Application of ALOGPS to predict 1-octanol/water distribution coefficients, *logP* & *logD*, of AstraZeneca in-house database

Igor V. Tetko and Pierre Bruneau

1-Biomedical Department, IBPC, Ukrainian Academy of Sciences, Murmanskaya 1,
Kyiv, 02094, Ukraine and 2 – AstraZeneca Centre de Recherche Parc Industriel Pompelle
BP 1050 – 51689 Reims Cedex 2

This is pre-print of the article (DOI link)
http://dx.doi.org/10.1002/jps.20217

Tetko, I. V.; Bruneau, P.
Application of ALOGPS to predict 1-octanol/water distribution coefficients, logP, and logD, of AstraZeneca in-house database,
J Pharm Sci, 2004, 93, 3103-10.

Current and contact address: Dr. Igor V. Tetko
Institute for Bioinformatics
GSF - Forschungszentrum für Umwelt und
Gesundheit, GmbH, Ingolstädter Landstraße 1,
D-85764 Neuherberg, Germany
Tel. +49-89-3187-3575 **Fax**. x.3585
itetko@vcclab.org

Running head: *logD* & *logP* prediction with ALOGPS 2.1 program

This manuscript contains: 19 pages including abstract, and 4 Tables.

Key words: *logP*, *logD*, computational ADME, drug design, drug like properties, associative neural network, QSAR, QSPR

ABSTRACT

The ALOGPS 2.1 was developed to predict 1-octanol/water partition coefficients, *logP*, and aqueous solubility of neutral compounds. An exclusive feature of this program is its ability to incorporate new user-provided data by means of self-learning properties of Associative Neural Networks. Using this feature, it calculated a similar performance, *RMSE*=0.7 and mean average error 0.5, for 2569 neutral *logP*, and 8122 *pH*-dependent $logD_{7.4}$, distribution coefficients from the AstraZeneca "in-house" database. The high performance of the program for the $logD_{7.4}$ prediction looks surprising, since this property also depends on ionization constants pK_a. Therefore, $logD_{7.4}$ is considered to be more difficult to predict than its neutral analog. We explain and illustrate this result and, moreover, discuss a possible application of the approach to calculate other pharmacokinetic and biological activities of chemicals important for drug development.

INTRODUCTION

The absorption of drugs is very important in rational drug design. Indeed, drugs have to cross a series of barriers either by a passive diffusion or a carrier-mediated uptake. The 1-octanol/water partition coefficient, *logP*, is accepted as one of the principal parameters to evaluate lipophilicity of chemical compounds that, to a large degree, determines these pharmaco-kinetic properties of drugs. LogP is used, for example, as a standard property determined for potential drug molecules in Lipinski's "rule of 5".[1] This property describes the partitioning of the neutral form only. If a molecule contains basic or acidic groups, it can be ionized and its distribution in 1-octanol/water mixture becomes *pH*-dependent. The *pH*-dependent distribution coefficient, *logD*, was shown to correlate with a number of biological parameters, such as the effective permeability in human jejunum,[2] blood-brain barrier permeability values[3] and volume of distribution.[4] Thus both coefficients are important parameters for drug development.[5]

The amount of reliable, publicly available compounds with *logP* data comprises tens of thousands of compounds.[6] These resources stimulated development of a number of programs to calculate it (see e.g., refs. [7,8]). The problem of predicting *logD* is more complicated. As a rule, it is computed from *logP* and pK_a assuming that only the neutral form of a molecule will partition into the organic phase as[8,9]

$$logD_{(pH)} = logP - log(1+10^{(pH-pKa)\delta i}), \qquad (1)$$

where $\delta_i = \{1,-1\}$ for acids and bases, respectively.

If several groups can be ionized, the equation is modified accordingly to incorporate correction terms for all of them. Thus the *logD* prediction potentially accumulates errors both due to the *logP* and pK_a predictions. The development of such approaches is further complicated due to an absence of large data sets with experimental *logD* and pK_a values.

