p*K*a Prediction for Acidic Phosphorus-Containing Compounds Using Multiple Linear Regression with Computational Descriptors

Donghai Yu, Ruobing Du, and Ji-Chang Xiao*

Ninety-six acidic phosphorus-containing molecules with pK_a 1.88 to 6.26 were collected and divided into training and test sets by random sampling. Structural parameters were obtained by density functional theory calculation of the molecules. The relationship between the experimental pK_a values and structural parameters was obtained by multiple linear regression fitting for the training set, and tested with the test set; the R^2 values were 0.974 and 0.966 for the training and test sets, respectively. This regression equation, which quantitatively describes the influence of structural parameters on pK_a, and can be used to predict pK_a values of similar structures, is significant for the design of new acidic phosphorus-containing extractants. © 2016 Wiley Periodicals, Inc.

DOI: 10.1002/jcc.24381

Introduction

Acidic phosphorus-containing compounds are an important class of chemicals, which are widely used as extractants in hydrometallurgy, especially in rare-earth extraction and separation. The extractability of an extractant is greatly influenced by its acid dissociation constant, i.e. pK_a. Generally, the lower the pK_a of the extractant, the higher the extraction rate. For example, the order of the pK_a values of di(2-ethylhexyl)phosphinic acid, 2-ethylhexylphosphonic acid mono-2-ethylhexyl ester, and di(2-ethylhexyl) phosphoric acid is the opposite of the order of their lanthanide extraction rates.[1,2] Additionally, the pK_a is a fundamental property of acidic compounds; it is mainly used to measure proton dissociation ability, and is important to acidity-related properties such as biocompatibility and activity.[3–7]

The pK_a is measured experimentally by acid–base titration, and the results are affected by the temperature, solvent, concentration, and other experimental conditions. Even for the same molecule measured under similar or identical conditions, different pK_a values are frequently reported. Theoretical methods are therefore employed to explore the relationships between pK_a and molecular structure.[3,7–9] Two techniques have been developed for this purpose, i.e., first-principle calculation and the quantitative structure–property relationship (QSPR) approach. First-principle calculation are based on thermodynamic energies; the Gibbs free energies of the deprotonation reaction are calculated, and the pK_a is then obtained using a formula.[9–11] This method depends on the deprotonation thermodynamic process, and is based on basic physical concepts.[9–15] Theoretically, first-principle calculation can be used to analyze any deprotonation process of any molecule. However, to achieve sufficient accuracy is a big challenge. For example, an accuracy of ± 1 pK_a unit is needed to control errors in the Gibbs free energy changes of deprotonation within ± 1.36 kcal/mol, but the errors in ion solvation are roughly 2 to 5 kcal/mol.[8]

The QSPR approach is based on linear free energy analysis and relies on molecular descriptors. Generally, three steps are involved: first, choosing training and test sets, and listing the pK_a values and molecule descriptors, e.g., bond lengths, energies, and charges; second, fitting equations to the relevant pK_a values and structural parameters in the training set; and third, examining the equations with the test set.[13–15] An effective relationship will be established if the equation is verified by the test set, and it can be used to predict the pK_a values of molecules with similar structures. In early QSPR approaches, a fitted numerical equation based on experimental molecular descriptors was used to represent the relationship between experimental data; some descriptors are difficult to obtain, therefore this approach is not useful. Additionally, abundant data are needed for fitting, which are suitable for similar structures, but not available for different types of molecule. Based on the respective characteristics of these two techniques, quantum QSPR methods have been developed in recent years, in which the molecular descriptors and detailed information are provided by quantum chemical computations.[16,17] These methods are not restricted by the experimental conditions and do not require the precision in pure first-principles methods.

