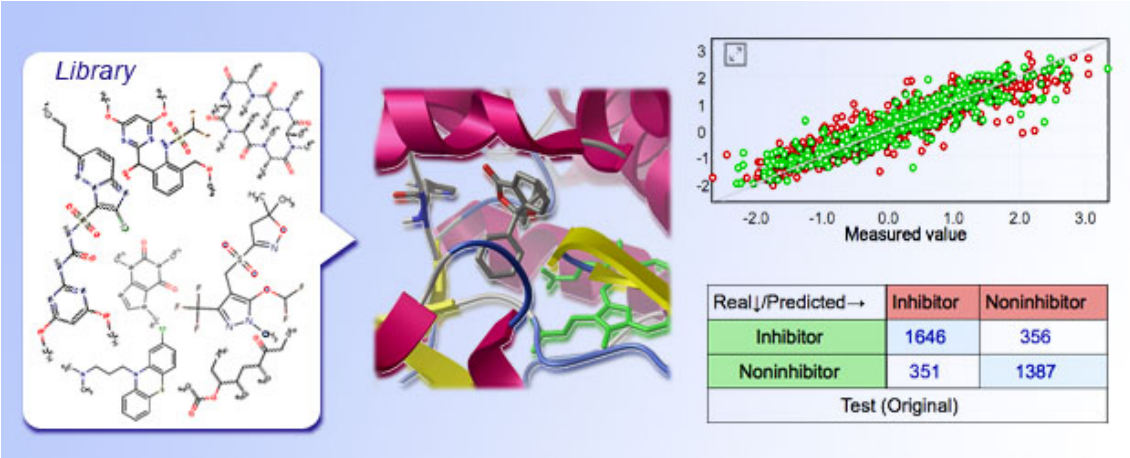


On-line Chemical Modeling Environment (OCHEM): an advanced platform for (multi)task properties analysis

Dr. Igor V. Tetko

BIGCHEM GmbH & Institute of Structural Biology, HelmholtzZentrum München

Wednesday, December 2, 2020

Institute of Structural Biology

HelmholtzZentrum münchen
German Research Center for Environmental Health

big chem

BigChem

http://bigchem.eu

big data in chemistry + informatics = chemoinformatics

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

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

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

big chem

BIGCHEM project publications http://bigchem.eu

BIGCHEM publications

FOLLOW

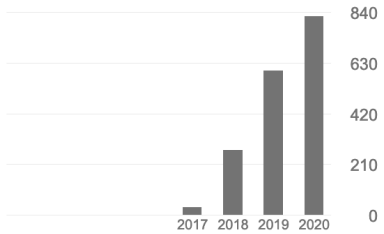
Horizon2020 Marie Skłodowska-Curie Innovative Training Network European Industrial Doctorate
Verified email at bigchem.eu

big data chemoinformatics cheminformatics

TITLE	CITED BY	YEAR
The rise of deep learning in drug discovery H Chen, O Engkvist, Y Wang, M Olivecrona, T Blaschke Drug discovery today 23 (6), 1241-1250	463	2018
Molecular de-novo design through deep reinforcement learning M Olivecrona, T Blaschke, O Engkvist, H Chen Journal of cheminformatics 9 (1), 48	293	2017
Automating drug discovery G Schneider Nature Reviews Drug Discovery 17 (2), 97	212	2018
Application of Generative Autoencoder in *De Novo* Molecular Design T Blaschke, M Olivecrona, O Engkvist, J Bajorath, H Chen Molecular informatics 37 (1-2), 1700123	153	2018
BIGCHEM: challenges and opportunities for big data analysis in chemistry IV Tetko, O Engkvist, U Koch, JL Reymond, H Chen Molecular informatics 35 (11-12), 615-621	67	2016
On the integration of in silico drug design methods for drug repurposing E March-Vila, L Pinzi, N Sturm, A Tinivella, O Engkvist, H Chen, G Rastelli Frontiers in pharmacology 8, 298	64	2017
Exploring the GDB-13 chemical space using deep generative models J Arús-Pous, T Blaschke, S Ulander, JL Reymond, H Chen, O Engkvist Journal of cheminformatics 11 (1), 1-14	45	2019
Randomized SMILES strings improve the quality of molecular generative models J Arús-Pous, SV Johansson, O Prykhodko, EJ Bjerrum, C Tyrchan, ... Journal of cheminformatics 11 (1), 1-13	40	2019

Cited by

	All	Since 2015
Citations	1754	1751
h-index	16	16
i10-index	24	24



Co-authors VIEW ALL

- Ola Engkvist, AstraZeneca R&D Gothenburg O…
- Hongming Chen, Astrazeneca R&D Mölndal
- Jürgen Bajorath, Professor of Life Science Inform…
- Thomas Blaschke, Phd student, AstraZeneca/Unive…
- Jean-Louis Reymond, University of Bern
- Igor V. Tetko, Group Leader at Helmholtz Zentr…

Up to now ~ 70 articles, including four highly cited (<1%) and one hot (<0.1%) article according to the Web of Science