As a result, only a few programs are available to evaluate the *logD*.[8] Prediction ability of such programs can be low. A recent evaluation of two commercial programs by Pfizer calculated *RMSE*=1.4–1.9 log units for a dataset of about 20000 compounds.[10,11] This accuracy is too low for practical use. Therefore, large pharmaceutical companies, e.g. AstraZeneca, Pfizer etc., have established their own centers to experimentally determine *logD* for their "in-house" series of compounds. Since its experimental determination is costly and time consuming, there is a need to develop computational approaches to estimate the *logD*, at least within the series of compounds. Unfortunately, the experimental data measured in the Pharma industry cannot be made available for public use due to confidentiality issues. Of course, the experimental values can be determined only for a small subset of compounds, while the computational approaches are used to calculate this property for compounds not yet synthesized. These computed values are used in different QSAR equations to predict the biological properties of the compounds and to prioritize the development process.

The ALOGPS program[12-14] (http://www.vcclab.org) was developed using Associative Neural Network (ASNN) method.[15,16] The ALOGPS program was programmed in C++ and it is available for Windows, LINUX and Mac Os X systems. The standalone versions are free of charge for academic and non-for-profit organizations and the Internet version is free for all users. The users from commercial organizations can request standalone demo versions on our web site. The current analysis was done on AMD 2.4 GH computer running Windows XP.

The ASNN provides a possibility to include new data without retraining the neural networks as a LIBRARY.[15] The LIBRARY dramatically improved prediction of the ALOGPS program for the *logP* prediction using in-house data of BASF,[15] Pfizer[17] and AstraZeneca.[18] Previous work demonstrated a dramatic improvement of the ASNN results in LIBRARY mode for prediction of NMR proton shifts.[19] The current study further confirms this result and demonstrates that the ALOGPS could be also used to predict the *pH* dependent distribution coefficient, *logD*.

DATA AND METHODS

The 1-octanol/water partition data contained in the internal database of AstraZeneca were used. The compounds with a comment about questionable purity or stability were removed. For the *logD* set, all the measurements at *pH* = 7.4 were selected, excluding the ones with an over range (< or >) information. The remaining data were averaged for identical ID's with the calculation of the range of measurements. The average standard deviation of 72 compounds having at least triplicate measurements was 0.37 log units. To obtain the final 'clean' database, the compounds having a range of measurements greater than 0.3 log unit were excluded. The final database contained 8122 $logD_{7.4}$ and 2569 *logP* measurements. The prediction performance of a standalone version of ALOGPS 2.1 was compared with CLOGP v. 4.71 (http://www.biobyte.com) and ACD Labs *logP*/*D* v. 7.0 (http://acdlabs.com/) programs.

The performance of programs was evaluated for the whole $logD_{7.4}$ dataset and also for different scaffolds of compounds. The classification of the $logD_{7.4}$ set in acid, base, zwitterions, and neutral categories was done according to two criteria. The first criterion used proprietary SMARTS definitions of common pre-classified sub-structures likely to be ionised at *pH* 7.4. The second criterion used the relative values of calculated *logP*, $logD_{7.4}$, and $logD_{6.5}$ computed with the ACD Labs program to determine if a compound was an acid ($logD_{7.4} – logD_{6.5} < - 0.1$), a base ($logD_{7.4} – logD_{6.5} > 0.1$), or neutral ($logP – logD_{7.4} \leq 0.1$ or $logP – logD_{6.5} \leq 0.1$). The other compounds were classified as zwitterionics. The compound was retained in its respective class only if the criteria were identical (i.e., if a compound was considered as basic according to SMART but it had the same *logP* and *logD* for *pH* 6.5 and 7.4, it was considered as mixed and excluded from the analysis). After this analysis, the bases and acids were further split into two categories according to the number of acidic or basic functions using the proprietary SMARTS definitions.

