Kazan Summer School of Chemoinformatics
Kazan, Russia, 6-9 July 2015

Applications of the mixtures representation approach in QSAR modeling

P. Polishchuk[1], E. Mokshyna[1], E. Varlamova[2], E. Muratov[3], T. Madzhidov[4], V. Kuz'min[1]

[1] A.V. Bogatsky Physico-Chemical Institute of National Academy of Sciences of Ukraine, Odessa, Ukraine;
[2] Laboratory for Molecular Modeling and Drug Design, Faculty of Pharmacy, Federal University of Goiás, Goiás Brazil;
[3] University of North Carolina, Chapel Hill, NC, USA;
[4] A.M. Butlerov Institute of Chemistry, Kazan Federal University, Kazan, Russia.
pavel_polishchuk@ukr.net

Mixture modeling tasks

Modeling of mixtures (examples):

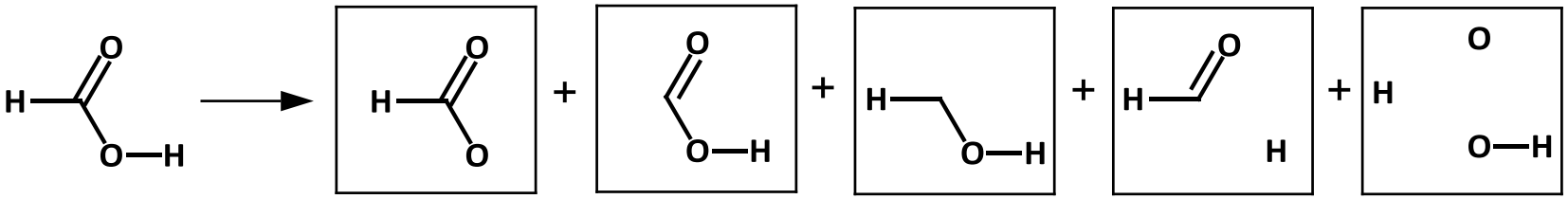
- Toxicity assessment (*Daphnia magna*, *Vibrio fischeri, Photobacterium phosphoreum, etc*)
- Joint effects of antivirals
- Physical properties of binary mixtures

Application of mixture modeling approach to other tasks:

- Chemical reactions
- Ligands binding to their targets
- Macroscopic properties of pure chemicals

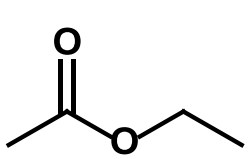
Simplex representation of molecular structure (SiRMS)

Simplex generation example

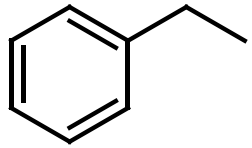


Kuz'min, V. E. et al, *Journal of Molecular Modeling* **2005**, 11, 457-467.
Kuz'min, V. E. et al, *Journal of Computer-Aided Molecular Design* **2008**, 22, 403-421.

Mixture representation (M-SiRMS)

