International Journal of Chemical Studies

P-ISSN: 2349-8528
E-ISSN: 2321-4902
Impact Factor (RJIF): 6.85
www.chemijournal.com
IJCS 2025; 13(6): 139-144

Received: 06-11-2025
Accepted: 09-12-2025

Assouma Dagri Cyrille
Département de Mathématiques Physiques Chimie, UFR des Sciences Biologiques, Université Peleforo GON COULIBALY de Korhogo, BP 1328 Korhogo, Côte d'Ivoire

Tognonvi Tohoué Monique
Département de Mathématiques Physiques Chimie, UFR des Sciences Biologiques, Université Peleforo GON COULIBALY de Korhogo, BP 1328 Korhogo, Côte d'Ivoire

Seka Yapoga Jean
Institut de Gestion Agropastorale, Université Peleforo GON COULIBALY de Korhogo, BP 1328 Korhogo, Côte d'Ivoire

Yacouba Zoungranan
Département de Mathématiques Physiques Chimie, UFR des Sciences Biologiques, Université Peleforo GON COULIBALY de Korhogo, BP 1328 Korhogo, Côte d'Ivoire

Souly Félix
Département de Mathématiques Physiques Chimie, UFR des Sciences Biologiques, Université Peleforo GON COULIBALY de Korhogo, BP 1328 Korhogo, Côte d'Ivoire

Corresponding Author:
Assouma Dagri Cyrille
Département de Mathématiques Physiques Chimie, UFR des Sciences Biologiques, Université Peleforo GON COULIBALY de Korhogo, BP 1328 Korhogo, Côte d'Ivoire

Development of QSPR models for the prediction of pKa values of benzoic acid and three of its derivatives based on quantum descriptors determined by DFT

Assouma Dagri Cyrille, Tognonvi Tohoué Monique, Seka Yapoga Jean, Yacouba Zoungranan and Souly Félix

DOI: https://www.doi.org/10.22271/chemi.2025.v13.i6b.12633

Abstract
The dissociation constant (pK_a) of acids is one of the most widely used physicochemical parameters, both in industry and in fundamental research. It allows to understand numerous chemical and biological phenomena such as polymer formation, adsorption, metabolism, distribution and biological transport. Several methods can be used to determine pK_a values. This study was carried out using QSPR (Quantitative Structure-Property Relationship) modelling. Each model was developed from a series of sets of three quantum descriptors, calculated in the gaseous state at the B3LYP level with the LANL2DZ basis set, using multiple linear regression. The descriptors are the energies of the E_{HOMO} (energy of the highest occupied molecular orbital) and E_{LUMO} (energy of the lowest unoccupied molecular orbital) orbitals, the dipole moment, and the energy gap ΔE (energy difference between E_{HOMO} and E_{LUMO}). Indeed, based on experimental pK_a values for benzoic, fluoro-benzoic, chloro-benzoic and bromo-benzoic acids and their isomers, four reliable, efficient and robust QSPR models were developed to theoretically predict the pK_a values of these acids in aqueous solution. In addition to their high predictive power for theoretical pK_a values, these models offer the possibility of explaining the chemical phenomena that accompany the dissociation reaction of these acids, in which the solvent plays an important role.

Keywords: Dissociation constant, quantum descriptors, QSPR modelling

1. Introduction
Carboxylic acids and their derivatives are used in the composition of several commonly used products such as medicines, food additives, cosmetics and plastics, as well as in organic and biological synthesis [1]. The physico-chemical property that characterises the dissociation state of these acids in solution is the dissociation constant, denoted pK_a. It plays a fundamental role in solvent extraction and purification processes [2, 3], in acid-base titrations [4], in complex formation [5] and even in ion transport [6]. It also allows to evaluate chemical and biological phenomena such as absorption, distribution, metabolism, elimination, toxicity, and solubility of compounds in biological cells [4]. Numerous experimental methods are used to determine pK_a values, including potentiometry [7], spectrophotometry [8], liquid chromatography [9], conductimetry [10], and capillary electrophoresis [11]. Not only these experimental determinations of pKa values are complex and onerous, but the values obtained are often associated with high uncertainties. These uncertainties in the pK_a values are also affected by the experimental method used, temperature control, solvent composition and the stability of the chemicals used [12]. Considering the above, research methods are oriented towards predicting the physicochemical properties of new molecules. Among these methods is QSPR modelling (Quantitative Structure-Property Relationship) [13, 14]. In QSPR models, a mathematical relationship is established between the physicochemical properties and the molecular structure of chemical or biochemical compounds. Once established, these models will allow to develop of new theories or to explain observed phenomena and to predict the physicochemical properties of compounds for which experimental data are not available. They are based on the characterisation of molecules by a set of descriptors measured or calculated from molecular structures. Few studies on predicting the pK_a values of benzoic acid and

