CONCERTREACH
CONCERTING EXPERIMENTAL DATA
AND IN SILICO MODELS FOR REACH

LIFE17 GIE/IT/00046

The open OCHEM QSAR models

Igor V. Tetko

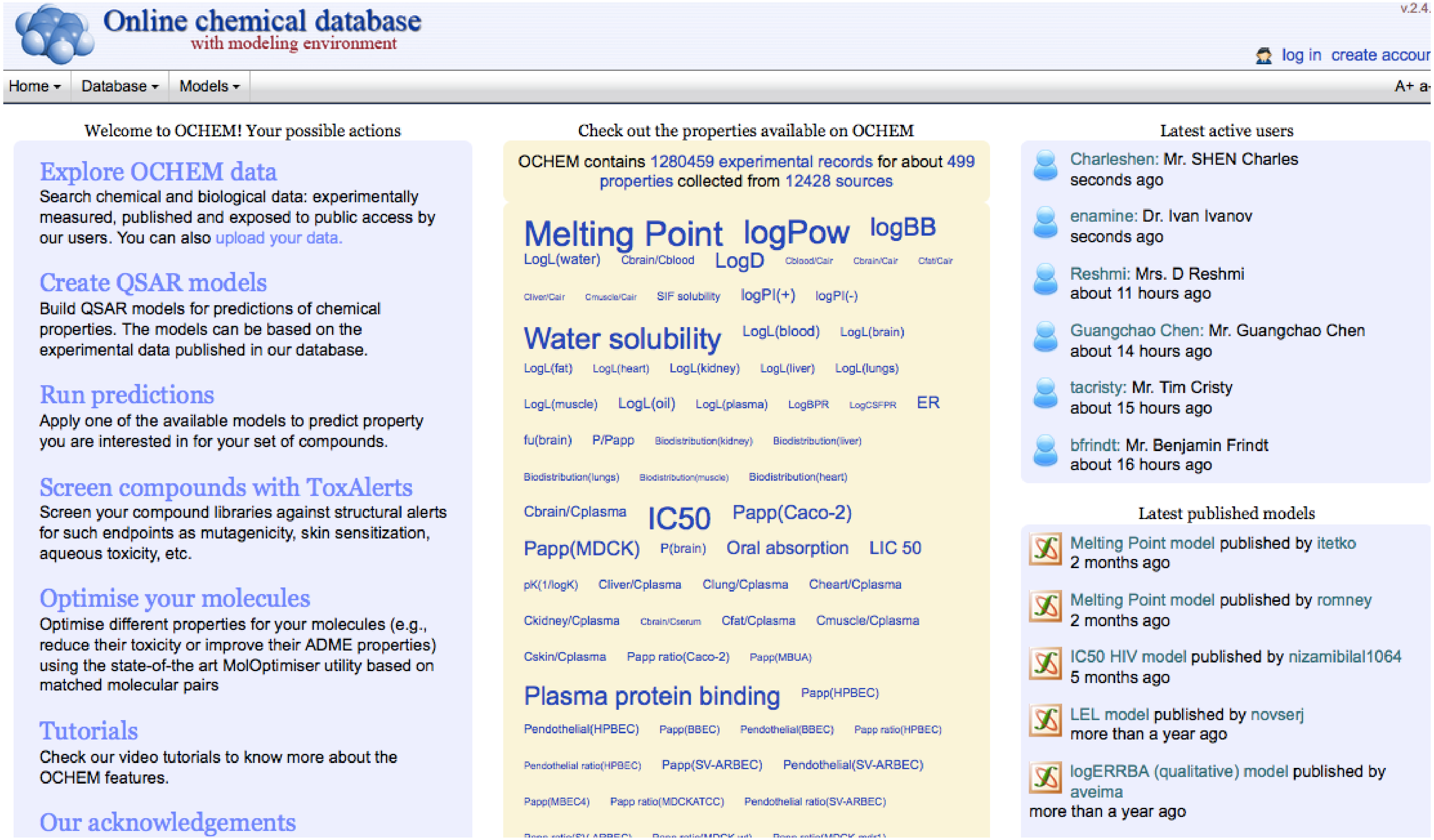
BIGCHEM GmbH and Helmholtz Munich

June 19, 2023, Final Workshop, Milano, Italy

HELMHOLTZ MUNICH

big chem

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



Online chemical database
with modeling environment

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.

Optimise your molecules

Optimise different properties for your molecules (e.g., reduce their toxicity or improve their ADME properties) using the state-of-the art MolOptimiser utility based on matched molecular pairs

Tutorials

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

Our acknowledgements

Check out the properties available on OCHEM

OCHEM contains 1280459 experimental records for about 499 properties collected from 12428 sources

Melting Point logPow logBB LogL(water) Cbrain/Cblood LogD Cblood/Cair Cbrain/Cair Cfat/Cair Cliver/Cair Cmuscle/Cair SIF solubility logPI(+) logPI(-) Water solubility LogL(blood) LogL(brain) LogL(fat) LogL(heart) LogL(kidney) LogL(liver) LogL(lungs) LogL(muscle) LogL(oil) LogL(plasma) LogBPR LogCSFPR ER fu(brain) P/Papp Biodistribution(kidney) Biodistribution(liver) Biodistribution(lungs) Biodistribution(muscle) Biodistribution(heart) Cbrain/Cplasma IC50 Papp(Caco-2) Papp(MDCK) P(brain) Oral absorption LIC 50 pK(1/logK) Cliver/Cplasma Clung/Cplasma Cheart/Cplasma Ckidney/Cplasma Cbrain/Cserum Cfat/Cplasma Cmuscle/Cplasma Cskin/Cplasma Papp ratio(Caco-2) Papp(MBUA) 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)

Latest active users

Charleshen: Mr. SHEN Charles
seconds ago

enamine: Dr. Ivan Ivanov
seconds ago

Reshmi: Mrs. D Reshmi
about 11 hours ago

Guangchao Chen: Mr. Guangchao Chen
about 14 hours ago

tacristy: Mr. Tim Cristy
about 15 hours ago

bfrindt: Mr. Benjamin Frindt
about 16 hours ago

Latest published models

Melting Point model published by itetko
2 months ago

Melting Point model published by romney
2 months ago

IC50 HIV model published by nizamibilal1064
5 months ago

LEL model published by novserj
more than a year ago

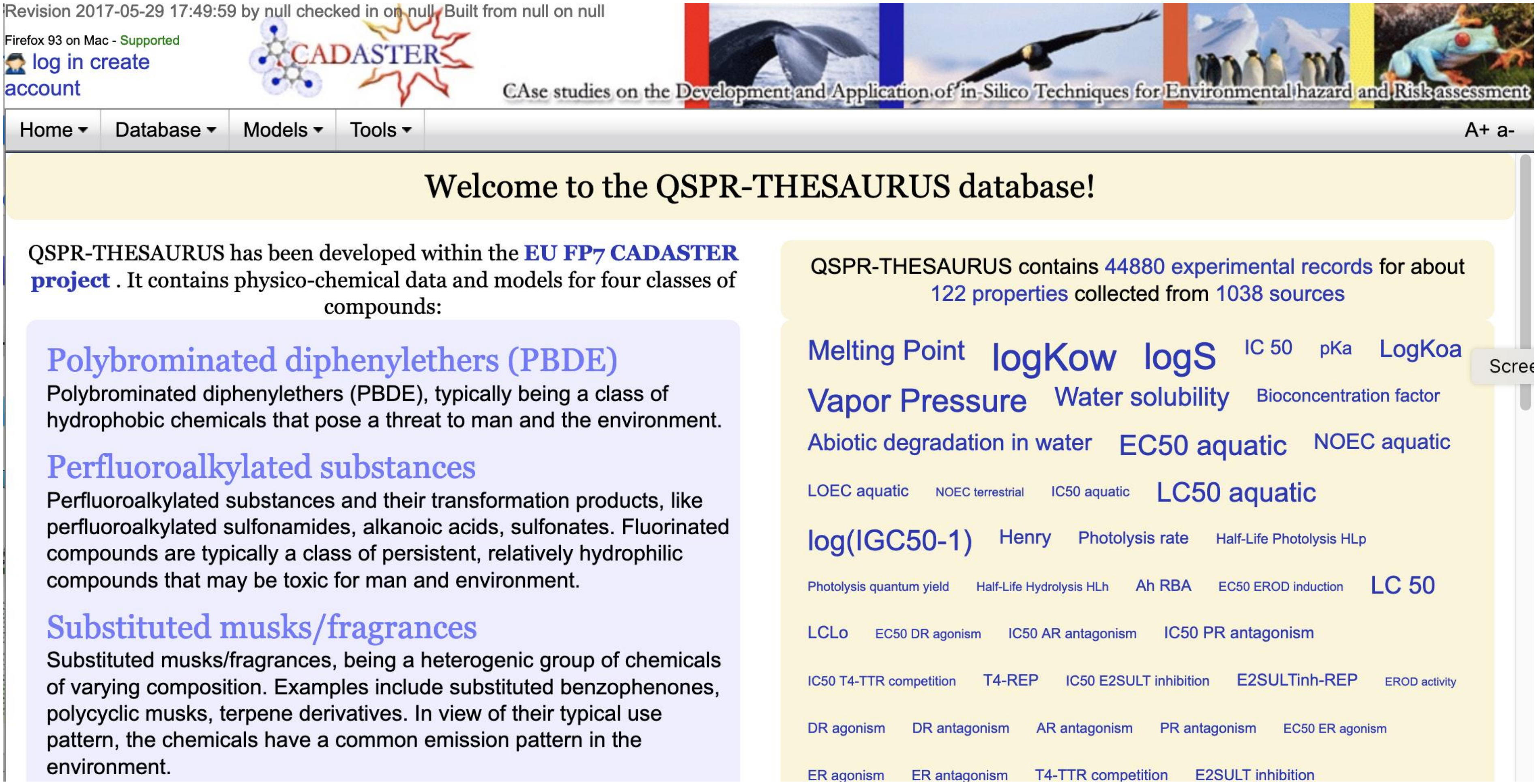
logERRBA (qualitative) model published by aveima
more than a year ago

HELMHOLTZ
MUNICH

German Research Center for Environmental Health

BIGCHEM GmbH is a spin-off of the center

OCHEM was partially developed within FP7 CADASTER


CADASTER
CAse studies on the Development and Application of in-Silico Techniques for Environmental hazard and Risk assessment
log in create account
Home
Database
Models
Tools
A+ a-
Scree

Revision 2017-05-29 17:49:59 by null checked in on null, Built from null on null

Firefox 93 on Mac - Supported

Welcome to the QSPR-THESAURUS database!

