HelmholtzZentrum münchen
German Research Center for Environmental Health

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

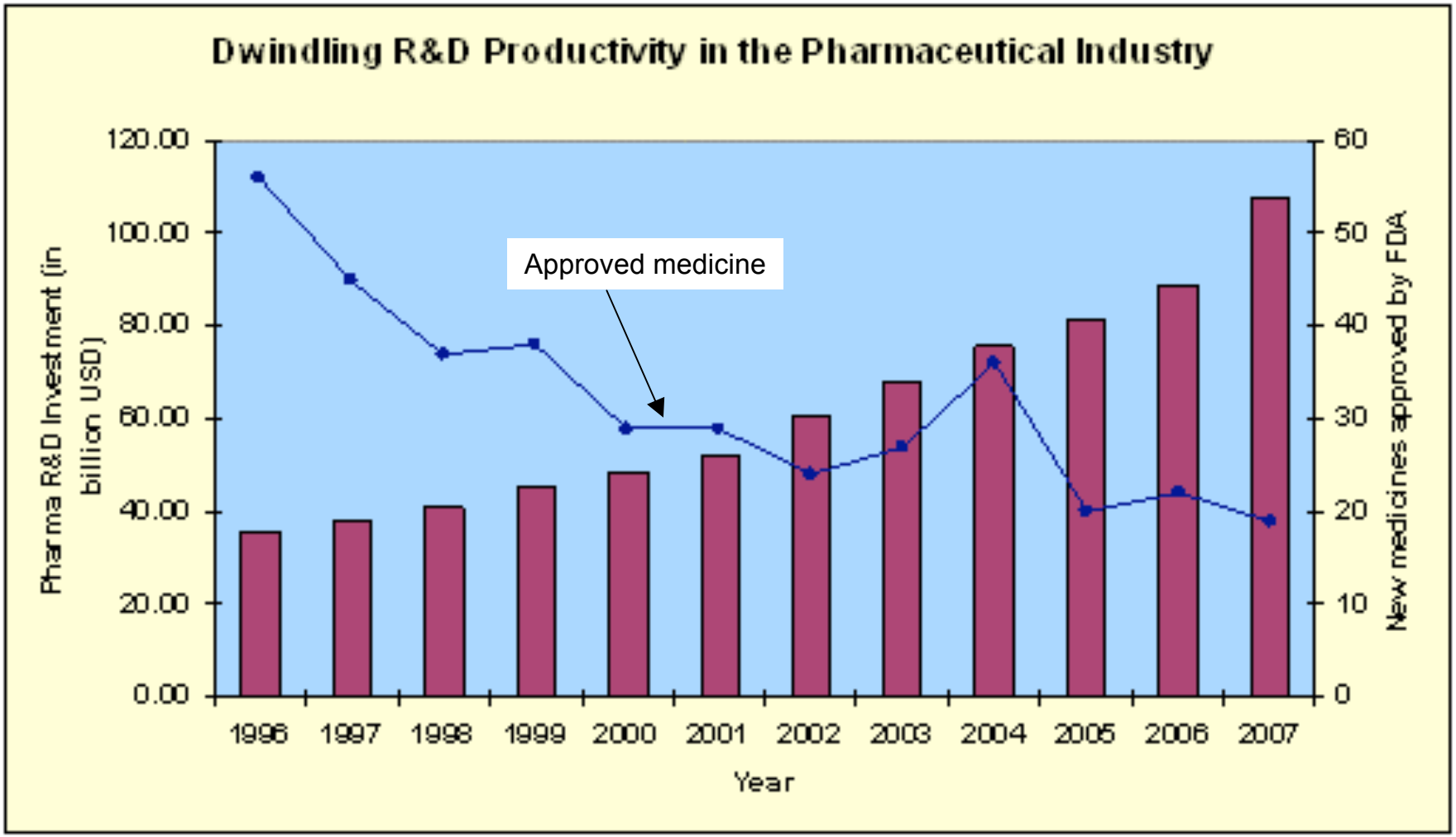
Igor V. Tetko

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

Berlin, 25 February 2009, ADMETox Europe

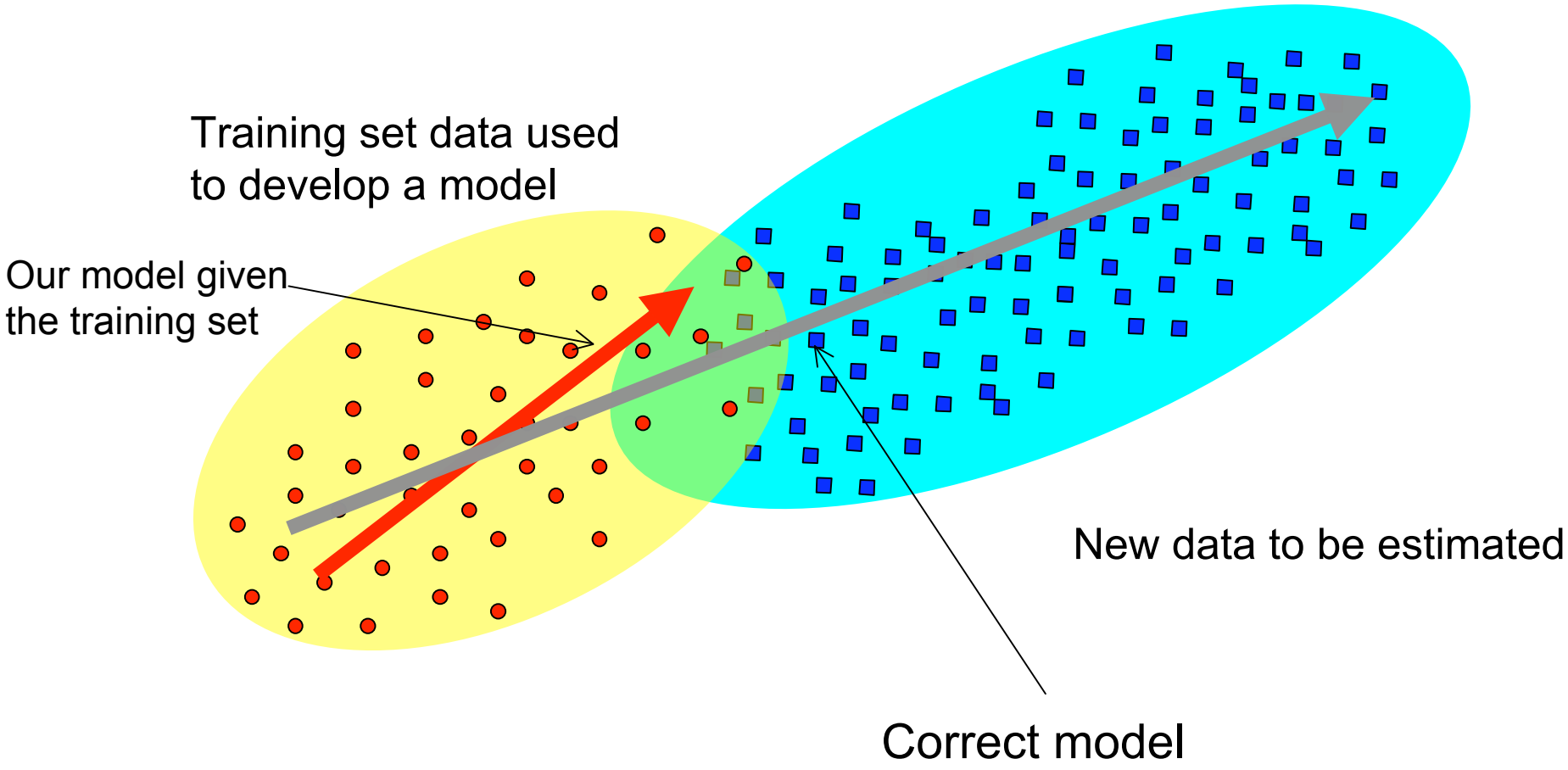
HELMHOLTZ ASSOCIATION

Declining R&D productivity in the pharmaceutical industry



http://www.frost.com/prod/servlet/market-insight-top.pag?docid=128394740

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



"One can not embrace the unembraceable."

