J Math Chem (2010) 47:891–909
DOI 10.1007/s10910-009-9609-2

ORIGINAL PAPER

# Outliers detection in the statistical accuracy test of a p$K_a$ prediction

**Milan Meloun · Sylva Bordovská · Karel Kupka**

Received: 24 November 2008 / Accepted: 8 September 2009 / Published online: 19 September 2009

**Abstract** The regression diagnostics algorithm REGDIA in S-Plus is introduced to examine the accuracy of p$K_a$ predicted with four programs: PALLAS, MARVIN, PERRIN and SYBYL. On basis of a statistical analysis of residuals, outlier diagnostics are proposed. Residual analysis of the ADSTAT program is based on examining goodness-of-fit via graphical diagnostics of 15 exploratory data analysis plots, such as bar plots, box-and-whisker plots, dot plots, midsum plots, symmetry plots, kurtosis plots, differential quantile plots, quantile-box plots, frequency polygons, histograms, quantile plots, quantile-quantile plots, rankit plots, scatter plots, and autocorrelation plots. Outliers in p$K_a$ relate to molecules which are poorly characterized by the considered p$K_a$ program. Of the seven most efficient diagnostic plots (the Williams graph, Graph of predicted residuals, Pregibon graph, Gray L–R graph, Index graph of Atkinson measure, Index graph of diagonal elements of the hat matrix and Rankit Q–Q graph of jackknife residuals) the Williams graph was selected to give the most reliable detection of outliers. The six statistical characteristics, $F_{\mathrm{exp}}$, $R^2$, $R_{\mathrm{P}}^2$, *MEP*, *AIC*, and $s$ in p$K_a$ units, successfully examine the specimen of 25 acids and bases of a Perrin's data set classifying four p$K_a$ prediction algorithms. The highest values $F_{\mathrm{exp}}$, $R^2$, $R_{\mathrm{P}}^2$ and the lowest value of *MEP* and $s$ and the most negative *AIC* have been found for PERRIN algorithm of p$K_a$ prediction so this algorithm achieves the best predictive power and the most accurate results. The proposed accuracy test of the REGDIA program can also be extended to test other predicted values, as log $P$, log $D$, aqueous solubility or some physicochemical properties.

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

K. Kupka
TriloByte Statistical Software, s.r.o., Jiráskova 21, 530 02 Pardubice, Czech Republic

**Keywords** p$K_a$ prediction · Dissociation constants · Outliers · Residuals · Goodness-of-fit · Williams graph

## 1 Introduction

The principle of the *structure–property relationship* is a basic concept in organic chemistry, as the properties of molecules are intrinsically determined by their structure. The macroscopic properties of chemical compounds clearly depend on their microscopic structural descriptors, and the development of a **Q**uantitative **S**tructure/**P**roperty **R**elationship QSPR on theoretical descriptors is a powerful tool for the prediction of the chemical, physical and biological properties of compounds. An enormous number of structural descriptors have been used by researchers to increase the ability to correlate various properties. A molecule is transformed into a sequence or a fixed-length vector of values before it can be used to conduct QSPR studies. Although molecular size may vary to a large extent, the vector representation must be in a fixed length for all the molecules in a data set in order to apply a data analysis method. Various approaches have been developed to represent the structure of molecules for QSPR studies. Since so many descriptors are available, the development and selection of appropriate descriptors in describing a selected property of molecule has become a Herculean task. An important role of the degree of ionization in the biological behaviour of chemical substances, namely drugs, is well established. One of the fundamental properties of an organic drug molecule, the p$K_a$ value, determines the degree of dissociation in solution [1–12]. To obtain a significant correlation and accurately predicted p$K_a$, it is crucial that appropriate structural descriptors be employed. In this context, the approach using a statistical accuracy examination of the predicted p$K_a$ is important.

Numerous studies have considered, and various approaches have been used in the prediction of p$K_a$, but mostly without a rigorous statistical test of p$K_a$ accuracy [13–35].

The goal of this paper is to develop a rigorous accuracy examination tool which is able to investigate weather a p$K_a$ prediction method leads to a sufficiently accurate estimate of p$K_a$ value, as the correlation between predicted p$K_{a,pred}$ and experimental value p$K_{a,exp}$ is usually very high. In this examination, the linear regression models are used for interpreting the essential features of a set of p$K_{a,pred}$ data. There are a number of common difficulties associated with real datasets. The first involves the detection and elucidation of outlying p$K_{a,pred}$ values in the predicted p$K_a$ data. A problem with p$K_{a,pred}$ outliers is that they can strongly influence the regression model, especially when using least squares criteria, so several steps are required: firstly to identify whether there are any p$K_{a,pred}$ values that are atypical of the dataset, then to remove them, and finally to interpret their deviation from the straight line regression model.

Because every prediction is based on a congeneric parent structure, p$K_a$ values can only be reliably predicted for compounds very similar to those in the training set, making it difficult or impossible to get good estimates for novel structures. A further disadvantage is the need to derive a very large number of fragment constants and correlation factors, a process which is complicated and potentially ambiguous. Although this is probably the most widely used method, the accuracy and extensibility of the predictions obtained have not been gratifying.

Authors usually evaluate model quality and outliers on the basis of fitted residues. A simple criticism criterion like, for example, "more than 80% of p$K_a$ in the training sample are predicted with an accuracy of within one log unit of their measurement, and 95% are within two log units of the accuracy," is often used, or "when the difference between the measured p$K_{a,exp}$ and predicted p$K_{a,pred}$ values is larger than 3 log units, it is used to denote a discrepancy". However, rigorous statistical detection, assessment, and understanding of the outliers in p$K_{a,pred}$ values are major problems of interests in an accuracy examination. The goal of any p$K_{a,pred}$ outlier detection is to find this true partition and, thus, separate good from outlying p$K_a$ values. A single case approach to the detection of outliers can, however, fail because of masking or swamping effects, in which outliers go undetected because of the presence of other, usually adjacent, p$K_a'$s. Masking occurs when the data contain outliers which we fail to detect; this can happen because some of the outliers are hidden by other outliers in the data. Swamping occurs when we wrongly declare some of the non-outlying points to be outliers [36]; this occurs because outliers tend to pull the regression equation toward them, thereby making other points further from the fitted equation. Masking is therefore a false negative decision, whereas swamping is a false positive. Unfortunately, a bewilderingly large number of statistical tests, diagnostic graphs and residual plots have been proposed for diagnosting influential points, namely outliers, and it is time to select those approaches that are appropriate for p$K_a$ prediction. This paper provides a critical survey of many outlier diagnostics, illustrated with data examples to show how they successfully characterize the joint influence of a group of cases and to yield a better understanding of joint influence.

## 2 Methods

### 2.1 Software and data used

Several software packages for p$K_a$ prediction were used and tested in this study. Most of the work was carried out on PALLAS [10], MARVIN [15], Perrin method [29] and SYBYL [14] software packages, based mostly on chemical structure, the reliability of which reflects the accuracy of the underlying experimental data. In most software the input is the chemical structure drawn in a graphical mode. For the accuracy examination of p$K_{a,pred}$ values we used Perrin's examples of different chemical classes as an external data set [29]. The model predicted p$K_a$ values were compared to Perrin's predictions, and the experimental measurements are listed in Table 1.

For the creation of regression diagnostic graphs and computation of the regression based characteristics, the REGDIA algorithm was written in *S-Plus* [37], and the Linear Regression module of our ADSTAT package [38] was used. We have tried to show some effective drawbacks of the statistical diagnostic tools in REGDIA which are, in our experience, able to correctly pinpoint influential points. One should concentrate on that diagnostic tool which measures the impact on the quantity of primary interest. The main difference between the use of regression diagnostics and classical

