Environmental ChemOinformatics

Grant agreement no.: 238701

Marie Curie Initial Training Center

Sub-Priority ENV2007 3.3.1.1: In-silico techniques for hazard-, safety-, and environmental risk-assessment

ECO – Deliverable: 11

Report about the results of STR fellowships

Due date of deliverable: 31.10.2011
Actual submission date: 31.10.2011

Start date of project: 1 October 2009 Duration: 4 years

Lead Contractor: Helmholtz Zentrum Muenchen
Corresponding author of document: A. Jan Hendriks[1] and Willie Peijnenburg[2]

1. Radboud University, Nijmegen, the Netherlands
2. Leiden University, the Netherlands

Deliverable no: 11
Nature: Report

Project co-funded by the EU Commission within the Seventh Framework Programme		
Dissemination Level		
PU	Public	X
PP	Restricted to other program participants	
CO	Confidential, only for members of the consortium (including the Commission Services	

Recruitment of STR

Despite intense and repeated announcements of positions in Euraxess and at several other job lists (e.g., Nature jobs, CCL list), we found that it was difficult to find suitable and qualified candidates who could fill-in the announced short-time fellowships. However, finally several fine candidates were recruited. During the current reporting period two STRs, namely Michał Świtnicki and Monika Gajewska are staying at the laboratories of Dr. Igor Tetko (Helmholtz Zentrum München, Germany) and Prof. Dr. Roberto Todeschini (University of Milano-Bicocca, Italy), respectively. The detailed progress reports of both STRs are attached as Annex Ia and Ib.

Training of STR

The primary objective of the training of both STRs is to contribute to their education in environmental sciences, in particular toxicological modelling.

Both students learn to access and critically collect data from databases, literature reviews and original papers. The data collected are used to derive QSAR-type of relationships, typically the basic modeling used in risk assessment, especially in the REACh framework. Both ECO students are attached to ongoing projects of the respective host institutes.

Additionally both students get training on different modeling tools. During their stay in their host labs, they are able to actively participate in crosstalk and collaboration, not only between students and their mentors in the respective host institutes but also with the other ECO participants.

Monika Gajewska already took part as an external participant at the Winter School 2011 in Idstein. During their fellowship, Michal Switnicki and Monika Gajewska participated in the Summer School 2011 in Leiden. Both students have worked according to their individual Career Development Plans (CDPs). They prepared and gave a presentation of their work during the Summer School 2011 in Leiden

(attached as Annex IIa en IIb). In addition, Monika Gajewska has given two presentations at international workshops, participated in ECO online training and attended various seminars (see Annex Ib). In addition, she took an internship at Dr. Igor Tetko's group (Institute of Bioinformatics and Systems Biology, Helmholtz Zentrum München, Germany).

The data collected by the STRs have been made publicly available at the web site of OCHEM http://ochem.eu and additionally will be published in international peer-reviewed scientific journals. This will provide an easy mechanism to measure the progress and success of the ECO-ITN.

Evaluation of the scientific results

Michał Świtnicki collected data for toxic effects of drugs and chemicals on mammals from various sources. Most data applied to rats and humans. He is now in the process of developing QSAR models from these data using advanced machine learning methods.

Monika Gajewska has also collected data on toxic effects, yet on 3 species of algae and 1 species of daphnids, using 7 differences sources of data. For different endpoints, she has obtained data on 14-644 compounds. While it proved to be difficult to find studies that are sufficiently reliable for QSAR development, she is now analyzing the sets using different statistical techniques, including PCA, Random Forest Regression etc. Currently, R-scripts are written to allow in-depth interpretation of model results.

Both STRs demonstrated very fast integration in the host labs and they also participated in several socially events of the respective laboratories.

Conclusions and outlook

In summary, both ECO short-term fellowships were a great success for both the selected fellows and their host institutes.

HelmholtzZentrum münchen
German Research Center for Environmental Health

Marie Curie Initial Training Network

Environmental Chemoinformatics (ECO)

Intermediate project report

13 October 2011

Quantitative Modelling Of Toxicological Data

Duration of Short Term fellowship:
15th June 2011 – 31st December 2011

Early stage researcher:
Michał Świtnicki

Project supervisor:
Igor Tetko and Monica Campillos

Research Institution:
Helmholtz Zentrum München

Introduction

The toxic effects of environmental chemicals and the adverse effects of drugs albeit probably caused by similar molecular mechanisms have been traditionally studied separately. The integration of the two data types will increase the coverage of chemical space on toxic effects and thus improve the applicability of predictive models. In this project we will collect and integrate toxicological data from environmental chemicals and drugs with the aim to build predictive models of chemical toxicological effects applicable to novel compounds.

Materials and Methods

In this project, the following workflow has been derived for acquiring the data:

1) Mine various publicly available datasets containing data about toxic effects of environmental chemicals and side effects of drugs in human.
2) Map the observations made in these datasets to common ontology. The resource used in case of this project was Unified Medical Language System (UMLS) with effects represented as concepts. Each concept is represented with unique ID and has a very defined location in the hierarchy of ontology.
3) Create a custom database containing parsed data to allow for proper comparison, analysis, and later reuse.

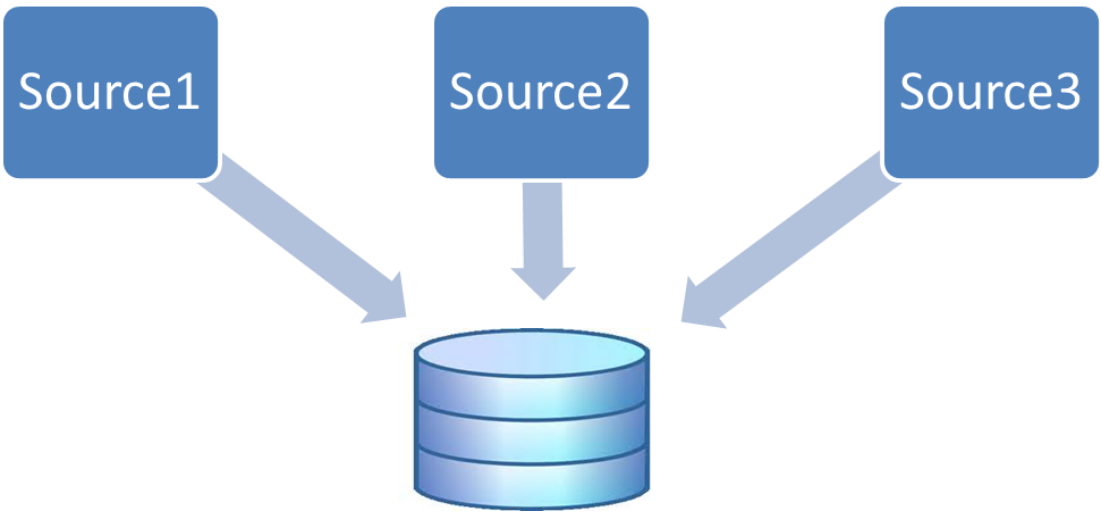


Fig.1 Data collection approach.

Currently, 3 resources have been mined so far, and these include:

- ToxRefDB
- SIDER
- Histopathology data from rat liver xenobiotic and pharmacology database (Ganter et al, 2005)

For the mapping of these data, the combination of the following dictionaries (within UMLS) was used: COSTAR (Computer-Stored Ambulatory Records), CHV (Consumer Health Vocabulary) and MSH (Medical Subject Headings).

In total, data was obtained for 1475 compounds, described in terms of 1748 non-redundant concepts (effects) for 4 organisms. Records total up to around 129 000.

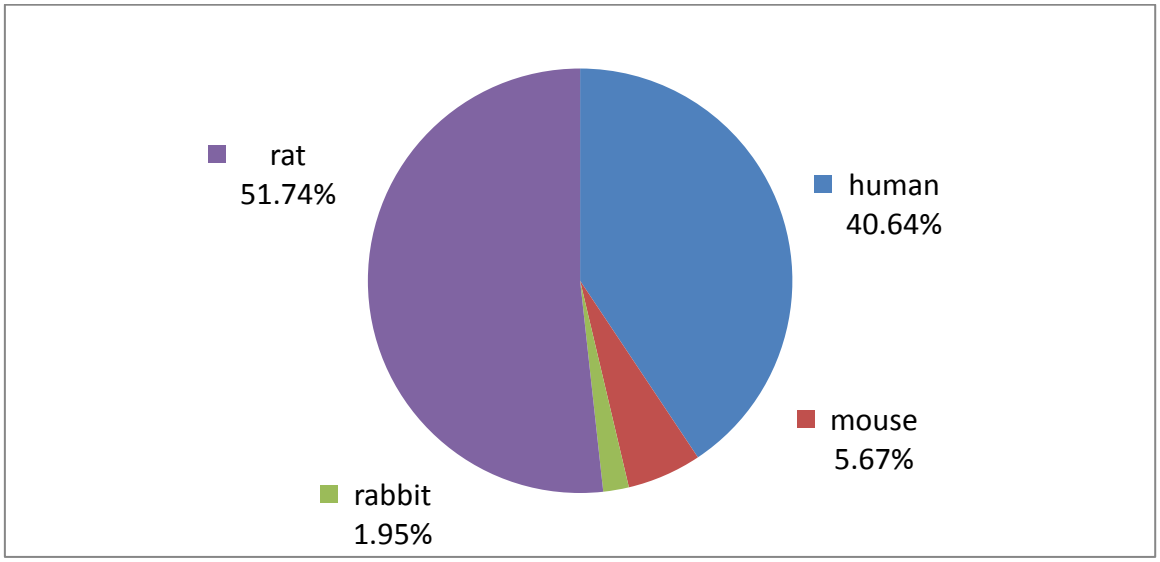


Fig. 2 Distribution of data across different organisms.

Then, assuming completeness of data (i.e. all compounds have been effectively tested for all collected side effects/toxicological end points), we tried applying Quantitative Structure-Activity Relationship (QSAR) modelling approaches to build predictive models for each concept (side effect/toxicological end point).

In this approach, compounds annotated with concept of interest are modelled against all other compounds from the entire dataset which are treated as negative control. This strategy has been tested utilizing OCHEM, a modelling environment developed in the group of Dr. Tetko, with several representative concepts and produced models with externally cross-validated accuracies ranging from 65% to 80%.

