

# Initial OpenTox Evaluation of ToxCast Phase 1 Datasets

ToxCast Data Analysis Summit

US EPA

Research Triangle Park, North Carolina, USA

14 - 15 May 2009

*Presented by Barry Hardy representing  
OpenTox Partners*



# Acknowledgements

- The OpenTox Community
- Nina and Vedrin Jeliaskova (Ideaconsult)
- Christoph Helma and Andreas Maunz (In Silico Toxicology and University of Freiburg)
- Romualdo Benigni (Istituto Superiore di Sanita)
- Haralambos Sarimveis (National Technical University of Athens)
- Vladimir Poroikov (Institute of Biomedical Chemistry of the Russian Academy of Medical Sciences)
- David Gallagher
- Nicki Douglas
- and all the group coworkers working on the project at [opentox.org](http://opentox.org)

# Warning

- The analysis summary presented here is a very preliminary one carried out by several OpenTox partners from their initial access to the project data in April 2009. No strong reliance on results and conclusions should be made at this point!
- Strong interest in understanding and developing the collaboration opportunities existing between OpenTox and ToxCast during phase 2 of the ToxCast project

# Presentation Outline

- About OpenTox
- Prediction of ToxRefDb *in vivo* data with existing models (including IZAS and ToxTree)
- Correlations between chemical structures, biological activities (predicted by PASS) and *in vitro* and *in vivo* ToxCast data
- Application of Pre-Processing, Feature Selection and Classification procedures to ToxCast datasets
- Data Management and Web Services approaches for access and manipulation of ToxCast data
- Impact of ToxCast on OpenTox Development and REACH-relevant risk assessment
- Recommendations and suggestions on datasets, assays and endpoints for ToxCast phase 2 data

# About OpenTox

The EC-funded FP7 project "OpenTox" commenced in September 2008 and is developing an Open Source-based integrating predictive toxicology framework that supports a unified access to toxicological data and (Q)SAR models. Initial research has defined the essential components of the framework architecture, approach to data access, schema and management, use of controlled vocabularies and ontologies, web service and communications protocols, and selection and integration of algorithms for predictive modeling. Analyses of use cases is in progress and includes cases for REACH-relevant risk assessment, chemical categorization and prioritisation, drug development, and food safety evaluation, with the resulting requirements guiding framework design and initial application development.



More Information at [Opentox.org](http://Opentox.org)

# OpenTox Partners

- Douglas Connect
- In Silico Toxicology
- Ideaconsult
- Istituto Superiore di Sanita'
- Technical University of Munich
- Albert Ludwigs University Freiburg
- National Technical University of Athens
- David Gallagher
- Institute of Biomedical Chemistry of the Russian Academy of Medical Sciences
- Seascape Learning
- Fraunhofer Institute for Toxicology & Experimental Medicine



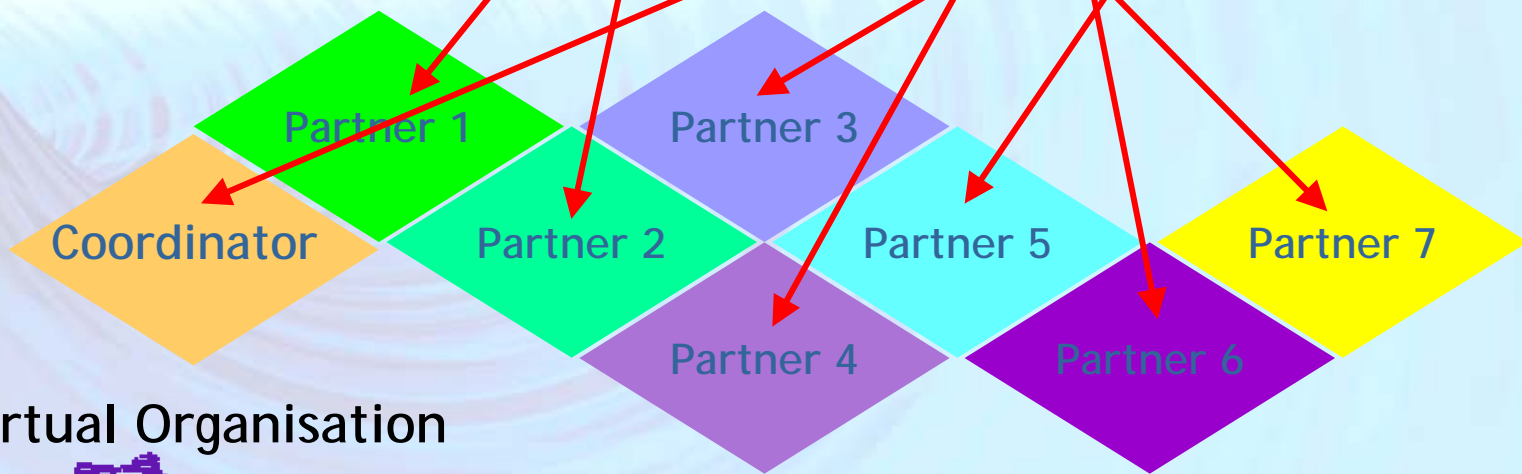
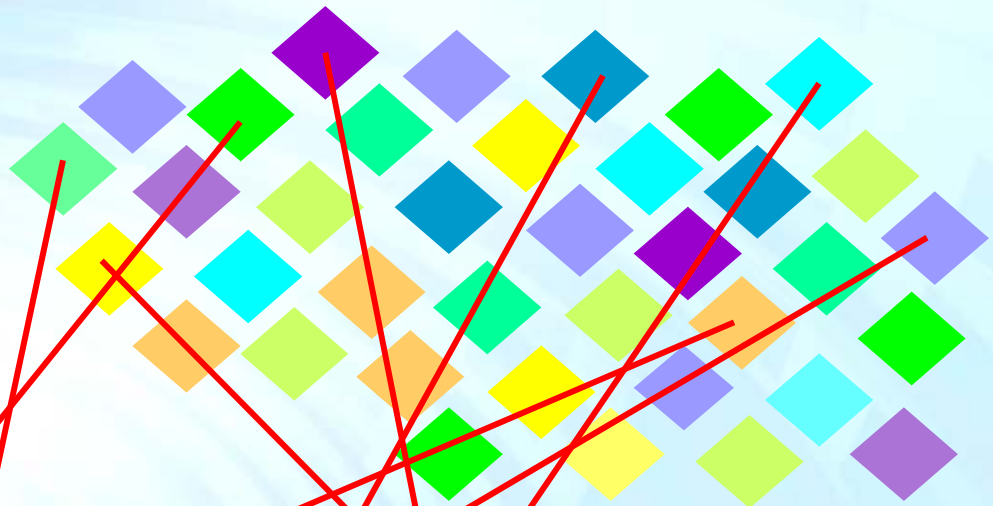
# OpenTox Advisory Board

- European Center for the Validation of Alternative Methods
- European Chemicals Bureau
- U.S Environmental Protection Agency
- U.S. Food & Drug Administration
- Nestle
- Roche
- AstraZeneca
- Lhasa
- Leadscope
- University of North Carolina
- EC Environment Directorate General
- Organisation for Economic Co-operation & Development
- CADASTER
- Bayer Healthcare

**Opportunity**

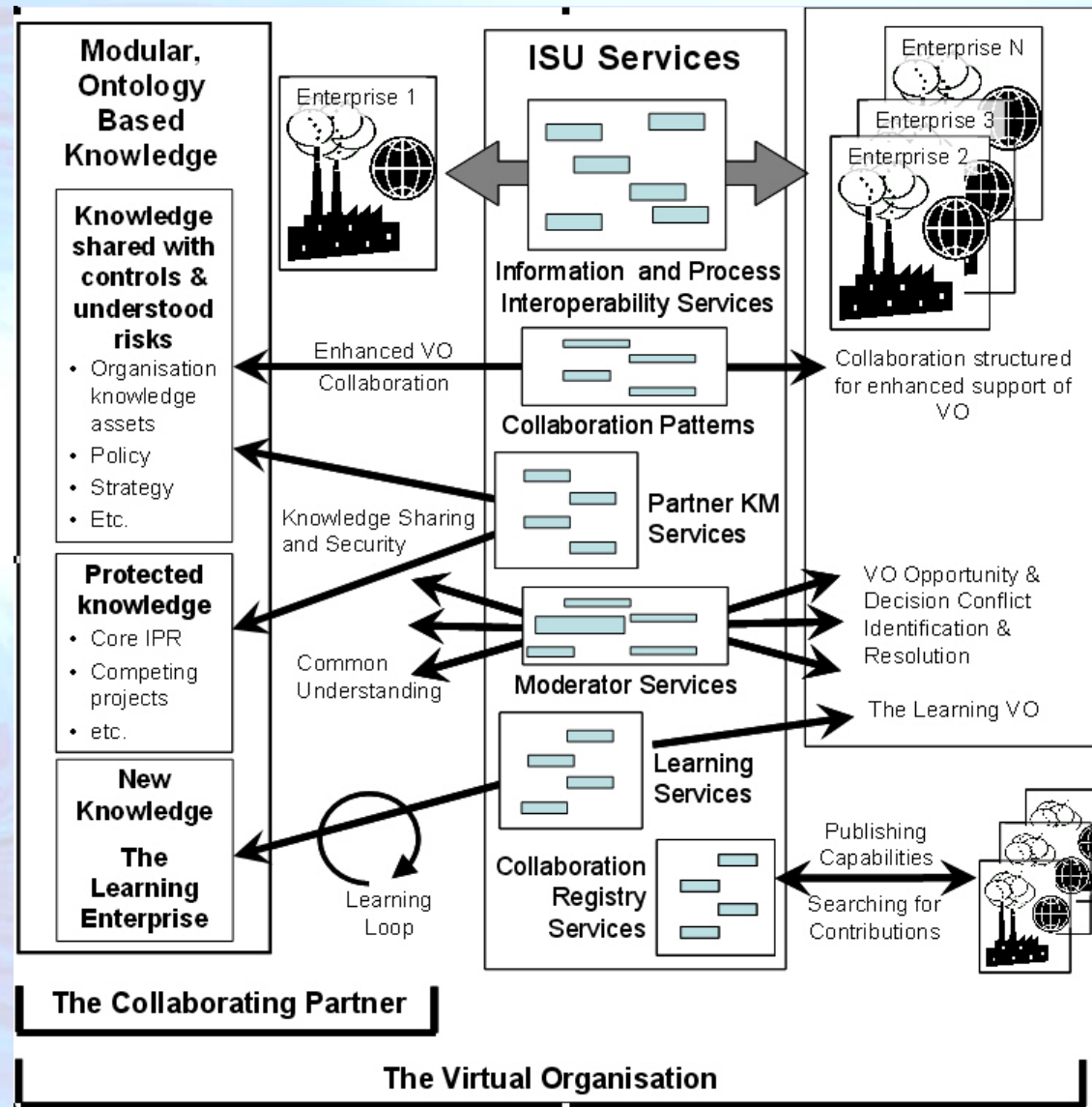


Call for Tender  
Need for joint effort  
Major project

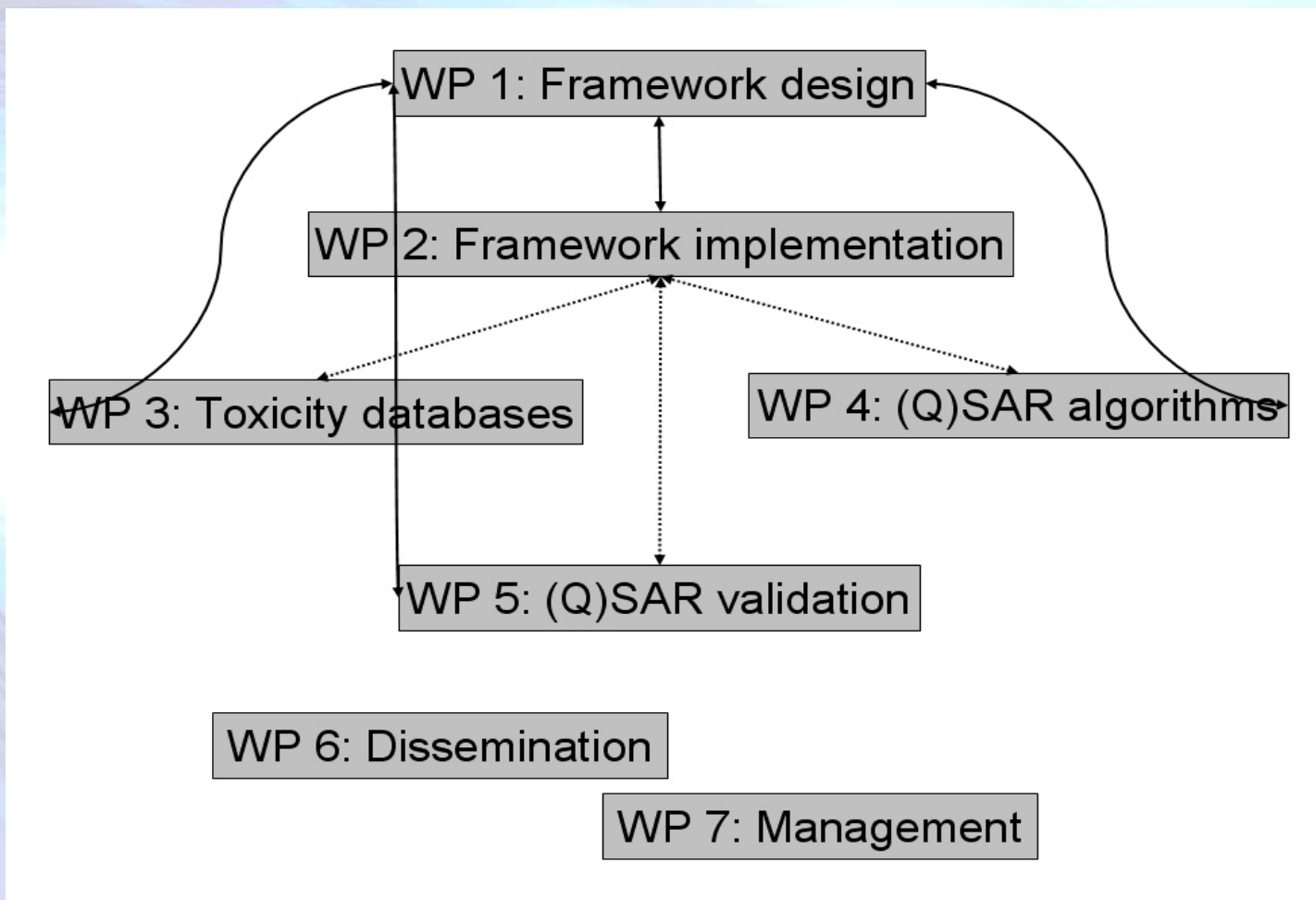


Virtual Organisation

# SYNERGY Collaboration Support Services



# OpenTox - Current Workpackages



# Prediction of ToxRefDb *in vivo* data with existing models (including Iazar and Toxtree)

- Toxtree::Benigni / Bossa rulebase (for mutagenicity and carcinogenicity) confusion matrix