QSPR-THESAURUS has been developed within the **EU FP7 CADASTER project** . It contains physico-chemical data and models for four classes of compounds:

Polybrominated diphenylethers (PBDE)

Polybrominated diphenylethers (PBDE), typically being a class of hydrophobic chemicals that pose a threat to man and the environment.

Perfluoroalkylated substances

Perfluoroalkylated substances and their transformation products, like perfluoroalkylated sulfonamides, alkanoic acids, sulfonates. Fluorinated compounds are typically a class of persistent, relatively hydrophilic compounds that may be toxic for man and environment.

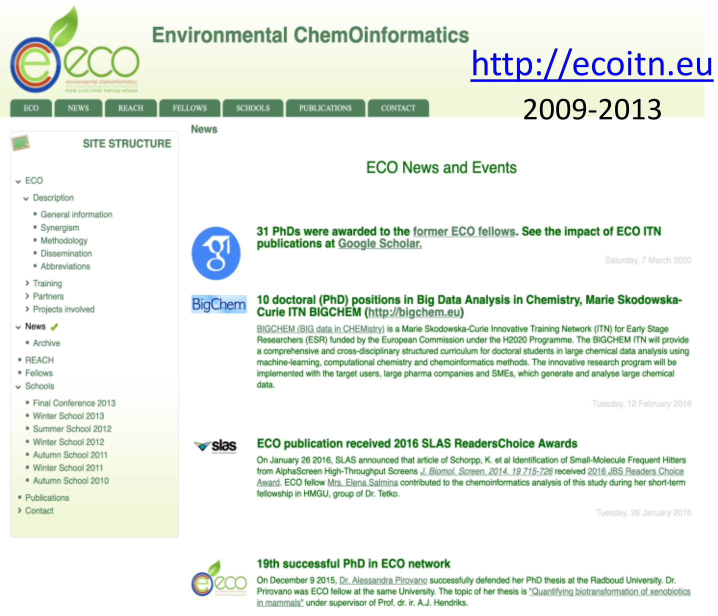
Substituted musks/fragrances

Substituted musks/fragrances, being a heterogenic group of chemicals of varying composition. Examples include substituted benzophenones, polycyclic musks, terpene derivatives. In view of their typical use pattern, the chemicals have a common emission pattern in the environment.

QSPR-THESAURUS contains 44880 experimental records for about 122 properties collected from 1038 sources

Melting Point logKow logS IC 50 pKa LogKoa Vapor Pressure Water solubility Bioconcentration factor Abiotic degradation in water EC50 aquatic NOEC aquatic LOEC aquatic NOEC terrestrial IC50 aquatic LC50 aquatic log(IGC50-1) Henry Photolysis rate Half-Life Photolysis HLp Photolysis quantum yield Half-Life Hydrolysis HLh Ah RBA EC50 EROD induction LC 50 LCLo EC50 DR agonism IC50 AR antagonism IC50 PR antagonism IC50 T4-TTR competition T4-REP IC50 E2SULT inhibition E2SULTinh-REP EROD activity DR agonism DR antagonism AR antagonism PR antagonism EC50 ER agonism ER agonism ER antagonism T4-TTR competition E2SULT inhibition

OCHEM in MSC ITN networks



Environmental ChemOinformatics

ECO | NEWS | REACH | FELLOWS | SCHOOLS | PUBLICATIONS | CONTACT

SITE STRUCTURE

- ECO
  - Description
    - General information
    - Synergism
    - Methodology
    - Dissemination
    - Abbreviations
  - Training
  - Partners
  - Projects involved
- News ✓
  - Archive
- REACH
- Fellows
- Schools
  - Final Conference 2013
  - Winter School 2013
  - Summer School 2012
  - Winter School 2012
  - Autumn School 2011
  - Winter School 2011
  - Autumn School 2010
- Publications
- Contact

News

ECO News and Events

31 PhDs were awarded to the former ECO fellows. See the impact of ECO ITN publications at Google Scholar.

Saturday, 7 March 2020

10 doctoral (PhD) positions in Big Data Analysis in Chemistry, Marie Skodowska-Curie ITN BIGCHEM (http://bigchem.eu)

BIGCHEM (BIG data in CHEMistry) is a Marie Skodowska-Curie Innovative Training Network (ITN) for Early Stage Researchers (ESR) funded by the European Commission under the H2020 Programme. The BIGCHEM ITN will provide a comprehensive and cross-disciplinary structured curriculum for doctoral students in large chemical data analysis using machine-learning, computational chemistry and chemoinformatics methods. The innovative research program will be implemented with the target users, large pharma companies and SMEs, which generate and analyse large chemical data.

Tuesday, 12 February 2016

ECO publication received 2016 SLAS ReadersChoice Awards

On January 26 2016, SLAS announced that article of Schorpp, K. et al Identification of Small-Molecule Frequent Hitters from AlphaScreen High-Throughput Screens *J. Biomol. Screen. 2014, 19 715-726* received 2016 JBS Readers Choice Award. ECO fellow Mrs. Elena Salmina contributed to the chemoinformatics analysis of this study during her short-term fellowship in HMGU, group of Dr. Tetko.

Tuesday, 26 January 2016

19th successful PhD in ECO network

On December 9 2015, Dr. Alessandra Pirovano successfully defended her PhD thesis at the Radboud University. Dr. Pirovano was ECO fellow at the same University. The topic of her thesis is "Quantifying biotransformation of xenobiotics in mammals" under supervisor of Prof. dr. ir. A.J. Hendriks.

http://ecoitn.eu

2009-2013

BigChem http://bigchem.eu

big data in chemistry + informatics = chemoinformatics

The **increasing volume of biomedical data** in chemistry and life sciences requires development of **new methods and approaches for their analysis.**

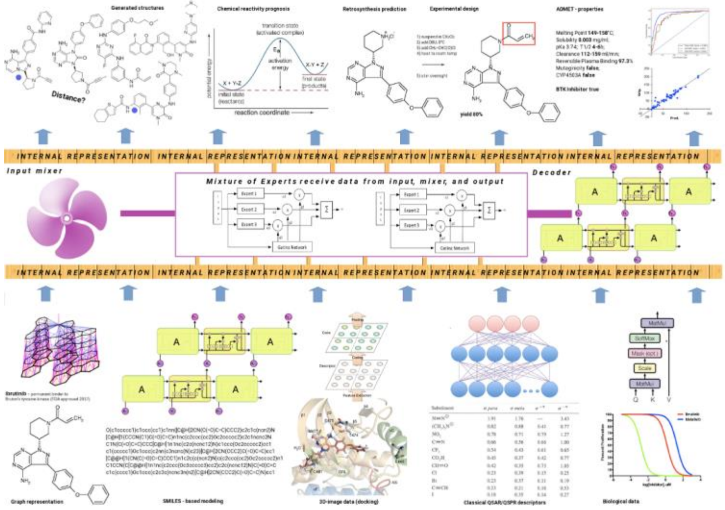
The BIGCHEM project will provide **innovative education in large chemical data analysis**. The innovative research program will be implemented with the target users, **large pharma companies and SMEs**, which generate and analyze large chemical data as well as will promote technology transfer from academy to industrial applications.

Marie Skłodowska-Curie Innovative Training Network
European Industrial Doctorate (2016-2019)

2021-2024

Advanced machine Learning for Innovative drug discovery

http://ai-dd.eu



AiChemist MSC DN

http://aichemist.eu

2023-2027

MSC ITN Project AiChemist

Optimising biological activity and ADME properties, while minimising toxicity, are objectives when developing new compounds. Advanced machine learning methods are indispensable to this process. The project will develop and benchmark representation learning approaches, addressing their accuracy and explainability, using public and *in-house* data for endpoints ranging from chemical reactions to toxicity. The program will be done with the target users: large companies, regulatory agencies and SMEs.

OCHEM statistics

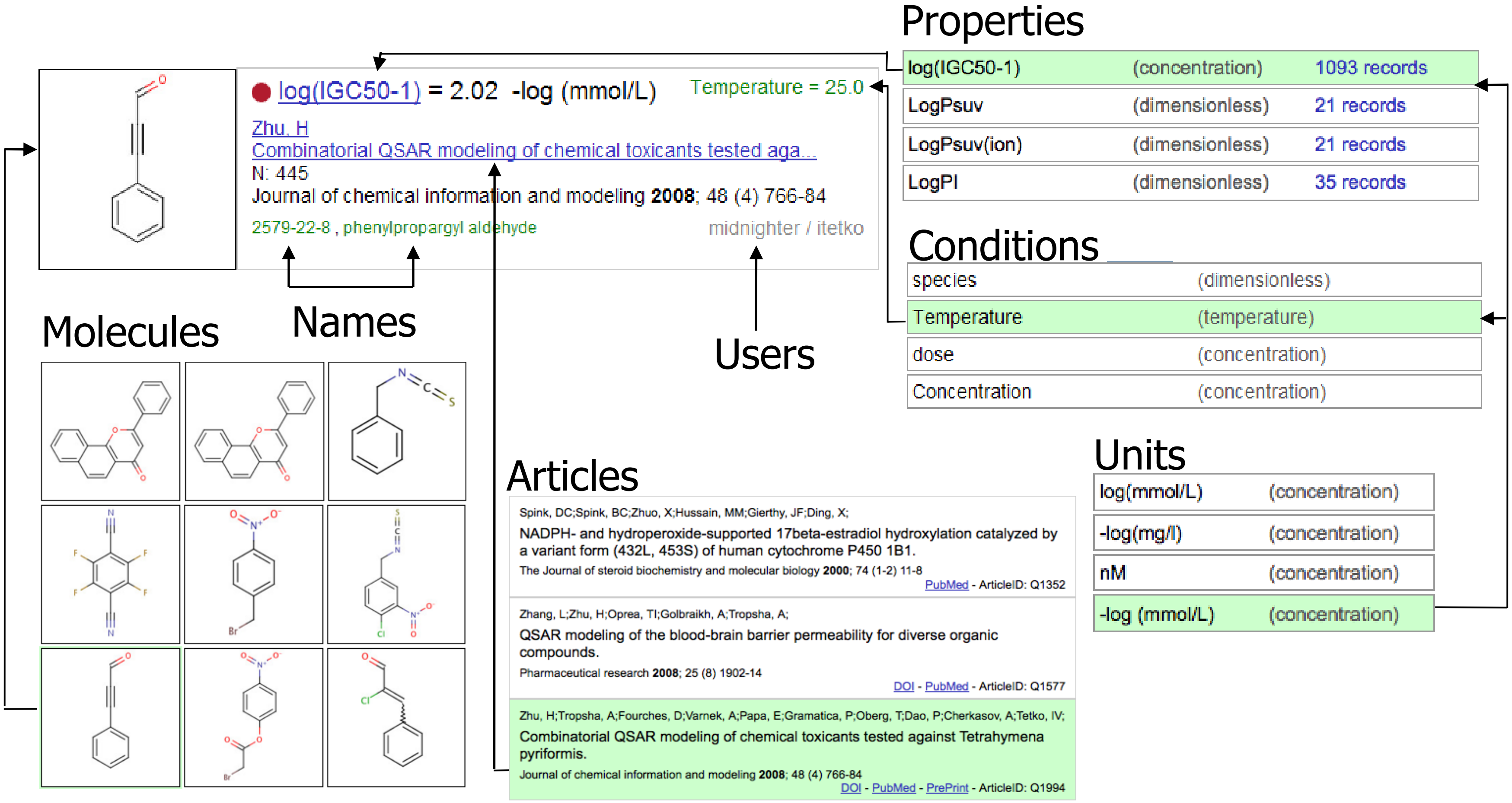
Physico-chemical properties: logP, water solubility, melting point, pyrolysis, vapor pressure, etc.