**Table 1** Perrin's p$K_a$ values data set of 25 organic molecules: (a) $i$ = 1–11 (p$K_a$ values of substituted aliphatic acids and bases), (b) $i$ = 12–18 (p$K_a$ values for phenols, aromatic carboxylic acids and aromatic amines), (c) $i$ = 19–21 (p$K_a$ values of heteroatomic acids and bases), (d) $i$ = 22–25 (p$K_a$ values of heterocycles) [29]

| $i$ | Name | p$K_{exp}$ | p$K_{pred}$ (Pallas) | p$K_{pred}$ (Marvin) | p$K_{pred}$ (Perrin) | p$K_{pred}$ (Sybyl) |
|---|---|---|---|---|---|---|
| 1 | Bis(2-Chloroethyl)(2-Methoxyethyl)Amine | 5.45 | 6.26 | 5.9 | 5.1 | 6.91 |
| 2 | 1-(4'-Hydroxycyclohexyl)-2-(Isopropylamino)Ethanol | 10.23 | **11.23** | 10.1 | 9.99 | 10.03 |
| 3 | 2-Aminocycloheptanol | 9.25 | 9.77 | 9.98 | **9.67** | 9.84 |
| 4 | N,N-Dimethyl-2-Butyn-1-Amine | 8.28 | 7.84 | **7.16** | 8.1 | **10.17** |
| 5 | 5-Chloro-3-Methyl-3-Aza-pentanol | 7.48 | 7.48 | 7.9 | 7.1 | **9.52** |
| 6 | 2-Acetylbutanedioic Acid | 2.86 | 2.89 | 3.66 | 3.15 | 2.35 |
| 7 | 2-(Methylamino)Acetamide | 8.31 | **4.93** | 8.81 | 8.43 | 8.11 |
| 8 | 2-(Dimethylamino)Ethyl Acetate | 8.35 | 8.72 | 8.42 | 8.26 | 8.6 |
| 9 | 2,3-Dihydroxy-2-Hydroxymethylpropanoic Acid | 3.29 | 3.28 | 3.32 | 3.01 | 3.85 |
| 10 | 1,8-Diamino-3,6-Dithiaoctane | 9.47 | 9.54 | 9.41 | 9.06 | 9.26 |
| 11 | 4-Morpholino-2,2-Diphenyl-pentanenitrile | 6.05 | 7.07 | 6.96 | 6.38 | 7.45 |
| 12 | Benzenehexol | 9.0 | 8.32 | 9.50 | 8.31 | 9.28 |
| 13 | Picric Acid | 0.33 | 0.91 | 1.35 | 0.91 | 1.18 |
| 14 | 2,6-Dichloro-1,4-Benzenediol | 7.3 | 6.82 | 6.99 | 6.82 | 7.6 |
| 15 | 4-Bromo-1,2-Benzenedicarboxylic Acid | 2.5 | 2.86 | 2.84 | 2.86 | 3.26 |
| 16 | 4-Hydroxy-3,5-Dimethoxy-benzoic Acid | 4.34 | 4.36 | 3.93 | 4.36 | 4.54 |
| 17 | 3-Iodo-4-Methylthioaniline | 3.44 | 3.34 | 3.85 | 3.34 | 3.29 |
| 18 | 4-Bromo-3-Nitroaniline | 1.8 | 1.82 | 1.68 | 1.82 | 1.78 |
| 19 | 3-Bromo-5-Methoxypyridine | 2.6 | 2.3 | 2.49 | 2.3 | 3.19 |
| 20 | 4-Aminopyridazine | 6.65 | **4.45** | 6.46 | **5.31** | 6.76 |
| 21 | 4-Amino-6-Chloropyrimidine | 2.1 | 1.99 | 3.19 | **1.41** | 1.68 |
| 22 | 4-Nitrothiophen-2-Carboxylic Acid | 2.68 | 2.67 | 3.26 | 2.7 | 2.58 |
| 23 | 4-Bromopyrrol-2-Carboxylic Acid | 4.06 | 2.93 | 3.6 | 4.05 | 4.22 |
| 24 | Furan-2,4-Dicarboxylic Acid | 2.63 | 3.13 | 3.06 | 2.77 | 2.22 |
| 25 | Pyrazole-3-Carboxylic Acid | 3.74 | 3.98 | 3.18 | 3.98 | 3.77 |

REGDIA indicated outliers are in bold

statistical tests in REGDIA is that there is no necessity for an alternative hypothesis, as all kinds of deviations from the ideal state are discovered. Seven diagnostic plots (the Graph of predicted residuals [36], Williams graph [36,39], Pregibon graph [36], Gray L–R graph [36,39–41], Scatter plot of classical residuals versus prediction

[36,39–41], Index graph of jackknife residuals [36], and Index graph of Atkinson distance [41]) were selected as the most efficient to give reliable influential point detection results and four being powerful enough to separate influential points into outliers and high-leverages.

### 2.2 Regression diagnostics for examining the p$K_a$ accuracy in REGDIA

The examination of p$K_a$ data quality involves detection of the *influential points* in the regression model proposed $\mathrm{p}K_{\mathrm{a,pred}} = \beta_0 + \beta_1 \mathrm{p}K_{\mathrm{a,exp}}$, which cause many problems in regression analysis by shifting the parameter estimates or increasing the variance of the parameters [36]: (i) $\mathrm{p}K_{\mathrm{a,pred}}$-*outliers*, which differ from the other points in their value on the $y$-axis, where $y$ stands in all of the following relations for $\mathrm{p}K_{\mathrm{a,pred}}$; (ii) *high-leverage points*, which differ from the other points in their value on the $x$-axis, where $x$ stands in all of the following relations for $\mathrm{p}K_{\mathrm{a,exp}}$, or (iii) both outliers and high-leverages, standing for a combination of both together. Analysis of various types of residuals in the REGDIA program is useful for detecting inadequacies in the model, or influential points in the data [36]:

(a) *Ordinary residuals* $\hat{e}_i$ are defined by $\hat{e}_i = y_i - x_i\beta$, where $x_i$ is the $i$th row of matrix $\mathbf{p}\boldsymbol{K}_{\mathrm{a,exp}}$.
(b) *Normalized residuals* $\hat{e}_{N,i} = \hat{e}_i/s(\hat{e})$ are often recommended for outlier detection.
(c) *Standardized residuals* $\hat{e}_{S,i} = \hat{e}_i/(s(\hat{e})\sqrt{1-\hat{h}_{ii}}$ exhibit constant unit variance, and their statistical properties are the same as those of ordinary residuals.
(d) *Jackknife residuals* $\hat{e}_{J,i} = \hat{e}_{S,i}\sqrt{\frac{n-m-1}{n-m-\hat{e}^2_{S,i}}}$ are residuals, where $n$ stands for the number of points and $m$ for the number of parameters, here $m - 2$ and for which a rule is valid: strongly influential points have squared jackknife residuals $\hat{e}^2_{J,i}$ greater than 10. The descriptive statistics of residuals can be used for a numerical goodness-of-fit evaluation in REGDIA program, *cf.* page 290 in Vol. 2 of [36]:

(1) The *residual bias* is the arithmetic mean of residuals $E(\hat{e})$ and should be equal to zero.
(2) The square-root of the residuals variance $s^2(\hat{e}) = RSS(\boldsymbol{b})/(n-m)$ is used to estimate of the *residual standard deviation*, $s(\hat{e})$, where $RSS(\boldsymbol{b})$ is the residual square-sum, should be of the same magnitude as the random error $s(\mathrm{p}K_{\mathrm{a,pred}})$ as it is valid that $s(\hat{e}) \approx s(\mathrm{p}K_{\mathrm{a,pred}})$.
(3) The *determination coefficient D* calculated from the *correlation coefficient R* and multiplied by 100% is interpreted as the percentage of points which correspond to proposed regression model.
(4) One of the most efficient criterion is the *mean quadratic error of prediction*

$$MEP = \frac{\sum_{i=1}^{n}(y_i - x_i^{\mathrm{T}} b_{(i)})^2}{n}, \tag{1}$$

where $\boldsymbol{b}_{(i)}$ is the estimate of regression parameters when all points except the $i$th were used and $\boldsymbol{x}_i$ (here p$K_{\mathrm{a,exp,i}}$) is the $i$th row of matrix $\mathbf{p}\boldsymbol{K}_{\mathrm{a,exp}}$. The statistic *MEP* uses a prediction $\hat{y}_{P,i}$ (here p$K_{\mathrm{a,pred,i}}$) from an estimate constructed without including the $i$th point.

(5) The *MEP* (Eq. 1) can be used to express the predicted determination coefficient,

$$\hat{R}_P^2 = 1 - \frac{n \times MEP}{\sum_{i=1}^{n} y_i^2 - n \times \bar{y}^2}. \tag{2}$$

(6) Another statistical characteristic is derived from information theory and entropy, and is known as the *Akaike information criterion*,

$$AIC = n \ln \left( \frac{RSS(b)}{n} \right) + 2m, \tag{3}$$

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 minimal value of *MEP* and *AIC* and the highest value of the $R_P^2$ are reached.

Individual estimates $\mathbf{b}$ of parameters $\boldsymbol{\beta}$ are then tested for statistical significance using the Student $t$-test. The *Fisher-Snedecor F-test of significance of the regression model proposed* is based on the testing criterion

$$F_R = \hat{R}^2(n-m)/\left[(1-\hat{R}^2)(m-1)\right] \tag{4}$$

which has a Fisher-Snedecor distribution with $(m-1)$ and $(n-m)$ degrees of freedom, where $R^2$ is the determination coefficient. With the use of $F_R$ the null hypothesis $\mathrm{H}_0 : R^2 = 0$ may be tested and concerns a test of significance of all regression parameters $\boldsymbol{\beta}$.

