On-line chemoinformatics tools for SAR/QSAR model development

Igor V. Tetko
Institute of Structural Biology
Helmholtz Zentrum Muenchen

13th October 2016, on-line course of the BIGCHEM project

HelmholtzZentrum münchen
German Research Center for Environmental Health

Outline

Overview of existing on-line services on the web

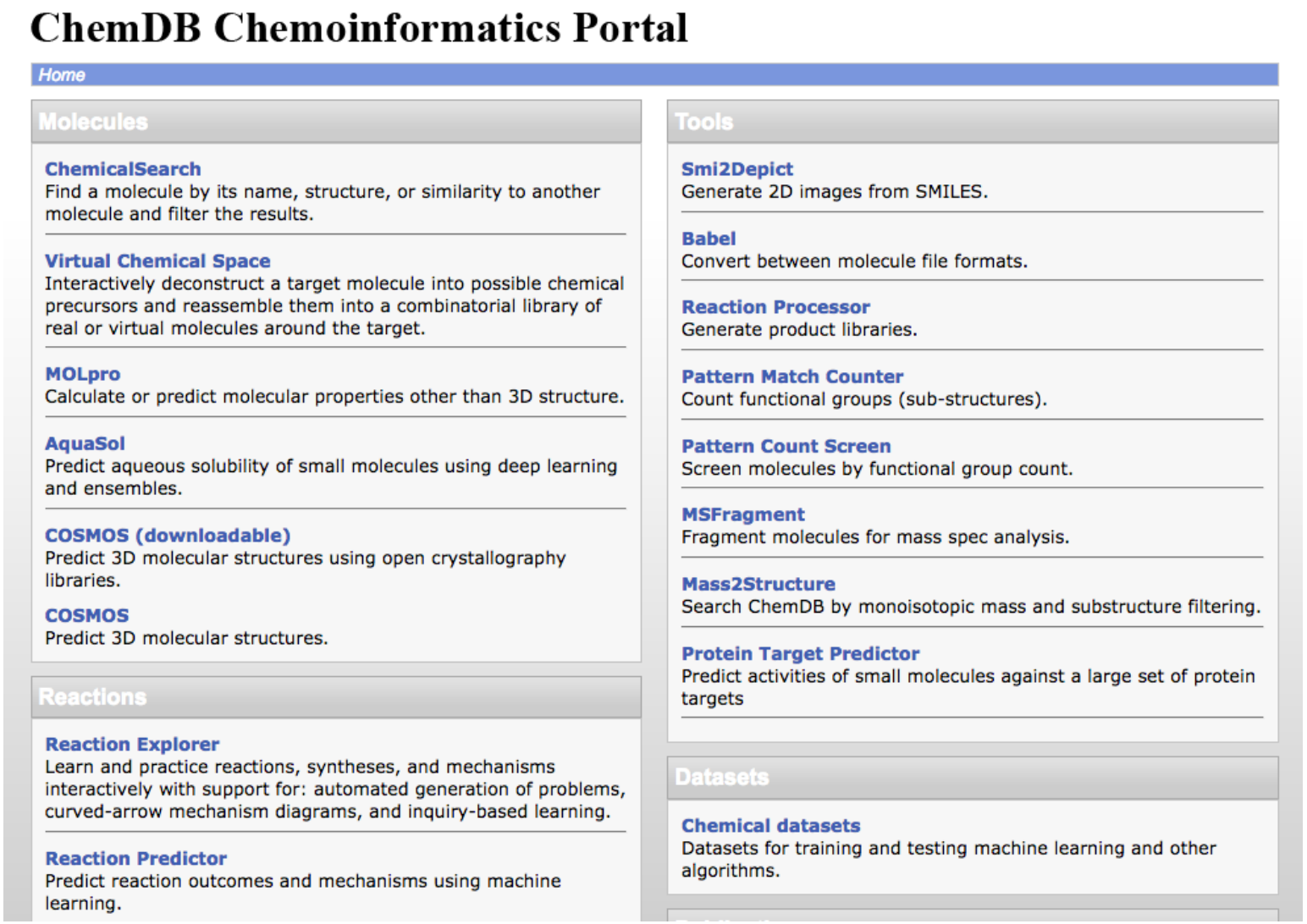
- Individual groups web servers
- Projects web servers
- Model repositories
- Model development platforms

OCHEM

- Typical uses of OCHEM
- Model development
- Model interpretation
- Model evaluation

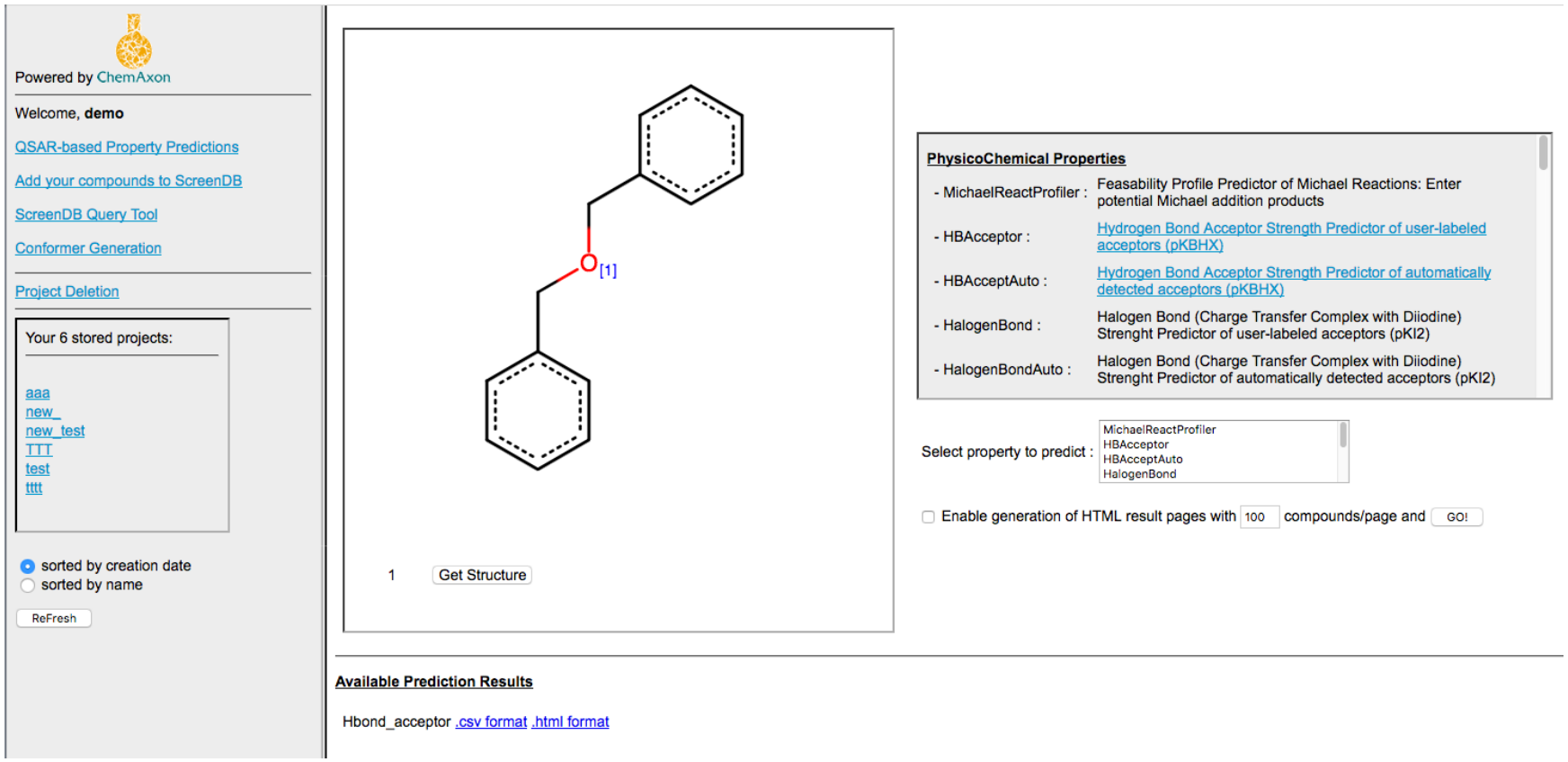
Individual groups web servers

http://cdb.ics.uci.edu



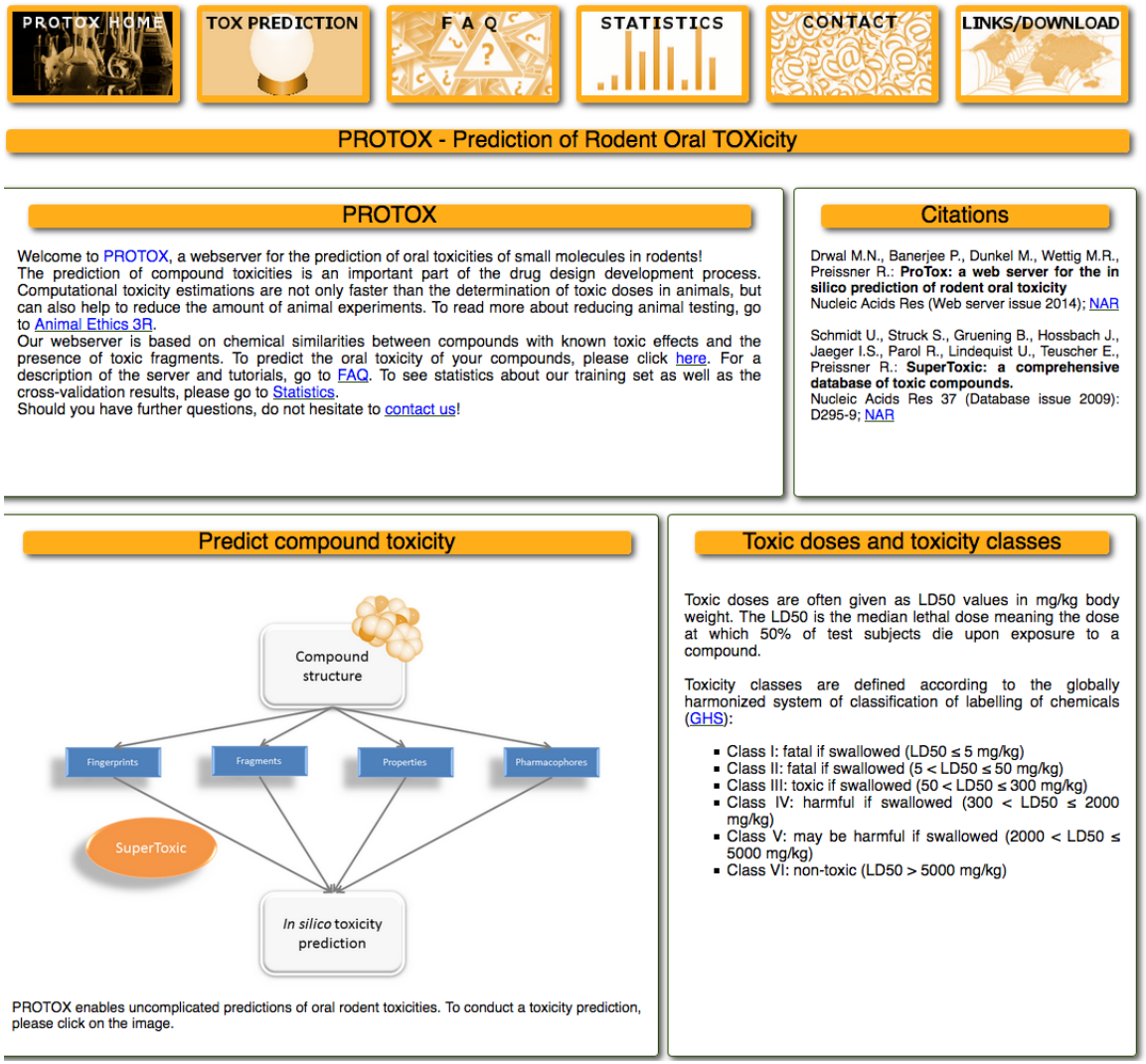
Prof. P. Baldi, UCI

http://infochim.u-strasbg.fr/webserv/VSEngine.html



Prof. A. Varnek, Strasbourg

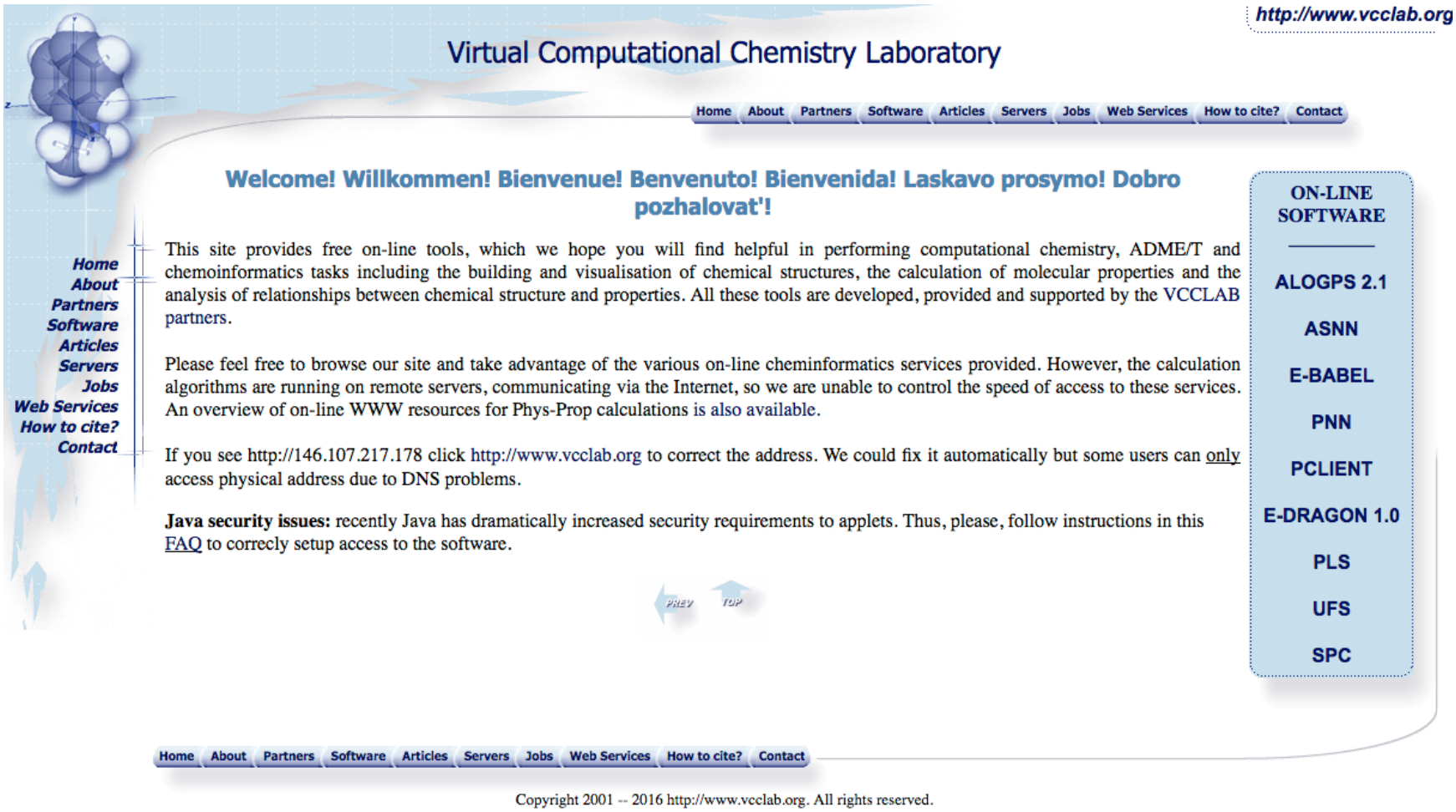
http://tox.charite.de/tox



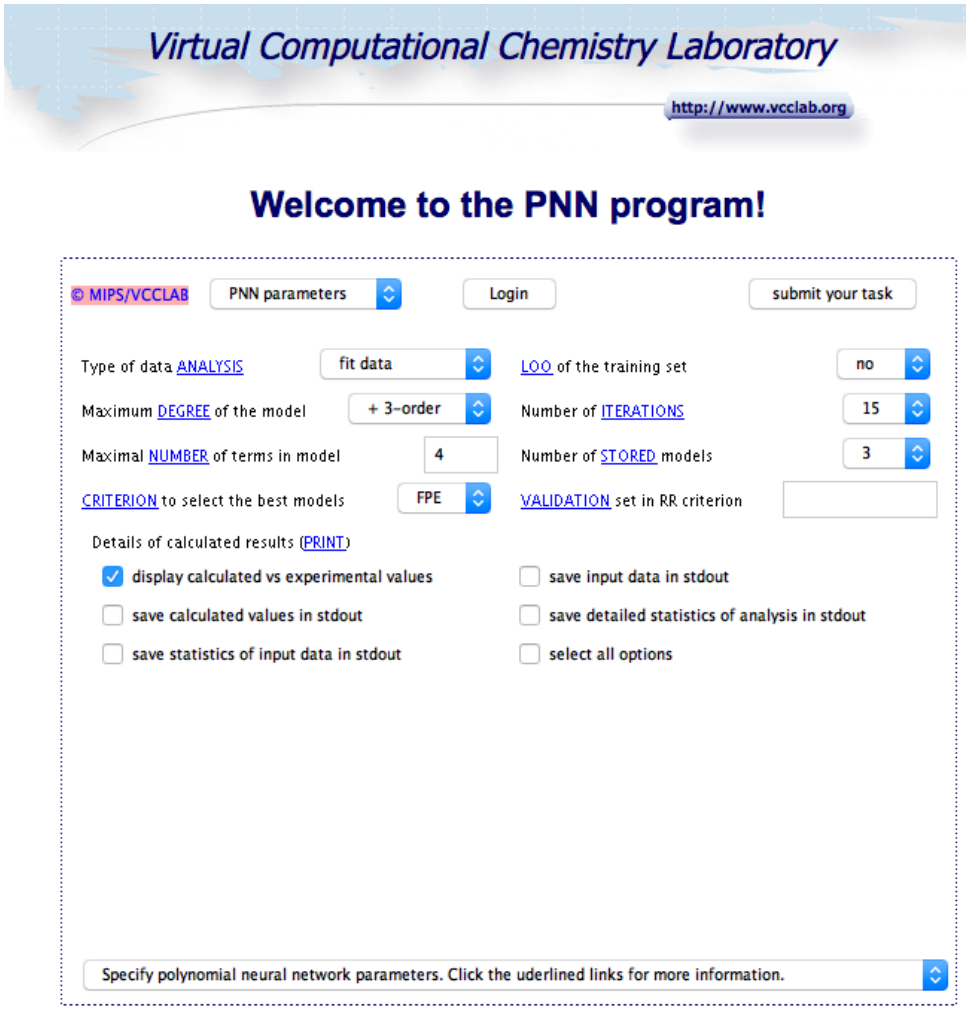
Dr. R. Preissner, Charite University of Medicine

