

## Three-Dimensional Correlation Analysis—A Novel Approach to the Quantification of Substituent Effects

Artem Cherkasov,\* Dennis G. Sprous, and Ridong Chen\*

APT Therapeutics, Inc., 893 North Warson Road, St. Louis, Missouri 63141

Received: April 16, 2003; In Final Form: July 31, 2003

A new method to quantify the substituent effect, called “3D correlation analysis” (3D-CAN), is presented. This approach employs an atomic level of consideration of substituent effects and is developed on the basis of empirical inductive and steric constants. However, unlike traditional correlation analysis, 3D-CAN takes into account the three-dimensional structure of substituents. Extensive datasets of experimental dissociation constants for a broad range of carboxylic acids and protonated amines (including a number of polypeptides) have been accurately reproduced in the framework of the developed novel technique. New formulas allowing calculation of  $pK_a$  values for 826 carboxylic acids and 802 protonated amines have been established, and the possibility of interpretation of the physical nature of the substituent effects within the framework of 3D-CAN is presented in the present paper. This validates the methodology as a powerful technique broadly applicable to general reaction phenomena.

### Introduction

Modeling and predicting dissociation constants of organic compounds has a long history.<sup>1–3</sup> Kirkwood and Westheimer<sup>1</sup> introduced one of the earliest methods in 1938 based on electrostatic theory. Shortly thereafter, the Hammett equation<sup>2</sup> was employed successfully for aromatic systems where extensive resonance is present. Aliphatic systems were later treated with the Taft<sup>3</sup> equation in the 1950s. However, the 1990s saw new emphasis on the need to model the problem from two sources: environmental regulation and the pharmaceutical industry. Under the U.S. Toxic Substances Control Act, every new chemical manufactured or used in the USA must undergo an environmental assessment. Knowledge of the  $pK_a$  or  $pK_b$  of an organic molecule can define the degree of soil/sediment absorption, mobility, reaction kinetics, and complexation. The brutal economics of pharmaceutical mass screening demands that compounds “Fail early. Fail cheaply”<sup>4</sup> with one of the key determinants in failing being poor oral absorption. As in the body as in the earth,  $pK_a$  is a key determinant of fate, as can be seen in the prominence of  $pK_a$  or  $pK_b$  dependent factors on oral bioavailability modeling.<sup>5,6</sup>

With the motive pressing, a variety of  $pK_a$  estimation studies have been done over the past decade. These include quantum based approaches, semiempirical based methods, and two chemoinformatic based approaches. A study done by da Silva et al.<sup>7</sup> at the 6-31G\*\*/HF theory level and a second study by Citra<sup>8</sup> using a semiempirical approach demonstrated  $pK_a$  can be accurately modeled from first principles. In the da Silva et al. study,<sup>7</sup> the  $pK_a$  values of seven compounds were predicted to within a single  $pK_a$  unit. Results from Citra<sup>8</sup> were likewise admirable. However, there is a large CPU-demand associated with such techniques, and for those who deal with even modest databases of hundreds of compounds in environmental assessment or in pharmaceutical research, this is too slow to consider. By necessity, such researchers must turn to methods requiring magnitudes less CPU time. One example is the SPARC

(SPARC: Performs Automated Reasoning in Chemistry) program developed by the U.S. Environmental Protection Agency.<sup>9</sup> The method requires only 2D structure input and can process hundreds of compounds in a minute. The initial study they presented modeled the  $pK_a$  values for 214 redundant dye molecules with a rms error of less than 0.62  $pK_a$  units. A second method was developed by Tsantili-Kakoulidou et al.<sup>10</sup> and is the basis of a commercial program. More recently, Xing and Glen<sup>11</sup> presented a novel 2D fingerprint method. This method was trained over 384 bases and 645 acids with a model Pearson's  $R^2 > 0.92$  in both cases. Both of these methods perform well and perform quickly but are more empirical fits than a physical method.

Cherkasov et al.<sup>12</sup> presented a general method for estimation of reaction phenomena and applied it at the time to model ionization potentials and gas-phase basicity. The present article extends this treatment to  $pK_a$  and  $pK_b$  in general. The theoretical foundation of our present method arises from correlation analysis and the Taft equation.<sup>3</sup> Correlation analysis is one of the most popular and reliable quantitative methods of estimation of practical quantitative structure–activity relationships (QSAR). Empirical correlations evaluating polar (inductive and resonance) and steric substituent effects are based on the principles of linearity of free energy (LFER) and multilinearity (PL) which make it possible to perform mathematical formalization of the relationship between structure and activity.<sup>3,13</sup>

Quantitative description of the polar influence of substituents first became possible within the framework of the approach developed by Hammett on the basis of the dissociation constants of substituted benzoic acids.<sup>2</sup> The difference between the logarithms of the dissociation  $K$  constant of substituted benzoic acid and the corresponding  $K^\circ$  value of the unsubstituted standard compound has been expressed by the empirical equation

$$\Delta\Delta G = \log \frac{K}{K^\circ} = \rho\sigma$$

in which two new quantities have been introduced:  $\sigma$  is the universal constant specific for a substituent in the benzene ring,

\* All correspondence should be addressed to R. Chen. E-mail: rchen@apttherapeutics.com. Telephone: (314) 812-8054. Fax: (314) 812-8127.

and  $\rho$  is the reaction series constant reflecting the sensitivity of the reaction center to variation of substituent influence.

Later, the Hammett equation was modified many times, but the vast majority of these modifications referred to the chemistry of aromatic compounds.<sup>14</sup> For a series of aliphatic compounds, the Hammett relation, as a rule, did not hold. Taft<sup>3</sup> suggested that in this case the steric substituent effects are significant and should be separated as

$$\Delta\Delta G = \rho \sum_i \sigma^* + \delta \sum_i E_s$$

where  $\sigma^*$  is a substituent constant depending only on its inductive influence, and  $E_s$  is the substituent constant reflecting its steric effect. Taft's inductive and steric constants are among the most reliable and widespread substituent parameters. At present, a large number of polar and steric substituent constants for hundreds of diverse substituents have been determined; these constants form dozens of scales, which are extensively used for analysis of molecular reactivity, bioactivity, physicochemical properties, and reaction mechanisms studies.<sup>3,15–19</sup>

In general, the nature of the steric effect is readily understood. An increasing of the bulk of the substituents leads to mechanical shielding of the reaction center from an attacking reagent (steric hindrance of motions), to an increase of steric repulsion in the reaction's transition state (steric strain), or to steric inhibition of solvation.<sup>20</sup> Thus, the methods of calculation of substituents' steric constants usually operate by different descriptors of effective atomic, group, or molecular sizes.<sup>20</sup>

In regard to the nature of the inductive effect, there is no unanimous opinion. There is still no strict mathematical description of the inductive influence, although this is generally reduced to the classic view that electron density distributes from the atom with a lower electronegativity to the atom with a higher electronegativity. Two possible mechanisms for transmission of this effect are discussed in the literature, both having their pros and cons, their adherents and opponents.

The first one, described long ago by Lewis, suggests that the influence is transmitted along the bonds by their consecutive polarization, that is, by a mechanism similar to the electrostatic induction.<sup>2,21</sup>

An alternative mechanism of the transfer of the inductive effect, proposed for the first time by Ingold,<sup>21</sup> involves interaction of functional groups through the space. This induction mechanism, called later the "field effect" (and which has been given the preference in recent years), is purely electrostatic and occurs via ion–ion, ion–dipole, and dipole–dipole interactions, the intensity of which is described by various functions of the distance  $r^{-n}$  ( $n = 1–4$ ).<sup>17,22–25</sup>

Numerous attempts of theoretical calculation of the inductive substituent constant have been historically addressed to electrostatic methods of theoretical evaluation of aqueous acidity. According to the popular Bjerrum–Kirkwood–Westheimer electrostatic theory, the energy of interaction of the anion resulting from the dissociation of an acid containing a polar substituent, with the substituent dipole  $\mu$ , is expressed by the following equation:<sup>26,27</sup>

$$\log \frac{K}{K^\circ} = \frac{1}{4\pi\epsilon_0\epsilon_{\text{eff}}} \frac{Z_1 e \mu \cos \theta}{2.303 k T r^2} \quad (1)$$

where  $K$  and  $K^\circ$  are the dissociation constants for the substituted and unsubstituted acids, respectively,  $Z_1$  is the charge on the reaction center,  $r$  is the distance from the substituent to the reaction center (usually it is an ionizable hydrogen atom),  $\theta$  is

the angle between the  $\mu$  and  $r$  vectors,  $\epsilon_0$  is the standard permittivity of the medium, and  $\epsilon_{\text{eff}}$  is the empirically selected effective dielectric permittivity.