Examination of data and model quality can be considered directly from the scatter plot of p$K_{\mathrm{a,pred}}$ vs. p$K_{\mathrm{a,exp}}$. For the analysis of residuals a variety of plots have been widely used in regression diagnostics of REGDIA program:

(a) the *overall index plot of classical residuals* gives an initial impression of the residuals trend in chronological order. If the straight line model is correct, the residuals $\mathbf{e}$ form a random pattern and should resemble values from a normal distribution with zero mean. To examine the normality of a residual distribution, the quantile-quantile (rankit) plot may be applied;

(b) the *graph of predicted residuals* indicates outliers as points located on the line $x = y$, i.e. here p$K_{\mathrm{a,pred}}$ = p$K_{\mathrm{a,exp}}$ but far from its central pattern;

(c) the *Williams graph* has two boundary lines, the first for outliers, $y = t_{0.95}(n - m - 1)$ and the second for 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;

(d) the *Pregibon graph* classifies two levels of influential points: strongly influential points are above the upper line, while medium influential points are located between the two lines;

(e) *Gray's L–R graph* indicates outliers as points situated close above the corner of the triangle;
(f) the *scatter plot of classical residuals* indicates only suspicious points which could be proven as outliers using other diagnostics;
(g) the *index graph of jackknife residuals* indicates outliers according to an empiric criterion which states: "strongly influential outliers reach a jackknife residual greater than 3";
(h) the *scatter plot of the Atkinson distance d* leads to numerically similar values as the jackknife residuals, and therefore its interpretation is similar.

### 2.3 Graphs for the exploratory analysis of residuals in ADSTAT [36,38]

Residual analysis is based on examining goodness-of-fit via graphical and/or numerical diagnostics in order to check the data and model quality. A variety of exploratory data analysis plots, such as *bar plots, box-and-whisker plots, dot plots, midsum plots, symmetry plots, kurtosis plots, differential quantile plots, quantile-box plots, frequency polygon, histogram, quantile plots, quantile-quantile plots, rankit plots, scatter plots*, and *autocorrelation plots*, have been introduced in the ADSTAT program [36,38], and widely used by authors such as Belsey, Kuh and Welsch [39], Cook and Weisberg [40], Atkinson [41], Chatterjee and Hadi [42], Barnett and Lewis [43], Welsch [44], Weisberg [45], Rousseeuw and Leroy [46]; others may be found Vol. 2 page 289 of [36]. The following plots are quite important as they give an initial impression of the $\mathrm{p}K_{\mathrm{a,pred}}$ residuals by using computer graphics [36,38]:

(a) The *autocorrelation scatter plot,* being an overall index plot of residuals checks whether there is evidence of any trend in a $\mathrm{p}K_{\mathrm{a,pred}}$ series. The ideal plot shows a horizontal band of points with constant vertical scatter from left to right and indicates the suspicious points that could be influential.
(b) The *quantile-box plot* for symmetrical distributions has a sigmoid shape, while for asymmetrical is convex or concave increasing. A symmetric unimodal distribution contains individual boxes arranged symmetrically inside one another, and the value of relative skewness is close to zero. Outliers are indicated by a sudden increase of the quantile function outside the quartile $F$ box.
(c) The *dot diagram and jittered-dot diagram* represent a univariate projection of a quantile plot, and give a clear view of the local concentration of points.
(d) The *notched box-and-whisker plot* permits determination of an interval estimate of the median, illustrates the spread and skewness of the sample data, shows the symmetry and length of the tails of distribution and aids the identification of outliers.
(e) The *symmetry plot* gives information about the symmetry of the distribution. For a fully symmetrical distribution, it forms a horizontal line $y = M$ (median).
(f) The *quantile-quantile* (*rankit*)*plot* has on the $x$-axis the quantile of the standardized normal distribution $u_{Pi}$ for $P_i = i/(n+1)$, and on the $y$-axis, has the ordered residuals $e_{(i)}$. Data points lying along a straight line indicate distributions of similar shape. This plot enables classification of a sample distribution according to its skewness, kurtosis and tail length. A convex or concave shape

indicates a skewed sample distribution. A sigmoidal shape indicates that the tail lengths of the sample distribution differ from those of a normal one.

(g) The *kernel estimation of probability density plot* and the *histogram* detect an actual sample distribution.

## 3 Experimental

### 3.1 Procedure of accuracy examination

The procedure for the examination of influential points in the data, and the construction of a linear regression model with the use of REGDIA and ADSTAT programs, consists of the following steps:

**Step 1** *Graphs for the exploratory indication of outliers*. This step carries out the goodness-of-fit test by a statistical examination of classical residuals in ADSTAT [36,38] for the identification of suspicious points (S) or outliers (O): the overall index plot of residuals trend, the quantile plot, the dot diagram and jittered-dot diagram, the notched box-and-whisker plot, the symmetry plot, the quantile-quantile rankit plot, the histogram and the Kernel estimation of the probability density function also prove a symmetry of sample distribution. Sample data lead to descriptive statistics, as they are the residual mean and the standard deviation of residuals. The Student $t$-test tests a null hypothesis of zero mean of the residuals bias, $\mathrm{H_0} : \mathrm{E}(\hat{e}) = 0$ vs. $\mathrm{H_A} :\ \mathrm{E}(\hat{e}) \neq 0$.

**Step 2** *Preliminary indication of suspicious influential points.* This step discovers suspicious points only. The index graph of classical residuals and the rankit plot also indicate outliers. Beside descriptive statistics $\mathrm{E}(\hat{e})$ and $s$ the Student $t$-test for a null hypothesis a null hypothesis $\mathrm{H_0} :\ \mathrm{E}(\hat{e}) = 0$ vs. $\mathrm{H_A} :\ \mathrm{E}(\hat{e}) \neq 0$ and $\alpha = 0.05$ is also examined.

**Step 3** *Regression diagnostics to detect suspicious points or outliers in REGDIA program:* The least squares straight-line fitting of the regression model proposed $\mathrm{p}K_{\mathrm{a,pred}} = \beta_0 + \beta_1 \mathrm{p}K_{\mathrm{a,exp}}$, with a 95% confidence interval, and regression diagnostics for the identification of outlying $\mathrm{p}K_{\mathrm{a,pred}}$ values detect suspicious points (S) or outliers (O) using the graph of predicted residuals indicates, the Williams graph, the Pregibon graph, the L–R graph indicates, the scatter plot of classical residuals *vs* prediction, the index graph of jackknife residuals and the index plot of the Atkinson distance.

**Step 4** *Interpretation of outliers.* The statistical significance of both parameters $\beta_0$ and $\beta_1$ of the straight-line regression model $\mathrm{p}K_{\mathrm{a,pred}} = \beta_0(s_0, \mathbf{A} \text{ or } \mathbf{R}) + \beta_1(s_1, \mathbf{A} \text{ or } \mathbf{R})\ \mathrm{p}K_{\mathrm{a,exp}}$ is tested in REGDIA program using the Student $t$-test, where **A** or **R** means that the tested null hypothesis $\mathrm{H_0} : \beta_0 = 0$ vs. $\mathrm{H_A} : \beta_0 \neq 0$ and $\mathrm{H_0} : \beta_1 = 1$ vs. $\mathrm{H_A} : \beta_1 \neq 1$ was either **A**ccepted or **R**ejected. The standard deviations $s_0$ and $s_1$ of the actual parameters $\beta_0$ and $\beta_1$ are estimated. A statistical test of total regression is performed using a Fisher-Snedecor $F$-test and the calculated significance level $P$ is enumerated. Outliers are indicated with the preferred Williams graph. The correlation coefficient $R$, the determination coefficient $R^2$ giving the regression rabat $D$

are computed. The mean quadratic error of prediction *MEP*, the Akaike information criterion *AIC* and the predictive coefficient of determination $R_{\mathrm{P}}^2$ as a percentage are calculated to examine the quality of the model. According to the test for the fulfilment of the conditions for the least-squares method, and the results of regression diagnostics, a more accurate regression model without outliers is constructed and statistical characteristics examined. Outliers should be elucidated.

### 3.2 Supporting information available

The complete computational procedures of the REGDIA program, input data specimens and corresponding output in numerical and graphical form are available free of charge via the internet at http://meloun.upce.cz in the blocks *DATA* and *ALGORITHMS*.

## 4 Results and discussion

The results of the p$K_{\mathrm{a}}$ prediction with the use of the four algorithms PALLAS [10], MARVIN [15], PERRIN [29] and SYBYL [13, 14] are compared, with the predicted values of the dissociation constants p$K_{\mathrm{a,pred}}$ are plotted against the experimental values p$K_{\mathrm{a,exp}}$ for the compounds of Perrin's data set from Table 1. Even given that PALLAS's performance might be somewhat less accurate for druglike compounds, there is overall a good agreement between the predicted p$K_{\mathrm{a,pred}}$ and experimental values p$K_{\mathrm{a,exp}}$.

