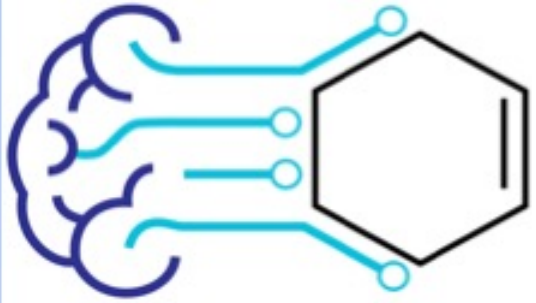


Institute of Structural Biology

OCHEM, openOCHEM and beyond: use of Advanced machIne learning (AI) in Chemistry

Igor V. Tetko

Helmholtz Munich and BIGCHEM GmbH

University of Münster, 12 December 2023

HELMHOLTZ MUNICH

big chem

Agenda

Overview of HMGU

Computational predictions in drug discovery

OCHEM

Examples of models

Consensus modelling

Multitask learning

Representation learning

Reaction predictions

Conclusions

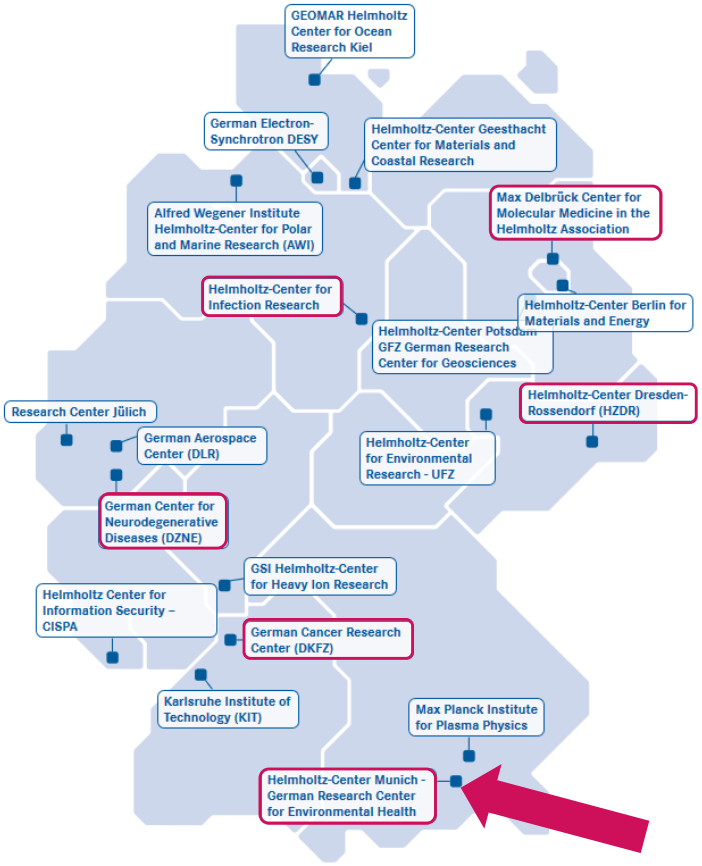
Helmholtz Association – Facts and Figures

Germany´s largest research organization

- 19 research centers
- Budget: 5 Billion €, more than 42.000 staff

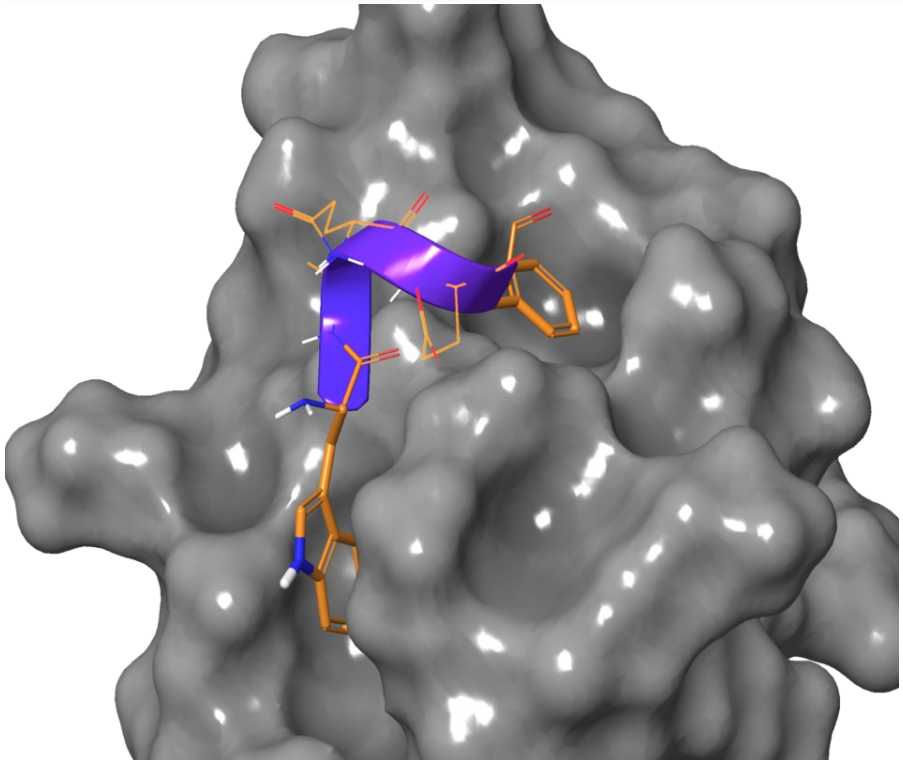
6 Research Fields

6 centers represent the Research Field Health.

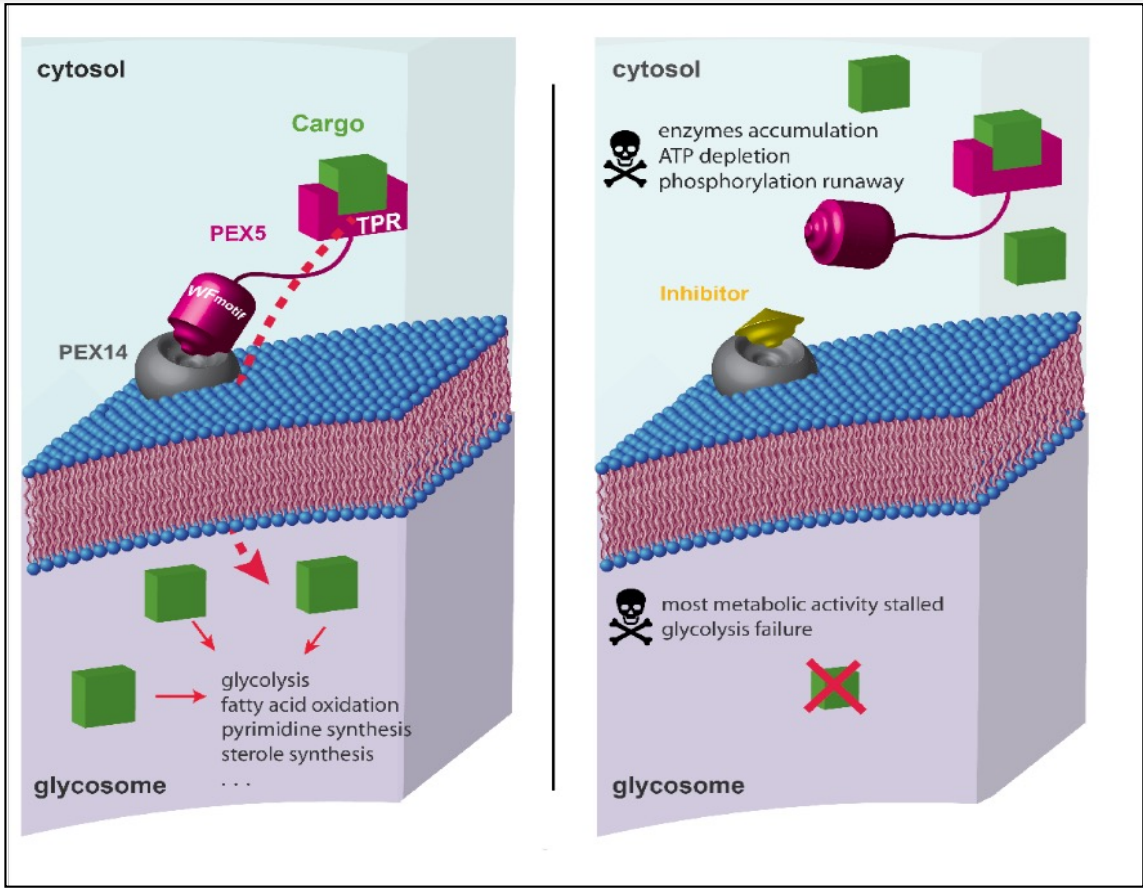


BIGCHEM GmbH is a spin-off of the center

Hypothesized Target for African sleeping sickness



PPI interface

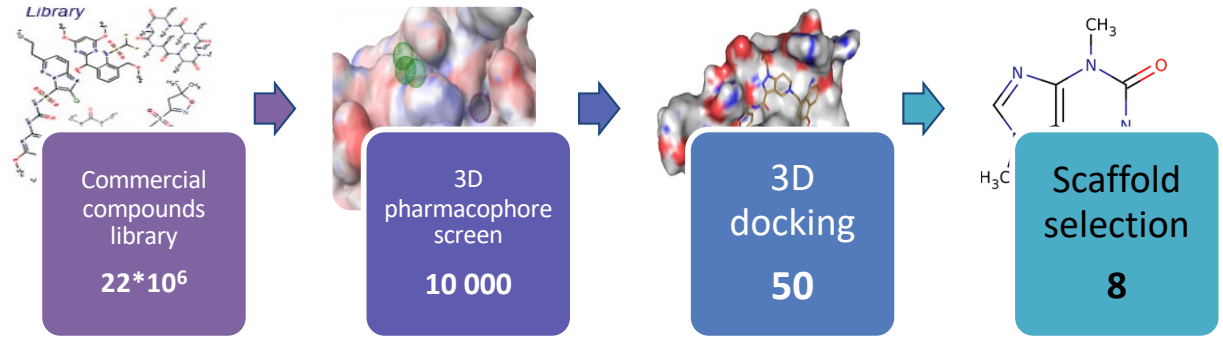


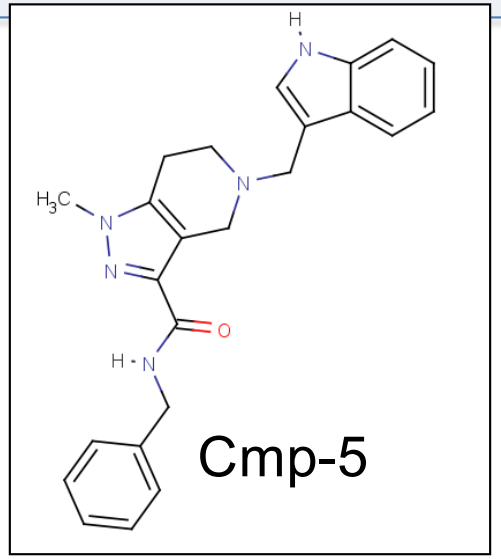
Neufeld, C. … Sattler, M. (2009) EMBO J. 28: 745-754

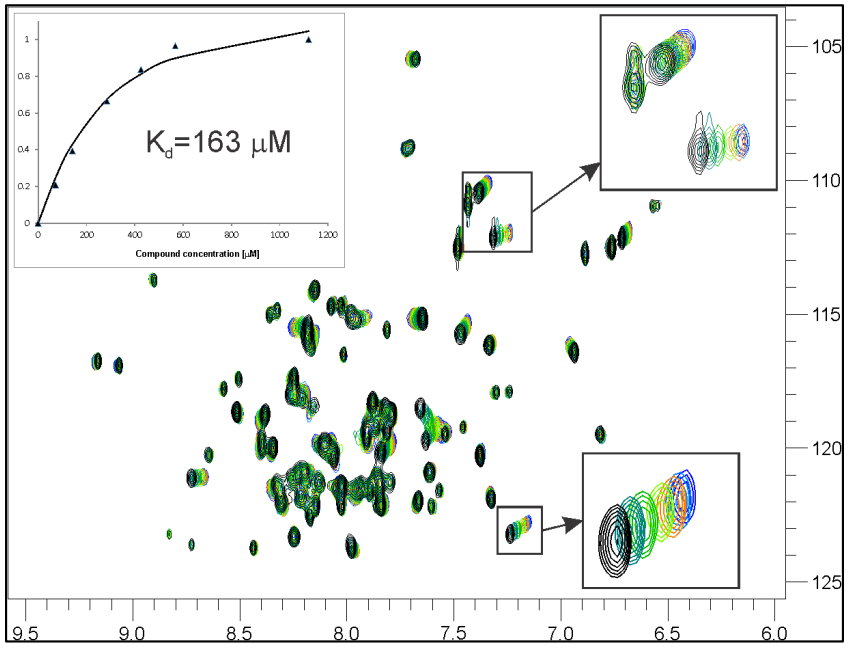
Peroxins Pex14/Pex5 are responsible for transport of glycosomal enzymes from cytoplasm to glycosomes for glucose metabolism

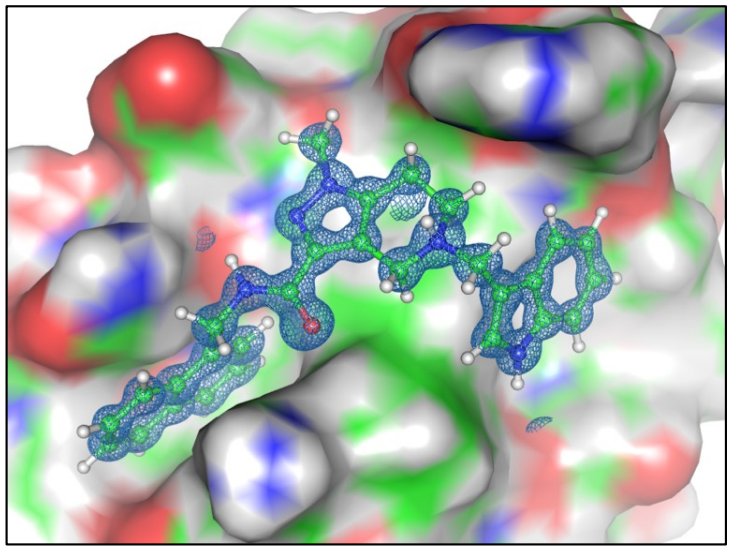
African trypanosomiasis: hit identification

G. Popowicz



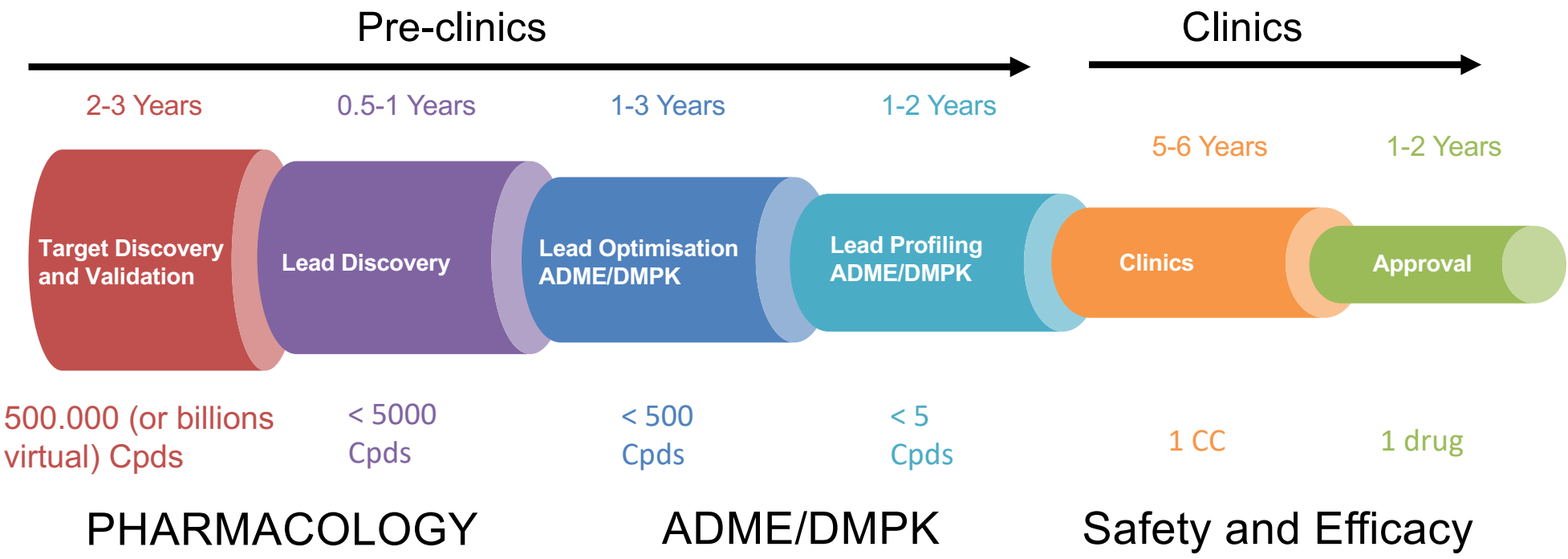






Dawidowski, M., … Popowicz, G. M. Science 2017, 355, 1416-1420.