Projects web servers

http://www.vcclab.org

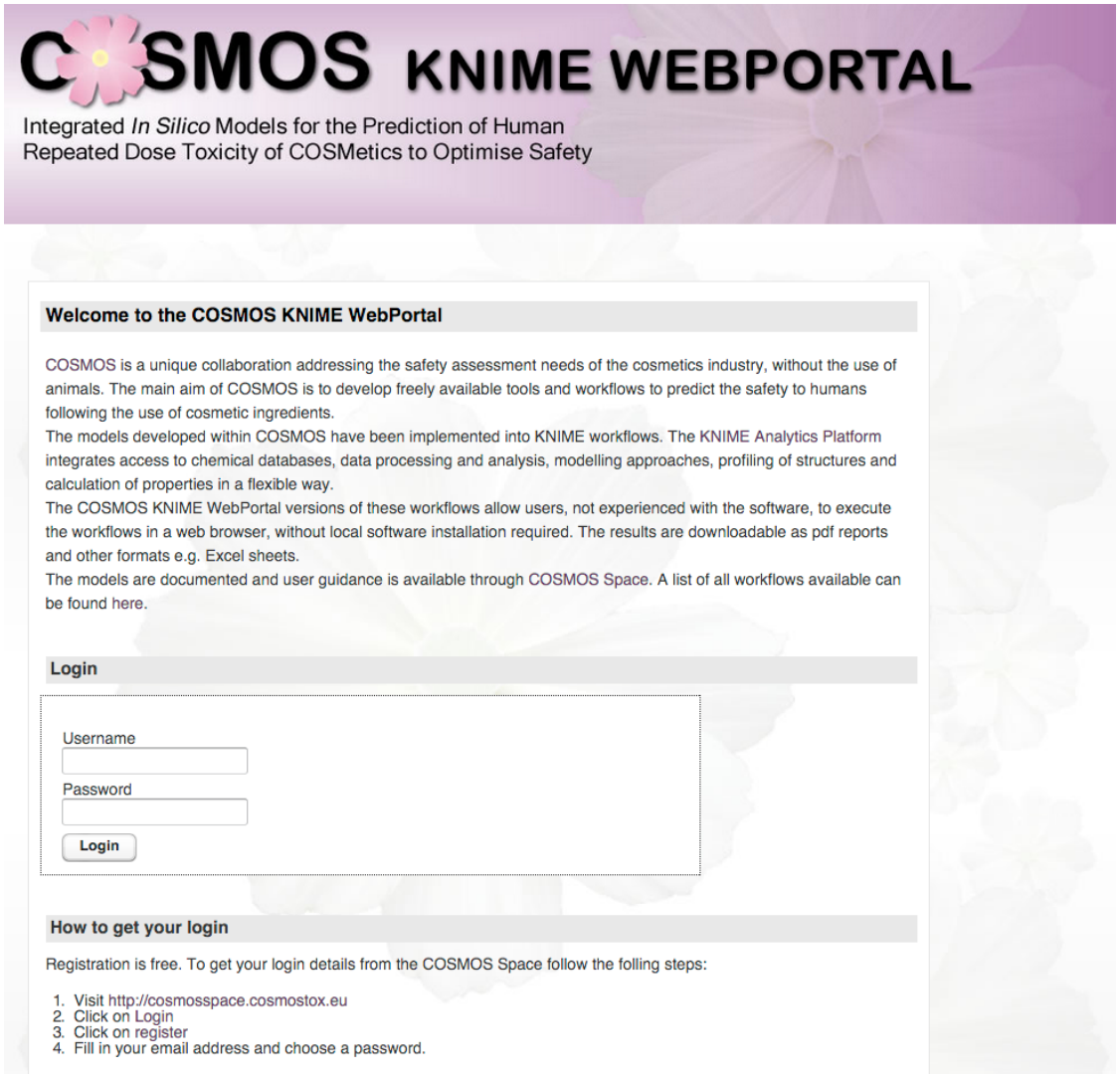


A first QSAR web tool: Polynomial Neural Networks

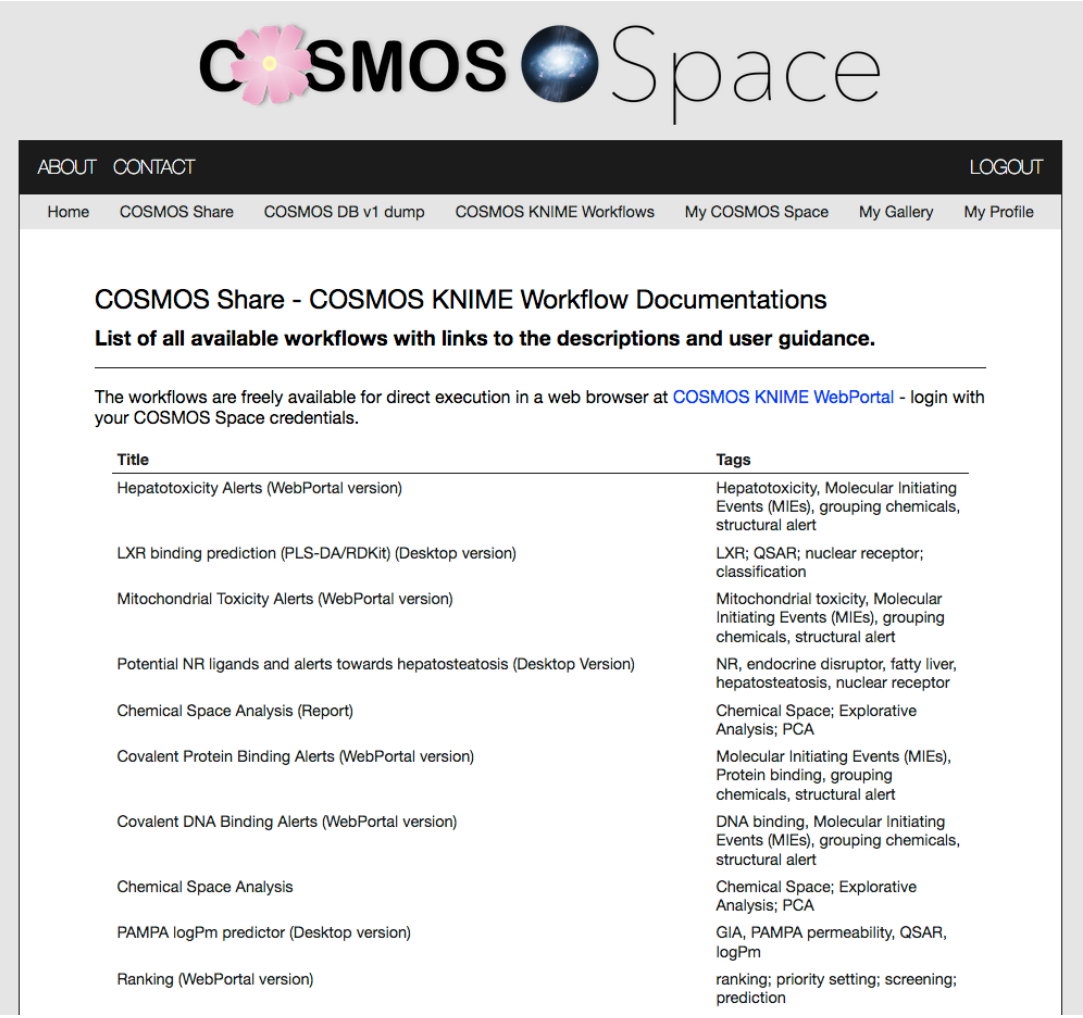


I. V. Tetko, T. I. Aksenova, V. V. Volkovich, T. N. Kasheva, D. V. Filipov, W. J. Welsh, D. J. Livingstone, A. E. P. Villa, *SAR QSAR Environ. Res.* **2000**, *11*, 263-280.

http://knimewebportal.cosmostox.eu

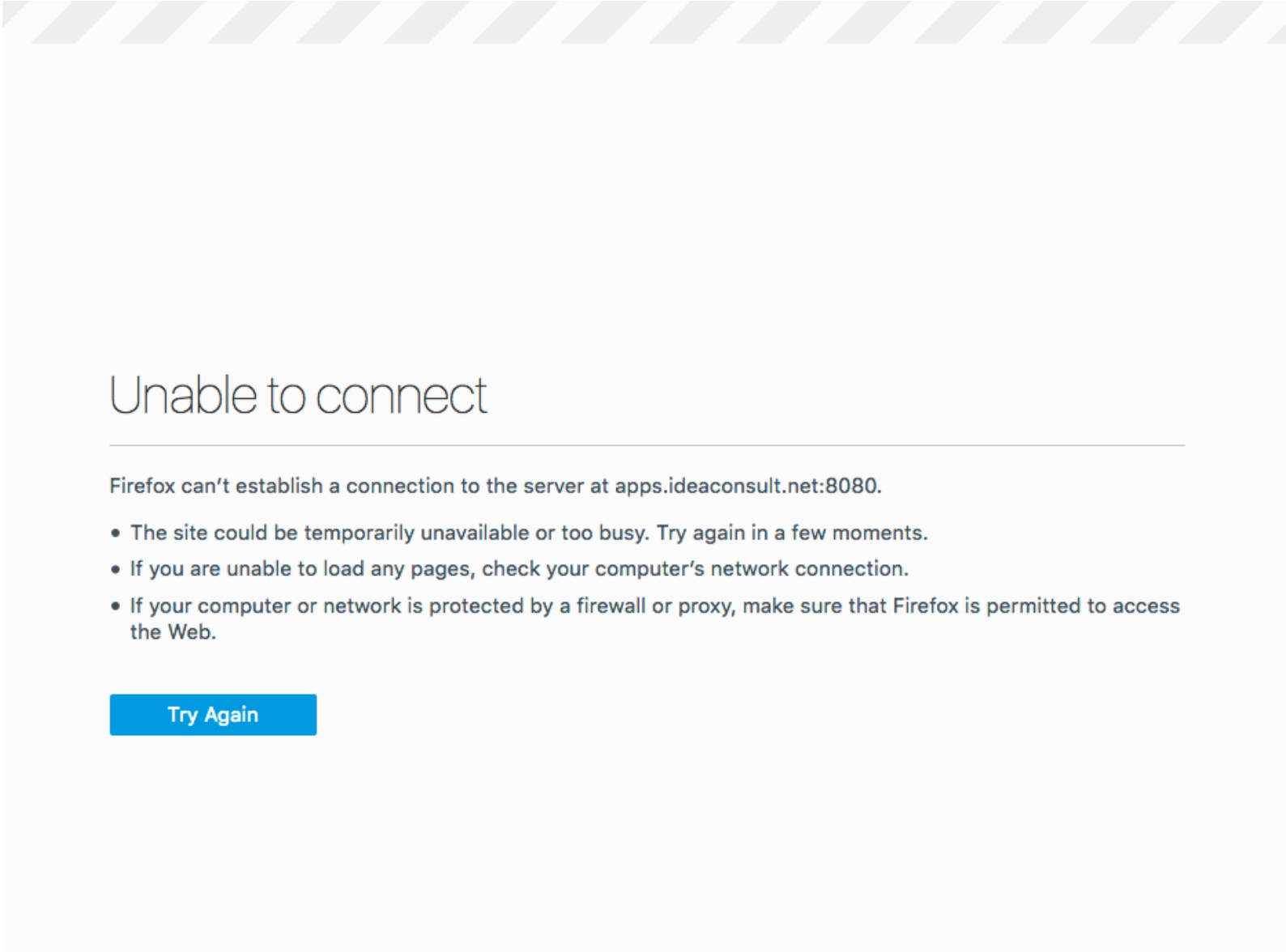


http://cosmosspace.cosmostox.eu/app/space/cosmosshare/ckwdprojects?pn=2



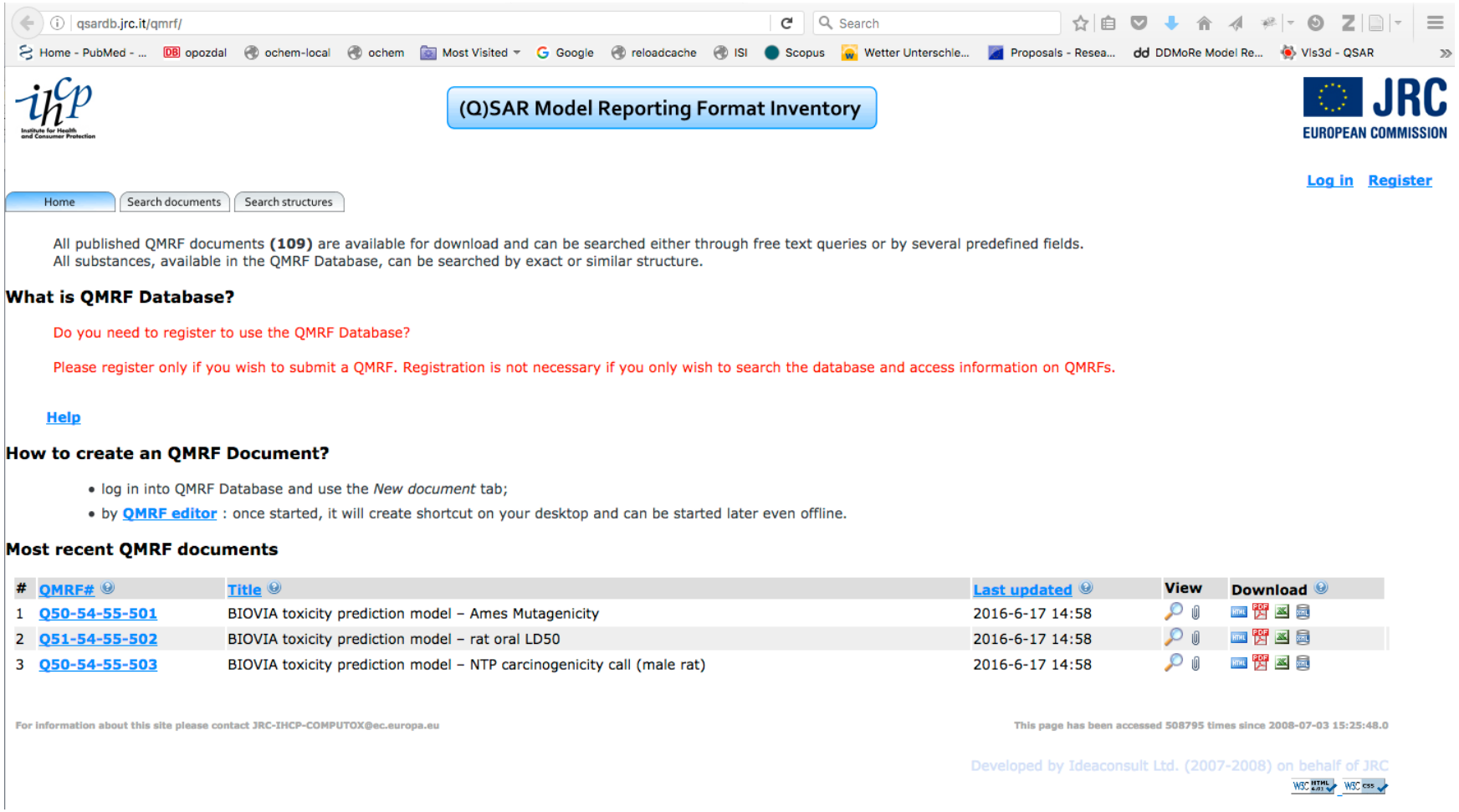
http://www.vega-qsar.eu/

http://toxpredict.org



Model repositories

http://qsardb.jrc.it/qmrf/



qsardb.jrc.it/qmrf/

Search

Home - PubMed - ... | opozdal | ochem-local | ochem | Most Visited | Google | reloadcache | ISI | Scopus | Wetter Unterschle... | Proposals - Resea... | DDMoRe Model Re... | Vis3d - QSAR

ihcp Institute for Health and Consumer Protection

(Q)SAR Model Reporting Format Inventory

JRC EUROPEAN COMMISSION

Log in Register

Home | Search documents | Search structures

All published QMRF documents **(109)** are available for download and can be searched either through free text queries or by several predefined fields.
All substances, available in the QMRF Database, can be searched by exact or similar structure.

What is QMRF Database?

Do you need to register to use the QMRF Database?

Please register only if you wish to submit a QMRF. Registration is not necessary if you only wish to search the database and access information on QMRFs.

Help

How to create an QMRF Document?

- log in into QMRF Database and use the *New document* tab;
- by QMRF editor : once started, it will create shortcut on your desktop and can be started later even offline.

Most recent QMRF documents

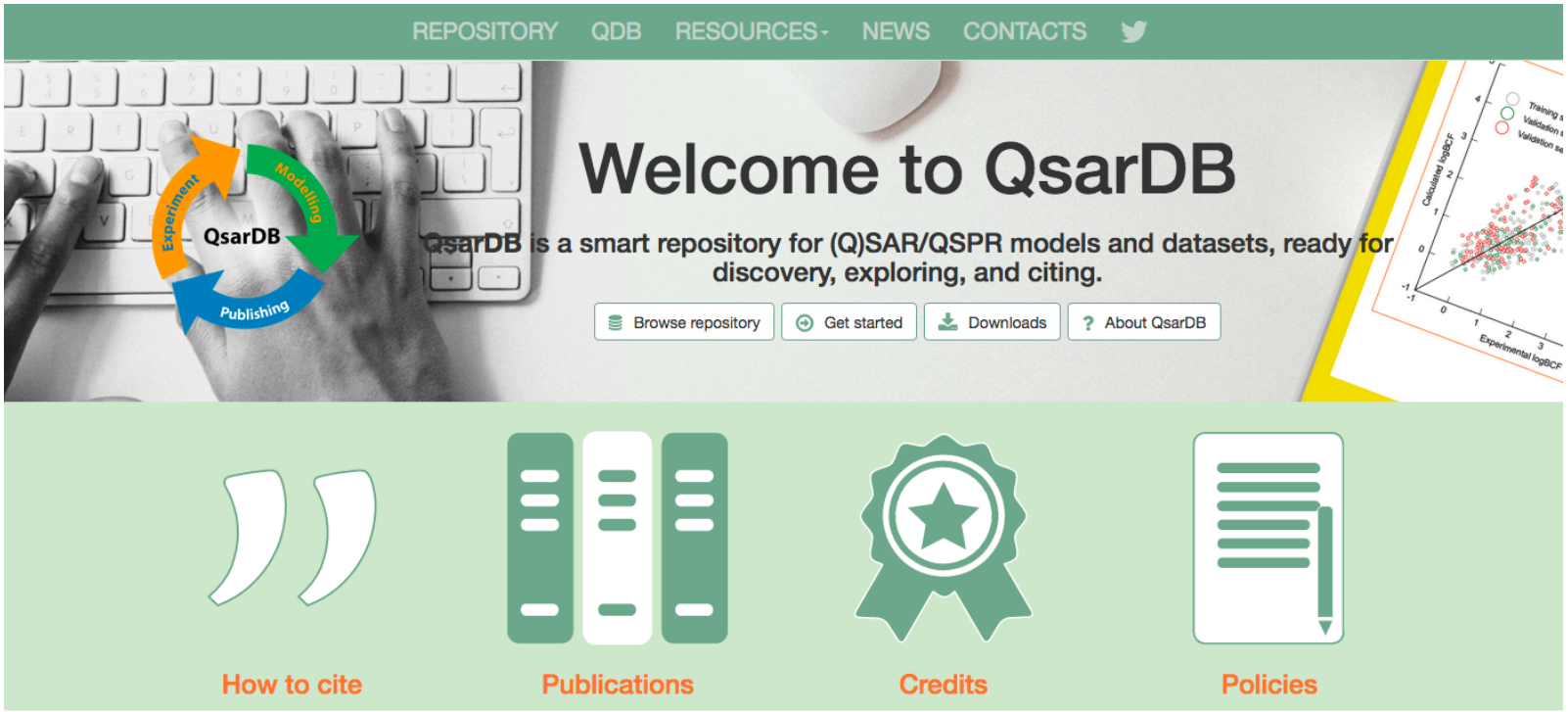
#	QMRF#	Title	Last updated	View	Download
1	Q50-54-55-501	BIOVIA toxicity prediction model – Ames Mutagenicity	2016-6-17 14:58		
2	Q51-54-55-502	BIOVIA toxicity prediction model – rat oral LD50	2016-6-17 14:58		
3	Q50-54-55-503	BIOVIA toxicity prediction model – NTP carcinogenicity call (male rat)	2016-6-17 14:58		

For information about this site please contact JRC-IHCP-COMPUTOX@ec.europa.eu

This page has been accessed 508795 times since 2008-07-03 15:25:48.0

Developed by Ideaconsult Ltd. (2007-2008) on behalf of JRC

http://qsardb.org



Recent submissions to the repository

07 Oct 2016	Ran, Y.; He, Y.; Yang, G.; Johnson, J. L. H.; Yalkowsky, S. H. Estimation of aqueous solubility of organic compounds by using the general solubility equation. Chemosphere 2002, 48, 487–509.
27 Sep 2016	Li, F.; Wu, H.; Li, L.; Li, X.; Zhao, J.; Peijnenburg, W. J. G. M. Docking and QSAR study on the binding interactions between polycyclic aromatic hydrocarbons and estrogen receptor. Ecotoxicology and Environmental Safety 2012, 80, 273 - 279.
20 Sep 2016	Papa, E.; Pilutti, P.; Gramatica, P. Prediction of PAH mutagenicity in human cells by QSAR classification. SAR and QSAR in Environmental Research 2008, 19, 115–127.
19 Sep 2016	Kar, S.; Deeb, O.; Roy, K. Development of classification and regression based QSAR models to predict rodent carcinogenic potency using oral slope factor. Ecotoxicology and Environmental Safety 2012, 82, 85–95.

Model development platforms

http://chembench.mml.unc.edu

Chembench

Home | My Bench | Datasets | Modeling | Prediction

Accelerating chemical genomics research

Chembench is a free portal that enables researchers to mine available chemical and biological data. Chembench can help researchers rationally design or select new compounds or compound libraries with significantly enhanced hit rates in screening experiments.

Chembench provides cheminformatics research support to molecular modelers, medicinal chemists and quantitative biologists by integrating robust model builders, property and activity predictors, virtual libraries of available chemicals with predicted biological and drug-like properties, and special tools for chemical library design. Chembench was initially developed to support researchers in the Molecular Libraries Probe Production Centers Network (MLPCN) and the Chemical Synthesis Centers.

Please log in

Username:

Password:

Log in

Or, log in as a guest

Forgot your password?
Register an account

Help & Resources

- Chembench Overview
- Workflows & Methodology

Site Stats

Total visitors: 1229488
Registered users: 503
Jobs completed: 59624
Compute time used: 76.94 years
Current users: 5
Running jobs: 34

We thank the following commercial sponsors for their support:

CHEMICAL COMPUTING GROUP
Scalable Software. Scalable Science.

