Reactivity Prediction Models Applied to the Selection of Novel Candidate Building Blocks for High-Throughput Organic Synthesis of Combinatorial Libraries

Mircea Braban, Iuliana Pop, Xavier Willard, and Dragos Horvath*

CEREP, 1 rue Calmette, 59019 Lille, France

Received August 15, 1999

Quantitative structure−property relationships (QSPRs) expressing the reactivity of compounds on the basis of molecular descriptors have been developed and applied to the computer-aided selection of synthons of appropriate reactivity for the high-throughput synthesis of combinatorial libraries. Our approach explicitly models the influence of substituents on the activity of the reactive center (RC), introducing specific molecular descriptors for their electronic, steric, and field effects (including the solvent effects) as a function of the 2D and 3D substituent−RC distances. Therefore, the approach requires a much smaller number of empirical "substituent constants" than the classical Hammett approach. These constants only depend on the chemical nature of the substituents and not on their relative position with respect to the RC. A general pK_a prediction model was obtained by calibrating the weighting factors that express the relative influences of the electronic and field effect descriptors on the acidity of functional groups, using a learning set of about 500 organic amines and acids. A QSPR model expressing the degrees of conversion of a reference amine in the amide synthesis reaction, in terms of the descriptors of the carboxylic acids, was then derived. The used learning set included 100 out of the 150 acids for which the conversions were experimentally determined at the first stage of a typical selection process of building blocks for combinatorial synthesis. The predicted percentages of conversion of the acids not included in the learning set showed (absolute) errors not exceeding ±20%. As a consequence, this model is a useful computational tool in discriminating between reactive and inappropriate compounds from molecular databases, retrieving the building blocks that are most likely to comply with the reactivity criteria.

I. INTRODUCTION

Automation of biological activity tests opened the way to the high-throughput screening (HTS) of large sets of molecules as a fast method of evaluating their biological activities, to discover new lead compounds and potential drugs. Under the pressure of important demands for diverse chemical libraries, high-throughput organic synthesis,[1−4] (HTOS) techniques, using robots to synthesize all the possible products A_i−B_j that result by combining each building block A_i of set A with every partner B_j of set B, have been developed. Not every chemical process can be, however, used in HTOS. An important bottleneck may be the low availability of building blocks that yield good conversions in the considered process.

Selecting the reactive, appropriate, building blocks for HTOS is an important and difficult step. A chemist becomes rapidly overwhelmed by the great number of commercially available compounds from electronic databases. Therefore, computer-aided approaches are required to perform this task. Several building block selection strategies aimed to either maximize the structural diversity[5−7] of the resulting combinatorial library or enrich the library in compounds that display the desired biological activity[8,9] are widely used nowadays. However, the equally important aspect of computer-based reactivity predictions of the potential building blocks has been less developed. A fast selection algorithm of reactive building blocks cannot be based on time-consuming quantum-mechanical models[10] of chemical reactivity, but must rely on empirical approaches such as quantitative structure−property relationships (QSPRs)[11] or neural networks (NN).[12,13]

QSPRs have been used to predict various molecular properties.[14−17] The first quantitative structure−reactivity model was proposed by Hammett,[18] who introduced eq 1

$$\log (k/k_0) = \sigma\rho \qquad (1)$$

relating the rate constant (or equilibrium constant) of a given reaction involving *meta*- or *para*-substituted aromatic compounds to the chemical nature of these substituents. The applicability of Hammett's equation is limited to aromatic compounds. The ρ values (*the reaction coefficients*) are known to depend on the experimental conditions,[19,20] while the σ values (*substituent coefficients*) differ as a function of the *meta* or *para* position of the substituent and are not always additive.[21−23] Steric effects are not accounted for explicitly, so the equation cannot be applied to *ortho*-substituted compounds.

Other authors proposed more elaborate and accurate "Hammett-like" equations[24−27] for different classes of organic compounds. In these approaches, a substituent is characterized by an empirical constant that is determined by calibration with respect to experimental data. Their main drawback is that they do not include any explicit description of the

* To whom correspondence should be addressed. Phone: +33.3.20.16.91.48. Fax: +33.3.20.58.07.36. E-mail: d.horvath@cerep.fr or Dragos.Horvath@pasteur-lille.fr.

10.1021/ci990104x CCC: $18.00
Published on Web 10/21/1999

variation of the effect of a substituent with its position with respect to the RC, requiring the recalibration of the "substituent constants" for every particular structural context. Therefore, the applicability of such a model will always be restricted to a single family of compounds sharing a common structural skeleton on which the substituents are attached in predefined positions. To our knowledge, no structure−reactivity models of general validity have been reported.

The analysis of the molecular topology and geometry by means of computational chemistry tools may explain a property in terms of fewer, more basic structural features[28] than the presence/absence of certain substituents at certain positions of a reference molecular scaffold. Reaction databases such as ChemInform RX[29] and reaction retrieval software[29] are nowadays widely used by chemists, while the computer-assisted retrosynthetic analysis of products or prediction of possible processes involving given synthons has also been developed recently[30,31] (also see references within ref 32). A modern and interesting approach is the analysis of the reactivity in terms of molecular topology, using neural networks.[32]

We introduce here a *general* relationship between chemical reactivity and molecular structure, which can be applied for a fast and reasonably accurate prediction of reactivity. The approach successfully classifies compounds with respect to a relative scale of reactivity, predicting whether a molecule will be a good, average, or poor reactant under the given conditions. Such a degree of accuracy is sufficient to turn the model into a useful tool for the preselection process of appropriate building blocks for HTOS.

The model is based on linear relations expressing the measure of chemical reactivity in terms of a set of geometric and topological descriptors representing *electronic, steric, and field effects* derived on the basis of multiconformational molecular models. It introduces *explicit laws of variation of the intensity of each of these effects in terms of the relative position of the substituents with respect to the reactive center (RC)*. Empirical equations have been assumed to provide an approximate description of these variations if no appropriate simple functional forms could be derived from fundamental physical principles.

In the first part of this work, we have developed a prediction model of the pK_a values of ionizable groups of the organic compounds. We have used the large set of available acidity constants from the literature[33,34] to select the relevant intensities of the electronic and field effects with respect to the experimental pK_a shifts, using a genetic algorithm[35,36]-based multilinear regression procedure. The subsequently developed reactivity prediction model uses both the electronic and field descriptors (the pK_a shifts calculated on the basis of the previously obtained models *were added as a new, synthetic, electronic descriptor*) and a set of steric descriptors. It was calibrated with respect to the experimental conversions of the amide formation reactions of a set of 100 aromatic and aliphatic carboxylic acids with a common amine, in the presence of *N,N′*-carbonyldiimidazole.

