



A.N. Frumkin

Institute of Physical Chemistry and Electrochemistry

Institution of Russian Academy of Sciences

31 Leninsky prospect, Moscow GSP-1, 119071 Russia

Substructural Molecular Fragments in Consensus QSPR Modeling

Vitaly Solov'ev

Laboratory of New Physico-Chemical Problems

Institute of Physical Chemistry and Electrochemistry, Moscow

E-mail: solovev-vp@mail.ru

INTRODUCTION

Substructural Molecular Fragments (SMF) as descriptors for QSPR modeling

Using of Molecular Fragments for ensemble modeling with combined applicability domain approach

Tools for compound design and property optimization on the basis of Fragments and their Contributions

QSPR modeling with fragment descriptors

PROGRAMS for Windows operating systems

ISIDA/QSPR program

Ensemble Multiple Linear Regression Analysis for QSPR modeling

Data manager EdiSDF

Editor of MDL Structure Data Files

2D sketcher EdChemS

Editor of Chemical Structures of MDL Molfiles

Collaborative Project ISIDA

The ISIDA/QSPR, EdiSDF and EdChemS programs The part of the ISIDA project

Laboratoire d'Infochimie, UMR 7177 CNRS, Université de Strasbourg, 4, rue
B.Pascal, Strasbourg, 67000, France
Prof. Alexandre Varnek

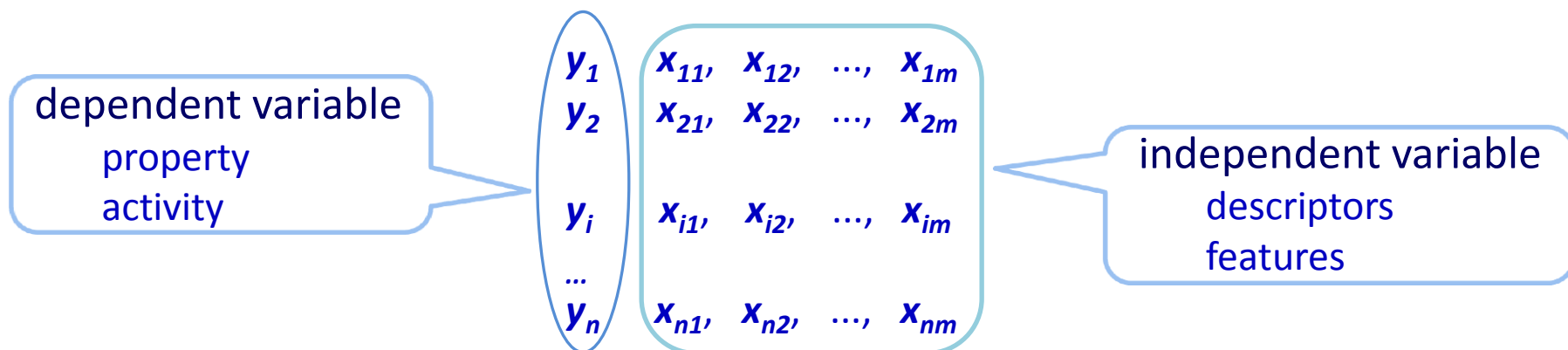
Laboratory of New Physico-Chemical Problems,
Institute of Physical Chemistry and Electrochemistry, Leninskiy prospect, 31a,
119991, Moscow, Russia
Acad. Aslan Tsivadze

<http://infochim.u-strasbg.fr/spip.php?rubrique53>
<http://vpsolovev.ru/programs/>

The ISIDA/QSPR Program

Substructural Molecular Fragments (SMF) as Descriptors

DATA:



Multiple Linear Regression model:

$$y = c_0 + c_1 \cdot x_1 + c_2 \cdot x_2 + \dots + c_m \cdot x_m$$

The ISIDA/QSPR Program

Counts of Substructural Molecular Fragments

id.	1	2	3	4	5	6	7	8	9
	C-N	C-C	C-O	N-C=O	C-C-O	C*C*C*C	C-C-C=C-C	C*C*C*C-C-C	O-C-C-O-C-N
Cmp 2	5	5	3	3	2	3	2	1	1
Cmp 3	5	4	3	3	2	6	0	1	1
Cmp 4	5	5	3	3	2	9	2	0	1
Cmp 5	5	4	3	3	2	6	1	0	1
Cmp 7	5	4	3	3	2	3	0	0	1
Cmp 8	5	3	5	3	2	3	0	0	1
Cmp 9	5	4	3	3	2	3	0	0	1
Cmp 10	5	5	3	3	2	3	0	0	1
Cmp 12	5	3	3	3	2	3	0	0	1
Cmp 13	6	3	3	3	2	3	0	0	1
Cmp 14	5	3	5	3	2	3	0	0	1
Cmp 15	5	5	3	3	2	3	0	0	1

independent variable
descriptors
features

Multiple Linear Regression model:

$$y = c_0 + c_1 \cdot x_1 + c_2 \cdot x_2 + \dots + c_m \cdot x_m$$

Descriptors for QSPR modeling

Two classes of Substructural Molecular Fragments

Molecular Fragments, Fragment Descriptors

sequences

topological paths
chains

augmented atoms

atoms with nearest neighbors
atoms with first shell

explicit representation of atoms and bonds?

atom sequences

$A_1 A_2 \dots A_n$

bond sequences

$B_1 B_2 \dots B_m$

atom/bond sequences

$A_1 B_1 A_2 B_2 \dots B_{n-1} A_n$

nearest atoms

$A(A_1; A_2; \dots A_k)$

nearest bonds

$A(B_1; B_2; \dots B_k)$

nearest atoms and bonds

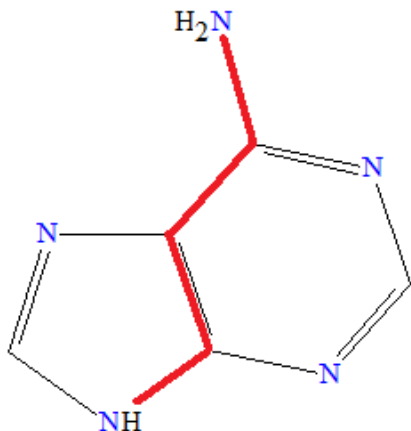
$A(B_1 A_1; B_2 A_2; \dots B_{k-1} A_k)$

Raevskii O.A., Sapegin A.M., Chistyakov V.V., Solov'ev V.P., Zefirov N.S. *Koord. Khim. (Rus)*, **1990**, *16*, 1175-1184

Solov'ev V. P., Varnek A. A. Wipff G. J. *Chem. Inf. Comput. Sci.*, **2000**, *40*, 847-858

Descriptors for QSPR modeling Substructural Molecular Fragments

sequences



atom sequence

N C C C N

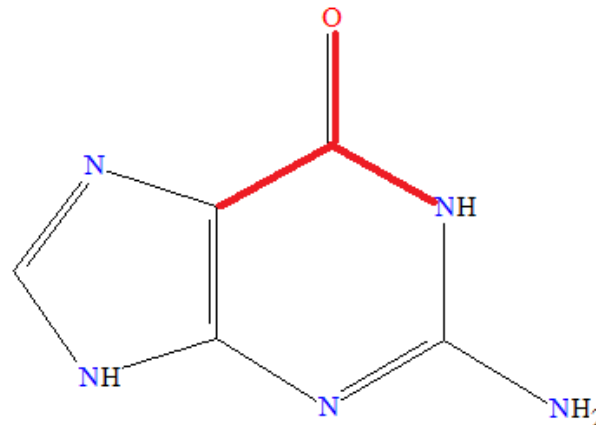
bond sequence

- = - -

atom/bond sequence

N - C = C - C - N

augmented atoms



nearest atoms

C(C; N; O)