– The Benigni / Bossa rulebase for mutagenicity and carcinogenicity - a module of Toxtree", by R. Benigni, C. Bossa, N. Jeliakova, T. Netzeva, and A. Worth. European Commission report EUR 23241 EN

No alerts for carcinogenic activity	CHR_Mouse_Tumorigen	Count
NO	Active	35
YES	Active	60
NO	Inactive	64
YES	Inactive	87

correct predictions  
in green

# Prediction of ToxRefDb *in vivo* data with existing models (including Iazar and Toxtree)

- Toxtree::Benigni / Bossa rulebase (for mutagenicity and carcinogenicity) confusion matrix

– The Benigni / Bossa rulebase for mutagenicity and carcinogenicity - a module of Toxtree", by R. Benigni, C. Bossa, N. Jeliakova, T. Netzeva, and A. Worth. European Commission report EUR 23241 EN

Structural Alert for genotoxic carcinogenicity	CHR_Mouse_Tumorigen	Count
NO	Active	70
YES	Active	25
NO	Inactive	100
YES	Inactive	51

correct predictions  
in green

# Prediction of ToxRefDb *in vivo* data with existing models (including Iazar and Toxtree)

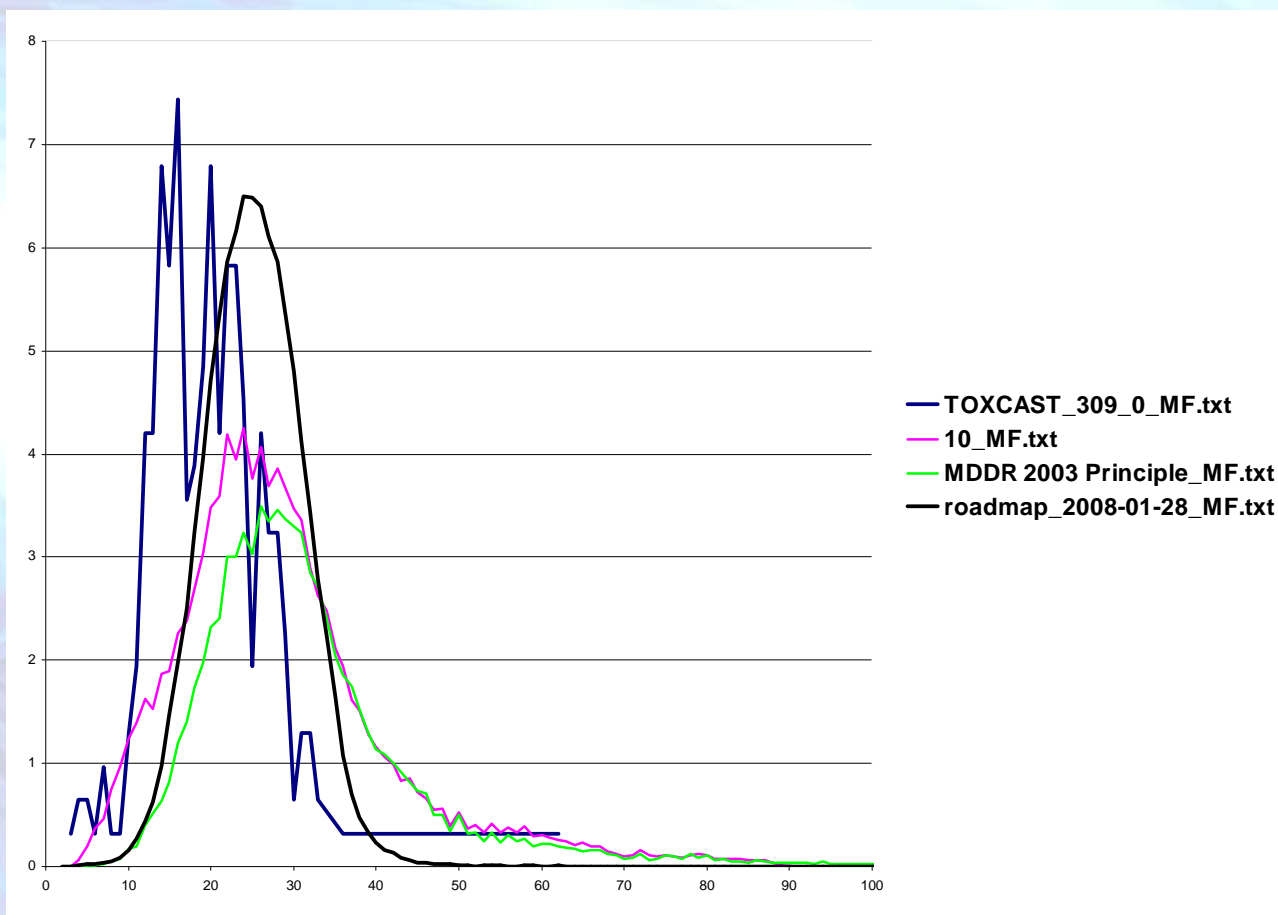
- Toxtree::Benigni / Bossa rulebase (for mutagenicity and carcinogenicity) confusion matrix

– The Benigni / Bossa rulebase for mutagenicity and carcinogenicity - a module of Toxtree", by R. Benigni, C. Bossa, N. Jeliakova, T. Netzeva, and A. Worth. European Commission report EUR 23241 EN

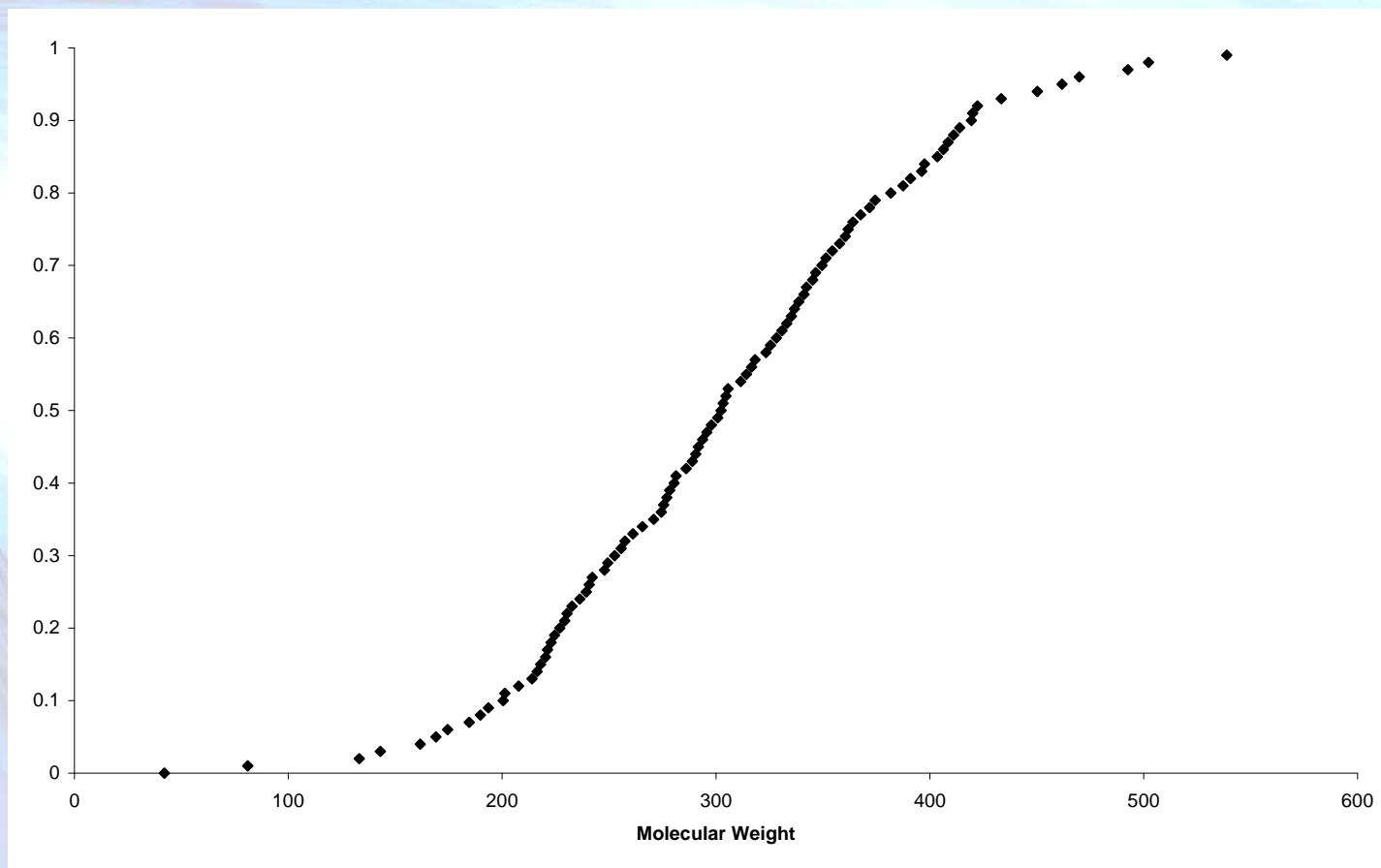
Structural Alert for nongenotoxic carcinogenicity	CHR_Mouse_Tumorigen	Count
NO	Active	85
YES	Active	10
NO	Inactive	138
YES	Inactive	13

correct predictions  
in green

# ToxCast Set : Distribution of Compounds vs. Non-Hydrogen Atoms' Amount



# ToxCast Set: Distribution of Compounds vs. Molecular Weights



# ToxCast Set: 157 Compounds Coincide with Molecules from PASS Training Set

The screenshot displays the ISIS/Base database interface for a specific compound. The main window shows the chemical structure of a pyridine derivative with a phosphorus-containing side chain. To the right of the structure, several data fields are visible:

ID	85
	40417
	C <sub>12</sub> H <sub>21</sub> N <sub>2</sub> O <sub>3</sub> PS
	304.3506
ToxDose_mkM	9.1028

Below the structure, a list of **PASS Activities** is provided:

- Teratogen
- Skin irritative effect
- Insecticide
- CYP1 substrate
- CYP1A substrate
- CYP1A2 substrate
- Non mutagenic, Salmonella
- Eye irritation, high
- Skin irritation, moderate
- CYP2 substrate
- CYP3A substrate
- CYP3A4 substrate
- CYP2C substrate
- CYP2B substrate
- CYP2B1 substrate
- CYP2B6 substrate
- CYP2C19 substrate
- CYP2B2 substrate
- CYP2C11 substrate

The interface includes a menu bar (File, Edit, Options, Object, Database, Search, List, Window, Help) and a toolbar with buttons for Forms, Query, Browse, and Update. The search domain is set to 'All' and the current view shows 50 of 157 records.

# Carcinogenicity Prediction by PASS

With the trained PASS program we predicted carcinogenicity for 306 compounds from ToxRefDB. Four compounds that have two components were excluded from the prediction. For 71 compounds mouse data were not available; for 62 compounds rat data were not available.

