Anal Bioanal Chem (2007) 389:1267–1281
DOI 10.1007/s00216-007-1502-x

ORIGINAL PAPER

Benchmarking and validating algorithms that estimate pK_a values of drugs based on their molecular structures

Milan Meloun · Sylva Bordovská

Received: 20 April 2007 / Revised: 4 June 2007 / Accepted: 11 July 2007 / Published online: 4 August 2007

Abstract The REGDIA regression diagnostics algorithm in S-Plus is introduced in order to examine the accuracy of pK_a predictions made with four updated programs: PALLAS, MARVIN, ACD/pKa and SPARC. This report reviews the current status of computational tools for predicting the pK_a values of organic drug-like compounds. Outlier predicted pK_a values correspond to molecules that are poorly characterized by the pK_a prediction program concerned. The statistical detection of outliers can fail because of masking and swamping effects. The Williams graph was selected to give the most reliable detection of outliers. Six statistical characteristics (F_{exp}, R^2, R^2_p, *MEP*, *AIC*, and $s(e)$ in pK_a units) of the results obtained when four selected pK_a prediction algorithms were applied to three datasets were examined. The highest values of F_{exp}, R^2, R^2_p, the lowest values of *MEP* and $s(e)$, and the most negative *AIC* were found using the ACD/pK_a algorithm for pK_a prediction, so this algorithm achieves the best predictive power and the most accurate results. The proposed accuracy test performed by the REGDIA program can also be applied to test the accuracy of other predicted values, such as log *P*, log *D*, aqueous solubility or certain physicochemical properties of drug molecules.

Keywords pK_a prediction · pK_a accuracy · Dissociation constants · Outliers · Influential points · Residuals · Goodness-of-fit · Williams graph

M. Meloun (✉) · S. Bordovská
Department of Analytical Chemistry,
Faculty of Chemical Technology, Pardubice University,
532 10 Pardubice, Czech Republic
e-mail: milan.meloun@upce.cz

Introduction

Predicting molecular properties and modeling chemical, biological and pharmaceutical effects are among the most challenging aims in modern chemistry and pharmacology. Effects are closely related to molecular properties, which can be calculated or predicted from the molecular structure using particular methods. The important influence of the degree of ionization on the biological behavior of chemical substances, namely drugs, is well established, and one of the fundamental properties of any organic drug molecule, the pK_a value, determines the degree of dissociation in solution—it is a measure of the strength of an acid or a base. Physicochemical properties such as acid pK_a value, solubility, permeability and protein binding are closely related to drug absorption, distribution, metabolism and excretion (ADME). During the drug development phase, timely knowledge of these properties of compounds aids candidate selection, formulation design and drug delivery. On the other hand, the pK_a value of an organic compound is also a vital piece of information in environmental exposure assessment, as it can be used to define the degree of ionization and the resulting propensity for sorption into soil and sediment which, in turn, can determine a compound's mobility, reaction kinetics, bioavailability, complexation, etc. In the world of chemometrics or chemoinformatics, there is immense interest in developing new and better software for pK_a prediction. To obtain a significant correlation and an accurate predicted pK_a value, it is crucial to employ the appropriate structure descriptors. Numerous studies have considered (and various approaches have been applied to) the prediction of pK_a, but mostly without a rigorous statistical test of pK_a accuracy. Efficient software packages have been implemented to predict the values; due to their fragment-based approach, however, they are

inadequate when the fragments present in the molecule under study are absent from the database. It is clear that such approaches to pK_a prediction are only accurate when the compounds that are under investigation are very similar to those available in the training set. Xing and Glen [1, 2] used molecular tree structured fingerprints of key fragments and atom types in a hierarchical tree form to correlate pK_a values with basic and acidic centers, a method based on the SYBYL informatics approach [1, 3, 4]. The ACD/pK_a module [5] uses fragment methods to build a large number of equations with experimental or calculated electronic constants that can be used to predict pK_a values [5–8]. The MARVIN software developed by ChemAxon [9] and the PALLAS software [10] are free of charge for academic use, and are therefore preferred in an academic setting to the advanced commercial software of ACD/Labs, provided that the performance is also satisfactory, while in an industrial setting other critical aspects of the software might be just as important as the prediction (which should be as accurate as possible), such as possibility of automation and batch processing, integration of in-house proprietary databases, reliability of the software, and long-term commitment and maintenance of the software producer. Comparative molecular field analysis (CoMFA) has been used to model pK_a values for small sets of structures of between 30 and 50 molecules drawn from specific chemical series [11–13]. In 1981, Perrin et al. [14] published a book on pK_a prediction which is still widely used. An artificial neural network (ANN) was successfully used to predict the pK_a values of various acids with diverse chemical structures using the QSPR relationship [15]. A method called quantum topological molecular similarity (QTMS) that can be used for the construction of a variety of medical, ecological and physical organic QSPRs and predicted pK_a values was proposed fairly recently [16]. The SPARC (SPARC Performs Automated Reasoning in Chemistry) program [17] predicts numerous physical properties and chemical reactivity parameters for a large number of organic compounds strictly from their molecular structures. SPARC applies a mechanistic perturbation method to estimate the pK_a value based on a number of models that account for electronic effects, salvation effects, hydrogen bonding effects, and the influence of temperature. The user only needs to know the molecular structure of the compound to predict the property of interest. SPARC web-based calculators have been used by many academics and the employees of chemical/pharmaceutical companies throughout the world. It has been announced that the free, web-based version of SPARC performs 50,000–100,000 calculations per month. The SPARC pK_a calculator has been highly refined and exhaustively tested. Unfortunately, to date no reliable method for predicting pK_a values over a wide range of molecular structures, including simple compounds and for complex molecules such as drugs and dyes, has been made available.

In this context, an examination of the statistical accuracy of the predicted pK_a value would appear to be an important approach. The regression diagnostics algorithm REGDIA [18] in S-Plus [19] has already been developed in order to examine the accuracy of the pK_a values predicted by four commonly used algorithms, PALLAS, MARVIN, PERRIN and SYBYL. Outlier predicted pK_a values correspond to molecules that are poorly characterized by the pK_a prediction program considered. The statistical detection of outliers can fail because of masking and swamping effects. Of the seven most efficient diagnostic plots, the Williams graph is considered to give the most reliable detection of outliers. Six statistical characteristics (F_{exp}, R^2, R^2_P, *MEP*, *AIC*, and $s(e)$ in pK_a units) of the results obtained when all four pK_a prediction algorithms were applied to three datasets were examined. The highest values of F_{exp}, R^2, and R^2_P, the lowest values of *MEP* and $s(e)$, and the most negative *AIC* were obtained for the PERRIN pK_a prediction algorithm, which indicates that this algorithm yields the best predictive power and the most accurate results. The proposed accuracy test performed by the REGDIA program can also be extended to test the accuracy of prediction for other values, such as log *P*, log *D*, aqueous solubility or other physicochemical properties.

The aim of this work was to compare the accuracy of the results from the four predictive algorithms when applied to three different literature datasets, using a tool to investigate whether the pK_a prediction method in question leads to a sufficiently accurate estimate of the pK_a value (i.e., the correlation between the predicted p$K_{a,pred}$ value and the experimental value p$K_{a,exp}$ is usually high). In this investigation, linear regression models were used to interpret the essential features of a p$K_{a,pred}$ dataset. Some difficulties associated with this investigation involve the detection and elucidation of outlying p$K_{a,pred}$ values in the predicted pK_a data; p$K_{a,pred}$ outliers can strongly influence the regression model, especially when using least squares criteria.

Methods

Software and data used

The ionization models were developed using a combination of descriptors that were mapped onto the molecular tree constructed around the ionizable center, using the four different algorithms studied. Most of the work was carried out in the PALLAS [10], MARVIN [9], ACD/pK_a [5] and

SPARC [17] software packages. These largely predict pK_a based on chemical structure, and so their reliability reflects the accuracy of the underlying experimental data. In most software, the input is the chemical structure drawn in graphical mode. The REGDIA algorithm in S-Plus [19] was applied to create regression diagnostic graphs and compute regression-based characteristics. Various diagnostic measures that were designed to detect individual p$K_{a,pred}$ outliers that may differ from the bulk of the data were used. The main difference between the use of regression diagnostics and classical statistical tests in REGDIA is that there is no need for an alternative hypothesis, because all types of deviations from the ideal state are discovered.

Regression diagnostics for examining the pK_a accuracy in REGDIA

The examination of pK_a data quality involves the detection of *influential points* in the proposed regression model p$K_{a,pred}$=β_0+β_1 p$K_{a,exp}$, which cause many problems in regression analysis by shifting the parameter estimates or increasing the variance of the parameters [20]. These points correspond to p$K_{a,pred}$ outliers, which differ from the other points in terms of their *y*-axis values, where *y* stands in all of the following relations for p$K_{a,pred}$. The benefits of analyzing various types of diagnostics graphs using the REGDIA program in order to detect inadequacies in the model or influential points in the data, have been described in detail previously [18, 20]. The following descriptive statistics of the residuals can be used for a numerical goodness-of-fit evaluation in REGDIA (see page 290 in volume 2 of [20]):

(1) The *residual bias* is the arithmetic mean of the residuals $E(\widehat{e})$ and should equal zero.

