

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

Год науки и технологий 21

5 - 7 Апрель, 2021

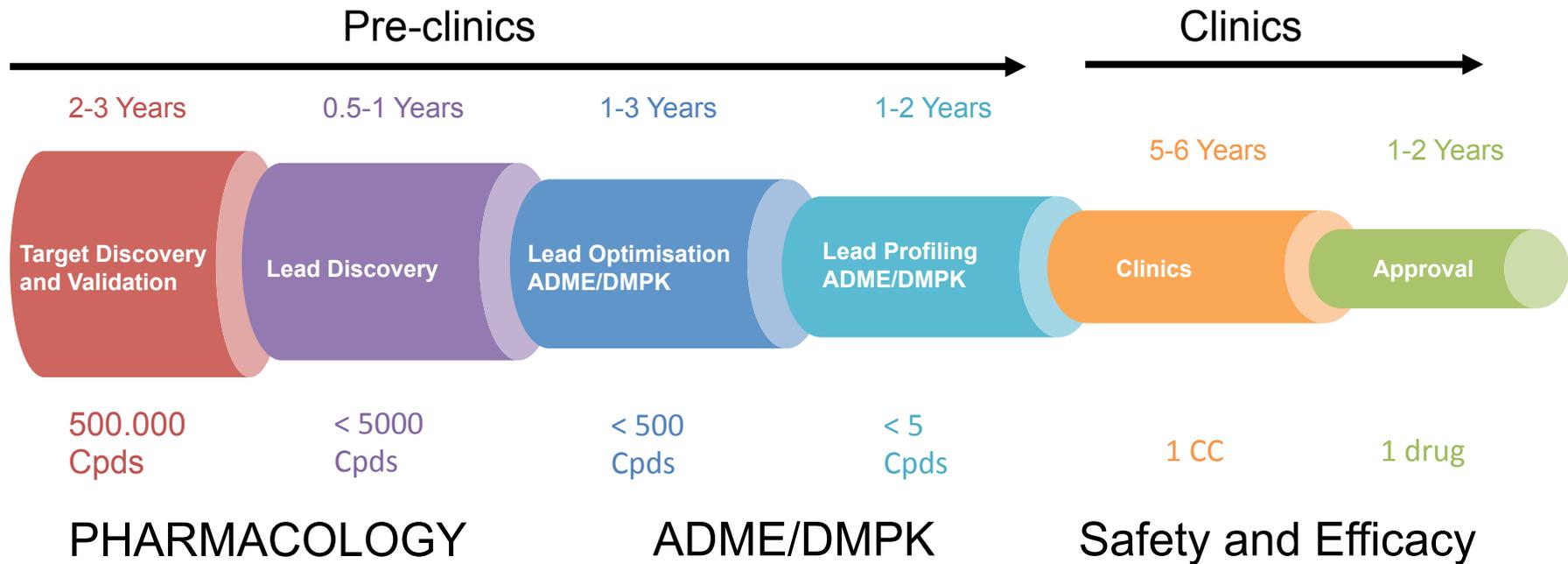
ONLINE

XXVII Симпозиум «Биоинформатика и компьютерное конструирование лекарств»

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

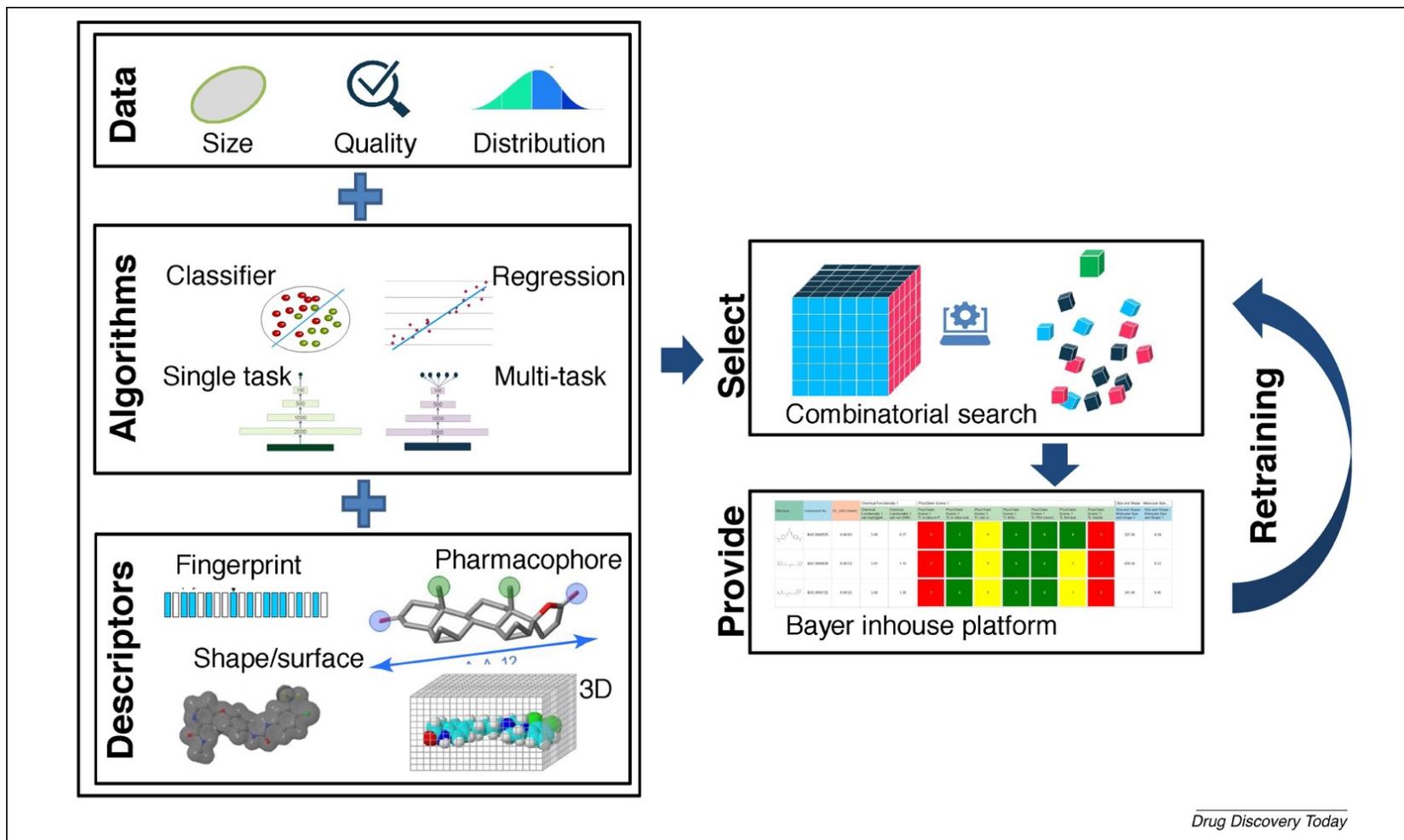
		Insufficient quality	First approach	Medium model	Good model	Robust model		
Endpoint		Model type	Data set size	2005	2009	2014	2019	Retraining
Absorption	Caco-2 permeation	C (N)	>10 000			RF	SVR	Weekly
	Caco-2 efflux	C (N)	>10 000			RF	SVR	Weekly
	Bioavailability (rat)	C	~2000				RF	On demand
Distribution	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.

06.04.2021

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]⁺.

