HelmholtzZentrum münchen
German Research Center for Environmental Health

Marie Curie Initial Training Network

Environmental Chemoinformatics (ECO)

Final project report

10 August 2012

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.

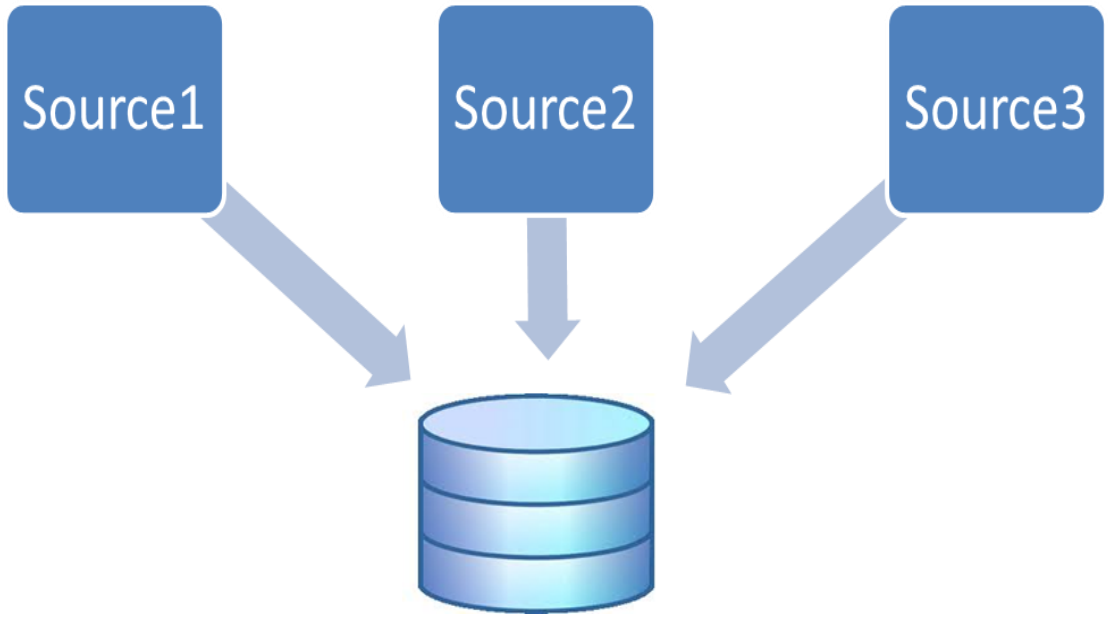

Fig.1 Data collection approach.

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

- ToxRefDB (http://actor.epa.gov/toxrefdb/faces/Home.jsp)

- SIDER (http://sideeffects.embl.de/)
- 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.

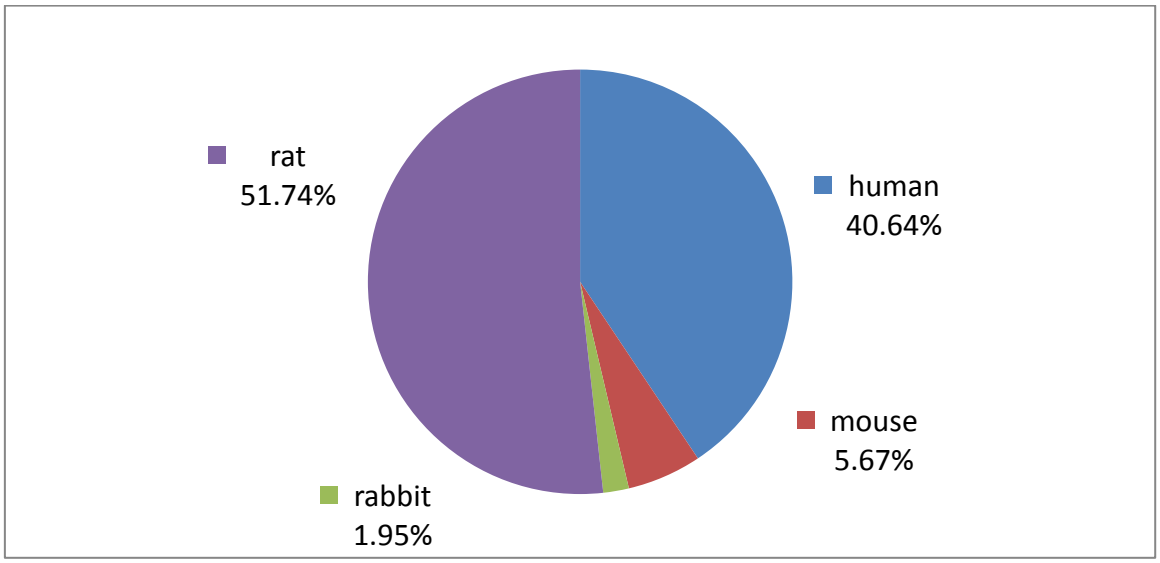

Fig. 2 Distribution of data across different organisms.

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

In this approach, the set of compounds annotated with a 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.