(2) The *square root of the residual variance* $s^2(\widehat{e}) = RSS(b)/(n-m)$ is used to estimate the *residual standard deviation*, $s(\widehat{e})$, where $RSS(\boldsymbol{b})$ is the residual square sum, and should be of the same magnitude as the random error s(p$K_{a,pred}$), as it is valid that $s(\widehat{e}) \approx s$ (p$K_{a,pred}$).

(3) The *determination coefficient* R^2, calculated from the *correlation coefficient* R and multiplied by 100%, is interpreted as the percentage of all of the points that agree with the proposed regression model.

(4) One of the most efficient criteria is the *mean quadratic error of prediction* $MEP = \frac{\sum_{i=1}^{n}\left(y_i - x_i^T b_{(i)}\right)^2}{n}$, where $\boldsymbol{b}_{(i)}$ represents the estimated regression parameters when all points except the *i*th are used and $\boldsymbol{x}_i$ (here p$K_{a,exp,i}$) is the *i*th row of matrix p$\boldsymbol{K}_{a,exp}$. The statistic *MEP* uses a predicted value $\widehat{y}_{P,i}$ (here p$K_{a,pred,i}$) obtained from an estimate derived without including the *i*th point.

(5) The *MEP* can be used to express the *predicted determination coefficient*, $\widehat{R}_P^2 = 1 - \frac{n \times MEP}{\sum_{i=1}^{n} y_i^2 - n \times \bar{y}^2}$.

(6) Another statistical characteristic is derived from information and entropy theory, and is known as the *Akaike information criterion*, $AIC = n \ln\left(\frac{RSS(b)}{n}\right) + 2m$, where n is the number of data points and m is the number of parameters (for a straight line, m=2). The best regression model is considered to be that in which the *MEP* and *AIC* values are minimized and the value of R_P^2 is maximized.

Individual estimates **b** of parameters **β** are then tested for statistical significance using the Student *t*-test. The *Fisher–Snedecor F-test of the significance of the proposed regression model* is based on the testing criterion $F_R = \widehat{R}^2(n-m)\Big/\left[\left(1-\widehat{R}^2\right)(m-1)\right]$ which has a Fisher–Snedecor distribution with $(m-1)$ and $(n-m)$ degrees of freedom, where R^2 is the determination coefficient. The null hypothesis H_0: R^2=0 may be tested using F_R, and this constitutes a test of the significance of all of the regression parameters **β**.

The quality of the data and the model can be assessed directly from a scatter plot of p$K_{a,pred}$ vs. p$K_{a,exp}$. A variety of plots have been widely used in REGDIA regression diagnostics [18], but the most efficient diagnostic seems to be the *Williams graph* with two boundary lines. The first line is horizontal, and points above this line are detected as the outliers: $y=t_{0.95}(n-m-1)$. The second line is vertical, and points located on its right side are detected as the high leverages: $x=2m/n$. Note that $t_{0.95}(n-m-1)$ is the 95% quantile of the Student distribution with $(n-m-1)$ degrees of freedom. The Williams graph contains the diagonal elements on its *x*-axis and the jackknife residuals on its *y*-axis.

Experimental

Procedure for examining the accuracy

The procedure for examining influential points in the data, and for constructing a linear regression model using the REGDIA program, has been described in detail previously [18]. The least squares straight-line fitting of the proposed regression model, p$K_{a,pred}$=β_0 + β_1 p$K_{a,exp}$ (with a 95% confidence interval) and the regression diagnostics for identifying outlying p$K_{a,pred}$ values detect suspicious points (S) or outliers (O) using the preferred Williams graph. The statistical significance of both parameters β_0 and β_1 of the straight-line regression model p$K_{a,pred}$=β_0(s_0, **A** or **R**) + β_1(s_1, **A** or **R**) p$K_{a,exp}$ is tested in REGDIA using the Student *t*-test, where **A** or **R** refer to whether the tested null

Table 1 pK_a values (experimental and predicted) of the compounds in the three validation datasets used in this study

i	Name	$pK_{a,exp}$	$pK_{a,pred}$(PALLAS)	$pK_{a,pred}$(MARVIN)	$pK_{a,pred}$(ACD)	$pK_{a,pred}$(SPARC)
Dataset a [6]						
1	Atropine	9.9	8.94	9.35	9.98	8.92
2	Chlorothiazide. pK_1	6.5	7.81	7.1	6.6	8.71
3	Chlorothiazide. pK_2	9.5	9.03	9.17	9.44	13.17
4	Chlorpromazine	9.3	9.71	9.2	9.43	9.36
5	Cimetidine	6.8	6.71	6.91	6.72	5.12
6	Diazepam	3.3	2.05	2.92	3.4	2.12
7	Diphenhydramine	9	9.62	8.87	8.76	8.91
8	Disopyramide	10.4	9.92	10.42	10.1	8.62
9	Flufenamic acid	3.9	3.92	3.88	3.65	3.47
10	Furosemide	3.9	4.06	4.25	3.04	2.64
11	Haloperidol	8.3	8.21	8.96	8.25	7.84
12	Imipramine	9.5	9.73	9.2	9.49	9.67
13	Lidocaine	7.94	8.03	7.45	8.53	7.86
14	Phenobarbital. pK_1	7.44	7.4	7.54	7.63	7.77
15	Phenobarbital. pK_2	12.2	***	11.2	12.23	12.14
16	Phenytoin	8.3	8.06	9.19	8.33	9.11
17	Procainamide	9.4	9.38	9.04	9.86	9.12
18	Propranolol	9.5	10.08	9.67	9.14	9.43
19	Tetracaine. pK_1	2.39	3.82	3.48	1.59	1.83
20	Tetracaine. pK_2	8.49	8.13	8.42	8.24	8.76
21	Trimethoprim	7.2	7.28	7.16	7.34	6.07
Dataset b [31]						
22	Benzoic acid	4.21	4.2	4.08	4.2	3.07
23	4-methoxyphenol	10.27	10.17	9.94	10.4	10.13
24	4-ethoxyphenol	10.25	10.46	9.93	10.44	10.11
25	4-propoxyphenol	10.27	10.23	9.93	10.34	10.11
26	4-butoxyphenol	10.26	10.3	9.93	10.33	10.11
27	4-pentoxyphenol	10.13	10.68	9.93	10.32	10.11
28	Phenol	10.01	9.92	10.02	9.86	10.01
29	4-chlorophenol	9.45	9.38	9.26	9.47	9.38
31	3,4-dichlorophenol	8.22	8.56	8.96	8.56	8.52
32	4-iodophenol	9.45	9.45	9.4	9.3	9.3
33	Quinoline	4.97	4.64	4.62	4.97	4.5
34	3-bromoquinoline	2.74	2.54	2.75	2.53	2.73
35	*N*-methylaniline	4.86	4.92	4.68	4.7	5.11
37	Butobarbitone	8	7.92	7.58	7.95	7.9
38	Amylobarbitone	8.07	7.9	7.58	7.94	7.9
39	Pentobarbitone	8.18	7.4	7.54	8	7.9
40	Quinalbarbitone	8.09	7.85	7.58	7.81	7.71
41	Chlorpromazine	9.24	9.71	9.2	9.43	9.36

42	Pericyazine	8.76	9.04	9.25	8.81	8.18
43	Ketoprofen	4.29	3.49	3.88	4.23	4.27
44	Celiprolol	9.66	10.42	9.66	9.11	9.01
45	Acebutolol	9.41	10.08	9.57	9.11	9.12
46	Propranolol	9.53	10.08	9.67	9.14	9.43
Dataset c [8, 30]						
47	Atenolol	9.6	10.08	9.67	9.16	9.28
48	Captopril	3.48	1.8	3.52	3.82	4.49
49	Diclofenac sodium	3.99	4.48	4	4.18	4.07
50	Diltiazem	8.02	8.41	8.18	8.91	8.14
51	Enalapril	5.5	1.8	5.19	5.58	5.06
52	Famotidine*	6.78	10.26	8.44	7.93	5.77
53	Flurbiprofen	4.33	3.03	4.37	4.14	4.19
54	Hydrochlorothiazide	10.17	9.03	9.96	9.49	10.74
56	Labetalol	9.42	10.05	9.8	9.2	10.02
57	Metoprolol	9.56	10.08	9.67	9.18	9.34
58	Nadolol	9.67	10.42	9.76	9.17	9.15
59	Naproxen	4.69	4.06	4.19	4.4	4.35
60	Naproxen sodium	4.74	4.06	4.19	4.4	4.35
61	Nortriptyline	10.11	9.98	10.47	10.08	10.12
62	Piroxicam*	2.33	4.16	1.79	3.6	3.05
63	Propoxyphene HCl	9.08	8.95	9.52	9.19	9.08
64	Propranolol	9.53	10.08	9.67	9.15	9.43

*** not estimated

* indicates that tautomeric forms may interfere

hypothesis H_0: $\beta_0=0$ vs. H_A: $\beta_0\neq0$ and H_0: $\beta_1=1$ vs. H_A: $\beta_1\neq1$ is either **A**ccepted or **R**ejected. The standard deviations s_0 and s_1 of the actual parameters β_0 and β_1 are estimated. A statistical test of the total regression is performed using a Fisher–Snedecor *F*-test, and the calculated significance level *P* is enumerated. The correlation coefficient *R* and the determination coefficient R^2 are computed. The mean quadratic error of prediction *MEP*, the Akaike information criterion *AIC* and the predictive coefficient of determination R^2_P (a percentage) are calculated to examine the quality of the model. Based on whether the conditions required for the least-squares method are fulfilled, and the results of the regression diagnostics, a more accurate regression model without outliers is constructed, and its statistical characteristics are examined. Outliers should also be elucidated.

