Methods and Applications of Computational Chemistry - 5
Kharkiv, Ukraine, 1 – 5 July 2013

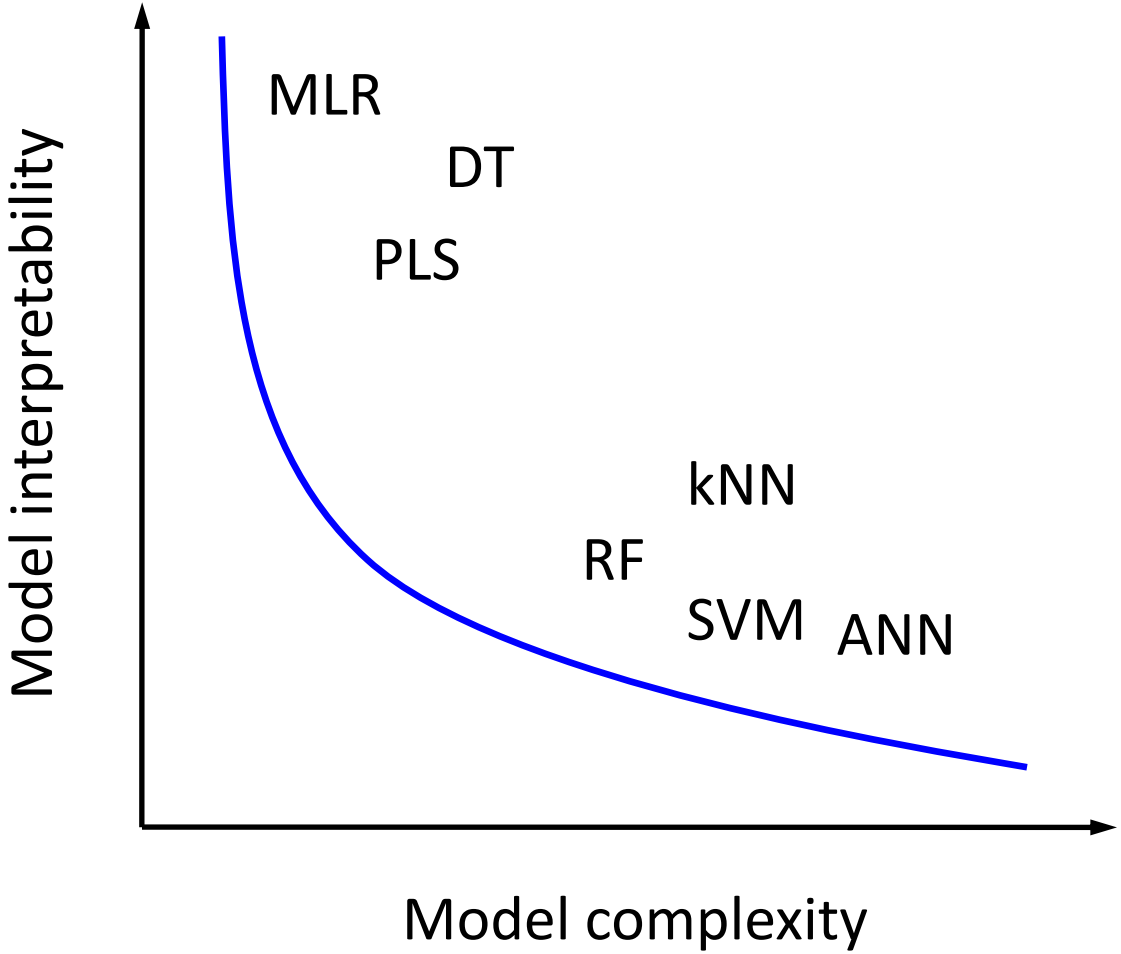
Structural interpretation of QSAR models – a universal approach

Victor Kuz'min, Pavel Polishchuk,
Anatoly Artemenko, Eugene Muratov

A.V. Bogatsky Physico-Chemical Institute
of National Academy of Sciences of Ukraine
Odessa, Ukraine
pavel_polishchuk@ukr.net

QSAR interpretation: interpretability *vs.* complexity

Popular misbelief



QSAR interpretation approaches

Model-specific approaches:

Rule-based (Decision tree)

Regression coefficients (MLR, PLS)

Latent variables (PLS)

Weights and biases (ANN)

Model-independent approaches:

Variable importance

Local gradients or partial derivatives

Model-independent interpretation approaches

Variable importance

$$\mathrm{Imp}_i = \mathrm{MSE}(x_i) - \mathrm{MSE}(x_i^{permut})$$

Local gradients or partial derivatives

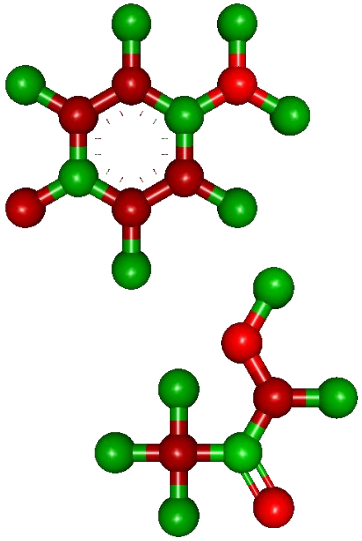
$$\mathrm{C}_i = \frac{f(x_i) - f(x_i + \Delta x_i)}{\Delta x_i}$$

QSAR interpretation: common workflow

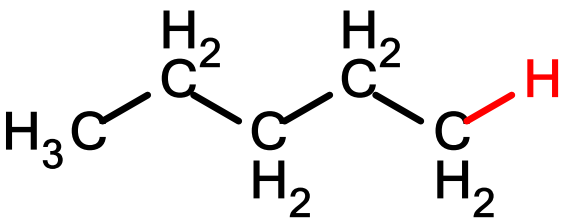
Model → **Variables contributions** → **Structure-property relationship**

