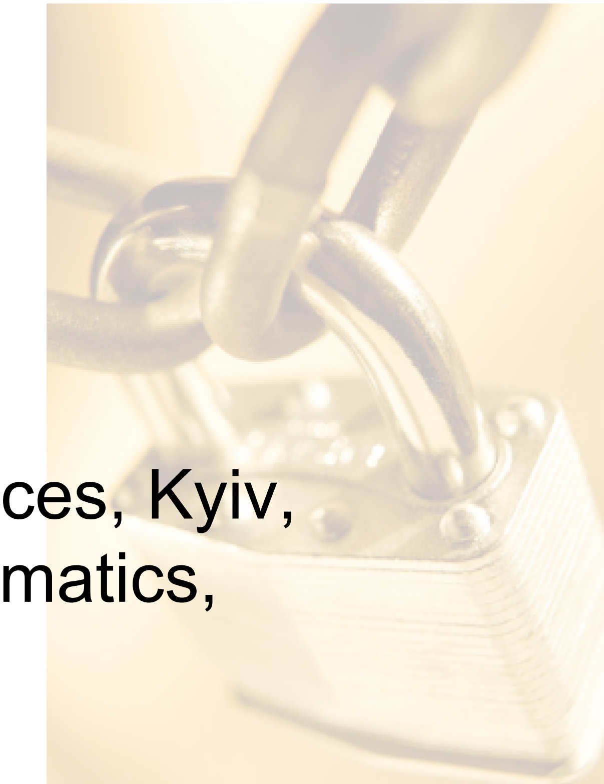
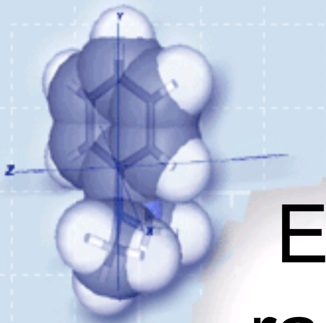
Encoding molecular structures as ranks of models: A new, secure way for sharing chemical data and development of ADME/T models

Igor V. Tetko

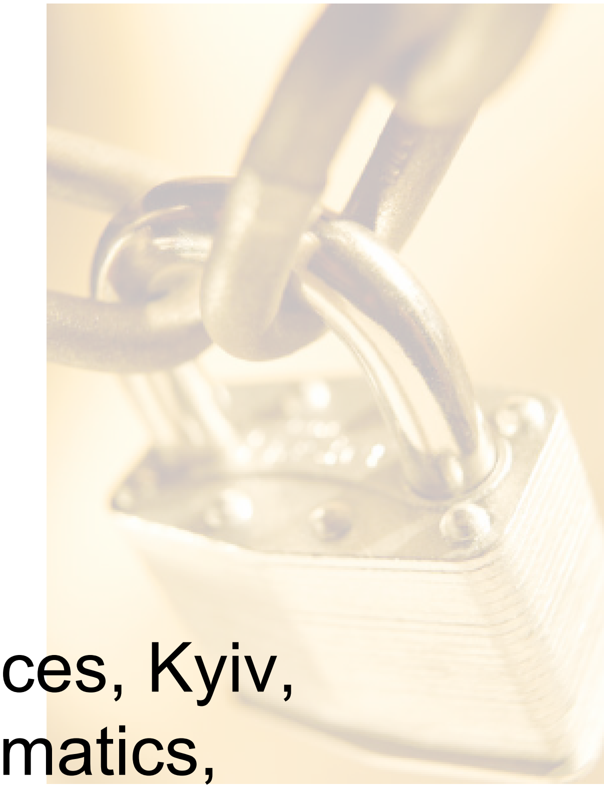
IBPC, Ukrainian Academy of Sciences, Kyiv, Ukraine and Institute for Bioinformatics, Munich, Germany



March 14th, ACS



Encoding molecular structures as SHUFFLED ranks of models: A new, secure way for sharing chemical data and development of ADME/T models



Igor V. Tetko

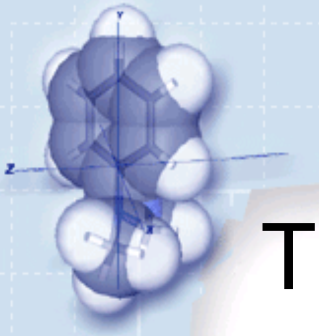
IBPC, Ukrainian Academy of Sciences, Kyiv, Ukraine and Institute for Bioinformatics, Munich, Germany

Structure-Property correlations

- Require representation (description) of the molecule in a format that can be used for machine learning methods, i.e. MLRA, neural network, PLS
- Two major systems: topological and 3D based

- Fragment-based indices
- topological indices
- E-state indices

- Quantum-chemical parameters
- VolfSurf descriptors
- Molecular shape parameters

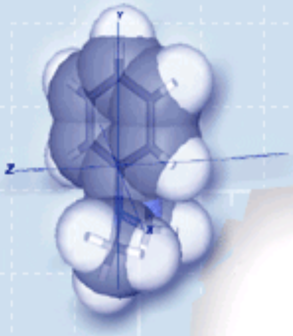


Three scenario for structure decoding

- Can we identify the molecule provided we have it in our portfolio? **-- the most difficult scenario**

- Can we do the same in knowledge that the molecule can be originated from one of several chemical series?

- Can we identify the molecule provided we do not know anything about it? **-- the practical scenario**



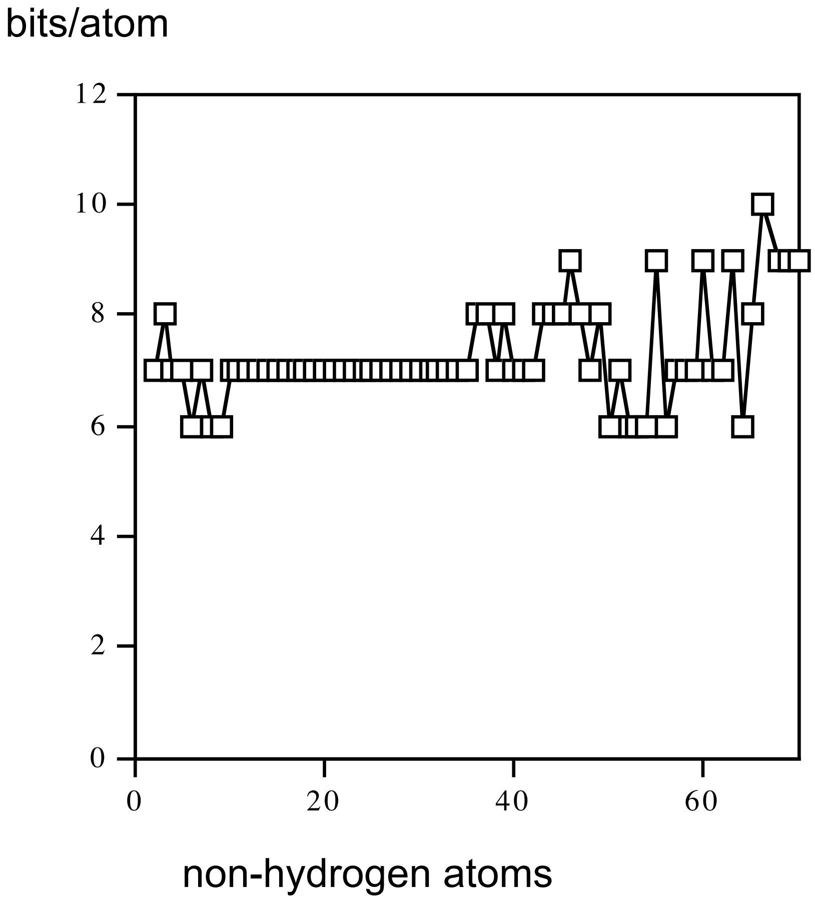
Can we identify the molecule provided we have it in our portfolio? Topological indices.

- The ability to unambiguously identify a molecule is limited to information content of indices
- If the indices contain sufficient information, the identification is possible
- Information content of a molecule:
- CCCCC -- 11111 (5 bits)
- C1CCCC1N -- 12111123 (11 bits)

C -- 1 bit
1 -- 2 bits
N -- 3 bits

Information content of molecules in set of 12908 molecules (PHYSPROP database)

Element	Frequency	Bits
C	78777	1
c	76965	2
)	42336	3
(	42336	4
O	29349	5
1	23648	6
=	20610	7
N	16156	8
2	12658	9



not optimal -- Huffman, arithmetic coding, other algorithms:
gz, zip -- 3.5 bits/atom, bzip2 -- 2.9 bits/atom

Information content of a molecule

- 30 -- 40 atoms -- 90 -- 110 bits

- 1 double value -- 32 bits, 3 -- 4 topological indices potentially contains sufficient information to unambiguously decode molecule with 40 atoms!

- In reality a larger number of indices can be required because of rounding effects, non-optimal storage of information

- Thus, the encoding of molecules using topological indices can be insecure.

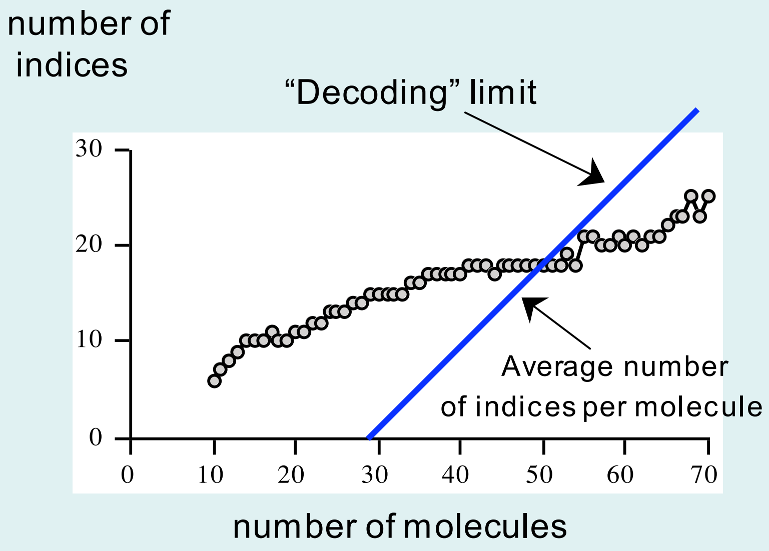
When reverse engineering is impossible? A practical scenario.

