EMBL-EBI Industry workshop: In Silico ADMET prediction
4-5 December 2014, Hinxton, UK

Exploring the black box: structural and functional interpretation of QSAR models.

(Automatic exploration of datasets using QSAR)

Pavel Polishchuk

A.V. Bogatsky Physico-Chemical Institute of
NAS of Ukraine, Odessa, Ukraine

pavel_polishchuk@ukr.net

Outline

Introduction and existed approaches

Structural QSAR interpretation

- theory and practical examples

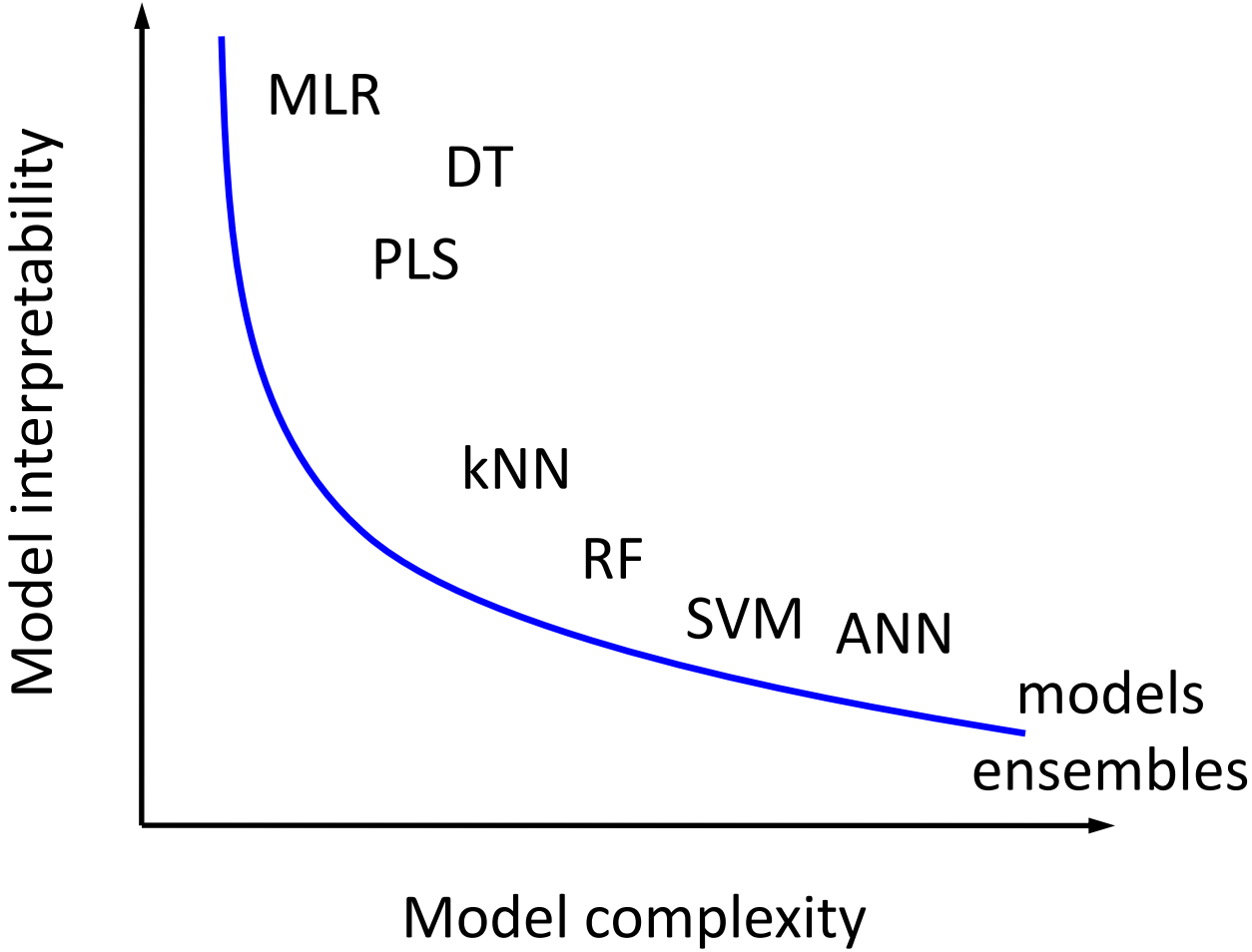
Functional QSAR interpretation

- theory, practical examples and comparison with docking studies

Automatic exploration of chemical dataset

QSAR interpretation: interpretability *vs.* complexity

Popular misbelief



Importance of QSAR structural interpretation

Extract SAR information in a chemically meaningful way

- detection of structural alerts, creation of structural filters or set of rules
- fragment-based drug design

Model validation

- interpretation results should not contradict with experimental observations

QSAR interpretation approaches

Model-specific approaches:

Rule-based (Decision tree)

Regression coefficients (MLR, PLS)

Latent variables (PLS)

Weights and biases (ANN)

…

Model-independent approaches:

Local gradients or partial derivatives

$$C_i = \frac{f(x_i) - f(x_i + \Delta x_i)}{\Delta x_i}$$

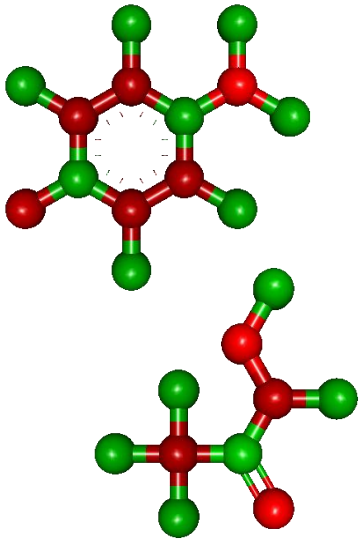
I.I. Baskin et al., SAR QSAR Environ Sci, **2002**, 35-41
G. Marcou et al., Molecular informatics, **2012**, 639-642

QSAR interpretation: common workflow

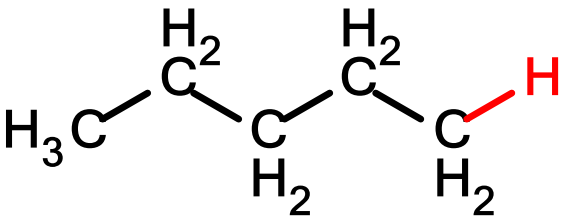
Model → **Variables contributions** → **Structure-property relationship**

$f(x)$

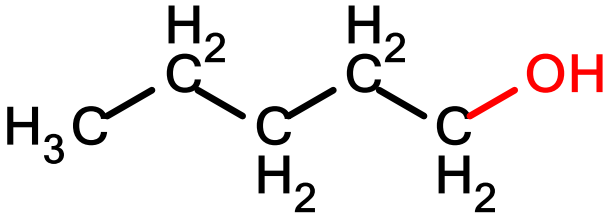
	Var_1	Var_2
Mol_1	-0.23	1.82
Mol_2	2.36	1.27
Mol_3	5.01	2.30
Mol_4	0.69	-0.58



Matched molecular pairs & molecular transformations



logS = -3.18

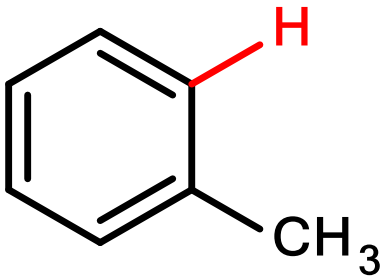


logS = -0.60

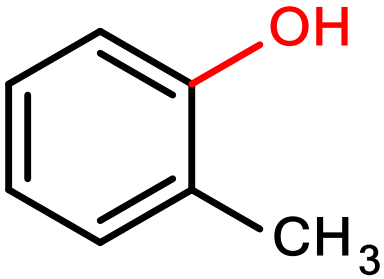
ΔlogS = 2.58

H → OH

ΔlogS = 1.59



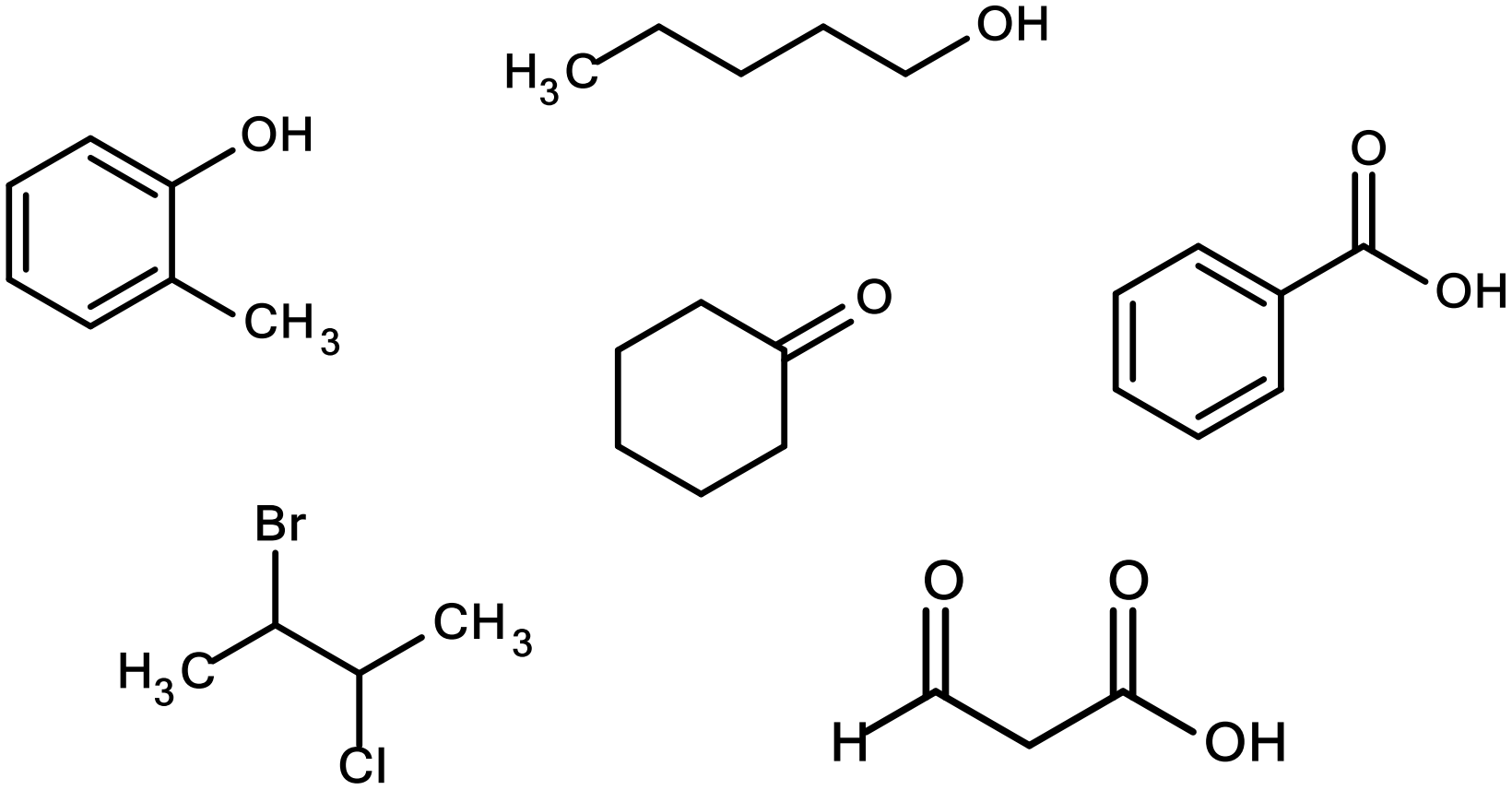
logS = -2.21



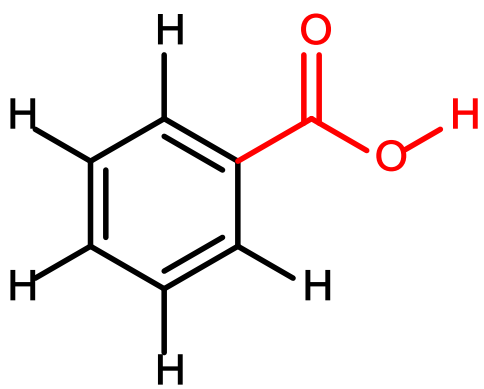
logS = -0.62

Leech A.G. et al, *J. Med. Chem.* **2006**, 49, 6672-6682
Sheridan R.P. et al., *J. Chem. Inf. Model.* **2006**, 46, 180-192