Advances in CPU power opened the way for Poisson–Boltzmann (PB) based approaches quantifying the variation of the electrostatic potential  $\phi(r)$  through space due to a system of point charges embedded in a continuum electrolyte.<sup>26–33</sup>

The PB calculations have been successfully used for the reinvestigation of pK shifts in small polysubstituted molecules (diamines, dicarboxylic acids).<sup>34,35</sup> It should, however, be noticed that the progress of electrostatic continuum models as yet did not implicate new developments in the methodologies of field effect and correlation analysis, which have been originated by survey of ionization constants. On the other hand, known attempts to relate inductive constants directly to atomic charges, accurately calculated by methods of quantum chemistry, did not achieve any general success either.<sup>22,24,36–40</sup> Thus, the methodology of correlation analysis still remained in the framework of the traditional 2D fragmental approach considering a molecule to be consisted of three virtual parts—an active site, a changing remote substituent, and a connecting skeleton. The empirically established correlations between the detected active site's quantities and the corresponding substituent inductive, steric, and resonance constants are still used for quantification of the substituent effect on the basis of linear relationships, which are often considered to be extrathermodynamic correlations, lacking any physical meaning.<sup>14</sup>

## Results and Discussion

### Mathematical Apparatus of 3D Correlation Analysis.

*Quantification of the Inductive Effect.* In the framework of our previous studies we have estimated the quantitative relationships between inductive and steric constants of a substituent, its group electronegativity, and its partial charge distribution. In this work our previous results have been developed into a nonfragmental method of correlation analysis, which allows the consideration of the real three-dimensional structure of substituents. The estimated relationships have also been used for interpretation of the nature of the inductive effect and clarification of the physical meaning of some extrathermodynamic correlations.

As has been found,<sup>18,39,40</sup> the inductive effect of a substituent can be determined by the sum of the inductive influences of its atoms:

$$\sigma^* = \sum_{i=1}^n \frac{\sigma_{A_i}}{r_i^2} \quad (2)$$

where  $\sigma^*$  is Taft's inductive constant of the substituent;  $n$  is the number of atoms in the substituent; and  $r_i$  is the distance from  $i$ -th atom to the reaction center. The introduced empirical parameter  $\sigma_A$  determines the capability of the  $i$ -th atom of exerting the inductive effect depending on the chemical nature of the element and on its valence state.

This semiempirical approach made it possible to describe with a high degree of accuracy the inductive constants of virtually all substituents for which the  $\sigma^*$  constants are available.

On the other hand, the estimated form of eq 2 obviously reflects the electrostatic nature of inductive interactions by underlining the importance of direct intramolecular distances. However, it not necessarily, as one would think, means that the inductive effect may be solely related to the energy of Coulomb electrostatic interactions. In our opinion, the transmission of inductive influence "along bonds" and "through space"

should not be considered as alternative mechanisms. It would be more worthwhile to imagine the united mechanism of inductive effect transmission as follows. Due to atomic electronegativity differences (driving force), the redistribution of electron density is occurring *along bonds*, but only to such a degree in which the arising charges can be effectively stabilized (already *through space*) by Coulomb interactions. On the simplest level, such a situation may be compared with an electric capacitor, where less distance between plates also leads to a more charge. However, in this case nobody says the charge is transferred through space.

It has been found for a broad range of elements that their  $\sigma_A$  constants correlate with the difference in electronegativity between a given element and the reaction center,  $\Delta\chi_{i-RC}$ , reflecting the driving force for the electron density displacement, and with the square of the covalent radius of the element,  $R_i$ , reflecting the ability to delocalize the charge ( $\sigma_A = 7.84\Delta\chi_{i-RC}R_i^2$ ).<sup>21</sup> Thus, Taft's  $\sigma^*$  constant can be presented by the following equation:<sup>40</sup>

$$\sigma^* = 7.84 \sum_{i=1}^n \frac{(\chi_i - 2.10)R_i^2}{r_i^2} = 7.84 \sum_{i=1}^n \frac{\Delta\chi_i R_i^2}{r_i^2} \quad (3)$$

which allows direct calculation of the inductive constant of any substituent at any reaction center from the fundamental characteristics of atoms. Later, the estimated eq 3 became a basis for elaboration of a wide range of other so-called "inductive" reactivity indexes, such as inductive electronegativity and inductive chemical hardness-softness parameters for atoms, groups, and molecules.<sup>41-43</sup> One of the most important developments was the elaboration of the procedure of calculation of the partial charges' distribution:

$$q_i = \alpha \sum_{j \neq i}^{N-1} \frac{(\chi_j - \chi_i)(R_j^2 + R_i^2)}{r_{j-i}^2}$$

where  $N$  is overall number of atoms in the molecule,  $q_i$  is the charge on the atom  $i$ ,  $\alpha$  is the scaling constant,  $R_j$  and  $R_i$  are the atomic radii for atoms  $j$  and  $i$ , respectively, and  $\chi_i$  and  $\chi_j$  are the electronegativity of atoms  $j$  and  $i$ , respectively.<sup>43</sup>

If we consider the estimated formula for the inductive constant in the context of electrostatic approaches, then the empirical atomic parameter  $\sigma_A$ , in a way, should have some relation to the local atomic dipole, while the function  $1/r^2$  represents the electrostatic interactions. But the operational  $\sigma_A$  parameter is the constant for a defined type of atom, depending only on its nature and a valent state, and the background of eq 3 appears to be different.

In the approximation "reaction center - the rest of the molecule", when all  $N - 1$  atoms of the molecule are considered as one sub-substituent, the overall inductive influence of the reaction center RC

$$\sigma_{MOL \rightarrow RC}^{**} = \alpha_0 \sum_{i \neq RC}^N \frac{\Delta\chi_{i-RC}(R_i^2 + R_{RC}^2)}{r_{i-RC}^2} \quad (4)$$

(where  $N$  is the number of atoms in the molecule, and  $R_i$  and  $R_{rc}$  are the atomic radii of a specific atom and of the reaction center), will then be proportional to the sum of the partial charges of the atoms of the rest of the molecule and, thus, the to charge of the reaction center (RC):

$$\sigma_{MOL \rightarrow RC}^{**} = \sum_{i \neq RC}^{N-1} \sigma_i^* = \frac{\alpha_0}{\alpha} \sum_{i \neq RC}^{N-1} q_i = - \frac{\alpha_0}{\alpha} q_{RC}$$

Here  $\sigma_i^*$  is the inductive influence on the  $i$ -th atom from the rest of the molecule, including the RD:

$$\sigma_i^* = \sum_{j \neq i}^{N-1} \frac{\Delta\chi_{ji}(R_i^2 + R_j^2)}{r_{ij}^2}$$

which under summarizing over

$$\sum_{i \neq RC}^{N-1}$$

can be reduced to eq 4:

$$\sum_{i \neq RC}^{N-1} \sigma_i^* = \sum_{i \neq RC}^{N-1} \sum_{j \neq i}^{N-1} \frac{\Delta\chi_{ji}(R_i^2 + R_j^2)}{r_{ij}^2} = \sum_{i \neq RC}^{N-1} \frac{\Delta\chi_{i-RC}(R_i^2 + R_{RC}^2)}{r_{RC-i}^2}$$

In the more traditional approximation "reaction center (RC) - skeleton - substituent (R)", the form of the equation will remain virtually the same

$$\sigma_{R \rightarrow RC}^* = \sum_{i \neq RC}^{N-1} \sigma_i^* = \sigma_{skel}^* + \sum_{j \in R}^n \sigma_j^* = \text{constant} + \frac{\alpha_0}{\alpha} \sum_{j \in R}^n q_j^* \quad (5)$$

where the unchanged part of molecules of the reaction series gives the constant term  $\sigma_{skel}^*$  of eq 5 while the inductive effect of the  $n$ -atomic substituent on the reaction center  $\sigma_{R \rightarrow RC}^*$  remains proportional to the charge of the substituent.