nearest bonds

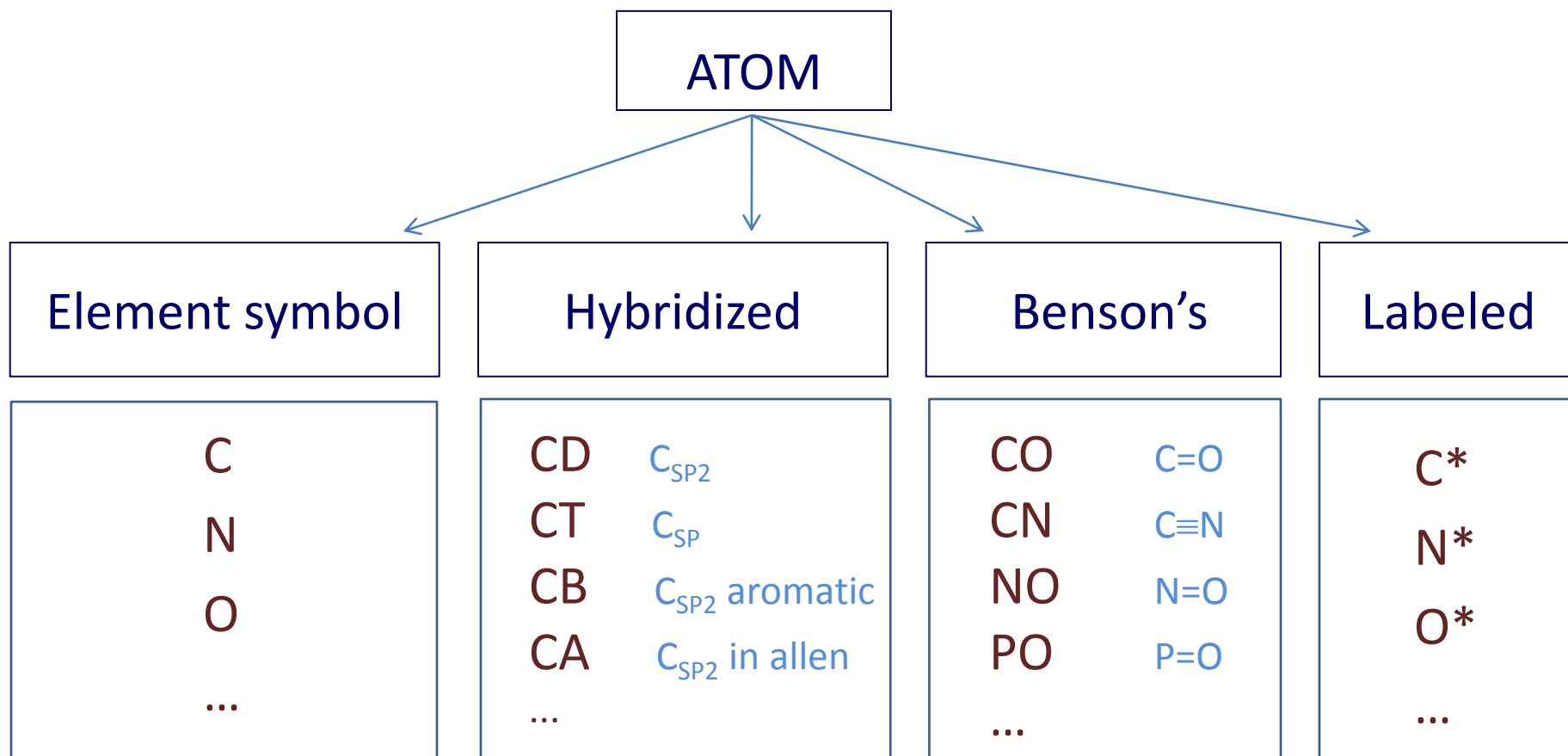
C(-; -; =)

nearest atoms and bonds

C(- C; - N; = O)