II. METHODS

In Hammett's equation (1), σ is a global parameter which simultaneously accounts for *all* the effects of a substituent of an aromatic ring on the reactivity. This parameter is misleadingly called a substituent constant, because it implicitly accounts for both the chemical nature *and* the relative position of the substituent with respect to the RC. If the same substituent is located elsewhere than in the *meta* or *para* position of an aromatic ring, its substituent constant will have to be refitted against new experimental reactivity data, since *Hammett's approach does not provide any information on how the effects caused by a substituent will depend on the topological and geometric distance to the reactive center*. In contrast to Hammett's approach, the following apply in our method.

(1) Electronic, field, and steric effects are modeled explicitly; e.g., a specific functional form is assigned to each one of the inductive, resonance, electrostatic, solvent, intramolecular hydrogen bond and steric effects. The intensity of each effect E is expressed as a product of a *context function*, e.g., a *descriptor* which expresses the variation of the effect with the position of a substituent with respect to the RC, and a *weighting factor* ($\sigma^E\rho^E$). The latter can be understood as the product of the intrinsic influence (σ^E) that the considered effect of that substituent exercises on the stereoelectronic features of the RC times the intrinsic sensitivity of the reactivity of the compound with respect to a variation of these stereoelectronic features (ρ^E). The σ^E coefficients will be referred to as the *effect* coefficients.

(2) The descriptor values are evaluated on the basis of multiconformational 3D models, generated for each of the studied molecules, according to a conformational sampling procedure outlined elsewhere.[37] Ionizable groups other than the RC itself have been represented in states of protonation according to the pH value at which the proteolytic reaction of the RC is expected to occur. For example, ethylenediamine was modeled as $H_2N^*-CH_2-CH_2-NH_3^+$ (* labels the reactive center) to properly explain the pK_a of the second ionization step, which corresponds to a proton uptake in the presence of a positive charge in the neighborhood of the RC.

A. Acidity Prediction Model. A set of 474 aromatic and aliphatic acids and amines with pK_a values available from the literature[33,34] were used for the calibration of the acidity prediction model. The pK_a shifts of the amines due to the influence of the rest of the molecule on the ionizable group were taken with respect to methylamine, while in the case of acids, the internal reference was the pK_a value of acetic acid.

$$\Delta pK_a = pK_a^{\text{amine}} - pK_a^{\text{methylamine}} \quad (2)$$

$$\Delta pK_a = pK_a^{\text{acid}} - pK_a^{\text{acetic acid}} \quad (3)$$

Since H^+ is the smallest existing electrophilic species, we adopted the working hypothesis that the steric effects have a negligible influence on the acidity of the compounds. The ΔpK_a values from eqs 2 and 3 were broken down into contributions from electronic and molecular field descriptors: inductive effect (I); resonance effect (R); electrostatic field effect (C); solvent effect (S); effect of hydrogen bonds (H).

$$\Delta pK_a = \Delta pK_a^I + \Delta pK_a^R + \Delta pK_a^C + \Delta pK_a^S + \Delta pK_a^H \quad (4)$$

The intrinsic sensitivities ρ of the proton exchange reaction with respect to the variations of the properties of the RC

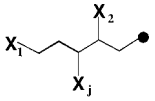

Figure 1. Inductive effects are considered as additive, topological distance-dependent contributions of the functional groups X_j located at each node of the molecular graph.

Table 1. Classes of Substituents Considered To Give Major Inductive Effects[a]

X^I	substituents	X^I	substituents
1	aliphatic C	6	Br, I, P, S
2	unsaturated C	7	carbonyl
3	any* N	8	C=O in esters/amides; CN
4	any* O	9	NO_2, SO_2, SO_3
5	F and Cl		

[a] *, except atoms encountered in the functional groups 7−9 explicitly listed. O−C=O in an ester is of type 4, while the group O−C=O is of type 8.

were by definition set to 1, so that the weighting factors of the different contributions of the effects discussed in the next paragraphs can be assimilated to the effect coefficients σ^E.

1. Definitions of the Electronic and Field Descriptors. For the effects which follow a well-defined physical law (electrostatic/reaction field effect) the mathematical expression describing this law is used as a "context function". Otherwise, empirical relationships were defined and tested, the final quality of the predictive models being the objective criterion of the validity of the postulated context functions.

I. Inductive Effect. The inductive effect of a substituent is considered to depend on its topological distance d_{top} (e.g., the minimal number of separating bonds) to the RC. Two hypotheses about the optimal functional form describing this dependence have been made: $f(d_{top}) = 1/d_{top}$ and $f(d_{top}) = \exp(-0.2d_{top})$. In a molecule with several substituents bound to its skeleton (Figure 1), it is judicious to regroup all the contributions of the substituents of the same type X into a common term, $i(X)$:

$$i(X) = \sum_{i \text{ atom of "X" type}} f(d_{top}(i,RC)) \qquad (5)$$

The considered types X of substituents correspond to families of functional groups with roughly similar electron-withdrawing features, for which a common $\sigma^I(X)$ value can be used without committing significant errors, and are defined in Table 1.

The total pK_a shift due to inductive effects, ΔpK_a^I, is a sum of the contributions per type of substituents, weighted by $\sigma^I(X)$ that stand for the intrinsic "*inductive power*" of a substituent X:

$$\Delta pK_a^I = \sum_X \sigma^I(X)\, i(X) \qquad (6)$$

ii. Resonance Effect. The contributions to the pK_a shift due to resonance effects are given by

$$\Delta pK_a^R = N_\Phi \sigma_\Phi + \langle \cos \alpha \rangle \Big(\sum_{\text{"}o\text{" positions}} \sigma^o(X_o) + \sum_{\text{"}p\text{" positions}} \sigma^p(X_p) \Big) \qquad (7)$$

where N_Φ is the number of the delocalized systems bound

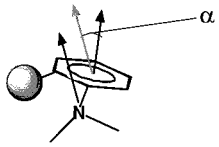

Figure 2. Definition of the collinearity parameter cos(α) as the cosine of the angle formed by the p orbital of the RC and the direction of the p orbitals of the delocalized system.

Table 2. Classes of Substituents Considered To Give Major Resonance Effects

X^R	substituents	X^R	substituents
1	N (amine)	5	N (pyridine)
2	O (hydroxy and ether)	6	C (unsaturated)
3	carbonyl, nitrile	7	COOR
4	NO_2, SO_2, SO_3		

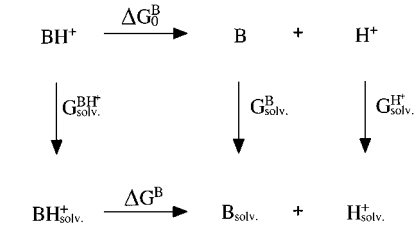

Figure 3. Thermodynamic cycle defining the solvation contribution to the free enthalpy of the proton exchange in solution.

to the RC, weighted by σ_Φ, and $\sigma^o(X)$ and $\sigma^p(X)$ are the specific intensities of the effects of a substituent X at the given (*ortho* or *para*) position. These specific intensities are corrected by the average degree of collinearity of the p orbitals of the RC and the π orbitals of the delocalized system ($\langle \cos a \rangle$; see Figure 2), averaged over all the conformations. Here, "*ortho*" and "*para*" denote the positions one and three bonds, respectively, away from the atom of the delocalized system carrying the RC. For example, in 1-aminonaphthalene, substituents in the 4 *or 8* position will be considered *para* with respect to the amino group.