$f(x)$

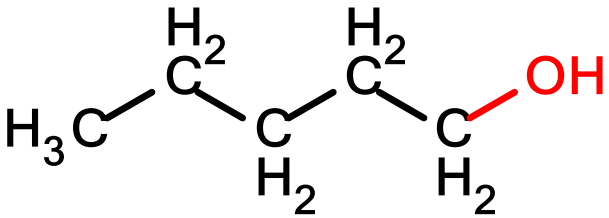
	Var_1	Var_2
Mol_1	-0.23	1.82
Mol_2	2.36	1.27
Mol_3	5.01	2.30
Mol_4	0.69	-0.58



Matched molecular pairs approach



logS = -3.18

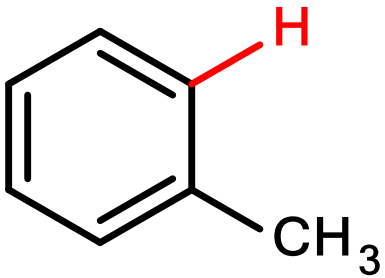


logS = -0.60

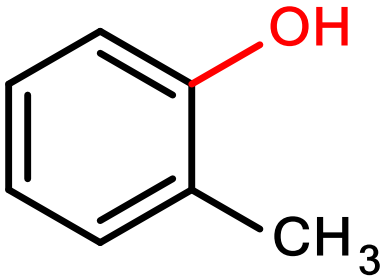
ΔlogS = 2.58

H → OH

ΔlogS = 1.59

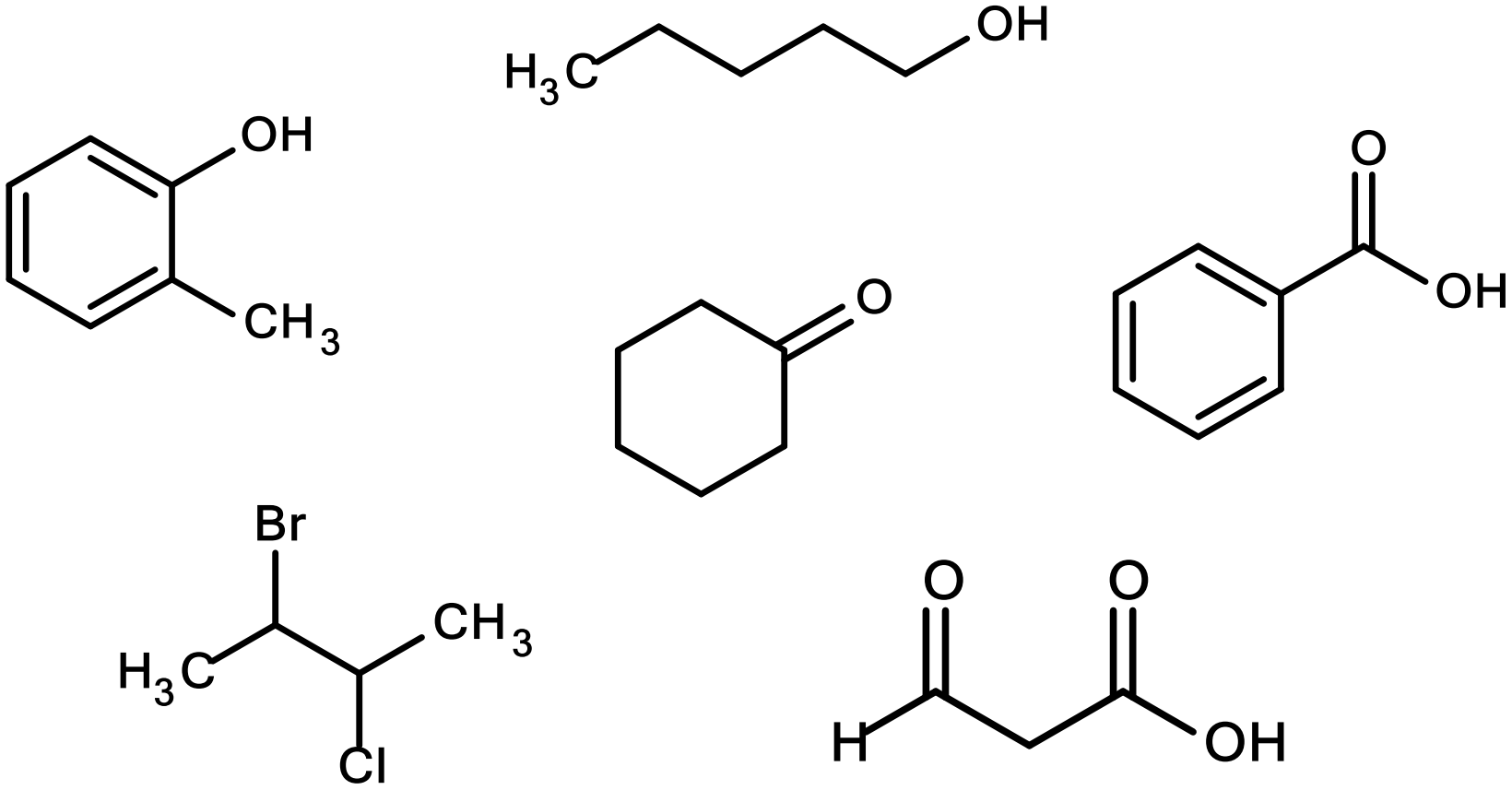


logS = -2.21

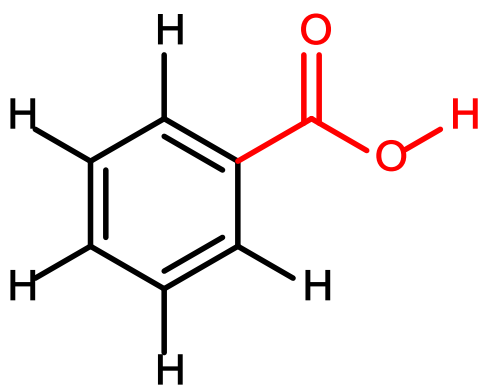


logS = -0.62

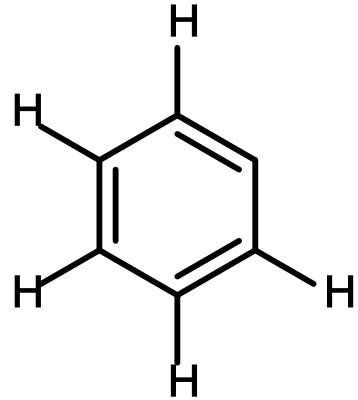
Exemplified dataset