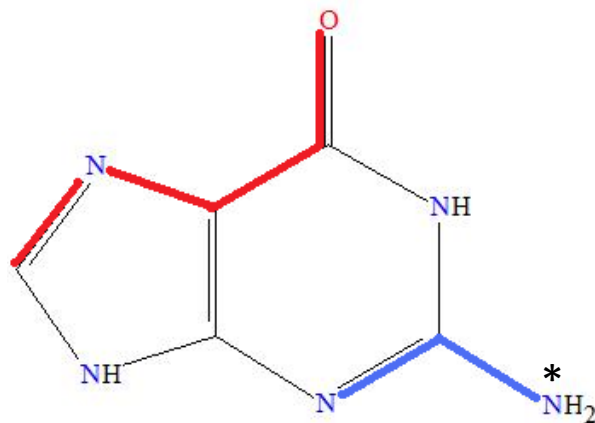
Substructural Molecular Fragments

Atomic attributes in molecular fragments



Substructural Molecular Fragments

Atomic attributes in molecular fragments



Element symbol

$C = N - C - C = O$

$N - C = N$

Hybridized

$CD = ND - CD - CD = OD$

$N - CD = ND$

Benson's

$CD - ND - C - CO$

$N - CD = ND$

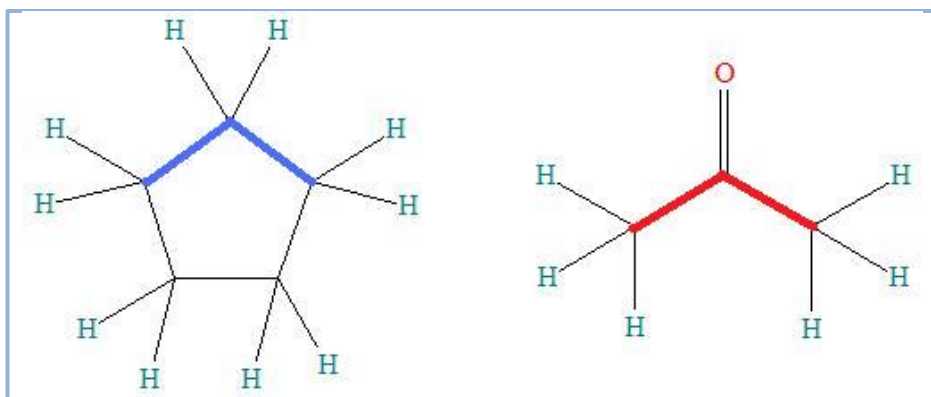
Labeled

$N^* - C = N$

Substructural Molecular Fragments

Labeled fragments and labeled atoms

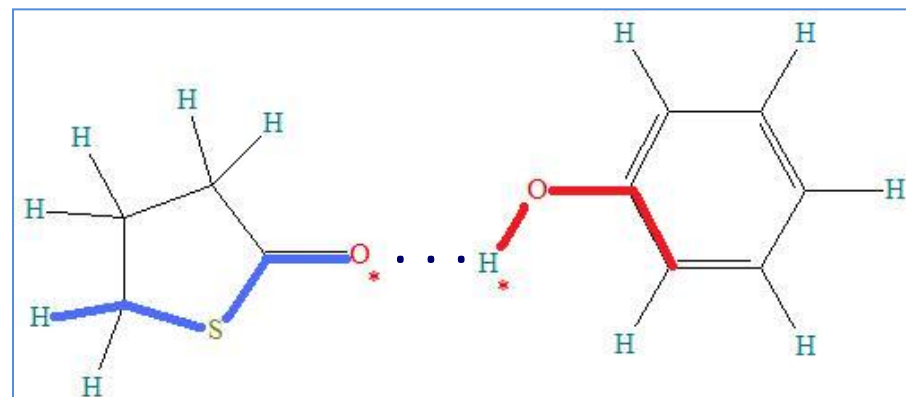
Labeled fragments of reagents



Reagent 1

Reagent 2

Labeled fragments and atoms of reagents



Reagent 1

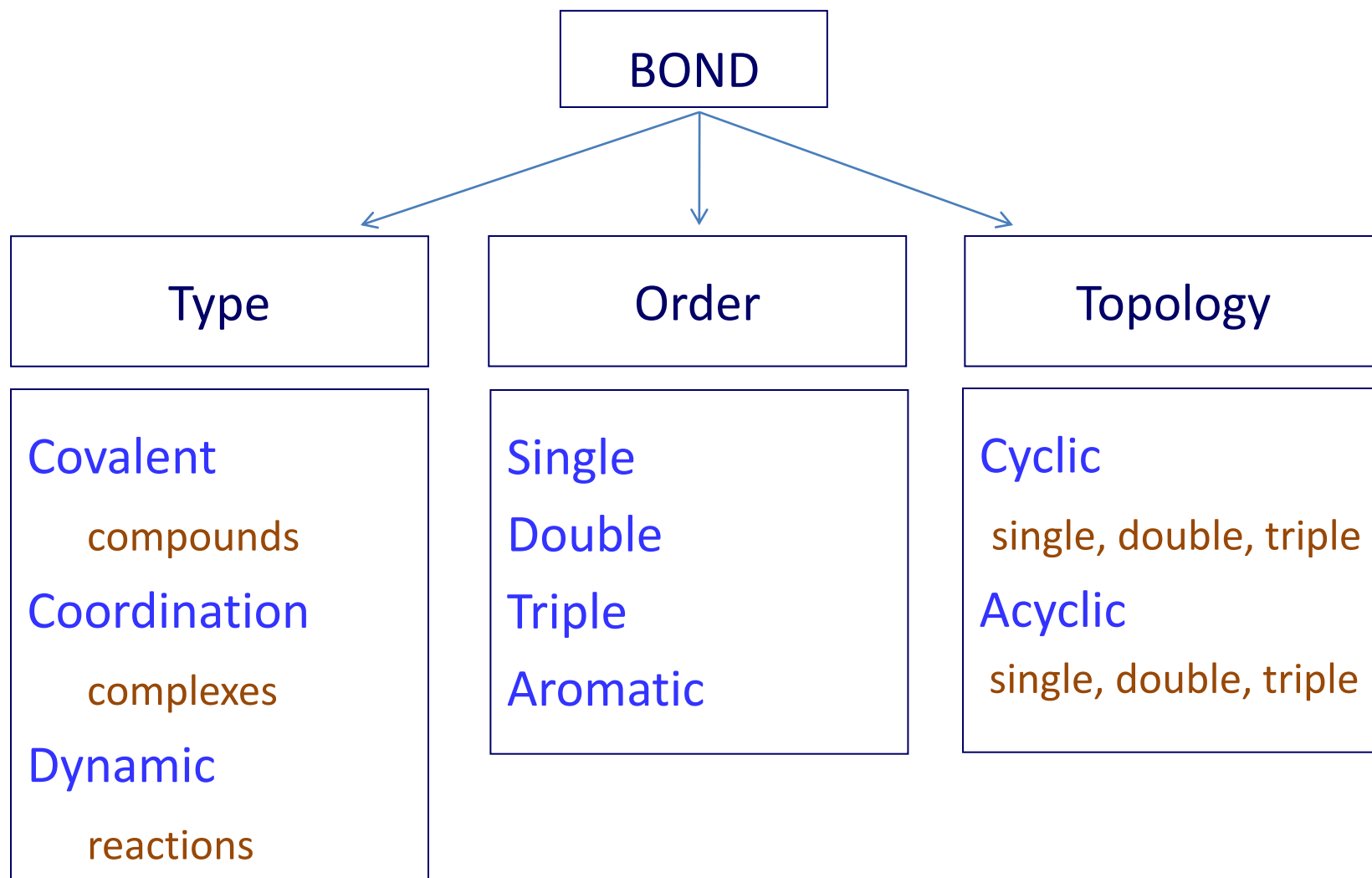
Reagent 2

Solov'ev V., Oprisiu I., Marcou G., Varnek A. *Ind. Eng. Chem. Res.*, **2011**, 50, pp. 14162–14167

Varnek A., Fourches D., Hoonakker F., Solov'ev V. P. *J. Comp.-Aided Mol. Design*, **2005**, 19, pp. 693-703

Substructural Molecular Fragments

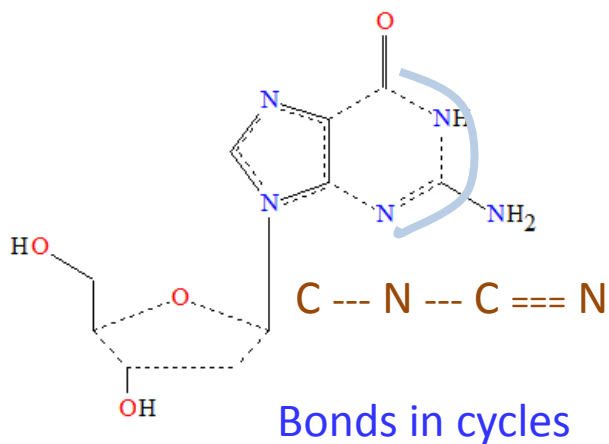
Bond attributes in molecular fragments



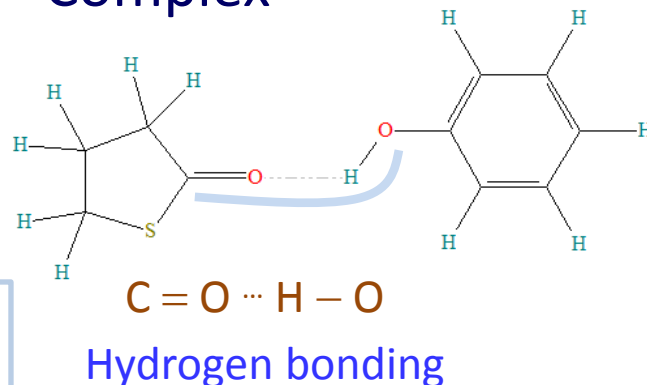
Substructural Molecular Fragments

Type, order and topology of bond in molecular fragments

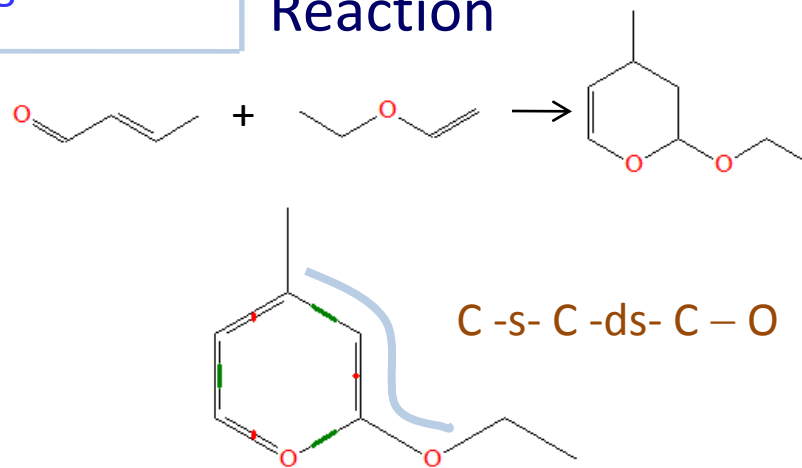
Compound



Complex



Reaction



Condensed Graphs of Reaction

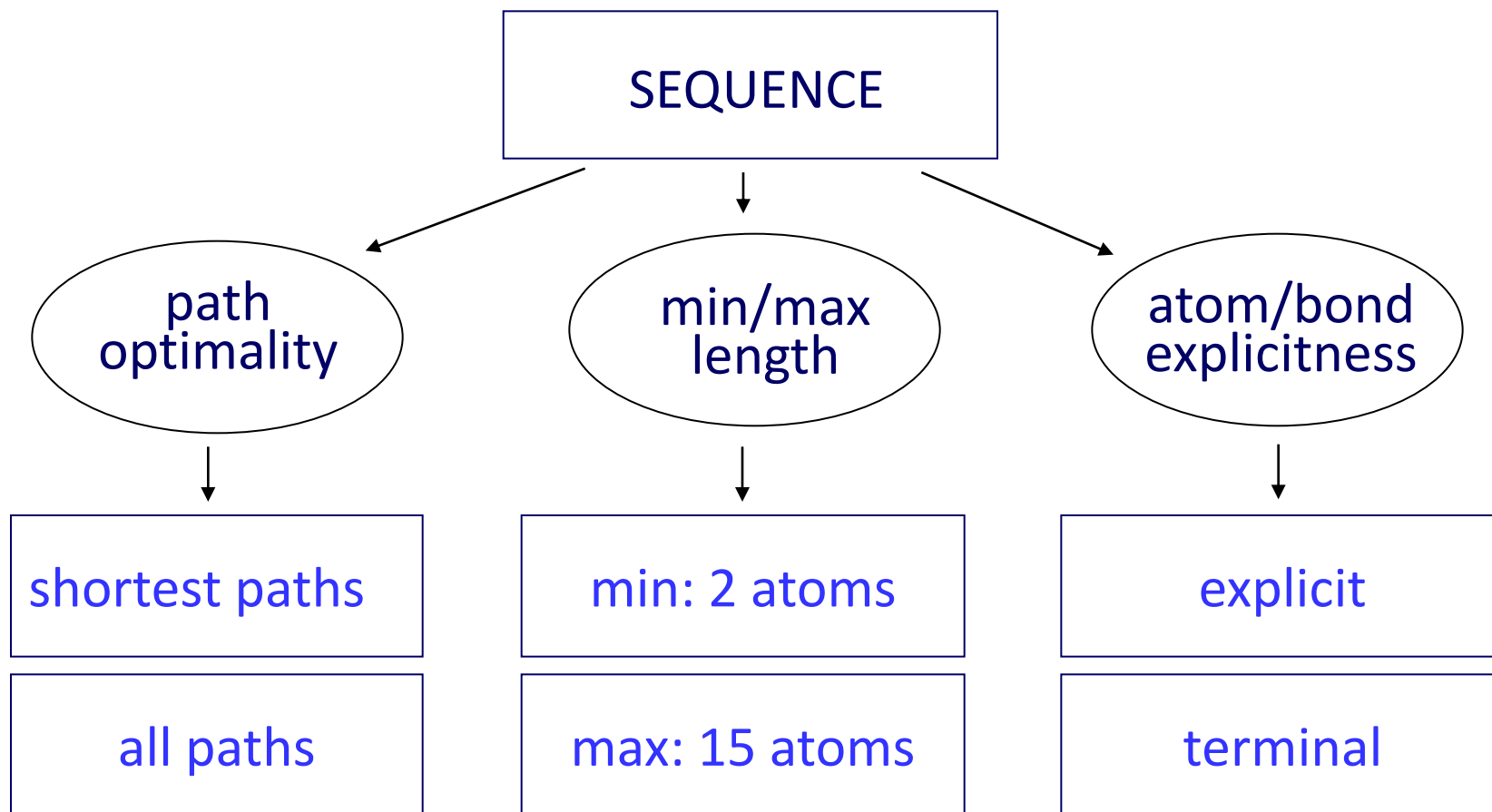
Solov'ev V. P., Tsivadze A. Yu., Varnek A. A. *Macroheterocycles*, **2012**, 5, pp. 404-410