Traditional Process of Drug Discovery



- Profiling and screening in the virtual space helps to identify the most promising candidates

Slide courtesy of Dr. C. Höfer, Merck

ADMETox filters in Bayer

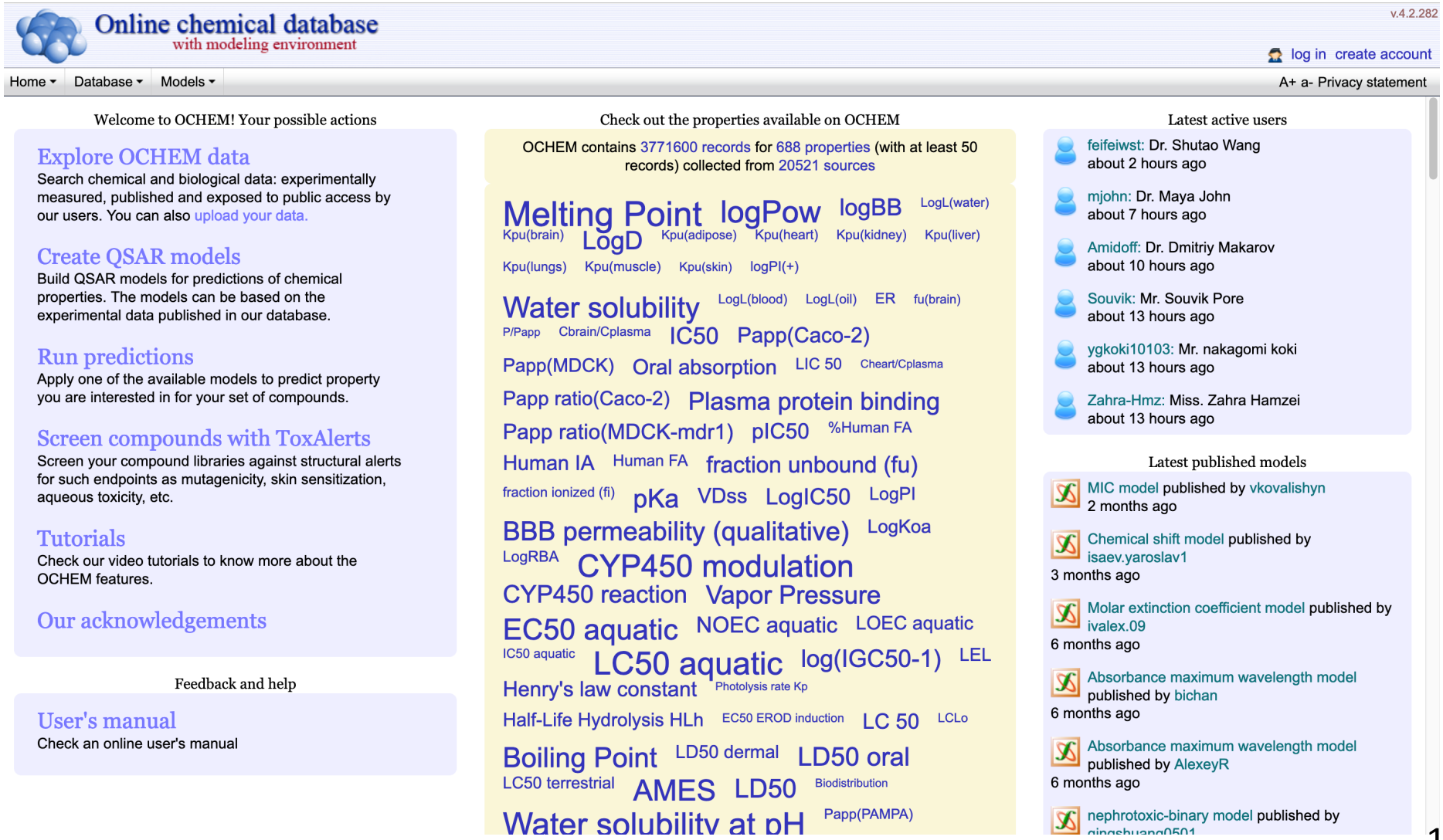
Insufficient quality	First approach	Medium model	Good model	Robust model

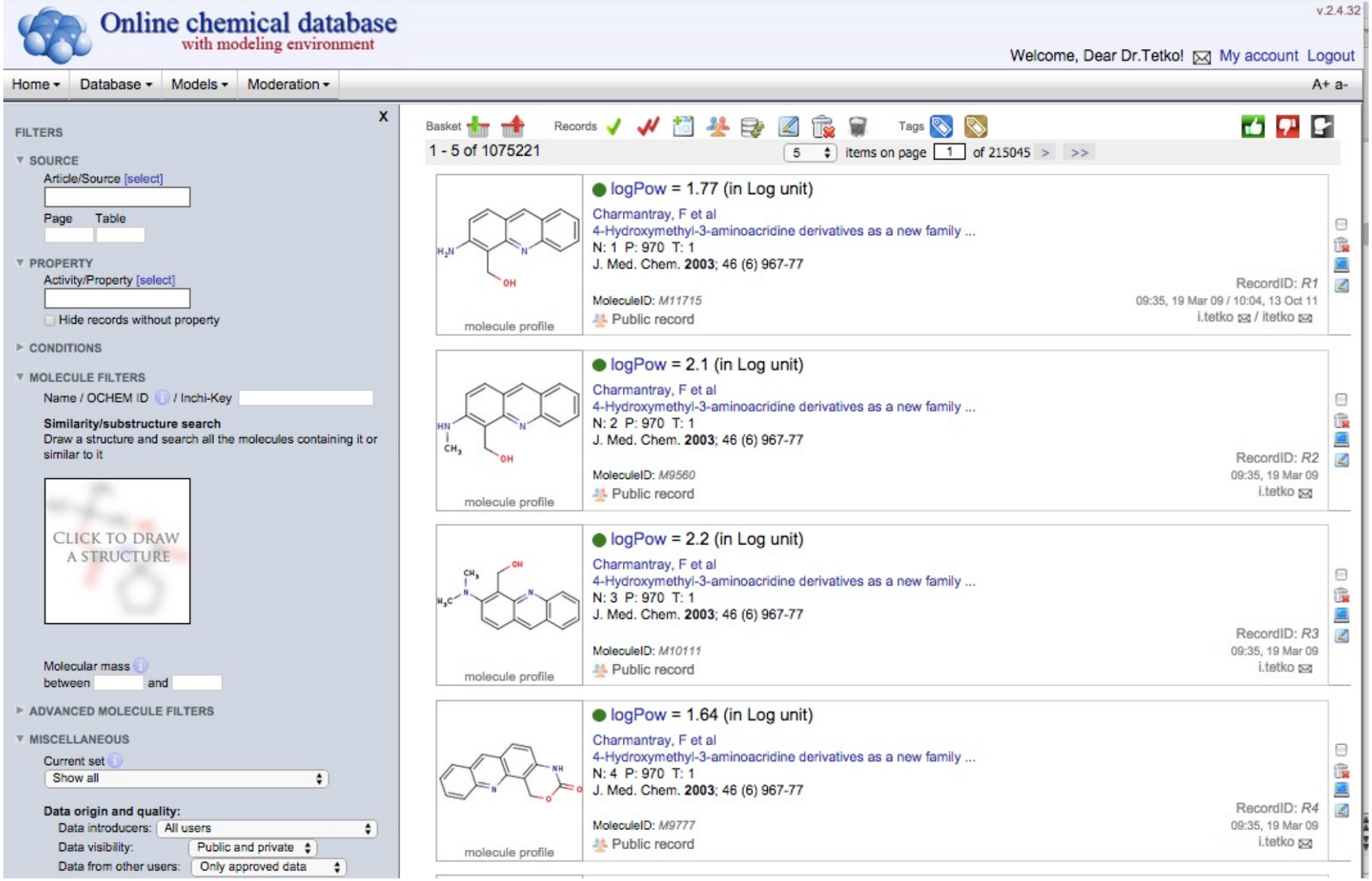
Endpoint		Model type	Data set size	2005	2009	2014	2019	Retraining
Absorption	Caco-2 permeation	C (N)	>10 000			RF	SVR	Weekly
	Caco-2 efflux	C (N)	>10 000			RF	SVR	Weekly
	Bioavailability (rat)	C	~2000				RF	On demand
Distribution	Human serum albumin	N	>30 000			PLS	MTNN	On demand
	Fraction unbound	N	>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

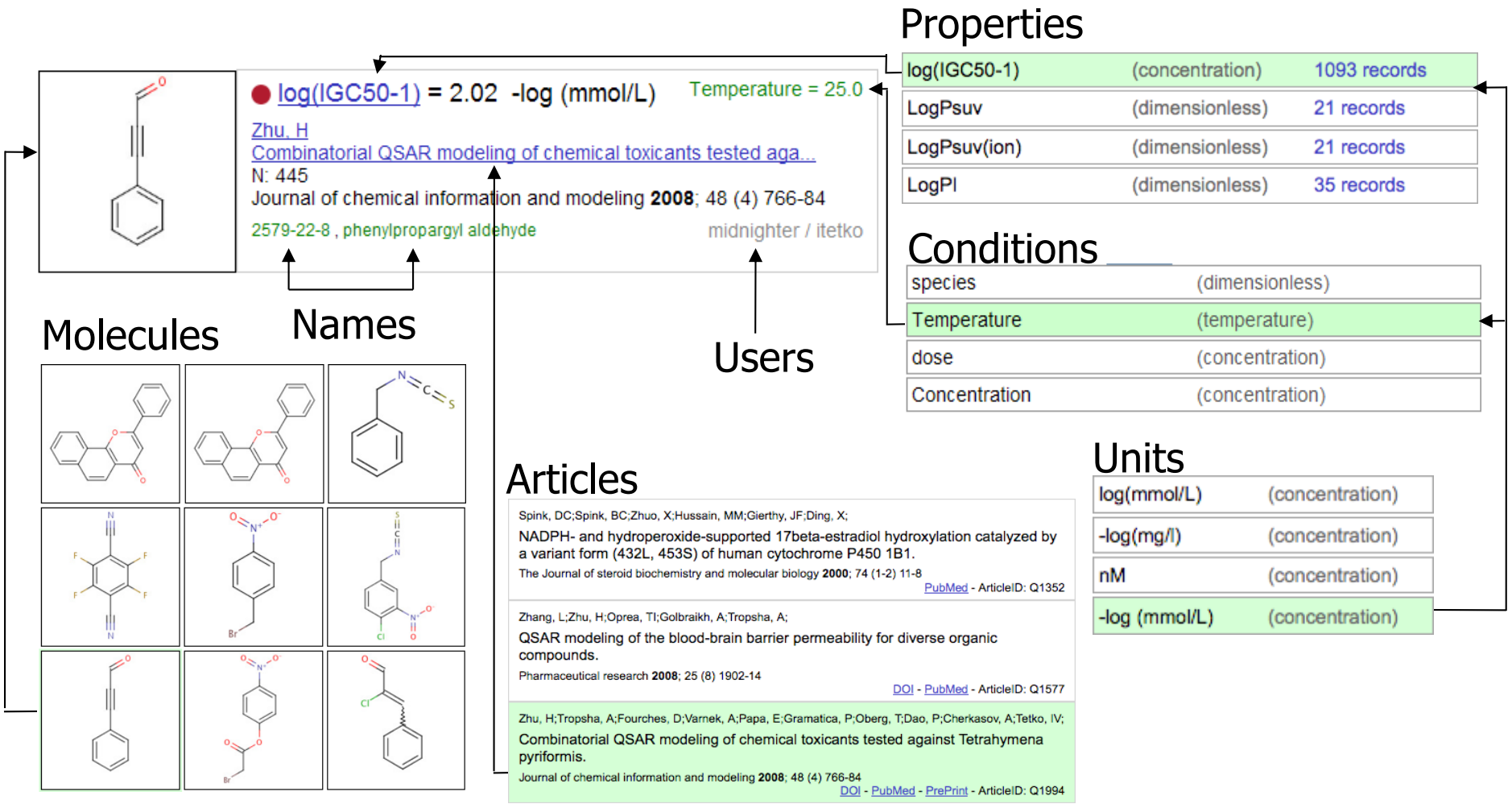
Göller, A.H. et al *Drug Discov. Today* **2020,** *25* (9), 1702-1709.