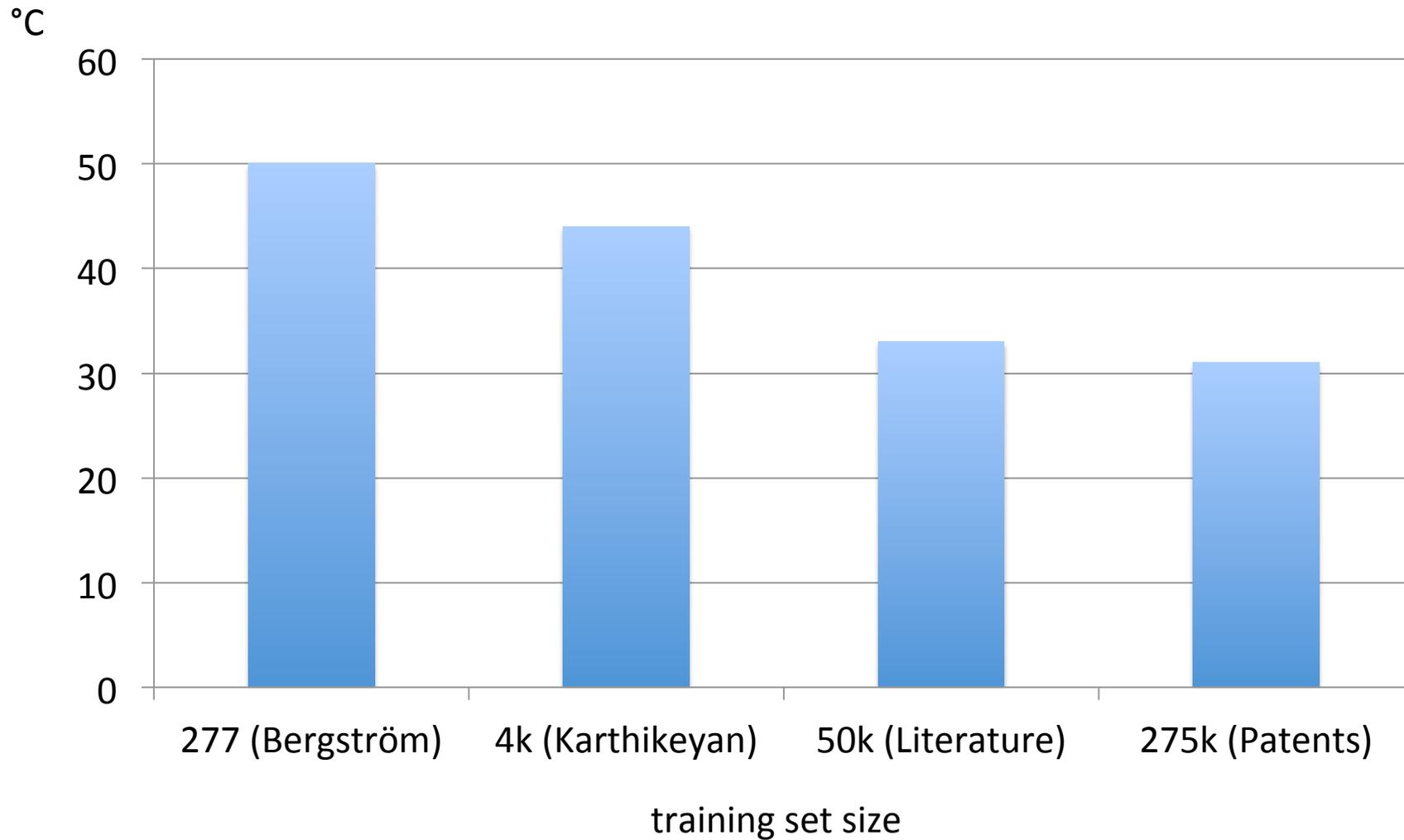
<http://ochem.eu>

Basket Records Tags

1 - 5 of 23 5 items on page 1 of 5 >>

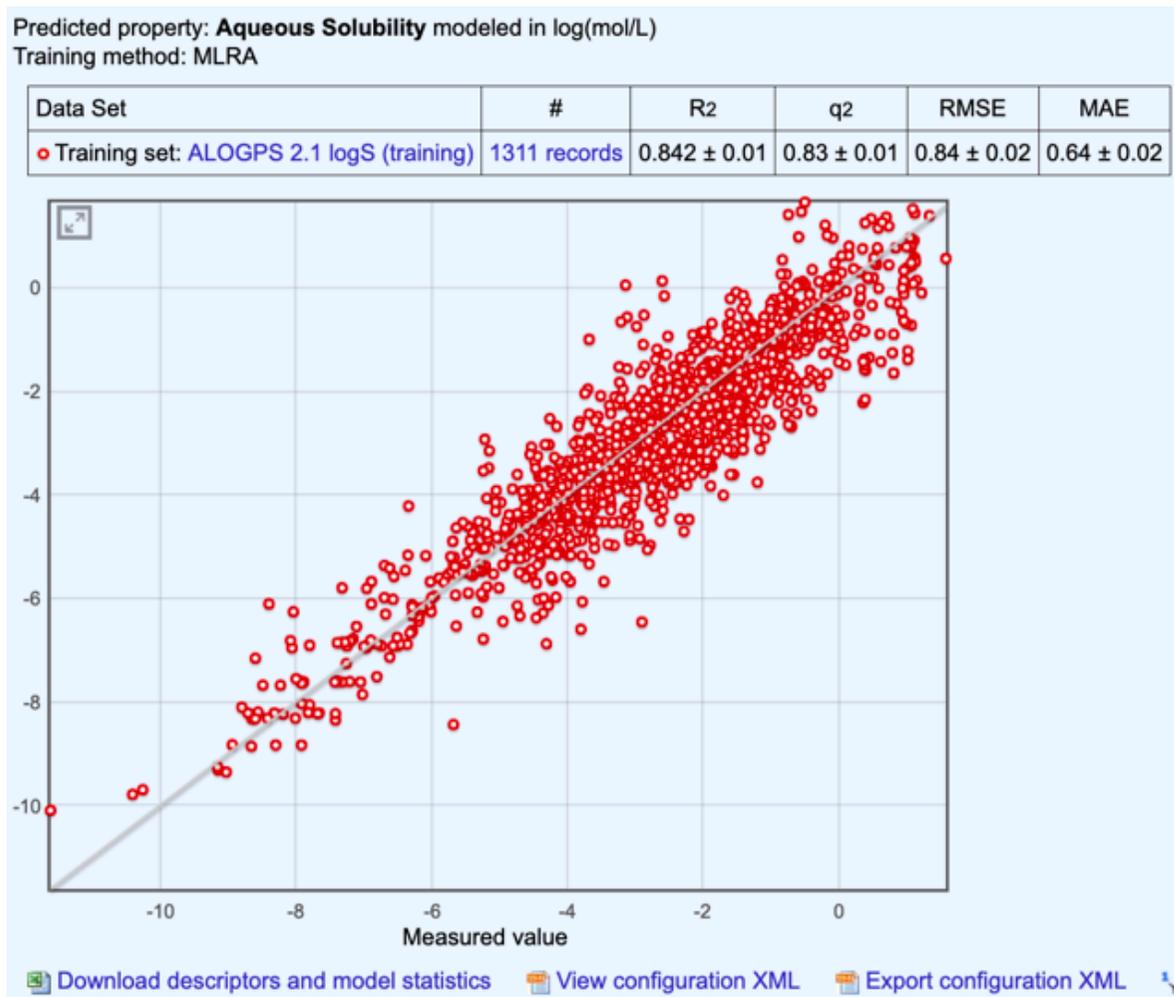
 molecule profile	<p>● Melting Point = 215.0 - 217.0 (in °C)</p> <p>Tetko, I.V. et al The development of models to predict melting and pyrolysis p... N: AUTO_261086 Journal of cheminformatics 2016; 8 () 2</p> <p>5,7-Dimethoxy-2-(2-(2-(pyrrolidin-1-yl)ethyl)-1H-indol-5-yl)quinazolin-4(3H)-one MoleculeID: M84181205</p> <p>Public and freely downloadable record</p> <p>RecordID: R21022022 02:50, 12 Aug 15 / 14:37, 3 Oct 17 dan2097 / published</p>
 molecule profile	<p>● Melting Point = 289.0 - 290.0 (in °C)</p> <p>Tetko, I.V. et al The development of models to predict melting and pyrolysis p... N: AUTO_265261 Journal of cheminformatics 2016; 8 () 2</p> <p>5,7-dimethoxy-2-(3-(5-(methylsulfinyl)thiophen-2-yl)phenyl)quinazolin-4(3H)-one MoleculeID: M83437784</p> <p>Public and freely downloadable record</p> <p>RecordID: R21026197 02:53, 12 Aug 15 / 14:37, 3 Oct 17 dan2097 / published</p>

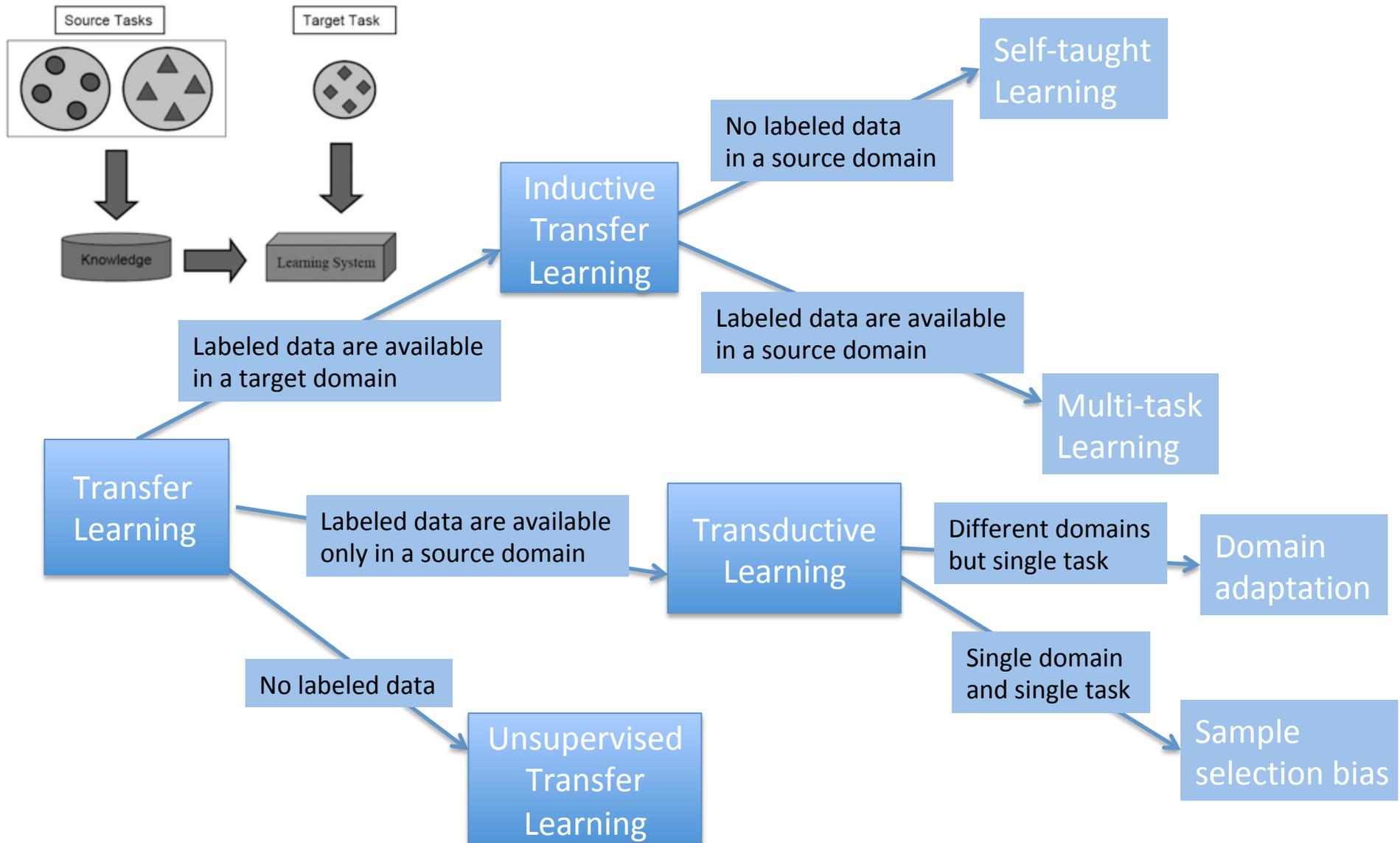
Prediction errors for a set of drugs (Bergström dataset) using models developed with different training sets



