COEFFICIENT PARTITION PREDICTION OF SATURATED MONOCARBOXYLIC ACIDS USING THE MOLECULAR DESCRIPTORS

*FAHIMEH MOHAMMAEI, ESMAT MOHAMMADINASAB**

Department of Chemistry, Arak Branch, Islamic Azad University, Arak, Iran[1]

ABSTRACT

Carboxylic acids have clearly been absent from the quantitative structure-property relationship literature. The studies of the quantitative structure–property relationships (QSPR) involve various chemometric methods in which the physico-chemical behavior of a compound is correlated with its structure represented by the structural indices. For example, QSPR methods are applied for the prediction of octanol-water partition coefficient of an organic compound. In this study, the relationship between the octanol/water coefficient partition and molecular descriptors was investigated. Also, the multiple linear-regression method based on QSPR methodology was applied to predict the Log P of saturated mono-carboxylic acids C_1-C_{22}. On the other hand, the relation [Log P = - 0.426 (Platt) + 0.190 (V/ $A^{\circ 3}$) - 0.155 (Max.P.A/ $A^{\circ 2}$) - 1.914 (X) - 1.576 ; N = 22, R^2 = 0.995 , F = 917.005, DW=1.391] was generated for selected mono-carboxylic acids. The results of study indicated that the Platt, Randic, Volume and Maximum-Projection-Area descriptors have an important role in predicting the octanol/water coefficient partition of saturated monocarboxylic acids (C_1- C_{22}).

Keywords: Octanol-Water Coefficient Partition, MLR Method, Saturated Monocarboxilic Acid.

INTRODUCTION

Carboxylic acids are pervasive in nature. They are frequently used to generate polymers, pharmaceuticals, solvents, esters and polyesters. Monocarboxylic acids are organic compounds that contain a carboxyl group. The combination of hydroxyl and carbonyl group form the functional group carboxyl. Carbonyl group of carboxylic acids is considerably different from its aldehydes or ketones sibling. However, when the hydroxyl group of a carboxylic acid is compared with that of alcohols or phenols, the same result can be achieved.

Carboxylic acids are polar molecules. Although 1-5 carbon carboxylic acids are soluble in water, the higher carbon carboxylic acids due to the increasing hydrophobic nature of the alkyl chain are rather soluble in less-polar solvents such as ethers and alcohols [1-3].

Log P (o/w) is considered as an essential property in new or problematic chemicals studies. It is mainly expressed as the n-octanol/water partition coefficient Log P(o/w) regarding the hydrophobicity of the compound. In recent years, it is known that the octanol/water partition coefficient has become a key parameter in the environmental science of organic compounds research. It should be noted that since the human body is made from water and lipids, therefore, Octanol/water partition coefficient is very important factor in biological, toxicological and environmental area [4-6]. As you know, the experimental methods of Log P, e. g. shake flask, reversed phase thin layer chromatography and high performance liquid chromatography (HPLC) are not always available. On the other hand, although there are many various software to calculate Log P of chemical compounds, almost all programs did not open the scheme and factors. Therefore, it seems that the use of statistic computational methods is essential.

Quantitative structure-activity relationships (QSAR) and quantitative structure-property relationships (QSPR) involve the statistical methods by which biological activities or physicochemical properties are related with structural elements [7-11]. We have used the multiple linear regression (MLR) technique for obtaining an appropriate QSPR model. Multiple linear regression (MLR) technique which is based on the least-squares procedures are very often used to estimate the coefficients involved in the model equation [12-16]. In the present research, we propose QSPR model to predict Log P of saturated monocarboxylic acids by describing the chemical structure with the aid of molecular descriptors. Molecular descriptors such as topological indices, geometric indices, etc. are numerical representations of the chemical structure computed on the basis of the molecular graph [17-18]. The values of the experimental Log P of saturated monocarboxylic acids are often scarce, and hence, molecular descriptors provide powerful tools for modeling and extrapolating experimental data.