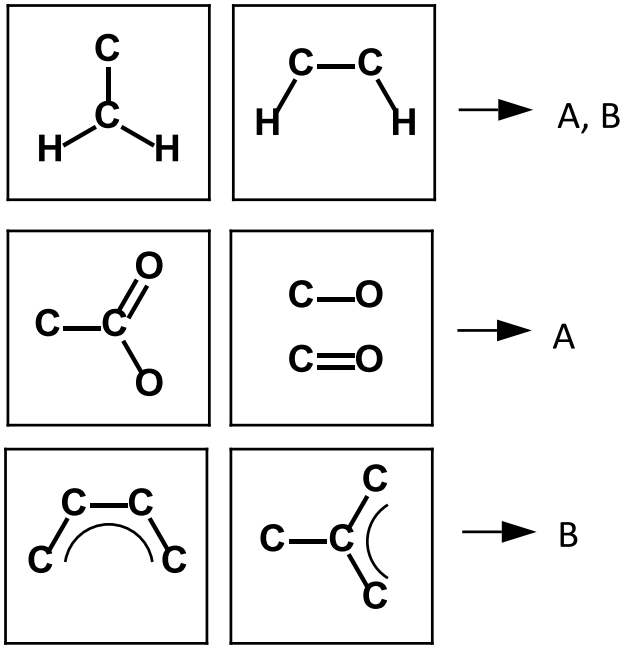


A

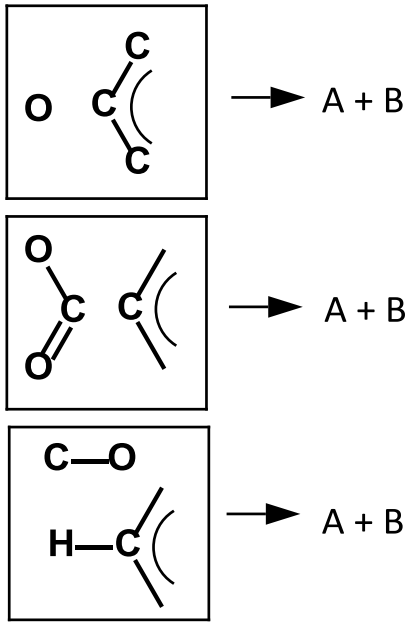


B

Single component



Mixture

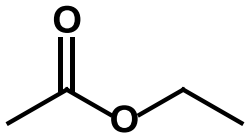


$x_1D_1 + x_2D_2$

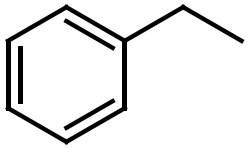
x_1D_{1+2} $(x_1 < x_2)$

Muratov, E. N. et al. *Molecular Informatics* **2012**, 31, 202-221.

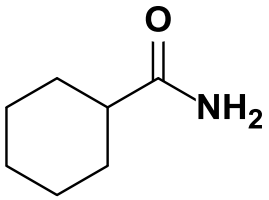
Mixture representation (M-SiRMS)



A



B



C

single compounds descriptors

mixture descriptors

A

B

C

A+B

B+C

A+C

D = aA + bB + cC

D = (A+B) + (B+C) + (A+C)

