline	95-76-1	7.57	0.0467	highAT	highAT	highAT
126	Allyl methacrylate	96-05-9	0.99	0.0079	highAT	highAT	highAT
128	2-Methylbutyraldehyde	96-17-3	9.97	0.1160	highAT	lowAT	lowAT
129	1,2,3-Trichloropropane	96-18-4	57.70	0.3910	highAT	highAT	highAT
130	3-Pentanone	96-22-0	1540	17.8800	lowAT	lowAT	lowAT
131	2-Butanone oxime	96-29-7	843	9.6760	lowAT	lowAT	lowAT
132	2-(Diisopropylamino)ethanol	96-80-0	201	1.3840	lowAT	lowAT	lowAT
133	2,4-Dinitroaniline	97-02-9	14.80	0.0808	highAT	highAT	highAT
134	2,2'-Methylenebis(4-chlorophenol)	97-23-4	0.31	0.0012	highAT	highAT	highAT
136	<i>p-tert</i> -Butylphenol	98-54-4	5.150	0.0343	highAT	highAT	highAT
137	Isopropylbenzene	98-82-8	6.32	0.0526	highAT	highAT	highAT
138	Acetophenone	98-86-2	162	1.3480	lowAT	highAT	highAT
139	Nitrobenzene	98-95-3	119	0.9670	lowAT	highAT	highAT
140	<i>m</i> -Aminoacetophenone	99-03-6	382	2.8260	lowAT	lowAT	lowAT
141	<i>m</i> -Nitrotoluene	99-08-1	25.60	0.1870	highAT	highAT	highAT
142	N,N-Dimethyl- <i>p</i> -toluidine	99-97-8	48.90	0.3620	highAT	highAT	highAT
143	<i>p</i> -Nitroaniline	100-01-6	125	0.9050	lowAT	lowAT	lowAT
144	<i>p</i> -Nitrophenol	100-02-7	44.80	0.3220	highAT	highAT	highAT
145	<i>p</i> -Dimethylaminobenzaldehyde	100-10-7	45.70	0.3060	highAT	highAT	highAT
146	1,4-Dinitrobenzene	100-25-4	0.71	0.0042	highAT	highAT	highAT
147	N,N-Diethyl ethanolamine	100-37-8	1780	15.1900	lowAT	lowAT	lowAT

TOX class model, training set: continued

No.	Substance chemical name	Substance CASRN	LC ₅₀ (mg/l)	LC ₅₀ (mmol/l)	Exp. Tox class	Pred. Tox class
148	Ethylbenzene	100-41-4	10.50	0.0989	highAT	highAT
149	Benzylamine	100-46-9	102	0.9520	lowAT	lowAT
150	Benzaldehyde	100-52-7	9.87	0.0930	highAT	highAT
151	N-Methylamine	100-61-8	100	0.9330	lowAT	lowAT
152	Cyclohexanone oxime	100-64-1	208	1.8380	lowAT	lowAT
153	2-Cyanopyridine	100-70-9	726	6.9730	lowAT	lowAT
154	2-Ethylpyridine	100-71-0	414	3.8640	lowAT	lowAT
155	Solketal	100-79-8	16700	126.4000	lowAT	lowAT
156	Hexamethylenetetramine (aliphatic)	100-97-0	49800	355.2000	lowAT	lowAT
157	Phenyl ether	101-84-8	4.00	0.0235	highAT	highAT
158	N-Ethyl- <i>m</i> -toluidine	102-27-2	49.50	0.3660	highAT	highAT
159	Tripropylamine	102-69-2	50.90	0.3550	highAT	highAT
160	Triethanolamine	102-71-6	11800	79.0900	lowAT	lowAT
161	Benzyl- <i>tert</i> -butanol	103-05-9	66.40	0.4040	highAT	highAT
163	1-(2-Hydroxyethyl)piperazine	103-76-4	6410	49.2400	lowAT	lowAT
164	N,N-Dimethylbenzylamine	103-83-3	37.80	0.2800	highAT	highAT
165	4-Acetamidophenol	103-90-2	814	5.3850	lowAT	lowAT
166	4-Butylaniline	104-13-2	10.20	0.0680	highAT	highAT
167	Nonylphenol (mixed)	25154-52-3	0.14	0.0006	highAT	highAT
169	2-Ethyl-1-hexanol	104-76-7	28.20	0.2170	highAT	highAT
170	4-Chlorobenzaldehyde	104-88-1	2.20	0.0156	highAT	highAT
171	5-Ethyl-2-methylpyridine	104-90-5	81.10	0.6690	lowAT	highAT

TOX class model, training set: continued

No.	Substance chemical name	Substance CASRN	LC ₅₀ (mg/l)	LC ₅₀ (mmol/l)	Exp. Tox class	Pred. Tox class
172	5-Diethylamino-2-pentanone	105-14-6	336	2.1370	lowAT	lowAT
173	Diethyl malonate	105-53-3	14.70	0.0918	highAT	lowAT
174	2,4-Dimethylphenol	105-67-9	16.60	0.1360	highAT	highAT
175	Dibutyl fumarate	105-75-9	0.63	0.0028	highAT	highAT
176	Dibutyl adipate	105-99-7	3.64	0.0141	highAT	highAT
177	<i>p</i> -Bromoaniline	106-40-1	47.50	0.2760	highAT	highAT
178	<i>p</i> -Xylene	106-42-3	8.87	0.0835	highAT	highAT
179	4-Methylphenol (<i>p</i> -cresol)	106-44-5	16.50	0.1530	highAT	highAT
180	4-Chloroaniline	106-47-8	31.40	0.2460	highAT	highAT
181	4-Chlorophenol	106-48-9	6.11	0.0475	highAT	highAT
182	4-Toluidine	106-49-0	160	1.4930	lowAT	lowAT
183	Isobutyl acrylate	106-63-8	2.10	0.0164	highAT	highAT
184	1-Bromopropane	106-94-5	67.30	0.5470	lowAT	highAT
185	Acrolein	107-02-8	0.017	0.0003	highAT	lowAT
186	1,2-Dichloroethane	107-06-2	136	1.3740	lowAT	lowAT
187	2-Chloroethanol	107-07-3	53.70	0.6670	lowAT	lowAT
188	Propylamine	107-10-8	308	5.2110	lowAT	lowAT
189	Propionitrile	107-12-0	1520	27.6000	lowAT	lowAT
190	Chloroacetonitrile	107-14-2	1.35	0.0178	highAT	lowAT
191	Ethylenediamine	107-15-3	220	3.6610	lowAT	lowAT
192	Allyl alcohol	107-18-6	0.32	0.0055	highAT	lowAT
193	2-Propyn-1-ol	107-19-7	1.48	0.0264	highAT	lowAT

TOX class model, training set: continued

No.	Substance chemical name	Substance CASRN	LC ₅₀ (mg/l)	LC ₅₀ (mmol/l)	Exp. Tox class	Pred. Tox class
194	Acetaldoxime	107-29-9	76	1.2870	lowAT	lowAT
195	2-Methyl-2,4-pentanediol	107-41-5	10700	90.5400	lowAT	lowAT
196	tert-Octylamine	107-45-9	24.60	0.1900	highAT	highAT
197	tert-Butyl sulfide	107-47-1	29.10	0.1990	highAT	highAT
198	2-Pentanone	107-87-9	1240	14.4000	lowAT	lowAT
199	4-Methyl-2-pentanone	108-10-1	522	5.2120	lowAT	lowAT
200	Isopropyl ether	108-20-3	786	7.6930	lowAT	lowAT
201	Toluene	108-88-3	33.90	0.3680	highAT	highAT
202	4-Picoline	108-89-4	403	4.3270	lowAT	lowAT
203	Chlorobenzene	108-90-7	16.90	0.1500	highAT	highAT
204	Cyclohexanol	108-93-0	704	7.0290	lowAT	lowAT
205	Cyclohexanone	108-94-1	621	6.3270	lowAT	lowAT
206	Phenol	108-95-2	32.70	0.3470	highAT	highAT
207	3-Picoline	108-99-6	144	1.5460	lowAT	lowAT
208	1-Methylpiperazine	109-01-3	2300	22.9600	lowAT	lowAT
209	2-Picoline	109-06-8	897	9.6320	lowAT	lowAT
210	2-Methylpiperazine	109-07-9	2240	22.3600	lowAT	lowAT
211	Propyl acetate	109-60-4	60	0.5870	lowAT	lowAT
212	1,3-Dibromopropane	109-64-8	2.09	0.0104	highAT	highAT
213	1-Bromobutane	109-65-9	36.70	0.2680	highAT	highAT
214	Butylamine	109-73-9	268	3.6640	lowAT	lowAT
215	Allyl cyanide	109-75-1	182	2.7130	lowAT	lowAT

TOX class model, training set: continued

No.	Substance chemical name	Substance CASRN	LC ₅₀ (mg/l)	LC ₅₀ (mmol/l)	Exp. class	Tox class	Pred. class	Tox class
216	1,3-Diaminopropane	109-76-2	1190	16.0500	lowAT	lowAT	lowAT	lowAT
217	Malonitrile	109-77-3	0.56	0.0085	highAT	highAT	lowAT	lowAT
218	2-Methoxyethylamine	109-85-3	524	6.9760	lowAT	lowAT	lowAT	lowAT
219	Diethylamine	109-89-7	855	11.6900	lowAT	lowAT	lowAT	lowAT
220	Pyrrrole	109-97-7	210	3.1300	lowAT	lowAT	lowAT	lowAT
221	Tetrahydrofuran	109-99-9	2160	29.9600	lowAT	lowAT	lowAT	lowAT
222	Furan	110-00-9	61	0.8960	lowAT	lowAT	lowAT	lowAT
223	<i>t</i> -Butyl disulfide	110-06-5	1.37	0.0077	highAT	highAT	highAT	highAT
224	5-Methyl-2-hexanone	110-12-3	159	1.3920	lowAT	lowAT	highAT	highAT
225	Dietyl sebacate	110-40-7	2.72	0.0105	highAT	highAT	highAT	highAT
226	2-Heptanone	110-43-0	131	1.1470	lowAT	lowAT	highAT	highAT
227	Hexane	110-54-3	2.50	0.0290	highAT	highAT	highAT	highAT
228	1,4-Dichlorobutane	110-56-5	51.60	0.4060	highAT	highAT	highAT	highAT
229	Amylamine	110-58-7	177	2.0310	lowAT	lowAT	lowAT	lowAT
230	Valeraldehyde	110-62-3	12.90	0.1500	highAT	highAT	highAT	highAT
231	2-Butyne-1,4-diol	110-65-6	53.60	0.6230	lowAT	lowAT	lowAT	lowAT
232	2-(Ethylamino)ethanol	110-73-6	1480	16.6000	lowAT	lowAT	lowAT	lowAT
233	Cyclohexane	110-82-7	4.53	0.0538	highAT	highAT	highAT	highAT
234	Pyridine	110-86-1	99.80	1.2620	lowAT	lowAT	lowAT	lowAT
235	<i>s</i> -Trioxane	110-88-3	5950	66.0500	lowAT	lowAT	lowAT	lowAT
236	6-Methyl-5-hepten-2-one	110-93-0	85.70	0.6790	lowAT	lowAT	lowAT	lowAT
237	2-Octanone	111-13-7	36	0.2810	highAT	highAT	highAT	highAT

TOX class model, training set: continued

No.	Substance chemical name	Substance CASRN	LC ₅₀ (mg/l)	LC ₅₀ (mmol/l)	Exp. Tox class	Pred. Tox class
238	2-Ethoxyethyl acetate	111-15-9	42.10	0.3190	highAT	lowAT
239	1-Bromohexane	111-25-1	3.45	0.0209	highAT	highAT
240	Hexylamine	111-26-2	56.60	0.5590	lowAT	lowAT
242	Diet hanolamine	111-42-2	4710	44.8000	lowAT	lowAT
243	2-Hydroxyethyl ether	111-46-6	75200	708.6000	lowAT	lowAT
244	<i>n</i> -Propyl sulfide	111-47-7	21.70	0.1840	highAT	highAT
245	<i>n</i> -Heptylamine	111-68-2	21.80	0.1890	highAT	highAT
246	1,4-Dicyanobutane	111-69-3	1930	17.8500	lowAT	lowAT
247	1-Heptanol	111-70-6	34.50	0.2970	highAT	highAT
248	1-Bromooctane	111-83-1	0.8380	0.0043	highAT	highAT
249	Octylamine	111-86-4	5.19	0.0402	highAT	highAT
250	1-Octanol	111-87-5	13.50	0.1040	highAT	highAT
251	2-(2-Ethoxyethoxy)ethanol	111-90-0	26500	197.5000	lowAT	lowAT
252	Nonanoic acid	112-05-0	104	0.6570	lowAT	highAT
253	2-Undecanone	112-12-9	1.50	0.0088	highAT	highAT
254	Nonylamine	112-20-9	2.16	0.0150	highAT	highAT
255	Triethylene glycol	112-27-6	68900	458.8000	lowAT	lowAT
256	1-Decanol	112-30-1	2.40	0.0152	highAT	highAT
258	Propoxur (Baygon)	114-26-1	8.80	0.0421	highAT	highAT
259	2-Methyl-3-butyn-2-ol	115-19-5	3290	39.1100	lowAT	lowAT
260	2,2,2-Trichloroethanol	115-20-8	299	2.0010	lowAT	lowAT