Datasets

Three different validation datasets (Dataset a, Dataset b and Dataset c), taken from the literature [21–23], were used to examine the accuracies of the four different algorithms. The authors then used the PALLAS, MARVIN and SPARC web calculators and predicted the pK_a values for 64 drugs (Table 1) from three datasets (see Table 1):

(a) The first validation data (Dataset a), assembled from several published studies [6, 24] of the accuracy of pK_a data, were taken from a paper by Rekker et al. [6]; the pK_a values for 21 drugs were available through the ACD/pK_a method and the results are summarized in Table 1. This physiochemical dataset has also been used in other papers [24–29].

(b) The second validation data (Dataset b) employed the ACD/pK_a approach, in which the experimental values reported by Slater et al. [7] were compared; the results are summarized in Table 1. This paper contains the pK_a values of 25 compounds, including six substituted phenols, two substituted quinolines, *N*-methylaniline, five barbiturate derivatives, two phenothiazines, and several other molecules of pharmaceutical interest, which were determined by a potentiometric technique at 25 °C and an ionic strength 0.1 M (KNO_3). These data were derived from the PHYSPROP database (http://www.syrres.com), a commercial dataset of experimental data for physical properties, which references the original papers that the data were compiled from. It has already been used successfully by other researchers to obtain a pK_a validation model. Engvist and Wrede [31] used several rules and filters to eliminate unwanted compounds from a group of

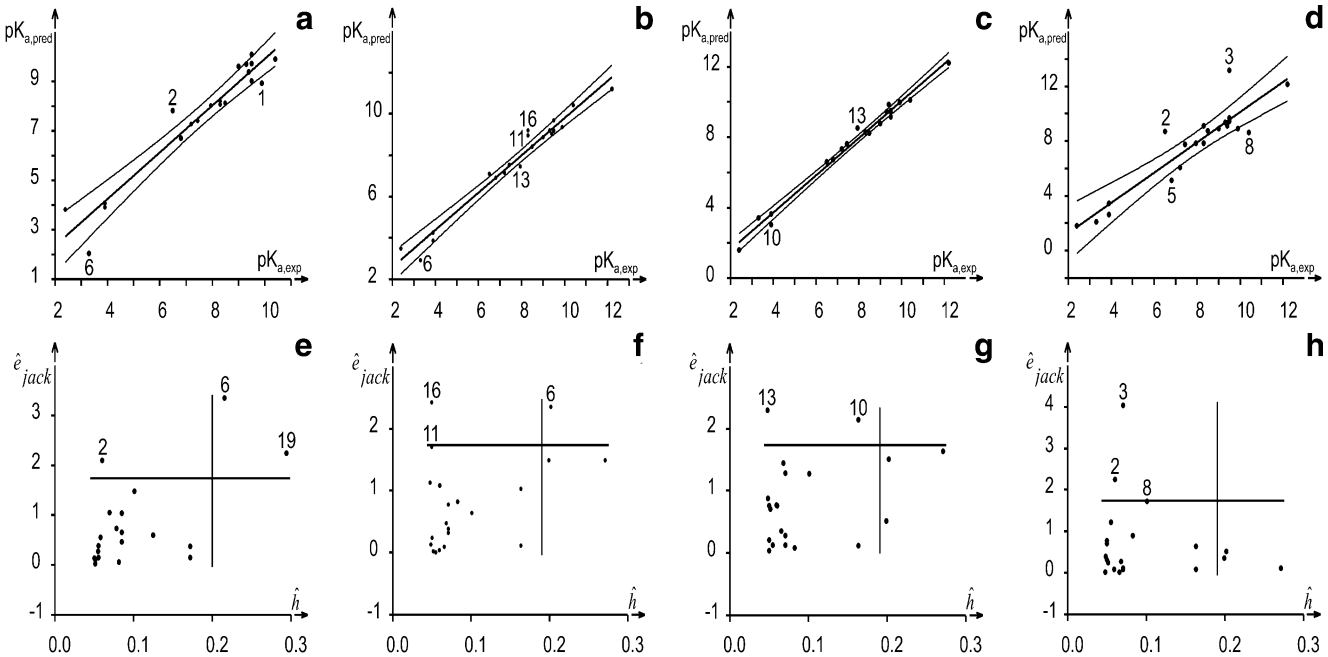

Fig. 1 Comparison of four programs in terms of the predictive ability of the proposed regression model p$K_{a,pred}$=$\beta_0(s_0$, **A** or **R**) + $\beta_1(s_1$, **A** or **R**) p$K_{a,exp}$. *Top*: scatter diagrams of the original data from Table 1 for **Dataset a**. *Bottom*: outlier detection with Williams graphs, with n=21 and α=0.05. **A** or **R** refer to whether the tested null hypothesis H_0: β_0=0 vs. H_A: $\beta_0\neq0$ and H_0: β_1=1 vs. H_A: $\beta_1\neq1$ was **A**ccepted or **R**ejected. The estimated standard deviation of the actual parameter is shown in parentheses. **a, e** PALLAS: $\beta_0(s_0)$=0.46 (0.49, **A**), $\beta_1(s_1)$=0.95 (0.06, **R**), R^2=92.8%, $s(e)$=0.64, F=233.1>4.38, P=9.6×10^{-12}, *MEP*=0.55, *AIC*=−15.6, R^2_P=80.3%, outliers indicated: 2, 6, 19. **b, f** MARVIN: $\beta_0(s_0)$=0.78 (0.32, **R**), $\beta_1(s_1)$=0.90 (0.04, **R**), R^2=96.6%, $s(e)$=0.50, F=537.2>4.38, P=2.15×10^{-15}, *MEP*=0.23, *AIC*=−32.4, R^2_P=91.2%, outliers indicated: 6, 11, 16. **c, g** ACD: $\beta_0(s_0)$=−0.50 (0.23, **R**), $\beta_1(s_1)$=1.06 (0.03, **R**), R^2=98.7%, $s(e)$=0.34, F=1408.7>4.38, P=2.8×10^{-19}, *MEP*=0.12, *AIC*=−45.9, R^2_P=96.6%, outliers indicated: 10, 13. **d, h** SPARC: $\beta_0(s_0)$=−0.93 (0.89, **A**), $\beta_1(s_1)$=1.10 (0.11, **R**), R^2=84.3%, $s(e)$=1.24, F=101.8>4.38, P=4.6×10^{-9}, *MEP*=1.64, *AIC*=11.2, R^2_P=66.6%, outliers indicated: 2, 3, 8

41040 compounds in order to obtain a data sample representing a drug-like chemical space, comprising compounds that were expected to be present in the drug manufacturing pipeline.

(c) The third set of validation data (Dataset c) [8, 30] comprised the results from titrometric measurements made on 18 selected drugs (which were compared to the ACD/pKa predictions for these drugs in [8].

Supporting information available

The complete computational procedures for the REGDIA program [18], input data specimens and corresponding outputs in numerical and graphical forms are available at http://meloun.upce.cz in the blocks *DATA* and *ALGORITHMS*.

Results and discussion

The pK_a values predicted using the four algorithms PALLAS [10], MARVIN [9], ACD/pKa [5] and SPARC [17] were compared with the predicted values of the dissociation constants p$K_{a,pred,}$ and plotted against the experimental p$K_{a,exp}$values for the compounds in the datasets described in Table 1; the resulting scatter plots are shown in Figs. 1, 2 and 3. Even given that SPARC may yield less accurate results for drug-like compounds, there is good agreement between the predicted p$K_{a,pred}$ values and the experimental p$K_{a,exp}$ values in general.

Predictability of pK_a and identifying outliers

The REGDIA program was used to investigate the regression analyses and to discover influential points in the p$K_{a,pred}$ data. The data shown in Table 1 provide a useful way to compare results, and to demonstrate the efficiency of each diagnostic tool for outlier detection. Most the outliers are obviously easier to spot using diagnostic plots than by performing statistical tests of the numerical diagnostic values in the table. These data have been analyzed many times in tests of outlier methods. Plots of the PALLAS-predicted p$K_{a,pred}$ values versus the experimentally observed p$K_{a,exp}$ values for the set of bases and acids examined are shown in Fig. 1a, while the MARVIN-predicted p$K_{a,pred}$ values are shown in Fig. 1b, the ACD/pK_a-predicted p$K_{a,pred}$ values are shown in Fig. 1c, and the SPARC-predicted p$K_{a,pred}$ values in Fig. 1d. The p$K_{a,pred}$ values are distributed evenly around the diagonal, implying