- ALOGPS program:
  75 indices per molecule for logP
  33 indices per molecule for logS

- We use decreased resolution of data, i.e to just 3 significant digits per index (7-10 bits instead of 32 bits)
- Additional bits are coming from range ~ 11 bits per index => 10-12 indices per molecule with 40 atoms

The information encoded in the indices could be (theoretically) adequate to decode the molecules with < 50 heavy atoms.

But, this can be too pessimistic conclusion. The theoretical possibility to decode does not propose a way how this can be done!



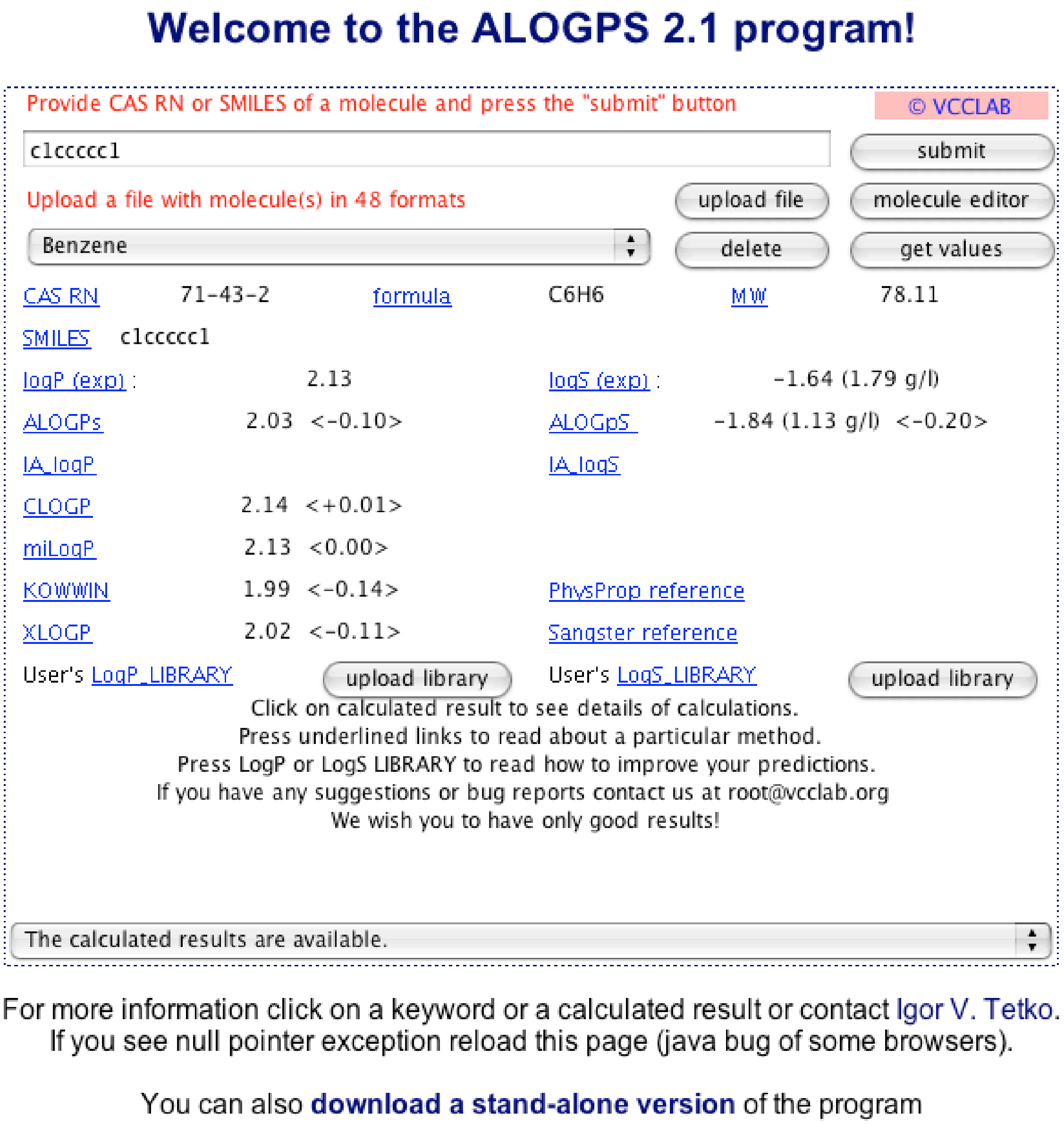
ALOGPS 2.1

- LogP: 75 input variables corresponding to electronic and topological properties of atoms (E-state indices), 12908 molecules in the database, 64 neural networks in the ensemble. Calculated results RMSE=0.35, MAE=0.26, n=76 outliers (>1.5 log units)

- LogS: 33 input E-state indices, 1291 molecules in the database, 64 neural networks in the ensemble. Calculated results RMSE=0.49, MAE=0.35, n=18 outliers (>1.5 log units)

- **Tetko, Tanchuk & Villa, JCICS, 2001, 41, 1407-1421.**
- **Tetko, Tanchuk, Kasheva & Villa, JCICS, 2001, 41, 1488-1493.**
- **Tetko & Tanchuk, JCICS, 2002, 42, 1136-1145.**

http://www.vcclab.org



Welcome to the ALOGPS 2.1 program!

Provide CAS RN or SMILES of a molecule and press the "submit" button

c1ccccc1

Upload a file with molecule(s) in 48 formats

Benzene

CAS RN	71-43-2	formula	C6H6	MW	78.11
SMILES	c1ccccc1				

logP (exp) :	2.13	logS (exp) :	-1.64 (1.79 g/l)
ALOGPs	2.03 <-0.10>	ALOGpS	-1.84 (1.13 g/l) <-0.20>
IA_logP		IA_logS	
CLOGP	2.14 <+0.01>		
miLogP	2.13 <0.00>		
KOWWIN	1.99 <-0.14>	PhysProp reference	
XLOGP	2.02 <-0.11>	Sangster reference	
User's LogP_LIBRARY		User's LogS_LIBRARY	

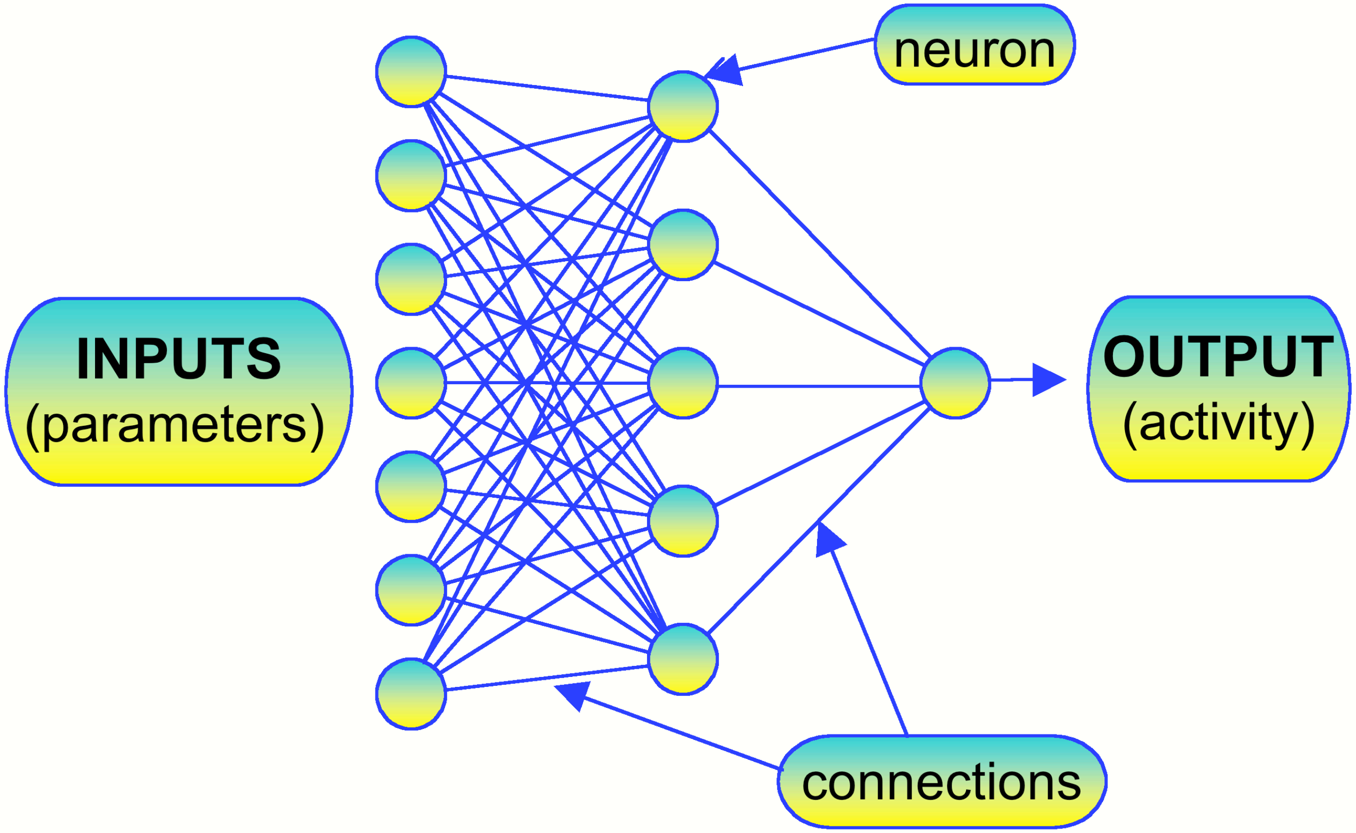
Click on calculated result to see details of calculations.
Press underlined links to read about a particular method.
Press LogP or LogS LIBRARY to read how to improve your predictions.
If you have any suggestions or bug reports contact us at root@vcclab.org
We wish you to have only good results!