Resonance effects over more than three separating alternating double/single or aromatic bonds are ignored. The different atom types or functional groups used in the description of the resonance effect are presented in Table 2.

iii. Electrostatic Field Effect. The electrostatic potential at the RC of the compound accounts for the free energy needed to transport a proton from "infinity" to the RC. The difference between the potential of the studied compound and that of the reference molecule is therefore related to the pK_a shift; see eq 8. The potentials at the RC are taken as the

$$\Delta pK_a^C = \sigma^C(\langle V_{coul} \rangle_{conf} - \langle V_{coul}^{ref} \rangle_{conf}) \qquad (8)$$

averages over all the sampled conformers, based on CVFF[38] partial charges, ignoring the contributions from the neutral charge group including the RC itself, with a relative dielectric constant $\epsilon_{int} = 2.0$.

iv. Solvent Effects. A solvation term accounts for the difference of polarization energy at the interface between the low-dielectric ($\epsilon_{int} = 2.0$) solute interior and the high-dielectric (ϵ_{solv}) solvent, at different protonation states of the solute. The free energy of deprotonation in a solvent can be established from the thermodynamic cycle in Figure 3.

$$\Delta G^B = \Delta G_0^B + G_{solv}^{H^+} + G_{solv}^B - G_{solv}^{BH^+} \qquad (9)$$

where ΔG_0^B is the deprotonation free energy in a hypothetical dielectrically homogeneous environment with a dielectric constant equal to that of the "interior" of the solute ($\epsilon_{int} = \epsilon_{ext} = 2.0$). It embodies the (inductive, resonance, electrostatic field) effects that are basically unrelated to the polarization effects caused by the reaction environment. G_{solv}^B and $G_{solv}^{BH^+}$ are the solvation energies (e.g., the transfer energies from the environment with $\epsilon_{ext} = \epsilon_{int}$ to the solvent where $\epsilon_{ext} = \epsilon_{solv} > \epsilon_{int}$) of the conjugated acid−base pair. They are calculated for each conformer with a fast "generalized Born" continuum approach called[39] the "Gilson−Still" model. $G_{solv}^{H^+}$ is a constant and can be ignored.

ΔG_{solv}^B is defined as the relative solvation energy of the basic species with respect to the protonated one.

$$\Delta G_{solv}^B = G_{solv}^B - G_{solv}^{BH^+} \tag{10}$$

This free energy difference represents the work required to change the state of polarization of the solvent surrounding the solute due to the rearrangement of the atomic charges during the loss of the proton by the solute. In the presence of substituents, this amount of work will differ with respect to that of the reference compound ΔG_{solv}^{ref}, and therefore

$$\Delta pK_a^S = \sigma^{solv}(\langle \Delta G_{solv}^B \rangle_{conf} - \langle \Delta G_{solv}^{ref} \rangle_{conf}) = \sigma^{solv} \Delta\Delta G_{solv} \tag{11}$$

In eq 11, the average variations of the solvation energy differences are calculated with respect to all the sampled conformers of the compounds. Dielectric constant values are set to $\epsilon_{int} = 2.0$ for the interior of the molecule and $\epsilon_{solv} = 80$ for the solvent (water). However, eq 11 may be used for other solvents, since the solvation energies are proportional to $1/\epsilon_{int} - 1/\epsilon_{solv}$ and weighing by σ^{solv} may compensate for the initial choice of ϵ.

v. Effect of Intramolecular Hydrogen Bonds. In this work, the pK_a shift due to intramolecular hydrogen bonds (HB) involving the RC is postulated to be proportional to the average number of potentially involved donor/acceptor groups. Potential HB partners are groups featuring free electron pairs or polar hydrogens more than two bonds from the RC, but with less than $d_{max} = 2.6$ Å between the involved heavy atoms. If the distance d exceeds d_{max}, the HB partner counter is incremented by a subunitary value, decreasing as a function of the violation of the d_{max} threshold.

$$\text{increment} = \min(1.0, e^{2.5(d_{max}-d)}) \tag{12}$$

The potential HB partners are counted in each sampled conformer of a compound to obtain the "fuzzy average count descriptors" N(X−H) and N(X:):

$$\Delta pK_a^H = \sigma_{don}^H N(X-H) + \sigma_{acc}^H N(X:) \tag{13}$$

2. Calibration of the Weighting Factors of the Acidity Prediction Model. A genetic algorithm (GA)[35,36] has been used to select the descriptors best suited to explain the 474 experimentally available ΔpK_a values. Chromosomes, defined as bit strings of length equaling the total number of potentially useful descriptors, encode a subset of "active" descriptors that enter the multilinear regression. This regression model is cross-validated (XV) by the leave-one-out

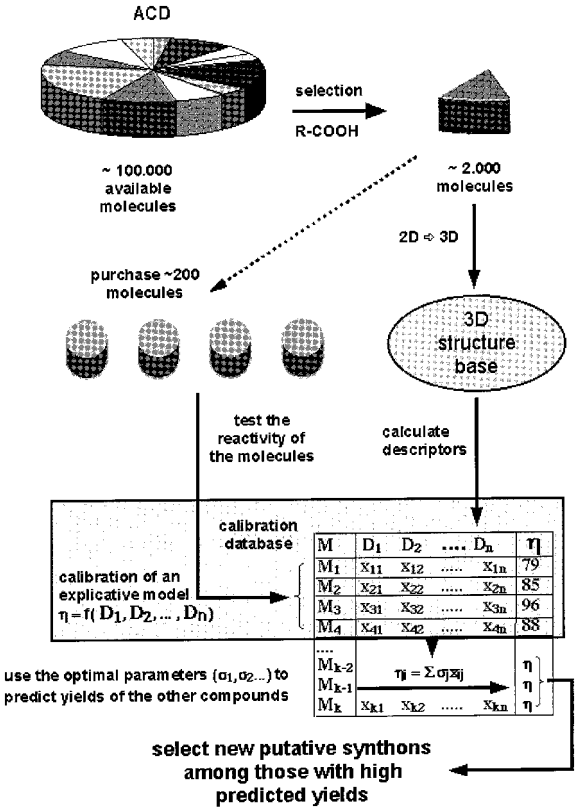

Figure 4. HTOS building block selection strategy.

method,[40] yielding the *correlation coefficient* r^2_{XV}—the "fitness function" to be maximized by the GA. The members of the current population of $N_{pop} = 100$ chromosomes are ranked with respect to decreasing fitness score. Mating pairs are constructed recursively: the not yet paired chromosome of highest fitness is associated with a partner among the most fit of the remaining unpaired chromosomes. $N_{pop}/2$ mating pairs encompassing the whole population are built. After the $N_{pop}/2$ crossovers are performed, during which point mutations (swapping) of a chromosome bit are allowed with a probability of 5%, the initial population doubles. The fitness of child chromosomes is subsequently evaluated, and the N_{pop} most fit chromosome out of $2N_{pop}$ form the new generation. This algorithm is run in parallel on networked workstations, and the runs are allowed to exchange chromosomes, on the average once every 10 generations. Whenever the evolution within one of these runs appears to reach a dead end, with no progress of the best score during 50 generations, its population is replaced with random chromosomes, except for the fittest member, which may be passed on to the new population with a probability of 80%. A total of 10 000 generations have been simulated within each of the 7 launched runs.

