QSAR and QSPR Studies of a Highly Structured Physicochemical Domain

Orazio Nicolotti and Angelo Carotti*

Dipartimento Farmaco−Chimico, University of Bari, via Orabona 4, I-70125 Bari, Italy

Received July 19, 2005

The relevance of terms other than linear when deriving quantitative structure−activity relationship/quantitative structure−property relationship (QSAR/QSPR) models has been rarely considered so far. In this study, the impact of quadratic and interacting terms has been taken into account. The first effect of including such highly structured terms is a significant extension of the parametric domain that moves from the initial N to $N(N + 3)/2$ parameters. This substantial enlargement over the conventional linear boundaries involves a higher computational cost due to the increased combinatorial number of resulting theoretical QSAR/QSPR models. To face this issue, novel genetic-algorithm-based software, MGZ (multigenetic zooming), was developed and used for both variable selection and model building. To speed up the entire process of domain searching, MGZ was supported with multiple independent evolving populations and genetic storms to further QSAR/QSPR analyses. In addition, a novel fitness function was developed to score models on the basis of their inner predictive capability, assessed on the training set, structure complexity, and presence of nonlinear terms. The models were further validated by monitoring model redundancy and performing intensive randomization runs. The Selwood data set was used as a reference set to derive QSAR models. Furthermore, a QSPR study was conducted on the solubility data set of a large array of organic compounds. The results reported in the present paper demonstrate that our approach is successful in finding linear models, which are at least as good as the models previously derived using standard statistical approaches, and in deriving new nonlinear models with good statistical figures.

INTRODUCTION

The ultimate goal of quantitative structure−activity relationship (QSAR) studies is to relate the biological activity (Y) of a series of compounds with some appropriate descriptors (X). More specifically, the aim of a QSAR study is the extraction of the most significant signals of the Y variance from the often undistinguishable fields of information and noise contained in the space of molecular physicochemical properties (X). Several closely related theoretical and methodological aspects underscoring the different approaches used in QSAR studies are still the object of a large debate and need to be discussed briefly herein.[1]

The number and nature of molecular descriptors employed to characterize any system under investigation indeed constitute a central issue in QSAR investigations. Descriptors represent the domain of information, or in other words, the source of theoretical and experimental knowledge from which the QSAR models should originate to explain the variation of biological data. As a result, one might think that the higher the number of molecular descriptors, the better. However, QSAR is not only a matter of data modeling, and therefore, consideration must be given to the validity and desirability of a given model. While the former feature can be evaluated with proper statistical parameters, the latter is perhaps questionable, since it has no mathematical definition but relies on the interpretability or applicability of the models and, ultimately, on the common sense of medicinal chemists. In addition, other critical aspects are related to the different techniques and methods used to elaborate appropriate mathematical statements that can opportunely represent QSAR models. Since Hansch's first approach,[2] considerable effort has been invested in improving the modeling of biological data, and as a result, many significant advances have been made. Presently, many molecular descriptors, such as the refractive index, octanol/water partition coefficient, water solubility, spectral data, and so forth, may be determined experimentally.[3] Alternatively, many descriptors can be estimated from theoretical calculations encoded in several commercially or publicly available computer programs.[4] Over the years, these descriptors constituted the starting materials of the so-called 2D-QSAR studies that are the principal focus of this study. The reader should keep in mind that a distinct but still important role has been played by grid-based descriptors (i.e., molecular interaction fields) from which 3D-QSAR investigations originated.[5] Despite some drawbacks mostly related to the arbitrary criteria chosen for molecular alignment, the user has the opportunity to visualize the main forces determining the biological activity at the 3D level.

In 2D-QSAR, it is easy to generate huge data matrixes in which each row represents a candidate compound and each column an experimental or computed descriptor so that the limit of current QSAR modeling is represented by the small number of dependent variables that constitute the biological data.[6] To fully understand the complexity of running a complete search of QSAR models,[7] it should be kept in mind that a combinatorial number of $2^N - 1$ models can theoretically result when N descriptors are available, irrespective of the degrees of freedom. For instance, a set of compounds described by only 53 properties involves 9×10^{15} combinations of descriptors. Even if these combinations could be analyzed in 1 second each, it would take 285420921 years to obtain them entirely.[8] Indeed, the need to keep things simple, as purported in Occam's Razor, which represents a

* Corresponding author e-mail: carotti@farmchim.uniba.it.

10.1021/ci050293l CCC: $33.50 © 2006 American Chemical Society
Published on Web 12/14/2005

well-known rule of thumb, suggests a drastic reduction in the combinatorial number of possible QSARs. Even when considering that a minimum set of five compounds is necessary to justify the presence of each descriptor in a given QSAR model, an enormous number of theoretical QSARs have yet to be generated to carry out a full systematic search. In fact, to explore all the theoretical three-term models derived from a data set of N molecular descriptors, a still large number of models [i.e., $\{N(N-1)(N-2)\}/\{3!\}$] must be calculated. In the case of the well-known Selwood data set, composed of 53 physicochemical properties, 23426 different three-term models exist.[9] In QSAR studies of large data sets, variable selection and model building become an even more difficult task when one considers that parametric space need not comprise only linear terms. Indeed, in many instances, the use of exponential terms (generally quadratic) in QSARs proved very useful as they could give rise to mathematical functions characterized by minima and maxima in the presence of the corresponding linear terms. In addition, interacting terms, more specifically, cross-product terms, could be relevant when the influence of two factors on a given response is not independent. An example of a quantitative structure−property relationship (QSPR) equation containing linear, nonlinear, and interacting terms is represented by the general solvation equation developed by Abraham and Joelle.[10] The equation, consisting of only linear and nonlinear terms, was significantly improved through the incorporation of an interacting term, $\sum\alpha_2^H\sum\beta_2^H$, representing the product of the hydrogen bond acidity (α_2^H) and basicity (β_2^H) of the solutes. For sake of completeness, the entire equation is reported below:

$$\log S_w = 0.518 - 1.004R_2 + 0.771\pi^2 + 2.168\sum\alpha_2^H + 4.238\sum\beta_2^H - 3.362\sum\alpha_2^H\sum\beta_2^H - 3.987V_x$$
$$r^2 = 0.920;\ n = 659 \qquad (1)$$

where R_2, π^2, $\sum\alpha_2^H$, $\sum\beta_2^H$, and V_x represent excess molar refraction, compound dipolarity/polarizability, the summation of both the hydrogen bond acidity and basicity, and the McGowan volume, respectively. The interested reader is referred to the original paper[10] for further details on the physicochemical descriptors and statistical attributes.

It is worth stating that the number of possible QSAR models dramatically increases as the inclusion of quadratic terms doubles the starting parametric space ($2N$), whereas the total number of interacting terms[11] is equal to $N(N-1)/2$. In other words, an even more overwhelming number of QSARs ($2^{N(N+3)/2}-1$) will result if one wants to explore the parametric space with the inclusion of quadratic and interacting terms. Since molecular parametric space is composed of a large number of chemical descriptors, attaining such a figure is unfeasible and time-consuming. Thus, considerable efforts have been made to develop innovative and efficient tools to cope with such a high combinatorial number of QSARs aimed at exploring and selecting the most informative models. The computational cost associated with feature selection has resulted in a number of different algorithms for feature selection and QSAR generation, such as principal component analysis (PCA), nonlinear mapping, partial least squares (PLS), neural networks, and evolutionary methods.[12]

Feature selection represents the most critical step when deriving QSAR/QSPR models from a large parametric space. The paper by Selwood et al.[13] provided a milestone data set for QSAR practitioners and has become a standard against which new methods are continuously tested. In the initial approach, Selwood et al. made use of forward-stepping multivariate regression analysis to obtain a three-descriptor model after generating a training and a test set. Wikel and Dow[14] constructed an improved three-descriptor model using a neural network for variable selection and presented the first result of an analysis in which the whole data set was used without setting apart a test set. Rogers and Hopfinger[15] analyzed the Selwood data set through the genetic function approximation (GFA) method, in which feature selection is performed using a genetic algorithm (GA) and QSAR models are derived through least-squares regression, with Friedman's lack of fit (LOF) being used to score QSAR models. The LOF function is based on the least-squares errors combined with a user-definable smoothing parameter that penalizes the inclusion of a high number of terms in a model. Three-descriptor models have been found to be the best models. In addition, relatively good two- and four-descriptor models have been developed. Interestingly, four- to six-descriptor models with modest enhancements in prediction were derived by reducing the smoothing parameter of the LOF function. The mutation and selection uncover models (MUSEUM) algorithm developed by Kubinyi[9] has also been applied to the study of the Selwood data set. MUSEUM is based on an evolutionary algorithm that uses only the mutation operator to generate offspring. The FIT value is used as the fitness criterion that, to some extent, is an improvement on the Fischer significance value in that FIT is better calibrated toward the change in number of independent variables selected in each model. A related method, called the evolutionary programming method, developed by Luke[16] has also been challenged by the Selwood data set. Its fitness function is defined by three terms. The first term is the root mean square between predicted and measured values. The second term is set to drive the solution toward a given number of descriptors, and the final term is designed to weigh the descriptors on the basis of their exponential values. For example, quadratic terms are penalized relative to linear terms. Both Kubinyi's and Luke's methods succeeded in finding new three-descriptor models. All previous methods have been oriented toward the development of conventional linear models. So and Karplus[17] proposed a hybrid method that combines a GA for selecting descriptors and an artificial neural network for deriving models. More recently, a novel algorithm based on the fast random elimination of descriptors (FRED) was proposed.[18] FRED has been able to find the same solutions for the Selwood data set as other previous methods. Cho and Hermsmeier[19] developed a novel GA-guided selection method (GAS) to simultaneously optimize a set of encoded variables that include both descriptors and compounds and to construct QSAR models that are comparable to those obtained with other methods. Inspired by studies of human sociality, Agrafitios and Cedeno[20] devised a new feature selection algorithm known as swarm particles. The search space is explored by a population of individuals, who adapt by returning toward previous successfully explored regions through semistochastic movements. More recently, a new approach called multiobjective QSAR

(MoQSAR) was elaborated upon on the basis of genetic programming to create a population of potential model solutions and on the basis of a multiobjective fitness function to handle multiple objectives independently without summation or weights.[21] By adopting Pareto ranking, MoQSAR allowed the discovery of families of tradeoff QSAR models, rather than a single solution, in only one run. According to this approach, the medicinal chemists were, therefore, given the option of choosing the most appropriate model for drug design from a pool of equivalent QSAR solutions.[22]

In this paper, we report a novel method based on a new implemented GA for QSAR variable selection that, in our opinion, may present some advantages over existing approaches. Because medicinal chemists are generally reluctant to consider complex nonlinear relationships, a novel fitness function was developed in which the occurrence of terms, other than those linears, was allowed only if the resulting QSAR/QSPR models showed significantly higher internal predictive power as measured by the squared correlation coefficient of predictions (q^2). A series of validation studies based on randomization analyses and measures of global correlation indexes was also conducted. Two different data sets, the well-known Selwood data set and the solubility data for a large array of organic compounds, were used as benchmarks to challenge our approach. In the latter data set, additional external predictions, expressed as $r_{trn/ts}^2$, were also advanced.