The ALOGPS 2.1 program was developed using ensemble of 64 neural networks, 12908 molecules with experimental *logP* values from PHYSPROP database, and 75 input parameters comprising two atom counts and E-state indices.[12,13] The ALOGPS used Parzen-window regression to correct a final prediction of the target molecule using errors of its nearest neighbors. The neural networks calculated 64 output values for the analyzed molecule and molecules from the LIBRARY. The molecules from the LIBRARY with maximum Spearman rank correlations to the analyzed molecule were selected as the nearest neighbors of the target molecule.[12,15] The parameters of the Parzen-window regression were optimized by minimization of the root mean squared error for the LIBRARY molecules. A more detailed description of the ALOGPS 2.1 can be found elsewhere.[12,15]

RESULTS

The ALOGPS program calculated the *logP* set with lower errors (Table 1), compared to other approaches in the blind prediction ("as is"), (i.e. when no one experimental value was used to improve the program performance). If all compounds were used in the LIBRARY, the accuracy of the program estimated by the Leave-One-Out (LOO) method significantly increased. In order to better validate the performance of the method, 50% of compounds selected by chance ("random library") were used as the LIBRARY, and the program predicted remaining 50% of compounds. The results calculated for both subsets were practically identical, thus indicating that the LOO results provided an unbiased estimation of the program performance.

Since the ALOGPS was developed to predict distribution coefficient of neutral compounds only, its "as is" prediction for the *logD* set, *RMSE*=1.63, was significantly more disperse than for the *logP* set, *RMSE*=0.84 (Table 2). The LIBRARY made it possible to incorporate the *pH* dependency and thereby dramatically increased the ALOGPS prediction performance. The obtained accuracy, *MAE* < 0.50 log units, is sufficient to evaluate other compounds from the same series for most projects. The new *logP*/*logD* models are developed by the ALOGPS in completely automatic way using ca. 5-10 minutes of CPU time. This speed allows an easy incorporation of new data and thus enhances model quality following daily experimental measurements.

The calculated results for different chemical subsets of the AstraZeneca *logD* set are shown in Table 3. The ALOGPS program applied in the "as is" mode calculated low errors (*MAE*=0.68) for neutral compounds but its performance for the charged compounds was much lower. The program results were significantly improved in the LIBRARY mode for all subsets. The ALOGPS program decreased its errors approximately three-fold for charged compounds. Its performance was highest for neutral compounds and compounds with one base. The high performance of the ALOGPS for the later series was presumably due to its large size. A further improvement (two last columns of the Table 3) was observed when compounds from each subset were analyzed in the LIBRARY mode separately. This later analysis created more homogeneous data sets for the ALOGPS program, and it decreases the rate of incorrect nearest neighbors, particularly for compounds from series with small number of chemicals as explained for the test series analyzed below.

In order to illustrate our results and to explore limitations of the ALOGPS program, we applied it to series of 31 compounds with experimental *logD* values.[20] The error of the ALOGPS program, *MAE*=1.79, in "as is" mode decreased to *MAE*=0.49 in the LIBRARY mode for this set. Thus the ALOGPS performance for this and AstraZeneca data sets in "as is" and LIBRARY mode was very similar. In "as is" mode the program calculated *MAE*< 0.50 log units for the core structures (1, 2), nitriles (3-6), amino (7), carbox-amides (8-11), sulfonamide, and methanesulfonylamido (13) derivatives. These classes of compounds are either neutral or uncharged at *pH* 7.4 since their *logP≈logD*. We will refer to these compounds as the "non-charged" subset. Other derivative series form the "charged" subset and are presumed partially or completely ionized at the same *pH*. The ionization produced significant shifts in the *logD* values of these compounds, and ALOGPS overestimated their lipophilicity values. The molecules were sorted in Table 4 according to the increase of their errors, and are approximately proportional to their ionization strength at *pH* 7.4. The improvement of the ALOGPS program in the LIBRARY mode is limited by its ability to detect nearest neighbors of the target compound that have similar *logP-logD* shifts. The ALOGPS used *k*=4 nearest neighbors in the LIBRARY mode that are listed in the Table 4. With the exception of carboxylic acids that had all nearest neighbors from its own derivative series, the ALOGPS determined more than 40% (34 out of 80) of the nearest neighbors from the same series of the target compound. Thus the similarity measure in the space of neural networks made it possible to detect nearest neighbors across four different core series used in this analysis. Moreover, more than 70% of all neighbors (58 out of 80) were from the same "charged" and "non-charged" set to which the target compound itself belonged. Thus nearest neighbor compounds in the lipophilicity space had similar *logP-logD* shifts that explains high prediction ability of ALOGPS in this mode. This result indicates that the same chemical groups important for ionization properties of molecules greatly determine the lipophilicity of the neutral compounds.