OCHEM https://ochem.eu

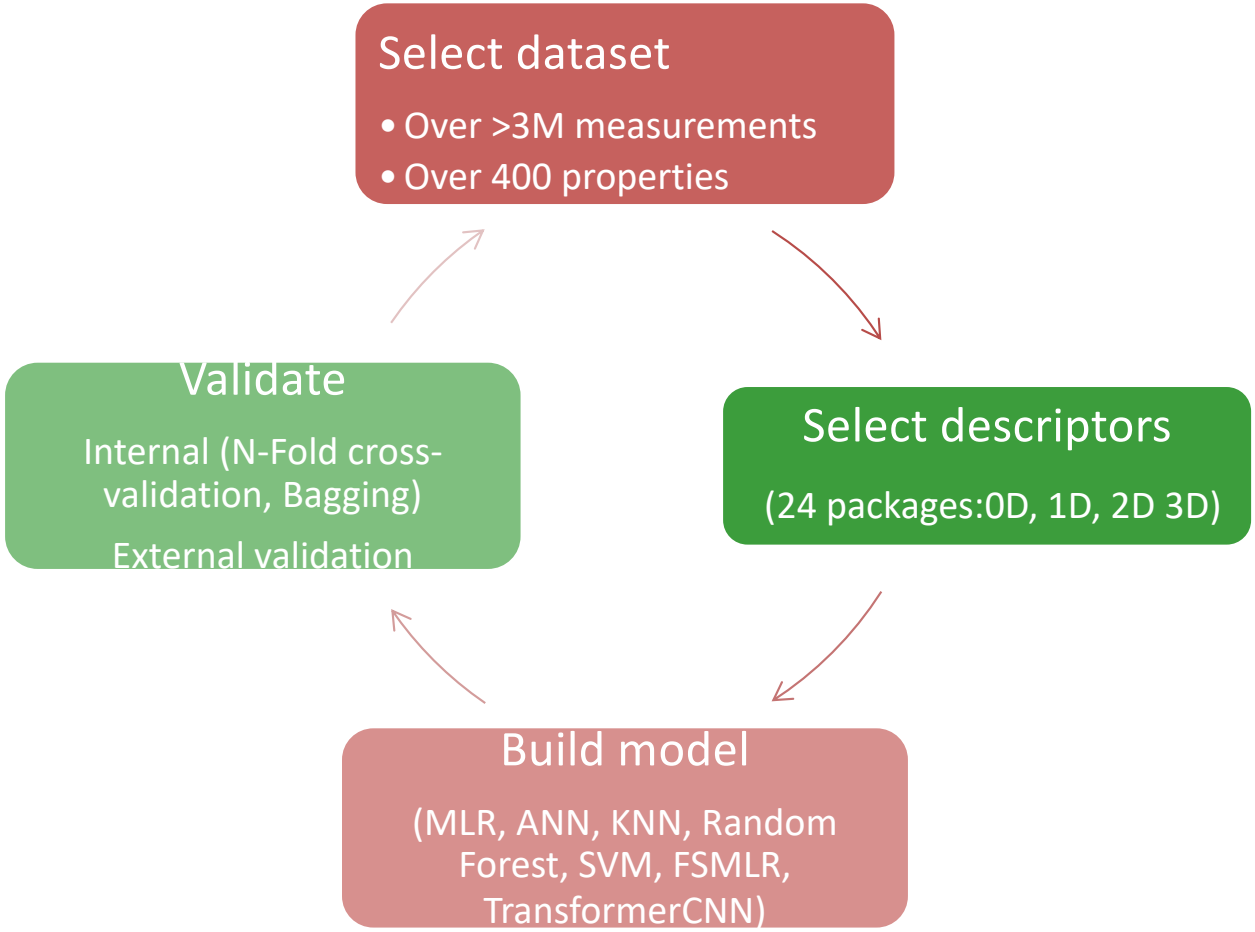



Online chemical database
with modeling environment
v.2.4.32
Welcome, Dear Dr.Tetko! My account Logout
Home Database Models Moderation
A+ a-
FILTERS
SOURCE
Article/Source [select]
Page Table
PROPERTY
Activity/Property [select]
Hide records without property
CONDITIONS
MOLECULE FILTERS
Name / OCHEM ID / Inchi-Key
Similarity/substructure search
Draw a structure and search all the molecules containing it or similar to it
CLICK TO DRAW A STRUCTURE
Molecular mass between and
ADVANCED MOLECULE FILTERS
MISCELLANEOUS
Current set
Show all
Data origin and quality:
Data introducers: All users
Data visibility: Public and private
Data from other users: Only approved data
Basket
Records
Tags
1 - 5 of 1075221
5 items on page 1 of 215045
logPow = 1.77 (in Log unit)
Charmantray, F et al
4-Hydroxymethyl-3-aminoacridine derivatives as a new family ...
N: 1 P: 970 T: 1
J. Med. Chem. 2003; 46 (6) 967-77
MoleculeID: M11715
Public record
RecordID: R1
09:35, 19 Mar 09 / 10:04, 13 Oct 11
i.tetko / itetko
molecule profile
logPow = 2.1 (in Log unit)
Charmantray, F et al
4-Hydroxymethyl-3-aminoacridine derivatives as a new family ...
N: 2 P: 970 T: 1
J. Med. Chem. 2003; 46 (6) 967-77
MoleculeID: M9560
Public record
RecordID: R2
09:35, 19 Mar 09
i.tetko
molecule profile
logPow = 2.2 (in Log unit)
Charmantray, F et al
4-Hydroxymethyl-3-aminoacridine derivatives as a new family ...
N: 3 P: 970 T: 1
J. Med. Chem. 2003; 46 (6) 967-77
MoleculeID: M10111
Public record
RecordID: R3
09:35, 19 Mar 09
i.tetko
molecule profile
logPow = 1.64 (in Log unit)
Charmantray, F et al
4-Hydroxymethyl-3-aminoacridine derivatives as a new family ...
N: 4 P: 970 T: 1
J. Med. Chem. 2003; 46 (6) 967-77
MoleculeID: M9777
Public record
RecordID: R4
09:35, 19 Mar 09
i.tetko
molecule profile

OCHEM Database schema



Modeling iterative workflow



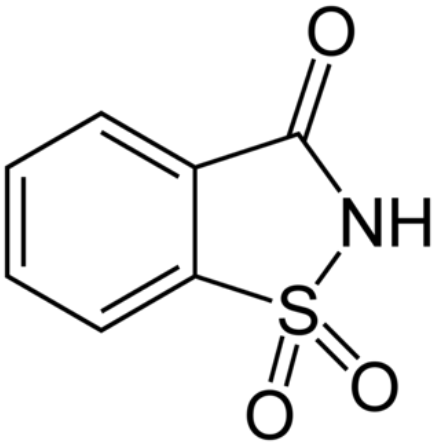
Representation of chemical structures

Saccharin

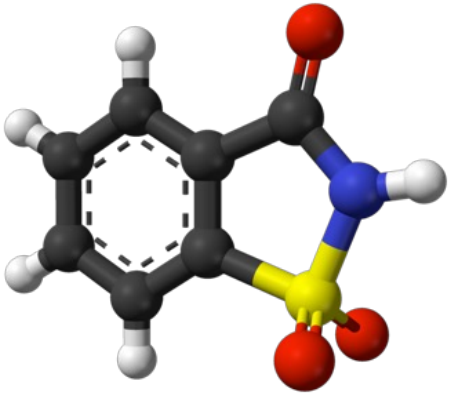
1D

$C_7H_5NO_3S$

2D



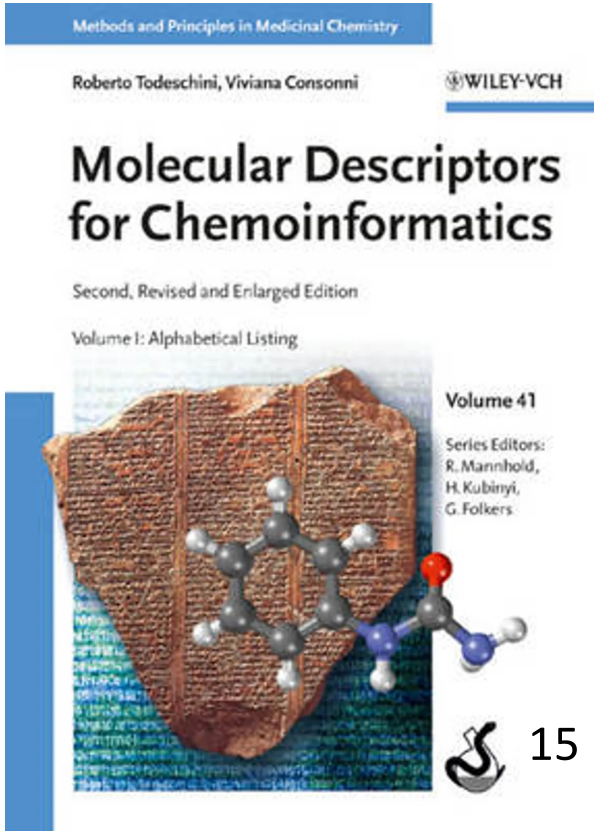
3D



Definition of molecular descriptors

The molecular descriptor is the final result of a logic and mathematical procedure which transforms chemical information encoded within a symbolic representation of a molecule into a useful number, or the result of some standardized experiment.

Roberto Todeschini & Viviana Consonni



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

Select the molecular descriptors

Recommended descriptor types (2D)

- [x] OEState
  - [x] Bonds Indices
  - [ ] Counts only
- [x] ALogPS (2)
- [ ] Mold2 (777)
- [ ] CDDD
- [ ] JPlogP
- [ ] SIRMS
- [ ] ISIDA fragments
- [ ] The in Hashed Atom Pair fingerprint (MAP4)
- [ ] GSFragment (1138)
- [ ] QNPR
- [ ] Multilevel Neighborhoods of Atoms (MNA)
- [ ] Structural alerts (ToxAlerts and Functinal Groups)

Recommended descriptor types (3D)

- [ ] alvaDesc v.2.0.4 (5666/3D)
- [ ] Dragon v. 7 (5270/3D)
- [ ] CDK 2.7.1 descriptors (256/3D)
- [ ] Chemaxon descriptors (499/3D)
- [ ] RDKit descriptors *(3D)*
- [ ] MORDRED descriptors (1826/3D)
- [ ] MOPAC2016 descriptors *(35/3D)*
- [ ] KrakenX descriptors (MOPAC2016 derived)*(124/3D)*
- [ ] PyDescriptor descriptors *(16251/3D)*
- [ ] MERA descriptors *(529/3D)*
- [ ] MERSY descriptors *(42/3D)*
- [ ] 'Inductive' descriptors *(54/3D)*
- [ ] Spectrophores (144/3D)

Predictions by OCHEM's featured models

- [ ] Ames levenberg
- [ ] Toxicity against T. Pyriformis
- [ ] ALogPS 3.0
- [ ] CYP1A2 Estate+ALogPS
- [ ] CYP2C9 Estate+ALogPS
- [ ] CYP2C19 Estate+ALogPS
- [ ] CYP2D6 Estate+ALogPS
- [ ] CYP3A4 Estate+ALogPS
- [ ] Pyrolysis point prediction (best Estate)
- [ ] Melting Point prediction (best Estate)
- [ ] Water solubility model based on logP and Melt
- [ ] ALOGPS 2.1 logP
- [ ] ALOGPS 2.1 logS

- [ ] Outputs of other OCHEM models

Obsolete/Additional descriptor types

- [ ] CDK 2.0 descriptors (256/3D)
- [ ] CDK 1.4.11 descriptors (256/3D)
- [ ] E-state
- [ ] Dragon v. 5.4 (1644/3D)
- [ ] Dragon v. 5.5 (3224/3D)
- [ ] Dragon v. 6 (4885/3D)
- [ ] MOPAC 7.1 descriptors *(25/3D)*

Special descriptors (scaffolds, fingerprints):

- [ ] Chemaxon Scaffolds
- [ ] Silicos-It Scaffolds
- [ ] ECFP Fingerprints
- [ ] MolPrint Fingerprints

Conditions of experiments

- [ ] pH
- [ ] Ionisable

Create a model

Select the training and validation sets, the machine learning method and the validation protocol

Select the training and validation sets:

Training set *(required)*: peptidesregr [details]

Add a validation set

The model will predict this property:

LogD using unit: Log unit

- [ ] Skip model configuration and use the predefined settings

Choose the learning method:

Suggested modeling methods:

- ( ) ASNN: ASsociative Neural Networks doi:10.1007/978-1-60327-101-1_10
- ( ) (New) Attentive FP doi: 10.1021/acs.jmedchem.9b00959
- ( ) ChemProp MPNN for property prediction (GPU) doi:10.1021/acs.jcim.9b00237
- ( ) CNF - Convolutional Neural Network Fingerprint (GPU) doi:10.1007/978-3-030-30493-5_79
- ( ) Transformer-CNF model
- ( ) Consensus model (based on models developed for the same set)
- ( ) DEEPCHEM: several methods from DeepChem (GPU) arXiv:1703.00564
- ( ) (New) DIMENET - Directional Message Passing Neural Network arXiv:2003.03123
- ( ) Deep Learning Consensus Architecture (DLCA) doi:10.1021/acs.jcim.9b00526
- ( ) DNN: Deep Neural Network (GPU) doi:10.1021/acs.jcim.8b00685
- ( ) EAGCNG - Edge Attention based Multi-relational Graph Convolutional Networks (GPU) arXiv:1802.04944
- ( ) FSMLR: Fast Stagewise Multiple Linear Regression doi:10.1134/S0012500807120026
- ( ) GNN - Graph Isomorphism Network (GPU) arXiv:1910.13124
- ( ) KNN: k - Nearest Neighbors
- ( ) KPLS - Kernel Partial Least Squares doi:10.1109/IJCNN.2006.246832
- ( ) LibSVM: grid-search parameter optimisation doi:10.1145/1961189.1961199
- ( ) LSSVMG: Least Squares Support Vector Machine (GPU) doi:10.1023/A:1018628609742
- ( ) MLR: Multiple Linear Regression
- ( ) PLS: Partial Least Squares doi:10.1016/S0169-7439(01)00155-1
- ( ) RFR: Random Forest regression and classification doi:10.1023/A:1010933404324
- ( ) Transformer-CNN - Transformer Convolutional Neural Network (GPU) doi:10.1186/s13321-020-00423-w
- ( ) Transformer-CNNi - faster Transformer-CNN (GPU) doi:10.1186/s13321-020-00423-w
- ( ) WEKA-J48: Weka C4.5 decision trees, only classification - use with bagging doi:10.1145/1656274.1656278
- ( ) WEKA-RF: Random Forest, only classification doi:10.1023/A:1010933404324
- ( ) XGBoost: Scalable and Flexible Gradient Boosting doi:10.1145/2939672.2939785

Model validation

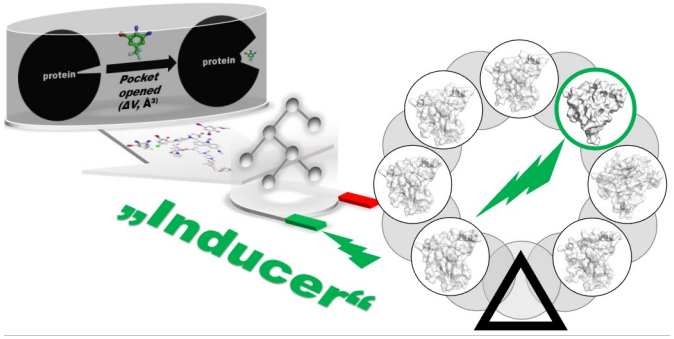
Validation method: N-Fold cross-validation

Number of folds: 5

- [ ] Stratified cross-validation (classification only)
- [ ] Treat each record as a new molecule

You can create a model from template: import an XML model template or use another model as a template

Examples of recent studies using OCHEM

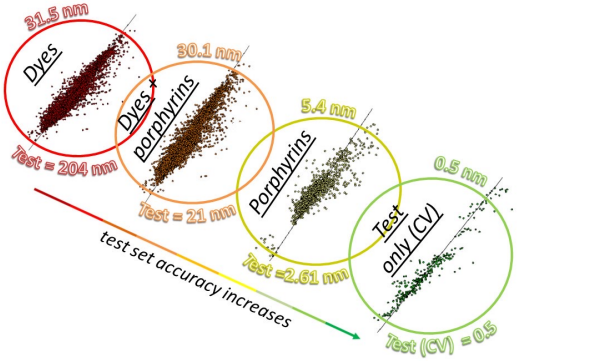


informatics
MDPI

Article

What Features of Ligands Are Relevant to the Opening of Cryptic Pockets in Drug Targets?

