



## Overview of Chemoinformatics group

Dr. Igor V. Tetko  
Institute of Structural Biology  
Helmholtz Zentrum Muenchen

Munich, 18/01/2016

Welcome to OCHEM! Your possible actions

### Explore OCHEM data

Search chemical and biological data: experimentally measured, published and exposed to public access by our users. You can also [upload your data](#).

### Create QSAR models

Build QSAR models for predictions of chemical properties. The models can be based on the experimental data published in our database.

### Run predictions

Apply one of the available models to predict property you are interested in for your set of compounds.

### Screen compounds with ToxAlerts

Screen your compound libraries against structural alerts for such endpoints as mutagenicity, skin sensitization, aqueous toxicity, etc.

### Tutorials

Check our video tutorials to know more about the OCHEM features.

### Our acknowledgements

Feedback and help

### User's manual

Check an online user's manual

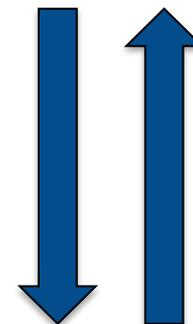
Check out the properties available on OCHEM

OCHEM contains 1151032 experimental records for about 490 properties collected from 12539 sources

**Melting Point** **logPow** **logBB** **LogL(water)** **Cbrain/Cblood**  
**LogD** **Cblood/Cair** **Cbrain/Cair** **Cfat/Cair** **Cover/Cair** **Cmuscle/Cair** **SIF solubility** **logPI(+)** **logPI(-)**  
**logS** **LogL(blood)** **LogL(brain)** **LogL(fat)** **LogL(heart)** **LogL(kidney)** **LogL(liver)** **LogL(lungs)**  
**LogL(muscle)** **LogL(oil)** **LogL(plasma)** **LogBPR** **LogCSFR** **ER** **fu(brain)** **PI/Papp**  
**Biodistributon(kidney)** **Biodistributon(liver)** **Biodistributon(lungs)** **Biodistributon(muscle)** **Biodistributon(heart)**  
**Cbrain/Cplasma** **IC 50** **Papp(Caco-2)** **Papp(MDCK)** **P(brain)**  
**Oral absorption** **LIC 50** **pK(1/10gK)** **Ctiver/Cplasma** **Clung/Cplasma** **Cheart/Cplasma**  
**Ckidney/Cplasma** **Cbrain/Cairum** **Cfat/Cplasma** **Cmuscle/Cplasma** **Cskin/Cplasma** **Papp ratio(Caco-2)**  
**Papp(MBIA)** **Plasma protein binding** **Papp(HPBEC)** **Pendothelial(HPBEC)**  
**Papp(BBEC)** **Pendothelial(BBEC)** **Papp ratio(HPBEC)** **Pendothelial ratio(HPBEC)** **Papp(SV-ARBEC)**  
**Pendothelial(SV-ARBEC)** **Papp(MBEC4)** **Papp ratio(MDCKATCC)** **Pendothelial ratio(SV-ARBEC)** **Papp ratio(SV-ARBEC)**  
**Papp ratio(MDCK-wt)** **Papp ratio(MDCK-mdr1)** **pIC50** **%Human FA** **Human IA**  
**Human FA** **ELogD** **fraction unbound (fu)** **fraction ionized (fi)** **pKa**  
**VDss** **%Human OB** **LogIC50** **Cgut/Cplasma** **Cbone/Cplasma** **APow** **APow** **LogD<sub>ov</sub>**  
**LogPI** **LogP<sub>ov</sub>(ion)** **LogP<sub>ov</sub>** **BBB permeability (qualitative)**  
**LogPeff(human jejunum)** **Peff(human jejunum)** **LogKoa** **LogRBA**  
**CYP450 modulation** **CYP450 reaction**  
**Vapor Pressure** **Water solubility** **Bioconcentration factor**  
**EC50 aquatic** **NOEC aquatic** **LOEC aquatic** **NOEC terrestrial**  
**IC50 aquatic** **LC50 aquatic** **log(IGC50-1)** **LEL**  
**Henry's law constant** **Photolysis rate Kp** **Half-Life Photolysis HLp** **Photolysis quantum yield**



