

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





# 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 Med	ium model	Good I	model	Robust model			
	Endpoint	Model type	Data set	size	2005	2009	2014	2019	Retraining
	Caco-2 permeation	C (N)	>10 00	0			RF	SVR	Weekly
Absorption	Caco-2 efflux	C (N)	>10 00	0			RF	SVR	Weekly
	Bioavailability (rat)	С	~2000					RF	On demand
Distribution	Human serum albumin	N	>30 00	0			PLS	MTNN	On demand
Distribution	Fraction unbound	N	>1000	)			PLS	MTNN	On demand
	Microsomal stability (hum)	C (N)	>10 00	0			RF	RF	Weekly
Motabolism	Microsomal stability (mouse)	C (N)	>10 00	0			RF	RF	Weekly
Wetabolish	Microsomal stability (rat)	C (N)	>10 00	0			RF	RF	Weekly
	Hepatocyte stability (rat)	C (N)	>30 00	0			RF	RF	Weekly
	hERG inhibition	С	>10 00	0			RF	SVM	Weekly
	Ames mutagenicity	С	>10 00	0			RF	RF	On demand
Toxicity	CYP inhibition isoforms	С	>10 00	0			RF	RF	On demand
	Phospholipidosis	С	<1000				SVM	SVM	On demand
	Structure filter tool	Score	n.a.		-	-	-	-	On demand
	Solubility (DMSO)	N	>30 ,00	0				MTNN	On demand
	Solubility (Powder)	N	<10 00	0			FL0	MTNN	On demand
	logD @ pH 7.5	N	>70 00	0			PLS	MTNN	On demand
PhysChem	Membrane affinity	N	<10 00	0			PLS	MTNN	On demand
	рКа	N	>10 00	0			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



Online chemical database with modeling environment

Home 
 Database 
 Models

#### Welcome to OCHEM! Your possible actions

Explore OCHEM data Search chemical and biological data: experimentally measured, published and exposed to public access by our users. You can also upload your data.

#### Create QSAR models

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

#### Run predictions Apply one of the available models to predict property

you are interested in for your set of compounds.

#### Screen compounds with ToxAlerts Screen your compound libraries against structural alerts for such endpoints as mutagenicity, skin sensitization,

aqueous toxicity, etc.

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

Our acknowledgements

Feedback and help

User's manual Check an online user's manual

#### Check out the properties available on OCHEM OCHEM contains 3642695 records for 639 properties (with at least 50 records) collected from 18178 sources Melting Point logPow Cotrain/Cplasma IC50 Papp(Caco-2) Papp(MDCK) Oral absorption LIC 50 CheckritCplasma Papp ratio(Caco-2) Plasma protein binding Papp ratio(MDCK-mdr1) pIC50 %Human FA Human IA Human FA fraction unbound (fu) fraction lonized (fi) pKa VDss LogIC50 LogPI BBB permeability (qualitative) LogKoa LogRBA CYP450 modulation CYP450 reaction Vapor Pressure EC50 aquatic NOEC aquatic LOEC ACCO L

 EC50 EROD induction
 LC 50
 C 50
 C 50
 LC 50
 LC 50
 C 50
 LC 50</td

#### Cblood/Cair(Human) Cfat/Cair(Rat) Cbrain/Cair(Rat) Cliver/Cair(Rat) Cmuscle/Cair(Rat) IC50 PDE4 % inhibition PDE4 IC50 inhibition Density pKa (smiles as ob. cond.) DMSO Solubility log Kb logk0 logLOAEL