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

$$\log S = 0.5 - 0.01(\text{MP}-25) - \log P$$





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

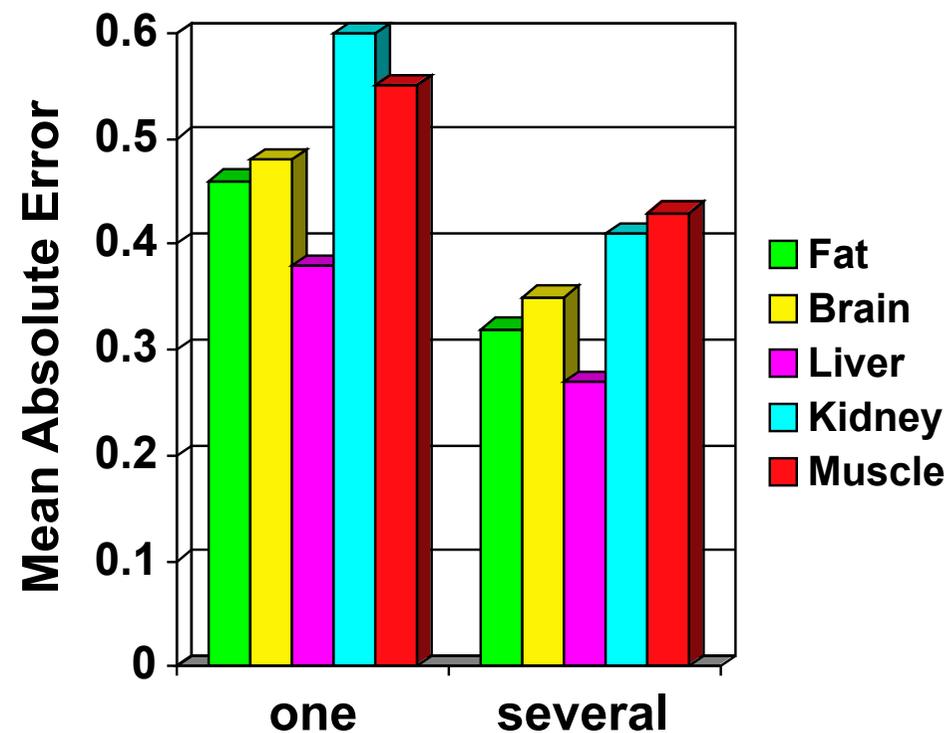
Multi-task learning

Problem:

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

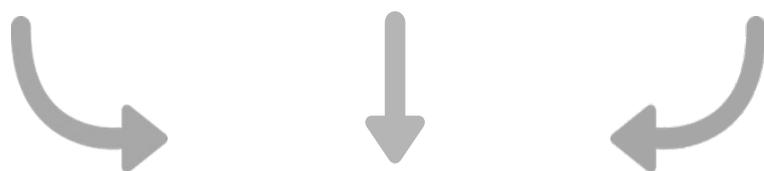
Results:

simultaneous prediction of several properties increased the accuracy of models



Analysis of toxicity of chemical compounds

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



LDLo

129 142 toxicity measurements
from RTECS*

87 064 unique molecular
structures

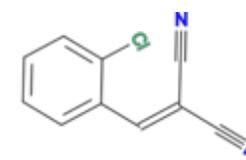
29 toxicity endpoints

*RTECS: Registry of Toxic Effects of Chemical Substances

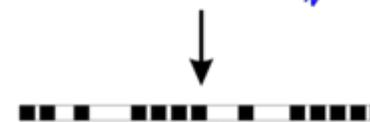
Sosnin, S. et al. *J. Chem. Inf. Model.*, **2019**, 59:1062-1072.

06.04.2021

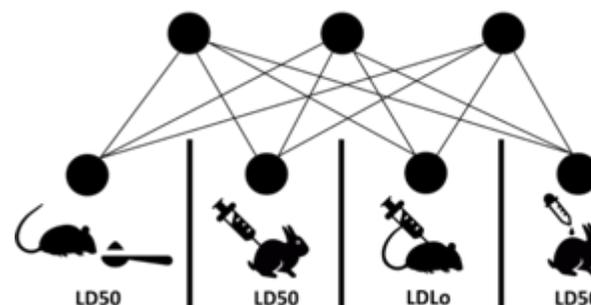
Jain, S. et al. *J. Chem. Inf. Model.* **2021**, 61 (2), 653-663 – extended to >60 endpoints.



a molecule



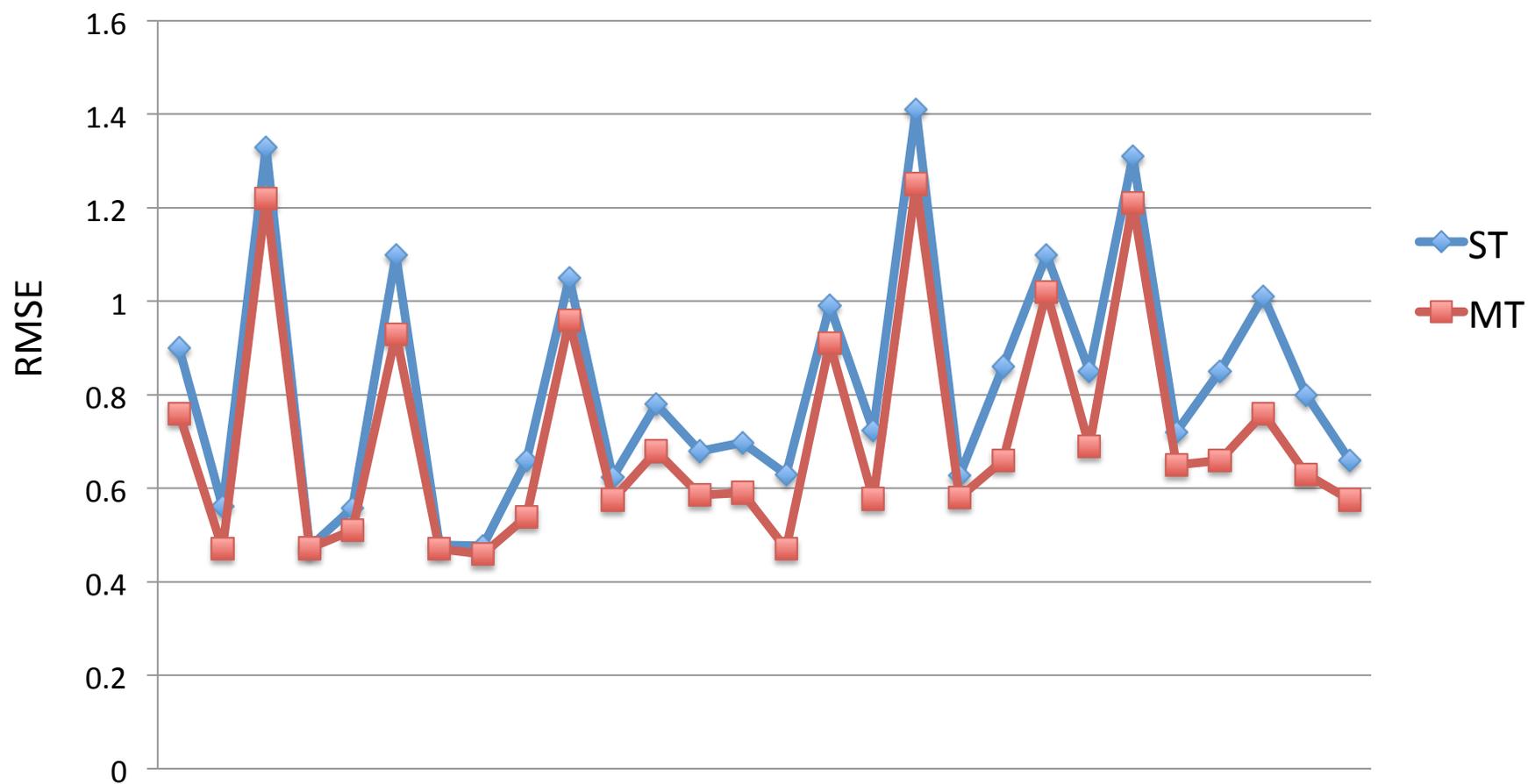
a vector of
descriptors



deep neural
network

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>

Online chemical database
with modeling environment

v.3.1.61

log in create account

Home Database Models A+ a- Privacy statement

Welcome to OCHEM! Your possible actions

Check out the properties available on OCHEM

OCHEM contains 3349504 records for 692 properties (with at least 50 records) collected from 15085 sources

Latest active users

Latest published models

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.

