JOURNAL OF
CHEMICAL INFORMATION
AND MODELING

ARTICLE

pubs.acs.org/jcim

Predicting pK_a Values of Substituted Phenols from Atomic Charges: Comparison of Different Quantum Mechanical Methods and Charge Distribution Schemes

Radka Svobodová Vařeková,[†] Stanislav Geidl,[†] Crina-Maria Ionescu,[†] Ondřej Skřehota,[†] Michal Kudera,[†] David Sehnal,[†] Tomáš Bouchal,[†] Ruben Abagyan,[‡] Heinrich J. Huber,[§] and Jaroslav Koča*,[†]

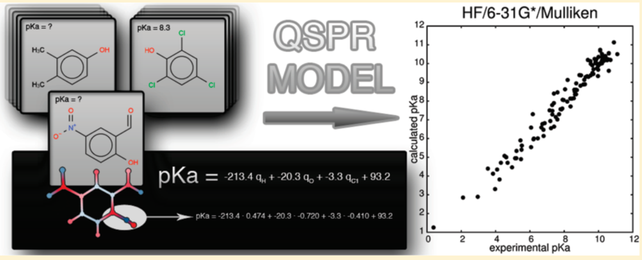
[†]National Centre for Biomolecular Research, Faculty of Science and CEITEC − Central European Institute of Technology, Masaryk University Brno, Kamenice 5, 625 00, Brno-Bohunice, Czech Republic

[‡]Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California, 9500 Gilman Drive, MC 0657, San Diego, California, United States

[§]Systems Biology Group, Royal College of Surgeons in Ireland, 123 St Stephens Green, Dublin 2, Ireland

S *Supporting Information*

ABSTRACT: The acid dissociation (ionization) constant pK_a is one of the fundamental properties of organic molecules. We have evaluated different computational strategies and models to predict the pK_a values of substituted phenols using partial atomic charges. Partial atomic charges for 124 phenol molecules were calculated using 83 approaches containing seven theory levels (MP2, HF, B3LYP, BLYP, BP86, AM1, and PM3), three basis sets (6-31G*, 6-311G, STO-3G), and five population analyses (MPA, NPA, Hirshfeld, MK, and Löwdin). The correlations between pK_a and various atomic charge descriptors were examined, and the best descriptors were selected for preparing the quantitative structure−property relationship (QSPR) models. One QSPR model was created for each of the 83 approaches to charge calculation, and then the accuracy of all these models was analyzed and compared. The pK_as predicted by most of the models correlate strongly with experimental pK_a values. For example, more than 25% of the models have correlation coefficients (R^2) greater than 0.95 and root-mean-square errors smaller than 0.49. All seven examined theory levels are applicable for pK_a prediction from charges. The best results were obtained for the MP2 and HF level of theory. The most suitable basis set was found to be 6-31G*. The 6-311G basis set provided slightly weaker correlations, and unexpectedly also, the STO-3G basis set is applicable for the QSPR modeling of pK_a. The Mulliken, natural, and Löwdin population analyses provide accurate models for all tested theory levels and basis sets. The results provided by the Hirshfeld population analysis were also acceptable, but the QSPR models based on MK charges show only weak correlations.



■ INTRODUCTION

The acid dissociation (ionization) constant pK_a is one of the fundamental properties of organic molecules determining the degree of dissociation at a given pH. Dissociation constants are of interest in chemical, biological, environmental, and pharmaceutical research because the important physicochemical properties, like lipophilicity, solubility, and permeability, are all pK_a dependent. The values of these constants are, e.g., essential for absorption, distribution, metabolism, elimination (ADME) profiling.[1] Ionization constants also provide an insight into interactions of drugs containing ionizable groups with a receptor. In drug formulation, pK_a is important for the choice of an appropriate excipient and counterion. Furthermore, pK_a is often used as a descriptor for quantitative structure−activity relationship (QSAR) models. For these reasons, there is a strong interest in the development of reliable methods for pK_a prediction.

Numerous pK_a prediction methods based on different approaches were developed. The linear free energy relationships (LFER) method,[2,3] applying the Hammett and Taft equations, is one of the first approaches used for pK_a prediction. LFER models are still used and have been implemented in popular software packages, such as ACD/pK_a,[4] EPIK,[5] and SPARC.[6] Database methods use similarity metrics[7] to assign the pK_a value of the molecule of interest to the pK_a value obtained from the most similar molecule found in dedicated databases. Likewise, the decision tree method uses similarity metrics and builds a tree which provides a decision path for processing a compound. Ab initio quantum mechanical (QM) methods have often been found to be the most accurate,[8] such as the Jaguar pK_a prediction module,[9] which performs geometry

Received: March 18, 2011
Published: July 17, 2011

optimization at the density functional theory (DFT) B3LYP/6-31G* level, or the approach of Shields et al.,[10] which used the CPCM[11] continuum solvation model in Gaussian 98.[12] On the other hand, the applicability of these approaches is limited due to their computational complexity. A popular way to benefit from quantum mechanical calculations while keeping lower computational costs is to use QM descriptors which have a strong correlation with pK_a. Such descriptors include, e.g., polarizability,[13] free energies [phenoxide highest occupied molecular orbital (HOMO) energy,[14] relative proton transfer energy,[14] minimum surface local ionization energy],[15] partial atomic charges,[16,17] group philicity,[18] molecular electrostatic potential,[19] etc. Information-based descriptors (i.e., molecular tree structured fingerprints or 2D substructure flags,[20] topological sphere descriptors,[21] steric descriptors,[21] etc.) are also applicable. One of the most common techniques that uses descriptors in pK_a prediction is QSAR or quantitative structure − property relationships (QSPR) in combination with partial least-squares (PLS) or multiple linear regression (MLR), while other methods include artificial neural networks (ANN).[8] Unfortunately, pK_a values remain one of the most challenging physicochemical properties to predict.

Using partial atomic charges to estimate the relative acidity or reactivity of organic compounds has a long history in organic chemistry. Specifically, the partial atomic charges concept allows the prediction of relative acidity or reactivity by estimating the extent of charge delocalization based on molecular structure information.

Therefore, the correlation between pK_a and relevant atomic charges calculated by different ab initio or semiempirical approaches has been analyzed. For example, Gross et al.[22] studied which population analyses provide a good correlation at the B3LYP/6-311G** level of theory for substituted phenols and anilines, and Kreye et al.[23] compared three different levels of theory for substituted phenols (RM1 with and without the SM8 solvent model and B3LYP/6-311G** and B3LYP/6-31+G* with the SM8 solvent model). Partial atomic charges are also often and successfully used as part of the descriptors set in QSAR/QSPR models. Dixon et al. calculated pK_a from σ and π partial charges,[17] Citra[16] used partial charges and bond order, Xing et al.[24] charges and polarizabilities, Soriano et al.[25] charges and frontier orbital energy, and Yangjeh[13] combined charges, polarizability, molecular weight, hydrogen-bond accepting capability, and partial-charge weighted topological electronic descriptors. The above studies demonstrate that charges are very powerful descriptors for pK_a modeling and show linear dependency between charges and pK_a. Charge utilization has been limited by the high computational cost of their quantum mechanical calculation.

Nowadays, computers are powerful enough to make QM charges accessible in a much shorter time. Moreover, empirical charge calculation approaches, like equalization methods,[26] are able to mimic QM methods with high accuracy, and such empirical approaches are even markedly faster than QM methods themselves. These facts create a good reason to develop accurate pK_a prediction models that are based on QM charges, because they can subsequently be used in techniques like virtual screening.

In the present study we report on the evaluation of pK_a prediction QSPR models based on 83 different charge calculation approaches. Specifically, we applied 83 combinations of theory levels (MP2, HF, B3LYP, BLYP, BP86, AM1, and PM3), basis sets (6-31G*, 6-311G, STO-3G) and population analyses (MPA, NPA, Hirshfeld, MK, and Löwdin). Then, we compared the correlations between experimental pK_a values and various atomic charge descriptors and used the best descriptors for designing the QSPR models. We created a model for each of the 83 approaches of charge calculation and subsequently analyzed the ability of these models to predict pK_a. The analysis was performed on phenol molecules, a class of compounds frequently used for the evaluation of pK_a prediction models.[16,22,23]

There are basically two possible ways to create a QSPR model of a feature to be predicted. The first is to create as general a model as possible, with the risk that the accuracy of such a model may not be high. The second approach is to develop more models, each of them being dedicated to a certain class of compounds. In our work, we follow the second approach and start with phenols.

