HelmholtzZentrum münchen
Deutsches Forschungszentrum für Gesundheit und Umwelt

Marie Curie Initial Training Network

Environmental Chemoinformatics (ECO)

Final report

14 July 2012

Modelling of acute aquatic toxicity on *Daphnia magna* and Fish

Early stage researcher:
Matteo Cassotti

Project supervisor:
Dr. Igor Tetko

Research Institution:
Helmholtz Zentrum München

Table of contents

Chapter 1

Introduction

1.1 Acute aquatic toxicity

Chemical substances can easily enter the environment via different routes, such as wastewaters and gas emissions in the air. The environmental fate of a substance is typically a function of its physical-chemical properties, such as $LogK_{ow}$, $LogK_{oc}$ and Henry's law constant. Depending on the environmental compartment(s) the substance is transported to, the ecotoxicological effects caused by the substance itself can be different.

Many chemicals eventually partition in water and can exert adverse effects on aquatic systems. Chemical substances that are toxic to aquatic organisms can cause serious damages not only to aquatic species themselves, but can also disrupt aquatic food webs and threaten the survival of other parts of these systems, such as birds and mammals[1]. In fact, since aquatic species comprise the components of food chains that lead eventually to man, the survival of terrestrial species is partially dependent upon aquatic organisms.

Lethality in aquatic organisms can be induced by means of both non-specific and specific mechanisms of reaction. Non-specific lethality, known as narcosis, is exerted by the majority of chemicals that are toxic to aquatic organisms. This type of lethality does not involve reactions with cellular macromolecules and occurs when the concentration of the chemical within a cell or in cellular membranes is high enough to cause non-specific perturbations in cellular function. Therefore, the ability of chemicals to diffuse across cellular membranes is the driving force to narcosis. Since cellular membranes are constituted by a lipid bilayer (Figure 1), they are more easily crossed by non-polar molecules. The relative toxic potency of non-polar substances that induce lethality by a narcosis mechanism is, thus, a function of their lipophilicity. Narcosis toxicity represents the baseline or minimum toxicity.

Some chemicals present an excess toxicity (compared to the baseline set by narcosis) which is due to the occurrence of specific reactions. These reactions usually take place between the toxicant (or its metabolites) and critical cellular macromolecules. An example is represented by reactions that lead to the formation of covalent bonds between the toxicant and enzymes.

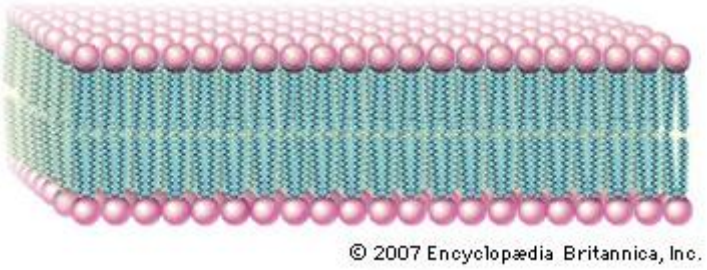


Figure 1: cellular lipid bilayer.

The assessment of aquatic toxicity of chemical substances is a primary aspect to be addressed to preserve not only the environment but also human health. Toxicity tests are typically divided in acute and chronic tests, according to the duration of the exposure to the toxicant the test organism is subject to.

Acute toxicity testing is the estimation of the hazard potential of a substance by determining its systemic toxicity in a test system, following a short-term exposure[7]. The test organism is subject to a single exposure or multiple events over a short period of time (hours or days). High doses of the toxicant, able to produce immediate effects are used[8]. These tests measure endpoints such as survival (or mortality), growth, reproduction, that are measured at each concentration in a gradient, along with a control test[6]. The assessment has traditionally been based on the median effective concentration (EC50), or lethal concentration (LC50) that has effect (or kills) 50% of test animals. Acute tests are not valid if mortality in the control sample is greater than 10%.

Chronic tests are long-term tests (weeks, months, years) relative to the test organism's life span. The test animals are subject to low, continuous doses of a toxicant. Chronic exposures may induce acute-like effects, but can also result in effects that develop slowly. These tests allow to evaluate the highest concentration that produced no observable effects (No Observed Effect Concentration, NOEC) and the lowest concentration that caused observable effects (Lowest Observed Effect Concentration, LOEC). Chronic tests are not considered valid if mortality in the control sample is greater than 20%.

There are different types of toxicity tests that can be performed on various test species. Different species differ in their susceptibility to chemicals, most likely due to differences in accessibility, metabolic rate, excretion rate, genetic factors, dietary factors, age, sex, health and stress level. Common standard test species are the fathead minnow (*Pimephales promelas*) and daphnids (*Daphnia magna*, *Daphnia pulex, Daphnia pulicaria, Ceriodaphnia dubia*)[8].

1.2 REACH requirements and QSAR

REACH (Registration, Evaluation, Authorization and Restriction of CHemicals) is a regulation of the European Union, adopted to improve the protection of human health and the environment from the risks that can be posed by chemicals, while enhancing the competitiveness of the EU chemical industry. REACH requires a huge amount of toxicological and ecotoxicological data for all chemicals manufactured and/or traded inside the European Community above 1 tonne/year.

In order to have a complete ecotoxicological profile of the substances subject to registration, REACH requires information on acute aquatic toxicity. Annex VII of the regulation, dealing with substances imported or manufactured above 1 tonne/year, states that the registrant must provide information on short-term, *i.e.* acute, aquatic toxicity to invertebrates, *Daphnia magna* being the preferred species.

In Annex VIII of REACH, which regards substances in the bandwidth 10-100 tonnes/year, the requirements for short-term aquatic toxicity on fish are reported.

In order to avoid unnecessary animal testing, according to the credo of the Three "R" (Replace, Reduce and Refine[2]) REACH promotes the adoption of alternative test methods, including *in-vitro* and computer based (also known as *in-silico)* methods[2]. Alternative test methods can be used both as key-studies, as well as in a weight of the evidence approach, *i.e.* as supporting information. Several guidelines have been released by the European Chemicals Agency (ECHA) to help companies adopt these methods. In particular Chapter R.6 of the "Guidance on information requirements and chemical safety assessment" document deals with QSAR methods and how to use them in compliance with REACH requirements[3].

According to the OECD principles, QSAR models can be applied in the framework of REACH if the following four conditions are met:

- results are derived from a (Q)SAR model whose scientific validity has been established;
- the substance falls within the applicability domain of the (Q)SAR model;
- results are adequate for the purpose of classification and labelling and/or risk assessment;
- adequate and reliable documentation of the applied method is provided.

1.3 Objectives

On the basis of the aforementioned importance of assessing the aquatic toxicity of chemical substances and the information requirements demanded by REACH, this project aims at developing mathematical models for acute aquatic toxicity on *Daphnia magna* and fish using QSAR methods. A workflow of the QSAR strategy is reported in Figure 2.

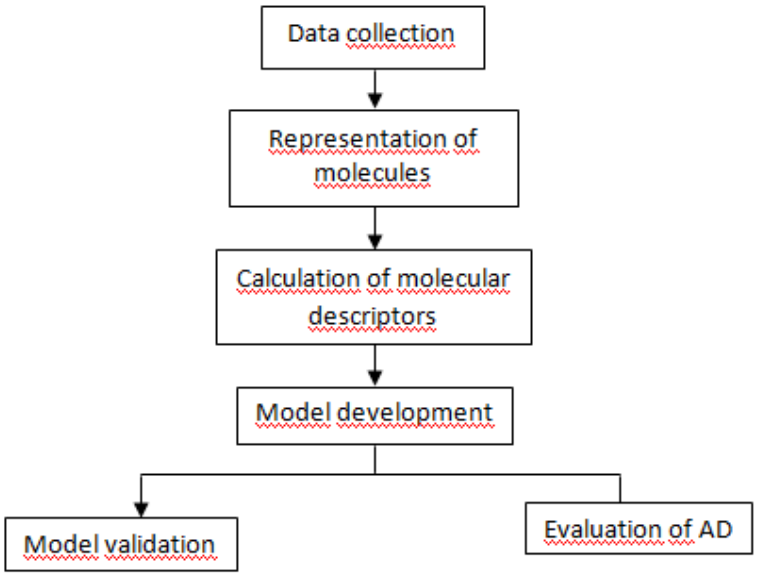


Figure 2: workflow of the strategy adopted to develop QSAR models.

The selected endpoint for both organisms is the LC50, *i.e.* the concentration of chemical that leads to the death of 50% of test organisms. A typical concentration-response curve is reported in Figure 3.

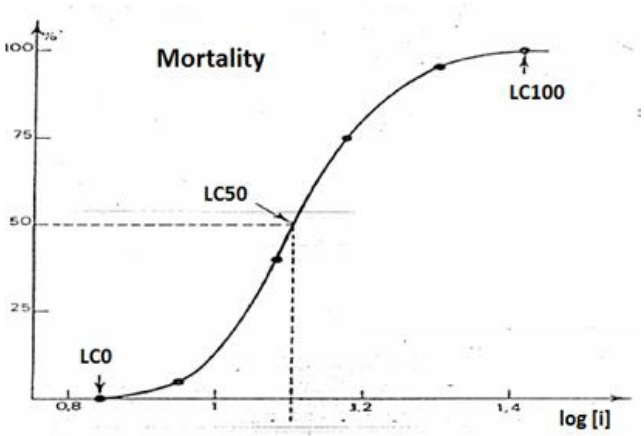


Figure 3: example of concentration-response curve showing the calculation of LC50.

Chapter 2

Methods

2.1 QSAR background

QSAR is the acronym for Quantitative Structure-Activity Relationship. QSAR analysis is based on the theory, according to which, biological activity, or a property (in this case referred to as QSPR), is directly related to molecular structure. According to the congenericity principle, molecules that feature similar structures will possess similar activities/properties, and changes in the structure are expressed by changes in activities/properties.

QSAR analysis implies:

- computational methods to calculate molecular descriptors;
- procedures to select relevant molecular descriptors;
- algorithms for model building.

2.2 Molecular descriptors

"The molecular descriptor is the final result of a logic and mathematical procedure which transforms chemical information encoded within a symbolic representation of a molecule into a useful number or the result of standardized experiment" [9]. The term "useful" has a double meaning: on one hand, it suggests that the number can allow to reach a deeper knowledge in the interpretation of molecular properties; on the other hand, this number can take part in determining a model to predict molecular properties. In fact, a molecular descriptor can be closely correlated to some molecular properties to give highly predictive models, even if its interpretation may be, sometimes, difficult.

The information content of a molecular descriptor depends not only the type of chemical representation from which it is calculated, but also on the algorithm defined for its calculation.

Descriptors are divided into two main categories: experimental measurements and theoretical molecular descriptors that derive from a symbolic representation of the molecule and can be further classified along with the different types of molecular representation. The most striking difference between theoretical and experimentally measured descriptors is that the former ones do not contain any statistical error due to experimental error.

A molecular descriptor must fulfil some mathematical requirements. In particular, basic properties that every descriptor must possess, are:

- invariance to atomic labelling and numeration;

- invariance to roto-translation of the molecule;
- an unambiguous definition, computable by means of algorithms;
- values in a suitable numeric interval for the set of molecules to which it is applicable.

Currently thousands of descriptors have been defined and they can be calculated by means of dedicated software. Each descriptor describes only a part of the whole chemical information included in the real molecule and, consequently, the high number of descriptors is increasing on and on with increasing the complexity of the analyzed chemical systems and properties.

2.3 Multiple Linear Regression (MLR) : OLS and PLS methods

Regression is a mathematical method able to search for the best quantitative functional relationship between a set of variables, x_i, that describe the objects under analysis, and a set of measured responses, y_i, for the objects themselves[10][11]:

$$y = f\,(x_1, x_2, \dots, x_p)$$

The resulting relationship provides information on how variables describing the system are related to the experimental measurement (fitting). Moreover, if the model is able to pass some statistical validation tests, it can be used for the prediction of responses of objects, for which only the independent variables are known. Regression is, therefore, structured in three main steps:

- definition of the model type;
- determination of model parameters;
- evaluation of model reliability.

Ordinary Least Squares (OLS) method provides a mathematical linear relationship between the *y* response and the independent variables x_i, expressed as:

$$\boldsymbol{y} = \boldsymbol{X_m \beta} + \boldsymbol{e}$$

where **β** is the vector of the true coefficients to be estimated, $\mathbf{X_m}$ is the model matrix and **e** is the vector of the errors. A regression model can be developed in the form:

$$y_i = b_o + b_1 x_{i1} + b_2 x_{i2} + \dots + b_p x_{ip}$$

where y_i is the response of the *i-th* object and x_{ij} is the value of the *j-th* variable for the *i-th* object. The regression model is therefore defined by the following equation:

$$\hat{\boldsymbol{y}} = \boldsymbol{X_m b_{ols}}$$

where **ŷ** is the vector of the responses calculated by the model.

Partial Least Squares (PLS) method is a biased regression method useful both when the ratio objects/variables is low (lower than 1), when variables correlated to each other are present and when dealing with several responses at the same time.

Basically, PLS method searches for pairs of principal components (latent variables) in $\mathbf{X_m}$ and **Y** (when **Y** contains more than one response) and searches for the maximum correlation between the principal components [12]. Each principal component is a linear combination of the original variables. Instead of solving the general form for any linear model for **b**:

$$\boldsymbol{y} = \boldsymbol{X_m b}$$

the following equation can be solved for **q**:

$$\boldsymbol{y} = \boldsymbol{X_m V q^t} + \boldsymbol{e}$$

where **V** is the loadings matrix.

Once the regression model has been obtained, some statistical tests to evaluate its reliability are required. Important parameters, for model quality evaluation, are:

Coefficient of determination (R^2)**:**

$$R^2 = 1 - \frac{RSS}{TSS}$$

where RSS is the sum of the residuals of the model and TSS a quantity referred to the average of the response, assumed as reference situation. R^2 expresses the correlation between experimental response and independent variables, represents the variance explained by the model and measures what is normally defined as fitting.

In order to have parameters that measure the predictive ability of a model, it is necessary to use validation and/or cross-validation techniques. There exist several validation methods, whose goal is to search for the optimal complexity of the model, *i.e.* the structure that maximizes its predictive power[10].

The general adopted scheme consists in splitting data into a training set, used to build one or more partial models, and an evaluation set, used to evaluate the model predictive power. These validation techniques differ in the way objects are split. For the purposes of this study, a 5-fold cross-validation was used.

The application of validation techniques makes it possible to calculate parameters to estimate the predictive power of the model. If, in the expression for the coefficient of determination (R^2), RSS is replaced with PRESS, *i.e.* sum of the residual of the model in prediction, one obtains the explained variance in prediction (Q^2):

$$Q^2 = 1 - \frac{PRESS}{TSS}$$

Q^2 value, unlike R^2, does not keep on increasing while increasing the number of variables included in the model (increase of complexity). Thus, the maximum value assumed by this parameter corresponds to the optimal model complexity.

In addition, two parameters, associated to RSS and PRESS, respectively, are:

Root Mean Squared Error in Calculation (RMSEC):

$$RMSEC = \sqrt{\frac{RSS}{n}}$$

Root Mean Squared Error in Prediction (RMSEP):

$$RMSEP = \sqrt{\frac{PRESS}{n}}$$

These two parameters measure the standard deviation of the error in fitting (RMSEC) and in prediction (RMSEP), respectively, and have the advantage of being dimensionally comparable to the studied response.

2.4 Artificial Neural Networks (ANN) and Associative Neural Networks (ASNN)

Artificial Neural Networks (ANNs) are computational models inspired from the brain. In most cases an ANN is an adaptive system that changes its structure based on external or internal information that flows through the network during the learning phase. ANNs can learn and generalize from experiences, and they can abstract essential information from data [13].

Based on the learning algorithm, neural networks can be classified into three major categories as follows:

- in supervised learning, pairs of input and target vectors are required to train networks, so that appropriate outputs that correspond to input signals are generated accordingly. When an input vector is applied, the error between the output of the neural network and its target output is calculated, which is used to tune weights in order to minimize the error.
- Unsupervised learning does not require target vectors for the outputs. Without input–output training pairs as external teachers, unsupervised learning is self-organized to produce consistent output vectors by modifying weights.
- Some neural networks employ hybrid learning. For example, counter-propagation networks and RBF networks use both supervised (at the output layer) and unsupervised learning (at the hidden layer).

Typically neural networks have a three-layer architecture that comprises a layer of input neurons connected to a layer of hidden neurons, which in turn are connected to an output layer (Figure 4).

Being a machine-learning method, neural networks need to be trained on the training data in order to define the vectors of the weights. This is typically accomplished by presenting the data to the network several times, in an iterative fashion, and allowing the network to learn from the data and set the weights. In case of supervised learning, the criterion that drives the setting of the weights is the minimization of the error at the output layer.

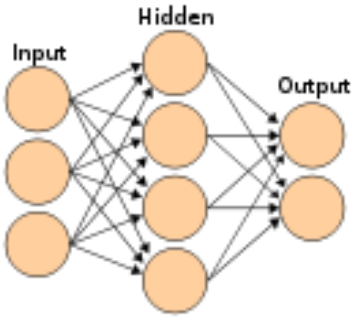


Figure 4[6]: architecture of a three-layer neural network.

An associative neural network (ASNN) is a combination of an ensemble of the feed-forward neural networks and the K-nearest neighbor technique[14]. ASNN, thus, combine a memory-less approach, provided by ANNs (after training is complete all information about the input patterns is stored in the neural network weights and input data are no longer needed), and methods such as the *K*-Nearest Neighbors (K-NN) that represent the memory-based approaches. These approaches keep in memory the entire database of examples and their predictions are based on some local approximation of the stored examples.

