Fingerprint-based detection of high acute aquatic toxicity

Pireddu, Luca[1]; Michielan, Lisa[2]; Floris, Matteo[3]; Rodriguez-Tomé, Patricia[1]; Moro, Stefano[2]

Email addresses: pireddu@crs4.it; stefano.moro@unipd.it

1. CRS4 – Molecular Informatics Group, Italy;

2. Molecular Modeling Section (MMS), Dep. of Pharmaceutical Sciences, U. of Padova, Italy;

3. CRS4 – Bioinformatics Laboratory, Italy

Introduction

Recent legislation is paying more attention to the dangers posed by chemicals to human and environmental health.

REACH regulation [1]:

- requires that industry provide information about the toxicity of the chemicals
- encourages reduction of animal testing
- encourages the use of existing data
- encourages alternative assessment approaches, such as QSAR modelling.

Consequently, there is a growing demand for in-silico tools for performing ecological risk assessments. With this work we aim to:

- develop an interpretable model to help determine the level of acute aquatic toxicity manifested by a chemical structure
- make this model available through a web interface
- integrate this tool with the large-scale chemoinformatics database MMsINC. [2]

Dataset

Our study is based on the well-known EPA Fathead Minnow dataset [3]:

- 617 industrial compounds
- 2D chemical structures
- measured 96-h LC_{50} values in mg/L and mmol/L

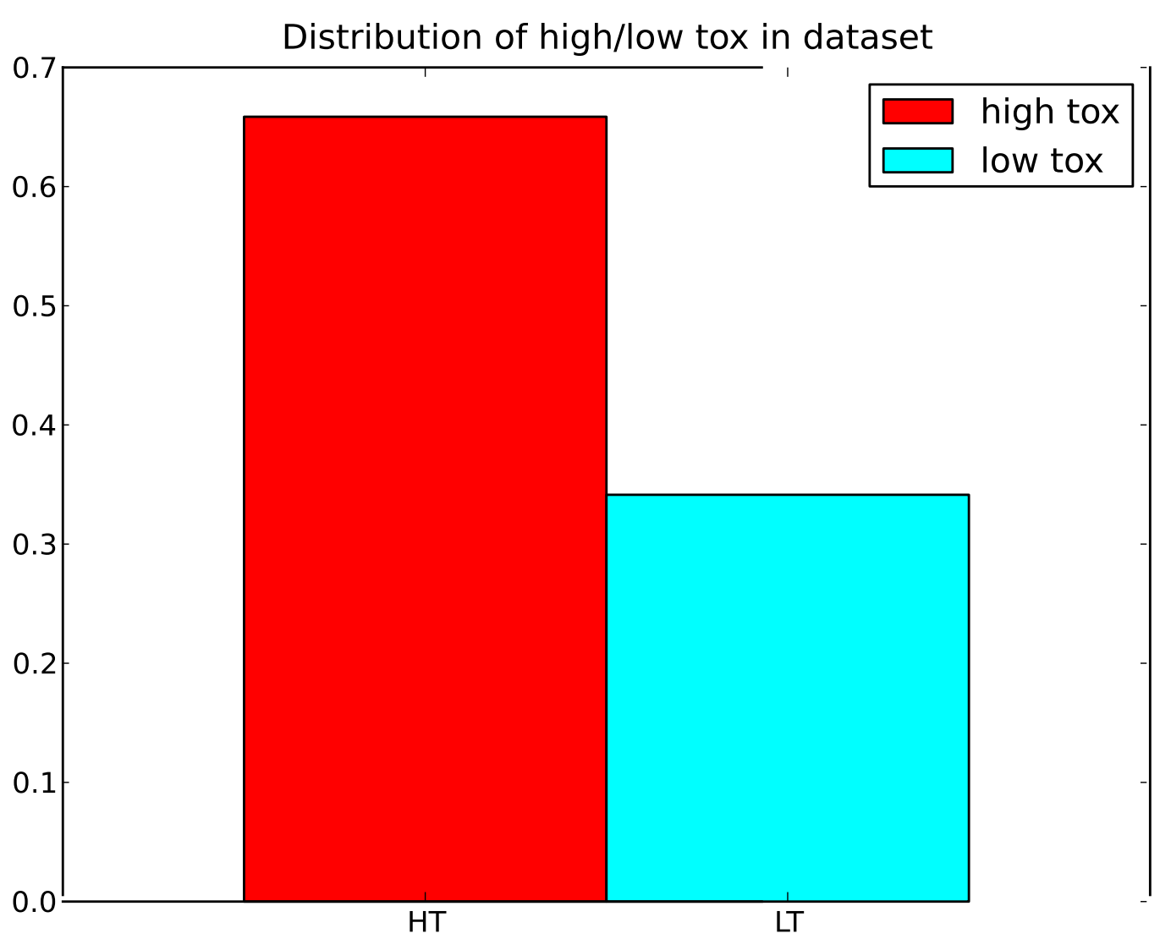
Compounds are classified as active,inactive, or inconclusive.