The calculated results are available.

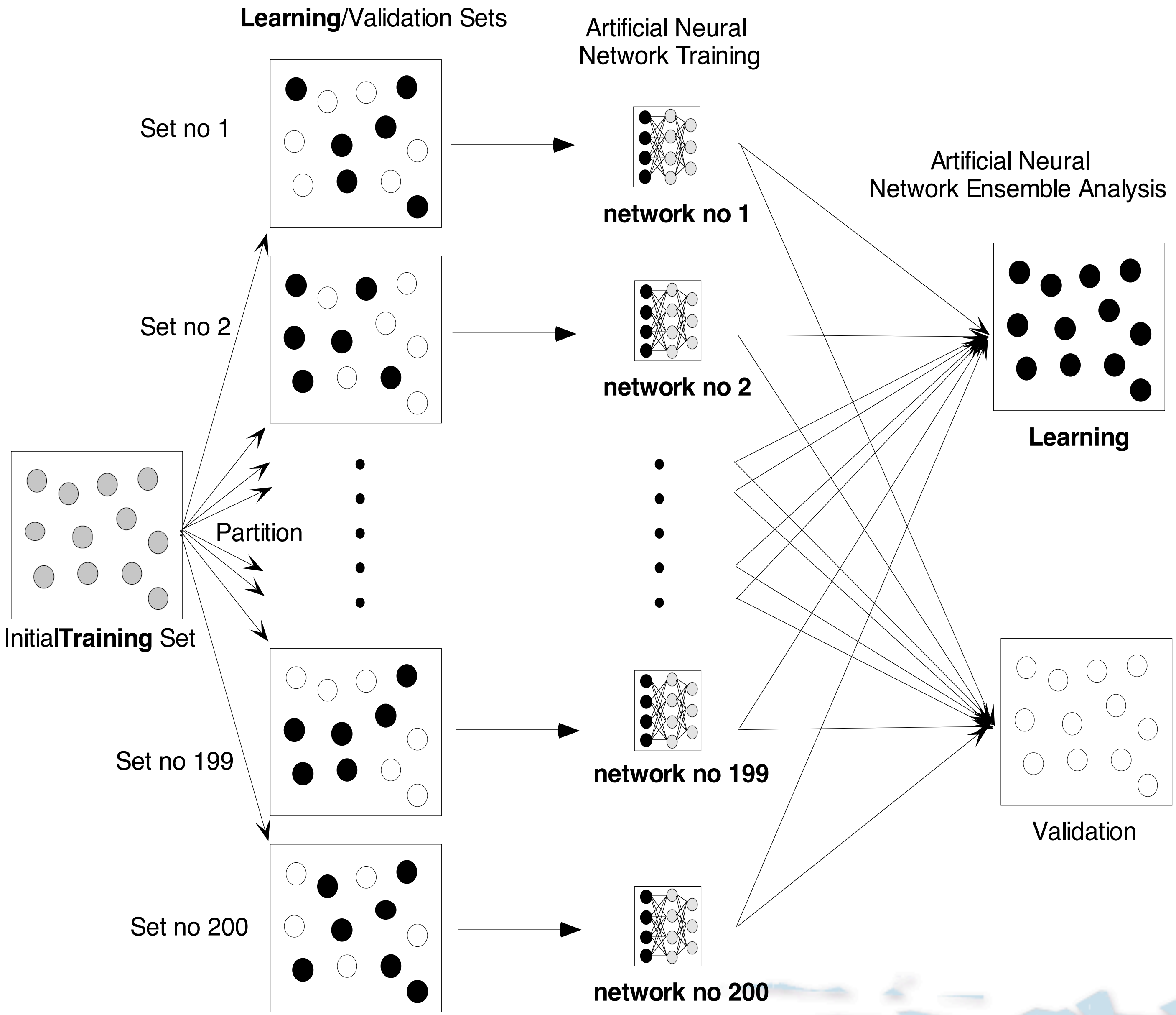
For more information click on a keyword or a calculated result or contact Igor V. Tetko.
If you see null pointer exception reload this page (java bug of some browsers).

You can also **download a stand-alone version** of the program

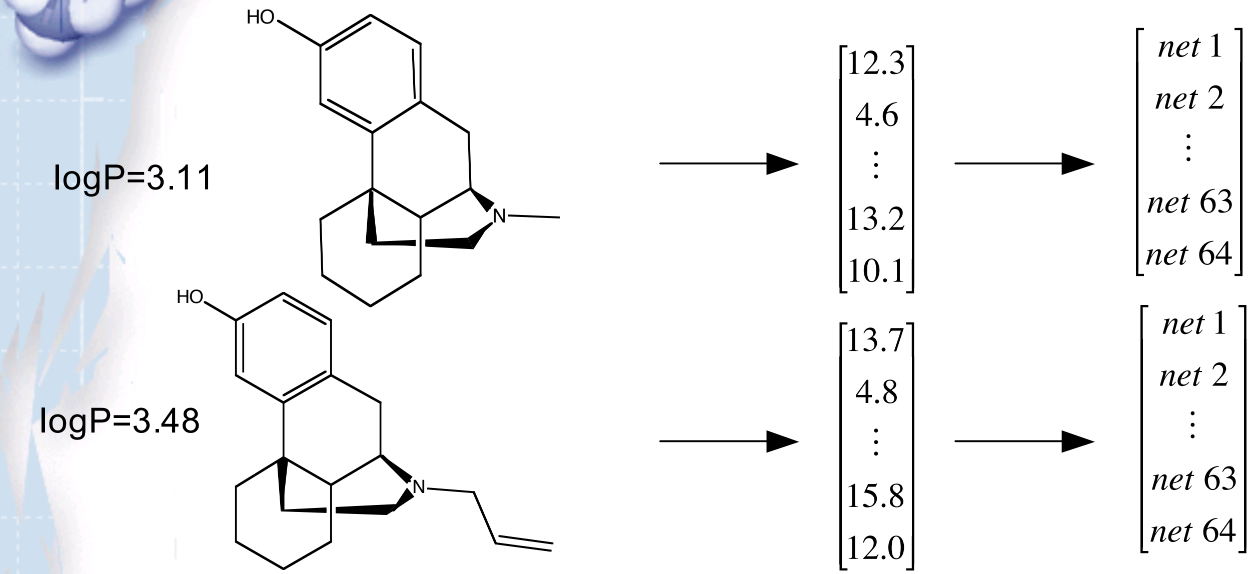
Artificial Feed-Forward Back-propagation Neural Network (FBNN)



Early Stopping Over Ensemble (ESE)



ASNN: an example correction



logP=3.11

$$\begin{bmatrix} 12.3 \\ 4.6 \\ \vdots \\ 13.2 \\ 10.1 \end{bmatrix} \rightarrow \begin{bmatrix} net\ 1 \\ net\ 2 \\ \vdots \\ net\ 63 \\ net\ 64 \end{bmatrix}$$

logP=3.48

$$\begin{bmatrix} 13.7 \\ 4.8 \\ \vdots \\ 15.8 \\ 12.0 \end{bmatrix} \rightarrow \begin{bmatrix} net\ 1 \\ net\ 2 \\ \vdots \\ net\ 63 \\ net\ 64 \end{bmatrix}$$

1-kNN correction

Morphinan-3-ol, 17-methyl-

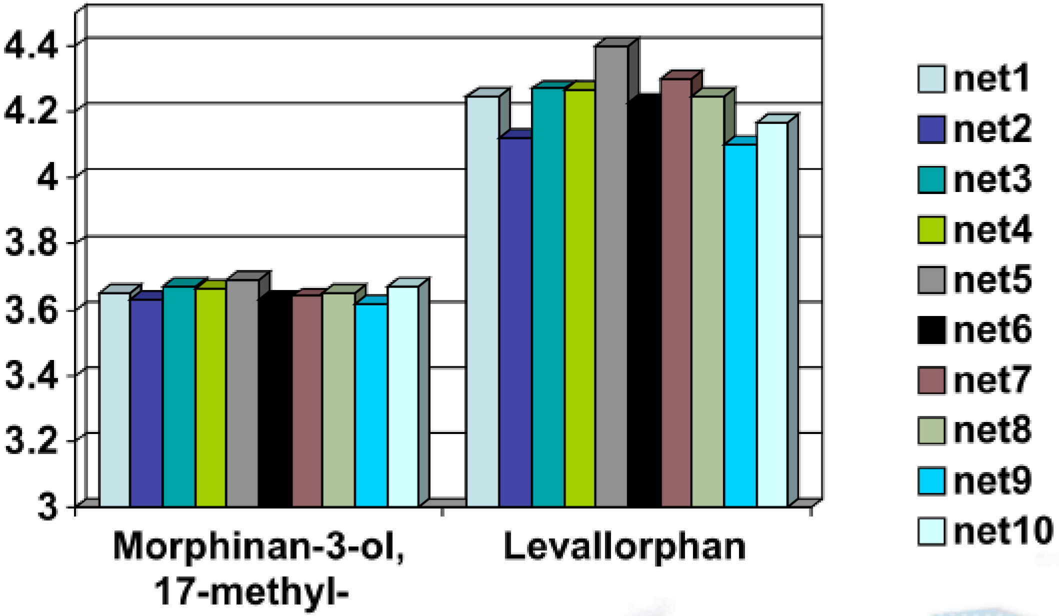
Calculated logP=3.65, δ=+0.54

--> 3.65-0.76=2.89 (δ=+0.22)

Levallorphan

Calculated logP=4.24, δ=+0.76

--> 4.24-0.54=3.70 (δ=+0.22)



-- both molecules are the nearest neighbors, r^2=0.47, in space of residuals!

