Prediction of *n*-Octanol/Water Partition Coefficients from PHYSPROP Database Using Artificial Neural Networks and E-State Indices

Igor V. Tetko,*[,†,‡] Vsevolod Yu. Tanchuk,[‡] and Alessandro E. P. Villa[†]

Laboratoire de Neuro-Heuristique, Institut de Physiologie, Université de Lausanne, Rue du Bugnon 7, Lausanne, CH-1005, Switzerland, and Biomedical Department, Institute of Bioorganic & Petroleum Chemistry, Murmanskaya 1, Kiev-660, 253660, Ukraine

Received March 8, 2001

A new method, ALOGPS v 2.0 (http://www.lnh.unil.ch/~itetko/logp/), for the assessment of *n*-octanol/water partition coefficient, log *P*, was developed on the basis of neural network ensemble analysis of 12 908 organic compounds available from PHYSPROP database of Syracuse Research Corporation. The atom and bond-type E-state indices as well as the number of hydrogen and non-hydrogen atoms were used to represent the molecular structures. A preliminary selection of indices was performed by multiple linear regression analysis, and 75 input parameters were chosen. Some of the parameters combined several atom-type or bond-type indices with similar physicochemical properties. The neural network ensemble training was performed by efficient partition algorithm developed by the authors. The ensemble contained 50 neural networks, and each neural network had 10 neurons in one hidden layer. The prediction ability of the developed approach was estimated using both leave-one-out (LOO) technique and training/test protocol. In case of interseries predictions, i.e., when molecules in the test and in the training subsets were selected by chance from the same set of compounds, both approaches provided similar results. ALOGPS performance was significantly better than the results obtained by other tested methods. For a subset of 12 777 molecules the LOO results, namely correlation coefficient $r^2 = 0.95$, root mean squared error, $RMSE = 0.39$, and an absolute mean error, $MAE = 0.29$, were calculated. For two cross-series predictions, i.e., when molecules in the training and in the test sets belong to different series of compounds, all analyzed methods performed less efficiently. The decrease in the performance could be explained by a different diversity of molecules in the training and in the test sets. However, even for such difficult cases the ALOGPS method provided better prediction ability than the other tested methods. We have shown that the diversity of the training sets rather than the design of the methods is the main factor determining their prediction ability for new data. A comparative performance of the methods as well as a dependence on the number of non-hydrogen atoms in a molecule is also presented.

INTRODUCTION

The *n*-octanol/water partition coefficient is the ratio of the concentration of a chemical in *n*-octanol to that in water in a two-phase system at equilibrium. The logarithm of this coefficient, log *P*, has been shown to be one of the key parameters in quantitative structure−activity relationship studies, and it is used to provide invaluable information for the overall understanding of the uptake, distribution, biotransformation, and elimination of a wide variety of chemicals. Thus, there is a need to have reliable programs that can be used to predict lipophilicity of chemical compounds by their structure.

Many approaches have been developed for the prediction of log *P* based on nonexperimental structural parameters. Most of these methods use substructures (fragments/atom fragments)[1−6] or quantum chemical parameters (charges, electronic potentials, molecular volumes, shape, etc.)[7−12] and multiple linear regression analysis to fit models to experimental data. There are also a number of atom-additive methods.[13−16] An exhaustive overview of different methods for estimation of octanol/water partition coefficient as well as of other physical properties was recently published by Katritzky et al.[17]

The fragment-based methods provide good results for a large number of compounds. However, difficulties can arise in decomposing some structures into appropriate fragments whose constants are available. The number of fragments in such methods can vary from several hundreds, in CLOGP[2] and KOWWIN[5] methods, up to several thousands in ACD/LogP[6] method. Such methods are considered to be the most precise in the field according to the comparison of 14 different methods performed by Mannhold and Dross.[18]

Quantum chemical parameters were used to predict only several hundreds of compounds, and it is not clear whether such approaches can be used as general estimation methods. In addition, the quantum chemical calculations are time-consuming, and this is actually a limiting factor for estimation of a large number of compounds.

The electrotopological state (E-state) indices were recently introduced by Hall and Kier[19,20] for the description of molecules. These indices combine together both electronic

* Corresponding author phone: ++41-21-692.5534; fax: ++41-21-692.5505; e-mail: itetko@eliot.unil.ch.
† Université de Lausanne.
‡ Institute of Bioorganic & Petroleum Chemistry.

10.1021/ci010368v CCC: $20.00
Published on Web 08/25/2001

and topological characteristics of the analyzed molecules. For each atom type in a molecule the E-state indices are summed and can be used in a group contribution manner. Such indices are known as atom-type E-state indices. In a similar way, the E-state indices can also be used to describe specific bonds between atoms, i.e., to describe two-atoms properties. Such indices are known as bond-type E-state indices. The calculation of the both types of E-state indices is very straightforward and simple. The number of successful application of these indices is rapidly growing, and more than 100 articles with application of these indices to predict physical properties and to correlate different kinds of biological activities were published.[21] Kier and Hall systematically documented the E-state indices in a new book.[22]

In our previous study with sets of 345 and 1754 organic compounds we found that the E-state indices can be successfully used to estimate the octanol/water partition coefficient.[23,24] The current study reports new results using PHYSPROP[25] that is the largest database of published log *P* coefficients.

DATA SET

The PHYSPROP database[25] used in our study included 13 360 compounds (May 2000) with experimental values for lipophilicity of diverse chemical compounds. Some of the chemicals, namely metal-containing compounds (190), all compounds that did not contain any carbons (11) and duplicates (251), such as L and D stereoisomers, were excluded from the analysis. Thus 12 908 molecules were used in the analysis.

Representation of Molecules. SMILES codes were used to represent molecular codes. All molecules were considered in their neutral form. The chlorides, bromides, and iodides of molecules were presented as one fragment with an atom of nitrogen with valence +5. For example, isoprenaline hydrochloride was represented not like a mixture of isoprenaline and hydrochloric acid, OC(C1=CC(O)=C(O)C=C1)CNC(C)C.[H]Cl or like ionized structure Oc1ccc(C(O)C-[NH2+]C(C)C)cc1O.[Cl-], but as one continuous fragment c1cc(O)c(O)cc1C(O)C[NH2](Cl)C(C)C, i.e., using the same representation that is adopted in KOWWIN.[5]

METHODS

Calculated Indices. Our previous analysis has indicated the importance of molecular weight for modeling of lipophilicity[23,24] and aqueous solubility[26] of chemical compounds. In this study, in addition to the molecular weight, two other simple parameters, namely the number of hydrogen and the number of non-hydrogen atoms, were also included into the analysis.

The input data for the program also included atom-type and bond-type E-state indices. Both sets of these indices were calculated using a program developed in-house (that was checked against the MolconnZ software[27]) with structure input for each analyzed compound using the SMILES line notation code.

Atom-Type E-State Indices. The basic set of indices consisted of electrotopological state indices proposed by Hall and Kier.[17,18] A set of extended indices for O and N atoms was also developed to take into account their functional groups and neighborhood.[24] The name of an extended index consisted of the name of original E-state index and name extension that depended on the atom neighborhood. All nitrogen atoms were divided into three groups. These groups included aliphatic (extension "(al)"), aromatic amines ("ar"), and other types of nitrogen ("oth"). In the same way oxygen atoms of OH groups (SsOH E-state index) were classified as atoms of alcohols ("alc"), phenols ("phen"), carboxylic acids ("acid"), and amino acids ("zwit"). Among the atoms of double-bonded oxygen (SdO E-state index) the atoms of ketones ("keto"), carboxylic acids ("acid"), esters ("ester"), amides ("amid"), nitro and nitroso groups ("nitro"), and sulfones and sulfoxides ("sulfo") were distinguished.

The scheme of calculation of atom-type E-state indices is shown in Table 1. A total of 72 atom-type E-state indices were calculated and were used as contributors in the regression and neural network models.

Bond-Type E-State Indices. The bond-type E-state indices were used to describe two atom estates. First, an intrinsic state value was assigned to each edge, and then the perturbation from each other edge was computed and added to the analyzed edge value (MolconnZ manual,[28] chapter 2). The bond E-state value was then computed as

$$I_{ij} = (I_i + I_j)/2 \tag{1}$$

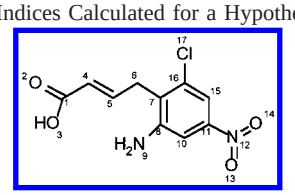
$$BES_{ij} = I_{ij} + \sum \Delta I_{ij}/(\bar{r}_{ij} + 1)^2 \tag{2}$$

where $\bar{r}_{ij}$ was computed as the average r_{ij} for the atoms in the two bonds. These values were computed for individual bonds and then were collected for each type of the bond in the molecule.[19]

Names of the bond type electrotopological state indices were different from those proposed by Hall and Kier,[18−20] although they were basically the same. The names also started with bond order indicator (e1, e2, e3, and ea for single, double, triple, and aromatic bonds, respectively) and were followed by two atom type names. Atom type names included atom name and the number of skeletal bonds for that atom. Finally an indication of the unused bonds[29] of each atom was given. The unused bonds were always mentioned in our implementation. This was different from the original version by Kier and Hall where an indication of unused bonds is given "if necessary".[28] For example, MolconnZ gives e1C2C3d for bond 3 in our example (see Table 1.4) while our program gives e1C2C3dd. A special type of nitrogen N== was introduced for the nitrogen with two double and one single bonds (ddsN type). This was done to avoid charged groups and to distinguish this type from the other nitrogen atoms bonded to three other skeletal atoms.

Efficient Partition Algorithm. The neural network ensemble (ANNE) of 50 networks was trained using efficient partition algorithm (EPA)[30−32] and the bias correction method.[33] All neural networks in the ensemble had the same architecture. The number of hidden neurons in one hidden layer was 10, and it was selected as described in the Results section. The EPA method was selected for neural network training, since the neural network trained with all 12 908 molecules fell in a local minimum. Thus no decrease of error below $RMSE = 1.5$ was achieved for the tested numbers of hidden neurons (5−15) and used training algorithm, SuperSAB,[34] for 25 000 iterations. On the contrary, no local minimum was observed if ANNs were trained using EPA

Table 1. E-State Indices Calculated for a Hypothetical Structure[a]



Part 1.1: E-State Indices for Skeletal Atoms

atom no.	index name	valence delta	electrotopological state S(i)	intrinsic state I(i)
1	dssC	4.00	−1.08	1.67
2	dO	6.00	10.26	7.00
3	sOH	5.00	8.41	6.00
4	dsCH	3.00	0.96	2.00
5	dsCH	3.00	1.37	2.00
6	ssCH2	2.00	0.20	1.50
7	aasC	4.00	0.47	1.67
8	aasC	4.00	0.17	1.67
9	sNH2	3.00	5.61	4.00
10	aaCH	3.00	1.19	2.00
11	aasC	4.00	−0.20	1.67
12	ddsN	5.00	−0.60	2.00
13	dO	6.00	10.52	7.00
14	dO	6.00	10.52	7.00
15	aaCH	3.00	1.18	2.00
16	aasC	4.00	0.15	1.67
17	sCl	0.78	5.83	4.11

Part 1.2: Atom-Type E-State Indices (Sums over All the Atoms of the Considered Type)

index no.	index name	value	index no.	index name	value
1	SaaCH	2.37	6	SdssC	−1.08
2	SaasC	0.58	7	SsCl	5.83
3	SdO	31.3	8	SsNH2	5.61
4	SddsN	−0.6	9	SsOH	8.41
5	SdsCH	2.32	10	SssCH2	0.20

