Examining the PM6 semiempirical method for pK_a prediction across a wide range of oxyacids

Sierra Rayne [a,*], Kaya Forest [b], and Ken J. Friesen [a]

[a] Department of Chemistry, The University of Winnipeg, Winnipeg, Canada

[b] Department of Chemistry, Okanagan College, British Columbia, Canada

[*] Corresponding author. Tel.: +1 204 786 9265; fax: +1 204 775 2114.

E-mail address: rayne.sierra@gmail.com (S. Rayne).

Abstract

The pK_a estimation ability of the semiempirical PM6 method was evaluated across a broad range of oxyacids and compared to results obtained using the SPARC software program. Compound classes under consideration included acetic acids, alicyclic and aromatic heterocyclic acids, benzoic acids, boronic acids, hydroxamic acids, oximes, peroxides, peroxyacids, phenols, α-saturated acids, α-saturated alcohols, sulfinic acids, α-unsaturated acids, and α-unsaturated alcohols. PM6 accurately predicts the acidity of acetic and benzoic acids and their derivatives, but is less reliable for alicyclic and aromatic heterocyclic acids and phenols. α-Saturated acids are reliably modeled by PM6 except for polyacid derivatives with α-alcohol moieties. α-Saturated alcohols only appear to yield reliable PM6 results where an α-hydroxy or α-alkoxy moiety is absent. Carboxylic acids with simple α-alkene unsaturation are well approximated by PM6 except where alkyne α-unsaturation or α-carboxylation are also present. The PM6 and SPARC methods exhibit approximately equal pK_a prediction performance for the acetic, alicyclic, and benzoic acids. SPARC outperforms PM6 on the peroxides, peroxyacids, phenols, and α-saturated acids and α-saturated alcohols. pK_a values for boron, nitrogen, and sulfur oxyacids do not appear to be reliably estimated by either the PM6 or SPARC methods. The findings will help guide the potential appropriateness of results from the PM6 pK_a estimation method for waste treatment and environmental fate investigations.

Keywords: pK_a prediction; PM6 method; semiempirical; validation; oxyacids; SPARC

1. Introduction

Predicting the acidity constant (pK_a value) of compounds is a critical task in designing waste treatment methods and understanding the environmental fate of both contaminants and natural compounds. Historical approaches to pK_a estimation typically involved linear free energy relationships [1-3], of which the Hammett-type correlations are perhaps the best known [4], and fragment-type, one- and two-dimensional, and topological/connectivity index methods [5,6]. However, these methods often have difficulty dealing with geometrical isomers, new substituent types, and intramolecular hydrogen-bonding effects due to the lack of dependence on three dimensional optimized molecular structures. Over the past few decades, and with the advent of lower cost, easy to use, and widely available computational methods, quantitative structure-property relationships (QSPRs) based on three-dimensional molecular structures have increased in popularity, accuracy, and applicability domain for estimating pK_a values [7-10]. Although ab initio computational approaches may offer the highest likelihood of accurate pK_a estimation [8,11,12], particularly for new compounds without prior class-based training sets for model validation and assessment, the computational cost of ab initio methods, requirement for specialized software knowledge, and lack of a commonly agreed upon basis set types and levels of calculation among the various options precludes widespread application of these computations for rapid screening in applications such environmental assessments.

The recent development of the lower computational cost semiempirical PM6 method with its built-in pK_a estimation function [13], and its application in new versions of the widely available MOPAC software packages (e.g., MOPAC 2007, MOPAC 2009) [14] may allow for accurate and

easily obtained reliable acidity constant estimates. While the PM6 method has been validated for a range of molecular properties [15,16], there have been no studies that assess the accuracy and range of its pK_a estimation function, unlike other studies on related computational programs such as ACD/pK_a, SPARC, COSMOtherm and others [5,17-19]. Although one recent study used the thermodynamic output from the PM6 method to calculate pK_a values for a range of 4-aryl-2,4-dioxobutanoic acids [20], the direct pK_a estimation function in PM6 was not used. In previous work, we have shown that the PM6 method likely underestimates the pK_a values for perfluorinated carboxylic acid contaminants [21] and likely overestimates the pK_a values for perfluorinated sulfonic acid contaminants [22]. Here we investigate the pK_a predictive capacity of the PM6 method across a wide range of carbon and non-carbon oxyacids (including environmental contaminants, natural products, industrial compounds, and medicinally active substances) in the hope of better defining the applicability domain of this computational approach.