MATERIALS AND METHODS

In this study, First, the structure and the values of experimental Log P of 22 different types of saturated carboxylic acids (C_1-C_{22}) were taken from the literature (Octanol-Water Partition Coefficients of Simple Organic Compounds, J. Phys. Chem. Ref. Data) [19]. Second, the used descriptors were obtained directly from the chemical structure and the values of topological descriptors , e. g. Platt (Platt), Balaban (J), Randic (χ), Harary (H), Wiener (W), Wiener Polarity (WP), Szeged (Sz) and HyperWiener (WW) indices [20-29] for 22 different types of saturated monocarboxylic acids (C_1-C_{22}) were calculated using the web chemicalize program and also the values of geometric descriptors, e. g. the minimal projection area (Min.P.A/$A^{\circ 2}$), the maximal projection area (Max.P.A/$A^{\circ 2}$), the minimal z length (Min.z.L/A°), the maximal z length (Max.z. L/A°), the van der Waals volume (V/$A^{\circ 3}$), the dreiding energy (E/kcalmol^{-1}) for 22 compounds of mentioned training set were taken from book and web book [30]. Thirdly, the relationship between experimental Log P with 14 different types of descriptors for mentioned saturated carboxylic acids using excel software was investigated and relevant equations were extracted. Fourth, the Log P estimation of used carboxylic acids was performed using SPSS software version 16 with multiple linear regression method and backward procedure. According to the key determining factors of this method, e. g. correlation coefficient (R), square correlation coefficient (R^2), adjust square correlation coefficient (R^2_{Adjust}), Fisher statistics (F), Durbin Watson (DW),…. the best topological indices were determined to predict Log P of used molecules.

RESULTS AND DISCUSSION

The values of experimental Log P of 22 different types of saturated monocarboxylic acids (C_1-C_{22}) were shown in Table 1.

The values of topological and geometric indices of all the mentioned compounds used were taken from the book and web book [24]. The relationship between experimental Log P and 14 different types of descriptors for saturated monocarboxylic acids mentioned was investigated using Excel software. (see equations 1-14)

Log P = 0.2708 F - 2.4323	R^2=0.9349	(1)
Log P = 1.1049 X - 2.6345	R^2=0.9690	(2)
Log P = 1.9415 J - 2.9576	R^2=0.0857	(3)
Log P = 0.1824 H - 1.2508	R^2=0.9620	(4)
Log P = 0.0057 W + 0.9315	R^2=0.8374	(5)
Log P = 0.0009 WW + 1.3382	R^2=0.7592	(6)
Log P = 0.5146 WP - 1.4305	R^2=0.8763	(7)
Log P = 0.0057 Sz - 0.9315	R^2=0.8374	(8)
Log P = 0.2842 DE - 1.0465	R^2=0.9628	(9)
Log P = 0.0323 V - 2.4511	R^2=0.9742	(10)
Log P = 0.2667 Min.P.A - 4.5875	R^2=0.6521	(11)

e-mail:e-mohammadinasab@iau-arak.ac.ir

Log *P* = 0.4206 Min.Z.L - 2.7585	R^2=0.9459	(12)
Log *P* = 0.0954 Max.P.A - 2.6519	R^2=0.9744	(13)
Log *P* = 1.5061 Max Z L - 6.1406	R^2=0.4638	(14)

According to equations (1-14) and their square correlation coefficients (R^2), it can be found that there is a significant linear correlation between Log *P* and some descriptors of this class of carboxylic acids. The following rank can be shown among Log *P* and descriptors:

Max.P.A > Volume > Randic > Deriding Energy > Harary > Min.Z.L > Platt. In equations (3, 5, 6, 7, 8, 11, 14), it can be seen that there is not a strong linear relationship between Max.Z.L, W, Min.P.A, Sz, WW, J, WP indices with Log *P*. In the next step, a multiple linear regression using seven independent variables and Log *P* as a dependent variable was made. Whether or not the regression model explains a statistically significant proportion of data was ascertained through the *ANOVA Table* of output based on the MLR model in terms of the relationship between coefficient partition and effective molecular indices. Then, different models were examined and the best model was defined using correlation coefficient (Pearson's r) and Fisher's coefficient and the associated significance values (Table.2).

TABLE 1. The experimental Log *P* values of the saturated carboxylic acids (C_1-C_{22}) training set.