Part 1.3: Extended Atom-Type E-State Indices (Sums over All Atoms of the Specified Type with Certain Chemical Functionality)

index no.	extended index name	value	index no.	extended index name	value
1	SdO(acid)	10.26	3	SsNH2(ar)	5.61
2	SdO(nitro)	21.03	4	SsOH(acid)	8.41

Part 1.4: Bond E-State Indices

bond no.	atom 1	atom 2	index name	intrinsic value, I	bond E-state value
1	1	2	e2C3O1s	4.33	6.03
2	1	3	e1C3O1d	3.83	4.95
3	4	1	e1C2C3dd	1.83	0.58
4	4	5	e2C2C2ss	2.00	1.61
5	5	6	e1C2C2ds	1.75	1.18
6	6	7	e1C2C3sa	1.58	0.70
7	7	8	eaC3C3aa	1.67	0.76
8	8	9	e1C3N1a	2.83	3.88
9	8	10	eaC2C3aa	1.83	1.18
10	10	11	eaC2C3aa	1.83	1.05
11	11	12	e1C3N==ad	1.83	0.35
12	12	13	e2NO2	4.50	6.67
13	13	14	e2NO2	4.50	6.67
14	11	15	eaC2C3aa	1.83	1.04
15	16	15	eaC2C3aa	1.83	1.17
16	16	17	e1C3Cl1a	2.89	4.02
17	16	7	eaC3C3aa	1.67	0.75

Part 1.5: Molecular Bond-Type E-State Indices (Sums of the Values for All the Bonds of the Considered Types)

index no.	name	value, sum of I values	index no.	name	value, sum of I values
1	Se1C2C2ds	1.18	7	Se1C3O1d	4.95
2	Se1C2C3dd	0.58	8	Se2C2C2ss	1.61
3	Se1C2C3sa	0.70	9	Se2C3O1s	6.03
4	Se1C3Cl1a	4.02	10	Se2NO2	13.3
5	Se1C3N1a	3.88	11	SeaC2C3aa	4.43
6	Se1C3N==ad	0.35	12	SeaC3C3aa	1.50

[a] SMILES code: C(=O)(O)C=CCc1C(N)Cc(N(=O)=O)cc1Cl; formula: $C_{10}H_9N_2O_4Cl$; molecular weight, MW = 256.65; number of nonhydrogen atoms, NA = 17; number of hydrogen atoms, NH = 9.

with only 2000−3000 iterations and the same training algorithm. The maximal number of iterations per one step of the EPA algorithm was fixed to be 15000, and it was used for all reported analyses.

In brief, for each ANN in the ensemble the first step of the computation by EPA began with two cases in the learning data set as well as two cases in the validated data set, selected by chance. At the end of this step, two additional cases for learning and two additional cases for validation were selected from the input data set as described in refs 30−32. Then, the training procedure of each ANN was based upon four cases in the learning data set and four cases in the validated data set at the second step. During each successive step, the number of cases in both data sets was doubled until it reached 50% from the initial training set for both learning and validation sets. The training of neural network was stopped, and neural network weights were saved when the minimum error for the validation set was calculated. The saved weights were loaded and again used for training at each consecutive step of EPA. This allowed EPA to avoid local minimum for training of a large data set. The selection of cases for a neural network was independent from all other networks. A more detailed description of EPA can be found in refs 30−32. The predictions of neural network ensemble were also corrected for systematic bias as described in ref 33.

The EPA algorithm used an automatic calculation of leave-one-out (LOO) results for the training set that was proposed in refs 35 and 36. In addition to the LOO results, a prediction ability of neural networks was estimated using several training/test sets protocol as described in the Results.

The prediction performance and comparison of methods analyzed in this study was performed using square of correlation coefficient, r^2, root mean squared error, *RMSE*, and the mean absolute error, *MAE*.

Analyzed Methods. A comparison of IA_LOGP, SciLogP Ultra, CLOGP, XLOGP, and KOWWIN programs was performed. The CLOGP program v. 4.0 was running under Mac OS operation system under evaluation license of BioByte Corporation.[37] The KOWWIN v. 1.66 was downloaded from the Internet,[38] and it was used under a user evaluation license from the Syracuse Inc. The XLOGP[66] program (v. 2.2) was freely distributed by the Institute of Physical Chemistry, Peking University, and it was downloaded from the ftp-server.[39] The XLOGP program used input data files in "mol2" Sybyl[40] format and relied on the internal types of atoms generated by Sybyl. We have programmed a conversion program that converted SMILES to "mol2".[41] A comparison of the developed algorithm and the IA_LOGP program was based on the published results of the IA_LOGP method[42] as well as using the online demo version of this program available at http://www.logp.com. The published results of SciLogP Ultra[43] were also used.

RESULTS

Several different calculations were performed. The main analysis included all 12 908 molecules, and it was used to build a powerful log *P* program for estimation of the *n*-octanol/water partition coefficient. The other analyses were performed to estimate prediction ability of the developed method and to compare its performance with other logP prediction programs.

Analysis of the Total Set of 12 908 Molecules. ANNs could provide accurate predictions, but this method was rather slow to select the pertinent indices to be used for the log *P* prediction. Thus, a preliminary analysis was performed by multiple linear regression analysis (MLRA).

As it was already mentioned in the Introduction, our previous studies[23,24] indicated that atom-type E-state indices generated significant models for prediction of lipophilicity and aqueous solubility of chemical compounds. A first study was performed to further evaluate performance of these indices for prediction of lipophilicity of chemical compounds from the PHYSPROP database.

Multiple Linear Regression Analysis. This section describes selection of parameters for the final MLRA model (MLR5), and results were calculated using several intermediate models (MLR1-MLR4).

MLR1. The total number of extended atom-type E-state indices generated for the analyzed data set included 72 indices (Table 2). Only indices that had frequency of at least 20 times were used at the first step of the analysis. In addition to the atom-type E-state indices, molecular weight and total numbers of hydrogen and non-hydrogen atoms were also analyzed. The analysis of the predicted results revealed a group of 292 molecules that mainly contributed to the large prediction errors by MLRA (*RMSE* = 0.95, indicated as MLR1 in Table 3). All molecules in this group contained a bond between quaternary ammonium and Cl, Br, or I.

MLR2. The detected group of compounds with a bond between a quaternary ammonium and halogens was characterized by specific index SssssN. However, an influence of N−Cl, N−Br, and N−I bond was rather different on the lipophilicity of these compounds and thus required specific explicit information about which type of interaction is present in the molecule. Indeed, including the bond-type indices, Se1N4Cl1a, Se1N5Br1s, Se1N4I1a, and Se1N5I1s, that had frequencies more than 20 decreased the MLRA error to *RMSE* = 0.76.

MLR3. In addition to these four indices there were other indices with a bond between quaternary ammonium and Cl, Br, or I, e.g. Se1N5Cl1s, Se1N4Cl1s, Se1N4Br1a, etc. The values of these rare indices were added ("aliased") to the values of the three largest indices related to them in properties, i.e., the values of Se1N5Cl1s were added to Se1N4Cl1a and the values of Se1N4Br1a were added to Se1N5Br1s. This procedure improved MLRA results, and *RMSE* = 0.72 was obtained.

MLR4. This result encouraged us to find aliases for atom-type E-state indices that had frequencies less than 20 and were excluded from the preliminary analysis. In addition, all pairs of indices that had large differences in frequency in data set but had similar regression coefficients (i.e., SdsN and SdNH, SsSH and SssS, etc.) were merged together, by adding an alias to the index with a smaller frequency. This procedure decreased the total number of used indices from 64 to 58 and did not change the statistical parameters of the model.

MLR5. Further analysis included careful examination of atom-type bond indices. All these indices were initially added one by one to the set of the selected indices. The index that provided the best improvement of the MLRA results was included in the set of selected indices. Such addition of indices was similar to the ascending procedure of regression analysis. However, after the addition of each new index, an attempt to "generalize" the detected index was performed. To do this, all indices with properties that were similar to the detected index were analyzed. For example, the index Se1C2O2ss was initially detected as significant. This index described a single bond, sCH_2-Os, between carbon and oxygen atoms. An analysis made possible to extend this index by indices Se1C1O2s (CH3−Os), Se1C3O2ss (ssCH−Os), and Se1C4O2ss (sssCH−Os). All these three indices were used as aliases of the index Se1C2O2ss. Thus, the "generalized" index described a single bond between the twice substituted atom of oxygen that has only single bonds to other atoms and the atom of carbon that also has only single bonds to other atoms. An attempt to extend this index by including carbon with aromatic or higher order bonds to other atoms increased *RMSE*, and thus was not continued. In general, whenever the extension of an index improved or did not change MLRA results it was considered as successful. Otherwise, an attempt was done to consider a different extension that was logical from a chemical point of view (i.e., in case of Se1C2O2ss it was possible, for example, to consider an alias from Se2C2O2s). The index SeaC2N4aa was another interesting example. This index was extended to include all single and aromatic bonds between nitrogen atom of tertiary and quaternary amino salts and a carbon atom. The analysis of bond-indices was terminated when the remaining indices did not improve MLRA results for more than 0.005 units ending up with 76 indices and *RMSE* = 0.63. The set of input parameters contained molecular weight, total number of carbons, and all atoms. It was also found that MLRA results were improved if the molecular weight was excluded from the set of indices, while the E-state indices were normalized according to its root of 1/3 degree. The normalization on other roots did not improve the results. The final MLRA model calculated *RMSE* = 0.61, and the regression coefficients are shown in Table 2.

Artificial Neural Network Analysis. The number of hidden neurons was determined using several EPA runs with three different numbers, 5, 10, and 15, of hidden neurons and all parameters for the whole training set of 12 908 molecules. It was found that the LOO error calculated for the training set decreased from $RMSE_{LOO} = 0.46$ to $RMSE_{LOO} = 0.45$ when the number of neurons in the hidden layer was changed from 5 to 10. $RMSE_{LOO}$ remained the same for 15 neurons in the networks, but training of such networks required 10 instead of 6 days using Celeron-500 MHz computer. Thus, the neural networks with 10 hidden neurons were selected for all studies.