Zhonghua Xia [1], Pavel Karpov [1], Grzegorz Popowicz [1], Michael Sattler [1,2] and Igor V. Tetko [1,3,*]

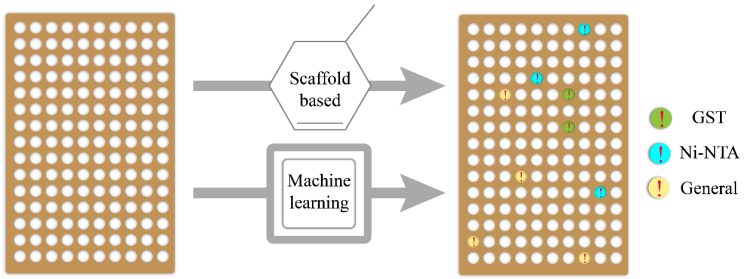


International Journal of Molecular Sciences
MDPI

Article

More Is Not Always Better: Local Models Provide Accurate Predictions of Spectral Properties of Porphyrins

Aleksey I. Rusanov [1], Olga A. Dmitrieva [1], Nugzar Zh. Mamardashvili [1] and Igor V. Tetko [1,2,3,*]



Research Article
www.molinf.com
molecular informatics
doi.org/10.1002/minf.202100151

Highly Accurate Filters to Flag Frequent Hitters in AlphaScreen Assays by Suggesting their Mechanism

Dipan Ghosh,[a] Uwe Koch,[a] Kamyar Hadian,[b] Michael Sattler,[c, d] and Igor V. Tetko*[d, e, f]

CATMoS: Collaborative Acute Toxicity Modeling Suite

Kamel Mansouri,[1,41] Agnes L. Karmaus,[1] Jeremy Fitzpatrick,[2] Grace Patlewicz,[3] Prachi Pradeep,[3,4] Domenico Alberga,[5] Nathalie Alepee,[6] Timothy E.H. Allen, [7] Dave Allen,[1] Vinicius M. Alves,[8,9] Carolina H. Andrade,[9] Tyler R. Auernhammer,[10] Davide Ballabio,[11] Shannon Bell,[1] Emilio Benfenati,[12] Sudin Bhattacharya,[13] Joyce V. Bastos,[9] Stephen Boyd,[14] J.B. Brown,[15] Stephen J. Capuzzi,[8] Yaroslav Chushak,[16,17] Heather Ciallella,[18] Alex M. Clark,[19] Viviana Consonni,[11] Pankaj R. Daga,[20] Sean Ekins,[19] Sherif Farag,[8] Maxim Fedorov,[21] Denis Fourches,[22,23] Domenico Gadaleta,[12] Feng Gao,[14] Jeffery M. Gearhart,[16,17] Garett Goh,[24] Jonathan M. Goodman,[7] Francesca Grisoni,[11] Christopher M. Grulke,[3] Thomas Hartung,[25] Matthew Hirn,[26] Pavel Karpov,[27] Alexandru Korotcov,[28] Giovanna J. Lavado,[12] Michael Lawless,[20] Xinhao Li,[22] Thomas Luechtefeld,[25] Filippo Lunghini,[29] Giuseppe F. Mangiatordi,[5] Gilles Marcou,[29] Dan Marsh,[25] Todd Martin,[30] Andrea Mauri,[31] Eugene N. Muratov,[8,9] Glenn J. Myatt,[32] Dac-Trung Nguyen,[33] Orazio Nicolotti,[5] Reine Note,[6] Paritosh Pande,[24] Amanda K. Parks,[10] Tyler Peryea,[33] Ahsan H. Polash,[15] Robert Rallo,[24] Alessandra Roncaglioni,[12] Craig Rowlands,[25] Patricia Ruiz,[34] Daniel P. Russo,[18] Ahmed Sayed,[35] Risa Sayre,[3,4] Timothy Sheils,[33] Charles Siegel,[24] Arthur C. Silva,[9] Anton Simeonov,[33] Sergey Sosnin,[21] Noel Southall,[33] Judy Strickland,[1] Yun Tang,[36] Brian Teppen,[14] Igor V. Tetko,[27,37] Dennis Thomas,[24] Valery Tkachenko,[28] Roberto Todeschini,[11] Cosimo Toma,[12] Ignacio Tripodi,[38] Daniela Trisciuzzi,[5] Alexander Tropsha,[8] Alexandre Varnek,[29] Kristijan Vukovic,[12] Zhongyu Wang,[39] Liguo Wang,[39] Katrina M. Waters,[24] Andrew J. Wedlake,[7] Sanjeeva J. Wijeyesakere,[10] Dan Wilson,[10] Zijun Xiao,[39] Hongbin Yang,[36] Gergely Zahoranszky-Kohalmi,[33] Alexey V. Zakharov,[33] Fagen F. Zhang,[10] Zhen Zhang,[40] Tongan Zhao,[33] Hao Zhu,[18] Kimberley M. Zorn,[19] Warren Casey,[41] and Nicole C. Kleinstreuer[41]

[1]Integrated Laboratory Systems, LLC, Morrisville, North Carolina, USA
[2]ScitoVation, Research Triangle Park, North Carolina, USA
[3]Center for Computational Toxicology and Exposure, U.S. Environmental Protection Agency, Research Triangle Park, North Carolina, USA
[4]Oak Ridge Institute for Science and Education (ORISE) Research Participation Program, U.S. Environmental Protection Agency, Research Triangle Park, North Carolina, USA
[5]Dipartimento di Farmacia-Scienze del Farmaco, Università degli Studi di Bari “Aldo Moro”, Bari, Italy
[6]L’Oréal Research & Innovation, Aulnay-sous-Bois, France

New low-molecular bioregulators as effective agents against multi-drug resistant Acinetobacter baumannii clinical isolate

Received: 18 October 2019 | Revised: 27 January 2020 | Accepted: 3 March 2020
DOI: 10.1111/cbdd.13678

RESEARCH ARTICLE

WILEY

In silico and in vitro studies of a number PILs as new antibacterials against MDR clinical isolate *Acinetobacter baumannii*

Maria M. Trush[1] | Vasyl Kovalishyn[1] | Diana Hodyna[1] | Olexandr V. Golovchenko[1] | Svitlana Chumachenko[1] | Igor V. Tetko[2,3] | Volodymyr S. Brovarets[1] | Larysa Metelytsia[1]

Figure 1. Inhibition zone diameters of the six studied PILs (content on a disk, 1.25 μmoles) of an MDR clinical isolate of *A. baumannii* on agar plates.

antibiotics



MDPI

Article

Theoretical and Experimental Studies of Phosphonium Ionic Liquids as Potential Antibacterials of MDR *Acinetobacter baumannii*

Larysa O. Metelytsia [1], Diana M. Hodyna [1], Ivan V. Semenyuta [1], Vasyl V. Kovalishyn [1], Sergiy P. Rogalsky [1], Kateryna Yu Derevianko [1], Volodymyr S. Brovarets [1] and Igor V. Tetko [2,3,*]

Overview | Applicability domain

Model name: M4_Consensus AcinBaum_Class - 334799 , published in In silico and in vitro studies of a number PILs as new antibacterials against MDR clinical isolate Acinetobacter baumannii.
Public ID is 783

Predicted property: **AcinBaum_Class** modeled in CLASS
Training method: Consensus

Data Set	#	Accuracy	Balanced Accuracy	MCC	AUC
Training set: A_Baumanii_Set1 (training)	210 records	83% ± 2.0	82% ± 3.0	0.66 ± 0.05	0.9 ± 0.02
Test set: A_Baumanii_Set1 (test) [x]	53 records	83% ± 5.0	83% ± 5.0	0.7 ± 0.1	0.92 ± 0.04

Show ROC curves

Real↓/Predicted→	inactive	active	Hit rate
inactive	69	21	0.77
active	14	106	0.88
Precision	0.83	0.83	
Training (Original)			

Real↓/Predicted→	inactive	active	Hit rate
inactive	22	4	0.85
active	5	22	0.81
Precision	0.81	0.85	
Test (Original)			

Journal of Molecular Liquids 344 (2021) 117722

Contents lists available at ScienceDirect

Journal of Molecular Liquids

journal homepage: www.elsevier.com/locate/molliq

Beware of proper validation of models for ionic Liquids!

D.M. Makarov [a,*], Yu.A. Fadeeva [a], L.E. Shmukler [a], I.V. Tetko [a,b,c]

[a] G. A. Krestov Institute of Solution Chemistry of the Russian Academy of Sciences, Ivanovo, Russia
[b] Institute of Structural Biology, Helmholtz Zentrum München-Research Center for Environmental Health (GmbH), Ingolstädter Landstraße 1, 85764 Neuherberg, Germany
[c] BIGCHEM GmbH, Valerystr. 49, 85716 Unterschleißheim, Germany

Title:	**Beware of proper validation of models for Ionic Liquids!**
Authors:	Makarov, D.M.; Fadeeva, Yu.A.; Shmukler, L.E.; Tetko, I.V.;
Journal reference:	Journal of Molecular Liquids, **2021**; 344 (); 117722
Internal identifier:	A135195

https://ochem.eu/article/135195

Data and models

This article is referenced from 249 experimental records This article is connected to **3 predictive model(s)**:

Melting Point for Ionic Liquids (TRANSNNI)	trained using the dataset MP ILs FULL set (view dataset profile or export the dataset) validated using dataset MP PILs (view dataset profile or export the dataset) validated using dataset MP_test set_Version2 (view dataset profile or export the dataset)
Melting Point for Ionic Liquids (RFR based on KrakenX descriptors)	trained using the dataset MP ILs FULL set (view dataset profile or export the dataset)
Melting Point for Ionic Liquids (RFR, based on ECFP4 descriptors)	trained using the dataset MP ILs FULL set (view dataset profile or export the dataset) validated using dataset MP_test set_Version2 (view dataset profile or export the dataset)

Open article editor | Attach files to this article

OCHEM modelling

- Comprehensive modeling
- Multitask learning (up to 100 properties)
- >20 descriptors blocks
- GPU + CPU modern methods Supports models
  - >1,000,000 compounds
  - >1,000 servers
  - up to 1GB in size (Java limit)
- Model private/publishing
- Conditions, external descriptors
- ToxAlerts
- Consensus models

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

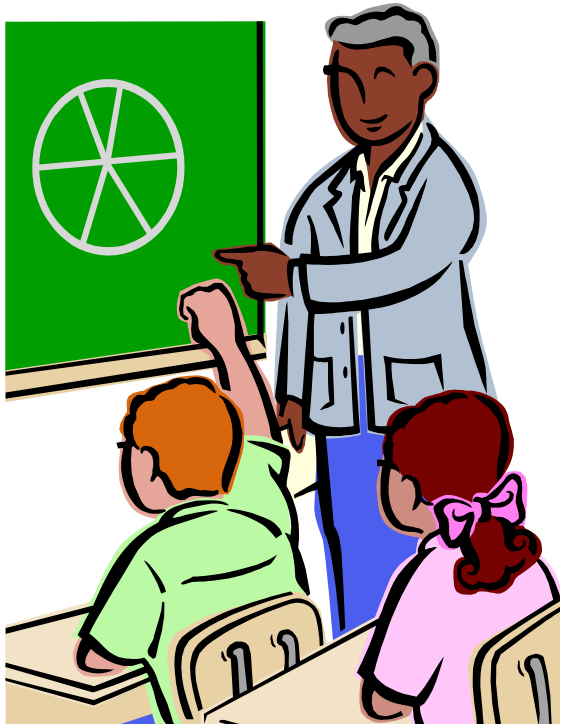
* Sparse format, DOI:10.1186/s13321-016-0113-y

Consensus modelling

Best method(s) are defined

Average prediction of models is used

The consensus prediction is more accurate and stable



www.epa.gov/ncct/challenges.html

EPA United States Environmental Protection Agency

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

ALL EPA THIS AREA Advanced Search SEARCH

Computational Toxicology Research

Contact Us

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

Key Links

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

ToxCast Chemical Data Challenges and Release

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

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

Key Links

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

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

Model to predict Lowest Effect Level (training set)

Predicted property: LEL
Training set: LEL (training)
Duplicate models: start

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

	ASNN (CHEMAXON)
Adriana (3D by Adriana) 3D:corina	0.93
CDK (cons,topol,geom,elect,hybr) 3D:corina	0.93
ChemaxonDescriptors (pH 0 - 14:1) 3D:corina	0.93
Dragon6 (3D blocks: (only) 1-29) 3D:corina	0.93
Fragmentor (length: 2-4)	0.98
GSFrag (F + L)	0.97
InductiveDescriptors 3D:corina	0.94
Mera, Mersy 3D:corina	0.93
OEstate	0.96
QNPR (length:1 - 3)	0.95

	Consensus:AVERAGE (CHEMAXON)
Final EPA ToxCast challenge model	0.88

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

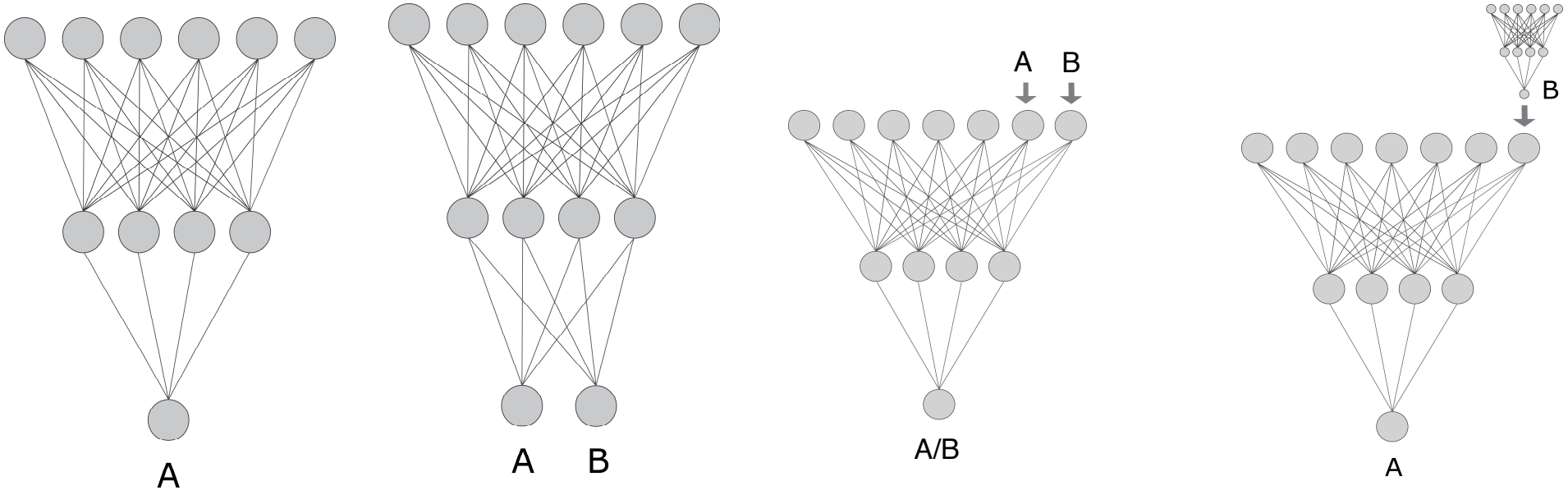
The 2014 Tox21 data challenge is designed to help scientists understand the potential of the chemicals and compounds being tested through the Toxicology in the 21st Century initiative to disrupt biological pathways in ways that may result in toxic effects.

