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:

dependent variable
- property
- activity

$$\begin{matrix} y_1 \\ y_2 \\ \\ y_i \\ \dots \\ y_n \end{matrix} \qquad \begin{matrix} x_{11}, & x_{12}, & \dots, & x_{1m} \\ x_{21}, & x_{22}, & \dots, & x_{2m} \\ \\ x_{i1}, & x_{i2}, & \dots, & x_{im} \\ \\ x_{n1}, & x_{n2}, & \dots, & x_{nm} \end{matrix}$$

independent variable
- descriptors
- features

Multiple Linear Regression model:

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

# 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:

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

# 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**

$A_1\ A_2 \ldots A_n$

**bond sequences**

$B_1\ B_2 \ldots B_m$

**atom/bond sequences**

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

**nearest atoms**

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

**nearest bonds**

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

**nearest atoms and bonds**

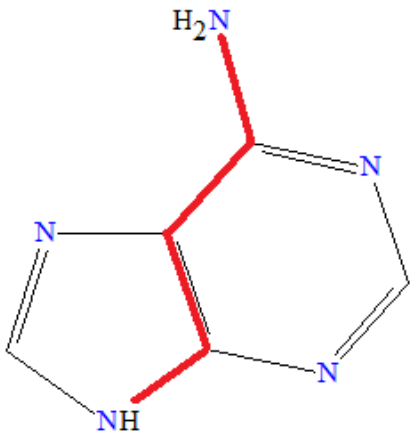
$A(B_1\ A_1; B_2\ A_2; \ldots B_{k-1}\ A_k)$

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

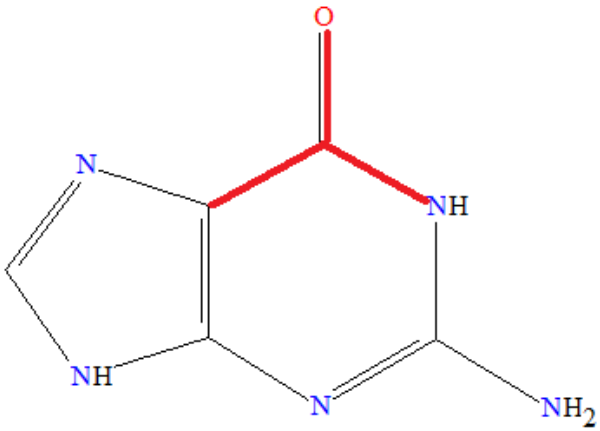
bond sequence

– = – –

atom/bond sequence

N – C = C – C – N

### augmented atoms




nearest atoms

C(C; N; O)

nearest bonds

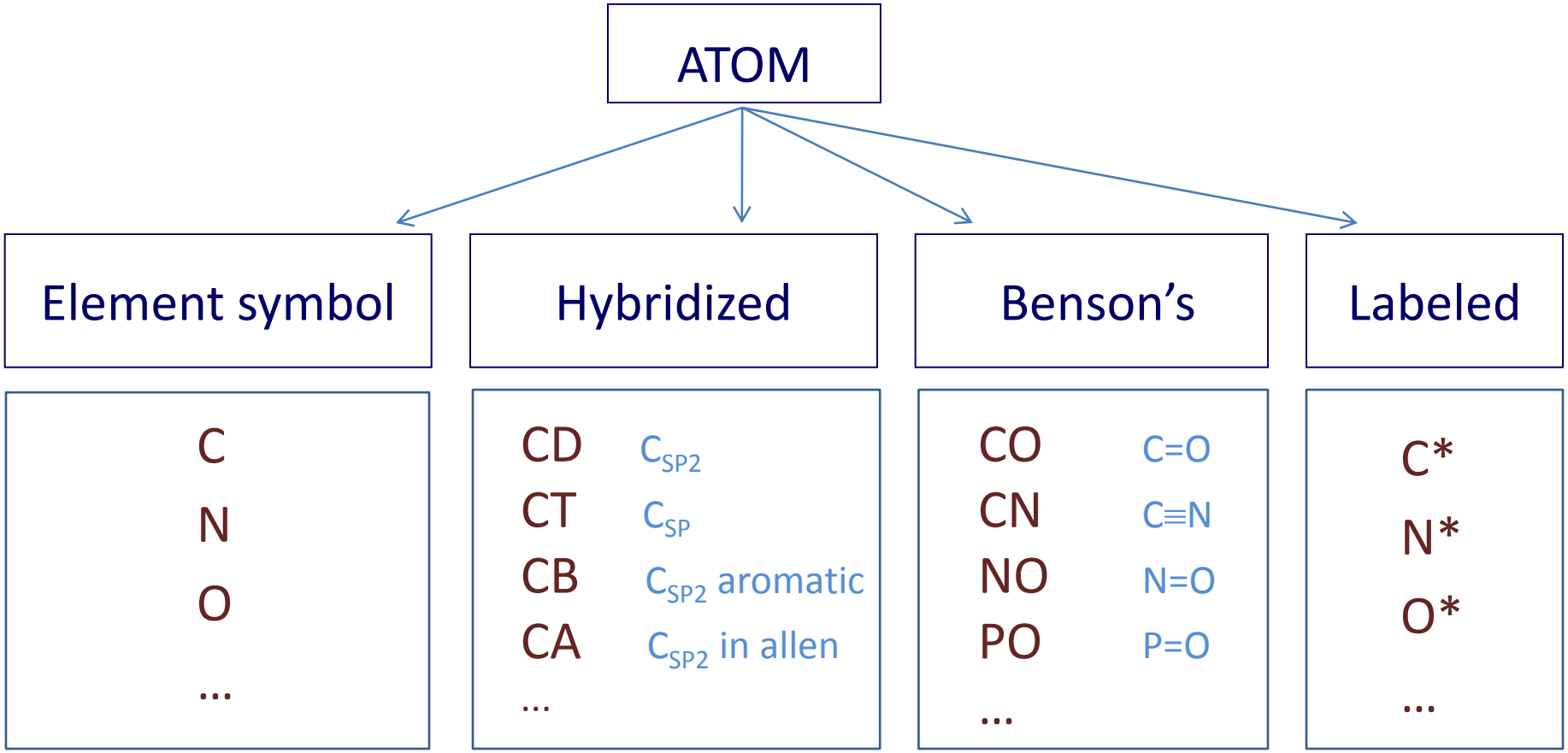
C(–; –; =)

nearest atoms and bonds

C(– C; – N; = O)

