



HelmholtzZentrum münchen
German Research Center for Environmental Health



**Marie Curie Initial Training Network
Environmental Chemoinformatics (ECO)**

Final report

**Environmental toxicity: In silico prediction accounting for bioavailability and
mechanism of action**

Early stage researcher:

Ahmed M. Abdelaziz

Project supervisor:

Igor V. Tetko

Table of Contents

1	About the project.....	4
2	Introduction.....	5
2.2	Human intestinal absorption model.....	5
2.3	QSAR model for plasma protein binding.....	8
2.4	QSAR model for hepatic clearance.....	8
2.4.1	Datasets and data handling.....	11
2.4.2	Methods.....	11
2.4.3	Discussion.....	13
2.5	AhR receptor activation model.....	21
2.5.1	In silico model building.....	21
2.5.2	Results:.....	22
2.5.3	Analysing most relevant Chemaxon descriptors:.....	23
2.5.4	Analysis of relevant fragments for AhR activation:.....	24
3	Conferences and Meetings attended.....	25
3.1.1	Chem/Bioinformatics.....	25
3.1.2	Entrepreneurship.....	26
3.1.3	DMPK/ADME.....	26
3.1.4	26
3.1.5	Trainings.....	26
3.1.6	Internship.....	27
3.2	Publications.....	27
3.2.1	Peer reviewed articles.....	27
3.2.2	Posters.....	27
3.2.3	Talks.....	28
3.2.4	Software Tools and Trainings.....	29

1 About the project

Research Institution:

HelmholtzZentrum Muenchen

Project description: The accuracy of prediction of in vivo toxicity endpoints of chemicals is expected to dramatically increase by an incorporation of information about mechanisms of action of molecules. Indeed, it is generally recognized that the toxicity of chemicals is the result of their influence on several toxicity mechanisms, which include for example the Wnt, Delta-Notch, Ras, TGF- β , and Hedgehog pathways. Therefore, the accurate use of the information about mechanism of action of molecules on these pathways (using, e.g. in vitro measurements or predicted with corresponding in silico models), and grouping of molecules according to their mode of action can increase the accuracy of in vivo toxicity predictions as well as provide a mechanistic explanation of the toxicity of chemicals. The information about the bioavailability of molecules can be also important to have better interpretation of the in vitro and in vivo correlations. This PhD will focus on the use of mechanism of action of molecules and will be complemented by work on bioavailability of molecules. The in vitro data from ToxCast™ project will be used. On the latest phase, the project will develop WWW tools for prediction of in vivo toxicity using the in vitro measurements and structural information of molecules.

2 Introduction

The European legislation on chemicals REACH (“REACH - Registration, Evaluation, Authorisation and Restriction of Chemicals,” n.d.) (Registration, Evaluation, Authorization and restriction of chemical) came into effect from 2007. The legislation assesses the risk of chemicals and aims to establish safe practices decreasing the impact of chemicals on human health, animal welfare, and the environment.

The legislation aims to collect all available information on a chemical substance to assist in identifying potential sources of hazard and further convey recommendations on risk management measures through supply chains. Responsibility for the management of substances’ risk is transmitted from the regulators to the manufacturers, importers, as well as the traders and users. This raises a huge need to provide accurate information on risk assessment to manufacturers and regulators alike.

In the course of the identification of big information gaps, the European Chemical Agency (ECHA) was established. ECHA aims to manage the databases, which are required to facilitate the information system. Additionally, it also coordinates the evaluation of suspicious chemicals and is building and managing a database for the collected hazard information, which will be kept public for consumers and professionals. The benefits from REACH would phase-in gradually as more substances get registered. REACH supports the use of animal testing only as a last resort, but encourages the justified use of well-established QSAR models, built with respect to the OECD principles, as a valid alternative.

With the evolution in the ‘omic’ approaches, the *in vitro* profiling of chemicals has been in focus over the previous years, as it appears to be offering a potential alternative to long-term *in vivo* animal testing.

WORK PACKAGE 1. The oral bioavailability of compounds is a function of absorption in gastrointestinal tract at different pH. The ChemAxon descriptors allow characterizing chemical structures at different pH and could be useful to predict oral bioavailability of chemicals. Therefore, these descriptors were implemented as part of the QSAR modeling platform (On-line CHEmical Modeling environment <http://ochem.eu>).

2.1.1.1 Results:

Chemaxon descriptors (also known as: Calculators) were implemented in OCHEM platform and can be calculated for any set of molecules. Descriptors implemented are those that return numerical or Boolean results. Unimplemented descriptors are those which return results not suitable for modeling purposes such as molecules or formula.

Also calculators that require specific input parameters are not implemented. Examples of these calculators are those that check whether certain atom is asymmetric, whether 2 atom are connected, or calculate the angle between 3 specified atoms.

The implemented descriptors are divided into 7 groups: Elemental Analysis, Charge, Geometry, Partitioning, Protonation, Isomers, and Others

A list of all implemented descriptors is available in Appendix A.


pH descriptors

Calculation of some descriptors requires consideration of pH value. User is allowed 3 options for pH:

- All: This calculates the value of the descriptor over the pH range from 0 to 14 taking 1 pH unit increments at a time. Additionally, the descriptor value at physiologic pH 7.4 is calculated.
- Specific value: calculates the value of the descriptor at the specified pH.
- Specific range: calculates the value of the descriptor over the pH range specified between "from" until "to" taking pH unit increments equal to the value specified in "step". Additionally, the descriptor value at pH 7.4 is calculated.

Descriptors, which consider pH value during their calculation, are: veragemolecularpolarizability, formalcharge, molecularpolarizability, molecularsurfacearea, polarsurfacearea, vdwsa, logd, acceptorcount, acceptorsitecount, donorcount, donorsitecount, hmopienergy, pienergy

An agreement was signed with Chemaxon SRL to allow the academic/non-commercial use of the implemented descriptor packages for scientists worldwide through the OCHEM platform. The descriptors can thus be used in model building via the URL www.ochem.eu

Chemaxon descriptors (499/3D) 

Please select pH range:

Please select descriptor groups you wish to calculate:

- Elemental Analysis
- Charge
- Geometry
- Partitioning
- Protonation
- Isomers
- Others

Please enter maximum time to allow for calculating descriptor values per mole
Time (in minutes):

Figure 1 Chemaxon in-silico descriptor package available to the scientific community as part of the OCHEM modeling framework

WORK PACKAGE 2: QSAR models for prediction of solubility and oral bioavailability of chemical compounds will be developed using data available at the OCHEM platform.

2.1.1.2 Results:

2.2 Human intestinal absorption model

Data for human intestinal absorption were based on *in vivo* permeability values measured in human subjects. Data from Zhao et al., 2002 were collected.

Chemaxon descriptors, ALOGPS and ESATE descriptors were calculated using OCHEM platform. The calculated descriptors were then used to build models using 7 different machine learning algorithms using the software Orange namely linear regression, PLS regression, kNN, SVM regression, random forest, regression tree, and Earth learner . 5-fold cross validation approach was applied. Figure shows the Orange workflow used to build the models.

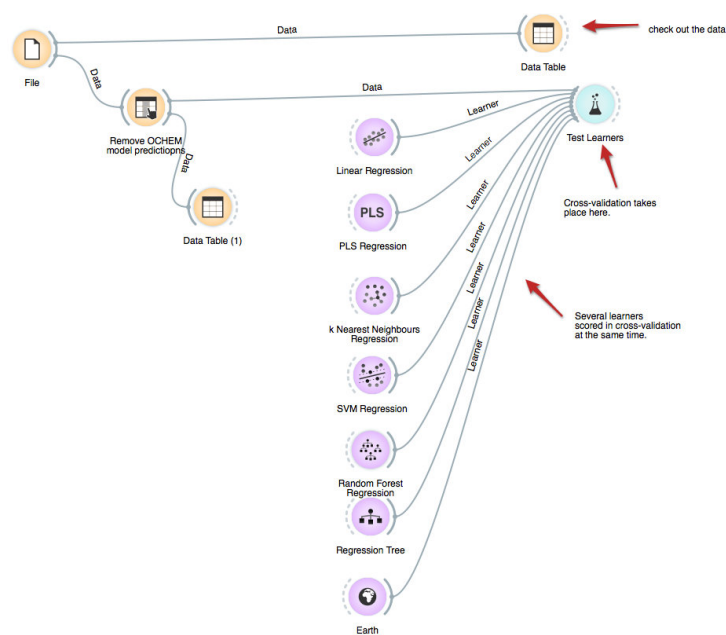


Figure 2 Workflow from orange software representing the model building process using 5-fold cross-validation approach

The 3 best performing methods according to RMSE and $cv-R^2$ were: Random Forest, PLS and kNN, the statistics of which are provided in table

	RMSE	MAE	R2
PLS Regression	23.13	17.92	0.42
kNN	26.12	18.56	0.26

Random Forest	21.56	16.04	0.50
---------------	-------	-------	------

OCHEM was also used to build models on the same data as well using different Algorithms and descriptor packages.

Models were developed using Multiple linear regression (MLRA), artificial neural networks (ANN), support vector machines (SVM), K-nearest neighbor (KNN), Fast Stagewise Multiple Linear Regression (FSMLR), and Partial Least Square (PLS). The chemaxon descriptors, which were integrated in **WP1**, were used together with other descriptor packages like AlogPS, and CDK, Dragon, OESTATE, Shape signatures, “ADRIANA.Code”, and inductive descriptors.

The models showed varied quality for the correlation between chemical structure and permeability/absorption. Below is the statistical analysis for the models that shows high quality.

Using OCHEM it is also possible to estimate the models applicability domain. The model showed a cross-validated $R^2 = 0.75 \pm 0.08$, Q^2 of 0.73 ± 0.09 (RMSE= 14 ± 1.7 and MAE= 11 ± 1.4)

A new approach was used for selecting best performing model based on consensus modeling. This lead to better predictivity and larger applicability domain estimation

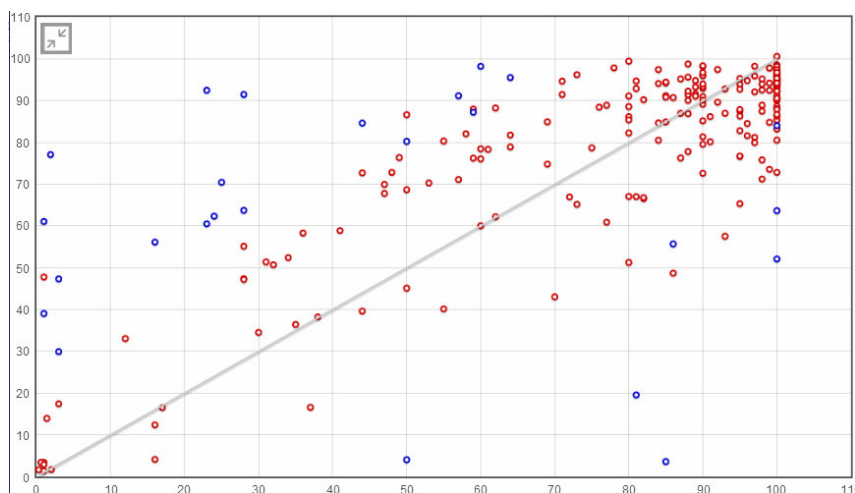


Figure 3 Regrsion line for the human intestinal absorption model showing the experimental (x-axis) vs. Predicted values (y-axis) with R^2 of 0.73