The ALOGPS program in the LIBRARY mode significantly decreased the errors for the compounds from the "charged" set. The most striking decrease of the error was observed for carboxylic acid derivatives. For these compounds, *MAE* decreased ten-fold from 3.4 to 0.33 log units. An improvement of the prediction for this set significantly enhanced the total performance of the ALOGPS method. The ALOGPS performance for the remaining set of 20 compounds was lower, *MAE*=0.57. It is interesting to note that if carboxylic acids were excluded, the error for these 20 molecules decreased to *MAE*=0.47. Of course, this LIBRARY provided a low prediction, *MAE*=2.76 for the carboxylic acids.

The nearest neighbors were detected sometimes incorrectly. For example, the prediction ability of ALOGPS the non-charged sub-set decreased from *MAE*=0.33 in "as is" to *MAE*=0.49 in the LIBRARY mode. This increase was mainly due to a detection of some compounds from the "charged" set, particularly carboxylic acids, as the nearest neighbors of molecules from the "un-charged" set. If both these sets were used in the LIBRARY mode separately, the program calculated lower errors, *MAE*=0.23 and

MAE=0.43 for "charged" and "non-charged" subsets, respectively. The detection of false nearest neighbors can have a large impact on chemical sets with small number of compounds. For example, the ALOGPS using LIBRARY of 8081 compounds calculated the largest errors for compounds with two or more acids and zwitterions. Both of these sets contained the lowest number of compounds. The ALOGPS performance for them had the largest improvement when each set was analyzed separately, as shown in two last columns of Table 3.

How many molecules are required for the LIBRARY mode to accurately predict *logD*? This question is impossible to answer in general because it completely depends on the diversity of molecules in the analyzed set, the accuracy of experimental measurements, and the required accuracy. For simplicity let us restrict our analysis to the carboxylic acids. In the "as is" mode, the ALOGPS error was *MAE*=3.40 log units. If only one molecule (cpd. 21) was used in the LIBRARY, the performance of the program for the remaining molecules increased five fold to *MAE*=0.74. The LIBRARY with two molecules (21 & 27) and four molecules (21, 22, 27 & 30) calculated *MAE*=0.66 and *MAE*=0.38, respectively. The use of all 11 molecules calculated LOO *MAE*= 0.31. Thus depending on project requirements the user can decide which number of experimental measurements is sufficient for the need of his particular project.

DISCUSSION

Similar high performance of ALOGPS for prediction of both *logP* and *logD* coefficients using LIBRARY is surprising. Indeed, since the latter coefficient is *pH* dependent, one could assume an explicit incorporation of the pK_a prediction module and/or use of Eq. 1 is required. This was not necessary because only a "one point" estimation of the *logD* coefficient for a fixed *pH* 7.4 was calculated.

An understanding of the high accuracy of the ALOGPS requires an analysis of the ASNN, as well as physico-chemical principles distinguishing *logP* and *logD* coefficients. Eq 1 indicates that the *logD* can be calculated from the logP by a simple correction. A presence of ionizable groups produces a shift in the distribution coefficient, while the general dependency of this parameter on the molecular structure remains similar for both neutral and *pH*-dependent coefficients. In general, not one but several ionizable groups, both acids and bases, can be present for a single chemical. The interaction of such groups could produce non-linear effects that may invalidate Eq. 1 and complicate calculation of *logD*. However, if a series of compounds has a similar set of ionizable groups, the shift will be about the same for the whole series. Therefore, if a method could correctly determine the nearest neighbors of the analyzed compound amid compounds with experimental values, it could reasonably estimate its *logD* value. The ASNN provides an accurate detection of the nearest neighbors in the property-based space (i.e., lipophilicity space for the ALOGPS program) as shown in our previous publications.[12,15,16] Therefore, ALOGPS correctly determined the required correction terms and provided high prediction performance.

