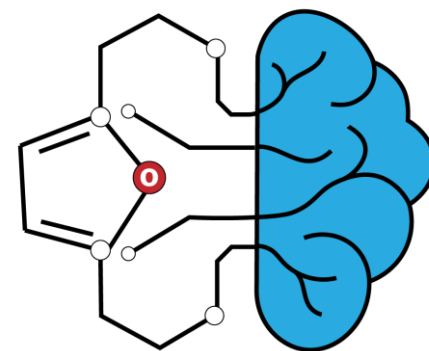
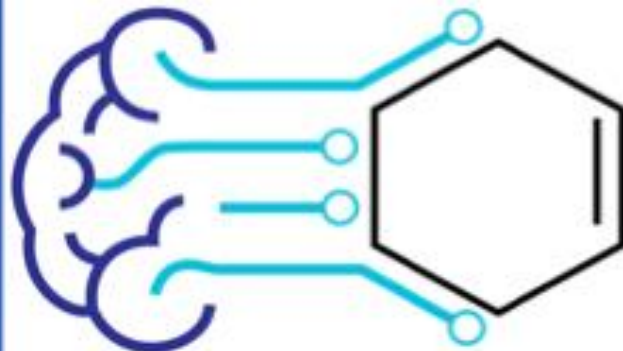




AiChemist

Developing and interpretation of models using OCHEM

19th June 2026

5th AiChemist School on Chemoinformatics

Igor V. Tetko
Institute of Structural Biology, HMGU

HELMHOLTZ MUNICH

Data storage and model development: <https://ochem.eu>

**Online chemical database**
with modeling environment

Home ▾ Database ▾ Models ▾ Tox24 Challenge ▾ EUOS25 Challenge ▾

log in create account

A+ a- Privacy statement

v.4.3.202

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

Announcements

Apply for a

Check out the properties available on OCHEM

OCHEM contains 5493507 records for 706 properties (with at least 50 records) collected from 23972 sources

Melting Point **logPow** **logBB** LogL(water) **LogD**
Kpu(tissue) 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**
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)**
IC50 CYP450 Inhibition **Ki CYP450** **logK' hsa** **Dissipation half-life DT50**
Freundlich coefficient Kf **BMF** **Atmospheric OH Rate Constant** **Ki** TDL0 **LD**
LDLo **Carcinogen** **Anti-inflammatory activity** **LogLD50** **MIC**
Retention Time **Surface tension** Critical micelle concentration
Cblood/Cair(Human) Cfat/Cair(Rat) Cbrain/Cair(Rat) Cliver/Cair(Rat) Cmuscle/Cair(Rat)

Latest active users

 jator: Dr. Jennifer Ator
seconds ago

 Citra: Dr. mario citra
12 minutes ago

 rajnemala: Dr. Appalaraju Nemala
about 4 hours ago

 huiminxuemilly@163.c: Miss. huimin xue
about 6 hours ago

 mutagenic: Dr. Manuel Martin
about 6 hours ago

 ladislav.janovec: Dr. Ladislav Janovec
about 7 hours ago

Latest published models

 logPow model published by tcirino
2 months ago

 Transmittance model published by itetko_euos
4 months ago

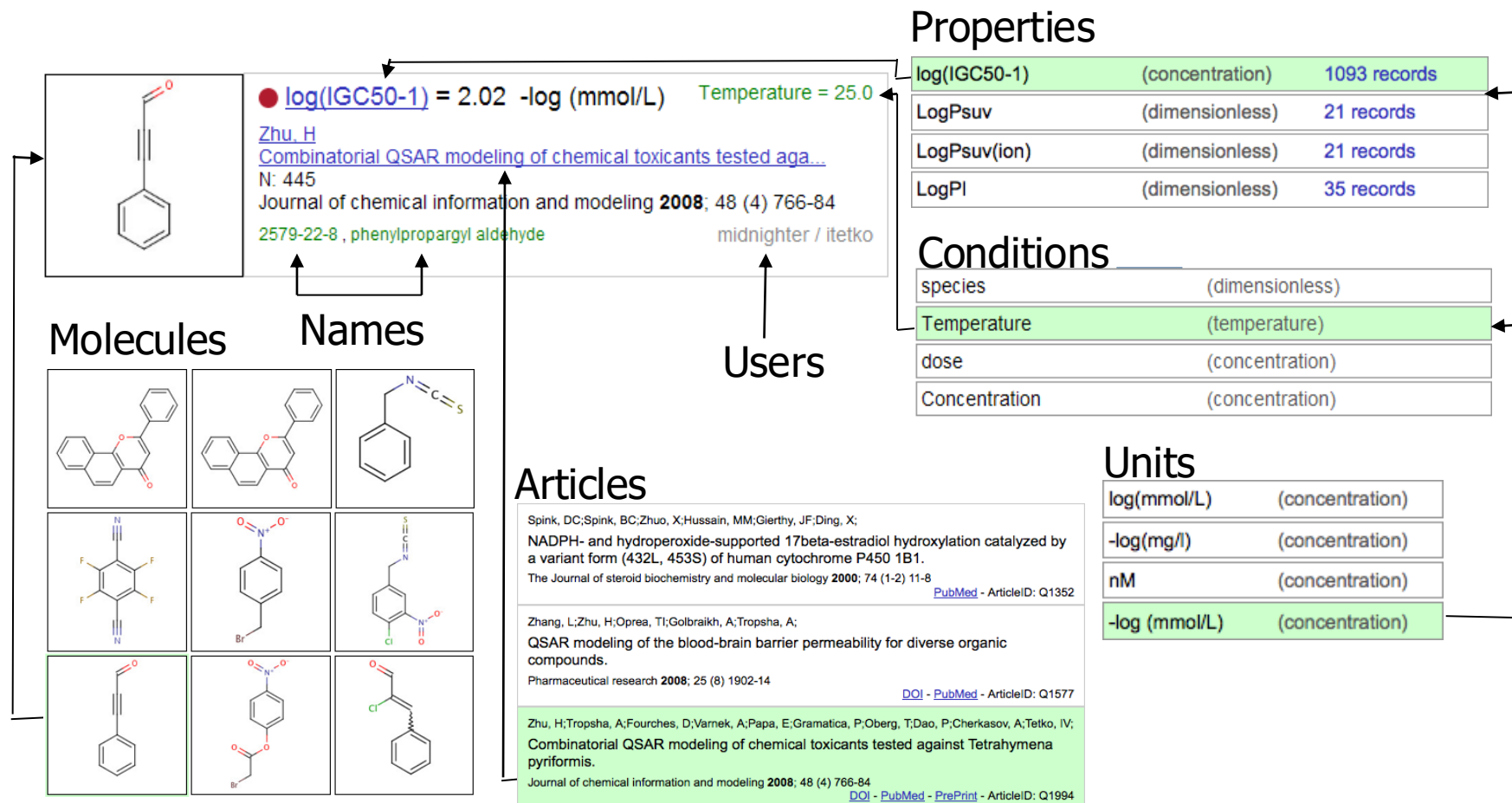
 Transmittance (qualitative) model published by lucic
4 months ago

 Growth inhibition model published by vkovalishyn
9 months ago

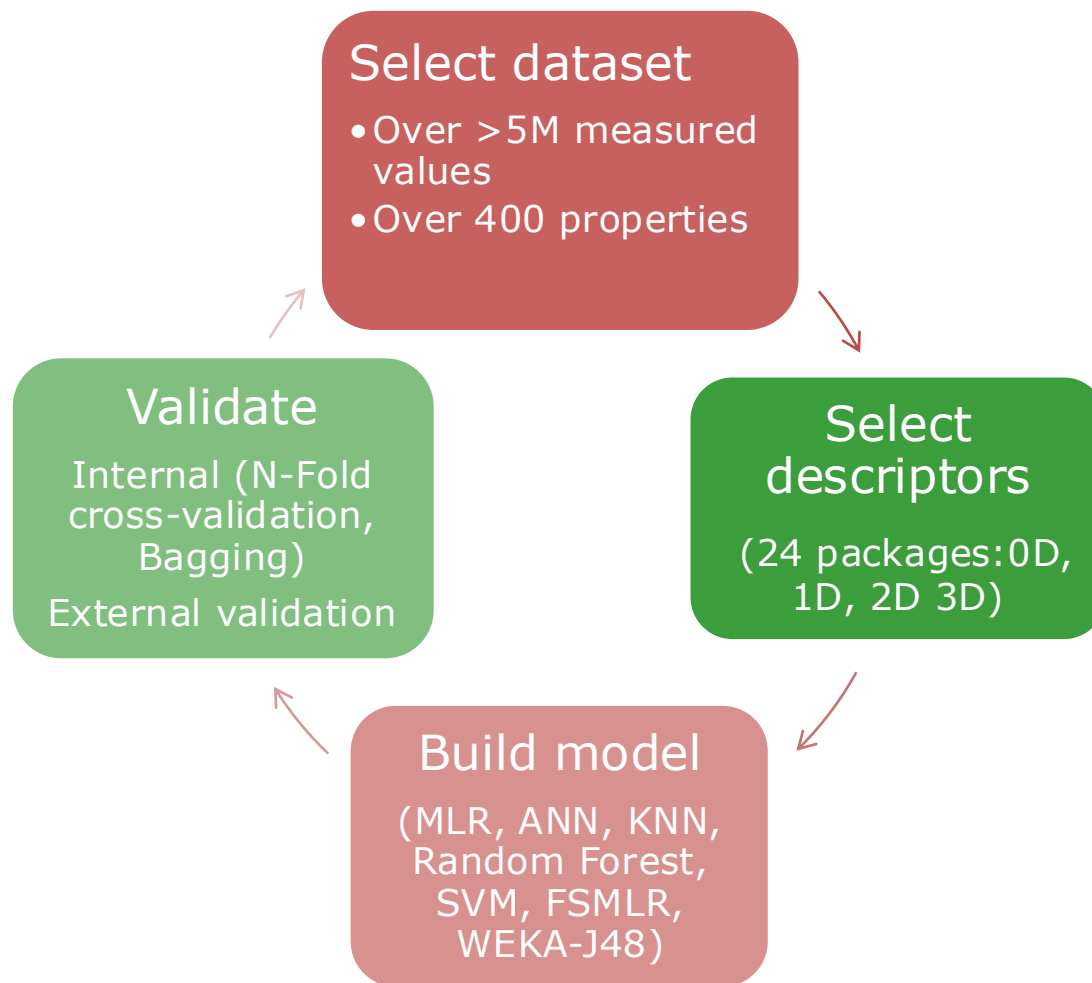
 TTR binding activity model published by itetko_acs
about a year ago

 heart block model published by ggbond
about a year ago

OCHEM Database schema



Modeling iterative workflow



QSPR/QSAR modelling in OCHEM

Select the molecular descriptors

Recommended descriptor types (2D)

- ☒ OEState
 - ☒ Bonds Indices
 - ☐ Counts only
- ☒ ALogPS (2)
 - ☐ Mold2 (777)
 - ☐ CDDD
 - ☐ JPlagP
 - ☐ SIRMS
 - ☐ ISIDA fragments
 - ☐ The in Hashed Atom Pair fingerprint (MAP4)
 - ☐ GSFragment (1138)
 - ☐ QNPR
 - ☐ Multilevel Neighborhoods of Atoms (MNA)
 - ☐ Structural alerts (ToxAlerts and Funtional 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

Predictions by OCHEM's featured models

- ☐ 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)

Create a model

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

Select the training and validation sets:

Training set (required): [peptidesregr \[details\]](#)
[Add a validation set](#)

The model will predict this property:

LogD using unit: Log unit

☐ Skip model configuration and use the predefined settings

Choose the learning method:

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: N-Fold cross-validation

Number of folds: 5

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

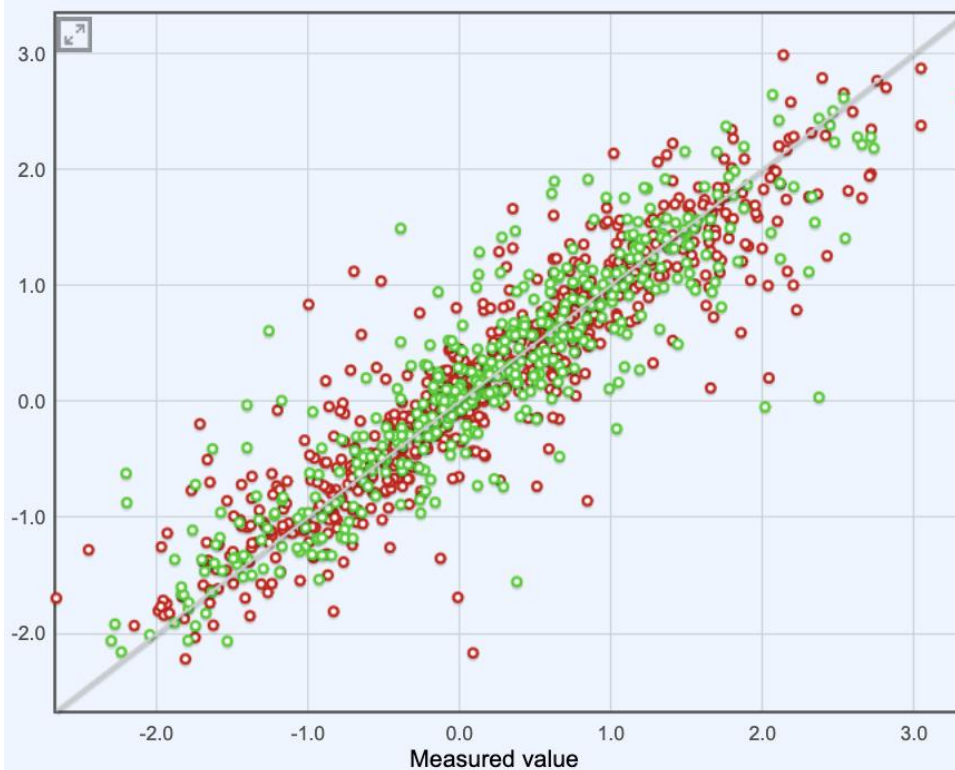
You can create a model from template: [import an XML model template](#) or [use another model as a template](#)

Confidence intervals

Predicted property: **log(IGC50-1)** modeled in -log(mmol/L)

Training method: ASNN

Data Set	#	R2	q2	RMSE	MAE
● Training set: <i>T. pyriformis</i> (training)	644 records	0.83 ± 0.02	0.83 ± 0.02	0.44 ± 0.02	0.31 ± 0.01
● Test set: <i>T. pyriformis</i> (test) [x]	449 records	0.81 ± 0.02	0.8 ± 0.02	0.46 ± 0.03	0.32 ± 0.02





Overfitting

Overfitting



We use cookies to enhance your experience on our website. By continuing to use our website, you are agreeing to our use of cookies
cookie settings at any time. [Continue](#) [Find out more](#)

Overfitting is incorrect reporting of a model performance which overestimates for data from the same distribution as used for model development.

overfitting    

NOUN

Statistics

The production of an analysis which corresponds too closely or exactly to a particular set of data, and may therefore fail to fit additional data or predict future observations reliably.

Origin

1930s; earliest use found in Quarterly Review of Biology. From over- + fitting.

Pronunciation 

overfitting /,əʊvə'fɪtɪŋ/

<https://en.oxforddictionaries.com/definition/overfitting>

Examples of overfitting by neural networks

J. Chem. Inf. Comput. Sci. **1995**, *35*, 826–833

Neural Network Studies. 1. Comparison of Overfitting and Overtraining

Igor V. Tetko,[†] David J. Livingstone,^{‡,§} and Alexander I. Luik*,[†]

Biomedical Department, Institute of Bioorganic and Petroleum Chemistry, Murmanskaya, 1, Kiev-660, 253660, Ukraine, ChemQuest, Cheyney House, 19-21 Cheyney Street, Steeple Morden, Herts, SG8 0LP, U.K., and Centre for Molecular Design, University of Portsmouth, Portsmouth, Hants, PO1 2EG, U.K.

Received January 27, 1995[®]

Neural Processing Letters **6**: 43–50, 1997.

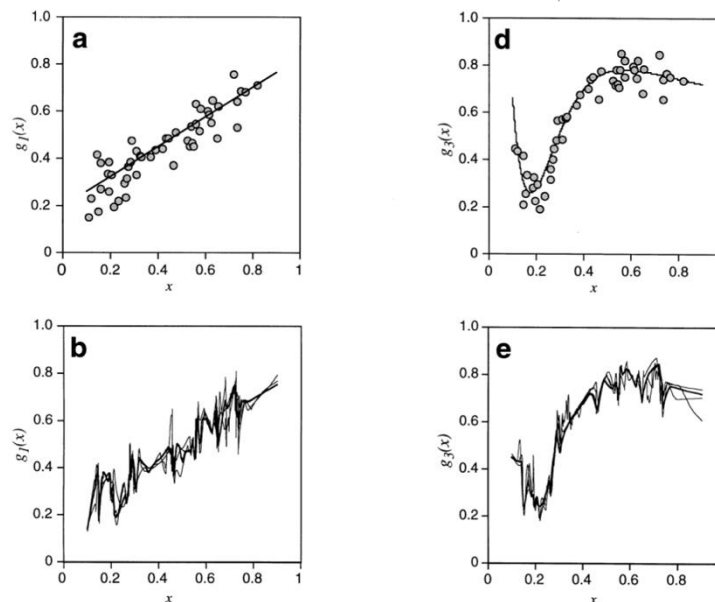
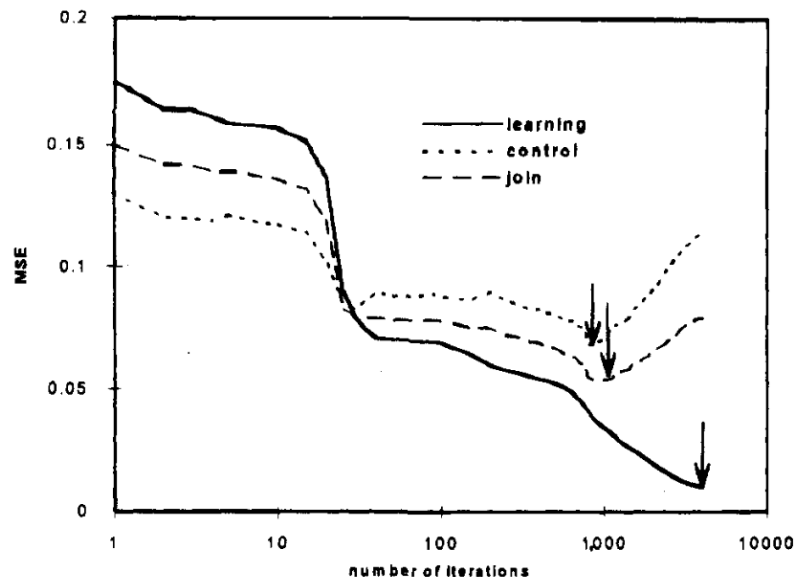
© 1997 Kluwer Academic Publishers. Printed in the Netherlands.

An Enhancement of Generalization Ability in Cascade Correlation Algorithm by Avoidance of Overfitting/Overtraining Problem

IGOR V. TETKO^{1,2} and ALESSANDRO E.P. VILLA¹

¹ Laboratoire de Neuro-heuristique, Institut de Physiologie, Université de Lausanne, Rue du Bugnon 7, Lausanne, CH-1005, Switzerland; ² Department of Biomedical Applications, Institute of Bioorganic and Petroleum Chemistry, Murmanskaya, 1, Kiev-660, 253660, Ukraine
E-mail: Alessandro.Villa@iphysiol.unil.ch

Key words: cascade correlation algorithm, early stopping, overfitting, overtraining



Examples of overfitting by neural networks

Neural Processing Letters **6**: 43–50, 1997.

© 1997 Kluwer Academic Publishers. Printed in the Netherlands.

43

An Enhancement of Generalization Ability in Cascade Correlation Algorithm by Avoiding Overfitting/Overtraining Problem

IGOR V. TETKO^{1,2} and ALESSANDRO E.P. VILLA¹

¹ Laboratoire de Neuro-heuristique, Institut de Physiologie, Université de Lausanne, Rue du Bugnon 7, Lausanne, CH-1005, Switzerland; ² Department of Biomedical Applications, Institute of Bioorganic and Petroleum Chemistry, Murmanskaya, 1, Kiev-660, 253660, Ukraine
E-mail: Alessandro.Villa@iphysiol.unil.ch

Key words: cascade correlation algorithm, early stopping, overfitting, overtraining

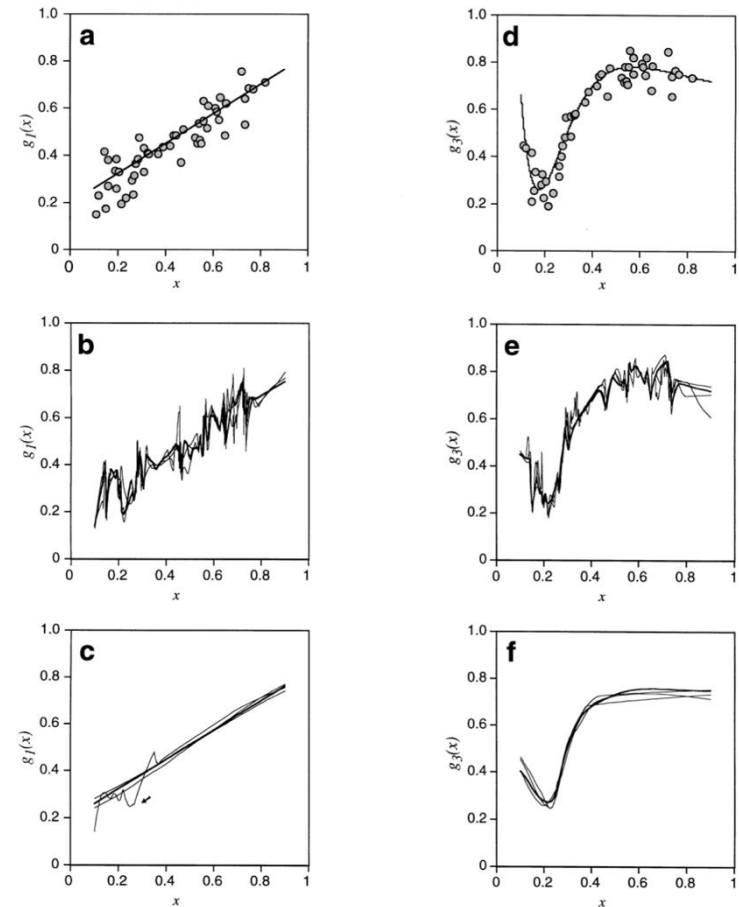


Figure 1. Training Data sets and Calculated Curves

Panels [a-c] refer to function $g_1(x)$ and panels [d-f] refer to function $g_3(x)$.

[a], [d]: circles represent 50 cases of the training data set and the thick line represent the function $g(x)$ over the interval [0.1, 0.9].

[b], [e]: three examples of overtrained ANNs (thin lines) and an ensemble average of 100 such ANNs (thick lines);

[c], [f]: three examples of early stopping ANNs (thin lines) and an ensemble average of 100 such ANNs (thick lines).