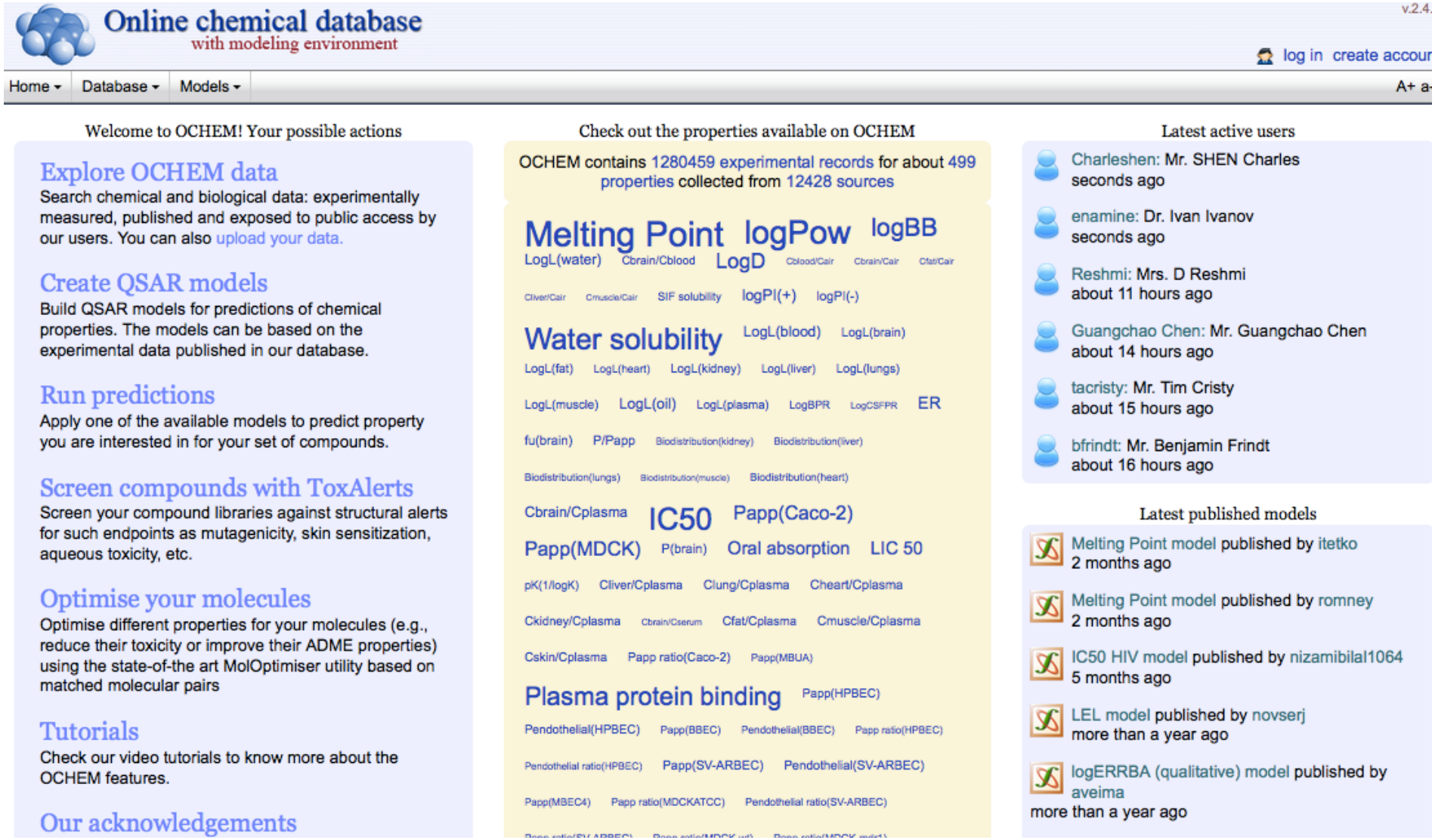
kode chemoinformatics

ChemAxon

eduSoft

SUNSET MOLECULAR

Data storage and model development: http://ochem.eu



Typical use of OCHEM

Online chemical database
with modeling environment

v.2.4.32

Welcome, Dear Dr.Tetko! My account Logout

Home ▾ | Database ▾ | Models ▾ | Moderation ▾ | A+ a-

FILTERS

▾ SOURCE
Article/Source [select]
Page Table

▾ PROPERTY
Activity/Property [select]
Hide records without property

▸ CONDITIONS

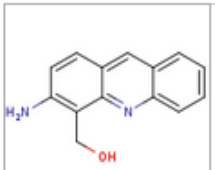
▾ MOLECULE FILTERS
Name / OCHEM ID / Inchi-Key
Similarity/substructure search
Draw a structure and search all the molecules containing it or similar to it
CLICK TO DRAW A STRUCTURE
Molecular mass between and

▸ ADVANCED MOLECULE FILTERS

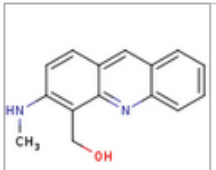
▾ MISCELLANEOUS
Current set
Show all
Data origin and quality:
Data introducers: All users
Data visibility: Public and private
Data from other users: Only approved data

Basket Records Tags

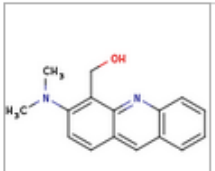
1 - 5 of 1075221 5 items on page 1 of 215045 > >>


molecule profile

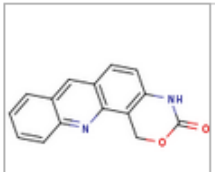
● logPow = 1.77 (in Log unit)
Charmantray, F et al
4-Hydroxymethyl-3-aminoacridine derivatives as a new family ...
N: 1 P: 970 T: 1
J. Med. Chem. **2003**; 46 (6) 967-77
MoleculeID: *M11715*
Public record
RecordID: *R1*
09:35, 19 Mar 09 / 10:04, 13 Oct 11
i.tetko / itetko


molecule profile

● logPow = 2.1 (in Log unit)
Charmantray, F et al
4-Hydroxymethyl-3-aminoacridine derivatives as a new family ...
N: 2 P: 970 T: 1
J. Med. Chem. **2003**; 46 (6) 967-77
MoleculeID: *M9560*
Public record
RecordID: *R2*
09:35, 19 Mar 09
i.tetko

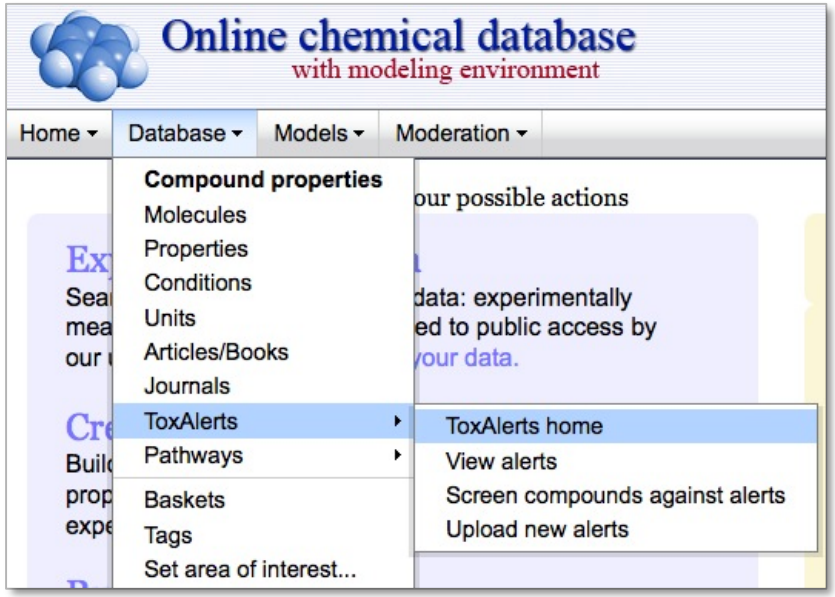

molecule profile

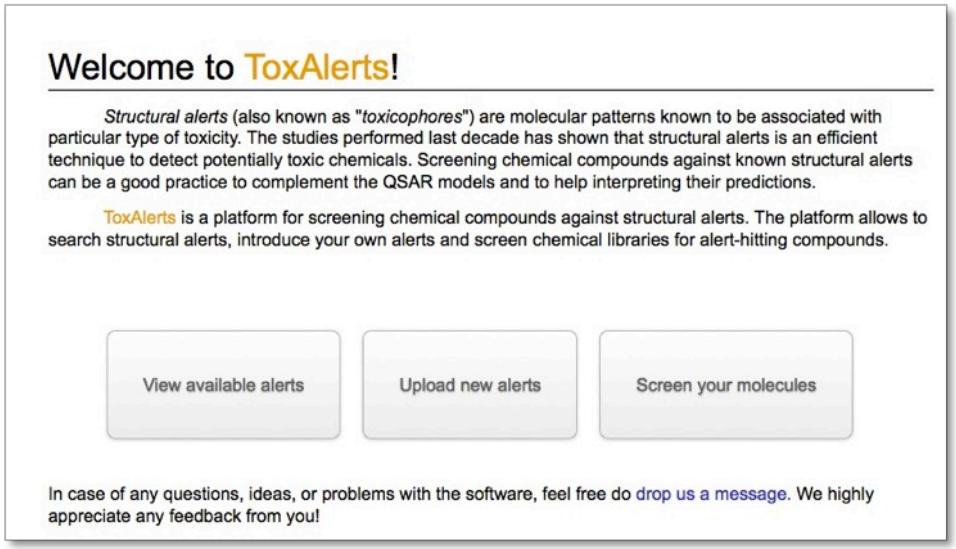
● logPow = 2.2 (in Log unit)
Charmantray, F et al
4-Hydroxymethyl-3-aminoacridine derivatives as a new family ...
N: 3 P: 970 T: 1
J. Med. Chem. **2003**; 46 (6) 967-77
MoleculeID: *M10111*
Public record
RecordID: *R3*
09:35, 19 Mar 09
i.tetko


molecule profile

● logPow = 1.64 (in Log unit)
Charmantray, F et al
4-Hydroxymethyl-3-aminoacridine derivatives as a new family ...
N: 4 P: 970 T: 1
J. Med. Chem. **2003**; 46 (6) 967-77
MoleculeID: *M9777*
Public record
RecordID: *R4*
09:35, 19 Mar 09
i.tetko

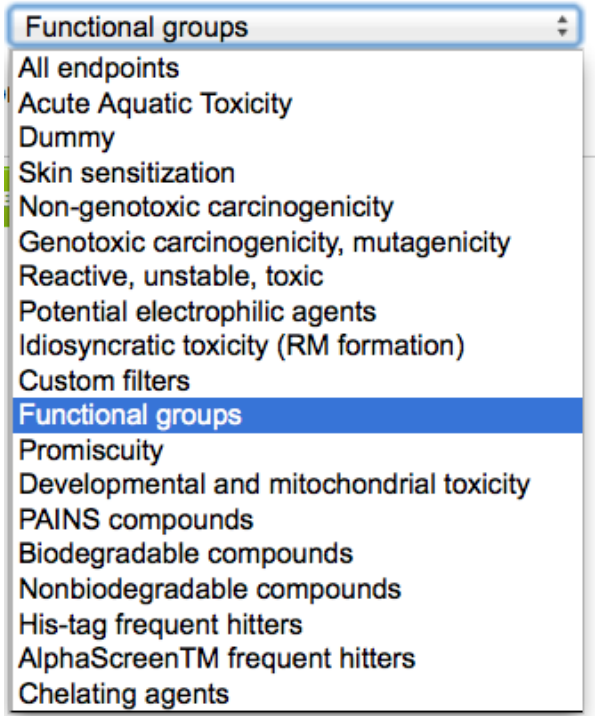
Working with ToxAlerts

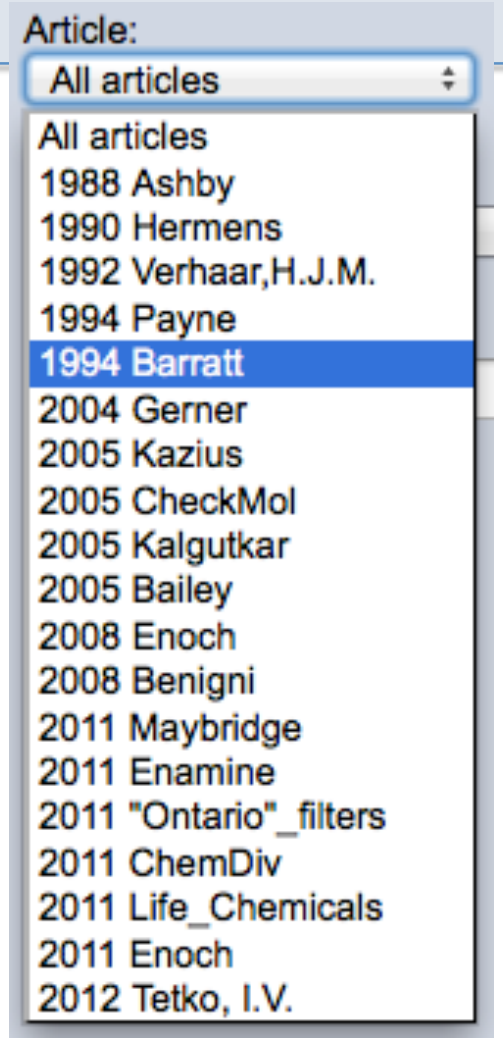




ToxAlerts

- Screening of compounds against published toxicity alerts, groups, frequent hitters
- Filter alerts by endpoints or publications
- Create or upload custom SMARTS rules





Sushko et al, JCIM, 2012, 52(8):2310-6.

Analysis of virtual libraries

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

ENDPOINTS

Functional groups		1890 compounds

PUBLICATIONS

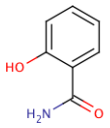
2005 CheckMol		1880 compounds
2012 Salmina		1890 compounds

DETECTED ALERTS

Hydroxy compounds: alcohols or phenols	2005 CheckMol	416 compounds
Phenols	2005 CheckMol	211 compounds
Carboxylic acid derivatives	2005 CheckMol	542 compounds
Carboxylic acid amides	2005 CheckMol	101 compounds
Carboxylic acid primary amides	2005 CheckMol	14 compounds
Aromatic compounds	2005 CheckMol	993 compounds
Arenes	2005 CheckMol	915 compounds
Nonmetals	2012 Salmina	1890 compounds
Chalcogens (oxygen group)	2012 Salmina	1433 compounds
Pnictogens (nitrogen group)	2012 Salmina	800 compounds
Tetragens (carbon group)	2012 Salmina	1890 compounds
Nitriles	2005 CheckMol	55 compounds
Carboxylic acid amidines	2005 CheckMol	4 compounds
Heterocyclic compounds	2005 CheckMol	309 compounds
Five-membered heterocycles (LS)	2012 Salmina	143 compounds
Five-membered heterocycles with two heteroatoms (LS)	2012 Salmina	61 compounds
Unsaturated five-membered heterocycles with two heteroatoms (LS)	2012 Salmina	6 compounds
Five-membered heterocycles (HS)	2012 Salmina	69 compounds
Five-membered heterocycles with two heteroatoms (HS)	2012 Salmina	31 compounds
Unsaturated five-membered heterocycles with two heteroatoms (HS)	2012 Salmina	5 compounds
Imidazolines (HS)	2012 Salmina	2 compounds
2-Imidazolines (HS)	2012 Salmina	2 compounds
Carbonyl compouns: aldehydes or ketones	2005 CheckMol	177 compounds
Aldehydes	2005 CheckMol	53 compounds
Thioethers	2005 CheckMol	14 compounds
Dialkylthioethers	2005 CheckMol	5 compounds
Carbonyl compounds: aldehydes and ketones	2005 CheckMol	170 compounds
Amines	2005 CheckMol	316 compounds
Primary aliphatic amines	2005 CheckMol	59 compounds
Primary amines	2005 CheckMol	180 compounds

View records for the filtered compounds | Tag the 1890 filtered molecules | Export the screening results

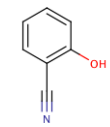
1 - 15 of 1890 — 15 items on page — 1 of 126 > >>



molecule profile

Hydroxy compounds: alcohols or phenols (for Functional groups in 2005 CheckMol)
Phenols (for Functional groups in 2005 CheckMol)
Carboxylic acid derivatives (for Functional groups in 2005 CheckMol)
Carboxylic acid amides (for Functional groups in 2005 CheckMol)
Carboxylic acid primary amides (for Functional groups in 2005 CheckMol)
Aromatic compounds (for Functional groups in 2005 CheckMol)
Arenes (for Functional groups in 2005 CheckMol)
Nonmetals (for Functional groups in 2012 Salmina)
Chalcogens (oxygen group) (for Functional groups in 2012 Salmina)
Pnictogens (nitrogen group) (for Functional groups in 2012 Salmina)
Tetragens (carbon group) (for Functional groups in 2012 Salmina)

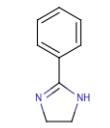
MoleculeID: *M10446*



molecule profile

Hydroxy compounds: alcohols or phenols (for Functional groups in 2005 CheckMol)
Phenols (for Functional groups in 2005 CheckMol)
Nitriles (for Functional groups in 2005 CheckMol)
Aromatic compounds (for Functional groups in 2005 CheckMol)
Arenes (for Functional groups in 2005 CheckMol)
Nonmetals (for Functional groups in 2012 Salmina)
Chalcogens (oxygen group) (for Functional groups in 2012 Salmina)
Pnictogens (nitrogen group) (for Functional groups in 2012 Salmina)
Tetragens (carbon group) (for Functional groups in 2012 Salmina)

MoleculeID: *M1364*



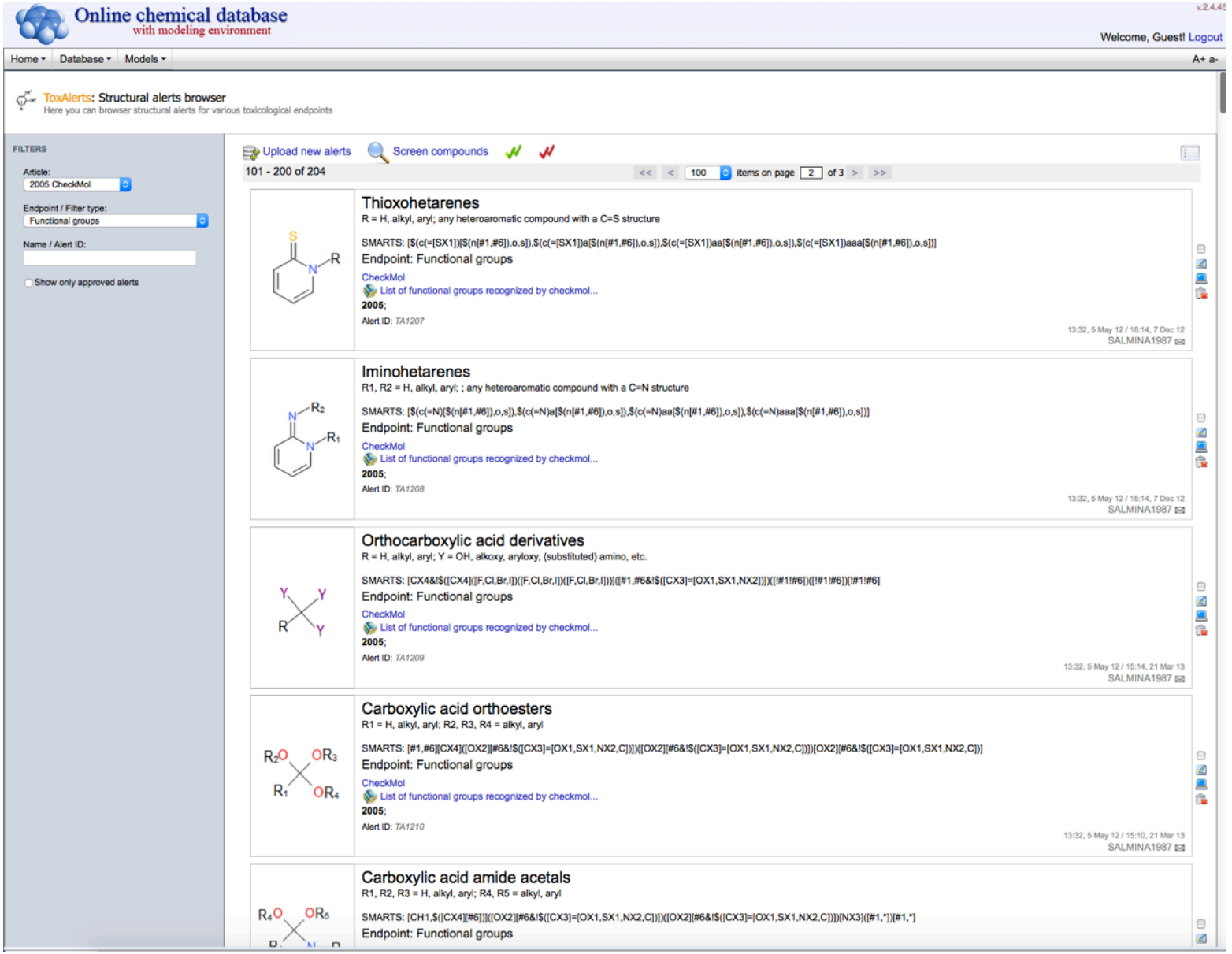
molecule profile

Carboxylic acid amidines (for Functional groups in 2005 CheckMol)
Aromatic compounds (for Functional groups in 2005 CheckMol)
Arenes (for Functional groups in 2005 CheckMol)
Heterocyclic compounds (for Functional groups in 2005 CheckMol)
Nonmetals (for Functional groups in 2012 Salmina)
Pnictogens (nitrogen group) (for Functional groups in 2012 Salmina)
Tetragens (carbon group) (for Functional groups in 2012 Salmina)
Five-membered heterocycles (LS) (for Functional groups in 2012 Salmina)
Five-membered heterocycles with two heteroatoms (LS) (for Functional groups in 2012 Salmina)
Unsaturated five-membered heterocycles with two heteroatoms (LS) (for Functional groups in 2012 Salmina)
Five-membered heterocycles (HS) (for Functional groups in 2012 Salmina)
Five-membered heterocycles with two heteroatoms (HS) (for Functional groups in 2012 Salmina)
Unsaturated five-membered heterocycles with two heteroatoms (HS) (for Functional groups in 2012 Salmina)
Imidazolines (HS) (for Functional groups in 2012 Salmina)
2-Imidazolines (HS) (for Functional groups in 2012 Salmina)

MoleculeID: *M24113*

An easy fast way to identify possibly problematic compounds
Comprehensive grouping of compounds according to a given feature