Carboxylic Acid	Formula	Log P_{ex}	Carboxylic Acid	Formula	Log P_{ex}
Formic acid	$C_1H_2O_2$	-0.54	Octanoic acid	$C_8H_{16}O_2$	3.05
Acetic acid	$C_2H_4O_2$	-0.31	2-Ethylhexanoic acid	$C_8H_{16}O_2$	2.64
Hydroxyacetic acid	$C_2H_4O_3$	-1.11	2-Propylpentanoic acid	$C_8H_{16}O_2$	2.75
Propanoic Acid	$C_3H_6O_2$	0.25	2-propylhexanoic acid	$C_9H_{18}O_2$	3.01
2-Hydroxypropanoic acid	$C_3H_6O_3$	-0.62	Decanoic acid	$C_{10}H_{20}O_2$	4.09
Butanoic acid	$C_4H_80_2$	0.79	2-Propylheptanoic acid	$C_{10}H_{20}O_2$	3.2
2-Hydroxybutanoic acid	$C_4H_80_3$	-0.36	Dodecanoic acid	$C_{12}H_{24}O_2$	4.2
Pentanoic acid	$C_5H_{10}O_2$	1.39	Tetradecanoic acid	$C_{14}H_{28}O_2$	6.11
Hexanoic acid	$C_6H_{12}\ O_2$	1.88	Hexadecanoic acid	$C_{16}H_{32}O_2$	7.17
2-Ethylbutanoic acid	$C_6H_{12}\ O_2$	1.68	Octadecanoic acid	$C_{18}H_{36}O_2$	8.23
2-Methylpentanoic acid	$C_6H_{12}\ O_2$	1.8	Eicosanoic acid	$C_{20}H_{40}O_2$	9.29

TABLE 2. Efficient output Paremeters of MLR Model in 4 models.

Model Number	Predictors	correlation coefficient (R)	correlation coefficient Square (R^2)	correlation coefficient Square Adjust (R^2_{Adjust})	STD. Error of the Estimate (σ)	Fisher Coefficient (F)	Mean Square (MS)	Significant (Sign)
1	Min.Z.L, H, DE Max.P.A, X, F, V,	0.998	0.996	0.994	0.21788	535.434	25.419	0.000
2	Min.Z.L, H, F, V Max.P.A, X	0.998	0.996	0.995	0.21135	663.834	29.654	0.000
3	Min.Z.L, Max.P.A, X, F, V	0.998	0.996	0.995	0.21256	787.364	35.574	0.000
4	Max.P.A, X, F, V	0.998	0.995	0.994	0.22015	917.005	44.443	0.000

To estimate the Log *P*, four models were used with sig =0.000, F: 535.434 < 663.834 < 787.364 <917.005, σ: 0.21788 > 0.21135 >0.21256 > 0.22015, respectively. Finally, the best model with R= 0.998, R^2 = 0.995, R^2_{Adjust} = 0.994, F = 917.005, σ =0.22015, MS=44.443, DW= 1.391 for estimating Log *P* was selected.

The significance is a coefficient which has been used in the statistical method. The more the significance level equal to zero, the lowest the significance level and the more meaningful the linear model will be. Therefore, a smaller significance level lead to a higher Fisher coefficient. As you know if the standard deviation of a set of data is close to zero, it indicates that the data have low dispersion and are close to the average, therefore, the values of standard deviation: 0.22015 in model 4 will be acceptable. One of the methods to examin autocorrelation in the residuals from a statistical regression analysis is Durbin Watson (DW) statistic. The statistical coefficient of Durbin Watson test is limited between 0 and 4. The value of statistical coefficient equal to 1.391 in this statistical analysis indicates there is no caution using the proposed models. Finally, model 4, with balance between the highest the correlation coefficient (R=0.998), the square correlation coefficient (R^2=0.995), the adjust square correlation coefficient (R^2_{Adjust}=0.994), Fisher coefficient (F=917.005), standard Error of Estimate (0.22015) with significance at the 0.000 level and the lowest number of descriptors was opted for further analysis, as reported in MLR Equation 15:

$$\text{Log } P = -\ 0.426\ (\text{ Platt }) + 0.190\ (\text{ V/ } A^{\circ 3}\ ) - 0.155\ (\text{ Max.P.A/ } A^{\circ 2}\ ) -1.914\ (X) -1.576$$

This equation has four common descriptors: Platt index, V/ $A^{\circ 3}$, Max.P.A/ $A^{\circ 2}$ and X index with high calibration statistics and prediction power. The predicted Log *P* of this equation is shown in Table 3. It is worth mentioning that there are many partial differences between the experimental and predicted Log *P* of the model. The residuals of Log *P* are depicted in Table 3. This table also indicates how the model is reliable in any one of the molecules.

