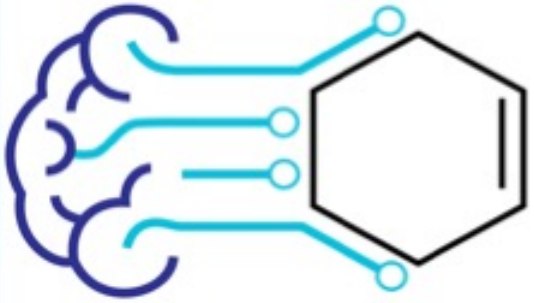




Development of QSAR/QSPR models using representation learning and descriptor based methods using openOCHEM. Which ones are better?

Igor V. Tetko

Helmholtz Munich and BIGCHEM GmbH

July 24, 2023, OpenTox Summer School 2023

HELMHOLTZ MUNICH

Agenda

Overview of OCHEM

Need for open-source publishing platforms

OCHEM data structure

OCHEM methods

Which method should I use?

Representation learning

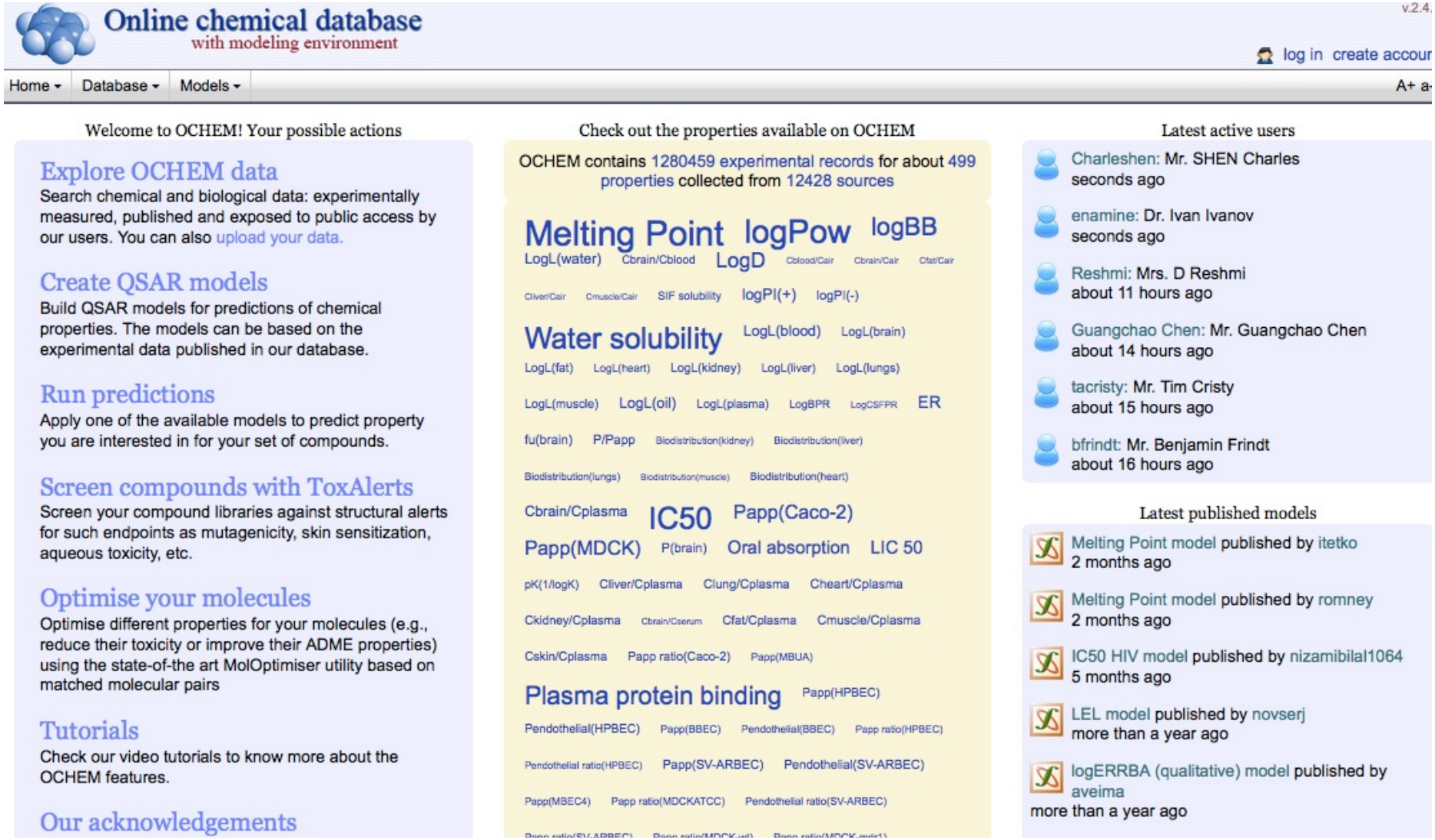
Analysis of SLAS challenge results

Applicability domain & detection of outliers

Multitask learning

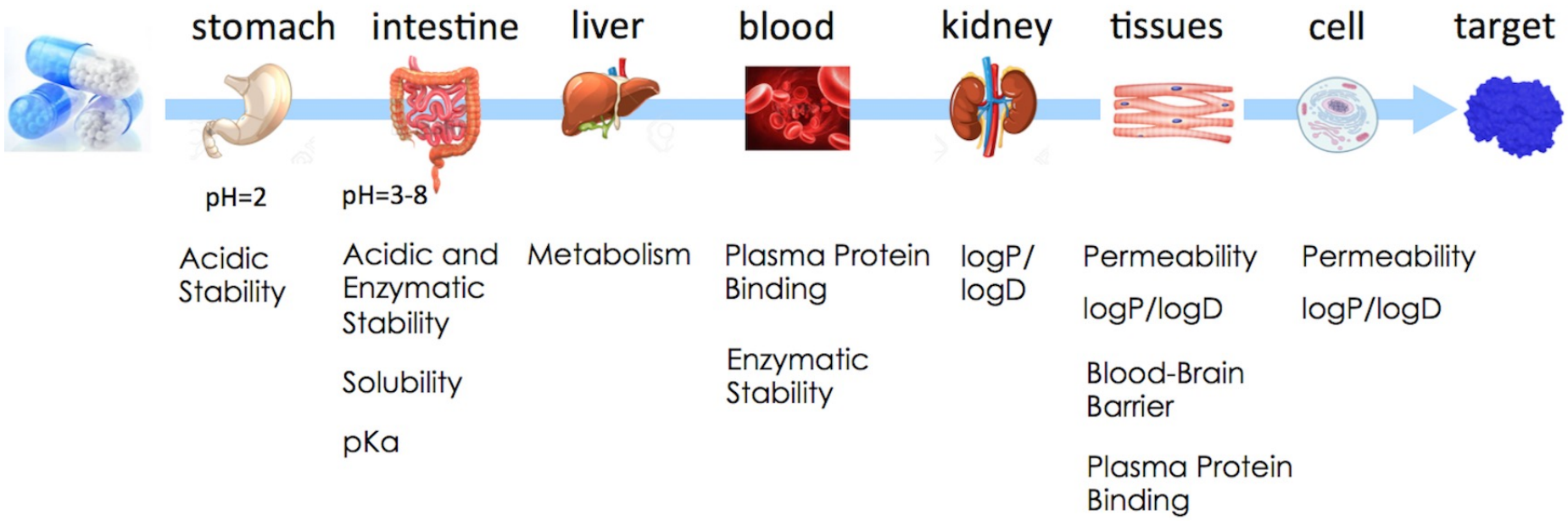
Tasks for hackathon

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



BIGCHEM GmbH is a spin-off of the center

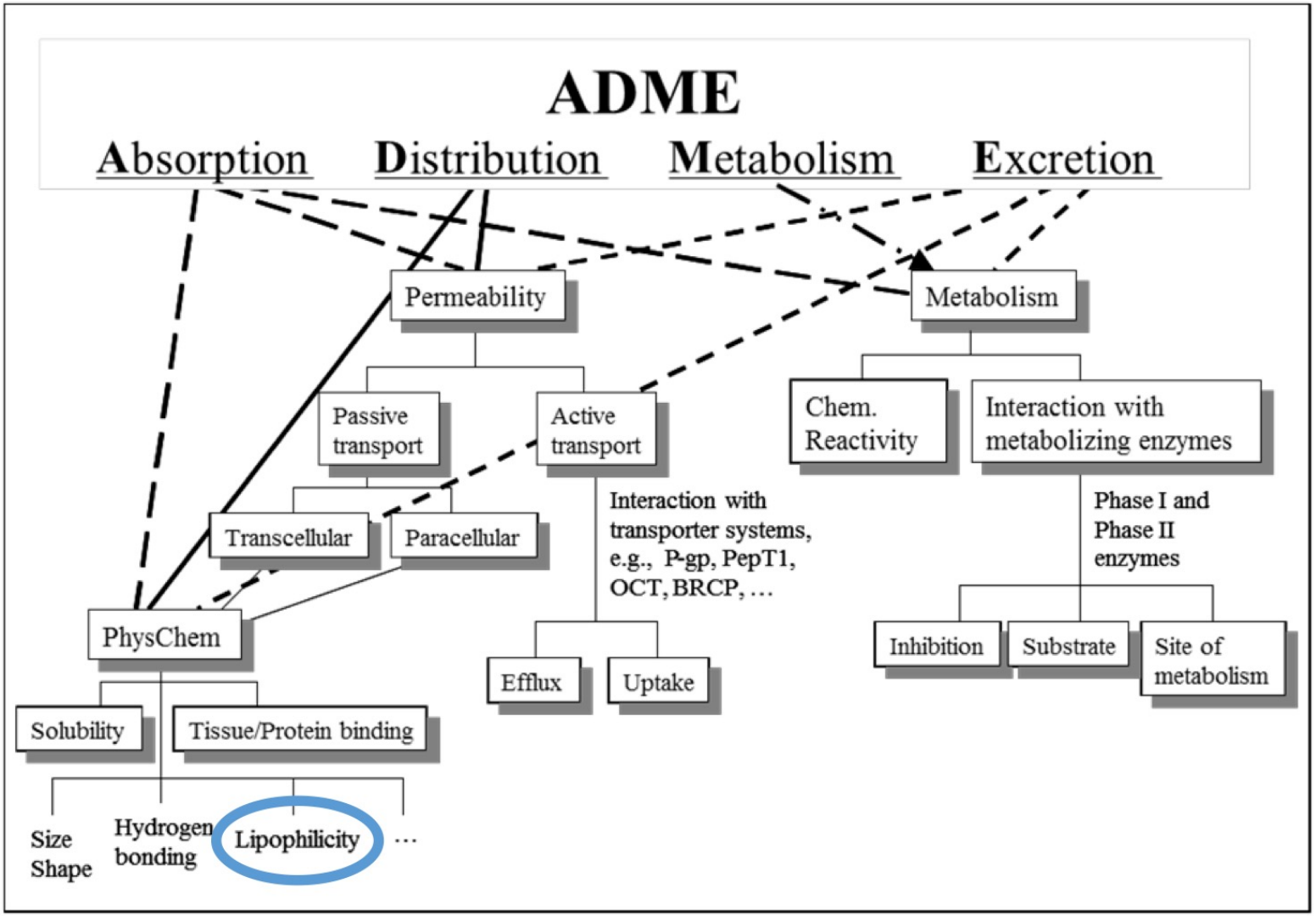
Physiological and physical–chemical barriers affecting a drug bioavailability.



Adapted from Kerns, E. H.; Di, L. Pharmaceutical Profiling in Drug Discovery. Drug Discov. Today 2003, 8, 316–323. Copyright (2003), with permission from Elsevier.

See also Ratkova, E. L. et al Empirical and Physics-Based Calculations of Physical–Chemical Properties. In *Comprehensive Medicinal Chemistry III*, Chackalamannil, S.; Rotella, D. P.; Ward, S., Eds.; Elsevier: Oxford, 2017; Vol. 3, pp 393-428.

Absorption Distribution Metabolism Excretion



see S.Winiwarter et al. *Use of Molecular Descriptors for ADME Predictions.* Compr. Med. Chem. II, D.J.Triggle & J.B.Taylor, Eds., Vol. 5, Elsevier, 531-554 (2007)

ADMETox filters in Bayer

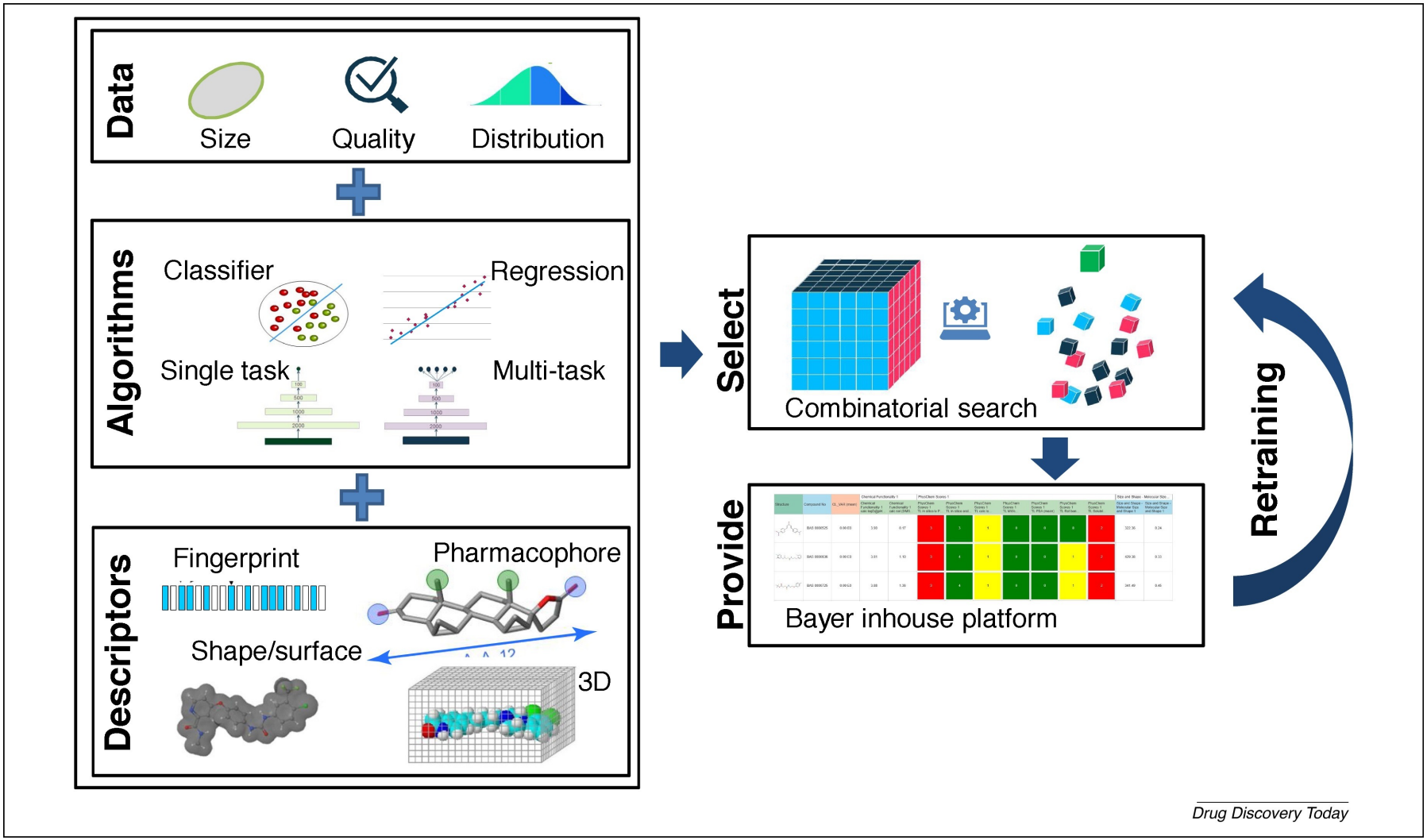
Insufficient quality | First approach | Medium model | Good model | Robust model

