



THE REFERENCE IN CHEMOINFORMATICS  
eADMET

## OCHEM

Product features and highlights

# Content

- OCHEM at a glance (components and Data upload)
- How to run models for ADME prediction?
- How to build models (Regression, Classification) and get Applicability domain?
- How to interpret models?
  - Tox Alerts
  - Set Compare
- System Architecture
- Case Study
- Our unique offer

# Content

- OCHEM at a glance (components and Data upload)
- How to run models for ADME prediction?
- How to build models (Regression, Classification) and get Applicability domain?
- How to interpret models?
  - Tox Alerts
  - Set Compare
- System Architecture
- Case Study
- Our unique offer

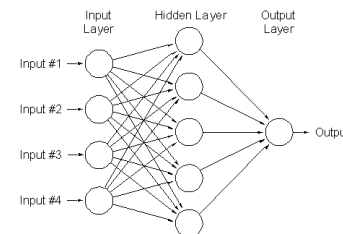
# What is OCHEM?

- It is a web-based modeling framework.
- Hosts Chemical and Biological data, ADME/T models together with the tools to build and validate these models.
- Allows collaboration between different scientists inter/intra organizations in storing, modeling and analyzing chemical related bioassays.

# How OCHEM parts interact?



0	3	2	5	4	7	6	9	8
3	0	1	2	3	4	5	6	7
2	1	0	3	2	5	4	7	6
5	2	3	0	1	2	3	4	5
4	3	2	1	0	3	2	5	4
7	4	5	2	3	0	1	2	3
6	5	4	3	2	1	0	3	2
9	6	7	4	5	2	3	0	1
8	7	6	5	4	3	2	1	0



## Database

Chemical structures  
(2D, 3D)

Experimental results,  
cached descriptors

Biological Pathways

Properties, their  
units, unit conversion  
rules, etc...

## Calculators

Chemical-structure-  
derived descriptor (in  
silico)

Biologically derived  
descriptor (in vitro)

Descriptors derived  
from Protein-ligand  
interaction  
(Autodock)

In research: Shape-  
derived descriptor  
(ex: for nanoparticles  
toxicity)

## Mining algorithms

Linear and non-linear

Applicability domain  
estimation

Apply bootstrapping  
(bagging) and cross  
validation cycles

Model multiple  
properties  
simultaneously

Ex: ANN, KNN, SVM,  
J48

# Data upload

- OCHEM supports data import through sdf/text/Excel or through Knime and PipelinePilot.
- Validity rules are automatically applied for Chemical Structures and Data.
- Duplicates Alert
- Automatic unit conversion throughout the system
- 2-level of data federation by Expert users (Property Moderators, System Admins)

# Batch Data Upload - Excel



## Batch upload browser

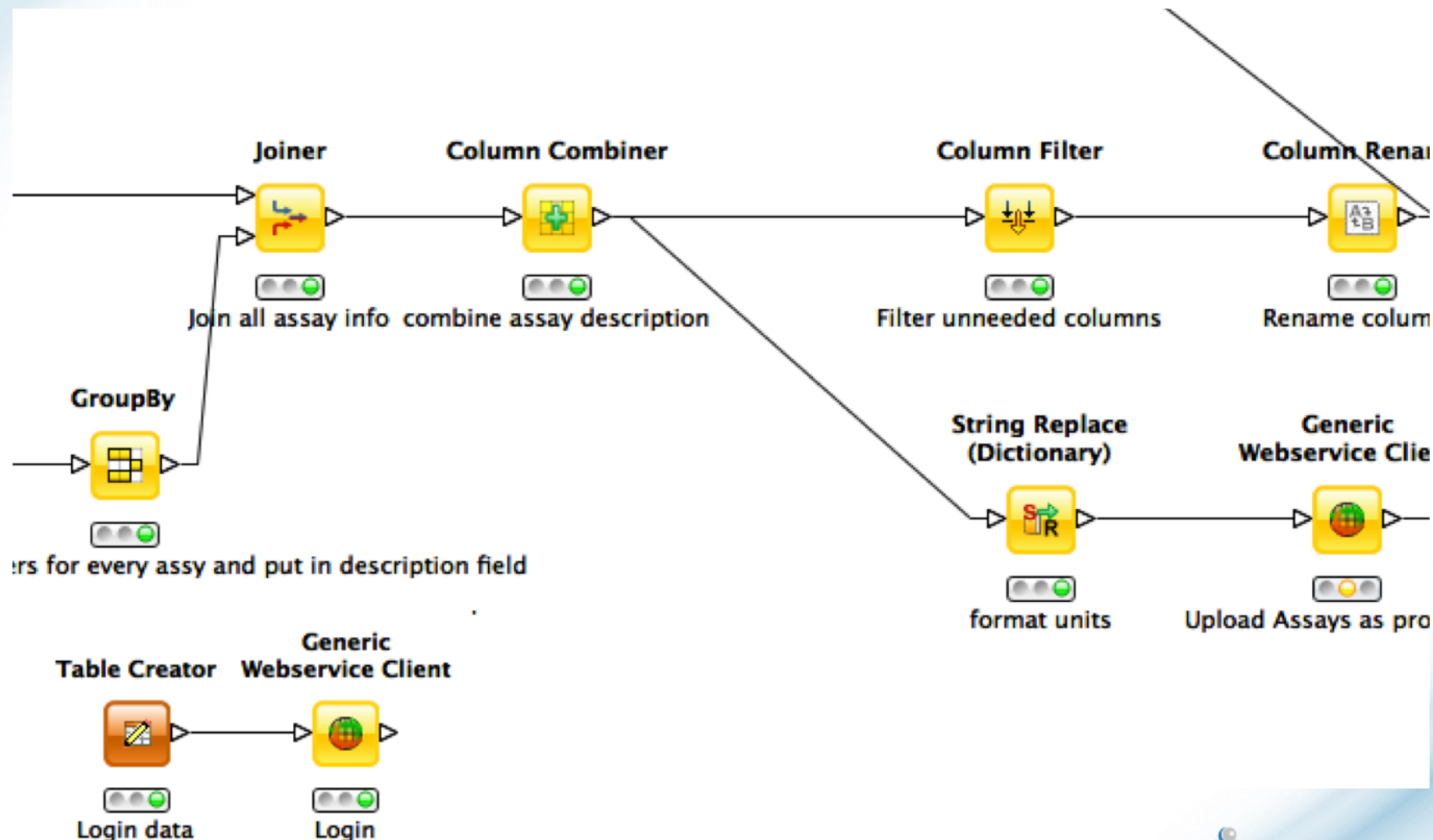
Please, select an XLS sheet you'd like to use to upload data

LogKOC

<input checked="" type="checkbox"/> casrn	<input checked="" type="checkbox"/> name	<input checked="" type="checkbox"/> LogKoc_Tutorial	<input checked="" type="checkbox"/> molecule
15972-60-8	alachlor	2.28	<chem>CCC1=CC=CC(CC)=C1N(COC)C(=O)CCl</chem>
23184-66-9	butachlor	2.86	<chem>CCCCOCN(C(=O)CCl)C1=C(CC)C=CC=C1CC</chem>
1918-16-7	propachlor	2.42	<chem>CC(C)N(C(=O)CCl)C1=CC=CC=C1</chem>
1646-88-4	Aldicarb Sulfone	0.42	<chem>CNC(=O)O\N=C\C(C)(C)S(C)(=O)=O</chem>
2008-41-5	butylate	2.11	<chem>CCSC(=O)N(CC(C)C)CC(C)C</chem>
63-25-2	carbaryl	2.4	<chem>CNC(=O)OC1=CC=CC2=CC=CC=C12</chem>
1563-66-2	Carbofuran	1.75	<chem>CNC(=O)OC1=CC=CC2=C1OC(C)(C)C2</chem>
101-21-3	Chlorpropham	2.53	<chem>CC(C)OC(=O)NC1=CC=CC(Cl)=C1</chem>
1134-23-2	cycloate	2.54	<chem>CCSC(=O)N(CC)C1CCCCC1</chem>
2303-16-4	diallate	3.28	<chem>CC(C)N(C(C)C)C(=O)SC\C(Cl)=C\Cl</chem>



# Batch Data upload – Sample Knime workflow





# Content

- ✓ OCHEM at a glance (components and Data upload)
- **How to run models for ADME prediction?**
- How to build models (Regression, Classification) and get Applicability domain?
- How to interpret models?
  - Tox Alerts
  - Set Compare
- System Architecture
- Case Study
- Our unique offer

# Applying a Model

- OCHEM hosts many Physicochemical (solubility, lipophilicity, DMSO solubility, Melting point, ...etc) and ADME/T models (CYP450 enzyme inhibition, AMES test, GI absorption,...etc) published in peer-reviewed journals.
- The system reports advanced applicability domain estimation on the estimated error in prediction together with the in/out of applicability domain flag.