Overfitting by variable selection: Simulated Dataset

N = 74 molecules (Boiling point of alkenes)

100 random variables

Software WEKA:

k-nearest neighbors

MLR (Multiple Linear Rgression)

Cross-validation strategies

- A. Descriptors (+ model hyperparameters) are selected using the whole set (training + validation set); CV is done after selection
- B. Descriptors and model are selected when estimating errors on the validation set; CV results are reported for the model optimized on the respective validation fold
- C. Validation set is kept hidden, descriptors (and model) are selected using respective training set

Selection of descriptors and hyperparameter optimization of model can be done using any available method.

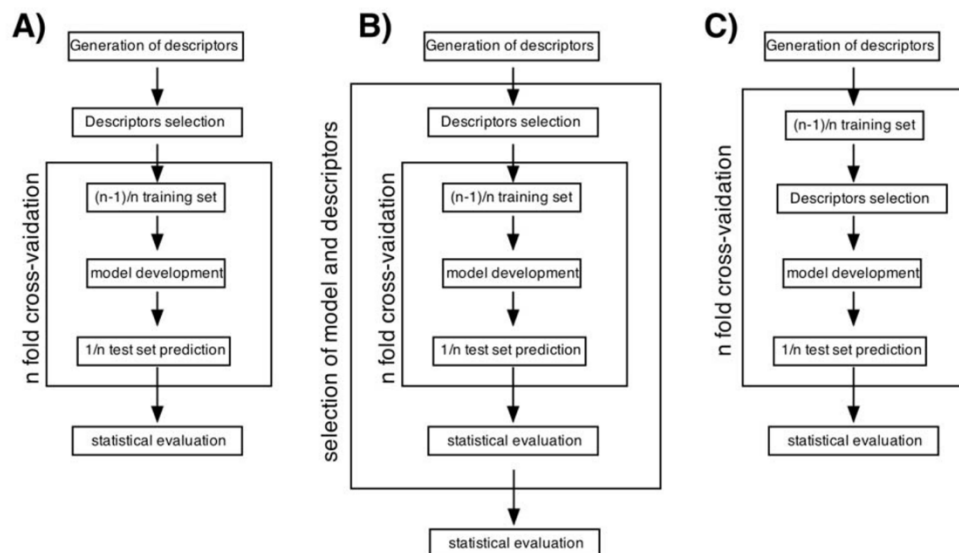


Figure 1. Internal (A, B) and External (C) cross-validation procedures.

Comparison of validation performances

Strategy A: correlation coefficient, $R = 0.82$

$RMSE = 27^{\circ} C$

Strategy B: $R = 0.55$

$RMSE = 39^{\circ} C$

Strategy C: $R = -0.28$

$RMSE = 54^{\circ} C \sim \text{random}$

Strategy C – double or nested cross-validation;
recommended!

The use of incorrect validation (strategy 1-2) contributes
overfitting!

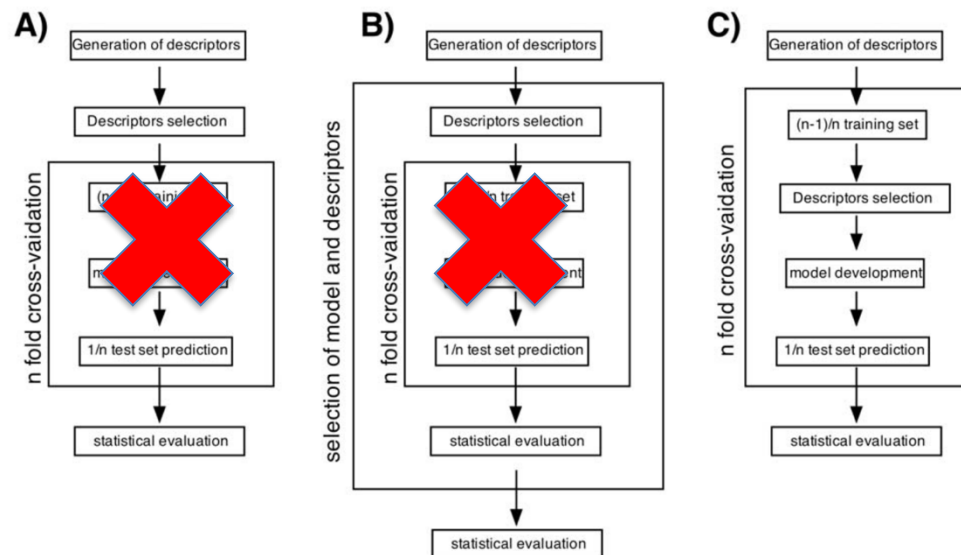
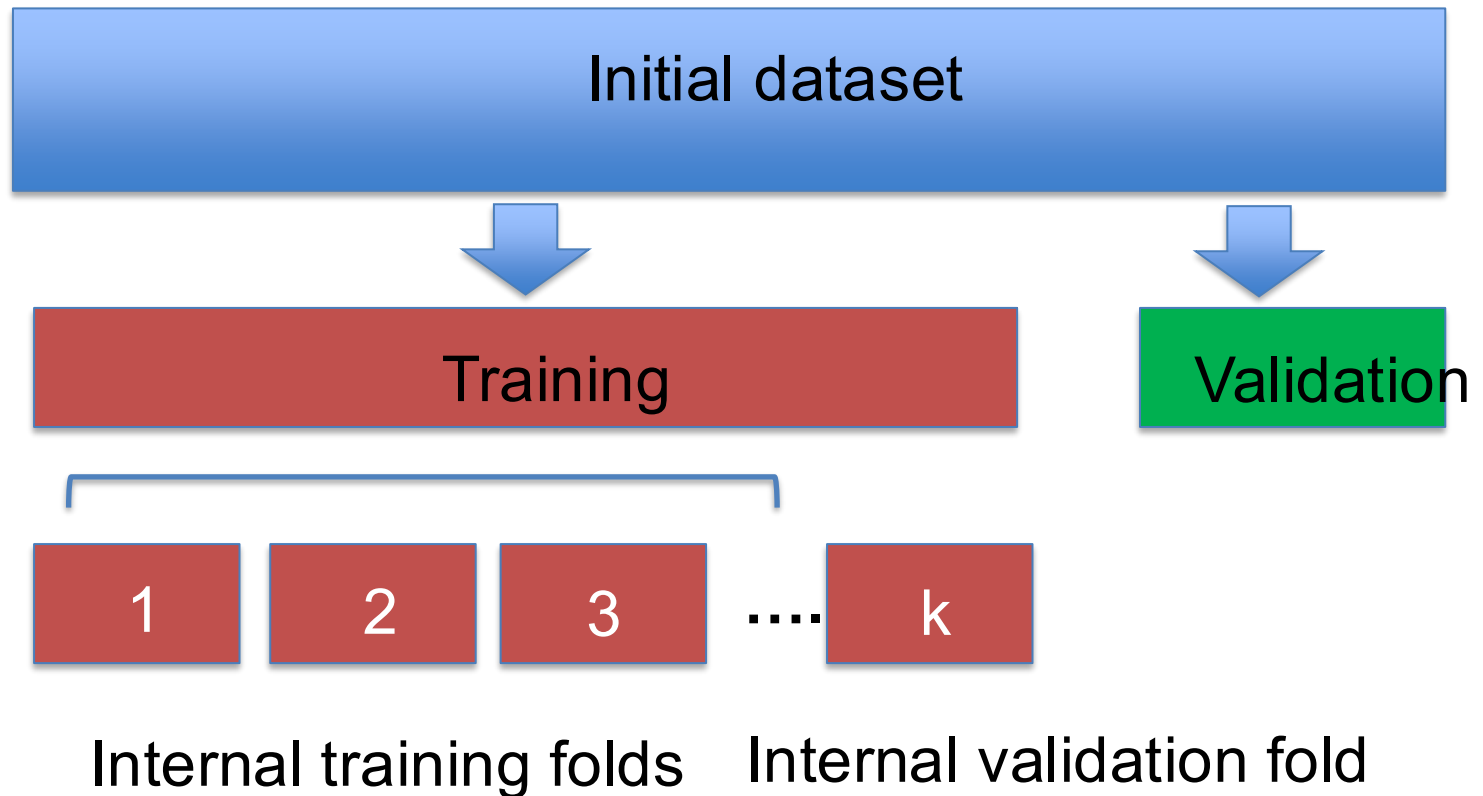


Figure 1. Internal (A, B) and External (C) cross-validation procedures.

Strategy C: Double cross-validation



But... is it always required?

scientific data



OPEN

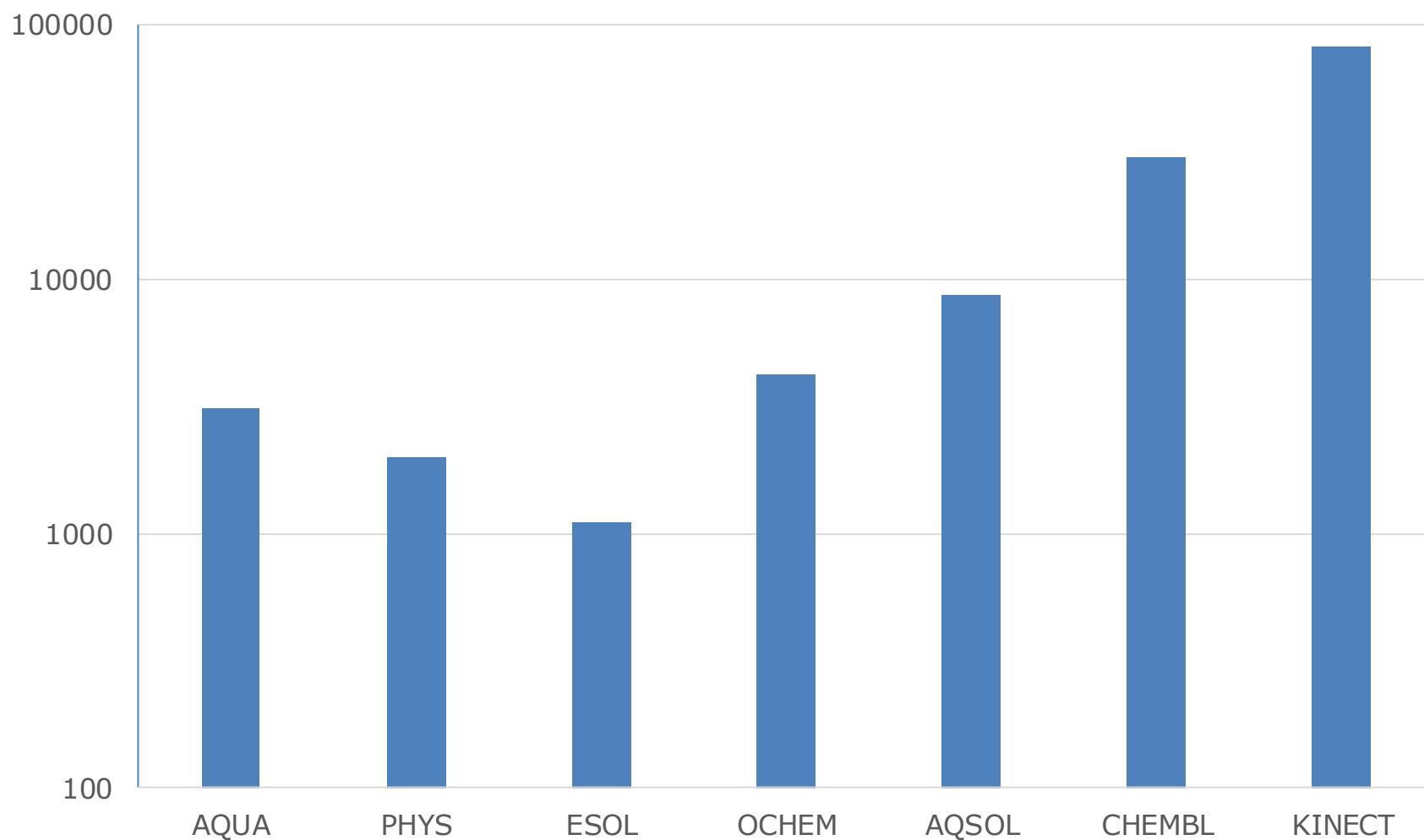
ANALYSIS

Boosting the predictive performance with aqueous solubility dataset curation

Jintao Meng^{1,5,6} , Peng Chen^{2,3} , Mohamed Wahib^{2,3}, Mingjun Yang⁴, Liangzhen Zheng¹, Yanjie Wei¹ , Shengzhong Feng⁵  & Wei Liu⁶

Intrinsic solubility is a critical property in pharmaceutical industry that impacts *in-vivo* bioavailability of small molecule drugs. However, solubility prediction with Artificial Intelligence(AI) are facing insufficient data, poor data quality, and no unified measurements for AI and physics-based approaches. We collect 7 aqueous solubility datasets, and present a dataset curation workflow. Evaluating the curated data with two expanded deep learning methods, improved RMSE scores on all curated thermodynamic datasets are observed. We also compare expanded *Chemprom* enhanced with curated data and state-of-art physics-based approach using pearson and spearman correlation coefficients. A similar performance on pearson with 0.930 and spearman with 0.947 from expanded *Chemprom* is achieved. A steadily improved pearson and spearman values with increasing data points are also illustrated. Besides that, the computation advantage of AI models enables quick evaluation of a large set of molecules during the hit identification or lead optimization stages, which helps further decision making within the time cycle at drug discovery stage.

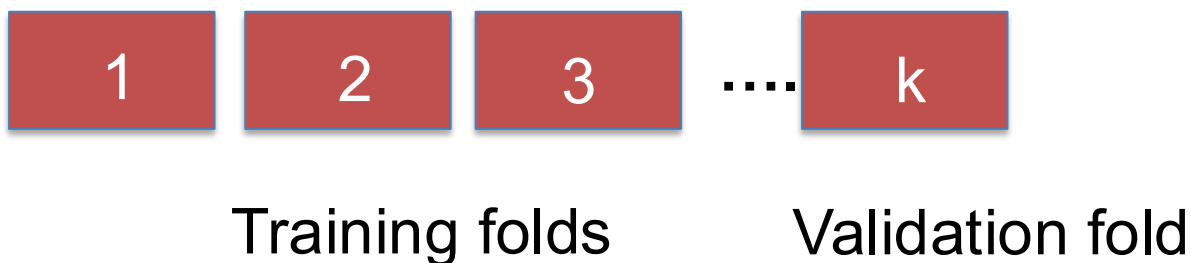
Data setup



Modelling approach

Grid hyperparameter optimization on each step – 108 combinations of parameters; the lowest RMSE is stored.

- Hidden neurons size, depth, dropout, number of feed-forward layers.
- Hyperparameter optimization script sees all the data given to it.
- Use of ensemble of 5 models to improve the model accuracy.
- Use 8 repetitions to obtain confidence intervals.
- Two weeks using 1200 compute nodes (38,200 cores and 4800 GPU accelerators)
- <https://github.com/Mengjintao/chemprop>

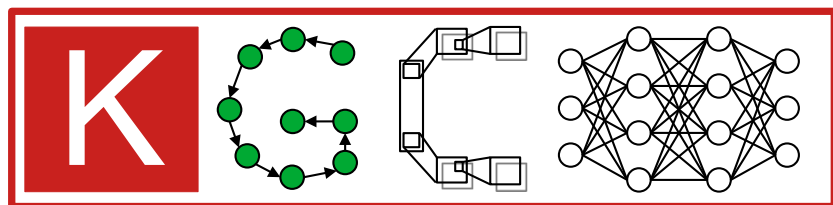


openOCHEM GitHub

Implementation:

<https://github.com/openochem/openochem>

<https://solub.ochem.eu>



KG CNN

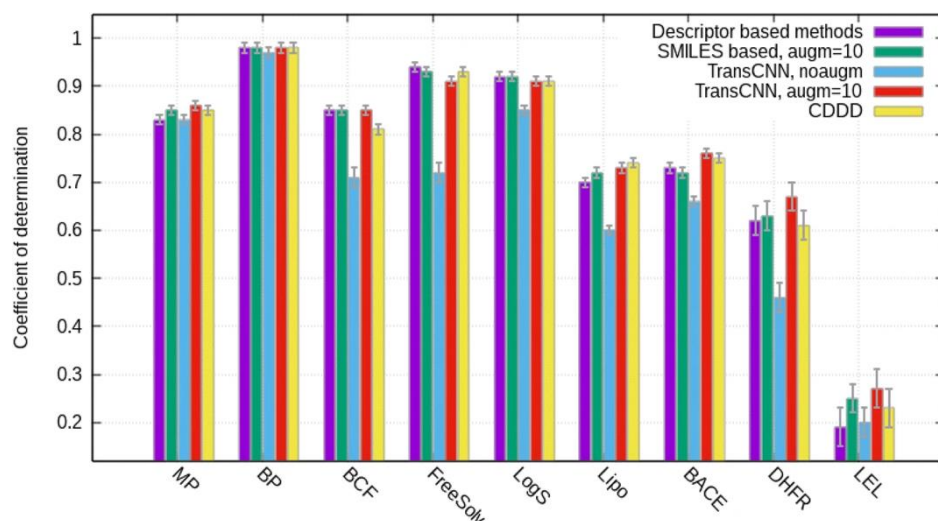
https://github.com/aimat-lab/gcnn_keras

Reiser, P.; Eberhard, A.; Friederich, P. Graph Neural Networks in TensorFlow-Keras with RaggedTensor Representation (Kgcnn). *Softw. Impacts* **2021**, 9, 100095.

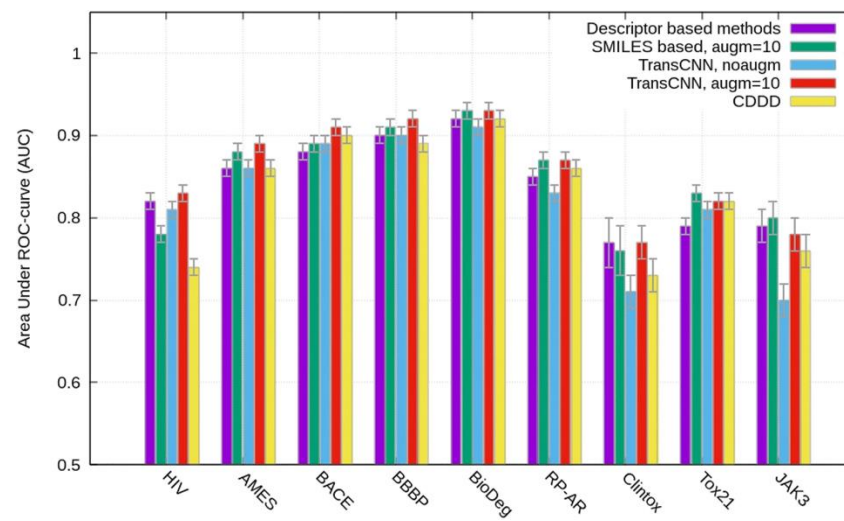
<https://doi.org/10.1016/j.simpa.2021.100095>.

Hypeparameter pre-optimisation

Regression tasks



Classification tasks



Tetko, I. V.; Karpov, P.; Bruno, E.; Kimber, T. B.; Godin, G. Augmentation Is What You Need! In *Artificial Neural Networks and Machine Learning – ICANN 2019: Workshop and Special Sessions*; Tetko, I. V., Kůrková, V., Karpov, P., Theis, F., Eds.; Lecture Notes in Computer Science; Springer International Publishing: Cham, 2019; pp 831–835. https://doi.org/10.1007/978-3-030-30493-5_79.

Karpov, P.; Godin, G.; Tetko, I. V. Transformer-CNN: Swiss Knife for QSAR Modeling and Interpretation. *J. Cheminformatics* 2020, 12 (1), 17. <https://doi.org/10.1186/s13321-020-00423-w>.

Comparison of Root Mean Squared Errors (RMSE) based on tuned and pre-set hyper-parameters („Org“ sets)

Dataset	Published by Meng et al (following hyperparameters optimisation)		Results from this study (using default hyperparameters of methods)		
			RMSE, validation by molecule		
	ChemProp	AttFP	ChemProp	AttFP	Transformer CNN
AQUA	0.58 ± 0.04	0.62 ± 0.03	0.58 ± 0.02	0.60 ± 0.02	$0.56 \pm 0.02^*$
PHYSP	0.55 ± 0.03	0.65 ± 0.02	0.57 ± 0.01	0.59 ± 0.01	$0.56 \pm 0.01^*$
ESOL	0.60 ± 0.08	0.64 ± 0.02	0.61 ± 0.01	0.59 ± 0.02	$0.58 \pm 0.02^*$
OCHEM	0.55 ± 0.02	0.60 ± 0.01	0.63 ± 0.01	0.65 ± 0.01	$0.59 \pm 0.01^{**}$
AQSOL	1.02 ± 0.04	0.83 ± 0.03	$1. \pm 0.02$	1.01 ± 0.02	$0.99 \pm 0.01^*$
CHEMBL	0.92 ± 0.02	n/a	0.88 ± 0.01	0.93 ± 0.01	$0.86 \pm 0.01^{**}$
KINECT	0.401 ± 0.001	n/a	0.434 ± 0.002	0.465 ± 0.001	$0.41 \pm 0.002^{**}$

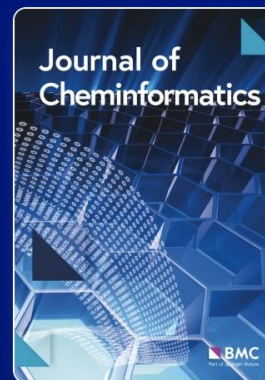
Pre-set hyperparameters provided better performance in 18 of 34 pairwise comparisons (and ca 10000x faster!)

[Home](#) > [Journal of Cheminformatics](#) > [Article](#)

Be aware of overfitting by hyperparameter optimization!

Research | [Open access](#) | Published: 09 December 2024

Volume 16, article number 139 (2024) [Cite this article](#)



[Journal of Cheminformatics](#)

[Aims and scope](#) →

[Submit manuscript](#) →

✓ You have full access to this [open access](#) article

Download PDF ↓

🔖 [Save article](#)

[Igor V. Tetko](#) ✉, [Ruud van Deursen](#) & [Guillaume Godin](#) ✉

👁 4010 Accesses 📄 43 Citations 📈 4 Altmetric [Explore all metrics](#) →

Part of a collection:
[AI in Drug Discovery](#)

Sections

Figures

References



Using power of OCHEM tools

Comprehensive modelling



The comprehensive modeling feature allows you to simultaneously run multiple models with different machine learning methods, molecular descriptors and validation protocols. Please note that running multiple models may require significant computational resources and time.

