HelmholtzZentrum münchen
German Research Center for Environmental Health

Speeding-up Drug Development with Confident Predictions of ADME/T properties

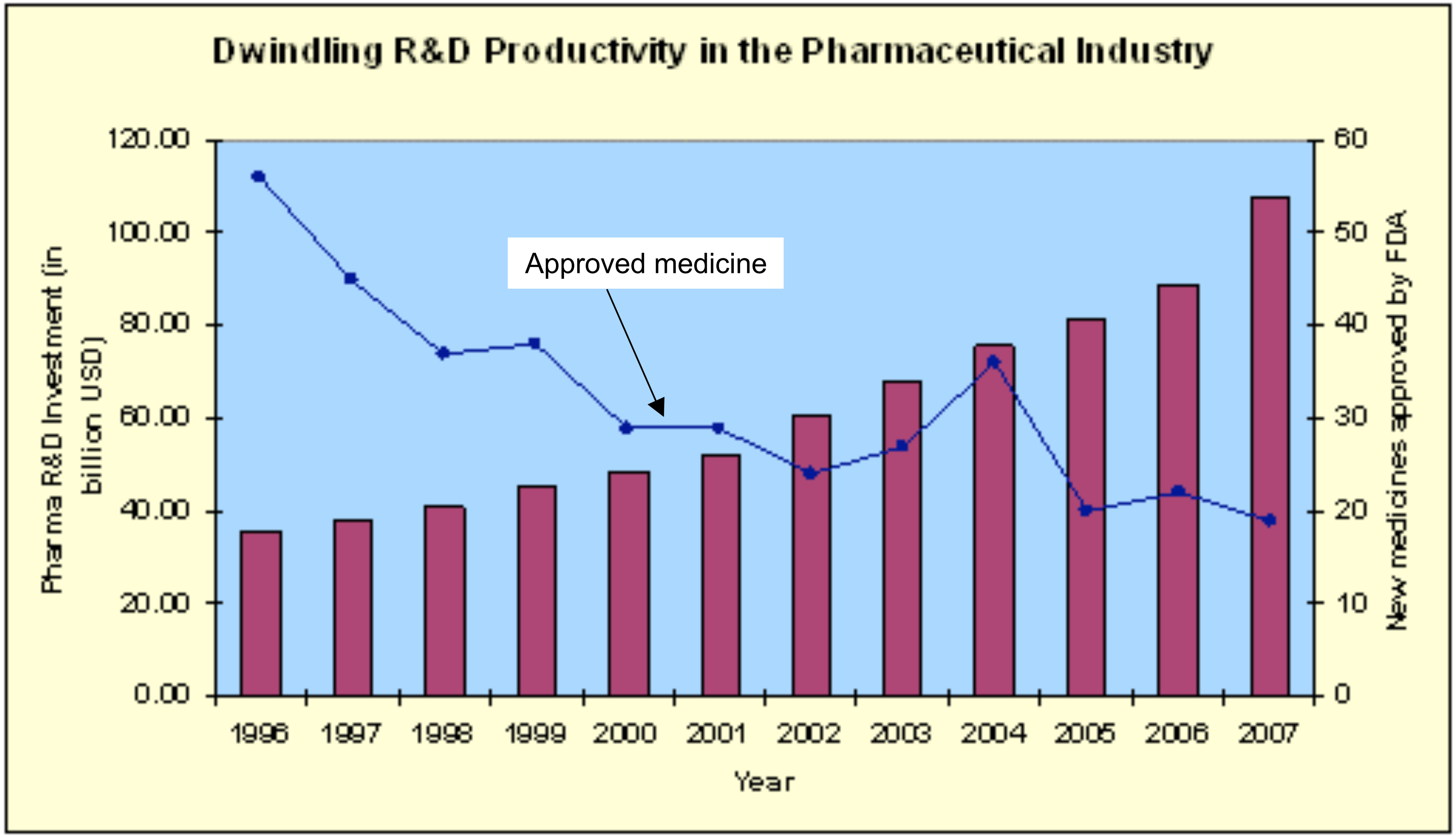
Igor V. Tetko

Helmholtz Zentrum München - German Research Center for Environmental Health (GmbH)
Institute of Bioinformatics & Systems Biology (HMGU)

HELMHOLTZ
ASSOCIATION

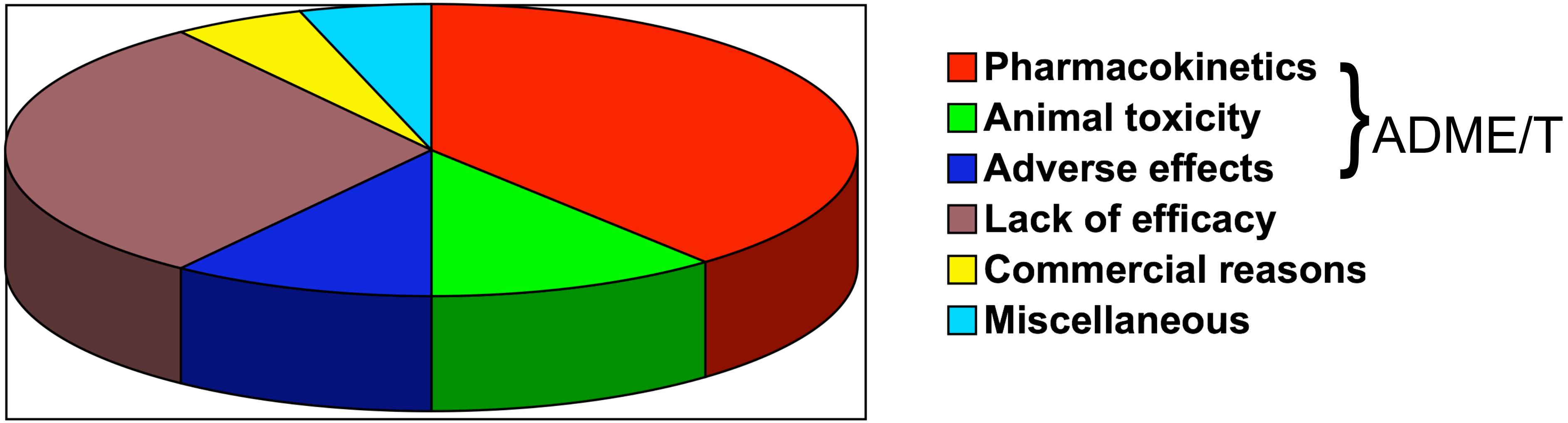
Heidelberg, 11 October 2008

Declining R&D productivity in the pharmaceutical industry



Source : PhRMA 2007, FDA

Reasons for failure in drug development



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

Chemical Universe and reasons for failure of ADME/T methods

“One can not embrace the unembraceable.”