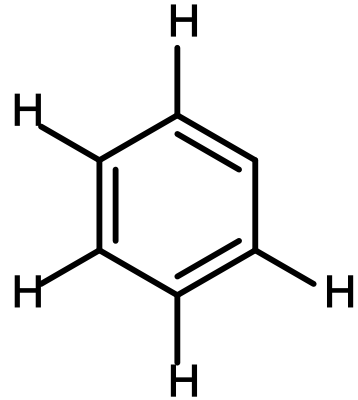
Exemplified dataset



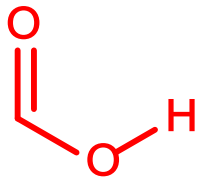
Structural QSAR interpretation



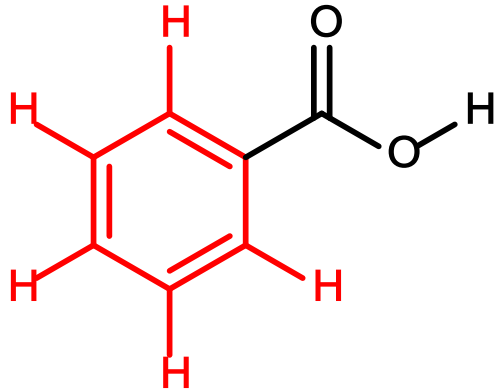
-



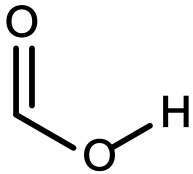
=



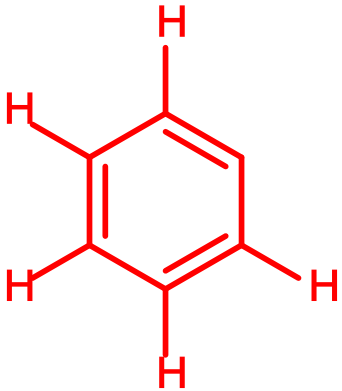
$logS_{pred}$ = -1.55 $logS_{pred}$ = -1.61 $\Delta logS_{pred}$ = 0.06



-



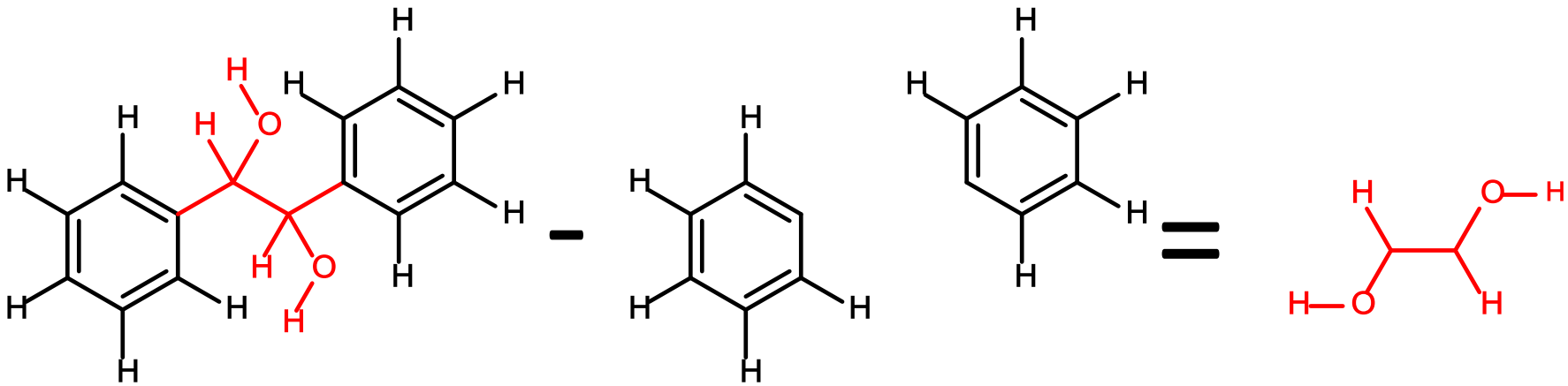
=



$logS_{pred}$ = -1.55 $logS_{pred}$ = -1.35 $\Delta logS_{pred}$ = -0.20

Polishchuk P.G. et al. *Molecular Informatics* **2013**,32, 843-853

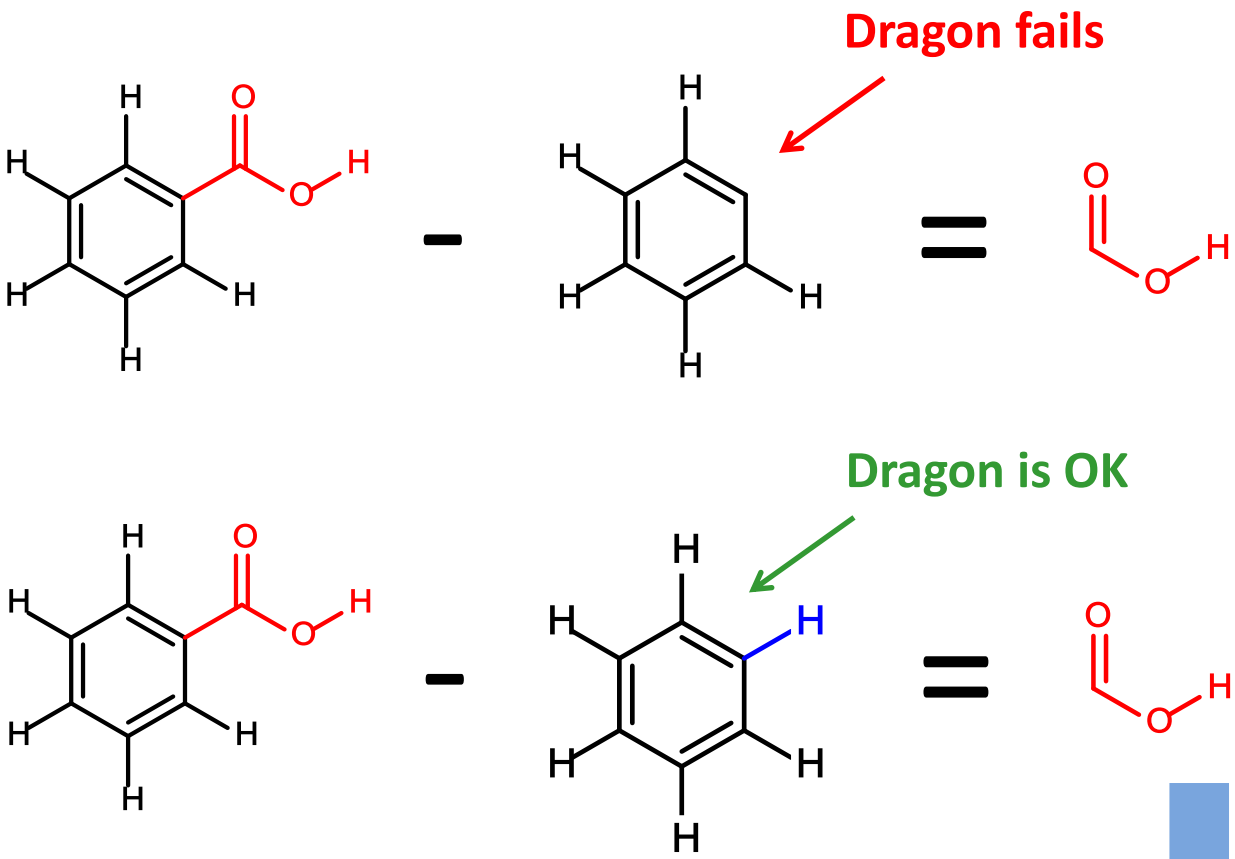
Structural QSAR interpretation



$\log S_{pred}$ = -1.93 $\log S_{pred}$ = -4.32 $\Delta \log S_{pred}$ = 2.39

Polishchuk P.G. et al. *Molecular Informatics* **2013**,32, 843-853

Limitations of existed descriptors (Dragon, etc)

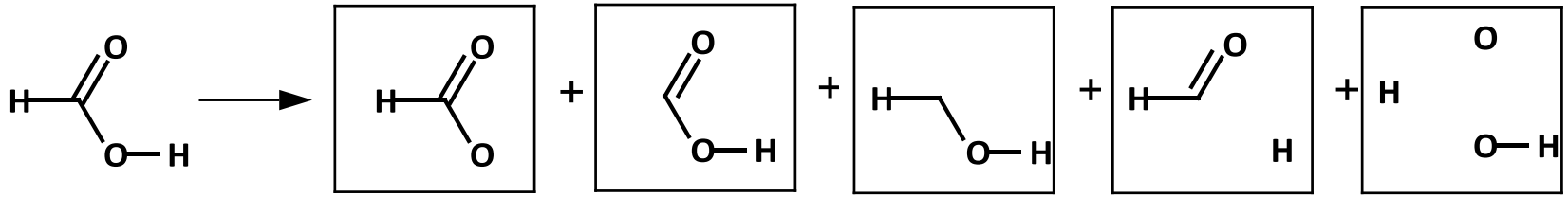


Computational MMP H → COOH

Polishchuk P.G. et al. *Molecular Informatics* **2013**,32, 843-853

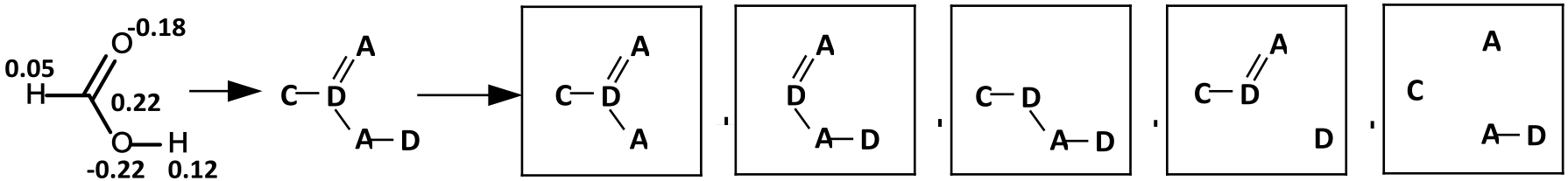
Simplex representation of molecular structure (SiRMS)

Simplex generation example



Atom-property labeling

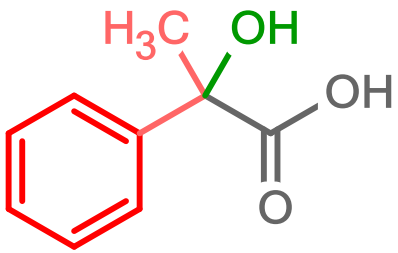
Labeling of simplex vertexes by atom properties
(for example by partial charge, groups are A ≤ -0.05 < B ≤ 0 < C ≤ 0.05 < D)

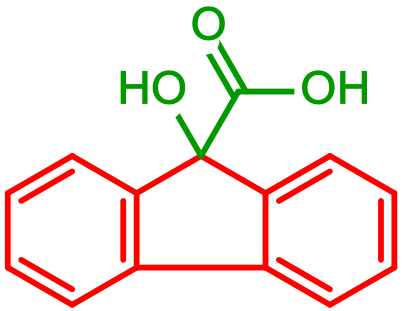


Kuz'min, V. E. et al, *Journal of Molecular Modeling* **2005**, 11, 457-467.
Kuz'min, V. E. et al, *Journal of Computer-Aided Molecular Design* **2008**, 22, 403-421.

