Stochastic versus Stepwise Strategies for Quantitative Structure−Activity Relationship Generation−How Much Effort May the Mining for Successful QSAR Models Take?[†]

Dragos Horvath,[*,‡] Fanny Bonachera,[‡] Vitaly Solov'ev,[§,||] Cédric Gaudin,[§,⊥] and Alexander Varnek[§]

UGSF-UMR 8576 CNRS/USTL, Université de Lille 1, Bât C9., 59650 Villeneuve d'Ascq, France,
Laboratoire d'Infochimie, UMR 7551 CNRS, Université Louis Pasteur, 4, rue B. Pascal, Strasbourg 67000,
France, Technologies Servier, 25-27 rue E. Vignat, 45000 Orléans, France, and Institute of
Physical Chemistry, Russian Academy of Sciences, Leninskiy prospect 31a, 119991 Moscow, Russia

Received October 31, 2006

Descriptor selection in QSAR typically relies on a set of upfront working hypotheses in order to boil down the initial descriptor set to a tractable size. Stepwise regression, computationally cheap and therefore widely used in spite of its potential caveats, is most aggressive in reducing the effectively explored problem space by adopting a greedy variable pick strategy. This work explores an antipodal approach, incarnated by an original Genetic Algorithm (GA)-based Stochastic QSAR Sampler (SQS) that favors unbiased model search over computational cost. Independent of a priori descriptor filtering and, most important, not limited to linear models only, it was benchmarked against the ISIDA Stepwise Regression (SR) tool. SQS was run under various premises, varying the training/validation set splitting scheme, the nonlinearity policy, and the used descriptors. With the considered three anti-HIV compound sets, repeated SQS runs generate sometimes poorly overlapping but nevertheless equally well validating model sets. Enabling SQS to apply nonlinear descriptor transformations increases the problem space: nevertheless, nonlinear models tend to be more robust validators. Model validation benchmarking showed SQS to match the performance of SR or outperform it in cases when the upfront simplifications of SR "backfire", even though the robust SR got trapped in local minima only once in six cases. Consensus models from large SQS model sets validate well−but not outstandingly better than SR consensus equations. SQS is thus a robust QSAR building tool according to standard validation tests against external sets of compounds (of same families as used for training), but many of its benefits/drawbacks may yet not be revealed by such tests. SQS results are a challenge to the traditional way to interpret and exploit QSAR: how to deal with thousands of well validating models, nonetheless providing potentially diverging applicability ranges and predicted values for external compounds. SR does not impose such burden on the user, but is "betting" on a single equation or a narrow consensus model to behave properly in virtual screening a sound strategy? By posing these questions, this article will hopefully act as an incentive for the long-haul studies needed to get them answered.

1. INTRODUCTION

Quantitative Structure−Activity Relationships[1−3] (QSAR) are empirical mathematical models (equations) returning an estimate of the activity level of a given molecule as a function of descriptors. Such models are obtained by calibration against a training set (TS) opposing activity values A_m of already tested molecules *m* (the explained variable) to their descriptor values (the explaining variables $D^1{}_m, D^2{}_m, ..., D^N{}_m$ where each element *i* of the vector stands for a certain structural or physicochemical aspect. Let *N* be the total number of available descriptors). Various data mining procedures may be used to complete the three main calibration steps:

I. Select a minimal number of explaining variables for actual use in model building. In the following, let *n* be the number of selected descriptors. Formally, the *N*-dimensional binary phase space associated with descriptor selection has each axis *i* associated with a binary variable $\delta_i = 0$ if the *i*th descriptor is be ignored and $\delta_i = 1$ if it enters the model.

II. Choose a functional form for the equation expected to optimally estimate activities as a function of the above-selected descriptors $Y_m = f(D^i{}_m;c^k)$−the simplest choice being a linear expression $Y_m = \sum_{i=1,N}^{\delta i>0} c^i D_m{}^i + c^0$.

III. Fit the coefficients c^k, $k = 1...n$, to obtain an equation minimizing the residual sum of squares $\Sigma(Y_m{-}A_m)^2$ between measured and predicted activities. If Y_m is a linear expression, c^k may be found by linear regression.

From a point of view of problem complexity, the descriptor selection step (I.) theoretically requires the exhaustive exploration of 2^N possible schemes. The function selection step (II.) offers virtually endless possibilities once nonlinear expressions are envisaged. Step (III.) may be quite time-consuming even with linear models: regression requires a $(n+1) \times (n+1)$ matrix inversion step scaling as $O(n^3)$ in terms of computational effort. Furthermore, successful completion of steps I.−III. is necessary but not sufficient: cross-validation[4,5] and external testing procedures have to be integrated into model buildup. However, typically,

[†] Dedicated to Professor Nenad Trinajstić on the occasion of his 70th birthday.

[*] Corresponding author phone +33.320.43.49.97; e-mail: dragos.horvath@univ-lille1.fr.

[‡] Université de Lille 1.

[§] Université Louis Pasteur.

