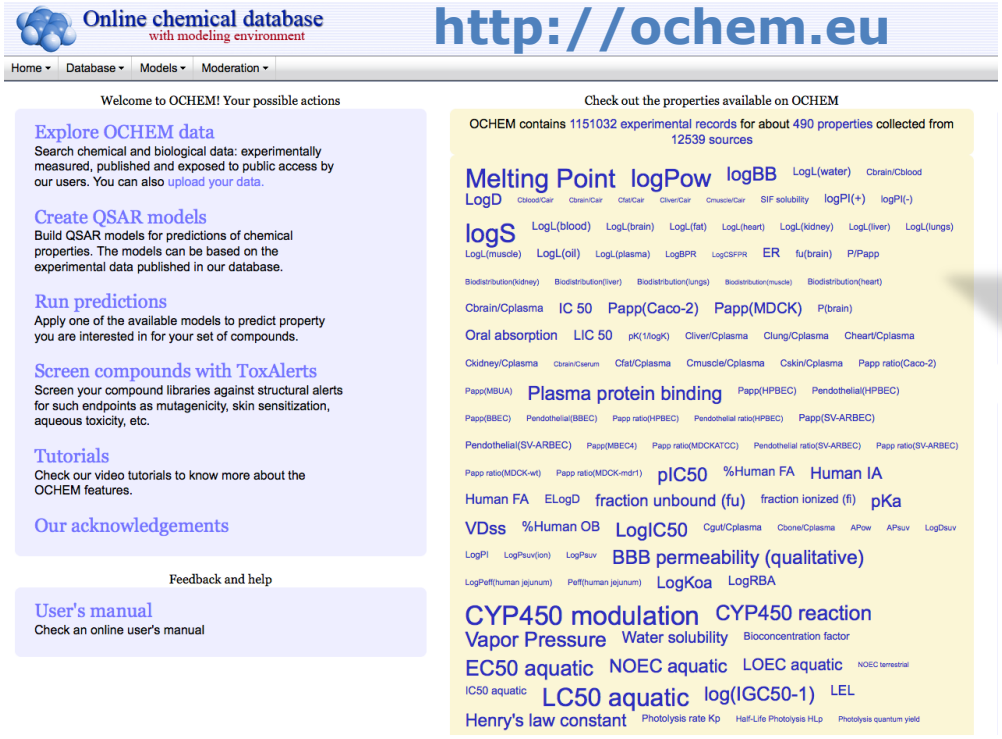
HelmholtzZentrum münchen
German Research Center for Environmental Health

Overview of Chemoinformarics group

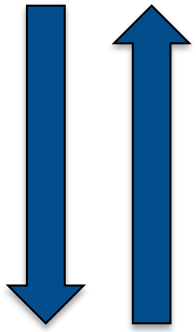
Dr. Igor V. Tetko
Institute of Structural Biology
Helmholtz Zentrum Muenchen