Biological activity: estrogen receptors; endocrine disruptors; AMES mutagenicity; *in vivo* toxicity

Environmental endpoints: ready biodegradability; fish toxicity; environmental toxicity, etc.

- In total ca 200 published models
- >10,000 registered users
  - 600 commercial
  - 450 governmental
- ca 38 M tasks were executed
- ca 3.4M data points for 692 properties
- >10M uploaded private data points
- Academic groups regularly contribute
- OCHEM is used for teaching
- Top-performing models in challenges (NIH, EPA ToxCast, SLAS Kaggle)

Database schema - Simplified overview



● log(IGC50-1) = 2.02 -log (mmol/L) Temperature = 25.0

Zhu, H

Combinatorial QSAR modeling of chemical toxicants tested aga...

N: 445

Journal of chemical information and modeling **2008**; 48 (4) 766-84

2579-22-8 , phenylpropargyl aldehyde

midnighter / itetko

Properties

log(IGC50-1)	(concentration)	1093 records
LogPsuv	(dimensionless)	21 records
LogPsuv(ion)	(dimensionless)	21 records
LogPI	(dimensionless)	35 records

Conditions

species	(dimensionless)
Temperature	(temperature)
dose	(concentration)
Concentration	(concentration)

Units

log(mmol/L)	(concentration)
-log(mg/l)	(concentration)
nM	(concentration)
-log (mmol/L)	(concentration)

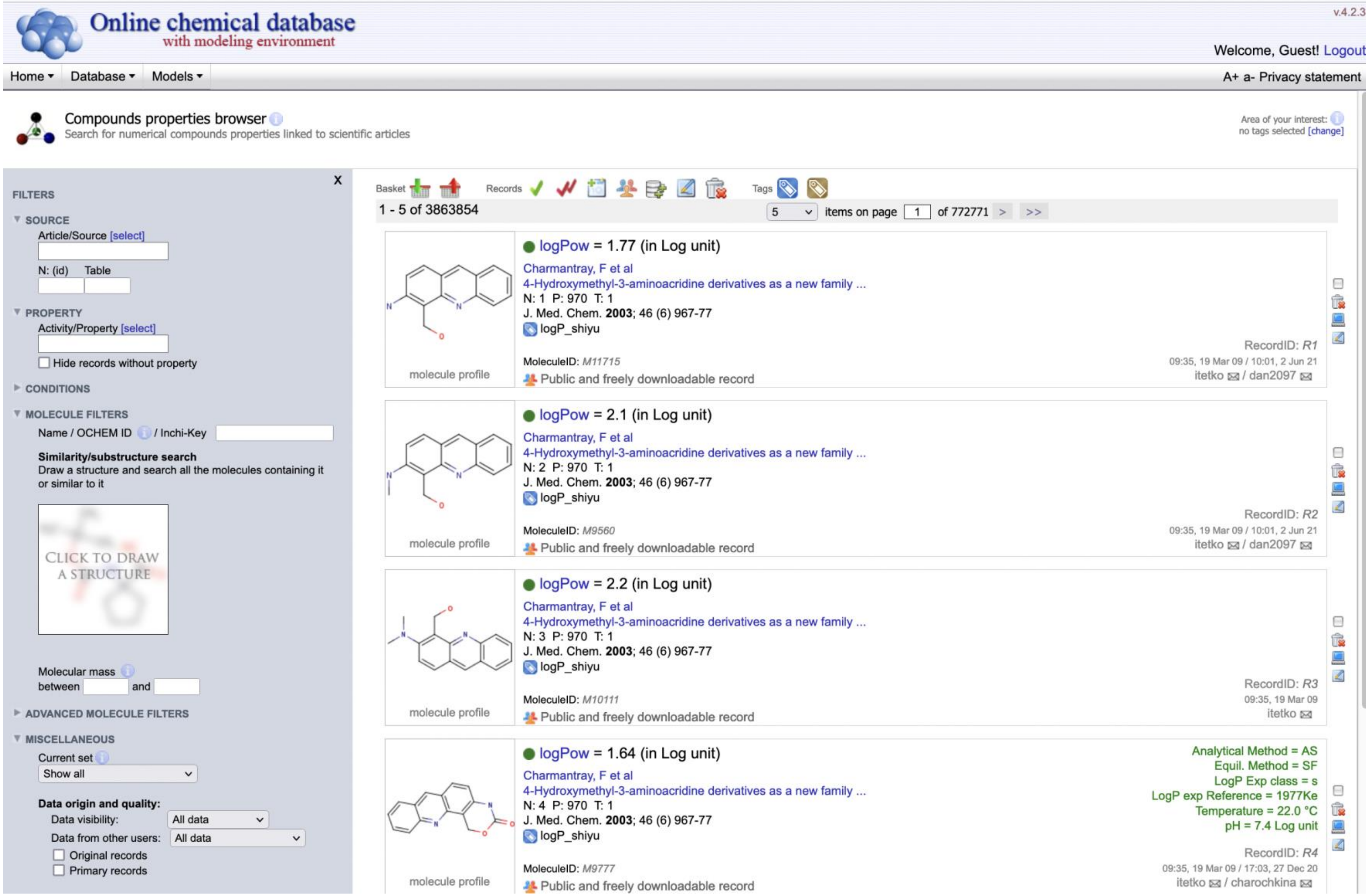
Articles

Spink, DC;Spink, BC;Zhuo, X;Hussain, MM;Gierthy, JF;Ding, X;
NADPH- and hydroperoxide-supported 17beta-estradiol hydroxylation catalyzed by a variant form (432L, 453S) of human cytochrome P450 1B1.
The Journal of steroid biochemistry and molecular biology **2000**; 74 (1-2) 11-8
PubMed - ArticleID: Q1352

Zhang, L;Zhu, H;Oprea, TI;Golbraikh, A;Tropsha, A;
QSAR modeling of the blood-brain barrier permeability for diverse organic compounds.
Pharmaceutical research **2008**; 25 (8) 1902-14
DOI - PubMed - ArticleID: Q1577

Zhu, H;Tropsha, A;Fourches, D;Varnek, A;Papa, E;Gramatica, P;Oberg, T;Dao, P;Cherkasov, A;Tetko, IV;
Combinatorial QSAR modeling of chemical toxicants tested against Tetrahymena pyriformis.
Journal of chemical information and modeling **2008**; 48 (4) 766-84
DOI - PubMed - PrePrint - ArticleID: Q1994

OCHEM database



Compounds properties browser

Search for numerical compounds properties linked to scientific articles

Area of your interest:
no tags selected [change]

FILTERS

▾ SOURCE

Article/Source [select]

N: (id) 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 visibility: All data

Data from other users: All data

□ Original records

□ Primary records

1 - 5 of 3863854 — 5 items on page 1 of 772771 > >>

● 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
logP_shiyu
MoleculeID: *M11715*
Public and freely downloadable record
RecordID: *R1*
09:35, 19 Mar 09 / 10:01, 2 Jun 21
iletko / dan2097
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
logP_shiyu
MoleculeID: *M9560*
Public and freely downloadable record
RecordID: *R2*
09:35, 19 Mar 09 / 10:01, 2 Jun 21
itetko / dan2097
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
logP_shiyu
MoleculeID: *M10111*
Public and freely downloadable record
RecordID: *R3*
09:35, 19 Mar 09
itetko
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
logP_shiyu
Analytical Method = AS
Equil. Method = SF
LogP Exp class = s
LogP exp Reference = 1977Ke
Temperature = 22.0 °C
pH = 7.4 Log unit
MoleculeID: *M9777*
Public and freely downloadable record
RecordID: *R4*
09:35, 19 Mar 09 / 17:03, 27 Dec 20
itetko / charochkina
molecule profile