### 4.1 Evaluating diagnostics in outlier detection

Regression analysis and the discovery of influential points in the p$K_{\mathrm{a,pred}}$ values of data have been investigated extensively using the REGDIA program. Perrin's literature data in Table 1 represent a useful medium for the comparison of results and demonstrating the efficiency of diagnostic tools for outliers detection. The majority of multiple outliers are better indicated by diagnostic plots than by statistical tests of the diagnostic values in the table. These data have been much analyzed as a test for outlier methods. The PALLAS-predicted p$K_{\mathrm{a,pred}}$ *vs* experimentally observed p$K_{\mathrm{a,exp}}$ values for the examined set for bases and acids are plotted in Fig. 1a. The p$K_{\mathrm{a,pred}}$ values are distributed evenly around the diagonal, implying consistent error behaviour in the residual values. The optimal slope $\beta_1$ and intercept $\beta_0$ of the linear regression model p$K_{\mathrm{a,pred}} = \beta_0 + \beta_1$p$K_{\mathrm{a,exp}}$ for $\beta_0 = 0.17(0.41)$ and $\beta_1 = 0.94(0.07)$ can be understood as 0 and 1, respectively, where the standard deviation of parameters appears in brackets.

Another way to evaluate a quality of the regression model proposed with the use of the PALLAS program is to examine its goodness-of-fit. Most of the acids and base in the examined sample are predicted with an accuracy of better than one log of their measurement. Detecting influential points, two figures, each of 8 diagnostics, were analyzed: (i) diagnostic plots based on exploratory data analysis (Fig. 2a–h), and (ii)

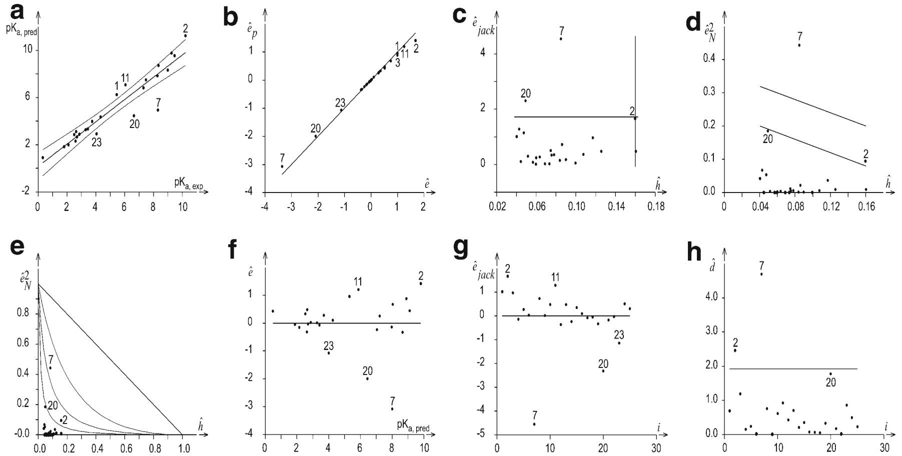

**Fig. 1** Diagnostics graphs for the identification of outlying p$K_{\mathrm{a,pred}}$ values detecting points which are suspicious values (S) or outliers (O) for the data of Table 1 with the PALLAS program: **a** The least squares straight-line fitting of the regression model p$K_{\mathrm{a,pred}} = 0.17(0.41) + 0.94(0.07)$ p$K_{\mathrm{a,exp}}$, with the 95% confidence interval and all the p$K_{\mathrm{a}}$ data indicates S: 1, 2, 7, 11, 20, 23. **b** The graph of predicted residuals indicates O: 1, 2, 3, 7, 11, 20, 23, **c** The Williams graph indicates O: 2, 7, 20, **d** The Pregibon graph indicates influential points (outliers and leverages): 2, 7, 20, **e** The L–R graph indicates O: 2, 7, 20, **f** The scatter plot of classical residuals vs prediction indicates S: 2, 7, 11, 20, 23, **g** The index graph of jackknife residuals indicates S: 2, 11, 20, 23 and O: 7, **h** The index plot of the Atkinson distance indicates S: 20 and O: 2, 7

diagnostic graphs based on the residuals and hat matrix elements (Fig. 1b–f) or vector and scalar influence measures (Fig. 1h) show that the five diagnostic plots (Fig. 1b–f, h) and the Q–Q graph of the jackknife residuals (Fig. 1g) indicate outlying points which obviously differ from the others. The statistical test criteria in the diagnostic plots of Fig. 1 were used to separate influential points into outliers and high-leverages.

The overall index plot of residuals in Fig. 2a indicates no trend in the residuals, as it shows a horizontal band of points with a constant vertical scatter from left to right, and found suspicious points 2, 7, 11, 12, 20 and 23. The quantile-box plot (Fig. 2b) proves the asymmetry of the sample distribution and detects suspicious points 2, 7, 11 and 20, with a sudden increase of the quantile function outside the quartile $F$ box. The dot and jittered-dot diagram (Fig. 2c) indicates suspicious points 7, 20 and 23. The notched box-and-whisker plot (Fig. 2d) illustrates the spread and skewness of the p$K_{\mathrm{a}}$ data, shows the symmetry and length of the tails of the distribution, and aids in the identification of outliers 7 and 20 while the symmetry plot (Fig. 2e) and the quantile-quantile or rankit plot (Fig. 2f) found suspiesious points 2, 7, 11 and 20. Both the histogram (Fig. 2g) and the Kernel estimation of the probability density function (Fig. 2h) prove asymmetry of the sample distribution. To test influential points which were still only suspicious, five diagnostic graphs for outlier detection were applied: the graph of predicted residuals (Fig. 1b) detects outliers 1, 2, 3, 7, 11, 20 and 23 while the most efficient tool, the Williams graph (Fig. 1c) indicates outliers 2, 7 and 20. The Pregibon graph (Fig. 1d) exhibits influential points (outliers and leverages together) 2, 7 and 20. The L–R graph (Fig. 1e) finds outliers 2, 7 and 20. The scatter plot of classical residuals *vs* prediction p$K_{\mathrm{a,pred}}$ (Fig. 1f) indicates suspicious points 2, 7, 11, 20

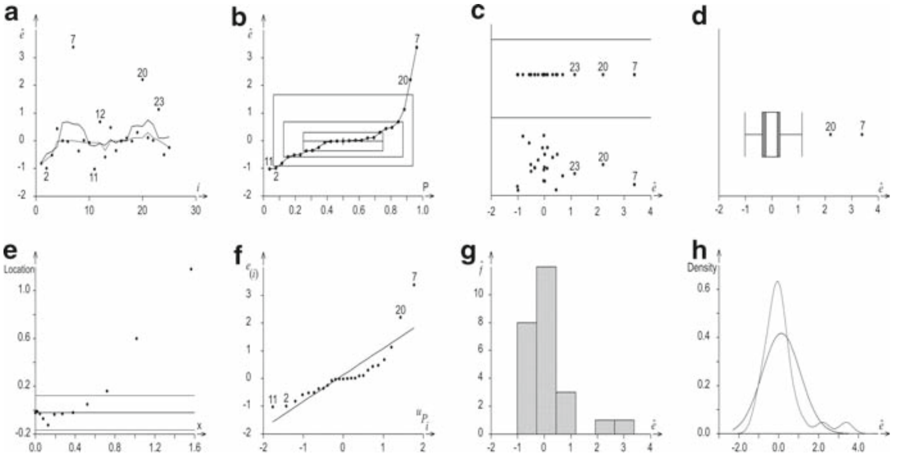

**Fig. 2** Exploratory data analysis graphs carrying out a statistical examination of classical residuals for the identification of outlying p$K_{\mathrm{a,pred}}$ values and detecting points which are suspicious values (S) or outliers (O) for the data of Table 1 with the PALLAS program. The data leads to the descriptive statistics: $n = 25, \alpha = 0.05$, the point estimate of the residual bias $\mathrm{E}(\hat{e}) = 0.13$, and for an interval estimate the lower limit $L_L = -0.26$ and the upper limity $L_U = 0.53$, the standard deviation $s = 0.96$. A Student $t$-test examines a null hypothesis $\mathrm{H}_0 : \mathrm{E}(\hat{e}) = 0$ vs. $\mathrm{H}_\mathrm{A} : \mathrm{E}(\hat{e}) \neq 0$ and leads to the statistical criterion $t_{\mathrm{exp}} = 0.69 < t_{\mathrm{crit}} = 2.06$ with a calculated significance level of $P = 0.23$ meaning that $\mathrm{H}_0$ is accepted: **a** The overall index plot of the residuals trend indicates no trend and S: 2, 7, 11, 12, 20, 23. **b** The quantile-box plot indicates S: 2, 7, 11, 20. **c** The dot and jittered-dot diagram indicates S: 7, 20, 23. **d** The notched box-and-whisker plot indicates O: 7, 20. **e** The symmetry plot, **f** The quantile-quantile (rankit) plot indicates S: 2, 7, 11, 20 **g** The histogram and **h** the Kernel estimation of probability density function prove an asymmetry of sample distribution

and 23 only. The index graph of jackknife residuals (Fig. 1g) also indicates suspicious points 2, 11, 20 and 23 and one outlier, 7, being greater than 3. The index plot of the Atkinson distance (Fig. 1h) shows one suspicious point, 20, and two outliers, 2 and 7 being outside the testing line in this graph. It may be concluded that one of the best diagnostic graphs for outlier detection is to be the Williams graph as it gives always clear detection of influential points and separate them on outliers and leverages.