Functional groups



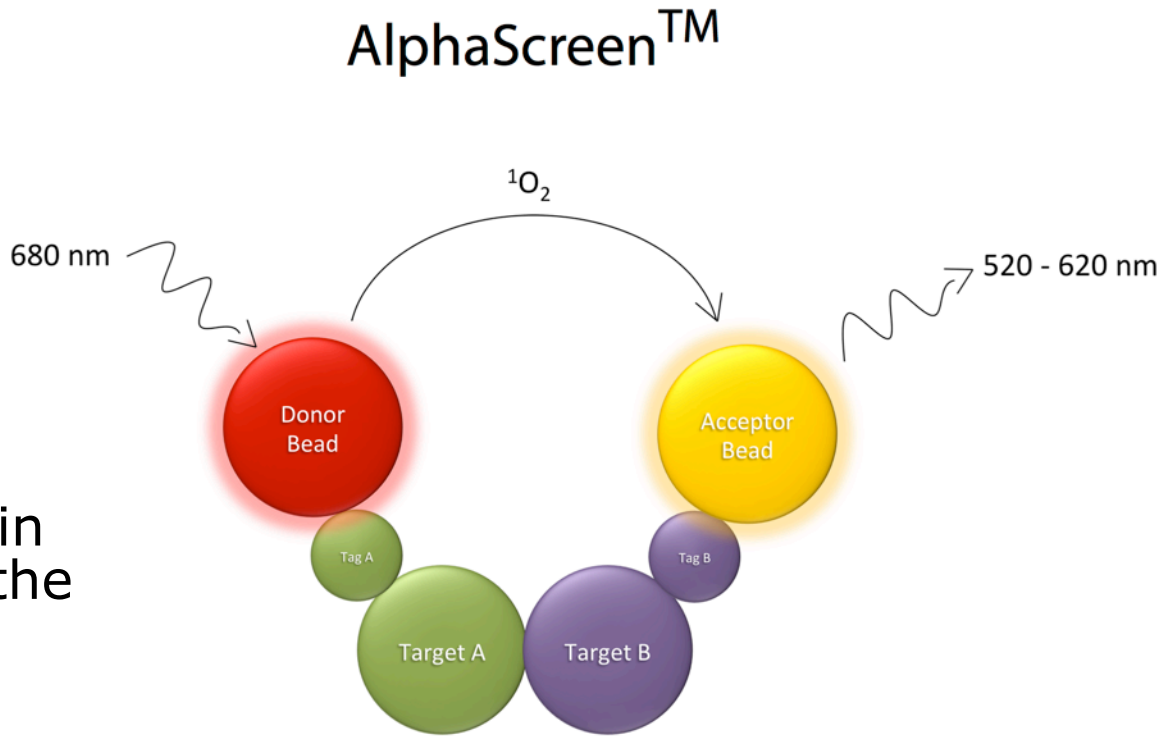
Screening Artefacts

singlet oxygen quenching

color quenching

auto-fluorescence

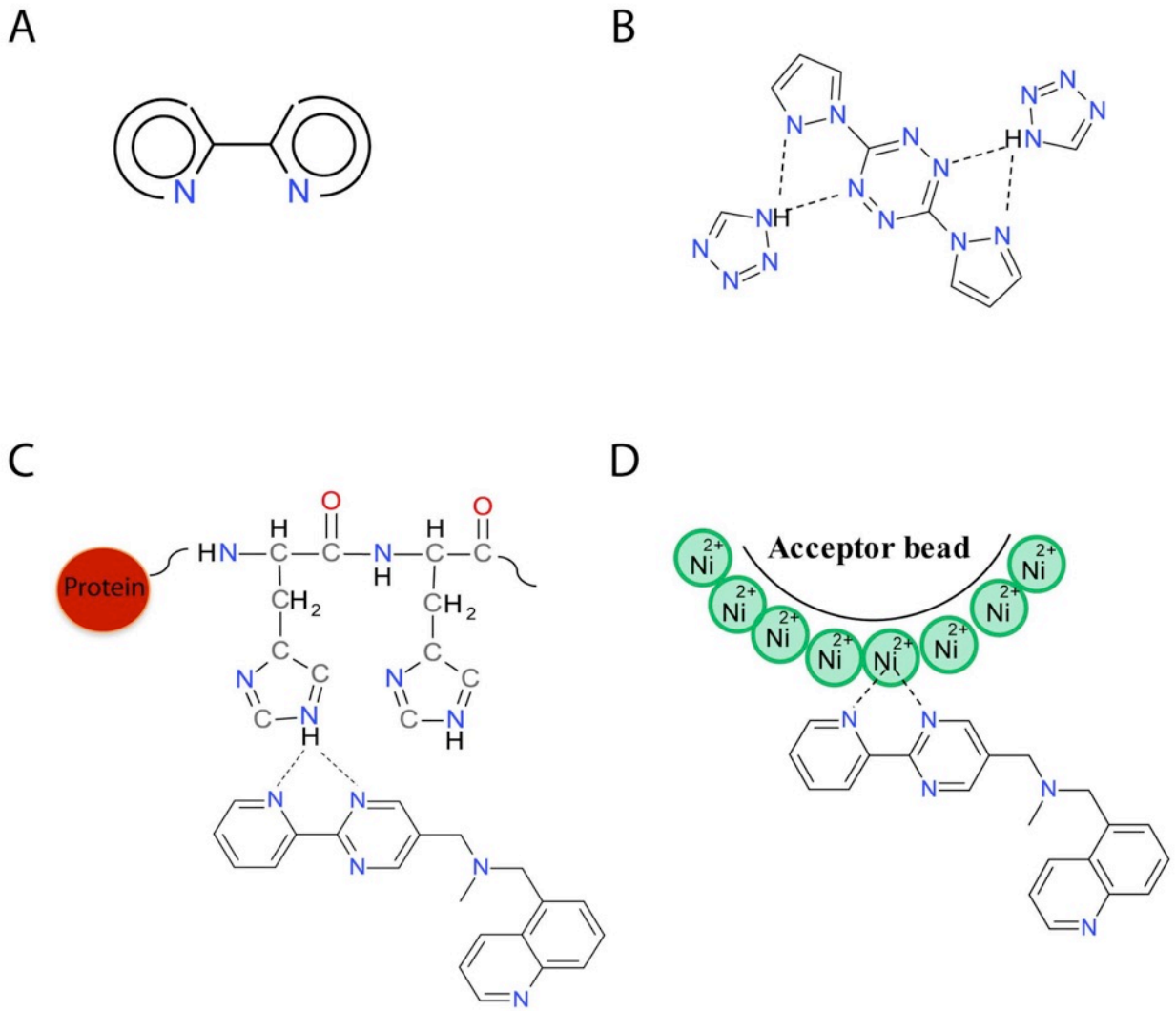
disruption of the interaction between the tag of the protein and binding site of the detection system



Pan- Assay Interference Compounds (PAINS) filters by Baell and Holloway, 2010.

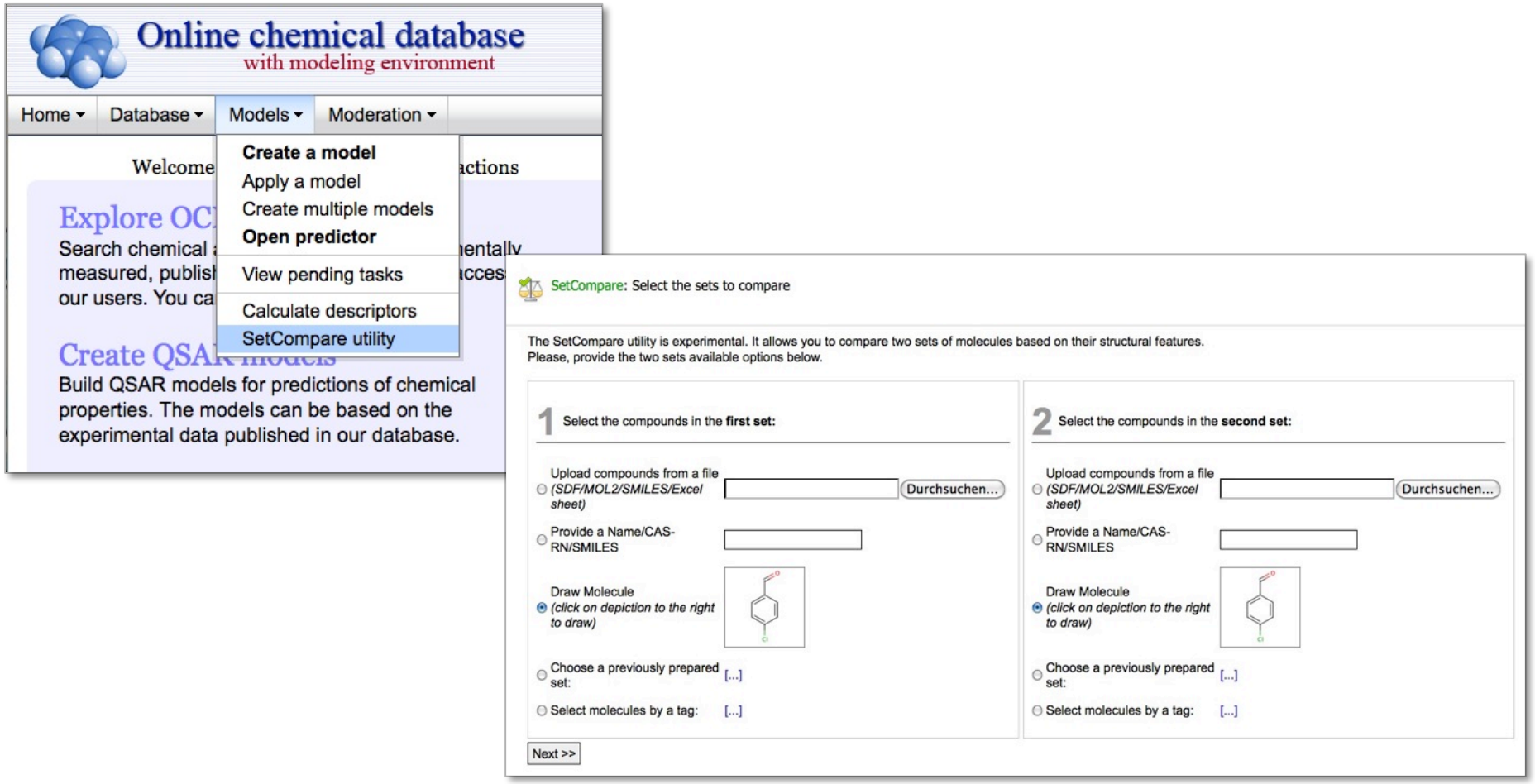
HIS tag frequent hitters, Schorpp, K. et al *J. Biomol. Screen.* 2014, 19(5):715-726.
GST tag frequent hitters, Brenke, J.K. et al *J. Biomol. Screen.* 2016, 21(6):596-607.

MOA of AlphaScreen™-HIS-FH



Schoorp et al, J Biomol Screen. 2014, 19(5):715-726.

SetCompare tool for data analysis



Examples of scaffolds analysis

SetCompare: Comparison results
The comparison summary of the two selected sets

The following table shows the features (molecular descriptors) that were significantly overrepresented in one of the tw
It includes appearance counts of the features in each set and the p-Value of such a distribution.
Export results as a CSV file

1 - 15 of 253 15 it

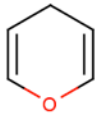
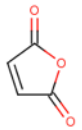
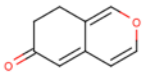
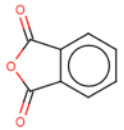
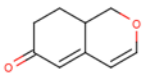
Descriptor	In set 1 (13785 molecules)	In set 2 (228174 molecules)	Enrichment factor	p-Value
	3232 (23.4%)	26196 (11.5%)	2.0	1.33E-315
Pnictogens Group 15: the nitrogen family **N P As Sb Bi**	13235 (96.0%)	202449 (88.7%)	1.1	1.42E-198
	3257 (23.6%)	79741 (34.9%)	1.5	-1.25E-172
	280 (2.0%)	352 (0.2%)	13.2	4.21E-172
	2063 (15.0%)	18761 (8.2%)	1.8	2.23E-140

Pyrolysis vs. melting point

SetCompare: Comparison results
The comparison summary of the two selected sets

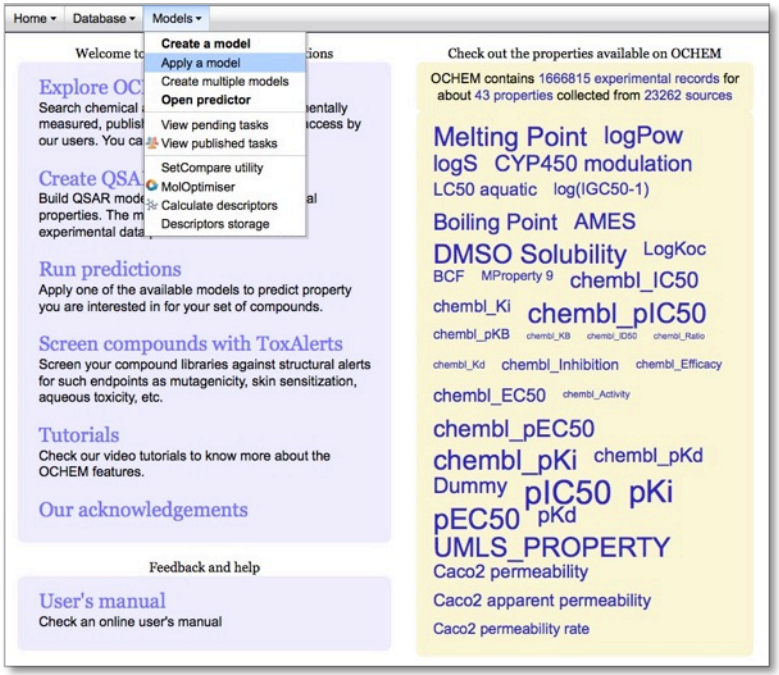
The following table shows the features (molecular descriptors) that were significantly overrepresented in on
It includes appearance counts of the features in each set and the p-Value of such a distribution.
Export results as a CSV file

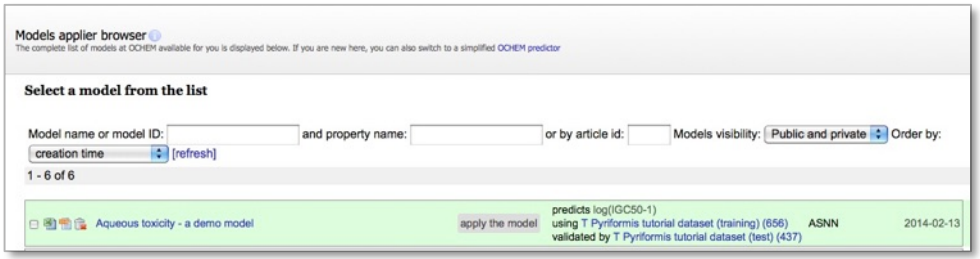
1 - 15 of 128

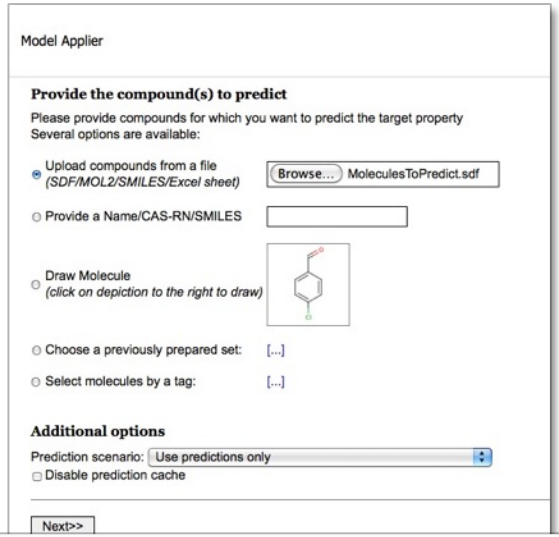
Descriptor	In set 1 (141 molecules)	In set 2 (20007 molecules)	Enrichment factor	p-Value
	11 (7.8%)	4 (0.0%)	390.2	1.77E-21
	9 (6.4%)	3 (0.0%)	425.7	6.72E-18
	7 (5.0%)	0 (0.0%)	Inf.	7.07E-16
	7 (5.0%)	2 (0.0%)	496.6	2.52E-14
	6 (4.3%)	0 (0.0%)	Inf.	1.06E-13

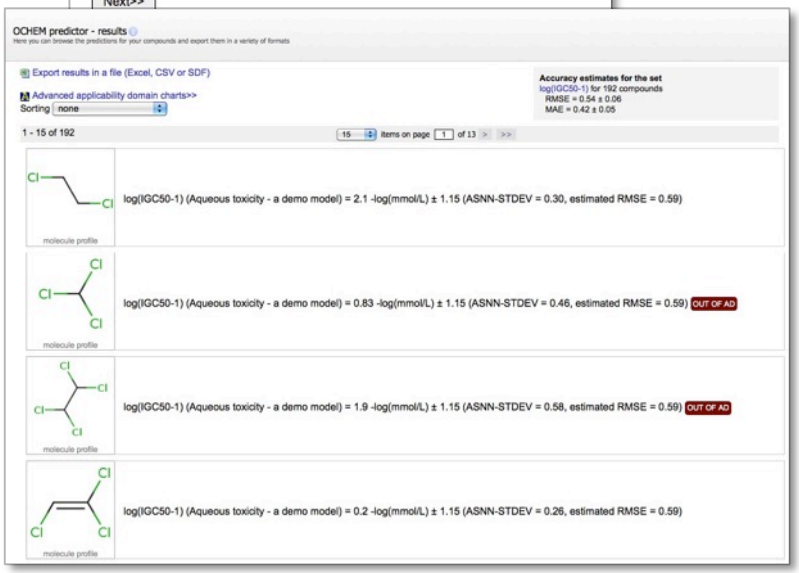
HIV Envelope glycoprotein GP120

Application of models









Modeling framework

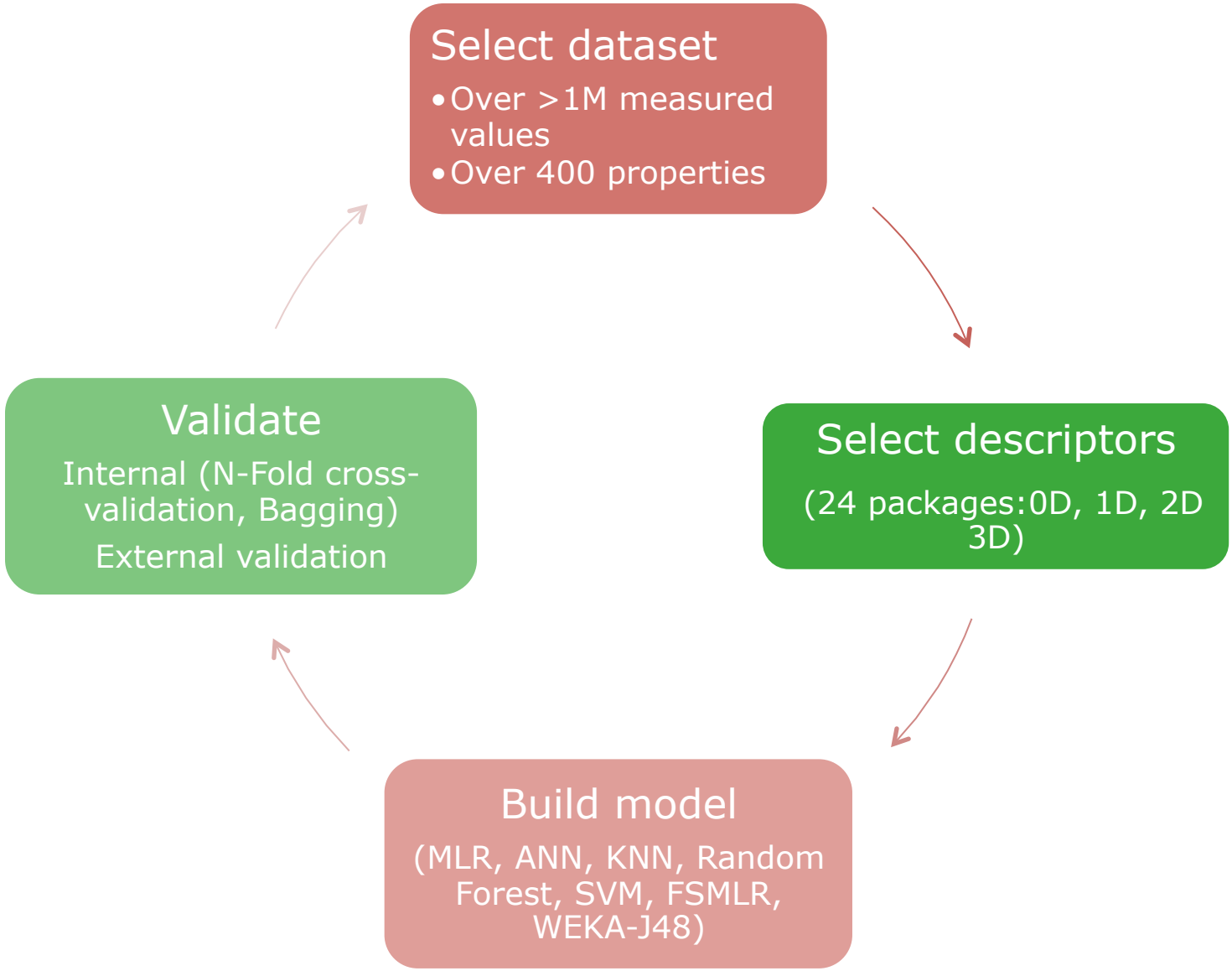
Creation of QSAR/QSPR models

- It is easy to create new models from uploaded data
- More than 10 Machine learning algorithms
  - Neural networks, KNN, SVM, MLR, PLS, random forests, J48
- More than 20 descriptor packages
  - 0D to 3D descriptors, academic and commercial
  - New descriptor packages can be easily integrated
  - Packages can be easily combined (mix & match)
  - Dragon, CDK, ADRIANA.CODE, Chemaxon, E-state, MERA, …