Representations of Chemical Structures

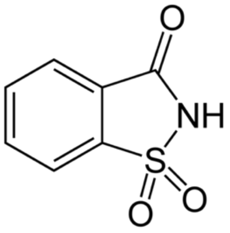
Saccharin

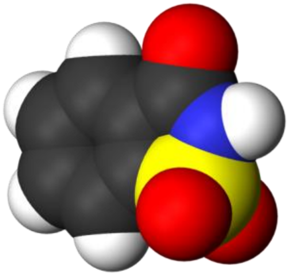
1D

2D

3D

$C_7H_5NO_3S$

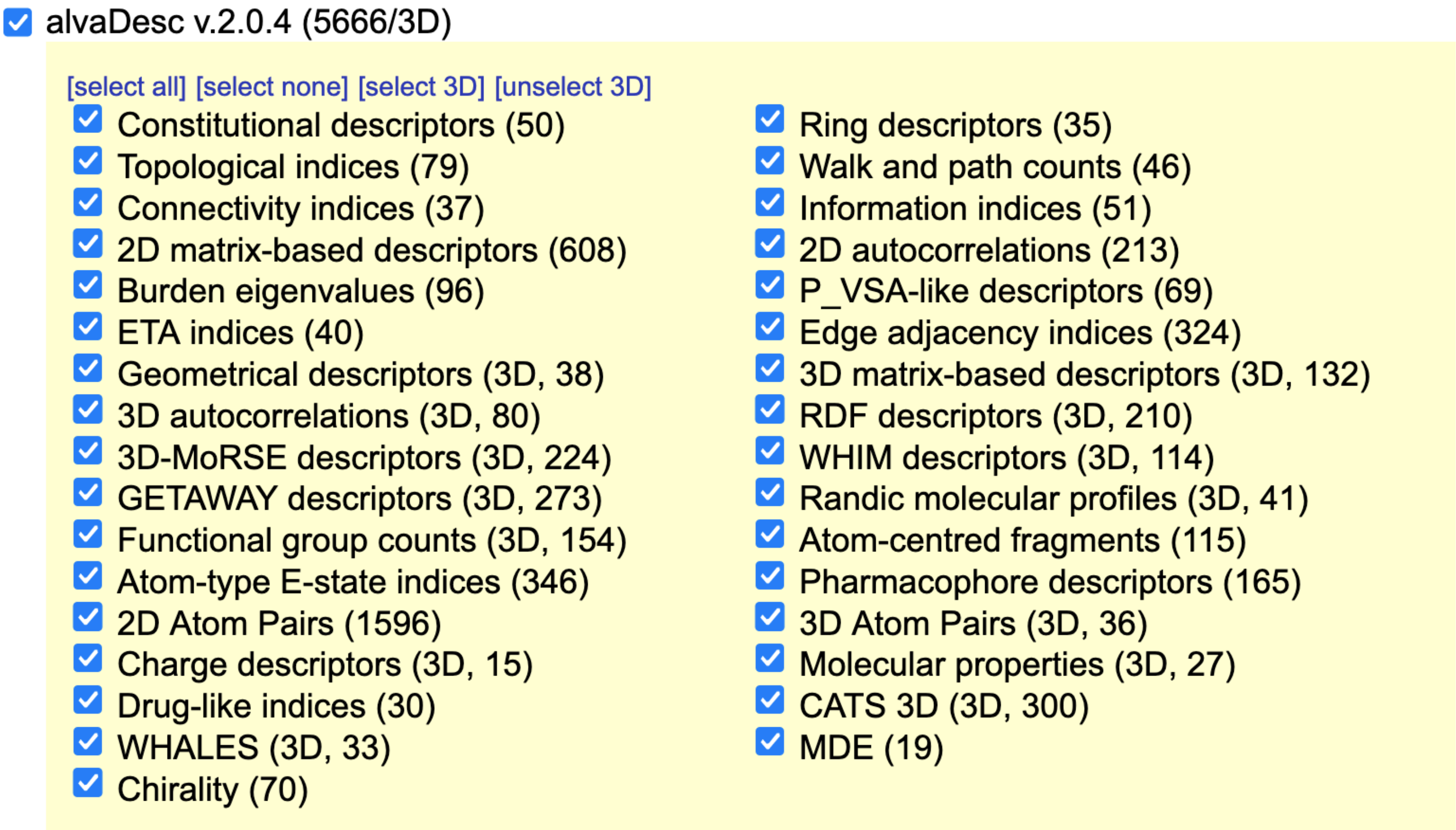




>20 descriptor packages, ToxAlerts

Representation learning: Smiles, Graphs, 3D …

Examples of descriptors



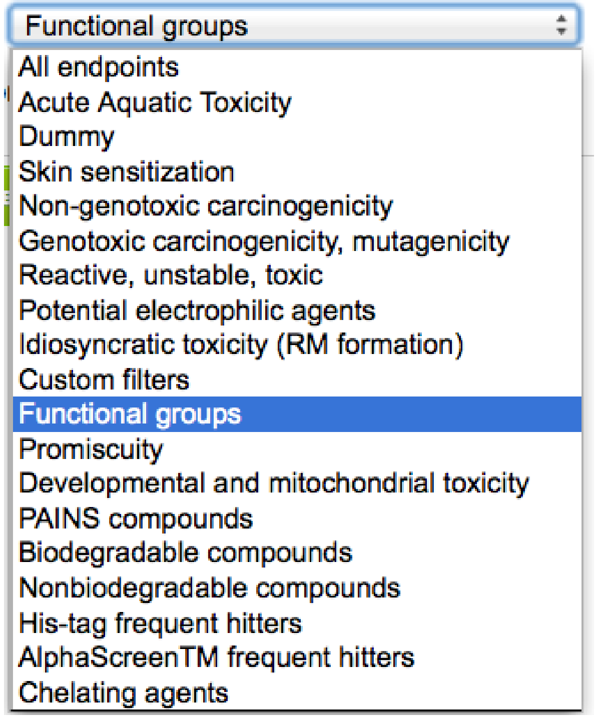
☑ alvaDesc v.2.0.4 (5666/3D)

[select all] [select none] [select 3D] [unselect 3D]

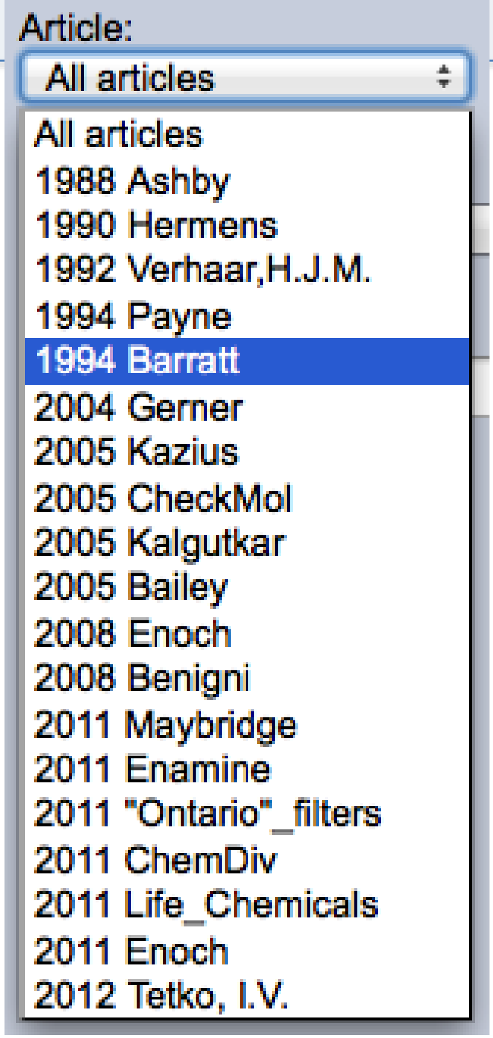
☑ Constitutional descriptors (50)
☑ Topological indices (79)
☑ Connectivity indices (37)
☑ 2D matrix-based descriptors (608)
☑ Burden eigenvalues (96)
☑ ETA indices (40)
☑ Geometrical descriptors (3D, 38)
☑ 3D autocorrelations (3D, 80)
☑ 3D-MoRSE descriptors (3D, 224)
☑ GETAWAY descriptors (3D, 273)
☑ Functional group counts (3D, 154)
☑ Atom-type E-state indices (346)
☑ 2D Atom Pairs (1596)
☑ Charge descriptors (3D, 15)
☑ Drug-like indices (30)
☑ WHALES (3D, 33)
☑ Chirality (70)
☑ Ring descriptors (35)
☑ Walk and path counts (46)
☑ Information indices (51)
☑ 2D autocorrelations (213)
☑ P_VSA-like descriptors (69)
☑ Edge adjacency indices (324)
☑ 3D matrix-based descriptors (3D, 132)
☑ RDF descriptors (3D, 210)
☑ WHIM descriptors (3D, 114)
☑ Randic molecular profiles (3D, 41)
☑ Atom-centred fragments (115)
☑ Pharmacophore descriptors (165)
☑ 3D Atom Pairs (3D, 36)
☑ Molecular properties (3D, 27)
☑ CATS 3D (3D, 300)
☑ MDE (19)

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

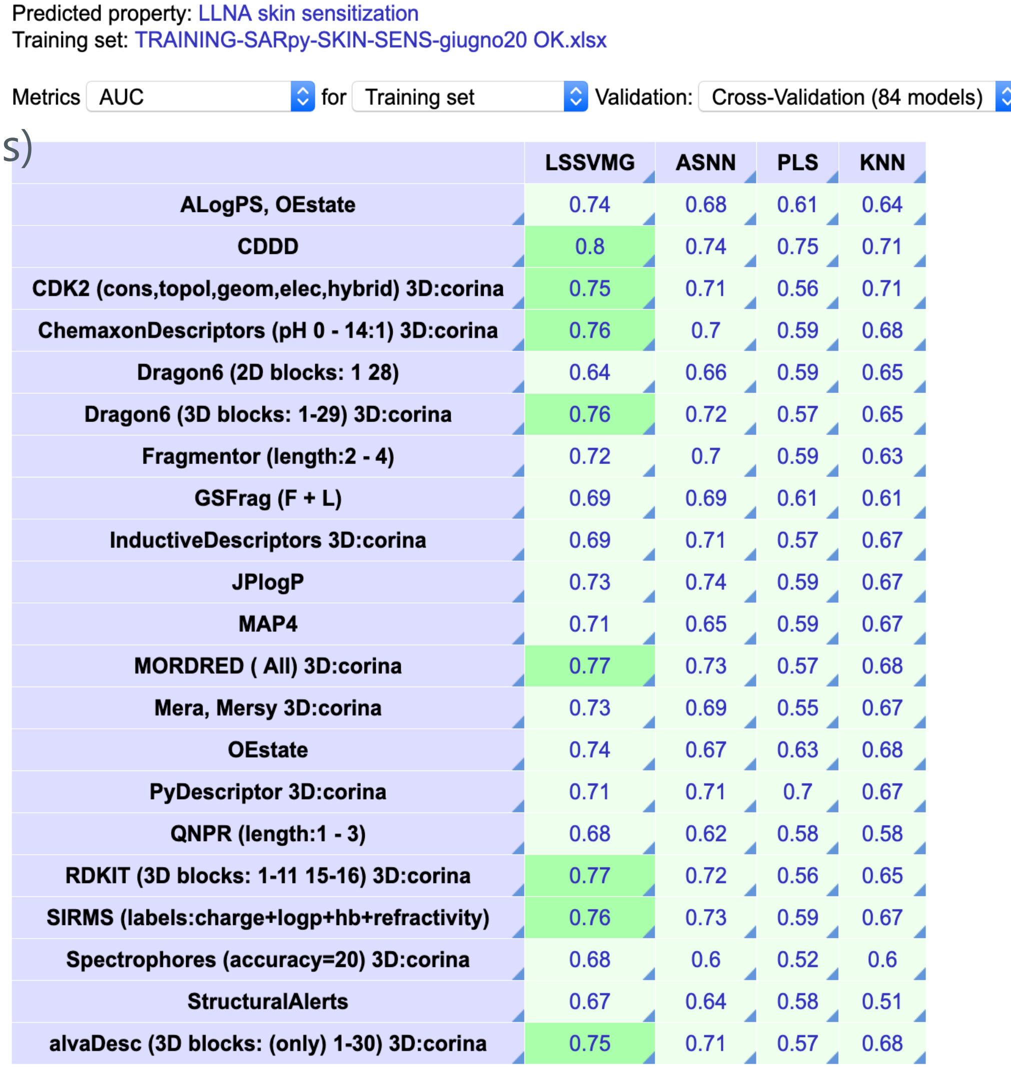
All endpoints
Acute Aquatic Toxicity
Dummy
Skin sensitization
Non-genotoxic carcinogenicity
Genotoxic carcinogenicity, mutagenicity
Reactive, unstable, toxic
Potential electrophilic agents
Idiosyncratic toxicity (RM formation)
Custom filters
Functional groups
Promiscuity
Developmental and mitochondrial toxicity
PAINS compounds
Biodegradable compounds
Nonbiodegradable compounds
His-tag frequent hitters
AlphaScreenTM frequent hitters
Chelating agents


Article: All articles

Sushko et al, JCIM, 2012, 52(8):2310-6. Salmina et al, Molecules, 2015, 21.

