HHS Public Access

Author manuscript

J Chem Inf Model. Author manuscript; available in PMC 2016 November 07.

Published in final edited form as:

J Chem Inf Model. 2015 June 22; 55(6): 1088–1097. doi:10.1021/ci500758w.

How Does the Methodology of 3D Structure Preparation Influence the Quality of pK_a Prediction?

Stanislav Geidl[†,‡], **Radka Svobodová Vařeková**[*,†,‡], **Veronika Bendová**[‡], **Lukáš Petrusek**[‡], **Crina-Maria Ionescu**[‡], **Zdeněk Jurka**[¶], **Ruben Abagyan**[§], 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, Faculty of Informatics, Masaryk University Brno, Botanická 68a, 602 00 Brno, Czech Republic, and Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California, 9500 Gilman Drive, MC 0657, San Diego, USA, svobodova@chemi.muni.cz; jkoca@chemi.muni.cz

Abstract

The acid dissociation constant is an important molecular property and it can be successfully predicted by Quantitative Structure-Property Relationship (QSPR) models, even for *in silico* designed molecules. We analyzed how the methodology of *in silico* 3D structure preparation influences the quality of QSPR models. Specifically, we evaluated and compared QSPR models based on six different 3D structure sources (DTP NCI, Pubchem, Balloon, Frog2, OpenBabel and RDKit) combined with four different types of optimization. These analyses were performed for three classes of molecules (phenols, carboxylic acids, anilines) and the QSPR model descriptors were quantum mechanical (QM) and empirical partial atomic charges. Specifically, we developed 516 QSPR models and afterwards systematically analyzed the influence of the 3D structure source and other factors on their quality.

Our results confirmed that QSPR models based on partial atomic charges are able to predict pK_a with high accuracy. We also confirmed that *ab-initio* and semiempirical QM charges provide very accurate QSPR models, and using empirical charges based on electronegativity equalization is also acceptable, as well as advantageous, since their calculation is very fast. On the other hand, Gasteiger-Marsili empirical charges are not applicable for pK_a prediction. We later found that QSPR models for some classes of molecules (carboxylic acids) are less accurate. In this context, we compared the influence of different 3D structure sources. We found that an appropriate selection of 3D structure source and optimization method is essential for the successful QSPR modeling of pK_a. Specifically, the 3D structures from the DTP NCI and Pubchem databases performed the best, as they provided very accurate QSPR models for all the tested molecular classes and charge calculation approaches, and they do not require optimization. Also Frog2 performed very well. Other 3D structure sources can also be used, but are not so robust, and an

[*]To whom correspondence should be addressed.

[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

[‡]NCBR & CEITEC

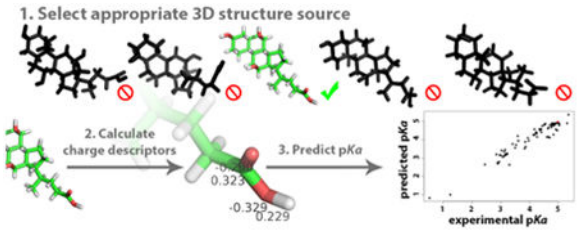
[¶]Faculty of Informatics

[§]Skaggs School of Pharmacy and Pharmaceutical Sciences

Supporting Information Available: This material is available free of charge via the Internet at http://pubs.acs.org/.

unfortunate combination of molecular class and charge calculation approach can produce weak QSPR models. Additionally, these 3D structures generally need optimization in order to produce good quality QSPR models.

Graphical abstract



Introduction

The acid dissociation constant, K_a, and its logarithmic version pK_a, is an important molecular property and its values are of interest in chemical, biological, environmental and pharmaceutical research.[1–3] Experimental pK_a values are usually unavailable for all compounds from the chemical catalogues. Therefore it cannot be used for example in virtual screening, which requires predictions of physico-chemical properties for large sets of *in silico* designed molecules. Several pK_a prediction methodologies have been published to date and they are summarized in review articles,[4–7] but reliable and accurate pK_a prediction is still a challenge and a topic of intensive research. [8–10]

A popular and frequently used pK_a prediction approach is based on the QSPR (Quantitative Structure-Property Relationship) methodology.[11–13] Various types of input values (so-called descriptors) can be used for the calculation of pK_a via QSPR models. Partial atomic charges are definitely relevant descriptors for pK_a calculations[12,14–17] and can be calculated directly from the 3D structure of the molecule. The partial atomic charges cannot be determined experimentally or derived from the results of quantum mechanics (QM) in a straightforward manner. For this reason, many different methods have been developed for their calculation. The most common method for charge calculation is using a quantum mechanical approach (a combination of a theory level and a basis set) and the subsequent application of a charge calculation scheme. For example for pK_a prediction via QSPR models, *ab-initio* QM charges calculated via HF or B3LYP theory levels and STO-3G or 6-31G* basis sets proved suitable. The most appropriate charge calculation schemes for these purposes seem to be MPA (Mulliken population analysis), NPA (Natural population analysis) and AIM (atoms in molecules).[8,15,17] Semiempirical QM charges have also been employed in QSPR models for pK_a prediction (e.g., AM1, PM3 or PM6 theory levels in combination with MPA).[11,14,17–19] A major drawback of the QM charges is the computational effort required for the calculation of the wave function. For this reason, the computational complexity of obtaining QM charges is at least $\Theta(B^4)$, where B is the number of basis functions. Therefore, the calculation of *ab-initio* QM charges is very time consuming, while the calculation of semiempirical QM charges is also relatively slow. The Electronegativity Equalization Method[20] is an empirical charge calculation approach which presents a faster alternative to

the QM methods. EEM is able to provide partial atomic charges with comparable accuracy to QM charges, and it is markedly less time consuming than QM charge calculation approaches. EEM is even able to mimic a certain QM charge calculation approach (i.e., the combination of a theory level, a basis set and a charge calculation scheme), because it includes parameters based on the QM charges. EEM charges also proved applicable for pK_a prediction via QSPR.[8] Last but not least, pK_a predicting QSPR models based on conformationally independent empirical charges (so called topological charges, e.g., Gasteiger-Marsili charges) have also been evaluated.[13,19]

Therefore, in principle, we can prepare a straightforward and time-efficient workflow for obtaining pK_a values for molecules designed *in silico*: use the 3D structures of molecules prepared *in silico*, calculate partial atomic charges for them, employ the charges as descriptors in QSPR models and predict the required pK_a values. Such a workflow can be applied in virtual screening. We can also design similar workflows for other biologically important properties such as log*P*, biodegradability, dioxin-like activity etc.