Select the training and validation sets:

Training set (required): [T. pyriformis \(training\)](#) [\[details\]](#)

Validation set #1: [T. pyriformis \(test\)](#) [\[x\]](#) [\[details\]](#)

[Add a validation set](#)

The model will predict this property:

[log\(IC50-1\)](#) using unit: -log(mol/L)

Select the methods you want to use for the modeling:

Method

[\[all\]](#) [\[none\]](#)

- ☐ FSMRLR (to cache descriptors)
- ☐ ASNN (MTL, bias correction)
- ☐ KNN
- ☐ MLRA
- ☐ PLS
- ☐ WEKA-RF (classification only)
- ☐ WEKA-J48 (classification only)
- ☐ LSSVMG (MTL, GPU)
- ☒ DNN (MTL, GPU)
- ☐ XGBOOST
- ☐ RFR
- ☐ DeepChem TexCNN (no descr, MTL, GPU)
- ☒ TRANSNNI 10/10 (no descr, MTL, GPU)
- ☒ Transformer CNF2 (no descr, MTL, GPU)
- ☐ TRANSNN [\[edit\]](#) [\[x\]](#)
- ☒ KGCNN AttFP (no descr, MTL, GPU)
- ☒ KGCNN ChemProp (no descr, MTL, GPU)
- ☐ Schnet [\[edit\]](#) [\[x\]](#)
- ☐ PAINN [\[edit\]](#) [\[x\]](#)
- ☐ GraphSAGE [\[edit\]](#) [\[x\]](#)
- ☐ GINE [\[edit\]](#) [\[x\]](#)
- ☐ GIN [\[edit\]](#) [\[x\]](#)

[+add a custom template](#)

Descriptors

[\[all\]](#) [\[none\]](#)

- ☐ alvaDesc (3D)
- ☐ CDK 2.7.1 (3D)
- ☐ Dragon v.7 (3D)
- ☒ ALogPS + OEstate
- ☐ ALogPS + StructuralAlerts (EFG)
- ☐ Inductive Descriptors (3D)
- ☐ QNPR (SMILES - length 1 - 3)
- ☐ ISIDA Fragments
- ☐ OSMORDRED
- ☐ RDKit (selected, 3D)
- ☐ JLogP
- ☐ MORDRED (3D)
- ☒ MOLD2
- ☐ ECFP4
- ☒ no desc. (Transformer, CGN, etc.)
- ☐ MW + # of carbons: (baseline model)
- ☐ EPA T.E.S.T. (2D)
- ☐ alvaDesc 2D [\[edit\]](#) [\[x\]](#)
- ☐ CDDD [\[edit\]](#) [\[x\]](#)
- ☐ GSFRAG [\[edit\]](#) [\[x\]](#)
- ☐ MNA [\[edit\]](#) [\[x\]](#)
- ☐ MORDRED 2D [\[edit\]](#) [\[x\]](#)
- ☐ RDKit 2D [\[edit\]](#) [\[x\]](#)
- ☐ SIRMES [\[edit\]](#) [\[x\]](#)
- ☐ OESTATE [\[edit\]](#) [\[x\]](#)
- ☐ ALogPS, StructuralAlerts [\[edit\]](#) [\[x\]](#)

[+add a custom template](#)

Descriptor selection

[\[all\]](#) [\[none\]](#)

- ☐ Unsupervised forward selection
- ☒ Pairwise de-correlation ($R < 0.95$)

[+add a custom template](#)

Model validation

[\[all\]](#) [\[none\]](#)

- ☒ 5-fold cross-validation
- ☐ 5-fold cross-validation (stratified - classification only)
- ☐ Bagging with 64 models
- ☐ Bagging with 64 models (stratified - classification only)

[+add a custom template](#)

[Show advanced options>>](#)

Considering the selection **15 models** are requested to be created. Only several models (depending on the number of users) will be processed simultaneously to decrease server load. Run several models with c

Create the models

Comprehensive modelling

**Comprehensive modeling**
Create multiple models simultaneously

The comprehensive modeling feature allows you to simultaneously run multiple models with different machine learning methods, molecular descriptors and validation protocols. Please note that running multiple models may require significant computational resources and time.

Select the training and validation sets:

Training set (required): [T. pyriformis \(training\)](#) [\[details\]](#)
Validation set #1: [T. pyriformis \(test\)](#) [\[x\]](#) [\[details\]](#)
[Add a validation set](#)

The model will predict this property:
[log\(IC50-1\)](#) using unit: [-log\(mol/L\)](#)

Select the methods you want to use for the modeling:

Method
[\[all\]](#) [\[none\]](#)
☐ FSMLR (to cache descriptors)
☐ ASNN (MTL, bias correction)
☐ KNN
☐ MLRA
☐ PLS
☐ WEKA-RF (classification only)
☐ WEKA-J48 (classification only)
☐ LSSVMG (MTL, GPU)
☒ DNN (MTL, GPU)
☐ XGBOOST
☐ RFR
☒ DeepChem TexCNN (no descr, MTL, GPU)
☒ TRANSNNI 10/10 (no descr, MTL, GPU)
☒ Transformer CNF2 (no descr, MTL, GPU)
☒ TRANSNN [\[edit\]](#) [\[x\]](#)
☒ KGCNN AttFP (no descr, MTL, GPU)
☒ KGCNN ChemProp (no descr, MTL, GPU)
☐ Schnet [\[edit\]](#) [\[x\]](#)
☐ PAINN [\[edit\]](#) [\[x\]](#)
☐ GraphSAGE [\[edit\]](#) [\[x\]](#)
☐ GINE [\[edit\]](#) [\[x\]](#)
☐ GIN [\[edit\]](#) [\[x\]](#)

[+add a custom template](#)

Descriptors
[\[all\]](#) [\[none\]](#)
☒ alvaDesc (3D)
☒ CDK 2.7.1 (3D)
☒ Dragon v.7 (3D)
☒ ALogPS + OEstate
☒ ALogPS + StructuralAlerts (EFG)
☒ Inductive Descriptors (3D)
☒ QNPR (SMILES - length 1 - 3)
☒ ISIDA Fragments
☒ OSMORDRED
☒ RDKit (selected, 3D)
☒ JPLogP
☒ MORDRED (3D)
☒ MOLD2
☒ ECFP4
☒ no desc. (Transformer, CGN, etc.)
☒ MW + # of carbons: (baseline model)
☒ EPA T.E.S.T. (2D)
☒ alvaDesc 2D [\[edit\]](#) [\[x\]](#)
☒ CDDD [\[edit\]](#) [\[x\]](#)
☒ GSFRAG [\[edit\]](#) [\[x\]](#)
☒ MNA [\[edit\]](#) [\[x\]](#)
☒ MORDRED 2D [\[edit\]](#) [\[x\]](#)
☒ RDKit 2D [\[edit\]](#) [\[x\]](#)
☒ SIRMS [\[edit\]](#) [\[x\]](#)
☒ OESTATE [\[edit\]](#) [\[x\]](#)
☒ ALogPS, StructuralAlerts [\[edit\]](#) [\[x\]](#)

[+add a custom template](#)

Descriptor selection
[\[all\]](#) [\[none\]](#)
☐ Unsupervised forward selection
☒ Pairwise de-correlation ($R < 0.95$)

[+add a custom template](#)

Model validation
[\[all\]](#) [\[none\]](#)
☒ 5-fold cross-validation
☐ 5-fold cross-validation (stratified - classification only)
☐ Bagging with 64 models
☐ Bagging with 64 models (stratified - classification only)

Please wait

Starting model 4 out of 31

[Cancel](#)

[Show advanced options>>](#)

Considering the selection **182 models** are requested to be created. Only several models (depending on the number of users) will be processed simultaneously to decrease server load. Run several models with

[Create the models](#)

Comprehensive modelling

Predicted property: [log\(IC50-1\)](#)
 Training set: [T. pyriformis \(training\)](#)
 Duplicate models: [start](#)

Metrics [RMSE - Root Mean Square Error](#) for [Training set](#) Validation: [Cross-Validation \(32\)](#)

	ASNN (CHEMAXON)	DNN	Represent.
ALogPS, OEstate	+	0.53	+
ALogPS, StructuralAlerts	+	0.48	+
CDDD	+	0.5	+
CDK23 (cons,topol,geom,elect,hybr) 3D:rdkit3d	+	0.52	+
Dragon6 (2D blocks: 1 28)	+	0.7	+
Dragon7 (3D blocks: 1-30) 3D:rdkit3d	+	0.51	+
EPA (T.E.S.T.)	+	0.51	+
Fragmentor (length: 2-4)	+	0.48	+
GSFrag (F + L)	+	0.62	+
InductiveDescriptors 3D:rdkit3d	+	0.51	+
JPlogP	+	0.51	+
MNA	+	0.79	+
MOLD2	+	0.52	+
MORDRED (2D)	+	0.53	+
MORDRED (3D) 3D:rdkit3d	+	0.56	+
OEstate	0.44 (+1 models)	0.48	+
OSMORDRED	+	0.51	+
QNPR (length:1 - 3)	+	0.58	+
RDKit (2D blocks: 1-3 5 10-11 15-16)	+	0.6	+
RDKit (3D blocks: 1-11 15-16) 3D:rdkit3d	+	0.56	+
RDKit (ECFP4)	+	0.85	+
SIRMS (labels: charge+logp+hb+refractivity)	+	0.54	+
alvaDesc (2D blocks: 1-12 22-25 29 32-33)	+	0.54	+
alvaDesc (3D blocks: 1-33) 3D:rdkit3d	+	0.54	+
CNF2 (F) (3D) 10/10	+	+	0.46
DEEPCHEM (F) (3D) TEXTCNN 10/10	+	+	0.47
KGCNN AttFP	+	+	0.42
KGCNN ChemProp	+	+	0.43
TRANSNN (F) (3D) 10/10	+	+	0.52
TRANSNNI (F) 10/10	+	+	0.43

Predicted property: [log\(IC50-1\)](#)
 Training set: [T. pyriformis \(training\)](#)
 Duplicate models: [start](#)

Metrics [RMSE - Root Mean Square Error](#) for [Training set](#) Validation: [Cross-Validation \(32\)](#)

	ASNN (CHEMAXON)	DNN	Represent.
ALogPS, OEstate	+	0.53	+
ALogPS, StructuralAlerts	+	0.48	+
CDDD	+	0.5	+
CDK23 (cons,topol,geom,elect,hybr) 3D:rdkit3d	+	0.52	+
Dragon6 (2D blocks: 1 28)	+	0.7	+
Dragon7 (3D blocks: 1-30) 3D:rdkit3d	+	0.51	+
EPA (T.E.S.T.)	+	0.51	+
Fragmentor (length: 2-4)	+	0.48	+
GSFrag (F + L)	+	0.62	+
InductiveDescriptors 3D:rdkit3d	+	0.51	+
JPlogP	+	0.51	+
MNA	+	0.79	+
MOLD2	+	0.52	+
MORDRED (2D)	+	0.53	+
MORDRED (3D) 3D:rdkit3d	+	0.56	+
OEstate	0.44 (+1 models)	0.48	+
OSMORDRED	+	0.51	+
QNPR (length:1 - 3)	+	0.58	+
RDKit (2D blocks: 1-3 5 10-11 15-16)	+	0.6	+
RDKit (3D blocks: 1-11 15-16) 3D:rdkit3d	+	0.56	+
RDKit (ECFP4)	+	0.85	+
SIRMS (labels: charge+logp+hb+refractivity)	+	0.54	+
alvaDesc (2D blocks: 1-12 22-25 29 32-33)	+	0.54	+
alvaDesc (3D blocks: 1-33) 3D:rdkit3d	+	0.54	+
CNF2 (F) (3D) 10/10	+	+	0.46
DEEPCHEM (F) (3D) TEXTCNN 10/10	+	+	0.47
KGCNN AttFP	+	+	0.42
KGCNN ChemProp	+	+	0.43
TRANSNN (F) (3D) 10/10	+	+	0.52
TRANSNNI (F) 10/10	+	+	0.43

Comprehensive modelling

Predicted property: [log\(IC50-1\)](#)
 Training set: [T. pyriformis \(training\)](#)
 Duplicate models: [start](#)

Metrics: RMSE - Root Mean Square Error for Training set Validation: Cross-Validation (41 models)

	ASNN (CHEMAXON)	DNN	Represent.	FSMLR	KNN	MLRA	PLS	LSSVMG	XGBOOST	RFR	SCI:CAT_BOOST	
ALogPS, OEstate	+	0.53	+	+	+	+	+	+	+	+	+	
ALogPS, StructuralAlerts	+	0.48	+	+	+	+	+	+	+	+	+	
CDDD	+	0.5	+	+	+	+	+	+	+	+	+	
CDK23 (cons,topol,geom,elect,hybr) 3D:rdkit3d	+	0.52	+	+	+	+	+	+	+	+	+	
Dragon6 (2D blocks: 1-28)	+	0.7	+	+	+	+	+	+	+	+	+	
Dragon7 (3D blocks: 1-30) 3D:rdkit3d	+	0.51	+	+	+	+	+	+	+	+	+	
EPA (T.E.S.T.)	+	0.51	+	+	+	+	+	+	+	+	+	
Fragmentor (length: 2-4)	running	0.48	+	0.48	0.67	0.59	0.49	0.62	0.51	0.5	0.43	
GSFrag (F + L)	+	0.62	+	+	+	+	+	+	+	+	+	
InductiveDescriptors 3D:rdkit3d	+	0.51	+	+	+	+	+	+	+	+	+	
JPlogP	+	0.51	+	+	+	+	+	+	+	+	+	
MNA	+	0.79	+	+	+	+	+	+	+	+	+	
MOLD2	+	0.52	+	+	+	+	+	+	+	+	+	
MORDRED (2D)	+	0.53	+	+	+	+	+	+	+	+	+	
MORDRED (3D) 3D:rdkit3d	+	0.56	+	+	+	+	+	+	+	+	+	
OEstate	0.44 (+1 models)	0.48	+	+	+	+	+	+	+	+	+	
OSMORDRED	+	0.51	+	+	+	+	+	+	+	+	+	
QNPR (length:1 - 3)	+	0.58	+	+	+	+	+	+	+	+	+	
RDKit (2D blocks: 1-3 5 10-11 15-16)	+	0.6	+	+	+	+	+	+	+	+	+	
RDKit (3D blocks: 1-11 15-16) 3D:rdkit3d	+	0.56	+	+	+	+	+	+	+	+	+	
RDKit (ECFP4)	+	0.85	+	+	+	+	+	+	+	+	+	
SIRMS (labels: charge+logp+hb+refractivity)	+	0.54	+	+	+	+	+	+	+	+	+	
alvaDesc (2D blocks: 1-12 22-25 29 32-33)	+	0.54	+	+	+	+	+	+	+	+	+	
alvaDesc (3D blocks: 1-33) 3D:rdkit3d	+	0.54	+	+	+	+	+	+	+	+	+	
CNF2 (F) (3D) 10/10	+	+	0.46	+	+	+	+	+	+	+	+	
DEEPCHEM (F) (3D) TEXTCNN 10/10	+	+	0.47	+	+	+	+	+	+	+	+	
KGCNN AttFP	+	+	0.42	+	+	+	+	+	+	+	+	
KGCNN ChemProp	+	+	0.43	+	+	+	+	+	+	+	+	
TRANSNN (F) (3D) 10/10	+	+	0.52	+	+	+	+	+	+	+	+	
TRANSNNI (F) 10/10	+	+	0.43	+	+	+	+	+	+	+	+	

Delete 1 matching models
 Create 29 missing models
 Create models for those saved in the first column

Comprehensive modelling

Predicted property: [log\(IC50-1\)](#)

Training set: [T. pyriformis \(training\)](#)

Duplicate models: [start](#)

Metrics: [RMSE - Root Mean Square Error](#) for [Training set](#) Validation: [Cross-Validation \(78 models\)](#)

	ASNN (CHEMAXON)	DNN	Represent.	FSMLR	KNN	MLRA	PLS	LSSVMG	XGBOOST	RFR	SCI:CAT_BOOST
ALogPS, OEstate	+	0.53	+	+	+	+	+	+	+	+	0.42
ALogPS, StructuralAlerts	0.42	0.48	+	0.52	0.6	0.56	0.49	0.56	0.46	0.45	0.43
CDDD	+	0.5	+	+	+	+	+	+	+	+	0.55
CDK23 (cons,topol,geom,elect,hybr) 3D:rdkit3d	+	0.52	+	+	+	+	+	+	+	+	0.47
Dragon6 (2D blocks: 1-28)	+	0.7	+	+	+	+	+	+	+	+	0.76
Dragon7 (3D blocks: 1-30) 3D:rdkit3d	+	0.51	+	+	+	+	+	+	+	+	0.49
EPA (T.E.S.T.)	+	0.51	+	+	+	+	+	+	+	+	0.46
Fragmentor (length: 2-4)	running	0.48	+	0.48	0.67	0.59	0.49	0.62	0.51	0.5	0.43
GSFrag (F + L)	+	0.62	+	+	+	+	+	+	+	+	0.57
InductiveDescriptors 3D:rdkit3d	+	0.51	+	+	+	+	+	+	+	+	0.55
JPlogP	+	0.51	+	+	+	+	+	+	+	+	0.46
MNA	+	0.79	+	+	+	+	+	+	+	+	0.64
MOLD2	+	0.52	+	+	+	+	+	+	+	+	0.48
MORDRED (2D)	+	0.53	+	+	+	+	+	+	+	+	0.49
MORDRED (3D) 3D:rdkit3d	+	0.56	+	+	+	+	+	+	+	+	0.49
OEstate	0.44	0.48	+	0.54	0.6	0.93	0.51	0.54	0.54	0.5	0.45
OSMORDRED	+	0.51	+	+	+	+	+	+	+	+	0.48
QNPR (length:1 - 3)	+	0.58	+	+	+	+	+	+	+	+	0.5
RDKit (2D blocks: 1-3 5 10-11 15-16)	+	0.6	+	+	+	+	+	+	+	+	0.47
RDKit (3D blocks: 1-11 15-16) 3D:rdkit3d	+	0.56	+	+	+	+	+	+	+	+	0.5
RDKit (ECFP4)	+	0.85	+	+	+	+	+	+	+	+	0.64
SIRMS (labels: charge+logp+hb+refractivity)	+	0.54	+	+	+	+	+	+	+	+	0.43
alvaDesc (2D blocks: 1-12 22-25 29 32-33)	+	0.54	+	+	+	+	+	+	+	+	0.48
alvaDesc (3D blocks: 1-33) 3D:rdkit3d	+	0.54	+	+	+	+	+	+	+	+	0.49
CNF2 (F) (3D) 10/10	+	+	0.46	+	+	+	+	+	+	+	+
DEEPCHEM (F) (3D) TEXTCNN 10/10	+	+	0.47	+	+	+	+	+	+	+	+
KGCNN AttFP	+	+	0.42	+	+	+	+	+	+	+	+
KGCNN ChemProp	+	+	0.43	+	+	+	+	+	+	+	+
TRANSNN (F) (3D) 10/10	+	+	0.52	+	+	+	+	+	+	+	+
TRANSNNI (F) 10/10	+	+	0.43	+	+	+	+	+	+	+	+

Selection of best models

Model profile

Statistical parameters, tables, charts - all the information related to the model.

Overview

Applicability domain

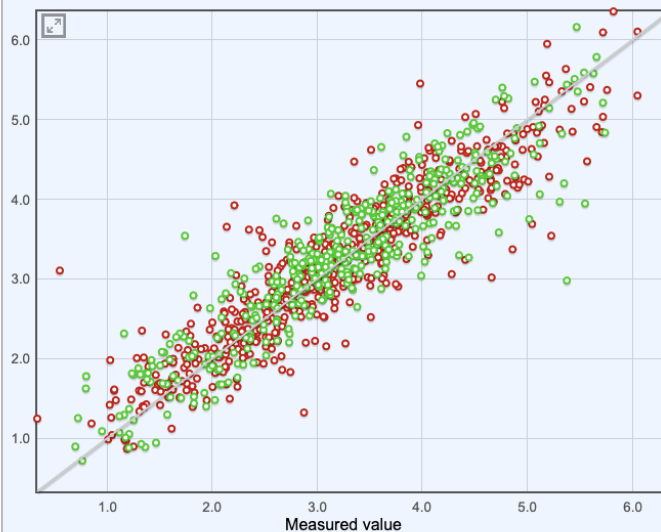
Model name: **log(IGC50-1)_KGCNN AttFP** [\[rename\]](#)

Temporal Public ID: **22966388** - use this link to share this temporal model (will be deleted after 3 months unless published)

Predicted property: **log(IGC50-1)** modeled in -log(mol/L)

Training method: KGCNN

Data Set	#	R2	q2	RMSE	MAE
Training set: <i>T. pyriformis</i> (training)	644 records	0.84 ± 0.01	0.84 ± 0.01	0.42 ± 0.02	0.3 ± 0.01
Test set: <i>T. pyriformis</i> (test) [x]	449 records	0.83 ± 0.02	0.83 ± 0.02	0.44 ± 0.02	0.33 ± 0.01



[\[Exclude duplicated records\]](#) [\[Exclude duplicated molecules\]](#)

[Export this model](#)

[Create a copy of this model](#)

[View configuration XML](#)

[Export configuration XML](#)

[MI](#)

PUBLISH THE MODEL

RECALCULATE MODEL AND STATISTICS

APPLY THE MODEL TO NEW COMPOUNDS

Consensus model development

Model creator

Select model template and training set

Choose the individual models for consensus

In order to build a consensus model, you must select several (at least two) individual models based on the selected training set **T. pyriformis (training)**.

[\[Add a model\]](#)

Model name	Method	
log(IGC50-1)_TRANSNNI (F) _ 10/10	TRANSNNI	[x]
log(IGC50-1)_KGCNN AttFP	KGCNN	[x]
log(IGC50-1)_KGCNN ChemProp	KGCNN	[x]
log(IGC50-1)_SCI:CAT_BOOST_[Fragmentor (length: 2-4)] - 554850	SKLEARN	[x]
log(IGC50-1)_SCI:CAT_BOOST_[ALogPS, StructuralAlerts]	SKLEARN	[x]
log(IGC50-1)_SCI:CAT_BOOST_[ALogPS, OEstate]	SKLEARN	[x]
log(IGC50-1)_SCI:CAT_BOOST_[SIRMS (labels: charge+logp+hb+refractivity)]	SKLEARN	[x]

Consensus type

☒ Ignore errors

✓ Simple average

Optimal average combination of models

Median values

Optimal median combination of models

Weighted by model accuracy

Weighted by molecule prediction accuracy

Best model for molecule

Consensus model

Model profile

Statistical parameters, tables, charts - all the information related to the model.

