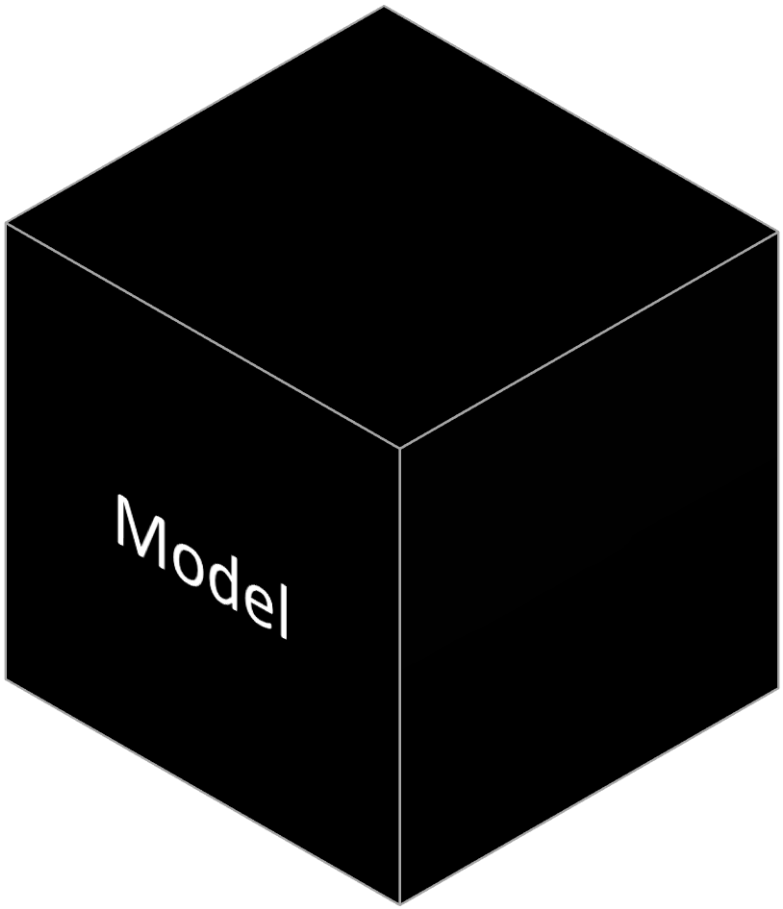
Explainable artificial intelligence: evolution, achievements and perspectives

Pavel Polishchuk

Institute of Molecular and Translational Medicine
Faculty of Medicine and Dentistry
Palacky University

pavlo.polishchuk@upol.cz


Model

plant growth inhibition activity of phenoxyacetic acids

$$1/C = 4.08\pi - 2.14\pi^2 + 2.78\sigma + 3.38$$

Hansch equation

rate of penetration of membranes in the plant cell

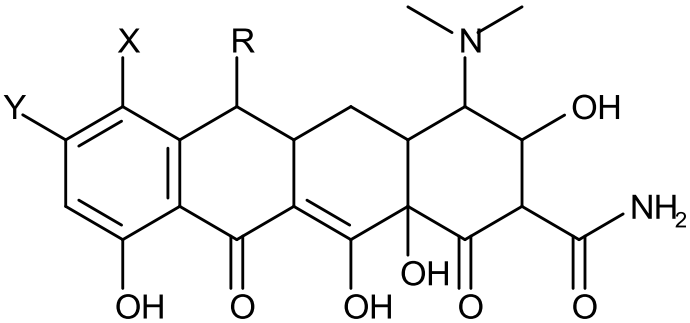
electronic factors

$\pi = \log P_X - \log P_H$

σ - Hammet constant

Hansch, C.; Fujita, T., ρ-σ-π Analysis. A Method for the Correlation of Biological Activity and Chemical Structure. *Journal of American Chemical Society* **1964**, 86, 1616-1626.

Free-Wilson models



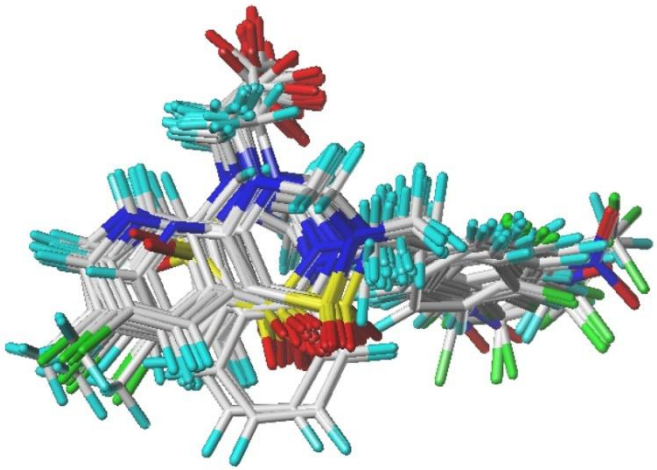
Inhibition activity of compounds against *Staphylococcus aureus*

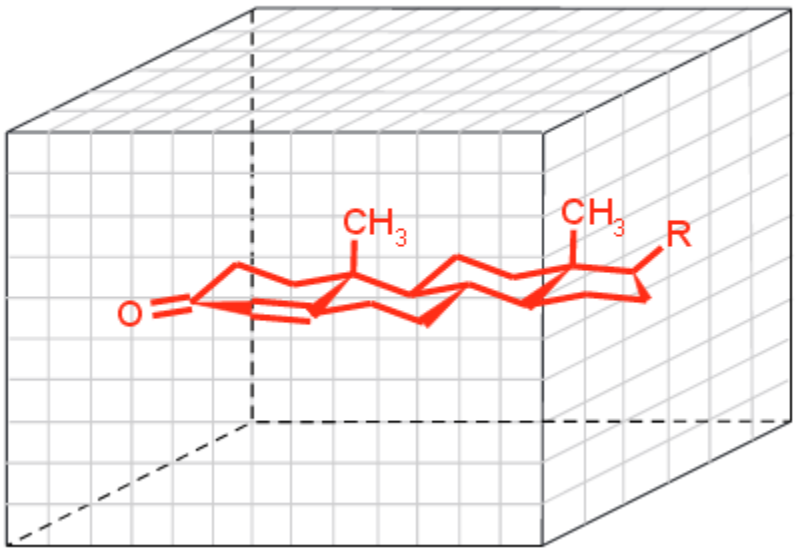
R is H or CH_3;
X is Br, Cl, NO_2 and
Y is NO_2, NH_2, $NHC(=O)CH_3$

$$Act = 75R_H - 112R_{CH3} + 84X_{Cl} - 16X_{Br} - 26X_{NO2} + 123Y_{NH2} + 18Y_{NHC(=O)CH3} - 218Y_{NO2}$$

Free, S. M.; Wilson, J. W., A Mathematical Contribution to Structure-Activity Studies. *Journal of Medicinal Chemistry* **1964**, 7, 396-399.

CoMFA: Comparative molecular field analysis





probe: electrostatic (H^+) steric (Csp^3)

PLS model

Y
2.4
3.7
…
8.1

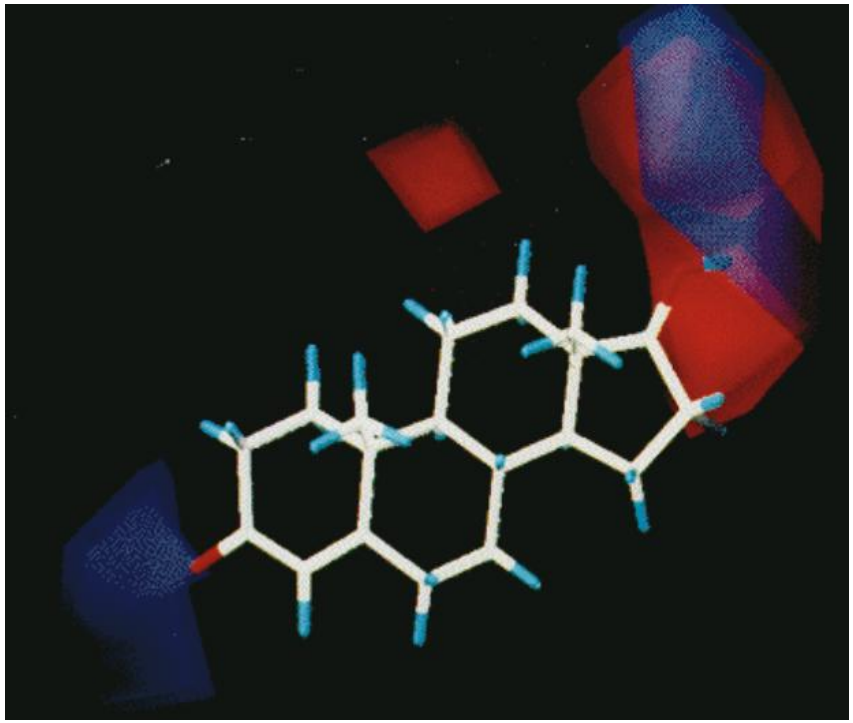
Mol	Electrostatic field descriptors								Steric field descriptors							
	(0,0,0)	(0,0,1)	…	…	…	…	…	(G,G,G)	(0,0,0)	(0,0,1)	…	…	…	…	…	(G,G,G)
1																
2																
…																
N																

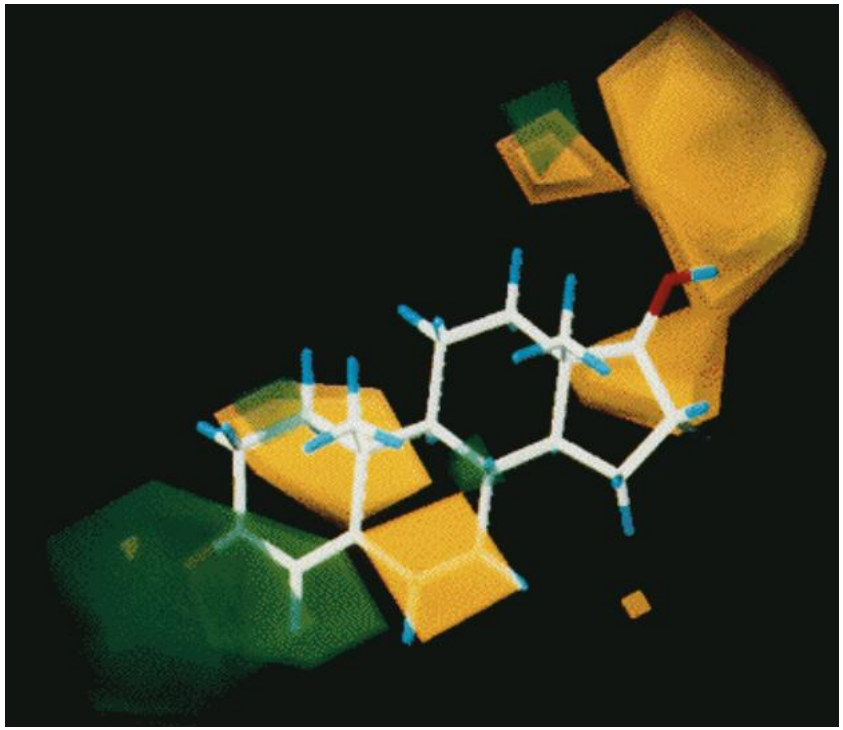
Cramer, R. D.; Patterson, D. E.; Bunce, J. D. Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *Journal of the American Chemical Society* **1988,** *110*, 5959-5967

CoMFA: Comparative molecular field analysis

$$Y = \sum_{i=1}^{n} b_i x_i + c$$

b – contribution of steric or electrostatic field in a particular cell

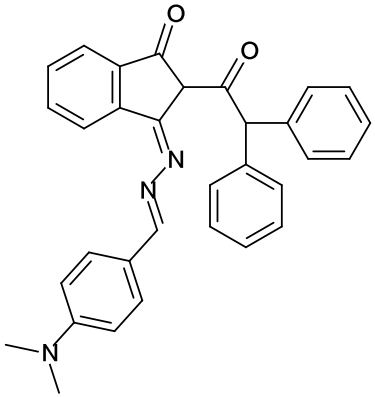


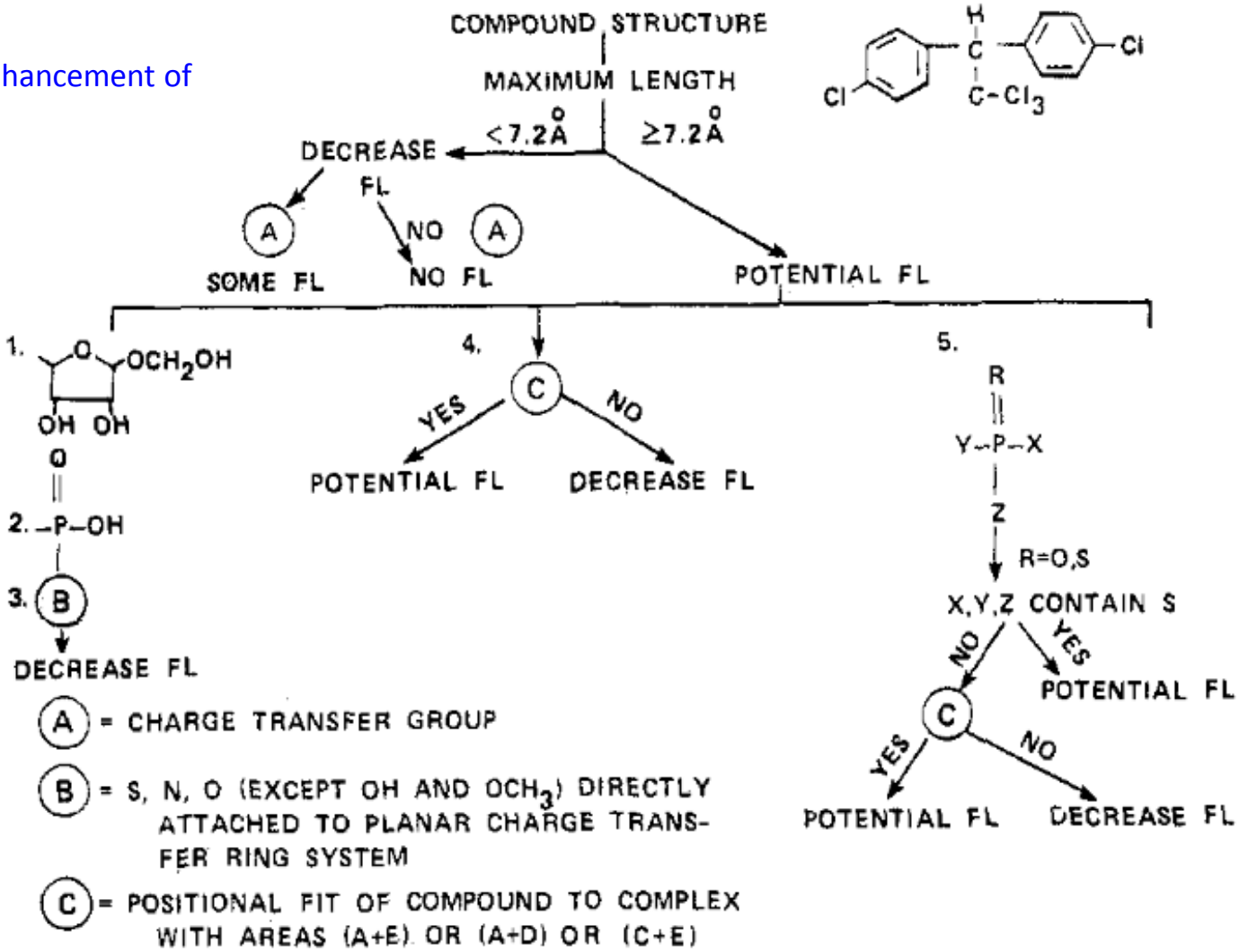


Cramer, R. D.; Patterson, D. E.; Bunce, J. D. Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *Journal of the American Chemical Society* **1988,** *110*, 5959-5967