About

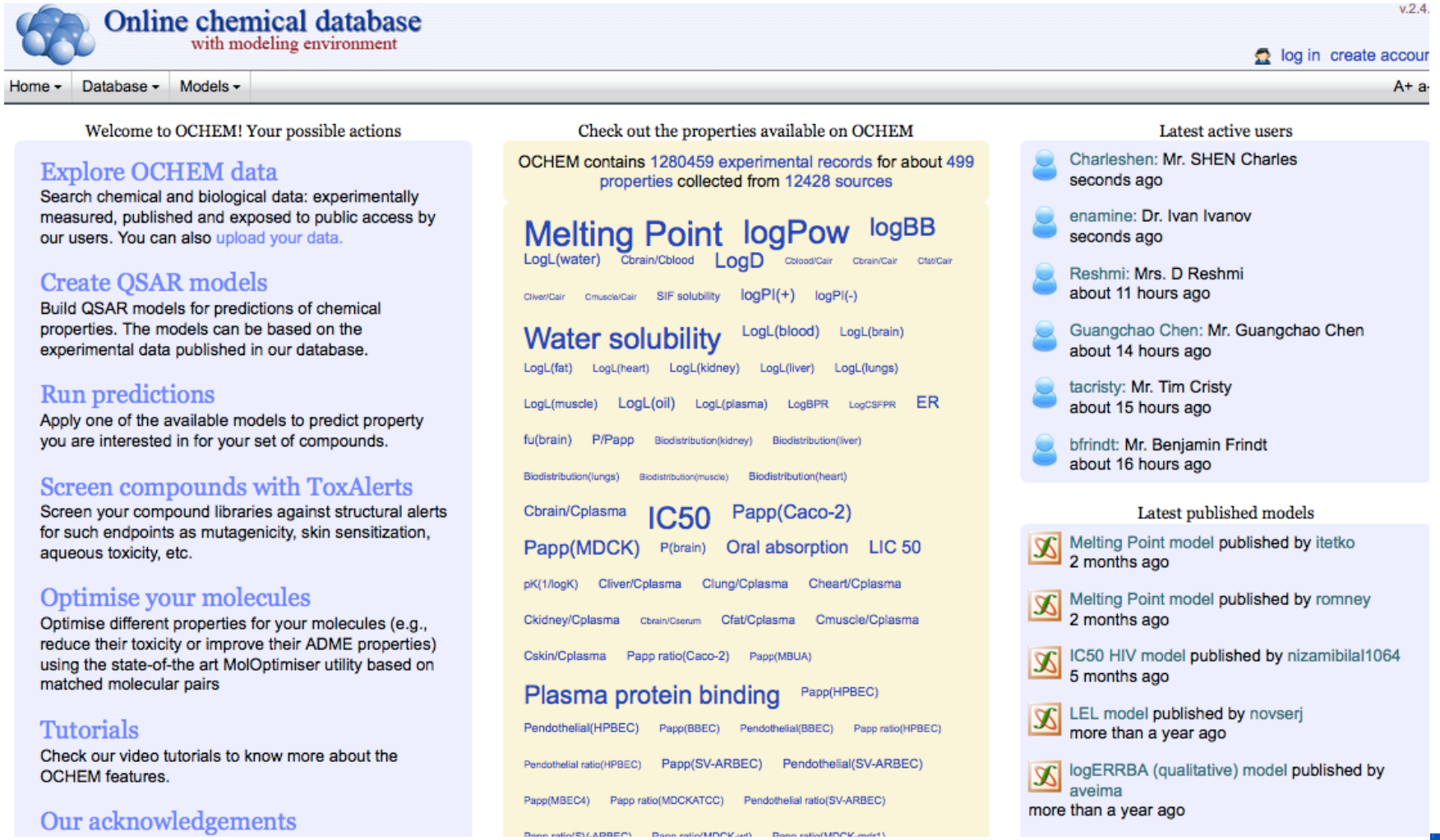
The dramatic increase in using of Artificial Intelligence (AI) and machine learning methods in different fields of science becomes an essential asset in the development of the chemical industry, including pharmaceutical, agro biotech, and other chemical companies. However, the application of AI in these fields is not straightforward and requires excellent knowledge of chemistry. Thus, there is a strong need to train and prepare a new generation of scientists who have skills both in machine learning and in chemistry and can advance medicinal chemistry, which is the prime goal of the AIDD proposal. Research WPs include sixteen topics selected to cover the key innovative directions in machine learning in chemistry. Fellows employed will be supervised by academics who have excellent complementary expertise and contributed some of the fundamental AI algorithms which are used billions of times per day in the world, and leading EU Pharma companies who are in charge of new medicine and public health. All developed methods can be used individually but will also contribute to an integrated "One Chemistry" model that can predict outcomes ranging from different properties to molecule generation and synthesis. Training on various modalities allows the model to understand how to intertwine chemistry and biology to develop a new drug making its design robust and explainable. All partners agreed to make their software open source. It will boost the field and will provide the broadest possible dissemination of the results both to the academy and industry, including SMEs. The network will offer comprehensive, structured training through a well-elaborated Curriculum, online courses, and six Schools. The IP policy and commercial exploitation of the project results have the highest priority supported by intellectual property asset management organizations. Comprehensive public engagement activities will complement the dissemination of results to the scientific community.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 956832.

The project will start on January 1st, 2021

http://ai-dd.eu fellowship applications will be announced soon

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



Some OCHEM statistics

- **Physico-chemical properties:** logP, water solubility, melting point, pyrolysis (decomposition), solubility in DMSO
- **Biological activity:** estrogen receptors binding; endocrine disruptors; anti-HIV activity; AMES mutagenicity
- **Environmental endpoints:** ready biodegradability; fish toxicity; toxicity against T. Pyriformis, daphnia, etc.

- More than >100 (>500) models were published
- >7000 registered users
- >36 millions tasks were executed since launch
- >3M data points for >500 properties from > 10,000 articles
- Several groups develop and regularly publish new models
- Top-performing models in challenges (NIH, EPA ToxCast)

big chem

275k Melting Point Datasets (Big Data)

