HelmholtzZentrum münchen
German Research Center for Environmental Health

Are Log P Calculators Accurate? Benchmarking on 96 000 Compounds

Igor V. Tetko,[1] Gennadiy I. Poda,[2] Claude Ostermann,[3] Raimund Mannhold[4,]*

1-Helmholtz Zentrum München, Neuherberg, Germany; 2-Pfizer Global R & D, Chesterfield, USA
3-Nycomed GmbH, Konstanz, Germany; 4-Heinrich-Heine-Universität, Düsseldorf, Germany

You can download this poster, preprint of the article and the ALOGPS program at http://www.vcclab.org

Introduction

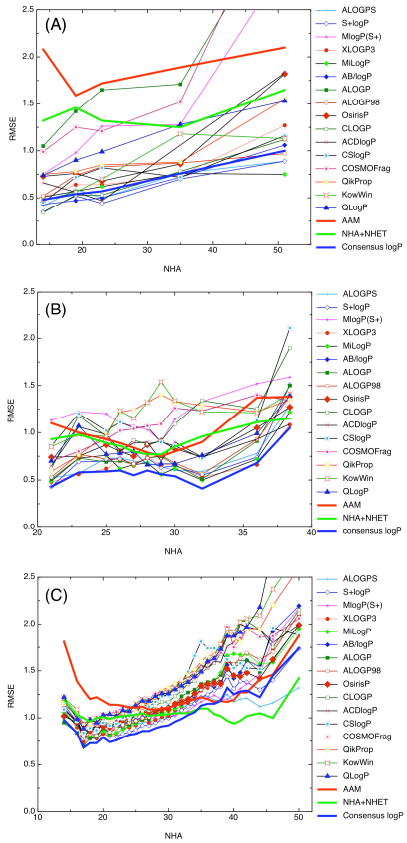
Lipophilicity, quantified as log *P*, is an important parameter daily monitored by medicinal chemists in drug discovery.

The application of log *P* prediction software tools deserves a comprehensive validity check on large, chemically diverse databases. Several evaluations of calculated *vs.* experimental log *P* values appeared in the literature. Most of them suffer from small size, limited chemical diversity, and/or inclusion of only a few log *P* calculation methods.

Here we review the state-of-the-art in development of log *P* prediction approaches including substructure- and property-based methods.

Then, we compare the predictive power of representative methods for one public (N = 266) and two *in-house* datasets from Nycomed (N = 882) and Pfizer (N = 95 809). 30 methods were tested for the public and 18 for the industrial datasets.[1]

[1] **Mannhold R, Poda G, Ostermann C, Tetko I (2008)** Calculation of Molecular Lipophilicity: State-of-the-Art and Comparison of Log *P* Methods on More Than 96,000 Compounds. *J. Pharm. Sci.*, in press.



Method performance as a function of the NHA for the public (A), Nycomed (B) and Pfizer (C) dataset. Each point on the graph shows the *RMSE* of the analyzed method for molecules with an NHA indicated on the x-axis (minimally 50 molecules where used per point for the public and the Nycomed and 500 molecules for the Pfizer dataset). The red bold line corresponds to the AAM. The blue bold line represents Consensus log *P*. The green bold line corresponds to the simple two-descriptor model (NC + NHET). Models with errors larger than those calculated with AAM fail to provide predictive model for molecules with a given NHA.

Log *P* programs used in benchmarking (Programs used for in-house datasets are given in bold)