its derivatives exist in the scientific literature. This study will consist of to develop QSPR models for the prediction of acidity constant values of benzoic acid and three of its derivatives (fluorobenzoic acid and isomers, chlorobenzoic acid and isomers, bromobenzoic acid and isomers) from quantum descriptors determined by Density Functional Theory (DFT). DFT is a quantum chemistry calculation method that uses electron density as a variable. The general aim of this work is to develop four QSPR models by establishing multilinear correlations between experimental pK_a values and three quantum descriptors.

2. Materials and methods

2.1 Materials

The material used in this study includes a Lenovo Thinkpad T 490 laptop with a 512 GB SSD hard drive and 16 GB DDR4 RAM.

Gaussian 09 software, specialised in optimising molecular structures and obtaining quantum molecular descriptors, was used.

The GaussView 6 module integrated into Gaussian 09 allows to construct 3D structures of molecules.

Microsoft Excel 2016 was used to develop the models.

The molecules studied are benzoic acid, fluoro-benzoic acid, chloro-benzoic acid, bromo-benzoic acid, and their isomers.

2.2. QSPR modelling method

In QSPR modelling, similarities between molecules are sought in a large database of known properties of existing molecules. Its purpose is to guide the synthesis of new molecules without having to produce them. Above all, it enables the analysis of entire families of chemical compounds whose properties are known [15].

2.2.1 Principle of QSPR methods

The principle of QSPR methods is to establish a mathematical relationship that quantitatively links descriptors from molecular structures with a physicochemical property for a series of similar chemical compounds using statistical data analysis methods. The general formula for QSPR models is:

$$Property = f(Descriptors) \quad (1)$$

The aim of QSPR models is therefore to analyse data derived from molecular structures in order to identify the factors that determine the measured property. Various modelling tools can be used for this purpose. The most used are simple linear regression (SLR) and multiple linear regression (MLR) [16, 17]. There are also partial least squares regression (PLS) [18], decision trees [19], neural networks [20] and genetic algorithms [21]. In this work, multiple linear regression was used. In practice, the model begins with the collection of experimental databases (pK_a values), which must be numerous. Then, we look for a series of quantum descriptors characteristic of the structure of the compounds in the database in order to link them to the experimental property studied. Finally, data analysis tools are used to help choose the appropriate descriptors and develop the model itself. When established and validated on a validation set, the model can then be used to predict the property of new molecules for which experimental values are not available or for molecules that have not yet been synthesised. The most widely used QSPR model validation tools are:

- The coefficient of determination, denoted R^2. It evaluates the proportion of variance explained by the model [22]. The value of R^2 is between 0 and 1. When it is close to 1 (ideal case), the predicted and experimental values are strongly correlated. However, a low R^2 value indicates that the model has low predictive power and that the descriptors have no effect on the response.
- The mean absolute error (MAE) is another statistical indicator [23]. However, standard deviation (SD) is preferred. These parameters allow the dispersion of the calculated properties relative to the experimental properties to be estimated.
- The Fisher F-index. It measures the degree of statistical significance of the model, i.e. the quality of the choice of parameters.

Methodology for computing quantum descriptors

To determine the descriptors derived from the molecular structures of benzoic acid, fluorobenzoic acid, chlorobenzoic acid, bromobenzoic acid and their isomers, quantum chemistry calculations are performed using a computer. We create a 3D graphical representation of the molecule using Gauss View 6 software while performing pre-optimisation. Then the structure of the molecule is optimised using density functional theory implemented in the Gaussian 09W software [24]. Molecular structure optimisation is performed using the B3LYP hybrid functional [25] with the LANL2DZ pseudo-potential orbital basis. After optimising the structure, it is possible to access descriptors that help to understand the electronic structure of the chemical system. In this work, all descriptors are calculated in the gaseous state. These descriptors are the energy of the highest occupied molecular orbital (E_{HOMO}), the energy of the lowest unoccupied molecular orbital (E_{LUMO}) and the dipole moment (μ).