Activity Class	Description
Active	Fatal to at least 50%
Inconclusive	Fatal to some, but less than 50%
Inactive	Fatal to none.

We excluded the 37 inconclusive or inactive compounds, leaving us with the 580 compounds that compose our dataset.

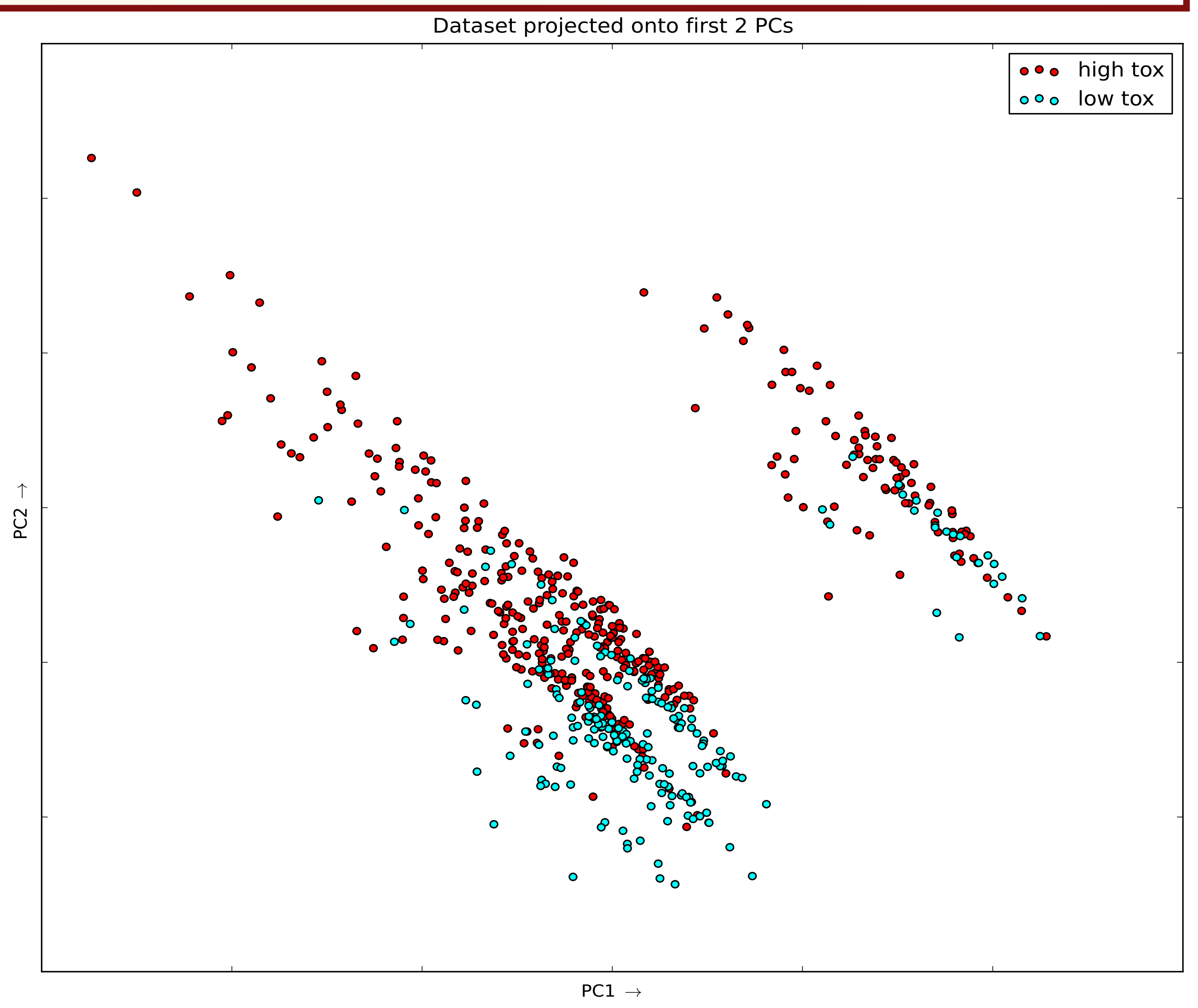
High/Low Acute Toxicity Labelling

Although the idea of classifying compounds by level of acute toxicity was inspired by the OECD Test Guideline 203 [1], the legislation defines acute toxicity as an LC50 <= 100 **mg/L**. We decided to adopt a more reasonable (chemically) molar LC50 threshold of 0.5 mmol/L, which matches the OECD separation OECD for most of the compounds in the dataset.



High/Low toxicity threshold

LC_{50} <= 0.5 mmol/L	high acute toxicity
LC_{50} > 0.5 mmol/L	low acute toxicity



Modelling Method

Molecular representation

We described molecular structures with our in-house implementation of the 881-bit PubChem structural fingerprints [2,4].

Feature selection

We applied a probabilistic filtering feature selection method to eliminate the less important bits from the fingerprints, eliminating all features Xi for which

$$|P(X_i = 1, X_j = 1|Y = v) - P(X_i = 1, X_j = 1)| < pmin$$

holds for all j and all possible values of Y=v

This approach considers the influence of combinations of two variables. In addition, it accommodates some noise by allowing an influence of up to pmin before deciding to keep the feature.

In this work, we empirically chose a pmin value of 2.5%. We apply this filter to the dataset, selecting 217 bits from the original 881.

Classification Model

