HelmholtzZentrum münchen
German Research Center for Environmental Health

Critical assessment of QSAR models to predict environmental toxicity against *T. pyriformis*: Applicability domain and overfitting by variable selection

I.V. Tetko,[1] I. Sushko,[1] A.K. Pandey,[1] H. Zhu,[2] A. Tropsha,[2] E. Papa,[3] T. Öberg,[4] R. Todeschini,[5] D. Fourches,[6] A. Varnek[6]

[1]Helmholtz Zentrum München (Germany), [2]University of North Carolina (USA), [3]University of Insubria (Italy), [4]University of Kalmar (Sweden), [5]University of Milano-Bicocca (Italy) and [6]Louis Pasteur University (France)

Neuherberg, 13 October 2008

HELMHOLTZ ASSOCIATION

REACH

Registration, **E**valuation, **A**uthorisation and Restriction of **Ch**emical substances

REACH Guidance
Registration, Evaluation, Authorisation and restriction of CHemicals

European **Ch**emicals **A**gency (ECHA) in Helsinki

ECHA

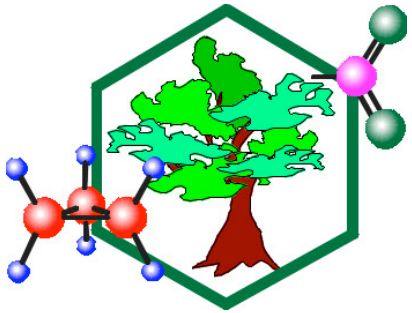
REACH and QSAR (Quantitative Structure Activity Relationship) models

> 30,000 chemicals to be registered … is a lot!

It is expensive to measure all of them ($200,000 per compound), a lot of animal testing

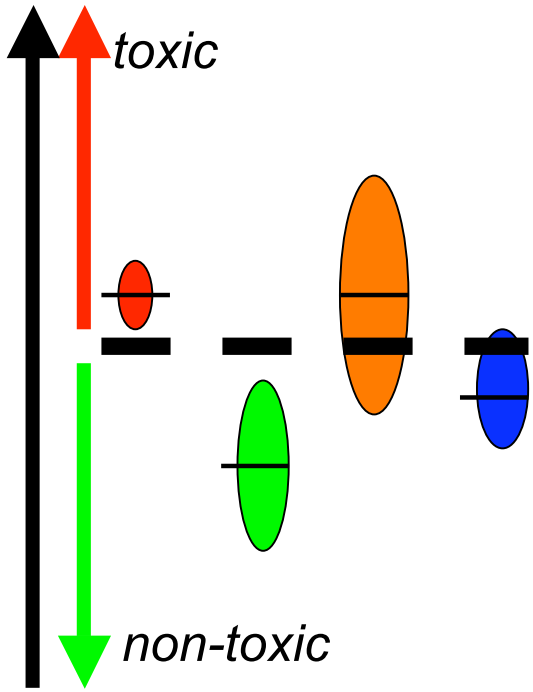
QSAR models can be used to prioritize compounds

- Compound is predicted to be toxic
  - Biological testing will be done to prove/ disprove the models
- Compound is predicted to be not toxic
  - tests can be avoided, saving money, animals
  - but ... only if we are confident in the predictions



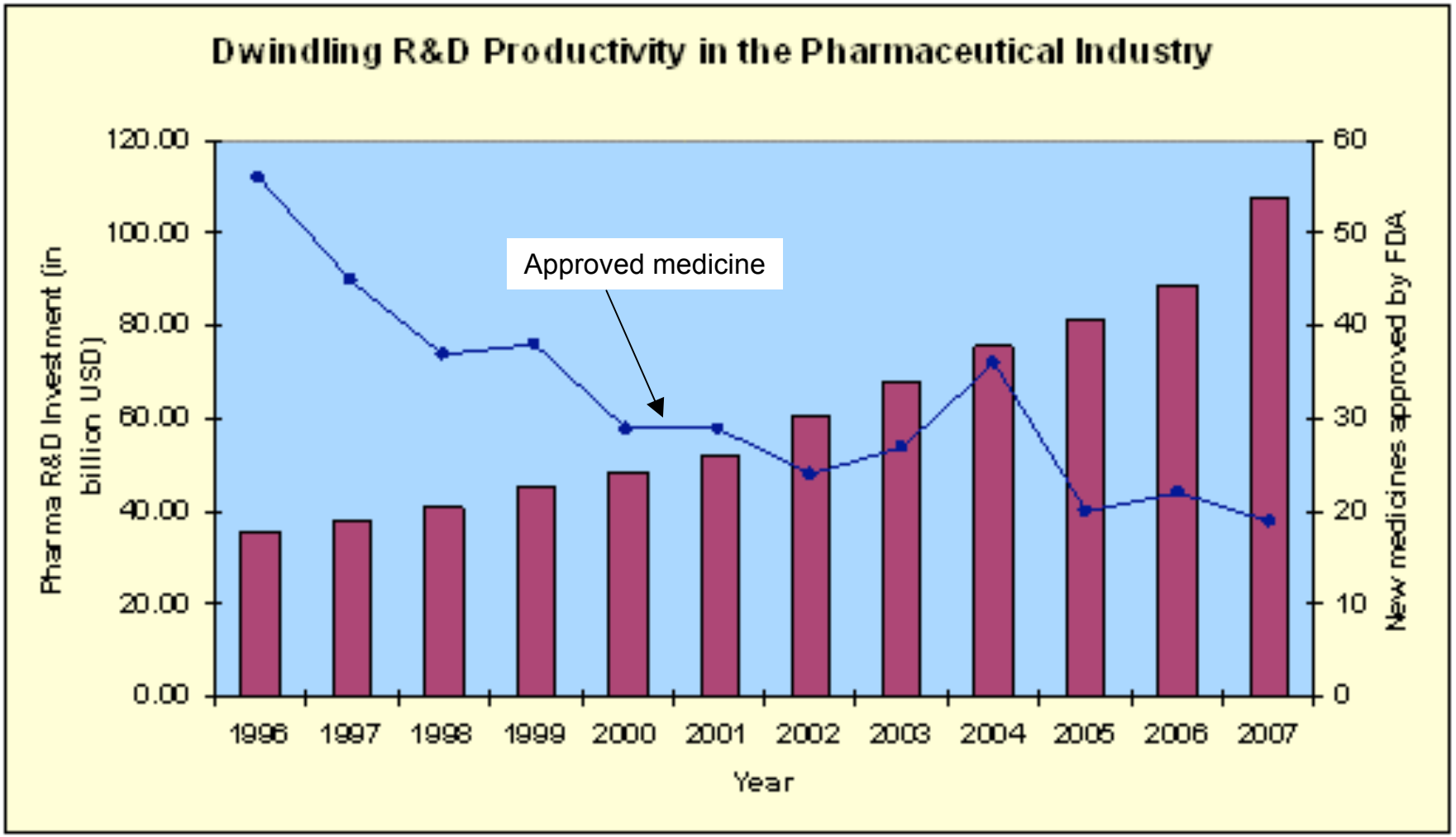
Requirements of biological testing following QSAR model prediction

model prediction	prediction confidence	
	high	low
toxic, IC50 > LIMIT	**strong need**	**moderate need (depends on other properties)**
non-toxic, IC50 < LIMIT	**no need**	**low need (depends on other properties)**



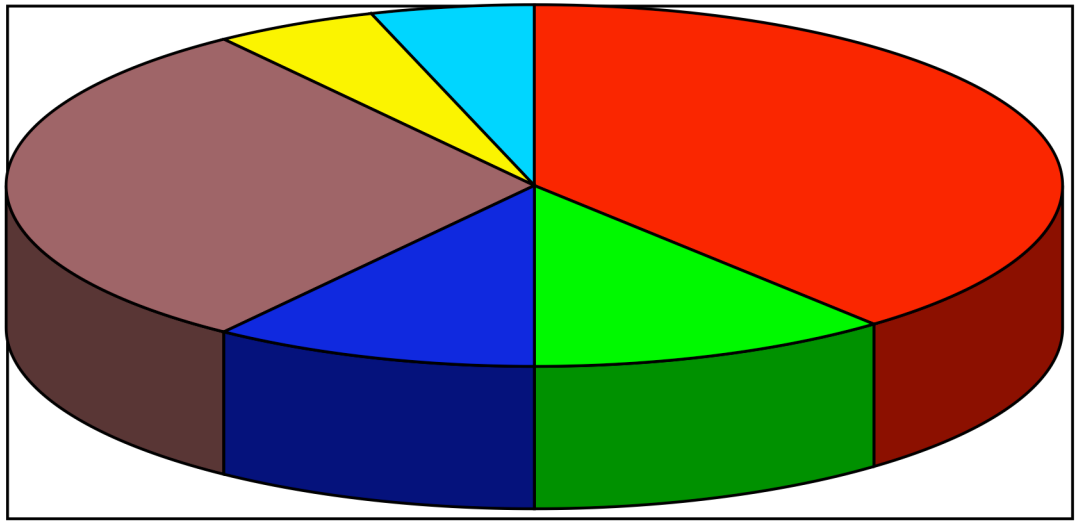
Acceptance of decisions will be more accurate if confidence intervals (prediction errors) are known and are taken into analysis: concept of applicability domain.