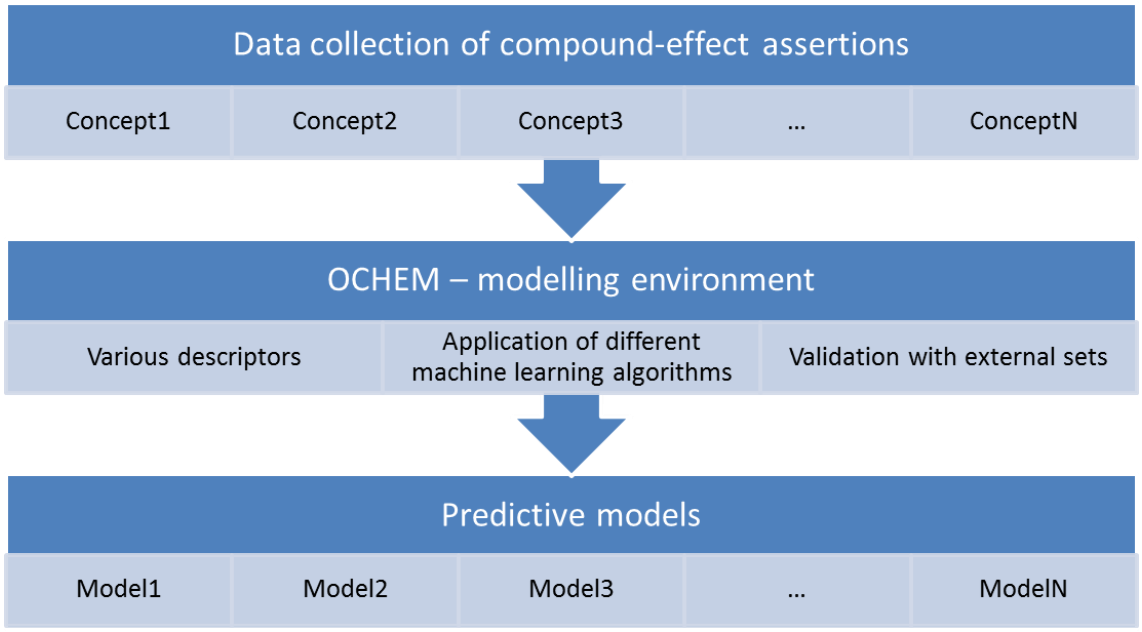


Fig. 3 Illustration of approach to build predictive models for the data.

Currently, we set a threshold of minimum 100 molecules when creating training sets for our models. This constraint helped increasing the statistical power of achieved models but on the other hand, limited number of modelable concepts to 266.

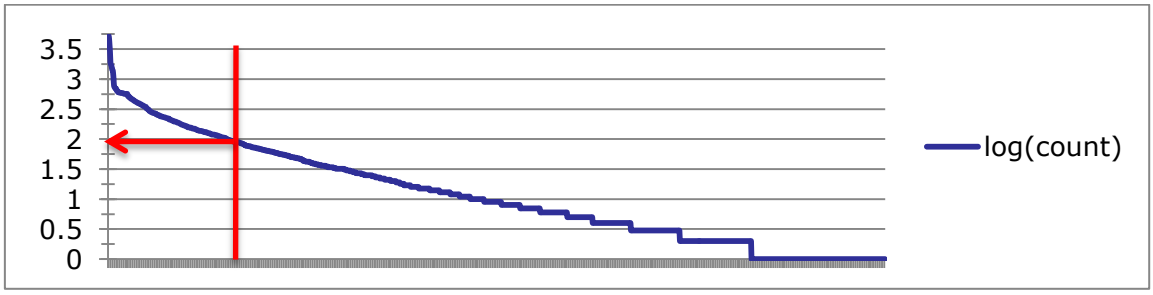


Fig. 4 After setting a threshold, 266 concepts were available to model.

Results

To date, it was possible to build models for chosen concepts, however, I did not manage to develop tools to analyse the obtained models. Also, there are still couple of small issues and bugs in the software repertoire I am using so the predictions are not entirely reliable. This problem will hopefully be overcame in the nearest future. For couple of representative concepts, the externally cross validated accuracies of models obtained by decision tree algorithm with E-State and ALogPS descriptors are as follows:

Effect	Hemolytic anemia	Urinary tract infection	Cardiac dieseases	Peripheral nephropathy	Eosinophilia
Accuracy	79.8%	71.9%	76.2%	78.3%	70.3%

Fig. 5 Example of obtained models with cross-validated accuracies.

Further results, along with statistical analysis of all obtained models will be achieved later on, when tools for these analyses will be developed.

Current goals

1. Perform statistical assessment of achieved models
2. Try combinations of different descriptors and machine learning algorithms
3. Add more data: Toxicology Literature Online (TOXLINE), EPA ACToR
4. Prepare manuscript for publication of the results

Bibliography

Ganter B, Tugendreich S, Pearson CI, Ayanoglu E, Baumhueter S,Bostian KA, Brady L, Browne LJ, Calvin JT, Day GJ, Breckenridge N,Dunlea S, Eynon BP, Furness LM, Ferng J, Fielden MR, Fujimoto SY,Gong L, Hu C, Idury R et al (2005) **Development of a large-scale chemogenomics database to improve drug candidate selection and to understand mechanisms of chemical toxicity and action.** J Biotechnol 119: 219–244

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

Marie Curie Initial Training Network
Environmental Chemoinformatics (ECO)

Project report
12 October 2011

The ECO Methods for selection of structural features that influence substance toxicities

Duration of Short Term fellowship:
June 2011 – March 2012

Early stage researcher:
Monika Gajewska

Project supervisor:
Prof. Roberto Todeschini

Research Institution:
University of Milano-Bicocca

INTRODUCTION

There are many commercial chemicals found in aquatic systems for which still either no information on toxicity exists or studies are quite limited. Recent legislation requires short-time assessment for their toxicity to aquatic organisms in order to determine which of these chemicals need to be further studied. As a result the new European Union chemical control system adoption, called Registration, Evaluation, and Authorization of Chemicals (REACH), Quantitative structure–activity relationship (QSAR) models are expected to play a crucial role in reducing a number of animals to be used for toxicity testing.Therefore the objective of the study is to collect and analyze available toxicity data for aquatic systems in order to choose the chemicals of interest, then to develop a QSAR models to predict acute in silico toxicity of these chemicals and finally, if possible, to compare the resulting models with the literature ones. However, the main interest of the project lays in the very strategy to construct these models. In order to generate robust, clear and simple predictive models having huge data matrix at one's disposal (up to by several thousands of molecular descriptors for a numerous observations) only the most relevant molecular descriptors should be selected. For this reason reliable and effective variable selection algorithm is needed, which as such, is not yet introduced in literature and there is still much of a controversy among modelers whether a mathematical technique, if any, can be beneficial for model construction.
In this project, the focus is on short term toxicity to aquatic invertebrate organisms (Daphnia, Algae).

MATERIALS AND METHODS

The initial task of the project was a careful collection of reliable experimental values presented in literature or in available on-line databases for acute aquatic toxicity. Only invertebrates (selected species of Algae and Daphnia Magna) were considered. These data would be later used in development and validation of feature selection algorithms.
After a profound insight into available feature selection (FS) methodologies final choice for a method suitable for the investigated problem, to analyze, implement and extend, if possible, has been done.
R software was used, as a tool for FS algorithms implementation. These were presented in a form of a script, incorporating functionalities of several packages and novel functions.
In the final stage QSAR models were to be developed and provided with their sufficient statistical and predictive characteristics, models validation, comparison with existing literature ones, applicability domain; conclusions, suggestions.
Lastly, an effort has been put in preparation of adequate documentation with a simple, straightforward and clear explanation of applied methods.

The software and services used in the project:

- Dragon 6 (http://www.talete.mi.it/products/dragon_description.htm)
- OCHEM portal (http://ochem.eu; online database with modeling environment)
- QSAR toolbox (http://www.qsartoolbox.org/)
- Mobydigs (http://michem.disat.unimib.it/chm/)
- R (http://www.r-project.org/) with the following packages: randomForest, Boruta, caret, party, clValid, cluster, subselect
- Tinn-R, R code editor (http://sourceforge.net/projects/tinn-r/)

Two algorithms, particularly famous in recent biostatistics have been chosen in this study: Random Forest (RF) and Genetic Algorithm (GA).

RESULTS AND DISCUSSION

Database on acute aquatic toxicity for selected invertebrates

Description:

Thorough review of REACH guidelines addressing safe use of chemicals, and in their accordance, careful choice made for the endpoint and species of interest. Subsequently, QSAR models development for a predicting aquatic toxicity. Collection of available experimental data; focus toward variety of organic compounds (industrial organic chemicals, pharmaceuticals, pesticides, surfactants).

Results:

Database on acute aquatic toxicity for Algae and Daphnia Magna with the endpoints: effective concentration (EC50) and lethal concentration (LC50)

SPECIES	ENDPOINT	TYPE AND NUMBER OF COMPOUNDS	DATA SOURCE
Daphnia Magna	48-h LC50	300 Various organics	U.S. EPA AQUIRE (2002)
	48-h LC50	222 Pharmaceuticals	Toxicology Letters 187 (2009) 84–93
	96-h LC50	262 Pesticides	Bioorganic & Medicinal Chemistry 14 (2006) 2779–2788
	48-h EC50 Immobilization	130 Various organics	Journal of Toxicology and Environmental Health, Part A, 72: 1181–1190, 2009
	48-h EC50 Immobilization	17 Substituted benzaldehydes	Chemosphere Vol. 37, No. 1, pp. 79-85, 1998
	48-h EC50 Immobilization	644 Various organics and inorganics	Ministry of the Environment in Japan: Eco-toxicity tests of chemicals (March 2011)
	48-h EC50 Immobilization	22 Benzoic acids	Chemosphere 59 (2005) 255–261
	48-h EC50 Immobilization	6 Anionic surfactants linear alkylbenzene sulphonates (LAS) and 21 ester sulphonates (ES)	Chemosphere 63 (2006)1443–1450
	48-h EC50 Immobilization	74 Organic, inorganic esters	Chemosphere 58 (2005) 559–570
	48-h EC50 Immobilization	40 Various organics	U.S. EPA database ECOTOX (+2000);
	48-h EC50 Immobilization	125 organic chemicals (derived from the European priority list in compliance with Directive 76/464/EEC)	Ecotoxicology and Environmental Safety 49, 206}220 (2001)
Chlorella Vulgaris	15 min EC50; Inhibition of enzyme activity (Fluorescein diacetate)	91 Diverse organic industrial compounds (aliphatic, aromatic)	Chem. Res. Toxicol. 2004, 17, 545-554
	96-h EC50; Inhibition of the activity of acetyl-CoA carboxylase	40 Herbicides	Ecotoxicology and Environmental Safety 51, 128 - 132 (2002)
	96-h EC50; Growth inhibition	14 Pesticide adjuvants	Ecotoxicology and Environmental Safety 58 (2004) 61–67
Pseudokirchneriella Subcapitata	48-h EC50; Biopopulation (Biomass-type based on the cell density)	108 Various organic compounds	Ecotoxicology and Environmental Safety 72 (2009) 1514–1522

	48-h EC50; Growth rate inhibition	20 Benzoic acids	Journal of Hazardous Materials 165 (2009) 156–161
	48-h EC50; Growth rate inhibition	13 Substituted anilines	Environmental Toxicology and Chemistry, Vol. 26, No. 6, pp. 1158–1164, 2007
Scenedesmus obliquus	48-h EC50; Growth rate inhibition	40 Substituted benzenes	Chemosphere 44 (2001) 437-440
	48-h EC50; Growth rate inhibition	25 Nitroaromatics	Chemosphere 59 (2005) 467–471

Problems and limitations:

There are limited resources including available scientific literature and online databases with experimental data on endpoints in question for aquatic toxicity. It is difficult to find a single or several comparable studies with numerous compounds investigated in experiment what would be satisfactory and necessary for reliable QSAR models construction.