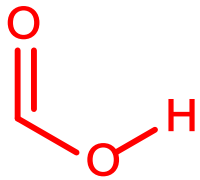
Universal structural QSAR interpretation



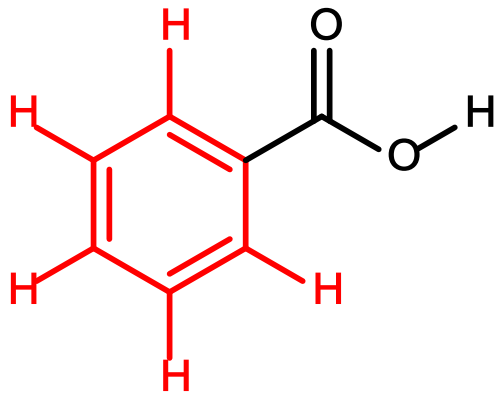
-



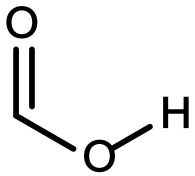
=



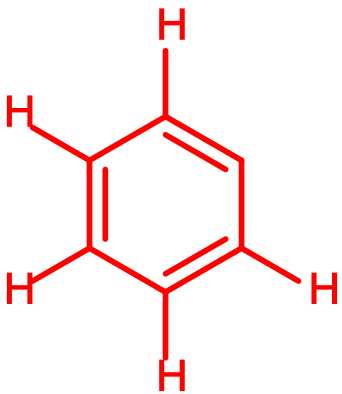
$\log S_{pred}$ = -1.55 $\log S_{pred}$ = -1.61 $\Delta\log S_{pred}$ = 0.06



-

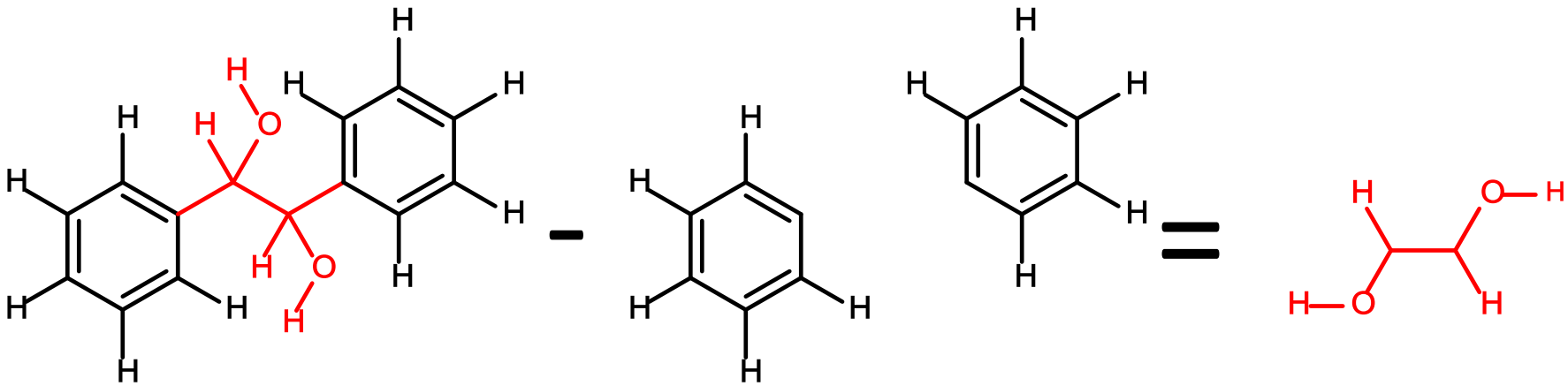


=



$\log S_{pred}$ = -1.55 $\log S_{pred}$ = -1.35 $\Delta\log S_{pred}$ = -0.20

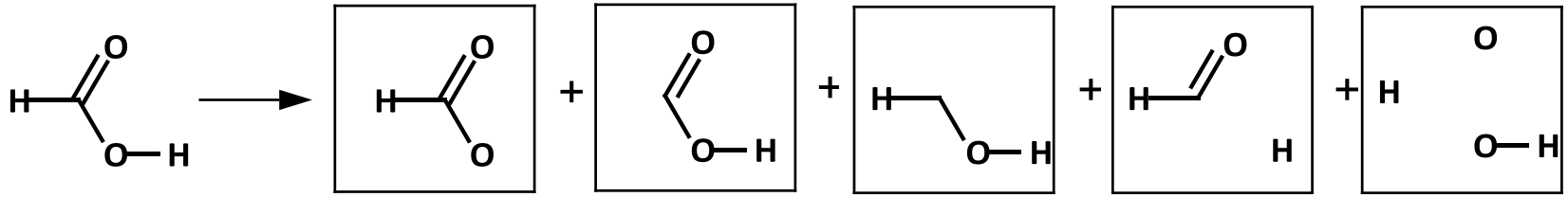
Universal structural QSAR interpretation