[||] Technologies Servier.

[⊥] Russian Academy of Sciences.

10.1021/ci600476r CCC: $37.00
Published on Web 05/05/2007

relatively little time and computer effort is invested in QSAR buildup, for two main reasons:

(1) Traditionally, the QSAR building problem is declared "solved" once acceptable models were found—the exploration of the entire phase space of the problem is not, at this time, seen as a goal per se—although this may be arguably wrong.

(2) Problem space pruning strategies were successfully developed. These emerged under the pressure of limited computer power in the early years of QSAR, and were not systematically challenged when RAM capacity and CPU time ceased to represent bottlenecks. Descriptors are routinely discarded for being "intercorrelated" although this correlation may (a) not extend beyond the training set or (b) hide an independent variable that happens to be a small difference of two large terms. Descriptor selection is routinely performed on hand of linear models, with selected variables being subsequently used as input for neural networks. They are not necessarily the best choice for nonlinear models, but the high cost of simultaneous descriptor selection coupled to nonlinear modeling justifies the "short-cut" as far as some valid equation is being obtained.

The goal of this publication is to explore whether a more computationally intensive (days on a typical dual processor PC) approach to QSAR buildup, involving distributed computation, may provide an in-depth exploration of the problem space and provide a deeper understanding of QSAR methodology. In particular, it may allow to situate typical equations from stepwise procedures in this broader context.

The recent development of high-dimensional (N=10^3...10^4) Fuzzy Pharmacophore Triplet (FPT) descriptors[6] represented an additional incentive to develop a powerful selection tool. Furthermore, we extended the selection phase space to include additional states encoding the choice of nonlinear functional forms. To this purpose, the degrees of freedom were allowed to take values beyond the two binary options $\delta_i = 0/1$. While $\delta_i = 0$ maintains its original meaning "ignore descriptor i", $\delta_i = 1,2,...,N_T$ selects one of the predefined N_T nonlinear transformation functions and enters the transformed descriptor i, according to rule $T_{\delta i}(D^i)$, in the model. The final functional form of the QSAR model will thus be

$$Y_m = c^0 + \sum_{i=1,N}^{\delta i>0} c^i T_{\delta(i)}(D^i_m) \qquad (1)$$

e.g. a linear combination of nonlinearly transformed descriptors, which is equivalent to a single-layer neural network. This does not cover all the possible nonlinear expressions but benefits from the relative facility of fitting the c^i by linear regression while nevertheless allowing for nonlinear treatment (the coefficients appearing within the transformation functions $T(D)$ are constant; they may be partially optimized by entering a same functional form with different coefficient values as independent choices). Any rules $T(D)$ may in principle be envisaged—including $T_1(D) = D$ and $T_2(D) = D^2$ (the linear and square functions being now particular "nonlinear" transformations)—at the cost of an "exploding" phase space volume of $(N_T+1)^N$.

In response to this challenge, the Stochastic QSAR Sampler (SQS) has been developed, based on a distributed, hybrid genetic algorithm-based descriptor selection procedure, the Model Builder (MB). SQS is inspired from an analogous conformational sampling tool[7] designed to handle folding problems of miniproteins and docking problems with full site side-chain flexibility. The evolutionist approach central to the descriptor and functional form selection strategies has been hybridized with alternative optimization techniques and driven by a meta-optimization loop choosing the MB control parameters.

Three different data sets of molecules of known anti-HIV activities have been used for benchmarking the potential benefits of the more aggressive SQS strategy. They are relatively small (~100 molecules each) and are known to permit successful stepwise QSAR model buildup. For each set, five different splitting schemes into Training (TS_k) and Validation (VS_k) Sets, $k = 1...5$, were considered such that to ensure that every compound in a series is once being kept for validation. This should avoid any potential bias due to any peculiar choice of validation molecules. Five independent molecular descriptor sets were used in this work, including Fuzzy Pharmacophore Triplets (FPT), ISIDA[8] fragment descriptors, and ChemAxon[9] (CAX) descriptors (including two-point pharmacophore fingerprint, BCUT descriptors etc.). Among the 75 combinations of 3 data sets × 5 splitting schemes × 5 molecular descriptor choices, the following QSAR build-up simulations were performed (and duplicated, in order to assess reproducibility): (a) stochastic searches for linear (N_T=1) and polynomial (including the squared transformation, N_T=2) models based on FPT, ISIDA, and ChemAxon descriptors, (b) stochastic searches of fully nonlinear FPT-based models, and (c) deterministic model buildup of linear QSARs using Stepwise Regression (SR), with FPT and ISIDA descriptors

The stochastic procedure typically generates ~10^5 models, out of which diverse representatives are selected among the best cross-validating ones. These are then systematically confronted to their respective validation sets in order to obtain validation correlation coefficients $R^2{}_V$. Out of the selected models, not all validate successfully, and the ones that do so cannot be known a priori.[10] Successful validation of SQS representative models will thus be treated like a probabilistic event. Density distribution histograms monitoring the probability to discover a model with $R^2{}_V$ within a given range $r \leq R^2{}_V < r + \epsilon$ will be traced and compared for various simulations—using different sets of descriptors, different approaches to nonlinearity, etc. Average validation scores $<R^2{}_V>$ taken over all selected models can thus serve as success criterion related to the specific setup(s) of the considered simulation(s).

