

Institute of
Structural Biology

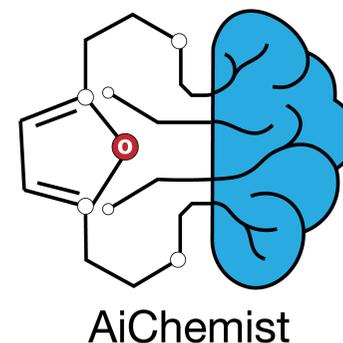
Advanced Machine Learning in Drug Discovery

Igor V. Tetko

Helmholtz Munich and BIGCHEM GmbH

Katholieke Universiteit te Leuven, 2 September 2024

HELMHOLTZ MUNICH



**HELMHOLTZ
MUNICH**



German Research Center for Environmental Health

Agenda

Computational predictions in drug discovery

OCHEM platform

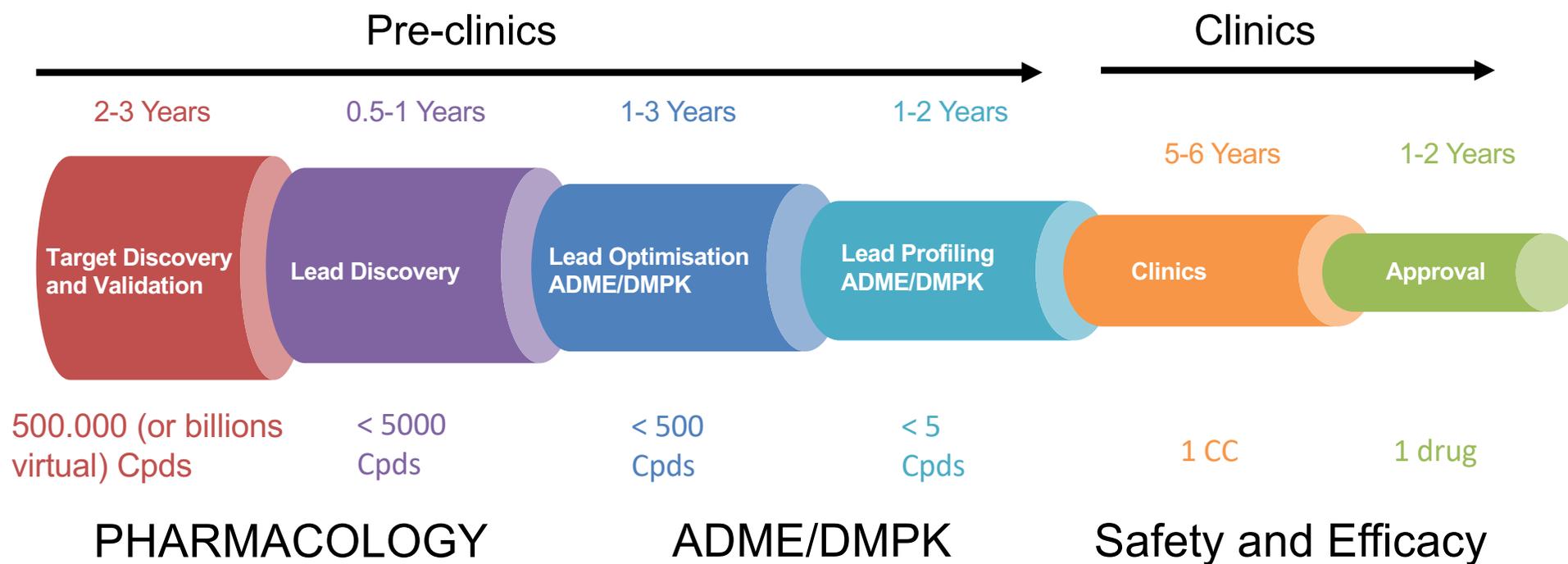
Consensus modelling as strategy to develop best models

Tox24 Challenge results

Uncertainty estimation and Applicability domain

Explainable AI

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

OCHEM <https://ochem.eu>

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 **3771600 records** for **688 properties** (with at least 50 records) collected from **20521 sources**

Melting Point **logPow** **logBB** LogL(water)
Kpu(brain) **LogD** Kpu(adipose) Kpu(heart) Kpu(kidney) Kpu(liver)
Kpu(lungs) Kpu(muscle) Kpu(skin) logPI(+)

Water solubility LogL(blood) LogL(oil) ER fu(brain)
P/Papp Cbrain/Cplasma **IC50** **Papp(Caco-2)**

Papp(MDCK) **Oral absorption** LIC 50 Cheart/Cplasma

Papp ratio(Caco-2) **Plasma protein binding**

Papp ratio(MDCK-mdr1) **pIC50** **%Human FA**

Human IA **Human FA** **fraction unbound (fu)**
fraction ionized (fi) **pKa** **VDss** **LogIC50** **LogPI**

BBB permeability (qualitative) **LogKoa**
LogRBA **CYP450 modulation**

CYP450 reaction **Vapor Pressure**

EC50 aquatic **NOEC aquatic** **LOEC aquatic**
IC50 aquatic **LC50 aquatic** **log(IGC50-1)** **LEL**

Henry's law constant Photolysis rate Kp

Half-Life Hydrolysis HLh EC50 EROD induction **LC 50** LCLo

Boiling Point **LD50 dermal** **LD50 oral**
LC50 terrestrial **AMES** **LD50** Biodistribution

Water solubility at pH Papp(PAMPA)

Latest active users

-  **feifeiwst:** Dr. Shutao Wang
about 2 hours ago
-  **mjohn:** Dr. Maya John
about 7 hours ago
-  **Amidoff:** Dr. Dmitry Makarov
about 10 hours ago
-  **Souvik:** Mr. Souvik Pore
about 13 hours ago
-  **ygkoki10103:** Mr. nakagomi koki
about 13 hours ago
-  **Zahra-Hmz:** Miss. Zahra Hamzei
about 13 hours ago

Latest published models

-  **MIC model** published by **vkovalishyn**
2 months ago
-  **Chemical shift model** published by **isaev.yaroslav1**
3 months ago
-  **Molar extinction coefficient model** published by **ivalex.09**
6 months ago
-  **Absorbance maximum wavelength model** published by **bichan**
6 months ago
-  **Absorbance maximum wavelength model** published by **AlexeyR**
6 months ago
-  **nephrotoxic-binary model** published by **ningshuang0501**



FILTERS

SOURCE

Article/Source [\[select\]](#)

Page Table

PROPERTY

Activity/Property [\[select\]](#)

Hide records without property

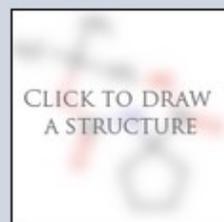
CONDITIONS

MOLECULE FILTERS

Name / OCHEM ID [\[i\]](#) / Inchi-Key

Similarity/substructure search

Draw a structure and search all the molecules containing it or similar to it



Molecular mass [\[i\]](#)

between and

ADVANCED MOLECULE FILTERS

MISCELLANEOUS

Current set [\[i\]](#)

Show all

Data origin and quality:

Data introducers: All users

Data visibility: Public and private

Data from other users: Only approved data

Basket Records

1 - 5 of 1075221 Items on page of 215045 > >>

 molecule profile	<p>logPow = 1.77 (in Log unit)</p> <p>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</p> <p>MoleculeID: M11715 Public record</p>	<p>RecordID: R1 09:35, 19 Mar 09 / 10:04, 13 Oct 11 i.tetko / itetko </p>
 molecule profile	<p>logPow = 2.1 (in Log unit)</p> <p>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</p> <p>MoleculeID: M9560 Public record</p>	<p>RecordID: R2 09:35, 19 Mar 09 i.tetko </p>
 molecule profile	<p>logPow = 2.2 (in Log unit)</p> <p>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</p> <p>MoleculeID: M10111 Public record</p>	<p>RecordID: R3 09:35, 19 Mar 09 i.tetko </p>
 molecule profile	<p>logPow = 1.64 (in Log unit)</p> <p>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</p> <p>MoleculeID: M9777 Public record</p>	<p>RecordID: R4 09:35, 19 Mar 09 i.tetko </p>

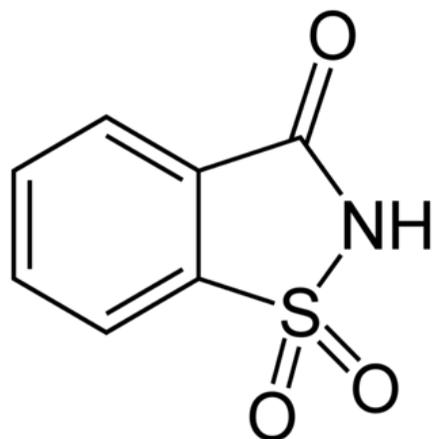
Traditional representation of chemical structures

Saccharin

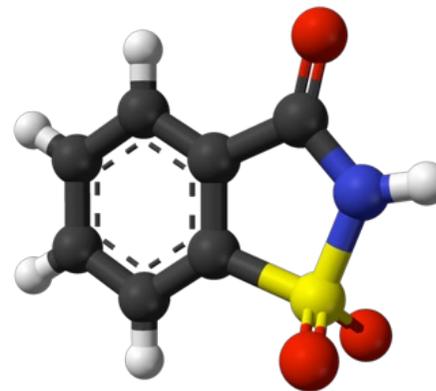
1D



2D



3D



Examples of descriptors

alvaDesc v.2.0.4 (5666/3D)

[\[select all\]](#) [\[select none\]](#) [\[select 3D\]](#) [\[unselect 3D\]](#)

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

QSPR/QSAR modelling in OCHEM

Select the molecular descriptors ¹

Recommended descriptor types (2D)

- OEState
 - Bonds Indices
 - Counts only
- ALogPS (2)
- Mold2 (777)
- CDDD
- JPligP
- 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)

Special descriptors (scaffolds, fingerprints):

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

Conditions of experiments

- pH
- Ionisable

Create a model ¹

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

Predictions by OCHEM's featured models ¹

- 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 Melti
- 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)

Select the training and validation sets:

Training set (*required*): [peptidesegr](#) [details]
[Add a validation set](#)

The model will predict this property:

LogD using unit:

Skip model configuration and use the predefined settings

Choose the learning method: ¹

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:

Number of folds:

- Stratified cross-validation (classification only) ¹
- Treat each record as a new molecule ¹

You can create a model from template: [import an XML model template](#) or [use another model as a template](#)

Each descriptor re-presentation sees only part of molecules



Blind monks examining an elephant, an [ukiyo-e](#) print by [Hanabusa Itchō](#) (1652–1724).

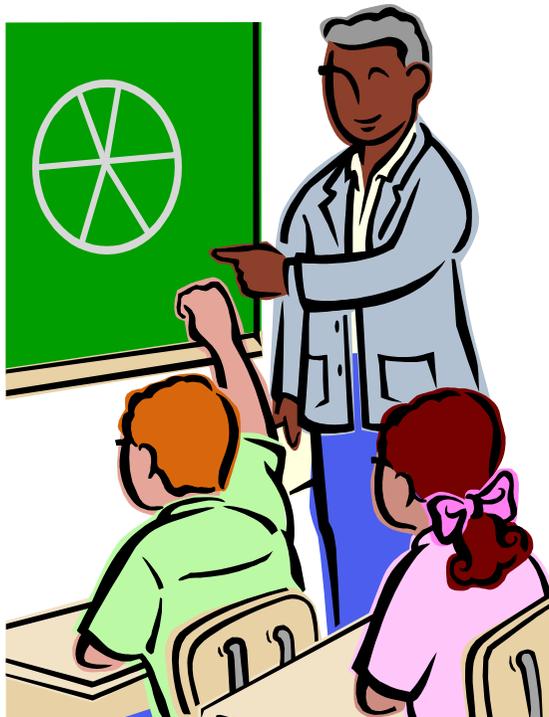
https://en.wikipedia.org/wiki/Blind_men_and_an_elephant

Consensus modelling

Best method(s) are defined

Average prediction of models is used

The consensus prediction is more accurate and stable



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](#)

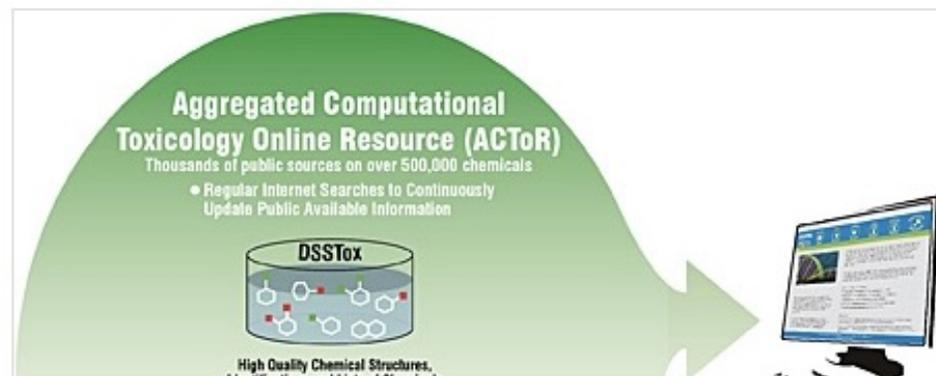
ToxCast Chemical Data Challenges and Release

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](#)



Aggregated Computational Toxicology Online Resource (ACToR)
Thousands of public sources on over 500,000 chemicals

- Regular Internet Searches to Continuously Update Public Available Information

DSSTox

High Quality Chemical Structures, Modifications and Lists of Chemicals

ToxCast Challenge

- OCHEM model (by Dr. S. Novotarskyi) got the the first position for the US Environmental Protection Agency (EPA) challenge in May 2014
- Prediction of systemic Lowest Effect Level (LEL)
 - lowest dose that shows adverse effects in animal toxicity tests
- Challenge: build a prediction model using data from high-throughput *in vitro* assays provided by EPA to quantitatively predict a chemical's systemic LEL.



<http://www.epa.gov/ncct/challenges.html>

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

ToxCast Challenge SetUp

- Training set of 483 compounds
- Test set of 1371
 - Leader board: 63
 - Private set: 80

EPA in vitro assays data were also provided (not used in Top I solution)

Model: Associative Neural Network (training with descriptor set optimization)

TOC

Lowest Effect Level (LEL) ToxCast EPA prediction challenge

in vitro + *in silico* → *in vivo* ≈ *in silico* → *in vivo*

*data upload, descriptors calculation, modeling,
consensus, Rank-1 submission, on-line available*

Table 1. Number of Descriptors and Models' Accuracy for the Prediction of the Test Set Compounds

descriptor set	number of selected descriptors	RMSE		
		whole test set (<i>n</i> = 143)	inside of AD ^a (<i>n</i> = 136)	outside of AD (<i>n</i> = 7)
CDK	159	1.13	1.01	2.4
Dragon	1824	1.15	1.05	2.4
Fragmentor	631	1.18	1.04	2.7
GSFrag	202	1.1	0.97	2.5
Mera, Mersy	242	1.04	0.96	2.1
Chemaxon	97	1.16	1.06	2.4
Inductive	39	1.17	1.03	2.7
Adriana	133	1.14	1.01	2.5
QNPR	381	1.12	1.02	2.7
E-state	185	1.16	1	2.8
<i>in vitro</i>	143	1.21	1.11	2.5
Consensus	4036	1.08	0.96	2.5

^aAD is the applicability domain of the model as defined by OCHEM⁸ (see also ref 20).

Statistical uncertainty

Table 2. Summary of the Performance of the Top-Ranked Models of the EPA ToxCast Challenge

model	training set ($n = 483$) ^a		test set					
	RMSE	R^2	provisional subset ($n = 63$)		final subset ($n = 80$)			full, $n = 143$
			RMSE	rank	RMSE	R^2	rank	RMSE
novserj	0.88 ± 0.04	0.27 ± 0.04	1.03 ± 0.08 ^b	8	1.12 ± 0.08 ^b	0.31	1	1.08 ± 0.07
NobuMiu			1.03	9	1.13	0.30	2	1.09
a9108tc			1.05	16	1.13	0.29	3	1.10
klo86 min			1.09	27	1.14	0.29	4	1.12
<i>in vitro</i> assays ^c	0.97 ± 0.04	0.11 ± 0.03						1.24 ± 0.09
MW + NC ^d	0.97 ± 0.04	0.11 ± 0.03						1.18 ± 0.08

^aPrediction accuracy for the “out-of-the-bag” samples. ^bConfidence intervals were estimated using the subsets, which were sampled from the training set, and each had the same size as the respective test set (see for more details ref 23). ^cBest model based on the *in vitro* assay descriptors developed using the LibSVM method (see also Table S1). ^dModel based on molecular weight (MW) and number of carbon atoms (NC) developed using the same approach as the above *in vitro* model.