## 4.2 Accuracy of p$K_\mathrm{a}$ prediction calculated with four algorithms

Four algorithms for p$K_\mathrm{a}$ prediction PALLAS [10], MARVIN [15], PERRIN [29] and SYBYL [13,14] were applied and their performance with effectiveness in the statistical accuracy test were compared extensively. As expected, the calculated values of p$K_{\mathrm{a,pred}}$ agree well with those of the experimental values p$K_{\mathrm{a,exp}}$.

Fitted residual evaluation can be quite an efficient tool in regression model building and testing. The correlations for all of the calculated values of p$K_\mathrm{a}$ from the four algorithms used and the experimental values using original data with outliers are as in Table 2. Figure 3 depicts a preliminary analysis of goodness-of-fit while Fig. 4 shows the Williams graph for identification and removal of outliers. In addition to the graphical analysis, the regression diagnostics of the fitness test prove the quality of p$K_\mathrm{a}$

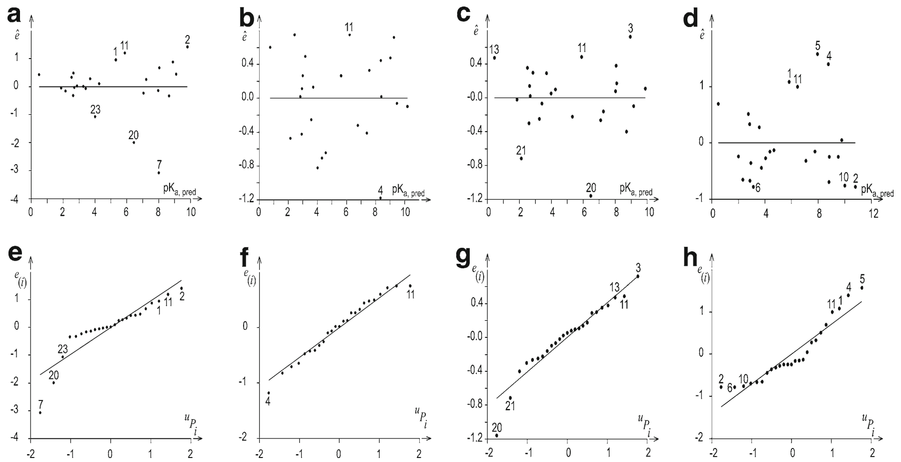

**Fig. 3** Comparison of four programs for the detection of outlying p$K_{\mathrm{a,pred}}$ values using the index graph of classical residuals (a, b, c, d in the upper part of figure) and the rankit Q–Q plot (e, f, g, h) for the data of Table 1. A Student $t$-test tests a null hypothesis $\mathrm{H}_0 : \mathrm{E}(\hat{e}) = 0$ vs. $\mathrm{H}_\mathrm{A} : \mathrm{E}(\hat{e}) \neq 0$ and for $n = 25$ and $\alpha = 0.05$ the descriptive statistics are calculated: **a** PALLAS: $\mathrm{E}(\hat{e}) = 0.13$, $s = 0.96$, test leading to $t_{\mathrm{exp}} = 0.69 < t_{\mathrm{crit}} = 2.06$, $P = 0.23$(H$_0$ is accepted), suspicious p$K_{\mathrm{a,pred}}$ values indicated 1, 2, 7, 11, 20, 23. **b** MARVIN: $\mathrm{E}(\hat{e}) = -0.19$, $s = 0.54$, test leading to $t_{\mathrm{exp}} = |-1.77| < t_{\mathrm{crit}} = 2.06$, $P = 0.05$ (H$_0$ is accepted), suspicious p$K_{\mathrm{a,pred}}$ values indicated: 4, 11. **c** Perrin: $\mathrm{E}(\hat{e}) = 0.12$, $s = 0.42$, test leading to $t_{\mathrm{exp}} = 1.42 < t_{\mathrm{crit}} = 2.06$, $P = 0.08$ (H$_0$ is accepted), suspicious p$K_{\mathrm{a,pred}}$ values indicated: 3, 11, 13, 20, 21. **d** SYBYL: $\mathrm{E}(\hat{e}) = -0.37$, $s = 0.70$, test leading to $t_{\mathrm{exp}} = |-2.64| > t_{\mathrm{crit}} = 2.06$, $P = 0.007$ (H$_0$ is rejected), suspicious p$K_{\mathrm{a,pred}}$ values indicated: 1, 2, 4, 5, 6, 10, 11

prediction. The highest values $R^2$, $R^2_\mathrm{P}$, the lowest value of *MEP* and $s$ and the more negative value of *AIC* in Fig. 5 and Table 2 are exhibited with the Perrin's algorithm of p$K_\mathrm{a}$ prediction, and this algorithm has the best predictive power and most accurate.

**Regression model**: The predicted versus the experimentally observed p$K_\mathrm{a}$ values for examined data set are plotted in Fig. 4a,b,c,d. The data points are distributed evenly around the diagonal in the figures, implying the consistent error behavior of the residual value. The slope and intercept of the linear regression are optimal; the slope estimates for the four algorithms used are $\beta_1(s_1) = 0.94(0.07, \mathbf{A}), 0.95(0.04, \mathbf{R}), 0.95(0.03, \mathbf{A})$, $1.04(0.05, \mathbf{A})$ where **A** or **R** means that the tested null hypothesis $\mathrm{H}_0 : \beta_0 = 0$ vs. $\mathrm{H}_\mathrm{A} : \beta_0 \neq 0$ and $\mathrm{H}_0 : \beta_1 = 1$ vs. $\mathrm{H}_\mathrm{A} : \beta_1 \neq 1$ was **A**ccepted or **R**ejected with the standard deviation of parameters estimates in brackets. Removing the outliers from the data set these estimates reach values $0.98(0.04, \mathbf{A}), 0.97(0.03, \mathbf{R}), 0.93(0.02, \mathbf{R}), 1.00(0.04, \mathbf{A})$. The intercept estimates are $\beta_0(s_0) = 0.17(0.41, \mathbf{A}), 0.43(0.23, \mathbf{R}), 0.12(0.17, \mathbf{A})$, $0.15(0.30, \mathbf{A})$ and after removing outliers from data set $0.14(0.22, \mathbf{A})$, $0.39(0.21, \mathbf{A})$, $0.28(0.11, \mathbf{R})$, $0.24(0.23, \mathbf{A})$. Here **A** or **R** means the tested null hypothesis $\mathrm{H}_0 : \beta_0 = 0$ vs. $\mathrm{H}_\mathrm{A} : \beta_0 \neq 0$ and $\mathrm{H}_0 : \beta_1 = 1$ vs. $\mathrm{H}_\mathrm{A} : \beta_1 \neq 1$ was **A**ccepted or **R**ejected. The slope is equal to one for 3 algorithms, excepting MARVIN, and the intercept is equal to zero for 3 algorithms, again excepting MARVIN. The Fisher-Snedecor $F$-test of overall regression in Fig. 5 leads to a calculated significance level of $P$ = 9.0E-13, 2.6E-18, 4.9E-21, 1.3E-16, and after removing the outliers from the data