COMPUTATIONAL METHODS

Genetic Algorithms for Variable Selection in QSAR. GAs are inspired by the evolution of DNA.[23] In short, individuals are represented as a 1-dimensional string of bits; an initial population of individuals is created, generally with random initial bits, and a fitness function is used to evaluate the quality of an individual so that the best individuals are assigned the best fitness scores. As a result, these individuals will more likely be chosen to propagate their genetic material to offspring through two basic recombination mechanisms. One is the mutation that causes a random change of separate elements within a chromosome. A mutation is generally considered to be a background operator as it ensures that the probability of searching a particular subspace is low and never zero. Mutations can generally result in pathological conditions; however, such an irregular change can also determine good results contributing to the evolution of the population through searches in new zones of the parametric space, thus avoiding deadlock situations. The other and more powerful genetic operator is crossover, in which a portion of genetic material is taken from each parent and recombined to create new child chromosomes. Among the different recombination schemes that can be adopted, MGZ performs a multipoint crossover that is intended to encourage the exploration of the search space rather than favor convergence to fit individuals early in the search.[24] Establishing a given number of mating steps, the average fitness of the individuals in a population increases as good combinations of genes are discovered and spread throughout the population. GAs are very suitable for searching high dimensional spaces, as they perform the sampling in very efficient and direct ways.

In MGZ, the user has the option of modifying the conventional GA scheme in two ways. In the first, MGZ was provided with an additional function that operates genetic

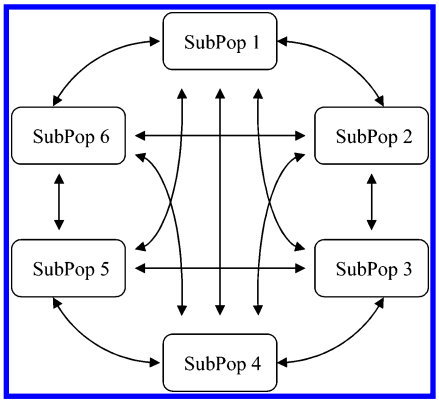

Figure 1. Scheme of the unrestricted migration topology adopted in MGZ.

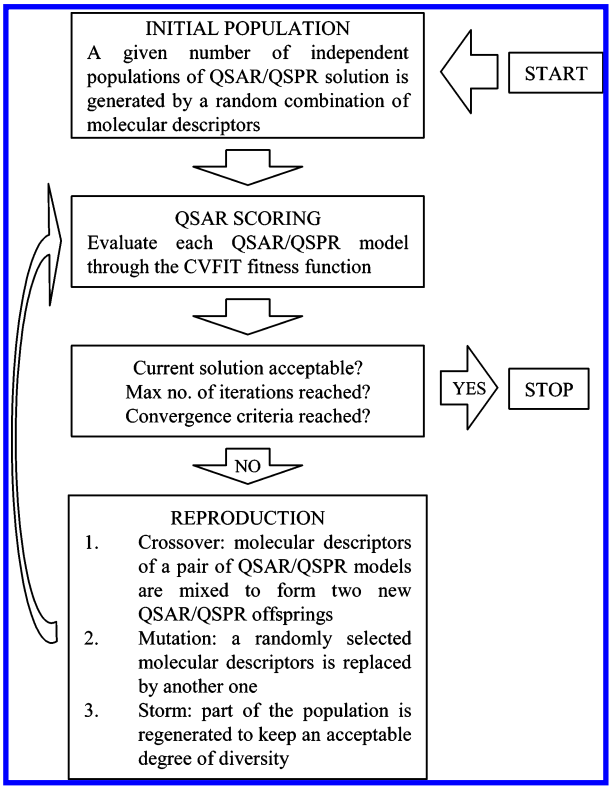

Figure 2. Flow diagram showing the GA search adopted in MGZ for QSAR/QSPR modeling.

storms to maintain an acceptable degree of diversity within the population; in other words, when the average fitness of a given evolving population is nearly constant, the replicated and worst performing chromosomes are destroyed by sudden death, favoring the birth of new individuals. In the second scheme, MGZ is allowed to operate with multiple independent evolving populations. By using this approach, it has been observed that the quality of results can be improved compared to the single-population GA approach. The adopted scheme is known as the *migration* or *island* model.[25] A single population was, therefore, divided into subpopulations that were independently evolved over generations by the GA. Individuals could migrate from one subpopulation to another according to an unrestricted migration topology, as shown in Figure 1. All the genetic steps adopted in the GA-based search are summarized in the flowchart of Figure 2.

Fitness Function for Scoring QSAR Models. The fitness function indeed plays the most critical role in evolutionary

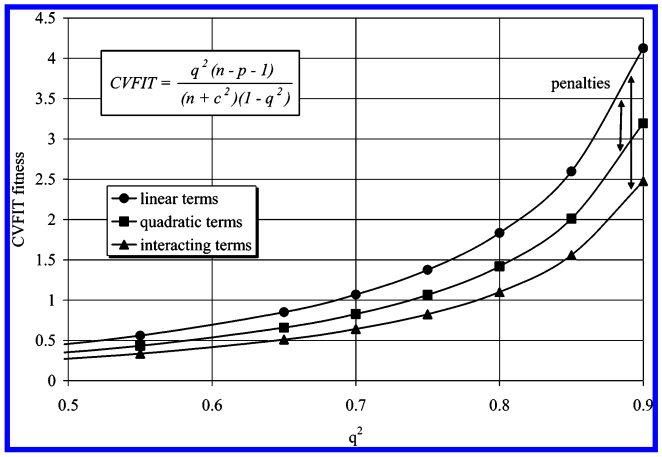

Figure 3. Simulation of the change of the CVFIT fitness at varying q^2 values. The impact on linear models of the inclusion of one quadratic or one interacting term is evaluated, keeping constant both the global number of terms (p) and the global number of objects (n). See text for additional comments and discussion.

methods for driving the sampling of the parametric space toward optimal solutions. To speed up the entire search process, the fitness function should take into account figures that are easy to calculate and are relevant for robust QSAR modeling. The principles of parsimony inspired a considerable number of different fitness functions described so far in the literature. The majority included parameters associated with model fitting and model structural complexity. In the present work, the fitness value of each chromosome (i.e., a QSAR model) was calculated using the following novel function:

$$\mathrm{CVFIT} = \frac{q^2(n - p - 1)}{(n + c^2)(1 - q^2)} \tag{2}$$

where q^2, n, and p represent the squared correlation coefficient of predictions of the training set, the number of compounds, and the number of molecular descriptors in the model, respectively. The term c is calculated as follows:

$$c = \text{linears} + 2 \times \text{quadratics} + 3 \times \text{interactings} \tag{3}$$

where linears, quadratics, and interactings represent the number of linear, quadratic, and interacting terms, respectively. Therefore, the presence of terms other than linear were penalized through the adoption of a weighting scheme. Assigning double and triple weights to quadratic and interacting terms proved to be particularly suitable for medium-sized molecular data sets that are the most diffuse in QSAR practice. More specifically, quadratic and interacting terms were included in a linear QSAR model only if the gain in internal prediction (q^2) was higher than the increment in model complexity. The effect of including one quadratic and one interacting term in a linear model is described in Figure 3. As can be observed, nonlinear models can compete with corresponding linear models only when they are provided with significantly better predictive statistics. Considering a constant number of terms, a model with an interacting term would be preferred to the corresponding linear term if its q^2 value is incremented by nearly 10%.

The fitness function proposed herein to some degree represents an improvement over the FIT function proposed by Kubinyi, derived in turn from the classic F Fischer significance value. Consequently, two main changes have been made to the FIT function:[26] one consisted in the option to also include nonlinear terms, whereas the other permitted the use of q^2 instead of r^2. The q^2 statistical parameter was first developed and used in PLS analysis to prevent the problem of overfitting and, thus, to determine the optimal number of model components.[27] Being a better statistical parameter, q^2 was subsequently considered the standard for quantifying the internal predictive power of a QSAR model, whereas r^2 is the standard measure of model fitting power. In our data analysis, leverage values (h_{ii}) were used to quickly compute the predicted value of the ith response ($y_{i/i,\mathrm{pred}}$) instead of the classic but unfeasible and highly time-consuming leave-one-out (LOO) procedure that is based on calculations over n cancellation groups, each one consisting of $n - 1$ objects. The concept of leverage is based on the intuitive assumption that the ability to make predictions is steadily reduced as the distance from the center of the experimental model space becomes greater.[28] At its core, leverage assesses how far away a value of prediction is from the domain of applicability: the farther away the assessment, the higher the leverage. More specifically, the difference between the calculated and predicted values of the ith response can be measured from the corresponding leverages using the following formula:

$$y_{i,\mathrm{calcd}} - y_{i/i,\mathrm{pred}} = (y_{i,\mathrm{obs}} - y_{i,\mathrm{calcd}})\frac{h_{ii}}{1 - h_{ii}} \tag{4}$$

By means of the above equation, the predicted values of the ith responses ($y_{i/i,\mathrm{pred}}$) can be easily computed from the observed values $y_{i,\mathrm{obs}}$, those calculated through the model $y_{i,\mathrm{calcd}}$, and the corresponding leverages h_{ii}. As a result, the predictive residual sum of square (PRESS) and q^2 can be calculated as follows:

$$\mathrm{PRESS} = \sum_{i=1}^{n}(y_{i/i,\mathrm{pred}} - y_{i,\mathrm{obs}})^2 \tag{5}$$

$$q^2 = 1 - \frac{\mathrm{PRESS}}{\sum_{i=1}^{n}(y_{i,\mathrm{obs}} - \bar{y})^2} \tag{6}$$

where $\bar{y}$ represents the average values of the responses.