TOX class model, training set: continued

No.	Substance chemical name	Substance CASRN	LC ₅₀ (mg/l)	LC ₅₀ (mmol/l)	Exp. class	Tox class	Pred. class
261	Dicofol (Kelthane)	115-32-2	0.6030	0.0017	highAT	highAT	highAT
262	Triphenyl phosphate	115-86-6	0.87	0.0027	highAT	highAT	highAT
264	Aldicarb	116-06-3	0.8610	0.0045	highAT	highAT	highAT
267	Phenyl salicylate	118-55-8	1.18	0.0055	highAT	highAT	highAT
268	Ethyl salicylate	118-61-6	20.20	0.1220	highAT	highAT	highAT
269	2,4,6-Tribromophenol	118-79-6	6.54	0.0198	highAT	highAT	highAT
270	4-Amino-2-nitrophenol	119-34-6	36.20	0.2350	highAT	highAT	highAT
271	Benzophenone	119-61-9	14.70	0.0807	highAT	highAT	highAT
272	N-Phenyldiethanolamine	120-07-0	735	4.0560	lowAT	lowAT	lowAT
273	4-(Dietylamino)benzaldehyde	120-21-8	23.90	0.1350	highAT	highAT	highAT
274	Catechol	120-80-9	9.22	0.0837	highAT	highAT	highAT
275	1,2,4-Trichlorobenzene	120-82-1	2.99	0.0165	highAT	highAT	highAT
276	2,4-Dichlorophenol	120-83-2	7.75	0.0475	highAT	highAT	highAT
277	2,4-Dinitrotoluene	121-14-2	24.30	0.1330	highAT	highAT	highAT
278	3-Ethoxy-4-hydroxybenzaldehyde	121-32-4	87.60	0.5270	lowAT	lowAT	lowAT
279	Vanillin	121-33-5	83.80	0.5510	lowAT	lowAT	lowAT
280	N,N-Dimethylaniline	121-69-7	64.10	0.5290	lowAT	lowAT	highAT
281	1-Chloro-3-nitrobenzene	121-73-3	18.80	0.1190	highAT	highAT	highAT
283	2-Chloro-4-nitroaniline	121-87-9	20.10	0.1160	highAT	highAT	highAT
284	<i>p</i> -Isopropyl benzaldehyde	122-03-2	6.62	0.0447	highAT	highAT	highAT
285	Diphenylamine	122-39-4	3.79	0.0224	highAT	highAT	highAT

TOX class model, training set: continued

No.	Substance chemical name	Substance CASRN	LC ₅₀ (mg/l)	LC ₅₀ (mmol/l)	Exp. Tox class	Pred. Tox class
286	2-Phenoxyethanol	122-99-6	344	2.4900	lowAT	lowAT
287	4-Ethylphenol	123-07-9	10.40	0.0851	highAT	highAT
288	2-Methylvaleraldehyde	123-15-9	18.80	0.1880	highAT	highAT
289	2,4-Pentanedione	123-54-6	135	1.3480	lowAT	lowAT
290	Ethyl hexanoate	123-66-0	8.90	0.0617	highAT	highAT
291	Butanal	123-72-8	14.70	0.2040	highAT	lowAT
292	Butyl acetate	123-86-4	18.00	0.1550	highAT	lowAT
293	1,4-Dioxane	123-91-1	10300	116.9000	lowAT	lowAT
294	Dodecylamine	124-22-1	0.1030	0.0006	highAT	highAT
296	Tributyl phosphate	126-73-8	9.48	0.0356	highAT	highAT
297	5,5-Dimethyl-1,3-cyclohexanedione	126-81-8	11500	82.0400	lowAT	lowAT
298	1-Chloro-2-propanol	127-00-4	245	2.5910	lowAT	lowAT
299	Tetrachloroethylene	127-18-4	16.50	0.0995	highAT	highAT
300	2-Phenyl-3-butyne-2-ol	127-66-2	113	0.7730	lowAT	highAT
301	2,6-Di- <i>tert</i> -butyl-4-methylphenol	28-37-0	0.3630	0.0017	highAT	highAT
302	Saccharin sodium salt hydrate	82385-42-0	8300	82.0000	lowAT	lowAT
303	Dibenzofuran	132-64-9	1.50	0.0089	highAT	highAT
304	Phenyl 4-aminosalicylate	133-11-9	4.76	0.0208	highAT	highAT
305	N,N-Diethyl- <i>m</i> -toluamide	134-62-3	110	0.5750	lowAT	highAT
306	Propionic acid, sodium salt	137-40-6	4790	49.8600	lowAT	lowAT
307	1-(2-Aminoethyl)piperazine	140-31-8	2190	16.9500	lowAT	lowAT
308	Dibutyl succinate	141-03-7	4.46	0.0194	highAT	highAT

TOX class model, training set: continued

No.	Substance chemical name	Substance CASRN	LC ₅₀ (mg/l)	LC ₅₀ (mmol/l)	Exp. class	Tox class	Pred. class
309	Diethyl adipate	141-28-6	18.00	0.0899	highAT	highAT	highAT
310	2-Aminoethanol	141-43-5	2070	33.8900	lowAT	lowAT	lowAT
311	Ethyl acetate	141-78-6	230	2.6100	lowAT	lowAT	lowAT
312	2,6-Dimethyl morpholine	141-91-3	387	3.3600	lowAT	lowAT	lowAT
313	<i>m</i> -Diethylbenzene	141-93-5	4.15	0.0309	highAT	highAT	highAT
314	1,3-Dichloropropane	142-28-9	111	0.9820	lowAT	lowAT	lowAT
315	Hexanoic acid	142-62-1	320	2.7550	lowAT	lowAT	highAT
316	Hexyl acetate	142-92-7	4.40	0.0305	highAT	highAT	highAT
317	Butyl ether	142-96-1	32.30	0.2480	highAT	highAT	highAT
318	1-Nonanol	143-08-8	5.70	0.0395	highAT	highAT	highAT
319	Di- <i>n</i> -hexylamine	143-16-8	0.78	0.0042	highAT	highAT	highAT
321	<i>o</i> -Vanillin	148-53-8	2.40	0.0158	highAT	highAT	highAT
322	3-Methoxyphenol	150-19-6	74	0.5960	lowAT	lowAT	lowAT
323	4-Methoxyphenol	150-76-5	110	0.8860	lowAT	lowAT	lowAT
324	<i>p</i> -Dimethoxybenzene	150-78-7	117	0.8470	lowAT	highAT	highAT
325	2,3-Benzofuran	271-89-6	14.00	0.1190	highAT	highAT	highAT
326	1,4-Diazabicyclo[2.2.2]octane	280-57-9	1730	15.4200	lowAT	lowAT	lowAT
327	Adamantane	281-23-2	0.28	0.0021	highAT	highAT	highAT
329	Secobarbital, sodium salt	309-43-3	23.60	0.0907	highAT	highAT	highAT
330	Bromacil	314-40-9	186	0.7120	lowAT	highAT	highAT
331	2,5-Dinitrophenol	329-71-5	3.36	0.0182	highAT	highAT	highAT

TOX class model, training set: continued

No.	Substance chemical name	Substance CASRN	LC ₅₀ (mg/l)	LC ₅₀ (mmol/l)	Exp. Tox class	Pred. Tox class
332	Diuron	330-54-1	14.20	0.0609	highAT	highAT
333	p-Fluorophenyl ether	330-93-8	1.13	0.0055	highAT	highAT
335	1-Fluoro-4-nitrobenzene	350-46-9	28.40	0.2010	highAT	highAT
336	alpha, alpha, alpha-Trifluoro-m-tolunitrile	368-77-4	47.70	0.2790	highAT	highAT
337	4-Fluoroaniline	371-40-4	16.90	0.1520	highAT	highAT
338	2-Chloro-6-fluorobenzaldehyde	387-45-1	9.41	0.0593	highAT	highAT
339	alpha, alpha, alpha-4-Tetrafluoro-o-toluidine	393-39-5	29.60	0.1650	highAT	highAT
340	2-Fluorobenzaldehyde	446-52-6	1.35	0.0109	highAT	highAT
341	alpha, alpha, alpha-Trifluoro-o-tolunitrile	447-60-9	42.20	0.2470	highAT	highAT
342	alpha, alpha, alpha-Trifluoro-m-tolualdehyde	454-89-7	0.9240	0.0053	highAT	highAT
343	4-Fluoro-N-methylaniline	459-59-6	38.40	0.3070	highAT	highAT
345	[(1S)-endo](-)-Borneol	464-45-9	63.20	0.4100	highAT	highAT
346	(1S)-(-)-Camphor	464-48-2	17.00	0.1120	highAT	highAT
347	Cineole	470-82-6	102	0.6610	lowAT	highAT
348	Neoabietic acid	471-77-2	1.49	0.0049	highAT	highAT
350	2,3-Dihydrobenzofuran	496-16-2	81.70	0.6800	lowAT	lowAT
351	exo-Norborneol	497-37-0	228	2.0330	lowAT	lowAT
352	Norbornylene	498-66-8	10.00	0.1060	highAT	highAT
353	2,6-Pyridinedicarboxylic acid	499-83-2	322	1.9270	lowAT	lowAT
354	3-Pyridinecarboxaldehyde	500-22-1	16.40	0.1530	highAT	lowAT
355	5-Nonanone	502-56-7	31.00	0.2180	highAT	highAT

TOX class model, training set: continued

No.	Substance chemical name	Substance CASRN	LC ₅₀ (mg/l)	LC ₅₀ (mmol/l)	Exp. Tox class	Pred. Tox class
356	2,3-Dimethyl-1,3-butadiene	513-81-5	6.91	0.0841	highAT	highAT
357	Abietic acid	514-10-3	2.38	0.0079	highAT	highAT
358	Flavone	525-82-6	3.50	0.0157	highAT	highAT
359	2,4,6-Trimethylphenol	527-60-6	13.00	0.0954	highAT	highAT
360	<i>o</i> -Tolunitrile	529-19-1	44.70	0.3820	highAT	highAT
361	<i>o</i> -Tolualdehyde	529-20-4	52.90	0.4400	highAT	highAT
362	Benzoic acid, sodium salt	532-32-1	484	3.3590	lowAT	highAT
363	4,6-Dinitro- <i>o</i> -cresol	534-52-1	1.73	0.0087	highAT	highAT
364	Amylbenzene	538-68-1	1.71	0.0115	highAT	highAT
365	<i>tert</i> -Butyl acetate	540-88-5	327	2.8150	lowAT	lowAT
366	1,3-Dichlorobenzene	541-73-1	8.03	0.0546	highAT	highAT
367	1,3-Dichloropropene	542-75-6	0.2390	0.0022	highAT	lowAT
368	<i>n</i> -Butyl sulfide	544-40-1	3.58	0.0245	highAT	highAT
369	2'-Hydroxy-4'-methoxyacetophenone	552-41-0	69.50	0.4180	highAT	highAT
370	<i>o</i> -Nitrobenzaldehyde	552-89-6	14.40	0.0953	highAT	highAT
371	4-Nitrobenzaldehyde	555-16-8	10.10	0.0668	highAT	highAT
372	3-Methyl-2-butanone	563-80-4	864	10.0300	lowAT	lowAT
373	2,6-Dinitrophenol	573-56-8	39.70	0.2160	highAT	highAT
374	1,2-Dibromobenzene	583-53-9	4.05	0.0172	highAT	highAT
375	<i>N</i> -Allylaniline	589-09-3	35.90	0.2700	highAT	highAT
376	4-Ethylaniline	589-16-2	73	0.6020	lowAT	highAT
377	Isovaleraldehyde	590-86-3	3.25	0.0377	highAT	lowAT

TOX class model, training set: continued

No.	Substance chemical name	Substance CASRN	LC ₅₀ (mg/l)	LC ₅₀ (mmol/l)	Exp. Tox class	Pred. Tox class
378	2-Hexanone	591-78-6	428	4.2730	lowAT	lowAT
379	2,4-Hexadiene	592-46-1	20.00	0.2430	highAT	highAT
380	2-Tridecanone	593-08-8	0.36	0.0018	highAT	highAT
382	Manool	596-85-0	0.12	0.0004	highAT	highAT
384	1,2-Dimethylpropylamine	598-74-3	284	3.2580	lowAT	lowAT
385	2,4-Dimethyl-3-pentanol	600-36-2	163	1.4030	lowAT	highAT
387	N,N-Diphenylformamide	607-00-1	30.40	0.1540	highAT	highAT
388	Dietyl benzylmalonate	607-81-8	5.43	0.0217	highAT	highAT
389	Pentabromophenol	608-71-9	0.0930	0.0002	highAT	highAT
390	2,4,6-Triiodophenol	609-23-4	1.21	0.0026	highAT	highAT
391	2,4-Dimethoxybenzaldehyde	613-45-6	20.10	0.1210	highAT	highAT
392	2-Acetamidophenol	614-80-2	27.00	0.1790	highAT	lowAT
393	2-Chloro-4-methylaniline	615-65-6	35.90	0.2540	highAT	highAT
394	4-Ethoxy-2-nitroaniline	616-86-4	26.00	0.1430	highAT	highAT
395	Methyl <i>p</i> -nitrobenzoate	619-50-1	23.80	0.1310	highAT	highAT
396	4-Nitrobenzamide	619-80-7	133	0.8010	lowAT	lowAT
397	4-Nitrophenyl phenyl ether	620-88-2	2.65	0.0123	highAT	highAT
398	Benzyl sulfoxide	621-08-9	80.10	0.3480	highAT	highAT
399	3-Acetamidophenol	621-42-1	1130	7.4750	lowAT	lowAT
400	4-(2-Hydroxyethyl)morpholine	622-40-2	2710	20.6600	lowAT	lowAT
401	<i>alpha, alpha'</i> -Dichloro- <i>p</i> -xylene	623-25-6	0.0390	0.0002	highAT	highAT
403	2,5-Dimethylfuran	625-86-5	71.10	0.7400	lowAT	lowAT