ideas & research;  
HMGU/STB

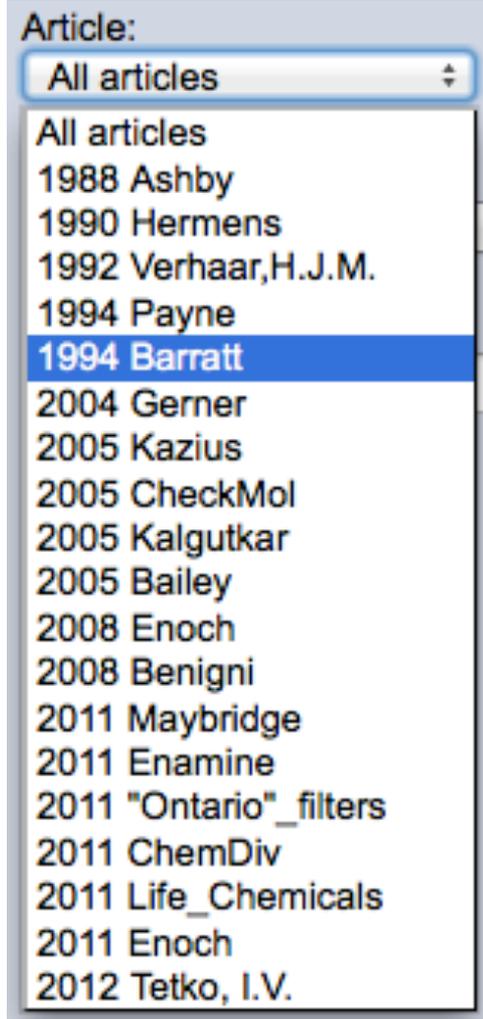
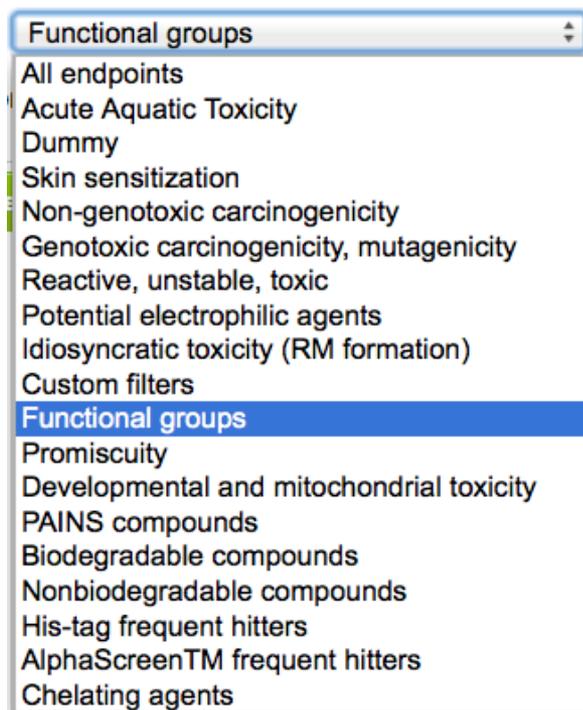


commercial exploitation;  
BigChem GmbH (spin-off)

- OCHEM platform was initially developed in HMGU and it is exclusively licensed to BigChem GmbH
- Over 750K experimental measurements with curated data
- Alerts database >2.3k filters for toxicity, reactivity, stability, frequent hitters, etc.
- About 40 models for various physico-chemical and biological properties
- Assessment of applicability domain and accuracy estimation for all models
- All models can be extended (self-learning) with new data, when they become available
- Modeling of complex properties like bioavailability and metabolic stability
- Virtual screening to identify and prioritize molecules

# ToxAlerts

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



# Functional groups

Online chemical database  
with modeling environment

v2.4.45  
Welcome, Guest! Logout

Home Database Models A+ a-

ToxAlerts: Structural alerts browser  
Here you can browse structural alerts for various toxicological endpoints

Upload new alerts Screen compounds

101 - 200 of 379

100 items on page 2 of 4

**HS**

**Four-membered heterocycles with one heteroatom (HS)**  
A = any atom except carbon; a dashed line indicates any type of covalent bonds  
High specificity (HS) pattern matches chemicals that include exact heterocyclic moiety as in the depiction (fusion with other ring(s) are not allowed)