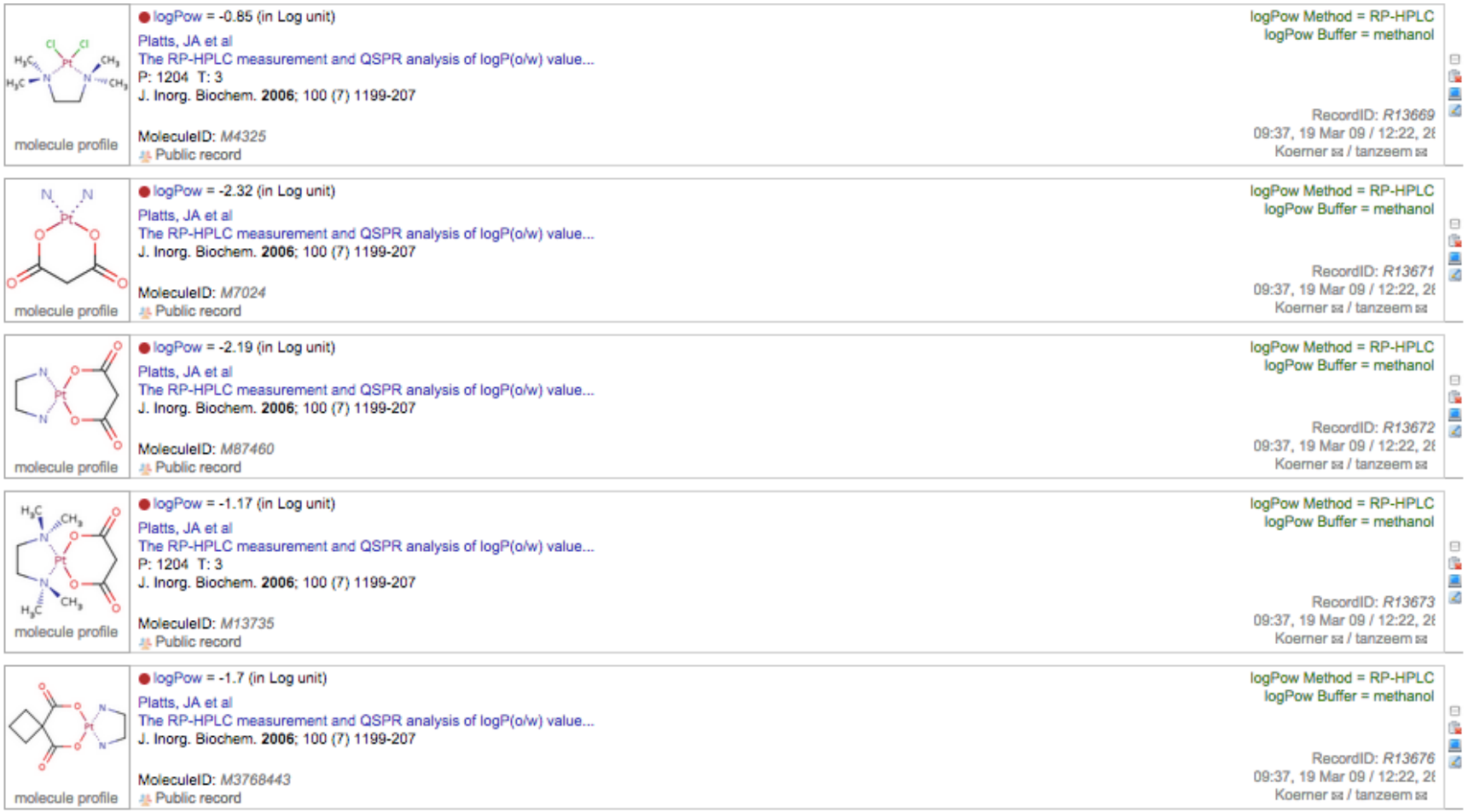
Models can be updated without complete rebuilt of model

Estimation of accuracy for each prediction

Modeling iterative workflow



LogP of Pt(II) + Pt(IV) complexes

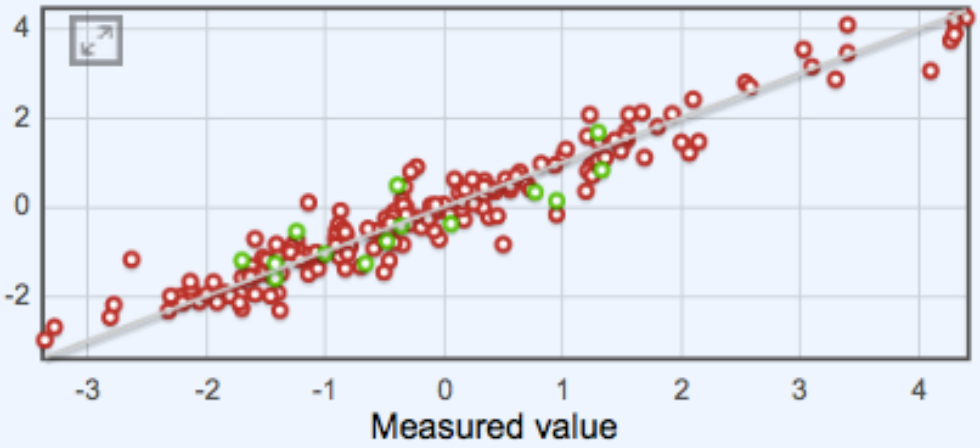


Tetko et al., J. Inorg. Biochem., 2016, 156(3), 1-13.

Model statistics

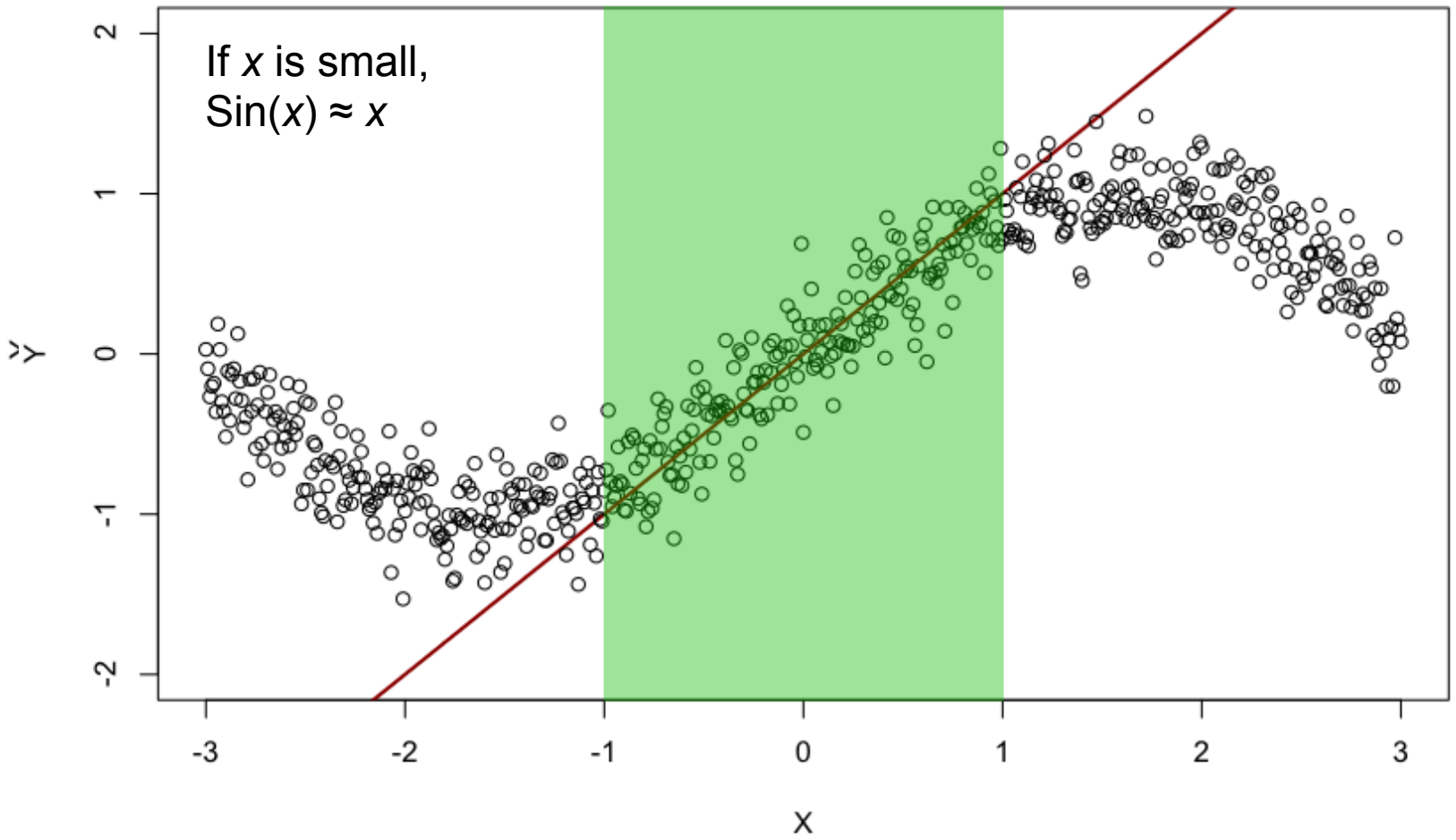
Predicted property: **logPow**
Training method: Consensus

Data Set	#	R_2	q_2	RMSE	MAE
Training set: LogPt final	187 records	0.93 ± 0.01	0.93 ± 0.01	0.41 ± 0.03	0.3 ± 0.02
Test set: Hristo [x]	14 records	0.76 ± 0.1	0.76 ± 0.1	0.5 ± 0.06	0.43 ± 0.07

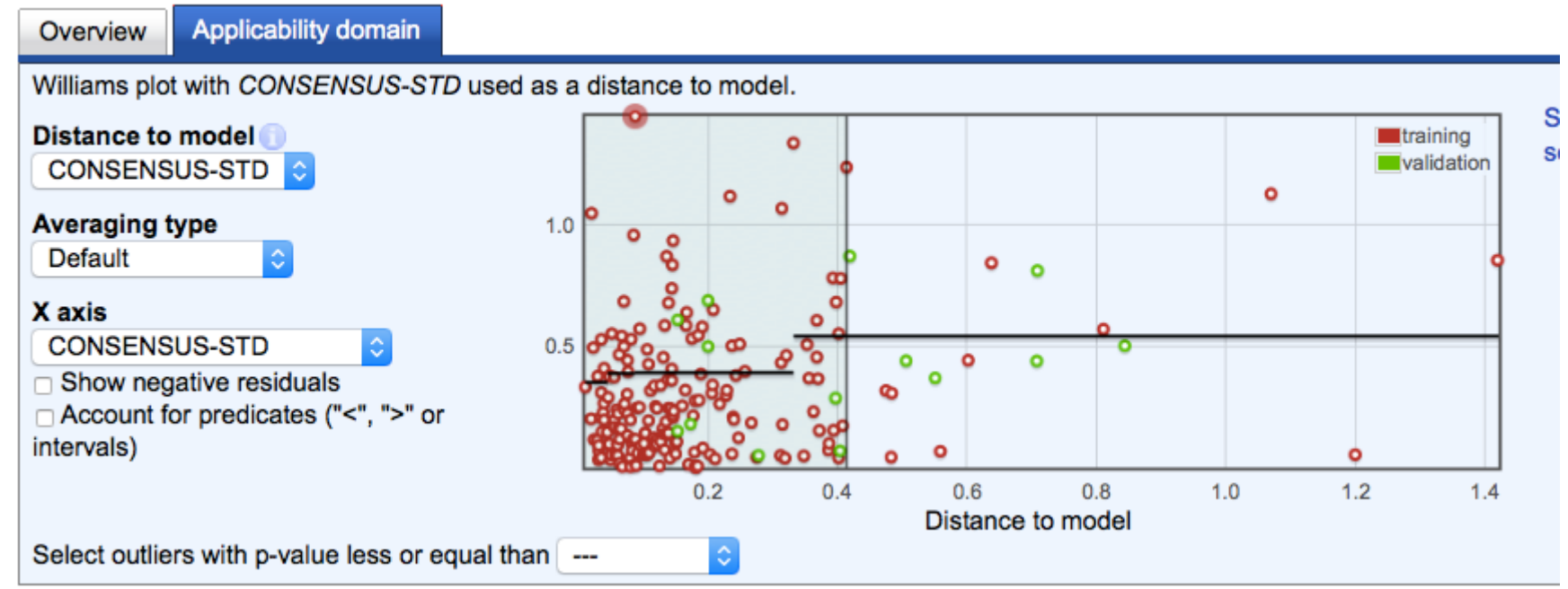


- Multiple statistical measures + confidence intervals
- Export for offline inspection
- Interactive scatter plot

Accuracy of prediction



Outlying point on the Applicability Domain plot



- Estimation of applicability domain of models
- Identification of outliers

Accuracy of predictions for classification model

Overview | Applicability domain

Model name: Ames levenberg , 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: 450 Kb

Data Set	#	Accuracy	Balanced accuracy
Training set: Ames challenge training	4357 records (4359 selected)	78.1 ± 1.2	77.9 ± 1.3
Test set: Ames challenge test [x]	2181 records	79.9 ± 1.7	79.8 ± 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)		

Overview | Applicability domain

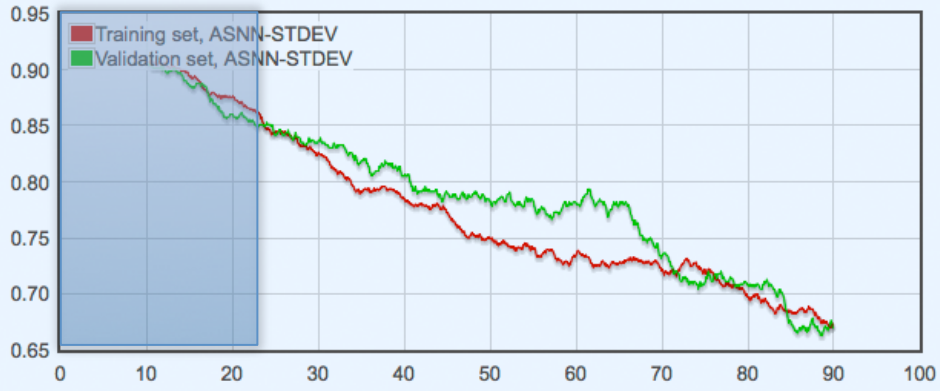
Wilkinson plot with *ASNN-STDEV* used as a distance to model.

Distance to model
ASNN-STDEV

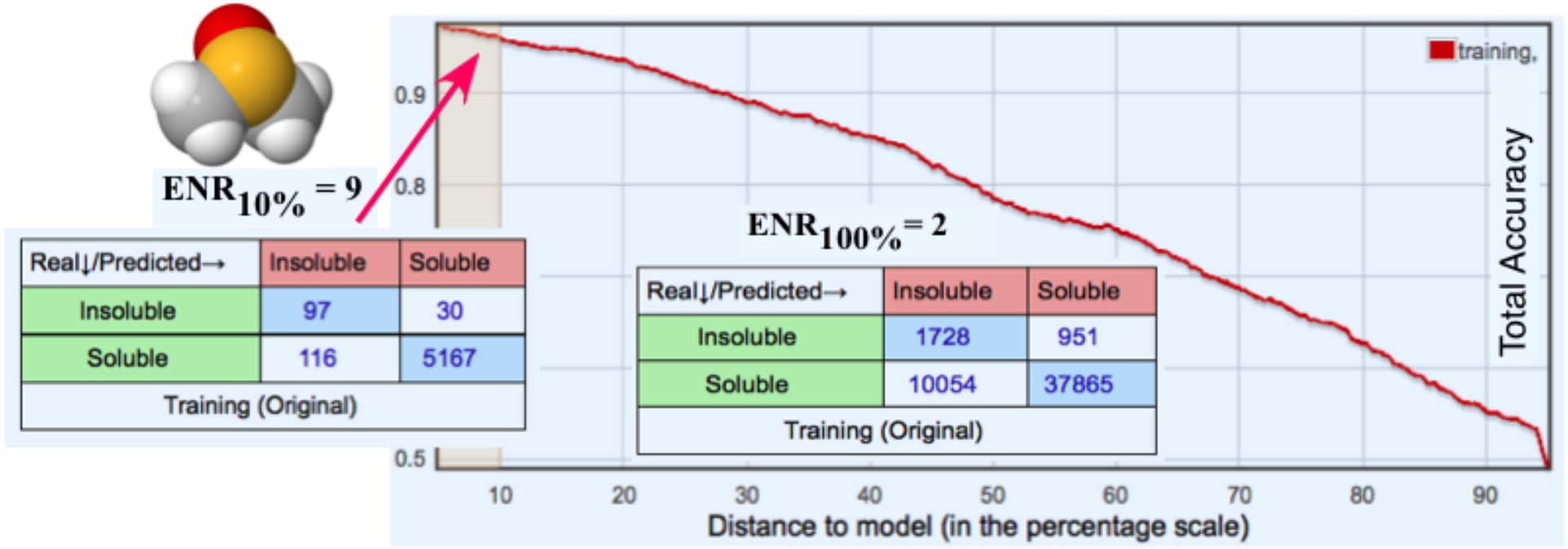
Averaging type
Default

Window size:
20% of the set

X axis
Percentage of compounds



Solubility in DMSO classification



Example of a published solubility in DMSO model

- ✓ based on >163k compounds (provided by UCB and Enamine)
- ✓ nine folds decrease number of non-soluble in DMSO compounds

Tetko et al *J. Chem. Inf. Model*. 2013, 53(8):1990-2000.

300k Melting Point Datasets

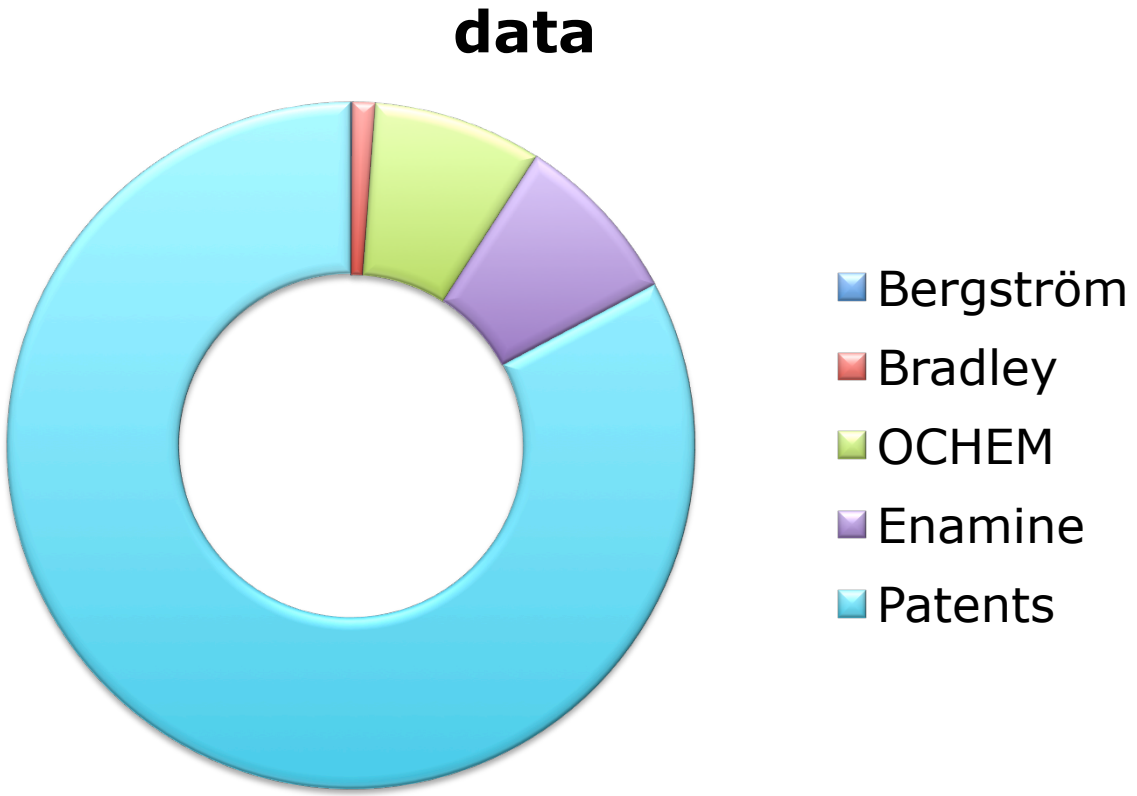
Bergström 277

Bradley 2886

OCHEM 22404

Enamine 21883

Patents 228079



Tetko et al *J. Chemoinformatics, J. Cheminform. 2016, 8, 2.*

Comprehensive modeling

Package name	Type of descriptors	Number of descriptors	Matrix size, billions	Non zero values, millions	Sparseness
Functional Groups	integer	595	0.18	3.1	33
QNPR	integer	1502	0.45	6.3	49
MolPrint	binary	688634	205	8.1	7200
Estate count	float	631	0.19	10	14
Inductive	float	54	0.02	11	1
ECFP4	binary	1024	0.31	12	25
Isida	integer	5886	1.75	18	37
ChemAxon	float	498	0.15	23	1.5
GSFrag	integer	1138	0.34	24	5.7
CDK	float	239	0.07	27	2
Adriana	float	200	0.06	32	1.3
Mera, Mersy	float	571	0.17	61	1.1
Dragon	float	1647	0.49	183	1.5

