Computational Determination of Aqueous pK_a Values of Protonated Benzimidazoles (Part 2)

Trevor N. Brown and Nelaine Mora-Diez*

Department of Chemistry, Thompson Rivers University, Kamloops, BC, V2C 5N3, Canada

Received: June 23, 2006; In Final Form: August 3, 2006

Our aim is to develop an effective computational procedure for predicting the aqueous acid equilibrium constants of protonated benzimidazoles at 298.15 K. The experimental determination of these values, apart from been laborious, is a challenge because of the low water solubility of these compounds. Using a variety of descriptors, quantitative structure−property relationships (QSPR) are explored between the experimental aqueous pK_a values of a group of fifteen benzimidazoles and descriptors calculated at the B3LYP/6-31+G(d,p) level of theory. Solvent effects are taken into account with the PCM solvation model through both single-point energy calculations (PCM(sp)), and in the geometry optimizations and frequency calculations (PCM(opt)). Descriptors considered are the Gibbs free-energy change of the acid equilibrium in water, the charges on the acidic hydrogen, and on the basic nitrogen, several orbital energies of the protonated and neutral species, and the volume of the solvent cavity. Multiple linear regressions are used to correlate descriptors to the experimental pK_a values. Several QSPR equations reproduce the experimental data more accurately, and show stronger correlations than previously attempted methodologies. The predictive capabilities of the QSPR methodologies are tested with four compounds that were not included in the set of benzimidazoles initially investigated. In addition, a correlation between experimental pK_a values in water and in a 50% ethanol−water solution is used to estimate aqueous pK_a values.

1. Introduction

Acid equilibrium constants (K_a, pK_a = −log K_a) are an important property of organic compounds, with extensive effects on many biological and chemical systems. A thorough discussion of the importance of aqueous pK_a values in general, and on the aqueous pK_a values of benzimidazoles specifically, was provided in a previous publication from our group.[1] In our previous work, a number of different approaches were considered with the aim of developing a methodology for the accurate prediction of the aqueous pK_a values of protonated benzimidazole derivatives at 298.15 K. Aqueous pK_a values were calculated for a group of thirteen benzimidazole derivatives using four equilibria and two equations at four levels of theory. Table 1 shows the experimental aqueous pK_a values of the benzimidazoles considered in our current publication.[2−5] Figure 1 shows the structure of benzimidazole, the parent compound, and its substitution positions.

The accuracy of the directly calculated pK_a values was good in some cases, but insufficient for practical applications.[1] To improve it, a linear correction was applied which produced pK_a values with a lower mean absolute deviation from the experimental values than the noncorrected calculations in all cases. The strongest correlations between experimental and calculated data were obtained at the B3LYP/6-31+G(d,p)-PCM(opt) level of theory (R^2 = 0.903−0.910), in which the PCM solvent model was included in the geometry optimizations and frequency calculations. The predictive capabilities of the methodologies attempted were tested with two compounds that were not included in the set of benzimidazoles initially investigated: 5-chlorobenzimidazole and 2-methoxybenzimidazole. In the direct and correlated pK_a calculations, the relative pK_a ordering of the two benzimidazoles was correct. However, the direct predictions fell outside the estimated upper and lower pK_a limits in every case. The correlation equations at the two levels of theory produced a pK_a value for 5-chlorobenzimidazole that fell well within the estimated upper and lower bounds of the aqueous pK_a value. The pK_a value for 2-methoxybenzimidazole was improved by using the linear correction, but the values still fell outside the estimated upper and lower bounds.

It is clear that some factors are not included in the calculation of the aqueous pK_a (ΔG_{aq}) values of the protonated benzimidazoles when using continuum solvation models. This fact is also evident in the pK_a order inversion between 5-nitrobenzimidazole and 2-chlorobenzimidazole. The results noted are true

TABLE 1: Experimental Aqueous pK_a Values of Protonated Benzimidazoles at 298.15 K[a]

molecule	pK_a	additional values[e]
5-nitrobenzimidazole	4.17[b]	
5-floro-6-chlorobenzimidazole	4.33[d]	
2-chlorobenzimidazole	4.68[c]	
5-chlorobenzimidazole	4.86[d]	
benzimidazole	5.48	5.41[b]
1-methylbenzimidazole	5.57	
1-ethylbenzimidazole	5.62	
4-methylbenzimidazole	5.67	
5-methylbenzimidazole	5.81	
5,6-dimethylbenzimidazole	5.98	5.89[b]
2-methylbenzimidazole	6.19	6.10[b]
5-aminobenzimidazole	6.11	
2-ethylbenzimidazole	6.20	
2-isopropylbenzimidazole	6.23	
2-aminobenzimidazole	7.18[c]	

[a] From ref 2, unless otherwise indicated. [b] Ref 3. [c] Ref 4. [d] Ref 5. [e] Additional experimental values not used in this work.

* Corresponding author phone: 250 828-5442; fax: 250 828 5450; e-mail: nmora@tru.ca.

10.1021/jp0639501 CCC: $33.50
Published on Web 09/15/2006

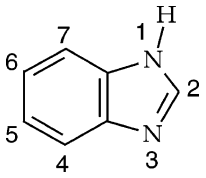

Figure 1. Benzimidazole and its substitution positions.

regardless of which pK_a equation, reaction scheme, or level of theory was used in the calculations. This situation led us to pursue a quantitative structure−property relationship (QSPR) approach to the calculation of pK_a values.

QSPR is a widely used procedure for relating molecular descriptors to the known properties of a set of compounds. The assumption is that once a QSPR correlation is obtained for a set of molecules, it can be used to predict the same property for other similar molecules.[6] A number of molecular descriptors have been used to predict pK_a values.[7] A recent paper used two molecular descriptors (the energy of the LUMO of the acid and the charge on the acidic hydrogen) and a third variable in a multiple linear regression.[8]

In this paper, a number of descriptors are calculated at the B3LYP/6-31+G(d,p)-PCM(sp) and B3LYP/6-31+G(d,p)-PCM-(opt) levels of theory and their correlations to experimental pK_a values are determined and compared. Electrostatic and orbitalic (i.e., descriptors that are derived from molecular orbitals) descriptors, as well as the aqueous Gibbs free-energy change of the acid−base equilibrium, and the volume of the solvent cavity, are considered for use in developing the QSPR equations. Several multivariable linear regressions are performed to correlate different combinations of variables to the experimental pK_a values. The predictive capabilities of the QSPR equations are tested with four compounds that were not included in the set of benzimidazoles initially investigated. In addition, a correlation between experimental pK_a values in water and in a 50% ethanol−water solution is used to estimate aqueous pK_a values.

2. Methodology

2.1. Calculation of Aqueous pK_a Values from Gibbs Free-Energy Changes. The aqueous pK_a of an acid is commonly calculated using eq 1, where ΔG_{aq} is the aqueous Gibbs free-energy change of the equilibrium considered. An equivalent expression that is rarely found in the literature regarding pK calculations uses molecular partition functions among other magnitudes. A more thorough discussion of these equations and their application is contained in our previous publication.[1] Our results showed that they could be used interchangeably in most cases. The equilibrium used to calculate absolute pK_a values is

$$pK_a = \frac{\Delta G_{aq}}{RT \ln 10} \tag{1}$$

the dissociation of an acid (a protonated benzimidazole, HBz^+) into its conjugate base (a neutral benzimidazole, Bz), and a free proton, as shown in Scheme 1. The absolute pK_a of HBz^+ is the pK of Scheme 1. Other reaction schemes have also been used to calculate the aqueous pK_a of benzimidazoles.[1] Changes in Gibbs free energies used in this paper are calculated using eq 1 and Scheme 1 with an experimentally determined value for the aqueous Gibbs free energy of the proton.[9,1]

$$HBz^+ \rightleftarrows Bz + H^+ \qquad \text{(Scheme 1)}$$

2.2. Treatment of Tautomeric Equilibria. Benzimidazole derivatives that are not substituted at positions 1 or 3 exhibit tautomeric equilibria when in their neutral forms.[10] Equation 2 defines the relative populations of the named benzimidazole tautomer (f_1) and the unnamed tautomer (f_2), and eqs 3 and 4 show their relation to the equilibrium constant of the tautomeric equilibrium (K_T). A more thorough discussion of these equations, and the tautomerism of benzimidazoles is provided in our previous paper.[1] In the QSPR analysis of the aqueous acidity

$$f_1 + f_2 = 1,\ 0 \le f_1 \le 1,\ 0 \le f_2 \le 1 \tag{2}$$

$$K_T = \frac{f_2}{f_1} \tag{3}$$

$$f_1 = \frac{1}{K_T + 1} \tag{4}$$

of benzimidazoles, molecular descriptors are calculated as weighted averages of the molecular descriptors of each tautomer, obtained by multiplying the molecular descriptor by the relative population.