Decision tree

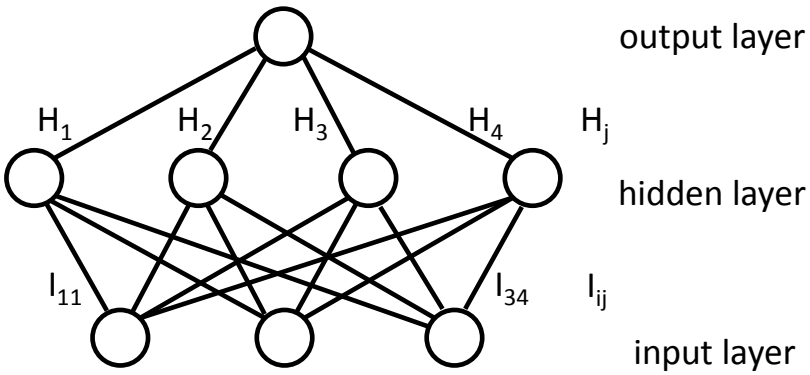
solid-phase fluorescence enhancement of





Ashman, W. P.; Lewis, J. H.; Poziomek, E. J., Decision tree for chemical detection applications. *Analytical Chemistry* **1985**, 57, 1951-1955

Neural networks: relative importance



$$P_{ij} = |I_{ij}| \times |H_j|$$

$$Q_{ij} = \frac{P_{ij}}{\sum_i P_{ij}}$$

$$S_i = \sum_j Q_{ij}$$

$$\text{relative importance}_i = \frac{S_i}{\sum_i S_i}$$

Garson, G. D. Interpreting neural-network connection weights. *AI Expert* **1991,** *6*, 46-51

adsorbability of 55 organic compounds on activated carbon fibers

$$\log K = 3.33 - 1.55\,{}^3\chi^v + 0.58\,{}^5\chi^v + 3.52\,{}^6\chi^v - 1.42\,{}^3\chi_c + 2.29\,{}^4\chi^v_{pc}$$

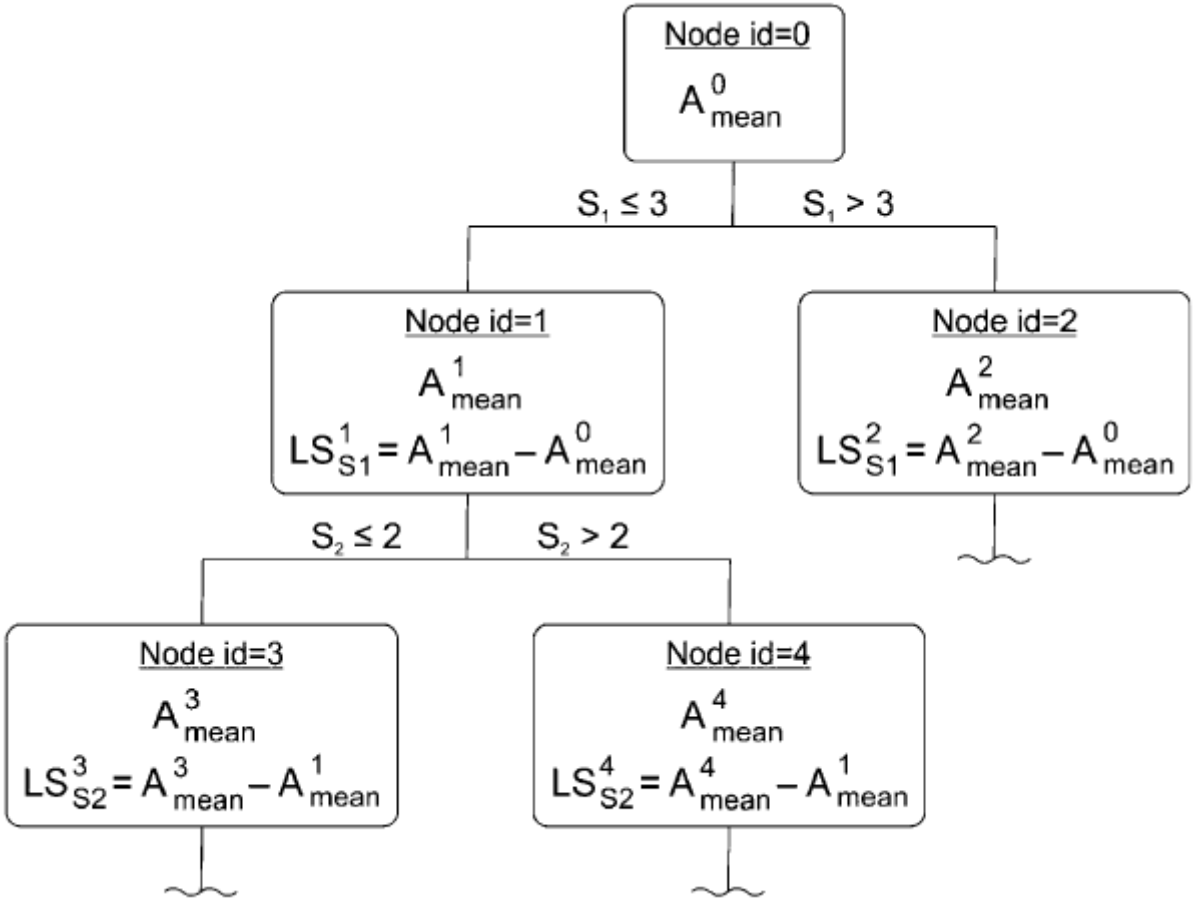
$n = 49$, $R^2_{adj} = 0.648$, SE = 0.199

	${}^3\chi^v$	${}^5\chi^v$	${}^6\chi^v$	${}^3\chi_c$	${}^4\chi^v_{pc}$
relative importance, %	20.3	17.3	34.4	11.6	16.4
influence on logK	↓	↑	↑	↓	↑

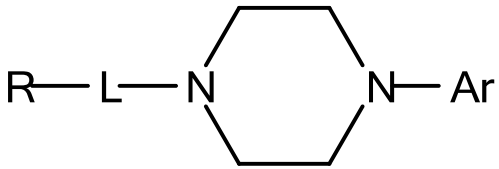
${}^3\chi^v$ is related to bulky and branched molecules
${}^5\chi^v$ is related to heteroatomic contents
${}^6\chi^v$ is related to highly branched compounds, like atrazin
${}^3\chi_c$ is related to highly substituted compounds comprising *tert*-butyl groups or having more than three substituents
${}^4\chi^v_{pc}$ is related to compounds with more than four substituents

Brasquet, C.; Bourges, B.; Le Cloirec, P. Quantitative Structure–Property Relationship (QSPR) for the Adsorption of Organic Compounds onto Activated Carbon Cloth: Comparison between Multiple Linear Regression and Neural Network. *Environ. Sci. Technol.* **1999,** *33*, 4226-4231.

Random Forest: descriptor contributions



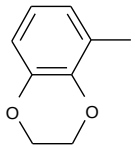
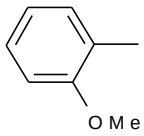
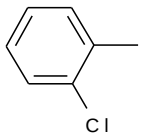
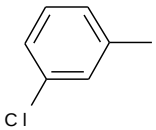
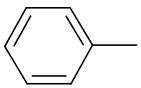
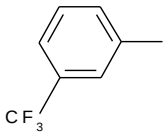
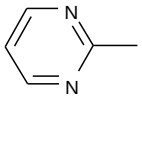
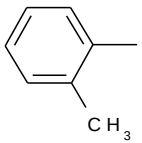
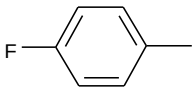
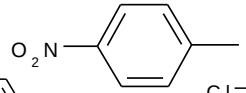
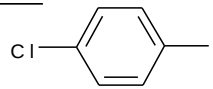
Kuz'min, V. E.; Polishchuk, P. G.; Artemenko, A. G.; Andronati, S. A., Interpretation of QSAR models based on Random Forest method. *Molecular Informatics* **2011, 30, 593-603.**



Ar - substituted (hetero)aryls
L - polymethylene chain
R - various (poly)cyclic residues

347 agonists of 5-HT_{1A} receptor

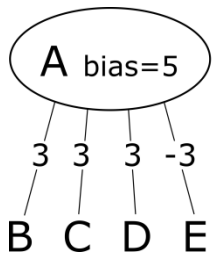
PLS, 72 descriptors, R^2_{5CV} = 0.64
RF, 2500 descriptors, R^2_{OOB} = 0.70

Ar											
PLS	0.84	0.18	0.03	-0.04	-0.06	-0.09	-0.11	-0.66	-0.73	-0.94	-0.96
RF	0.27	0.24	0.04	0.07	-0.02	0.11	0.04	-0.04	-0.55	-0.66	-0.66

L	$-(CH_2)_6-$	$-(CH_2)_5-$	$-(CH_2)_4-$	$-(CH_2)_3-$	$-(CH_2)_2-$	$-CH_2-$
PLS	0.8	0.71	0.81	0.08	-0.04	0.06
RF	0.14	0.19	0.14	-0.01	-0.03	0.05

Kuz'min, V. E.; Polishchuk, P. G.; Artemenko, A. G.; Andronati, S. A., Interpretation of QSAR models based on Random Forest method. *Molecular Informatics* **2011, 30, 593-603.**

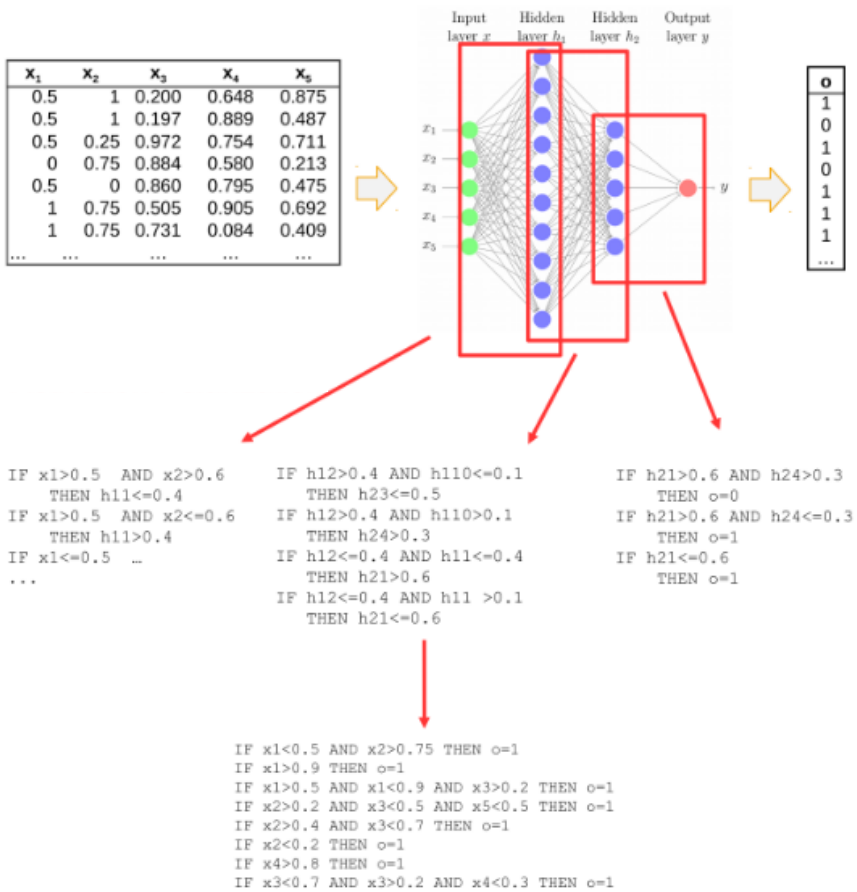
Rule extraction approaches
Decompositional (use knowledge about internal structure of a model, e.g. NN)



if B, C and not(E), then A
if B, D and not(E), then A
if C, D and not(E), then A
if B, C, D, then A