TABLE 3. The predicted Log *P*, the Residule values,of the saturated monocarboxylic acids (C_1-C_{22}) training set

Carboxylic Acid	Log P_{Pred}	Residual	Carboxylic Acid	Log P_{Pred}	Residual
Formic acid	-0.38685	-0.15315	Octanoic acid	2.904783	0.145217
Acetic acid	-0.31755	0.007546	2-Ethylhexanoic acid	2.654286	-0.01429
Hydroxyacetic acid	-1.06837	-0.04163	2-Propylpentanoic acid	2.700682	0.049318
Propanoic Acid	0.146445	0.103555	2-propylhexanoic acid	3.120591	-0.11059
2-Hydroxypropanoic acid	-0.63462	0.014622	Decanoic acid	3.924567	0.165433
Butanoic acid	0.763115	0.026885	2-Propylheptanoic acid	3.593787	-0.39379
2-Hydroxybutanoic acid	-0.44274	0.082738	Dodecanoic acid	4.871448	-0.67145
Pentanoic acid	1.254652	0.135348	Tetradecanoic acid	6.066341	0.043659
Hexanoic acid	1.777962	0.102038	Hexadecanoic acid	7.11828	0.05172
2-Ethylbutanoic acid	1.601481	0.078519	Octadecanoic acid	8.196993	0.033007
2-Methylpentanoic acid	1.700657	0.099343	Eicosanoic acid	9.044061	0.245939

Figure 1 shows the strong linear correlation between the experimental and the predicted Log *P* values obtained using equation 15.

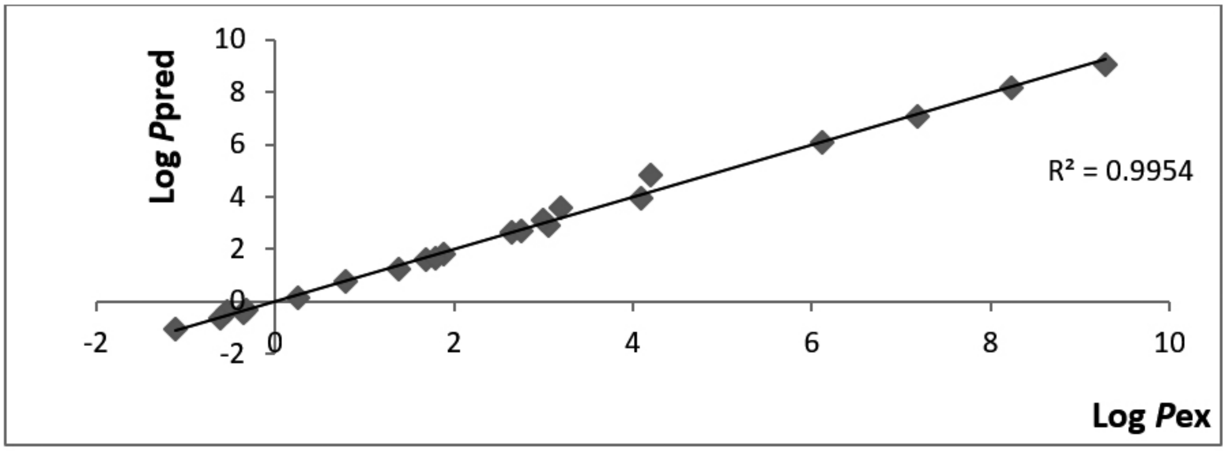

Fig.1. The diagram of the Experimental Log *P* versus the Predicted Log *P*

Comparison between the experimental and the perdicted Log *P* values by MLR model indicate that the equation 15 might be used successfully to predict the Log *P* of studied carboxylic acids (see figure .2).