Declining R&D productivity in the pharmaceutical industry



Source : PhRMA 2007, FDA

Reasons for failure in drug development



- Pharmacokinetics
- Animal toxicity
- Adverse effects
- Lack of efficacy
- Commercial reasons
- Miscellaneous

} ADME/T

> 60% of drug failures are due to absorption, distribution, metabolism, excretion and **toxicology** (ADME/T) problems

"One can not embrace the unembraceable."

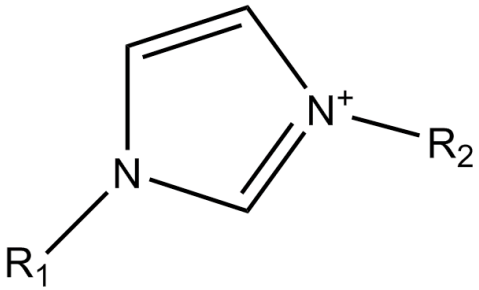
Possible: 10^{60} - 10^{100} molecules theoretically exist
(> 10^{80} atoms in the Universe)

Achievable: 10^{20} - 10^{24} can be synthesized now
(weight of the Moon is ca 10^{23} kg)

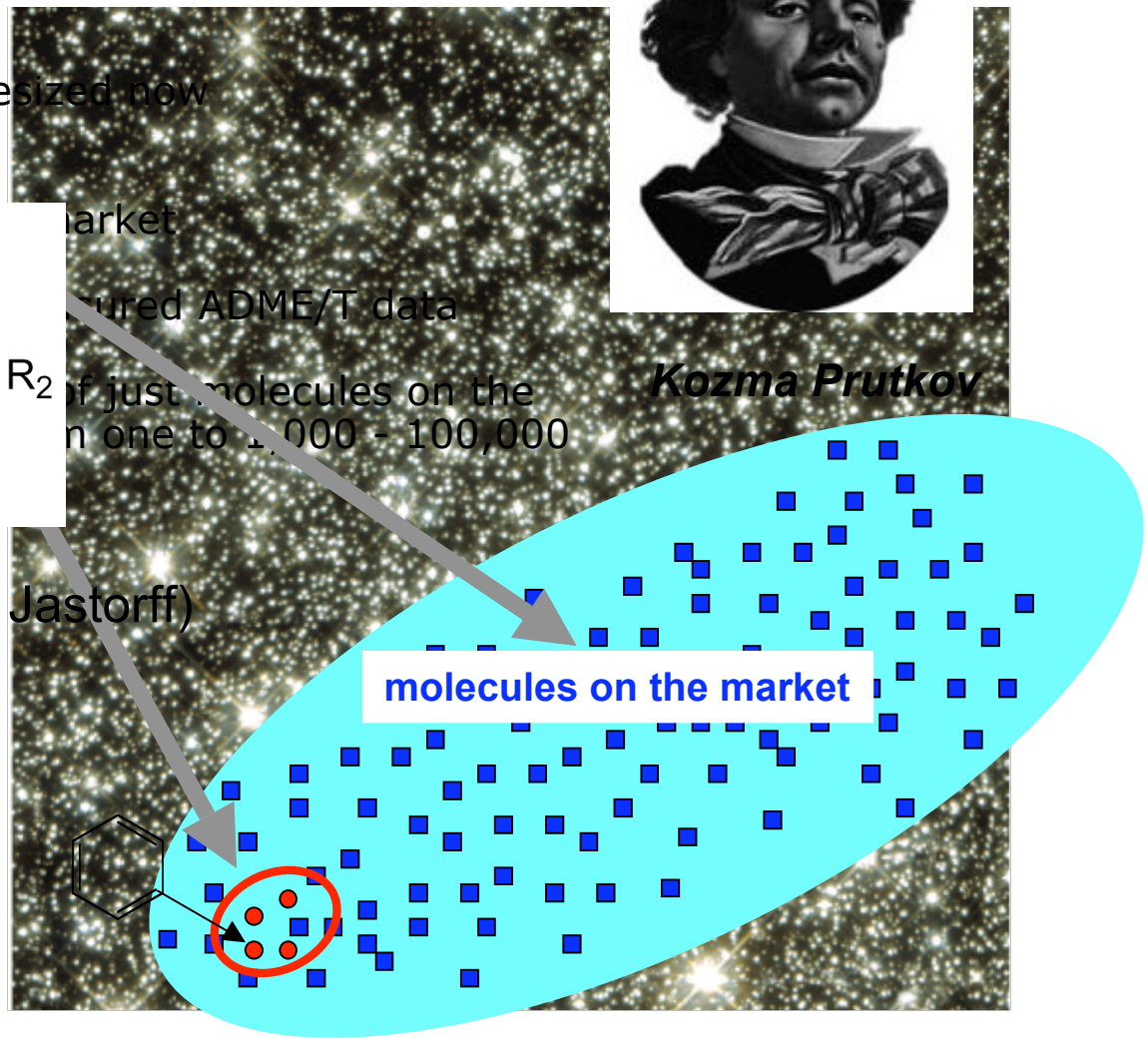
Available: 2*10 [illegible] market

Measured: 10^{2} - [illegible] ured ADME/T data

Problem: To pre [illegible] f just molecules on the market we m [illegible] n one to 1,000 - 100,000 molecules!



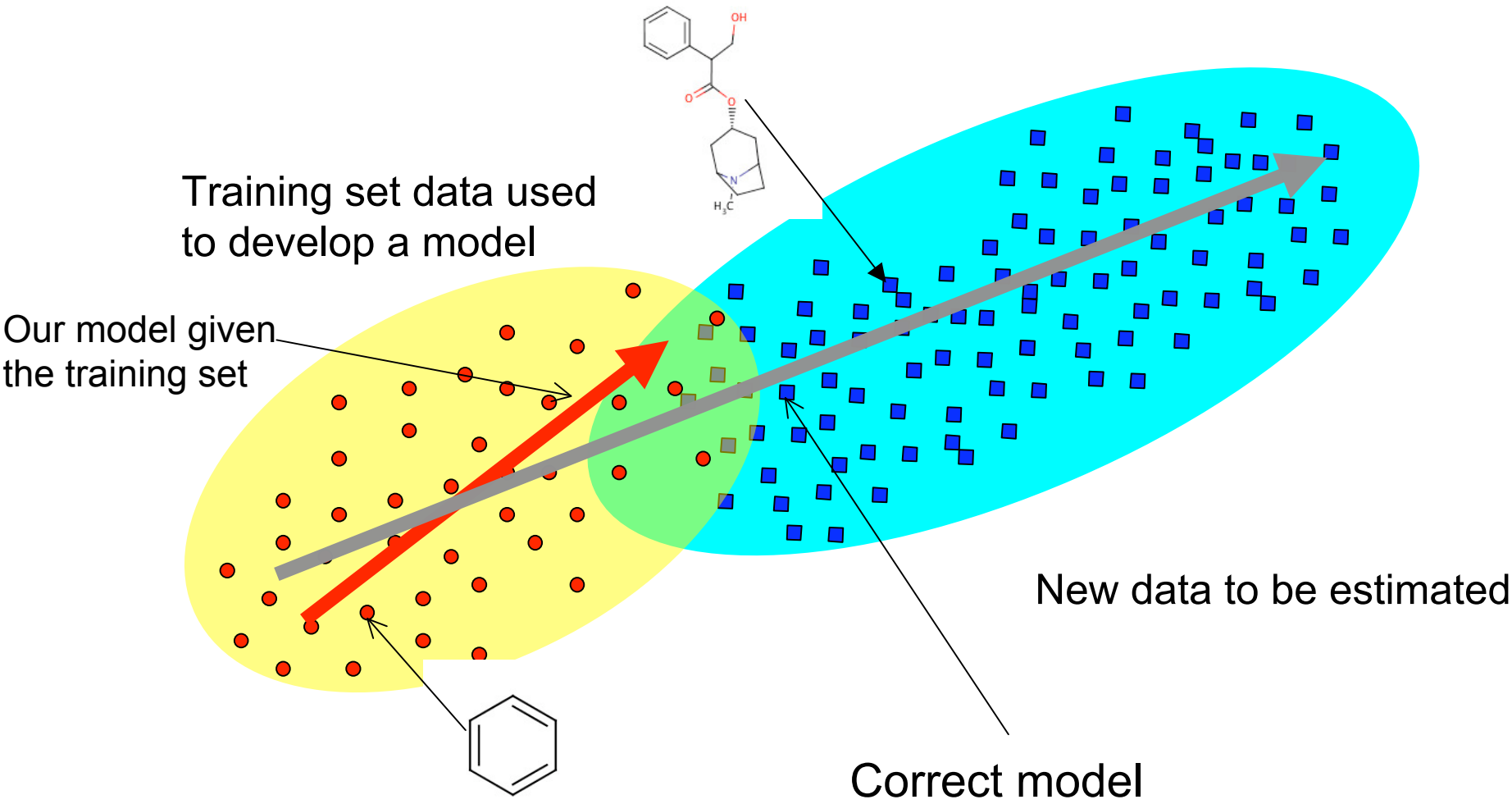
Ionic Liquids ca 10^{18} (Prof. Jastorff)