Review of feature selection methodologies. Regression by Random Forest.

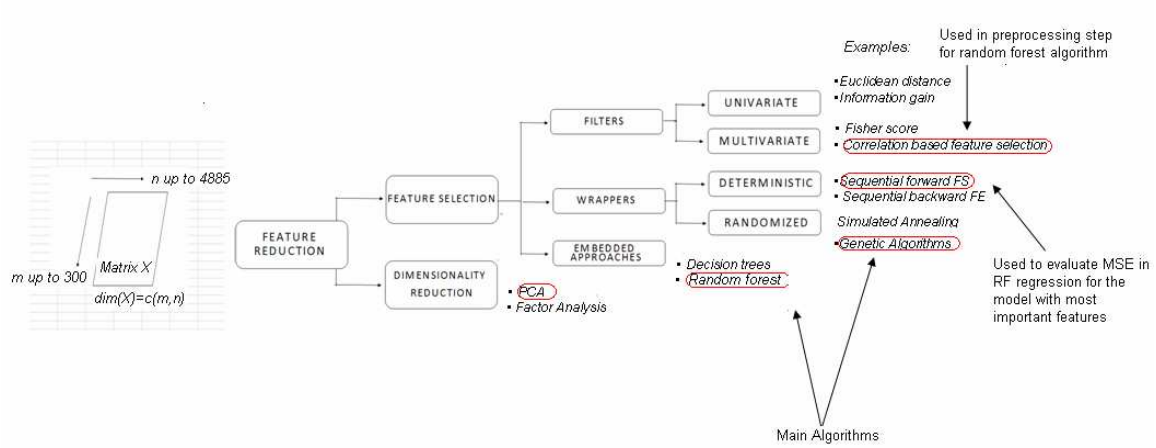
Description:

Development, elaboration and implementation of random forest based R scripts for supervised and unsupervised feature selection; application to collected datasets, performance evaluation, comparison with existing methodologies in OCHEM website.
Initial work toward in silico QSAR modeling to predict environmental toxicity of collected chemicals for Algae and Daphnia. In the final stage, their proper validation, sufficient statistical properties and comparison with the literature models.

Results:

a) Two separate functions implemented in R for dimensionality reduction and information visualization for exploring similarities or dissimilarities in data: principal component analysis and multidimensional scaling

b) Three algorithms for feature selection based on Random Forest: collection of functions for data handling, visualization, storage, generation of clear, comprehensible results.
Short summary of all three algorithms is presented below.



Supervised Random Forest regression

1. A matrix with molecular descriptors (up to 18 descriptors blocks: 0,1,2 D descriptors) for number of observations (compounds) considered as training set created was by means of Dragon 6.
2. Data input into R script, check for correlations, near to zero variance (pre-processing), statistical tests to remove the most irrelevant molecular descriptors.
3. Random Forest variable importance measure (two types) , stepwise addition of variables from the most to the least important one in a model to evaluate mean square error (MSE), choice for a model with the smallest MSE.
4. In search for a model with lower number of variables and lower or comparable MSE error, calculate this error for a model with all possible combinations of up to 10 most important variables (due to high computational burden).
5. Return of a new matrix with all statistically important descriptors, final decision on their acceptance and possible relevance with a measured activity is made by a modeler prior to model build-up.
6. Random Forest model construction, prediction evaluation, application to a previously designated test set, statistics

Supervised Random Forest Regression with Conditional Variable Importance

1. A matrix with molecular descriptors (up to 18 descriptors blocks: 0,1,2 D descriptors) for number of observations (compounds) considered as training set created was by means of Dragon 6.
2. Data input into R script, check for correlations, near to zero variance (pre-processing).
3. Unconditional variable importance measure used to define a threshold for variables pre-selection for conditional measure (statistical test implemented in “party” R package).
4. Variables are checked for existing trends and eventual monotonicity to distinguish from random fluctuations. Only descriptors stated: important and with non-random trend are returned.
5. Modeler makes final decision about the variables used for model build-up.
6. Random Forest model construction, prediction evaluation, application to a previously designated test set, statistics

Unsupervised Random Forest Regression with Cluster Analysis

1. A matrix with molecular descriptors (up to 18 descriptors blocks: 0,1,2 D descriptors) for number of observations (compounds) considered as training set created was by means of Dragon 6.
2. Data input into R script, check for correlations, near to zero variance (pre-processing).
3. RF proximity matrix replaced by dissimilarity matrix. Variable importance calculation.
4. Internal Validation for existence of clusters and appropriate clustering method. In case of clustering, their geometrical interpretation, otherwise choice for representatives on a basis of dissimilarity level.
5. Return a matrix of variables stated important and adequately dissimilar
6. Further model construction possible by any available method (not only random forest)

Problems and limitations:

- For each of the algorithms separate R script must be developed.
- Programming is time consuming and needs constant improvements, modifications and check for correctness.
- High dimensionality problem (far too higher number of molecular descriptors than observations), this exclude the application of many FS algorithms, however Random Forest is said to perform well in such difficulty.

- Problem of correlated descriptors. This affects robustness and performance of a model. An effort is done to reduce their number in a final matrix. Therefore, conditional variable importance is used next traditional approach.
- Problem of overfitting. This is almost unavoidable problem in regression, however, an effort in first two approaches is done to limit it. Last, unsupervised method is implemented where no experimental values (thus no regression) are considered so overfitting is not there a limitation.

Future tasks:

- Finalization of these three (supervised and unsupervised) R scripts for feature selection. They should be clear, simple and comprehensible, easily used by anyone. Few more novel additions and detailed check for the scripts functionality have to be done.
- Careful comparison with methodologies implemented in OCHEM services
- Separate attention given to Genetic Algorithm (MobyDigs and R packages) and its comparison with Random Forest approach.
- QSAR Models; their construction and prediction performance is not yet well completed and validated. For each of species and a given endpoint, model is to be created with a satisfactory explanation and literature comparison.
- Proper data directory and documentation on algorithms functionality must be provided.
- Possible work toward applicability domain, further work suggestions

REFERENCES

1. Hartmann W. M. Abstract: **"**Dimension Reduction vs. Variable Selection"; SAS Institute, Inc., Cary NC, USA
2. Dudek A. Z., Arodz T., Gálvez J. (2006) Combinatorial Chemistry & High Throughput Screening, 9, 213-228
3. Guyon I., Elisseeff A. (2003) Journal of Machine Learning Research 3 1157-1182
4. Kursa M.B., Rudnicki W. R.,(2010) Journal of Statistical Software Vol. 36, Issue 11
5. Liaw A. and Wiener M. (2002) "Classification and Regression by Random Forest" Vol. 2/3
6. Svetnik V., Liaw A., (2003) J. Chem. Inf. Comput. Sci.**,** 43, 1947-1958
7. Janecek G.K., Gansterer W.N, JMLR: Workshop and Conference Proceedings 4: 90-105
8. Strobl C., Boulesteix A.-L. (2008); Technical Report Number 23, Department of Statistics University of Munich
9. Polishchuk P.G., Muratov E.N. (2009); J. Chem. Inf. Model.**,** 49, 2481–2488

TRAININGS & SCIENTIFIC MEETINGS & SCHOOLS

- Attendance in seminars organized within the group to present the research results; Two presentations given: " Database setup for QSAR studies" and "Introduction to R"
- COSMOS Workshop ; 16th June 2011; JRC Ispra, Italy; Introduction to molecular (systems biology models), cellular (DEBTox models),organs (2D liver model) and organisms (PBTK models)
- ECO project online training: 1st June, 3rd August 2011
- Participation to doctoral seminar, presentation given by Faizan Sahigara; 6th July 2011 and Kamel Mansouri; 14th September 2011
- August 2011 Internship at Dr Igor Tetko's group at the Institute of Bioinformatics and SystemsBiology, Helmholtz Zentrum München- German Research Center for Environmental Health
- Participation in OpenTox InterAction Meeting: Innovation in Predictive Toxicology, In Vitro and
- In Silico Modelling, Applications, REACH, Risk Assessment - 9-12th August 2011
- 19th-30th September 2011, participation in Environmental ChemOinformatics Summer School at Leiden University (LU). http://www.eco-itn.eu/node/86

HelmholtzZentrum münchen
German Research Center for Environmental Health

Quantitative modelling of toxicological data

Michał Świtnicki

Helmholtz Zentrum München
Institute of Bioinformatics and Systems Biology

Leiden, 26/09/11

HELMHOLTZ ASSOCIATION

Couple of words about me

UNIVERSITY
of
GLASGOW

Institute of Biochemistry
and Biophysics
Polish Academy of Sciences

The project

- Joint effort between Chemoinformatics team and Systems Biology of Small Molecules group of Dr Mónica Campillos, IBIS
- A pilot study with the aim to build predictive models of chemical toxicological effects
- Anticipated end of project in the end of year

Motivation for the research

The toxic effects of environmental chemicals and the adverse effects of drugs albeit probably caused by similar molecular mechanisms have been traditionally studied separately.