2.3. Computational Details. Electronic structure calculations have been performed using the Gaussian 03[11] series of programs on a PQS Quantum Cube with 2.8 GHz Xeon processors running the Mandrake 9.2 Linux operating system. The DFT method B3LYP[12] and the 6-31+G(d,p) basis set are used for all calculations, along with the PCM[13] solvation model. The PCM solvation model is used in single-point energy calculations (PCM(sp)), and during the geometry optimizations and frequency calculations (PCM(opt)). All the calculations using the PCM solvent model employ the UAHF atomic radii when constructing the solvent cavity, as recommended in the Gaussian 03 user's reference when the "scfvac" keyword is used to obtain the free-energy of solvation, as is the case in this study.[14] All the geometries are fully optimized, and the character of the stationary points found is confirmed by a harmonic frequency calculation at the same level of theory to ensure a minimum is located. Multiple linear regressions are calculated and analyzed using the MINITAB 14.1 statistical software.[15]

3. Results and Discussion

3.1. Using the Experimental pK_a Values in 50% Water−Ethanol Solution. Hofman has reported a series of pK_a values of protonated benzimidazoles in both aqueous solution and in a 50% water−ethanol solution.[2] The derivatives for which both pK_a values have been determined are shown in Table S1. Since there is more pK_a data available for protonated benzimidazoles in 50% ethanol solution than in water, it is hoped that a correlation between the two sets of data in Table S1 could be used to determine aqueous pK_a values from the 50% ethanol data, so that they could be compared to the calculated pK_a values. The mean absolute deviation (AD) between the pairs of pK_a values in water and 50% ethanol is of 0.586 pK_a units. A reasonably good correlation factor of 0.973 is obtained from the linear correlation between these values. The correlation equation obtained is also shown in Table S1 and the correlation plot is displayed in Figure S1 of the Supporting Information.

The accuracy of the predictions made by this correlation is verified with two molecules: 5-chlorobenzimidazole and 2-ethylbenzimidazole. The linear correlation is performed once again after excluding the pK_a values of these two molecules, and then the new correlation is used to predict their aqueous pK_a values. The prediction for 5-chlorobenzimidazole is 4.89 and for

TABLE 2: Predicted pK_a Values of Protonated Benzimidazoles from Experimental Data[a]

molecule	experimental 50% water−ethanol pK_a[b]	predicted aqueous pK_a
5-nitro-1-methylbenzimidazole	2.67	4.05
4,6-dichlorobenzimidazole	2.76	4.11
5-chloro-1-methylbenzimidazole	2.88	4.19
5,6-dichlorobenzimidazole	3.26	4.44
5-nitro-2-methylbenzimidazole	3.37	4.52
6-chloro-1-methylbenzimidazole	3.88	4.85
2-phenylbenzimidazole	4.51	5.27
1-allylbenzimidazole	4.58	5.32
5-chloro-2-methylbenzimidazole	4.71	5.40
5-chloro-1,2-dimethylbenzimidazole	4.75	5.43
4-methoxybenzimidazole	4.98	5.58
5-methoxybenzimidazole	5.07	5.64
5-methoxy-1-methylbenzimidazole	5.07	5.64
2-phenyl-5,6-dimethylbenzimidazole	5.10	5.66
1,6-dimethylbenzimidazole	5.17	5.71
1,5-dimethylbenzimidazole	5.22	5.74
1,5,6-trimethylbenzimidazole	5.45	5.89
4,6-dimethylbenzimidazole	5.46	5.90
2-*n*-propylbenzimidazole	5.66	6.03
2-*tert*-butylbenzimidazole	5.76	6.10
5-methoxy-1,2-dimethylbenzimidazole	5.86	6.16
5-methoxy-2-methylbenzimidazole	5.93	6.21
2,5-dimethylbenzimidazole	6.03	6.27
1,2,5-trimethylbenzimidazole	6.07	6.30
2,5,6-trimethylbenzimidazole	6.29	6.45
	aqueous pK_a	50% water−ethanol pK_a
5-floro-6-chlorobenzimidazole	4.33[c]	3.09
2-chlorobenzimidazole	4.68[d]	3.61
5-aminobenzimidazole	6.11	5.78

[a] Correlation equation: (H_2O-pK_a) = 0.668 (50% EtOH-pK_a) + 2.27. [b] From ref 2, unless otherwise indicated. [c] ref 5. [d] ref 4.

2-ethylbenzimidazole 6.06, corresponding to experimental values of 4.86 and 6.20, respectively. The AD of the new correlation is 0.600 and the correlation factor is 0.975.

Table 2 shows additional data (mostly) from Hofman.[2] In these data, experimental 50% water−ethanol pK_a values (and three aqueous pK_a values) are reported. These values are used together with the correlation equation previously obtained to predict the aqueous pK_a values (and three 50% water-ethanol pK_a values) of the listed compounds at 298.15 K. These results are also displayed in Table 2. A linear correlation of pK_a values in different solvents has been successfully performed by Ruiz et al. for estimating aqueous pK_a values from varying mixtures of methanol and water.[16]

3.2. Calculating Aqueous pK_a Values with a Quantitative Structure−Property Relationship. As has been previously noted, continuum solvation models, such as those used in our previous[1] and current studies, are limited in their ability to accurately predict Gibbs free-energies of solvation in polar solvents because of the absence of specific solute−solvent interactions such as hydrogen bonding. Several methods have been employed that specifically include these effects,[17,18] or account for them with an empirical correction, such as using experimental Gibbs free-energy values as we did in our previous paper.[1] On a variety of occasions, explicit solvent molecules have been added in addition to considering continuum solvation models in an attempt to take into account specific solute−solvent interactions.[19] It was decided for this study that we would attempt to account for them in a general way by using molecular descriptors (calculated using a continuum solvation model) to correct for the deviations seen in the calculated aqueous Gibbs free energies. To accomplish this, aqueous Gibbs free-energy changes are included as possible descriptors in the development of the QSPR in the hopes that a multiple linear regression that includes additional descriptors might give a more complete description of all the molecular contributions to the pK_a values. To the best of our knowledge, the inclusion of calculated aqueous Gibbs free energy changes in a QSPR approach for the determination of aqueous pK_a values has not previously been attempted, though changes in total energy have been employed.[6] Among other descriptors, theoretically derived acidities and basicities, have been used to develop theoretical linear solvation energy relationships.[20]

In our previous publication,[1] the mean absolute deviations between the calculated and experimental values of the four levels of theory considered were similar to each other (when referring to the same reaction scheme). However, the calculations at the B3LYP/6-31+G(d,p)-PCM(opt) level of theory had the strongest correlation with experimental aqueous pK_a values. For comparison purposes we decided to also consider calculations at the B3LYP/6-31+G(d,p)-PCM(sp) level of theory in the QSPR study; these results and related comments are shown in the Supporting Information. Equivalent calculations to those reported in ref 1 at this level of theory are displayed in Table S2. The raw data (including the total aqueous-phase energy, the Gibbs free-energy of solvation, and the relative population of one of the tautomers) are reported in Table S3.