	NA	TP	TN	FP	FN	Sensitivity	Specificity	Accuracy
CHR_Mouse_LiverTumors	71	31	108	41	64	0.33	0.72	0.57
CHR_Mouse_LungTumors	71	11	138	11	84	0.12	0.93	0.61
CHR_Mouse_Tumorigen	71	25	129	41	49	0.34	0.76	0.63
CHR_Rat_LiverTumors	62	8	159	28	58	0.12	0.85	0.66
CHR_Rat_TesticularTumors	62	12	133	101	7	0.63	0.57	0.57
CHR_Rat_ThyroidTumors	62	3	212	20	18	0.14	0.91	0.85
CHR_Rat_Tumorigen	62	8	184	46	15	0.35	0.80	0.76

NA - data not available ; TP - true positive; TN - true negative; FP - false positive; FN - false negative



Sensitivity -  $TP/(TP+FN)$ ;

Specificity -  $TN/(TN+FP)$ ;

Accuracy -  $(TP+TN)/(TP+TN+FP+FN)$

# More Information..

For more information on PASS Analysis please see supplementary information and visit their poster at the meeting:

## Abstract 53

### (Q)SAR and (Q)AAR analysis of ToxCast Dataset Using PASS and GUSAR approaches

Vladimir Poroikov, Dmitry Filimonov, Alexey Zakharov, Alexey Lagunin,  
Sergey Novikov\*

Institute of Biomedical Chemistry of Rus. Acad. Med. Sci.;

\*A.N. Sysin Institute of Human Ecology and Environmental Health of  
Rus. Acad. Med. Sci., 10, Pogodinskaya Str., Moscow, 11912, Russia

# lazar predictions of mouse carcinogenicity

ToxRefDb	lazar		
	inactive	active	
inactive	111	40	151
active	66	29	95
	177	69	246

Sensitivity 0.31  
Specificity 0.74  
Accuracy 0.57

Andreas Maunz<sup>1</sup>  
Christoph Helma<sup>1,2</sup>

<sup>1</sup>) FDM Universität Freiburg (D)  
<sup>2</sup>) in-silico toxicology Basel (CH)

# lazar prediction of mouse carcinogenicity within applicability domain

ToxRefDb	lazar					
	inactive	active				
inactive	70	17	87			
active	45	10	55			
	115	27	142			
Sensitivity	0.182					
Specificity	0.805					
Accuracy	0.563					

# Lazar observations

- Poor sensitivity of Lazar predictions and ToxCast carcinogenicity results
- Specificity increases for compounds within the applicability domain, sensitivity decreases
- CPDB/Toxcast carcinogenicity concordance 72 % (rat) and 76 % (mouse), low sensitivity (~ 40%)

# Conclusions - possible explanations for prediction failures

- Structural dissimilarities between ToxCast and CPDB compounds: unlikely (in this case compounds within AD would perform better)
- Different experimental protocols/evaluation schemes: possible (low sensitivity of CPDB vs. ToxCast)
- Linear fragments miss crucial properties of ToxCast structures: possible (improvements should be possible with more expressive descriptors, e.g. BBRC analysis from Andreas Maunz)
- ToxCast compounds act by special mechanisms that are not covered by CPDB compounds: possible

# Conclusions

- Linear fragments miss crucial properties of ToxCast structures: possible (improvements should be possible with more expressive descriptors, e.g. BBRC from A. Maunz)
- ToxCast compounds act by special mechanisms that are not covered by CPDB compounds: possible

Andreas Maunz<sup>1</sup>, Christoph Helma<sup>1,2</sup>, and Stefan Kramer<sup>3</sup>:

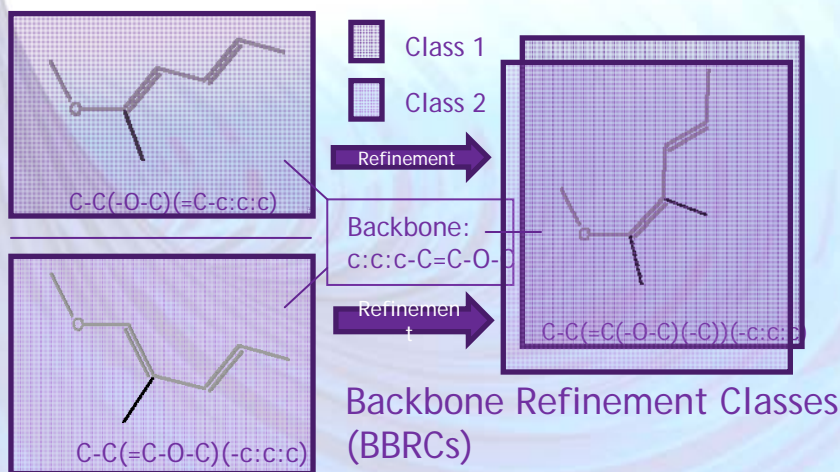
# Large-Scale Graph Mining using Backbone Refinement Classes

In KDD '09: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

Mining structurally diverse 2D-descriptors from large class-labelled graph databases.

Specialize on tree-shaped fragments!

- Efficient to mine.
- Considers branched substructures.
- **Method: Backbone Refinements** partition the search space **structurally** in contrast to open/closed fragments.



Mine most significant representative of classes

## BBRC-Representatives:

- Significantly improve accuracy in classification tasks compared to open/ closed fragments.
  - *Sensitivity >75% for carcinogenicity*
- Drastically reduce feature set sizes and running times (dynamic vs. static upper bound pruning).
  - *23,400 compounds in <5min, yielding 31,450 descriptors.*
- yield high descriptor coverage despite high min. frequencies.

C++ library implementation:

<http://www.maunz.de/libfminer-doc>

- 1) FDM Universität Freiburg (D)
- 2) in-silico toxicology Basel (CH)
- 3) Technische Universität München (D)

# Sensitivity

- Typical example: DEV\_Rat\_General\_GeneralFetal Pathology (10-fold Crossvalidation)
- Randomized, *not balanced* (n = 129 + 40):

Acc: 84.94 %  
Spec: 94.52 %  
Sens: 43.50 %

Classified: inactive	active	
207	12	inactive
27	13	active

- Randomized, *balanced by downsampling* (n = 41+ 40):

Acc: 77.78 %  
Spec: 63.42 %  
Sens: 92.50 %

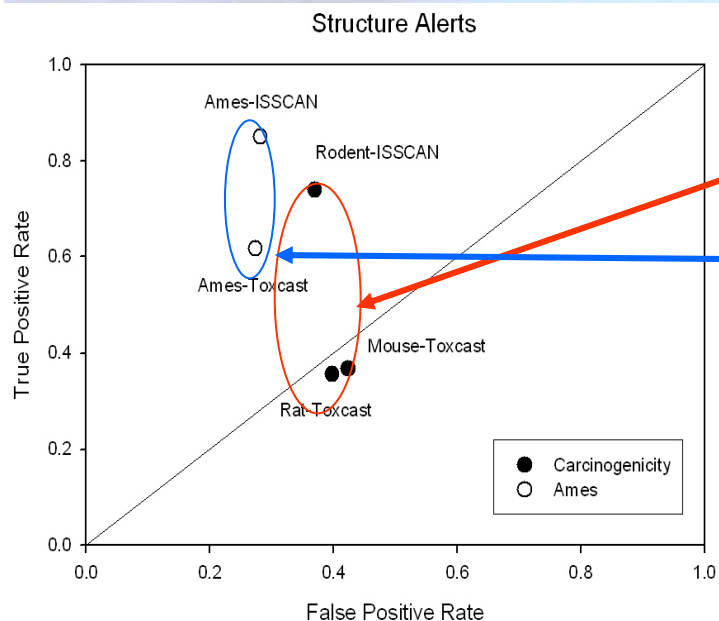
Classified: inactive	active	
26	15	inactive
3	37	active

# Preliminary evaluation of Toxcast Phase 1: sketching the landscape

Romualdo Benigni, Cecilia Bossa, Alessandro Giuliani, and Ann M. Richard<sup>1</sup>

Istituto Superiore di Sanita' - Rome Italy; <sup>1</sup> US Environmental Protection Agency, Research Triangle Park

## Characterization of Toxcast carcinogenicity data: Mechanisms of carcinogenesis



The **SAs for carcinogenicity / mutagenicity** in Toxtree code for genotoxic and (a few) nongenotoxic mechanism of carcinogenesis.

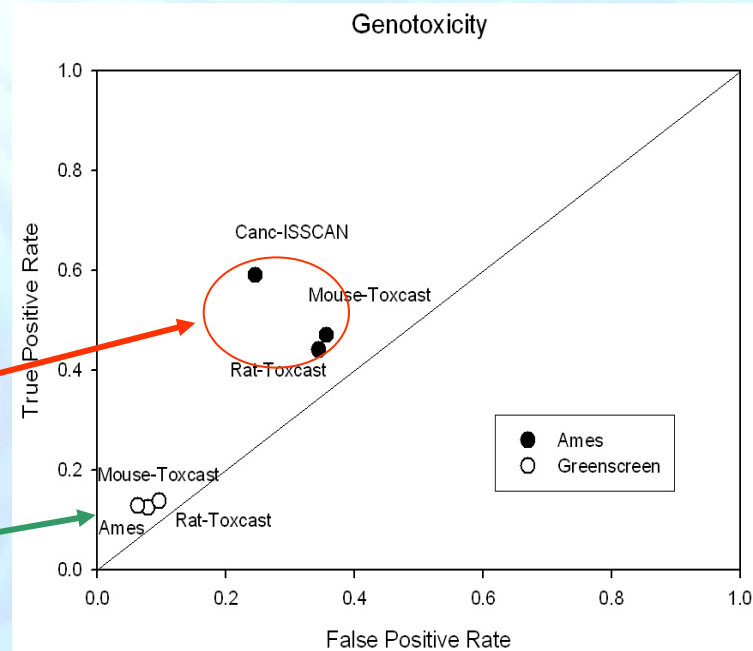
The SAs correlate with carcinogenicity in ISSCAN v3a ( $\chi = 152.6149$ ;  $p < 0.0001$ ), but not in Toxcast (mouse:  $\chi = 0.7448$   $p = 0.3881$ ; rat:  $\chi = 0.4368$   $p = 0.5086$ ).

The SAs correlate with the Ames test both in ISSCAN v3a ( $\chi = 203.5227$ ;  $p < 0.0001$ ) and Toxcast ( $\chi = 12.88$ ;  $p = 0.0003$ )

The **Ames test** is an experimental model for genotoxic carcinogenicity.

It correlates with carcinogenicity in ISSCAN v3a ( $\chi = 93.3330$ ,  $p < 0.0001$ ), but not in Toxcast (mouse:  $\chi = 1.0012$ ,  $p = 0.3170$ ; rat:  $\chi = 0.6808$ ,  $p = 0.4093$ ).

In Toxcast, the genotoxicity assay **Greenscreen** does not correlate neither with carcinogenicity (mouse:  $\chi = 1.4535$ ,  $p = 0.2280$ ; rat  $\chi = 1.1041$ ,  $p = 0.2934$ ), nor with the Ames test ( $\chi = 1.4299$ ,  $p = 0.2318$ ).

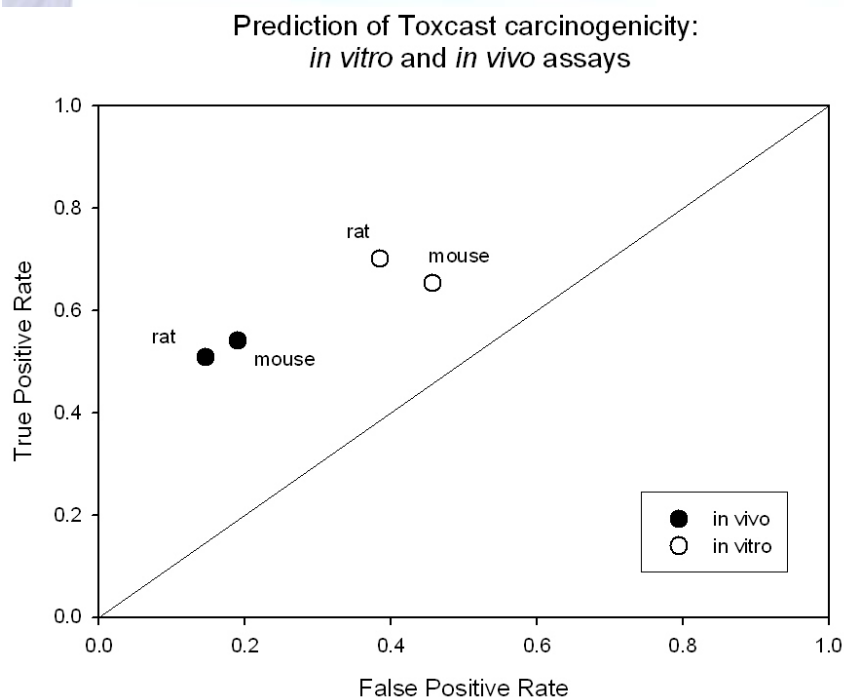


Toxcast database appears to be different from the general carcinogenicity database (e.g., ISSCAN) in many respects. In particular, the lack of correlation between Toxcast carcinogenicity data, and both the SAs and the Ames test points to a poor presence in Toxcast of the mechanisms of carcinogenesis predominant in the general carcinogenicity database.

## Do the Toxcast *in vitro* and *in vivo* results correlate with the Toxcast carcinogenicity data ? The general trend.

From the statistical analysis (Cluster and Principal Component Analysis, Stepwise Discriminant Analysis) of *in vitro* and *in vivo* toxicity assays, it appears that the *in vivo* toxicity measures correlate with rodent carcinogenicity to a level (around 20% explained variance) considerably higher than that of the *in vitro* assays (5 to 10% explained variance).

The pattern of *in vivo* toxicity variables entered into the model has no direct link with, and explanatory power for the carcinogenicity process. On the contrary, they should be considered -as a whole- as probes for ADME effects typical of the whole animal.



The largest difference between *in vivo* and *in vitro* assays is specificity (X-axis, False Positive rate =  $1 - \text{specificity}$ ).

The *in vitro* assays have a very low specificity (many false positives). Thus, they point only to potential effects.

The higher specificity of the *in vivo* toxicity assays can be attributed to their ability to discriminate the chemicals in terms of ADME characteristics (thus discriminating between potential and actual carcinogenicity).

## Conclusions from ISSCAN Analysis and Comparisons

The Toxcast carcinogenicity data are different from the classic carcinogenicity database in many respects. This peculiarity demonstrates that yet, the chemical biological interactions leading to carcinogenicity have not been sufficiently explored in large regions of the chemical space.

Thus the study of the Toxcast data may lead to new rules and structure alerts.

The overriding importance of in vivo phenomena in cancer has been confirmed. The in vitro assays appear to have quite a low explanatory and predictive power in respect to rodent carcinogenicity. This result indicates that the basic cellular processes coded for by the in vitro assays, are limitedly correlated with the biology of carcinogenesis, since they do not take into account crucial phenomena taking place at the level of organs and whole body organization.

The development of models for in vivo ADME phenomena should be considered as a priority.

More refined analyses are necessary to better qualify the present results. It should be emphasized that models providing mechanistic clues - and not only statistical measures- are to be preferred.

## More Information..

For more information on the ISS Analysis please visit their poster at the meeting:

Abstract 23

Preliminary Evaluation of ToxCast Phase 1: Sketching the Landscape

Romualdo Benigni, Cecilia Bossa, Alessandro Giuliani  
Istituto Superiore di Sanita - Rome, Italy

# Correlation Analysis (IST)

When an initial correlation analysis was carried out between the *in vitro* and *in vivo* datasets the following observations were made:

- 14 *in vivo* ToxRefDb endpoints had no correlated *in vitro* end points
- 118 *in vitro* assays had correlations with *in vivo* ToxRefDb endpoints

(See supplementary information for details.)