■ METHODS

Data Sets. Our data set contains the 3D structures of 124 distinct phenol molecules. The list of the molecules, including their experimental pK_a values, can be found in the Supporting Information. This data set is of high structural diversity, meaning it contains a wide range of electron-withdrawing and electron-donating substituents, covering a pK_a range of about 10 log units. The molecules were obtained from the NCI Open Database Compounds.[27] This database consists of organic molecules tested against cancer, and it includes their two-dimensional (2D) structures and also their 3D structures predicted by CORINA 2.6.[28] The main reason we used the CORINA approach is speed and compatibility with some other studies. The key point is speed. Our final goal with the approach is to use it when searching large databases for virtual screening purposes. CORINA provides an approximation of the global minimum conformation very quickly. Moreover, it is quite a common software for the preparation of 3D structures used in the validation of pK_a prediction models.[20,21,29]

pK_a Values. The experimental pK_a values were taken from the Physprop database.[30] The structures of phenol molecules from the NCI Open Database and their Physprop pK_a values were paired using the CAS registry numbers, which are unique identifiers in both databases.

Atomic Charge Calculation. All atomic charge calculations were carried out using Gaussian03 from Gaussian Inc.[31] The merely inputs for charge calculations were the 3D coordinates generated by CORINA, i.e., without any further geometry optimization (in a similar way as Ertl et al.).[20] The reason why QM optimization was skipped is again the speed of the approach. An optimization procedure even based on a QM method would bring the problem to a different level of computational complexity and related cost, which would not allow it to be used for our intended purposes, i.e., searching large databases.

Five ab initio levels of theory were examined. The first two were the standard Hartree−Fock (HF) method and the second-order Møller−Plesset (MP2) perturbation theory, which includes more sophisticated approximations of the Hamiltonian compared to HF. A computational cost of HF and MP2 is $\theta(N^4)$ and $\theta(N^5)$, respectively, where N is the number of basis functions. The other three were the DFT methods with BLYP, BP86, and B3LYP functionals. BLYP is a representative of gradient corrected functionals and is denoted according to its authors (Becke, Lee, Yang and Parr). BP86 (Becke Perdew 1986) is similar to BLYP but uses an older correlation functional (Perdew-86). B3LYP (Becke, three-parameter, Lee–Yang–Parr) is a hybrid functional constructed as a linear combination of the HF and BLYP functionals. A computational cost of all these DFT methods is $\theta(N^3)$. The basis sets STO-3G, 6-31G*, and 6-311G were used for each level of theory, therefore 15 combinations of theory levels and basis sets were studied (HF/STO-3G, HF/6-31G*, HF/6-311G, ..., BP86/STO-3G, BP86/6-31G*, and BP86/6-311G). Five types of charges were calculated for each of these 15 pairs of theory levels and basis sets—charges derived from: natural population analysis (NPA), Mulliken charges (MPA), Löwdin charges, Hirshfeld charges, and Merz–Singh–Kollman charges fit to the electrostatic potential (MK). Moreover, the application of two semiempirical methods Austin model 1 (AM1) and parameterization method 3 (PM3) with four PA (Mulliken, Löwdin, Hirshfeld, and MK) was analyzed. Both AM1 and PM3 exhibit computational cost of $\theta(N^3)$. Consequently, this publication examines 83 approaches for charge calculation and analyzes their relevance for pK_a calculation.

Descriptors. The selection of appropriate descriptors that are significantly related to the property of interest is very important for predictive QSPR models. The descriptors can be chosen using domain knowledge about the examined property, or the mathematical methods for the selection of descriptors can be applied. In our work, we have utilized both approaches. We have focused on atomic charges and their ability to estimate the pK_a of phenols. Therefore, atomic charges and their sums and differences have been employed as the descriptors. First, according to traditional knowledge about atomic charge influence on pK_a in phenols, we selected the atomic charge of the hydrogen atom from the phenolic OH group (q_{H}) and the atomic charges of the atoms close to this hydrogen as descriptors. These atoms and their denotations are shown in Figure 1. We also verified in all our molecules that this hydrogen is the most positively charged hydrogen, and therefore this hydrogen will dissociate first. Further descriptors are therefore the charge on the oxygen atom (q_{O}), the charge on the C1 carbon atom (q_{C1}), and the charge on the C4 carbon atom (q_{C4}). Because it is not possible to distinguish between the charges on C2 and C6, the sum of these charges was used as a descriptor ($q_{\mathrm{C2+C6}}$) and the same for C3 and C5 ($q_{\mathrm{C3+C5}}$). We further evaluated as descriptors also the sums and the differences of these atomic charges—the difference between the O and H charge ($q_{\mathrm{O-H}}$), the sum of charges on C1, C2, and C6 ($q_{\mathrm{C1+C2+C6}}$), the sum of charges on C3, C4, and C5 ($q_{\mathrm{C3+C4+C5}}$), and the sum of charges on all carbons in the phenolic group (q_{phe}). After this we evaluated the correlation between these 10 descriptors and the experimental pK_a values using the squared Pearson correlation coefficient (R^2) and the Student's statistic of the regression (t) in order to find descriptors significantly correlating with pK_a. These descriptors were used to establish the QSPR models.

QSPR Models: Parametrization and Quality Evaluation. The general equation for our QSPR models is

$$\mathrm{p}K_{\mathrm{a}} = param_1 \cdot descr_1 + param_2 \cdot descr_2 + ... + param_n \cdot descr_n + param_{n+1} \qquad (1)$$

where descr$_1$, descr$_2$, ..., descr$_n$ are the descriptors mentioned above; param$_1$, param$_2$, ..., param$_{n+1}$ are parameters of the QSPR model (i.e., constants derived by multiple linear regression), and n is the number of descriptors in the QSPR model. The parametrization of the QSPR models was done by MLR. We prepared one model for each procedure of charge calculation; therefore 83 different QSPR models were generated. The quality of the QSPR models, i.e., the correlation between experimental pK_a and the pK_a calculated by the model, was evaluated using the squared Pearson correlation coefficient (R^2), root-mean-square error (RMSE), average absolute pK_a error ($\overline{\Delta}$), standard deviation of the estimation (s), and Fisher's statistics of the regression (F). The robustness of the models was tested by cross-validation. Details about this procedure and its results are described in the following text.

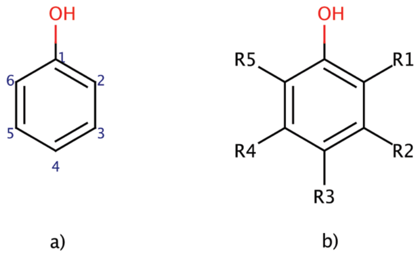


Figure 1. (a) The atom enumeration in phenols. (b) Markush structure of molecules from the data set, where R1, R2, ... R5 = $-CH_3$, $-CH{=}O$, $-C_6H_5$, $-O-CH_2-CH_3$, $-CH(CH_3)_2$, $-O-CH_3$, $-C{=}O-CH_3$, $-C{=}O-NH_2$, $-CH_2-OH$, ... $-Cl$, $-Br$, $-F$, and $-NO_2$.

RESULTS AND DISCUSSION

Evaluation of Descriptors. As the first step of our study, we investigated the pK_a predicting capabilities of all 10 suggested descriptors: q_{H}, q_{O}, q_{C1}, $q_{\mathrm{C2+C6}}$, $q_{\mathrm{C3+C5}}$, q_{C4}, ..., q_{phe}. Consequently, we calculated the atomic charges of all 124 phenol molecules from the data set via 83 combinations of theory levels (HF, MP2, B3LYP, BLYP, BP86, AM1, and PM3), population analyses (natural, Mulliken, Löwdin, Hirshfeld, and MK) and basis sets (STO-3G, 6-311G, and 6-31G*). And afterward we calculated the squared Pearson coefficients (R^2) and Student's t-value (t) for the correlations between each descriptor and experimental pK_a values for all 83 procedures of charge calculation.