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.

Challenge setup

Subchallenge Overview

Subchallenges 1-12

Predict the compound activity outcome (active or inactive) in one or more of the 12 pathway assays based on the chemical structure information for the following assays:

- estrogen receptor alpha, LBD ([ER, LBD](#))
- estrogen receptor alpha, full ([ER, full](#))
- [aromatase](#)
- aryl hydrocarbon receptor ([AhR](#))
- androgen receptor, full ([AR, full](#))
- androgen receptor, LBD ([AR, LBD](#))
- peroxisome proliferator-activated receptor gamma ([PPAR-gamma](#))
- nuclear factor (erythroid-derived 2)-like 2/antioxidant responsive element ([Nrf2/ARE](#))
- heat shock factor response element ([HSE](#))
- [ATAD5](#)
- mitochondrial membrane potential ([MMP](#))
- [p53](#)

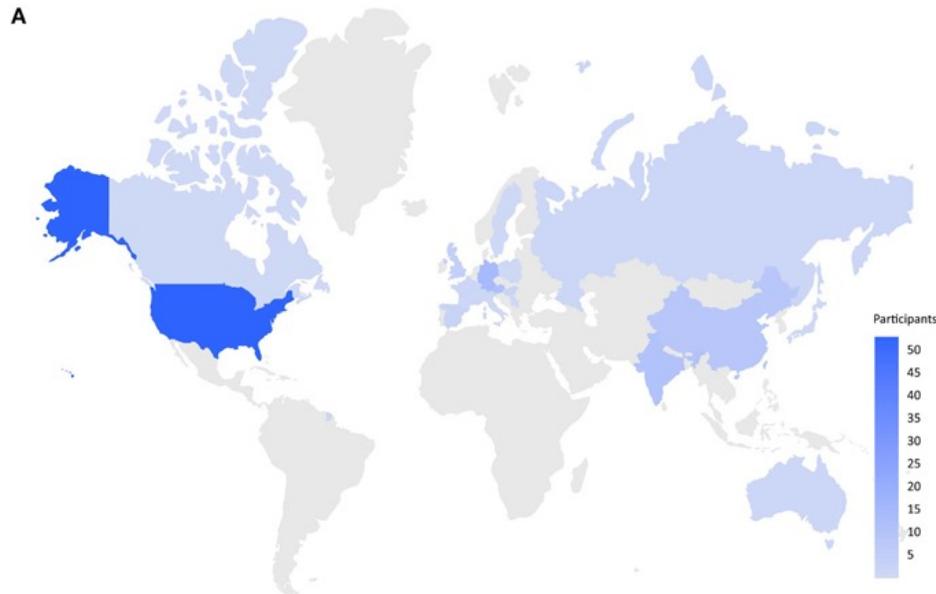
Grand Challenge (All 12)

Subchallenge 13 (all nuclear receptor)

Subchallenge 14 (all stress response pathways)

Set	Class	AhR	AR	AR-LBD	ARE	Aromatase	ATAD5	ER	ER-LBD	HSE	MMP	p53	PPAR.g
Train	Inactive	7219	8982	8296	6069	6866	8753	6760	8307	7722	6178	8097	7962
Train	Active	950	380	303	1098	360	338	937	446	428	1142	537	222
Leader	Inactive	241	289	249	186	196	247	238	277	257	200	241	252
Leader	Active	31	3	4	48	18	25	27	10	10	38	28	15

Tox21 Challenge winners



ROC-AUC:

Mayr et al

DeepTox: Multi-task deep neural network

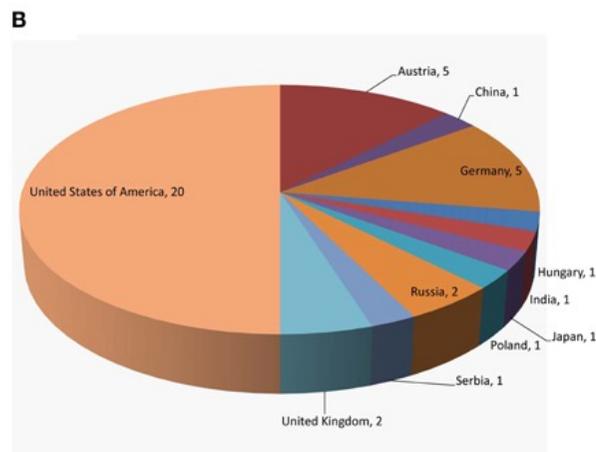
<https://doi.org/10.3389/fenvs.2015.00080>

Best balanced accuracy:

Abdelaziz et al.

ASNN: Associative neural network

<https://doi.org/10.3389/fenvs.2016.00002>



Unbalanced Data? Stop Using ROC-AUC and Use AUPRC Instead

Advantages of AUPRC when measuring performance in the presence of data imbalance — clearly explained



Daniel Rosenberg · Follow

Published in Towards Data Science · 6 min read · Jun 7, 2022



181

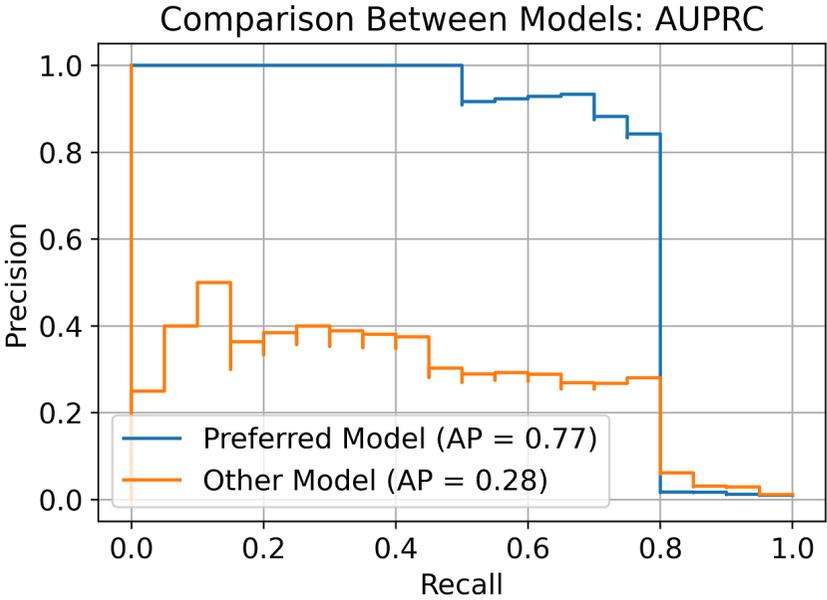
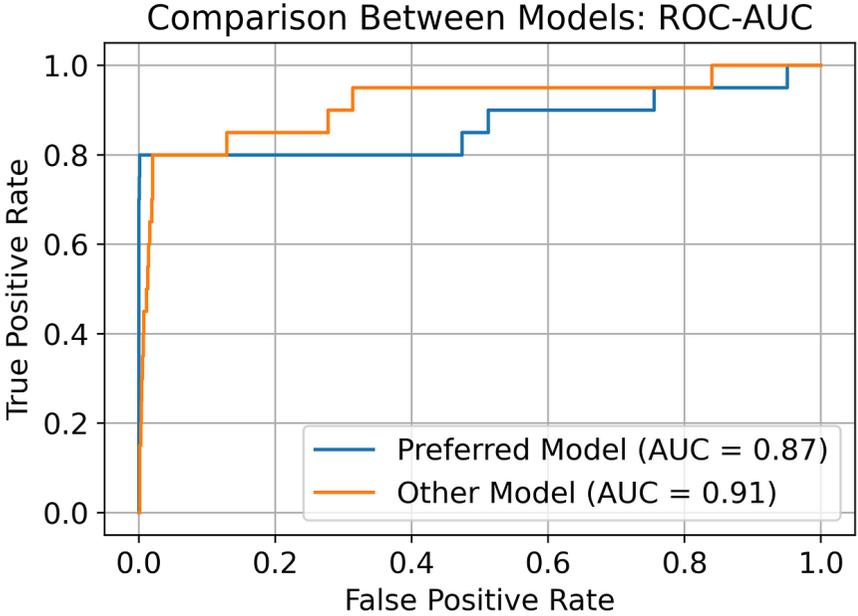


6

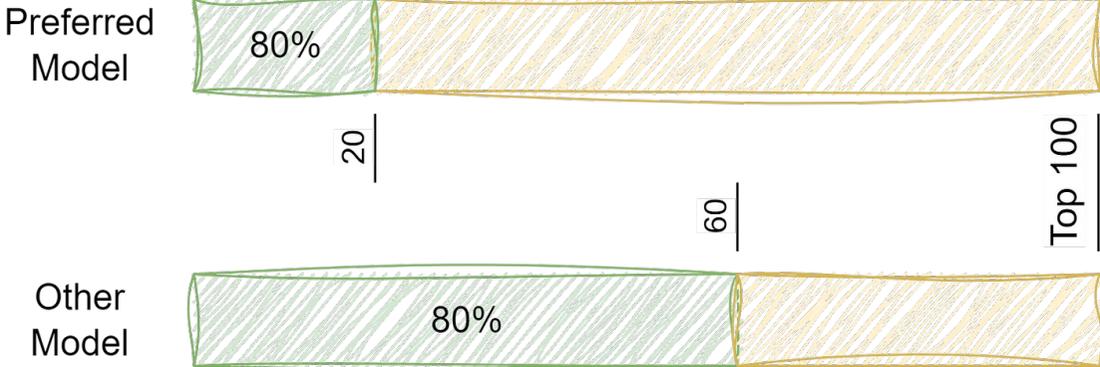


<https://towardsdatascience.com/imbalanced-data-stop-using-roc-auc-and-use-auprc-instead-46af4910a494>

Area Under the Precision-Recall Curve (AUPRC)



20 positives
200 negatives



***Winning model:
OCHEM-generated consensus
model***

Andrea Kopp (Hunklinger)

SLAS Europe 2023

25.05.2023

**HELMHOLTZ
MUNICH**

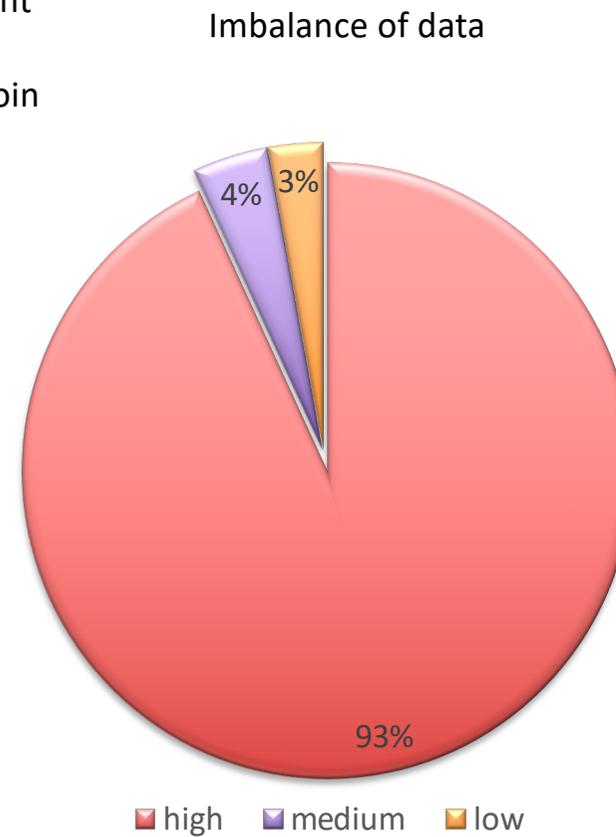
Together with Peter Hartog, Martin Šícho and Guillaume Godin



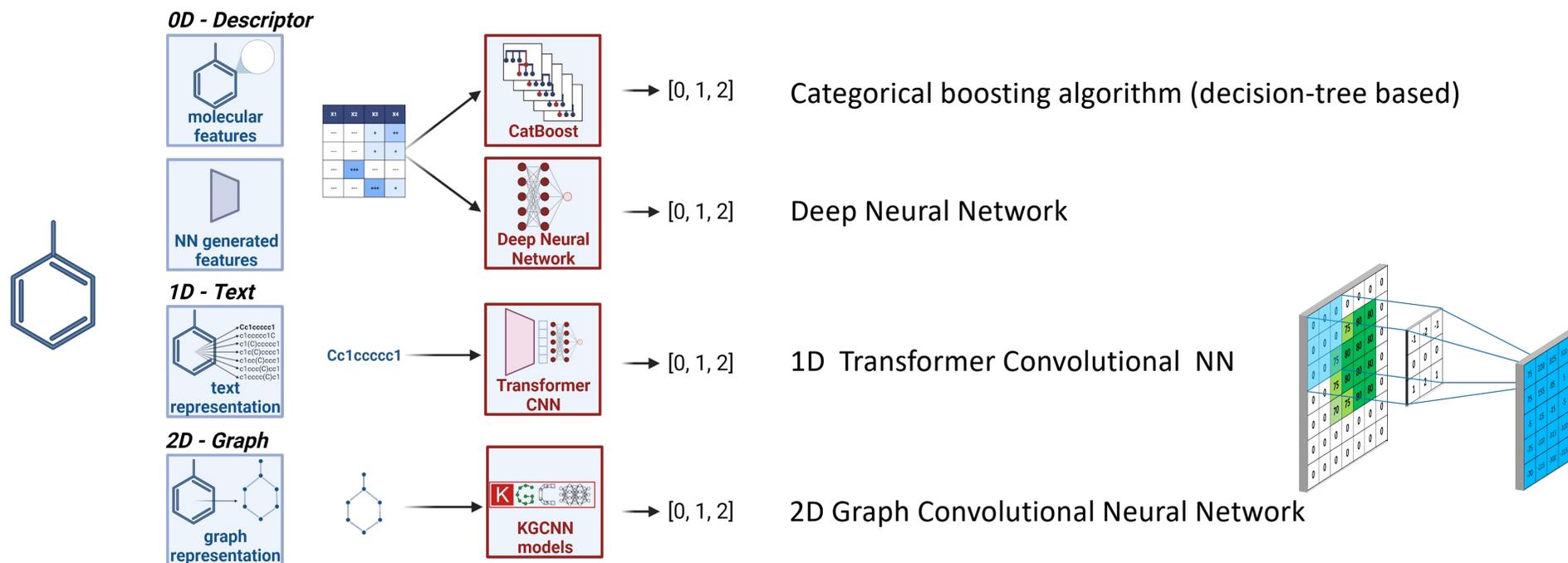
Hunklinger et al, DOI: [10.1016/j.slasd.2024.01.005](https://doi.org/10.1016/j.slasd.2024.01.005)