TOX class model, training set: continued

No.	Substance chemical name	Substance CASRN	LC ₅₀ (mg/l)	LC ₅₀ (mmol/l)	Exp. Tox class	Pred. Tox class
404	1,5-Dichloropentane	628-76-2	25.30	0.1790	highAT	highAT
405	1-Bromoheptane	629-04-9	1.47	0.0082	highAT	highAT
406	Propyl disulfide	629-19-6	2.62	0.0170	highAT	highAT
407	1,6-Dicyanohexane	629-40-3	528	3.8770	lowAT	lowAT
408	2,3,4-Trichloroaniline	634-67-3	3.64	0.0185	highAT	highAT
409	5-Chlorosalicylaldehyde	635-93-8	0.77	0.0049	highAT	highAT
410	4-Propylphenol	645-56-7	11.00	0.0808	highAT	highAT
411	Pentafluorobenzaldehyde	653-37-2	1.10	0.0056	highAT	highAT
412	2,2-Dichloroacetamide	683-72-7	241	1.8830	lowAT	lowAT
413	1-Methyl heptylamine	693-16-3	5.19	0.0402	highAT	highAT
414	2-Decanone	693-54-9	4.83	0.0309	highAT	highAT
415	Pentyl ether	693-65-2	3.14	0.0198	highAT	highAT
416	4-Methylloxazole	693-93-6	1390	16.7300	lowAT	lowAT
417	2-Methylimidazole	693-98-1	286	3.4830	lowAT	lowAT
418	2-Adamantanone	700-58-3	60.80	0.4050	highAT	lowAT
419	<i>gamma</i> -Decanolactone	706-14-9	18.00	0.1060	highAT	highAT
420	4,6-Dimethoxy-2-hydroxybenzaldehyde	708-76-9	2.68	0.0147	highAT	highAT
421	Propanil	709-98-8	8.60	0.0394	highAT	highAT
422	2,4,6-Tri- <i>tert</i> -butylphenol	732-26-3	0.0609	0.0002	highAT	highAT
423	3,4-Dichloro-1-butene	760-23-6	8.18	0.0654	highAT	highAT
424	N,N-Dibutylformamide	761-65-9	89.30	0.5680	lowAT	highAT
425	2-Butyn-1-ol	764-01-2	10.10	0.1440	highAT	lowAT

TOX class model, training set: continued

No.	Substance chemical name	Substance CASRN	LC ₅₀ (mg/l)	LC ₅₀ (mmol/l)	Exp. Tox class	Pred. Tox class
426	2,5-Dimethyl-2,4-hexadiene	764-13-6	3.78	0.0343	highAT	highAT
427	1-Adamantanamine	768-94-5	25.00	0.1650	highAT	highAT
428	3-Cyano-4,6-dimethyl-2-hydroxypyridine	769-28-8	157.00	1.0600	lowAT	highAT
429	2,3,4,5,6-Pentafluoroaniline	771-60-8	37.10	0.2030	highAT	highAT
431	Triphenylphosphine oxide	791-28-6	53.70	0.1930	highAT	highAT
433	2-Hydroxyethyl acrylate	818-61-1	4.80	0.0414	highAT	lowAT
434	1-Octyne-3-ol	818-72-4	0.4130	0.0033	highAT	highAT
435	2-Nonanone	821-55-6	15.20	0.1070	highAT	highAT
436	trans-1,2-Dichlorocyclohexane	822-86-6	18.40	0.1200	highAT	highAT
437	p-Phenoxyphenol	831-82-3	4.95	0.0266	highAT	highAT
438	2-Hydroxyethyl methacrylate	868-77-9	227	1.7440	lowAT	lowAT
439	3-Bromothiophene	872-31-1	6.19	0.0380	highAT	highAT
440	2,4-Dichlorobenzaldehyde	874-42-0	1.80	0.0103	highAT	highAT
443	Ethyl 3-amino benzoate methanesulfonic acid salt	886-86-2	79.00	0.3020	highAT	highAT
444	1,1,1,3,3,3-Hexafluoro-2-propanol	920-66-1	244	1.4520	lowAT	lowAT
445	1,5-Hexadien-3-ol	924-41-4	38.10	0.3880	highAT	lowAT
447	cis-3-Hexen-1-ol	928-96-1	381	3.8040	lowAT	lowAT
448	trans-3-Hexen-1-ol	928-97-2	271	2.7060	lowAT	lowAT
449	2-Acetyl-1-methylpyrrole	932-16-1	157	1.2750	lowAT	lowAT
450	4-Phenylpyridine	939-23-1	16.10	0.1040	highAT	highAT
451	Phenyl sulfoxide	945-51-7	87.30	0.4320	highAT	highAT
453	2-Hydroxypropyl acrylate	999-61-1	3.34	0.0260	highAT	lowAT

TOX class model, training set: continued

No.	Substance chemical name	Substance CASRN	LC ₅₀ (mg/l)	LC ₅₀ (mmol/l)	Exp. class	Tox class	Pred. class
454	2-Amino-5-bromopyridine	1072-97-5	177	1.0230	lowAT	lowAT	lowAT
455	Diethyl benzylphosphonate	1080-32-6	336	1.4720	lowAT	lowAT	highAT
457	4-Acetyl pyridine	1122-54-9	168	1.3870	lowAT	lowAT	lowAT
458	Methyl <i>p</i> -chlorobenzoate	1126-46-1	11.00	0.0640	highAT	highAT	highAT
459	Butyl phenyl ether	1126-79-0	5.77	0.0380	highAT	highAT	highAT
460	Methyl 4-cyanobenzoate	1129-35-7	46.80	0.2900	highAT	highAT	highAT
461	Tetrachlorocatechol	1198-55-6	1.27	0.0051	highAT	highAT	highAT
462	<i>alpha</i> -Bromo-2',5'-dimethoxyacetophenone	1204-21-3	0.0660	0.0026	highAT	highAT	highAT
464	3,4-Dimethyl-1-pentyn-3-ol	1482-15-1	205	1.8280	lowAT	lowAT	lowAT
465	N-Vinylcarbazole	1484-13-5	0.0032	0.0000	highAT	highAT	highAT
466	3-Benzylloxylaniline	1484-26-0	9.14	0.0459	highAT	highAT	highAT
467	Carbofuran	1563-66-2	0.8440	0.0038	highAT	highAT	highAT
468	<i>tert</i> -Butyl methyl ether	1634-04-4	672	7.6230	lowAT	lowAT	lowAT
469	1,9-Decadiene	1647-16-1	0.29	0.0021	highAT	highAT	highAT
470	<i>p</i> -Phenylazophenol	1689-82-3	1.17	0.0060	highAT	highAT	highAT
471	3,5-Diodo-4-hydroxybenzotrile	1689-83-4	6.80	0.0183	highAT	highAT	highAT
472	3,5-Dibromo-4-hydroxybenzotrile	1689-84-5	12.60	0.0455	highAT	highAT	highAT
473	Dehydroabietic acid	1740-19-8	2.10	0.0070	highAT	highAT	highAT
474	2-Allylphenol	1745-81-9	15.00	0.1120	highAT	highAT	highAT
475	5-Bromosalicylaldehyde	1761-61-1	1.30	0.0065	highAT	highAT	highAT
476	3-Chloro-2-chloromethyl-1-propene	1871-57-4	0.19	0.0015	highAT	highAT	highAT
477	3,5-Dichloro-4-hydroxybenzotrile	1891-95-8	24.30	0.1290	highAT	highAT	highAT

TOX class model, training set: continued

No.	Substance chemical name	Substance CASRN	LC ₅₀ (mg/l)	LC ₅₀ (mmol/l)	Exp. Tox class	Pred. Tox class
478	Di- <i>n</i> -butylterephthalate	1962-75-0	0.59	0.0021	<i>high</i> AT	<i>high</i> AT
479	4,4'-Dihydroxydiphenylether	1965-09-9	5.78	0.0286	<i>high</i> AT	<i>high</i> AT
480	2,6-Dichloro benzamide	2008-58-4	469	2.4680	<i>low</i> AT	<i>high</i> AT
481	<i>n</i> -Decylamine	2016-57-1	1.03	0.0066	<i>high</i> AT	<i>high</i> AT
482	Aminocarb	2032-59-9	1.95	0.0094	<i>high</i> AT	<i>high</i> AT
483	2,4,5-Tribromoimidazole	2034-22-2	6.12	0.0201	<i>high</i> AT	<i>high</i> AT
485	(+)-4-Pentyn-2-ol	2117-11-5	35.10	0.4170	<i>high</i> AT	<i>low</i> AT
486	4-Chlorocatechol	2138-22-9	1.58	0.0109	<i>high</i> AT	<i>high</i> AT
487	Methyl 2,4-dihydroxybenzoate	2150-47-2	45.80	0.2720	<i>high</i> AT	<i>high</i> AT
489	Pentachloropyridine	2176-62-7	0.47	0.0019	<i>high</i> AT	<i>high</i> AT
490	(1R,2S,5R)-(-)-menthol	2216-51-5	18.90	0.1210	<i>high</i> AT	<i>high</i> AT
491	1-(<i>p</i> -Toluenesulfonyl)imidazole	2232-08-8	41.80	0.1880	<i>high</i> AT	<i>high</i> AT
492	2,4'-Dichloroacetophenone	2234-16-4	11.70	0.0689	<i>high</i> AT	<i>high</i> AT
493	<i>n</i> -Octyl cyanide	2243-27-8	5.25	0.0377	<i>high</i> AT	<i>high</i> AT
494	<i>alpha, alpha, alpha</i> -4-Tetrafluoro- <i>m</i> -toluidine	2357-47-3	30.10	0.1680	<i>high</i> AT	<i>high</i> AT
495	<i>trans</i> -2-Phenyl-1-cyclohexanol	2362-61-0	44.40	0.2520	<i>high</i> AT	<i>high</i> AT
496	2-Ethoxyethyl methacrylate	2370-63-0	27.70	0.1750	<i>high</i> AT	<i>high</i> AT
497	2,3,6-Trimethylphenol	2416-94-6	8.20	0.0602	<i>high</i> AT	<i>high</i> AT
498	<i>n</i> -Undecyl cyanide	2437-25-4	0.43	0.0024	<i>high</i> AT	<i>high</i> AT
499	<i>o</i> -Methoxybenzamide	2439-77-2	120	0.7940	<i>low</i> AT	<i>low</i> AT
501	Tetrahydrofurfuryl methacrylate	2455-24-5	34.70	0.2040	<i>high</i> AT	<i>high</i> AT
502	4,5-Dichloroguaiacol	2460-49-3	4.47	0.0232	<i>high</i> AT	<i>high</i> AT

TOX class model, training set: continued

No.	Substance chemical name	Substance CASRN	LC ₅₀ (mg/l)	LC ₅₀ (mmol/l)	Exp. Tox class	Pred. Tox class
503	Benzyl methacrylate	2495-37-6	4.67	0.0265	highAT	highAT
504	hexyl acrylate	2499-95-8	1.11	0.0071	highAT	highAT
505	<i>p</i> -(<i>tert</i> -butyl)-Phenyl-N-methylcarbamate	2626-83-7	10.00	0.0482	highAT	highAT
507	1-Benzylpiperazine	2759-28-6	47.40	0.2690	highAT	lowAT
508	3-(3-Pyridyl)-1-propanol	2859-67-8	150	1.0930	lowAT	lowAT
509	Tridecylamine	2869-34-3	0.0654	0.0003	highAT	highAT
510	2-Amino-4'-chloro benzophenone	2894-51-1	2.12	0.0092	highAT	lowAT
511	Methyl 2,5-dichloro benzoate	2905-69-3	14.00	0.0683	highAT	highAT
513	5-Bromovanillin	2973-76-4	59.70	0.2580	highAT	highAT
514	Cyclohexyl acrylate	3066-71-5	1.48	0.0096	highAT	highAT
516	Triethyl nitrilotricarboxylate	3206-31-3	13.30	0.0570	highAT	highAT
517	4,5-Dichlorocatechol	3428-24-8	0.89	0.0050	highAT	highAT
518	2,3,5,6-Tetrachloroaniline	3481-20-7	0.27	0.0012	highAT	highAT
519	2,6-Diphenylpyridine	3558-69-8	0.21	0.0009	highAT	highAT
520	1,3-Dichloro-4,6-dinitro benzene	3698-83-7	0.0456	0.0002	highAT	highAT
523	2,3-Dimethylvaleraldehyde	3944-76-1	16.00	0.1400	highAT	highAT
524	2-Decyn-1-ol	4117-14-0	1.07	0.0069	highAT	highAT
525	5-Chloro-2-pyridinol	4214-79-3	1140	8.8000	lowAT	lowAT
526	Isopropyl disulfide	4253-89-8	8.31	0.0553	highAT	highAT
527	2,4,5-Trimethoxybenzaldehyde	4460-86-0	49.50	0.2520	highAT	highAT
528	Isopropyl methacrylate	4655-34-9	38.00	0.2960	highAT	lowAT
529	1-Hexen-3-ol	4798-44-1	30.40	0.3040	highAT	lowAT