Nonetheless, before implementing the workflow we need to answer a key question: How does the methodology of *in silico* 3D structure preparation influence the quality of QSPR models for pK_a prediction? In previous works focused on pK_a prediction via QSPR,[8,17,19,21,22] 3D structures were mainly obtained from the DTP NCI database[23] (which uses CORINA to generate the 3D structures) or directly designed by CORINA.[24] But there are other tools and databases which are often used as sources of 3D structures. For example, the database Pubchem[25] (employing the software Omega[26]) or software tools such as Balloon,[27] Frog2,[28] OpenBabel[29] or RDKit.[30] These tools create 3D structures via a data or knowledge-based approach (CORINA, OpenBabel, Omega), distance geometry approach (Balloon, RDKit) or other approaches (Frog2). Specifically, Frog2 first generates a graph of rings and acyclic elements, and afterwards performs a Monte Carlo search. Can we use any of these 3D structure sources for the QSPR modeling of pK_a? Or is it that only some methodologies for 3D structure preparation provide acceptable QSPR models? In parallel, another important question is whether the 3D structures need to be optimized before they can be used in QSPR models or not. Some articles on this topic use optimization,[14,15,22,31,32] while some provide accurate models even without it.[8,11,17]

In this study, we addressed the above questions. Specifically, we evaluated and compared QSPR models based on six different 3D structure sources combined with four different types of optimization. The 3D structure sources were the databases DTP NCI and Pubchem, and the software tools Balloon, Frog2, OpenBabel and RDKit. The optimization was either skipped or done by molecular mechanics (MMFF94 for all 3D structure sources, MM-UFF for RDKit) or quantum mechanics (B3LYP/6-31G*). These analyses were performed for three classes of molecules (phenols, carboxylic acids, anilines). We mainly focused on *ab-initio* QM charges, which provide the most accurate pK_a predicting QSPR models, and on empirical EEM charges, which are a faster and comparably accurate alternative to *ab-initio* QM charges. Specifically, we used four types of QM charges (HF/STO-3G/MPA, B3LYP/6-31G*/MPA, B3LYP/6-31G*/NPA, and B3LYP/6-31G*/AIM) and four corresponding types of EEM charges. To create a complete overview, we provide also QSPR models based on semiempirical charges (i.e., PM6 charges) and on conformationally independent

empirical charges (i.e., Gasteiger-Marsili charges). Thus we developed 516 QSPR models, and afterwards systematically analyzed the influence of the 3D structure source and other factors on their quality.

Methods

Data sets

Our training data set is composed of three classes of molecules (i.e., phenols, anilines and carboxylic acids), which represent common classes of organic molecules. These types of molecules are also frequently used for the evaluation of QSPR models.[8,11,14–17,19,22,31] The data set contains 190 molecules: 60 phenols, 82 carboxylic acids and 48 anilines. Additionally, we used a test data set containing 53 phenols which were not included in the training data set. The list of molecules including their figures, NCS numbers and CAS numbers can be found in the Supporting Information (Table S1).

pK_a values

The experimental pK_a values were taken from the Physprop database.[33] The pK_a values of all molecules can be found in the Supporting Information (Table S1).

2D structure of molecules

Information about the 2D structure of individual molecules was obtained from the DTP NCI database. The 2D structures were described in SMILES format. The files with the SMILES of all molecules are in the Supporting Information.

Sources of 3D structure of molecules

For each molecule, the 3D structure was obtained from six different sources. Specifically, the structure was obtained from two databases (Pubchem, DTP NCI) and in parallel generated by four different freely available software tools (Balloon, Frog2, OpenBabel and RDKit). These sources were selected because they appear to be the most popular, and they also represent the main approaches for 3D structure preparation.

Optimization

Each molecule was thus associated with six different 3D structures, obtained by the six approaches described above. Afterwards, each 3D structure was processed in two different ways. Specifically, two types of optimization were performed – an optimization via quantum mechanics, and an optimization via molecular mechanics (MM). The QM optimization was performed by Gaussian 09[34] using B3LYP/6-31G*, and the MM optimization was done with RDKit using MMFF94. These approaches were selected because they are common and frequently used representatives of QM and MM optimization. Additionally, we also performed an optimization via the MM force field UFF (Universal Force Field) for structures prepared with RDKit. The reason is that the RDKit developers recommend applying this particular force field for the structures generated with RDKit.

3D structures in the training and test data sets

Each molecule in our training data set was associated with 19 different structures, because there were 6 sources of 3D structure and 3 types of optimization for each (no optimization, QM optimization and MM optimization) plus an additional UFF optimization for RDKit. The test data set contained only phenol molecules. Each molecule was associated with 2 different structures, because we selected 2 sources of 3D structure (i.e., DTP NCI and RDKit) and one type of optimization for each (no optimization).

In our QSPR models, we used neutral forms of all the molecules and also dissociated forms of phenols and carboxylic acids and associated forms of anilines (see Figure 1). The dissociated forms of molecules were created by removing the hydrogen atom of the dissociating group. The associated forms of anilines were created by adding one hydrogen atom to the amino group. The adding of the atom was done via an in-house script which applies the Bioshell library,[35,36] and a detailed description of the procedure is given in the Supporting information.

In this way, our training data set contained 19 (6*3+1) different structures for each molecule, and 7220 (=19*190*2) structures in total. In parallel, our test data set included 2 different structures for each molecule, therefore 212 (=2*53*2) structures in total.

QM charges

For each of the 7220 structures from the training set, we calculated *ab-initio* QM partial atomic charges via 4 QM charge calculation approaches (i.e., HF/STO-3G/MPA, B3LYP/6-31G*/MPA, B3LYP/6-31G*/NPA, and B3LYP/6-31G*/AIM) and semiempirical QM charges using PM6. These approaches were selected, because they represent the main types of charge calculation approaches which have been reported as successful for pK_a prediction via QSPR.[8,15,17] The second reason for selection of the *ab-initio* QM approaches was that corresponding EEM parameters are available for them. For each of the 212 structures from the test set, we calculated *ab-initio* QM charges via B3LYP/6-31G*/NPA. This charge calculation approach was selected based on the results obtained on the training set. All the *ab-initio* and semiempirical QM charges were calculated by Gaussian 09.[34]

EEM charges

For each of the 7220 structures in our dataset, the EEM charges were calculated by the program EEM SOLVER[37] using the 4 EEM parameter sets described in Table 1. EEM charges calculated using these parameter sets should mimic QM charges calculated by the relevant QM charge calculation approaches.

Gasteiger-Marsili charges

We calculated also empirical Gasteiger-Marsili charges for all the molecules from the training set, including their dissociated or associated forms, therefore for 380 (=2*190) molecules. Gasteiger-Marsili charges are based on 2D structure, therefore they do not depend on the source of 3D structure and on the optimization. All these charges were calculated by RD-Kit.[30]

Descriptors and QSPR models

The descriptors used for QSPR modeling were partial atomic charges from atoms that are close to the dissociation or association site. We employed both charges from neutral and from dissociated (or associated) molecules. The linear model is justified by the linear relationship between pK_a and the electrostatic potential at the protonation site combined with the linear dependence of the potential on the surrounding charges. The distance dependances are absorbed by the p coefficients derived from the experimental data.

Thus, the QSPR model employed in this study for phenol molecules has the following equation:

$$\mathrm{p}K_a = p_{p(H)} \cdot q_H + p_{p(O)} \cdot q_O + p_{p(C1)} \cdot q_{C1} + p_{p(OD)} \cdot q_{OD} + p_{p(C1D)} \cdot q_{C1D} + p_p \quad (1)$$

where q_H is the atomic charge of the hydrogen atom from the phenolic OH group of the neutral molecule, q_O is the charge on the oxygen atom from the phenolic OH group of the neutral molecule, q_{C1} is the charge on the carbon atom binding the phenolic OH group of the neutral molecule, q_{OD} is the charge on the phenoxide O^- from the dissociated molecule, and q_{C1D} is the charge on the carbon atom binding this oxygen in the dissociated molecule (see Figure 1 a)). The symbols $p_{p(H)}$, $p_{p(O)}$, $p_{p(C1)}$, $p_{p(OD)}$, $p_{p(C1D)}$ and p_p are parameters of the QSPR model.