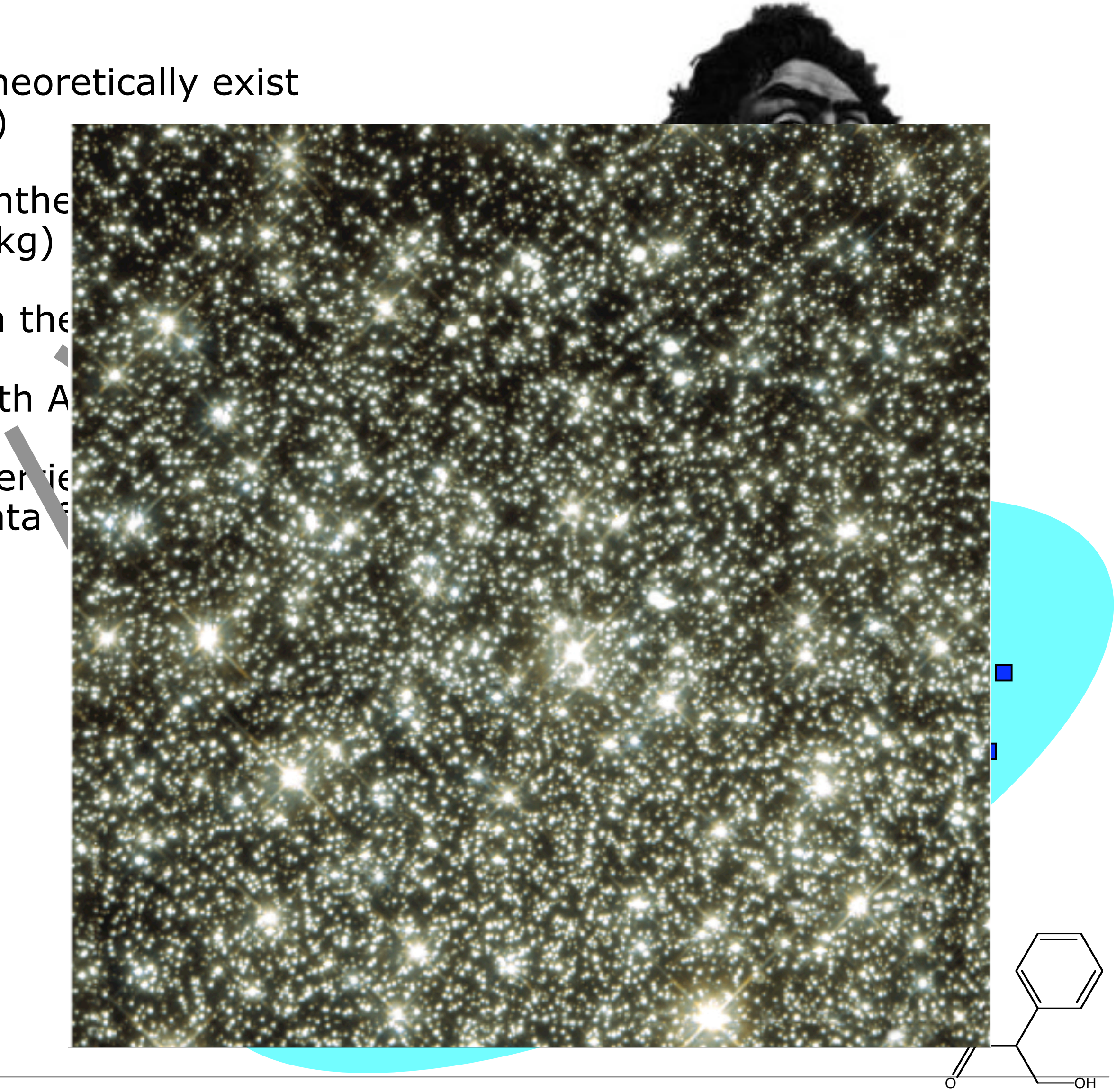
Possible: 10^{60} - 10^{100} molecules theoretically exist
(> 10^{80} atoms in the Universe)

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

Available: $2*10^{7}$ molecules are on the

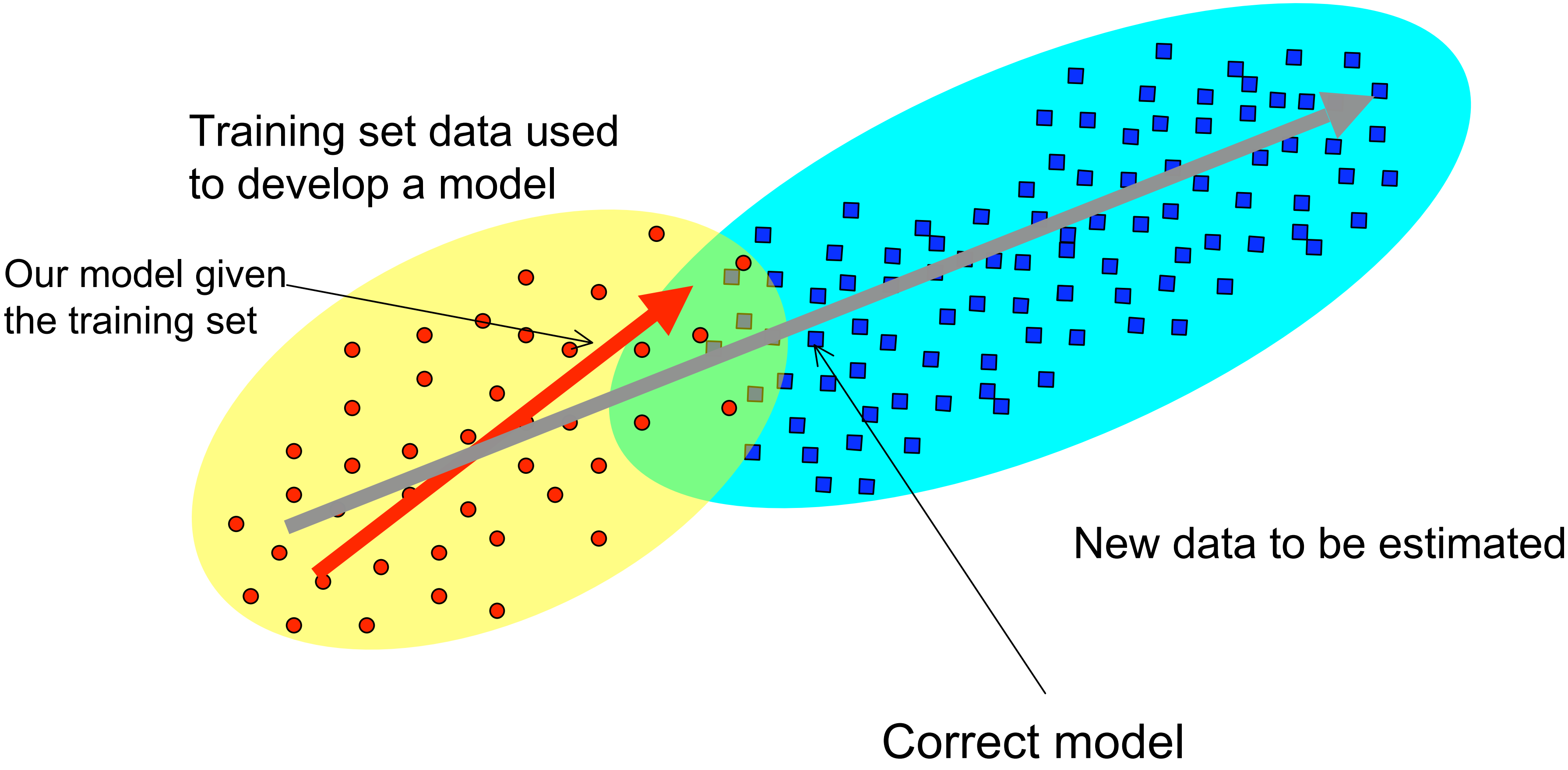
Measured: 10^{2} - 10^{4} molecules with A