TOX class model, training set: continued

No.	Substance chemical name	Substance CASRN	LC ₅₀ (mg/l)	LC ₅₀ (mmol/l)	Exp. Tox class	Pred. Tox class
530	2,3,4,5-Tetrachlorophenol	4901-51-3	0.41	0.0018	highAT	highAT
531	1,2-bis(4-Pyridyl)ethane	4916-57-8	151	0.8200	lowAT	lowAT
532	1,3-Diethyl-2-thiobarbituric acid	5217-47-0	4510	22.5200	lowAT	lowAT
533	Dimethyl nitroterephthalate	5292-45-5	6.52	0.0273	highAT	highAT
534	5-Chloro-2-mercaptopbenzotiazole	5331-91-9	3.21	0.0159	highAT	highAT
535	Dimethyl aminoterephthalate	5372-81-6	8.94	0.0427	highAT	highAT
536	3,6-Dithiaoctane	5395-75-5	60.20	0.4010	highAT	highAT
537	Dimethylaminopropyl chloride, hydrochloride	5407-04-5	133	0.8410	lowAT	lowAT
538	4'-Chloro-3'-nitroacetophenone	5465-65-6	5.50	0.0276	highAT	highAT
539	2-Amino-4-chloro-6-methylpyrimidine	5600-21-5	141	0.9820	lowAT	highAT
540	2,6-Dimethoxytoluene	5673-07-4	20.20	0.1330	highAT	highAT
541	2-Dimethylaminopyridine	5683-33-0	127	1.0400	lowAT	lowAT
542	2,2-Dimethyl-1-propylamine	5813-64-9	475	5.4490	lowAT	lowAT
543	Isopimaric acid	5835-26-7	0.87	0.0029	highAT	highAT
544	2-Amino-5-chlorobenzonitrile	5922-60-1	28.60	0.1870	highAT	highAT
545	1,1,1-Trichloro-2-methyl-2-propanol(hydrate) (2:1)	6001-64-5	135	0.3620	highAT	highAT
546	2-Dodecanone	6175-49-1	1.18	0.0064	highAT	highAT
547	4-Dimethylaminocinnamaldehyde	6203-18-5	5.90	0.0367	highAT	highAT
549	1,3,5-Trichloro-2,4-dinitrobenzene	6284-83-9	0.2220	0.0008	highAT	highAT
550	2-Chloro-5-nitrobenzaldehyde	6361-21-3	3.87	0.0209	highAT	highAT
551	2-Chloro-6-methylbenzonitrile	6575-09-3	15.10	0.0996	highAT	highAT
552	2-Bromo-3-pyridinol	6602-32-0	469	2.6950	lowAT	lowAT

TOX class model, training set: continued

No.	Substance chemical name	Substance CASRN	LC ₅₀ (mg/l)	LC ₅₀ (mmol/l)	Exp. class	Tox class	Pred. class
553	2-Chloro-3-pyridinol	6636-78-8	622	4.8010	lowAT	lowAT	lowAT
554	Tripropargylamine	6921-29-5	296	2.2560	lowAT	lowAT	lowAT
555	N,N-bis(2,2-Diethoxyethyl)methylamine	6948-86-3	635	2.4110	lowAT	lowAT	lowAT
556	1,4-bis(3-Aminopropyl)piperazine	7209-38-3	3100	15.4700	lowAT	lowAT	lowAT
557	3-Hydroxy-3,7,11-trimethyl-1,6,10-dodecatriene	7212-44-4	1.43	0.0064	highAT	highAT	highAT
558	1-(2-Chloroethyl)pyrrolidine·HCl	7250-67-1	153	0.9000	lowAT	lowAT	lowAT
559	n-Undecylamine	7307-55-3	0.21	0.0012	highAT	highAT	highAT
560	1-Heptyn-3-ol	7383-19-9	1.76	0.0157	highAT	highAT	highAT
561	p-Ethoxybenzaldehyde	10031-82-0	28.10	0.1870	highAT	highAT	highAT
562	[1(R)-endo](+)-3-Bromocamphor	10293-06-8	68.50	0.2960	highAT	highAT	highAT
563	Resmethrin	10453-86-8	0.0062	0.0000	highAT	highAT	highAT
565	alpha, alpha, alpha', alpha'-Tetrabromo-o-xylene	13209-15-9	0.4370	0.0010	lowAT	highAT	highAT
566	2',3',4'-Trichloroacetophenone	13608-87-2	2.00	0.0090	highAT	highAT	highAT
567	2',3',4'-Trimethoxyacetophenone	13909-73-4	199	0.9470	lowAT	highAT	highAT
568	Diethyl chloromalonate	14064-10-9	0.95	0.0049	highAT	highAT	highAT
569	N-Ethylbenzylamine	14321-27-8	57.10	0.4220	highAT	highAT	highAT
571	4-Bromophenyl 3-pyridyl ketone	14548-45-9	20.40	0.0778	highAT	highAT	highAT
572	4-Benzoylpyridine	14548-46-0	103	0.5620	lowAT	highAT	highAT
573	2,2,5,5-Tetramethyltetrahydrofuran	15045-43-9	168	1.3100	lowAT	highAT	highAT
574	3-Hydroxy-2-nitropyridine	15128-82-2	167	1.1920	lowAT	lowAT	lowAT
575	Alachlor	15972-60-8	5.00	0.0185	highAT	highAT	highAT
576	4-Octylaniline	16245-79-7	0.12	0.0006	highAT	highAT	highAT

TOX class model, training set: continued

No.	Substance chemical name	Substance CASRN	LC ₅₀ (mg/l)	LC ₅₀ (mmol/l)	Exp. Tox class	Pred. Tox class
577	Methomyl	16752-77-5	2.11	0.0130	highAT	lowAT
578	6-Chloro-2-pyridinol	16879-02-0	214	1.6520	lowAT	highAT
579	3-Amino-5,6-dimethyl-1,2,4-triazine	17584-12-2	952	7.6680	lowAT	lowAT
581	4-(Dietylamino)salicylaldehyde	17754-90-4	5.36	0.0277	highAT	highAT
582	6-Chloro-2-picoline	18368-63-3	232	1.8190	lowAT	highAT
583	3,6-Dimethyl-1-heptyn-3-ol	19549-98-5	49.00	0.3490	highAT	highAT
584	2,4,5-Trimethylloxazole	20662-84-4	449	4.0400	lowAT	lowAT
585	4-Dimethylamino-3-methyl-2-butanone	22104-62-7	8.50	0.0658	highAT	lowAT
588	Oxamyl	23135-22-0	6.78	0.0309	highAT	highAT
589	2,6-Diisopropylaniline	24544-04-5	15.30	0.0863	highAT	highAT
593	2-Methyl-3,3,4,4-tetrafluoro-2-butanol	29553-26-2	582	3.6350	lowAT	lowAT
594	(+)- <i>sec</i> -Butylamine	13952-84-6	275	3.7600	lowAT	lowAT
595	N-(3-Methoxypropyl)-3,4,5-trimethoxybenzylamine	34274-04-9	136	0.5050	lowAT	lowAT
596	2-(Bromomethyl)tetrahydro-2H-pyran	34723-82-5	205	1.1450	lowAT	highAT
597	4-Decylaniline	37529-30-9	0.0620	0.0003	highAT	highAT
598	4-Hexyloxylaniline	39905-57-2	3.01	0.0156	highAT	highAT
599	Methyl 4-chloro-2-nitrobenzoate	42087-80-9	27.70	0.1280	highAT	highAT
600	5-Hydroxy-2-nitrobenzaldehyde	42454-06-8	41.90	0.2510	highAT	highAT
601	Feuvalerate	51630-58-1	0.0051	0.0000	highAT	highAT
602	Permethrin	52645-53-1	0.0160	0.0000	highAT	highAT
603	3,8-Dithiadecane	54576-32-8	6.06	0.0340	highAT	highAT

TOX class model, training set: continued

No.	Substance chemical name	Substance CASRN	LC ₅₀ (mg/l)	LC ₅₀ (mmol/l)	Exp. Tox class	Pred. Tox class
604	2-(Octyloxy)-acetanilide	55792-61-5	0.45	0.0017	highAT	highAT
605	<i>p</i> -(tert-Butyl)benzamide	56108-12-4	31.90	0.1800	highAT	highAT
606	2,9-Dithiadecane	56348-40-4	10.10	0.0566	highAT	highAT
607	<i>dl</i> -3-Butyn-2-ol	65337-13-5	11.70	0.1670	highAT	lowAT
610	3-(4- <i>tert</i> -Butylphenoxy)benzaldehyde	69770-23-6	0.37	0.0015	highAT	highAT
611	Flucythrinate	70124-77-5	0.0002	0.0000	highAT	highAT
613	3-(3,4-Dichlorophenoxy)benzaldehyde	79124-76-8	0.30	0.0011	highAT	highAT
614	2,4-Dinitro-1-naphthol sodium salt dihydrate (Martius yellow)	101836-92-4	4.24	0.0145	highAT	highAT
615	<i>t</i> -Butylstyrene	1746-23-2	0.49	0.0031	highAT	highAT
617	Chloromethyl styrene	30030-25-2	0.31	0.0020	highAT	highAT

F.2 MOA class model: Training set

No.	Substance chemical name	Substance CASRN	Exp. class	MOA	Pred. class	MOA
4	<i>p</i> -Chlorophenyl- <i>o</i> -nitrophenyl ether	39145-47-6	A		A	A
6	Di- <i>n</i> -butylisophthlate	3126-90-7	A		A	A
7	1,1-Diphenyl-2-propyn-1-ol	3923-52-2	A		A	A
8	4,7-Dithia decane	22037-97-4	A		A	A
15	Phenobarbital	50-06-6	E		E	E
16	2,4-Dinitrophenol	51-28-5	F		F	F
17	Urethane	51-79-6	A		A	A
21	1,1-Dimethyl hydrazine	57-14-7	D		D	D
22	Pentobarbital sodium	57-33-0	E		E	E
23	Amobarbital	57-43-2	E		E	E
24	Caffeine	58-08-2	G		G	G
25	2-Methyl-1,4-naphthoquinone	58-27-5	D		D	D
27	4-Chloro-3-methyl phenol	59-50-7	B		B	B
28	Tolazoline hydrochloride	59-97-2	A		A	A
29	Amphetamine sulfate	60-13-9	G		G	G
30	Diethyl ether	60-29-7	A		A	A
31	Strychnine hemisulphate salt	60-41-3	G		G	G
32	Aniline	62-53-3	B		B	B
33	Carbaryl (sevin)	63-25-2	H		H	H
34	Ethanol	64-17-5	A		A	A
35	Nicotine sulfate	65-30-5	G		G	G

MOAclass model, training set: *continued*

No.	Substance chemical name	Substance CASRN	Exp. class	MOA	Pred. class	MOA
38	Hexanal	66-25-1	D		D	
41	Methanol	67-56-1	A		A	
42	2-Propanol	67-63-0	A		A	
46	Hexachloroethane	67-72-1	A		A	
48	4'-Aminopropiophenone	70-69-9	A		A	
49	1-Propanol	71-23-8	A		A	
50	<i>n</i> -Butanol	71-36-3	A		A	
51	1-Pentanol	71-41-0	A		A	
53	1,1,1-Trichloroethane	71-55-6	A		A	
54	Thiopental, sodium salt	71-73-8	E		E	
57	Dichloromethane	75-09-2	A		A	
60	2-Methyl-2-propanol	75-65-0	A		A	
63	Pentachloroethane	76-01-7	A		A	
65	3-Methyl-3-pentanol	77-74-7	A		A	
66	3-Methyl-1-pentyn-3-ol	77-75-8	A		A	
67	1-Ethynyl-cyclohexanol	78-27-3	A		A	
69	2-Methyl-1-propanol	78-83-1	A		A	
70	1,2-Dichloropropane	78-87-5	A		A	
71	1,2-Diaminopropane	78-90-0	B		B	
72	2-Butanol	78-92-2	A		A	
75	1,1,2-Trichloroethane	79-00-5	A		A	
76	Trichloroethylene	79-01-6	A		A	

MOAclass model, training set: continued

No.	Substance chemical name	Substance CASRN	Exp. class	MOA class	Pred. class	MOA
77	Methyl acetate	79-20-9	C		C	
78	1,1,2,2-Tetrachloroethane	79-34-5	A		A	
79	<i>beta</i> -Ionone	79-77-6	A		A	
85	3-Methylindole	83-34-1	D		A	
86	Rotenone	83-79-4	I		I	
95	Hexachloro-1,3-butadiene	87-68-3	D		D	
96	Pentachlorophenol	87-86-5	F		F	
97	2,4,6-Trichlorophenol	88-06-2	B		B	
98	3-Trifluoromethyl-4-nitrophenol	88-30-2	B		B	
103	1-Naphthol	90-15-3	B		B	
107	Naphthalene	91-20-3	A		A	
110	N,N-Diethylaniline	91-66-7	A		A	
111	2-(N-Ethyl- <i>m</i> -toluidino)ethanol	91-88-3	A		A	
114	Ethyl <i>p</i> -aminobenzoate	94-09-7	C		C	
118	<i>o</i> -Cresol	95-48-7	B		B	
120	2-Chloroaniline	95-51-2	B		B	
122	2-Chlorophenol	95-57-8	B		B	
125	3,4-Dichloroaniline	95-76-1	B		B	
126	Allyl methacrylate	96-05-9	D		D	
129	1,2,3-Trichloropropane	96-18-4	A		A	
132	2-(Diisopropylamino)ethanol	96-80-0	A		A	
133	2,4-Dinitroaniline	97-02-9	F		F	

MOAclass model, training set: *continued*

No.	Substance chemical name	Substance CASRN	Exp. class	MOA class	Pred. class	MOA
138	Acetophenone	98-86-2	A		A	A
142	N,N-Dimethyl- <i>p</i> -toluidine	99-97-8	A		A	A
144	<i>p</i> -Nitrophenol	100-02-7	B		B	B
147	N,N-Diethylethanolamine	100-37-8	A		A	A
150	Benzaldehyde	100-52-7	D		D	D
153	2-Cyanopyridine	100-70-9	B		B	B
154	2-Ethylpyridine	100-71-0	A		A	A
161	Benzyl- <i>tert</i> -butanol	103-05-9	A		A	A
169	2-Ethyl-1-hexanol	104-76-7	A		A	A
171	5-Ethyl-2-methylpyridine	104-90-5	A		A	A
174	2,4-Dimethylphenol	105-67-9	B		B	B
177	<i>p</i> -Bromoaniline	106-40-1	B		B	B
180	4-Chloroaniline	106-47-8	B		B	B
182	4-Toluidine	106-49-0	B		B	B
183	Isobutyl acrylate	106-63-8	D		D	D
184	1-Bromopropane	106-94-5	A		A	A
185	Acrolein	107-02-8	D		D	D
186	1,2-Dichloroethane	107-06-2	A		A	A
187	2-Chloroethanol	107-07-3	D		D	D
189	Propionitrile	107-12-0	A		A	A
192	Allyl alcohol	107-18-6	D		D	D
193	2-Propyn-1-ol	107-19-7	D		D	D