Methods that can estimate the accuracy of predictions are required.

Both environmental & health sciences have similar problems!

Models can fail due to chemical diversity of training & test sets



It is easy to build a QSAR model

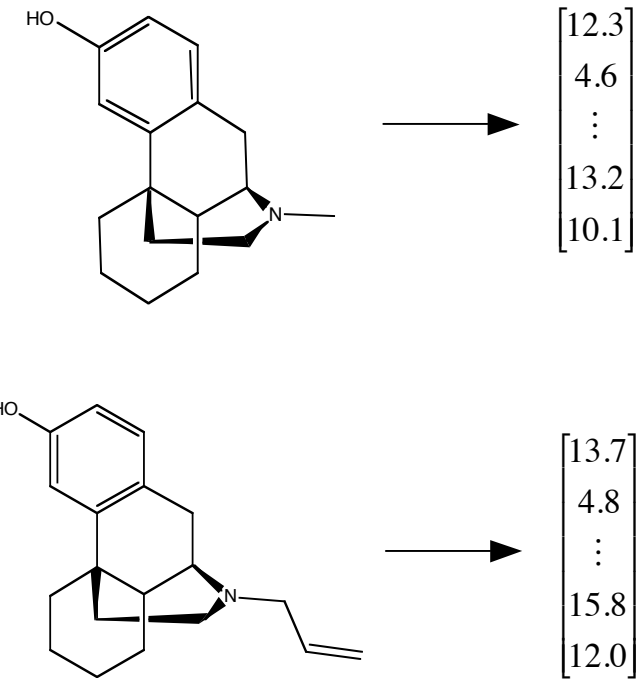
but it is much more difficult to estimate its accuracy for new data

Representation of Molecules for Quantitative Structure-Activity Relationship (QSAR)

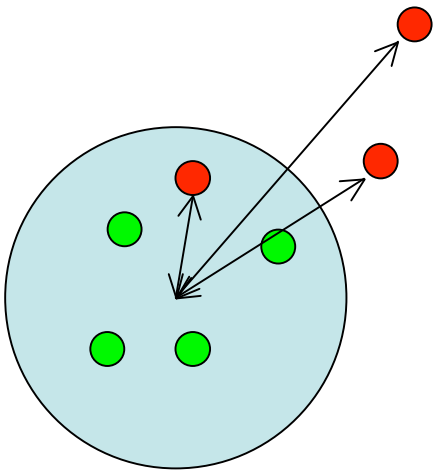
Can be defined with calculated properties (logP, quantum-chemical parameters, etc.)

Can be defined with a set of structural descriptors (topological 2D, 3D, etc.).

One of these sets of descriptors is usually used for determination the applicability domain of models.



Distance to model:



Goals of this study

- Develop new models for prediction of environmental toxicity against *T. pyriformis*
- Benchmark different applicability domains (distances to models)
- Is accuracy of predictions limited by the approach or by the data themselves?
- Is there a best ("universal") AD?

Estimation toxicity of *T. pyriformis*

Initial Dataset[1,2]

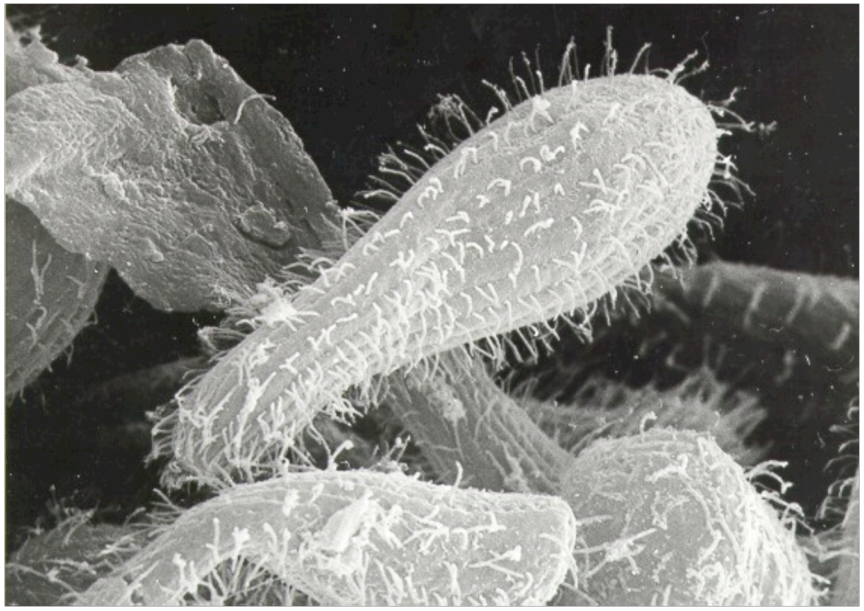
n=983 molecules

n=644 training set

n=339 test set 1

Test set 2:

n=110 molecules[1,2]



The overall goal is to predict (and to assess the reliability of predictions) toxicity against T. pyriformis for chemicals directly from their structure.

[1]Zhu et al, *J. Chem. Inf. Comput. Sci*, **2008**, 48(4), 766-784.

[2]Schultz et al, *QSAR Comb Sci, **2007**, 26(2), 238-254.*

Overview of analyzed QSAR approaches and distances to models

country	modeling techniques	descriptors	abbreviation	distances to models (in space)	
				descriptors	property-based
(UNC)	ensemble of 192 kNN models	MolconnZ	kNN-MZ	EUCLID	STD
	ensemble of 542 kNN models	Dragon	kNN-DR	EUCLID	STD
	SVM	MolconnZ	SVM-MZ		
	SVM	Dragon	SVM-DR		
(ULP)	SVM	Fragments	SVM-FR		
	kNN	Fragments	kNN-FR	EUCLID, TANIMOTO	
	MLR	Fragments	MLR-FR	EUCLID, TANIMOTO	
	MLR	Molec. properties (CODESSA-Pro)	MLR-COD		
(UI)	OLS	Dragon	OLS-DR	LEVERAGE	
(UK)	PLS	Dragon	PLS-DR	LEVERAGE	PLSEU
(HMGU)	ensemble of 100 neural networks	E-state indices	ASNN-ESTATE		CORREL, STD
	consensus model	-	CONS		STD

Tetko et al, *J Chem Inf Model*, **2008,** 48(9):1733-46.

