Institute of Structural Biology

HelmholtzZentrum münchen
German Research Center for Environmental Health

Recent advances in machine learning for ADMETox prediction

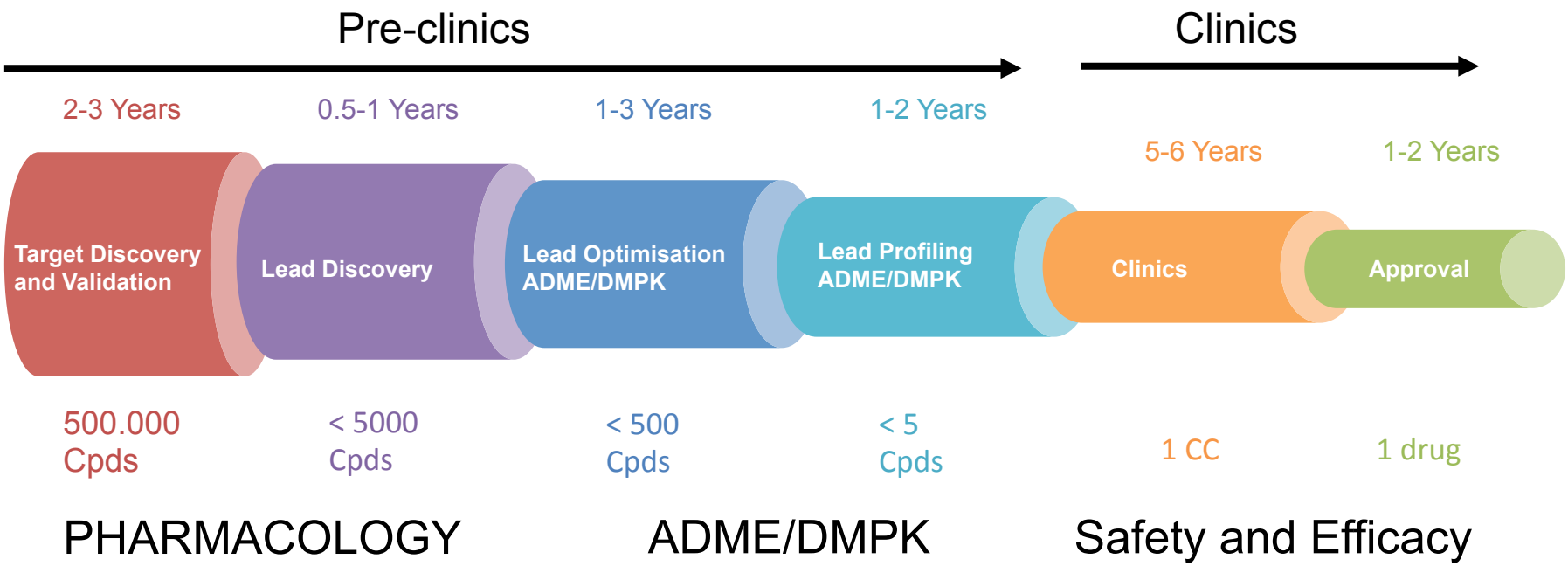
Dr. Igor V. Tetko

BIGCHEM GmbH, Institute of Structural Biology, Helmholtz Zentrum München, Germany and A. Krestov Institute of Solution Chemistry of the Russian Academy of Sciences, Ivanovo, Russia

Agenda

- Use of ADMETOx in industry
- New sources of data
- Inductive learning
- Interpretation of predictions
  - Design of interpretable descriptors
  - Use of (more) interpretable methods
- Descriptor-less methods
  - Explainable AI
- Accuracy of prediction
- Conclusion and perspectives

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, Sandoz

ADMETox filters in Bayer

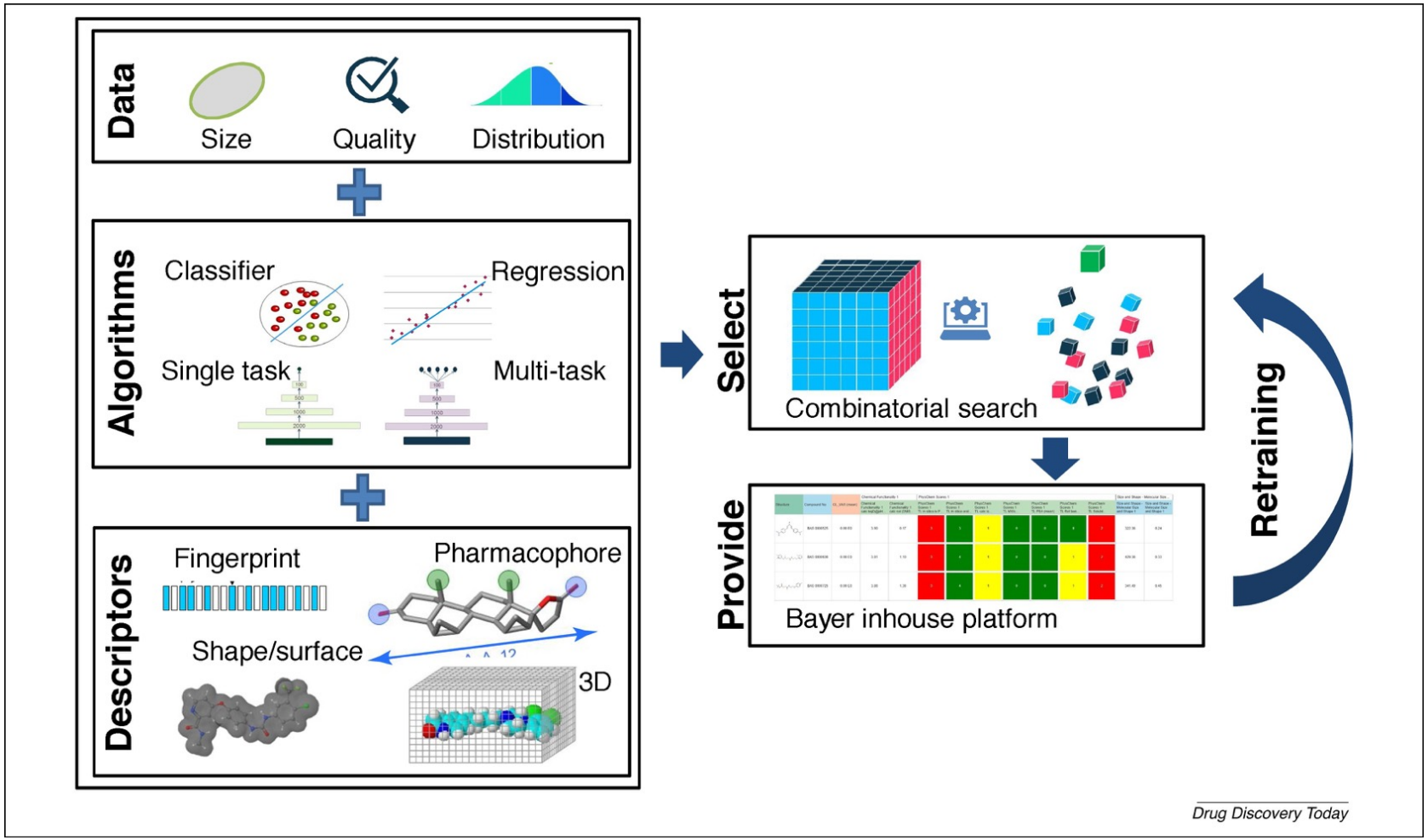
Insufficient quality | First approach | Medium model | Good model | Robust model

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

Drug Discovery Today

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

Bayer workflow for model life cycle



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

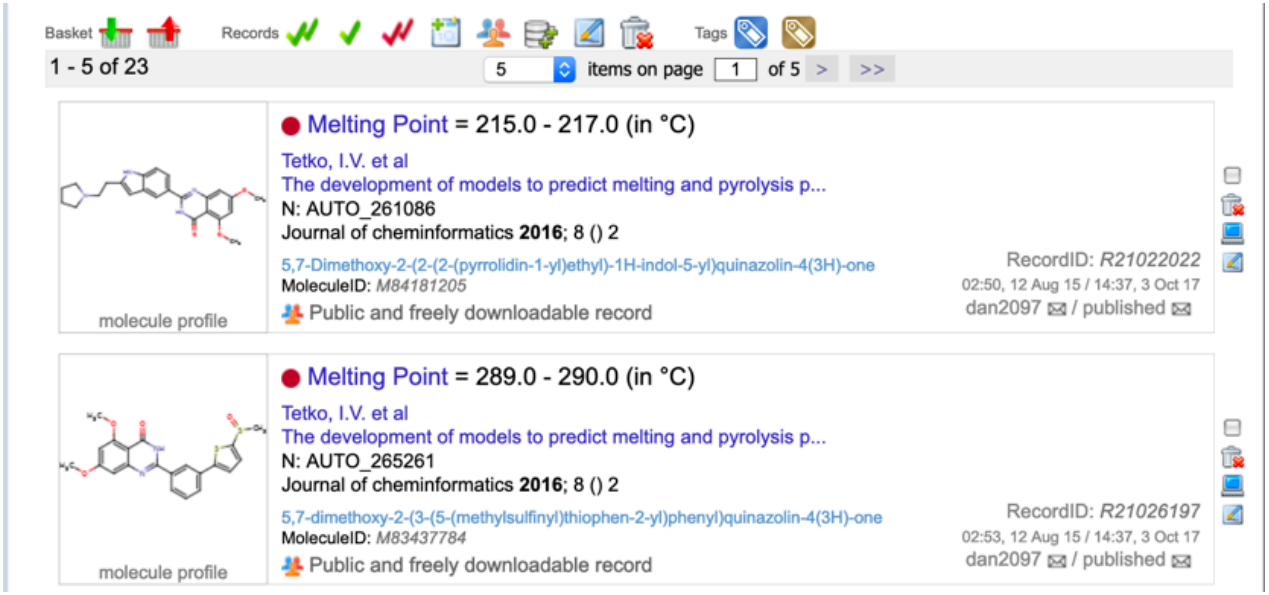
Challenges - Extraction of information from patents

[0835] To a solution of 2-amino-4,6-dimethoxybenzamide (0.266 g, 1.36 mmol) and 3-(5-(methylsulfinyl)thiophen-2-yl)benzaldehyde (0.34 g, 1.36 mmol) in N,N-dimethylacetamide (17 mL) was added NaHSO3 (0.36 g, 2.03 mmol) and p-toluenesulfonic acid monohydrate (0.052 g, 0.271 mmol) at rt. The reaction mixture was heated at 120° C. for 12.5 h. After that time the reaction was cooled to rt, concentrated under reduced pressure and diluted with water (20 mL). The precipitated solids were collected by filtration, washed with water and dried. The product was purified by flash column chromatography (silica gel, 95:5 chloroform/methanol) to give 5,7-dimethoxy-2-(3-(5-(methylsulfinyl)thiophen-2-yl)phenyl)quinazolin-4(3H)-one (0.060 g, 10%) as a light yellow solid: mp 289-290° C.; 1H NMR (400 MHz, DMSO-d6) δ 12.19 (br s, 1H), 8.48 (s, 1H), 8.18 (d, J=7.81 Hz, 1H), 7.90 (d, J=8.20 Hz, 1H), 7.72 (d, J=3.90 Hz, 1H), 7.55-7.64 (m, 2H), 6.77 (d, J=2.34 Hz, 1H), 6.54 (d, J=1.95 Hz, 1H), 3.88 (s, 3H), 3.84 (s, 3H), 2.96 (s, 3H); ESI MS m/z 427 [M+H]+.