We built Support Vector Machine [5] (C-SVC) classifiers from the 580-molecule training set, using linear and Radial Basis Function (RBF) kernels. We performed a parameter search as summarized below. We evaluated each combination of parameter values with a 5-fold stratified cross-validation.

We also tested performance with the polynomial, Tanimoto, and exponential Tanimoto kernels. However, they did not show any advantage over the linear RBF kernels, so we refrained from thoroughly evaluating those options.

All SVMs were built using the LIBSVM software package. [6]

Parameter search

SVM cost (C)	values 1 to 1024 by powers of 2
relative weight on each class	from +5 to the high tox class to +5 for the low tox class, in steps of 1
gamma (RBF only)	from 1/1024 to 1, by powers of 2

Validation

After selecting model parameters by estimating classification performance through 5-fold cross validation, we evaluated two models with Leave-One-Out (LOO) cross validation. We measured the following:

TP	No. of high tox recognized
FP	No. low tox incorrectly classified
TN	No. of low tox recognized
FN	No. of high tox incorrectly classified
F-measure	2*Precision*Recall / (Precision+Recall)
Precision	TP / (TP+FP)
Recall	TP / (TP+FN)
Accuracy	(TP+TN) / (TP+FP+TN+FN)
Avg nSV	Avg no. of support vectors in 5-fold CV
StdDev nSV	StdDev in no. of support vectors in 5-fold CV

Results and Discussion

Explaining Predictions

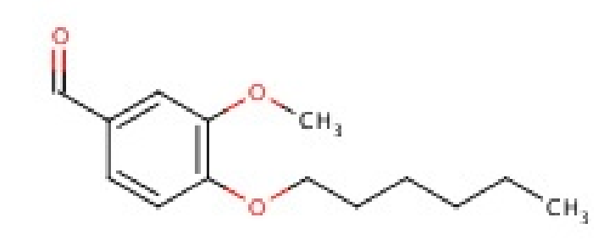
We do not expect QSAR models to replace chemists. Rather, we expect them to be a helpful decision-making tool. To achieve this goal, it is important for a user to understand why the model predicts that a molecule is more toxic or less toxic. To this end, we are implementing the EXPLAIN decision exploration methodology [7] for our linear models.