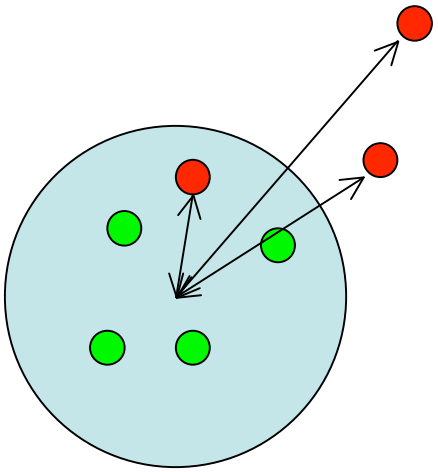
Overview of analyzed distances to models (DMs)

EUCLID $EU_m = \frac{\sum_{j=1}^{k} d_j}{k}$ $EUCLID = \overline{EU}_m$ k is number of nearest neighbors, m index of model	**TANIMOTO** $Tanimoto(a,b) = \frac{\sum x_{a,i} x_{b,i}}{\sum x_{a,i} x_{a,i} + \sum x_{b,i} x_{b,i} - \sum x_{a,i} x_{b,i}}$ $x_{a,i}$ and $x_{b,i}$ are fragment counts
LEVERAGE $LEVERAGE = \mathbf{x}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}$	**PLSEU** (DModX) Error in approximation (restoration) of the vector of input variables from the latent variables and PLS weights.
STD $STD = \frac{1}{N-1}\sum(y_i - \bar{y})^2$ y_i is value calculated with model i and $\bar{y}$ is average value	**CORREL** $CORREL(a) = \max_j CORREL(a,j) = R^2(\mathbf{Y}^a_{calc}, \mathbf{Y}^j_{calc})$ $\mathbf{Y}^a = (y_1, \ldots, y_N)$ is vector of predictions of molecule i

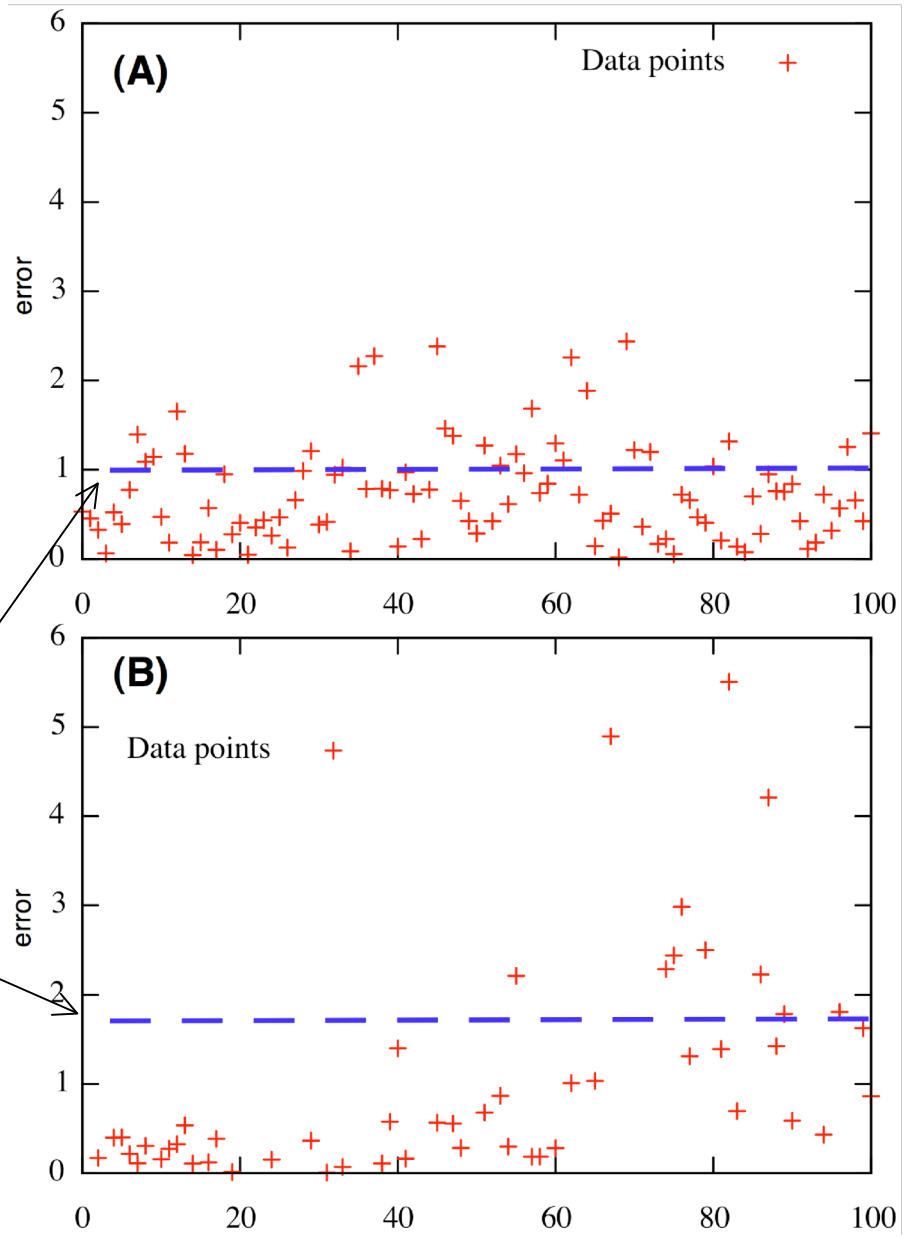
Analysis of two simulated datasets

A) Errors do not depend on the distance to model (DM)

B) Errors depend on the DM



σ

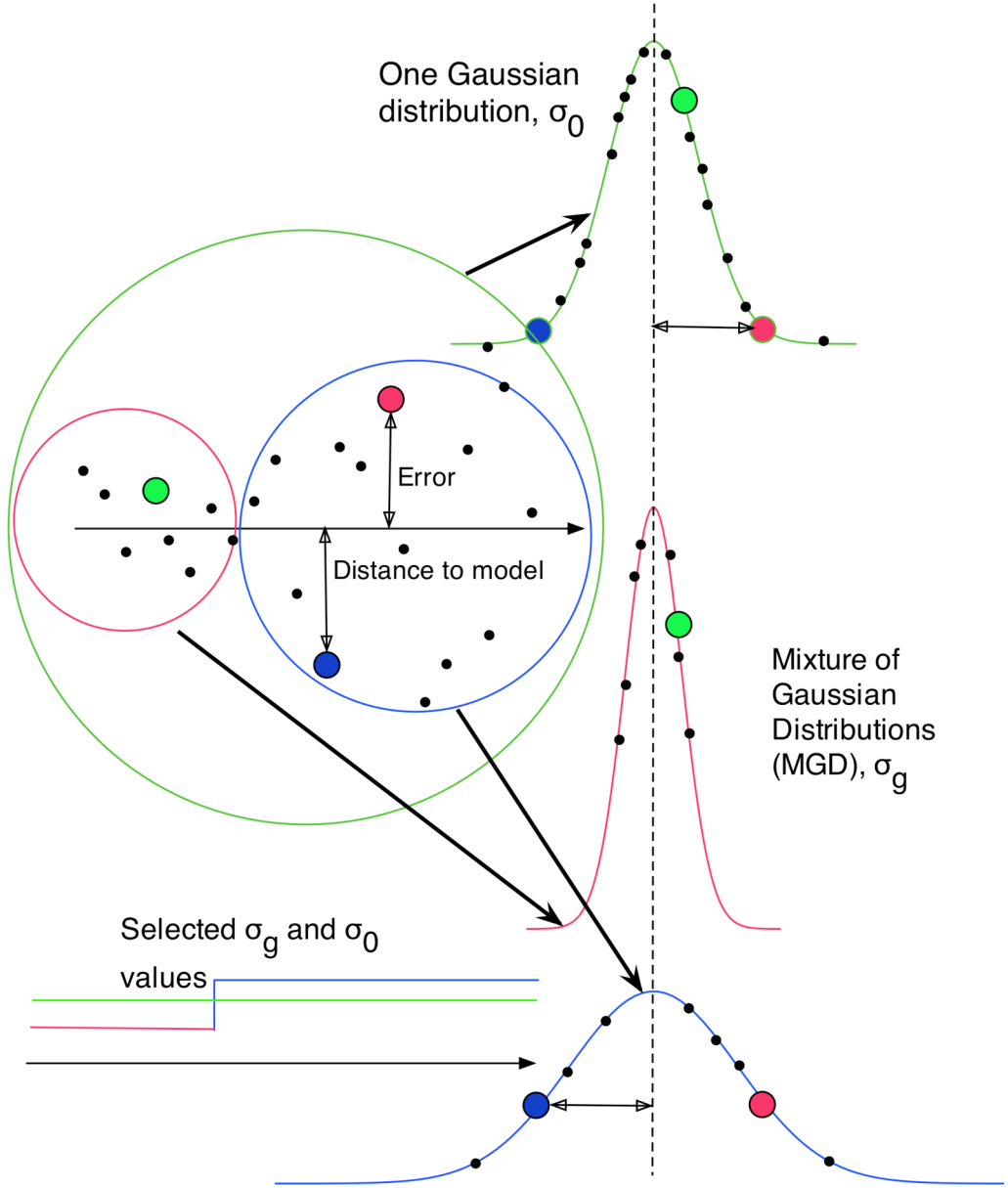


Mixture of Gaussian Distributions (MGD)