ASNN have been proposed with the aim of reducing the high bias that can be associated to certain regions of the space. In order to improve the performance of a neural network in such regions of the space, the ensemble predictions $\bar{z}_i$ are corrected according to the following formula:

$$\bar{z}'_i = \bar{z}_i + \frac{1}{k} \sum_{j \in N_k(\boldsymbol{x})} (y_j - \bar{z}_j)$$

where y_i are the experimental values, $N_k(\mathbf{x})$ is the collection of the k nearest neighbours of $\boldsymbol{x}$ among the input vectors in the training set $\{\mathbf{x}_i\}^N_{i=1}$ determined using Spearman non-parametric rank correlation coefficient r_{ij}.

2.5 *K*-Nearest Neighbours (*K*-NN)

K-NN is a non-parametric method based on the concept of similarity. The predicted value for any new object is computed from the values of its *k* nearest neighbours. *K*-NN method does not provide a global mathematical model, *i.e.* a function to be applied to unknown objects, rather it is a local approach. Typically, the formula adopted for quantitative responses is in the form:

$$y_{(k)} = \frac{1}{k} \sum_{j \in N_k(\boldsymbol{x})} y(\boldsymbol{x_j})$$

where $y_{(k)}$ is the predicted value for $\mathbf{x}$, $N_k(\mathbf{x})$ is the collection of the k nearest neighbours of $\boldsymbol{x}$ among the input vectors in the training set $\{\mathbf{x}_i\}^N_{i=1}$ using Euclidean distance and $y(\boldsymbol{x_j})$ are the experimental values of the k neighbours.

2.6 Support Vector Machines (SVM)

Support Vector Machines are a set of supervised learning methods used in classification and regression analysis. In its original definition a SVM is a binary classifier that is able to recognize the border between objects belonging to two different classes. As well as ANN, given a set of training data, defined by independent and dependent variables, an SVM tries to derive a mathematical function to correctly classify (qualitative response) or calculate (quantitative response) the dependent variable, in such a way to minimize the bias. Once the SVM has been trained, new objects can be given as input and the output is the membership to a particular class, or a numerical value for a continuous variable.
The basic functioning of SVM consists in a projection of the objects in a multidimensional space and a search for the separation hyper plane in this space. The separation hyper plane maximizes the distance between the two classes, considering the closest objects.
The initial *d*-dimensional space R^d is transformed into a $\hat{d}$-dimensional space (where $\hat{d} > d$, $d<\infty$ and the transformation can be both linear or not linear) and in this new space a quadratic problem subject to some constraints is faced.
From a theoretical point of view[15], let's start considering the training set defined as:

$$(x_i, y_i), \dots, (x_l, y_l)\ \mathrm{C}\ XR$$

where *X* represents the space of input patterns R^d, ***x*** the input vector and ***y*** the target vector. The regression with ϵ-SV consists in finding the function *f(x)* which is as "flat" as possible and whose maximum distance from the target patterns y_i belonging to the entire training set is less than or equal to ϵ. This means that a criterion on acceptability or tolerance of the error is defined. Figure 5 provides a depiction of what aforementioned, in which the points outside the dark region contribute to the cost-function, since the error is higher than ϵ.

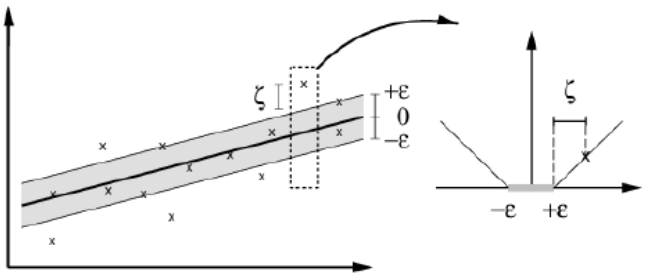


Figure 5[58]: cost function.

Chapter 3

Data

Four databases were identified as sources of experimental data and are briefly described and referenced below.

ECOTOXicology database (ECOTOX): the ECOTOXicology database is a source for locating single chemical toxicity data for aquatic life, terrestrial plants and wildlife. ECOTOX was created and is maintained by the U.S.-EPA, Office of Research and Development (ORD), and the National Health and Environmental Effects Research Laboratory (NHEERL)[16].

ECETOC: the ECETOC Aquatic Toxicity (EAT) database has been updated, mainly from data published between 1992 and 2000, to include information on the toxicity of substances to aquatic species in fresh and saline waters. It was created by the European Centre for Ecotoxicology and Toxicology of Chemicals[17].

OASIS: the OASIS database contains measured data for aquatic species. Experimental results for aquatic toxicity are developed and donated by the Laboratory of Mathematical Chemistry, Bulgaria, U.S.-EPA, University of Knoxville, Tennessee and MITI Japan[18]. This database was downloaded from the QSAR Toolbox[19] version 2.3.

Aquatic Japan MoE: this database contains experimental results on aquatic toxicity based on tests performed within the Japanese Existing Chemicals Programme. The results are also published in the Japanese Chemical Risk Information Platform (CHRIP)[20]. This database was downloaded from the QSAR Toolbox[19] version 2.3.

The downloaded databases were processed by means of *ad-hoc* designed workflows of KNIME[21] for each database. The workflows were designed in order to retain only information defining relevant experimental conditions, such as test species, duration, temperature and pH (where available). Moreover, the web service to ChemSpider[22] was used to retrieve the SMILES of every compound. The queries were performed giving both CAS registry numbers and names as input. In case the results did not match, a comment to warn the user of this situation was added.

Data were uploaded on the OCHEM platform[4] and made publicly available. Since the upload procedure is highly automated, not only values of LC50 were uploaded, but also of other short-term and long-term properties. Table 1 summarizes the uploaded data.

The design of the KNIME workflow and the upload on OCHEM of the ECOTOX database was performed by Kamel Mansouri.

Table 1: Number of data uploaded on OCHEM for each source database.

	EC50	**LC50**	**NOEC**	**LOEC**
ECETOC	4342	3167	816	827
OASIS	265	2180	-	-
Japan MoE	1230	370	1156	29

The actual number of records for each property in the source databases is slightly higher than the number of records uploaded on OCHEM, the reason being an automatic check for internal and external duplicates performed by OCHEM during the upload steps.

In the ECETOC database the concentration of chemical resulting in the death of 50% of test organism (LC50) is reported as EC50 with mortality as observed effect. These data have been re-uploaded on OCHEM as LC50 values in order to combine them with data from other sources.

Before uploading the data, more than 400 scientific publications, referenced in the databases, were retrieved and uploaded on OCHEM. This allowed most of the data to be linked to the original study where they were published. The availability of the original articles is useful since it allows to check the values used to compile the databases for input errors.

Chapter 4

Results and discussion

4.1 Preparation of data on *Daphnia magna*

As aforementioned, the endpoint considered in this study is the LC50. In order to have consistent data for modelling, it is important to define the test organism and the test duration. For *Daphnia magna* the chosen duration was 48 hours, according to the most commonly used experimental condition. The dataset for LC50 on *Daphnia magna* with a test duration of 48 hours consists in 1459 records, comprising 536 unique molecules. This means that for some compounds more than one experimental value was available. This dataset contains also 29 records previously uploaded on OCHEM by other users.

A filtering stage was then applied to this set in order to have a more robust set for modelling. For 62 records no structure was provided during the upload and no matching structure was retrieved from PubChem[23] by OCHEM. These records were excluded from the set.

Afterwards, the application of a set of structural alerts on OCHEM provided insight into the composition of the dataset, which is reported in Table 2.

Table 2: composition of the dataset on Daphnia magna.

Organic chemistry molecules[1]	Non-organic chemistry molecules	Perfluoroalkylated compounds (PFCs)	Metals	Molecules not supported in ALOGPS program
452	84	5	75	85

1 Molecules including H, C, N, O, S, P, Si, F, Cl, Br and I atoms.

Only compounds including atoms typical of organic chemistry were retained at this stage. Also NH_3, $NH_4^+Cl^-$, I_2, Cl_2 and Br_2, and were removed.

Since there was not enough time to manually check all the experimental values, it was decided to develop a preliminary model and check only outliers, both in the prediction and in the descriptors space. A useful feature of OCHEM is the possibility to automatically retrieve all the records for the same molecule. This allowed to identify some values which were wrongly input in the source database. If the value or the units in the original publication were different from those reported in the database, this information was corrected on OCHEM.

Few other records were removed because of three distinct reasons:

- the observed effect reported in the original publication was different than the one input in the database;
- the molecule was not present in the original publication;
- the values were expressed as higher than (> [i]).

Table 3 summarizes the number of removed and corrected values for the aforementioned reasons.

Table 3: number of corrected and removed records because of errors in the source database.

Not in reference	Different observed effect	Corrected	Higher than (>[i])
12	5	7	5

The refined dataset used for modelling contains 876 records for 435 unique compounds.

4.2 QSAR analysis of toxicity on *Daphnia magna*

The online platform OCHEM was used to carry out the QSAR analysis. The dataset includes few salts and disconnected structures. The approach implemented in OCHEM to treat these species implies that only the bigger substructure (on the basis of the atom count, hydrogen atoms ignored) is retained and used to calculate molecular descriptors.

Initially all the records were considered for the modelling stage, meaning that some molecules were present in the dataset with more than one experimental value.

The experimental response was transformed in logarithmic scale of molarity (Log(mol/L)) for the subsequent QSAR analysis.

Different types of molecular descriptors implemented in OCHEM were calculated and used to derive QSAR models. In particular, the following descriptors were used:

- CDK[24];
- DRAGON[25];
- ALogPS[26] and OEstate[27];
- ISIDA fragments[28];
- Mera and Mersy[29];
- ChemAxon descriptors[30];
- Inductive descriptors[31];
- Adriana[32];
- Spectrophores[33];
- Shape Signatures[34];
- QNPR[35].

A variable reduction step was carried out before the development of the models in order to reduce the number of analyzed molecular descriptors. This filtering procedure was based on 4 criteria, which are listed below:

- absolute values: only descriptors with absolute values lower than 999999 were retained;
- variance: descriptors with variance lower than 0.01 were excluded;
- unique values: only descriptors with more than 2 unique values were retained;

- pairwise correlation: if the Pearson's correlation coefficient was higher than 0.95, the descriptors were grouped.

Several different methods (see Chapter 2) representing different approaches (linear regression, machine-learning, local approaches) were used to derive QSAR models on the OCHEM platform.

The derived models were validated using a 5-fold cross-validation. As aforementioned, all the records were used. This fact should not affect the cross-validation since OCHEM splits the dataset into training and validation sets using the molecule ID rather than the record ID. This means that all the records of one molecule are assigned to the same set. This avoids the bias that can be originated if the same molecule is present in both sets. In this case the molecule being predicted to test the predictive power of the model, had already been used to train the model itself, leading to an overestimation of the predictive power.

The results presented below refer to models develop using the aforementioned descriptors separately, because a combination of them did not produce any improvement in the results. Table 4, 5 and 6 report the R^2, Q^2 and RMSECV values, respectively, for the complete list of models obtained on the OCHEM platform using all records. Afterwards, the best model is analyzed more in details.

Table 4: R^2 of the models developed on OCHEM using all records.

	ANN	ASNN	*K*-NN	SVM	FSMLR	MLRA	PLS
CDK	0.62	0.65	0.54	0.58	0.53	0.43	0.18
DRAGON	0.64	0.64	0.62	0.63	0.54	0.45	0.04
ALogPS,OEstate	0.58	0.58	0.55	0.56	0.39	0.47	0.51
ISIDA	0.56	0.56	0.49	0.49	0.33	0.37	0.47
Mera, Mersy	0.59	0.62	0.55	0.59	0.48	0.53	0.13
ChemAxon	0.62	0.63	0.47	0.61	0.46	0.51	0.56
Inductive	0.57	0.60	0.49	0.54	0.06	0.34	0.39
Adriana	0.61	0.62	0.50	0.39	0.40	0.34	0.00
Spectrophores	0.45	0.49	0.47	0.53	0.16	0.19	0.23
ShapeSignatures	0.21	0.27	0.59	0.47	0.25	0.41	0.39
QNPR	0.52	0.52	0.43	0.44	0.44	0.43	0.48

Table 5: Q^2 of the models developed on OCHEM using all records.

	ANN	ASNN	*K*-NN	SVM	FSMLR	MLRA	PLS
CDK	0.62	0.63	0.52	0.57	0.49	0.38	-0.09
DRAGON	0.64	0.64	0.60	0.61	0.49	0.44	-9.64
ALogPS,OEstate	0.57	0.57	0.52	0.54	0.25	0.46	0.47
ISIDA	0.56	0.56	0.48	0.45	0.25	0.29	0.44
Mera, Mersy	0.59	0.61	0.54	0.59	0.45	0.52	-0.18
ChemAxon	0.62	0.60	0.38	0.59	0.45	0.50	0.55
Inductive	0.57	0.59	0.46	0.53	-1.97	0.31	0.37
Adriana	0.61	0.62	0.46	0.28	0.35	0.20	-15.74
Spectrophores	0.45	0.48	0.41	0.52	0.09	0.12	0.20
ShapeSignatures	-0.07	-0.05	0.58	0.46	-0.46	0.41	0.39
QNPR	0.51	0.51	0.40	0.43	0.40	0.41	0.46

Table 6: RMSECV of the models developed on OCHEM using all records.

	ANN	ASNN	*K*-NN	SVM	FSMLR	MLRA	PLS
CDK	1.10	1.08	1.24	1.18	1.28	1.40	1.86
DRAGON	1.06	1.06	1.12	1.10	1.26	1.33	5.78
ALogPS,OEstate	1.17	1.17	1.23	1.21	1.54	1.31	1.29
ISIDA	1.18	1.18	1.28	1.31	1.54	1.49	1.32
Mera, Mersy	1.14	1.11	1.21	1.14	1.31	1.22	1.93
ChemAxon	1.10	1.12	1.39	1.14	1.32	1.25	1.19
Inductive	1.17	1.14	1.30	1.21	3.06	1.47	1.41
Adriana	1.11	1.10	1.31	1.50	1.44	1.59	7.27
Spectrophores	1.32	1.28	1.36	1.23	1.69	1.67	1.58
ShapeSignatures	1.82	1.80	1.15	1.31	2.12	1.35	1.37
QNPR	1.24	1.24	1.37	1.34	1.37	1.36	1.30

These models indicate that the methods that give better results are ANN and ASNN; from the descriptors side, the best results are obtained with DRAGON descriptors, followed by CDK and ChemAxon descriptors. Since the aim is to develop predictive QSAR models, the model with highest Q^2 was chosen as best.

ANN and ASNN models on DRAGON descriptors gave the same results. For the sake of simplicity, the ANN model was considered for further analysis.

ANN with DRAGON descriptors: the neural network was trained with 1000 iterations, using the SuperSAB training method and 3 neurons in the hidden layer. The number of retained descriptors after the variable reduction step used to train the network was 1413.

2 molecules comprising a total of 5 records were ignored during the development of the model because of failure during the optimization of the structure (1,2,3,4,5,6-hexachlorocyclohexane, 3 records) or the calculation of the descriptors (fullerene, 2 records). Thus, 871 records, comprising 434 unique compounds was used to develop the model.

The plot of measured versus calculated values for LC50 is reported in Figure 6.

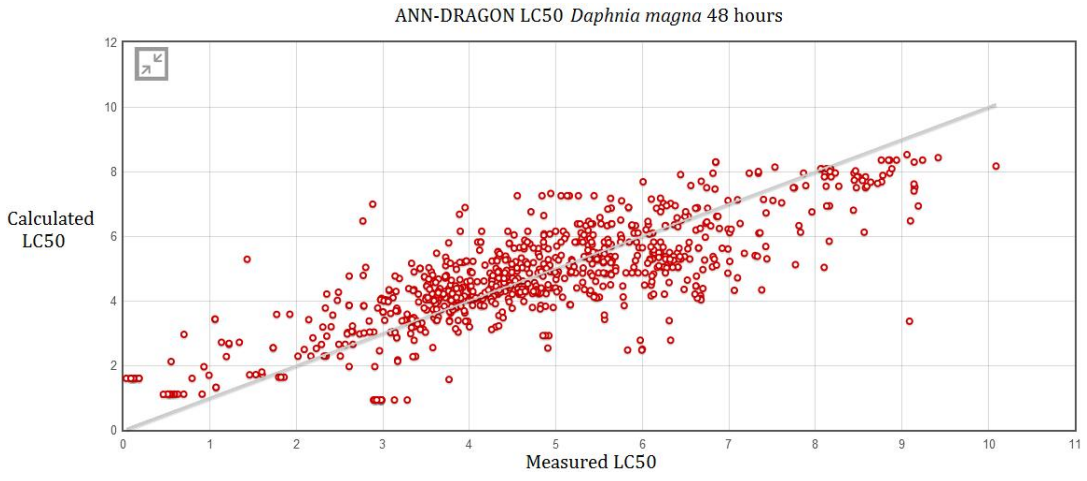


Figure 6: calculated versus measured values for LC50 on Daphnia magna (48 hours) obtained with the ANN model.

The statistics for the model are provided in Tables 4, 5 and 6.

Even though this is the best model developed, it can be seen that the statistics both for fitting and prediction are not very high.