D. Yu, R. Du, Ji-Chang Xiao
Key Laboratory of Organofluorine Chemistry, Shanghai Institute of Organic Chemistry, Chinese Academy of Sciences, Shanghai, China
E-mail: jchxiao@sioc.ac.cn

Contract grant sponsor: National Basic Research Program of China; Contract grant number: 2012CBA01200 and 2015CB931900; Contract grant sponsor: National Natural Science Foundation; Contract grant numbers: 21172240; 21421002; 21472222; Contract grant sponsor: Chinese Academy of Sciences; Contract grant number: XDA02020105 and XDA02020106

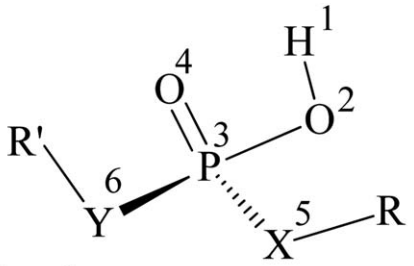
Both QSPR approaches are widely used to investigate various properties, including the pK_a.[18] Selecting appropriate descriptors is critical for developing a successful equation. The fittings of various descriptors such as empirical atomic charges, electrostatic energies, cavity energies, dispersion energies, repulsion interactions, and bond lengths to pK_a have been reported in the literature.[15,17,19] A broad range of substrates have also been investigated, including alcohols, phenols, and carboxylic acids with different pK_a ranges. Both single and multiple variable regressions have been reported.[14,15] The prediction accuracies vary depending on the method or substrate; the R^2 values range from less than 0.85 to higher than 0.99, and the root-mean-square errors vary from higher than 0.75 pK_a units to lower than 0.40 units.[13–16,20] Harding et al. reviewed the descriptors related to pK_a.[3,14,15] The results reported in the literature show that multiple linear regression (MLR) generally gives better results than simple linear regression, for both empirical and quantum QSPR approaches.

The relationship between pK_a and the molecular structures of compounds used as extractants was early studied experimentally using empirical QSPR approaches.[21,22] However, the results are difficult to apply in the prediction of new molecules. In this work, MLR was used to fit numerical equations for experimental pK_a values and quantum descriptors for a set of acidic phosphorus-containing molecules (training set). The equation was used to predict the pK_a values of another set of acidic phosphorus-containing molecules (test set). The results indicate that the equation can be used to estimate the pK_a values of these molecules, and can provide theoretical support for the design of new extractants.

Methods

Model structures

Ninety-six acidic phosphorus-containing molecules were collected from literature reports, namely 5 dialkylphosphoric acids, 86 alkyl phosphonates, 5 dialkyl phosphates, and 2 alkylphosphoric acids (only the pK_{a1} values were considered). All the data were obtained in 75% (volume ratio with water) ethanol solution by Yuan's group.[2,21–28] The pK_a of acidic phosphorus-containing compounds mainly depend on the electronic effect and steric effect of the phosphorus centers, the former is dominant. The alkoxy and aryl are stronger electron-withdrawing than alkyl, which are benefit for proton dissociating. The structure that containing more electron-withdrawing group own lower pK_a value. However, the proton dissociation is inhibited by steric effect, the structures that side chain is closer the phosphorus center own higher pK_a value. A test set and a training set were obtained by random sampling; the maximum and minimum pK_a values were in the training set, as shown in the left half of Table (see Supporting Information for details). The geometries were optimized using a density functional theory method, and several structural parameters were collected, as shown in Scheme 1.



Abbreviations of structure parameters:
$Q_{E,x}$, $Q_{M,x}$, $Q_{N,x}$ represent the ESP charge, Mulliken charge, NBO charge of atom x, x=1, 2, … , 6.
$R_{x,y}$ represent the distance between atom x and y, x,y=1, 2, … , 6.
H and L represent the HOMO and LUMO of molecules.
X, Y= Carbon, Oxygen.

Scheme 1. The atomic number and the descriptor abbreviations.