The Hirshfeld PA demonstrates an untypical correlation between descriptors and pK_a for the basis sets STO-3G and 6-311G with all levels of theory, where the set contains eight strong outliers, all bromophenol molecules in the data set, and there is no reasonable correlation (Figure 2a). When the outliers were removed, the correlations became similar to those for Mulliken, natural, or Löwdin population analyses (Figure 2b). When the polarization basis set 6-31G* is used or when the semiempirical methods are applied, the charges obtained via the Hirshfeld PA do not contain the outliers. Therefore, we removed the bromophenols from the data set and recalculated the correlation coefficients and Student's t-values for the Hirshfeld PA and the basis sets STO-3G and 6-311G using this reduced set of 116 molecules.

The values of R^2 and t for all charge calculation procedures and all descriptors are summarized in the Supporting Information (Table S1), and a set of selected values of R^2 are visualized in Figures 3–5. These results show that q_{H} and q_{O} have a high correlation with experimental pK_a, i.e., $R^2 > 0.8$ for most charge calculation approaches. Almost all these correlation coefficients are statistically significant at p = 0.05. It is worth mentioning that, for the sets with 124 or 116 molecules, the R^2 is statistically significant (at $p = 0.05$) when $t > 1.66$. Also q_{C1} exhibits a good correlation, i.e., $R^2 > 0.5$ for some approaches

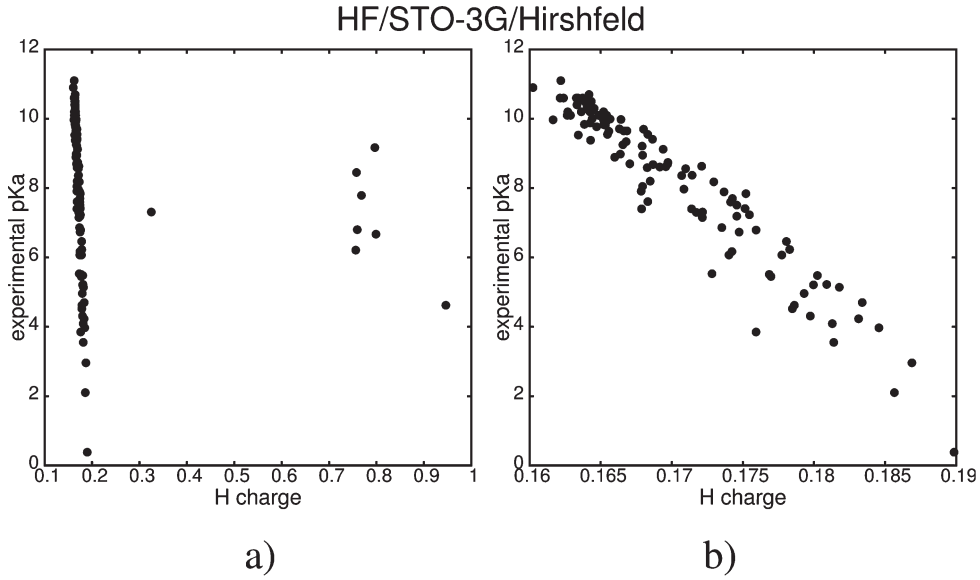


Figure 2. Correlations between q_H and pK_a for HF, STO-3G, and Hirshfeld PA. Graph (a) with and (b) without outliers.

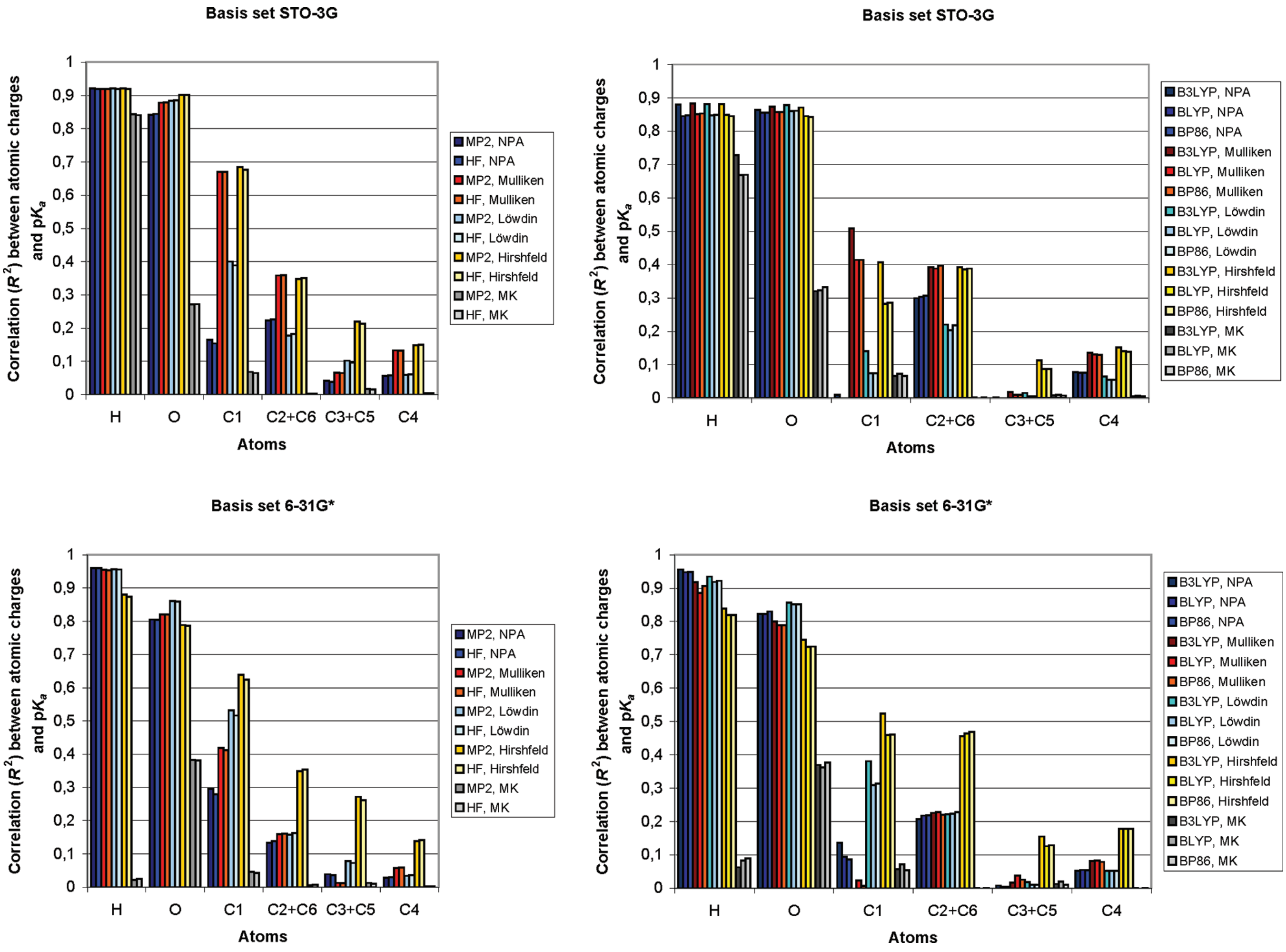


Figure 3. Correlations between descriptors and experimental pK_a.

and $t > 1.66$ for most approaches. The q_{O-H} shows a good correlation ($R^2 > 0.5$ and $t > 1.66$ for many approaches) too, but this descriptor is only a combination of q_O and q_H, and both q_O and q_H are better descriptors than q_{O-H}. Therefore, it