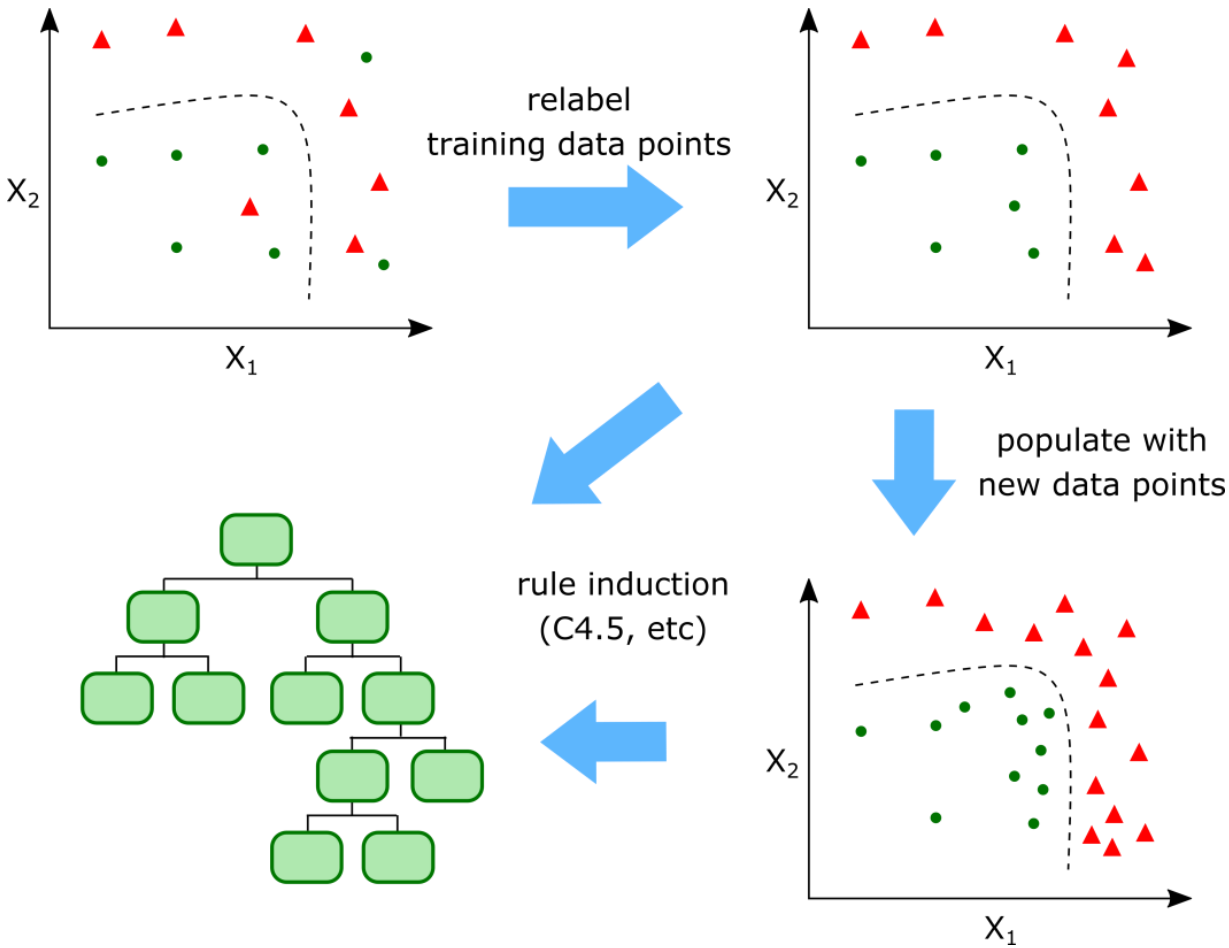
Fu, L., Rule learning by searching on adapted nets. In *Proceedings of the ninth National conference on Artificial intelligence - Volume 2*, AAAI Press: Anaheim, California, **1991**; pp 590-595.

DeepRED



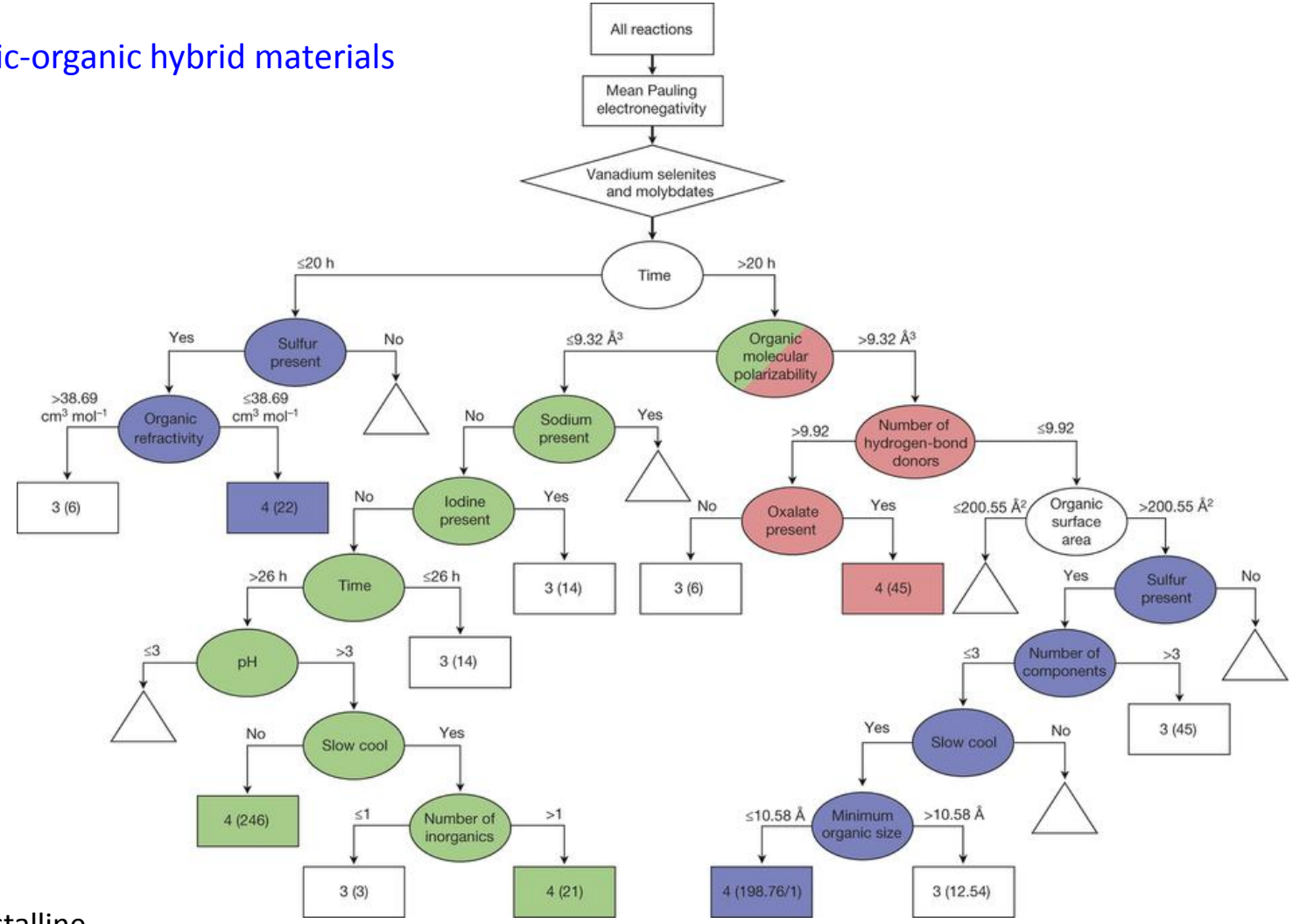
Zilke, J. R.; Loza Mencía, E.; Janssen, F. DeepRED – Rule Extraction from Deep Neural Networks. In *Discovery Science: 19th International Conference, DS 2016, Bari, Italy, October 19–21, 2016, Proceedings*, Calders, T.; Ceci, M.; Malerba, D., Eds. Springer International Publishing: Cham, **2016**; pp 457-473.

Rule extraction approaches
Pedagogical / surrogate modeling (treat a model as a black box)



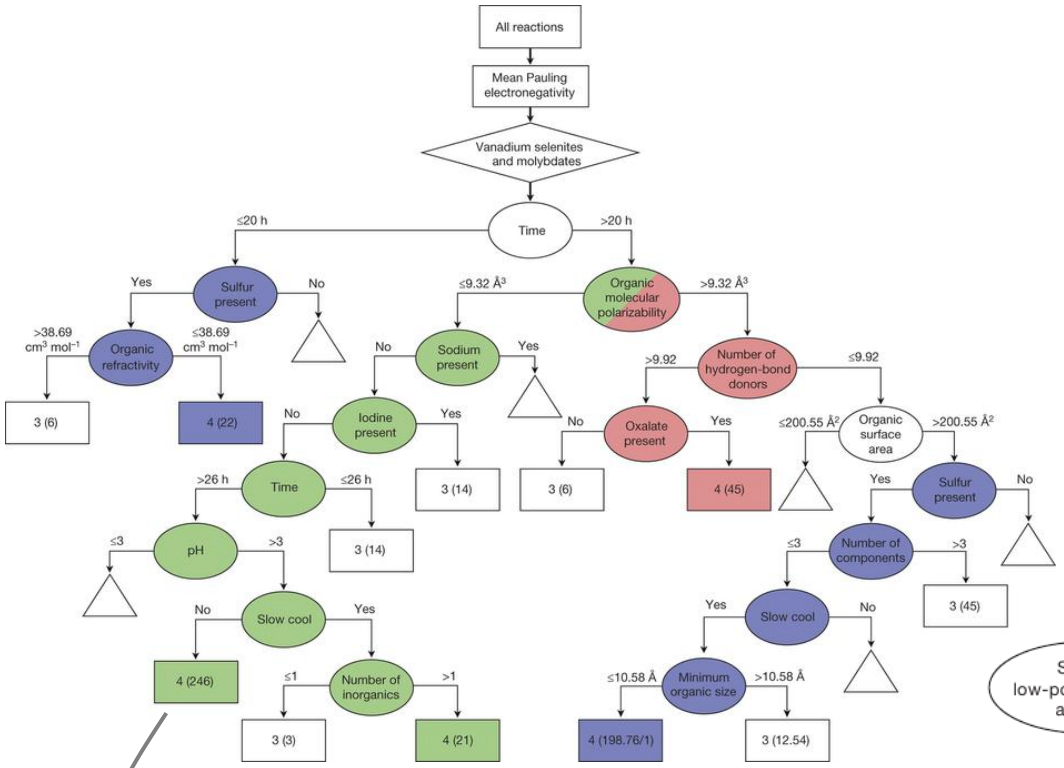
Martens, D.; Baesens, B.; Gestel, T. V., Decompositional Rule Extraction from Support Vector Machines by Active Learning. *IEEE Transactions on Knowledge and Data Engineering* **2009**, 21, 178-191.

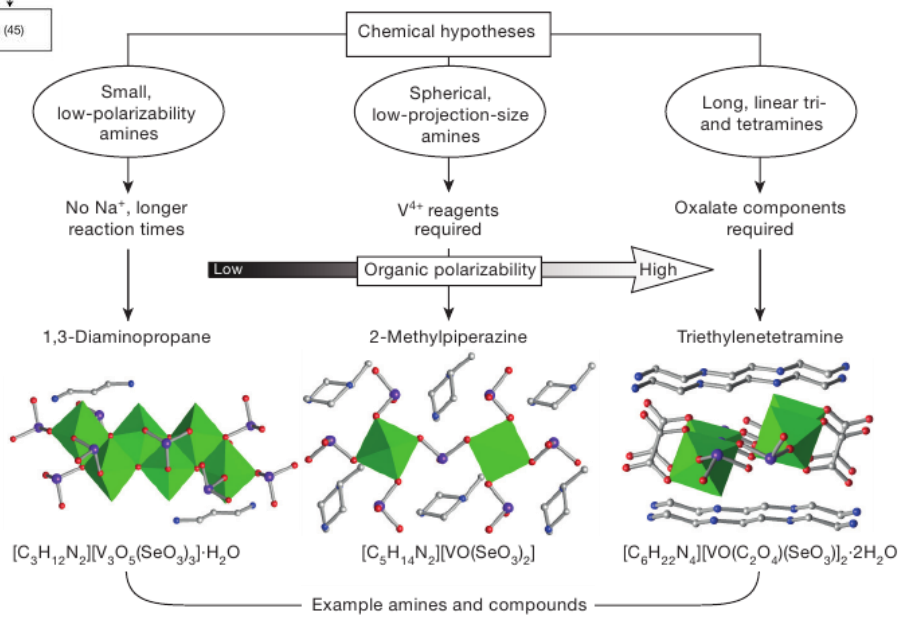
inorganic-organic hybrid materials



output:
3 – polycrystalline
4 – single-crystal

Raccuglia, P.; Elbert, K. C.; Adler, P. D. F.; Falk, C.; Wenny, M. B.; Mollo, A.; Zeller, M.; Friedler, S. A.; Schrier, J.; Norquist, A. J., Machine-learning-assisted materials discovery using failed experiments. *Nature* **2016**, 533, 73.