Computational details

All geometries were fully optimized, without any constraints, at the M062X theory level in water, using the conductor-like polarizable continuum model solvation model with default calculation condition of package (the static dielectric constant of water is 78.3533);[29,30] the CC-PVTZ basis set was used for all atoms.[31] The vibrational frequency was computed for each structure to determine whether it was a minimum point (no imaginary frequency). The conformation with the lowest free energy was selected for molecules that had more than one possible conformation. The different conformations were only result from change of side chain direction, which is a little variation to whole molecule. So the free energies contributions of vibration were neglected. Natural bond orbital (NBO) analysis was performed at the same theory level.[32,33] All calculations were performed using the Gaussian 09, Revision D.01 software package.[34]

The structural parameters, listed in Supporting Information Table S2, and assigned as shown in Scheme 1, were the electrostatic potential (ESP), Mulliken, NBO charges, bond lengths of key atoms, and the highest occupied molecular orbitals (HOMOs) and lowest unoccupied molecular orbitals (LUMOs) (see Supporting Information for details). The test and training sets were selected by random sampling to avoid subjective errors. The training set contained 51 structures, and the test set contained 45 structures. The numerical equation describing the relationship between the experimental pK_a values and the structural parameters was obtained by MLR, performed using the SPSS 19.0 (Statistical Package for the Social Sciences) software package; the results were assessed based on the p value and R^2. The test set was also tested using SPSS.

Table 1. The parameters of $r^2 > 0.5$.

Parameters	$Q_{N,4}$	$Q_{M,2}$	$R_{3,4}$	$R_{2,3}$	$Q_{N,1}$	$Q_{N,2}$	L
R^2	0.8616	0.6474	0.6470	0.6000	0.5994	0.5706	0.5597

Table 2. The MLR results with constant and without.

Parameters	Coefficients	Standard error	*p*	Coefficients	Standard error	*p*
Intercept	−20.2272	9.6952	0.0425	0	0	\
$Q_{E,4}$	5.7303	0.7723	2.14E-09	5.3403	0.7756	1.22E-08
$Q_{N,1}$	−83.6396	11.0230	1.21E-09	−104.6919	4.5923	4.69E-27
$Q_{N,4}$	−62.3397	4.3680	1.64E-18	−54.2396	2.0716	1.12E-29
L	9.0450	0.8786	1.60E-13	9.5744	0.8706	1.36E-14

Results and Discussion

Fitting of equation for training set

The r^2 values for the relationships between the structural parameters and experimental pK_a values were determined; the parameters with $r^2 > 0.5$ are listed in Table 1.

Among the parameters listed in Table 1, the most relevant structural parameters are $Q_{N,4}$, $Q_{M,2}$, $R_{3,4}$, $Q_{N,1}$, $R_{2,3}$, $Q_{N,2}$, and *L*, according to the r^2 values; this is in agreement with reports of similar studies.[3] However, the highest r^2 value is only 0.86, obtained from the pK_a and NBO charge on the phosphoryl oxygen. The values for $R_{3,4}$ and $R_{2,3}$ are 0.65 and 0.60, respectively; these are not in agreement with those for carboxylic acids in Alkorta et al.'s study,[17] in which the relationship between C═O and C–OH carboxyl bond lengths and pK_a had high R^2 values. This is because all the carboxylic acids in Alkorta et al.'s study have similar structures, whereas for the acidic phosphorus-containing molecules in this work, the different phosphorus centers, i.e. two P–C, P–C and P–O, two P–O, and P–Ph bonds, were taken into account, and these may have different relationships. These results indicate that a single descriptor is insufficient for interpreting the trends in pK_a changes. The poor fitting results for the training set were unsuitable for the test set, therefore MLR was used.

$$\begin{aligned} \mathrm{p}K_a = {} & (-20.2272 \pm 9.6952) + (5.7303 \pm 0.7723)Q_{E,4} \\ & + (9.0450 \pm 0.8786)L + (-83.6396 \pm 11.0230) \\ & Q_{N,1} + (-62.3397 \pm 4.3680)Q_{N,4}, R^2 = 0.9742. \end{aligned} \quad (1)$$

$$\begin{aligned} \mathrm{p}K_a = {} & (5.3403 \pm 0.7756)Q_{E,4} + (9.5744 \pm 0.8706)L \\ & + (-104.6919 \pm 4.5923)Q_{N,1} + \\ & (-54.2396 \pm 2.0716)Q_{N,4}, R^2 = 0.9990. \end{aligned} \quad (2)$$