Associative Neural Network (ASNN)

Ensemble approach:

A prediction of case i: $$[\mathbf{x}_i] \bullet [\mathbf{ANNE}]_M = [\mathbf{z}_i] = \begin{bmatrix} z_1^i \\ \vdots \\ z_k^i \\ \vdots \\ z_M^i \end{bmatrix}$$

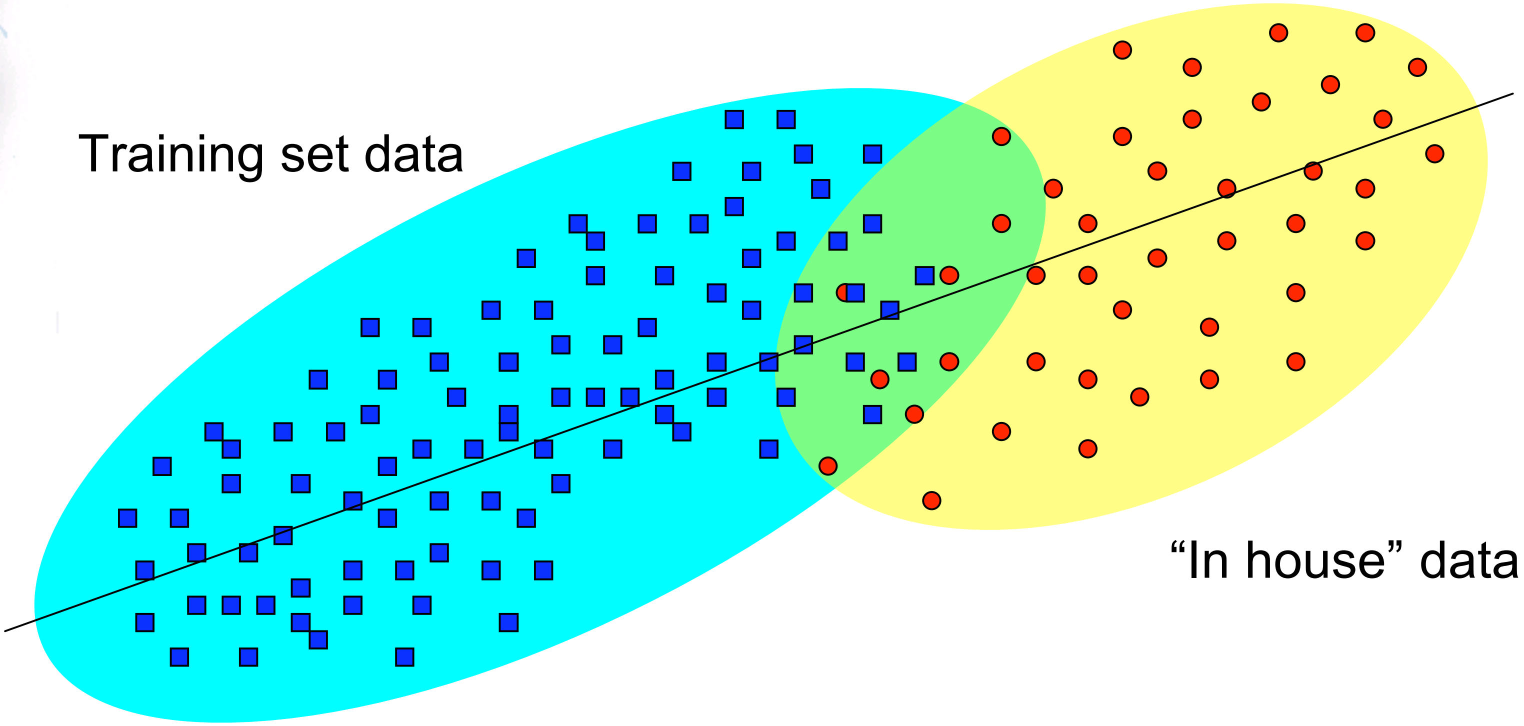
$$\bar{z}_i = \frac{1}{M} \sum_{k=1,M} z_k^i$$

Pearson's (Spearman) correlation coefficient $r_{ij}=R(z_i,z_j)>0$ ***in space of residuals***

$$\bar{z}_i' = \bar{z}_i + \tfrac{1}{k} \sum_{j \in N_k(\mathbf{x}_i)} \left(y_j - \bar{z}_j\right)$$ **<<= ASNN bias correction**

The correction of neural network ensemble value is performed using errors (biases) calculated for the neighbor cases of analyzed case $\mathbf{x}_i$ detected in space of neural network models

Prediction Space of the model does not cover the “in house” compounds



Each new molecule is encoded as rank of models

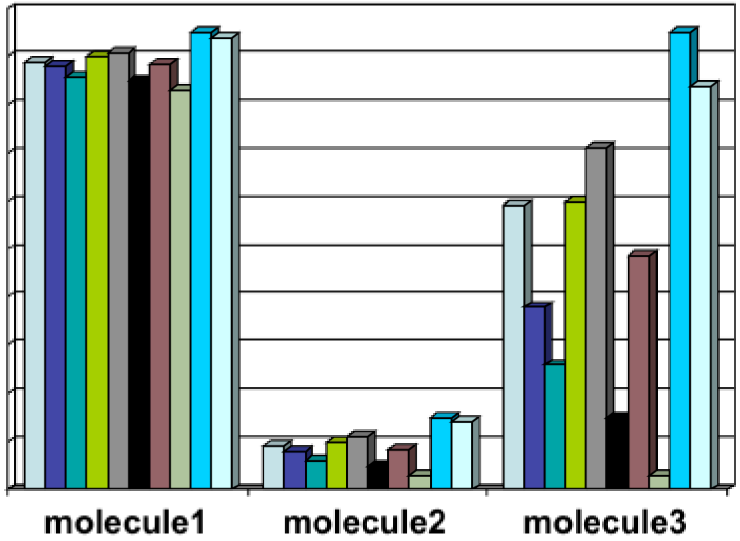
Encoding of a molecule as rank of models

- ΔlogP=logPexp-logPcalc
- 64 values, ranks of NN

0.89	0.88	0.86	0.90	0.91	0.85	.885	0.83	0.95	0.94
5	7	8	4	3	9	6	10	1	2

0.09	0.08	0.06	0.10	0.11	0.05	.085	0.03	0.15	0.14
5	7	8	4	3	9	6	10	1	2

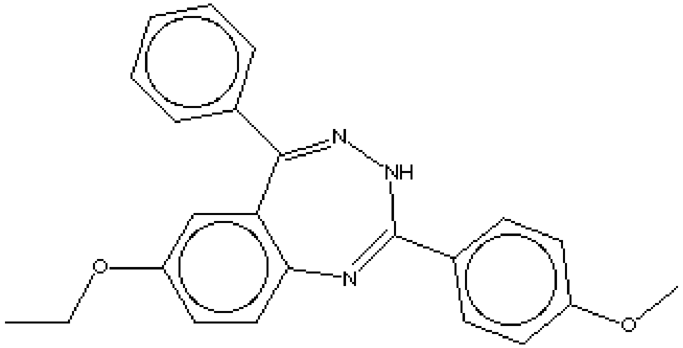
0.59	0.38	0.26	0.60	0.71	0.15	.485	0.03	0.95	0.84
5	7	8	4	3	9	6	10	1	2

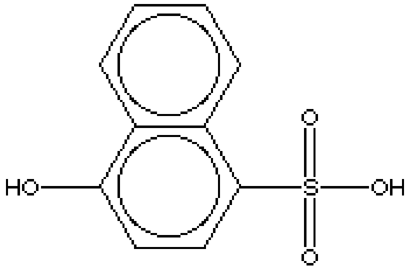


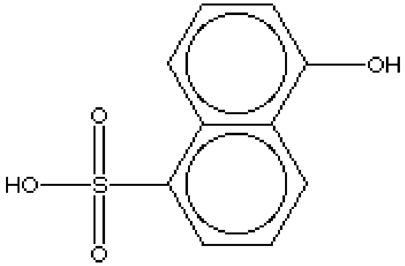
Millions of solutions provide the same ranks of NN responses -->
no way to decode -- previous name of the paper, but…

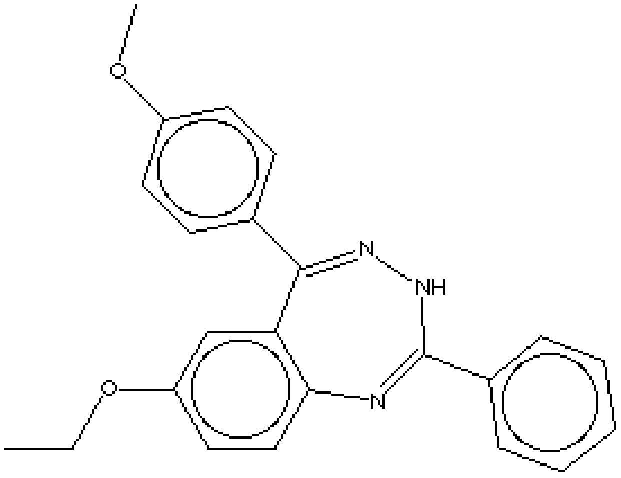
How selective is rank coding?