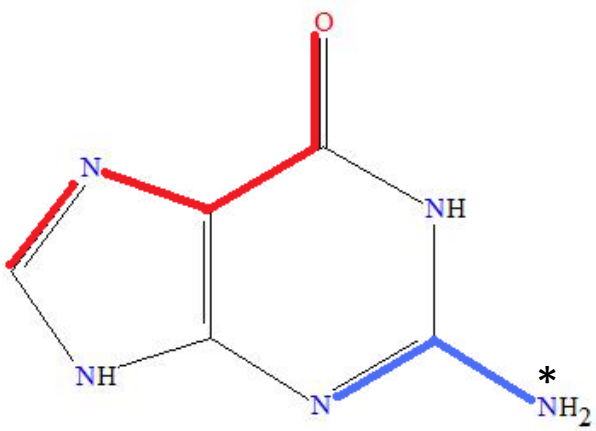
# Substructural Molecular Fragments

## Atomic attributes in molecular fragments



# Substructural Molecular Fragments

## Atomic attributes in molecular fragments

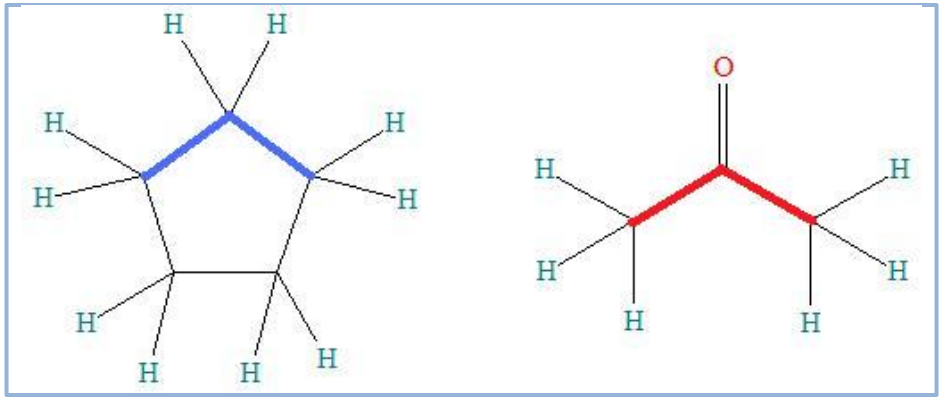



| Element symbol | Hybridized | Benson's | Labeled |
|---|---|---|---|
| C = N – C – C = O | CD = ND – CD – CD = OD | CD – ND – C – CO | N* – C = N |
| N – C = N | N – CD = ND | N – CD = ND | |

# Substructural Molecular Fragments

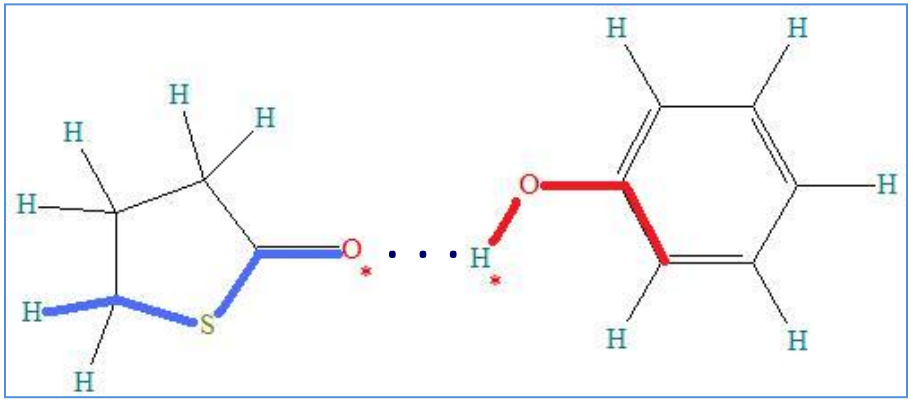
## Labeled fragments and labeled atoms

Labeled fragments of reagents




Reagent 1 Reagent 2

Labeled fragments and atoms of reagents



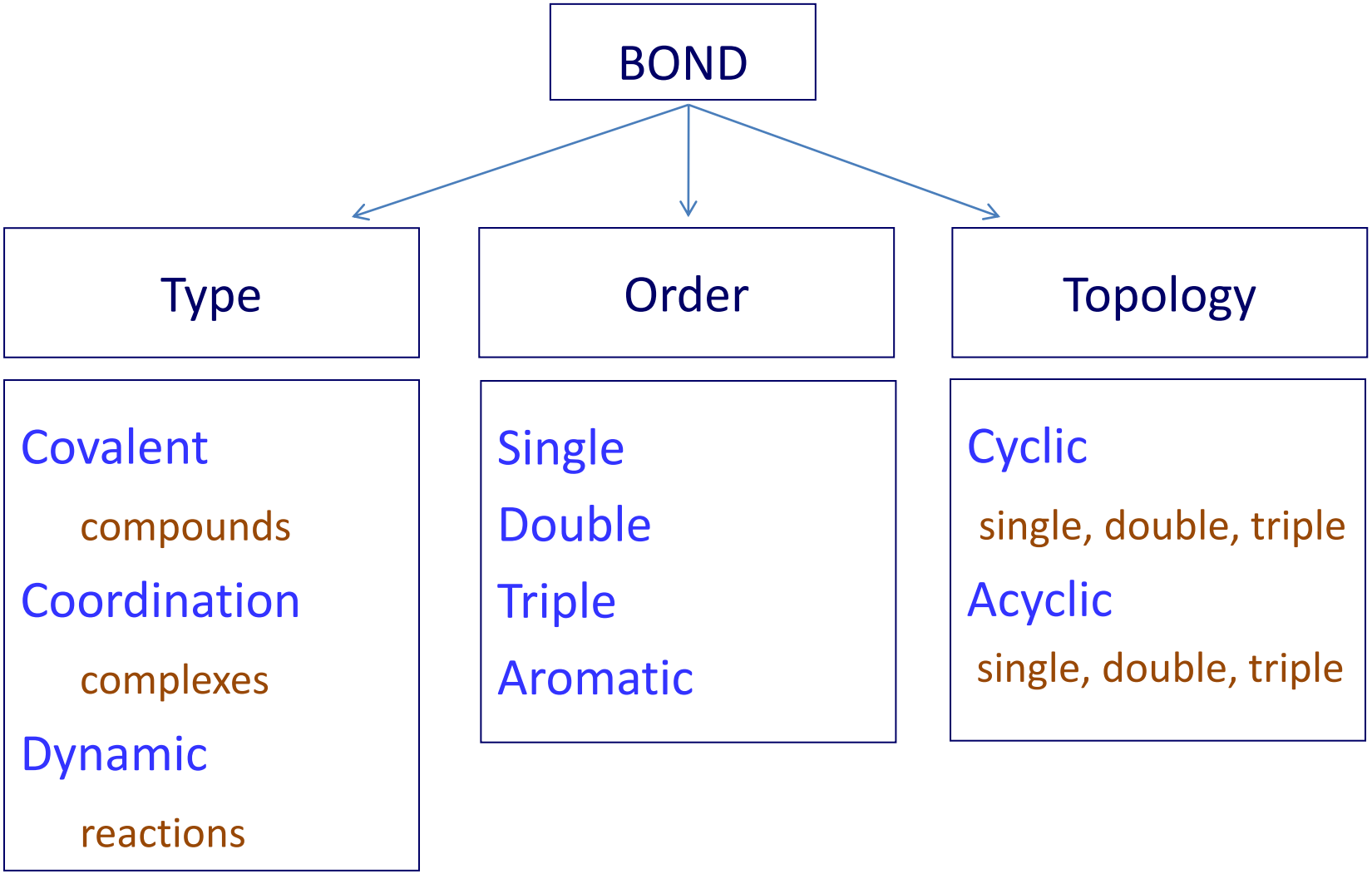

Reagent 1 Reagent 2


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

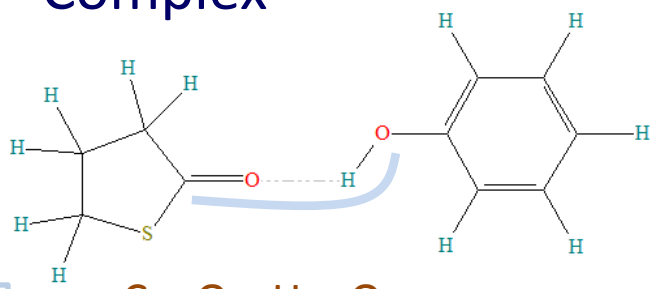
## Bond attributes in molecular fragments