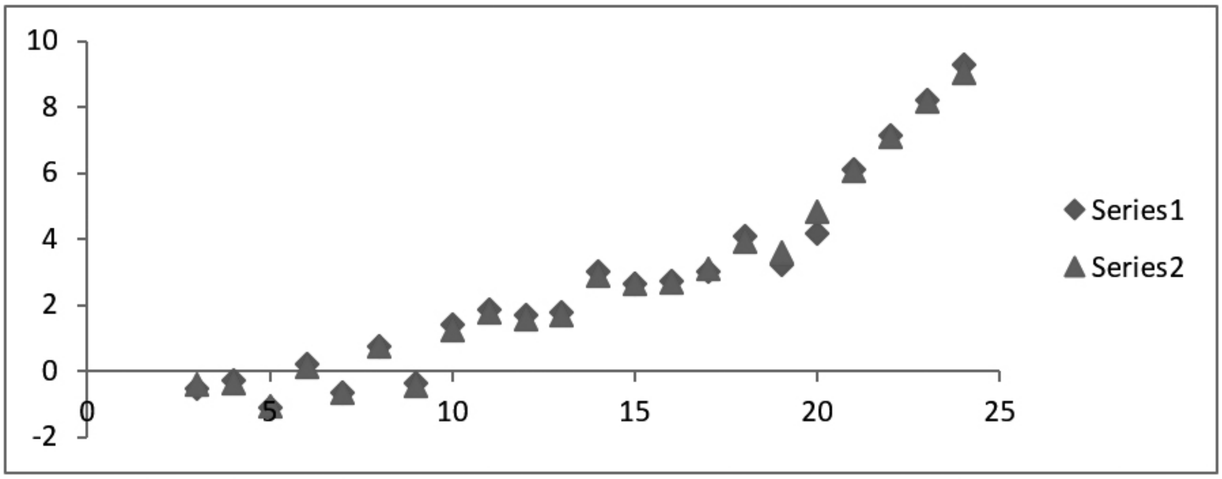

Fig. 2. Comparison between the experimental and perdicted Log *P*

The residual values show a fairly random pattern (see Figure 3). This random pattern indicates that a linear model provides a decent fit to the data, therefore, the result is very satisfactory.

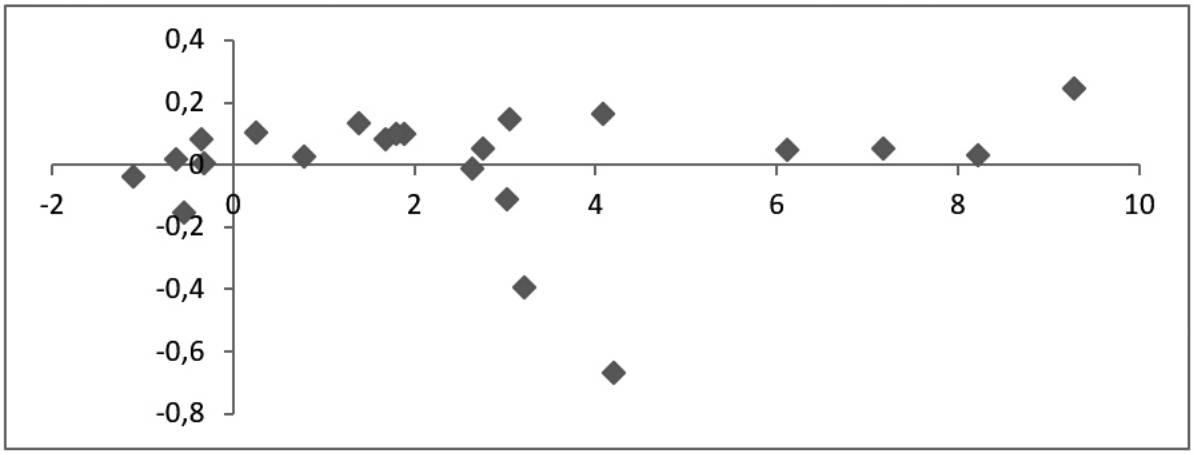

Fig. 3. The plot of experimental values against residuals of Log *P*

CONCLUSION

In present work, least square as a linear regression modeling method was used to construct a relation between coefficient partition of saturated monocarboxylic acids and their topological and geometric descriptors. The obtained results demonstrated that there are strong relationships between some molecular descriptors of this class carboxylic acids. According to the result, it can be seen that there is good correlation between the Log *P* discussed in this report with values of the Platt, V/A$^{°3}$, Max.P.A/A$^{°2}$ and X indices of mentioned molecules. The results of experimental for used molecules were compared with the results of multiple linear regression calculations and was represented that Platt, Randic topological indices and Volume, Maximum Projection Area are the best descriptors for predicting the values of Log *P* of saturated monocarboxylic acids. According to the literature search, to the best of our knowledge this is the first report on saturated monocarboxilic acids which their Log *P* in contrast to mentioned descriptors is prediced by SPSS software and linear multiple regression model.