Problem: To predict ADME/T proper ie
market we must extrapolate data
molecules!



We need methods which can estimate the accuracy of predictions!

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



Training set data used to develop a model

Our model given the training set

New data to be estimated

Correct model

Benchmarking of logP methods for in-house data of Pfizer & Nycomed

LogP - octanol/water partition coefficient

One of the most important descriptors in the drug discovery

Correlates with many biological and ADME/T properties of molecules

Supported with one of the largest experimental database

3rd dedicated conference will be in Zurich next year (logP2009)

Performance of algorithms for *in-house* datasets

	Pfizer set (*N* = 95 809)							Nycomed set (*N* = 882)				
	RMSE	Failed[1]	**rank**	% in error range			*RMSE, zwitterions excluded*[2]	*RMSE*	**rank**	% in error range		
Method				<0.5	0.5-1	>1				<0.5	0.5-1	>1
Consensus log *P*	0.95		I	48	29	24	0.94	0.58	I	61	32	7
ALOGPS	1.02		I	41	30	29	1.01	0.68	I	51	34	15
S+logP	1.02		I	44	29	27	1.00	0.69	I	58	27	15
NC+NHET	1.04		II	38	30	32	1.04	0.88	III	42	32	26
MLOGP(S+)	1.05		II	40	29	31	1.05	1.17	III	32	26	41
XLOGP3	1.07		II	43	28	29	1.06	0.65	I	55	34	12
MiLogP	1.10	27	II	41	28	30	1.09	0.67	I	60	26	14
AB/LogP	1.12	24	II	39	29	33	1.11	0.88	III	45	28	27
ALOGP	1.12		II	39	29	32	1.12	0.72	II	52	33	15
ALOGP98	1.12		II	40	28	32	1.10	0.73	II	52	31	17
OsirisP	1.13	6	II	39	28	33	1.12	0.85	II	43	33	24
AAM	**1.16**		**III**	**33**	**29**	**38**	**1.16**	**0.94**	**III**	**42**	**31**	**27**
CLOGP	1.23		III	37	28	35	1.21	1.01	III	46	28	22
ACD/logP	1.28		III	35	27	38	1.28	0.87	III	46	34	21
CSlogP	1.29	20	III	37	27	36	1.28	1.06	III	38	29	33
COSMOFrag	1.30	1088[3]	III	32	27	40	1.30	1.06	III	29	31	40
QikProp	1.32	103	III	31	26	43	1.32	1.17	III	27	24	49
KowWIN	1.32	16	III	33	26	41	1.31	1.20	III	29	27	44
QLogP	1.33	24	III	34	27	39	1.32	0.80	II	50	33	17
XLOGP2	1.80		III	15	17	68	1.80	0.94	III	39	31	29
MLOGP(Dragon)	2.03		III	34	24	42	2.03	0.90	III	45	30	25

[1] Nr of molecules with calculations failures due to errors or crash of programs. All methods predicted all molecules for the Nycomed dataset. [2] *RMSE* calculated after excluding of 769 zwitterionic compounds from the Pfizer dataset. [3] Most molecules failed by COSMOFrag are zwitterions.

Mannhold, R. et al, J. Pharm. Sci., 2008, in press.