Endpoint		Model type	Data set size	2005	2009	2014	2019	Retraining
Absorption	Caco-2 permeation	C (N)	>10 000			RF	SVR	Weekly
	Caco-2 efflux	C (N)	>10 000			RF	SVR	Weekly
	Bioavailability (rat)	C	~2000				RF	On demand
Distribution	Human serum albumin	N	>30 000			PLS	MTNN	On demand
	Fraction unbound	N	>1000			PLS	MTNN	On demand
Metabolism	Microsomal stability (hum)	C (N)	>10 000			RF	RF	Weekly
	Microsomal stability (mouse)	C (N)	>10 000			RF	RF	Weekly
	Microsomal stability (rat)	C (N)	>10 000			RF	RF	Weekly
	Hepatocyte stability (rat)	C (N)	>30 000			RF	RF	Weekly
Toxicity	hERG inhibition	C	>10 000			RF	SVM	Weekly
	Ames mutagenicity	C	>10 000			RF	RF	On demand
	CYP inhibition isoforms	C	>10 000			RF	RF	On demand
	Phospholipidosis	C	<1000			SVM	SVM	On demand
	Structure filter tool	Score	n.a.	-	-	-	-	On demand
PhysChem	Solubility (DMSO)	N	>30 ,000			PLS	MTNN	On demand
	Solubility (Powder)	N	<10 000				MTNN	On demand
	logD @ pH 7.5	N	>70 000			PLS	MTNN	On demand
	Membrane affinity	N	<10 000			PLS	MTNN	On demand
	pKa	N	>10 000			ANN	ANN	On demand
	Oral PhysChem score	Score	n.a.	-	-	-	-	On demand
	i.v. PhysChem score	Score	n.a.	-	-	-	-	On demand

Drug Discovery Today

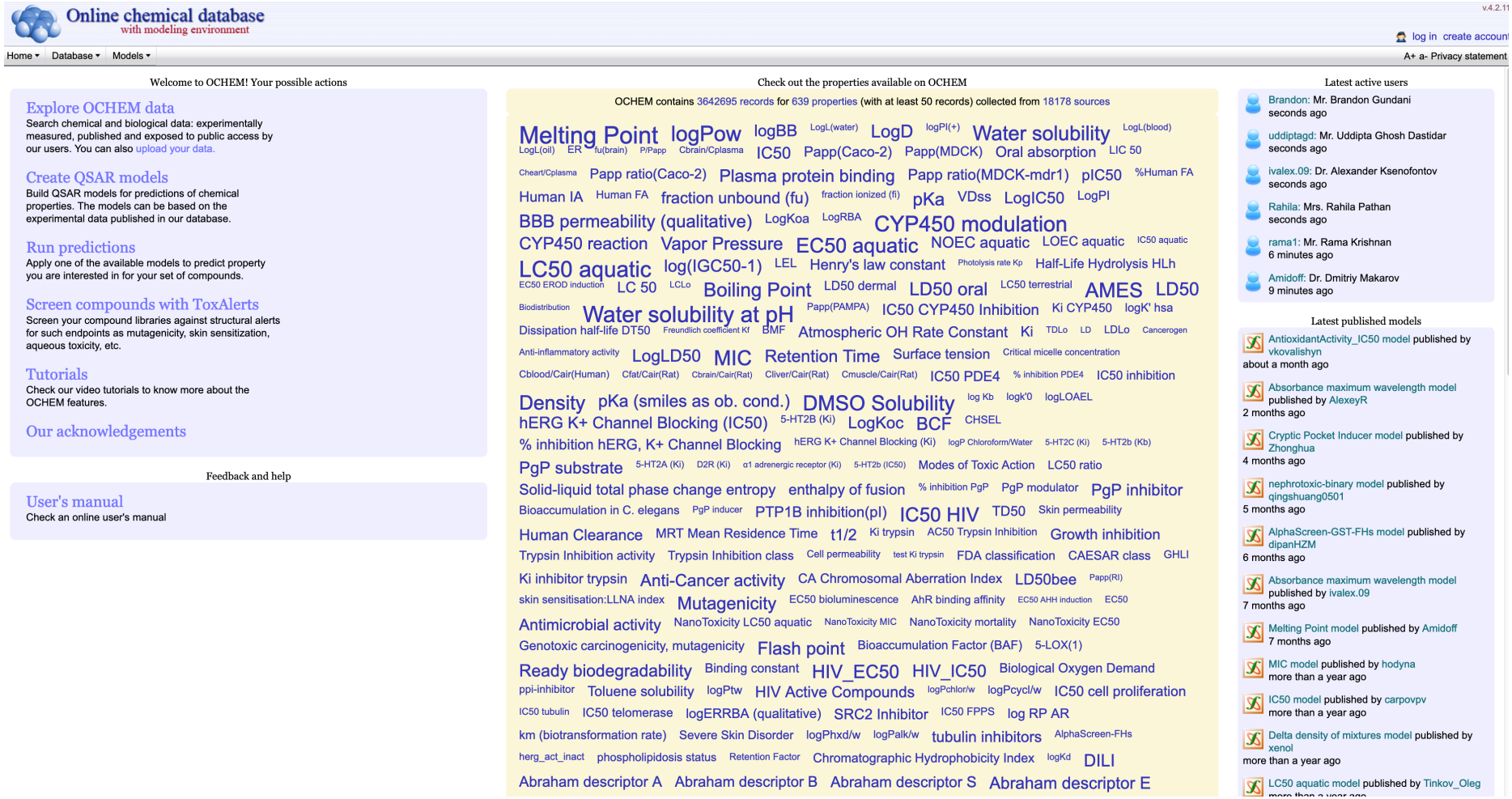
Göller, AH et al *Drug Discov. Today* **2020,** *25* (9), 1702-1709.

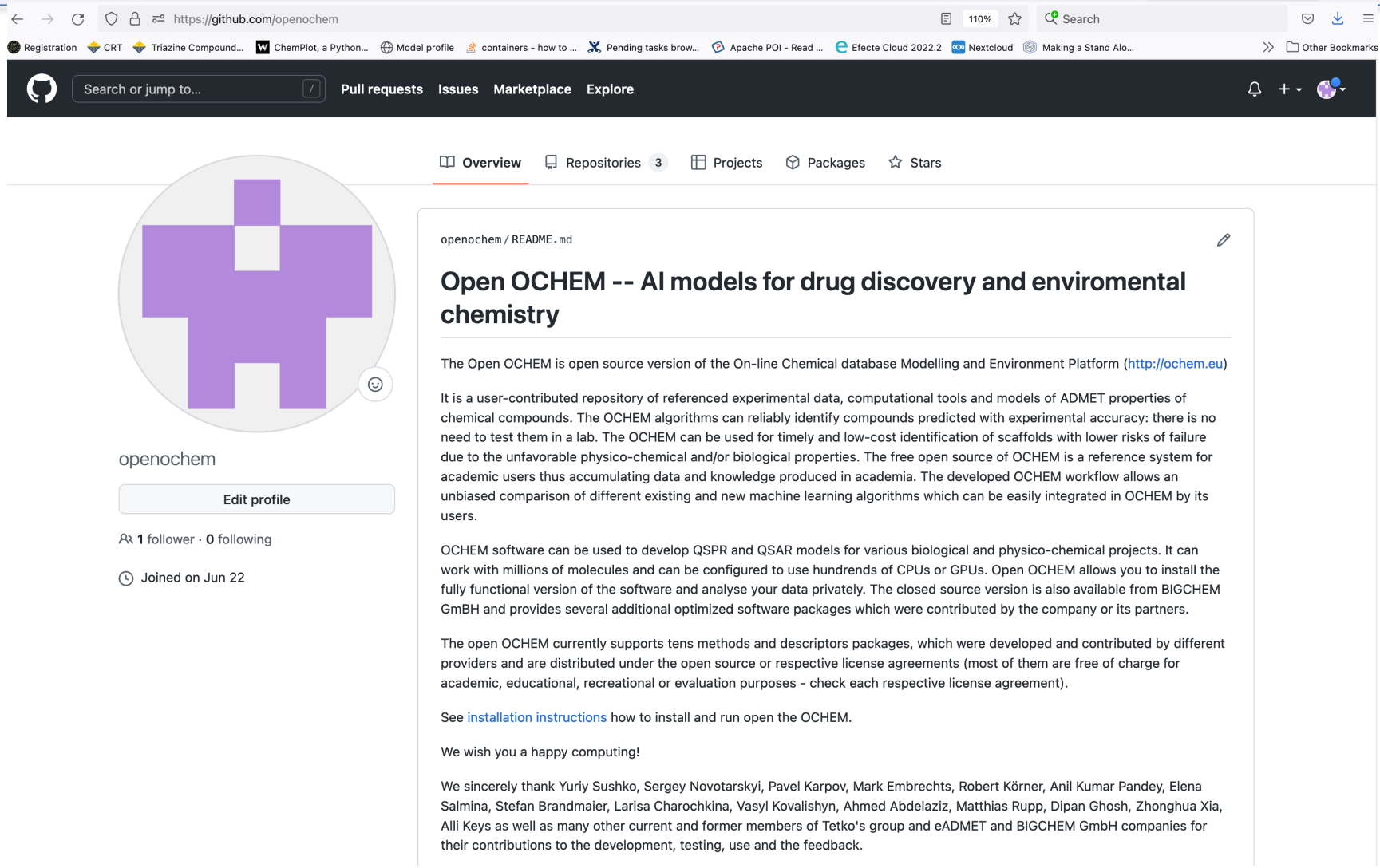
Bayer workflow for model life cycle



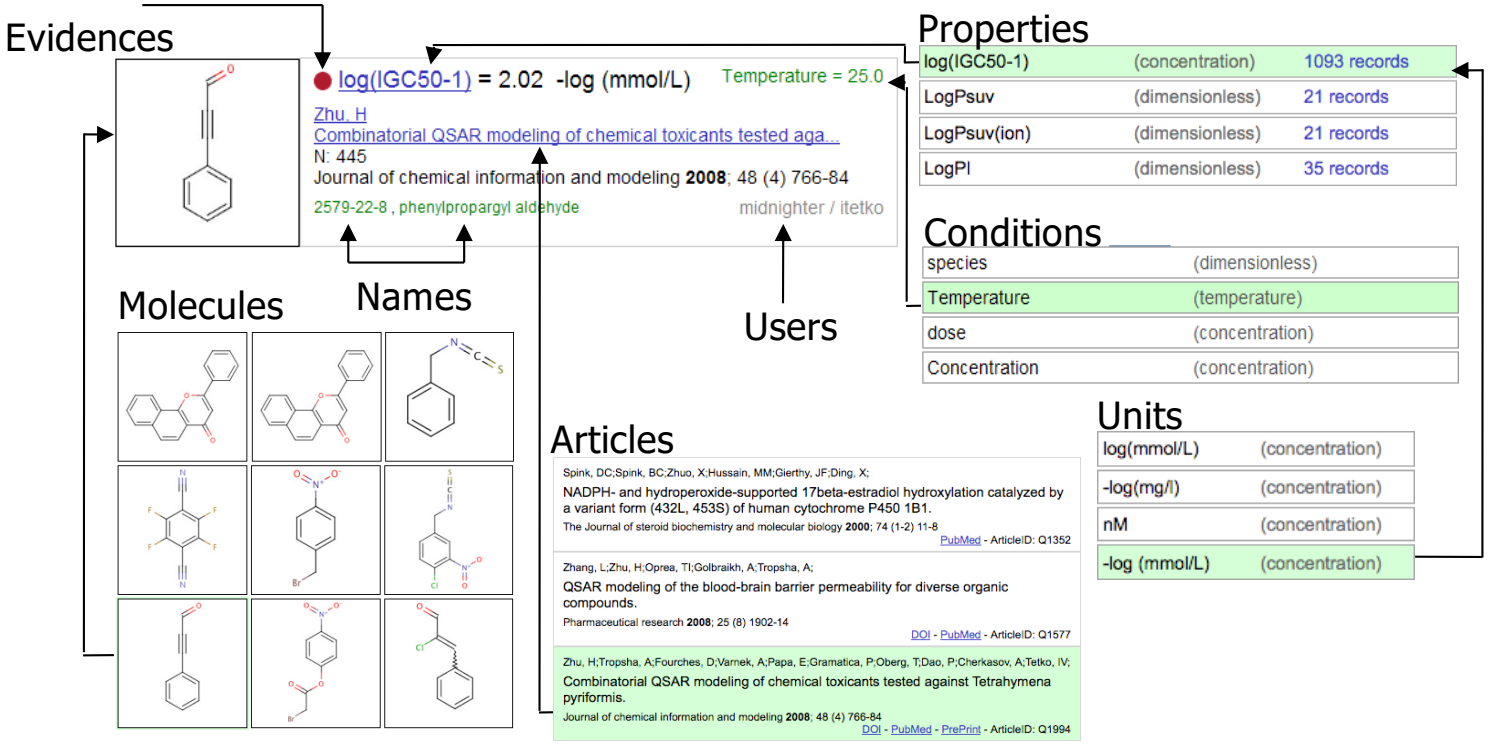
Göller, A.H. et al. *Drug Discov. Today* **2020,** *25* (9), 1702-1709.

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




https://github.com/openochem
110%
Search
Registration
CRT
Triazine Compound...
ChemPlot, a Python...
Model profile
containers - how to ...
Pending tasks brow...
Apache POI - Read ...
Efecte Cloud 2022.2
Nextcloud
Making a Stand Alo...
Other Bookmarks
Search or jump to...
Pull requests
Issues
Marketplace
Explore
Overview
Repositories 3
Projects
Packages
Stars
openochem
Edit profile
1 follower · 0 following
Joined on Jun 22
openochem / README.md
Open OCHEM -- AI models for drug discovery and enviromental chemistry
The Open OCHEM is open source version of the On-line Chemical database Modelling and 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 hundrends 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, 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.