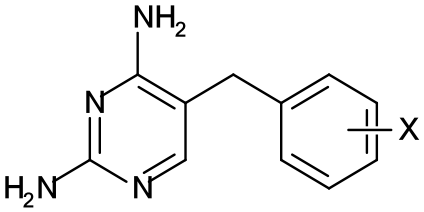




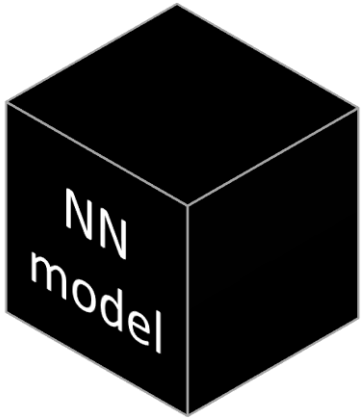
output:
3 – polycrystalline
4 – single-crystal

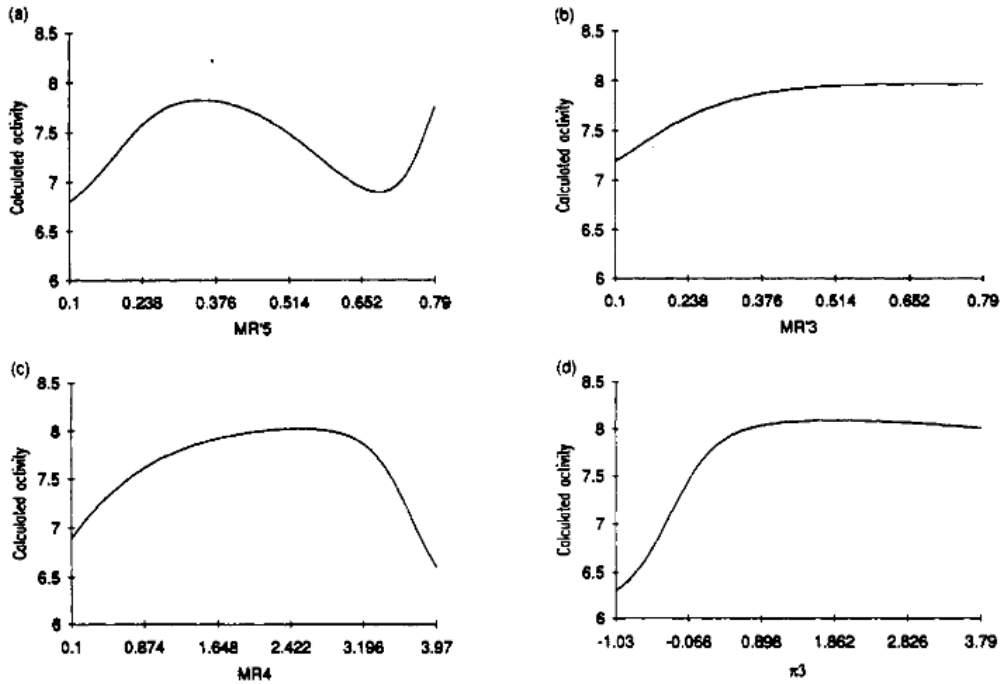
Raccuglia, P.; Elbert, K. C.; Adler, P. D. F.; Falk, C.; Wenny, M. B.; Mollo, A.; Zeller, M.; Friedler, S. A.; Schrier, J.; Norquist, A. J., Machine-learning-assisted materials discovery using failed experiments. *Nature* **2016**, 533, 73.

Sensitivity analysis



DHFR inhibitors





$$\log(1/K) = 0.95MR_5 + 0.89MR_3 + 0.80MR_4 - 0.21MR_4^2 + 1.58\pi_3 - 1.77\log(\beta\times10^{\pi3} + 1) + 6.65$$

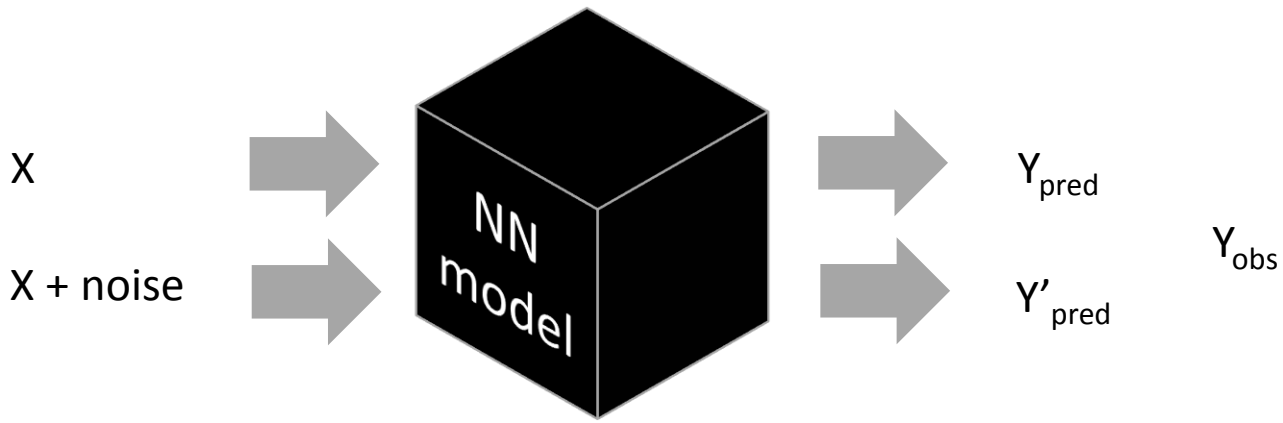
RMSE = 0.093

$$\log(1/K) = 11.79MR_5^3 - 15.74MR_5^2 + 6.55MR_5 + 0.89MR_3 + 0.80MR_4 - 0.21MR_4^2 + 1.58\pi_3 - 1.77\log(\beta\times10^{\pi3} + 1) + 6.24$$

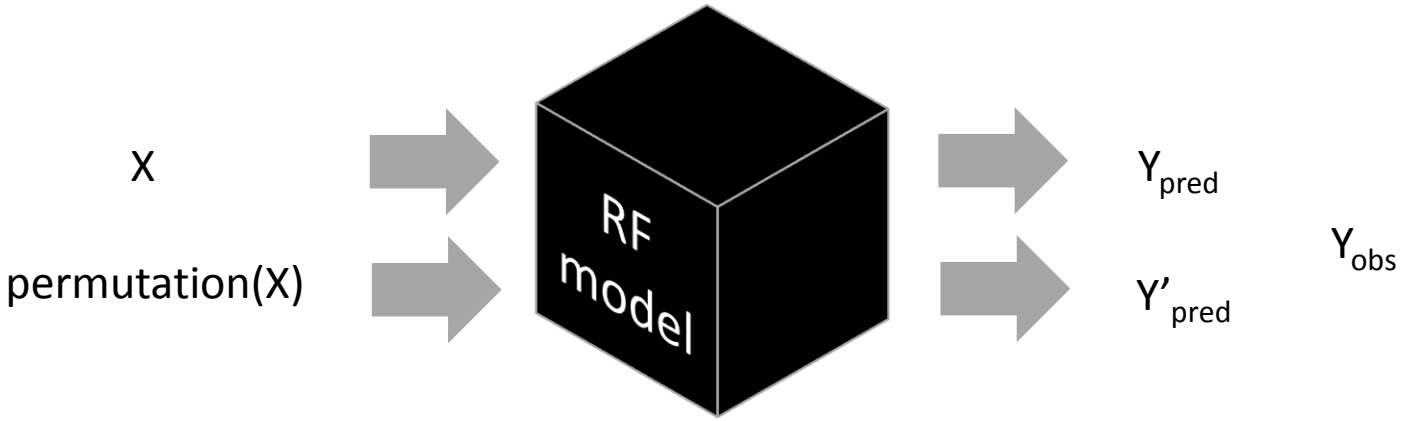
RMSE = 0.074

So, S. S.; Richards, W. G. Application of neural networks: quantitative structure-activity relationships of the derivatives of 2,4-diamino-5-(substituted-benzyl)pyrimidines as DHFR inhibitors. *J. Med. Chem.* **1992,** *35*, 3201-3207.

Variable importance



Györgyi, G. Inference of a rule by a neural network with thermal noise. *Phys. Rev. Lett.* **1990,** *64*, 2957-2960.



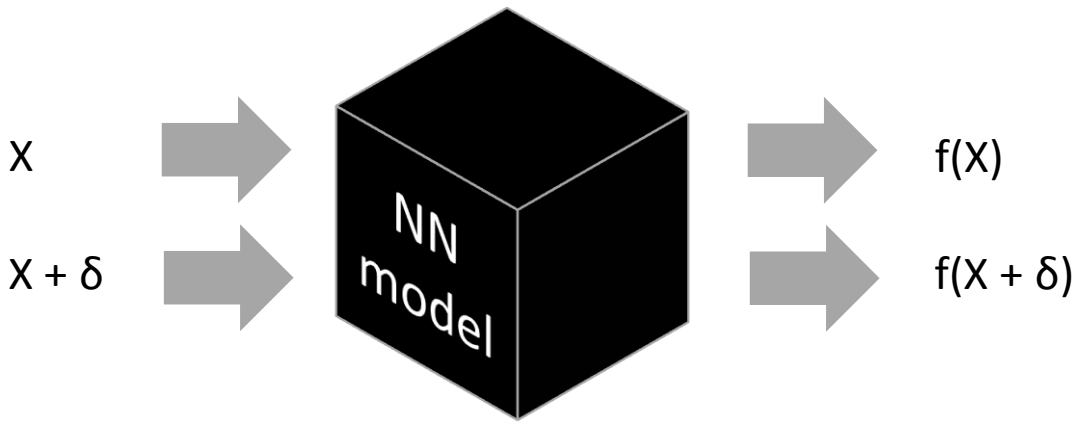
Breiman, L., Random Forests. *Machine Learning* **2001**, 45, 5-32.

Partial derivatives

$$\frac{\partial Y}{\partial x_i} = \sum_j f'(y_j) I_{ij} g'(y) H_j$$

analytical derivatives

Aoyama, T.; Ichikawa, H. Neural networks as nonlinear structure-activity relationship analyzers. Useful functions of the partial derivative method in multilayer neural networks. *J. Chem. Inf. Comput. Sci.* **1992,** *32*, 492-500.



$$C = \frac{f(x + \delta) - f(x)}{\delta}$$

forward difference

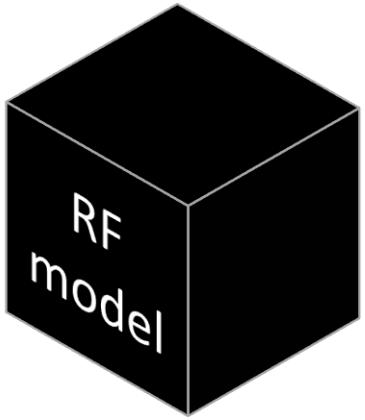
$$C = \frac{f(x) - f(x - \delta)}{\delta}$$

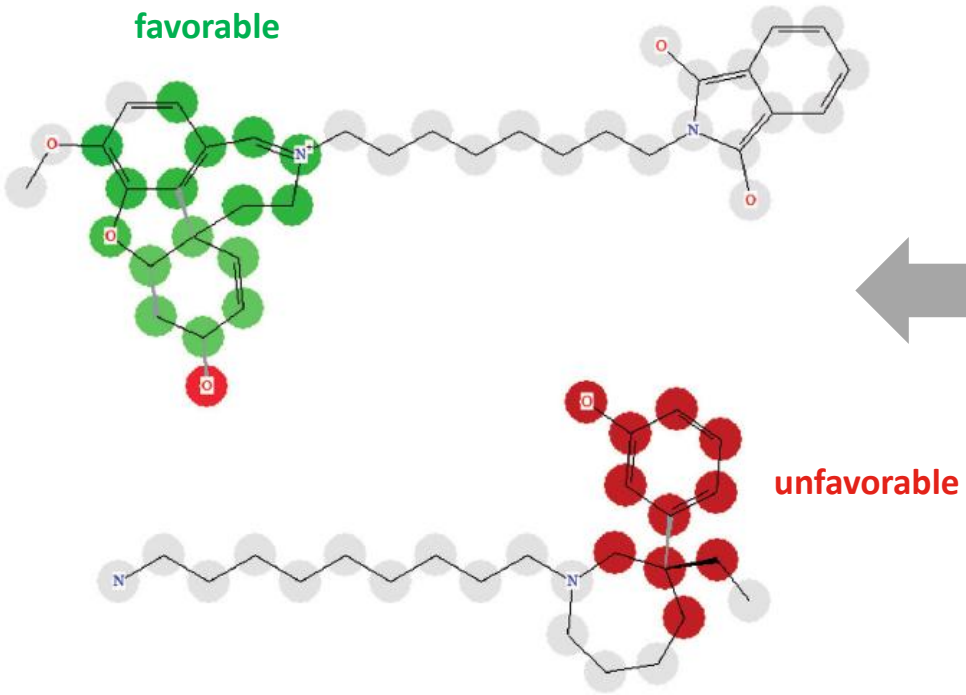
backward difference

$$C = \frac{f(x + \delta/2) - f(x - \delta/2)}{\delta}$$

central difference

AChE inhibitors → fragmental descriptors

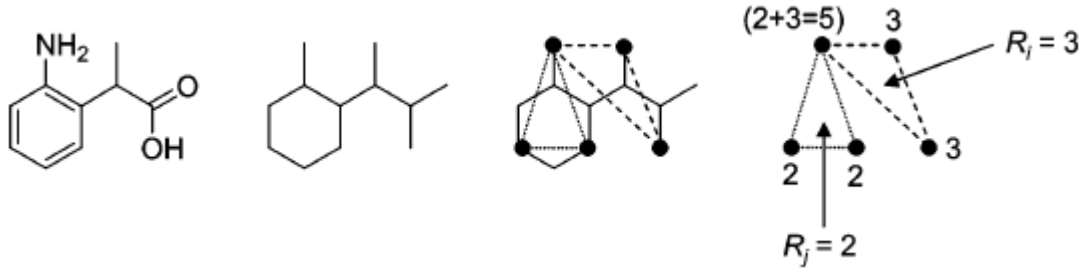


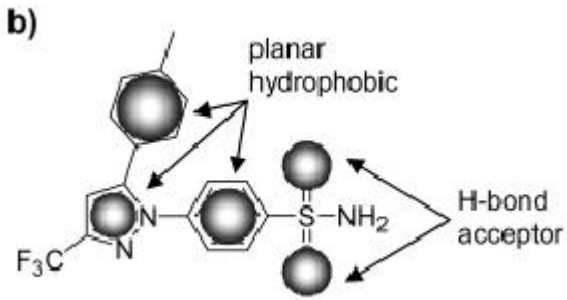


Marcou, G.; Horvath, D.; Solov'ev, V.; Arrault, A.; Vayer, P.; Varnek, A. Interpretability of SAR/QSAR Models of any Complexity by Atomic Contributions. *Molecular Informatics* **2012,** *31*, 639-642

COX2 inhibitors

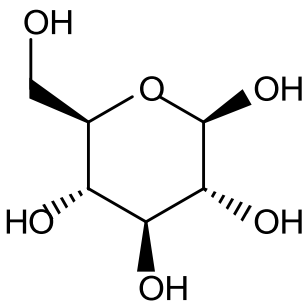
$$R_i = f(\mathbf{x}(F_i = 1)) - f(\mathbf{x}(F_i = 0))$$



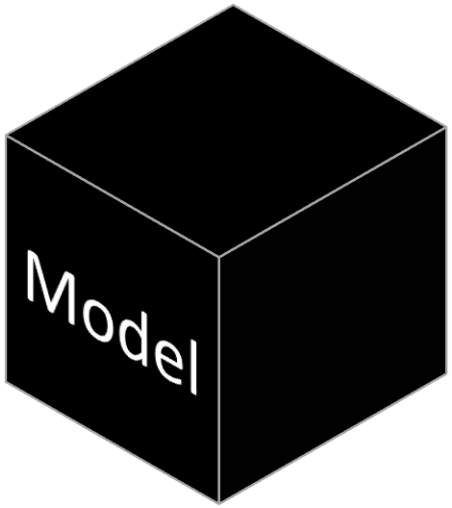


Franke, L.; Byvatov, E.; Werz, O.; Steinhilber, D.; Schneider, P.; Schneider, G., Extraction and Visualization of Potential Pharmacophore Points Using Support Vector Machines: Application to Ligand-Based Virtual Screening for COX-2 Inhibitors. *Journal of Medicinal Chemistry* **2005**, 48, 6997-7004.