Munich, 18/01/2016

HELMHOLTZ ASSOCIATION



ideas & research;
HMGU/STB

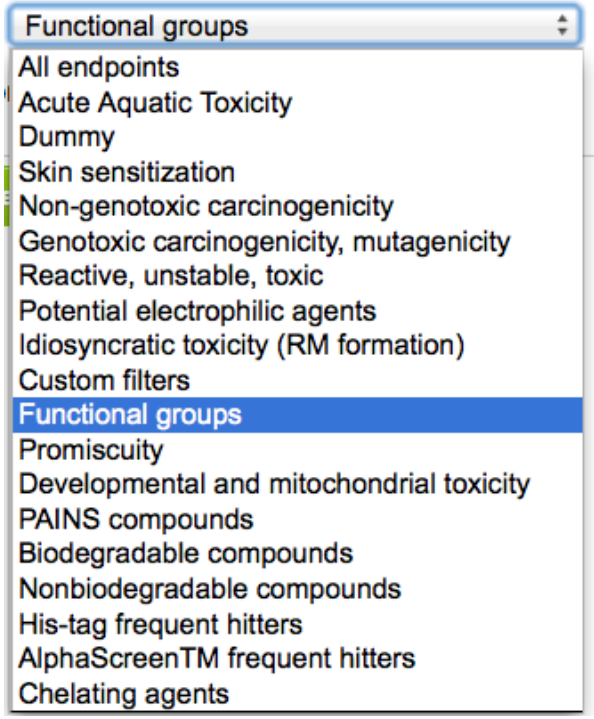


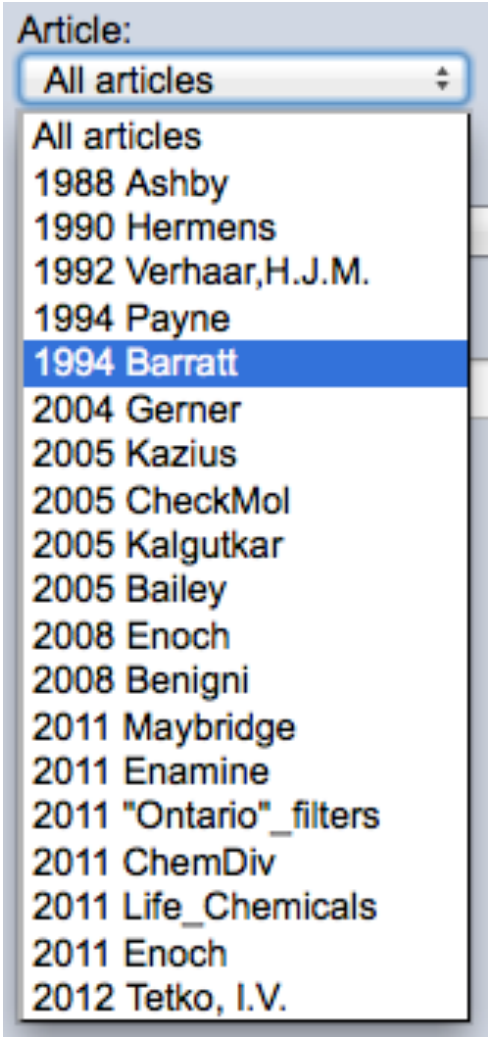
commercial exploitation;
BigChem GmbH (spin-off)

- 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