To determine which molecular descriptors to use for the QSPR approach, sixteen descriptors per compound are initially considered (at each of the two levels of theory previously mentioned), and their correlation to experimental pK_a values are determined. These descriptors are as follows for the protonated and neutral benzimidazoles: six orbital energies (HOMO-2 through LUMO+2), four atomic charges (charges on N1, N3, H1, H3; refer to Figure 1) calculated using a natural bond order analysis (NBO) and also Mulliken atomic charges, the aqueous Gibbs free energy of the species, and the volume of the solvent cavity. Due to limitations in the way that multiple linear regressions are calculated, the number of descriptors considered may not be larger than the number of available data points, so it was not possible to simultaneously consider all 32 descriptors for the multiple linear regression of the fifteen benzimidazoles.[21] To circumvent this problem a reduced set of descriptors is created and considered in the "Best Subsets Analysis" in MINITAB.

The energies of the frontier orbitals (HOMO and LUMO) of each species are included in the reduced set of descriptors. These are arguably the most important orbitals, but in cases where lower-energy occupied orbitals are close in energy to the HOMO, and higher energy orbitals are close in energy to the LUMO, other orbitals may also be important. By visualization of the molecular orbitals calculated at the B3LYP/6-31+G(d,p)-PCM(opt) level of theory, it can be noted that the HOMO-2 orbital of the base is localized primarily on the basic nitrogen atom (N3). This is the molecular orbital where the lone pair of electrons is localized and should be more important in the acid−base equilibrium than the HOMO of the base. This orbital has a significantly higher correlation to the experimental pK_a values (R^2 = 0.77) than the HOMO (R^2 = 0.66). The energies of the molecular orbitals with a higher correlation to the experimental pK_a values than the HOMO and LUMO of each species are included in the reduced set of descriptors. At the B3LYP/6-31+G(d,p)-PCM(opt) level of theory, the other orbital energies chosen are those of HOMO-1 and LUMO+2 of the acid, and HOMO-2 and HOMO-1 of the base. At the B3LYP/6-31+G-(d,p)-PCM(sp) level of theory the other orbital energies chosen

TABLE 3: Correlation of the Reduced Set of B3LYP/6-31+G(d,p)-PCM(opt) Descriptors to the Experimental Aqueous pK_a Values of Benzimidazoles at 298.15 K[a]

molecule	ϵ^{acid}_{HOMO-1}	ϵ^{acid}_{HOMO}	ϵ^{acid}_{LUMO}	ϵ^{acid}_{LUMO+2}	ϵ^{base}_{HOMO-2}	ϵ^{base}_{HOMO-1}	ϵ^{base}_{HOMO}	ϵ^{base}_{LUMO}	Q^{acid}_H	Q^{base}_N	ΔG_{aq}	ΔV_{sc}
5-nitrobenzimidazole	−0.2996	−0.2924	−0.1244	−0.0308	−0.3060	−0.2644	−0.2575	−0.1092	0.5374	−0.5361	3578	−4.051
5-floro-6-chlorobenzimidazole	−0.2809	−0.2725	−0.0782	−0.0249	−0.3019	−0.2482	−0.2431	−0.0452	0.5339	−0.5474	15902	−1.270
2-chlorobenzimidazole	−0.2753	−0.2742	−0.0744	−0.0235	−0.3083	−0.2446	−0.2405	−0.0348	0.5349	−0.5414	−1159	−3.732
5-chlorobenzimidazole	−0.2814	−0.2681	−0.0743	−0.0219	−0.2983	−0.2472	−0.2385	−0.0396	0.5318	−0.5509	20548	−2.810
benzimidazole	−0.2749	−0.2714	−0.0681	−0.0156	−0.2933	−0.2404	−0.2367	−0.0315	0.5288	−0.5571	28189	−3.872
1-methylbenzimidazole	−0.2744	−0.2699	−0.0683	−0.0136	−0.2922	−0.2393	−0.2343	−0.0325	0.5276	−0.5569	29597	−3.885
1-ethylbenzimidazole	−0.2739	−0.2693	−0.0680	−0.0148	−0.2914	−0.2385	−0.2337	−0.0322	0.5273	−0.5563	30773	−3.739
4-methylbenzimidazole	−0.2732	−0.2619	−0.0662	−0.0112	−0.2913	−0.2380	−0.2299	−0.0283	0.5266	−0.5557	29432	−2.679
5-methylbenzimidazole	−0.2703	−0.2622	−0.0650	−0.0134	−0.2918	−0.2364	−0.2300	−0.0283	0.5276	−0.5592	31114	−3.834
5,6-dimethylbenzimidazole	−0.2625	−0.2575	−0.0626	−0.0120	−0.2904	−0.2303	−0.2264	−0.0261	0.5262	−0.5613	34412	−3.782
5-aminobenzimidazole	−0.2685	−0.2225	−0.0610	−0.0037	−0.2897	−0.2354	−0.2010	−0.0256	0.5249	−0.5611	36972	−3.948
2-methylbenzimidazole	−0.2685	−0.2679	−0.0613	−0.0125	−0.2895	−0.2361	−0.2320	−0.0268	0.5219	−0.5584	37536	−3.966
2-ethylbenzimidazole	−0.2682	−0.2674	−0.0619	−0.0132	−0.2888	−0.2360	−0.2318	−0.0278	0.5217	−0.5555	38060	−4.383
2-aminobenzimidazole	−0.2618	−0.2429	−0.0380	−0.0059	−0.2893	−0.2328	−0.2104	−0.0157	0.5149	−0.6225	51815	−4.121
5-chloro-1-methylbenzimidazole	−0.2801	−0.2676	−0.0744	−0.0200	−0.2975	−0.2444	−0.2378	−0.0408	0.5306	−0.5511	22341	−1.211
5,6-dichlorobenzimidazole	−0.2779	−0.2713	−0.0800	−0.0269	−0.3026	−0.2472	−0.2437	−0.0471	0.5343	−0.5456	14657	−1.152
5-methoxy-2-methylbenzimidazole	−0.2693	−0.2406	−0.0571	−0.0058	−0.2888	−0.2364	−0.2136	−0.0240	0.5207	−0.5602	40548	−3.144
2-methoxybenzimidazole	−0.2687	−0.2599	−0.0491	−0.0048	−0.2982	−0.2390	−0.2239	−0.0214	0.5223	−0.5945	27444	−4.640
R^2	0.744	0.474	0.708	0.843	0.768	0.728	0.656	0.525	0.955	0.644	0.876	0.269
A[b]	75.08	35.04	37.35	101.52	110.13	81.65	48.91	26.75	−136.44	−33.14	5.47E−05	−0.52
B[b]	26.12	14.82	8.15	7.13	37.99	25.20	16.90	6.52	77.54	−12.95	4.05	3.68

[a] The symbols ϵ and Q indicate orbital energies and atomic charges, respectively (both in au), ΔG_{aq} is the change in Gibbs free-energy (in J, using eq 1 and Scheme 1), and ΔV_{sc} is the change in the volume of the solvent cavity going from the protonated to the neutral species (in Å^3). [b] Coefficients of the correlation equation $pK_{a\ exp} = A \cdot \text{Descriptor} + B$.

are those of HOMO-2 and HOMO-1 of the base. Orbital energies will be denoted $\epsilon^{species}_{MO}$, so the symbol ϵ^{acid}_{HOMO} refers to the energy of the HOMO of the acid.