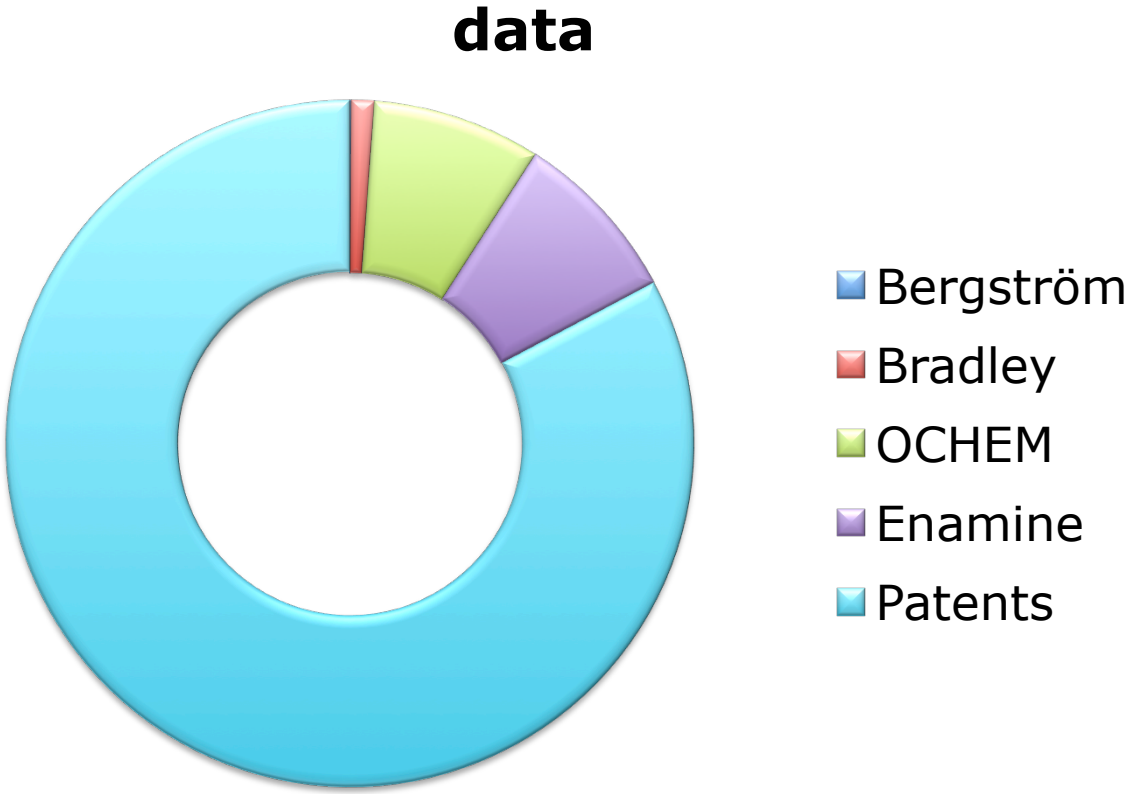
Bergström 277

Bradley 2886

OCHEM 22404

Enamine 21883

PATENTS 228079



COMBINED: OCHEM + Enamine + Bradley + Bergström

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

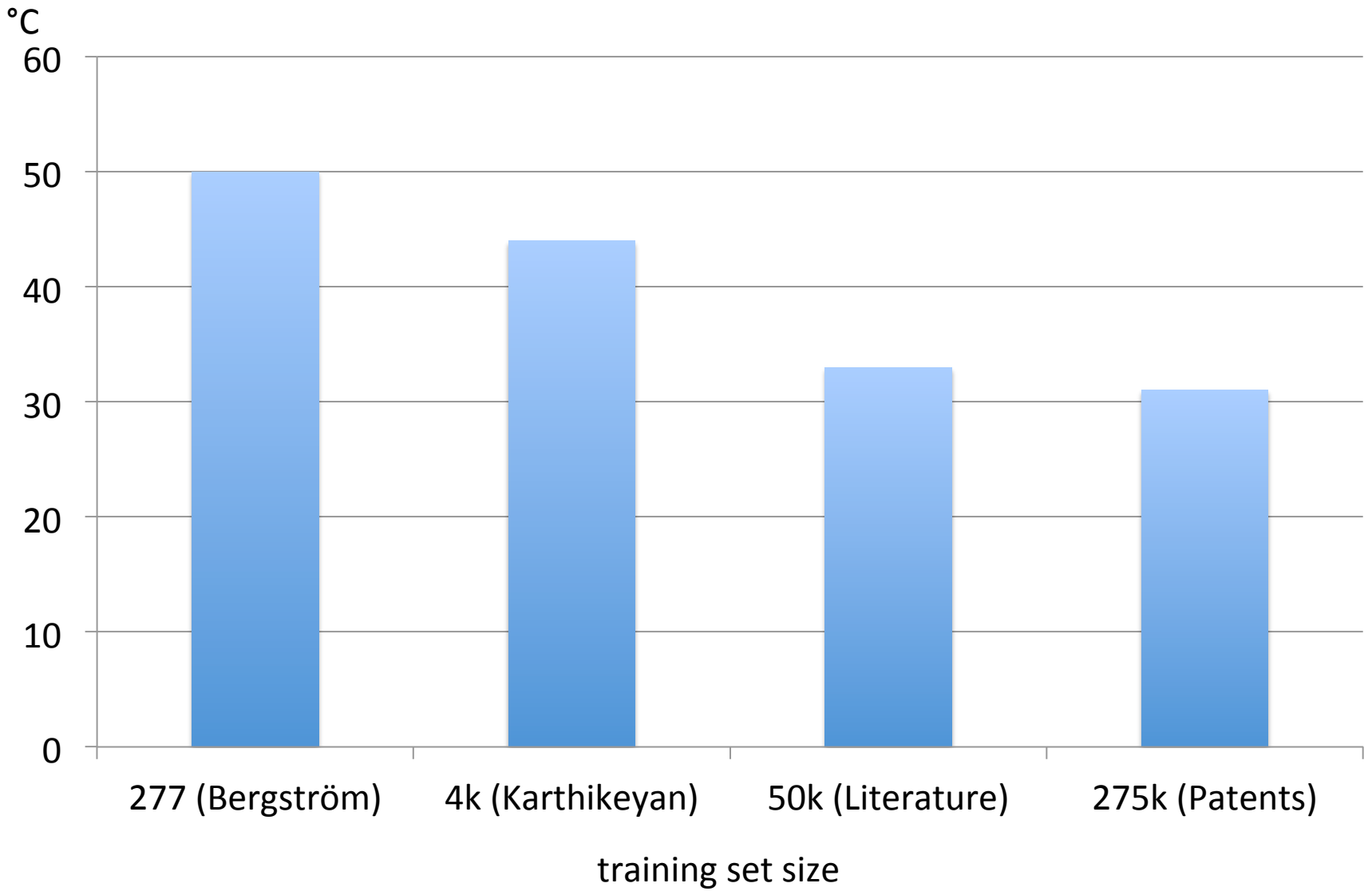
big chem

Modeling of MP data

Package name	Type of descriptors	Number of descriptors	Matrix size, billions	Non zero values, millions	Sparseness
Functional Groups	integer	595	0.18	3.1	33
QNPR	integer	1502	0.45	6.3	49
MolPrint	binary	688634	205	8.1	7200
Estate count	float	631	0.19	10	14
Inductive	float	54	0.02	11	1
ECFP4	binary	1024	0.31	12	25
Isida	integer	5886	1.75	18	37
ChemAxon	float	498	0.15	23	1.5
GSFrag	integer	1138	0.34	24	5.7
CDK	float	239	0.07	27	2
Adriana	float	200	0.06	32	1.3
Mera, Mersy	float	571	0.17	61	1.1
Dragon	float	1647	0.49	183	1.5

big chem

Prediction errors for a set of drugs using models developed with different training sets