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



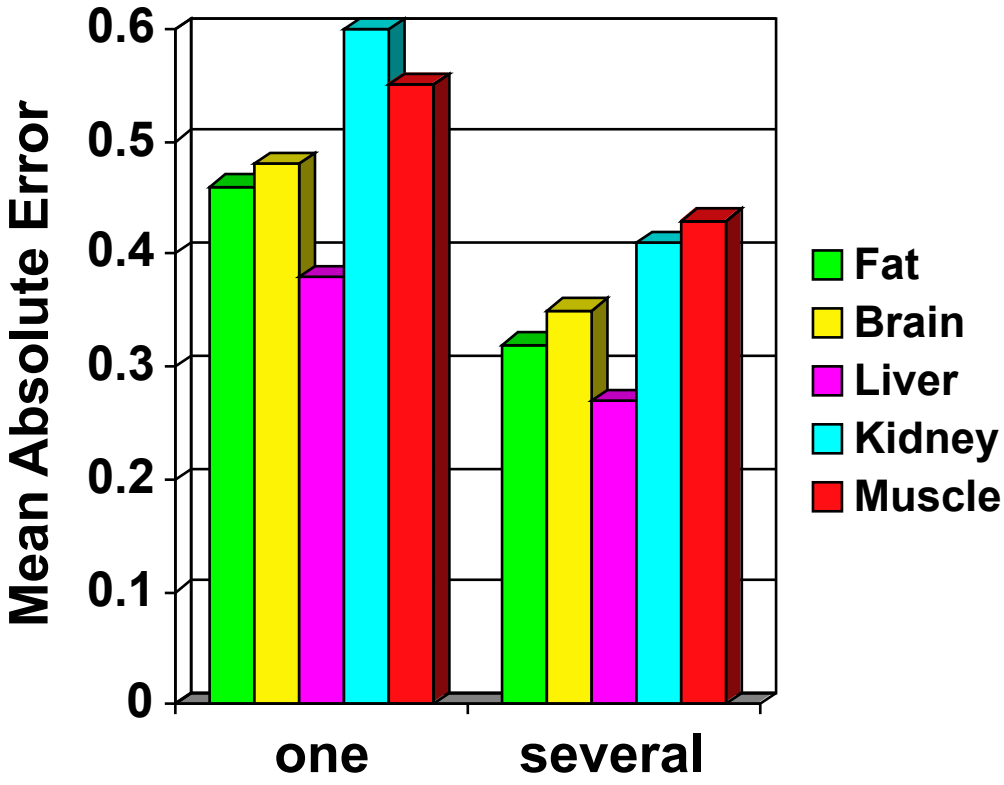
Multi-task learning

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.

Analysis of toxicity of chemical compounds

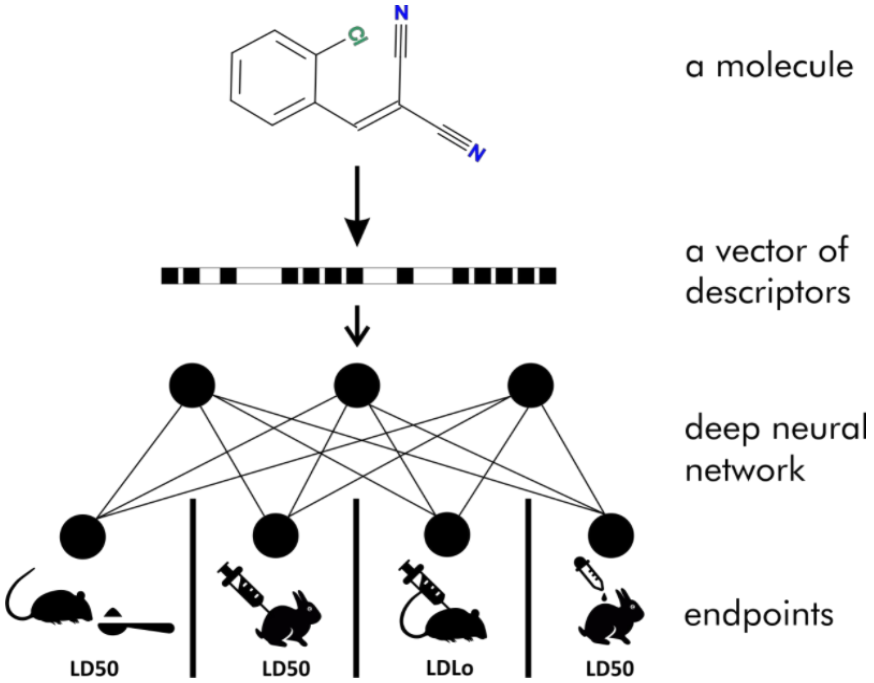
{ , , ...} { , , ,...} {LD50, TDLo, LDLo}

LDLo

129 142 toxicity measurements from RTECS*

87 064 unique molecular structures

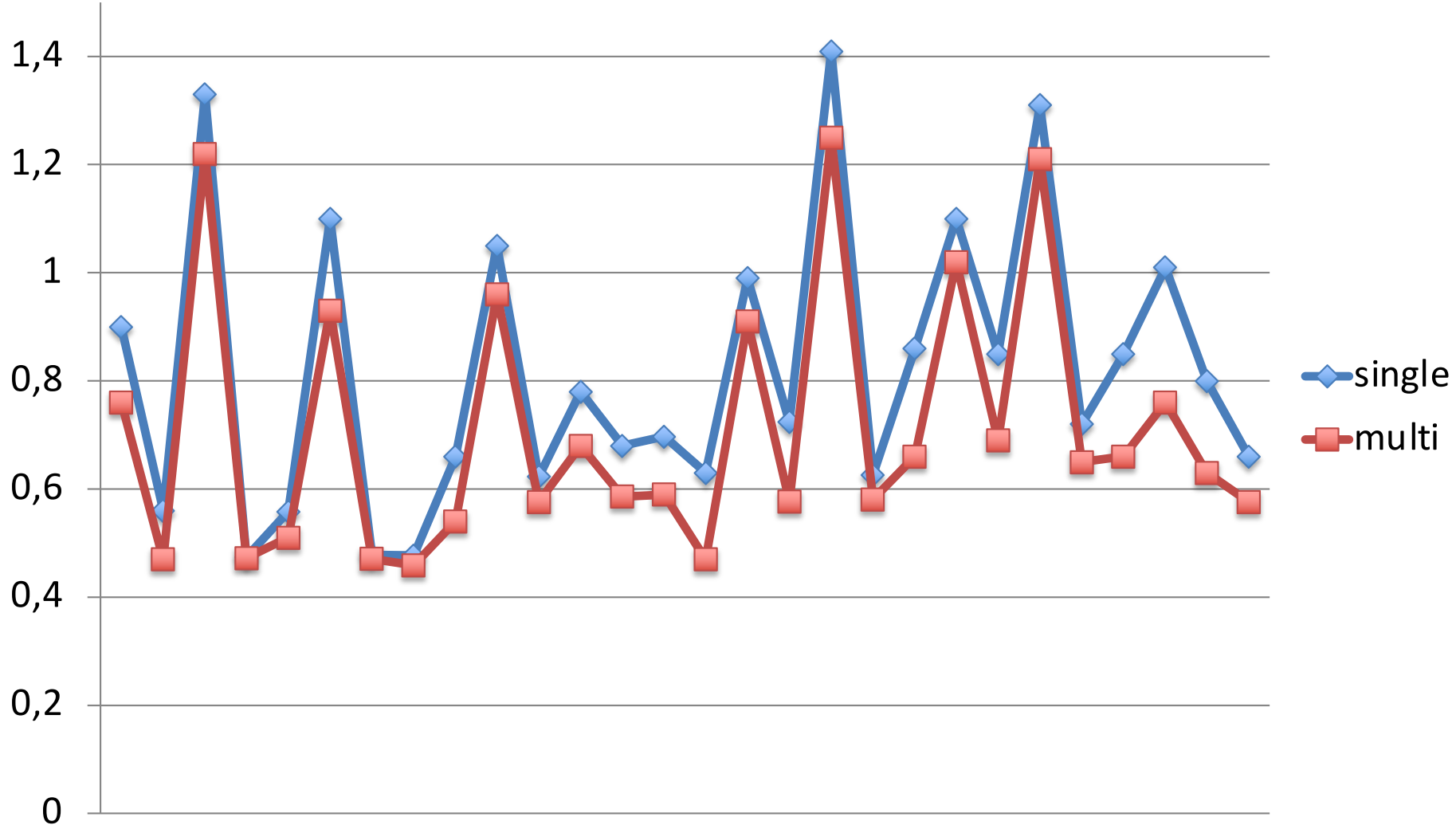
29 toxicity endpoints



*RTECS: Registry of Toxic Effects of Chemical Substances

Sosnin, S. et al. *J. Chem. Inf. Model.*, **2018**, 59:1062-1072.

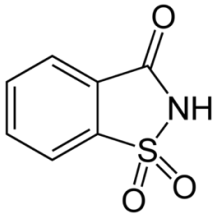
RMSE for different toxicities using CDK descriptors and DNN



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.

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

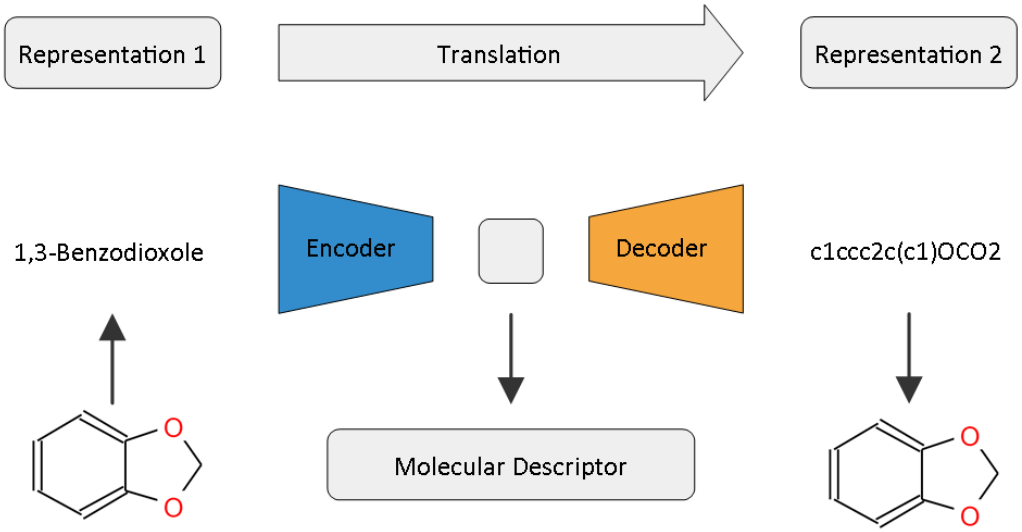
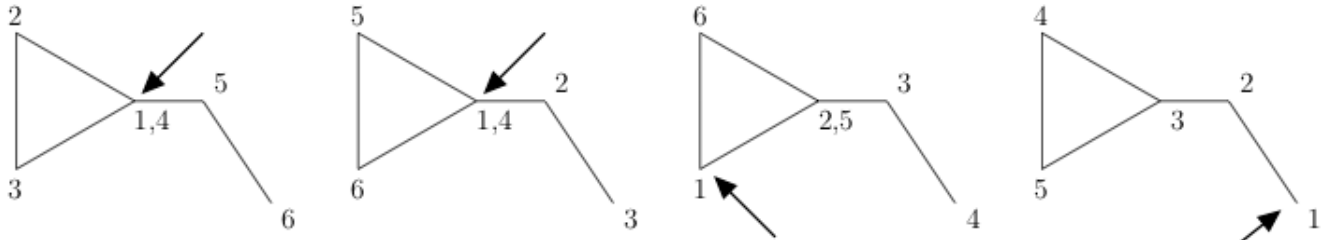