Overview

Applicability domain

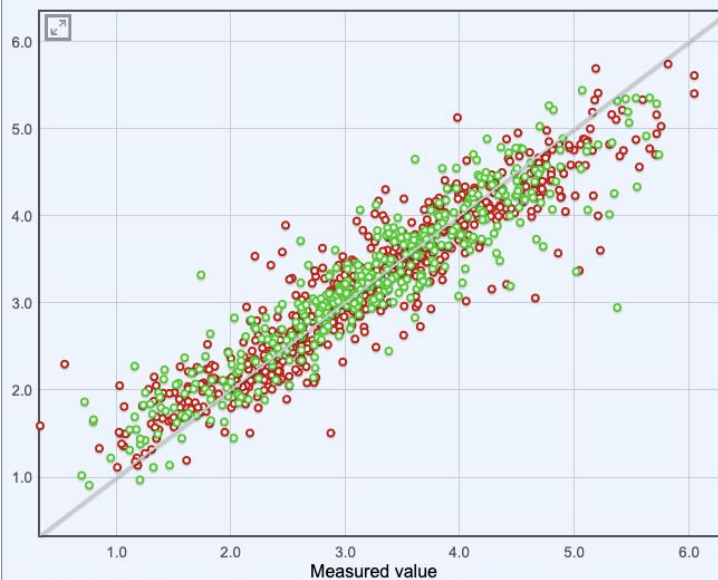
Model name: Consensus log(IGC50-1) [\[rename\]](#)

Temporal Public ID: [19562889](#) - use this link to share this temporal model (will be deleted after 3 months unless published)

Predicted property: **log(IGC50-1)** modeled in -log(mol/L)

Training method: Consensus

Data Set	#	R2	q2	RMSE	MAE
● Training set: T. pyriformis (training)	644 records	0.87 ± 0.01	0.87 ± 0.01	0.38 ± 0.02	0.27 ± 0.01
● Test set: T. pyriformis (test) [x]	449 records	0.86 ± 0.02	0.85 ± 0.02	0.4 ± 0.02	0.29 ± 0.01



[\[Exclude duplicated records\]](#) [\[Exclude duplicated molecules\]](#)

Each descriptor/method sees only part of a molecule



Blind monks examining an elephant, an [ukiyo-e](#) print by [Hanabusa Itchō](#) (1652–1724).
https://en.wikipedia.org/wiki/Blind_men_and_an_elephant

OCHEM modeling

- Comprehensive modeling
- Multitask learning (up to 100 properties)
- Feature net ("model in model")
- Consensus models
- GPU + CPU modern methods (~20)
- Supports models
 - >1,000,000 compounds
 - >200,000,000,000 descriptors*
 - >1,000 servers
 - up to 1GB in size (Java limit)
- Model private/publishing
- Export, import, web/REST services
- Conditions, external descriptors
- ToxAlerts screening

Predicted property: LLNA skin sensitization

Training set: TRAINING-SARpy-SKIN-SENS-giugno20 OK.xlsx

Metrics for Validation:

	LSSVMG	ASNN	PLS	KNN
ALogPS, OEstate	0.74	0.68	0.61	0.64
CDDD	0.8	0.74	0.75	0.71
CDK2 (cons,topol,geom,elec,hybrid) 3D:corina	0.75	0.71	0.56	0.71
ChemaxonDescriptors (pH 0 - 14:1) 3D:corina	0.76	0.7	0.59	0.68
Dragon6 (2D blocks: 1 28)	0.64	0.66	0.59	0.65
Dragon6 (3D blocks: 1-29) 3D:corina	0.76	0.72	0.57	0.65
Fragmentor (length:2 - 4)	0.72	0.7	0.59	0.63
GSFrag (F + L)	0.69	0.69	0.61	0.61
InductiveDescriptors 3D:corina	0.69	0.71	0.57	0.67
JPIlogP	0.73	0.74	0.59	0.67
MAP4	0.71	0.65	0.59	0.67
MORDRED (All) 3D:corina	0.77	0.73	0.57	0.68
Mera, Mersy 3D:corina	0.73	0.69	0.55	0.67
OEstate	0.74	0.67	0.63	0.68
PyDescriptor 3D:corina	0.71	0.71	0.7	0.67
QNPR (length:1 - 3)	0.68	0.62	0.58	0.58
RDKit (3D blocks: 1-11 15-16) 3D:corina	0.77	0.72	0.56	0.65
SIRMS (labels:charge+logp+hb+refractivity)	0.76	0.73	0.59	0.67
Spectrophores (accuracy=20) 3D:corina	0.68	0.6	0.52	0.6
StructuralAlerts	0.67	0.64	0.58	0.51
alvaDesc (3D blocks: (only) 1-30) 3D:corina	0.75	0.71	0.57	0.68

* Sparse format, DOI:10.1186/s13321-016-0113-y

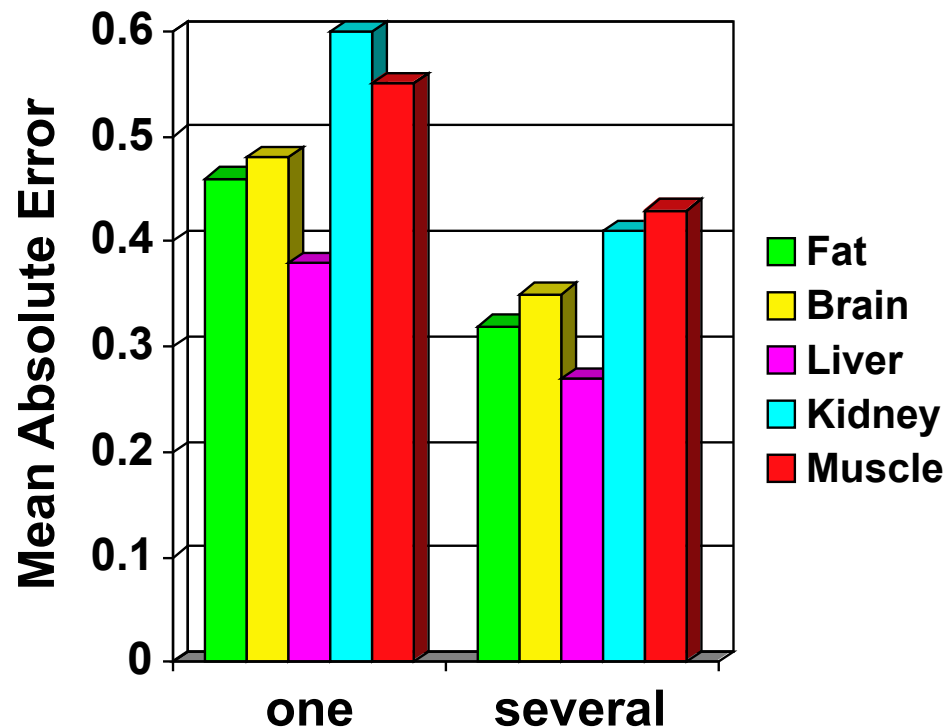
Multi-task learning

Problem:

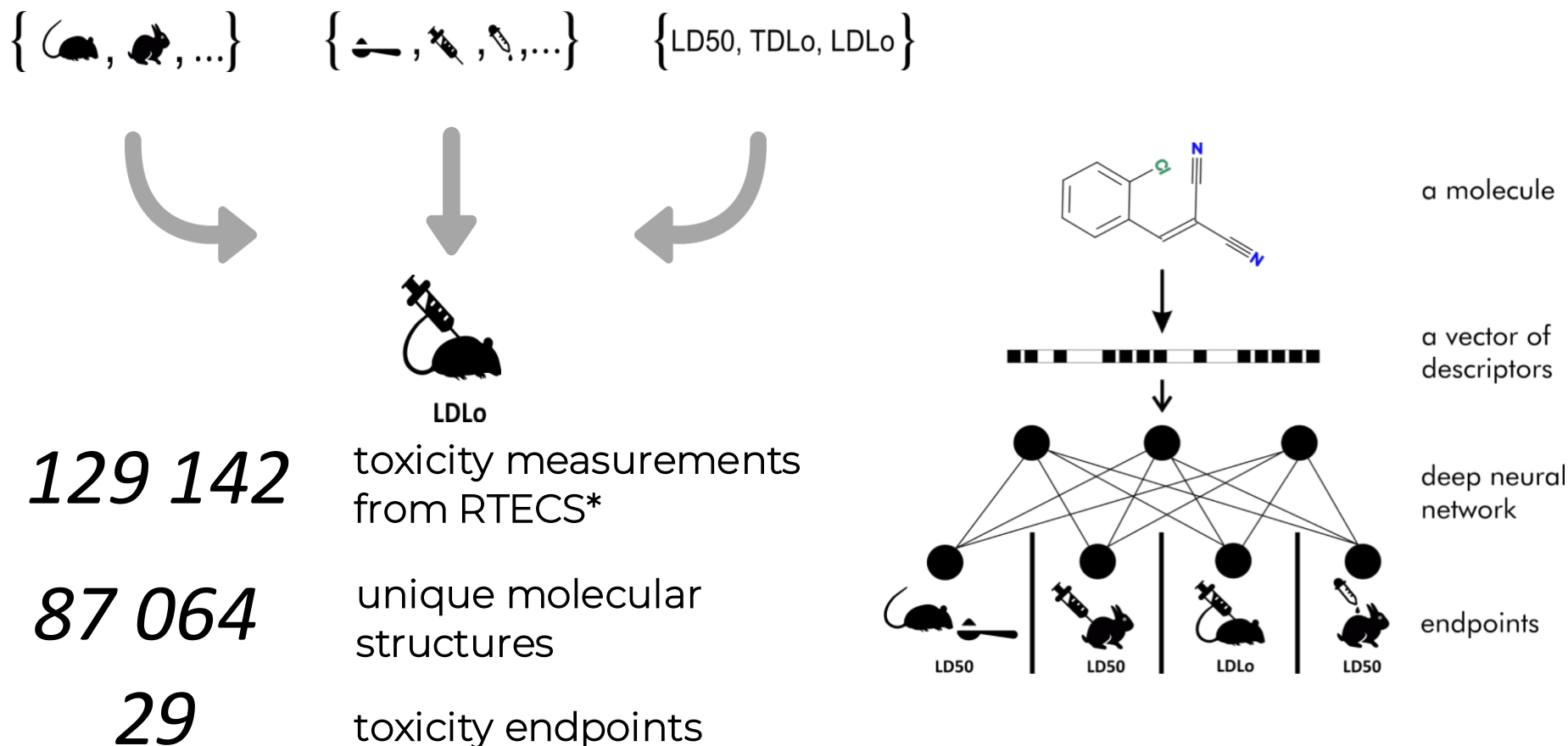
- prediction of tissue-air partition coefficients
- small datasets 30-100 molecules (human & rat data)

Results:

simultaneous prediction of several properties increased the accuracy of models

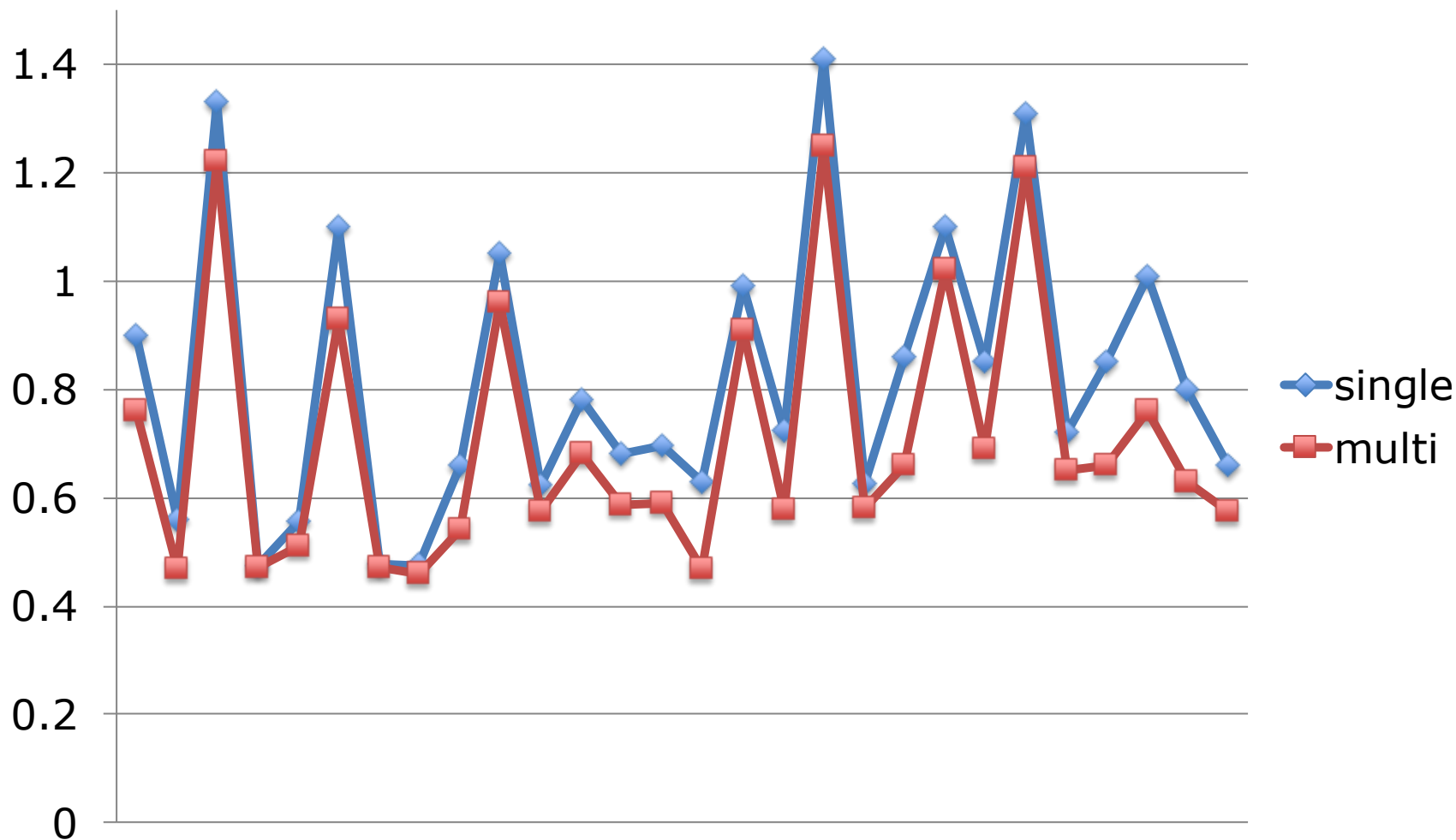


Analysis of toxicity of chemical compounds



*RTECS: Registry of Toxic Effects of Chemical Substances

RMSE for different toxicities using CDK descriptors and DNN



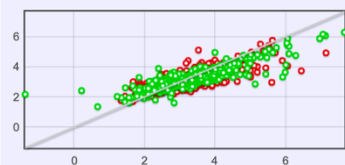
Sosnin, S.; Karlov, D.; Tetko, I.V.; Fedorov, M.V. A comparative study of prediction of multi-target toxicity for a broad chemical space. *J Chem Inf Model.* **2018**, **59**, 1062-1072.

Comprehensive model view

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

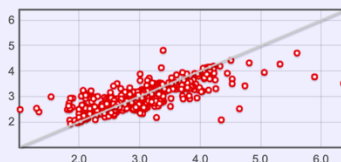
Property: mouse_intraperitoneal_LD50 measured in -log(M) ([Details..](#))

Dataset	R ₂	RMSE	MAE
data_upload_new.csv(36295)	0.65	0.41	0.26
uniques in RTECS(5189)	0.77	0.38	0.25



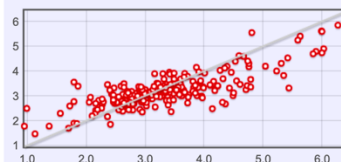
Property: mammal (species unspecified)_intraperitoneal_LD50 measured in -log(M) ([Details..](#))

Dataset	R ₂	RMSE	MAE
data_upload_new.csv(545)	0.61	0.42	0.26
uniques in RTECS(0)	0.00	0.00	0.00



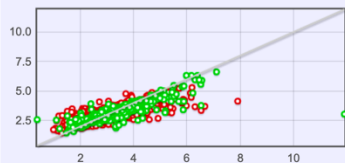
Property: guinea pig_intraperitoneal_LD50 measured in -log(M) ([Details..](#))

Dataset	R ₂	RMSE	MAE
data_upload_new.csv(248)	0.66	0.60	0.44
uniques in RTECS(0)	0.00	0.00	0.00



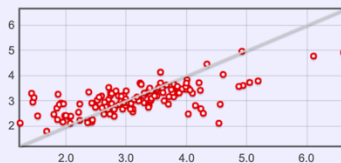
Property: rat_intraperitoneal_LD50 measured in -log(M) ([Details..](#))

Dataset	R ₂	RMSE	MAE
data_upload_new.csv(5021)	0.63	0.53	0.36
uniques in RTECS(933)	0.72	0.47	0.25



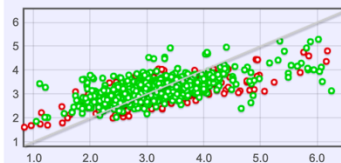
Property: rabbit_intraperitoneal_LD50 measured in -log(M) ([Details..](#))

Dataset	R ₂	RMSE	MAE
data_upload_new.csv(131)	0.46	0.69	0.50
uniques in RTECS(0)	0.00	0.00	0.00



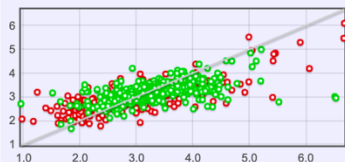
Property: mouse_intraperitoneal_LDL0 measured in -log(M) ([Details..](#))

Dataset	R ₂	RMSE	MAE
data_upload_new.csv(266)	0.54	0.62	0.48
uniques in RTECS(2678)	0.44	0.52	0.38



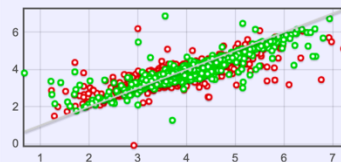
Property: rat_intraperitoneal_LDL0 measured in -log(M) ([Details..](#))

Dataset	R ₂	RMSE	MAE
data_upload_new.csv(318)	0.61	0.55	0.40
uniques in RTECS(757)	0.35	0.46	0.30



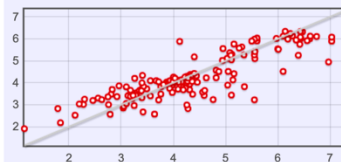
Property: mouse_intravenous_LD50 measured in -log(M) ([Details..](#))

Dataset	R ₂	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	R ₂	RMSE	MAE
data_upload_new.csv(153)	0.75	0.64	0.47
uniques in RTECS(0)	0.00	0.00	0.00



Comparison of different models to predict toxicity (RMSE)

	single	multi	single
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, OEstimate	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)

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.

Incorrect multitask validation

Model validation

Validation method: N-Fold cross-validation ▾

Number of folds: 5

- ☐ Stratified cross-validation (classification only) ⓘ
- ☐ Treat each record as a new molecule ⓘ
- ☐ (Experimental) Use scaffold-based split ⓘ

Multiple models overview

Predicted property: [Cblood/Cair\(Human\)](#)

Training set: [blood_tissue](#)

Duplicate models: [start](#)

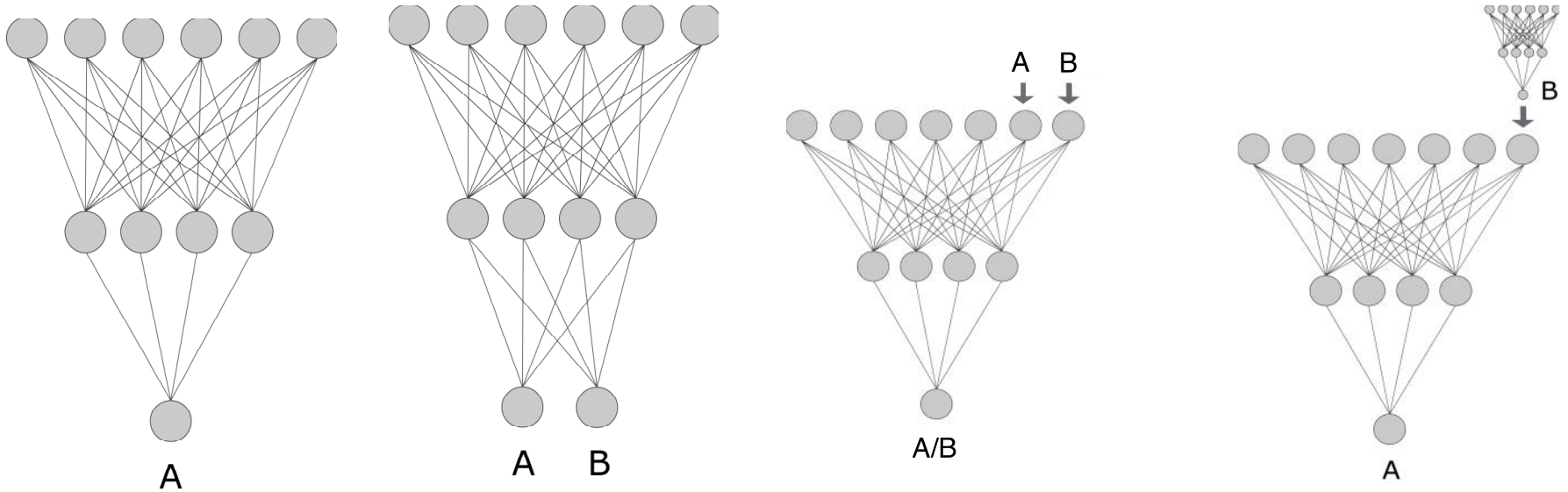
Metrics **RMSE - Root Mean Square Error** ▾ for **Training set** ▾ Validation: **Cross-Validation (3 models)** ▾

	ASNN	ASNN (STL)	ASNN (record)
Fragmentor (length: 2-4)	0.52 0.4 0.25 0.28 0.32 0.3 0.37 0.44 0.45 0.41 0.38 1.4 (0.46)	0.59 0.55 0.35 0.64 0.48 0.44 0.42 0.61 0.51 0.51 0.4 1.3 (0.5667)	0.47 0.3 0.15 0.16 0.24 0.2 0.3 0.27 0.3 0.3 0.39 1.3 (0.365)

⌂ Predicates (" $<$ ", " $>$ ") and/or optimal classification thresholds are NOT used for statistics (click to change)

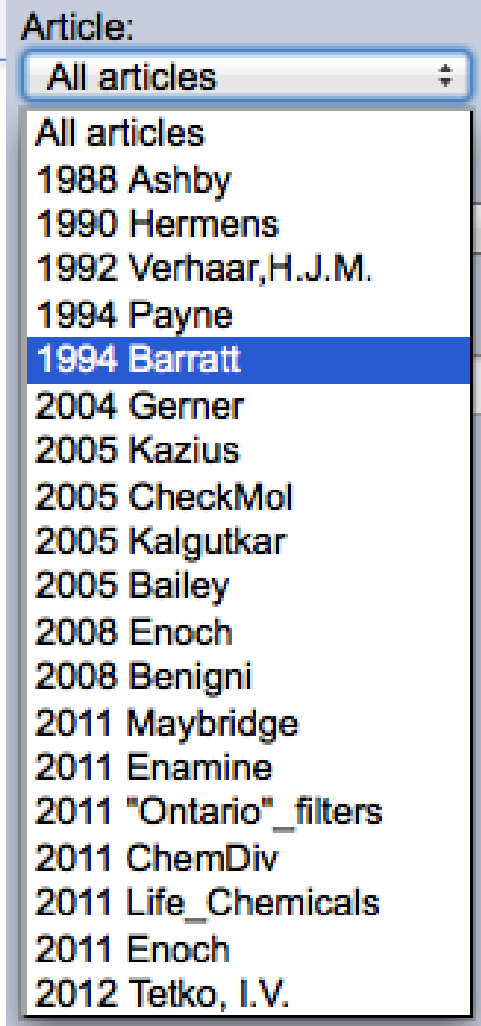
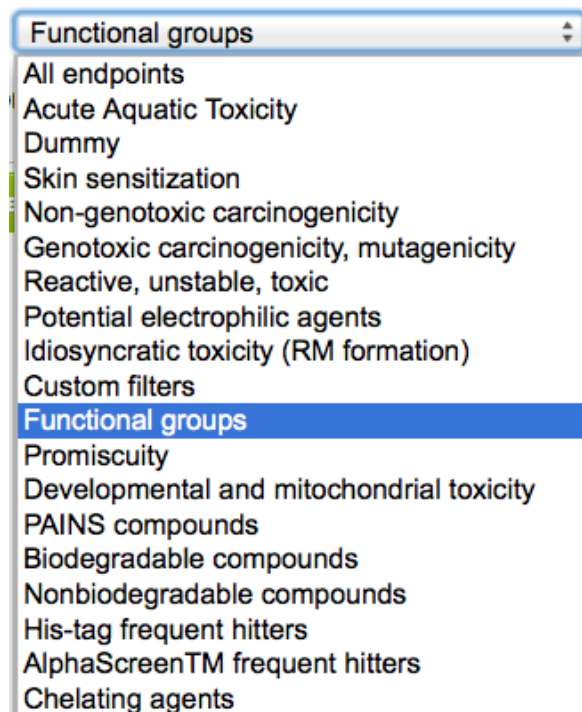
⌂ Refresh

Multi-task learning



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

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

FILTERS

Article:
2015 Salmina, E.