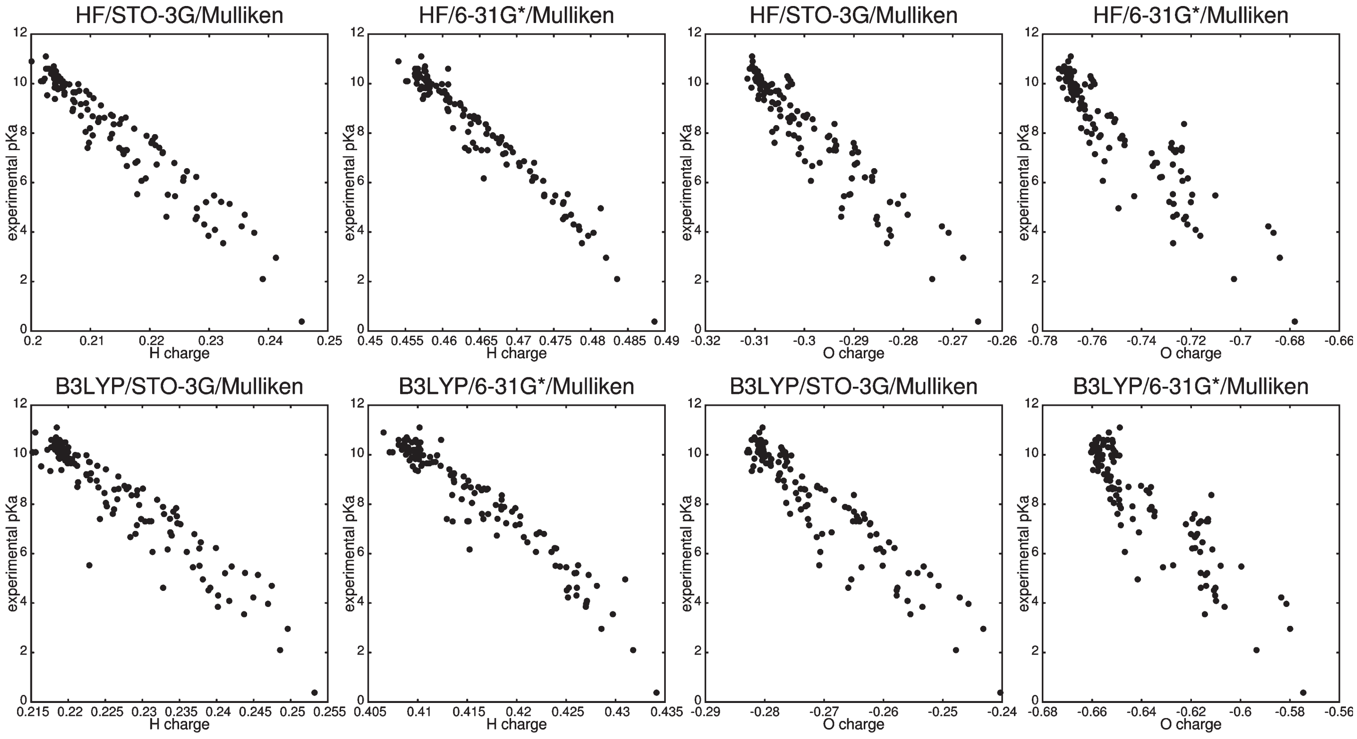


Figure 4. Correlations between q_H, q_O, and experimental pK_a for Mulliken PA and some selected basis sets and theory levels.

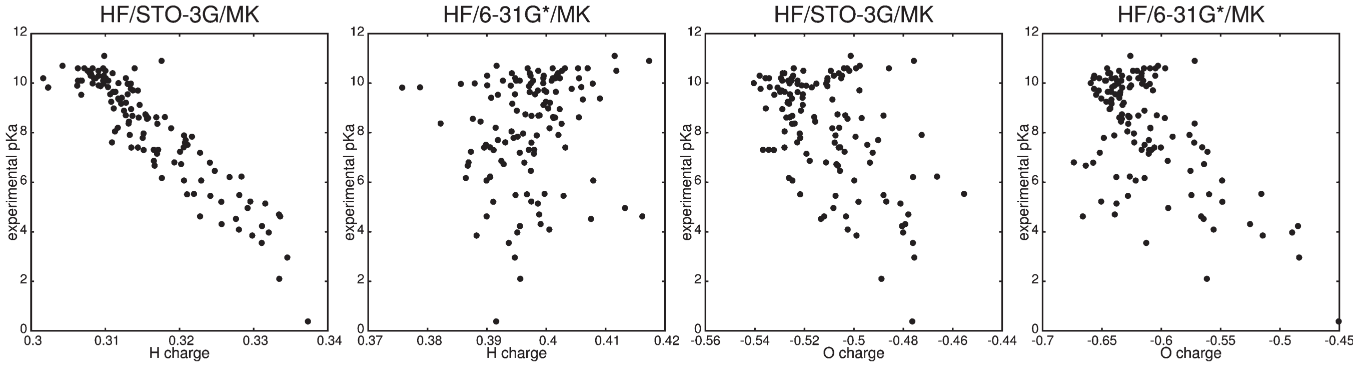


Figure 5. Correlations between q_H, q_O, and experimental pK_a for MK PA and some selected basis sets and theory levels.

makes no sense to introduce q_{O-H} into the models. Further descriptors show only a weak correlation (<0.5). For these reasons, the descriptors q_H, q_O, and q_{C1} were selected to create the QSPR models.

Figures 3−5 and Supporting Information (Table S1) also help us to recognize basic trends in the relevance of the charge calculation approaches for pK_a prediction. All seven theory levels seem applicable for pK_a prediction. Specifically, most of the charge calculation approaches utilizing these theory levels provide q_H or q_O correlating with pK_a with $R^2 > 0.8$ and statistically significant (at $p = 0.05$). In addition, all three basis sets and four out of five examined population analyses seem applicable for the same considerations. However, the MK PA demonstrates only weak correlation with pK_a and most approaches using MK provide q_H or q_O correlating with pK_a with $R^2 < 0.5$. Examples of correlation graphs for the Mulliken and MK population analyses are shown in Figures 4 and 5.

The trends for Mulliken, natural, Löwdin, and MK PA are in agreement with the study of Gross et al.,[22] who examined the correlation between q_O, q_H, and pK_a for different population analyses on the B3LYP/6-311G level of theory on a set of 19 substituted phenols. An interesting discovery of these analyses was also the fact that the correlation between pK_a and the atomic charge descriptors decreases linearly with the distance from the hydrogen atom of the phenolic OH group.

Parameterization and Validation of QSPR Models. The descriptors q_H, q_O, and q_{C1} were selected as inputs for the QSPR models, therefore the models have used the following equation for pK_a calculation:

$$\mathrm{p}K_a = param_H \cdot q_H + param_O \cdot q_O + param_{C1} \cdot q_{C1} + constant \qquad (2)$$

where param$_H$, param$_O$, param$_{C1}$, and constant are the parameters of the model. The parametrization of the QSPR models

Table 1. Quality Criteria and Statistical Criteria for All the QSPR Models[a]