A,B,C

A+B+C

descriptors of A, B and C

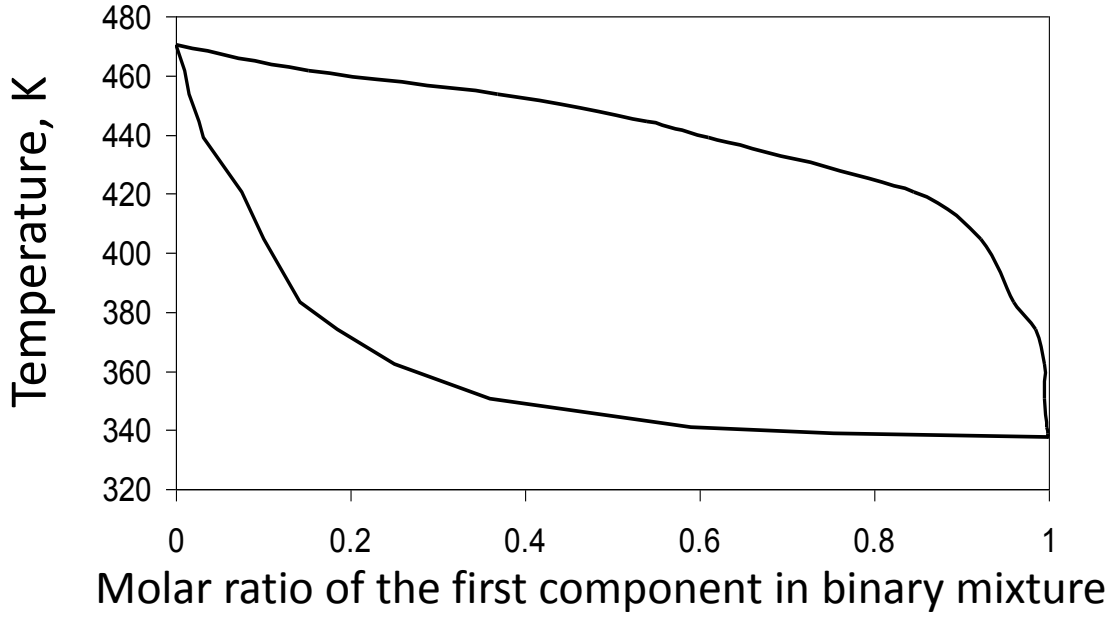
Task

Vapor-liquid equilibrium of binary mixtures of organic solvents

Modeling of vapor-liquid equilibrium curves

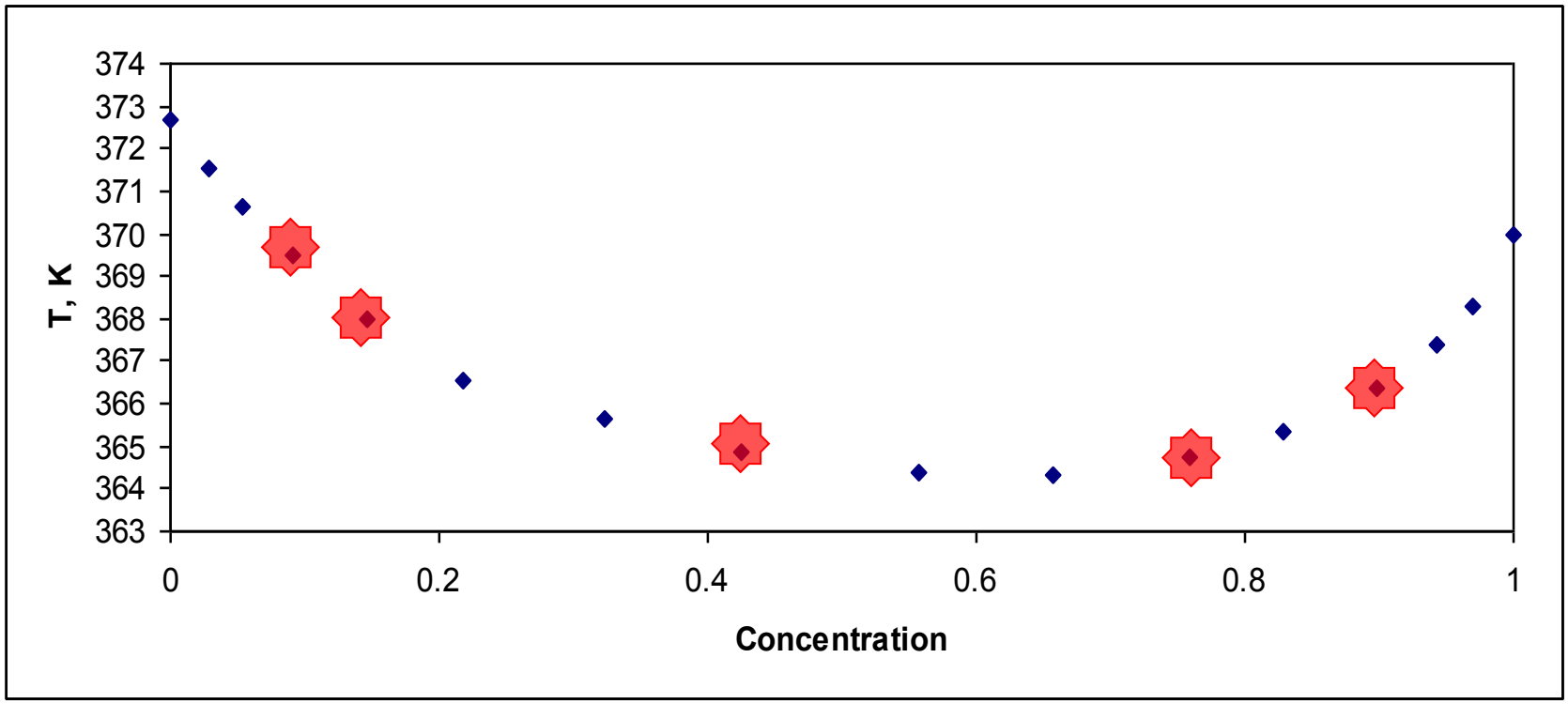
Compounds = 67
Mixtures = 167
Data points = 3185

Hydrocarbons
Alcohols
Esters
Ethers
Aromatic compounds and their derivatives



I. Oprisiu, et al. *Molecular Informatics,* **2012.** 31(6-7): p. 491-502.

Validation strategies: Points out



Estimates accuracy of prediction of missing points on curves

Validation strategies: Mixtures out

Number of data points for each mixture

Compounds	1	2	3	4	5
1	0	12	0	32	0
2		0	0	0	14
3			0	13	18
4				0	0
5					0

Estimates accuracy of prediction of new combination of available compounds

67 compounds → 2211 possible mixtures

167 available binary mixtures → 2044 new combinations of 67 available compounds

Validation strategies: Compounds out

Number of data points for each mixture

Compounds	1	2	3	4	5
1	0	12	0	32	0
2		0	0	0	14
3			0	13	18
4				0	0
5					0

Estimates accuracy of prediction of mixtures of new compounds

Modeling of vapor-liquid equilibrium curves

Random Forest models – external cross-validation results

	Q^2	RMSE
Points out	0.98	3.2 K
Mixtures out	0.90	6.9 K
Compounds out	0.79	10.3 K

Prediction of external test set containing compounds unavailable in the training set