# Substructural Molecular Fragments

## Type, order and topology of bond in molecular fragments

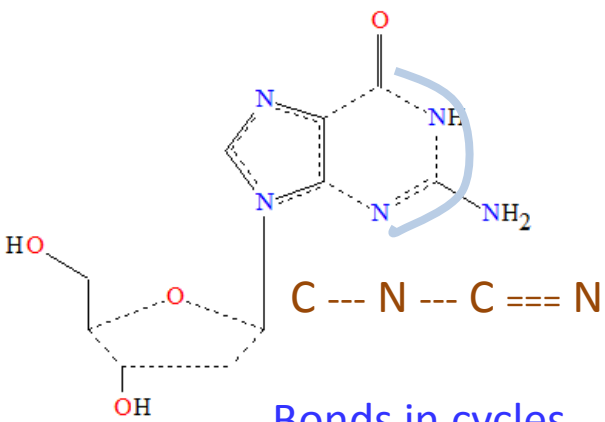
### Complex



C = O ··· H – O

Hydrogen bonding

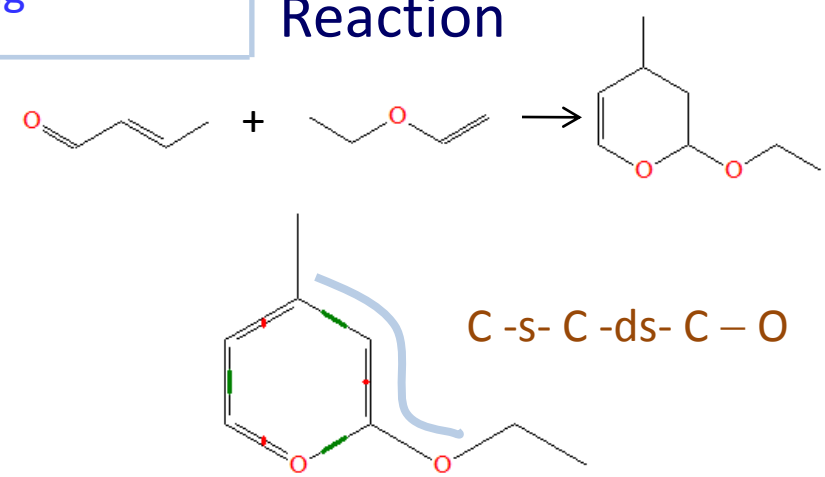
### Compound



C --- N --- C === N

Bonds in cycles

### Reaction



C -s- C -ds- C – O

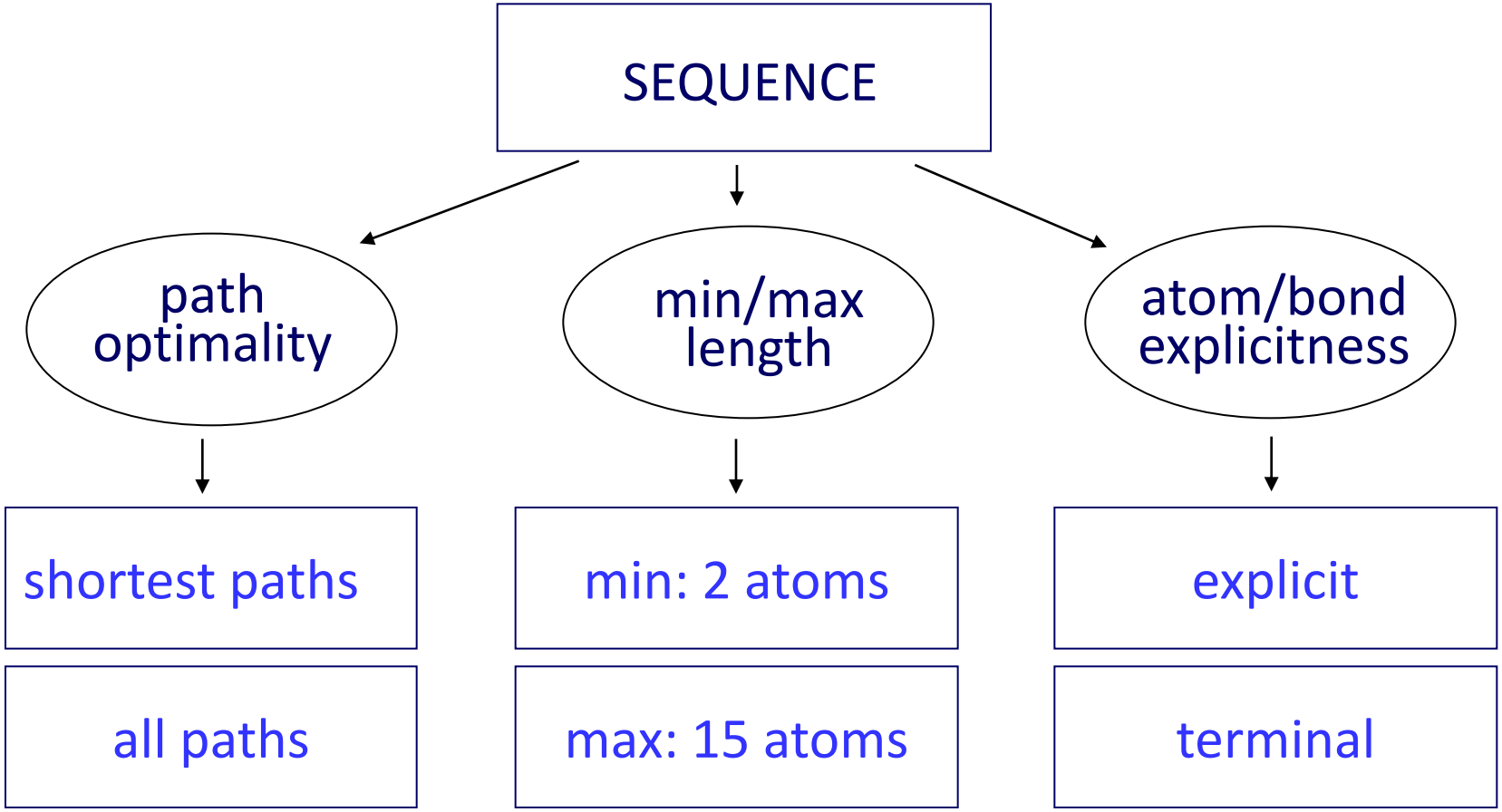
Condensed Graphs of Reaction


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

# 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* $Q^2 > Q^2_{lim} = 0.5 \ldots 0.99$

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

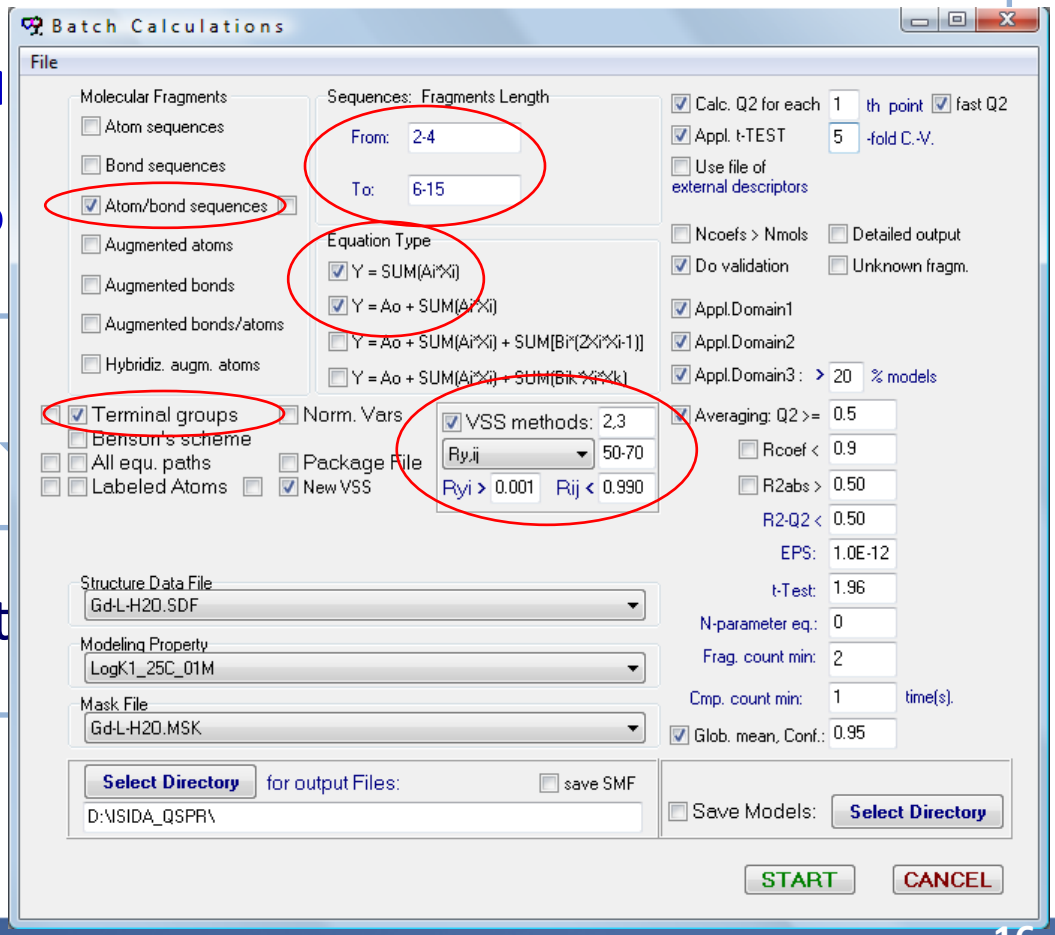
# Ensemble modeling by ISIDA/QSPR

## 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 descripto
2 types of linear equations

The accept



# Consensus model for predicted compound

## Model applicability domain and excluding of outlying predictions

Acceptable QSPR models

*Applicability domain (AD)*

*Bounding Box*

*Fragment Control*

*Quorum Control*

Applicable individual models

*Excluding of outlying predictions*