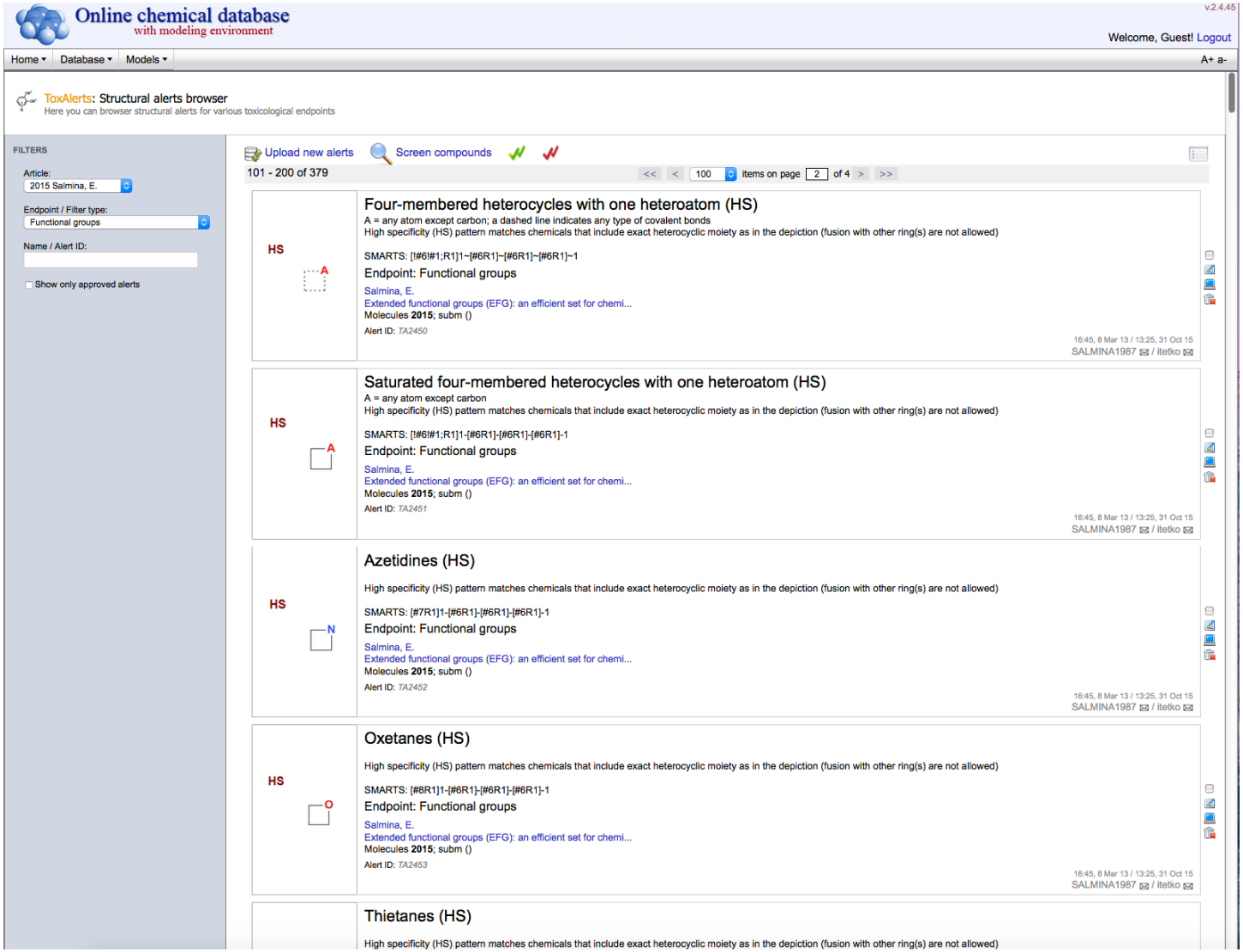
- Screening of compounds against published toxicity alerts, groups, frequent hitters
- Filter alerts by endpoints or publications
- Create or upload custom SMARTS rules