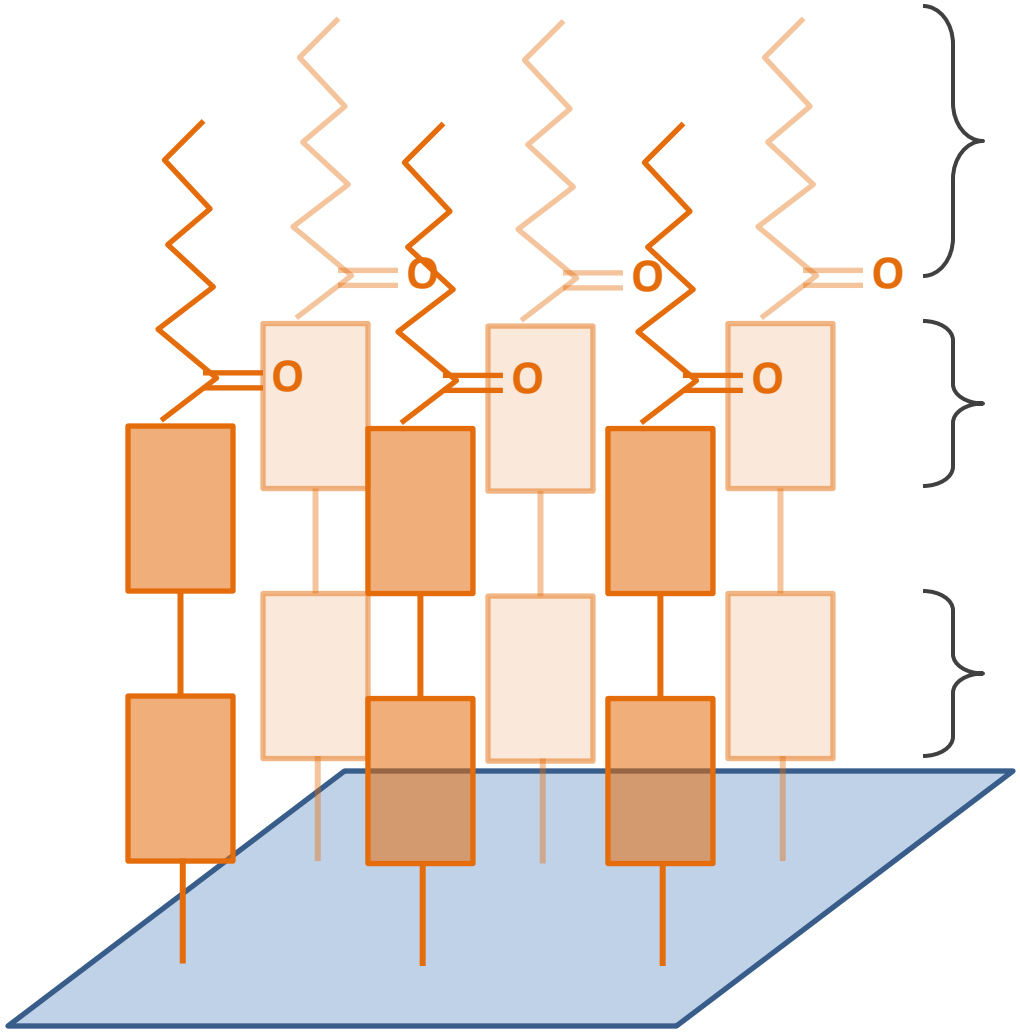
67 mixtures (1421 data points)