From the plot of calculated versus measured values, it is clear that the model does not perform well on certain molecules, for which the error in calculation is very high. Therefore, the model cannot be said to provide reliable predictions over the entire chemical space represented in the training set. In order to provide an estimation of the applicability domain of the model in terms of chemical space, the correlation between function groups and performance of the model was analyzed.

The dataset was screened against the library of functional groups implemented on OCHEM ToxAlerts [Shusko Y. JCIM, in press] and published by Haider[36]. 87 functional groups were identified in the training set. Some analyses were carried out. First, the number of molecules featuring each functional group was calculated. This number was correlated with the RMSEC for each functional group. Figure 7 reports a bubble plot to highlight the correlation between number of molecules for each functional and the RMSEC.

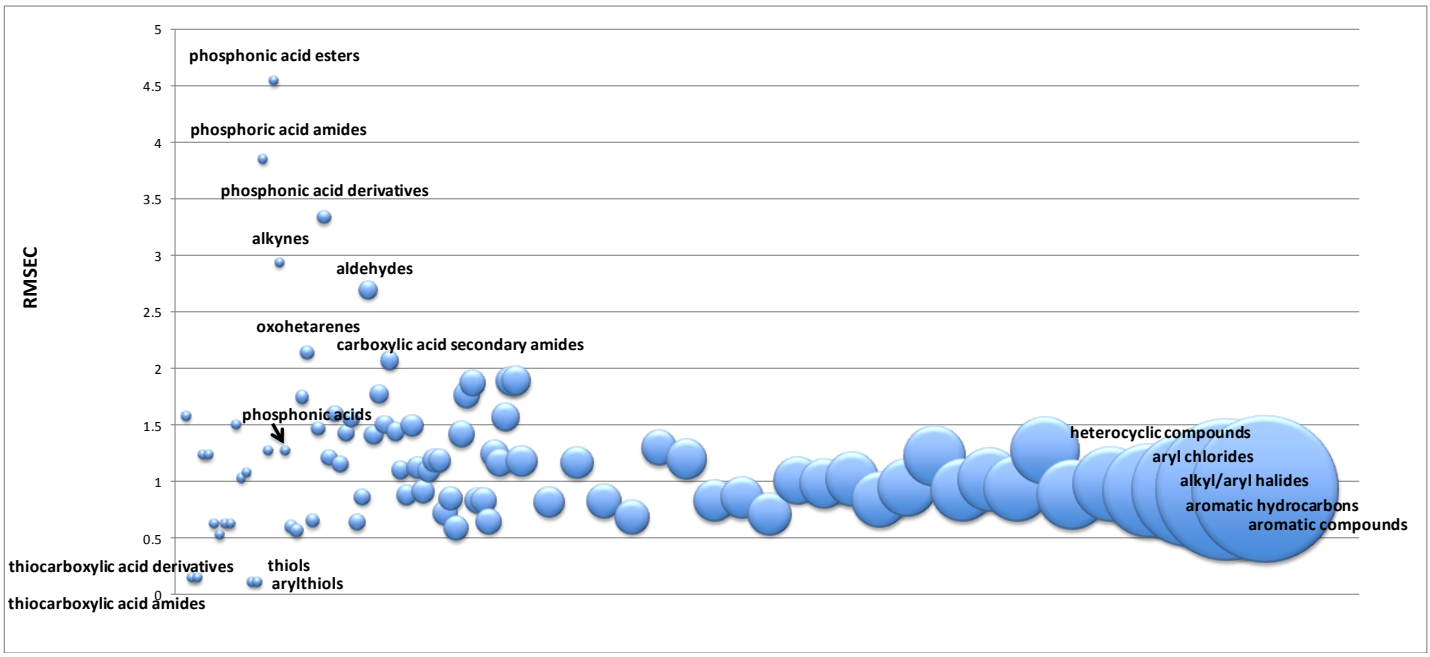


Figure 7: bubble plot of number of compounds for each functional group and RMSEC values. The size of the bubbles is proportional to the number of molecules. The x axis is an enumerator used to sort the bubbles in ascending order of the number of molecules.

Some considerations can be drawn from the bubble plot. One can see that the RMSEC values for all the functional groups that are well represented in the dataset (on the right side of the plot) is approximately equal to or lower than the RMSEC on the entire dataset (1.06). These well represented functional groups include aromatic and heterocyclic compounds as well as halides. One can also see that all the functional groups possessing a very high RMSEC are not well represented in the dataset. The largest errors are made on derivatives of phosphoric and phosphonic acids. The RMSEC

on phosphonic acids (1.28) is, instead, only slightly larger than that on the entire dataset. Also carboxylic acid secondary amides are not well predicted (RMSEC = 2.06). 3 more functional groups have a much larger RMSEC than that on the entire dataset, namely alkynes, aldehydes and oxohetarenes. This set of functional groups can be considered outside the applicability domain of the model and includes mainly derivatives of acidic functions. On one hand, as expected, all functional groups with very large RMSEC values are not well represented in the dataset; on the other hand, it is not true the opposite, *i.e.* that all functional groups not well represented are associated with a large RMSEC. In fact the predictions on thiols and thiocarboxylic amides is very good and the RMSEC values are much lower than the RMSEC on the entire dataset, 0.11 and 0.15 for thiols and thiocarboxylic amides, respectively. However, the molecule featuring the thiol and arylthiol structural alerts (Figure 8) is included also in the aromatic compounds, heterocyclic compounds, aromatic heterocyclic compounds, alcohols/phenols and phenols groups. These latter are well represented in the dataset and this fact can explain the good performance of the model on this molecule. On the contrary, the molecule featuring the thiocarboxylic acid derivative and thiocarboxylic acid amide structural alerts possesses no other functional groups. Therefore the good performance of the model cannot be explained in terms of other well represented functional groups within the molecule. However, it must be considered that for these functional groups only one or two molecules are present. Therefore, it is not possible to derive general conclusions. The good or poor performance of the model on these functional groups can be also imputed to chance.

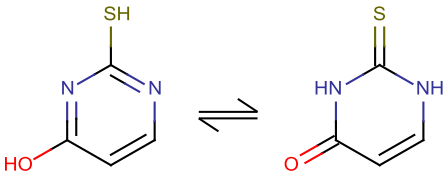
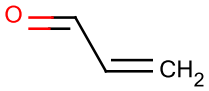
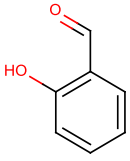
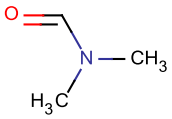
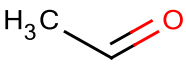
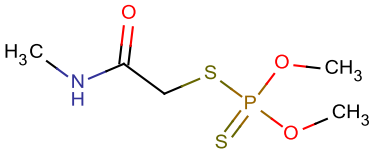
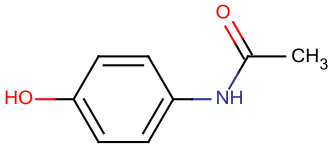
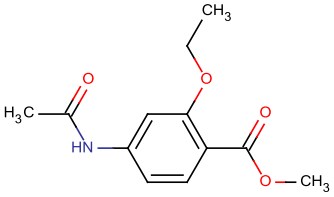
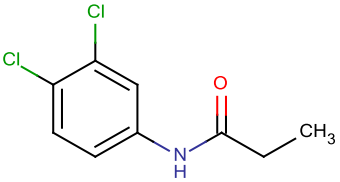
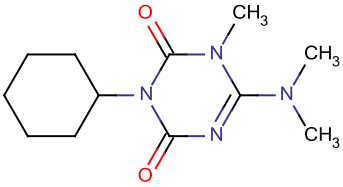
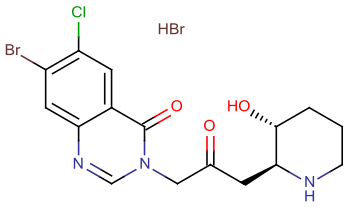
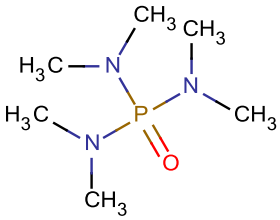
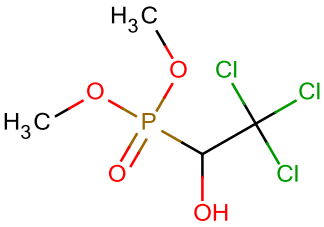
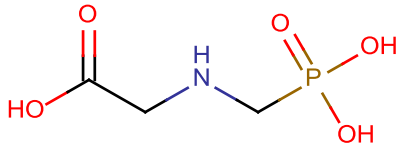
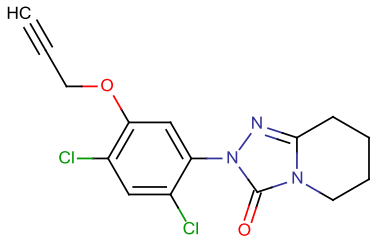


Figure 8: molecule with thiol moiety.

Most of the molecules belonging to the 7 outlier functional groups include also well represented moieties, such as aromatic/heterocyclic rings and halides, but still the models' performance is very poor. Table 7 reports structures, names and average LC50 values for the 14 molecules belonging to the 7 groups outside the applicability domain.

Table 7: structure, name and LC50 (in –Log(mol/L)) values of the 14 molecules possessing the functional groups outside the applicability domain of the model.

acrolein	salicylaldehyde
	
LC50 = 5.94	LC50 = 4.45

dimethylformamide  LC50 = 0.70	acetaldehyde  LC50 = 0.55
dimethoate  LC50 = 5.00	acetaminophen  LC50 = 6.67
ethopabate  LC50 = 3.07	propanil  LC50 = 4.64
hexazinone  LC50 = 5.56	halofuginone hydrobromide  LC50 = 7.44
hexamethylphosphoramide  LC50 = 1.43	trichlorfon  LC50 = 7.70
glyphosate  LC50 = 3.91	azafenidin  LC50 = 3.95

In order to analyze whether there is a mechanistic reason that can explain the poor performance of the model on the molecules belonging to the 7 outlier functional groups, these molecules were screened against the structural alerts for aquatic toxicity

of the Verhaar scheme[37] using the software ToxTree[38]. The results, reported in Table 8, do not provide an insight into the reactivity of these molecules. In fact, it was not possible to classify most of the molecules. All aldehydes were classified as acting by non specific mechanisms, while only one carboxylic acid amide was identified as molecule acting through a specific mechanism.

Table 8: screening of molecules belonging to the outlier functional groups using the Verhaar scheme.

Molecule	Functional group	Classification
acrolein	aldehydes	unspecific reactivity
salicylaldehyde	aldehydes	unspecific reactivity
N,N-dimethylformamide	aldehydes	unspecific reactivity
acetaldehyde	aldehydes	unspecific reactivity
dimethoate	carboxylic acid secondary amides	specific reactivity
acetaminophen	carboxylic acid secondary amides	not classifiable
ethopabate	carboxylic acid secondary amides	not classifiable
propanil	carboxylic acid secondary amides	not classifiable
hexazinone	oxohetarenes	not classifiable
Halofuginone hydrobromide	oxohetarenes	not classifiable
hexamethylphosphoramide	phosphoric acid amides	not classifiable
trichlorfon	phosphonic acid derivatives, phosphonic acid esters	not classifiable
glyphosate	phosphonic acid derivatives	not classifiable
azafenidin	alkynes	not classifiable

However, two considerations can be drawn. First, aldehydes are in general reactive species, thus they can undergo a number of reactions resulting in a toxic action that may deviate from that of narcotic toxicants. Secondly, all the other molecules (not belonging to the aldehydes class) are pesticides, herbicides, chemosterilant or analgesic substances. Therefore, they represent a peculiar class of compounds. For example, Kim *et al.*[39] report *Daphnia magna* to be particularly sensitive to certain drugs including acetaminophen (paracetamol), which in fact has a high toxicity. Some of these molecules, indeed, have high toxicity (acrolein, acetaminophen, halofuginone hydrobromide and trichlorfon) compared to the average of the dataset (4.76). On the contrary, 3 molecules have a very low toxicity, namely acetaldehyde, dimethylformamide and hexamethylphosphoramide. The toxicity values of the remainders are close to the mean LC50. However, specific mechanisms of actions may be the cause of their deviation from the general behaviour.

The second analysis that was carried out concerns the single molecules, instead of functional groups. In this case, the correlation between the number of functional groups possessed by each compound and the RMSEC for that compound was analyzed. Figure 9 reports a bar plot of the RMSEC values for each molecule. What emerges from this plot is that, in practice, there is no correlation between the number of functional groups within each molecule and the RMSEC. It was expected that the RMSEC values would increase together with an increasing number of functional groups per molecule. Unexpectedly, the molecules featuring the largest RMSEC values have only a moderate

number of moieties. However the bar plot of the average RMSEC values for each block (Figure 10) seems to show that the average RMSEC follows a parabolic trend, indicating that the error is smaller for molecules featuring 2 to 6 functional groups. This is reasonable because most of the molecules in the dataset have such number of functional groups. Molecules with just one or many moieties, instead are less frequent and had therefore less weight in the model calibration. Moreover, very complex molecules, as well as simple molecules with particular moieties, can show a different toxic behaviour and therefore the prediction of their toxic activity is more difficult.
However, it must be kept in mind that for this analysis only the number of different functional groups within each molecule was considered, irrespective of the number of instances.
A further analysis that could not be carried out for timing issues would be to analyze the correlation between the RMSEC of each molecule and the overall number of functional groups, which can be calculated as:

$$Nf = \sum_{i=1}^{N} Tf * Nf$$

where the sum runs over the number of functional groups, *Tf* is a binary variable that codes the presence/absence of a particular functional group and *Nf* is the number of instances of that functional group.

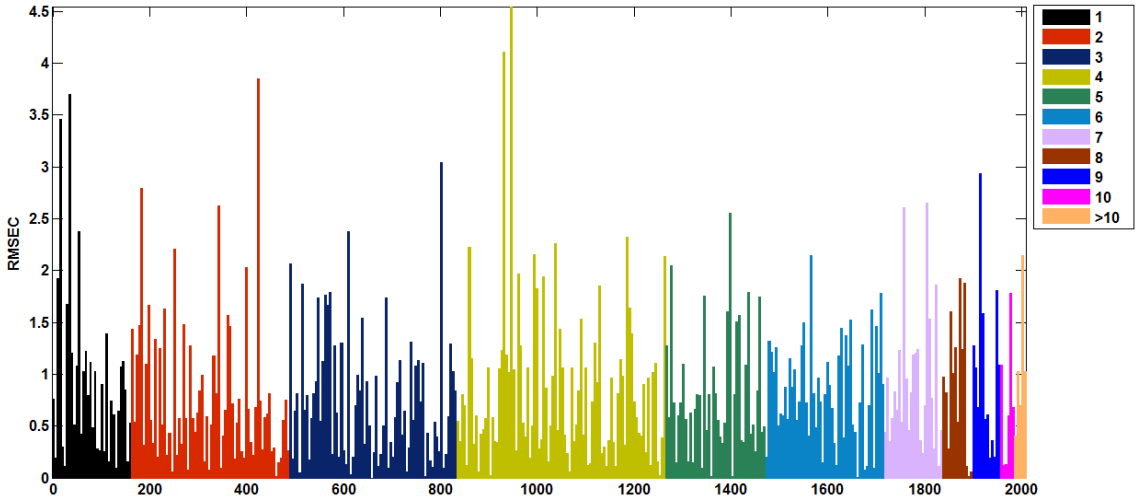


Figure 9: bar plot of RMSEC values for each molecule. Colours indicate the number of functional groups per molecule according to the legend.

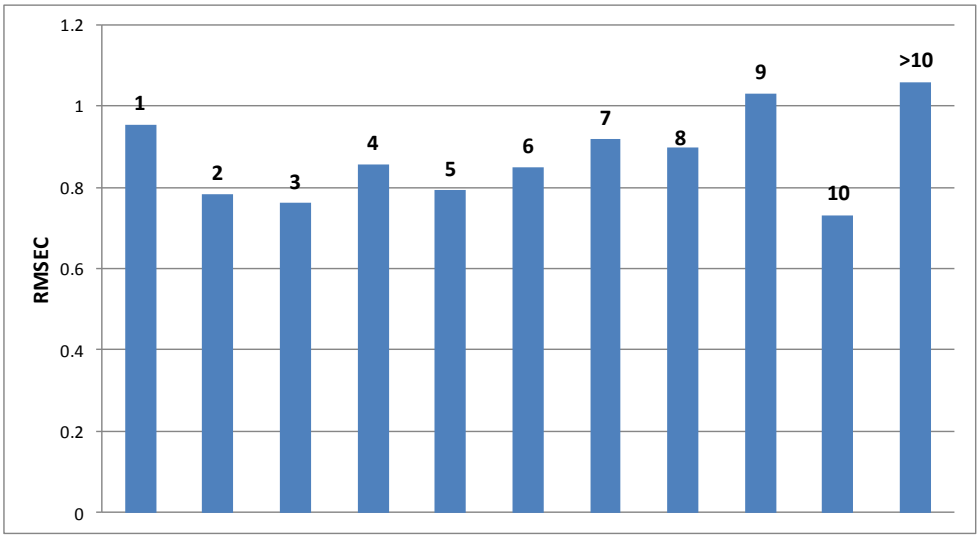


Figure 10: bar plot of the average RMSEC for each block of number of functional groups. The number of functional groups is indicated on top of the bars.