# SVM QSAR Models for the prediction of CHR\_Mouse\_KidneyPathology

The procedure:

- Only the compounds for which some information (numerical value) for CHR\_Mouse\_KidneyPathology end-point is included in the TOXRefDB are taken into account (246 instances)
- The class 0 was assigned to all 199 non-toxic compounds (value of 1000000 in the database), while the class 1 was assigned to the remaining 47 compounds.
- The same procedure was followed for all in-vitro descriptors
- An arff file file was developed containing all the in-vitro and chemical descriptors and the end-point values for 246 instances.
- The Weka library was used to select a subset among the 990 variables. Most algorithms failed, but the BestFirst method with the CfsSubsetEval attribute evaluator selected 39 descriptors. All of them contain only discrete 0 or 1 values.
- The data file containing only the 39 descriptors was used to develop an SVM model. The SVM method was validated by Leave-One-Out cross-validation and by splitting the data into training and validation files. Initially, we assumed the first 200 compounds as training examples and the remaining as test examples. We also used several random partitions into 200 training and 46 validation data was used.

# SVM QSAR Models for the prediction of CHR\_Mouse\_KidneyPathology

The results using the linear kernel:

For the entire set:

=====  
Classification Statistics  
=====

Correctly Classified Molecules	: 229 out of 246
Incorrectly Classified Molecules	: 17
Overall Success Rate	: 93.0895%
Error Percentage	: 6.910500000000001%

=====  
Confusion Matrix  
=====

*0*	*1*		<== Correct Class
196	14	*0*	
3	33	*1*	

=====  
Per-class Success Rates (%)  
=====

Class 0 -->	98.492
Class 1 -->	70.213



Work of Haralambos Sarimveis and Georgia Melagraki  
National Technical Univeristy of Athen

# SVM QSAR Models for the prediction of CHR\_Mouse\_KidneyPathology

The results using the linear kernel:

Cross validation:

```
=====  
Classification Statistics  
Correctly Classified Molecules : 210 out of 246  
Incorrectly Classified Molecules : 36  
Overall Success Rate : 85.3658%  
Error Percentage : 14.6342%
```

```
=====  
Confusion Matrix  
*0*      *1*      <== Correct Class  
197      34      | *0*  
2        13      | *1*
```

```
=====  
Per-class Success Rates (%)  
Class 0 --> 98.995  
Class 1 --> 27.659
```



Work of Haralambos Sarimveis and Georgia Melagraki  
National Technical Univeristy of Athen

# SVM QSAR Models for the prediction of CHR\_Mouse\_KidneyPathology

The results using the linear kernel:

For the validation set that contains the last 46 examples:

```
=====  
Classification Statistics =====  
Correctly Classified Molecules      : 36 out of 46  
Incorrectly Classified Molecules    : 10  
Overall Success Rate                : 78.26089999999999%  
Error Percentage                    : 21.7391%  
=====  
Confusion Matrix =====  
*0*      *1*      | *0*      <== Correct Class  
33       10       | *0*  
0        3        | *1*  
=====  
Per-class Success Rates (%) =====  
Class 0 --> 100.0  
Class 1 --> 23.077
```



Work of Haralambos Sarimveis and Georgia Melagraki  
National Technical Univeristy of Athen

# Prediction of ToxRefDb *in vivo* data with existing models (including Iazar and Toxtree)

- Toxtree::Cramer tree (threshold of toxicological concern)

– Cramer G. M., R. A. Ford, R. L. Hall, Estimation of Toxic Hazard - A Decision Tree Approach, J. Cosmet. Toxicol., Vol.16, pp. 255-276, Pergamon Press, 1978

Low (Class I)	Intermediate (Class II)	High (Class III)
9	5	305

# Prediction of ToxRefDb *in vivo* data with existing models (including Iazar and Toxtree)

- Toxtree::Verhaar scheme (toxicity mode of action)

- Verhaar H.J.M., Van Leeuwen C., Hermens J.L.M., Classifying Environmental Pollutants. 1: Structure-Activity Relationships for Prediction of Aquatic Toxicity, Chemosphere, Vol.25, No.4, pp.471-491, 1992.

Class 1 (narcosis or baseline toxicity)	Class 2 (less inert compounds)	Class 3 (unspecific reactivity)	Class 4 (compounds and groups of compounds acting by a specific mechanism)	Class 5 (Not possible to classify according to these rules)
1	5	27	1	284

## Prediction of ToxRefDb *in vivo* data with existing models (including Iazar and Toxtree)

- Toxtree::START (Structural Alerts for Reactivity in Toxtree – biodegradation & persistence decision tree)

Class 1 (easily biodegradable chemical)	Class 2 (persistent chemical)	Class 3 (unknown biodegradability)
131	154	34

The START rulebase estimates potential biodegradability or environmental persistence, by using a series of Structural Alerts (SAs) in combination with a decision tree. If the substance contains degradable functional groups, it usually can be considered as having the potential to degrade in the compartments to which it partitions. Depending on the weight of evidence that can be gathered on the potential for degradation, a determination can be made as to whether a substance is persistent according to the Persistence and Bioaccumulation Regulations (Government of Canada 2000).

NOTE: New expt data on biodegradability of ToxCast set is needed to test predictions.

## More Information on Classification Work..

For more information on Classification work there is a talk at the meeting on Friday morning from Nina Jeliaskova:

Abstract 15

Hierarchical Multi-label Classification of ToxCast Datasets

Nina Jeliaskova and Vedrin Jeliaskov

Ideaconsult Ltd., Angel Kanchev Str 4, 1000 Sofia, Bulgaria

## Prediction of ToxRefDb *in vivo* data with existing models (including lazarus and Toxtree)

lazarus predictions, covered endpoints:

- Carcinogenicity:
  - Multi cell call
  - Mouse
  - Rat
- Salmonella Mutagenicity (Kazius/Bursi dataset)
- Human Liver Toxicity
- Fathead minnow LC50
- Maximum recommended daily dose

## Prediction of ToxRefDb *in vivo* data with existing models (including IZAR and ToxTree)

- Many of the existing models perform poorly when predicting ToxRefDb *in vivo* data, especially in terms of “false negatives” and “applicability domain”;
- There is clear evidence that new models should be developed, taking into account the classes of chemicals in ToxRefDb as well as peculiarities of the *in vitro* data like complexity in data set, skewed distributions, etc.
- ToxRefDb currently does not provide data for some important endpoints like “toxicity mode of action” or “biodegradability”; it would be nice if such data could be gathered and provided in the future.

# Application of Pre-Processing, Feature Selection and Classification procedures to ToxCast datasets

- Need for feature selection, pre-processing and prediction / classification techniques that are suitable for:
  - High dimensional data;
  - Missing values;
  - Analysis of rare events
    - Most *in-vitro* data is distributed like ~10% active and 90% inactive

# Data Management and Web Services approaches for access and manipulation of ToxCast data

- Data management
  - Provide more convenient access than text files
  - Provide easy access to different slices of the data (a module for Toxcast data RESTful access would be a nice example)
  - Data exploration, visualization
  - Ontology
    - Map ToxCast terms into an ontology or existing data formats
    - Example:
      - CHR\_Rat\_Tumorigen
        - » is "Endpoint"
        - » is "Chronic Endpoint"
        - » is "Endpoint for All neoplastic lesions"
        - » target (any target)
        - » species rat
    - Domain experts knowledge required

# ToxCast Data Analysis

Before QSAR descriptors can be calculated reliably, chemical structures need to be checked for chemical correctness, ionization state, stereochemistry, consistency, duplicates, and conformation, etc. Also the data needs to be checked for range and evenness of spread.

For high-throughput calculations and for novice users this process needs to be automated in a consistent and chemically-intelligent way.

The ToxCast samples highlight a number of issues, potential errors and inconsistencies, which could potentially affect a QSAR correlation. A number of points are summarized in the supplementary slides by David Gallagher.

**OpenTox services will automate data cleanup, integration and reduce time for dataset and run preparation as far as practical ... and will strive to apply to new ToxCast datasets.**



# Collaboration, OpenTox Development and REACH risk assessment: OpenTox - CADASTER Collaboration

The FP7-funded CADASTER project (<http://www.cadaster.eu/>) will provide practical guidance to integrated risk assessment by carrying out a full hazard and risk assessment for industrial chemicals. The project will develop a Decision Support System that will be updated on a regular basis in order to accommodate and integrate emerging practices and procedures for alternative non-animal based testing methods. OpenTox and CADASTER partners will collaborate closely so as to promote and develop common practices, standards and procedures in the area of *in silico* based predictive toxicology approaches responding to user requirements in the area of REACH-relevant risk assessment. The collaboration should enable the development of a leading platform supporting the safety evaluation and regulatory compliance needs of industry operating in the European marketplace.



# CADASTER

## Goals:

Exemplify the integration of information, models, strategies for safety-, hazard-, risk assessment for large numbers of substances

Carry out “real” risk assessment for large numbers of substances according to the basic philosophy of REACH: < costs, animal testing, time

Exemplify how to increase non-testing information whilst quantifying and reducing uncertainty

See Poster 52, Igor Tetko

EU project CADASTER: Case studies on the Development and Application of in-Silico Techniques for Environmental hazard and Risk assessment

# CADASTER

## Aim:

Provide full environmental hazard and risk assessment according to the REACH philosophy for chemicals belonging to 4 classes of emerging chemicals:

- 1 – Polybrominated diphenylethers (PBDE), typically class of hydrophobic chemicals that pose a threat to man and the environment.
- 2 - Perfluoroalkylated substances and their transformation products, like perfluoroalkylated sulfonamides, alkanolic acids, sulfonates. Persistent hydrophilic compounds that may be toxic for man and environment.
- 3 – Substituted musks/fragrances; a heterogenic group of chemicals of varying composition like substituted benzophenones, polycyclic musks, terpene derivatives. Common emission pattern in the environment.
- 4 - Triazoles/benzotriazoles: increasingly used as pesticides and anti-corrosives.

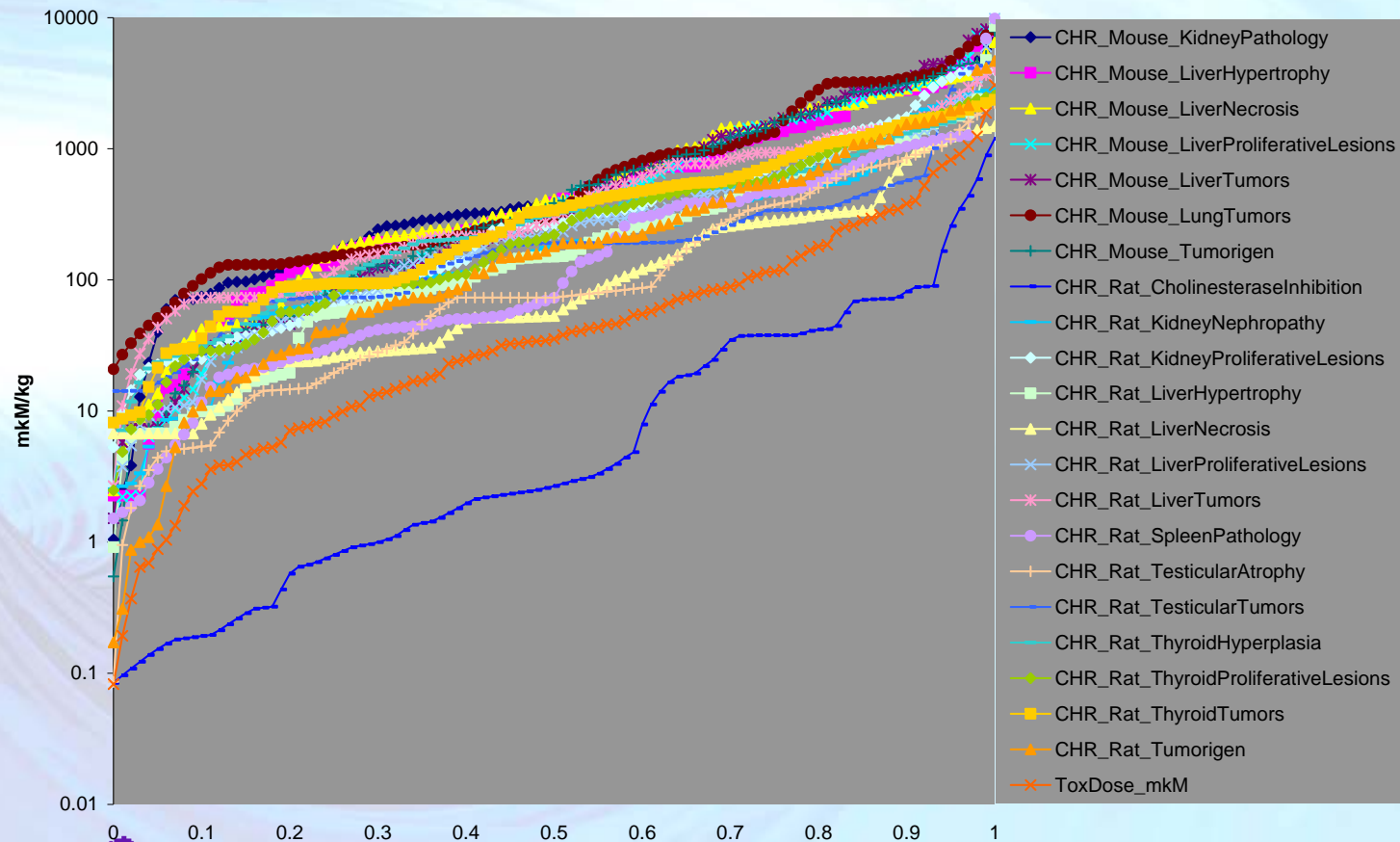
# Attempt of Creation of Integral Parameter Characterized the Compound's Hazard

Using dosage characteristics for all 75 end-points, experimental data for which were obtained in vivo, we calculated the ToxDose values for 283 compounds:

$$\text{ToxDose} = \frac{m}{n \sum_k \frac{1}{D_k}}$$

Here: D is the LEL value; m is the number of end-points for a particular compound; n is the total number of tests.