OCHEM modeling

- Comprehensive modeling
- Multitask learning (up to 100 properties)
- Feature net ("model in model")
- Consensus models
- GPU + CPU modern methods (~20)
- Supports models
  - >1,000,000 compounds
  - >200,000,000,000 descriptors*
  - >1,000 servers
  - up to 1GB in size (Java limit)
- Model private/publishing
- Export, import, web/REST services
- Conditions, external descriptors
- ToxAlerts screening



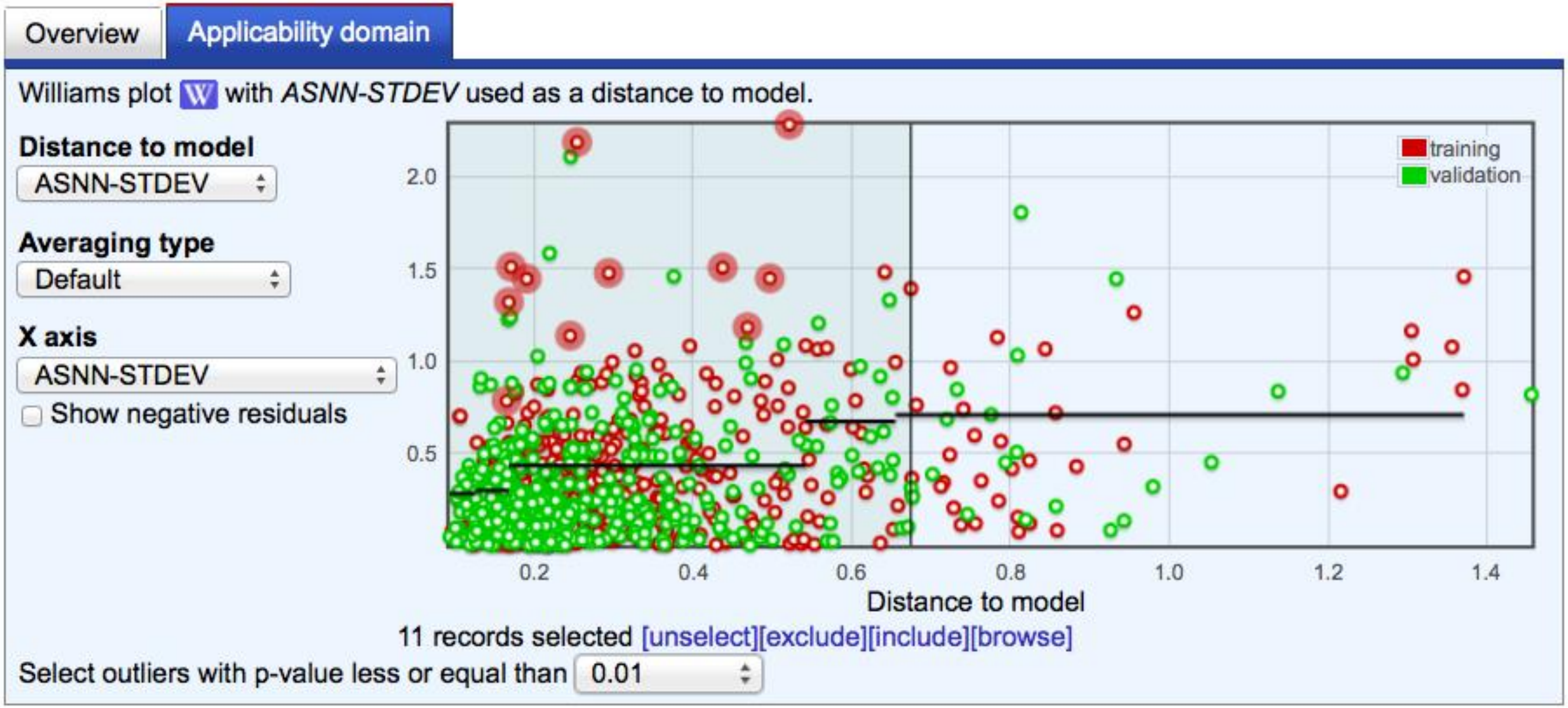
Predicted property: LLNA skin sensitization
Training set: TRAINING-SARpy-SKIN-SENS-giugno20 OK.xlsx

Metrics: AUC for: Training set Validation: Cross-Validation (84 models)

	LSSVMG	ASNN	PLS	KNN
ALogPS, OEstate	0.74	0.68	0.61	0.64
CDDD	0.8	0.74	0.75	0.71
CDK2 (cons,topol,geom,elec,hybrid) 3D:corina	0.75	0.71	0.56	0.71
ChemaxonDescriptors (pH 0 - 14:1) 3D:corina	0.76	0.7	0.59	0.68
Dragon6 (2D blocks: 1 28)	0.64	0.66	0.59	0.65
Dragon6 (3D blocks: 1-29) 3D:corina	0.76	0.72	0.57	0.65
Fragmentor (length:2 - 4)	0.72	0.7	0.59	0.63
GSFrag (F + L)	0.69	0.69	0.61	0.61
InductiveDescriptors 3D:corina	0.69	0.71	0.57	0.67
JPlogP	0.73	0.74	0.59	0.67
MAP4	0.71	0.65	0.59	0.67
MORDRED (All) 3D:corina	0.77	0.73	0.57	0.68
Mera, Mersy 3D:corina	0.73	0.69	0.55	0.67
OEstate	0.74	0.67	0.63	0.68
PyDescriptor 3D:corina	0.71	0.71	0.7	0.67
QNPR (length:1 - 3)	0.68	0.62	0.58	0.58
RDKIT (3D blocks: 1-11 15-16) 3D:corina	0.77	0.72	0.56	0.65
SIRMS (labels:charge+logp+hb+refractivity)	0.76	0.73	0.59	0.67
Spectrophores (accuracy=20) 3D:corina	0.68	0.6	0.52	0.6
StructuralAlerts	0.67	0.64	0.58	0.51
alvaDesc (3D blocks: (only) 1-30) 3D:corina	0.75	0.71	0.57	0.68

* Sparse format, DOI:10.1186/s13321-016-0113-y

Applicability domain assessment (regression)



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

Distance to model
ASNN-STDEV

Averaging type
Default

X axis
ASNN-STDEV

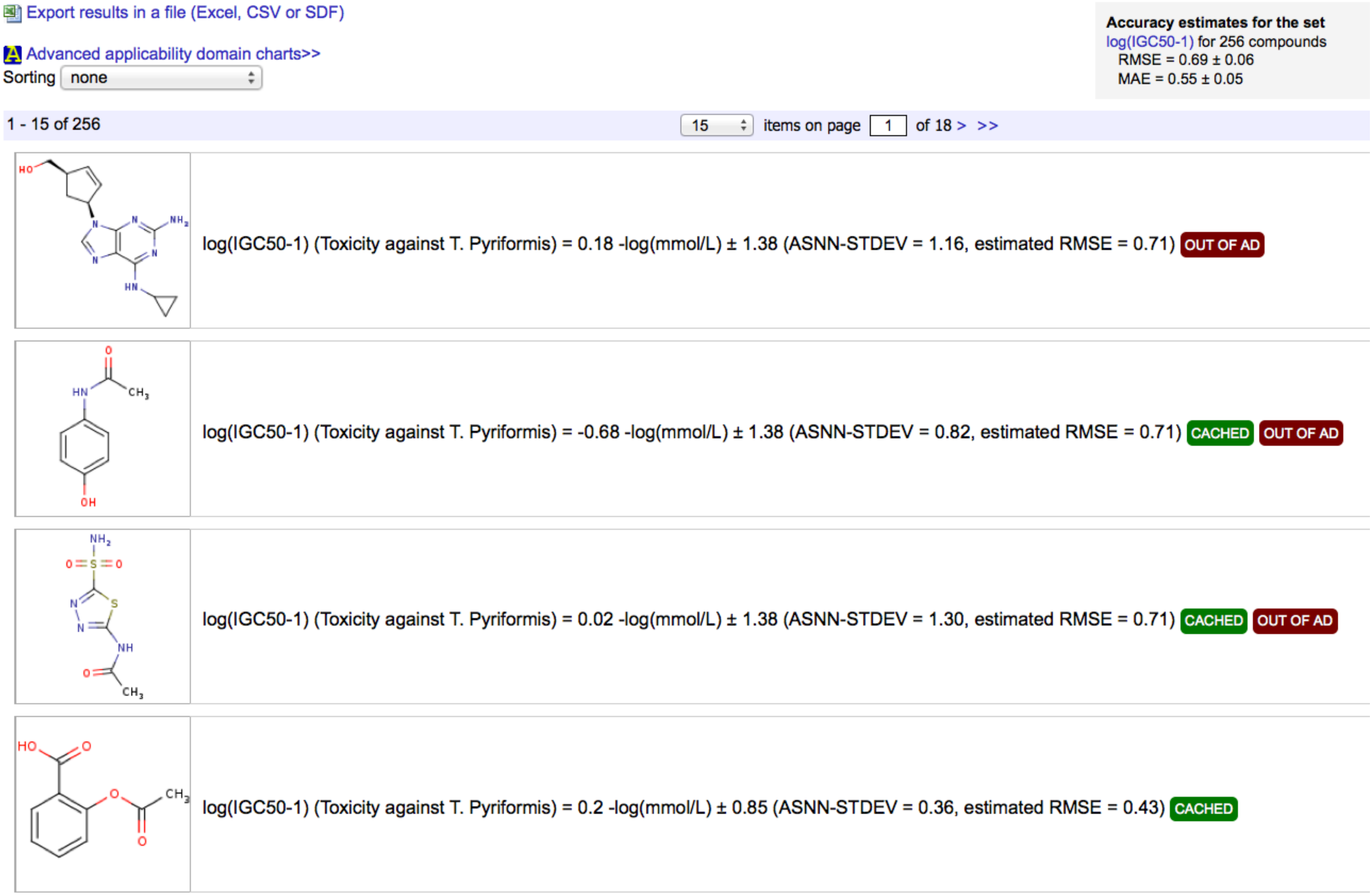
□ Show negative residuals

11 records selected [unselect][exclude][include][browse]