Solubility 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

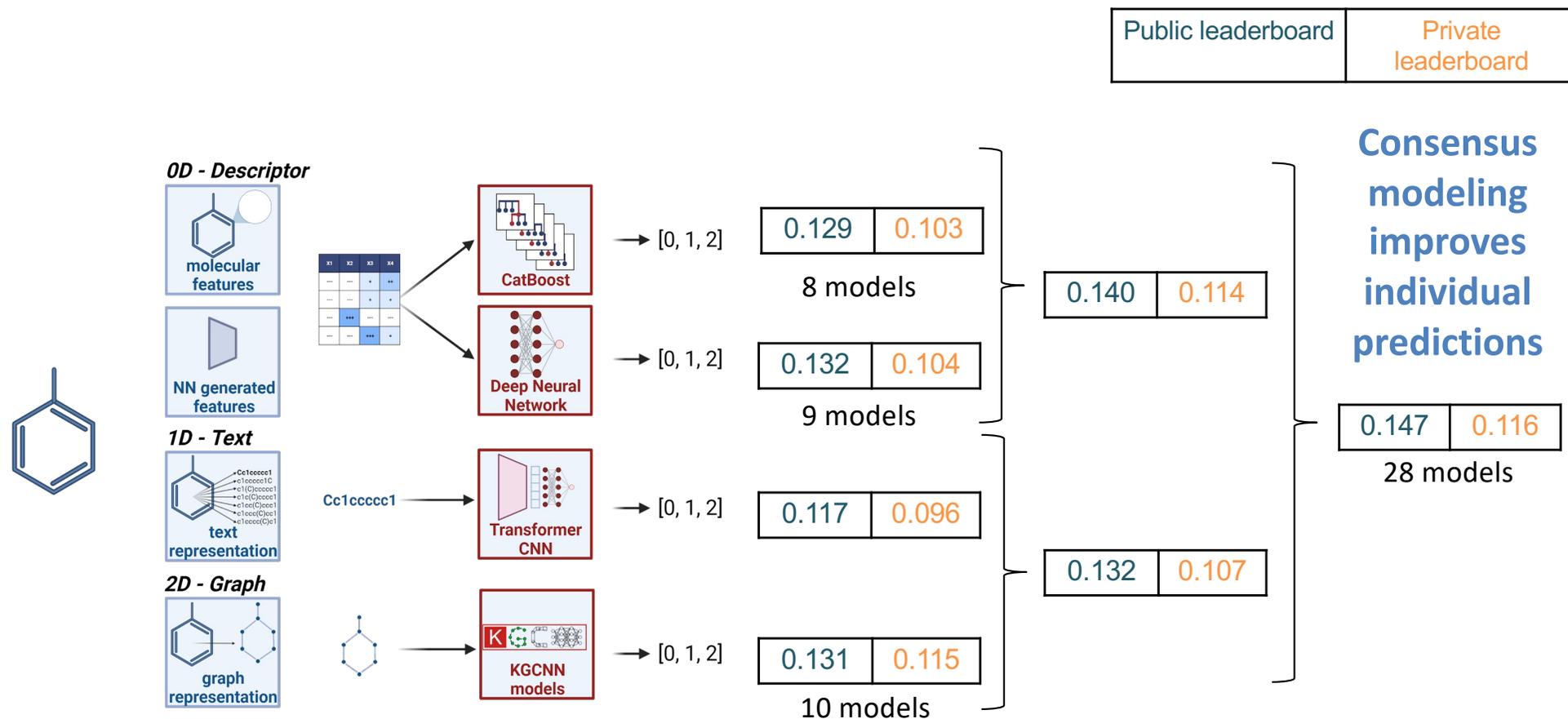


Molecular representation



@Peter Hartog with BioRender.com

Quadratic kappa metric scores



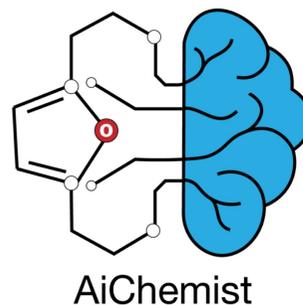
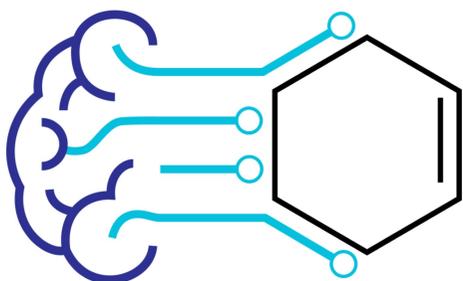
@Peter Hartog with BioRender.com

ICANN24

33rd International Conference on Artificial Neural Networks

Tox24 Challenge: How accurately can we predict binding to transthyretin?

Start:17/05 » Submit:31/08 » Winner:18/09 » Article:31/12



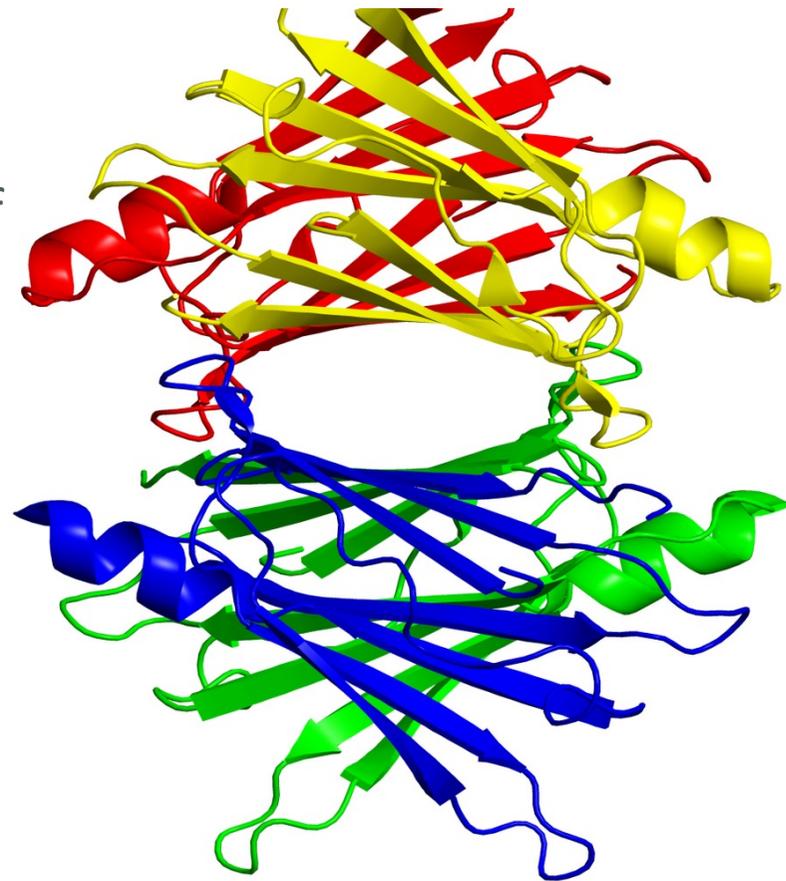
<https://ochem.eu>

<https://e-nns.org/icann2024>

Transthyretin binding and EDC

TTR is one of the serum binding proteins responsible for delivering thyroid hormones (THs) to target tissues and maintaining the balance of free versus bound THs.

The binding of compounds to TTR and subsequent displacement of TH is important to identify potential interference of the thyroid system which are endocrine disrupting chemicals (EDCs).



Tox24 Challenge

EPA submitted an article in February with screening of TTR compounds

March-April: negotiation with EPA, ChemResTox, ICANN2024, AIDD to organize the challenge

May 17th – data are publicly available

- 1012 training set
- 200 LeaderBoard set
- 300 Blind Test set

August 15th

- LeaderBoard set is available

September 1st

- Results announced

ICANN24

33rd International Conference on Artificial Neural Networks

Home **Conference** ▾ Registration Contributors ▾ Organisation ▾ 1



Tox24 Challenge

The **Tox24** challenge is designed to assess the progress in computational methods for predicting in vitro activity of compounds. All ML experts are strongly invited to participate in it and compete for a prize of **1000€**, to be awarded for the winning model.

[Tetko, I. V. Tox24 Challenge. *Chem. Res. Toxicol.* **2024**, *37* \(6\), 825–826. <https://doi.org/10.1021/acs.chemrestox.4c00192>.](https://doi.org/10.1021/acs.chemrestox.4c00192)

Tox24 Challenge results

Tox24 final ranking based on the blind set - congratulations to the winning team Amidoff!

In case of equal performances the earlier entry was used. Models with lower RMSE (Root Mean Squared Error) are better.

You can access the [winning model here](#)

rank	User	submission id	*** Blind test set RMSE ***	Leaderboard set RMSE	Submission time (GMT-12)
1	Amidoff	1148	20.5	7.8	25-08-2024 06:13:02
2	tcirino	1098	20.7	4.1	21-08-2024 08:53:04
3	znavoyan	1309	20.7	5.7	30-08-2024 07:16:19
4	uesawa	1149	20.8	21.2	25-08-2024 14:03:09
5	YingkaiZhangLab	1388	20.8	20.7	31-08-2024 06:49:33
6	AntonijaBoss	1400	21.2	20.6	31-08-2024 09:07:16
7	SankalpJain	1284	21.3	5.1	30-08-2024 02:46:16
8	alx.dga	1097	21.4	0.0	21-08-2024 08:45:46
9	luispintoc	1209	21.4	15.0	28-08-2024 06:00:04
10	GAT_Wang	1253	21.4	6.0	29-08-2024 15:26:28
11	vchupakhin	1404	21.4	3.2	31-08-2024 13:34:03

Winning model (Makarov, D.; Ksenofontov, A.)

Model name: Consensus TTR binding activity , published in [Tox24 Challenge](#)
Public ID is [1149](#)

Predicted property: **TTR binding activity** modeled in %
Training method: Consensus

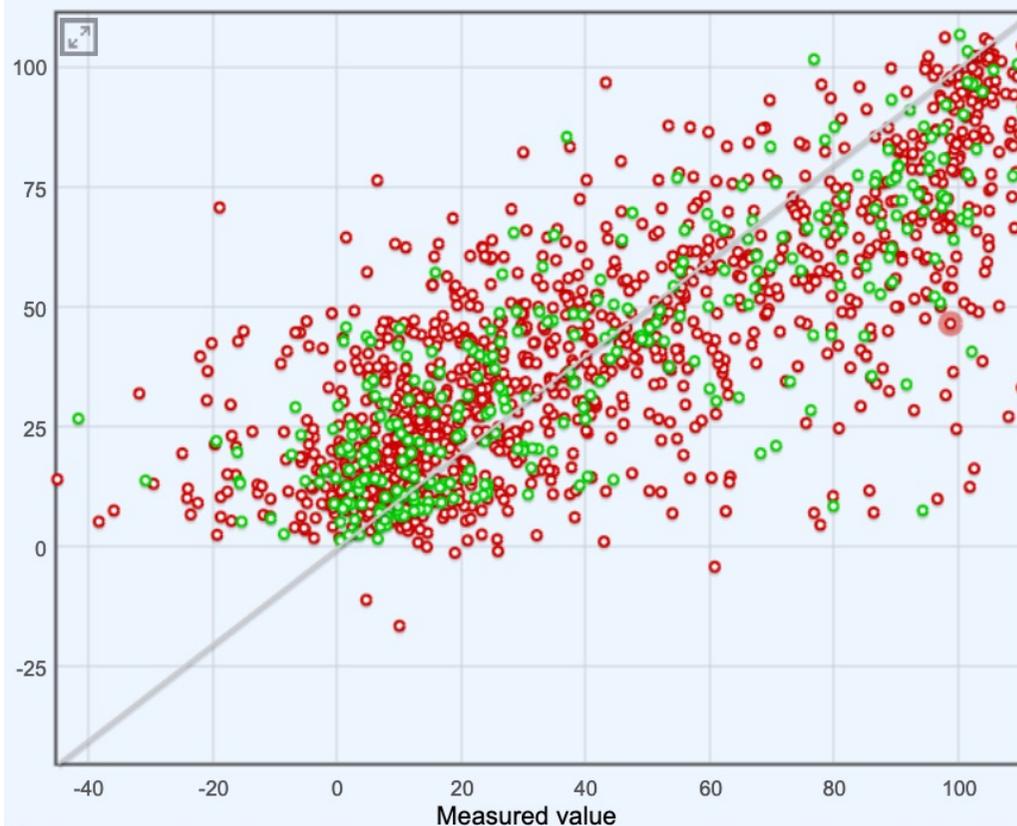
4 individual models:

ModelID: 1145
ModelID: 1146
ModelID: 1147
ModelID: 1148
AVERAGE

No validation

[Size: 1 Kb](#)

Data Set	#	R2	q2	RMSE	MAE
Training set: TTR_train	1208 records	0.62 ± 0.02	0.62 ± 0.02	22.3 ± 0.6	16.6 ± 0.4
Test set: tox24_challenge_test_data [x]	300 records	0.67 ± 0.04	0.67 ± 0.03	21 ± 1	15.6 ± 0.8



Number of compounds ignored because of errors in original model = 4

Winning model (Makarov, D.; Ksenofontov, A.)

Consensus model:

- Transformer CNN
- Transformer CNF2
- Cat Boost based on Mold2 descriptors
- Cat Boost based on ALOGPS + OESTATE descriptors

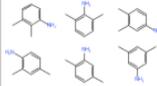
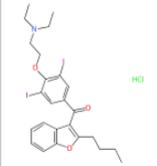
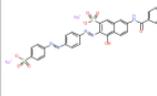
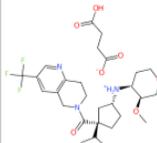
And using mixture descriptors

Compounds properties browser
Search for numerical compounds properties linked to scientific articles

Area of your interest: disconnected, [change]

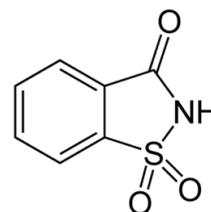
Basket Records Tags

6 - 10 of 129 5 Items on page 2 of 26

 molecule profile	<p>● TTR binding activity = 33.9 (in %)</p> <p>Eytcheson, S.A. et al Screening the ToxCast Chemical Libraries for Binding to Tran... N: 791 ChemRxiv 2024; ()</p> <p>Dimethylaniline ; 1300-73-8 MoleculeID: M16097837 DTXSID8027377</p> <p>Public and freely downloadable record (awaiting approval)</p>	<p>Dataset = Train</p> <p>RecordID: R56041905 10:37, 16 May 24 / 11:23, 20 Aug 24 itetko_acs / published</p>
 molecule profile	<p>● TTR binding activity = 48.4 (in %)</p> <p>Eytcheson, S.A. et al Screening the ToxCast Chemical Libraries for Binding to Tran... N: 468 ChemRxiv 2024; ()</p> <p>Amiodarone hydrochloride ; 19774-82-4 MoleculeID: M118366 DTXSID7037185</p> <p>Public and freely downloadable record (awaiting approval)</p>	<p>Dataset = Train</p> <p>RecordID: R56041685 10:36, 16 May 24 / 15:46, 16 May 24 itetko_acs / published</p>
 molecule profile	<p>● TTR binding activity = 103.4 (in %)</p> <p>Eytcheson, S.A. et al Screening the ToxCast Chemical Libraries for Binding to Tran... N: 577 ChemRxiv 2024; ()</p> <p>2610-11-9 ; C.I. Direct red 81 disodium salt MoleculeID: M51977 DTXSID5041726</p> <p>Public and freely downloadable record (awaiting approval)</p>	<p>Dataset = Train</p> <p>RecordID: R56041755 10:37, 16 May 24 / 11:23, 20 Aug 24 itetko_acs / published</p>
 molecule profile	<p>● TTR binding activity = 20.9 (in %)</p> <p>Eytcheson, S.A. et al Screening the ToxCast Chemical Libraries for Binding to Tran... N: 1099 ChemRxiv 2024; ()</p> <p>851916-42-2 ; MK-812 MoleculeID: M3407009 DTXSID4047335</p> <p>Public and freely downloadable record (awaiting approval)</p>	<p>Dataset = Train</p> <p>RecordID: R56042116 10:37, 16 May 24 / 15:46, 16 May 24 itetko_acs / published</p>

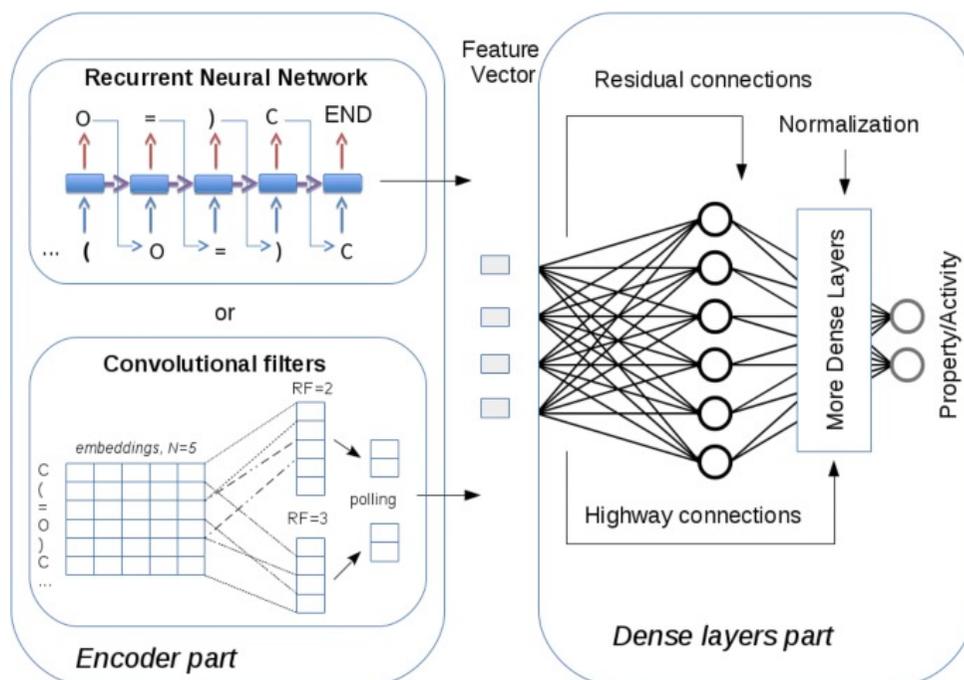
Machine Learning directly from chemical structures