3. Results

3.1 Analysis of descriptors

The geometries of benzoic acid molecules and their derivatives are optimised using the LANL2DZ basis set at the B3LYP functional level. They are all shown in Figure 1. These optimised structures are used to obtain molecular descriptors from data extracted from calculations performed with Gaussian 09W [24]. These descriptors are the energies of the frontier orbitals (E_{HOMO} and E_{LUMO}) and the dipole moment (μ). We have an additional descriptor derived from the combination of the energies of the frontier orbitals. This is the LUMO-HOMO energy gap ($\Delta E = E_{LUMO} - E_{HOMO}$). The latter is a descriptor indicating the global electronic stability of the molecule. A lower ΔE suggests that the molecule is more polarizable and therefore potentially more reactive. The dipole moment reflects the global electronic distribution of the molecule. E_{HOMO} and E_{LUMO} energies are respectively related to the ionisation energy and electron affinity of the molecule. The values of the acid descriptors studied are grouped together in Table 1. We note that the values of E_{HOMO} and E_{LUMO} energies, which also represent the energies of the electronic states of the most stable geometric structures of molecules, are all negative. In general, the energy values of LUMO orbitals are higher than those of HOMO orbitals. This leads to a positive energy gap ($\Delta E = E_{LUMO} - E_{HOMO}$). Table 1 also contains the experimental pKa values of the acids studied. Some values, notably those for benzoic acid and ortho-halobenzoic acid isomers, are taken from our previous work [26]. However, the values for other acids are taken from the literature [27].

3.2 Development of QSPR models

In establishing QSPR models, molecular descriptors (x_i) are linearly related to a response variable (y) according to the mathematical relationship:

$$y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \cdots \quad (2)$$

The constants α_i are unknowns of the models, also called regression coefficients, and the objective of regression analysis is to estimate them. A series of sets of three descriptors were used to develop a model. We were able to establish four models based on the experimental pKa values of benzoic, fluorobenzoic, chlorobenzoic and bromobenzoic acids and a series of combinations of values for three descriptors from Table 1. The equations of the four QSPR models developed have the general formula:

$$pK_a = AX_1 + BX_2 + CX_3 + D \quad (3)$$

X_1 , X_2 , X_3 are the descriptors contained in table 1. A , B, C et D are unknowns that will be determined by multilinear regression using Microsoft Excel 2016 software installed on a Lenovo Thinkpad T490 computer. The equations of the four QSPR models developed and their validation tools are represented in table 2.