SMARTS: [\*6][\*1;R]1-[\*6R1]-[\*6R1]-[\*6R1]-1  
Endpoint: Functional groups

Salmima, E.  
Extended functional groups (EFG): an efficient set for chemi...  
Molecules 2015; subm ()  
Alert ID: TA2450

16:45, 8 Mar 13 / 13:25, 31 Oct 15  
SALMINA1987 / Itelko

**HS**

**Saturated four-membered heterocycles with one heteroatom (HS)**  
A = any atom except carbon  
High specificity (HS) pattern matches chemicals that include exact heterocyclic moiety as in the depiction (fusion with other ring(s) are not allowed)

SMARTS: [\*6][\*1;R]1-[\*6R1]-[\*6R1]-[\*6R1]-1  
Endpoint: Functional groups

Salmima, E.  
Extended functional groups (EFG): an efficient set for chemi...  
Molecules 2015; subm ()  
Alert ID: TA2451

16:45, 8 Mar 13 / 13:25, 31 Oct 15  
SALMINA1987 / Itelko

**HS**

**Azetidines (HS)**  
High specificity (HS) pattern matches chemicals that include exact heterocyclic moiety as in the depiction (fusion with other ring(s) are not allowed)

SMARTS: [\*7R]1-[\*6R1]-[\*6R1]-[\*6R1]-1  
Endpoint: Functional groups

Salmima, E.  
Extended functional groups (EFG): an efficient set for chemi...  
Molecules 2015; subm ()  
Alert ID: TA2452

16:45, 8 Mar 13 / 13:25, 31 Oct 15  
SALMINA1987 / Itelko

**HS**

**Oxetanes (HS)**  
High specificity (HS) pattern matches chemicals that include exact heterocyclic moiety as in the depiction (fusion with other ring(s) are not allowed)

SMARTS: [\*8R]1-[\*6R1]-[\*6R1]-[\*6R1]-1  
Endpoint: Functional groups

Salmima, E.  
Extended functional groups (EFG): an efficient set for chemi...  
Molecules 2015; subm ()  
Alert ID: TA2453