Database schema: importance to store comprehensive information



Quantitative Structure Activity (QSAR)

Basic QSAR (Hansch, 1962):

activity = function (structure)

y = f (descriptors(structure))

f – linear, descriptors – complex, derived from structure

Machine learning:

y = F(Representation(structure))

F – complex, non-linear functions, learnable
Representation – learnable descriptors (“descriptor-less”)

Traditional representation of chemical structures

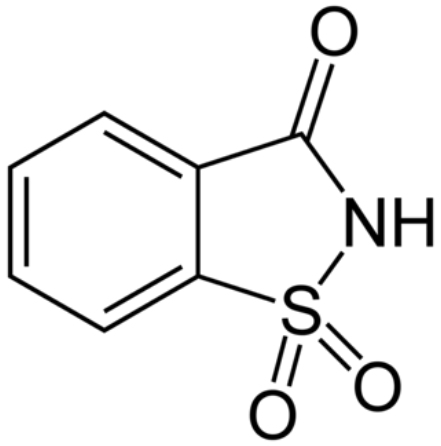
Saccharin

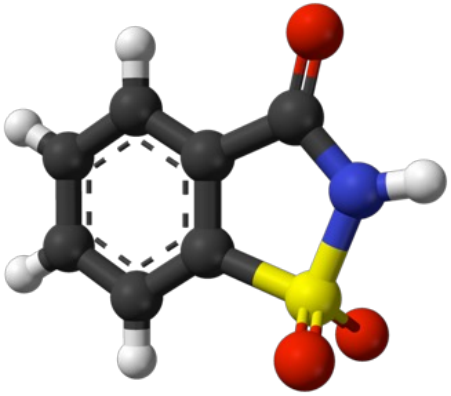
1D

2D

3D

$C_7H_5NO_3S$





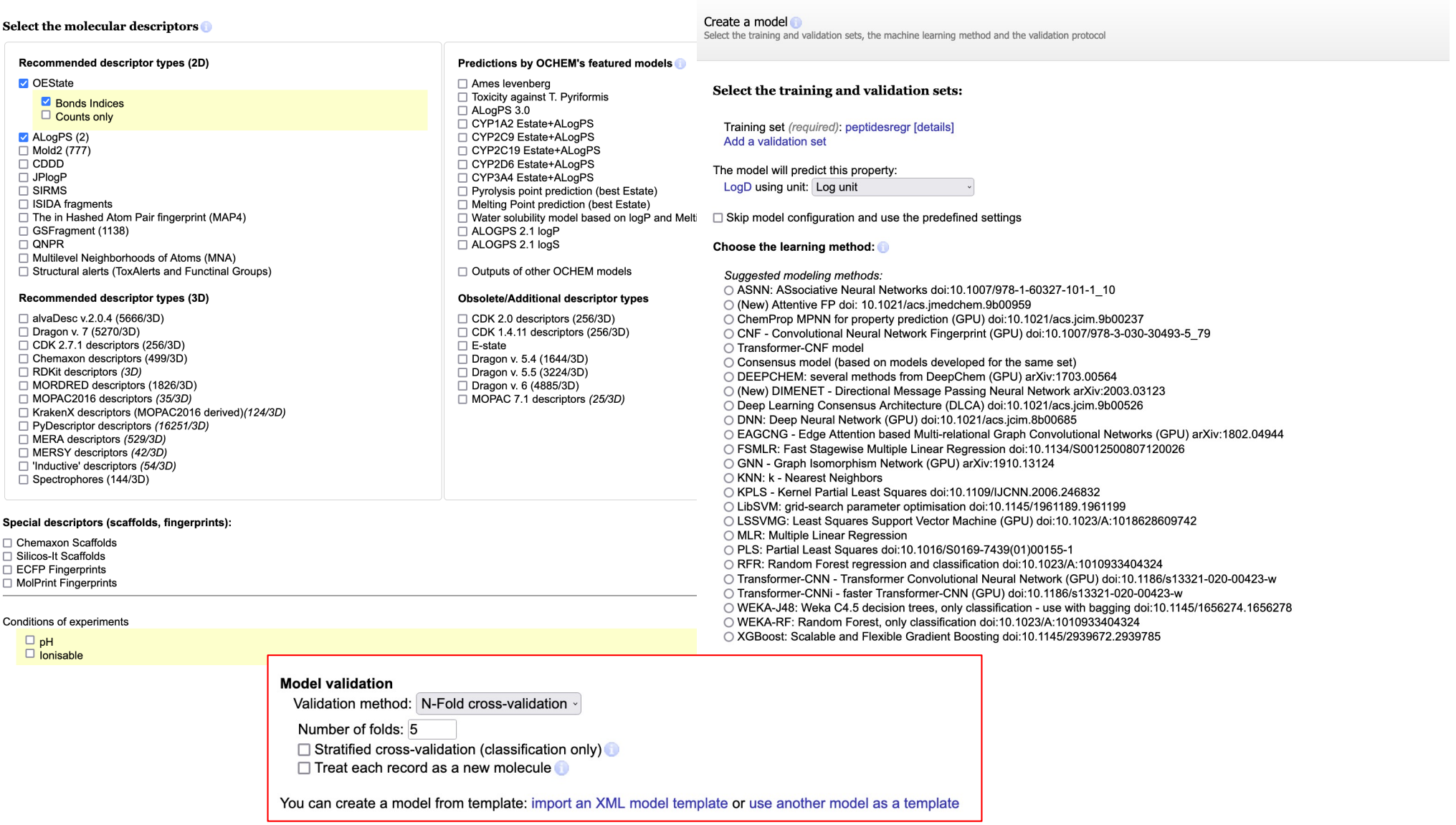
Examples of descriptors

- [x] alvaDesc v.2.0.4 (5666/3D)

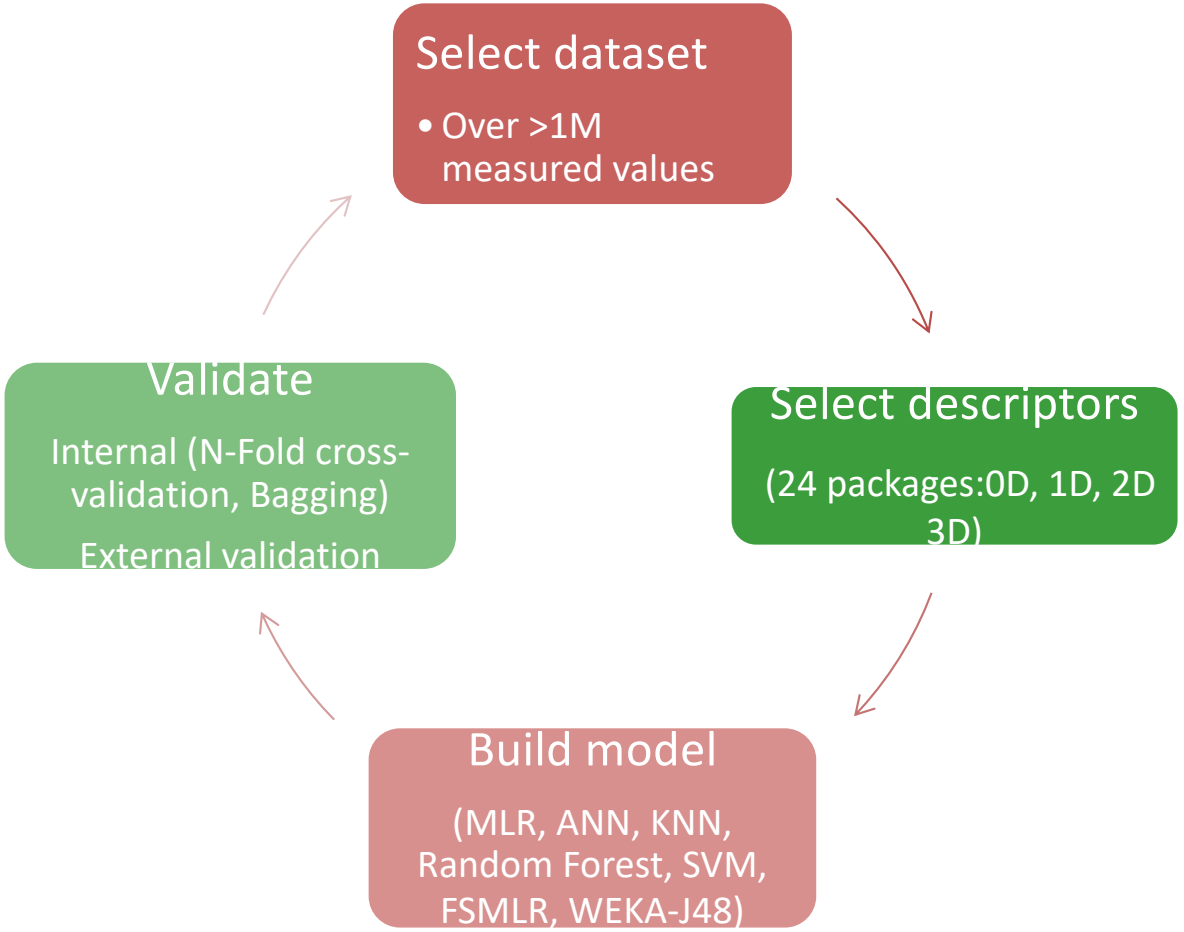
[select all] [select none] [select 3D] [unselect 3D]

