Prediction of the Basicities of Pyridines in the Gas Phase and in Aqueous Solution

Glenn I. Hawe,[†] Ibon Alkorta,*,[‡] and Paul L. A. Popelier*,[†]

Manchester Interdisciplinary Biocentre (MIB), 131 Princess Street, Manchester M1 7DN, Great Britain, and School of Chemistry, University of Manchester, Oxford Road, Manchester M13 9PL, Great Britain, Instituto de Química Médica (CSIC), Juan de la Cierva 3, 28006-Madrid, Spain

Received October 13, 2009

The basicities of 125 pyridine derivatives in the gas phase and in water have been correlated with the electron density properties within the framework of quantum topological molecular similarity (QTMS). We used the theory of quantum chemical topology (QCT) to provide ab initio descriptors that are able to predict pK_b values. Partial least squares (PLS) and the machine-learning technique Kriging generated validated models. Properties were considered for systems in their neutral and protonated forms. The compounds were divided into a training set, used to develop the models, and a test set, for which the predicted values of the different models were compared with the experimental ones. The results were found to be good for those compounds with substituents in the meta and para positions, whereas the use of Kriging was required to obtain reasonable results when ortho derivatives were included. The basicity was found to be better described in the gas phase than in water. Special attention was paid to external validation.

1. INTRODUCTION

The existence of links between structural characteristics of a series of compounds and their biological or physicochemical properties is well-known.[1] These links, in the form of quantitative models, are used to understand and predict the properties of molecules that have not been measured or synthesized. Thus, such models are of main interest in a number of scientific fields because their costs are relatively low compared with those of the synthesis of new compounds or the evaluation of their biological activity. The process of developing such models can be divided into two parts. The first part is concerned with the generation of molecular descriptors, whereas the second part involves construction of a mathematical model correlating the descriptors with the (measured) property of interest.

A large variety of molecular descriptors have been described in the literature,[2] from topological descriptors to quantum mechanical ones. In fact, commercial programs (e.g., DRAGON[3]) are available that can calculate thousands of descriptors for a given molecule with the limitation that those molecules should have chemical moieties that are included in the original parametrization of the descriptors. The quantum topological molecular similarity (QTMS) method[4−10] developed in our group uses descriptors based on quantum chemical calculations that do not require any previous parametrization to model properties and biological activities for which electronic factors are important.

The mathematical models used range from simple linear regression to more sophisticated techniques such as artificial neural networks and Kriging. Even though the more sophisticated models are able to provide better predictive results, they often lack a simple interpretation of the most important descriptors used in the correlation.

Pyridines, which are structurally analogous to benzene derivatives,[11] are present in a large assortment of bioactive systems. A number of studies have been devoted to the physicochemical properties of these compounds, and several models have been proposed to correlate the structural characteristics with their basicity in solution. The pK_a values of a series of methyl-substituted pyridines were found to correlate with the number and position(s) of methyl moieties attached to pyridine.[12] The substituent polarizability and field-inductive and resonance effects were invoked to analyze the differences between the free energies of protonation in the gas phase and in solution.[13] Voelkel et al.[14] applied several topological descriptors to model the pK_a values in water of mono- and disubstituted pyridines. Tehan et al.[15] used properties based on semiempirical molecular orbital methods to develop correlations with the basicities of a large number of pyridines using multiple linear regression. More recently, Habibi-Yangjeh et al.[16] calculated 1481 descriptors with the computer package DRAGON in combination with a genetic algorithm, principal component analysis, and artificial neural networks to model the pK_a values of 91 pyridines.

In this article, predictive models of the basicities of pyridines in the gas phase and in water are constructed using QTMS descriptors in conjunction with partial least squares (PLS) and Kriging. The data comprise a training set used to develop the models and a test set to validate those models. In contrast to a typical application of predictive models, the pK_b range of the current predictive models is narrower than the range used to validate it.

2. MATERIAL AND METHODS

2.1. Data Set. All of the pyridines derivatives for which the pK_b values at 25 °C in water are listed in ref 17 were

* Corresponding authors e-mail: ibon@iqm.csic.es (I.A.), pla@manchester.ac.uk (P.L.A.P.).

[†] Manchester Interdisciplinary Biocentre (MIB) and University of Manchester.

[‡] Instituto de Química Médica (CSIC).

10.1021/ci900396k © 2010 American Chemical Society
Published on Web 01/05/2010

chosen to form the training set, which consists of 84 compounds. In addition, those derivatives for which the experimental pK_b was determined at different experimental conditions, mainly at 20 and 30 °C, were assigned to the test set. This can be justified because these pK_b values presumably do not differ significantly from the corresponding values at 25 °C.[17] The test set comprises a total of 41 compounds. Finally, the proton affinities (PAs) in the gas phase for the compounds in the training and test sets were collected from the NIST database.[18]

2.2. Ab Initio Calculations. The geometries of the neutral and protonated forms of all compounds were optimized with the hybrid HF/DFT B3LYP[19] computational method using the 6-31+G(d,p) basis set[20] within the Gaussian 03 package.[21] Frequency calculations were carried out at the same computational level to verify that the geometries obtained correspond to energetic minima. The calculated proton affinity ($\mathrm{PA}_{\mathrm{calc}}$) was obtained as the enthalpy difference between the neutral and the corresponding protonated forms. For the neutral molecules, the molecular electrostatic potential (MEP) was evaluated at the B3LYP/6-31+G(d,p) level, and the minimum in the MEP associated with the lone pair of the pyridine nitrogen was located using the Gaussian 03 program.

2.3. Electron Density Analysis. The electron densities of the molecules obtained at the B3LYP/6-31+G(d,p) level were analyzed with the AIMPAC package[22] to locate and characterize bond critical points (BCPs)[23,24] surrounding the nitrogen atom and ring critical points (RCPs). BCPs and RCPs are special topological points in 3D space at which the gradient of the electron density vanishes. They are the only two types of saddle points (as opposed to maxima or minima) that can exist in three dimensions. A BCP is a minimum in one direction and a maximum in each of the two remaining orthogonal directions. An RCP embodies the opposite situation: a maximum in one direction and a minimum in each of the two remaining orthogonal directions. The Hessian is evaluated and diagonalized, leading to three eigenvalues denoted by λ_1, λ_2, and λ_3 that quantify the local curvature of the electron density at a given point in space. At all of these points, the electron density ρ; the three curvatures λ_1, λ_2, and λ_3; and the kinetic energies G and K were calculated. Hence, for each type of critical point, six descriptors were used in the construction of the predictive model. Two additional descriptors were added for each BCP, thus totalling eight descriptors. They are the bond length (corresponding to the BCP) and the distance of the nitrogen atom from the BCP position, denoted by R_{BCP}. In addition, the Laplacian minimum[25] associated with the position of the lone pair of the pyridine nitrogen atom was located in the neutral molecules, and the electron density properties were derived from it where calculated.