B. Reactivity Prediction Model. The present work illustrates a building block selection strategy (Figure 4) applied to the robotized synthesis of amides, from amines and carboxylic acids, activated by *N*,*N*′-carbonyldiimidazole,[41] in tetrahydrofuran−dimethylformamide. First, a structural query featuring the required functional group(s) retrieved the set of available building block candidates out of in-house or commercial molecular databases (ACD).[42] Out of these, a learning set was selected, its size representing a compromise between the need to cover a maximum of compound classes and the effort required to perform the

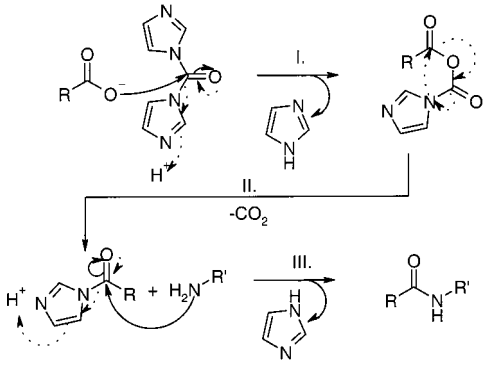

Figure 5. Reaction mechanism of the acylation of the amines with carboxylic acids in the presence of the activating agent *N*,*N*′-carbonyldiimidazole:[41] (I) addition−elimination of the carboxylic acid to the carbonyldiimidazole ("conversion-controlling" step); (II) rearrangement of the addition product; (III) addition−elimination of the amine to the acylimidazole intermediate. Full curved arrows show the electron density rearrangements at addition steps, while dotted curved arrows display the subsequent eliminations.

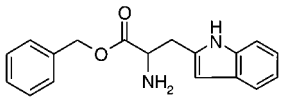

Figure 6. Reference amine used as a common reaction partner to evaluate the reactivity of the acids used to calibrate the reactivity model.

reactivity tests. A diversity[43]-biased selection strategy was applied to maximize coverage at a given set size. Eventually, the reactivity of the learning set members was measured, their molecular descriptors were calculated, and a linear relation between the reactivity and its explaining variables was established using the GA-driven variable selection scheme.

1. A "High-Throughput" Measure of Chemical Reactivity. Reactivity may be analyzed from both kinetic and thermodynamic points of view, by determining the activation free energy or the standard free energy variation. However, in an HTOS laboratory, the physicochemical study of hundreds of different reactions is prohibitive. Due to time and cost constraints, it is also impossible to assess the feasibility for each of the pairwise reactions between the building blocks (A_i, B_j) of a combinatorial library. In a building block reactivity test session, a chemist typically performs reactions between 100 and 200 building blocks A_i (carboxylic acids) and a *common* reference partner, B_{ref} (in the present work, the benzyl ester of tryptophan, Figure 6), chosen to represent the "average" (or slightly lower than average) reactivity of the partners B_j. The total conversion of B_{ref} is monitored by comparing the area of the HPLC peak of this reference product in the mixture prior to and after reaction. These experimental conversion values are not straightforwardly related to the thermodynamics and kinetics of the studied chemical reaction and may be biased by secondary processes involving the monitored reaction partner. They are low-cost, low-quality criteria of chemical reactivity. However, in terms of the quality/cost ratio, they represent a quite effective source of information used to construct a *relative scale* of reactivity of building blocks to select the most appropriate candidates in the considered HTOS process. Obviously, such a relative reactivity scale depends on the choice of the reference partner and does not ensure that all the combinatorial reactions between pairs of building blocks will proceed smoothly. Reactivity is not the only criterion in building block selection, so that less reactive synthons adding diversity to the collection can be tolerated even if they fail to react with some of their partners. In most cases of combinatorial library design, an average evaluation of reactivity is extremely helpful in building block selection.

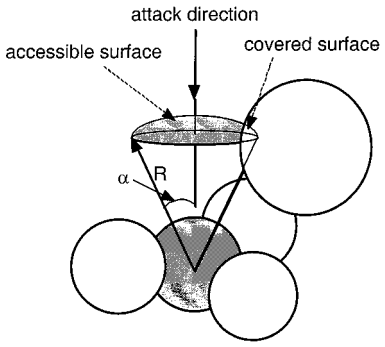

Figure 7. Definition of the accessibility descriptors $p(\alpha,R)$ with respect to a most probable attack direction.

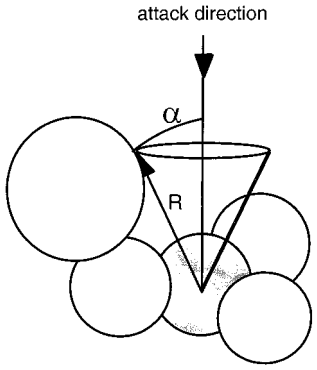

Figure 8. Maximal opening angle $A(R)$ with respect to the most probable attack direction, at a given distance from the reactive center.

2. Choice of Descriptors. In principle, descriptors of *both reaction partners* or those of the *product* molecules[32] need to be used to understand reactivity in a bimolecular process. However, due to the peculiar design of the HTOS reactivity tests, the descriptors of the common reaction partner are constants and can be ignored.

3. Steric Descriptors. The steric descriptors defined in this work are a measure of the hindrance due to the atomic van der Waals spheres (Figures 7 and 8) with respect to the most probable attack directions on the RC. In the case of −COOH, these are orthogonal to the plane of the group.

(i) The "*accessibility descriptors*" $p(\alpha,R)$ encode the fraction of the accessible surfaces of a series of spherical caps centered on the most probable attack direction (Figure 7) and corresponding to different values of radii and solid angles. Their radii R span a range from the van der Waals radius r of the RC to $r + 5$ Å, with an increment of 1 Å, while the opening angle of the associated cones ranges from 20° to 90°, with an increment of 10°.