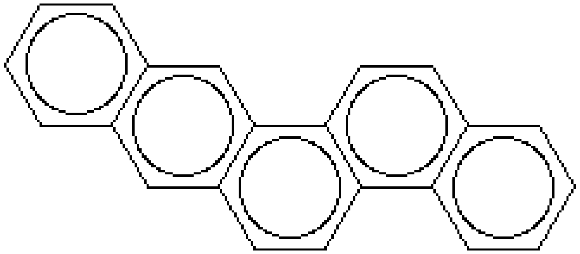
- 8x64 = 512 bits (comparable to MDL keys)
- 126 out of 121281 Asiprox (0.1%)
- 12 out of 12908 PHYSPROP (0.1%)

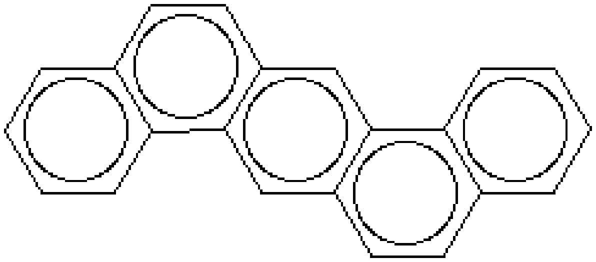




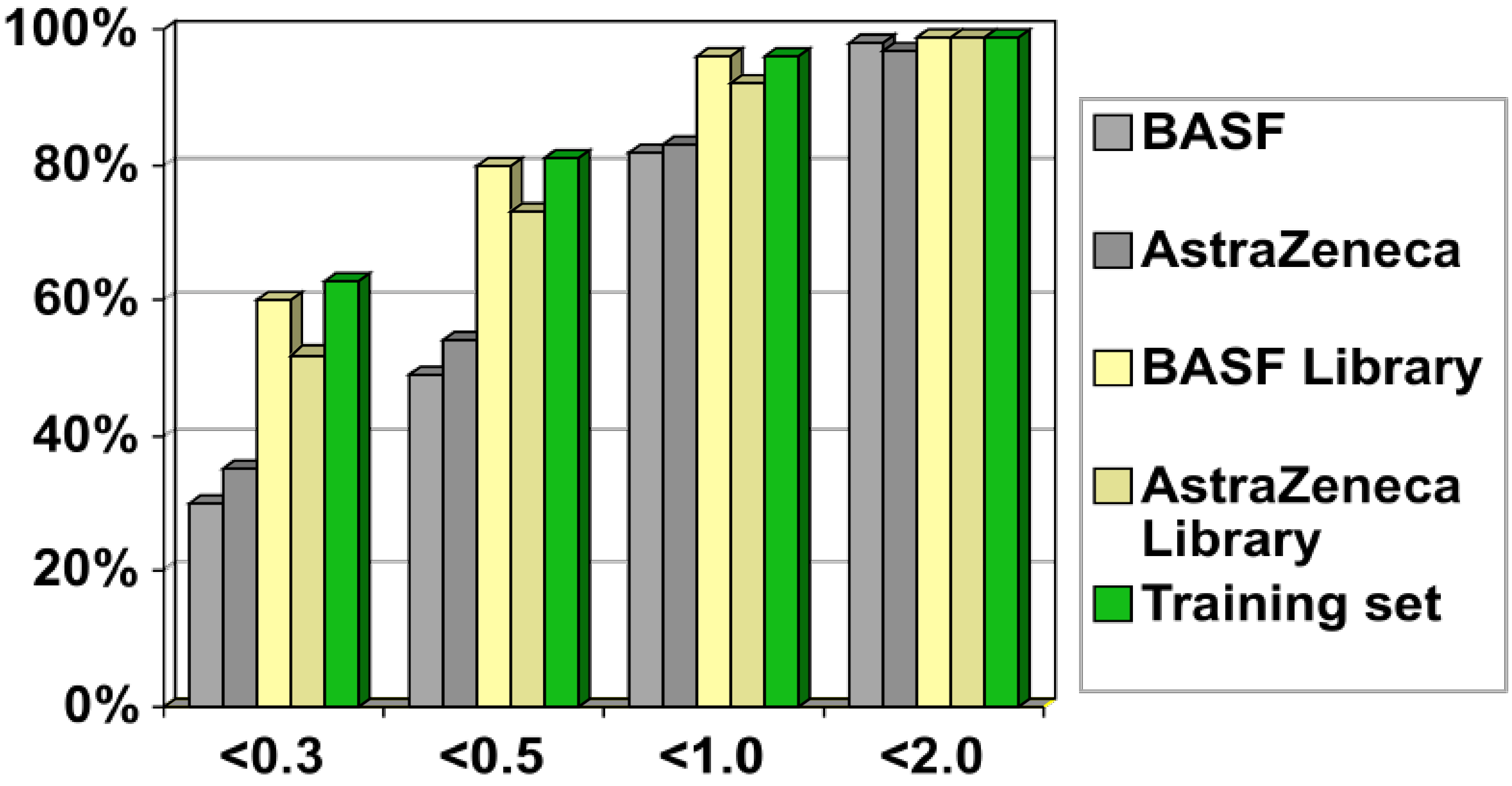








ALOGPS: Extrapolation vs Interpolation

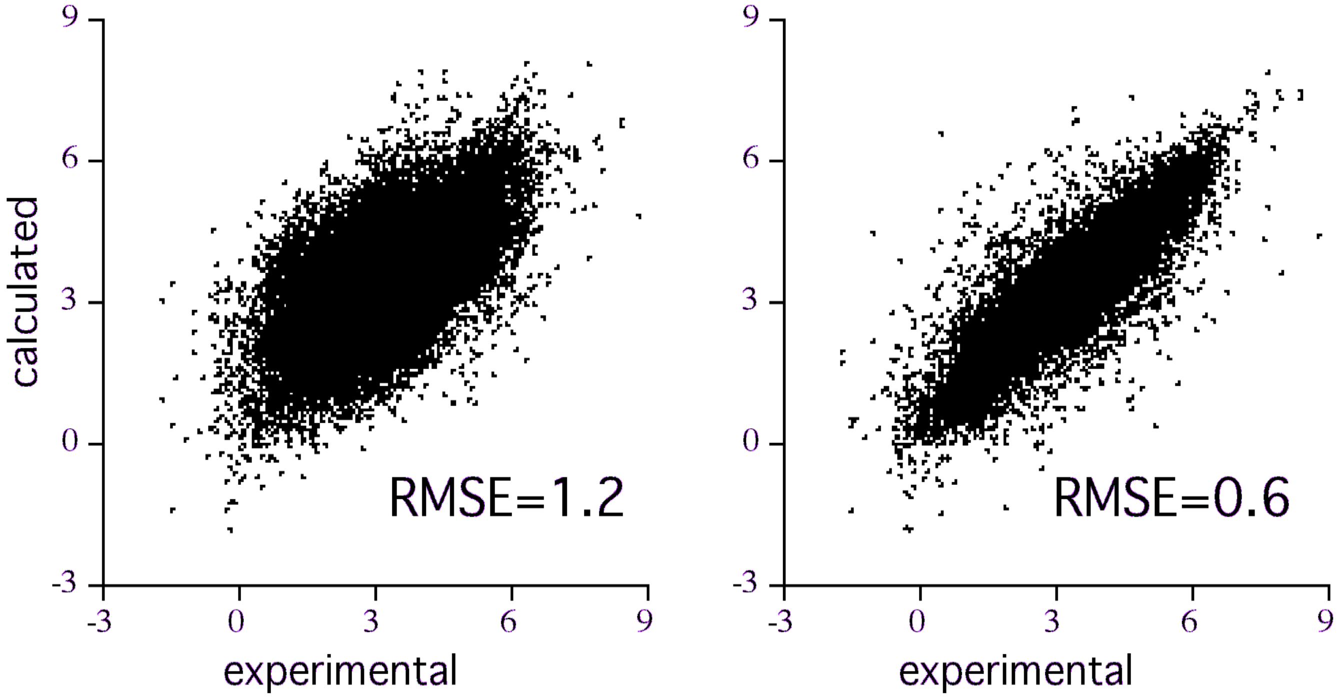


ALOGPS logP (blind) **:***MAE = 1.27, RMSE=1.63*
ALOGPS logP (LIBRARY):*MAE = 0.49, RMSE=0.70*

Tetko, JCICS, 2002, 42, 717-742.
Tetko & Bruneau, J. Pharm. Sci., 2004, 94, 3103-3110.

Analysis of Pfizer data

ALOGPS prediction for ElogD set of 17,861 compounds



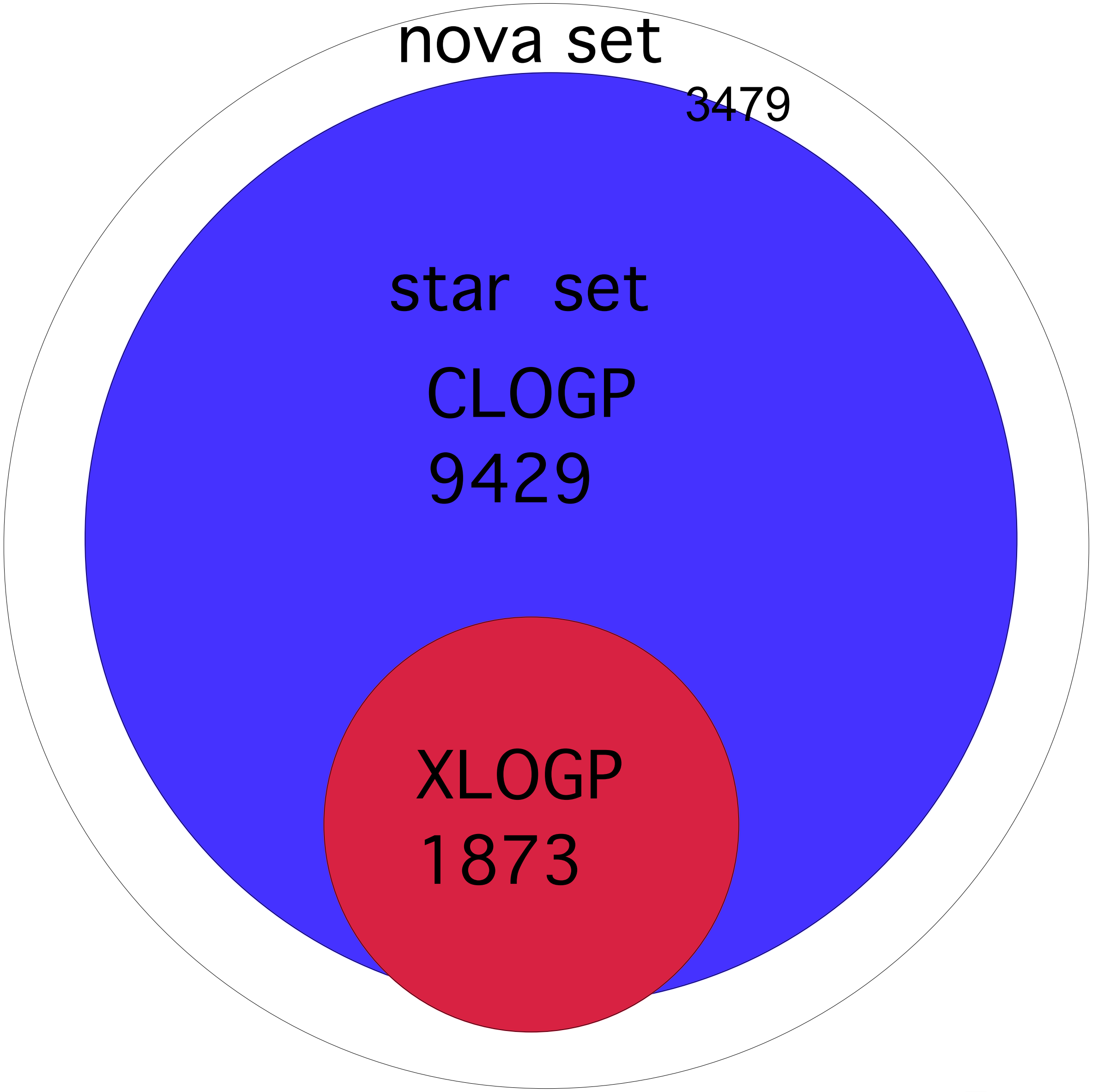
ALOGPS "as is" → ALOGPS LIBRARY

Pallas PrologD :	*MAE = 1.06, RMSE=1.41*
ACDlogD (v. 7.19):	*MAE = 0.97, RMSE=1.32*
ALOGPS:	*MAE = 0.92, RMSE=1.17*
ALOGPS LIBRARY:	*MAE = 0.43, RMSE=0.64*

Tetko & Poda, J. Med. Chem., 2004, 94, 5601-5604.