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

Functional groups



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

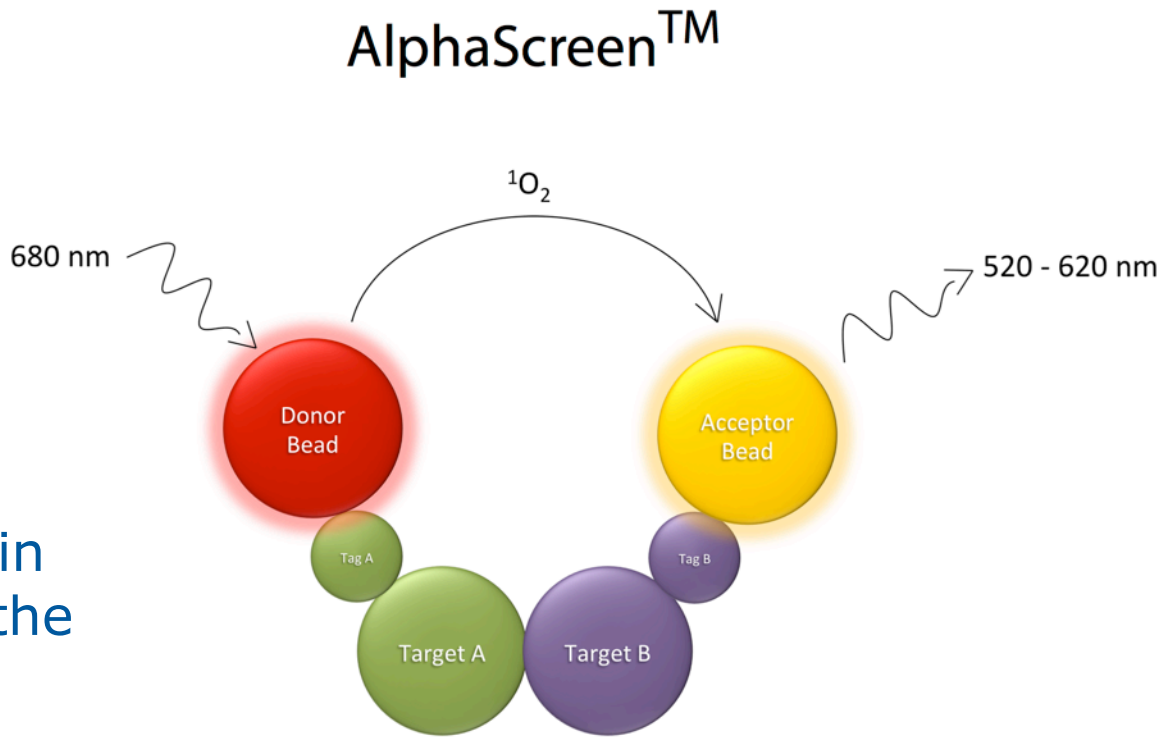
Screening Artefacts

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 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

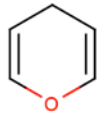
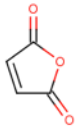
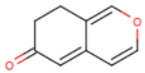
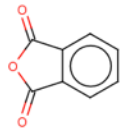
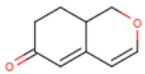
Descriptor	In set 1 (13785 molecules)	In set 2 (228174 molecules)	Enrichment factor	p-Value
	3232 (23.4%)	26196 (11.5%)	2.0	1.33E-315
Pnictogens Group 15: the nitrogen family **N P As Sb Bi**	13235 (96.0%)	202449 (88.7%)	1.1	1.42E-198
	3257 (23.6%)	79741 (34.9%)	1.5	-1.25E-172
	280 (2.0%)	352 (0.2%)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

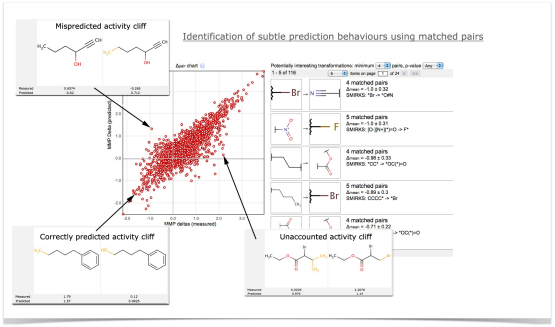
The following table shows the features (molecular descriptors) that were significantly overrepresented in on
It includes appearance counts of the features in each set and the p-Value of such a distribution.
Export results as a CSV file

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 GP1

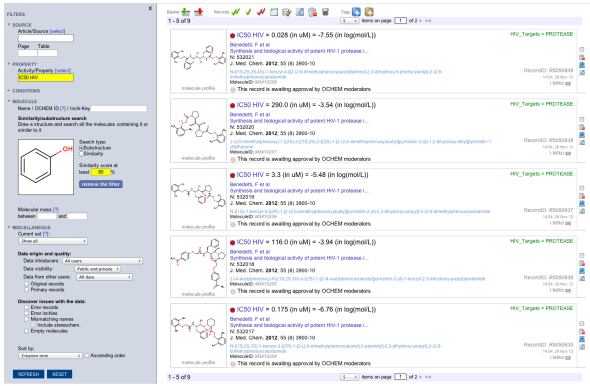
Modeling iterative workflow



Apply and interpret

Select dataset

- Over 800K measured values
- Over 400 property



Prepare and review

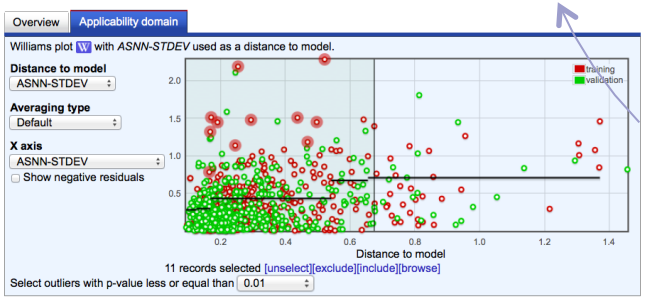
Validate

Internal (N-Fold cross-validation, Bagging)