http://www.google.com/patents/US20140140956

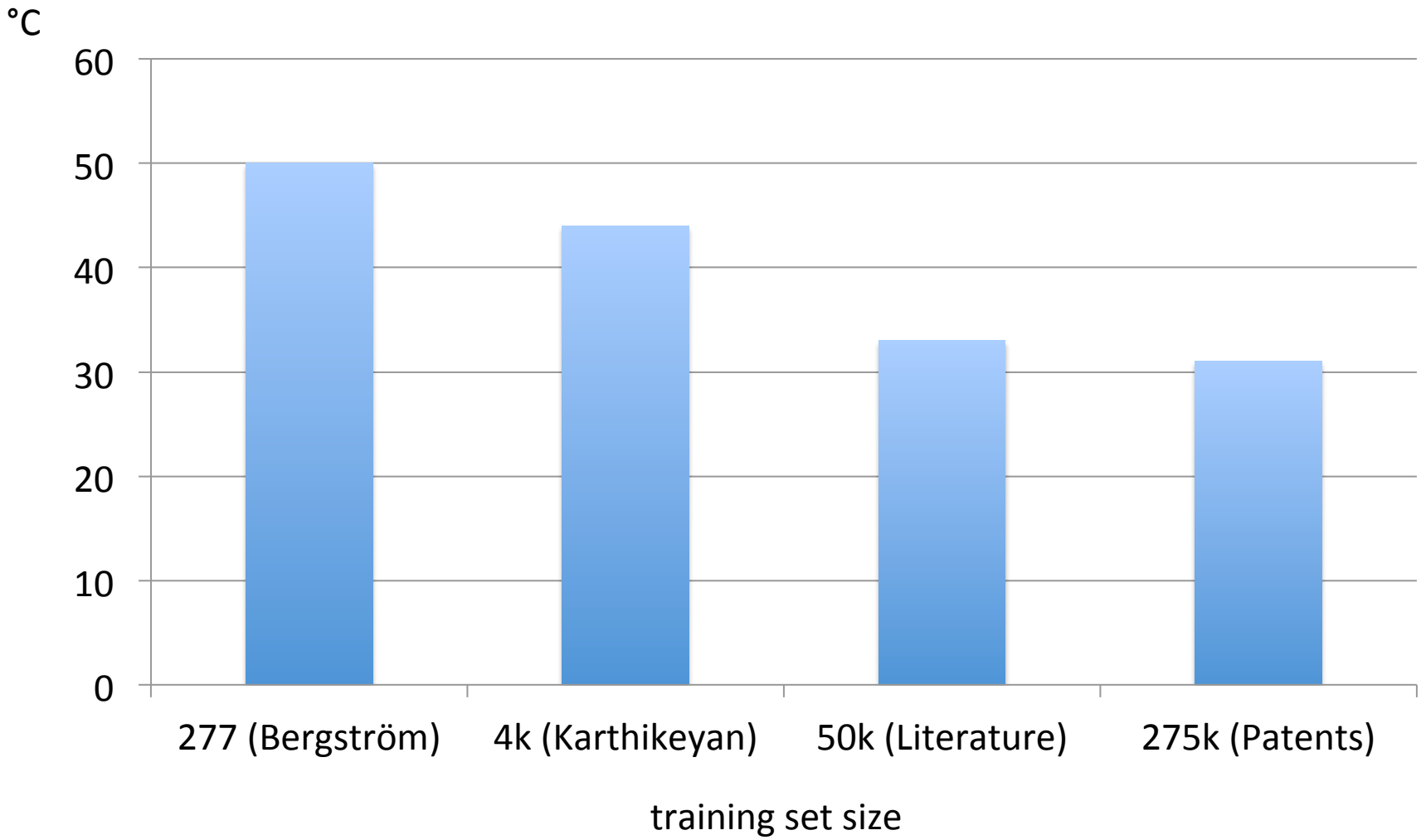
Challenges - Extraction of information from patents

[0835] To a solution of 2-amino-4,6-dimethoxybenzamide (0.266 g, 1.36 mmol) and 3-(5-(methylsulfinyl)thiophen-2-yl)benzaldehyde (0.34 g, 1.36 mmol) in N,N-dimethylacetamide (17 mL) was added NaHSO3 (0.36 g, 2.03 mmol) and p-toluenesulfonic acid monohydrate (0.052 g, 0.271 mmol) at rt. The reaction mixture was heated at 120° C. for 12.5 h. After that time the reaction was cooled to rt, concentrated under reduced pressure and diluted with water (20 mL). The precipitated solids were collected by filtration, washed with water and dried. The product was purified by flash column chromatography (silica gel, 95:5 chloroform/methanol) to give 5,7-dimethoxy-2-(3-(5-(methylsulfinyl)thiophen-2-yl)phenyl)quinazolin-4(3H)-one (0.060 g, 10%) as a light yellow solid: mp 289-290° C.; 1H NMR (400 MHz, DMSO-d6) δ 12.19 (br s, 1H), 8.48 (s, 1H), 8.18 (d, J=7.81 Hz, 1H), 7.90 (d, J=8.20 Hz, 1H), 7.72 (d, J=3.90 Hz, 1H), 7.55-7.64 (m, 2H), 6.77 (d, J=2.34 Hz, 1H), 6.54 (d, J=1.95 Hz, 1H), 3.88 (s, 3H), 3.84 (s, 3H), 2.96 (s, 3H); ESI MS m/z 427 [M+H]+.



http://ochem.eu

Prediction errors for a set of drugs (Bergström dataset) using models developed with different training sets

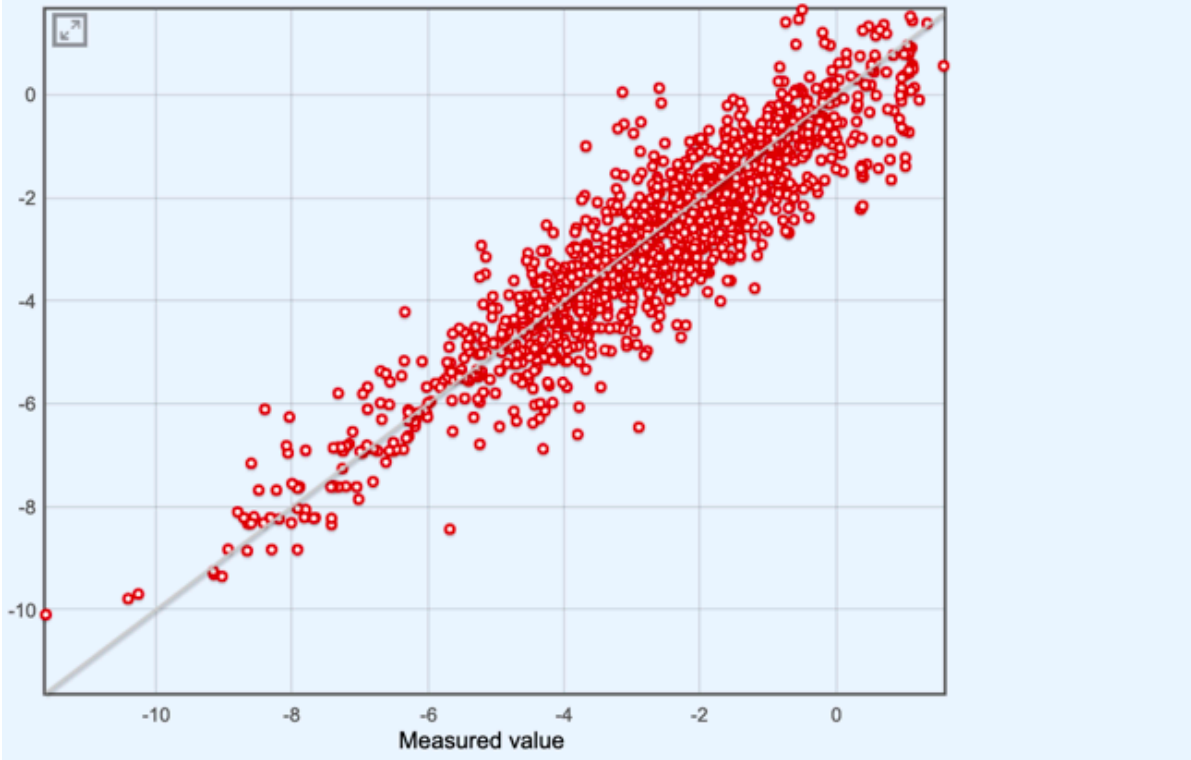


Prediction of solubility using logP and melting point (MP) based on 230k measurements

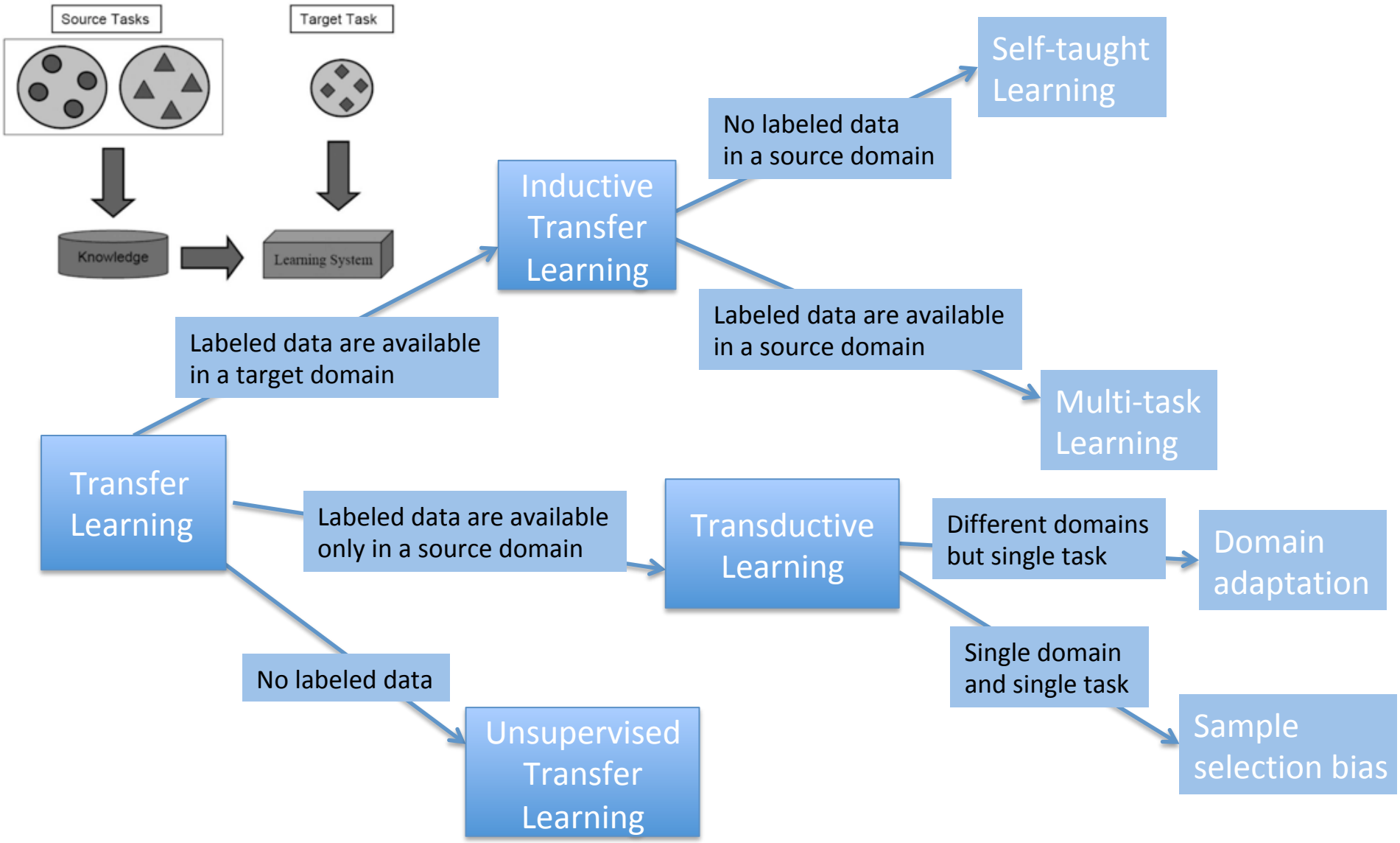
logS = 0.5 – 0.01(MP-25) – logP

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

Data Set	#	R_2	q_2	RMSE	MAE
Training set: ALOGPS 2.1 logS (training)	1311 records	0.842 ± 0.01	0.83 ± 0.01	0.84 ± 0.02	0.64 ± 0.02