Eventually, the correlation between the standard deviation of the experimental response within each functional group and the RMSEC for that functional group was studied. Figure 11 reports the RMSEC values versus the standard deviations for each functional group. It can be seen that there is no clear correlation (R^2 = 0.05). This result was expected since the molecules possessing a common functional group can be very different in their structure and therefore possess also very different LC50 values. In this case the standard deviation is large, but this does not imply that the model should have poor performance on this subset of molecules. The lack of correlation is highlighted also by the previously identified 7 outlier functional groups. In fact, their RMSEC values are very high irrespective of their standard deviation, that ranges from 0 for alkynes and phosphoric acid amides, to 2.62 for aldehydes.

In the same way, also the correlation between RMSEC and standard deviation of the experimental response for each molecule was studied (Figure 12). Many molecules have only one experimental value, therefore the standard deviation associated with them is 0. The red line in the plot indicates a region in which it seems there is a certain degree of correlation: for the molecules lying along this line the RMSEC increases together with the standard deviation. This situation was expected, because if the standard deviation of the experimental values is high (very different experimental values), necessarily also the prediction error will be high. This is because the error in prediction cannot be lower than the deviation of the experimental values.
However, there are many molecules that do not follow this trend, thus it is not possible to state that there is an overall correlation between RMSEC and standard deviation.

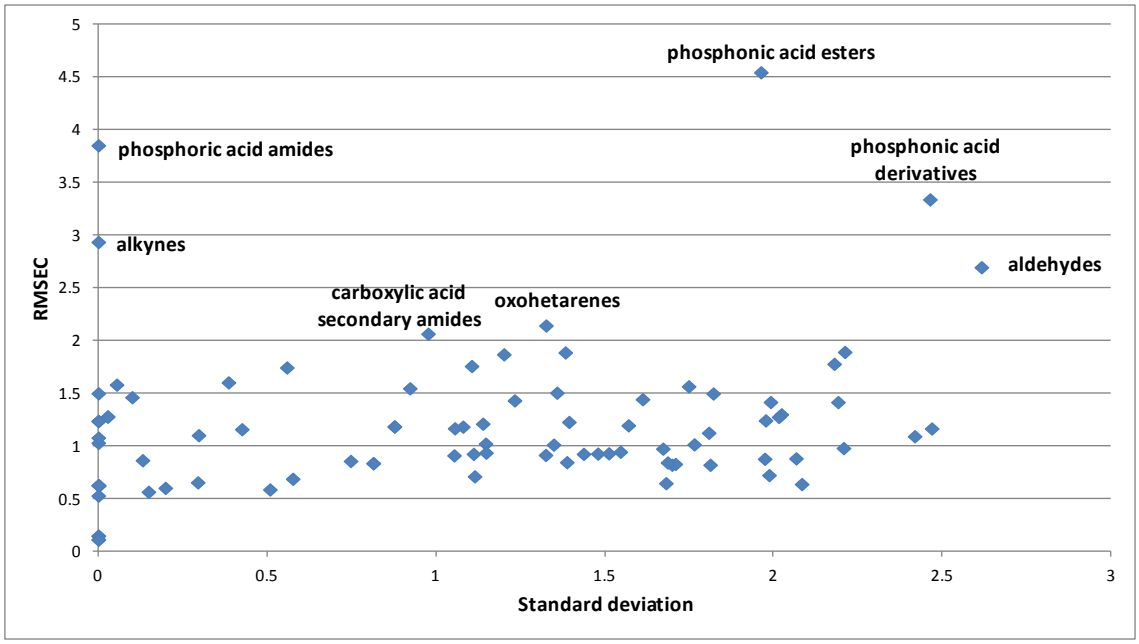


Figure 11: RMSEC values versus the standard deviation for each functional group.

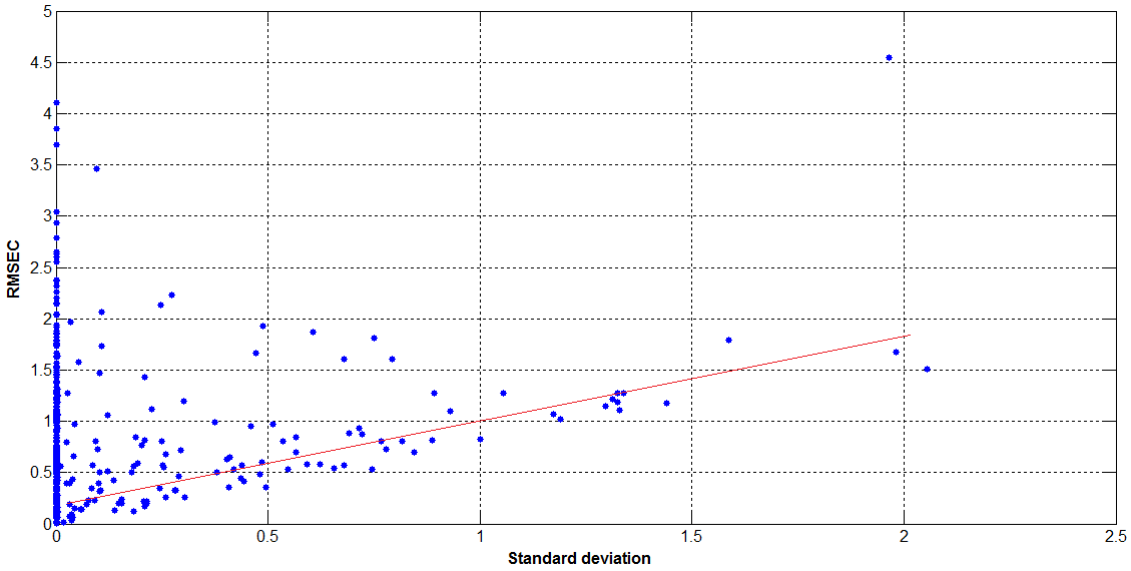


Figure 12: RMSEC versus standard deviation for each molecule.

In addition to the previously presented models, developed using all the records for each molecule, another set of models was developed using only one experimental value for each molecule. Since there was no time to analyze all the experimental values and choose the most reliable, the retained measurement was randomly chosen using OCHEM.

The response was transformed in logarithmic scale of molarity (Log(mol/L)) and the same methods and descriptors aforementioned were used. The statistics of the developed models are presented in Tables 9, 10 and 11.

Table 9: R^2 of the models developed on OCHEM using only one record per molecule.

	ANN	ASNN	*K*-NN	SVM	FSMLR	MLRA	PLS
CDK	0.49	0.51	0.50	0.50	0.34	0.37	0.27
DRAGON	0.54	0.54	0.51	0.53	0.36	0.36	0.10
ALogPS,OEstate	0.51	0.51	0.48	0.45	0.40	0.14	0.48
ISIDA	0.46	0.46	0.43	0.46	0.31	0.25	0.42
Mera, Mersy	0.49	0.50	0.47	0.46	0.37	0.34	0.07
ChemAxon	0.46	0.50	0.40	0.55	0.36	0.38	0.35
Inductive	0.36	0.47	0.33	0.43	0.23	0.27	0.28
Adriana	0.45	0.47	0.45	0.35	0.32	0.24	0.05
Spectrophores	0.31	0.39	0.39	0.36	0.21	0.17	0.14
ShapeSignatures	0.29	0.39	0.35	0.26	0.06	0.27	0.22
QNPR	0.44	0.43	0.44	0.37	0.28	0.42	0.40

Table 10: Q^2 of the models developed on OCHEM using only one record per molecule.

	ANN	ASNN	*K*-NN	SVM	FSMLR	MLRA	PLS
CDK	0.49	0.48	0.48	0.49	0.22	0.35	0.26
DRAGON	0.54	0.52	0.50	0.52	0.31	0.33	-0.33
ALogPS,OEstate	0.51	0.51	0.46	0.44	0.39	-0.45	0.46
ISIDA	0.46	0.46	0.41	0.46	0.14	0.10	0.41
Mera, Mersy	0.49	0.50	0.47	0.45	0.35	0.32	-0.07
ChemAxon	0.46	0.47	0.38	0.54	0.36	0.37	0.33
Inductive	0.36	0.45	0.31	0.42	0.21	0.25	0.26
Adriana	0.45	0.45	0.44	0.30	0.32	0.13	-1.04
Spectrophores	0.31	0.36	0.39	0.35	0.15	0.10	0.08
ShapeSignatures	0.29	0.37	0.34	0.24	-0.21	0.27	0.22
QNPR	0.44	0.42	0.40	0.32	0.25	0.42	0.39

Table 11: RMSEC of the models developed on OCHEM using only one record per molecule.

	ANN	ASNN	*K*-NN	SVM	FSMLR	MLRA	PLS
CDK	1.19	1.20	1.20	1.19	1.48	1.35	1.44
DRAGON	1.12	1.14	1.17	1.14	1.37	1.36	1.91
ALogPS,OEstate	1.16	1.16	1.22	1.24	1.29	1.99	1.21
ISIDA	1.21	1.22	1.28	1.22	1.53	1.57	1.27
Mera, Mersy	1.18	1.17	1.20	1.22	1.33	1.36	1.71
ChemAxon	1.21	1.20	1.30	1.12	1.32	1.31	1.35
Inductive	1.32	1.22	1.37	1.25	1.47	1.44	1.42
Adriana	1.23	1.23	1.23	1.39	1.37	1.54	2.36
Spectrophores	1.37	1.32	1.29	1.33	1.52	1.56	1.59
ShapeSignatures	1.39	1.30	1.33	1.44	1.81	1.41	1.45
QNPR	1.24	1.26	1.28	1.36	1.43	1.26	1.29

The statistics of the models obtained using only one record for each molecule are slightly worse than those obtained on all available experimental values. This result is quite surprising since the presence of multiple values was expected to negatively affect the results, this being mainly due to large standard deviations of the experimental

response for some molecules. Some hypotheses can be raised to explain the worsening of the models on this dataset:

1. prior published data provide a benchmark for new measurements. Thus, the availability of already existing measurements can help in the detection of errors during the measurement itself. The reference provided by prior data generates a process that can lead to good quality data. On the contrary, errors present in measurements taken for the first time are harder to detect and these data may be regarded as reliable.
2. Important molecules were present with several experimental values. When developing models using all the records, these molecules acquired increased weight and helped in the definition of some crucial features.
3. somehow the cross-validation step introduces a bias when using multiple values for one molecule. The splitting in training and validation sets is done in such a way that all the records (experimental values) of one molecule are assigned to the same set. This avoids that one molecule is used to build the model but also to test its predictive power, leading to an overestimation of the predictive ability of the model. However, if the same molecule was provided in different forms, *e.g.* tautomers or mesomeric forms (Figure 8), these are recognized as being different molecules (the values of some descriptors are different) and thus they may be assigned to both training and validation set.

Figure 13 reports the number of molecules for each number of replicates (multiple values). It is clear that most of the molecules have only one experimental value, but there is still a quite large number of molecules with 2 (92), 3 (36) and 4 (23) records. The percentage of molecules with multiple records is 41%.

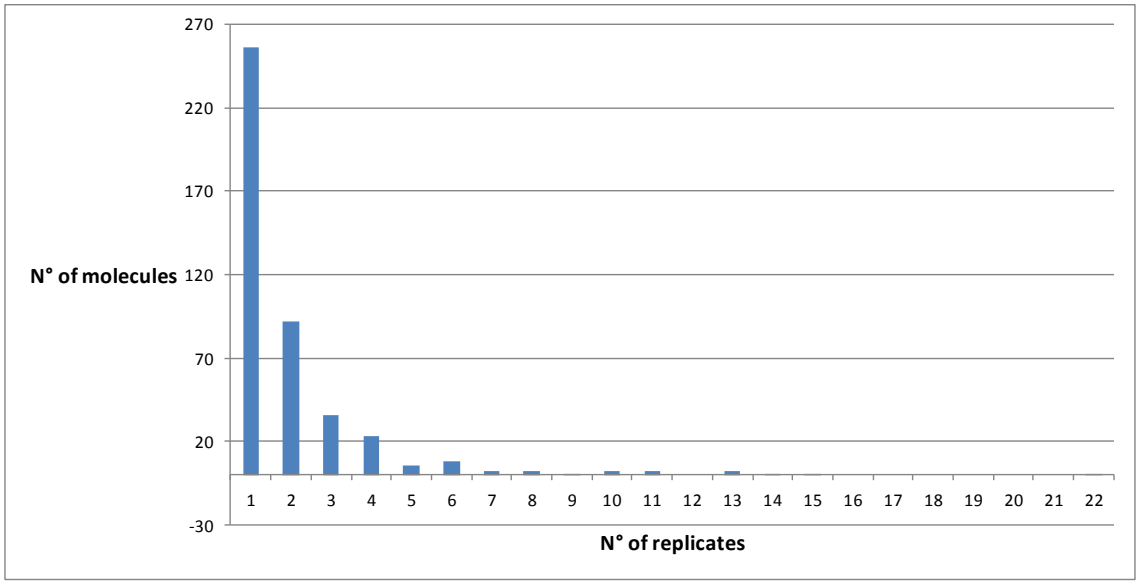


Figure 13: number of molecules for each number of replicates (multiple records).

The best result was obtained with a Support Vector Machine using ChemAxon descriptors (R^2=0.55, Q^2=0.54), followed by an Artificial Neural Network using DRAGON descriptors R^2=0.54, Q^2=0.54). Since these models are worse than the previous ones, only the SVM model is briefly commented below.

SVM with ChemAxon descriptors: the SVM was run using a Radial Basis Function kernel. The number of retained descriptors after the variable reduction step was 83.

3 molecules were ignored during the development of the model because of failure during the optimization of the structure (1,2,3,4,5,6-hexachlorocyclohexane) or the calculation of the descriptors (fullerene and acenaphthene).

The plot of calculated versus measured values for LC50 is reported in Figure 14.

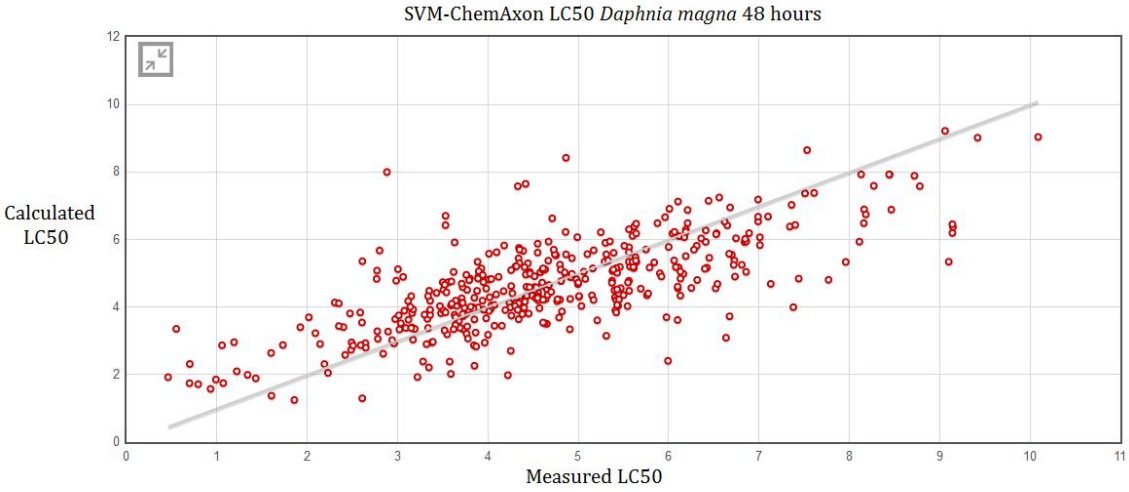


Figure 14: calculated versus measured values for LC50 on Daphnia magna (48 hours) obtained with the SVM model.

The evaluation of the AD in terms of functional groups was undertaken in the same fashion as for the previous model. Figure 15 reports a bubble plot for the RMSEC versus the number of compounds for each functional group. Also for this model, the functional groups that are well represented in the training set have RMSEC values lower or equal to the RMSEC of the model on the entire dataset (1.12). These functional groups include aromatic compounds as well as heterocyclic compounds and halides. Large RMSEC values are associated with poorly represented functional groups. Some of these are the same as those of the ANN on DRAGON descriptors, namely aldehydes and oxohetarenes, while other functional groups outside the AD are different, *i.e.* thiols, tertiary aliphatic amines, phosphoric acid esters and derivatives. It is interesting to note that thiols and arylthiols were very well predicted by the previous model while they are outside the AD of the present model. This fact supports what aforementioned, that is that the good or poor performance of the model on functional groups possessed only by one or two molecules can be due to chance and general conclusions may be misleading. Differently, thiocarboxylic acid amides are well predicted by both models nevertheless only few molecules possess this moiety.

A bar plot of RMSEC versus number of functional groups for each molecule is reported in Figure 16. Also in this case there is no clear correlation between the number of functional groups and the RMSEC value for each molecule. Still, it must be kept in mind that only the number of different functional groups was considered, regardless of the number of instances.