Optimise your molecules

Optimise different properties for your molecules (e.g., reduce their toxicity or improve their ADME properties) using the state-of-the-art MolOptimiser utility based on matched molecular pairs

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

Melting Point logPow logBB LogL(water) LogD logPI(+)
Water solubility LogL(blood) LogL(oil) ER Cbrain/Cplasma IC50 Papp(Caco-2)
Papp(MDCK) Oral absorption LIC 50 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(IC50-1) LEL
Henry's law constant EC50 EROD induction LC 50 Boiling Point LD50 dermal
LD50 oral LC50 terrestrial AMES LD50 Biodistribution
Water solubility Kinetic Papp(PAMPA) IC50 CYP450 Inhibition Ki CYP450
logK^h hsa Dissipation half-life DT50 Freundlich coefficient Kf BMF
Atmospheric OH Rate Constant Ki TDL_o LDLo Cancerogen Anti-inflammatory activity
Methanol solubility LogLD50 MIC Retention Time Surface tension Cblood/Cair(Human)
Cfat/Cair(Rat) Cbrain/Cair(Rat) Cliver/Cair(Rat) Cmuscle/Cair(Rat) IC50 PDE4 % inhibition PDE4
IC50 inhibition Density pKa (smiles as ob. cond.) DMSO Solubility
log Kb logk⁰ logLOAEL hERG K+ Channel Blocking (IC50) 5-HT2B (Ki) LogKoc
BCF CHSEL % inhibition hERG, K+ Channel Blocking hERG K+ Channel Blocking (Ki)
logP Chloroform/Water 5-HT2C (Ki) 5-HT2b (Kb) Pgp substrate 5-HT2A (Ki) D2R (Ki) α1 adrenergic receptor (Ki)
5-HT2b (IC50) Modes of Toxic Action LC50 ratio Solid-liquid total phase change entropy
enthalpy of fusion % inhibition Pgp Pgp modulator Pgp inhibitor
Bioaccumulation in C. elegans Pgp inducer PTP1B inhibition(pl) IC50 HIV TD50
Skin permeability Human Clearance MRT Mean Residence Time t1/2 Ki trypsin
AC50 Trypsin Inhibition Growth inhibition Trypsin Inhibition activity Trypsin Inhibition class
Cell permeability test Ki trypsin FDA classification CAESAR class GHLI Ki inhibitor trypsin

AVolta: Dr. Anna Volta
seconds ago

msoskic: Dr. Milan Soskic
seconds ago

thom040: Ms. Allison Thompson
seconds ago

martinkrauss: Dr. Martin Krauss
seconds ago

GSelvestrel: Dr. Gianluca Selvestrel
seconds ago

marco.torge: Mr. Marco Torge
seconds ago

IC50 model published by carpovpv
1 months ago

delta_density_mix model published by xenol
7 months ago

AntimycoticActivity model published by vkovalishyn
9 months ago

Lethal Concentrations Fish Cronin model published by Tinkov_Oleg
about a year ago

Critical micelle concentration model published by echnstry
more than a year ago

Drug-Induced Rhabdomyolysis model published by qingshuang0501
more than a year ago

guinea pig_oral_LD50 model published by pirotex
more than a year ago

LogIC50 model published by amitju
more than a year ago

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

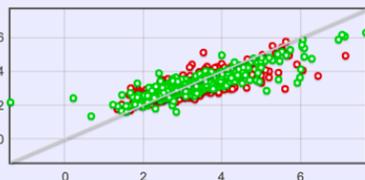
multi

Comprehensive model view

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

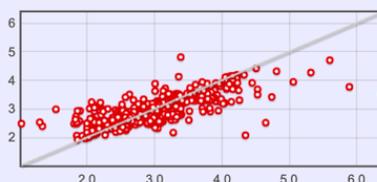
Property: mouse_intraperitoneal_LD50 measured in -log(M) ([Details..](#))

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



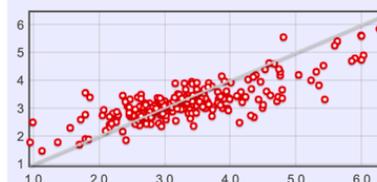
Property: mammal (species unspecified)_intraperitoneal_LD50 measured in -log(M) ([Details..](#))

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



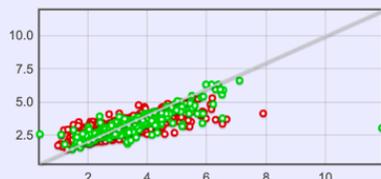
Property: guinea pig_intraperitoneal_LD50 measured in -log(M) ([Details..](#))

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



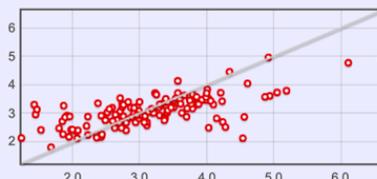
Property: rat_intraperitoneal_LD50 measured in -log(M) ([Details..](#))

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



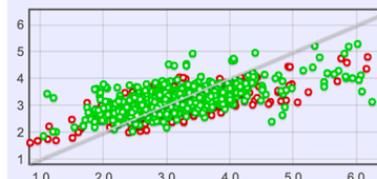
Property: rabbit_intraperitoneal_LD50 measured in -log(M) ([Details..](#))

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



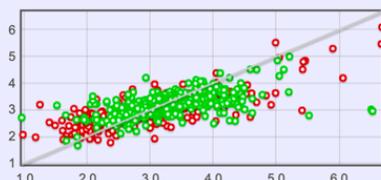
Property: mouse_intraperitoneal_LDLo measured in -log(M) ([Details..](#))

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



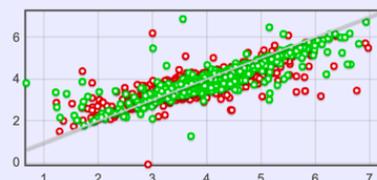
Property: rat_intraperitoneal_LDLo measured in -log(M) ([Details..](#))

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



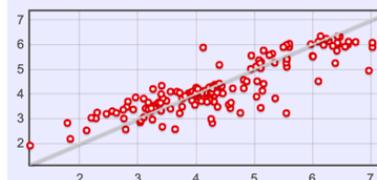
Property: mouse_intravenous_LD50 measured in -log(M) ([Details..](#))

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

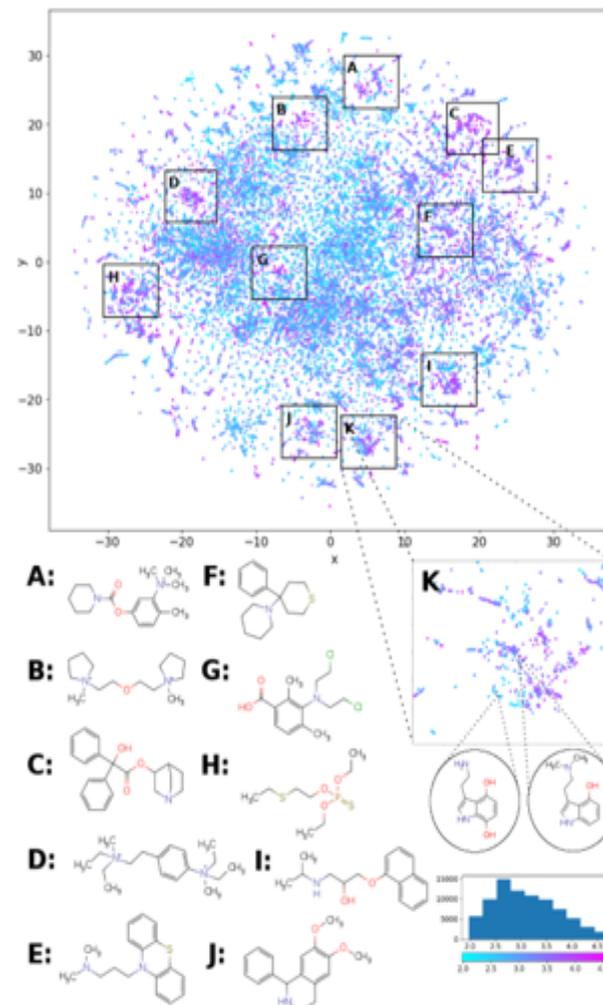
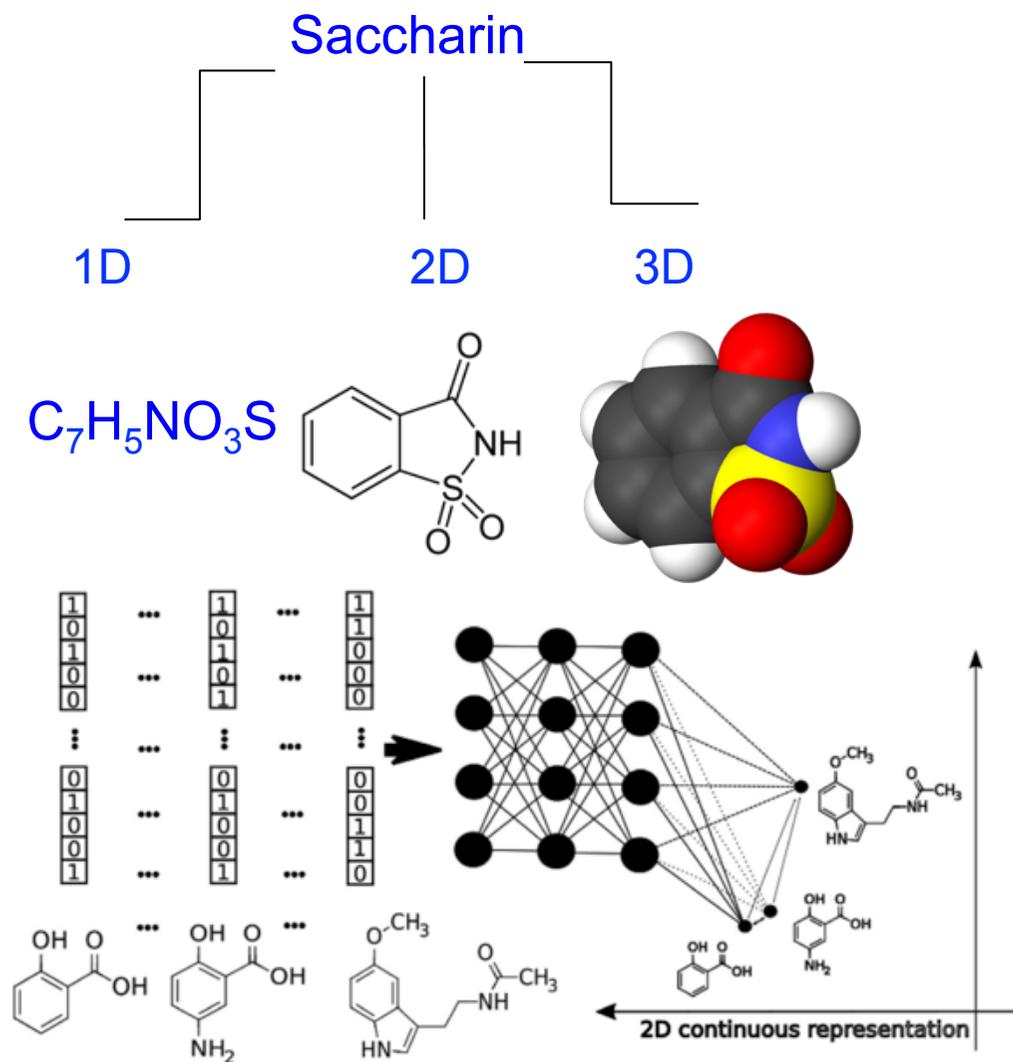