The application of ANNs significantly increased the prediction ability of the MLRA, and $r^2 = 0.94$, $RMSE_{LOO} = 0.45$, and $MAE_{LOO} = 0.31$ were calculated for all analyzed molecules. The consensus results of the program were $r^2 = 0.95$, *RMSE* = 0.42, and *MAE* = 0.29. These results were calculated when the trained ANNE was used to predict all 12 908 molecules. However, as it is shown later in the article, the LOO and not the consensus results provided a correct estimation of the prediction performance of the method. The ANNs weights were saved and were included into ALOGPS (artificial neural network program for calculation of log *P* and log *S*) v. 2.0 program, that is freely available for both single and batch modes (about 1000 compounds/minute using Celeron 500 MHz computer) at

Table 2. E-State Indices Used in the Analysis

no.	index	frequency of the index in the whole set	frequency of the index in the star set	frequency of the index in the nova set	used in MLRA model	coeff in MLR5	no.	index	frequency of the index in the whole set	frequency of the index in the star set	frequency of the index in the nova set	used in MLRA model	coeff in MLR5
	constant term					−1.42							
1	no. of non-H atoms, NA					0.052							
2	no. of H atoms, NH					−0.043							
						Atom-Type E-State Indices							
3	SsNH2(al)	292	218	74	1−5	−0.07	29	SsssCH	3753	2512	1241	1−5	0.082
4	SsNH2(ar)	818	614	204	1−5	−0.083	30	SssssC	1980	1406	574	1−5	0.091
5	SsNH2(oth)	778	638	140	1−5	−0.032	31	StsC	619	437	182	1−5	0.030
6	SssNH(al)	461	282	179	1−5	−0.172	32	SaaN	2851	2096	755	1−5	
7	SssNH(ar)	522	358	164	1−5	−0.048	33	SaaNH	452	324	128	1−5	
8	SssNH(oth)	3419	2593	826	1−5	−0.175	34	SaasN	1044	705	339	1−5	−0.113
9	SsssN(al)	1121	730	391	1−5	−0.225	35	SaassN	170	0	170	1−5	−0.678
10	SsssN(ar)	347	235	112	1−5	−0.079	36	SaadN	118	89	29	1−5	−0.915
11	SsssN(oth)	1942	1400	542	1−5	−0.294	37	SssssN	104	0	104	1−5	−0.312
12	SsOH(alc)	1616	1138	478	1−5	−0.011	38	SdsN	1644	1073	571	1−5	−0.017
13	SsOH(phen)	804	528	276	1−5	0.029	39	SdssN	25	8	17	1−5	−1.89
14	SsOH(acid)	922	582	340	1−5		40	SddsN	1208	953	255	1−5	0.214
15	SsOH(zwit)	86	67	19	1−5	−0.117	41	StN	510	360	150	1−5	0.012
16	SdO(keto)	1251	918	333	1−5	0.0094	42	StdN	34	10	24	1−5	0.098
17	SdO(acid)	1005	649	356	1−5	0.018	43	SaaO	406	332	74	1−5	0.035
18	SdO(ester)	1992	1485	507	1−5	0.011	44	SssO	5081	3675	1406	1−5	0.035
19	SdO(amid)	4049	3097	952	1−5	−0.005	45	SaaS	385	212	173	1−5	0.271
20	SdO(nitro)	1371	1078	293	1−5	0.028	46	SdS	308	253	55	1−5	0.156
21	SdO(sulfo)	969	674	295	1−5	−0.032	47	SdssS	173	57	116	1−5	−0.142
22	SaaCH	10049	7276	2773	1−5	0.119	48	SddssS	856	617	239	1−5	−0.183
23	SaasC	10102	7330	2772	1−5	0.082	49	SssS	854	657	197	1−5	0.216
24	SaaaC	1216	866	350	1−5	0.122	50	SsBr	543	346	197	1−5	0.233
25	SdsCH	2111	1456	655	1−5	0.093	51	SsCl	2331	1552	779	1−5	0.103
26	SdssC	7669	5686	1983	1−5		52	SsF	1067	821	246	1−5	0.046
27	SsCH3	8538	6180	2358	1−5	0.170	53	SdsssP	304	199	105	1−5	0.083
28	SssCH2	7830	5418	2412	1−5	0.141	54	SsI	294	134	160	1−5	0.357
						Bond-Type E-State Indices							
55	Se1N4Cl1a	117	0	117	3−5	−0.49	66	Se1C3N2as	1713	1284	429	5	0.108
56	Se1N5Br1s	51	0	51	3−5	−1.00	67	Se1C3N2ds	3345	2529	816	5	0.136
57	Se1N5I1s	46	0	46	3−5	−1.31	68	Se1C2O2ss	2533	1789	744	5	−0.042
58	Se1C1C2s	2779	1949	830	5	0.145	69	Se1C2S2ss	385	268	117	5	−0.106
59	Se1C1C3s	1479	1017	462	5	0.077	70	Se1C3Cl1a	1776	1257	519	5	0.031
60	Se1C2C2ss	4178	2777	1401	5	0.119	71	SeaN2N2aa	262	220	42	5	−0.052
61	Se1C2C3ss	2777	1858	919	5	0.098	72	Se2N2N2ss	135	91	44	5	0.159
62	Se1C3C3ss	1214	835	379	5	−0.123	73	Se1N2N2ds	177	81	96	5	0.105
63	SeaC2N2aa	1885	1376	509	5	−0.048	74	Se1O1S4d	12	6	6	5	−0.178
64	Se1C2N==dd	41	34	7	5	−0.529	75	Se1S3I1d	55	0	55	5	−1.725
65	SeaC2N4aa	169	0	169	5	−0.312							
						Atom Type E-State Indices Used as Aliases							
76[b]	StCH	116	83	33	1−5	StsC	86	SaasNH	2	0	2	4−5	SaassN
77[b]	SdCH2	264	156	108	1−5	SdsCH	87[b]	SaaaN	21	18	3	1−5	SaasN
78[b]	SddC	47	35	12	1−5	SdsCH	88[b]	SsSH	27	18	9	1−5	SssS
79	SdsssN	8	3	5	4−5	SsssssN	89	SssssssS	2	2	0	4−5	SssS
80	SsssNH2	8	0	8	4−5	SsssssN	90	SaaaS	6	6	0	4−5	SaaS
81	SdsNH2	1	0	1	4−5	SsssssN	91	SddsI	1	1	0	4−5	SsI
82	SssssNH	5	0	5	4−5	SsssssN	92	SdsI	1	0	1	4−5	SsI
83	SssNH3	12	0	12	4−5	SsssssN	93	SsssP	4	4	0	4−5	SdsssP
84[b]	SdNH	79	40	39	1−5	SdsN	94	SdssPH	2	1	1	4−5	SdsssP
85	SaasNH2	1	0	1	4−5	SaassN	95	SssssP	1	0	1	4−5	SdsssP
						Bond Type E-State Indices Used as Aliases							
96	Se1N4I1a	35	0	35	2−5	Se1N5I1s	110	Se1C1C4s	1013	747	266	5	Se1C1C3s
97	Se1N2Cl1s	11	0	11	3−5	Se1N4Cl1a	111	Se1C1C1	1	1	0	5	Se1C1C2s
98	Se1N3Cl1a	2	0	2	3−5	Se1N4Cl1a	112	Se1C1C2d	35	13	22	5	Se1C1C2s
99	Se1N3Cl1s	7	0	7	3−5	Se1N4Cl1a	113	Se1C1C2t	10	6	4	5	Se1C1C2s
100	Se1N4Cl1d	3	0	3	3−5	Se1N4Cl1a	114	Se1C1S2s	194	165	29	5	Se1C2S2ss
101	Se1N4Cl1s	4	0	4	3−5	Se1N4Cl1a	115	Se1C2S1s	10	9	1	5	Se1C2S2ss
102	Se1N5Cl1s	5	0	5	3−5	Se1N4Cl1a	116	Se1C3S1s	3	3	0	5	Se1C2S2ss
103	Se1N3Br1s	1	0	1	3−5	Se1N5Br1s	117	Se1C3S2ss	158	115	43	5	Se1C2S2ss
104	Se1N4Br1a	15	0	15	3−5	Se1N5Br1s	118	Se1C4S1s	1	0	1	5	Se1C2S2ss
105	Se1C2C2dd	88	71	17	5	Se1C2C2ss	119	Se1C4S2ss	51	43	8	5	Se1C2S2ss
106	Se1C2C2ds	312	210	102	5	Se1C2C2ss	120	Se2C3S1s	175	146	29	5	Se1C2S2ss
107	Se1C2C2dt	14	5	9	5	Se1C2C2ss	121	Se1C3C4ss	682	466	216	5	Se1C3C3ss
108	Se1C2Cst	169	130	39	5	Se1C2C2ss	122	Se1C1N4a	89	0	89	5	SeaC2N4aa
109	Se1C2C2tt	11	2	9	5	Se1C2C2ss	123	Se1C1N4d	2	0	2	5	SeaC2N4aa

Table 2 (Continued)

no.	index	frequency of the index in the whole set	star set	nova set	used in MLRA model	coeff in MLR5	no.	index	frequency of the index in the whole set	star set	nova set	used in MLRA model	coeff in MLR5
						Bond Type E-State Indices Used as Aliases							
124	Se1C1N4s	3	0	3	5	SeaC2N4aa	143	Se1C4N==sd	17	12	5	5	Se1C2N==dd
125	Se1C1N5s	99	0	99	5	SeaC2N4aa	144	Se2C2N==sd	11	10	1	5	Se1C2N==dd
126	Se1C2N4da	4	0	4	5	SeaC2N4aa	145	Se2C3N==sd	4	1	3	5	Se1C2N==dd
127	Se1C2N4sa	76	0	76	5	SeaC2N4aa	146	Se1C3N1a	818	614	204	5	Se1C3N2as
128	Se1C2N4sd	6	3	3	5	SeaC2N4aa	147	Se1C3N3da	82	70	12	5	Se1C3N2ds
129	Se1C2N4ss	5	0	5	5	SeaC2N4aa	148	Se1C3N3ds	1891	1369	522	5	Se1C3N2ds
130	Se1C2N5ss	95	0	95	5	SeaC2N4aa	149	Se1C1O1	1	1	0	5	Se1C2O2ss
131	Se1C3N4ad	1	0	1	5	SeaC2N4aa	150	Se1C1O2s	1748	1320	428	5	Se1C2O2ss
132	Se1C3N4sa	1	0	1	5	SeaC2N4aa	151	Se1C3O2ss	1109	718	391	5	Se1C2O2ss
133	Se1C3N4ss	4	0	4	5	SeaC2N4aa	152	Se1C4O2ss	296	212	84	5	Se1C2O2ss
134	Se1C3N5as	8	0	8	5	SeaC2N4aa	153	Se1O2S3sd	3	2	1	5	Se1O1S4d
135	Se1C3N5ss	1	0	1	5	SeaC2N4aa	154	Se1O2S4sd	27	18	9	5	Se1O1S4d
136	Se1C4N5ss	1	0	1	5	SeaC2N4aa	155	Se1N2N2dd	39	23	16	5	Se1N2N2ds
137	Se2C3N4ss	3	0	3	5	SeaC2N4aa	156	Se1N2N==dd	5	3	2	5	Se2N2N2ss
138	SeaC3N4aa	131	0	131	5	SeaC2N4aa	157	Se1N2N==sd	2	0	2	5	Se2N2N2ss
139	Se1C1N==d	2	1	1	5	Se1C2N==dd	158	Se1N3N==sd	1	0	1	5	Se2N2N2ss
140	Se1C2N==sd	15	11	4	5	Se1C2N==dd	159	Se2N2Nst	18	0	18	5	Se2N2N2ss
141	Se1C3N==dd	11	10	1	5	Se1C2N==dd	160	Se2N2N==sd	32	32	0	5	Se2N2N2ss
142	Se1C3N==sd	9	6	3	5	Se1C2N==dd							

[a] See description of MLR1-MLR5 models in the text. If the index was used as an alias, the last column contains the target index for the alias. The alias indices were not used in the calculations, but their values were added to the value of the target indices. [b] Used as an index in the MLR1-MLR3 and as an alias in the MLR4-MLR5 models.

Table 3. Statistical Parameters of MLRA Models[a]

model	description of used params	number of params initial	number of params significant	r^2	RMSE
MLR1	58 atom type E-state indices with frequencies = 20 and MW, NA, NH	61	55	0.74	0.95
MLR2	+ 4 nitrogen−halogen bond indices	65	60	0.84	0.80
MLR3	+ aliases to all bond indices describing bond between nitrogen and halogens	64	56	0.85	0.72
MLR4	+ aliases to missed and similar atom-type indices	58	51	0.85	0.72
MLR5	+ aliases to bond-type E-state indices + indices were normalized on $0.1*\sqrt[3]{MW}$	75	71	0.89	0.61

[a] MW: molecular weight; NA: number of non-hydrogen atoms; NH: number of hydrogen atoms in a molecule.

http://www.lnh.unil.ch/~itetko/logp. A description of the WWW interface of the program was recently provided[41] for ALOGPS v. 1.0, and it applies to the current version too.