Idea is to find a MGD, which maximize likelihood (probability)

$\Pi\, N(0,\sigma^2(e_i))$

of the observed distribution of errors

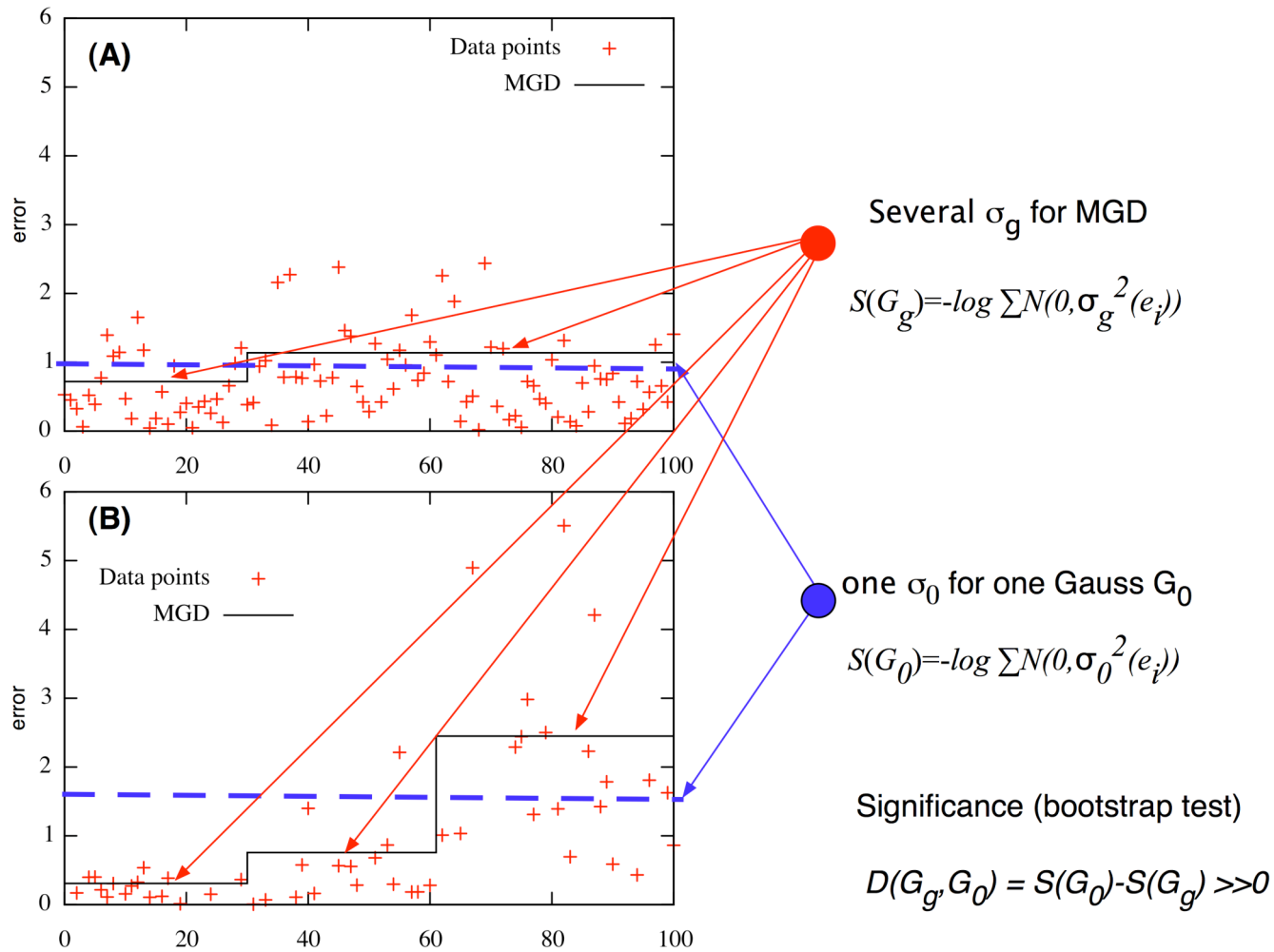


$$N\left(0,\sigma^2(e)\right) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{e^2}{2\sigma^2}\right) \qquad S(G_g)=\Sigma \log N(0,\sigma_g^2(e_i)))$$

MGDs for the simulated datasets

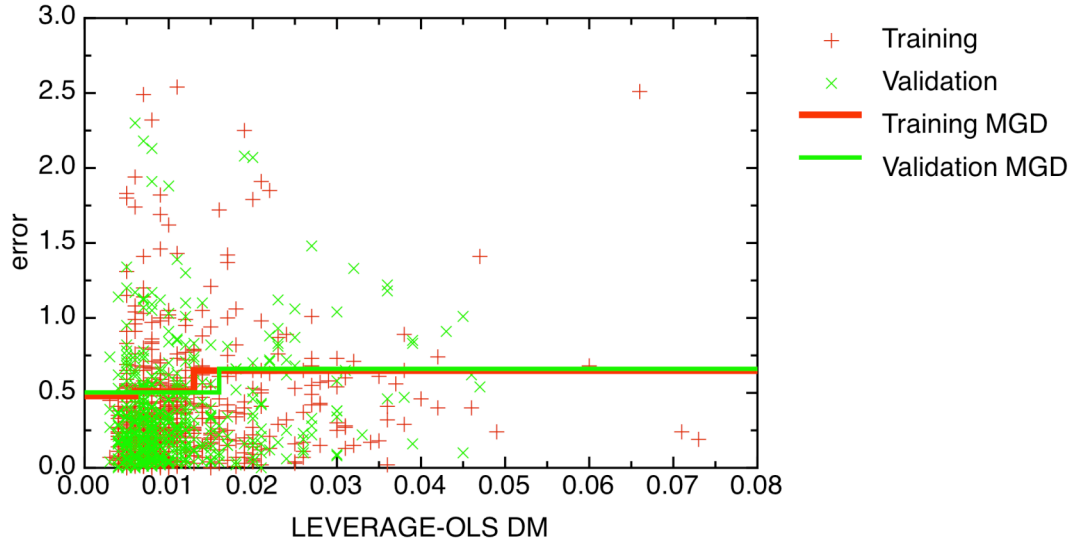
A) Non significant MGD was found

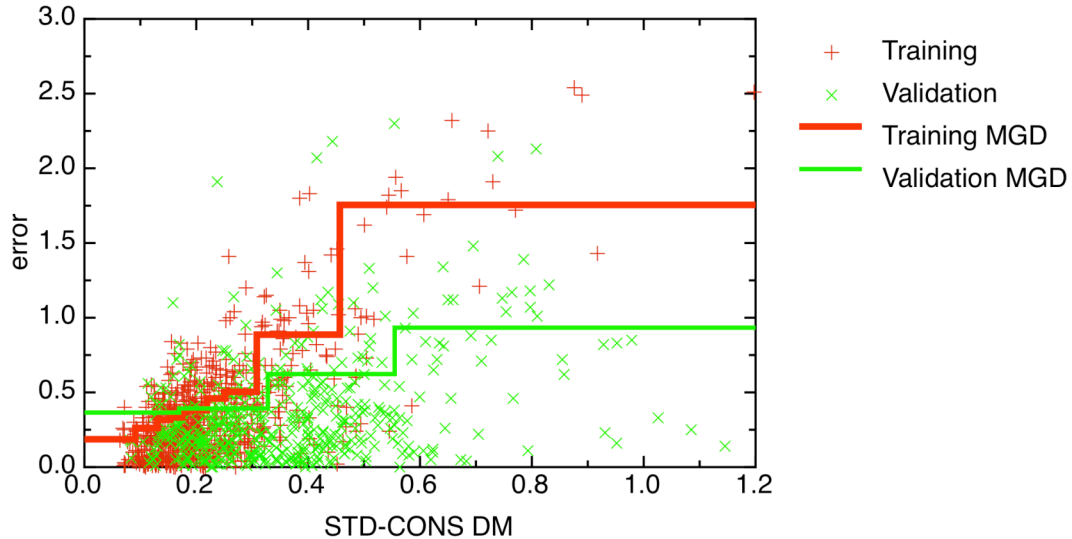
B) A MGD composed of 3 Gaussian distributions was found



Analysis of DMs for a linear model

$$\text{Log}(IGC_{50}^{-1}) = -18(\pm 0.7) + 0.065(\pm 0.002)\mathbf{AMR} - 0.50(0.04)\mathbf{O56} - 0.30(0.03)\mathbf{O58} - 0.29(0.02)\mathbf{nHAcc} + 0.046(0.005)\mathbf{H\text{-}046} + 16(0.7)\mathbf{Me}$$





The use of various DM provides different discrimination of molecules with low and large errors.