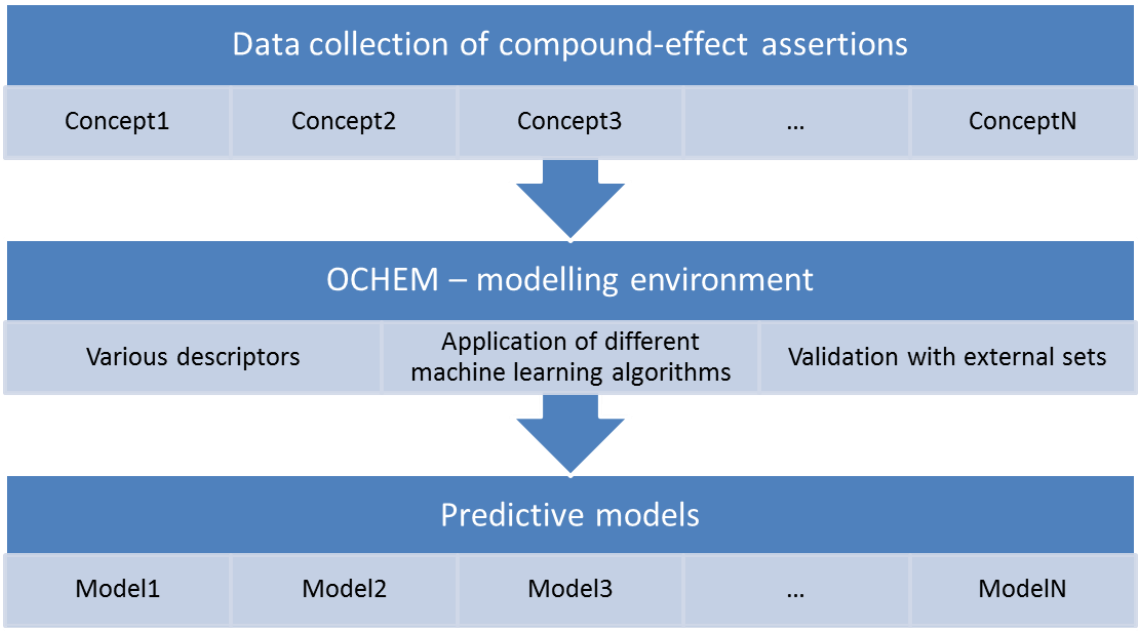

Fig. 3 Illustration of approach to build predictive models for the data.

Currently, we set a threshold of minimum 100 active molecules per concept when creating training sets for our models. This constraint helped increasing the statistical power of achieved models but on the other hand, limited the number of concepts to be modelled to 233.

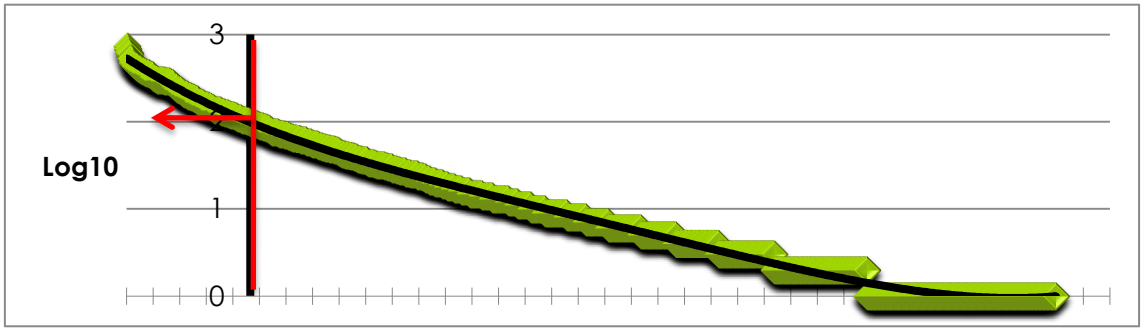

Fig. 4 After setting a threshold on a minimum number of molecules per concept to 100, 233 concepts were available to model.

Within the chosen concepts, majority of the indications (over 77%) stemmed from SIDER, with the following distribution across all sources:

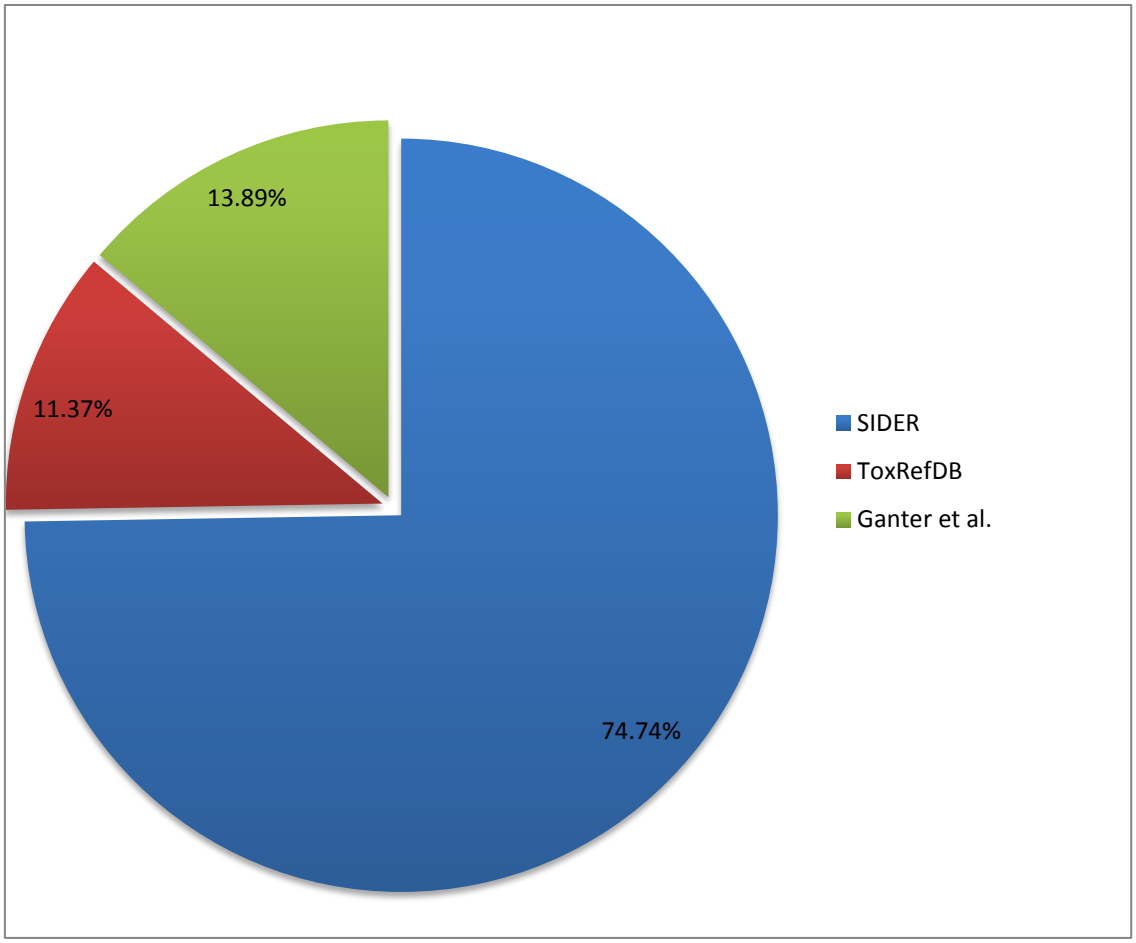

Fig. 5 Distribution of data according to source of indication for the chosen side-effects.

SIDER, being a database of human side effects, heavily contributes to an overall change of species distribution in our training set of selected side effects. 74.74% of indications at hand come from observations in human with the remaining 25.26% almost entirely in rat with negligible contribution from mouse and rabbit.

Curation of molecules

A careful consideration was given to the quality of obtained molecule structures via automated means of OCHEM (http://ochem.eu/) which relied heavily on PubChem (http://pubchem.ncbi.nlm.nih.gov/) in this matter. The obtained structures were validated against structures obtained from ChemSpider (http://www.chemspider.com/) via use of web services. There were some 180 molecules where suspicion was raised due to disagreement between alternatively fetched indications. For these, a manual curation was done and revealed that in more than 34% (62 molecules) the discrepancy was true and was not due to i.e. slightly different stereochemistry. At this I have also found that in some cases none of the sources provided the correct structures. All erroneous structures were manually corrected. In disputable cases, ChemSpider provided correct structures for 42 compounds, while PubChem did it for 20. Thus, ChemSpider is generally more reliable.

At this point, EState and AlogPS descriptors as well as those from Dragon 6.0 package were calculated to find out which of the molecules are suitable for modelling. Following this analysis we exclude very large molecules (like insulin) or molecules without carbon component (inorganic) for which calculation of descriptors is currently not possible/supported. Thanks to this approach, I found that 27 of the molecules in the dataset are not suitable for QSAR due to aforementioned reasons, and excluded them from further consideration in modelling.

Molecule Set at hand

The set of EState and AlogPS descriptors will be referred to as 2D while Dragon 6.0 descriptors as 3D due to the requirement of the 3D structure of molecules for the calculation of some descriptors. Three dimensional molecule structures were obtained using CORINA. Additionally, standard filters and procedures that are routinely applied at this point were used, i.e. pre-processing of molecules (standardization, neutralization, removal of salts), unsupervised filtering of redundant descriptors (variance smaller than 0.01, grouping descriptors that have pair-wise Pearson's correlation coefficient R larger than 0.95, elimination of descriptors with less than 2 unique values). On average, 263 descriptors were calculated in 2D set, and 1673 for 3D one.