The calculated results contained a number of outliers, i.e., molecules with large prediction errors. The limit of ±1.5 log units prediction error was used to identify $n = 131$ outliers. This number approximately corresponded to the expected number by chance for normal distribution with $\sigma = 0.45$. The calculated results without outliers were improved thus giving $RMSE_{LOO} = 0.39$ and $MAE_{LOO} = 0.29$. About 20−50% of outliers detected in our model were also outliers using the same criterion in the IA_LOGP, XLOGP, CLOGP, and KOWWIN programs. Thus, other methods also had difficulty in predicting the lipophilicity of these molecules. The relative statistic for outliers is provided in Table 4, and their complete list including values calculated for these molecules with different programs are available at http://www.lnh.unil.ch/~itetko/logp/outliers.html.

It was very clear that the outliers, even if their number was small (about 1%), biased the prediction ability of the method. The outliers seriously influence the least mean square analysis,[44] and such molecules were usually excluded.[45] Thus, for further analysis and especially for comparison of performance of different methods the results calculated without outliers were used, and the number of outliers, identified as molecules with more than ±1.5 log unit prediction errors, were provided.

Table 4. Prediction Performances of Different Methods for ALOGPS Outliers

method	outliers[a]	r^2	RMSE	MAE
ALOGPS − consensus	86	0.59	1.88	1.73
IA_LOGP − consensus	51	0.64	1.82	1.39
CLOGP	45	0.58	2.11	1.45
KOWWIN	18	0.85	1.16	0.78
XLOGP	70	0.53	2.15	1.79

[a] Number of molecules (out of 131 ALOGPS outliers for LOO procedure) that were also outliers for the considered method. Notice, that only 86 molecules were outliers for ALOGPS consensus model.

In the following sections we provide a comparison of our new algorithm with several popular lipophilicity calculation programs. Special care is taken in order to account for different training sets used to develop these methods.

ALOGPS vs IA_LOGP. The first comparison of results was done with the IA_LOGP program developed by M. Parham et al.[42] The authors used an ensemble of 10 neural networks that were applied to the initial training set of 12 942 organic compounds. The compounds used in their study were taken from the PHYSPROP database as well as from some

Table 5. Comparison of IA_LOGP and ALOGPS[c]

	molecules	outliers[a]	r^2	RMSE	MAE
IA_LOGP consensus	12942	0	0.96	n/a	0.29
IA_LOGP validation	10353	0	0.94	n/a	0.36
IA_LOGP test set	1258	0	0.96	n/a	0.31
ALOGPS consensus	12908	86	0.96 (0.95)[a]	0.38 (0.42)	0.28 (0.29)
ALOGPS validation	12908	130	0.95 (0.94)	0.39 (0.45)	0.28 (0.31)
test set[b]	1174[b]	1	0.96 (0.96)	0.36 (0.36)	0.27 (0.27)
ALOGPS LOO	12908	131	0.95 (0.94)	0.39 (0.45)	0.29 (0.31)
ANN1 LOO	6439	80	0.95 (0.93)	0.42 (0.49)	0.32 (0.34)
ANN1 test	6469	90	0.94 (0.92)	0.42 (0.49)	0.32 (0.34)
ANN2 LOO	6469	101	0.95 (0.93)	0.43 (0.51)	0.34 (0.36)
ANN2 test	6439	98	0.95 (0.93)	0.43 (0.51)	0.33 (0.35)
ANN1 + ANN2 LOO	12908	181	0.95 (0.93)	0.42 (0.49)	0.32 (0.34)
ANN1 + ANN2 test	12908	188	0.95 (0.93)	0.43 (0.50)	0.33 (0.35)

[a] The molecules with prediction error above ±1.5 log units were considered as outliers. Results calculated with outliers are indicated in parentheses. No outliers were reported for IA_LOGP program, but Table 4 suggestes that such molecules have been presumably excluded before the final analysis. [b] Only 1174 molecules from the test set of IA_LOGP method were from the PHYSPROP database and were available for our analysis. Neural networks were trained using 11 734 molecules. The predicted results for the test set of 1174 molecules are shown. [c] ALOGPS: neural networks were trained using 12908 molecules. ANN1: neural networks were trained using 6439 molecules and were used to predict 6469 molecules from the test set. ANN2: training and test sets were interchanged in comparison to ANN1. LOO: leave-one-out results calculated for the molecules from the training set. Test: predicted results for the test set. ANN1+ANN2: combined results for all analyzed molecules.

other commercial databases.[46] The authors excluded compounds that

1. could not be read by MolconnZ or ChemOffice;

2. organometallics that could not be used with MolconnZ (Fe, Hg, Se, etc.), except for Na and K;

3. compounds that belonged to such small classes that they could not be modeled, e.g. boron containing compounds, etc.

4. duplicates (e.g. D and L structures).

The authors did not report how many compounds in their set were selected in the PHYSPROP database.[47] The input parameters for artificial neural networks included 107 E-state indices that were selected from the original set of 224 MolconnZ indices using Interactive Analysis software. Thus, both ALOGPS and IA_LOGP basically used the same database, the same method, and the same indices. However, even for such similar programs a comparison of calculated results should be done with a great care.

To better understand the comparison between methods, let us consider the training procedure of the IA_LOGP program. Each neural network in the IA_LOGP program had its own learning and validation sets.[48] The learning sets included 80% of the molecules (i.e., 9319 to 9895 molecules), while the validation sets included 10% of the molecules (i.e., 1789 to 2365) from the initial training set. The remaining 10% of the molecules (1258 molecules) were used to test performance of the ensemble of neural networks after termination of training. The authors indicated that all predicted values by their method were within 1.5 log units. Since 51 of the outliers in our model were also outliers in IA_LOGP (Table 4), it is likely that molecules with absolute errors more than 1.5 log units were excluded by the authors. Thus, a proper comparison of both methods should be performed without the outliers.

The authors reported consensus results, results calculated for the validation, and the test set. The consensus results were calculated by the ANNE applied to the whole set of 12 942 molecules. The validation results were calculated for 90% of the molecules that were in the 10 validation sets. And, finally, the prediction ability of ANNE for the test set of 1258 was also reported.

The consensus results, however, are influenced by the number of molecules in the training and in the validated sets, training algorithm, and neural network architecture. In fact, the larger number of molecules is in the learning set compared to the validation set (and in the limit zero validation set), and the more similar consensus results to the fitted and not to the predicted results can be obtained. The learning set of the IA_LOGP neural network included 80% of the molecules from their initial training set compared to 50% of the molecules used in the EPA algorithm. Thus, in general, the percentage of molecules in the learning set of IA_LOGP was higher than that of ALOGPS, i.e., the IA_LOGP calculated consensus results were more similar to the fitted results than the ALOGPS.

The number of hidden neurons could also influence the consensus results. For example, $RMSE = 0.43$, 0.42, and 0.41 were calculated by ALOGPS for neural networks with 5, 10, and 15 hidden neurons. The error continued to decrease with an increase in the number of hidden neurons, and the use of larger neural networks could, probably, provide even better consensus results. Thus, even if consensus results of ALOGPS were better in comparison to the IA_LOGP results, this comparison did not provide any idea about the predictive performance of these both methods.

The results calculated for the validation sets, i.e., the sets that were used to terminate neural network training, were in favor of our model (Table 2). The validation results were less influenced by ANNs architecture, i.e., $RMSE = 0.46$, 0.45, and 0.45, were calculated for neural networks with 5, 10, and 15 hidden neurons. These results were similar to the LOO estimation.

The predicted results calculated for a test set of 1258 molecules were reported by IA_LOGP program. Only 1174 of these molecules were from the PHYSPROP dataset, while the other molecules were from commercial databases and were not available for our analysis.[47] The EPA algorithm was trained using the remaining set of 11734 molecules, and, after termination of training, the molecules from the test set were predicted. The calculated results for the test set (Table 5) were better than the results calculated by the IA_LOGP program. These results were also better than the LOO results

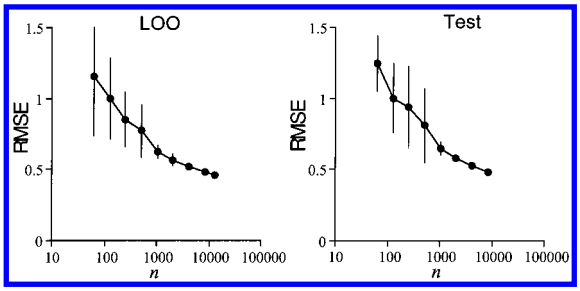


Figure 1. Leave-one-out (LOO) and predicted error (test) calculated using training sets selected by chance from the initial set of 12 908 molecules. The LOO error was estimated using *n* molecules from the training set. The predicted error (test) was calculated for the remaining molecules that were left apart from the neural network learning. The error bars indicate confidence intervals at $p = 0.95$ estimated using 10 different runs of algorithm with training sets selected by chance for each run. The results are reported without excluding the outliers.

calculated by the ALOGPS model using the whole training set. Thus, the use of this test set overestimated generalization ability of the ALOGPS method compared to the LOO results. Overall the ALOGPS method provided better statistical parameters compared to the IA_LOGP for the analyzed protocols.

Validation of ALOGPS Method Using Different Training/Test Set Procedures. An additional study using the test/training set protocol was performed in order to better estimate how the LOO and results calculated for the test set might correspond to one another. All data corresponding to the initial training set were divided in two sets at random. Both sets contained approximately 50% of the data points, namely 6439 and 6469 molecules. An ensemble of 50 ANNs was trained by EPA algorithm using either one of the two sets. After the termination of training the molecules from the remaining test set were predicted. The prediction ability of ANNs estimated by LOO parameters coincided with the actual predictions for the test sets (Table 5). Thus, the LOO results provided reliable estimations of the prediction accuracy of EPA method.

It is important to notice that the prediction ability estimated by LOO results for this analysis was lower compared to the LOO results calculated using all molecules in the training set (Table 5). It is easy to anticipate that a large dataset should provide better generalization ability of the model, but it also indicated that the prediction ability of ALOGPS could be further increased by using a larger number of molecules in the training set.

To evaluate the extent of the gain in accuracy as a function of the sample size of the training set we used sets of 64, 128, ..., 4096, molecules that were selected by chance from the initial training set. The prediction ability was measured using LOO results and by the prediction of the remaining molecules that did not participate in the neural network training.

The calculated results indicated that the *RMSE* decreased approximately linearly with the logarithm $log_{10}(n)$ of the sample size *n* of the training set (Figure 1). However, there was an interesting behavior and dramatic change in the slope of decrease of the error for about 1000 molecules in the training set. It looked like 1000 was a critical "mass" of molecules required for a neural network to learn the diversity of the investigated data set. Both LOO and prediction errors of neural networks had a very large variation before this critical number of samples in the training set. However, for sample sizes larger than 1000 the confidence of neural network predictions increased dramatically, and the slope of the *RMSE* error changed from −0.4 to −0.2. Another important result was that both estimated LOO error and actual ANNs prediction for the test set were within the confidence limits estimated using 10 runs of algorithm with different training sets selected by chance.

This analysis raised the following question: which generalization ability of the developed method was actually tested. In fact, the performance of a method inside the homologous series of compounds, i.e., the interseries prediction, was measured. Indeed, the molecules for the training and for the test sets were selected randomly, and the representatives of all series of compounds (in general their number was not known to us) were used for both training and for testing.

The prediction within the same series of compounds could be of practical interest, for example, when developing new drugs with a common basic structure. However, there is also a big demand to develop a method that could reliably predict lipophilicity for molecules from the other series of compounds that were not belonging to the training set, i.e., to perform cross-series prediction, as described in the next section.