The correlations of the NBO atomic charges are consistently higher than the Mulliken charges in all cases, so the Mulliken charges are eliminated from the reduced set of descriptors. In the neutral benzimidazoles there is no hydrogen at position 3, and there is also no hydrogen at position 1 in the benzimidazoles substituted there, so the atomic charges on H1 and H3 of the neutral species are not included in the reduced set of descriptors. Because of the nomenclature rule by which benzimidazoles are named, it is always the hydrogen (H3) attached to the nitrogen at position 3 (N3) of the cation which is lost in its acid dissociation, leaving the remaining hydrogen (H1) or other substituent on the nitrogen at position 1 (N1). This nitrogen is most likely involved in the acid equilibrium with the benzimidazole anion (defined by a second equilibrium constant, pK_{a-2}), but because we are considering the acid dissociation of the benzimidazole cation the atomic charge on N1 is not included in the reduced set of descriptors, only the charge on N3 of the neutral benzimidazoles is included, and it is denoted Q^{base}_N. The other charges not included in the reduced set of descriptors all have lower R^2 values when correlated to the experimental pK_a values.

In a protonated benzimidazole not substituted at position 1 either of the hydrogen atoms (H1 or H3) can be lost in its acid dissociation, and by definition the nitrogen to which the lost proton was attached becomes N3 of the neutral species. To account for this a composite charge is calculated using the relative populations of the neutral tautomers and the charge on the hydrogen atoms of the protonated species (or the nitrogen atoms). Each charge is multiplied by the relative population of the neutral species resulting from the loss of the proton, and the results are summed yielding a composite atomic charge. The composite nitrogen charge of the protonated benzimidazole is not considered in the reduced set of descriptors because we are concerned only with the loss of the attached hydrogen. Only the charge on the acidic hydrogen of the protonated benzimidazoles is included, and it is denoted Q^{acid}_H.

The aqueous Gibbs free-energy of each species is meaningless individually, the correlations of each to the experimental pK_a values are identical ($R^2 = 0.520$), and this increases to $R^2 = 0.876$ when they are combined in the calculation of ΔG_{aq} using Scheme 1. For demonstrative purposes, the Gibbs free-energy of each species is included separately in a two-variable multiple linear regression; the calculated coefficients of each variable in this regression are equal in magnitude and opposite in sign, which we consider is statistical evidence that they should be summed together, so the ΔG_{aq} is included in the reduced set of variables. A similar case can be made for combining together the volume of the solvent cavity of each species. The correlation of each volume alone is approximately $R^2 = 0.02$, but when both are included in a two-variable multiple linear regression the correlation increases to 0.269 and the coefficients of each are approximately equal in magnitude and opposite in sign, so the change in the volume of the solvent cavity (ΔV_{sc}) is also included in the reduced set of descriptors.

The unmodified descriptors and relative population of the tautomers are shown in Tables S4 and S5. As explained before, molecular descriptors are calculated as weighted averages of the molecular descriptors of each tautomer, obtained by multiplying the molecular descriptor by the relative population.

Following the procedure outlined above a reduced set of descriptors for each level of theory is obtained: 12 for the B3LYP/6-31+G(d,p)-PCM(opt) level of theory and 10 for the B3LYP/6-31+G(d,p)-PCM(sp) level of theory. The reduced sets of descriptors are shown in Tables 3 and S6, along with their identifying symbols, which will be used throughout the rest of the paper, and details of their correlation to the experimental aqueous pK_a values. From the B3LYP/6-31+G(d,p)-PCM(opt) results it can be seen that the NBO charge on the acidic hydrogen of the protonated benzimidazoles, Q^{acid}_H, has the highest correlation to the experimental values ($R^2 = 0.955$), followed by the correlation of ΔG_{aq} ($R^2 = 0.876$). Of the orbital eigenvalues, the energy of the LUMO+2 of the protonated benzimidazoles, ϵ^{acid}_{LUMO+2}, shows the strongest correlation to the experimental data ($R^2 = 0.843$), followed by the energy of the HOMO-2 orbital of the base ($R^2 = 0.768$). The ΔV_{sc} shows

TABLE 4: Inter-variable Correlations (R^2 Values) of the B3LYP/6-31+G(d,p)-PCM(opt) Descriptors for the Set of Benzimidazoles under Study

	ϵ^{acid}_{HOMO-1}	ϵ^{acid}_{HOMO}	ϵ^{acid}_{LUMO}	ϵ^{acid}_{LUMO+2}	ϵ^{base}_{HOMO-2}	ϵ^{base}_{HOMO-1}	ϵ^{base}_{HOMO}	ϵ^{base}_{LUMO}	Q^{acid}_{H}	Q^{base}_{N}	ΔG_{aq}	ΔV_{sc}
ϵ^{acid}_{HOMO-1}		0.483	0.891	0.738	0.570	0.965	0.619	0.856	0.672	0.404	0.612	0.074
ϵ^{acid}_{HOMO}	0.483		0.491	0.723	0.358	0.462	0.961	0.412	0.403	0.349	0.444	0.023
ϵ^{acid}_{LUMO}	**0.891**	0.491		0.704	0.528	0.853	0.622	0.930	0.699	0.534	0.637	0.020
ϵ^{acid}_{LUMO+2}	**0.738**	**0.723**	0.704		0.786	0.788	0.855	0.616	0.749	0.440	0.786	0.136
ϵ^{base}_{HOMO-2}	0.570	0.358	0.528	**0.786**		0.681	0.485	0.456	0.740	0.304	0.903	0.120
ϵ^{base}_{HOMO-1}	**0.965**	0.462	**0.853**	**0.788**	0.681		0.597	0.865	0.639	0.322	0.646	0.071
ϵ^{base}_{HOMO}	0.619	**0.961**	0.622	**0.855**	0.485	0.597		0.512	0.575	0.462	0.584	0.072
ϵ^{base}_{LUMO}	**0.856**	0.412	**0.930**	0.616	0.456	**0.865**	0.512		0.491	0.283	0.465	0.004
Q^{acid}_{H}	0.672	0.403	0.699	**0.749**	**0.740**	0.639	0.575	0.491		0.682	0.899	0.176
Q^{base}_{N}	0.404	0.349	0.534	0.440	0.304	0.322	0.462	0.283	0.682		0.585	0.066
ΔG_{aq}	0.612	0.444	0.637	**0.786**	**0.903**	0.646	0.584	0.465	**0.899**	0.585		0.104
ΔV_{sc}	0.074	0.023	0.020	0.136	0.120	0.071	0.072	0.004	0.176	0.066	0.104	
SIVC[a]	6.884	5.109	6.907	7.322	5.933	6.887	6.343	5.889	6.726	4.431	6.664	0.866

[a] Sum of the inter-variable correlations.

a poor correlation on its own. The fact that the LUMO+2 orbital energies show a better correlation than the LUMO ones should be an artifact of the level of theory.

From the B3LYP/6-31+G(d,p)-PCM(sp) results it can be seen that the patterns of correlation strength are similar to those noted for the B3LYP/6-31+G(d,p)-PCM(opt) level of theory. The correlations of the descriptors are weaker in all cases, corresponding to the weaker correlation of the pK_a values calculated at the B3LYP/6-31+G(d,p)-PCM(sp) level of theory to the experimental pK_a values.