Download descriptors and model statistics View configuration XML Export configuration XML



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

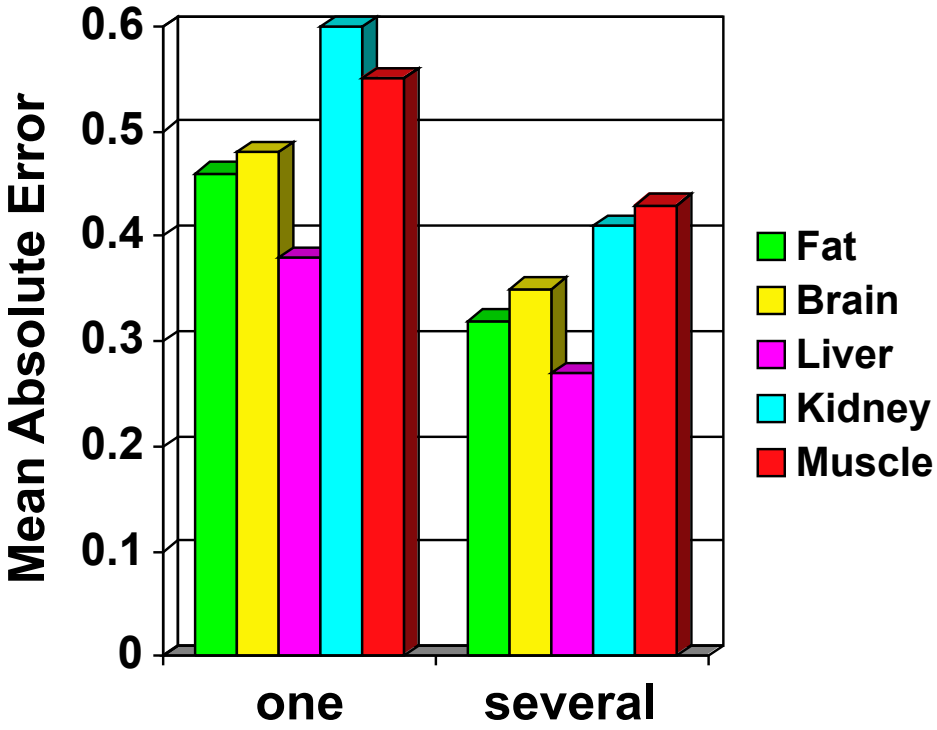
Multi-task learning

Problem:

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

Results:

simultaneous prediction of several properties increased the accuracy of models

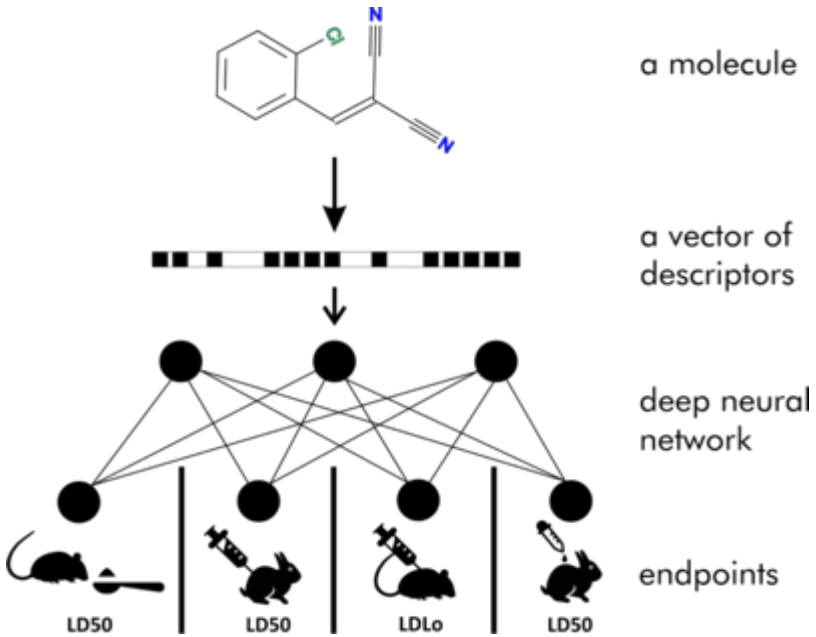


Varnek, A. et al. *J. Chem. Inf. Model.* **2009**, 49, 133-44.

Analysis of toxicity of chemical compounds

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

LDLo



129 142 **toxicity measurements from RTECS***

87 064 **unique molecular structures**

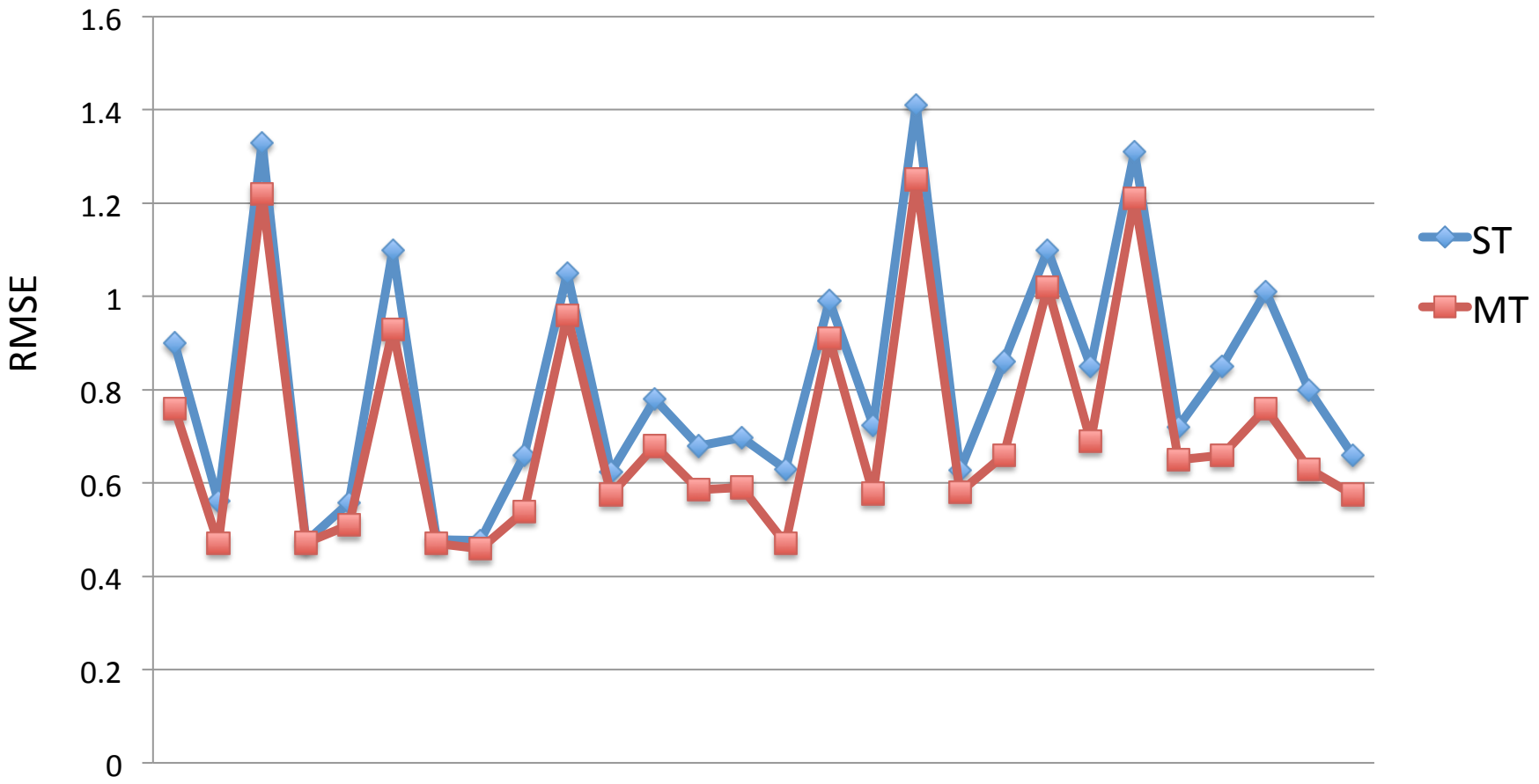
29 **toxicity endpoints**

*RTECS: Registry of Toxic Effects of Chemical Substances

Sosnin, S. et al. *J. Chem. Inf. Model.*, **2019**, 59:1062-1072.

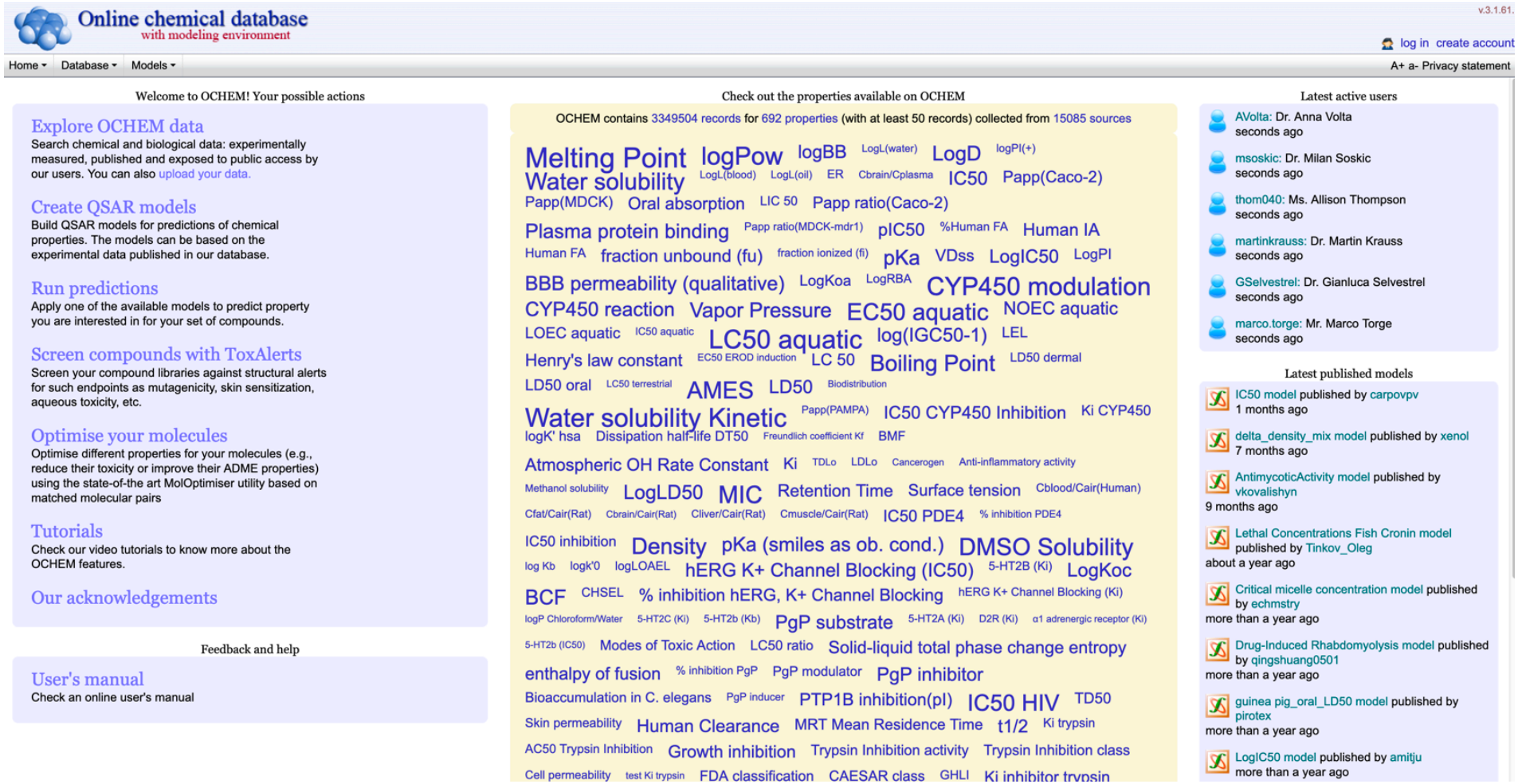
Jain, S. et al. *J. Chem. Inf. Model.* **2021,** *61* (2), 653-663 − extended to >60 endpoints.