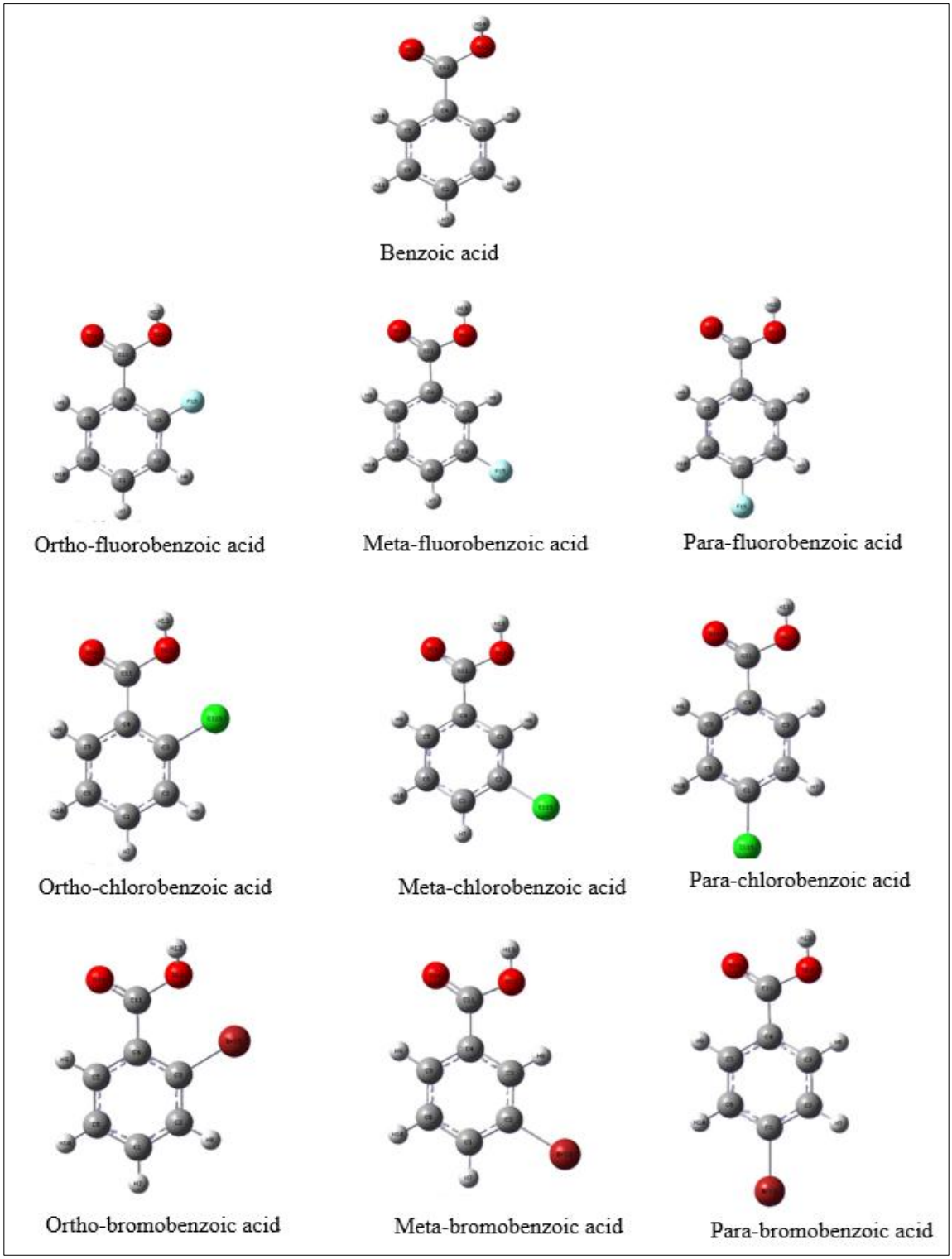


Fig 1: Optimised structures of acid molecules (B3LYP/LANL2DZ)

Table 1: Molecular descriptor values and experimental pKa values of the acids studied [26, 27]

Names of acids	μ (D)	E_{HOMO} (ev)	E_{LUMO} (ev)	ΔE (ev)	pK_a
Benzoic acid	2,5996	-7,3630	-1,8577	5,5052	4,19
Ortho-fluorobenzoic acid	3,0649	-7,4914	-2,1126	5,3788	3,27
Ortho-chlorobenzoic acid	2,8409	-7,3124	-2,1180	5,1943	2,90
Ortho-bromobenzoic acid	2,5766	-7,1011	-2,1074	4,9936	2,85
Metha-fluorobenzoic acid	1,4144	-7,4976	-2,2124	5,2852	2,85
Meta-chlorobenzoic acid	1,4107	-7,3646	-2,2165	5,1481	3,84
Meta-bromobenzoic acid	1,4492	-7,1718	-2,1822	4,9895	3,81
Para-fluorobenzoic acid	1,5470	-7,5703	-2,0737	5,4965	4,15
Para-chlorobenzoic acid	1,5486	-7,4372	-2,1615	5,2757	4
Para-bromobenzoic acid	1,5918	-7,2482	- 2,1479	5,1002	3,96

Table 2: Different models developed

The Models	Model equations	R^2	*SD*	*F*
Model 1 (E_{HOMO}, E_{LUMO}, μ)	$pK_a = -0{,}2649E_{HOMO} + 1{,}3036E_{LUMO} - 0{,}1773\mu + 5{,}1221$	0,998	0,019	5285,62
Model 2 (E_{HOMO}, ΔE, μ)	$pK_a = 1{,}0418E_{HOMO} + 1{,}306\Delta E - 0{,}1779\mu + 5{,}1334$	0,997	0,035	1287,15
Model 3 (E_{LUMO}, ΔE, μ)	$pK_a = 1{,}0393E_{LUMO} + 0{,}2647\Delta E - 0{,}1774\mu + 5{,}1244$	0,997	0,023	2832,16
Model 4 (E_{HOMO}, E_{LUMO}, ΔE)	$pK_a = -363{,}7584E_{HOMO} + 363{,}3255E_{LUMO} - 363{,}3102\Delta E + 1{,}9970$	0,997	0,023	3667,61