Select outliers with p-value less or equal than 0.01

- 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 none

Accuracy estimates for the set
log(IGC50-1) for 256 compounds
RMSE = 0.69 ± 0.06
MAE = 0.55 ± 0.05

1 - 15 of 256 15 items on page 1 of 18 > >>

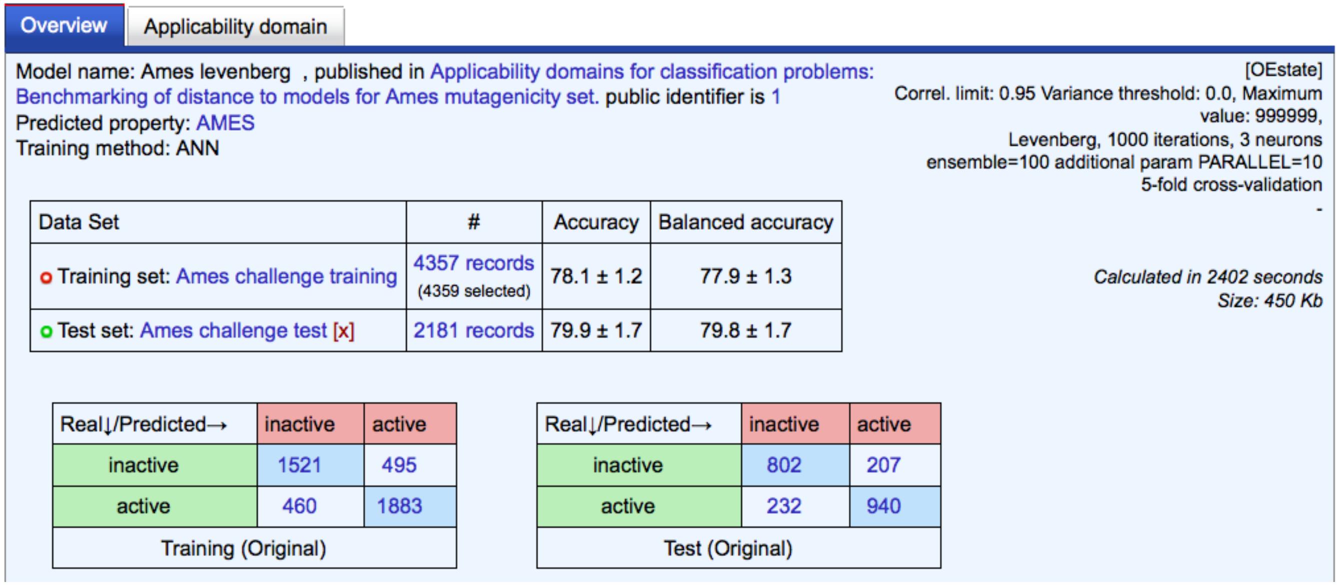
log(IGC50-1) (Toxicity against T. Pyriformis) = 0.18 -log(mmol/L) ± 1.38 (ASNN-STDEV = 1.16, estimated RMSE = 0.71) OUT OF AD

log(IGC50-1) (Toxicity against T. Pyriformis) = -0.68 -log(mmol/L) ± 1.38 (ASNN-STDEV = 0.82, estimated RMSE = 0.71) CACHED OUT OF AD

log(IGC50-1) (Toxicity against T. Pyriformis) = 0.02 -log(mmol/L) ± 1.38 (ASNN-STDEV = 1.30, estimated RMSE = 0.71) CACHED OUT OF AD

log(IGC50-1) (Toxicity against T. Pyriformis) = 0.2 -log(mmol/L) ± 0.85 (ASNN-STDEV = 0.36, estimated RMSE = 0.43) CACHED

Accuracy of predictions for classification model



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

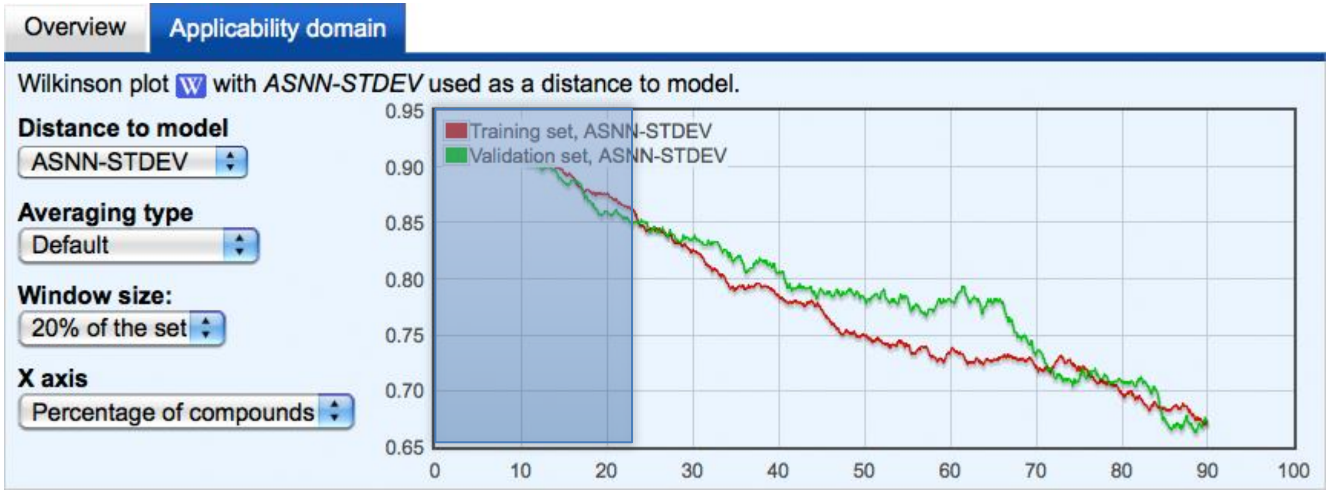
[OEstate]
Correl. limit: 0.95 Variance threshold: 0.0, Maximum value: 999999,
Levenberg, 1000 iterations, 3 neurons
ensemble=100 additional param PARALLEL=10
5-fold cross-validation
-

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

Calculated in 2402 seconds
Size: 450 Kb

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

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



Wilkinson plot W 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

Example of OCHEM models for REACH endpoints

7.2 Melting point

7.3 Boiling point

7.7 Water solubility

7.8 LogP

8.4.1 AMES test

9.2.1.1 Ready biodegradability

9.3.2 BCF

7.5 Vapor pressure

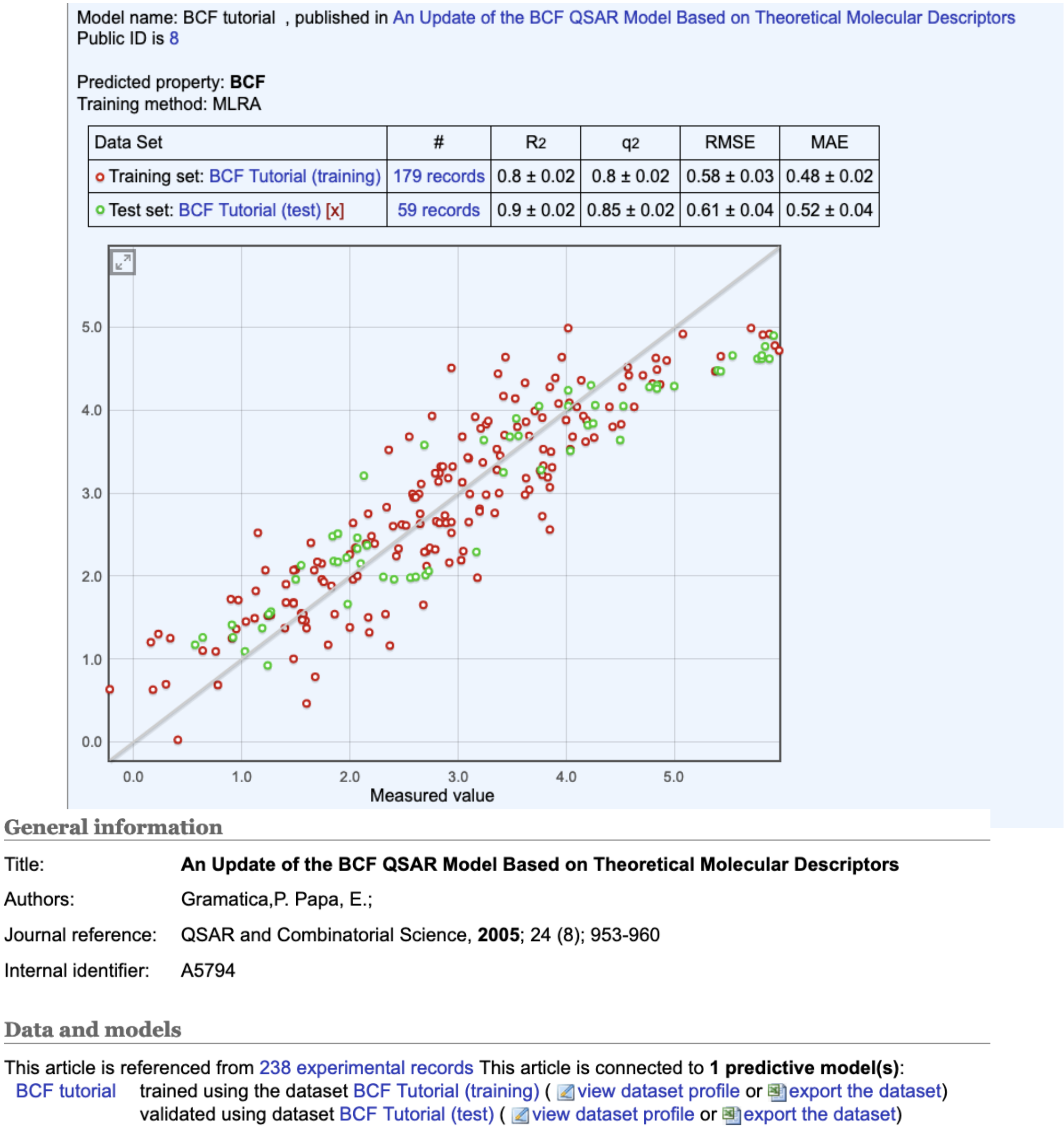
8.1 Skin irritation

8.2 Eye irritation

8.3 Skin sensitization

- CERAPP: Estrogen receptors
- COMPARA: Androgen receptor
- CATMoS: Acute toxicity