The detection of nearest neighbors in the current study was performed in the lipophilicity space, since the ALOGPS program was developed to predict *logP*. It is possible that application of the ASNN in the space of ionizable groups, following a development of a program to predict pK_a, may provide better results. The pK_a space would probably detect correct nearest neighbors for tetrazole (19), which erroneously enclosed three nearest neighbors from the "un-charged" set and did not comprise another tetrazole (18). This program may have some other limitations that could not be foreseen in advance.

It was shown that a pre-processing of molecules by separation on "charged" and "un-charged" sets, or exclusion of large dominating series (such as carboxylic acids in Table 4 or separate analysis of each series in Table 3) before their use in the LIBRARY mode, may further improve prediction ability of the ALOGPS. Indeed, such pre-processing decreases heterogeneity of molecules in the LIBRARY, and the correction procedure works more accurately. A separation of initial data set and creation of "targeted" libraries corresponds to incorporation of *a priori* information about the expected chemical properties of the series in question. For example, the carboxylic LIBRARY is expected to be used for compounds whose *logD* values are dominated by carboxylic acids. The "charged" LIBRARY assumes that the prediction set will contain molecules predominantly charged under experimental conditions. The targeted libraries, contrary to static *logD* prediction methods, may provide a flexible tool that can be used by an expert (i.e. computational chemist) to fine-tune the performance of the ALOGPS program for the investigated series of compounds. The success of the "targeted" libraries would depend, as is the case for all computational approaches, on the correspondence of chemical spaces covered by the LIBRARY and the test set. For example, an attempt to predict carboxylic acids using LIBRARY that do not contain such molecules will fail. Thus the ALOGPS program used in the LIBRARY mode represents both a simple (i.e., the user uses all his compounds as one LIBRARY) and advanced tool (the expert can design his own targeted LIBRARY and get better results for his series). Further development of the ALOGPS program will make it possible to create the targeted libraries in an automatic mode.

An important remark concerns results for chemical series reported in Table 3. It should be mentioned that the ALOGPS program was used for this analysis with all compounds that were not *a priory*, before the analysis, separated on classes by exception of the two last columns discussed in previous paragraph. The classification on series was used mainly to provide a deeper insight into performance of the ALOGPS program for different scaffolds of compounds. Thus the performance of ALOGPS program does not critically depend on SMARTS scripts used for identification of acid and basic groups or prediction of *logD* values using the ACDlogD program. However, as it is demonstrated in the last two columns of Table 3, this classification may improve in its prediction ability. To perform analysis of ALOGPS performance for different groups of chemicals, we had to identify them. Since it was not feasible to classify all 8122 compounds by manual inspection of each molecule, we developed the automatic scheme described in the method section. The ACD classification was done by comparing ACDLogP, ACDLogD7.4, and ACDLogD6.5 values, and it was not 100% proof. The classification using predefined SMARTS definition was also not 100% proof. By combining the two lines of evidence and retaining only compounds that are assigned in the same category by both criteria we ensure a much cleaner classification. While application of ALOGPS program to each classified series improved its prediction ability this result does not influence other conclusions of the article.

A limitation of the current version of the program is the use of a single number of nearest neighbors for all compounds. For example, a use of *k*=1 instead of *k*=4 would increase the performance of the method for prediction of the core structures (cmpds 1 and 2) without a requirement to separate them in “non-charged” set LIBRARY. At the same time, *k*=6 instead of *k*=4 could further decrease error for the carboxylic acids set to

MAE=0.31. The development of an averaging procedure that uses *k* numbers selected according to the size and properties of each particular series may provide another way to further improve prediction of the method.

The performance of the ALOGPS program was significantly improved compared to other analyzed programs in the LIBRARY mode, i.e., when some fresh user-specific data were used in the memory of neural networks. This option is very easy to employ within the ALOGPS program, and its usage is clear to the end-user. Actually, by our experience, the main effort is preparation of the input data file. This file contains SMILES codes followed by the respective experimental *logD* values of molecules, which is uploaded into the ALOGPS program. Calculation of the LIBRARY is performed in completely automatic mode. The calculation of LIBRARY for ca. 10,000 compounds requires less than 10 minutes of CPU time of AMD 2.4 GHz computer. Thus in less than 10 minutes the user can create his own LIBRARY and dramatically improve prediction for his series. The LIBRARY remains in program memory only. If a user wants to analyze his compounds in "as is" mode he just needs to re-start the program. Therefore, the ALOGPS does not have any difficulty in the user training that is, according to one of the reviewers, is most concern to people working in this area with similar software packages.