Toxicity prediction of single vs. multitask



Sosnin, S. et al. *J. Chem. Inf. Model.*, **2019**, 59:1062-1072.

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



Tetko, I.V.; Maran, U.; Tropsha, A. *Mol. Inform.* **2017**, *36*.

Comparison of different models, RMSE

Metrics 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 0.466 0.439 0.56 1.04 0.584 0.75 0.6 0.65 0.59 0.95 0.66 1.33 0.585 0.75 1.08 0.764 1.3 0.67 0.81 0.88 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 0.47 0.439 0.59 1.02 0.588 0.73 0.61 0.66 0.602 0.94 0.67 1.33 0.585 0.76 1.09 0.77 1.38 0.68 0.82 0.88 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 0.469 0.44 0.56 1.03 0.58 0.73 0.581 0.65 0.61 0.95 0.64 1.34 0.59 0.77 1.11 0.79 1.33 0.69 0.8 0.81 0.75 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 1.07 0.476 0.436 0.58 1.09 0.592 0.75 0.61 0.67 0.59 0.94 0.67 1.3 0.589 0.77 1.14 0.79 1.43 0.69 0.83 0.82 0.77 0.64 1.29 (0.797)

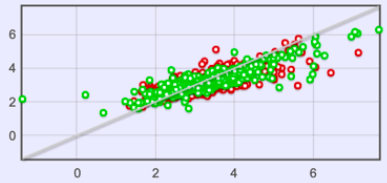
single multi

Comprehensive model view

Model name: Consensus all[apply to new compounds]
Training method: Consensus

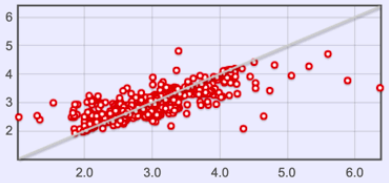
Property: mouse_intraperitoneal_LD50 measured in -log(M) (Details..)

Dataset	R2	RMSE	MAE
data_upload_new.csv(36295)	0.65	0.41	0.26
uniques in RTECS(5189)	0.77	0.38	0.25



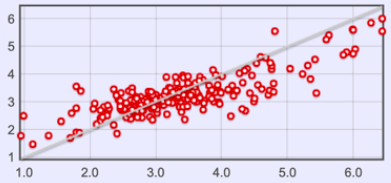
Property: mammal (species unspecified)_intraperitoneal_LD50 measured in -log(M) (Details..)

Dataset	R2	RMSE	MAE
data_upload_new.csv(545)	0.61	0.42	0.26
uniques in RTECS(0)	0.00	0.00	0.00



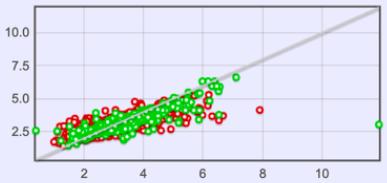
Property: guinea pig_intraperitoneal_LD50 measured in -log(M) (Details..)

Dataset	R2	RMSE	MAE
data_upload_new.csv(248)	0.66	0.60	0.44
uniques in RTECS(0)	0.00	0.00	0.00



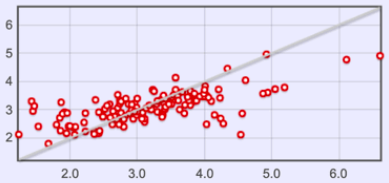
Property: rat_intraperitoneal_LD50 measured in -log(M) (Details..)

Dataset	R2	RMSE	MAE
data_upload_new.csv(5021)	0.63	0.53	0.36
uniques in RTECS(933)	0.72	0.47	0.25



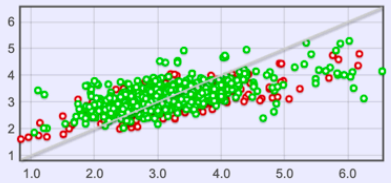
Property: rabbit_intraperitoneal_LD50 measured in -log(M) (Details..)

Dataset	R2	RMSE	MAE
data_upload_new.csv(131)	0.46	0.69	0.50
uniques in RTECS(0)	0.00	0.00	0.00



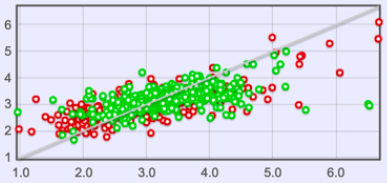
Property: mouse_intraperitoneal_LDLo measured in -log(M) (Details..)

Dataset	R2	RMSE	MAE
data_upload_new.csv(266)	0.54	0.62	0.48
uniques in RTECS(2678)	0.44	0.52	0.38



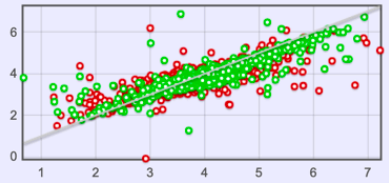
Property: rat_intraperitoneal_LDLo measured in -log(M) (Details..)

Dataset	R2	RMSE	MAE
data_upload_new.csv(318)	0.61	0.55	0.40
uniques in RTECS(757)	0.35	0.46	0.30



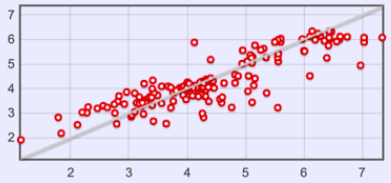
Property: mouse_intravenous_LD50 measured in -log(M) (Details..)

Dataset	R2	RMSE	MAE
data_upload_new.csv(16978)	0.67	0.42	0.28
uniques in RTECS(4151)	0.82	0.35	0.22

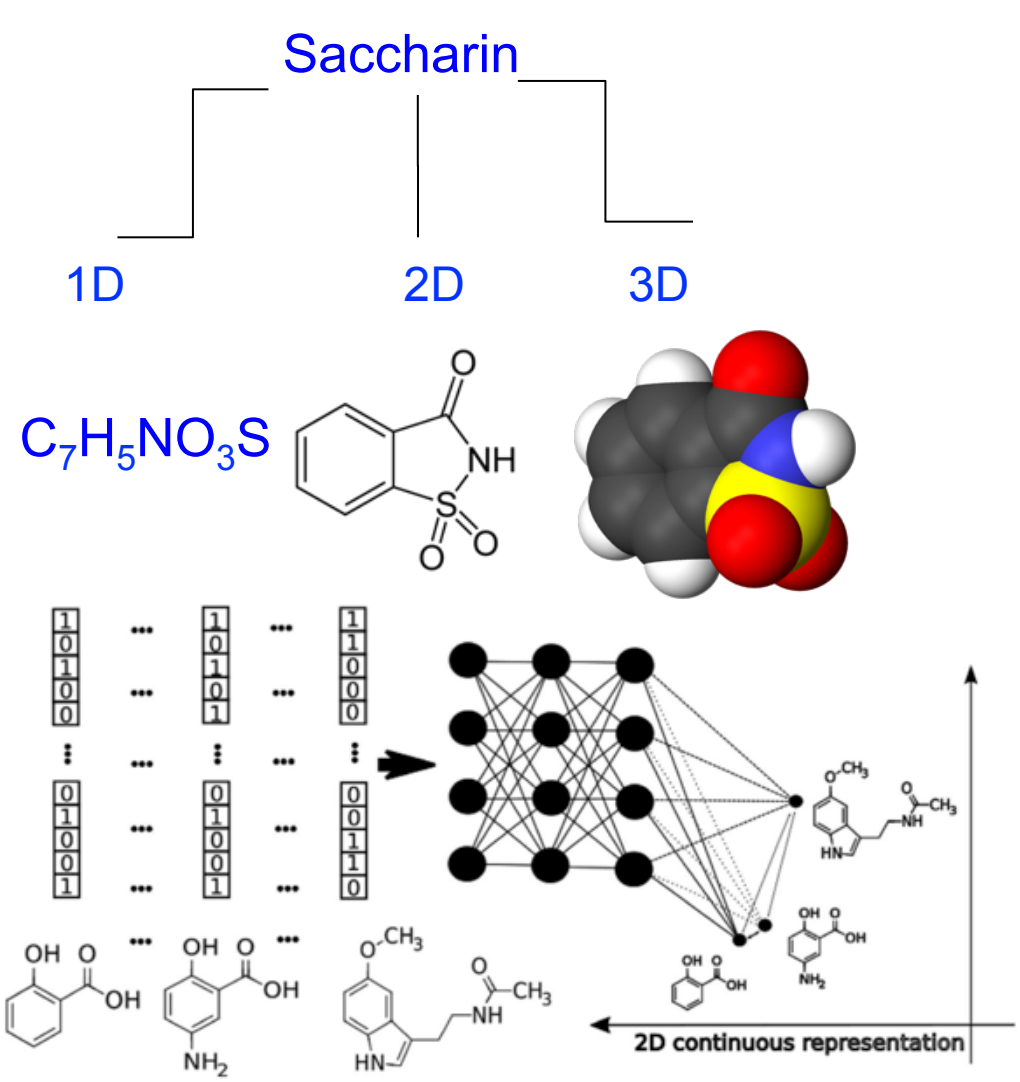


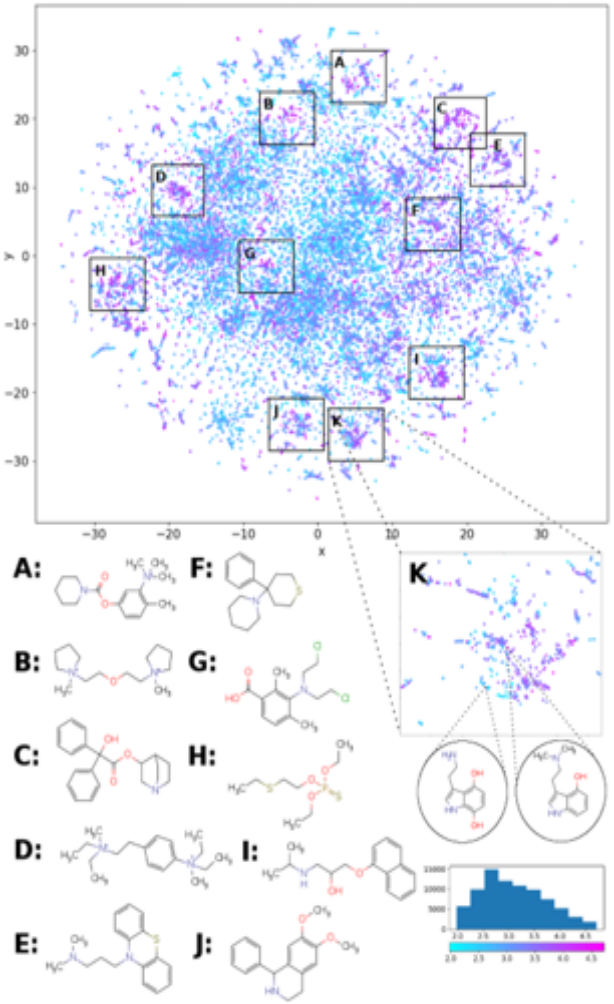
Property: guinea pig_intravenous_LD50 measured in -log(M) (Details..)