PHYSPROP data set

Total: 12908

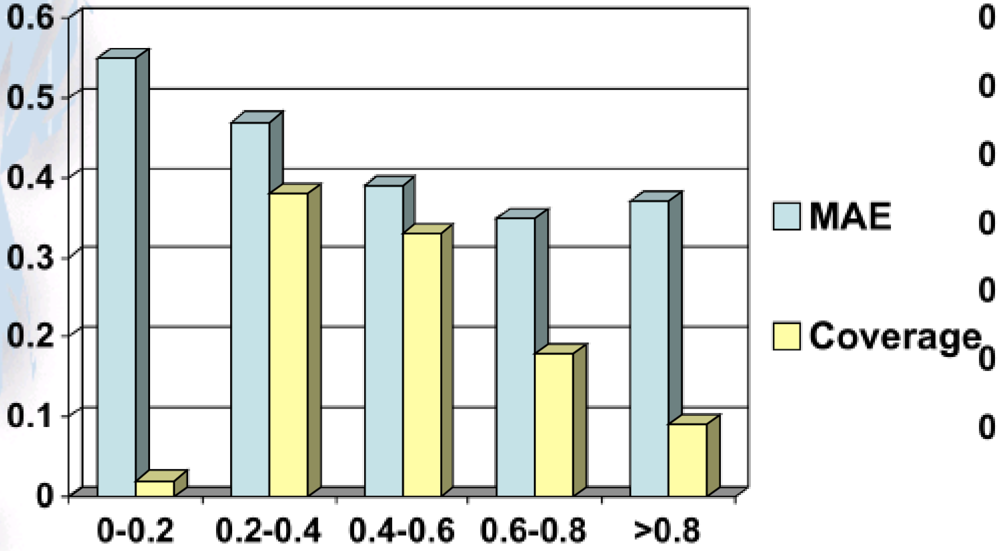


training "nova" --> prediction star set

of 12 908

Prediction performance as function of similarity in space of models of “star” set

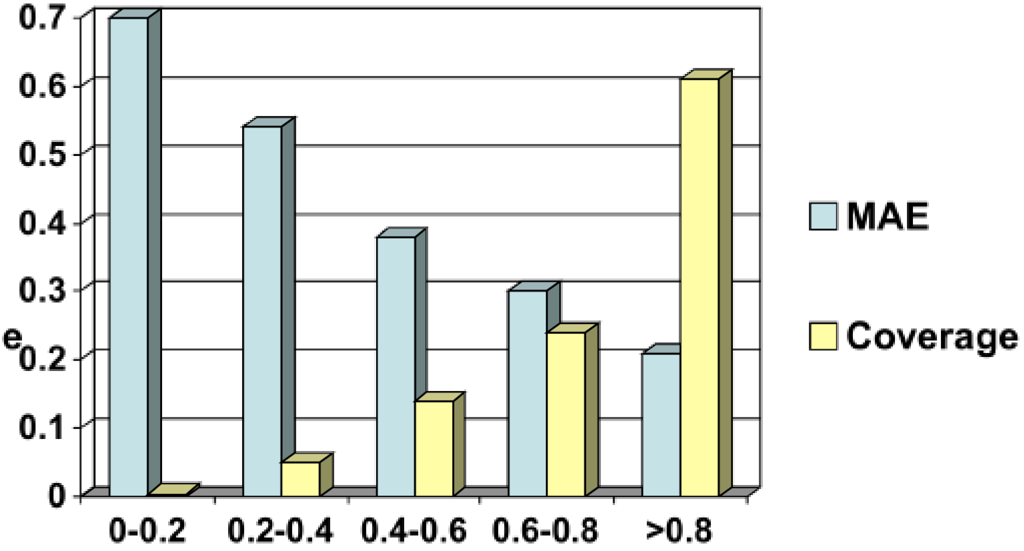
Blind prediction



max correlation coefficient
of a test compound to training
set compounds

MAE=0.43

LIBRARY mode

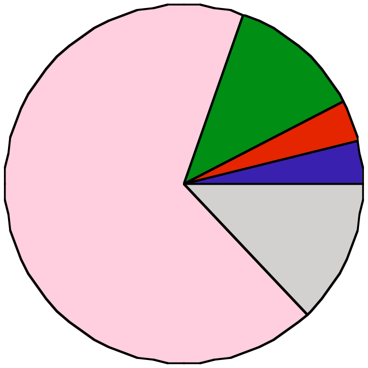


max correlation coefficient
of a test compound to
LIBRARY compounds

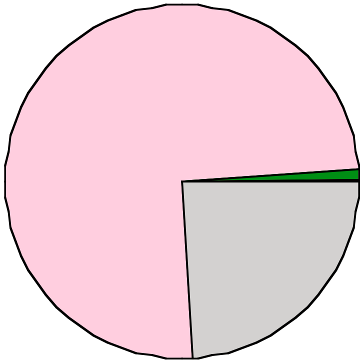
MAE=0.28 (0.26)

Reliability of new compound predictions

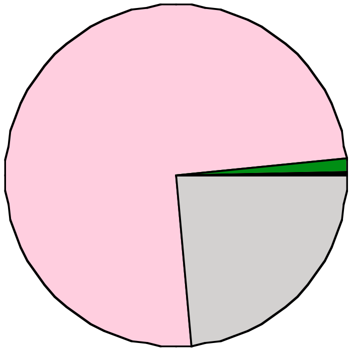
r²		error
0-0.2	⬜	>0.7
0.2-0.4	🟪	~0.6
0.4-0.6	🟩	~0.5
0.6-0.8	🟥	~0.4
0.8-1	🟦	<0.3



NCI,
250,000



http://asinex.com
120,000

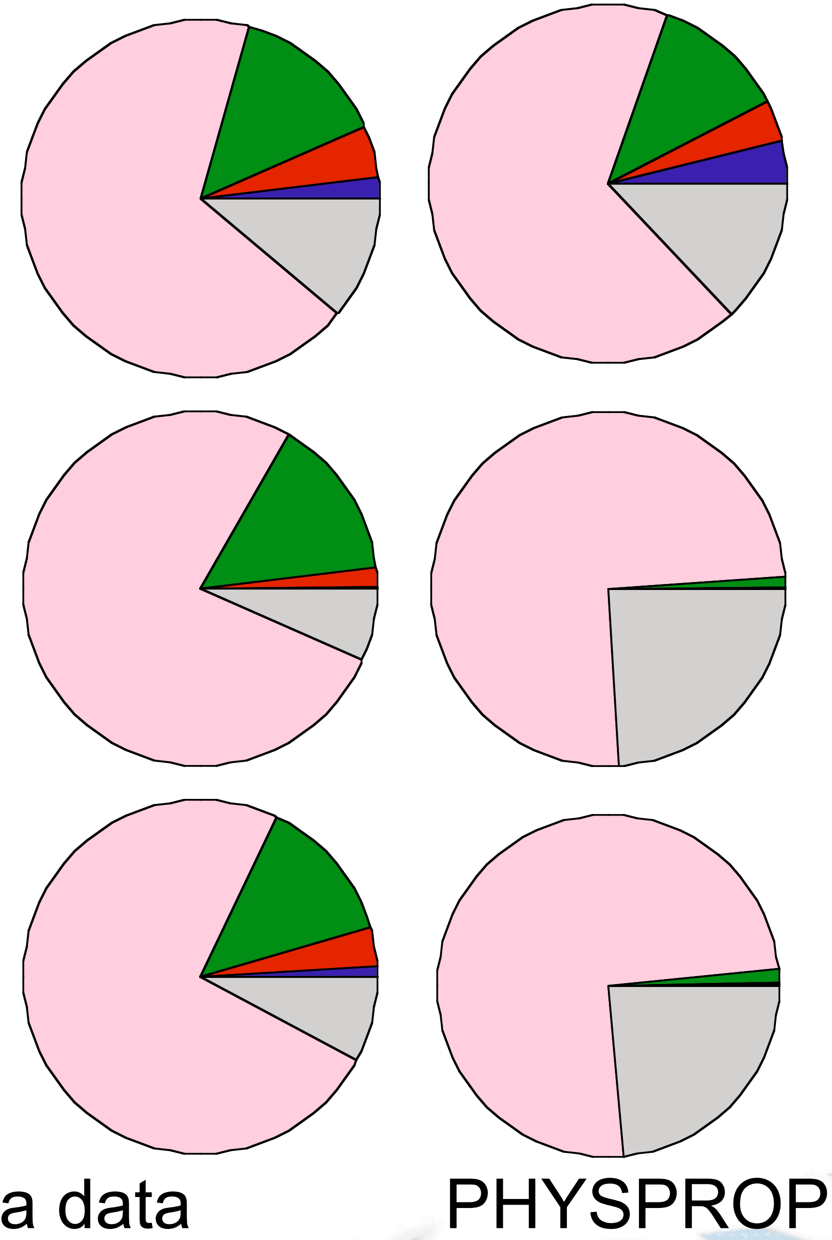


http://ambinter.com
650,000

PHYSPROP

Reliability of new compound predictions

r^2		error
0-0.2	⬜	>0.7
0.2-0.4	🟪	~0.6
0.4-0.6	🟩	~0.5
0.6-0.8	🟥	~0.4
0.8-1	🟦	<0.3