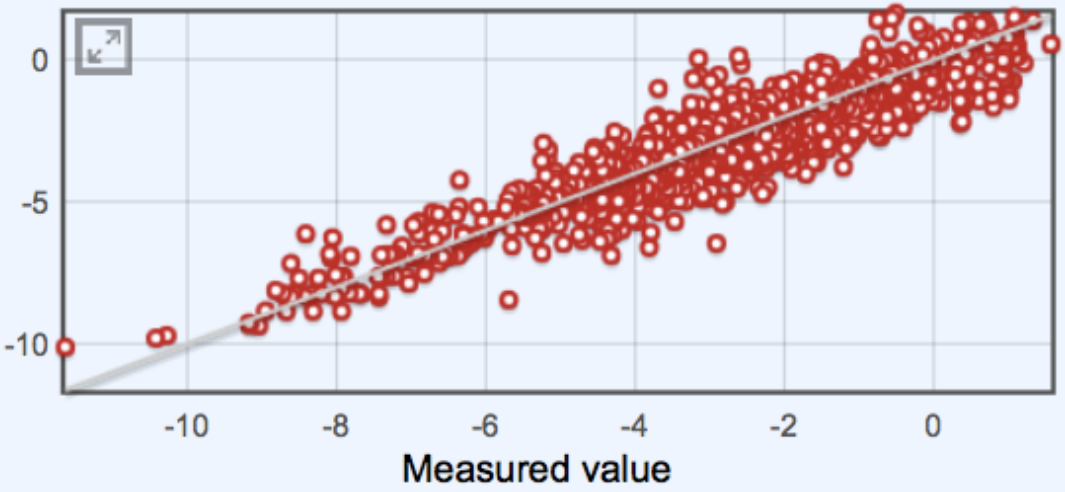


Prediction of Huuskonen set using ALOGPS logP and MP based on 230k measurements

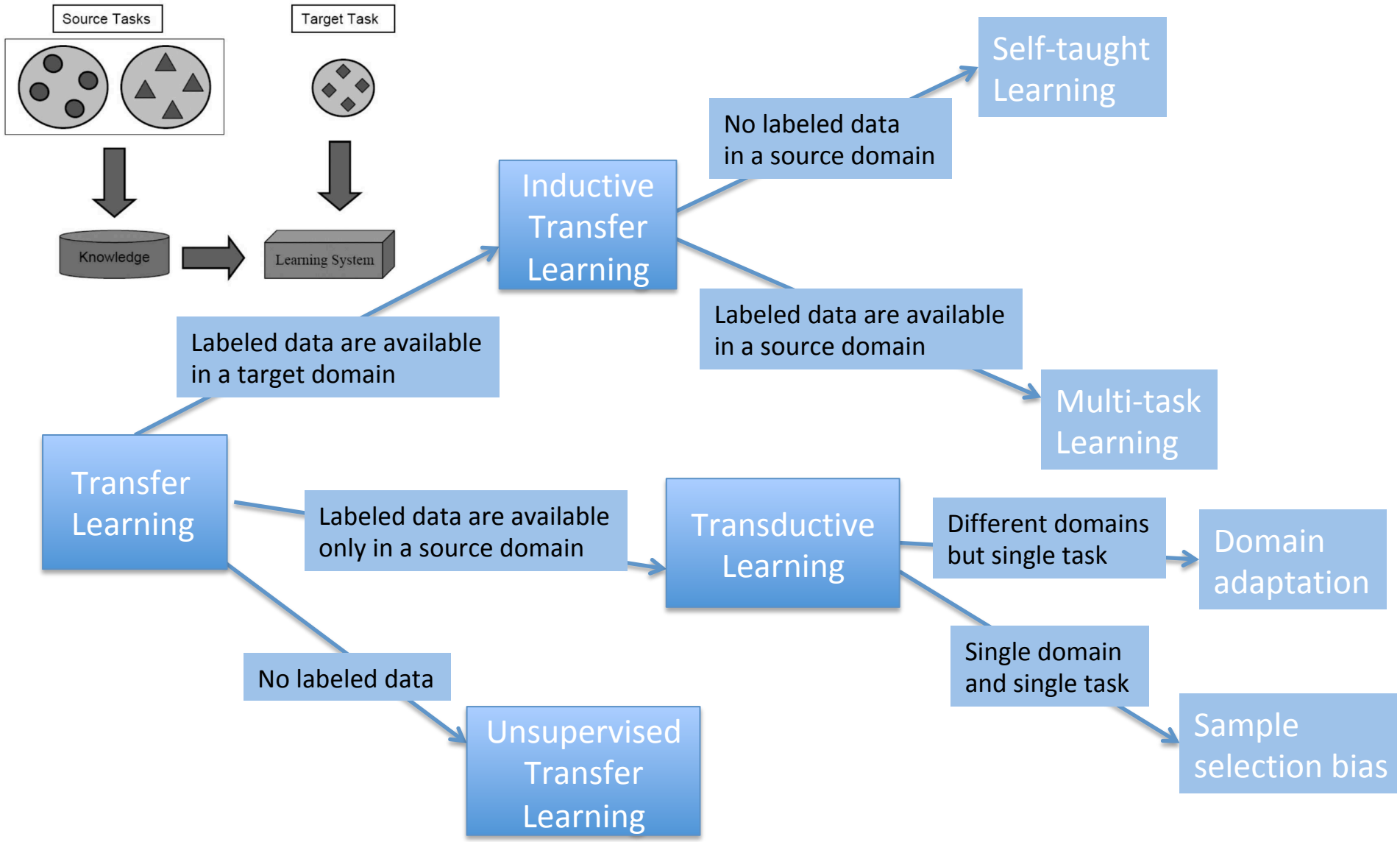
$$logS = 0.5 - 0.01(MP-25) - log\ Kow$$

Predicted property: **Aqueous Solubility** modeled in log(mol/L)
Training method: MLRA

Data Set	#	R_2	q_2	RMSE	MAE
Training set: logS set	1311 records	0.842 ± 0.009	0.83 ± 0.01	0.84 ± 0.02	0.64 ± 0.02