Later, this 3D set of descriptors was normalized, and used for distance matrix calculation between the molecules in the set (Euclidean) and unsupervised clustering (average linkage). This analysis was applied to obtain a picture of heterogeneity of assembled molecule set. The following heat map was calculated according to the described procedure:

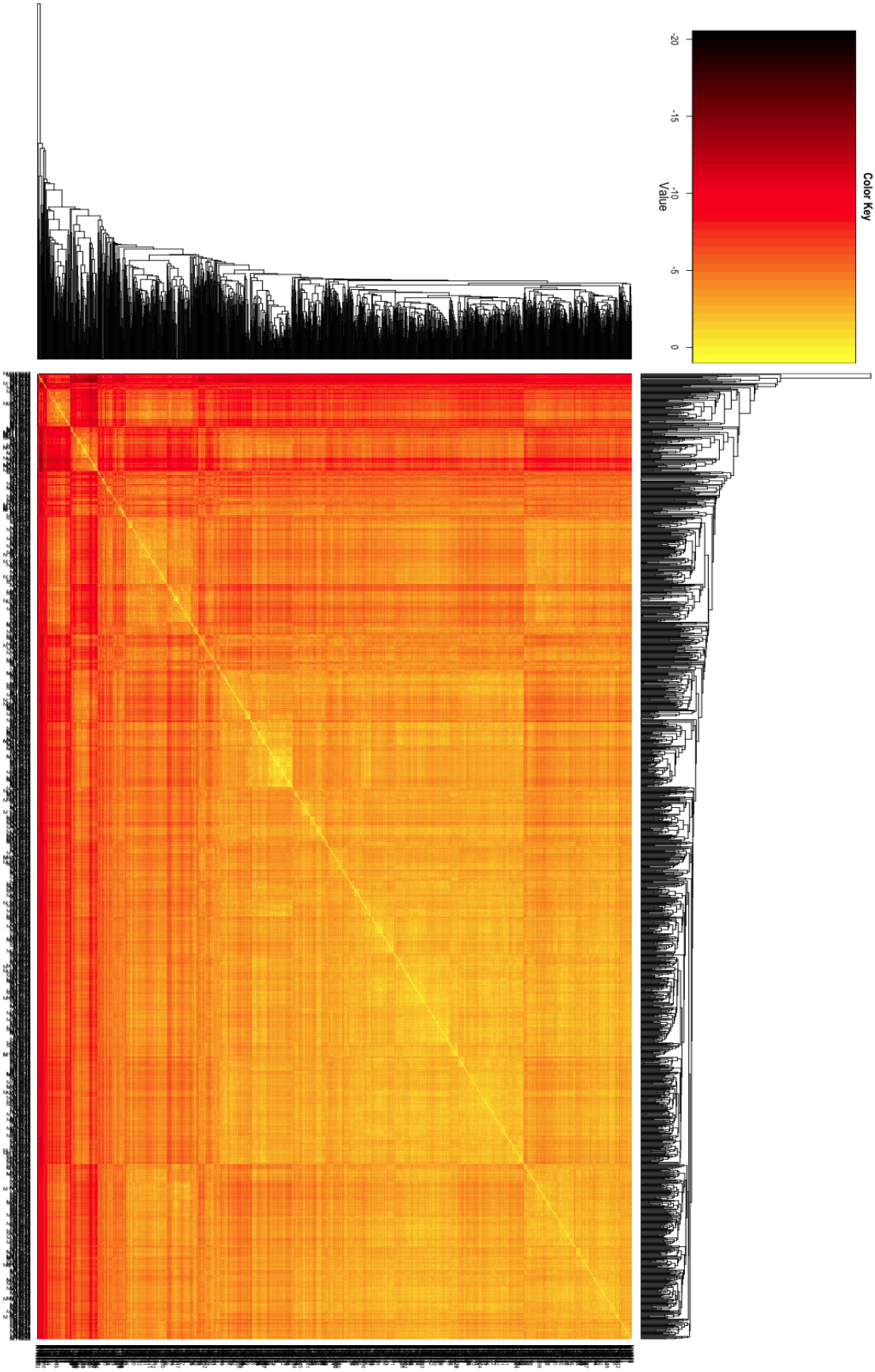

Fig. 6 Heatmap of molecules in dataset with visible dendrograms showing several clusters of highly similar molecules.

An analysis of enrichment of categorical annotations was also performed for molecules in the dataset utilizing the MBRole utility (http://csbg.cnb.csic.es/mbrole/index.jsp). The most interesting (from the perspective of this research) annotations were used for this analysis, including Biological role and Chemical role and application (based on ChEBI http://www.ebi.ac.uk/chebi/ and KEGG http://www.genome.jp/kegg/). The functions of molecules, which appeared most frequently in the dataset, are shown in Fig.7.