The estimated relationships underline a general similarity between the influence of substituents on the free energy change and that on the electron density distribution of the molecule. It should also be stressed that the idea that the inductive constant can be directly related to the partial charge has been repeatedly pronounced<sup>22,36-38</sup> (although it has also been pointed out that such relations should be considered merely as empirical, unjustified by physical laws).<sup>24</sup>

Thus, if we present the energy in the traditional way,  $\sigma\rho$ , when  $\sigma$  is identified with partial charges

$$\Delta G = \Delta G^\circ + \rho \sum_i \sigma_i^* = \Delta G^\circ + (\text{const})\rho \sum_i q_i$$

then the assumption of the merely electrostatic nature of  $\Delta G$  implies that the  $\rho$  parameter should also be a function of intermolecular distances.

In articles that followed the publication of our method of calculation of inductive constants,<sup>24,25,44</sup> authors have statistically tested  $\rho$  reaction series constants in a similar manner to that with which we have previously examined  $\sigma$  values. The  $\rho$  constants for various ionization equilibria have also been estimated as functions of  $1/r^2$  and  $1/n^2$ , where  $r$  represents the direct distance between the substituent and the reaction site and  $n$  is the number of bonds between them (which generally should be well proportional to  $r$ ). It should be emphasized that the  $\mu$

and  $\cos \theta$  parameters have been neglected in these studies (in contrast to numerous similar, but less successful, previous attempts). We could assume, however, that the accuracy of the corresponding procedure may be improved by using some function of averaged distances from the reaction site to all atoms of substituents  $f(1/r)$ , rather than just  $1/r^2$ .

Indeed, if we identify the inductive constant of a substituent with the sum of the partial charges on its atoms, then the  $\rho\sigma$  expression for free energy change may have the form

$$\Delta\Delta G = \Delta G - \Delta G^\circ = \rho \sum \sigma^* \approx \left( \frac{q_{RC}}{\epsilon(\bar{r})\bar{r}} \right) \sum_{i \neq RC}^{N-1} q_i \quad (6)$$

(where  $q_i$  itself depends on the distance to the reaction center,  $r$  is the distance from the reaction center, and  $\epsilon(r)$  is the distance dependent permittivity function). However, it should be clearly distinguished that the parameter of the averaged distance is included in eq 6 as the constant, which does not change its proportionality to the  $1/r^2$  factor. When varying substituents are insulated from the indicating group (RC) by the bulk skeleton, the distances from the RC to substituent atoms will not significantly deviate from their mean value and, approximately, eq 6 should hold.

Thus, the electrostatic change of free energy may be approximated by the sum of the averaged energies of atomic Coulomb interactions. In the presented eq 6 we have used the distance dependent form of the effective dielectric media permittivity  $\epsilon(r)$ , as is required by the Laplace equation.<sup>45</sup> Thus, the general form of the Coulomb equation, containing the distance dependent parameter  $\epsilon(r)$

$$\Delta G = \sum_{i \neq j} \frac{q_i q_j}{\epsilon(r_{ij}) r_{ij}}$$

makes the analysis of  $\rho$  in terms of intramolecular distances even less obvious, especially considering the relativity of identification of different parts of a molecule as its reaction center, skeleton, and substituent(s).

It remains unclear, though, whether the applicability of the inductive constants can be validated in a similar formal way for other reaction series where the free energy change may be controlled by other than electrostatic factors. The investigation of possible relations between the  $\sigma^*$  constants and the molecular electronic density distribution and integration of the elaborated approaches with advanced PB methods are underway.

**Quantification of the Steric Effect.** The practical application of eq 2 allowed application of multiple linear regression (MLR) to interpret a variety of gas phases containing up to several hundred entries: ionization potentials and electron affinities of C-, N-, O-, and S-centered free radicals,<sup>46,47</sup> and ionization energies and gas basicities of organic amines have been successfully quantified in terms of inductive interactions.<sup>12</sup> In a similar way, the substituent effect of the energies of CH bonds in a wide series of substituted carbohydrogens has also been described.<sup>48</sup> However, the broader application of the developed MLR technique (not restricted by the gas-phase data, mainly controlled by field effects) requires additional consideration of steric and resonance effects.

This problem can be solved by using a previously suggested model of the frontier steric effect, which assumes the frontal character of steric interactions<sup>33,49</sup> and specifies the steric constant  $R_s$  as the specific surface, screened on the reaction center by all atoms of the substituent:

$$R_s = -30 \log \left( 1 - \sum_{i=1}^n \frac{R_i^2}{4r_i^2} \right) \quad (7)$$

where  $n$  is the number of atoms in the substituent,  $R_i$  is the radius of the  $i$ -th atom, and  $r_i$  is the direct distance between the  $i$ -th atom and the reaction center. The normalizing coefficient 30 was introduced for transformation of  $R_s$  values to the scale of steric Taft constants. The calculated  $R_s$  substituent parameters correlate well with the steric empirical scales  $E_s$ ,  $E_s^\circ$ , and  $V_x$  for all the possible ranges of changes of the steric effect.

**Resonance Contributions.** The presented models allow quantification of only steric and inductive effects, and resonance effects cannot be directly accounted for. However, our previous investigations show that the additive model for the inductive effect describes fairly well substituent effects in conjugated systems though some of them can be treated as exceptional due to strong direct polar conjugation or saturation effects.<sup>14</sup>

**General Consideration of the Substituent Effect.** The fact that both the steric and inductive effects of substituents have been established as functions of the inverse square of the distance between the reaction center and the substituent atom can be used in the following manner. It is known that if the argument  $x$  is small enough, then the function  $\lg(1-x)$  is linearly related to  $x$ . Consequently, eq 7 can be transformed into a simpler one for the  $R'_s$  steric parameter:

$$R'_s = \sum_{i=1}^n \frac{R_i^2}{4r_i^2} \quad (8)$$

At the same time, according to the LFER principle, inductive and steric effects normally do not interfere and are used as independent contributions to the two-parameter Taft equation (eq 1).

Therefore, when in the system "reaction center (RC) – the rest of the molecule", inductive and steric effects on the RC are explored in terms of distance dependent atomic contributions. Formulas 2 and 8 and eq 1 can be written as

$$\Delta\Delta G = \text{const}_1 \sum_{i \neq RC}^{N-1} \frac{\sigma_{Ai}}{r_{RC-i}^2} + \text{const}_2 \sum_{i \neq RC}^{N-1} \frac{R_i^2}{r_{RC-i}^2} = \sum_{i \neq RC}^{N-1} \frac{g_i}{r_{RC-i}^2} \quad (9)$$

where  $N$  is the number of atoms in the molecule,  $r_{RC-i}$  is the distance between atom  $i$  and the reaction center (RC), and  $g_i$  is the ability of an atom of a certain type to contribute to the overall  $\Delta\Delta G = \Delta G - \Delta G^\circ$  value.