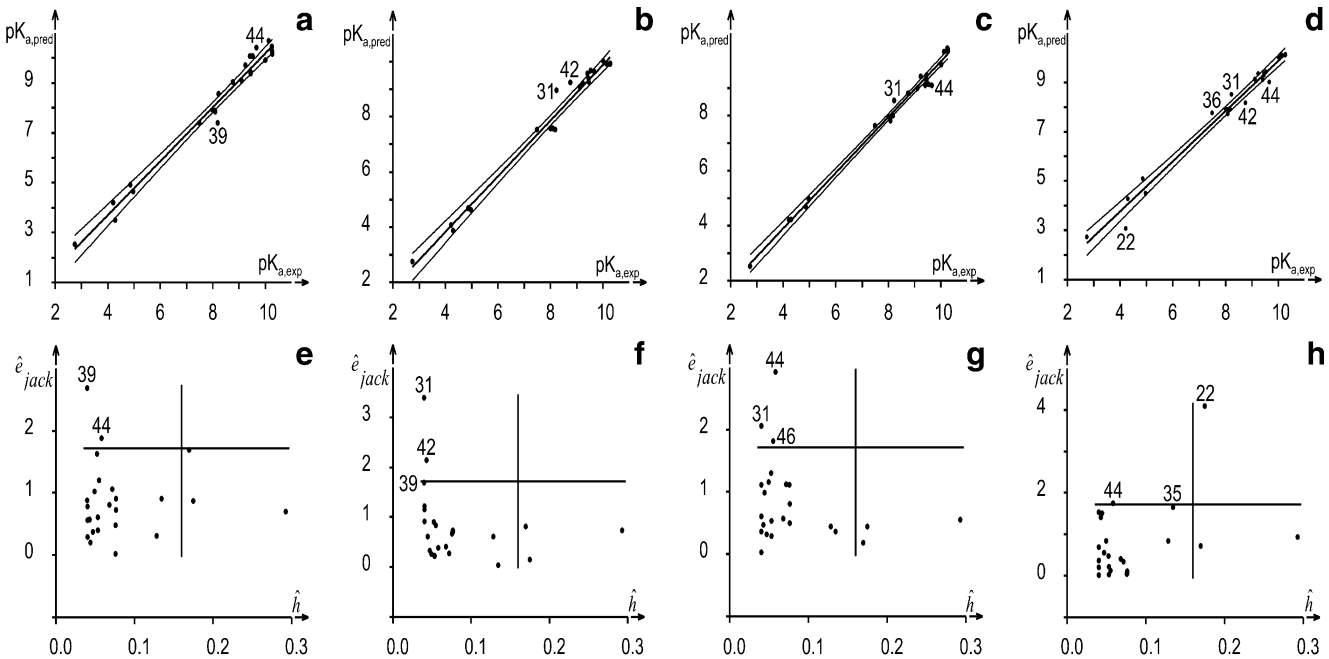

Fig. 2 Comparison of four programs in terms of the predictive ability of the proposed regression model p$K_{a,pred}$=$\beta_0(s_0$, **A** or **R**) + $\beta_1(s_1$, **A** or **R**) p$K_{a,exp}$. *Top*: scatter diagrams of the original data from Table 1 for **Dataset b**. *Bottom*: outlier detection with Williams graphs with n=25 and α=0.05, where **A** or **R** refer to whether the tested null hypothesis H_0: β_0=0 vs. H_A: $\beta_0\neq0$ and H_0: β_1=1 vs. H_A: $\beta_1\neq1$ was **A**ccepted or **R**ejected. The estimated standard deviation of the actual parameter is shown in parentheses. **a, e** PALLAS: $\beta_0(s_0)$=−0.63 (0.27, **R**), $\beta_1(s_1)$= 1.08 (0.03, **R**), R^2=98.0%, $s(e)$=0.39, F=1130.3>4.12, P=4.3×10^{-21}, MEP=0.13, AIC=−50.7, R^2_P=95.4%, outliers indicated: 39, 44. **b, f** MARVIN: $\beta_0(s_0)$=−0.22 (0.25, **A**), $\beta_1(s_1)$=1.01 (0.03, **R**), R^2=98.1%, $s(e)$=0.31, F=1194.1>4.12, P=2.3×$^{-21}$, MEP=0.10, AIC=−55.5, R^2_P= 95.8%, outliers indicated: 31, 39, 42. **c, g** ACD: $\beta_0(s_0)$=−0.14 (0.16, **A**), $\beta_1(s_1)$=1.01 (0.02, **A**), R^2=99.2%, $s(e)$=0.20, F=2787.5>4.12, P=1.5× 10^{-25}, MEP=0.05, AIC=−76.6, R^2_P=98.2%, outliers: indicated 31, 44, 46. **d, h** SPARC: $\beta_0(s_0)$=−0.31 (0.25, **A**), $\beta_1(s_1)$=1.02 (0.03, **R**), R^2= 98.2%, $s(e)$=0.31, F=1218.4>4.12, P=1.8×10^{-21}, MEP=0.12, AIC= −55.6, R^2_P=95.3%, outliers: indicated 22, 35, 44

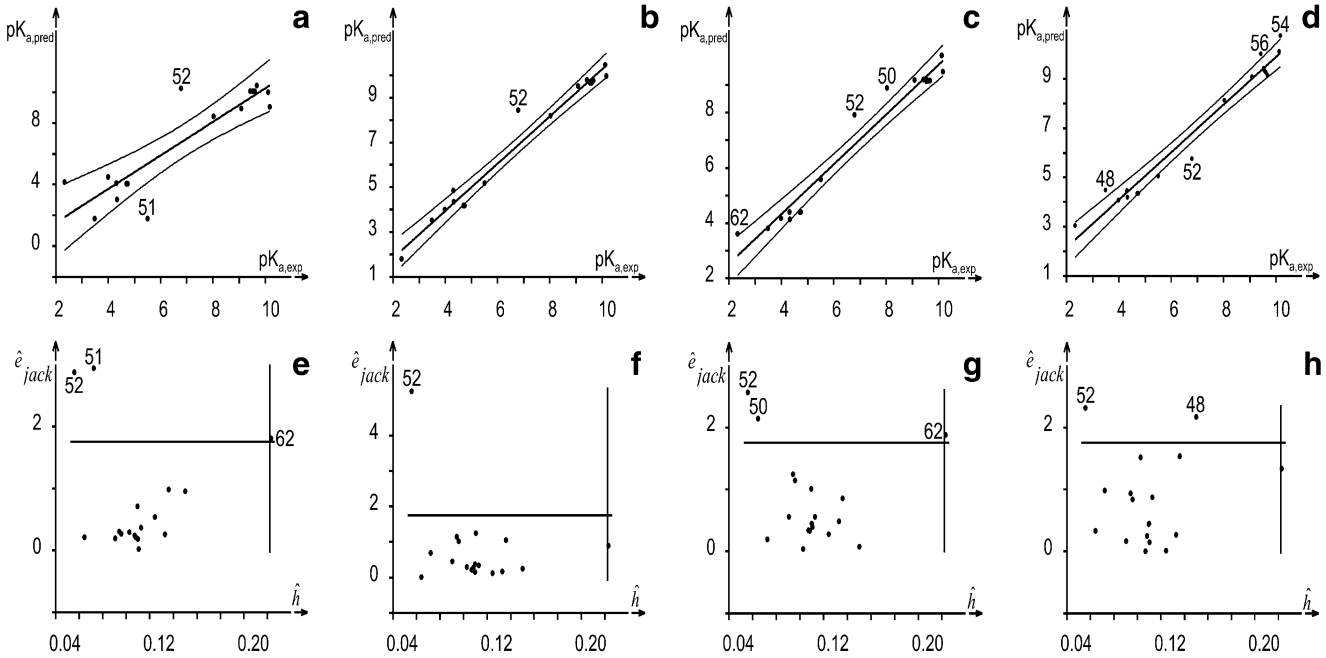

Fig. 3 Comparison of four programs in terms of the predictive ability of the proposed regression model $pK_{a,pred}=\beta_0(s_0$, **A** or **R**) + $\beta_1(s_1$, **A** or **R**) $pK_{a,exp}$. *Top*: scatter diagrams of the original data from Table 1 for **Dataset c**. *Bottom*: outlier detection with Williams graphs, with $n=18$ and $\alpha=0.05$ where **A** or **R** refers to whether the tested null hypothesis H_0: $\beta_0=0$ *vs.* H_A: $\beta_0\neq0$ and H_0: $\beta_1=1$ *vs.* H_A: $\beta_1\neq1$ was **A**ccepted or **R**ejected. The estimated standard deviation of the actual parameter is shown in parentheses. **a, e** PALLAS: $\beta_0(s_0)=-0.64$ (1.00, **A**), $\beta_1(s_1)=1.09$ (0.13, **R**), $R^2=80.4\%$, $s(e)=1.50$, $F=65.6>4.49$, $P=4.7\times10^{-7}$, *MEP*=2.6, *AIC*=17.0, $R^2_P=57.1\%$, outliers indicated: 51, 52, 62. **b, f** MARVIN: $\beta_0(s_0)=-0.27$ (0.33, **A**), $\beta_1(s_1)=1.05$ (0.04, **R**), $R^2=97.3\%$, $s(e)=0.51$, $F=569.4>4.49$, $P=6.2\times10^{-14}$, *MEP*=0.26, *AIC*=−23.1, $R^2_P=93.61\%$, outliers indicated: 52. **c, g** ACD: $\beta_0(s_0)=0.71$ (0.34, **R**), $\beta_1(s_1)=0.90$ (0.05, **R**), $R^2=96.2\%$, $s(e)=0.56$, $F=402.2>4.49$, $P=9.2\times10^{-13}$, *MEP*=0.29, *AIC*=−22.4, $R^2_P=90.7\%$, outliers indicated: 50, 52, 62. **d, h** SPARC: $\beta_0(s_0)=0.22$ (0.33, **A**), $\beta_1(s_1)=0.97$ (0.04, **R**), $R^2=96.7\%$, $s(e)=0.50$, $F=472.5>4.49$, $P=2.6\times10^{-13}$, *MEP*=0.29, *AIC*=−22.8, $R^2_P=91.8\%$, outliers indicated: 48, 52

consistent error behavior for the residual values. The optimal slope β_1 and the intercept β_0 of the linear regression model $pK_{a,pred}=\beta_0+\beta_1\ pK_{a,exp}$ for $\beta_0=0.46$ (0.49, A) and $\beta_1=0.95$ (0.06, A) in the case of PALLAS can be taken to be 0 and 1, respectively, where the standard deviations of the parameters appear in parentheses, and A means that the tested null hypothesis H_0: $\beta_0=0$ vs. H_A: $\beta_0\neq0$ and H_0: $\beta_1=1$ vs. H_A: $\beta_A\neq1$ was accepted.