Dataset	R2	RMSE	MAE
data_upload_new.csv(153)	0.75	0.64	0.47
uniques in RTECS(0)	0.00	0.00	0.00



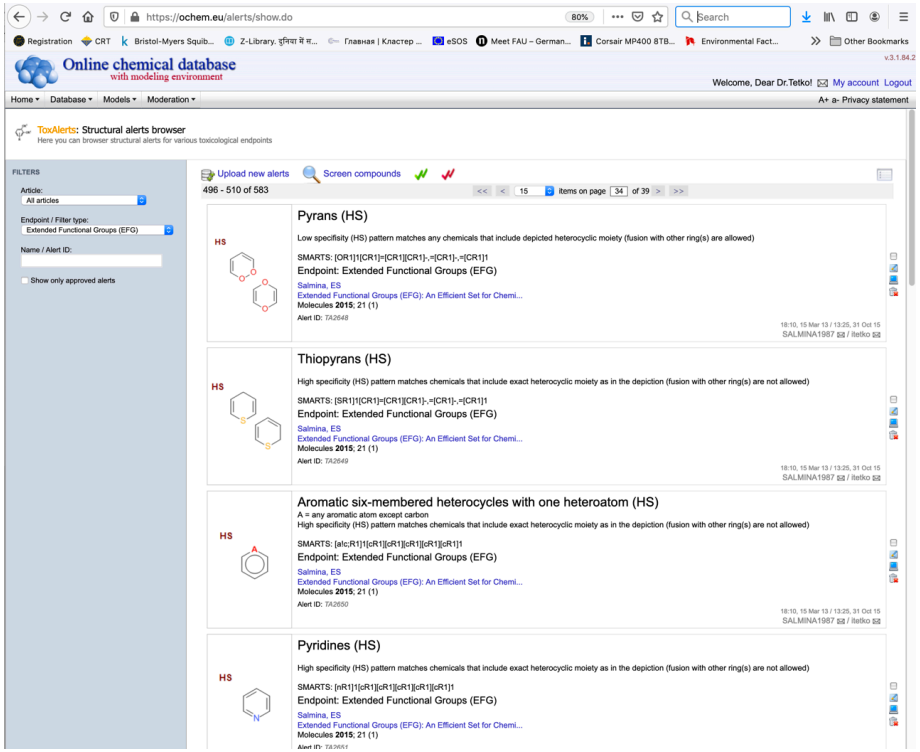
Traditional Representations of Chemical Structures





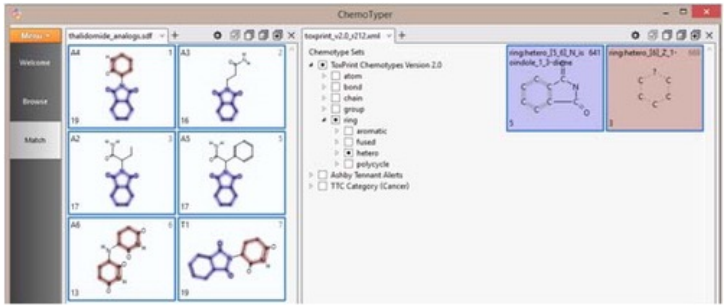
Karlov, D.S. et al. *RSC Advances* **2019**, *9*, 5151-5157.

Design of interpretable descriptors

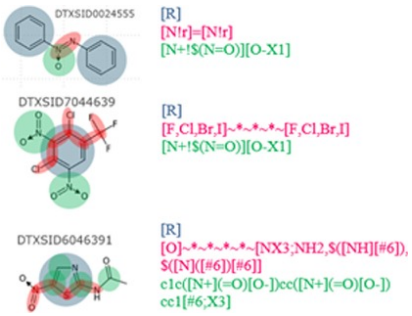


Structural alerts and extended functional groups (**EFG**) http://ochem.eu
Salmina, E.S. et al. *Molecules* **2016**, *21*, 1

Toxprints

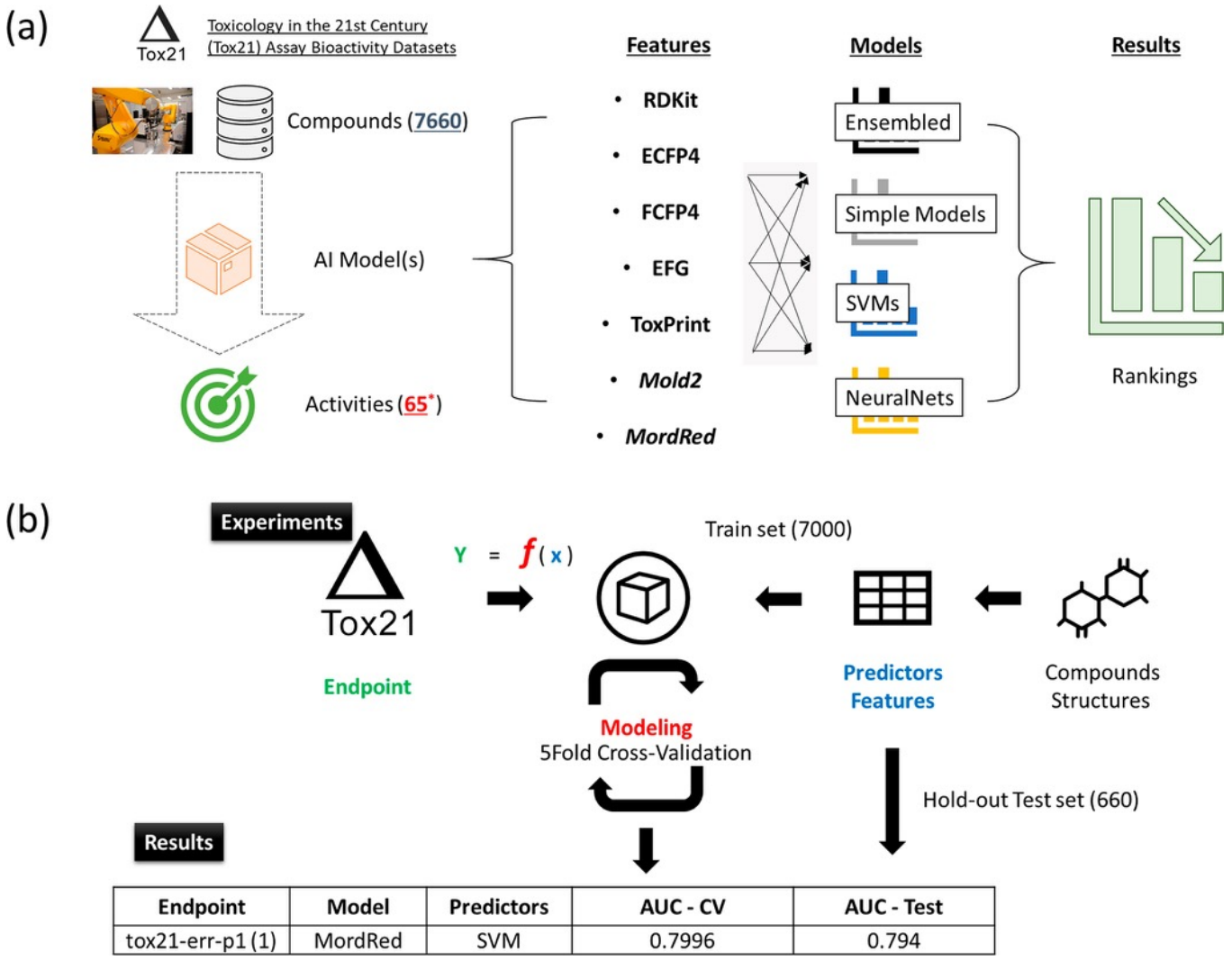


Yang, C. et al. *J. Chem. Inf. Model.* **2015**, *55*, 510-528.



Sedykh, A. et al. *Chem. Res. Toxicol.* **2021**, *34*, 634-640.

06.04.2021

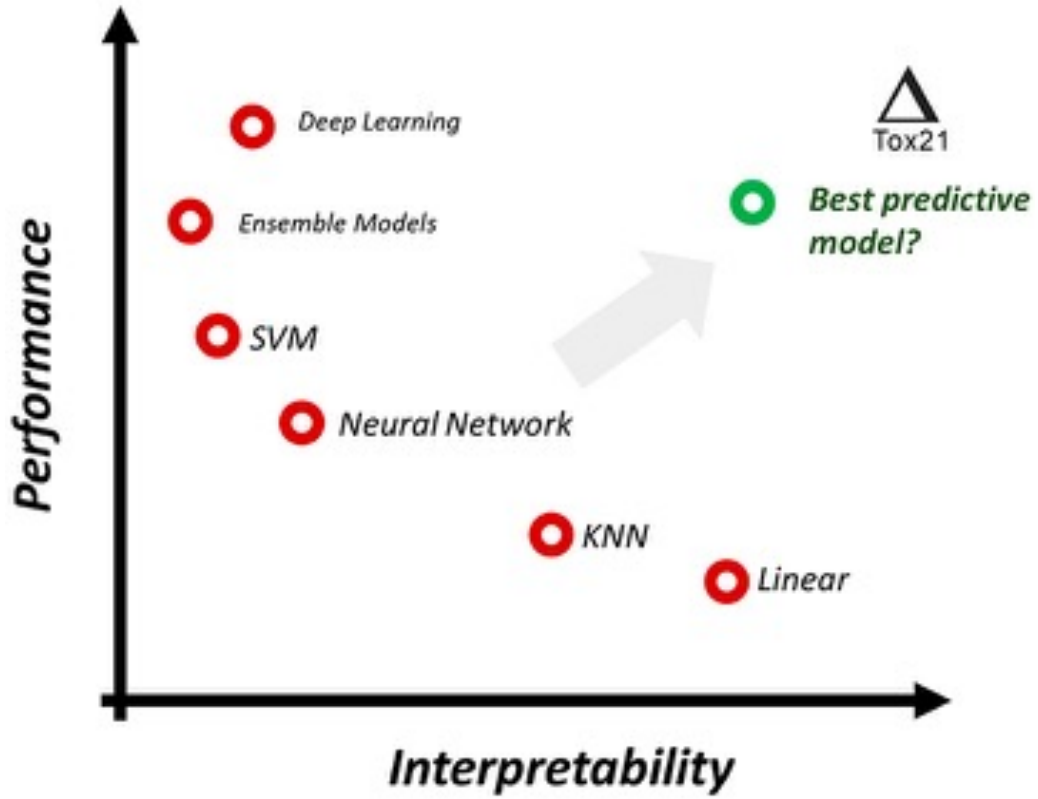


Endpoint	Model	Predictors	AUC - CV	AUC - Test
tox21-err-p1 (1)	MordRed	SVM	0.7996	0.794

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

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

06.04.2021



Overall best performance
5CV/TEST

RF + all: 0.84/0.84

LS-SVM + MORDRED: 0.87/0.88

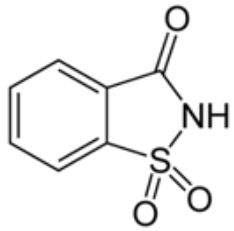
RF + EFG: 0.81/0.86

All + MOLD2: 0.82/0.84

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

Machine Learning directly from chemical structures

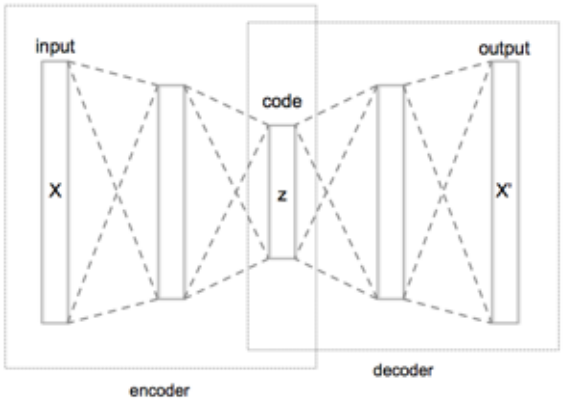
Saccharin: c1ccc2c(c1)C(=O)NS2(=O)=O

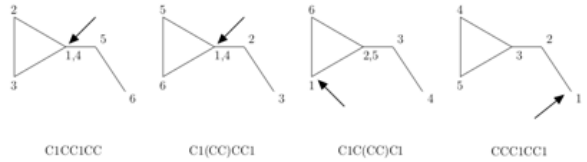


Text processing: convolutional neural networks, Transformers, LSTM

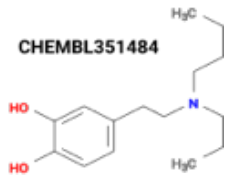
Graph processing: message passing neural networks