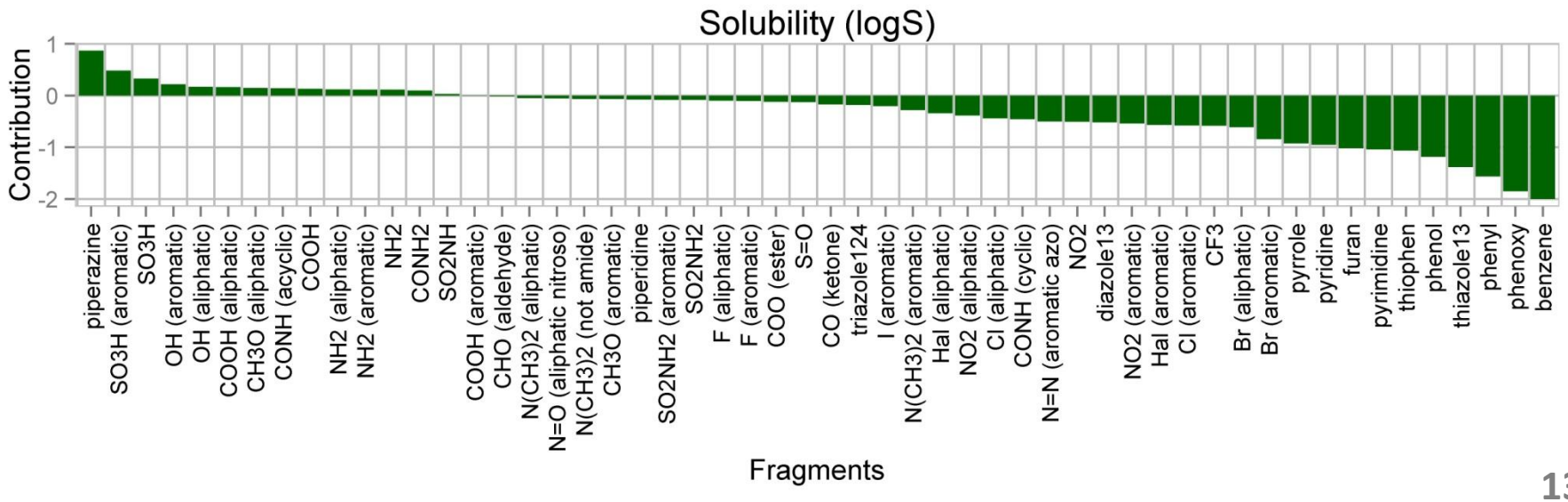
Local and global interpretation

Local interpretation – analysis of single compounds





Global interpretation – reveal trends



Interpretation: fragmentation

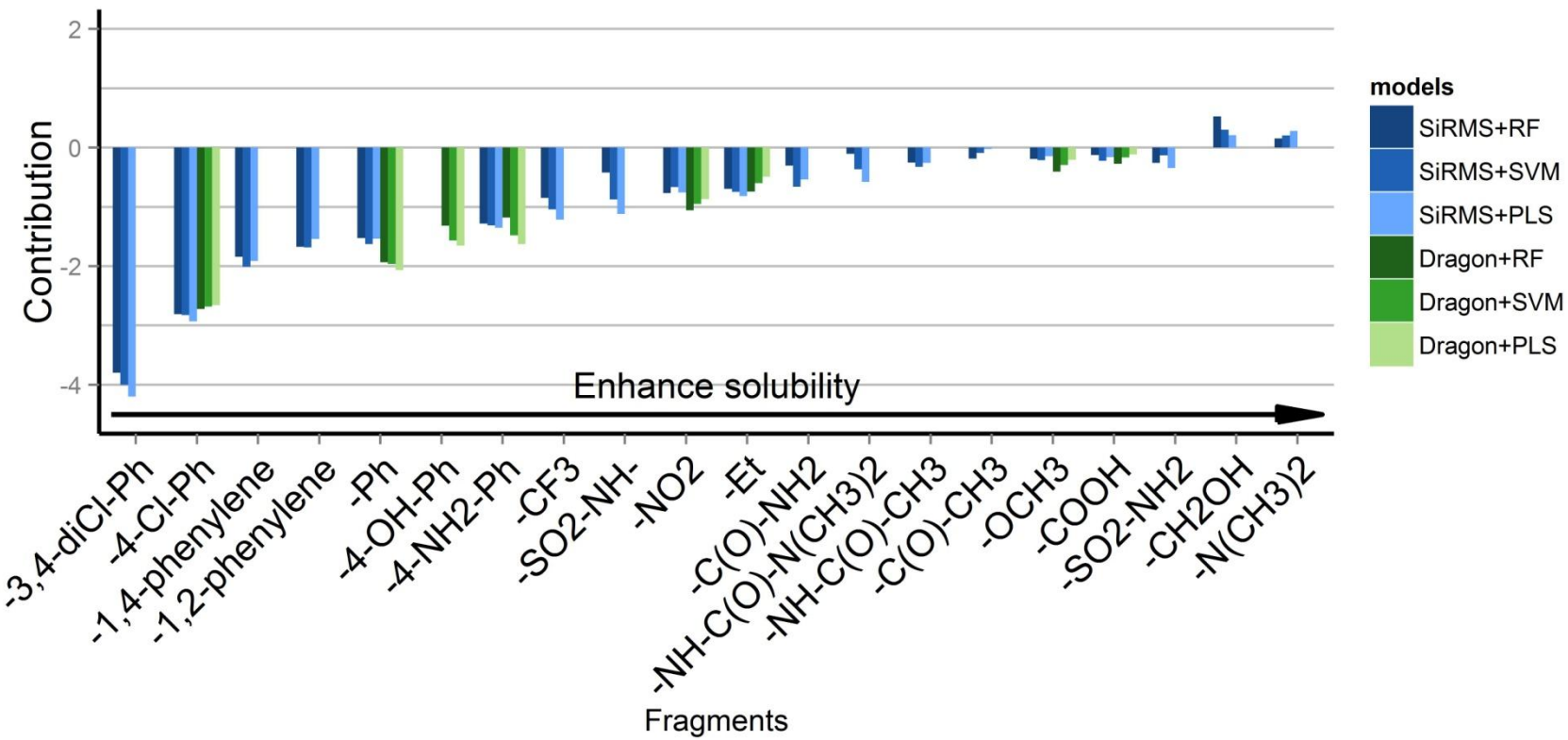
Case number	Do specific interactions of a ligand with its target exist or important?	Is an orientation of a ligand relatively its target known?	Fragments selection and grouping
1	NO (e.g. passive diffusion through membranes, solubility, lipophilicity, etc)	not relevant	can be done by the researcher based on his own knowledge
2	YES (ligand-receptor interactions, host-guest complexes, etc)	YES	consider fragments' positions relatively to the target and observed or predicted interactions
3		NO	MMP can be applied, silently assumed that all compounds have the same interaction mode

Examples of structural interpretation

Solubility (1033 compounds)

5-fold external cross validation results

Endpoint	Model	SiRMS		Dragon	
		R^2_{CV}	RMSE	R^2_{CV}	RMSE
Solubility, logS	PLS	0.84	0.82	0.91	0.60
	RF	0.88	0.71	0.91	0.62
	SVM	0.87	0.72	0.92	0.59

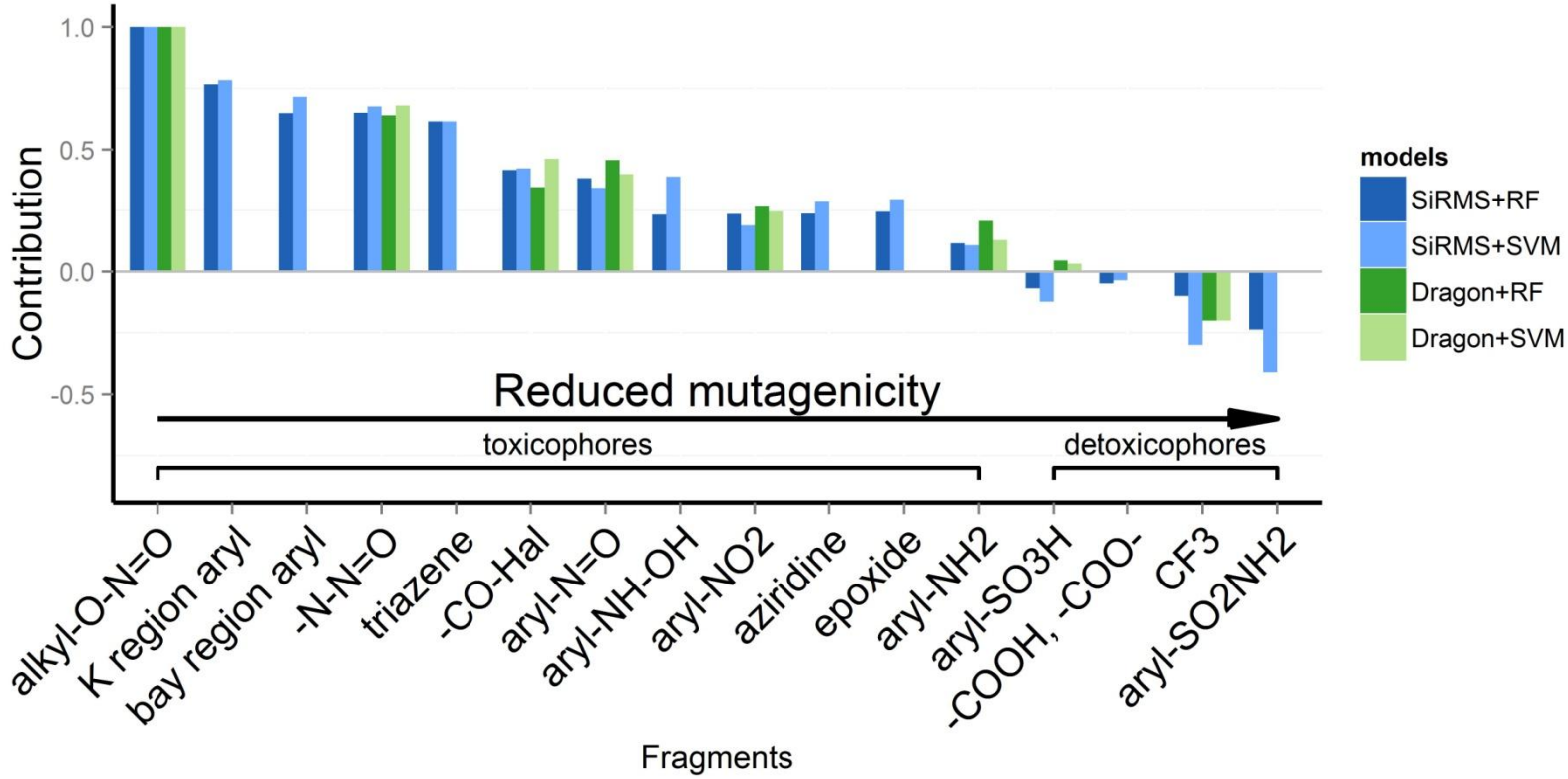


Polishchuk P.G. et al., *Molecular Informatics*, **2013**, 843-853

Mutagenicity (Ames, 4361 compounds)

5-fold external cross validation results

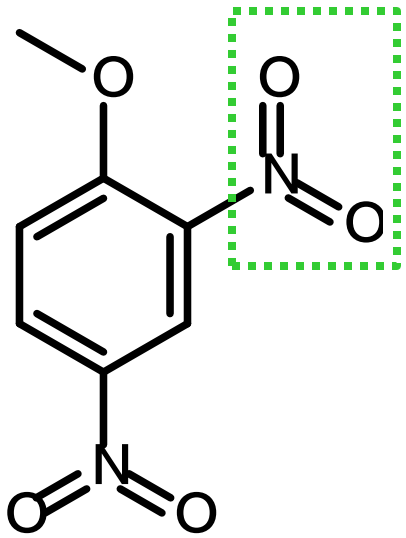
Descriptors	Algorithm	Balanced Accuracy
SiRMS	RF	0.817
	SVM	0.800
Dragon	RF	0.816
	SVM	0.793