It should be mentioned that ACD Labs software has a user training option that could improve prediction ability of this program for the *logD* prediction. However, the version that was available to us did not include an option for batch analysis. Thus we were unable to provide a fair comparison of both programs in similar conditions. Thus no conclusions on the relative performances of the *logD* software in "trained" mode can be drawn from this work. However, an example of such comparison for the *logP* prediction can be found elsewhere.[12]

Practically all ADME/T and many biological properties of compounds important for drug development critically depend on their lipophilicity. One possibility is to measure or predict 1-octanol/water partition coefficients and to use these values explicitly in the model. At the same time, these activities also depend on other physico-chemical properties of compounds that are difficult to formalize and to explicitly use in the model. These properties introduce shifts of the ADME/T values similar to the influence of pK_a contributions on *logD* in Eq. 1. If experimental ADME/T values are measured for some compounds, a detection of the nearest neighbors of a target compound in the lipophilicity space will make it possible to estimate its activity in a model-free way.

This study was partially supported with "Virtual Computational Chemistry Laboratory" INTAS grant 00-0363. We thank reviewers, Drs. A. Davis and N. McElroy from AstraZeneca and the associate editor, Prof. B. Anderson, for their remarks that significantly improved our study.

REFERENCES

1. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ 2001. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Adv Drug Deliv Rev 46(1-3):3-26.

2. Winiwarter S, Bonham NM, Ax F, Hallberg A, Lennernas H, Karlen A 1998. Correlation of human jejunal permeability (in vivo) of drugs with experimentally and theoretically derived parameters. A multivariate data analysis approach. J Med Chem 41(25):4939-4949.

3. Liu X, Tu M, Kelly RS, Chen C, Smith BJ 2004. Development of a computational approach to predict blood-brain barrier permeability. Drug Metab Dispos 32(1):132-139.

4. Poulin P, Theil FP 2002. Prediction of pharmacokinetics prior to in vivo studies. 1. Mechanism-based prediction of volume of distribution. J Pharm Sci 91(1):129-156.

5. Kerns EH 2001. High throughput physicochemical profiling for drug discovery. J Pharm Sci 90(11):1838-1858.

6. Tetko IV 2003. The WWW as a tool to obtain molecular parameters. Mini Rev Med Chem 3(8):809-820.

7. Japertas P, Didziapetris R, Petrauskas A 2003. Fragmental methods in the analysis of biological activities of diverse compound sets. Mini Rev Med Chem 3(8):797-808.

8. Livingstone DJ 2003. Theoretical property predictions. Curr Top Med Chem 3(10):1171-1192.

9. Hansch C, Leo AJ. 1979. Subsistent constants for correlation analysis in chemistry and biology. ed., New York: Wiley.

10. Lombardo F, Shalaeva MY, Bissett BD, Chistokhodova N. LogP2004 The 3rd Lipophilicity Symposium, Zurich, Switzerland, 2004, pp L-22.

11. Tetko IV, Poda GI 2004. Application of ALOGPS 2.1 to Predict LogD Distribution Coefficient for Pfizer Proprietary Compounds. *in prep.*

12. Tetko IV, Tanchuk VY 2002. Application of associative neural networks for prediction of lipophilicity in ALOGPS 2.1 program. J Chem Inf Comput Sci 42(5):1136-1145.

13. Tetko IV, Tanchuk VY, Villa AE 2001. Prediction of n-octanol/water partition coefficients from PHYSPROP database using artificial neural networks and E-state indices. J Chem Inf Comput Sci 41(5):1407-1421.

14. Tetko IV, Tanchuk VY, Kasheva TN, Villa AE 2001. Estimation of aqueous solubility of chemical compounds using E-state indices. J Chem Inf Comput Sci 41(6):1488-1493.