$R^2_{test} = 0.48, \quad RMSE_{test} = 18.5$ K

Task

Binding of compounds to artificial receptors

Modeling of compounds binding to artificial receptors



palmitic acid
ricinoleic acid
capric acid

Ala, Glu, His, Lys, Pro, Ser, Arg

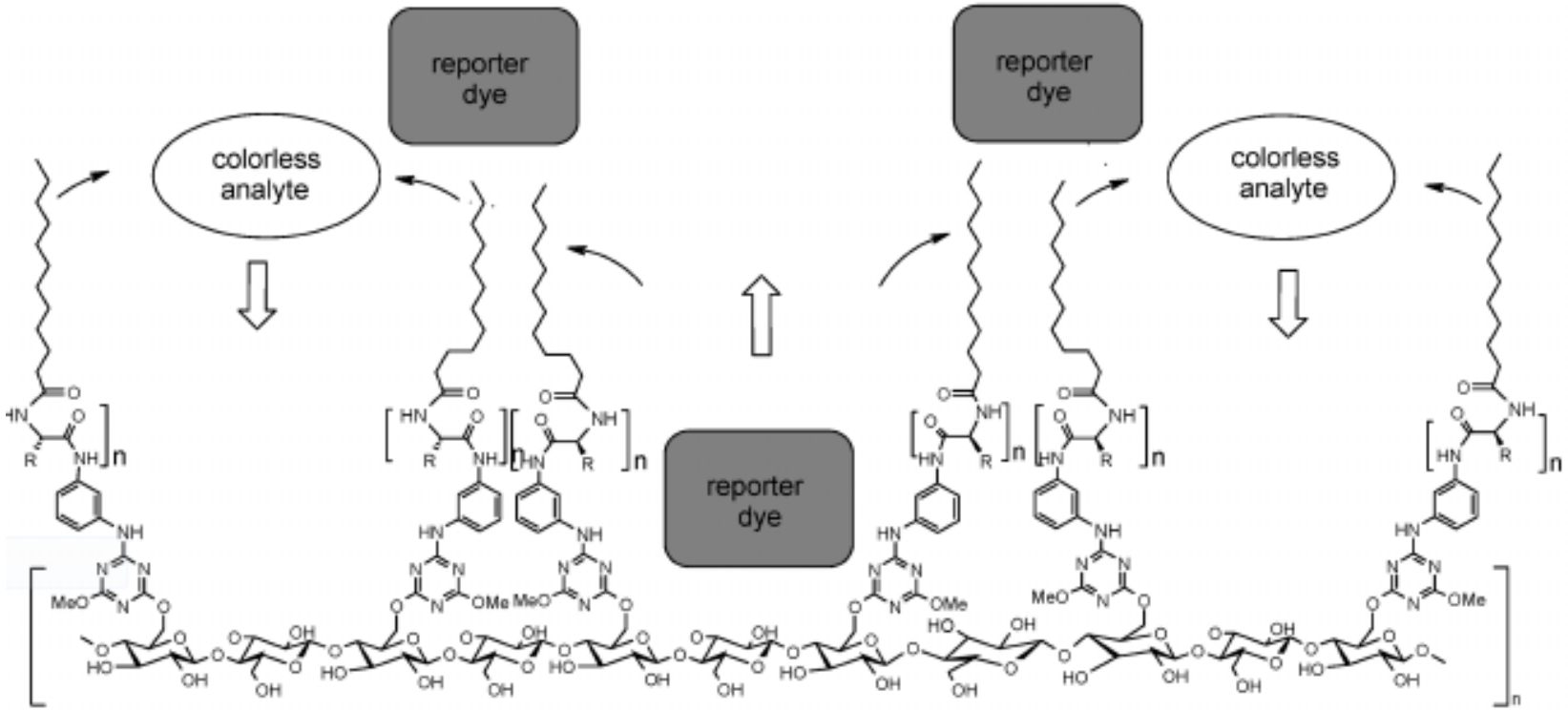
Ala, Glu, His, Lys, Pro, Ser, Trp

147 artificial receptors
x
33 ligands

4851 complexes

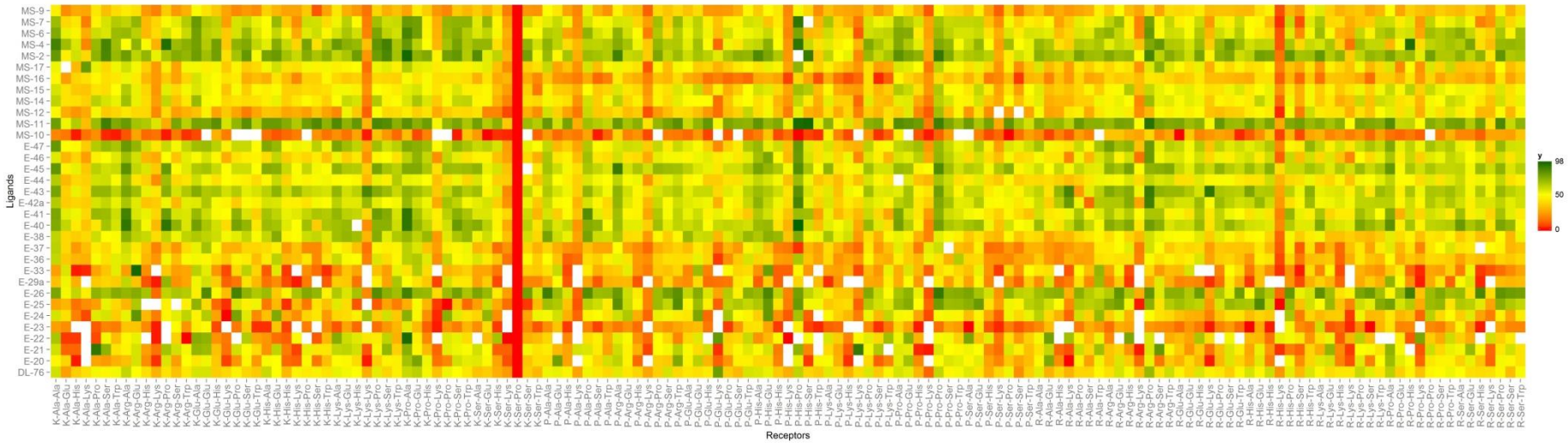
Bak A. et al. *Comb Chem High Throughput Screen*. **2014** Feb;17(2):141-156.