# Distribution of ToxDose Values for Different End-Points



# Results of PASS Training for Predicting Different Categories of ToxDose

No	Num	IEP, %	Activity Type
1	301	5.151	ToxCast
2	9	25.689	CHR_ToxDose < 1.0 mkM/kg
3	19	28.623	CHR_ToxDose < 3.16 mkM/kg
4	51	28.052	CHR_ToxDose < 10.0 mkM/kg
5	89	25.008	CHR_ToxDose < 31.6 mkM/kg
6	145	27.272	CHR_ToxDose < 100.0 mkM/kg
7	176	24.130	CHR_ToxDose < 316.0 mkM/kg
8	196	23.499	CHR_ToxDose < 1000.0 mkM/kg
9	15	33.091	CHR_ToxDose 0.316-3.16 mkM/kg
10	42	26.576	CHR_ToxDose 1.0-10.0 mkM/kg
11	70	28.502	CHR_ToxDose 3.16-31.6 mkM/kg
12	94	38.760	CHR_ToxDose 10.0-100.0 mkM/kg
13	87	33.645	CHR_ToxDose 31.6-316.0 mkM/kg
14	51	40.901	CHR_ToxDose 100.0-1000.0 mkM/kg
15	26	34.029	CHR_ToxDose 316.0-3160.0 mkM/kg



---

Num is the number of compounds in the training set; IEP is Independent Error of Prediction.

# Final words...

<http://www.opentox.org/>

Contact Information:  
Barry.Hardy -(at)-  
douglasconnect.com  
Tel: +41 61 851 0170



# Supplementary Information

Supplementary Information  
from our initial evaluation  
is provided in the  
following slides



# OpenTox Partners

Org	Name	Project Role
Douglas Connect	Barry Hardy	Project Coordinator, WP7 Leader
Douglas Connect	Nicki Douglas	LEAR, WP6 Leader, Marketing, Project Administration
In silico toxicology	Christoph Helma	LEAR, WP1 Leader, OpenTox Framework Design & Development
In silico toxicology	Andreas Maunz	Developer
In silico toxicology	Michael Rautenberg	Web Resources
In silico toxicology	Marek Kralewski	Web Resources
In silico toxicology	David Vorgrimmler	Developer

# OpenTox Partners

Org	Name	Project Role
Ideaconsult	Nina Jeliaskova	WP2 Leader, OpenTox Framework Design & Development
Ideaconsult	Vedrin Jeliaskov	LEAR, Developer
Ideaconsult	Luben Boyanov	Developer
Ideaconsult	Chelsea Jiang	Technical Writer
Ideaconsult	Martin Martinov	Developer
Istituto Superiore di Sanita	Romualdo Benigni	WP3 Leader, Ontologies and Data Resources
Istituto Superiore di Sanita	Olga Tcheremenskaia	Developer

# OpenTox Partners

Org	Name	Project Role
Technische Universität München (TUM)	Stefan Kramer	WP4 Leader, Algorithm Development
TUM	Ulrike Ronchetti	LEAR
TUM	Tobias Girschick	Developer
TUM	Fabian Buchwald	Developer
TUM	Jörg Wicker	Developer
Albert Ludwigs University Freiburg (ALU)	Andreas Karwath	WP5 Leader, Testing & Validation
ALU	Martin Gütlein	Developer, Testing & Validation

# OpenTox Partners

Org	Name	Project Role
National Technical University of Athens (NTUA)	Haralambos Sarimveis	NTUA Leader, Algorithm Development
NTUA	Georgia Melagraki	Algorithm Developer
NTUA	Antreas Afantitis	Algorithm Developer
NTUA	Yannis Polyzos	OpenTox CA Signatory
NTUA	Despina I. Alatopoulou	LEAR
David Gallagher	David Gallagher	DG Leader, User Requirements, Use Cases, Usability, LEAR
David Gallagher	Christina Golden	DG Deputy Leader

# OpenTox Partners

Org	Name	Project Role
Institute of Biomedical Chemistry of the Russian Academy of Medical Sciences	Vladimir Poroikov	IBMC Leader, OpenTox CA Signatory, Algorithm and Data Resource Development
IBMC	Dmitry Filimonov	IBMC Deputy Leader, Algorithm and Data Resource Development
IBMC	Alexey Zakharov	Algorithm and Data Resource Development
IBMC	Alexey Lagunin	Algorithm and Data Resource Development
IBMC	Tatyana Glorizova	Algorithm and Data Resource Development
IBMC	Sergey Novikov	Algorithm and Data Resource Development
IBMC	Skvortsova Natalia	Algorithm and Data Resource Development

# OpenTox Partners

Org	Name	Project Role
SL	Sunil Chawla	LEAR, SL Leader, Algorithm Project Leader
SL	Meera Suri	Deputy SL Leader
SL	Steve Bowlus	SL Consultant
SL-JNU	Indira Ghosh	SL Scientific Leader & Advisor, Algorithm Development
SL-JNU	Surajit Ray	Algorithm and Data Resource Development
SL-JNU	Gaurav Singhai	Algorithm and Data Resource Development
SL-JNU	Om Prakash	Advisor Algorithm Development

# OpenTox Partners

Org	Name	Project Role
Fraunhofer Institute for Toxicology & Experimental Medicine	Sylvia Escher	ITEM Leader, Ontologies & Data Resource Development
Fraunhofer Institute for Toxicology & Experimental Medicine	Sara Weiss	ITEM Deputy Leader, Ontologies & Data Resource Development

# Supplementary Information

## Toxtree Analysis

# Prediction of ToxRefDb *in vivo* data with existing models (including Iazar and Toxtree)

- Toxtree::Benigni / Bossa rulebase (for mutagenicity and carcinogenicity) confusion matrix

– The Benigni / Bossa rulebase for mutagenicity and carcinogenicity - a module of Toxtree", by R. Benigni, C. Bossa, N. Jeliazkova, T. Netzeva, and A. Worth. European Commission report EUR 23241 EN

No alerts for carcinogenic activity	CHR_Mouse_Tumorigen	Count
NO	Active	35
YES	Active	60
NO	Inactive	64
YES	Inactive	87

correct predictions  
in green

# Prediction of ToxRefDb *in vivo* data with existing models (including Iazar and Toxtree)

- Toxtree::Benigni / Bossa rulebase (for mutagenicity and carcinogenicity) confusion matrix

– The Benigni / Bossa rulebase for mutagenicity and carcinogenicity - a module of Toxtree", by R. Benigni, C. Bossa, N. Jeliazkova, T. Netzeva, and A. Worth. European Commission report EUR 23241 EN

Structural Alert for genotoxic carcinogenicity	CHR_Mouse_Tumorigen	Count
NO	Active	70
YES	Active	25
NO	Inactive	100
YES	Inactive	51

correct predictions  
in green

# Prediction of ToxRefDb *in vivo* data with existing models (including Iazar and Toxtree)

- Toxtree::Benigni / Bossa rulebase (for mutagenicity and carcinogenicity) confusion matrix

– The Benigni / Bossa rulebase for mutagenicity and carcinogenicity - a module of Toxtree", by R. Benigni, C. Bossa, N. Jeliakova, T. Netzeva, and A. Worth. European Commission report EUR 23241 EN

Structural Alert for nongenotoxic carcinogenicity	CHR_Mouse_Tumorigen	Count
NO	Active	85
YES	Active	10
NO	Inactive	138
YES	Inactive	13

correct predictions  
in green

# Prediction of ToxRefDb *in vivo* data with existing models (including Iazar and Toxtree)

- Toxtree::Benigni / Bossa rulebase (for mutagenicity and carcinogenicity) confusion matrix

– The Benigni / Bossa rulebase for mutagenicity and carcinogenicity - a module of Toxtree", by R. Benigni, C. Bossa, N. Jeliakova, T. Netzeva, and A. Worth. European Commission report EUR 23241 EN

Potential <i>S. typhimurium</i> TA100 mutagen based on QSAR	CHR_Mouse_Tumorigen	Count
NO	Active	94
YES	Active	1
NO	Inactive	146
YES	Inactive	5

correct predictions  
in green

# Prediction of ToxRefDb *in vivo* data with existing models (including Iazar and Toxtree)

- Toxtree::Benigni / Bossa rulebase (for mutagenicity and carcinogenicity) confusion matrix

– The Benigni / Bossa rulebase for mutagenicity and carcinogenicity - a module of Toxtree", by R. Benigni, C. Bossa, N. Jeliazkova, T. Netzeva, and A. Worth. European Commission report EUR 23241 EN

Unlikely to be a <i>S. typhimurium</i> TA100 mutagen based on QSAR	CHR_Mouse_Tumorigen	Count
NO	Active	87
YES	Active	8
NO	Inactive	132
YES	Inactive	19

correct predictions  
in green

# Prediction of ToxRefDb *in vivo* data with existing models (including Iazar and Toxtree)

- Toxtree::Benigni / Bossa rulebase (for mutagenicity and carcinogenicity) confusion matrix

– The Benigni / Bossa rulebase for mutagenicity and carcinogenicity - a module of Toxtree", by R. Benigni, C. Bossa, N. Jeliazkova, T. Netzeva, and A. Worth. European Commission report EUR 23241 EN

Potential carcinogen based on QSAR	CHR_Mouse_Tumorigen	Count
NO	Active	93
YES	Active	2
NO	Inactive	149
YES	Inactive	2

correct predictions  
in green

# Prediction of ToxRefDb *in vivo* data with existing models (including Iazar and Toxtree)

- Toxtree::Benigni / Bossa rulebase (for mutagenicity and carcinogenicity) confusion matrix

– The Benigni / Bossa rulebase for mutagenicity and carcinogenicity - a module of Toxtree", by R. Benigni, C. Bossa, N. Jeliazkova, T. Netzeva, and A. Worth. European Commission report EUR 23241 EN

Unlikely to be a carcinogen based on QSAR	CHR_Mouse_Tumorigen	Count
NO	Active	93
YES	Active	2
NO	Inactive	147
YES	Inactive	4

correct predictions  
in green

# Prediction of ToxRefDb *in vivo* data with existing models (including Iazar and Toxtree)

- Toxtree::Benigni / Bossa rulebase (for mutagenicity and carcinogenicity) confusion matrix

– The Benigni / Bossa rulebase for mutagenicity and carcinogenicity - a module of Toxtree", by R. Benigni, C. Bossa, N. Jeliakova, T. Netzeva, and A. Worth. European Commission report EUR 23241 EN

No alerts for carcinogenic activity	CHR_Rat_Tumorigen	Count
NO	Active	36
YES	Active	65
NO	Inactive	62
YES	Inactive	94

correct predictions  
in green

# Prediction of ToxRefDb *in vivo* data with existing models (including Iazar and Toxtree)

- Toxtree::Benigni / Bossa rulebase (for mutagenicity and carcinogenicity) confusion matrix

– The Benigni / Bossa rulebase for mutagenicity and carcinogenicity - a module of Toxtree", by R. Benigni, C. Bossa, N. Jeliazkova, T. Netzeva, and A. Worth. European Commission report EUR 23241 EN

Structural Alert for genotoxic carcinogenicity	CHR_Rat_Tumorigen	Count
NO	Active	73
YES	Active	28
NO	Inactive	113
YES	Inactive	43

correct predictions  
in green

# Prediction of ToxRefDb *in vivo* data with existing models (including Iazar and Toxtree)

- Toxtree::Benigni / Bossa rulebase (for mutagenicity and carcinogenicity) confusion matrix

– The Benigni / Bossa rulebase for mutagenicity and carcinogenicity - a module of Toxtree", by R. Benigni, C. Bossa, N. Jeliaskova, T. Netzeva, and A. Worth. European Commission report EUR 23241 EN

Structural Alert for nongenotoxic carcinogenicity	CHR_Rat_Tumorigen	Count
NO	Active	93
YES	Active	8
NO	Inactive	137
YES	Inactive	19

correct predictions  
in green

# Prediction of ToxRefDb *in vivo* data with existing models (including Iazar and Toxtree)

- Toxtree::Benigni / Bossa rulebase (for mutagenicity and carcinogenicity) confusion matrix

– The Benigni / Bossa rulebase for mutagenicity and carcinogenicity - a module of Toxtree", by R. Benigni, C. Bossa, N. Jeliakova, T. Netzeva, and A. Worth. European Commission report EUR 23241 EN

Potential <i>S. typhimurium</i> TA100 mutagen based on QSAR	CHR_Rat_Tumorigen	Count
NO	Active	100
YES	Active	1
NO	Inactive	151
YES	Inactive	5

correct predictions  
in green

# Prediction of ToxRefDb *in vivo* data with existing models (including Iazar and Toxtree)

- Toxtree::Benigni / Bossa rulebase (for mutagenicity and carcinogenicity) confusion matrix

– The Benigni / Bossa rulebase for mutagenicity and carcinogenicity - a module of Toxtree", by R. Benigni, C. Bossa, N. Jeliakova, T. Netzeva, and A. Worth. European Commission report EUR 23241 EN

Unlikely to be a <i>S. typhimurium</i> TA100 mutagen based on QSAR	CHR_Rat_Tumorigen	Count
NO	Active	87
YES	Active	14
NO	Inactive	143
YES	Inactive	13

correct predictions  
in green

# Prediction of ToxRefDb *in vivo* data with existing models (including Iazar and Toxtree)

- Toxtree::Benigni / Bossa rulebase (for mutagenicity and carcinogenicity) confusion matrix

– The Benigni / Bossa rulebase for mutagenicity and carcinogenicity - a module of Toxtree", by R. Benigni, C. Bossa, N. Jeliazkova, T. Netzeva, and A. Worth. European Commission report EUR 23241 EN