# Available models

## Physicochemical properties

- LogP, LogS, Solubility in DMSO
- Melting point, Boiling point

## Toxicity

- CYP 450 Inhibition, AhR Activation
- AMES Test, BioConcentration factor, T. Pyriformis toxicity

## Bioavailability

- Gastrointestinal absorption
- BBB permeability, CACO-2

# Building new models

- It is easy to create new models from existing data.
- More than 10 Machine learning algorithms (Neural networks, KNN, SVM, MLR, PLS, random forests, J48)
- More than 15 descriptor packages (Academic and Commercial)

# Content

- ✓ OCHEM at a glance (components and Data upload)
- ✓ How to run models for ADME prediction?
- How to build models (Regression, Classification) and get Applicability domain?
- How to interpret models?
  - Tox Alerts
  - Set Compare
- System Architecture
- Case Study
- Our unique offer

## > 20 Descriptors packages

- Commercial and academic packages
- Thousands of 0D to 3D descriptors
- Experimental data, conditions of experiment or predictions of other models are used as descriptors
- New descriptor packages can be easily integrated
- Packages can be easily combined (mix & match)

### *Suggested descriptors:*

- ☐ E-state [W](#)
- ☒ ALogPS (2) [W](#)
- ☐ GSFragment (1138) [W](#)
- ☐ Dragon v. 6.0 (4885/3D) [W](#)
- ☐ ISIDA fragments [W](#)
- ☐ ADRIANA.Code (211/3D) [W](#)
- ☐ CDK descriptors (246/3D) [W](#)
- ☐ 'Inductive' descriptors (54/3D) [W](#)
- ☐ MERA descriptors (529/3D) [W](#)
- ☐ MERSY descriptors (42/3D) [W](#)
- ☐ Chemaxon descriptors (499/3D) [W](#)
- ☐ QNPR [W](#)
- ☐ Spectrophores (144/3D) [W](#)

### *Additional or obsolete descriptors:*

- ☐ OEState [W](#)
- ☐ MolPrint [W](#)
- ☐ Dragon v. 5.4 (1630/3D) [W](#)
- ☐ Dragon v. 5.5 (3190/3D) [W](#)
- ☐ Structural Alerts [W](#)
- ☐ MOPAC descriptors (21/3D) [W](#)
- ☐ ShapeSignatures (3D) [W](#)

# Featured descriptor package: Dragon

- Dragon 6 calculates 4885 molecular descriptors [[list](#)]
- Divided into 29 descriptor blocks (2D and 3D)
- Different versions available for backward compatibility (5.4, 5.5 and 6)
- Widely used by the industry and Academia for chemical modeling
- No extra fees for users that already posses a Dragon license
- Developed and Licensed by Talete srl<sup>[1]</sup>

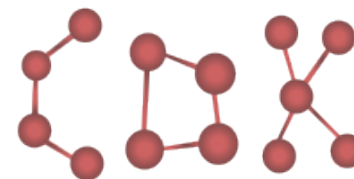


[1] R.Todeschini and V.Consonni: "Molecular Descriptors for Chemoinformatics", (2 volumes), WILEY-VCH, Weinheim (Germany) 2009, 1257 pp.



# Featured descriptor package: CDK

- Open-source Java library for Chemoinformatics and Bioinformatics
- Offers variety of Constitutional, Topological, electronic, Geometrical and Hybrid descriptors
- Has both 2D and 3D descriptors
- Distributed under the GNU LGPL license



# Featured descriptor package: ADRIANA.CODE

- ADRIANA.Code comprises a unique combination of methods for calculating molecular structure descriptors on a sound geometric and physicochemical basis.
- Physicochemical properties (global descriptors)
- Shape- and size-related
- Autocorrelation of 2D interatomic distance distributions
- Autocorrelation of 3D interatomic distance distributions
- Radial distribution functions (RDF) of 3D interatomic distances
- Autocorrelation of distances between surface points
- Developed and licensed by: Molecular Networks GmbH
- No extra fees for users that already posses a license



# Featured descriptor package: Chemaxon Calculators

- Implemented calculations and property predictions efficiently evaluate pharmaceutically relevant physico-chemical properties and molecular descriptors
  - Physico-chemical property predictors
  - Structural property calculations
  - Molecular Modelling
- Developed and licensed by: Chemaxon Kft.
- No extra fees for users that already possess calculators license



# Featured descriptor package: ESTATE

- Electro-topological state indices combine electronic and graph-topological information about a compound.
- EState indices are separated on atom/bond type. In addition to indices, it is also possible to select E-state counts, which correspond to counts of atom or bond types according to the respective indices.
- EState indices encode electronic and topological information, and have proven useful in the establishment of QSAR and QSPR models.
- Implementation by eADMET GmbH, introduced by Lowel H. Hall and Lemont B. Kier in the 1990s.

# Featured descriptor package: AlogPS

- Provides predictions for lipophilicity and water solubility
- Ranked first for the prediction of logPow in an independent study by Pfizer and Nycomed.
- logPow is known to correlate well with many biological processes.
- Usually used in combination with other descriptor packages.

Performance of algorithms for *in-house* datasets

Method	Pfizer set (N = 95 899)							Nycomed set (N = 852)								
	RMSE	Failed	rank	% in error range			RMSE	rank	% in error range			RMSE	rank	% in error range		
				<0.5	0.5 - 1	>1			<0.5	0.5 - 1	>1					
ALOGPS	1.02		I	41	30	29	1.01	0.68	I	51	34	15				
S*logP	1.02		I	44	29	27	1.00	0.69	I	58	27	15				
NC+NHET	1.04		II	38	30	32	1.04	0.88	III	42	32	26				
MLOGP(S+)	1.05		II	40	29	31	1.05	1.17	III	32	26	41				
XLOGP3	1.07		II	43	28	29	1.06	0.65	I	55	34	12				
MiLogP	1.10	27	II	41	28	30	1.09	0.67	I	60	26	14				
ABLogP	1.12	24	II	39	29	33	1.11	0.88	III	45	28	27				
AlogP	1.12		II	39	29	32	1.12	0.72	II	52	33	15				
AlogP98	1.12		II	40	28	32	1.10	0.73	II	52	31	17				
OsirisP	1.13	6	II	39	28	33	1.12	0.85	II	43	33	24				
AAM	1.16		III	33	29	38	1.16	0.94	III	42	31	27				
CLOGP	1.23		III	37	28	35	1.21	1.01	III	46	28	22				
ACDlogP	1.28		III	35	27	38	1.28	0.87	III	46	34	21				
CSlogP	1.29	20	III	37	27	36	1.28	1.06	III	38	29	33				
COSMOFrag	1.30	1088 <sup>3</sup>	III	32	27	40	1.30	1.06	III	29	31	40				
QikProp	1.32	103	III	31	26	43	1.32	1.17	III	27	24	49				
KowWIN	1.32	16	III	33	26	41	1.31	1.20	III	29	27	44				
QLogP	1.33	24	III	34	27	39	1.32	0.80	II	50	33	17				
XLOGP2	1.80		III	15	17	68	1.80	0.94	III	39	31	29				
MLOGP(Dragon)	2.03		III	34	24	42	2.03	0.90	III	45	30	25				

# Featured descriptor package: Chiral descriptors

- Physicochemical atomic stereodescriptors (PAS) that represent the chirality of an atomic chiral center on the basis of empirical physicochemical properties of the ligands
- The learned models could make correct predictions about the preferred enantiomer, from the molecular structure of the substrate.
- Developed by Dr. Qingyou Zhang and Dr. Joao Aires de Sousa

Zhang, Q.-Y., & Aires-de-Sousa, J. (2006). Physicochemical stereodescriptors of atomic chiral centers. *Journal of chemical information and modeling*, 46(6), 2278–87. doi:10.1021/ci600235w

# Featured descriptor package: ISIDA fragments

- ISIDA Fragments is a category of fragments descriptors. They use 2D Lewis graph representation of the compounds but do not consider stereoisomerism. They also are tautomer- and mesomer-dependant.
- ISIDA Fragments uses three modes of fragmentations: Paths, Trees, Neighbours
- These fragmentation modes are completed by several representations of the atoms of the molecule.
- Developed by: Laboratoire d'Infochimie, UMR 7177 Université de Strasbourg-CNRS



# Featured descriptor package: ToxAlerts

- Structural alerts (also known as "toxicophores") are molecular patterns known to be associated with particular type of toxicity.
- efficient technique to detect potentially toxic chemicals.
- Screening chemical compounds against known structural alerts can be a good practice to complement the QSAR models and to help interpreting their predictions.