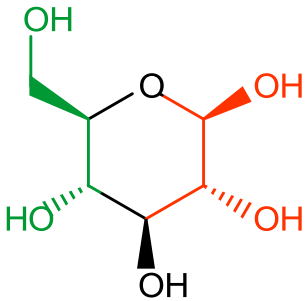
Learning



D_1	D_2	D_3	…	D_N
1	0	9	…	1
4	0	1	…	1
0	2	3	…	3
…	…	…	…	…
4	0	0	…	1



Output



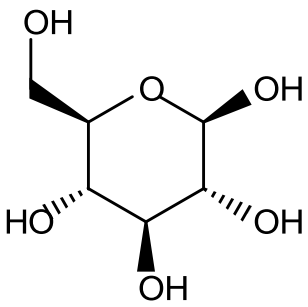
(optional)

Descriptor contributions

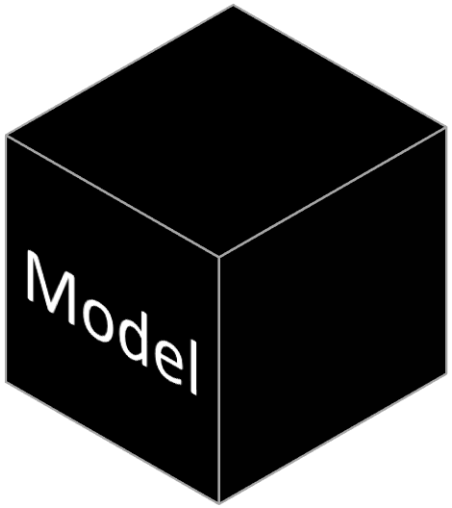
Interpretation

“model → descriptors → structure” paradigm

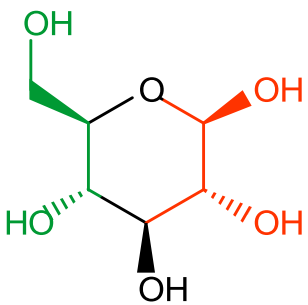
Learning



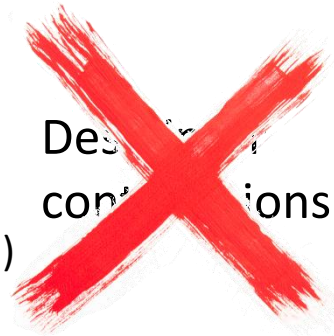
D_1	D_2	D_3	…	D_N
1	0	9	…	1
4	0	1	…	1
0	2	3	…	3
…	…	…	…	…
4	0	0	…	1



Output



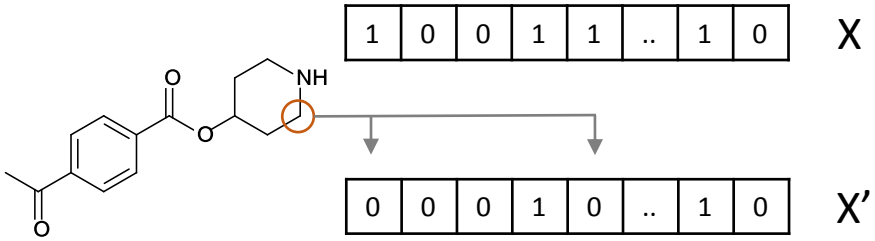
(optional)



Interpretation

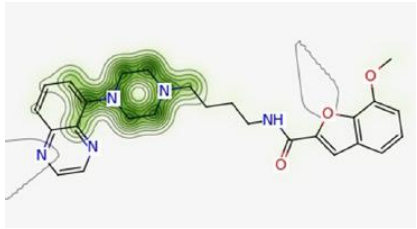
“model → structure” paradigm

Similarity maps

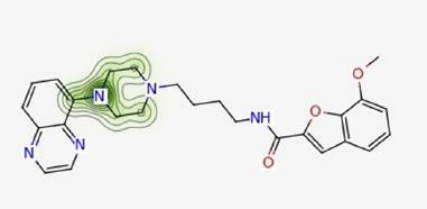


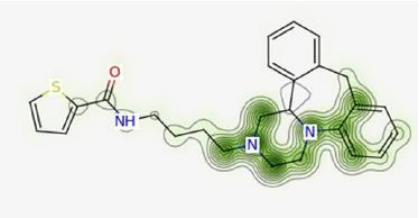
Contribution = f(X) – f(X')

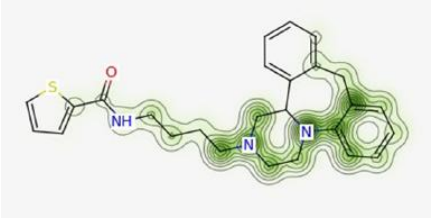
RF



NB



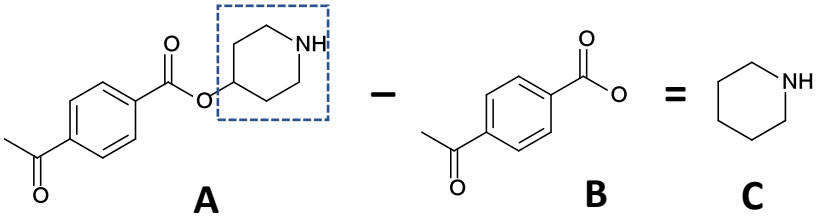




dopamine D3 ligands

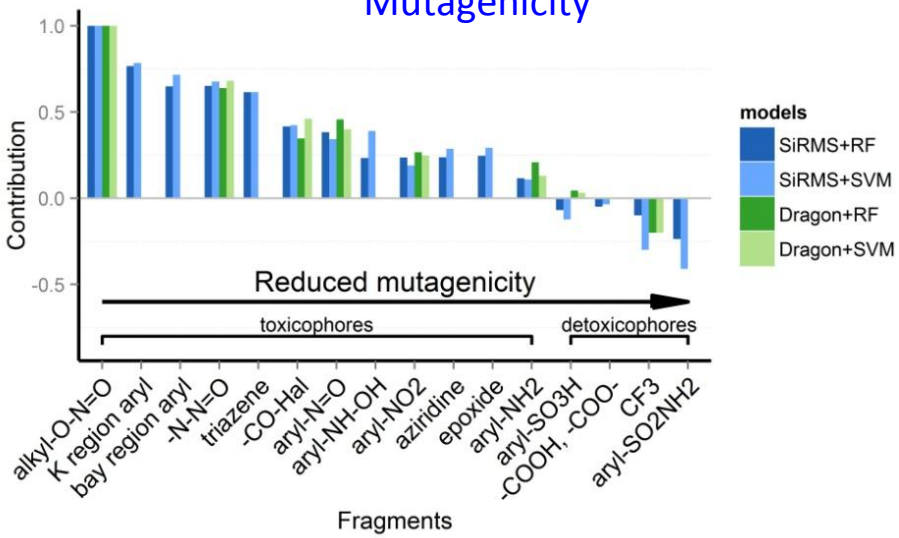
Riniker, S.; Landrum, G. Similarity maps - a visualization strategy for molecular fingerprints and machine-learning methods. *J. Cheminf.* **2013,** *5*, 43.

Universal interpretation



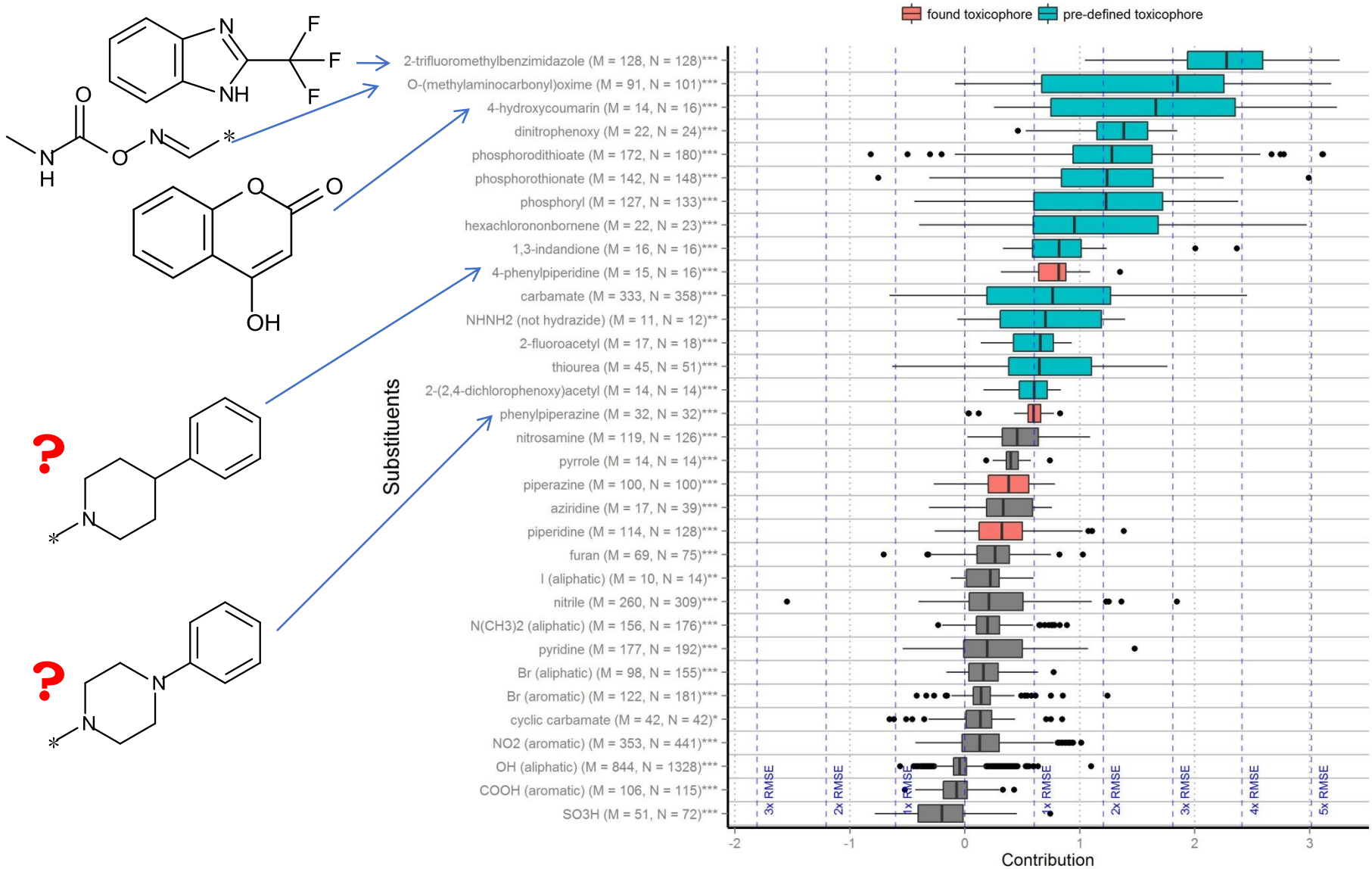
$Activity_{pred}(A)$	$Activity_{pred}(B)$	Contribution(C)
$f(A)$	$f(B)$	$W(C) = f(A) - f(B)$