De Luca A., Horvath D., Marcou G., Solov'ev V., Varnek A. *J. Chem. Inf. Model.*, **2012**, 52, pp. 2325-2338

EdChemS, EdiSDF, ISIDA QSPR

Substructural Molecular Fragments

Types of molecular sequences



Ensemble modeling by ISIDA/QSPR

The ISIDA/QSPR program can generate more than 12000 MLR models

- > 250 types of the SMF descriptors
- 5 forward variable selection algorithms
- 5 pools of preselected descriptors
- 2 types of linear equations



Selection by $Q^2 > Q^2_{lim} = 0.5 \dots 0.99$

The acceptable QSPR models

Varnek A. A., Wipff G., Solov'ev V. P., Solotnov A. F. *J. Chem. Inf. Comput. Sci.*, **2002**, 42, pp. 812-829

Solov'ev V. P., Varnek A. A. *Rus. Chem. Bull.*, **2004**, 53, pp. 1434-1445

Solov'ev V. P., Kireeva N. V., Tsivadze A. Y., Varnek A. A. *J. Struct. Chem.*, **2006**, 47, pp. 298-311

Ensemble modeling by ISIDA/QSPR

Recommended parameters of ISIDA/QSPR program

The ISIDA/QSPR program can generate more than 12000 MLR models

- > 250 particular types of the SM
- 5 variable selection techniques
- 5 pools of preselected descriptors
- 2 types of linear equations

The accepted

The screenshot shows the 'Batch Calculations' dialog box with the following parameters highlighted by red circles:

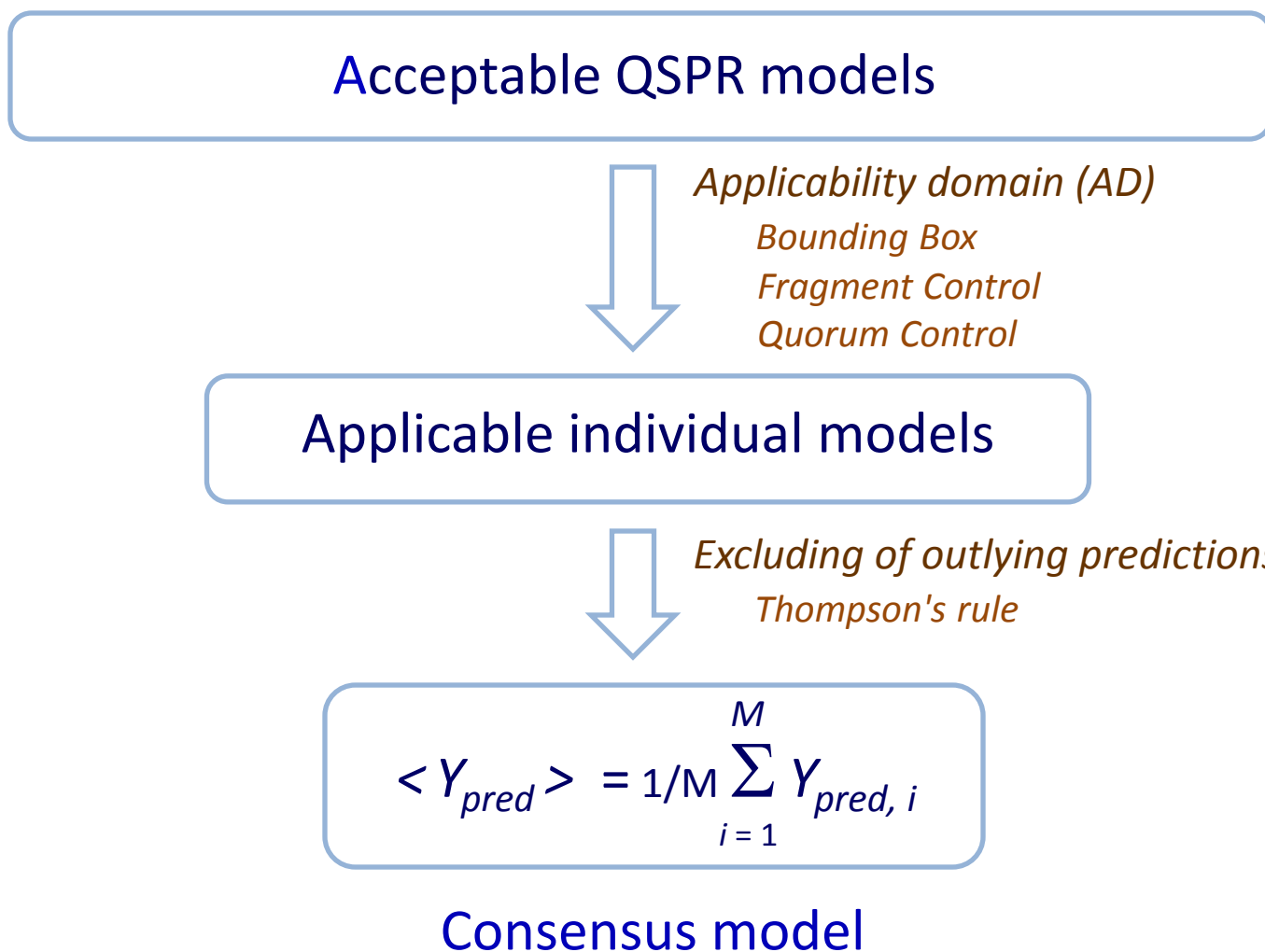
- Sequences: Fragments Length**: From: 2-4, To: 6-15
- Atom/bond sequences**: Atom/bond sequences
- Equation Type**: Y = SUM(Ai*Xi), Y = Ao + SUM(Ai*Xi)
- Terminal groups**: Terminal groups
- VSS methods**: VSS methods: 2,3; R_{yi} > 0.001, R_{ij} < 0.990

Other visible parameters include:

- Structure Data File**: Gd-L-H2O.SDF
- Modeling Property**: LogK1_25C_01M
- Mask File**: Gd-L-H2O.MSK
- Output Directory**: D:\SIDA_QSPR\
- Calc. Q2 for each**: 1 th point, fast Q2
- Appl. t-TEST**: 5 -fold C.-V.
- Do validation**: Do validation
- Averaging: Q2 >=**: 0.5
- Rcoef <**: 0.9
- R2abs >**: 0.50
- R2-Q2 <**: 0.50
- EPS:**: 1.0E-12
- t-Test:**: 1.96
- N-parameter eq.:**: 0
- Frag. count min:**: 2
- Cmp. count min:**: 1 time(s)
- Glob. mean, Conf.:**: 0.95