4. Discussion

In Table 2, all models show:

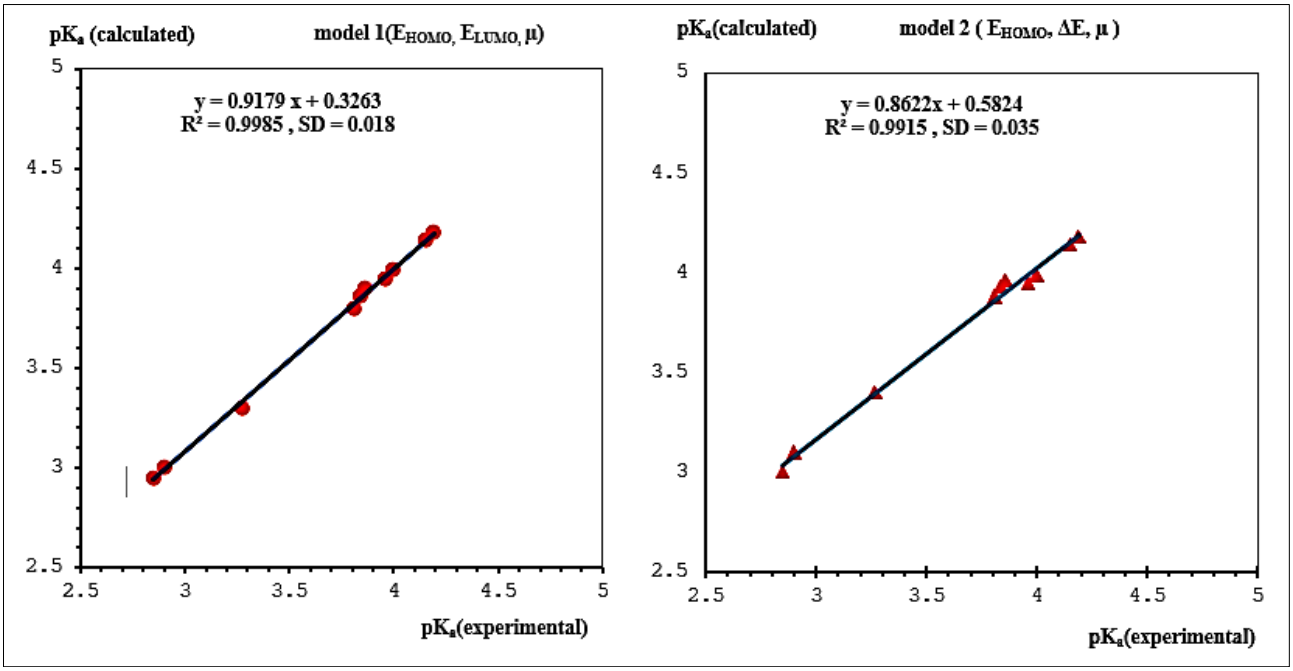
- Correlation coefficients R^2 tending towards 1 ($R^2 > 0{,}99$)
- High Fisher's index values ($F > 100$)
- Low standard deviation values ($SD < 0{,}05$)

In Figure 2, the correlation between the calculated pKa values and the experimental pKa values is also represented for each model.

From the validation tool values, we can say that in all four models, there is a strong correlation between the calculated pK_a values of the ten benzoic acids and each set of three molecular descriptors calculated in the gas phase at the B3LYP/LANL2DZ level. Also, as shown in figure 2, the values of the acidity constants obtained with the different models are identical to the experimental pKa values. In addition, these different models, which show good correlations ($R^2 > 0.99$; $SD < 0.1$) are robust. These QSPR models, with three quantum descriptors, as developed could be used to predict theoretically the values of acid constants for numerous other derivatives of benzoic acids in a reliable manner [28].