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

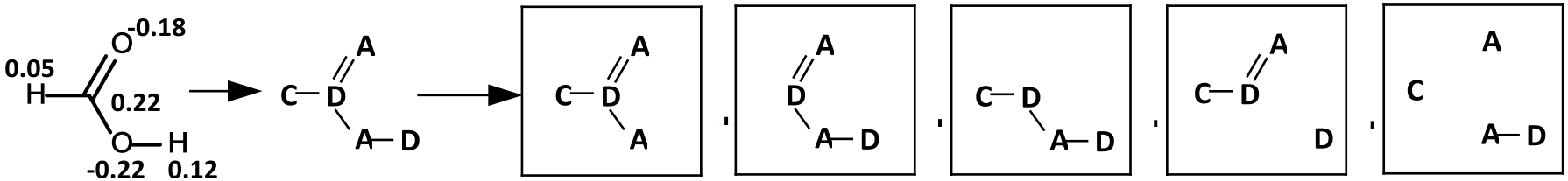
Simplex representation of molecular structure (SiRMS)

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

Case studies

End points:

Solubility (regression)

Inhibition of Transglutaminase 2 – TG2 (regression)

Mutagenicity (binary classification)

Descriptors:

Simplex representation of molecular structure (SiRMS)

Dragon

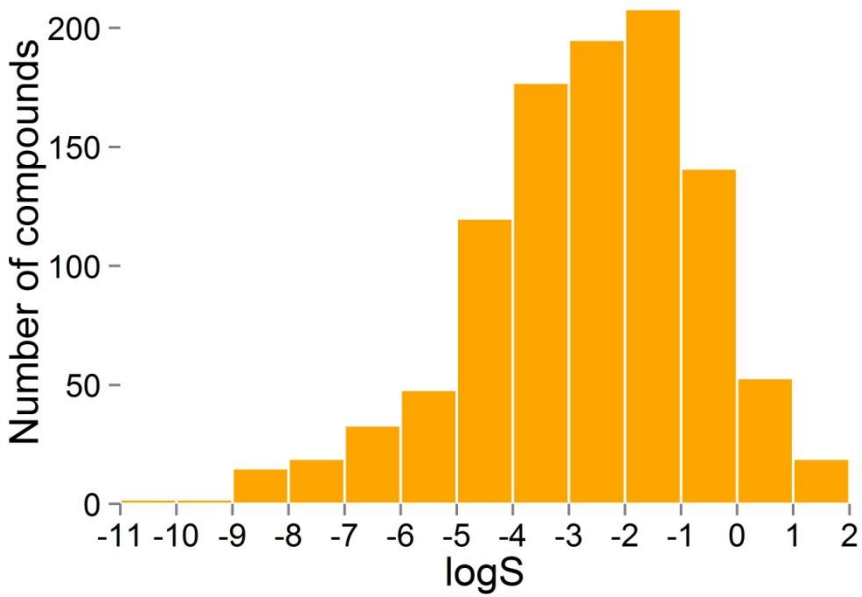
Machine learning methods:

Random Forest (RF)

Support vector machine (SVM)

Projects to latent structures (PLS)