Consensus model for predicted compound

Model applicability domain and excluding of outlying predictions



Substructural Molecular Fragments

applicable descriptors for modeling of different properties

Physical properties

Melting point temperature: organic compounds, ionic liquids

Boiling point temperature: organic compounds, binary azeotropes

Chemical properties

Stability constants: metal-ligand and H-bonding complexation

Solvent extraction: separation factor, distribution coefficient

Similarity search and classification of chemical reactions

Biological properties

Anti-Human immunodeficiency virus activity

Blood/air, tissue/air partition coefficients

Skin permeation rate

Blood/brain penetration

Varnek A., Solov'ev V. Rev. in Book: *Ion Exchange and Solvent Extraction, A Series of Advances*, **2009**, 19, pp. 319-358

Katritzky A., Dobchev D., Fara D., Hur E., Tamm K., Kurunczi L., Karelson M., Varnek A., Solov'ev V. *J. Med. Chem.*, **2006**, 49, p. 3305

De Luca A., Horvath D., Marcou G., Solov'ev V., Varnek A. *J. Chem. Inf. Model.*, **2012**, 52, pp. 2325–2338

Prediction of stability constants

Complexation of organic ligands with metal ions



H	39 studied metals 2090 organic ligands 6749 logK values																He	
Li	Be											B	C	N	O	F	Ne	
Na	Mg											Al	Si	P	S	Cl	Ar	
K	Ca	Sc	Ti	V	Cr	Mn	Fe	Co	Ni	Cu	Zn	Ga	Ge	As	Se	Br	Kr	
Rb	Sr	Y	Zr	Nb	Mo	Tc	Ru	Rh	Pd	Ag	Cd	In	Sn	Sb	Te	I	Xe	
Cs	Ba	La	Hf	Ta	W	Re	Os	Ir	Pt	Au	Hg	Tl	Pb	Bi	Po	At	Rn	
Fr	Ra	Ac																
			Ce	Pr	Nd	Pm	Sm	Eu	Gd	Tb	Dy	Ho	Er	Tm	Yb	Lu		
			Th	Pa	U	Np	Pu	Am	Cm	Bk	Cf	Es	Fm	Md	No	Lr		

Solov'ev V. P., Tsvadze A. Yu., Varnek A. A. *Macroheterocycles*, **2012**, 5, No. 4-5, pp 404-410

Solov'ev V., Marcou G., Tsvadze A. Yu., Varnek A. *Ind. Eng. Chem. Res.*, **2012**, 51, pp 13482-13489

Application of developed QSPR models

Tools for design of new compounds

Property predictors

Generator of virtual combinatorial libraries

Interactive designer of compounds

Application of developed QSPR Models

Tools for design of new compounds

ISIDA predictor

Generator of virtual combinatorial libraries

Interactive designer of compounds

ISIDA Predictor

ISIDA Predictor

Laboratoire Infochimie UMR 7177, ULP, STRASBOURG

ISIDA Predictor

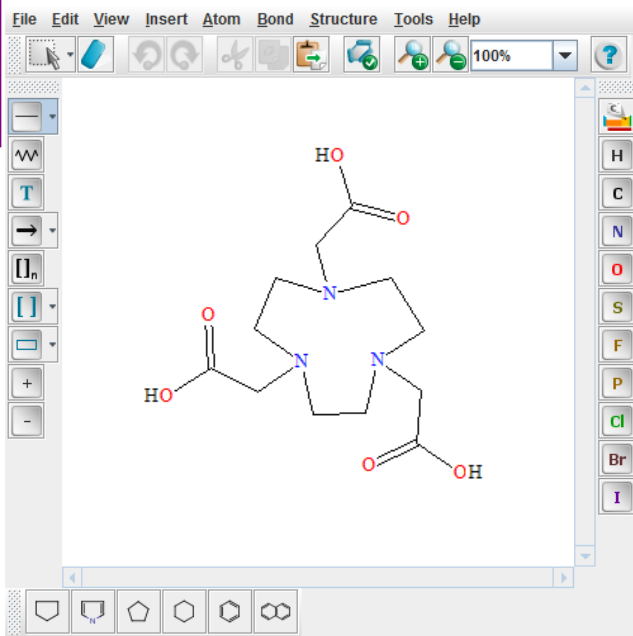
ISIDA Predictor is based on [ISIDA project](#)

Select a general kind of property : CoMet

Select a property to model : Zn

Sketcher provided by Chemaxon

Draw a molecule



CoMet Predictor

Prediction of stability constants $\log K$

metal - ligand complexes

27 metal ions

- Consensus Model predictions
simultaneous application of hundreds models
- Combined applicability domain approach
- Only INTERNET browser is required

<http://infochim.u-strasbg.fr/cgi-bin/predictor.cgi>

Varnek A., Fourches D., Kireeva N., Klimchuk O., Marcou G., Tsvadze A., Solov'ev V. *Radiochim. Acta*, **2008**, 96, 505-511 .

Application of developed QSPR Models

Tools for design of new compounds

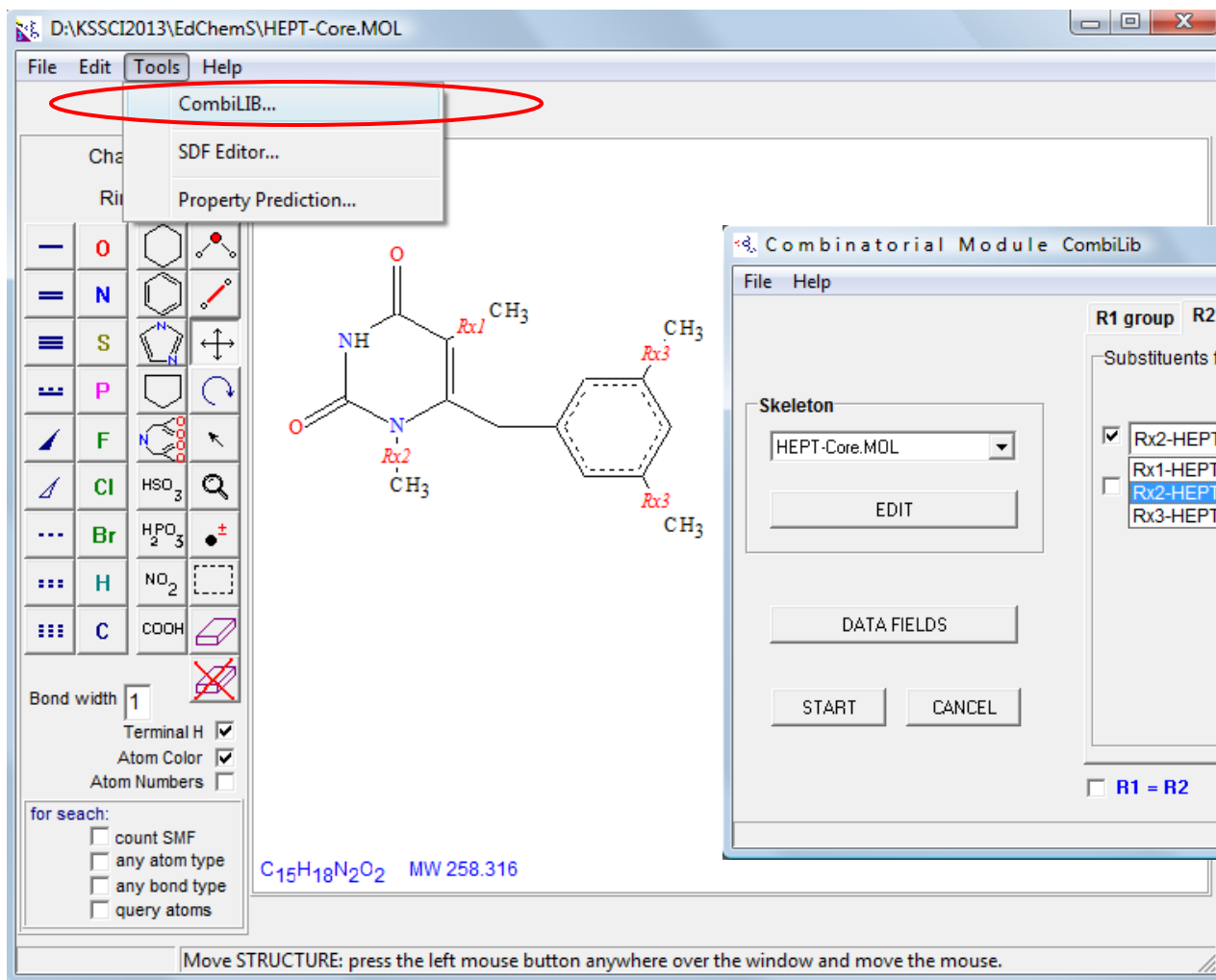
ISIDA predictor available via Internet

Generator of virtual combinatorial libraries

Interactive designer of compounds

Chemical editor EdChemS

Tools for generation of virtual combinatorial libraries



The 'Combinatorial Module CombiLib' dialog box is shown. It has a 'File' menu and a 'Help' button. The 'R1 group' tab is selected. The 'Skeleton' dropdown is set to 'HEPT-Core.MOL'. The 'Substituents for selected R2 group' table is as follows:

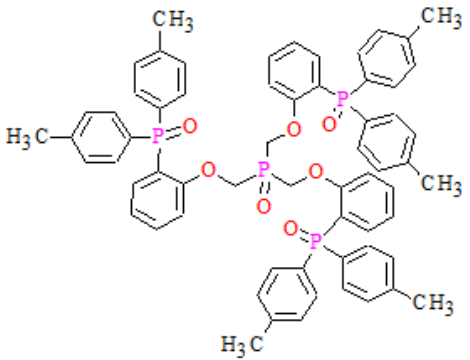
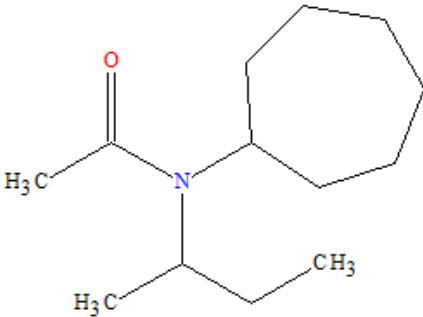
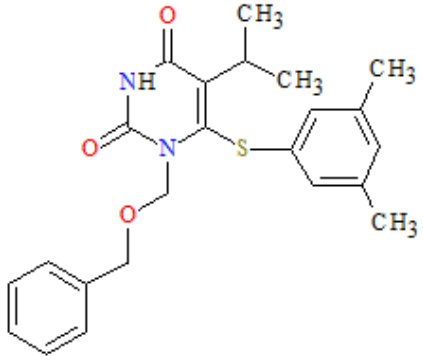
	NSubsts	Number of C atoms		Edit Substituents Set
		min	max	
<input checked="" type="checkbox"/> Rx2-HEPT	10	0	6	Edit set
<input type="checkbox"/> Rx1-HEPT	10	0	6	Edit set
<input type="checkbox"/> Rx2-HEPT	10	0	6	Edit set
<input type="checkbox"/> Rx3-HEPT				

At the bottom of the dialog, there are checkboxes for 'R1 = R2', 'R2 = R3', and 'R1 = R3', all of which are currently unchecked. Buttons for 'EDIT', 'DATA FIELDS', 'START', and 'CANCEL' are also present.

Varnek A.; Solov'ev V. P. *Combinat. Chem. High Throughput Screening*, **2005**, *8*, 403-416

Virtual combinatorial libraries

Extractants of uranyl ion and anti-HIV compounds

Activity	Extractant of the uranyl ion		Anti-HIV
Number of generated compounds	2020	10460	2640
Hits			
Predicted activity	$\lg D_{pred}$ 2.03 ± 0.06 (35)	$\lg D_{pred}$ 1.09 ± 0.02 (397)	$\lg(1/EC_{50})_{pred}$ 9.8 ± 0.2 (285)
Experimentally tested activity	$\lg D_{exp}$ 1.81 ± 0.05	$\lg D_{exp}$ 1.19	

1. Varnek A.; Fourches D.; Solov'ev V. P.; Baulin V. E.; Turanov A. N.; Karandashev V. K.; Fara D.; Katritzky A. R. J. *Chem. Inf. Comp. Sci.*, **2004**, *44*, 1365-1382
2. Varnek A.; Fourches D.; Solov'ev V.; Klimchuk O.; Ouadi A.; Billard I. *Solv. Extr. Ion Exch.*, **2007**, *25*, 433-462
3. Varnek A.; Solov'ev V. P. *Combinat. Chem. High Throughput Screening*, **2005**, *8*, 403-416