Sushko, I. et al, 2012. ToxAlerts: A Web Server of Structural Alerts for Toxic Chemicals and Compounds with Potential Adverse Reactions. Journal of chemical information and modeling. doi:10.1021/ci300245

# Featured descriptor package: Experimental Properties

- Allows modeler to use experimental data from other modeled properties to model the property of interest.
- For example: Correlate HTS in-vitro assay data to a toxicological endpoint.
- Optimized for handling big Data with millions of points
- Specially suitable for Genomic/Proteomic approaches
- Developed by: eADMET GmbH

# Other descriptor packages

AMBIT Descriptors  
Chemaxon scaffolds  
Chiral Descriptors  
Custom descriptors from a file  
Docking descriptors  
ECFP fingerprints  
ETM  
Functional groups  
GSFrag  
Inductive Descriptors

MERA & MERSY descriptors  
MolPrint  
MOPAC-derived descriptors  
OEState  
QNPR  
Scaffold Hunter descriptors  
Shape Signatures  
Silicos-It scaffolds  
Spectrophores

# Pre-processing of descriptors

- Unsupervised methods
  - Eliminate near constant descriptors
  - Group highly-correlated descriptors
  - Use PCA
  - Use Unsupervised Forward Selection
  - Select descriptors manually or/and from a list
- Supervised methods
  - MLRA
  - PLS
- In development
  - Genetic algorithms
  - Variable importance estimation

# Model statistics (regression)

Overview    Applicability domain

Model name: Melting Point - 3D (Dragon 6 + Corina) [rename] , published in [Sample OCHEM models](#) public identifier is 67

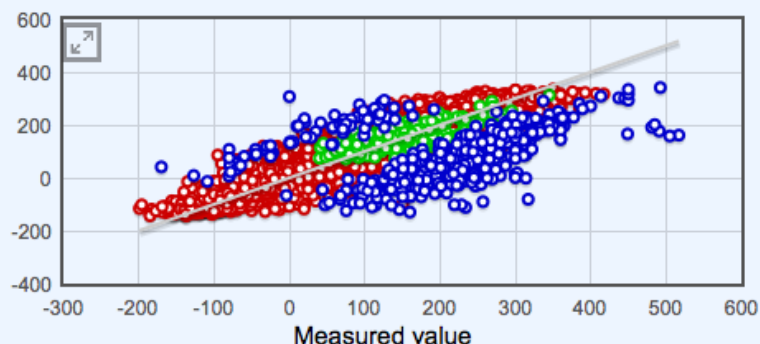
Predicted property: [Melting Point](#)

Training method: ANN  
modeled in °C

[Dragon6 (blocks: 1-29)]  
Correl. limit: 0.95 Variance threshold: 0.01,  
Maximum value: 500000, using UFS  
Supersab, 1000 iterations, 3 neurons  
ensemble=64 additional param  
PARTITION=3,SELECTION=2  
5-fold cross-validation  
-

Data Set	#	R2	q2	RMSE	MAE
Training set: <a href="#">A MP new</a>	<a href="#">25547 records</a>	0.821 ± 0.005	0.82 ± 0.005	37 ± 0.4	28 ± 0.3
Test set: <a href="#">Bergstrom</a> [x]	<a href="#">277 records</a>	0.53 ± 0.09	0.5 ± 0.1	38 ± 3.1	30 ± 2.6
Excluded from training set	<a href="#">472 records</a>	0.11 ± 0.06	0 ± 0.4	168 ± 5.8	159 ± 5.1

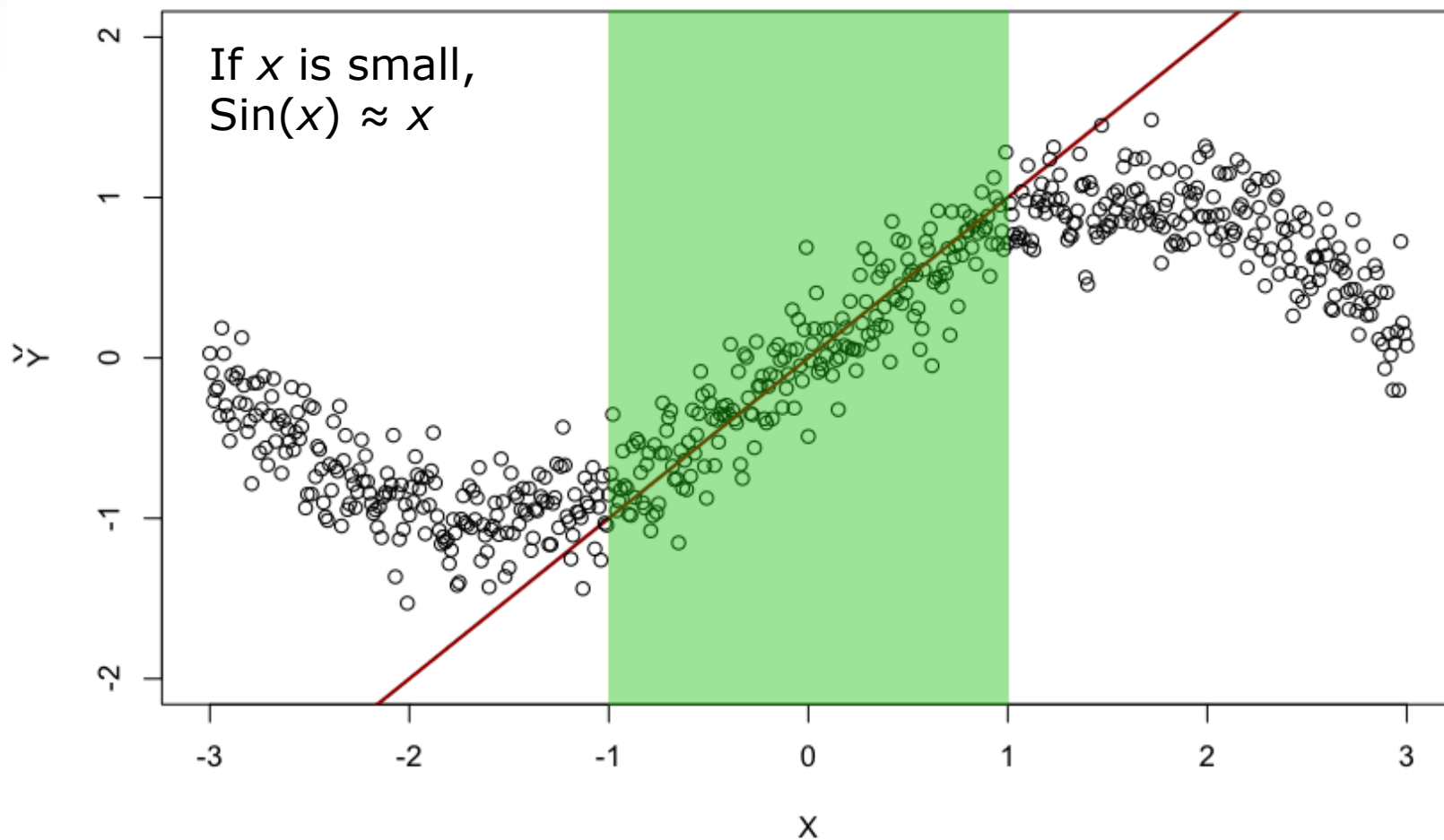
Calculated in 132989 seconds  
Size: 25788 Kb



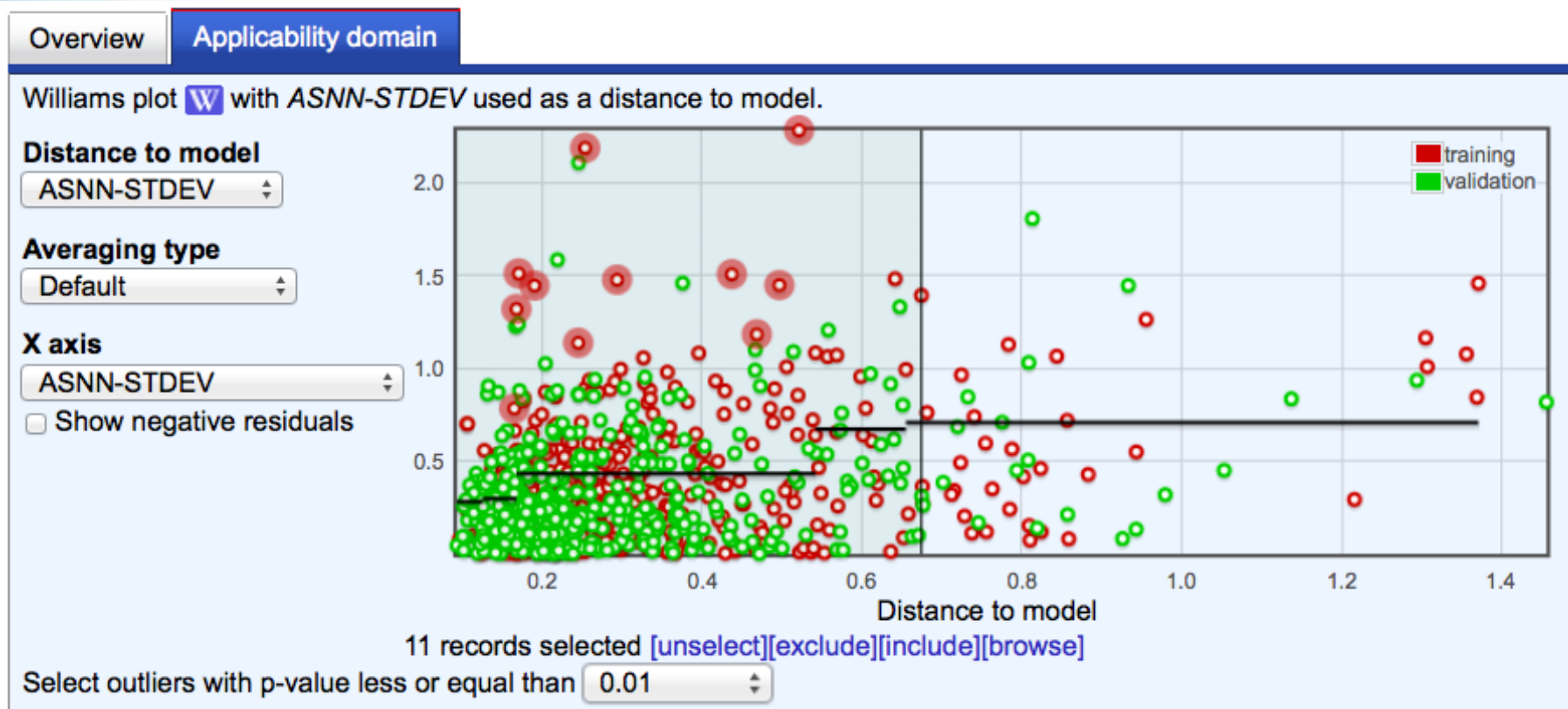
[Download model statistics in Excel format](#)    [Create a copy of this model](#)    [View configuration XML](#)    [Export configuration XML](#)