It is appropriate to briefly review the quantum topological descriptors employed in this work. Some time ago, Bader et al.[26] proposed to relate the electron density evaluated at the BCP, ρ_b, to the bond order by an exponential relationship. Howard and Larmarche[27] suggested that this interpretation seems to be successful for carbon−carbon bonds only. However, this assertion still vindicates the use of ρ_b as a quantum chemical descriptor; only its meaning for a general (non-CC) bond might not be straightforward. However, a multiple linear relationship of the type $a + b\rho_b + c\lambda_3 + d(\lambda_1 + \lambda_2)$ has been recommended[27] to predict a topological bond order. In the work making that recommendation,[27] ρ_b and λ_3 were interpreted as measures of σ character, whereas $\lambda_1 + \lambda_2$ was considered a measures of the degree of π character. The kinetic energy densities, denoted by $K(\mathbf{r})$ and $G(\mathbf{r})$, are discussed in great detail in ref 28. Interpreting $K(\mathbf{r})$ in chemical terms is not straightforward, although useful formulas describing its link to the Laplacian and the more "classical" kinetic energy $G(\mathbf{r})$ can be found in ref 28. The electronic characteristics of the RCP have been associated with molecular properties such as the strength of the intramolecular hydrogen bond[29] and have been proposed as aromaticity indexes of the whole system.[30] The Laplacian of the electron density, which corresponds to the scalar derivative of the gradient vector field, determines where electronic charge is locally concentrated (<0) or depleted (>0).[23] The presence of local minima of the Laplacian, in nonbonding regions, has been associated with the presence of lone pairs.[31,32]

2.4. PLS and Kriging. The following properties of the BCPs, RCPs, and Laplacian minima, in the neutral molecules, feature in the construction of the predictive models: ρ, λ_1, λ_2, λ_3, G, and K. In addition, for those bonds considered, the interatomic distance and the distance between the BCP and the nitrogen atom were used as descriptors. In each case, descriptors were scaled according to eq 1 separately for the sets of neutral and protonated compounds

$$\bar{X}_i = \frac{X_i - X_{i,\mathrm{min}}}{X_{i,\mathrm{max}} - X_{i,\mathrm{min}}} \qquad (1)$$

PLS. The program SIMCA[33] was used to perform the PLS analysis of the data set with the default parameters. The PLS procedure generalizes and combines features from principal component analysis and multiple regression.[34,35] The principal components that PLS generates are called latent variables (LVs). To avoid overfitting, a strict test for the significance of each consecutive PLS component is necessary. Two statistics provided by SIMCA-P are the squared correlation coefficient, R^2, and the cross-validated R^2, denoted by Q^2. The generated Q^2 is based on a "leave-one-seventh-of-the-data-out" approach rather than the popular "leave-one-out" approach, which is not recommended because of its overly optimistic assessment. The value of Q^2 for a newly constructed LV is calculated, and if this value is smaller than 0.097, the LV is considered not significant and, consequently, not added to the model. Cross-validation is a practical and reliable method of testing this significance.[35] Cross-validation excludes parts of the data from the model development and compares the values predicted by the corresponding model with the actual values. This procedure is repeated seven times until every observation has been excluded once and only once.

The relative importance of each descriptor to the model was checked using a parameter called the variable importance in the projection (VIP).[36] The VIP is a weighted sum of squares of the PLS weights, taking into account the amount of explained Y variance in each dimension (where Y represents the property of interest). Terms with largest VIP values (>1) are the most relevant for explaining Y.

The final models were also subjected to a randomization test. In this test, the property data (Y) were randomly permuted 20 times keeping the descriptor matrix intact, after which a PLS run was executed. Each randomization and subsequent PLS analysis generated a new set of R^2 and Q^2 values, which were plotted against the correlation coefficient between the original Y values and the permuted Y values. The intercepts for the R^2 and Q^2 lines in this plot, R^2_{int} and Q^2_{int}, are a measure of overfitting. A model is considered valid[37] if $R^2_{\text{int}} < 0.4$ and $Q^2_{\text{int}} < 0.05$.

Kriging Models. Kriging (also known as Gaussian process regression) is a machine-learning method[38] with its origins in geostatistics.[39] It has already received some attention in the QSAR/QSPR literature.[40−44] Here, we summarize the salient features of the Kriging approach, along with details of a novel multiobjective approach to determine the Kriging hyperparameters, presented for the first time in this article. Although we present the Kriging methodology in a self-contained way herein, more details can be found in ref 45.

In the Kriging approach, the following general equation is used for making predictions

$$y(\mathbf{x}) = \sum_{i=1}^{m} \beta_i f_i(\mathbf{x}) + z(\mathbf{x}) \tag{2}$$

where the vector $\mathbf{x} = [x_1\ x_2\ ...\ x_d]^{\mathrm{T}}$ with T indicating the transpose and d representing the number of molecular descriptors. The right-hand side of this equation comprises two parts. The first is a "global term", which is a sum of m basis functions f_i ($i = 1, ..., m$), usually taken to be polynomial terms of the form $x_1^{g_1}x_2^{g_2}\cdots x_d^{g_d}$ with $g_1 + g_2 + \cdots + g_d \le G$ (where G is the order of the polynomial). The exact value of m depends on the order of polynomial used and the number of features (molecular descriptors) in the problem, d. The second part of eq 2, $z(\mathbf{x})$, can be viewed as an "error term", compensating for the inability of the global term to model observed data exactly. It is modeled as a Gaussian process with zero mean and some unknown variance σ^2. The error terms for two different molecules are modeled as being correlated in a way that can be expressed through the difference in the values of their molecular descriptors. Typically, the following correlation function is used

$$\mathrm{Cor}[z(\mathbf{x}^i), z(\mathbf{x}^j)] = \exp(-\sum_{h=1}^{d} \theta_h |x_h^i - x_h^j|^{p_h}) \tag{3}$$

where $\theta_h > 0$ and $0 < p_h \le 2$, $h = 1, 2, ..., d$. The indices i and j refer to individuals of the training set, which are molecules in this study. Both indices run from 1 to n, where n is the number of individuals in the training set. It is convenient to introduce the (symmetric) $n \times n$ matrix $\mathbf{R}$ whose $i-j$th entry is given by eq 3, which gives the correlation between the ith and jth training-set examples. Note that the correlation is expressed at the level of the "y variable" (the dependent variable) rather than the "x variable" (the independent variable). The unknown parameters of the Kriging model are thus β_j, σ^2, θ_i, and p_i. The β_j values are found by fitting to the data using the generalized least-squares method

$$\boldsymbol{\beta} = (\mathbf{F}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{F})^{-1}\mathbf{F}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{y} \tag{4}$$

where $\mathbf{y}$ is an $n \times 1$ vector containing the n observed data. $\mathbf{F}$ is the $n \times m$ matrix whose $i-j$th entry is $f_j(\mathbf{x}^i)$. The values of σ^2, θ_i, and p_i are found by maximizing the likelihood L of the observations, given by

$$L(\boldsymbol{\theta}, \mathbf{p}, \sigma^2 | y^i, i = 1, 2, ..., n) = \frac{1}{(2\pi)^{n/2}(\sigma^2)^{n/2}|\mathbf{R}|^{1/2}} \times \exp\left[-\frac{(\mathbf{y} - \mathbf{F}\boldsymbol{\beta})^{\mathrm{T}}\mathbf{R}^{-1}(\mathbf{y} - \mathbf{F}\boldsymbol{\beta})}{2\sigma^2}\right] \tag{5}$$