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

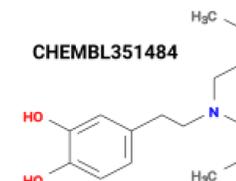


Text processing: convolutional neural networks, transformers, LSTM

Graph processing: message passing neural networks



non-canonical	>>	canonical
<chem>c1e(ccc(O)c1O)CCN(CCCC)CCC</chem>	>>	<chem>CCCCN(CCc1ccc(O)c1)O)CCC</chem>
<chem>c1(ccc(O)c1)O)CCN(CCC)CCCC</chem>	>>	<chem>CCCCN(CCc1ccc(O)c1)O)CCC</chem>
<chem>c1e(ccc(O)c1)CCN(CCCC)CCC)O</chem>	>>	<chem>CCCCN(CCc1ccc(O)c1)O)CCC</chem>
<chem>c1(CCN(CCCC)CCC)ccc(O)c1</chem>	>>	<chem>CCCCN(CCc1ccc(O)c1)O)CCC</chem>
<chem>CCCCN(CCc1ccc(O)c1)O)CCC</chem>	>>	<chem>CCCCN(CCc1ccc(O)c1)O)CCC</chem>
<chem>CCCCN(CCc1ccc(O)c1)O)CCC</chem>	>>	<chem>CCCCN(CCc1ccc(O)c1)O)CCC</chem>
<chem>N(CCCC)(CCc1ccc(O)c1)O)CCC</chem>	>>	<chem>CCCCN(CCc1ccc(O)c1)O)CCC</chem>
<chem>C(N(CCc1ccc(O)c1)O)CCCC)CC</chem>	>>	<chem>CCCCN(CCc1ccc(O)c1)O)CCC</chem>
<chem>N(CCc1ccc(O)c1)O)CCCC)CCC</chem>	>>	<chem>CCCCN(CCc1ccc(O)c1)O)CCC</chem>
<chem>c1e(O)c(ccc1CCN(CCC)CCCC)O</chem>	>>	<chem>CCCCN(CCc1ccc(O)c1)O)CCC</chem>
<chem>c1(c(ccc1)CCN(CCC)CCCC)O)O</chem>	>>	<chem>CCCCN(CCc1ccc(O)c1)O)CCC</chem>



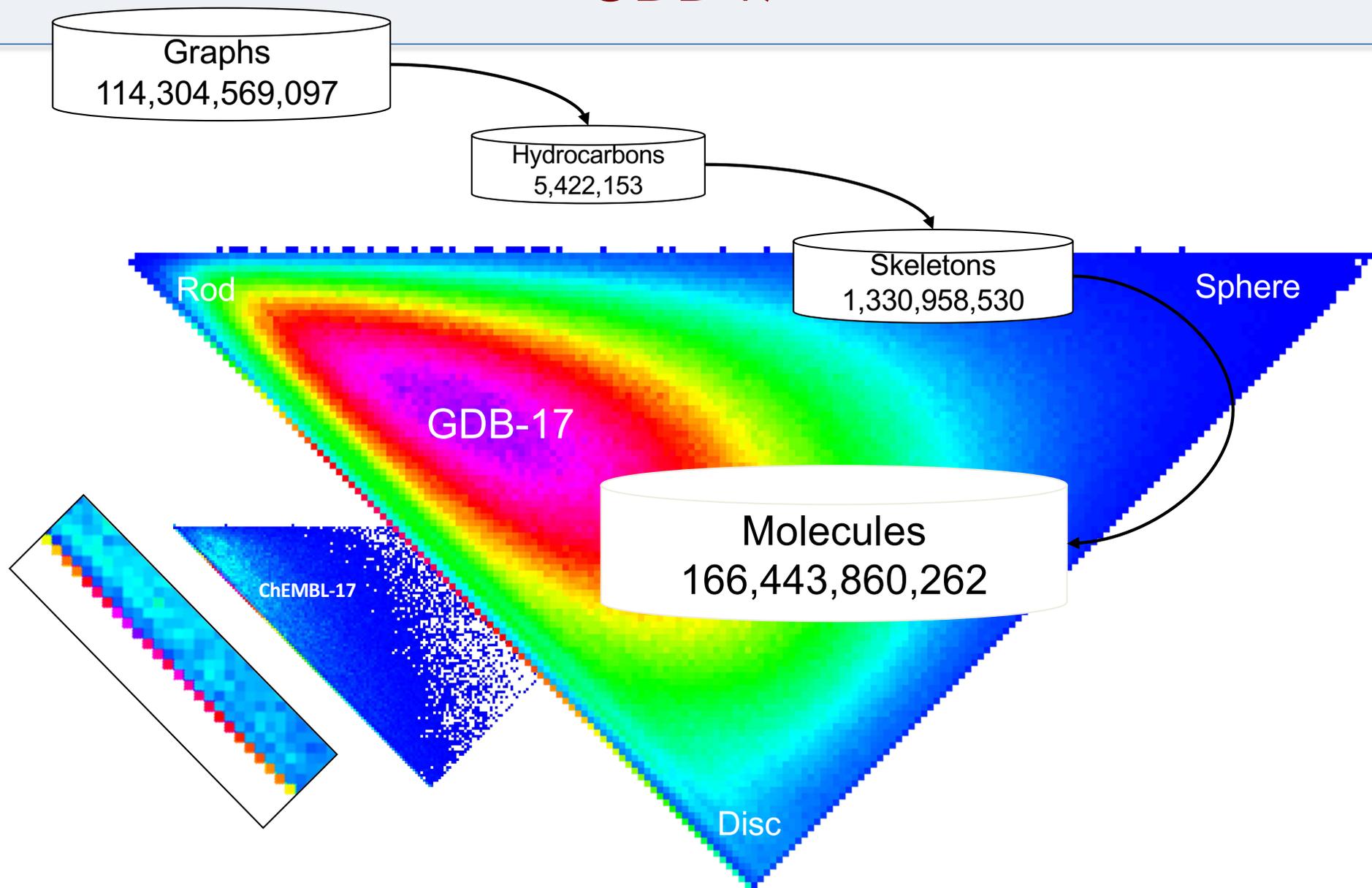
CheMBL database was used for pretraining

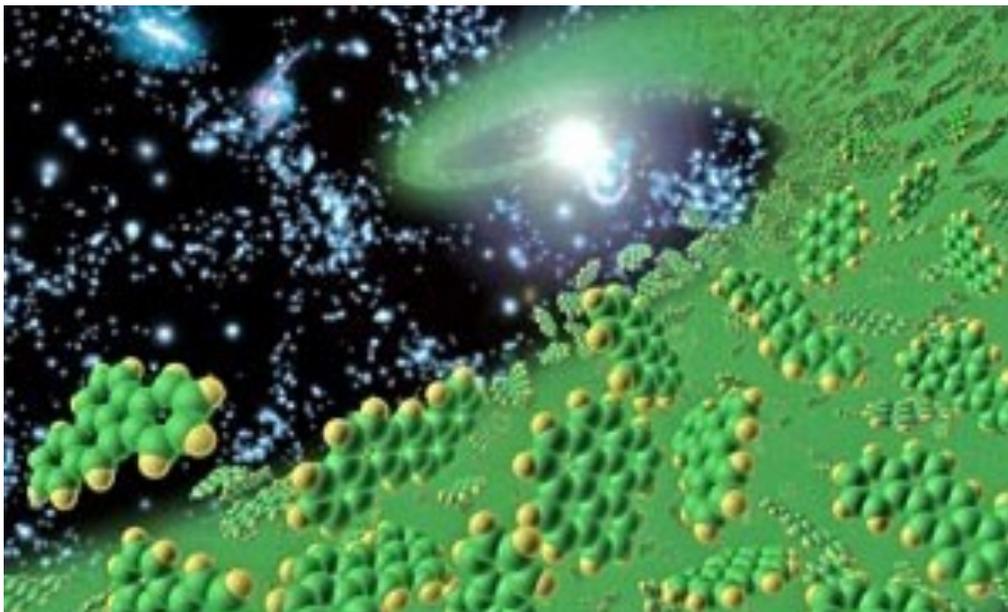
Image augmentation



<https://github.com/aleju/imgaug>

GDB-17

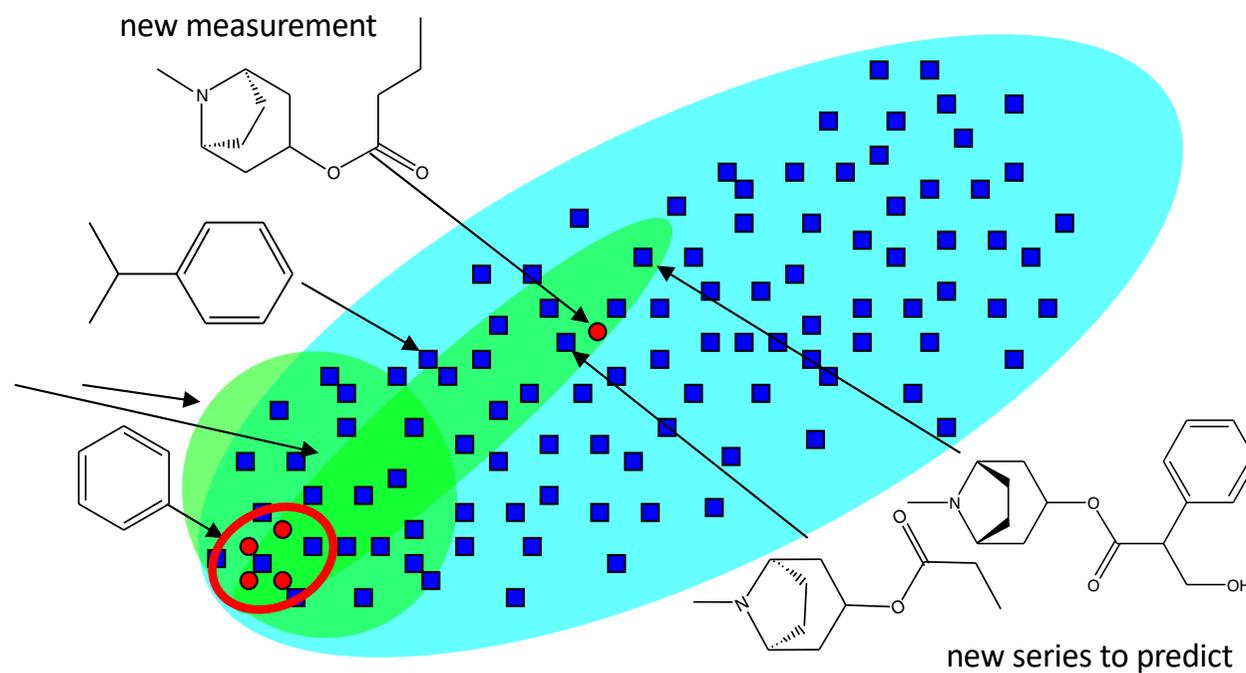




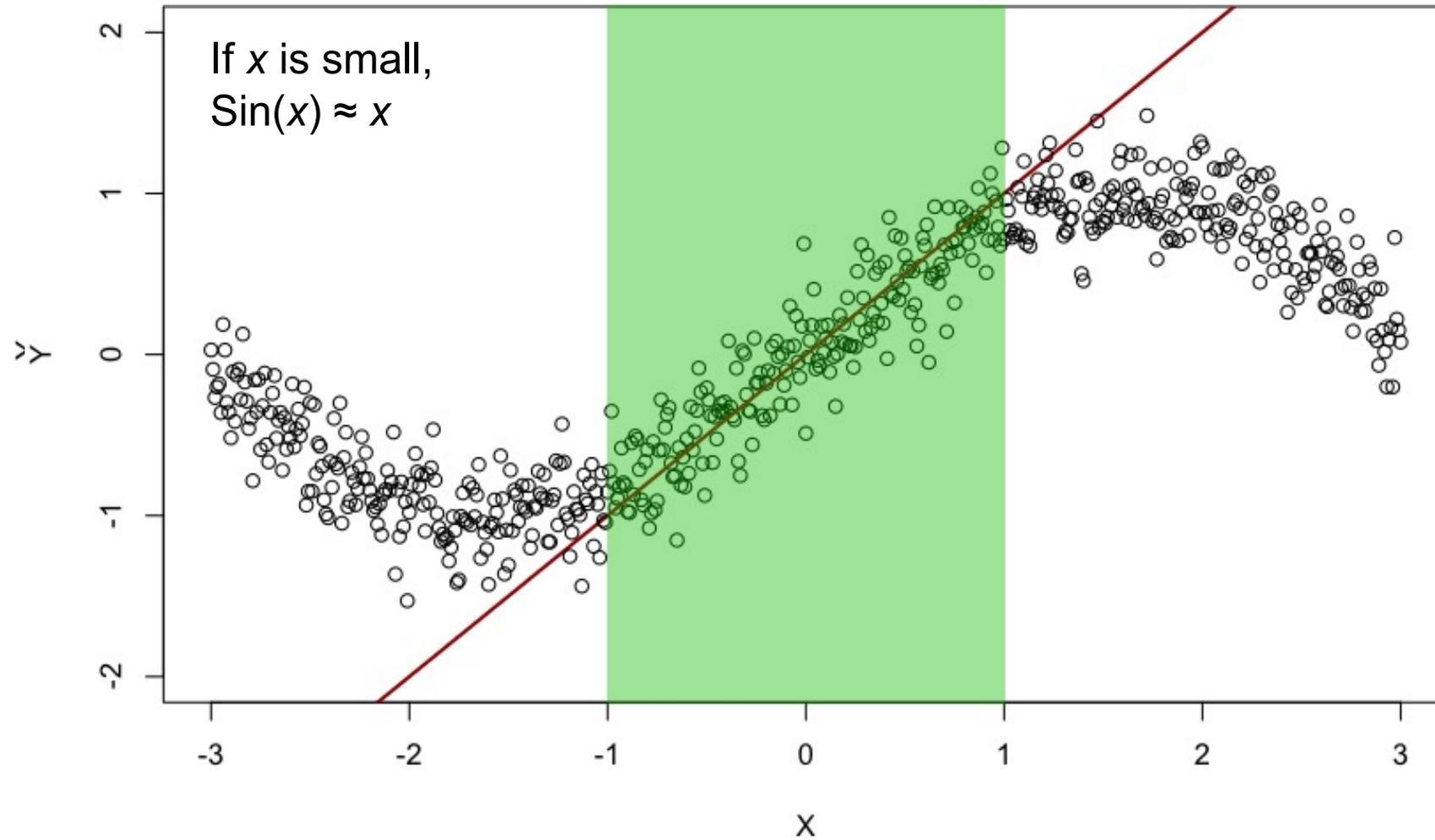
But: we can't predict unpredictable!