A Comparison of ALOGPS, CLOGP, and XLOGP for Nonhomologous Sets of Compounds. The PHYSPROP database contained many compounds that were included in BioByte Starlist.[37] A run of the CLOGP program over all 12 908 compounds from the PHYSPROP provided experimental values for 9429 compounds. This and the remaining set of 3479 compounds were used to compare performances of the different methods. For the sake of simplicity we referred to the set of 9429 compounds as a "star" set (i.e., since these data were coming from BioByte Starlist of log *P* values) and to the remaining set of compounds as the "nova" set (i.e., the name of a new star in astrophysics).

The frequencies of E-state indices calculated for both sets were rather different, including some indices that were completely absent in one of the sets. For example, there were no indices corresponding to a bond between quaternary ammonium and Cl, Br, and by consequence the indices SaassN and SssssN were also absent. Indeed, the CLOGP program does not predict this kind of compounds. Another index, Se1S3I1d, that was absent in the star set corresponded to the series of thiouronium compounds as well as there were some indices describing nitrogen bonds. In total, there were 384 compounds with the E-type indices that were found to be significant for the whole initial training set of 12 908 molecules but were presented only in the nova set.

The star set compounds were "more easy compounds" for training and prediction by all investigated methods. Indeed, all methods calculated better statistical coefficients and less number of outliers for this set compared to the nova set. The results of CLOGP method were slightly improved (*RMSE* = 0.35, *MAE* = 0.24 and 56 outliers) if it was applied using the experimental values provided by the CLOGP program. The ALOGPS calculated the same statistical coefficients using values provided by the CLOGP program, but the number of outliers decreased from 68 to 64.

Table 6. Comparison of CLOGP, KOWWIN, XLOGP, and ALOGPS Methods[c]

	star set, 9429 molecules				nova set, 3479 molecules			
method	r^2	RMSE	MAE	outliers	r^2	RMSE	MAE	outliers[a]
CLOGP	0.96	0.36	0.25	74	0.92	0.62	0.50	558 (356)
KOWWIN	0.95	0.40	0.29	56	0.96	0.46	0.35	55 (10)
XLOGP	0.89 {0.95}[b]	0.56 {0.33}	0.44 {0.25}	448 {4}	0.91	0.64	0.52	690 (366)
ALOGPS LOO	0.95	0.37	0.27	68	0.96	0.44	0.33	63 (15)
ANN1 + ANN2 LOO	0.94	0.41	0.31	82	0.96	0.48	0.37	99 (18)
ANN1 + ANN2 test	0.94	0.41	0.31	81	0.96	0.48	0.37	107 (23)
ANN3	0.94	0.41	0.31	76	0.92	0.57	0.45	480 (357)
ANN4	0.89 {0.95}[b]	0.55 {0.33}	0.43 {0.25}	358 {14}	0.91	0.61	0.49	576 (357)
ANN5	0.90	0.52	0.41	205	0.96	0.47	0.36	81 (15)

[a] The number in parentheses indicates the number of outliers for a subset of 384 molecules containing significant E-state indices that were absent for all molecules in the star set. Notice, that for cross-series predictions (XLOGP, CLOGP, ANN3, and ANN4) 90% of these molecules were outliers for all methods. [b] Results for 1853 molecules from the XLOGP training set are shown in brackets. [c] ALOGPS, ANN1+ANN2: see Table 5 for description of the used training/test protocols. ANN3: neural networks were trained using 9429 molecules from the star set. The LOO results are reported for the star set and the predicted results for the nova set. ANN4: neural networks were trained using XLOGP training set of 1853 molecules. The LOO results for the molecules from the XLOGP training sets (in brackets) and actual prediction for the remaining molecules in the star and nova sets were reported. ANN5: neural networks were trained using the nova set. LOO results were reported for the nova set and predicted results for the star set. All reported results were obtained after excluding the outliers.

Out of 1853 molecules from the XLOGP training set, 1564 were in star and 111 were in the nova sets. These molecules were removed from both these sets, and the prediction results of the XLOGP method for the remaining molecules in both sets were calculated (Table 6). The results for the star and nova sets using XLOGP were much worse than with the other reported methods.

There was also a dramatic difference in the results calculated by CLOGP compared to KOWWIN and ALOGPS program for the nova set. While KOWWIN and ALOGPS programs provided only a slight decrease in the prediction performance for this set, the CLOGP performed almost twice as worse, and, in addition, 558 outliers were sorted out. Almost 65% of these outliers (356 molecules) were, in fact, molecules with the significant E-state indices that were presented in the nova and were missed in the star set.

This finding raised the question why KOWWIN and ALOGPS programs were so good for the prediction of both sets and why the other two programs failed to do it. Is the difference in the prediction power of all methods due to some specific design of these methods or to some other factors, e.g., due to the various diversities of molecules in their training sets? To investigate this question, the ALOGPS program was trained using the star set and the XLOGP training sets, and its prediction ability was tested on the remaining compounds.

ANNs trained only with the star set (ANN3) provided a performance similar to CLOGP for the nova set. Notice, that the actual predictions of networks, ANN1 and ANN2, trained with 50% molecules from the PHYSPROP database were the same as the LOO results calculated for the star set. However, the results of these networks for the nova set were considerably better. Another interesting fact was that the LOO results calculated for the molecules of the star set were better for ANNs trained using all 12 908 molecules compared to the LOO results calculated with ANNs trained using only the star set. Thus, the presence of molecules from the nova set allowed the ANNs to improve their prediction ability for the star set molecules.

ANNs trained using the XLOGP training set (ANN4) provided similar results for the prediction of the remaining compounds from the star and from the nova sets. The only exception was a large number of outliers calculated for nova in comparison to the star set. The statistical coefficients of ANNs were rather similar to those of XLOGP. It was also interesting that the LOO results of neural networks trained using the XLOGP training set were biased and did not correspond to the actual predictions. For example, $RMSE = 0.39$ (0.33 without 14 outliers) was calculated by ANNs for the molecules from the XLOGP set. However, the actual prediction for the star set, $RMSE = 0.71$ (0.54 without 358 outliers), and for the nova set, $RMSE = 1.20$ (0.61 without 934 outliers), were approximately twice as worse. Notice that for the training sets of the same size ($n = 1853$) but selected at random from the whole set of 12 908 molecules, both LOO $RMSE_{LOO} = 0.59 \pm 0.03$ and the actual prediction of the remaining 11 055 molecules $RMSE = 0.60 \pm 0.03$ (0.49 ± 0.02 without 250 ± 20 outliers) were significantly better. Thus, a use of sets that were selected at random provided a considerable improvement of the performance of ANNs over the remaining molecules compared to the use of the XLOGP set. The XLOGP set was not diverse enough to represent all molecules of the star set. The LOO results, $RMSE_{LOO} = 0.50 \pm 0.02$, were calculated for the training sets of 1853 molecules that were selected at random from the star set. These results corresponded to the actual prediction of the remaining molecules $RMSE = 0.53 \pm 0.02$ (0.46 ± 0.02 without 140 ± 20 outliers).

The last analysis (ANN5) was performed when ANNs were trained using the nova set, and the molecules from the star set were predicted. The LOO results calculated for the nova set were quite similar to the results calculated in the train/test set protocols using 50% of the molecules and were only about 8% worse than ALOGPS LOO results calculated for this subset. The prediction ability for the star set was very similar to the predicted results, and the same percentage of outliers (2%) was calculated for the training and test sets. The results for the star set were worse compared to ALOGPS LOO results calculated for this set, but they were in agreement with the results predicted for the used training set. Thus, the star set had a rather similar diversity given the nova set but not vice versa.

This suggests that the poor performance of ANN3, XLOGP, and CLOGP programs for prediction of molecules

Table 7. List of the Compounds Having a Calculated and Observed Log *P* Difference Greater than 2.0 by the ANN3 Neural Network for the Star Set[a]

				experimental log *P*			
no.	CAS RN	compound name	NA	PHYSPROP	BioByte	CLOGP	ANN3
1	2517−04−6	2-azetidinecarboxylic acid	7	−2.84	−2.84	−2.97	−0.74
2	1118−68−9	*N*,*N*-dimethylglycine	7	−2.91	−2.91	−2.37	−0.26
3	147−85−3	proline	8	−2.54	−2.5	−2.41	−0.43
4	535−75−1	2-piperidinecarboxylic acid	9	−2.31	−2.31	−1.85	0.06
5	498−95−3	nipecotic acid	9	−2.89	−2.89	−2.2	−0.23
6	498−94−2	isonipecotic acid	9	−3.05	−3.05	−2.85	−0.25
7	26537−53−1	sydnone, 3-(carboxymethyl)-	10	−1.66	−1.66	−1.66	0.53
8	1071−83−6	glyphosate	10	−4	−3.39	−3.69	−1.52
9	56−91−7	4-aminomethylbenzoic acid	11	1.03	−1.55	−1.43	−1.77
10	SRC 1−25−8	propionic acid, 3-(*N*-piperidinyl)	11	−2.45	−2.45	−1.01	0.96
11	1918−02−1	picloram	13	0.3	2.3	2.39	2.37
12	SRC 2−82−7	phenylalaninenmethyl	13	−1.53	−1.53	−1.5	1.39
13	SRC 2−09−3	hexamethyl-methanetricarboxamide	16	−3.09	−3.09	−0.99	−0.53
14	SRC 2−88−9	hexanoicacid62phenylethylamino	17	−1.3	−1.3	0.01	2.53
15	355−68−0	perfluorocyclohexane	18	2.91	2.91	2.75	5.13
16	SRC 2−26−9	trityldifluoroamine	22	3.73	3.73	3.73	6.06
17	68694−11−1	triflumizole	23	1.4	4.5	4.05	3.97
18	7696−12−0	tetramethrin	24	4.73	4.73	4.33	2.46
19	123331−83−9	dibenzo[b,d]pyran-9-carboxaldehyde derivative	28	8.03	8.02	7.2	6
20	60−54−8	tetracycline	32	−1.3	−1.47	−0.91	1.36
21	2030−63−9	clofazimine	33	7.66	7.48	6.69	5.52
22	94050−52−9	flucycloxuron, (E)-isomer	34	6.97	6.97	6.58	4.85
23	SRC 3−29−8	perfluoropmethylcyclohexylpiperidine	36	7.1	7.1	8.99	9.46
24	SRC 2−83−5	O156	38	9.07	9.07	9.24	6.51
25	SRC 3−25−6	peptidebenzyloxycarbonyl analog	56	5.66	5.66	4.91	3.01
26	68325−31−5	digoxin-16′-glucuronide	66	−1.77	−1.77	1.57	0.55

[a] The names of molecules and CAS RN are given according to PHYSPROP database of Syracuse Corp. For some molecules CAS RN are not known, and only internal numbers in this database are provided (SRC #-##-#). It is interesting that 4-aminomethylbenzoic acid, picloram, and triflumizole are not outliers according to the experimental values provided by BioByte Corp.

from the test sets was mainly due to the limited diversity of compounds in the training sets used to develop these methods.

Outlier Detection Procedure. The CLOGP program provides a built-in warning if some of its fragment values were estimated "from scratch". Indeed, 376 out of 558 outliers (and all 274 outliers with "N"-halogen bond) of the CLOGP method were marked with "calculated fragment value". This message indicated the absence of one or of several fragments in the database of the CLOGP program. The CLOGP v. 4.0 estimated the values of such fragments ab initio, i.e., from scratch as described in ref 49. This method also provided other warning messages. In total there were 1387 out of 3479 compounds that had one or another kind of message covering 450 out of 558 outliers. Thus, in general the CLOGP user could be aware about possible problems with the analyzed molecules.