Analyses for QSAR Validation. The LOO procedure is a powerful tool that provides direct information on the internal predictive power of the models obtained. However, one should be aware that LOO results could be too optimistic as a result of the presence of chance correlation between the set of molecular descriptors forming a model and its response. A serious risk of chance correlation can indeed result when terms other than the linear term are included in a model. In fact, such terms could lead to the emergence of noisy and often meaningless QSAR models. While the latter statement is always questionable and dependent on the experience of a given QSAR researcher, the noise in the data can be controlled. For instance, an interacting term should be considered eligible as a useful molecular descriptor only when it is the product of two linear terms that are both different and minimally correlated. If the terms were highly correlated, it would be theoretically better to take only one

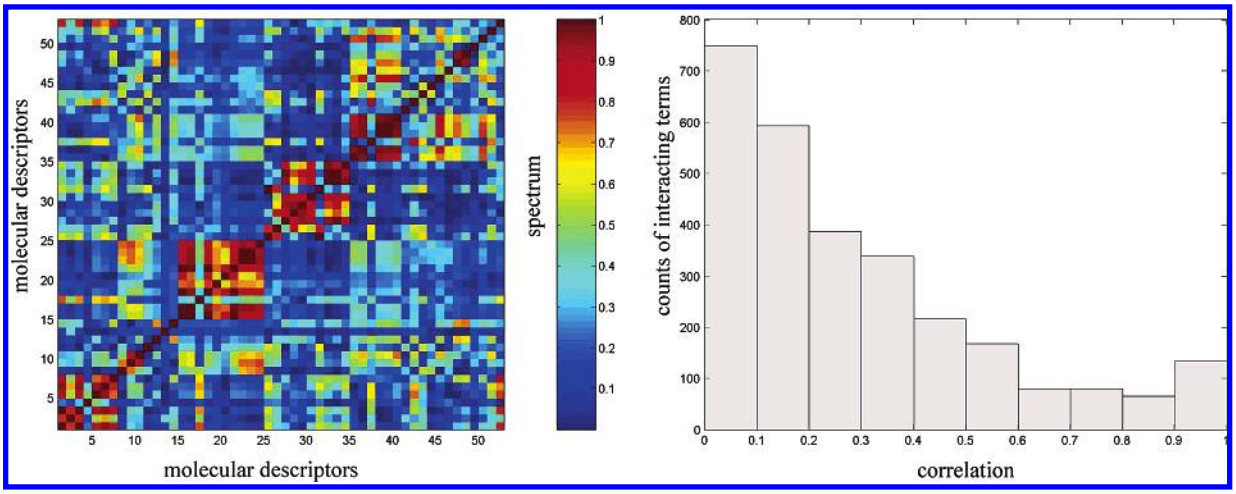

Figure 4. Symmetric color map and histogram of the correlation of the Selwood data set on the left and right side, respectively.

of the two; on the other hand, a quadratic term would result by combining two identical descriptors, yielding a much greater gain in chemical interpretability, albeit at the expense of the correlation. From this standpoint, the statistical validity of interacting terms can be monitored through the use of a symmetric correlation map whereby the amount of collinearity for each pair of molecular descriptors can be efficiently represented by a spectrum of colors ranging from blue to red to indicate a shift from minimum to maximum correlation, respectively. As shown on the left-hand side of Figure 4, where the symmetric correlation map of the Selwood data set was reported as an example, most regions are in blue, indicating the presence of interacting terms formed by mutually independent variables; the smaller red-like portion suggests the presence of highly correlated interacting terms. On the right-hand side of Figure 4, a quantitative measure of the above-mentioned concept was reported by unfolding the color map in a histogram with bins of descriptor pair correlations sized to 0.1.

Nonetheless, the global informative content of a set of molecular descriptors can be measured calculating a redundancy index, known as the *K* index,[29] which represents the total amount of correlation contained in each QSAR model. Such an index was evaluated from the decomposition of the eigenvalues of the corresponding correlation matrix. The explained variance EV_i derived from the *i*th principal component was first calculated according to the following formula:

$$EV_i = \frac{\lambda_i}{\sum_{i=1}^{p} \lambda_i} \tag{7}$$

where $\lambda_1 - \lambda_p$ represents the eigenvalues derived by means of a PCA of the molecular descriptors in a given QSAR model; the *K* index was calculated as follows:

$$K = \frac{\sum_{i=1}^{p} \left| EV_i - \frac{1}{p} \right|}{\frac{2(p-1)}{p}} \tag{8}$$

K ranges from 0 to 1 and represents a measure of collinearity of the set of *p* molecular descriptors constituting a given QSAR model. In addition, the robustness of all models found by MGZ and reported in the present paper were further challenged by routinely performing two different types of full randomization analyses with the aim to assess the risk of chance correlations between the values of the molecular descriptors and biological activities. In the first approach, the *y* values were shuffled 50 times and each of these scrambled vectors was correlated to each QSAR model previously found. In the second randomization study, *y* was scrambled 500 times and each of these vectors was used for the derivation of new QSAR random models by resubmitting the data to MGZ for an evolutionary search. The rationale behind these analyses is that, if the statistics of models arising from randomization are comparable to those of the original QSAR equations, then the probability of chance correlations will be very high. On the basis of our first randomization study, the highest and, thus, more optimistic value of q_{SCR}^2 found after 50 randomization runs is reported for each model in addition to the q^2 value of the corresponding unscrambled QSAR model.

All of these procedures have the sole aim to increase the confidence that one should have in the model's predictivity; however, they do not guarantee any reliable extrapolation outside the explored physicochemical domain. Indeed, the only way to challenge the predictive power of a model is by predicting the activity of a true external set of molecules.[30] No relationship between internal and external predictions has yet been described, and very often, even higher internal predictivity may result in low external predictivity and vice versa. This dilemma, also known as *Kubinyi's paradox*,[31] represents the ultimate and perhaps unsolvable challenge for QSAR practitioners. In the present paper, external predictions were made for the nitrogen data set.

Software Implementation. MGZ, written in Matlab code,[32] is a customizable toolbox available as a stand-alone system or with a user-friendly graphic interface. Besides the options already described, MGZ provides many additional features. Briefly, the well-known fitness functions FIT, AIK, q^2, r^2, and *F* are available in addition to CVFIT. A series of options for data transformation are also available such as the calculation of quadratic or interacting terms, the extraction of principal components, *y*-scrambling analyses, the

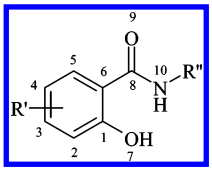

Figure 5. General structure of the 31 compounds of the Selwood data set.

computation of *K* index values, and diverse tools for visual inspection. Parameters for the GA runs are also user-adjustable according to the complexity of the QSAR design. In particular, the user can select the multiple-population option. In our laboratory, MGZ is usually run on a Pentium 4 2.6 GHz PC with 1 GB RAM and on a five-node-cluster Linux openMosix[33] machine (AMD Athlon XP 2400+ CPU). Three different and independent populations were adopted for the evolutionary search with the number of individuals being three times the pool of initial descriptors of each given data set. Mutation and crossover genetic operators were set at 0.1 and 0.9, respectively. MGZ requires about 1.5 h of computing time to perform 5000 iterations on the Pentium PC in the analysis of the Selwood data set.

RESULTS AND DISCUSSION

Selwood Data Set. The Selwood data set[13] contains 31 antimycin A1 analogues tested for their in vitro antifilarial activity. These compounds, whose general structure is reported in Figure 5, have been described using the following 53 different physicochemical parameters from Molconn-Z:[34] partial atomic charge for atoms 1−10 (ATCH1−ATCH10), dipole vectors (DIPV_X, DIPV_Y, and DIPV_Z), dipole moment (DIPMOM), electrophilic superdelocalizability for atoms 1−10 (ESDL1−ESDL10), nucleophilic superdelocalizability for atoms 1−10 (NSDL1−NSDL10), van der Waals volume (VDWVOL), surface area (SURF_A), principal moments of inertia (MOFI_X, MOFI_Y, and MOFI_Z), principal ellipsoid axes (PEAX_X, PEAX_Y, and PEAX_Z), molecular weight (MOL_WT), substituent dimensions (S8_1DX, S8_1DY, and S8_1DZ), substituent centers (S8_1CX, S8_1CY, and S8_1CZ), partition coefficient (LOGP), melting point (M_PNT), and sums of the *F* and *R* substituent constants (SUM_F and SUM_R).

Nitrogen Data Set. This data set, kindly supplied by McElroy and Jurs,[35] was selected for a QSPR study. It consists of 176 organic compounds with a minimum of one nitrogen atom, zero or more oxygen atoms, and zero or more halogens per molecule. As reported in the original paper, the geometries of the analyzed compounds were optimized by means of the PM3 and AM1 Hamiltonians to calculate structural and chemical parameters. The solubility (mol/L) values, ranging from −7.41 to 0.96 log units, and 12 descriptors were also made available. As done in a previous work,[36] 28 additional descriptors were calculated using Molconn-Z to obtain a total of 40 descriptors. The original 12 descriptors were as follows: MDE_14, the molecular distance edge between all primary and quaternary carbons; GEOM_1, the first geometric moment; GEOM_3, the third geometric moment; PPSA_1, the summation of the positive surface area; FPSA_1, the positive surface area divided by the total surface area; SCDH_2, the average surface area times charge on donatable hydrogens; NN, the number of nitrogens; NSB, the number of single bonds; WTPT_2, the sum of unique weighted paths divided by the total number of atoms; EAVE_2, the average E-state value over all heteroatoms; FPSA_2, the fractional charged partial surface area; and CTDH, the number of donatable hydrogens. The reader is referred to the Molconn-Z manual for a description of the remaining 28 parameters which are as follows: x0, x1, x2, xv0, xv1, xv2, xvp3, dx0, dx1, dx2, dxp3, dxv0, dxv1, dxv2, dxvp3, k0, k1, k2, k3, ka1, ka2, ka3, Si, Totop, SumI, SumdelI, tets2, and Phia.