In models (1, 2, 3) developed, the dipole moment coefficients are negative. Therefore, an increase in dipole moment values would lead to a decrease in pK_a values. The dipole moment therefore contributes efficiently and at any time to the dissociation of the benzoic acids studied. The importance of the dipole moment in the prediction of pK_a values is due to the fact that the molecules of these acids are highly polarised by the presence of the hydroxyl group (O-H). In contact with the water solvent, these molecules will easily dissociate, liberating the hydrogen atom H [28].

Unlike the dipole moment, the regression coefficients of the LUMO energies are positive (models 1, 3, 4). This indicates that the dissociation of the acids studied is strongly influenced by the values of these LUMO energies. In fact, negative E_{LUMO} values (Table 1) also lead to a decrease in pK_a values. The energies of the LUMO orbitals therefore contribute to the dissociation of benzoic acids.



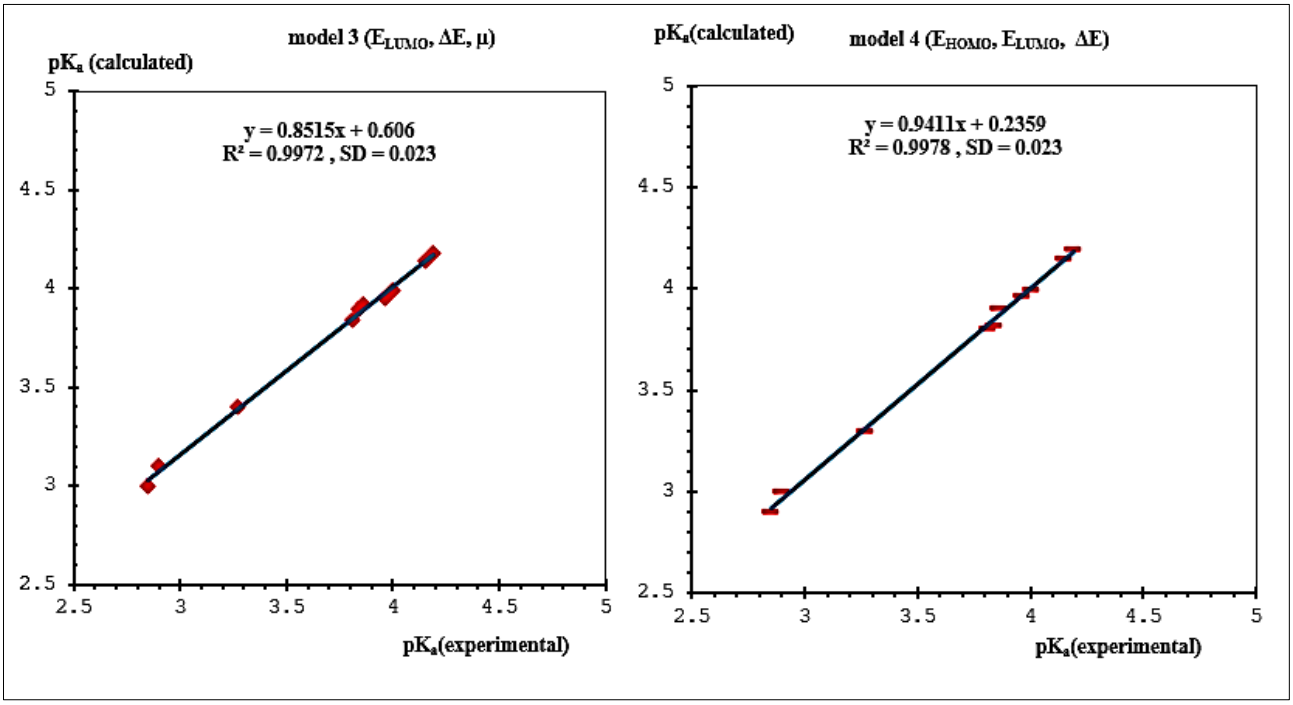


Fig 2: Correlations between calculated pK_a values and experimental pK_a values

In the prediction of pK_a values by model 1, the contribution of LUMO energies to acid dissociation, modulated by the dipole moment, is slightly attenuated by the presence of HOMO energies. Substituting LUMO energies with ΔE in model 1 leads to model 2, in which HOMO energies and dipole moment contribute to acid dissociation. However, ΔE introduced in model 2 has an antagonistic effect on the dissociation. Model 3 is also obtained from Model 1 by substituting ΔE by the descriptor E_{HOMO}. While the LUMO energies and dipole moment facilitate acid dissociation, ΔE slightly opposes it. Models 2 and 3 not only have the same predictive power, but their descriptors contribute by the identical manner to the dissociation of acids.