Kozma Prutkov

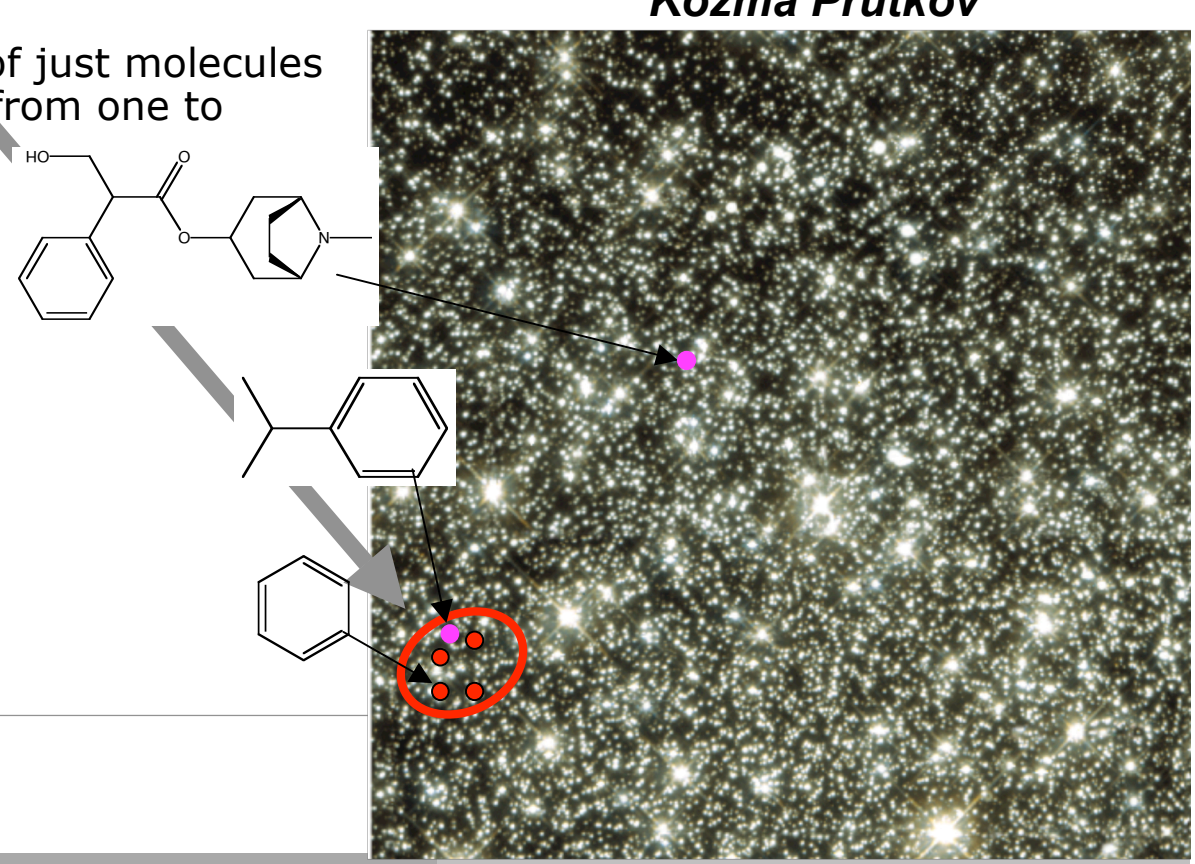
Possible: 10^{60} - 10^{100} molecules theoretically exist
(> 10^{80} atoms in the Universe)

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

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

Measured: 10^{2} - 10^{4} molecules with ADME/T data

<u>Problem:</u> To predict ADME/T properties of just molecules on the market we must extrapolate data from one to 1,000 - 100,000 molecules!



There is a need for methods which can estimate the accuracy of predictions!

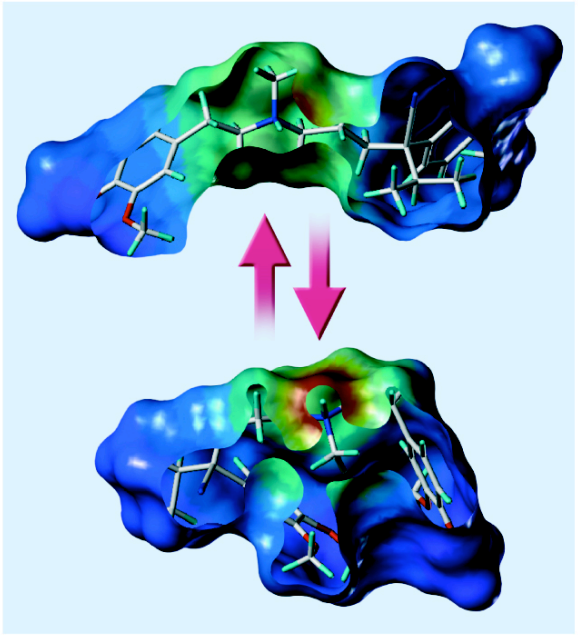
Methods and Principles in Medicinal Chemistry

Edited by
Raimund Mannhold

WILEY-VCH

Molecular Drug Properties

Measurement and Prediction



Volume 37

Series Editors:
R. Mannhold,
H. Kubinyi,
G. Folkers

Performance of algorithms for the public dataset

AAM = average logP used as predicted value for all molecules $R^2=0$

Bootstrap test:
- **rank I** - similar to "best model"
- rank II -- better than AAM
- **rank III** - similar to AAM

[1]Provided by A. Avdeef, *Absorption and drug development. Solubility, permeability and charge state*, ed. Hoboken, NJ: Wiley–Interscience, 2003.

Method	Star set (*N* = 223)					Non-Star set (*N* = 43)				
			% within error range					% within error range		
	RMSE	rank	<0.5	0.5-1	>1	*RMSE*	rank	<0.5	0.5-1	>1
AB/LogP	0.41	I	84	12	4	1.00	I	42	23	35
S+logP	0.45	I	76	22	3	0.87	I	40	35	26
ACD/logP	0.50	I	75	17	7	1.00	I	44	33	23
Consensus log *P*	0.50	I	74	18	8	0.80	I	47	28	26
CLOGP	0.52	II	74	20	6	0.91	I	47	28	26
VLOGP OPS	0.52	II	64	21	7	1.07	I	33	28	26
ALOGPS	0.53	II	71	23	6	0.82	I	42	30	28
MiLogP	0.57	II	69	22	9	0.86	I	49	30	21
XLOGP	0.62	II	60	30	10	0.89	I	47	23	30
KowWIN	0.64	II	68	21	11	1.05	I	40	30	30
CSlogP	0.65	II	66	22	12	0.93	I	58	19	23
ALOGP (Dragon)	0.69	II	60	25	16	0.92	I	28	40	33
MolLogP	0.69	II	61	25	14	0.93	I	40	35	26
ALOGP98	0.70	II	61	26	13	1.00	I	30	37	33
OsirisP	0.71	II	59	26	16	0.94	I	42	26	33
VLOGP	0.72	II	65	22	14	1.13	I	40	28	33
TLOGP	0.74	II	67	16	13	1.12	I	30	37	30
ABSOLV	0.75	II	53	30	17	1.02	I	49	28	23
QikProp	0.77	II	53	30	17	1.24	II	40	26	35
QuantlogP	0.80	II	47	30	22	1.17	II	35	26	40
SLIPPER-2002	0.80	II	62	22	15	1.16	II	35	23	42
COSMOFrag	0.84	II	48	26	19	1.23	II	26	40	33
XLOGP2	0.87	II	57	22	20	1.16	II	35	23	42
QLOGP	0.96	II	48	26	25	1.42	II	21	26	53
VEGA	1.04	II	47	27	26	1.24	II	28	30	42
CLIP	1.05	II	41	25	30	1.54	III	33	9	49
LSER	1.07	II	44	26	30	1.26	II	35	16	49
MLOGP (Sim+)	1.26	II	38	30	33	1.56	III	26	28	47
NC+NHET	1.35	III	29	26	45	1.71	III	19	16	65
SPARC	1.36	III	45	22	32	1.70	III	28	21	49
MLOGP(Dragon)	1.52	III	39	26	35	2.45	III	23	30	47
LSER UFZ	1.60	III	36	23	41	2.79	III	19	12	67
AAM	**1.62**	**III**	**22**	**24**	**53**	**2.10**	**III**	**19**	**28**	**53**
VLOGP-NOPS	1.76	III	1	1	7	1.39	III	7	0	7
HINT	1.80	III	34	22	44	2.72	III	30	5	65
GBLOGP	1.98	III	32	26	42	1.75	III	19	16	65

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

- Large number of methods could not perform better than AAM model
- Best results are calculated using Consensus logP model
- log P = 1.46 + 0.11 (NC - NHET)
  N=95 809, RMSE=1.04, R^2=0.2

Different MlogP implementations demonstrate very different performances for both sets

N.B! Do we really compare methods or their implementations?

Performance of algorithms for *in-house* datasets

Method	Pfizer set (*N* = 95 809) RMSE	Failed[1]	rank	% in error range <0.5	0.5-1	>1	RMSE, zwitterions excluded[2]	Nycomed set (*N* = 882) RMSE	rank	% in error range <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., 2009, 98(3), 861-893.