Endpoint / Filter type:
Functional groups

Name / Alert ID:

☐ Show only approved alerts

Upload new alerts Screen compounds

101 - 200 of 379

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)

HS

SMARTS: [*]1-[*]1-[*]1-[*]1-1

Endpoint: Functional groups

Salmina, E.
Extended functional groups (EFG): an efficient set for chemi...
Molecules 2015; subm ()
Alert ID: TA2450

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)

HS

SMARTS: [*]1-[*]1-[*]1-[*]1-1

Endpoint: Functional groups

Salmina, E.
Extended functional groups (EFG): an efficient set for chemi...
Molecules 2015; subm ()
Alert ID: TA2451

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)

HS

SMARTS: [*]1-[*]1-[*]1-[*]1-1

Endpoint: Functional groups

Salmina, E.
Extended functional groups (EFG): an efficient set for chemi...
Molecules 2015; subm ()
Alert ID: TA2452

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)

HS

SMARTS: [*]1-[*]1-[*]1-[*]1-1

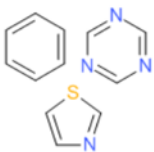
Endpoint: Functional groups

Salmina, E.
Extended functional groups (EFG): an efficient set for chemi...
Molecules 2015; subm ()
Alert ID: TA2453

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 for 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 	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

Using SetCompare functionality and selecting EFG as descriptors

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

$$Y = 0.5372 + 0.1612 * c_{10} + 0.1309 * p_{1-1N} - 0.1134 * p_{2B} + 0.05943 * c_3 + 0.05349 * c_9$$

E-state descriptors

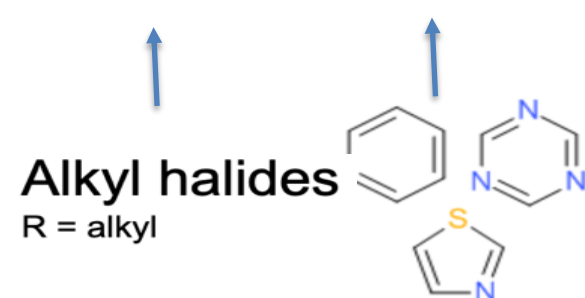
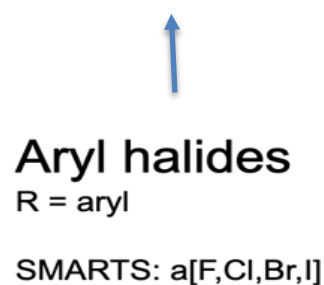
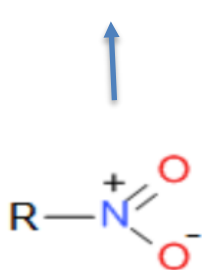
$$Y = 0.5375 + 0.09956 * \text{PSA} + 0.08731 * a_{\text{CNOS}} - 0.08703 * \text{DONORS} - 0.06814 * S_{\text{SCH3}} - 0.04474 * S_{\text{ssO}}$$

Dragon descriptors

$$Y = 0.5375 - 0.1173 * \text{GATS1m} + 0.0954 * \text{MATS1e} - 0.06558 * \text{SpMax_AEA(dm)} + 0.05933 * J_{\text{D/Dt}} + 0.05496 * n_{\text{R03}}$$

Functional groups

$$Y = 0.4733 + 0.191 * \text{Alert146} - 0.00411 * \text{Alert238} - 0.1024 * \text{Alert213} + 0.1912 * \text{Alert214} + 0.0199 * \text{Alert196}$$



Tox24 Challenge

EPA submitted an article in February with results of screening of TTR-interfering 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.

Eytcheson, S. A.; Tetko, I. V. Which Modern AI Methods Provide Accurate Predictions of Toxicological End Points?

Analysis of Tox24 Challenge Results. *Chem. Res. Toxicol.* **2025**, 38 (9), 1443–1451. <https://doi.org/10.1021/acs.chemrestox.5c002>

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

Follow up study by the EPA (JCIM)



pubs.acs.org/jcim

Article

Combined *In Vitro* and *In Silico* Workflow to Deliver Robust, Transparent, and Contextually Rigorous Models of Bioactivity

Nathaniel Charest,* Gabriel Sinclair, Stephanie A. Eytcheson, Daniel T. Chang, Todd M. Martin, Charles N. Lowe, Katie Paul Friedman, and Antony J. Williams



Cite This: *J. Chem. Inf. Model.* 2025, 65, 4426–4441



Read Online

ACCESS |



Metrics & More

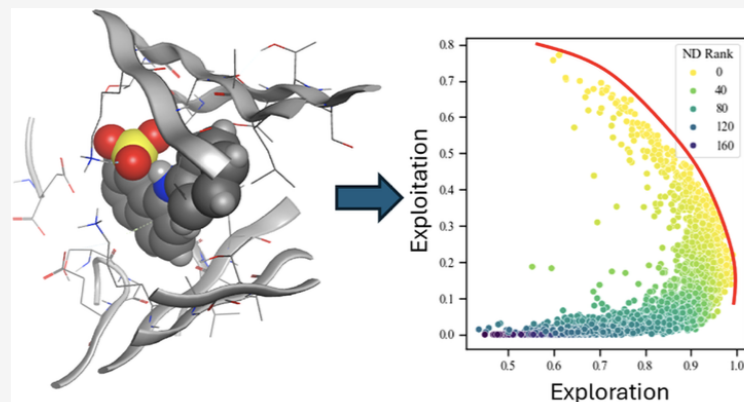


Article Recommendations



Supporting Information

ABSTRACT: New approach methodologies (NAMs) are an increasing priority in the field of toxicology to fill data gaps and reduce time and resources in chemical safety assessment. We describe an NAMs workflow that integrates an *in vitro* high-throughput bioassay with an *in silico* computational model. In defining this workflow, we propose, as a crucial step of *in silico* development, the identification of explicit “purpose contexts”: *a priori* definitions of the scope and intent of an *in silico* solution, which provide natural targets for the mechanistic interpretation, validation, and output design of the model. By inspecting data from an *in vitro* assay measuring the displacement of fluorescent probe 8-anilino-1-naphthalenesulfonic acid (ANSA) from the serum transport protein transthyretin (TTR) as a proxy for potential



Descriptors from JCIM article

FEATURE	DESCRIPTOR MEANING
CrippenLogP	A group contribution model of predicted water-octanol partition coefficient
ATSC3c	The autocorrelations of lag 3 within the molecular graph weighted by charge. Useful for capturing local structure motifs or patterns that encode bioactivity.
ATSC5c	The autocorrelations of lag 5 within the molecular graph weighted by charge. Useful for capturing local structure motifs or patterns that encode bioactivity.
C1SP3	Number of SP3 hybridized carbons.
ETA_BetaP_s	A measure of electrons in resonance.
naAromAtom	Number of aromatic atoms in the molecular graph
ZMIC5	A measure of neighborhood in the molecular graph symmetry. ZMIC(n) where n is the radius of the neighborhood unit.
ATSC4m	The autocorrelations of lag 4 within the molecular graph weighted by mass. Useful for capturing local structure motifs or patterns that encode bioactivity.
ZMIC1	A measure of neighborhood in the molecular graph symmetry. ZMIC(n) where n is the radius of the neighborhood unit.
C1SP2	Number of SP2 hybridized carbons.
hmax	Maximum electronic state of the molecule.
ATSC2m	The autocorrelations of lag 2 within the molecular graph weighted by mass. Useful for capturing local structure motifs or patterns that encode bioactivity.
hmin	Minimum electronic state of the molecule.
ATSC0m	The autocorrelations of lag 0 within the molecular graph weighted by mass. Useful for capturing local structure motifs or patterns that encode bioactivity.
ZMIC2	A measure of neighborhood in the molecular graph symmetry. ZMIC(n) where n is the radius of the neighborhood unit.
ATSC5m	The autocorrelations of lag 5 within the molecular graph weighted by mass. Useful for capturing local structure motifs or patterns that encode bioactivity.
ATSC4c	The autocorrelations of lag 4 within the molecular graph weighted by charge. Useful for capturing local structure motifs or patterns that encode bioactivity.
VE1_DzZ	The sum of coefficients from the first eigenvector of the Barysz matrix. Very abstract, but a summary descriptor for capturing heteroatomic topological similarity. Useful for global molecular similarity.
nHBAcc	Number of hydrogen bond acceptors.
ETA_dAlpha_A	Measure of number of heteroatoms
ZMIC3	A measure of neighborhood in the molecular graph symmetry. ZMIC(n) where n is the radius of the neighborhood unit.
ETA_Beta_ns_d	A set of descriptors capturing, in order, a measure of non-hydrogen heteroatoms; a measure of hydrogen bonding propensity; and electrons within a resonance structure. Specifically, the number of non-hydrogen heteroatoms, the number of hydrogen bond donors, and the number of hydrogen bond acceptors.
ATSC1m	The autocorrelations of lag 1 within the molecular graph weighted by mass. Useful for capturing local structure motifs or patterns that encode bioactivity.
ATSC2c	The autocorrelations of lag 2 within the molecular graph weighted by charge. Useful for capturing local structure motifs or patterns that encode bioactivity.
ATSC3m	The autocorrelations of lag 3 within the molecular graph weighted by mass. Useful for capturing local structure motifs or patterns that encode bioactivity.
maxdssC	An electrotopological index capturing carbons with a double bond and two single bonds. Included due to sensitivity to aromatized carbons with substituent bonds.
ATSC0c	The autocorrelations of lag 0 within the molecular graph weighted by charge. Useful for capturing local structure motifs or patterns that encode bioactivity.
ETA_Shape_Y	A measure of hydrogen bonding propensity of the molecule.
ZMIC4	A measure of neighborhood in the molecular graph symmetry. ZMIC(n) where n is the radius of the neighborhood unit.
nHBDon	Number of hydrogen bond donors.
ATSC1c	The autocorrelations of lag 1 within the molecular graph weighted by charge. Useful for capturing local structure motifs or patterns that encode bioactivity.

“Data matters!”

#	Training set size	RMSE, 5CV training set	RMSE, test set	Sensitivity, %	Balance accuracy, ¹ %	Model
0	639	-	-	77	80	Original study by EPA
I	639	22.0±0.6	21 ± 1	88	87	https://ochem.eu/model/1266
II	1298	22.7±0.6	19 ± 1	91	89	https://ochem.eu/model/1212
III	639	23.5±0.6	21 ± 1	81	83	https://ochem.eu/model/1268
IV	1298	24.3±0.6	20 ± 1	84	85	https://ochem.eu/model/1267

I-II developed using the winning model protocol

III & IV are developed using logP/S + Extended Functional Groups*

Salmina, E. S.; Haider, N.; Tetko, I. V. Extended Functional Groups (EFG): An Efficient Set for Chemical Characterization and Structure-Activity Relationship Studies of Chemical Compounds. *Molecules*. 2015, 21 (1), E1.

<https://doi.org/10.3390/molecules21010001>.

Model built using functional groups

MLRA model

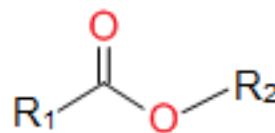
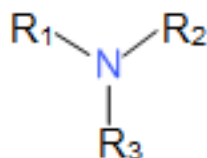
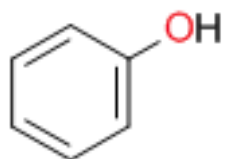
$$Y = 27.47 + 12.72 \cdot \text{ALogPS_logP} + 16.58 \cdot \text{TA1139} + 1.051 \cdot \text{TA1301} - 1.289 \cdot \text{TA1340} + 5.556 \cdot \text{TA1343} + 9.262 \cdot \text{TA1169} - 6.594 \cdot \text{TA1315} + 9.814 \cdot \text{TA1152} - 5.707 \cdot \text{TA1178}$$

TA1139 – Phenols; TA1301 – Aromatic compounds;

TA1340 – Nonmetals; TA1343 – Chalcogens;

TA1169 – Aryl chlorides; TA1315 – Aryl halides;

TA1152 – Amines; TA1178 – Carboxylic acid esters;

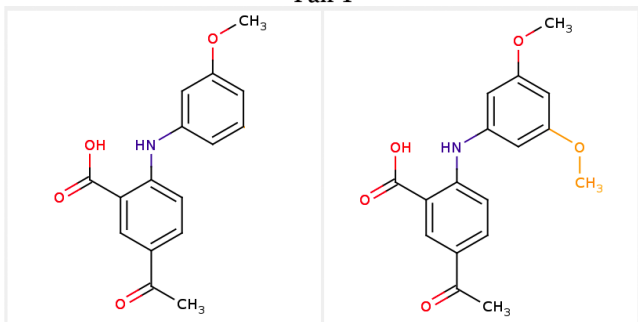


MMP definition

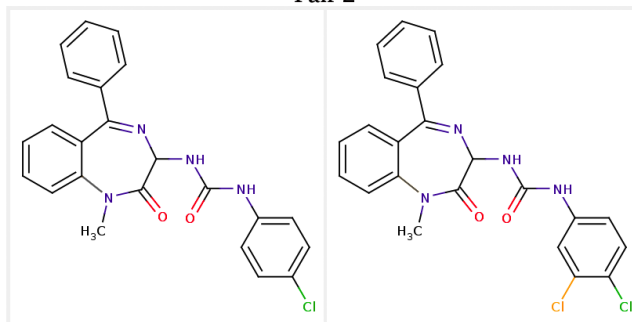
A molecular matched pair (MMP) is a pair of molecules that have only a (minor) single-point difference.

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

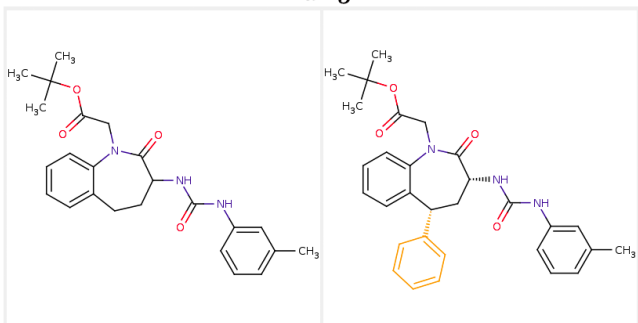
Pair 1



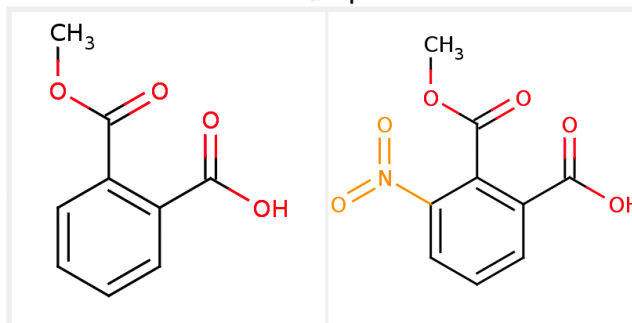
Pair 2



Pair 3



Pair 4



View historical facts

- Term was first introduced in 2005 by Kenny and Sadowski
 - Popularized by works of A. Leach, J. Bajorath, etc.
 - Commercially distributed by MedChemica
 - Publicly available in KNIME, OCHEM, etc.
 - 2013: Roche, AZ and MedChemica formed consortium with a goal
- “extracts the pre-competitive knowledge of how to design drugs, and combines these into a large database, which all the members of the consortium use”.

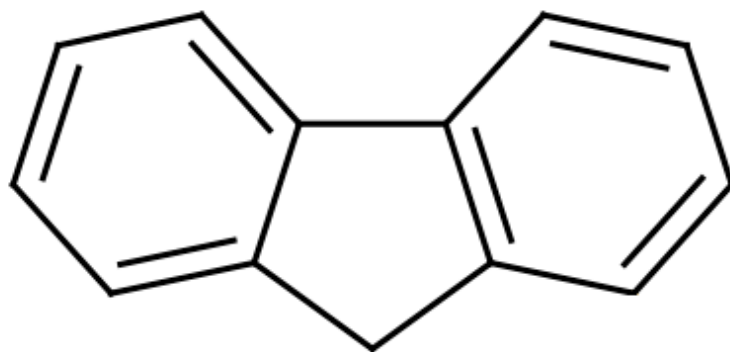
Why MMPs?

Keyword: **interpretation**

Small localized changes allow for interpreting activity

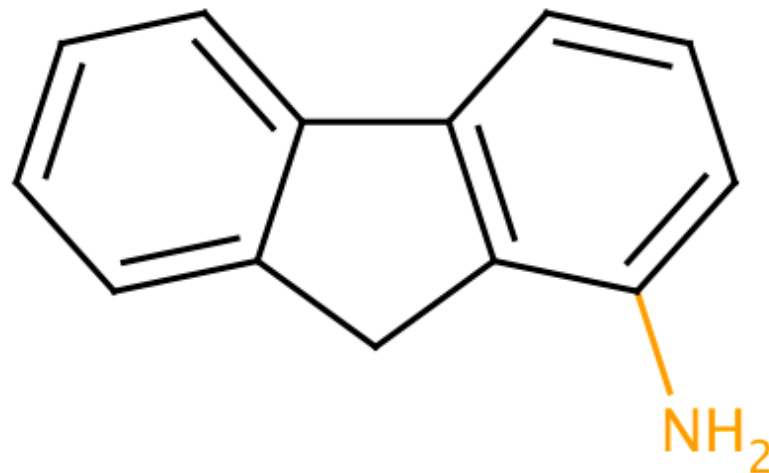
- Identification of interpretable activity cliffs
- Identification of “significant” transformations
- Interpretation of models

MMP example



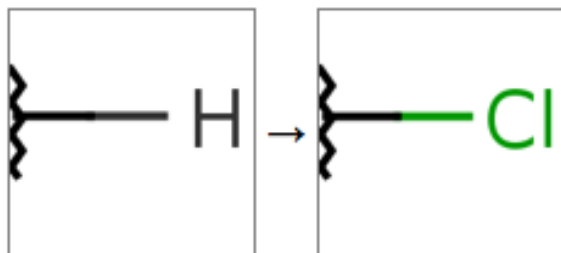
Melting Point
AMES
logPow

= 112.0 – 116.0 (in °C)
= inactive
= 4.18 (in Log unit)

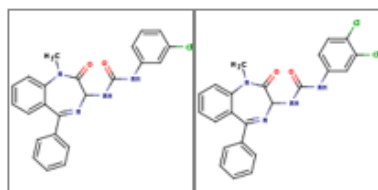


= 125.5 (in °C)
= active
= 3.18 (in Log unit)

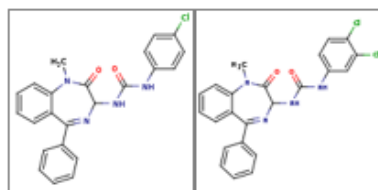
MMPs and transformations



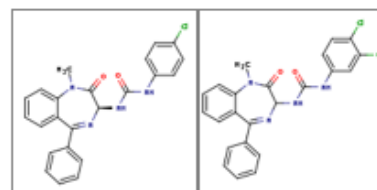
Transformation ID: TR3651887
SMIRKS: *[H] -> *Cl



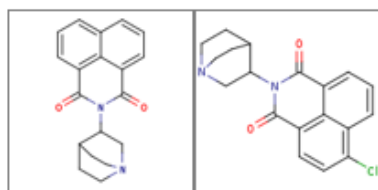
M1000289022 M1000288950



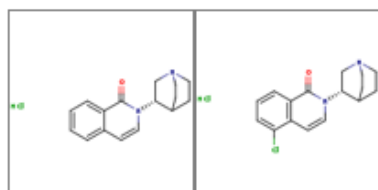
M1000289025 M1000288950



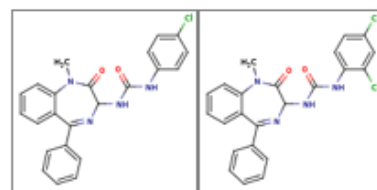
M1000289110 M1000288950



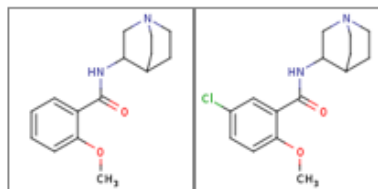
M1000303314 M1000303365



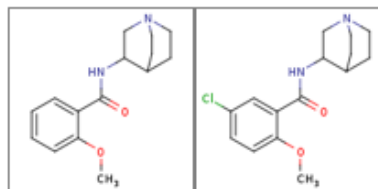
M1000303309 M1000303347



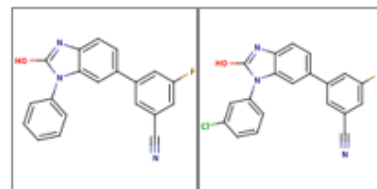
M1000289025 M1000288930



M1000303332 M1000303429

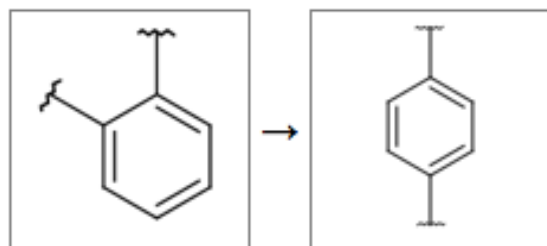


M1000303332 M1000303429



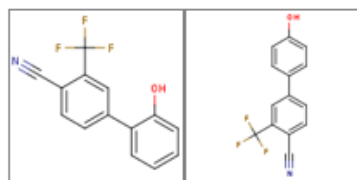
M1000050303 M1000050306

MMPs and transformations



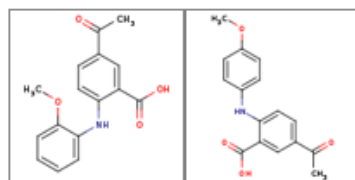
Transformation ID: TR3651270

SMIRKS: *C1=CC=CC=C1* -> *C1=CC=C(*)C=C1



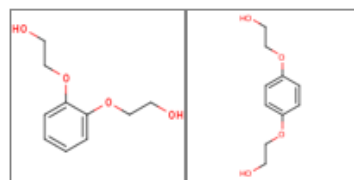
M1000050341

M1000050340



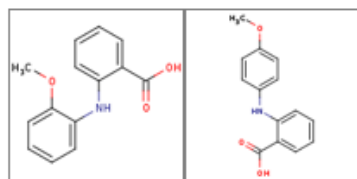
M1002111542

M1002111562



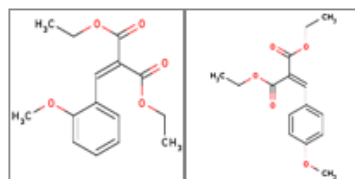
M108286

M65324



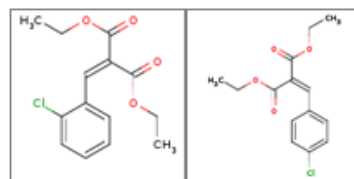
M1002111554

M1001278883



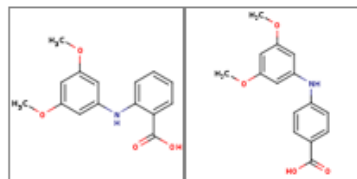
M31447

M31448



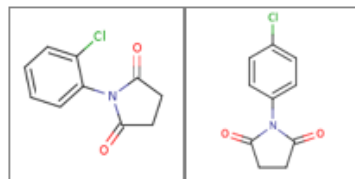
M31445

M31477



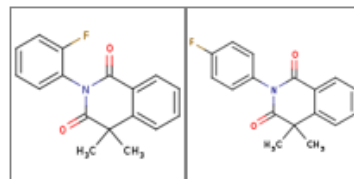
M1002111513

M1002111550



M31913

M31550



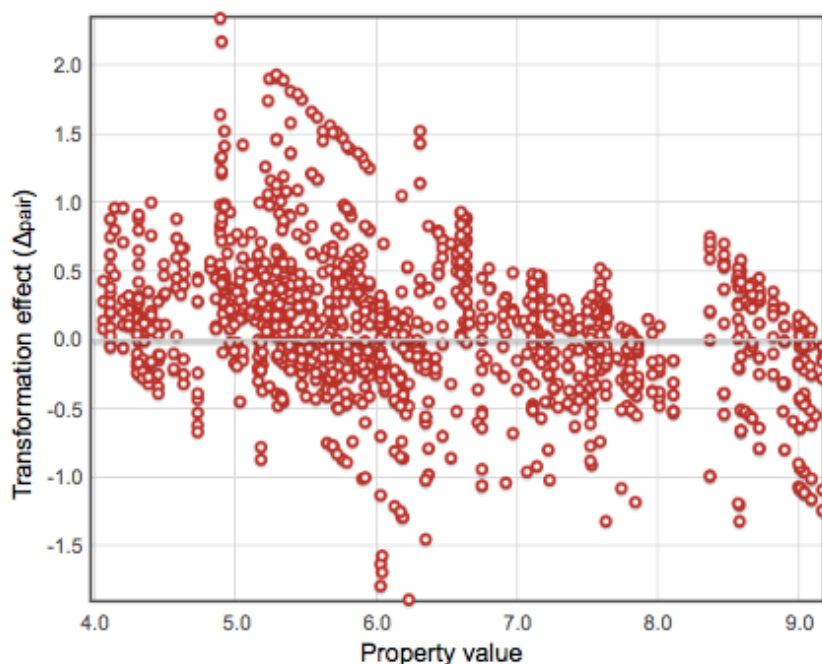
M1000050431

M1000050429

MMPs for quantitative activity data

The info is based on dataset **ic50_RT HIV** for property **IC50 HIV**

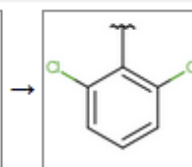
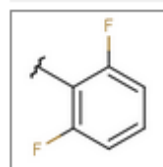
Minimal pair similarity: 50