2. Materials and methods

Two dimensional molecular structures were drawn in ACD/ChemSketch v. 11.02 (Build 25941, 21 May 2008; Advanced Chemistry Development, Inc., Toronto, ON, Canada), exported into ACD/3D Viewer v. 11.01 (Build 22009, 04 Oct 2007; Advanced Chemistry Development, Inc., Toronto, ON, Canada), converted to three dimensional structures using the 3D Optimization function, and saved as MOPAC Z-matrix files. Geometry optimizations and pK_a estimates in MOPAC 2009 v. 8.345W [14] were conducted using the PM6 method [13] with the following keywords in the input file header: PM6; PKA; BONDS; CHARGE=0; SINGLET; LET;

GNORM=0; CYCLES=10000; GRAPHF. pK_a estimates using SPARC (v. March 2008 release w4.2.1405-s4.2.1408; The University of Georgia, Athens, GA, USA) [23] with its acidity function estimation algorithm [5] and SMILES [24,25] input structures based on the 2D structures in ACD/ChemSketch. Univariate regression analyses were performed using the KyPlot v. 2.0 b. 15 statistical package (Dr. Koichi Yoshioka, Department of Biochemistry and Biophysics, Graduate School of Allied Health Sciences, Tokyo Medical and Dental University, Tokyo, Japan).

3. Results and discussion

A comparison of pK_a values for 68 compounds from the source validation set in the MOPAC 2009 manual [14] and the current work is shown in Figure 2 and given in Electronic Supplementary Material Table S1. Excellent agreement exists between the two datasets, with a slope equal to unity (1.01±0.02 (±std. error)) and y-intercept of zero (-0.02±0.12) within the respective error ranges. A substantial deviation (ΔpK_a >2) between the two datasets was observed for only the following two compounds whose estimated pK_a values as stated in the MOPAC 2009 manual we had difficulty reproducing: citric acid (ΔpK_a=-2.1, $pK_{a,MOPAC\ 2009}$=2.6; **1** in Figure 3) and salicylaldehyde (ΔpK_a=2.6, $pK_{a,MOPAC\ 2009}$=7.5; **2**). The ionization of both compounds can be substantially influenced by intramolecular hydrogen bonding, as has been studied extensively for salicylaldehyde [26]. This process cannot be readily accounted for in the PM6 optimization process, and any such modeled effects will likely be dependent on starting geometries. We also note that the literature pK_a value for salicylaldehyde from ref. [14] is 6.8, whereas the literature value we used from ref. [27] is 8.4.

Any small deviations between the two validation datasets shown in Figure 2 are likely due to possible differences between the starting geometries in both approaches. The starting geometries from ref. [14] are not available, but may be geometries from a centralized crystallographic database. However, the merit in the PM6 method is the capacity to rapidly (and ideally, reliably) predict pK_a values for new compounds for which crystallographic data is not available. Thus, our validation approach approximates a real-world application of the software package, whereby a starting molecular geometry needs to be approximated using a readily available and rapid technique, after which the PM6 method can be applied for pK_a estimation.

Having calibrated our validation approach against this source dataset, we then proceeded to calculate pK_a values for a total of 284 oxyacids from the following compound classes using both the PM6 method and the well-established SPARC program: acetic acids, alicyclic and aromatic heterocyclic acids, benzoic acids, boronic acids, hydroxamic acids, oximes, peroxides, peroxyacids, phenols, α-saturated acids, α-saturated alcohols, sulfinic acids, α-unsaturated acids, and α-unsaturated alcohols. Summary statistics for the validation efforts are provided in Table 1, and comparisons between the predicted and literature pK_a values are shown in Figure 4 and given in Electronic Supplementary Information Table S2. We note that not only are carbon oxyacid classes included in our investigation (acetic acids, alicyclic and aromatic heterocyclic acids, benzoic acids, phenols, α-saturated acids, α-saturated alcohols, α-unsaturated acids, and α-unsaturated alcohols), but so are nitrogen oxyacids (hydroxamic acids and oximes), oxygen oxyacids (peroxides and peroxyacids), sulfur oxyacids (sulfinic acids), and boron oxyacids (boronic acids) whereby the heteroatoms are connected to an organic carbon substituent.

In general, the PM6 method adequately estimates the pK_a values of carbon oxyacids for most subclasses over at least a substantial portion of the pK_a range within a compound class (typically the mid-range of experimental pK_a values within a class). For acetic, alicyclic, and benzoic acids, the PM6 predictive capacity is approximately equally distributed across the experimental pK_a range, with average unsigned errors of 0.40 (n=34), 0.63 (n=10), and 0.46 (n=52), respectively. In contrast, the PM6 reliability decreases considerably with increasing experimental acidity for phenols, α-saturated alcohols and acids, and α-unsaturated acids, with average unsigned errors of 1.06 (n=59), 0.49 (n=56), and 0.84 (n=8), and 0.78 (n=17), respectively. The PM6 method also has difficulty estimating the pK_a values of the two α-unsaturated alcohols we examined (ΔpK_a of 0.7 for propargyl alcohol and -1.5 for allyl alcohol), with a similarly poor performance quality by SPARC (ΔpK_a of 1.4 for propargyl alcohol and -0.3 for allyl alcohol). Weak PM6 predictive ability was found for the large polycyclic phytochemical oleanolic acid (**3**; ΔpK_a=2.4; $pK_{a,exp}$=2.5) and the strained 1,1-cyclopropanedicarboxylic acid (**4**; ΔpK_a=-1.3; $pK_{a,exp}$=1.8). SPARC also was not able to model these compounds effectively, with ΔpK_a values of 2.29 and 1.20 for **3** and **4**, respectively. However, increasing molecular size does not necessarily diminish the PM6 predictive ability within the alicyclic acid class, as the pK_a of the plant hormone gibberellic acid (**5**) is reasonably approximated by both the PM6 (ΔpK_a=0.4; $pK_{a,exp}$=4.0) and SPARC (ΔpK_a=-0.1) methods.