(ii) The "*maximal opening angle*" $A(R)$ characterizes the opening of a cone that can be generated around the most probable attack direction of the RC, with the property that its base circle at radius R is tangent to at least one sphere, but does not intersect any of the spheres, of neighboring atoms (Figure 8). These angles $A(R)$ are calculated at different radii $R = (r, r + 1$ Å, ..., $r + 5$ Å) starting from the van der Waals radius of the RC.

For each conformer, these descriptors are calculated with respect to both of the two possible attack directions that correspond to the two (unequally crowded) faces of the −COOH plane, and the values corresponding to the less crowded face are kept. Their averages taken over all the conformers represent the steric descriptors used in this model.

4. Calibration of the Reactivity Prediction Model. Equation 14 shows the proposed relationship between the acylation conversions and the defined descriptors, where

$$\eta = \rho^{ac}\Delta pK_a + \sum_{\text{electronic/field effect terms}}^{\text{“}E\text{”}} \sigma^E(\rho^E - \rho^{ac})E + \sum_{\text{steric descriptors}}^{\text{“}S\text{”}} \xi^S S \quad (14)$$

$\Delta pK_a = \sum \sigma^E E$, E is a generic name given to the electronic and field effect descriptors from eqs 5−8, 11, and 13, while S stands for the steric descriptors. The latter are weighted by the coefficients ξ^S, while the factors of the electronic descriptors are now formulated as the product between their intrinsic intensities σ^E and the sensitivity ρ^E of the studied reaction to that particular effect.

It is considered that each electronic and field effect may be the source of both *nonspecific* and *specific* changes of reactivity. The nonspecific changes are represented by the $\rho^{ac}\sigma^E E$ that make up the first term $\rho^{ac}\Delta pK_a$ of eq 14 and express the overall dependence of the reactivity on the electron density and electrostatic property shifts at the RC. However, these static properties of the unperturbed RC represent only one of the multitude of kinetic and thermodynamic factors, including possible unwanted secondary processes, controlling the experimentally measured overall conversions. The specific changes in reactivity $(\rho^E - \rho^{ac})\sigma^E E$ were introduced to empirically correct for the factors that cannot be explicitly modeled, assuming that the variations in the experimental conversions can be statistically correlated with the presence of certain functional groups at certain locations in the molecule, as encoded by the effect descriptors E.

The ΔpK_a descriptor in eq 14 appears to be redundant, since it is a linear combination of several of the electronic and field descriptors that constitute the second term of the equation. However, the use of such a synthetic term that regroups all the electronic and field effects leads to a more robust model with less parameters. The GA only selects the specific electronic and field contributions E for which the impact they have on the reactivity ($\rho^E\sigma^E E$) strongly differs from the one related to the perturbation of the acidic−basic properties of the reactive group ($\rho^{ac}\sigma^E E$).

III. RESULTS AND DISCUSSIONS

A. Acidity Prediction Model. The best linear model includes 19 explaining variables and predicts the pK_a shifts of the 474 compounds in the learning set, with a cross-validated RMS error of 0.76 pK_a unit ($r^2 - r^2_{xv} = 0.932$) and a maximal error of 3.07 pK_a units. A plot of predicted vs experimental pK_a shifts of the compounds in the learning set is shown in Figure 9. Table 3 shows the selected descriptors and their associated weighting factors σ^E. Table 4 shows the distribution of the prediction errors of the pK_a shifts of the molecules in the learning set. The calculated pK_a values may not be precise enough to serve in the prediction of the ionization status of the compounds (although in 70% of the cases, they erred by less than 0.5 unit), but they are extremely useful in providing a rapid, automated discrimination between “electron-rich” and “electron-poor” reactive centers within a large series of building blocks.

Most of the coefficients reported in Table 3 have signs and orders of magnitude that agree well with the expected

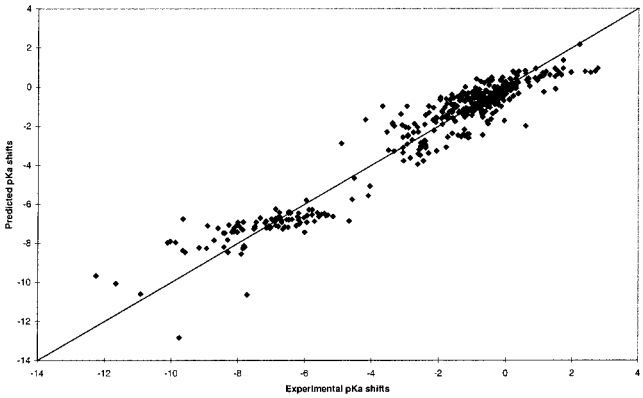

Figure 9. Predicted pK_a shifts versus experimental values for the 474 organic acids and amines used as a learning set (RMS = 0.76 pK_a unit, $r^2 = 0.932$).

Table 3. Descriptors Selected by the Genetic Algorithm and Their Weighting Factors in the Acidity Prediction Model[a]

substituent classes yielding inductive effects		σ^I (applied function of d_{top})	
XI	type	$(1/d_{top})\ \sigma_1^I(X)$	$\exp(-0, 2d_{top}))\ \sigma_1^I(X)$
1	aliphatic C	*	0.14
2	unsaturated C	−0.10	*
3 and 4	N and O	−6.15	3.71
5	F and Cl	−7.31	3.57
6	Br, I, P, S	−2.12	*
7 and 8	−CO−, ester, amide, nitrile	−0.91	−0.82
9	NO_2, NO, SO_2	−14.34	6.60

substituent classes yielding resonance effects		σ^R	
X^R	type	(*ortho*) $\sigma_o^R(X)$	(*para*) $\sigma_p^R(X)$
1	NR_2	*	+1.07
2	−OR	0.68	*
4	NO_2, NO, SO_2	−1.70	−1.18

solvation parameter $\sigma^{solv} = 0.037$

RC-type dependent parameters	nucleophilic RC	electrophilic RC
σ^Φ	−5.80	*
σ^H_{don}	*	−1.70
σ^H_{acc}	−1.70	*

[a] *, term not used in the model. d_{top} = topological distance from the RC.

Table 4. Error Distribution for the ΔpK_a Prediction of 474 Compounds

$\Delta\Delta pK_a$	no. of molecules	$\Delta\Delta pK_a$	no. of molecules
2.3−1.6	14	0.7	17
1.3−1.5	10	0.6	35
1.2	13	0.5	45
1.1	12	0.4	38
1.0	8	0.3	52
0.9	14	0.2	47
0.8	17	<0.1	152

relative intensities of the electronic effects. For the aromatic amines, a decrement of $\sigma^{\Phi} = -5.80$ corresponds to the loss of basic character associated with the presence of a phenyl ring attached to the nitrogen atom, while no such decrement needed to be introduced for the aromatic acids.