Another way to evaluate the quality of the regression model proposed by the prediction algorithm is to examine its goodness-of-fit. Most of the acids and bases in the examined sample were predicted with an accuracy of better than one log of their measured values.

Diagnostic graphs for outlier detection can usually be applied to test suspected influential points, and the most efficient of these tools, the Williams graph (Fig. 1e–h), indicates three outliers: 2, 6 and 19. It has previously been concluded [18] that the Williams graph is one of the best diagnostic graphs for outlier detection.

Benchmarking the predicted pK_a values obtained using the four algorithms

Four algorithms—PALLAS [10], MARVIN [9], ACD/pK_a [5] and SPARC [17]—were applied to the datasets in order to predict pK_a values, and their performances in statistical accuracy tests were compared. As expected, the calculated values of $pK_{a,pred}$ agreed well with the experimental values of $pK_{a,exp}$.

Fitted residual evaluation can be an efficient tool to use when building and testing a regression model. The predictive power of each prediction algorithm was evaluated by comparison with experimental data taken from the literature. Altogether, 64 drugs and other organic molecules with complex and diverse structural patterns were used as an external and realistic test set. The quality of the prediction models used by the algorithm was measured using six main statistical parameters, F_{exp}, R^2, R^2_P, *MEP*, *AIC*, and $s(e)$ in pK_a units. The results are presented in Table 2.

Analysis of dataset a

The correlations between the values of pK_a calculated by each of the four algorithms used and the experimental pK_a values with outliers are shown in Table 2. Figure 1a–d illustrate preliminary analyses of the goodness-of-fit for each model, while Fig. 1e–h show the Williams graphs used to identify and remove outliers. In addition to these graphical analyses, the regression diagnostics for the fitness

Table 2 Accuracies of the predicted pK_a values calculated using the four algorithms PALLAS, MARVIN, ACD and SPARC (evaluated using the REGDIA program)

Statistic used	PALLAS		MARVIN		ACD		SPARC	
	With outliers	Without outliers	With outliers	Without outliers	With outliers	Without outliers	With outliers	Without outliers
Regression model proposed: $pK_{a,pred}=\beta_0 + \beta_1\ pK_{a,exp}$								
Intercept $\beta_0(s_0$, A or R)	0.46(0.49, A)	0.28(0.45, A)	0.78(0.32, R)	1.07(0.24, R)	−0.50(0.23, R)	−0.35(0.20, A)	−0.93(0.89, A)	−1.24(0.44, R)
Slope $\beta_1(s_1$, A or R)	0.95(0.06, R)	0.96(0.05, A)	0.90(0.04, R)	0.86(0.03, R)	1.06(0.03, R)	1.04(0.02, R)	1.10(0.11, R)	1.12(0.06, R)
F_{exp} versus $F_{0.95}$(2-1, 21-2)=4.38	233.13	320.17	537.15	889.98	1408.67	1802.41	101.75	408.81
P versus α=0.05 and H_0: regression model is accepted or rejected	9.58×10^{-12}	1.57×10^{-11}	2.15×10^{-15}	1.87×10^{-15}	2.75×10^{-19}	1.08×10^{-18}	4.58×10^{-9}	8.09×10^{-13}
Correlation								
Determination coefficient, R^2 (%)	92.83	95.52	96.58	98.23	98.67	99.07	84.26	96.23
Predicted determination coefficient, R^2_P [%]	80.28	89.33	91.19	95.07	96.57	97.25	66.57	91.13
Prediction ability criteria								
Mean error of prediction, *MEP*	0.55	0.18	0.23	0.11	0.12	0.09	1.64	0.38
Akaike information criterion, *AIC*	−15.58	−28.88	−32.37	−41.47	−45.92	−49.26	11.15	−16.63
Goodness-of-fit test								
E(Λ)	−0.05	0.03	−0.01	0.05	0.07	0.06	0.12	0.37
$s(e)$ in log units of pK_a	0.64	0.40	0.50	0.46	0.34	0.27	1.24	0.65
t_{exp} versus $t_{0.95}(n-m)$=2.09	−0.32	0.34	−0.10	0.49	0.90	0.97	0.44	2.38
P versus α=0.05 and H_0: E(Λ)=0 is accepted or rejected	0.37, A	0.37, A	0.46, A	0.32, A	0.19, A	0.17, A	0.33, A	0.01, R
Outlier detection using the Williams plot								
Number of outliers detected	3	0	3	0	2	0	3	0
Indices of outliers detected	2, 6, 19	–	6, 11, 16	–	10, 13	–	2, 3, 8	–
Regression model proposed: $pK_{a,pred}=\beta_0 + \beta_1\ pK_{a,exp}$								
Intercept $\beta_0(s_0$, A or R)	−0.63(0.27, R)	−0.56(0.23, R)	−0.22(0.25, A)	−0.23(0.16, A)	−0.14(0.16, A)	−0.21(0.12, A)	−0.31(0.25, A)	−0.12(0.20, A)
Slope $\beta_1(s_1$, A or R)	1.08(0.03, R)	1.07(0.03, R)	1.01(0.03, R)	1.01(0.02, R)	1.01(0.02, R)	1.02(0.01, R)	1.02(0.03, R)	1.00(0.02, R)
F_{exp} versus $F_{0.95}$(2-1, 25-2)=4.12	1130.30	1584.81	1194.05	2739.10	2787.45	5517.83	1218.40	1879.07
P versus α=0.05 and H_0: regression model is accepted or rejected	4.29×10^{-21}	2.90×10^{-21}	2.31×10^{-21}	7.08×10^{-23}	1.52×10^{-25}	6.66×10^{-26}	1.84×10^{-21}	2.97×10^{-21}
Correlation								
Determination coefficient, R^2 (%)	98.01	98.69	98.11	99.28	99.18	99.64	98.15	98.95
Predicted determination coefficient, R^2_P [%]	95.36	96.82	95.79	98.26	98.15	99.16	95.27	97.51
Prediction ability criteria								
Mean error of prediction, *MEP*	0.13	0.09	0.10	0.05	0.05	0.02	0.12	0.05
Akaike information criterion, *AIC*	−50.68	−55.02	−55.46	−67.07	−76.63	−82.57	−55.58	−65.64
Goodness-of-fit test								
E(Λ)	−0.04	−0.04	0.13	0.18	0.05	0.03	0.16	0.11
$s(e)$ in log units of pK_a	0.39	0.33	0.31	0.20	0.20	0.15	0.31	0.21
t_{exp} versus $t_{0.95}(n-m)$=2.06	−0.47	−0.59	2.15	4.11	1.31	1.04	2.59	2.54
P versus α=0.05 and H_0: E(Λ)=0 is Accepted or Rejected	0.31, A	0.28, A	0.02, R	0.0002, R	0.10, A	0.16, A	0.008, R	0.010, R

Table 2 (continued)

Statistic used	PALLAS		MARVIN		ACD		SPARC	
	With outliers	Without outliers	With outliers	Without outliers	With outliers	Without outliers	With outliers	Without outliers
Outlier detection using the Williams plot								
Number of outliers detected	2	0	3	0	3	0	3	0
Indices of outliers detected	39, 44	–	31, 39, 42	–	31, 44, 46	–	22, 35, 44	–
Regression model proposed: $pK_{a,pred}=\beta_0 + \beta_1\ pK_{a,exp}$								
Intercept $\beta_0(s_0$, A or R)	−0.64(1.00, A)	−1.29(0.53, R)	−0.27(0.33, A)	−0.37(0.20, A)	0.71(0.34, R)	0.28(0.18, A)	0.22(0.33, A)	0.04(0.28, A)
Slope $\beta_1(s_1$, A or R)	1.09(0.13, R)	1.16(0.07, R)	1.05(0.04, R)	1.06(0.03, R)	0.90(0.05, R)	0.94(0.02, R)	0.97(0.04, R)	0.99(0.04, R)
F_{exp} versus $F_{0.95}$(2-1, 18-2)=4.49	65.58	289.12	569.39	1518.21	402.18	1529.54	472.54	738.56
P versus α=0.05 and H_0: regression model is accepted or rejected	4.73×10^{-7}	2.91×10^{-10}	6.19×10^{-14}	1.73×10^{-16}	9.18×10^{-13}	7.17×10^{-15}	2.64×10^{-13}	1.63×10^{-13}
Correlation								
Determination coefficient, R^2 (%)	80.39	95.70	97.27	99.02	96.17	99.16	96.72	98.14
Predicted determination coefficient, R^2_P [%]	57.10	88.22	93.61	97.49	90.66	97.75	91.78	94.82
Prediction ability criteria								
Mean error of prediction, *MEP*	2.56	0.58	0.26	0.11	0.29	0.07	0.29	0.19
Akaike information criterion, *AIC*	16.99	−9.43	−23.09	−38.31	−22.38	−40.35	−22.83	−28.12
Goodness-of-fit test								
E(Λ)	0.03	0.14	−0.11	−0.02	−0.04	0.18	0.01	0.01
$s(e)$ in log units of pK_a	1.50	0.78	0.51	0.34	0.56	0.29	0.50	0.38
t_{exp} versus $t_{0.95}(n-m)$=2.06	0.08	0.69	−0.90	−0.19	−0.29	2.32	0.10	0.15
P versus α=0.05 and H_0: E(Λ)=0 is accepted or rejected	0.47, A	0.25, A	0.19, A	0.42, A	0.39, A	0.02, R	0.46, A	0.44, A
Outlier detection using the Williams plot								
Number of outliers detected	3	0	1	0	3	0	2	0
Indices of outliers detected	51, 52, 62	–	52	–	50, 52, 62	–	48, 52	–