Modeling of compounds binding to artificial receptors



J. Fraczyk et al., *J. Comb. Chem.* **2009**, 11, 446–451

Modeling of compounds binding to artificial receptors



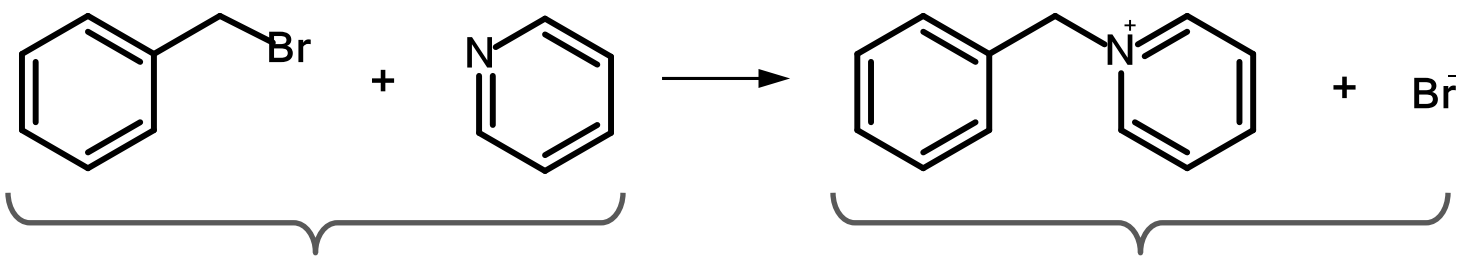
Results of LOO cross-validation
(Random Forest)

Method	Q^2	RMSE
Ligand out	0.47	16 %
Receptor out	0.64	13 %

Task

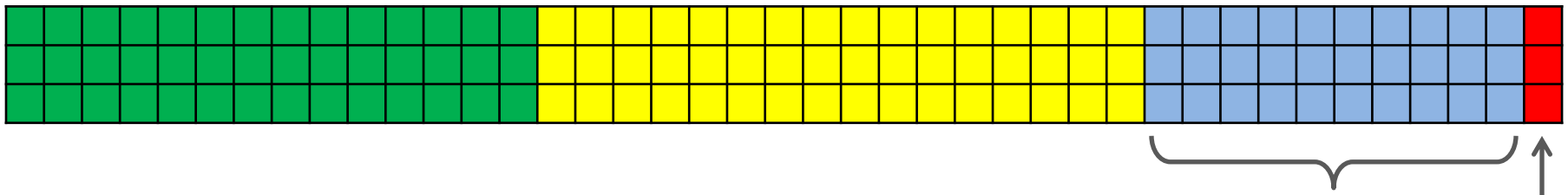
Rate constants of chemical reactions

Modeling of rate constants of chemical reactions



reactants mixture

products mixture



solvent parameters

temperature

Implicitly encoded reaction center.

No need of atom-to-atom mapping.