Application of MGZ **to the Selwood Data Set.** The Selwood data set was analyzed using MGZ with CVFIT as the fitness function. MGZ was first applied to construct linear models containing up to five descriptors. As shown in Table 1, the best model (CVFIT = 1.236) was the three-term linear model consisting of LOGP, MOFI_Y, and SUM_F as descriptors, often found in previous studies.[8,9,13−17,21,26] The second- (CVFIT = 1.223) and third-best (CVFIT = 1.218) QSAR equations were still three-term models, with LOGP as the common descriptor. A series of four-term models, ranking from 4th to 11th place, yielded CVFIT values ranging from 1.100 to 1.033. All these models can be reasonably considered as an extension of the previous three-term models in which an additional descriptor (DIPV_Y, DIPV_X, DIPMOM, or others) is present. The best five-term model was placed in the 13th row of the table (CVFIT = 1.022); other three- and four-term models followed. The best two-term (LOGP and MOFI_Z) model was recorded as the 42nd best (CVFIT = 0.914) overall. Interestingly and as wished, the fitness function CVFIT tends to award models with a limited structural complexity, that is, with fewer descriptors. The results of the MGZ run (listed in Table 1) confirmed some QSAR models already found by others through different methods.[8,9,13−17,21,26] In particular, the top three models in Table 1 were also derived through the GFA, FIT and MoQSAR approaches, although with a different model ranking. Comparing our model fitness values with those obtained in previous studies using the FIT scoring function,[9] we noticed that the best model on the basis of FIT was a five-term model (FIT = 2.127) similar to our 13th model (CVFIT = 1.022) and that our overall best model (CVFIT = 1.236) was the 3rd best FIT model (FIT = 1.745). Unlike previous studies, the incorporation of q^2 in the fitness function kept the number of terms low, resulting in a number of new four-term models that emerged with higher q^2 values than the threshold values of 0.636 (ATCH4, ESDL3, PEAX_X, and LOGP) and 0.646 (MOFI_Y, MOL_WT, LOGP, and SUM_F) attained for a four-term model by MUSEUM and MoQSAR, respectively.

In a second phase, MGZ was allowed to also include quadratic terms with the consequent effect of doubling the initial parametric domain. A preliminary analysis of the derived models listed in Table 2 revealed that CVFIT indeed limited the number of nonlinear terms in the model. In fact, only two models contained more than one quadratic term. The first (LOGP, SUM_F, SURF_A∧2, and SUM_R∧2) is ranked in the ninth row (CVFIT = 1.168), whereas the second (LOGP, SUM_F, MOFI_Y∧2, and SUM_R∧2) is in the 15th row (CVFIT = 1.121). However, it is worth noting that the inclusion of quadratic terms led to an increase in the fitness value for the best overall QSAR model (CVFIT = 1.319), which was a four-term model closely related to

Table 1. Selwood Data Set, Best 20 QSARs Derived from the Pool of Linear Terms[a]

CVFIT	q^2	r^2	q_{SCR}^2	K	id1	id2	id3	id4	id5
1.236	0.647	0.721	−0.017	0.450	MOFI_Y	LOGP	SUM_F		
1.223	0.644	0.719	0.057	0.474	ESDL3	SURF_A	LOGP		
1.218	0.643	0.718	−0.073	0.439	MOFI_Z	LOGP	SUM_F		
1.100	0.665	0.745	0.070	0.405	DIPV_Y	MOFI_Y	LOGP	SUM_F	
1.072	0.660	0.755	−0.048	0.407	DIPV_X	MOFI_Y	LOGP	SUM_F	
1.068	0.659	0.741	0.030	0.482	ESDL3	MOFI_Y	LOGP	SUM_F	
1.053	0.656	0.733	0.143	0.398	ESDL3	NDSL6	SURF_A	LOGP	
1.044	0.654	0.740	−0.014	0.425	DIPMOM	MOFI_Y	LOGP	SUM_F	
1.042	0.653	0.740	0.104	0.417	DIPV_Y	ESDL3	SURF_A	LOGP	
1.041	0.653	0.737	0.075	0.406	DIPV_Y	MOFI_Z	LOGP	SUM_F	
1.033	0.651	0.735	0.035	0.480	ESDL3	MOFI_Z	LOGP	SUM_F	
1.032	0.604	0.702	0.053	0.475	ESDL3	MOFI_Y	LOGP		
1.022	0.696	0.791	0.414	0.495	ATCH7	DIPV_X	MOFI_Y	LOGP	SUM_F
1.019	0.648	0.741	0.089	0.426	ESDL3	ESDL9	MOFI_Y	LOGP	
1.017	0.648	0.745	−0.048	0.398	DIPV_X	MOFI_Z	LOGP	SUM_F	
1.015	0.601	0.697	0.058	0.462	ESDL3	MOFI_Z	LOGP		
1.012	0.647	0.735	0.194	0.567	ATCH7	MOFI_Y	LOGP	SUM_F	
1.011	0.646	0.740	0.021	0.415	MOFI_Y	MOL_WT	LOGP	SUM_F	
1.005	0.645	0.742	0.040	0.519	ESDL3	VDVWOL	MOL_WT	LOGP	
1.004	0.598	0.688	0.048	0.398	PEAX_X	LOGP	SUM_F		

[a] CVFIT, q^2, r^2, q_{SCR}^2, K, and id1−5 indicate, respectively, the fitness value, the squared correlation coefficient of predictions, the squared correlation coefficient, the best squared correlation coefficient of predictions over 50 runs using *y*-scrambled data, the K index value, and the descriptors occurring in each given QSAR model.

Table 2. Selwood Data Set, Best 20 QSARs Derived from the Pool of Linear and Quadratic Terms[a]

CVFIT	q^2	r^2	q_{SCR}^2	K	id1	id2	id3	id4	id5
1.319	0.740	0.799	−0.013	0.449	MOFI_Y	LOGP	SUM_F	SUM_R∧2	
1.291	0.692	0.751	−0.025	0.419	LOGP	SUM_F	MOFI_Y∧2		
1.281	0.690	0.749	−0.007	0.410	LOGP	SUM_F	MOFI_Z∧2		
1.236	0.647	0.721	−0.017	0.450	MOFI_Y	LOGP	SUM_F		
1.223	0.644	0.719	0.057	0.474	ESDL3	SURF_A	LOGP		
1.218	0.643	0.718	−0.073	0.439	MOFI_Z	LOGP	SUM_F		
1.197	0.762	0.825	−0.082	0.367	ESDL6	MOFI_Y	LOGP	SUM_F	SUM_R∧2
1.193	0.720	0.789	−0.030	0.446	MOFI_Z	LOGP	SUM_F	SUM_R∧2	
1.168	0.751	0.807	0.052	0.447	LOGP	SUM_F	SURF_A∧2	SUM_R∧2	
1.140	0.753	0.822	−0.033	0.394	ESDL4	MOFI_Y	LOGP	SUM_F	SUM_R∧2
1.132	0.709	0.783	−0.025	0.587	NDSL8	MOFI_Y	LOGP	NDSL9∧2	
1.126	0.751	0.819	−0.003	0.401	ESDL1	MOFI_Y	LOGP	SUM_F	SUM_R∧2
1.126	0.751	0.827	0.006	0.415	DIPV_X	MOFI_Y	LOGP	SUM_F	SUM_R∧2
1.124	0.751	0.819	−0.002	0.399	ESDL7	MOFI_Y	LOGP	SUM_F	SUM_R∧2
1.121	0.743	0.810	−0.013	0.428	LOGP	SUM_F	MOFI_Y∧2	SUM_R∧2	
1.120	0.750	0.816	0.065	0.433	ESDL10	SURF_A	LOGP	SUM_F	SUM_R∧2
1.115	0.749	0.815	0.047	0.407	ESDL2	MOFI_Y	LOGP	SUM_F	SUM_R∧2
1.100	0.665	0.745	0.070	0.405	DIPV_Y	MOFI_Y	LOGP	SUM_F	
1.099	0.657	0.729	0.009	0.427	LOGP	SUM_F	SURF_A∧2		
1.098	0.746	0.815	0.074	0.394	MOFI_Y	S8_1CZ	LOGP	SUM_F	SUM_R∧2

[a] For descriptions of CVFIT, q^2, r^2, q_{SCR}^2, K, and id1−5, see footnote of Table 1.

the best overall model developed, considering linear terms only. The addition of the square of SUM_R to the optimal combination of descriptors—LOGP, MOFI_Y, and SUM_F—yielded, in fact, a significant increase in model prediction power ($q^2 = 0.740$) with an acceptable degree of model complexity. The second-best model (CVFIT = 1.291) can be considered a polynomial extension of the overall best three-term linear model, in which the term MOFI_Y now occurrs as a square. A slight modification (i.e., the introduction of the square MOFI_Z to replace the square MOFI_Y) resulted in the third-best model (CVFIT = 1.281). The top three linear models previously found now ranked from the fourth to the sixth line, confirming the ability of CVFIT to select models that are both simpler and highly predictive. From the seventh row downward, a number of five-term models emerged. Among these, the best one can be considered as an expansion of the overall best model to which ESDL6 was added at the expense of model quality (CVFIT = 1.197). Finally, MGZ was run to also investigate the inclusion of interacting terms with linear and quadratic terms. As a result, the chemical space was further enlarged and the number of variables became 1484. Table 3 reveals that the top model was a four-term model, surprisingly different from those obtained so far. Such a model, comprising two linear descriptors (ATCH4 and DIPV_X) and two interacting terms (ATCH5LOGP and PEAX_YS8_1CX), displayed very impressive internal predictive statistics ($q^2 = 0.829$), leading to the best fitness values (CVFIT = 1.331) obtained so far for the Selwood data set. In addition to the randomization tests routinely performed for each model, the validity of this QSAR was further demonstrated by a more intensive randomization analysis consisting of 1000 *y*-scrambling runs. The best squared correlation of prediction ($q_{SCR}^2 = 0.250$) of the scrambled models was indeed far from the squared correlation of prediction ($q^2 = 0.829$) of the original model. Such evidence is confirmed by the histogram of q_{SCR}^2 versus