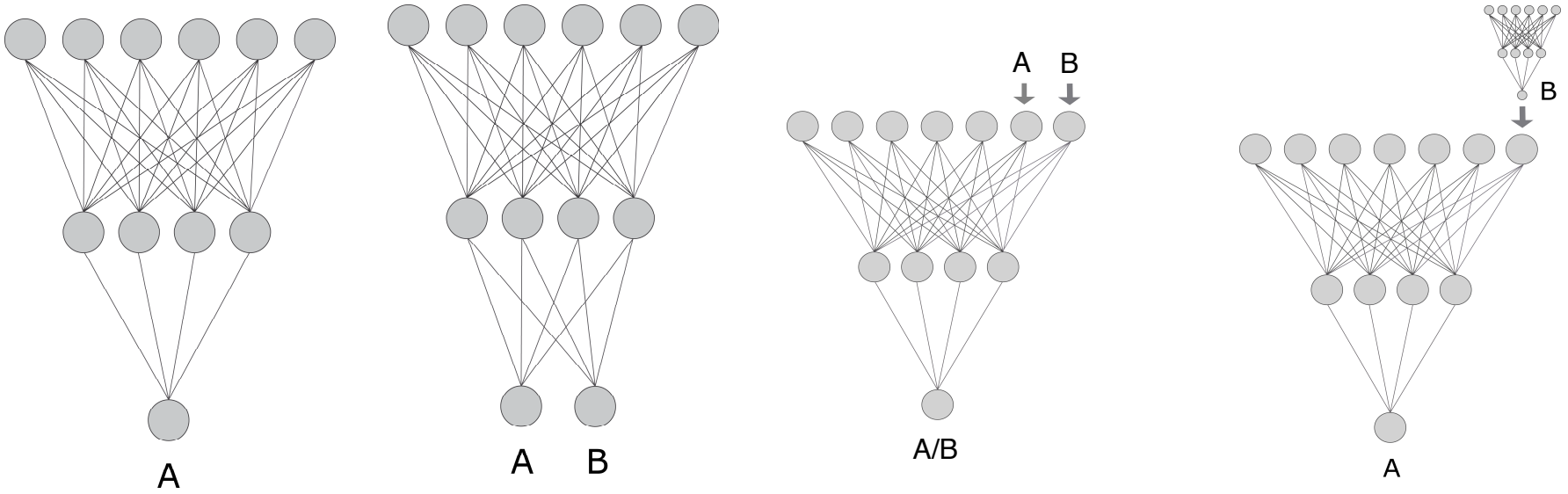


"model in model"



Adapted from: Pan, S.J.; Yang, Q. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering* **2010**, *22*, 1345-1359.

Multi-task learning



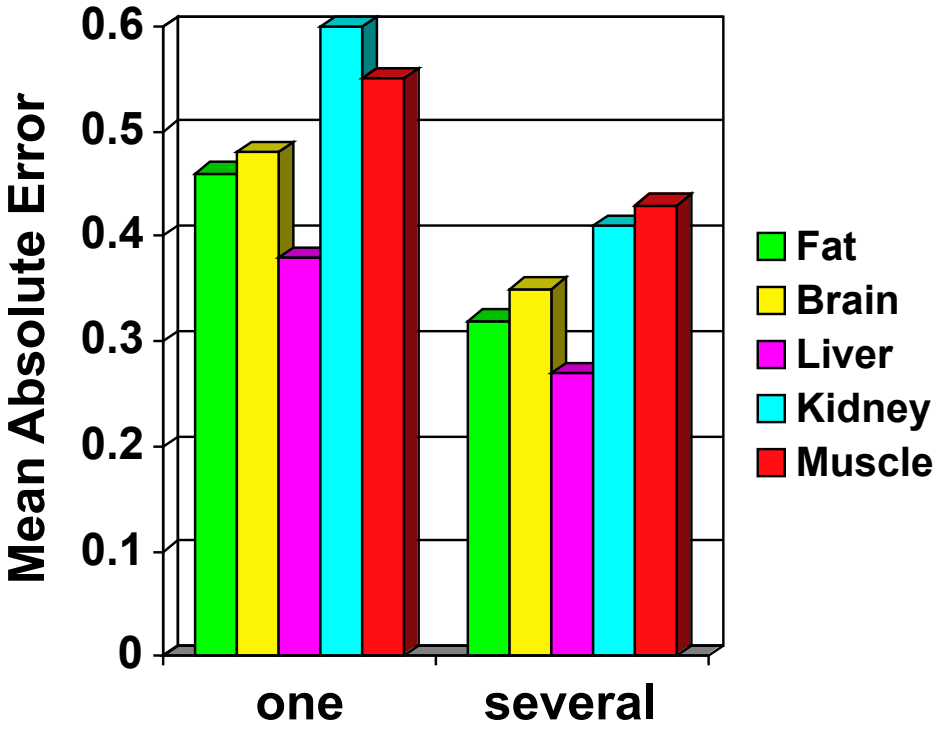
Multi-task learning

Problem:

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

Results:

simultaneous prediction of several properties increased the accuracy of models



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

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

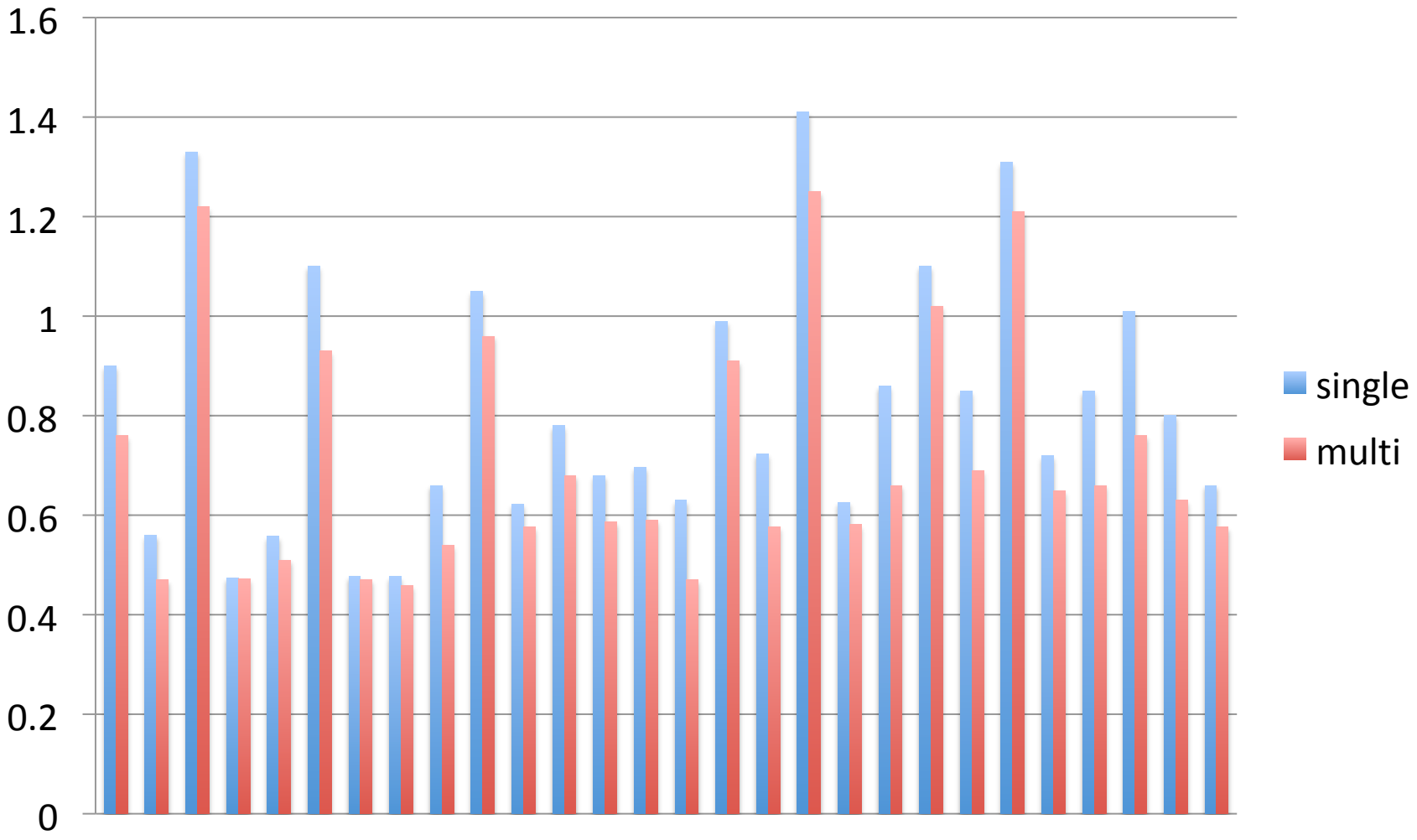
Different species
- Rat
- Mouse
- Rabbit
- …
- Human

~ 129k records
~ 87k compounds
29 properties

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

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

RMSE for different toxicities using CDK descriptors and DNN



Sosnin, S. et al. A comparative study of prediction of multi-target toxicity for a broad chemical space. *J Chem Inf Model.* 2019, **59**, 1062-1072.

Comparison of different models to predict toxicity (RMSE)

single multi single

s RMSE - Root Mean Square Error for Training set Validation: Cross-Validation (63 models)

	DNN	DNN(2)	XGBOOST
CDK2 (constitutional, topological, geometrical, electronic, ...	0.9 0.56 1.33 0.474 0.56 1.1 0.478 0.477 0.66 1.05 0.623 0.78 0.68 0.7 0.63 0.99 0.724 1.41 0.63 0.86 1.1 0.85 1.31 0.72 0.85 1.01 0.8 0.66 1.27 (0.834)	0.76 0.47 1.22 0.472 0.51 0.93 0.471 0.459 0.54 0.96 0.576 0.68 0.59 0.591 0.47 0.91 0.577 1.25 0.581 0.66 1.02 0.69 1.21 0.65 0.66 0.76 0.63 0.58 1.14 (0.725)	0.8 0.47 1.29 0.454 0.5 1.02 (0.439 0.56 1.04 0.584 0.75 0.65 0.59 0.95 0.66 1.33 0.5 0.75 1.08 0.764 1.3 0.67 0.81 0.76 0.63 1.2 (0.779)
Dragon6 (blocks: 1-29)	0.89 0.58 1.3 0.458 0.56 1.06 0.481 0.472 0.6 1.06 0.63 0.74 0.66 0.686 0.63 0.97 0.69 1.32 0.622 0.82 1.09 0.83 1.33 0.76 0.83 0.98 0.8 0.7 1.24 (0.82)	0.78 0.44 1.31 0.445 0.474 0.96 0.461 0.446 0.52 1 0.555 0.68 0.55 0.581 0.47 0.95 0.57 1.31 0.574 0.65 1.08 0.68 1.2 0.68 0.67 0.74 0.64 0.59 1.22 (0.732)	0.8 0.49 1.3 0.454 0.523 1.01 0.439 0.59 1.02 0.588 0.73 (0.66 0.602 0.94 0.67 1.33 0. 0.76 1.09 0.77 1.38 0.68 0.82 0.74 0.63 1.24 (0.786)
ALogPS, OEstate	0.91 0.61 1.32 0.461 0.54 1.1 0.478 0.469 0.6 1.1 0.617 0.75 0.7 0.652 0.64 1 0.69 1.36 0.617 0.84 1.11 0.87 1.43 0.76 0.85 0.95 0.8 0.71 1.2 (0.832)	0.79 0.44 1.23 0.447 0.49 0.94 0.467 0.444 0.53 0.99 0.554 0.66 0.55 0.59 0.49 0.9 0.58 1.21 0.571 0.65 1.05 0.69 1.22 0.65 0.7 0.74 0.64 0.6 1.17 (0.724)	0.84 0.5 1.42 0.456 0.519 1 0 0.44 0.56 1.03 0.58 0.73 0.5 0.65 0.61 0.95 0.64 1.34 0.59 1.11 0.79 1.33 0.69 0.8 0.81 0.63 1.21 (0.786)
Fragmentor (Length 2 - 4)	0.96 0.61 1.43 0.463 0.542 1.14 0.491 0.484 0.62 1.1 0.647 0.81 0.71 0.71 0.64 1.04 0.74 1.38 0.643 0.79 1.14 0.86 1.33 0.82 0.86 0.94 0.84 0.66 1.22 (0.849)	0.73 0.45 1.25 0.44 0.48 0.95 0.465 0.448 0.502 0.99 0.554 0.65 0.55 0.56 0.46 0.92 0.575 1.28 0.564 0.63 1.07 0.69 1.24 0.7 0.66 0.73 0.63 0.62 1.2 (0.724)	0.78 0.45 1.38 0.447 0.52 1 0.476 0.436 0.58 1.09 0.592 0.61 0.67 0.59 0.94 0.67 1.3 (0.77 1.14 0.79 1.43 0.69 0.83 0.77 0.64 1.29 (0.797)