Tetko et al, *J Chem Inf Model*, **2008,** 48(9):1733-46.

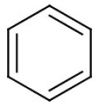
Performances of MGDs calculated with different definitions of Distance to Models (DM)

DM	average rank			highest rank[1]		
	LOO	5-CV	**Valid.***	LOO	5-CV	Valid.
STD-CONS	1	1.8	**1.1**	12	2	11
STD-ASNN	2	1.2	**2.5**		10	1
STD-kNN-DR	6.6	4.3	**4.1**			
STD-kNN-MZ	9.2	8.3	**5.3**			
EUCLID-kNN-DR	7.1	4.9	**5.4**			
LEVERAGE-PLS	8.4	5	**6.3**			
EUCLID-kNN-MZ	7.5	7.1	**6.4**			
TANIMOTO-kNN-FR	7	6.1	**6.8**			
TANIMOTO-MLR-FR	8.3	8.3	**9**			
CORREL-ASNN	10.7	10.8	**9.4**			
LEVERAGE-OLS-DR	12.3	12.6	**11.1**			
EUCLID-MLR-FR	7	9.3	**11.5**			
PLSEU-PLS	11.1	11.8	**11.5**			
EUCLID-kNN-FR	12.1	13.3	**12.1**			

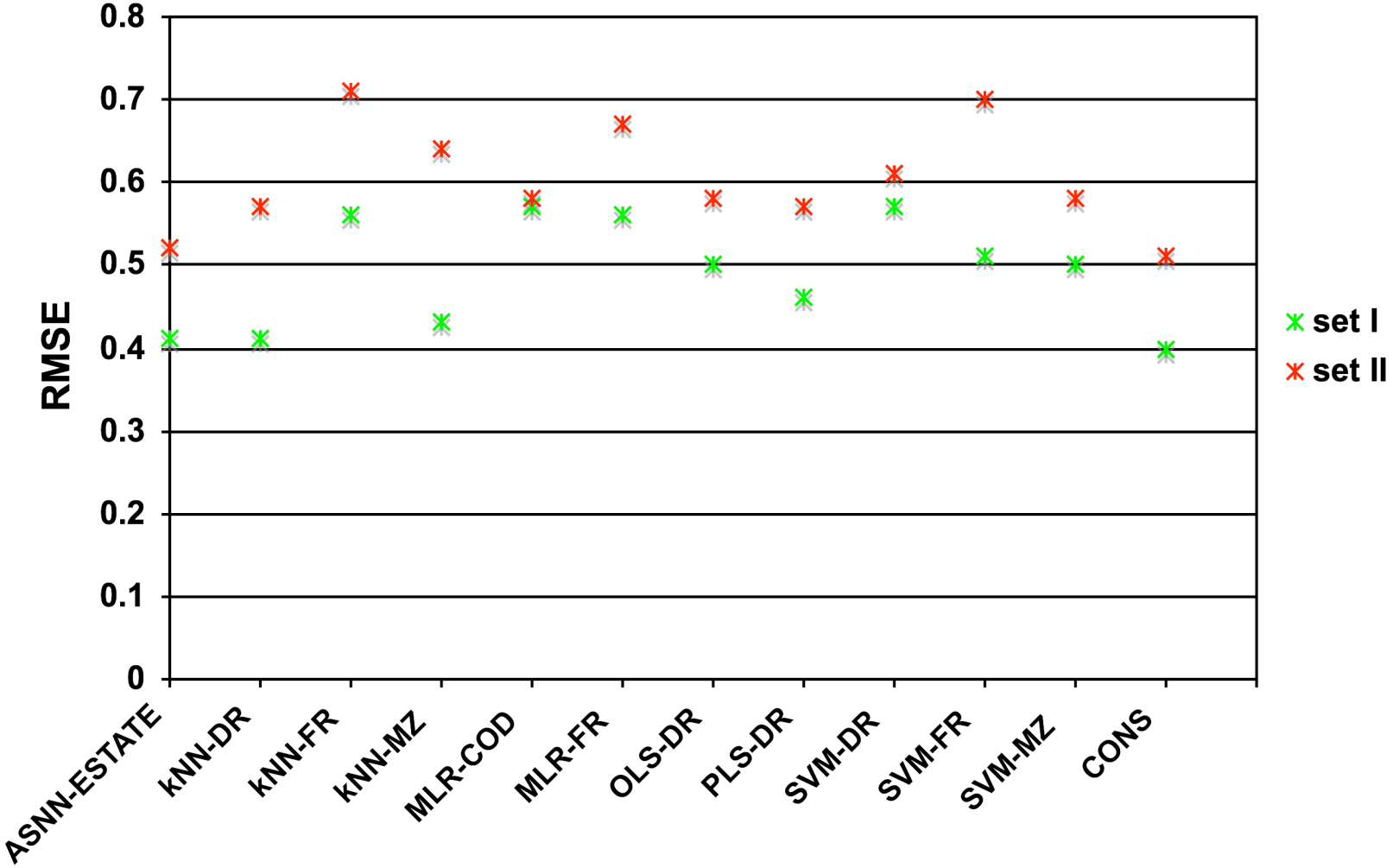
*Ordered by performance of the DMs on the validation dataset

Tetko et al, *J Chem Inf Model*, **2008,** 48(9):1733-46.

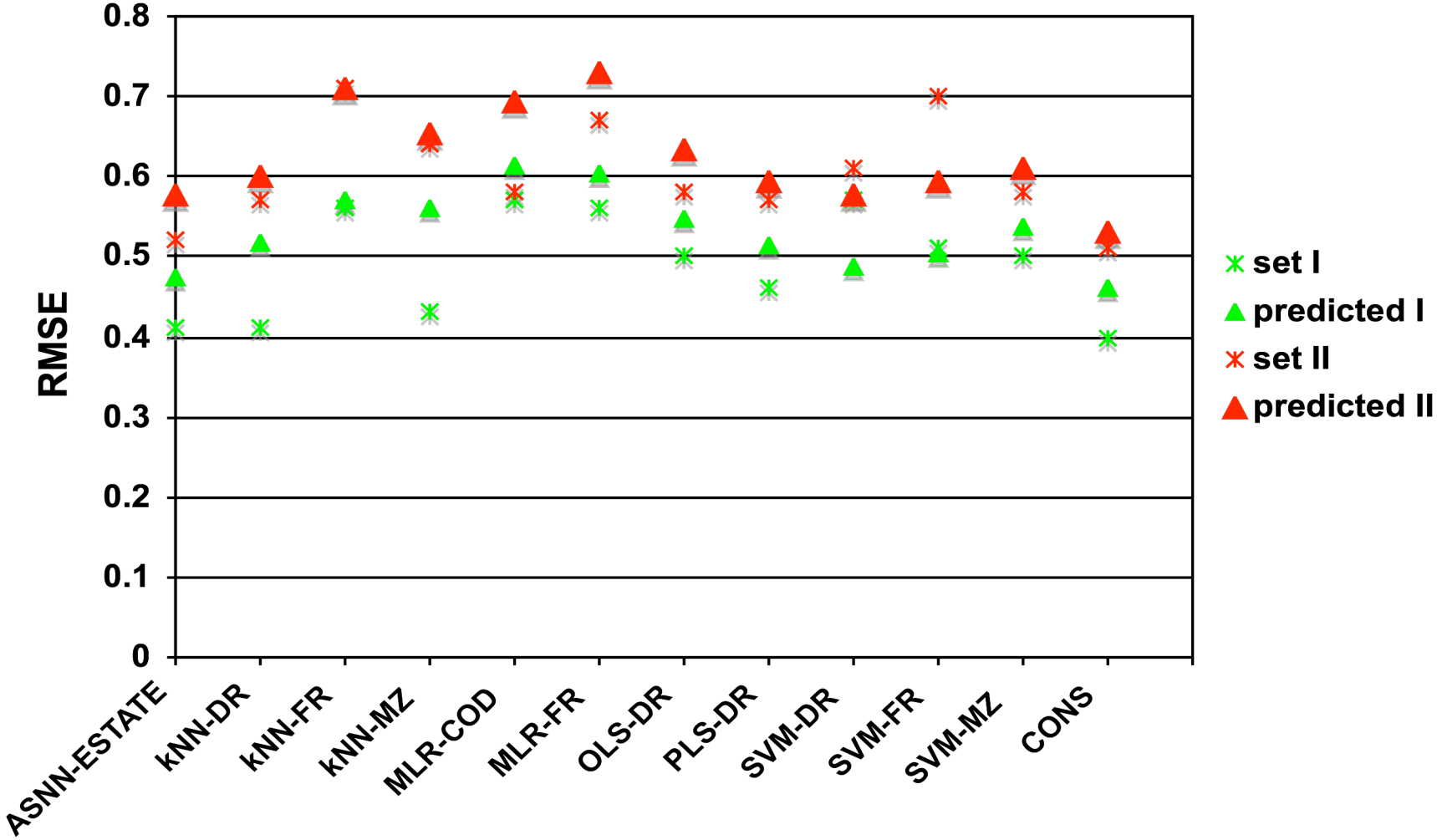
Standard Deviation of Models (STD)