Solubility: dataset and models



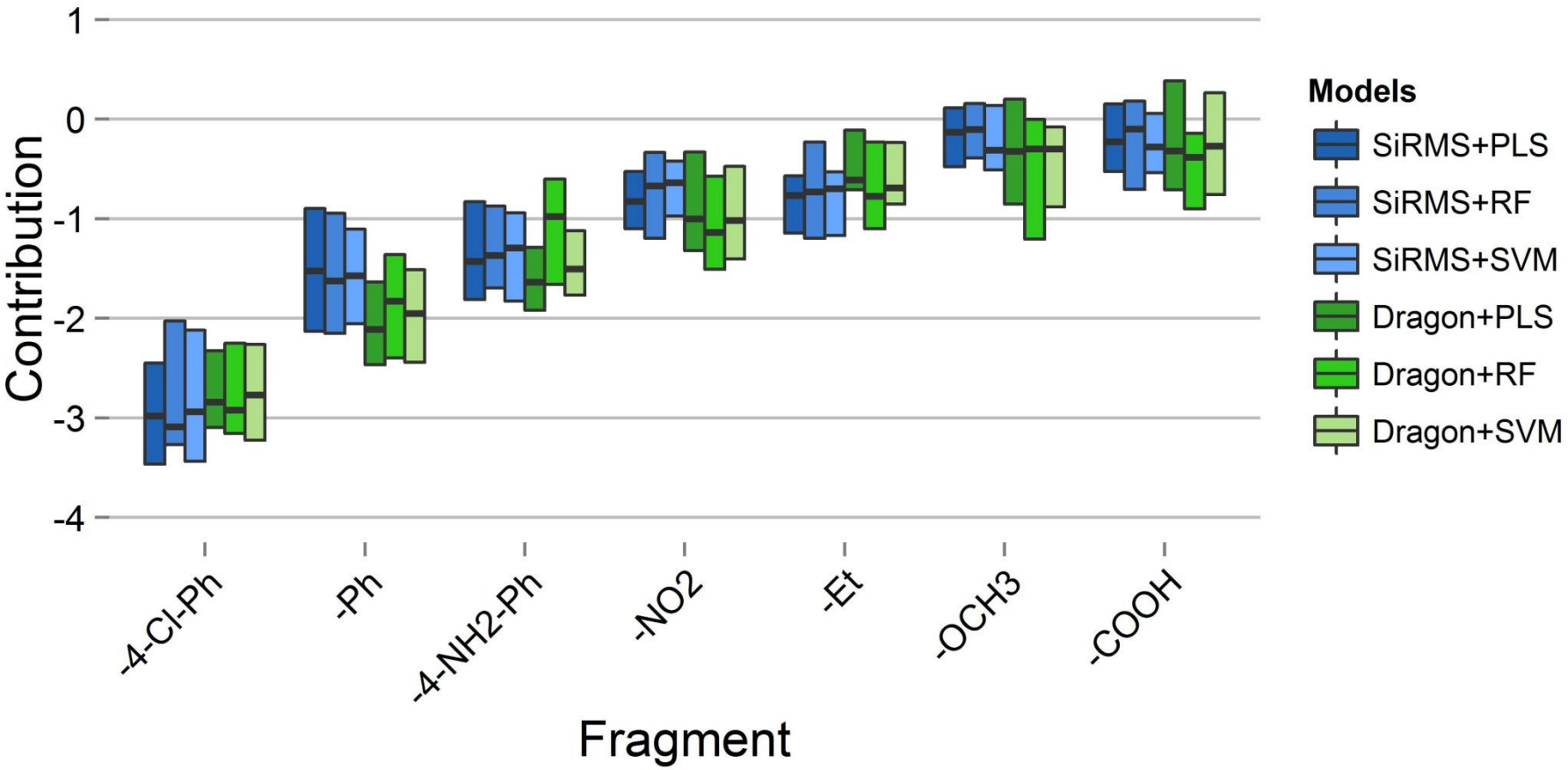
Overall number of compounds
1033

Huuskonen, J.
J. Chem. Inf. Comp. Sci. ***2000***, *40, 773-777*

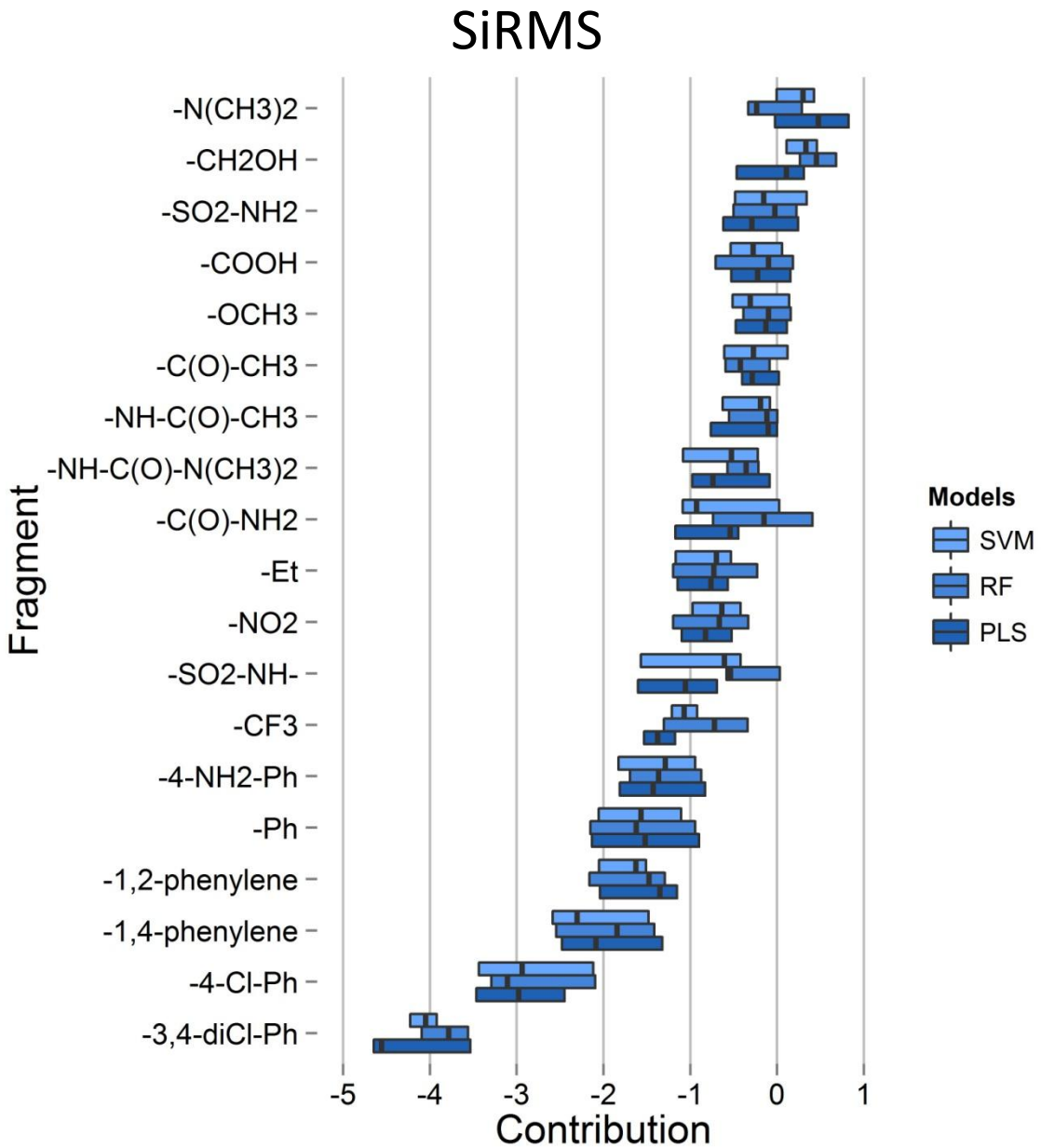
5-fold external cross validation results (10 runs)