For intercept and slope estimates, the letters **A** or **R** refer to whether the tested null hypothesis H_0: β_0=0 vs. H_A: $\beta_0\neq0$ and H_0: β_1=1 vs. H_A: $\beta_1\neq1$ was accepted or rejected for the proposed regression model $pK_{a,pred}=\beta_0 + \beta_1\ pK_{a,exp}$. The estimated standard deviation of the parameter is given in parentheses

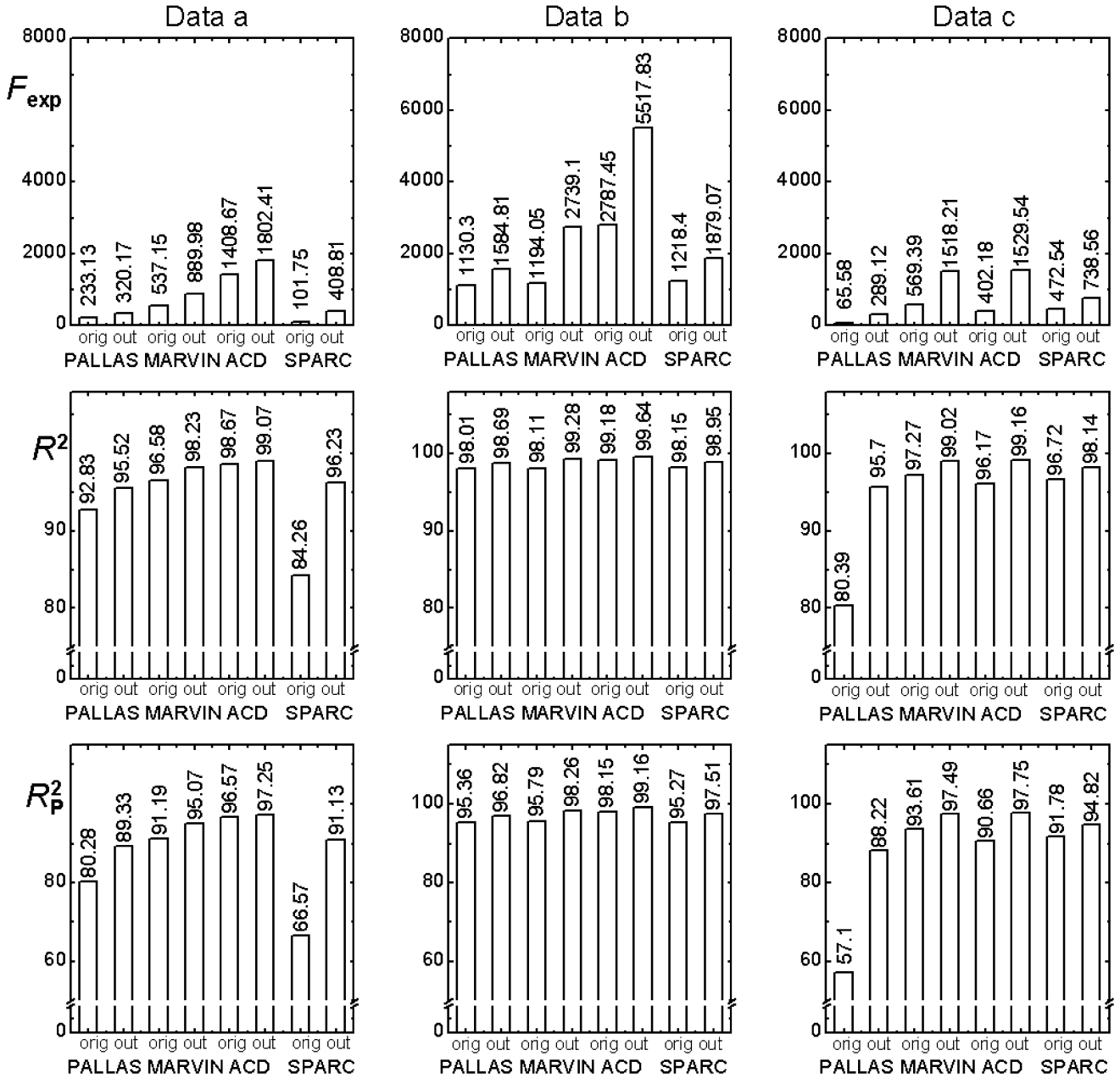

Fig. 4 Resolution abilities of the three regression diagnostic criteria F_{exp}, R^2 and R^2_P in relation to examining the accuracies of pK_a predictions made by the four algorithms PALLAS, MARVIN, ACD/pK_a and SPARC. Here *orig* refers to the original dataset from Table 1, and *out* refers to the dataset without outliers

tests shown in Table 2 show the quality of pK_a prediction: the highest values of R^2 and R^2_P in Fig. 4, the lowest values of *MEP* and *s*(*e*), and the most negative value of *AIC* in Fig. 5 and Table 2 all show that the ACD/pK_a algorithm used for pK_a prediction offers the best predictive power and the most accurate results.

Proposed regression model The predicted vs. experimentally observed pK_a values for the dataset examined are plotted in Fig. 1a–d, while the numerical results are shown in Table 2. The data points are distributed evenly around the diagonal in the figures, implying consistent error behavior for the residual value. The slope and the intercept of the linear regression are optimal; the slope estimates for the four algorithms used are $\beta_1(s_1)$=0.95 (0.06, **R**), 0.90 (0.04, **R**), 1.06 (0.03, **R**), 1.10 (0.11, **R**), where **A** or **R** mean that the tested null hypothesis H_0: β_0=0 vs. H_A: $\beta_0 \neq 0$ and H_0: β_1=1 vs. H_A: $\beta_1 \neq 1$ was **A**ccepted or **R**ejected; the standard deviation of the estimated parameter is given in parentheses. Upon removing the outliers from the dataset, these estimates improve to 0.96 (0.05, **R**), 0.86 (0.03, **R**), 1.04 (0.02, **R**), 1.12 (0.06, **R**). The estimated intercepts are