The QSPR model employed in this study for carboxylic acids uses the following equation:

$$\mathrm{p}K_a = p_{c(H)} \cdot q_H + p_{c(O1)} \cdot q_{O1} + p_{c(O2)} \cdot q_{O2} + p_{c(C1)} \cdot q_{C1} + p_{c(O1D)} \cdot q_{O1D} + p_{c(O2D)} \cdot q_{O2D} + p_{c(C1D)} \cdot q_{C1D} + p_c$$

(2)

where q_H and q_{O1} are the atomic charge of the hydrogen and oxygen atoms from the OH group of the neutral molecule, respectively; q_{O2} is the charge on the oxygen atom from the carbonyl group of the neutral molecule; q_{C1} is the charge on the carbon atom binding in the COOH group of the neutral molecule; q_{O1D} is the charge on the O^- oxygen from the dissociated molecule; q_{O2D} is the charge on the oxygen atom from the carbonyl group of the dissociated molecule; and q_{C1D} is the charge on the carbon atom in the carboxyl group of the dissociated molecule (see Figure 1b)). Because the structures of dissociated carboxylic acid molecules were created by removing the H atom with no further correction of the structure, the values q_{O1D}, q_{O2D} and q_{C1D} describe charge distribution immediately after removing of this hydrogen atom. The symbols $p_{c(H)}$, $p_{c(O1)}$, $p_{c(O2)}$, $p_{c(C1)}$, $p_{c(O1D)}$, $p_{c(O2D)}$, $p_{c(C1D)}$, and p_c are parameters of the QSPR model.

The QSPR model employed in this study for anilines is based on the following equation:

$$\mathrm{p}K_a = p_{a(H)} \cdot q_H + p_{a(N)} \cdot q_N + p_{a(C1)} \cdot q_{C1} + p_{a(HA)} \cdot q_{HA} + p_{a(NA)} \cdot q_{NA} + p_{a(C1A)} \cdot q_{C1A} + p_a \quad (3)$$

where q_H is the average of charges located on both hydrogens in the amino group of the neutral molecule; q_N is the charge of the nitrogen from the amino group of the neutral molecule; q_{C1} is the charge on the carbon atom binding the amino group in the neutral molecule; q_{HA} is the average of charges located on the three hydrogens in the amino group of the associated molecule; q_{NA} is the charge on the nitrogen from the amino group of the associated molecule and q_{C1A} is the charge on the carbon atom binding the amino group in the associated molecule (see Figure 1 c)). The symbols $p_{a(H)}$, $p_{a(N)}$, $p_{a(C1)}$, $p_{a(HA)}$, $p_{a(NA)}$, $p_{a(C1A)}$, and p_a are parameters of the QSPR model.

The QSPR model equations (1) and (2) originate from,[8] and they proved useful for pK_a prediction based on QM and EEM charges. Equation (3) was inspired by these two equations.

In this way we created one QSPR model for each of our 3 classes of molecules (phenols, carboxylic acids, anilines), 19 types of structures (6 sources of 3D structures * 3 methods of optimization + RDKit with MM-UFF) and 9 types of charges (5 types of QM charges and 4 types of EEM charges). For each class of molecules, we additionally created one QSPR model based on Gasteiger-Marsili charges. Thus we created 516 (=3*19*9+3) QSPR models. Specifically, 228 QSPR models based on *ab-initio* QM charges (denoted QM QSPR models), 57 models based on semiempirical charges (denoted semiempirical QM QSPR models), 228 models based on EEM charges (denoted EEM QSPR models) and 3 models based on Gasteiger-Marsili charges (GM QSPR models). The parameterization of the QSPR models was done by multiple linear regression (MLR) using the software QSPR Designer.[42]

Cross-validation

The robustness of all 516 QSPR models was tested by cross-validation. The *k*-fold cross-validation procedure was used,[43,44] where $k = 5$. Specifically, for each QSPR model, its training data set was divided into five parts (each contained 20% of the molecules). This division was done randomly, and included stratification by pK_a value. Afterwards, 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 cross-validation steps were prepared in a similar manner.

Results and discussion

The quality of the QSPR models, i.e. the correlation between experimental pK_a and the pK_a calculated by each model, was evaluated using the squared Pearson correlation coefficient (R^2), root mean square error (RMSE), and average absolute pK_a error ($\bar{\Delta}$), while the statistical criteria were the standard deviation of the estimation (s) and Fisher's statistics of the regression (F).

Tables 2, 11 and S2 in Supporting Information summarize the squared Pearson correlation coefficients for all QSPR models based on QM charges (QM QSPR models) and for all QM QSPR models, EEM QSPR models and semiempirical QM QSPR models, respectively. Table S3 in the Supporting Information contains all the quality criteria (R^2, RMSE, $\bar{\Delta}$) and statistical criteria (s and F) for all the QSPR models analyzed. All these models are statistically significant at $p = 0.01$. Since our data sets contained 60 phenols, 82 carboxylic acids and 48 anilines, the appropriate F values to consider were those for 60 samples, 80 samples and 50 samples, respectively. The QSPR models for phenols, carboxylic acids and anilines contained 5, 7 and 6 descriptors, respectively. Thus, the QSPR models for phenols are statistically significant (at $p = 0.01$) when $F > 3.34$, the QSPR models for carboxylic acids when $F > 2.87$ and the QSPR models for anilines when $F > 3.19$.

The parameters of the QSPR models are summarized in the Supporting Information (Table S4).

Quality of QM QSPR models – general summary

The results summarized in Tables 2 and 3 confirmed that the QSPR models based on QM charges are able to predict pK_a with high accuracy. Specifically, about 24% of the models have excellent quality ($R^2 \geq 0.95$), close to 40% have very good quality ($R^2 \geq 0.9$), 30% have lower quality, but are still applicable ($R^2 \geq 0.8$), and only about 6% have low quality ($R^2 < 0.8$).

Predictivity of QM QSPR models

In general, the predictivity of QSPR models calculating pK_a based on charges was shown in the literature (e.g.[11–13]). Additionally, high quality of QM QSPR models based on the same charge descriptors as our models was shown by Svobodová Vařeková et al..[17] To confirm the predictivity, we did a cross-validation for all our QSPR models. Cross-validation results for selected QSPR models are in Table 4 (i.e., based on B3LYP/6-31G*/NPA charges and non-optimized OpenBabel 3D structures, which show average quality in comparison with other QM QSPR models). All the cross-validation results can be found in the Supporting Information (Table S5). These results showed that the values of R^2 are similar for the test set, the training set and the complete set, therefore the models are stable.

For further confirmation of our QSPR models predictivity, we tested selected QSPR models on an independent test data set prepared only for testing purposes, with a size comparable to that of training data set and which was. Specifically, the test data set includes 53 phenol molecules and we used it for testing two selected QM QSPR models for phenols, namely, one of the best quality models (B3LYP/6-31G*/NPA charges and non-optimized 3D structures from NCI) and one of the worst quality models (HF/STO-3G/MPA charges and non-optimized 3D structures from RDKit). The quality criteria for the test set and the training set are in Table 5. These results demonstrate that the QSPR models perform comparably for the test set and the training set.

Influence of *ab-initio* QM charge calculation approach

The results (Tables 2 and 6) show that all four of the *ab-initio* QM charge calculation approaches tested here provide a comparable quality of pK_a prediction. These results therefore confirmed, that all the selected charge calculation approaches are suitable for the QSPR prediction of pK_a. Additionally, all the charge calculation approaches are applicable for all three classes of molecules. Specifically, for each class of molecules, any *ab-initio* QM charge calculation approach provides good quality QSPR models (R^2 close to 0.9) at least for some sources of 3D structures. An interesting finding is that the suitability of a certain charge calculation approach strongly depends on the class of molecules. For example, B3LYP/6-31G*/MPA charges work very well for anilines and markedly poorer for carboxylic acids. The next interesting finding is that the charge calculation approach HF/STO-3G/MPA, which uses the smallest basis set (STO-3G) and the simplest population analysis (MPA), performs very well.