In fact, it is the logarithm of the likelihood, readily obtained from eq 5, that is maximized. By setting the derivative of log L with respect to σ^2 equal to zero and solving this new equation, the optimal value of σ^2 can be written in terms of $\boldsymbol{\theta}$ and $\mathbf{p}$. Hence, L only needs to be optimized with respect to $\boldsymbol{\theta}$ and $\mathbf{p}$, which amounts to optimizing $2d$ parameters.

As noted in ref 41, care should be taken when implementing this approach, as the matrix $\mathbf{R}$ (and the matrix $\mathbf{F}$) can sometimes become almost singular. This is characterized by $\mathbf{R}$ having a very large condition number.[46] Indeed, for the data sets used in this work, a tradeoff was observed between the likelihood of observations and the condition number of $\mathbf{R}$. In such scenarios, the parameters maximizing the likelihood might not be desirable because they give the worst condition number. Instead, it might be beneficial to trade some likelihood for a reduction in the condition number of $\mathbf{R}$. In this work, we tackled this problem by simultaneously maximizing the likelihood of the data and minimizing the condition number of $\mathbf{R}$. Optimization was carried out using the multiobjective particle-swarm algorithm MOPSO-CD,[47] with a Pareto-archive size of 40. Taking this approach, a Pareto-optimal set of Kriging parameters was returned by MOPSO-CD, representing different levels of tradeoff between likelihood and condition number. The best set of parameters was selected through testing on the external test set.

Two further points about our approach should also be noted. First, as in refs 42 and 43, we extend the model in eq 2 to deal with noise in the data

$$y(\mathbf{x}) = \sum_{i=1}^{m} \beta_i f_i(\mathbf{x}) + z(\mathbf{x}) + \varepsilon(\mathbf{x}) \tag{6}$$

This has the effect of adding an extra parameter to the diagonal of the correlation matrix $\mathbf{R}$, which has to be fit during the maximization of likelihood and minimization of condition number. Second, we make use of the fact that the Kriging parameters θ_h can be examined to determine which features (molecular descriptors) are most important for making predictions.[48] After building a model using all molecular descriptors, we sequentially eliminate molecular descriptors from the training set, in order of increasing importance, rebuilding the Kriging model at each stage. The results reported in this article are for the best Pareto-optimal Kriging models, built using this feature-selection method.

2.5. External Validation. Once the models were generated, the test sets were used as external validation. The validation strategies checked the reliability of the developed models for their application to a new set of data; confidence of prediction could thus be judged. For external validation, a predictive coefficient R_{pred}^2 was calculated as

$$R_{\text{pred}}^{2} = 1 - \frac{\sum (Y_{\text{obs}} - Y_{\text{pred}})^2}{\sum (Y_{\text{obs}} - \bar{Y}_{\text{Training}})^2} \qquad (7)$$

where Y_{obs} and Y_{pred} represent the observed and predicted property values of the test-set compounds, respectively, whereas $\bar{Y}_{\text{training}}$ represents the mean observed value of the training set. The R_{pred}^{2} value is in part controlled by the magnitude of $(Y_{\text{obs}} - \bar{Y}_{\text{Training}})^2$. The larger this difference, the smaller the quantity subtracted from unity and, hence, the closer R_{pred}^{2} to unity. This difference is, in turn, dependent on the selection of training-set members. Therefore, squared correlation coefficient values between the observed and predicted values of the test-set compounds with intercept (r^2) and without intercept (r_0^2) can be calculated to assess the quality of the prediction. Moreover, the squared regression coefficient (r^2) between observed and predicted values of the test-set compounds does not necessarily indicate that the predicted values are very near to observed property values (there can be considerable numerical differences between observation and prediction even if a good overall intercorrelation is maintained). To better gauge the external predictive capacity of a model, a modified r^2 term (r_m^2), defined before,[49] is used

$$r_{\text{m}}^{2} = r^2\left[1 - \sqrt{(r^2 - r_0^{2})}\right] \qquad (8)$$

Note that r^2 is always larger than r_0^2. In the case of good external predictions, predicted values will be very close to observed property values. In this case, the r^2 value will be very near to the r_0^2 value, and in the best case, r_m^2 will be equal to r^2. An additional statistical parameter called the root-mean-square error of prediction (RMSEP) was calculated as

$$\text{RMSEP} = \sqrt{\frac{\sum (Y_{\text{obs}} - Y_{\text{pred}})^2}{N_{\text{Test}}}} \qquad (9)$$

where N_{Test} indicates the number of test-set compounds.

3. RESULTS AND DISCUSSION

3.1. Prediction of the Proton Affinity. The compounds selected in the present article are collected in Table 1, along with their corresponding experimental pK_b and PA values. In addition, the experimental conditions used for the compounds included in the test set are indicated in the same table.

The first step in this study was the geometry optimization of the molecules in their neutral and protonated forms. In general, the presence of the pyridine ring and the nature of the substituents reduce the possibility of alternative minimum-energy configurations. However, there are two possible complications: tautomeric forms can occur in the neutral derivatives, and alternative protonation sites can exist in the charged species. The former complication has been widely investigated in the literature, especially for the hydroxypyridine/pyridine equilibrium.[50] Here, we looked at the hydroxyl/oxo, amino/imino, and thiol/thione tautomerisms for the monosubstituted compounds in the 2, 3, and 4 positions of the pyridine ring. Table 2 shows that, in two cases (2-pyridone and 2-pyridithione), the C═X arrangement is energetically more stable, whereas in the rest of the cases, the C─XH arrangement is favored. Two additional molecules considered in the data set might exhibit 2-hydroxy/2-oxo tautomerism. They are 2-hydroxy-4-methylpyridine/4-methyl-2-pyridone and 3-ethyl-2-hydroxypyridine/3-ethyl-2-pyridone. The calculations predicted that the most stable form for both is the 2-oxo form, by 3.5 and 5.5 kJ/mol, respectively. Only the most stable tautomer will be considered. Finally, it should be noted that, in the protonated form, the problem of tautomerism disappears because the protonations of both tautomers provide the same structure.