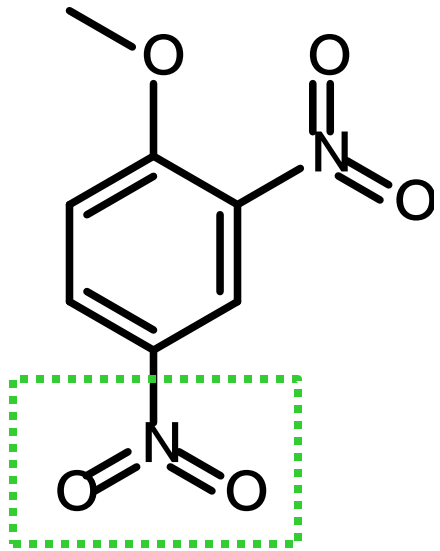
Polishchuk P.G. et al. *Molecular Informatics*, **2013**, 843-853

Combined contribution (effect) of fragments

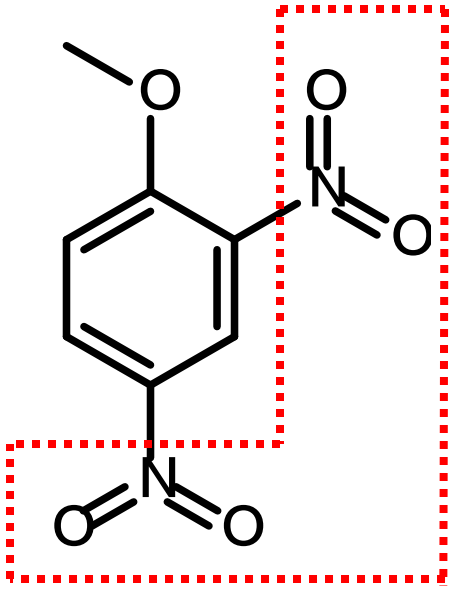
RF+SiRMS



Contribution = **0**
(non-mutagen)



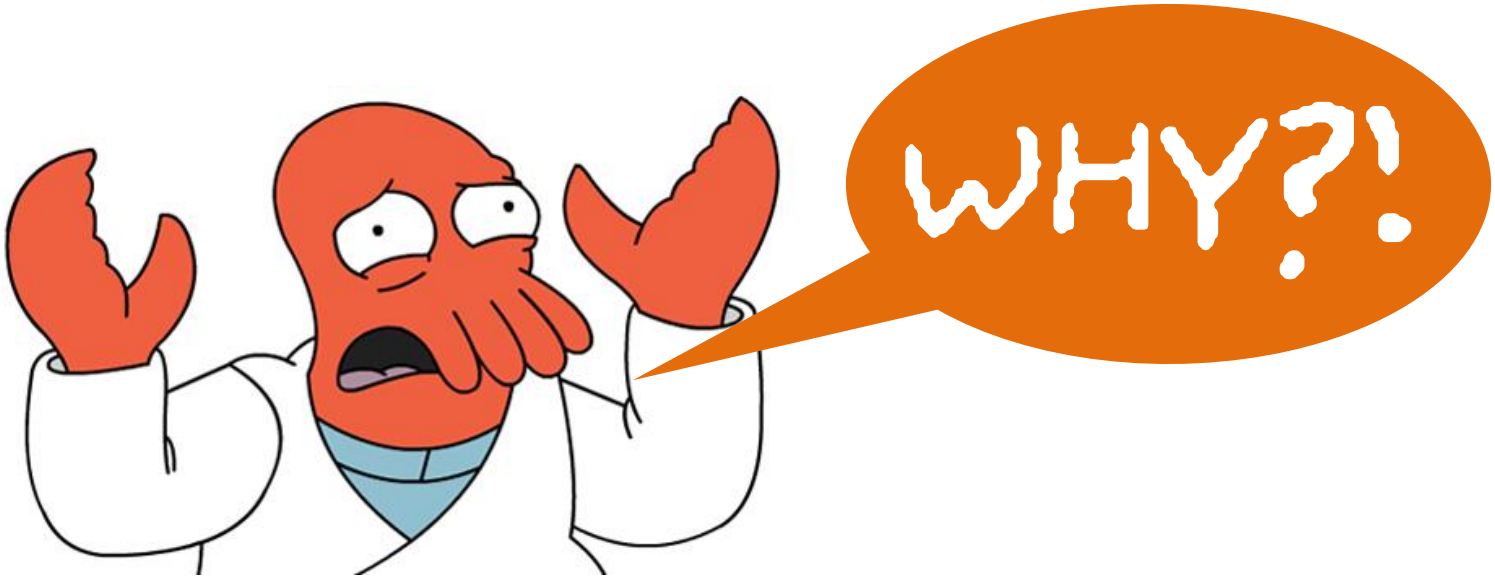
Contribution = **0**
(non-mutagen)



Contribution = **1**
(mutagen)

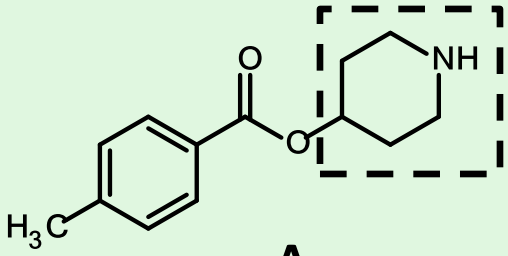
Questions

How does the fragment influence the property?



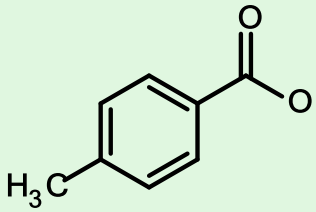
Functional interpretation of QSAR models

Structural interpretation



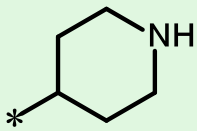
A

-



B

=



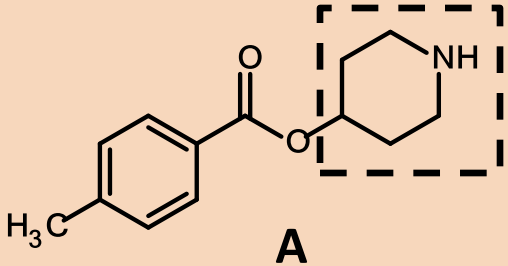
C

$pIC_{50} = f(A_1, A_2, A_3) = x$

$pIC_{50} = f(B_1, B_2, B_3) = y$

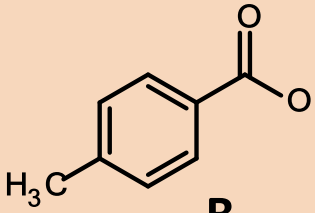
$Contribution(C) = x - y$

Functional interpretation



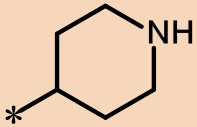
A

-



B

=



C

$pIC_{50} = f(A_1, A_2, A_3) = x$

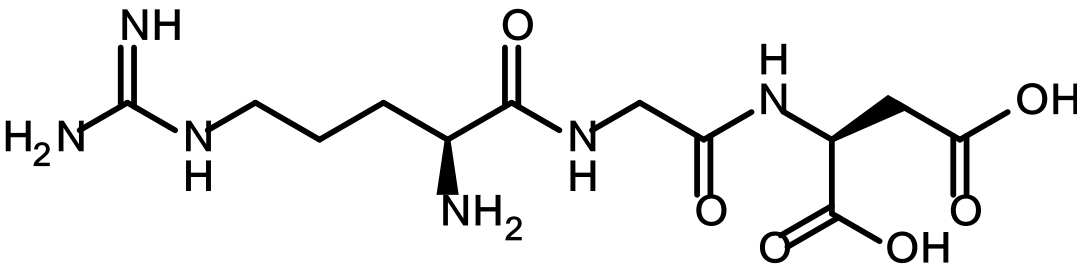
$pIC_{50} = f(A_1, A_2, B_3) = y$

$Contribution_3(C) = x - y$

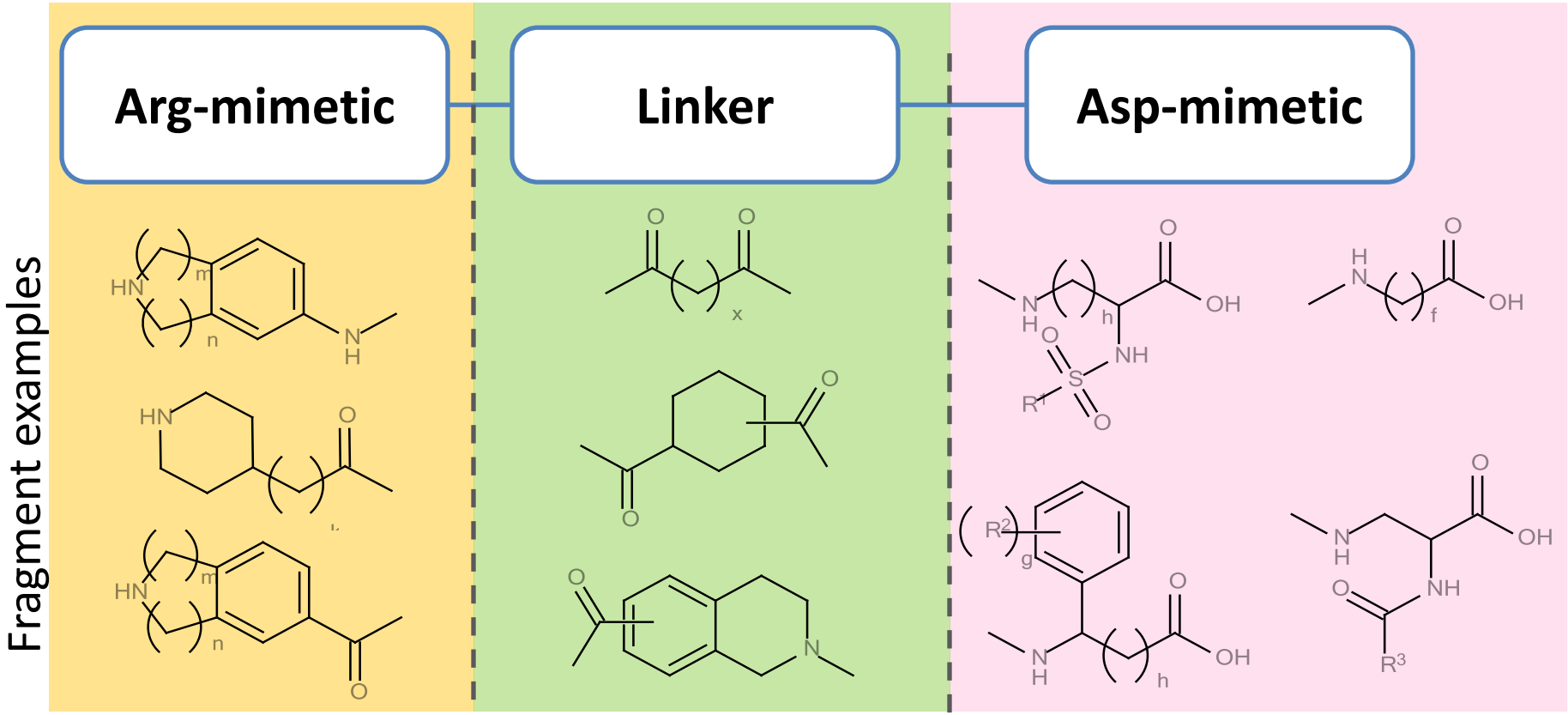
1, 2, 3 – groups of descriptors represented different physico-chemical factors (charge, H-bonding, etc) of compound A and B.