**Table 2** Accuracy of p$K_a$ prediction of the REGDIA program which were calculated using the four algorithms: PALLAS, MARVIN, PERRIN and SYBYL

| Statistic used | PALLAS |  | MARVIN |  | PERRIN |  | SYBYL |  |
|---|---|---|---|---|---|---|---|---|
|  | 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**) | 0.17(0.41, A) | 0.14(0.22, A) | 0.43(0.23, R) | 0.39(0.21, A) | 0.12(0.17, A) | 0.28(0.11, R) | 0.15(0.30, A) | 0.24(0.23, A) |
| Slope $\beta_1 (s_1$, **A** or **R**) | 0.94(0.07, A) | 0.98(0.04, A) | 0.95(0.04, R) | 0.97(0.03, R) | 0.95(0.03, A) | 0.93(0.02, R) | 1.04(0.05, A) | 1.00(0.04, A) |
| $F_{exp}$ versus $F_{0.95}$(2-1, 25-2) = 4.12 | 195.7 | 631.6 | 638.4 | 782.2 | 1116.2 | 2323.5 | 446.7 | 634.5 |
| $P$ versus $\alpha = 0.05$ and $H_0$ : regression model is Accepted or Rejected | 9.0E-13, A | 1.3E-16, A | 2.6E-18, A | 1.1E-18, A | 4.9E-21, A | 3.6E-22, A | 1.3E-16, A | 3.6E-17, A |
| **Correlation** |  |  |  |  |  |  |  |  |
| Determination coefficient, $R^2$ [%] | 89.5 | 96.9 | 96.5 | 97.3 | 98.0 | 99.1 | 95.1 | 96.8 |
| Predicted determination coefficient, $R^2_P$ [%] | 76.6 | 92.8 | 91.97 | 93.6 | 95.3 | 97.9 | 88.8 | 92.6 |
| **Prediction ability criteria** |  |  |  |  |  |  |  |  |
| Mean error of prediction, *MEP* | 1.01 | 0.27 | 0.32 | 0.26 | 0.18 | 0.07 | 0.54 | 0.32 |
| Akaike information criterion, *AIC* | 0.02 | −28.4 | −28.94 | −32.8 | −42.9 | −57.8 | −15.6 | −26.0 |
| **Goodness-of-fit test** |  |  |  |  |  |  |  |  |
| $E(\hat{e})$ | 0.13 | −0.06 | −0.19 | −0.25 | 0.12 | 0.06 | −0.37 | −0.23 |
| $s$ in log units of p$K_a$ | 0.96 | 0.49 | 0.54 | 0.48 | 0.42 | 0.31 | 0.70 | 0.53 |
| $t_{exp}$ versus $t_{0.95}(n - m) = 2.06$ | 0.69 | −0.55 | −1.77 | −2.52 | 1.42 | 0.94 | −2.64 | −2.08 |
| $P$ versus $\alpha = 0.05$ and $H_0 : E(\hat{e}) = 0$ is Accepted or Rejected | 0.23, A | 0.29, A | 0.05, R | 0.009, R | 0.08, A | 0.18, A | 0.007, R | 0.02, R |
| **Outlier detection using the Williams plot** |  |  |  |  |  |  |  |  |
| Number of outliers detected | 3 | 0 | 1 | 0 | 3 | 0 | 2 | 0 |
| Indices of outliers detected | 2, 7, 20 | — | 4 | — | 3, 20, 21 | — | 4, 5 | — |

In intercept and slope estimates the letters **A** or **R** mean that the tested null hypothesis $H_0 : \beta_0 = 0$ vs. $H_A : \beta_0 \neq 0$ and $H_0 : \beta_1 = 1$ vs. $H_A : \beta_1 \neq 1$ was **A**ccepted or **R**ejected for the regression model proposed $pK_{a,pred} = \beta_0 + \beta_1 pK_{a,exp}$. In brackets are the standard deviation of the parameters estimates

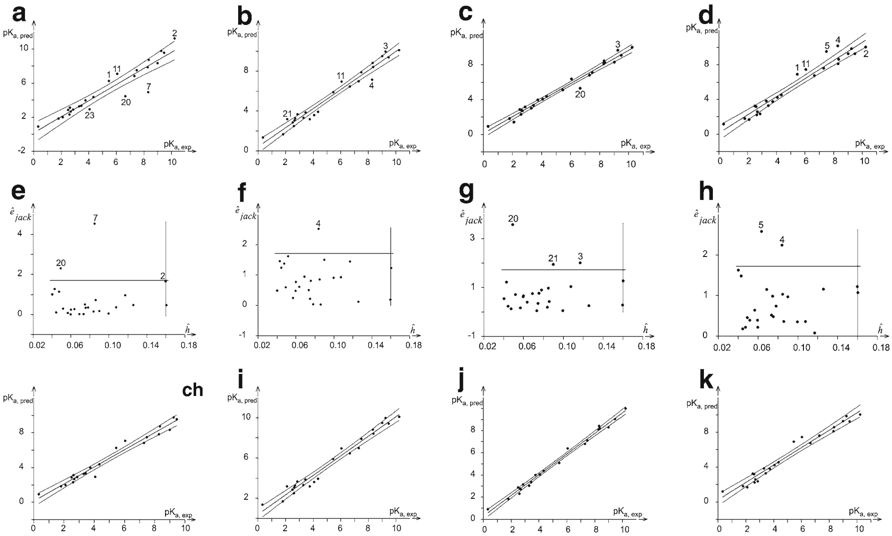

**Fig. 4** Accuracy examination and comparison of four programs for the predictive ability of the proposed regression model $\mathrm{p}K_{\mathrm{a,pred}} = \beta_0(s_0, \mathbf{A}$ or $\mathbf{R}) + \beta_1(s_1, \mathbf{A}$ or $\mathbf{R})\mathrm{p}K_{\mathrm{a,exp}}$ in a scatter diagram of the original data of Table 1 (upper part of figure denoted **a, b, c, d**), outlier detection with a Williams graph (middle part of figure denoted **e, f, g, h**) and in the scatter diagram after removing outliers from the data (lower part of figure **ch, i, j, k**) from the data of Table 1 with $n = 25$ and $\alpha = 0.05$ where **A** or **R** means that the tested null hypothesis $\mathrm{H}_0 : \beta_0 = 0$ vs. $\mathrm{H}_\mathrm{A} : \beta_0 \neq 0$ and $\mathrm{H}_0 : \beta_1 = 1$ vs. $\mathrm{H}_\mathrm{A} : \beta_1 \neq 1$ was **A**ccepted or **R**ejected. In brackets the standard deviation of an actual parameter is estimated. Data with outliers: (a) PALLAS: $\beta_0(s_0) = 0.17(0.41, \mathbf{A}), \beta_1(s_1) = 0.94(0.07, \mathbf{A}), R^2 = 89.5\%,\ s = 0.96,\ F = 195.7 > 4.12,\ P = 9.0\mathrm{E}-13, MEP = 1.01, AIC = 0.02,\ R_\mathrm{P}^2 = 76.6\%$, indicated outliers: 2, 7, 20. (b) MARVIN: $\beta_0(s_0) = 0.43(0.23, \mathbf{R}), \beta_1(s_1) = 0.95(0.04, \mathbf{R}), R^2 = 96.5\%, s = 0.54,\ F = 638.4 > 4.12,\ P = 2.6\mathrm{E}-18, MEP = 0.32, AIC = -28.94,\ R_\mathrm{P}^2 = 91.97\%$, indicated outliers: 4. (c) Perrin: $\beta_0(s_0) = 0.12(0.17, \mathbf{A}), \beta_1(s_1) = 0.95(0.03, \mathbf{A}), R^2 = 98.0\%, s = 0.42,\ F = 1116.2 > 4.12,\ P = 4.9\mathrm{E}-21,\ MEP = 0.18, AIC = -42.9,\ R_\mathrm{P}^2 = 95.3\%$, indicated outliers: 3, 20, 21. (d) SYBYL: $\beta_0(s_0) = 0.15(0.30, \mathbf{A}), \beta_1(s_1) = 1.04(0.05, \mathbf{A}), R^2 = 95.1\%, s = 0.70,\ F = 446.7 > 4.12,\ P = 1.3\mathrm{E}-16,\ MEP = 0.54, AIC = -15.60,\ R_\mathrm{P}^2 = 88.8\%$, indicated outliers: 4, 5. Data after removing outliers from data: (a) PALLAS: $\beta_0(s_0) = 0.14(0.22, \mathbf{A}), \beta_1(s_1) = 0.98(0.04, \mathbf{A}), R^2 = 96.9\%, s = 0.49,\ F = 631.6 > 4.35,\ P = 1.3\mathrm{E}-16,\ MEP = 0.27, AIC = -28.4,\ R_\mathrm{P}^2 = 92.8\%$, (b) MARVIN: $\beta_0(s_0) = 0.39(0.21, \mathbf{A}), \beta_1(s_1) = 0.97(0.03, \mathbf{R}), R^2 = 97.3\%, s = 0.48,\ F = 782.2 > 4.30,\ P = 1.1\mathrm{E}-18,\ MEP = 0.26, AIC$ =-32.8, $R_\mathrm{P}^2 = 93.6\%$, (c) Perrin: $\beta_0(s_0) = 0.28(0.11, \mathbf{R}), \beta_1(s_1) = 0.93(0.02, \mathbf{R}), R^2 = 99.1\%, s = 0.31,\ F = 2323.5 > 4.35,\ P = 3.6\mathrm{E}-22,\ MEP = 0.07, AIC = -57.8,\ R_\mathrm{P}^2 = 97.9\%$, (d) SYBYL: $\beta_0(s_0) = 0.24(0.23, \mathbf{A}), \beta_1(s_1) = 1.00(0.04, \mathbf{A}), R^2 = 96.8\%, s = 0.53,\ F = 634.5 > 4.32,\ P = 3.6\mathrm{E}-17,\ MEP = 0.32, AIC = -26.0,\ R_\mathrm{P}^2 = 92.6\%$

set of $P$ = 1.3E-16, 1.1E-18, 3.6E-22, 3.6E-17, meaning that all four algorithms proposed a significant regression model. The highest value of $F$-test is exhibited by the PERRIN algorithm.