Neither of the two hypothesized context functions of the inductive effect seems to properly describe its decay in terms of the topological distance when used alone. However, if inductive effects estimated with both of these context functions are *simultaneously* allowed to enter the pool of putative descriptors, the GA selects in many cases *both* descriptors associated with the same substituent type, generating a composite context function of d_{top}. For example, the optimal function describing the pK_a^I shift due to the inductive effect of N or O atoms is found to be $\Delta pK_a^I(d_{top}) = -6.15/(d_{top}) + 3.71\exp(-0.2d_{top})$, e.g., $\Delta pK_a^I(1) \approx -3$ (strong acidifying effect of the N/O atom bound directly to the RC of hydrazines/hydroxylamines), but $\Delta pK_a^I(3) \approx 0$. This function yields a sharp decrease of the effect over the first two coordination spheres around the RC, more like $1/d_{top}^2$ than $1/d_{top}$ or $\exp(-0.2d_{top})$.

Alkyl groups were the only ones displaying a positive, "electron-enriching" inductive effect. The independent types 3 and 4 and, respectively, 7 and 8 entered the model with almost identical weighing factors, so that they could be merged to reduce the total number of parameters.

The Coulomb energy associated with the transfer of a proton from infinity to the RC is mostly compensated by the opposite variation of the solvent reaction field. By weighing the solvation field by $\sigma^{solv} = 0.037$, the linear model actually evaluates the effective ΔpK_a due to the sum of the Coulomb and solvation terms. σ^{solv} perfectly matches the theoretical value of $2.3RT\epsilon_{int}/(\epsilon_{solv} - \epsilon_{int}) = 0.036$ estimated on the basis of the discussed compensation effect.

The resonance effect of a *p*-amino group was found to decrease the acidity by 1 pK_a unit. However, the *p*-hydroxy and alkoxy groups did not enter the model, apparently due to the "noise" stemming from erratic pK_a shifts of the polyhydroxy/methoxybenzoic acids and amines available for calibration. The resonance effects of carbonyl and related groups and unsaturated carbon (vinyl/phenyl substituents) appear to be quite weak, not affecting the acidity constants by more than the typical imprecision of the model and therefore were not selected. Models including the mesomeric effect due to pyridine nitrogens in the aromatic systems carrying the RC, exemplified only by the *o*- and *p*-pyridinecarboxylic acids, failed the cross-validation tests. Ionization constants of the *amino* groups of *o*/*p*-aminopyridines are not available (protonation exclusively occurs at the pyridine N).

No hydrogen-bonding term is selected unless the use of family specific coefficients for the acids and amines is allowed. In this case, the presence of an HB *donor* group in the neighborhood of an *acidic* RC was found to expectedly decrease the pK_a by $(\sigma_{don}^H)_{acidic\ RC} = -1.7$ pK_a units. However, an HB donor does not appear to affect the acidity of a *basic* RC, and an HB *acceptor* appears to have no influence on the acidity of *acidic* groups, but *decreases* the pK_a of amines by $(\sigma_{acc}^H)_{basic\ RC} = -1.7$ pK_a units.

B. Evaluation of the Reactivity Data of the Building Blocks. The conversion of the acylation of tryptophan benzyl ester significantly depends on the structures of the carboxylic acids. By contrast, when a set of aliphatic primary and secondary amines were tested against a common acid, the conversions were in most cases superior to 90% and showed no significant spread. These results are consistent with the interpretation that step I of the mechanism in Figure 5 appears to control the final conversion of the process. The acylimidazole intermediate will quantitatively convert into an amide irrespectively of the structure of the added amine, unless the latter is not totally devoid of nucleophilic character.

Table 5. Weighting Factors of the Electronic and Steric Descriptors Entering the Reactivity Prediction Model[a]

Electronic Descriptors
(Global Weight of the pK_a Shift Descriptor, $\rho_{ac} = 9.98$)

effective weight $(\rho^E - \rho_{ac})\sigma^E$	selected effects E
0.32	solvation term $\Delta\Delta G_{solv}$
−20.8	inductive effect of oxygen
11.90	resonance effect of *p*-OR/OH

Steric Descriptors

term S	weighting factor ξ^S	term S	weighting factor ξ^S
$A(r+1)$	−0.35	$p(40°, r+2)$	+0.75
$A(r+2)$	+0.47	$p(90°, r)$	+1.81
$p(20°, r+2)$	−0.77	$p(90°, r+3)$	−1.14

free intercept = 58.9

[a] The nomenclature of the steric descriptors S is outlined in the Methods, section II.2.B.

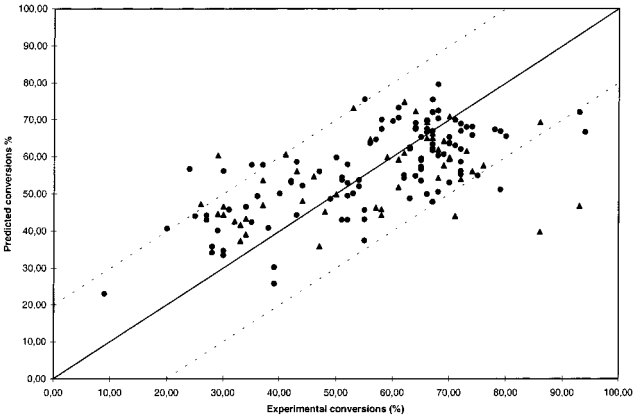

Figure 10. Predicted versus experimental acylation conversions for the studied carboxylic acids. The molecules used in the learning set are plotted with circles and those of the validation set with triangles.

C. Reactivity Prediction Model. Table 5 reports the 10 most relevant descriptors and their associated weighting factors entering the reactivity prediction model. Figure 10 shows the predicted conversions versus the measured values for both the molecules in the learning set (106 compounds, cross-validated correlation coefficient $r^2 - CV = 0.353$, and a cross-validated RMS error of 13.3%) and, respectively, in the validation set (52 molecules, $r^2_{pred} = 0.175$, $RMS_{pred} = 15.4\%$). The prediction error for the conversion of the reference amine was in 90% of the cases smaller than 20%. The few mispredicted molecules are shown in Figures 11 and 12.

Interestingly, the positive weighting factor of the ΔpK_a descriptor implies that the reactivity of an acid in the considered process *decreases* with its acidity. This is

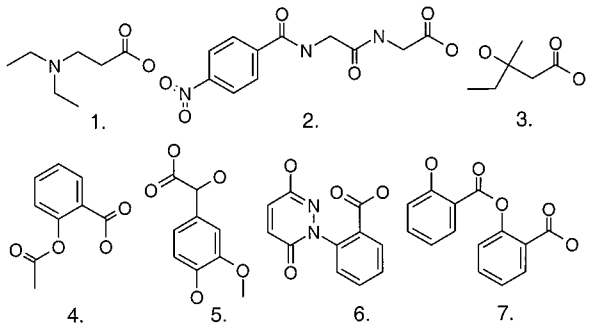

Figure 11. False negative outliers for which the model underestimated the experimental conversions by more than 20%.