- Multiple statistical measures + confidence intervals
- Export for offline inspection
- Interactive scatter plot

# Accuracy of prediction



# Applicability domain assessment (regression)



- Several applicability domain measures (bagging-based for all methods; standard deviation, correlation in the property space, leverage, etc.)
- Automatic exclusion of outliers based on  $p$ -value



# Prediction of new molecules (regression)

 Export results in a file (Excel, CSV or SDF)

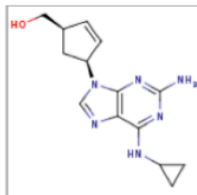
 Advanced applicability domain charts>>

Sorting

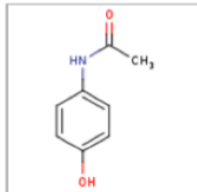
Accuracy estimates for the set  
log(IC50-1) for 256 compounds  
RMSE = 0.69 ± 0.06  
MAE = 0.55 ± 0.05

1 - 15 of 256

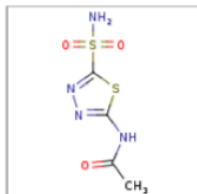
15 items on page 1 of 18 > >>



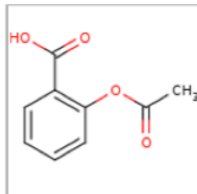
log(IC50-1) (Toxicity against T. Pyriformis) = 0.18 -log(mmol/L) ± 1.38 (ASNN-STDEV = 1.16, estimated RMSE = 0.71) **OUT OF AD**



log(IC50-1) (Toxicity against T. Pyriformis) = -0.68 -log(mmol/L) ± 1.38 (ASNN-STDEV = 0.82, estimated RMSE = 0.71) **CACHED** **OUT OF AD**



log(IC50-1) (Toxicity against T. Pyriformis) = 0.02 -log(mmol/L) ± 1.38 (ASNN-STDEV = 1.30, estimated RMSE = 0.71) **CACHED** **OUT OF AD**



log(IC50-1) (Toxicity against T. Pyriformis) = 0.2 -log(mmol/L) ± 0.85 (ASNN-STDEV = 0.36, estimated RMSE = 0.43) **CACHED**

# Model statistics (classification)

Overview
Applicability domain

Model name: Ames levenberg [rename] , published in [Applicability domains for classification problems: Benchmarking of distance to models for Ames mutagenicity set](#). public identifier is 1

Predicted property: AMES

Training method: ANN

[OEstate]

Correl. limit: 0.95 Variance threshold: 0.0,  
Maximum value: 999999,  
Levenberg, 1000 iterations, 3 neurons  
ensemble=100 additional param  
PARALLEL=10  
5-fold cross-validation  
-

Calculated in 2402 seconds  
Size: 948 Kb

Data Set	#	Accuracy	Balanced accuracy
Training set: Ames challenge training	4357 records	78.1% $\pm$ 1.2	77.9% $\pm$ 1.2
Test set: Ames challenge test [x]	2181 records	79.9% $\pm$ 1.7	79.8% $\pm$ 1.7

Real↓/Predicted→	inactive	active
inactive	1521	495
active	460	1883
Training (Original)		

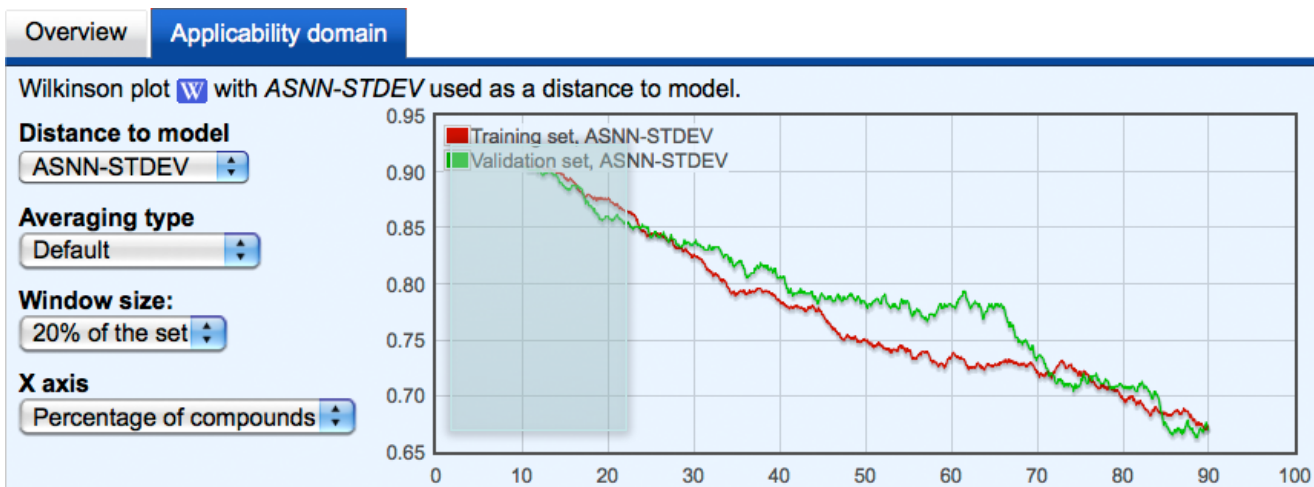
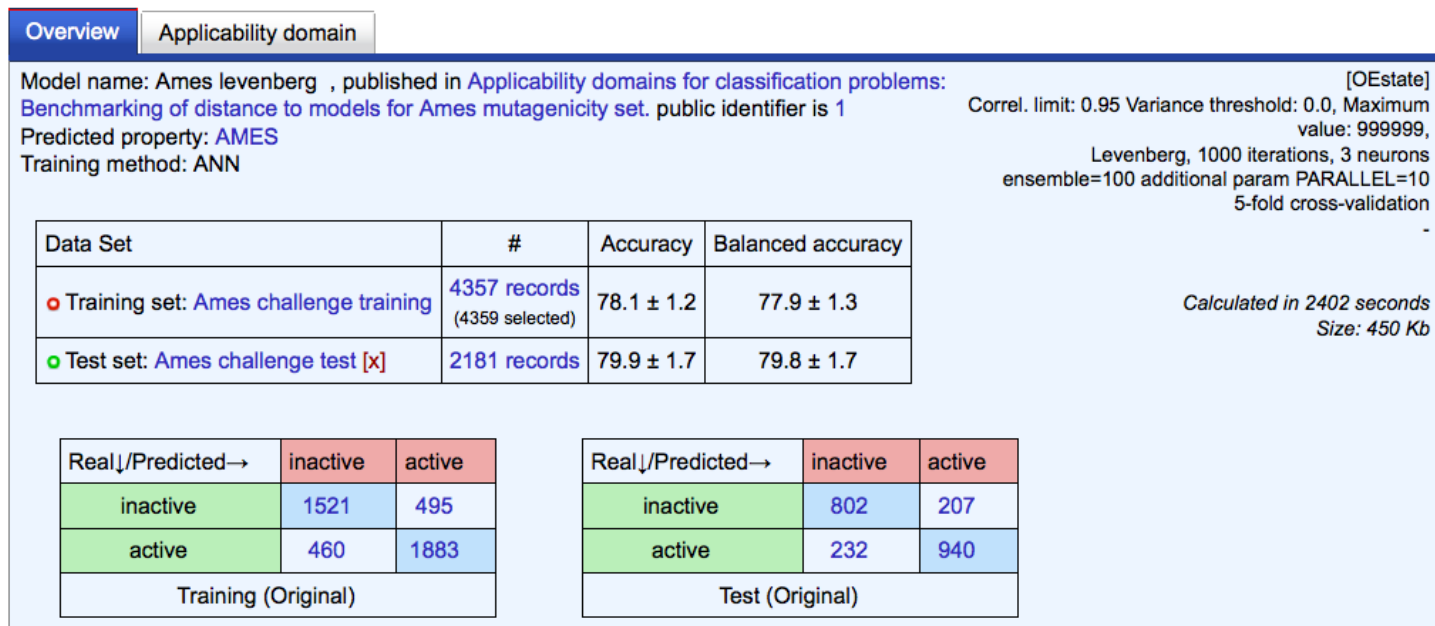
Real↓/Predicted→	inactive	active
inactive	802	207
active	232	940
Test (Original)		