The problematic molecules in the ALOGPS model were identified in a similar way to the CLOGP method but considering E-state indices instead of fragments. The implemented "alarm" procedure warned a user if the analyzed molecule contained one or several E-state indices that were absent in the training set or were calculated only for molecules that were outliers in this set. A set of 394 molecules was identified using this method as possibly problematic in the nova set according to the indices calculated for the star set. Indeed, 357 of these molecules were outliers for ANN3 (when neural networks were trained using the star set) and 356 were outliers for the CLOGP method (Table 6). A similar analysis using indices from the XLOGP training set indicated 1316 problematic molecules in the star and nova sets, i.e., 561 and 501 out of 1138 and 934 outliers calculated for XLOGP and ANN4 (when neural networks were trained using the XLOGP set) methods, respectively. This simple method correctly predicted about 50%−70% of possible outliers.

Tables 7 and 8 show molecules with large prediction errors (more than 2 log units) that were outliers for ANN3 method for star and nova sets. Since the number of such molecules, 391, was quite large for the nova set, only those compounds that were not predicted by the outlier detection procedure described above are shown. These outliers should be considered as true outliers for the interseries prediction, since they did not contain some specific indices that were completely absent in the training set. Some of these outliers were also unsuccessfully predicted by CLOGP method. Indeed, five out of 26 (star set) and 16 out of 32 (nova set) molecules were also outliers for the CLOGP method.

Prediction Ability of Investigated Methods as a Function of a Number of Non-Hydrogen Atoms. The development of modern drug design approaches and especially the use of combinatorial libraries tend to result in compounds with a higher number of atoms in the molecule. It was found that the prediction ability of all methods linearly decreased with the number of non-hydrogen atoms in a molecule (Figure 2). The highest slope 0.010 (log units)/(non-hydrogen atom) was detected for the CLOGP program, while other methods had a similar slope of about 0.007. The results for more than 40 and even 30 atoms had large variability since there were only 163 (1%) and 570 (4%) molecules with more than 40 and 30 non-hydrogen atoms. The data points shown by a circle and a cross corresponded to the inter- and cross-series predictions of different methods, respectively.

Table 8. List of the Compounds Having a Predicted and Observed Log *P* Difference Greater than 2.0 by the ANN3 Neural Network for the Nova Set

no.	CAS RN	compound name	NA	PHYSPROP	CLOGP	ANN3
1	4144−64−3	1H−benzotriazole-1-acetic acid	13	−1.88	0.83	0.37
2	4144−68−7	2-carboxymethylbenzotriazole	13	−1.64	1.14	0.88
3	145−73−3	endothal	13	1.91	−0.34	−0.21
4	83040−20−4	1-pyrrolidinyloxy, 3,4-dicarboxy-2,2,5,5-tetramethyl-, cis	16	0.46	0.57	2.62
5	SRC 3-21-3	5-chlorocyclocytosine	16	−3.1	−1.99	−0.48
6	61566−10−7	ambazone [semicarbazone]	16	1.56	0.46	−1.03
7	113−45−1	methylphenidate	17	0.2	2.56	2.42
8	100853−65−4	TEPA,N−p-Cl benzylidene-N-hydrazin-1-yl	18	0.13	2.22	2.29
9	10054−21−4	1-lauryl-4-carboxy-2-pyrrolidone	21	2.6	5.71	4.84
10	60−92−4	cyclic AMP	22	−2.96	−2.58	−0.91
11	74618−18−1	hydrazinecarboximidamide, N-phenyl-2-(2-quinolinylmethylene)-	22	0.99	2.05	3.19
12	SRC 1−60−7	Br−N(4-^{i}PrPh)carbamoylMe iminodiacetic acid	23	−0.08	2.59	2.08
13	3290−92−4	22Bis(methacryOMe)Bu MeAcrylate	24	4.39	4.31	1.93
14	77−09−8	1(3H)-isobenzofuranone, 3,3-bis(4-hydroxyphenyl)-	24	2.41	2.63	4.72
15	100853−82−5	TEPA,N−p-Cl-benzyliden-N−Ph-hydrazin1yl	24	0.35	4.41	2.64
16	SRC 1−60−6	Br−N(4-BuPh)carbamoylMe iminodiacetic acid	24	0.1	3.25	2.68
17	13936−01−1	indene-13-dion,2[35bis(11diMeEt)Ph]-	25	6.86	6.46	4.8
18	174814−97−2	nalidixic amide,N(2-i-amyl acid), Me ester	25	3.61	1.67	1.55
19	132169−99−4	Br−N(2,6-^{i}PrPh)carbamoylMe iminodiacetic acid	26	0.2	2.22	3.27
20	32487−38−0	8-S−benzyl cyclic AMP	30	−1.15	0.18	0.93
21	41941−66−6	8-(4-Cl−Ph)-S cyclic AMP	30	−0.98	1.1	1.31
22	2488−80−4	digitoxigenin-3-α-sulfate	31	−0.21	2.62	2.29
23	57−62−5	chlortetracycline	33	−0.62	−0.09	1.62
24	79−57−2	oxytetracycline	33	−0.9	−1.28	1.11
25	SRC 3−67−5	D6-anthrquinone dye	33	5.3	3.15	3.2
26	55−56−1	chlorhexidine	34	0.08	4.81	3.07
27	67485−29−4	hydramethylnon	35	2.31	10.74	5.69
28	78182−93−1	(des-^{i}Pr) *N*,*N*-diEtAminoEt-Clofazimine	37	7.97	7.63	5.96
29	132213−90−2	1,4-isoquinolinedione derivative	40	9.07	9.36	6.32
30	103620−84−4	5,12-naphthacenedione derivative	45	0.03	−0.97	2.43
31	149151−33−7	L-phenylalaninamide derivative	60	7.29	7.43	4.5
32	68325−34−8	card-20(22)-enolide derivative	66	−1.1	2.59	1.35

The decrease observed in the accuracy of all methods as the number of non-hydrogen atoms increases is partially associated with the difference in performance of all methods for the star and the nova sets. Indeed, the average numbers of such atoms in the star and in the nova sets were 15.88 and 18.69, respectively. Thus, a molecule of the nova set was on average about 20% larger than a molecule from the star set. Therefore, the estimated error for this set was expected to be on average about 0.02 *MAE* log units larger. The actual differences of about 0.06 log units (calculated by KOWWIN and ALOGPS methods) is three times more than the expected value. Such a large difference could be due to some systematic differences in the experimental protocols used for the sets. Also during the development of the ALOGPS method the SMILES codes of approximately 10 200 compounds from the PHYSPROP database were verified against public databases, including ChemFinder,[50] ChemExper,[51] ChemIdPlus,[52] Sigma/Aldrich online catalog of chemicals,[53] database of the National Institute of Cancer,[54] and Chapman & Hall/CRC Combined Chemical Dictionary as well as the Available Chemicals Directory available at ChemWeb.[55] The detected differences in SMILES codes were reported to the Syracuse Inc. The remaining compounds either did not have a Chemical Abstract Service Registration Number (CAS RN) or were absent in the public databases and therefore could not be checked. We found that more than 30% of the compounds from the nova set could not be verified in the public databases compared to about 20% of the nonverified compounds from the star set. Errors in the SMILES codes could also contribute to the higher prediction errors of all analyzed methods for the nova set. While, these results are rough, they still could provide useful information especially if screening of a large number of data is anticipated. Notice, that the large outliers reported in Tables 7 and 8 had average numbers of 22 and 28 atoms for molecules from the star and nova sets, respectively.

The molecules from the XLOGP training set had an average number of 11.96 non-hydrogen atoms. To some extent this explained the better results (i.e., these molecules were more easy to learn) calculated for this set by XLOGP and ANNs methods. However, due to the same fact the diversity of this set was lower compared to both star and nova sets. Therefore the prediction ability for these sets either by XLOGP or ANNs developed using XLOGP training set was rather low.

ALOGPS vs SciLogP Ultra. SciLogP Ultra[43] is a program developed by SCIVISION[56] using ANNE of 10 networks and 154 E-state indices. The neural networks were trained using 95% of compounds (8837) from the BioByte'97 Starlist (9332). The training set of SciLogP Ultra corresponded to the star set, and, theoretically, a comparison of ANN3 and SciLogP Ultra should be the most adequate. The SciLogP Ultra reported consensus results for the training set and prediction of the independent set of 896 compounds.

As it was already mentioned for other comparisons the consensus results did not provide a proper evaluation of the predictive power of ANNs algorithm. Thus, we can only mention that ANN3 consensus results for the star set $n = 9429$, $r^2 = 0.95$, and $MAE = 0.28$ ($r^2 = 0.95$, $MAE = 0.27$ without 49 outliers) were comparable to the SciLogP Ultra results ($n = 8837$, $r^2 = 0.96$, $MAE = 0.26$) calculated for their training set.

The ANN3 results for the test set, $r^2 = 0.90$, $MAE = 0.42$ ($r^2 = 0.92$, $MAE = 0.39$ without 20 outliers), were similar

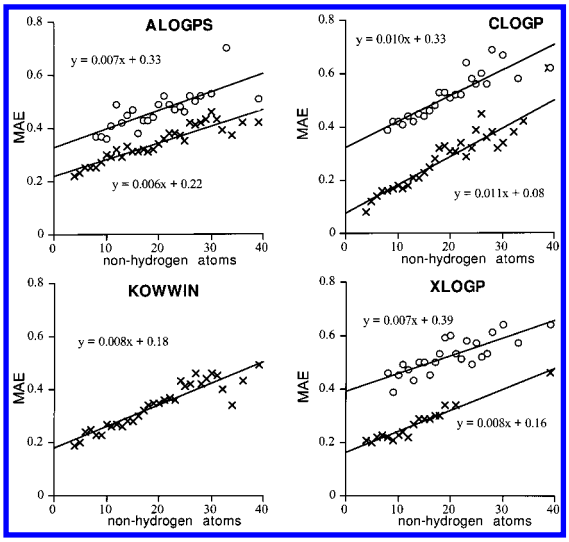


Figure 2. Mean average error (MAE) of different methods as the function of the number of non-hydrogen atoms in a molecule. The results are shown after excluding the outliers identified as molecules with predicted errors above ±1.5 log units. A minimum number of 50 molecules per point was used. If count of molecules for a given number of non-hydrogen atoms was smaller, the results were added to the next number. The crosses and circles correspond to the inter- and to the cross-series analysis, respectively. ALOGPS: crosses corresponded to the predicted results for the 12 908 molecules from the test sets in ANN1/ANN2 analysis. Circles corresponded to the predicted results calculated for the nova set using neural networks trained with the star set (ANN3). CLOGP: crosses and circles corresponded to the CLOGP results for the star and the nova sets, respectively. KOWWIN: calculated results for the 12 908 molecules. XLOGP: crosses and circles corresponded to the results for the XLOGP training set and for the molecules from the nova set.

to the SciLogP results. The SCIVISON also reported results of CLOGP ($r^2 = 0.87$, $MAE = 0.48$) and ACD/logP ($r^2 = 0.81$ $MAE = 0.48$) (see, however, ref 57). A number of compounds, $n = 385$, from the SciLogP Ultra test set had the same SMILES with the molecules from our star set, while the other compounds belonged to the nova set. Thus, it was quite difficult to derive which performance of the analyzed methods has been actually tested by this comparison in terms of the "inter-" and "cross-" series predictions considered above. The performance of ANN3, SciLogP Ultra, and ACD/logP was very similar. All these methods were basically developed using the same training star set except ACD/logP whose training set included 3601 compounds.