Table 3. Selwood Data Set, Best 20 QSARs Derived from the Pool of Linear, Quadratic, and Interacting Terms[a]

CVFIT	q^2	r^2	q_{SCR}^2	K	id1	id2	id3	id4	id5
1.331	0.829	0.871	0.250	0.227	ATCH4	DIPV_X	ATCH5LOGP	PEAX_YS8_1CX	
1.319	0.740	0.799	−0.013	0.449	MOFI_Y	LOGP	SUM_F	SUM_R∧2	
1.291	0.692	0.751	−0.025	0.419	LOGP	SUM_F	MOFI_Y∧2		
1.281	0.690	0.749	−0.007	0.410	LOGP	SUM_F	MOFI_Z∧2		
1.238	0.761	0.821	0.061	0.447	MOFI_Y	M_PNT	SUM_F	LOGPM_PNT	
1.236	0.647	0.721	−0.017	0.450	MOFI_Y	LOGP	SUM_F		
1.223	0.644	0.719	0.057	0.474	ESDL3	SURF_A	LOGP		
1.218	0.643	0.718	−0.073	0.439	MOFI_Z	LOGP	SUM_F		
1.197	0.762	0.825	−0.082	0.367	ESDL6	MOFI_Y	LOGP	SUM_F	SUM_R∧2
1.193	0.720	0.789	−0.030	0.446	MOFI_Z	LOGP	SUM_F	SUM_R∧2	
1.192	0.786	0.843	0.065	0.556	ATCH4	LOGP	MOFI_Z∧2	ATCH4ATCH5	
1.192	0.842	0.887	0.440	0.487	ATCH4	DIPV_Y	ATCH4∧2	MOFI_Z∧2	ATCH1LOGP
1.177	0.752	0.814	0.055	0.451	MOFI_Z	M_PNT	SUM_F	LOGPM_PNT	
1.176	0.752	0.815	0.100	0.399	VDVWOL	M_PNT	SUM_F	LOGPM_PNT	
1.175	0.752	0.840	0.202	0.217	ATCH4	DIPV_X	S8_1CX	ATCH5LOGP	
1.168	0.751	0.807	0.052	0.447	LOGP	SUM_F	SURF_A∧2	SUM_R∧2	
1.160	0.839	0.885	0.448	0.486	ATCH4	DIPV_Y	ATCH4∧2	MOFI_Y∧2	ATCH1LOGP
1.140	0.753	0.822	−0.033	0.394	ESDL4	MOFI_Y	LOGP	SUM_F	SUM_R∧2
1.132	0.709	0.783	−0.025	0.587	NDSL8	MOFI_Y	LOGP	NDSL9∧2	
1.132	0.777	0.838	0.060	0.559	ATCH4	LOGP	MOFI_Y∧2	ATCH4ATCH5	

[a] For descriptions of CVFIT, q^2, r^2, q_{SCR}^2, K, and id1−5, see footnote of Table 1.

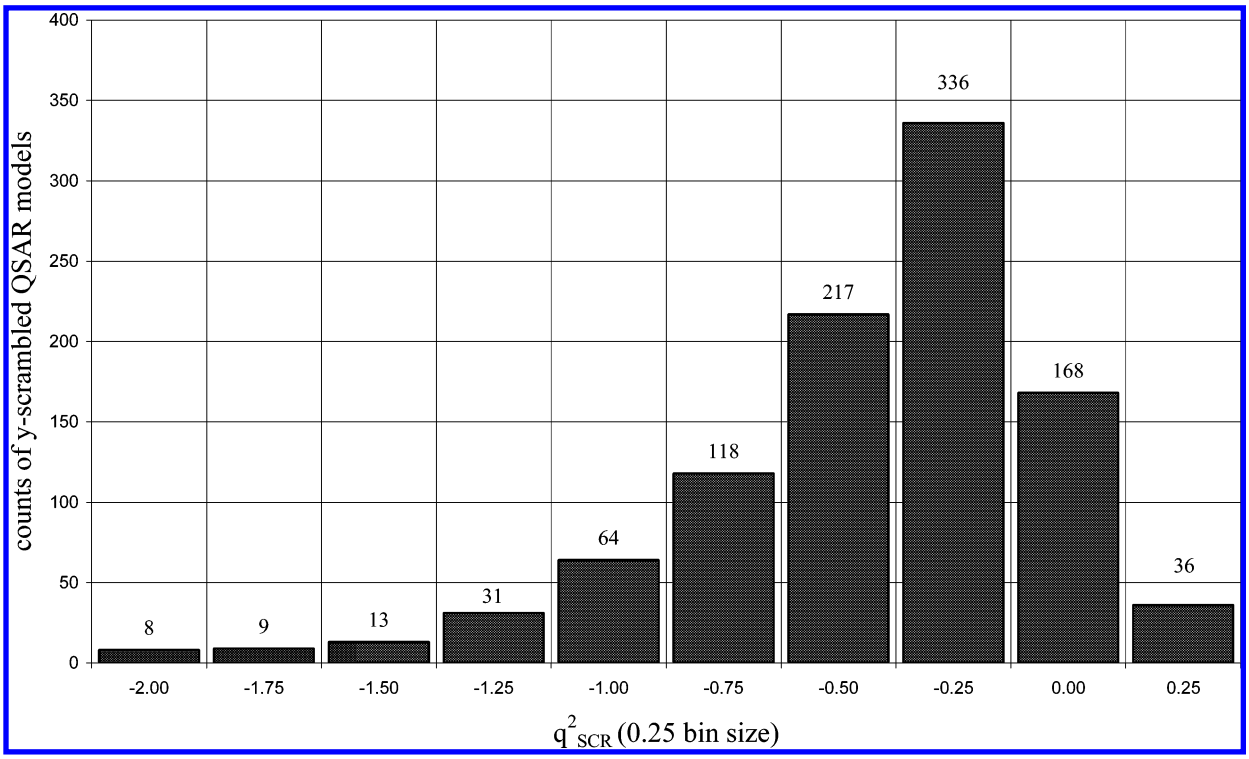

Figure 6. Histogram of q_{SCR}^2 vs frequency of occurrence over 1000 *y*-scrambling runs. The statistics were recalculated using the top Selwood QSAR model shown in Table 3 [−log(EC_{50}) = f(ATCH4, DIPV_X, ATCH5LOGP, PEAX_YS8_1CX); CVFIT = 1.331; q^2 = 0.829; r^2 = 0.871; q_{SCR}^2 = 0.250; K = 0.227; n = 31).

the frequency of occurrence (Figure 6) where one can observe that the great majority of scrambled models were associated with very low or even negative values of the squared correlation of internal prediction. Interestingly, the amount of correlation related to the independent variables was quite low (K = 0.222) and this indeed increases model confidence.

Models from the second to fourth rows of Table 3 remained unchanged compared to the best three models of Table 2 obtained from the combination of linear and quadratic terms. After this series of QSARs, another interesting and highly predictive four-term model occurred, which included one interacting term (LOGPM_PNT). The model presented a remarkable internal predictive power (q^2 = 0.761), ensuring a high fitness function value (CVFIT = 1.238). The successive interacting five-term model attained a significant rank (CVFIT = 1.192) because of the highest value of the squared correlation coefficient of internal predictions (q^2 = 0.842), a value never observed in the modeling of the Selwood data set. Such a model had an acceptably low correlation (K = 0.487) and included two linear terms (ATCH4 and DIPV_X), two quadratic terms (ATCH4∧2 and MOFI_Z∧2), and one interacting term (ATCH1LOGP). Scrolling the remaining models in Table 3, one may observe that the majority of them resulted from slight modifications of the interacting model discussed above.

Models constituted of interacting terms were not available for the Selwood data set; thus, it was not possible to make direct comparisons with previous approaches. However, some of our proposed quadratic QSAR models, scored using CVFIT as a fitness function, were found in a recent work based on a multiobjective optimization approach (the MoQSAR program).[21,36] Both methods were successful in identifying new QSAR models for the Selwood data set. In addition, the use of CVFIT, which is based on q^2 and interacting terms, permitted the finding of a consistent number of more highly predictive models not considered in the families of tradeoff models identified by the multiobjective optimization scheme. Interestingly, cross-product terms of LOGP with a number of different descriptors, such as ATCH5, M_PNT, and ATCH1, recurred many times, giving rise to a great variety of good predictive models ($0.752 < q^2 < 0.842$). This fact should not be so surprising if one considers the multicomposite nature of LOGP as a molecular descriptor.