**Correlation**: The quality of the regression models yielded by the four algorithms was measured using the two statistical characteristics of correlation in Fig. 5, *i.e.* $R^2 = 89.5, 96.5, 98.0$, and $95.1\%$ and after removing the outliers from the data set $R^2 = 96.9$, 97.3, 99.1, and 96.8% and $R_\mathrm{P}^2 = 76.6$, 92.0, 95.3, and 88.8% and after removing the

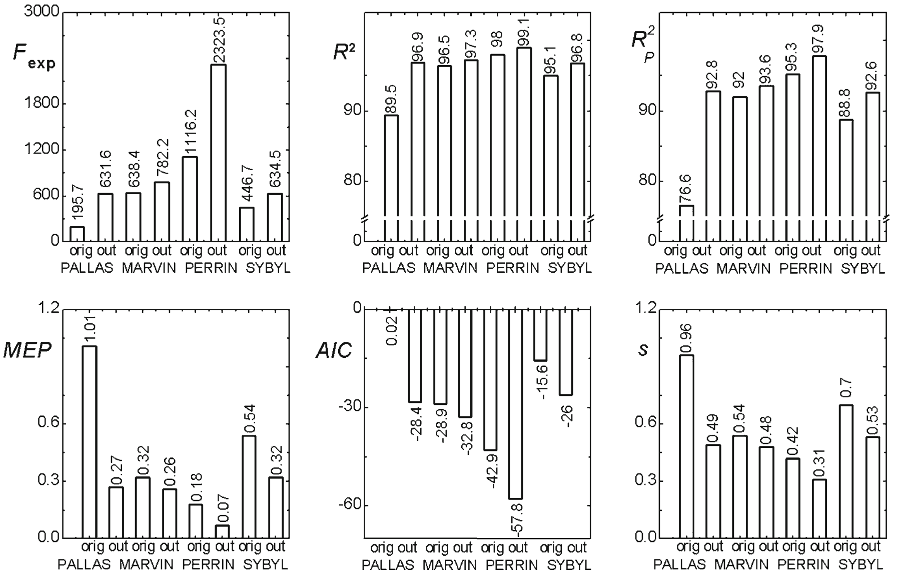

**Fig. 5** Resolution capability of the six regression diagnostic criteria $F_{\text{exp}}$, $R^2$, $R^2_{\text{P}}$, *MEP*, *AIC* and $s$ for an accuracy examination of p$K_{\text{a}}$ prediction when the four algorithms PALLAS, MARVIN, PERRIN and SYBYL are tested and compared. Here *orig* means the original data set and *out* means the data set without outliers

outliers from the data set $R^2_{\text{P}} = 92.8$, 93.6, 97.9, and 92.6%. $R^2$ is prominently high for all four algorithms and indicates an algorithm's ability to interpolate within the range of p$K_{\text{a}}$ values in the examined data set. The highest value is exhibited by the PERRIN algorithm.

**Prediction ability criteria**: The most efficient criteria of a goodness-of-fit test to also express a predictive ability are the mean error of prediction *MEP* and the Akaike information criterion *AIC* in Fig. 5. Calculated *MEP* lead to values of 1.01, 0.32, 0.18, 0.54, and after removing the outliers from the data set *MEP* reaches the lower values of 0.27, 0.26, 0.07, 0.32. The Akaike information criterion *AIC* yields values of 0.02, −28.94, −42.9, −15.6, and after removing the outliers from the data set *AIC* reaches more negative values of −28.4, −32.8, −57.8, −26.0. This shows that both *MEP* and *AIC* classify the predictive ability of the 4 algorithms well. The lowest value of $MEP = 0.18$ is attained for the Perrin method, while the most negative value of $AIC = -42.9$ is also attained for the Perrin method. This criteria used efficiently classify predictive ability of the regression model, and classify the four algorithms compared from best to worst. The regression models are predictive enough, i.e. are also able to extrapolate beyond the training set.

**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 having a normal distribution N(0, $s^2$) with the residual mean equal to 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 criteria value for the four algorithms in the form of the calculated significance levels $P = 0.23, 0.05, 0.08, 0.007$. Three algorithms give a residual bias equal to zero, the exception being the SYBYL. The estimated standard deviation of regression straight line in Fig. 5 is $s = 0.96, 0.54, 0.42, 0.70$ log units p$K_a$, and after removing the outliers from the data set $s = 0.49, 0.48, 0.31, 0.53$ log units p$K_a$, the lowest value being attained for the Perrin method.

**Outlier detection**: The detection, assessment, and understanding of outliers in p$K_{a,pred}$ values are major areas of interest in an accuracy examination. If the data contains a single outlier p$K_{a,pred}$, the problem of identifying such a p$K_{a,pred}$ value is relatively simple. If the p$K_{a,pred}$ data contains more than one outlier (which is likely to be the case in most data), the problem of identifying such p$K_{a,pred}$ values becomes more difficult, due to the masking and swamping effects [36]. *Masking* occurs, when an outlying p$K_{a,pred}$ goes undetected because of the presence of another, usually adjacent, p$K_{a,pred}$ subset. *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$K_{a,pred}$. Statistical tests are needed to decide how to use the real data, in order approximately to satisfy the assumptions of the hypothesis tested. In the PALLAS straight line model three outliers, 2, 7 and 20 were detected. In the MARVIN straight line model only one outlier, 4, was detected. In Perrin's straight line three outliers, 3, 20 and 21, were detected, while in the SYBYL straight line model only two outliers, 4 and 5, were detected.

**Outlier interpretation and removal:** Poorest molecular p$K_a$ predictions are indicated as 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 elucidated and removed from the data: here with the use of the Williams plot three outliers, i.e. outlier no. 2 (1-4'-hydroxycyclohexyl-2-isopropylaminoethanol), outlier no. 7, (2-methylaminoacetamide) and outlier no. 20, (4-aminopyridazine) were detected in the PALLAS regression model (Fig. 4e), one outlier no. 4 (N,N-dimethyl-2-butyn-1-amine) in the MARVIN model (Fig. 4f), three outliers, i.e. outlier no. 3, (2-aminocycloheptanol), outlier no. 20, (4-aminopyridazine) and outlier no. 21, (4-amino-6-chloropyrimidine) in Perrin's model (Fig. 4g) and two outliers, i.e. outlier no. 4 N,N-dimethyl-2-butyn-1-amine, outlier no. 5 (5-chloro-3-methyl-3-azapentanol) in the SYBYL model (Fig. 4h). Removing the outlying values of p$K_a$ poorly predicted molecules, all the remaining data points were statistically significant (Fig. 4ch,i,j,k). Outliers frequently turned out to be either misassignment of p$K_a$ values or suspicious molecular structure. The fragment based approach is inadequate when fragments present in a molecule under study are absent in the database. Such p$K_a$ prediction only depends on the compounds very similar to those available in the training set. Suitable corrections are made where possible, but in some cases the corresponding data had to be omitted from the training set. In other cases, outliers served to point up a 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 is embedded.

## 5 Conclusions

Most poorly predicted molecular p$K_\text{a}$ are indicated as outliers. Seven selected diagnostic plots (the Graph of predicted residuals, Williams graph, Pregibon graph, Gray L–R graph, Scatter plot of classical residuals versus prediction, Index graph of jackknife residuals, Index graph of Atkinson distance) were chosen as the best and most efficient to give reliable detection of outlying p$K_\text{a}$ values. The proposed accuracy test of the REGDIA program can also be extended for other predicted values, as log $P$, log $D$, aqueous solubility, and some physicochemical properties.

**Novelty:** Many structure-property algorithms have been used to predict p$K_\text{a}$ but mostly without a rigorous test of p$K_\text{a}$ accuracy. The regression diagnostics algorithm REGDIA in S-Plus is introduced here to test and to compare the accuracy of p$K_\text{a}$ predicted with four programs, PALLAS, MARVIN, PERRIN and SYBYL. Indicated outliers in p$K_\text{a}$ relate to molecules which are poorly characterized by the considered p$K_\text{a}$ algorithm. Of the seven most efficient diagnostic plots the Williams graph was selected to give the most reliable detection of outliers. The six statistical characteristics, $F_\text{exp}$, $R^2$, $R_\text{P}^2$, *MEP*, *AIC*, and $s$ in p$K_\text{a}$ units, successfully examine the p$K_\text{a}$ data. The proposed accuracy test can also be applied to test any other predicted values, such as log $P$, log $D$, aqueous solubility or some physicochemical properties.