The best MLR results were usually obtained from parameters with high R^2 values; however, the square of the multiple correlation coefficient among seven variables with pK_a was 0.946. This is because these variables are strongly related (see Supporting Information for details). To achieve the best fitting results, automatic regressions were performed for all parameters; the best results are listed in Table 2. Because the *p* value, 0.04, is close to the default critical value of 0.05 for the constant term, the zero constant result was also investigated. As shown in the right half of Table 2, the R^2 values were 0.9742 and 0.9990 with and without a constant, respectively. As can be seen from the squares of the correlation coefficients, R^2 for multivariate analysis was much higher than that for a single variable, whether or not a constant was present. This indicates that MLR is more effective than single-parameter regression. The pK_a is influenced by various substituent effects, which do not necessarily reflect the same parameter, therefore it is reasonable that changes in the pK_a trends cannot be explained by a single variable. The *p* values are much lower than 0.05 for all the parameters in Table 2, both with and without a constant. This shows that these parameters are effective in the equations, and the regression results are not significantly changed by using different training sets.[35,36]

The regression results listed in Table 2 can be transformed into numerical equations, i.e. eqs. (1) and (2), for constant and zero constant cases, respectively. These more directly reflect the relationships between the experimental pK_a and theoretical parameters. The physical meaning of the coefficients is the change in the value of pK_a when the variable changes by one unit.[36] The absolute values of the coefficients indicate the influences on pK_a of the variables. The greatest influence on the pK_a is exerted by the hydroxyl hydrogen atom, followed by the phosphoryl oxygen; this is in agreement with previous reports that these two groups were frequently used to explore the pK_a.[3] It can be seen from the numerical coefficients that the pK_a is higher when $Q_{N,4}$ is lower, *L* is higher, $Q_{E,4}$ is higher, and $Q_{N,1}$ is lower. The NBO and ESP charges on phosphoryl oxygen play different roles in the equations because they are obtained using different algorithms. The NBO charge is obtained from natural orbital population analysis at each atomic center,[37] and the ESP charge is based on an exact one-electron property calculated from the molecular wavefunction of space.[33] Additionally, the r^2 value for $Q_{N,4}$ and $Q_{E,4}$ is 0.187, therefore it is statistically reasonable to use the two charge parameters in the same equation. Based on eqs. (1) and (2), for a higher pK_a, with a more negative $Q_{N,4}$, which results from the electron-donating effect of the group bonded to the phosphorus atom, the hydroxyl oxygen is also affected, and the O–H bond is strengthened; otherwise a more negative $Q_{N,4}$ leads to a stronger hydrogen bond between the phosphoryl oxygen and hydroxyl hydrogen, which inhibits proton dissociation.[13,38] For a higher pK_a, a higher LUMO energy may lead to a higher-energy dissociated anion, which is more unstable.[39] The ESP charge of the phosphoryl oxygen is different to its NBO charge. A more positive $Q_{E,4}$ gives a higher pK_a, because the ESP charge is calculated from the electron density; a more positive $Q_{E,4}$ leads to a stronger O–H bond through an inductive effect, and disfavors proton dissociation.[40] For the hydroxyl hydrogen, a more positive $Q_{N,1}$ leads to a lower pK_a, because the more positive $Q_{N,1}$ indicates higher ionicity of the hydroxyl bond, with closer dissociation states.

Testing of test set

The square of the multiple correlation coefficient R^2 in the zero constant of eq. (2) is higher than that in eq. (1). However, the R^2 value for eq. (2) is 0.971, and that for eq. (1) is 0.974 for the predicted *versus* experimental pK_a values for the training set. Equation (1) is therefore slightly better; the results are displayed in Supporting Information Table S5. For the training set, the maximum error is 0.285, at No. 60, and seven structures have errors greater than 0.200; the standard error is 0.128, and R^2 is 0.974, i.e. 97.4% of the change in pK_a can be interpreted, based on the four parameters in eq. (1). This is statistically effective regression and comparable to previously reported results.