REFERENCES

1. W. Riemenschneider, Carboxylic Acids, Aliphatic, Ullmann's Encyclopedia of Industrial Chemistry. Weinheim: Wiley-VCH, 2002.
2. J. March, Advanced organic chemistry—reactions, mechanisms and structure. 4th edn, Wiley Interscience, New York, 8, 1992.
3. J. R. Seward, T. W. Schultz, QSAR analyses of the toxicity of aliphatic carboxylic acids and salts to Tetrahymena pyriformis. SAR QSAR Environ Res. 10, 557, (1999)
4. S. P. Torres, J. Sales, M. Rosés, C. Ràfols, E. Bosch , Journal of Chromatography A., 1217 (18), 3026, (2010)
5. R. Smith and C. Tanford, Proc. Nat. Acad. Sci. USA, 70, (2), 289, (1973)
6. F. Spafiu, A. Mischie, P. Ionita, A. Beteringhe, T. Constantinescu and A. T. Balabanb,. Arkivoc, General Papers, (x) 174, (2009)
7. J. Devillers, A. T. Balaban, Topological indices and related descriptors in QSAR and QSPR. Gordon and Breach Science, Netherlands, 1999.
8. E. Mohammadinasab, and M. Goodarzi, J. Fullerenes, Nanotubes, and Carbon Nanostructures., 19, 550, (2011)
9. M. Goodarzi, E. Mohammadinasab, Fullerenes, Nanotubes and Carbon Nanostructures. 21, 2 ,102 (2013)
10. M. Zanoozi and Z. Bayat, Pelagia Research Library, Der Chemica Sinica. 2(6), 288. (2011)
11. Z. Bayat, J. Movaffagh, Russian Journal of Physical Chemistry A. **84**, (13), 2293, (2010)
12. G.W. Snedecor, W. G. Cochran, Statistical methods, Oxford and IBH, New Delhi, 1967.
13. S. Wold, M. Sjöström, L. Eriksson. Chemometrics and Intelligent Laboratory Systems, 58, 109, (2001).
14. A. Agresti, An introduction to categorical data analysis, Wiley, Hoboken, 1996.
15. J. G. Topliss, R. J. Costello, J. Med Chem. 15, 1066, (1972).
16. I. Gutman and O.E. Polansky, Mathematical Concepts in Organic Chemistry, Springer-Verlag, Berlin, 1986.
17. R. Todeschini, V. Consonni, Handbook of Molecular Descriptors, Methods and Principles in Medicinal Chemistry, WILEY-VCH Verlag GmbH, 2008.
18. I. Gutman, B. Furtula(eds), novel molecular structure descriptors-theory and applications I and II, University of Kragujevac and Faculty of Science Kragujevac, 2010.
19. J. Sangster, Octanol-Water Partition Coefficients of Simple Organic Compounds, J. Phys. Chem. Ref. Data. 18, 3, (1989)
20. H. Wiener, J. Am. Chem. Soc. 69, 17, (1947)
21. G. Cash, S. Klavzar, Marko Petkovsek, J. Chem. Inf. Comput. Sci. 42, 571, (2002)
22. X. Li, Y. Shi, MATCH Communications in Mathematical and in Computer Chemistry. 59 (1) 127, (2008).
23. M. Randic, MATCH Commun. Math. Comput. Chem. 7, 5, (1979).
24. B. Liu, I. Gutman, MATCH Communications in Mathematical and in Computer Chemistry., 58 (1), 147, (2007)
25. M. Randic, J. Am. Chem. 97(23), 6609, (1975)
26. A. T. Balaban and T. S. Balaban, Math. Chem. 8, 383, (1991)
27. K. C. Das, B. Zhou and N. TrinajstiQ, J. Math. Chem. 1369, (2009)
28. I. Gutman, Graph Theory Notes of New York. 27, 9, 1994.
29. P. V. Khadikar, N. Deshpande, P. P. Dobrynin, J. Chem. Inf. Compt. Sci.35, 547 (1995).
30. Web search engine developed by ChemAxon; software available at, http:// WWW. Chemicalize. Org.