16:45, 8 Mar 13 / 13:25, 31 Oct 15  
SALMINA1987 / Itelko

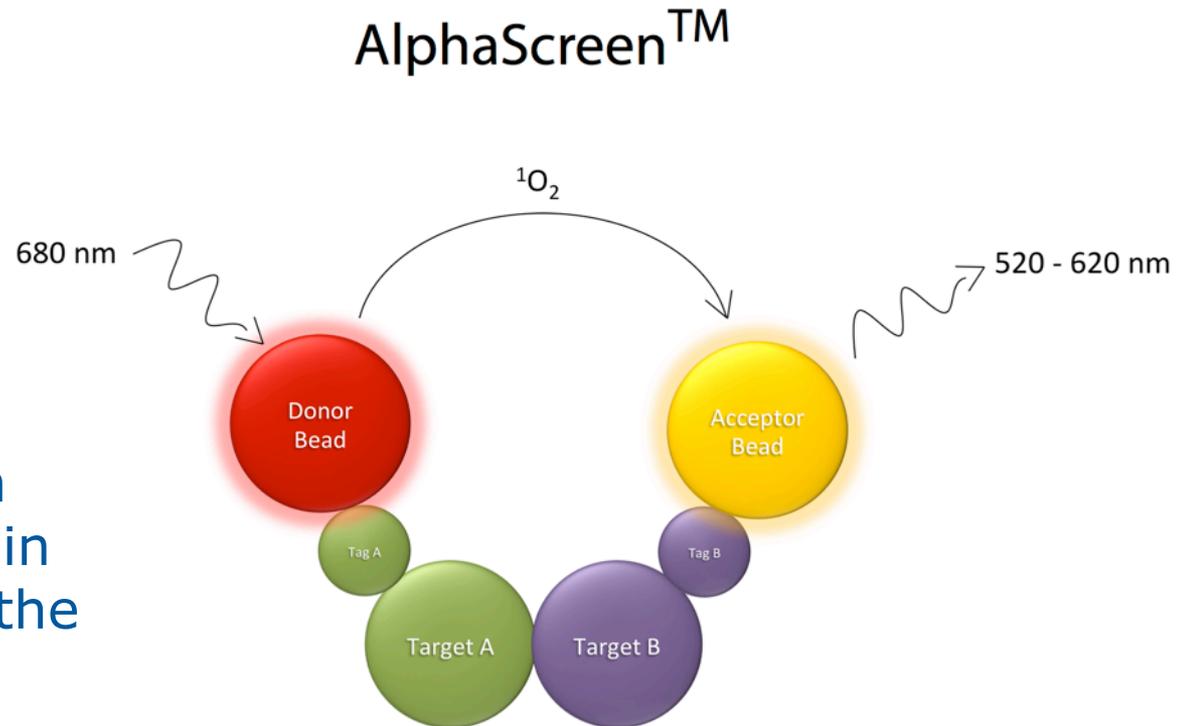
**HS**

**Thietanes (HS)**  
High specificity (HS) pattern matches chemicals that include exact heterocyclic moiety as in the depiction (fusion with other ring(s) are not allowed)

*Salmima et al, Molecules, 2016, 21(1), 1-16.*

# 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 by Schoorp et al, **2014**, 19(5):715-726.
- GST tag frequent hitters, Brenke et al, **2016**, in press.

# 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 two. 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
<b>Pnictogens</b> Group 15: the nitrogen family				
<b>N P As</b>	13235 (96.0%)	202449 (88.7%)	1.1	1.42E-198
<b>Sb Bi</b>				
	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

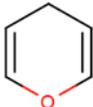
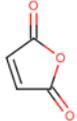
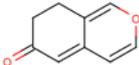
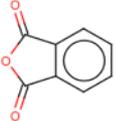
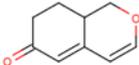
Pyrolysis vs. melting point

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

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

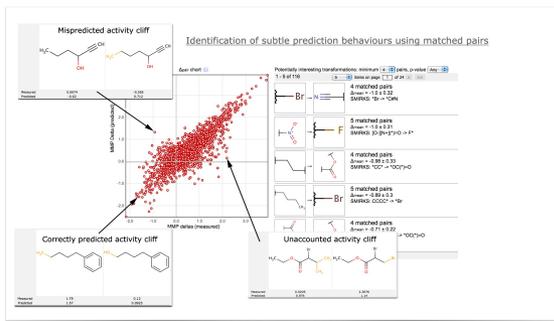
[Export results as a CSV file](#)

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

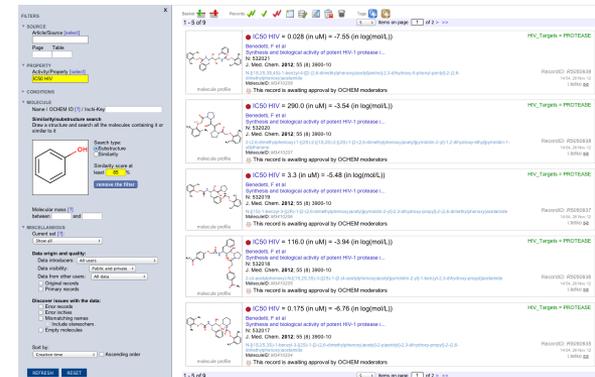
HIV Envelope glycoprotein GP1

# Modeling iterative workflow



Select dataset

- Over 800K measured values
- Over 400 property

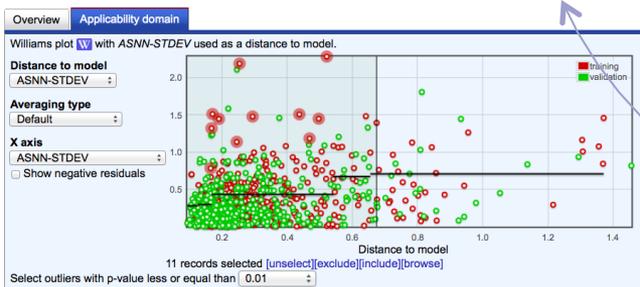


Apply and interpret

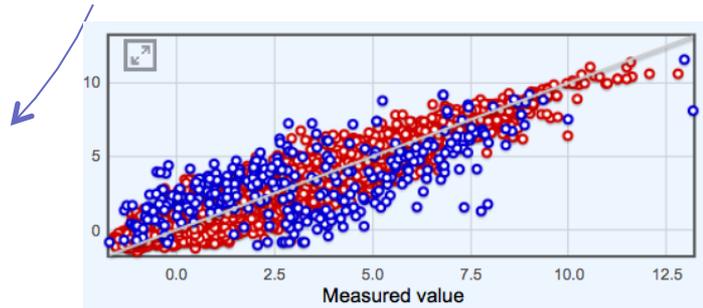
Prepare and review

Validate  
Internal (N-Fold cross-validation, Bagging)  
External validation

Select descriptors  
(24 packages: 0D, 1D, 2D 3D)