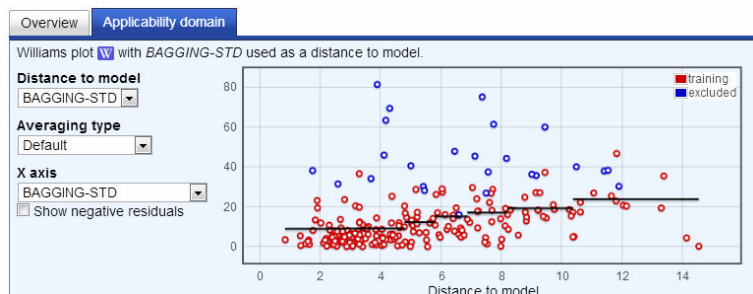


Figure 4 Williams plot representing the applicability domain off he HIA model. The distance to model is represented on the x-axis while error is represented on the y-axis

Models were developed based on literature data published in the following articles:

Articles	Compound s
MDCK (Madin-Darby canine kidney) cells: A tool for membrane permeability screening (Irvine et al., 1999).	55
Molecular hashkeys: a novel method for molecular characterization and its application for predicting important pharmaceutical properties of molecules (Ghuloum, Sage, & Jain, 1999).	20
In silico ADME modeling 3: Computational models to predict human intestinal absorption using sphere exclusion and kNN QSAR methods(Gunturi & Narayanan, 2007).	174
CODES/neural network model: A useful tool for in silico prediction of oral absorption and blood-brain barrier permeability of structurally diverse drugs (Dorransoro et al., 2004).	28
Physicochemical high throughput screening: parallel artificial membrane permeation assay in the description of passive absorption processes (Kansy, Senner, & Gubernator, 1998).	25
ADME evaluation. 2. A computer model for the prediction of intestinal absorption in humans (Klopman, Stefan, & Saiakhov, 2002).	49
Toward minimalistic modeling of oral drug absorption (Oprea & Gottfries, 1999).	85
Experimental and computational screening models for the prediction	20

of intestinal drug absorption (Bergström, Norinder, Luthman, & Artursson, 2002).

Functional role of P-glycoprotein in limiting intestinal absorption of drugs: contribution of passive permeability to P-glycoprotein mediated efflux transport (Varma, Sateesh, & Panchagnula, 2005).	88
Prediction of human intestinal absorption of drug compounds from molecular structure (Wessel, Jurs, Tolan, & Muskal, 1998).	86
Rate-limited steps of human oral absorption and QSAR studies (Zhao et al., 2002).	237
Drug liposome partitioning as a tool for the prediction of human passive intestinal absorption (Balon, Riebesehl, & Müller, 1999).	21

2.3 QSAR model for plasma protein binding

Many models were built to predict the percentage of drug that is free in plasma using 2D descriptors. Models were trained with experimental data gathered from the literature for 112 well-characterized drugs. 3D models of this endpoint did not offer significantly better performance. The model was tested with literature data from a set of 66 drugs that were not applied to the training process in any way.

The 2 best performing models were:

1- A support vector machine model built using Chemaxon descriptors with $R^2 = Q^2 = 0.94$, RMSE=0.22 and MAE=0.13 (Figure 5)

2- Associative neural network model built using AlogPS and OESTATE descriptors yielded the following statistics $R^2 = 0.83$, Q^2 of 0.82 (RMSE=13.32 and MAE=10.74) (Figure 6)

Both models were built using the same dataset with 5-fold cross validation

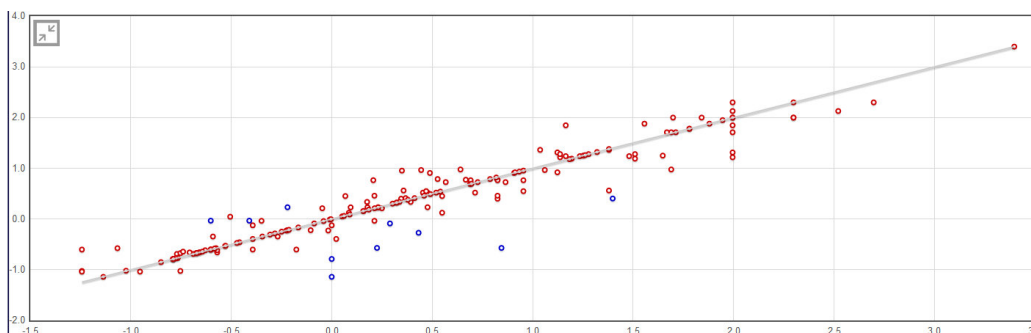


Figure 5 measured vs predicted (by svm model) Plasma protein binding in logit units

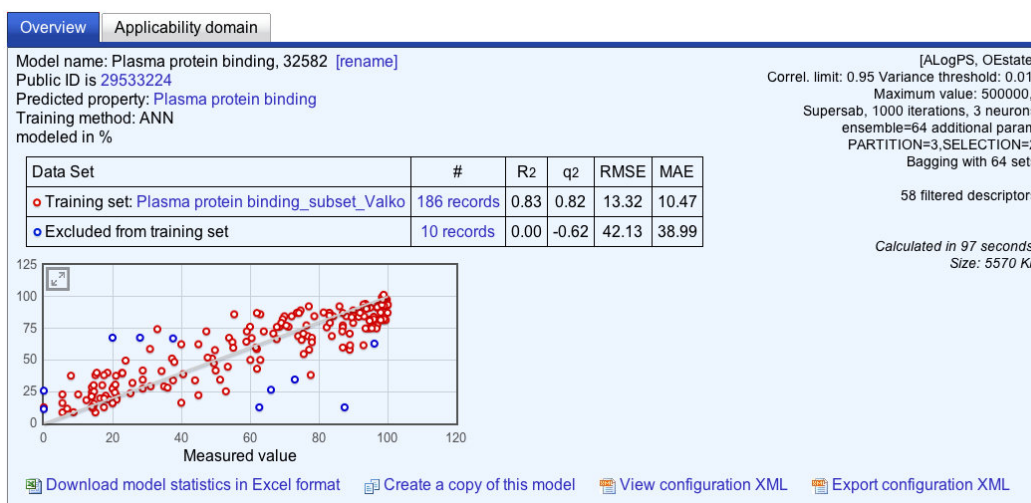


Figure 6 Statistics for the plasma protein binding model built using ALOGPS and OESTATE descriptors and ASNN.

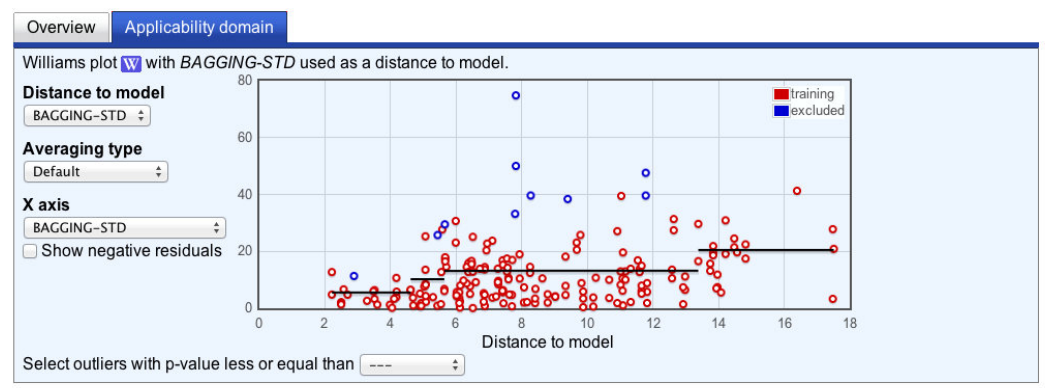


Figure 7 applicability domain for the AHR activators classification model. Plot shows % of compounds (x-axis) against balanced accuracy (y-axis)

2.4 QSAR model for hepatic clearance

A model for human excretion has $R^2 = 0.51$, Q^2 of 0.50 (RMSE=14.73 and MAE=10.73)

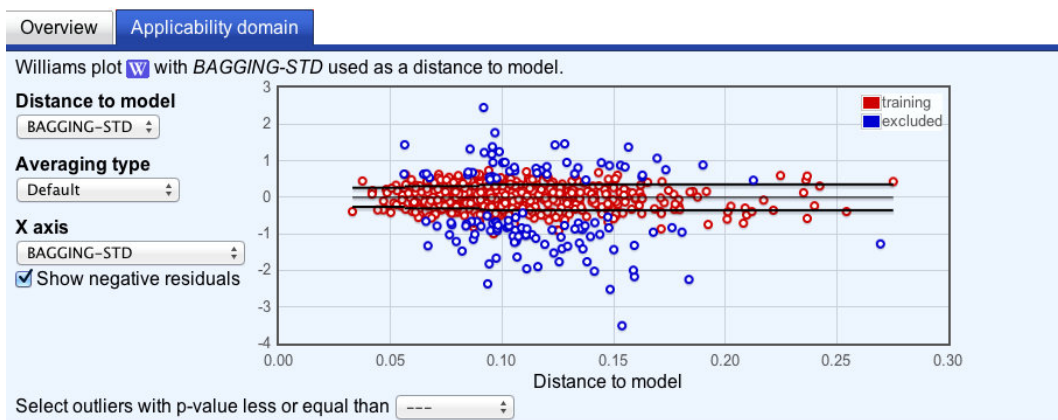


Figure 8 Applicability domain for the hepatic clearance model.

Work packages 3: Integration of data available from TOXCAST project to be used within the project. The data will be uploaded from PubChem/ACToR database.

Results:

ACToR is EPA's online warehouse of all publicly available chemical toxicity data and can be used to find all publicly available data about potential chemical risks to human health and the environment. ACToR aggregates data from over 500 public sources on over 500,000 environmental chemicals searchable by chemical name, other identifiers and by chemical structure.

The data warehouse:

- Allows users to search and query data from other EPA chemical toxicity databases including:
 - ToxRefDB (30 years and \$2 billion worth of animal toxicity studies).
 - ToxCastDB (data from screening 1,000 chemicals in over 500 high-throughput assays).
 - DSSTox (provides high quality chemical structures and annotations).

A framework was built using Knime (Figure 9) for introducing the data into our modeling environment's (OCHEM) database. This framework was then used to transfer all data from the databases to OCHEM. The use of such workflow allows continuous update of the OCHEM database with new data released from EPA as long as the 3 databases maintain their structure. The data from ACToR system have been uploaded in a test version of our database and has been analyzed with respect to their mapping to Properties and Conditions existing in our database and their incorporation in our main database.

Naturally, some of the data published through the ACTOR system of databases about the chemicals are not related to toxicity and are thus irrelevant to the current work. Such data was filtered out and will not be included in the final release. Such workflow is useful for the future upload (e.g: ToxCast Phase II data; expected for release in late 2013). It also permits the flexibility of filtering data to be passed through to the modeling framework OCHEM. This framework will be the base for testing the effect of bioavailability and biochemical pathways on quality of in vitro to in vivo correlations.