Property: guinea pig_intravenous_LD50 measured in -log(M) ([Details..](#))

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



Design of interpretable descriptors

The screenshot shows the 'ToxAlerts: Structural alerts browser' interface. It lists several alerts with their corresponding chemical structures and SMARTS patterns:

- Pyrans (HS)**: Low specificity (HS) pattern matches any chemicals that include depicted heterocyclic moiety (fusion with other rings) are allowed. SMARTS: [OR1][CR1][CR1][CR1]-[CR1]-[CR1]1. Endpoint: Extended Functional Groups (EFG). Molecules 2015, 21 (1).
- Thiopyrans (HS)**: High specificity (HS) pattern matches chemicals that include exact heterocyclic moiety as in the depiction (fusion with other rings) are not allowed. SMARTS: [SR1][CR1][CR1][CR1]-[CR1]-[CR1]1. Endpoint: Extended Functional Groups (EFG). Molecules 2015, 21 (1).
- Aromatic six-membered heterocycles with one heteroatom (HS)**: High specificity (HS) pattern matches chemicals that include exact heterocyclic moiety as in the depiction (fusion with other rings) are not allowed. SMARTS: [a;c:R1][cR1][cR1][cR1][cR1]1. Endpoint: Extended Functional Groups (EFG). Molecules 2015, 21 (1).
- Pyridines (HS)**: High specificity (HS) pattern matches chemicals that include exact heterocyclic moiety as in the depiction (fusion with other rings) are not allowed. SMARTS: [pR1][pR1][cR1][cR1][cR1]1. Endpoint: Extended Functional Groups (EFG). Molecules 2015, 21 (1).

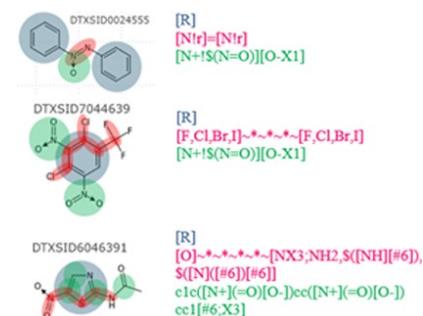
Structural alerts and extended functional groups (EFG) <http://ochem.eu>
Salmina, E.S. et al. *Molecules* **2016**, *21*, 1

Toxprints

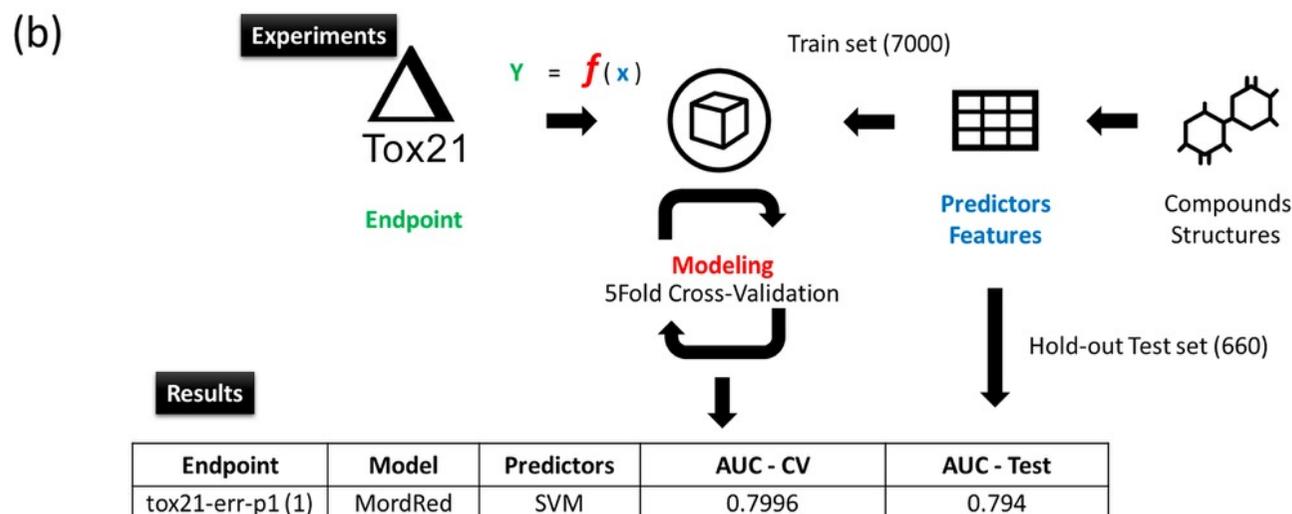
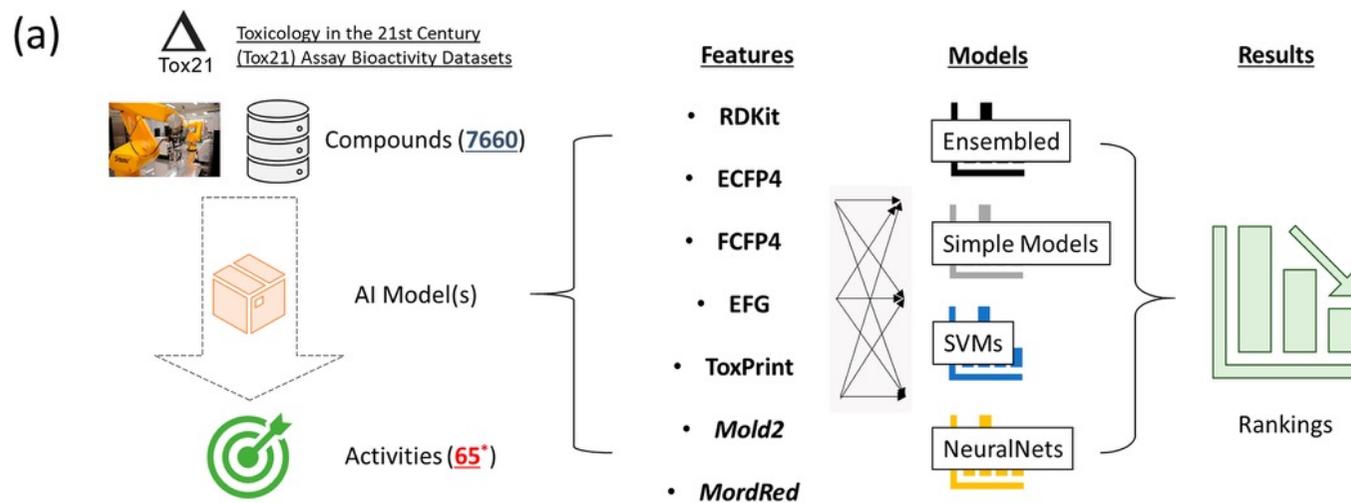
The screenshot shows the ChemoType software interface. It displays a grid of chemical structures (A4, A5, A6, A7, A8, A9, A10, A11) and a list of Toxprint categories on the right, including:

- atom
- bond
- chain
- group
- ring
- aromatic
- fused
- hetero
- polycyclic
- Ashby Toxprint Alerts
- TTC Category (Cancel)

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

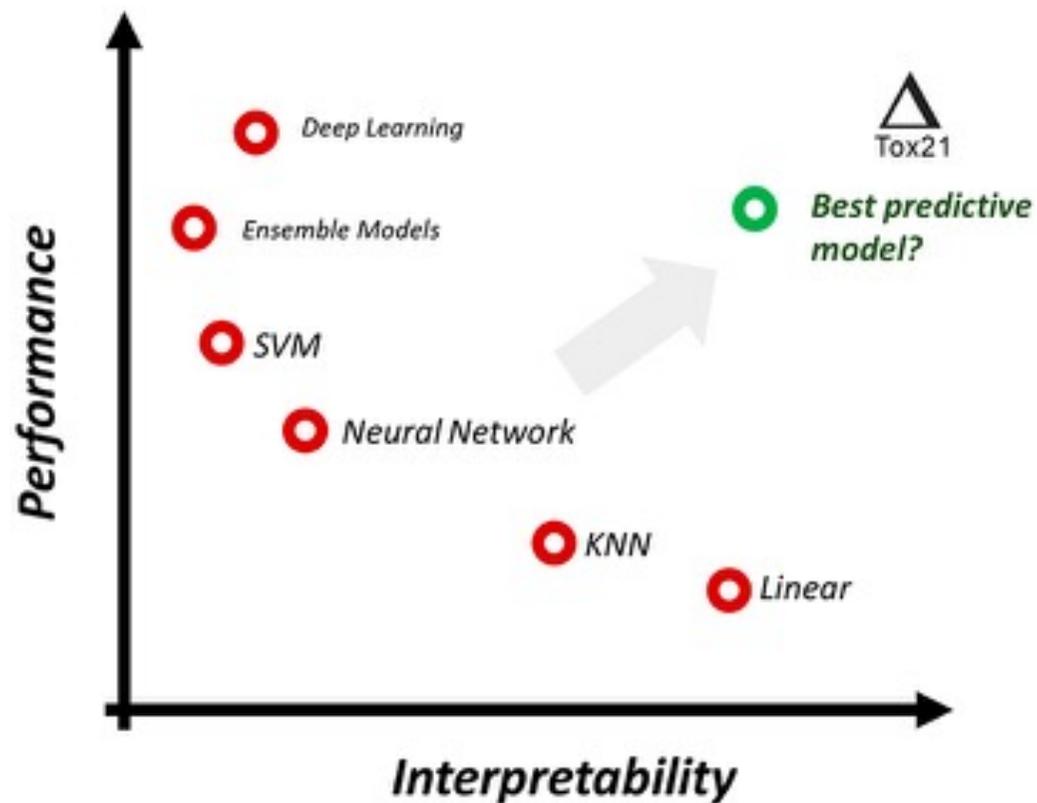


Sedykh, A. et al. *Chem. Res. Toxicol.* **2021**, *34*, 634-640.