Name	Provider	URL/E-mail
ABLogP v. 2.0	Pharma Algorithms, Lithuania/Canada	http://www.ap-algorithms.com
ABSOLV, LSER	Pharma Algorithms, Lithuania/Canada	http://www.ap-algorithms.com
ACDlogP v. 11	Advanced Chemistry Development, USA	http://www.acdlabs.com
ALOGP (DragonX 1.4)	Talete SRL, Milano, Italy	http://www.talete.mi.it
ALOGP98	Accelrys Software Inc., USA	http://www.accelrys.com
ALOGPS v. 2.1	Virtual Computational Chemistry Laboratory, Germany	http://www.vcclab.org
CLIP	University of Geneva, Switzerland	pierre-alain.carrupt@pharm.unige.ch
CLOGP v. 4.3 (v. 5.0)	BioByte Inc., USA	http://www.biobyte.com
COSMOFrag v. 2.3	COSMOlogic GmbH & Co. KG, Germany	http://www.cosmologic.de
CSlogP v. 2.2.0.0	ChemSilico LLC, USA	http://www.chemsilico.com
GBLOGP	Max Totrov, USA	maxt@molsoft.com
HINT	EduSoft, LC, USA	http://www.eduSoft-lc.com
KowWIN v. 1.67	Syracuse Inc., USA	http://www.syrres.com
LSER UFZ	Helmholtz Centre for Environm. Research UFZ, Germany	ralph.kuehne@ufz.de
MiLogP v. 2.2	Molinspiration Cheminformatics, Slovak Republic	http://www.molinspiration.com
MLOGP (DragonX 1.4)	Talete SRL, Milano, Italy	http://www.talete.mi.it
MLOGP(S+), ADMET 2.3	Simulations Plus, Inc., USA	http://www.simulations-plus.com
MolLogP	MolSoft LLC, USA	http://www.molsoft.com
NC+NHET	Virtual Computational Chemistry Laboratory, Germany	http://www.vcclab.org
OsirisP	Actelion, Switzerland	http://www.actelion.com
QikProp v. 3.0	Schrödinger, LLC, USA	http://www.schrodinger.com
QLOGP	University of Miami, USA	PBuchwald@med.miami.edu
QuantlogP	Quantum Pharmaceuticals, Russia	http://q-pharm.com
S+logP, ADMET 2.3	Simulations Plus, Inc., USA	http://www.simulations-plus.com
SLIPPER-2002	Institute of Physiol. Active Compounds, Russia	http://camd.ipac.ac.ru
SPARC	University of Georgia, USA	http://ibmlc2.chem.uga.edu
TLOGP	Upstream Solutions, Switzerland	http://www.upstream.ch
VEGA	University of Milan, Italy	http://www.ddl.unimi.it
VLOGP	TOPKAT, Accelrys Software Inc., USA	http://www.accelrys.com
XLOGP2	Inst. of Physical Chemistry, Peking University, China	ftp://ftp2.ipc.pku.edu.cn
XLOGP3	Institute of Organic Chemistry, Shanghai, China	http://sioc-cbg.ac.cn

Which Factors Determine Method Accuracy?

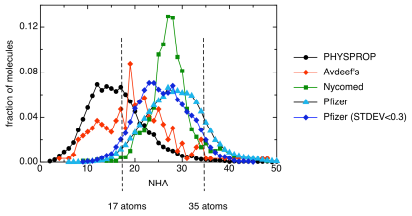
Previous studies showed accuracy in log P prediction to degrade linearly with increasing numbers of Non-Hydrogen Atoms (NHA). Our study confirmed these findings: for the public and the Pfizer dataset all methods produced larger RMSE for molecules with larger numbers of NHA. This dependency was less pronounced for the Nycomed set. The Pfizer set shows that prediction is challenging for molecules with NHA<17. Lower accuracy for these molecules could be due to a higher spread of log P values, which was minimal for molecules with 17-20 NHA.

Following the concept of MLOGP we developed a simple two-descriptor model (NC and NHET are the number of carbons and heteroatoms):

$$\log P = 1.46(\pm 0.02) + 0.11(\pm 0.001)\,NC - 0.11(\pm 0.002)\,NHET$$
$$N=95{,}823,\ RMSE=1.04,\ R^2=0.2$$

This simple equation was almost as accurate as the best models for the Pfizer set and shows an accuracy similar to that of other top methods tested on this set. This model had a rather constant accuracy of RMSE=1.00 that was independent on molecular complexity/size. It provided higher accuracy compared to any of the analyzed models for molecules with NHA>35.

Size-related distribution of molecules in the analyzed datasets



The PHYSPROP dataset and its various subsets were frequently used to develop computational methods analyzed in this study. The PHYSPROP and the public Avdeef dataset have considerably different distribution of molecules compared to the *in-house* datasets from Pfizer and Nycomed.

The first dashed line (17 atoms) indicates molecules for which the analyzed methods calculated lowest errors for the Pfizer dataset while the second line (35 atoms) indicates molecules for which all methods failed to calculate errors lower than that of the AAM model. The distribution of molecules with most confident predictions in the Pfizer dataset (*StdDev* < 0.3) is shifted toward molecules with smaller NHA.