Influence of the class of molecules

We can see (Table 2 and Table 7), that some classes of molecules are more easily handled by QSPR modeling, while some are more challenging. Specifically, QSPR models work very well for anilines and phenols. These models have high R^2 for all charge calculation approaches and for most of the 3D structure sources. On the other hand, QSPR models provide markedly weaker pK_a predictions for carboxylic acids. Namely, only a few 3D structure sources are applicable for QSPR modeling for carboxylic acids. One reason for the lower quality of QSPR models for the carboxylic acids is, that the carboxyl group bound some arbitrary chemical scaffold. In contrast, the –OH group of phenols and $–NH_2$ group of anilines have the same, conserved neighborhood – the phenolic ring. In parallel, the phenolic ring also allows higher de-localization of electrons, which is better suited for the calculation of QM descriptors than the more rigid electron localization in carboxylic acids.

Influence of 3D structure preparation methodology on the quality of the QM QSPR model

Tables 2, 8 and 9 show that an appropriate selection of 3D structure source and optimization method is essential for the QSPR modeling of pK_a.

These results imply that the most appropriate 3D structures were obtained from the DTP NCI and Pubchem databases (i.e., structures prepared with the tools CORINA and Omega, respectively). The QSPR models based on these structures are very accurate, and these 3D structures do not require optimization. A great feature of these 3D structures was that they performed very well for all the tested QM charge calculation approaches and classes of molecules. An interesting finding is that the QM optimization of such 3D structures can markedly decrease the accuracy of the models.

Frog2 also seems to be applicable. QSPR models based on 3D structures from Frog2 are accurate even when the structures were not optimized, and the MM optimization of these structures mainly improves the models. They can be successfully used for all the classes of molecules and all the QM charge calculation approaches tested here.

RDKit, OpenBabel and Balloon are slightly troublesome sources of 3D structures. They can provide accurate QSPR models ($R^2 > 0.9$) for some classes of molecules. In this case, the MM optimization of 3D structures improves the models. But when we process other classes of molecules (carboxylic acids), the QSPR models are weak ($R^2 \sim 0.85$) for most of the charge calculation approaches. And for certain charge calculation approaches the QSPR models can even be unsatisfactory ($R^2 < 0.7$). An interesting fact is that the structures generated by RDKit with no optimization provide the worst performing QSPR models of the whole study. The explanation is clear, these 3D structures are just the raw results of RDKit and, as mentioned in its manual, they need to be optimized by RDKit's internal force field UFF. This case study shows how weak QSPR models can be when based on problematic structures.

Particular geometrical properties, which are incorrectly modelled in certain 3D structure preparation methodologies and which cause worse performance of QSPR models are summarized in Supporting Information.

Semiempirical QM QSPR models – quality, predictivity and influences

The results summarized in Table 10 and Supplementary Table S2 show that the quality of these models is comparable to the quality of QSPR models based on ab-initio QM charges, just slightly lower for phenols and anilines and slightly better for carboxylic acids. The cross-validation results (see Supplementary Table S5) confirmed the robustness of the semiempirical QM models. When we evaluated the influence of the class of molecules and the 3D structure preparation methodology, we saw the same trends as for the *ab-initio* QM QSPR models (see Table 10 and S2).

Quality of EEM QSPR models – general summary

The results summarized in Tables 11 and 12 show that the quality of EEM QSPR models is in general lower than for QM QSPR models, but still sufficient. Specifically, about 36% of the models are very good quality ($R^2 \geq 0.9$), most of the models are acceptable quality (R^2 between 0.9 and 0.8) and only about 2% are low quality ($R^2 < 0.8$). On the other hand, the number of weak models is lower than for QM QSPR models, and there are no models with ($R^2 < 0.75$).

Predictivity of EEM QSPR models

A high quality of EEM QSPR models based on the same charge descriptors as our models was shown in.[8] We tested the predictivity of our EEM QSPR models the same way as we did for the QM QSPR models – by cross-validation and by testing on a larger set of independent molecules. These results are summarized in Supporting Information (Table S5 and S6, respectively), and confirm that our EEM QSPR models are robust and can handle molecules outside the training set.

Influence of EEM parameter set

The results (Table 11 and Supplementary Table S7) show that all four EEM parameter sets tested here are applicable for pK_a prediction. The quality of the QSPR models obtained by

all the EEM parameter sets is comparable. The parameter set Chaves2006 (mimicking B3LYP/6-31G*/MPA charges) performed slightly better than the remaining sets.

Influence of the class of molecules

As with QM charges, some classes of molecules are more challenging for the QSPR modeling of pK_a (carboxylic acids), see Table 11 and Supplementary Table S8. Nonetheless, the differences between the quality of EEM QSPR models for various classes of molecules are markedly smaller than for the QM QSPR models.

Influence of 3D structure preparation methodology on the quality of the EEM QSPR model

Table 8 and Supplementary Table S6 show that EEM QSPR models are markedly less sensitive to the selection of 3D structure source and optimization method.

As with QM QSPR models, 3D structures from DTP NCI and Pubchem can be successfully used for all of the tested molecular classes and all EEM parameter sets, even without optimization (i.e., more than 90% of EEM QSPR models based on non-optimized NCI 3D structures and all EEM QSPR models based on non-optimized Pubchem 3D structures have $R^2 > 0.85$).

Frog2 also preforms very well. More than 80% of EEM QSPR models based on non-optimized Frog2 3D structures have $R^2 > 0.85$. Additionally, these models seem to be applicable for all molecular classes and all EEM parameter sets tested here.

For the other four tools, the accuracy of EEM QSPR models depends on the molecular class and EEM parameter set, as certain combinations of these can produce lower accuracy QSPR models.

For all six sources of 3D structures tested in this study, QM optimization produces an improvement in the EEM QSPR models in most cases.

Quality of GM QSPR models

Gasteiger-Marsili charges does not depend on the 3D structure of molecules, therefore we prepared only one model for each class of molecules. The R^2 values of these models are given in Table 13 and further quality criteria are available in Supplementary Table S3. These results show that GM QSPR models are markedly less accurate than EEM QSPR models and therefore, GM charges are not applicable for pK_a prediction. These conclusions are in agreement with results published in the past.[15]

Conclusion

Our results confirmed that QSPR models based on QM and EEM partial atomic charges are are able to predict pK_a with high accuracy. Specifically, more than 60% of *ab-initio* and semiempirical QM QSPR models and nearly 40% of EEM QSPR models are very good quality ($R^2 \geq 0.9$). We also confirmed that *ab-initio* and semiempirical QM charges provide very accurate QSPR models and using EEM charges is also acceptable, and moreover advantageous because their calculation is very fast. Afterwards, we evaluated the predictivity

of our QM, semiempirical QM and EEM QSPR models via cross-validation and via testing on an independent test data set. This way, we verified that all the types of *ab-initio* and semiempirical and EEM charges used are applicable for QSPR modeling. On the contrary, QSPR models based on empirical Gasteiger-Marsili charges showed low quality, suggesting that Gasteiger-Marsili charges are not suitable descriptors for the prediction of pK_a.