The possibility of alternative protonation sites was checked in the simplest molecules with additional nitrogen atoms, such as aminopyridines and cyanopyridines. The results (Table 3) indicate that, in both cases, protonation on the pyridine ring is favored over protonation on the additional nitrogen atom.

The calculated enthalpy differences between the neutral and protonated forms can be directly compared with the experimental proton affinity (PA) for those systems that are available in the literature (38 molecules). The calculated results (Figure 1 and Table 1) nicely reproduce the experimental data, which is an indication of the good quality of the present calculations. In addition, because of the small differences found between the experimental and calculated PA values, the latter were used for the whole data set (125 molecules) in the development of predictive models.

The calculated geometries of some of the systems considered here were compared with those obtained experimentally in a previous article,[11] showing very good correspondence, especially for those cases where the experimental geometry was obtained in the gas phase.

3.2. Prediction of the Basicity Using the Properties of the Neutral Systems. The systems at hand were divided into two subsets. The first subset contained those molecules with substituents in the position ortho to the pyridine nitrogen atom and is named Ortho. The second subset, with only substituents in the meta and para positions relative to the pyridine nitrogen atom, is designated as the Meta/Para group. The combination of the Ortho and Meta/Para groups is marked as All. The numbers of molecules in the training and test sets of each group are indicated in Table 4. In addition, in the same table the ranges, of PA and pK_b values are indicated.

First, we discuss the results of the PLS analysis, which, at all times, was separate from the subsequent analysis by means of Kriging. The PLS analysis initially considered a set of 6 + 6 + (2 × 8) = 28 descriptors, associated with the Laplacian CP representing the lone pair of the nitrogen of the pyridine (6 descriptors), the aromatic RCP (6 descriptors), and the N1−C2 and N1−C6 BCPs (8 descriptors for each BCP). Additional PLS analyses were carried out considering the most important descriptors for the initial model (VIP > 1). Table 5 and Figure 2 show PLS models evaluated with only the descriptors from each type of descriptor, that is, the Laplacian CP, RCP, or each BCP, in turn. In this discussion, the square correlation coefficient between the predicted and observed values in the test set is used as a measure of the quality of the model. A complete statistical description of each model is given in the Supporting Information.

It is interesting to note that the most important features in the Meta/Para subset, for both the PA pK_b, are those associated with the Laplacian CP positions associated with

Table 1. Experimental (pK_b and PA) and Calculated (PA) Values for the Systems Considered in the Training and Test Sets

		PA (kJ/mol)	
	pK_b (exp)	experimental	calculated
	Molecules in the Training Set		
2,4,6-trimethylpyridine	6.57		980.5
2,4-dimethylpyridine	7.26	962.9	966.0
2,5-dimethylpyridine	7.57	958.8	961.4
2,6-dimethylpyridine	7.29	963	964.8
2,6-di-*tert*-butylpyridine	10.42	982.9	989.3
2-acetylpyridine	11.36		901.2
2-amino-3-methylpyridine	6.76		960.9
2-amino-4-methylpyridine	6.52		966.1
2-amino-5-methylpyridine	6.78		962.7
2-amino-6-methylpyridine	6.59		964.6
2-aminopyridine	7.29	947.2	950.2
2-benzylpyridine	8.87		957.8
2-bromopyridine	13.29	904.8	905.6
2-chloropyridine	13.51	900.9	899.4
2-cyanopyridine	14.26	872.9	872.6
2-ethylpyridine	8.11	952.4	952.2
2-fluoropyridine	14.45	884.6	882.7
2-hydroxy-4-methylpyridine	9.47		930.9
2-hydroxypyridine	12.75		895.8
2-isopropylpyridine	8.17		955.5
2-methoxycarbonylpyridine	11.79		933.8
2-methoxypyridine	10.94	934.7	935.1
2-methylpyridine	8.04	949.1	948.1
2-(N-ethylmethanesulfonamido)pyridine	12.27		928.6
2-(N-methoxyacetamido)pyridine	11.99		911.1
2-(N-methylbenzamido)pyridine	12.56		981.2
2-propylpyridine	7.7		955.8
2-pyridinealdoxime	10.44		941.7
2-pyridinecarbaldehyde	10.16		887.2
2-*tert*-butylpyridine	8.24	961.7	963.3
2-vinylpyridine	9.02		943.3
3,4-dimethylpyridine	7.53	957.3	961.8
3,5-dimethylpyridine	7.91	955.4	957.8
3-acetamidopyridine	9.63		924.4
3-acetylpyridine	10.74		919.7
3-aminopyridine	7.97	954.4	958.4
3-bromopyridine	11.15	910	909.1
3-chloropyridine	11.16	903.4	905.3
3-cyanopyridine	12.55	877	873.8
3-ethoxycarbonylpyridine	10.65		928.0
3-ethyl-2-hydroxypyridine	9		911.1
3-fluoropyridine	11.03	902	900.7
3-hydroxypyridine	9.2	929.5	927.0
3-methoxycarbonylpyridine	10.87		923.6
3-methoxypyridine	9.09	942.7	943.3
3-methylpyridine	8.32	943.4	944.5
3-(N-ethylmethanesulfonamido)pyridine	10.06		932.8
3-(N-methoxyacetamido)pyridine	10.48		915.1
3-(N-methylbenzamido)pyridine	10.34		937.7
3-pyridinealdoxime	9.93		933.6
3-pyridinecarbaldehyde	10.2		901.7
3-pyridinecarbamidenicotinamide_	10.67	918.3	922.7
3-*tert*-butylpyridine	8.18		957.2
4-acetylpyridine	10.5		916.7
4-amino-3-bromomethylpyridine	6.53		969.0
4-amino-3-methylpyridine	4.57		995.2
4-aminopyridine	4.89	979.7	984.6
4-bromopyridine	10.29	917.8	919.0
4-chloropyridine	10.17	916.1	915.3
4-cyanopyridine	12.1		880.6
4-ethoxycarbonylpyridine	10.55		926.8
4-ethoxypyridine	7.33		969.2
4-ethylpyridine	8.13		952.5
4-formyl-3-hydroxypyridine	9.95		902.1
4-hydroxypyridine	10.77		946.9
4-isopropylpyridine	7.98	955.7	956.1
4-methoxycarbonylpyridine	10.74	926.6	922.7
4-methoxypyridine	7.53	961.7	962.9
4-methylpyridine	8	947.2	949.3