Build model  
(MLR, ANN, KNN, KRR, SVM, FSMLR, KPLS, LogP, WEKA-J48)

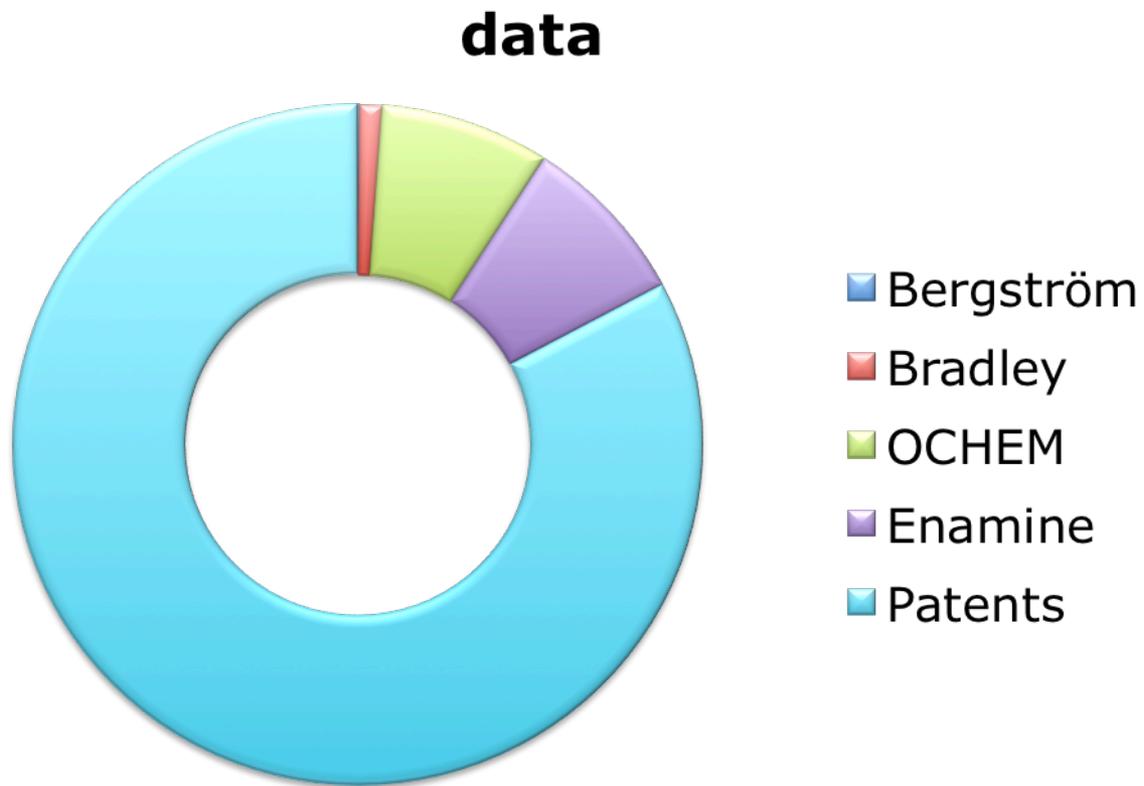


Validate and estimate

Develop and analyze