Potential carcinogen based on QSAR	CHR_Rat_Tumorigen	Count
NO	Active	98
YES	Active	3
NO	Inactive	155
YES	Inactive	1

correct predictions  
in green

# Prediction of ToxRefDb *in vivo* data with existing models (including Iazar and Toxtree)

- Toxtree::Benigni / Bossa rulebase (for mutagenicity and carcinogenicity) confusion matrix

– The Benigni / Bossa rulebase for mutagenicity and carcinogenicity - a module of Toxtree", by R. Benigni, C. Bossa, N. Jeliaskova, T. Netzeva, and A. Worth. European Commission report EUR 23241 EN

Unlikely to be a carcinogen based on QSAR	CHR_Rat_Tumorigen	Count
NO	Active	98
YES	Active	3
NO	Inactive	153
YES	Inactive	3

correct predictions  
in green

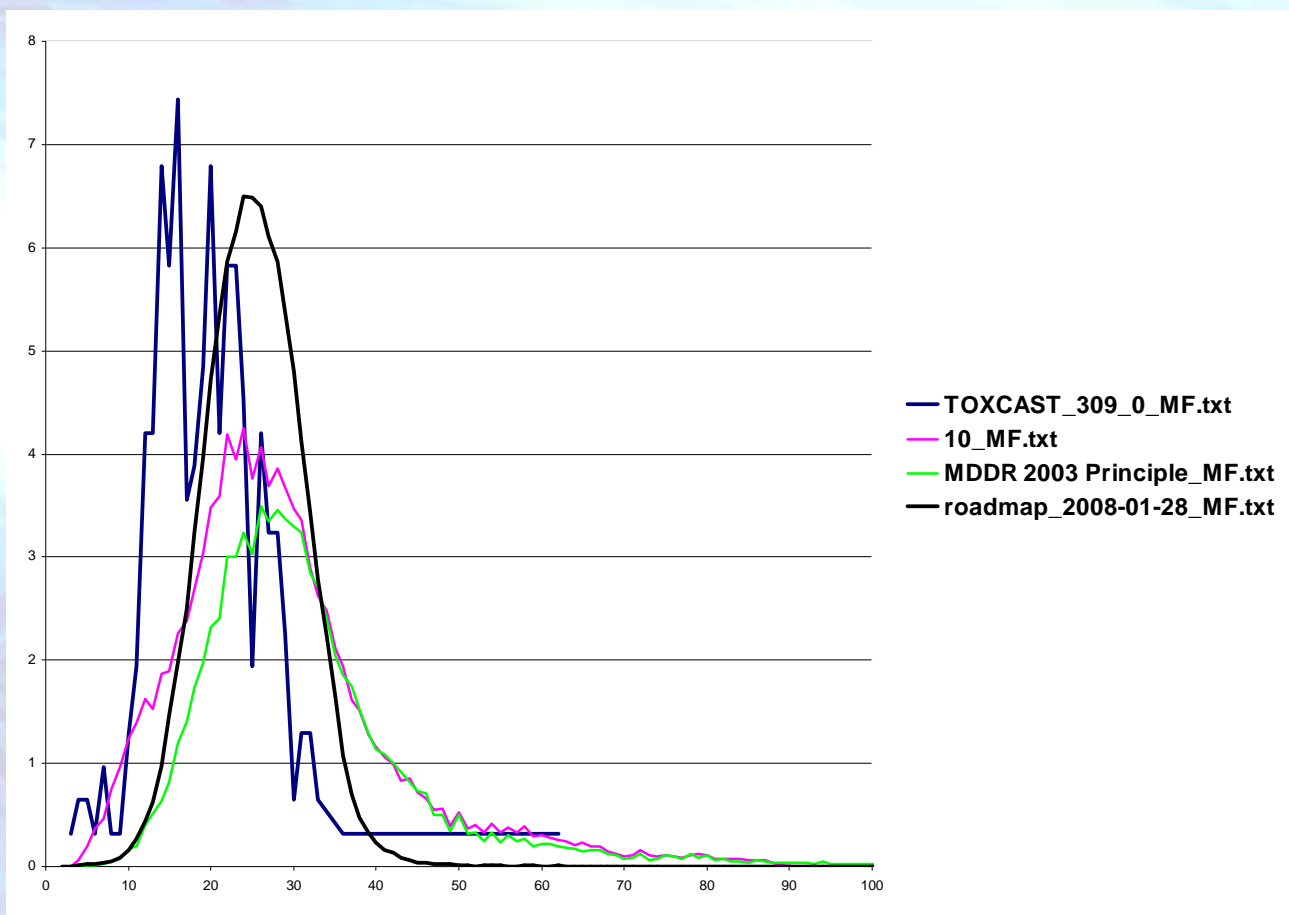
# Supplementary Information

## IBMC Analysis

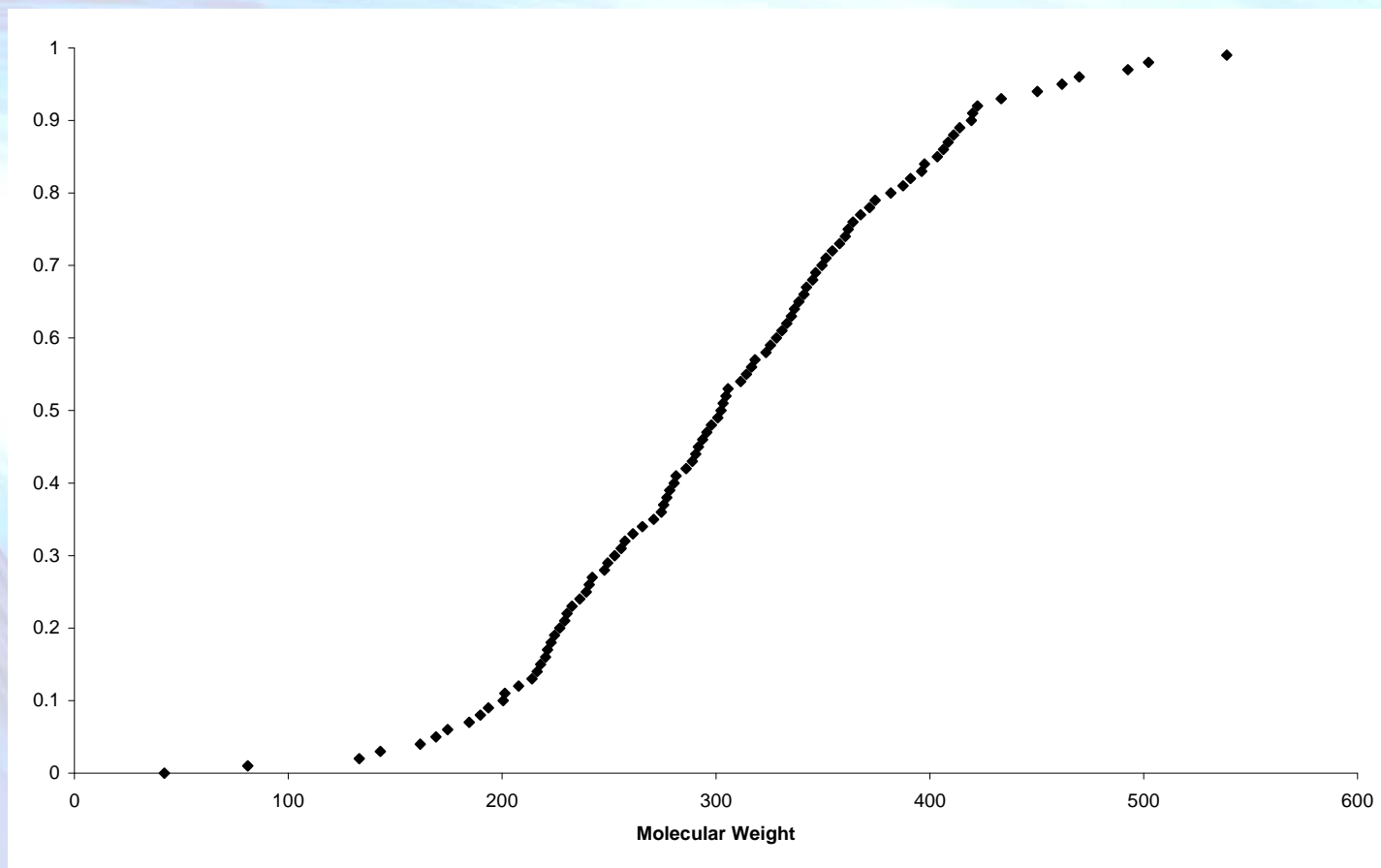
**OpenTox** 

**ИБМХ**   
ИНСТИТУТ  
БИОМЕДИЦИНСКОЙ  
ХИМИИ

# ToxCast Set : Distribution of Compounds vs. Non-Hydrogen Atoms' Amount



# ToxCast Set: Distribution of Compounds vs. Molecular Weights



# ToxCast Set: 157 Compounds Coincide with Molecules from PASS Training Set

The screenshot displays the ISIS/Base database interface for a specific compound. The main window shows a chemical structure on a yellow background, which is a pyridine ring substituted with a methyl group, an isopropyl group, and a diethyl phosphorothioate group. To the right of the structure, a table lists the following data:

ID	85
	40417
	C <sub>12</sub> H <sub>21</sub> N <sub>2</sub> O <sub>3</sub> PS
	304.3506
ToxDose_mkM	9.1028

Below the structure, the "PASS Activities" section lists the following activities:

- Teratogen
- Skin irritative effect
- Insecticide
- CYP1 substrate
- CYP1A substrate
- CYP1A2 substrate
- Non mutagenic, Salmonella
- Eye irritation, high
- Skin irritation, moderate
- CYP2 substrate
- CYP3A substrate
- CYP3A4 substrate
- CYP2C substrate
- CYP2B substrate
- CYP2B1 substrate
- CYP2B6 substrate
- CYP2C19 substrate
- CYP2B2 substrate
- CYP2C11 substrate

The interface also shows a menu bar with "Forms", "Query", "Browse", and "Update" options, and a status bar at the bottom with the text "50 of 157" and "Search Domain: All". The Windows taskbar at the bottom shows several open applications, including "пуск", "JAN-2009...", "CMTPI 200...", "IBMC", "TOXcastIB...", "ToxCast To...", "ISIS/Base...", and "Безьянн...". The system clock shows "19:38".

# PASS Training on CPDB Data

The procedure:

- The data from CPDB 2007 was used as the training set of PASS.

SAR Base information is presented:

Activity Type	Number	IAP, %
Carcinogenic, mouse	424	73.1
Carcinogenic, mouse, liver	236	74.2
Carcinogenic, mouse, lung	106	74.9
Carcinogenic, rat	553	70.0
Carcinogenic, rat, liver	198	75.3
Carcinogenic, rat, testes	19	71.1
Carcinogenic, rat, thyroid gland	35	78.2



*IAP* - Independent Accuracy of Prediction calculated by leave-one-out cross-validation procedure

# Carcinogenicity Prediction by PASS

With the trained PASS program we predicted carcinogenicity for 306 compounds from ToxRefDB. Four compounds that have two components were excluded from the prediction. For 71 compounds mouse data were not available; for 62 compounds rat data were not available.

	NA	TP	TN	FP	FN	Sensitivity	Specificity	Accuracy
CHR_Mouse_LiverTumors	71	31	108	41	64	0.33	0.72	0.57
CHR_Mouse_LungTumors	71	11	138	11	84	0.12	0.93	0.61
CHR_Mouse_Tumorigen	71	25	129	41	49	0.34	0.76	0.63
CHR_Rat_LiverTumors	62	8	159	28	58	0.12	0.85	0.66
CHR_Rat_TesticularTumors	62	12	133	101	7	0.63	0.57	0.57
CHR_Rat_ThyroidTumors	62	3	212	20	18	0.14	0.91	0.85
CHR_Rat_Tumorigen	62	8	184	46	15	0.35	0.80	0.76

NA - data not available ; TP - true positive; TN - true negative; FP - false positive; FN - false negative

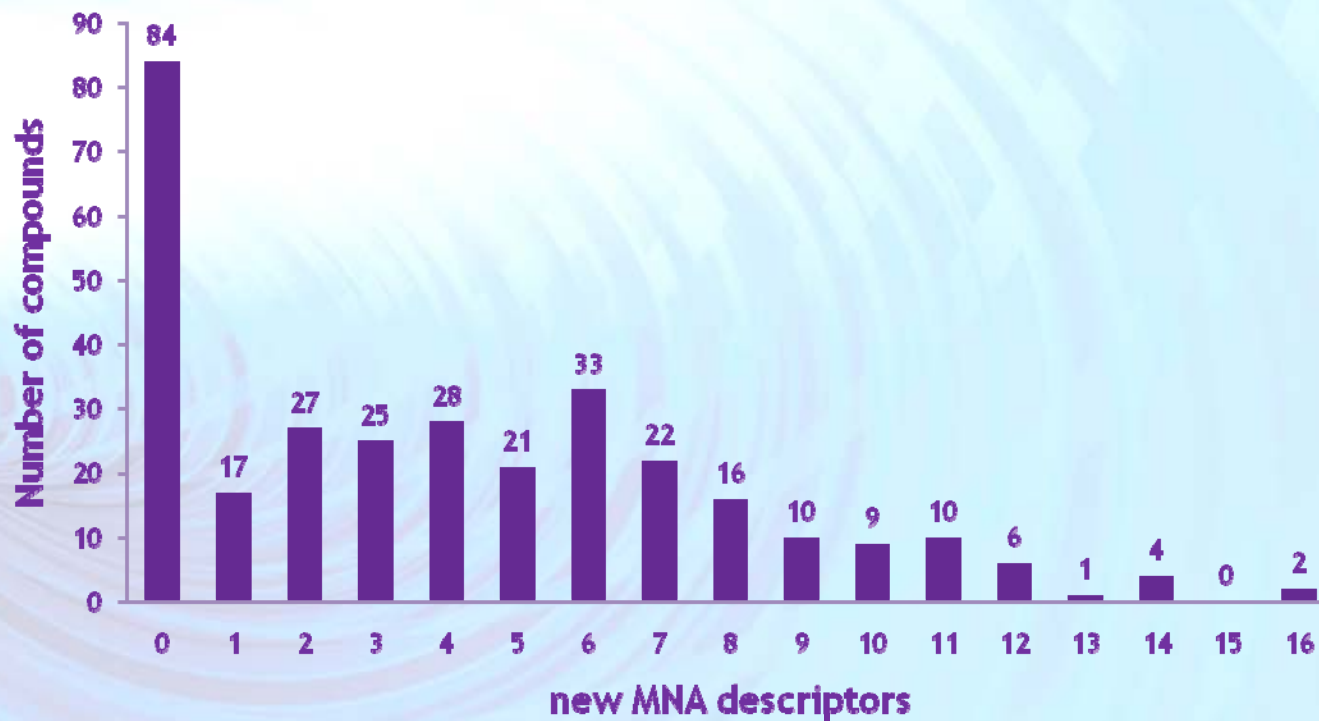


Sensitivity -  $TP/(TP+FN)$ ;