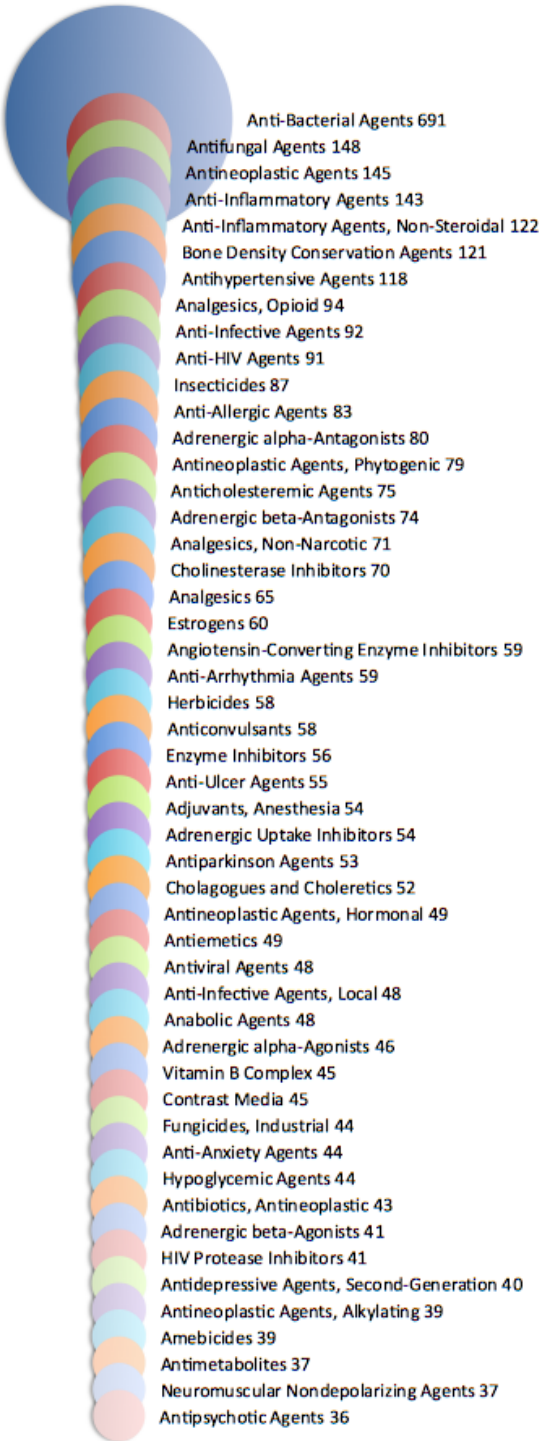

Fig. 7 Some of the most common annotated functions of molecules in the dataset.

Results

To date, a number of models were built utilizing different combinations of 2D and 3D descriptors and machine learning algorithms. Amongst machine learning algorithms we find Associating Neural Networks (ANN), Decision Trees (DT), K-Nearest Neighbours (KNN) and Random Forests (RF). For each combination of algorithm and set of descriptors, 233 models were trained on the same data and model statistics were obtained via use of automated workflow developed in KNIME (http://www.knime.com/). Predictions were cross-validated using stratified bagging with 64 models per training. Below are presented results averaged across each combination with highlight of best- and worse- approaches. On top of standard model statistics like Accuracy, positive predictive value (PPV), log odds ratio (LOR), sensitivity, specificity, false positive ratio (FPV); we also calculated balanced accuracy. It is to account for the fact that error for specificity of models is likely to be two times lower that we observe due to our assumption about non-active compounds. In many cases, discovery of a serious deleterious effect in animal prevents from further testing for human effects. It has also been reported that pharmaceutical companies, being the main source of data for this study, often conceal inconvenient results of drug tests [Angell M.] such as adverse effects that we are trying to predict with our approach. Nevertheless, this assumption had to be made but we are accounting for this in our analysis by inclusion of the error-balancing score:

Model cohort name	Feature						
	Accuracy	PPV	LOR	Sensitivity	Specificity	FPR	Balanced accuracy
ann+2d	68.58	0.271	10.13	57.64	70.59	0.294	71.47
ann+3d	72.02	0.306	10.24	62.55	73.54	0.265	74.66
dt+2d	73.41	0.311	10.40	58.20	75.88	0.241	73.07
dt+3d	75.46	0.323	10.55	55.56	78.74	0.213	72.47
knn+2d	68.21	0.274	10.07	55.38	70.29	0.297	70.26
knn+3d	52.60	0.238	8.71	79.61	48.06	0.519	76.82
rf+2d	79.29	0.360	10.98	47.49	84.81	0.152	69.95
rf+3d	79.74	0.362	11.07	44.27	85.98	0.140	68.63
average	71.16	0.31	10.27	57.59	73.49	0.27	72.17

Fig. 8 Summary of built models. Green and yellow backgrounds highlight best- and worst-performing approaches respectively within category given in the header of table.