External validation

Select descriptors

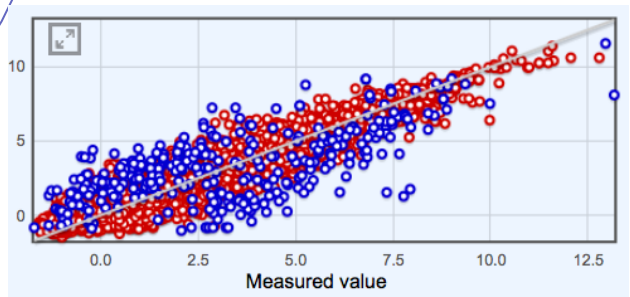
(24 packages:0D, 1D, 2D 3D)



Validate and estimate

Build model

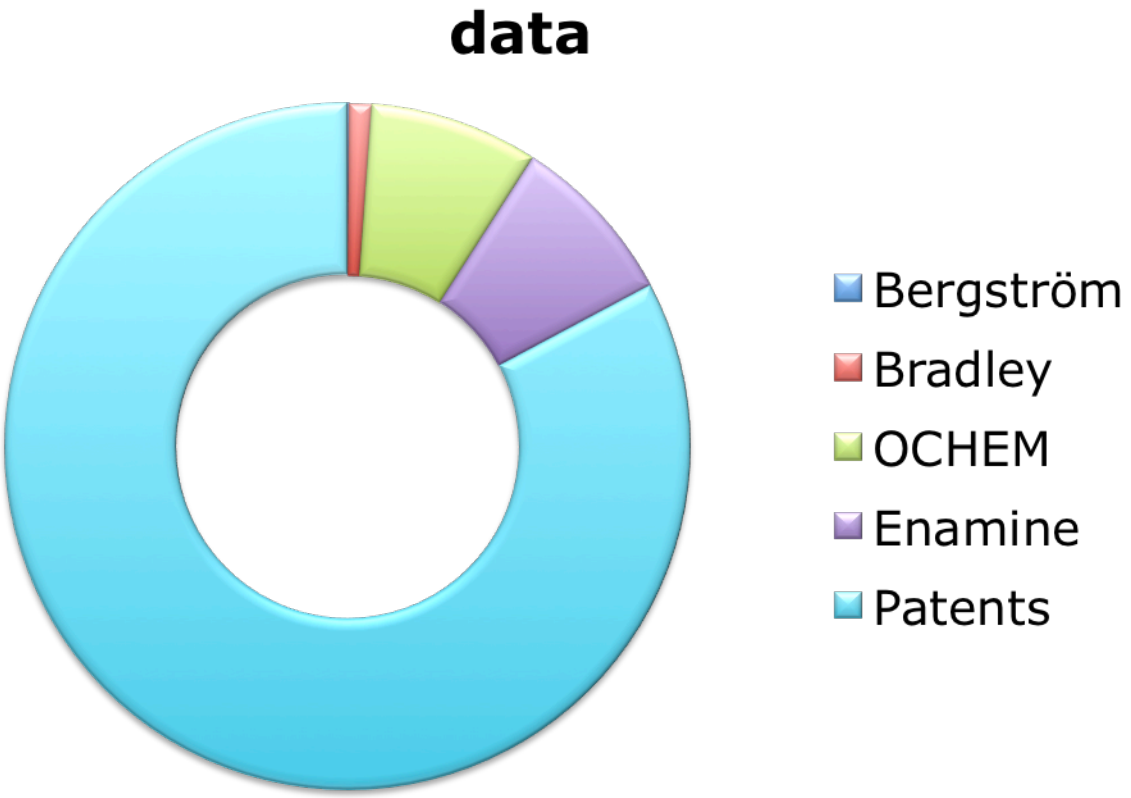
(MLR, ANN, KNN, KRR, SVM, FSMLR, KPLS, LogP, WEKA-J48)