# 300k Melting Points Dataset

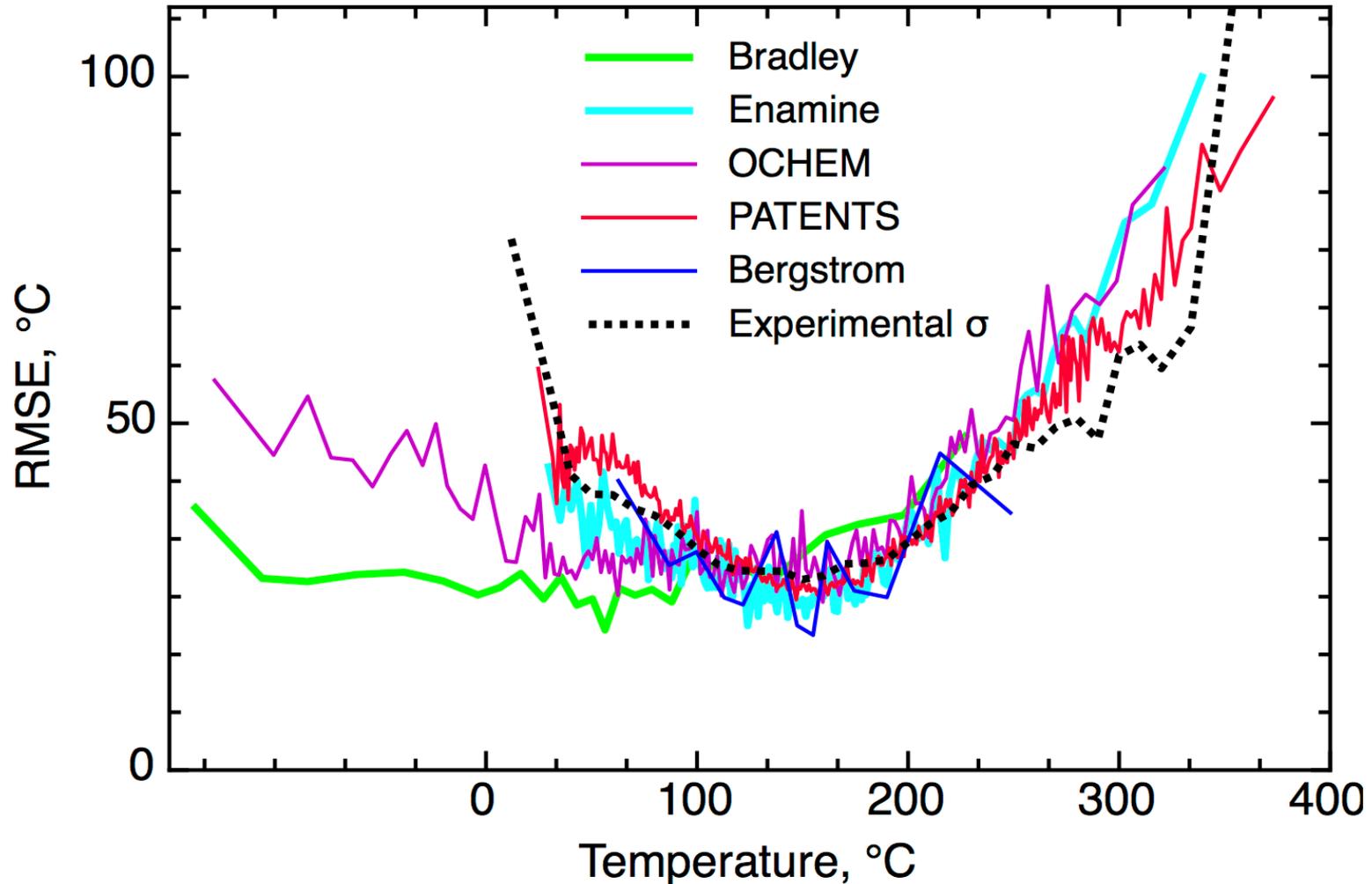
Bergström	277
Bradley	2886
OCHEM	22404
Enamine	21883
Patents	228079