Conclusions

Most methods produced reliable results for the public dataset, but for the *in-house* datasets only a few were superior to AAM. Among the best methods, ALOGPS, S+logP, XLOGP3, OsirisP, ALOGP, and ALOGP98 were consistent for both *in-house* datasets. Surprisingly, these methods include the classical approaches ALOGP and MLOGP. Best performances for the *in-house* datasets were calculated with Consensus log *P*.

Only Consensus log *P* calculated an *RMSE*<1.00 for the Pfizer set. Despite low average accuracy of prediction, the confidence of prediction and the analysis of molecular size serve to distinguish reliable from non-reliable predictions. E.g., molecules with NHA = 18-20 or molecules with a small *StdDev* had an *RMSE*<0.75 for Consensus log *P*. All models poorly predicted molecules with NHA > 30-35. Particular for these molecules we propose a simple two-descriptors-equation (based on NC and NHET) as a base-line estimation of log *P* with an *RMSE* ≈ 1.00.

When prediction accuracy is low, measuring log *P* is highly recommended. Such new values allow to improve prediction power via a user-training option without retraining as provided by, e.g. ALOGPS, SLIPPER, KowWIN, AB/logP, ACD/logP, XLOGP3, and good local models for improved predictions within homologous series can be derived.[2]

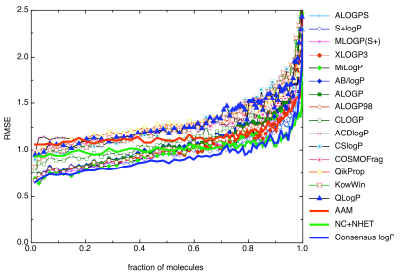
In summary, despite log *P* calculation is commonly viewed as simple, our analysis showed low prediction accuracy for most of the existing calculation methods.
Inaccurate log *P* prediction can mislead selection of chemical series to follow up from virtual screening, HTS triage and new analog design since it may lead to discarding potentially promising structures due to "poor" physico-chemical properties. Studying the confidence of log *P* prediction allows to distinguish reliable from non-reliable predictions and, thus, helps to avoid this problem.

2. Tetko, I. V.; Bruneau, P.; Mewes, H. W.; Rohrer, D. C.; Poda, G. I., Can we estimate the accuracy of ADME-Tox predictions? *Drug Discov. Today* **2006**, 11, (15-16), 700-707.

Reliable *versus* non-reliable predictions

Can we *a priori* distinguish reliable from non-reliable log P predictions? We estimated the prediction accuracy of the tested methods by their standard deviations (StdDev) compared to Consensus logP; accuracy gradually decreased with StdDev. For example, 10% of molecules with most confident predictions (i.e., StdDev < 0.3) had log P values predicted with an average RMSE of 0.73 and 0.75 for Consensus logP and ALOGPS, respectively. While, there was a shift in the distribution of these 10% of molecules towards smaller NHA, this set also included molecules with different molecular sizes. Thus, analysing the confidence of predictions allowed accurate predictions for molecules outside the range of optimal size, i.e. 16<NHA<21.

Comparison of methods using the total RMSE may lead to wrong conclusions about method performance. For example, ClogP produced an RMSE of 1.29 for the Pfizer set, i.e. had a larger RMSE compared to those of the NC+NHET and AAM models. However, for the same 10% threshold of the most confident predictions ClogP produced an RMSE of 0.87, indicating a better performance than the AAM (RMSE=1.07) and NC+NHET (RMSE=0.95) models. Thus, despite the fact that the NC+NHET equation had a lower RMSE compared to ClogP for the entire set, the latter model provided higher accuracy for the molecules with most confident predictions.



The *RMSE* of methods for Pfizer data as function of the fraction of molecules sorted along increasing *StdDev* values. Each point (at least 500 molecules) averages errors of methods with the same (or very similar) *StdDev* values. The first 10% of molecules have a *StdDev* < 0.3.

Benchmarking of methods using the public dataset

Datasets and definitions

30 methods compared in separate analysis of StarList and Non-Star List.