The pK_a values in the training set ranged from 1.88 to 6.26, and are nearly all for acidic phosphorus-containing structures, as well as being applicable to a wider range of similar molecules. The test set of 45 structures gained from random sampling was used to test the method. Equation (1) was more effective than eq. (2) for prediction using the training set, therefore eq. (1) was also used to predict the pK_a values for the test set, without fitting. The results are shown in the right half of Supporting Information Table S5. For the test set, the maximum error was 0.344, at No. 27, and five structures had errors greater than 0.200. The standard error was 0.126, and R^2 was 0.966, which is statistically sufficient for predictions.[14,15,41] MLR is therefore an efficient method for predicting the pK_a values of acidic phosphorus-containing structures.

Conclusions

This work represents the first study of the relationship between the experimental pK_a values and structural parameters of acidic phosphorus-containing molecules using MLR. A numerical equation was fitted from the training set, and tested on the test set. Efficient results were obtained for both the training and test sets, with R^2 values of 0.97 and 0.96, respectively, and maximum errors of 0.285 and 0.344, respectively. This accuracy is similar to those in previous reports. In this work, random sampling was used to select the training and test sets, to avoid subjective errors. Almost the same accuracies were obtained for the two sets, therefore the equations are suitable for predicting pK_a values of these types of structure. This method will therefore be useful in designing new acidic phosphorus-containing extractants.

Acknowledgment

Computing resources were provided by the National Supercomputing Center of China in Shenzhen.

Keywords: pK_a prediction · multiple linear regression · acidic phosphorus-containing compounds · computational descriptors

How to cite this article: D., Yu, R., Du, J.-C., Xiao *J. Comput. Chem.* **2016**, *37*, 1668–1671. DOI: 10.1002/jcc.24381

Additional Supporting Information may be found in the online version of this article.