The inclusion of quadratic terms in the top 20 models derived in the three MGZ experiments discussed above led to an increment of q^2 values. This result should be considered quite convenient, on the whole, although at the expense of model structure complexity. A major penalty was assigned to interacting terms, as formalized in the fitness function. Consequently, their occurrence was accepted only when the increment of q^2 was consistently high. A close analysis of the top four-term models can indeed help to better understand the role of CVFIT in selecting QSAR equations. The best four-term linear models are shown below:

$$-\log(EC_{50}) = -2.545 + 0.040\ \text{DIPV\_Y} - 0.000066\ \text{MOFI\_Y} + 0.572\ \text{LOGP} + 1.445\ \text{SUM\_F}$$

$$\text{CVFIT} = 1.100;\ q^2 = 0.665;\ r^2 = 0.745;\ q_{SCR}{}^2 = 0.070;\ K = 0.405;\ n = 31 \quad (9)$$

The inclusion of one quadratic term only determined a significant increase (+11%) of q^2 that was also worthy of a higher CVFIT value:

$$-\log(EC_{50}) = -3.903 - 0.000087\ \text{MOF\_Y} + 0.695\ \text{LOGP} + 2.088\ \text{SUM\_F} + 17.828\ \text{SUM\_R\^2}$$

$$\text{CVFIT} = 1.319;\ q^2 = 0.740;\ r^2 = 0.799;\ q_{SCR}{}^2 = -0.013;\ K = 0.449;\ n = 31 \quad (10)$$

The occurrence of one interacting term, LOGPM_PNT, yielded an even better value of q^2 over those of the linear and quadratic models (+14% and +3%, respectively). However, the increase of prediction power resulted in a CVFIT value higher than the value of the linear model but lower than the value of the quadratic model.

$$-\log(EC_{50}) = -0.234 - 0.000041\ \text{MOFI\_Y} - 0.0209\ \text{M\_PNT} + 1.96\ \text{SUM\_F} + 0.00376\ \text{LOGPM\_PNT}$$

$$\text{CVFIT} = 1.238;\ q^2 = 0.761;\ r^2 = 0.821;\ q_{SCR}{}^2 = 0.061;\ K = 0.447;\ n = 31 \quad (11)$$

As shown in Figure 3, a gain in prediction of at least 10% is generally considered as the minimum threshold for a reasonable chance to improve the CVFIT at the expense of model complexity. Indeed, some interacting terms occurred with such strikingly increased internal predictive statistics as to render them worthy of the top positions. Interestingly, such models were also provided with low K values. Nevertheless, this aspect is, to some extent, still difficult to address inasmuch as K represents a measure of the correlation between the molecular descriptors, yet no direct connection can be found between K and either the types of descriptors in the models or the values of q^2. As anticipated, model desirability remains a controversial problem depending on the nature/type of chemical descriptors used, with the result that the widely used LOGP (or π) is by far preferred to some obscure topological index. Moreover, linear terms are always favored since they are easier to handle than quadratic terms, with the latter being more easily accepted than interacting terms. Particular attention needs to be devoted to parabolic relationships, which frequently occur in classic QSAR studies and are, thus, highly desirable by medicinal chemists, along with bilinear relationships.[37] Well-known examples are models that relate the LOGP to the in vivo activity of drugs. In this regard, the Selwood data set yielded quite disappointing statistics ($q^2 = 0.196$, $r^2 = 0.331$) with a parabolic model based on LOGP. However, parabolas with well-shaped descending sides are not so common, and often, other parameters are needed to obtain better correlations. For this reason, the impact of an additional third descriptor was also evaluated setting MGZ to conduct an exhaustive search of the 1482 three-term models with LOGP and LOGP∧2 combined to one linear, quadratic, or interacting term. As observed in Figure 7, the best predictive statistics ($q^2 = 0.606$) were associated with a model having an interacting term. A total of 12 out of the top 20 models showed interacting terms. Only 4 of 20 models included linear terms (MOFI_Y, MOFI_Z, SURF_A, and PEAX_X), but these were provided with a lower power of prediction ($0.479 < q^2 < 0.504$). The poor performance of LOGP can, to some extent, be related to the fact that the compounds of the Selwood data set were tested for their in vitro antifilarial activity, whereas parabolic functions are generally more suitable to model in vivo processes. However, these results reinforce the notion that biological activity is often governed by forces that cannot be correctly and sufficiently described

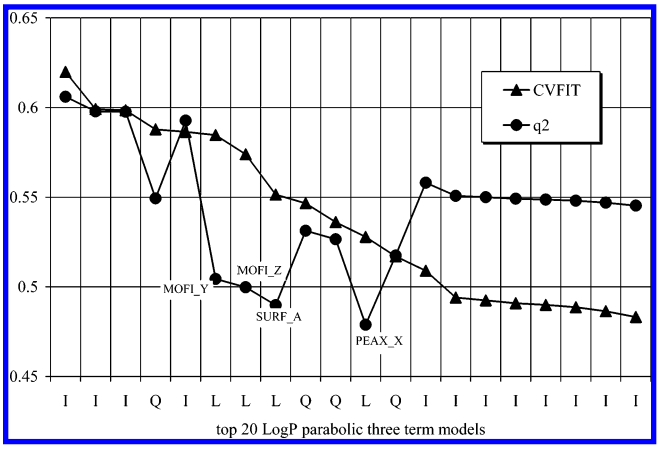

Figure 7. CVFIT and q^2 values of the top 20 best three-term QSARs (see text) in the modeling of the Selwood data set. LOGP and LOGP∧2 occurred in all models combined to one linear (L), quadratic (Q), or interacting (I) term.

Table 4. Nitrogen Data Set, Best 20 QSPRs Derived from the Pool of Linear, Quadratic, and Interacting Terms[a]

CVFIT	q^2	r^2	$r_{trn/ts}^2$	q_{SCR}^2	K	id1	id2	id3	id4	id5
1.924	0.703	0.726	0.701	0.038	0.502	MDE_14	GEOM_3	SCDH_2	xv0_4	dx2_10
1.905	0.714	0.735	0.660	0.053	0.592	MDE_14	GEOM_3	CTDH	xv0_4	dx2_10∧2
1.899	0.728	0.749	0.710	0.421	0.499	MDE_14	GEOM_3	xv0_4	dx2_10	SCDH_2phia_43
1.890	0.699	0.722	0.646	0.036	0.494	MDE_14	GEOM_3	SCDH_2	xv1_5	dx2_10
1.868	0.697	0.721	0.693	0.040	0.518	MDE_14	GEOM_3	SCDH_2	x2_3	xv0_4
1.867	0.724	0.748	0.703	0.044	0.519	MDE_14	GEOM_3	x2_3	xv0_4	SCDH_2phia_43
1.861	0.696	0.724	0.566	0.058	0.591	MDE_14	GEOM_3	CTDH	xv0_4	totop_39
1.859	0.696	0.721	0.649	0.057	0.565	MDE_14	GEOM_3	CTDH	xv1_5	dx2_10
1.853	0.695	0.720	0.699	0.059	0.576	MDE_14	GEOM_3	CTDH	xv0_4	dx2_10
1.851	0.695	0.719	0.652	0.036	0.511	MDE_14	GEOM_3	SCDH_2	x2_3	xv1_5
1.851	0.695	0.722	0.659	0.060	0.581	MDE_14	GEOM_3	CTDH	x2_3	xv1_5
1.837	0.693	0.720	0.693	0.062	0.590	MDE_14	GEOM_3	CTDH	x2_3	xv0_4
1.829	0.720	0.743	0.714	0.039	0.504	MDE_14	GEOM_3	xv0_4	dx2_10	SCDH_2ka2_36
1.828	0.720	0.741	0.706	0.041	0.490	MDE_14	GEOM_3	xv0_4	dx2_10	SCDH_2ka3_37
1.821	0.691	0.719	0.656	0.058	0.595	MDE_14	GEOM_3	CTDH	x2_3	xv2_6
1.820	0.691	0.717	0.646	0.033	0.520	MDE_14	GEOM_3	SCDH_2	x2_3	xv2_6
1.820	0.691	0.717	0.654	0.037	0.493	MDE_14	GEOM_3	SCDH_2	x2_3	dxv0_27
1.819	0.691	0.717	0.655	0.044	0.450	MDE_14	GEOM_3	PPSA_1	SCDH_2	x2_3
1.815	0.719	0.740	0.707	0.039	0.492	MDE_14	GEOM_3	xv0_4	dx2_10	SCDH_2k3_34
1.809	0.676	0.699	0.681	0.049	0.406	MDE_14	GEOM_3	SCDH_2	xv0_4	

[a] CVFIT, q^2, r^2, $r_{trn/ts}^2$, q_{SCR}^2, and id1−5 indicate the fitness value, the squared correlation of predictions, the squared correlation coefficient, the squared correlation coefficient between observed and experimental LOGS values of the test set, the best squared correlation of predictions over 50 runs using *y*-scrambled data, the *K* index value, and the descriptors occurring in each given QSAR model.

by means of conventional or theoretical descriptors. Within this still obscure picture, it might not be surprising to learn that terms other than linear, even without a definite physicochemical connotation, may furnish better predictive results. An example is the solvation equation of Abraham previously cited.

For the sake of completeness, other MGZ runs were also carried out (data not shown), setting the FIT instead of the CVFIT fitness function. The resulting models were generally provided with a greater structure complexity as they presented a higher number of terms and a significant occurrence of terms other than linear. Useful indications about model reliability and robustness were also obtained by runs in which 500 *y*-scrambled vectors were processed by MGZ to conduct an evolutionary search of the space of the corresponding 500 optimal random models. Among these, the best random model was provided with apparently acceptable statistics ($q^2 = 0.636$, $r^2 = 0.703$) and a reasonable structural complexity ($p = 3$, $c = 6$). However, its value of fitness (CVFIT = 0.705) was far from the corresponding fitness values derived from the unscrambled models described previously; this may be considered as indirect proof of the strength of CVFIT in selecting QSAR models.

Moreover, an exhaustive search of the linear and quadratic three-term models, with a total number of 192920 (i.e., 106 × 105 × 104/6) subsets for the Selwood data set, was also conducted to determine the effectiveness and accuracy of MGZ. Interestingly, the top six models found after this exhaustive search (data not shown) were the same as those found by MGZ during genetic sampling. An equivalent analysis based on a systematic search for three-term linear models, but limited to linear terms only, was already reported;[26] the same best linear three-term models were found by MGZ. Similar outcomes were also found for the 82160 (i.e., 80 × 78 × 77/6) linear and quadratic three-term models relative to the nitrogen data set, which will be examined later in this paper.