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

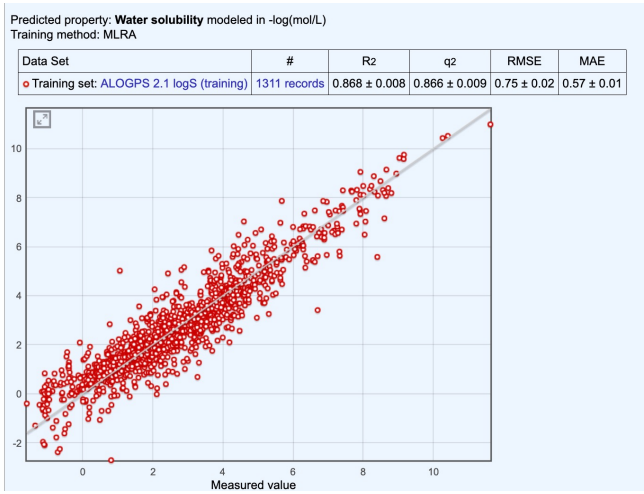
QSPR/QSAR modelling in OCHEM

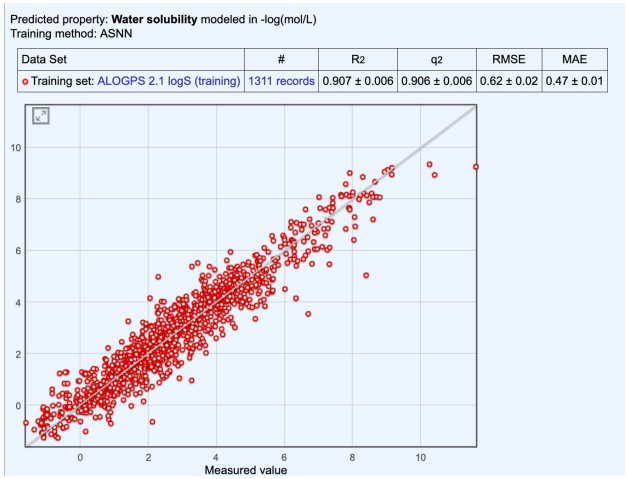


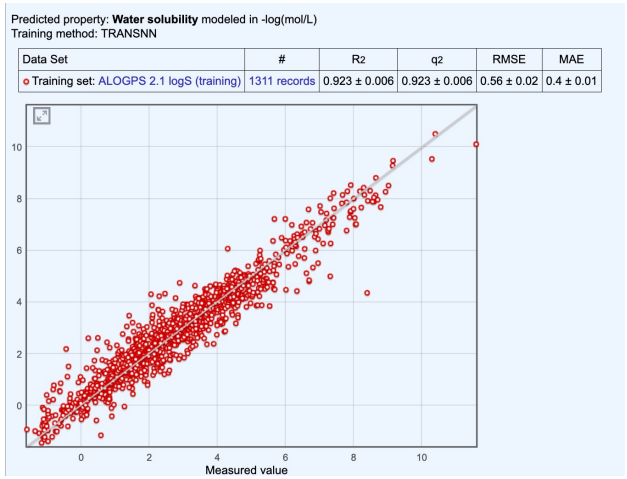
Modeling iterative workflow



Examples of models for water solubility







Linear

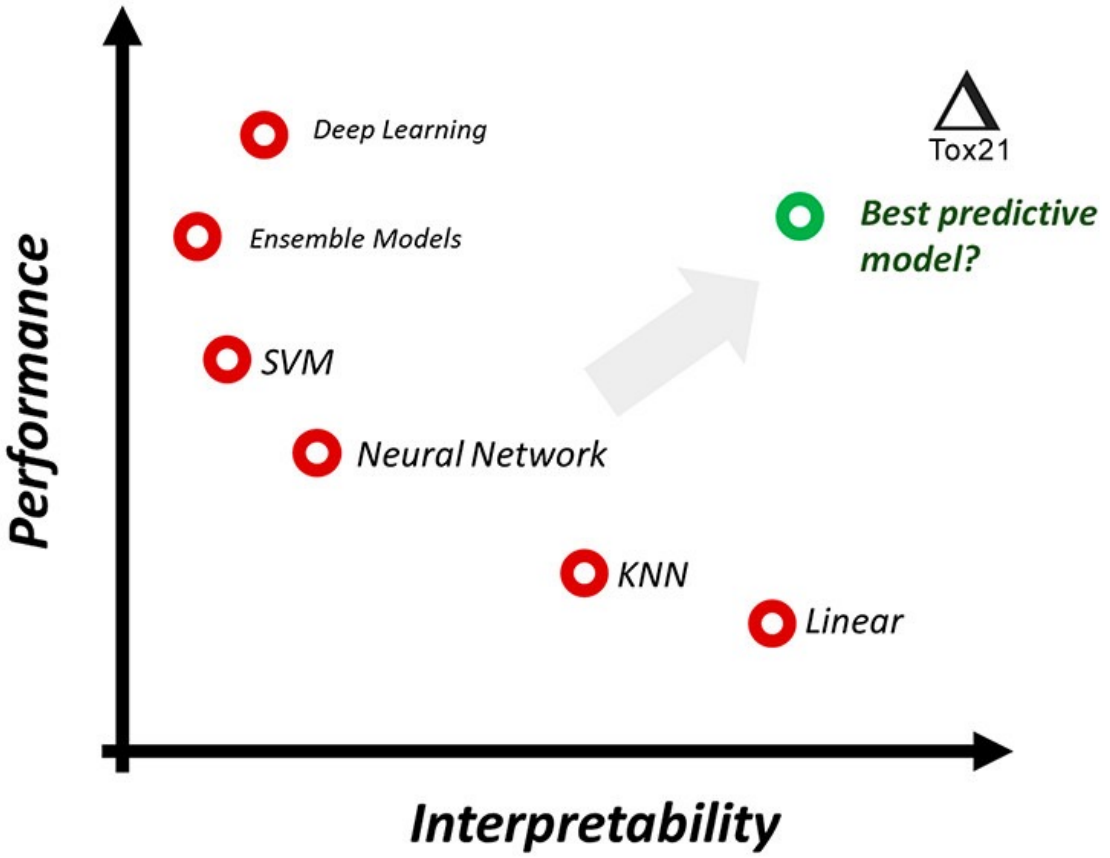
Standard (shallow) NN

Advanced NN

Interpretability

higher

lower



Overall best performance
5CV/TEST

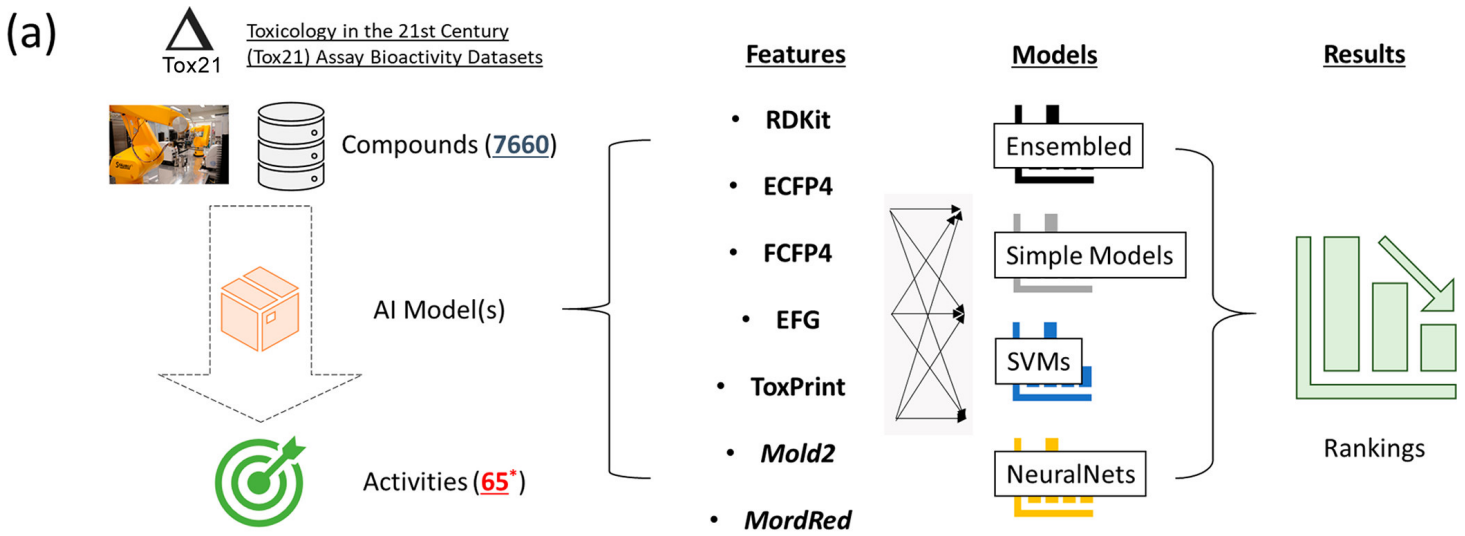
RF + all: 0.84/0.84

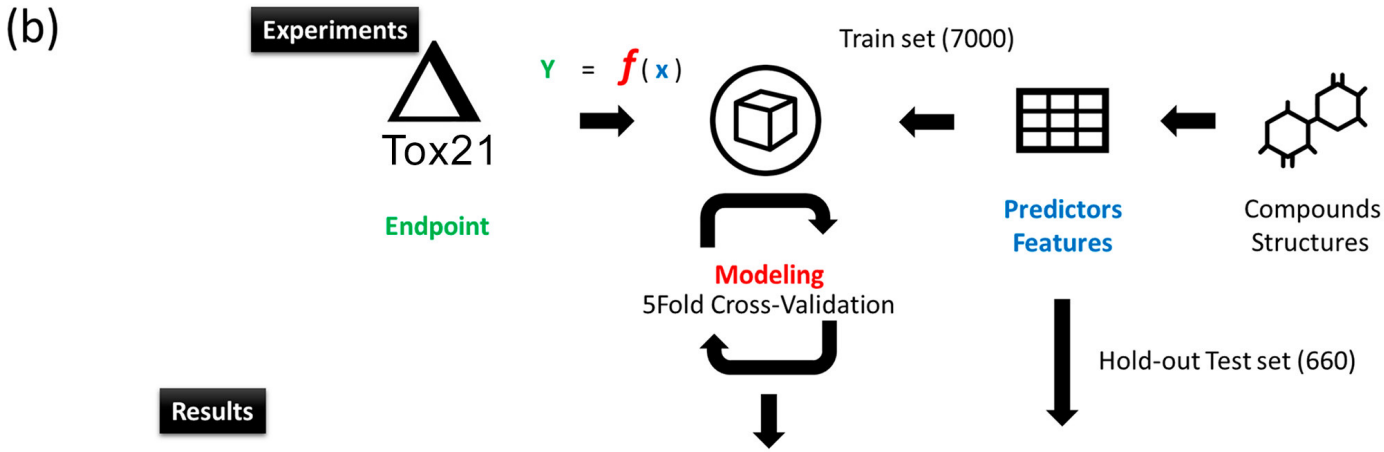
LS-SVM + MORDRED: 0.87/0.88

RF + EGF: 0.81/0.86

All + MOLD2: 0.82/0.84

Wu, L. et al **Trade-off Predictivity and Explainability for Machine-Learning...** *Chem. Res. Toxicol.* **2021**, *34*, 541-549





Results

Endpoint	Model	Predictors	AUC - CV	AUC - Test
tox21-err-p1 (1)	MordRed	SVM	0.7996	0.794

Overview of the workflow used to analyze the Tox21 450k dataset. (a) Overall study design. (b) Construct and evaluate predictive model with selected predictor, modeling algorithm, and end point.

Wu, L. et al *Chem. Res. Toxicol.* **2021**, *34*, 541-549.

Prediction of AMES mutagenicity

Predicted property: AMES
Training set: Ames challenge (training)

Metrics AUC for Training set Validation: Cross-Validation (69 models)

	MLRA (CHEMAXON)	MLRA (CHEMAXON)(2)	MLRA (CHEMAXON)(3)	LSSVMG (CHEMAXON)	DNN (CHEMAXON)	RFR (CHEMAXON)
CDK23 (cons,topol,geom,elect,hybr) 3D:corina	0.71	0.77	0.82	0.85	0.85	0.88
ChemaxonDescriptors (pH 0 - 14:1) 3D:corina	0.69	0.74	0.76	0.84	0.83	0.85
Dragon7 (3D blocks: 1-30) 3D:corina	0.72	0.75	0.83	0.86	0.84	0.88
EPA (T.E.S.T.)	0.66	0.74	0.83	0.87	0.84	0.88
Fragmentor (length: 2-4)	0.71	0.75	0.82	0.85	0.83	0.88
GSFrag (F + L)	0.72	0.77	0.79	0.81	0.82	0.85
JPlogP	0.74	0.76	0.8	0.84	0.83	0.86
MOPAC2016 (MOPAC basic) 3D:corina	0.72	0.73	0.73	0.78	0.81	0.82
OEstate	0.67	0.72	0.83	0.85	0.83	0.88
StructuralAlerts	0.72	0.75	0.76	0.82	0.81	0.86
alvaDesc (3D blocks: 1-33) 3D:corina	0.7	0.79	0.83	0.86	0.85	0.88
Descriptors:	*5*	*10*	*all*	*all*	*all*	*all*

Classification model for AMES test

Mopac2016 descriptors

- Y = 0.5378 - 0.3411*MullikenElectronegativity - 0.3277*LumoEnergy + 0.2389*IonisationPotential + 0.1178*FinalHeat + 0.05051*DipolPointCharge

GSFRAG

- Y = 0.5372 + 0.1612*c10 + 0.1309*p1-1N - 0.1134*p2B + 0.05943*c3 + 0.05349*c9

E-state descriptors

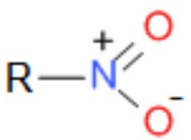
- Y = 0.5375 + 0.09956*PSA + 0.08731*aCNOS - 0.08703*DONORS - 0.06814*SsCH3 - 0.04474*SssO

Dragon descriptors

- Y = 0.5375 - 0.1173*GATS1m + 0.0954*MATS1e - 0.06558*SpMax_AEA(dm) + 0.05933*J_D/Dt + 0.05496*nR03

Structural Alerts

- Y = 0.4733 + 0.191*Alert146 - 0.004113*Alert238- 0.1024*Alert213 + 0.1912*Alert214+ 0.01988*Alert196



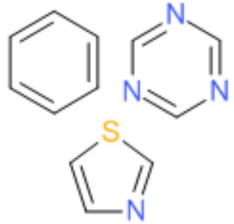
Aryl halides

R = aryl

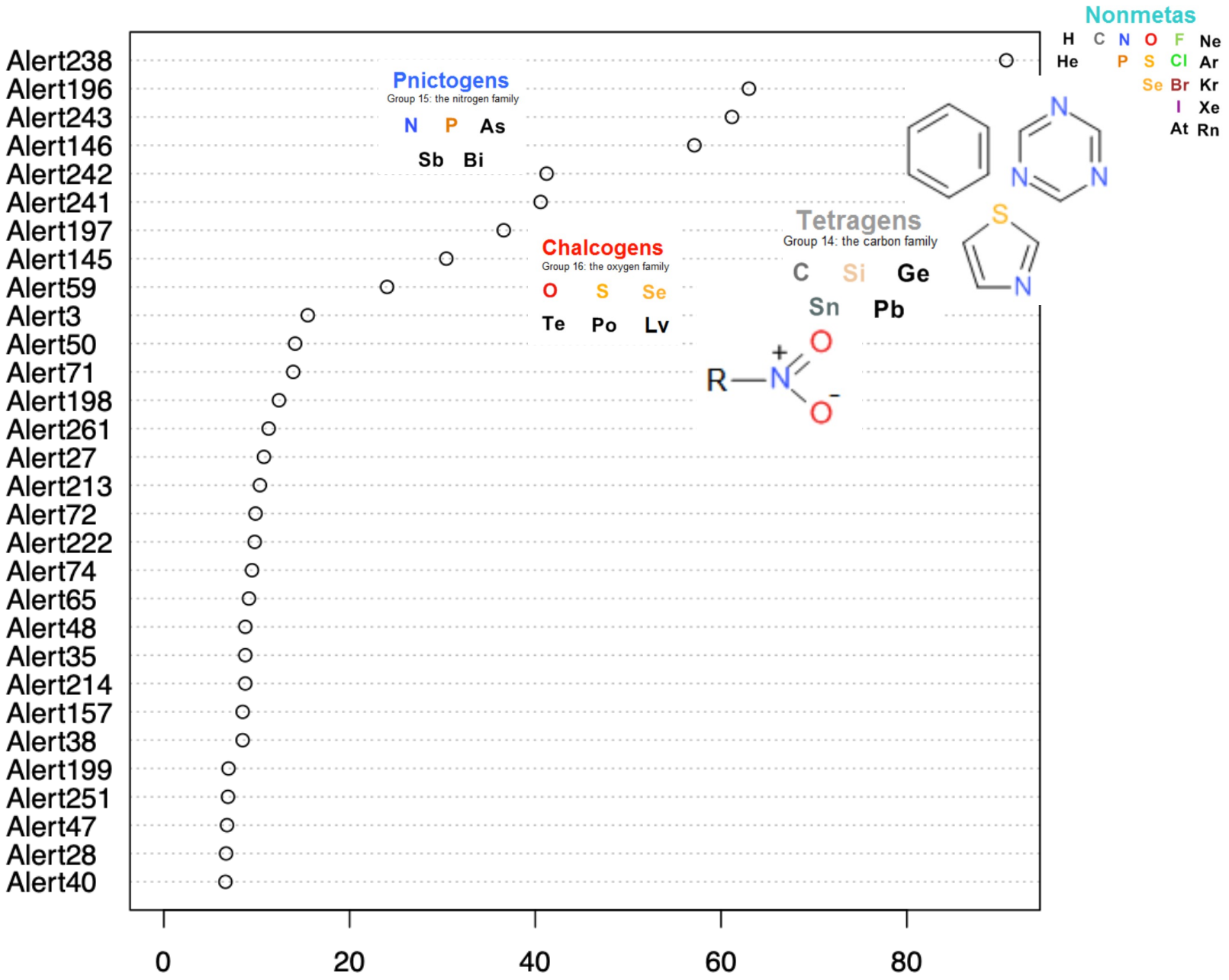
SMARTS: a[F,Cl,Br,I]

Alkyl halides

R = alkyl

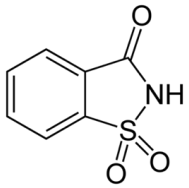


Importance of ToxAlerts in Random Forest model



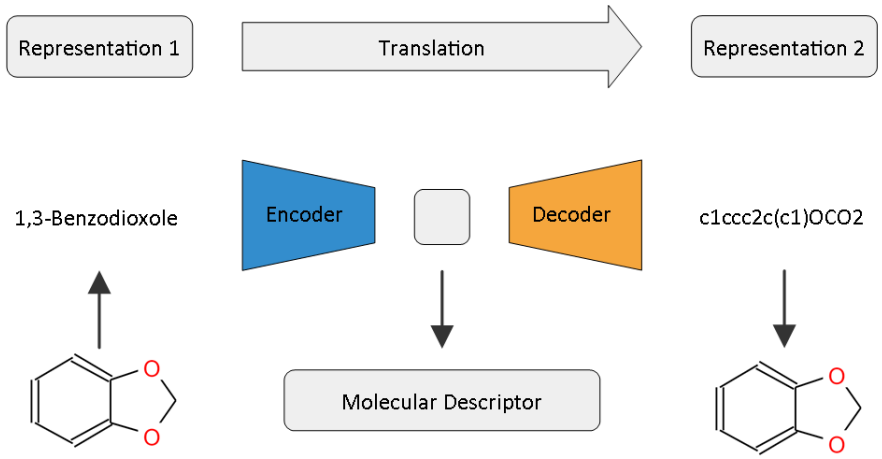
Machine Learning directly from chemical structures

Saccharin: c1ccc2c(c1)C(=O)NS2(=O)=O

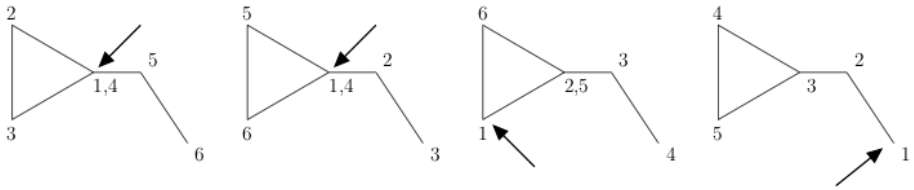


Text processing: convolutional neural networks, transformers, LSTM

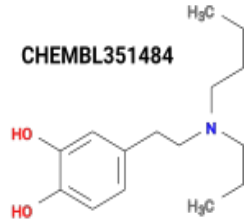
Graph processing: message passing neural networks



Machine Learning to canonise chemical structures



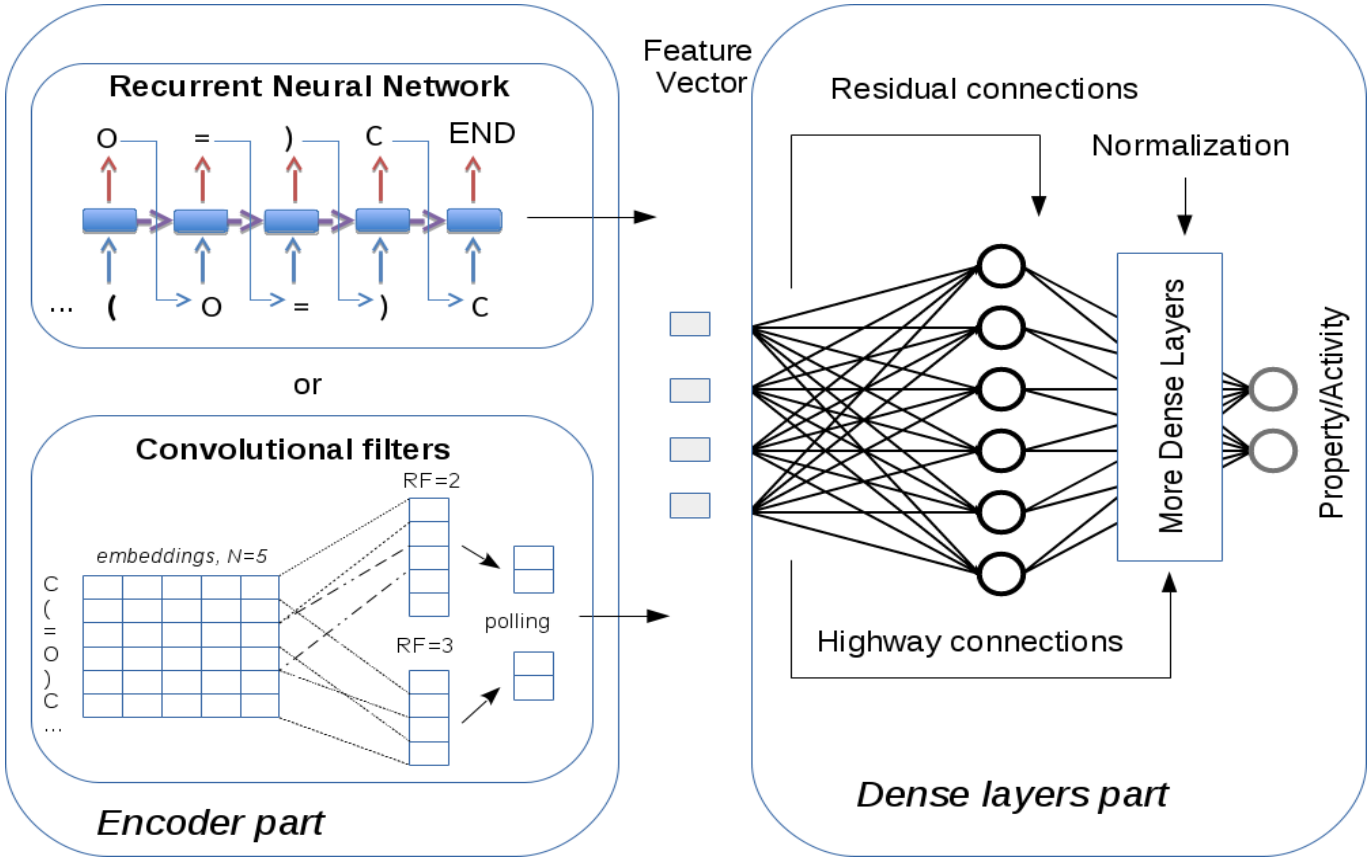
non-canonical		canonical
c1c(ccc(O)c1O)CCN(CCCC)CCC	>>	CCCCN(CCc1ccc(c(c1)O)O)CCC
c1(ccc(c(c1)O)O)CCN(CCC)CCCC	>>	CCCCN(CCc1ccc(c(c1)O)O)CCC
c1c(c(cc(c1)CCN(CCCC)CCC)O)O	>>	CCCCN(CCc1ccc(c(c1)O)O)CCC
c1(CCN(CCCC)CCC)ccc(O)c(O)c1	>>	CCCCN(CCc1ccc(c(c1)O)O)CCC
CCCCN(CCc1cc(c(cc1)O)O)CCC	>>	CCCCN(CCc1ccc(c(c1)O)O)CCC
CCCCN(CCc1ccc(c(c1)O)O)CCC	>>	**CCCCN(CCc1ccc(c(c1)O)O)CCC**
N(CCCC)(CCc1ccc(c(O)c1)O)CCC	>>	CCCCN(CCc1ccc(c(c1)O)O)CCC
C(N(CCc1ccc(O)c(c1)O)CCCC)CC	>>	CCCCN(CCc1ccc(c(c1)O)O)CCC
N(CCc1ccc(c(O)c1)O)(CCCC)CCC	>>	CCCCN(CCc1ccc(c(c1)O)O)CCC
c1c(O)c(ccc1CCN(CCC)CCCC)O	>>	CCCCN(CCc1ccc(c(c1)O)O)CCC
c1(c(cc(cc1)CCN(CCC)CCCC)O)O	>>	CCCCN(CCc1ccc(c(c1)O)O)CCC