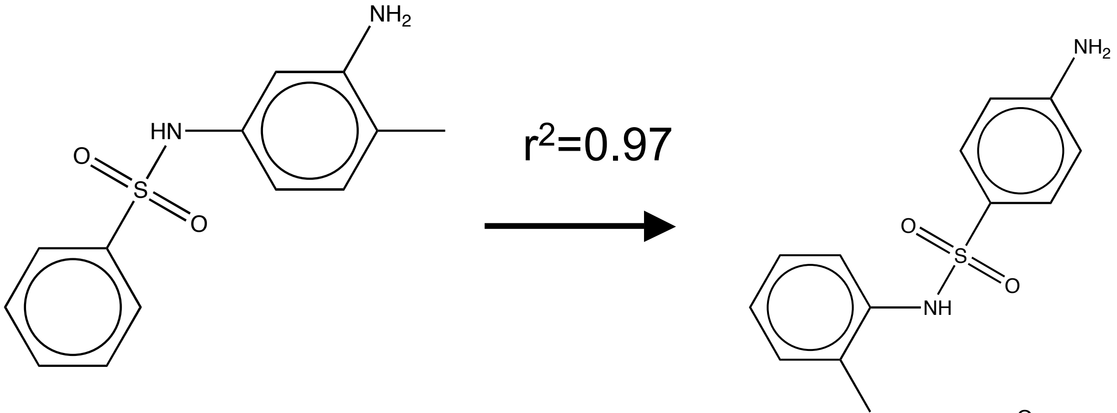


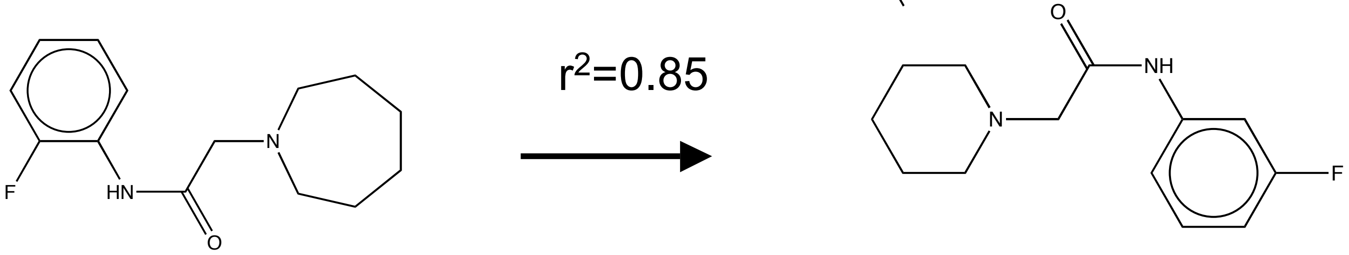
NCI,
250,000

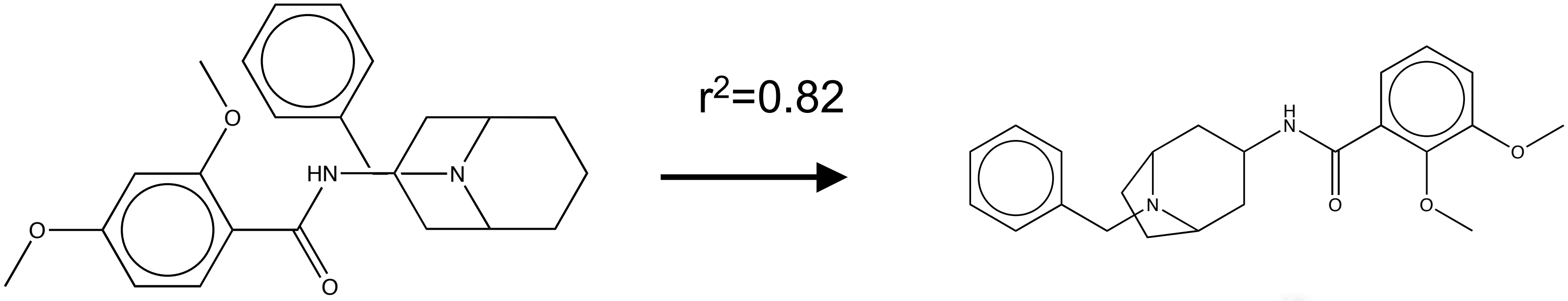
http://asinex.com
120,000

http://ambinter.com
650,000

Is identification possible? PHYSPROP -- Asinex study

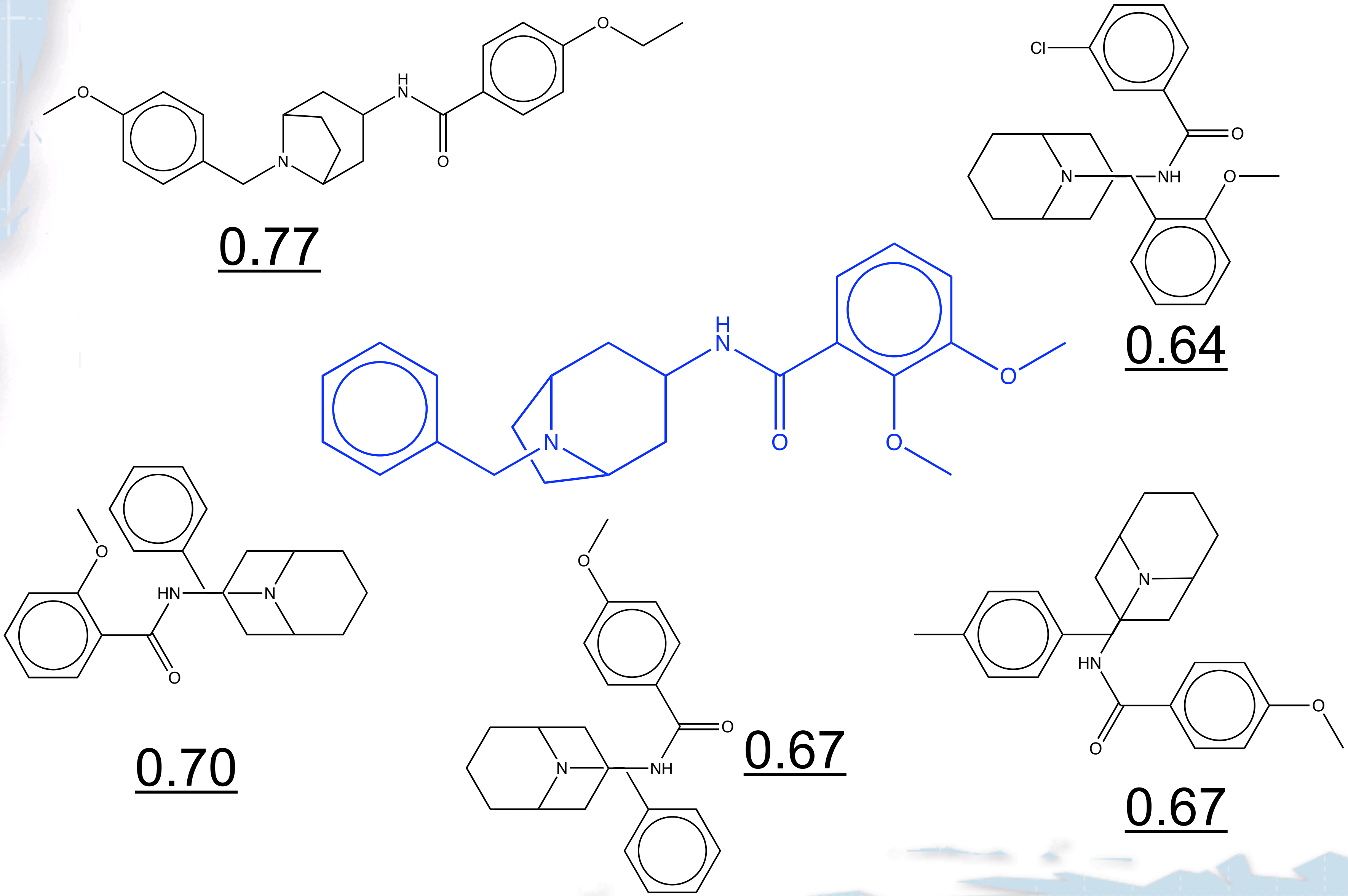




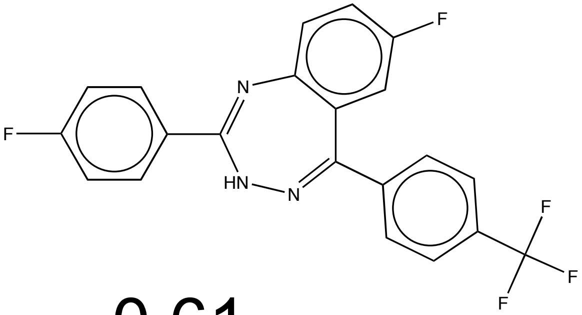


http://vcclab.org

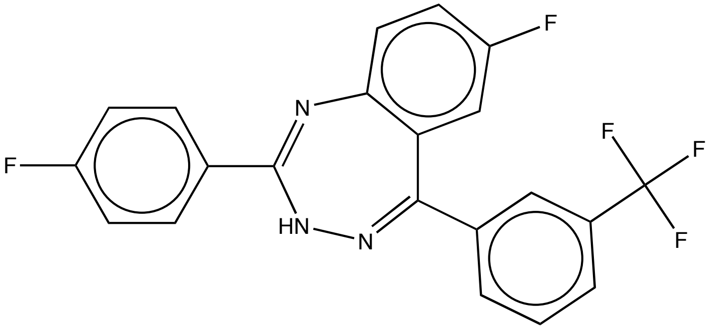
Is identification possible? PHYSPROP -- Asinex study