*Thompson's rule*

$$< Y_{pred} > = 1/M \sum_{i=1}^{M} Y_{pred,\, i}$$

Consensus model

# 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

**$M^{n+} + L = (M^{n+})L$, log*K* in water**

39 studied metals
2090 organic ligands
6749 log*K* values

| | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| H | | | | | | | | | | | | | | | | | 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., Tsivadze A. Yu., Varnek A. A. *Macroheterocycles*, **2012,** *5*, No. 4-5, pp 404-410
Solov'ev V., Marcou G., Tsivadze 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

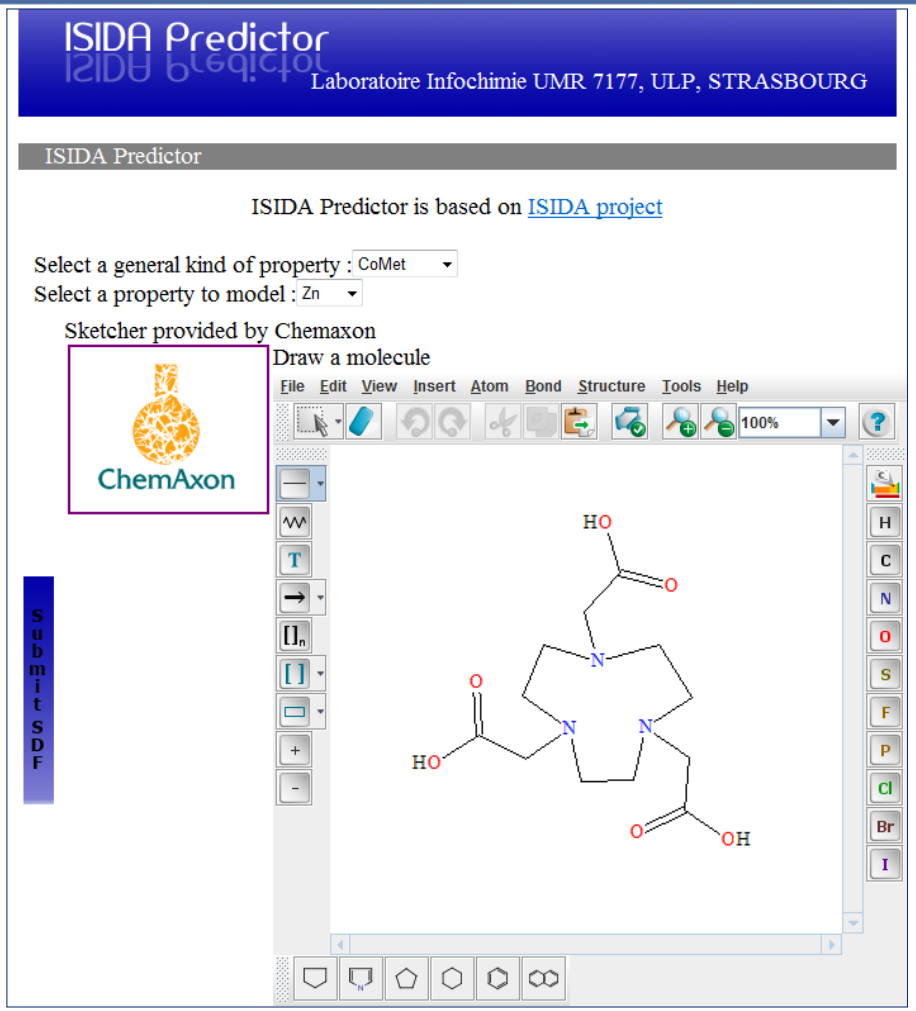
## Tools for design of new compounds

ISIDA predictor

Generator of virtual combinatorial libraries

Interactive designer of compounds

# ISIDA Predictor




## CoMet Predictor

Prediction of stability constants log*K*

metal - ligand complexes

27 metal ions

- Consensus Model predictions

  simulteneous 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., Tsivadze A., Solov'ev V. *Radiochim. Acta*, **2008**, *96*, 505-511 .