SMILES canonization can be done by machine learning!

ChEMBL database (1.7M) was used, >95% accuracy

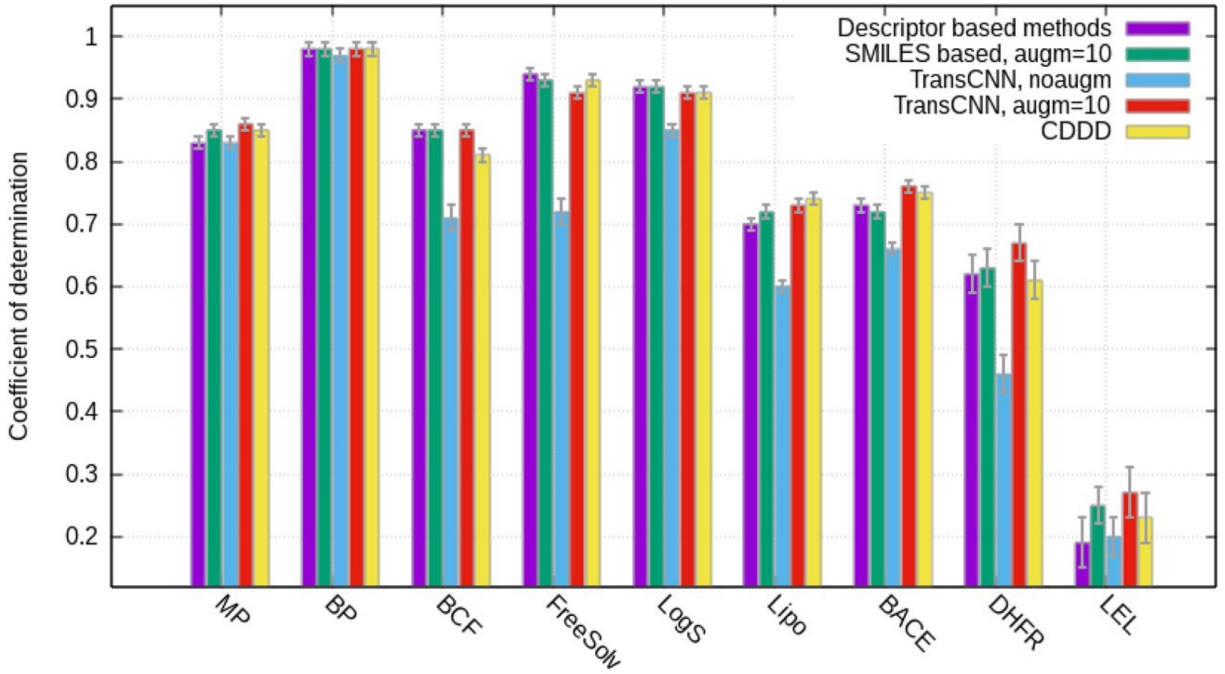
Machine Learning directly from chemical structures



P. Karpov, G. Godin, I. V. Tetko, *J. Cheminform.* **2020**, *12*, 17.

https://github.com/bigchem/transformer-cnn

Convolutional vs. Descriptor-based Neural Neural Networks



Coefficient of determination, r^2. Transformer CNN provides similar or better accuracy compared to traditional methods based on descriptors **even for small datasets (few hundrends compounds!)**.
P. Karpov, G. Godin, I. V. Tetko, *J. Cheminform.* **2020**, *12*, 17.

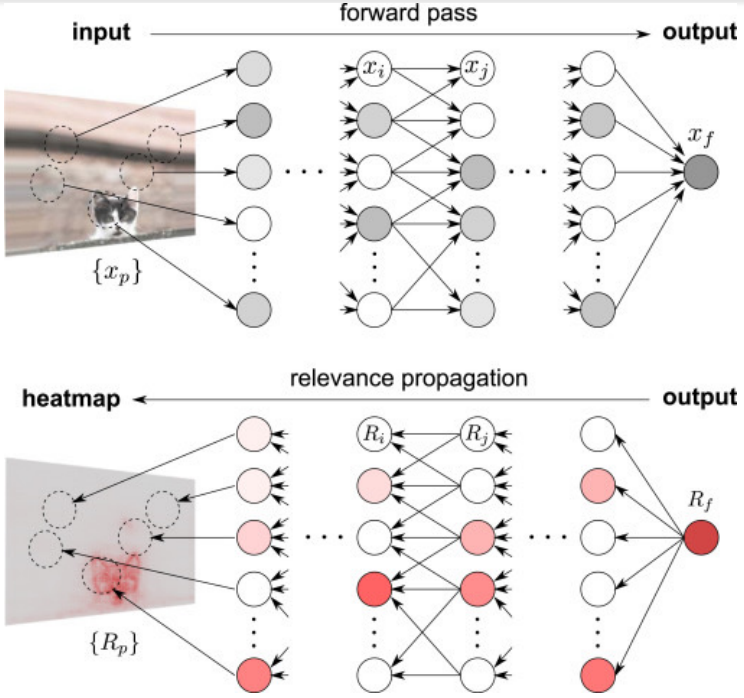
AMES mutagenicit

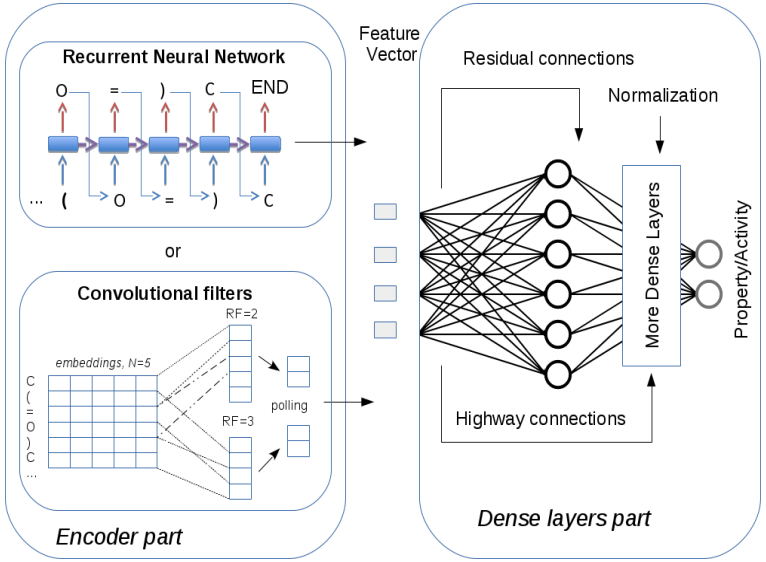
Predicted property: AMES
Training set: Ames challenge (training)

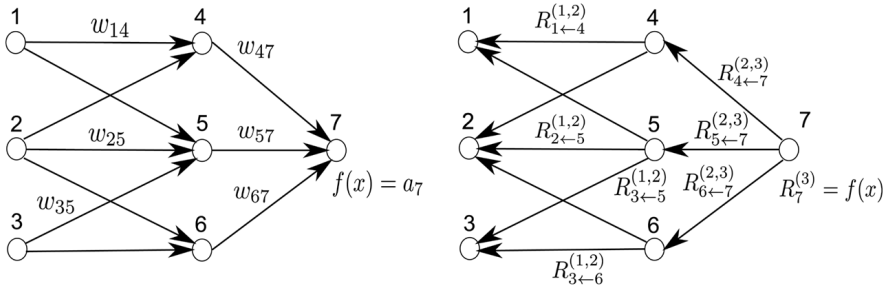
Metrics AUC for Training set Validation: Cross-Validation (67 models)

	MLRA (CHEMAXON)	MLRA (CHEMAXON)(2)	MLRA (CHEMAXON)(3)	LSSVMG (CHEMAXON)	DNN (CHEMAXON)	RFR (CHEMAXON)	TRANSNNI (F) (CHEMAXON)
CDK23 (cons,topol,geom,elect,hybr) 3D:corina	0.71	0.77	0.82	0.85	0.85	0.88	+
ChemaxonDescriptors (pH 0 - 14:1) 3D:corina	0.69	0.74	0.76	0.84	0.83	0.85	+
Dragon7 (3D blocks: 1-30) 3D:corina	0.72	0.75	0.83	0.86	0.84	0.88	+
EPA (T.E.S.T.)	0.66	0.74	0.83	0.87	0.84	0.88	+
Fragmentor (length: 2-4)	0.71	0.75	0.82	0.85	0.83	0.88	+
GSFrag (F + L)	0.72	0.77	0.79	0.81	0.82	0.85	+
JPlogP	0.74	0.76	0.8	0.84	0.83	0.86	+
MOPAC2016 (MOPAC basic) 3D:corina	0.72	0.73	0.73	0.78	0.81	0.82	+
OEstate	0.67	0.72	0.83	0.85	0.83	0.88	+
StructuralAlerts	0.72	0.75	0.76	0.82	0.81	0.86	+
alvaDesc (3D blocks: 1-33) 3D:corina	0.7	0.79	0.83	0.86	0.85	0.88	+
SMILES 10/10	+	+	+	+	+	+	0.88

Layerwise Relevance Propagation (LRP)







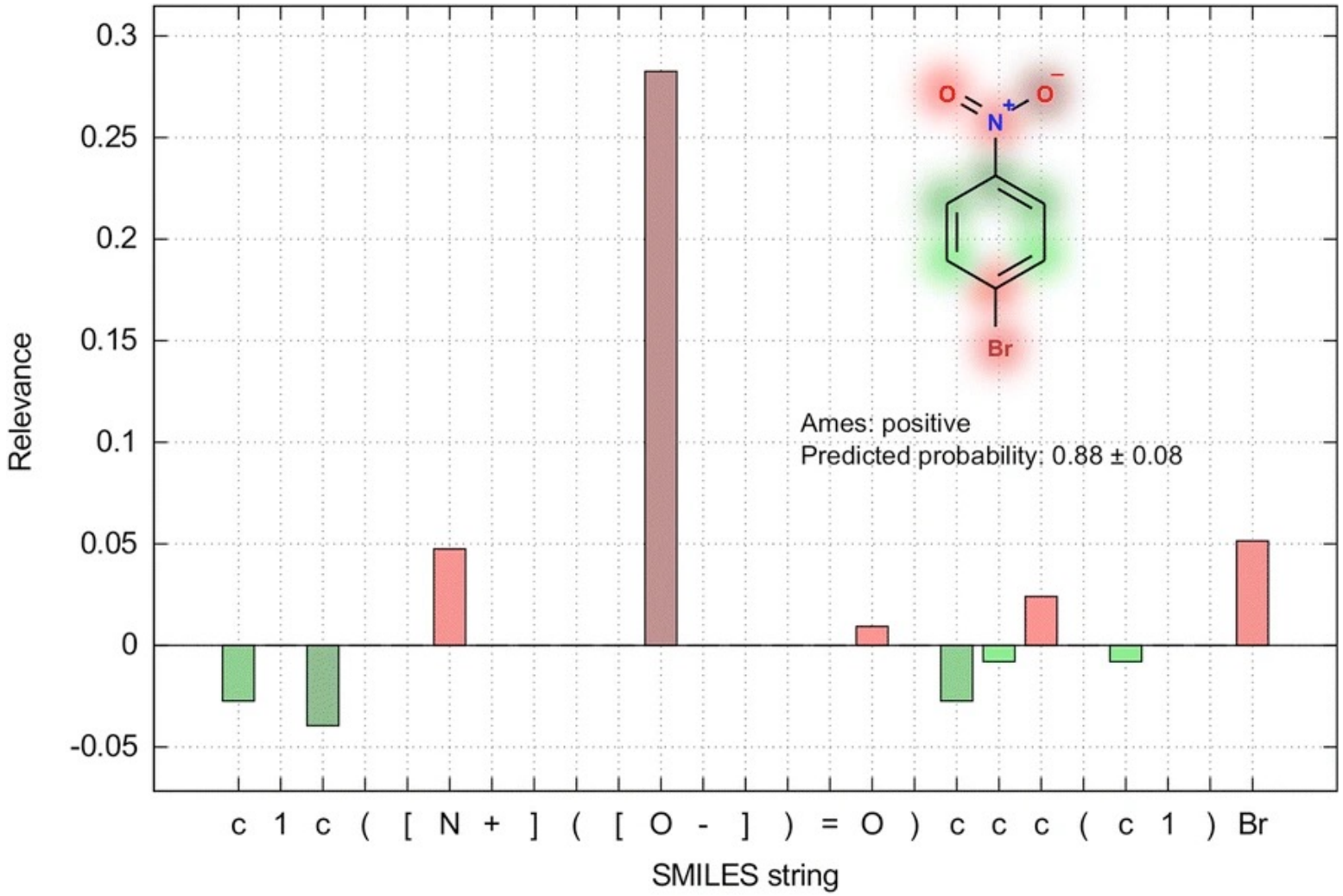
$$f(x) \approx f(x_0) + Df(x_0)[x - x_0]$$

$$= f(x_0) + \sum_{d=1}^{V} \frac{\partial f}{\partial x_{(d)}}(x_0)(x_{(d)} - x_{0(d)})$$

Bach, S. et al. *PloS One* **2015**, *10*, e0130140.