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

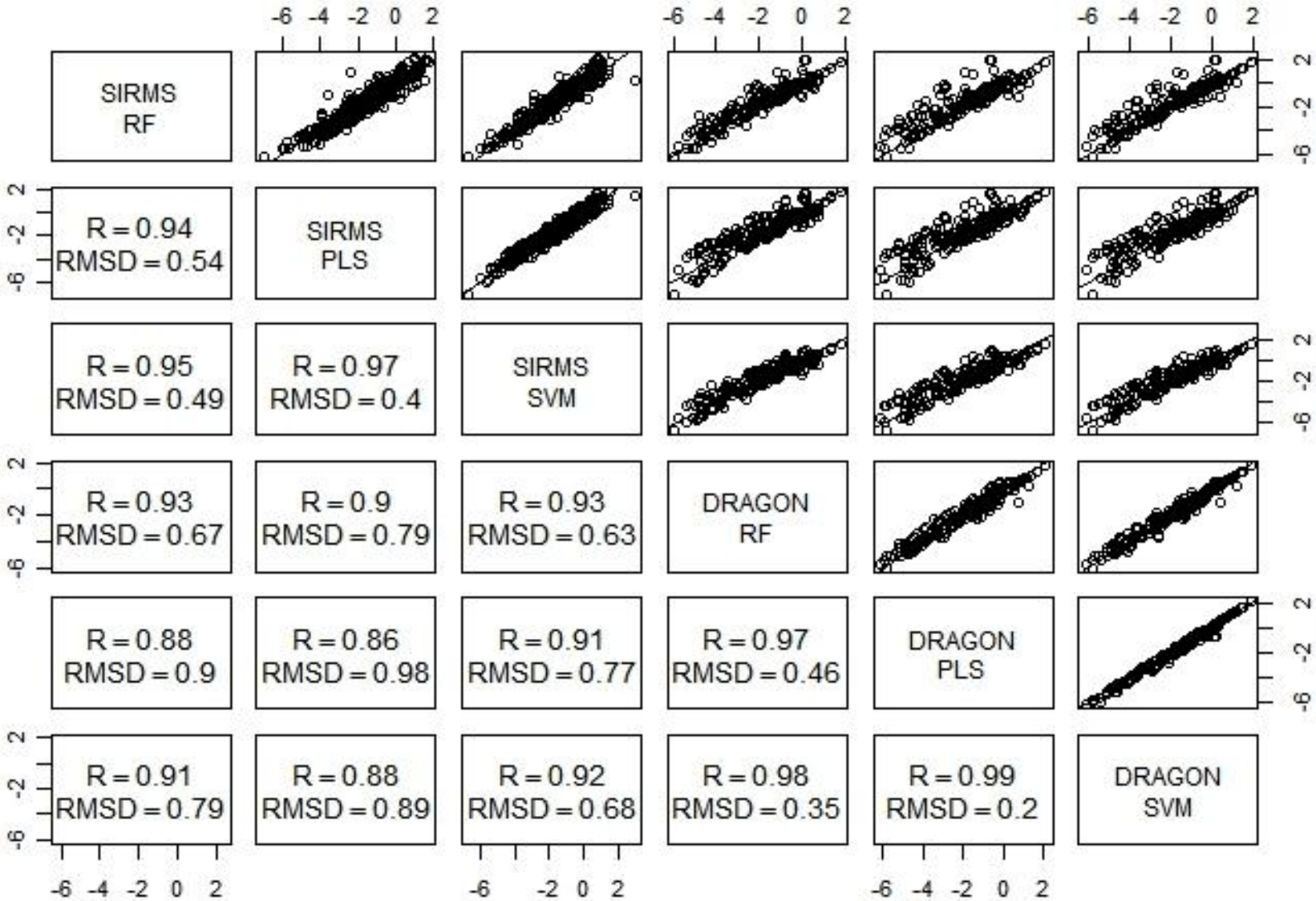
Solubility: interpretation SiRMS *vs.* Dragon



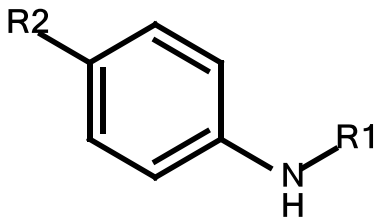
Solubility: fragment ranking



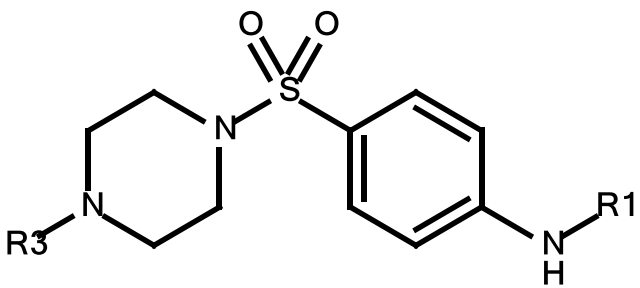
Solubility: pair-wise contribution correlations



Transglutaminase 2 inhibition: dataset and models



R1 = acyl groups(preferably acryl);
R2 = NO_2, F, Br, CF_3, CH_3, OCH_3.