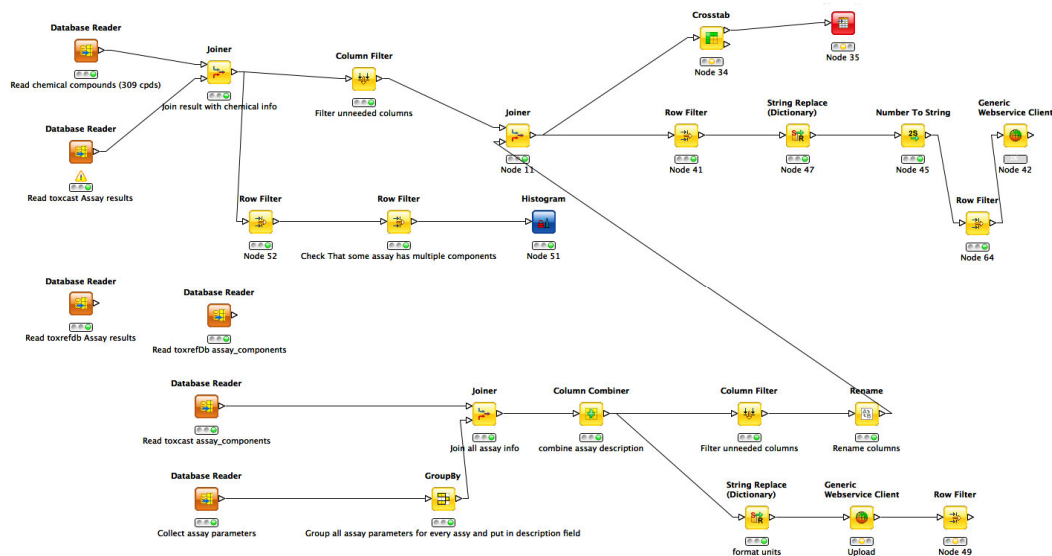


Figure 9 graphical representation of the Knime workflow used for semi-automatic data upload from ToxCast database into the OCHEM platform.

The Knime workflow has been validated for data upload. Data from ToxCast and ToxrefDb databases were uploaded to a new instance of OCHEM modeling platform (<http://iprior.eadmet.com>). All records were successfully introduced as expected (207270 records for ToxCast and 111440 records of ToxrefDb). The process lasted around 36 hours. Uploaded data is available for scientists through our system as part of Work Package 7. Knime was also used to upload the chemical structures of 309 chemicals of the Phase I ToxCast project.

Milestone 1: QSAR models were developed for classification of chemicals. Analysis and categorization of molecules from ToxCast project according to their solubility and permeability.

WORK PACKAGES 4, 5 and 6: The U.S. Environmental Protection Agency has established a number of programs for *in vitro* bioactivity profiling, including ToxCast, (Dix et al., 2007; Judson et al., 2010) Tox21, (Betts, 2013; Raymond Tice Robert Kavlock, Ph.D., and Christopher Austin, M.D., Tice, Kavlock, & Austin, n.d.) and EDSP21. ToxCast is the largest in terms of number of *in vitro* assays. It covers more than 450 assays. In its Phase I, the program covered 309 chemicals (mostly food pesticides for which thorough animal toxicity studies are available).

Multiple previous studies evaluated the ability of the *in vitro* assays for predicting selected *in vivo* endpoints (Kleinstreuer et al., 2011; M T Martin et al., 2011; Shah et al., 2011) and analyzed the biochemical pathways that could be involved with the

observed toxicity. Most studies focused on a single *in vivo* toxicity endpoint. A comprehensive analysis of the *in vitro*-to-*in vivo* predictive capability of the ToxCast high-throughput screening effort has also been independently presented.(Thomas et al., 2012)

In this study we assess the predictive ability of the HTS *in vitro* assays in constructing a toxicity signature. We aimed to provide an exhaustive overview of the provided data and its effectiveness and limitations within QSAR studies. We also investigate different *in silico* descriptor packages regarding their ability to represent the *in vitro* assays. We also investigate to what extent can *in silico* descriptors represent the information in the *in vitro* assays by building *in silico* QSAR models for the prediction of *in vitro* assays output.

We finally identified 6 *in vivo* endpoints, which appear to be predictive with balanced accuracy of more than 0.65 (at 95% confidence interval). Furthermore, we found a number of five assays for which a high balanced accuracy (0.75) was achievable by *in silico* descriptors, which enables an improved approach towards *in silico* modeling towards toxicity in general.



Figure 10. The aim of using *in vitro* profiling of chemicals in combination with knowledge about pathways of toxicity, bioavailability of compounds, as well as machine learning algorithms to reduce long term animal toxicity studies.

2.4.1 Datasets and data handling

In vitro assays

Toxminer v17 was downloaded from ToxCast EPA website as SQL dump script, together with its enhanced entity relationship (EER) diagram. The data were rebuilt into a local MySQL database before being imported into iPRIOR (“iPrior -

Prioritization and estimation of toxicity of chemical compounds,” n.d.), using Knime (see software section for details).

The database included information on biochemical pathways, processes, assay-gene, and gene-pathway mappings. Correlations between genes and pathways were collected from Gene Ontology (GO), (“Gene Ontology Documentation,” n.d.) Kyoto Encyclopedia of Genes and Genomes (KEGG), (“KEGG: Kyoto Encyclopedia of Genes and Genomes,” n.d.) Ingenuity Pathways analysis (IPA, Ingenuity systems Inc, Redwood city, CA), (“Ingenuity IPA - Integrate and understand complex ’omics data,” n.d.) pathway commons, (“Pathway Commons,” n.d.) and the OMIM(Boyadjiev & Jabs, 2000) phenotype databases.

The extracted data included the chemical structure files (sdf) for all 309 compounds in the database. The *in vitro* information consisted of 467 assays, some of which evaluates multiple time points, resulting in 669 assay endpoints. It is worth mentioning that the response of the ToxCast phase I chemicals (309 compounds) varies significantly across different *in vitro* assay categories. Figure 1 shows the response of the ToxCast phase I chemicals to the 669 endpoints measured. The assays cover nine technologies: cell-free HTS assays; multiplexed transcription reporter; biologically multiplexed activity profiling; high-content cell imaging; multiplexed gene expression; cell-based HTS; phase I and II XME cytotoxicity; real-time cell electronic sensing; and HTS genotoxicity. The assays measure both direct interactions between chemicals and identified receptors and enzymes, as well as downstream events on receptor gene activity or cellular consequence.

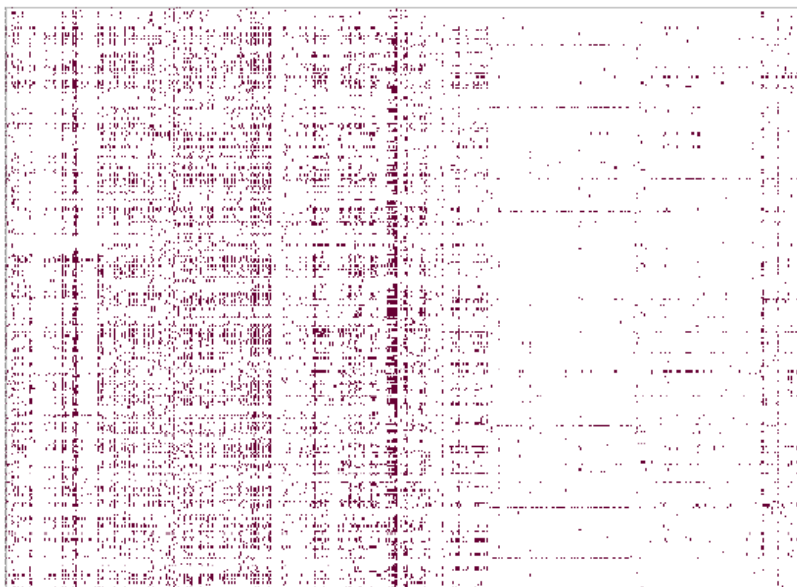


Figure 11. Heat map of 669 assay endpoint measurements (including multiple time points where available) in the ToxCast phase I data set. The assays are arranged from the left to the right, and chemicals are arranged top to bottom. The data values were discretized for the analysis, thus higher or lower values of AC50/LEC are not differentiable.

EPA database reported a half maximum activity concentration (AC_{50}) or lowest effective concentration (LEC) for assay responses. However, due to the comparably low accuracy associated with HTS settings, under which these experiments were conducted, we focused to calculate classification models. If such models deliver reasonable results, more detailed regression models could be interesting for a further exploration of the underlying endpoints. Therefore, all assay results were discretized into (response/no response) values.

At a rough estimate, only 7% of the assay/chemical interaction matrix showed any response. Another approach we considered, in terms of data consolidation, is to analyze the liability of a chemical to cause a perturbation in a given pathway, regardless of which gene it affects to cause such perturbation.

In vivo animal studies

The ToxMiner v17 included subset of the toxicity reference database (ToxRefDB) that is relevant for the chemicals of the ToxCast study. The database had results for 461 animal studies conducted. Again, all results were discretized to whether a chemical exerts the toxicity of question or not. The score of a chemical or toxicity was considered irrelevant for the subsequent analysis.

2.4.2 Methods

In Silico Descriptors

iPrior web platform("iPrior - Prioritization and estimation of toxicity of chemical compounds," n.d.) was used to calculate *in silico* descriptors from eight different commercial and academic providers. These considered packages are GSfrag(Aires-de-Sousa & Gasteiger, 2001), ISIDA fragments(Varnek et al., 2008), Chemaxon descriptors("Calculator Plugins «ChemAxon – cheminformatics platforms and desktop applications," n.d.), Estate indices(Hall, Kier, & Brown, 1995) & AlogPS(I V Tetko, Tanchuk, Kasheva, & Villa, 2001; Igor V Tetko, Tanchuk, & Villa, 2001), CDK(Steinbeck et al., 2003), inductive descriptors(Cherkasov et al., 2008), Dragon 6(Todeschini & Consonni, 2009), Adriana.Code("ADRIANA.Code - Calculation of Molecular Descriptors | Inspiring Chemical Discovery," n.d.). The descriptor values are available in the supplementary materials.

To calculate chemical-pathway perturbations, 1456 pathways were correlated to 299 chemical structures. We considered the correlation of pathways to their respective genes then investigated whether a compound had a positive hit to any assay associated with these genes. If a chemical shows activity in any assay associated with these genes then it was considered perturbing the investigated pathway. Subsequently we built of a chemical/pathway-perturbation matrix that showed that 14% of potential interactions were positive.

The iPrior online platform (“iPrior - Prioritization and estimation of toxicity of chemical compounds,” n.d.), containing an implementation of the Chemaxon Standardizer, was used for the preprocessing of chemical compounds. The standardization process included a salt counter-ion removal, charge neutralization, and recalculation of 3D structures, using CORINA (Sadowski, Gasteiger, & Klebe, 1994), and the standardization of nitro groups and aromatic ring representations. The iPrior implementation additionally was used to calculate descriptor values for the packages listed above.

In total 9691 descriptors were derived for 299 compounds. The calculation for ten structures failed, as these compounds were inorganics, organometalics, mixtures or large macrocyclic compounds. A list of the disregarded molecules is available in the supplementary materials. Furthermore, descriptors with a variance < 0.01 were removed, which resulted in a total number of 6318 relevant descriptors.

Modeling in vivo animal toxicity

Models were built using 9 different machine learning algorithms provided by Knime (“KNIME | KNIME Desktop,” n.d.). The used approaches consisted of: Probabilistic Neural Network (PNN); Support vector machines (LIBSVM v2.89); multilayer feedforward networks (RPROP); a decision tree learner; the k-nearest neighbor approach; Random forests; and three algorithms based on WEKA (Holmes, Donkin, & Witten, 1994) v3.6: J48 (Java implementation of C4.5 decision tree); LADTree; and REPTree. All models were built based on a 5-fold cross-validation. Supplementary materials include the parameters for all used algorithms.