Table 1. Continued

	pK_b (exp)	PA (kJ/mol)	
		experimental	calculated
4-(N-ethylmethanesulfonamido)pyridine	8.86		967.5
4-(N-methoxyacetamido)pyridine	9.38		968.8
4-(N-methylbenzamido)pyridine	9.32		969.4
4-pyridinealdoxime	9.27		936.9
4-pyridinecarbaldehyde	9.26	904.6	901.1
4-*tert*-butylpyridine	8.01	957.7	959.4
4-vinylpyridine	8.38		951.7
6-methylpyridine-2-carboxylic acid	8.17		932.3
pyridine	8.83	930	930.1
pyridine-2,3-dicarboxylic acid	11.64		883.4
pyridine-2,4-dicarboxylic acid	11.77		890.8
pyridine-2,6-dicarboxylic acid	11.84		911.9
pyridine-2-carboxylic acid	12.99		913.8
pyridine-3-carboxylic acid	11.93		910.8
pyridine-4-carboxylic acid	12.16		910.2
Molecules in the Test Set			
2,3,6-trimethylpyridine ($\mu = 0.5$)	6.40		976.4
2,4-dihydroxypyridine (20 °C)	12.63		927.4
2-(2-hydroxyphenyl)pyridine (20 °C)	9.81		954.3
2,3,5,6-tetramethyl-4-amino-pyridine (20 °C)	3.42		1028.8
2,3,5,6-tetramethyl-4-methylaminopyridine (20 °C)	3.94		1028.6
2,3,5,6-tetramethylpyridine (20 °C)	6.10		986.8
2-aminomethylpyridine ($\mu = 0.5$)	11.69		982.4
2-carbamoylpyridine (20 °C)	11.90		917.4
2-mercaptopyridine (20 °C)	15.07		914.5
2-methylthiopyridine (20 °C)	10.41	937.8	933.5
2-nitropyridine (μ^{a}=0.02)	16.06		865.9
3,5-dimethyl-4-(dimethylamino)-pyridine (20 °C)	5.88		1005.0
3,5-dimethyl-4-methylamino)pyridine (20 °C)	4.04		1011.3
3-aminomethyl-6-methylpyridine (30 °C)	5.30		976.6
3-bromo-4-(dimethylamino)pyridine(20 °C)	7.48		971.4
3-bromo-4-methylaminopyridine (20 °C)	6.51		974.2
3-ethyl-4-methylamino)pyridine (20 °C)	4.10		1009.5
3-ethyl-6-methylpyridine (20 °C)	7.49		964.3
3-ethylpyridine (20 °C)	8.20		948.0
3-hydroxy-2-hydroxymethylpyridine (20 °C, $\mu = 0.2$)	9.00		935.4
3-hydroxy-4-hydroxymethylpyridine (20 °C, $\mu = 0.2$)	9.00		955.4
3-isopropyl-4-methylamino)pyridine (20 °C)	4.04		1014.2
3-isopropylpyridine (20 °C)	8.28		951.2
3-mercaptopyridine (20 °C)	11.74		924.5
3-(methylamino)pyridine (30 °C)	5.30		967.5
3-methylthiopyridine (20 °C)	9.58	936.5	941.7
3-nitropyridine ($\mu = 0.02$)	13.21		866.2
4-amino-3,5-dimethylpyridine (20 °C)	4.46		1005.0
4-amino-3-ethylpyridine (20 °C)	4.49		998.0
4-amino-3-isopropylpyridine (20 °C)	4.46		1002.8
4-carbamoylpyridine (20 °C)	10.39		918.7
4-(dimethylamino)pyridine (20 °C)	7.91		1005.9
4-(dimethylamino-3,5-dimethylpyridine (20 °C)	5.85		999.5
4-(dimethylamino)-3-ethylpyridine (20 °C)	5.34		1003.5
4-(dimethylamino)-3-isopropylpyridine (20 °C)	5.73		1002.9
4-(dimethylamino)-3-methylpyridine (20 °C)	5.32		1000.3
4-mercaptopyridine (20 °C)	12.57		942.8
4-(methylamino)pyridine (20 °C)	4.35		996.9
4-methylamino-3-methylpyridine (20 °C)	4.17		1006.8
4-methylthiopyridine (20 °C)	8.06	955.2	960.9
4-nitropyridine ($\mu = 0.02$)	12.77	874.3	868.2

[a] μ represents the ionic strength of the experimental measurement.

the lone pair of the pyridine nitrogen atom. In contrast, the models for the Ortho and All groups do not contain any significant contribution of the Laplacian CP. Indeed, in several cases, the SIMCA program is not able to develop any reasonable model using only those features for these two sets. The best results were obtained for the Meta/Para subset, especially in the case of the PA. The PA values were obtained in the gas phase, but the pK_b values can be influenced by

Table 2. Relative Energies of the Tautomers (kJ/mol)

position of the substituent	OH vs =O	NH_2 vs =NH	SH vs =S
2	–2[a]	59	–12[a]
3	52	124	44
4	7	71	6

[a] Negative value indicates that the C=X tautomer is more stable.

Table 3. Relative Energies (kJ/mol) of the Alternative Protonated Positions Compared to Protonation on the Nitrogen of the Pyridine Ring

position of the substituent	aminopyridines	cyanopyridines
	pyridineH$^+$ vs NH$_3$$^+$	pyridineH$^+$ vs CNH$^+$
2	82	74
3	112	84
4	156	107

the solvent, in this case water, whereas the descriptors used correspond to the molecules in vacuum.

The properties of the RCPs are the most important ones in the Ortho and All sets. Finally, it is interesting to note that, in all cases, the N1−C6 BCP descriptors show better correlations with basicity than the descriptors associated with the N1−C2 BCP. In previous studies,[5,10] it has also been noticed that descriptors of distant BCPs are important, although they were considered as a "contamination" in the model. However, here, the proximity of the BCPs to the protonation site indicates that there is some underlying difference in the discrimination between the information of the N1−C2 and N1−C6 bonds.

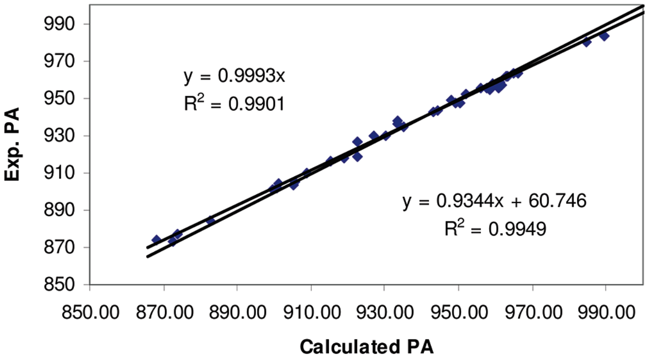

Figure 1. Calculated vs experimental proton affinities (kJ/mol).