Application of developed QSPR Models

Tools for design of new compounds

ISIDA predictor available via Internet

Generator of virtual combinatorial libraries

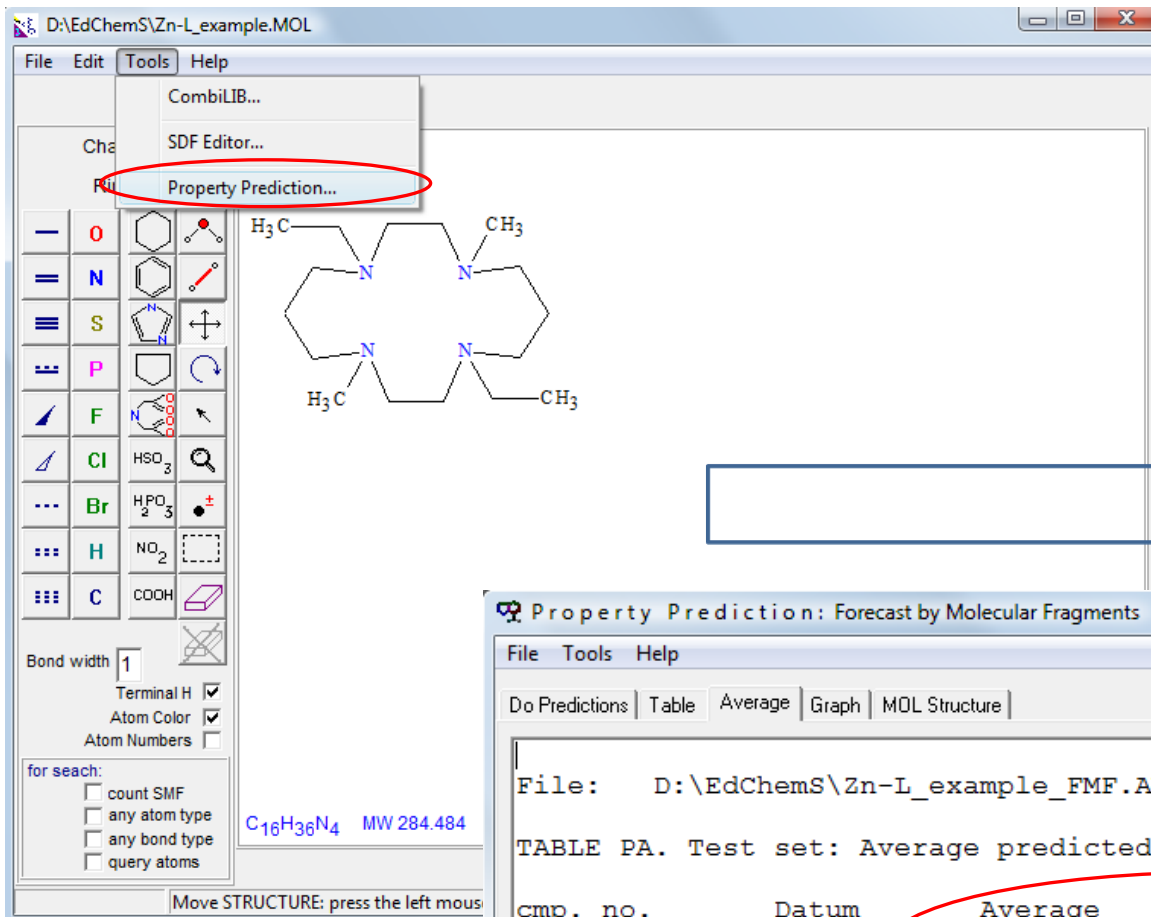
Interactive designer of compounds

by interaction of editor EdChemS with predictor FMF

Interactive design of compounds

Editor EdChemS + Predictor FMF

Drawing by EdChemS



Property Prediction by FMF

Property Prediction: Forecast by Molecular Fragments

File Tools Help

Do Predictions | Table | Average | Graph | MOL Structure

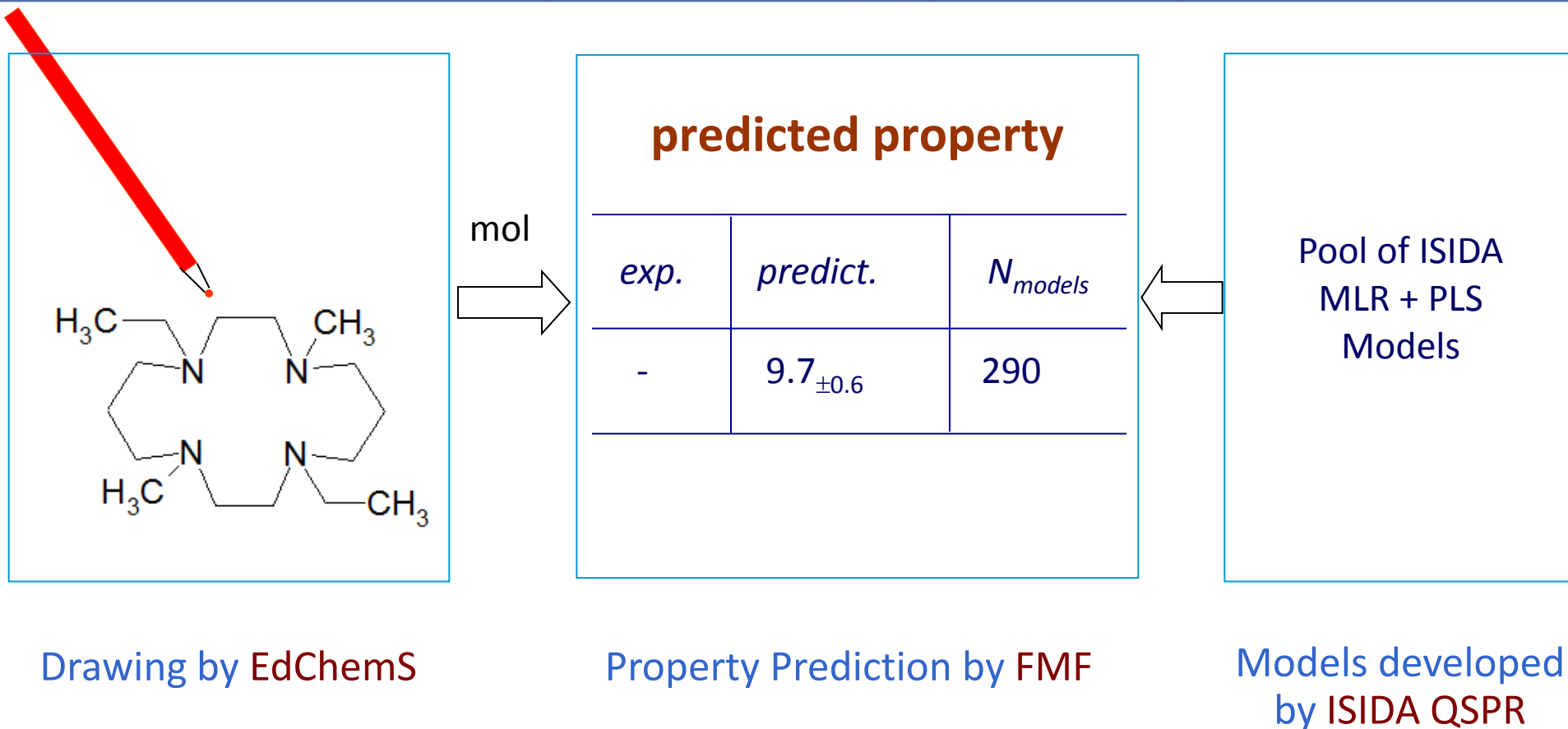
File: D:\EdChemS\Zn-L_example_FMF.AVE

TABLE PA. Test set: Average predicted property

cmp. no.	Datum	Average	STDEV	Nm	Dat.- Ave.
1	10.4	9.73E+00	5.5E-01	290	6.7E-01

Interactive design of compounds

Chemical editor EdChemS + Predictor FMF



Solov'ev V., Sukhno I., Buzko V., Polushin A., Marcou G., Tsivadze A., Varnek A.
J. Incl. Phenom. Macrocycl. Chem., **2012**, 72, 309–321

Interactive design of compounds

Tool for mean fragment contributions

EdiSDF File: Mn2_L_H2O.SDF; Current MOL: 6; All MOLs: 261

File Edit Tools Options Search Help

- MOL Editor
- Text Editor
- Run Linear Regression Data
- Fragment Contributions...**
- Paint Over Atom Contributions...
- View Classified Structures...
- CFR to SDF...
- Replace LF by CR+LF

$C_9H_{15}NO_6P_2$ MW 295.168

FIELD NAME	PROPERTY VALUE
LogK1_25C_01M	7.28
IonStr	0.2
TEMP	25°C
LogK1	6.96
EXP_NO	64650

1

Edit... Replace Mol Extract Mol Add Record Del. Record New SD File

Edit

Fragment Contributions of ISIDA Consensus Model

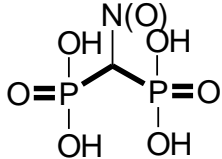
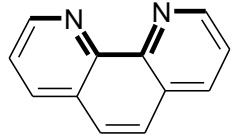
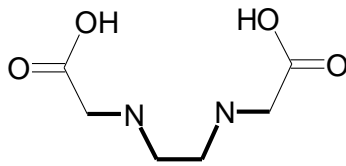
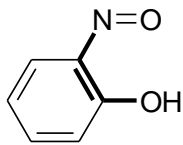
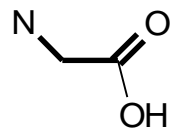
Options Individual Models Consensus Model

File: _MeanPSE_Mn2_L_H2O_261s1.TXT (merged 315 files)

C	DC	models	Frag_sumMols
1.235983	3.86E-001	78	C.C.N-C-C=O_16
0.556091	3.09E-001	3	C.C.N-C-C-O_20
-1.291552	6.16E-001	14	C.C.O.C.C.N.C.C.O.C_6
-0.247508	0.00E+000	1	C.C.O.C.C.N.C.C.O_6
0.481485	1.98E-001	16	C.C.O.C.C.N.C_23
0.924465	1.05E-001	10	C.C.O.C.C.N_23
1.107325	4.45E-001	34	C.C.O.C.C.N-C-C=O_8
0.846665	2.48E-001	16	C.C.O.C.C.O.C.C.N_7
1.843918	0.00E+000	1	C.C.O.C.C.O.C.C.N-C-C-O_3
0.625372	1.72E+000	3	C.C.O.C.C.O.C.C_14
0.856737	8.01E-001	2	C.C.O.C.C.O.C_16
0.381350	1.25E+000	10	C.C.O.C.C_24
-0.294877	6.45E-002	2	C.C.O_49
-0.799863	2.13E-002	2	C.C.S_11
2.332881	5.87E-001	12	C.C:C_14
-0.245603	0.00E+000	1	C.C_87
1.426841	3.45E-001	2	C.C=O_16
1.435773	7.17E-001	3	C.C-C=O_8
1.204567	1.95E-001	5	C.C-N_25
1.240225	0.00E+000	1	C.N.C.C+N+C.C.N.C_2

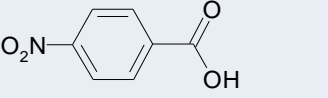
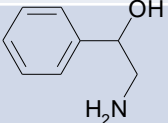
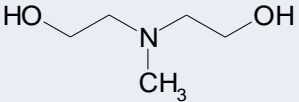
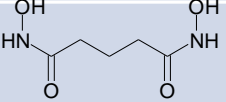
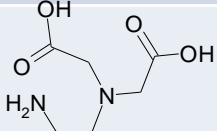
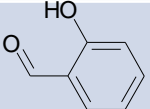
Mean SMF contributions: effective fragments

$$M + L = ML, \log K$$

M	SMF	$\langle a_i \rangle$	N_{model}	N_{mol}	Ligand
Mn ²⁺	O=P-C-P=O	3.74 ± 0.54	54	5	
Fe ²⁺	C _{ar} -N _{ar} -C _{ar} -C _{ar} -N _{ar} -C _{ar}	3.13 ± 0.28	37	6	
Y ³⁺	C-N-C-C-N-C	1.48 ± 0.10	25	5	
La ³⁺	N-C _{ar} -C _{ar} -O	3.39 ± 0.77	20	12	
UO ₂ ²⁺	N-C-C=O	4.63 ± 0.13	35	14	