MMP transformations: minimum 4 pairs, p-value 0.05, max. 4

1 - 5 of 9

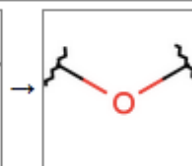
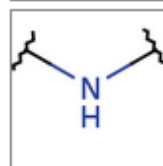
5 items on page 1 of 2 > >>



9 matched pairs

$\Delta \text{ mean} = -0.79 \pm 0.22$

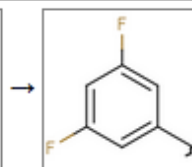
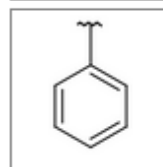
SMIRKS: FC1=CC=CC(F)=C1* -> *C1=C(Cl)C=CC=C1Cl



9 matched pairs

$\Delta \text{ mean} = -0.32 \pm 0.19$

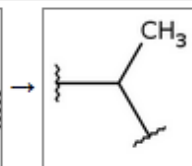
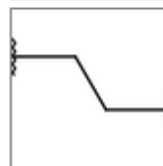
SMIRKS: *N* -> *O*



16 matched pairs

$\Delta \text{ mean} = 0.33 \pm 0.12$

SMIRKS: *C1=CC=CC=C1 -> FC1=CC(*)=CC(F)=C1



9 matched pairs

$\Delta \text{ mean} = 0.43 \pm 0.18$

SMIRKS: *CC* -> CC(*)*

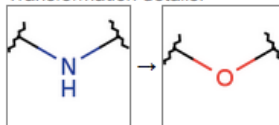
MMPs for quantitative activity data

The info is based on dataset **ic50_RT HIV** for property **IC50 HIV**

Minimal pair similarity:  



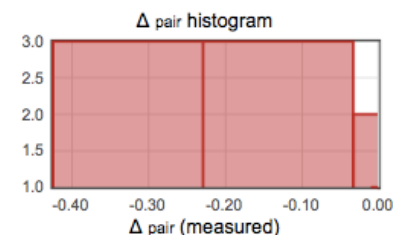
Transformation details:



Transformation ID: **TR 5719**

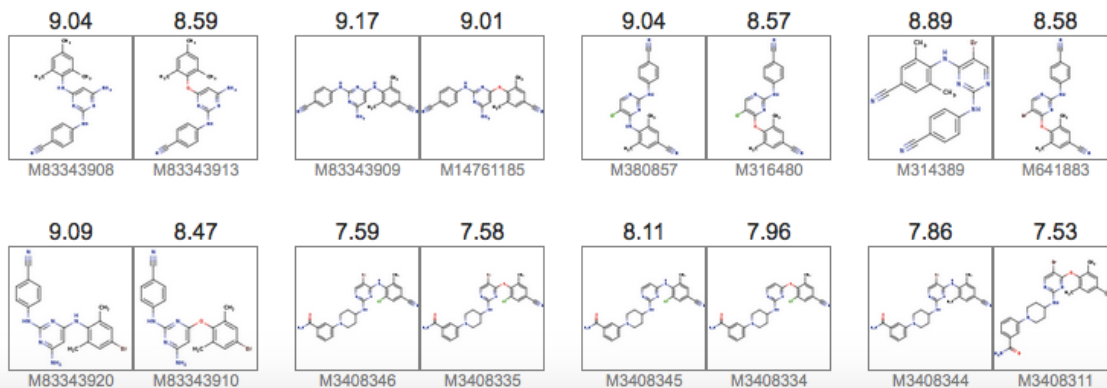
SMIRKS: ***N* -> *O***

[Back to all transformations](#)



Pairs having the same transformation as the selected one:

1 - 9 of 9



[Resubmit all molecules](#)

[Show fragment graph](#)

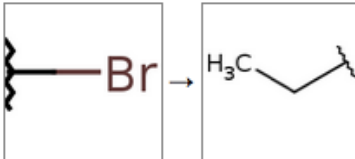
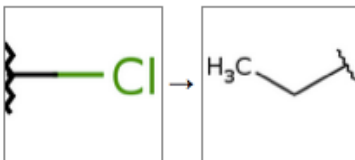
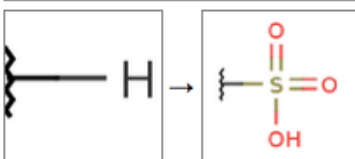
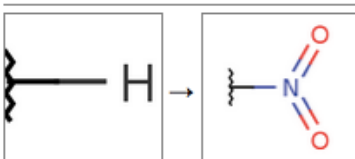
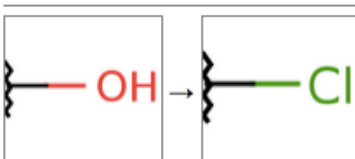
[Save transformations](#)

[Export pairs](#)

MMPs for classification data (AMES mutagenicity)

Transformations: minimum pairs, p-value

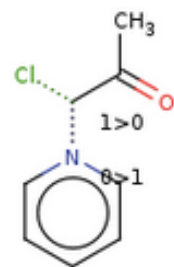
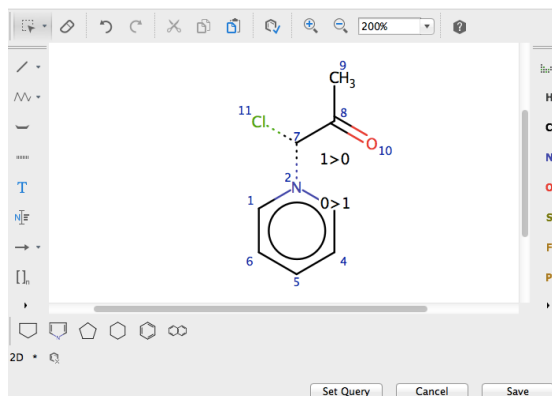
1 - 5 of 32 items on page of 7 > >>

	12 matched pairs ↑ 0 pairs, ↓ 7 pairs 2 negative and 3 positive molecules unaffected SMIRKS: <chem>*Br -> CC*</chem>
	21 matched pairs ↑ 0 pairs, ↓ 9 pairs 8 negative and 4 positive molecules unaffected SMIRKS: <chem>*Cl -> CC*</chem>
	17 matched pairs ↑ 1 pairs, ↓ 8 pairs 5 negative and 3 positive molecules unaffected SMIRKS: <chem>*[H] -> OS(*) (=O) = O</chem>
	204 matched pairs ↑ 57 pairs, ↓ 8 pairs 26 negative and 113 positive molecules unaffected SMIRKS: <chem>*[H] -> *N (=O) = O</chem>
	69 matched pairs ↑ 21 pairs, ↓ 4 pairs 25 negative and 19 positive molecules unaffected SMIRKS: <chem>O* -> *Cl</chem>

1 - 5 of 32

[Resubmit all molecules](#)[Show fragment graph](#)[Save transformations](#)[Export pairs](#)

Storage of Chemical Reaction as Condensed Graph of Reaction (CRG)



molecule profile

● **SN2 reaction**
rate constant =
-4.72 (in log(L*mol⁻¹*s⁻¹))

Nugmanov, R. et al
 Development of "structure-property" models in nucleophilic s...
 N: 8 P: 270
 Journal of Structural Chemistry **2014**; 55 (6) 1026-1032

MoleculeID: M84326300

Private record

Temperature= 55.6 °C
 Solvent = Ethanol

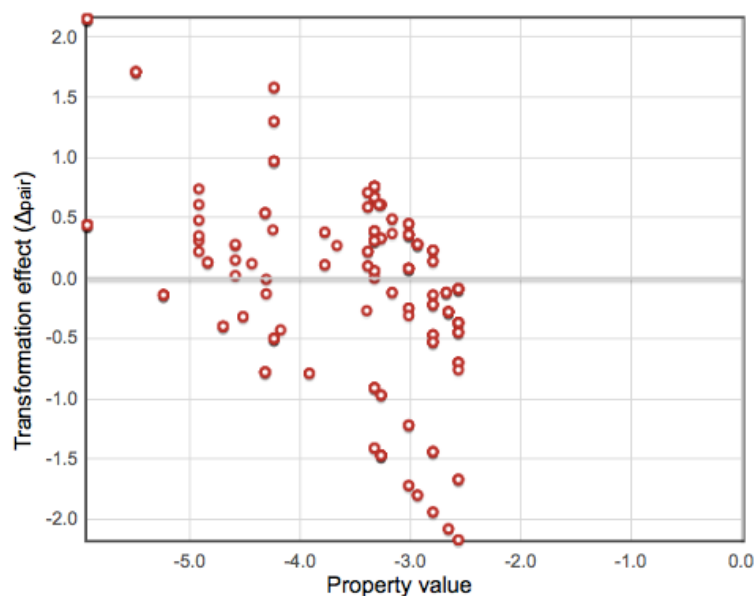
RecordID: R19277000
 23:25, 4 Jul 15 / 20:32, 6 Jul 15
 itetko
 Only visible to itetko



MMPs for CGRs

The info is based on dataset **methanol - 25** for property **SN2 reaction rate constant** (excluding 15 molecules not indexed for MMP analysis)

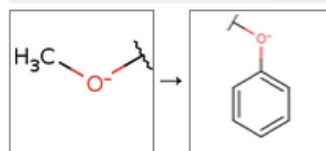
Minimal pair similarity: 50



MMP transformations: minimum 2 pairs, p-value Any, max. atoms Any

1 - 5 of 22

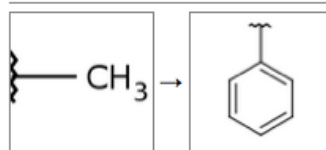
5 items on page 1 of 5



2 matched pairs

$\Delta \text{mean} = -0.59 \pm 0.27$

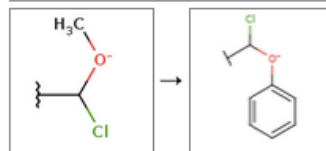
SMIRKS: [CH3:12][O-:13] -> *[O-:8][C:13]1=[CH:12][CH:11]=[CH:10][CH:15]=[CH:14]1



2 matched pairs

$\Delta \text{mean} = -0.59 \pm 0.27$

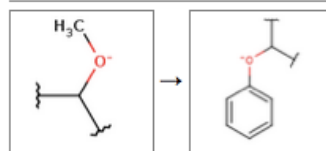
SMIRKS: C* -> *C1=CC=CC=C1



2 matched pairs

$\Delta \text{mean} = -0.59 \pm 0.27$

SMIRKS: [CH3:9][O-:8][CH:7](*)[Cl:11] -> *[CH:7]([Cl:19])[O-:8][C:13]1=[CH:12][CH:11]=[CH:10][CH:15]=[CH:14]1



2 matched pairs

$\Delta \text{mean} = -0.59 \pm 0.27$

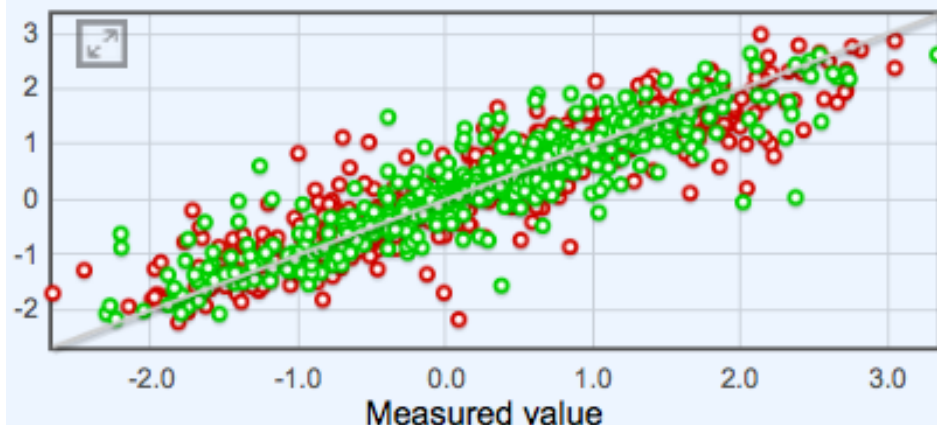
SMIRKS: [CH3:6][O-:5][CH:4](*) -> *[CH:7](*)[O-:8][C:13]1=[CH:12][CH:11]=[CH:10][CH:15]=[CH:14]1

Regression Model for environmental toxicity

Predicted property: **log(IGC50-1)** modeled in -log(mmol/L)

Training method: ASNN

Data Set	#	R2	q2	RMSE	MAE
● Training set: <i>T. pyriformis</i> train	644 records	0.83 ± 0.02	0.83 ± 0.02	0.44 ± 0.02	0.31 ± 0.01
● Test set: <i>T. pyriformis</i> test [x]	449 records	0.81 ± 0.02	0.8 ± 0.02	0.46 ± 0.03	0.32 ± 0.02



 [Download model statistics](#)

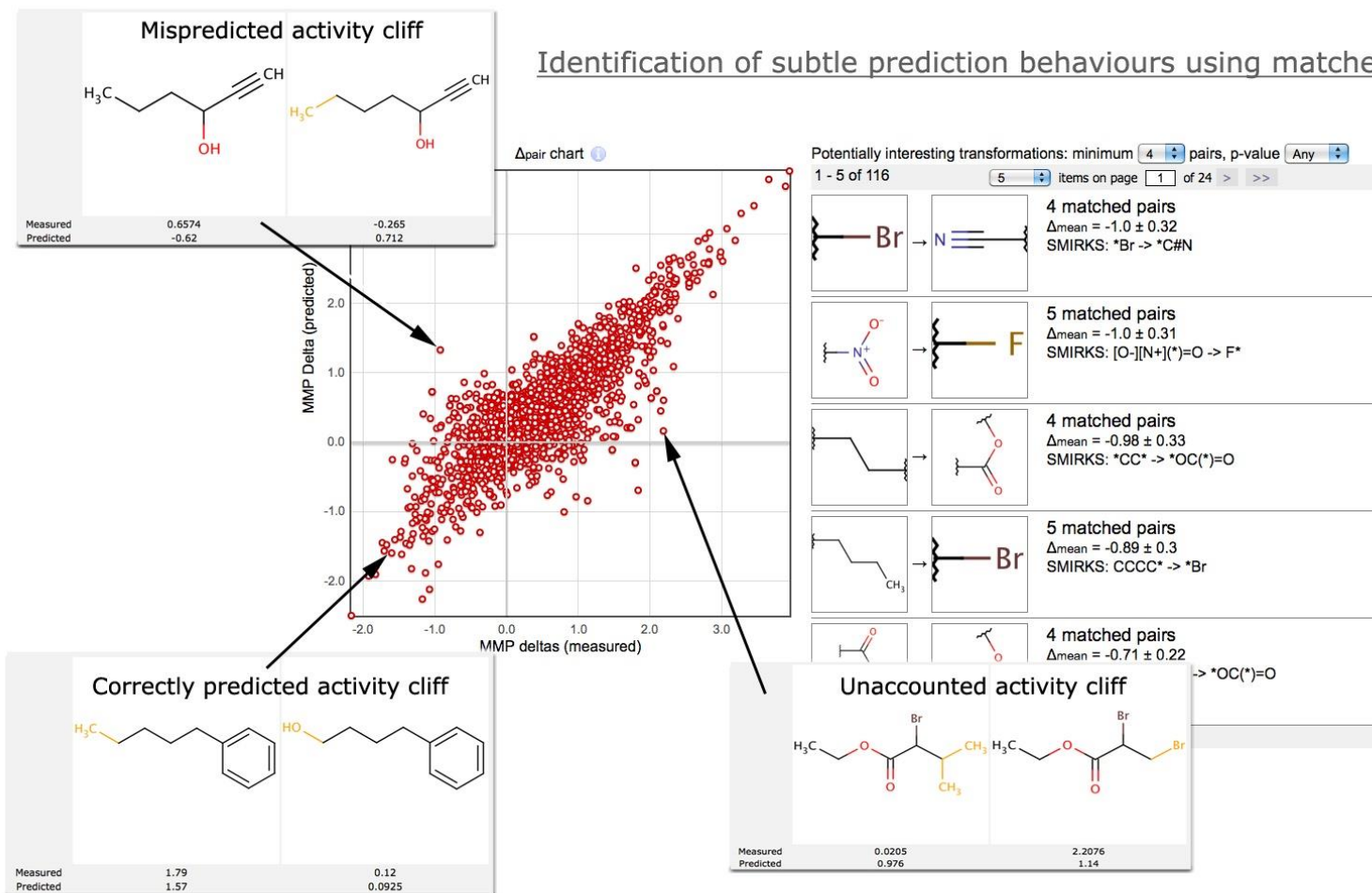
 [View configuration XML](#)

 [Export configuration XML](#)

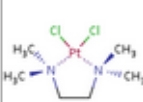
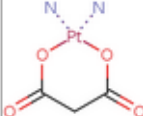
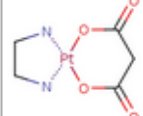
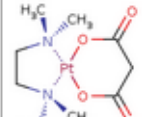
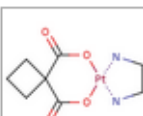
 [MMP-based analysis](#)

Identification of predicted MMPs

Identification of subtle prediction behaviours using matched pairs



LogP of Pt(II) + Pt(IV) complexes

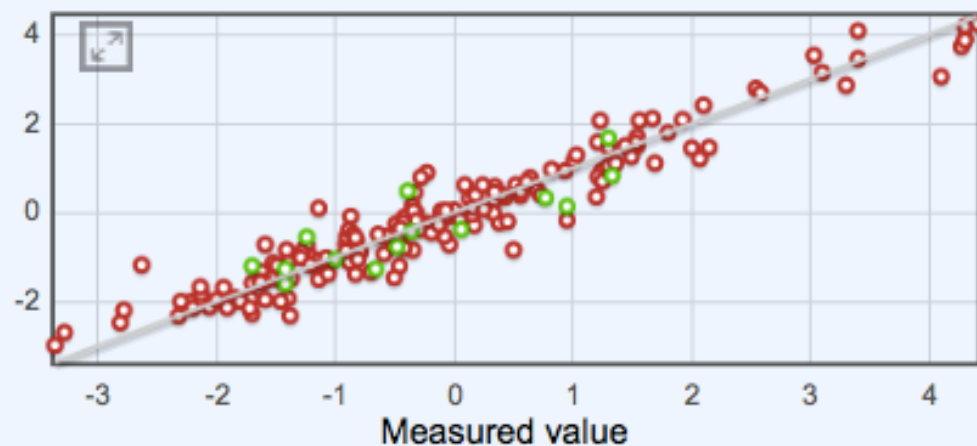
 molecule profile	● logPow = -0.85 (in Log unit) Platts, JA et al The RP-HPLC measurement and QSPR analysis of logP(o/w) value... P: 1204 T: 3 J. Inorg. Biochem. 2006; 100 (7) 1199-207 MoleculeID: M4325 Public record	logPow Method = RP-HPLC logPow Buffer = methanol RecordID: R13669 09:37, 19 Mar 09 / 12:22, 26 Koerner sz / tanzeem sz
 molecule profile	● logPow = -2.32 (in Log unit) Platts, JA et al The RP-HPLC measurement and QSPR analysis of logP(o/w) value... J. Inorg. Biochem. 2006; 100 (7) 1199-207 MoleculeID: M7024 Public record	logPow Method = RP-HPLC logPow Buffer = methanol RecordID: R13671 09:37, 19 Mar 09 / 12:22, 26 Koerner sz / tanzeem sz
 molecule profile	● logPow = -2.19 (in Log unit) Platts, JA et al The RP-HPLC measurement and QSPR analysis of logP(o/w) value... J. Inorg. Biochem. 2006; 100 (7) 1199-207 MoleculeID: M87460 Public record	logPow Method = RP-HPLC logPow Buffer = methanol RecordID: R13672 09:37, 19 Mar 09 / 12:22, 26 Koerner sz / tanzeem sz
 molecule profile	● logPow = -1.17 (in Log unit) Platts, JA et al The RP-HPLC measurement and QSPR analysis of logP(o/w) value... P: 1204 T: 3 J. Inorg. Biochem. 2006; 100 (7) 1199-207 MoleculeID: M13735 Public record	logPow Method = RP-HPLC logPow Buffer = methanol RecordID: R13673 09:37, 19 Mar 09 / 12:22, 26 Koerner sz / tanzeem sz
 molecule profile	● logPow = -1.7 (in Log unit) Platts, JA et al The RP-HPLC measurement and QSPR analysis of logP(o/w) value... J. Inorg. Biochem. 2006; 100 (7) 1199-207 MoleculeID: M3768443 Public record	logPow Method = RP-HPLC logPow Buffer = methanol RecordID: R13676 09:37, 19 Mar 09 / 12:22, 26 Koerner sz / tanzeem sz