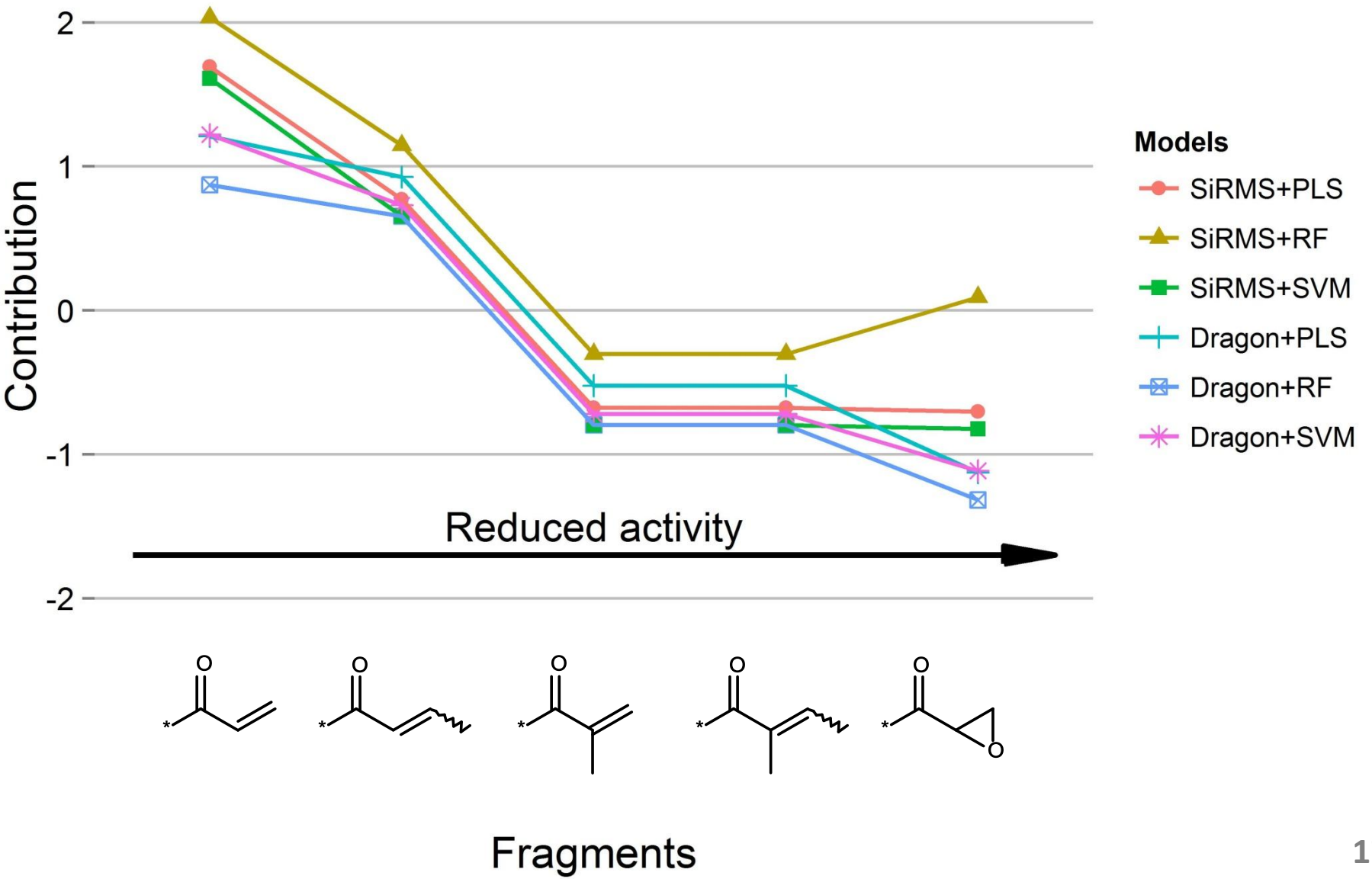
R1 = acyl groups (preferably acryl and its derivatives);
R3 = acyl groups (preferably Boc, Cbz and its derivatives), substituted phenyl and pyridyl.

Prime, M. E. et al, *J. Med. Chem.* **2012**, 55, 1021-1046.

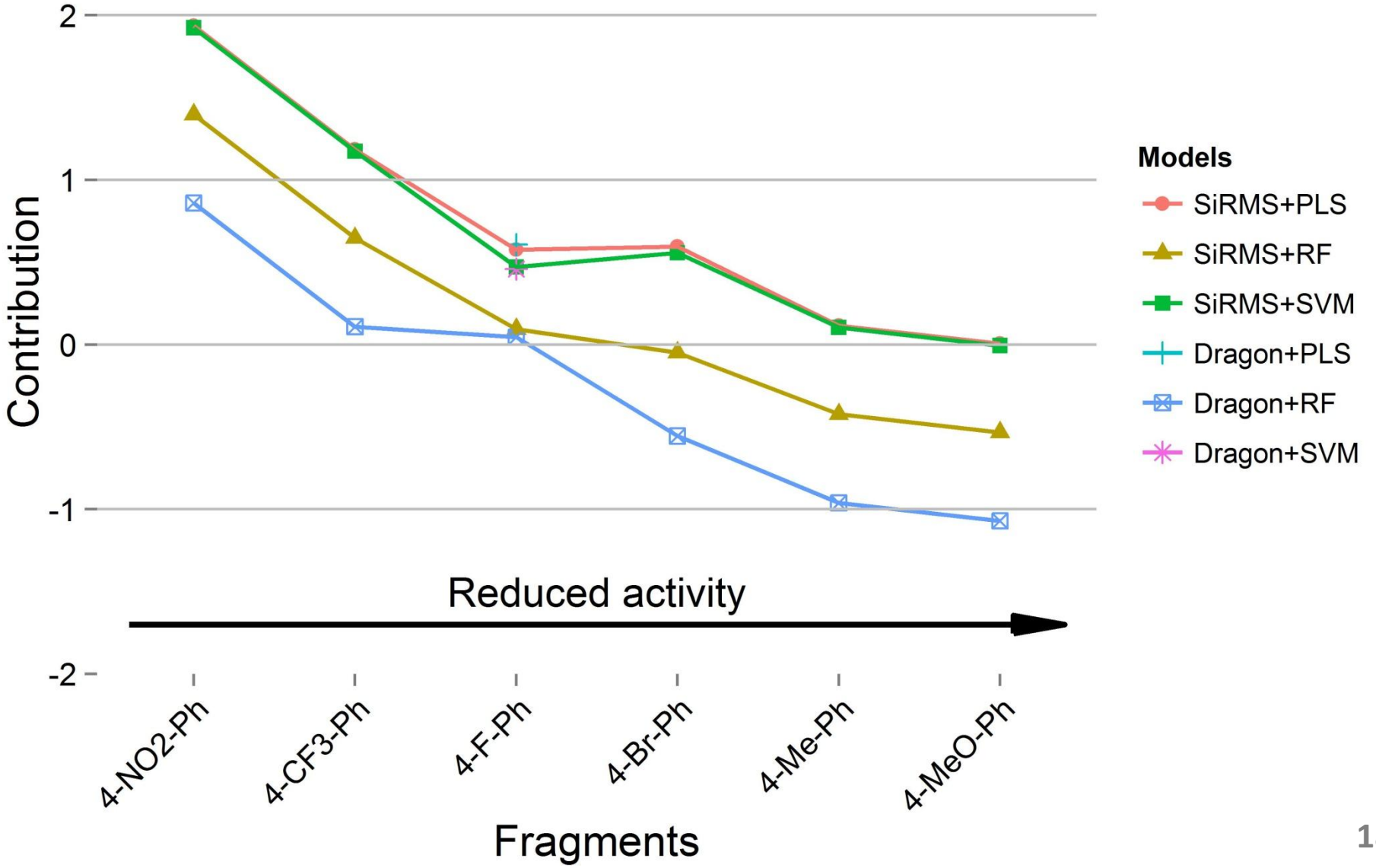
5-fold external cross validation results (10 runs)