MOAclass model, training set: continued

No.	Substance chemical name	Substance CASRN	Exp. class	MOA	Pred. class	MOA
195	2-Methyl-2,4-pentanediol	107-41-5	A		A	A
199	4-Methyl-2-pentanone	108-10-1	A		A	A
200	Isopropyl ether	108-20-3	A		A	A
202	4-Picoline	108-89-4	B		A	A
204	Cyclohexanol	108-93-0	A		A	A
206	Phenol	108-95-2	B		B	B
209	2-Picoline	109-06-8	A		A	A
212	1,3-Dibromopropane	109-64-8	D		A	A
213	1-Bromobutane	109-65-9	A		A	A
216	1,3-Diaminopropane	109-76-2	B		B	B
226	2-Heptanone	110-43-0	A		A	A
228	1,4-Dichlorobutane	110-56-5	A		A	A
230	Valeraldehyde	110-62-3	D		D	D
231	2-Butyne-1,4-diol	110-65-6	D		D	D
234	Pyridine	110-86-1	B		B	B
237	2-Octanone	111-13-7	A		A	A
238	2-Ethoxyethyl acetate	111-15-9	C		C	C
239	1-Bromohexane	111-25-1	A		A	A
241	1-Hexanol	111-27-3	A		A	A
243	2-Hydroxyethyl ether	111-46-6	A		A	A
247	1-Heptanol	111-70-6	A		A	A
248	1-Bromooctane	111-83-1	A		A	A

MOAclass model, training set: *continued*

No.	Substance chemical name	Substance CASRN	Exp. class	MOA	Pred. class	MOA
250	1-Octanol	111-87-5	A		A	A
251	2-(2-Ethoxyethoxy)ethanol	111-90-0	A		A	A
253	2-Undecanone	112-12-9	A		A	A
255	Triethylene glycol	112-27-6	A		A	A
256	1-Decanol	112-30-1	A		A	A
258	Propoxur (Baygon)	114-26-1	H		H	H
259	2-Methyl-3-butyn-2-ol	115-19-5	A		A	A
260	2,2,2-Trichloroethanol	115-20-8	A		A	A
261	Dicofol (Keltane)	115-32-2	G		G	G
264	Aldicarb	116-06-3	H		H	H
268	Ethyl salicylate	118-61-6	C		C	C
269	2,4,6-Tribromophenol	118-79-6	A		A	A
270	4-Amino-2-nitrophenol	119-34-6	B		B	B
272	N-Phenyldiethanolamine	120-07-0	A		A	A
273	4-(Diethylamino)benzaldehyde	120-21-8	A		A	A
274	Catechol	120-80-9	B		B	B
275	1,2,4-Trichlorobenzene	120-82-1	A		A	A
276	2,4-Dichlorophenol	120-83-2	B		B	B
280	N,N-Dimethylaniline	121-69-7	A		A	A
283	2-Chloro-4-nitroaniline	121-87-9	B		B	B
284	<i>p</i> -Isopropyl benzaldehyde	122-03-2	A		A	A
286	2-Phenoxyethanol	122-99-6	A		A	A

MOAclass model, training set: continued

No.	Substance chemical name	Substance CASRN	Exp. class	MOA	Pred. class	MOA
288	2-Methylvaleraldehyde	123-15-9	D		D	
296	Tributyl phosphate	126-73-8	A		A	
298	1-Chloro-2-propanol	127-00-4	D		D	
299	Tetrachloroethylene	127-18-4	A		A	
300	2-Phenyl-3-butyne-2-ol	127-66-2	A		A	
304	Phenyl 4-aminosalicylate	133-11-9	C		C	
314	1,3-Dichloropropane	142-28-9	A		A	
317	Butyl ether	142-96-1	A		A	
318	1-Nonanol	143-08-8	A		A	
322	3-Methoxyphenol	150-19-6	B		B	
323	4-Methoxyphenol	150-76-5	B		B	
324	<i>p</i> -Dimethoxybenzene	150-78-7	A		A	
325	2,3-Benzofuran	271-89-6	D		A	
329	Secobarbital, sodium salt	309-43-3	E		E	
331	2,5-Dinitrophenol	329-71-5	D		D	
339	<i>alpha, alpha, alpha</i> -4-Tetrafluoro- <i>o</i> -toluidine	393-39-5	A		A	
341	<i>alpha, alpha, alpha</i> -Trifluoro- <i>o</i> -tolunitrile	447-60-9	A		A	
347	Cineole	470-82-6	A		A	
350	2,3-Dihydrobenzofuran	496-16-2	A		A	
356	2,3-Dimethyl-1,3-butadiene	513-81-5	D		A	
358	Flavone	525-82-6	A		A	
359	2,4,6-Trimethylphenol	527-60-6	A		A	

MOAclass model, training set: *continued*

No.	Substance chemical name	Substance CASRN	Exp. class	MOA class	Pred. class	MOA
363	4,6-Dinitro- <i>o</i> -cresol	534-52-1	F		F	
367	1,3-Dichloropropene	542-75-6	D		D	
371	4-Nitrobenzaldehyd	555-16-8	D		D	
373	2,6-Dinitrophenol	573-56-8	F		F	
377	Isovaleraldehyde	590-86-3	D		D	
378	2-Hexanone	591-78-6	A		A	
379	2,4-Hexadiene	592-46-1	A		A	
385	2,4-Dimethyl-3-pentanol	600-36-2	A		A	
389	Pentabromophenol	608-71-9	F		F	
393	2-Chloro-4-methylaniline	615-65-6	B		B	
394	4-Ethoxy-2-nitroaniline	616-86-4	A		A	
396	4-Nitrobenzamide	619-80-7	B		B	
398	Benzyl sulfoxide	621-08-9	A		A	
400	4-(2-Hydroxyethyl)morpholine	622-40-2	A		A	
403	2,5-Dimethylfuran	625-86-5	A		A	
404	1,5-Dichloropentane	628-76-2	A		A	
405	1-Bromoheptane	629-04-9	A		A	
406	Propyl disulfide	629-19-6	D		D	
415	Pentyl ether	693-65-2	A		A	
416	4-Methylloxazole	693-93-6	A		A	
417	2-Methylimidazole	693-98-1	D		D	
419	<i>gamma</i> -Decanolactone	706-14-9	A		A	

MOAclass model, training set: continued

No.	Substance chemical name	Substance CASRN	Exp. class	MOA	Pred. class	MOA
423	3,4-Dichloro-1-butene	760-23-6	D		D	
425	2-Butyn-1-ol	764-01-2	D		D	
426	2,5-Dimethyl-2,4-hexadiene	764-13-6	A		A	
428	3-Cyano-4,6-dimethyl-2-hydroxypyridine	769-28-8	B		B	
429	2,3,4,5,6-Pentafluoroaniline	771-60-8	A		A	
433	2-Hydroxyethyl acrylate	818-61-1	D		D	
434	1-Octyn-3-ol	818-72-4	D		D	
435	2-Nonanone	821-55-6	A		A	
436	<i>trans</i> -1,2-Dichlorocyclohexane	822-86-6	A		A	
439	3-Bromothiophene	872-31-1	D		A	
443	Ethyl 3-aminobenzoate methanesulfonic acid salt	886-86-2	C		C	
444	1,1,1,3,3-Hexafluoro-2-propanol	920-66-1	A		A	
445	1,5-Hexadien-3-ol	924-41-4	D		D	
446	3-Butyn-1-ol	927-74-2	D		D	
447	<i>cis</i> -3-Hexen-1-ol	928-96-1	A		D	
448	<i>trans</i> -3-Hexen-1-ol	928-97-2	A		A	
450	4-Phenylpyridine	939-23-1	A		A	
451	Phenyl sulfoxide	945-51-7	A		A	
453	2-Hydroxypropyl acrylate	999-61-1	D		D	
454	2-Amino-5-bromopyridine	1072-97-5	B		B	
455	Diethyl benzylphosphonate	1080-32-6	A		A	
457	4-Acetyl pyridine	1122-54-9	B		B	

MOAclass model, training set: continued

No.	Substance chemical name	Substance CASRN	Exp. class	MOA class	Pred. class	MOA
458	Methyl <i>p</i> -chlorobenzoate	1126-46-1	C		C	
459	Butyl phenyl ether	1126-79-0	A		A	
461	Tetrachlorocatechol	1198-55-6	F		F	
464	3,4-Dimethyl-1-pentyn-3-ol	1482-15-1	A		A	
465	N-Vinylcarbazole	1484-13-5	D		D	
467	Carbofuran	1563-66-2	H		H	
469	1,9-Decadiene	1647-16-1	D		D	
470	<i>p</i> -Phenylazophenol	1689-82-3	F		F	
476	3-Chloro-2-chloromethyl-1-propene	1871-57-4	D		D	
480	2,6-Dichlorobenzamide	2008-58-4	A		A	
482	Aminocarb	2032-59-9	H		H	
485	(+)-4-Pentyn-2-ol	2117-11-5	D		D	
486	4-Chlorocatechol	2138-22-9	F		F	
490	(1R,2S,5R)-(-)-menthol	2216-51-5	A		A	
492	2',4'-Dichloroacetophenone	2234-16-4	A		A	
493	<i>n</i> -Octyl cyanide	2243-27-8A	A		A	
494	<i>alpha, alpha, alpha</i> -4-Tetrafluoro- <i>m</i> -toluidine	2357-47-3	A		A	
495	<i>trans</i> -2-Phenyl-1-cyclohexanol	2362-61-0	A		A	
496	2-Ethoxyethyl methacrylate	2370-63-0	C		C	
501	Tetrahydrofurfuryl methacrylate	2455-24-5	C		C	
503	Benzyl methacrylate	2495-37-6	C		C	
510	2-Amino-4'-chlorobenzophenone	2894-51-1	A		A	

MOAclass model, training set: continued

No.	Substance chemical name	Substance CASRN	Exp. class	MOA class	Pred. class	MOA
514	Cyclohexyl acrylate	3066-71-5	D		D	
518	2,3,5,6-Tetrachloroaniline	3481-20-7	F		F	
526	Isopropyl disulfide	4253-89-8	A		A	
528	Isopropyl methacrylate	4655-34-9	C		C	
529	1-Hexen-3-ol	4798-44-1	D		D	
532	1,3-Diethyl-2-thiobarbituric acid	5217-47-0	E		E	
536	3,6-Dithiaoctane	5395-75-5	A		A	
539	2-Amino-4-chloro-6-methylpyrimidine	5600-21-5	A		A	
540	2,6-Dimethoxytoluene	5673-07-4	A		A	
541	2-Dimethylaminopyridine	5683-33-0	A		A	
545	1,1,1-Trichloro-2-methyl-2-propano(hydrate) (2:1)	6001-64-5	A		A	
546	2-Dodecanone	6175-49-1	A		A	
547	4-Dimethylaminocinnamaldehyde	6203-18-5	D		D	
554	Tripropargylamine	6921-29-5	A		A	
555	N,N-bis(2,2-Diethoxyethyl)methylamine	6948-86-3	A		A	
557	3-Hydroxy-3,7,11-trimethyl-1,6,10-dodecatriene	7212-44-4	A		A	
558	1-(2-Chloroethyl)pyrrolidine.HCl	7250-67-1	A		A	
560	1-Heptyn-3-ol	7383-19-9	D		D	
563	Resmethrin	10453-86-8	G		G	
567	2',3',4'-Trimethoxyacetophenone	13909-73-4	A		A	
571	4-Bromophenyl 3-pyridyl ketone	14548-45-9	A		A	
573	2,2,5,5-Tetramethyltetrahydrofuran	15045-43-9	A		A	

MOAclass model, training set: continued

No.	Substance chemical name	Substance CASRN	Exp. class	MOA	Pred. class	MOA
574	3-Hydroxy-2-nitropyridine	15128-82-2	A		A	
577	Methomyl	16752-77-5	H		H	
578	6-Chloro-2-pyridinol	16879-02-0	B		B	
581	4-(Diethylamino)salicylaldehyde	17754-90-4	A		A	
583	3,6-Dimethyl-1-heptyn-3-ol	19549-98-5	A		A	
584	2,4,5-Trimethyloxazole	20662-84-4	A		A	
587	<i>m</i> -Bromobenzamide	22726-00-7	A		A	
588	Oxamyl	23135-22-0	H		H	
589	2,6-Diisopropylaniline	24544-04-5	A		A	
593	2-Methyl-3,3,4-tetrafluoro-2-butanol	29553-26-2	A		A	
596	2-(Bromomethyl)tetrahydro-2H-pyran	34723-82-5	A		A	
598	4-Hexyloxyaniline	39905-57-2	A		A	
601	Fevalerate	51630-58-1	G		G	
602	Permethrin	52645-53-1	G		G	
605	<i>p</i> -(<i>tert</i> -Butyl)benzamide	56108-12-4	A		A	
607	<i>dl</i> -3-Butyn-2-ol	65337-13-5	D		D	
610	3-(4- <i>tert</i> -Butylphenoxy)benzaldehyde	69770-23-6	A		A	
611	Flucythrinate	70124-77-5	G		G	
613	3-(3,4-Dichlorophenoxy)benzaldehyde	79124-76-8	A		A	
614	2,4-Dinitro-1-naphthol sodium salt dihydrate (Martius yellow)	101836-92-4	F		F	
615	<i>t</i> -Butylstyrene	1746-23-2	D		D	
617	Chloromethyl styrene	30030-25-2	D		D	