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.



Overall best performance
5CV/TEST

RF + all: 0.84/0.84

LS-SVM + MORDRED: 0.87/0.88

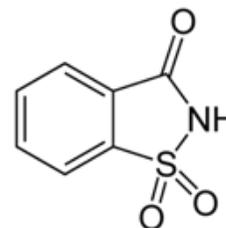
RF + EFG: 0.81/0.86

All + MOLD2: 0.82/0.84

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

Machine Learning directly from chemical structures

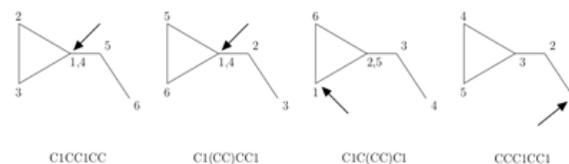
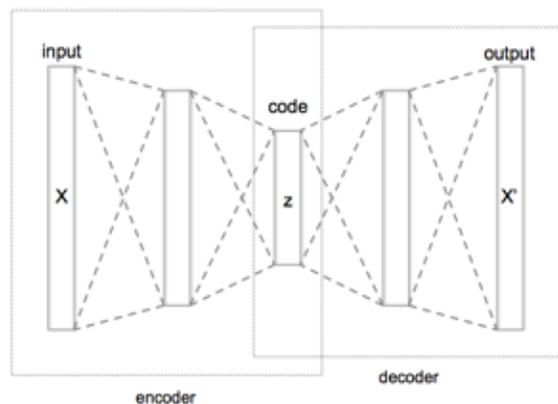
Saccharin: c1ccc2c(c1)C(=O)NS2(=O)=O



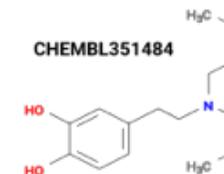
Text processing: convolutional neural networks, Transformers, LSTM

Graph processing: message passing neural networks

Auto-encoder:

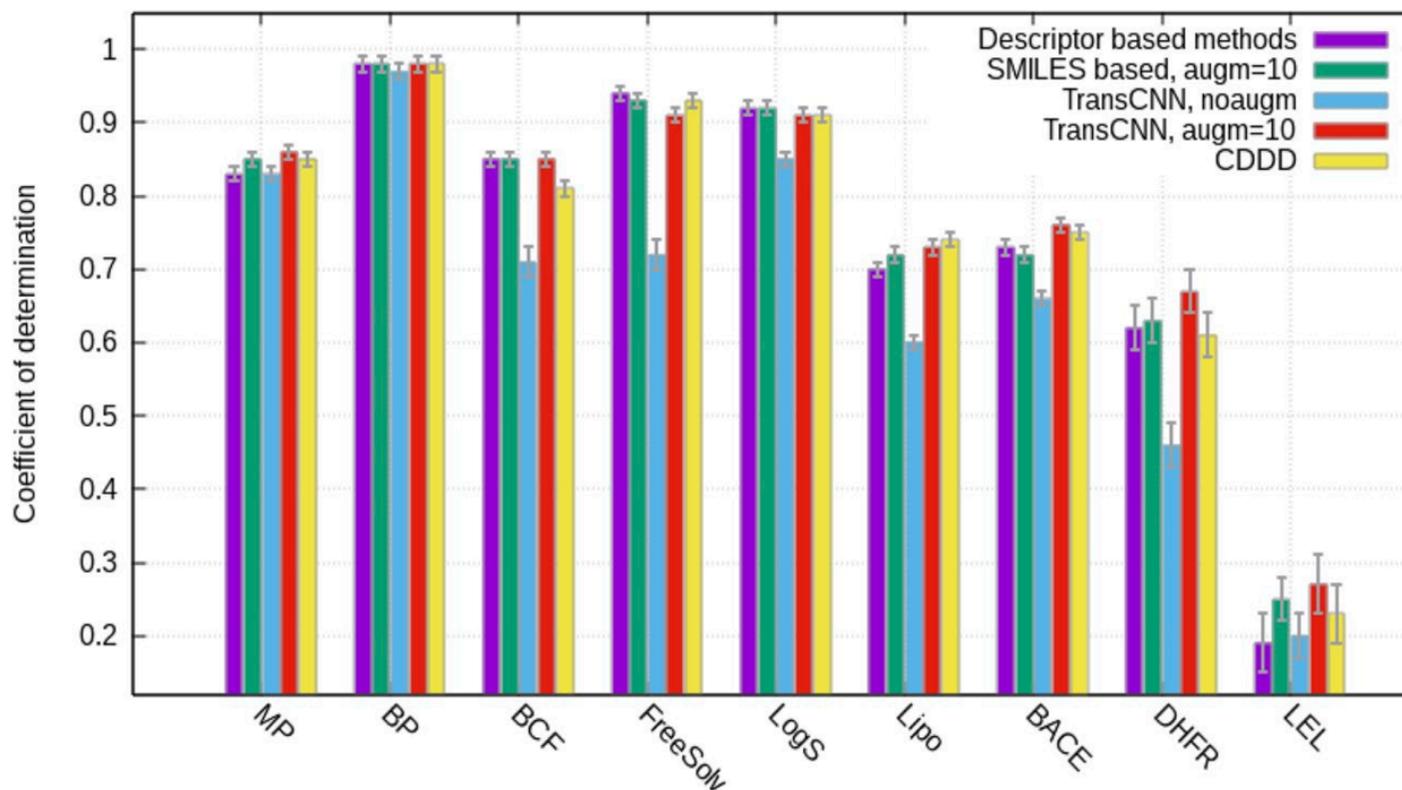


non-canonical	canonical
<chem>c1c(ccc(O)c1O)CCN(CCCC)CCC</chem>	<chem>CCCCN(Cc1ccc(O)c1)O)CCC</chem>
<chem>e1(ccc(O)c1)O)CCN(CCC)CCCC</chem>	<chem>CCCCN(Cc1ccc(O)c1)O)CCC</chem>
<chem>e1c(cc(O)c1)CCN(CCCC)CCC)O</chem>	<chem>CCCCN(Cc1ccc(O)c1)O)CCC</chem>
<chem>e1(CCN(CCCC)CCC)ccc(O)c1</chem>	<chem>CCCCN(Cc1ccc(O)c1)O)CCC</chem>
<chem>CCCCN(Cc1ccc(O)c1)O)CCC</chem>	<chem>CCCCN(Cc1ccc(O)c1)O)CCC</chem>
<chem>CCCCN(Cc1ccc(O)c1)O)CCC</chem>	<chem>CCCCN(Cc1ccc(O)c1)O)CCC</chem>
<chem>N(CCCC)(CCc1ccc(O)c1)O)CCC</chem>	<chem>CCCCN(Cc1ccc(O)c1)O)CCC</chem>
<chem>C(N(Cc1ccc(O)c1)O)CCC)CC</chem>	<chem>CCCCN(Cc1ccc(O)c1)O)CCC</chem>
<chem>N(Cc1ccc(O)c1)O)(CCC)CCC</chem>	<chem>CCCCN(Cc1ccc(O)c1)O)CCC</chem>
<chem>e1c(O)c(ccc1CCN(CCC)CCCC)O</chem>	<chem>CCCCN(Cc1ccc(O)c1)O)CCC</chem>
<chem>e1(c(cc1)CCN(CCC)CCCC)O</chem>	<chem>CCCCN(Cc1ccc(O)c1)O)CCC</chem>



SMILES canonization by machine learning
 → transfer learning to new data

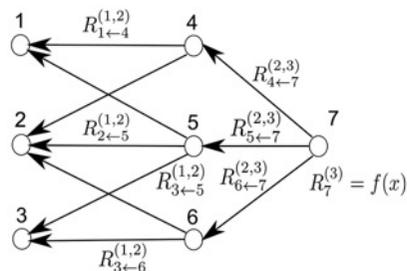
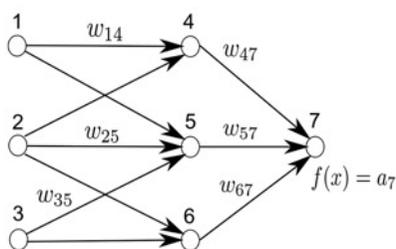
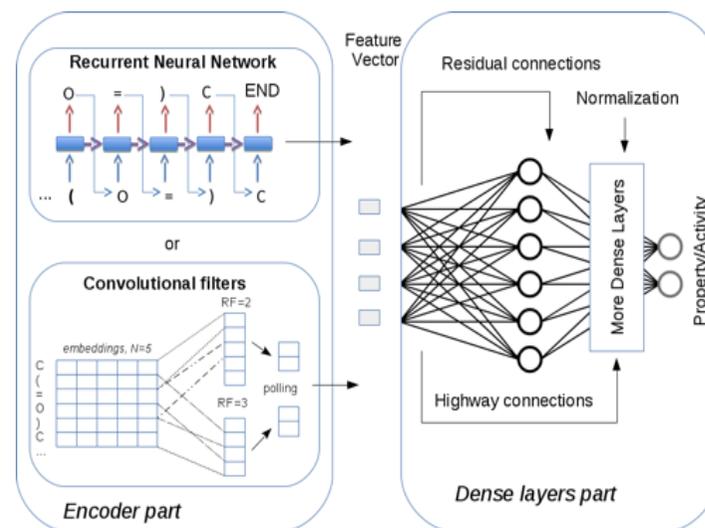
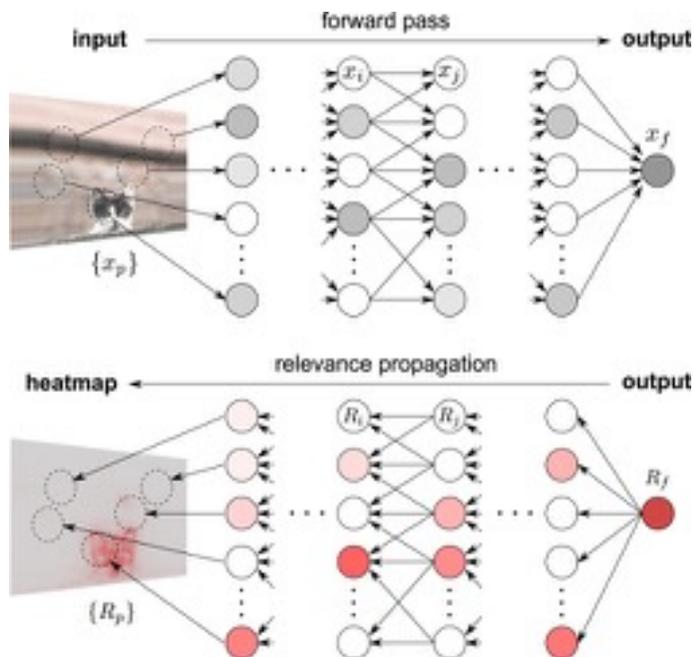
Convolutional vs. Descriptor-based Neural Neural Networks