This approximation enables us to check if there exists  $g_i$  parameter for a given atom. If these exist, the next question is to analyze the possible physical meaning. The formalism of the proposed technique suggests that each atom, common for all molecules of a reaction series, can be explored as a hypothetical reaction center (RC). In this case, for every single compound of a series, all its  $N-1$  atoms, except the RC, can be taken as one sub-substituent, which can be treated by eq 9. It should also be highlighted that the elaborated procedure of 3D correlation analysis allows consideration of any free energy related quantities  $\Delta G$  without a priori specification of the standard value  $\Delta G^\circ$ , which in this case can be statistically established as the intercept of the regression

$$\Delta G = \Delta G^\circ + \sum_{i \neq \text{RC}}^{N-1} \frac{g_i}{r_{\text{RC}-i}^2} \quad (10)$$

The procedure of 3D correlation analysis consists of three simple consecutive steps:

1. *Input.* Structural files for the optimized geometries of the molecules of the reaction series should be prepared, where each atom is specified with its number and three spatial coordinates. The atomic types, for which operational parameters  $g$  in eq 9 will be estimated, should also be specified in accordance with the atom's nature and valent state. The ionization state and specific molecular environment of atoms may be taken into consideration by the introduction of the corresponding atomic types. If a reaction series contains  $M$  molecules, then the input of  $M$  structural files should be prepared. For each molecule  $j$ , its atom reaction center ( $\text{RC}_j$ ) needs to be specified by placing the corresponding atomic number into the  $[\text{RC}_1, \text{RC}_j, \text{RC}_M]$  vector.

2. *R Matrix.* The next step of the procedure is composition of the  $\mathbf{R}$  matrix containing sums of the  $\sum_k (1/r_{\text{RC}-m_k}^2)$  terms, related to certain types of atoms. When there are  $K$  atomic types presented in molecules of the reaction series, the  $[M \times K]$   $\mathbf{R}$  matrix is formed by the developed RMA routine. For each structural file the program sorts the atoms according to specified atomic types and calculates the sums  $\sum_k (1/r_{\text{RC}-m_k}^2)$ , where  $r$  is the direct distance between atoms of  $m$ -type in molecule  $j$  and the atom reaction center and  $k$  is the number of atoms of type  $m$  in the molecule  $j$ :

$$R = \begin{bmatrix} \left( \sum_k \frac{1}{r_{\text{RC}-m_k}^2} \right)_{1,1} & \left( \sum_k \frac{1}{r_{\text{RC}-m_k}^2} \right)_{1,2} & \dots & \left( \sum_k \frac{1}{r_{\text{RC}-m_k}^2} \right)_{1,K} \\ \left( \sum_k \frac{1}{r_{\text{RC}-m_k}^2} \right)_{j,1} & \left( \sum_k \frac{1}{r_{\text{RC}-m_k}^2} \right)_{j,2} & \dots & \left( \sum_k \frac{1}{r_{\text{RC}-m_k}^2} \right)_{j,K} \\ \left( \sum_k \frac{1}{r_{\text{RC}-m_k}^2} \right)_{M,1} & \left( \sum_k \frac{1}{r_{\text{RC}-m_k}^2} \right)_{M,2} & \dots & \left( \sum_k \frac{1}{r_{\text{RC}-m_k}^2} \right)_{M,K} \end{bmatrix}$$

In the absence of atom(s) of  $m$ -type in the molecule  $n$ , the corresponding matrix element is set equal to 0.

3. *PLS Analysis.* The final step in this procedure is estimation of whether the  $\Delta G$  dataset can be treated as a set of dependent parameters of multilparameter regression with an intercept equal to  $\Delta G^\circ$ . When the experimental parameters of free energy changes are taken as the vector  $\Delta G$

$$\Delta G = \begin{bmatrix} \Delta G_1 \\ \Delta G_2 \\ \dots \\ \Delta G_M \end{bmatrix}$$

eq 10 can be written in matrix notation as the following:

$$R\mathbf{g} = \Delta G$$

where  $\mathbf{g}$  is the solution vector

$$\begin{bmatrix} g_1 \\ g_2 \\ \dots \\ g_K \end{bmatrix}$$

containing  $K$  values of operational atomic parameters  $g_i$ , corresponding to all atomic types specified.

When  $M > K$  (i.e. the number of molecules in a reaction series is greater than the number of atomic types presented), the system is consistent and  $R\mathbf{g} = \Delta G$  can be solved.

An approximate solution of eq 10 can be achieved by multi-parameter regression (MLR), when the columns of the  $\mathbf{R}$  matrix are considered as sets of independent variables and set  $\Delta G$  values are considered as dependent parameters. If such a regression can be estimated with high accuracy, its linear coefficients can be taken as the operational atomic  $g_i$  parameters, corresponding to defined types of atoms. They can be readily used for simple estimations of unknown  $\Delta G$  parameters for similar molecular systems, composed by atoms with empirically determined  $g_i$  values.

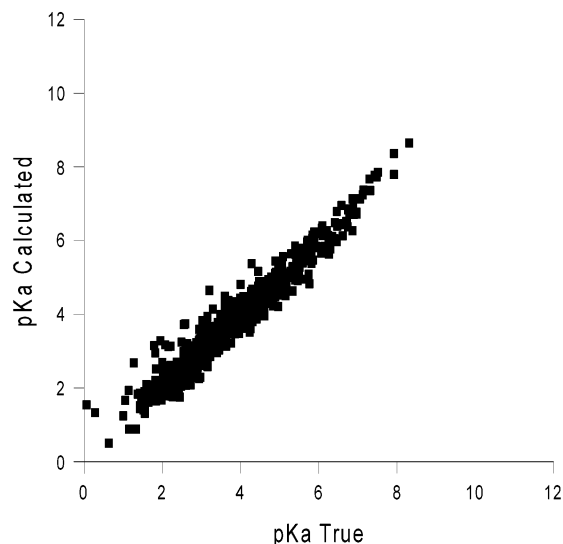
The elaborated approach should work for any reaction series which in principle can be quantitatively described by inductive and steric constants, since the  $\sigma^*$  and  $R'_s$  parameters calculated by eqs 2 and 8 are in excellent agreement with the literature data. However, the developed technique not only brings 3D formalism into correlation analysis but also may entirely transform its methodology from old-fashioned rummaging for published substituent constants to suitably fitting one's experimental data into a regression, making the method a modern and powerful iterative computational technique. In the present work, we have applied such a technique for quantification of the "classic" reaction series of dissociation of carboxylic acids and protonated aliphatic amines. By addressing the very basis of the traditional correlation analysis, we intend to demonstrate the practical application, possibilities, and advances of the new 3D methodology.

**Use of the 3D-CAN Approach for Quantification of Dissociation Constants of Molecules Containing a Carboxylic Group.** Values of ionization constants for 827 various carboxylic acids (including small polypeptides), taken from ref 50, have been extrapolated to 25 °C and zero ionic strength according to ref 51.

The structures of acid molecules have been optimized within the MM+ routine of the *Hyperchem* software package, allowing simple estimation of the standard geometries in the gas phase. After we have assumed ionizable oxygen as the reaction center, we have composed a  $[827 \times 21]$   $\mathbf{R}$  matrix for 827 compounds containing 21 types of substituent atoms. The following atomic types—H, C sp<sup>3</sup>, C sp<sup>2</sup>, C sp, C<sub>aromatic</sub>, N sp<sup>3</sup>, N sp (CN group), O sp<sup>2</sup>, O sp<sup>3</sup>, F, Cl, Br, I, S sp<sup>3</sup>, S4 (from —SO<sub>2</sub>—) Si, Se, N<sup>+</sup>, O<sup>-</sup>, and N<sup>+</sup> sp<sup>2</sup>—have been specified. The nitro group in nitro-substituted compounds has been considered as a subatomic unit, and the corresponding  $r$  parameters have been taken as the distances between the reaction center and the nitrogen of the NO<sub>2</sub>. Ionized carboxylic groups have been considered as having a full negative charge on one of oxygen atoms, while the other is in the O sp<sup>2</sup> configuration.