model number	theory level	population analysis	basis set	R^2	RMSE	$\overline{\Delta}$	s	F	number of molecules
1	MP2	Mulliken	6-31G*	0.966	0.403	0.315	0.410	1136	124
2	HF	Mulliken	6-31G*	0.966	0.403	0.315	0.410	1136	124
3	MP2	Löwdin	6-31G*	0.966	0.405	0.315	0.412	1136	124
4	HF	Löwdin	6-31G*	0.966	0.406	0.315	0.413	1136	124
5	MP2	NPA	6-31G*	0.964	0.419	0.325	0.426	1071	124
6	B3LYP	Löwdin	6-31G*	0.963	0.421	0.325	0.428	1041	124
7	HF	NPA	6-31G*	0.963	0.421	0.329	0.428	1041	124
8	B3LYP	NPA	6-311G	0.962	0.428	0.336	0.435	1013	124
9	MP2	NPA	6-311G	0.961	0.432	0.344	0.439	986	124
10	B3LYP	NPA	6-31G*	0.961	0.433	0.332	0.440	986	124
11	HF	NPA	6-311G	0.961	0.434	0.346	0.441	986	124
12	BLYP	NPA	6-311G	0.96	0.437	0.341	0.444	960	124
13	BP86	NPA	6-311G	0.959	0.443	0.342	0.450	936	124
14	B3LYP	Mulliken	6-31G*	0.959	0.443	0.355	0.450	936	124
15	BLYP	NPA	6-31G*	0.959	0.444	0.34	0.451	936	124
16	BLYP	Löwdin	6-31G*	0.959	0.444	0.34	0.451	936	124
17	BP86	Löwdin	6-31G*	0.959	0.445	0.341	0.452	936	124
18	BP86	NPA	6-31G*	0.959	0.447	0.342	0.454	936	124
19	BP86	Mulliken	6-31G*	0.954	0.471	0.374	0.479	830	124
20	BLYP	Mulliken	6-31G*	0.953	0.477	0.377	0.485	811	124
21	MP2	Löwdin	6-311G	0.951	0.486	0.375	0.494	776	124
22	HF	Löwdin	6-311G	0.95	0.491	0.38	0.499	760	124
23	HF	Mulliken	6-311G	0.945	0.513	0.399	0.521	687	124
24	MP2	Mulliken	6-311G	0.945	0.514	0.401	0.522	687	124
25	BP86	Mulliken	6-311G	0.939	0.541	0.429	0.550	616	124
26	B3LYP	Mulliken	6-311G	0.938	0.547	0.433	0.556	605	124
27	B3LYP	Löwdin	6-311G	0.937	0.551	0.417	0.560	595	124
28	BLYP	Mulliken	6-311G	0.932	0.573	0.452	0.582	548	124
29	BP86	Löwdin	6-311G	0.931	0.577	0.443	0.587	540	124
30	MP2	Hirshfeld	STO-3G	0.929	0.594	0.467	0.605	488	116
31	HF	Hirshfeld	STO-3G	0.928	0.597	0.47	0.608	481	116
32	BLYP	Löwdin	6-311G	0.926	0.596	0.451	0.606	501	124
33	AM1	Mulliken	–	0.924	0.605	0.452	0.615	486	124
34	AM1	Löwdin	–	0.924	0.605	0.452	0.615	486	124
35	MP2	Mulliken	STO-3G	0.922	0.615	0.502	0.625	473	124
36	MP2	Löwdin	STO-3G	0.921	0.617	0.505	0.627	466	124
37	MP2	NPA	STO-3G	0.921	0.618	0.501	0.628	466	124
38	HF	Mulliken	STO-3G	0.92	0.619	0.508	0.629	460	124
39	HF	Löwdin	STO-3G	0.92	0.621	0.51	0.631	460	124
40	HF	NPA	STO-3G	0.92	0.622	0.506	0.632	460	124
41	AM1	Hirshfeld	–	0.917	0.631	0.499	0.641	442	124
42	MP2	Hirshfeld	6-31G*	0.912	0.652	0.529	0.663	415	124
43	MP2	Hirshfeld	6-311G	0.91	0.667	0.534	0.679	377	116
44	HF	Hirshfeld	6-31G*	0.908	0.665	0.538	0.676	395	124
45	HF	Hirshfeld	6-311G	0.907	0.678	0.541	0.690	364	116
46	B3LYP	Mulliken	STO-3G	0.904	0.68	0.558	0.691	377	124
47	B3LYP	Hirshfeld	STO-3G	0.902	0.698	0.536	0.710	344	116
48	B3LYP	Hirshfeld	6-31G*	0.897	0.705	0.546	0.717	348	124
49	BP86	Mulliken	STO-3G	0.896	0.707	0.575	0.719	345	124
50	BLYP	Mulliken	STO-3G	0.896	0.709	0.581	0.721	345	124
51	B3LYP	Löwdin	STO-3G	0.895	0.71	0.565	0.722	341	124
52	PM3	Hirshfeld	–	0.895	0.711	0.561	0.723	341	124
53	B3LYP	NPA	STO-3G	0.894	0.715	0.567	0.727	337	124
54	BP86	Hirshfeld	6-31G*	0.89	0.729	0.553	0.741	324	124

Table 1. Continued

model number	theory level	population analysis	basis set	R^2	RMSE	$\overline{\Delta}$	s	F	number of molecules
55	BLYP	Hirshfeld	6-31G*	0.886	0.741	0.567	0.753	311	124
56	BLYP	Hirshfeld	STO-3G	0.886	0.75	0.571	0.763	290	116
57	BP86	Hirshfeld	STO-3G	0.882	0.763	0.578	0.777	279	116
58	B3LYP	Hirshfeld	6-311G	0.882	0.764	0.599	0.778	279	116
59	PM3	Mulliken	–	0.88	0.76	0.581	0.773	293	124
60	PM3	Löwdin	–	0.88	0.76	0.581	0.773	293	124
61	BLYP	Löwdin	STO-3G	0.879	0.764	0.599	0.777	291	124
62	BP86	Löwdin	STO-3G	0.878	0.766	0.597	0.779	288	124
63	BLYP	NPA	STO-3G	0.877	0.769	0.604	0.782	285	124
64	BP86	NPA	STO-3G	0.876	0.772	0.601	0.785	283	124
65	BP86	Hirshfeld	6-311G	0.874	0.789	0.613	0.803	259	116
66	MP2	MK	STO-3G	0.869	0.795	0.634	0.808	265	124
67	BLYP	Hirshfeld	6-311G	0.868	0.807	0.627	0.821	245	116
68	HF	MK	STO-3G	0.867	0.8	0.641	0.813	261	124
69	BLYP	MK	6-311G	0.826	0.917	0.714	0.932	190	124
70	BP86	MK	6-311G	0.825	0.919	0.714	0.934	189	124
71	B3LYP	MK	6-311G	0.822	0.926	0.721	0.941	185	124
72	B3LYP	MK	STO-3G	0.817	0.939	0.749	0.955	179	124
73	BP86	MK	6-31G*	0.813	0.949	0.716	0.965	174	124
74	BLYP	MK	6-31G*	0.813	0.95	0.72	0.966	174	124
75	MP2	MK	6-311G	0.812	0.951	0.746	0.967	173	124
76	HF	MK	6-311G	0.811	0.954	0.747	0.970	172	124
77	B3LYP	MK	6-31G*	0.808	0.962	0.728	0.978	168	124
78	MP2	MK	6-31G*	0.788	1.011	0.773	1.028	149	124
79	HF	MK	6-31G*	0.788	1.012	0.773	1.029	149	124
80	BP86	MK	STO-3G	0.787	1.014	0.8	1.031	148	124
81	BLYP	MK	STO-3G	0.786	1.016	0.799	1.033	147	124
82	AM1	MK	–	0.447	1.633	1.247	1.660	32	124
83	PM3	MK	–	0.445	1.636	1.249	1.663	32	124

[a] The models are sorted according their R^2 (descending) and afterwards according their RMSE (ascending) and $\overline{\Delta}$ (ascending).

was performed for all 83 charge calculation approaches via MLR. The complete data set of 124 phenol molecules was used for the parametrization, and the obtained models were validated for all molecules in the data set. The only exceptions were the charge calculation procedures containing the Hirshfeld PA and basis sets 6-311G and STO-3G. In these cases, only 116 phenols were used for the parametrization and evaluation of the models, because 8 molecules from the original set were strong outliers. Table 1 contains the quality criteria (R^2, RMSE, and $\overline{\Delta}$) and the statistical criteria (s and F) for all the models. The models are sorted according to their quality. The parameters of the models are summarized in the Supporting Information (Table S2). The most relevant graphs of correlation between experimental and calculated pK_a are visualized in Figure 6. Tables 2–4 provide a clue for the comparison of QSPR models. Table 2 summarizes the R^2 of all models, Table 3 contains the average values of R^2 for all QSPR models which use a specific theory level, basis set, or PA, and Table 4 summarizes the quality of the charge calculation approaches which use a specific theory level, basis set or PA, and whose R^2 are in a certain interval.

The results provided in Tables 1–4 and Figure 6 lead us to the following conclusions regarding the relevance of the charge calculation method to the ability of the QSPR model to predict pK_a for phenolic compounds.

Comparison of All Models. All the presented models are statistically significant at $p = 0.01$. For the sets of 124 or 116 molecules, the models with three descriptors are statistically significant (at $p = 0.01$) when $F > 3.949$. The best models are MP2/6-31G*/Mulliken and HF/6-31G*/Mulliken ($R^2 = 0.966$, RMSE = 0.403, $\overline{\Delta} = 0.315$, $s = 0.410$, and $F = 1136$). More than 25% of the analyzed models (22 out of 83) have excellent quality and statistical criteria ($R^2 \geq 0.95$, RMSE ≤ 0.491, $\overline{\Delta} \leq 0.38$, $s \leq 0.5$, and $F \geq 760$), and more than 50% (47 out of 83) have very good statistical criteria ($R^2 > 0.9$, RMSE ≤ 0.698, $\overline{\Delta} \leq 0.54$, $s \leq 0.71$, and $F \geq 344$). About 80% of the models are able to predict pK_a with acceptable quality ($R^2 > 0.85$, RMSE ≤ 0.8, $\overline{\Delta} \leq 0.641$, $s \leq 0.813$, and $F \geq 261$). Only less than 20% of the models show a week correlation.