In terms of sensitivity and balanced accuracy, the 2 characteristics most unbiased by negatives, the K-Nearest Neighbours approach with the use of 3D descriptors provided on average the

best solution. However, usable models for the following side effects were built with all combinations of machine learning algorithm and set of descriptors used:

Consistently well predicted side effect	Sources of indication			ANN+3D characteristics		
	SIDER	ToxRefDB	Ganter et al.	Sensitivity	Specificity	Balanced Accuracy
Excessive body weight gain	0	108	292	83.75	86.77	88.57
Absolute organ weight change, NOS	0	319	0	83.65	85.06	88.09
Body weight decrease	1	381	0	82.29	85.93	87.63
Decreased RBC	0	166	0	81.44	80.80	85.92
Hypertrophy	0	190	0	77.42	80.99	83.96
Nausea	685	0	0	78.66	78.37	83.92
Cholesterol increase	0	158	0	75.95	80.62	83.13
Enlargement, NOS	0	131	0	75.94	79.62	82.88
Discoloration, NOS	0	148	0	76.51	77.59	82.65
Headache	680	1	0	76.50	74.93	81.98
Emesis	629	0	0	75.92	75.33	81.79
Platelet change, NOS	0	102	0	73.27	79.66	81.55
Diarrhoea	588	2	0	73.31	74.88	80.38

Fig. 9 Details on consistently well-predicted side effects and example statistics for the ANN+3D approach.

In total, setting an arbitrary threshold of 80 for balanced accuracy, I found 79 individual models to speak of numbers. The top scoring effects (**Fig. 9**) represent a good mixture of the 3 sources of annotations with predominant SIDER annotations and least contribution from the rat liver xenobiotic response repertoire database. This gives some idea about the quality of the annotations from sources at hand. Unfortunately, the overlap between the sources was not

very significant with only one side effect from this list ("Excessive body weight gain") observed to share annotations across 2 sources.

From the 233 adverse effects chosen to model, only a few were predicted very poorly (i.e. below arbitral threshold of 60 for balanced accuracy in), with the worst performance achieved when using 2D descriptors and Random Forest algorithm (data not shown). These included such effects as: Ulcers, Cysts, Ecchymosis, Neck pain, Dysuria and Nocturia – with most annotations from human. These adverse effects are probably caused by a very diverse set of compounds that perhaps is too complicated to be captured within the descriptors used. Also, the very complex mechanisms behind occurrence of these effects might contribute to the overall inability of QSAR here.

Correlations in models

When taking into consideration individual models for chosen combinations of machine learning approach and set of descriptors, one could try to look for correlations between model statistics and number of active compounds:

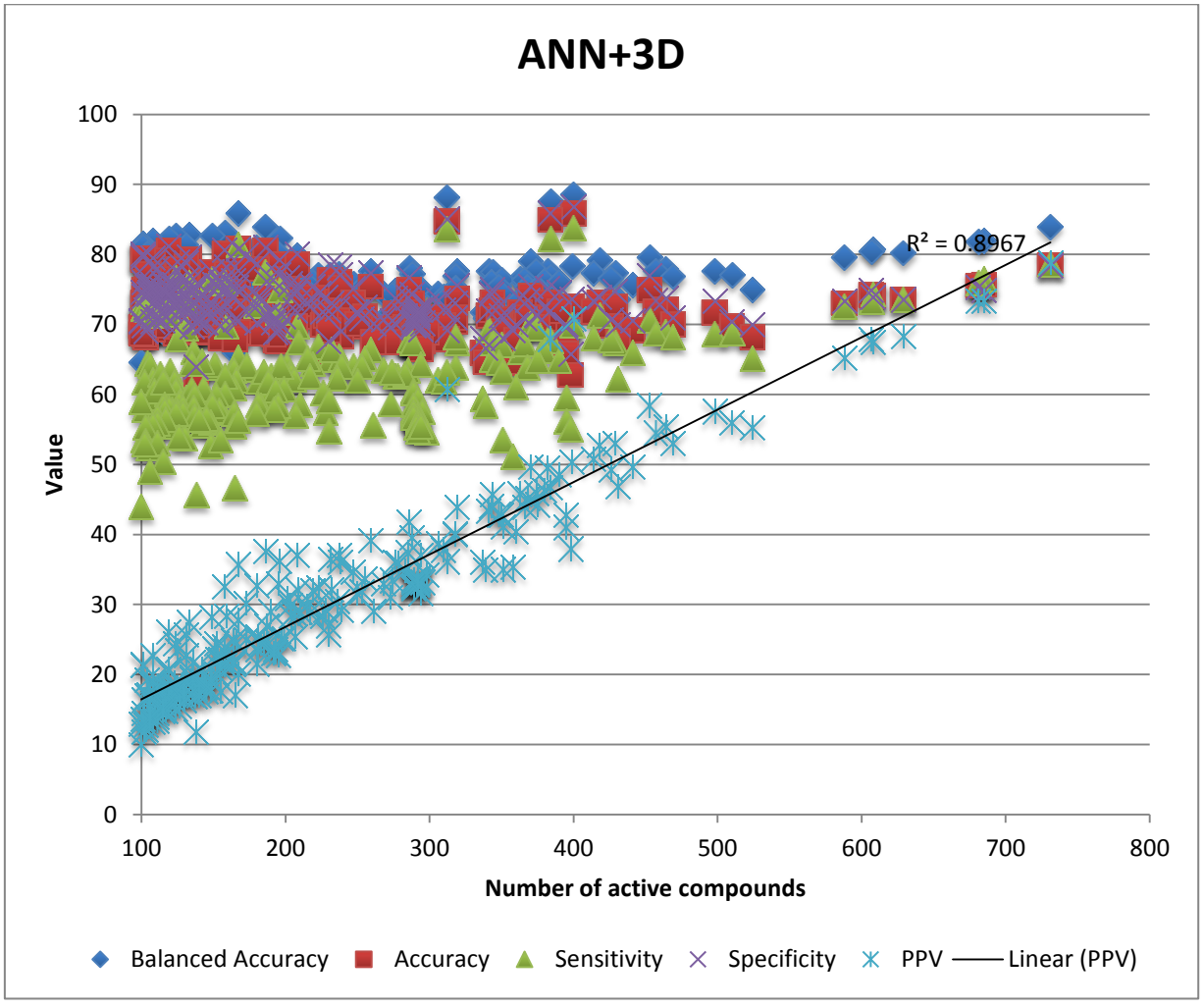

Fig. 10 Seeking for possible relationships between model statistics and number of active compounds using example of ANN trained on 3D descriptors' values.

Presented above is an example plot with summary values of models trained using combination of ANN and 3D descriptors - similar picture is seen for other combinations of algorithm and descriptor set. The positive correlation seems apparent for the positive predictive value (PPV in the plot) with R^2 = ~0.90 when approximating this relationship with least squares linear regression. PPV is the relation of true positives to all positive calls. Based on that definition and pattern seen in our models, we conclude that number of active compounds is absolutely decisive for the final quality of the model in terms of PPV with best models having most number of active compounds. This model behaviour could be explained by the more comprehensive compound annotation for these popular concepts or underreporting of less popular effects.

In similar fashion, one could plot individual models onto specificity against sensitivity graph:

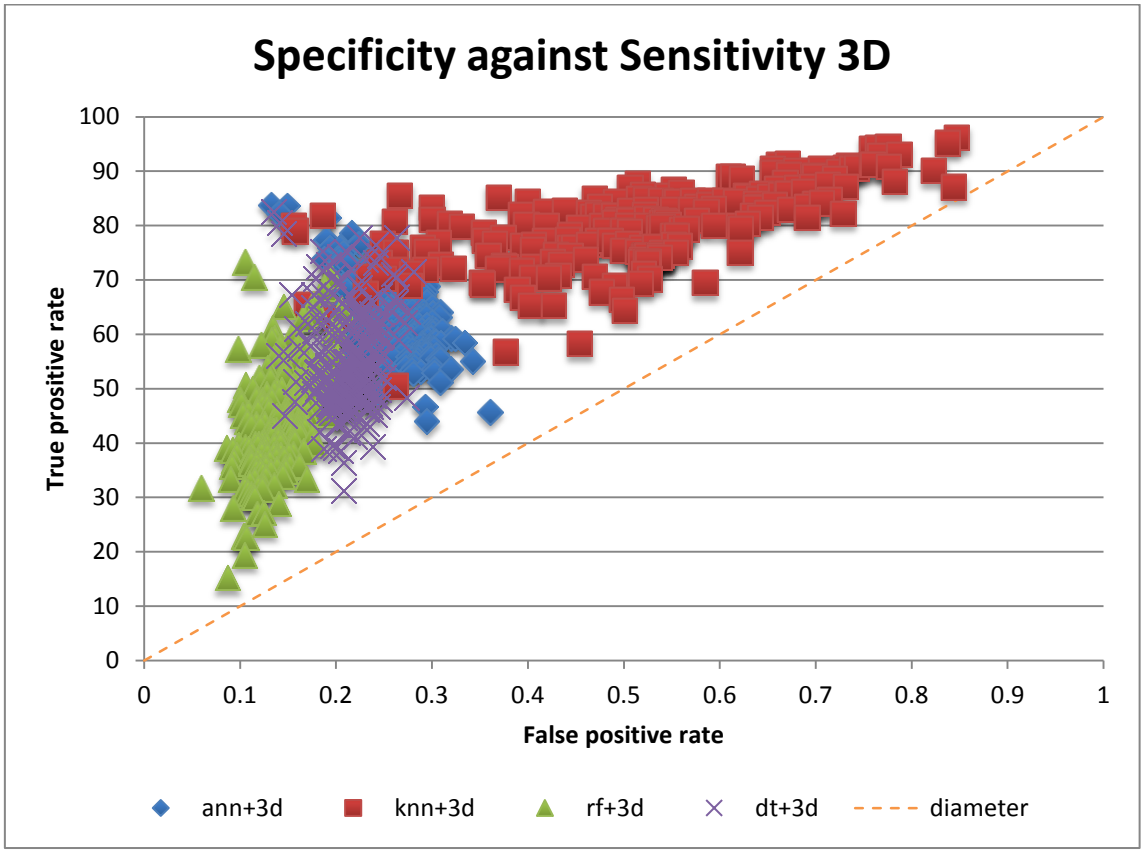

Fig.11 Specificity against sensitivity comparison dot plot for models built based on 3D descriptors.