Auto-encoder:



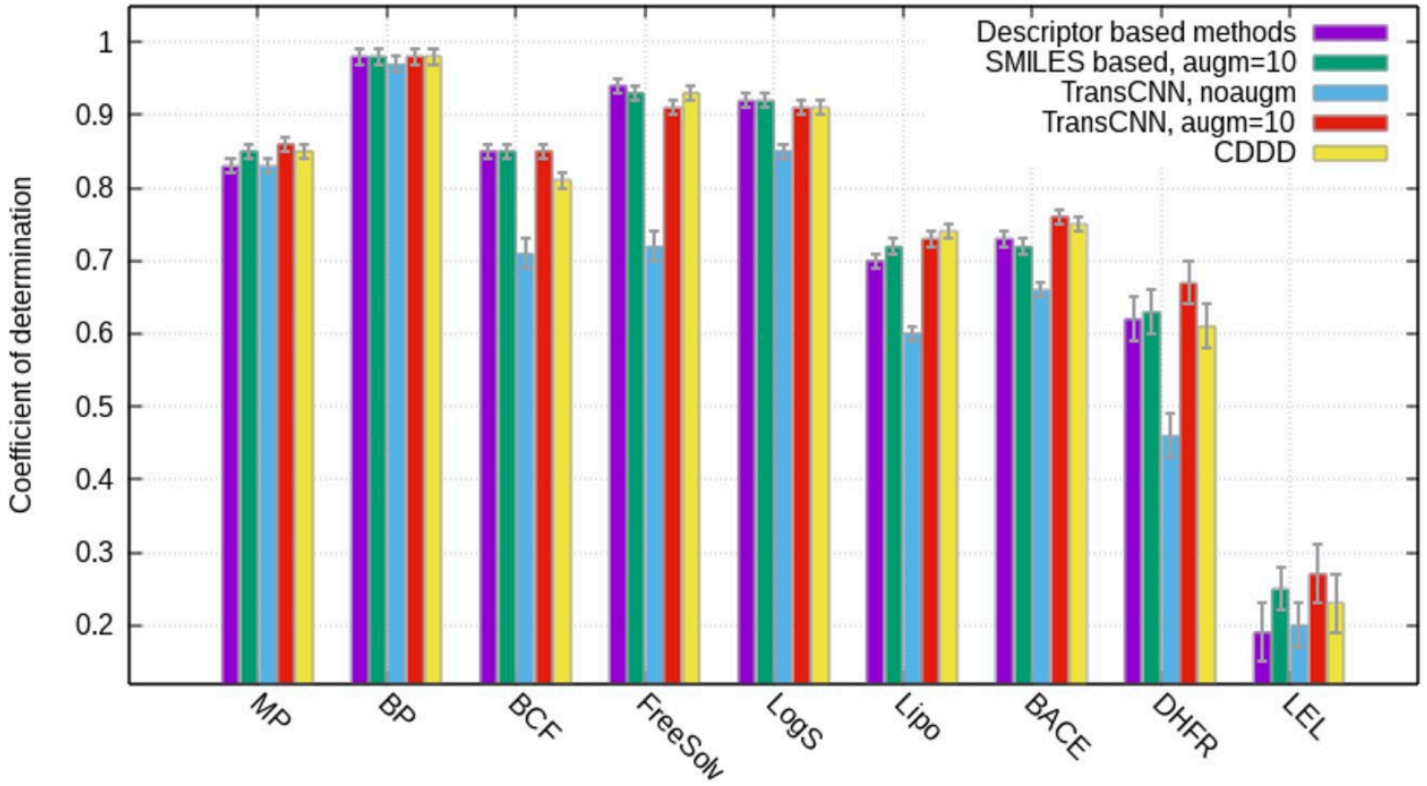


non-canonical		canonical
c1c(ccc(O)c10)CCN(CCCC)CCC	>>	CCCCN(CCc1ccc(c(c1)O)O)CCC
c1(ccc(c(c1)O)O)CCN(CCC)CCCC	>>	CCCCN(CCc1ccc(c(c1)O)O)CCC
c1c(c(cc(c1)CCN(CCCC)CCC)O)O	>>	CCCCN(CCc1ccc(c(c1)O)O)CCC
c1(CCN(CCCC)CCC)ccc(O)c(O)c1	>>	CCCCN(CCc1ccc(c(c1)O)O)CCC
CCCCN(CCc1cc(c(cc1)O)O)CCC	>>	CCCCN(CCc1ccc(c(c1)O)O)CCC
CCCCN(CCc1ccc(c(c1)O)O)CCC	>>	**CCCCN(CCc1ccc(c(c1)O)O)CCC**
N(CCCC)(CCc1ccc(c(O)c1)O)CCC	>>	CCCCN(CCc1ccc(c(c1)O)O)CCC
C(N(CCc1ccc(O)c(c1)O)CCCC)CC	>>	CCCCN(CCc1ccc(c(c1)O)O)CCC
N(CCc1ccc(c(O)c1)O)(CCCC)CCC	>>	CCCCN(CCc1ccc(c(c1)O)O)CCC
c1c(O)c(ccc1CCN(CCC)CCCC)O	>>	CCCCN(CCc1ccc(c(c1)O)O)CCC
c1(c(cc(cc1)CCN(CCC)CCCC)O)O	>>	CCCCN(CCc1ccc(c(c1)O)O)CCC



SMILES canonization by machine learning
→ transfer learning to new data

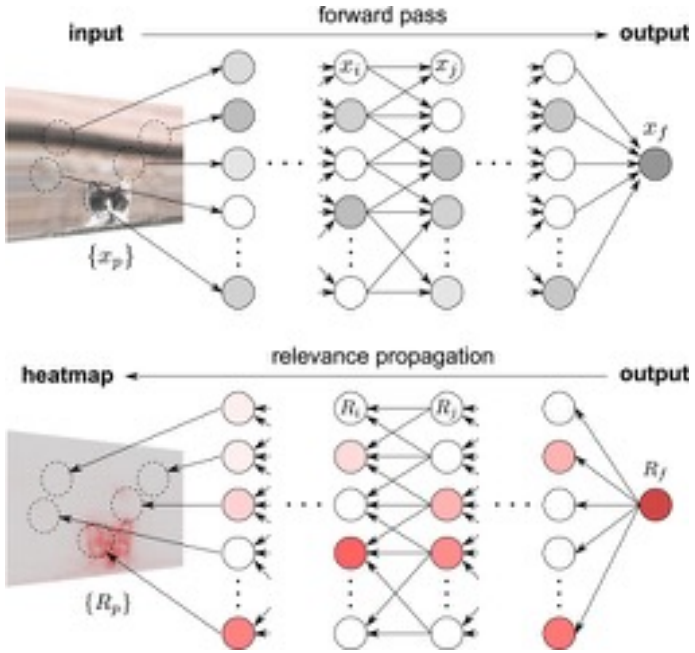
Convolutional vs. Descriptor-based Neural Neural Networks

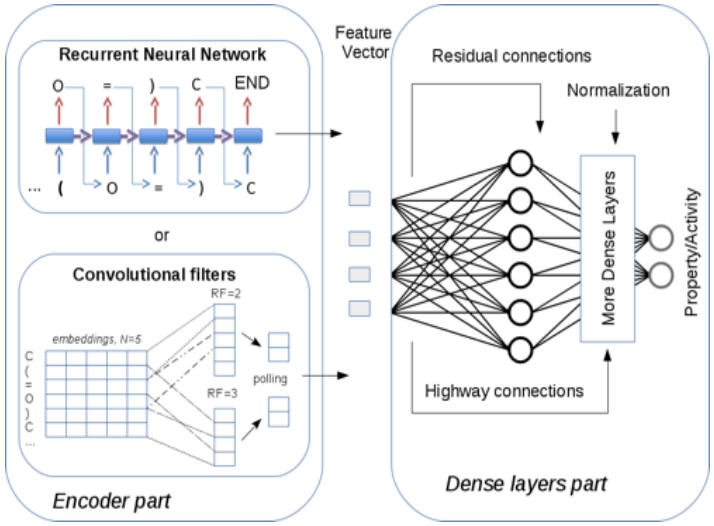


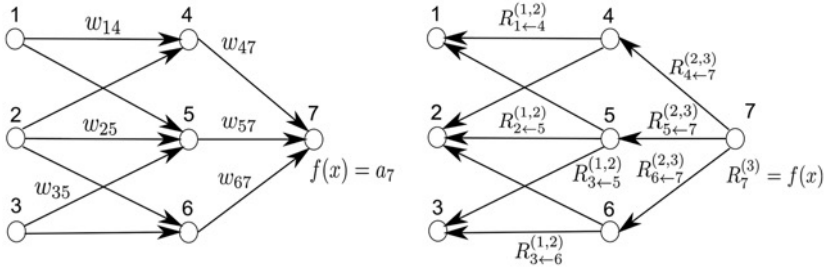
Coefficient of determination, r^2. Transformer CNN provides similar or better accuracy compared to traditional methods based on descriptors even for small datasets (hundreds compounds!). Karpov, P et al. *J. Cheminform.* **2020**, *12*, 17.

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

Layer wise Relevance Propagation (LRP)