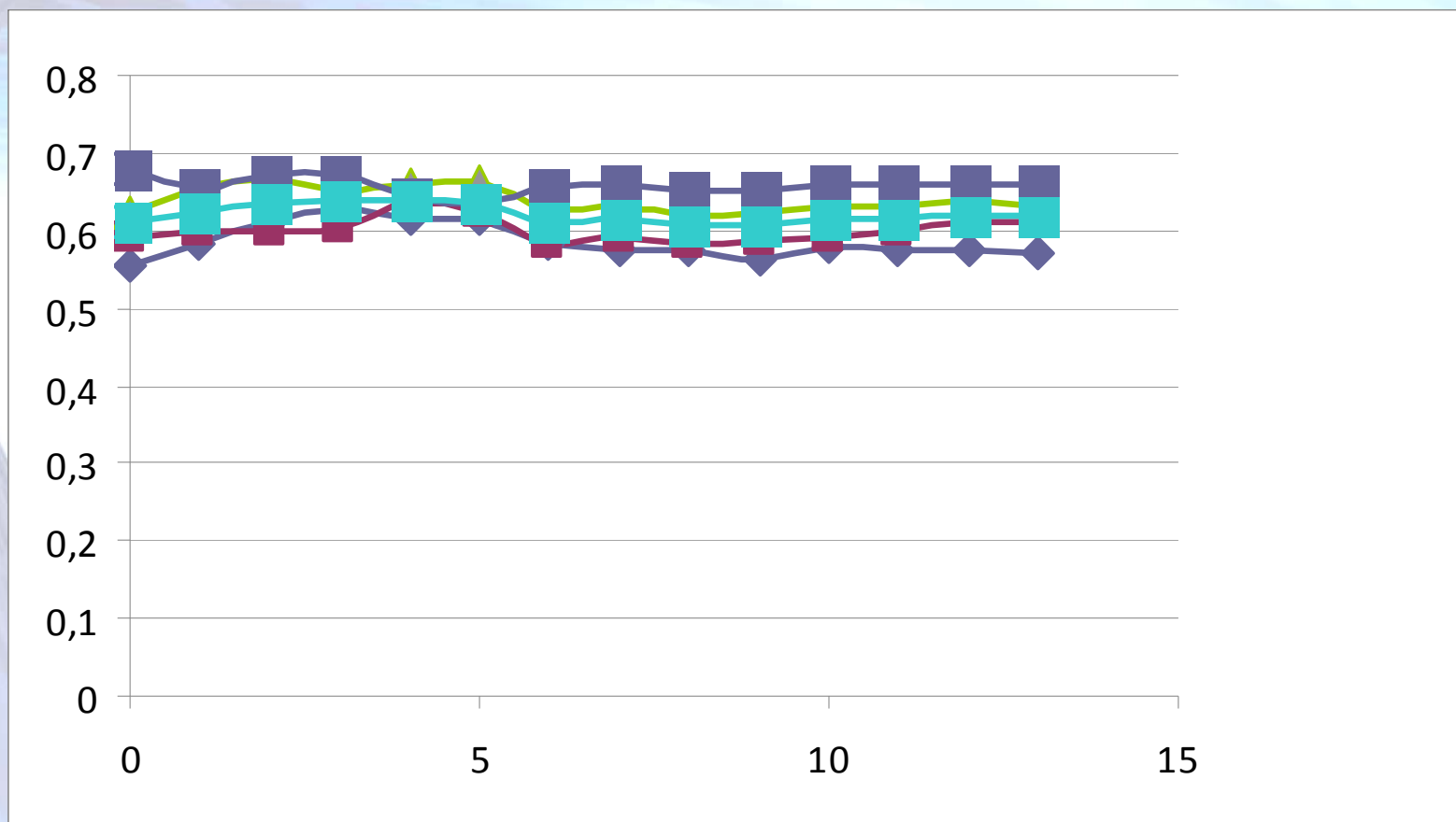
Specificity -  $TN/(TN+FP)$ ;

Accuracy -  $(TP+TN)/(TP+TN+FP+FN)$

# New MNA Descriptors in ToxCast set



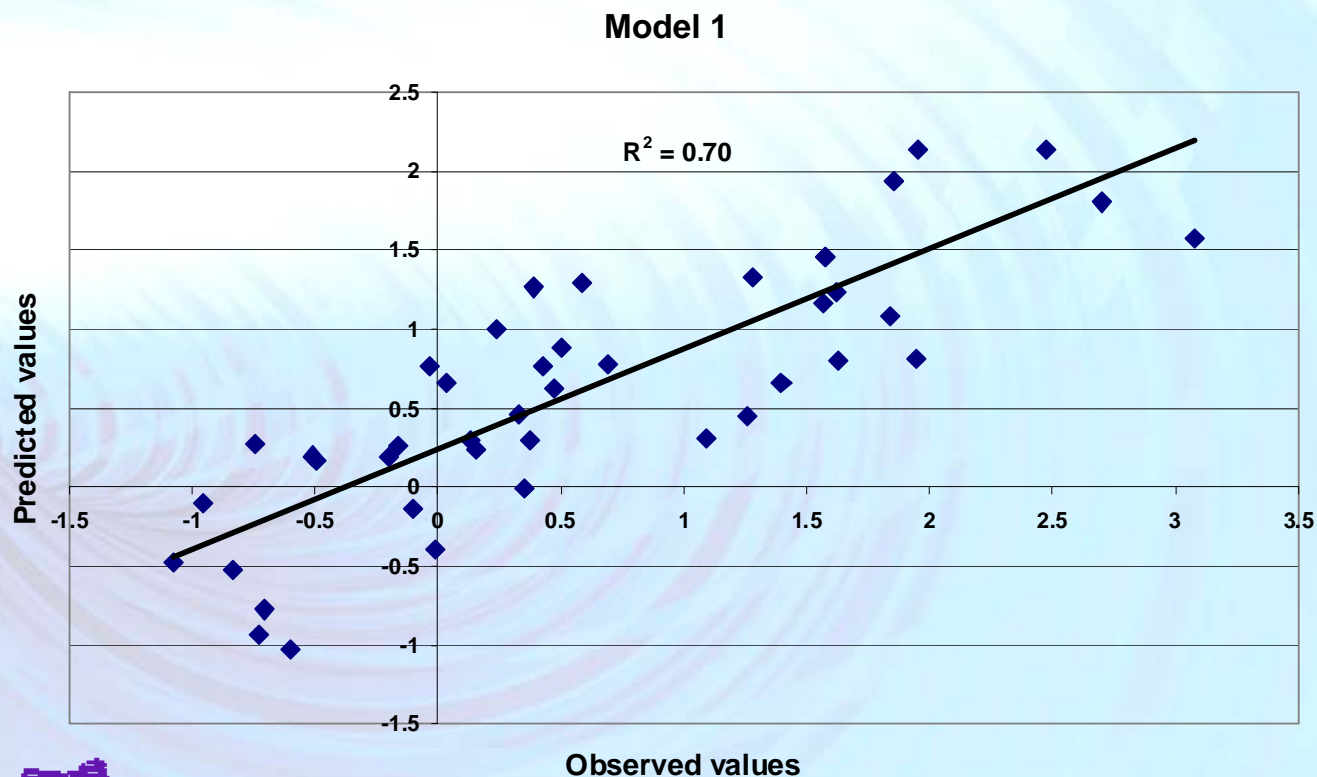
# Accuracy of Carcinogenicity Prediction vs. the Number of New Descriptors



# QSAR Models for Rat's Cholinesterase Inhibitors

ID	Num.	R2	Q2	Fisher	SD	Variables	L10%OCV
model 1	45	0.697	0.567	10.706	0.699	8	0.71
model 2	45	0.684	0.559	10.073	0.706	8	0.7
model 3	45	0.676	0.553	11.366	0.709	7	0.65
model 4	45	0.686	0.52	7.96	0.744	10	0.524

# QSAR Models for Rat's Cholinesterase Inhibitors



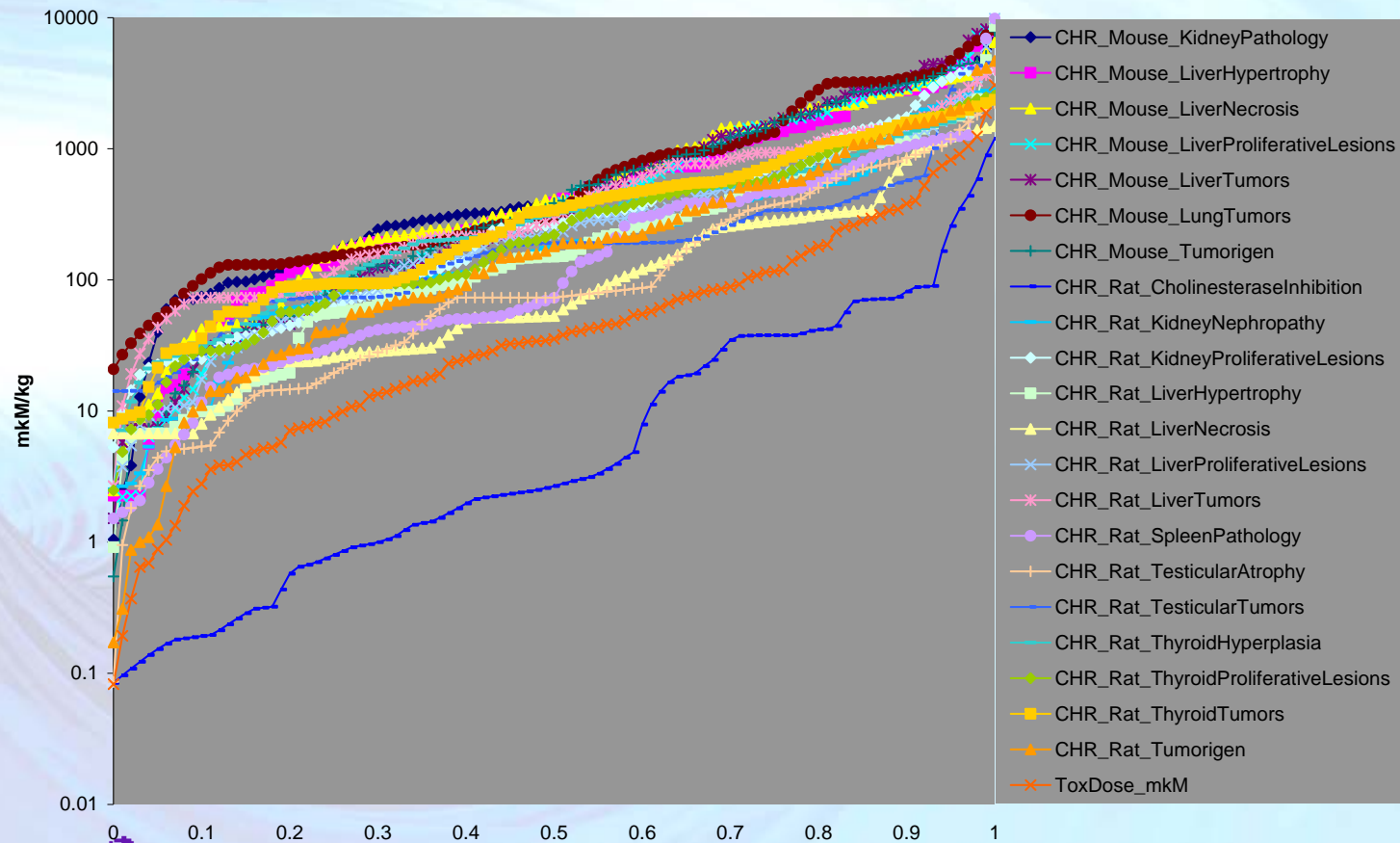
# Attempt of Creation of Integral Parameter Characterized the Compound's Hazard

Using dosage characteristics for all 75 end-points, experimental data for which were obtained in vivo, we calculated the ToxDose values for 283 compounds:

$$\text{ToxDose} = \frac{m}{n \sum_k \frac{1}{D_k}}$$

Here: D is the LEL value; m is the number of end-points for a particular compound; n is the total number of tests.

# Distribution of ToxDose Values for Different End-Points



# Results of PASS Training for Predicting Different Categories of ToxDose

No	Num	IEP, %	Activity Type
1	301	5.151	ToxCast
2	9	25.689	CHR_ToxDose < 1.0 mkM/kg
3	19	28.623	CHR_ToxDose < 3.16 mkM/kg
4	51	28.052	CHR_ToxDose < 10.0 mkM/kg
5	89	25.008	CHR_ToxDose < 31.6 mkM/kg
6	145	27.272	CHR_ToxDose < 100.0 mkM/kg
7	176	24.130	CHR_ToxDose < 316.0 mkM/kg
8	196	23.499	CHR_ToxDose < 1000.0 mkM/kg
9	15	33.091	CHR_ToxDose 0.316-3.16 mkM/kg
10	42	26.576	CHR_ToxDose 1.0-10.0 mkM/kg
11	70	28.502	CHR_ToxDose 3.16-31.6 mkM/kg
12	94	38.760	CHR_ToxDose 10.0-100.0 mkM/kg
13	87	33.645	CHR_ToxDose 31.6-316.0 mkM/kg
14	51	40.901	CHR_ToxDose 100.0-1000.0 mkM/kg
15	26	34.029	CHR_ToxDose 316.0-3160.0 mkM/kg



---

Num is the number of compounds in the training set; IEP is Independent Error of Prediction.