We then focused on the influence of molecular class. We found that some molecular classes are more amenable to QSPR modeling (phenols and anilines), while some are more challenging (carboxylic acids).

In this context, we compared the influence of the different 3D structure sources. We found that the selection of 3D structure source and optimization method can strongly influence the quality of QSPR models for pK_a prediction. The 3D structures from the DTP NCI and Pubchem databases, i.e. structures generated by CORINA and Omega, respectively, exhibited the best performance. These 3D structures provided very accurate QSPR models for all the tested molecular classes and charge calculation approaches, and they do not require optimization. Frog2 also performed very well for all of the tested molecular classes and charge calculation approaches. Other 3D structure sources can also be used, but they are not so robust, and an unlucky combination of molecular class and charge calculation approach can lead to weak QSPR models. Additionally, these structures generally need to be optimized in order to produce high quality QSPR models. Specifically, the best approach is to apply MM optimization to 3D structures used with QM QSPR models, and QM optimization to 3D structures used with EEM QSPR models.

The main point of this article is that a workflow for the fast and accurate prediction of pK_a or other important properties for *in silico* designed molecules can be as follows: Preparation of 3D structures by CORINA or Omega (with no further optimization), calculation of EEM charges for these structures and then the EEM QSPR calculation of pK_a.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by the Ministry of Education, Youth and Sports of the Czech Republic (LH13055); the European Community's Seventh Framework Programme (CZ.1.05/1.1.00/02.0068) from the European Regional Development Fund and the Capacities specific program (286154); and by the European Social Fund and the state budget of the Czech Republic (CZ.1.07/2.3.00/20.0042, CZ.1.07/2.3.00/30.0009).

This work was also supported in part by NIH grants R01 GM071872, U01 GM094612, and U54 GM094618 to R.A.. The access to MetaCentrum supercomputing facilities provided under the research intent MSM6383917201 is greatly appreciated.

References

1. Comer, J.; Tam, K. Pharmacokinetic Optimization in Drug Research: Biological, Physicochemical, and Computational Strategies. Verlag Helvetica Chimica Acta, Post-fach; CH-8042 Zürich, Switzerland: 2001.

2. Klebe G. Recent Developments in Structure-Based Drug Design. J Mol Med. 2000; 78:269–281. [PubMed: 10954199]
3. Kim JH, Gramatica P, Kim MG, Kim D, Tratnyek PG. QSAR Modelling of Water Quality Indices of Alkylphenol Pollutants. SAR QSAR Environ Res. 2007; 18:729–743. [PubMed: 18038370]
4. Lee AC, Crippen GM. Predicting pK_a. J Chem Inf Model. 2009; 49:2013–2033. [PubMed: 19702243]
5. Rupp M, Körner R, Tetko IV. Predicting the pK_a of Small Molecules. Comb Chem High Throughput Screen. 2010; 14:307–327.
6. Ho J. Predicting pK_a in Implicit Solvents: Current Status and Future Directions. Aust J Chem. 2014; 67:1441–1460.
7. Balogh GT, Tarcsay Á, Keserű GM. Comparative Evaluation of pK_a Prediction Tools on a Drug Discovery Dataset. J Pharm Biomed Anal. 2012; 6768:63–70.
8. Svobodová Vařeková R, Geidl S, Ionescu CM, Skřehota O, Bouchal T, Sehnal D, Abagyan R, Koča J. Predicting pK_a Values From EEM Atomic Charges. J Cheminf. 2013; 5:18–34.
9. Fraczkiewicz R, Lobell M, Gller AH, Krenz U, Schoenneis R, Clark RD, Hillisch A. Best of Both Worlds: Combining Pharma Data and State of the Art Modeling Technology to Improve *in silico* pK_a Prediction. J Chem Inf Model. 2015; 55:389–397. [PubMed: 25514239]
10. Settimo L, Bellman K, Knegtel RMA. Comparison of the Accuracy of Experimental and Predicted pK_a Values of Basic and Acidic Compounds. Pharmaceut Res. 2014; 31:1082–1095.
11. 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. [PubMed: 17381168]
12. Dixon SL, Jurs PC. Estimation of pK_a for Organic Oxyacids Using Calculated Atomic Charges. J Comput Chem. 1993; 14:1460–1467.
13. 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. [PubMed: 17125168]
14. Citra MJ. Estimating the pK_a of Phenols, Carboxylic Acids and Alcohols From Semi-empirical Quantum Chemical Methods. Chemosphere. 1999; 1:191–206.
15. Gross KC, Seybold PG, Hadad CM. Comparison of Different Atomic Charge Schemes for Predicting pK_a Variations in Substituted Anilines and Phenols. Int J Quantum Chem. 2002; 90:445–458.
16. Kreye WC, Seybold PG. Correlations Between Quantum Chemical Indices and the pK_as of a Diverse Set of Organic Phenols. Int J Quantum Chem. 2009; 109:3679–3684.
17. Svobodová Vařeková R, Geidl S, Ionescu CM, Skřehota O, Kudera M, Sehnal D, Bouchal T, Abagyan R, Huber HJ, Koča J. Predicting pK_a Values of Substituted Phenols From Atomic Charges: Comparison of Different Quantum Mechanical Methods and Charge Distribution Schemes. J Chem Inf Model. 2011; 51:1795–1806. [PubMed: 21761919]
18. Rayne, S.; Forest, K.; Friesen, K. Examining the PM6 Semiempirical Method for pK_a Prediction Across a Wide Range of Oxyacids. Available from Nature Precedings. http://hdl.handle.net/10101/npre.2009.2981.1
19. 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. [PubMed: 17381172]
20. Mortier WJ, Ghosh SK, Shankar S. Electronegativity Equalization Method for the Calculation of Atomic Charges in Molecules. J Am Chem Soc. 1986; 108:4315–4320.
21. Czodrowski P, Dramburg I, Sotriffer CA, Klebe G. Development, Validation, and Application of Adapted PEOE Charges to Estimate pK_a Values of Functional Groups in Protein–Ligand Complexes. Proteins Struct Funct Bioinf. 2006; 65:424–437.
22. Tehan BG, Lloyd EJ, Wong MG, Pitt WR, Montana JG, Manallack DT, Gancia E. Estimation of pK_a Using Semiempirical Molecular Orbital Methods. Part 1: Application to Phenols and Carboxylic Acids. Quant Struct-Act Relat. 2002; 21:457–472.
23. NCI Open Database Compounds. Retrieved from http://cactus.nci.nih.gov/ on August 10, 2010