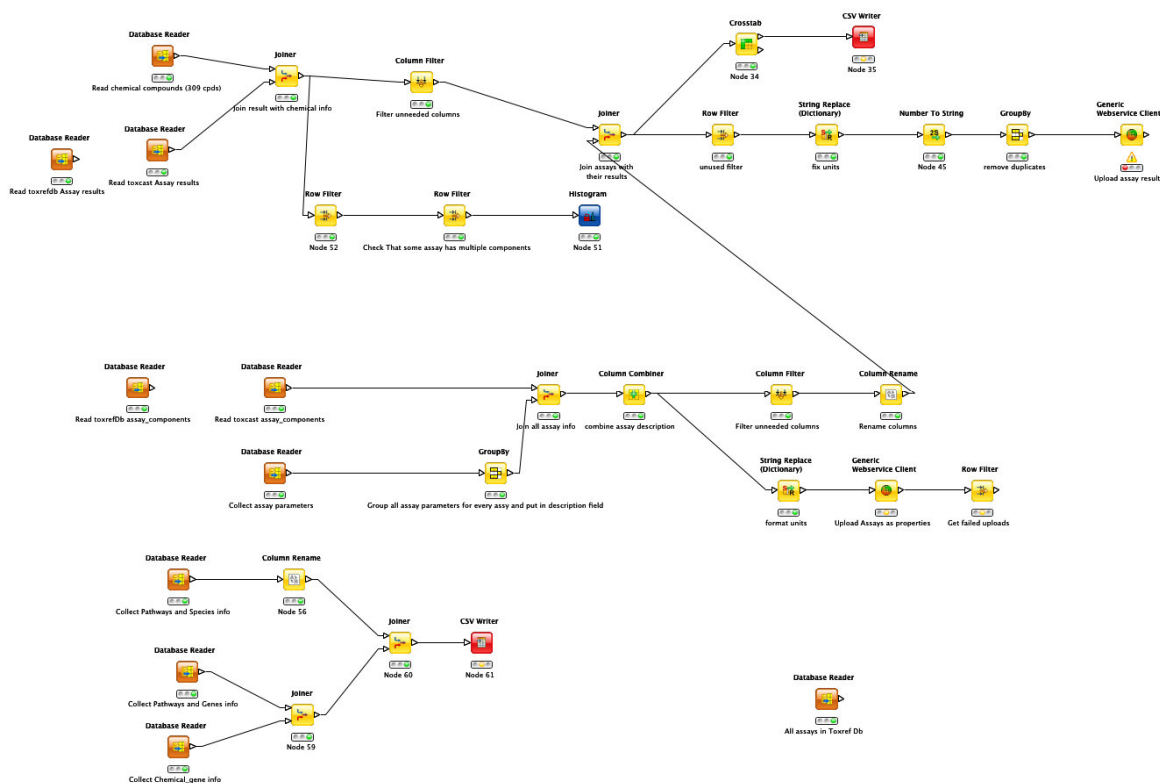


Figure 12 Knime workflow used to build QSAR models for *in vitro* to *in vivo* correlation using Phase I ToxCast data

As descriptors with low variance are likely to degrade the performance of certain learning algorithms (in particular those which are distance based), all *in vitro* assays that have low hit ratio, 3 or less compounds (i.e. <1%), were removed from the list of relevant biological descriptors prior to modeling (as done with the *in silico* descriptors). The list of all *in vitro* assays with the number of their hit compounds is listed in the supplementary materials.

Descriptors used for the calculation of the models were: *in silico* descriptors from the eight descriptor packages separately; discretized *in vitro* assays separately; discretized pathway correlations separately; and a combination of all descriptors.

Out of the 461 available animal studies, only 61 showed toxicity for 35 or more chemicals, which was used as a tentative threshold for conducting proper 5-fold cross validation. For every endpoint the total number of tested compounds was between 234-251. Animal toxicity studies were conducted in rats, rabbits, and mice. For each study only one animal species was used. The *in vivo* toxicology assays are those derived from ToxRefDB. The majority of data currently in the ToxRefDB database, a component of the larger ACToR system, contains summary results of primary toxicology studies submitted to the EPA on pesticide active ingredients (M T Martin, Houck, McLaurin, Richard, & Dix, 2007). Typically these data have been extracted

from EPA Office of Pesticide Programs (OPP) evaluations of studies, based on EPA Office of Prevention, Pesticides and Toxic Substances (OPPTS) harmonized test guidelines. Full details of the collected data has been described in literature(Matthew T Martin, Judson, Reif, Kavlock, & Dix, 2009).

The Toxicity Reference Database (ToxRefDB) has been the primary tool for storing and accessing high-quality toxicology studies and is available online for search and download(US EPA, n.d.). ToxRefDB has characterized thousands of studies using a standardized vocabulary, a uniform structure across study types, and a high level of internal and external quality control (QC) for the extraction of endpoints useful in developing predictive models(Matthew T Martin et al., 2009). Full list of *in vivo* toxicology assays are available in the supplementary materials.

In total 6039 models were built on 61 endpoints with nine machine learning approaches applied to eleven feature combinations. The selected *in vivo* animal toxicities for modeling together with their respective number of toxic compounds are listed in the supplementary materials.

Modeling *in vitro* assays

An interesting exploration was to figure out to which extent could *in silico* descriptors represent the information represented in the *in vitro* assays. To investigate this, we evaluated approaches to model the *in vitro* assays using different *in silico* packages.

Most of the *in vitro* assays show activity for only few compounds or even none at all. Therefore they cannot be modeled with the available data. From the available 669 *in vitro* assay endpoints, only 148 contained 35 or more active hits with the tested compounds. For these endpoints, all 299 concerned compounds were used to build QSAR models. List of all *in vitro* assays with their description is available in the supplementary materials.

The same learning algorithms were used as with the *in vivo* animal experiment modeling. The *in silico* descriptors from the eight descriptor packages, which were used to model each assay, were applied separately or all combined in one descriptor set. In total 11988 models were built on 148 endpoints applying nine machine learning approaches to nine feature combinations).

2.4.3 Discussion

Modeling *in vivo* animal toxicity

Fig. 2 shows the balanced accuracy of all 6039 models built. Different statistical parameters for all the models including: sensitivity; specificity; balanced accuracy; as well as Matthews correlation coefficient (MCC) are included in the supplementary materials.

Table 1. The five best and the four worst predicted in vitro assays based on the median balanced accuracy of the respective models.

	Mean BA	Median BA	Max. BA	Min. BA
CHR_Rat_CholinesteraseInhibition	0.74	0.77	0.92	0.5
MGR_Rat_Liver	0.6	0.61	0.7	0.49
CHR_Rat_Liver_1_AnyLesion	0.58	0.58	0.66	0.5
DEV_rat_Developmental_Skeletal_Axial	0.55	0.56	0.65	0.46
DEV_rat_Developmental_Skeletal	0.55	0.55	0.67	0.45
MGR_Rat_LitterSize	0.5	0.5	0.56	0.46
MGR_Rat_Ovary	0.5	0.5	0.56	0.46
CHR_Rat_Tumorigen	0.5	0.5	0.56	0.42
MGR_Rat_Testis	0.5	0.5	0.61	0.45

The mean, median, maximum as well as minimum balanced accuracy of the five best and the four worst predicted animal toxicity studies are provided in table 1. The ranking was based on the median balanced accuracy of their respective models.

Table 2 The All toxicity endpoints that have a significant probability for getting a balanced accuracy of 0.65 for their best model together with their respective probabilities

Toxicity end point	Probability of the best model having BA > 0.65
CHR_Rat_CholinesteraseInhibition	1
Multigeneration Rat Endpoint for Liver microscopic and gross pathologies and weight changes	0.9997
Multigeneration Rat Endpoint for Kidney microscopic and gross pathologies and weight change	0.9990
Developmental rat Developmental Skeletal	0.9876
Developmental rabbit Developmental Skeletal	0.9827
CHR_Rat_Liver_1_AnyLesion	0.9806

To better understand the reason behind the success in modeling of cholinesterase inhibition in rats, the “Set Compare” tool from iPrior was used. The 2 sets of compounds (toxic vs. non-toxic) were compared using ISIDA fragmental descriptors. All toxic compounds had common phosphorus-containing scaffolds. Indeed, Cholinesterase inhibition is the main mechanism of action by which phosphorus insecticides perform their function. 32 of the 45 toxic compounds for this endpoint were organophosphorus compounds. Only one non-toxic compound was a phosphorus derivative. This simple scaffold is easy to capture for any descriptor package that accounts for fragments or atom counts while becomes harder for in vitro assays to

indirectly capture the presence of that scaffold. **Table 3** shows the most common scaffolds and their respective p-value.

Table 3 Most common ISIDA fragments in the toxic cholinesterase inhibitors showing clear indication of organphosphorus compounds

Descriptor	# in toxic set (45 molecules)	# in non-toxic set (212 molecules)	p-value
SdsssP	30	1	8.94E-28
Se1O2P4sd	28	1	2.17E-25
Se2P4S1s	16	1	4.53E-13
Se2O1P4s	14	0	2.13E-12

It was possible to obtain, at least, one model for each end point that shows statistically significant prediction higher than random guess balanced accuracy (0.5) with 99.5% confidence interval. For the chronic rat cholinesterase inhibition, the predictive accuracy reached more than 90% for some models. The provided statistics reveal that only six end points has any models that exceeded a balanced accuracy of 0.65 (with 95% balanced accuracy). **Table 5** shows the performance of different descriptor packages regarding the median confidence interval of having a significant model. While ISIDA descriptors showed best values, the *in vitro* assays and pathway correlations performed worst. The use of *in vitro* assays, as biological descriptors, did not result in significantly better results than the use of *in silico* descriptors. Neither the combination of both had significant improvement to the prediction quality. The supplementary materials show the average performance of models built using different descriptor combinations.

Table 4. Performance of different algorithms in constructing QSAR models for both *in vitro* assay endpoints and *in vivo* toxicity. For each algorithm, the median probability of building a model that shows statistically significant prediction higher than random guess balanced accuracy (0.5) is shown. All algorithms were challenged by the same assay endpoints and descriptors.

Algorithm	Median Probability (in <i>in vivo</i> toxicity models)	Median Probability (in <i>in vitro</i> assays modeling)
KNN-5cv	0.822	0.998
RProp-MLP-5cv	0.797	0.986
J48-5cv	0.781	0.990
LADTree-5cv	0.760	0.991
RandomForest-5cv	0.756	0.987
PNN-5cv	0.756	0.941
Decision-Tree-5cv	0.712	0.976
LIBSVM-5cv	0.683	0.965
REPTree-5cv	0.667	0.922

Table 5. Performance of different descriptor packages in constructing QSAR models for both in vitro assay endpoints and in vivo toxicity. For each descriptor package, the median probability of building a model that shows statistically significant prediction higher than random guess balanced accuracy (0.5) is shown. Compared models were built for the same assay endpoints and using the same algorithms.

Descriptrs	Median Probability (in vivo)	Median Probability (in vitro)
ISIDA Fragments	0.841	0.989
All-Combined	0.829	0.990
CDK	0.816	0.990
Estate-AlogPS	0.801	0.990
Dragon6	0.794	0.987
GSFrag	0.756	0.973
Adriana-CODE	0.754	0.984
Inductive Descriptors	0.744	0.962
Chemaxon Descriptor	0.742	0.988
Pathways correlations	0.624	-
Toxcast in vitro assays	0.597	-

Table 4 compares different algorithms on their performance, the observation that a comparably simple method, such as the kNN approach shows the best average performance, compared to non-linear high resolution methods, such as support vector machines or neural networks indicates that the information contained in the majority of in vivo experiments provided in the data is not informative by classical QSAR modeling techniques.