# 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

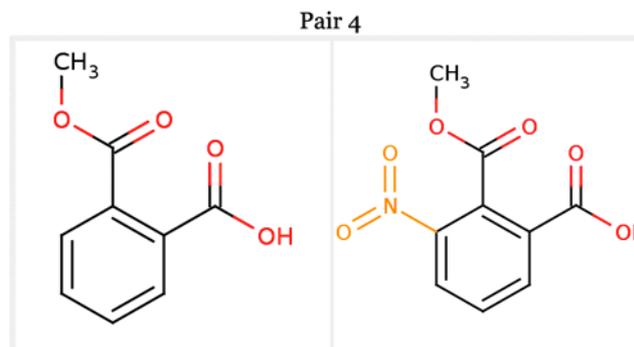
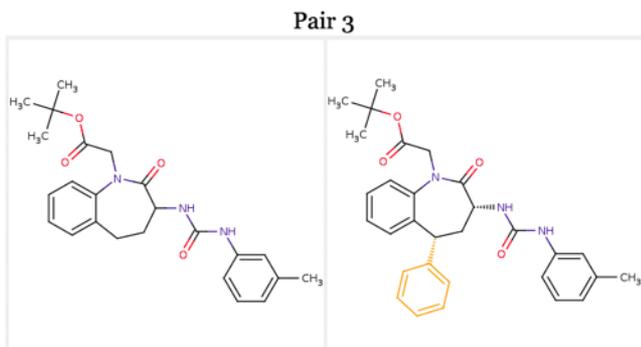
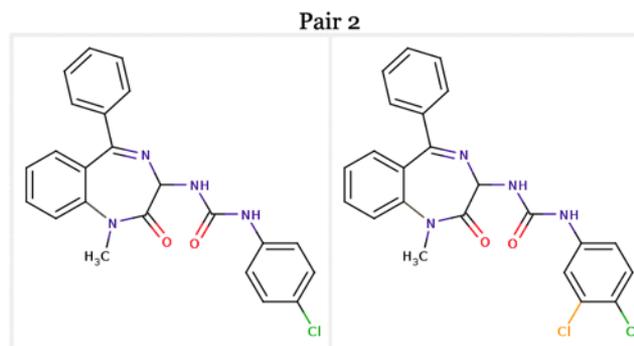
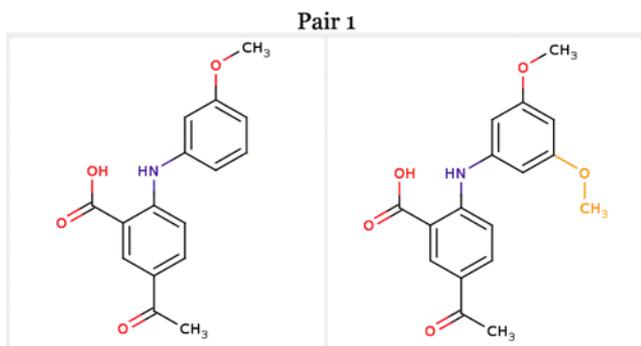
# Prediction error as function of experimental MP for analyzed datasets



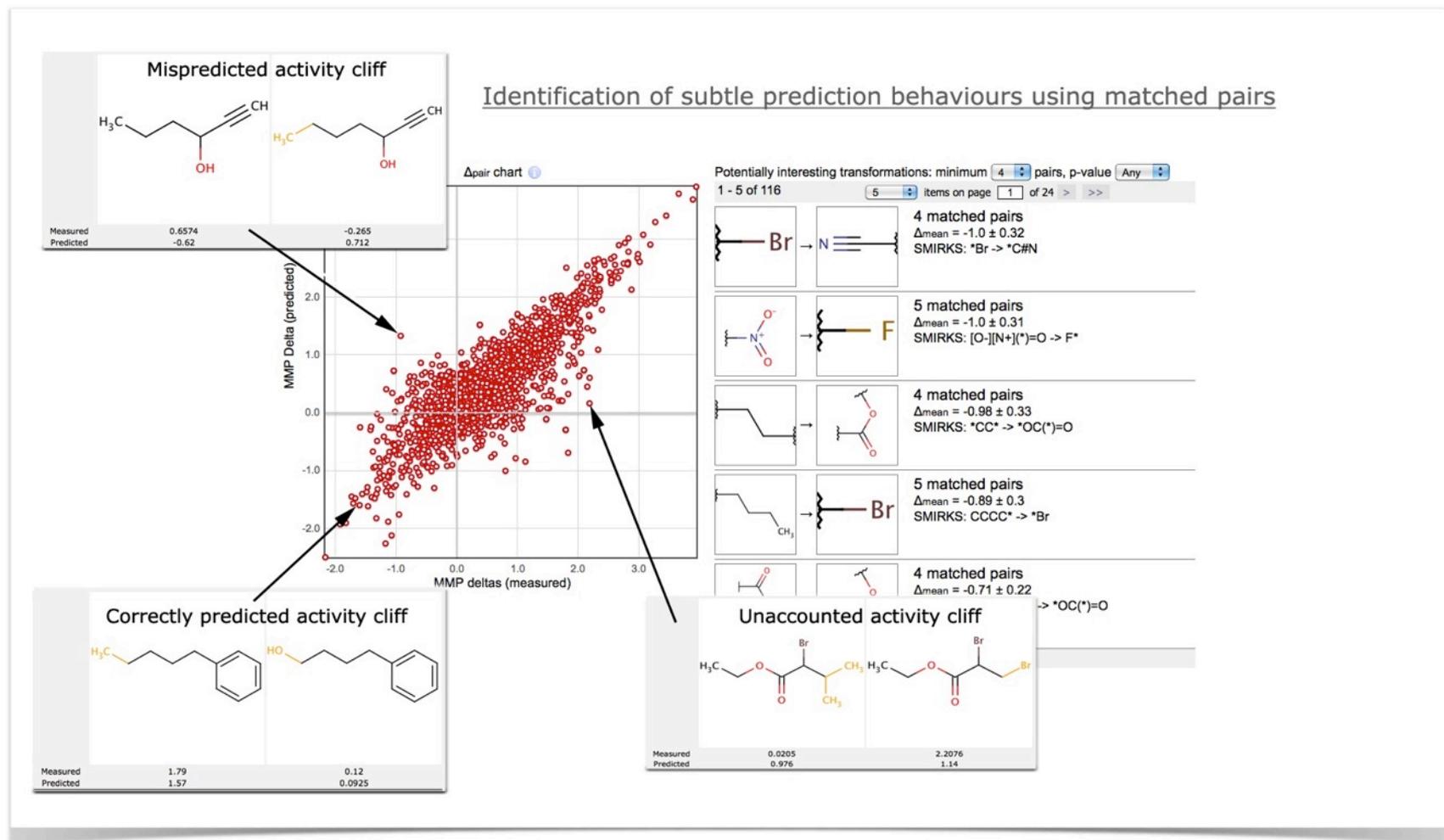
Experimental  $\sigma$  is based on  $N = 18058$  duplicated measurements