Highlighted Recipes

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

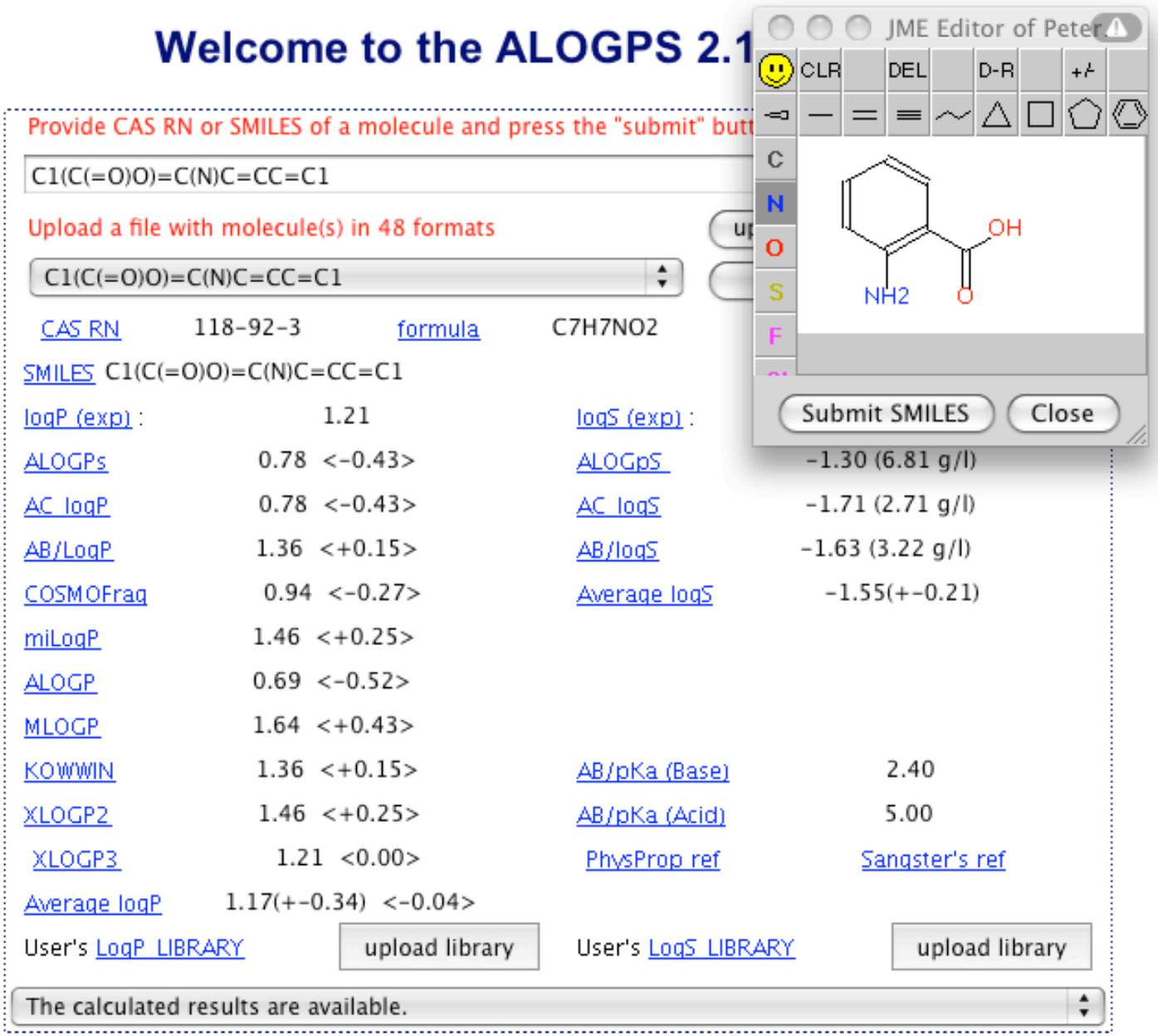
This model does not work for these data…

Is it possible to improve it by using new measurements?

ALOGPS 2.1

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

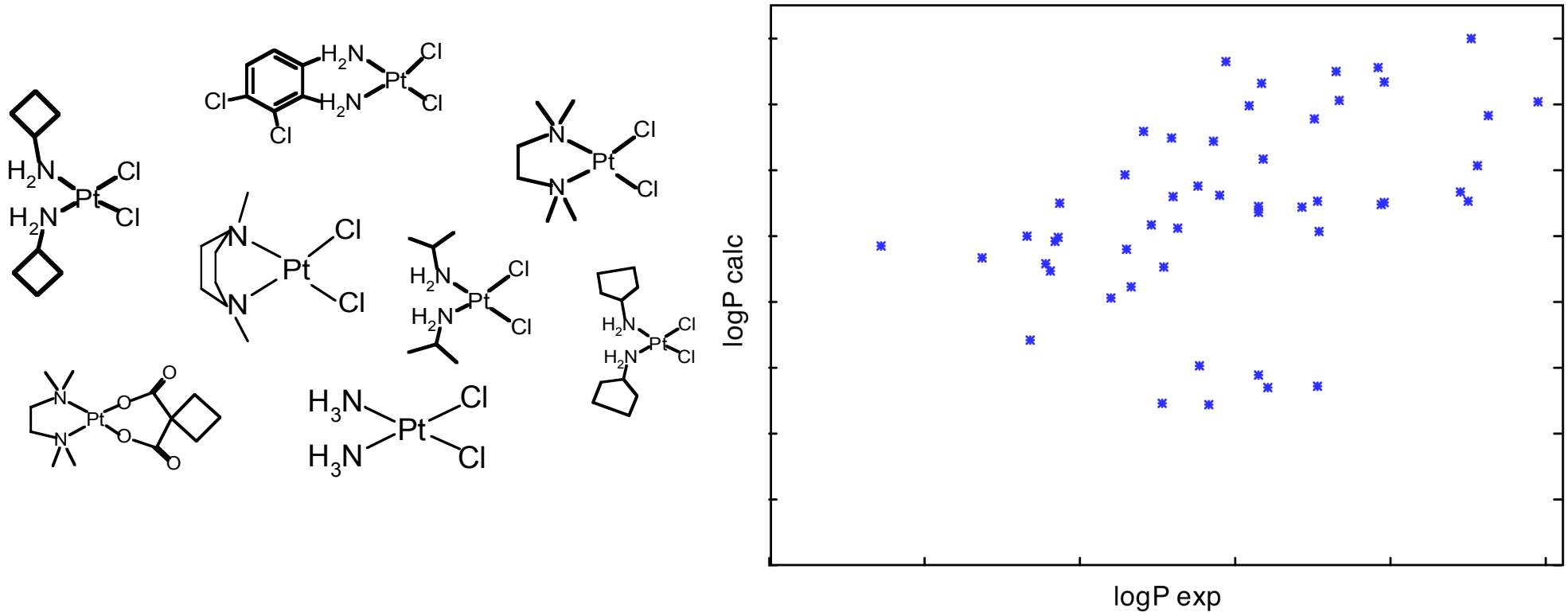
- LogS: 33 variables,
  1291 molecules,
  RMSE=0.49,
  MAE=0.35



Tetko et al, J. Comput. Aided Mol. Des. 2005, 19, 453-463.

Tetko & Tanchuk, *J. Chem. Info. Comput. Sci.*, 2002, 42, 1136-1145.