Influence of Theory Level. Ab initio theory levels: All five examined ab initio theory levels (MP2, HF, B3LYP, BLYP, and BP86) are applicable to pK_a prediction using charges. The best QSPR models are provided by MP2 and HF (models 1 and 2). Surprisingly, the differences between MP2 and HF were very small (illustrated by Tables 2–4). The pK_a values calculated from DFT charges have a slightly weaker correlation with experimental pK_a compared to MP2 and HF. The best performing DFT functional has been B3LYP, the models created by BLYP and BP86 have been less accurate.

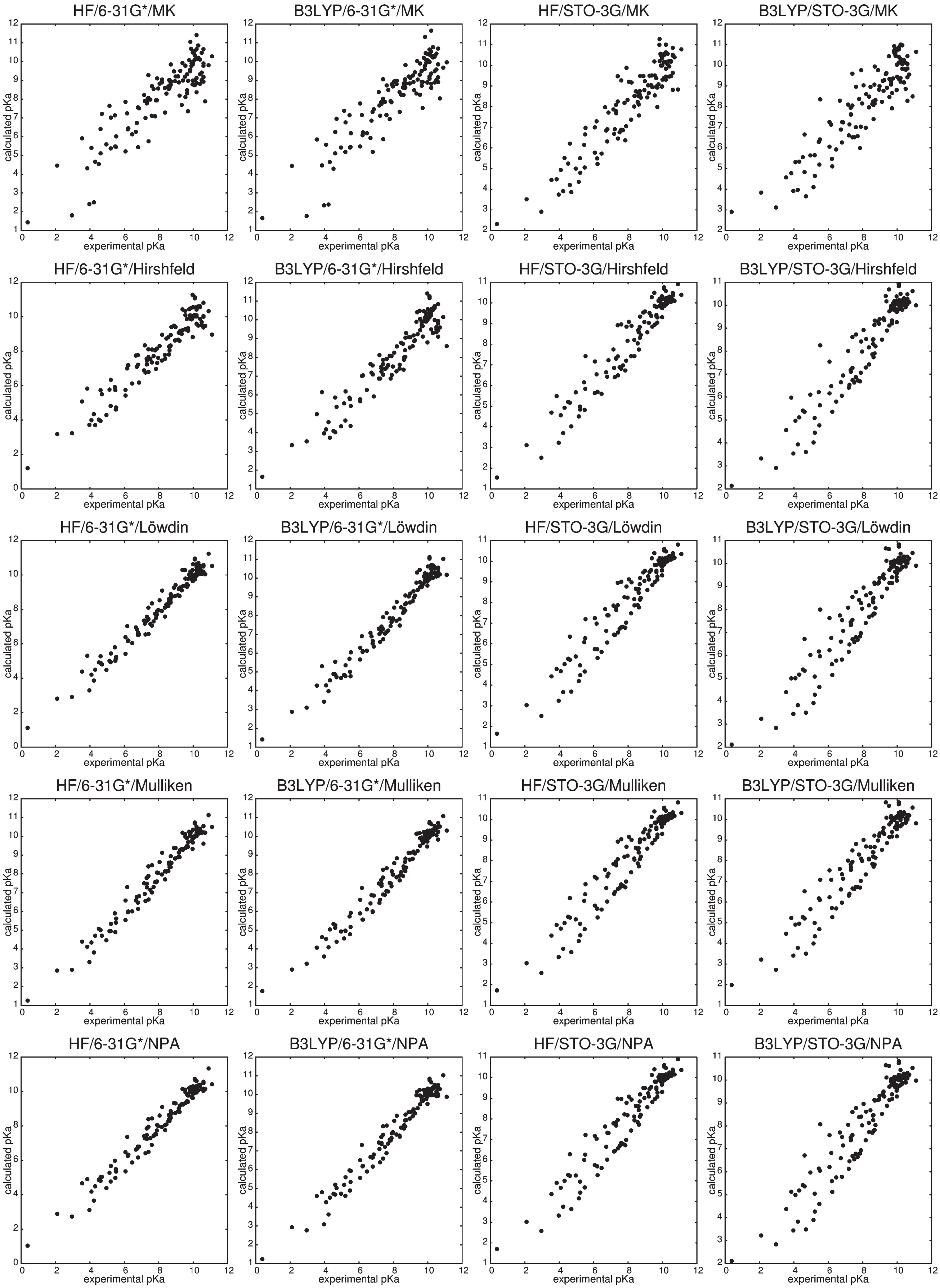


Figure 6. Graphs showing the correlation between experimental and calculated p*K*a for some selected charge calculation procedures.

Table 2. Squared Pearson Coefficients between Calculated and Experimental pK_a

R^2 for the basis set 6-31G*

Theory level	Population analysis				
	MK	Hir.	Löw.	MPA	NPA
BLYP	0.813	0.886	0.959	0.953	0.959
BP86	0.813	0.890	0.959	0.954	0.959
B3LYP	0.808	0.897	0.963	0.959	0.961
HF	0.788	0.908	0.966	0.966	0.963
MP2	0.788	0.912	0.966	0.966	0.964

R^2 for the basis set 6-311G

Theory level	Population analysis				
	MK	Hir.	Löw.	MPA	NPA
BLYP	0.826	0.868	0.926	0.932	0.960
BP86	0.825	0.874	0.931	0.939	0.959
B3LYP	0.822	0.882	0.937	0.938	0.962
HF	0.811	0.907	0.950	0.945	0.961
MP2	0.812	0.910	0.951	0.945	0.961

R^2 for the basis set STO-3G

Theory level	Population analysis				
	MK	Hir.	Löw.	MPA	NPA
BLYP	0.786	0.886	0.879	0.896	0.877
BP86	0.787	0.882	0.878	0.896	0.876
B3LYP	0.817	0.902	0.895	0.904	0.894
HF	0.867	0.928	0.920	0.920	0.920
MP2	0.869	0.929	0.921	0.922	0.921

Legend

	R^2	$RMSE$	$\bar{\Delta}$
excellent	0.950–0.970	0.40–0.50	0.32–0.38
very good	0.920–0.950	0.50–0.63	0.38–0.51
good	0.900–0.920	0.63–0.70	0.51–0.54
acceptable	0.850–0.900	0.70–0.80	0.54–0.64
weak	0.800–0.850	0.80–0.97	0.71–0.73

R^2 for semiempirical methods

Theory level	Population analysis				
	MK	Hir.	Löw.	MPA	NPA
AM1	0.447	0.917	0.924	0.924	–
PM3	0.445	0.895	0.880	0.880	–

Table 3. Average Squared Pearson Coefficients for All Charge Calculation Approaches Which Use a Specific Theory Level, PA, or Basis Set

theory level	BLYP	BP86	B3LYP	HF	MP2	AM1	PM3
average R^2	0.894	0.895	0.903	0.915	0.916	0.803	0.775

population analysis	MK	Hirshfeld	Löwdin	Mulliken	NPA
average R^2	0.772	0.898	0.930	0.932	0.940

basis set	6-31G*	6-311G	STO-3G
average R^2	0.917	0.909	0.887

Table 4. Percentage of Charge Calculation Approaches Which Use a Specific Theory Level, PA, or Basis Set and Whose Squared Pearson Coefficients Are in a Certain Interval[a]

theory level	interval	BLYP	BP86	B3LYP	HF	MP2	AM1	PM3
	$R^2 \geq 0.95$	27	27	29	33	33	0	0
	$0.95 > R^2 \geq 0.9$	13	13	29	47	47	75	0
	$0.9 > R^2 \geq 0.85$	40	40	21	7	7	0	75
	$R^2 < 0.85$	20	20	21	13	13	25	25

population analysis	interval	MK	Hirshfeld	Löwdin	Mulliken	NPA
	$R^2 \geq 0.95$	0	0	41	29	67
	$0.95 > R^2 \geq 0.9$	0	47	35	53	13
	$0.9 > R^2 \geq 0.85$	12	53	24	18	20
	$R^2 < 0.85$	88	0	0	0	0

basis set	interval	6-31G*	6-311G	STO-3G
	$R^2 \geq 0.95$	60	28	0
	$0.95 > R^2 \geq 0.9$	12	40	40
	$0.9 > R^2 \geq 0.85$	8	12	48
	$R^2 < 0.85$	20	20	12

[a] The percentages are calculated from total number of approaches with the defined theory level, basis set, or PA.