Phenols with pK_a values >7 are generally reliably estimated by PM6. Although some of the largest phenols, such as the more acidic visual acid-base indicators bromophenol blue (**6**; ΔpK_a=4.4; $pK_{a,exp}$=4.0), bromocresol green (**7**; ΔpK_a=3.8; $pK_{a,exp}$=4.7), bromocresol purple (**8**;

ΔpK_a=2.6; $pK_{a,exp}$=6.3), bromothymol blue (**9**; ΔpK_a=3.1; $pK_{a,exp}$=7.0), and cresol red (**10**; ΔpK_a=2.1; $pK_{a,exp}$=8.3), and the anticoagulant rodenticide bromadiolone (**11**; ΔpK_a=4.7; $pK_{a,exp}$=4.0), are poorly modeled by PM6, other large phenols - particularly the more basic visual acid-base indicator such as phenol red (**12**; ΔpK_a=1.2; $pK_{a,exp}$=7.9), the dye 2-cresolphthalein (**13**; ΔpK_a=0.9; $pK_{a,exp}$=9.4), as well as the chemotherapeutic teniposide (**14**; ΔpK_a=0.6; $pK_{a,exp}$=10.1), are reasonably well approximated with the PM6 method. SPARC also performs weakly on these compounds, with ΔpK_a values of 2.6 for **6**, 2.1 for **7**, 2.7 for **8**, 1.8 for **9**, 1.8 for **10**, 1.8 for **11**, 1.8 for **12**, 0.6 for **13**, and -0.8 for **14**.

For α-saturated acids, the PM6 method displays strong predictive ability at pK_a values >4, but generally underpredicts the pK_a substantially (by up to several units) for polyacids with α-alcohol moieties (e.g., hydroxypropanedioic acid **15**, ΔpK_a=-2.2; $pK_{a,exp}$=2.4; tartaric acid **16**, ΔpK_a=-2.8; $pK_{a,exp}$=3.0; and isocitric acid **17**, ΔpK_a=-3.0; $pK_{a,exp}$=3.3). By comparison, SPARC does not have difficulty accurately predicting the pK_a values for any particular acidity range of α-saturated acids. Where an α-hydroxy or alkoxy group is present (i.e., glycerol, 1,2,3,4-butanetetrol, ethylene glycol and its monomethyl ether), the α-saturated alcohols are not very well modeled by the PM6 method throughout their acidity range, although the approach accurately predicts the pKa values of class members having simple hydrocarbon (i.e., methanol and ethanol) or halohydrocarbon (i.e., 2,2,2-trichloro- and trifluro-ethanols) α-substitution. Carboxylic acids with simple α-alkene unsaturation are very well modeled by PM6, but alkyne α-unsaturation (e.g., 2-propynoic acid **18**, ΔpK_a=1.6; $pK_{a,exp}$=1.8; 2-butynoic acid **19**, ΔpK_a=2.2; $pK_{a,exp}$=2.6) or α-carboxylation (*Z*-1-propene-1,2,3-tricarboxylic acid **20**, ΔpK_a=-2.8; $pK_{a,exp}$=2.0) confound obtaining reliable results. SPARC performs very well for all α-saturated alcohols (including the

α-hydroxy and alkoxy members), performs better than PM6 for the α-unsaturated acids as a whole, but also has difficulty with the alkyne α-unsaturation for compounds **18** (ΔpK_a=2.3), **19** (ΔpK_a=1.9), and **20** (ΔpK_a=1.1), as well as maleic acid (**21**, ΔpK_a=1.6).

For all carbon oxyacids, we stress that for a number of compounds, there still remains debate in the literature regarding the acidity constants. Thus, the comparative analyses presented here may need to be refined as more accurate experimental data becomes available. For some specific sub-classes where both the PM6 and SPARC methods agree with each other, but differ substantially from the experimental data in ref. [27] (e.g., sp-hybridized α-unsaturation on carboxylic acids), future studies and consensus in the literature may reveal that the computation methods were more accurate than the existing experimental data.