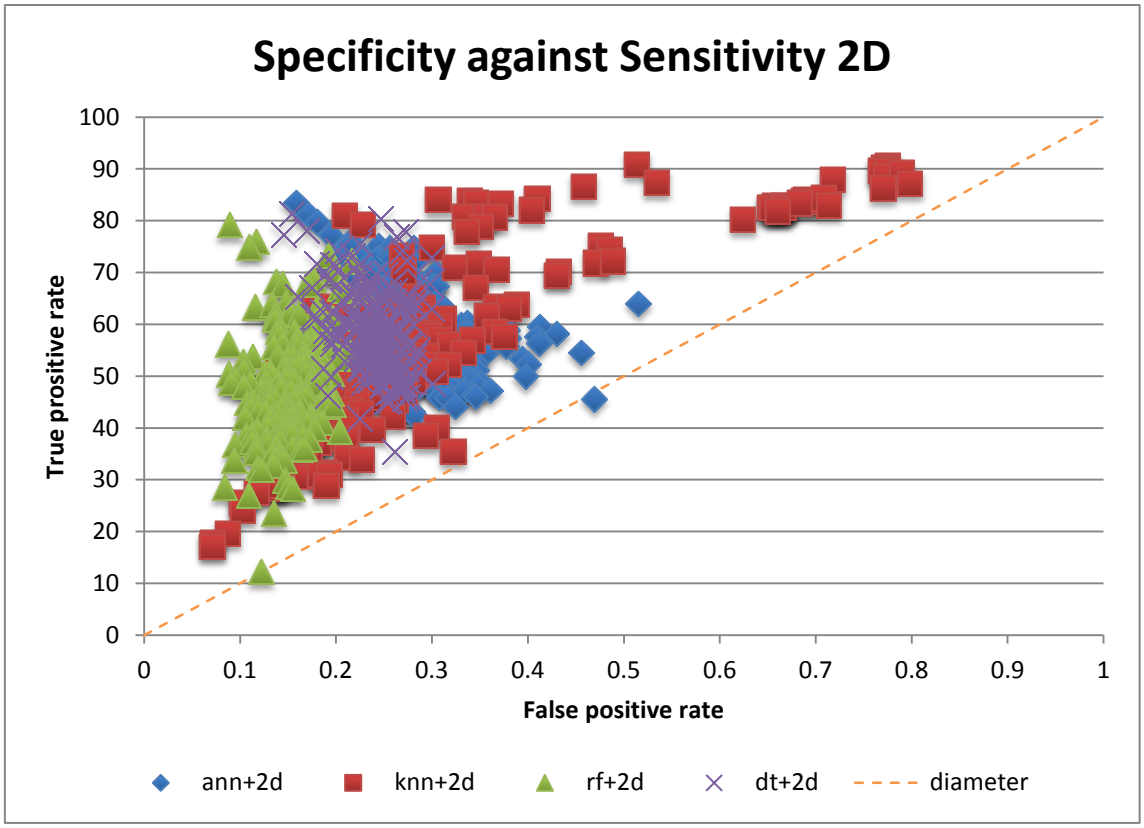

Fig. 12 Specificity against sensitivity comparison dot plot for models built based on 2D descriptors.

According to the above analysis, all models using 3D descriptors, and most using 2D ones (2 outliers), predict better than random (points above hyphened orange diagonal, where completely randomly predicting models would align) with different approaches occupying different regions of the plot. For our use, models having best true positive rate, are favoured. This analysis further supported K-Nearest Neighbours approach as the best solution in terms of average performance. It was also found that augmentation of 3D descriptors for training using KNN algorithm provided relatively most substantial improvements in the models' fit (**Fig. 8**) as indicated by occupancy of more favoured areas of the sensitivity versus specificity plot by the resulting predictive models (**Fig. 11** vs **Fig. 12**).

Discussion

The goal of this study was to find out if QSAR, a chemoinformatics approach, is feasible for answering questions of Systems Biology of Small Molecules. Combination of data mining and computational chemistry modelling was proved working together as intended by production of several good quality models. This study has also shed light onto which of the available machine learning algorithms perform best in this setting and which compound features (types of descriptors) provide best basis for training of models, namely KNN and ANN in combination with Dragon 6.0 descriptors. However, the approach is not flawless and contains a big area for improvements within. Better mapping strategies, inclusion of hierarchy information of adverse

effects in the modelling process, reduction of redundancy due to highly similar molecules, definition of optimal distance of molecules from the model, to mention just a few potential traits to explore in follow-up studies.

In the future, when the method is refined, one could implement this approach into an easy to use tool that would, given a structure of novel compound, be able to give reliable predictions about potential adverse effects it may cause for prioritization in testing. Ideally, one would like to predict very reliably, however, a living organism response to a compound is often too complex to predict solely based on molecular structure. Into play come genetic makeup and environment, 2 extremely important factors that are hard to account for in QSAR. Nevertheless, our attempts give better results day by day as the state of the art improves and we thereby hope in many beneficial uses for this rather successful approach.

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

Chagoyen, M. and Pazos, F. (2011) '**MBRole: enrichment analysis of metabolomic data**' *Bioinformatics* 27: 730-731

Angell M. (2004) **The Truth About Drug Companies** *Random House, 1st edition*