On the contrary, the application of ALOGPS provided much better LOO results for the same test set of compounds $r^2 = 0.94$, $MAE = 0.32$ ($r^2 = 0.95$, $MAE = 0.30$ without seven outliers).

DISCUSSION

A new method for calculation of lipophilicity of chemical compounds was introduced in this article, and its performance was compared to other known approaches used in this field.

We emphasize the importance of careful consideration of many different factors taken into account before performing an unbiased comparison between different methods. This is especially important when comparing results calculated by ANNs and traditional methods, such as MLRA. Since the computational power and fitting abilities of ANNs are considerably better than those of traditional methods, only results predicted by both methods should be taken into account. The new method, ALOGPS 2.0, calculated better results compared to the IA_LOGP and SciVision Ultra using their test sets. The results calculated for the validated set would also provide an estimation of ANNs performance; however, such results could be biased if small data samples are available for ANNs training, as it was discussed in ref 35. On the contrary, the consensus results are depending on the training algorithms and on the number of ANNs parameters, including neural network architecture and the number of molecules in the training/validated sets. Thus, by our opinion, the consensus results should not be used to compare predictive performance even between two neural network methods. This is a very important remark, because ANNs become a standard method in chemistry and in drug design.[58,59]

The performance of neural network algorithm for the interseries prediction (i.e. to predict molecules with the same diversity that were used to develop the method) was better or similar to the performance of methods that were developed using the same sets. In fact, ALOGPS LOO results calculated for 12 908 molecules were better than KOWWIN results. The LOO results calculated for the star and XLOGP training sets were similar to the results reported for such sets by CLOGP and XLOGP. This comparison was biased in favor of other methods, since the neural network LOO results were calculated when ANNS **were predicting** all molecules, while the regressed/fitting results were reported for some of these data by KOWWIN, CLOGP, and XLOGP methods. A more careful estimation could provide further 3−5% increase of the predictive ability of our method. For example, the LOO results of XLOGP method were $RMSE = 0.37$ compared to $RMSE = 0.35$ ($RMSE = 0.33$ without four outliers) reported in this study.

The most adequate comparison of ALOGPS method could be carried out with KOWWIN. KOWWIN was initially developed using 2351 molecules selected from the set of 8506 molecules.[5] The diversity of molecules in this training set was adequate to represent all molecules both from the nova and the star sets. Thus, the predicted and not fitted results were reported by this method for more than 80% of analyzed molecules. A better comparison between ALOGPS and KOWWIN should include the KOWWIN training set for neural network learning.

ALOGPS results calculated for the cross-series predictions were better than the results obtained by CLOGP and XLOGP. It should be mentioned that CLOGP and XLOGP did not predict chlorides, iodides, and bromides. However, all these molecules were outliers for all methods in the cross-series predictions and were separately treated in the last column of Table 6. Thus, the exclusion of such molecules would not change the statistical results but would merely decrease the number of detected outliers. In fact by using the same training data sets, the prediction abilities for cross-series analysis for different methods are quite similar. Notice, that XLOGP and ANNs results for the nova set are more similar if neural networks were trained with XLOGP training set (ANN4) compared to ANNs trained using star set (ANN3). A 5-fold increase of the training set sample size (i.e., use of star set of 9429 molecules compared to XLOGP set of 1853 molecules) provided less than 10% gain in the prediction performance of ANNs for the nova set. Actually, for the

prediction of the nova set, the predictions of all methods were quite similar.

The model developed in this article is rather statistical. Indeed, we tried to keep the number of variables as small as possible, and only some simple physical chemistry knowledge was used to determine aliases and to group the E-state indices. It is possible that a more detailed and careful grouping of bond-type indices could further improve the results using the same molecules from the PHYSPROP database.

To our knowledge, only several methods, namely CLOGP,[2] VLOGP,[45] and ACD/logP,[6] provide error estimates. The ALOGPS program does not provide such an estimate in a statistical sense (confidence level, etc.). However, like the VLOGP or the CLOGP method, it generates warnings to a user if new molecules contain indices that were never used to train the program or if all molecules with such indices were outliers in the training set. About 50−70% molecules detected by this simple approach were indeed outliers for cross-series prediction including CLOGP and XLOGP methods that do not use the E-state indices. Since the number of possible bond E-state indices is limited, then by the time of development of ALOGPS all molecules with all possible indices will be used. This method will not predict outliers anymore. However, it will be still be possible that molecules with indices whose values are out of the range used in the training set could be still predicted rather poorly. Some of the outliers in the nova set had indices that were out of range for indices given the star set. Such outliers could be detected using, e.g. the optimum prediction space (OPS) of VLOGP method,[45,60] that was shown to be a reliable approach to assess the reliability of the predicted log *P* values according to the distribution of indices from the training set. The methods used to calculate confidence intervals for neural networks[61,62] provide another possibility to detect molecules with large prediction errors. These methods like OPS will predict the outlier defined as the molecules for which the input parameters are out of range (i.e., their own OPS) and are used to develop ANNs, i.e., when extrapolation and not interpolation of new data is required. It is known that in general all methods, and in particular nonparametric methods such as ANNs, provide a poor performance for extrapolation.[63] This can be a considerable problem for the analysis of large molecules, since the increase in size of molecules tends in general to increase the values of the E-state indices. Then, it can be expected that large molecules will be outside of the OPS space, and their prediction will be unreliable. To some extent, this prediction explains the poor performance to the nova set of several methods analyzed in this study, which were developed using the star set. Indeed, since the molecules in the nova set had, on average, a larger size, an extrapolation rather than an interpolation of some molecules was performed. The normalization of indices according to the root of 1/3 degree of molecular weight decreased the dependency of indices on the size of the molecules and provided an important extension of OPS and improved prediction ability of ANNs. It will be interesting to see if a normalization of the indices on molecular weight (or on other similar molecular parameters, such as the number of atoms, etc.) could also improve the prediction ability of large molecules by fragment-based approaches.

It is very clear that an increase in size of the training data set by including new compounds could further improve prediction performance of the ALOGPS as well as of other methods. However, the use of E-state indices provides a much simpler and faster way to develop such programs compared to the other methods. The largest databases of lipophilicity of chemical compounds usually remain the property of pharmaceutical firms and are not accessible for analysis due to confidence restrictions. To this extent E-state indices provide an extraordinary possibility to analyze such compounds without the need to disclose their chemical structure. Indeed, once indices were calculated, there is no need to know the underlining chemical structure of the compounds. The disclosure of the molecular structure knowing that the E-state indices is impossible[22,42] attributes to such a procedure an extremely high level of confidentiality.

The prediction performance of all methods almost linearly decreased with the number of non-hydrogen atoms in the analyzed molecule in a similar way for all methods. Such a behavior was also observed for the prediction of the aqueous solubility[26] of chemical compounds. A similar tendency was recently reported for ALOGP and CLOGP by Ghose et al.[15] Then, this dependency is rather general and should be taken into account when applying the investigated approaches to compounds with a large number of atoms. Thus, even for the interseries analysis, a correct prediction of lipophilicity of large molecules is a difficult task. To this extent, the smallest slope of the error increase calculated for the ALOGPS method has significant practical importance.

The decrease in the prediction power of the methods with an increase in the number of non-hydrogens is likely to be due to an increase of intermolecular interactions in complex structures, such as intramolecular hydrogen bonds and folding of molecules due to London dispersion forces.[64] As a result of these interactions a part of the molecule becomes unavailable for interaction, thus provoking a decrease of prediction ability of the method. This problem could be probably addressed with molecular dynamics models and conformational analysis aimed to detect atoms that are not accessible for the interactions with solvents. A progress in this field can be also expected with the development of the theory of mobile order and disorder (MOD) that has recently successfully predicted physical properties of simple aliphatic compounds in H-bonded solvents,[65] but this theory could not be easily applied to large compounds.

CONCLUSIONS

This article provided a statistical interpretation of the problem of prediction the lipophilicity of chemical compounds. It did not try to enter complex problems of solvation theory, such as the influence of H-bonds, interactions of molecular groups, tautomerization effects, etc. on the log *P* calculations. We just took experimental data from the PHYSPROP database, predicted them using E-state indices and ANNs, and compared the calculated results with other methods.

This heuristic procedure let us obtain important results about the dependency of the prediction power of different methods depending on the training sets. Our results suggested that the prediction performance of methods critically depended on the diversity of the molecular structures used to

develop such methods. Thus, a fair comparison of the predictive power of different methods is impossible without the knowledge of the diversity of their training sets. Otherwise, rather than comparing different methods and approaches there is a risk in comparing the diversities of the different training sets. The methods that use large databases of structures are the most probable winners in the competition for the cross-series predictions. The use of E-state indices and ANNs make it possible to develop such a method in a fast straightforward way even without a knowledge of the analyzed structures.

APPENDIX 1. NEURAL NETWORK TERMINOLOGY

In this section we summarized ANNs terminology[48] that was used in the article.

EPA − efficient partition algorithm. This algorithm provided an improved performance of ANNs by sophisticated selection of learning and validated sets as described in refs 30−32. This algorithm was used to calculate all results reported in this study.

Initial training set − the whole set of molecules used to train neural network. During training a part of this set was used as learning and validation sets.

Learning set − the set used to train (i.e., adjust) neural network weights.

Validation set − the set used to terminate ANNs learning. This set was not used to adjust neural network weights but only to evaluate neural network performance.

Early stopping point − the iteration of ANNs that provided a minimum neural network error for the validation set. The neural network weights were saved, and further neural network training was terminated.

LOO − leave-one-out, the method used to estimate performance of ANNs. The traditional implementation of LOO procedure required very long calculation. However, the fast automatic LOO procedure described in ref 35 does not require additional overhead work to perform this estimation.

ACKNOWLEDGMENT

This study was partially supported by INTAS 97-0168 and 00-363 grants. We thank Dmitry Shakhnin and Tamara Kasheva for their help with verification of SMILES codes of the analyzed molecules. We are grateful to Luhua Lai, Gao Ying, and Renxiao Wang (Peking University) for providing us the source code of their program and fruitful feed-back for development of SMILES to mol2 conversion program. Many thanks to Albert Leo who provided us a free license for use of CLOGP v. 4.0 program and Bill Meylan for his advice and feedback about the verification of the SMILES codes in the PHYSPROP database. We are grateful to Marc Parham and Joseph Votano for giving us the test sets of the IA_LOGP and the SciLogP Ultra programs, respectively. We thank Javier Iglesias (Computer Science Institute, University of Lausanne) for providing us an access to a cluster of Linux computers used in this study.