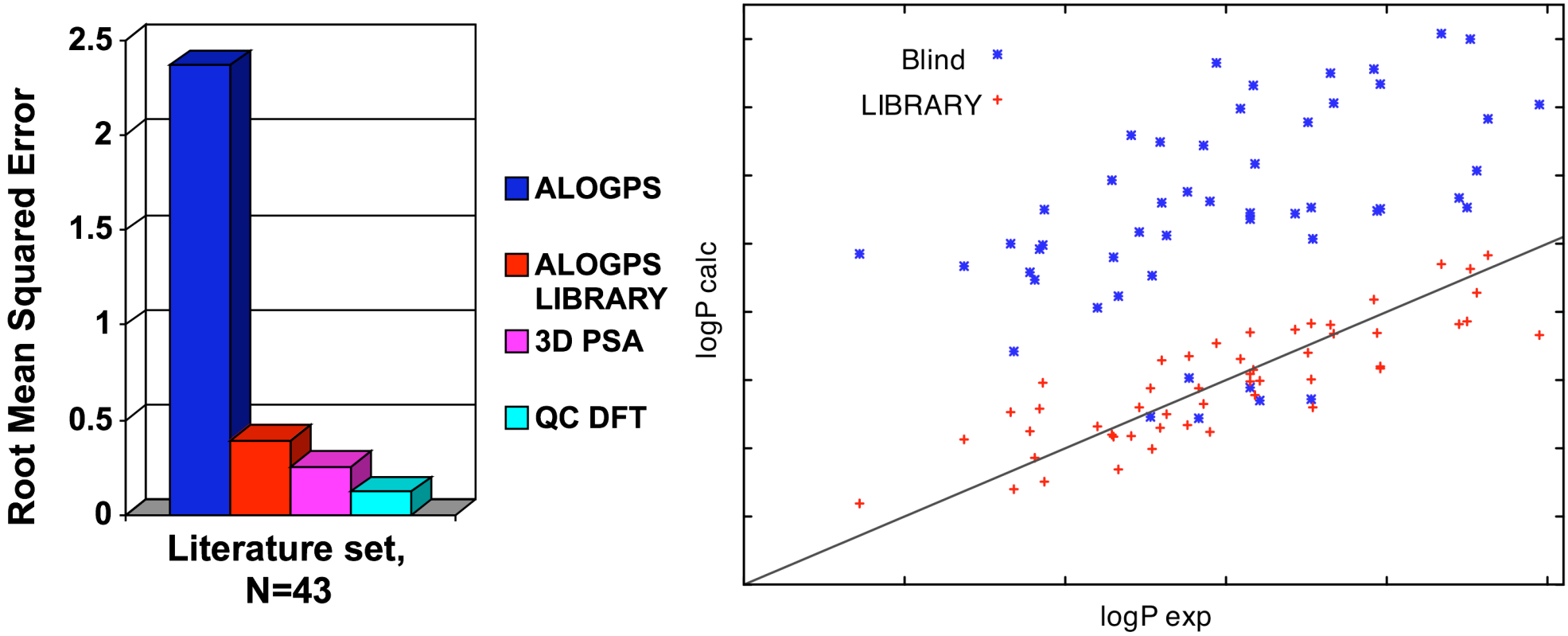
Local models: Instance learning of logP for Pt(II) molecules



Prediction of new classes of compounds can be extremely difficult as exemplified by an absence of correlations between predicted and experimental values using the ALOGPS program.

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

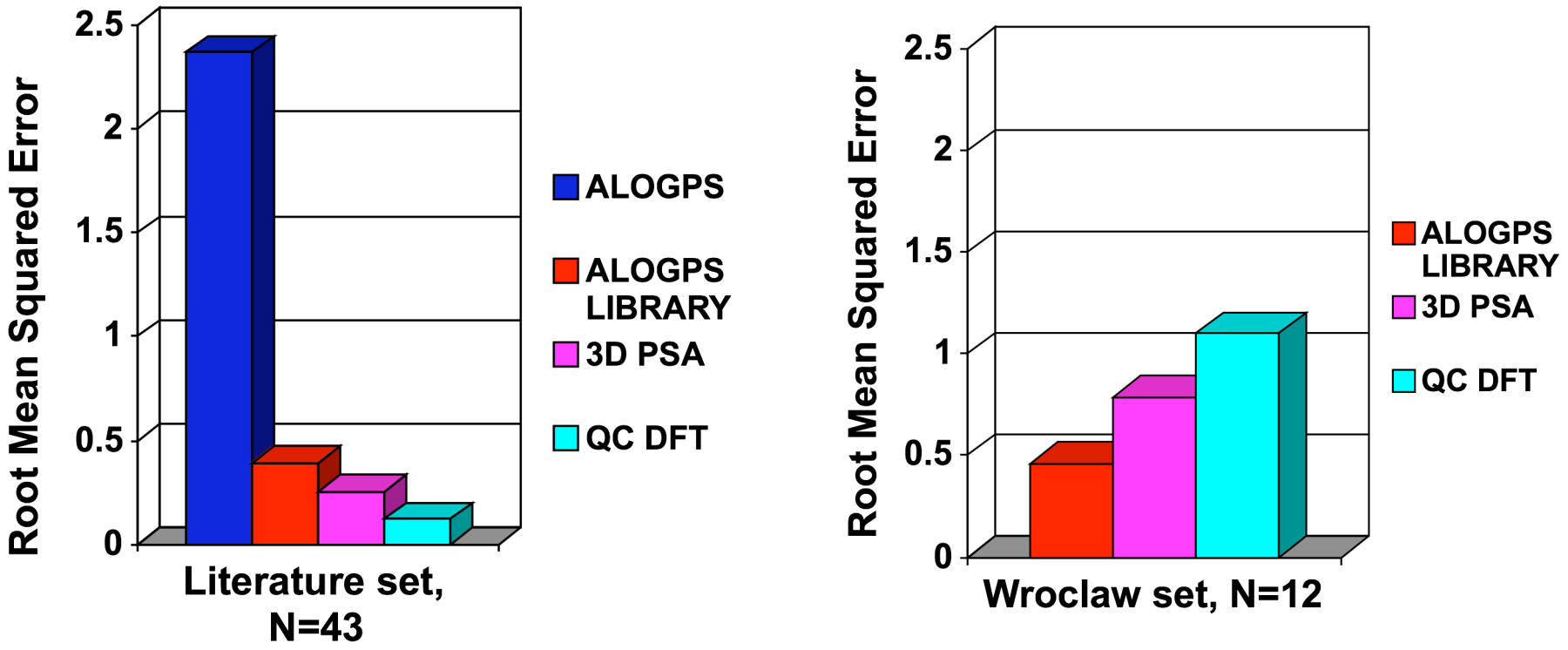
Local models: Instance learning by knowledge transfer



The use of LIBRARY mode (local correction of the global model) dramatically (5 times!) decreased logP errors, but other models look better...

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

Local models: Instance learning by knowledge transfer



The right panel shows that correction of global ALOGPS model with local Pt(II) data (red column) allowed us to calculate superior prediction (lower errors) compared to the models developed using Pt(II) data only (3D PSA & QC DFT) .

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

Performance of algorithms for *in-house* datasets

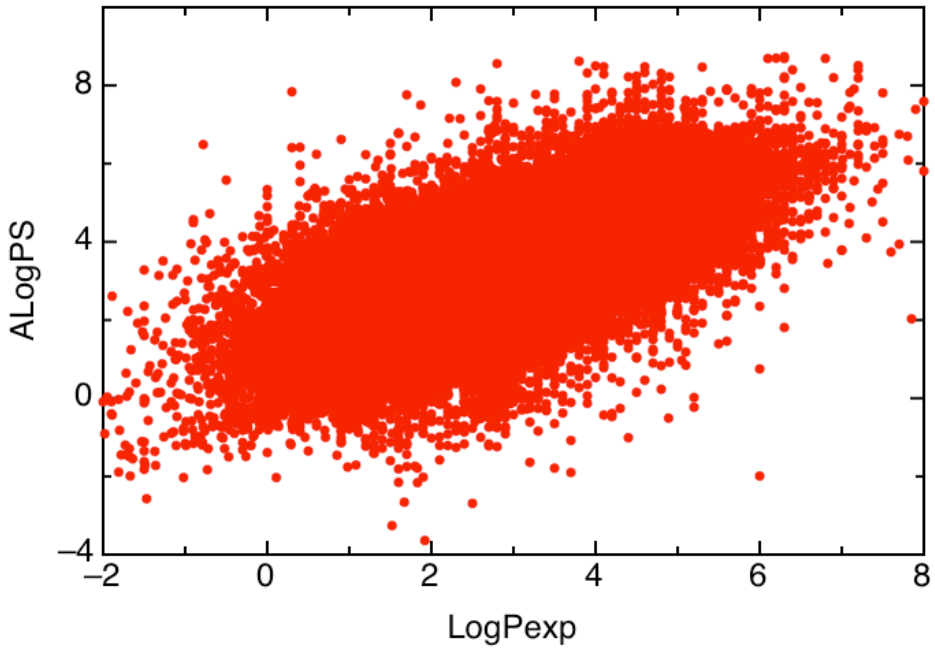
Method	Pfizer set (*N* = 95 809)							Nycomed set (*N* = 882)				
	RMSE	Failed[1]	rank	% in error range			*RMSE*, *zwitterions excluded*[2]	*RMSE*	rank	% in error range		
				<0.5	0.5-1	>1				<0.5	0.5-1	>1
Consensus log *P*	0.95		I	48	29	24	0.94	0.58	I	61	32	7
ALOGPS	1.02		I	41	30	29	1.01	0.68	I	51	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.