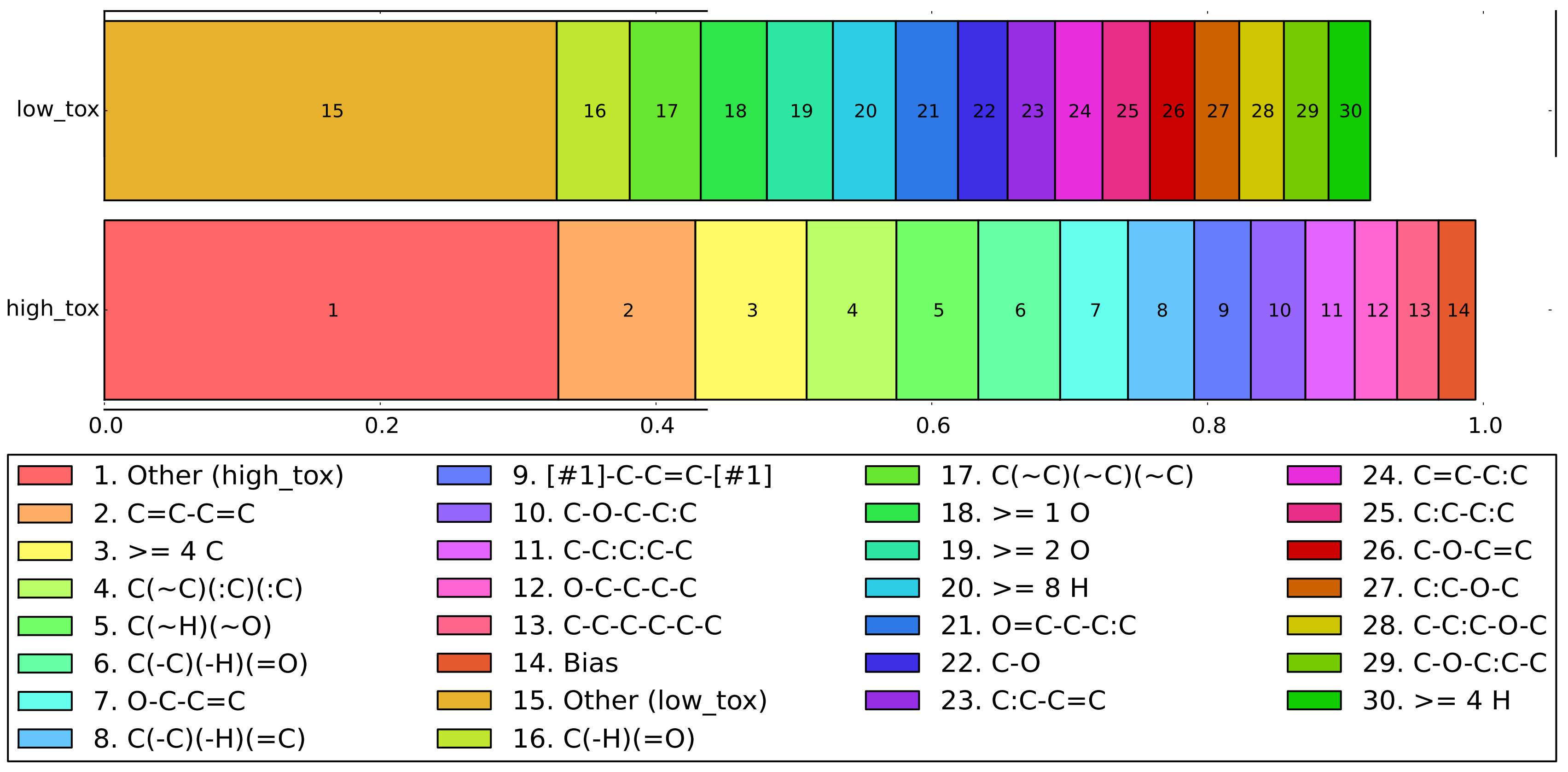
In a linear SVM classifier, like all additive binary classifiers, each feature of a query molecule contributes a weight in favour of either class. The final decision depends on the sum of all weights and the model bias:

- if sum >= 0 then positive class
- if sum < 0 then negative class

The Explain bar below shows the contribution of the most important features in a classification.

C1=CC(C=O)=CC(OC)=C1OCCCCCC





Validation Results

5-fold cross validation

Kernel	TP	FP	TN	FN	F-measure	Precision	Recall	Accuracy	Avg nSV	StdDev nSV
Linear	320	42	156	62	0.86	0.88	0.84	0.82	136.4	3.32
RBF	352	54	144	30	0.89	0.87	0.92	0.86	228.2	5.78

LOO cross validation

Kernel	TP	FP	TN	FN	F-measure	Precision	Recall	Accuracy	Avg nSV	StdDev nSV
Linear	327	42	156	55	0.87	0.89	0.86	0.83	164.93	2.27
RBF	350	55	143	32	0.89	0.86	0.92	0.85	270.29	1.56

Comparison

	TP	FP	TN	FN	F-measure	Precision	Recall	Accuracy
Michielan [9]	323	43	144	44	0.88	0.88	0.88	0.77
Maunz [10]	317	64	102	53	0.84	0.83	0.86	0.78

This simple comparison was performed to see if our model is on par with others that have been published for this dataset.