When comparing different descriptor packages, some were better than others in capturing toxicity events. **Table 5** shows that ISIDA fragments performed best, while the use of in vitro assays was considerably less successful. It is also possible to notice that the use of pathway correlations slightly improved the probability to obtain a predictive model (higher than random guess balanced accuracy of 0.5).

Modeling in vitro assays

Fig. 3 shows the balanced accuracy of all 11988 models built to predict the *in vitro* assays. As with the animal toxicity modeling, different statistical parameters are reported in the supplementary materials. In comparison to the prediction of *in vivo* experiments a significant improvement in accuracy is observable. For numerous assays, such as those listed in **Table 6**, we were able to build predictive models exceeding a balanced accuracy of 0.75. The ten best predicted *in vitro* assays based on the median balanced accuracy of their respective 81 models are listed in table 3. Considering the best models built for each endpoint, almost all (147 out of 148) *in vitro* assays showed an ability to build *in silico* models that are statistically significant than random guess balanced accuracy (0.5) (95% confidence level). 68 endpoints showed balanced accuracy of >0.6 while for four assays it was possible to get a

balanced accuracy of 0.75 at that confidence level. These assays are related to change in expression of different isoforms of the liver metabolizing enzymes CYP450, namely (2B6 in humans, 2B1 and 2C11 in rats) as well as the cell loss count after 72 hours. This also agrees with previous in silico studies that reported success in building in silico QSAR models for prediction of CYP450 expression change of different isoforms (Lewis, Modi, & Dickins, 2002; Novotarskyi, Sushko, Korner, Pandey, & Tetko, 2011; Roy & Roy, 2009).

Table 6 shows the best predicted in vitro assay endpoints and their confidence intervals for a balanced accuracy of 0.75

Assay name	Probability of a balanced accuracy >0.75
CellzDirect CYP2B6 24hr	0.993
Novascreen Rat CYP2B1	0.986
Cellumen Cell Number 72hr	0.960
Novascreen Human CYP2B6	0.953
Novascreen Rat CYP2C11	0.947

Table 7. Performance of the ten best predicted in vitro assays, based on the median balanced accuracy performance of their respective 81 models.

	Mean	Median	Maximum	Minimum
CLM_CellLoss_72hr	0.71	0.71	0.79	0.61
CLZD_CYP3A4_48	0.69	0.70	0.77	0.52
CLZD_CYP2B6_24	0.69	0.70	0.80	0.51
BSK_SAg_Proliferation_down	0.69	0.69	0.76	0.57
NVS_ADME_rCYP2B1	0.68	0.69	0.84	0.50
BSK_3C_Proliferation_down	0.68	0.69	0.78	0.50
NVS_ADME_rCYP2C11	0.66	0.68	0.80	0.49
BSK_hDFCGF_Proliferation_down	0.67	0.68	0.75	0.50
CLZD_CYP2B6_6	0.66	0.67	0.74	0.50
NVS_ADME_hCYP3A5	0.65	0.66	0.78	0.50

Regarding the descriptor collections, the observations were different from the case with the in vivo experiments. In this case, there was no significant difference between the performances of different descriptor packages as shown in **Table 5**. Analogously, **Table 4** provides a comparison of different machine learning algorithms. Also in this case all algorithms performed comparably good.

Conclusion

The comprehensive analysis of Phase I compounds shows that, with the exceptions of few in vivo toxicity end points, it is still challenging to build predictive toxicity

models for replacement of animal testing. The only end point with possible prediction power for such replacement was the acetyl cholinesterase inhibition. In a way, the limited chemical diversity of the dataset, consisting mainly of insecticides and pesticides, could have been responsible for both the success of modeling for this endpoint as well as the failure for modeling others.

The *in vitro* profiling of chemicals didn't have a significant improvement for model statistics. It was however possible to get better predictive models representing the *in vitro* assays using exclusively *in silico* descriptors. This might be due to the fact that each *in vitro* assay is typically measuring a small number of interacting genes and pathways, which is insufficient, when considering the more complex requirements needed to model a toxicity phenotype.

We also found a significant correlation, again with the exception of acetyl cholinesterase inhibition, between the number of toxic compounds and the median ability of algorithms and descriptors to build a predictive model as shown in Figure 13.

Many challenges remain in place: first of all, the use of a statistical approach, such as QSAR modeling, requires a considerable amount of data. The comparably low number of training instances limits the possibility to model the data in an appropriate way; secondly, the *in vitro* representation is probably too simple to address the complexity of the interactions in the organism. Properties, such as bioavailability and biotransformation play a significant role in terms of the toxic effect of a compound; thirdly and finally, it is possible that the assays conducted are not enough to capture biochemical events on the molecular level that can describe the pathways responsible for toxicity.

A consequence arising from this should be the careful investigation and analysis of potentially useful *in vitro* assays in terms of specific toxicity endpoints, as well as the identification of those *in vitro* assays, which enable proper QSAR modeling.

With that taken into consideration, ToxCast Phase I still provided useful overview of the chemical initiating events that could be useful for further investigation with a higher number compounds. For example, many assays may be redeemed unnecessary in future tests, as they were focused on promiscuous dormant endpoints. This initial phase offered to implement the required workflows and modeling infrastructure and enables to experience the needs and challenges of developing predictive biological signatures. Such infrastructure is now available for the analysis of future data releases.

As more data become available with the progress of the next phases of ToxCast and similar projects, it could be possible to build statistical models that support the prediction of toxicity and therefore reduce the number of animal experiments. Till then, *in vitro* assays for chemical profiling remains a useful investigation and exploratory tool. Previous study(Thomas et al., 2012) showed similar results and suggested that *in vitro* profiling could be useful for the prioritization of compounds, rather than the replacement of animal testing.

Example toxicity end point: Liver toxicity

The rat liver neoplastic lesions end point was selected from the EPA ToxcastTM/ToxRefDB in vivo assay endpoints as an example for modeling.

In silico descriptors

Models were built using Estate Indices, Dragon 6, AlogPS, ISIDA, chemaxon, ADRIANA.Code and CDK descriptor packages. All models were internally validated using 5-fold cross validation. Models were compared in their ability to predict the presence of rat liver neoplastic lesions.

Biologically-derived descriptors

In vitro assay responses were discretized using OCHEM software into a binary format (response/no response). The resulting binary experimental values were used as biologically derived descriptors to predict the in vivo endpoint (either alone or in combination with in silico descriptors)

Machine-learning algorithm

Models were built using FSMLR, SVM, KNN, and ASNN. Thus, linear and non-linear algorithms were evaluated. In general ASNN showed the best performance.

Consensus model

A consensus model was built between in vitro assay and the best performing in silico model (using Dragon 6 descriptor package)

Model access

Every model has a public id (see results table) which can be used to access the model on the iPrior system online at <http://iprior.ochem.eu> once it becomes approved and published.

Results:

Descriptor packages	True positive	False Positive	True Negative	False negative	Sensitivity	Specificity (%)	Balanced accuracy (%)	Accuracy
					(%)			

ALogPS	14	166	60	7	66.7	26.5	46.6	30
Dragon 6	14	109	118	7	66.7	52	59.3	53.2
ISIDA	6	97	130	15	28.6	57.3	42.9	54.8
Chemaxon	7	64	160	14	33.3	71.4	52.4	68.2
CDK	14	125	97	7	66.7	43.7	55.2	45.7
ADRIANA. CODE	13	144	79	8	61.9	35.4	48.7	37.7
ALOGPS+ ESTATE	4	115	111	17	19	49.1	34.1	46.6
In vitro assays	17	141	86	4	81	37.9	59.4	41.5
In vitro + Dragon 6	13	64	163	8	61.9	71.8	66.9	71

TP: True positive; **FP:** False positive; **TN:** True negative; **FN:** false negative; **SN:** sensitivity; **SP:** specificity; **BA:** Balanced accuracy; **Acc:** Accuracy.

Balanced accuracy was used as the measure for performance comparison between different models. The best performing model was a consensus built on models: (45096241; Dragon 6 in silico descriptors) and (id:39831812; biologically-derived HTS in vitro assays).

Conclusion

ToxCast Phase I dataset is particularly challenging for modeling. The data include only 309 different compounds and more than 500 in vitro assay endpoints which form a large number of descriptors that overwhelm machine learning algorithms. The data is also highly unbalanced making it difficult for parameter selection methods. However, Our study demonstrate that hybrid models, which incorporate both Toxcast™ in vitro parameters and in silico descriptors, provided higher accuracy for prediction of liver carcinogenicity compared to the separate use of individual descriptors. The in vitro parameters also expand the applicability domain of models.

Figure 13 Chart showing the number of toxic compounds and the median balanced accuracy for all modeled toxicity endpoints. With the exception of rat cholinesterase inhibition, there is a significant correlation between the performance of the model and the number of compounds.

Figure 14. Overview of 6039 models built for the prediction of 61 animal toxicity endpoints from the toxicity reference database. The endpoints names are shown on the x-axis ordered alphabetically based on the format in ToxRefDB database: study_type_species_organ_effect_category. The full list of endpoints and their description is provided in supplementary material. Study type: DV, developmental; CR, chronic; MGR, multigenerational. Species: Rt, rat; Rb, rabbit; Ms, mouse. Effect and category: Mat, maternal; GL-Mt, general maternal; Dev, developmental; PregRel, pregnancy related; PregLoss, pregnancy loss; AnyLes, any lesion; Skel, skeletal; PreneoplastLes, preneoplastic lesion; GenFetal, general fetal; Prolif-eratLes, proliferative lesion; WghtReg, weight reduction; NeoplastLes, neoplastic lesion; Reproduct, reproductive; ThyroidGld, thyroid gland; ReproductTract, reproductive tract; Perform, performance; Cholinester, cholinesterase; Inhibit, inhibition.

Figure 15 Overview of 11988 models built for the prediction of 148 in vitro assay endpoints from the toxicity reference database. Full description of each assay is available in the supplementary materials. ACEA: ACEA - Real-time Cell Electronic Sensing; ATG: Attagene - Transcription factor assays; BSK: BioSeek - Cell-based protein level assays; CLM: Cellumen - Cell imaging assays; CLZD: CellzDirect - Transcription assays; NCGC: NCGC - nuclear receptor assays; NVS: Novascreen / Caliper - receptor binding and enzyme inhibition assays; Solidus: Solidus - P450 vs. cytotoxicity assays

As more data become available from future ToxCast data release it might be possible to extend the current models. However, The concept of building an in silico model for an in vitro assay was still interesting for investigation using a bigger more homogenous dataset. For this, the activation of the Aryl Hydrocarbon receptor was investigated below.

2.5 AhR receptor activation model

The Aryl hydrocarbon receptor (AhR) is a ligand-dependent transcription factor. It regulates the expression of a battery of genes in a wide range of species and tissues. Among the most characterized chemical classes that are known to be ligands for AhR are environmental toxins, such as the halogenated aromatic hydrocarbons (HAHs) and nonhalogenated polycyclic aromatic hydrocarbons (PAHs). Endogenous ligands have also been reported. Exposure to TCDD (dioxin), the prototypical and most potent HAH, and related compounds produces a diverse array of species- and tissue- specific toxic and biological effects, the majority of which are AhR dependent

In vitro assay:

Primary cell-based high throughput screening assay to identify activators of the Aryl Hydrocarbon Receptor (AHR) was conducted by The Scripps Research Institute Molecular Screening Center (SRIMSC). Overall, 324751 compounds were tested, of which 7988 compounds were active. A mathematical algorithm was used to determine nominally activating compounds in the primary screen. Two values were calculated: (1) the average percent activation of all compounds tested, and (2) three times their standard deviation. The sum of these two values was used as a cutoff parameter, i.e. any compound that exhibited greater % activation than the cutoff parameter was declared active. The data were made available through the pubchem bioassay database (AID: 2796). The in vitro testing of such a large number of compounds provides the potential for building a predictive model for the AHR activation.