The inter-variable correlations of the descriptors are shown in Tables 4 and S7. Values of R^2 greater than 0.700 are shown in bold. The sum of the inter-variable correlations (SIVC) is also shown. Although this value has no specific defined statistical meaning, it could be viewed as a general measure of how strongly each variable is correlated to the other variables in the reduced set. In general it is observed that descriptors with a strong correlation to the experimental pK_a values have a high SIVC value as well. We interpret this to mean that descriptors which correlate highly to the experimental pK_a values contain relevant thermodynamic information on the acid dissociation equilibrium; additionally, a high inter-variable correlation between descriptors indicates that the descriptors contain overlapping information regarding the equilibrium considered. A chemically significant observation is that the highest inter-variable correlations are between ϵ^{acid}_{HOMO} and ϵ^{base}_{HOMO} (R^2 = 0.961), ϵ^{acid}_{LUMO} and ϵ^{base}_{LUMO} (R^2 = 0.930), and ϵ^{acid}_{HOMO-1} and ϵ^{base}_{HOMO-1} (R^2 = 0.965) at the B3LYP/6-31+G(d,p)-PCM(opt) level of theory. This can be interpreted to mean that the molecular orbitals of the acids and their conjugate bases remain largely unchanged after the acid dissociation, or that the pairs of orbitals in both species change in energy in an almost constant way among the molecules of a family of compounds. The correlations of these orbital descriptors to the experimental pK_a values vary significantly despite their strong inter-variable correlations which indicate that some molecular orbitals are more involved than others in the acid−base equilibrium. From the correlations in Tables 3, it appears that the LUMO+2 of the acids and the HOMO-2 of the bases are more involved in the equilibrium under study at the B3LYP/6-31+G(d,p)-PCM-(opt) level of theory. It makes sense to think that an empty orbital(s) of an acid will be more involved in its deprotonation process, while an occupied orbital of the base (that of the lone pair of electrons that could be donated) will be more involved in its protonation process.

Relatively high PCM(opt) inter-variable correlations are also observed between ΔG_{aq} and Q^{acid}_{H} (R^2 = 0.899), ϵ^{base}_{HOMO-2} (R^2 = 0.903), and ϵ^{acid}_{LUMO+2}(R^2 = 0.786), also between Q^{acid}_{H} and ϵ^{base}_{HOMO-2} (R^2 = 0.740), and ϵ^{acid}_{LUMO+2}(R^2 = 0.749). The PCM-(sp) inter-variable correlations are lower than the PCM(opt) ones. On the other hand, ΔV_{sc} is not strongly correlated to any of the descriptors considered.

In a recent paper, Soriano et al.[8] developed a QSPR equation to calculate the aqueous pK_a in a series of fifteen N3-protonated imidazole-1-ylalcanoic acid derivatives using two molecular descriptors: the natural atomic charge on the N3 hydrogen of the acid, the energy of the LUMO of the protonated species, and an indicator variable equal to the number of ester groups. They derived their descriptors from optimizations done in the gas phase at the HF/6-31G(d) level of theory. As noted in their work, both of these descriptors are easily related to the acidity of a chemical species. The charge on the acidic hydrogen is related to the delocalization of charge over the molecule and in turn to the acidity; a lower value of Q^{acid}_{H} indicates good delocalization and should produce a higher pK_a. This result could also be explained by stating that the lower the positive charge on the acidic hydrogen, the less polarized the N−H bond and the less acidic the compound (higher pK_a). The energy of the LUMO of the protonated species is related to the formation of hydrogen bonds with the solvent molecules and the subsequent deprotonation. A lower value for ϵ^{acid}_{LUMO} means that hydrogen bonds will form with greater ease, and allow for easier deprotonation and a lower pK_a.

From the data reported in Tables 3 (and S6), it is possible to rationalize the trends of the descriptors considered as the experimental pK_a values of the protonated benzimidazoles increase (acidity decreases). As previously discussed, Q^{acid}_{H} and ϵ^{acid}_{LUMO} show negative and positive correlations (slopes), respectively, and ϵ^{acid}_{LUMO+2} follows the same trend as ϵ^{acid}_{LUMO}. The values of Q^{base}_{N} become more negative with increasing pK_a, (negative correlation), which indicates that the larger the

TABLE 5: Correlations Between the Experimental Aqueous pK_a Values (at 298.15 K) of the Protonated Benzimidazoles and the Calculated Values from the Multiple Linear Regressions with the Selected Descriptors at the B3LYP/6-31+G(d,p)-PCM(opt) Level of Theory

molecules	pK_a	ΔG_{aq}-Q_H^{acid}	ΔG_{aq}-ϵ_{LUMO}^{acid}	Q_H^{acid}-ϵ_{LUMO}^{acid}	ΔG_{aq}-Q_H^{acid}-ϵ_{LUMO}^{acid}	ΔV_{sc}-Q_H^{acid}-ϵ_{LUMO+2}^{acid}
5-nitrobenzimidazole	4.17	4.21	3.90	4.12	4.10	4.21
5-floro-6-chlorobenzimidazole	4.33	4.71	4.96	4.72	4.73	4.37
2-chlorobenzimidazole	4.68	4.48	4.28	4.60	4.53	4.63
5-chlorobenzimidazole	4.86	5.00	5.20	5.00	5.01	4.86
benzimidazole	5.48	5.41	5.60	5.41	5.42	5.48
1-methylbenzimidazole	5.57	5.57	5.66	5.56	5.56	5.66
1-ethylbenzimidazole	5.62	5.61	5.71	5.60	5.60	5.63
4-methylbenzimidazole	5.67	5.70	5.67	5.70	5.69	5.67
5-methylbenzimidazole	5.81	5.58	5.76	5.57	5.58	5.66
5,6-dimethylbenzimidazole	5.98	5.76	5.93	5.75	5.76	5.82
5-aminobenzimidazole	6.11	6.32	6.08	6.31	6.28	6.24
2-methylbenzimidazole	6.19	5.94	6.06	5.92	5.94	6.23
2-ethylbenzimidazole	6.20	6.35	6.09	6.33	6.31	6.29
2-aminobenzimidazole	7.18	7.26	6.95	7.27	7.24	7.13
R^2		0.957	0.901	0.957	0.958	0.990
AD[a]		0.122	0.131	0.119	0.114	0.068

[a] Mean absolute deviation of the predicted values from experimental pK_a values.

TABLE 6: Best Four Results of the "Best Subsets" Regression Analysis of the MINTAB Program for Two-, Three-, and Four-variable Correlation Equations at the B3LYP/6-31+G(d,p) Level of Theory

# variables	R^2	ϵ_{HOMO-1}^{acid}	ϵ_{HOMO}^{acid}	ϵ_{LUMO}^{acid}	ϵ_{LUMO+2}^{acid}	ϵ_{HOMO-2}^{base}	ϵ_{HOMO-1}^{base}	ϵ_{HOMO}^{base}	ϵ_{LUMO}^{base}	Q_H^{acid}	Q_N^{base}	ΔG_{aq}	ΔV_{sc}
2	0.976				X					X			
2	0.970						X			X			
2	0.970									X			X
2	0.967	X								X			
3	0.990				X					X			X
3	0.988						X			X			X
3	0.985	X								X			X
3	0.983							X		X			X
4	0.993	X			X					X			X
4	0.993				X		X			X			X
4	0.992						X	X		X			X
4	0.992		X				X			X			X

negative charge on the basic nitrogen of the neutral benzimidazole, the stronger it will be as a base (accepting a proton more readily), hence a weaker acid once protonated. For additional comments that show how the descriptors considered provide an example of how electrostatic and orbitalic factors help explain the acidity (basicity) order of the protonated benzimidazoles (neutral benzimidazoles), refer to the Appendix section of the Supporting Information.

While the ΔG_{aq} (and calculated pK_a) order of 5-nitrobenzimidazole and 2-chlorobenzimidazole is inverted with respect to the experimental data available (as previously stated in Section 1 and ref 1), the acidity order predicted for these two compounds using most of the other descriptors is in agreement with the experimental values. An inverted order is obtained from the ϵ_{HOMO-2}^{base} descriptor, and also from the Q_N^{base} descriptor at the B3LYP/6-31+G(d,p)-PCM(sp) level of theory.

Two different approaches for defining the multiple linear regressions are attempted. The first approach employs the two descriptors used by Soriano et al., that in our case include (continuum) solvent effects, and ΔG_{aq}, instead of the structural variable. These three descriptors happen to also be ones with the best correlations to the experimental pK_a values (with the exception of ϵ_{LUMO}^{acid} at the B3LYP/6-31+G(d,p)-PCM(opt) level of theory). Tables 5 and S8 show the results of four different multiple linear regressions at both levels of theory: three two-variable combinations and the combination of all three variables. The calculated pK_a values are obtained using the multiple linear correlation equations between the experimental data and the corresponding descriptors. The R^2 and the AD values refer to the linear correlations between the experimental and calculated pK_a values.