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)



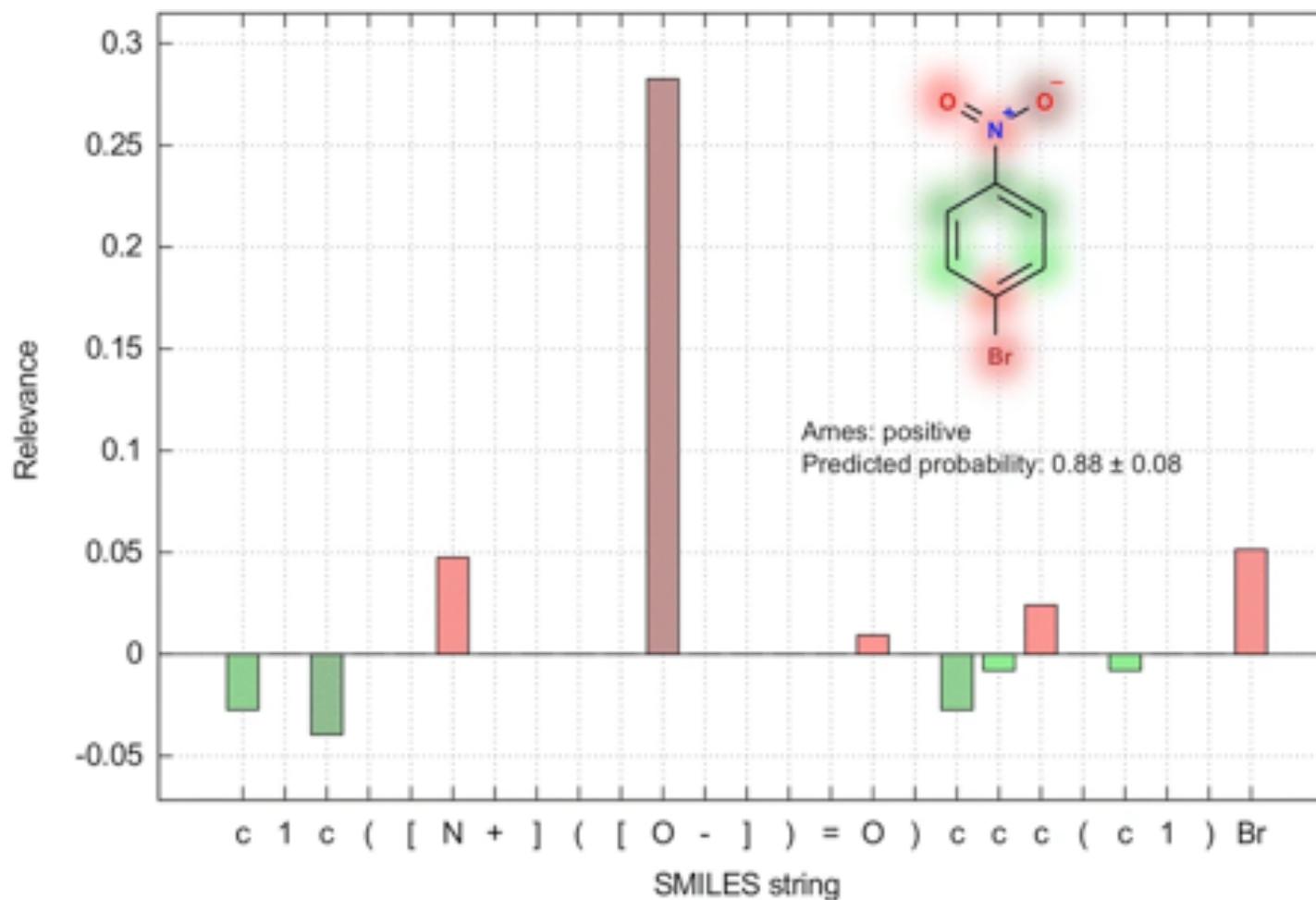
$$f(x) \approx f(x_0) + Df(x_0)[x - x_0]$$

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

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

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

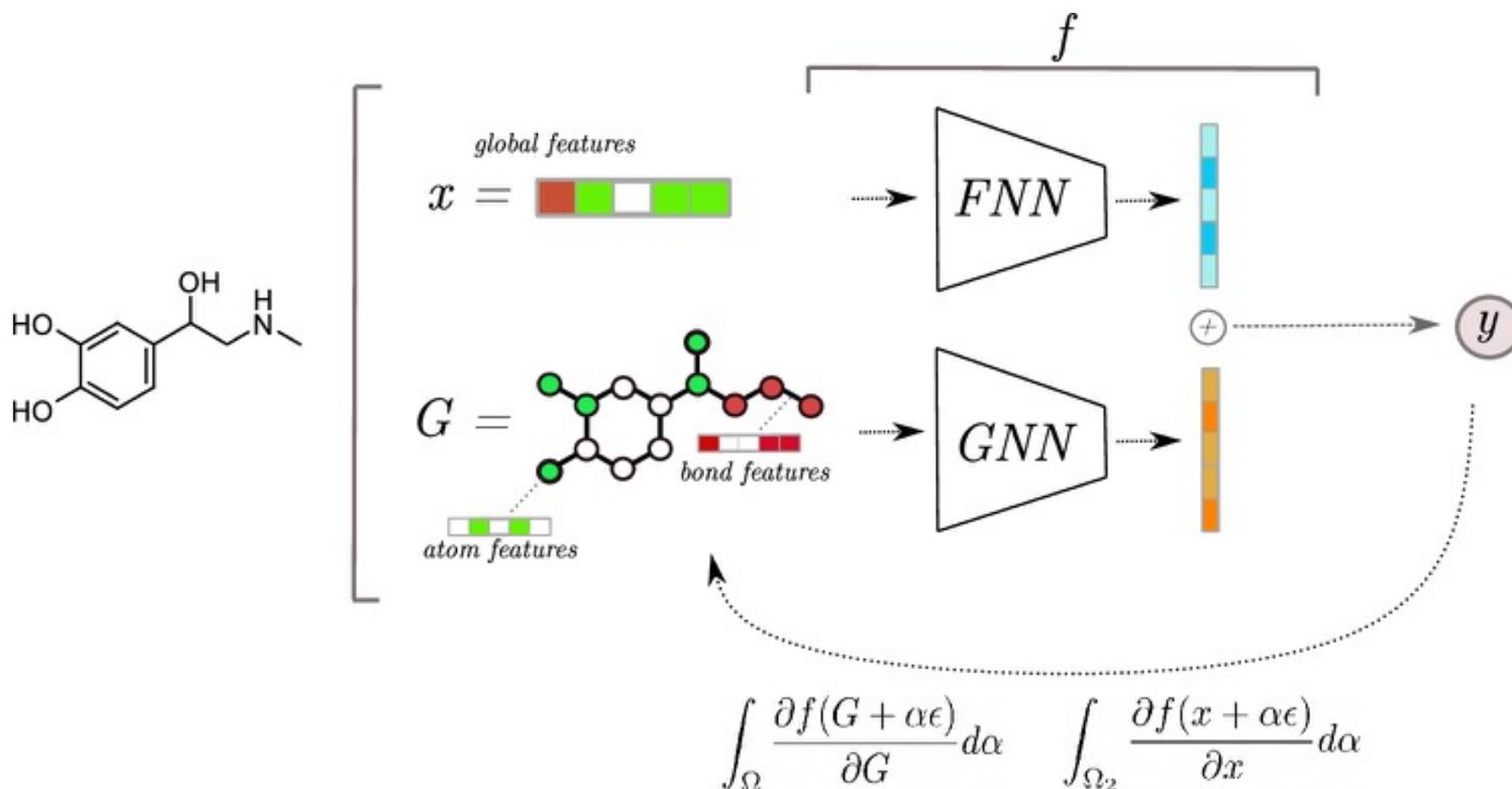
Interpretation of models for Transformer-CNN