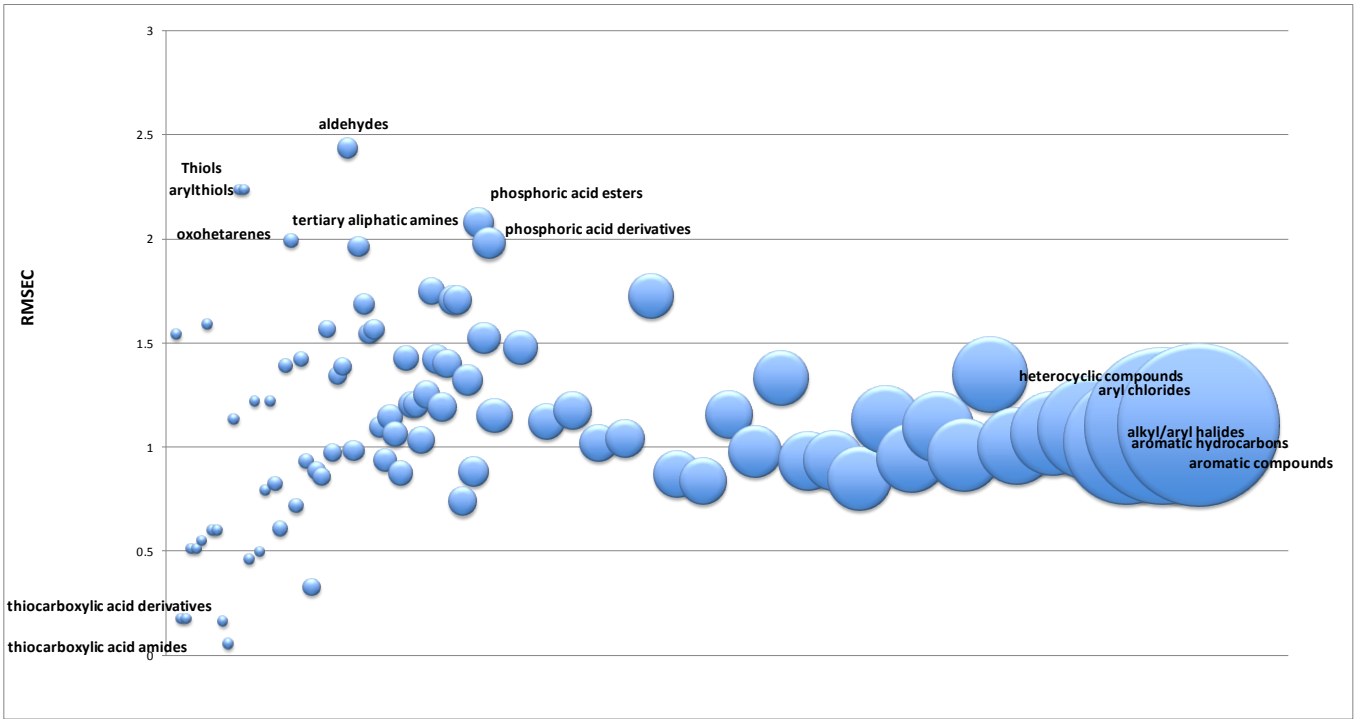


Figure 15: bubble plot of number of compounds for each functional group and RMSEC values. The size of the bubbles is proportional to the number of molecules. The x axis is an enumerator used to sort the bubbles in ascending order of the number of molecules.

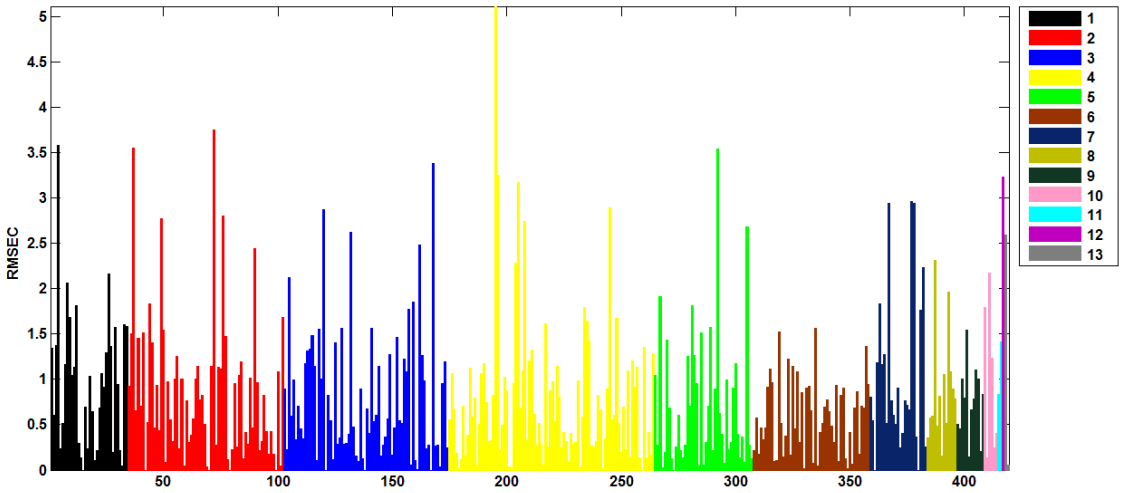


Figure 16: bar plot of RMSEC values for each molecule. Colours indicate the number of functional groups per molecule according to the legend.

The statistics show that the predictive power of the developed models is not very high. These models were compared with those implemented in the software T.E.S.T.[40] of the U.S.-EPA. For the development of the models, the authors used data taken from the ECOTOX database; the database was filtered in order to retain only measurements taken under similar experimental conditions and the median value for each substance was used. The final dataset comprised 337 molecules. Different methods were used to develop the models, which were validated using an external set of 68 compounds. The

performance on the external set for the various approaches are reported in Table 12 along with the statistics of the two best models developed with OCHEM..

Table 12: statistics of the models implemented in T.E.S.T. and the two best developed models using all records and only one record.

Model	Q^2ext	RMSEP
Single model	0.51	1.09
Consensus	0.56	1.09
FDA	0.57	0.94
Nearest neighbour	0.57	1.10
Hierarchical	0.66	0.90
ANN DRAGON	0.64	1.06
SVM ChemAxon	0.54	1.12

It can be seen that also the predictive power of the models implemented in T.E.S.T. is poor. This leads to the conclusion that the LC50 on *Daphnia magna* is a difficult endpoint to model. The best model developed on OCHEM using all the available records seem to be better than the T.E.S.T. models, with the exclusion of the hierarchical model. The SVM with ChemAxon descriptors obtained using only one record for each molecule has slightly lower performance of almost all the T.E.S.T. models, but the single model. A further filtering of the experimental data, combined with the use of some average value or a non-random selection of the measurement to retain and the development of consensus models may lead to a further improvement of the present models using only one record.

4.3 Preparation of data on fish

The dataset for LC50 on fish was prepared in the same fashion as that for *Daphnia magna*. Only one species, namely *Pimephales promelas* (Fathead minnow), and one test duration (96 hours) were considered .

After removing 99 records for which no structure was available, the initial dataset consisted in 3359 records comprising 1048 unique compounds. Thus, also for this property some molecules have more than one experimental value associated. This dataset contains also 919 records previously uploaded on OCHEM by other users.

The same filtering procedure based on structural alerts was applied. The distribution of the initial dataset over the 5 classes is reported in Table 13. Still, only compounds including atoms of organic chemistry were retained. As in the case of *Daphnia magna*, some other compounds and mixtures were also removed, namely NH_3, $HClO_2$, NH_4^+ Cl^-, HS, N_2, $[NH_4^+]_4$ $[HPO_4^{2-}]$ $[SO_4^{2-}]$ and $[NH_4^+]2$ $[SO_4^{2-}]$. The filtered dataset consists in 2592 records for 949 unique chemical compounds.

Table 13: composition of the dataset on Pimephales promelas.

Organic chemistry molecules[1]	Non-organic chemistry molecules	Perfluoroalkylated compounds (PFCs)	Metals	Molecules not supported in ALOGPS program
958	90	3	86	91

1 Molecules including H, C, N, O, S, P, Si, F, Cl, Br and I atoms.

Due to lack of time, a preliminary model was developed in order to highlight the presence of outliers possibly due to errors in the databases. This step allowed to identify some errors and few outliers. The detected errors were of different type, namely:

- the molecule was not present in the original publication;
- the experimental values in the original publication referred to mixtures of compounds but were input in the database as referred to a single chemical;
- the values in the original publications were expressed as intervals (< [i] >) or as higher than (> [i]) but were uploaded as single value;
- the values in the database were different from those reported in the original publication;
- one value (or few values) was highly deviating from the average of the other values for the same molecule. Different test conditions, not reported in the database, may be the cause for these differences;
- 2 molecules with values highly deviating from the trend of all other compounds were considered as outliers and removed.

Table 14 summarizes the number of removed records for the aforementioned reasons.

The refined dataset used for modelling contains 2537 records for 927 unique compounds.

Table 14: number of removed records.

Not in reference	Mixture	Intervals	Different values	Records highly deviating	Outliers (molecules highly deviating)
9	5	18	5	3	16

4.4 QSAR analysis of toxicity on fish

The online platform OCHEM was used to carry out the QSAR analysis. The same approach used to model short-term toxicity on *Daphnia magna* was applied.

Initially all the records were considered for the modelling stage and the response was transformed in logarithmic scale of molarity (Log(mol/L)). The descriptors and methods used to model the LC50 on *Daphnia magna* were also used for the QSAR analysis of short-term toxicity on *Pimephales promelas*.

A variable reduction step was carried out before the development of the models. This procedure was based on absolute values, variance, unique values and pairwise correlation as previously described. The derived models were validated using a 5-fold cross-validation.

Table 15, 16 and 17 report the R^2, Q^2 and RMSECV values, respectively, for the complete list of models obtained on the OCHEM platform using all records. Afterwards, the best model is analyzed more in details.

These models indicate that machine-learning methods (ANN, ASNN and SVM) provide the best results; from the descriptors side, the best results are obtained with DRAGON descriptors, followed by CDK and ChemAxon descriptors. Since the aim is to develop predictive QSAR models, the model with highest Q^2 was chosen as best.

Table 15: R^2 of the models developed on OCHEM using all records.

	ANN	**ASNN**	***K*-NN**	**SVM**	**FSMLR**	**MLRA**	**PLS**
CDK	0.73	0.72	0.66	0.74	0.20	0.56	0.62
DRAGON	0.74	0.74	0.61	0.77	0.58	0.57	0.56
ALogPS,OEstate	0.71	0.71	0.52	0.68	0.66	0.67	0.68
ISIDA	0.67	0.67	0.50	0.63	0.61	0.53	0.63
Mera, Mersy	0.70	0.71	0.54	0.72	0.36	0.52	0.44
ChemAxon	0.73	0.75	0.54	0.69	0.59	0.57	0.60
Inductive	0.66	0.68	0.49	0.63	0.26	0.50	0.57
Adriana	0.66	0.67	0.55	0.56	0.58	0.57	0.48
Spectrophores	0.51	0.49	0.33	0.49	0.30	0.35	0.35
ShapeSignatures	0.62	0.61	0.40	0.59	0.02	0.42	0.49
QNPR	0.66	0.67	0.57	0.61	0.56	0.52	0.60

Table 16: Q^2 of the models developed on OCHEM using all records.

	ANN	**ASNN**	***K*-NN**	**SVM**	**FSMLR**	**MLRA**	**PLS**
CDK	0.73	0.72	0.63	0.73	-1.05	0.56	0.61
DRAGON	0.74	0.74	0.58	0.76	0.53	0.56	0.51
ALogPS,OEstate	0.70	0.70	0.49	0.66	0.65	0.66	0.67
ISIDA	0.65	0.65	0.46	0.61	0.60	0.53	0.62
Mera, Mersy	0.70	0.70	0.48	0.72	0.15	0.49	0.43
ChemAxon	0.73	0.74	0.48	0.68	0.47	0.57	0.59
Inductive	0.66	0.68	0.40	0.62	-0.97	0.49	0.57
Adriana	0.66	0.67	0.53	0.53	0.57	0.57	0.46
Spectrophores	0.51	0.46	0.18	0.47	0.24	0.34	0.34
ShapeSignatures	0.62	0.60	0.28	0.58	-6.83	0.40	0.49
QNPR	0.66	0.67	0.55	0.59	0.56	0.51	0.59

Table 17: RMSEC of the models developed on OCHEM using all records.

	ANN	ASNN	*K*-NN	SVM	FSMLR	MLRA	PLS
CDK	0.77	0.79	0.90	0.76	2.12	0.99	0.93
DRAGON	0.76	0.76	0.97	0.72	1.03	0.99	1.05
ALogPS,OEstate	0.81	0.81	1.07	0.87	0.89	0.87	0.86
ISIDA	0.88	0.88	1.10	0.93	0.94	1.03	0.92
Mera, Mersy	0.81	0.81	1.07	0.78	1.37	1.06	1.13
ChemAxon	0.78	0.75	1.07	0.84	1.09	0.98	0.95
Inductive	0.87	0.85	1.16	0.92	2.10	1.07	0.99
Adriana	0.87	0.86	1.02	1.02	0.98	0.98	1.09
Spectrophores	1.05	1.10	1.35	1.09	1.30	1.21	1.21
ShapeSignatures	0.93	0.95	1.28	0.97	4.20	1.16	1.07
QNPR	0.87	0.86	1.00	0.96	1.00	1.04	0.96

SVM with DRAGON descriptors: the SVM was run using a Radial Basis Function kernel. The number of retained descriptors after the variable reduction step was 1505.

1 molecule (1,2,3,4,5,6-hexachlorocyclohexane) comprising 4 records was ignored during the development of the model because of failure during the optimization of the structure. Thus, 2533 records, comprising 926 unique compounds was used to develop the model.

The plot of calculated versus measured values is reported in Figure 17.

The statistics for the model are provided in Tables 15, 16 and 17. Considering the variability in the experimental conditions – for example, no filter was applied on the type of water (fresh/sea water), test location (laboratory or field), life stage of test organism – it can be said that the model has satisfactory performance. In fact, from the plot of calculated versus measured values, it is clear that for most of the molecules the error in prediction is small. However the model cannot be said to predict very well all the training molecules: for few of them the error in prediction is quite large. The applicability domain of the model was assessed in terms of functional groups in the same way as for the models on *Daphnia magna*.

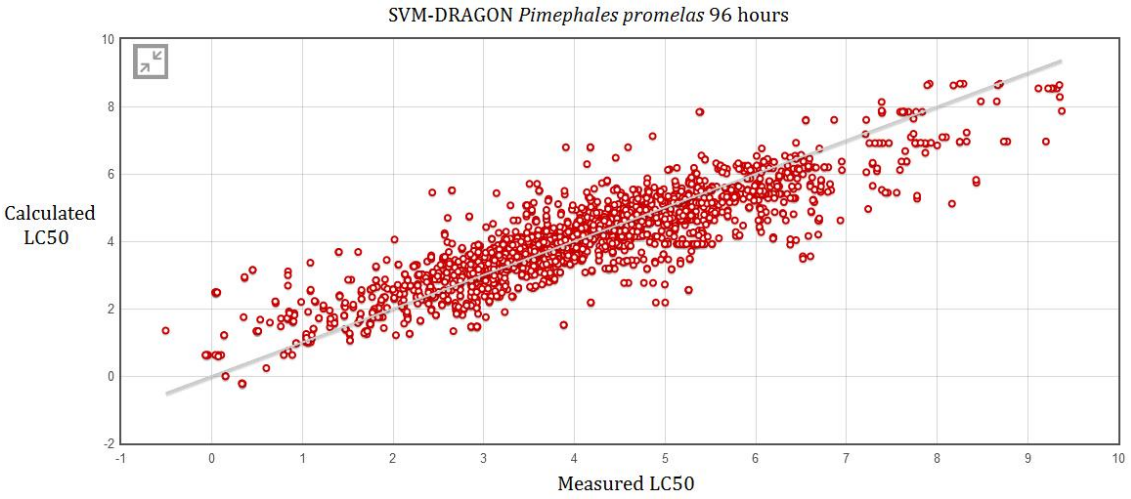


Figure 17: calculated versus measured values for LC50 on Pimephales promelas (96 hours) obtained with the SVM model.

The dataset was screened against the library of functional groups implemented on OCHEM ToxAlerts [Shusko Y. JCIM, in press] and published by Haider[36]. 100 functional groups were identified in the training set.

First, the number of molecules featuring each functional group was calculated. This number was correlated with the RMSEC for each functional group. Figure 18 reports a bubble plot to highlight the correlation between number of molecules for each functional group and the RMSEC.

The correlation RMSEC-number of compounds for each functional group is similar to that obtained on *Daphnia magna*, *i.e.* the most common functional groups (right side of the plot) possess RMSEC values very close to the RMSEC on the entire dataset (0.72). This set of functional groups include aromatic compounds, halides, amines, carboxylic acid derivatives and heterocyclic compounds. The RMSEC for heterocyclic compounds (0.88) is slightly higher than that on the entire dataset. 7 functional groups were identified as being outside the applicability domain of the model (upper left corner of the bubble plot), since the RMSEC is larger than 2 times the RMSEC on the entire dataset. These outlier functional groups are aminals, hydrazine derivatives, nitrates, α-amino acids, phosponic acids, isothiocyanates and phosphinoxides. As expected, these functional groups are not well represented in the dataset. Similarly to the models on *Daphnia magna*, some of the functional groups which are better predicted are also not well represented. This is the case of enamines, secondary aromatic amines, ketene acetal derivatives, carboxylic acid amidines and acyl bromides. However, it must be noted that all the molecules belonging to these 5 functional groups, have also moieties that are well represented in the dataset and this can explain the good performance of the model. Table 18 reports the molecules belonging to the enamines, secondary aromatic amines, ketene acetal derivatives, carboxylic acid amidines and acyl bromides functional groups together with the other moieties they possess.