Finally, in the prediction of pK_a values by model 4, the contribution of LUMO orbital energies to the dissociation of carboxylic acids is compensated by the presence of HOMO orbital energies. Only ΔE contributes to dissociation.

In all the different models developed, the dipole moment and LUMO orbital energies contribute to the dissociation of acids in solutions. On the other hand, depending on the model, descriptors such as HOMO energies and ΔE can contribute to the phenomenon of acid dissociation on the one hand and oppose it on the other. To conclude, we have developed four models that can effectively and reliably predict the theoretical pK_a values of benzoic acid and its derivatives considered in this study. These models were developed based on experimental pK_a values collected in our previous work [26] and in the literature [27] for a series of ten benzoic acids with three quantum descriptors calculated in the gaseous state at the B3LYP level using the LANL2DZ basis set. The mathematical tool used is multiple linear regression.

However, the study has some insufficiencies due to the limited number of benzoic acid molecules. We will then propose to extend the modelling to a numerous number of benzoic acid molecules, while considering the molecules of the aqueous solvent in the descriptor calculations.

5. Conclusion

Benzoic acid and its derivatives are used in the formulation of many widely used products including medicines, food additives, cosmetics and plastics, as well as in various organic and biological synthesis. The dissociation constant of these acids plays a fundamental role in several analytical procedures such as extractions and solvent purifications, acid-base titrations, complex formation and in ion transport. Based on experimental pK_a values for benzoic, fluoro-benzoic, chloro-benzoic and bromo-benzoic acids and their isomers, four reliable, efficient and robust QSPR models were developed to theoretically predict the pK_a values of these acids in aqueous solution. Each model was developed from a series of sets of three quantum descriptors calculated in the gaseous state at the B3LYP level with the LANL2DZ basis set using multiple linear regression. The descriptors are the HOMO and LUMO orbital energies, the dipole moment, and the energy gap ΔE. With, $(R^2 > 0{,}99), F > 100$ et $(SD < 0{,}05)$, all models obtained have a high predictive power for pKa values. They also help explain the chemical phenomena that accompany the dissociation reaction of these acids, in which the solvent plays an important role.

However, the study presents some limitations due to the small number of benzoic acid derivatives considered. We therefore propose to extend the modelling to a larger set of benzoic acid molecules, while explicitly including aqueous solvent molecules in the descriptor calculations. In addition, quantum calculations will be performed using several exchange-correlation functionals (B3PW91, B3LYP, O3LYP, and PBEO) and different orbital basis sets to improve the predictive performance of the QSPR models.

References

1. Gooßen LJ, Rodríguez N, Gooßen K. Carboxylic acids as substrates in homogeneous catalysis. Angewandte Chemie International Edition. 2008;47(17):3100-3120.
2. Vysotsky YB, Kartashynska ES, Vollhardt D, Fainerman VB. Surface pKa of saturated carboxylic acids at the air/water interface: a quantum chemical approach. Journal of Physical Chemistry C. 2020;124(25):13809-13818.
3. Mebarka T. Etude experimentale et modelisation des systemes (eau + acide carboxylique + solvants). 2015.