Mutagenicity



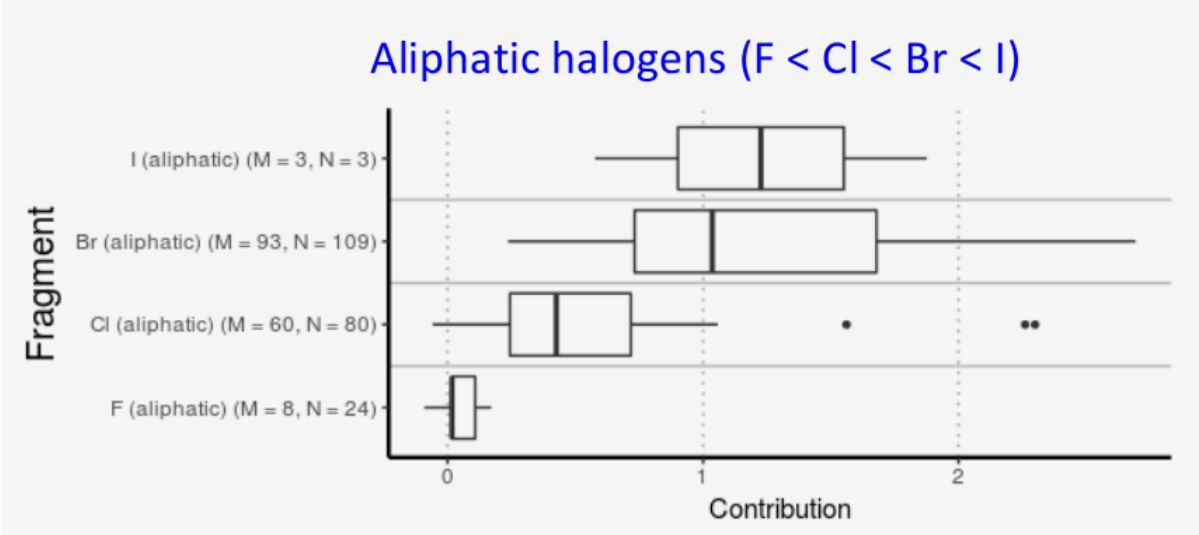
Polishchuk, P. G.; Kuz'min, V. E.; Artemenko, A. G.; Muratov, E. N. Universal Approach for Structural Interpretation of QSAR/QSPR Models. *Molecular Informatics* **2013,** *32*, 843-853
https://github.com/DrrDom/spci

Acute oral toxicity on rats



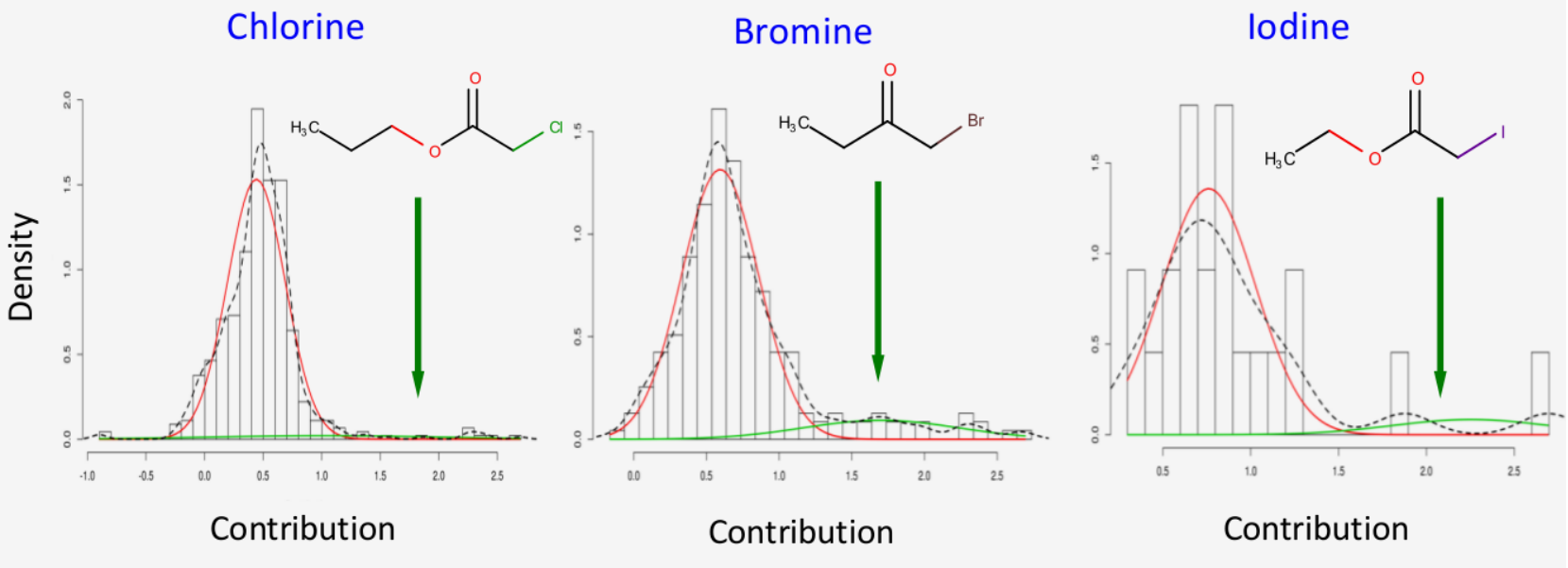
Polishchuk P, Tinkov O, Khristova T, Ognichenko L, Kosinskaya A, Varnek A, Kuz'min V. *Journal of Chemical Information and Modeling* **2016**, 56, 1455-1469.

Toxicity towards *Tetrahymena Pyriformis*

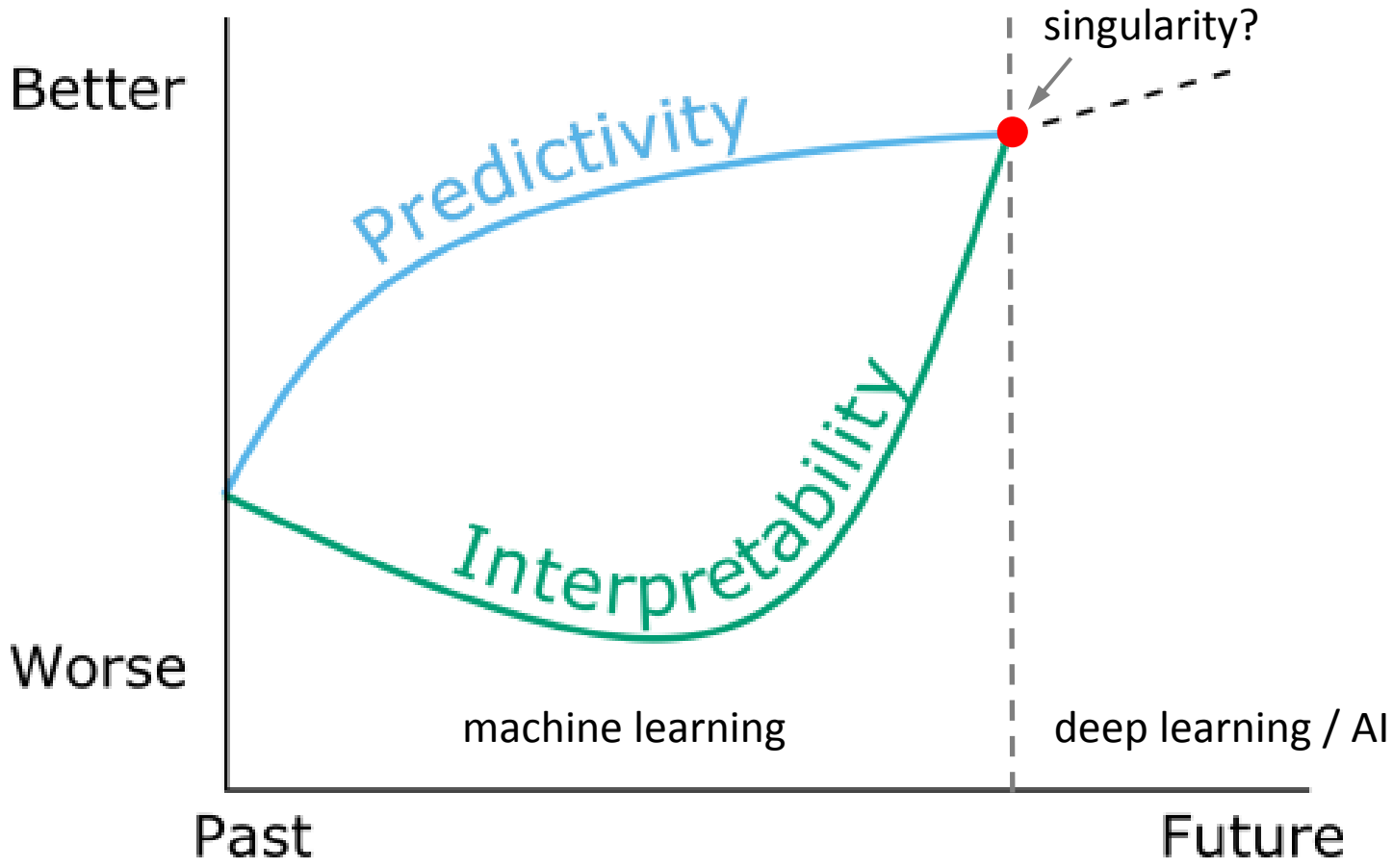


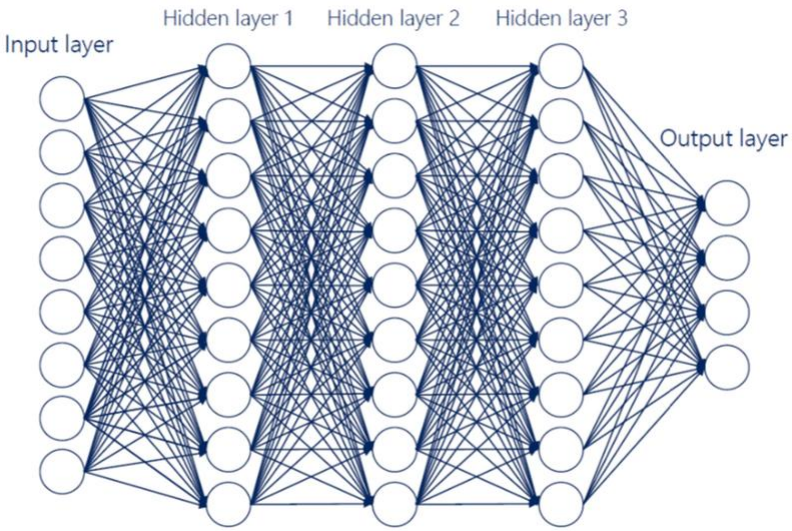
Gaussian Mixture Modeling

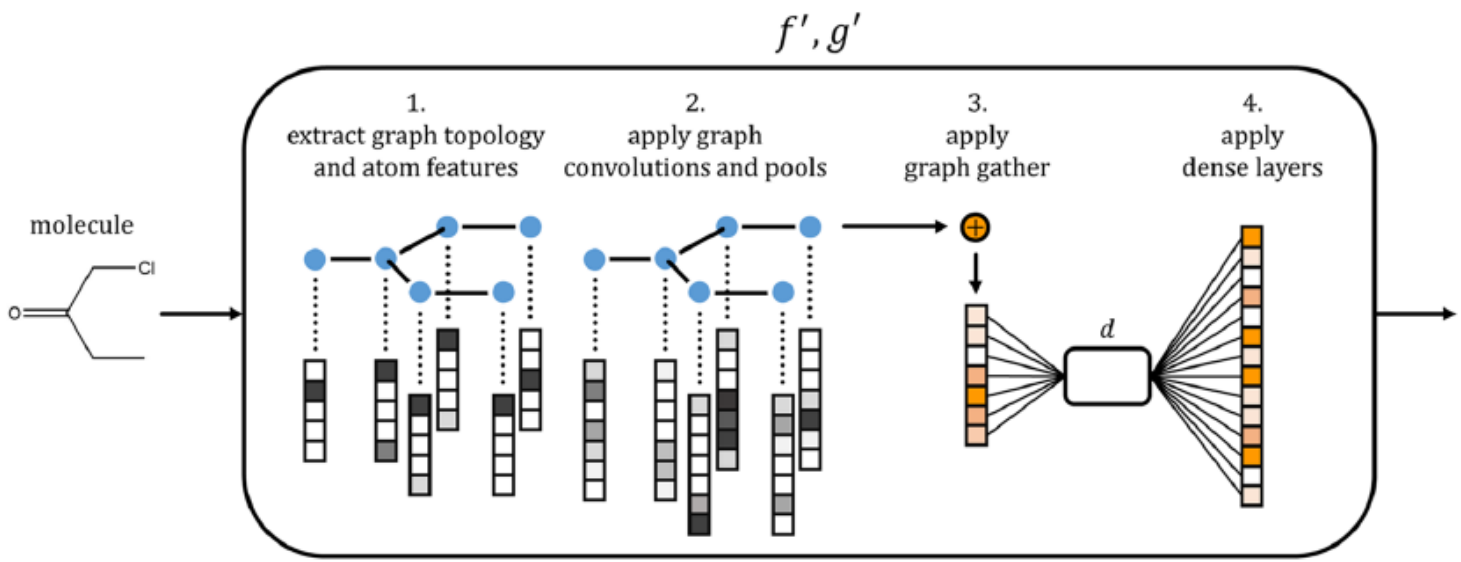
SMARTSminer



Matveieva, M.; Cronin, M. T. D.; Polishchuk, P., Interpretation of QSAR Models: Mining Structural Patterns Taking into Account Molecular Context. *Molecular Informatics* **2018**, 38, 1800084..


singularity?
Better
Predictivity
Interpretability
Worse
machine learning
deep learning / AI
Past
Future


Input layer
Hidden layer 1
Hidden layer 2
Hidden layer 3
Output layer


f', g'
molecule
1.
extract graph topology
and atom features
2.
apply graph
convolutions and pools
3.
apply
graph gather
4.
apply
dense layers
d

Self-explaining

attention-based approaches

Perturbation-based

SmoothGrad

Gradient-based

GradCAM

GradInput

Integrated Gradients

Class activation map (CAM)

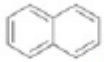
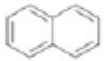
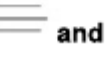
Layer-wise propagation

GNNExplainer

Surrogate modeling

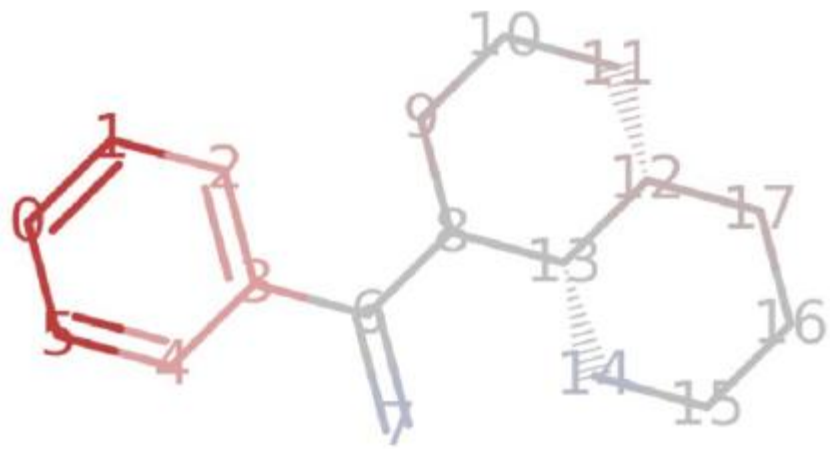
Local Interpretable Model-Agnostic Explanation (LIME)

Shapley additive explanation (SHAP)

Logic number	Synthetic binding logic	GC Zinc AUC	GC Attribution AUC
0.		1.000	0.980
1.		0.995	0.980
2.		1.000	1.000
3.	no 	1.000	0.970
4.	 or (no )	0.992	0.910
5.	 and (no )	0.999	0.890
6.	and 	1.000	0.770
7.	 and 	1.000	0.790
8.	and  and (no)	1.000	0.930
9.	and  and 	0.995	0.700
10.	(or no ) and (no )	0.999	0.860
11.	and (no) and (no )	1.000	0.880
12.	and  and (or no)	0.999	0.670
13.	( and no ) or ( and no )	1.000	0.700
14.	( or no) and  and (no )	1.000	0.750
15.	 and (no ) and and 	0.996	0.760

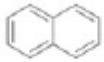
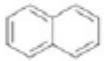
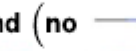
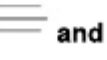
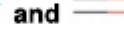
McCloskey, K.; Taly, A.; Monti, F.; Brenner, M. P.; Colwell, L. J., Using attribution to decode binding mechanism in neural network models for chemistry. *Proceedings of the National Academy of Sciences* **2019**, 116, 11624-11629.