The integration of the two data types will increase the coverage of chemical space on toxic effects and thus improve the applicability of predictive models.

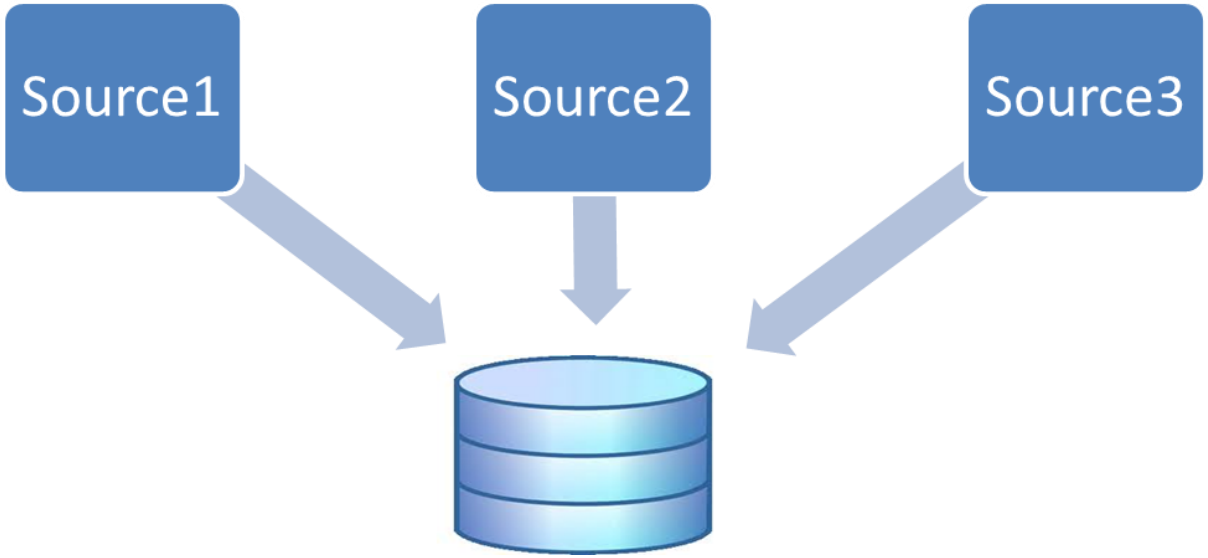
Appropriateness for REACH

Achieving the goal for this project would fulfill these goals of REACH:

- To limit vertebrate animal testing
- To define structural alerts for screening new compounds against undesired effects in human

Data collection approach

1) Mine publicly available datasets containing both types of data and map it to a common ontology (Unified Medical Language System – UMLS).
2) Create a custom database containing parsed data.



Current collection of data

3 sources have been mined so far:

- ToxRefDB (in vivo animal toxicology data): exact match mapping,
- SIDER (side effects of current or withdrawn drugs in human): reuse of existing mappings,
- histopathology data from rat liver xenobiotic and pharmacology database (Ganter et al, 2005): manual mapping.

The most used ontologies (dictionaries) include COSTAR (Computer-Stored Ambulatory Records), CHV (Consumer Health Vocabulary) and MSH (Medical Subject Headings) accessed via UMLS.

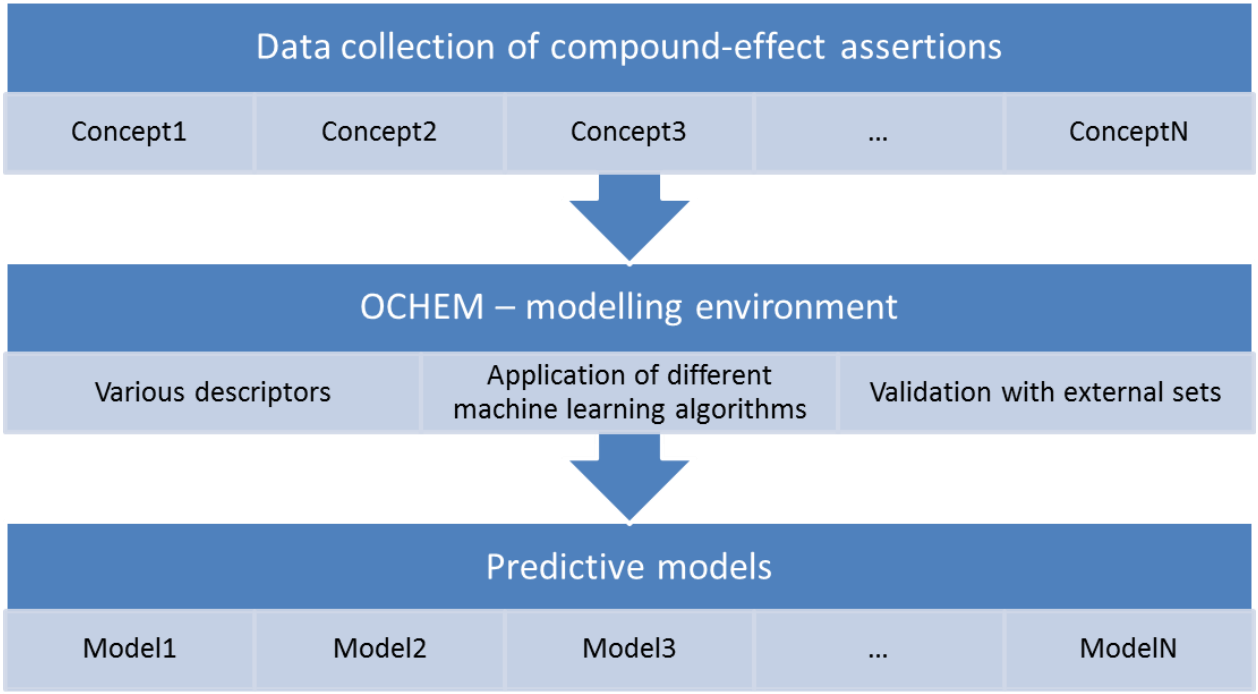
Summary:

- 1660 compounds
- 1748 concepts (effects)
- ~129 000 records.

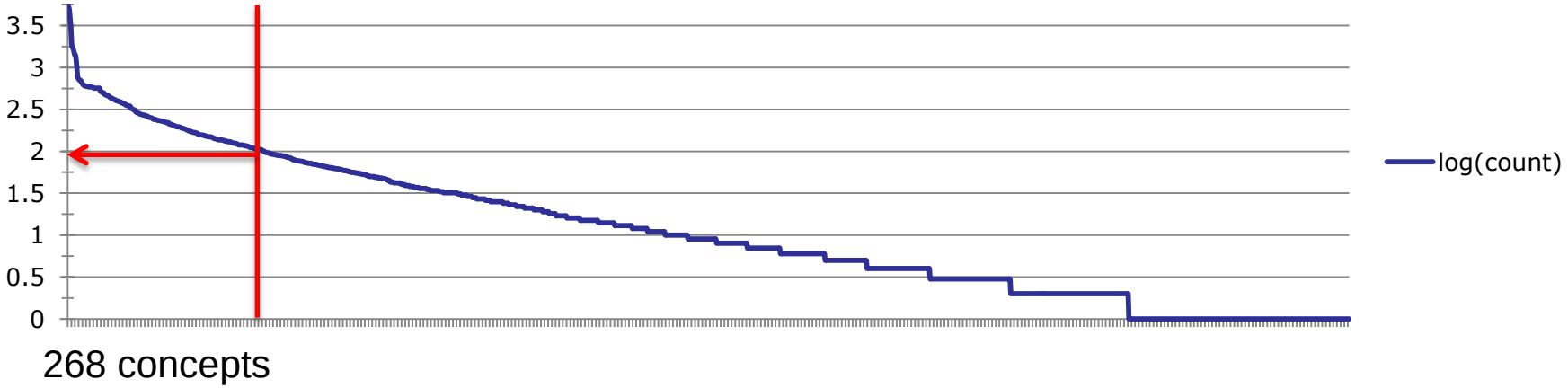
Approaches to build predicitve models

Assumption : completeness of data

Utilizing OCHEM, apply Quantitative Structure-Activity Relationship (QSAR) modelling approaches to build predictive models for each concept (side effect/tox end point).



Cont. - distribution of concepts



Cont.

Compounds annotated with concept of interest are modelled against all other compounds from the entire dataset which are treated as negative control.

This approach has been tested with several representative concepts and produced models with externally cross-validated accuracies ranging from 63.4% to 84.1%.

Current goals

- Add more data.
  Toxicology Literature Online (TOXLINE), EPA ACToR

- Proceed with modelling of the entire list of concepts.

- Obtain statistical assessment of achieved models and try combinations of different machine learning algorithms with different descriptors in order to obtain optimal results.

Further development

(i.e. other possible approaches to explore with little chances for proceeding within the current contract)

Provided that the naïve approach will continue to produce sensible results for the rest/considerable part of dataset, proceed to stage 2 of the project.

In this stage, a more sophisticated approach would be exploited. Here, one would have to define a similarity measure between concepts (i.e. based on hierarchy of used ontologies) and annotate every compound with respective scores for each concept used. This would result in an extended effect profile for every compound in the dataset.

Compound	Concept1	Concept2	Concept3	…	ConceptN
Comp1	Score1	Score2	Score3	…	ScoreN
Comp2	Score1	Score2	Score3	…	ScoreN
…	Score1	Score2	Score3	…	ScoreN
CompM	Score1	Score2	Score3	…	ScoreN

These profiles, would then be used to inform the machine learning approaches and thus improve the predictive power of obtained models. This would allow to predict spectres of effects rather than single concepts.

Learning outcomes & impact on career

- Valuable insight into the field of chemoinformatics and systems biology of small molecules.
- Introduction to new resources for biological data.
- Application of machine learning approaches for data modelling.
- Understanding the idea of building predictive models.
- Improving personal network of professional connections.