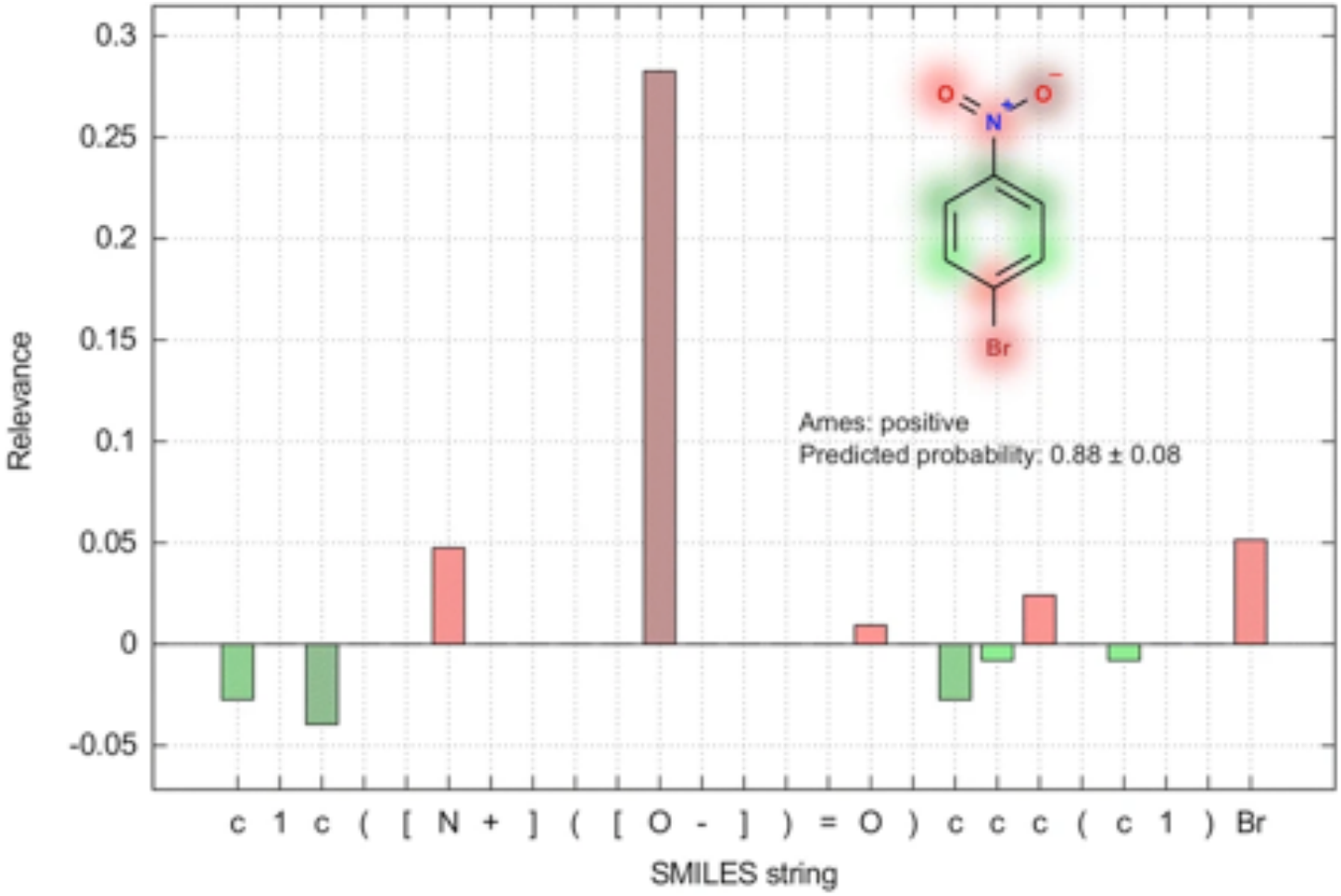


$$f(x) \approx f(x_0) + Df(x_0)[x - x_0]$$
$$= f(x_0) + \sum_{d=1}^{V} \frac{\partial f}{\partial x_{(d)}}(x_0)(x_{(d)} - x_{0(d)})$$

Tetko, I.V. et al. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 794-803.

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

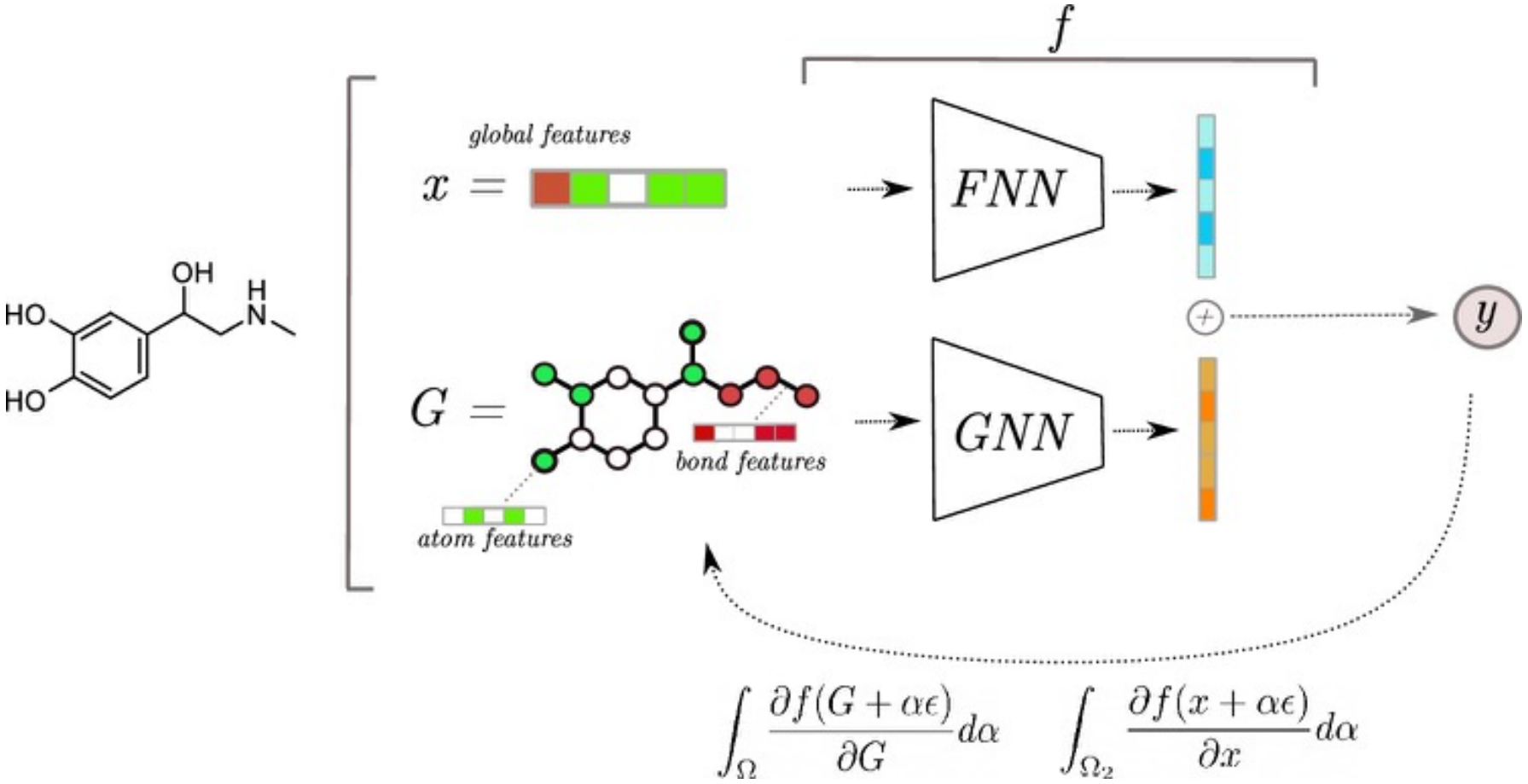
Interpretation of models for Transformer-CNN



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

06.04.2021

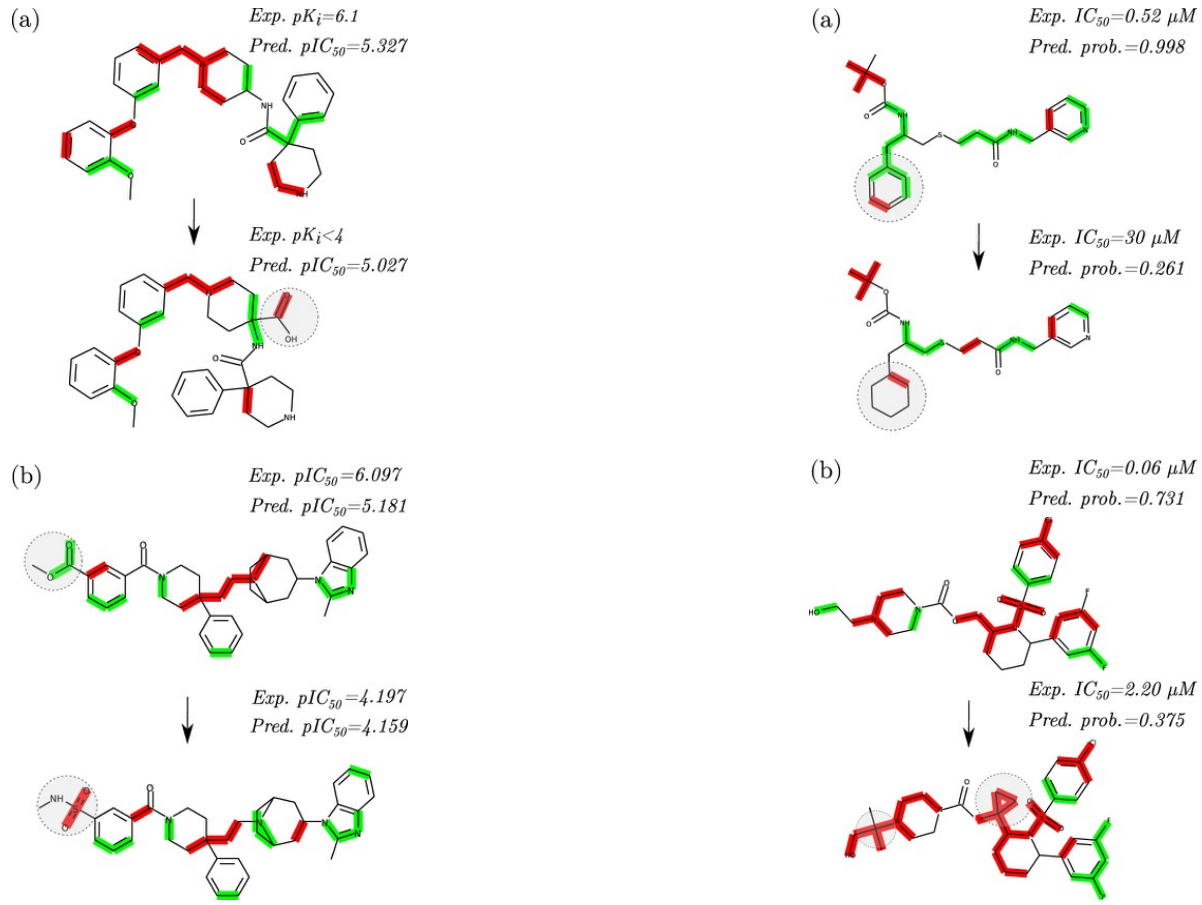
https://github.com/bigchem/transformer-cnn



Schematic of the XAI methodology and neural network architecture. A message-passing graph neural network (GNN) and a forward fully connected neural network (FNN) were combined to process an input presented as a molecular graph with atom, bond, and computed global properties (*e.g.*, octanol–water partition coefficient, topological polar surface area). The **integrated gradients method** was applied to compute atom, bond, and global importance scores.

Jiménez-Luna, J.; et al *J. Chem. Inf. Model.* **2021**, 10.1021/acs.jcim.0c01344.

06.04.2021

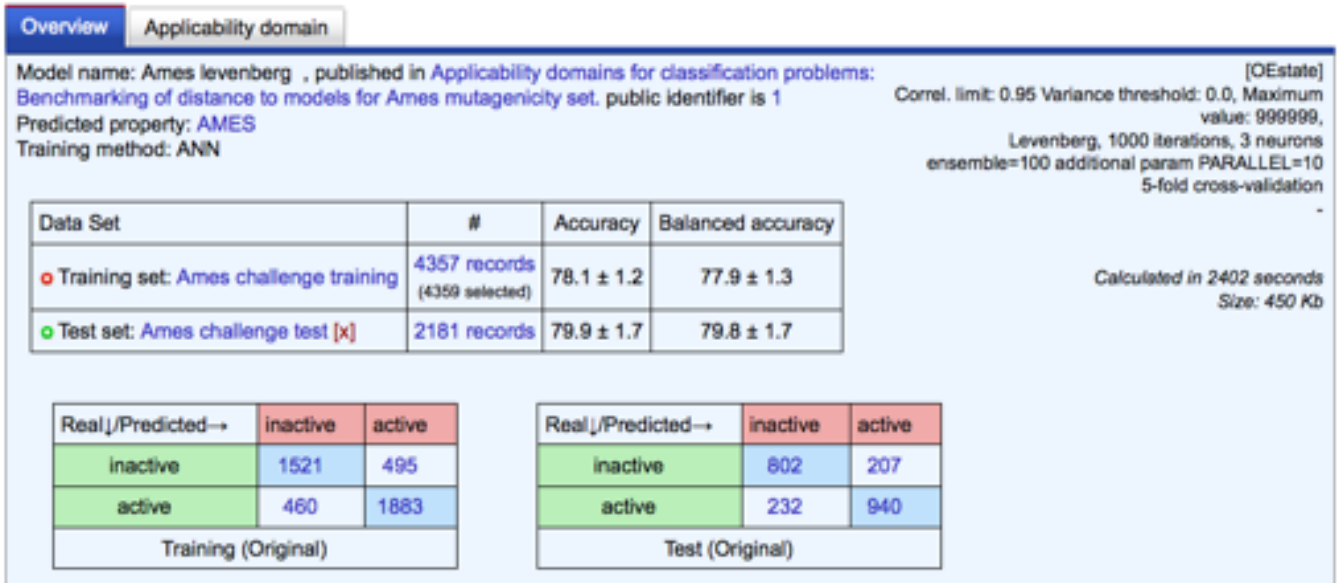


Examples of motifs indicating a) hERG and b) CYP450 inhibition.