Application of MGZ **to the Nitrogen Data Set.** The nitrogen data set is composed of 176 organic compounds and 40 initial descriptors, resulting in a parametric space of 860 variables of which 780 are interacting terms. Such parameters described the main features of this large series of compounds whose water solubility, expressed as LOGS, was the target property. The objective of the present study was the modeling of this property by analyzing a 176 × 860 data matrix with MGZ. Given the relatively high number of candidate compounds, the data set was divided, as routinely done in these cases, by a blind rational partition (*y* sampling) into a training set and a test set comprising 141 and 35 compounds, respectively. The training set was subjected to MGZ to perform QSPR modeling, whereas the test set was used for external model validation.[38] The success of external (and widely accepted) validations is indeed a very advantageous prerequisite for designing novel compounds with a desired target feature. It is worth noting that the values of q^2 from the LOO procedure may sometimes furnish overly optimistic predictions, whereas the adoption of the leave-more-out approach for defining cancellation groups might lead to more reliable but, to some degree, not reproducible results. For this reason, the correlation coefficient $r_{trn/ts}^2$ between the experimental and predicted LOGS of the test set compounds is reported in Table 4 in addition to q^2 to better prove the statistical reliability of the models developed. At first glance, Table 4 reveals a tendency to select longer models. In fact, 19 of the 20 models showed five terms, which was the maximum model complexity allowed by us during the evolutionary search with MGZ. Unlike previous results from the Selwood data set, the behavior of CVFIT seemed more strictly dependent on the number of physical terms included in a given model. As shown in Figure 8, the CVFIT values were almost constantly higher for each given model size. This was indeed further proven by the fact that the q^2 values followed a quite similar trend. Interestingly, a close look at Figure 9 showed a significant increase of q^2 for some QSPR models that are reported with the number 2, 6, 13, 14, and 19. As noted in Table 4, such models were those including interacting terms which proved to play a definite role in the internal prediction of the data contained

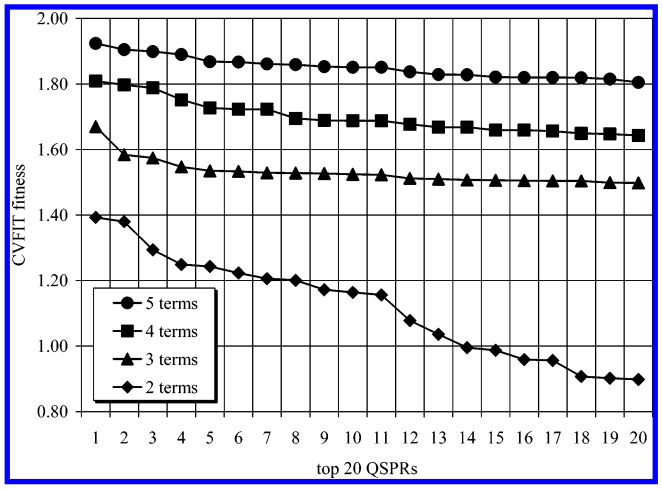

Figure 8. CVFIT fitness values of the top 20 best QSPR models for the given model size returned by MGZ in the modeling of the nitrogen training set.

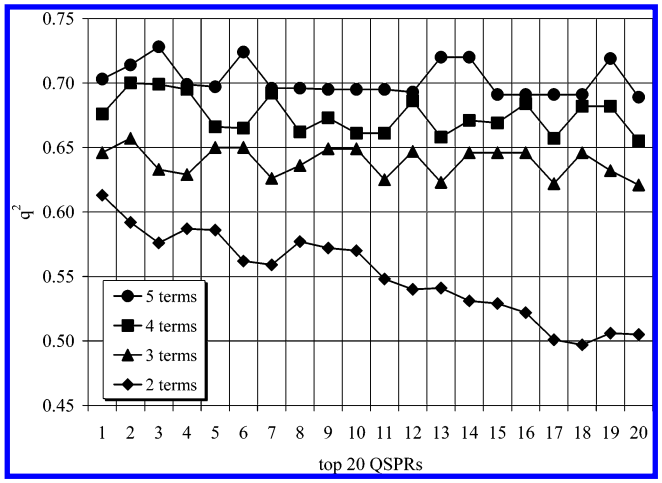

Figure 9. q^2 values of the top 20 best QSPR models for each given model size returned by MGZ in the modeling of the nitrogen training set.

in the training set. In fact, models with interacting terms were provided with q^2 values ranging from 0.719 to 0.728. Such a remarkable result was further supported by the ability of interacting term models to furnish excellent predictions for the 35 molecules of the test set as well. As reported in Table 4, their correlation coefficients $r_{trn/ts}^2$ ranged from 0.703 to 0.710. The model with the best external predictive ability is reported below:

$$\text{LOGS(mol/L)} = 1.510 + 0.166\ \text{MDE\_14} + 1.213\ \text{GEOM\_3} - 0.465\ \text{xv0\_4} - 0.705\ \text{dx2\_10} + 0.0364\ \text{SCDH\_2ka2\_36}$$

$$\text{CVFIT} = 1.829;\ q^2 = 0.720;\ r^2 = 0.743;\ r_{trn/ts}{}^2 = 0.714;\ q_{SCR}{}^2 = 0.039;\ K = 0.504;\ n = 141 \quad (12)$$

As observed in the plot of Figure 10, the calculated and predicted values of the training and test sets, computed from the equation above, were quite close to the experimental values. This finding indicated the importance of considering, in some cases, terms other than those linears in developing QSA(P)R models. The importance of interacting terms increases when dealing with QSPR studies that are aimed at modeling physical properties, such as solubility and permeability that may influence the ADME-Tox properties of drugs.[39] In the list of the top 20 models of Table 4, only one included a quadratic term (dx2_10). Though such a model was the second best (CVFIT = 1.905) and was also provided with a remarkable predictive statistic ($q^2 = 0.714$) for the training set, a similarly satisfactory result was not observed when forecasting the solubility of the organic compounds in the test set ($r_{trn/ts}^2 = 0.660$). This can be considered another unfortunate example of the Kubinyi paradox.[31] Linear models showed a uniform trend in both internal ($0.691 < q^2 < 0.703$) and external ($0.655 < r_{trn/ts}^2 < 0.701$) predictions.

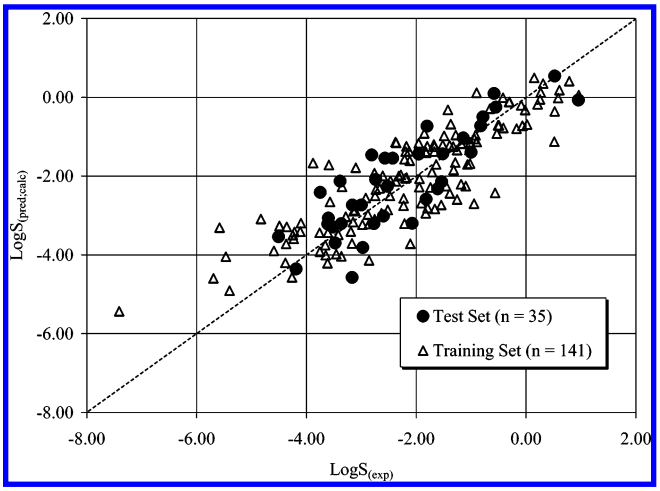

Figure 10. Plot of the experimental vs predicted (test set) and calculated (training set) solubility values of organic compounds of the nitrogen test and training sets, respectively. The line represents the case of a perfect correlation. The plot is associated with the equation (12) reported in the text.

Among the four-variable models, the best ones with linear (CVFIT = 1.809, $q^2 = 0.676$, and $r_{trn/ts}^2 = 0.681$), quadratic (CVFIT = 1.689, $q^2 = 0.673$, and $r_{trn/ts}^2 = 0.520$), and interacting (CVFIT = 1.797, $q^2 = 0.700$, and $r_{trn/ts}^2 = 0.690$) terms were ranked 19th, 22nd, and 29th, respectively. As for the three-variable models, the best linear (CVFIT = 1.670, $q^2 = 0.646$, and $r_{trn/ts}^2 = 0.642$), quadratic (CVFIT = 1.528, $q^2 = 0.636$, and $r_{trn/ts}^2 = 0.689$), and interacting (CVFIT = 1.584, $q^2 = 0.657$, and $r_{trn/ts}^2 = 0.635$) terms were ranked 33rd, 42nd, and 48th, respectively. Finally, among the two variable models, the best linear (CVFIT = 1.380, $q^2 = 0.592$, and $r_{trn/ts}^2 = 0.483$), quadratic (CVFIT = 1.078, $q^2 = 0.540$, and $r_{trn/ts}^2 = 0.318$), and interacting (CVFIT = 1.393, $q^2 = 0.613$, and $r_{trn/ts}^2 = 499$) terms were ranked 62nd, 72nd, and 61st, respectively. An analysis of model composition revealed that the combination between variables MDE_14 and xv0_4 resulted in the best models for each given model size. Finally, one may observe that the quadratic models were generally the worst in predicting the solubility values of the organic compounds belonging to the test set, whereas models with the interacting terms yielded the best performance.