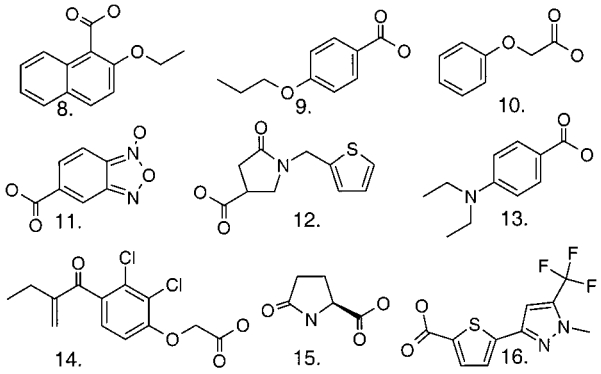

Figure 12. False positive outliers for which the model overestimated the experimental conversions by more than 20%.

somehow unexpected, since acids usually represent the *electrophilic* species in acylation. However, the proposed mechanism (Figure 5) in which the carboxylate acts as a nucleophilic reagent in the conversion-controlling step I explains this apparent contradiction.

The influence of the solvation term on the reactivity of the acids is more important than that expected on the basis of its influence on the pK_a shifts, the underlying physical reasons for this behavior being unclear.

Surprisingly, many phenoxy/alkoxyacetic acids were found to be much less reactive than expected on the basis of the inductive effect of the O atom only. While no explanation could yet be found for these findings, the reactivity model introduced a correction term by enhancing the weight of this inductive effect beyond its contribution to the DpK_a descriptor. Even so, phenoxyacetic acids **8**, **10**, and **14** (Figure 12) figure on the list of "false positives" (overestimated reactivity). The extrapolation of this empirical reactivity rule to good acylation reactants with −OH or −O−C═O groups in the vicinity of the RC (Figure 11) led to their erroneous classification as "false negatives". The model could be improved if distinct parameters are introduced for the OR and OH groups.

The resonance effect of *p*-hydroxy/methoxy groups, which had not been selected by the acidity prediction model, enters the reactivity model with a positive coefficient, in contrast to what has been observed for the pK_a prediction model (see discussions in section II.A).

Several steric descriptors enter the model, but it is difficult to find a rationale for the weights associated with each one of them. For acids with highly accessible carboxyl groups, such as linear or *ortho*-unsubstituted aromatic compounds, the global steric contribution $\sum \xi^S S$ is, as expected, positive (up to +25%), while for molecules with crowded RCs a negative steric contribution decreases the expected conversion values by up to −14%.

IV. CONCLUSIONS AND PERSPECTIVES

The design and synthesis of combinatorial libraries, one of the most promising approaches to drug discovery, highly benefits from rational selection strategies of building blocks. Deciding upon the building blocks that should be used in the synthesis of a library is a problem involving a manifold of structural, physicochemical, pharmaceutical, toxicological, and economic constraints. Computer programs aimed to guide the choice of the chemist with respect to different important criteria (molecular diversity, toxicity, ...) have been developed, while relatively few efforts have been registered in the equally important field of the reactivity-oriented computer-aided selection.

In the present work, a multilinear reactivity prediction model, based on a set of herein developed electronic and steric descriptors, has been calibrated to reproduce the relative reactivity scales that are experimentally established to discriminate between suitable and unsuitable building blocks for a given chemical synthesis. The model proved able to successfully estimate the reactivity of candidate building blocks in more than 90% of the cases. Therefore, it can be used to rapidly browse through large databases of 3D structures and to retrieve the subset of the compounds that are most likely to display the desired reactivity. Less testing time and reagents will be lost on unreactive building blocks than in the situations when the primary database query is based on a criterion other than reactivity.

The main drawback of the approach lies in the loss of chemically diverse families of synthons that are erroneously predicted to be nonreactive. An optimal strategy would consist in a selection according to a weighted average criterion of the predicted reactivity and the calculated gain in structure-space coverage associated with the candidate synthon.

ACKNOWLEDGMENT

We thank Professor André Tartar and Dr. Benoît Deprez for helpful discussions and suggestions. Dr. Brian Fulton is acknowledged for carefully proof-reading this text.

REFERENCES AND NOTES

(1) Chaiken, I. M., Kim, D. J., Eds. *Molecular Diversity and Combinatorial Chemistry: Libraries and Drug Discovery*; American Chemical Society: Washington, DC, 1996.

(2) Williard, X.; Pop, I.; Bourel, L.; Horvath, D.; Baudelle, R.; Deprez, B.; Melnyk, P.; Tartar, A. Combinatorial chemistry: a rational approach to chemical diversity. *Eur. J. Med. Chem.* **1996**, *31*, 87−98.

(3) Gallop, M. A.; Barret, R. W.; Dower, W. J.; Fodor, S. P. A.; Gordon, E. M. Applications of Combinatorial Technologies to Drug Discovery. 1. Background and Peptide Combinatorial Libraries. *J. Med. Chem.* **1994**, *37*, 1233−1251.

(4) Gordon, E. M.; Barret, R. W.; Dower, W. J.; Fodor, S. P. A.; Gallop, M. A. Applications of Combinatorial Technologies to Drug Discovery. 2. Combinatorial Organic Synthesis, Library Screening Strategies and Future Directions. *J. Med. Chem.* **1994**, *37*, 1385−1401.

(5) Lewis, R. A.; Mason, J. S.; McLay, I. M. Similarity Measures for Rational Set Selection and Analysis of Combinatorial Libraries: The Diverse Property-Derived (DPD) Approach. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 599−614.