# Correlation between the *in vivo* and *in vitro* data

		1	2	3	4	5	6	7	8	9
1	ToxDose_mkM	1								
2	ACEA_ED	-0.003	1							
3	ATG_ED	0.094	<b>0.253</b>	1						
4	BSK_ED	0.043	<b>0.358</b>	<b>0.322</b>	1					
5	CLM_ED	<b>0.201</b>	<b>0.418</b>	<b>0.608</b>	<b>0.404</b>	1				
6	CLZD_ED	-0.007	0.054	<b>0.247</b>	0.094	<b>0.229</b>	1			
7	NCGC_ED	0.109	<b>0.225</b>	<b>0.467</b>	<b>0.269</b>	<b>0.399</b>	<b>0.184</b>	1		
8	NVS_ED	<b>0.260</b>	0.137	<b>0.394</b>	<b>0.230</b>	<b>0.443</b>	<b>0.198</b>	<b>0.272</b>	1	
9	Solidus_ED	0.083	<b>0.182</b>	<b>0.235</b>	<b>0.221</b>	<b>0.347</b>	0.095	0.181	<b>0.332</b>	1

# Supplementary Information

*in silico* toxicology

lazar Analysis -  
confusion matrix  
comparisons of lazarus  
predictions with  
ToxRefDb in vivo data

Andreas Maunz<sup>1</sup>  
Christoph Helma<sup>1,2</sup>

<sup>1</sup>) FDM Universität Freiburg (D)

<sup>2</sup>) in-silico toxicology Basel (CH)

# Rat carcinogenicity

ToxRefDb	lazar		
	inactive	active	
inactive	120	36	156
active	80	21	101
	200	57	257

Sensitivity 0.21  
Specificity 0.77  
Accuracy 0.55

# Rat carcinogenicity within AD

ToxRefDb	lazar		
	inactive	active	
inactive	58	12	70
active	40	5	45
	98	17	115

Sensitivity 0.11  
Specificity 0.83  
Accuracy 0.55

# Rat carcinogenicity Toxcast/CPDB

ToxRefDb	CPDB		
	inactive	active	
inactive	21	2	23
active	8	5	13
	29	7	36

Sensitivity 0.38  
Specificity 0.91  
Accuracy 0.72

# Mouse carcinogenicity

ToxRefDb	lazar		
	inactive	active	
inactive	111	40	151
active	66	29	95
	177	69	246

Sensitivity 0.31  
Specificity 0.74  
Accuracy 0.57

# Mouse carcinogenicity within AD

ToxRefDb	lazar		
	inactive	active	
inactive	70	17	87
active	45	10	55
	115	27	142

Sensitivity 0.18  
Specificity 0.8  
Accuracy 0.56

# Mouse carcinogenicity Toxcast/CPDB

ToxRefDb	CPDB		
	inactive	active	
inactive	22	0	22
active	10	9	19
	32	9	41

Sensitivity 0.47  
Specificity 1  
Accuracy 0.76

# Toxcast PHASE I: in-vivo data (ToxRefDB)

Modelling with Backbone  
Refinement Class Descriptors

Andreas Maunz<sup>1</sup>  
Christoph Helma<sup>1,2</sup>

<sup>1</sup>) FDM Universität Freiburg (D)

<sup>2</sup>) in-silico toxicology Basel (CH)

# Data Preprocessing

- 320 Toxcast PHASE I chemicals on 76 rodent endpoints (ToxRefDB)
- Inactive chemical-assay combinations indicated by a value of 1,000,000; led to binary classification:

MG/KG/DAY	1,000,000	else
CLASSIFICATION	inactive	active

- Used random order on chemicals

# Descriptor Calculation

- Backbone Refinement Classes <sup>1</sup>
- 2D-fragments (tree-shaped)
- Each fragment represents a class of structurally similar descriptors
- Minimum frequency: 2; significance threshold: 95%

<sup>1</sup> A. Maunz, C. Helma, and S. Kramer: Large Scale Graph Mining using Backbone Refinement Classes. In KDD '09: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (*to be published*).

# Sampling from Data

Problem: class distribution highly to fairly skewed (more inactive instances). Possible solutions:

1. **Downsampling:** reduce no of inactive samples by removing instances. *Drawback:* loss of inactive instances.
2. **Oversampling:** add active samples multiple times. *Drawback:* possible presence of the same sample in prediction and test set, therefore overfitting.
3. **Weighted Loss:** increase training weight for active instances. *Drawback:* limited effect.

# Crossvalidation

- WEKA machine learning package on randomized data, converted to ARFF format
- Best results were obtained with balanced classes obtained by downsampling; mean overall accuracy of 72.5% (+- 10.2 %)
- Model: SMO (Support Vector Machine)

# Sensitivity

- Typical example: DEV\_Rat\_General\_GeneralFetal Pathology (10-fold Crossvalidation)
- Randomized, *not balanced* (n = 129 + 40):

Acc: 84.94 %  
Spec: 94.52 %  
Sens: 43.50 %

Classified: inactive	active	
207	12	inactive
27	13	active

- Randomized, *balanced by downsampling* (n = 41+ 40):

Acc: 77.78 %  
Spec: 63.42 %  
Sens: 92.50 %

Classified: inactive	active	
26	15	inactive
3	37	active

# Supplementary Information

*in silico* toxicology

## In Vitro In Vivo Dataset - Initial Correlation Analysis

Andreas Maunz<sup>1</sup>  
Christoph Helma<sup>1,2</sup>

<sup>1</sup>) FDM Universität Freiburg (D)

<sup>2</sup>) in-silico toxicology Basel (CH)

# Qualitative correlations between HTS assays and ToxRefDB

Dataset: 20090303

Activities:

1000000 inactive

others active

Statistics:

Fisher exact test

# 14 ToxRefDB endpoints without correlated HTS assays (Fisher $p < 0.05$ )

- CHR\_Mouse\_LungTumors
- DEV\_Rabbit\_General\_GeneralFetalPathology
- DEV\_Rabbit\_Orofacial\_CleftLipPalate
- DEV\_Rabbit\_Orofacial\_JawHyoid
- DEV\_Rabbit\_Trunk\_SplanchnicViscera
- DEV\_Rabbit\_Urogenital\_Renal
- DEV\_Rat\_Cardiovascular\_Heart
- DEV\_Rat\_Cardiovascular\_MajorVessels
- DEV\_Rat\_General\_GeneralFetalPathology
- DEV\_Rat\_Trunk\_BodyWall
- MGR\_Rat\_Adrenal
- MGR\_Rat\_Mating
- MGR\_Rat\_Pituitary
- MGR\_Rat\_Prostate

# 118 HTS assays correlated with ToxRefDB endpoints (Fisher $p < 0.05$ )

ABCB11_24	ACTIN_24	ATG_PPARGa_TRANS	ATG_TGFB_CIS
ABCB11_6	ACTIN_6	ATG_PPARGd_TRANS	ATG_THRA1_TRANS
ABCB1_24	ATG_M_32_TRANS	ATG_PPARGg_TRANS	ATG_VDRE_CIS
ABCB1_48	ATG_M_61_CIS	ATG_PPARG_CIS	ATG_VDR_TRANS
ABCB1_6	ATG_M_61_TRANS	ATG_PXRE_CIS	ATG_XBP1_CIS
ABCG2_24	ATG_NFI_CIS	ATG_PXR_TRANS	BSK_3C_Eselectin
ABCG2_48	ATG_NF_kB_CIS	ATG_RARA_TRANS	BSK_3C_ICAM1
ABCG2_6	ATG_NRF1_CIS	ATG_RARB_TRANS	BSK_3C_IL8
ACEA_IC50	ATG_NRF2_ARE_CIS	ATG_RXRa_TRANS	BSK_3C_MCP1
ACEA_LOC2	ATG_NURR1_TRANS	ATG_RXRb_TRANS	BSK_3C_MIG
ACEA_LOC3	ATG_Oct_MLP_CIS	ATG_SREBP_CIS	BSK_3C_Proliferation
ACEA_LOC4	ATG_PBREM_CIS	ATG_Sp1_CIS	BSK_3C_SRB

# 118 HTS assays correlated with ToxRefDB endpoints (Fisher $p < 0.05$ )

BSK_3C_Thrombomodulin	BSK_BE3C_SRB	BSK_LPS_VCAM1
BSK_3C_TissueFactor	BSK_BE3C_TGFb1	BSK_SAg_CD38
BSK_3C_VCAM1	BSK_BE3C_hLADR	BSK_SAg_CD40
BSK_3C_Vis	BSK_KF3CT_ICAM1	BSK_SAg_CD69
BSK_3C_uPAR	BSK_KF3CT_IL1a	BSK_SAg_Eselectin
BSK_4H_Eotaxin3	BSK_KF3CT_IP10	BSK_SAg_IL8
BSK_4H_Pselectin	BSK_KF3CT_MCP1	BSK_SAg_MCP1
BSK_4H_VEGFR1I	BSK_KF3CT_MMP9	BSK_SAg_Proliferation
BSK_BE3C_IL1a	BSK_KF3CT_SRB	BSK_SAg_SRB
BSK_BE3C_MIG	BSK_LPS_CD40	BSK_SM3C_HLADR
BSK_BE3C_MMP1	BSK_LPS_IL1a	BSK_SM3C_IL_6
BSK_BE3C_PAI1	BSK_LPS_IL8	BSK_SM3C_IL_8

# 118 HTS assays correlated with ToxRefDB endpoints (Fisher $p < 0.05$ )

BSK_SM3C_MCP1	CYP1A2_24	MicrotubuleCSK_24hr
BSK_SM3C_MCSF	CYP1A2_48	MicrotubuleCSK_Destabilizer_1hr
BSK_hDFCGF_CollagenIII	CYP1A2_6	MicrotubuleCSK_Destabilizer_24hr
BSK_hDFCGF_EGFR	CYP2B6_24	MitoMass_1hr
BSK_hDFCGF_IP10	CYP2B6_48	MitoMembPot_1hr
BSK_hDFCGF_MCSF	CYP2B6_6	MitoMembPot_24hr
BSK_hDFCGF_Proliferation	CYP2C19_24	MitoticArrest_24hr
BSK_hDFCGF_TIMP1	CYP2C19_48	OxidativeStress_1hr
BSK_hDFCGF_VCAM1	CYP2C9_24	OxidativeStress_72hr
CYP1A1_24	CellCycleArrest_1hr	p53Act_1hr
CYP1A1_48	GAPDH_24	
CYP1A1_6	MicrotubuleCSK_1hr	

# OpenTox: Sample Clean-up for QSAR

David Gallagher



# ToxCast Data Analysis

Before QSAR descriptors can be calculated reliably, chemical structures need to be checked for chemical correctness, ionization state, stereochemistry, consistency, duplicates, and conformation, etc. Also the data needs to be checked for range and evenness of spread .

For high-throughput calculations and for novice users this process needs to be automated in a consistent and chemically-intelligent way.

The ToxCast samples highlight a number of issues, potential errors and inconsistencies, which could potentially affect a QSAR correlation. These are summarized next

**OpenTox project will automate data clean up as far as practical**

# Number of molecules per chemical sample

**5% (17 out of 320) of Toxcast samples contain more than one molecule**

Most QSAR descriptor algorithms are designed to work with a single molecule, however, some of the ToxCast chemical samples contain multiple molecules, additional water molecules, duplicate molecules, ions and ion pairs.

This can produce unreliable results for some descriptor calculations depending on how the situation is handled (which molecule do you use? or use the average of all? or the sum of all, etc?)

# Number of molecules per chemical sample

00066	2	acid and separate alcohol (should this be an ester?)
00104	2	includes methyl sulphate counter anion
00111	3	includes two Br <sup>-</sup> counter anions
00116	2	includes benzoate counter anion
00164	2	second molecule is HCl (should this be a hydrochloride?)
00179	2	includes Na <sup>+</sup> counter cation
00190	4	two identical molecules, plus Zn and Mn cations
00191	2	includes Mn counter cation
00193	2	includes Cl <sup>-</sup> counter anion
00197	5	includes Na <sup>+</sup> counter cation and three water molecules
00207	2	includes Zn <sup>++</sup> counter cation
00229	3	neutral molecule, but includes two water molecules
00249	3	two identical anions and one Ca <sup>++</sup> counter cation
00252	2	second molecule is HCl (should this be a hydrochloride?)
00259	2	includes Na <sup>+</sup> counter cation
00270	2	includes Na <sup>+</sup> counter cation
00314	2	includes Na <sup>+</sup> counter cation

# 2D to 3D structure conversion

Most databases store molecules as 2D structures, however, many QSAR descriptors (e.g. quantum mechanics, 3D substructure searching) require reasonable 3D starting structures.

Hence, 2D-3D conversion using standard rules and templates and saturation with hydrogen needs to be applied as a preprocessing step, followed by a molecular mechanics cleanup.

**Geometry**

**Stereochemistry and chirality**

**Conformation**

**Ionization state**

# Other Issues

**Duplicate structures ?**

**Data from consistent mechanism classes or modes of action ?**

**Singletons and chemical space**

limit the prediction reliability

**Data skew ?**

the data set needs to be checked for even spread (i.e data skewness is below a specified threshold, to avoid the potential for a misleadingly high  $r^2$ ).

**Descriptor parameterization and molecule size ?**