Develop and analyze

300k Melting Points Dataset

Bergström	277
Bradley	2886
OCHEM	22404
Enamine	21883
Patents	228079

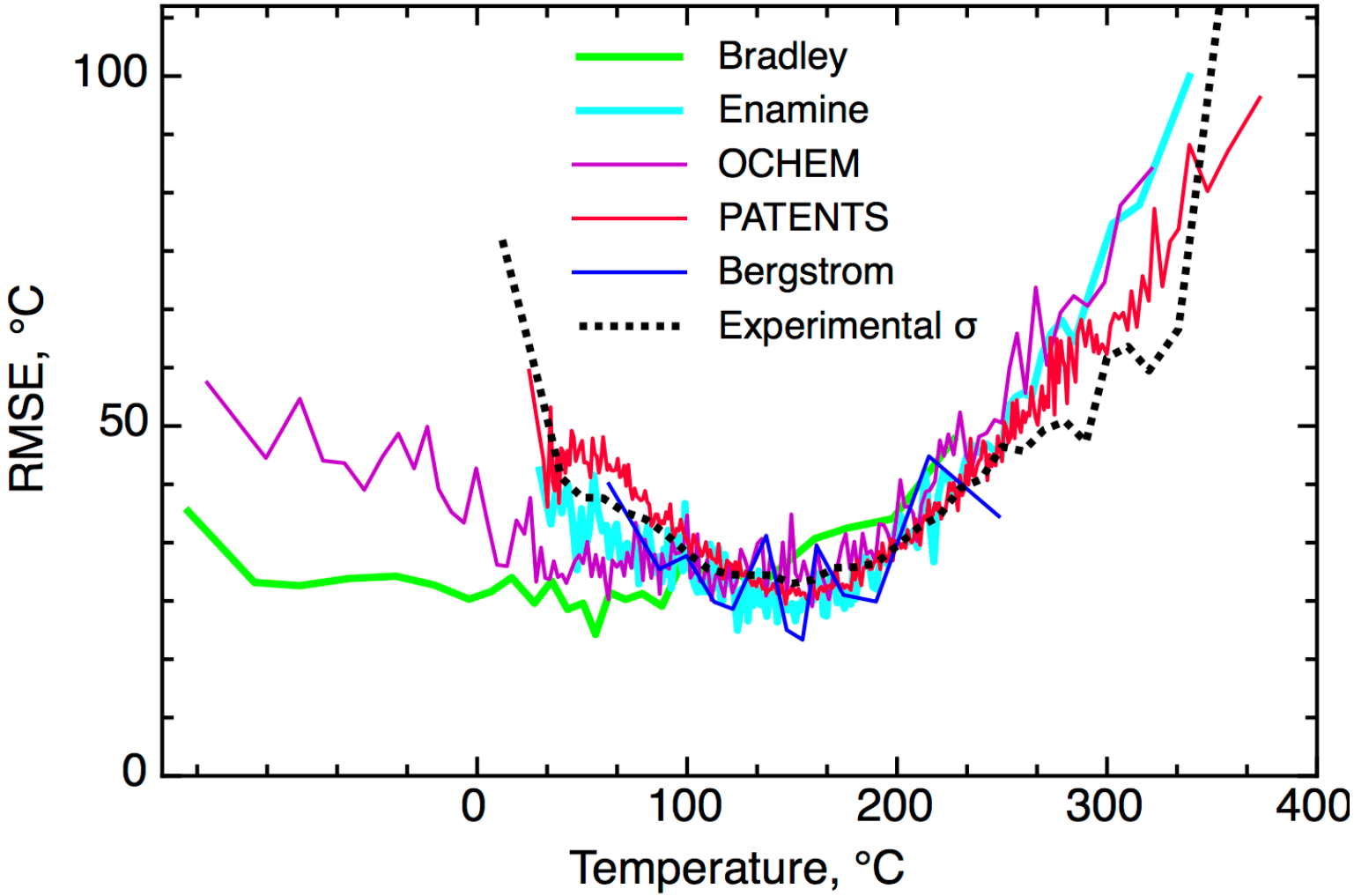


Tetko et al *J. Chemoinformatics,* **2016**, *Jan 22;8:2.*