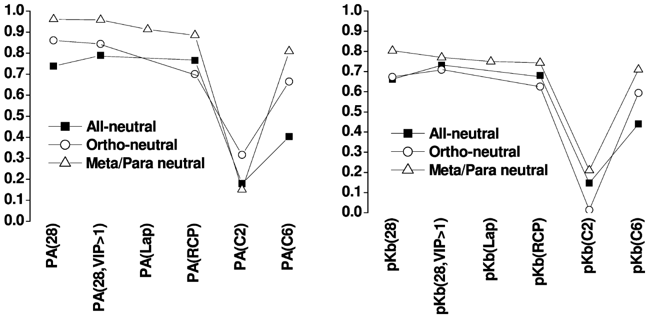

Figure 2. r^2 correlation coefficients between the observed and predicted values of the test set for various PLS models developed using the training test compounds and the descriptors derived from the neutral compounds.

Table 6 reports on the predictivity of the Kriging models using the feature selection applied on the 28 descriptors mentioned above. The results obtained with polynomial degree 1 were always better than those obtained with polynomial degree 0. This fact can be explained by the behavior of Kriging models when used for extrapolation because, in that case, the error term in the Kriging prediction formula (eq 2) tends toward zero. The accuracy of the prediction then depends on the accuracy of the global polynomial term. Hence, the models with polynomial degree 1 are more accurate when extrapolation is required, as is the case for several compounds included in the test set.

Table 4. Number of Molecules in Each Group and Range of PA (kJ/mol) and pK_b Values

	number of cases		PA range		pK_b range	
	training	test	training	test	training	test
Ortho	37	14	989.3–872.6	1028.8–865.9	14.45–6.52	16.06–3.42
Meta/Para	47	27	995.2–873.7	1014.2–866.2	12.55–4.57	13.21–4.04
All	84	41	995.2–872.6	1028.8–865.9	14.45–4.57	16.06–3.42

Table 5. r^2 Correlation Coefficients between the Observed and Predicted Values of the Test Set for Various PLS Models Developed Using the Training Test Compounds and the Descriptors Derived from the Neutral Compounds[a])

property and descriptors used	All	Ortho	Meta/Para	property and descriptors used	All	Ortho	Meta/Para
PA (28)	0.738	**0.861**	**0.961**	pK_b (28)	0.661	0.673	**0.803**
PA (28, VIP > 1)	**0.789**	0.844	0.958	pK_b (28, VIP > 1)	**0.731**	**0.708**	0.770
PA (Lap)	—[b]	—[b]	0.913	pK_b (Lap)	—[b]	—[b]	0.749
PA (RCP)	0.765	0.700	0.885	pK_b (RCP)	0.681	0.625	0.743
PA (C2)	0.179	0.316	0.151	pK_b (C2)	0.147	0.013	0.210
PA (C6)	0.403	0.665	0.809	pK_b (C6)	0.440	0.593	0.709

[a] Model with the largest r^2 value for each subset indicated in bold. [b] SIMCA program is not able to develop any model using these descriptors for the training set.

Table 6. r^2 Correlation Coefficients between the Observed and Predicted Values of the Test Set for the Multiobjective Kriging Models and Polynomial Degrees 0 and 1, Applied to the Set of Neutral Molecules[a]

	PA			pK_b		
polynomial degree	All	Ortho	Meta/Para	All	Ortho	Meta/Para
0	0.796	0.754	0.957	0.729	0.480	0.824
1	**0.889**	**0.882**	**0.970**	**0.782**	**0.759**	**0.843**

[a] Model with the largest r^2 value for each subset indicated in bold.

A comparison with the PLS models shows that the best Kriging model was always significantly better than the corresponding best PLS model for each given molecular subset and property considered. As before, the Meta/Para subset gave the best predictive values. The number of optimal features varied from 15 for the prediction of pK_b for the Ortho set to 28 for the prediction of pK_b for the All set. A minimum of two features was selected for a given molecular feature (Laplacian CP, RCP, and each BCP) in each of the optimal models.

3.3. Prediction of the Basicity Using the Properties of the Protonated Molecules. Predictive models were built using as descriptors the properties of the RCP (6 descriptors), and the N1−H, N1−C2 and N1−C6 BCPs (8 descriptors for each BCP) of the protonated molecules. Thus, the total number of descriptors per molecule in this case was 6 + (3 × 8) = 30. The main difference with the descriptors in the neutral molecules is that, for the protonated molecules, the N1−H BCP values were used whereas, for the neutral molecules, we used properties evaluated at the Laplacian CP positions. The PLS models obtained in these cases provided the r^2 values reported in Table 7 and Figure 3. As before, the best models were obtained for the Meta/Para subset. Note the poor predictivity for the pK_b values of the Ortho subset (the best model provided an r^2 value of 0.36). For the Meta/Para subset, the NH descriptors were the most important ones. This is reminiscent of the neutral compounds, for which the descriptors evaluated at the Laplacian CP were the most important ones.

As in the case of the neutral molecules, the Kriging models showed better predictive r^2 values (Table 8) than the corresponding PLS models. The improvement for the Ortho subset in the prediction of both PA and pK_b values was dramatic. Both models used a global term of polynomial degree 0. The procedure of feature selection (described in section 2.4) yielded a pK_b model with four features, namely, three descriptors associated with the N1−C6 BCP and one associated with the N1−H BCP. On the other hand, the procedure of feature selection generated 22 features for the PA model. The rest of the models, which correspond to polynomial degree 1, used between 10 and 21 features after feature selection.

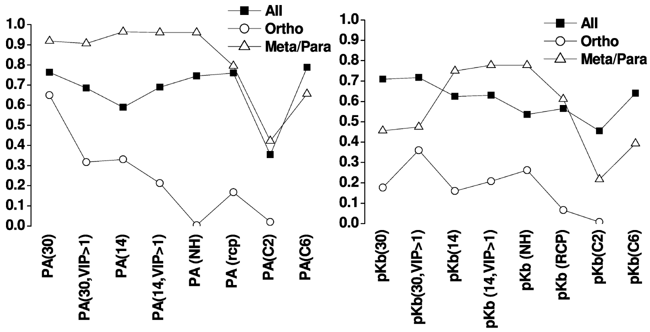

Figure 3. r^2 correlation coefficients between the observed and predicted values of the test set for the PLS model developed using the training test compounds and the descriptors derived from the protonated compounds.

A comparison of the r^2 values of the Kriging models using the neutral and protonated properties shows that, in three cases, the neutral properties were slightly better than the protonated ones whereas the opposite occurred in another three cases. The largest difference between neutral and protonated molecules was observed in the predictability of the PA values. The All set yielded an r^2 value of 0.970 (Table 8) for the protonated molecules and a value of 0.889 (Table 6) for the neutral compounds.