The second approach employs the "Best Subsets" regression analysis of the MINTAB program to identify the strongest multivariable correlations. Tables 6 and S9 show the best four results of this analysis for two-, three-, and four-variable correlation equations at the two levels of theory. At the B3LYP/6-31+G(d,p)-PCM(opt) level of theory the best two-variable correlation includes the Q_H^{acid} and ϵ_{LUMO+2}^{acid} descriptors, the best three-variable correlation includes these two plus ΔV_{sc}, and the best four-variable correlation includes these three plus ϵ_{HOMO-1}^{acid}. Adding the fourth variable only increases the R^2 value very slightly so the three-variable correlation equation is used to calculate the pK_a values, and these results are shown in Table 5. The coefficients and R^2 values of what we think are the best multiple linear regression equations between the experimental aqueous pK_a values and the descriptors under study, are reported in Table S10. These equations will be used in the next section to test their predictive capabilities.

The multiple linear regressions using B3LYP/6-31+G(d,p)-PCM(opt) descriptors have R^2 values (0.901−0.990) that are higher than those of the individual variables correlated to the pK_a values. The multi-variable correlations of our first approach that include Q_H^{acid} have R^2 values of $\sim$ 0.95 and a mean absolute deviation from experiment of $\sim$ 0.12. The addition of ΔG_{aq} to the multivariable correlation that includes Q_H^{acid} and ϵ_{LUMO}^{acid} does almost nothing to the quality of the correlation and the magnitude of AD. A similar situation is found when instead of ϵ_{LUMO}^{acid}, ϵ_{LUMO+2}^{acid} is used, but the R^2 in the two- and three-

TABLE 7: Selected Equations of the Multivariable Linear Regressions Obtained Using normalized Descriptors

B3LYP/6-31+G(d,p)-PCM(opt)
$pK_a = -0.1054 \times \Delta V_{sc} - 0.5547 \times Q_H^{acid} + 0.2319 \times \epsilon_{LUMO+2}^{acid} + 5.5608$
$pK_a = -0.0921 \times \Delta G_{aq} - 0.6628 \times Q_H^{acid} + 0.2592 \times \epsilon_{LUMO+2}^{acid} + 5.5608$
$pK_a = -0.5964 \times Q_H^{acid} + 0.2350 \times \epsilon_{LUMO+2}^{acid} + 5.5607$
$pK_a = 0.0734 \times \Delta G_{aq} - 0.6768 \times Q_H^{acid} + 0.0640 \times \epsilon_{LUMO}^{acid} + 5.5607$

descriptor equations are higher. The best three-descriptor correlation obtained with the second method (ΔV_{sc}-Q_H^{acid}-ϵ_{LUMO+2}^{acid}) has a significantly higher correlation to experimental pK_a values ($R^2 = 0.990$) and a significantly lower mean absolute deviation from experimental values (AD = 0.068).

Equation 5 is obtained when performing the multiple linear regression of pK_a with the ΔG_{aq}, Q_H^{acid} and ϵ_{LUMO}^{acid} descriptors calculated at the B3LYP/6-31+G(d,p)-PCM(opt) level of theory. As expected, the coefficients of ΔG_{aq} and ϵ_{LUMO}^{acid} are positive and the coefficient of Q_H^{acid} is negative. This expression can be rearranged (see eq 6) to show its relationship with eq 1. In this rearrangement, the additional terms could be seen as electrostatic and orbitalic corrections to the calculated pK_a values using continuum solvation models. The best three-descriptor multivariable linear regression equation obtained by the Minitab "Best Subsets" analysis is shown in eq 7

$$pK_a = 5.500 \times 10^{-6} \times \Delta G_{aq} - 115.2 \times Q_H^{acid} + 3.344 \times \epsilon_{LUMO}^{acid} + 66.39 \quad (5)$$

$$pK_a = 0.0314 \times \frac{\Delta G_{aq}}{RT \ln 10} + (-115.2 \times Q_H^{acid} + 3.344 \times \epsilon_{LUMO}^{acid} + 66.39) \quad (6)$$

$$pK_a = -0.1299 \times \Delta V_{sc} - 94.70 \times Q_H^{acid} + 31.27 \times \epsilon_{LUMO+2}^{acid} + 55.54 \quad (7)$$

$$pK_a = 0.0734 \times \Delta G_{aq} - 0.677 \times Q_H^{acid} + 0.0640 \times \epsilon_{LUMO}^{acid} + 5.561 \quad (8)$$

If the values of the descriptors are normalized, their weight in the pK_a determination can be interpreted in a better way. The normalized data at both levels of theory are obtained by subtracting each descriptor the mean value and dividing by the standard deviation. Equation 5 becomes eq 8 when using normalized descriptors. Tables 7 and S11 show equations equivalent to those reported in Table S10 using normalized descriptors. The greater the absolute value of a normalized coefficient the greater the descriptor's contribution to the overall correlation. From eq 8 it can be inferred that Q_H^{acid} has the largest contribution to the pK_a, followed by ΔG_{aq} and ϵ_{LUMO}^{acid} which are similar in the weight of their contributions. In the best three-variable correlation equation at the B3LYP/6-31+G(d,p)-PCM(opt) level of theory, Q_H^{acid} once again has the largest contribution, followed by ϵ_{LUMO+2}^{acid} and then ΔV_{sc}.The independent terms in the nonnormalized multi-descriptor equations are relatively large when compared to the pK_a values the equations are used to predict, and they are of similar magnitude when using normalized descriptors. One might wonder about their meaning and possible elimination in order to come up with QSPR equations that only contain weighted descriptors. Indeed, the independent terms could be eliminated by forcing the correlations to have zero intercept. When doing so, the R^2 values using nonnormalized data significantly decrease, but remain unchanged if the descriptors are normalized, and in both cases the predictive capabilities of the QSPR equations are negatively affected. The explanation for this lies in the fact that the individual descriptors when linearly correlated to the experimental pK_a values, have an intercept (independent term) that should not disappear when several of them are combined in a multivariable linear correlation. Having no statistical or mathematical justification for eliminating the independent term in the QSPR equations reported, they are left unchanged.

The two correction terms in eq 5 refer to electrostatic (charge on the acidic hydrogen) and orbitalic properties (LUMO energy) of the acid (the protonated benzimidazoles), respectively. A similar analogy can be made with eq 7. There exists some similarity between these equations that refer to a thermodynamic equilibrium and the equation for estimating chemical reactivity proposed by Klopman[23] and Salem[24] in which electrostatic charges and frontier orbital energies of the interacting species must be considered when estimating kinetic aspects of a given chemical reaction. This equation is composed of terms that describe electron densities, orbital energies, and atomic charges.[22] From the equation that refers to reactivity, we understand that in some cases reactivity is electrostatically controlled (between hard species) while in others it is determined by orbitalic (frontier-orbital interactions, between soft species). Cases between are also to be expected. The electrostatic and orbitalic criteria for reactivity have been widely used to predict sites of attack between interacting species.

Substituent effects modify the way molecules interact with each other (e.g., in solute−solvent interactions) from electrostatic and/or orbitalic points of view. This situation seems to indicate that certain orbitalic and electrostatic factors that influence the thermodynamics of a chemical (in this case, acid−base) equilibrium are not taken into account when performing direct pK calculations in polar solvents if continuum solvation models are used. In general, one might be able to quantify the effects that electrostatic and orbitalic interactions have in determining the extent of a given chemical equilibrium.