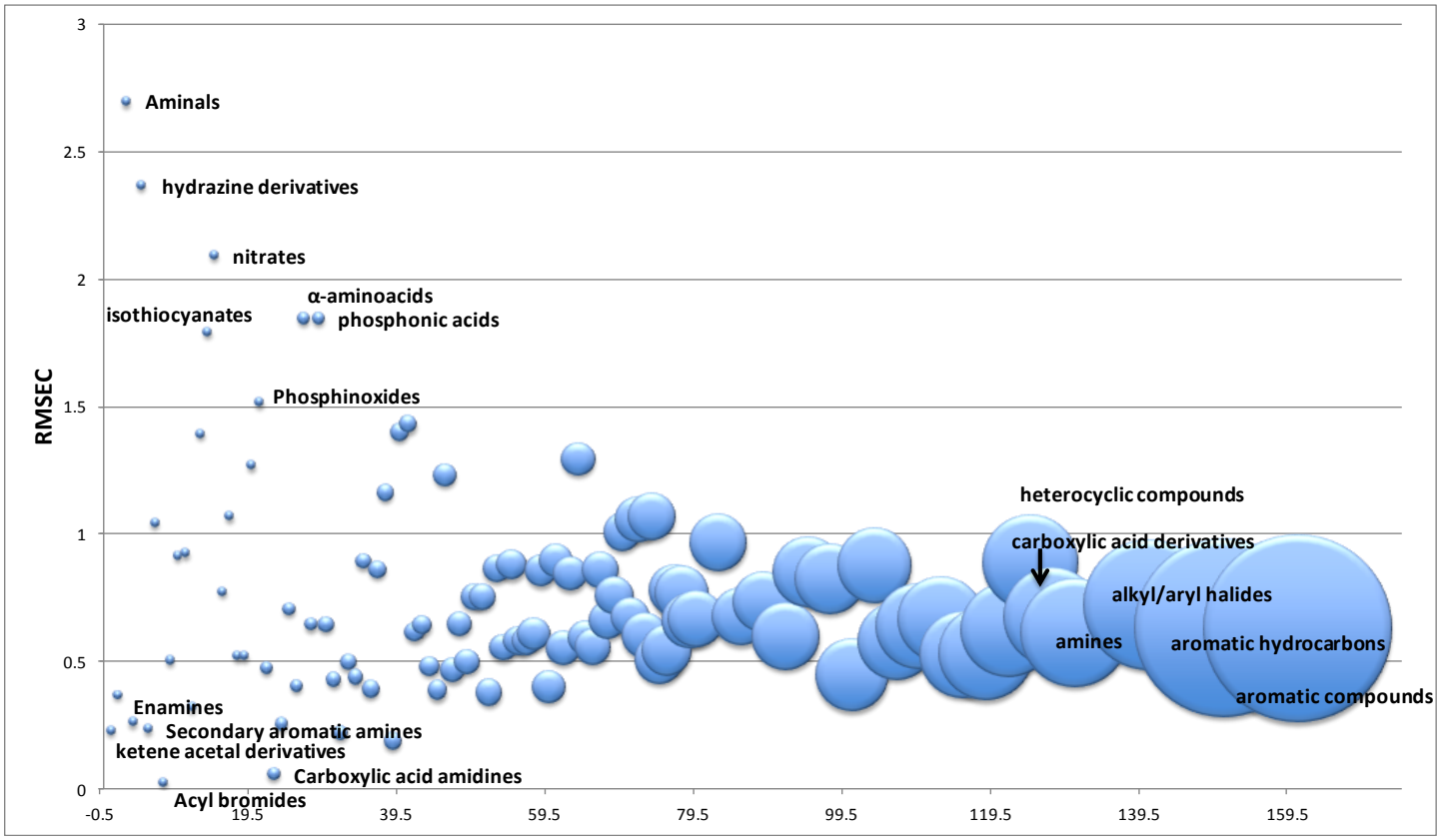
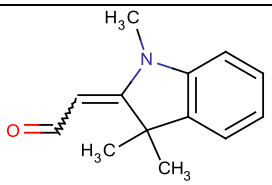
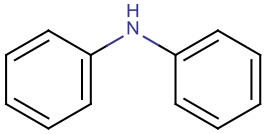
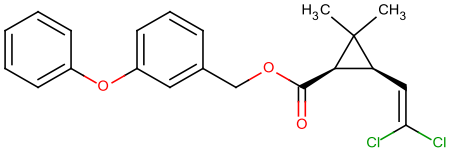
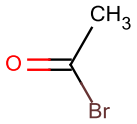
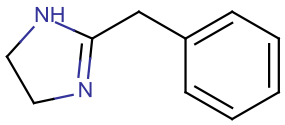


Figure 18: bubble plot of number of compounds for each functional group and RMSEC values. The size of the bubbles is proportional to the number of molecules. The x axis is an enumerator used to sort the bubbles in ascending order of the number of molecules.

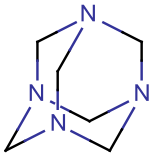
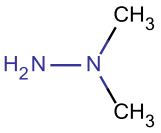
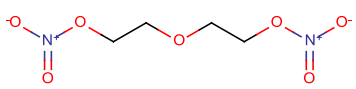
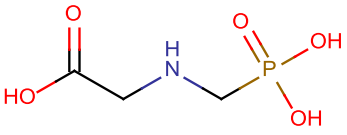
Table 18: name, structure and complete list of functional groups for the molecules belonging to the 5 well predicted but not well represented functional groups.

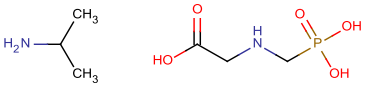
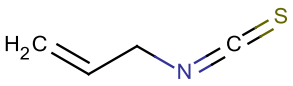
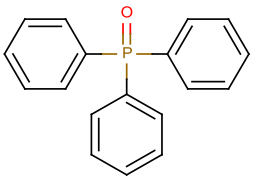
Name	Structure	List of functional groups
2-(1,3,3-trimethyl-2,3-dihydro-1H-indol-2-ylidene)acetaldehyde		aldehydes aromatic compounds aromatic hydrocarbons heterocyclic compounds enamines
N-phenylaniline		aromatic compounds aromatic hydrocarbons amines secondary aromatic amines
Permethrin		aromatic compounds aromatic hydrocarbons ethers diaryl ethers carboxylic acid derivatives carboxylic acid esters ketene acetal derivatives

acetyl bromide		carboxylic acid derivatives acyl halides acyl bromides
2-benzyl-4,5-dihydro-1H-imidazole		aromatic compounds aromatic hydrocarbons heterocyclic compounds carboxylic acid amidines

Some of the molecules belonging to the 7 outlier functional groups also include well represented moieties, such as aromatic/heterocyclic rings and carboxylic acid derivatives, but still the models' performance is very poor. Table 19 reports the structure, name, list of detected functional groups and LC50 values for the 7 molecules belonging to the 7 groups outside the applicability domain. As previously mentioned for the models on *Daphnia magna*, it must be considered that for these functional groups only one or two molecule is present. Therefore, general conclusions may be misleading. The good or poor performance of the model can be also imputed to chance.

Table 19: name, structure, complete list of functional groups and average LC50 values (in –Log(mol/L)) for the molecules belonging to the 7 functional groups outside the applicability domain of the model.

Name	Structure	List of functional groups	LC50
hexamethylene tetramine		heterocyclic compounds amines tertiary aliphatic amines aminals	0.45
1,1-dimethylhydrazine		hydrazine derivatives	3.88
diethylene diglycol dinitrate		ethers cations anions dialkyl ethers nitrates	2.60
glyphosate		carboxylic acid derivatives amines secondary aliphatic amines phosphonic acid derivatives phosphonic acids α-amino acids	4.40

glyphosate isopropylamine		carboxylic acid derivatives amines primary aliphatic amines secondary aliphatic amines phosphonic acid derivatives phosphonic acids α-amino acids	5.00
allyl isothiocyanato		alkenes isothiocyanates	6.06
triphenylphosphine oxide		aromatic compounds aromatic hydrocarbons phosphinoxides	3.71

As for the case of *Daphnia* magna, in order to analyze whether there is a mechanistic reason that can explain the poor performance of the model on the molecules belonging to the 7 outlier functional groups, these molecules were screened against the structural alerts for aquatic toxicity of the Verhaar scheme[37] using the software ToxTree[38]. The results are reported in Table 20.

Table 20: screening of molecules belonging to the outlier functional groups using the Verhaar scheme.

Molecule	Functional group	Classification
hexamethylene tetramine	aminals	not classifiable
1,1-dimethylhydrazine	hydrazine derivatives	not classifiable
diethylene diglycol dinitrate	nitrates	not classifiable
glyphosate	phosphonic acids α-amino acids	not classifiable
glyphosate isopropylamine	phosphonic acids α-amino acids	not classifiable
allyl isothiocyanate	isothiocyanates	unspecific reactivity
triphenylphosphine oxide	phosphinoxides	not classifiable

Unfortunately all the molecules but one were fired as not classifiable according to the rules of the decision tree. Unlike the outlier molecules for the ANN model on *Daphnia magna*, which belonged to particular chemical classes or had specific major uses, the outliers for the model on *Pimephales promelas* belong to different chemical classes and have various uses. In fact these molecules are used as vulcanizing agents (hexamethylene tetramine), propellants and stabilizers for plant growth regulators (1,1-dimethylhydrazine), plasticizers (diethylene diglycol dinitrate), herbicides and pesticides (glyphosate, glyphosate isopropylamine and allyl isothiocyanates) and crystallizing agents (triphenylphosphine oxide). In addition, their LC50 values are close to the mean LC50 over the entire dataset (4.05), with the exception of hexamethylene tetramine, which has a low toxicity, and allyl isothiocyanato, which has a large LC50.

The second analysis that was carried out concerns the single molecules, instead of functional groups. The correlation between the number of functional groups possessed by each compound and the RMSEC for that compound was analyzed. Figure 19 reports a bar plot of the RMSEC of each molecule. Also in this case there is no clear correlation. Molecules with large RMSEC values are present in all the blocks and also the baseline (darker area close to the x axis) seems to follow no general trend. However, the bar plot on the average RMSEC values for each block (Figure 20) shows an interesting feature. The average RMSEC values follow approximately a parabolic trend, with smaller values for blocks 2 to 6. This is reasonable because most of the molecules in the dataset have such number of functional groups. Molecules with just one or many moieties, instead are less frequent and had therefore less weight in the model calibration. However, it must be kept in mind that for this analysis only the number of different functional groups within each molecule was considered, irrespective of the number of instances.

This parabolic trend is even more clear reporting the average RMSEC versus the number of atoms (Figure 21). Molecules with an average size are more represented in the dataset and influenced the development of the model to a major extent. Instead, small and very large molecules are less common and had minor weight in the calibration of the model. It is not surprising thus that the model has lower performance on these molecules which represent more "extreme" situations. An analogous situation had already been found by Mannhold *et al.*[41] when modelling LogP on a dataset comprising more than 96000 compounds.

Again it would have been interesting to study the correlation between the RMSEC of each molecule and the overall number of functional groups, which can be calculated as per the aforementioned formula.

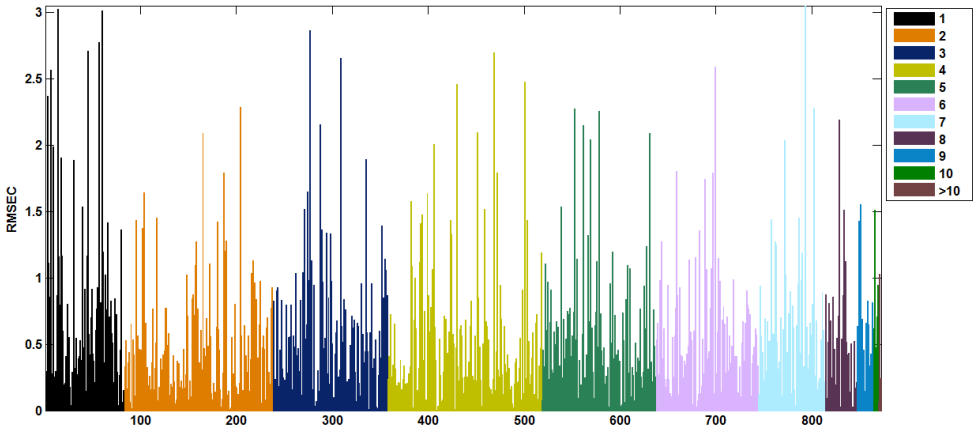


Figure 19: RMSEC for each molecule. Colours indicate the number of functional groups per molecule according to the legend.

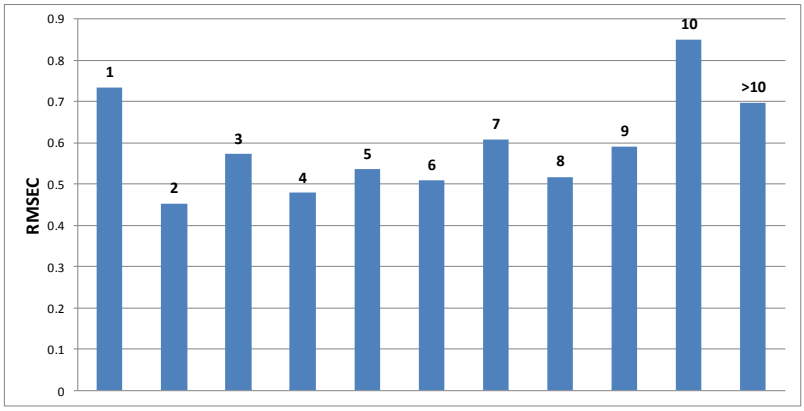


Figure 20: bar plot of the average RMSEC for each block of number of functional groups. The number of functional groups is indicated on top of the bars.

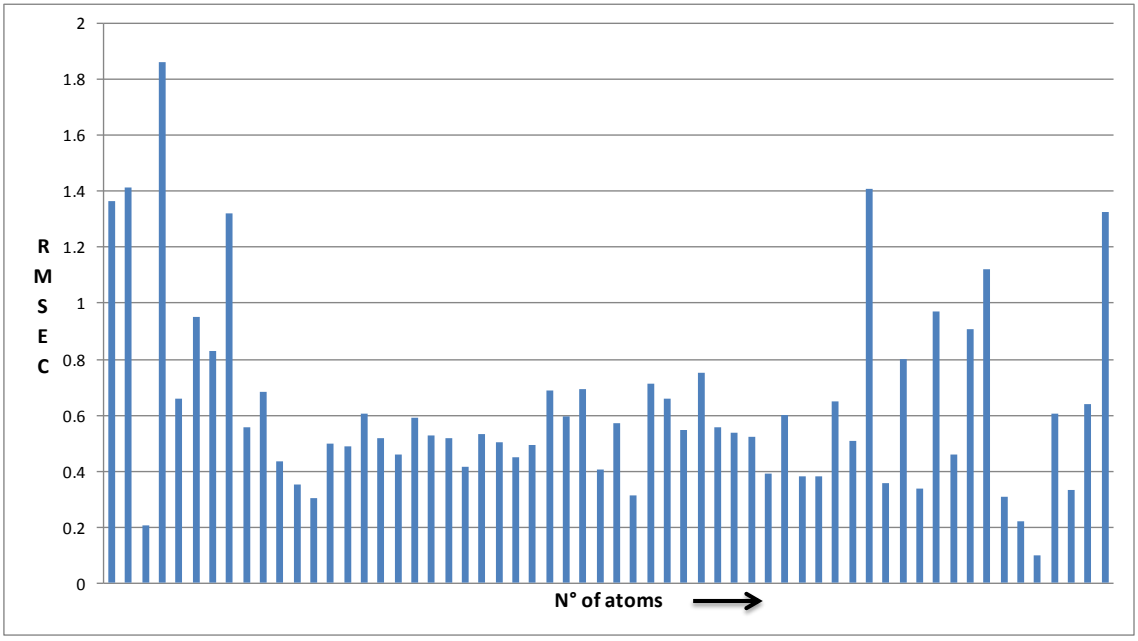


Figure 21: bar plot of the average RMSEC for each block of number of atoms.

As for the model on *Daphnia magna*, also the correlation between the standard deviation of the experimental response within each functional group and the RMSEC for that functional group was studied.

Figure 22 reports the RMSEC values versus the standard deviations for each functional group. The functional groups identified as being outside the applicability domain of the model are highlighted. Also in this case is no clear correlation (R^2 = 0.02) and this result can be explained again considering that molecules featuring a common functional group can be very different in their structure and thus also in their experimental values (large standard deviation), but this does not imply that the model should have poor performance. Unlike the case of *Daphnia magna*, where the outlier functional groups were scattered along the x axis, in this case the outlier moieties have moderate or zero (only one compound features that moiety) standard deviations .

In the same way, also the correlation between RMSEC and standard deviation of the experimental response for each molecule was studied (Figure 23). A picture similar to that found for the model on *Daphnia magna* was obtained, with many molecules having standard deviation equal to 0 (only one experimental value) and some compounds lying along a line (red line) indicating a certain degree of correlation. The same comments made for the model on *Daphnia magna* apply here too.