Environmental ChemOinformatic (ECO)
Marie Curie Initial Training Network

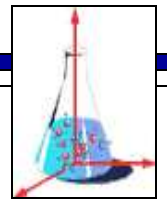
MONIKA GAJEWSKA

Short-term project:
The ECO Methods for selection of structural features that influence substance toxicities

Supervisor: Prof. Roberto Todeschini

Milano Chemometrics and QSAR Research Group,
Department of Environmental Sciences,
University of Milano Bicocca,
P.za della Scienza, 1-20126 Milano, Italy

Educational background and relevant scientific work



1. **Master's degree in Chemical Engineering and Chemical Technology with the specialization in Fine Chemicals**

Master's thesis: "On characteristic times in parabolic diffusion in polymeric membranes"

Silesian University of Technology (Politechnika Śląska);
Faculty of Chemistry in Gliwice, Poland

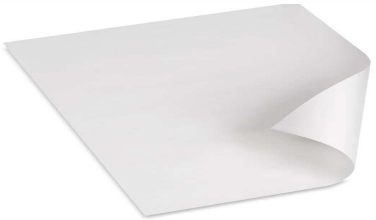
Joint Research Centre, Institute for Heath and Consumer Protection;
One- year Internship project :
"Preprocessing of chromatographic data with the application in Metabolomics"

JRC
EUROPEAN COMMISSION

ihcp
Institute for Health and Consumer Protection

Work towards REACH implementation

REACH Registration, Evaluation, Authorisation of Chemicals

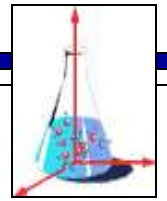


- **Quantitative structure-activity relationship (QSAR)** models development aimed at promotion of the alternative (non-animal) methods expansion for hazard assessment of chemicals

- Collection of available experimental data on target species of interest and the endpoint for acute aquatic toxicity tests for invertebrates performed according to:
  - OECD 201(Algae, Growth Inhibition Test)
  - OECD 202 (*Daphnia* sp., Acute Immobilisation Test and Reproduction Test)

- Feature selection methods, their unbiased algorithms, properly validated and described; simplicity, transparency to the users, directed against weaknesses and drawbacks of existing approaches

Main goals of the project



1. Database set-up for acute aquatic toxicity for selected algae types and Daphnia Magna

2. Feature selection methods: random forest (R) and genetic algorithm (MobyDigs)

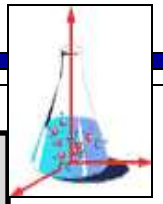
   a) Supervised random forest regression approach

   b) Supervised random forest incorporating conditional variable importance

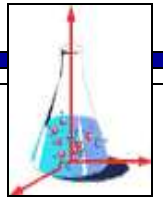
   c) Unsupervised random forest and cluster analysis

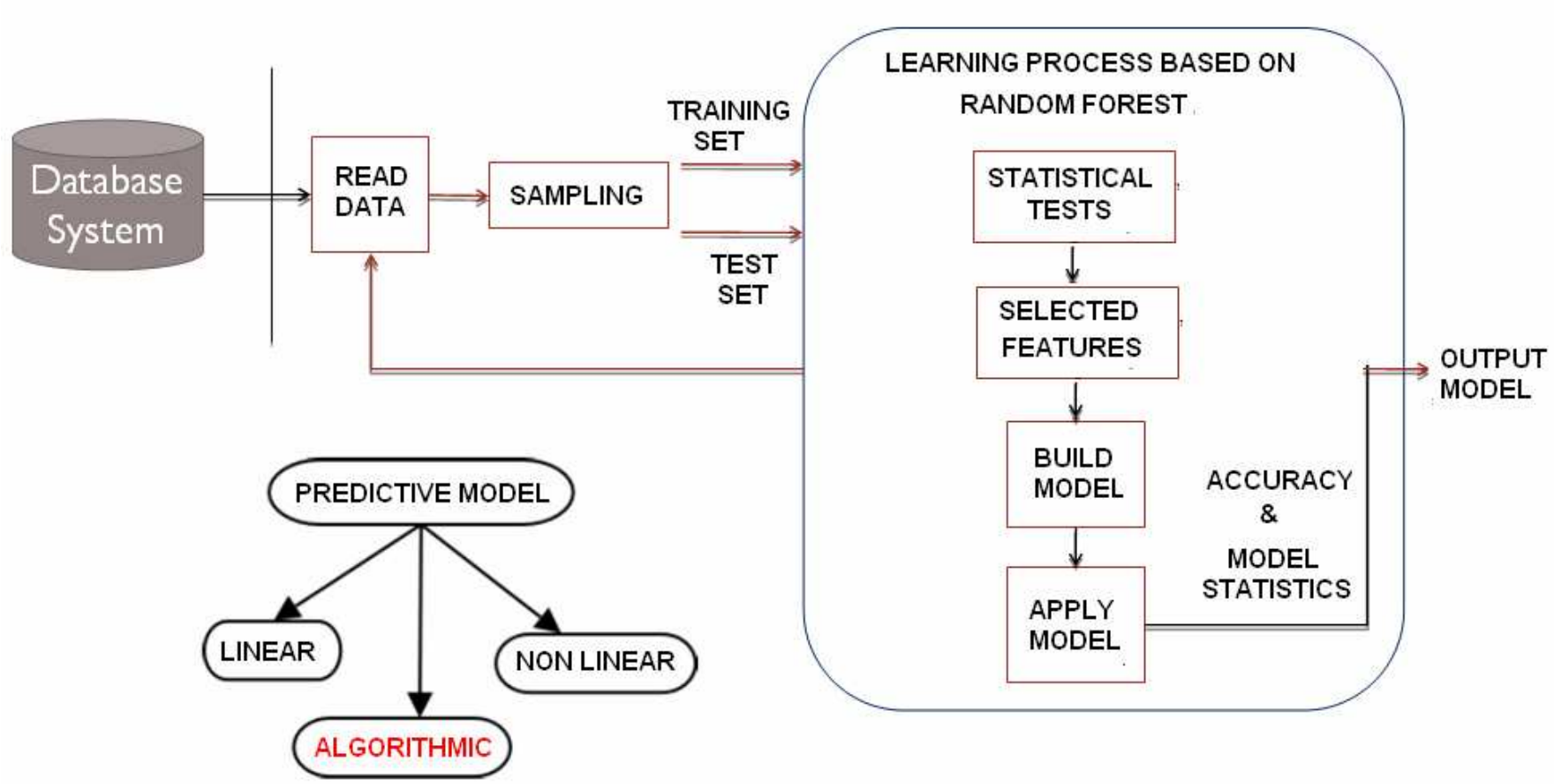
3. QSAR models, their evaluation and comparison with the already proposed models; literature and available on-line database review

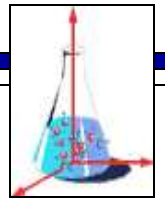
Selected datasets on acute, aquatic toxicity



SPECIES	TYPE OF COMPOUNDS	ENDPOINT	REF.
Chlorella vulgaris 	Diverse organic industrial compounds (aliphatic, aromatic) **91**	Inhibition of enzyme activity (Fluorescein diacetate)	Chemical Research in Toxicology (2004)
Pseudokirchneriella subcapitata 	Diverse organic industrial compounds (aliphatic, aromatic) including benzoic acid derivatives **110**	Biomass-type based on the cell density	Environmental Toxicology and Chemistry (2007); Journal of Hazardous Materials (2009)
Scenedesmus obliquus 	Substituted benzenes; phenols and anilines **61**	Growth inhibition	Chemosphere (2001); Journal of Environmental Sciences (2008)
Daphnia Magna 	Diverse organic, industrial compounds; organic esters; substituted benzaldehydes and benzoic acids **800** Pharmaceuticals **222**	Mortality; Immobilization	Chemosphere (2005); US.EPA database AQUIRE (+1953); Journal of Toxicology and Environmental Health (2009); Toxicology Letters (2009)




Database System
READ DATA
SAMPLING
TRAINING SET
TEST SET
LEARNING PROCESS BASED ON RANDOM FOREST
STATISTICAL TESTS
SELECTED FEATURES
BUILD MODEL
APPLY MODEL
ACCURACY & MODEL STATISTICS
OUTPUT MODEL
PREDICTIVE MODEL
LINEAR
NON LINEAR
ALGORITHMIC

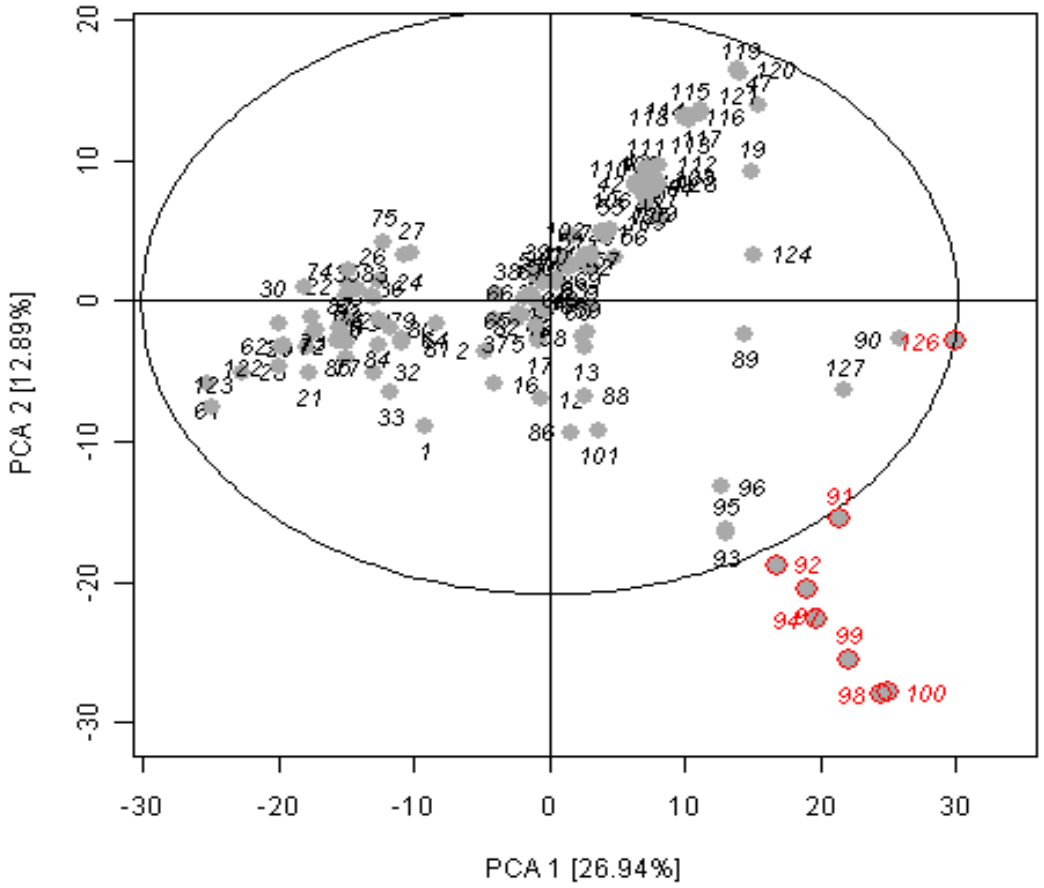


Test species: Pseudokirchneriella Subcapitata

Endpoint: 48-h, Effective concentration (EC50)