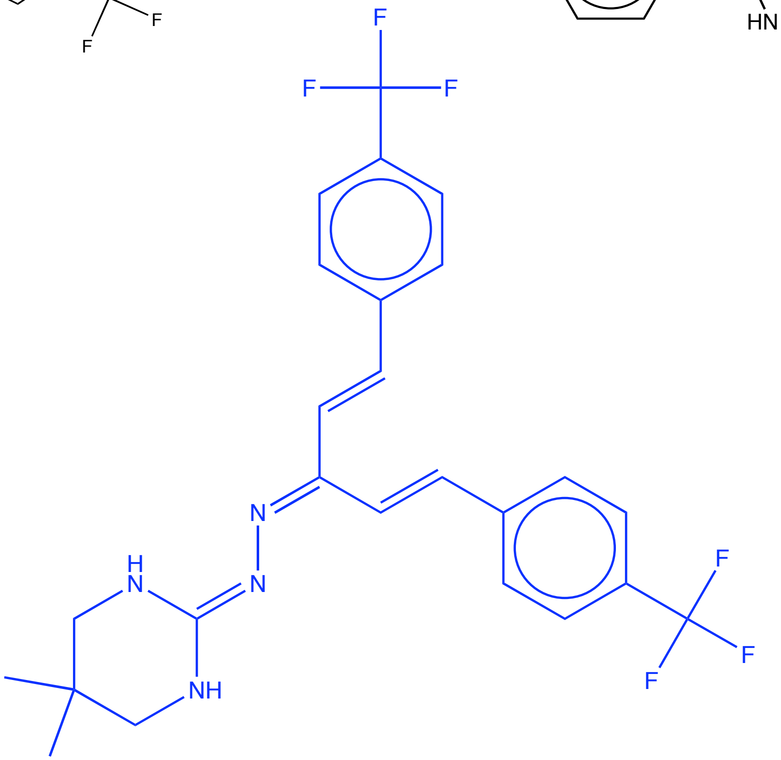
Is identification possible? PHYSPROP -- Asinex study

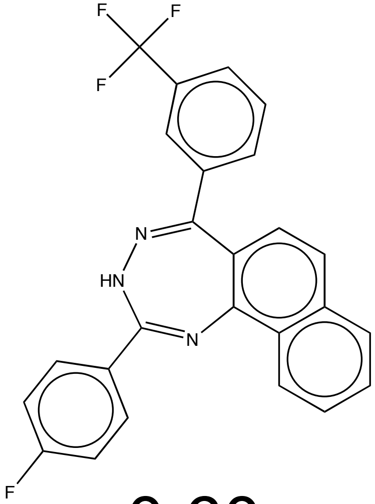


0.61



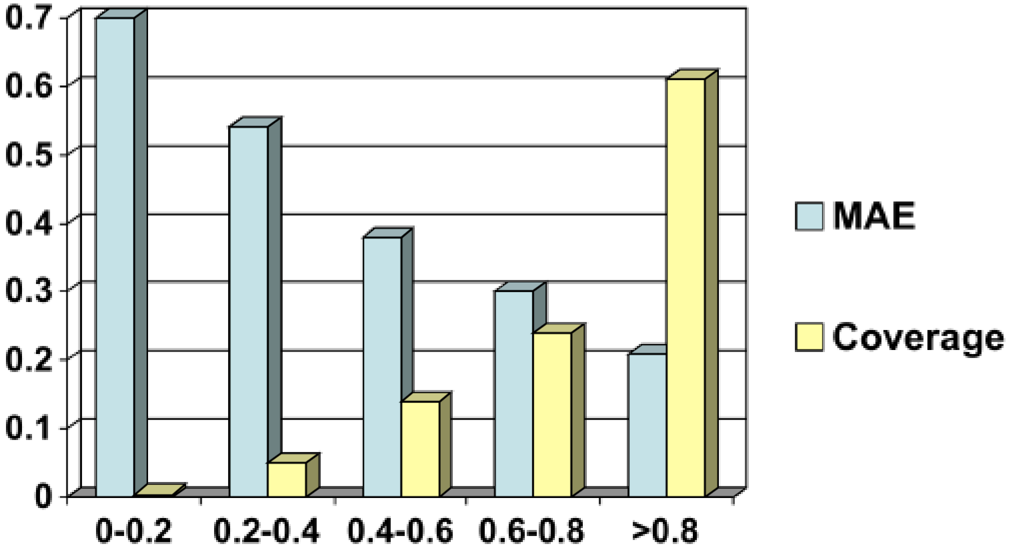
0.60



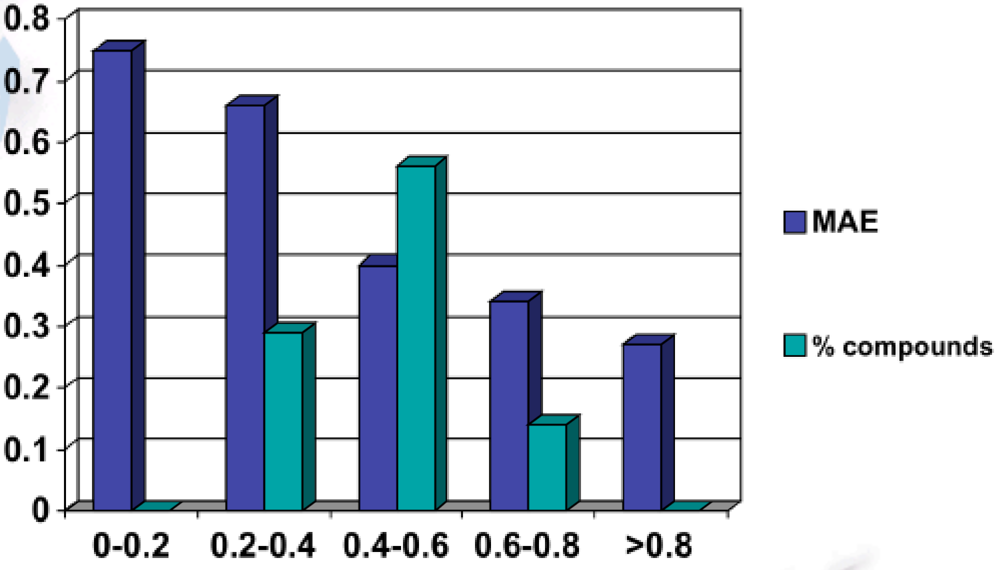


0.60

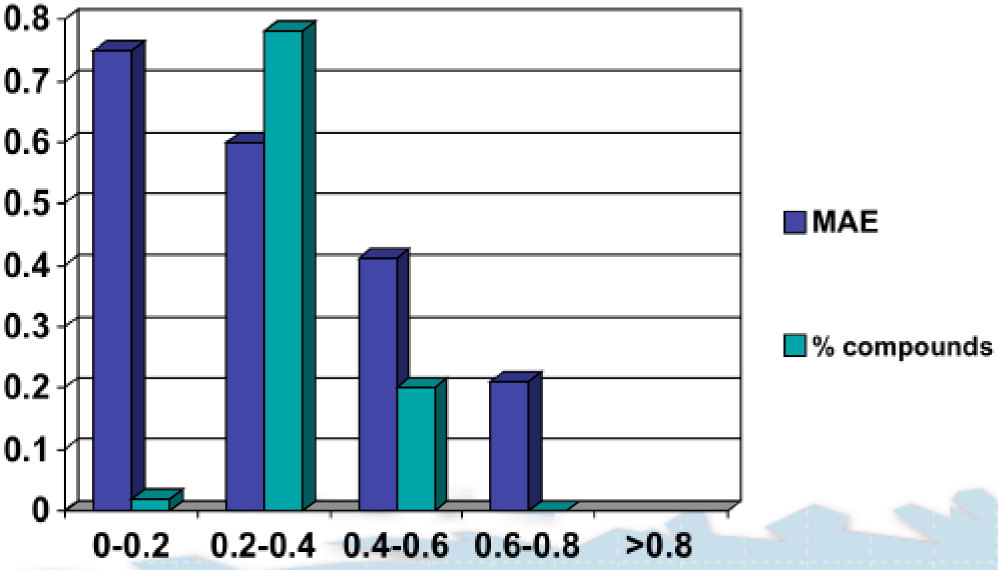
Securing the data -- shuffling ranks!



Shuffle r^2=0.8



Shuffle r^2=0.6



Rank shuffling

- Shuffled rank molecule is less similar to itself than the molecules from the other series wiil be pick-upped --> secure encoding
- Different molecules will have different distribution of neighbors as function of similarity=> lower level of security (e.g. 1 in 10^5, 1 in 10^6) can be determined individually for each single compound using an external library (e.g. complete enumeration, compilation of public libraries)
- Everything can be done in completely automatic mode

Possible approaches

Raw topological indices

- Development of new global models, after the development the data can be discarded
- There is a theoretical possibility to decode the structure, particular for smaller number of atoms in a molecule (not clear if such algorithm can be realized)
- One-to-one contract may be required…

Rank of models

- Allows to incorporate explicit structural parameters as feature elements
- No limitation on the number of indices
- The quality of local correction is comparable to retraining
- Very appealing to share on the WWW
- Security can be controlled by shuffling but will deteriorate prediction quality of model

Development of new models

- Develop new models in-house
- Provide them to be included in the set of models
- Predict new data using an ensemble of diverse models (ASNN in space of models of different companies)
- A complete set of automated tools to develop them can be provided

Acknowledgement

Part of this presentation was done thanks to Virtual Computational Chemistry Laboratory INTAS-INFO 00-0363 project (http://www.vcclab.org).

I thank Prof Hugo Kubinyi, Drs Pierre Bruneau and Gennadiy Poda for collaboration and Prof. Tudor Oprea for inviting me to participate in this conference.

Thank you for your attention!