Antagonists of fibrinogen receptor
(functional interpretation example)

Antagonists of fibrinogen receptor: dataset



Arg-Gly-Asp



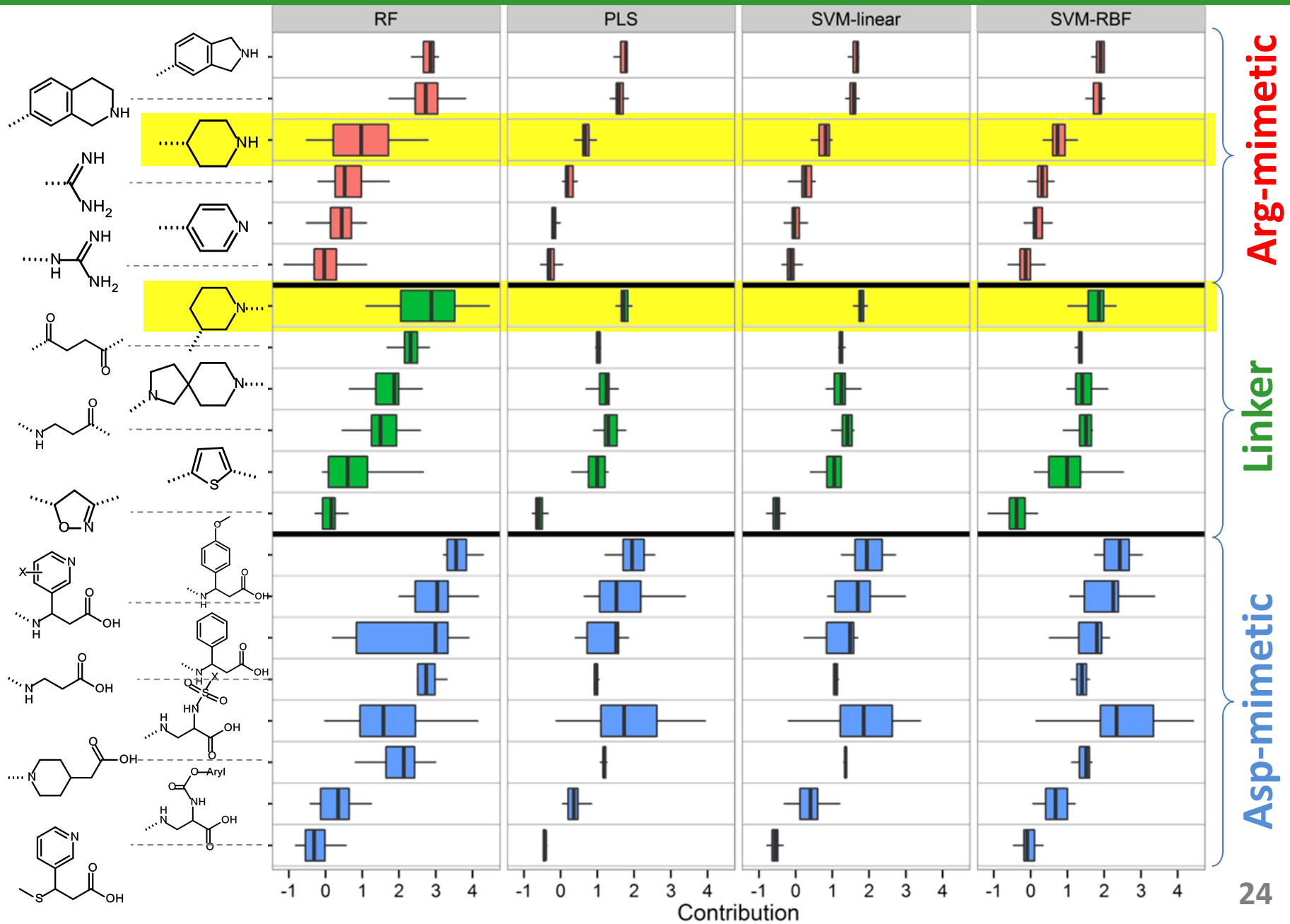
338 compounds

Antagonists of fibrinogen receptor: models

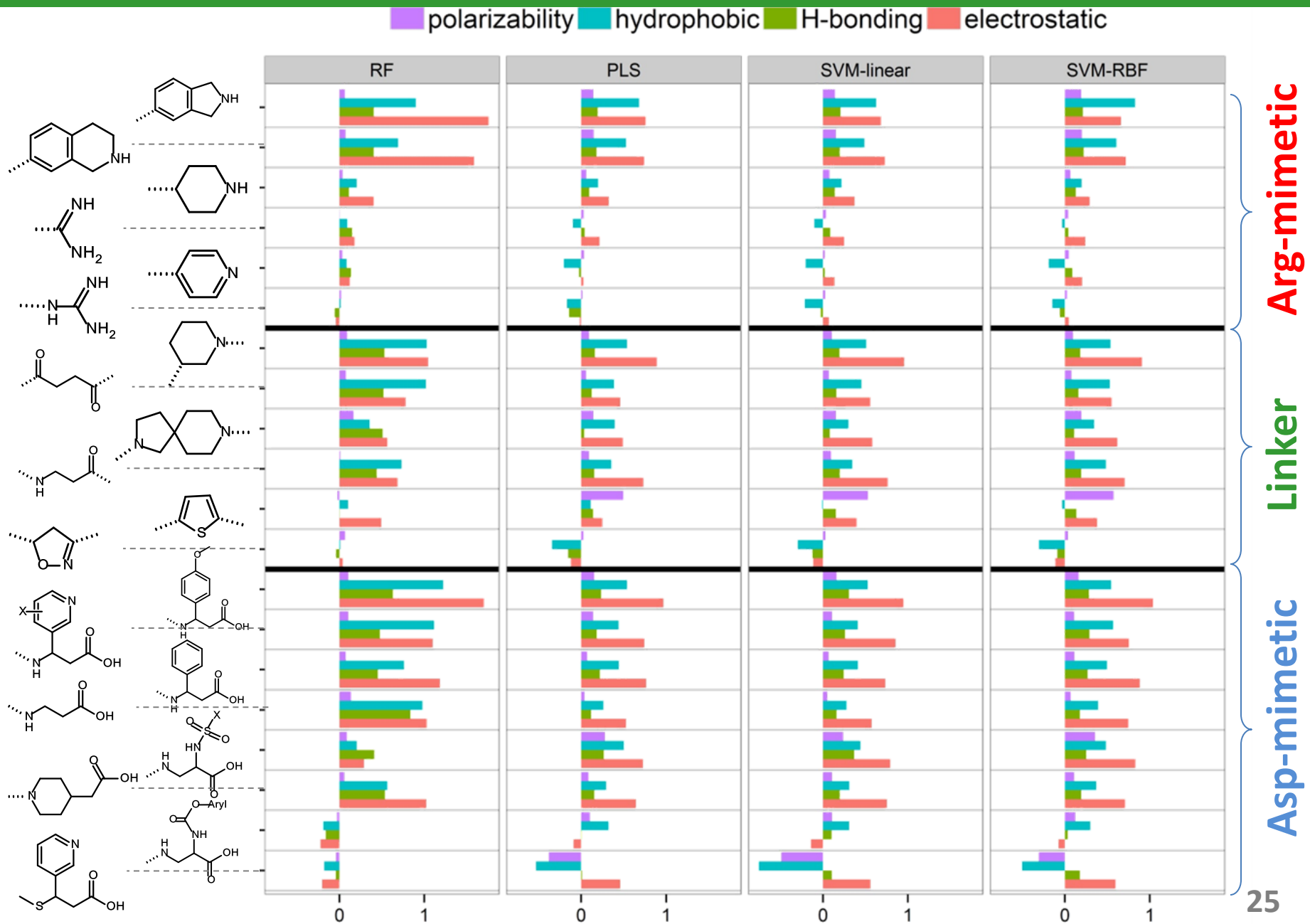
5-fold external cross validation results

Algorithm	R^2	RMSE
RF	0.72	0.80
SVM (RBF kernel)	0.69	0.84
SVM (linear)	0.67	0.88
PLS	0.67	0.87

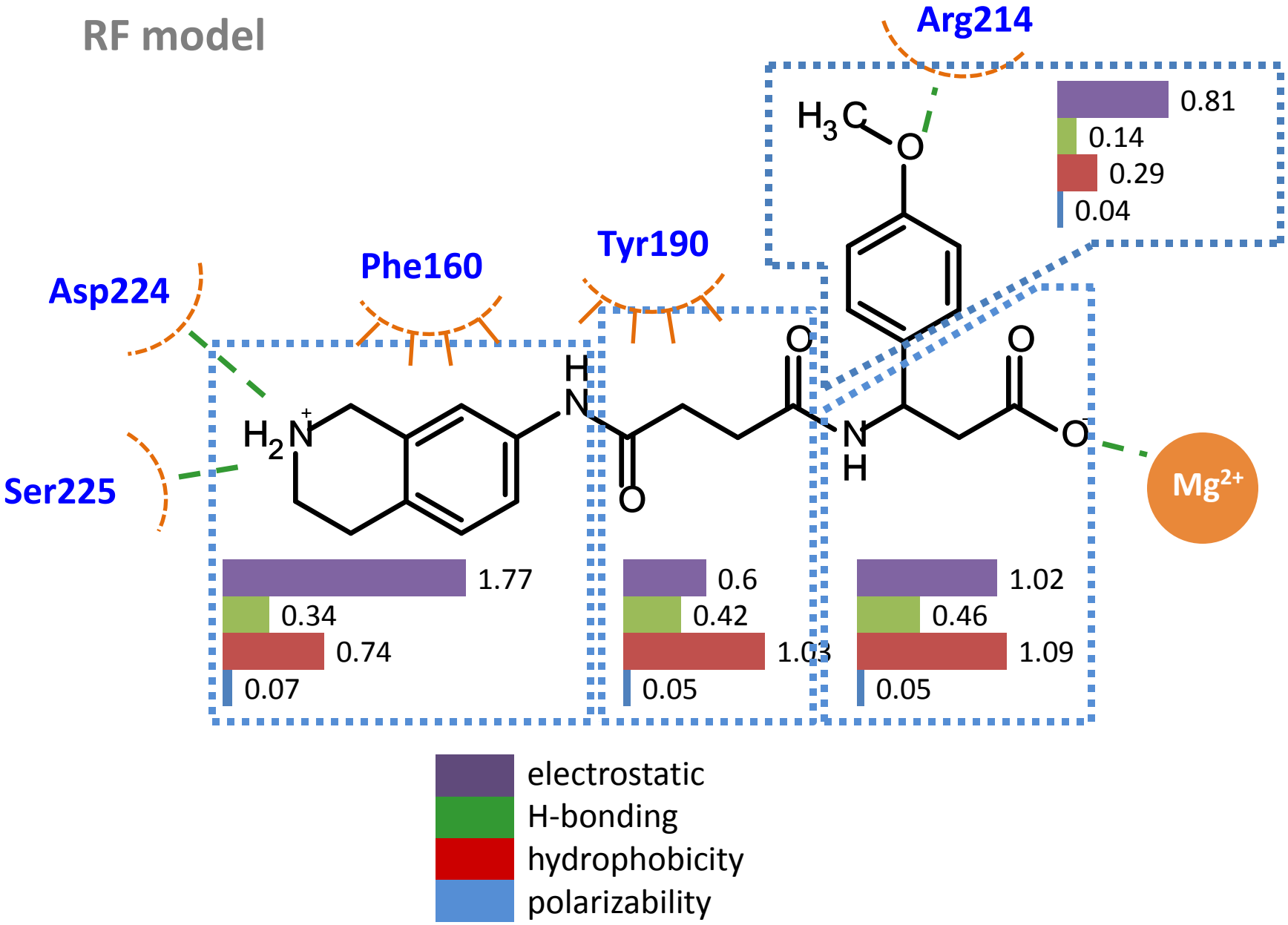
Structural interpretation (global)