P. Karpov, G. Godin, I. V. Tetko, *J. Cheminform.* **2020**, *12*, 17.

Interpretation of models



P. Karpov, G. Godin, I. V. Tetko, *J. Cheminform.* **2020**, *12*, 17.

https://github.com/bigchem/transformer-cnn

AiChemist MSC DN

https://aichemist.eu/

Home

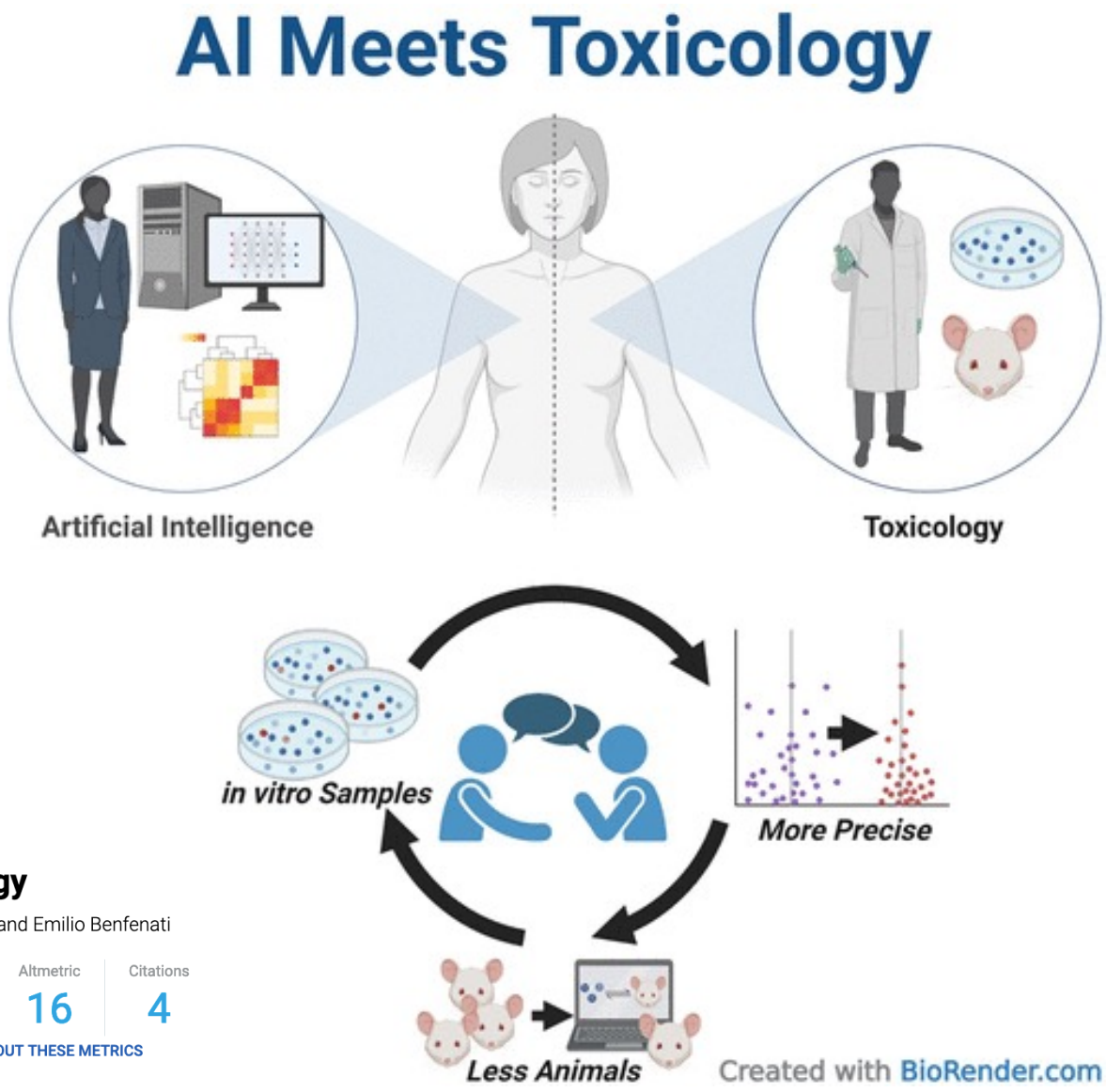
MSC ITN Project AiChemist

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.

15 positions will be soon announced; first selection will be in September

Also see Twitter: https://twitter.com/AiddOne

Interested about AI in toxicology? Read ChemResTox August SI



Artificial Intelligence Meets Toxicology

Igor V. Tetko*, Günter Klambauer, Djork-Arné Clevert, Imran Shah, and Emilio Benfenati

Cite this: *Chem. Res. Toxicol.* 2022, 35, 8, 1289–1290
Publication Date: August 15, 2022
https://doi.org/10.1021/acs.chemrestox.2c00196

Winning model: OCHEM-generated consensus model

Andrea Kopp

SLAS Europe 2023

25.05.2023

HELMHOLTZ MUNICH

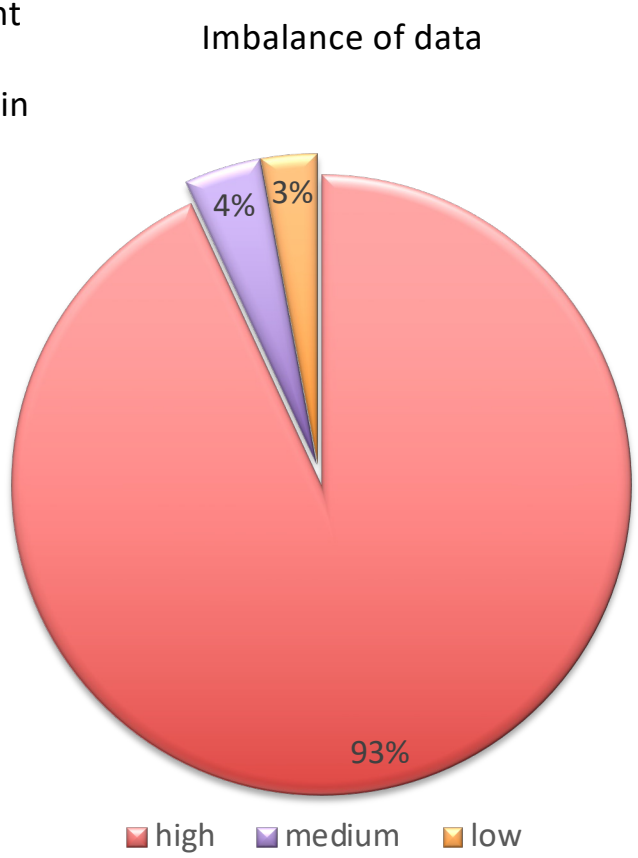
Team of Igor Tetko with Peter Hartog, Martin Šícho and Guillaume Godin

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

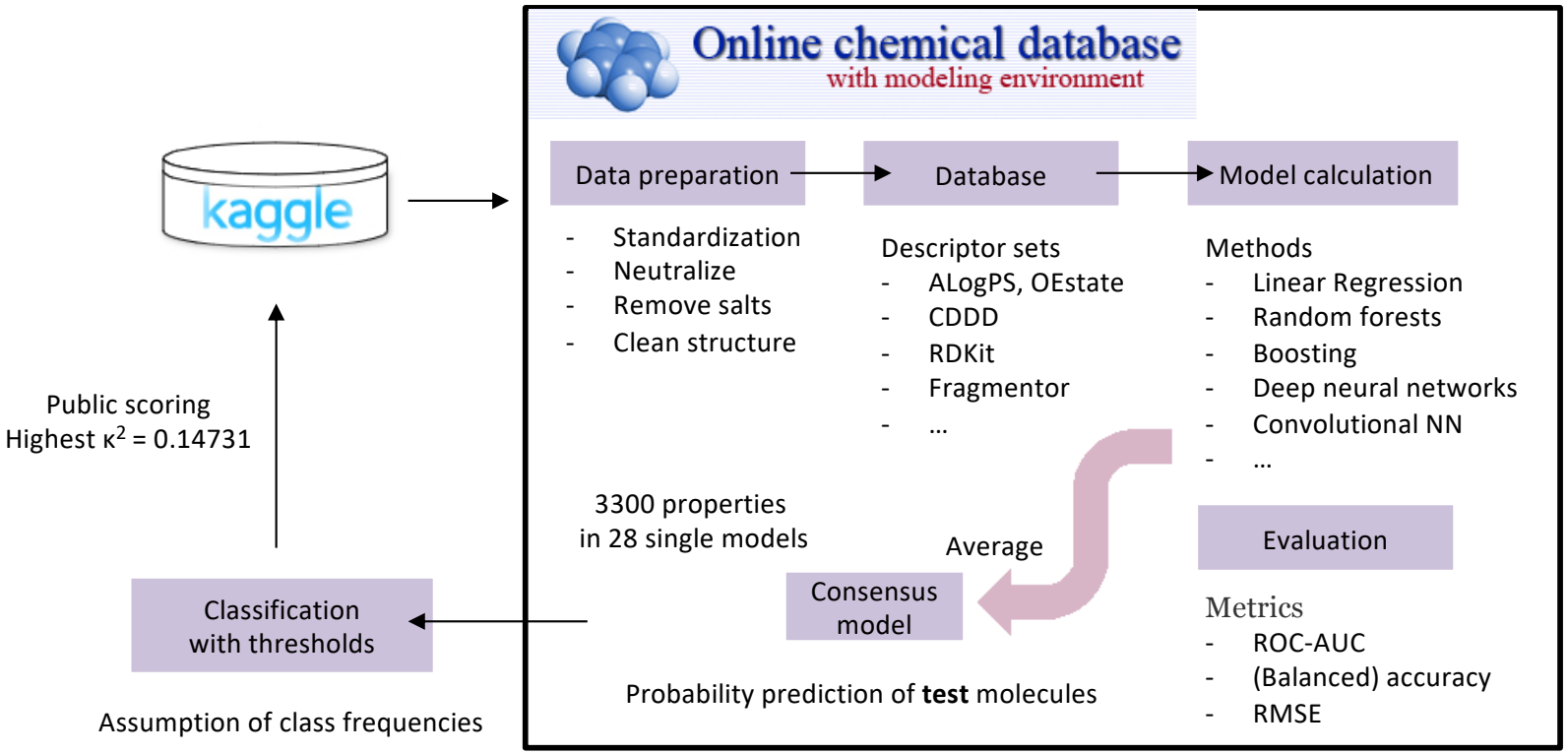
Kopp at al, DOI: 10.26434/chemrxiv-2023-p8qcv

Challenge set-up

- Experimentally: Nephelometer measures undissolved sediment
- Classification into *low*, *medium* and *high* soluble with phenytoin and amiodarone as thresholds
- 70k training datapoints, 15k public leaderboard, 15k private leaderboard
- Stratified random sampling



Workflow with OCHEM



OCHEM for modeling

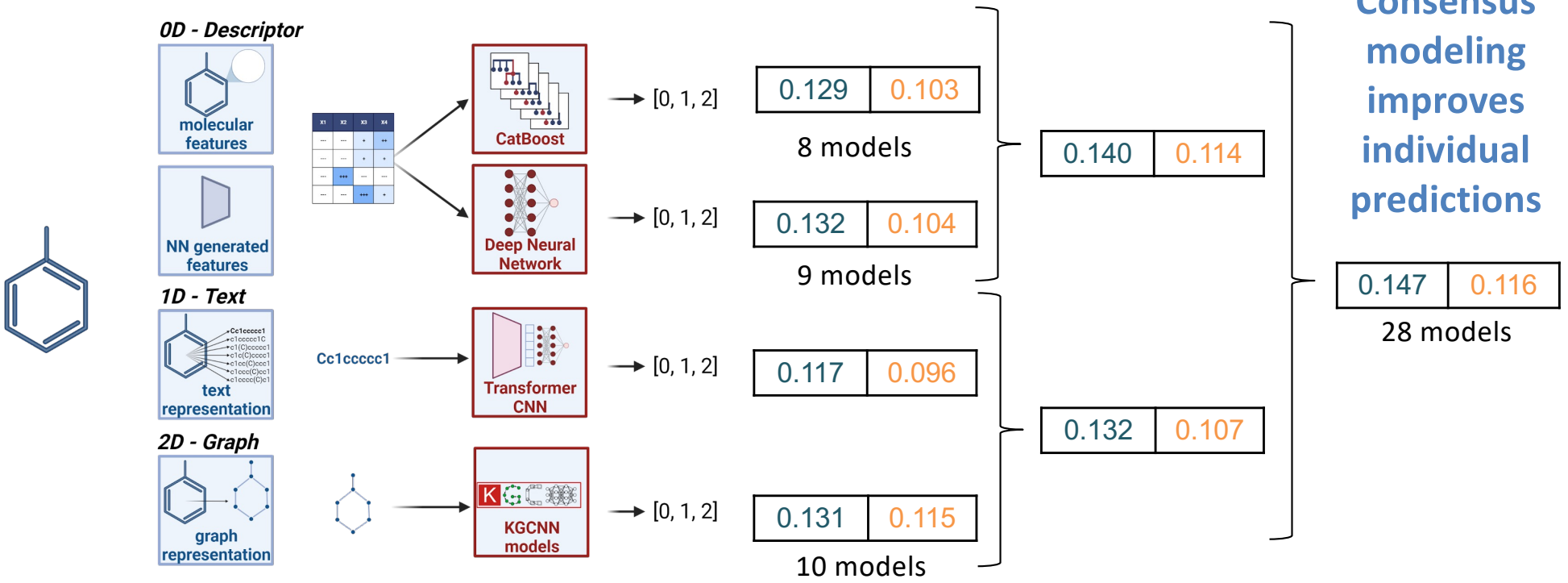
- Graphical interface allows comprehensive modeling without explicit coding
- Implementation for GPU and CPU use
- Consensus models:
  - Average over multiple models to improve prediction
  - Orthogonal models
  - Various descriptor sets/ molecular representations

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

Quadratic kappa metric scores →

Public leaderboard	Private leaderboard



Kopp at al, DOI: [10.26434/chemrxiv-2023-p8qcv](https://doi.org/10.26434/chemrxiv-2023-p8qcv)

www.epa.gov/ncct/challenges.html

EPA United States Environmental Protection Agency

ALL EPA THIS AREA Advanced Search SEARCH

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

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

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

National Center for Advancing Translational Sciences

Tox21 Data Challenge 2014

Login

» Home | About | Registration | Data/Resources | Submissions | Discussion | Leaderboard | Contact Us | Survey NEW

Open call ends: November 14, 2014

About the Data

The Challenge

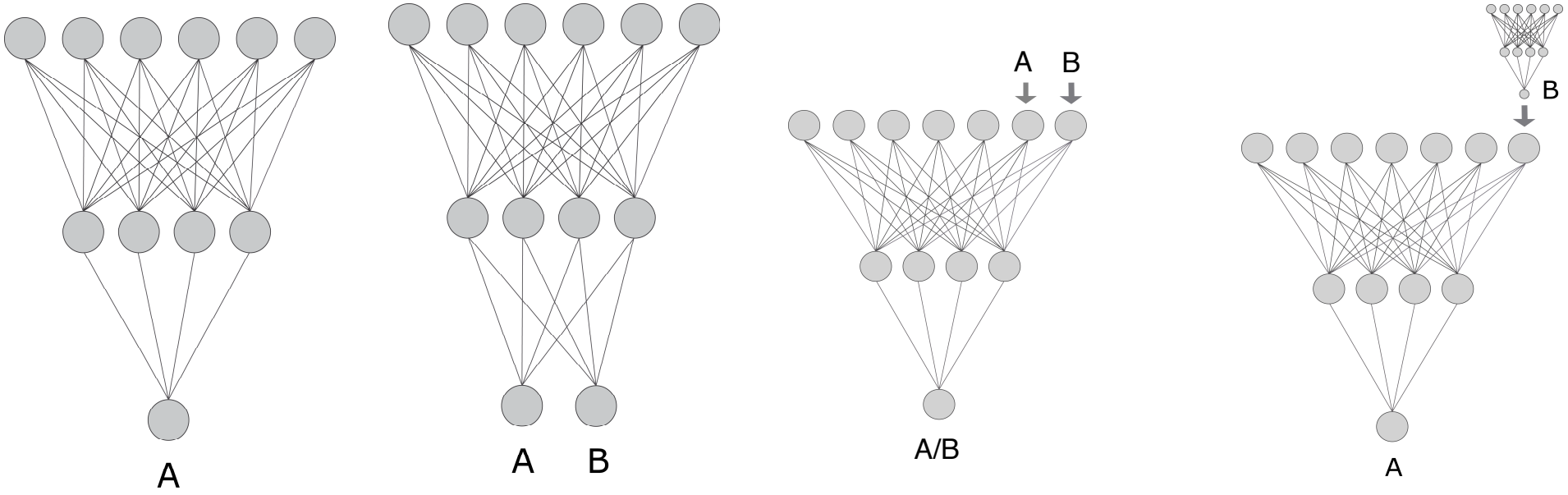
The 2014 Tox21 data challenge is designed to help scientists understand the potential of the chemicals and compounds being tested through the Toxicology in the 21st Century initiative to disrupt biological pathways in ways that may result in toxic effects.