Number of compounds ignored because of errors in original model = 2

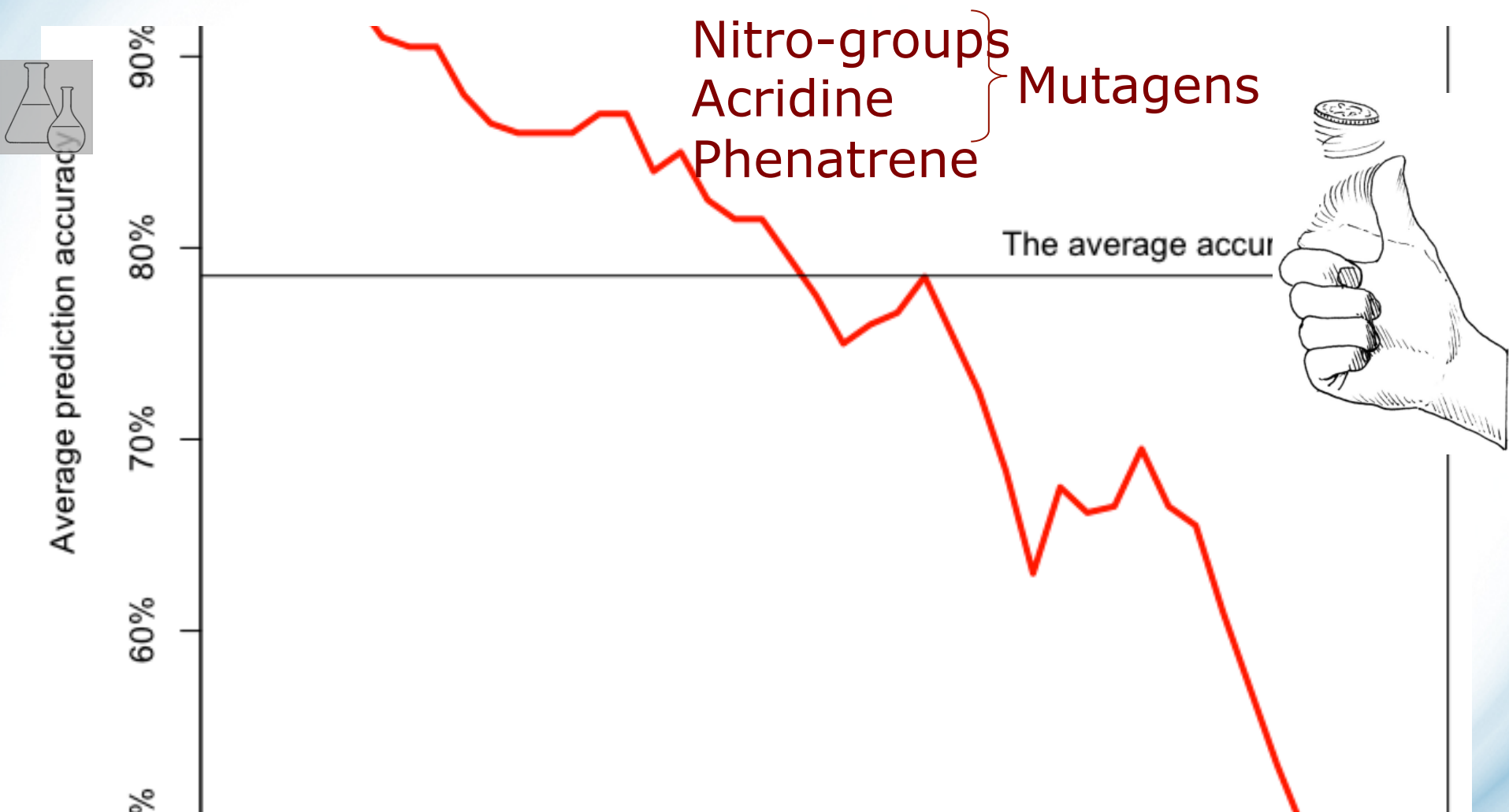
[Download model statistics in Excel format](#)
[Create a copy of this model](#)
[View configuration XML](#)
[Export configuration XML](#)

- Confusion matrices
- Support for binary and multi-class classification

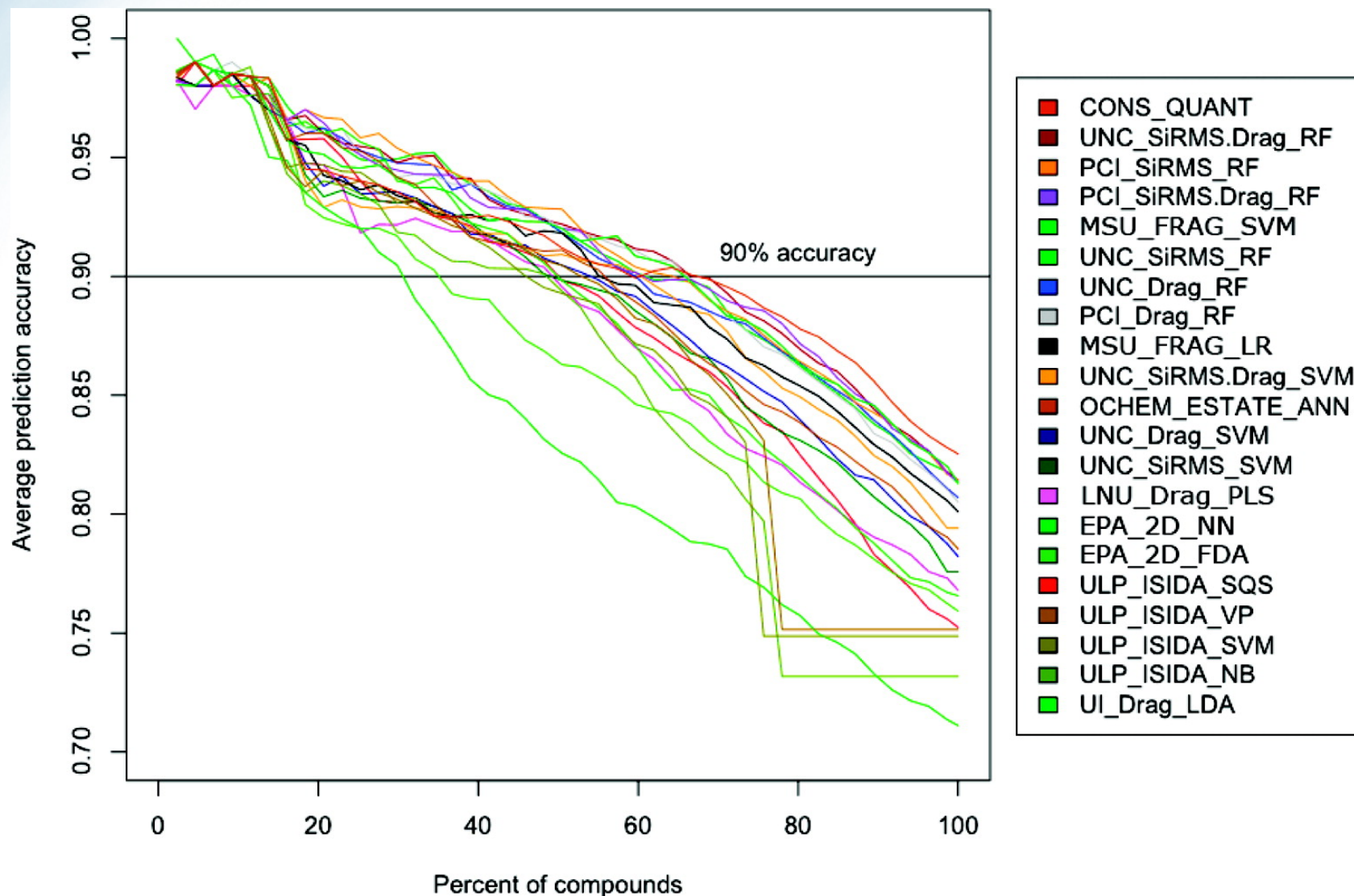
# Accuracy of predictions for classification model



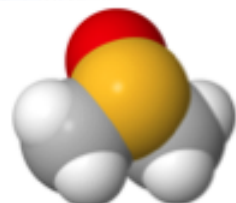
# Very large differences in the accuracy of prediction