15. Tetko IV 2002. Neural network studies. 4. Introduction to associative neural networks. J Chem Inf Comput Sci 42(3):717-728.

16. Tetko IV 2002. Associative neural network. Neural Proc Lett 16(2):187-199.

17. Tetko IV, Tanchuk VY, Poda GI. LogP2004 The 3rd Lipophilicity Symposium, Zurich, Switzerland, 2004, pp C-30.

18. Tetko IV, Bruneau P. LogP2004 The 3rd Lipophilicity Symposium, Zurich, Switzerland, 2004, pp C-17.

19. Binev Y, Corvo M, Aires-De-Sousa J 2004. The impact of available experimental data on the prediction of (1)h NMR chemical shifts by neural networks. J Chem Inf Comput Sci 44(3):946-949.

20. Fichert T, Yazdanian M, Proudfoot JR 2003. A structure-permeability study of small drug-like molecules. Bioorg Med Chem Lett 13(4):719-722.

Table 1. Performance of methods for prediction of *logP* dataset

program	n	*MAE*	*RMSE*	r^2	% of compounds within *RMSE*			
					0–0.3	0–0.5	0–1.0	0 – 2.0
ACDLogP 7.0	2569	0.86	1.20	0.40	28	40	67	82
ClogP 4.72	2569	0.71	1.07	0.49	38	55	79	90
ALOGPS 2.1								
"as is" blind prediction	2567	0.60	0.84	0.53	35	54	83	97
LOO, all compounds are used in the LIBRARY	2567	0.45	0.68	0.65	52	73	92	99
LOO for 50% compounds used in the LIBRARY	1268	0.46	0.68	0.64	47	69	93	100
prediction of 50% remaining compounds	1299	0.46	0.67	0.64	47	70	93	100

MAE is mean average absolute error; *RMSE* is root mean squared error; r^2 is Pearson correlation coefficient.

Table 2. Performance of methods for prediction of *logD* dataset

program	*n*	*MAE*	*RMSE*	r^2	% of compounds within *RMSE*			
					0–0.3	0–0.5	0–1.0	0 – 2.0
ACDLogD 7.0	7845	1.05	1.62	0.42	25	39	64	86
ALOGPS 2.1								
“as is” blind prediction	8122	1.27	1.63	0.23	18	28	50	77
LOO, all compounds are used in the LIBRARY	8081	0.49	0.70	0.60	45	65	88	98
LOO for 50% compounds used in the LIBRARY	4060	0.55	0.78	0.53	42	60	85	97
prediction of 50% remaining compounds	4062	0.53	0.77	0.53	42	61	86	98

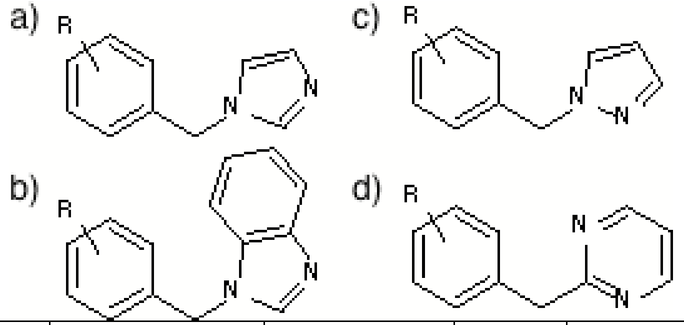
MAE is mean average absolute error; *RMSE* is root mean squared error; r^2 is Pearson correlation coefficient.