http://www.vcclab.org

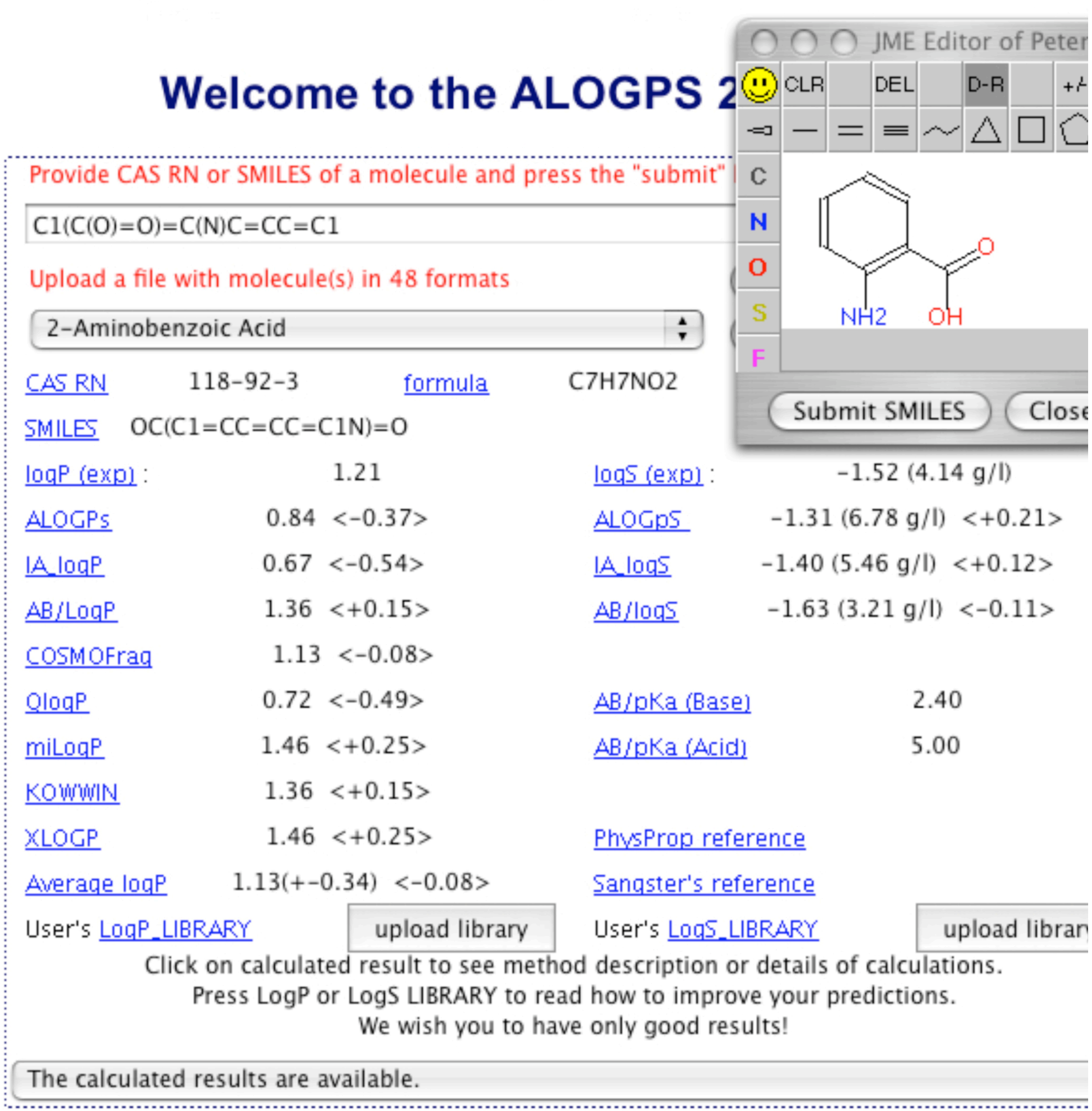
Virtual Computational Chemistry Laboratory

ALOGPS 2.1

- LogP: 75 variables,
  12908 molecules,
  RMSE=0.35,
  MAE=0.26

- LogS: 33 variables,
  1291 molecules,
  RMSE=0.49,
  MAE=0.35

***Tetko et al, J. Comput. Aided Mol. Des. 2005,* 19, 453-463.**
Tetko & Tanchuk, J. Chem. Info. Comput. Sci., 2004, 2002, 42, 1136-1145.



Welcome to the ALOGPS 2

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

C1(C(O)=O)=C(N)C=CC=C1

Upload a file with molecule(s) in 48 formats

2-Aminobenzoic Acid

CAS RN	118-92-3	formula	C7H7NO2
SMILES	OC(C1=CC=CC=C1N)=O		
logP (exp) :	1.21	logS (exp) :	-1.52 (4.14 g/l)
ALOGPs	0.84 <-0.37>	ALOGpS	-1.31 (6.78 g/l) <+0.21>
IA_logP	0.67 <-0.54>	IA_logS	-1.40 (5.46 g/l) <+0.12>
AB/LogP	1.36 <+0.15>	AB/logS	-1.63 (3.21 g/l) <-0.11>
COSMOFrag	1.13 <-0.08>		
QlogP	0.72 <-0.49>	AB/pKa (Base)	2.40
miLogP	1.46 <+0.25>	AB/pKa (Acid)	5.00
KOWWIN	1.36 <+0.15>		
XLOGP	1.46 <+0.25>	PhysProp reference	
Average logP	1.13(+-0.34) <-0.08>	Sangster's reference	
User's LogP_LIBRARY	upload library	User's LogS_LIBRARY	upload librar

Click on calculated result to see method description or details of calculations.
Press LogP or LogS LIBRARY to read how to improve your predictions.
We wish you to have only good results!

The calculated results are available.

Methodology: Associative Neural Network (ASNN)

Some software tools rely just on one “best” model.

Other software tools rely on the ensemble average (“panel of experts”).

ASNN explores disagreement of individual models in the ensemble to improve its accuracy and to derive a confidence score.