The validation dataset for the PM6 pK_a method in ref. [14] is dominantly comprised of aliphatic and aromatic carbon oxyacids (107 of 109 compounds listed), although one hydroxamic acid (benzohydroxamic acid **22**) and one oxime (benzophenone oxime **23**) are also given (both nitrogen oxyacids) with good agreements between their experimental and estimated pK_a values (ΔpK_a=0.01 [$pK_{a,exp}$=8.9] and ΔpK_a=-0.12 [$pK_{a,exp}$=11.3] for **22** and **23**, respectively, from ref. [14]; we obtained respective values of ΔpK_a=+0.00 and ΔpK_a=0.65 for **22** and **23**). The reported estimates for the pK_a values of these two compounds led us to examine what other classes of non-carbon oxyacids (e.g., boronic, sulfinic, and peroxy acids, as well as peroxides), including other members of the hydroxamic acids and oximes, may be amenable to reliable pK_a prediction using the PM6 method. Based on our studies, the PM6 method is not suitable for reliable estimation of any of these non-carbon oxyacids, including hydroxamic acids or oximes for which

we used a broader validation set than was given in ref. [14]. PM6 overestimates the pK_a values of all non-carbon oxyacids, in some cases by >10 units for compounds such as the boronic and sulfinic acids. For the hydroxamic acids, PM6 overestimates the pK_a at experimental values <9, and underestimates the pK_a at values >9, with benzohydroxamic acid **22** being the only member of this class that is reliably modeled. Similarly, oximes appear to display the opposite error trendings about an experimental pK_a value of 11. While SPARC cannot calculate the pK_a of sulfinic acids (precluding a comparative analysis with the PM6 method), SPARC does perform reasonably well with the peroxides and peroxyacids (average unsigned errors of 0.53 and 0.43, respectively), but is also not well suited for acidity prediction of the boron or nitrogen acids.

4. Conclusion

The pK_a estimation ability of the semiempirical PM6 method was evaluated across a broad range of oxyacids and compared to results obtained using the SPARC software program. The acidity of acetic and benzoic acids and their derivatives are well modeled by the PM6 method across their pK_a ranges, with weaker predictive capacity for alicyclic and aromatic heterocyclic acids, and phenols. α-Saturated acids are reliably modeled except for polyacid derivatives with α-alcohol moieties. α-Saturated alcohols only appear to yield reliable PM6 results where an α-hydroxy or α-alkoxy moiety is absent. Carboxylic acids with simple α-alkene unsaturation are very well modeled, but alkyne α-unsaturation or α-carboxylation confound obtaining reliable results. pK_a values for non-carbon oxyacids (e.g., boronic, sulfinic, hydroxamic, and peroxy acids, as well as oximes and peroxides) do not appear to be reliably modeled by the PM6 method. The PM6 and SPARC methods exhibit approximately equal pK_a prediction performance for the acetic,

alicyclic, and benzoic acids. SPARC outperforms PM6 on the peroxides, peroxyacids, phenols, and α-saturated acids and alcohols, is not capable of estimating acidity constants for sulfinic acids, and is also not suitable for reliable pK_a estimation for boron and nitrogen oxyacids. The findings from the current study will help constrain and validate efforts at using the PM6 method in estimating pK_a values of environmentally relevant compounds for the design and optimization of waste treatment methods, environmental fate investigations, and toxicological studies.

Acknowledgements

S.R. thanks the Natural Sciences and Engineering Research Council (NSERC) of Canada for financial support.

References

[1] H.H. Jaffe, A reexamination of the Hammett equation, Chem. Rev. 53 (1953) 191-261

[2] C. Hansch, A. Leo, Substituent Constants for Correlation Analysis in Chemistry and Biology, Wiley, New York, 1979.

[3] J. Clark , D.D. Perrin, Prediction of the strengths of organic bases, Q. Rev. Chem. Soc. 18 (1964) 295-320.

[4] D.D. Perrin, B. Dempsey, E.P. Serjeant, pK_a Prediction for Organic Acids and Bases, Chapman and Hall, London, 1981.

[5] S.H. Hilal, S.W. Karickhoff, A rigorous test for SPARC's chemical reactivity models: Estimation of more than 4300 ionization pK_as, Quant. Struc.-Act. Relat. 14 (1995) 348-355.

[6] A.C. Lee, J.Y. Yu, G.M. Crippen, pK_a prediction of monoprotic small molecules the SMARTS way, J. Chem. Inf. Model. 48 (2008) 2042-2053.

[7] S.L. Dixon, P.C. Jurs, Estimation of pK_a for organic oxyacids using calculated atomic charges, J. Comp. Chem. 14 (1993) 1460-1467.

[8] G. Schuurmann, Modelling pK_a of carboxylic acids and phenols, Quant. Struc.-Act. Relat. 15

(1996) 121-132.

[9] M.J. Citra, Estimating the pK_a of phenols, carboxylic acids and alcohols from semi-empirical quantum chemical methods, Chemosphere 38 (1999) 191-206.

[10] E. Soriano, S. Cerdan, P. Ballestros, Computational determination of pK_a values. A comparison of different theoretical approaches and a novel procedure, J. Mol. Struct. (Theochem) 684 (2004) 121-128.