# Application of developed QSPR Models

## Tools for design of new compounds

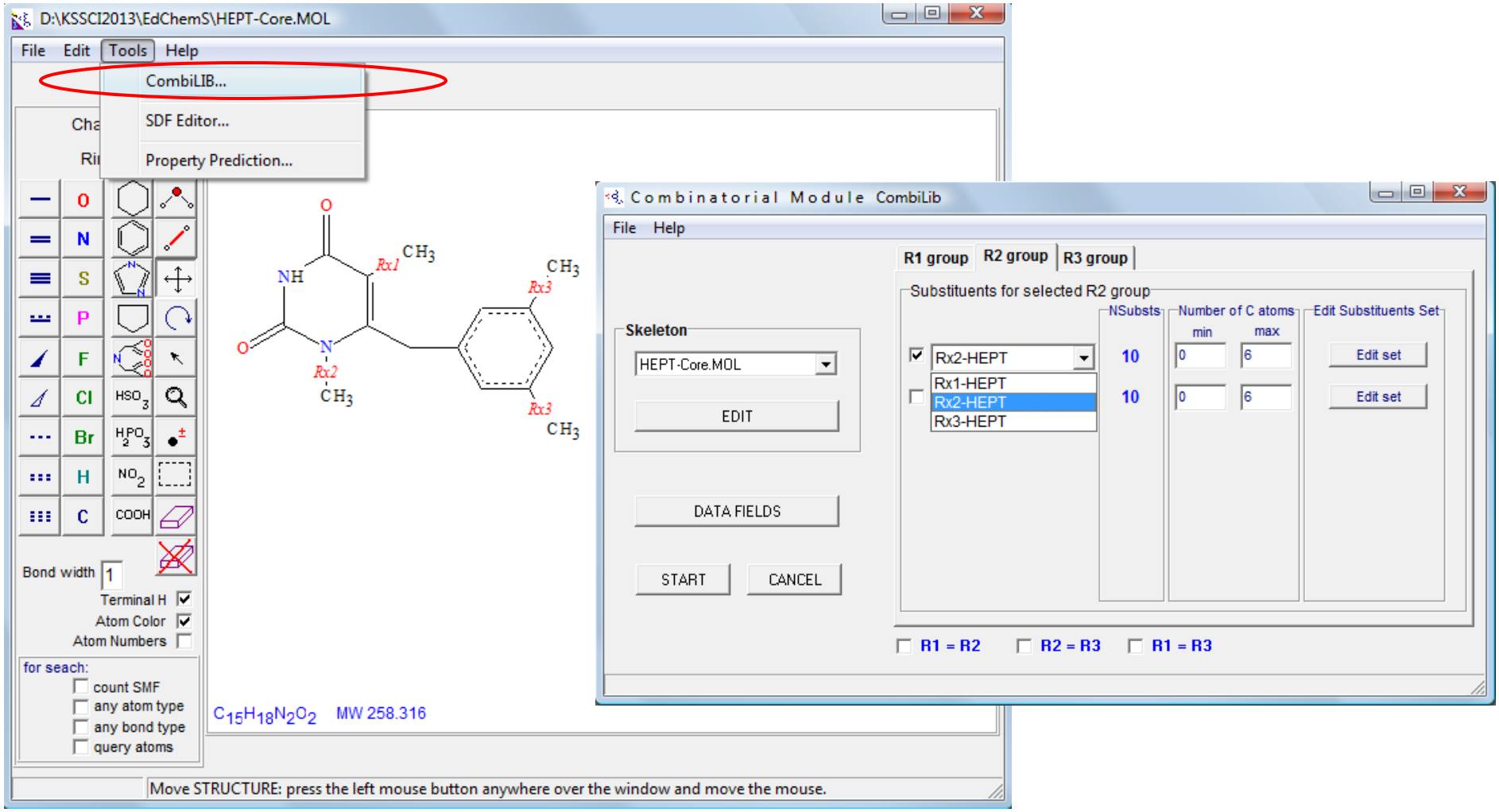
ISIDA predictor available via Internet

Generator of virtual combinatorial libraries

Interactive designer of compounds

# Chemical editor EdChemS

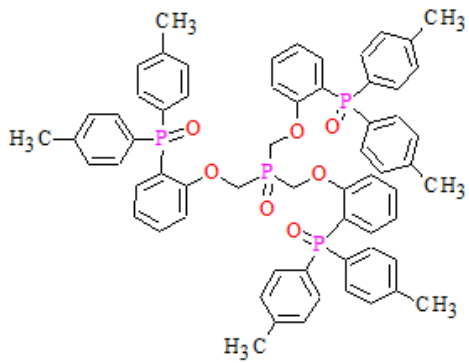
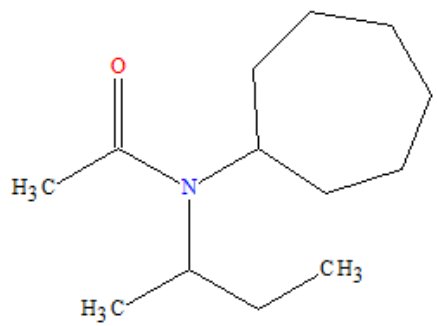
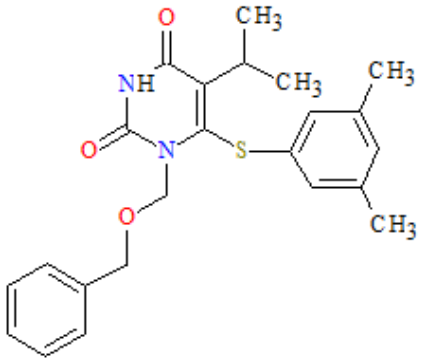
## Tools for generation of virtual combinatorial libraries




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 | $\lg D_{pred}$<br>$2.03 \pm 0.06$ (35) | $\lg D_{pred}$<br>$1.09 \pm 0.02$ (397) | $\lg(1/EC_{50})_{pred}$<br>$9.8 \pm 0.2$ (285) |
| Experimentally tested activity | $\lg D_{exp}$<br>$1.81 \pm 0.05$ | $\lg D_{exp}$<br>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