2.5.1 In silico model building

Data were uploaded to the OCHEM modeling framework and multiple classification models were developed using different descriptor packages and a dataset of **15974 compounds** (all 7988 active compounds and equal number of randomly selected inactive compounds). Models were built using Estate Indices, Dragon 5&6, AlogPS, ISIDA, CDK, MERAnMERSY, and chemaxon descriptors. Linear and nonlinear algorithms were evaluated: Ann, Knn, SVM, J48, Random forests. 2D and 3D structure were evaluated (The software package CORINA, integrated into OCHEM, was used for 3D structure generation).

Parameter selection was performed using 90% correlation cut-off. Cross-validation used stratified bagging validation with 64 bagging models

The machine learning algorithms had little impact on the model quality. By evaluating the applicability domain of the developed model, one can reach an accuracy of > 90% for the top 50% of compounds.

A quantitative model was built for modeling the % inhibition. The cross-validated Q squared was > 0.53 (using Dragon or fragmental descriptors) and neural networks. The quality of the model deteriorated with MLR and Dragon to 0.28. It was also affected in case of fragment descriptors but not as bad (0.42).

To standardize the model development process, Chemaxon package was used to preprocess the molecules by removing salt counter ions, neutralizing ions and standardizing the chemical structures regarding nitro-group representation and aromaticity. Below is a summary of the sequence of chemical structure handling performed through OCHEM.

Molecules preprocessing:

- Aromatize structures for compatibility with certain descriptors
- Standardize molecular structure by given molecular templates
- Remove counter ions prior to descriptor calculation
- Neutralize compounds
- Generate 3D structures (by Corina or Mopac)

Descriptor selection was done by eliminating non-useful descriptors (with less than 2 unique values), deleting descriptors that have failed in calculation by dragon (reported with a value of 999999), deleting descriptors that have variance smaller than 0.01 and grouping descriptors, that have pair-wise correlations Pearson's correlation coefficient R larger than 0.95

Internal validation was performed using stratified bagging validation with 64 bagging models and showed similar results.

The online chemical modeling framework (OCHEM) was used as it is a robust system to handle the chemical structures, descriptors calculation and model validation. Data were introduced to using Knime (Workflow management tool).

Figure 16. Diagram representing the workflow process for the model development, starting with data download from Pubchem Bioassay followed by data upload to OCHEM using Knime and finally the QSAR modeling using OCHEM

2.5.2 Results:

The three best performing models were built using Chemaxon descriptors implemented in WPI. The balanced accuracy for the three best performing models is shown in Table 8

Algorithm	Balanced accuracy
Random forests	71.8%
J48	73.1%
ASNN	72.9%

Table 8 Balanced accuracy for the 3 best performing models built using Chemaxon descriptors

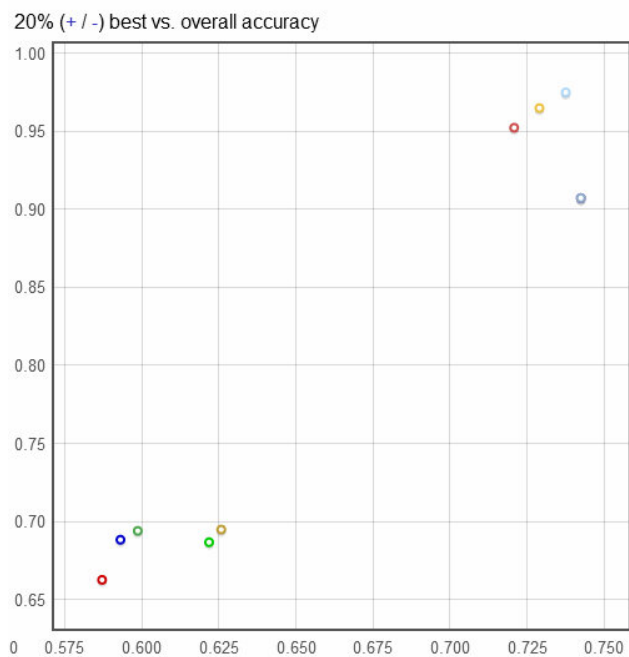


Figure 17. Graph representing a comparison between overall accuracy and accuracy of top 20% subset of the dataset. Each dot represents a model. Models lying in the top right corner are considered more predictive as they would have better overall accuracy as well as better accuracy for the 20% of compounds nearest to the model from a distance-to-model perspective.

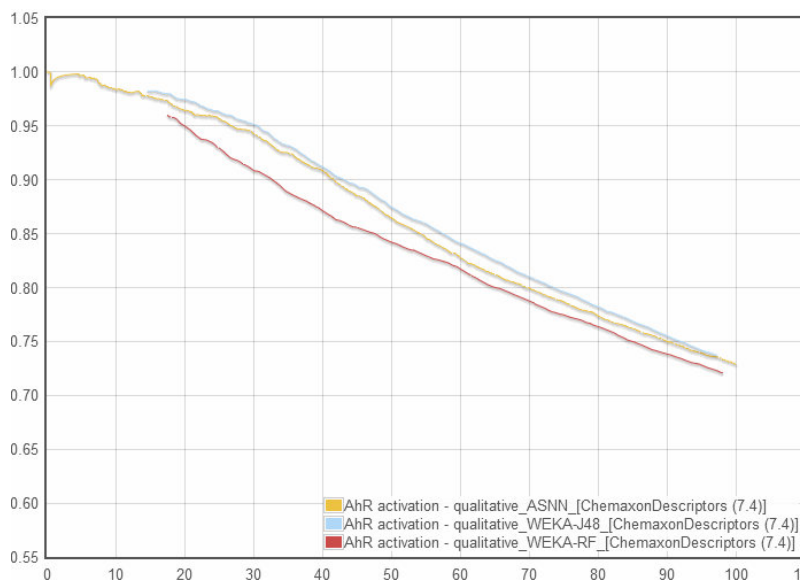


Figure 18 Graph representing the overall applicability domain performance of 3 models developed using chemaxon descriptors with different machine learning algorithms (Associative neural networks, J48 WEKA implementation, and random forests)

Best performed model using single descriptor package

<input type="checkbox"/> Experimental / prediction <input type="checkbox"/>	Inactive (Training set)	Active (Training set)	Inactive (Test set)	Active (Test set)
Inactive	4478	1912	1119	464
active	1520	4846	404	1207

Table 9 Confusion matrix for the best performing model, based on balanced accuracy, built using the descriptor package Chemaxon implemented in WP1 and the J48 machine-learning algorithm.

Dataset	# compounds	Balanced accuracy
Training set	12733	73.1% ± 0.8
Test set	3193	72.8% ± 1.5

Table 10 balanced accuracy for the training and test sets for the best performing model.

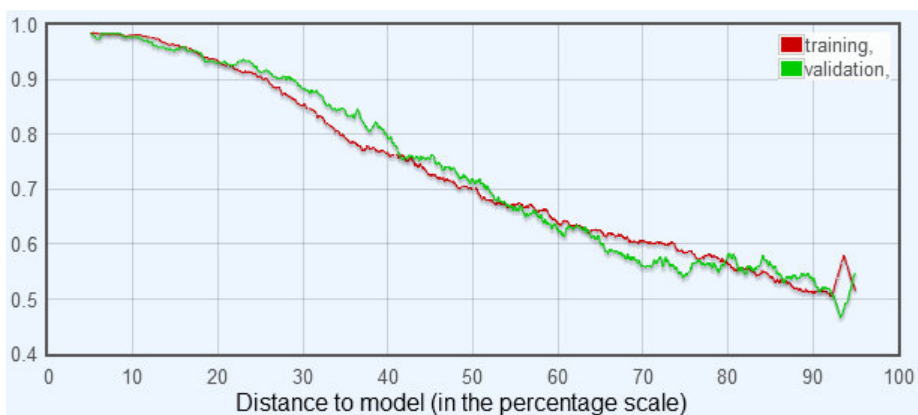


Figure 19 Williams plot for the AHR activation classification model representing the training set (red) and the test set (green) showing balanced accuracy on the y-axis and the distance to model for % of compounds on x-axis. It shows that the 20% nearest compounds to the model had a balanced accuracy >90%

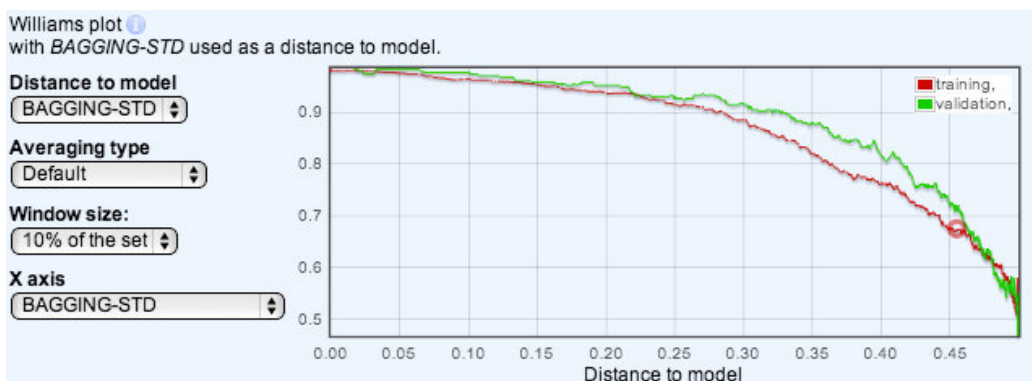


Figure 20 Williams plot for the applicability domain of AHR activation classification model representing the training set (red) and the test set (green) showing balanced accuracy on the y-axis and the distance to model in terms of Bagging standard deviation on x-axis. It shows that prediction accuracy decreases (moving left to right) as the distance to the model increases.

Consensus model:

We investigated the possibility to further improve the model performance using consensus modeling. The 3 best performing models were used to build a simple average consensus model. The consensus model showed better balanced accuracy than any single model as shown below:

□ Experimental / prediction □	Inactive (Training set)	Active (Training set)	Inactive (Test set)	Active (Test set)
Inactive	4816	1574	1185	398
active	1712	4654	435	1176

Table 11 Confusion matrix for the consensus model developed using Chemaxon descriptors and J48 algorithm.

Dataset	# compounds	Balanced accuracy
Training set	12756	74.2% ± 0.8
Test set	3194	73.9% ± 1.5

Table 12 Statistics for the consensus model showing better balanced accuracy for both training and test sets than individual models.