Comprehensive modeling

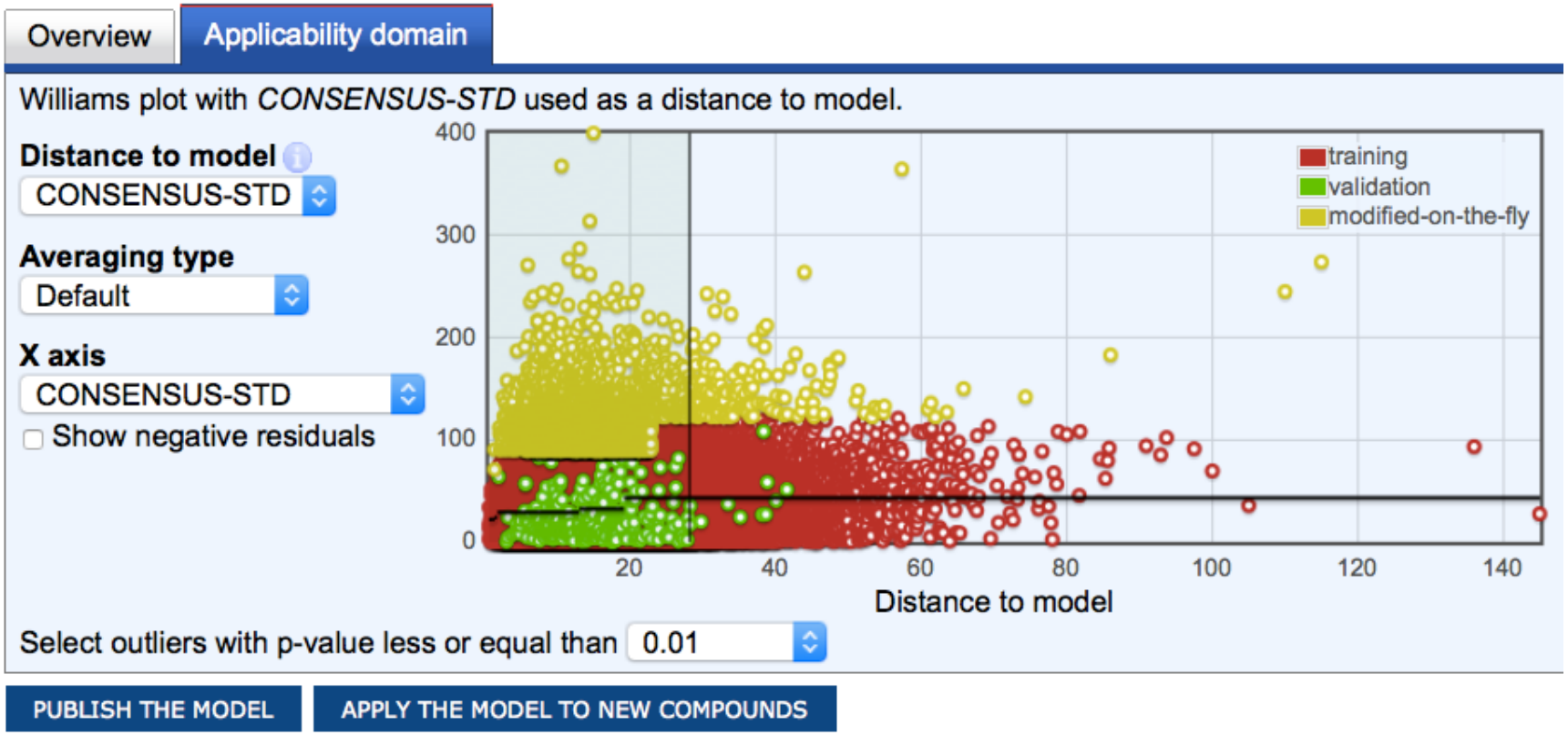
Predicted property: Melting Point
Training set: patents - decomposition (4 different versions detected)

Metrics RMSE - Root Mean Square Error for Training set Validation: All validation protocols

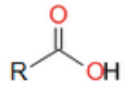
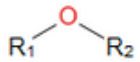
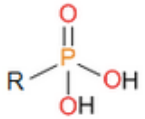
	LibSVM
GSFrag	41.85
ChemaxonDescriptors (7.4)	39.63
InductiveDescriptors	49.54
MolPrint (Length 2)	61.2
StructuralAlerts	41.92
ECFP (4_1024)	61.46
OEstate	37.8
Fragmentor (Length 2 - 4)	38.54
QNPR (SMILES - length 1 - 3 threshold 5)	39.75
CDK (constitutional, topological, geometrical, electronic, hybrid)	38.45
Mera, Mersy	42.8
Dragon (blocks: 1-20)	43.04
Adriana	41.91
SIRMS (LABELING=ELM noH type=EXTENDED)	40.36
MolPrint (Length 3)	60.89

	Consensus
Misc.	36.01

Outliers identified with applicability domain (AD) plot



Functional group analysis of pyrolysis and MP data

Descriptor	In set 1 (13785 molecules)	In set 2 (228174 molecules)	Enrichment factor	p-Value
	3232 (23.4%)	26196 (11.5%)	2.0	1.33E-315
Pnictogens Group 15: the nitrogen family **N P As** **Sb Bi**	13235 (96.0%)	202449 (88.7%)	1.1	1.42E-198
	3257 (23.6%)	79741 (34.9%)	1.5	-1.25E-172
	280 (2.0%)	352 (0.2%)	13.2	4.21E-172

Outliers have higher percentage of predicted decomposing compounds

20% of ~5k outliers are predicted as pyrolysis

vs.

17% of ~223k PATENTS compounds

Why do we need the Melting Point (MP)?

Use of MP for water solubility prediction

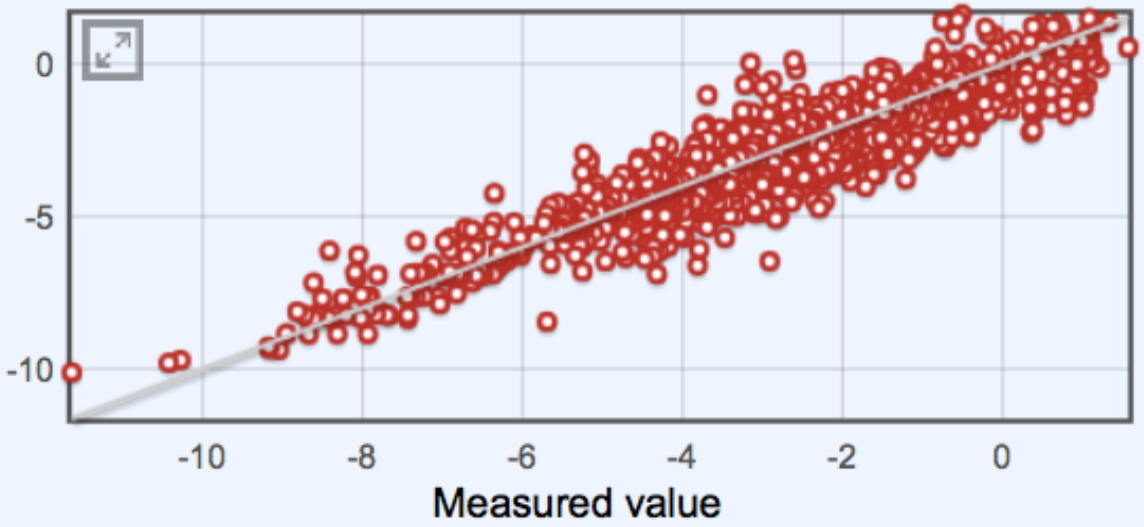
Yalkowsky equation:

$$\log S = 0.5 - 0.01(MP-25) - \log Kow$$

Prediction of Huuskonen set using ALOGPS logP and MP based on 230k measurements

Predicted property: **Aqueous Solubility** modeled in log(mol/L)
Training method: MLRA

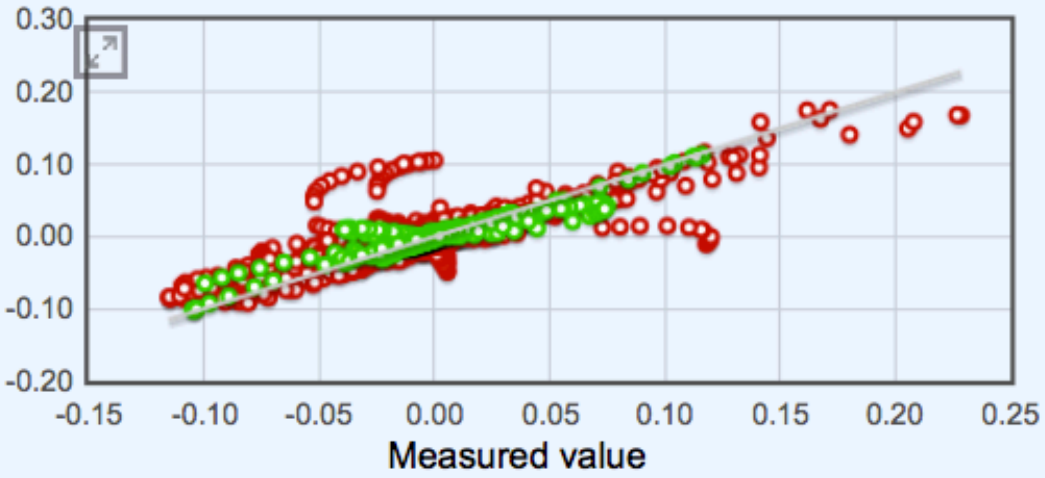
Data Set	#	R_2	q_2	RMSE	MAE
Training set: logS set	1311 records	0.842 ± 0.009	0.83 ± 0.01	0.84 ± 0.02	0.64 ± 0.02



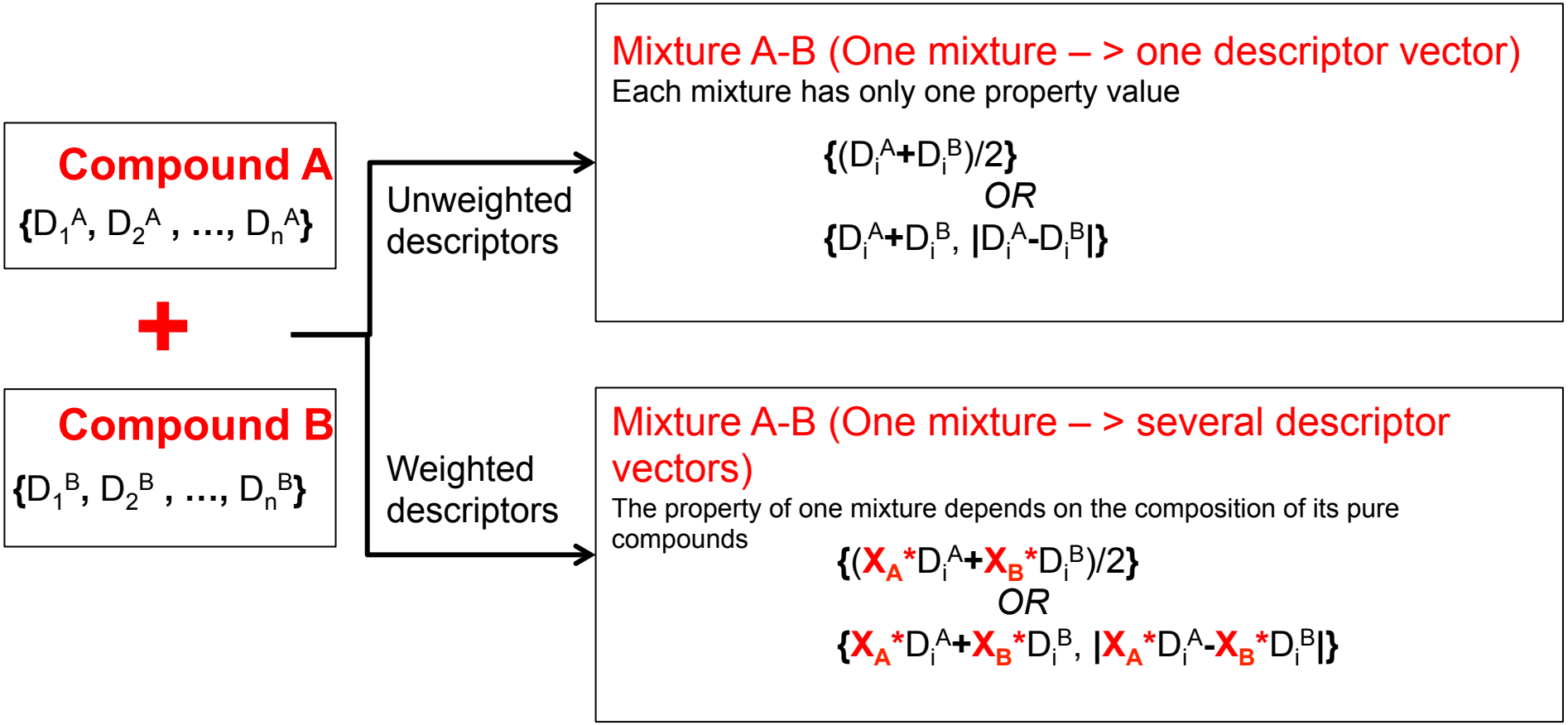
Best models based on this set have RMSE ~ 0.6

Modeling of mixtures

Data Set	#	R_2	q_2	RMSE	MAE
Training set: Ajmani_delta_train	3857 records	0.69 ± 0.06	0.69 ± 0.06	0.014 ± 0.001	0.0064 ± 0.0004
Test set: Ajmani_delta_test [x]	672 records	0.88 ± 0.04	0.87 ± 0.04	0.0089 ± 0.001	0.0041 ± 0.0006



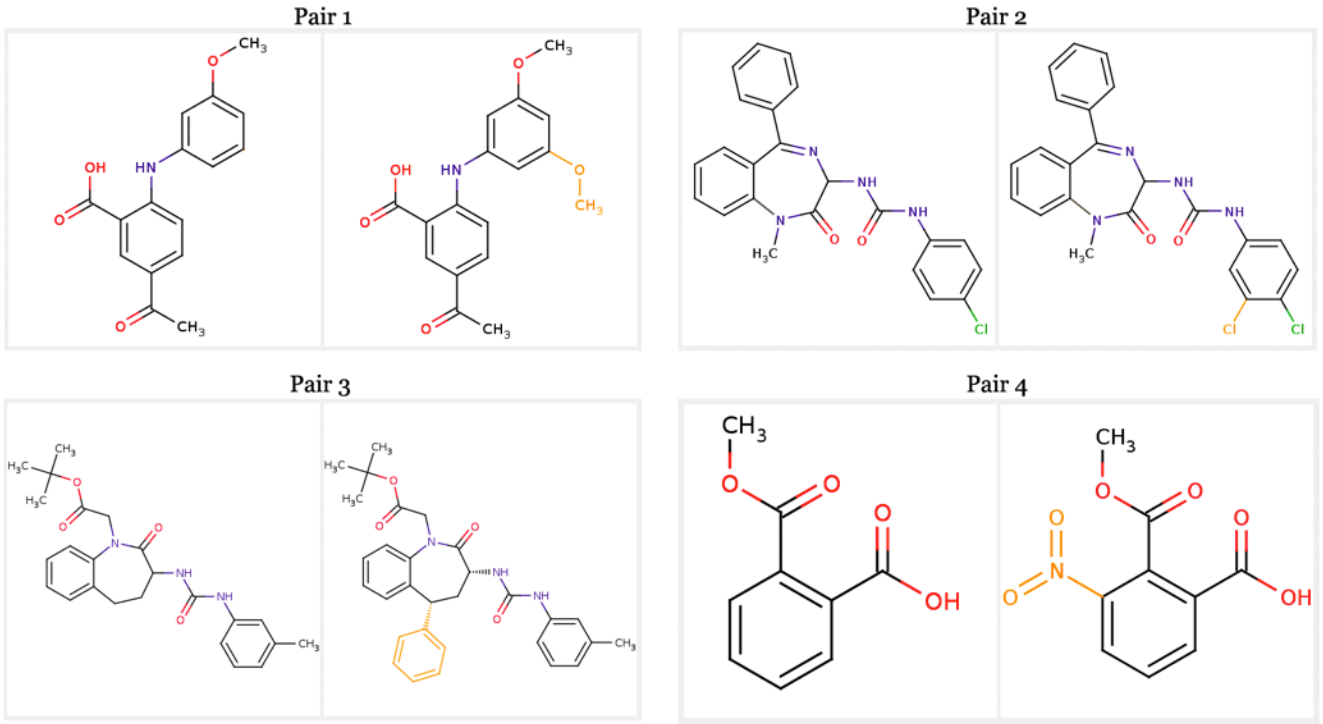
Mixtures' descriptors



Interpretation of models

Molecular Matched Pairs

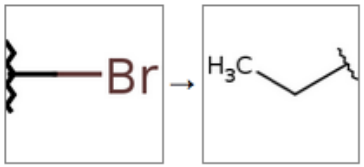
A molecular matched pair (MMP) is a pair of molecules that have only a (minor) single-point difference.
The typical way is to define a minor difference as a changed molecular fragment with less than 10 atoms.



MMPs for classification data (AMES mutagenicity)

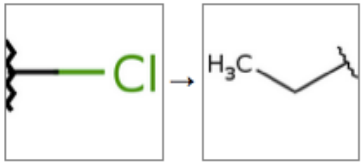
Transformations: minimum 10 pairs, p-value 0.01

1 - 5 of 32 | 5 items on page | 1 of 7 > >>



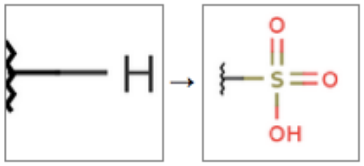
12 matched pairs
↑ 0 pairs, ↓ 7 pairs
2 negative and 3 positive molecules unaffected

SMIRKS: *Br -> CC*



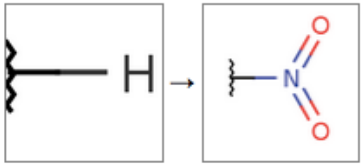
21 matched pairs
↑ 0 pairs, ↓ 9 pairs
8 negative and 4 positive molecules unaffected

SMIRKS: *Cl -> CC*



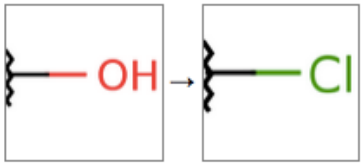
17 matched pairs
↑ 1 pairs, ↓ 8 pairs
5 negative and 3 positive molecules unaffected

SMIRKS: *[H] -> OS(*)(=O)=O



204 matched pairs
↑ 57 pairs, ↓ 8 pairs
26 negative and 113 positive molecules unaffected

SMIRKS: *[H] -> *N(=O)=O



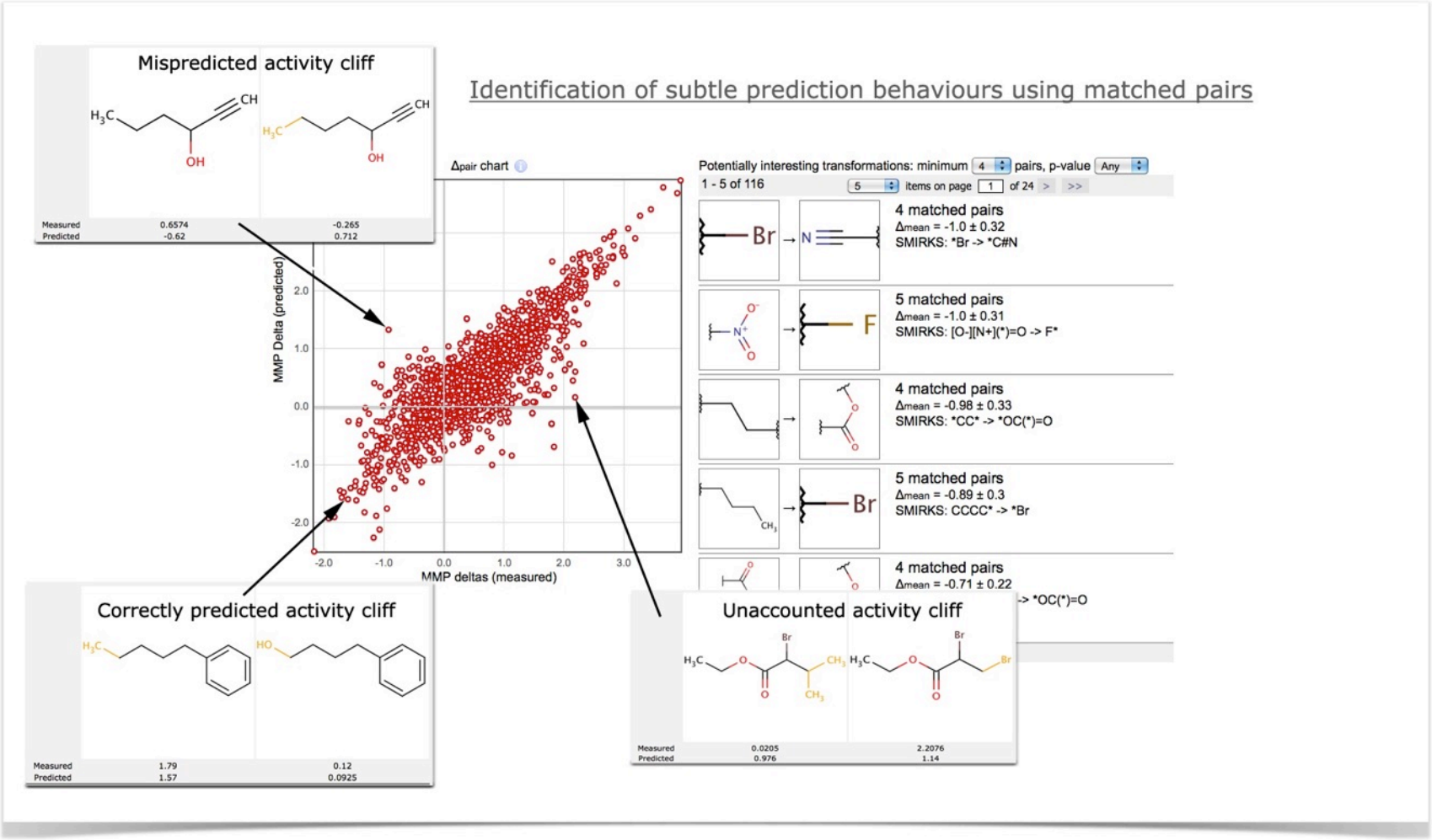
69 matched pairs
↑ 21 pairs, ↓ 4 pairs
25 negative and 19 positive molecules unaffected