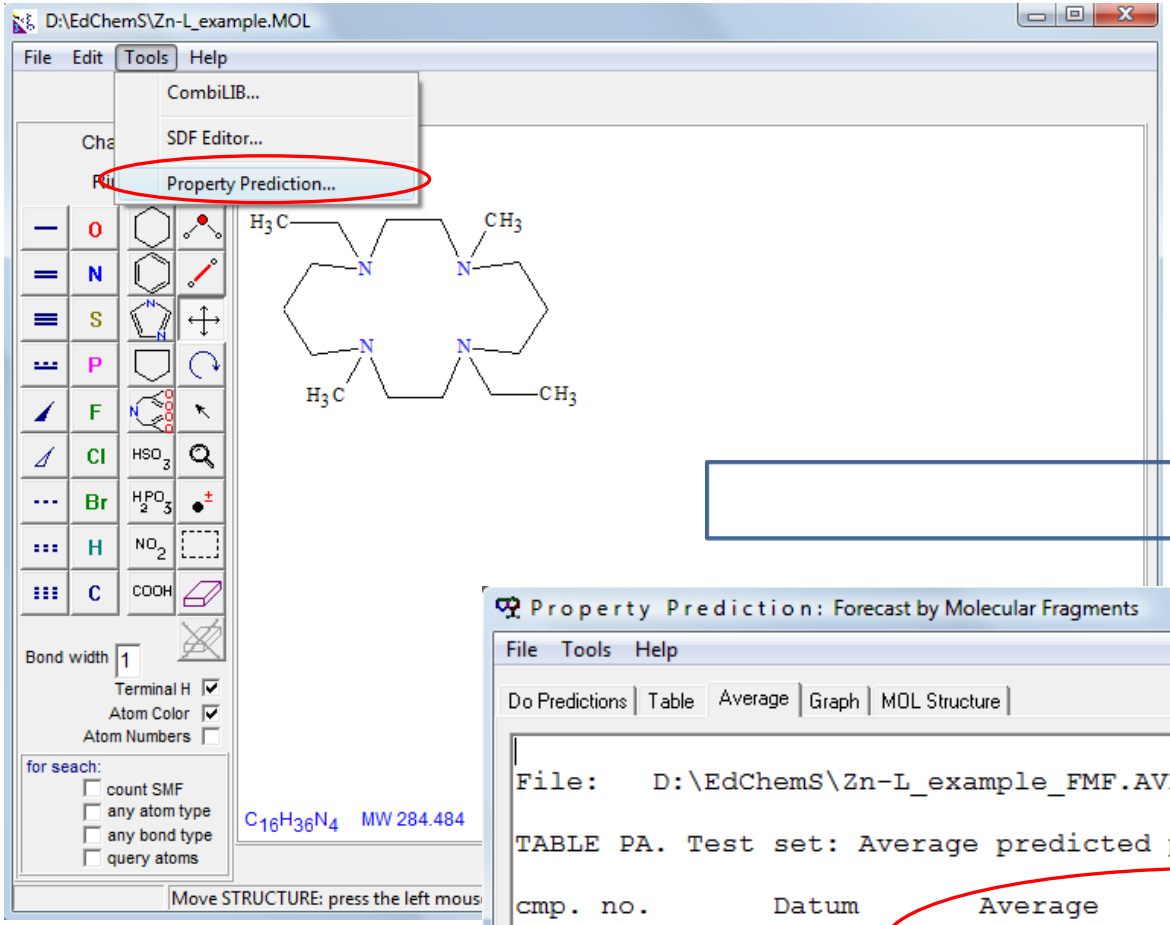
Interactive designer of compounds

by interaction of editor EdChemS with predictor FMF

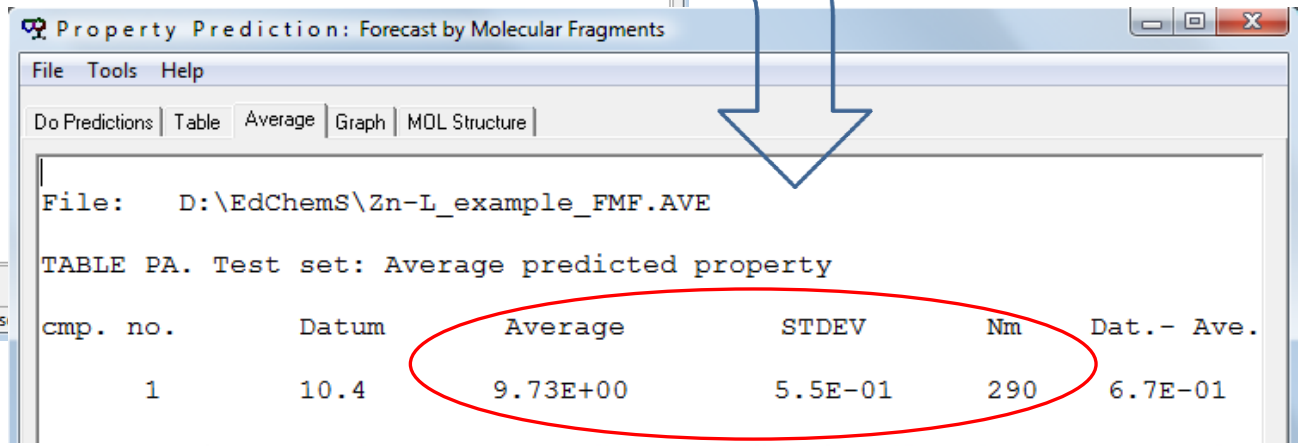
# Interactive design of compounds

## Editor EdChemS + Predictor FMF

Drawing by EdChemS

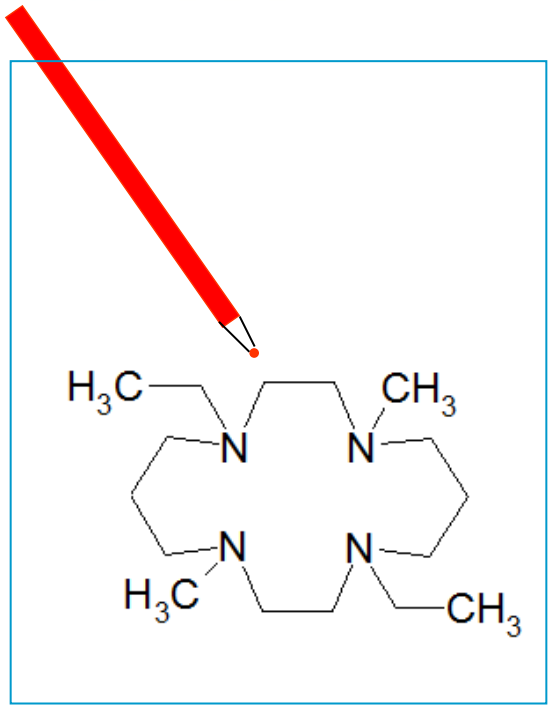



Property Prediction by FMF



# Interactive design of compounds

## Chemical editor EdChemS + Predictor FMF




mol

**predicted property**

| *exp.* | *predict.* | $N_{models}$ |
|---|---|---|
| - | $9.7_{\pm 0.6}$ | 290 |