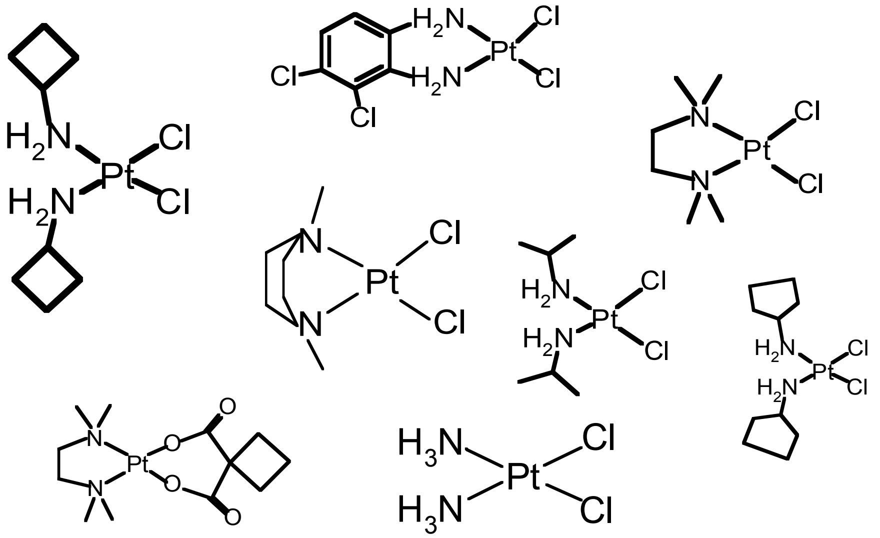
Highlighted Examples

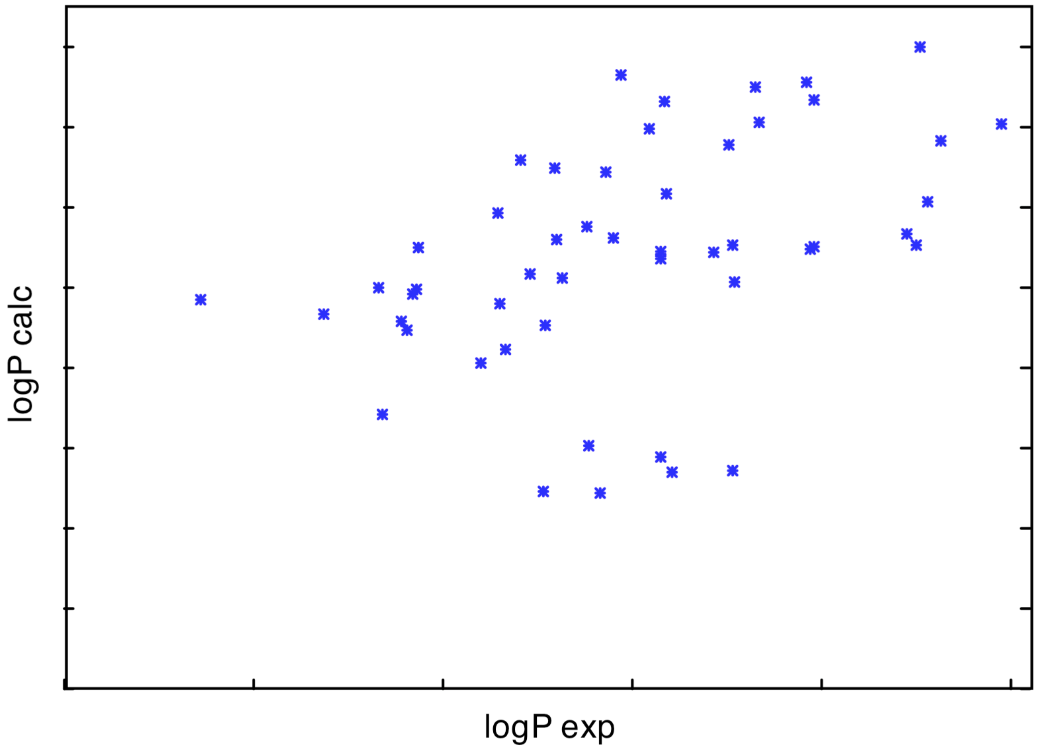
- Development of focused (local) models
- Estimation of accuracy of predictions
- Multi-task learning

This model does not work for these data…

Is it possible to improve it by using new measurements?

Local models: Instance learning of logP for PtII molecules

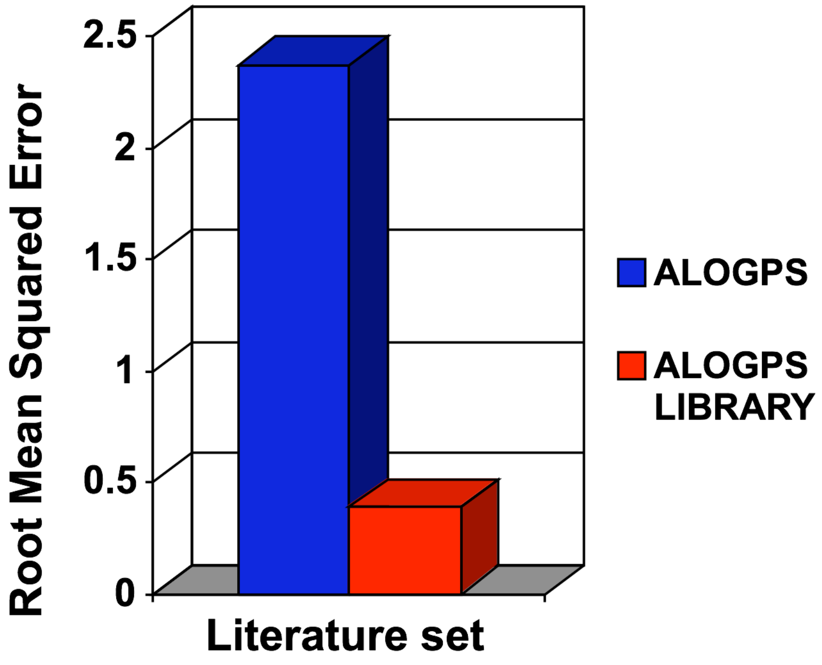


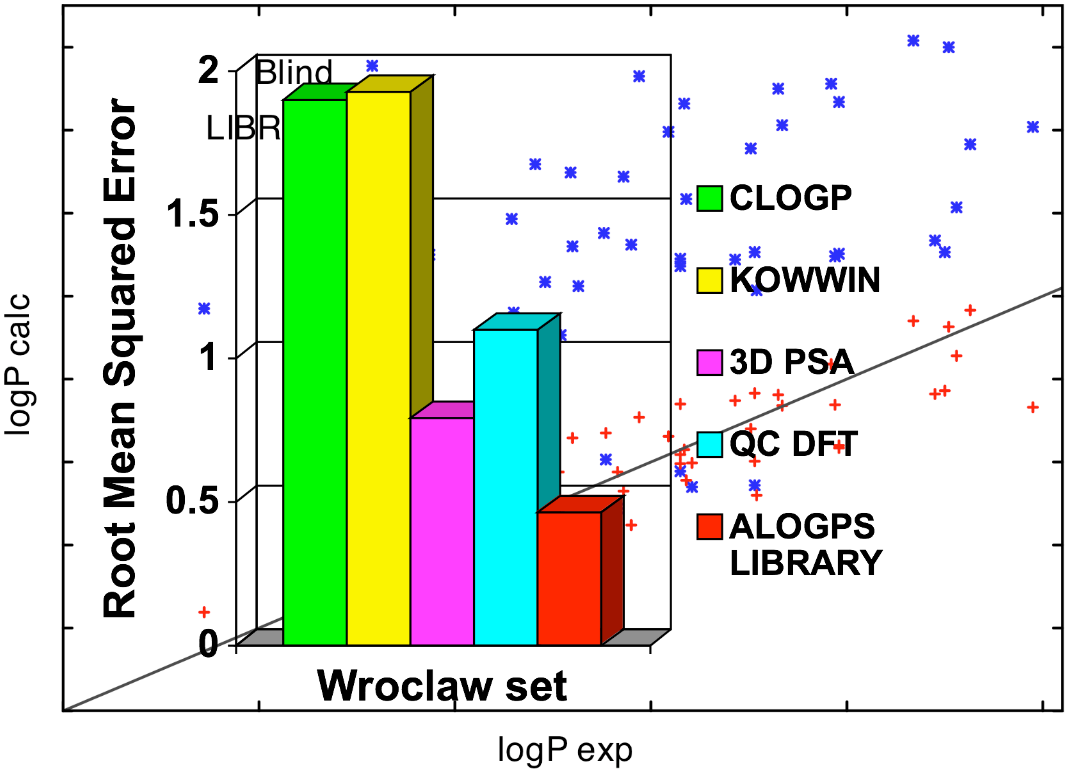


The Figure shows that prediction of new classes of compounds can be extremely difficult as exemplified by an absence of correlations between predicted and experimental values.