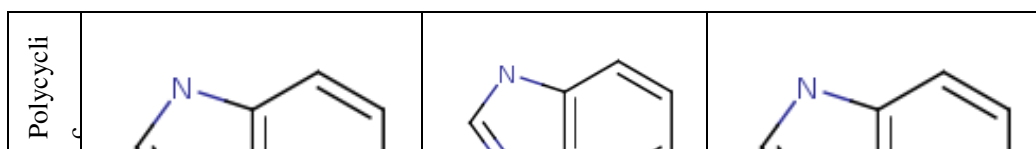
2.5.3 Analysing most relevant Chemaxon descriptors:

MLR model was used to investigate the descriptors that have the highest influence on the model (ordered by their relevance). These descriptors were found to be the most influential:

1. Smallest ring system size
2. Formalcharge at acidic PH
3. Molecular polarizability
4. Carboaromatic ring count
5. Formalcharge at acidic PH
6. Aliphatic ring count of size 4 and 8
7. Donor site count
8. Minimal projection size and area

2.5.4 Analysis of relevant fragments for AhR activation:

In order to get a better understanding of the models for AhR activation, the setCompare utility on OCHEM.eu was used to compare the set of activators and non-activators to determine the fragments which are significantly more abundant in the set of activators but absent in the set of non-activators. The fragments were found from chemical classes that are well-known to activate the AhR. Below are some of these fragments with their associated p-values:



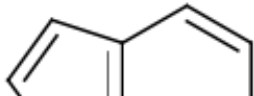
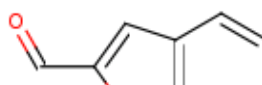
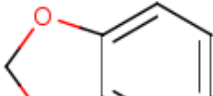
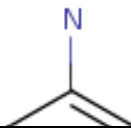
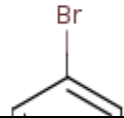
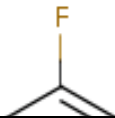
	P-value: $10^{-3.6}$	P-value: 10^{-6}	P-value: $10^{-2.5}$
benzofurans and			
	P-value: $10^{-12.8}$	P-value: $10^{-4.9}$	P-value: $10^{-5.6}$
Halogenated			
	P-value: 10^{-15}	P-value: 10^{-5}	P-value: 10^{-6}

Table 13 fragments significantly associated with the activation of the AhR showed together with their respective P-values

WORK PACKAGE 7: Developing of web tools for utilizing the models developed during the project.

Bioavailability models:

All bioavailability models built using OCHEM are accessible online via the OCHEM platform www.ochem.eu which is hosted for free non-commercial use by the Helmholtz Zentrum Muenchen. Academics all over the EU and the world are able to access the models, apply them for prediction of their compounds or use them as an input for building their own models.

Public links for the models developed are:

- Model for human intestinal absorption: <https://ochem.eu/model/4887243>
- Model for plasma protein binding: <https://ochem.eu/model/29533224>
- Model for hepatic clearance (by hepatic suspension cells):
<https://ochem.eu/model/27662546>
- Model for the Aryl hydrocarbon activation: <https://ochem.eu/model/46085288>

Orange and Knime workflow are available at: www.amaziz.com/eco/workflows

ToxCast analysis: A special modified version of OCHEM dedicated for the prioritization and estimation of toxicity of chemical compounds was deployed online

to act as a repository for storing all data and models generated. The new modeling framework is called iPRIOR and can be accessed online through the URL (<http://iprior.eadmet.com>). For dissemination of information, scientists can create free accounts on the framework and access models and their statistics. They can apply these models to new chemical structures, or use them to build even more complex models

WORK PACKAGE 8: Document the project achievements and report future developments and suggestions as well as the potential for commercialization.

The research conducted in this project will be documented in a PhD thesis to be submitted by the researcher to the technical University in Munich (TUM). A version of the thesis will also be available online.

The iPrior platform will be further hosted by eADEMT GmbH, a spin-off from the host organization, which will also explore the commercialization possibilities of the reported research results.

The Marie Curie Initial Training Network (ECO) offered the researcher an exceptional opportunity to attend multiple conferences and network with the chemoinformatics community. Not only scientifically, but also on the self-development level, the network provided the researcher with language courses that allowed him to better integrate into the European host country. The industrial internship at Pfizer Inc global research site was of particular importance as it allowed a much better understanding of the Pharmaceutical industry needs and how the in silico models are utilized in practice. Below is a list of the activities that the researcher took part between the periods (March 2010 – April 2013).

The researcher is currently employed in eADMET GmbH and wishes to continue his research activities in Europe.

Milestone 3: Summing the work done on the project by providing a thesis describing the methods and techniques developed as well as web tools exposing the technologies used to other scientists who can benefit from it.

3 Conferences and Meetings attended

3.1.1 Chem/Bioinformatics

- German conference in chemoinformatics (Goslar, 2010)
- Chemaxon European user group meeting (Budapest, 2011)
- OpenTox meeting (Munich, 2011)
- German Conference on Bioinformatics (Weihenstephan, 2011)
- 7th German conference in chemoinformatics (Goslar, 2011)
- BioTech NetWorkshop 2012 (Schloss Ringberg, January 2012)

- Munich Interact (Munich, March 2012)
- 12th symposium on ePhyschem (April, 2012)
- Chemaxon European Group Meeting 2012 (May, 2012)
- Biovaria 2012 (May, 2012)
- ASTP annual conference (May, 2012)
- 3rd Strasbourg summer school on chemoinformatics (June 2012)
- 244th American Chemical Society meeting (Philadelphia, US, August, 2012)
- Second Cadaster workshop (Munich , October 2012)
- International conference for the information Technology (Berlin October 2012)
- German Conference on Chemoinformatics (Goslar, November 2012)

3.1.2 Entrepreneurship

- Innovation Days 2012 (Munich, November 2012)
- SACHS Biotechnology investment Forum (Zurich, October 2012)

3.1.3 DMPK/ADME

- ADMET Europe (Munich, 2010)
- International pharmaceutical federation congress (Lisbon, 2010)
- ADMET Europe (Munich, 2011)
- International Pharmaceutical Federation congress (Amsterdam, 3-8 October 2012)
- ADME and predictive Toxicology Europe (Munich, March 2012)

3.1.4

3.1.5 Trainings

- Spring school in Bioinformatics (Hohenkammer, 2010)
- European Patent Academy Workshop (Munich, 2010)
- SimCYP PBPK modeling (Konstanz, 2011)
- Research management training (ReMaT)
- German language courses (2010, 2011)

3.1.6 Internship

- In the group of Prof. Hilde Spahn-Langguth (Mainz, Germany, July 2012)
- In the Institute of Environmental sciences (CML), university of Leiden under the supervision of Prof. dr. ir. W.J.G.M. (Willie) Peijnenburg (Leiden, Netherlands, November 2012)
- Pfizer Inc. Global research site (Groton, CT, USA, March-April 2013)

3.2 Publications

3.2.1 Peer reviewed articles

- Sushko, I. et al, 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.
- Stefan Brandmaier, Ahmed Abdelaziz, Igor V. Tetko, Balance in the chemical space: Cleaning datasets from structurally similar compounds (in review)
- Stefan Brandmaier et al, The QSPR-Thesaurus: The online platform of the CADASTER project (in review)

3.2.2 Posters

- Abdelaziz A.; Alexander Safanayev; Tetko, I., "Building QSAR for HTS in vitro assays – A study for the prediction of Aryl hydrocarbon receptor activators" German Conference on Chemoinformatics (Goslar, November 2012)
- Abdelaziz A.; Tetko, I., "Combining HTS in vitro assays with in silico descriptors for Liver toxicity modeling" 244th American Chemical Society meeting (Philadelphia, US, August, 2012)
- Abdelaziz A.; Alexander Safanayev; Tetko, I., "QSAR modeling for the evaluation of Aryl Hydrocarbon receptor activators" 12th symposium on ePhyschem, Budapest, Hungary, March 2012
- Abdelaziz A.; Tetko, I., "Using Toxcast™ HTS assays as biologically derived descriptors in QSAR" 3rd Strasbourg summer school on chemoinformatics, Strasbourg, France, June 2012
- Abdelaziz A.; Alexander Safanayev; Tetko, I., "QSAR modeling for the evaluation of Aryl Hydrocarbon receptor activators" ADME and predictive Toxicology Europe & Munich Interact, Munich, Germany, March 2012
- Ahmed Abdelaziz, Iurii Sushko, Wolfram Teetz, Robert Körner, Sergii Novotarskyi, Igor V. Tetko "QSAR modeling for In vitro assays: linking ToxCast™ database to the integrated modeling framework, OCHEM" German Conference on Chemoinformatics, Goslar, Germany, 6-8 November 2011
- Abdelaziz A.; Körner R.; Novotarskyi S.; Teetz W.; Sushko I.; Tetko, I., "QSAR modeling for In vitro assays: linking ToxCast™ database to the integrated modeling framework-OCHEM" German Conference on Bioinformatics, Weihenstephan, Germany, 7-9 September 2011
- "Active and Reactive Metabolites Formed During Hepatic First-Pass: Simulations Featuring Their Contribution to the Overall Effect in Altered Liver Clearance and Drug-Drug Interactions" OpenTox 2011 InterAction Meeting Program, Munich, Germany, 9-12 August 2011
- Abdelaziz A.; Körner R.; Novotarskyi S.; Teetz W.; Pandey A.; Sushko I.; Rupp M.; Tetko, I.:OCHEM: "public QSAR framework with integrated measurements database" Chemaxon eUGM 2011, Budapest, Hungary, 15-20 May 2011

- Brandmaier, S.; Abdelaziz, A.; Sahlin, U.; Oberg, T.; Tetko, I. Stepwise D-Optimal design based on latent variables, interact 2011 Munich, Munich, Germany, April 7, 2011
- Abdelaziz A.; Tetko, I.; Spahn-Langguth H., "Prediction of kinetic characteristics of drug metabolites in-silico: The distribution characteristics of beta-adrenoceptor antagonists" ADMET Europe 2011, Munich, Germany, 28-29 March 2011
- Abdelaziz A.; Körner R.; Novotarskyi S.; Teetz W; Pandey A.; Sushko I.; Rupp M.; Tetko, I., "OCHEM: public QSAR framework for modeling PK/PD parameters" ADMET Europe 2011, Munich, Germany, 28-29 March 2011

3.2.3 Talks

- Abdelaziz A., "Kinetics of active metabolites: Compartmental approach and in-silico predictions accounting for first-pass metabolism" Karl-Franzens-University Graz, Austria, June 17, 2010
- Abdelaziz A.; Alexander Safanayev; Tetko, I., "QSAR modeling for the evaluation of Aryl Hydrocarbon receptor activators" 244th American Chemical Society meeting (Philadelphia, US, August, 2012)
- Abdelaziz A.; Tetko, I., " Combining HTS in vitro assays with in silico descriptors for Liver toxicity modeling" 244th American Chemical Society meeting (Philadelphia, US, August, 2012)

3.2.4 Software Tools and Trainings

In the field of Chemoinformatics, it is essential to receive training on as many informatics/computational tools. During the course of the project, I was keen to receive training on many valuable tools. Below is a summary of these tools

Software	Developer	Training
OCHEM	Helmholtz Zentrum Muenchen & eADMET GmbH	* Internal tool used during the whole project * Implemented the Chemaxon descriptor package
SimCYP simulator	SimCYP Limited	* One week intensive training course on Model-based drug development: Incorporating population variability into mechanistic prediction of PK and modelling of PK-PD * One year academic license Course details: http://www.simcyp.com/ProductServices/Workshops/20110411_Konstanz.htm?p=1
Instant JChem, Marvin Sketch, Marvin beans	Chemaxon SRL	* 2011 and 2012 Chemaxon European user group meeting developer training and End-user training * 3 year academic license
Schrödinger	Schrödinger http://www.schrodinger.com/about/	* One-day user training at Helmholtz Zentrum Muenchen * License through the Technical University of Munich (TUM) Schrödinger http://www.schrodinger.com/about/
OpenToxLab	BioGraf3R	* User manuals * 2 Year academic license
Knime	KNIME.com AG	* User manuals * Free license
GastroPlus, ADMET predictor	SimulationsPlus inc	* User manuals and online webinars * Access during the internship at Prof. Hilde-Spahn Langguth in Mainz