Chemicals: 128 Industrial organic chemicals, including aliphatic and aromatic compounds

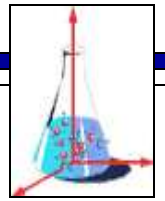
Molecular Descriptors: 548



Variable selection should aim at:

- Better interpretation of a model and structure of a data
- Reduction of irrelevant, redundant and noisy features
- Reliable prediction model formation

Supervised random forest regression implemented in R



DATA INPUT AND PREPROCESSING; CORRELATIONS, NEAR TO ZERO VARIANCE

BORUTA ALGORITHM FOR IRRELEVANT MDs ELIMINATION

RF VARIABLE IMPORTANCE; STEPWISE VARIABLE ADDITION AND MSE

MSE FOR COMBINATIONS OF 10 MOST IMPORTANT VARIABLES

CHOICE TOWARDS LOWEST MSE AND NUMBER OF MDs

MODEL BUILD-UP AND OUT-OF-BAG PREDICTION STATISTICS

INTERPRETATION AND FINAL DECISION

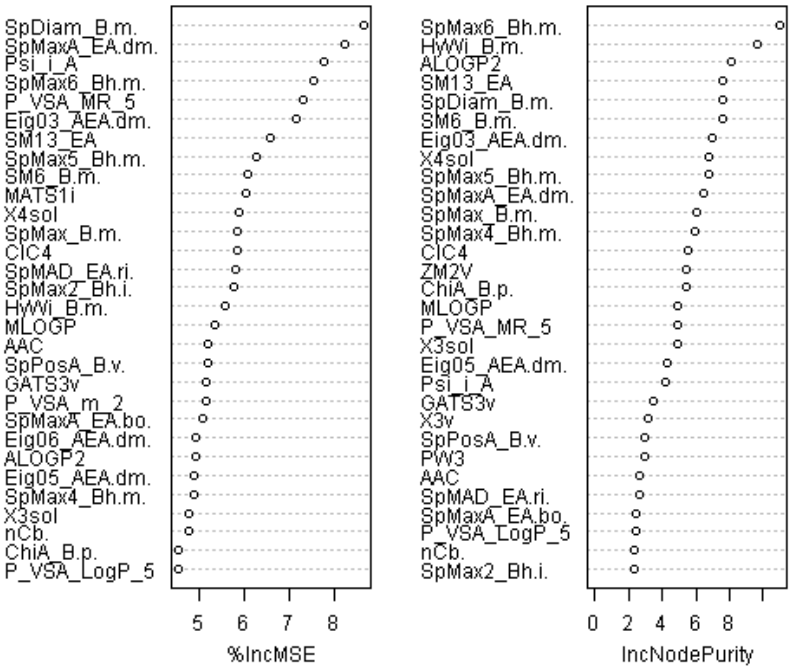
DATA OUTPUT; RF MODEL AND STATISTICS

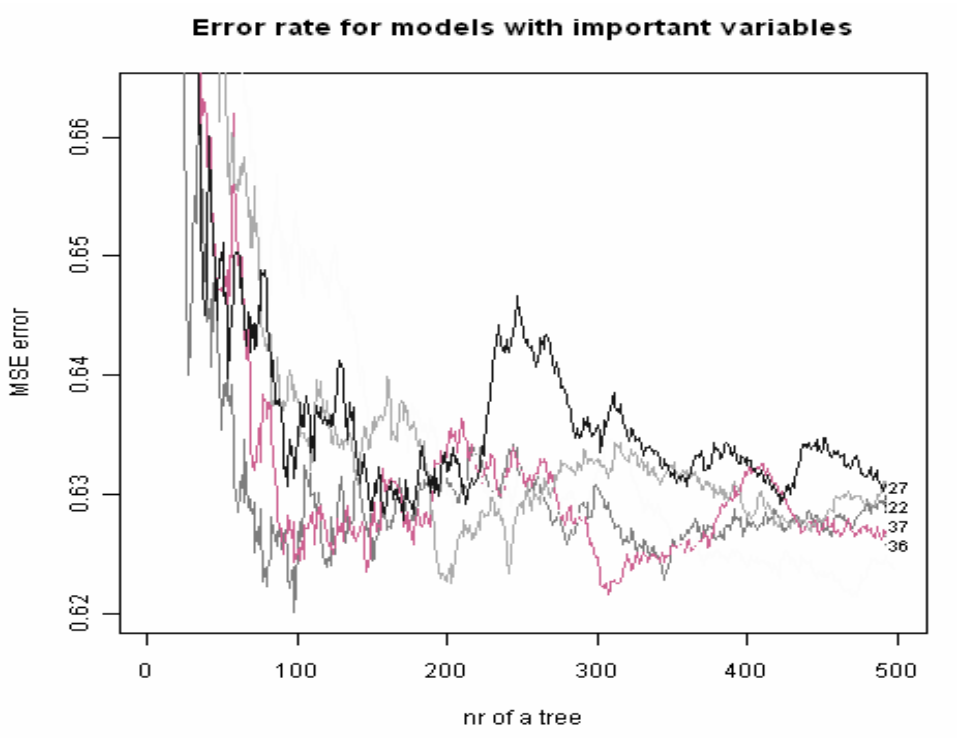
MD = MOLECULAR DESCRIPTORS

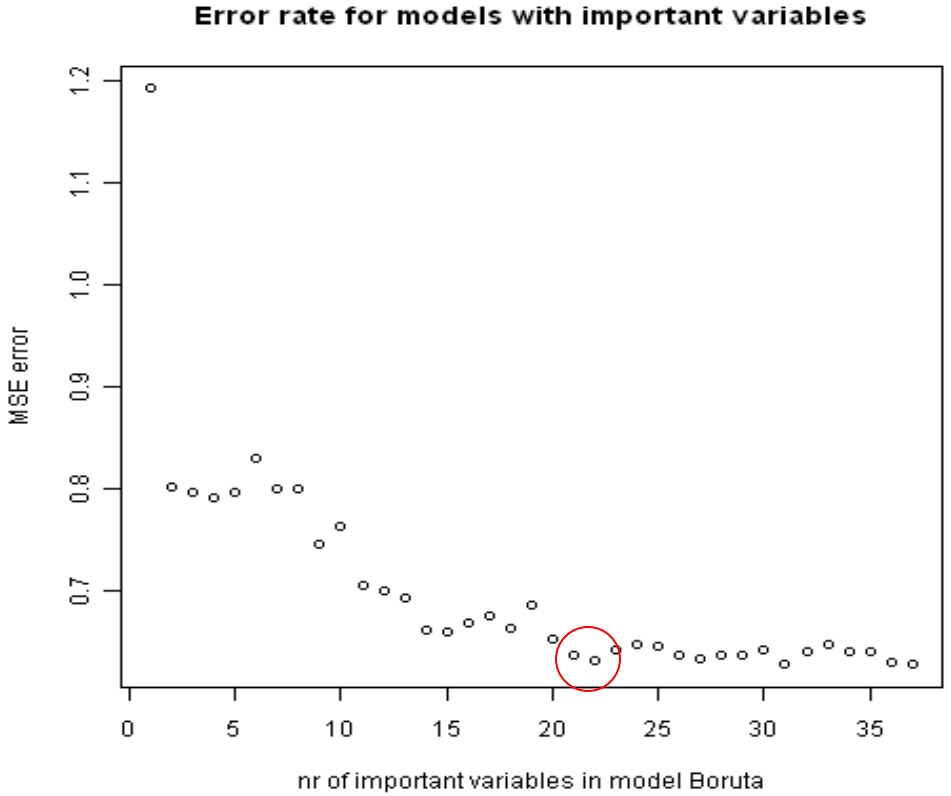
RF = RANDOM FOREST

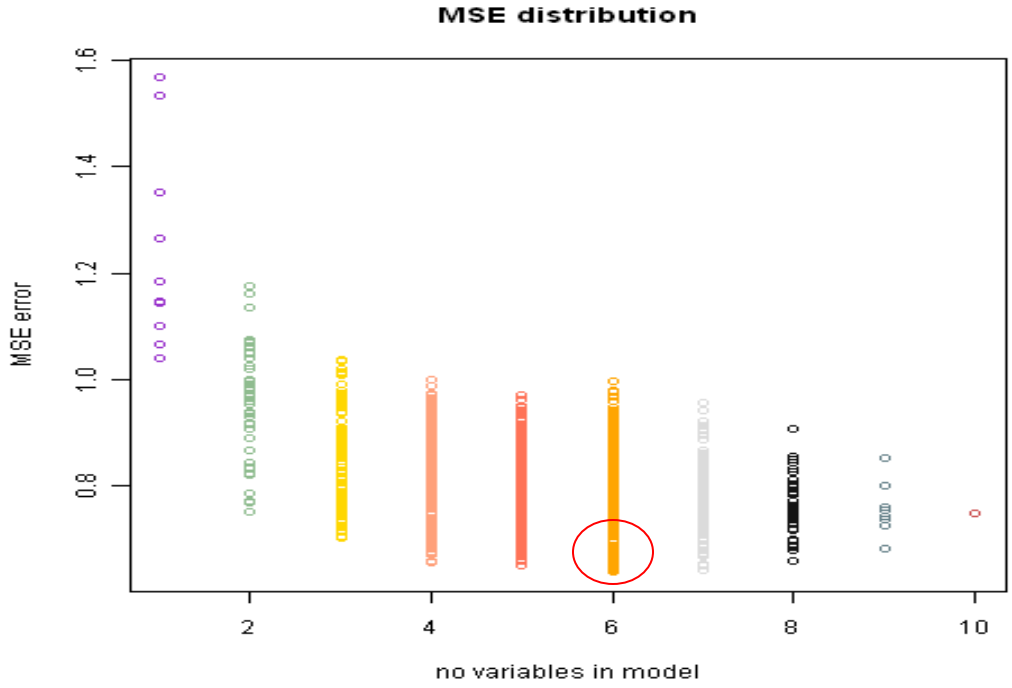
MSE = MEAN SQUARED ERROR

RF VARIABLE IMPORTANCE







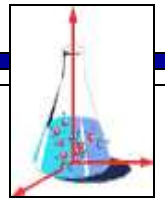