Pool of ISIDA MLR + PLS Models

Drawing by EdChemS

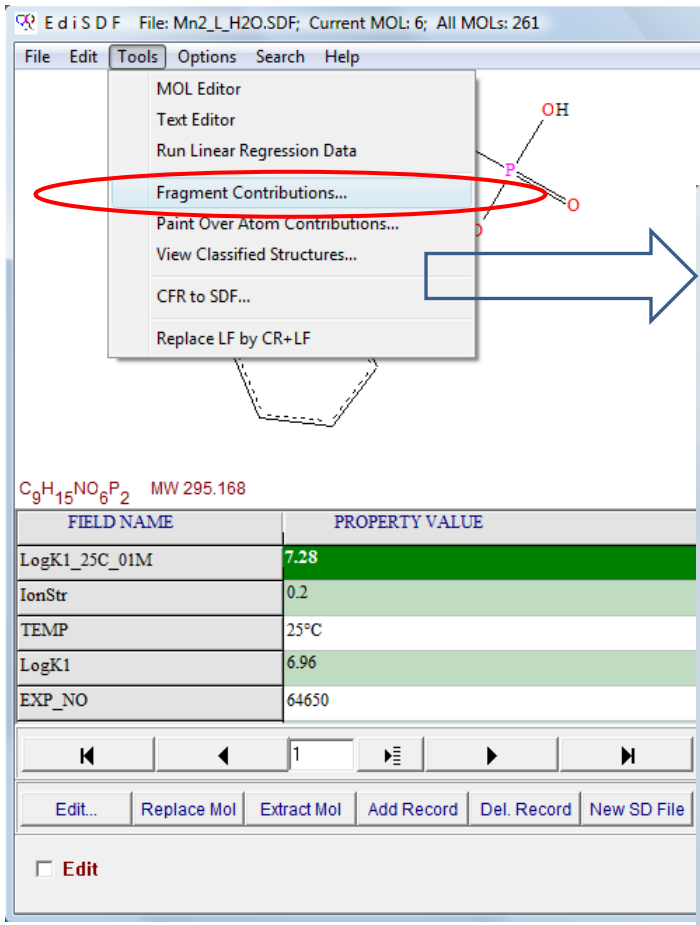
Property Prediction by FMF

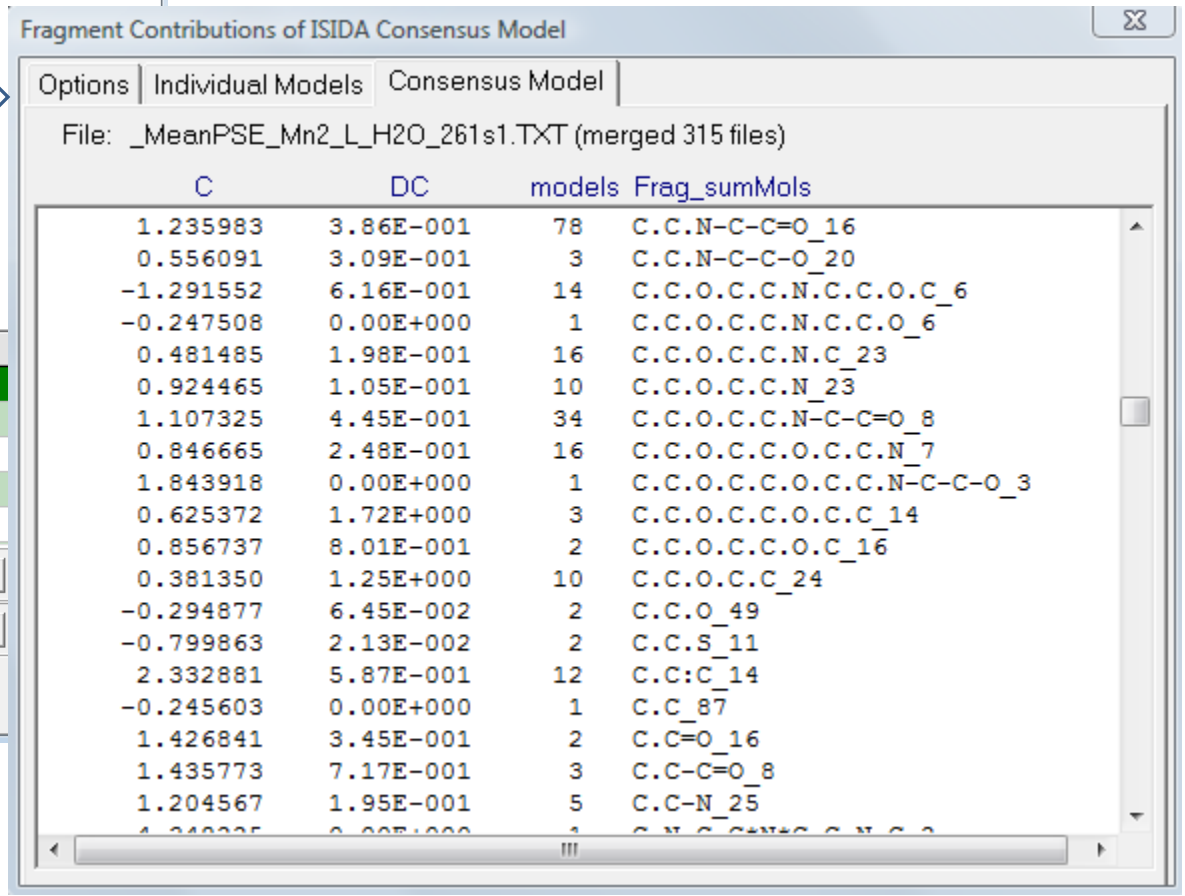
Models developed by ISIDA QSPR

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

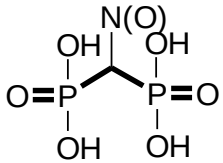
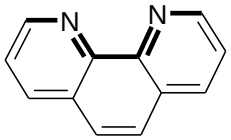
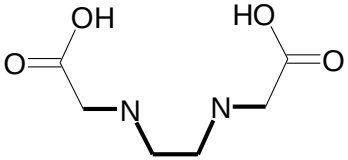
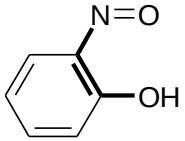
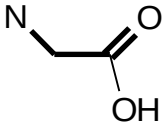



| C | DC | models | Frag_sumMols |
|---|---|---|---|
| 1.235983 | 3.86E-001 | 78 | C.C.N-C-C=O_16 |
| 0.556091 | 3.09E-001 | 3 | C.C.N-C-C-O_20 |
| -1.291552 | 6.16E-001 | 14 | C.C.O.C.C.N.C.C.O.C_6 |
| -0.247508 | 0.00E+000 | 1 | C.C.O.C.C.N.C.C.O_6 |
| 0.481485 | 1.98E-001 | 16 | C.C.O.C.C.N.C_23 |
| 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 |

# Mean SMF contributions: effective fragments

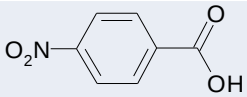
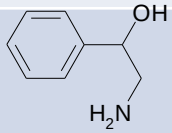
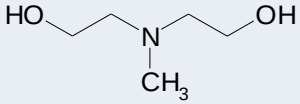
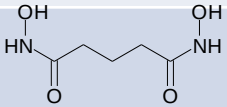
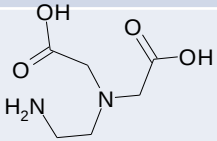
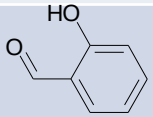
## M + L = ML, log*K*

| M | SMF | $< a_i >$ | $N_{model}$ | $N_{mol}$ | 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 |  |