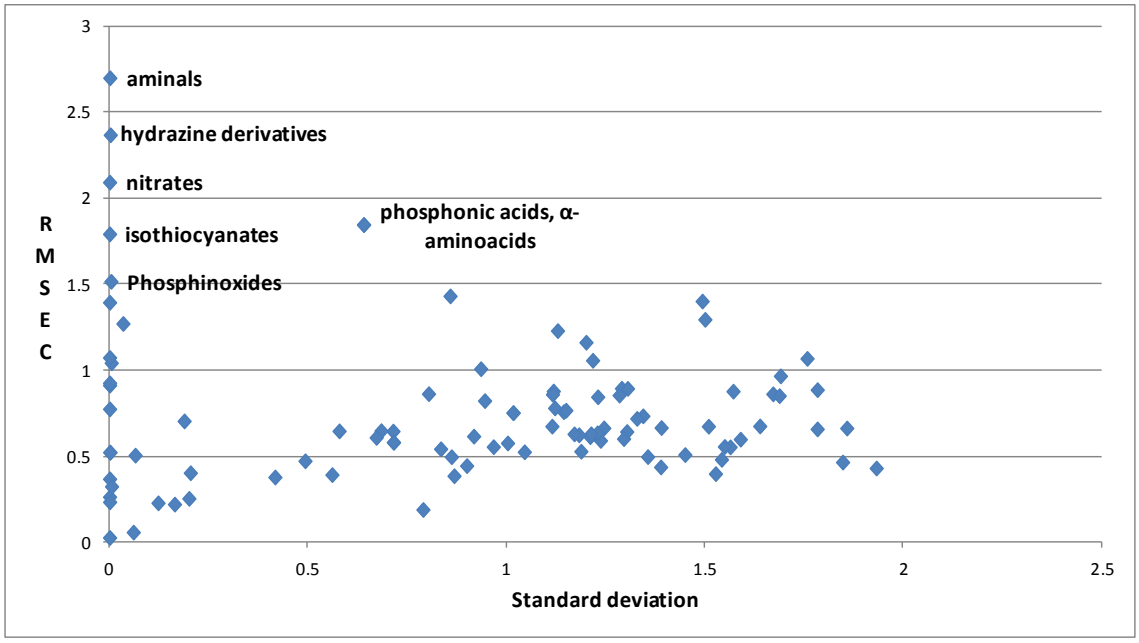


Figure 22: RMSEC values versus the standard deviation for each functional group.

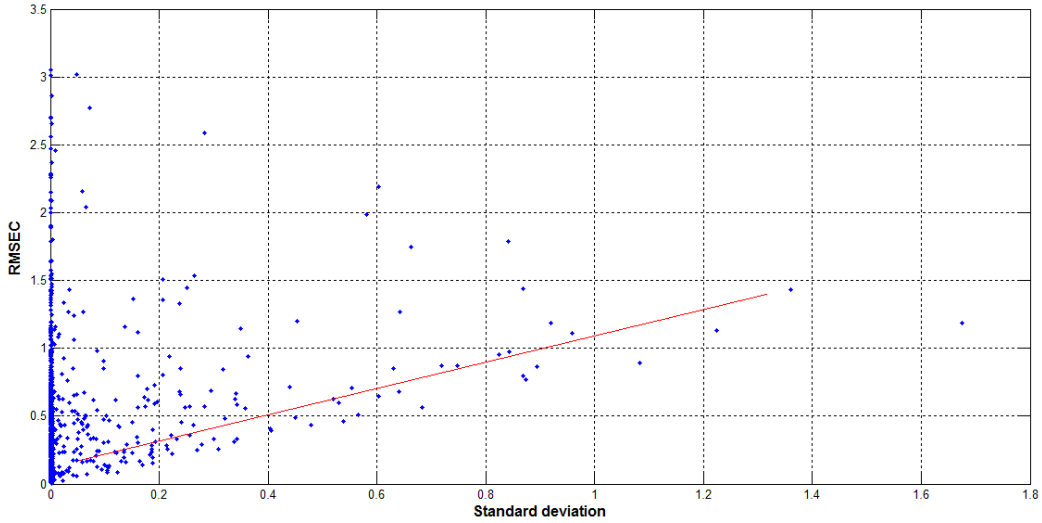


Figure 23: RMSEC versus standard deviation for each molecule.

Additionally another set of models was developed using only one experimental value for each molecule. Since there was no time to analyze all the experimental values and choose the most reliable, the retained measurement was randomly chosen using OCHEM.

The response was transformed in logarithmic scale of molarity (Log(mol/L)) and the same methods and descriptors aforementioned were used. The statistics of the developed models are presented in Tables 21, 22 and 23.

Table 21: R^2 of the models developed on OCHEM using only one record per molecule.

	ANN	ASNN	*K*-NN	SVM	FSMLR	MLRA	PLS
CDK	0.63	0.65	0.55	0.61	0.55	0.58	0.57
DRAGON	0.66	0.65	0.50	0.57	0.37	0.56	0.52
ALogPS,OEstate	0.68	0.67	0.50	0.61	0.55	0.50	0.57
ISIDA	0.57	0.57	0.40	0.63	0.41	0.41	0.57
Mera, Mersy	0.60	0.60	0.51	0.60	0.55	0.54	0.20
ChemAxon	0.60	0.68	0.52	0.65	0.55	0.58	0.57
Inductive	0.58	0.60	0.50	0.54	0.41	0.40	0.32
Adriana	0.62	0.66	0.52	0.61	0.50	0.60	0.48
Spectrophores	0.43	0.47	0.38	0.38	0.31	0.32	0.31
ShapeSignatures	0.50	0.51	0.42	0.43	0.39	0.35	0.33
QNPR	0.23	0.28	0.46	0.58	0.46	0.49	0.50

Table 22: Q^2 of the models developed on OCHEM using only one record per molecule.

	ANN	ASNN	*K*-NN	SVM	FSMLR	MLRA	PLS
CDK	0.63	0.64	0.55	0.61	0.53	0.58	0.57
DRAGON	0.66	0.65	0.50	0.54	0.16	0.56	0.47
ALogPS,OEstate	0.67	0.67	0.48	0.59	0.55	0.50	0.56
ISIDA	0.57	0.57	0.37	0.63	0.36	0.41	0.56
Mera, Mersy	0.60	0.59	0.50	0.59	0.55	0.53	0.19
ChemAxon	0.60	0.68	0.51	0.64	0.54	0.58	0.57
Inductive	0.58	0.60	0.50	0.54	0.40	0.37	0.25
Adriana	0.62	0.66	0.52	0.61	0.48	0.60	0.47
Spectrophores	0.43	0.46	0.38	0.38	0.31	0.32	0.31
ShapeSignatures	0.50	0.50	0.42	0.43	0.39	0.34	0.33
QNPR	-0.11	-0.07	0.44	0.58	0.46	0.49	0.50

Table 23: RMSEC of the models developed on OCHEM using only one record per molecule.

	ANN	ASNN	*K*-NN	SVM	FSMLR	MLRA	PLS
CDK	0.86	0.84	0.94	0.88	0.96	0.91	0.92
DRAGON	0.83	0.83	1.00	0.96	1.30	0.94	1.03
ALogPS,OEstate	0.81	0.81	1.02	0.90	0.95	1.00	0.93
ISIDA	0.93	0.93	1.12	0.86	1.13	1.09	0.93
Mera, Mersy	0.89	0.90	1.00	0.90	0.94	0.96	1.26
ChemAxon	0.89	0.80	0.98	0.84	0.95	0.92	0.93
Inductive	0.92	0.89	1.00	0.96	1.10	1.12	1.22
Adriana	0.87	0.82	0.98	0.88	1.01	0.89	1.02
Spectrophores	1.06	1.04	1.11	1.12	1.17	1.17	1.17
ShapeSignatures	0.99	0.99	1.07	1.06	1.10	1.14	1.15
QNPR	1.49	1.46	1.05	0.92	0.46	1.01	0.99

Also for the QSAR analysis of LC50 on *Pimephales promelas* the results using only one record for each molecule are worse. The same hypotheses raised for the QSAR analysis on *Daphnia magna* apply here too. The number of molecules for each number of replicates (records) is presented in Figure 24.

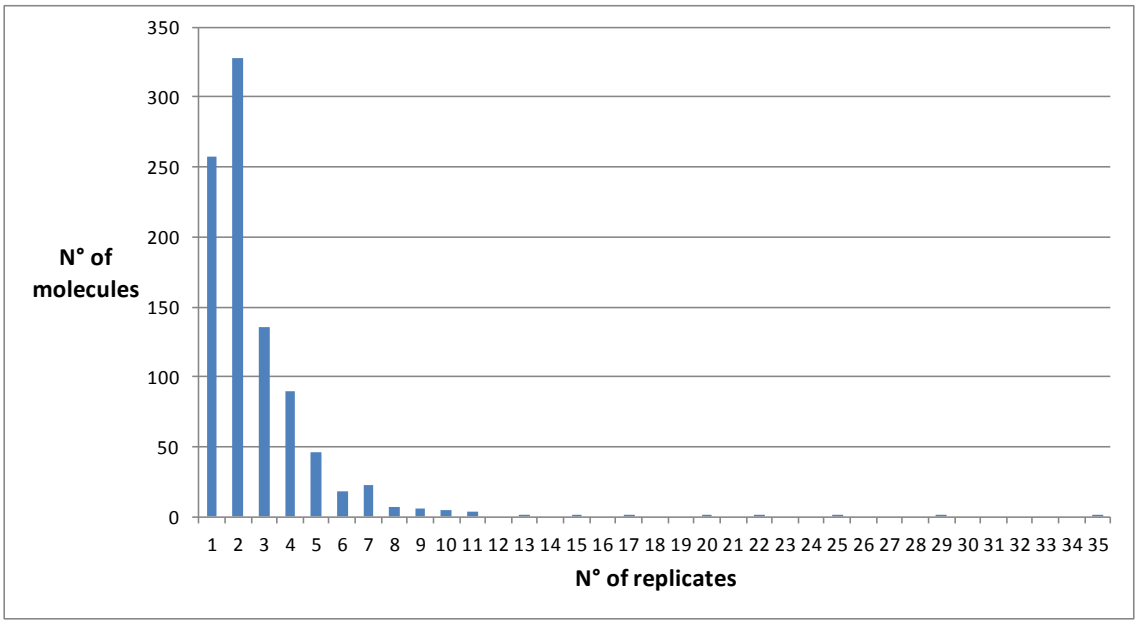


Figure 24: number of molecules for each number of replicates (records).

For this dataset most of the molecules have multiple records. The percentage of molecules with only one experimental measurement is 28%.

The best result was provided by an associative neural network using ChemAxon descriptors (R^2=0.68, Q^2=0.68, RMSEC=0.80). However, ten compounds were ignored during the development of the model because of errors during the optimization of the structure or the calculation of the descriptors. The ANN model with ALogPS and OEstate descriptors has a very similar performance (R^2=0.68, Q^2=0.67, RMSEC=0.81) but only one molecule was ignored during the development of the model because of failure in the calculation of the descriptors. This model is briefly commented below.

ANN with ALogPS and OEstate descriptors: the neural network was trained with 1000 iterations, using the SuperSAB training method and 3 neurons in the hidden layer. The number of retained descriptors after the variable reduction step used to train the network was 142.

The plot of calculated versus measured values for LC50 is reported in Figure 25.

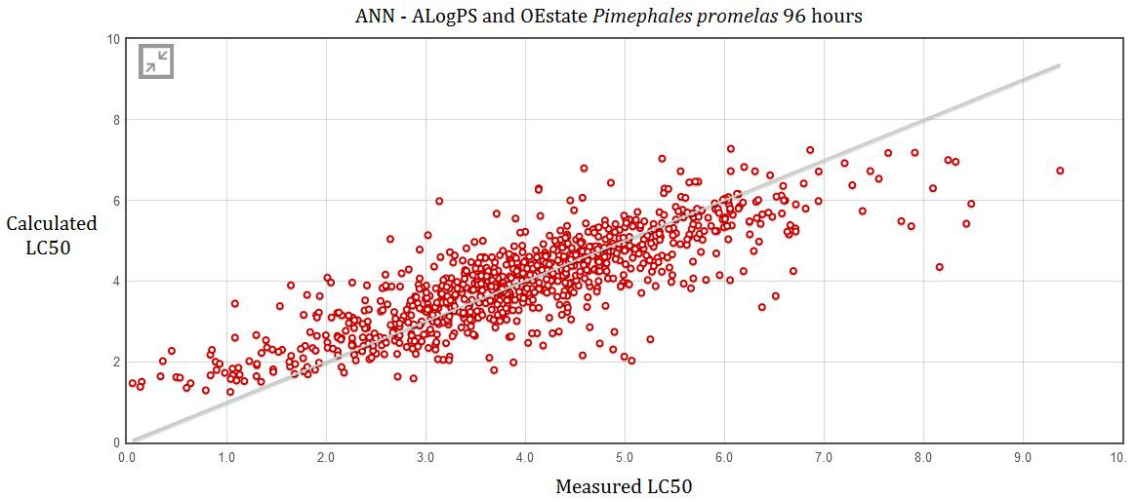


Figure 25: plot of calculated versus measured values for LC50 (96 hours) on Pimephales promelas.

The evaluation of the AD in terms of functional groups was undertaken in the same fashion as for the previous models. Figure 26 reports a bubble plot for the RMSEC versus the number of compounds for each functional group. Again it is possible to notice that the well represented functional groups are associated to an RMSEC lower than that on the entire dataset (0.81). These functional groups are the same observed in the previous models, namely aromatic compounds, alky/aryl halides and carboxylic acid derivatives. In addition, also amines are well represented and predicted. 9 functional groups can be said to be outside the AD of the model since their RMSEC values are higher than 2 times the RMSEC on the entire dataset. Some of these functional groups were detected also in the previous models as outliers (phosphonic acids and their derivatives, aminals, hydrazine derivatives, phosphinoxides, α-amino acids and isothiocyanates). For this model, also isocyanates and ketene acetal derivatives can be considered as outliers. As for the previous models, all of these functional groups are poorly represented in the dataset. Other not well represented functional groups are instead well predicted by the model. Some of the molecules featuring these groups present also well represented moieties. Again, the good or poor performance of the model on such rare functional groups can be due to chance and not to a general behavior of the model.

A bar plot of RMSEC versus number of functional groups for each molecule is reported in Figure 27. Also in this case there is no clear correlation between the number of functional groups and the RMSEC value for each molecule. Still, it must be kept in mind that only the number of different functional groups was considered, regardless of the number of instances.

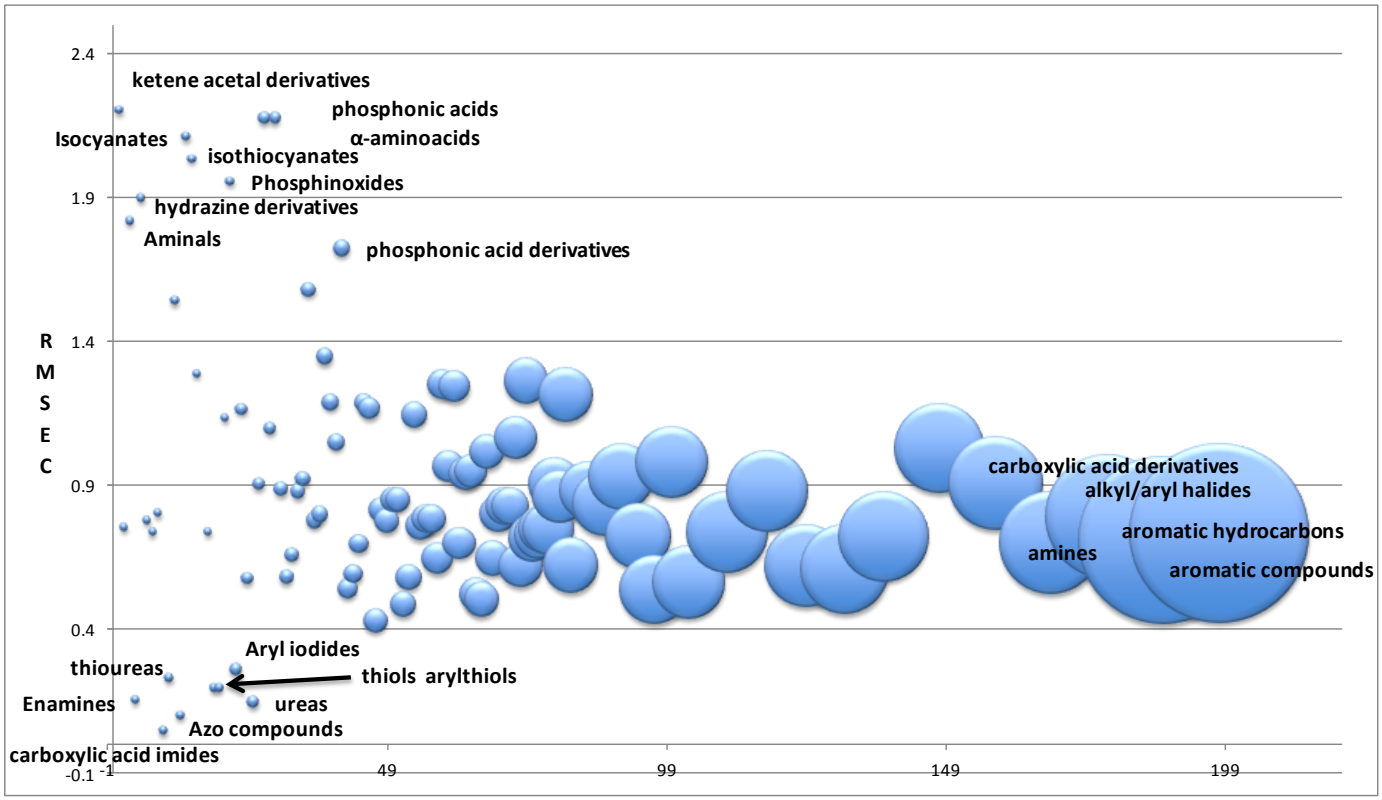


Figure 26: bubble plot of number of compounds for each functional group and RMSEC values. The size of the bubbles is proportional to the number of molecules. The x axis is an enumerator used to sort the bubbles in ascending order of the number of molecules.