Logic number	Synthetic binding logic	GC Zinc AUC	GC Attribution AUC
0.		1.000	0.980
1.		0.995	0.980
2.			
3.			
4.			
5.			
6.			
7.			
8.			
9.			
10.			
11.			
12.			
13.			
14.			
15.			



Atom index	Attribution scores ranked in decreasing order	Involved in ground truth binding logic
1	0.29	1
5	0.29	1
0	0.28	1
2	0.09	1
4	0.09	1
3	0.07	1
9	0.03	0
11	0.02	0
...	...	...

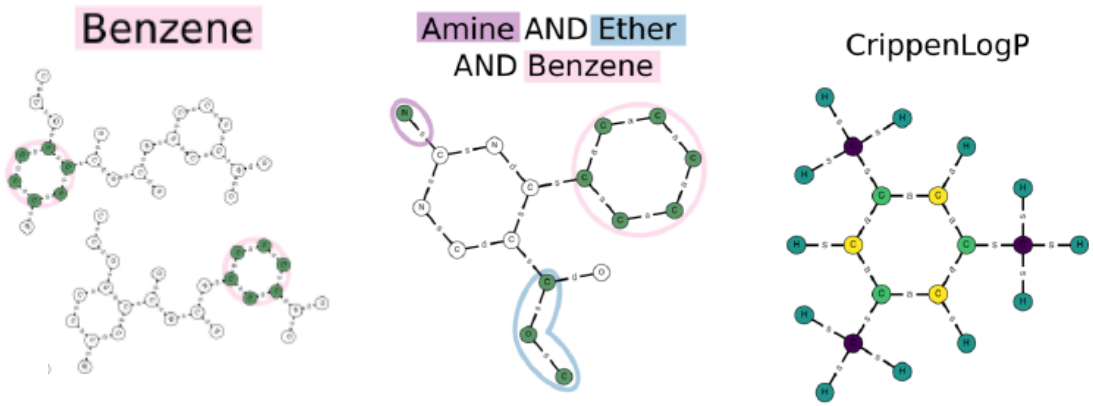
McCloskey, K.; Taly, A.; Monti, F.; Brenner, M. P.; Colwell, L. J., Using attribution to decode binding mechanism in neural network models for chemistry. *Proceedings of the National Academy of Sciences* **2019**, 116, 11624-11629.

Logic number	Synthetic binding logic	GC Zinc AUC	GC Attribution AUC
0.		1.000	0.980
1.		0.995	0.980
2.		1.000	1.000
3.	no 	1.000	0.970
4.	 or (no )	0.992	0.910
5.	 and (no )	0.999	0.890
6.	and 	1.000	0.770
7.	 and 	1.000	0.790
8.	and  and (no)	1.000	0.930
9.	and  and 	0.995	0.700
10.	(or no ) and (no )	0.999	0.860
11.	and (no) and (no )	1.000	0.880
12.	and  and (or no)	0.999	0.670
13.	( and no) or ( and no )	1.000	0.700
14.	( or no ) and and (no)	1.000	0.750
15.	 and (no ) and and	0.996	0.760

McCloskey, K.; Taly, A.; Monti, F.; Brenner, M. P.; Colwell, L. J., Using attribution to decode binding mechanism in neural network models for chemistry. *Proceedings of the National Academy of Sciences* **2019**, 116, 11624-11629.

binary classification

regression



	Benzene				Amine AND Ether AND Benzene				CrippenLogP			
	GCN	MPNN	GraphNets	GAT	GCN	MPNN	GraphNets	GAT	GCN	MPNN	GraphNets	GAT
Random Baseline	0.61	0.61	0.61	0.61	0.5	0.5	0.5	0.5	0.13	0.13	0.13	0.13
GradInput	0.72	0.54	0.54	0.56	0.52	0.53	0.55	0.41	0.12	0.09	0.13	0.1
SmoothGrad(GI)	0.71	0.54	0.54	0.53	0.51	0.55	0.59	0.38	0.15	0.11	0.15	0.11
GradCAM-last	0.74	0.72	0.66	0.66	0.54	0.74	0.55	0.46	0.04	0.33	0.24	0.07
GradCAM-all	0.75	0.68	0.84	0.62	0.54	0.62	0.7	0.44	0.05	0.27	0.27	0.09
IG	0.97	0.89	0.94	0.95	0.69	0.59	0.72	0.54	0.31	0.24	0.24	0.27
CAM	0.98	0.96	0.76	0.99	0.75	0.76	0.6	0.65	0.2	0.37	0.28	0.23
Attention Weights	--	--	--	0.51	--	--	--	0.51	--	--	--	-0.06

AUC

$R_{Kendall}$

Sanchez-Lengeling, B.; Wei, J.; Lee, B.; Reif, E.; Wang, P.; Qian, W.; McCloskey, K.; Colwell , L.; Wiltschko, A., Evaluating Attribution for Graph Neural Networks. In *Advances in Neural Information Processing Systems, Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M. F.; Lin, H., Eds.* **2020**; Vol. 33, pp 5898--5910.

Benchmark suite for interpretation approaches

regression data sets:

N data set:
N = +1
others = 0

N-O data set:
N = +1
O = -1
others = 0

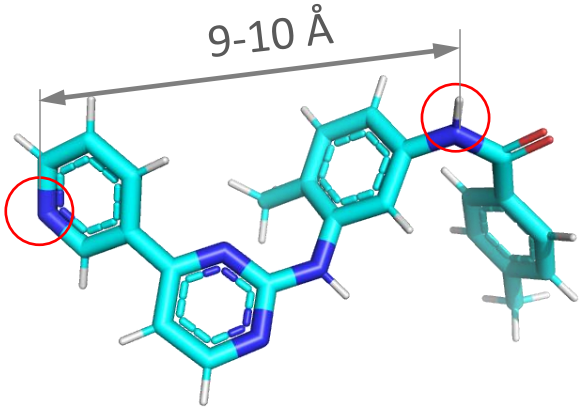
N+O data set:
N = +0.5
O = +0.5
n(N) = n(O) in molecules

amide data set:
amide group = +1

classification data sets:

amide data set:
presence of an amide group

pharmacophore data set:
H-bond donor ... 9-10Å ... H-bond acceptor
(for at least one conformer)



training: 7000
test: 3000

Control possible correlations in data sets
to avoid hidden biases

https://github.com/ci-lab-cz/ibenchmark

Matveieva, M.; Polishchuk, P., Benchmarks for interpretation of QSAR models. *Journal of Cheminformatics* **2021**, 13, 41.

Benchmark suite for interpretation approaches

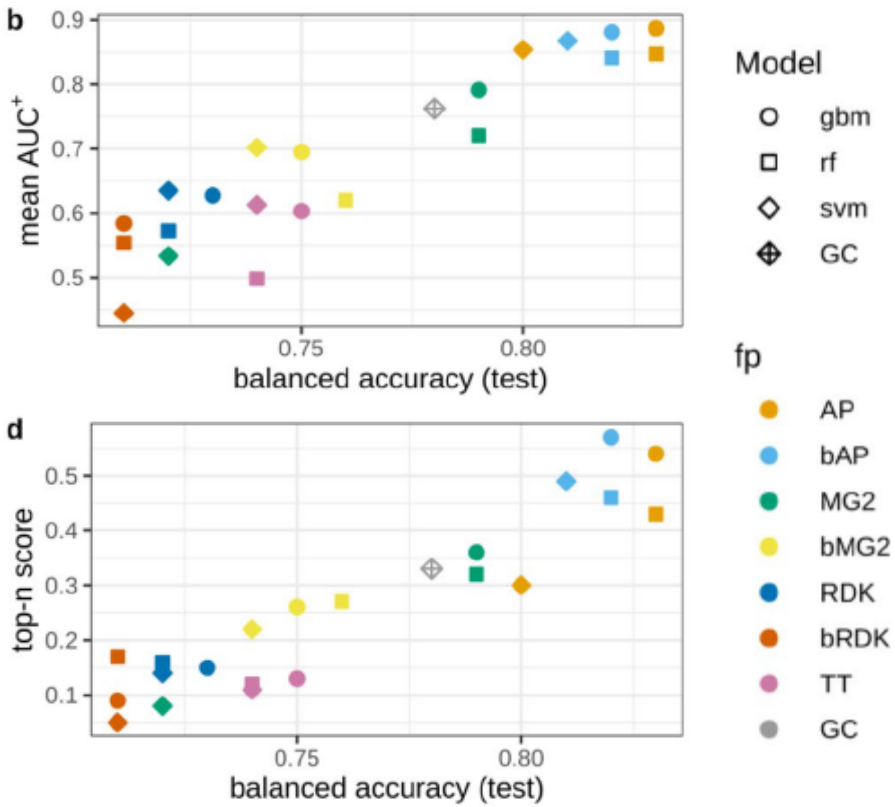
metrics:

AUC – how well a model/approach rank all atoms

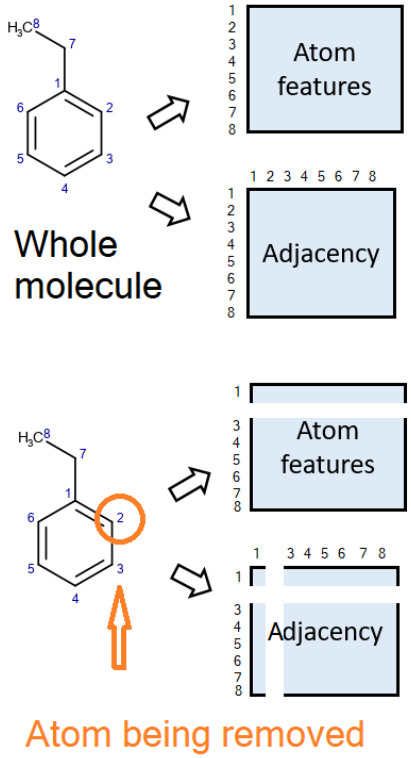
top-n – how well a model/approach rank ground truth atoms on top

RMSE = how precisely a model/approach calculate atom contribution

pharmacophore data set



GCN interpretation



Matveieva, M.; Polishchuk, P., Benchmarks for interpretation of QSAR models. *Journal of Cheminformatics* **2021**, 13, 41.

JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

Cite This: *J. Chem. Inf. Model.* XXXX, XXX, XXX−XXX

pubs.acs.org/jcim

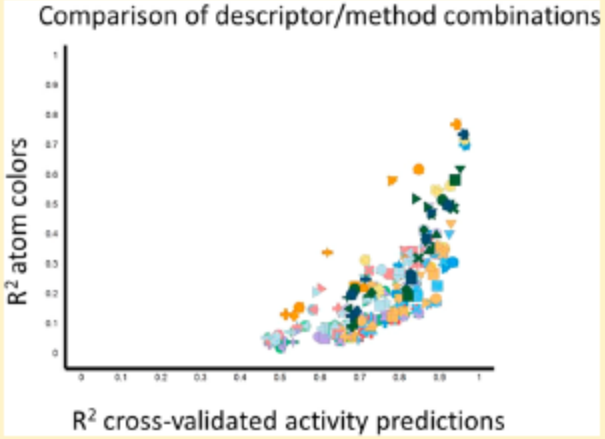
Interpretation of QSAR Models by Coloring Atoms According to Changes in Predicted Activity: How Robust Is It?

Robert P. Sheridan*

Modeling and Informatics, Merck & Co. Inc., Kenilworth, New Jersey 07065, United States

Supporting Information

ABSTRACT: Most chemists would agree that the ability to interpret a quantitative structure−activity relationship (QSAR) model is as important as the ability of the model to make accurate predictions. One type of interpretation is coloration of atoms in molecules according to the contribution of each atom to the predicted activity, as in "heat maps". The ability to determine which parts of a molecule increase the activity in question and which decrease it should be useful to chemists who want to modify the molecule. For that type of application, we would hope the coloration to not be particularly sensitive to the details of model building. In this Article, we examine a number of aspects of coloration against 20 combinations of descriptors and QSAR methods. We demonstrate that atom-level coloration is much less robust to descriptor/method combinations than cross-validated predictions. Even in ideal cases where the contribution of individual atoms is known, we cannot always recover the important atoms for some descriptor/method combinations. Thus, model interpretation by atom coloration may not be as simple as it first appeared.