Tetko et al, J. Inorg. Biochem, 2008, 102, 1424-37.

Local models: Instance learning by knowledge transfer

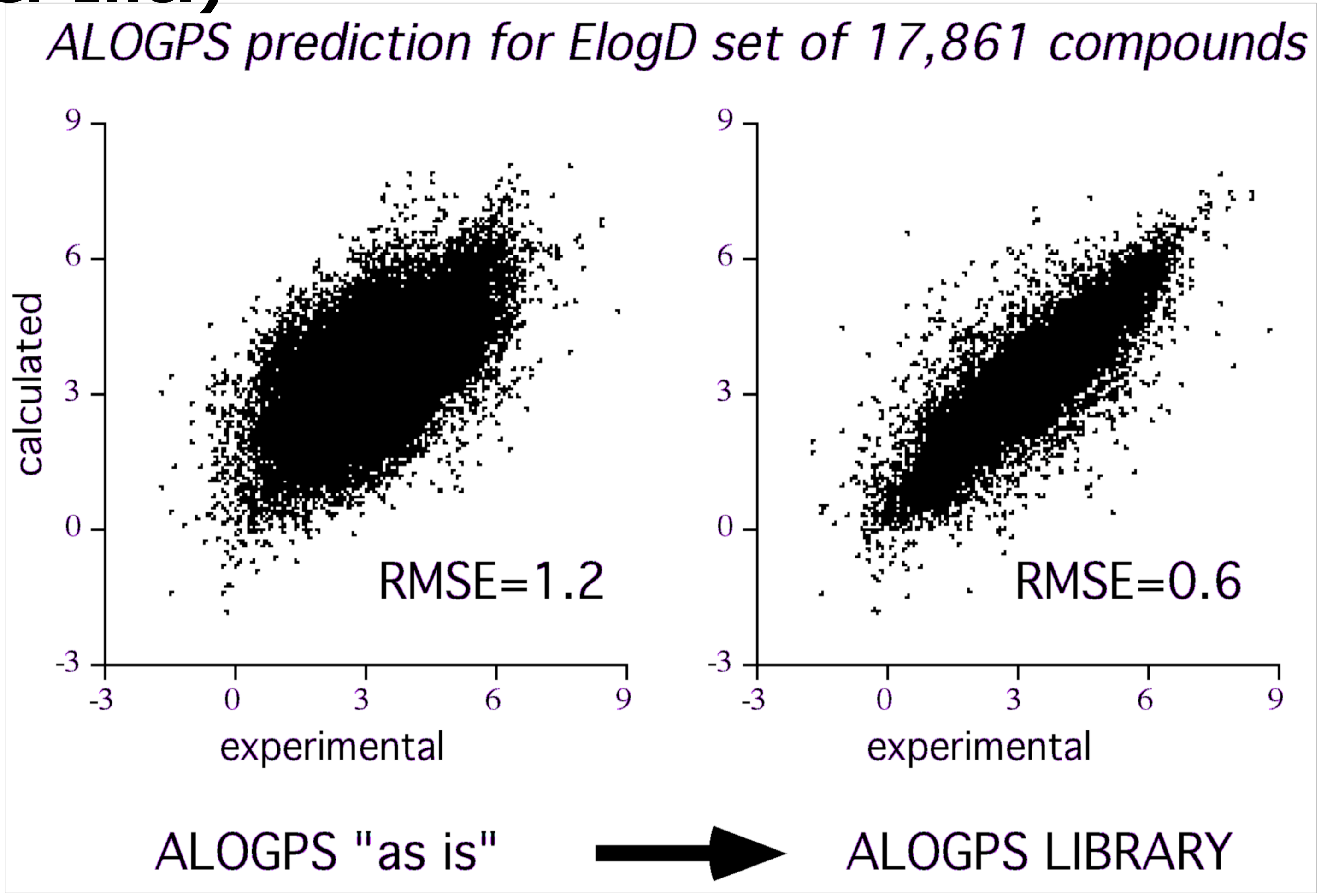




The right panel shows that our methodology (red column) allowed to calculate superior prediction (lower errors) compared to traditional methods.

Tetko et al, J. Inorg. Biochem, 2008, 102, 1424-37.

Local models: Instant learning of in-house data (Pfizer Inc.)