4. CONCLUSIONS

Electron density properties of neutral and protonated pyridines have been used to construct models of their basicities in the gas phase and in water. For that purpose, PLS and Kriging models were used. The compounds were divided into two subsets: those with substituents in the ortho position, the Ortho subset, and the rest, the Meta/Para subset. The Meta/Para set was properly described by PLS and Kriging, whereas the Ortho set was only properly described using Kriging. The prediction of the basicities was found to be more difficult in water than in the gas phase. In general, the descriptors derived from the neutral or protonated molecules provided predictive models of similar quality.

Table 7. r^2 Correlation Coefficients between the Observed and Predicted Values of the Test Set for Various PLS Models Developed Using the Training Test Compounds and the Descriptors Derived from the Pronated Compounds[a])

	All	Ortho	Meta/Para		All	Ortho	Meta/Para
PA (30)	0.763	**0.650**	0.919	pK_b (30)	0.710	0.177	0.456
PA (30, VIP > 1)	0.685	0.317	0.906	pK_b (30, VIP > 1)	**0.718**	**0.360**	0.475
PA (NH)	0.745	0.004	**0.961**	pK_b (NH)	0.536	0.262	**0.778**
PA (RCP)	0.760	0.167	0.795	pK_b (RCP)	0.565	0.066	0.611
PA (C2)	0.355	0.020	0.423	pK_b (C2)	0.455	0.008	0.217
PA (C6)	**0.788**	–[b]	0.656	pK_b (C6)	0.640	–[b]	0.393

[a] Model with the largest r^2 value for each subset indicated in bold. [b] SIMCA program is not able to develop any model using these descriptors for the training set.

Table 8. r^2 Correlation Coefficients between the Observed and Predicted Values of the Test Set for the Multiobjective Kriging Models and Polynomial Degrees 0 and 1, Applied to the Set of Protonated Molecules[a]

	PA			pK_b		
polynomial degree	All	Ortho	Meta/Para	All	Ortho	Meta/Para
0	0.861	**0.834**	0.963	0.732	**0.632**	0.766
1	**0.970**	0.806	**0.974**	**0.784**	0.460	**0.786**

[a] Model with the largest r^2 value for each subset indicated in bold.

ACKNOWLEDGMENT

This work was supported by the Spanish DGI Project CTQ2009-13129-C02-02 and Project MADRISOLAR, ref S-0505/PPQ/0225, of the Comunidad Autónoma de Madrid. Thanks are given to the CTI (CSIC) for the allocation of computer time. I.A. thanks the CSIC for a travel grant that allowed him to stay at the University of Manchester during the development of this article. Alex Harding is thanked for his useful comments.

Supporting Information Available: Values of all statistical parameters employed in this study for all models generated by PLS and Kriging for the proton affinity or pK_b of neutral and protonated molecules. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Gupta, S. P. *QSAR and Molecular Modeling Studies in Heterocyclic Drugs I*; Springer-Verlag: Berlin, 2006.

(2) Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; Wiley-VCH: Weinheim, Germany, 2000.

(3) Tetko, I. V.; Gasteiger, J.; Todeschini, R.; Mauri, A.; Livingstone, D.; Ertl, P.; Palyulin, V. A.; Radchenko, E. V.; Zefirov, N. S.; Makarenko, A. S.; Tanchuk, V. Y.; Prokopenko, V. V. Virtual computational chemistry laboratory—Design and description. *J. Comput.-Aided Mol. Des.* **2005**, *19*, 453–463.

(4) Chaudry, U. A.; Popelier, P. L. A. Ester hydrolysis rate constant prediction from quantum topological molecular similarity (QTMS) descriptors. *J. Phys. Chem. A* **2003**, *107*, 4578–4582.

(5) O'Brien, S. E.; Popelier, P. L. A. Quantum Molecular Similarity. Part 3: QTMS Descriptors. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 764–775.

(6) O'Brien, S. E.; Popelier, P. L. A. Quantum Topological Molecular Similarity. Part 4: A QSAR study of Cell Growth Inhibitory Properties of Substituted (*E*)-1-Phenylbut-1-en-3-ones. *J. Chem. Soc., Perkin Trans. 2* **2002**, 478–483.

(7) Smith, P. J.; Popelier, P. L. A. Quantum Chemical Topology (QCT) Descriptors as Substitutes for Appropriate Hammett Constants. *Org. Biomol.Chem.* **2005**, *3*, 3399–3407.

(8) Smith, P. J.; Popelier, P. L. A. Quantitative structure−activity relationships from optimized ab initio bond lengths: Steroid binding affinity and antibacterial activity of nitrofuran derivatives. *J. Comput.-Aided Mol. Des.* **2004**, *18*, 135–143.

(9) Popelier, P. L. A.; Smith, P. J. QSAR Models based on Quantum Topological Molecular Similarity. *Eur. J. Med. Chem.* **2006**, *41*, 862–873.

(10) Popelier, P. L. A.; Chaudry, U. A.; Smith, P. J. Quantum Topological Molecular Similarity. Part 5. Further Development with an Application to the Toxicity of Polychlorinated Dibenzo-*p*-dioxins (PCDDs). *J. Chem. Soc., Perkin Trans. 2* **2002**, *2*, 1231–1237.

(11) Blanco, F.; O' Donovan, D.; Alkorta, I.; Elguero, J. Substitution effects on neutral and protonated pyridine derivatives along the periodic table. *Struct. Chem.* **2008**, *19*, 339–352.

(12) Gero, A.; Markham, J. J. Studies on Pyridines: I. The Basicity of Pyridine Bases. *J. Org. Chem.* **1951**, *16*, 1835–1838.

(13) Abboud, J. L. M.; Catalan, J.; Elguero, J.; Taft, R. W. Polarizability effects on the aqueous solution basicity of substituted pyridines. *J. Org. Chem.* **2002**, *53*, 1137–1140.

(14) Jan, S.; Borowiak-Resterna, A.; Voelkel, A. Structure−basicity relationships for pyridine extractants. *J. Chem. Technol. Biotechnol.* **1995**, *62*, 233–240.

(15) Tehan, B. G.; Lloyd, E. J.; Wong, M. G.; Pitt, W. R.; Gancia, E.; Manallack, D. T. Estimation of pKa Using Semiempirical Molecular Orbital Methods. Part 2: Application to Amines, Anilines and Various Nitrogen Containing Heterocyclic Compounds. *Quant. Struct.−Act. Relat.* **2002**, *21*, 473–485.

(16) Habibi-Yangjeh, A.; Pourbasheer, E.; Danandeh-Jenagharad, M. Prediction of basicity constants of various pyridines in aqueous solution using a principal component-genetic algorithm-artificial neural network. *Monatsh. Chem.* **2008**, *139*, 1423–1431.

(17) Dean, J. A. *Lange's Handbook of Chemistry*, 15th ed.; McGraw-Hill: New York, 1999.