F.3 Test set

No.	Substance chemical name	Substance CASRN	LC ₅₀ (mg/l)	LC ₅₀ (mmol/l)	Exp. Tox class	Pred. Tox class	MOA confidence	Exp. MOA class	Pred. MOA class
1	4-(Hexyloxy)- <i>m</i> -anisaldehyde	61096-84-2	2.67	0.0113	<i>high</i> AT	<i>high</i> AT	Low	D	A
2	5-Bromo-2-nitrovanillin	98434-34-5	73.3	0.266	<i>high</i> AT	<i>high</i> AT	Low	D	E
5	3'-Chloro- <i>o</i> -formotoluidide	71862-02-7	46.6	0.275	<i>high</i> AT	<i>high</i> AT	Moderate	A	A
11	4,9-Dithiadioecane	56348-39-1	2.99	0.0145	<i>high</i> AT	<i>high</i> AT	-	nd	A
14	2-chloroethyl-N-cyclohexyl carbamate	31502-57-5	35	0.17	<i>high</i> AT	<i>high</i> AT	Moderate	D	A
18	Salicylic acid Na+	54-21-7	1720	12.5	<i>low</i> AT	<i>low</i> AT	-	nd	B
19	Benzamide	55-21-0	661	5.46	<i>low</i> AT	<i>low</i> AT	Moderate	A	A
26	2,3,4,6-Tetrachlorophenol	58-90-2	1.03	0.0044	<i>high</i> AT	<i>high</i> AT	Moderate	A	F
36	2-Hydroxybenzamide	65-45-2	101	0.736	<i>low</i> AT	<i>low</i> AT	Low	A	A
39	Dicumaryl	66-76-2	5.11	0.0152	<i>high</i> AT	<i>high</i> AT	-	nd	A
40	<i>p</i> -Phenoxybenzaldehyde	67-36-7	4.6	0.0232	<i>high</i> AT	<i>high</i> AT	Moderate	A	C
43	Acetone	67-64-1	7160	123	<i>low</i> AT	<i>low</i> AT	Moderate	A	A
44	Chloroform	67-66-3	70.7	0.592	<i>low</i> AT	<i>low</i> AT	-	nd	A
45	Methyl sulfoxide	67-68-5	34000	435	<i>low</i> AT	<i>low</i> AT	Moderate	A	D
47	2,2'-Methylene <i>bis</i> (3,4,6-trichlorophenol)	70-30-4	0.021	5.16E-05	<i>high</i> AT	<i>high</i> AT	Moderate	D	G
52	Benzene	71-43-2	17.6	0.225	<i>high</i> AT	<i>high</i> AT	-	nd	A
55	Acetonitrile	75-05-8	1644	40.05	<i>low</i> AT	<i>high</i> AT	Moderate	A	A
56	Ethanal	75-07-0	33.8	0.767	<i>low</i> AT	<i>low</i> AT	Moderate	D	D
58	Iodoform	75-47-8	2.92	0.0074	<i>high</i> AT	<i>high</i> AT	-	nd	A
61	2,2,2-Trifluoroethanol	75-89-8	119	1.19	<i>low</i> AT	<i>low</i> AT	Moderate	D	A

Test set: continued

No.	Substance chemical name	Substance CASRN	LC ₅₀ (mg/l)	LC ₅₀ (mmol/l)	Exp. Tox class	Pred. Tox class	MOA confidence	Exp. MOA class	Pred. MOA class
62	3,3-Dimethyl-2-butanone	75-97-8	87	0.869	lowAT	lowAT	Moderate	A	A
64	5,5-Dimethylhydantoin	77-71-4	16460	128.5	lowAT	lowAT	-	nd	A
68	Tris(2-butoxyethyl) phosphate	78-51-3	11.2	0.0281	highAT	highAT	Moderate	A	G
73	2-Butanone	78-93-3	3220	44.66	lowAT	lowAT	Moderate	A	A
74	1-Amino-2-propanol	78-96-6	2520	33.55	lowAT	lowAT	-	nd	D
80	4,4'-Isopropylidenebis(2,6-dichlorophenol)	79-95-8	1.33	0.0036	highAT	highAT	Moderate	A	G
81	<i>p</i> -tert-Pentylphenol	80-46-6	2.59	0.0158	highAT	highAT	High	A; B	A
82	1,8-Diamino- <i>p</i> -menthane	80-52-4	65.30	0.383	highAT	highAT	-	nd	A
83	<i>alpha</i> , <i>alpha</i> '-2,6-Tetrachlorotoluene	81-19-6	0.97	0.0042	highAT	highAT	Moderate	A	A
84	Acenaphthene	83-32-9	1.73	0.0112	highAT	highAT	Moderate	A	A
88	Diphenyl pht halate	84-62-8	0.08	0.0002	highAT	highAT	-	nd	G
90	Diethyl pht halate	84-66-2	31.8	0.143	highAT	highAT	Moderate	A	A
91	Di- <i>n</i> -butylorthophthalate	84-74-2	1	0.0036	highAT	highAT	Moderate	A	A
94	Salicylanilide	87-17-2	3.95	0.0185	highAT	highAT	Moderate	A	A
99	Anthranilamide	88-68-6	395	2.901	lowAT	lowAT	Low	A	A
100	2-Nitrophenol	88-75-5	160	1.15	lowAT	highAT	Moderate	B	B
101	2- <i>sec</i> -Butyl-4,6-dinitrophenol (dinoseb)	88-85-7	0.535	0.0022	highAT	highAT	Moderate	F	A
102	Salicylaldehyde	90-02-8	2.3	0.0188	highAT	lowAT	Moderate	D	B
104	2-Phenylphenol	90-43-7	6.15	0.0361	highAT	highAT	High	A; B	D
106	3,5-Dibromosalicylaldehyde	90-59-5	0.85	0.0030	highAT	highAT	Moderate	D	A
108	Quinoline	91-22-5	77.80	0.602	lowAT	highAT	-	nd	A

Test set: continued

No.	Substance chemical name	Substance CASRN	LC ₅₀ (mg/l)	LC ₅₀ (mmol/l)	Exp. Tox class	Pred. Tox class	MOA confidence	Exp. MOA class	Pred. MOA class
109	N,N-Diethylcyclohexylamine	91-65-6	21.4	0.138	highAT	highAT	-	nd	A
113	1-Benzoylacetone	93-91-4	1.1	0.0068	highAT	highAT	Low	D	A
115	Piperine (aliphatic)	94-62-2	7.84	0.0275	highAT	highAT	Moderate	A	A
116	2,4-Dihydroxybenzaldehyde	95-01-2	13.1	0.095	highAT	lowAT	Moderate	D	B
117	<i>o</i> -Xylene	95-47-6	16.4	0.154	highAT	highAT	Moderate	A	A
119	1,2-Dichlorobenzene	95-50-1	9.47	0.064	highAT	highAT	-	nd	A
121	2-Fluorotoluene	95-52-3	19.4	0.176	highAT	highAT	Low	D	D
123	1,2,4-Trimethylbenzene	95-63-6	7.72	0.0642	highAT	highAT	Moderate	A	A
124	3,4-Dichlorotoluene	95-75-0	2.91	0.0181	highAT	highAT	-	nd	A
127	2,3-Dibromopropanol	96-13-9	71	0.326	highAT	lowAT	Moderate	D	A
128	2-Methylbutyraldehyde	96-17-3	9.97	0.116	highAT	lowAT	Moderate	D	D
130	3-Pentanone	96-22-0	1540	17.88	lowAT	lowAT	Moderate	A	A
131	2-Butanone oxime	96-29-7	843	9.676	lowAT	lowAT	-	nd	A
134	2,2'-Methylenebis(4-chlorophenol)	97-23-4	0.31	0.0011	highAT	highAT	Moderate	D	A
136	<i>p</i> - <i>tert</i> -Butylphenol	98-54-4	5.15	0.0343	highAT	highAT	Moderate	A; B	A
137	Isopropylbenzene	98-82-8	6.32	0.0526	highAT	highAT	-	nd	D
139	Nitrobenzene	98-95-3	119	0.967	lowAT	highAT	Moderate	A	D
140	<i>m</i> -Aminoacetophenone	99-03-6	382	2.826	lowAT	lowAT	Moderate	A	A
141	<i>m</i> -Nitrotoluene	99-08-1	25.6	0.187	highAT	highAT	Moderate	A	B
143	<i>p</i> -Nitroaniline	100-01-6	125	0.905	lowAT	lowAT	-	nd	B
145	<i>p</i> -Dimethylaminobenzaldehyde	100-10-7	45.70	0.306	highAT	highAT	Moderate	A	A

Test set: continued

No.	Substance chemical name	Substance CASRN	LC ₅₀ (mg/l)	LC ₅₀ (mmol/l)	Exp. Tox class	Pred. Tox class	MOA confidence	Exp. MOA class	Pred. MOA class
146	1,4-Dinitrobenzene	100-25-4	0.71	0.0042	highAT	highAT	-	nd	D
148	Ethylbenzene	100-41-4	10.5	0.0989	highAT	highAT	Moderate	A	A
149	Benzylamine	100-46-9	102	0.952	lowAT	lowAT	-	nd	A
151	N-Methylamine	100-61-8	100	0.933	lowAT	lowAT	Moderate	A; B	A
152	Cyclohexanone oxime	100-64-1	208	1.838	lowAT	lowAT	-	nd	A
155	Solketal	100-79-8	16700	126.4	lowAT	lowAT	Moderate	A	A
156	Hexamethylenetetramine (aliphatic)	100-97-0	49800	355.2	lowAT	lowAT	-	nd	C
157	Phenyl ether	101-84-8	4	0.0235	highAT	highAT	Moderate	A	A
158	N-Ethyl- <i>m</i> -toluidine	102-27-2	49.5	0.366	highAT	highAT	Moderate	A	A
159	Tripropylamine	102-69-2	50.90	0.355	highAT	highAT	-	nd	A
160	Triethanolamine	102-71-6	11800	79.09	lowAT	lowAT	-	nd	A
163	1-(2-Hydroxyethyl)piperazine	103-76-4	6410	49.24	lowAT	lowAT	-	nd	A
164	N,N-Dimethylbenzylamine	103-83-3	37.8	0.28	highAT	highAT	-	nd	A
165	4-Acetamidophenol	103-90-2	814	5.385	lowAT	lowAT	-	nd	A
166	4-Butylaniline	104-13-2	10.2	0.068	highAT	highAT	Moderate	A; B	A
167	Nonylphenol (mixed)	25154-52-3	0.14	0.0006	highAT	highAT	Moderate	A; B	A
170	4-Chlorobenzaldehyde	104-88-1	2.2	0.0156	highAT	highAT	Moderate	D	D
172	5-Diethylamino-2-pentanone	105-14-6	336	2.137	lowAT	lowAT	-	nd	A
173	Diethyl malonate	105-53-3	14.7	0.0918	highAT	highAT	-	nd	C
175	Dibutyl fumarate	105-75-9	0.63	0.0028	highAT	highAT	Moderate	C	A
176	Dibutyl adipate	105-99-7	3.64	0.0141	highAT	highAT	Moderate	C	A

Test set: continued

No.	Substance chemical name	Substance CASRN	LC ₅₀ (mg/l)	LC ₅₀ (mmol/l)	Exp. Tox class	Pred. Tox class	MOA confidence	Exp. MOA class	Pred. MOA class
178	<i>p</i> -Xylene	106-42-3	8.87	0.0835	<i>high</i> AT	<i>high</i> AT	Moderate	A	A
179	4-Methylphenol (<i>p</i> -cresol)	106-44-5	16.5	0.153	<i>high</i> AT	<i>high</i> AT	-	nd	B
181	4-Chlorophenol	106-48-9	6.11	0.0475	<i>high</i> AT	<i>high</i> AT	Moderate	B	B
188	Propylamine	107-10-8	308	5.211	<i>low</i> AT	<i>low</i> AT	-	nd	D
190	Chloroacetonitrile	107-14-2	1.35	0.0178	<i>high</i> AT	<i>low</i> AT	Moderate	I	D
191	Ethylenediamine	107-15-3	220	3.661	<i>low</i> AT	<i>low</i> AT	-	nd	B
194	Acetaldoxime	107-29-9	76	1.287	<i>low</i> AT	<i>low</i> AT	-	nd	A
196	<i>tert</i> -Octylamine	107-45-9	24.6	0.19	<i>high</i> AT	<i>high</i> AT	-	nd	A
197	<i>tert</i> -Butyl sulfide	107-47-1	29.1	0.199	<i>high</i> AT	<i>high</i> AT	Moderate	A	D
198	2-Pentanone	107-87-9	1240	14.4	<i>low</i> AT	<i>low</i> AT	-	nd	A
201	Toluene	108-88-3	33.90	0.368	<i>high</i> AT	<i>high</i> AT	-	nd	A
203	Chlorobenzene	108-90-7	16.9	0.15	<i>high</i> AT	<i>high</i> AT	-	nd	D
205	Cyclohexanone	108-94-1	621	6.327	<i>low</i> AT	<i>low</i> AT	Moderate	A	A
207	3-Picoline	108-99-6	144	1.546	<i>low</i> AT	<i>low</i> AT	Moderate	B	A
208	1-Methylpiperazine	109-01-3	2300	22.96	<i>low</i> AT	<i>low</i> AT	-	nd	A
210	2-Methylpiperazine	109-07-9	2240	22.36	<i>low</i> AT	<i>low</i> AT	-	nd	D
211	Propyl acetate	109-60-4	60	0.587	<i>low</i> AT	<i>low</i> AT	Moderate	C	A
214	Butylamine	109-73-9	268	3.664	<i>low</i> AT	<i>low</i> AT	-	nd	D
215	Allyl cyanide	109-75-1	182	2.713	<i>low</i> AT	<i>low</i> AT	-	nd	D
217	Malonitrile	109-77-3	0.56	0.0085	<i>high</i> AT	<i>high</i> AT	Moderate	I	A
218	2-Methoxyethylamine	109-85-3	524	6.976	<i>low</i> AT	<i>low</i> AT	-	nd	D