The procedure of composition of the  $\mathbf{R}$  matrix has been performed by a MATLAB routine, which imports atomic types and coordinates from the *Hyperchem* structural file, arranges atoms according to the types specified, and calculates intramolecular distances. After atoms/reaction centers have been indicated for all molecules of a reaction series, the routine has composed the corresponding  $\mathbf{R}$  matrix. The columns of such a  $[827 \times 21]$  matrix of the reaction series have been taken as the sets of independent variables, and the corresponding thermodynamic pK values have been considered as dependent parameters of the polynomial equation

$$\text{pK}(\text{RCOOH}) = \sum_i^{N-1} \frac{\delta_i^a}{r_i^2} + \text{constant} \quad (11)$$



**Figure 1.** Experimental vs estimated pK values of acids.

**TABLE 1. Operational Atomic Constants  $\delta_i^a$  and  $\delta_i^b$ , Estimated from pK Parameters of Carboxylic Acids and Protonated Amines, Respectively, the Corresponding Values, Predicted by Correlations 12 and 11, and Parameters of Atomic “Inductive” Electronegativities and Radii, Used in These Correlations**

	$\chi$	$R$	$\delta_i^a$	$\pm$	$\delta_i^a$	$\delta_i^b$	$\pm$	$\delta_i^b$
H	2.10	0.30	0.95	0.17	0.15	0.76	0.06	0.22
C4	2.10	0.77	0.48	0.24	0.99	0.08	0.04	1.48
C3	2.25	0.67	0.56	0.20	-0.23	-2.54	0.27	-1.05
C2	2.65	0.60	-5.07	1.25	-4.88	-8.66	0.45	-11.26
C ar	2.45	0.67	-0.45	0.11	-1.56	-2.46	0.11	-4.01
N3	2.56	0.70	-3.34	0.33	-2.45	-5.15	0.26	-6.03
N1	6.76	0.55	-18.24	2.55	-19.95	-42.00	1.34	-44.56
O2	3.05	0.66	-5.61	0.25	-5.28	-9.54	0.24	-12.22
F	3.93	0.64	-2.88	0.28	-8.32	0.46		
Cl	3.09	0.99	-12.59	0.55	-12.44	-23.77	0.31	-28.75
Br	2.96	1.14	-14.60	0.83	-14.05	-36.59	0.64	-32.70
I	2.80	1.33	-8.90	1.88	-16.52	4.67		
S2	2.69	1.04	-6.19	0.50	-7.45	-14.85	4.30	-17.82
Si	2.06	1.11	2.86	1.49	2.77	1.36	0.84	4.65
N+	4.33	0.70	-20.33	0.42	-15.04	-41.29	0.72	-33.91
O-	1.85	0.70	28.61	0.60		9.44	0.46	5.19
N2						2.05	3.47	
N2+			-16.71	2.02	-13.28	-30.67	16.33	
O1	4.60	0.62	-6.25	0.32	-13.30	-9.82	0.71	
S6			-3.64	1.36				
Se	2.54	1.17	-16.30	3.69				
nitro			-9.02	2.04				

where  $\delta_i^a$  is the introduced atomic operational parameter, reflecting the ability of atoms of one type to contribute to the pK value of an N atomic carboxylic acid RCOOH where R represents the molecular environment of the carboxylic group. A multilinear regression has then been established with high accuracy (constant =  $4.84 \pm 0.12$ ;  $N = 827$ ;  $R^2(\text{mult}) = 0.941$ ;  $S = 0.1035$ ). The interrelation between estimated and experimental pK values is presented graphically in Figure 1. The estimated results demonstrate that the suggested approach allows for accurate quantitative interpretation dissociation constants of a wide range of various carboxylic acids. The values of the estimated atomic operational contributions in eq 11 can be used for an accurate enough prediction of unknown pK values of molecules, constituted from the atom types presented in Table 1.

**Quantitative Assessment of pK Values of Amines.** It is a matter of common knowledge that the basicity of amines can be interpreted in terms of polar substituent constants. Numerous

authors have proposed different LFER equations describing a limited series of basicity data for primary, secondary, and tertiary amines.<sup>51</sup>

We have not separated experimental data into several reaction series and have considered a dataset of 802 pK values for various amines in which ionizing nitrogen was not engaged in conjugation interactions.

The structures of organic amines have been optimized within the MM+ routine of the *Hyperchem* software package, allowing simple estimation of the standard geometries in the gas phase. After we have assumed ionizable nitrogen as the reaction center, a  $[802 \times 19]$  **R** matrix for 802 compounds containing 19 types of substituent atoms has been composed. The following atomic types—H, C sp<sup>3</sup>, C sp<sup>2</sup>, C sp, C<sub>aromatic</sub>, N sp<sup>3</sup>, N sp<sup>2</sup>, N sp (CN group), O sp<sup>2</sup>, O sp<sup>3</sup>, F, Cl, Br, I, S, Si, N<sup>+</sup>, O<sup>-</sup>, and N<sup>+</sup> sp<sup>2</sup>—have been specified. Ionized carboxylic groups have been considered as having a full negative charge on one of the oxygens, while the other is in the O sp<sup>2</sup> configuration.

The columns of the  $[802 \times 19]$  **R** matrix have been taken as the sets of independent variables. The values of pK taken from ref 45 have been extrapolated to 25 °C and zero ionic strength according to ref 44. When experimental details were insufficient, the corresponding pK values were accepted as given (which in some cases might lead to uncertainties up to 0.1 pK units). Then, the corrected pK<sub>a</sub> parameters were considered as dependent parameters of the polynomial equation

$$pK(R_3N) = \sum_i^{N-1} \frac{\delta_i^b}{r_i^2} + \text{constant} \quad (12)$$

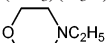
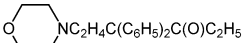
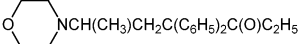
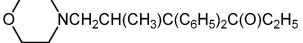

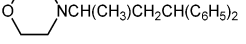

where  $N$  is the number of atoms in the amine and  $\delta_i^b$  is the introduced atomic operational parameter reflecting the ability of atoms of one type to contribute to the amine's pK<sub>a</sub>. A multilinear regression (eq 12) has been established with high accuracy (constant =  $9.12 \pm 0.19$ ;  $N = 802$ ;  $R^2(\text{mult}) = 0.933$ ;  $S = 0.1819$ ) which allows the usage of the estimated operational atomic parameters for amine basicity predictions:

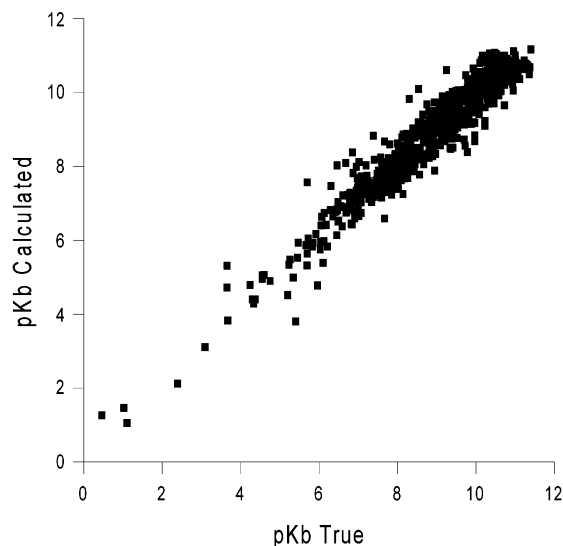
$$pK(R_3N) = 9.12 + \sum_i^{N-1} \frac{\delta_i^b}{r_i^2} \quad (13)$$

The estimated pK<sub>a</sub> values of the amines are presented in Table 2 along with the corresponding experimental data. The operational atomic parameters  $\delta_i^b$  for the 19 atomic types used, taken as the multiple coefficients of eq 13, are collected in Table 1. Interrelation between estimated and experimental pK values is presented graphically in Figure 2.

Thus, the suggested approach allows for accurate quantitative interpretation of basicity data of a wide range of primary, secondary, and tertiary amines. The values of the estimated atomic operational contributions in eq 13 can hence be used for prediction of unknown pK values for amines, constituted from the atom types presented in Table 2. The large uncertainties in the operational parameters  $\delta_i^b$  estimated for O sp<sup>2</sup>, F, and I are due to the lack of data (column elements of the **R** matrix) for these atoms, which leads to significant statistical deviations. A reviewer has noted that this protocol predicts dissociation well but fails to accurately predict operational parameters as accurately. The specific cases where the fit is poor are due to lack of data. That these are poorly fitted is unsurprising, as is the fact that they have little impact on the overall fit of the dissociations, since they are not well represented in the population of the training set.