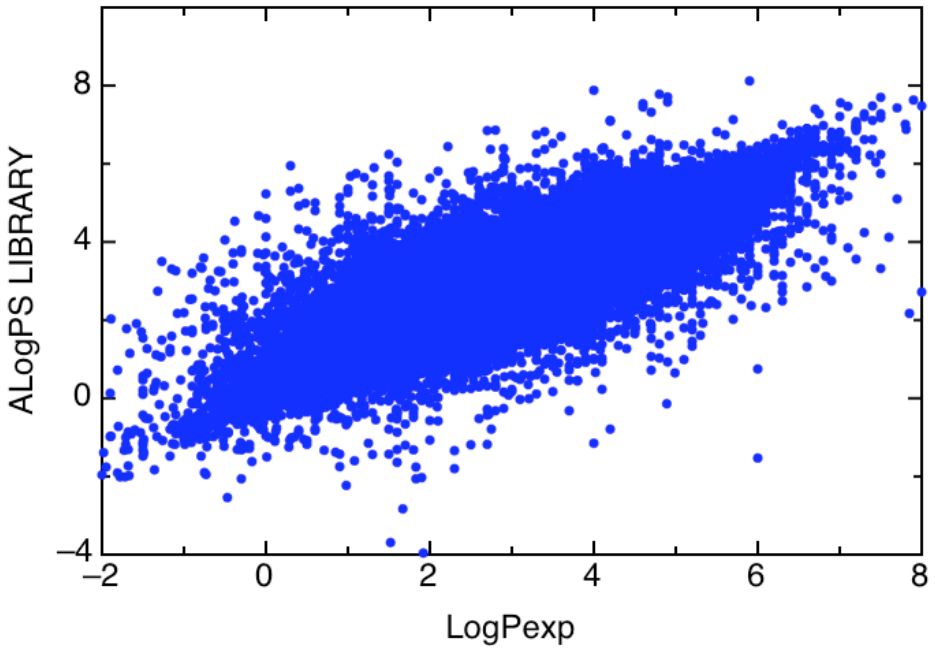
Mannhold, R. et al, J. Pharm. Sci., 2009, 98(3), 861-893.

Local models: Instant learning of in-house data (Pfizer Inc.), *N*=95809

ALOGPS Blind prediction



ALOGPS LIBRARY



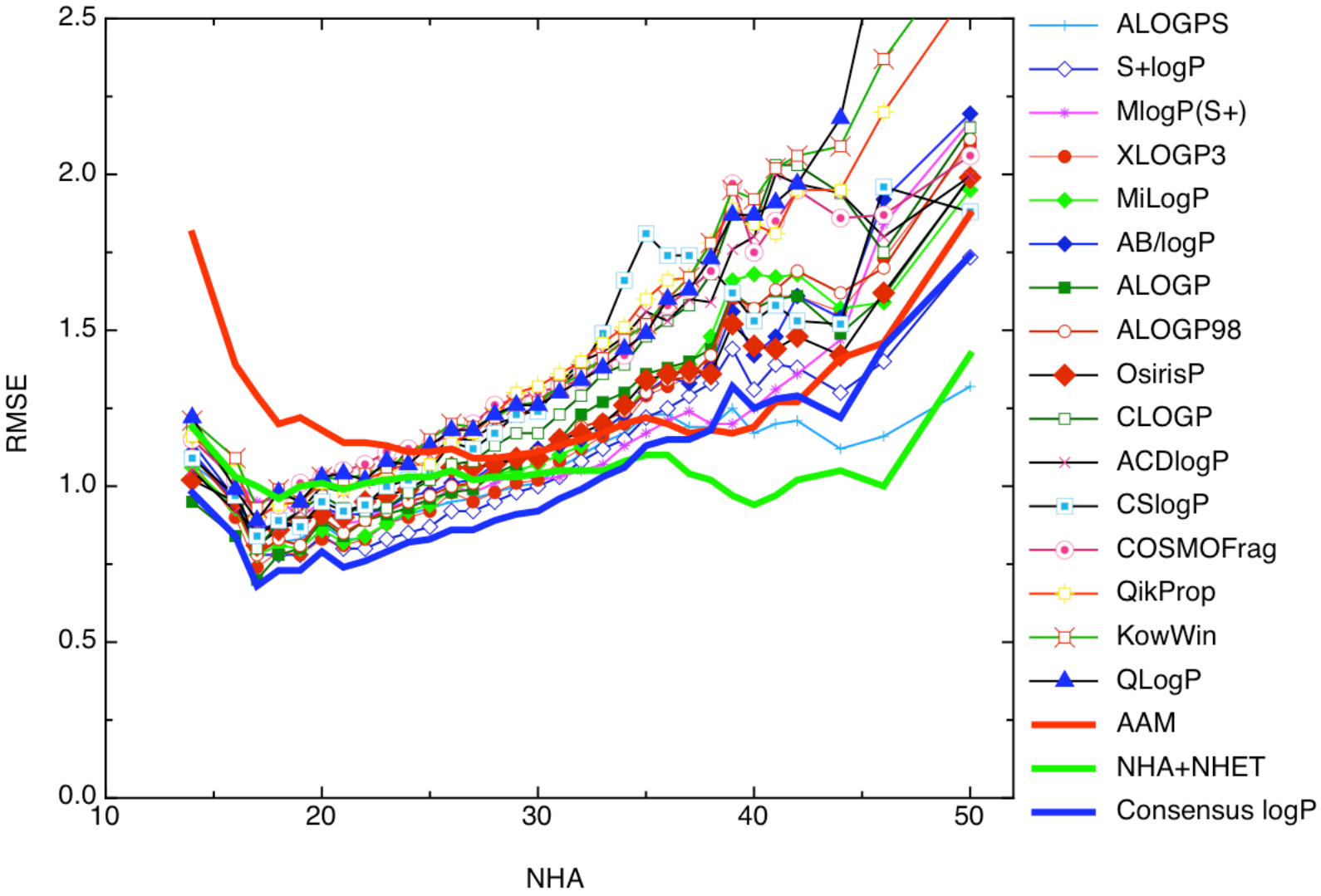
RMSE=1.02 → RMSE=0.59

in less than 30 minutes of calculations on a notebook!

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

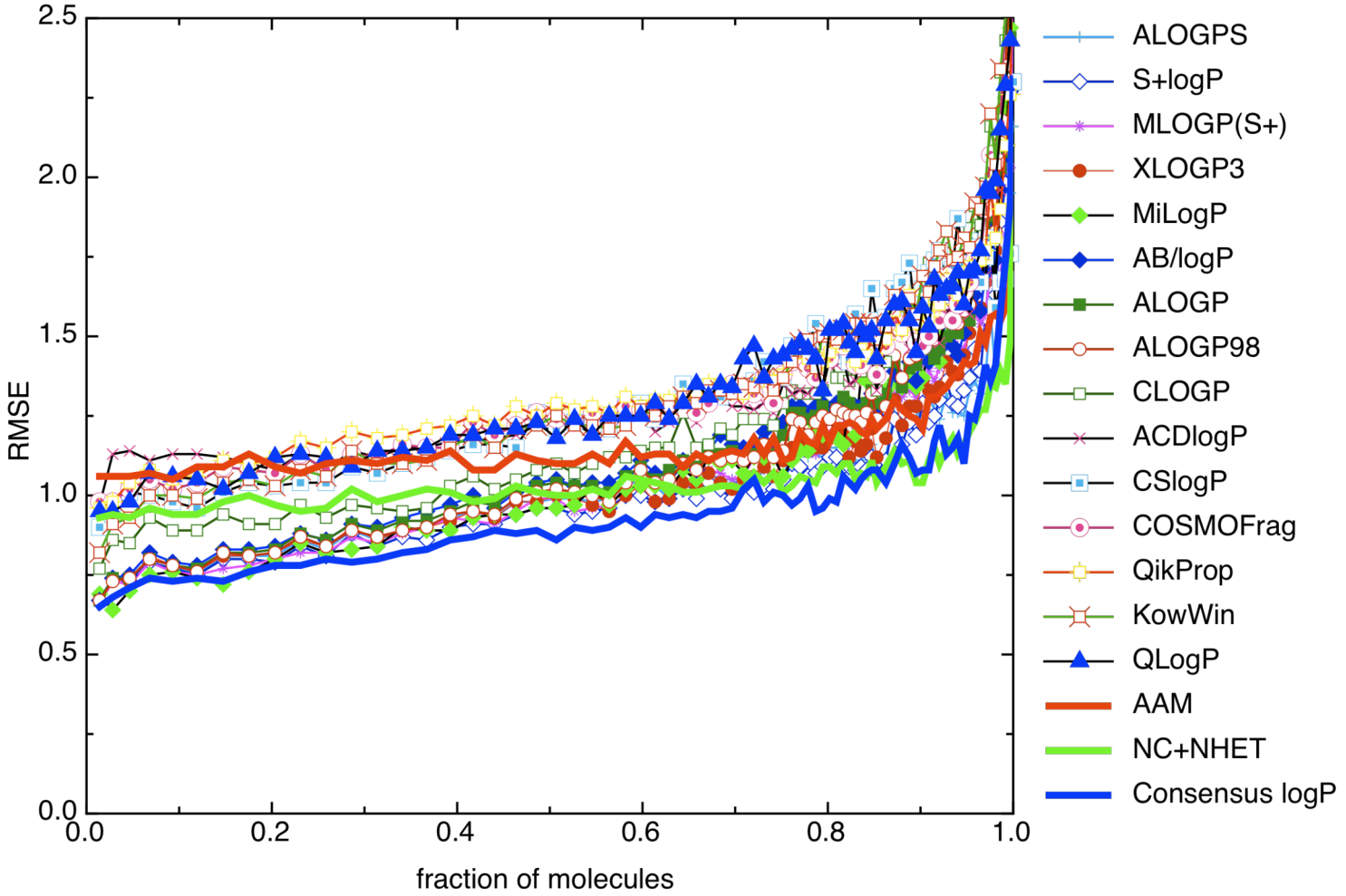
Is it possible to save costs by skipping measurements of some molecules?