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

# From effective fragments to selective ligands

## M + L = ML, log*K*

| 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 |

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

# Application of developed QSPR Models

## Tools for design of new compounds

ISIDA predictor available via Internet

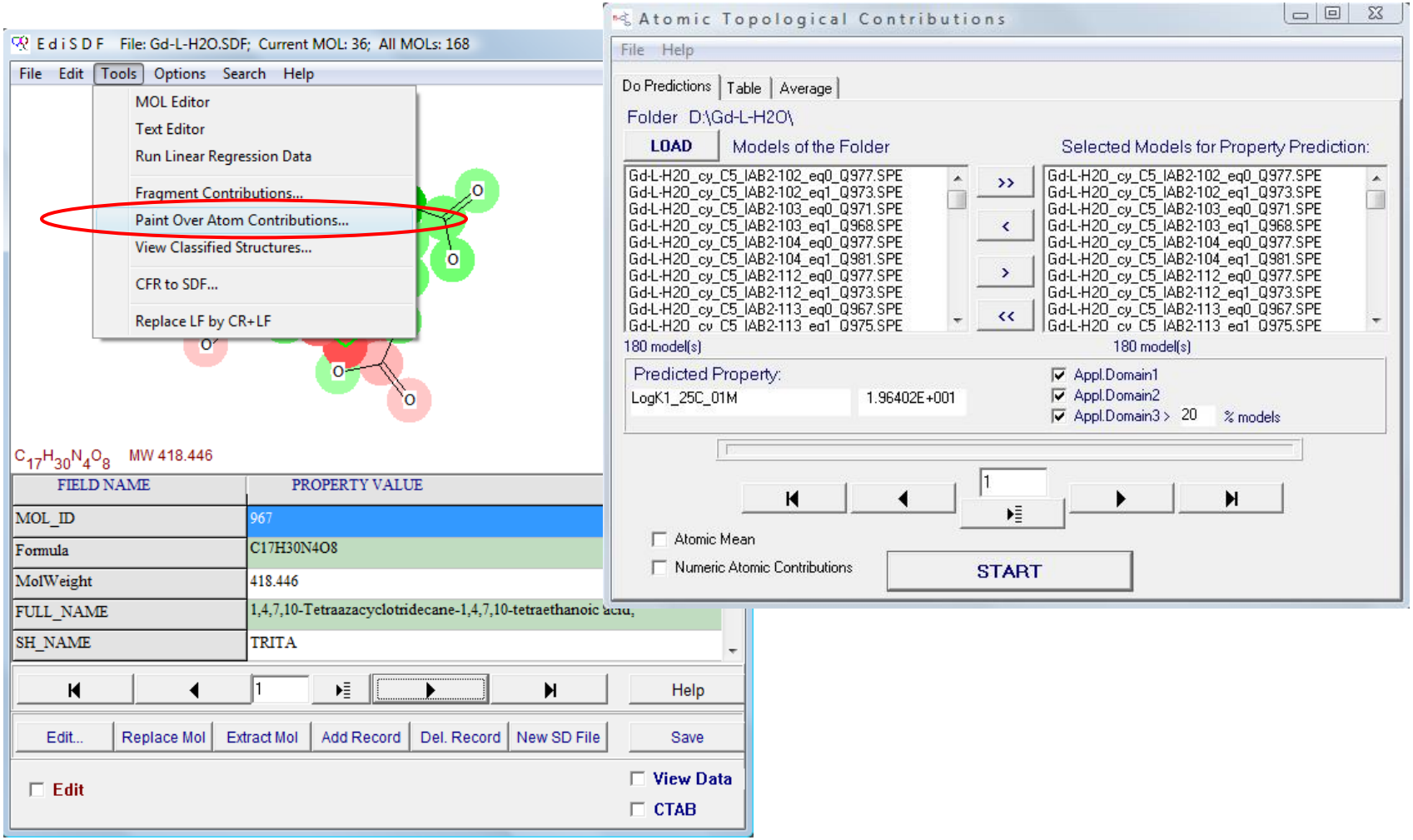
Generator of virtual combinatorial libraries

Interactive designer of compounds

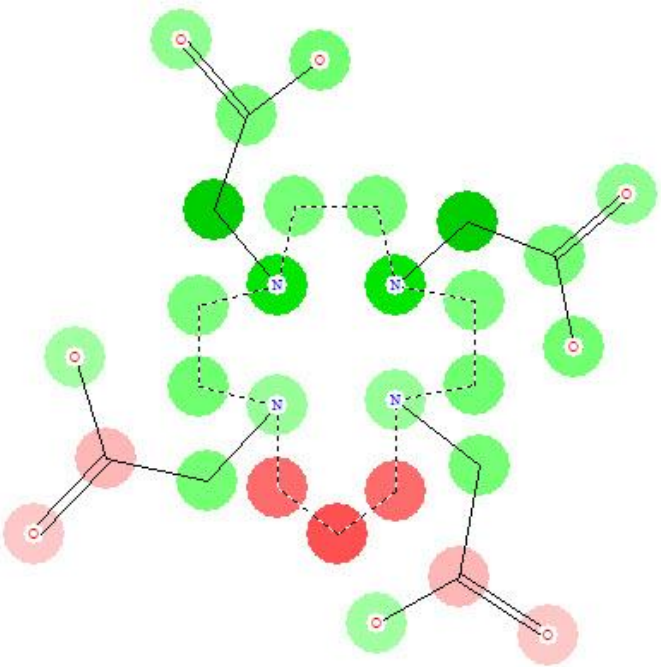
by coloring of atoms of chemical formula according fragment contributions

# Interactive design of compounds

## Tool for coloration of atoms of chemical formula



# Coloration of atoms according fragment contributions of consensus model



Color depth of atom

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

equilibrium

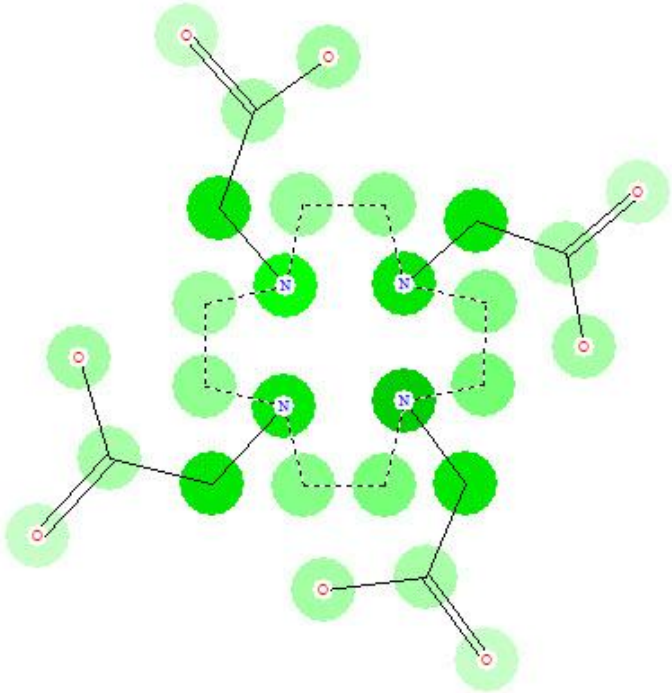
$Gd^{3+} + L = (Gd^{3+})L$

stability constant

$\log K_{exp} = 19.2$

topological contributions of ligand atoms

# Coloration of atoms according fragment contributions of consensus model



Color depth of atom

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

equilibrium

$Gd^{3+} + L = (Gd^{3+})L$

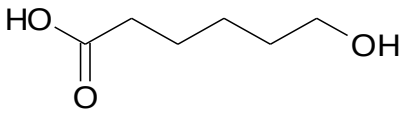
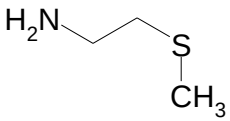
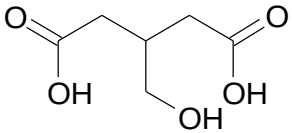
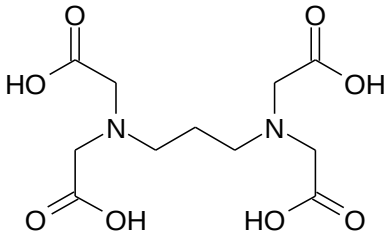
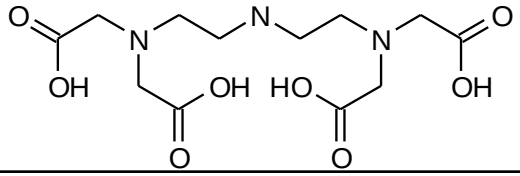
stability constant

$\log K_{exp} = 24.7$

topological contributions of ligand atoms

# Interactive design of compounds

## The log*K* tuning of ($Zn^{2+}$)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**