Sosnin, S. et al. A comparative study of prediction of multi-target toxicity for a broad chemical space. *J Chem Inf Model.* 2019, **59**, 1062-1072.

big chem

Comparison of MTL and STL

Multiple models overview

Predicted property: Cblood/Cair(Human)
Training set: tissue/air set

Metrics RMSE - Root Mean Square Error for Training set Validation: Cross-Validation (16 models)

	ASNN	MTL	DNN	ASNN(2)	STL	DNN(2)
CDK2 (constitutional, topological, geometrical, electronic, ...	0.45 0.28 0.21 0.29 0.39 0.33 0.28 0.32 0.4 0.33 0.4 (0.335)		0.54 0.33 0.38 0.35 0.4 0.45 0.321 0.43 0.44 0.49 0.52 (0.423)	0.41 0.41 0.45 0.42 0.44 0.56 0.279 0.5 0.39 0.37 0.44 (0.424)		0.549 0.45 0.54 0.48 0.71 0.66 0.35 0.6 0.46 0.44 0.71 (0.541)
OEstate	0.44 0.35 0.31 0.33 0.4 0.44 0.32 0.33 0.33 0.31 0.36 (0.356)		0.42 0.29 0.31 0.32 0.38 0.41 0.31 0.33 0.41 0.37 0.4 (0.359)	0.41 0.47 0.44 0.51 0.66 0.6 0.37 0.57 0.5 0.39 0.48 (0.491)		0.44 0.35 0.46 0.41 0.4 0.46 0.38 0.48 0.47 0.41 0.57 (0.439)

	DAG	GRAPH_CONV	TEXTCNN	WEAVE
MTL	0.75 0.55 0.6 0.35 0.94 0.67 0.44 0.64 0.58 0.57 0.92 (0.637)	0.93 0.64 0.8 0.58 1 1 0.6 0.79 0.85 0.89 0.8 (0.807)	0.53 0.4 0.43 0.33 0.48 0.53 0.35 0.53 0.47 0.48 0.5 (0.457)	0.7 0.69 0.8 0.61 0.9 0.64 0.41 0.74 0.57 0.61 0.7 (0.67)
STL	0.63 0.52 0.9 0.47 1.1 1 0.38 0.8 0.62 0.62 1 (0.731)	0.8 0.61 0.9 0.7 0.9 0.78 0.65 0.8 0.86 0.92 0.9 (0.802)	0.58 0.54 0.57 0.51 0.7 0.63 0.39 0.66 0.51 0.62 0.48 (0.563)	0.62 0.52 0.7 0.59 0.8 1.1 0.48 0.71 0.72 0.72 0.8 (0.705)

big chem

Support of mixtures

Basket Records Tags

1 - 5 of 465 5 items on page 1 of 93 > >>

molecule profile

● Azeo = non azeotrope

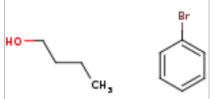
MIXTURES = ClC(Cl)Cl;0.5 C1CCCCC1;0.5

Horsley, L.
Table of Azeotropes and Nonazeotropes
N: AUTO_400
Anal. Chem. **1947**; 19 (8) 508 - 600

Chloroform ; CYCLOHEXANE
MoleculeID: *M96691339*

Public and freely downloadable record (awaiting approval)

RecordID: *R32620625*
11:08, 2 Jul 18
xenol



molecule profile

● Azeo = non azeotrope

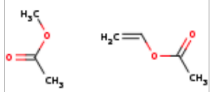
MIXTURES = CCCCO;0.5 BrC1=CC=CC=C1;0.5

Horsley, L.
Table of Azeotropes and Nonazeotropes
N: AUTO_399
Anal. Chem. **1947**; 19 (8) 508 - 600

butanol ; BROMOBENZENE
MoleculeID: *M96691338*

Public and freely downloadable record (awaiting approval)

RecordID: *R32620624*
11:08, 2 Jul 18
xenol



molecule profile

● Azeo = non azeotrope

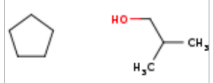
MIXTURES = COC(C)=O;0.5 CC(=O)OC=C;0.5

Horsley, L.
Table of Azeotropes and Nonazeotropes
N: AUTO_398
Anal. Chem. **1947**; 19 (8) 508 - 600

methylacetate ; VINYLACETATE
MoleculeID: *M96691337*

Public and freely downloadable record (awaiting approval)