Endpoint	Model	SiRMS		Dragon	
		R^2_{CV}	RMSE	R^2_{CV}	RMSE
TG2 inhibition, pIC_{50}	PLS	0.70	0.67	0.65	0.72
	RF	0.74	0.62	0.64	0.74
	SVM	0.70	0.67	0.68	0.70

TG2 inhibition: ranking R1 substituents



TG2 inhibition: ranking R2 substituents



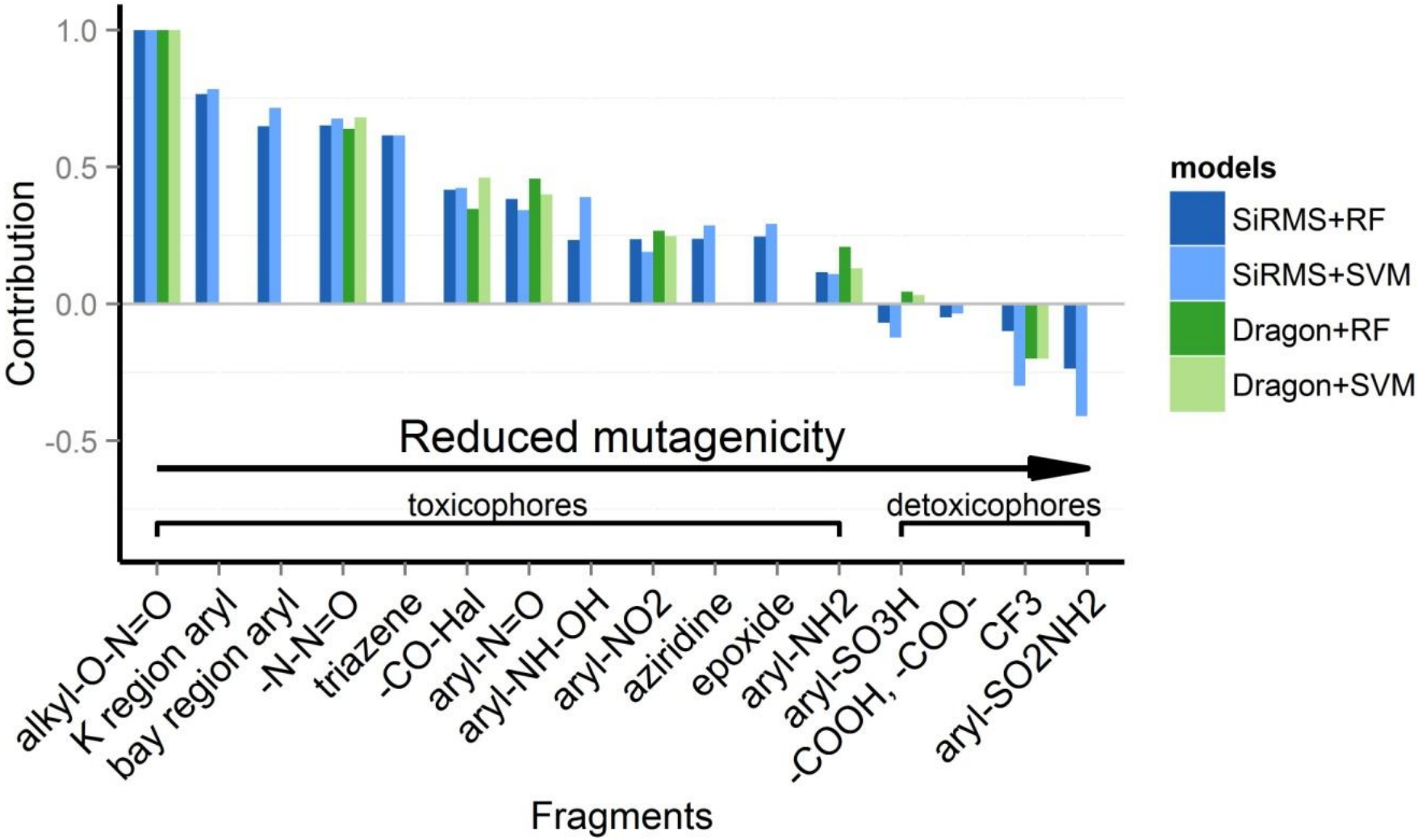
Ames mutagenicity: dataset and models

	2344	mutagens
+	2017	non-mutagens
	4361	overall

5-fold external cross validation results (10 runs)

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

Ames mutagenicity: fragments ranking



Universal structural QSAR interpretation: benefits

Estimation of contribution of fragments with single (terminal groups) and multiple attachment points (scaffolds or linkers)

Non-additivity of calculated contributions (depends on an investigated property)

Estimation of mutual fragment influence on a property

Calculated fragment contributions are independent from used descriptors and machine learning methods

Related projects

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

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