TABLE 2. Specific Outliers from the Basic Amine Dataset

tag	formula	$pK_b(\text{true})$	$pK_b(\text{calc})$	residual
18	$\text{Cl}_3\text{CC}_2\text{H}_4\text{NH}_2$	5.40	3.80	1.60
19	$\text{F}_3\text{CC}_2\text{H}_4\text{NH}_2$	5.70	7.57	-1.87
48	$\text{Cl}_3\text{CC}_3\text{H}_6\text{NH}_2$	9.78	8.39	1.39
133	$\text{F}_3\text{C}_6\text{CH}_2\text{NH}_2$	7.67	6.59	1.08
622	$\text{H}_5\text{C}_2\text{OC}(\text{O})\text{C}(\text{C}_6\text{H}_5)_2\text{C}_2\text{H}_4(\text{H}_3\text{C})_2\text{N}$	9.72	8.56	1.15
623	$\text{H}_5\text{C}_2\text{OC}(\text{O})\text{C}(\text{C}_6\text{H}_5)_2\text{CH}_2\text{CH}(\text{CH}_3)(\text{H}_3\text{C})_2\text{N}$	9.97	8.69	1.27
703		7.67	8.67	-1.00
704		6.95	7.99	-1.04
705		6.68	8.10	-1.42
706		7.02	8.12	-1.10
707		7.38	8.83	-1.45
708		6.85	8.38	-1.53
709		10.95	10.06	0.89
125	$\text{C}_6\text{H}_3(2\text{-OCH}_3, 3\text{-OCH}_3)\text{CH}_2\text{NH}_2$	9.41	8.51	0.90
126	$\text{C}_6\text{H}_3(3\text{-OCH}_3, 4\text{-OCH}_3)\text{CH}_2\text{NH}_2$	9.39	8.75	0.64
127	$\text{C}_6\text{H}_4(2\text{-OCH}_3)\text{CH}_2\text{NH}_2$	9.70	8.73	0.97

Figure 2. Experimental vs estimated  $pK$  values of amines.

**Deviations from the General Trend.** It should be stressed that the accuracy of most of the experimental points used in the analysis is not very high, which may significantly contribute to prediction errors. On the other hand, together with numerous merits of the elaborated 3D approach, it possesses some drawbacks, which may lead to deviation of certain points from the general prediction trend.

Thus, the  $pK$  values of amines 18, 19, 48, and 133 from Table 2 and carboxylic acids 241, 267, and 269 from Table 3, containing halogen atoms, have been established with lower accuracy due to the presence of a nonadditive saturation effect. Rather strong resonance interactions also cannot be taken into account comprehensively by the method, which probably caused deviations for amines 622 and 623, containing several phenyl fragments (Table 2). The predicted  $pK$  values of cyclic compounds 703–709 from Table 2 generally disagree with the corresponding experimental values, perhaps due to additional interaction of oxygen and nitrogen atoms in the cyclic structure;

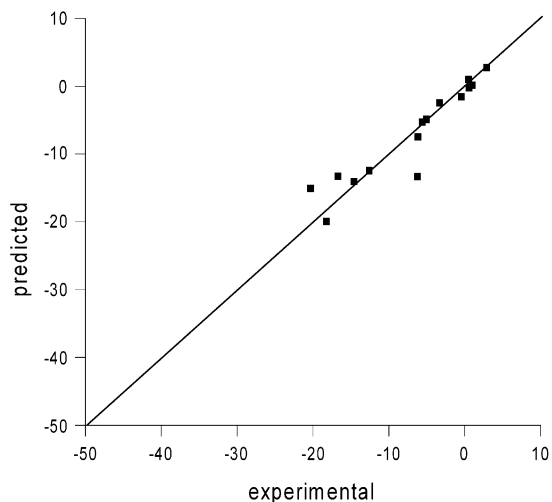
TABLE 3. Specific Outliers from the Carboxylic Acid Dataset

tag	formula	$pK_a(\text{true})$	$pK_a(\text{calc})$	residual
241	$\text{FCH}_2\text{COOH}$	2.59	3.74	-1.15
267	$\text{H}_2\text{C}=\text{FCOOH}$	2.55	3.72	-1.16
269	$\text{F}_2\text{C}=\text{FCOOH}$	1.79	3.15	-1.36
21	$\text{HOCCOOH}$	1.27	2.68	-1.41
49	$\text{HOCC}(\text{C}_2\text{H}_5)_2\text{COOH}$	2.21	3.13	-0.92
51	$\text{HOCC}(\text{C}_2\text{H}_5)(\text{C}_3\text{H}_7)\text{COOH}$	2.15	3.12	-0.98
53	$\text{HOCC}(\text{C}_3\text{H}_7)_2\text{COOH}$	2.07	3.17	-1.10
128	$\text{C}_3\text{H}_4(\text{cyclo})\text{-1-COOH, 1-COOH}$	1.82	2.95	-1.13

however, they may probably be corrected if we specify new atomic types for these atoms. The presence of a strong ortho effect in amino derivatives 125 and 127 also caused the deviation of the corresponding calculated  $pK$  parameters. Rather significant (about 1  $pK$  unit) overestimation of the dissociation constants of carboxylic acids 21, 49, 51, 53, and 128 compared to the corresponding experimental values may be attributed to the presence of intramolecular hydrogen bonding. Therefore, the mentioned deviations illustrate that the developed approach correctly describes merely inductive (and possibly steric) interactions and any extra effects can be readily identified.

On the other hand, the accuracy of the described procedure depends on the conformations of the molecules of the series, and therefore, more detailed geometry optimization is required to achieve higher accuracy of the correlation procedure. At the same time, if the failure of some points from dependence is caused by the geometry of the corresponding compounds but regression (eq 10) is well established on the basis of extensive experimental data, then this general experimental trend may guide further geometry optimization of deviating members. Similarly, a "solid" regression (eq 10) may help to establish a correct order of ionization of competing groups in polyfunctional systems, which perhaps was not always the case for the present study and has caused some random deviations.

**Physical Meaning of the Operational Atomic Contributions.** When exploring the  $\Delta G$  dependent characteristics of molecules of a reaction series by eq 10, the physical meaning of the corresponding atomic  $g$  parameters remains unclear and



**Figure 3.** Experimental vs estimated  $\delta_i^a$ —operational atomic parameters.

the overall substituent effects cannot be divided into particular electronic and steric components. Thus, once the atomic  $g$  parameters are established, further interpretation of their physical meaning is the actual goal. Simple treatment of operational atomic parameters can be performed on the basis of the previously established eqs 3 and 8 for the inductive ( $\sigma^*$ ) and steric ( $R_s$ ) constants, linear combination of which leads to the general eq 1. Therefore, the operational atomic parameter  $g_i$  estimated on the basis of eq 1 can be considered as the following:

$$g_i = a'\Delta\chi_i R_i^2 + b'R_i^2 \quad (14)$$

where the coefficient  $b'$  contains the electronegativity of the reaction center.

Thus, the correct separation of the inductive and steric contributions to substituent effects in terms of the elaborated technique is problematic without the knowledge of the nature (electronegativity) of the reaction center. Once  $\chi_{RC}$  is assumed, the eq 15 may be divided into increments reflecting the contributions of inductive and steric effects:

$$g_i = a\Delta\chi_{i-RC}R_i^2 + bR_i^2 \quad (15)$$

Consequently, eq 15 can be divided into the inductive and steric components, presented by eqs 3 and 8, respectively.