10^5 (Soulbility set) \rightarrow 10^{11} (GDB17)

1 \rightarrow 1,000,000



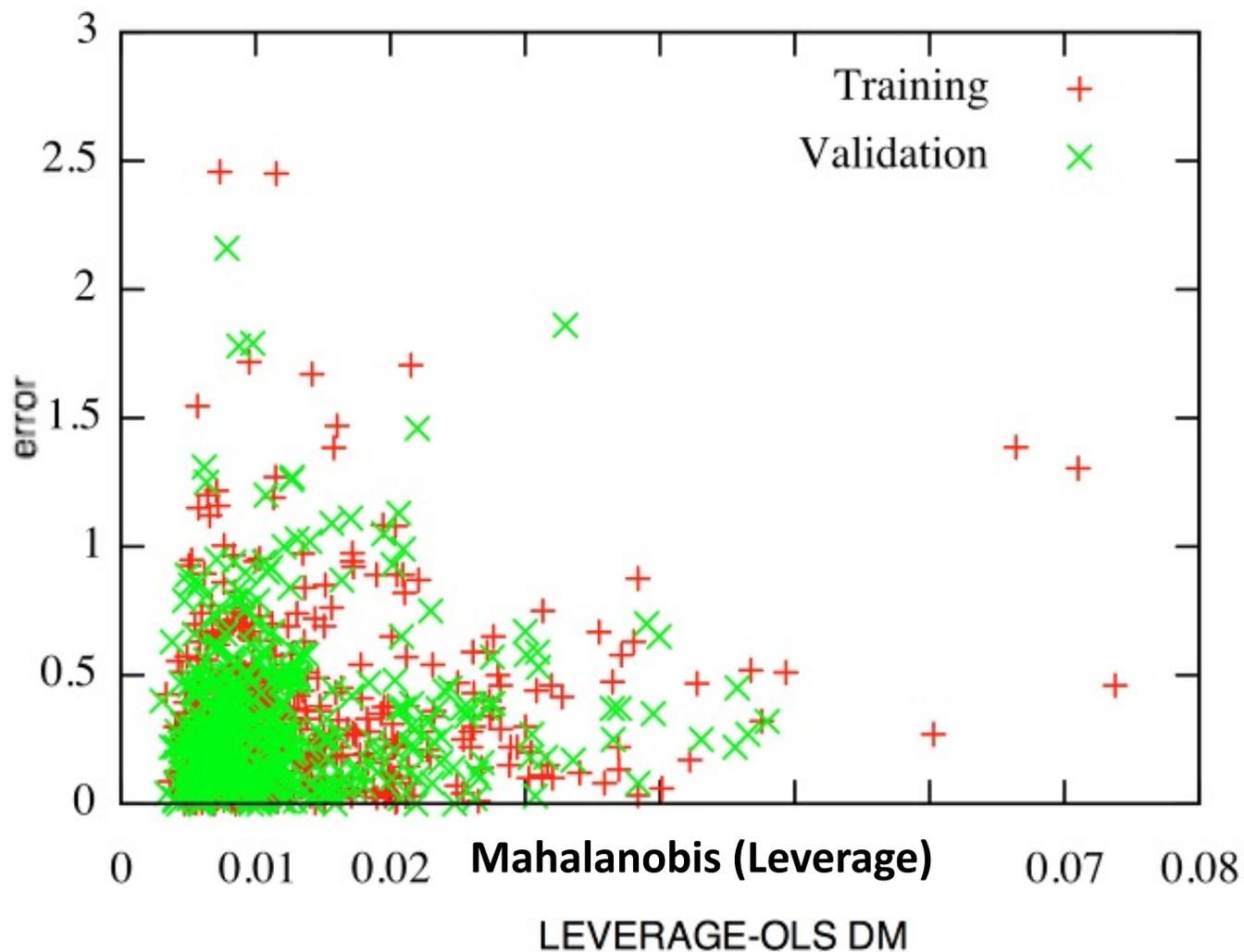
Accuracy of prediction



Overview of analyzed distances to models (DMs)

<p>EUCLID</p> $EU_m = \frac{\sum_{j=1}^k d_j}{k}$ <p>k is number of nearest neighbors, m index of model</p> $EUCLID = E\bar{U}_m$	<p>TANIMOTO</p> $Tanimoto(a,b) = \frac{\sum x_{a,i}x_{b,i}}{\sum x_{a,i}x_{a,i} + \sum x_{b,i}x_{b,i} - \sum x_{a,i}x_{b,i}}$ <p>$x_{a,i}$ and $x_{b,i}$ are fragment counts</p>
<p>LEVERAGE</p> $LEVERAGE = \mathbf{x}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}$	<p>PLSEU (DModX)</p> <p>Error in approximation (restoration) of the vector of input variables from the latent variables and PLS weights.</p>
<p>STD</p> $STD = \frac{1}{N-1} \sum (y_i - \bar{y})^2$ <p>y_i is value calculated with model i and \bar{y} is average value</p>	<p>CORREL</p> $CORREL(a) = \max_j CORREL(a,j) = R^2(\mathbf{Y}^a_{calc}, \mathbf{Y}^j_{calc})$ <p>$\mathbf{Y}^a = (y_1, \dots, y_N)$ is vector of predictions of molecule i</p>

Descriptor space, ASNN model: DM does not work



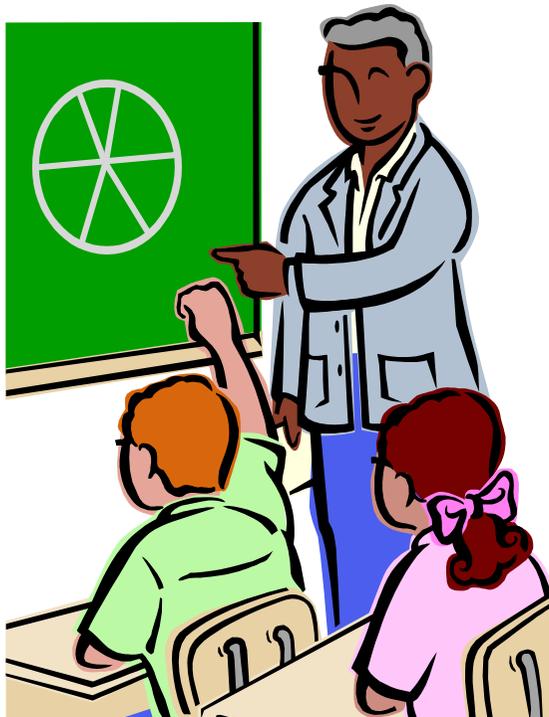
Tetko et al, *J. Chem. Inf. Model*, **2008**, 48, 1733-46.

Consensus modelling

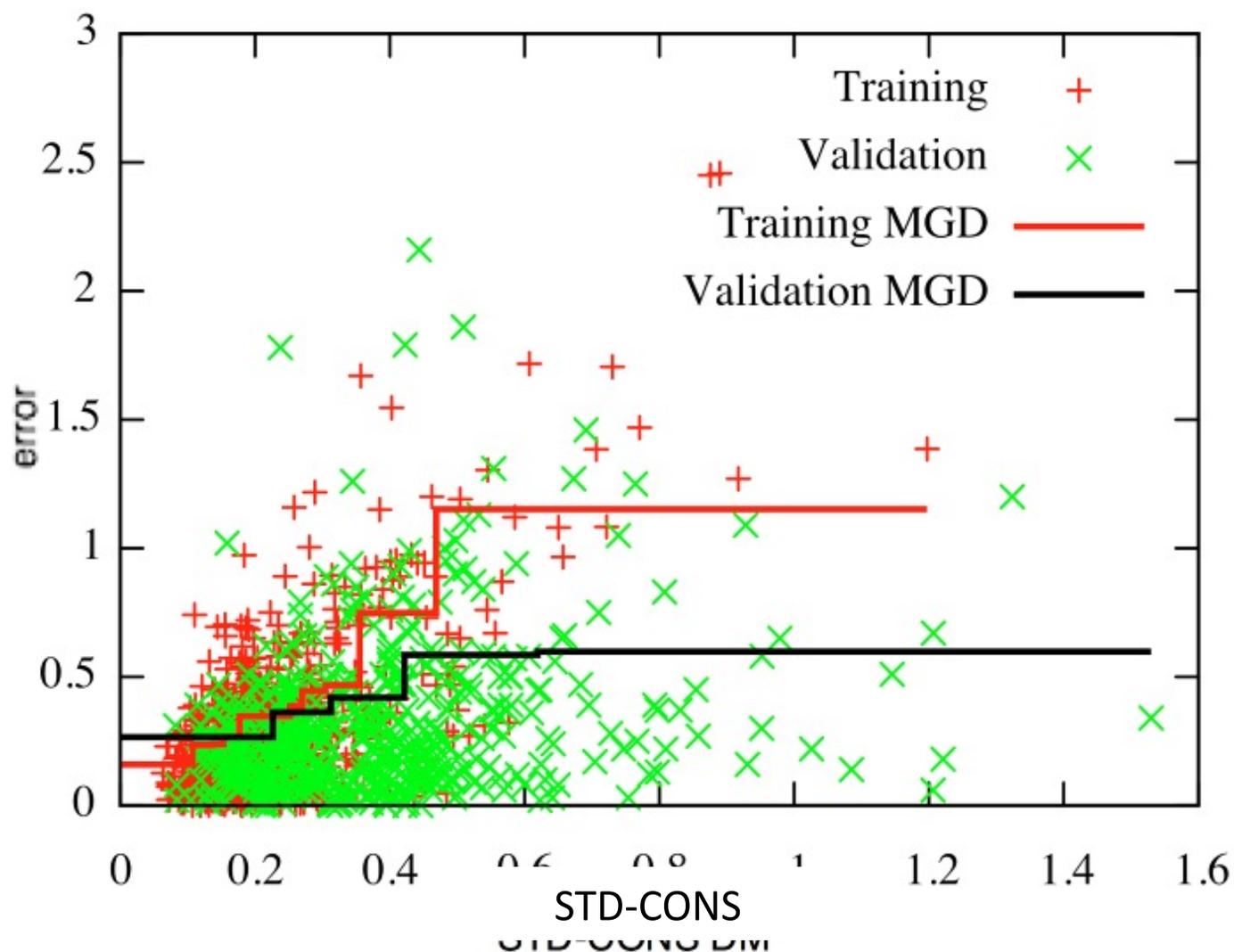
Best method(s) are defined

Average prediction of models is used

The consensus prediction is more accurate and stable

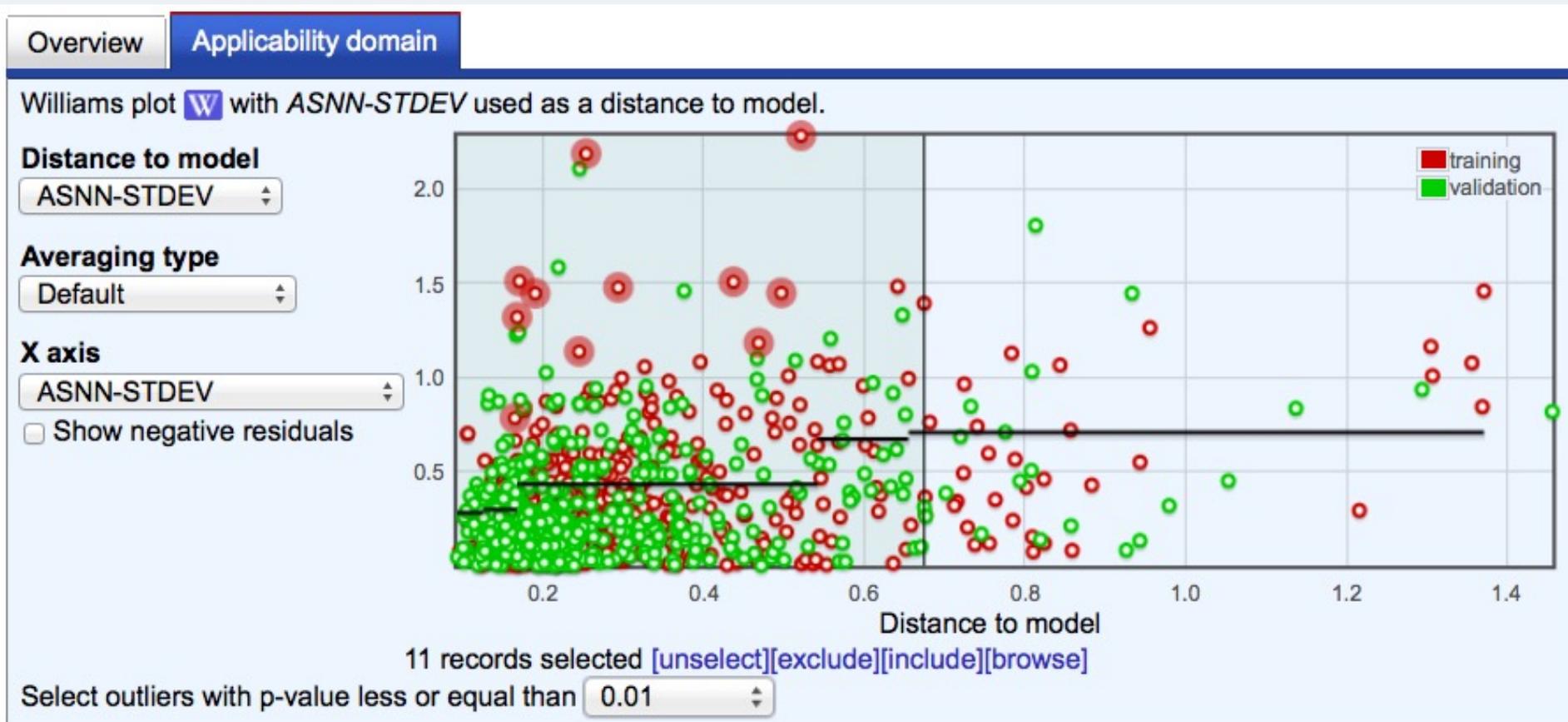


Property-based, ASNN model: DM does work!



Tetko et al, *J. Chem. Inf. Model*, **2008**, 48, 1733-46.

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*

Accuracy of predictions for classification model

Overview

Applicability domain

Model name: Ames levenberg , published in [Applicability domains for classification problems: Benchmarking of distance to models for Ames mutagenicity set](#). public identifier is 1

Predicted property: AMES

Training method: ANN

[OEstate] Correl. limit: 0.95 Variance threshold: 0.0, Maximum

value: 999999,

Levenberg, 1000 iterations, 3 neurons

ensemble=100 additional param PARALLEL=10

5-fold cross-validation

Data Set	#	Accuracy	Balanced accuracy
Training set: Ames challenge training (4359 selected)	4357 records	78.1 ± 1.2	77.9 ± 1.3
Test set: Ames challenge test [x]	2181 records	79.9 ± 1.7	79.8 ± 1.7

Calculated in 2402 seconds

Size: 450 Kb

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

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

Overview

Applicability domain

Wilkinson plot [W](#) with ASNN-STDEV used as a distance to model.

Distance to model

ASNN-STDEV

Averaging type

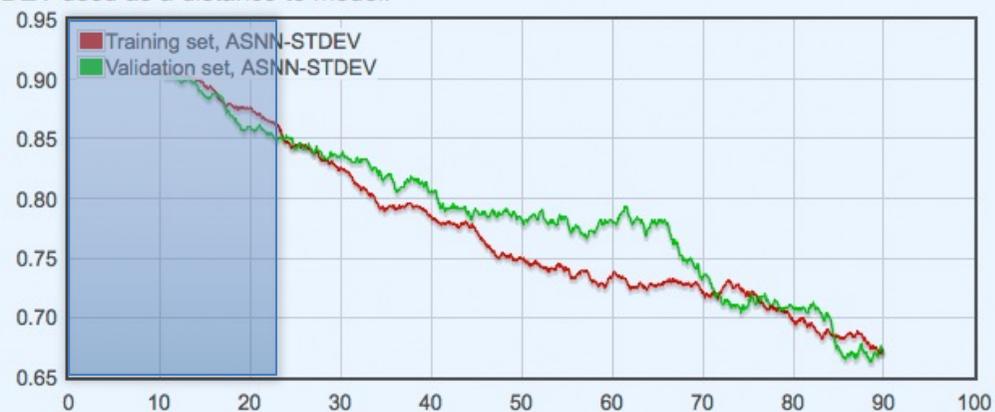
Default

Window size:

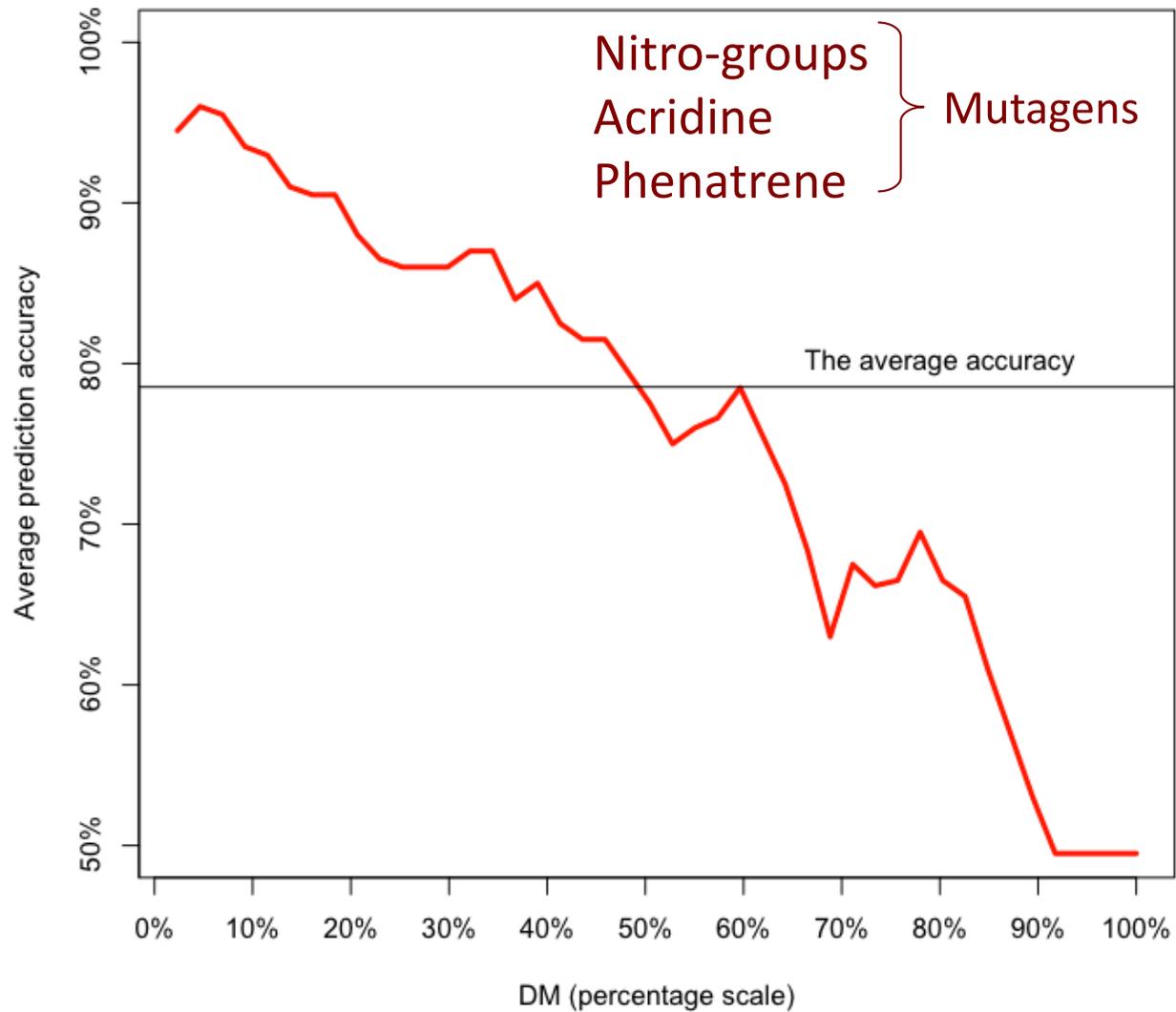
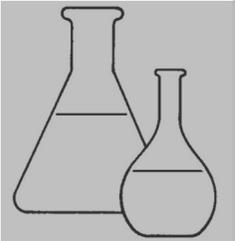
20% of the set

X axis

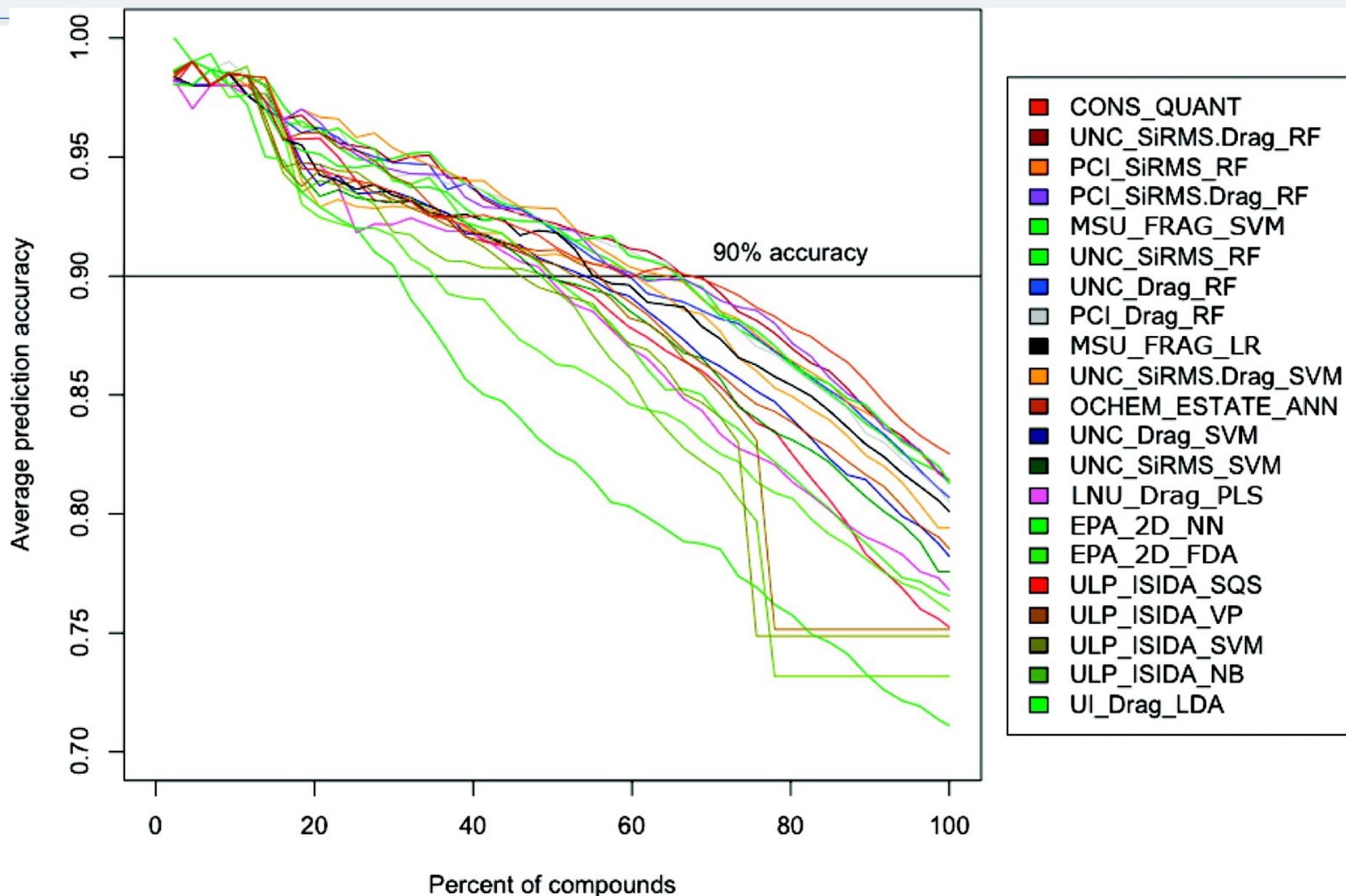
Percentage of compounds



Discriminative power of model for AMES test



Accuracy of all models for AMES test set



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

Model explanations



> A to Z

Google Custom search



OECD Home

About

Countries ▾

Topics ▾

COVID-19

Ukraine

> Français

[OECD Home](#) > [Chemical safety and biosafety](#) > [Assessment of chemicals](#) > Validation of (Q)SAR Models

Validation of (Q)SAR Models

Although a variety of (Q)SAR models have been developed, and some models have been used in assessment of chemicals in some countries for many years, transparent validation process and objective determination of the reliability of (Q)SAR models are crucial in order to further enhance the regulatory acceptance of (Q)SAR models.

In November 2004, the OECD member countries agreed on the principles for validating (Q)SAR models for their use in regulatory assessment of chemical safety. The agreed principles provide member countries with basis for evaluating regulatory applicability of (Q)SAR models and will contribute to their enhanced use for more efficient assessment of chemical safety.

[OECD principles for the Validation, for Regulatory Purpose, of \(Q\)SAR Models](#)

A full report from the OECD Expert Group on (Q)SARs was also published in 2004:

[The report from the Expert Group on \(Q\)SARs on the validation of \(Q\)SARs](#)

In February 2007, the OECD published a "Guidance Document on the Validation of (Q)SAR Models" with the aim of providing guidance on how specific (Q)SAR models can be evaluated with respect to the OECD principles. A check list for the validation, a reporting format for the validation and validation case studies are attached as annexes:

[Guidance Document on the Validation of \(Q\)SAR Models](#)

In November 2004, the 37th OECD's Joint Committee and the Working Party on Chemicals, Pesticides and Meeting of the Chemicals Biotechnology (Joint Meeting) agreed on the OECD Principles for the Validation, for Regulatory Purposes, of (Q)SAR Models.

The OECD Principles of (Q)SAR Validation

To facilitate the consideration of a (Q)SAR model for regulatory purposes, it should be associated with the following information:

- 1.a defined endpoint;
- 2.an unambiguous algorithm;
- 3.a defined domain of applicability;
- 4.appropriate measures of goodness-of-fit, robustness and predictivity;
- 5.a mechanistic interpretation, if possible.**

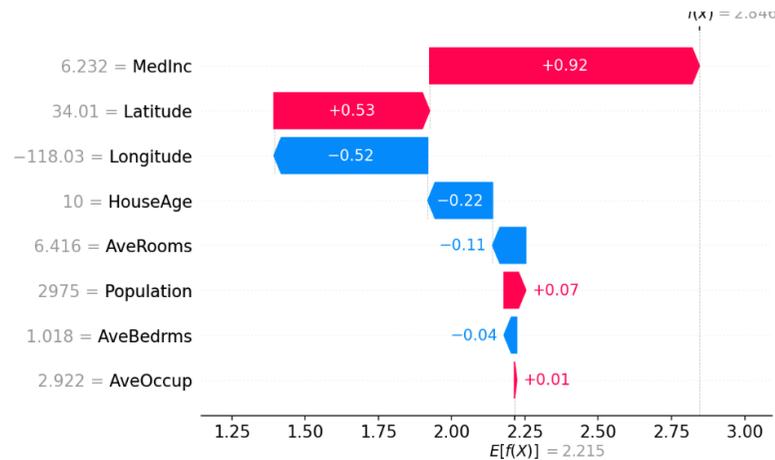
Model agnostic methods: SHAP values, LIME

EPFL

SHAP values

Lundberg, Scott M., and Su-In Lee. "A unified approach to interpreting model predictions." *Advances in neural information processing systems* 30 (2017).

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} [f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S)]$$



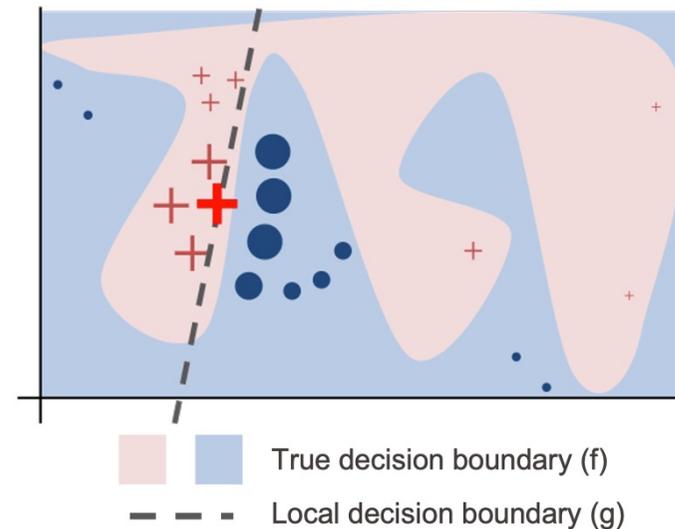
<https://shap.readthedocs.io/>

LIAC
Geemi Wellawatte

LIME

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "Why should i trust you?" Explaining the predictions of any classifier." *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 2016.

$$\xi(x) = \operatorname{argmin}_{g \in G} \mathcal{L}(f, g, \pi_x) + \Omega(g)$$



See full lecture of Dr. Wellawatte at <https://ai-dd.eu/lectures>

Linear classification model for AMES test

Mopac2016 descriptors

$$Y = 0.5378 - 0.3411 * \text{MullikenElectronegativity} - 0.3277 * \text{LumoEnergy} + 0.2389 * \text{IonisationPotential} + 0.1178 * \text{FinalHeat} + 0.05051 * \text{DipolPointCharge}$$

GSFRAG

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

E-state descriptors

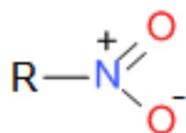
$$Y = 0.5375 + 0.09956 * \text{PSA} + 0.08731 * a\text{CNOS} - 0.08703 * \text{DONORS} - 0.06814 * \text{SsCH3} - 0.04474 * \text{SssO}$$

Dragon descriptors

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

Structural Alerts

$$Y = 0.4733 + 0.191 * \text{Alert146} - 0.004113 * \text{Alert238} - 0.1024 * \text{Alert213} + 0.1912 * \text{Alert214} + 0.01988 * \text{Alert196}$$



Nonmetals

H	C	N	O	F	Ne
He		P	S	Cl	Ar
		Se	Br		Kr
			I		Xe
			At		Rn

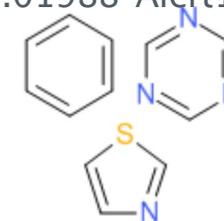
Aryl halides

R = aryl

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

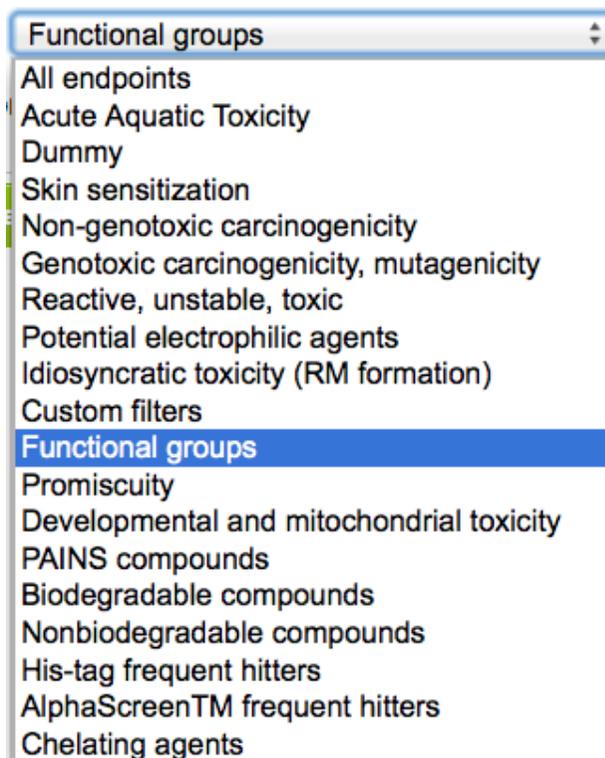
Alkyl halides

R = alkyl



ToxAlerts

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



Functional groups

Online chemical database
with modeling environment

Welcome, Guest! [Logout](#)

Home Database Models A+ a-

ToxAlerts: Structural alerts browser
Here you can browse structural alerts for various toxicological endpoints

[Upload new alerts](#) [Screen compounds](#)