[11] W.H. Richardson, C. Peng, D. Bashford, L. Noodleman, D.A. Case, Incorporating solvation effects into density functional theory: Calculation of absolute acidities, Int. J. Quant. Chem. 61 (1997) 207-217.

[12] A. Klamt, F. Eckert, M. Diedenhofen, M.E. Beck, First principles calculations of aqueous pK_a values for organic and inorganic acids using COSMO-RS reveal an inconsistency in the slope of the pK_a scale, J. Phys. Chem. A 107 (2003) 9380-9386.

[13] J.J.P. Stewart, Optimization of parameters for semiempirical methods V: Modification of NDDO approximations and application to 70 elements, J. Mol. Model. 13 (2007) 1173-1213.

[14] J.J.P. Stewart, MOPAC 2009, http://openmopac.net, Accessed 11 Dec 2008.

[15] T. Puzyn, N. Suzuki, M. Haranczyk, J. Rak, Calculation of quantum-mechanical descriptors

for QSPR at the DFT Level: Is it necessary?, J. Chem. Inf. Model. 48 (2008) 1174-1180.

[16] A. Alparone, V. Librando, Z. Minniti, Validation of semiempirical PM6 method for the prediction of molecular properties of polycyclic aromatic hydrocarbons and fullerenes, Chem. Phys. Lett. 460 (2008) 151-154.

[17] J.C. Dearden, M.T.D. Cronin, D.C. Lappin, A comparison of commercially available software for the prediction of pK_a, J. Pharm. Pharmacol. 59 (2007) 1-16.

[18] B. Slater, A. McCormack, A. Avdeef, J.E.A. Commer, pH-Metric log P. 4. Comparison of partition coefficients determined by HPLC and potentiometric methods to literature values, Pharm. Sci. 83 (1994) 1280-1283.

[19] M. Meloun, S. Bordovska, Benchmarking pK_a prediction and algorithm validation for accurate pK_a of drugs estimated from their molecular structures, Anal. Bioanal. Chem. 389 (2007) 1267-1281.

[20] T.Z. Verbic, B.J. Drakulic, M.F. Zloh, J.R. Pecelj, G.V. Popovic, I.O. Juranic IO, An LFER study of 4-aryl-2,4-dioxobutanoic acids protolytic equilibria in aqueous solutions, J. Serb. Chem. Soc. 72 (2007) 1201-1216.

[21] S. Rayne, K. Forest, K.J. Friesen, Computational approaches may underestimate pK_a values of longer-chain perfluorinated carboxylic acids: Implications for assessing environmental and

biological effects, J. Env. Sci. Health A 44 (2009) DOI: 10.1080/10934520802659620.

[22] S. Rayne, K. Forest, K.J. Friesen, Extending the semi-empirical PM6 method for carbon oxyacid pK_a prediction to sulfonic acids: Application towards congener-specific estimates for the environmentally and toxicologically relevant C_1 through C_8 perfluoroalkyl derivatives, Lett. Org. Chem., submitted.

[23] L.A. Carreira, S. Hilal, S.W. Karickhoff, Estimation of chemical reactivity parameters and physical properties of organic molecules using SPARC, in: P. Politzer, J. S. Murray (Eds.), Theoretical and Computational Chemistry - Quantitative Treatment of Solute/Solvent Interactions, Elsevier Publishers, St. Louis, MO, USA, 1994.

[24] D. Weininger, SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules, J. Chem. Inf. Comp. Sci. 28 (1988) 31-36.

[25] D. Weininger, A. Weininger, J.L. Weininger, SMILES. 2. Algorithm for generation of unique SMILES notation, J. Chem. Inf. Comp. Sci. 29 (1989) 97-101.

[26] A. Ebrahimi, S.M. Habibi, R.S. Neyband, Substituent effect on intramolecular hydrogen bonding in 2-hydroxybenzaldehyde, Int. J. Quant. Chem. (2009) DOI:10.1002/qua.21947.

[27] D.R. Lide, CRC Handbook of Chemistry and Physics, 87th edition, Taylor and Francis, Boca Raton, FL, USA, 2007.

Figure Captions

Fig. 1. General structures for the compounds classes under consideration.

Fig. 2. Comparison between the source validation predicted pK_a values using the PM6 method in MOPAC 2009 from ref. [14] and the current work. A 1:1 line is shown for comparison.

Fig. 3. Structures of compounds discussed in the text.

Fig. 4. Comparison between experimentally observed (x-axis) and predicted (y-axis) pK_a values using the semiempirical PM6 method in MOPAC 2009 (▢) and the SPARC method (○) by compound class. 1:1 lines are shown in each plot.