The model built by Michielan [9] is based on a subset of the structures in our dataset (559 of our 580), and uses the same toxicity criterion. It is based on AutoMEP, Sterimol, and logP molecular descriptors. The results reported are for LOO validation on the 559 molecules.

The model by Maunz [10] is actually a regression, using a fragment-based approach, and is trained on 568 structures from the same EPAFHM dataset. [3] The results were compiled by querying the published web application and transforing the regression value into a label based on our high/low acute toxicity threshold.

Notice that both these models could not provide predictions for all the structures we queried (559 for [9], 536 for [10]).

Future directions

Validation

There remain a few validation steps to be performed in order to ascertain the validity of our SVM model, as suggested by Tropsha et al. [8] Performing these steps is currently a priority for this project.

Domain of applicability

Establishing the domain of applicability of a QSAR model is as essential as the modelling activity itself. [8] However, measuring the distance-to-model is a still a topic of research [11], especially with respect to binary fingerprint-based methods and classification. We are actively working in this domain.

Web application

We are planning to create a web-accessible application to provide to the world explained predictions by this type of model. We have already implemented a prototype that applies a selected prediction profile to a number of molecular structures, returning for each molecule a card with its results.

Integration with MMsINC

We are working on the integration of predictive models such as the one presented in this poster with our MMsINC database [2], as to provide predicted molecular properties and activity to query and examine for each of the 3M compounds, tautomers, and ionic states in the DB.

References

[1] EU. Corrigendum to Regulation (EC) No 1907/2006 of the European Parliament and of the Council of 18 December 2006 concerning the Registration, Evaluation, Authorization and Restriction of Chemicals (REACH). (2007) Off. J. Eur. Union, L136, 50.
[2] Masciocchi J., Frau G., Fanton M., Sturlese M., Floris M., Pireddu L., Palla P., Cedrati F., Rodriguez-Tomé P., Moro S., MMsINC: a large-scale chemoinformatics database, Nucleic Acids Res. 2009 Jan;37(Database issue):D284-90. Epub 2008 Oct 17
[3] Russom C.L., Bradbury S.P. , Broderius S.J., Hammermeister D.E. and Drummond R.A, Predicting modes of action from chemical structure: Acute toxicity in the fathead minnow (Pimephales promelas), Environmental Toxicology and Chemistry 16(5): 948-967
[4] http://pubchem.ncbi.nlm.nih.gov/help.html
[5] Drucker H., Burges C.J.C., Kaufman L., Smola A. and Vapnik V., Support Vector Regression Machines, Advances in Neural Information Processing Systems 9, NIPS 1996, 155-161, MIT Press
[6] Chang C.C. and Lin CJ., LIBSVM: a library for support vector machines, 2001, http://www.csie.ntu.edu.tw/~cjlin/libsvm
[7] Poulin B., Eisner R., Szafron D., Lu P., Greiner R., Wishart D.S., Fyshe A., Pearcy B., MacDonell C. and Anvik J., Visual Explanation of Evidence in Additive Classifiers, 18th Conference on Innovative Applications of Artificial Intelligence (IAAI), July 2006
[8] Tropsha A., Gramatica P., Gombar V.K., The Importance of Being Earnest: Validation is the Absolute Essential for Successful Application and Interpretation of QSPR Models, QSAR & Combinatorial Science, Volume 22 Issue 1, Pages 69 - 77, Published Online: 16 Apr 2003
[9] Michielan L., Pireddu L., Floris M., Bacilieri M., Rodriguez-Tomé P., Moro S., Support Vector Machine (SVM) as alternative tool to assign acute aquatic toxicity warning labels to chemicals. Submitted. 2009
[10] Maunz A., Helma C., Prediction of chemical toxicity with local support vector regression and activity-specific kernels. SAR QSAR Environ. Res. 2008, 19, 413-431.
[11] Tetko I.V., et al., Critical assessment of QSAR models of environmental toxicity against Tetrahymena pyriformis: Focusing on applicability domain and overfitting by variable selection, J. Chem. Inf. Model. 2008, 48 (9), 1733-46