Method performance as a function of the Number of Non Hydrogen Atoms (NHA)



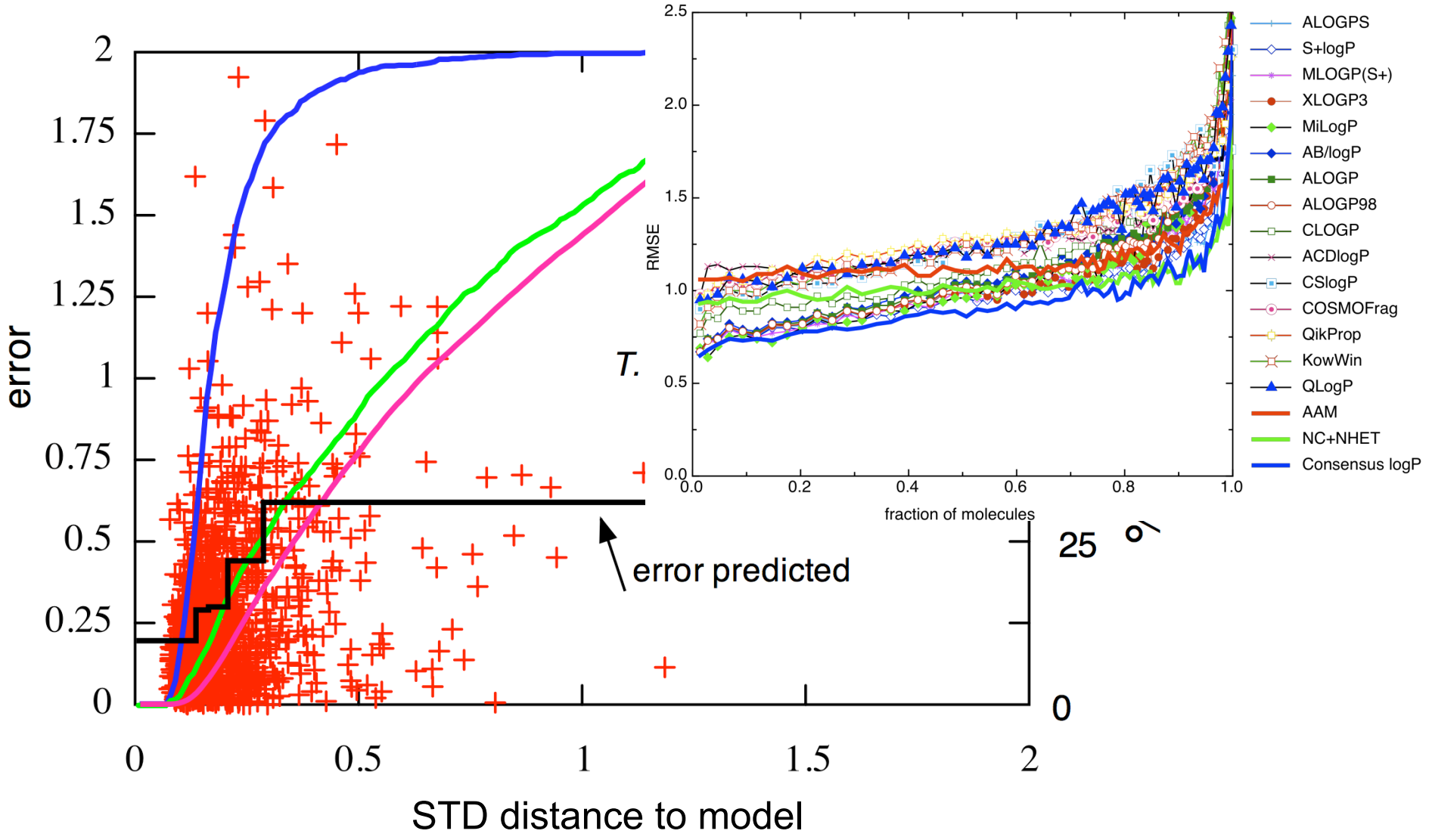
Mannhold, R. et al, *J. Pharm. Sci.*, 2009, 98(3), 861-893.

Global model: Accuracy of logP predictions as function of standard deviation (STD) of models



Mannhold, R. et al, J. Pharm. Sci., 2009, 98(3), 861-893.

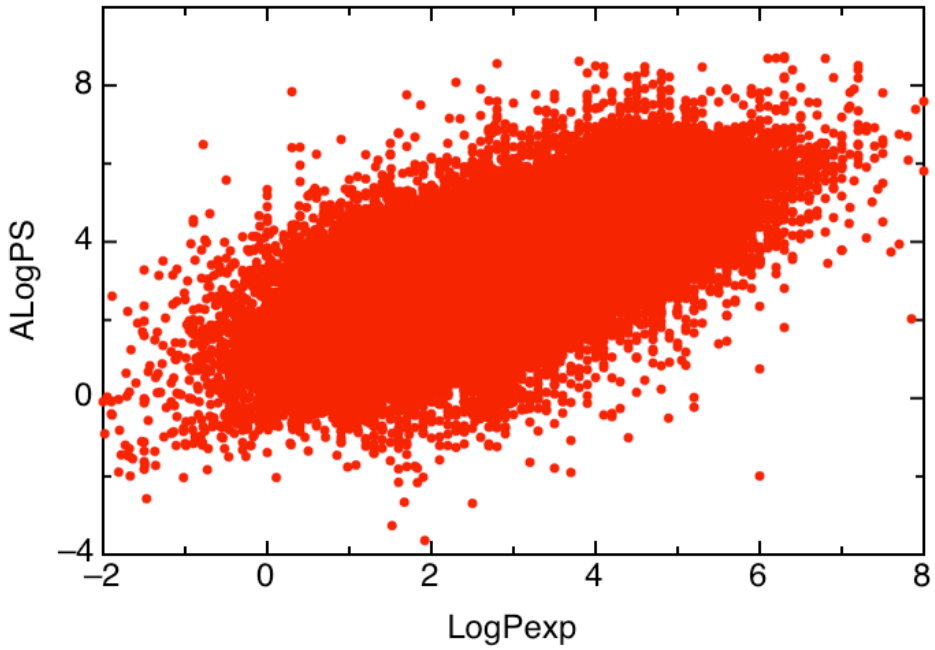
Prioritizations of chemicals: Accuracy of environmental toxicity prediction against *T. pyriformis*



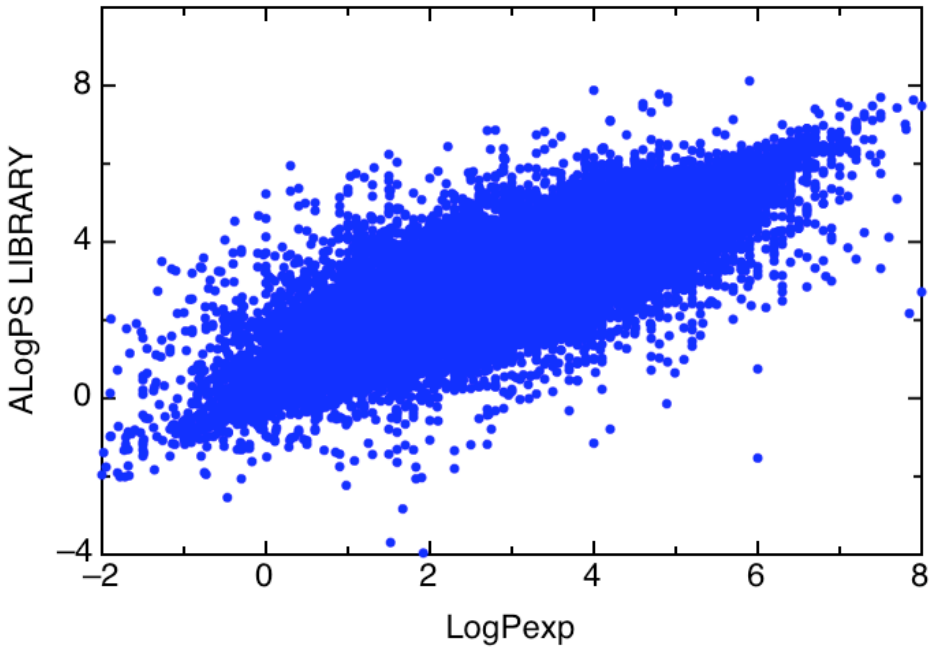
Tetko et al, *J Chem Inf Model*, 2008, 48(9):1733-46.