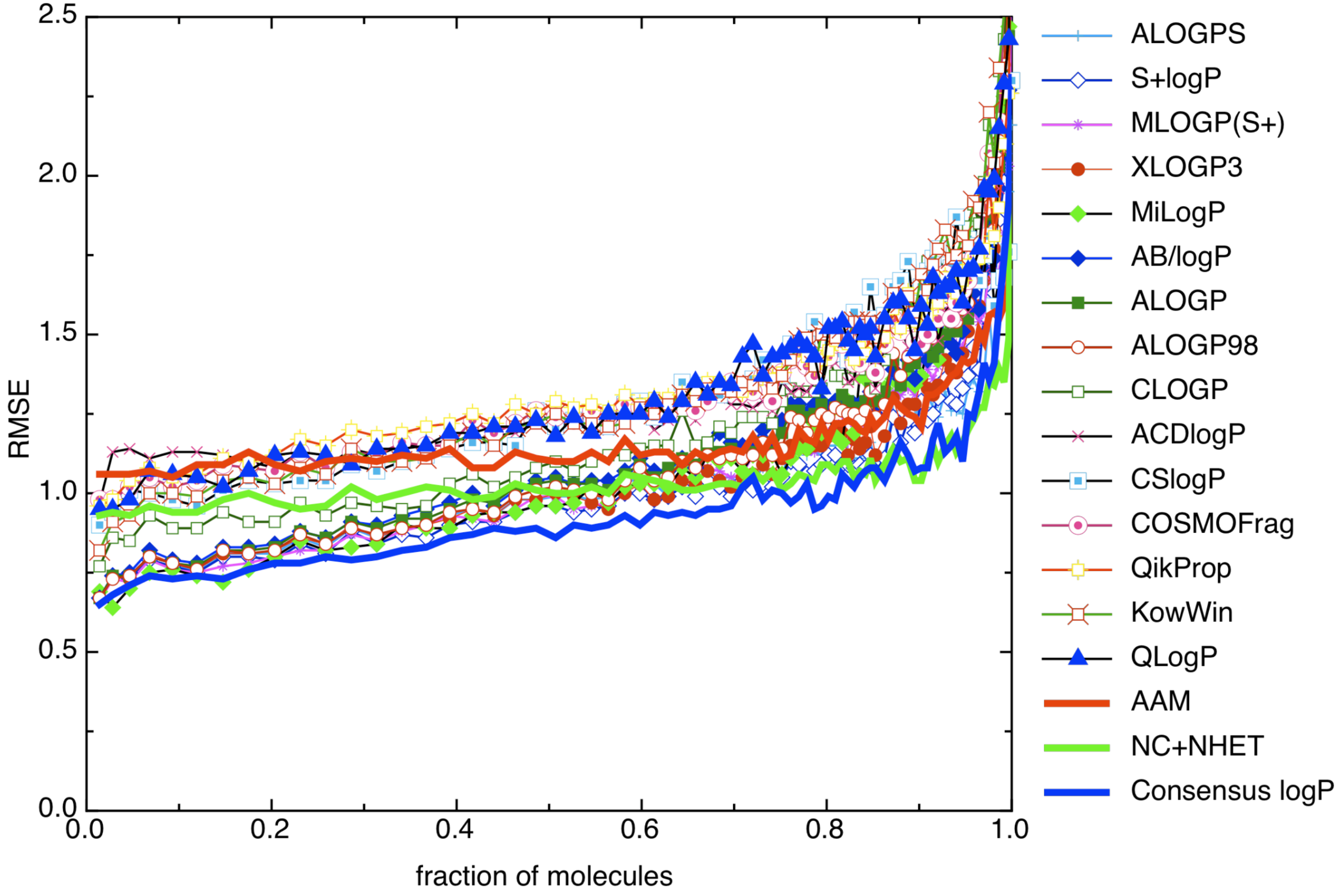
The LIBRARY mode produced local models and dramatically decreased the error for a very large set of compounds in just less than 10 minutes of calculations.

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

Is it possible to distinguish reliable vs. non-reliable predictions?

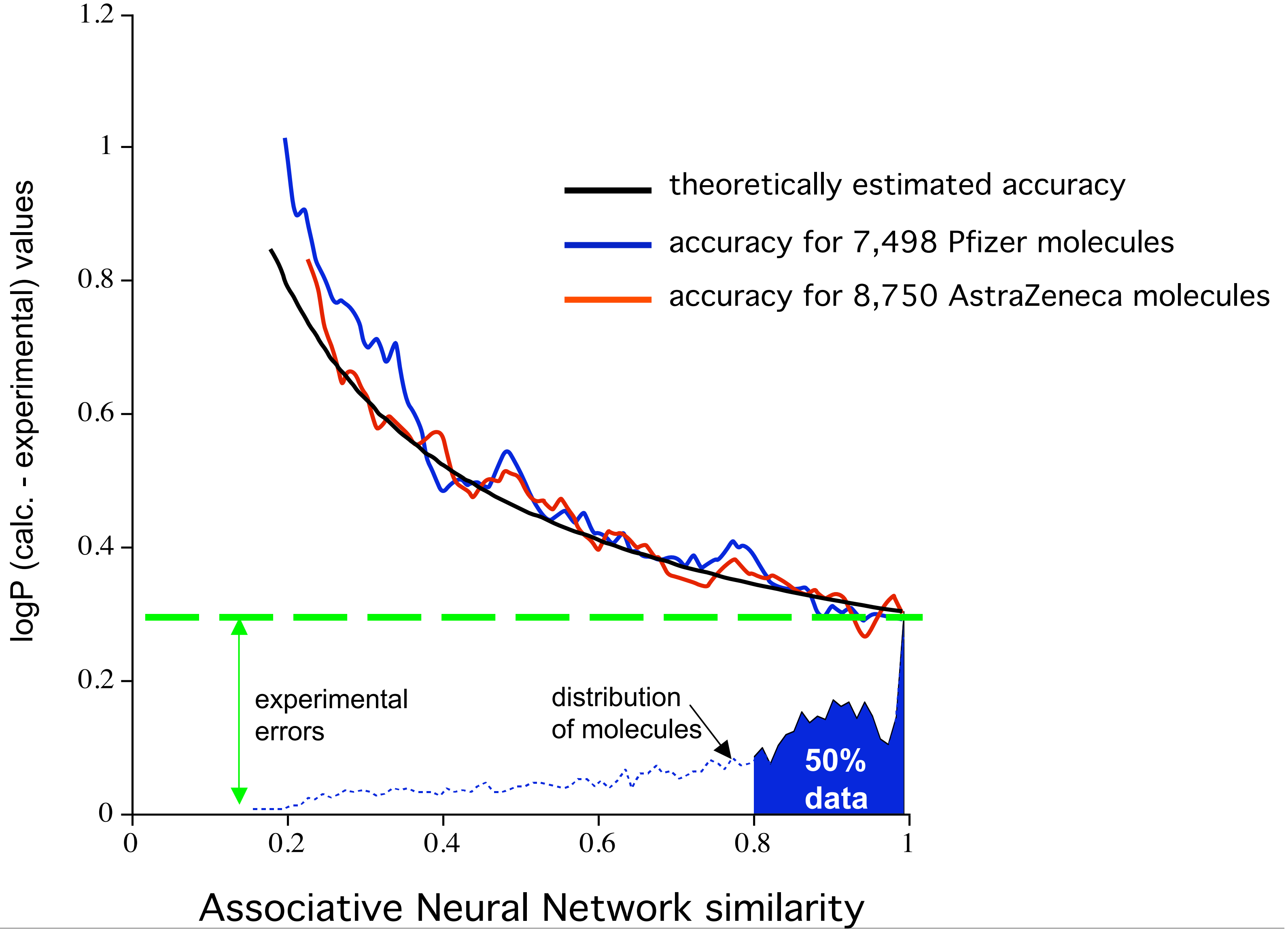
Is it possible to save costs by skipping measurements of some molecules?

Global model: Accuracy of logP predictions for 96,000 molecules



Mannhold, R. et al, *J. Pharm. Sci.*, 2008, in press.