24. Sadowski J, Gasteiger J. From Atoms and Bonds to Three-Dimensional Atomic Coordinates: Automatic Model Builders. Chem ReV. 1993; 93:2567–2581.
25. Bolton, EE.; Wang, Y.; Thiessen, PA.; Bryant, SH. In Annual Reports in Computational Chemistry. Wheeler, R.; Spellmeyer, D., editors. Vol. 4; Chapter 12. Elsevier; 2008.
26. Hawkins PCD, Skillman AG, Warren GL, Ellingson BA, Stahl MT. Conformer Generation with OMEGA: Algorithm and Validation Using High Quality Structures from the Protein Databank and Cambridge Structural Database. J Chem Inf Model. 2010; 50:572–584. [PubMed: 20235588]
27. Vainio MJ, Johnson MS. Generating Conformer Ensembles Using a Multiobjective Genetic Algorithm. J Chem Inf Model. 2007; 47:2462–2474. [PubMed: 17892278]
28. Leite TB, Gomes D, Miteva M, Chomilier J, Villoutreix B, Tuffry P. Frog: a FRee Online DruG 3D Conformation Generator. Nucleic Acids Res. 2007; 35:W568–W572. [PubMed: 17485475]
29. O'Boyle N, Banck M, James C, Morley C, Vandermeersch T, Hutchison G. Open Babel: An Open Chemical Toolbox. J Cheminf. 2011; 3:33–47.
30. Landrum, G. RDKit: Open-Source Cheminformatics. Retrieved from http://www.rdkit.org on January 10, 2014
31. Gross KC, Seybold PG. Substituent Effects on the Physical Properties and pK_a of Phenol. Int J Quantum Chem. 2001; 85:569–579.
32. Habibi-Yangjeh A, Danandeh-Jenagharad M, Nooshyar M. Application of Artificial Neural Networks for Predicting the Aqueous Acidity of Various Phenols Using QSAR. J Mol Model. 2006; 12:338–347. [PubMed: 16344950]
33. Howard, P.; Meylan, W. Physical/Chemical Property Database (PHYSPROP). Syracuse Research Corporation, Environmental Science Center; North Syracuse NY: 1999.
34. Frisch, MJ.; Trucks, GW.; Schlegel, HB.; Scuseria, GE.; Robb, MA.; Cheeseman, JR.; Montgomery, JA., Jr; Vreven, T.; Kudin, KN.; Burant, JC.; Millam, JM.; Iyengar, SS.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, GA.; 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, JE.; Hratchian, HP.; Cross, JB.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, RE.; Yazyev, O.; Austin, AJ.; Cammi, R.; Pomelli, C.; Ochterski, JW.; Ayala, PY.; Morokuma, K.; Voth, GA.; Salvador, P.; Dannenberg, JJ.; Zakrzewski, VG.; Dapprich, S.; Daniels, AD.; Strain, MC.; Farkas, O.; Malick, DK.; Rabuck, AD.; Raghavachari, K.; Foresman, JB.; Ortiz, JV.; Cui, Q.; Baboul, AG.; Clifford, S.; Cioslowski, J.; Stefanov, BB.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, RL.; Fox, DJ.; Keith, T.; Al-Laham, MA.; Peng, CY.; Nanayakkara, A.; Challacombe, M.; Gill, PMW.; Johnson, B.; Chen, W.; Wong, MW.; Gonzalez, C.; Pople, JA. Gaussian 09, Revision E.01. Gaussian, Inc.; Wallingford, CT: 2004.
35. Gront D, Kolinski A. BioShell – a Package of Tools for Structural Biology Computations. Bioinformatics. 2006; 22:621–622. [PubMed: 16407320]
36. Gront D, Kolinski A. Utility Library for Structural. Bioinformatics. 2008; 24:584–585. [PubMed: 18227118]
37. 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.
38. Svobodová Vařeková R, Jiroušková Z, Vaněk J, Suchomel S, Koča J. Electronegativity Equalization Method: Parameterization and Validation for Large Sets of Organic, Organohalogene and Organometal Molecule. Int J Mol Sci. 2007; 8:572–582.
39. Chaves J, Barroso JM, Bultinck P, Carbo-Dorca R. Toward an Alternative Hardness Kernel Matrix Structure in the Electronegativity Equalization Method (EEM). J Chem Inf Model. 2006; 46:1657–1665. [PubMed: 16859297]
40. Bultinck P, Langenaeker W, Lahorte P, De Proft F, Geerlings P, Van Alsenoy C, Tollenaere JP. The Electronegativity Equalization Method II: Applicability of Different Atomic Charge Schemes. J Phys Chem A. 2002; 106:7895–7901.
41. Bultinck P, Vanholme R, Popelier PLA, De Proft F, Geerlings P. High-speed Calculation of AIM Charges Through the Electronegativity Equalization Method. J Phys Chem A. 2004; 108:10359–10366.

42. Skřehota O, Svobodová Vařeková R, Geidl S, Kudera M, Sehnal D, Ionescu CM, Koča J. QSPR Designer – a Program to Design and Evaluate QSPR models. Case Study on pK_a Prediction. J Cheminf. 2011; 3(Suppl 1):16.

43. Lemm S, Blankertz B, Dickhaus T, Müller KR. Introduction to Machine Learning for Brain Imaging. NeuroImage. 2011; 56:387–399. [PubMed: 21172442]

44. Katritzky AR, Lobanov VS, Karelson M. QSPR: The Correlation and Quantitative Prediction of Chemical and Physical Properties From Structure. Chem Soc Rev. 1995; 24:279–287.

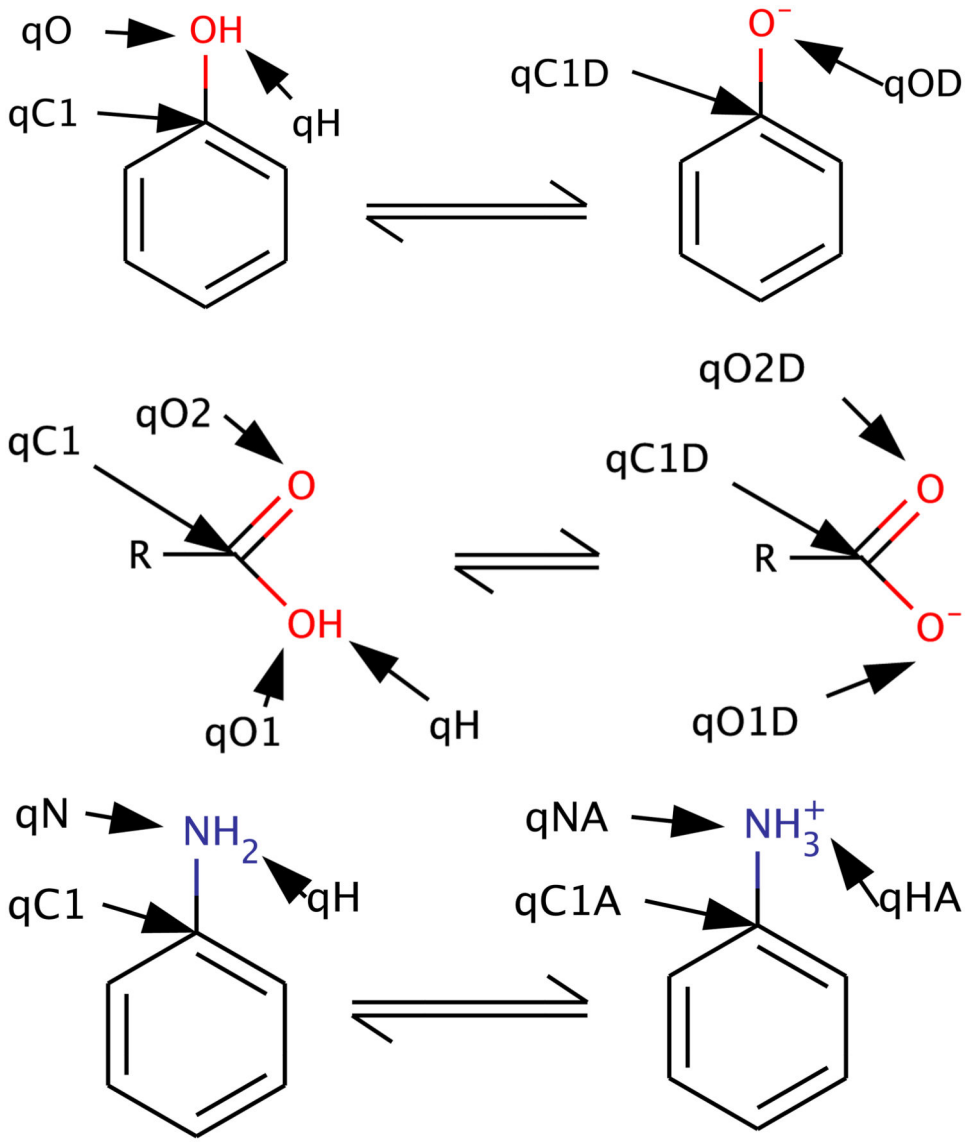


Figure 1.

a) dissociation of phenols, b) dissociation of carboxylic acids and c) association of anilines. The particular atomic charges used in our QSPR models are marked by their denotations.