Test set: continued

No.	Substance chemical name	Substance CASRN	LC ₅₀ (mg/l)	LC ₅₀ (mmol/l)	Exp. Tox class	Pred. Tox class	MOA confidence	Exp. MOA class	Pred. MOA class
219	Diethylamine	109-89-7	855	11.69	<i>low</i> AT	<i>low</i> AT	-	nd	A
220	Pyrrrole	109-97-7	210	3.13	<i>low</i> AT	<i>low</i> AT	Moderate	A	B
221	Tetrahydrofuran	109-99-9	2160	29.96	<i>low</i> AT	<i>low</i> AT	Moderate	A	A
222	Furan	110-00-9	61	0.896	<i>low</i> AT	<i>low</i> AT	Moderate	D	B
223	<i>t</i> -Butyl disulfide	110-06-5	1.37	0.0077	<i>high</i> AT	<i>high</i> AT	Moderate	A	D
224	5-Methyl-2-hexanone	110-12-3	159	1.392	<i>low</i> AT	<i>high</i> AT	Moderate	A	D
225	Diethyl sebacate	110-40-7	2.72	0.0105	<i>high</i> AT	<i>high</i> AT	-	nd	A
227	Hexane	110-54-3	2.5	0.029	<i>high</i> AT	<i>high</i> AT	Moderate	A	A
229	Amylamine	110-58-7	177	2.031	<i>low</i> AT	<i>low</i> AT	-	nd	A
232	2-(Ethylamino)ethanol	110-73-6	1480	16.6	<i>low</i> AT	<i>low</i> AT	Moderate	A	D
233	Cyclohexane	110-82-7	4.53	0.0538	<i>high</i> AT	<i>high</i> AT	Moderate	A	A
235	<i>s</i> -Trioxane	110-88-3	5950	66.05	<i>low</i> AT	<i>low</i> AT	Moderate	A	A
236	6-Methyl-5-hepten-2-one	110-93-0	85.7	0.679	<i>low</i> AT	<i>low</i> AT	Moderate	A	A
240	Hexylamine	111-26-2	56.6	0.559	<i>low</i> AT	<i>high</i> AT	-	nd	D
242	Dietanolamine	111-42-2	4710	44.8	<i>low</i> AT	<i>low</i> AT	Moderate	A	D
244	<i>n</i> -Propyl sulfide	111-47-7	21.7	0.184	<i>high</i> AT	<i>high</i> AT	Moderate	A	A
245	<i>n</i> -Heptylamine	111-68-2	21.8	0.189	<i>high</i> AT	<i>high</i> AT	-	nd	A
246	1,4-Dicyanobutane	111-69-3	1930	17.85	<i>low</i> AT	<i>high</i> AT	Moderate	A	A
249	Octylamine	111-86-4	5.19	0.0402	<i>high</i> AT	<i>high</i> AT	-	nd	A
252	Nonanoic acid	112-05-0	104	0.657	<i>low</i> AT	<i>high</i> AT	-	nd	D
254	Nonylamine	112-20-9	2.16	0.015	<i>high</i> AT	<i>high</i> AT	-	nd	A

Test set: continued

No.	Substance chemical name	Substance CASRN	LC ₅₀ (mg/l)	LC ₅₀ (mmol/l)	Exp. Tox class	Pred. Tox class	MOA confidence	Exp. MOA class	Pred. MOA class
262	Triphenyl phosphate	115-86-6	0.87	0.0027	highAT	highAT	-	nd	G
267	Phenyl salicylate	118-55-8	1.18	0.0055	highAT	highAT	Moderate	C	A
271	Benzophenone	119-61-9	14.7	0.0807	highAT	highAT	Moderate	A	A
277	2,4-Dinitrotoluene	121-14-2	24.3	0.133	highAT	highAT	Low	D	F
278	3-Ethoxy-4-hydroxybenzaldehyde	121-32-4	87.6	0.527	lowAT	lowAT	Moderate	A	H
279	Vanillin	121-33-5	83.8	0.551	lowAT	lowAT	Moderate	D	H
281	1-Chloro-3-nitrobenzene	121-73-3	18.8	0.119	highAT	highAT	Moderate	A	B
284	<i>p</i> -Isopropyl benzaldehyde	122-03-2	6.62	0.0447	highAT	highAT	Moderate	A	A
287	4-Ethylphenol	123-07-9	10.4	0.0851	highAT	highAT	-	nd	A
289	2,4-Pentanedione	123-54-6	135	1.348	lowAT	lowAT	Moderate	D	A
290	Ethyl hexanoate	123-66-0	8.9	0.0617	highAT	highAT	Moderate	C	A
291	Butanal	123-72-8	14.7	0.204	highAT	lowAT	Moderate	D	D
292	Butyl acetate	123-86-4	18	0.155	highAT	lowAT	Moderate	C	A
293	1,4-Dioxane	123-91-1	10300	116.9	lowAT	lowAT	Moderate	A	A
294	Dodecylamine	124-22-1	0.103	0.0006	highAT	highAT	-	nd	A
297	5,5-Dimethyl-1,3-cyclohexanedione	126-81-8	11500	82.04	lowAT	lowAT	Low	A	A
301	2,6-Di- <i>tert</i> -butyl-4-methylphenol	128-37-0	0.363	0.0016	highAT	highAT	Moderate	A	G
302	Saccharin sodium salt hydrate	82385-42-0	18300	82	lowAT	lowAT	Moderate	A	A
303	Dibenzofuran	132-64-9	1.5	0.0089	highAT	highAT	-	nd	F
305	N,N-Diethyl- <i>m</i> -toluamide	134-62-3	110	0.575	lowAT	highAT	Moderate	A	A
306	Propionic acid,sodium salt	137-40-6	4790	49.86	lowAT	lowAT	-	nd	A

Test set: continued

No.	Substance chemical name	Substance CASRN	LC ₅₀ (mg/l)	LC ₅₀ (mmol/l)	Exp. Tox class	Pred. Tox class	MOA confidence	Exp. MOA class	Pred. MOA class
307	1-(2-Aminoethyl)piperazine	140-31-8	2190	16.95	lowAT	lowAT	-	nd	B
308	Dibutyl succinate	141-03-7	4.46	0.0194	highAT	highAT	Moderate	C	A
309	Diethyl adipate	141-28-6	18	0.0899	highAT	highAT	-	nd	A
310	2-Aminoethanol	141-43-5	2070	33.89	lowAT	lowAT	-	nd	D
311	Ethyl acetate	141-78-6	230	2.61	lowAT	lowAT	Moderate	C	A
312	2,6-Dimethyl morpholine	141-91-3	387	3.36	lowAT	lowAT	Moderate	A	A
313	<i>m</i> -Diethylbenzene	141-93-5	4.15	0.0309	highAT	highAT	Moderate	A	D
315	Hexanoic acid	142-62-1	320	2.755	lowAT	highAT	-	nd	D
316	Hexyl acetate	142-92-7	4.4	0.0305	highAT	highAT	Moderate	C	D
319	Di- <i>n</i> -hexylamine	143-16-8	0.78	0.0042	highAT	highAT	-	nd	A
321	<i>o</i> -Vanillin	148-53-8	2.40	0.0158	highAT	highAT	Moderate	D	A
326	1,4-Diazabicyclo[2.2.2]octane	280-57-9	1730	15.42	lowAT	lowAT	-	nd	A
327	Adamantane	281-23-2	0.28	0.0021	highAT	highAT	-	nd	A
330	Bromacil	314-40-9	186	0.712	lowAT	highAT	Low	A	H
332	Diuron	330-54-1	14.2	0.0609	highAT	highAT	Moderate	A	A
333	<i>p</i> -Fluorophenyl ether	330-93-8	1.13	0.0055	highAT	highAT	Moderate	D	A
335	1-Fluoro-4-nitrobenzene	350-46-9	28.4	0.201	highAT	highAT	Moderate	A	D
336	<i>alpha, alpha, alpha</i> -Trifluoro- <i>m</i> -tolunitrile	368-77-4	47.70	0.279	highAT	highAT	Moderate	A	D
337	4-Fluoroaniline	371-40-4	16.9	0.152	highAT	highAT	Low	D	B
338	2-Chloro-6-fluorobenzaldehyde	387-45-1	9.41	0.0593	highAT	highAT	Moderate	D	A
340	2-Fluorobenzaldehyde	446-52-6	1.35	0.0109	highAT	highAT	Moderate	D	D

Test set: continued

No.	Substance chemical name	Substance CASRN	LC ₅₀ (mg/l)	LC ₅₀ (mmol/l)	Exp. Tox class	Pred. Tox class	MOA confidence	Exp. MOA class	Pred. MOA class
342	<i>alpha, alpha, alpha</i> -Trifluoro- <i>m</i> -tolualdehyde	454-89-7	0.924	0.0053	<i>high</i> AT	<i>high</i> AT	Moderate	D	D
343	4-Fluoro- <i>N</i> -methylaniline	459-59-6	38.4	0.307	<i>high</i> AT	<i>high</i> AT	Moderate	A; B	A
345	[(1 <i>S</i>)-endo]-(-)-Borneol	464-45-9	63.2	0.41	<i>high</i> AT	<i>high</i> AT	-	nd	A
346	(1 <i>S</i>)-(-)-Camphor	464-48-2	17	0.112	<i>high</i> AT	<i>high</i> AT	-	nd	A
348	Neobietic acid	471-77-2	1.49	0.0049	<i>high</i> AT	<i>high</i> AT	-	nd	A
351	exo-Norborneol	497-37-0	228	2.033	<i>low</i> AT	<i>low</i> AT	-	nd	A
352	Norbornylene	498-66-8	10	0.106	<i>high</i> AT	<i>high</i> AT	-	nd	A
353	2,6-Pyridinedicarboxylic acid	499-83-2	322	1.927	<i>low</i> AT	<i>low</i> AT	-	nd	D
354	3-Pyridinecarboxaldehyde	500-22-1	16.4	0.153	<i>high</i> AT	<i>low</i> AT	Moderate	D	D
355	5-Nonanone	502-56-7	31	0.218	<i>high</i> AT	<i>high</i> AT	Moderate	A	A
357	Abietic acid	514-10-3	2.38	0.0079	<i>high</i> AT	<i>high</i> AT	-	nd	A
360	<i>o</i> -Tolunitrile	529-19-1	44.7	0.382	<i>high</i> AT	<i>high</i> AT	Moderate	A	A
361	<i>o</i> -Tolualdehyde	529-20-4	52.9	0.44	<i>high</i> AT	<i>high</i> AT	Moderate	A	D
362	Benzoic acid,sodium salt	532-32-1	484	3.359	<i>low</i> AT	<i>high</i> AT	-	nd	B
364	Amylbenzene	538-68-1	1.71	0.0115	<i>high</i> AT	<i>high</i> AT	Moderate	A	D
365	<i>tert</i> -Butyl acetate	540-88-5	327	2.815	<i>low</i> AT	<i>low</i> AT	Moderate	A	A
366	1,3-Dichlorobenzene	541-73-1	8.03	0.0546	<i>high</i> AT	<i>high</i> AT	-	nd	A
368	<i>n</i> -Butyl sulfide	544-40-1	3.58	0.0245	<i>high</i> AT	<i>high</i> AT	-	nd	A
369	2'-Hydroxy-4'-methoxyacetophenone	552-41-0	69.5	0.418	<i>high</i> AT	<i>high</i> AT	Moderate	A	A
370	<i>o</i> -Nitrobenzaldehyde	552-89-6	14.4	0.0953	<i>high</i> AT	<i>high</i> AT	Moderate	D	B
372	3-Methyl-2-butanone	563-80-4	864	10.03	<i>low</i> AT	<i>low</i> AT	Moderate	A	A

Test set: continued

No.	Substance chemical name	Substance CASRN	LC ₅₀ (mg/l)	LC ₅₀ (mmol/l)	Exp. Tox class	Pred. Tox class	MOA confidence	Exp. MOA class	Pred. MOA class
374	1,2-Dibromobenzene	583-53-9	4.05	0.0172	highAT	highAT	Moderate	A	A
375	N-Allylaniline	589-09-3	35.9	0.27	highAT	highAT	Moderate	A	A
376	4-Ethylaniline	589-16-2	73	0.602	lowAT	highAT	High	A; B	B
380	2-Tridecanone	593-08-8	0.36	0.0018	highAT	highAT	Moderate	A	A
382	Manool	596-85-0	0.12	0.0004	highAT	highAT	-	nd	A
384	1,2-Dimethylpropylamine	598-74-3	284	3.258	lowAT	lowAT	-	nd	A
387	N,N-Diphenylformamide	607-00-1	30.4	0.154	highAT	highAT	Moderate	A	A
388	Diethyl benzylmalonate	607-81-8	5.43	0.0217	highAT	highAT	-	nd	A
390	2,4,6-Triodophenol	609-23-4	1.21	0.0026	highAT	highAT	Moderate	A	F
391	2,4-Dimethoxy benzaldehyde	613-45-6	20.1	0.121	highAT	highAT	Moderate	D	A
392	2-Acetamido phenol	614-80-2	27	0.179	highAT	lowAT	-	nd	A
395	Methyl <i>p</i> -nitrobenzoate	619-50-1	23.8	0.131	highAT	highAT	Moderate	C	A
397	4-Nitrophenyl phenyl ether	620-88-2	2.65	0.0123	highAT	highAT	Moderate	A	A
399	3-Acetamido phenol	621-42-1	1130	7.475	lowAT	lowAT	-	nd	A
401	<i>alpha, alpha</i> '-Dichloro- <i>p</i> -xylene	623-25-6	0.039	0.0002	highAT	highAT	Low	D	A
407	1,6-Dicyanohexane	629-40-3	528	3.877	lowAT	lowAT	Moderate	A	A
408	2,3,4-Trichloroaniline	634-67-3	3.64	0.0185	highAT	highAT	High	A; B	B
409	5-Chlorosalicylaldehyde	635-93-8	0.77	0.0049	highAT	highAT	-	nd	B
410	4-Propylphenol	645-56-7	11	0.0808	highAT	highAT	-	nd	A
411	Pentafluoro benzaldehyde	653-37-2	1.1	0.0056	highAT	highAT	Moderate	D	A
412	2,2-Dichloroacetamide	683-72-7	241	1.883	lowAT	lowAT	Moderate	D	D
413	1-Methyl heptylamine	693-16-3	5.19	0.0402	highAT	highAT	-	nd	A