Local models: Instant learning of in-house data (Pfizer Inc.), *N*=95809 ?

ALOGPS Blind prediction

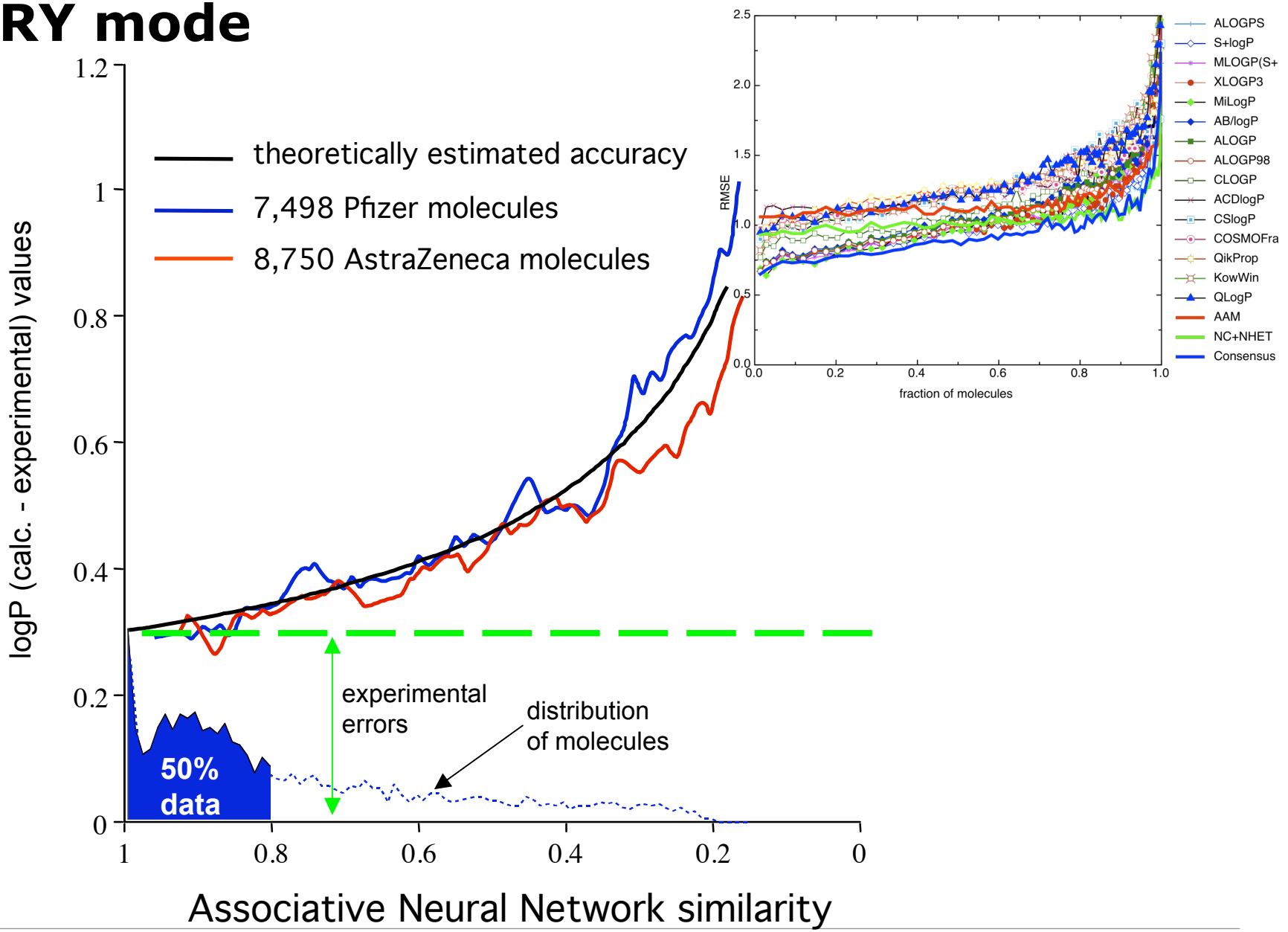


ALOGPS LIBRARY



RMSE=1.02 → RMSE=0.59

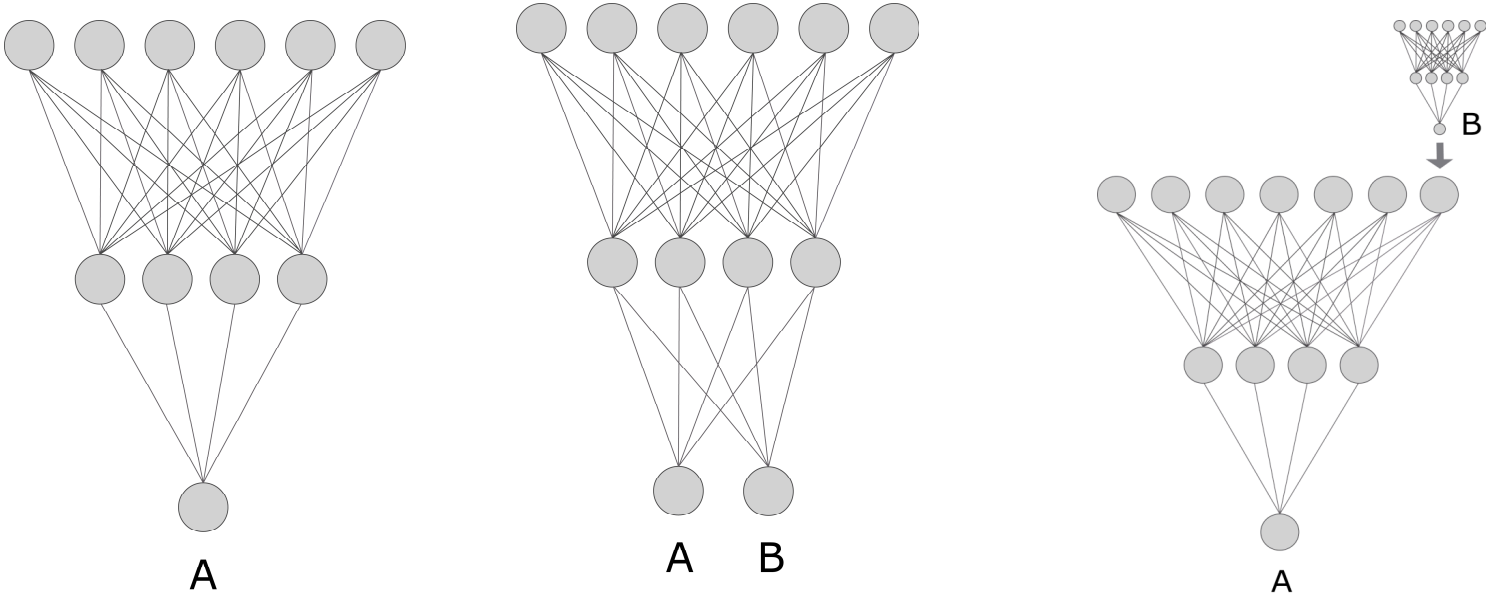
Local model: Accuracy of logP predictions in LIBRARY mode