REFERENCES AND NOTES

(1) Hansch, L.; Leo, A. *Substitutient Constants for Correlation Analysis in Chemistry and Biology*; Wiley: New York, 1979.
(2) Leo, A. Calculating logPoct from Structures. *Chem. Rev.* **1993**, *93*, 1281−1306.
(3) Rekker, R. E. *Hydrophobic Fragment Constant*; Elsevier: New York, 1977.
(4) Klopman, G.; Li, J.-Y.; Wang, S.; Dimayuga, M. Computer Automated logP Calculations Based on an Extended Group Contribution Approach. *J. Chem. Inf. Comput. Sci*. **1994**, *34*, 752−781.
(5) Meylan, W. M.; Howard, P. H. Atom/Fragment Contribution Method for Estimating Octanol−Water Partition Coefficients. *J. Pharm. Sci*. **1995**, *84*, 83−92.
(6) Petrauskas, A. A.; Kolovanov, E. A. ACD/LogP Method Description. *Persp. Drug Discov. Design* **2000**, *19*, 1−19.
(7) Klopman, G.; Iroff, L. Calculation of Partition Coefficients by the Charge Density Method. *J. Comput. Chem*. **1981**, *2*, 157.
(8) Bodor, N.; Huang, M.-J. An Extended Version of a Novel Method for Estimation of Partition Coefficients. *J. Pharm. Sci*. **1992**, *81*, 272−281.
(9) Haeberlin, M.; Brinck, T. Prediction of Water-Octanol Partition Coefficients Using Theoretical Descriptors Derived from the Molecular Surface Area and the Electrostatic Potential. *J. Chem. Soc., Perkin Trans*. 2 **1997**, 289−294.
(10) Bodor, N.; Buchwald, P. Molecular Size Based Approach to Estimate Partition Properties for Organic Solutes. *J. Phys. Chem.* **1997**, *101*, 3404−3412.
(11) Breindl, A.; Beck, N.; Clark, T.; Glen, R. C. Prediction of the *n*-Octanol/Water Partition Coefficient, logP, Using a Combination of Semiempirical MO−Calculations and a Neural Network. *J. Mol. Model.* **1997**, *3*, 142−155.
(12) Buchwald, P.; Bodor, N. Octanol−Water Partitioning: Searching for Predictive Models. *Current. Med. Chem.* **1998**, *5*, 353−380.
(13) Broto, P.; Moreau, G.; Vandycke, C. Molecular Structures, Perception, Autocorrelation Descriptor and SAT Studies; System of Atomic Contributions for the Calculation of Octanol−Water Partition Coefficient. *Eur. J. Med. Chem.* **1984**, *19*, 71−78.
(14) Ghose, A. K.; Pritchett, A.; Grippen, G. M. Atomic Physicochemical Parameters for Three-Dimensional Structure Directed Quantitative-Activity Relationships. III. Modeling Hydrophobic Interactions. *J. Chem. Inf. Comput. Sci*. **1988**, *9*, 80−90.
(15) Ghose, A. K.; Viswanadhan, V. N.; Wendoloski, J. J. Prediction of Hydrophobic (Lipophilic) Properties of Small Organic Molecules Using Fragmental Methods: An Analysis of ALOGP and CLOGP Methods. *J. Phys. Chem.* **1998**, *102*, 3762−3772.
(16) Wang, R.; Fu, Y.; Lai, L. A New Atom-Additive Method for Calculating Partition Coefficients. *J. Chem. Inf. Comput. Sci*. **1997**, *37*, 615−621.
(17) Katritzky, A. R.; Maran, U.; Lobanov, V. S.; Karelson, M. Structurally Diverse Qunatitative Structure−Property Relationship Correlations of Technologically Relevant Physical Properties. *J. Chem. Inf. Comput. Sci*. **2000**, *40*, 1−18.
(18) Mannhold, R.; Dross, K. Calculation Procedures for Molecular Lipophilicity: A Comparative Study. *Quant. Struct.-Act. Relat.* **1996**, *15*, 403−409.
(19) Kier, L. B.; Hall, L. H. An Electrotopological State Index for Atoms in Molecules. *Pharm. Res.* **1990**, *7*, 801−807.
(20) Hall, L. H.; Kier, L. B. Electrotopological State Indices for Atom Types: A Novel Combination of Electronic, Topological, and Valence State Information. *J. Chem. Inf. Comput. Sci*. **1995**, *35*, 1039−1045.
(21) A continuously growing list of these publications is available from the home page of MolconnZ software at http://www.eslc.vabiotech.com/molconn/mconpubs.html.
(22) Kier, L. B.; Hall, L. H. *Molecular Structure Description: The Electrotopological State;* Academic Press: London, 1999.
(23) Huuskonen, J. J.; Villa, A. E. P.; Tetko, I. V. Prediction of Partition Coefficient Based on Atom-Type Electrotopological State Indices. *J. Pharm. Sci*. **1999**, *88*, 229−233.
(24) Huuskonen, J. J.; Livingstone, D. J.; Tetko, I. V. Neural Network Modeling for Estimation of Partition Coefficient Based on Atom-type Electrotopological State Indices. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 947−955.
(25) Syracuse Research Corporation. Physical/Chemical Property Database (PHYSPROP); SRC Environmental Science Center: Syracuse, NY, 1994.
(26) Tetko, I. V.; Tanchuk, V. Yu.; Kasheva, T. N.; Villa, A. E. P.; Estimation of Aqueous Solubility of Chemical Compounds Using E-state Indices. *J. Chem. Inf. Comput. Sci.* **2001**, in press.
(27) Hall Associated Consulting, Quincy, MA.
(28) http://www.eslc.vabiotech.com/molconn.
(29) The "unused" bond is the bond of the highest order (3≫2≫a≫s, i.e., triple≫double≫aromatic≫single) of the analyzed atom except the analyzed bond.
(30) Tetko, I. V.; Villa, A. E. P. Unsupervised and Supervised Learning: Cooperation toward a Common Goal. ICANN95, International Conference on Artificial Neural Networks NEURONIMES'95; Fogelman-Soulie, F., Ed.; EC2 & Cie: Paris, France, 1995; Vol. 2, pp 105−110.

(31) Tetko, I. V.; Villa, A. E. P. Efficient Partition of Learning Datasets for Neural Network Training. *Neural Networks* **1997**, *10*, 1361−1374.
(32) Tetko, I. V.; Villa, A. E. P. An Efficient Partition of Training Data Set Improves Speed and Accuracy of Cascade-Correlation Algorithm. *Neural Processing Lett.* **1997**, *6*, 51−59.
(33) Tetko, I. V. Associative Neural Network. *Neural Processing Lett.*, submitted.
(34) Tollenaere, T. SuperSAB: Fast Adaptive Back Propagation with Good Scaling Properties. *Neural Networks* **1990**, *3*, 561−573.
(35) Tetko, I. V.; Livingstone, D. J.; Luik, A. I. Neural Network Studies. 1. Comparison of Overfitting and Overtraining. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 826−833.
(36) Tetko, I. V.; Villa, A. E. P. An Enhancement of Generalization Ability in Cascade Correlation Algorithm by Avoidance of Overfitting/Overtraining Problem. *Neural Processing Lett.* **1997**, *6*, 43−50.
(37) BioByte Corp., 201 W. Fourth Street, Claremont, CA.
(38) http://esc-plaza.syrres.com/interkow/DLForm.htm.
(39) XLOGP v2.0 is available by anonymous FTP to ftp2.ipc.pku.edu.cn, directory "pub/software/xlogp".
(40) Sybyl is software package of Tripos, http://www.tripos.com.
(41) Tetko, I. V.; Tanchuk, V. Yu.; Kasheva, T. N.; Villa, A. E. P. Internet Software for Calculation of Lipophilicity and Aqueous Solubility of Chemical Compounds. *J. Chem. Inf. Comput. Sci*. **2001**, *41*, 246−252.
(42) Parham, M.; Hall, L.; Kier. L. Accurate Prediction of LogP Using E-State Indices with Neural Network Analysis. Poster presentation at ACS Meeting, August 22, 2000, Washington, DC, available as Microsoft Power Point file at http://www.logp.com.
(43) Votano, J. R. Use of E-State Atom Indices and Neural Network Algorithms in SciLogP Ultra for the Prediction of *n*-Octanol/Water Partition coefficient, logP. The Second logP Symposium. Lipophilicity in Drug Disposition; Lausanne, March 5−9, 2000, P-B19.
(44) Rousseeuw, P. J.; Leroy, A. M. *Robust Regression & Outlier Detection*; John Wiley & Sons: New York, 1987.
(45) Gombar, V. K.; Enslein, K. Assessment of *n*-Octanol/Water Partition Coefficient: When is the Assessment Reliable? *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 1127−1134.
(46) Parham, M. personal communications.
(47) In the WWW version of Power Point presentation it was indicated that IA_LOGP was developed using 12942 compounds from Syracuse Research Corporation. It is not clear if the authors used the PHYSPROP databases of this company or if some other compounds from other internal databases of the Syracuse Corp. were also analyzed.
(48) In the description of IA_LOGP results, the names of the validation and test sets were interchanged compared to their meaning in the present article. Thus, in their article the tests sets were used to select the best ANNs models, while the validation set (sometimes also named as the "independent validation set") was predicted after termination of the neural network learning. It should be noted, that the ANNs terminology is not yet completely defined and, sometimes, indeed the set used to select the ANNs model is referred to as the test set (Reed, R.; Marks II, R. J., Neurosmithing: Improving Neural Network Learning. In *The Handbook of Brain Theory and Neural Networks*; Arbib, M., Ed.; MIT: Cambridge, 1995, 1118pp.). However, to our knowledge the terminology used in this article is generally accepted in the neural network literature (see, e.g. Prechelt, L. Automatic Early Stopping Using Cross Validation: Quantifying the Criteria. *Neural Networks* **1998**, *11*, 761−767. Anders, U.; Korn, O. Model Selection in Neural Networks. *Neural Networks* **1999**, *12*, 309−322.). Many definitions of terms used in neural network field can be also found at http://www.faqs.org/faqs/ai-faq/neural-nets/part3/
(49) Leo, A. J.; Hoekman, D. Calculating log P(oct) with no Missing Fragments; The problem of Estimating New Interaction Parameters. *Persp. Drug Discov. Design* **2000**, *18*, 19−38.
(50) http://www.chemfinder.com.
(51) http://www.chemexper.com.
(52) http://chem.sis.nlm.nih.gov/chemidplus.
(53) http://www.sigma-aldrich.com.
(54) http://www2.chemie.uni-erlangen.de/ncidb2/index.html.
(55) http://www.chemweb.com/databases.
(56) SciVision, Inc.: Burlington, MA.
(57) Williams, A.; Kolovanov, E. A Rebuttal Regarding Recent Comparison of SciLogP Ultra (SciVision Inc.) with ACD/LogP Prediction Software, March 2000, available at http://www.acdlabs.com/publish/publ_pres00.html.
(58) Zupan, J.; Gasteiger, J. *Neural Networks for Chemistry and Drug Design: An Introduction*, 2nd ed.; VCH: Weinheim, 1999.
(59) Devillers, J. *Neural Networks in QSAR and Drug Design*; Academic Press: London, 1996.
(60) Gombar, V. K. Reliable Assessment of logP of Compounds of Pharmaceutical Relevance. *SAR QSAR Environ. Stud.* **1999**, *10*, 371−380.
(61) Rivals, I.; Personnaz, L. Construction of Confidence Intervals for Neural Networks Based on Least Square Estimation. *Neural Networks* **2000**, *13*, 463−484.
(62) Heskes, T. Practical Confidence and Prediction Intervals. In *Advances in neural information processing system*; Mozer, M., Jordan, M., Petsche, T., Eds.; MIT Press: Cambridge, MA, Vol. 9, pp 176−182.
(63) Geman, S.; Bienenstok, E.; Doursat, R. Neural Networks and the Bias/Variance Dilemma. *Neural Computation* **1992**, *4*, 1−58.
(64) Leo, A. J.; Hansch, H. Role of Hydrophobic Effects in Mechanistic QSAR. *Persp. Drug Discov. Design* **1999**, *17*, 1−25.
(65) Ruelle, P. Towards a Comprehensive Non-Ergodic Treatment of H−Bonds and Hydrophobicity in Real Solutions: The Mobile Order and Disorder Theory. *Persp. Drug Discov. Design* **1999**, *17*, 61−96.
(66) Wang, R.; Gao, Y.; Lai, L. Calculating partition coefficient by atom-additive method. *Persp. Drug Discov. Design* **2000**, *19*, 47−66.

CI010368V