Image augmentation

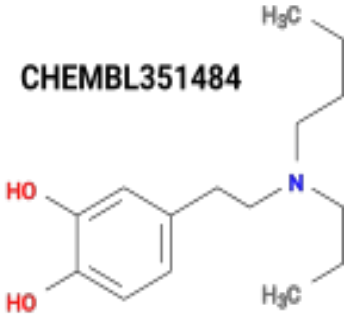
https://github.com/aleju/imgaug

Machine Learning to canonise chemical structures



C1CC1CC C1(CC)CC1 C1C(CC)C1 CCC1CC1

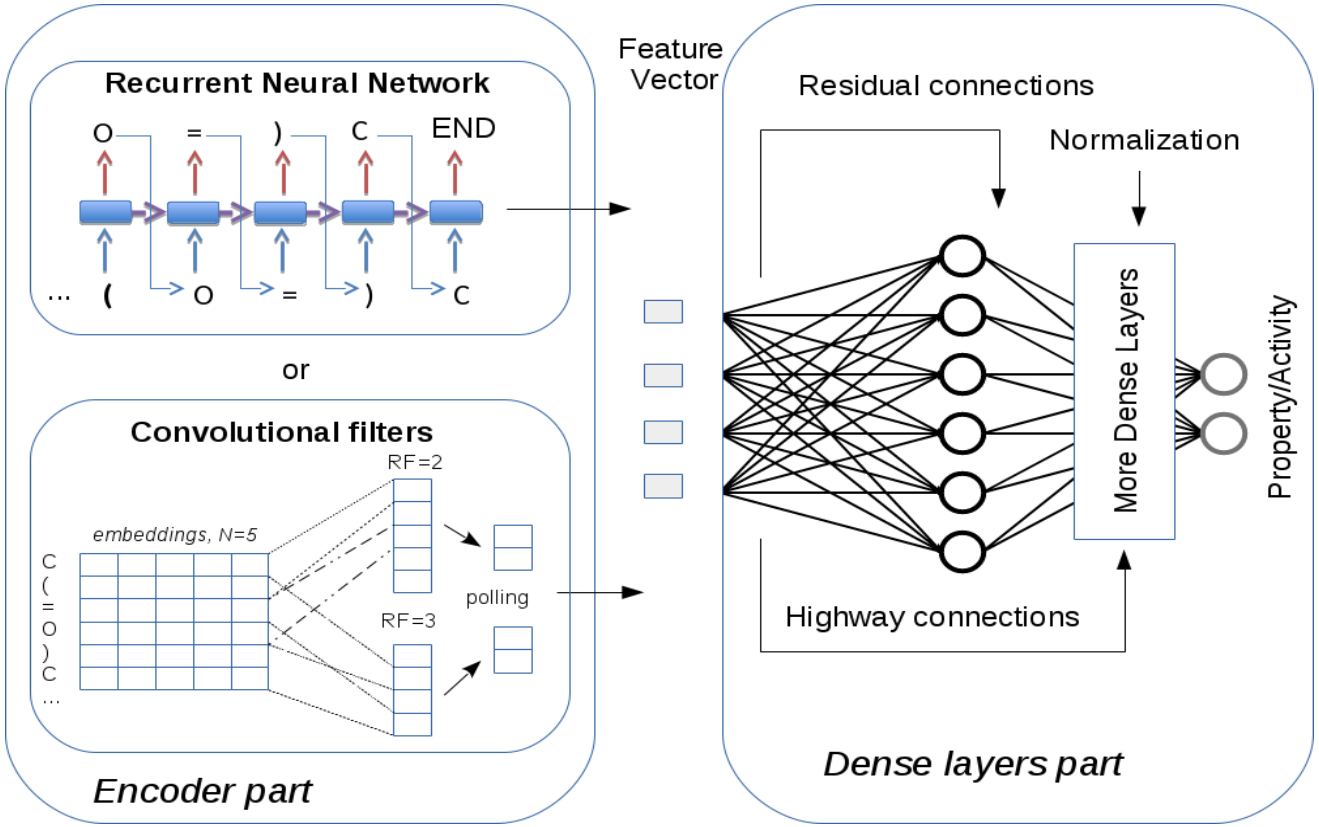
non-canonical		canonical
c1c(ccc(O)c1O)CCN(CCCC)CCC	>>	CCCCN(CCc1ccc(c(c1)O)O)CCC
c1(ccc(c(c1)O)O)CCN(CCC)CCCC	>>	CCCCN(CCc1ccc(c(c1)O)O)CCC
c1c(c(cc(c1)CCN(CCCC)CCC)O)O	>>	CCCCN(CCc1ccc(c(c1)O)O)CCC
c1(CCN(CCCC)CCC)ccc(O)c(O)c1	>>	CCCCN(CCc1ccc(c(c1)O)O)CCC
CCCCN(CCc1cc(c(cc1)O)O)CCC	>>	CCCCN(CCc1ccc(c(c1)O)O)CCC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

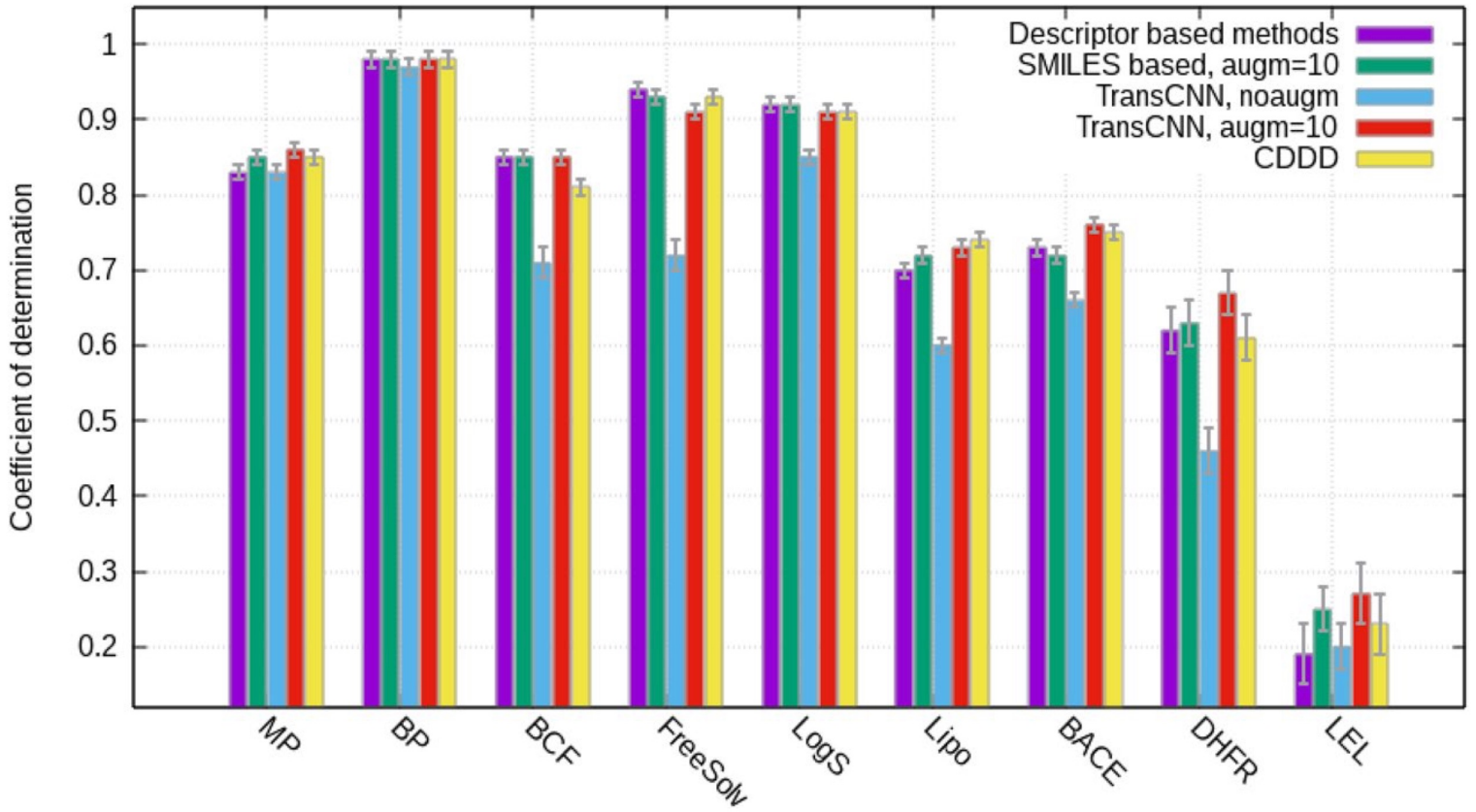
Machine Learning directly from chemical structures



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.

Winning model: OCHEM-generated consensus model

Andrea Kopp

SLAS Europe 2023

25.05.2023

HELMHOLTZ MUNICH

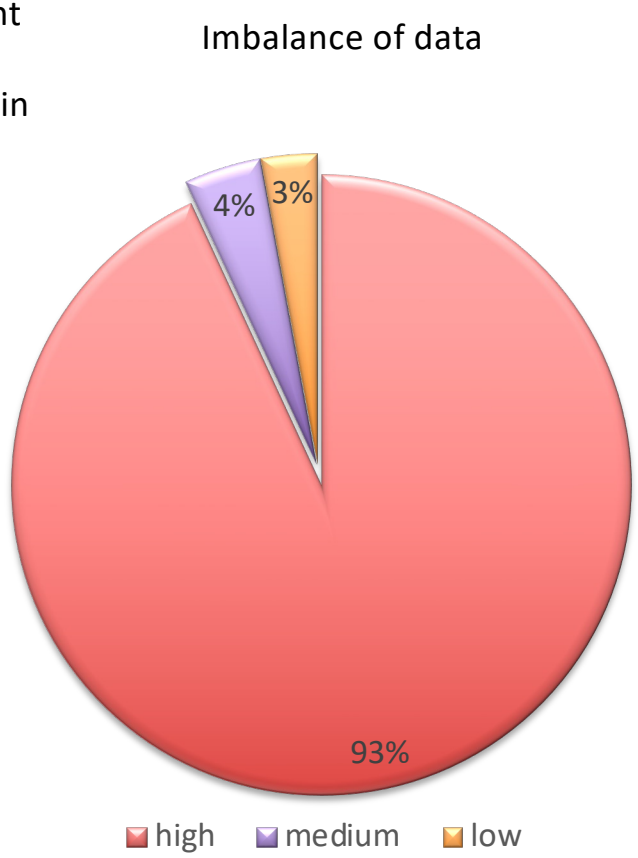
Team of Igor Tetko with Peter Hartog, Martin Šícho and Guillaume Godin

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

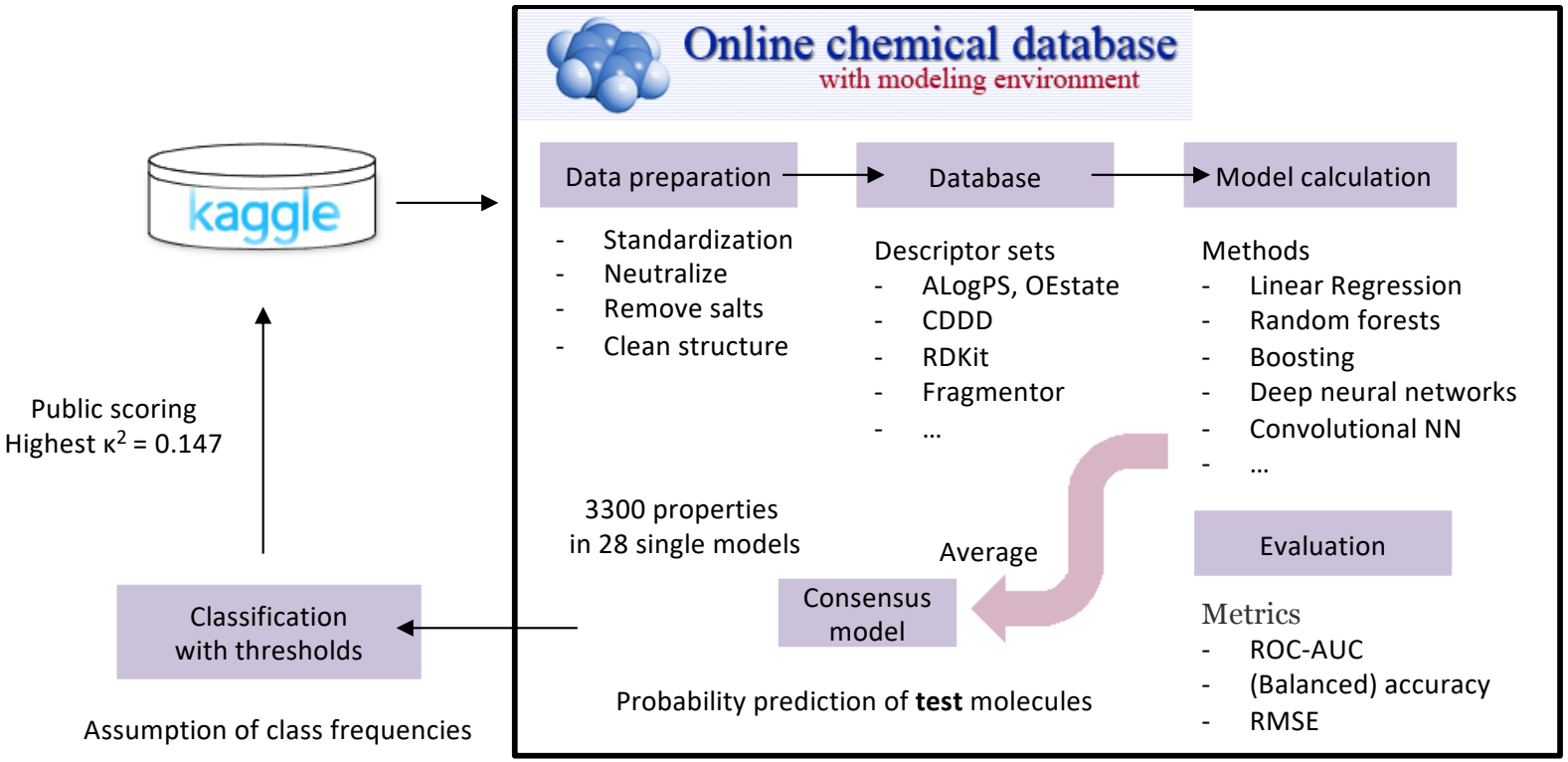
Kopp at al, DOI: 10.26434/chemrxiv-2023-p8qcv

Challenge set-up

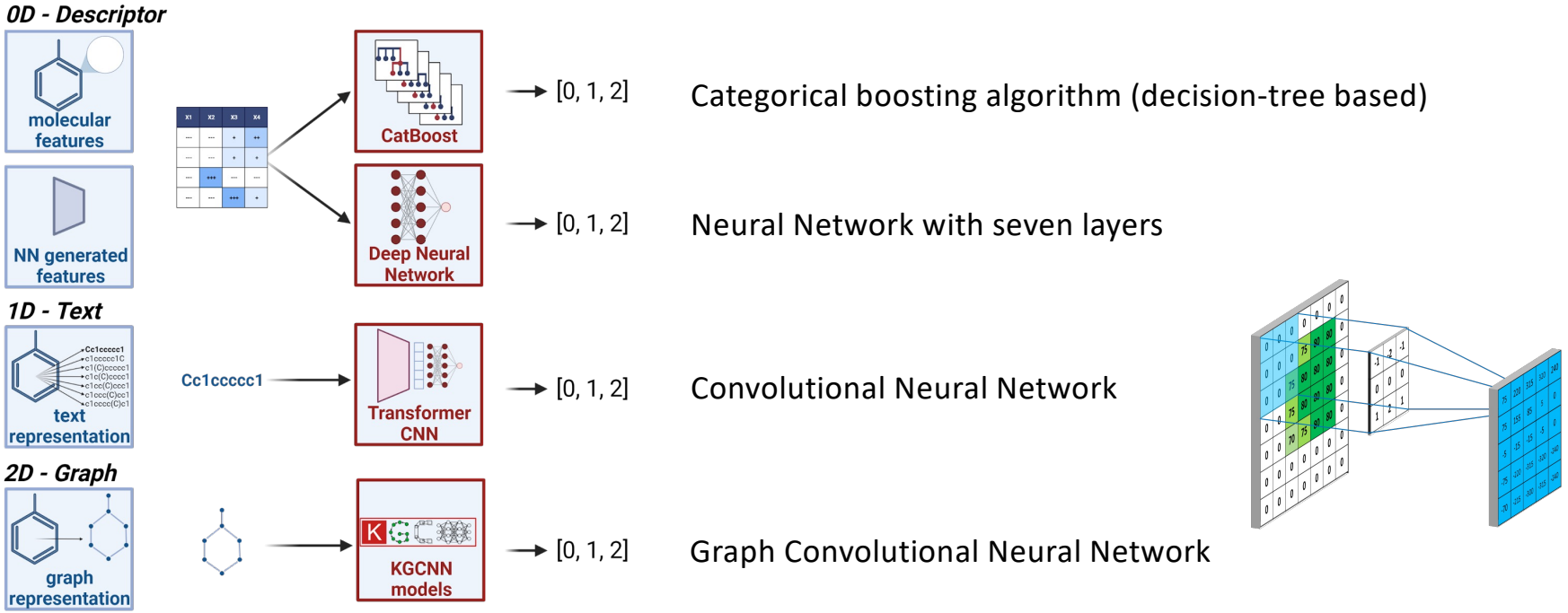
- Experimentally: Nephelometer measures undissolved sediment
- Classification into *low*, *medium* and *high* soluble with phenytoin and amiodarone as thresholds
- 70k training datapoints, 15k public leaderboard, 15k private leaderboard
- Stratified random sampling