Modeling of rate constants of chemical reactions

Results of 5-fold cross-validation
(dataset: **1522 S_N2 reaction**, method Random Forest)

		Q^2	RMSE	Q^2_{AD}	$RMSE_{AD}$	AD coverage
Mixture	Reaction out	0.68	0.62	0.70	0.60	89%
	Product out	0.58	0.71	0.57	0.68	75%
CGR	Reaction out	0.70	0.60	0.72	0.58	88%
	Product out	0.58	0.71	0.56	0.68	76%

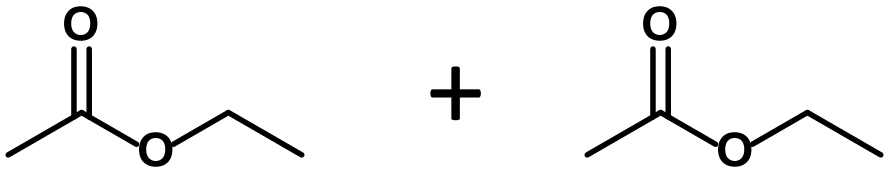
Results of 5-fold cross-validation
(dataset: **342 E2 reaction**, method Random Forest)

		Q^2	RMSE	Q^2_{AD}	$RMSE_{AD}$	AD coverage
Mixture	Reaction out	0.71	0.79	0.76	0.74	83%
	Product out	0.36	1.18	0.73	0.69	11%
CGR	Reaction out	0.72	0.79	0.73	0.77	93%
	Product out	0.40	1.14	0.40	0.98	40%

Task

Critical properties of organic compounds

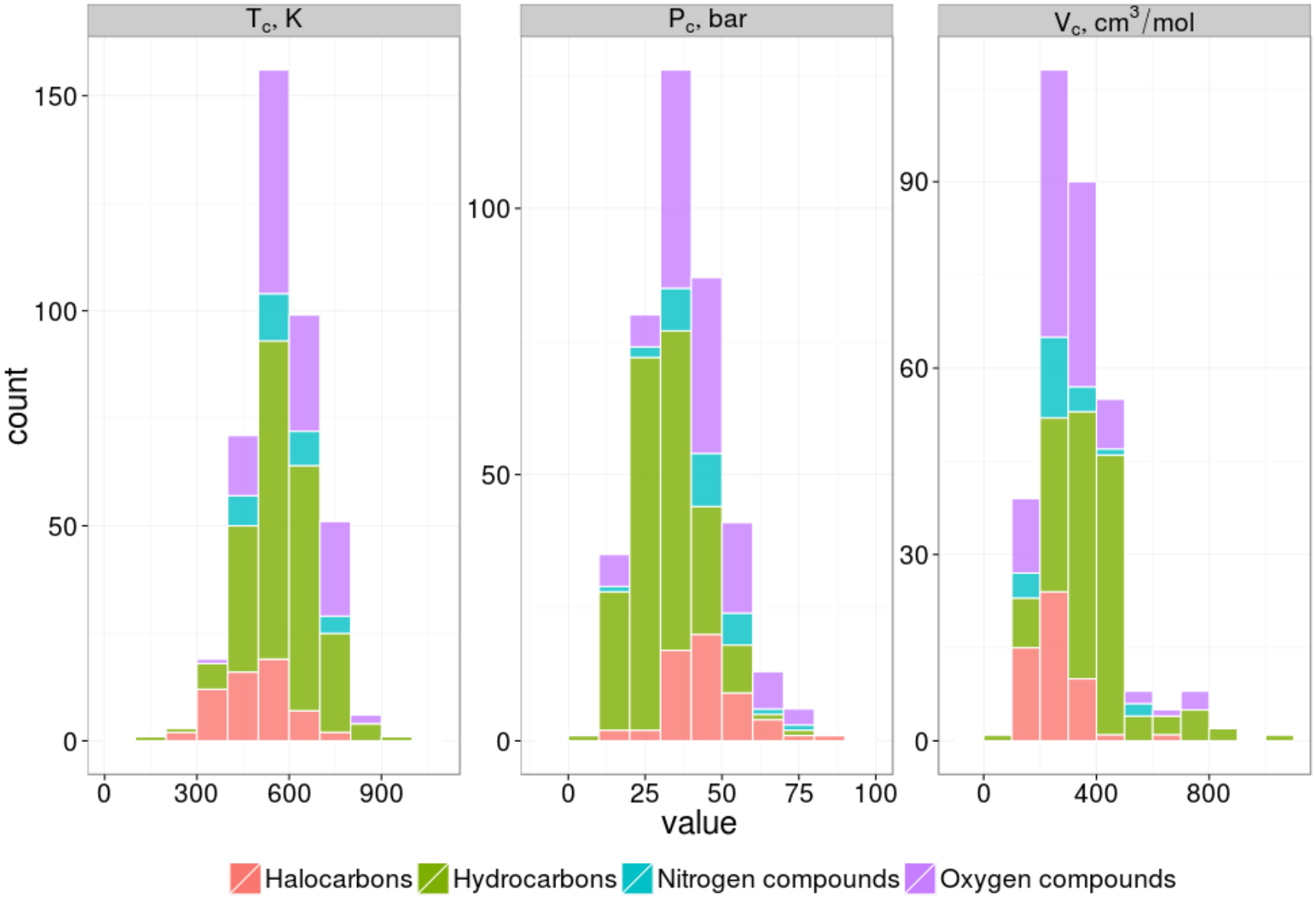
Quasi-mixture approach for modeling of critical properties