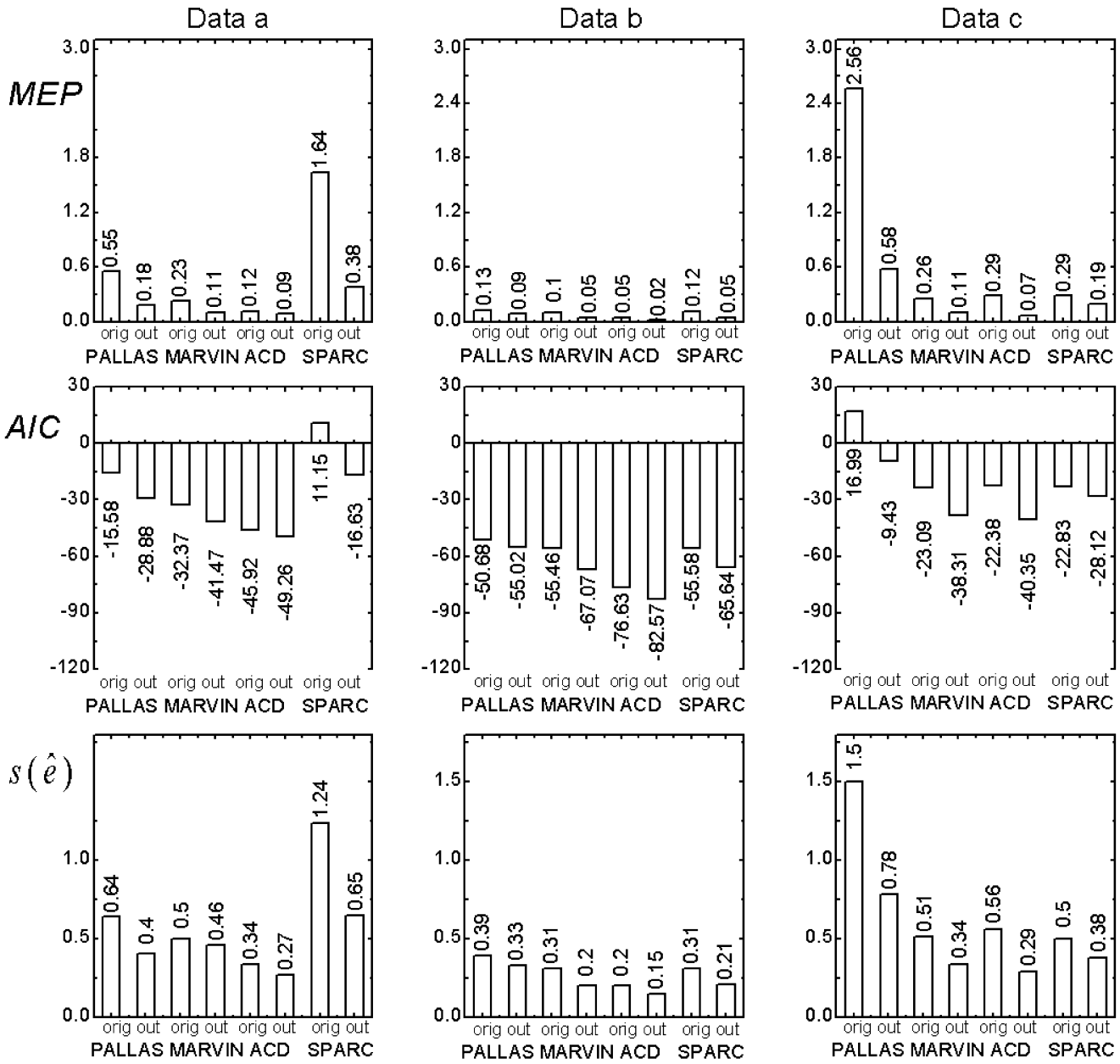

Fig. 5 Resolution abilities of the three regression diagnostic criteria *MEP, AIC* and *s*(*e*) in relation to examining the accuracies of pK_a predictions made by the four algorithms PALLAS, MARVIN, ACD/pKa and SPARC. Here *orig* refers to the original dataset from Table 1, and *out* refers to the dataset without outliers

$\beta_0(s_0)$=0.46 (0.49, **A**), 0.78 (0.32, **R**), −0.50 (0.23, R), −0.93 (0.89, A), and after removing outliers from the dataset they are 0.28(0.45, **A**), 1.07(0.24, R), −0.35(0.20, A), −1.24(0.44, R). Here **A** or **R** mean that the tested null hypothesis H_0: β_0=0 vs. H_A: $\beta_0\neq0$ and H_0: β_1=1 vs. H_A: $\beta_1\neq1$ was **A**ccepted or **R**ejected. The slope is almost equal to one for all four algorithms, and the intercept is almost equal to zero for all four algorithms. The Fisher–Snedecor *F*-test of overall regression for the four prediction algorithms in Fig. 4 yields calculated significance levels of $P=9.58\times10^{-12}$, 2.15×10^{-15}, 2.75×10^{-19}, 4.58×10^{-9}, and after removing the outliers from the dataset these significance levels changed to $P=1.57\times10^{-11}$, 1.87×10^{-15}, 1.08×10^{-18}, 8.09×10^{-13}, meaning that all four algorithms proposed significant regression models. The highest *F*-test value was obtained using the ACD/pK_a algorithm, which therefore gave the best prediction of p$K_{a,pred}$.

Correlation The quality of the regression models yielded by the four algorithms was measured using the two

statistical characteristics of correlation shown in Fig. 4 and Table 2; i.e., R^2=92.83%, 96.58%, 98.67%, 84.26% and R^2 =95.52%, 98.23%, 99.07%, 96.23% before and after removing the outliers from the dataset, respectively, while the predicted determination coefficient R^2_P=80.28%, 91.19%, 96.57%, 66.57% and R^2_P=89.33%, 95.07%, 97.25%, 91.13% before and after removing the outliers from the dataset, respectively. R^2 is high for all four algorithms and this indicates that they are all able to interpolate within the range of pK_a values associated with the examined dataset. The highest values of R^2 and R^2_P are exhibited by the ACD/pK_a algorithm.

Criteria for expressing the prediction ability The goodness-of-fit test criteria that best express the predictive ability are the mean error of prediction *MEP* and the Akaike information criterion *AIC*, as shown in Fig. 5 and Table 2. Calculated *MEP* values were 0.55, 0.23, 0.12, 1.64, but after removing the outliers from the dataset these *MEP* values dropped to 0.18, 0.11, 0.09, 0.38. The Akaike information criterion *AIC* yielded values of −15.58, −32.37, −45.92, 11.15, but after removing the outliers from the dataset they dropped to −28.88, −41.47, −49.26, −16.63. A numerical comparison shows that both *MEP* and *AIC* classify the predictive abilities of the four algorithms well, and are able to distinguish between them. The lowest value of *MEP* was 0.12 and the most negative value of *AIC* was −45.92, both of which were attained for the ACD/pK_a method. These criteria were used to classify the predictive abilities of the regression models, and to rank the four algorithms from best to worst. The regression models were sufficiently predictive, i.e., they were able to extrapolate beyond the range of the training set values.

Goodness-of-fit test The best way to evaluate the four regression models is to examine the fitted residuals. If the proposed model represents the data adequately, the residuals should form a random pattern with a normal distribution N(0, s^2) and the residual mean of zero, $E(\hat{e})$= 0. A Student *t*-test examines the null hypothesis H_0: $E(\hat{e})$= 0 vs. H_A: $E(\hat{e})\neq 0$ and gives the values of the criteria for the four algorithms in the form of the calculated significance levels *P*=0.37, 0.46, 0.19, 0.33. All four algorithms give a residual bias of zero. The estimated standard deviation of the regression straight line $s(e)$ in Fig. 5 and Table 2 is $s(e)$= 0.64, 0.50, 0.34, 1.24 log units pK_a, and after removing the outliers from the dataset they became $s(e)$=0.40, 0.46, 0.27, 0.65 log units pK_a, with the lowest value attained for the ACD/pK_a method. Previously, Hilal et al. [17] used the SPARC calculator to estimate the 4338 pK_a values for some 3685 compounds, including multiple pK_a values up to the sixth pK_a, and the overall standard deviation $s(e)$ for this large test set of compounds was found to be 0.37 pK_a units. For complicated structures where a molecule has multiple ionization sites, such as azo dyes, the expected SPARC error was ±0.65 pK_a units. SPARC was used to estimate 358 pK_a values for 214 azo dyes, and the SPARC standard deviation was found to be 0.63 pK_a units. The reported IUPAC RMS interlaboratory deviation between observed values of pK_a for azo dyes, when more than one measurement was reported, was 0.64. The error in the SPARC-calculated values was comparable to the experimental error and perhaps better for these complicated molecules.

Outlier detection The detection, assessment, and understanding of outlier p$K_{a,pred}$ values are major areas of interest when examining accuracy. If the data contains a single outlier p$K_{a,pred}$, it is relatively simple to identify this p$K_{a,pred}$ value. If the p$K_{a,pred}$ data contain more than one outlier (which is likely to be the case in most data), it becomes more difficult to identify such p$K_{a,pred}$ values, due to masking and swamping effects [20]. *Masking* occurs when an outlying p$K_{a,pred}$ value goes undetected because of the presence of another, usually adjacent, subset of p$K_{a,pred}$ values. *Swamping* occurs when "good" p$K_{a,pred}$ values are incorrectly identified as outliers because of the presence of another, usually remote, subset of p$K_{a,pred}$ values. Statistical tests are needed in order to decide how to use the real data such that the assumptions of the hypothesis tested are approximately satisfied. In the PALLAS straight line model, three outliers (2, 6 and 19) were detected. In the MARVIN straight line model, three outliers (6, 11, and 16) were detected. In the ACD/pK_a straight line model, only two outliers (10 and 13) were indicated, while in the SPARC straight line model, three outliers (2, 3, and 8) were found.

Outlier interpretation and removal The poorest molecular pK_a predictions correspond to outliers. Outliers are molecules which belong to the most poorly characterized class considered, so it is no great surprise that they are also the most poorly predicted. Outliers should therefore be analyzed, elucidated and then removed from the data. In our study, the use of the Williams plot revealed three outliers, nos. 2 (chlorothiazide, pK_1), 6 (diazepam) and 19 (tetracaine), in the PALLAS regression model (Fig. 1e); three outliers, nos. 6 (diazepam), 11 (haloperidol) and 16 (phenytoin), in the MARVIN model (Fig. 1f); two outliers, nos. 10 (furosemide) and 13 (lidocaine), in ACD/pK_a (Fig. 1g); and three outliers, nos. 2 (chlorothiazide, pK_1), 3 (chlorothiazide, pK_2) and 8 (disopyramide), in the SPARC model (Fig. 1h). After removing the outlying values of pK_a for the poorly predicted molecules, all of the remaining data points were statistically significant (Table 2). Outliers frequently turned out to be either

misassignments of pK_a values or suspicious molecular structures. Due to their fragment-based approach, the methods proved to be inadequate when fragments present in the molecule under investigation were absent from the database. Such pK_a prediction methods require that the compounds being studied are very similar to those available in the training set. Suitable corrections were made where possible, but in some cases the corresponding data had to be omitted from the training set. In other cases, the outliers served to highlight the need to split one class of molecules into two or more subclasses based on the substructure in which the acidic or (more often) basic center was embedded.