Model for LogP for Pt complexes

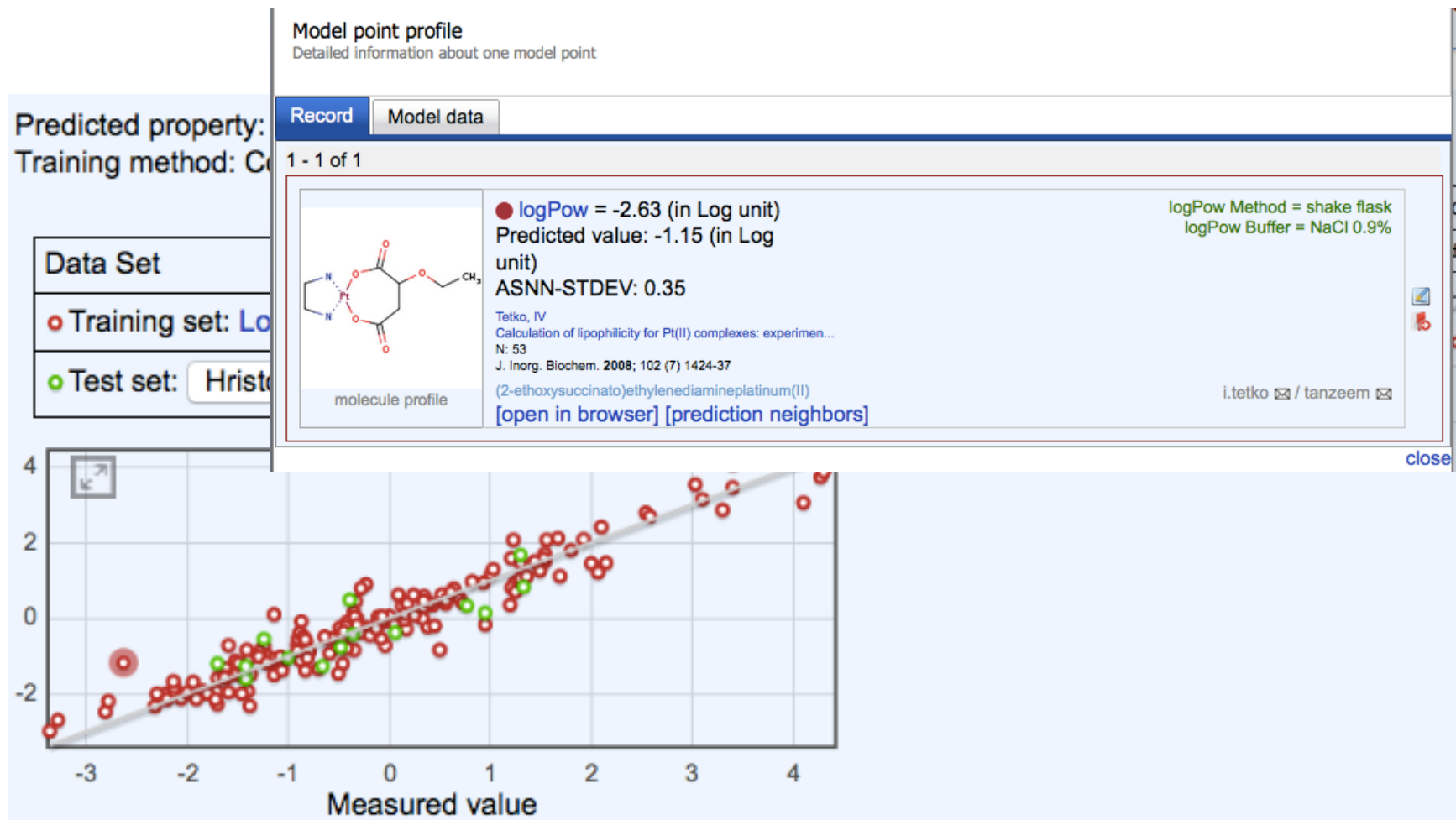
Predicted property: **logPow**

Training method: Consensus

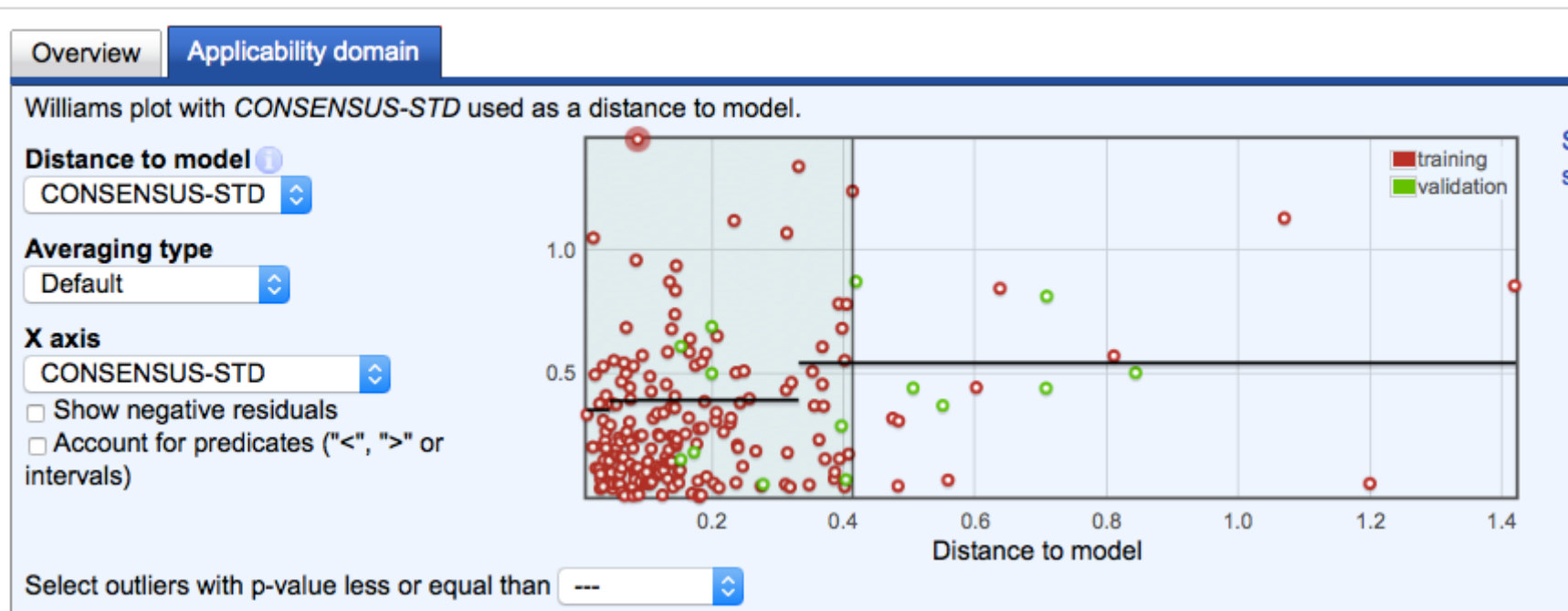
Data Set	#	R ²	q ²	RMSE	MAE
● Training set: LogPt final	187 records	0.93 ± 0.01	0.93 ± 0.01	0.41 ± 0.03	0.3 ± 0.02
● Test set: Hristo	14 records	0.76 ± 0.1	0.76 ± 0.1	0.5 ± 0.06	0.43 ± 0.07



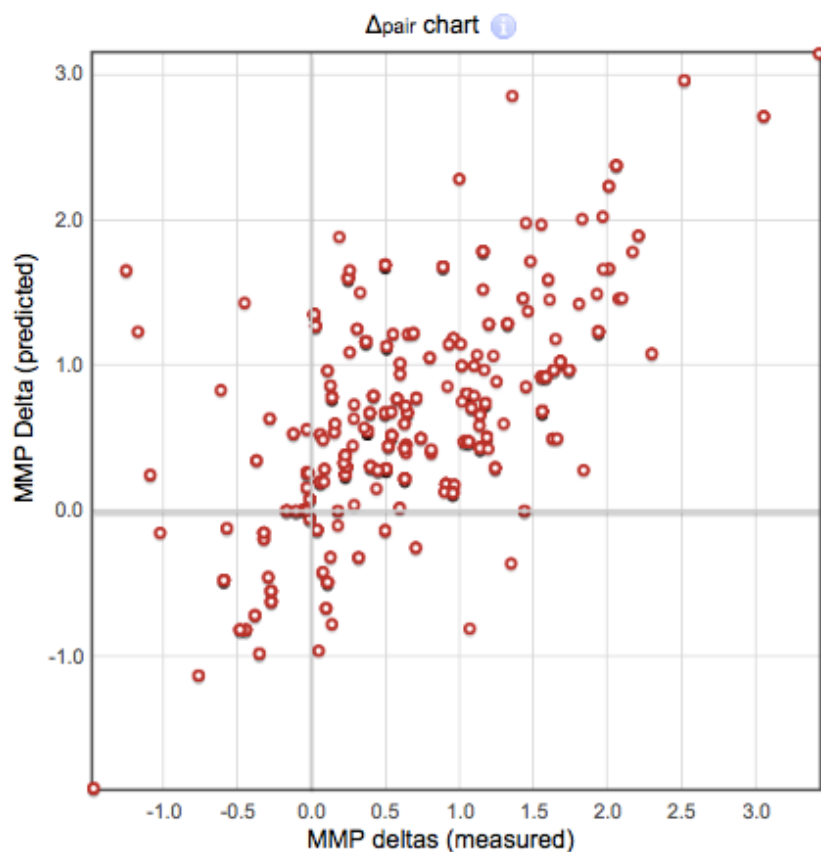
Model for LogP for Pt complexes



Outlying point on the Applicability Domain plot



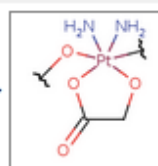
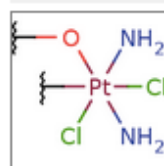
Prediction of logP for Pt complexes



Transformations: minimum 4 pairs, p-value Any

1 - 5 of 60

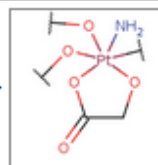
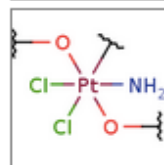
5 items on page 1 of 12 > >>



6 matched pairs

$\Delta_{\text{mean}} = -0.27 \pm 0.19$

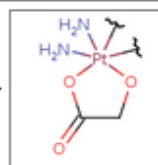
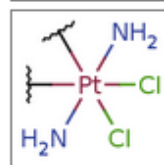
SMIRKS: N[Pt](N)(*)(Cl)(Cl)O* -> N[Pt]1(N)(*)(O*)OCC(=O)O1



6 matched pairs

$\Delta_{\text{mean}} = -0.27 \pm 0.19$

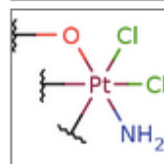
SMIRKS: N[Pt](*)(Cl)(Cl)(O*)O* -> N[Pt]1(*) (O*)(O*)OCC(=O)O1



7 matched pairs

$\Delta_{\text{mean}} = -0.24 \pm 0.2$

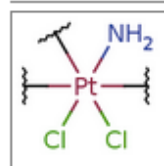
SMIRKS: N[Pt](N)(*)(*)(Cl)Cl -> N[Pt]1(N)(*)(*)(OCC(=O)O)1



7 matched pairs

$\Delta_{\text{mean}} = -0.24 \pm 0.2$

SMIRKS: N[Pt](*)(*)(Cl)(Cl)O* -> N[Pt]1(*) (*)(O*)OCC(=O)O1

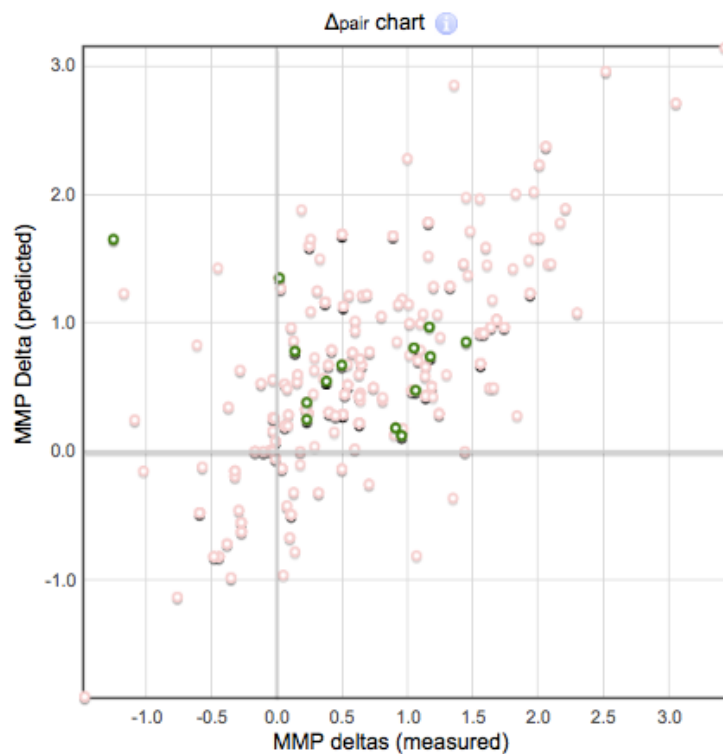


8 matched pairs

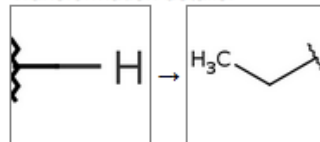
$\Delta_{\text{mean}} = -0.17 \pm 0.26$

SMIRKS: N[Pt](*)(*)(*)(Cl)Cl -> N[Pt]1(*) (*)(*)(OCC(=O)O)1

Outliers of the model (functional groups descriptors)

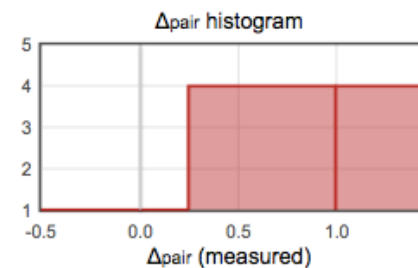


Transformation details:



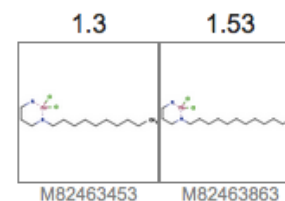
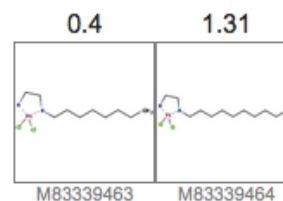
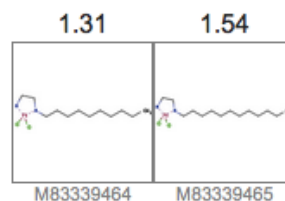
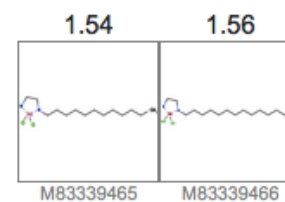
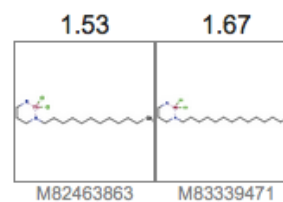
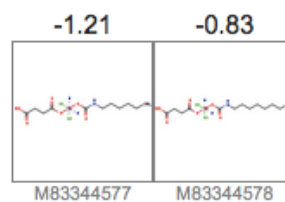
Transformation ID: **TR181**
SMIRKS: *[H] -> CC*

[Back to all transformations](#)

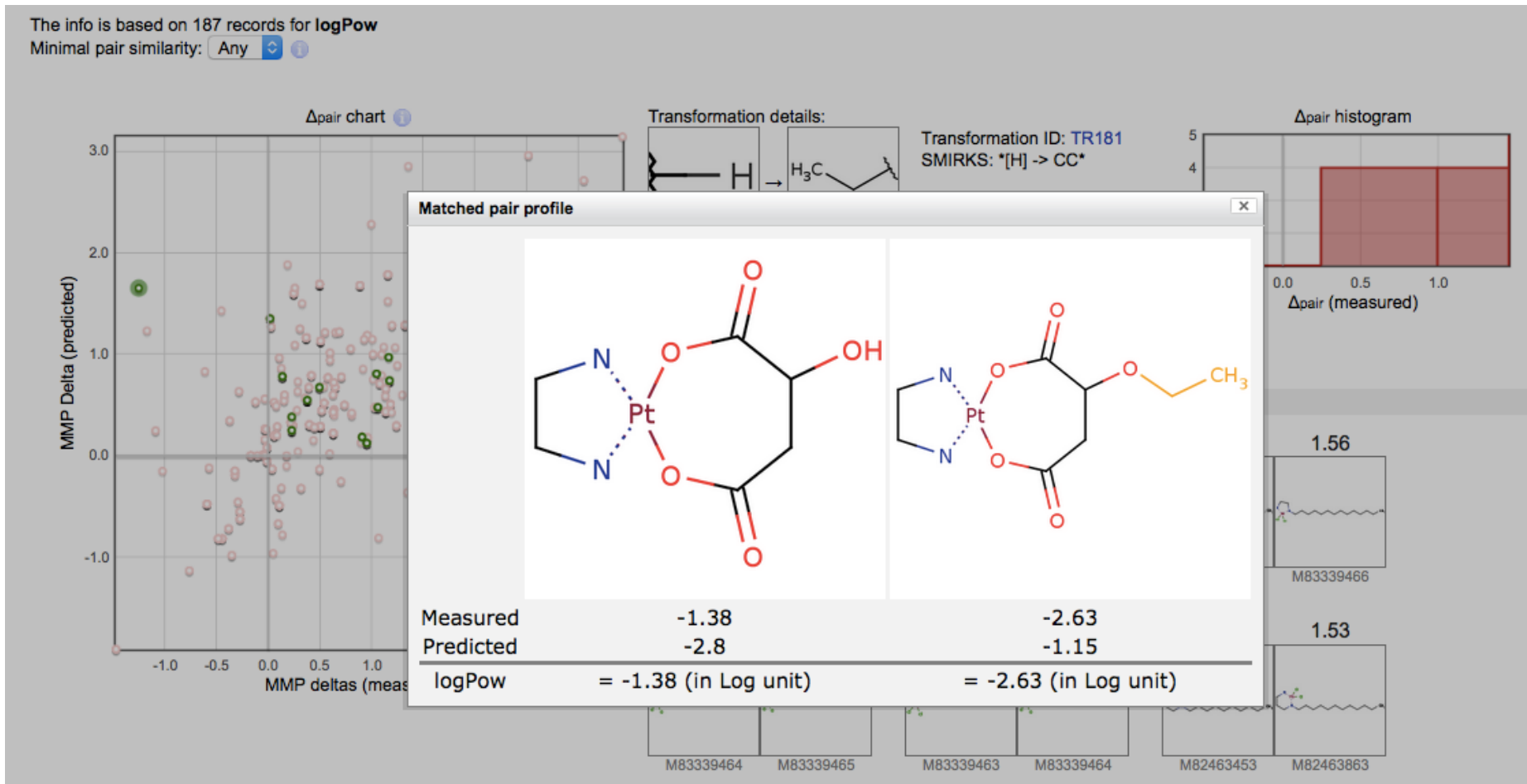


Pairs having the same transformation as the selected one:

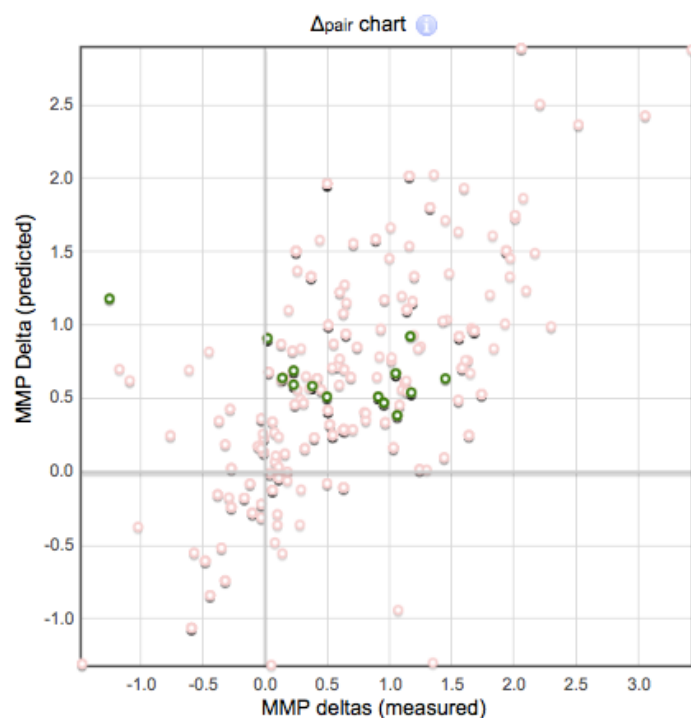
1 - 14 of 14



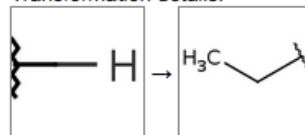
Outliers of the model (Functional groups descriptors)



Outliers of the model (CDK descriptors)

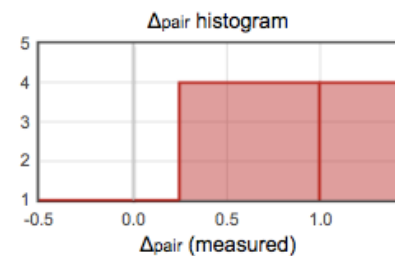


Transformation details:



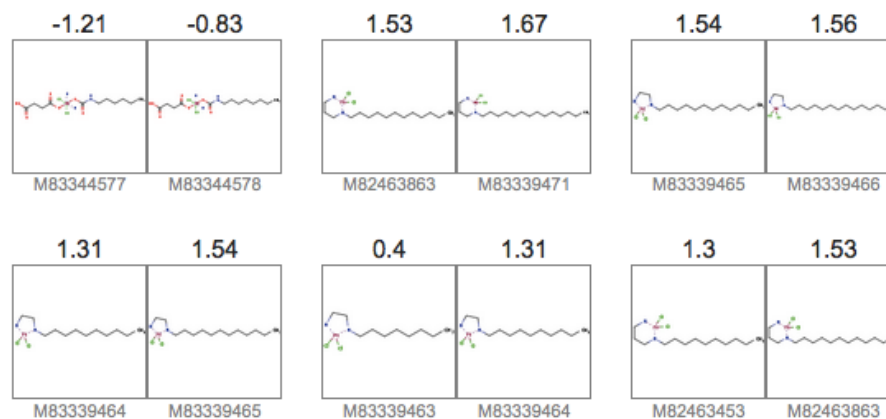
Transformation ID: [TR181](#)
SMIRKS: *[H] -> CC*

[Back to all transformations](#)