hERG K+ Channel Blocking (IC50) 5-HT2B (KI) LogKoc BCF CHSEL % inhibition hERG, K+ Channel Blocking hERG K+ Channel Blocking (Ki) logP Chloroform/Water 5-HT2C (Ki) 5-HT2b (Kb) PaP substrate 5-HT2A (Ki) D2R (Ki) a1 adrenergic receptor (Ki) 5-HT2b (IC50) Modes of Toxic Action LC50 ratio Solid-liquid total phase change entropy enthalpy of fusion % inhibition PgP PgP modulator PgP inhibitor Bioaccumulation in C. elegans PgP inducer PTP1B inhibition(pl) IC50 HIV TD50 Skin permeability Human Clearance MRT Mean Residence Time t1/2 Ki trypsin AC50 Trypsin Inhibition Growth inhibition Trypsin Inhibition activity Trypsin Inhibition class Cell permeability test Ki trypsin FDA classification CAESAR class GHLI Ki inhibitor trypsin Anti-Cancer activity CA Chromosomal Aberration Index LD50bee Papp(RI) skin sensitisation:LLNA index Mutagenicity EC50 bioluminescence AhR binding affinity EC50 AHH induction EC50 Antimicrobial activity NanoToxicity LC50 aquatic NanoToxicity MIC NanoToxicity mortality NanoToxicity EC50 Genotoxic carcinogenicity, mutagenicity Flash point Bioaccumulation Factor (BAF) 5-LOX(1) Ready biodegradability Binding constant HIV EC50 HIV IC50 Biological Oxygen Demand ppi-inhibitor Toluene solubility logPtw HIV Active Compounds logPchlor/w logPcycl/w IC50 cell proliferation IC50 tubulin IC50 telomerase logERRBA (qualitative) SRC2 Inhibitor IC50 FPPS log RP AR km (biotransformation rate) Severe Skin Disorder logPhxd/w logPalk/w tubulin inhibitors AlphaScreen-FHs herg\_act\_inact phospholipidosis status Retention Factor Chromatographic Hydrophobicity Index logKd DILL Abraham descriptor A Abraham descriptor B Abraham descriptor S Abraham descriptor E

#### Latest active users Brandon: Mr. Brandon Gundani seconds ago uddiptagd: Mr. Uddipta Ghosh Dastidar seconds ago Ivalex.09: Dr. Alexander Ksenofontov seconds ago Rahila: Mrs. Rahila Pathan seconds ago rama1: Mr. Rama Krishnan

v.4.2.1

👨 log in create accoun

A+ a- Privacy statement

6 minutes ago

Amidoff: Dr. Dmitriy Makarov 9 minutes ago

Latest published models

AntioxidantActivity\_IC50 model published by
kovalishyn
about a month ago

Absorbance maximum wavelength model published by AlexeyR 2 months ago

Cryptic Pocket Inducer model published by Zhonghua 4 months ago

nephrotoxic-binary model published by qingshuang0501 5 months ago

AlphaScreen-GST-FHs model published by dipanHZM

Absorbance maximum wavelength model published by ivalex.09 7 months ago

Melting Point model published by Amidoff 7 months ago

MIC model published by hodyna more than a year ago

IC50 model published by carpovpv more than a year ago

Delta density of mixtures model published by xenol more than a year ago

LC50 aquatic model published by Tinkov\_Oleg





```
openochem
```

Edit profile

R 1 follower · 0 following

( Joined on Jun 22

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)

ß

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!

openochem / README.md

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



### 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)
- **2** 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)