[1] C. Y. Yuan, *Chin. J. Org. Chem.* **1979**, 43.
[2] C. Y. Yuan, J. Y. Yan, H. Z. Feng, H. Y. Long, P. B. Wu, P. L. Jin, *Sci. Chin. Ser. B* **1986**, 1150.
[3] A. P. Harding, D. C. Wedge, P. L. A. Popelier, *J. Chem. Inf. Model* **2009**, *49*, 1914.
[4] T. Meyer, E. W. Knapp, *J. Chem. Theor. Comput.* **2015**, *11*, 2827.
[5] B. Thapa, H. B. Schlegel, *J. Phys. Chem. A* **2015**, *119*, 5134.
[6] M. Zrnčić, S. Babić, D. Mutavdžić Pavlović, *J. Sep. Sci.* **2015**, *38*, 1232.
[7] A. C. Lee, G. M. Crippen, *J. Chem. Inf. Modell.* **2009**, *49*, 2013.
[8] P. G. Seybold, G. C. Shields, *WIREs Comput. Mol. Sci.* **2015**, *5*, 290.
[9] S. Kheirjou, A. Abedin, A. Fattahi, *Comput. Theor. Chem.* **2012**, *1000*, 1.
[10] C. Lim, D. Bashford, M. Karplus, *J. Phys. Chem.* **1991**, *95*, 5610.
[11] Y. Zeng, H. Qian, X. Chen, Z. Li, S. Yu, X. Xiao, *Chin. J. Chem.* **2010**, *28*, 727.
[12] J. Hong, *Aust. J. Chem.* **2014**, *67*, 1441.
[13] S. L. Dixon, P. C. Jurs, *J. Comput. Chem.* **1993**, *14*, 1460.
[14] J. Zhang, T. Kleinöder, J. Gasteiger, *J. Chem. Inf. Modell.* **2006**, *46*, 2256.
[15] J. Ghasemi, S. Saaidpour, S. D. Brown, *J. Mol. Struct. Theochem.* **2007**, *805*, 27.
[16] M. Namazian, H. Heidary, *J. Mol. Struct. Theochem.* **2003**, *620*, 257.
[17] I. Alkorta, M. Z. Griffiths, P. L. A. Popelier, *J. Phys. Org. Chem.* **2013**, *26*, 791.
[18] P. G. Seybold, In *Advance Quantum Chemistry*, Vol. *64*; R. S. John, J. B. Erkki, Eds.; Academic Press, New York, American; **2012**; pp. 83–104.
[19] M. S. Bodnarchuk, D. M. Heyes, D. Dini, S. Chahine, S. Edwards, *J. Chem. Theor. Comput.* **2014**, *10*, 2537.
[20] R. Svobodová Vařeková, S. Geidl, C. M. Ionescu, O. Skřehota, M. Kudera, D. Sehnal, T. Bouchal, R. Abagyan, H. J. Huber, J. Koča, *J. Chem. Inf. Modell.* **2011**, *51*, 1795.
[21] C. Y. Yuan, S. S. Hu, *Acta Chim. Sin.* **1986**, 590.
[22] C. Y. Yuan, S. S. Li, W. X. Hu, H. Z. Fen, *Heteroat. Chem.* **1993**, *4*, 23.
[23] C. Y. Yuan, Z. C. Sheng, W. Z. Ye, *At. Energy Sci. Technol.* **1965**, 870.
[24] C. Y. Yuan, *J. Chin. Rare Earth Soc.* **1983**, *1*, 13.
[25] C. Y. Yuan, H. Y. Long, E. X. Ma, W. H. Cheng, X. M. Yan, *J. Chin. Rare Earth Soc.* **1985**, *3*, 13.
[26] C. Y. Yuan, S. S. Hu, *Sci. Chin. Ser. B* **1987**, 27.
[27] D. Z. Shen, J. Y. Yan, C. Y. Yuan, *J. Chin. Rare Earth Soc.* **1990**, *10*, 293.
[28] Z. F. Gao, R. Y. Xie, C. Y. Yuan, *Chin. J. Org. Chem.* **1994**, *14*, 200.
[29] M. Cossi, N. Rega, G. Scalmani, V. Barone, *J. Comput. Chem.* **2003**, *24*, 669.
[30] Y. Zhao, D. G. Truhlar, *Theor. Chem. Acc.* **2008**, *120*, 215.
[31] R. A. Kendall, T. H. Dunning, R. J. Harrison, *J. Chem. Phys.* **1992**, *96*, 6796.
[32] J. A. Bohmann, F. Weinhold, T. C. Farrar, *J. Chem. Phys.* **1997**, *107*, 1173.
[33] C. M. Breneman, K. B. Wiberg, *J. Comput. Chem.* **1990**, *11*, 361.
[34] Frisch, M. e. et. al, Gaussian 09, Revision D.01. Gaussian, Inc. Wallingford, CT: 2013.
[35] D. Taeger, S. Kuhnt, In *Statistical Hypothesis Testing with SAS and R*; Wiley, New York, American; **2014**; pp. 3–16.
[36] F. J. Fabozzi, S. M. Focardi, S. T. Rachev, B. G. Arshanapalli, In *The Basics of Financial Econometrics*; Wiley, New York, American; **2014**; pp. 41–80.
[37] A. E. Reed, L. A. Curtiss, F. Weinhold, *Chem. Rev.* **1988**, *88*, 899.
[38] J. Gasteiger, M. Marsili, *Tetrahedron* **1980**, *36*, 3219.
[39] H. Yu, K. Ralph, E. Ralf-Uwe, S. Gerrit, *J. Chem. Inf. Modell.* **2011**, *51*, 2336.
[40] G. Galstyan, E. W. Knapp, *J. Comput. Chem.* **2015**, *36*, 69.
[41] G. Schüürmann, R. U. Ebert, J. Chen, B. Wang, R. Kühne, *J. Chem. Inf. Model* **2008**, *48*, 2140.

Received: 30 September 2015
Revised: 12 January 2016
Accepted: 5 March 2016
Published online on 24 May 2016