Example of an external model integration



Model name: BCF tutorial , published in An Update of the BCF QSAR Model Based on Theoretical Molecular Descriptors
Public ID is 8

Predicted property: **BCF**
Training method: MLRA

Data Set	#	R2	q2	RMSE	MAE
○ Training set: BCF Tutorial (training)	179 records	0.8 ± 0.02	0.8 ± 0.02	0.58 ± 0.03	0.48 ± 0.02
○ Test set: BCF Tutorial (test) [x]	59 records	0.9 ± 0.02	0.85 ± 0.02	0.61 ± 0.04	0.52 ± 0.04

General information

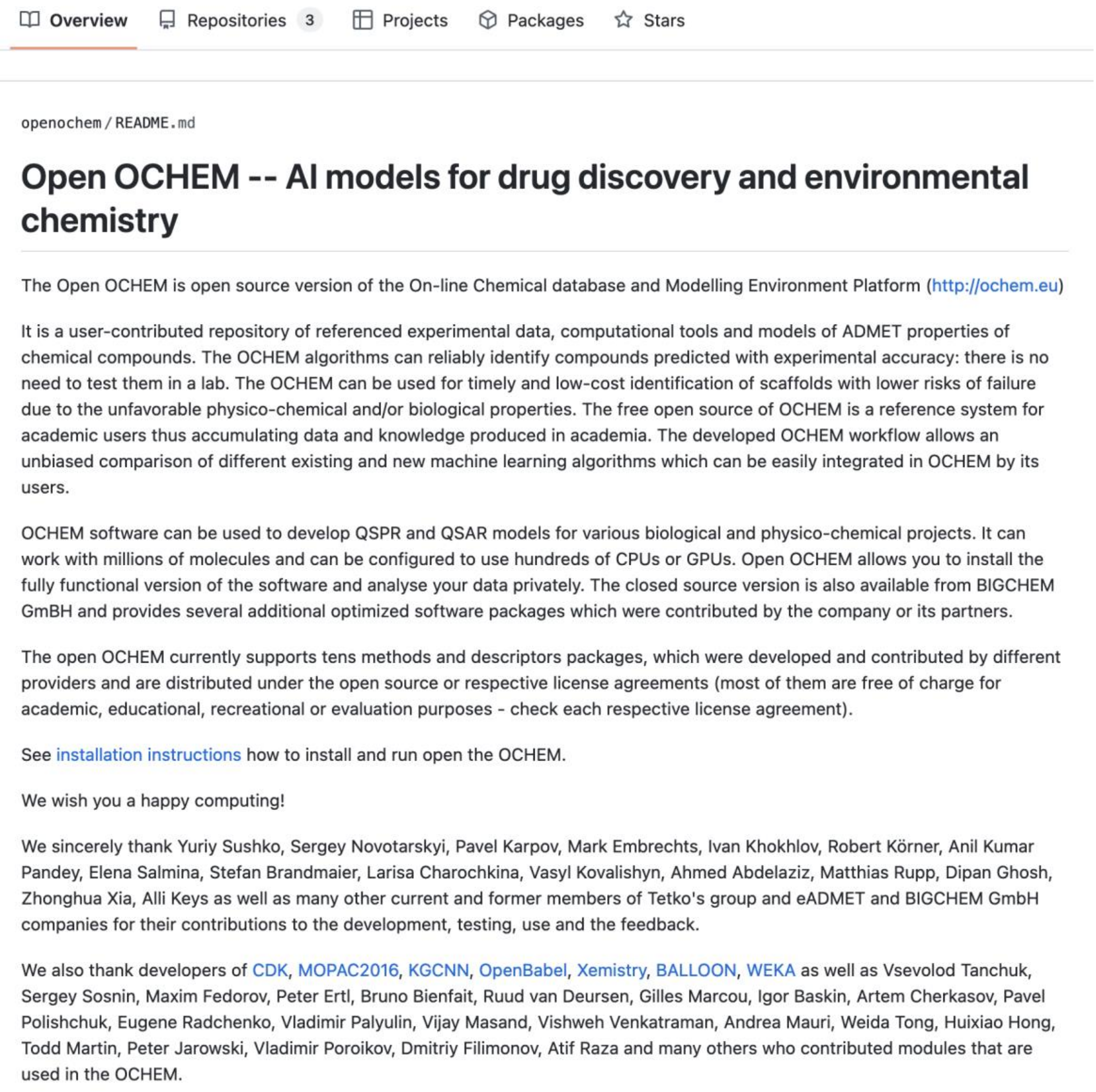
Title:	**An Update of the BCF QSAR Model Based on Theoretical Molecular Descriptors**
Authors:	Gramatica,P. Papa, E.;
Journal reference:	QSAR and Combinatorial Science, **2005**; 24 (8); 953-960
Internal identifier:	A5794

Data and models

This article is referenced from 238 experimental records This article is connected to **1 predictive model(s)**:

BCF tutorial trained using the dataset BCF Tutorial (training) (view dataset profile or export the dataset)
validated using dataset BCF Tutorial (test) (view dataset profile or export the dataset)

openOCHEM https://github.com/openochem



openochem / README.md

Open OCHEM -- AI models for drug discovery and environmental chemistry

The Open OCHEM is open source version of the On-line Chemical database and Modelling Environment Platform (http://ochem.eu)

It is a user-contributed repository of referenced experimental data, computational tools and models of ADMET properties of chemical compounds. The OCHEM algorithms can reliably identify compounds predicted with experimental accuracy: there is no need to test them in a lab. The OCHEM can be used for timely and low-cost identification of scaffolds with lower risks of failure due to the unfavorable physico-chemical and/or biological properties. The free open source of OCHEM is a reference system for academic users thus accumulating data and knowledge produced in academia. The developed OCHEM workflow allows an unbiased comparison of different existing and new machine learning algorithms which can be easily integrated in OCHEM by its users.

OCHEM software can be used to develop QSPR and QSAR models for various biological and physico-chemical projects. It can work with millions of molecules and can be configured to use hundreds of CPUs or GPUs. Open OCHEM allows you to install the fully functional version of the software and analyse your data privately. The closed source version is also available from BIGCHEM GmBH and provides several additional optimized software packages which were contributed by the company or its partners.

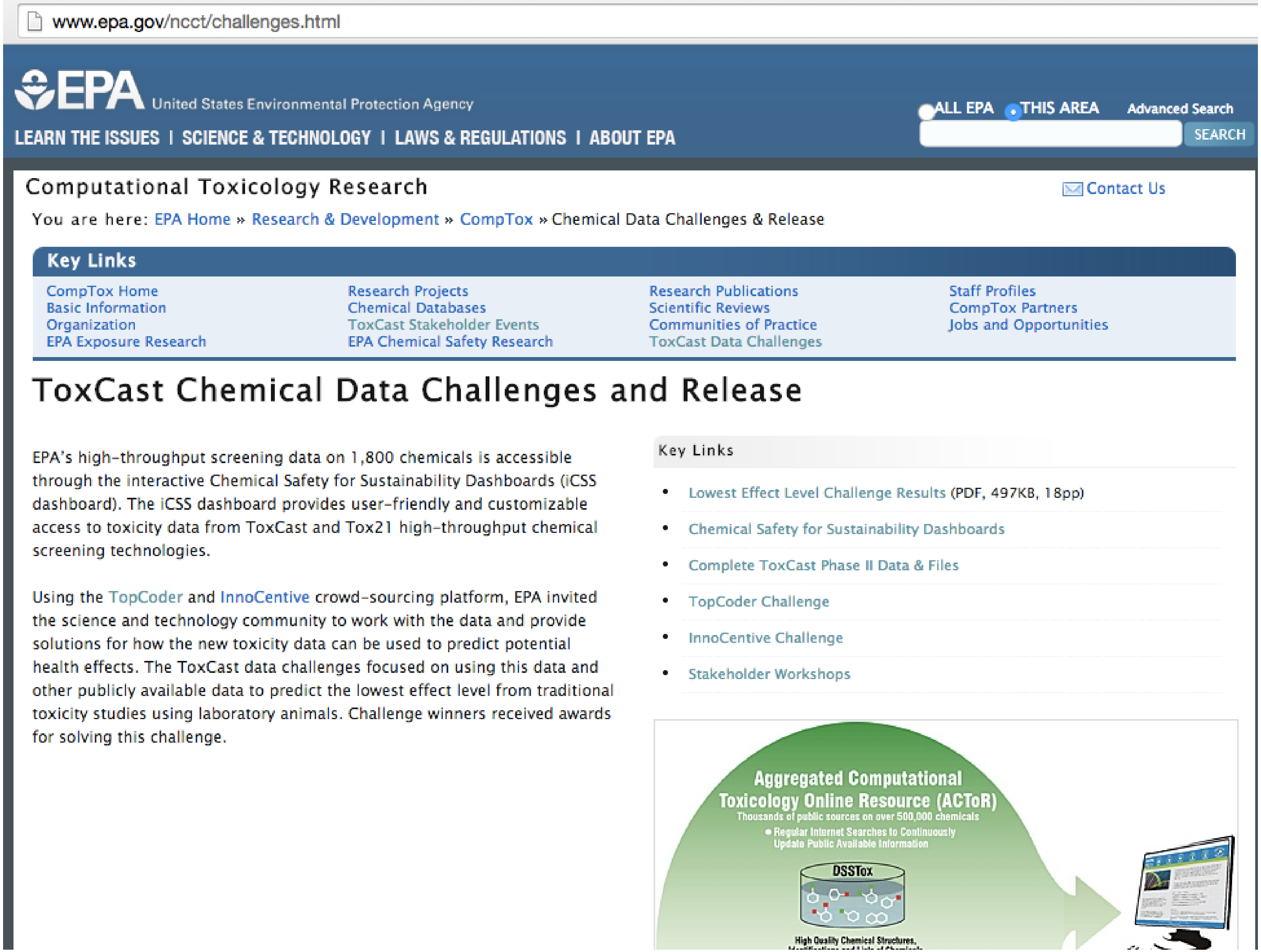
The open OCHEM currently supports tens methods and descriptors packages, which were developed and contributed by different providers and are distributed under the open source or respective license agreements (most of them are free of charge for academic, educational, recreational or evaluation purposes - check each respective license agreement).