Entire dataset 266 molecules of a dataset from Alex Avdeef
StarList subset 223 molecules also present in BioByte list
Non-Star List subset 43 molecules outside BioByte list (NCEs)

Arithmetic Average Model (AAM): mean log P, used as model that predicts the same value for all dataset molecules; mean log P of entire dataset = 2.32, R^2=0 between predicted and experimental values

Rank III: models with root mean squared errors (RMSE) close to or larger than that of the AAM, i.e. models are non-predictive
Rank I: methods with statistical results identical or close to AB/LogP and ALOGPS (smallest RMSE for Star and Non-Star set)
Rank II: remaining models

Consensus logP: average of predicted log P from all rank I and II methods

Benchmarking Results

Models are ranked according to their predictive performance for both datasets. Prediction results for the Star set were statistically almost identical for the top-ranked methods AB/LogP, S+logP and ACD/logP.

For the Non-Star set, 17 methods non-significantly differed in performance from top-ranking ALOGPS. Its small size (N=43) excludes statistical differentiation between methods. For all tested calculation methods, the prediction performance was on average by 0.35 log units lower for the Non-Star set versus the Star set.

6 methods for the Star set and 8 methods for the Non-Star set showed results statistically similar to the AAM model. Thus, these methods failed to produce accurate results for these datasets.

Counting acceptable (Δ<0.5), disputable (Δ=0.5-1.0), and unacceptable (Δ>1.0) predictions allows accurate ranking of predictive power and quick checking of the general applicability of the methods.

Performance of algorithms for the public dataset

Method	Star set (N = 223) RMSE	rank	% within error range <0.5	0.5-1	>1	Non-Star set (N = 43) RMSE	rank	% within error range <0.5	0.5-1	>1
ABLogP[1]	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	8	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[2]	0.52	II	64	21	7	0.89	I	33	28	40
ALOGPS[1]	0.53	II	71	23	6	0.82	I	42	30	28
MLogP[1]	0.57	II	69	22	9	0.86	I	49	30	21
XLOGP[1]	0.62	II	60	30	10	0.89	I	47	23	30
KowWIN[1]	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[1]	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
OsiriP[1]	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[1]	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[1]	0.84	II	48	26	19	1.23	II	26	40	33
XLOGP2[1]	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[4]	1.04	II	47	27	26	1.24	II	28	30	42
CLIP[1]	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]	1.36	III	45	22	32	1.70	III	28	21	49
MLOGP(Dragon)[1]	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

[1]calculated at http://www.vcclab.org. [2]VLOGP results for 241 molecules inside and outside OPS (NOPS). [3]85 molecules excluded (75 with experimental values in TLOGP and 10 molecules that were not calculated). [4]ionized forms used to estimate log *P*. [5]14 molecules not calculated. [6]Zidovudine not calculated. [7]17 molecules not analyzed (error message).

Benchmarking of methods using *in-house* datasets

Datasets

18 methods compared in separate analysis of Pfizer and Nycomed set. Due to the large size of the *in house* datasets we included those methods that could be run in batch mode.

Pfizer dataset
95 809 experimental log P values generated at a number of legacy sites. We retained all experimental log P data points and added log D values for species predominantly neutral at pH 7.4. Multiple log P/D data points for compounds with identical structures were averaged. Salts were stripped off. The original tautomers and stereoisomers were preserved (as deposited in the proprietary database).

Nycomed dataset
The dataset of 882 compounds was prepared in a similar fashion.

Benchmarking Results

Mean log P values of 2.92 and 3.15 were calculated as AAMs for compounds from the Pfizer and Nycomed datasets, respectively.

We observed dramatic differences in method performance between public and *in house* datasets. Based on RSME, out of 18 methods only 8 (Nycomed) and 9 (Pfizer) produced results significantly better than AAM. Following methods coincided for both sets: ALOGPS and S+logP, XLOGP3, ALOGP, OsiriP, MiLogP and ALOGP98.

MLOGP(S+) and AB/LogP produced good results for the Pfizer dataset, but failed to produce models with RMSE significantly lower than that of AAM for the Nycomed dataset. QLOGP showed the opposite behavior.

In summary, atom-based approaches coupled with non-linear neural network learning techniques exhibited higher prediction accuracy in this validity check than fragment-based approaches. The highest accuracy for both *in house* datasets was calculated using Consensus logP.

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
OsiriP	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.

Institute of Bioinformatics and System Biology is supported by BMBF Go-Bio project 0313883.

HELMHOLTZ ASSOCIATION