The measurements are very expensive…

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

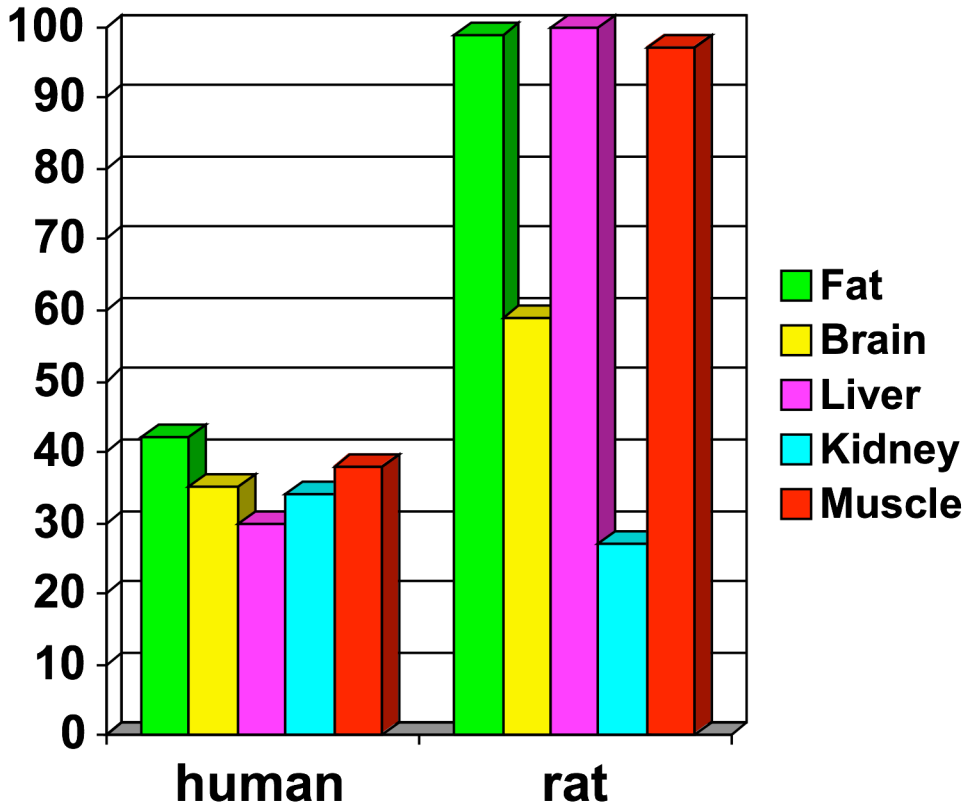
Multi-task learning



Multi-task learning: unequal number of data

Problem:

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



Varnek, et al, JCIM, 2009, 49(1):133-144.

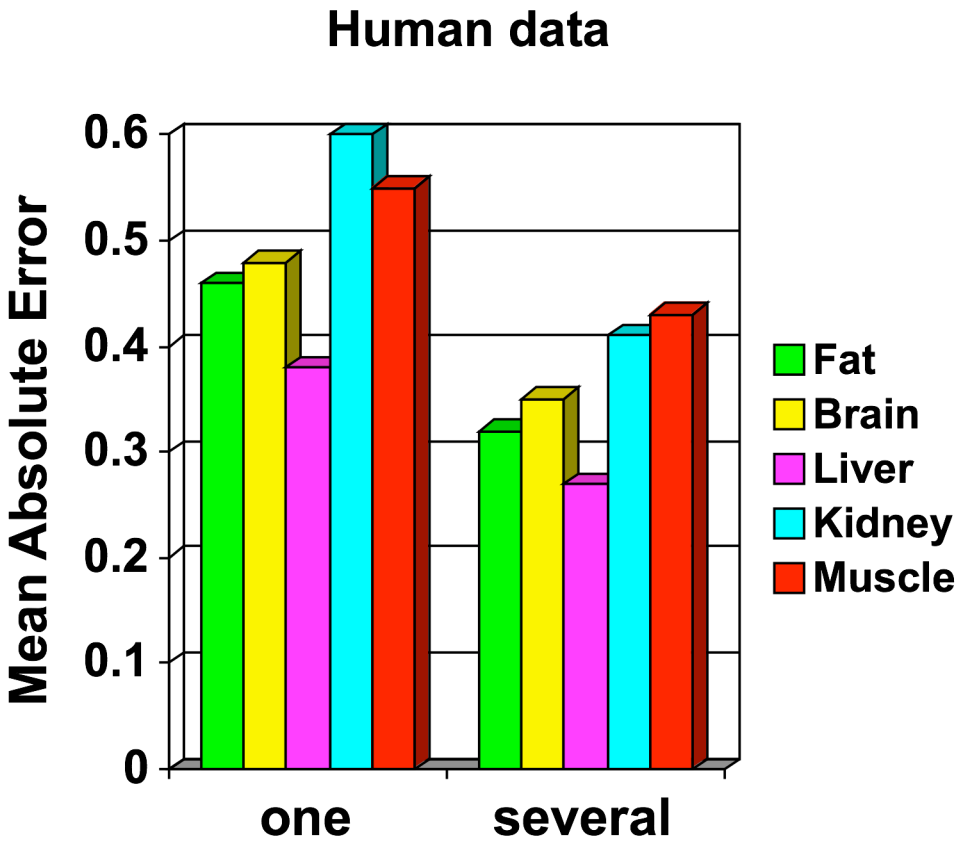
Multi-task learning can improve models for small sets

Problem:

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

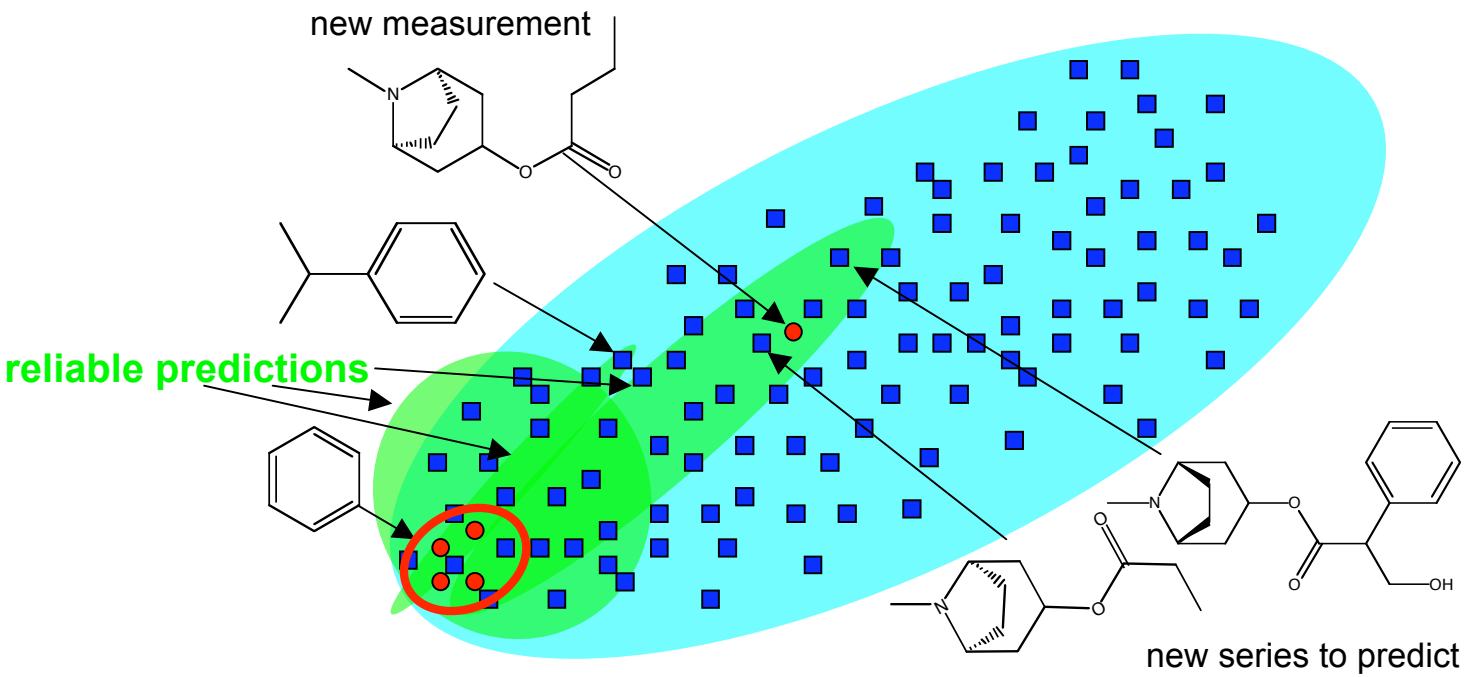
Results:

simultaneous prediction of several properties increased the accuracy of models



Varnek, et al, JCIM, 2009, 49(1):133-144.

ADMETox Challenges



We need methodology allows navigation in space of molecules with a confidence.

- ✓ to develop targeted (local) models to cover specific series.
- ✓ to reliably estimate which compounds can/can't be reliably predicted.
- ✓ to provide experimental design and to minimize costs of new measurements.
- ✓ to integrate data from similar properties.

Acknowledgement

Dr. G. Poda (Pfizer)

Dr. C. Ostermann (Nycomed)

Prof. R. Mannhold (Dusseldorf University)

+ many other colleagues & co-authors

Go-Bio BMBF

FP6 INTAS (VCCLAB, http://www.vcclab.org or just "google" **tetko**)