See installation instructions how to install and run open the OCHEM.

We wish you a happy computing!

We sincerely thank Yuriy Sushko, Sergey Novotarskyi, Pavel Karpov, Mark Embrechts, Ivan Khokhlov, Robert Körner, Anil Kumar Pandey, Elena Salmina, Stefan Brandmaier, Larisa Charochkina, Vasyl Kovalishyn, Ahmed Abdelaziz, Matthias Rupp, Dipan Ghosh, Zhonghua Xia, Alli Keys as well as many other current and former members of Tetko's group and eADMET and BIGCHEM GmbH companies for their contributions to the development, testing, use and the feedback.

We also thank developers of CDK, MOPAC2016, KGCNN, OpenBabel, Xemistry, BALLOON, WEKA as well as Vsevolod Tanchuk, Sergey Sosnin, Maxim Fedorov, Peter Ertl, Bruno Bienfait, Ruud van Deursen, Gilles Marcou, Igor Baskin, Artem Cherkasov, Pavel Polishchuk, Eugene Radchenko, Vladimir Palyulin, Vijay Masand, Vishweh Venkatraman, Andrea Mauri, Weida Tong, Huixiao Hong, Todd Martin, Peter Jarowski, Vladimir Poroikov, Dmitriy Filimonov, Atif Raza and many others who contributed modules that are used in the OCHEM.


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
Contact Us

Computational Toxicology Research

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

Key Links

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

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,

Novotarskyi, S. et al. *Chem. Res. Toxicol*. 2016, 29, 768-75.

slas

Come Transform Research

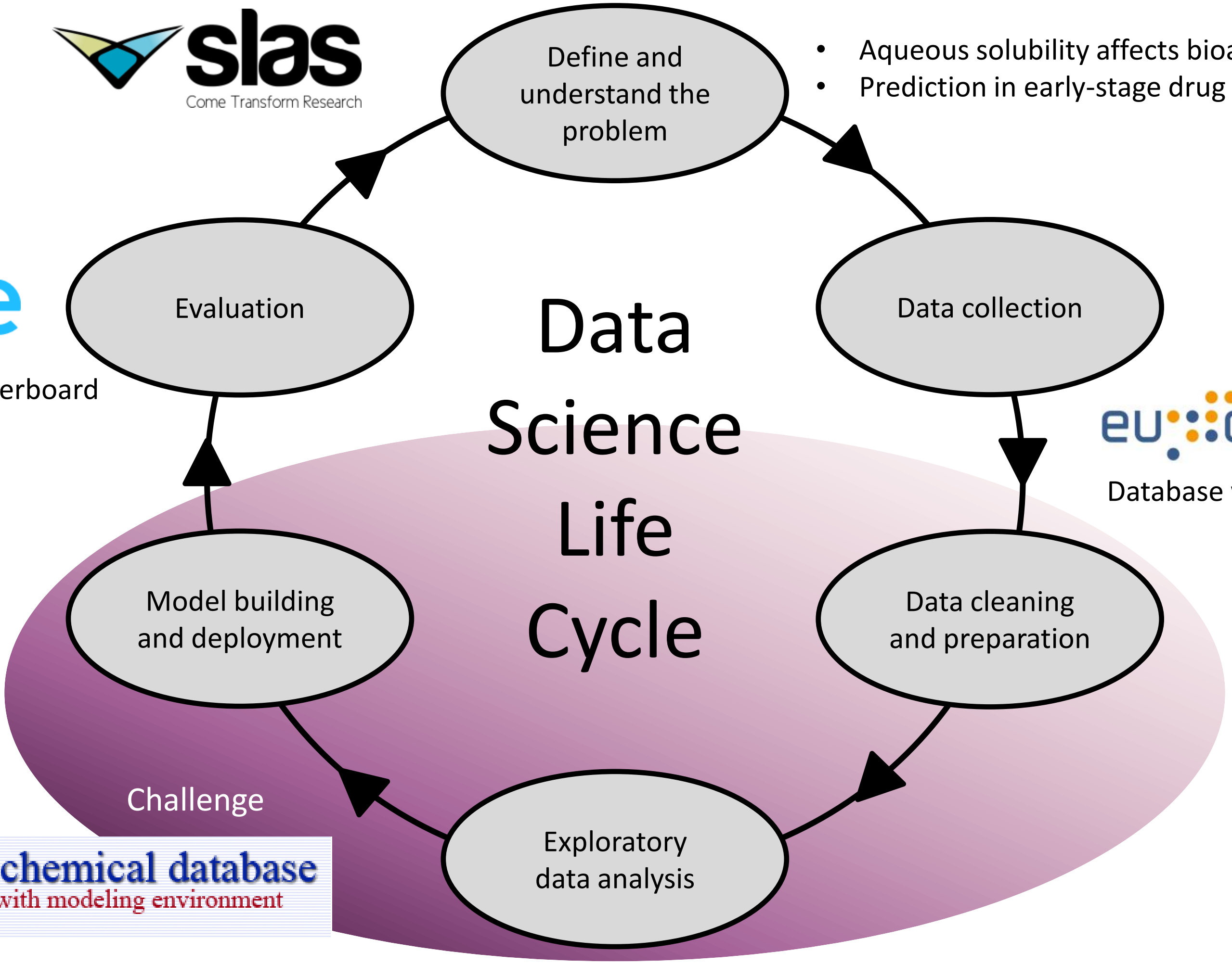
- Aqueous solubility affects bioavailability/ bioactivity
- Prediction in early-stage drug development

kaggle

Public & private leaderboard

eu::openscreen

Database with 100k compounds


Define and understand the problem
Data collection
Data cleaning and preparation
Exploratory data analysis
Model building and deployment
Evaluation
Data Science Life Cycle
Challenge

Online chemical database

with modeling environment

A. Kopp et al, *SLAS Discovery*, 2023, in press

Acknowledgements

Andi Kopp
Peter Hartog
Fabian Krüger
Paula Torren-Peraire
Varvara Voinarovska
Katya Ahmad
Marchela Pandelova
Nesma Mousa
Mark Embrechts

Emilio Benfenati
and all colleagues from
CONCERT REACH

Guillaume Godin (Firmenich)
Ruud van Deursen (Firmenich)

Michael Sattler (HMGU)

Alexander von Humboldt
Stiftung/Foundation

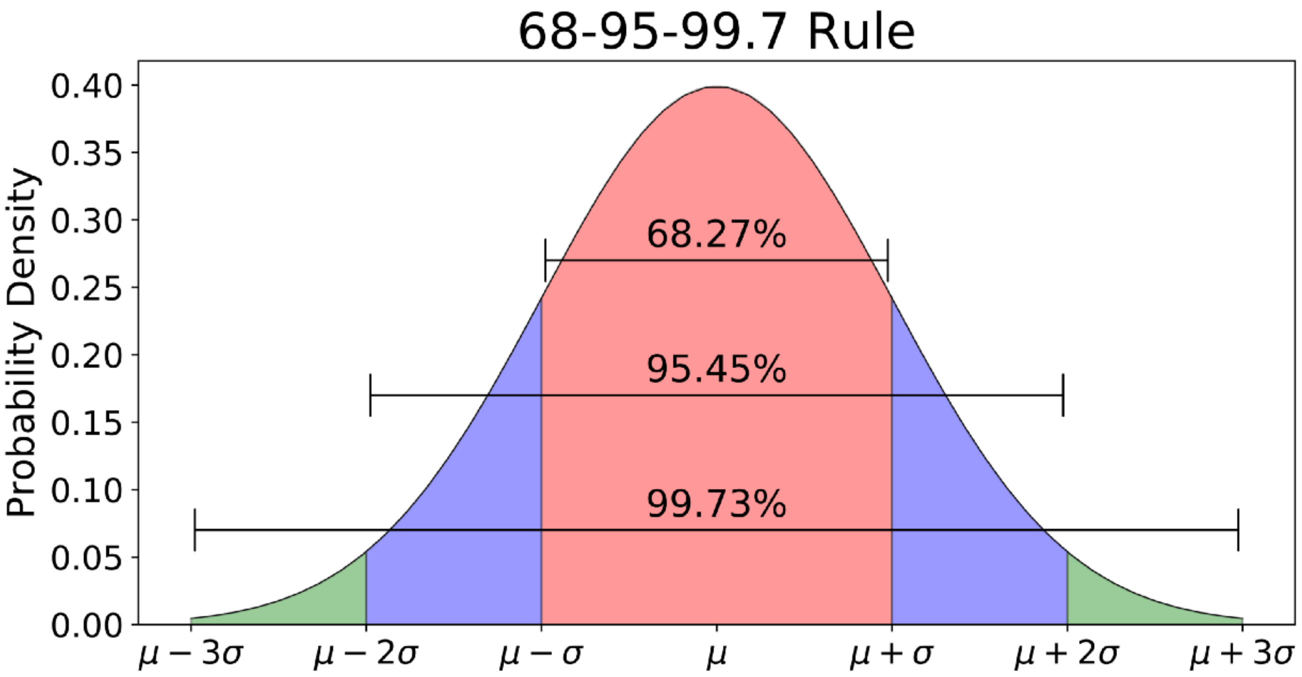
CONCERTREACH
CONCERTING EXPERIMENTAL DATA
AND IN SILICO MODELS FOR REACH

eco
environmental chemoinformatics
marie curie initial training network

MEGA GRANTS

BigChem

Gaussian distribution and outliers detection


68-95-99.7 Rule
Probability Density
$\mu - 3\sigma$
$\mu - 2\sigma$
$\mu - \sigma$
μ
$\mu + \sigma$
$\mu + 2\sigma$
$\mu + 3\sigma$