(6) Gillet, V. J.; Willett, P.; Bradshaw, J. The Effectiveness of Reactant Pools for Generating Structurally-Diverse Combinatorial Libraries. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 731−740.
(7) Menard, P. R.; Mason, J. S.; Morize, I.; Bauerschmidt, S. Chemistry Space Metrics in Diversity Analysis, Library Design and Compound Selection. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 1204−1213.
(8) Li, S.; Gao, J.; Satoh, T.; Friedman, T. M.; Edling, A.; Koch, U.; Choksi, S.; Han, X.; Korngold, R.; Huang, Z. A computer screening approach to immunoglobulin superfamily structures and interactions: Discovery of small non-peptide CD4 inhibitors as novel immunotherapeutics. *Proc. Natl. Acad. Sci. U.S.A.* **1997**, *94*, 73−78.
(9) Kaminski, J. J.; Rane, D. F.; Snow, M. E.; Weber, L.; Rothofsky, M. L.; Anderson, S. D.; Lin, S. L. Identification of Novel Farnesyl Protein Transferase Inhibitors Using Three-Dimensional Database Searching Methods. *J. Med. Chem.* **1997**, *40*, 4103−4112.
(10) Damborsky, J.; Kuty, M.; Nemec, M.; Koca, J.; A Molecular Modeling Study of the Catalytic Mechanism of Haloalkane Dehalogenase: 1. Quantum Chemical Study of the First Reaction Step. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 562−568.
(11) Katritzky, A. R.; Gordeeva, E. V. Traditional topological indices vs electronic, geometrical, and combined molecular descriptors in QSAR/QSPR research. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 835−857.
(12) Zupan, J.; Gasteiger, J. *Neural Networks for Chemists - An Introduction*; VCH Verlagsgesellschaft mbH: Weinheim, Germany, 1993.
(13) So, S.-S.; Karplus, M. Genetic Neural Networks for Quantitative Structure−Activity Relationships: Improvements and Application of Benzodiazepine Affinity for Benzodiazepine/$GABA_A$ Receptors. *J. Med. Chem.* **1996**, *39*, 5246−5256.
(14) Klopman, G.; Ding, C.; Macina, O. T. Computer Aided Olive Oil-Gas Partition Coefficient Calculations. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 569−575.
(15) Rossiter, K. J. Structure-Odor Relationships. *Chem. Rev.* **1996**, *96*, 3201−3240.
(16) Katrizky, A. R.; Mu, L. QSPR Treatment of the Unified Nonspecific Solvent Polarity Scale. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 756−761.
(17) Mitchell, B. E.; Jurs, P. C. Prediction of Autoignition Temperatures of Organic Compounds from Molecular Structure. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 538−547.
(18) Jaffé, H. H. A reexamination of the Hammett equation. *Chem. Rev.* **1953**, *53*, 191−261.
(19) Ruasse, M. F.; Argile, A.; Dubois, J.-E. Selectivity Relationships and Substituent-Substituent Interactions in Carbocation-Forming Bromination. The Transition-State Contribution to the ρ Variation. *J. Am. Chem. Soc.* **1984**, *106*, 4846−4849.
(20) Lee, I.; Shim, C. S.; Chung, S. Y.; Kim, H. I.; Lee, H. W. Cross-interaction Constants as a Measure of the Transition-state Structure. Part 1. The Degree of Bond Formation in Nucleophilic Substitutions Reactions. *J. Chem. Soc., Perkin Trans. 2* **1988**, 1919−1923.
(21) Brown, H. C.; Okamoto, I.; Inukai, T. Rates of Solvolysis of the m- and p-Phenyl-, m- and p-Methylthio-, and m- and p-Trimethylsilylphenyldimethylcarbinyl Chlorides. Steric Inhibition of Resonance as a Factor in Electrophilic Substituent Constants. *J. Am. Chem. Soc.* **1958**, *80*, 4964−4968.
(22) Okamoto, I.; Inukai, T.; Brown, H. C. Rates of Solvolysis of Phenyldimethylcarbinyl Chlorides Containing Meta Directing Substituents. *J. Am. Chem. Soc.* **1958**, *80*, 4969−4979.
(23) Okamoto, I.; Brown, H. C. Electrophilic Substituent Constants. *J. Am. Chem. Soc.* **1958**, *80*, 4979−4987.
(24) Taft, W. R.; Lewis, C. I. Evaluation of Resonance Effects on Reactivity by Application of the Linear Inductive Energy Relationship. V. Concerning a σ_R Scale of Resonance Effects. *J. Am. Chem. Soc.* **1959**, *81*, 5343−5352.
(25) Swain, C. G.; Lupton, E. C. Field and Resonance Components of Substituent Effects. *J. Am. Chem. Soc.* **1968**, *89*, 4328−4337.
(26) Marriott, S., Topsom, R. D. A Theoretical Scale of Substituent Resonance Parameters (s_R^o). *J. Chem. Soc., Perkin Trans. 2* **1985**, 1045−1047.
(27) de Ligny, C. L.; van Houwelingen, H. C. The Tree-parameter Nieuwdorp Equation is the Optimal Linear Free Energy Relationship for the Prediction of Missing Data. *J. Chem. Soc., Perkin Trans. 2* **1987**, 559−562.
(28) Robinson, D.; Barlow, T. W.; Richards, W. G. The Utilization of Reduced Dimensional Representations of Molecular Structure for Rapid Molecular Similarity Calculations. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 943−950.
(29) ChemInform RX and ISIS are marketed by MDL Information Systems, San Leandro, CA.
(30) Gasteiger, J.; Rose, J. R. In *Software Development in Chemistry 8*; Jochum, C., Ed.; Springer: Berlin, 1994; pp 29−58.
(31) Gasteiger, J.; Rose, J. R. In *Software Development in Chemistry 9*; Moll, R., Ed.; Springer: Berlin, 1995; pp 129−139.
(32) Chen, L.; Gasteiger, J. Knowledge Discovery in Reaction Databases: Landscaping Organic Reactions by a Self-Organizing Neural Network. *J. Am. Chem. Soc.* **1997**, *119*, 4033−4042.
(33) *CRC Handbook of Chemistry and Physics*, 77th ed.; CRC Press: New York, 1996−1997; p 8-45−55.
(34) Albert, A.; Serjeant, E. P. *Ionization Constants of Acids and Bases A Laboratory Manual*; Methuen & Co. Ltd.: London; John Wiley & Sons Inc.: New York, 1962; Chapter 8, pp 124−148.
(35) Rogers, D. R.; Hopfinger, A. J. Applications of genetic function approximation to quantitative structure−activity relationships and quantitative structure−property relationships. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 854−866.
(36) Luke, B. T. Evolutionary programming applied to the development of quantitative structure−activity relationships and quantitative structure−property relationships. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1279−1287.
(37) Horvath, D. A Virtual Screening Approach Applied to the Search for Trypanothione Reductase Inhibitors. *J. Med. Chem.* **1997**, *40*, 2412−2423.
(38) Ermer, O. Calculation of molecular properties using force fields. Applications in organic chemistry. *Struct. Bonding*, **1976**, *27*, 167−211.
(39) Horvath, D.; van Belle, D.; Lippens, G.; Wodak, S. J. Development and parametrization of continuum solvent models. II. A unified approach to the solvation problem. *J. Chem. Phys.* **1996**, *105*, 4197−4210.
(40) Everitt, B. S.; Dunn, G. *Applied Multivariate Data Analysis*; Oxford University Press: New York, 1992.
(41) Staab, H. A. New Methods of Preparative Organic Chemistry IV. Syntheses Using Heterocyclic Amides (Azolides). *Angew. Chem., Int. Ed. Engl.* **1962**, *1*, 351−367.
(42) ACD−Available Chemicals Directory. Copyright 1995 MDL Information Systems Inc., San Leandro, CA.
(43) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching, *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983−996.

CI990104X