SMIRKS: O* -> *Cl

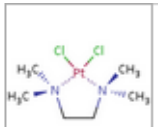
1 - 5 of 32

Resubmit all molecules | Show fragment graph | Save transformations | Export pairs

Analysis of rules that were learnt by models



LogP of Pt(II) + Pt(IV) complexes



molecule profile

● logPow = -0.85 (in Log unit)
Platts, JA et al
The RP-HPLC measurement and QSPR analysis of logP(o/w) value...
P: 1204 T: 3
J. Inorg. Biochem. **2006**; 100 (7) 1199-207

MoleculeID: *M4325*
Public record

logPow Method = RP-HPLC
logPow Buffer = methanol

RecordID: *R13669*
09:37, 19 Mar 09 / 12:22, 2
Koerner / tanzeem

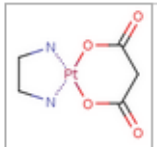
molecule profile

● logPow = -2.32 (in Log unit)
Platts, JA et al
The RP-HPLC measurement and QSPR analysis of logP(o/w) value...
J. Inorg. Biochem. **2006**; 100 (7) 1199-207

MoleculeID: *M7024*
Public record

logPow Method = RP-HPLC
logPow Buffer = methanol

RecordID: *R13671*
09:37, 19 Mar 09 / 12:22, 2
Koerner / tanzeem



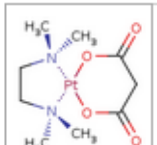
molecule profile

● logPow = -2.19 (in Log unit)
Platts, JA et al
The RP-HPLC measurement and QSPR analysis of logP(o/w) value...
J. Inorg. Biochem. **2006**; 100 (7) 1199-207

MoleculeID: *M87460*
Public record

logPow Method = RP-HPLC
logPow Buffer = methanol

RecordID: *R13672*
09:37, 19 Mar 09 / 12:22, 2
Koerner / tanzeem



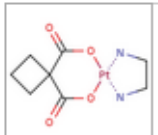
molecule profile

● logPow = -1.17 (in Log unit)
Platts, JA et al
The RP-HPLC measurement and QSPR analysis of logP(o/w) value...
P: 1204 T: 3
J. Inorg. Biochem. **2006**; 100 (7) 1199-207

MoleculeID: *M13735*
Public record

logPow Method = RP-HPLC
logPow Buffer = methanol

RecordID: *R13673*
09:37, 19 Mar 09 / 12:22, 2
Koerner / tanzeem



molecule profile

● logPow = -1.7 (in Log unit)
Platts, JA et al
The RP-HPLC measurement and QSPR analysis of logP(o/w) value...
J. Inorg. Biochem. **2006**; 100 (7) 1199-207

MoleculeID: *M3768443*
Public record

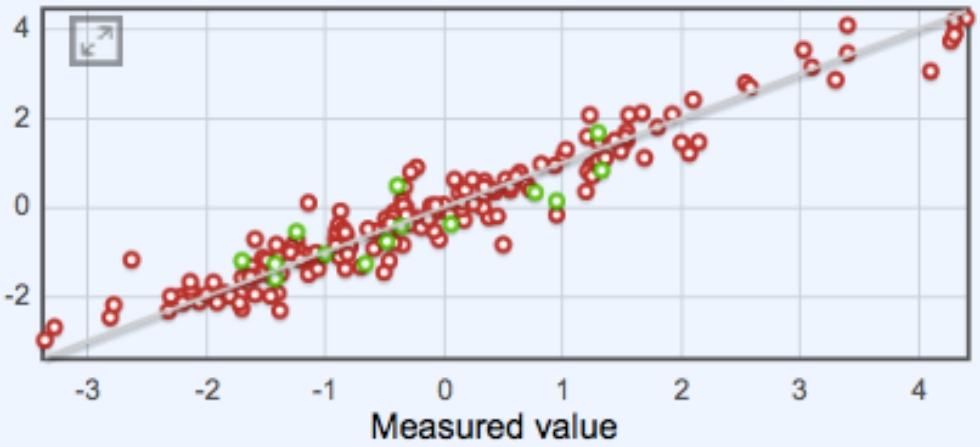
logPow Method = RP-HPLC
logPow Buffer = methanol

RecordID: *R13676*
09:37, 19 Mar 09 / 12:22, 2
Koerner / tanzeem

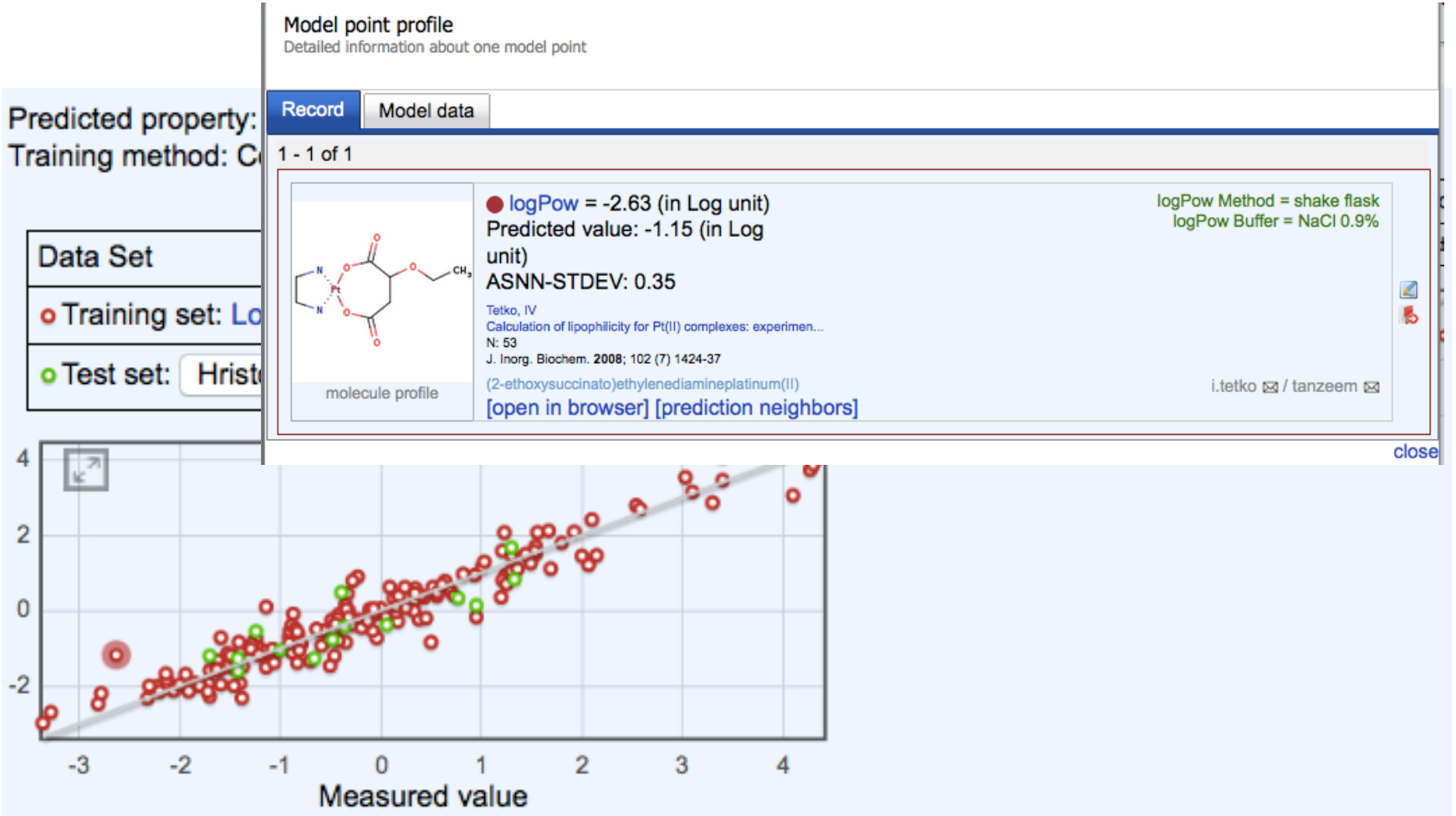
Model for LogP for Pt complexes

Predicted property: **logPow**
Training method: Consensus

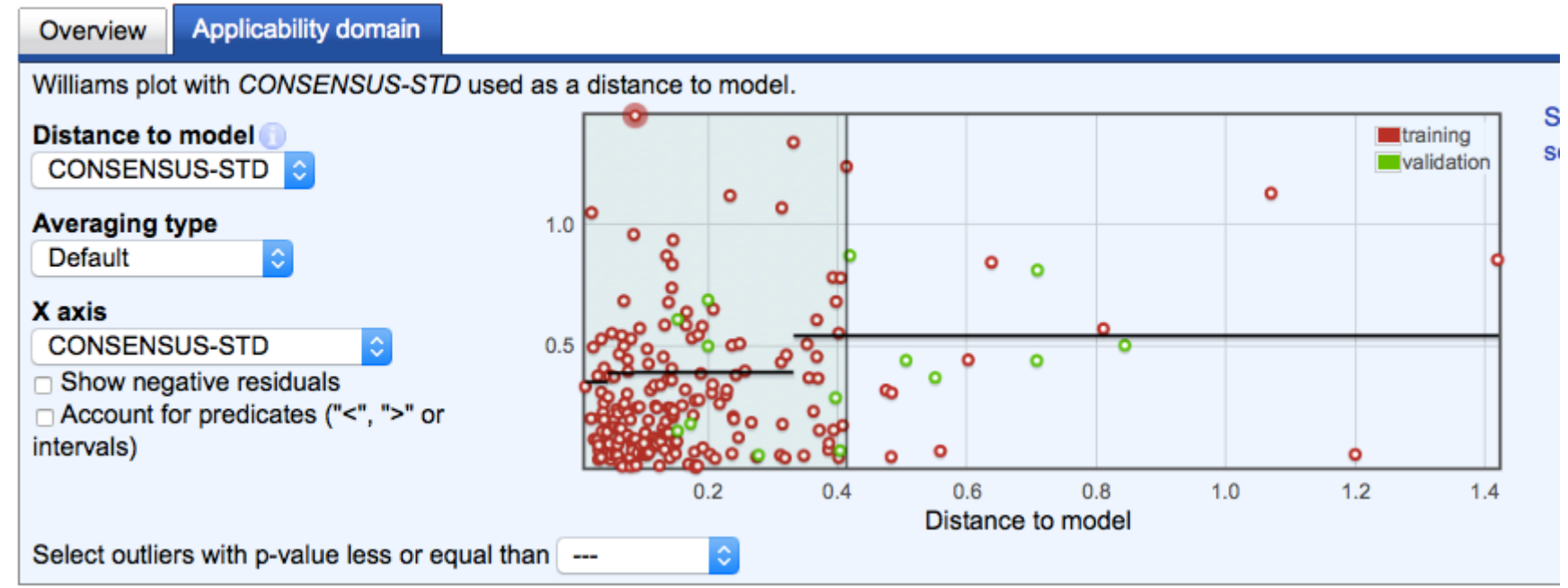
Data Set	#	R^2	q^2	RMSE	MAE
Training set: LogPt final	187 records	0.93 ± 0.01	0.93 ± 0.01	0.41 ± 0.03	0.3 ± 0.02
Test set: Hristo [x]	14 records	0.76 ± 0.1	0.76 ± 0.1	0.5 ± 0.06	0.43 ± 0.07



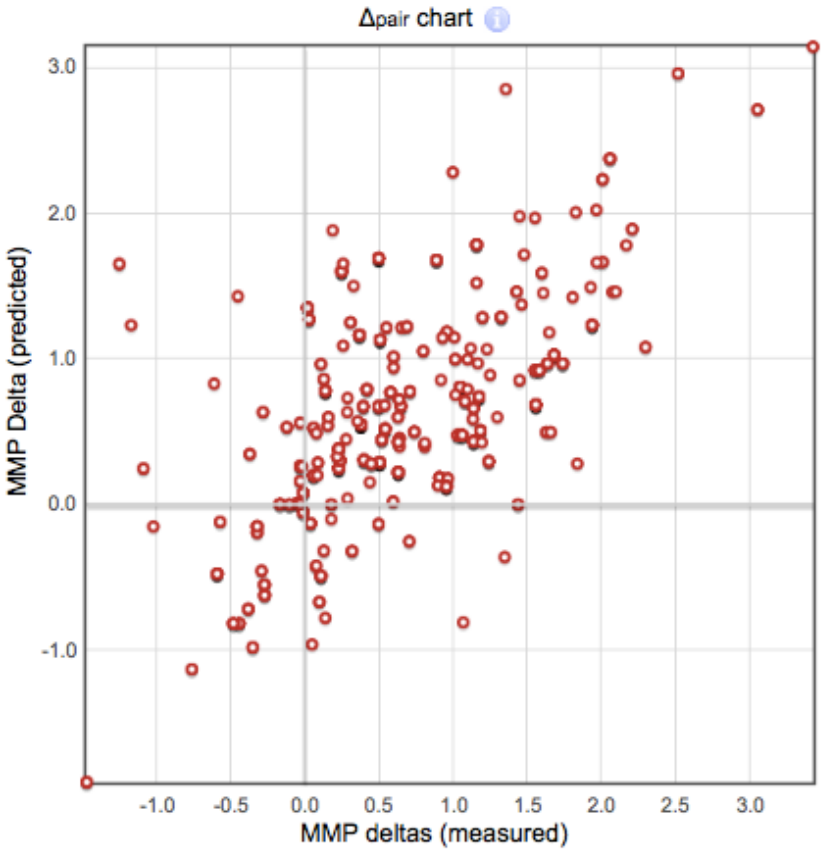
Model for LogP for Pt complexes



Outlying point on the Applicability Domain plot

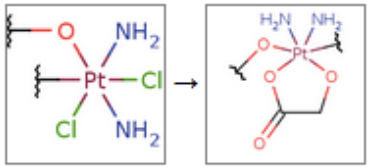


Prediction of logP for Pt complexes

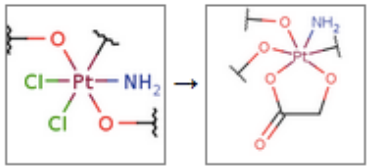


Transformations: minimum 4 pairs, p-value Any

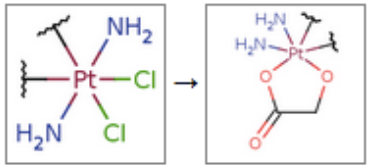
1 - 5 of 60 5 items on page 1 of 12 > >>



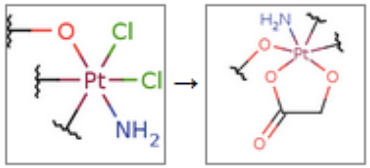
6 matched pairs
Δmean = -0.27 ± 0.19
SMIRKS: N[Pt](N)(*)(Cl)(Cl)O* -> N[Pt]1(N)(*)(O*)OCC(=O)O1



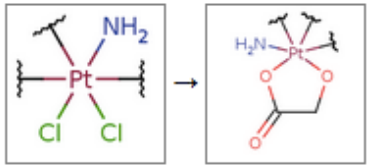
6 matched pairs
Δmean = -0.27 ± 0.19
SMIRKS: N[Pt](*)(Cl)(Cl)(O*)O* -> N[Pt]1(*)(O*)(O*)OCC(=O)O1



7 matched pairs
Δmean = -0.24 ± 0.2
SMIRKS: N[Pt](N)(*)(*)(Cl)Cl -> N[Pt]1(N)(*)(*)OCC(=O)O1

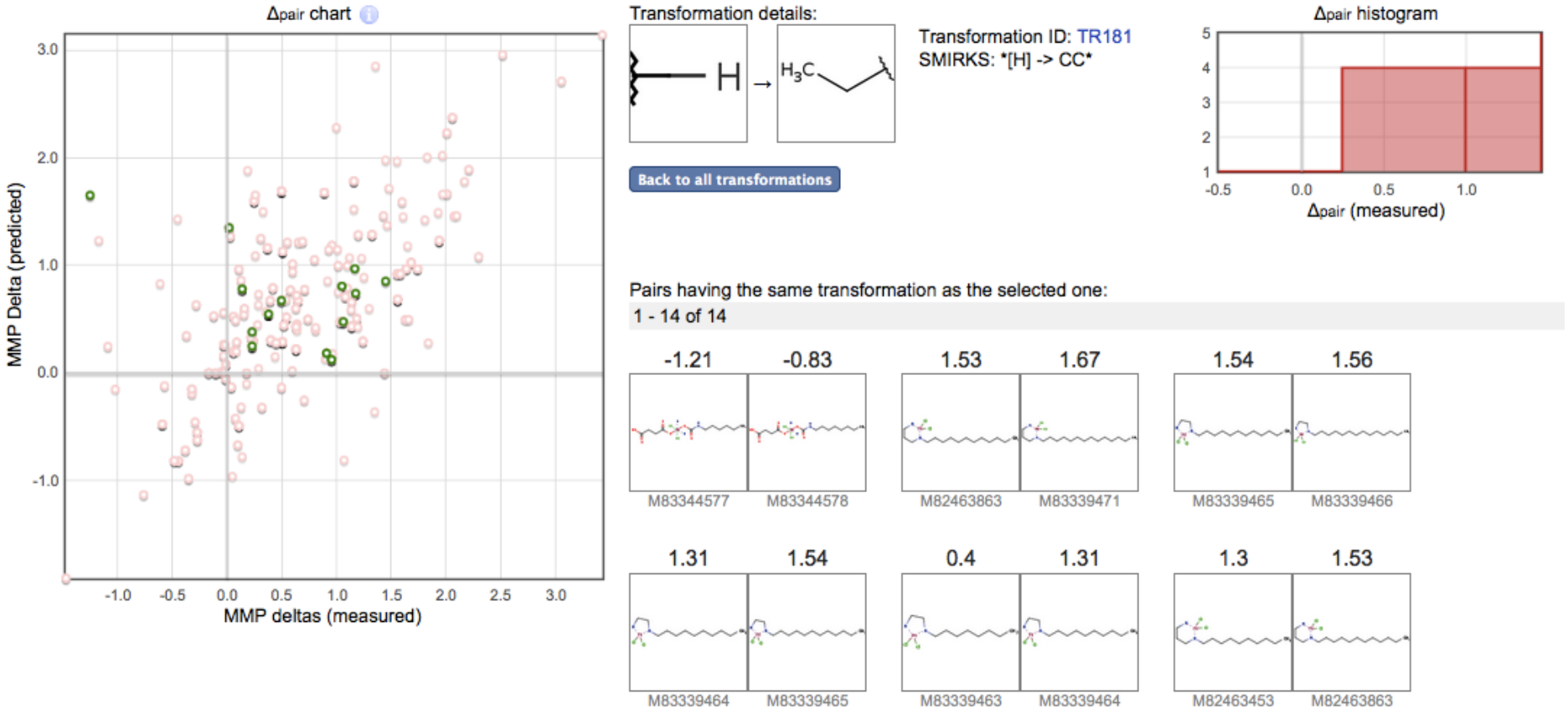


7 matched pairs
Δmean = -0.24 ± 0.2
SMIRKS: N[Pt](*)(*)(Cl)(Cl)O* -> N[Pt]1(*)(*)(O*)OCC(=O)O1