Workflow with OCHEM



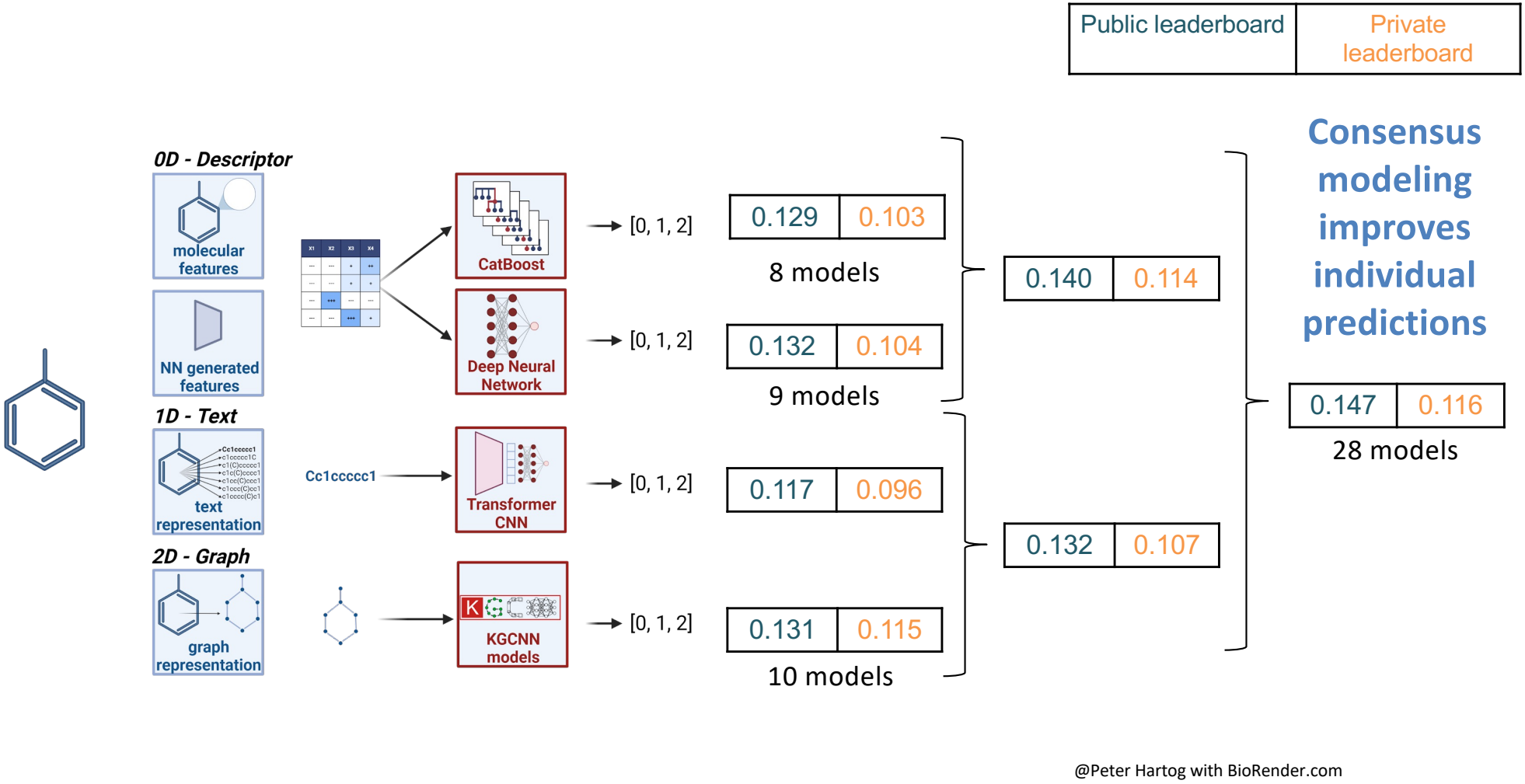
Molecular representation



@Peter Hartog with BioRender.com

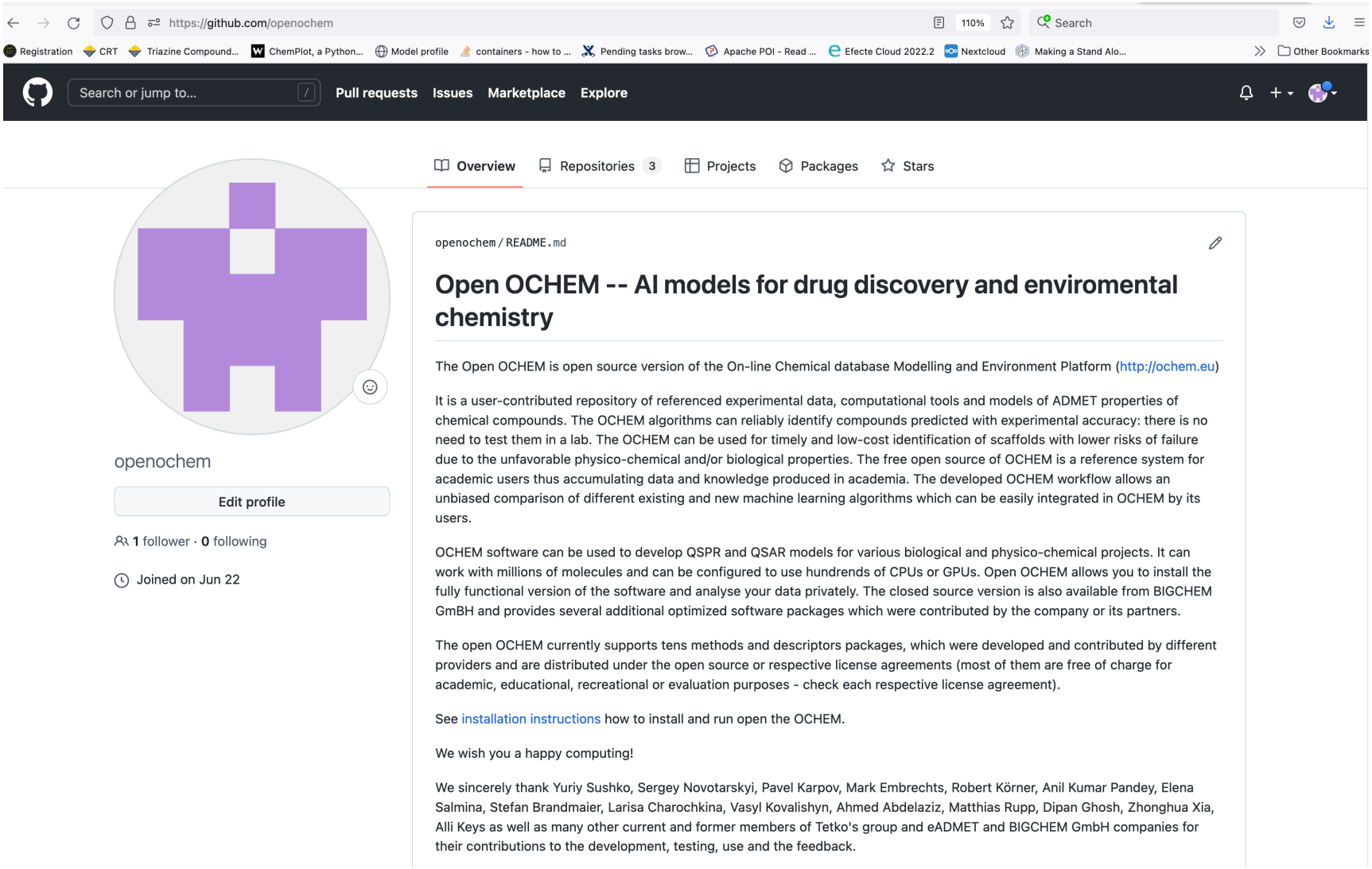
Kopp at al, DOI: 10.26434/chemrxiv-2023-p8qcv

Quadratic kappa metric scores



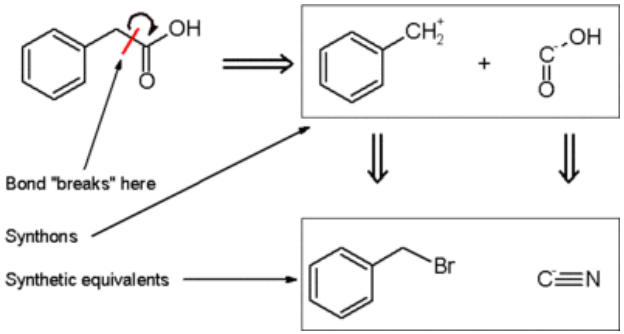
Kopp at al, DOI: [10.26434/chemrxiv-2023-p8qcv](https://doi.org/10.26434/chemrxiv-2023-p8qcv)

openOCHEM


https://github.com/openochem
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.

Elias James Corey
Nobel Prize 1990

Retrosynthesis – it is a problem solving technique for transforming the structure of **synthetic target molecule (TM)** to a sequence of progressively simpler structures along the pathway which ultimately leads to simple or commercially available starting materials for a chemical synthesis.

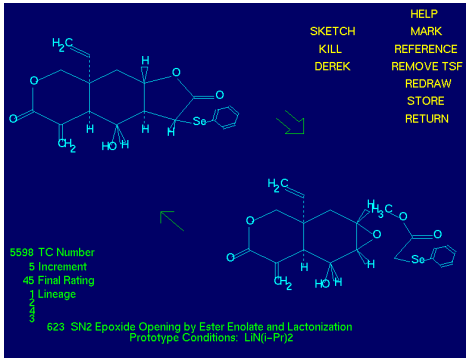


<u>TM Transformations:</u>

FGA – function group addition
FGI - function group interconversion
FGR – function group removing

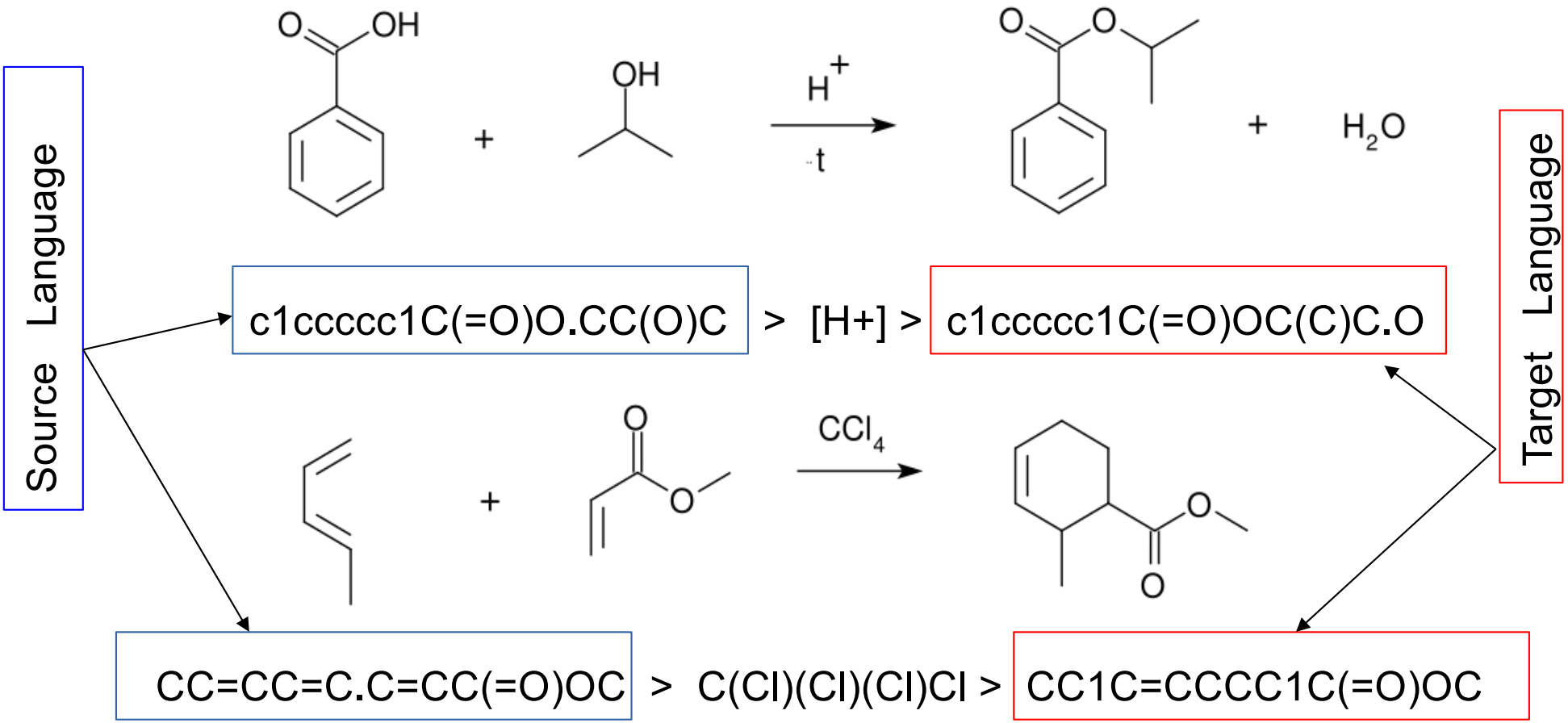
Cyclisation, Fisher indole, Mannich, Michael, Oxidation, Rearrangement, and many others.