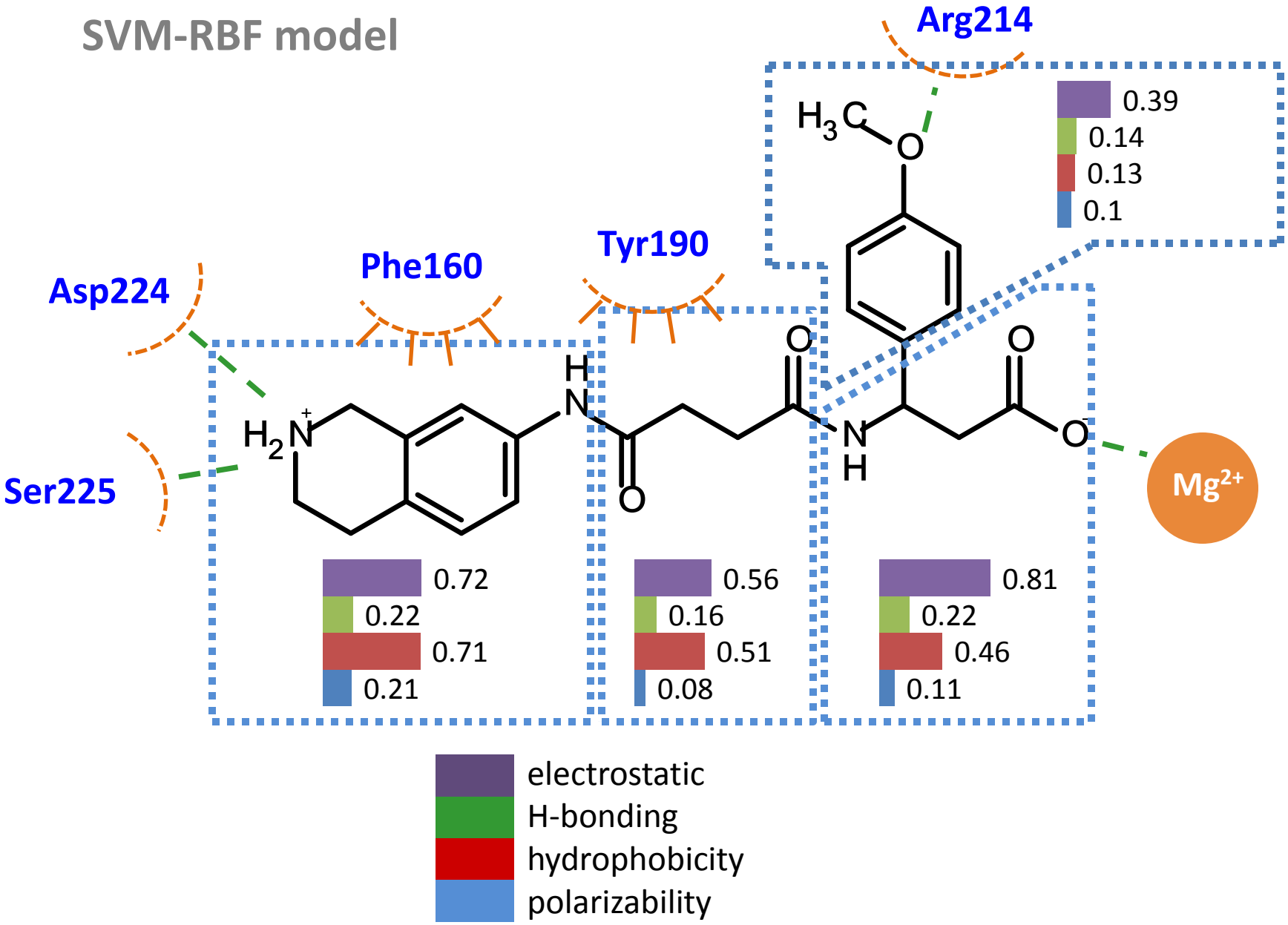
Functional interpretation (global)



Functional interpretation of RF model (local)



Functional interpretation of SVM model (local)

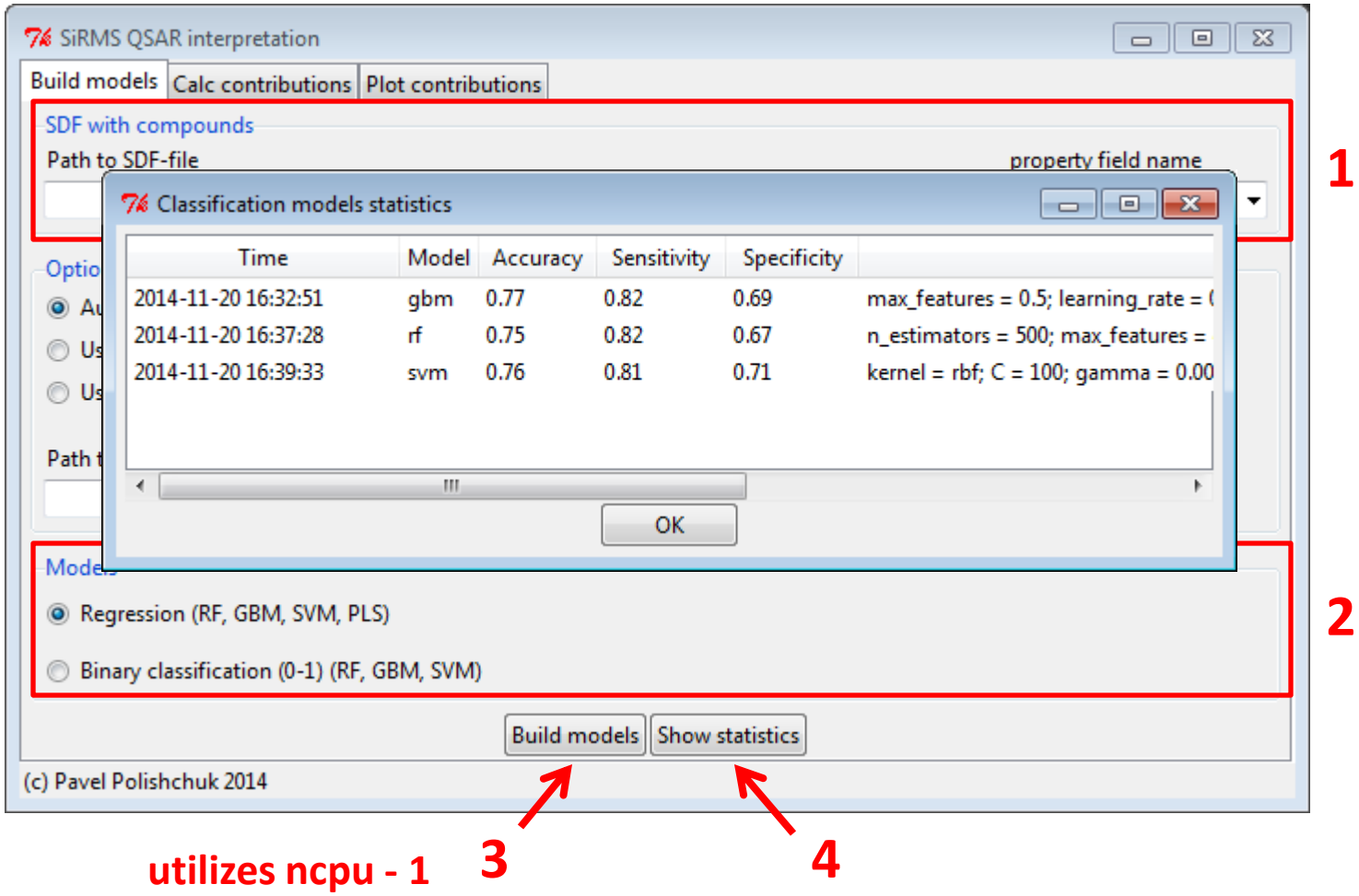


Automatic exploration of datasets of chemical compounds

(dataset mining)