country	modeling techniques	descriptors	abbreviation	
(UNC)	kNN ensemble	MolconnZ	kNN-MZ	1.12
	kNN ensemble	Dragon	kNN-DR	1.02
	SVM	MolconnZ	SVM-MZ	0.97
	SVM	Dragon	SVM-DR	0.91
(ULP)	SVM	Fragments	SVM-FR	0.88
	kNN	Fragments	kNN-FR	0.95
	MLR	Fragments	MLR-FR	0.99
	MLR	CODESSA-Pro	MLR-COD	1.14
(UI)	OLS	Dragon	OLS-DR	1.06
(UK)	PLS	Dragon	PLS-DR	1.08
(HMGU)	neural networks ensemble	E-state indices	ASNN-ESTATE	1.10
consensus (average)			CONS	1.02
STD			STD-CONS	**0.09**

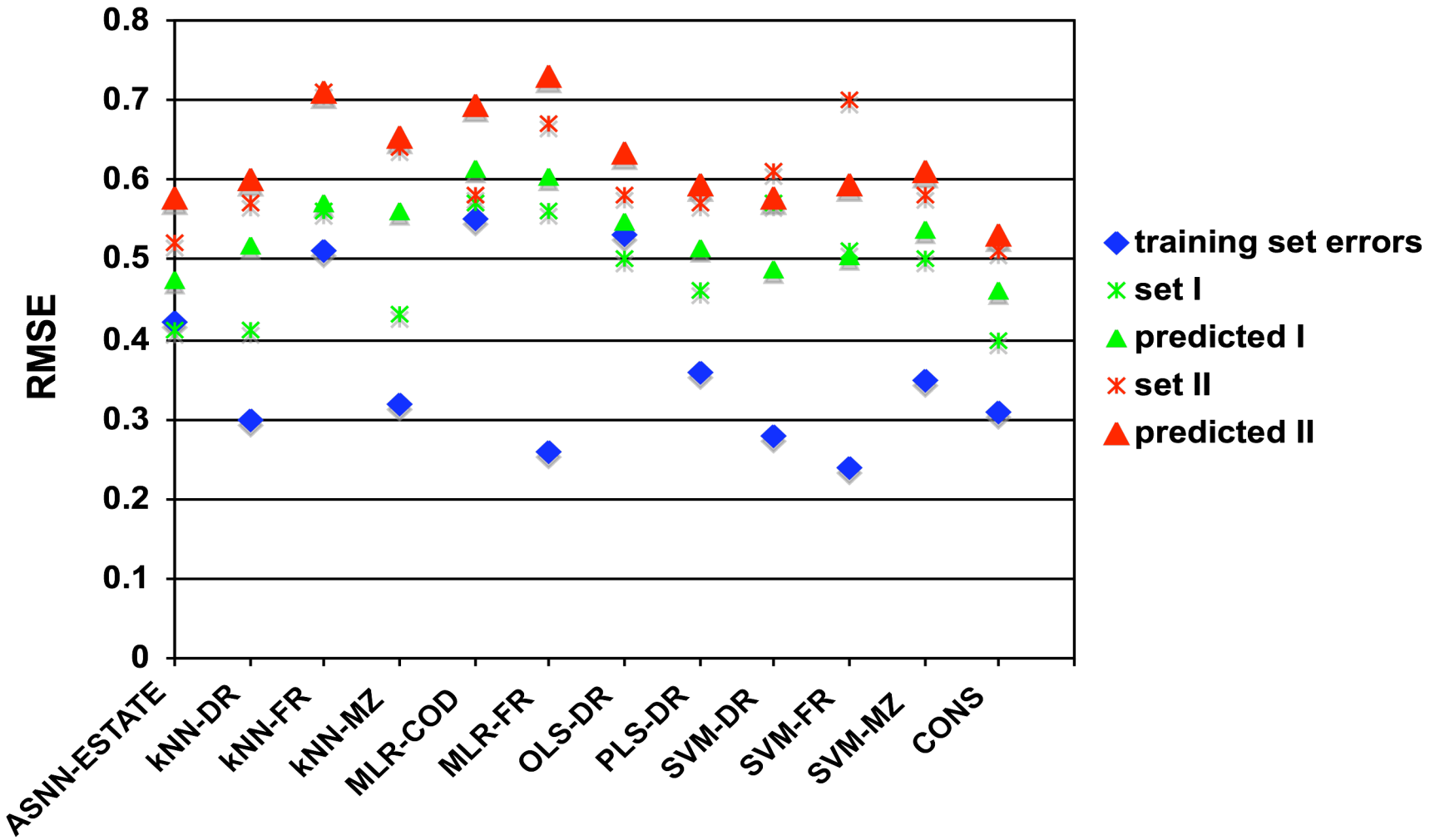
Errors using MGD & STD distance to models



Estimations of errors using STD distance to models



Estimations based on training set errors calculated with incorrect validation protocol

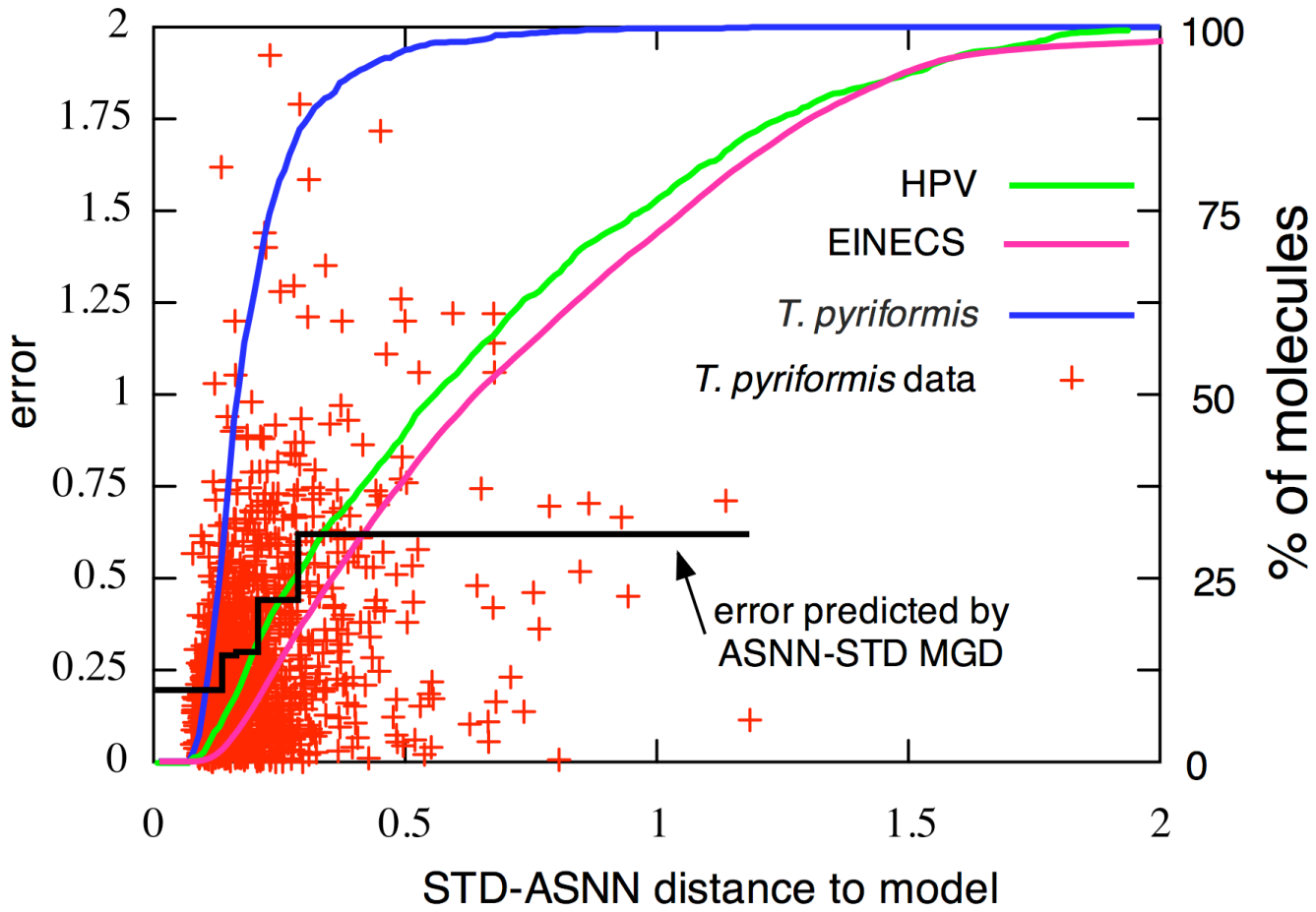


Tetko et al, *J Chem Inf Model*, **2008,** 48(9):1733-46.