From effective fragments to selective ligands



no.	ligand	$\log K_{pred}$					
		Mn ²⁺	Fe ²⁺	Y ³⁺	La ³⁺	Pb ²⁺	UO ₂ ²⁺
1		4.8	0.5	2.5	1.8	2.3	2.6
2		2.8	6.1	2.0	1.6	2.6	2.5
3		2.5	1.1	6.1	1.6	3.8	2.6
4		2.5	0.7	5.7	8.7	4.0	5.6
5		6.7	6.8	7.6	6.9	10.5	8.0
6		3.8	3.3	4.3	3.4	2.0	8.9

Application of developed QSPR Models

Tools for design of new compounds

ISIDA predictor available via Internet

Generator of virtual combinatorial libraries

Interactive designer of compounds

by coloring of atoms of chemical formula
according fragment contributions

Interactive design of compounds

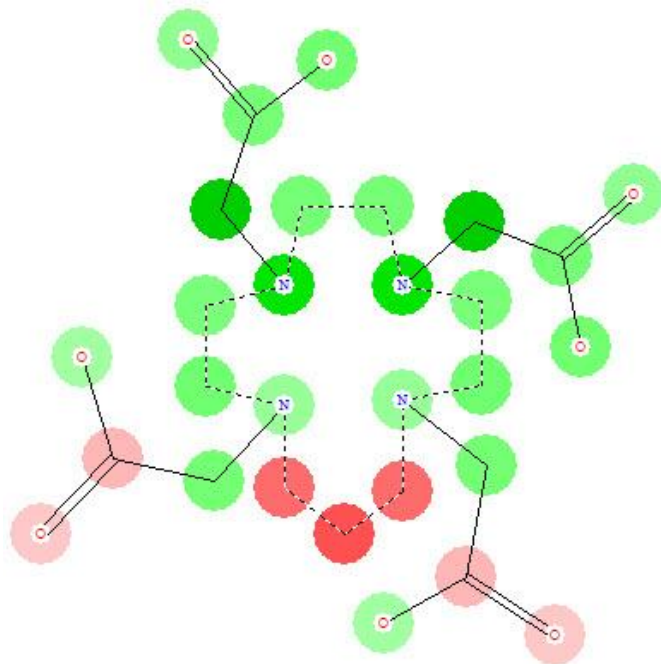
Tool for coloration of atoms of chemical formula

The image shows two overlapping windows from the EdiSDF software. The background window is the main application, displaying a chemical structure with atoms highlighted in green and red. The 'Tools' menu is open, and 'Paint Over Atom Contributions...' is highlighted with a red oval. Below the menu, the molecular formula $C_{17}H_{30}N_4O_8$ and molecular weight 418.446 are shown. A table of properties is visible, and a toolbar with navigation buttons is at the bottom.

FIELD NAME	PROPERTY VALUE
MOL_ID	967
Formula	C17H30N4O8
MolWeight	418.446
FULL_NAME	1,4,7,10-Tetraazacyclotridecane-1,4,7,10-tetraethanoic acid
SH_NAME	TRITA

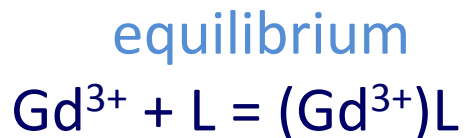
The foreground window is titled 'Atomic Topological Contributions'. It features a 'Do Predictions' section with 'Table' and 'Average' options. The folder path is 'D:\Gd-L-H2O\'. A list of 180 model files is shown, with 'Selected Models for Property Prediction' containing the same 180 models. The 'Predicted Property' section shows 'LogK1_25C_01M' with a value of 1.96402E+001. There are checkboxes for 'Atomic Mean' and 'Numeric Atomic Contributions', and a 'START' button.

Coloration of atoms according fragment contributions of consensus model



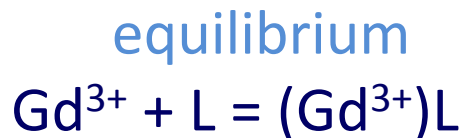
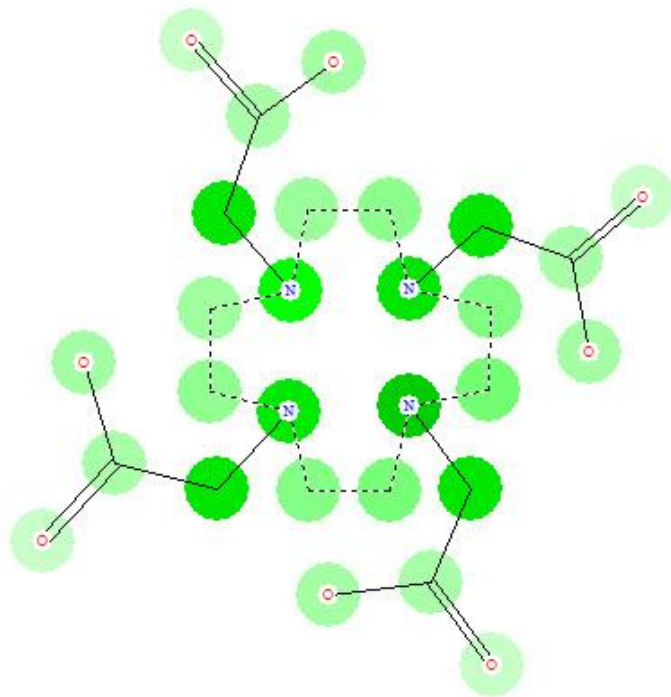
Color depth of atom

$$c_A = \sum_{j=1}^M \sum_{A \in F_j} a_{ij}$$



topological contributions of ligand atoms

Coloration of atoms according fragment contributions of consensus model



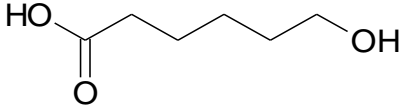
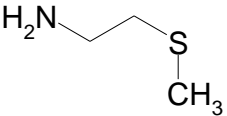
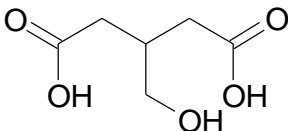
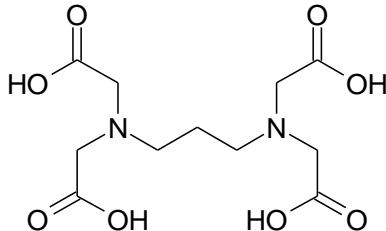
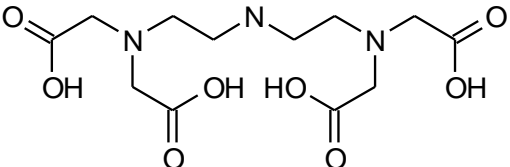
Color depth of atom

$$c_A = \sum_{j=1}^M \sum_{A \in F_j} a_{ij}$$

topological contributions of ligand atoms

Interactive design of compounds

The $\log K$ tuning of $(\text{Zn}^{2+})\text{L}$ complexes

no.	Ligand	$\log K_{pred}$	s	N_m
1		1.2	0.3	337
2		2.1	0.5	362
3		2.9	0.5	355
...
14		14.0	0.4	320
15		15.4	0.5	331

Solov'ev V., Sukhno I., Buzko V., Polushin A., Marcou G., Tsivadze A., Varnek A.
J. Incl. Phenom. Macrocycl. Chem., **2012**, 72, 309–321

CONCLUSIONS

- Substructural Molecular Fragments are applicable descriptors for QSPR modeling of physical, chemical and biological properties.
- Molecular Fragments are suitable for ensemble modeling and combined applicability domain approach ensuring good accuracy of predictions.
- Fragments and their Contributions are convenient building blocks for compound design and property optimization.

ACKNOWLEDGMENTS

Prof. Alexandre Varnek

Acad. Aslan Tsivadze

Dr. Timur Madzhidov

Dr. Gilles Marcou

Dr. Natalia Kireeva

**Organizers of Kazan Summer School on Chemoinformatics,
GDR PARIS, GDRE SupraChem, ARCUS, RFBR**