# 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.



# Analysis of rules that were learnt by models



## 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](#)

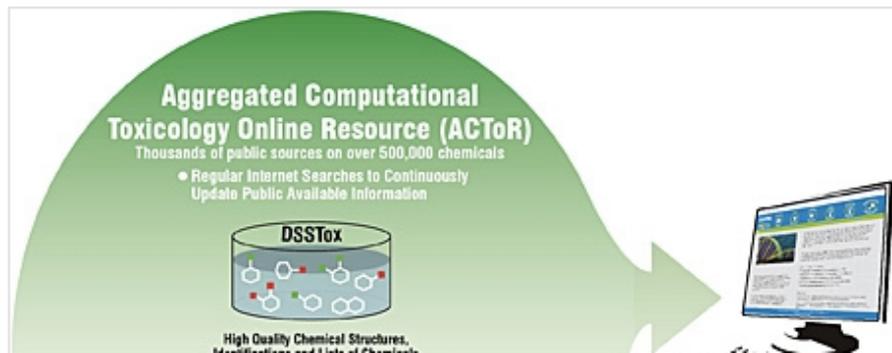
## ToxCast Chemical Data Challenges and Release

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](#)



**Aggregated Computational Toxicology Online Resource (ACToR)**  
Thousands of public sources on over 500,000 chemicals

- Regular Internet Searches to Continuously Update Public Available Information

**DSSTox**

High Quality Chemical Structures, Modified Chemical Structures

The diagram illustrates the flow of data from the DSSTox database (represented by a beaker with chemical structures) through a large green arrow to the ACToR online resource, which is shown as a computer monitor displaying a search interface.



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.

# OCHEM Modeling capabilities

- Top-1 rank submission model (May 2014) – entry by Sergii Novotarskyi<sup>1</sup>
- Two Top-1 rank individual sub-challenges and overall best balanced accuracy for all targets (January 2015) – entry by Ahmed Abdelaziz<sup>2</sup>

<sup>1</sup>Novotarskyi et al, submitted

<sup>2</sup>Abdelaziz et al, *Front. Environ. Sci.*, **2016**, 4:2.



# Participation in BigChem

- ◆ ESR4 Development of frequent hitters filters for HTS screening (with LDC)
- ◆ ESR8 Accessing new chemical space for lead optimization based on QSAR models (with AZ)
- ◆ ESR10 Secure sharing of information using ensemble of machine learning methods and surrogate data (with AZ)