548 VARIABLES

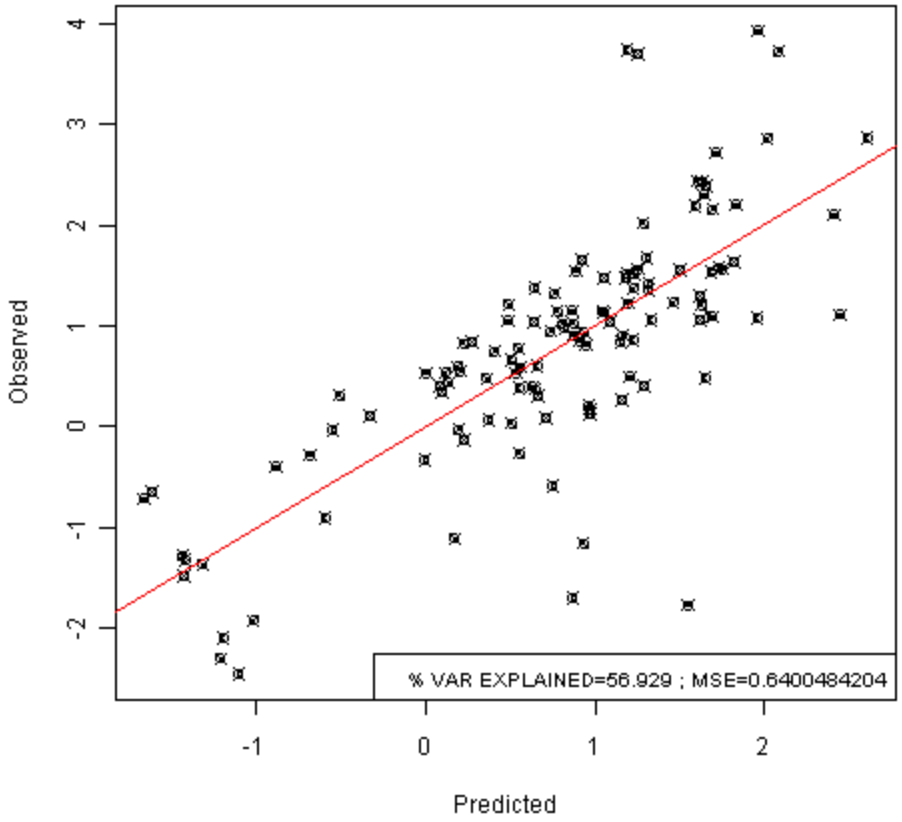
22 VARIABLES

6 VARIABLES

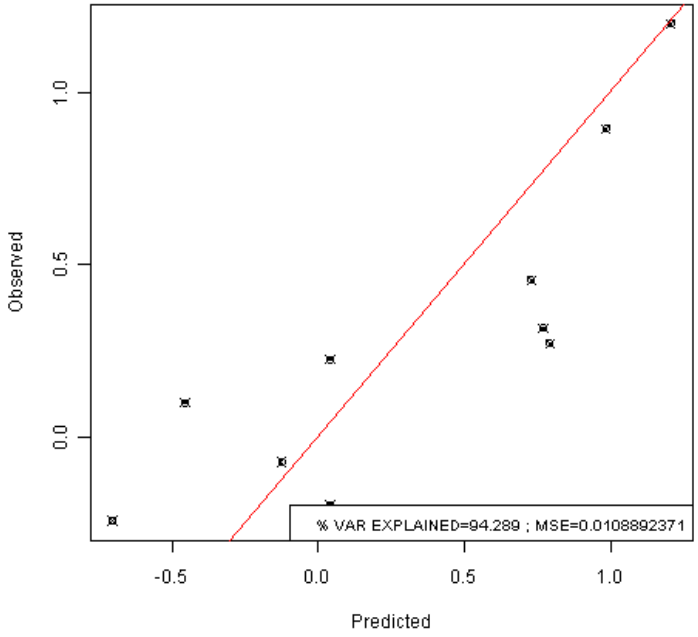
Supervised random forest: model with 12 variables



Molecular properties (MLOGP2), 2D matrix-based descriptors, 2D autocorrelations, Constitutional indices, Walk and path counts, Connectivity indices, Burden eigenvalues

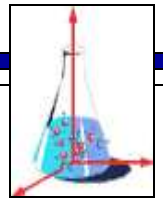


Application of the model to the test set of 10 compounds



	RMSE	MAE	% Variance explained	q squared
Values	0.1043515	0.033	94.289	77.328

Conditional variable importance in random forest regression



DATA INPUT AND PREPROCESSING;
CORRELATIONS,
NEAR TO ZERO VARIANCE

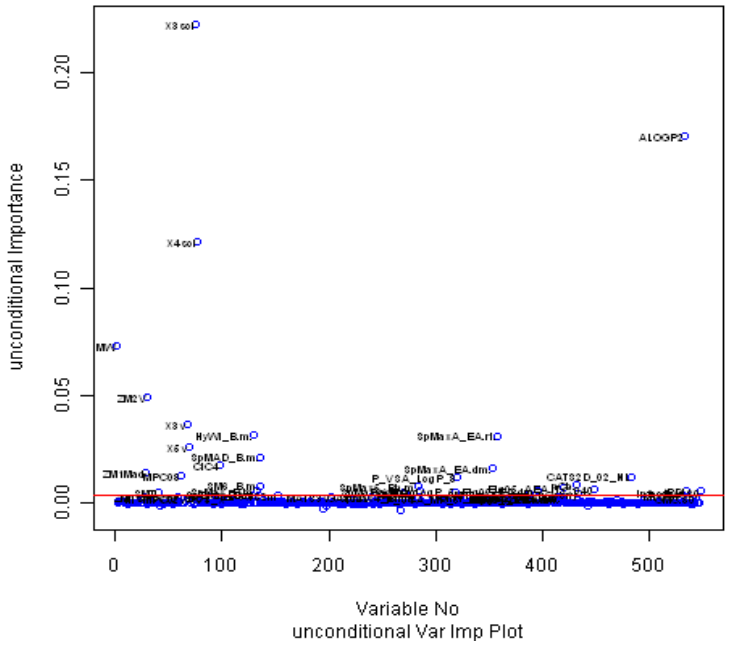
UNCONDITIONAL VARIABLE IMPORTANCE
THRESHOLD FOR CONDITIONAL VARIABLE IMPORTANCE CALCULATION

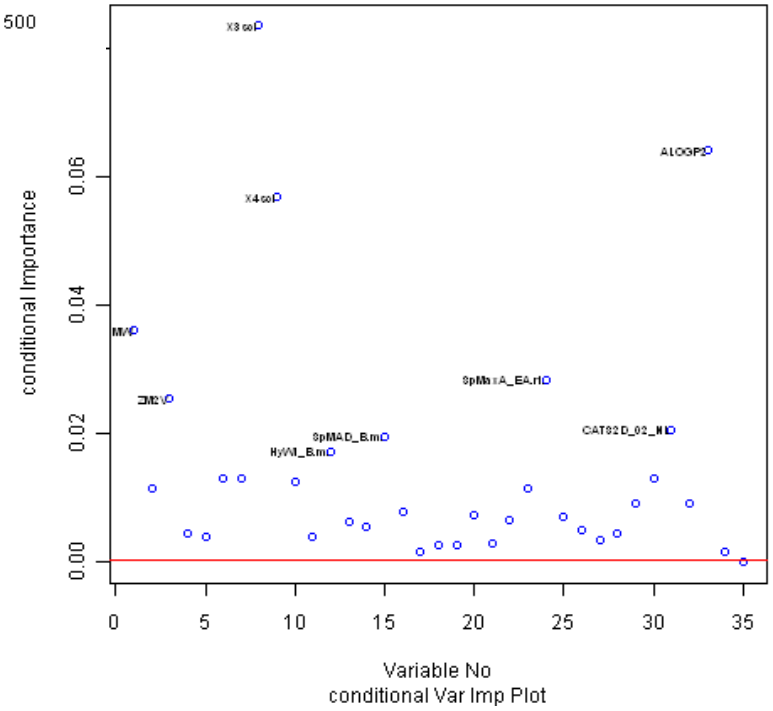
TREND AND DISTANCE ANALYSIS

FINAL SELECTION OF FEATURES AND
CONDITIONAL RF MODEL BUILD-UP

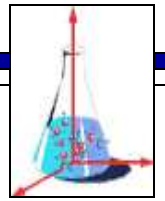
INTERPRETATION AND
FINAL DECISION

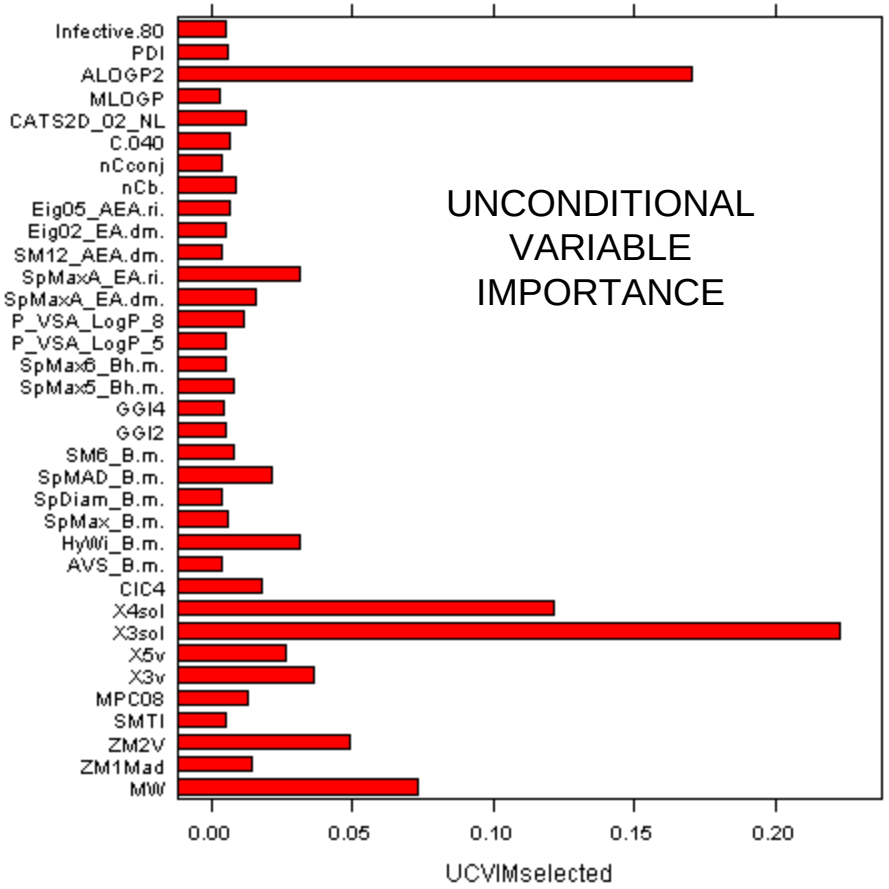
DATA OUTPUT;
CONDITIONAL RF MODEL AND STATISTICS