According to the described procedure, we have analyzed the physical meaning of the operational atomic parameters  $\delta_i^a$  and  $\delta_i^b$ , which have been correlated with the corresponding  $\chi R^2$  and  $R^2$  magnitudes as the following:

$$\delta_i^a = (-14.51 \pm 1.33)\chi_i R_i^2 + (32.14 \pm 3.89)R_i^2 \quad (16)$$

$$R = 0.9395; S^\circ = 2.7478; N = 15$$

$$\delta_i^b = (-32.14 \pm 1.97)\chi_i R_i^2 + (69.99 \pm 5.53)R_i^2 \quad (17)$$

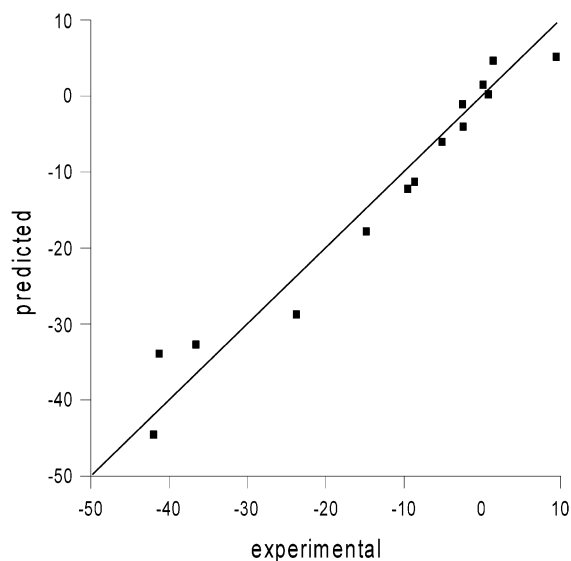
$$R = 0.9781; S^\circ = 3.6558; N = 14$$

Correlations 16 and 17 can thus be written as

$$\delta_i^a = -14.51(\chi_i - 2.21)R_i^2 \quad (18)$$

$$\delta_i^b = -32.14(\chi_i - 2.17)R_i^2 \quad (19)$$

Correlation 18 is presented graphically in Figure 3, and correlation 19 is presented in Figure 4.



**Figure 4.** Experimental vs estimated  $\delta_i^b$ —operational atomic parameters.

The superposition of eqs 3 and 18 allows us to present the  $pK$  value of carboxylic acid  $RCOOH$  (when the ionizing oxygen of the carboxylic group is considered as a reaction center) as the following:

$$pK(RCOOH) = 4.84 - 14.51 \sum_i^{N-1} \frac{(\chi_i - 2.21)R_i^2}{r_i^2} \quad (20)$$

Similar combination of eqs 3 and 19 gives the equation for the amine's aqueous basicity, explored at atomic consideration level, when ionizing nitrogen is considered as the reaction center and all other atoms are considered as the whole sub-substituent:

$$pK(R_3N) = 9.12 - 32.14 \sum_i^{N-1} \frac{(\chi_i - 2.17)R_i^2}{r_i^2} \quad (21)$$

In the above equations, the intercepts reflect a baseline  $pK_a$  or  $pK_b$  for an unsubstituted acid or base. The intercepts do, however, reflect the nature of the training set. Hence, the 4.84 intercept suggests that the unsubstituted acid is acetic acid rather than formic acid. This reflects the aliphatic nature of the dataset.

It is a remarkable fact that the statistically established parameters of "inductive" electronegativity of the reaction centers in eqs 20 and 21 are virtually the same (2.21 and 2.17, respectively). Generally speaking, this means that all considered substituents demonstrate the same type of electron sharing (same sign of the  $\sigma^*$  inductive constant) in both reaction series.

Bearing in mind that oxygen and nitrogen have been originally assumed as the corresponding reaction centers, the estimated eqs 20 and 21 made us wonder if ionizing hydrogen ( $\chi_H = 2.10$ ) should have been taken as the reaction center for the reaction series of the dissociation constants. However, when we considered the hydrogen to be the reaction center and conducted the corresponding 3D correlation analysis procedures for the  $pK$  datasets under investigation, we were not able to establish the corresponding correlations (eq 9) with reasonable accuracy. We can suggest two possible explanations for this observed phenomenon. First of all, we could misidentify the parameter 2.10 with the "inductive" electronegativity of  $sp^3$  carbon, considered as the reaction center (this assumption justified zero inductive effect of alkyls and sign of  $\sigma^*$  constants



for other substituents). Probably, the nature of the corresponding constants in eqs 20 and 21 is different and needs to be studied in more detail. Or, as the second explanation (and, probably, the most reasonable one), we could admit the significance of the steric effect on ionization properties of amines and acids, when ionizing atoms are considered as reaction centers and no insulating fragments are used. In this case, we cannot correctly separate the estimated operational parameters  $\delta_i^a$  and  $\delta_i^b$  into inductive and steric components and should use them as they are for prediction of pK parameters on the basis of eqs 12 and 13. Then, it should be fair to assume that Taft's inductive constants are also not free from a partial steric contribution.

Nevertheless, if we neglect the difference in the estimated constants 2.21 and 2.17 in eqs 20 and 21, and the value 2.10 used in eq 3 for Taft's inductive parameters, then we can express the pK values of carboxylic acids on the basis of Taft's  $\sigma^*$  constant

$$\text{pK}(\text{RCOOH}) = 4.84 - 1.85\sigma^* \quad (22)$$

(where  $\sigma^*$  is Taft's inductive constant of the molecular environment of ionizable oxygen in the molecule) and the pK values of organic amines as

$$\text{pK}(\text{R}_3\text{N}) = 9.12 - 4.1\sigma^* \quad (23)$$

(where  $\sigma^*$  is Taft's inductive constant of the molecular environment of ionizable nitrogen).

It is not surprising that eq 23 agrees with the previously established LFER correlation  $\text{pK}(\text{RCOOH}) = 4.66 - 1.62\sigma^{*1,53}$  and eq 23 is in reasonable agreement with previously established LFER correlations for primary [ $\text{pK}(\text{RNH}_2) = 10.15 - 31.4\sigma^*$ ], secondary [ $\text{pK}(\text{R}'\text{R}''\text{NH}) = 10.59 - 3.23\sigma^*$ ], and tertiary [ $\text{pK}(\text{R}'\text{R}''\text{R}'''\text{N}) = 9.61 - 3.30\sigma^*$ ] amines.<sup>51</sup> At the same time, it should be stressed that the way eqs 22 and 23 have been estimated is completely different from the procedures of classic correlation analysis.

On the basis of eq 23 and the estimated  $\text{pK}_a$  values, we have calculated Taft's inductive constants  $\sigma^*$ , corresponding to nitrogen's molecular environment  $R_n$  in the studied amines  $R_n\text{NH}_{3-n}$  ( $3 - n$  hydrogens of the ionizing amino group have not been considered). The corresponding overall  $\sigma^*$  parameters are presented in Table 3. The calculated values  $\sigma^*$  clearly demonstrate that the developed 3D approach allows the quantification of the inductive effect and estimation of inductive constants even for rather complex molecular systems which normally could not be considered by conventional approaches of the correlation analysis.

## Conclusions

In this paper we have demonstrated that the application of 3D correlation analysis to extensive aqueous acidity and basicity data leads us to new formulas which can be used in direct calculations of  $\text{pK}_a$  values of carboxylic acids and protonated amines of any complexity and in molecular modeling of compounds with desired acidic and basic functions. The physical meaning of the estimated operational atomic constants  $\delta_i^b$  and  $\delta_i^a$  has been identified and allows us to express aqueous acidity/basicity in terms of atomic electronegativities, covalent radii, and interatomic distances. It should be stressed, especially, that the obtained results not only demonstrate the practical usefulness of the elaborated 3D technique but also highlight its numerous substantial advantages:

1. There are no limitations in the choice of appropriate substituent scales, since none are required—scaled inductive constants can be readily calculated for any substituent.

2. The developed approximation “reaction center – the rest of the molecule”, when the skeleton and indicative group are all included in the substituent, resolves the problem of choice of the standard of reaction series (which is often a problem for classic correlation analysis).

3. The developed approach takes into account actual 3D molecular structures and allows calculation of substituent effects of conformers.

4. The approach possesses significant versatility, allowing consideration of different atomic types related to atom's type, valent state, ionization, and molecular environment.

5. It is an important feature of the developed technique that eqs 13 and 14 make it possible to carry on the search over several potential reaction centers of the series.