# Accuracy of models for AMES test set

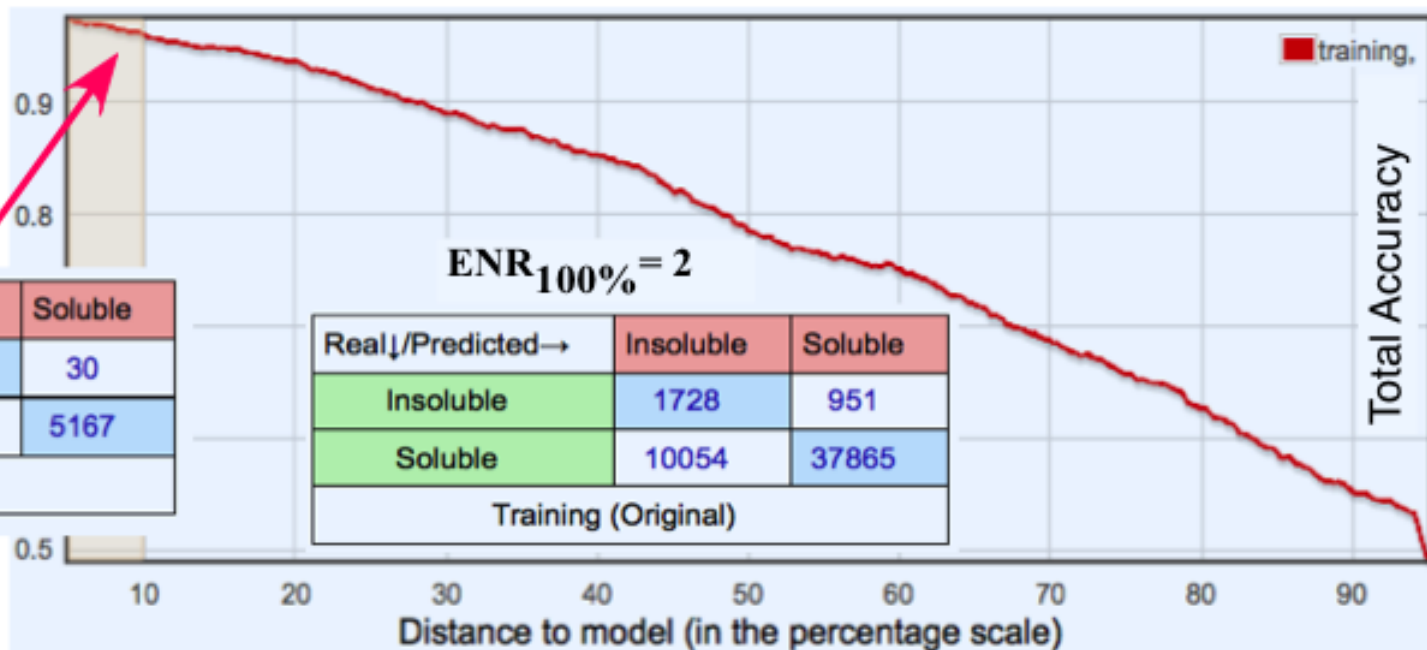


# Prediction of compound solubility in DMSO



$ENR_{10\%} = 9$

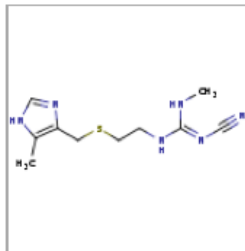
Real↓/Predicted→	Insoluble	Soluble
Insoluble	97	30
Soluble	116	5167
Training (Original)		



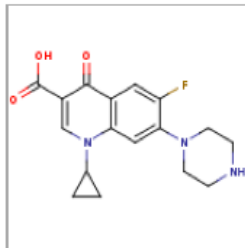
The final model is based on >163k compounds – the largest published model



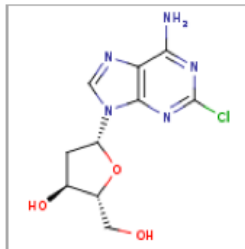
# Prediction of new molecules (classification)



AMES (Ames levenberg) = inactive (68.0% accuracy) **CACHED**



AMES (Ames levenberg) = active (53.0% accuracy) **CACHED** **OUT OF AD**



AMES (Ames levenberg) = active (53.0% accuracy) **OUT OF AD**



# Comprehensive modeling

- Thousand of models can be simultaneously built based on preconfigured modeling templates:
- Using KNIME or Pipeline Pilot with pre-saved XML configurations exported from OCHEM
- Using the web-interface:

The comprehensive modeling feature allows you to simultaneously run multiple models with different machine learning methods, molecular descriptors and validation protocols. Please note that running multiple models may require significant computational resources and time.

**Select the training and validation sets:**  
Training set (required): [...] [Add a validation set](#)

**Select the methods you want to use for the modeling:**

Method	Descriptors	Descriptor selection	Model validation
<a href="#">[all]</a> <a href="#">[none]</a> <input type="checkbox"/> ANN <input checked="" type="checkbox"/> ASNN (with Library mode) <input type="checkbox"/> KNN <input type="checkbox"/> LibSVM <input checked="" type="checkbox"/> FSMLR <input type="checkbox"/> MLRA <input checked="" type="checkbox"/> PLS <a href="#">+add a custom template</a>	<a href="#">[all]</a> <a href="#">[none]</a> <input checked="" type="checkbox"/> CDK <input type="checkbox"/> Dragon v.6 (all blocks) <input checked="" type="checkbox"/> OEstate and ALogPS <input type="checkbox"/> ISIDA Fragments (Length 2 - 4) <input checked="" type="checkbox"/> GSfrag <input type="checkbox"/> Mera and Mersy <input type="checkbox"/> Chemaxon descriptors <input type="checkbox"/> Inductive Descriptors <input type="checkbox"/> Adriana <input type="checkbox"/> Spectrophores <input type="checkbox"/> Shape Signatures <input type="checkbox"/> QNPR (SMILES - length 1 - 3 threshold 5) <input type="checkbox"/> Two simple descriptors (MW+Number of carbons) <a href="#">+add a custom template</a>	<a href="#">[all]</a> <a href="#">[none]</a> <input checked="" type="checkbox"/> Unsupervised forward selection <input type="checkbox"/> Simple pairwise decorrelation ( $r < 0.95$ ) <input type="checkbox"/> DiffVals=1--Corr=0.95--UFS=False0--std=0.01--max=999999 <a href="#">[edit]</a> <a href="#">[x]</a> <a href="#">+add a custom template</a>	<a href="#">[all]</a> <a href="#">[none]</a> <input checked="" type="checkbox"/> 5-fold cross-validation <input type="checkbox"/> 5-fold cross-validation (stratified) <input type="checkbox"/> Bagging with 64 models <a href="#">+add a custom template</a>

[Show advanced options>>](#)

Considering the selection above, **9 models** will be created.

[Create the models](#)

# Comprehensive modeling

- Multiple models overview for:
  - Comparison of hundreds of models
  - Batch operations with multiple models

\*The screenshot shows the analysis of multiple models for melting point

Multiple models overview

Predicted property: **Melting Point**  
Training set: **MP clean**

Metrics: **RMSE - Root Mean Square Error** for Training set Validation: **Cross-validation**

	ANN	ASNN	KNN	LibSVM	FSMLR	MLRA	PLS	ASNN(2)	LibSVM(2)	ASNN(3)	ASNN(4)	ASNN(5)	ASNN(6)
CDK	46.6	41.3	51.9	57.8	57.7	56.6	60.1	39.3	56.9	39.4	+	38.3	39.7
Dragon6 (blocks: 1-29)	42.7	40.3	54.6	78.7	51.7	49.0	64.9	38.1	78.5	37.6	37.2	38.1	38.1
OEstate, ALogPS	48.7	43.0	53.2	70.8	58.8	58.3	61.4	41.4	62.9	41.0	+	41.7	41.8
Fragmentor (Length 2 - 4)	46.6	42.8	60.7	65.8	58.1	56.1	+	41.2	65.1	40.7	40.6	38.6	38.7
GSFrag	56.8	51.2	59.6	70.8	69.1	68.5	74.7	51.1	69.7	49.4	+	43.2	49.9
Mera, Mersy	49.7	44.7	56.9	63.3	69.3	56.4	79.9	43.1	62.4	42.7	+	41.3	43.0
ChemaxonDescriptors (7.4)	48.2	42.7	50.7	58.5	68.1	58.3	60.1	40.8	57.6	40.5	+	38.8	41.0
InductiveDescriptors	59.6	50.2	59.5	73.1	98.8	68.8	70.4	48.7	72.5	48.3	+	47.9	48.7
Adriana	50.1	43.9	52.4	60.2	60.5	59.4	65.6	42.6	60.2	42.0	+	40.0	42.2
Spectrophores (accuracy=20 Stereospecificity=0 resolution=3.0)	72.5	68.2	71.1	77.8	78.3	77.6	78.1	67.2	77.1	67.1	+	64.8	65.8
ShapeSignatures	70.9	68.0	70.9	79.3	165.0	76.0	78.1	67.1	79.0	67.2	+	66.8	66.6
QNPR (SMILES - length 1 - 3 threshold 5)	48.7	45.0	64.3	64.8	60.2	58.3	60.6	43.3	63.9	42.7	42.7	40.4	43.0
Dragon6 (blocks: 1 28)	76.6	76.1	76.6	82.4	82.2	82.2	82.2	75.5	82.0	75.2	+	75.4	75.6
OEstate, ALogPS	+	+	+	+	55.5	+	+	+	+	+	+	38.8	38.8