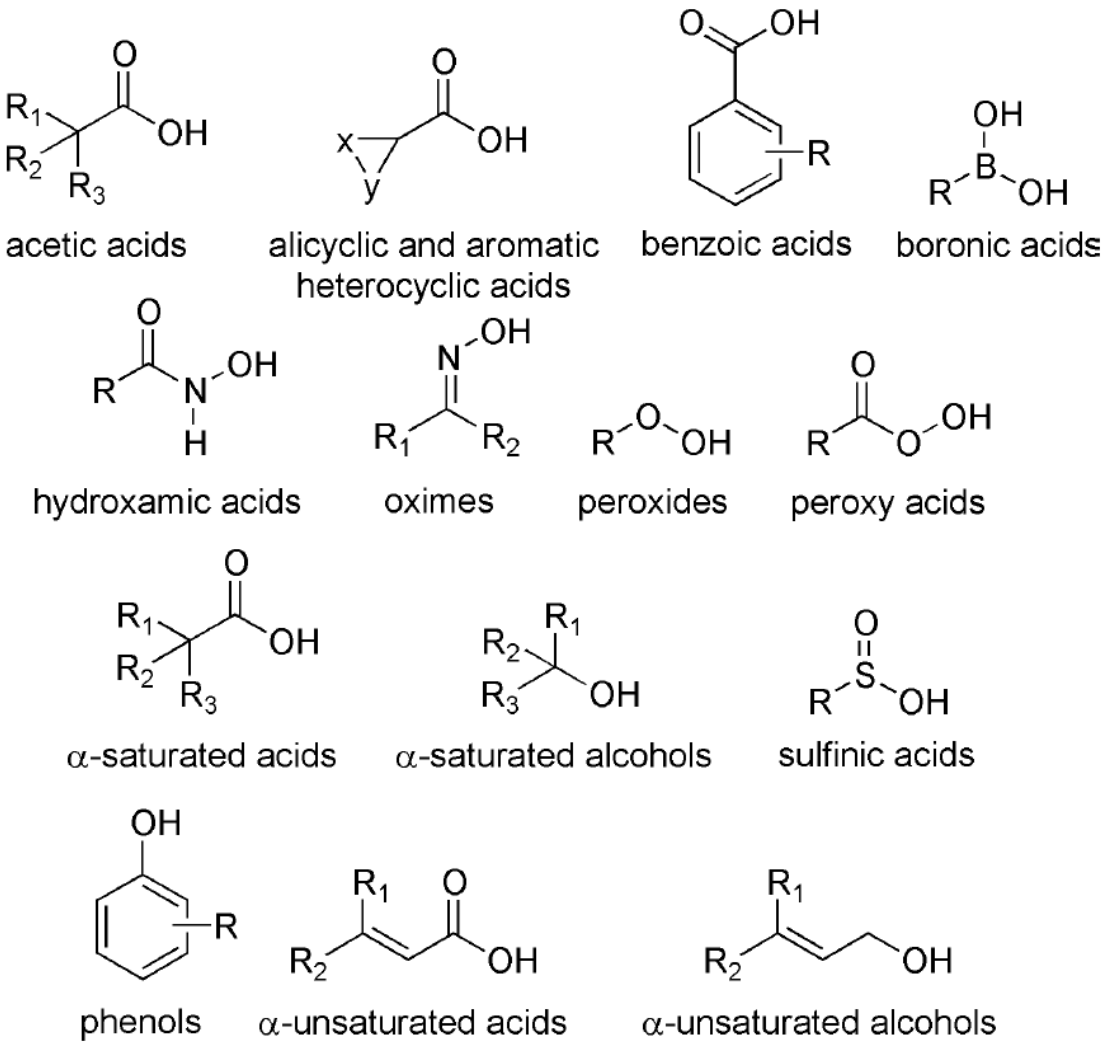

Fig. 1