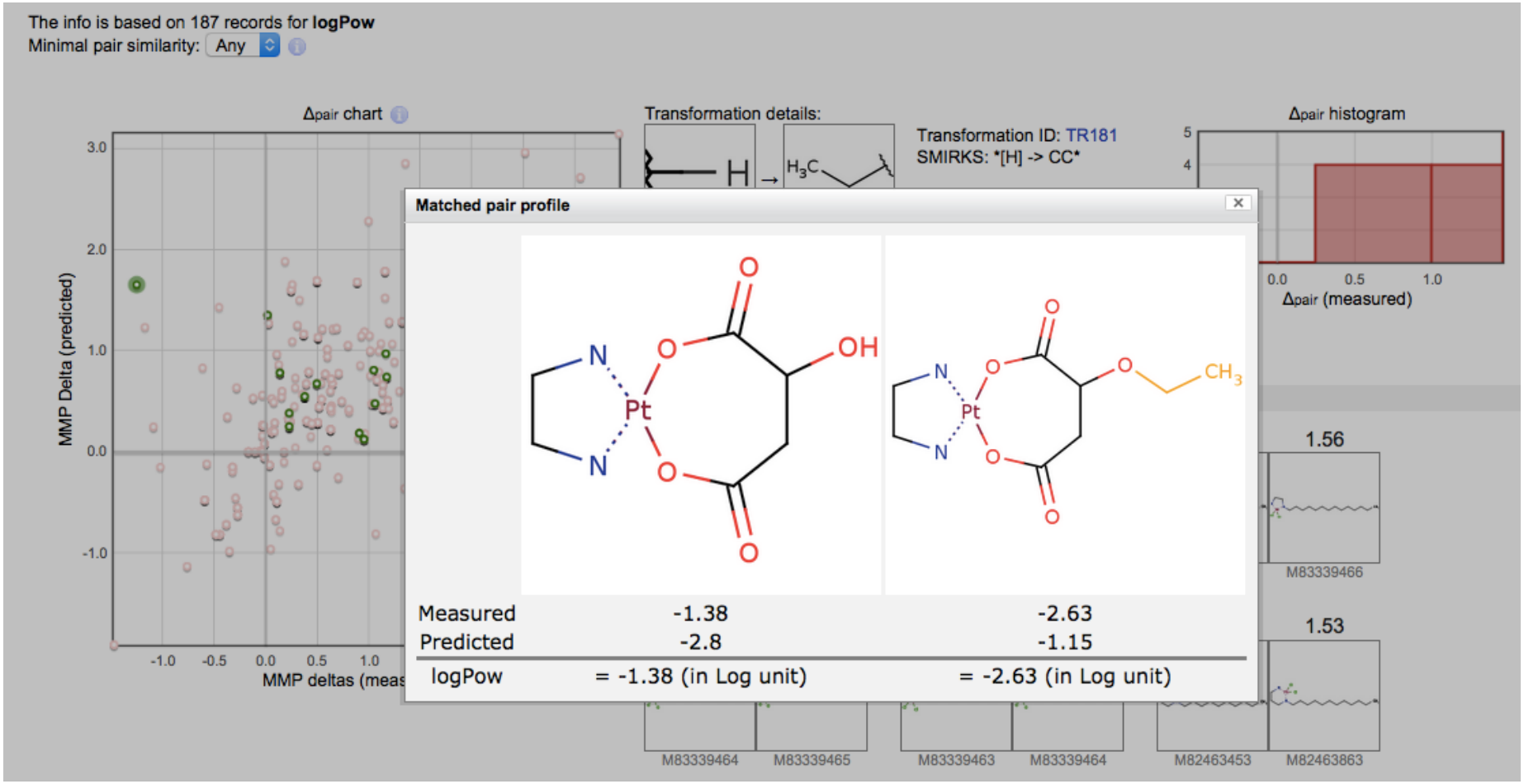


8 matched pairs
Δmean = -0.17 ± 0.26
SMIRKS: N[Pt](*)(*)(*)(Cl)Cl -> N[Pt]1(*)(*)(*)OCC(=O)O1

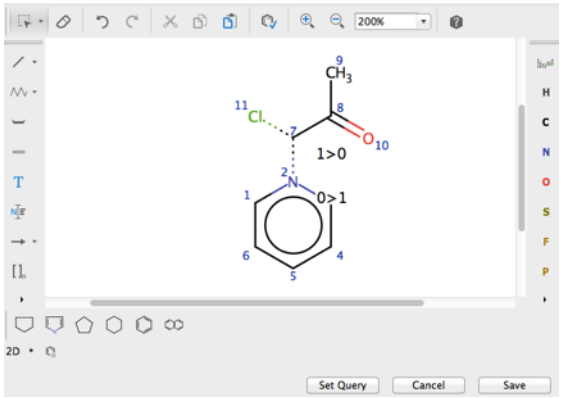
Outliers of the model (functional groups descriptors)

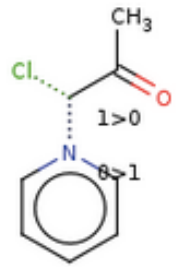


Outliers of the model (Functional groups descriptors)



Storage of Chemical Reaction as Condensed Graph of Reaction (CRG)





molecule profile

● SN2 reaction

rate constant =

-4.72 (in log(L*mol^(-1)*s^(-1)))

Nugmanov, R. et al

Development of "structure-property" models in nucleophilic s...

N: 8 P: 270

Journal of Structural Chemistry **2014**; 55 (6) 1026-1032

MoleculeID: *M84326300*

Private record

Temperature= 55.6 °C

Solvent = Ethanol

RecordID: *R19277000*

23:25, 4 Jul 15 / 20:32, 6 Jul 15

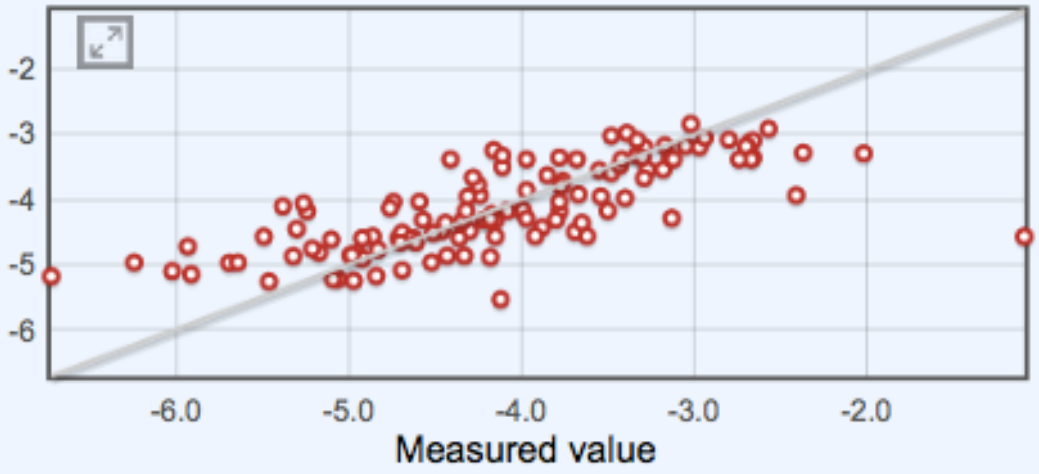
itetko

Only visible to itetko

Modeling of SN2 reactions using OCHEM

Predicted property: **SN2 reaction rate constant** modeled in log(L*mol^(-1)*s^(-1))
Training method: Consensus

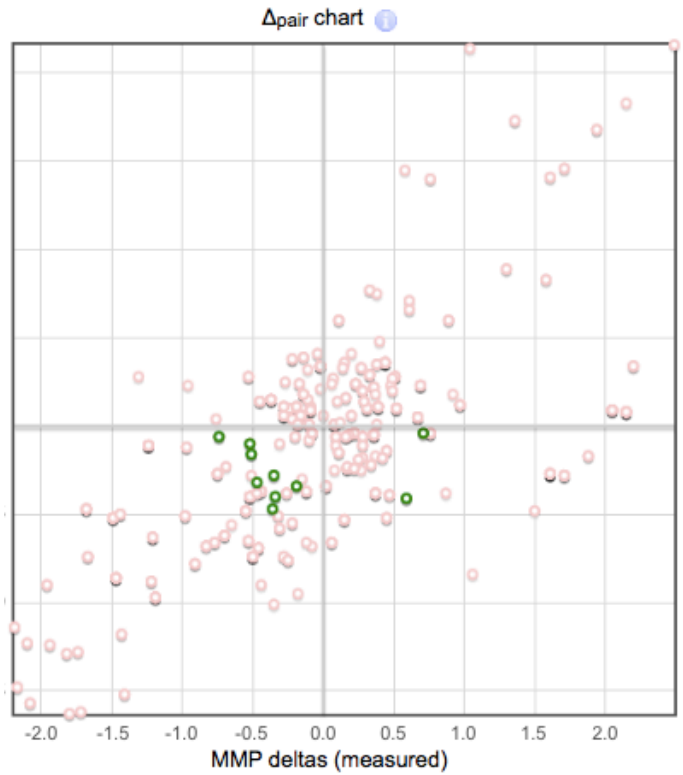
Data Set	#	R_2	q_2	RMSE	MAE
Training set: SN2 methanol 20-40	125 records	0.54 ± 0.08	0.54 ± 0.08	0.64 ± 0.08	0.44 ± 0.04



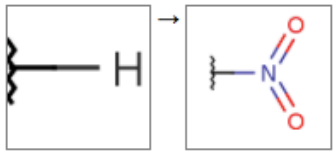
Consistent data: t = 20-40, solvent = methanol

based on 125 records for **SN2 reaction rate constant** (excluding 20 not indexed molecules)

ir similarity: Any

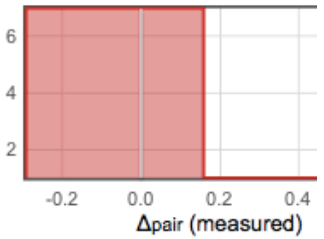


Transformation details:



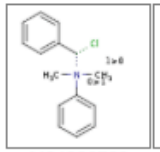
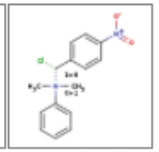
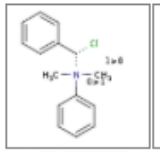
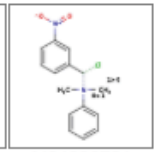
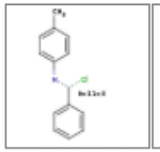
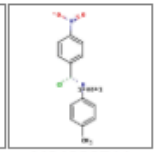
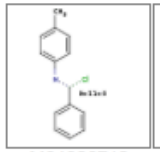
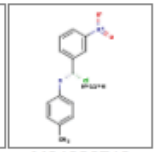
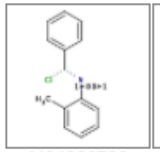
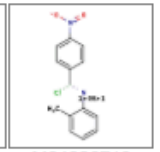
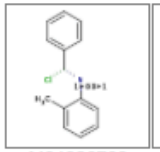
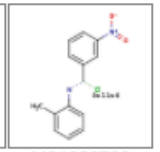
Transformation ID: TR291
SMIRKS: *[H] -> *N(=O)=O

Back to all transformations

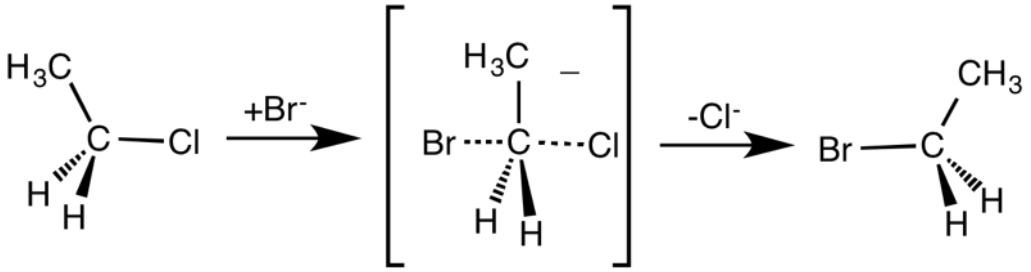


Pairs having the same transformation as the selected one:

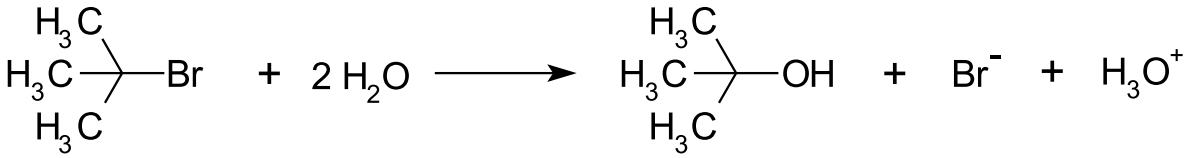
1 - 10 of 10

-4.43	-5.17	-4.43	-4.95	-3.99	-4.34
					
M84326720	M84326722	M84326720	M84326721	M84326712	M84326714
-3.99	-4.18	-4.35	-4.86	-4.35	-4.69
					
M84326712	M84326713	M84326708	M84326710	M84326708	M84326709

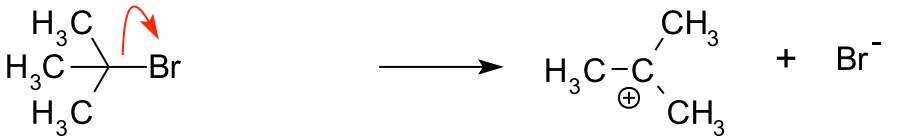
SN2 vs. SN1 reactions



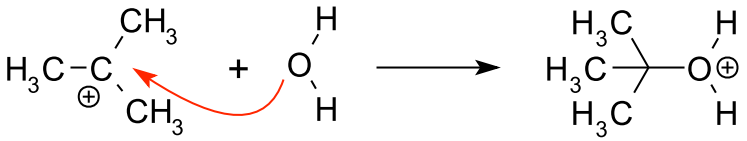
SN2: one step

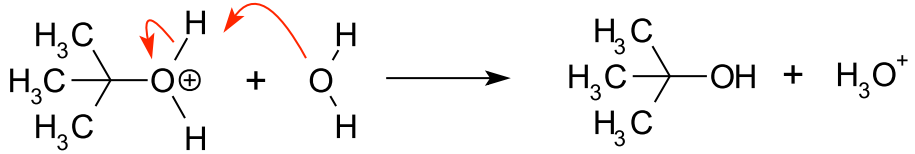


SN1: one step



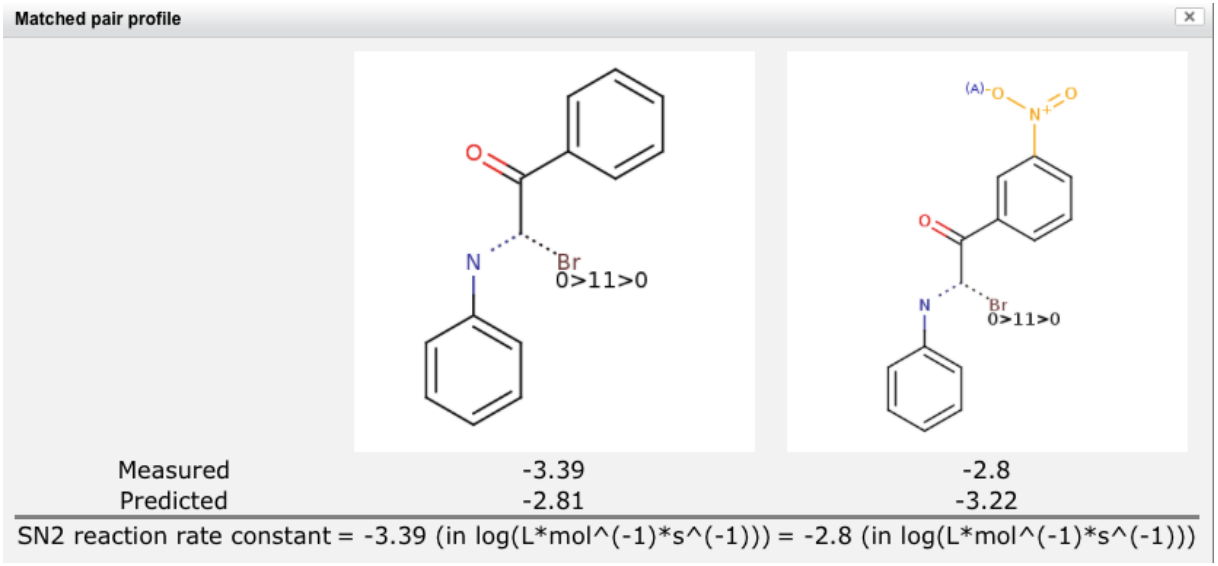
rate determining step



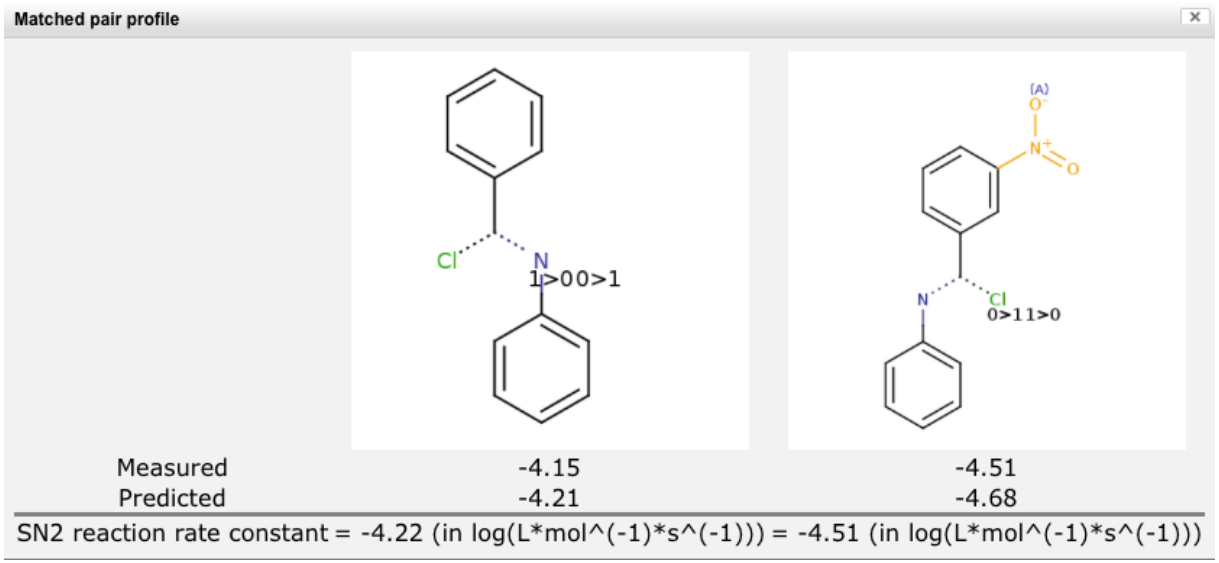


http://www.wikipedia.org/

SN2 vs. SN1 reaction



SN2



SN1

Evaluation in challenges

www.epa.gov/ncct/challenges.html

EPA United States Environmental Protection Agency

LEARN THE ISSUES | SCIENCE & TECHNOLOGY | LAWS & REGULATIONS | ABOUT EPA

ALL EPA THIS AREA Advanced Search SEARCH

Computational Toxicology Research

Contact Us

You are here: EPA Home » Research & Development » CompTox » Chemical Data Challenges & Release

Key Links

CompTox Home
Basic Information
Organization
EPA Exposure Research
Research Projects
Chemical Databases
ToxCast Stakeholder Events
EPA Chemical Safety Research
Research Publications
Scientific Reviews
Communities of Practice
ToxCast Data Challenges
Staff Profiles
CompTox Partners
Jobs and Opportunities

ToxCast Chemical Data Challenges and Release

EPA's high-throughput screening data on 1,800 chemicals is accessible through the interactive Chemical Safety for Sustainability Dashboards (iCSS dashboard). The iCSS dashboard provides user-friendly and customizable access to toxicity data from ToxCast and Tox21 high-throughput chemical screening technologies.

Using the TopCoder and InnoCentive crowd-sourcing platform, EPA invited the science and technology community to work with the data and provide solutions for how the new toxicity data can be used to predict potential health effects. The ToxCast data challenges focused on using this data and other publicly available data to predict the lowest effect level from traditional toxicity studies using laboratory animals. Challenge winners received awards for solving this challenge.

Key Links

- Lowest Effect Level Challenge Results (PDF, 497KB, 18pp)
- Chemical Safety for Sustainability Dashboards
- Complete ToxCast Phase II Data & Files
- TopCoder Challenge
- InnoCentive Challenge
- Stakeholder Workshops

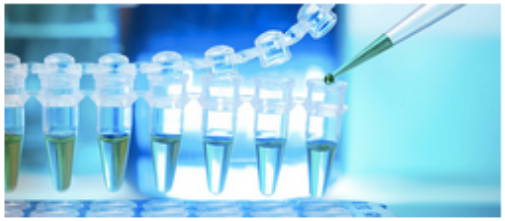
National Center for Advancing Translational Sciences

Tox21 Data Challenge 2014

Login

Open call ends: November 14, 2014

About the Data



The Challenge

The 2014 Tox21 data challenge is designed to help scientists understand the potential of the chemicals and compounds being tested through the Toxicology in the 21st Century initiative to disrupt biological pathways in ways that may result in toxic effects.

The goal of the challenge is to "crowdsource"

All challenge winners will receive the opportunity to submit a paper for publication in a special thematic issue of *Frontiers in Environmental Science* and recognition on the NCATS website and via social media.

Modeling capabilities

- Top-I rank submission model (May 2014) – entry by Sergii Novotarskyi*

- Two Top-I rank individual sub-challenges and overall best balanced accuracy for all targets (January 2015) – entry by Ahmed Abdelaziz**

*Novotarskyi, S. et al. Chem. Res. Toxicol. 2016, 29, 768-75.
**Abdelaziz, A. et al. Front. Environ. Sci. 2016, 4, 2.

Conclusions

- **On-line tools are important for dissemination**
- OCHEM
  - Can used for chemical data searching
  - ToxAlerts for understanding of hemical data
  - SetCompare can be useful for comparison of sets of molecules
  - Has a powerful modeling framework
  - Easily handles millions of compounds
  - Contribute highly predictive models
  - MMP analysis can complement modeling efforts

Acknowledgements

Dr. Y. Sushko
Dr. S. Novotarskyi
Mr. R. Körner
Mr. A. Abdelaziz
Mrs. E. Salmina

Prof. M. Sattler (HMGU)
Prof. G. Wess (HMGU)

Dr. K. Hadian (TOX, HMGU)
Dr. A. Williams (USA)
Dr. D. Lowe (UK)
Dr. H. Varbanov (Switzerland)
Prof. N. Haider (Austria)

MARIE CURIE ACTIONS

THE FRAMEWORK PROGRAMME FOR RESEARCH AND INNOVATION
HORIZON 2020

eco
environmental chemoinformatics
marie curie initial training network

Alexander von Humboldt
Stiftung/Foundation

Thank you for your attention!