Semiempirical theory levels: The semiempirical approaches tested here provide weaker correlation than ab initio methods but are still applicable for pK_a prediction. The models with AM1 and Mulliken, Hirshfeld, or Löwdin PA show good correlation ($R^2 \geq 0.917$). The models using PM3 and Mulliken, Hirshfeld, or Löwdin PA also demonstrate acceptable correlation ($R^2 \geq 0.88$). The combination of semiempirical approaches with MK PA gives the worst models in this study (models 82 and 83).

Influence of Basis Set. The charges most appropriate for QSPR modeling of pK_a are provided by the 6-31G* basis set, and the accuracy of these models is very high. For example, the model with HF/6-31G*/Mulliken charges shows $R^2 = 0.966$. The results for the 6-311G basis set are slightly weaker. Unexpectedly, also the charges obtained using the STO-3G basis set are suitable for QSPR modeling, and the quality of these models is acceptable. For example, the model employing MP2/STO-3G/Hirshfeld charges exhibits $R^2 = 0.929$.

Influence of Population Analysis. Mulliken, natural, and Löwdin PAs with all levels of theory and basis sets provide the charges that are appropriate for pK_a prediction. The Hirshfeld PA with the STO-3G basis set provides results similar to the Mulliken, natural, or Löwdin PA with the same basis set. Nevertheless, the Hirshfeld PA with the 6-31G* or 6-311G basis sets lead to less accurate models than the above-mentioned population analyses employing these basis sets. Moreover, the occurrence of strong outliers complicates the application of the Hirshfeld PA. The charges calculated by the MK PA show only weak correlation with pK_a, and the QSPR models based on these charges have low accuracy, i.e., the best of such QSPR models employs HF/STO-3G/MK charges and shows $R^2 = 0.867$.

Table 5. Comparison of the Presented QSPR Models with Previous Work

theory level	PA	basis set	descriptors	R^2	s	F	number of molecules	source
B3LYP	NPA	6-311G**	q_{O-H}	0.789	1.300	48	15	Kreye and Seybold,[23, a]
B3LYP	NPA	6-311G**	q_O	0.731	1.500	38	15	Kreye and Seybold,[23, a]
B3LYP	NPA	6-31+G*	q_{O-H}	0.880	0.970	95	15	Kreye and Seybold,[23, b]
B3LYP	NPA	6-31+G*	q_O	0.865	1.000	38	15	Kreye and Seybold,[23, b]
B3LYP	NPA	6-311G(d,p)	q_{O^-}	0.911	0.252	173	19	Gross and Seybold[14]
B3LYP	NPA	6-311G(d,p)	q_H	0.887	0.283	134	19	Gross and Seybold[14]
B3LYP	NPA	6-31G*	q_H, q_O, q_{C1}	0.961	0.440	986	124	this work, model 10
B3LYP	NPA	6-311G	q_H, q_O, q_{C1}	0.962	0.435	1013	124	this work, model 8
B3LYP	MPA	6-311G(d,p)	q_H	0.913	0.248	179	19	Gross and Seybold[14]
B3LYP	MPA	6-311G(d,p)	q_{O^-}	0.894	0.274	144	19	Gross and Seybold[14]
B3LYP	MPA	6-311G	q_H, q_O, q_{C1}	0.938	0.556	605	124	this work, model 26
B3LYP	MPA	6-31G*	q_H, q_O, q_{C1}	0.959	0.450	936	124	this work, model 14
B3LYP	MK	6-311G(d,p)	q_H	0.344	0.682	9	19	Gross and Seybold[14]
B3LYP	MK	6-311G(d,p)	q_{O^-}	0.692	0.467	38	19	Gross and Seybold[14]
B3LYP	MK	6-311G	q_H, q_O, q_{C1}	0.822	0.941	185	124	this work, model 71
B3LYP	MK	6-31G*	q_H, q_O, q_{C1}	0.808	0.978	168	124	this work, model 77

[a] With solvent model SM5.4. [b] With solvent model SM8.

Table 6. Comparison of R^2 and RMSE for Test, Training, and Complete Sets for Model 2 (employing HF, Mulliken, 6-31G*) Charge Calculation Approaches

complete set				
R^2	RMSE	s	F	number of molecules
0.966	0.403	0.410	1136	124

cross validation										
	training set					test set				
cross-validation step	R^2	RMSE	s	F	number of molecules	R^2	RMSE	s	F	number of molecules
1	0.965	0.405	0.413	873	99	0.973	0.405	0.442	252	25
2	0.970	0.382	0.390	1024	99	0.930	0.489	0.534	93	25
3	0.964	0.415	0.424	848	99	0.977	0.357	0.390	297	25
4	0.967	0.394	0.402	928	99	0.966	0.444	0.484	199	25
5	0.968	0.403	0.411	968	100	0.957	0.442	0.484	148	24

Table 7. Comparison of R^2 and RMSE for Test, Training, and Complete Sets for Model 14 (employing B3LYP, Mulliken, 6-31G*) Charge Calculation Approaches

complete set				
R^2	RMSE	s	F	number of molecules
0.959	0.443	0.450	936	124

cross validation										
	training set					test set				
cross-validation step	R^2	RMSE	s	F	number of molecules	R^2	RMSE	s	F	number of molecules
1	0.958	0.441	0.450	722	99	0.963	0.452	0.493	182	25
2	0.962	0.434	0.443	802	99	0.925	0.509	0.555	86	25
3	0.955	0.462	0.472	672	99	0.975	0.358	0.391	273	25
4	0.961	0.425	0.434	780	99	0.956	0.516	0.563	152	25
5	0.962	0.435	0.444	810	100	0.950	0.506	0.554	127	24

Comparison with Previous Work. QSPR models similar to those presented in this paper were previously published by Gross and Seybold[14] and also by Kreye and Seybold.[23] Table 5 shows a comparison of these models with our models. It is seen therein that our models show markedly higher R^2 and F values, even for simpler basis sets. The reason may be that they employ more descriptors and were parametrized within a larger training set.

Cross-Validation. The robustness of the models was tested by cross-validation. The set of phenol molecules was divided into five parts (each contained 20% of the molecules). Afterward, five cross-validation steps were performed. In the first step, the first part was selected as a test set, and the remaining four parts were taken together as the training set. The test and training sets for the other steps were prepared in a similar manner by subsequently considering one part as a test set and the remaining parts served as a training set. For each step, the QSPR model was parametrized on the training set. Afterward, the pK_a values of the respective test molecules were calculated via this model and compared with experimental pK_a values. The cross-validation was performed for all 83 analyzed charge calculation approaches. The results are summarized in the Supporting Information (Table S3), and a part of these results is shown in Tables 6 and 7. The cross-validation showed that the models are stable, and the values of R^2 and RMSE are similar for the test, training, and complete sets.