Test set: continued

No.	Substance chemical name	Substance CASRN	LC ₅₀ (mg/l)	LC ₅₀ (mmol/l)	Exp. Tox class	Pred. Tox class	MOA confidence	Exp. MOA class	Pred. MOA class
414	2-Decanone	693-54-9	4.83	0.0309	highAT	highAT	Moderate	A	A
418	2-Adamantanone	700-58-3	60.8	0.405	highAT	lowAT	-	nd	A
420	4,6-Dimethoxy-2-hydroxybenzaldehyde	708-76-9	2.68	0.0147	highAT	highAT	Moderate	D	A
421	Propanil	709-98-8	8.6	0.0394	highAT	highAT	Moderate	A	A
422	2,4,6-Tri- <i>tert</i> -butylphenol	732-26-3	0.0609	0.0002	highAT	highAT	-	nd	A
424	N,N-Dibutylformamide	761-65-9	89.3	0.568	lowAT	highAT	Moderate	A	A
427	1-Adamantanamine	768-94-5	25	0.165	highAT	highAT	-	nd	A
431	Triphenylphosphine oxide	791-28-6	53.7	0.193	highAT	highAT	Moderate	A	G
437	<i>p</i> -Phenoxyphenol	831-82-3	4.95	0.0266	highAT	highAT	High	A; B	A
438	2-Hydroxyethyl methacrylate	868-77-9	227	1.744	lowAT	lowAT	Moderate	C	D
440	2,4-Dichlorobenzaldehyde	874-42-0	1.8	0.0103	highAT	highAT	Moderate	D	A
449	2-Acetyl-1-methylpyrrole	932-16-1	157	1.275	lowAT	lowAT	Moderate	D	B
460	Methyl 4-cyano benzoate	1129-35-7	46.8	0.29	highAT	highAT	Moderate	A	C
462	<i>ortho</i> -Bromo-2',5'-dimethoxyacetophenone	1204-21-3	0.066	0.0025	highAT	highAT	Moderate	D	A
466	3-Benzoyloxyaniline	1484-26-0	9.14	0.0459	highAT	highAT	Moderate	B	A
468	<i>tert</i> -Butyl methyl ether	1634-04-4	672	7.623	lowAT	lowAT	Moderate	A	A
471	3,5-Diiodo-4-hydroxybenzonitrile	1689-83-4	6.8	0.0183	highAT	highAT	Moderate	A	F
472	3,5-Dibromo-4-hydroxybenzonitrile	1689-84-5	12.6	0.0455	highAT	highAT	-	nd	A
473	Dehydroabietic acid	1740-19-8	2.1	0.0070	highAT	highAT	-	nd	A
474	2-Allylphenol	1745-81-9	15	0.112	highAT	highAT	-	nd	A
475	5-Bromosalicylaldehyde	1761-61-1	1.3	0.0065	highAT	highAT	-	nd	B

Test set: continued

No.	Substance chemical name	Substance CASRN	LC ₅₀ (mg/l)	LC ₅₀ (mmol/l)	Exp. Tox class	Pred. Tox class	MOA confidence	Exp. MOA class	Pred. MOA class
477	3,5-Dichloro-4-hydroxybenzoxonitrile	1891-95-8	24.3	0.129	highAT	highAT	Moderate	A	A
478	Di- <i>n</i> -butylterephthalate	1962-75-0	0.59	0.0021	highAT	highAT	Moderate	A	A
479	4,4'-Dihydroxydiphenyl ether	1965-09-9	5.78	0.0286	highAT	highAT	High-Moderate	A; B	A
481	<i>n</i> -Decylamine	2016-57-1	1.03	0.0065	highAT	highAT	-	nd	A
483	2,4,5-Tribromoimidazole	2034-22-2	6.12	0.0201	highAT	highAT	Moderate	D	F
486	4-Chlorocatechol	2138-22-9	1.58	0.0109	highAT	lowAT	Low	D	B
487	Methyl 2,4-dihydroxybenzoate	2150-47-2	45.8	0.272	highAT	lowAT	Moderate	C	C
491	1-(<i>p</i> -Toluenesulfonyl)imidazole	2232-08-8	41.8	0.188	highAT	highAT	-	nd	G
497	2,3,6-Trimethylphenol	2416-94-6	8.2	0.0602	highAT	highAT	Moderate	A	B
498	<i>n</i> -Undecyl cyanide	2437-25-4	0.43	0.0024	highAT	highAT	Moderate	A	A
499	<i>o</i> -Methoxybenzamide	2439-77-2	120	0.794	lowAT	lowAT	Moderate	D	A
500	2,4-Dichlorobenzamide	2447-79-2	95.6	0.503	lowAT	highAT	Low	D	A
502	4,5-Dichloroguaiacol	2460-49-3	4.47	0.0232	highAT	highAT	Moderate	A	B
504	Hexyl acrylate	2499-95-8	1.11	0.0071	highAT	highAT	-	nd	D
505	<i>p</i> -(<i>tert</i> -butyl)-Phenyl- <i>N</i> -methylcarbamate	2626-83-7	10	0.0482	highAT	highAT	Moderate	D	A
507	1-Benzylpiperazine	2759-28-6	47.4	0.269	highAT	lowAT	-	nd	A
508	3-(3-Pyridyl)-1-propanol	2859-67-8	150	1.093	lowAT	lowAT	Moderate	B	A
509	Tridecylamine	2869-34-3	0.0654	0.0003	highAT	highAT	-	nd	A
511	Methyl 2,5-dichlorobenzoate	2905-69-3	14	0.0683	highAT	highAT	Moderate	C	A
513	5-Bromovanillin	2973-76-4	59.7	0.258	highAT	highAT	Moderate	D	B

Test set: continued

No.	Substance chemical name	Substance CASRN	LC ₅₀ (mg/l)	LC ₅₀ (mmol/l)	Exp. Tox class	Pred. Tox class	MOA confidence	Exp. MOA class	Pred. MOA class
516	Triethyl nitritotricarboxylate	3206-31-3	13.3	0.057	highAT	highAT	-	nd	C
517	4,5-Dichlorocatechol	3428-24-8	0.89	0.0050	highAT	highAT	Low	D	B
519	2,6-Diphenylpyridine	3558-69-8	0.21	0.0009	highAT	highAT	Moderate	A	D
520	1,3-Dichloro-4,6-dinitrobenzene	3698-83-7	0.0456	0.0002	highAT	highAT	Moderate	D	D
523	2,3-Dimethylvaleraldehyde	3944-76-1	16	0.14	highAT	highAT	Moderate	D	A
524	2-Decyn-1-ol	4117-14-0	1.07	0.0070	highAT	highAT	-	nd	A
525	5-Chloro-2-pyridinol	4214-79-3	1140	8.8	lowAT	lowAT	-	nd	B
527	2,4,5-Trimethoxybenzaldehyde	4460-86-0	49.5	0.252	highAT	highAT	Moderate	D	A
530	2,3,4,5-Tetrachlorophenol	4901-51-3	0.41	0.0018	highAT	highAT	Moderate	A	F
531	1,2-bis(4-Pyridyl)ethane	4916-57-8	151	0.82	lowAT	highAT	Moderate	A	C
533	Dimethyl nitroterephthalate	5292-45-5	6.52	0.0273	highAT	highAT	Low	C	A
534	5-Chloro-2-mercaptobenzothiazole	5331-91-9	3.21	0.0159	highAT	highAT	Moderate	A	A
535	Dimethyl aminoterephthalate	5372-81-6	8.94	0.0427	highAT	highAT	Moderate	A	A
537	Dimethylamino propyl chloride,hydrochloride	5407-04-5	133	0.841	lowAT	lowAT	-	nd	A
538	4'-Chloro-3'-nitroacetophenone	5465-65-6	5.5	0.0276	highAT	highAT	Low	D	B
542	2,2-Dimethyl-1-propylamine	5813-64-9	475	5.449	lowAT	lowAT	-	nd	A
543	Isopimaric acid	5835-26-7	0.87	0.0029	highAT	highAT	-	nd	A
544	2-Amino-5-chlorobenzonitrile	5922-60-1	28.6	0.187	highAT	highAT	Moderate	A	B
549	1,3,5-Trichloro-2,4-dinitrobenzene	6284-83-9	0.222	0.0008	highAT	highAT	Moderate	D	F
550	2-Chloro-5-nitro benzaldehyde	6361-21-3	3.87	0.0209	highAT	highAT	Moderate	D	D
551	2-Chloro-6-methylbenzonitrile	6575-09-3	15.1	0.0996	highAT	highAT	-	nd	A

Test set: continued

No.	Substance chemical name	Substance CASRN	LC ₅₀ (mg/l)	LC ₅₀ (mmol/l)	Exp. Tox class	Pred. Tox class	MOA confidence	Exp. MOA class	Pred. MOA class
552	2-Bromo-3-pyridinol	6602-32-0	469	2.695	lowAT	lowAT	Moderate	B	B
553	2-Chloro-3-pyridinol	6636-78-8	622	4.801	lowAT	lowAT	Moderate	B	B
556	1,4-bis(3-Aminopropyl)piperazine	7209-38-3	3100	15.47	lowAT	lowAT	-	nd	A
559	n-Undecylamine	7307-55-3	0.21	0.0012	highAT	highAT	-	nd	A
561	p-Ethoxybenzaldehyde	10031-82-0	28.1	0.187	highAT	highAT	Moderate	D	A
562	[1(R)-endo](+)-3-Bromocamphor	10293-06-8	68.5	0.296	highAT	highAT	-	nd	A
565	<i>alpha, alpha, alpha</i> ,-Tetrabromo- <i>o</i> -xylene	13209-15-9	0.437	0.0010	highAT	highAT	Low	D	A
566	2',3',4'-Trichloroacetophenone	13608-87-2	2	0.0089	highAT	highAT	Moderate	A	A
568	Diethyl chloromalonate	14064-10-9	0.95	0.0049	highAT	highAT	-	nd	C
569	N-Ethylbenzylamine	14321-27-8	57.1	0.422	highAT	highAT	-	nd	A
572	4-Benzoylpyridine	14548-46-0	103	0.562	lowAT	highAT	Moderate	A	A
575	Alachlor	15972-60-8	5	0.0185	highAT	highAT	Moderate	A	G
576	4-Octylaniline	16245-79-7	0.12	0.0006	highAT	highAT	Low	A	A
579	3-Amino-5,6-dimethyl-1,2,4-triazine	17584-12-2	952	7.668	lowAT	lowAT	Moderate	B	H
582	6-Chloro-2-picoline	18368-63-3	232	1.819	lowAT	highAT	-	nd	A
585	4-Dimethylamino-3-methyl-2-butanone	22104-62-7	8.5	0.0658	lowAT	highAT	-	nd	A
594	(+)- <i>sec</i> -Butylamine	13952-84-6	275	3.76	lowAT	lowAT	-	nd	D
595	N-(3-Methoxypropyl)-3,4,5-trimethoxybenzylamine	34274-04-9	136	0.505	lowAT	lowAT	-	nd	H
597	4-Decylaniline	37529-30-9	0.062	0.0003	highAT	highAT	Low	A	A

Test set: continued

No.	Substance chemical name	Substance CASRN	LC ₅₀ (mg/l)	LC ₅₀ (mmol/l)	Exp. Tox class	Pred. Tox class	MOA confidence	Exp. MOA class	Pred. MOA class
599	Methyl 4-chloro-2-nitrobenzoate	42087-80-9	27.7	0.128	<i>high</i> AT	<i>high</i> AT	Moderate	C	A
600	5-Hydroxy-2-nitrobenzaldehyde	42454-06-8	41.9	0.251	<i>high</i> AT	<i>high</i> AT	-	nd	B
603	3,8-Dithiadecane	54576-32-8	6.06	0.034	<i>high</i> AT	<i>high</i> AT	Moderate	A	A
604	2'-(Octyloxy)-acetanilide	55792-61-5	0.45	0.0017	<i>high</i> AT	<i>high</i> AT	Moderate	A	A
606	2,9-Dithiadecane	56348-40-4	10.1	0.0566	<i>high</i> AT	<i>high</i> AT	-	nd	A

F.4 References

- [1] Russom, C. L.; Bradbury, S. P.; Broderius, S. J.; Hammermeister, D. E.; Drummond, R. A. Predicting modes of toxic action from chemical structure: acute toxicity in the fathead minnow (*Pimephales promelas*). *Environ. Toxicol Chem.* **1997**, *16*, 948-967. Dataset available on line at http://www.epa.gov/nct/dsstox/sdf_epafhm.html.