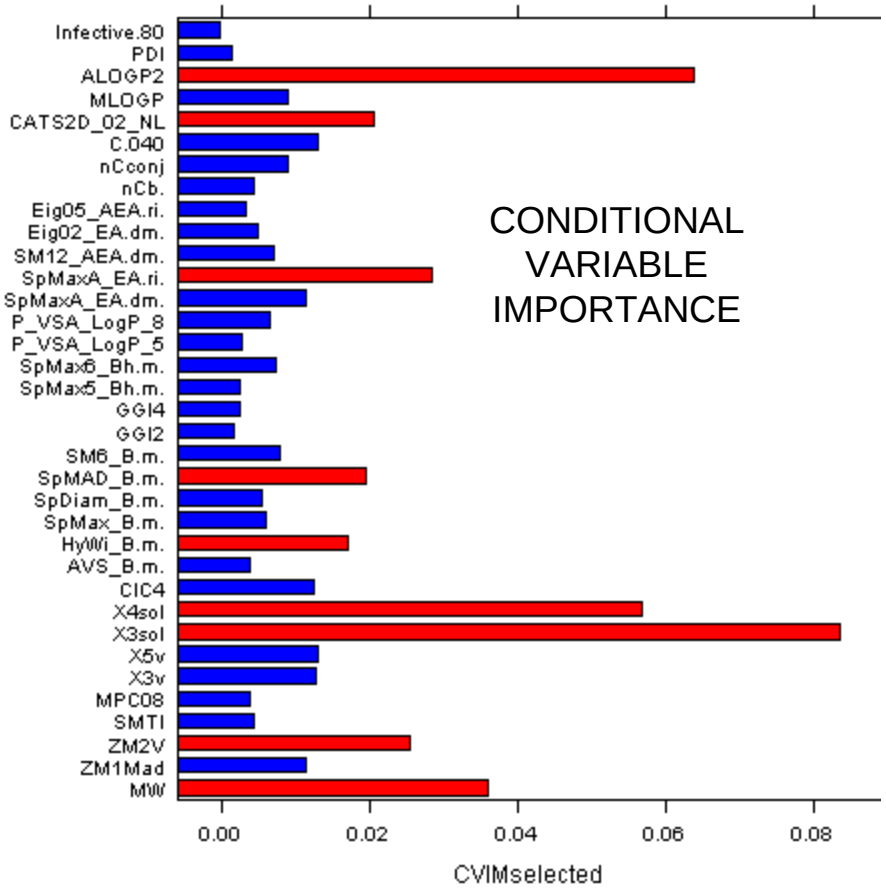


Conditional variable importance in random forest regression





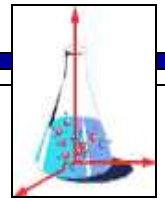
UNCONDITIONAL VARIABLE IMPORTANCE



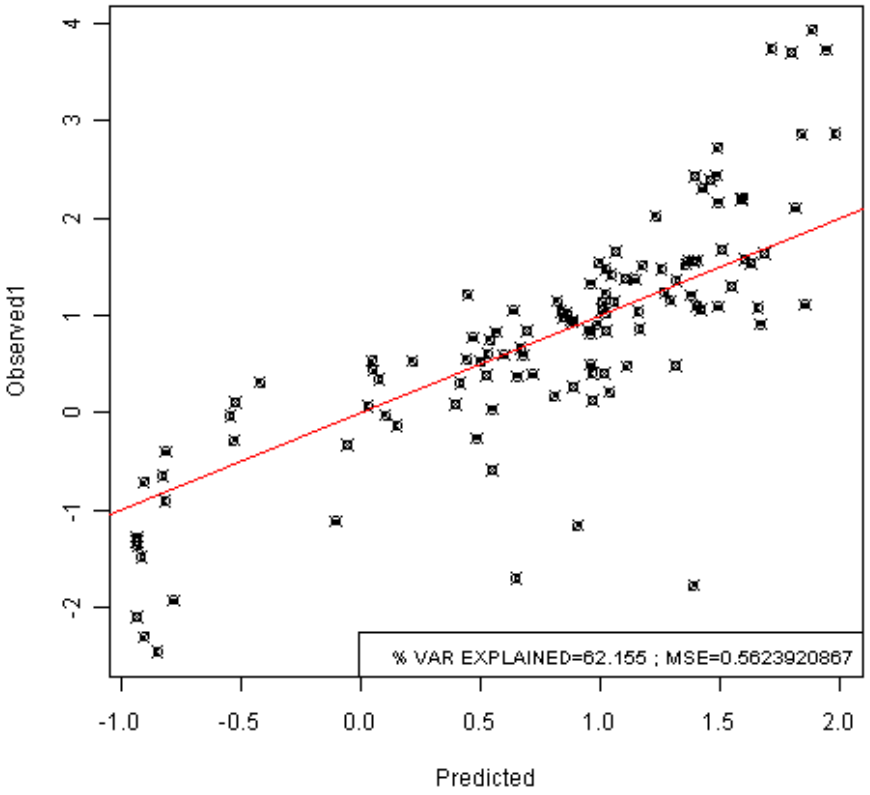
CONDITIONAL VARIABLE IMPORTANCE

548 initial variables → **35 variables selected by unconditional variable importance** → **9 variables selected by unconditional variable importance**

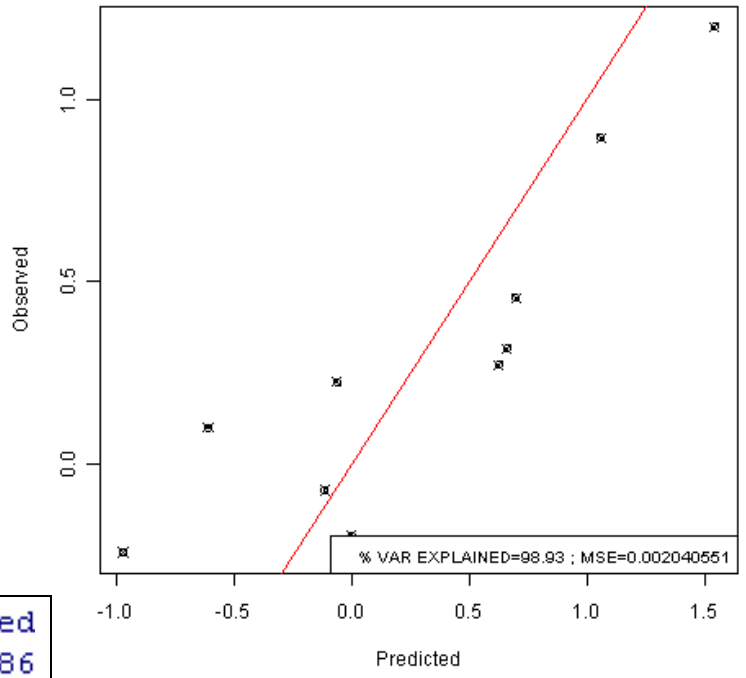
Conditional variable importance in random forest regression



Model with 9 most important indicated variables:

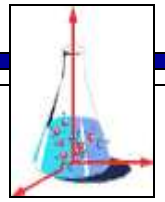


Application of the model to the test set of 10 compounds



	RMSE	MAE	% Variance explained	q squared
Values	0.04517246	-0.01428	98.93	90.186

Unsupervised random forest regression



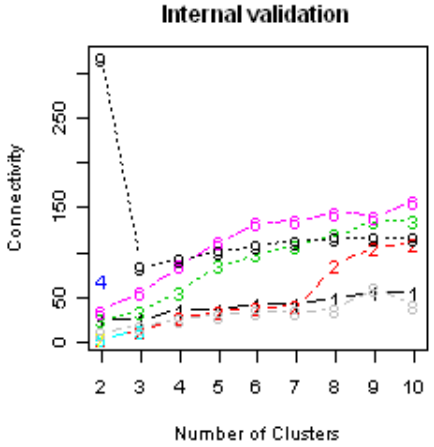
DATA INPUT
AND PREPROCESSING;
CORRELATIONS,
NEAR TO ZERO VARIANCE

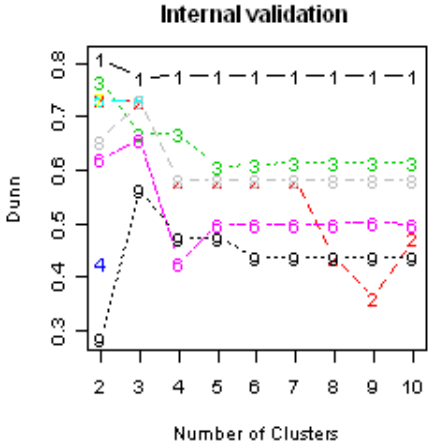
RANDOM FOREST PROXIMITY
MEASURE
AND VARIABLE IMPORTANCE

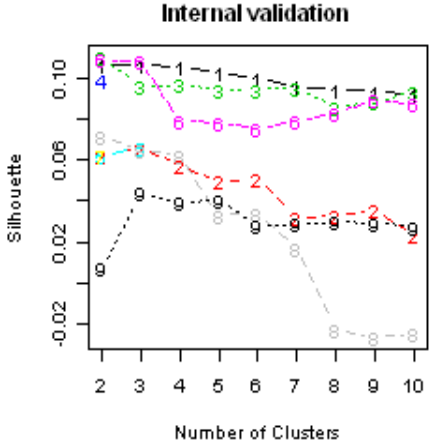
INTENRNAL VALIDATION
FOR NUMBER OF CLUSTERS
AND CLUSTERING METHOD

CLUSTER ANALYSIS TOWARDS
SELECTION OF DISTINCT,
IMPORTANT FEATURES

MODEL AND STATISTICS







Thank You for Your attention