4. Manallack DT, Prankerd RJ, Yuriev E, Oprea TI, Chalmers DK. The significance of acid/base properties in drug discovery. Chemical Society Reviews. 2013;42(2):485-496.
5. Bas DC, Rogers DM, Jensen JH. Very fast prediction and rationalization of pKa values for protein-ligand complexes. Proteins: Structure, Function, and Bioinformatics. 2008;73(3):765-783.
6. Mihichuk LM, Driver GW, Johnson KE. Brønsted acidity and the medium: fundamentals with a focus on ionic liquids. ChemPhysChem. 2011;12(9):1622-1632.
7. Beltrán JL, Sanli N, Fonrodona G, Barrón D, Özkan G, Barbosa J. Spectrophotometric, potentiometric and chromatographic pKa values of polyphenolic acids in water and acetonitrile-water media. Analytica Chimica Acta. 2003;484(2):253-264.
8. Reijenga J, van Hoof A, van Loon A, Teunissen B. Development of methods for the determination of pKa values. Analytical Chemistry Insights. 2013;8:ACI.S12304. Available from: http://journals.sagepub.com/doi/10.4137/ACI.S12304
9. Erdemgil F, Sanli S, Sanli N, Ozkan G, Barbosa J, Guiteras J, *et al*. Determination of pKa values of some hydroxylated benzoic acids in methanol-water binary mixtures by LC methodology and potentiometry. Talanta. 2007;72(2):489-496.
10. Santos T de AD dos, Costa DO da, Pita SS da R, Semaan FS. Potentiometric and conductimetric studies of chemical equilibria for pyridoxine hydrochloride in aqueous solutions: simple experimental determination of pKa values and analytical applications to pharmaceutical analysis. Eclética Química. 2010;35(4):81-86.
11. Poole SK, Patel S, Dehring K, Workman H, Poole CF. Determination of acid dissociation constants by capillary electrophoresis. Journal of Chromatography A. 2004;1037(1-2):445-454.
12. Sooväli L, Kaljurand I, Kütt A, Leito I. Uncertainty estimation in measurement of pKa values in nonaqueous media: a case study on basicity scale in acetonitrile medium. Analytica Chimica Acta. 2006;566(2):290-303.
13. Jover J, Bosque R, Sales J. QSPR prediction of pKa for benzoic acids in different solvents. QSAR & Combinatorial Science. 2008;27(5):563-581.
14. Kim J, Kim S. State of the art in the application of QSAR techniques for predicting mixture toxicity in environmental risk assessment. SAR and QSAR in Environmental Research. 2015;26(1):41-59.
15. Wadhwa P, Mittal A. Quantitative structure-property relationship (QSPR) modeling applications in formulation development. In: Computer Aided Pharmaceutics and Drug Delivery. Singapore: Springer Nature Singapore; 2022. p. 543-560.
16. Ghasemi JB, Ahmadi S, Ayati M. QSPR modeling of stability constants of the Li-hemispherands complexes using MLR: a theoretical host-guest study. Macroheterocycles. 2010;3(4):234-242.
17. Ghasemi J, Saaidpour S, Brown SD. QSPR study for estimation of acidity constants of some aromatic acids derivatives using multiple linear regression (MLR) analysis. Journal of Molecular Structure: THEOCHEM. 2007;805(1-3):27-32.
18. Geladi P, Kowalski BR. Partial least-squares regression: a tutorial. Analytica Chimica Acta. 1986;185:1-17.
19. Myles AJ, Feudale RN, Liu Y, Woody NA, Brown SD. An introduction to decision tree modeling. Journal of Chemometrics. 2004;18(6):275-285.
20. Gasteiger J, Zupan J. Neural networks in chemistry. Angewandte Chemie International Edition in English. 1993;32(4):503-527.
21. Leardi R. Genetic algorithms in chemometrics and chemistry: a review. Journal of Chemometrics. 2001;15(7):559-569.
22. Mehdi K, Oussama A. Etude des relations quantitatives structure-activité (QSAR) de plusieurs familles de composés sur des données biologiques. 2019.
23. Willmott C, Matsuura K. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Climate Research. 2005;30:79-82.
24. Frisch MJ, Trucks GW, Schlegel HB, Scuseria GE, Robb MA, Cheeseman JR, *et al*. Gaussian 09, Revision B.01. Gaussian, Inc, Wallingford CT. 2009;1-20. Available from: citeulike-article-id:9096580
25. Becke AD. Correlation energy of an inhomogeneous electron gas: a coordinate-space model. Journal of Chemical Physics. 1988;88(2):1053-1062. Available from: https://pubs.aip.org/jcp/article/88/2/1053/222755/Correlation-energy-of-an-inhomogeneous-electron
26. Assouma DC. Détermination et modélisation des constantes thermodynamiques de dissociation de quelques acides carboxyliques dans les mélanges eau-éthanol. Thèse; 2016. 238 p.
27. Perrin DD. Dissociation constants of organic bases in aqueous solution. 1972.
28. Assouma DC, Kouassi K koffi E, Niamien PM, Kouassi HA. Acid dissociation constant pKa of some carboxylic acids: correlation between experimental and theoretical values using density functional theory and QSPR. 2015;9(October):110-118.