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

06.04.2021

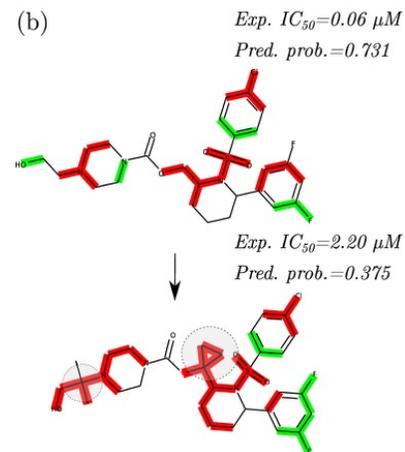
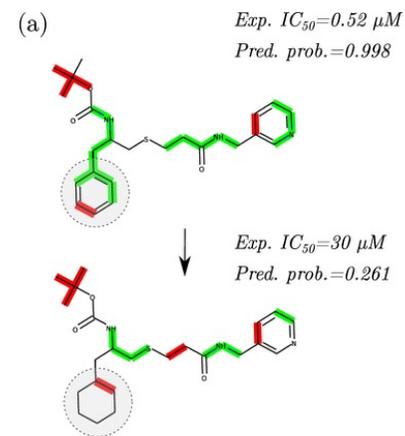
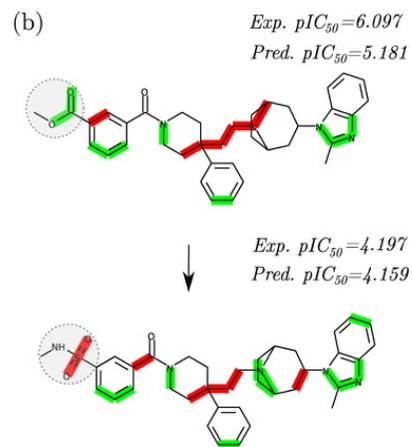
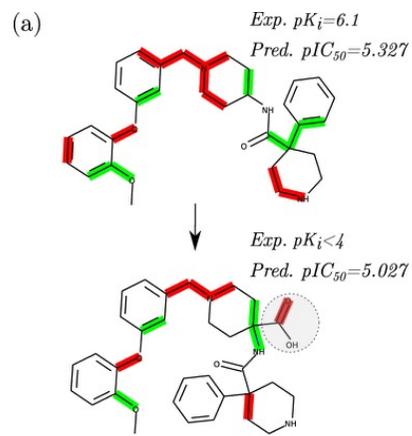
<https://github.com/bigchem/transformer-cnn>



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.

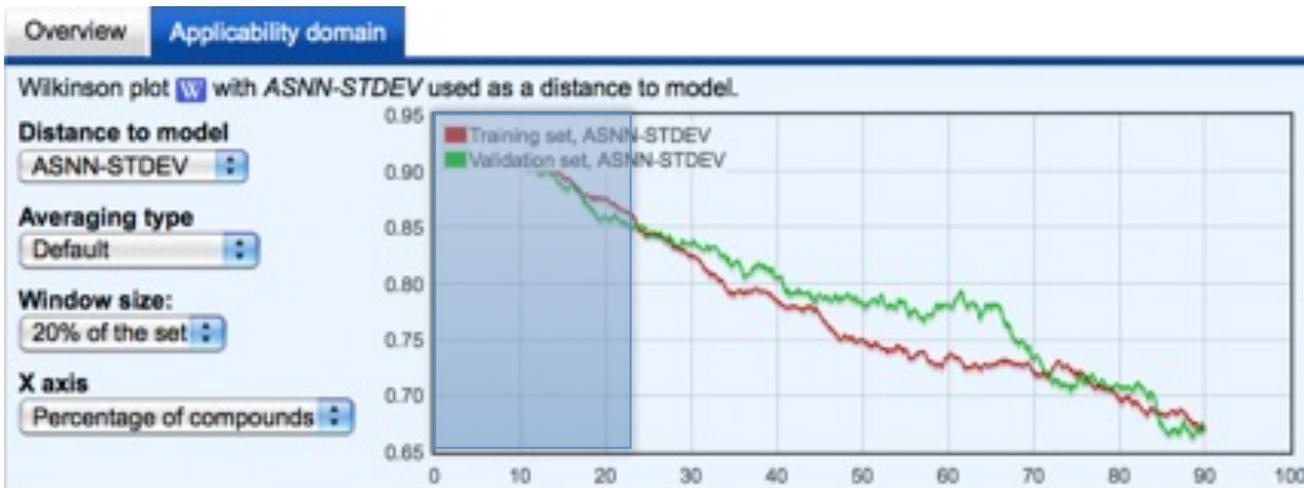
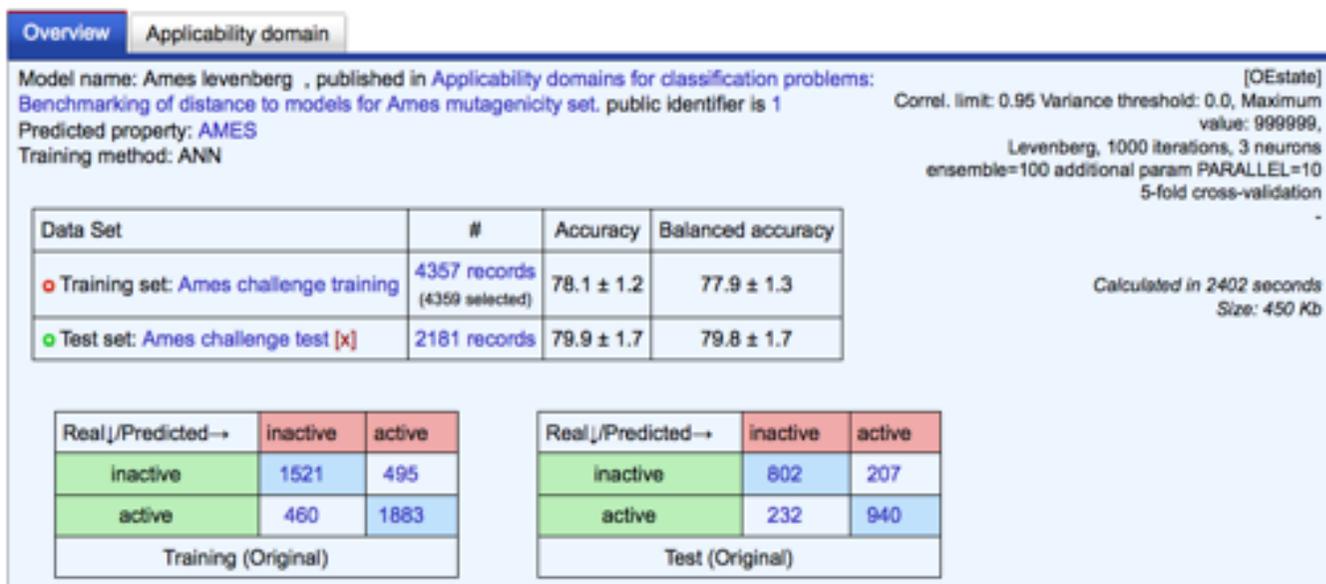
06.04.2021



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.

Accuracy of predictions for classification model



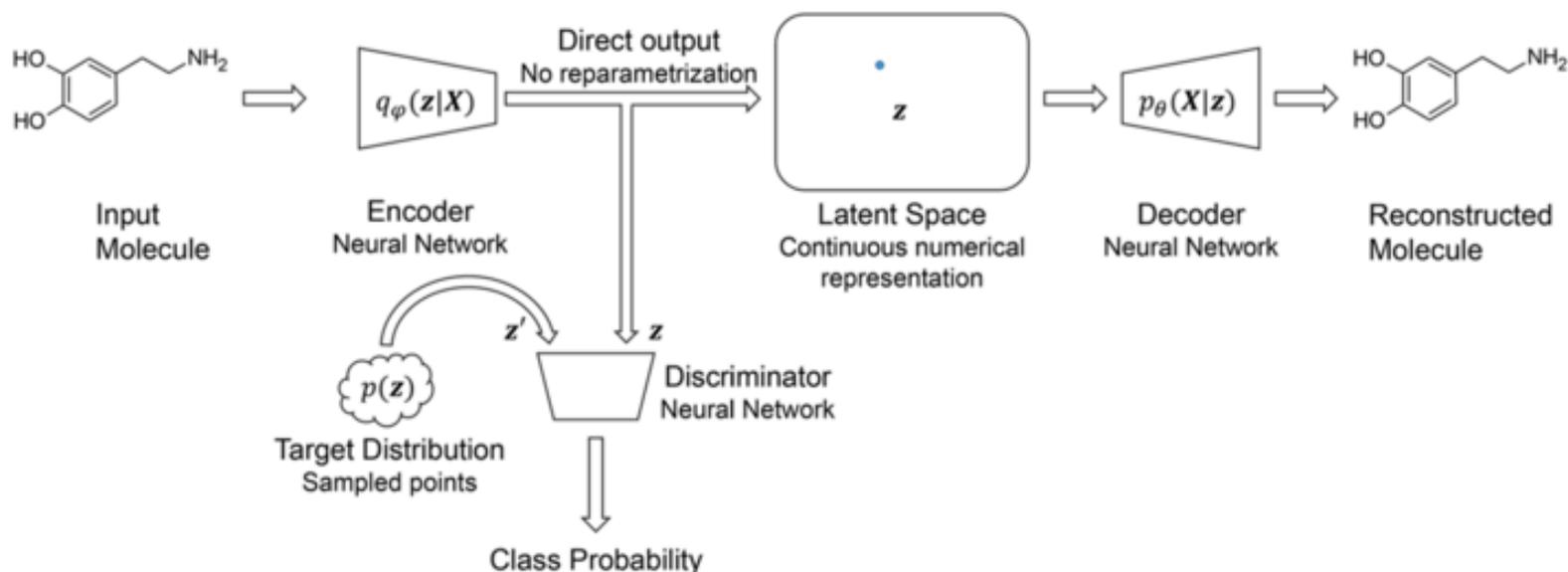
Application of Generative Autoencoder in de Novo Molecular Design

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



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 matured 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.**”**

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)