Table 1

Summary information about the EEM parameter sets used in this study.

Parameter set name	QM charge calculation approach	Published by
Svob2007_chal2	HF/STO-3G/MPA	Svobodova et al.[38]
Chaves2006	B3LYP/6-31G*/MPA	Chaves et al.[39]
Bult2002_npa	B3LYP/6-31G*/NPA	Bultinck et al.[40]
Bult2004_aim	B3LYP/6-31G*/AIM	Bultinck et al.[41]

Table 2

R^2 describing the correlation between calculated and experimental pK_a for QM QSPR models.

R^2		Class of molecules	Phenols				Carboxylic acids				Anilines				Average
		Charge calculation approach	HF, STO-3G, MPA	B3LYP, 6-31G*, MPA	B3LYP, 6-31G*, NPA	B3LYP, 6-31G*, AIM	HF, STO-3G, MPA	B3LYP, 6-31G*, MPA	B3LYP, 6-31G*, NPA	B3LYP, 6-31G*, AIM	HF, STO-3G, MPA	B3LYP, 6-31G*, MPA	B3LYP, 6-31G*, NPA	B3LYP, 6-31G*, AIM	
Source + Optimization	Balloon	none	0.896	0.939	0.908	0.904	0.823	0.720	0.819	0.846	0.836	0.903	0.912	0.805	0.859
		MM	0.917	0.881	0.933	0.891	0.867	0.587	0.805	0.843	0.874	0.953	0.927	0.921	0.867
		QM	0.915	0.871	0.901	0.856	0.890	0.618	0.824	0.807	0.948	0.967	0.933	0.921	0.871
	Frog2	none	0.894	0.912	0.906	0.891	0.896	0.876	0.876	0.884	0.934	0.911	0.924	0.916	0.902
		MM	0.967	0.931	0.907	0.938	0.907	0.830	0.903	0.922	0.958	0.973	0.965	0.926	0.927
		QM	0.969	0.963	0.953	0.939	0.917	0.853	0.906	0.917	0.875	0.973	0.911	0.853	0.919
	NCI	none	0.947	0.971	0.960	0.973	0.931	0.891	0.911	0.910	0.951	0.970	0.966	0.903	0.940
		MM	0.958	0.963	0.959	0.936	0.938	0.889	0.929	0.922	0.954	0.955	0.967	0.914	0.940
		QM	0.891	0.935	0.861	0.902	0.925	0.854	0.903	0.921	0.942	0.959	0.937	0.892	0.910
	OpenBabel	none	0.955	0.961	0.957	0.963	0.869	0.658	0.845	0.876	0.952	0.973	0.966	0.930	0.909
		MM	0.961	0.965	0.959	0.961	0.863	0.665	0.841	0.875	0.958	0.975	0.967	0.927	0.910
		QM	0.955	0.957	0.956	0.936	0.845	0.674	0.804	0.827	0.874	0.974	0.928	0.880	0.884
	PubChem	none	0.960	0.950	0.935	0.900	0.909	0.873	0.891	0.907	0.938	0.939	0.921	0.937	0.922
		MM	0.963	0.911	0.927	0.864	0.916	0.885	0.892	0.916	0.942	0.979	0.966	0.916	0.923
		QM	0.943	0.936	0.922	0.886	0.901	0.871	0.896	0.908	0.934	0.974	0.885	0.828	0.907
	RDKit	none	0.782	0.895	0.796	0.882	0.780	0.723	0.804	0.817	0.853	0.816	0.851	0.796	0.816
		MM-UFF	0.947	0.961	0.941	0.934	0.894	0.821	0.842	0.860	0.965	0.979	0.973	0.980	0.925
		MM	0.931	0.909	0.934	0.950	0.902	0.750	0.797	0.862	0.959	0.976	0.967	0.927	0.905
		QM	0.935	0.944	0.933	0.922	0.861	0.696	0.814	0.855	0.940	0.964	0.927	0.908	0.892
Average			0.931	0.934	0.924	0.917	0.886	0.776	0.858	0.878	0.926	0.953	0.936	0.899	
Legend			$R^2 \geq 0.95$		$R^2 \geq 0.9$		$R^2 \geq 0.866$		$R^2 \geq 0.833$		$R^2 \geq 0.8$		$R^2 \geq 0.7$		$R^2 < 0.7$

Table 3

Number and percentage of QM QSPR models with R^2 higher than a defined limit.

R^2	≥ 0.95	(0.95, 0.9>	(0.9, 0.8>	< 0.8
Number of models	55	90	69	14
Percentage of models	24%	39%	30%	6%

Table 4

R^2 values for cross-validation of selected QM QSPR models.

QSPR model description: phenols, charges: B3LYP/6-31G*/NPA, 3D structure: OpenBabel with no optimization					
Cross-validation step	**1**	**2**	**3**	**4**	**5**
R^2 for training set	0.955	0.956	0.964	0.959	0.957
R^2 for test set	0.956	0.967	0.939	0.952	0.957
R^2 for complete set	0.957				

QSPR model description: carboxylic acids, charges: B3LYP/6-31G*/NPA, 3D structure: OpenBabel with no optimization					
Cross-validation step	**1**	**2**	**3**	**4**	**5**
R^2 for training set	0.818	0.825	0.889	0.863	0.852
R^2 for test set	0.928	0.785	0.609	0.850	0.816
R^2 for complete set	0.845				

QSPR model description: anilines, charges: B3LYP/6-31G*/NPA, 3D structure: OpenBabel with no optimization					
Cross-validation step	**1**	**2**	**3**	**4**	**5**
R^2 for training set	0.966	0.965	0.973	0.963	0.970
R^2 for test set	0.937	0.925	0.910	0.988	0.932
R^2 for complete set	0.966				

Table 5

Quality criteria for testing of selected QM QSPR models.

QSPR model description: phenols, charges: B3LYP/6-31G*/NPA, 3D structure: NCI with no optimization			
Quality criteria	R^2	**RMSE**	Δ
Training set	0.960	0.415	0.333
Test set	0.948	0.532	0.437

QSPR model description: phenols, charges: HF/STO-3G/MPA, 3D structure: RDKit with no optimization			
Quality criteria	R^2	**RMSE**	Δ
Training set	0.782	1.067	0.896
Test set	0.715	0.421	0.328

Table 6

Number and percentage of QM QSPR models with R^2 higher than a defined limit for individual charge calculation approaches.

QM charge calculation approach	R^2			R^2_{chrg}
	≥ 0.9	(0.9, 0.8>	< 0.8	
HF/STO-3G/MPA	67%	30%	4%	0.914
B3LYP/6-31G*/MPA	60%	25%	16%	0.888
B3LYP/6-31G*/NPA	68%	28%	4%	0.906
B3LYP/6-31G*/AIM	60%	39%	2%	0.898

Note: R^2_{chrg} is the average value of R^2 for all QSPR models, which use charges calculated by a given QM charge calculation approach.