One could be adventurous in stating that, for the equilibrium under study with the acid dissociation of protonated benzimidazoles, electrostatic factors play a more important role than the orbitalic ones in determining the equilibrium state, since the normalized coefficient for Q_H^{acid} is larger than that of ϵ_{LUMO}^{acid} at both levels of theory explored. This statement would be in agreement with the fact that these compounds are relatively hard acids since the protonated site is a hard (polarizing) atom, nitrogen. One might expect that for a similar family of compounds having sulfur as the protonated atomic center, orbitalic factors should play a more important role. Possibly an equivalent equation to eq 8 could be obtained in which the ϵ_{LUMO}^{acid} (or another orbitalic descriptor) coefficient is larger than the Q_H^{acid} (or another electrostatic descriptor) coefficient.

3.3. Predicting Aqueous pK_a Values of Protonated Benzimidazoles. Aqueous pK_a calculations are performed for four additional benzimidazole derivatives not shown in Table 1: 5-chloro-1-methylbenzimidazole, 5,6-dichlorobenzimidazole, 5-methoxy-2-methylbenzimidazole, and 2-methoxybenzimidazole. As explained in ref 1, the aqueous pK_a of 2-methoxybenzimidazole is predicted to be in the range 6.10−7.18. pK_a ranges can also be predicted for 5-chloro-1-methylbenzimidazole, 5,6-dichlorobenzimidazole and 5-methoxy-2-methylbenzimidazole.

5-chloro-1-methylbenzimidazole will be more acidic than

TABLE 8: Predictions for the Aqueous pK_a Values (at 298.15 K) of Selected Benzimidazoles Using Various Methodologies

level of theory		5,6-dichloro benzimidazole	5-chloro-1-methyl benzimidazole	5-methoxy-2-methyl benzimidazole	2-methoxy benzimidazole
	predicted range	4.17−4.86	4.86−5.57	6.10−7.18	6.10−7.18
	50% ethanol−water correlation	4.45	4.19	6.23	-
B3LYP/6-31+G(d,p)-PCM(opt)	ΔV_{sc} -Q_H^{acid}- ϵ_{LUMO+2}^{acid}	4.25	4.82	6.46	6.53
	ΔG_{aq} -Q_H^{acid}- ϵ_{LUMO+2}^{acid}	4.48	5.09	6.58	6.52
	Q_H^{acid}- ϵ_{LUMO+2}^{acid}	4.51	5.11	6.56	6.43
	ΔG_{aq} -Q_H^{acid}- ϵ_{LUMO}^{acid}	4.67	5.15	6.46	6.22
	Correlation: S1,[a] eq 1	4.84	5.35	6.25	5.54
	Correlation: S3,[a] eq 1	4.85	5.36	6.26	5.54
	Direct: S1, eq 1	2.27	3.91	6.81	4.51
	Direct: S3, eq 1	3.11	4.76	7.65	5.35
B3LYP/6-31+G(d,p)-PCM(sp)	ΔG_{aq} -Q_H^{acid}- ϵ_{HOMO}^{base}	4.70	5.20	6.57	6.08
	ΔG_{aq} -Q_H^{acid}- ϵ_{LUMO}^{acid}	4.76	5.22	6.36	5.98
	Correlation: S1,[b] eq 1	4.88	5.40	6.15	5.44
	Correlation: S3,[b] eq 1	4.90	5.42	6.18	5.47
	Direct: S1, eq 1	1.72	3.52	6.13	3.67
	Direct: S3, eq 1	3.08	4.88	7.49	5.03

[a] The correlation equations reported in Table 9 of ref 1 have been updated to include two more compounds: ($pK_{a\ exp} = 0.3097 \cdot pK_{a\ cal} + 4.14$) for S1 eq 1, and ($pK_{a\ exp} = 0.3098 \cdot pK_{a\ cal} + 3.89$) for S3 eq 1; S1 and S3 are Schemes 1 and 3 of ref 1. [b] Using the correlation equations reported in Table S2 of the Supporting Information (including the fifteen compounds).

1-methylbenzimidazole, so the maximum pK_a will be 5.57, and it will be more basic than 5-chlorobenzimidazole so the minimum pK_a will be 4.86. 5,6-dichlorobenzimidazole will be more acidic than 5-chlorobenzimidazole, so the maximum pK_a is 4.86, and will probably be less acidic than 5-nitrobenzimidole so the lower pK_a boundary is 4.17. 5-methoxy-2-methylbenzimidazole will be more basic than 2-methylbenzimidazole so the lower pK_a bound is 6.10, and more acidic than 2-aminobenzimidazole so the upper bound will be 7.18.

Predictions are made for the pK_a values of these compounds using the selected multi-variable linear QSPR equations (see Table S10) and are summarized in Table 8. Some of the correlations obtained in our previous publication and direct calculations are shown for comparison purposes,[1] together with predictions using the 50% water-ethanol data (see Tables S1 and 2). We judge the accuracy of the predictions based on the ability of the method to predict values within the estimated upper and lower bounds of the pK_a, and to correctly predict the pK_a ordering of the four test molecules. The molecule with the lowest pK_a should be 5,6-dichlorobenzimdazole, followed by 5-chloro-1-methylbenzimidazole. Predicting the pK_a ordering for the other molecules is more difficult as the pK_a values are most likely very close, but in 2-methoxybenzimidazole the electron-donating methoxy group is much closer to the acidic site, so we predict that 2-methoxybenzimidazole will have the highest pK_a, with 5-methoxy-2-methylbenzimidazole slightly less basic.

When using the multivariable linear equations calculated at the B3LYP/6-31+G(d,p)-PCM(opt) and PCM(sp) levels of theory, the predicted values fall mostly within the predicted upper and lower bounds of the pK_a. Direct calculation of pK_a values gives predictions that are outside of the predicted pK_a range in almost every case, but when the values are predicted with the corresponding correlation equation, the values calculated for 5-chloro-1-methylbenzimidazole and 5-methoxy-2-methylbenzimidazole are within the predicted range, and the values for 5,6-dichlorobenzimidazole and 2-methoxybenzimidazole are just above and below their predicted pK_a ranges, respectively.

The pK_a ordering of 5,6-dichlorobenzimidazole and 5-chloro-1-methylbenzimimidazole is predicted correctly by every method attempted except for the prediction using the 50% water-ethanol pK_a correlation. The 50% water-ethanol predictions fall within the predicted bounds for 5,6-dichlorobenzimidazole and 5-methoxy-2-methylbenzimidazole, and do not differ greatly from the three-descriptor QSPR predictions at the two levels of theory considered. However, this predicting approach produces a pK_a well below the expected minimum value for 5-chloro-1-methylbenzimidazole. A reason for this situation might be the fact that none of the 1-substituted benzimidazoles used in obtaining this correlation equation have any other substituents present, and none of the compounds in Table S1 contain substituents in the two rings of benzimidazole simultaneously.

With respect to the QSPR approach, we have more confidence in the PCM(opt) predictions with the equations that contain ϵ_{LUMO+2}^{acid}, since they produce the best correlations with the experimental data. The pK_a range predicted for the four test compounds using these equations is of 0.10−0.29 pK_a units, which is not significant from an experimental point of view. ΔG_{aq} and the other two descriptors, Q_H^{acid} and ϵ_{LUMO+2}^{acid}, are not linearly independent, R^2 values of 0.899 and 0.786 exist between them, respectively (PCM(opt)). Clearly, there is some overlap between the relevant thermodynamic information they contain regarding this equilibrium. Equation 7 might be better from a statistical point of view because Q_H^{acid} and ϵ_{LUMO+2}^{acid} correlate very poorly with ΔV_{sc}, indicating a much higher linear independence between them. In addition to this, a much better correlation and accurate pK_a values are obtained with this equation.