### QSPR/QSAR modelling in OCHEM

Select the molecular descriptors 🕕		Create a model () Select the training and validation sets, the machine learning method and the validation protocol
Recommended descriptor types (2D)	Predictions by OCHEM's featured models 🕕	
OEState     Reads Indices	Ames levenberg     Toxicity against T. Pyriformis	Select the training and validation sets:
Counts only	ALogPS 3.0     CYP1A2 Estate+ALogPS	Training set (required): peptidesrear [details]
✓ ALogPS (2) ☐ Mold2 (777)	CYP2C9 Estate+ALogPS	Add a validation set
DPlogP	CYP2Db Estate+ALogPS CYP3A4 Estate+ALogPS Pvrolvsis point prediction (best Estate)	The model will predict this property: LogD using unit: Log unit
ISIDA fragments The in Hashed Atom Pair fingerprint (MAP4)	Melting Point prediction (best Estate)     Water solubility model based on logP and Melti	Skip model configuration and use the predefined settings
GSFragment (1138) QNPR	ALOGPS 2.1 logP	Choose the learning method:
<ul> <li>Multilevel Neighborhoods of Atoms (MNA)</li> <li>Structural alerts (ToxAlerts and Functinal Groups)</li> </ul>	Outputs of other OCHEM models	Suggested modeling methods:
Recommended descriptor types (3D)	Obsolete/Additional descriptor types	<ul> <li>ASNN: ASsociative Neural Networks doi:10.1007/978-1-60327-101-1_10</li> <li>(New) Attentive FP doi: 10.1021/acs.jmedchem.9b00959</li> </ul>
<ul> <li>alvaDesc v.2.0.4 (5666/3D)</li> <li>Dragon v. 7 (5270/3D)</li> <li>CDK 2.7.1 descriptors (256/3D)</li> <li>Chemaxon descriptors (499/3D)</li> </ul>	CDK 2.0 descriptors (256/3D) CDK 1.4.11 descriptors (256/3D) E-state Dragon v. 5.4 (1644/3D)	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 Conservation and characteristic developed for the same set)
<ul> <li>RDKit descriptors (3D)</li> <li>MORDRED descriptors (1826/3D)</li> <li>MOPAC2016 descriptors (35/3D)</li> </ul>	<ul> <li>Dragon v. 5.5 (3224/3D)</li> <li>Dragon v. 6 (4885/3D)</li> <li>MOPAC 7.1 descriptors (25/3D)</li> </ul>	<ul> <li>Consensus model (based on models developed for the same set)</li> <li>DEEPCHEM: several methods from DeepChem (GPU) arXiv:1703.00564</li> <li>(New) DIMENET - Directional Message Passing Neural Network arXiv:2003.03123</li> <li>Deep Learning Consensus Architecture (DI CA) doi:10.1021/acs.icim 9b00526</li> </ul>
<ul> <li>□ KrakenX descriptors (MOPAC2016 derived)(124/3D)</li> <li>□ PyDescriptor descriptors (16251/3D)</li> <li>□ MERA descriptors (529/3D)</li> <li>□ MERDY descriptors (200)</li> </ul>		<ul> <li>DNN: Deep Neural Network (GPU) doi:10.1021/acs.jcim.8b00685</li> <li>EAGCNG - Edge Attention based Multi-relational Graph Convolutional Networks (GPU) arXiv:1802.04944</li> <li>ESMI B: Fast Stagewise Multiple Linear Begression doi:10.1134/S0012500807120026</li> </ul>
<ul> <li>MERST descriptors (42/3D)</li> <li>Inductive' descriptors (54/3D)</li> <li>Spectrophores (144/3D)</li> </ul>		<ul> <li>GNN - Graph Isomorphism Network (GPU) arXiv:1910.13124</li> <li>KNN: k - Nearest Neighbors</li> <li>KPL S - Kernel Partial Least Squares doi:10.1109/LICNN 2006 246832</li> </ul>
Special descriptors (scaffolds fingerprints):		<ul> <li>LibSVM: grid-search parameter optimisation doi:10.1145/1961189.1961199</li> <li>LSSVMG: Least Squares Support Vector Machine (GPU) doi:10.1023/A:1018628609742</li> </ul>
Chemaxon Scaffolds Cicenterational Content of Content		<ul> <li>MLR: Multiple Linear Regression</li> <li>PLS: Partial Least Squares doi:10.1016/S0169-7439(01)00155-1</li> <li>RFR: Random Forest regression and classification doi:10.1023/A:1010933404324</li> </ul>
MolPrint Fingerprints		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
Conditions of experiments		<ul> <li>WEKA-J48: Weka C4.5 decision trees, only classification - use with bagging doi:10.1145/1656274.1656278</li> <li>WEKA-RF: Random Forest, only classification doi:10.1023/A:1010933404324</li> <li>XGBoost: Scalable and Flexible Gradient Boosting doi:10.1145/2939672.2939785</li> </ul>
Model validation Validation method	N-Fold cross-validation ~	
□ Stratified cross	-validation (classification only)	

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

### Modeling iterative workflow



### Examples of models for water solubility



Performance



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



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 T	raining set	Validation: Cross-Va	alidation (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



### 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



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<sup>2</sup>. Transformer CNN provides similar or better accuracy compared to traditional methods based on descriptors <u>even for small datasets (few hundrends compounds!)</u>. 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)



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: <u>https://twitter.com/AiddOne</u>

### Interested about AI in toxicology? Read ChemResTox August SI

# **AI Meets Toxicology**



**RIGHTS & PERMISSIONS V**Subscribed

**RETURN TO ISSUE** 

1289-1290

Society

# Winning model: OCHEM-generated consensus model

Andrea Kopp SLAS Europe 2023

25.05.2023

# HELMHOLTZ MUNICH Team of Igor Tetko with

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



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

### **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 ᅌ for Training set	ᅌ Valida	tion: Cros	s-Validatio	on (84 model	s)
	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	



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

	rironmental Protection Agency TECHNOLOGY I LAWS & REGULATIONS I A	BOUT EPA	ALL EPA THIS AREA Advanced Se
Computational Toxico You are here: EPA Home » Re	logy Research search & Development » CompTox » Chemi	cal Data Challenges & Release	⊠ Contact Us
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

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.



About

### Tox21 Data Challenge 2014



Contact Us

» Home

Registration

Data/Resources

Submissions

Discussion

Leaderboard

Survey





#### 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



### **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 multitarget 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 multitarget 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



Х

### 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.



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

15

 $\Diamond$ 

1 - 15 of 266

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

H<sub>3</sub>C

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)







Stiftung/Foundation

Alexander von Humboldt