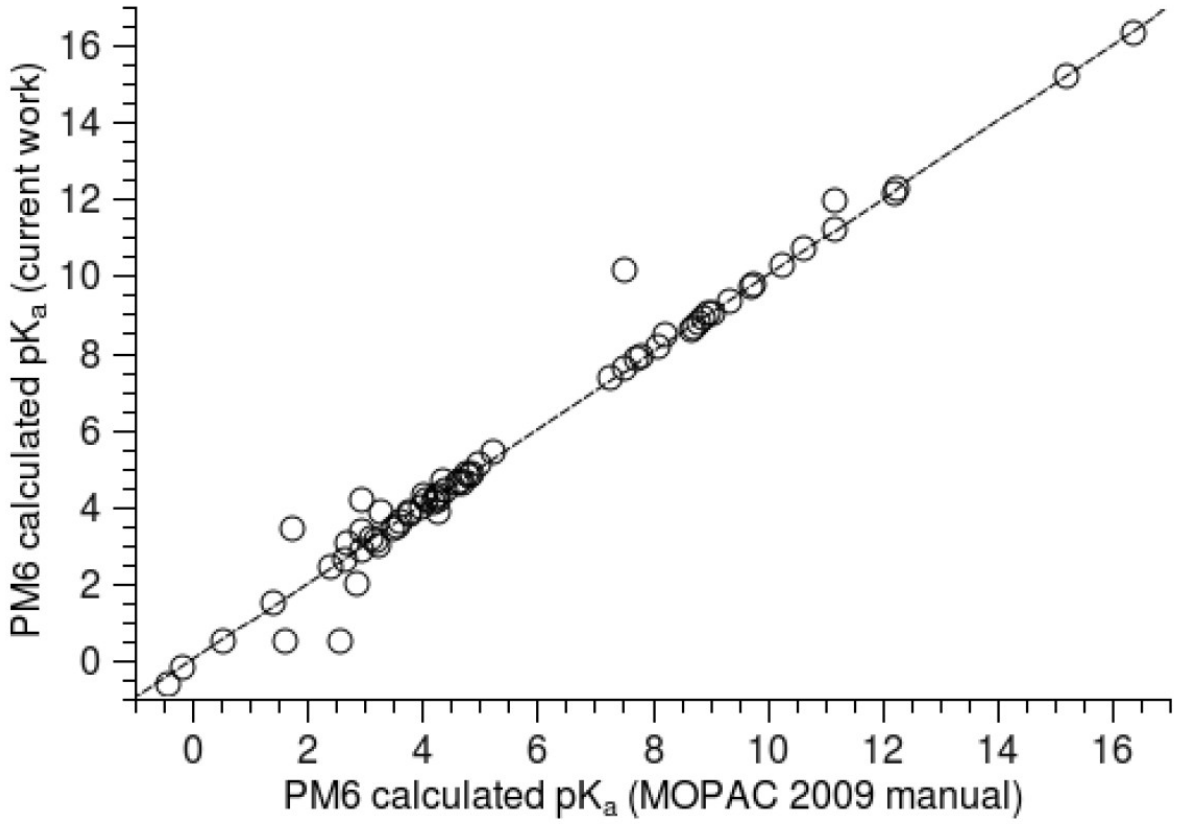

Fig. 2

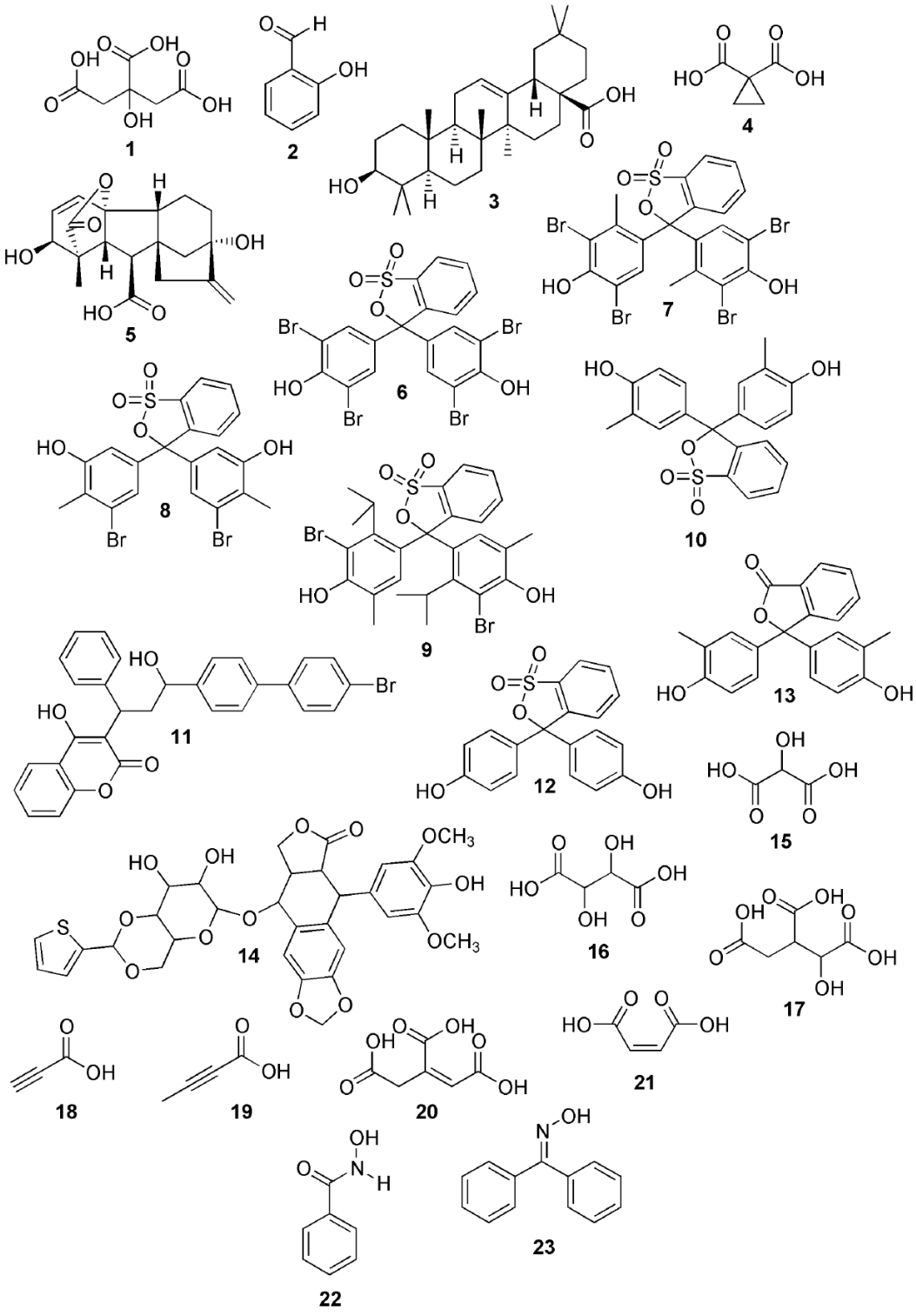