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.

Best Balanced accuracy - Abdelaziz, A. et al. Front. Environ. Sci. 2016, 4, 2.

Multi-task learning



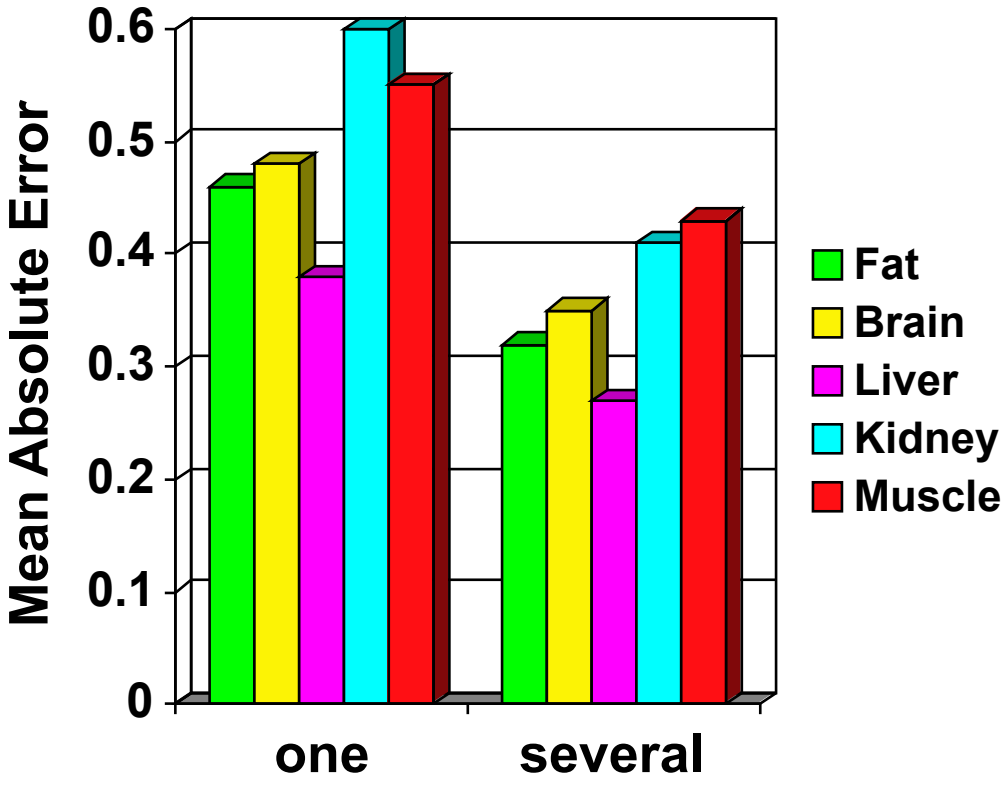
Multi-task learning

Problem:

• prediction of tissue-air partition coefficients
• small datasets 30-100 molecules (human & rat data)

Results:

simultaneous prediction of several properties increased the accuracy of models



Varnek, A. et al J. Chem. Inf. Model. 2009, 49, 133-44.

Prediction of toxicity of chemical compounds:
REGISTRY OF TOXIC EFFECTS OF CHEMICAL SUBSTANCES (RTECS®)

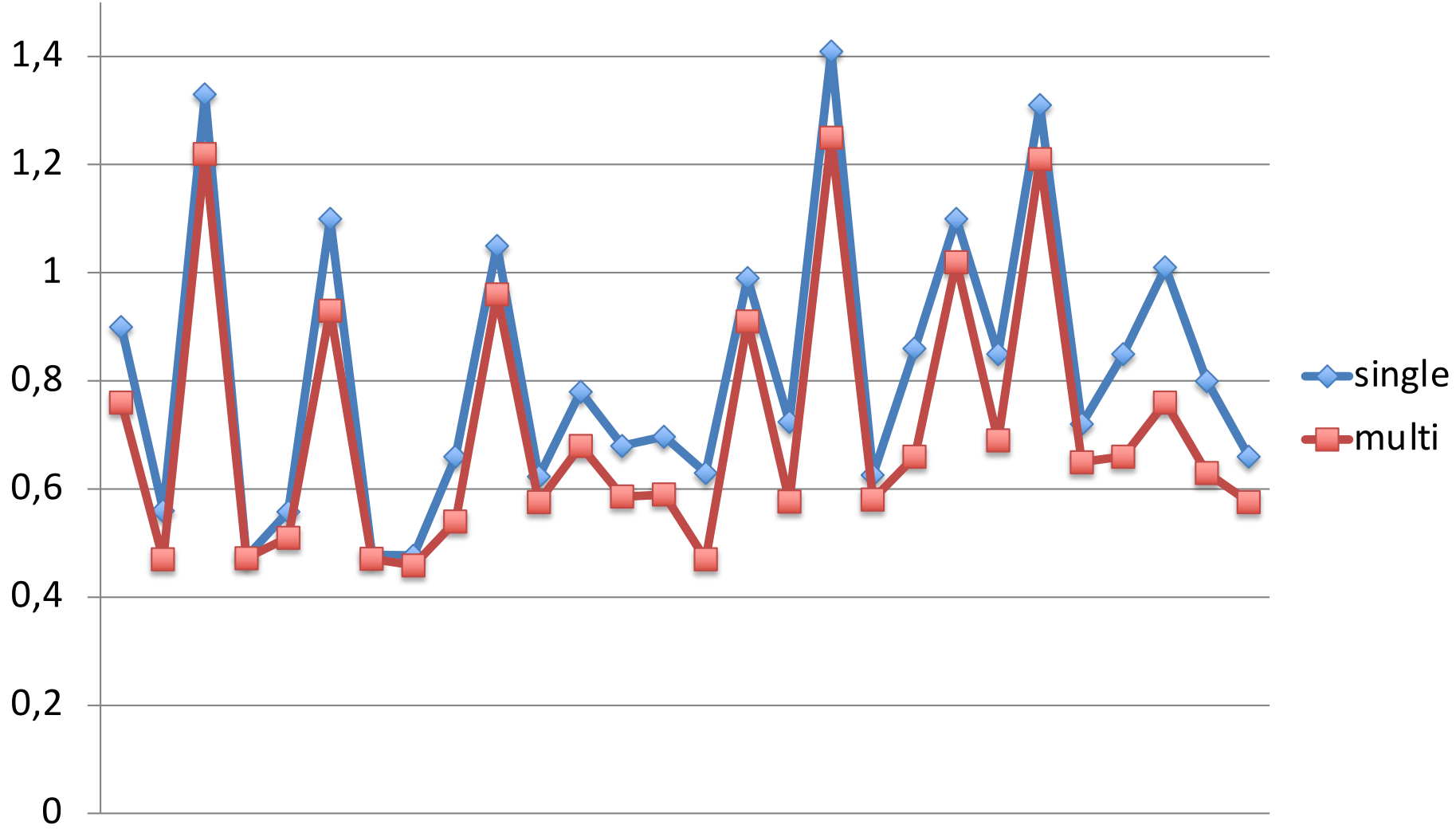
Different species
- Rat
- Mouse
- Rabbit
- …
- Human

~ 129k records
~ 87k compounds
29 properties

- Different toxicities
  - LD50
  - TDL
  - NOEL
  - LDLo
- Administartion
  - Oral
  - IPR (intraperitoneal)
  - IVR (intravenous)

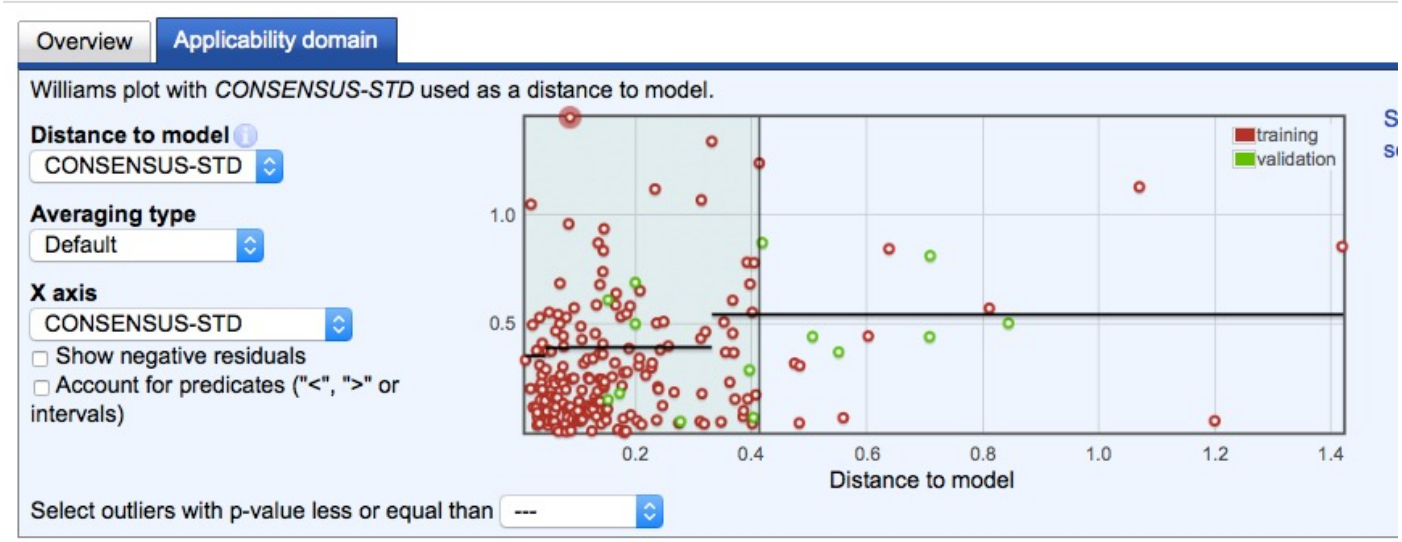
Sosnin, S.; Karlov, D.; Tetko, I.V.; Fedorov, M.V. A comparative study of prediction of multi-target toxicity for a broad chemical space. *J Chem Inf Model.* ***2018,*** **59**, 1062-1072.

RMSE for different toxicities using CDK descriptors and DNN



Sosnin, S.; Karlov, D.; Tetko, I.V.; Fedorov, M.V. A comparative study of prediction of multi-target toxicity for a broad chemical space. *J Chem Inf Model.* ***2018,*** **59**, 1062-1072.

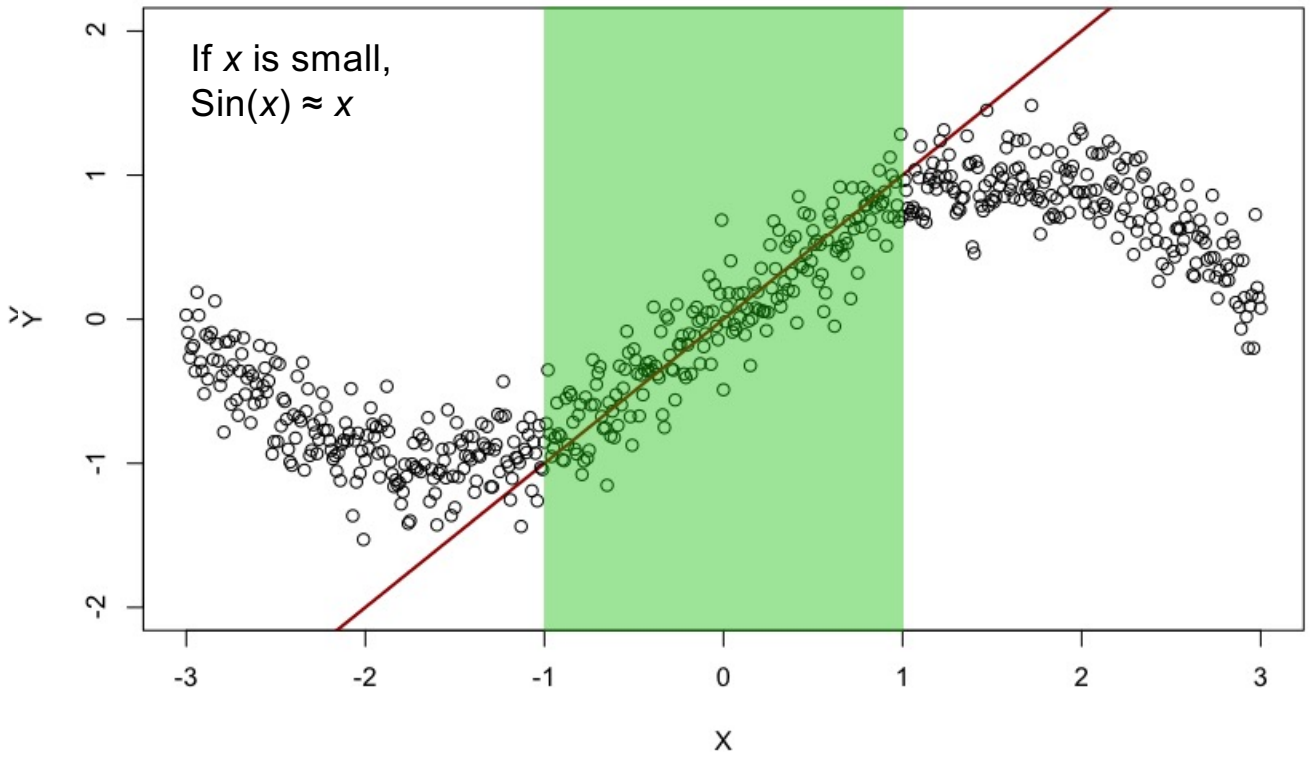
Outlying point on the Applicability Domain plot



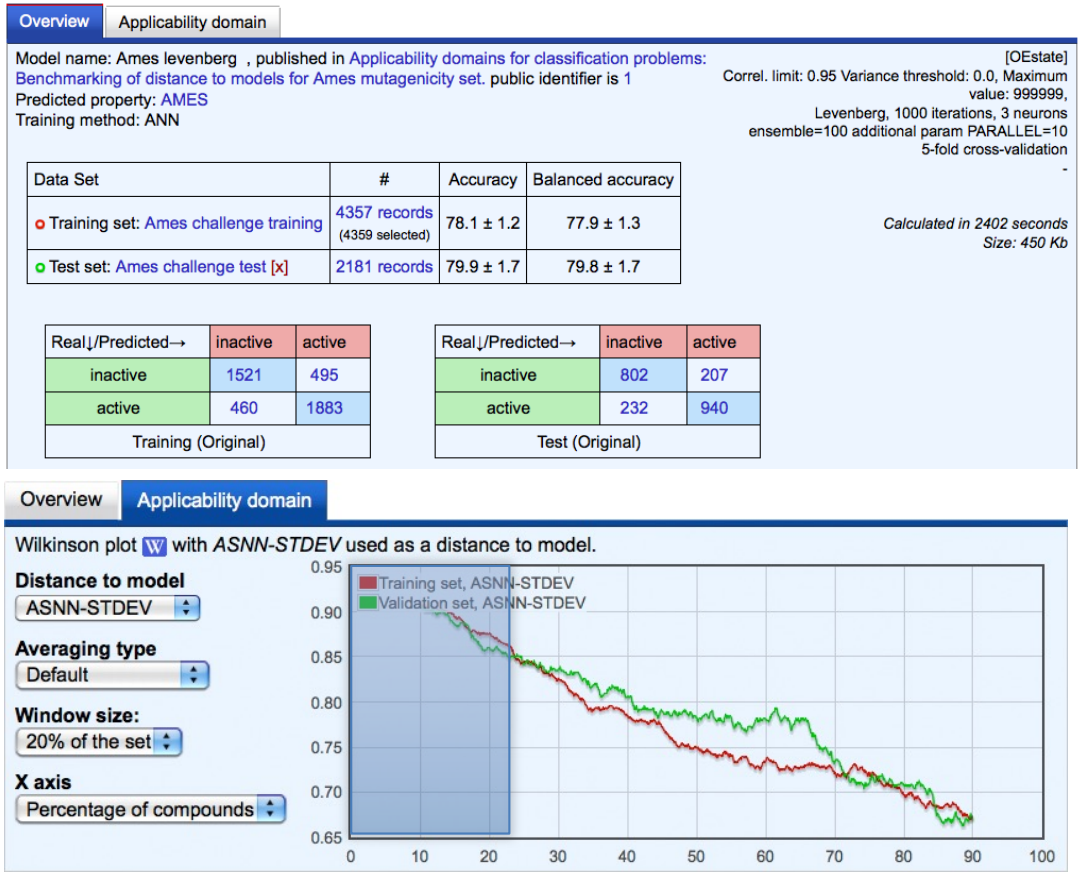
- Estimation of applicability domain of models
- Identification of outliers

Tetko et al, *J Chem Inf Model*, **2008,** 48(9):1733-46.