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

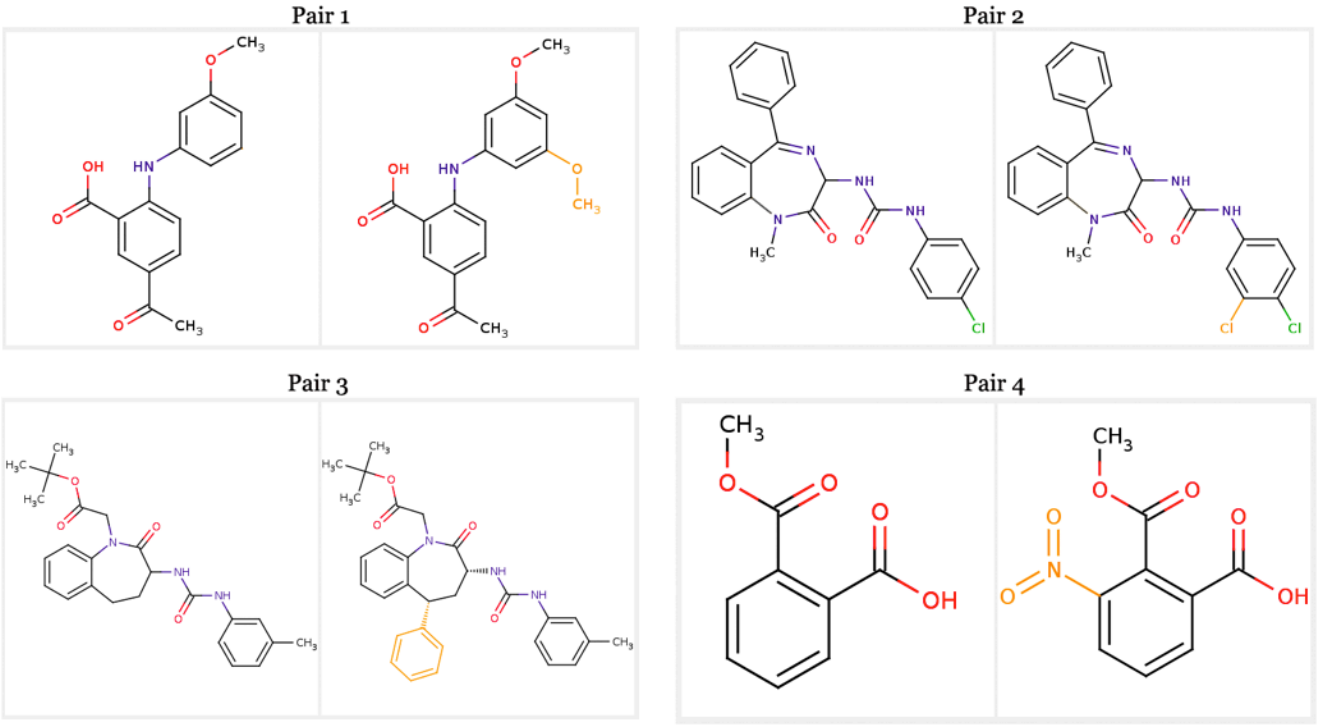
Prediction error as function of experimental MP for analyzed datasets