single molecule descriptors	quasi-mixture descriptors
C—C, H, H	C—O, C—O
C—O, C=O	C=O, C=O
O, C—C, O	O=C=O, O

Mokshyna E. et al. *Molecular Informatics*, **2014**, 33, 647-654

Quasi-mixture approach for modeling of critical properties

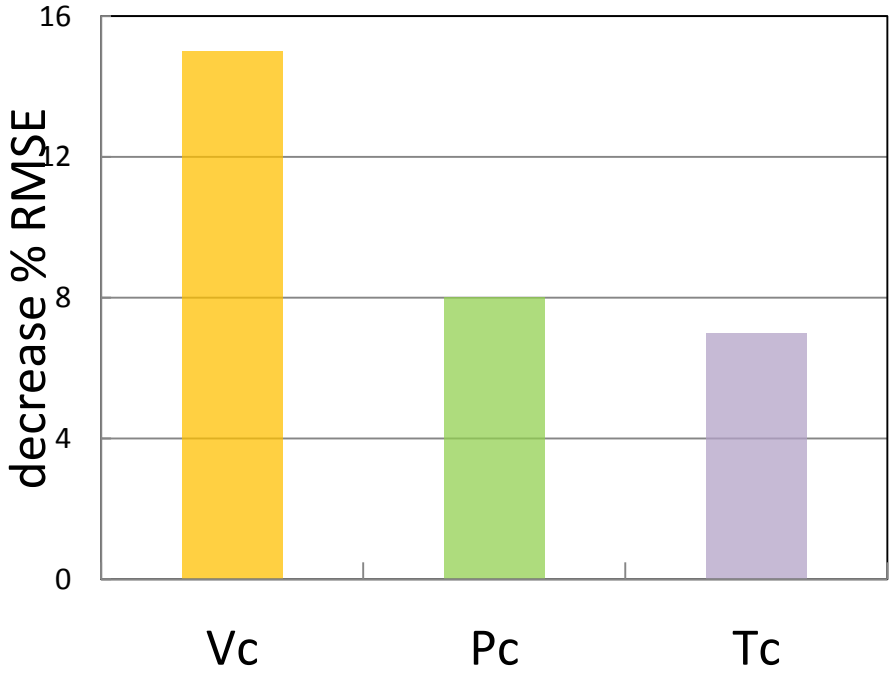


Mokshyna E. et al. *Molecular Informatics*, **2014**, 33, 647-654

Quasi-mixture approach for modeling of critical properties

Results of 5-fold cross-validation
(Random Forest)

		V_c	P_c	T_c
"Single molecule"	Q^2_{cv}	0.96 ± 0.01	0.88 ± 0.03	0.87 ± 0.04
	$RMSE_{cv}$	28.0 ± 0.10 (cm^3/mol)	4.45 ± 0.07 (bar)	40.1 ± 0.7 (K)
"Quasi-mixture"	Q^2_{cv}	0.96 ± 0.01	0.90 ± 0.02	0.89 ± 0.02
	$RMSE_{cv}$	24.0 ± 0.12 (cm^3/mol)	4.17 ± 0.05 (bar)	37.2 ± 0.7 (K)



Mokshyna E. et al. *Molecular Informatics*, **2014**, 33, 647-654

Acknowledgment

Strasbourg University (France)
Prof. A. Varnek

Bogatsky Physico-Chemical Institute (Ukraine)
Prof. V. Nedostup

Kazan Federal University (Russia)
Prof. I. Antipin
Dr. R. Nougmanov
A. Bodrov
A. Lin
M. Zeldi
T. Gimadiev
G. Sabirova

Lodz University of Technology (Poland)
Prof. Z. Kaminski

University of Silesia (Poland)
Prof. J. Polanski
Dr. A. Bak

Resources

Chemoinformatic group of A.V. Bogatsky Physico-Chemical Institute of NAS of Ukraine

http://qsar4u.com

Simplex representation of molecular structure (SiRMS)

https://github.com/DrrDom/sirms