Accuracy of prediction

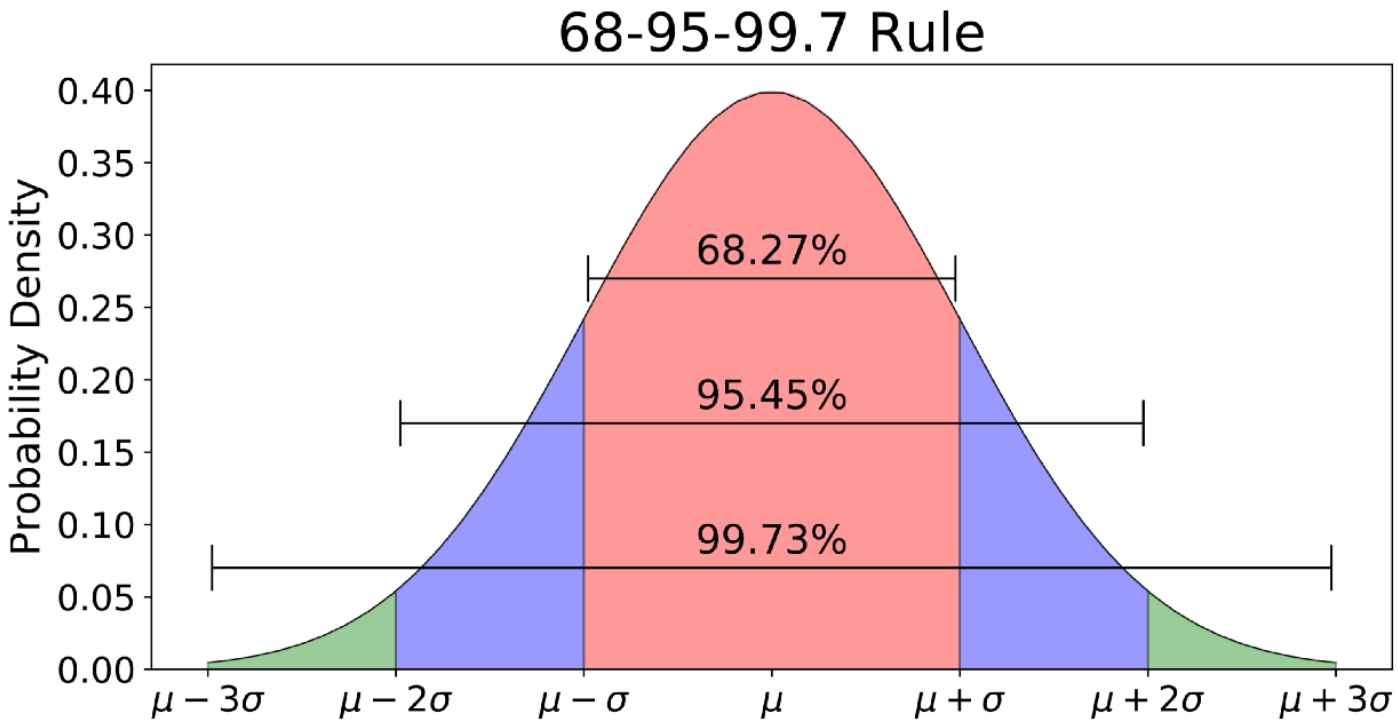


Accuracy of predictions for classification model

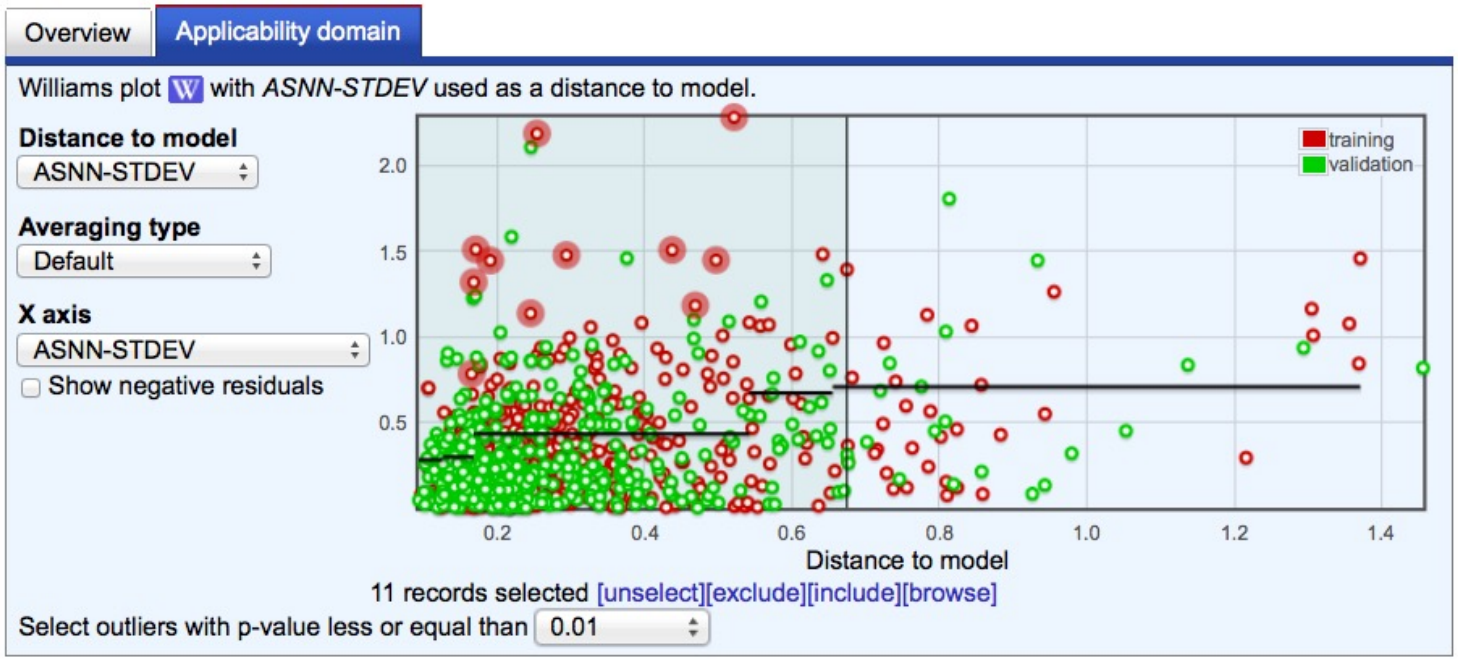


Sushko et al, JCIM, 2010, 50, 2094 - 2111.

Gaussian distribution and outliers detection



Applicability domain assessment (regression)



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

275k Melting Point Dataset

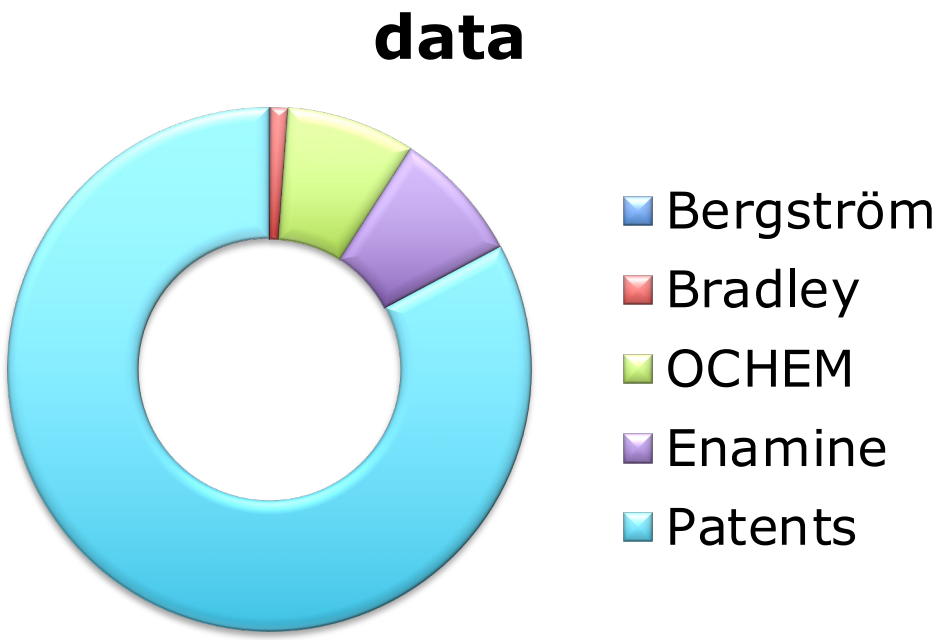
Bergström 277

Bradley 2886

OCHEM 22404

Enamine 21883

PATENTS 228079



COMBINED: OCHEM + Enamine + Bradley + Bergström

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

Outliers identified with applicability domain (AD) plot

Model applier X | Model profile X

Model profile

Statistical parameters, tables, charts - all the information related to the model.

Overview | Applicability domain

Williams plot with *CONSENSUS-STD* used as a distance to model.

Distance to model

CONSENSUS-STD

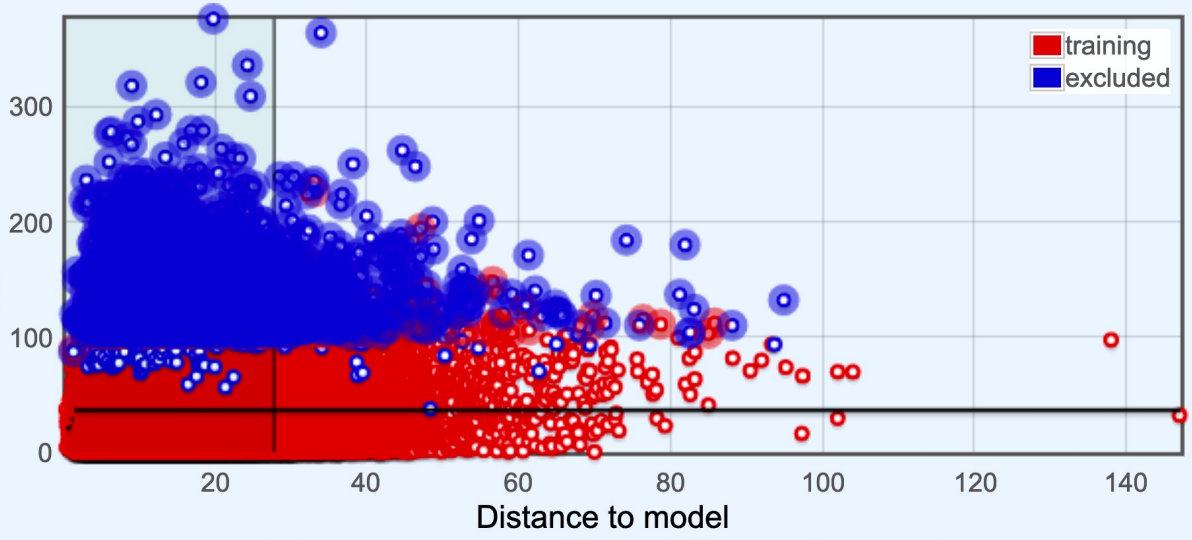
Averaging type

Default

X axis

CONSENSUS-STD

Show negative residuals



4118 records selected; S/N = 3 [unselect][exclude][include][browse]

Select outliers with p-value less or equal than 0.005

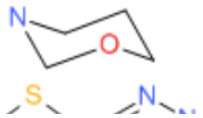
APPLY THE MODEL TO NEW COMPOUNDS

Functional group analysis of pyrolysis and MP data

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 sets. 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 266 — 15

Descriptor	In set 1 (13769 molecules)	In set 2 (275133 molecules)	Enrichment factor	p-Value
Pnictogens Group 15: the nitrogen family **N P As Sb Bi**	13222 (96.0%)	236886 (86.1%)	1.1	1.2E-319
	11839	204774	1.2	1.64E-230

Typical outliers

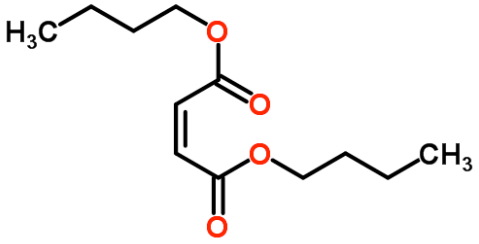
Polymers instead of monomers

Salts (chloride, bromide, etc.)

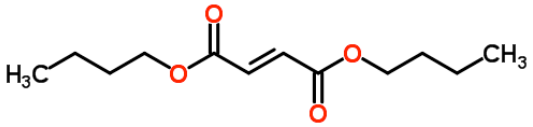
Text mining processing errors (chemical naming, MP errors)

Stereoisomers

Pyrolysis instead of melting point



Dibutyl(2Z)-2-butenedioate, -85 °C



n-Butyl fumarate, -18 °C

Tasks for hackathon

Datasets:

1) BCF Tutorial (training)
2) T. pyriformis (training)
3) Ames challenge (training)
4) Tissue/air set -- multi-task learning

Tasks:

1) Develop models using different methods (descriptor and representation learning)
2) Make consensus model (each method)
3) Compare results

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

Martin Šícho (Uni Leiden)
Guillaume Godin (Firmenich)
Ruud van Deursen (Firmenich)

Michael Sattler (HMGU)

Alexander von Humboldt
Stiftung/Foundation

eco
environmental chemoinformatics
marie curie initial training network

CONCERTREACH
CONCERTING EXPERIMENTAL DATA
AND IN SILICO MODELS FOR REACH

BigChem