101 - 200 of 379 << < 100 items on page 2 of 4 > >>

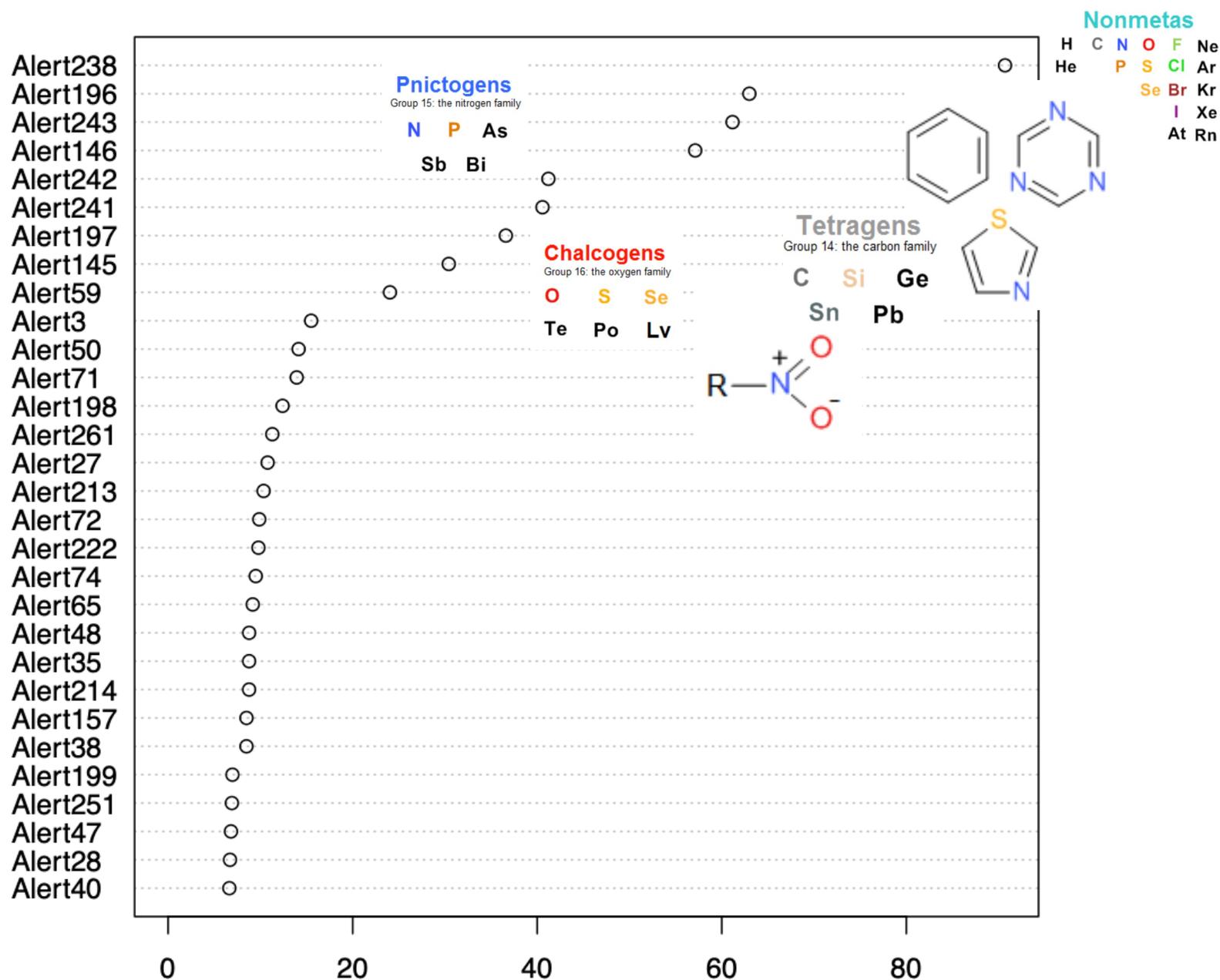
HS 	Four-membered heterocycles with one heteroatom (HS) A = any atom except carbon; a dashed line indicates any type of covalent bonds High specificity (HS) pattern matches chemicals that include exact heterocyclic moiety as in the depiction (fusion with other ring(s) are not allowed) SMARTS: <chem>[*6]#1;R1]1-[*6R1]-[*6R1]-[*6R1]-1</chem> Endpoint: Functional groups Salmina, E. Extended functional groups (EFG): an efficient set for chemi... Molecules 2015 ; subm () Alert ID: TA2450	16:45, 8 Mar 13 / 13:25, 31 Oct 15 SALMINA1987 / itelko
HS 	Saturated four-membered heterocycles with one heteroatom (HS) A = any atom except carbon High specificity (HS) pattern matches chemicals that include exact heterocyclic moiety as in the depiction (fusion with other ring(s) are not allowed) SMARTS: <chem>[*6]#1;R1]1-[*6R1]-[*6R1]-[*6R1]-1</chem> Endpoint: Functional groups Salmina, E. Extended functional groups (EFG): an efficient set for chemi... Molecules 2015 ; subm () Alert ID: TA2451	16:45, 8 Mar 13 / 13:25, 31 Oct 15 SALMINA1987 / itelko
HS 	Azetidines (HS) High specificity (HS) pattern matches chemicals that include exact heterocyclic moiety as in the depiction (fusion with other ring(s) are not allowed) SMARTS: <chem>[*7R1]1-[*6R1]-[*6R1]-[*6R1]-1</chem> Endpoint: Functional groups Salmina, E. Extended functional groups (EFG): an efficient set for chemi... Molecules 2015 ; subm () Alert ID: TA2452	16:45, 8 Mar 13 / 13:25, 31 Oct 15 SALMINA1987 / itelko
HS 	Oxetanes (HS) High specificity (HS) pattern matches chemicals that include exact heterocyclic moiety as in the depiction (fusion with other ring(s) are not allowed) SMARTS: <chem>[*8R1]1-[*6R1]-[*6R1]-[*6R1]-1</chem> Endpoint: Functional groups Salmina, E. Extended functional groups (EFG): an efficient set for chemi... Molecules 2015 ; subm () Alert ID: TA2453	16:45, 8 Mar 13 / 13:25, 31 Oct 15 SALMINA1987 / itelko
	Thietanes (HS) High specificity (HS) pattern matches chemicals that include exact heterocyclic moiety as in the depiction (fusion with other ring(s) are not allowed)	

Overrepresented functional groups (AMES)

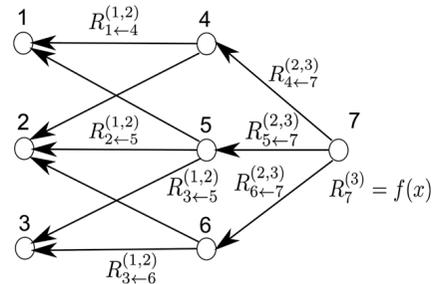
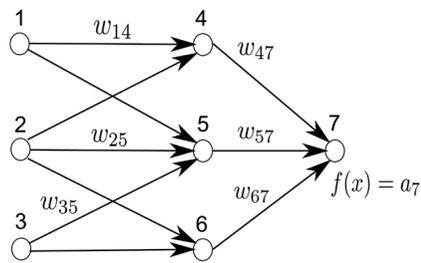
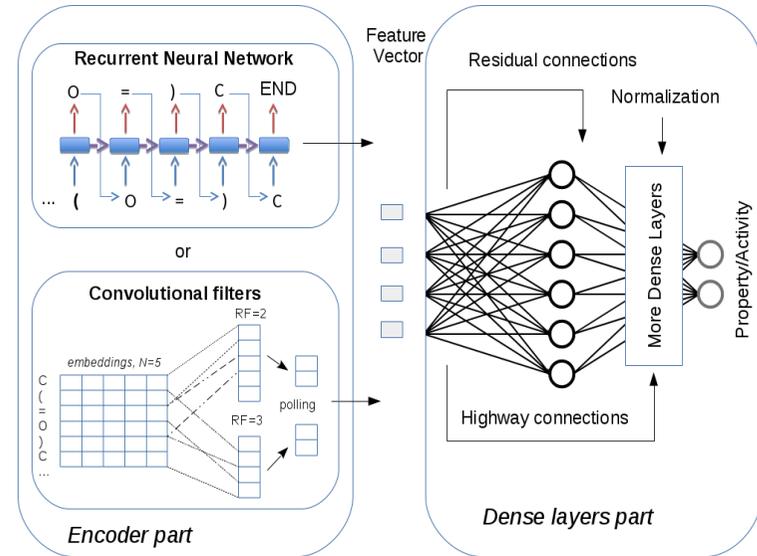
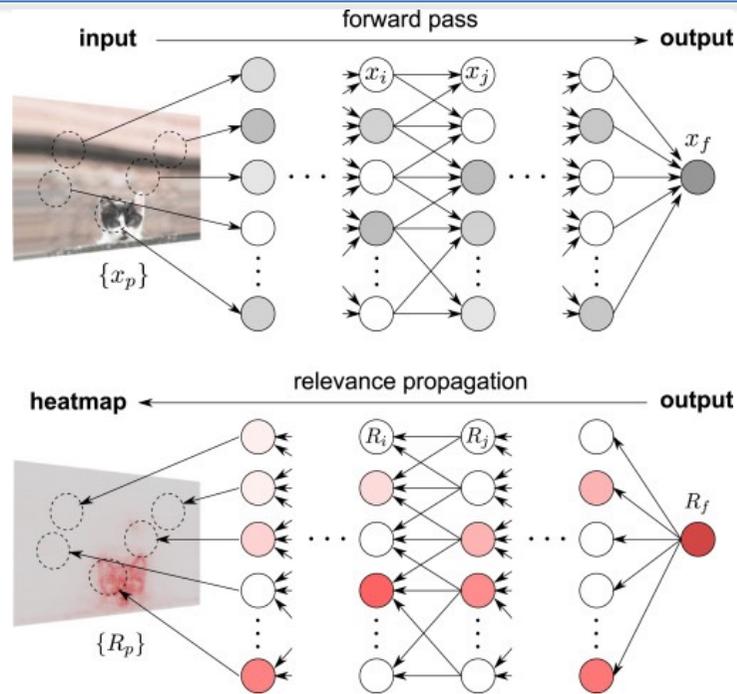
Active Inactive

Descriptor	In set 1 (3514 unique molecules)	In set 2 (3023 unique molecules)	Enrichment factor	p-Value
	808 (23.0%)	167 (5.5%)	4.2	1.545E-94
Pnictogens Group 15: the nitrogen family N P As Sb Bi	2498 (71.1%)	1714 (56.7%)	1.3	5.687E-34
	219 (6.2%)	28 (0.9%)	6.7	2.796E-33
	2546 (72.5%)	1782 (58.9%)	1.2	7.944E-31
	2770 (78.8%)	2004 (66.3%)	1.2	3.461E-30

Importance of ToxAlerts in Random Forest model



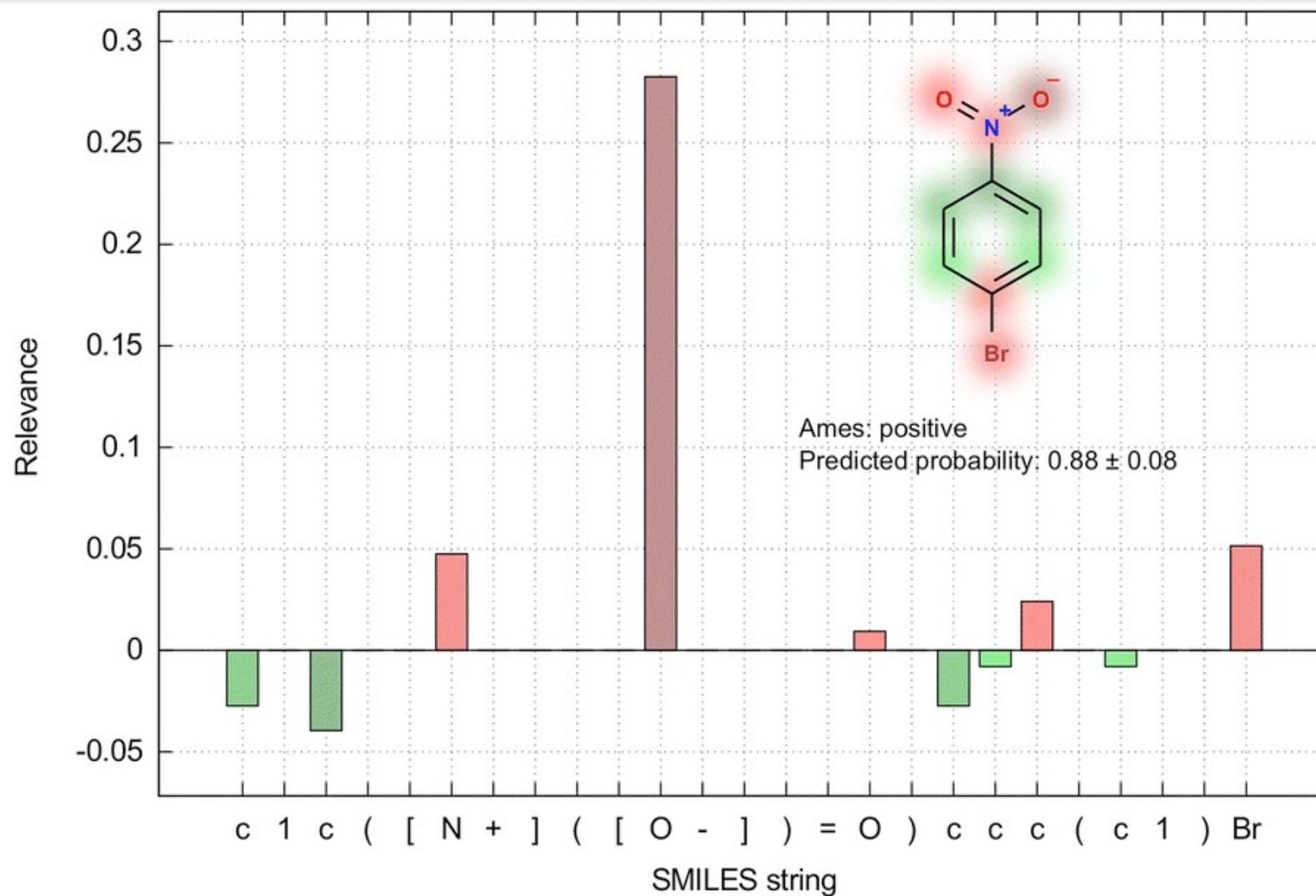
Layerwise Relevance Propagation (LRP)



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

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

Interpretation of models



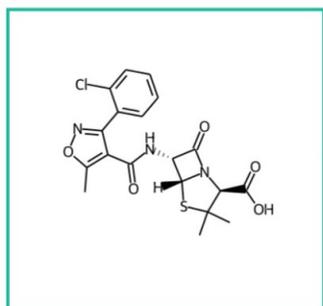
P. Karpov, G. Godin, I. V. Tetko, *J. Cheminform.* **2020**, *12*, 17.

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

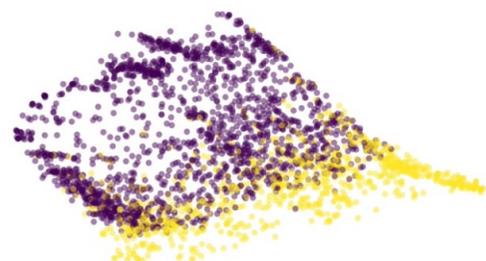
Contrafactual examples – Molecular Model Agnostic Counterfactual Explanations

EPFL

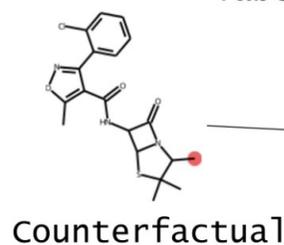
1. Molecule being predicted: base



2. Expand chemical space around base

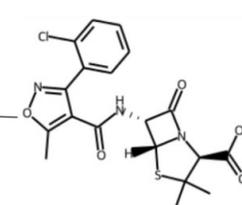


3. Identify most similar molecule with changed label: counterfactual



● Counterfactual
● Same Class

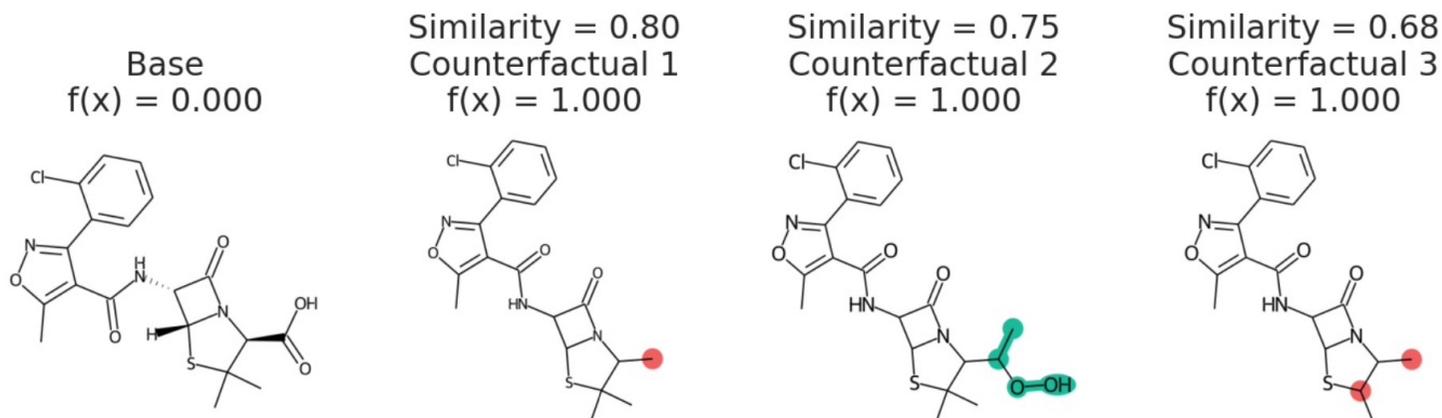
Base



Wellawatte, G. P., Seshadri, A., & White, A. D. (2022). *Chemical science*, 13(13), 3697-3705.

Example of interpretation of RF model for BBP

EPFL



Explanation: The negative example can be made to cross the blood brain barrier if the carboxylic group is altered.