# Content

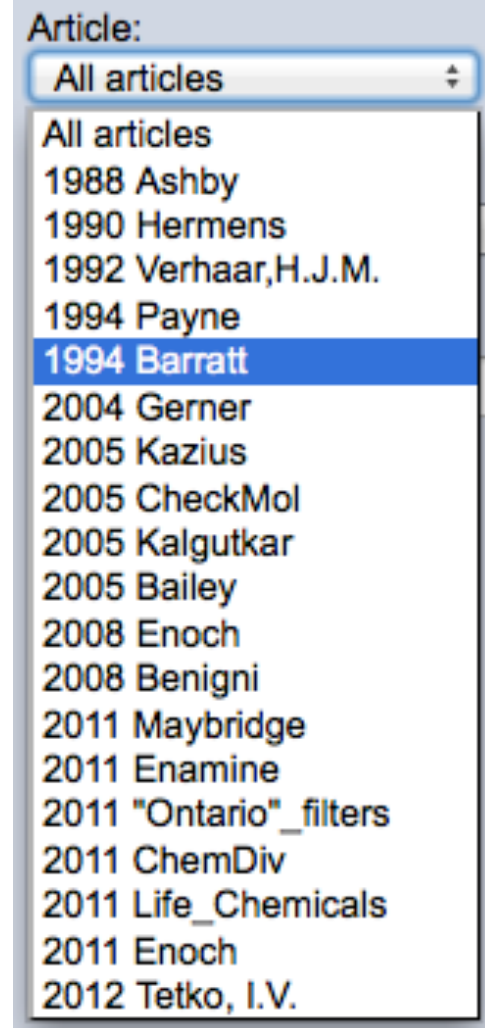
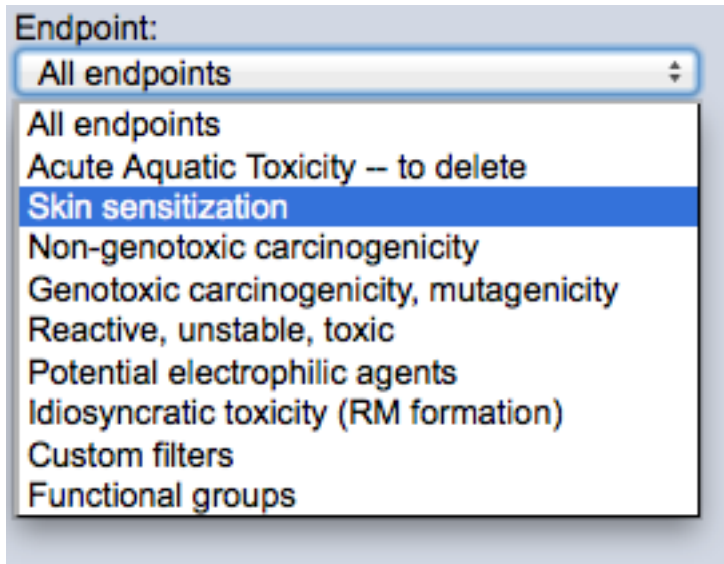
- ✓ OCHEM at a glance (components and Data upload)
- ✓ How to run models for ADME prediction?
- ✓ How to build models (Regression, Classification) and get Applicability domain?
- How to interpret models?
  - Tox Alerts
    - Set Compare
  - System Architecture
  - Case Study
  - Our unique offer

# How to interpret models?

- Multi-learning – development of models for several properties simultaneously (any combination of quantitative and qualitative properties)
- Stratified learning (bagging, cross-validation) for imbalanced datasets
- Feature nets – use prediction of other models as descriptors
- Use experimental (or calculated) values as descriptors
- Use externally calculated descriptors

# ToxAlerts

- Screening millions of compounds against published toxicity alerts
- Filter alerts by endpoints or publications
- Create or upload custom SMARTS rules



# Screening of virtual libraries

**ToxAlerts: Screening results**  
The compounds that matched any alerts grouped by endpoints, publications and by alerts themselves

**ENDPOINTS**

☐ Reactive, unstable, toxic 52021 compounds

**PUBLICATIONS**

☐ 2011 ChemDiv 21989 compounds

☐ 2011 Enamine 33145 compounds

☐ 2011 Life\_Chemicals 26415 compounds

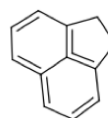
**DETECTED ALERTS**

<input type="radio"/> Allenes	2011 ChemDiv	1 compounds
<input type="radio"/> Cumulated double bonds	2011 Enamine	2 compounds
<input type="radio"/> C, N, O, P and S atoms in unusual valence states	2011 Enamine	40 compounds
<input type="radio"/> Acrylates and similar	2011 ChemDiv	1944 compounds
<input type="radio"/> Positively charged N-heterocycles	2011 ChemDiv	674 compounds
<input type="radio"/> Activated haloaromatics	2011 Enamine	6133 compounds
<input type="radio"/> N,N-Dialkyl aniline derivatives (3)	2011 Enamine	96 compounds
<input type="radio"/> $\beta$ -hydroxy substituted carbonyls	2011 Life_Chemicals	209 compounds
<input type="radio"/> p-Aminoanilines	2011 Enamine	104 compounds
<input type="radio"/> Activated halides ( $\alpha$ -halogen substituted N-heterocycles)	2011 Life_Chemicals	2466 compounds
<input type="radio"/> Over halogenated rings	2011 Life_Chemicals	403 compounds
<input type="radio"/> $\alpha$ -Halogen substituted N-heterocycles	2011 Enamine	2466 compounds
<input type="radio"/> Iodine	2011 Life_Chemicals	566 compounds
<input type="radio"/> Polycyclic 4 fused rings and more	2011 Life_Chemicals	1326 compounds
<input type="radio"/> Ketones	2011 Enamine	2081 compounds
<input type="radio"/> Imines	2011 Life_Chemicals	5298 compounds
<input type="radio"/> Other undesirable polycyclic (adamantane derivatives)	2011 Life_Chemicals	1551 compounds
<input type="radio"/> Acrylamides	2011 ChemDiv	3845 compounds
<input type="radio"/> Michael acceptors ( $\alpha,\beta$ -unsaturated carbonyls)	2011 Life_Chemicals	3813 compounds
<input type="radio"/> Aldimines and ketimines	2011 ChemDiv	1178 compounds
<input type="radio"/> Hydrazones and similar	2011 ChemDiv	4167 compounds
<input type="radio"/> N-N Single bound not in a ring	2011 Life_Chemicals	4225 compounds
<input type="radio"/> Ketal	2011 Life_Chemicals	265 compounds
<input type="radio"/> Singel acyclic N-N bonds	2011 Enamine	5724 compounds
<input type="radio"/> Lipophilic	2011 ChemDiv	4505 compounds

View records for the filtered compounds Tag the 52021 filtered molecules

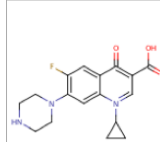
1 - 15 of 52021

15 items on page 1 of 3469 > >>



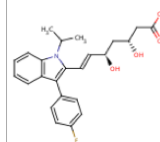
**Allen**es (for Reactive, unstable, toxic in 2011 ChemDiv)  
**Cumulated double bonds** (for Reactive, unstable, toxic in 2011 Enamine)  
**C, N, O, P and S atoms in unusual valence states** (for Reactive, unstable, toxic in 2011 Enamine)

MoleculeID: M1711



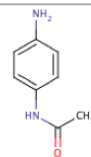
**Acrylates and similar** (for Reactive, unstable, toxic in 2011 ChemDiv)  
**Positively charged N-heterocycles** (for Reactive, unstable, toxic in 2011 ChemDiv)  
**Activated haloaromatics** (for Reactive, unstable, toxic in 2011 Enamine)  
**Cumulated double bonds** (for Reactive, unstable, toxic in 2011 Enamine)  
**C, N, O, P and S atoms in unusual valence states** (for Reactive, unstable, toxic in 2011 Enamine)  
**N,N-Dialkyl aniline derivatives (3)** (for Reactive, unstable, toxic in 2011 Enamine)

MoleculeID: M7123



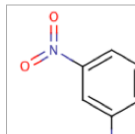
**$\beta$ -hydroxy substituted carbonyls** (for Reactive, unstable, toxic in 2011 Life\_Chemicals)

MoleculeID: M2647



**p-Aminoanilines** (for Reactive, unstable, toxic in 2011 Enamine)

MoleculeID: M1350



**C, N, O, P and S atoms in unusual valence states** (for Reactive, unstable, toxic in 2011 Enamine)

MoleculeID: M12971

# ToxAlerts

- Extension with > 300 functional groups
  - Filters for frequent hitters
  - 158 for promiscuous compounds [1];
  - 480 for Pan Assay Interference (PAIN) compounds [2].
- 
- Collaboration with European ScreeningPort GmbH and HMGU to develop new filters for alpha screens frequent hitters
- 
- [1] B. C. Pearce, *et al.* *J. Chem. Inf. Model.* **2009**, 46, 1060-1068.
  - [2] J. B. Baell, G. A. Holloway. *J. Med. Chem.* **2010**, 53, 2719-2740.

# Content

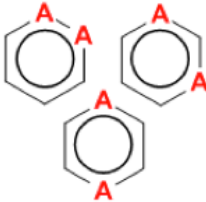
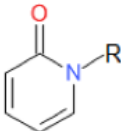
- ✓ OCHEM at a glance (components and Data upload)
- ✓ How to run models for ADME prediction?
- ✓ How to build models (Regression, Classification) and get Applicability domain?
- How to interpret models?
  - ✓ Tox Alerts
    - Set Compare
  - System Architecture
  - Case Study
  - Our unique offer



# SetCompare

- Comparison of structural features/toxic groups

## Soluble and in-soluble in DMSO

Descriptor	In set 1 (2681 molecules)	In set 2 (47939 molecules)	p-Value
<b>LS</b> 	1129	7551	7.49E-216
	985	6612	3.63E-180

## Biodegradable and not

Descriptor	In set 1 (1221 molecules)	In set 2 (717 molecules)	p-Value
<b>Halogens</b> <b>F</b> <b>Cl</b> <b>Br</b> <b>I</b> <b>At</b>	384	49	4.08E-41
R— <b>X</b>	355	40	1.03E-40

# Content

- ✓ OCHEM at a glance (components and Data upload)
- ✓ How to run models for ADME prediction?
- ✓ How to build models (Regression, Classification) and get Applicability domain?
- ✓ How to interpret models?
  - ✓ Tox Alerts
  - ✓ Set Compare
- System Architecture
  - Case Study
  - Our unique offer