Table 3. Results calculated for different chemical series

compounds	*n*	ACD *logD*		ALOGPS 2.1					
				as is[1]		LIBRARY[2]		libraries[3]	
		MAE	*RMSE*	*MAE*	*RMSE*	*MAE*	*RMSE*	*MAE*	*RMSE*
one acid	760	1.40	2.10	1.94	2.19	0.64	0.95	0.51	0.72
two and more acids	54	3.40	4.80	1.27	1.54	0.66	0.96	0.58	0.89
one base	2616	0.85	1.17	1.57	1.82	0.47	0.64	0.47	0.65
two and more bases	477	1.11	2.04	1.33	1.55	0.51	0.71	0.49	0.68
zwitterions	124	2.44	3.41	2.18	2.57	0.66	0.90	0.51	0.78
neutral	2645	0.84	1.23	0.68	0.93	0.46	0.66	0.44	0.65

1-blind prediction; 2-LOO results for LIBRARY of 8081 compounds from Table 2; 3-LOO results for each chemical series used as a LIBRARY (i.e., LOO results for compounds with one acid were calculated using LIBRARY of 760 molecules)

Table 4. ALOGPS 2.1 results for logD[a] prediction of four chemical series



n	core structure	fragment	*logD*	"as is"	LIBRARY LOO	nearest neighbors
1	a	H	1.78±0.25	1.62	0.96	2,8,24,28
2	b	H	2.91±0.05	3.21	2.72	1,7,4,16
				0.23[b]	0.51[b]	
3	a	m-CN	1.01±0.04	1.33	1.19	5,4,8,10
4	b	m-CN	2.47±0.01	2.82	2.70	5,3,2,9
5	c	m-CN	1,74±0.02	1.64	1.30	3,4,6,8
6	d	m-CN	1.46±0.03	1.74	1.46	5,12,3,4
				0.26[b]	0.21[b]	
7	b	m-NH2	1.92±0.06	2.47	1.57	9,25,2,10
8	a	m-CONH2	-0.01±0.10	0.4	0.32	10,9,1,7
9	b	m-CONH2	1.7±0.02	1.9	1.51	10,8,7,21
10	c	m-CONH2	0.71±0.05	0.7	-0.14	8,9,21,7
11	d	m-CONH2	0.28±0.01	0.89	0.61	6,8,9,22
				0.36[b]	0.41[b]	
12	b	p-SO2NH2	1.04±0.09	1.86	1.40	13,14,6,16
13	b	m-NHSO2CH3	2.08±0.04	2.28	1.23	12,25,23,10
				0.51[b]	0.61[b]	
14	b	m-CH2NH2	0.21±0.03	2.05	0.58	15,2,12,16
15	c	m-CH2NH2	-0.77±0.00	0.88	-0.95	14,16,2,13
16	d	m-CH2NH2	-1.11±0.03	1.07	-0.04	2,15,23,14
17	d	m-C(NH)NH2	-1.03±0.03	0.81	-2.26	29,1,23,18
				1.88[b]	0.71[b]	
18	b	m-tetrazole	0.27±0.06	1.78	-0.46	19,27,7,31
19	d	m-tetrazole	-1.28±0.01	0.68	0.23	10,9,8,27
				1.73[b]	1.12[b]	
20	b	p-SO3H	-1.78±0.02	1.09	-0.57	15,14,16,2
				2.87[b]	1.21[b]	
21	a	p-CH2CO2H	-2.37±0.11	1.66	-1.88	24,28,25,22
22	b	p-CH2CO2H	-1.13±0.11	3.16	-0.35	25,21,28,24
23	c	p-CH2CO2H	-1.8±0.08	1.44	-1.58	26,30,29,24
24	a	p-CO2H	-2.1±0.08	1.23	-2.45	28,21,25,22
25	b	p-CO2H	-0.61±0.01	2.74	-1.04	22,21,28,24
26	c	p-CO2H	-1.65±0.02	1.31	-1.79	30,23,29,24
27	d	p-CO2H	-2.2±0.14	1.37	-2.10	31,22,25,21
28	a	m-CO2H	-2.07±0.10	1.2	-2.50	24,21,25,22
29	b	m-CO2H	-0.72±0.11	2.43	-0.62	30,26,23,24
30	c	m-CO2H	-1.62±0.13	1.28	-1.84	26,23,29,24
31	d	m-CO2H	-1.98±0.12	1.33	-2.36	27,22,25,21
				3.4[b]	0.33[b]	
Total	*MAE*			1.79	0.49	

[a]The $logD_{7.4}$ values are means of six experiments ± standard deviation from ref. [20]. [b]*MAE* calculated for each corresponding derivative series.