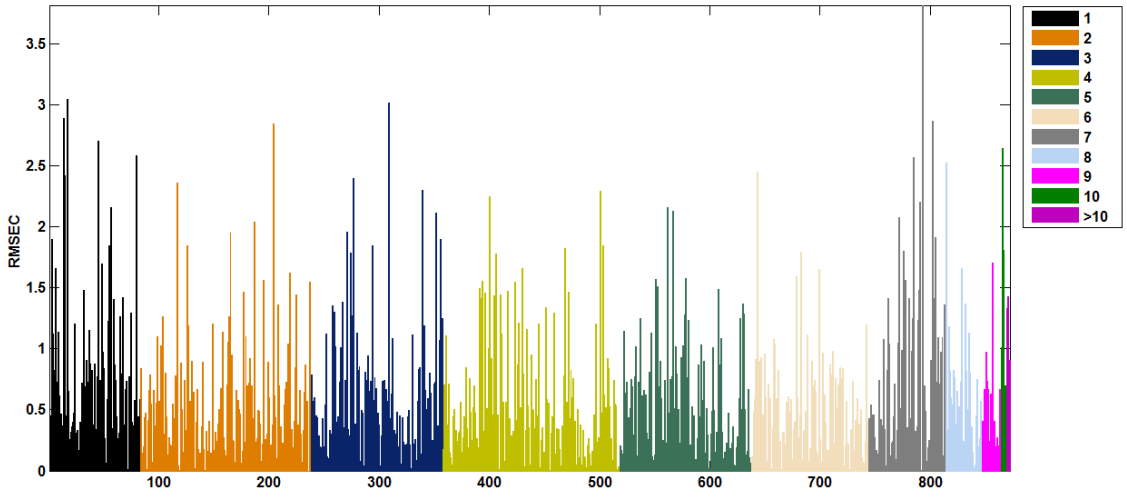


Figure 27: RMSEC for each molecule. Colours indicate the number of functional groups per molecule according to the legend.

The developed models have satisfactory statistics, especially considering the variability in the experimental conditions. Again, the developed models were compared with those implemented in the software T.E.S.T.[40]. The procedure applied by the U.S.-EPA for the definition of the dataset on *Pimephales promelas* was exactly the same as that used for *Daphnia magna*. The final dataset comprised 816 molecules. Several approaches were implemented and the different methods were validated using an

external set of 164 compounds. The performance on the external set for the various approaches are reported in Table 24 along with the statistics of the two best models developed with OCHEM.

Table 24: statistics of the models implemented in T.E.S.T. and of the two best models using all records and only one record.

Model	**Q^2ext**	**RMSEP**
Single model	0.63	0.89
Consensus	0.72	0.78
FDA	0.67	0.81
Nearest neighbour	0.64	0.86
Hierarchical	0.68	0.83
Group contribution	0.64	0.84
SVM DRAGON	0.76	0.72
ANN ALogPS, OEstate	0.67	0.81

The SVM model on DRAGON descriptors using all available records has better parameters than all the models implemented in T.E.S.T.. The question regarding the influence of using multiple records was not clarified. Assuming that the use of multiple records might lead to an overoptimistic estimation of the model predictive power, it can be seen that also the best model developed using only 1 record for each molecule (ANN with ALogPS and OEstate descriptors) has a Q^2 value higher than all the models of T.E.S.T., with the exception of the consensus and the hierarchical models. It should be kept in mind that the choice of the value to retain was randomly undertaken, while for the development of T.E.S.T. models the median was used. The use of the median can balance the experimental variance and avoid extreme or incorrect values. Therefore, it can be stated that there is still possibility to improve the models developed in this study by means of 3 processes:

1. further filtering of data on the basis of the experimental conditions used to measure the property;
2. not random choice of the value to retain and/or calculation of some average value, such as mean or median;
3. develop consensus models.

Chapter 5

Conclusions

Four databases were identified as source of experimental data on aquatic toxicity. *Ad-hoc* designed workflows of KNIME were used to process the databases in order to prepare them for the upload on OCHEM. These data, together with few experimental measurements already uploaded on OCHEM by other users, were used to derive QSAR models for short-term aquatic toxicity. The selected endpoint was the LC50 with a test duration of 48 hours for *Daphnia magna* and 96 hours for *Pimephales promelas*.

For both QSAR studies the data have been filtered in order to have a robust and consistent dataset. Moreover, few values were corrected or excluded due to errors in the data present in the databases. Several different types of descriptors and methods were used for the QSAR study. For each endpoint, the best model was chosen on the basis of the predictive power, indicated by the Q^2 statistics. For both QSAR studies 2 sets of models were developed:

1. using all the available records (experimental values);
2. using only one record for each molecule.

The best QSAR model on *Daphnia magna* using all records is a neural network based on 1413 descriptors of the software DRAGON. The statistics (R^2=0.64 , Q^2=0.64 and RMSEC=1.06) are not very good, but are better than those of the models implemented in software T.E.S.T., with the exception of the hierarchical model. The evaluation of the applicability domain of the model highlighted 7 functional groups that can be considered outside the AD. As expected, these moieties are not well represented in training set. These moieties are aldehydes, carboxylic acid secondary amides, oxohetarenes, phosphoric acid amides, phosphonic acid derivatives, phosphonic acid esters and alkynes. The molecules featuring these groups are or aldehydes, known to be reactive species, or substances used as pesticide, herbicide and chemosterilant. Thus, they represent a peculiar class of compounds (designed to be active against some organisms) that may explain their different toxic behaviour. On the contrary, all well represented functional groups (aromatic and heterocyclic compounds as well as halides) have an RMSEC lower or equal to that on the entire dataset. No correlation was found between the RMSEC of each molecule and the number of different functional groups. However, when the average RMSEC is calculated from the molecules that have certain number of functional groups, a slight parabolic trend is observed, indicating that the error of the model is lower on molecules with an average number of moieties. This is due to the fact that most of the molecules used to develop the model have 2 to 6 different functional groups. No correlation was also found between the RMSEC of each functional group and the standard deviation of the experimental responses of the molecules featuring that group. This result was also expected since the molecules that feature a common functional group can be very different in their structure and possess, therefore, very different LC50 values, but this does not necessarily imply that the model should have poor performance. For many molecules a slight correlation was found between the RMSEC and the standard deviation of the experimental response.

The models developed using only one record (randomly chosen) gave worse results. Some hypothesis were proposed regarding the quality of the experimental data, the sampling step and the cross-validation when the same molecule is provided in different forms to explain these results. The best model developed using only one record is a SVM based on 83 ChemAxon descriptors (R^2=0.55, Q^2=0.54 and RMSEC=1.22). The evaluation of the AD showed again that the most common functional groups (aromatic compounds, heterocyclic compounds and halides) have RMSEC values lower or equal to the RMSEC of the model on the entire dataset (1.12). Some of the functional groups that can be considered outside the AD are the same indentified for the model on all records, namely aldehydes and oxohetarenes, while others are different, *i.e.* thiols, tertiary aliphatic amines, phosphoric acid esters and derivatives. Interestingly thiols and arylthiols were very well predicted by the ANN model while they are outside the AD of the present model.

The best QSAR model on *Pimephales promelas* using all the records is a support vector machine with 1505 DRAGON descriptors (R^2=0.77, Q^2=0.76 and RMSEC=0.72). Also for this model the most common functional groups (aromatic compounds, halides, amines, carboxylic acid derivatives and heterocyclic compounds) have an RMSEC lower or equal to that on the entire dataset. 7 not well represented functional groups (aminals, hydrazine derivatives, nitrates, α-amino acids, phosponic acids, isothiocyanates and phosphinoxides) associated with large RMSEC values were identified as being outside the AD of the model. No clear correlation was observed between the RMSEC and the number of functional groups of each molecule because molecules associated with large RMSEC values are present in all the blocks. However, the bar plot on the average RMSEC values for each block shows that the average RMSEC values follow approximately a parabolic trend, with smaller values for blocks 2 to 6 (corresponding to the most common number of functional groups per molecule in the dataset). This trend is even more clear when the correlation between the average RMSEC and the number of atoms is analyzed. This is due to the fact that most of the molecules in the dataset have an average size and influenced to a major extent the development of the model. Again, no correlation was also found between the RMSEC of each functional group and the standard deviation of the experimental responses of the molecules featuring that group. The same comments made for the model on *Daphnia magna* hold here too. The analysis of the correlation between the RMSEC and the standard deviation of the experimental response for each molecule highlighted that for many molecules there is a trend. This result was expected since the performance of the model cannot be very good if multiple experimental values largely disagree (large standard deviation).

As for the case of *Daphnia magna*, also the models developed on *Pimephales promelas* using only one records gave worse results than those obtained using multiple records. The influence of the presence of multiple experimental measurements was not clarified. The best model developed using only one record for molecule (R^2=0.67, Q^2=0.67, RMSEC=0.81) is a ANN based on 142 descriptors (LogPS and OEstate). The most common organic functional groups (aromatic rings, halides and amines) are associated with RMSEC values lower or equal to the RMSEC on the entire dataset. 9 functional groups were identified as being outside the AD of the model. These moieties include phosphonic acids and their derivatives, aminals, hydrazine

derivatives, phosphinoxides, ketene acetal derivatives, α-amino acids, isocyanates and isothiocyanates.

The SVM model with DRAGON descriptors seem to be better than all the models implemented in the software T.E.S.T.. If one assumes that the presence of multiple values lead to an overestimation of the predictive power, it can be seen that, anyway, the ANN model developed using only one record per molecule has also better results than the models of T.E.S.T., with the exception of the consensus model. A further filtering of the experimental data, combined with the use of some average value or a non-random selection of the measurement to retain and the development of consensus models may lead to a further improvement of the present models.

Additional activities

During my fellowship I participated to the following schools:

- ECO Summer School 2012, 11-15 June 2012, Verona (Italy);
- Strasbourg Summer School on Chemoinformatics 2012, 25-29 June 2012, Strasbourg (France).

Acknowledgements

I acknowledge my supervisor, Dr. Igor Tetko, for giving me the possibility to pursue this study, guiding me throughout my fellowship and letting me attend the school on chemoinformatics in Strasbourg. I also thank my colleagues at the Helmholtz Zentrum München, Kamel Mansouri, Ioana Oprisiu, Pantelis Sopasakis, Stefan Brandmaier, Kai Zillessen, Ahmed Abdelaziz, Wolfram Teetz, Yurii Sushko, Sergii Novotarskyi, Robert Körner and Pankaj Yadav for their help and suggestions.
I offer thanks to Prof. Roberto Todeschini and Milano Chemometrics and QSAR Research Group for the organization of the 2nd ECO Summer School in Verona.

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under Grant Agreement n. 238701.

Bibliography

[1] Newsome L. D., Nabholz J. V. and Kim A). Designing Aquatically Safer Chemicals. Book chapter in *Designing Safer Chemicals. Green Chemistry for Pollution Prevention*. DeVito S. C. and Garrett R. L. editors. American Chemical Society, Washington, DC, 1996.

[2] Regulation (EC) No 1907/2006

[3] Guidance on information requirements and chemical safety assessment. Chapter R.6: QSARs and Grouping of Chemicals. ECHA. http://echa.europa.eu/documents/10162/13632/information_requirements_r6_en.pdf

[4] Sushko I., Novotarskyi S., Körner R., Pandey A. K., Rupp M., Teetz W., Brandmaier S., Abdelaziz A., Prokopenko V. V., Tanchuk V. Y., Todeschini R., Varnek A., Marcou G., Ertl P., Potemkin V., Grishina M., Gasteiger J., Schwab C., Baskin I. I., Palyulin V. A., Radchenko E. V., Welsh W. J., Kholodovych V., Chekmarev D., Cherkasov A., Aires-de-Sousa J., Zhang Q. Y., Bender A., Nigsch F., Patiny L., Williams A., Tkachenko V., Tetko I. V. Online chemical modeling environment (OCHEM): web platform for data storage, model development and publishing of chemical information. *J Comput Aided Mol Des*. 2011; 25(6):533-54.

[5] Todeschini R. and Consonni V. *Molecular Descriptors for Chemoinformatics*. Mannhold R., Kubinyi H., Folkers G. editors. WILEY-VCH, 2009.

[6] http://www.wikipedia.org/

[7] http://alttox.org

[8] Rand G. M., Petrocelli S. R. *Fundamentals of aquatic toxicology: Methods and applications*. Washington: Hemisphere Publishing, 1985.

[9] Todeschini R. and Consonni V. *Handbook of molecular descriptors*. Wiley-VCH, 2000.

[10] Frank I.E. and Friedman J.H. A statistical view of some chemometrics regression tools. *Technometrics.* 1993; 35, 109-135.

[11] Todeschini R. *Introduzione alla Chemiometria*. Edises, 1998.

[12] Andersson M. A comparison of nine PLS1 algorithms. *Journal of Chemometrics*, 2009; 23, 518-529.

[13] *Encyclopedia of Computer Science and Engineering*, Wah (editor). Wiley, 2008.

[14] Tetko I. V. Associative Neural Networks. *Neural Processing Letters,* 2002, 16, 187-199.

[15] Tavella M. Regressione con SVM e Backpropagation. 2005: files.mtvl.org/studies/mt_TTR1_SVM_2005.pdf

[16] ECOTOX: http://cfpub.epa.gov/ecotox/

[17] ECETOC: http://www.ecetoc.org/

[18] OASIS: http://www.oasis-lmc.org/

[19] QSAR Toolbox: www.qsartoolbox.org/

[20] Aquatic Japan MoE: http://www.safe.nite.go.jp/english/db.html

[21] KNIME: http://www.knime.org

[22] ChemSpider: http://www.chemspider.com/

[23] PubChem: http://pubchem.ncbi.nlm.nih.gov/

[24] Steinbeck, C.; Han, Y. Q.; Kuhn, S.; Horlacher, O.; Luttmann, E.; Willighagen, E.L. The Chemistry Development Kit (CDK): An open-source Java library for chemo- and bioinformatics. *Journal of Chemical Information and Computer Sciences.* 2003, 43, 493-500.

[25] Talete srl (2012). *DRAGON (Software for Molecular Descriptor Calculation)*, 6.0.

[26] Tetko, I. V.; Gasteiger, J.; Todeschini, R.; Mauri, A.; Livingstone, D.; Ertl, P.; Palyulin, V. A.; Radchenko, E. V.; Zefirov, N. S.; Makarenko, A. S.; Tanchuk, V. Y.; Prokopenko, V. V. Virtual computational chemistry laboratory - design and description, *J. Comput. Aid. Mol. Des*., 2005, 19, 453-63.

[27] Kier L.B. and Hall L.H. An Electrotopological-State Index for atoms in molecules. *Pharm.Res.* 1990; *7*, 801-807.

[28] Solov'ev P., Varnek A., Wipff G. Modelling of Ion Complexation and Extraction of Organic Molecules Using Substructural Molecular Fragments. *Chem. Inf. Comp. Sci.*, 2000, 40 , 847-858.

[29] Potemkin VA, Grishina MA. A new paradigm for pattern recognition of drugs. *J Comput Aided Mol Des*. 2008;22:489–505.

[30] ChemAxon: http://www.chemaxon.com

[31] Cherkasov A. Inductive QSAR Descriptors. Distinguishing Compounds with Antibacterial Activity by Artificial Neural Networks. *Int. J. Mol. Sci*. 2005, 6, 63-86.

[32] ADRIANA.Code: http://www.molecular-networks.com.

[33] OpenBabel: http://openbabel.org/docs/2.3.1/Fingerprints/spectrophore.html

[34] Zauhar R. J., Moyna G., Tian L., Li Z., and Welsh W. J.. Shape signatures: a new approach to computer-aided ligand- and receptor-based drug design. *J. Med. Chem* 2003; 46:5674–5690.

[35] Thormann M., Vidal D., Almstetter M., Pons M. Nomen Est Omen: Quantitative Prediction of Molecular Properties Directly from IUPAC Names. *The Open Applied Informatics Journal.* 2007; 1 (1), 28-32.

[36] Haider, N., Functionality Pattern Matching as an Efficient Complementary Structure/Reaction Search Tool: an Open-Source Approach. *Molecules*; 2010, 15, 5079-5092.

[37] Verhaar H.J.M., Van Leeuven C., Hermens J.L.M.,Classifying Environmental Pollutants. 1: Structure-Activity Relationships for Prediction of Aquatic Toxicity. *Chemosphere*, 1992; 25, 4, 471-491.

[38] Ideaconsult Ltd. Toxtree (Estimation of Toxic Hazard - A Decision Tree Approach), 2.5.1.

[39] Kim Y., Choi K., Jung J., Park S., Kim P. G., Park J. Aquatic toxicity of acetaminophen, carbamazepine, cimetidine, diltiazem and six major sulfonamides, and their potential ecological risks in Korea. *Environ Int.;* 2007, 33(3):370-5.

[40] T.E.S.T. Version 4.0.1. US-EPA.

[41] Mannhold R., Poda G. I., Ostermann C., Tetko I. V. Calculation of Molecular Lipophilicity: State-of-the-Art and Comparison of LogP Methods on More Than 96,000 Compounds. *Journal of Pharmaceutical Sciences*; 2009, 98, 861-893.