RecordID: *R32620623*
11:08, 2 Jul 18
xenol



molecule profile

● Azeo = non azeotrope

MIXTURES = C1CCCC1;0.5 CC(C)CO;0.5

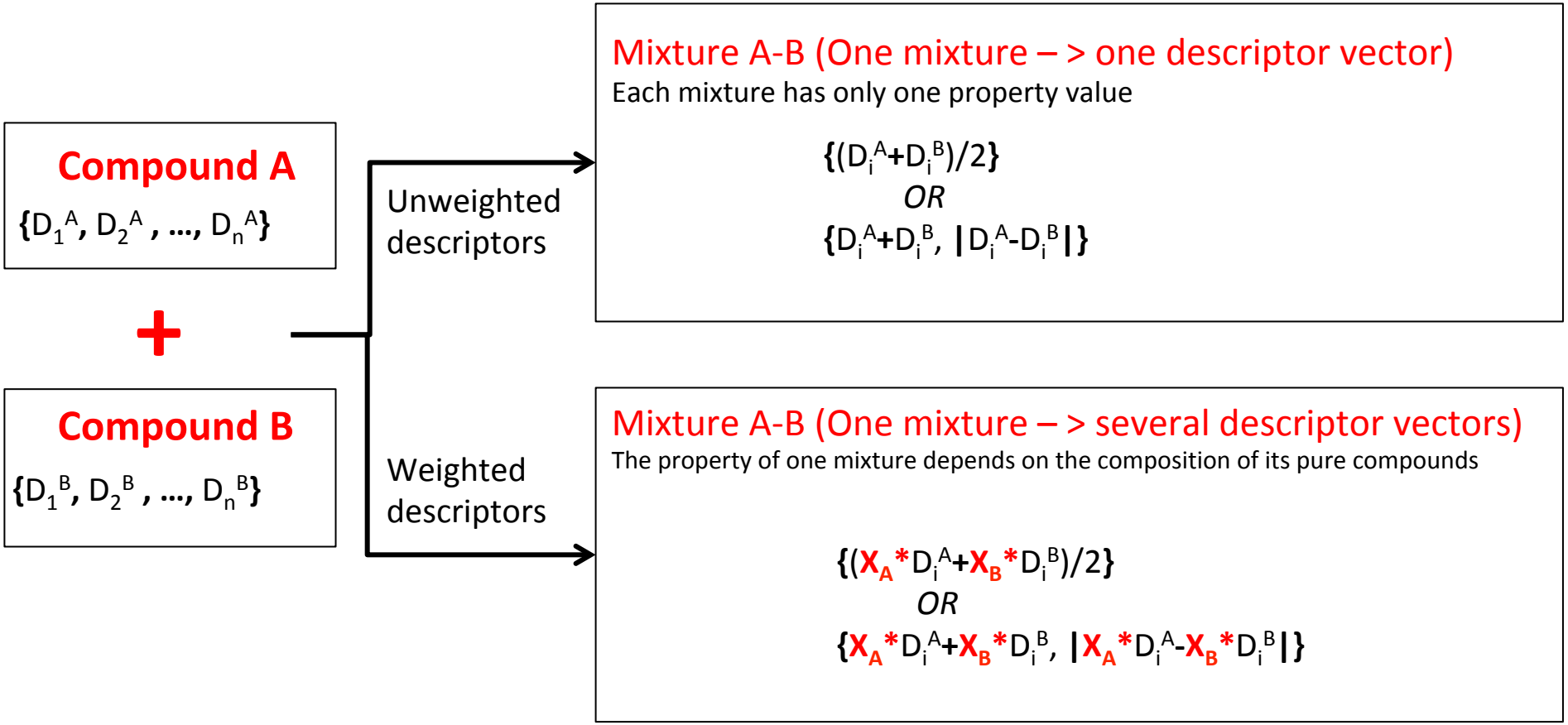
Horsley, L.
Table of Azeotropes and Nonazeotropes
N: AUTO_397
Anal. Chem. **1947**; 19 (8) 508 - 600

Isobutanol ; CYCLOPENTANE
MoleculeID: *M96691336*

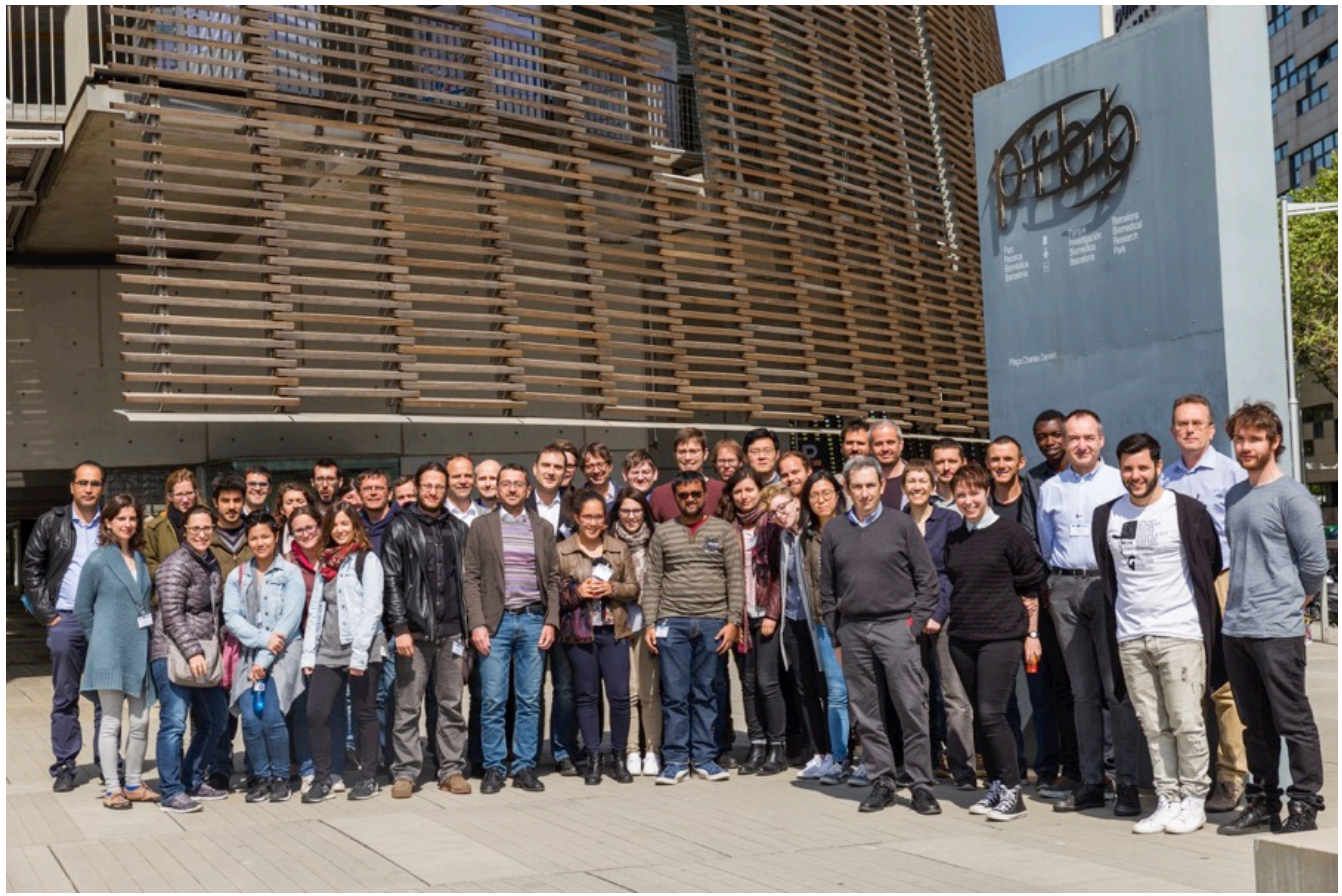
Public and freely downloadable record (awaiting approval)

RecordID: *R32620622*
11:08, 2 Jul 18
xenol

Mixtures' descriptors



Acknowledgements



Pavel Karpov
Dipan Ghosh
Michael Withnall
Zhonghua Xia
Barbara Gasset

Sergey Sosnin (Skoltech)
Maxim Fedorov (Skoltech)

Yura Sushko
Sergey Novotarskyi
Robert Körner

M. Sattler (HMGU)