The pK_a ordering of 5-methoxy-2-methylbenzimidazole and 2-methoxybenzimidazole is incorrectly predicted in almost every case. Only at the B3LYP/6-31+G(d,p)-PCM(opt) level of theory in the three multi-variable correlations using the Q_H^{acid} and ϵ_{LUMO+2}^{acid} descriptors are the pK_a values close to each other. For this reason, we conclude that the Q_H^{acid}, and ϵ_{LUMO+2}^{acid} descriptors have most of the predictive power, with additional descriptors accounting for smaller variations. Only when these two descriptors are paired with ΔV_{sc}, the expected pK_a ordering is obtained. When using this QSPR equation, all the predicted values fall within the expected upper and lower pK_a limits, except the value for 5-chloro-1-methylbenzimidazole that is 0.04 units below the lower expected limit. It should be taken into account that the

predicted ranges have been determined from the available experimental data that, as usual, are not perfect.

4. Conclusions

A QSPR approach is undertaken in an attempt to correct for specific solute−solvent (electrostatic and orbitalic) interactions not included in the direct and correlated pK_a calculations using continuum solvation models. The QSPR approach using three descriptors reproduces the experimental aqueous pK_a values of the protonated benzimidazoles under study with good accuracy and shows a strong correlation to the experimental data.

It is shown that from 50% water−ethanol pK_a values of protonated benzimidazoles, aqueous equilibrium constants can be estimated with reasonable accuracy in most cases. The pK_a values predicted using this method are in good agreement with the predicted pK_a range and also with the values predicted with the QSPR methodology. Discrepancies are found with benzimidazoles with substitution patterns not present in the set of compounds used to develop this correlation equation.

The predictive capabilities of all the methodologies attempted are tested with four compounds that are not included in the set of benzimidazoles initially investigated. The QSPR equation with the ΔV_{sc}, Q_H^{acid} and ϵ_{LUMO+2}^{acid} descriptors at the B3LYP/6-31+G(d,p)-PCM(opt) level of theory has the highest correlation to the experimental values, is better from a statistical point of view, and gives the most reasonable predictions.

Electrostatic and orbitalic factors influence the acid−base properties of the species involved in the equilibrium under study. It is observed that the acid dissociation of the protonated benzimidazoles appears to be primarily electrostatically controlled, but that the highest correlations with the experimental aqueous pK_a values are obtained when both electrostatic and orbitalic descriptors are included in the correlation. Even though more descriptors of the individual species or of complexes with a solvent molecule could be investigated, the QSPR equations reported in this work are useful from the practical point of view of predicting the desired property.

One could hypothesize that for any chemical equilibrium it could be possible to come up with electrostatic and/or orbitalic descriptors that characterize the interaction between the reactant and product species, and that could allow the accurate calculation of the equilibrium constant. Since rate constants can be related to the expression derived by Klopman[23] and Salem,[24] and equilibrium constants are a ratio of rate constants, the development of a generalized expression for pK calculations should be possible. The influence of electrostatic and orbitalic factors in determining the magnitude of equilibrium constants (in a variety of media) is yet to be exhaustively explored.

Acknowledgment. We gratefully acknowledge the Natural Sciences and Engineering Research Council of Canada (NSERC) and Thompson Rivers University (CUEF program) for financial support. We are also thankful to Prof. Shane Rollans at TRU for valuable discussions regarding the multivariable linear regression analysis.

Supporting Information Available: The experimental pK_a values in water and 50% water−ethanol used to develop a correlation, and the correlation plot, are shown in Table S1 and Figure S1, respectively. Equivalent calculations to those reported in ref 1 at the B3LYP/6-31+G(d,p)-PCM(sp) level of theory, are displayed in Table S2; the raw data (including the total aqueous-phase energy, the Gibbs free-energy of solvation, and the relative populations of one of the tautomers) are reported in Table S3. The unmodified descriptors and the relative populations are reported in Tables S4 and S5. Tables S6 to S9, and S11, show the B3LYP/6-31+G(d,p)-PCM(sp) results analogous to the B3LYP/6-31+G(d,p)-PCM(opt) results reported in this paper. The correlation equations of the best multivariable linear regressions using nonnormalized descriptors, and the multivariable correlation coefficients, as calculated by Minitab, are reported in Table S10. Additional comments on the results obtained appear in the Appendix. This information is available free of charge via the Internet at http://pubs.acs.org.

References and Notes

(1) Brown, T. N.; Mora-Diez, N. *J. Phys. Chem. B* **2006**, *110*, 9270.
(2) Hofman, K. *Imidazole and its Derivatives*; Interscience Publishers: New York, **1953**.
(3) Donkor, K. K.; Kratochvil, B. *J. Chem. Eng. Data*, **1993**, *38*, 569.
(4) Schenkeveld, S.; Donkor, K. K. *Spectrometric Determination of Aqueous Ionization Constants (pK_a) of Benzimidazoles, Directed Studies Report*; Thompson Rivers University Library: Kamloops, Canada, **2002**.
(5) Kapinos, L. E.; Song, B.; Sigel, H. *Chem. Eur. J.* **1999**, *5*, 1794.
(6) Karelson, M.; Lobanov, V. S.; Katritzky, A. R. *Chem. Rev.* **1996**, *96*, 1027.
(7) Some examples: (a) Duchowicz, P. R.; Castro, E. A. *Mendeleev Comm.* **2002**, *5*, 187. (b) Citra, M. J. *Chemosphere* **1999**, *38*, 191. (c) Adam, K. R. *J. Phys. Chem. A* **2002**, *106*, 11963. (d) Hennemann, M.; Clark, T. *J. Mol. Model* **2002**, *8*, 95.
(8) Soriano, E.; Cerdán, S.; Ballesteros, P. *J. Mol. Struct. (THEOCHEM)* **2004**, *684*, 121.
(9) Palascak, M. W.; Shields, G. C. *J. Phys. Chem. A* **2004**, *108*, 3692.
(10) Eicher, T.; Hauptmann, S. *The Chemistry of Heterocycles*; Wiley-VCH: Weinheim, 2003.
(11) 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.
(12) (a) B3 functional: Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648. (b) LYP functional: Lee, C.; Yang, W. T.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785.
(13) Cances, E.; Mennucci, B.; Tomasi, J. *J. Chem. Phys.* **1997**, *107*, 3032.
(14) Frisch, Æ; Frisch, M.; Trucks, G. W. *Gaussian 03 User*'s *Reference*; Gaussian Inc.: Wallingford, CT, **2003**.
(15) MINITAB State College, PA Minitab, Inc.
(16) Ruiz, R.; Rosés, M.; Ràfols, C.; Bosch, E. *Anal. Chim. Acta* **2005**, *550*, 210.
(17) Cramer, C. J.; Truhlar, D. G. *Chem. Rev.* **1999**, *99*, 2161.
(18) Some examples: (a) Fdez Galvan, I.; Sanchez, M. L.; Martin, M. E.; Olivares del Valle, F. J.; Aguilar, M. A. *Comput. Phys. Comm.* **2003**, *155*, 244. (b) Fdez Galvan, I.; Sanchez, M. L.; Martin, M. E.; Olivares del Valle, F. J.; Aguilar, M. A. *J. Chem. Phys.* **2003**, *118*, 255. (c) Riccardi, D.; Schaefer, P.; Cui, Q. *J. Phys. Chem. B* **2005**, *109*, 17715. (d) Kaminski, G. A. *J. Phys. Chem. B* **2005**, *109*, 5884.
(19) For example: Mora-Diez, N.; Senent, M. L.; García, B. *Chem. Phys.* **2006**, *324*, 350−358.
(20) Famini, G. R.; Wilson, L. Y. *J. Phys. Org. Chem.* **1999**, *12*, 645.
(21) Tabachnick, B. G.; Fidell, L. S. *Using Multivariate Statistics*, 4th ed.; Allyn and Bacon: Needham Heights, MA, **2001**.
(22) Fleming, I. *Frontier Orbitals and Organic Chemical Reactions*; John Wiley & Sons: New York, **1976**.
(23) Klopman, G. *J. Am. Chem. Soc.* **1968**, *90*, 223.
(24) Salem, L. *J. Am. Chem. Soc.* **1968**, *90*, 543.