Table 7

Number and percentage of QM QSPR models with R^2 higher than a defined limit for individual classes of molecules.

Class of molecules	R^2			R^2_{mol}
	≥ 0.9	(0.9, 0.8>	< 0.8	
Phenols	32%	49%	17%	0.927
Carboxylic acids	0%	29%	57%	0.849
Anilines	41%	41%	17%	0.929

Note: R^2_{mol} is the average value of R^2 for all QSPR models, which were built for a given class of molecules.

J Chem Inf Model. Author manuscript; available in PMC 2016 November 07.

Table 8

Percentage of QM QSPR models with given R^2 for individual 3D structure sources.

Source	Optimization	R^2				
		≥ 0.95	**(0.95, 0.9>**	**(0.9, 0.85>**	**(0.85, 0.8>**	**< 0.8**
Balloon	none	0%	42%	8%	42%	8%
	MM	8%	33%	33%	17%	8%
	QM	**8%**	**42%**	**25%**	**17%**	**8%**
Frog2	none	0%	50%	50%	0%	0%
	MM	**33%**	**58%**	**0%**	**8%**	**0%**
	QM	33%	42%	25%	0%	0%
NCI	**none**	**50%**	**42%**	**8%**	**0%**	**0%**
	MM	**50%**	**42%**	**8%**	**0%**	**0%**
	QM	8%	58%	33%	0%	0%
OpenBabel	**none**	**58%**	**8%**	**17%**	**8%**	**8%**
	MM	**58%**	**8%**	**17%**	**8%**	**8%**
	QM	33%	17%	17%	25%	8%
PubChem	**none**	**8%**	**75%**	**17%**	**0%**	**0%**
	MM	**25%**	**50%**	**25%**	**0%**	**0%**
	QM	8%	50%	33%	8%	0%
RDKit	none	0%	0%	33%	25%	42%
	UFF	**42%**	**25%**	**17%**	**17%**	**0%**
	MM	25%	50%	8%	0%	17%
	QM	8%	58%	17%	8%	8%

Note: The optimization procedures which produce the best QSPR models for each source of 3D structures are marked in bold font.

Table 9

Sensitivity of a 3D structure source to a change of molecular class.

Percent of insensitive QSPR models	Source					
Optimization	**Balloon**	**Frog2**	**NCI**	**OpenBabel**	**PubChem**	**RDKit**
None	50%	100%	25%	0%	75%	75%
MM	50%	25%	75%	0%	50%	0%
QM	25%	50%	75%	0%	75%	25%
UFF	-	-	-	-	-	25%
Total	42%	58%	58%	0%	67%	31%

Note: The sensitivity of a particular QSPR model to a change coefficient of molecular class was analyzed via a statistical test, which compared correlation coefficient of three independent populations (i.e. molecular classes), employed Fisher's z-transformation and used the significance level 0.05. Detailed information about this statistical test are in Supporting Information.

Table 10

Number and percentage of semiempirical QM QSPR models with R^2 higher than a defined limit.

R^2	≥ 0.95	(0.95, 0.9 >	(0.9, 0.8 >	< 0.8
Number of models	15	25	17	0
Percentage of models	26%	44%	30%	0%

Table 11

R^2 describing the correlation between calculated and experimental pK_a for EEM QSPR models.

R^2		Class of molecules	Phenols				Carboxylic acids				Anilines				Average
		Charge calculation approach	HF, STO-3G, MPA	B3LYP, 6-31G*, MPA	B3LYP, 6-31G*, NPA	B3LYP, 6-31G*, AIM	HF, STO-3G, MPA	B3LYP, 6-31G*, MPA	B3LYP, 6-31G*, NPA	B3LYP, 6-31G*, AIM	HF, STO-3G, MPA	B3LYP, 6-31G*, MPA	B3LYP, 6-31G*, NPA	B3LYP, 6-31G*, AIM	
Source + Optimization	Balloon	none	0.873	0.904	0.903	0.888	0.832	0.924	0.888	0.853	0.806	0.847	0.826	0.870	0.868
		MM	0.852	0.906	0.907	0.885	0.800	0.917	0.883	0.837	0.867	0.845	0.855	0.880	0.870
		QM	0.869	0.908	0.906	0.890	0.772	0.917	0.889	0.851	0.953	0.930	0.908	0.945	0.895
	Frog2	none	0.907	0.897	0.898	0.858	0.832	0.875	0.831	0.870	0.894	0.879	0.904	0.887	0.878
		MM	0.918	0.906	0.917	0.868	0.859	0.888	0.860	0.848	0.863	0.857	0.852	0.902	0.878
		QM	0.921	0.907	0.918	0.869	0.841	0.898	0.866	0.874	0.939	0.926	0.907	0.939	0.900
	NCI	none	0.906	0.906	0.899	0.890	0.875	0.926	0.891	0.879	0.870	0.852	0.839	0.882	0.884
		MM	0.891	0.926	0.926	0.916	0.860	0.920	0.888	0.829	0.844	0.834	0.848	0.889	0.881
		QM	0.896	0.924	0.925	0.912	0.821	0.923	0.884	0.834	0.921	0.884	0.869	0.920	0.893
	OpenBabel	none	0.900	0.920	0.912	0.908	0.830	0.898	0.848	0.826	0.860	0.849	0.851	0.899	0.875
		MM	0.900	0.919	0.911	0.907	0.827	0.903	0.849	0.835	0.858	0.851	0.857	0.897	0.876
		QM	0.896	0.917	0.911	0.904	0.807	0.911	0.856	0.851	0.946	0.935	0.939	0.934	0.901
	PubChem	none	0.896	0.918	0.913	0.902	0.888	0.891	0.866	0.873	0.874	0.881	0.874	0.907	0.890
		MM	0.887	0.917	0.915	0.899	0.874	0.902	0.876	0.871	0.886	0.852	0.872	0.900	0.888
		QM	0.898	0.921	0.925	0.899	0.825	0.923	0.894	0.892	0.890	0.905	0.867	0.927	0.897
	RDKit	none	0.894	0.907	0.904	0.885	0.836	0.932	0.889	0.874	0.832	0.842	0.840	0.857	0.874
		MM-UFF	0.923	0.917	0.912	0.895	0.801	0.919	0.866	0.844	0.838	0.845	0.843	0.875	0.873
		MM	0.899	0.908	0.902	0.892	0.823	0.907	0.871	0.852	0.846	0.852	0.854	0.897	0.875
		QM	0.909	0.919	0.916	0.895	0.753	0.915	0.881	0.851	0.933	0.892	0.869	0.923	0.888
Average			0.897	0.913	0.911	0.893	0.829	0.910	0.872	0.855	0.880	0.871	0.867	0.902	
Legend		$R^2 \geq 0.95$		$R^2 \geq 0.9$		$R^2 \geq 0.866$		$R^2 \geq 0.833$		$R^2 \geq 0.8$		$R^2 \geq 0.7$			

Table 12

Number and percentage of EEM QSPR models with R^2 higher than a defined limit.

R^2	≥ 0.95	(0.95, 0.9>	(0.9, 0.8>	< 0.8
Number of models	82	106	38	2
Percentage of models	36%	46%	17%	1%

Table 13

R^2 describing the correlation between calculated and experimental pK_a for GM QSPR models.

Class of molecules	Phenols	Carboxylic acids	Anilines
R^2	0.747	0.737	0.870