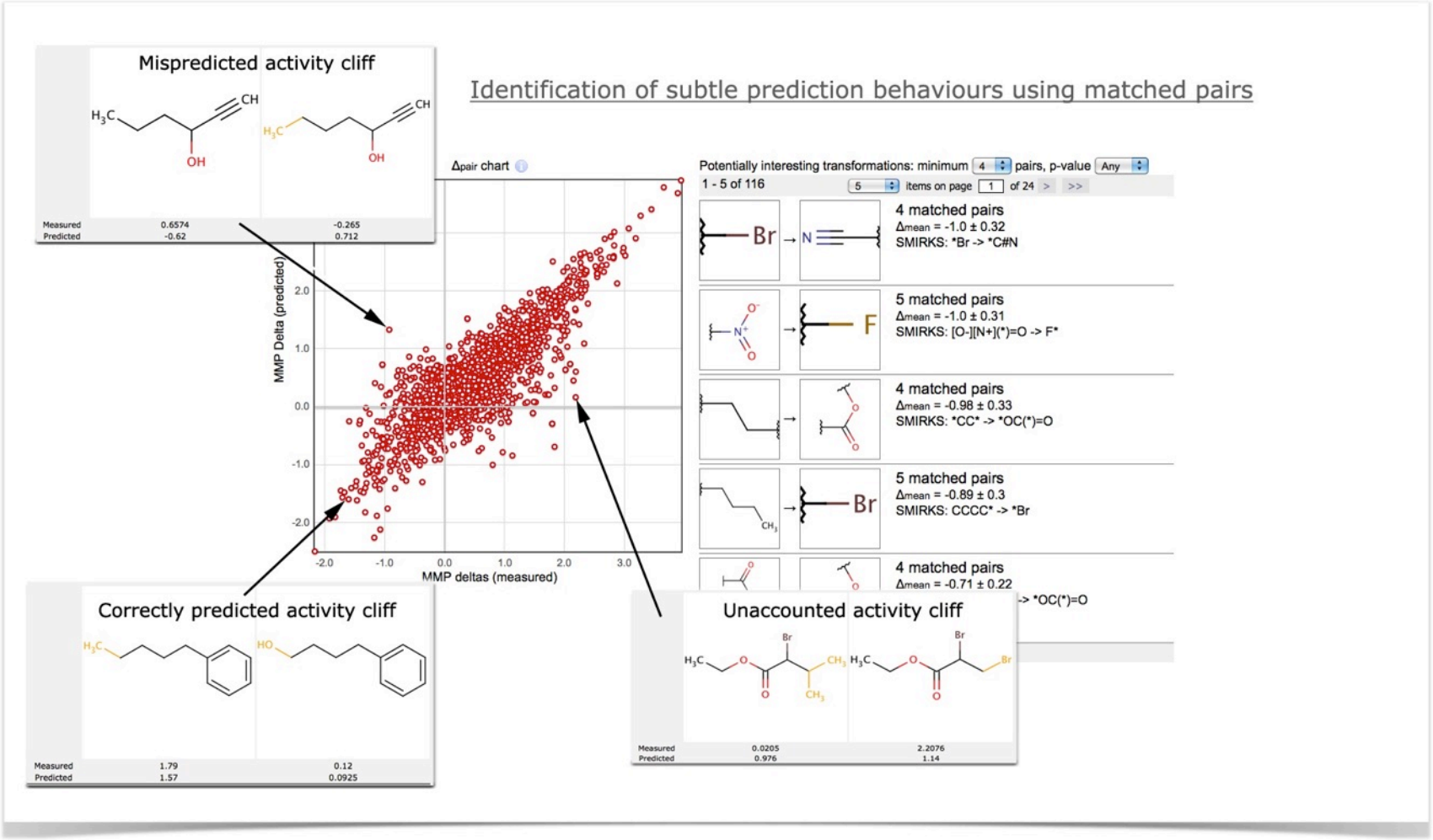
Experimental σ is based on *N* = 18058 duplicated measurements

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



Sushko et al, J Cheminform. 2014 Dec 11;6(1):48.

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

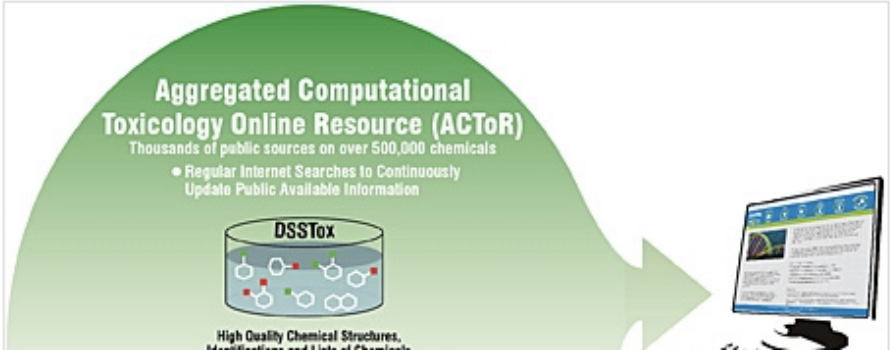
ToxCast Chemical Data Challenges and Release

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.

OCHEM Modeling capabilities

- Top-I rank submission model (May 2014) – entry by Sergii Novotarskyi[1]

- Two Top-I rank individual sub-challenges and overall best balanced accuracy for all targets (January 2015) – entry by Ahmed Abdelaziz[2]

[1]Novotarskyi et al, submitted
[2]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)