**Acknowledgements** The financial support of the 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. L. Xing, R.C. Glen, Novel methods for the prediction of log$P$, p$K$ and log$D$. J. Chem. Inf. Comput. Sci. **42**, 796–805 (2002)
2. L. Xing, R.C. Glen, R.D. Clark, Predicting p$K_\text{a}$ by molecular tree structured fingerprints and PLS. J. Chem. Inf. Comput. Sci. **43**, 870–879 (2003)
3. J. Zhang, T. Kleinöder, J. Gasteiger, Prediction of p$K_\text{a}$ values for aliphatic carboxylicv acids and alcohols with empirical atomic charge descriptors. J. Chem. Inf. Model. **46**, 2256–2266 (2006)
4. N.T. Hansen, I. Kouskoumvekaki, F.S. Jorgensen, S. Brunak, S.O. Jonsdottir, Prediction of pH-dependent aqueous solubility of druglike molecules. J. Chem. Inf. Model. **46**, 2601–2609 (2006)
5. ACD/Labs™, p$K_\text{a}$ Predictor 3.0, Advanced Chemistry Development Inc. 133 Richmond St. W. Suite 605, Toronto
6. R.F. Rekker, A.M. ter Laak, R. Mannhold, Prediction by the ACD/p$K_\text{a}$ method of values of the acid-base dissociation constant (p$K_\text{a}$) for 22 drugs. Quant. Struct. Act. Relat. **12**, 152 (1993)
7. B. Slater, A. McCormack, A. Avdeef, J.E.A. Commer, Comparison of ACD/p$K_\text{a}$ with experimental values. Pharm. Sci. **83**, 1280–1283 (1994)
8. Results of titrometric measurements on selected drugs compared to ACD/p$K_\text{a}$ September 1998 predictions, (Poster), AAPS, Boston, November 1997
9. P. Fedichev, L. Menshikov, Long-range interactions of macroscopic objects in polar liquids, Quantum p$K_\text{a}$ calculation module, QUANTUM pharmaceuticals, http://www.q-lead.com
10. Z. Gulyás, G. Pöcze, A. Petz, F. Darvas, Pallas cluster—a new solution to accelerate the high-throughut ADME-TOX prediction, ComGenex-CompuDrug, PKALC/PALLAS 2.1 CompuDrug Chemistry Ltd., http://www.compudrug.com
11. J. Kenseth, Ho-ming Pang, A. Bastin, Aqueous p$K_\text{a}$ determination using the p$K_\text{a}$ Analyzer Pro™, http://www.CombiSep.com

12. V. Evagelou, A. Tsantili-Kakoulidou, M. Koupparis, Determination of the dissociation constants of the cephalosporins cefepime and cefpirome using UV spectrometry and pH potentiometry. J. Pharm. Biomed. Anal. **31**, 1119–1128 (2003)
13. E. Tajkhorshid, B. Paizs, S. Suhai, Role of isomerization barriers in the p$K_a$ control of the retinal Schiff base: a density functional study. J. Phys. Chem. B **103**, 4518–4527 (1999)
14. SYBYL is distributed by tripos, Inc., St. Louis MO 63144, http://www.tripos.com
15. Marvin: http://www.chemaxon.com/conf/Prediction_of_dissociation_constant_using_microcon-stants.pdf and http://www.chemaxon.com/conf/New_method_for_pKa_estimation.pdf
16. W.A. Shapley, G.B. Bacskay, G.G. Warr, Ab initio quantum chemical studies of the p$K_a$ values of hydroxybenzoic acids in aqueous solution with special reference to the hydrophobicity of hydroxybenzoates and their binding to surfactants. J. Phys. Chem. B **102**, 1938–1944 (1998)
17. G. Schueuermann, M. Cossi, V. Barone, J. Tomasi, Prediction of the p$K_a$ of carboxylic acids using the ab initio Continuum-Solvation Model PCM-UAHF. J. Phys. Chem. A **102**, 6707–6712 (1998)
18. C.O. da Silva, E.C. da Silva, M.A.C. Nascimento, Ab initio calculations of absolute p$K_a$ values in aqueous solution I. Carboxylic acids. J. Phys. Chem. A **103**, 11194–11199 (1999)
19. N.L. Tran, M.E. Colvin, The prediction of biochemical acid dissociation constants using first principles quantum chemical simulations. Theochem **532**, 127–137 (2000)
20. M.J. Citra, Estimating the p$K_a$ of phenols, carboxylic acids and alcohols from semiempirical quantum chemical methods. Chemosphere **38**, 191–206 (1999)
21. I.J. Chen, A.D. MacKerell, Computation of the influence of chemical substitution on the p$K_a$ of pyridine using semiempirical and ab initio methods. Theor. Chem. Acc. **103**, 483–494 (2000)
22. D. Bashford, M. Karplus, p$K_a$'s of ionizable groups in proteins: atomic detail from a continuum electrostatic model. Biochemistry **29**, 10219–10225 (1990)
23. H. Oberoi, N.M. Allewell, Multigrid solution of the nonlinear Poison–Boltzmann equation and calculation of titration curves. Biophys. J. **65**, 48–55 (1993)
24. J. Antosiewicz, J.A. McCammon, M.K. Gilson, Prediction of pH-dependent properties of proteins. J. Mol. Biol. **238**, 415–436 (1994)
25. Y.Y. Sham, Z.T. Chu, A. Warshel, Consistent calculation of p$K_a$'s of ionizable residues in proteins: semi-microscopic and microscopic approaches. J. Phys. Chem. B **101**, 4458–4472 (1997)
26. K.H. Kim, Y.C. Martin, 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 (1991)
27. K.H. Kim, Y.C. Martin, 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 (1991)
28. R. Gargallo, C.A. Sotriffer, K.R. Liedl, B.M. Rode, Application of multivariate data analysis methods to comparative molecular field analysis (CoMFA) data: proton affinities and p$K_a$ prediction for nucleic acids components. J. Comput. Aided Mol. Des. **13**, 611–623 (1999)
29. D.D. Perrin, B. Dempsey, E.P. Serjeant, *p$K_a$ prediction for organic acids and bases* (Chapman and Hall Ltd., London, 1981)
30. CompuDrug NA Inc., pKALC version 3.1, (1996)
31. ACD Inc. ACD/p$K_a$ version 1.0, (1997)
32. http://chemsilico.com/CS_prpKa/PKAhome.html. Accessed Aug 2006
33. A. Habibi-Yangjeh, M. Danandeh-Jenagharad, M. Nooshyar, Prediction acidity constant of various benzoic acids and phenols in water using linear and nonlinear QSPR models. Bull. Korean Chem. Soc. **26**, 2007–2016 (2005)
34. P.L.A. Popelier, P.J. Smith, QSAR models based on quantum topological molecular similarity. Eur. J. Med. Chem. **41**, 862–873 (2006)
35. G.H. Schmid et al., The application of iterative optimization techniques to chemical kinetic data of large random error. Can. J. Chem. **54**, 3330–3341 (1976)
36. M. Meloun, J. Militký, M. Forina, Chemometrics for Analytical Chemistry, Vol. 2. PC-Aided Regression and Related Methods (Ellis Horwood, Chichester, 1994), and Vol. 1. PC-Aided Statistical Data Analysis (Ellis Horwood, Chichester, 1992)
37. S-PLUS: MathSoft, Data Analysis Products Division, 1700 Westlake Ave N, Suite 500, Seattle, WA 98109, USA, http://www.insightful.com/products/splus (1997)
38. ADSTAT: ADSTAT 1.25, 2.0, 3.0 (Windows 95), TriloByte Statistical Software Ltd., Pardubice, Czech Republic

39. D.A. Belsey, E. Kuh, R.E. Welsch, *Regression Diagnostics: Identifying Influential data and Sources of Collinearity* (Wiley, New York, 1980)
40. R.D. Cook, S. Weisberg, *Residuals and Influence in Regression* (Chapman & Hall, London, 1982)
41. A.C. Atkinson, *Plots, Transformations and Regression: An Introduction to Graphical Methods of Diagnostic Regression Analysis* (Claredon Press, Oxford, 1985)
42. S. Chatterjee, A.S. Hadi, *Sensitivity Analysis in Linear Regression* (Wiley, New York, 1988)
43. V. Barnett, T. Lewis, *Outliers in Statistical Data*, 2nd edn. (Wiley, New York, 1984)
44. R.E. Welsch, Linear Regression Diagnostics, Technical Report 923-77, Sloan School of Management, Massachusetts Institute of Technology, (1977)
45. S. Weisberg, *Applied Linear Regression* (Wiley, New York, 1985)
46. P.J. Rousseeuw, A.M. Leroy, *Robust Regression and Outlier Detection* (Wiley, New York, 1987)