Prediction of data from the training and two external database

Experimental accuracy:

Estimated experimental accuracy:[1]
SE = 0.38 reactive
and
SE = 0.21 narcosis mechanism of action



[1]Seward et al, *Aquat Toxicol*. 2001, 53(1), 33–47.

Sustainable or Green Chemistry

Twelve Principles

- Prevent waste
- Design safer chemicals and products
- Use renewable feedstocks
- Use catalysts, not stoichiometric reagents
- Avoid chemical derivatives
- Maximize atom economy
- Use safer solvents and reaction conditions
- Increase energy efficiency
- Design chemical and products to degrade after use
- Analyze in real time to prevent pollution
- Minimize the potential for accidents

QSAR for Sustainable or Green Chemistry

Twelve Principles

- Prevent waste
- ✓ **Design safer chemicals and products**
- Use renewable feedstocks
- Use catalysts, not stoichiometric reagents
- Avoid chemical derivatives
- Maximize atom economy
- Use safer solvents and reaction conditions
- Increase energy efficiency
- ✓ **Design chemical and products to degrade after use**
- Analyze in real time to prevent pollution
- Minimize the potential for accidents

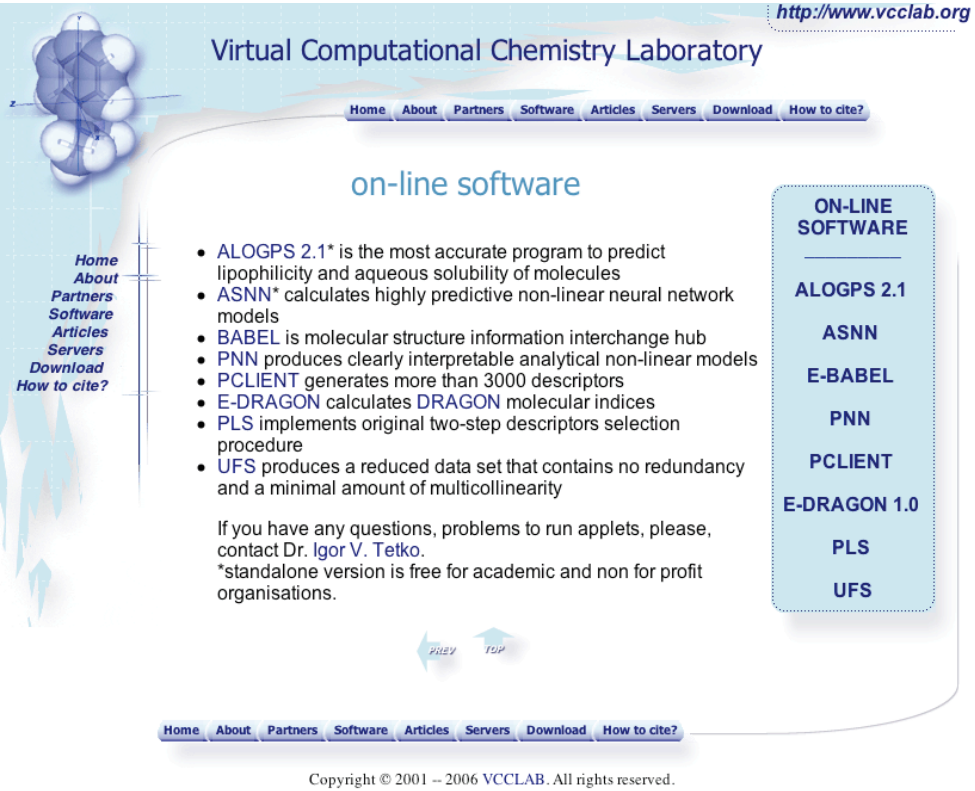
Conclusions

- Development of green chemistry (environmental sciences) and discovery of drugs (health sciences) share similar problems
- The use of QSAR approaches can help to identify toxic/non-toxic compounds before start of their commercial exploitation in chemical industry or clinical testing in the drug discovery

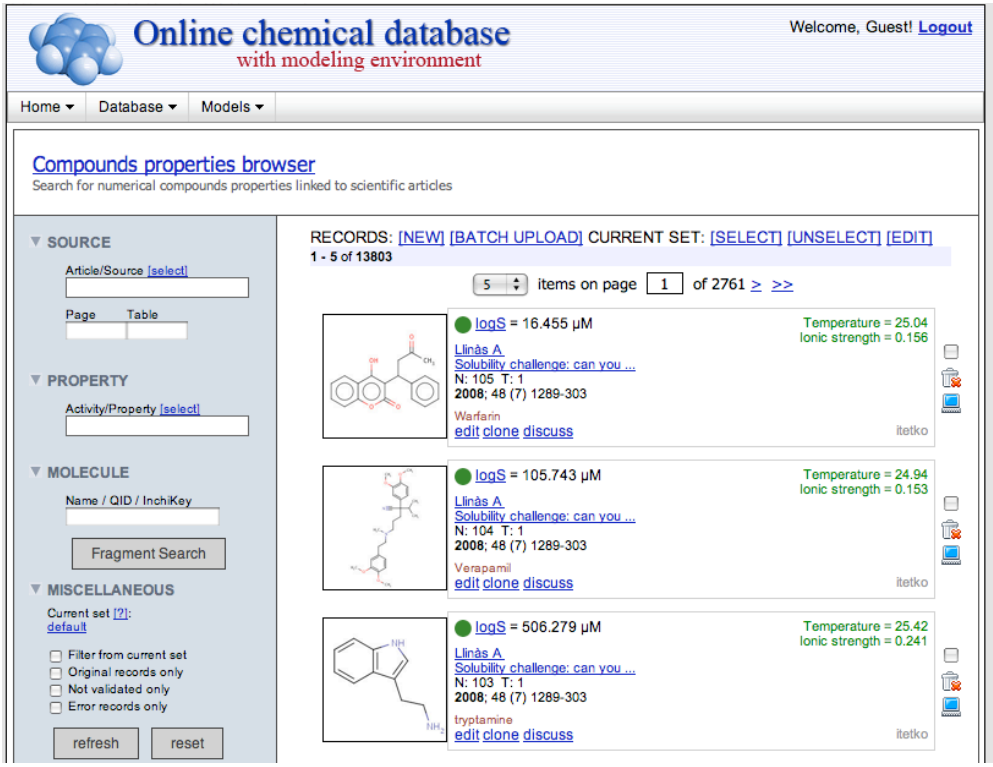
- Data (diversity, accuracy) but not the methods dominate in determination of the accuracy of model predictions
- The standard deviation of models provided the best discrimination of molecules with low and high prediction accuracy
- Models are available at http://www.qspr.org (in development)
- Models can reliably predict only small % of molecules from the REACH-like database

Do you need more information?

http://www.vcclab.org



http://www.qspr.eu*



Tetko et al, *J Chem Inf Model*, **2008,** 48(9):1733-46.

**to be available later this year.*

Acknowledgements

All collaborators

- Ester Papa
- Tomas Öberg
- Roberto Todeschini
- Alexander Tropsha & Hao Zhu
- Alexandre Varnek & Denis Fourches

Our team

- Iurii Sushko
- Robert Koerner
- Sergii Novatarskyi
- Anil Kumar Pandey

Paola Gramatica

Mark Hewitt

Mark Cronin

Terry Schultz

Johann Gasteiger & Molecular Networks GmbH

This study has been partially supported with BMBF GoBio and FP7 CADASTER projects.

Thank you very much for your attention!