■ CONCLUSION

The quantum chemical partial atomic charges have been shown to provide very good QSPR models for the estimation of pK_a. More than 25% of the analyzed models (22 out of 83) have excellent quality and statistical criteria (e.g., $R^2 \geq 0.95$), and more than 50% (47 out of 83) have very good statistical criteria (e.g., $R^2 > 0.9$). The descriptors used in the models we developed are the atomic charges of the hydrogen and oxygen from the phenolic OH group and the charge of the carbon binding to the OH group. Other atomic charges show only a weak correlation with pK_a. All seven examined theory levels (MP2, HF, B3LYP, BLYP, BP86, AM1, and PM3) are applicable to predicting pK_a from charges. The best results have been obtained for MP2 and HF. Utilizing DFT also provides good correlation. Semiempirical methods have generated weaker but acceptable models. The most suitable basis set was 6-31G*, while 6-311G provided slightly weaker correlations, and unexpectedly also the STO-3G basis set proved applicable to the QSPR modeling of pK_a. The Mulliken, natural, and Löwdin population analyses provided accurate models for all tested theory levels and basis sets. The Hirshfeld PA has been also useful, but the QSPR models based on MK charges showed only weak correlations. It is thus clear from our study that it is possible to predict pK_a values with very good accuracy using only partial atomic charges, and even unsophisticated theory levels and basis sets can provide good descriptors.

■ ASSOCIATED CONTENT

Ⓢ **Supporting Information.** List of phenol molecules employed in this study, including their experimental pK_a, the table of R^2 values for all charge calculation procedures and all descriptors (Table S1), the table with parameters of all QSPR models (Table S2) and the table of cross-validation results (Table S3). This material is available free of charge via the Internet at http://pubs.acs.org/.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: jkoca@chemi.muni.cz.

■ REFERENCES

(1) Wan, H.; Ulander, J. High-throughput pK_a screening and prediction amenable for ADME profiling. *Expert Opin. Drug Metab. Toxicol.* **2006**, *2*, 139–155.

(2) Clark, J.; Perrin, D. D. Prediction of the strengths of organic bases. *Q. Rev., Chem. Soc.* **1964**, *18*, 295–320.

(3) Perrin, D. D.; Dempsey, B.; Serjeant, E. P. *pKa Prediction for Organic Acids and Bases*; Chapman and Hall: New York, 1981.

(4) *ACD/pK_a*; Advanced Chemistry Development, Inc.: Toronto, Ontario, Canada, 2009.

(5) Shelley, J. C.; Cholleti, A.; Frye, L. L.; Greenwood, J. R.; Timlin, M. R.; Uchimaya, M.; Epik, M. A software program for pK_a prediction and protonation state generation for druglike molecules. *J. Comput.-Aided Mol. Des.* **2007**, *21*, 681–691.

(6) Hilal, S. H.; Karickhoff, S. W. A rigorous test for SPARC's chemical reactiity models: Estimation of more than 4300 ionization pK_as. *Quant. Struct.-Act. Relat.* **1995**, *14*, 348–355.

(7) Sayle, R. *Physiological ionization and pK_a prediction*; Metaphorics LLC: Santa Fe, NM, 2000; http://www.daylight.com/meetings/emug00/Sayle/pkapredict.html. Accessed January 24, 2011.

(8) Lee, A. C.; Crippen, G. M. Predicting pK_a. *J. Chem. Inf. Model.* **2009**, *49*, 2013–2033.

(9) *Jaguar*, version 4.2; Schrödinger, Inc.: New York, 2010.

(10) Liptak, M. D.; Gross, K. C.; Seybold, P. G.; Feldgus, S.; Shields, G. Absolute pK_a determinations for substituted phenols. *J. Am. Chem. Soc.* **2002**, *124*, 6421–6427.

(11) Barone, V.; Cossi, M. Quantum calculaton of molecular energies and energy gradients in solution by a conductor solvent model. *J. Phys. Chem. A* **1988**, *102*, 1995–2001.

(12) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Zakrzewski, V. G.; Montgomery, J. A., Jr.; Stratmann, R. E.; Burant, J. C.; Dapprich, S.; Millam, J. M.; Daniels, A. D.; Kudin, K. N.; Strain, M. C.; Farkas, O.; Tomasi, J.; Barone, V.; Cossi, M.; Cammi, R.; Mennucci, B.; Pomelli, C.; Adamo, C.; Clifford, S.; Ochterski, J.; Petersson, G. A.; Ayala, P. Y.; Cui, Q.; Morokuma, K.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Cioslowski, J.; Ortiz, J. V.; Baboul, A. G.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Gomperts, R.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Gonzalez, C.; Challacombe, M.; Gill, P. M. W.; Johnson, B. G.; Chen, W.; Wong, M. W.; Andres, J. L.; Head-Gordon, M.; Replogle, E. S.; Pople, J. A. *Gaussian 98*, revision A.6; Gaussian, Inc.: Pittsburgh, PA, 1998.

(13) Habibi-Yangjeh, A. Application of artificial neural networks for predicting the aqueous acidity of various phenols using QSAR. *J. Mol. Model.* **2006**, *12*, 338–347.

(14) Gross, K. C.; Seybold, P. G. Substituent effects on the physical properties and pK_a of phenol. *Int. J. Quantum Chem.* **2001**, *85*, 569–579.

(15) Brinck, T.; Murray, J. S.; Politzer, P. Molecular surface electrostatic potentials and local ionization energies of group V - VII hydrides and their anions: Relationships for aqueous and gas-phase acidities. *Int. J. Quantum Chem.* **1993**, *48*, 73–88.

(16) Citra, M. J. Estimating the pK_a of phenols, carboxylic acids and alcohols from semi-empirical quantum chemical methods. *Chemosphere* **1999**, *1*, 191–206.

(17) Dixon, S. L.; Jurs, P. C. Estimation of pK_a for organic oxyacids using calculated atomic charges. *J. Comput. Chem.* **1993**, *14*, 1460–1467.

(18) Parthasarathi, R.; Padmanabhan, J.; Elango, M.; Chitra, K.; Subramanian, V.; Chattaraj, P. K. pK_a prediction using group philicity. *J. Phys. Chem. A* **2006**, *110*, 6540–6544.

(19) Liu, S.; Pedersen, L. G. Estimation of molecular acidity via electrostatic potential at the nucleus and valence natural atomic orbitals. *J. Phys. Chem. A* **2009**, *113*, 3648–3655.

(20) Jelfs, S.; Ertl, P.; Selzer, P. Estimation of pK_a for druglike compounds using semiempirical and information-based descriptors. *J. Chem. Inf. Model.* **2007**, *47*, 450–459.

(21) Zhang, J.; Kleinöder, T.; Gasteiger, J. Prediction of pK_a values for aliphatic carboxylic acids and alcohols with empirical atomic charge descriptors. *J. Chem. Inf. Model.* **2006**, *46*, 2256–2266.

(22) Gross, K. C.; Seybold, P. G.; Hadad, C. M. Comparison of different atomic charge schemes for predicting pK_a variations in substituted anilines and phenols. *Int. J. Quantum Chem.* **2002**, *90*, 445–458.

(23) Kreye, W. C.; Seybold, P. G. Correlations between quantum chemical indices and the pK_as of a diverse set of organic phenols. *Int. J. Quantum Chem.* **2009**, *109*, 3679–3684.

(24) Xing, L.; Glen, R. C. Novel methods for the prediction of logP, pK_a, and logD. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 796–805.

(25) Soriano, E.; Cerdán, S.; Ballesteros, P. Computational determination of pK_a values. A comparison of different theoretical approaches and a novel procedure. *J. Mol. Struct.: THEOCHEM* **2004**, *684*, 121–128.

(26) Svobodová Vařeková, R.; Koča, J. Optimized and parallelized implementation of the electronegativity equalization method and the atom-bond electronegativity equalization method. *J. Comput. Chem.* **2006**, *3*, 396–405.

(27) *NCI Open Database Compounds*; National Cancer Institute, National Institutes of Health: Bethesda, MD; http://cactus.nci.nih.gov/. Accessed August 10, 2010.

(28) Sadowski, J.; Gasteiger, J. From atoms and bonds to three-dimensional atomic coordinates: Automatic model builders. *Chem. Rev.* **1993**, *93*, 2567–2581.

(29) Gieleciak, R.; Polanski, J. Modeling Robust QSAR. 2. Iterative Variable Elimination Schemes for CoMSA: Application for Modeling Benzoic Acid pK_a Values. *J. Chem. Inf. Model.* **2007**, *47*, 547–556.

(30) Howard, P.; Meylan, W. *Physical/chemical property database (PHYSPROP)*. Syracuse Research Corporation, Environmental Science Center: North Syracuse, NY, 1999.

(31) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03*, revision C.02; Gaussian, Inc.: Wallingford, CT, 2004.