Sheridan, R. P., Interpretation of QSAR Models by Coloring Atoms According to Changes in Predicted Activity: How Robust Is It? *Journal of Chemical Information and Modeling* **2019.**

1. Do we need new interpretation approaches?

Yes, but they should be properly validated and compared with state-of-the-art approaches.

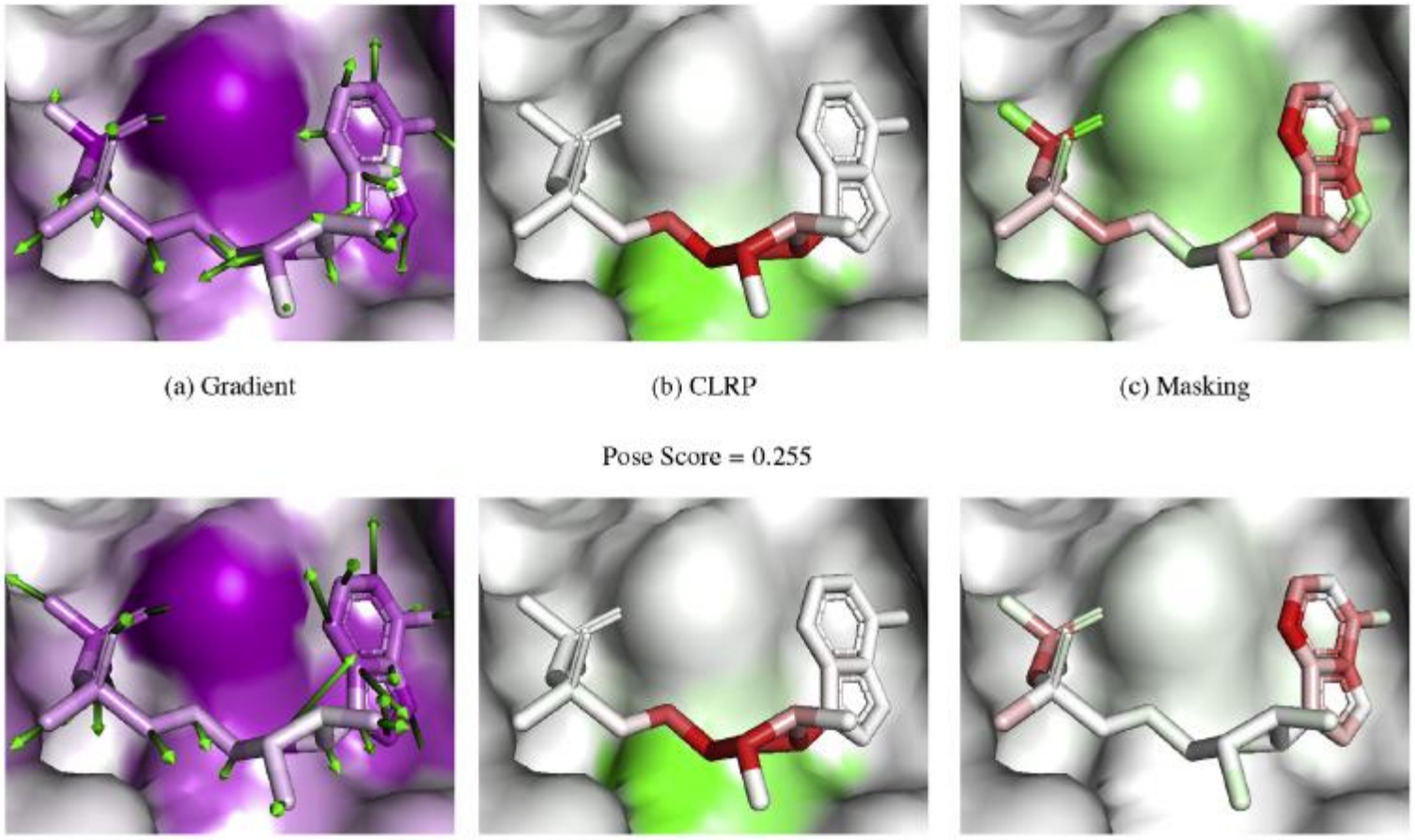
2. What should be the main focus of further research?

Development of approaches able to retrieve new types of knowledge from models (we can already retrieve atom contributions by different methods, let's search for something new and useful)

There is a need to introduce interpretation into routine decision making pipelines to supports decisions of medicinal chemists and other researchers

Gnina interpretation

Affinity Prediction Score = 2.698



(a) Gradient (b) CLRP (c) Masking

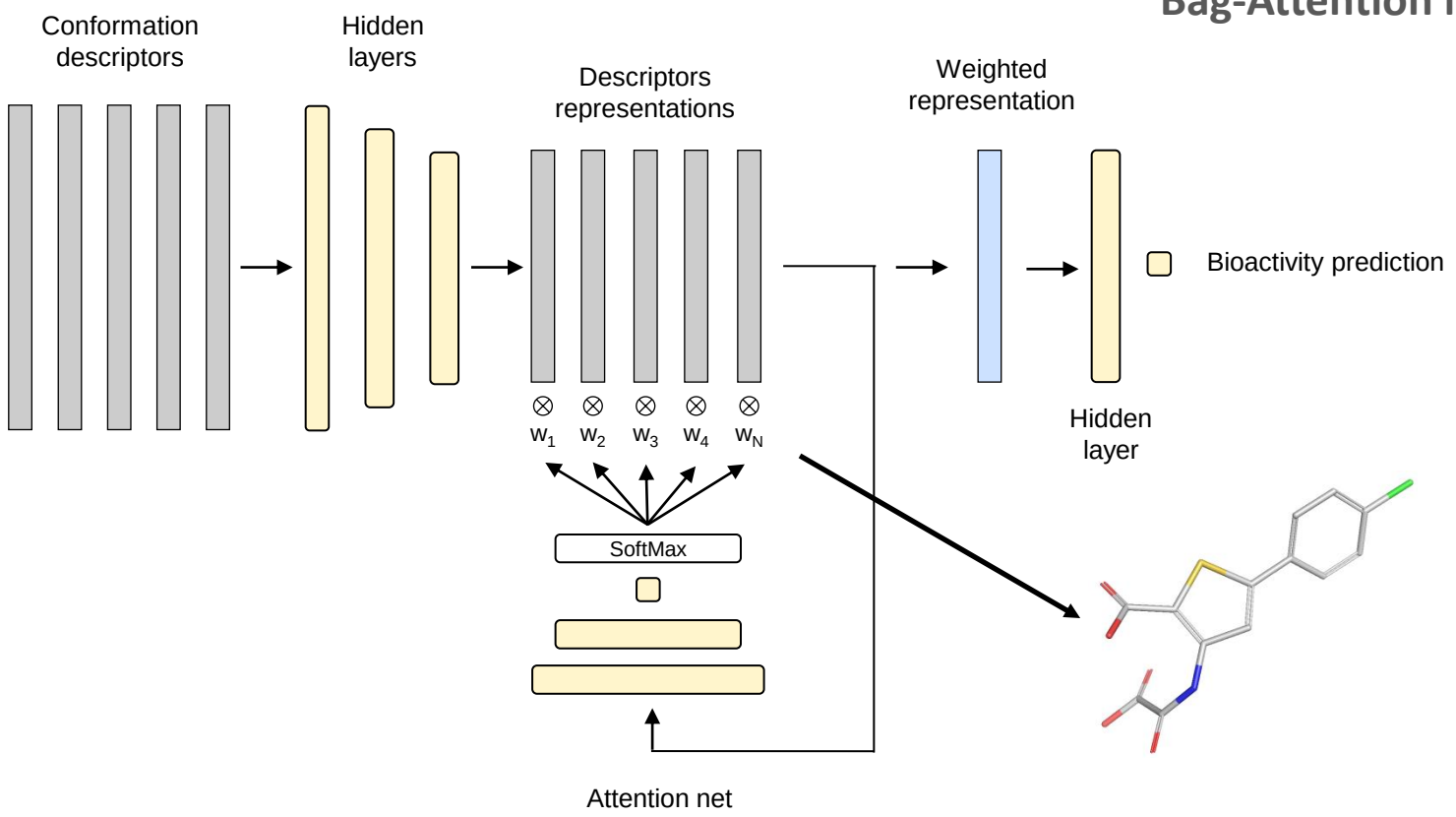
Pose Score = 0.255

(d) Gradient (e) CLRP (f) Masking

Hochuli, J.; Helbling, A.; Skaist, T.; Ragoza, M.; Koes, D. R., Visualizing convolutional neural network protein-ligand scoring. *Journal of Molecular Graphics and Modelling* **2018**, 84, 96-108.

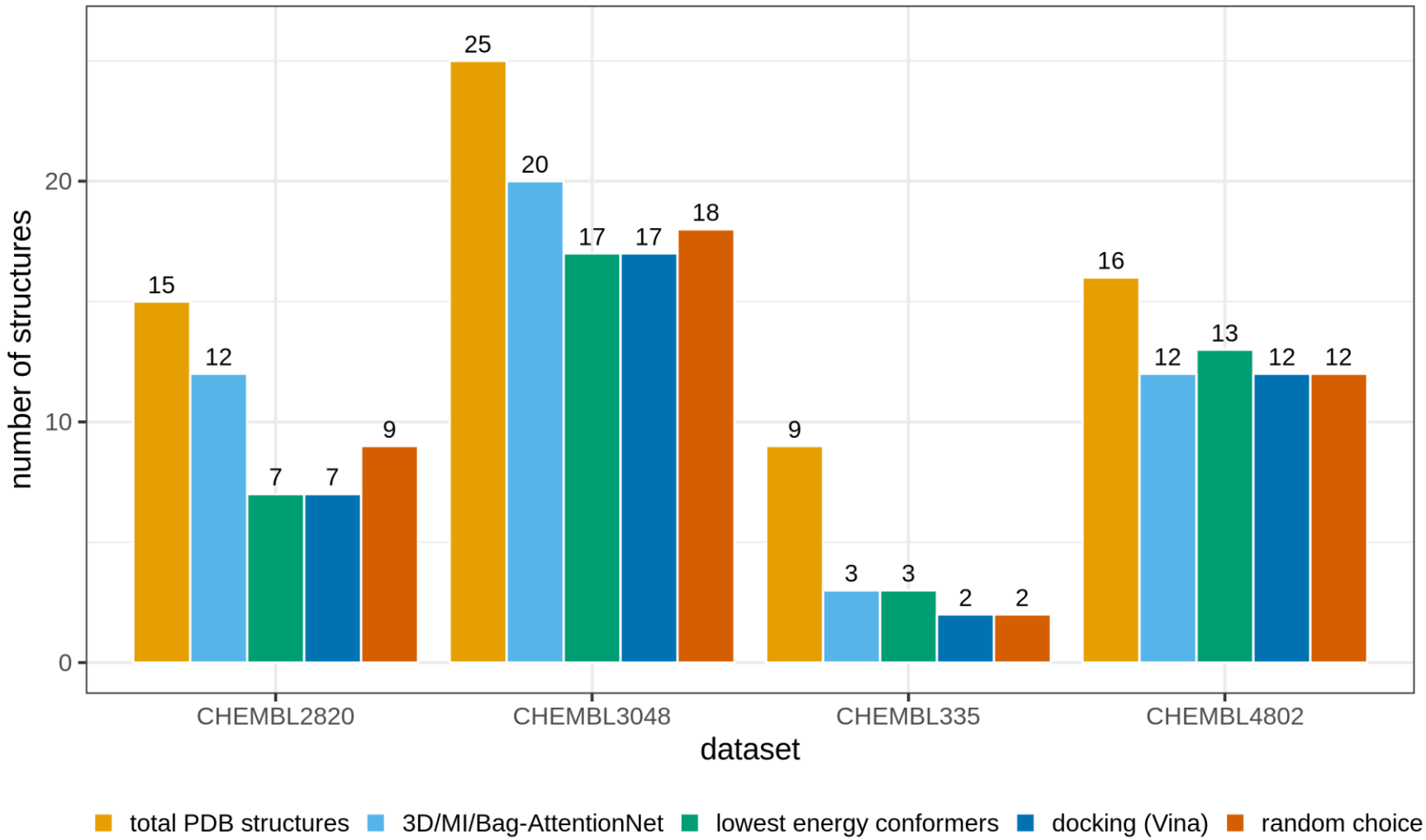
Identification of bioactive conformers

Bag-Attention Net



Zankov, D. V.; Matveieva, M.; Nikonenko, A. V.; Nugmanov, R. I.; Baskin, I. I.; Varnek, A.; Polishchuk, P.; Madzhidov, T. I., QSAR Modeling Based on Conformation Ensembles Using a Multi-Instance Learning Approach. *Journal of Chemical Information and Modeling* **2021**, 61, 4913-4923.

Selected compounds from test set had average RMSD of generated conformers > 2A relative to PDB structure



Zankov, D. V.; Matveieva, M.; Nikonenko, A. V.; Nugmanov, R. I.; Baskin, I. I.; Varnek, A.; Polishchuk, P.; Madzhidov, T. I., QSAR Modeling Based on Conformation Ensembles Using a Multi-Instance Learning Approach. *Journal of Chemical Information and Modeling* **2021**, 61, 4913-4923.

Take-home messages

- There are a lot of hidden treasures in models which can improve our understanding of complex phenomena and augment our knowledge and our goal is to retrieve them

- Do not use anecdotal evidence for evaluation of interpretation performance, do systematic evaluation (use available benchmarks or develop your own)

- The more predictive a model the better interpretation performance, however, even for well predictive models interpretation may be rather low

- Not all interpretation approaches are applicable to chemical problems

- Any predictive model is interpretable ("model ⟶ structure" paradigm)

pharmaceuticals

an Open Access Journal by MDPI

IMPACT FACTOR 5.863

Covered in: PubMed

Pharmacophore Modeling and Applications in Drug Discovery: Challenges and Recent Advances

Guest Editors

Dr. Pavel Polishchuk, Dr. Thomas Seidel

Deadline

28 February 2022

Special Issue

mdpi.com/si/91899

Invitation to submit