These rules make core of the first programs **OCSS** (organic chemical simulation of synthesis) and **LHASA** (logic and heuristics applied to synthetic analysis)

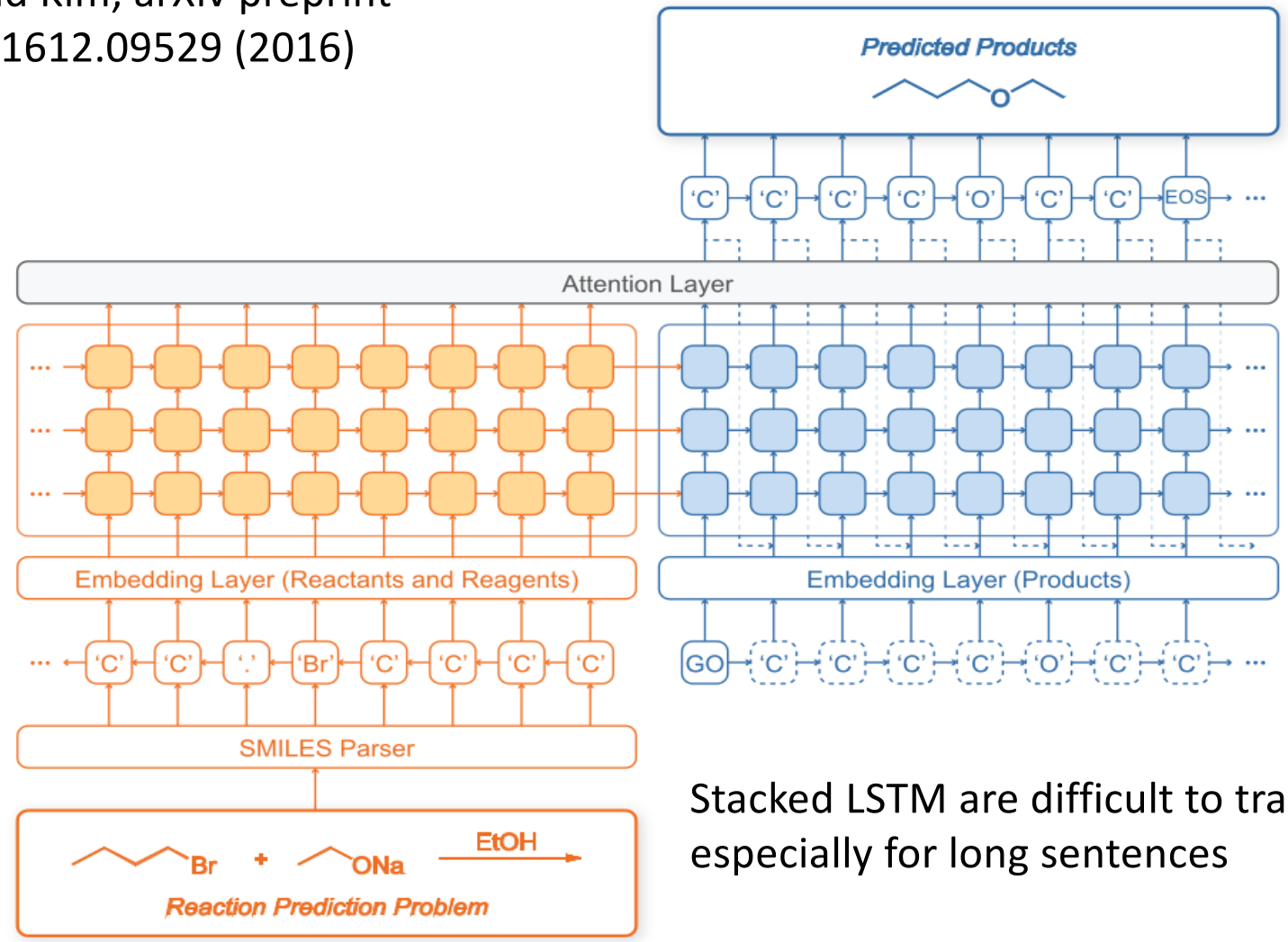


Natural Language Processing (NLP) for reaction predictions

Reaction prediction task can be treated as a neural machine translation

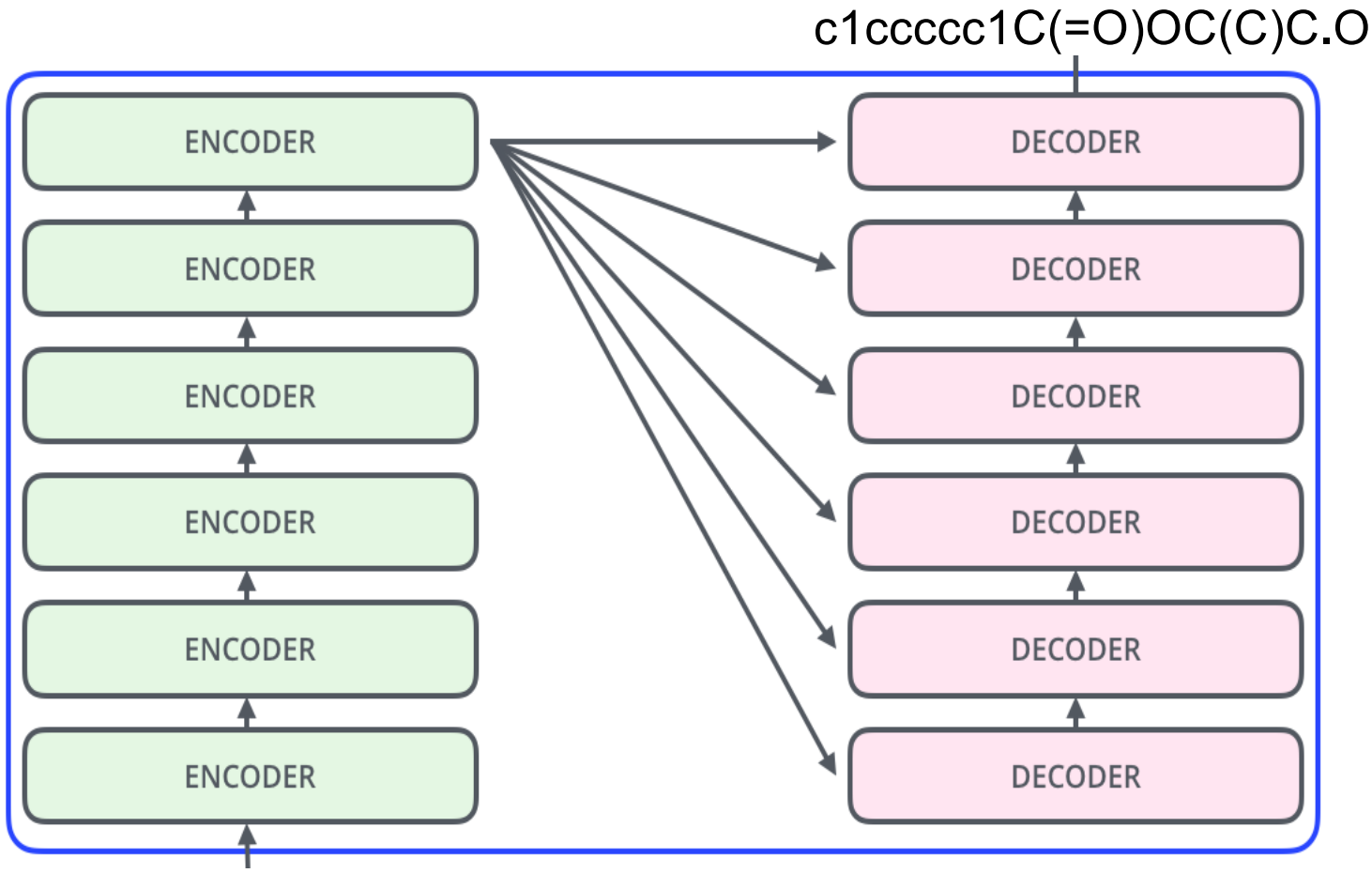


Nam and Kim, arXiv preprint
arXiv:1612.09529 (2016)



Stacked LSTM are difficult to train especially for long sentences

The Transformer architecture is currently the state of art for organic reactions



https://arxiv.org/abs/1706.03762 (Attention is all you need)

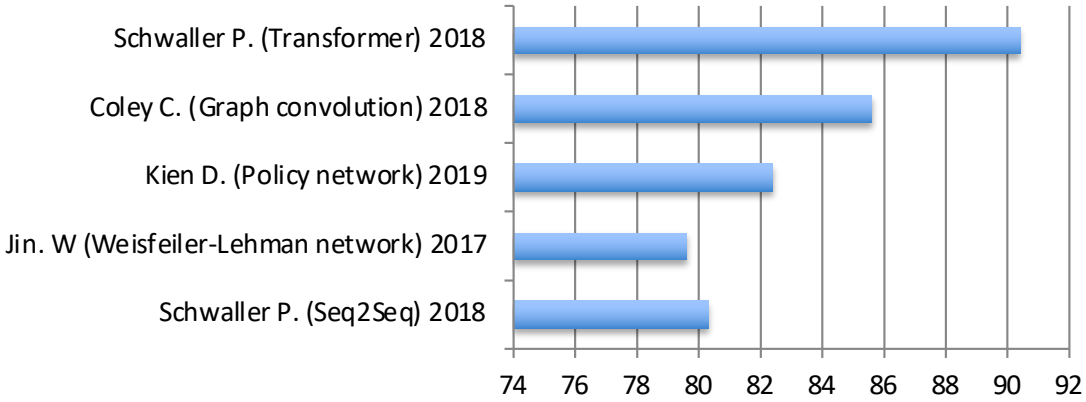
Synthesis planning (top-1)

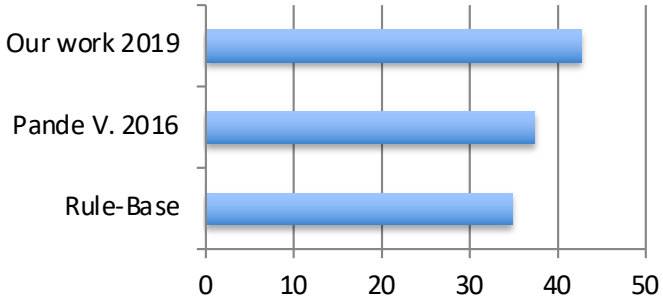
Direct synthesis USPTO-500k

Schwaller P. (Seq2Seq) 2018	80,3
Jin. W (Weisfeiler-Lehman network) 2017	79,6
Kien D. (Policy network) 2019	82,4
Coley C. (Graph convolution) 2018	85,6
Schwaller P. (Transformer) 2018	90,4

Retro-synthesis USPTO 50k

Rule-Base	34,8
Pande V. et al 2016	37,4
ICANN 2019 (our work)	42,7





Karpov, P.; Godin, G.; Tetko, I. V. In *A Transformer Model for Retrosynthesis*, Artificial Neural Networks and Machine Learning – ICANN 2019: Workshop and Special Sessions, Münich, 17th - 19th September 2019; Tetko, I. V.; Kůrková, V.; Karpov, P.; Theis, F., Eds. Springer International Publishing: Münich, 2019; pp 817-830.

Image augmentation

https://github.com/aleju/imgaug

Synthesis planning (top-1)

Retro-synthesis (50k), USPTO-50

Rule-Base	34,8
Seq2Seq, Pande V. 2016	37,4
Graph Logic Networks, Coley C. 2019	52,5
Our work: ICANN2019	42,7
Our work: (Augmented Transformer, 2020)*	**53,5**

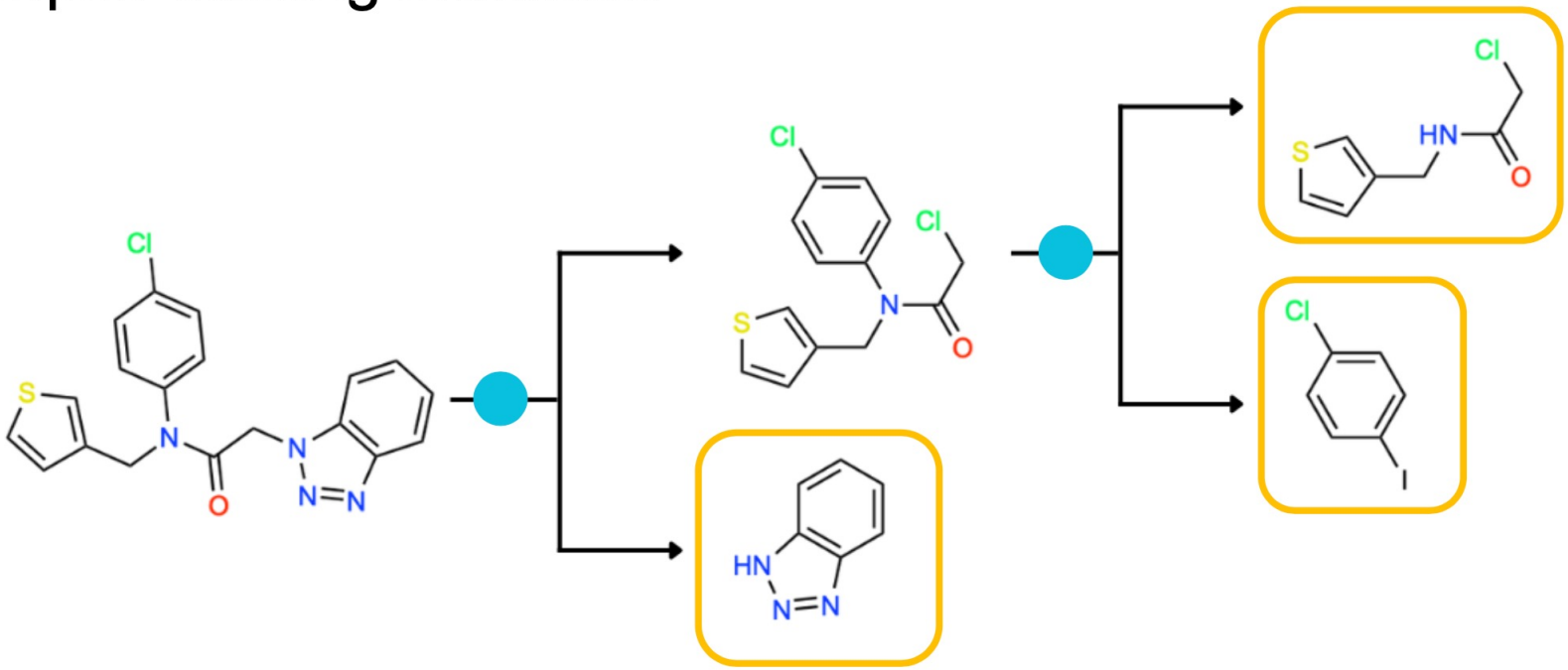
Direct synthesis (500k), USPTO-MIT

Schwaller P. (Seq2Seq) 2018	80,3
Jin. W (Weisfeiler-Lehman network) 2017	79,6
Kien D. (Policy network) 2019	82,4
Coley C. (Graph convolution) 2018	85,6
Schwaller P. (Transformer) 2019	90,4
Our work (Augmented Transformer, 2020)*	**91,9**

*Tetko et al *Nature Comm*. 2020. 11, 11, 1-11.

Retrosynthesis planning

- Design a synthetic route for a target molecule
- Working backward to identify a sequence of reactions to reach simpler starting materials



P. Torren-Peraire et al Models Matter: The Impact of Single-Step Retrosynthesis on Synthesis Planning, https://arxiv.org/abs/2308.05522

Conclusions

- Computational tools are essential for drug discovery
- Use of diverse methods and descriptors can help to identify the best models for the given task
- Representation learning methods provide similar or better performance than traditional models based on descriptors
- Multitask learning can further improve model performances
- Combination of models as consensus usually contributes best performing approaches
- Publishing model on-line allows their wide promotion and acceptance by the scientific community
- Multistep reaction predictions is a new challenge to be addressed with AI

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.

6 out of 14 positions are still available

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

https://icann2024.org

Welcome to ICANN 2024

The 33rd International Conference on Artificial Neural Networks.
A conference of the European Neural Network Society.

The 2024 edition of the ICANN will be organized by the Dalle Molle Institute for Artificial Intelligence Research (IDSIA USI-SUPSI) in Lugano, Switzerland in collaboration with AIDD and AiChemist Horizon MSCA projects.

Conference venue: **USI-SUPSI Campus Est, Via la Santa 1, 6962 Lugano-Viganello, Switzerland**

Conference dates: **September 17 to September 20, 2024.**

Special session: AI in Drug Discovery
Big Data and advanced machine learning in chemistry
eXplainable AI (XAI) in chemistry
Chemoinformatics
Use of deep learning to predict molecular properties
Modeling and prediction of chemical reaction data
Generative models

Dalle Molle Institute for Artificial Intelligence
IDSIA
USI SUPSI

Acknowledgements

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

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

AiChemist

eco
environmental chemoinformatics
marie curie initial training network

BigChem