Analysis of dataset b

The regression data treatment of Dataset b was performed in the same way as for Dataset a, and individual blocks of Table 2 associated with Dataset b were interpreted in a similar way to the blocks associated with Dataset a in Table 2. It is obvious that the line fits are much better in Fig. 2 (for Dataset b) than in Fig. 1 (for Dataset a). While R^2 and R^2_p cannot be so clearly discriminated among the different correlation of variables {pK_a, p$K_{a,pred}$} leading to the straight line, the statistical test criterion F_{exp} exhibits much better resolution power. All three statistics *MEP, AIC* and $s(e)$ in Fig. 5 show good resolution, as differences between the various line-fittings are obviously pronounced. Figures 4 and 5 enable us to classify not only the three datasets but also the performances of the four prediction algorithms. The numerical values of all of the statistics mentioned are given in Table 2. The algorithms indicate some outliers, and so these outliers should be analyzed and removed from the data. The Williams plots revealed two outliers, nos. 39 (pentobarbitone), 44 (celiprolol), in the PALLAS regression model (Fig. 2e); three outliers, nos. 31 (3,4-dichlorophenol), 39 (pentobarbitone) and 42 (pericyazine), in the MARVIN model (Fig. 2f); three outliers, nos. 44 (celiprolol), 31 (3,4-dichlorophenol) and 46 (propranolol), in the ACD/pK_a model (Fig. 2g), and three outliers, nos. 44 (celiprolol), 22 (benzoic acid) and 35 (*N*-methylaniline), in the SPARC model (Fig. 2h). All of these outlying molecules belong to the poorly characterized class in the training set of the algorithm's database.

Analysis of dataset c

The regression data treatment of Dataset c was performed in the same way as for Dataset a, and the individual blocks associated with Dataset c in Table 2 were interpreted in a similar way to those of Dataset a in Table 2. The Williams plots revealed three outliers, nos. 51 (enalapril), 52 (famotidine), and 62 (piroxicam), in the PALLAS regression model (Fig. 3e); one outlier, no. 52 (famotidine), in the MARVIN model (Fig. 3f); three outliers, nos. 52 (famotidine), 62 (piroxicam) and 50 (diltiazem), in the ACD/pK_a model (Fig. 3g); and two outliers, nos. 52 (famotidine) and 48 (captopril), in the SPARC model (Fig. 3h). Most of the poorly predicted molecules, which were outliers in relation to the regression line, were important pharmacologically but were also poorly represented, and were the most poorly characterized classes considered in the algorithm's training set. A criterion that describes the similarity of the molecules under investigation to those in the training database would be very useful.

One may also question, however, whether a failure to make predictions for unusual compounds is a particularly bad thing. When predictions are not obtained for some molecules, this means that the training set does not contain any molecules that are similar to them. This would explain the absence of the required fragments in the training set. However, the authors also note that the diversity and complexity of the molecules used for pK_a model development and testing has dramatically increased in the last few years, which should lead to greater robustness.

Conclusions

Researchers should use rigorous statistical rules and regression prediction models with caution, and should always validate these models with known experimental data (using the REGDIA algorithm for example) before making any critical decisions. The most poorly predicted molecular pK_a values correspond to outliers. The Williams graph is the preferred tool for the reliable detection of outlying pK_a values. Regression diagnostics analysis ensures that outliers in the predicted pK_a dataset are found, and this represents a critical step in explicitly manipulating the degree of ionization in order to improve solubility, permeability, protein binding and blood–brain permeation. The ACD/pK_a program proved to be the most accurate method of predicting pK_a values for the three datasets tested. The proposed accuracy test provided by the REGDIA program can also be extended to other predicted values, such as log *P*, log *D*, aqueous solubility, and some physiochemical properties.

Acknowledgements The financial support of the Czech Ministry of Education (Grant No MSM0021627502) and of the Grant Agency of the Czech Republic (Grant No NR 9055-4/2006) is gratefully acknowledged.

References

1. Xing L, Glen RC (2002) Novel methods for the prediction of logP, pK and logD. J Chem Inf Comput Sci 42:796–805
2. Xing L, Glen RC, Clark RD (2003) Predicting pK_a by molecular tree structured fingerprints and PLS. J Chem Inf Comput Sci 43:870–879
3. Tajkhorshid E, Paizs B, Suhai (1999) Role of isomerization barriers in the pK_a control of the retinal schiff base: a density functional study. J Phys Chem B 103:4518–4527
4. Tripos (2007) SYBYL software. Tripos, Inc., St. Louis, MO (http://www.tripos.com, cited 25 July 2007)
5. ACD/Labs (2007) pK_a Predictor 3.0. Advanced Chemistry Development Inc., Toronto, Canada (http://www.acdlabs.com, cited 25 July 2007)
6. Rekker RF, ter Laak AM, Mannhold R (1993) Prediction by the ACD/pK_a method of values of the acid–base dissociation constant (pK_a) for 22 drugs. Quant Struct–Act Relat 12:152
7. Slater B, McCormack A, Avdeef A, Commer JEA (1994) Comparison of ACD/pK_a with experimental values. Pharm Sci 83:1280–1283
8. ACD/Labs (1997) Results of titrometric measurements on selected drugs compared to ACD/pK_a September 1998 predictions (poster). In: AAPS, 1–6 November 1997, Boston, MA
9. Szegezdi J, Csizmadia F (2004) Marvin plug-in. In: Prediction of dissociation constant using microconstants. http://www.chemaxon.com/conf/Prediction_of_dissociation_constant_using_microconstants.pdf. Cited 25 July 2007
10. Gulyás Z, Pöcze G, Petz A, Darvas F PALLAS cluster—a new solution to accelerate the high-throughput ADME-TOX prediction. CompuDrug Chemistry Ltd., Sedona, AZ (see http://www.compudrug.com, last cited 25 July 2007)
11. Kim KH, Martin YC (1991) Direct prediction of linear free energy substituent effects from 3D structures using comparative molecular field effect. 1: Electronic effect of substituted benzoic acids. J Org Chem 56:2723–2729
12. Kim KH, Martin YC (1991) Direct prediction of dissociation constants of clonidine-like imidazolines, 2-substituted imidazoles, and 1-methyl-2-substituted imidazoles from 3D structures using a comparative molecular field analysis (CoMFA) approach. J Med Chem 34:2056–2060
13. Gargallo R, Sotriffer CA, Liedl KR, Rode BM (1999) Application of multivariate data analysis methods to comparative molecular field analysis (CoMFA) data: proton affinities and pK_a prediction for nucleic acids components. J Comput Aided Mol Des 13:611–623
14. Perrin DD, Dempsey B, Serjeant EP (1981) pK_a prediction for organic acids and bases. Chapman and Hall, London
15. Habibi-Yangjeh A, Danandeh-Jenagharad M, Nooshyar M (2005) Prediction acidity constant of various benzoic acids and phenols in water using linear and nonlinear QSPR models. Bull Korean Chem Soc 26:2007–2016
16. Popelier PLA, Smith PJ (2006) QSAR models based on quantum topological molecular similarity. European J Med Chem 41:862–873
17. Hilal SH, Karickhoff SW, Carreira LA (2003) Prediction of chemical reactivity parameters and physical properties of organic compounds from molecular structure using SPARC (EPA/600/R-03/030 March 2003). National Exposure Research Laboratory, Office of Research and Development, US Environmental Protection Agency, Research Triangle Park, NC
18. Meloun M, Bordovská S, Kupka K (2007) Outliers detection in the statistical accuracy test of a pK_a prediction. Anal Chim Acta (in press)
19. MathSoft (1997) S-PLUS. MathSoft, Seattle, WA (see http://www.insightful.com/products/splus, cited 25 July 2007)
20. Meloun M, Militký J, Forina M (1992–1994) Chemometrics for analytical chemistry, vols 1–2. Ellis Horwood, Chichester, UK
21. ACD/Labs (2007) ACD/pK_a DB vs. experiment: a comparison of predicted and experimental values. http://www.acdlabs.com/products/phys_chem_lab/pka/exp.html. Cited 25 July 2007
22. Lombardo F, Obach RS, Shalaeva MY, Feng G (2004) Prediction of human volume of distribution values for neutral and basic drugs. 2: Extended dataset and leave-class-out statistics. J Med Chem 47:1242–1250
23. Luan F, Ma W, Zhang H, Zhang X, Liu M, Hu Z, Fan B (2005) Prediction of pK_a for neutral and basic drugs based on radial basis function neutral networks and the heuristic method. Pharm Research 22:1454–1460
24. Masuda T, Jikihara T, Nakamura K, Kimura A, Takagi T, Fujiwara H (1997) Introduction of solvent-accessible surface area in the calculation of the hydrophobicity parameter log *P* from an atomistic approach. J Pharm Sciences 86:57–63
25. Moriguchi I, Hirono S, Nakagome I, Hirano H (1994) Comparison of reliability of log *P* values for drugs calculated by several methods. Chem Pharm Bull 42:976–978
26. Leo AJ (1995) Critique of recent comparison of log *P* calculation methods. Chem Pharm Bull 43:512–513
27. Suzuki T, Kudo Y (1990) Automatic log *P* estimation based on combined additive modeling methods. J Comput Aided Mol Design 4:155–198
28. Kolovanov EA, Petrauskas AA (2007) Comparison of the accuracy of log *P* and log *D* calculations for 22 drugs. http://www.acdlabs.com/publish/acc_logp.html. Cited 25 July 2007
29. Kolovanov EA, Petrauskas AA (2007) Re-evaluation of log *P* data for 22 drugs and comparison of six calculation methods. http://www.acdlabs.com/publish/ac_logp.html. Cited 25 July 2007
30. Hansen NT, Kouskoumvekaki I, Jorgensen FS, Brunak S, Jonsdottir SO (2006) Prediction of pH-dependent aqueous solubility of druglike molecules. J Chem Inf Model 46:2601–2609
31. Engkvist O, Wrede P (2002) High-throughput, in silico prediction of aqueous solubility based on one- and two-dimensional descriptors. J Chem Inf Comput Sci 42:1247–1249