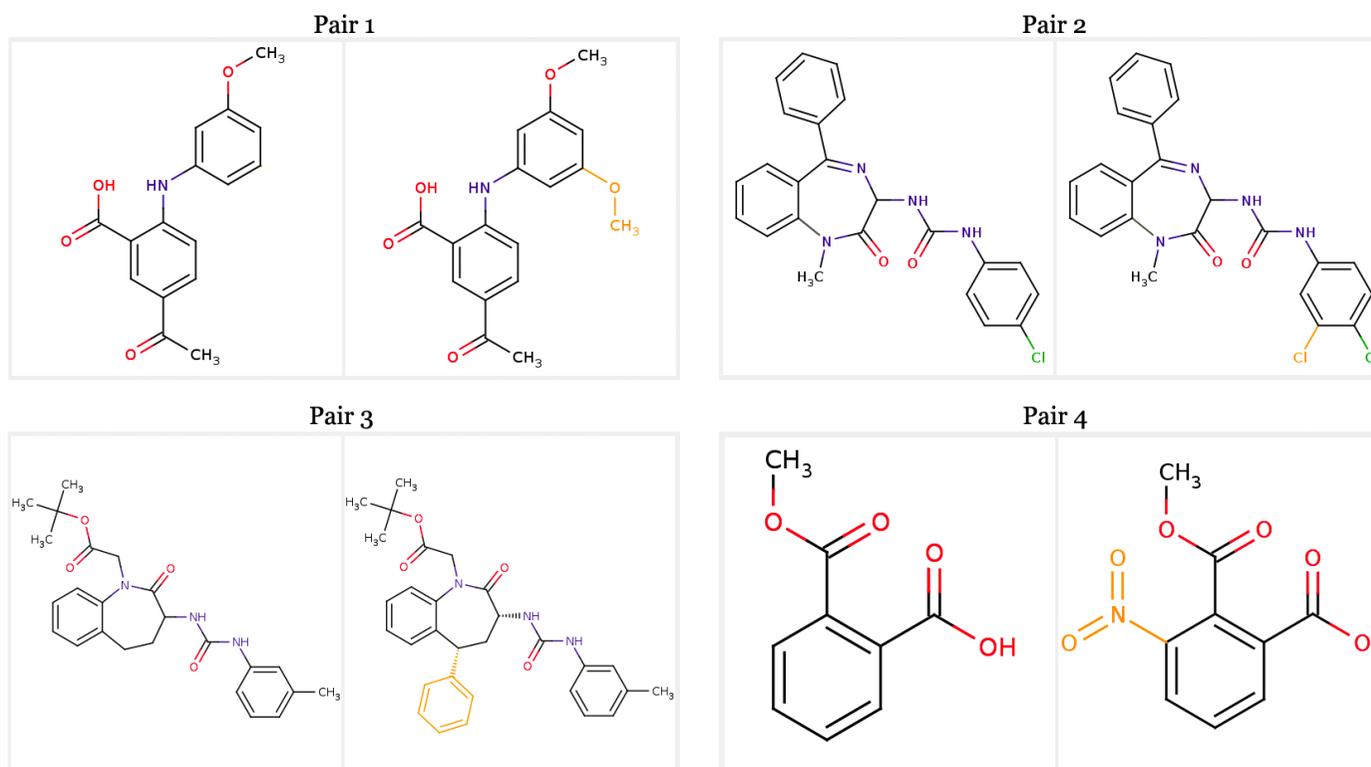
Experimental observations: hydrophobic interactions and surface area govern BBB permeation (Boobier S, *et al.*, *Nat Commun.* 2020)

Wellawatte, G. P., Seshadri, A., & White, A. D. (2022). *Chemical science*, 13(13), 3697-3705.

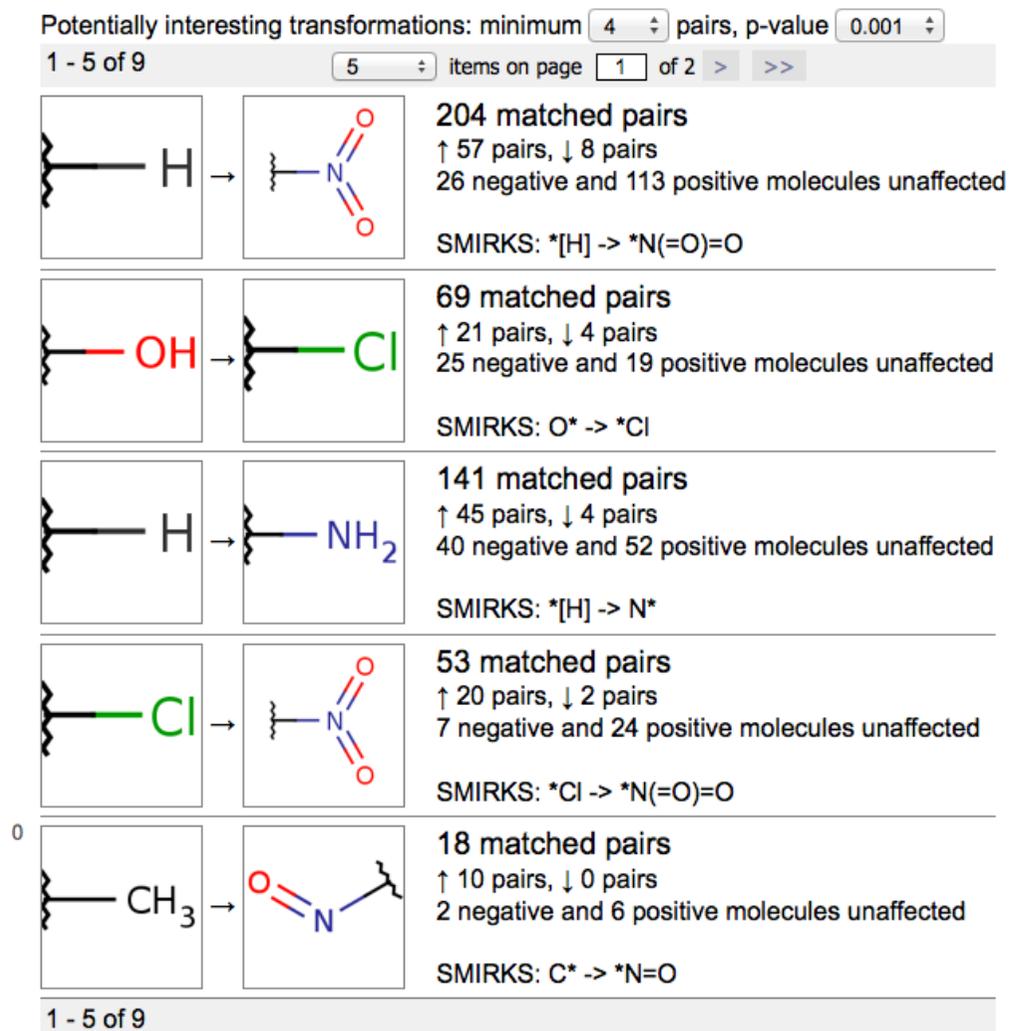
MMP definition

A **molecular matched pair** (MMP) is a pair of molecules that have only a (minor) single-point difference.

The typical way is to define a minor difference as a changed molecular fragment with less than 10 atoms.

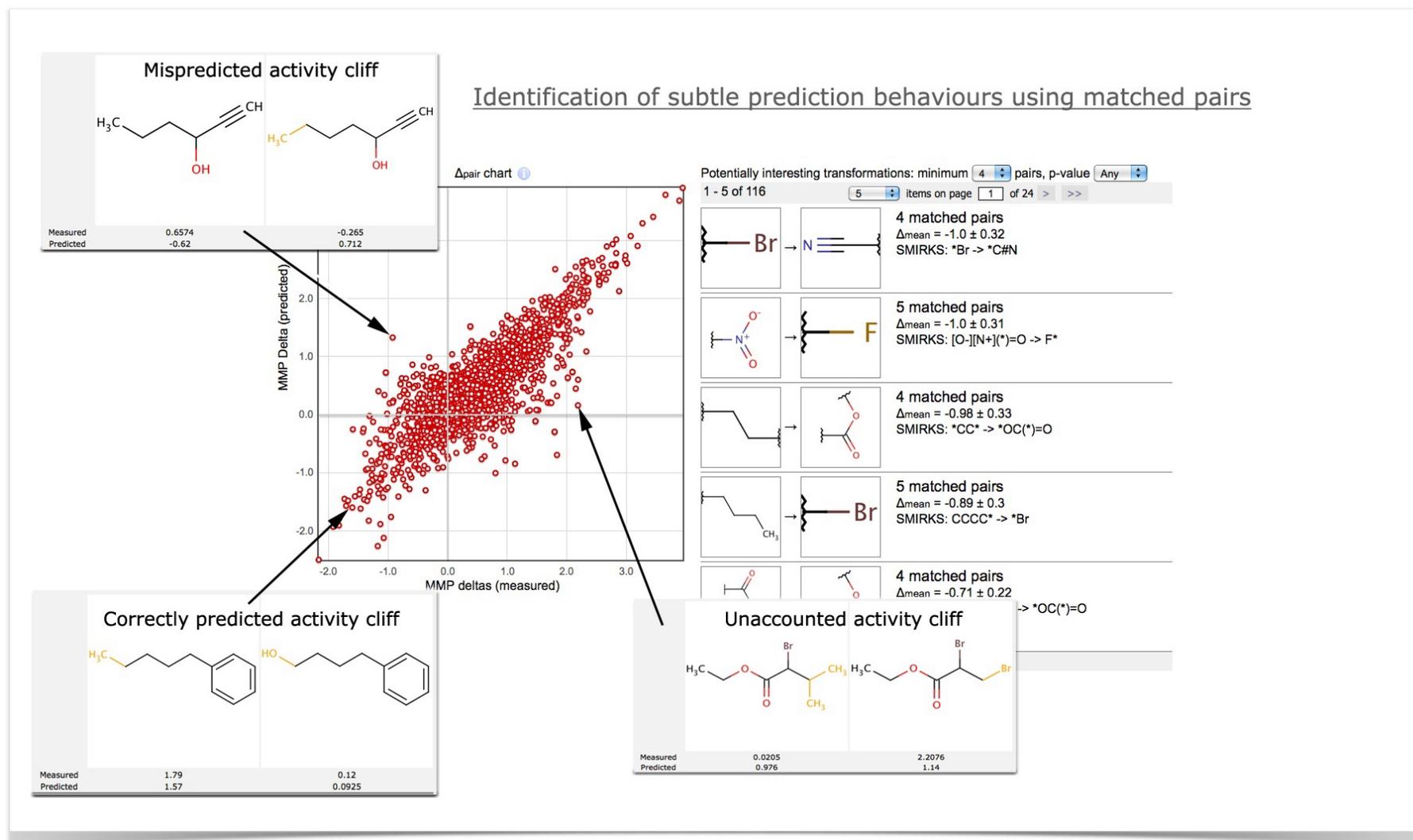


Data analysis using Matched Molecular Pairs



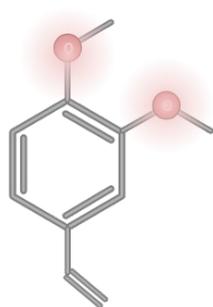
Identify molecular transformations that lead to significant change of activity (AMES test data are shown)

Identification of predicted activity cliffs

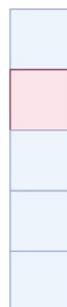


XAI methods

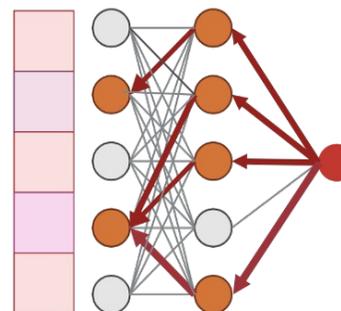
Perturbation-based XAI



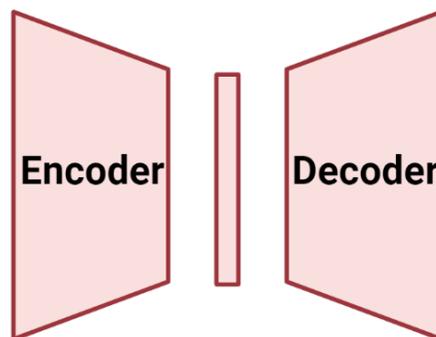
SHAP



Gradient-based XAI



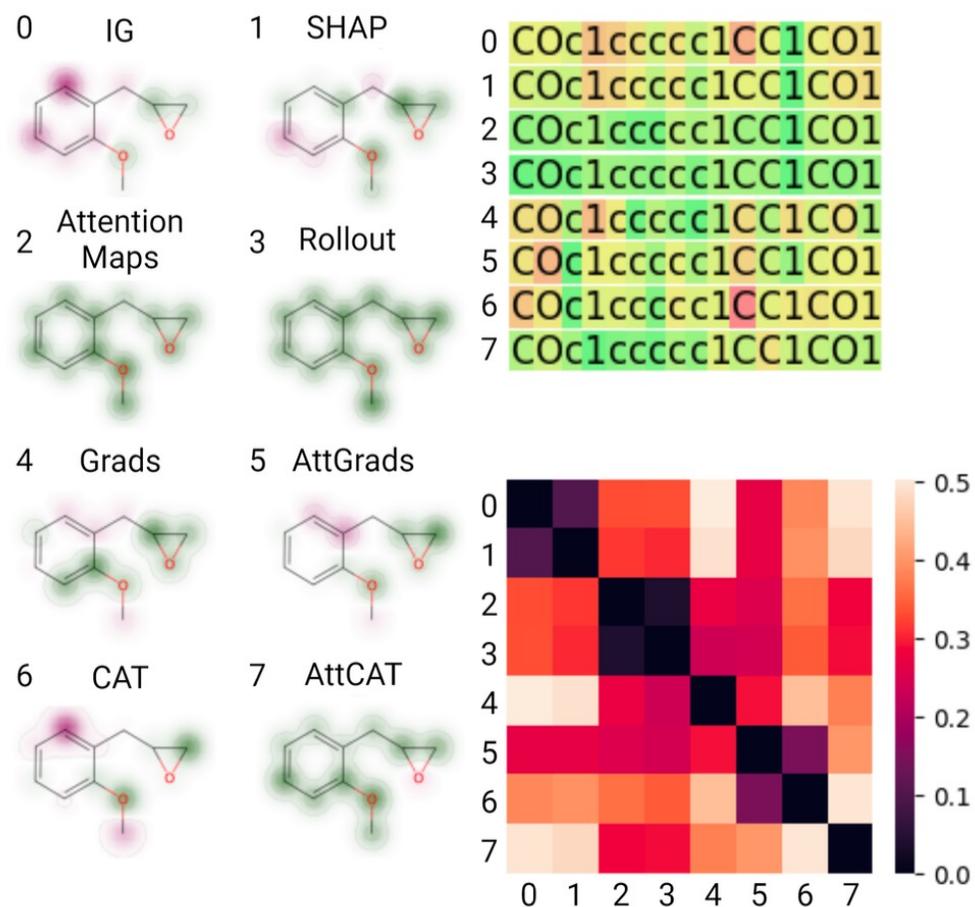
Integrated Gradients (IG)



Attention Maps, Rollout, Grads, AttGrads, CAT and AttCAT

Hartog P et al, Using test-time augmentation to investigate explainable AI: inconsistencies between method, model and human intuition *J. Cheminformatics*, **2024**, 16 (1), 39.

Importance of features across XAI methods



Hartog P et al, Using test-time augmentation to investigate explainable AI: inconsistencies between method, model and human intuition *J. Cheminformatics*, **2024**, 16 (1), 39.

AiChemist – Explainable AI for molecules

Ai Chemist
91 posts

AiChemist
@AiChemist_DN

AiChemist project is funded by the EU Horizon Europe under the Marie Skłodowska-Curie grant agreement No 101120466 . #machinelearning #drugdesign #AI #DN

<https://aichemist.eu/lectures> aichemist.eu Joined August 2023

40 Following 350 Followers

Edit profile

Twitter: aichemist_dn or Bluesky: aichemist

Take home message

New methods based on representation learning successfully compete with traditional ones

Consensus modelling is a best approach to develop models with the highest prediction accuracy

Applicability domain of models and accuracy of predictions are crucial for their use and interpretation predictions

Model interpretation is essential for their acceptance by the end users (sometimes legally required)

Different XAI explanations do not always overlap; statistical evaluation is strongly required

Acknowledgements

Katya Ahmad
Thalita Chirino
Mark Embrechts

Andi Kopp
Peter Hartog
Fabian Krüger
Paula Torren-Peraire
Varvara Voinarovska
Nesma Mousa

Guillaume Godin (ex Firmenich)
Ruud van Deursen (DSM-Firmenich)

Michael Sattler (HMGU)