(18) NIST Chemistry WebBook; NIST Standard Reference Database Number 69; National Institute of Standards and Technology (NIST): Gaithersburg, MD, 2005; available at http://webbook.nist.gov/chemistry/ (accessed August 3, 2009).

(19) Becke, A. D. Density-Functional Thermochemistry. 3. The Role of Exact Exchange. *J. Chem. Phys.* **1993**, *98* (7), 5648–5652.

(20) Hariharan, P. C.; Pople, J. A. *Theor. Chim. Acta* **1973**, *28*, 213.

(21) 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 E.01; Gaussian, Inc.: Pittsburgh, PA, 2004.

(22) Biegler-Koenig, F. W.; Bader, R. F. W.; Tang, T. H. Calculation of the Average Properties of Atoms in Molecules. 2. *J. Comput. Chem.* **1982**, *3* (3), 317–328.

(23) Bader, R. F. W. *Atoms in Molecules: A Quantum Theory*; Oxford University Press: 1990.

(24) Popelier, P. L. *Atoms in Molecules: An Introduction*; Pearson Education: London, 2000.

(25) Popelier, P. L. A. On the Full Topology of the Laplacian of the Electron Density. *Coord. Chem. Rev.* **2000**, *197*, 169–189.

(26) Cremer, D.; Kraka, E.; Slee, T. S.; Bader, R. F. W.; Lau, C. D. H.; Nguyendang, T. T.; Macdougall, P. J. Description of Homoaromaticity in Terms of Electron Distributions. *J. Am. Chem. Soc.* **1983**, *105* (15), 5069–5075.

(27) Howard, S. T.; Lamarche, O. Description of covalent bond orders using the charge density topology. *J. Phys. Org. Chem.* **2003**, *16*, 133–141.

(28) Bader, R. F. W.; Preston, H. J. T. The Kinetic Energy of Molecular Charge Distributions and Molecular Stability. *Int. J. Quantum Chem.* **1969**, *3*, 327–347.

(29) Grabowski, S. J. Properties of a Ring Critical Pointas Measures of Intramolecular H-Bond Strength. *Monatsh. Chem.* **2002**, *133*, 1373–1380.

(30) Marcin, P.; Tadeusz, M. K. Application of AIM Parameters at Ring Critical Points for Estimation of π-Electron Delocalization in Six-Membered Aromatic and Quasi-Aromatic Rings. *Chem.—Eur. J.* **2007**, *13*, 7996–8006.

(31) Bader, R. F. W.; Heard, G. L. The mapping of the conditional pair density onto the electron density. *J. Chem. Phys.* **1999**, *111*, 8789–8798.

(32) Blanco, F.; Alkorta, I.; Elguero, J. Barriers about Double Carbon−Nitrogen Bond in Imine Derivatives (Aldimines, Oximes, Hydrazones, Azines). *Croat. Chem. Acta* **2009**, *82*, 173–183.

(33) *SIMCA-P 10.0*; Umetrics AB: Umeå, Sweden, 2002; www.umetrics.com.

(34) Wold, S.; Sjostrom, M.; Eriksson, L. *Chemom. Intell. Lab. Syst.* **2001**, *58*, 109–130.

(35) Wold, S. *PLS for Multivariate Linear Modeling* in *Chemometric Methods in Molecular Design*; van de Waterbeemd, H. E., Ed.; VCH: Weinheim, Germany, 1995; pp 195−218.

(36) Chong, I.-G.; Jun, C.-H. Performance of some variable selection methods when multicollinearity is present. *Chemom. Intell. Lab. Syst.* **2005**, *78*, 103–112.

(37) Wold, S.; Sjostrom, M.; Eriksson, L. Partial Least Squares (PLS) in Chemistry. In *Encyclopedia of Computational Chemistry*; Schleyer, P. v. R., Ed.; Wiley: Chichester, U.K., 1998; Vol. 3, pp 2006−2021.

(38) Rasmussen, C. E.; Williams, C. K. I. *Gaussian Processes for Machine Learning*; The MIT Press: Cambridge, MA, 2006.

(39) Cressie, N. *Statistics for Spatial Data*; Wiley: New York, 1993.

(40) Burden, F. R. Quantitative Structure−Activity Relationship Studies Using Gaussian Processes. *J. Chem. Inf. Model.* **2001**, *41*, 830–835.

(41) Fang, K.-T.; Yin, H.; Liang, Y.-Z. New Approach by Kriging Models to Problems in QSAR. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 2106–2113.

(42) Yin, H.; Runze, L.; Fang, K.-T.; Liang, Y.-Z. Empirical Kriging models and their applications to QSAR. *J. Chemom.* **2007**, *21*, 43–52.

(43) Obrezanova, O.; Csanyi, G.; Gola, J. M. R.; Segall, M. D. Gaussian Processes: A Method for Automatic QSAR Modeling of ADME Properties. *J. Chem. Inf. Model.* **2007**, *47*, 1847–1857.

(44) Zhou, P.; Xiang, C.; Wu, Y.; Shang, Z. Gaussian process: An alternative approach for QSAM modeling of peptides. *Amino Acids*, published online Jan 4, 2009; http://dx.doi.org/10.1007/s00726-008-0228-1.

(45) Handley, C. M.; Hawe, G. I.; Kell, D. B.; Popelier, P. L. A. Optimal Construction of a Fast and Accurate Polarisable Water Potential based on Multipole Moments trained by Machine Learning. *Phys. Chem. Chem. Phys.* **2009**, *11*, 6365–6376.

(46) Press, W. H.; Teukolsky, S. A.; Vetterling, W. T.; Flannery, B. P., *Numerical Recipes: The Art of Scientific Computing*, 3rd ed.; Cambridge University Press: Cambridge, U.K., 2007.

(47) Raquel, C. R.; Naval, P. C. J. An Effective Use of Crowding Distance in Multiobjective Particle Swarm Optimization. In *Proceedings of Genetic and Evolutionary Computation Conference (GECCO)*; Beyer, H.-G., O'Reilly, U.-M., Eds.; Association for Computing Machinery: New York, 2005; pp 257−264.

(48) Welch, W. J.; Buck, R. J.; Sacks, J.; Wynn, H. P.; Mitchell, T. J.; Morris, M. D. Screening, Predicting, and Computer Experiments. *Technometrics* **1992**, *34*, 15–25.

(49) Roy, P. P.; Leonard, J. T.; Roy, K. Exploring the impact of the size of training sets for the development of predictive QSAR models. *Chemom. Intell. Lab. Sys.* **2008**, *90*, 31–42.

(50) Alkorta, I.; Elguero, J. Influence of Intermolecular Hydrogen Bonds on the Tautomerism of Pyridine Derivatives. *J. Org. Chem.* **2002**, *67*, 1515–1519.

CI900396K