Fig. 3

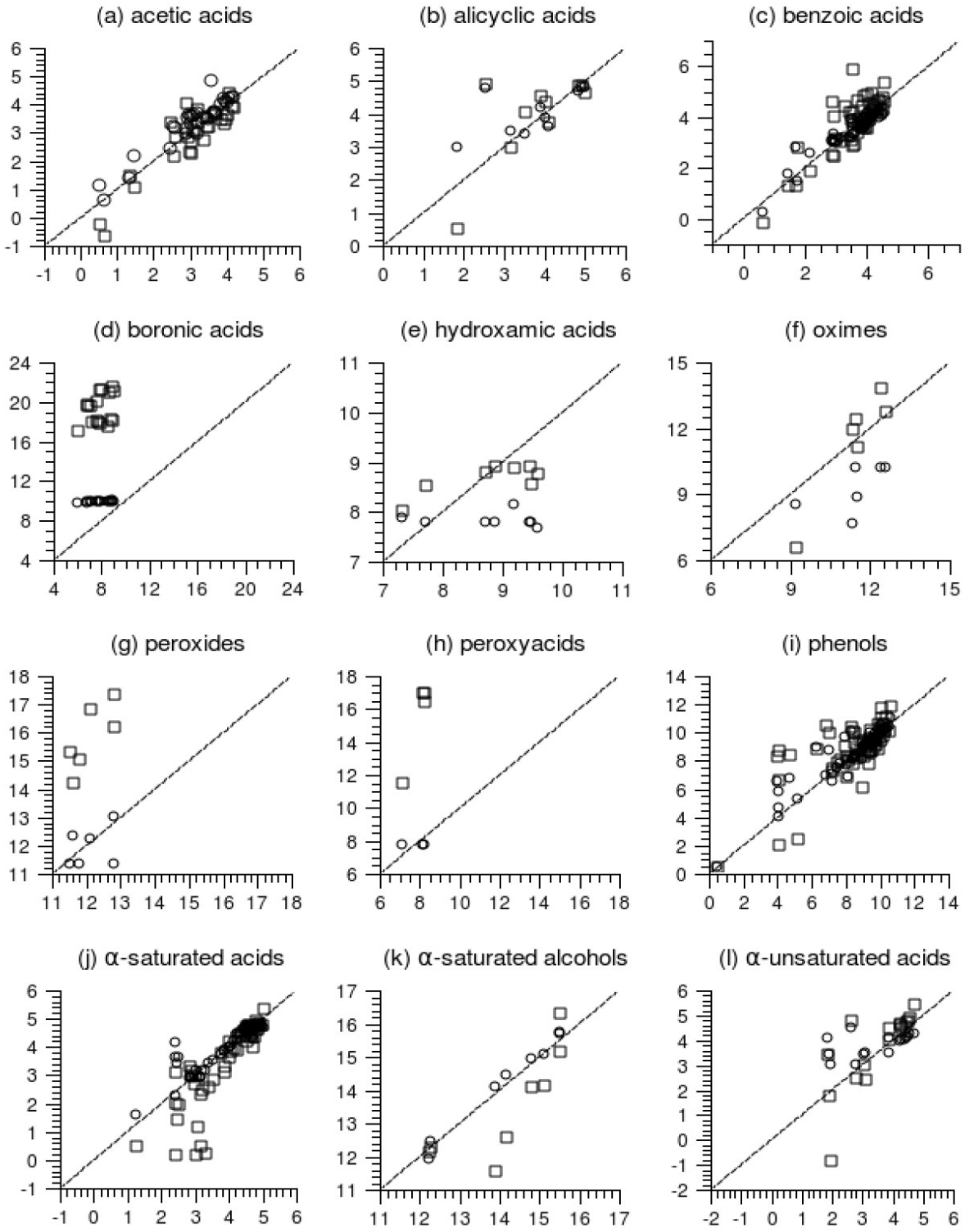

Fig. 4