Jiménez-Luna, J.; et al. *J. Chem. Inf. Model.* **2021**, 10.1021/acs.jcim.0c01344.

06.04.2021

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

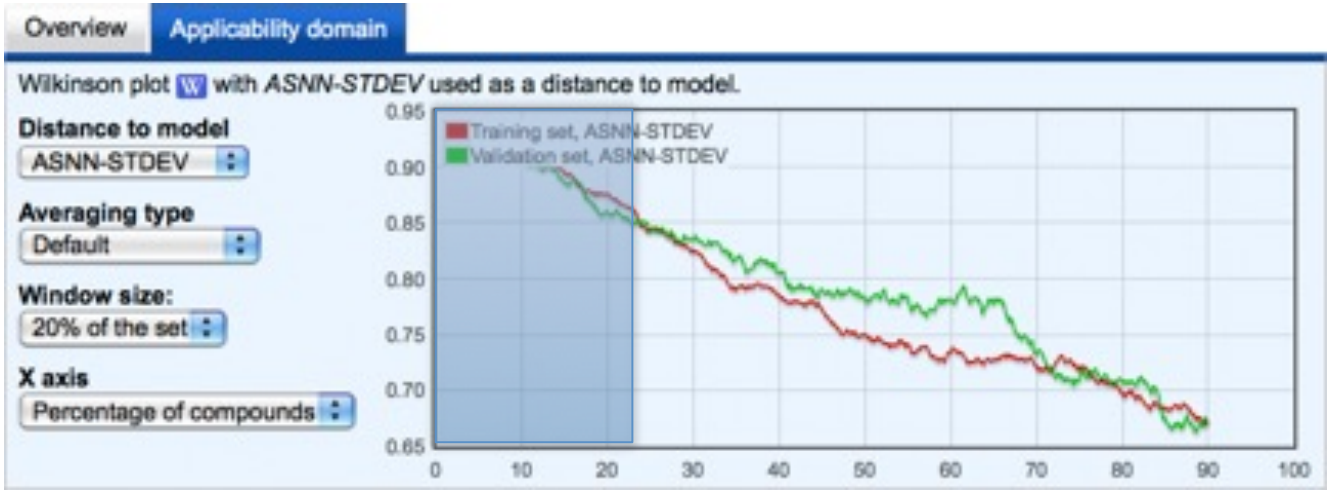
[OEstate]
Correl. limit: 0.95 Variance threshold: 0.0, Maximum value: 999999,
Levenberg, 1000 iterations, 3 neurons
ensemble=100 additional param PARALLEL=10
5-fold cross-validation
-

Calculated in 2402 seconds
Size: 450 Kb

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

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

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



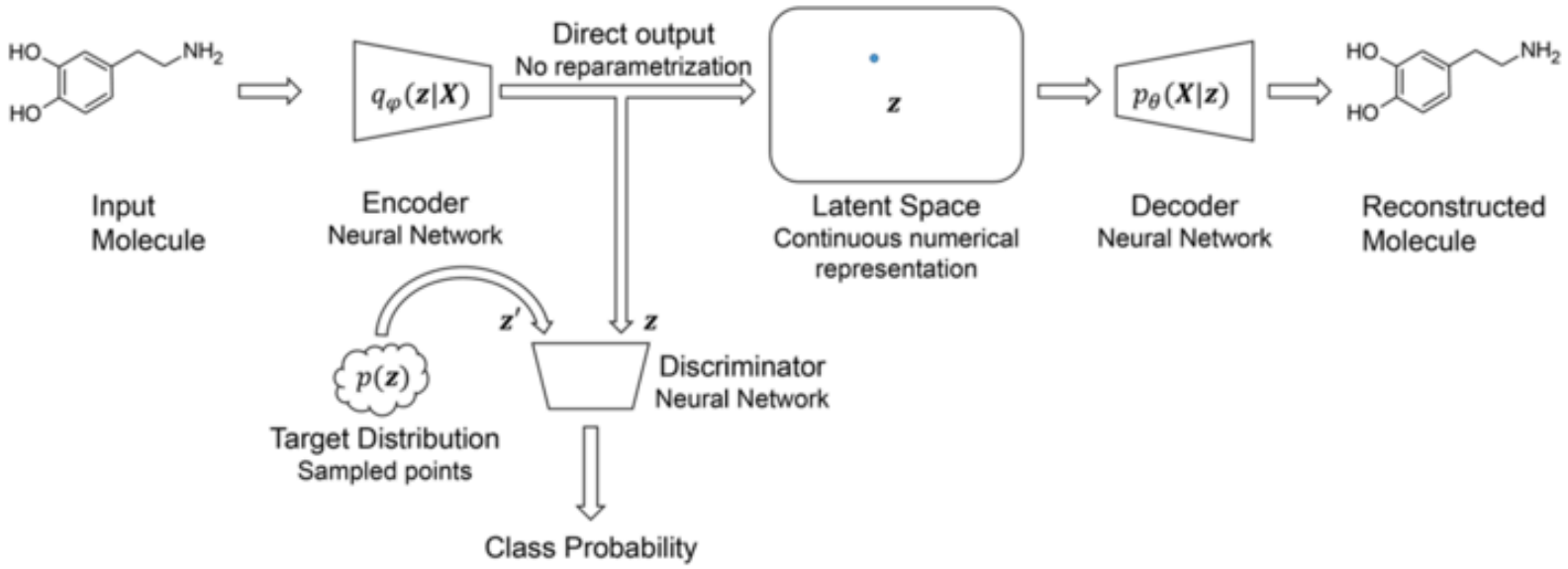
DOI: 10.1002/minf.201700123

Application of Generative Autoencoder in de Novo Molecular Design

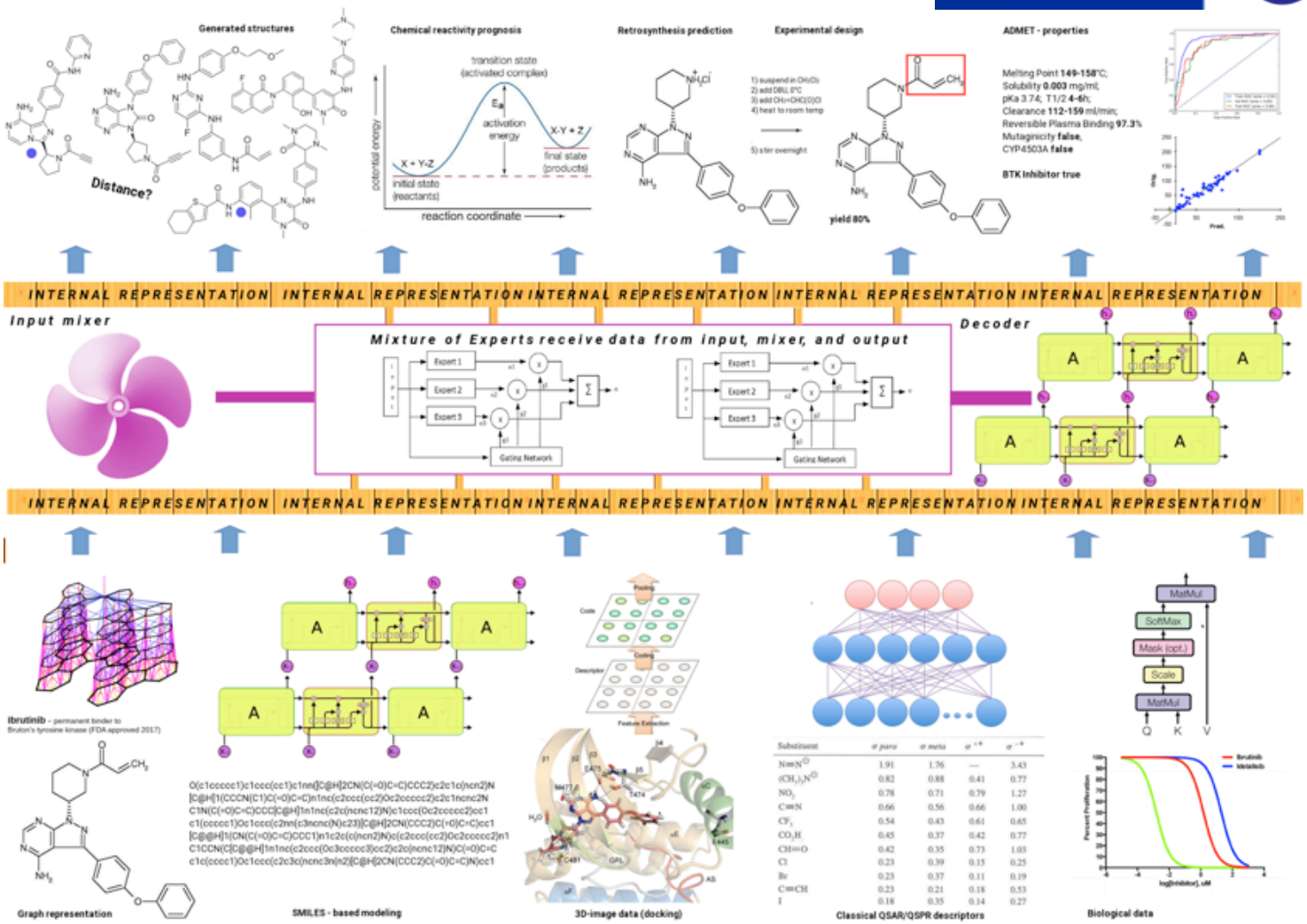
Thomas Blaschke,*[a, b] Marcus Olivecrona,[a] Ola Engkvist,[a] Jürgen Bajorath,[b] and Hongming Chen*[a]

Abstract: A major challenge in computational chemistry is the generation of novel molecular structures with desirable pharmacological and physiochemical properties. In this work, we investigate the potential use of autoencoder, a deep learning methodology, for de novo molecular design. Various generative autoencoders were used to map molecule structures into a continuous latent space and vice versa and their performance as structure generator was assessed. Our results show that the latent space preserves chemical similarity principle and thus can be used for the generation of analogue structures. Furthermore, the latent space created by autoencoders were searched systematically to generate novel compounds with predicted activity against dopamine receptor type 2 and compounds similar to known active compounds not included in the trainings set were identified.

Keywords: Autoencoder · chemoinformatics · de novo molecular design · deep learning · inverse QSAR



Advanced machine Learning for Innovative drug discovery



06.04.2021

15 PhD positions at http://ai-dd.eu *deadline April 18th!*

Take home message

- ADMETox modeling depends on the quality and amount of data
- Text processing methods can automatically generate data
  - Data extraction from Patents is a maturated technology
  - Image processing methods can extract data from books, reports, pdf
- Simultaneous modeling of related properties increase model quality
- Use of interpretable descriptors and interpretable methods should not be neglected
- Use of descriptor-less methods contributes highly predictive models
- Explainable Artificial Intelligence (XAI) to explain models is on the rise

*"Compared to Big Data challenges, "how to best analyze the Big Data", **the future progress is linked to the need for explainable "chemistry aware" methods.**"**

*Tetko, I.V.; Engkvist, O. *J. Cheminform.* **2020**, *12*, 74.

Acknowledgement

Pavel Karpov
Zhonghua Xia
Mark Embrechts
Dipan Ghosh
Joseph Yap
Elena Dracheva
Genny Cau
Monica Campillos

Guillaume Godin (Firmenich)
Sergey Sosnin (Skoltech)
Maxim Fedorov (Skoltech)

Yura Sushko (Google)
Sergey Novotarskyi (Facebook)
Robert Körner

M. Sattler (HMGU)