Local model: Accuracy of logP predictions for a subset of data



Tetko, I.V. et al, DDT, 2006, 11, 700-7.

The measurements are very expensive…

Is it possible to use some related measurements to develop a better model?

Multi-task learning

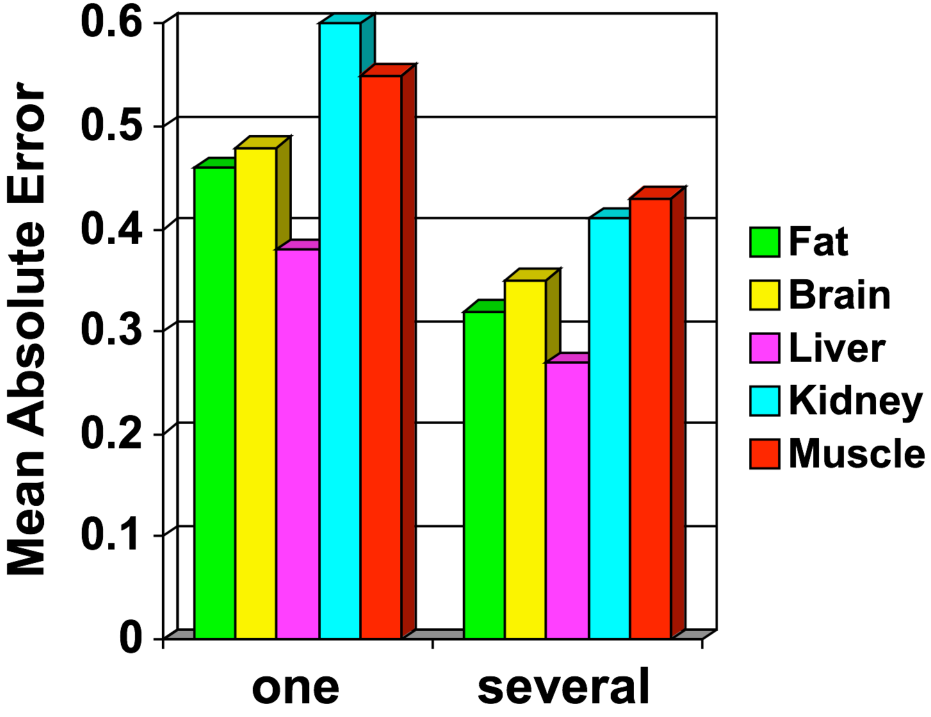
Problem:

- prediction of tissue-air partition coefficients
- small datasets 30-100 molecules (human & rat data)

Results:

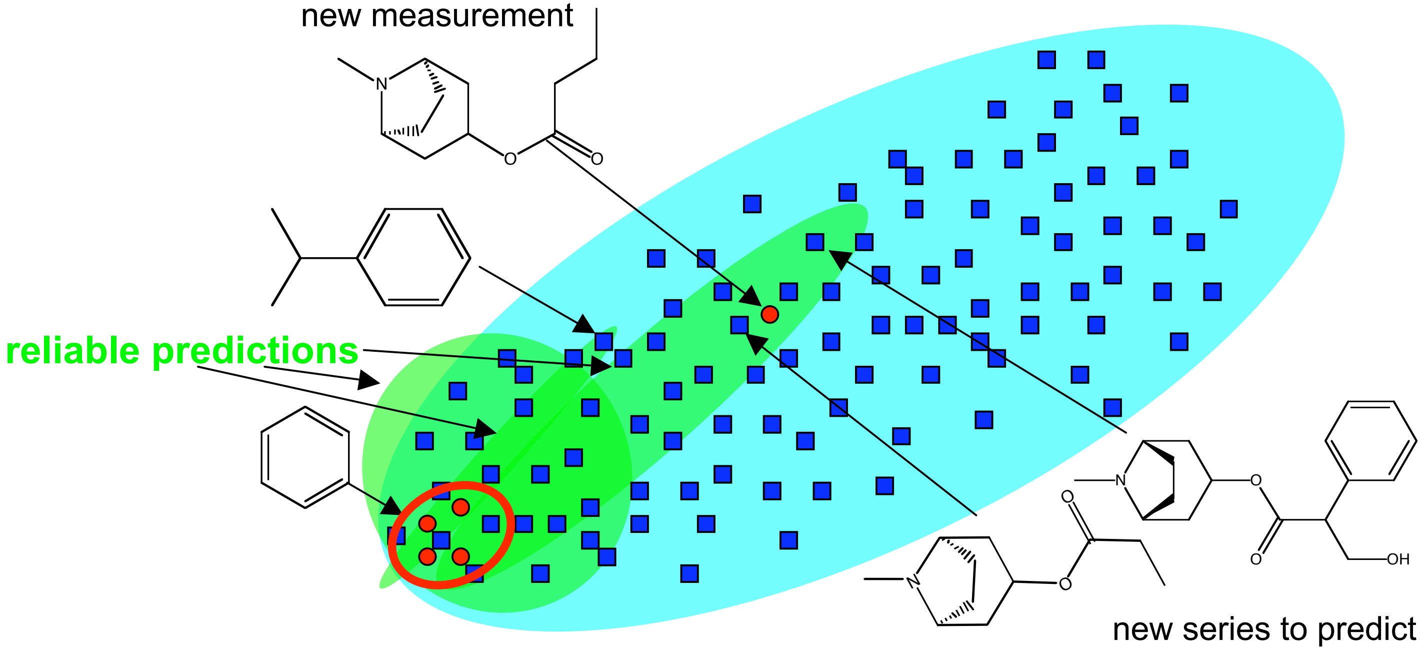
simultaneous prediction of several properties increased the accuracy of models

Human data



Varnek et al, 2008, submitted.

Challenges and solutions



New methodology allows navigation in space of molecules with a confidence.

- ✓ It can be used to develop targeted (local) models to cover specific series.
- ✓ It can be used to reliably estimate which compounds can/can’t be reliably predicted.
- ✓ It can be used to provide experimental design and to minimize costs of new measurements.

Acknowledgement

Dr. G. Poda (Pfizer)

Dr. P. Bruneau (AstraZeneca)

Dr. C. Ostermann (Nycomed)

Prof. R. Mannhold (Dusseldorf University)

+ many other colleagues & co-authors

Prof. G. Wess

Prof. H.W. Mewes

Go-Bio BMBF

FP6 INTAS (VCCLAB, http://www.vcclab.org)

INTAS