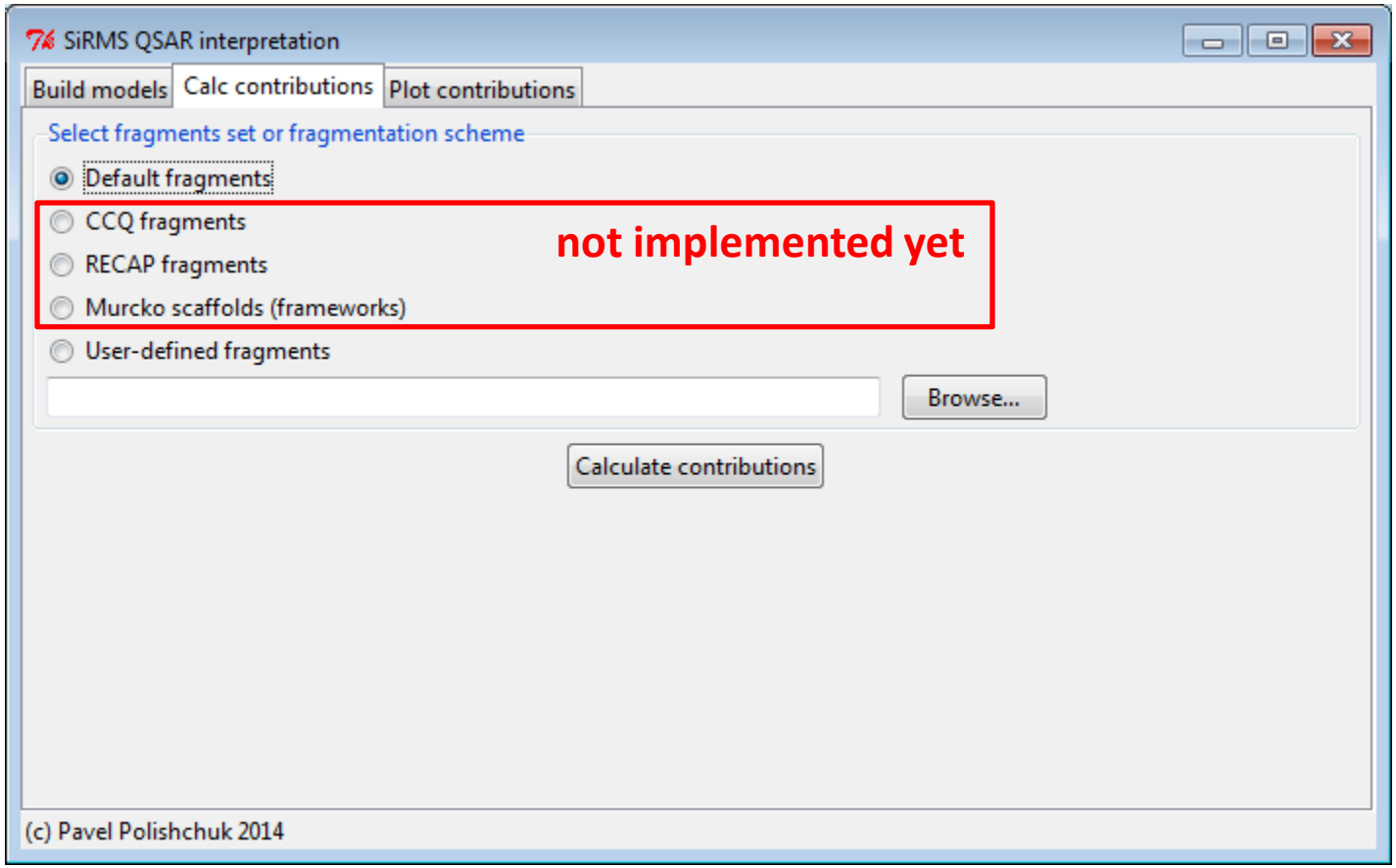
http://qsar4u.com/pages/sirms_qsar.php

QSAR model building



utilizes ncpu - 1

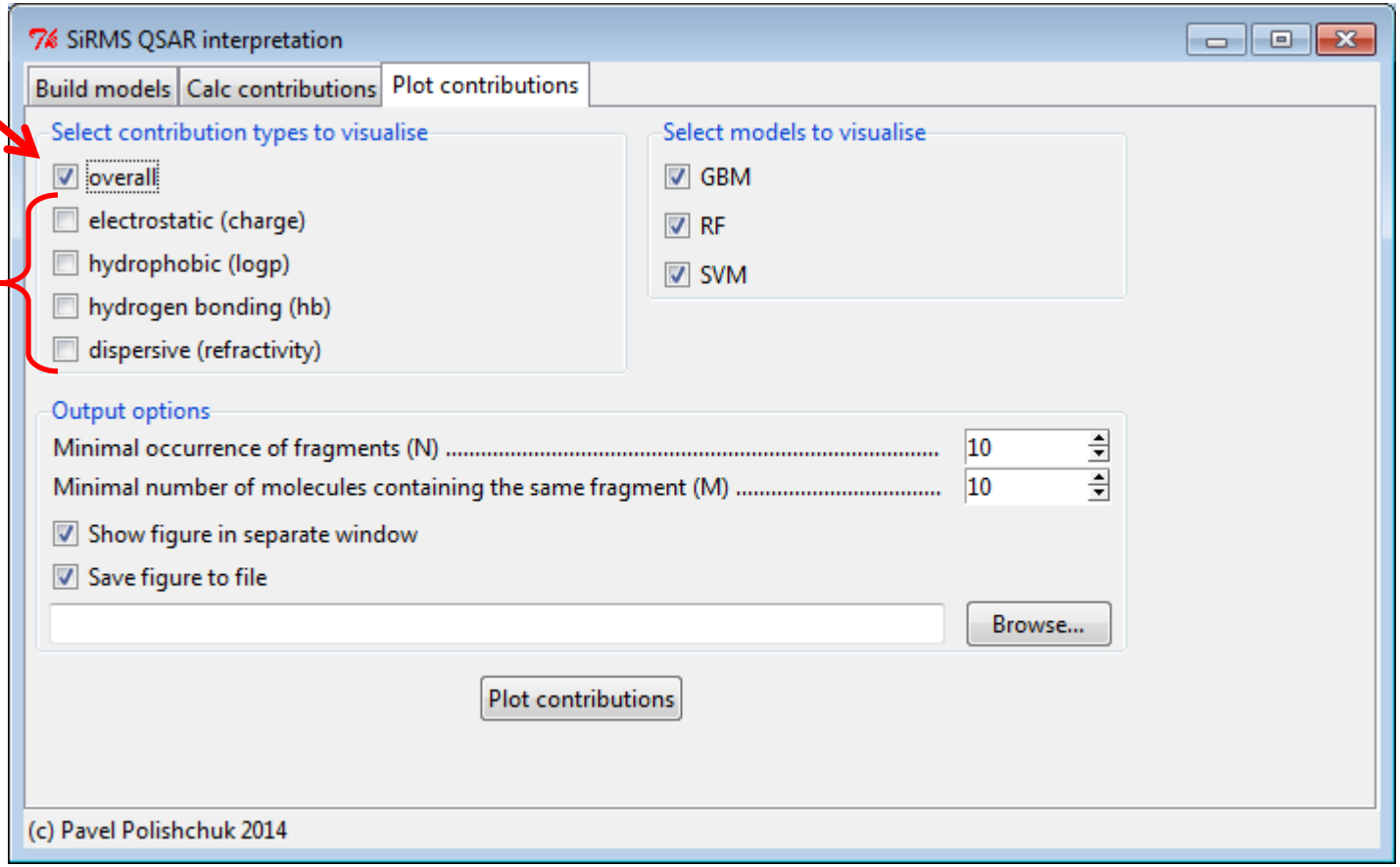
Calculation of fragments' contributions



Plot fragments' contributions

structural interpretation

functional interpretation



External visualization tool

https://pavel.shinyapps.io/sirms-qsar-vis/

Visualization of fragment contributions (demo) - Go to the full version

Input options

Plot type

barplot boxplot

Plot title

Blood-brain barrier

Choose models

- consensus
- gbm
- knn
- rf
- svm

Choose contributions

- overall
- hydrogen bonding
- dispersion
- electrostatic
- hydrophobic

Order by

Model: consensus

Contribution: overall

Filter

Min compounds (M): 10

Min fragments (N): 10

Output options

Width: 35

Height: 28

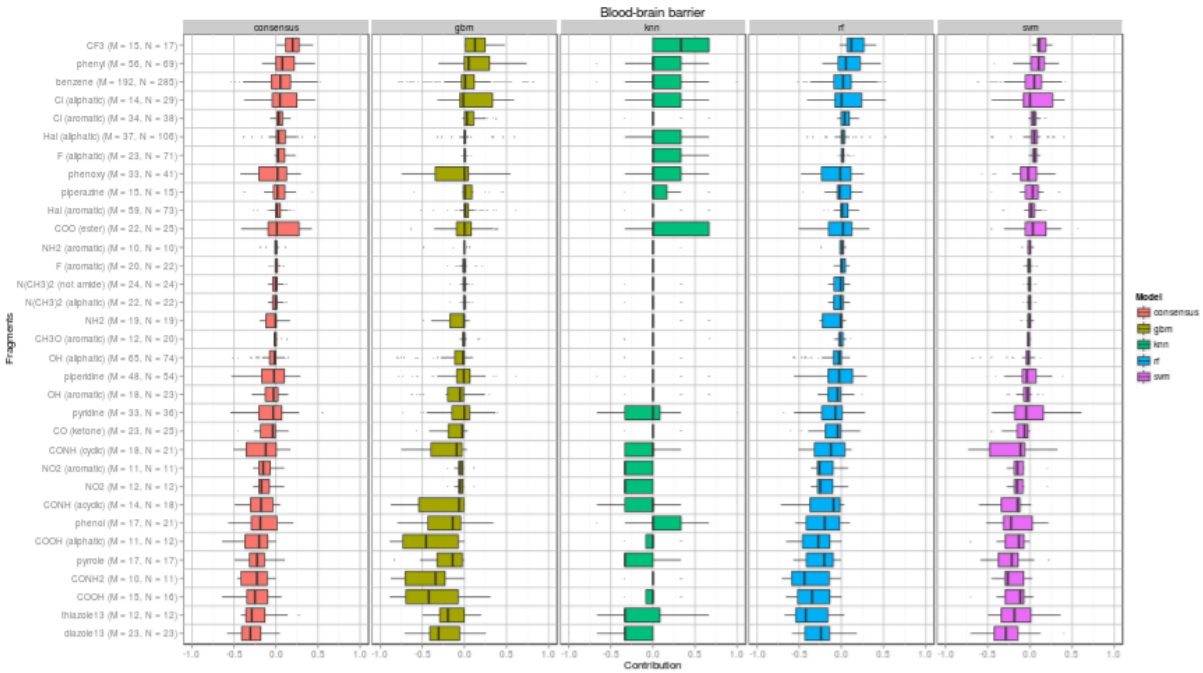
Units: cm

dpi: 300

Download figure

File format

png jpg pdf tiff svg



Interpretation workflow scheme

Create sdf file with property values

Build models (regression or classification)

Look at models stat (if all models are bad reconsider dataset)

Calculate fragment contributions

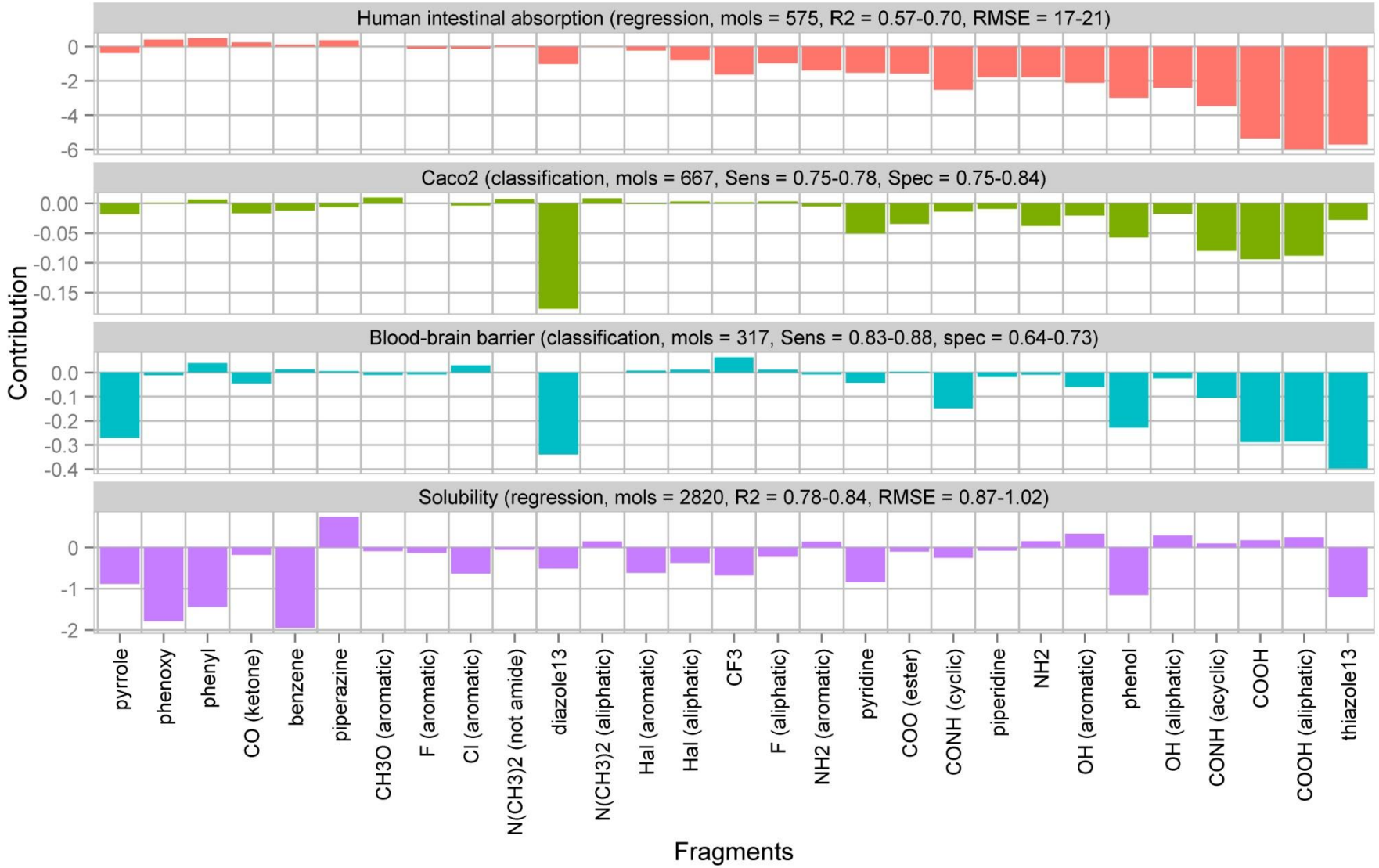
Plot contributions of desired models selected from statistically significant ones

ADME/Tox examples
(SAR trends, global interpretation)

Datasets taken from:

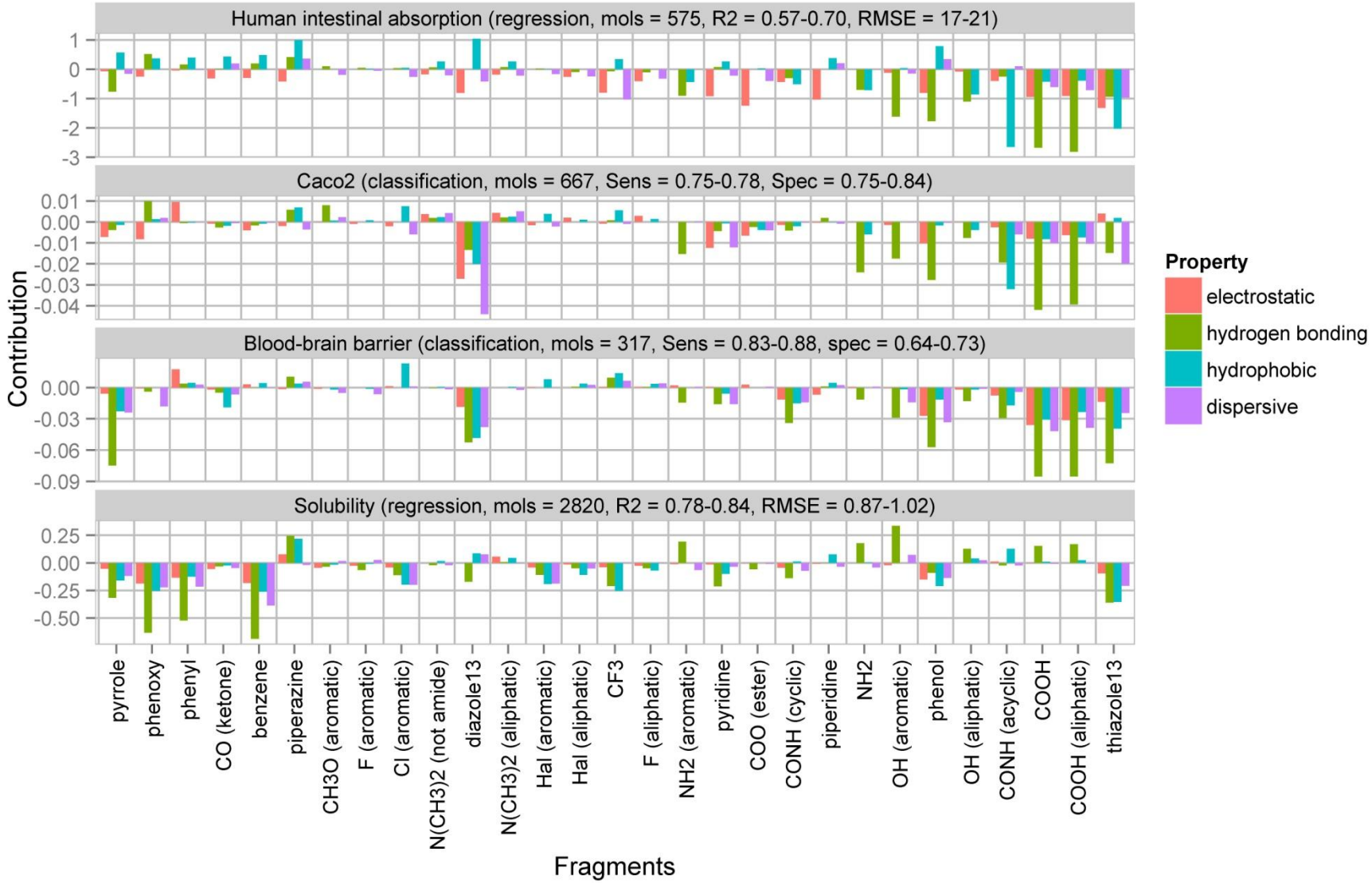
1) Cheng W. et al., J. Chem. Inf. Model., **2012**, 3099-3105
2) Kovdienko N.A. et al., Molecular informatics, **2010**, 394-406
3) Polishchuk P.G. et al., J. Chem. Inf. Model., **2009**, 2481-2488
4) in-house data

Permeability (structural interpretation)



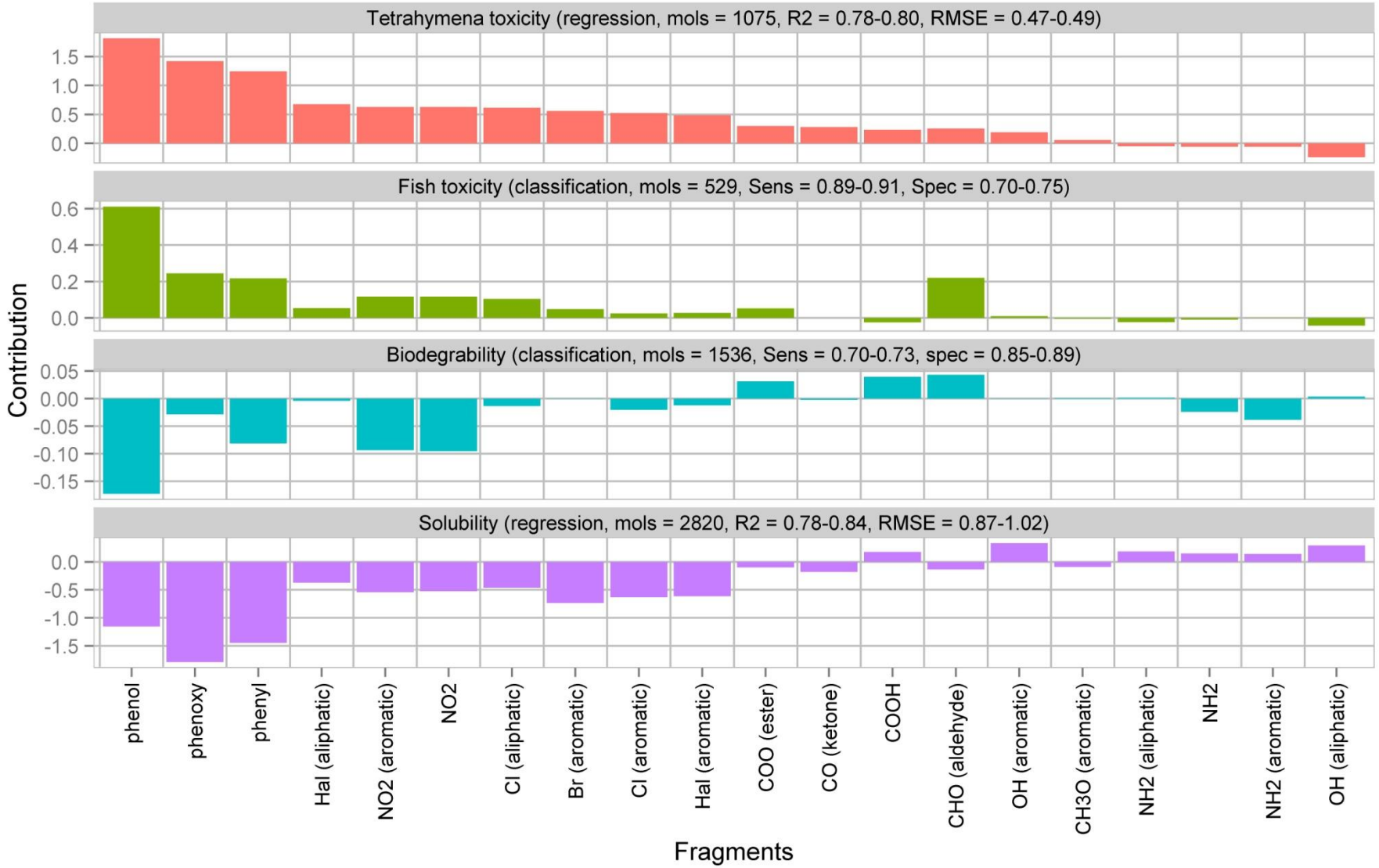
consensus of RF, GBM, SVM models

Permeability (functional interpretation)



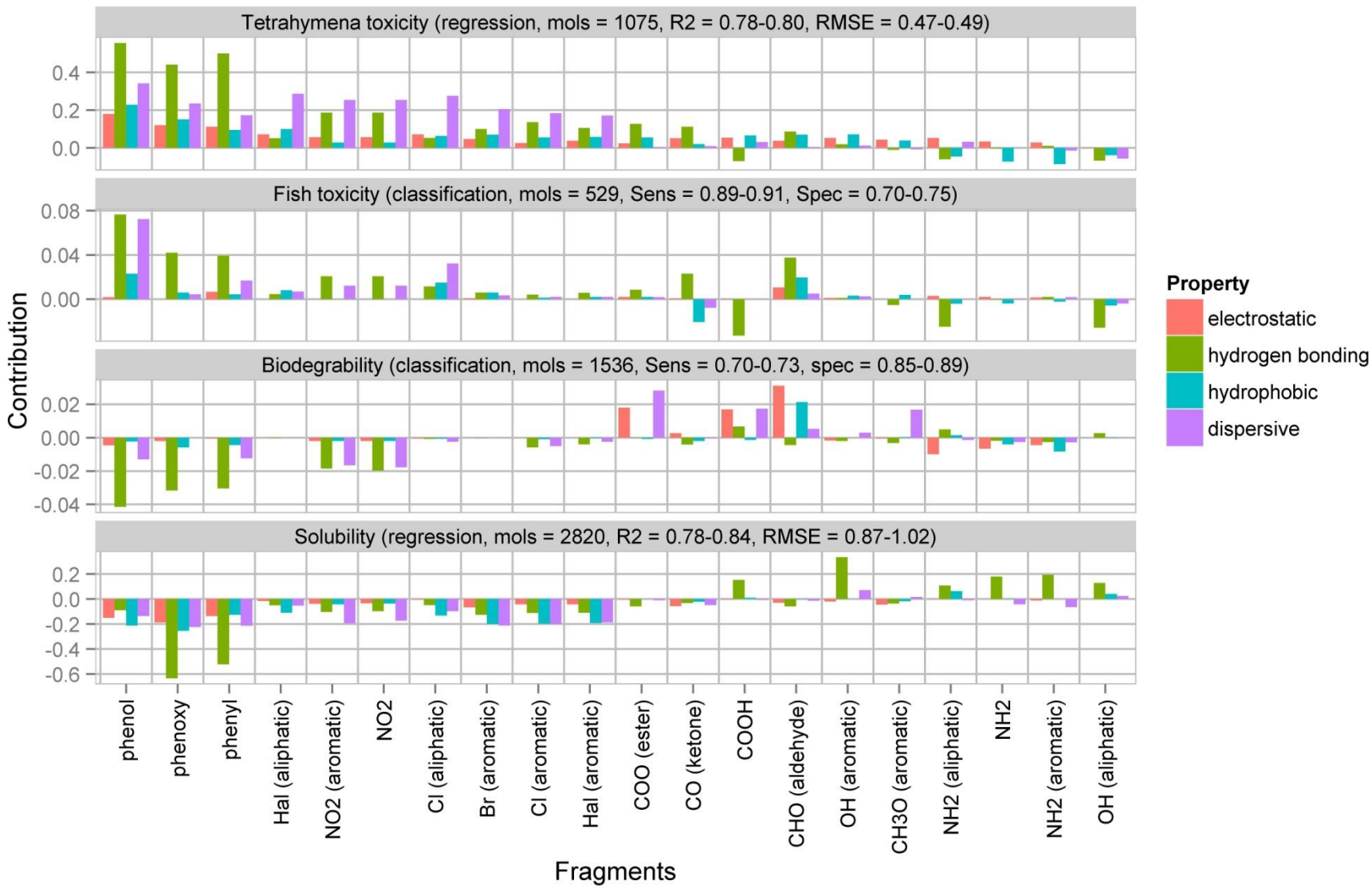
consensus of RF, GBM, SVM models

Toxicity (structural interpretation)



consensus of RF, GBM, SVM models

Toxicity (functional interpretation)



consensus of RF, GBM, SVM models

Summary

		Descriptors	
		SiRMS	Others (Dragon, CDK, etc)
Models	Regression	+	+
	Classification	+	+
Fragments	Terminal (substituent)	+	+
	Scaffold/linker	+	-
Interpretation	Structural	+	+
	Functional	+	?

Conclusions

Almost any QSAR model can be interpreted using the proposed schemes.

Results of structural and functional interpretation obtained from different models are well correlated between models and correspond to observed trends.

Structural interpretation allows to reveal trends in SAR, rank fragments, find potential structural alerts, etc.

Functional interpretation may provide a guess about factors which are dominated and influence on the investigated property.

Perspectives

Smart automatic fragmentation approaches

Detection of potential activity cliffs in local interpretation

Testing on other types of descriptors

Usage of datasets which include mixtures of compounds

Application of this approach for wider range of structurally diverse datasets with different end-points and comparison to MMP

Useful web links

A.V. Bogatsky Physico-Chemical Institute, Chemoinformatic group:
http://qsar4u.com

SiRMS project on GitHub:
https://github.com/DrrDom/sirms

SiRMS-QSAR (dataset analysis):
http://qsar4u.com/pages/sirms_qsar.php

External web-based visualization:
https://pavel.shinyapps.io/sirms-qsar-vis/

Acknowledgement

A.V. Bogatsky
Physico-Chemical Institute
(Odessa, Ukraine)

Prof. V. Kuz'min
Dr. T. Khristova
Dr. L. Ognichenko
A. Kosinskaya
E. Mokshina
M. Kulinskiy

Strasbourg University
(France)

Prof. A. Varnek
Dr. D. Horvath