Orange	University of Ljubljana, Slovenia	* User manual and tutorials * Free license
WEKA	The University of Waikato	* User manual and tutorials * Free
R – The statistics package	Statistics Department of the University of Auckland	* 2-day training at the technical university in Munich (“Using R for statistical data analysis II”) * training course during the Kalmar winter school (“Advances methods for regression and classification, and how to use them in R”, Peter Filzmoser) * Free
Matlab, SimBiology	MathWorks	* User manual and online webinars * Trial license

References

- ADRIANA.Code - Calculation of Molecular Descriptors | Inspiring Chemical Discovery. (n.d.). Retrieved September 28, 2013, from <http://www.molecular-networks.com/products/adriana-code>
- Aires-de-Sousa, J., & Gasteiger, J. (2001). New description of molecular chirality and its application to the prediction of the preferred enantiomer in stereoselective reactions. *Journal of chemical information and computer sciences*, *41*(2), 369–375.
- Balon, K., Riebesehl, B. U., & Müller, B. W. (1999). Drug Liposome Partitioning as a Tool for the Prediction of Human Passive Intestinal Absorption. *Pharmaceutical Research*, *16*(6), 882–888. doi:10.1023/a:1018882221008
- Bergström, C. A. S., Norinder, U., Luthman, K., & Artursson, P. (2002). Experimental and computational screening models for prediction of aqueous drug solubility. *Pharmaceutical Research*, *19*(2), 182–188.
- Betts, K. S. (2013). Tox21 to date: steps toward modernizing human hazard characterization. *Environmental health perspectives*, *121*(7), A228. doi:10.1289/ehp.121-a228
- Boyadjiev, S. A., & Jabs, E. W. (2000). Online Mendelian Inheritance in Man (OMIM) as a knowledgebase for human developmental disorders. *Clinical genetics*, *57*(4), 253–66. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/10845565>
- Calculator Plugins « ChemAxon – cheminformatics platforms and desktop applications. (n.d.). Retrieved September 28, 2013, from <https://www.chemaxon.com/marvin/help/calculations/calculator-plugins.html>
- Cherkasov, A., Ban, F., Santos-Filho, O., Thorsteinson, N., Fallahi, M., & Hammond, G. L. (2008). An updated steroid benchmark set and its application in the discovery of novel nanomolar ligands of sex hormone-binding globulin. *Journal of medicinal chemistry*, *51*(7), 2047–2056.
- Dix, D. J., Houck, K. A., Martin, M. T., Richard, A. M., Setzer, R. W., & Kavlock, R. J. (2007). The ToxCast program for prioritizing toxicity testing of environmental chemicals. *Toxicol Sci*, *95*(1), 5–12. doi:kf1103 [pii] 10.1093/toxsci/kf1103
- Dorronsoro, I., Chana, A., Abasolo, M. I., Castro, A., Gil, C., Stud, M., & Martinez, A. (2004). CODES/Neural Network Model: a Useful Tool for in Silico Prediction of Oral Absorption and Blood-Brain Barrier Permeability of Structurally Diverse Drugs. *QSAR & Combinatorial Science*, *23*(2-3), 89–98. doi:10.1002/qsar.200330858
- Gene Ontology Documentation. (n.d.). Retrieved September 21, 2013, from <http://www.geneontology.org/GO.contents.doc.shtml>
- Ghuloum, A. M., Sage, C. R., & Jain, A. N. (1999). Molecular hashkeys: a novel method for molecular characterization and its application for predicting important pharmaceutical properties of molecules. *J Med Chem*, *42*(10), 1739–1748. doi:10.1021/jm980527a jm980527a [pii]
- Gunturi, S. B., & Narayanan, R. (2007). In Silico ADME Modeling 3: Computational Models to Predict Human Intestinal Absorption Using Sphere Exclusion and kNN QSAR Methods. *QSAR & Combinatorial Science*, *26*(5), 653–668. doi:10.1002/qsar.200630094
- Hall, L. H., Kier, L. B., & Brown, B. B. (1995). Molecular similarity based on novel atom-type electrotopological state indices. *Journal of chemical information and computer sciences*, *35*(6), 1074–1080.
- Holmes, G., Donkin, A., & Witten, I. H. (1994). WEKA: a machine learning workbench (pp. 357–361).
- Ingenuity IPA - Integrate and understand complex 'omics data. (n.d.). Retrieved September 21, 2013, from <http://www.ingenuity.com/products/ipa>

- iPrior - Prioritization and estimation of toxicity of chemical compounds. (n.d.). Retrieved September 24, 2013, from <http://iprior.eadmet.com/home/show.do>
- Irvine, J. D., Takahashi, L., Lockhart, K., Cheong, J., Tolan, J. W., Selick, H. E., & Grove, J. R. (1999). MDCK (Madin-Darby canine kidney) cells: A tool for membrane permeability screening. *J Pharm Sci*, *88*(1), 28–33. doi:10.1021/js9803205 10.1021/js9803205 [pii]
- Judson, R. S., Houck, K. A., Kavlock, R. J., Knudsen, T. B., Martin, M. T., Mortensen, H. M., ... Dix, D. J. (2010). In vitro screening of environmental chemicals for targeted testing prioritization: the ToxCast project. *Environmental health perspectives*, *118*(4), 485–492. doi:10.1289/ehp.0901392
- Kansy, M., Senner, F., & Gubernator, K. (1998). Physicochemical high throughput screening: parallel artificial membrane permeation assay in the description of passive absorption processes. *J Med Chem*, *41*(7), 1007–1010. doi:10.1021/jm970530e jm970530e [pii]
- KEGG: Kyoto Encyclopedia of Genes and Genomes. (n.d.). Retrieved September 21, 2013, from <http://www.genome.jp/kegg/>
- Kleinstreuer, N. C., Judson, R. S., Reif, D. M., Sipes, N. S., Singh, A. V, Chandler, K. J., ... Knudsen, T. B. (2011). Environmental Impact on Vascular Development Predicted by High Throughput Screening. *Environmental health perspectives*. doi:10.1289/ehp.1103412
- Klopman, G., Stefan, L. R., & Saiakhov, R. D. (2002). ADME evaluation. 2. A computer model for the prediction of intestinal absorption in humans. *Eur J Pharm Sci*, *17*(4-5), 253–263. doi:S0928098702002191 [pii]
- KNIME | KNIME Desktop. (n.d.).
- Lewis, D. F. V, Modi, S., & Dickins, M. (2002). Structure-activity relationship for human cytochrome P450 substrates and inhibitors. *Drug metabolism reviews*, *34*(1-2), 69–82.
- Martin, M T, Houck, K. A., McLaurin, K., Richard, A. M., & Dix, D. J. (2007). Linking regulatory toxicological information on environmental chemicals with high-throughput screening (HTS) and genomic data. *The Toxicologist CD-An official Journal of the Society of Toxicology*, *96*, 219–220.
- Martin, M T, Knudsen, T. B., Reif, D. M., Houck, K. A., Judson, R. S., Kavlock, R. J., & Dix, D. J. (2011). Predictive Model of Rat Reproductive Toxicity from ToxCast High Throughput Screening. *Biology of reproduction*, *85*(2), 327–339. doi:10.1095/biolreprod.111.090977
- Martin, Matthew T, Judson, R. S., Reif, D. M., Kavlock, R. J., & Dix, D. J. (2009). Profiling chemicals based on chronic toxicity results from the US EPA ToxRef Database. *Environmental health perspectives*, *117*(3), 392.
- Novotarskyi, S., Sushko, I., Korner, R., Pandey, A. K., & Tetko, I. V. (2011). A comparison of different QSAR approaches to modeling CYP450 1A2 inhibition. *Journal of chemical information and modeling*, *51*(6), 1271–1280. doi:10.1021/ci200091h
- Oprea, T. I., & Gottfries, J. (1999). Toward minimalistic modeling of oral drug absorption. *Journal of molecular graphics & modelling*, *17*(5-6), 261–274,329. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/10840686>
- Pathway Commons. (n.d.). Retrieved September 21, 2013, from <http://www.pathwaycommons.org/about/>
- Raymond Tice Robert Kavlock, Ph.D., and Christopher Austin, M.D., P. D., Tice, R., Kavlock, R., & Austin, C. (n.d.). The U.S. “Tox21 Community” and the Future of Toxicology.
- REACH - Registration, Evaluation, Authorisation and Restriction of Chemicals. (n.d.). Retrieved September 21, 2013, from http://ec.europa.eu/enterprise/sectors/chemicals/reach/index_en.htm

- Roy, K., & Roy, P. P. (2009). QSAR of cytochrome inhibitors.
- Sadowski, J., Gasteiger, J., & Klebe, G. (1994). Comparison of automatic three-dimensional model builders using 639 X-ray structures. *Journal of chemical information and computer sciences*, 34(4), 1000–1008.
- Shah, I., Houck, K., Judson, R. S., Kavlock, R. J., Martin, M. T., Reif, D. M., ... Dix, D. J. (2011). Using nuclear receptor activity to stratify hepatocarcinogens. *PLoS One*, 6(2), e14584. doi:10.1371/journal.pone.0014584
- Steinbeck, C., Han, Y., Kuhn, S., Horlacher, O., Luttmann, E., & Willighagen, E. (2003). The Chemistry Development Kit (CDK): An open-source Java library for chemo- and bioinformatics. *Journal of chemical information and computer sciences*, 43(2), 493–500.
- Tetko, I V, Tanchuk, V. Y., Kasheva, T. N., & Villa, A. E. (2001). Estimation of aqueous solubility of chemical compounds using E-state indices. *Journal of chemical information and computer sciences*, 41(6), 1488–1493. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/11749573>
- Tetko, Igor V, Tanchuk, V. Y., & Villa, A. E. P. (2001). Prediction of n-octanol/water partition coefficients from PHYSPROP database using artificial neural networks and E-state indices. *Journal of chemical information and computer sciences*, 41(5), 1407–1421.
- Thomas, R. S., Black, M. B., Li, L., Healy, E., Chu, T.-M., Bao, W., ... Wolfinger, R. D. (2012). A comprehensive statistical analysis of predicting in vivo hazard using high-throughput in vitro screening. *Toxicological sciences : an official journal of the Society of Toxicology*, 128(2), 398–417. doi:10.1093/toxsci/kfs159
- Todeschini, R., & Consonni, V. (2009). *Molecular descriptors for chemoinformatics*. Wiley.com.
- US EPA, O. of W. C. and O. of E. I. (n.d.). ToxRefDB (Toxicity Reference Database). Retrieved from <http://actor.epa.gov/toxrefdb/>
- Varma, M. V, Sateesh, K., & Panchagnula, R. (2005). Functional role of P-glycoprotein in limiting intestinal absorption of drugs: contribution of passive permeability to P-glycoprotein mediated efflux transport. *Mol Pharm*, 2(1), 12–21. doi:10.1021/mp0499196
- Varnek, A., Fourches, D., Horvath, D., Klimchuk, O., Gaudin, C., Vayer, P., ... Marcou, G. (2008). ISIDA-Platform for virtual screening based on fragment and pharmacophoric descriptors. *Current Computer-Aided Drug Design*, 4(3), 191–198.
- Wessel, M. D., Jurs, P. C., Tolan, J. W., & Muskal, S. M. (1998). Prediction of human intestinal absorption of drug compounds from molecular structure. *Journal of chemical information and computer sciences*, 38(4), 726–735. Retrieved from http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=9691477
- Zhao, Y. H., Abraham, M. H., Le, J., Hersey, A., Luscombe, C. N., Beck, G., ... Cooper, I. (2002). Rate-limited steps of human oral absorption and QSAR studies. *Pharmaceutical Research*, 19(10), 1446–1457. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/12425461>