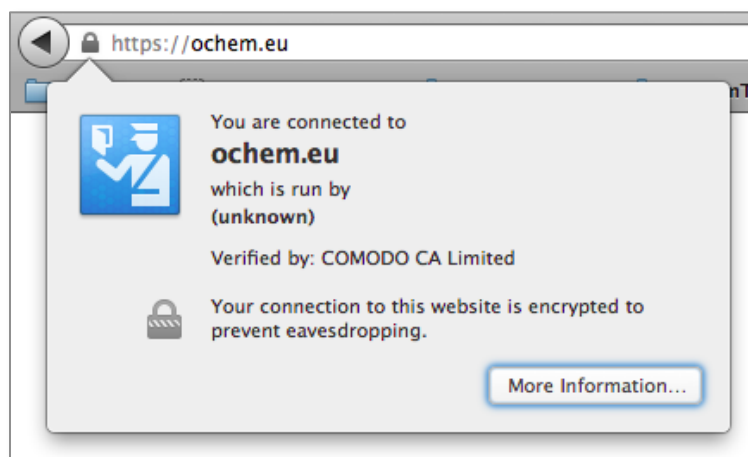
# System Architecture

- Powerful engines integrated
- Automatic control of workflow and parallelization of tedious computational tasks



# Engines and Protocols

- Xemistry substructure search is integrated into OCHEM
- Chemaxon Standardizer is used for standardizing chemical structures and handling different file formats
- MySQL and NoSQL are used as database engines
- All communications with the online platform is protected by SSL



# Calculations distribution

- The system can be implemented on any Grid, Cloud or distributed calculation environment.
- Current implementations include Amazon Cloud, VMWare ESX and Sun Grid Engines
- Implementation on more than 500 CPU cores
- Completely automated distribution of tasks, collection of results and update of the distributed code to the latest versions.
- Simple interface to control the number of running calculation instances
- Configurable tasks priority

# Content

- ✓ OCHEM at a glance (components and Data upload)
- ✓ How to run models for ADME prediction?
- ✓ How to build models (Regression, Classification) and get Applicability domain?
- ✓ How to interpret models?
  - ✓ Tox Alerts
  - ✓ Set Compare
- ✓ System Architecture
- Case Study
  - Our unique offer

# Case Study: Comparison of Different Algorithms

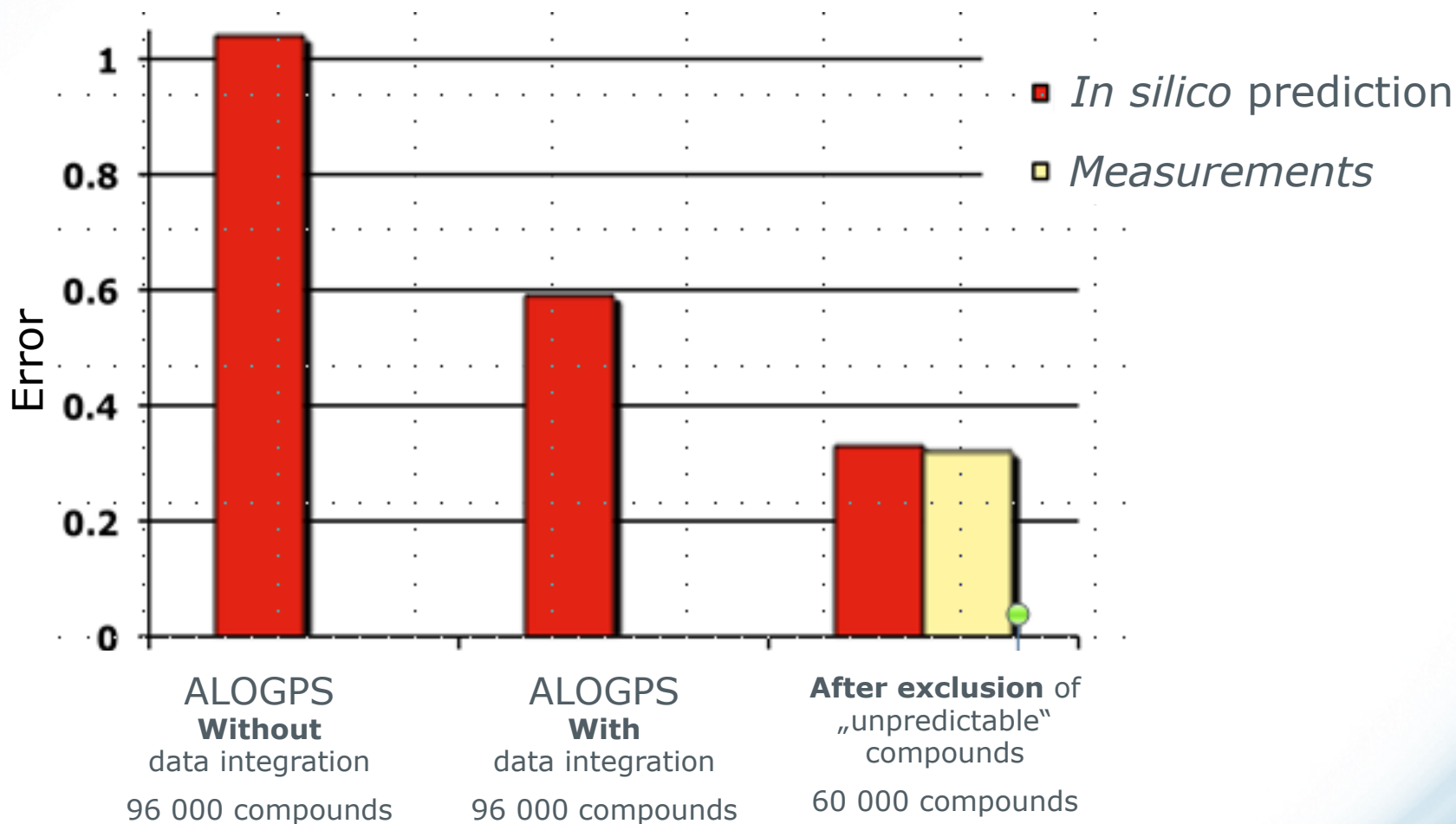
**Algorithms**  
(without data integration)



Method	RMSE	Failed <sup>1</sup>	rank	% in error range			RMSE, zwitterions excluded <sup>2</sup>	RMSE	rank	% in error range		
				<0.5	0.5-1	>1				<0.5	0.5-1	>1
<b>eADMET</b> → <b>ALOGPS</b>	1.02		I	41	30	29	1.01	0.68	I	51	34	15
<b>Competitor</b> → <b>S+logP</b>	1.02		I	44	29	27	1.00	0.69	I	58	27	15
NC+NHET	1.04		II	38	30	32	1.04	0.88	III	42	32	26
<b>MLOGP(S+)</b>	1.05		II	40	29	31	1.05	1.17	III	32	26	41
XLOGP3	1.07		II	43	28	29	1.06	0.65	I	55	34	12
<b>MiLogP</b>	1.10	27	II	41	28	30	1.09	0.67	I	60	26	14
<b>AB/LogP</b>	1.12	24	II	39	29	33	1.11	0.88	III	45	28	27
<b>ALOGP</b>	1.12		II	39	29	32	1.12	0.72	II	52	33	15
<b>ALOGP98</b>	1.12		II	40	28	32	1.10	0.73	II	52	31	17
OsirisP	1.13	6	II	39	28	33	1.12	0.85	II	43	33	24
AAM	1.16		III	33	29	38	1.16	0.94	III	42	31	27
<b>CLOGP</b>	1.23		III	37	28	35	1.21	1.01	III	46	28	22
<b>Competitor</b> → <b>ACD/logP</b>	1.28		III	35	27	38	1.28	0.87	III	46	34	21
<b>CSlogP</b>	1.29	20	III	37	27	36	1.28	1.06	III	38	29	33
<b>COSMOFrag</b>	1.30	1088 <sup>3</sup>	III	32	27	40	1.30	1.06	III	29	31	40
<b>QikProp</b>	1.32	103	III	31	26	43	1.32	1.17	III	27	24	49
KowWIN	1.32	16	III	33	26	41	1.31	1.20	III	29	27	44
QLogP	1.33	24	III	34	27	39	1.32	0.80	II	50	33	17
XLOGP2	1.80		III	15	17	68	1.80	0.94	III	39	31	29
<b>MLOGP(Dragon)</b>	2.03		III	34	24	42	2.03	0.90	III	45	30	25

Decreasing prediction error,  
Increasing performance

# Case Study: Accuracy of eADMET's Predictions





# Content

- ✓ OCHEM at a glance (components and Data upload)
- ✓ How to run models for ADME prediction?
- ✓ How to build models (Regression, Classification) and get Applicability domain?
- ✓ How to interpret models?
  - ✓ Tox Alerts
  - ✓ Set Compare
- ✓ System Architecture
- ✓ Case Study
- Our unique offer

# Our unique offer

- collaborative approach to data handling  
users can add, modify and delete data; can use, create and publish model; create hidden and public data
- Mandatory reference to an origin of information  
each record in a database should contain a reference to a source (article, book, proceeding or even personal communications), where the data were published
- Storing rich meta-information  
measurement conditions to increase data quality
- Several tools to automatize and support decision making  
integration with Knime (Pipe-Line pilot), provides and consumes web services, estimates accuracy of prediction; manage duplicated records
- Aimed at model building  
filter by property, article, substructure, export data either to internal modelling tools or download as Excel file; models with >150,000 molecules; sparse data format (millions of descriptors); support of various descriptors and machine learning tools
- Advanced Analysis tools  
chemogenomic approaches; structural alerts; SetCompare utility; support of mixtures



**Thank you**