Thus, the new powerful QSAR technique called “3D correlation analysis”, which allows quantification of the substituent effect without use of pre-established substituent constants and possesses numerous advantages, has been elaborated.

Together with previously developed “inductive” reactivity indices, 3D correlation analysis may become a very effective molecular modeling and data mining tool for chem- and bioinformatics. The developed technique allows avoiding any limitations related to molecular size and makes it possible to readily treat very massive sets of experimental data, which makes it especially important for bioactivity-related studies.

The broader practical application and development of the methodology of 3D correlation analysis is underway, and its possibilities are being explored for structure–activity studies of a wide range of proteins.

**Acknowledgment.** This work has been supported by Medical Research Council of Canada Grant MT-14306 to R.C.

## References and Notes

- (1) Kirkwood, J. G.; Westheimer, F. H. *J. Chem. Phys.* **1938**, *6*, 506.
- (2) Johnson, C. D. *The Hammett Equation*; Cambridge University Press: Cambridge, 1973.
- (3) Taft, R. W. *J. Am. Chem. Soc.* **1953**, *75*, 4538.
- (4) Clark, D. E.; Pickett, S. D. *Drug Discuss. Today* **2000**, *5*, 49.
- (5) Yoshida, F.; Topliss, J. G. *J. Med. Chem.* **2000**, *43*, 2575.
- (6) Winiwater, S.; Bonham, N. M.; Ax, F.; Hallberg, F.; Halberg, A.; Lennernas, H.; Karlen, A. *J. Med. Chem.* **1998**, *41*, 4939.
- (7) da Silva, C. O.; da Silva, E. C.; Nascimento, M. A. C. *J. Phys. Chem.* **1999**, *103*, 11194.
- (8) Citra, M. J. *Chemosphere* **1999**, *38*, 191.
- (9) Hillal, S. H.; Carreira, L. A.; Baughman, G. L.; Karickhoff, S. W.; Melton, C. M. *J. Phys. Org. Chem.* **1994**, *7*, 122.
- (10) Tsantili-Kakoulidou, A.; Panderi, I.; Cszimadia, F.; Darvas, F. *J. Pharm. Sci.* **1997**, *86*, 1173.
- (11) Li, X.; Glen, R. C. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 796.
- (12) Cherkasov, A.; Jonsson, M.; Galkin, V. I. *J. Mol. Graphics Modell.* **1999**, *17*, 28.
- (13) Pal'm, V. A. *Osnovy Kolichestvennoi Teorii Organicheskikh Reaktsii* (Fundamentals of the Quantitative Theory of Organic Reactions); Khimiya: Leningrad, 1977.
- (14) Exner, O. In *Advances in Linear Free Energy Relationships*; Chapman, N. B., Shorter, J., Eds.; Plenum Press: London, New York, 1972, pp 1–71.
- (15) Hansch, C.; Leo, A. *Substituents Constants for Correlation Analysis in Chemistry and Biology*; Wiley–Interscience: New York, 1979.
- (16) Vereshchagin, A. N. *Konstanty Zamestitelei dlya Korrelyatsionnogo Analiza* (Substituent Constants for Correlation Analysis); Nauka: Moscow, 1988.
- (17) Charton, M. *Adv. Quant. Struct.–Prop. Relat.* **1996**, *1*, 171.
- (18) Cherkasov, A. R.; Galkin, V. I.; Cherkasov, R. A. *Russ. Chem. Rev.* **1996**, *65*, 641.
- (19) Wells, P. R. *Linear Free Energy Relationships*; Academic Press: London, 1968.
- (20) Jaffe, H. H. *Chem. Rev.* **1953**, *53*, 191.

- (21) Ingold, C. K. *Structure and Mechanism in Organic Chemistry*; Cornell University Press: Ithaca, London, 1969.
- (22) Vereshchagin, A. N. *Induktivnyi Effekt (Inductive Effect)*; Nauka: Moscow, 1987.
- (23) Hansch, C.; Leo, A.; Taft, R. W. *Chem. Rev.* **1991**, *91*, 165.
- (24) Exner, O. *J. Phys. Org. Chem.* **1999**, *12*, 265.
- (25) Charton, M. *J. Phys. Org. Chem.* **1999**, *12*, 275.
- (26) Yang, A. S.; Honig, B. *J. Mol. Biol.* **1993**, *231*, 459.
- (27) Gilson, M. K.; Honig, B. H. *Nature* **1987**, *330*, 84.
- (28) Gilson, M. K.; Honig, B. H. *Proteins: Struct., Funct., Genet.* **1988**, *3*, 32.
- (29) Bashford, D.; Karplus, M. *Biochemistry* **1990**, *29*, 10219.
- (30) Berioza, P.; Fredkin, D. R.; Okamura, M. Y.; Feher, G. *Proc. Natl. Acad. Sci. U.S.A.* **1991**, *88*, 5804.
- (31) Bashford, D.; Gerwert, K. *J. Mol. Biol.* **1992**, *224*, 473.
- (32) Yang, A. S.; Gunner, M. R.; Sampogna, R.; Sharp, K.; Honig, B. *Proteins: Struct., Funct., Genet.* **1993**, *15*, 252.
- (33) Oberoi, H.; Allewell, N. M. *Biophys. J.* **1993**, *65*, 48.
- (34) Potter, M. J.; Gilson, M. G.; McCammon, J. A. *J. Am. Chem. Soc.* **1994**, *116*, 10298.
- (35) Rajasekaran, E.; Jayaram, B.; Honig, B. *J. Am. Chem. Soc.* **1994**, *116*, 8238.
- (36) Marriotti, S.; Reynolds, W. F.; Taft, R. W.; Topsom, R. *J. Org. Chem.* **1984**, *49*, 959.
- (37) Reynolds, W. F.; Mezey, P. G.; Hamer, P. G. *Can. J. Chem.* **1977**, *55*, 522.
- (38) Exner, O.; Ingr, M.; Carsky, P. *THEOCHEM* **1997**, *397*, 231.
- (39) Cherkasov, A. R.; Galkin, V. I.; Sibgatullin, I. M.; Cherkasov, R. A. *Phosphorus, Silicon, Sulp.* **1996**, *111*, 141.
- (40) Cherkasov, A. R.; Galkin, V. I.; Cherkasov, R. A. *J. Phys. Org. Chem.* **1998**, *11*, 437.
- (41) Cherkasov, A. R.; Galkin, V. I.; Zueva, E. M.; Cherkasov, R. A. *Russ. Chem. Rev.* **1998**, *67*, 375.
- (42) Cherkasov, A. R.; Galkin, V. I.; Cherkasov, R. A. *THEOCHEM* **1999**, *489*, 43.
- (43) Cherkasov, A. R.; Galkin, V. I.; Cherkasov, R. A. *THEOCHEM* **2000**, *497*, 115.
- (44) Charton, M.; Charton, B. *J. Chem. Soc., Perkin Trans. 2* **1999**, *13*, 2203.
- (45) Kovetz, A. *Electromagnetic Theory*; Oxford University Press: New York, 2000.
- (46) Cherkasov, A.; Jonsson, M. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 1151.
- (47) Cherkasov, A.; Jonsson, M. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1057.
- (48) Cherkasov, A.; Jonsson, M. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1222.
- (49) Galkin, V. I.; Sayakhov, R. D.; Cherkasov, R. A. *Russ. Chem. Rev.* **1991**, *60*, 815.
- (50) Kortum, G.; Vogel, W.; Andrussov, K. *Dissociation Constants of Organic Acids in Aqueous Solution*; Butterworth: London, 1961.
- (51) Perrin, D. D.; Dempsey, B.; Serjeant, E. P. *pK<sub>a</sub> Prediction for Organic Acids and Bases*; Chapman & Hall: London, New York, 1981.
- (52) Perrin, D. D. *Dissociation Constants of Organic Bases in Aqueous Solution*; Butterworth: London, 1965.
- (53) Hall, H. K. *J. Am. Chem. Soc.* **1957**, *79*, 5441.