CONLCUSIONS

In the present paper, we proposed a novel approach to deriving QSAR models that makes use of a newly implemented GA to explore a parametric space composed of not only linear but also quadratic and interacting terms. For the sake of completeness, one should say that, in the presence of their corresponding linear terms, quadratic terms are

indeed important to model several processes as they represent a way of detecting minima or maxima through curvature. Most chemical reactions attain an optimal yield under a particular set of experimental conditions such as pH, reactant concentrations, solvent, and so on, and almost all enzymatic reactions function in a similar manner. Also, the typical crossing of a drug through biological membranes can be conveniently described by parabolic or bilinear functions of LOGP. On the other hand, the influence of two factors on a given response is rarely independent. For example, in many chemical and biological processes, the optimal pH may vary at different temperatures. However, the use of interacting terms can cause a serious risk of overfitting and collinearity, especially in those cases where the number of molecular descriptors are much higher than the number of compounds. A symmetric correlation map could indeed furnish an example of this and could be used to detect those interacting terms resulting from highly correlated variables. As already anticipated, a real-life example of the importance of interacting terms is represented by the amended solvation energy relationship developed by Abraham. The addition of nonlinear terms by far extends the chemical space that, in our approach, was searched by evolutionary-independent populations and in which the balance between model predictivity and complexity was controlled via the use of penalties in a novel fitness function. Using multiple independent populations allowed an extensive exploration of the fitness landscape. The risk of trapping in local minima was further limited by genetic storms that increased the degree of diversity within each population. Of course, all the methods developed thus far led to a single *best* model or a given population of *best* models. The meaning of *best* is closely dependent on the criterion used to score models. Consequently, it is not surprising that the identification of a single *best* model in absolute terms is generally impossible in QSAR studies. In the present work, we used CVFIT as a fitness function, which we believe to be an improvement over the FIT function already proposed by Kubinyi.[9] Unlike FIT, CVFIT uses q^2 instead of r^2 for scoring QSAR/QSPR models, thus ensuring better chances of finding reliable models at the same computational cost since leverage values were used to calculate q^2. More specifically, CVFIT represents a steady balance between the ability to make internal predictions on one hand and the number of molecular descriptors in each QSAR/QSPR model on the other. Since the QSAR community might not be so enthusiastic in accepting the inclusion of quadratic terms and might be even more skeptic in considering interacting terms that, to some extent, could be difficult to understand from a physicochemical point of view, a weighting scheme was adopted to limit their frequency in QSAR models. Therefore, quadratic and interacting terms were penalized by being assigned double and triple weights whenever they occurred. This may delight medicinal chemists who are traditionally happier when making direct relationships between the response and a small number of widely accepted and interpretable physicochemical descriptors. In light of this, one should observe that the availability, at present, of a consistent number of easy-to-calculate molecular descriptors allowed an over-riding of the traditional basic philosophy that related the structural changes affecting the biological activities of a set of congeners with a very limited group of electronic, hydrophobic, and steric parameters. The demand for new descriptors, of even obscure meaning, is in fact principally due to the need for deriving more predictive models for drug design. Besides the reproduction of already well-known QSAR models, a series of new QSARs for the Selwood data set was developed through the incorporation of nonlinear terms. The risk of selecting redundant QSAR variables was monitored by calculating the *K* index correlation, which quantifies the global amount of collinearity within the selected molecular descriptors in a model. A series of randomization studies based on extensive *y* scrambling was also conducted. Finally, the efficacy of our approach was further demonstrated by validations conducted on an external set of compounds as was done for the nitrogen data set. In view of all the above-mentioned details, we believe that the most significant way to obtain a predictive QSAR is still to work with reliable and large data matrixes, hopefully composed of a high number of compounds and meaningful descriptors. The appropriate exploration of the physicochemical parametric domain may pave the way to a deeper understanding of the nature of a given biological response by deriving interpretable and predictive QSAR models. From this standpoint, one can say that, beginning with the first CoMFA studies,[5] 3D-QSAR techniques constitute the logical result of years of effort in this field.

ACKNOWLEDGMENT

We are grateful to Nathan McElroy and Peter Jurs for providing the nitrogen data set and to the Italian Ministry for Education, Universities and Research (Rome, Italy), for providing financial support.

REFERENCES AND NOTES

(1) Mazzatorta, P.; Benfenati, E.; Lorenzini, P.; Vighi, M. QSAR in Ecotoxicity: An Overview of Modern Classification Techiniques. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 105−112.

(2) Hansch, C.; Maloney, P. P.; Fujita, T.; Muir R. M. *Nature* **1962**, *194*, 178−180.

(3) Kubinyi, H. *QSAR: Hansch analysis and related approaches*; VCH: Weinheim, Germany, 1993.

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

(5) Cramer, R. D., III; Patterson, D. E.; Bunce, J. D. *J. Am. Chem. Soc.* **1988**, *110*, 5959−5967.

(6) Hawkins, D. M. The Problem of Overfitting. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1−12.

(7) Gillet, V. J.; Nicolotti, O. Evaluation of reactant-based and product-based approaches to the design of combinatorial libraries. *Perspect. Drug Discovery Des.* **2000**, *20*, 265−287.

(8) McFarland, J. W.; Guns, D. J. On Identifying Likely Determinants of Biological Activity in High Dimensional QSAR Problem. *Quant. Struct.-Act. Relat.* **1994**, *13*, 11−17.

(9) Kubinyi, H. Variable Selection in QSAR Studies. I. An Evolutionary Algorithm. *Quant. Struct.-Act. Relat.* **1994**, *13*, 285−294.

(10) Abraham, M. H.; Joelle, L. The Correlation and Prediction of the Solubility of Compounds in Water Using an Amended Solvation Energy Relationships. *J. Pharm. Sci.* **1999**, *88*, 868−880.

(11) Wold, S.; Trygg, J.; Berglund, A.; Antti, H. Some recent developments in PLS modeling. *Chemom. Intell. Lab. Syst.* **2001**, *58*, 131−150.

(12) Brenton, R. G. *Chemometrics Data Analysis for the Laboratory and Chemical Plant*; Wiley & Sons: New York, 2003.

(13) Selwood, D. L.; Livingstone, D. J.; Comley, J. C.; O'Dowd, B. A.; Hudson, A. T.; Jackson, P.; Jandu, K. S.; Rose, V. S.; Stables, J. N. Structure−Activity Relationships of Antifilarial Antimycin Analogues: A Multivariate Pattern Recognition Study. *J. Med. Chem.* **1990**, *33*, 136−142.

(14) Wikel, J. H.; Dow, E. R. The Use of Neural Networks for Variable Selection in QSAR. *Bioorg. Med. Chem. Lett.* **1993**, *3*, 645−651.

(15) Rogers, D. R.; Hopfinger, A. J. Application of Genetic Function Approximation to Quantitative Structure−Activity Relationships and Quantitative Structure−Property Relationships. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 854−866.
(16) Luke, B. T. Comparison of Different Data Set Screening Methods for Use in QSAR/QSPR Generation Studies. *THEOCHEM* **2000**, *507*, 229−238.
(17) So, S.; Karplus, M. Evolutionary Optimisation in Quantitative Structure−Activity Relationship: An Application of Genetic Neural Networks. *J. Med. Chem.* **1996**, *39*, 1521−1530.
(18) Waller, C. L.; Bradley, M. P. Development and Validation of a Novel Variable Selection Technique with Application to Multidimensional Quantitative Structure−Activity Relationship Studies. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 345−355.
(19) Cho, S. J.; Hermsmeier, M. A. Genetic Algorithm Guided Selection: Variable Selection and Subset Selection. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 927−936.
(20) Agrafitios, D. K.; Cedeno, W. Feature Selection for Structure−Activity Correlation Using Binary Particle Swarms. *J. Med. Chem.* **2002**, *45*, 1098−1107.
(21) Nicolotti, O.; Gillet, V. J.; Fleming, P.; Green, D. A Novel Approach to Deriving Accurate and Chemically Intuitive QSAR Models. *J. Med. Chem.* **2002**, *45*, 5069−5080.
(22) Nicolotti, O.; Altomare, C.; Pellegrini-Calace, M.; Carotti, A. Neuronal Nicotinic Acetylcholine Receptor Agonists: Pharmacophores, Evolutionary QSAR and 3D-QSAR Models. *Curr. Top. Med. Chem.* **2004**, *4*, 335−360.
(23) So, S. Quantitative Structure−Activity Relationships. In *Evolutionary Algorithms in Molecular Design*; Clark, D. E., Ed.; Wiley-VCH: Weinheim, Germany, 2000; pp 71−97.
(24) Spears, W. M.; De Jong, K. A. An analysis of Multi-Point Crossover. In *Foundations of Genetic Algorithms*; Rawlins, J. E., Ed.; Morgan Kaufmann Publishers Inc.: San Francisco, CA, 1991; pp 301−315.
(25) Goldberg, D. E. *The Design of Innovation: Lessons from and for Competent Genetic Algorithms*; Kluwer Academic Publishers: Norwell, MA, 2002.
(26) Kubinyi, H. Variable Selection in QSAR Studies. II. A Highly Efficient Combination of Systematic Search and Evolution. *Quant. Struct.-Act. Relat.* **1994**, *13*, 393−401.
(27) Wold, S.; Eriksson, L. In *Chemometric Methods in Molecular Design*; Waterbeemed, H., Ed.; VCH: Weinheim, Germany, 1995; pp 309−318.
(28) Tropsha, A.; Gramatica, P.; Gombar, V. K. The Importance of Being Earnest: Validation is the Absolute Essential for Successful Application and Interpretation of QSPR Model. *Quant. Struct.-Act. Relat.* **2003**, *22*, 69−77.
(29) Todeschini, R.; Consonni, V.; Maiocchi, A. The K correlation index: theory and development and its application in chemometrics. *Chemom. Intell. Lab. Syst.* **1999**, *46*, 13−29.
(30) Guha, R.; Jurs, P. Determining the Validity of a QSAR Model − A Classification Approach. *J. Chem. Inf. Comput. Sci.* **2005**, *45*, 65−73.
(31) van Drie, J. H. Pharmacophore discovery − lessons learned. *Curr. Pharm. Des.* **2003**, *9*, 1649−1664
(32) *MATLAB The Language of Technical Computing*, version 7.0; The MathWorks; Natick, MA, 2004.
(33) The openMosix Project Homepage. http://openmosix.sourceforge.net/.
(34) Molconn-Z is available from eduSoft, LC., P. O. Box 1811, Ashland, VA 23005.
(35) McElroy, N. R.; Jurs, P. C. Prediction of aqueous solubility of heteroatom-containing organic compounds from molecular structure. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1237−47.
(36) Nicolotti, O.; Gillet, V. J.; Fleming, P.; Green, D. A Multi-Objective Approach to Deriving QSAR Models. In *Designing Drug and Crop Protectants: Processes, Problems and Solutions*; Blackwell Publishing: Cambridge, MA, 2002; pp 264−267.
(37) Kubinyi, H. The Bilinear Model, a New Model for Nonlinear Dependence of Biological Activity on Hydrophobic Character. *J. Med. Chem.* **1977**, *20*, 625−629.
(38) Sheridan, P. S.; Fesuton, B. P.; Maiorov, V. N.; Kearsley, S. K. Similarity to Molecules in the Training Set is a Good Discriminator for Prediction Accuracy in QSAR. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1912−1928.
(39) Waterbeemd, H.; Gifford, E. ADMET in Silico Modelling: Towards Prediction Paradise? *Nature* **2003**, *2*, 192−204.

CI050293L