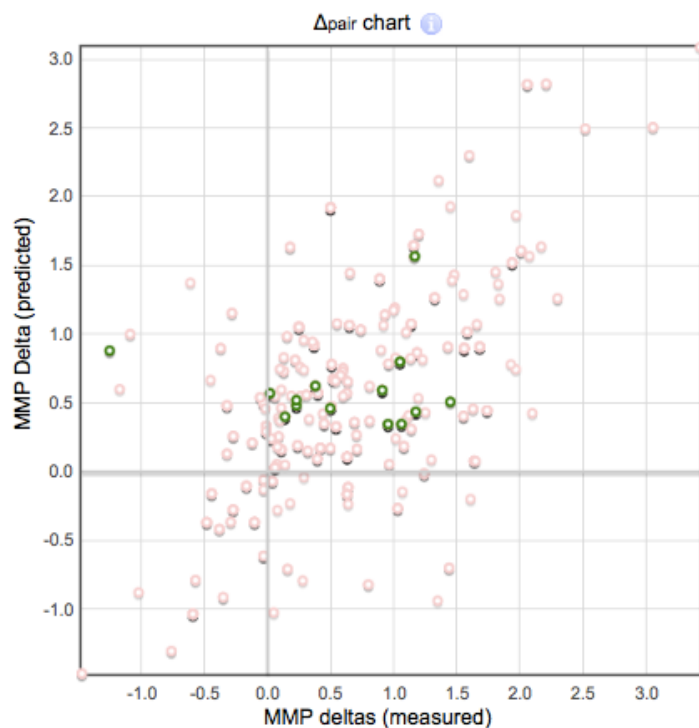


Pairs having the same transformation as the selected one:

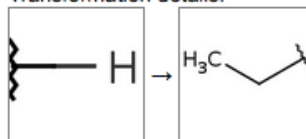
1 - 14 of 14



Outliers of the model (Estate descriptors)

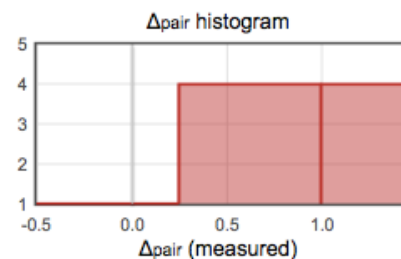


Transformation details:



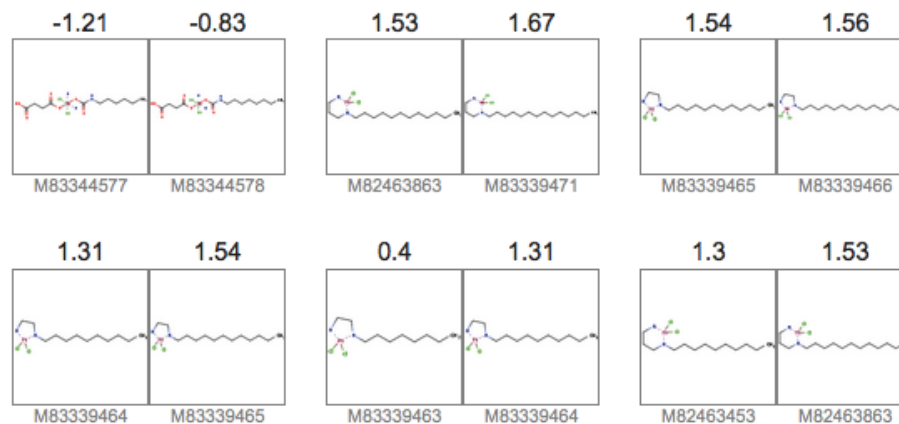
Transformation ID: [TR181](#)
SMIRKS: *[H] -> CC*

[Back to all transformations](#)



Pairs having the same transformation as the selected one:

1 - 14 of 14



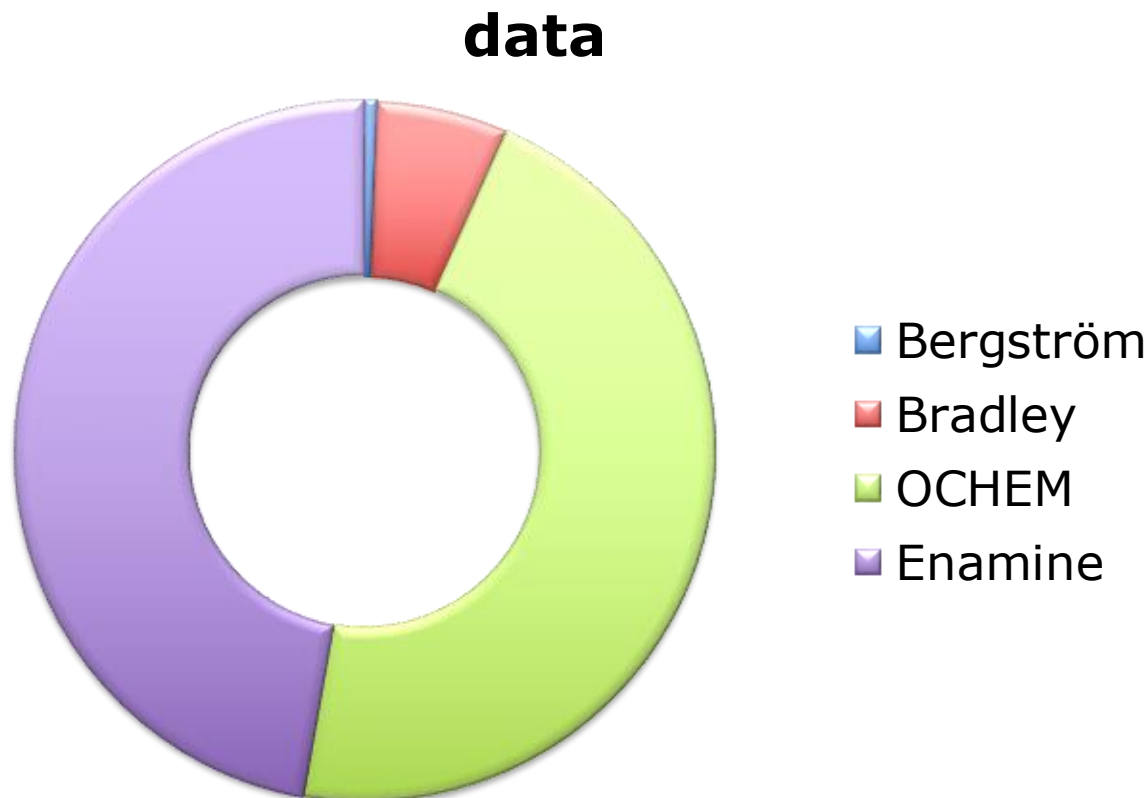
Model predictions (Bergström set)

Bergström 277

Bradley 2886

OCHEM 22404

Enamine 21883



Model predictions (Bergström set)

Method	RMSE	MAE	R ²
Estate	36 ± 2	31 ± 3	0.49 ± 0.09
Consensus	34 ± 1	26 ± 2	0.63 ± 0.07
Karthikeyan et al	41.4		0.66
Bergström et al	49.8		0.54
Nigsch et al	46.3		0.30
Hughes et al	52.8	40.7	0.46

Bergstrom, et al *J. Chem. Inf. Comput. Sci.* **2003**, 43, 1177-1185.

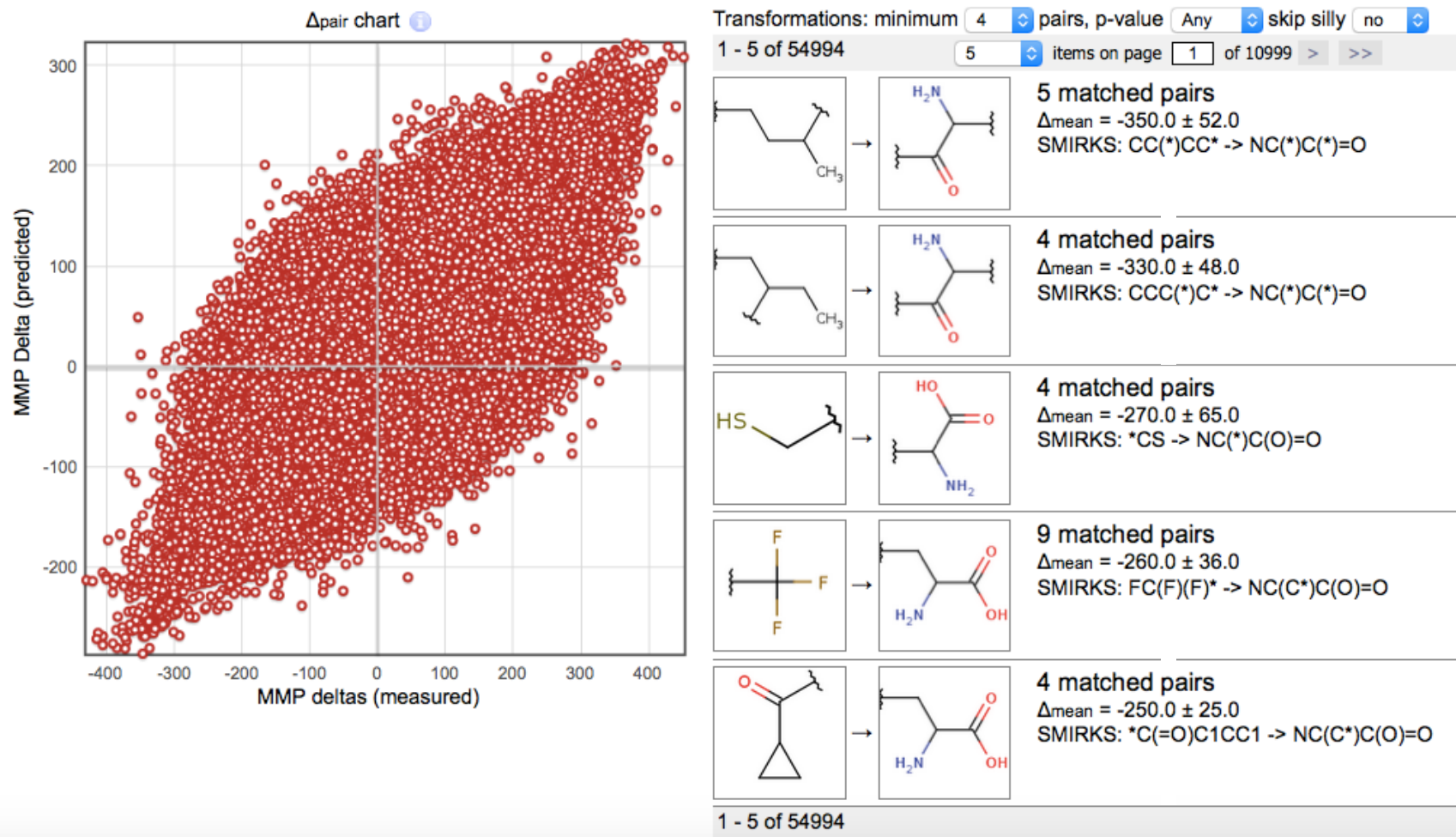
Karthikeyan, et al *J. Chem. Inf. Model.* **2005**, 45, 581-590.

Nigsch, et al *J. Chem. Inf. Model.* **2006**, 46, 2412-2422.

Hughes et al *J. Chem. Inf. Model.* **2008**, 48, 220-232.

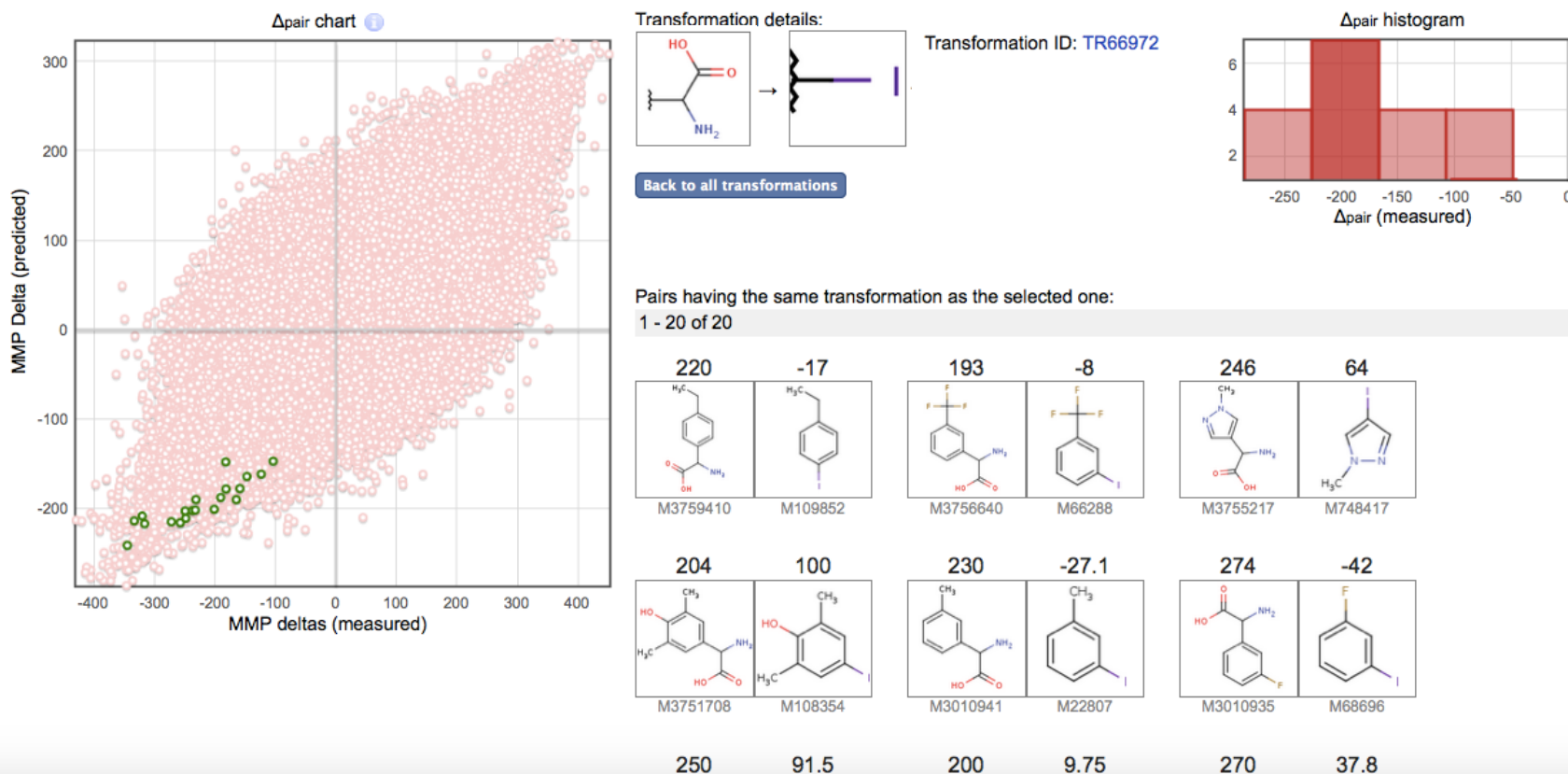
Tetko, I. V.; M Lowe, D.; Williams, A. J. The Development of Models to Predict Melting and Pyrolysis Point Data Associated with Several Hundred Thousand Compounds Mined from PATENTS. *J. Cheminformatics* **2016**, 8, 2. <https://doi.org/10.1186/s13321-016-0113-y>.

Predicted MMPs for Melting Point



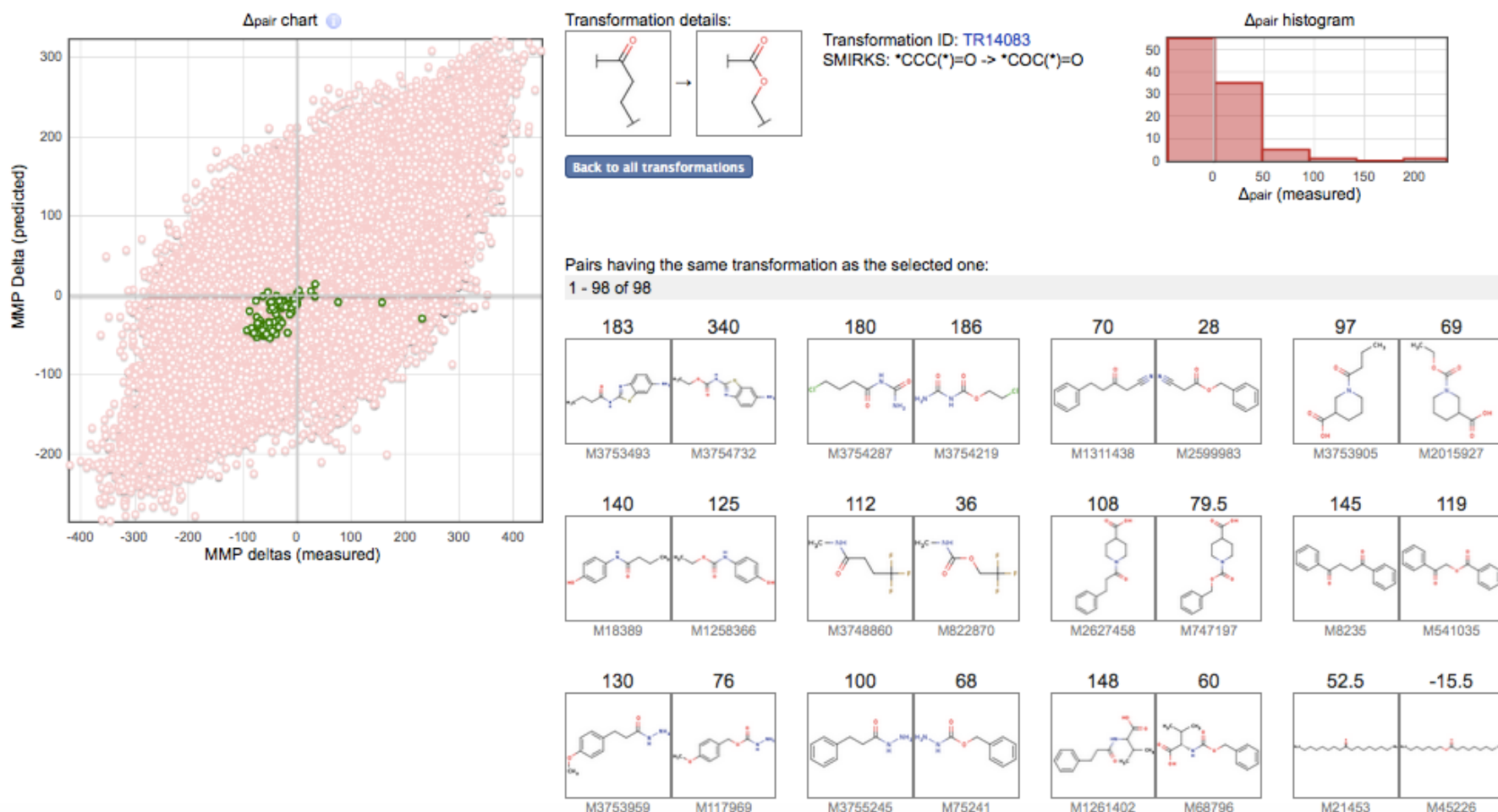
Withnall, M.; Chen, H.; Tetko, I. V. Matched Molecular Pair Analysis on Large Melting Point Datasets: A Big Data Perspective. *ChemMedChem* **2018**, 13 (6), 599–606. <https://doi.org/10.1002/cmdc.201700303>.

An example of MMP transformation



Withnall, M.; Chen, H.; Tetko, I. V. Matched Molecular Pair Analysis on Large Melting Point Datasets: A Big Data Perspective. *ChemMedChem* **2018**, 13 (6), 599–606. <https://doi.org/10.1002/cmdc.201700303>.

An example of MMP transformation



Withnall, M.; Chen, H.; Tetko, I. V. Matched Molecular Pair Analysis on Large Melting Point Datasets: A Big Data Perspective. *ChemMedChem* **2018**, 13 (6), 599–606. <https://doi.org/10.1002/cmdc.201700303>.

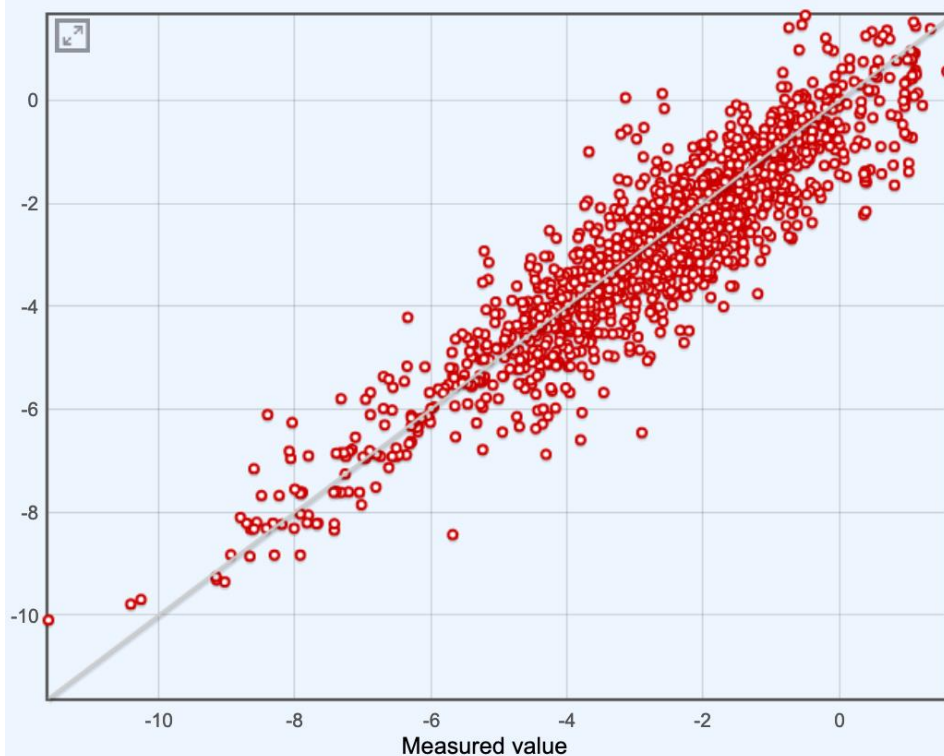
Prediction of solubility using Yalkowsky's equation

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

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

Training method: MLRA

Data Set	#	R2	q2	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](#)

1/2

Conclusions

- OCHEM is powerful platform for model development
- It provides best strategies for model development
- Cautions should be taken to avoid overfitting
- Validation techniques should match intended model use
- Use interpretable descriptors to interpret models
- MMP are useful for model interpretation

Acknowledgements



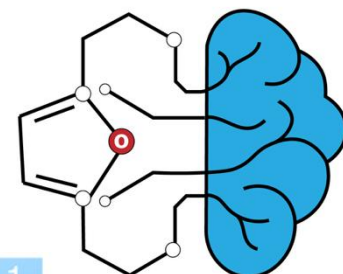
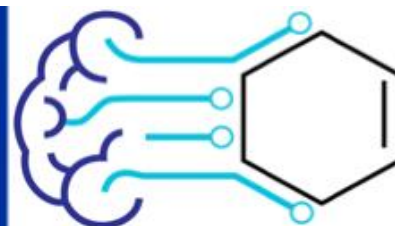
Katya Ahmad
Fabian Krüger
Dina Khasanova
Mark Embrechts
Michael Sattler

Peter Hartog
Thalita Cirino
Nesma Mousa
Massa Zahdeh
Andi Hunklinger
Alessandro Dipalma
Paula Torren-Peraire
Varvara Voinarovska

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



Alexander von Humboldt
Stiftung / Foundation



AiChemist