The first subject of this paper is the study of the intrinsic behavior of SQS:

1. Reproducibility: to what extent is the overall performance of the final models subject to fluctuations? Will models be found again when repeating a stochastic search?

2. SQS response with respect to a phase space volume increase: How does the introduction of nonlinearity impact model validation propensity? Is SQS sensitive to the size of the initial descriptor pool?

Next, linear SQS models will be compared to equations obtained by a deterministic QSAR build-up procedure: the Stepwise Regression tool (SR) of the ISIDA package.

3. Comparing individual SQS and SR models: how many individual models with $R^2{}_V$ above a given threshold were

Table 1. Considered Transformation Functions[a]

code	function	remark
0	$T_0(D) = 0$	null function (ignore *D*)
1	$T_1(D) = D$	identity function
2	$T_2(D) = D^2$	squared descriptor
3	$T_3(D) = \exp\{ - [(D - \langle D \rangle)/\sigma(D)]^2\}$	broad Gaussian (zexp)
4	$T_4(D) = \exp\{-3[(D - \langle D \rangle)/\sigma(D)]^2\}$	sharp Gaussian (zexp3)
5	$T_5(D) = \{1 + \exp[(D - \langle D \rangle)/\sigma(D)]\}^{-1}$	flat sigmoid (zsig)
6	$T_6(D) = \{1 + \exp[3(D - \langle D \rangle)/\sigma(D)]\}^{-1}$	steep sigmoid (zsig3)

[a] Nonlinear functions 3−6 require the input of the expectation values (averages $<D>$) and variances $\sigma(D)$ for the concerned descriptors calculated on hand of 2200 currently marketed drugs and reference compounds.

produced by every approach—and what fraction do these make out of the respective sets of models? Are there study cases in which SR fails to discover well validating models, while SQS succeeds? Reversely, can the fluctuation-prone SQS be seen to miss valid models picked up by the deterministic approach?

4. Comparing SQS and SR consensus models[11] (CM), reportedly a safer choice, in terms of validation propensity, than individual QSAR equations: does this still apply with extensive averaging of thousands of sampled parent equations?

This work is structured as follows: in the Methods section, a description of SQS and SR procedures precedes an introduction of the different descriptors and compound sets followed by a presentation of the statistical tools utilized to tackle the four above-mentioned key questions. The Results section will sequentially address these questions and lead to the Conclusion paragraph.

2. METHODS

2.1. The Stochastic QSAR Sampler (SQS). SQS is aimed to provide an effective, combined descriptor and nonlinear function selection procedure for the fitting of QSAR models according to eq 1. Table 1 enumerates the $N_T = 6$ currently implemented predefined transformation functions. The use of sigmoids and Gaussians requires the input of nominal expectation values $\langle D_i \rangle$ and variances $\sigma(D_i)$ for each descriptor, which in the present work were sought to be representative of the "universe" of drugs and druglike molecules. $\langle D_i \rangle$ and $\sigma(D_i)$ were calculated on hand of an independent set of 2200 drugs and druglike molecules and are constants throughout this work, e.g., independent of the processed training/validation sets.

2.1.1. Chromosome Definition and Prefiltering of Allowed Transformations. The Genetic Algorithm (GA)-driven Model Builder (MB) features chromosomes of size *N*, where every locus *i* codes $\delta(i)$—a natural number standing for the function $T_{\delta(i)}$ to be applied to D_i. Setting an upper threshold for acceptable δ values limits the allowed transformations to 1 (linear), 2 (polynomial, involving descriptors and/or their squares), or 6 (fully nonlinear regime). At input of training set molecules, the appropriateness of using a function T_k in conjunction with D_i is evaluated: the minimal variance of the vector of transformed descriptor values $T_k(D_i^m)$ over selected subsets including 2/3 of the training set molecules is required to exceed a user-defined threshold. If this test fails, then *k* will be deleted from the list of integer values chromosome locus *i* is allowed to adopt—the opposite would likely have caused a cross-validation failure anyway. Also, if the transformed descriptor is too strongly correlated with an already accepted transformation p—e.g. $T_k(D_i^m) \approx aT_p(D_i^m) + b$ for all *m*—then *k* will be rejected. For example, consider binary descriptor $D_i = 0/1$. A Gaussian transformation $T_3(D) = \exp[-(D-0.5)^2]$ makes no sense, since $T_3(0) = T_3(1)$. Also, using $T_1(D) = D$ is sufficient—replacing 0 by $T(0)$ and 1 by $T(1) \neq T(0)$ will have no other effect but adjusting of the associated linear coefficient.

No attempt is made at this stage to discard descriptor columns D_i that are (themselves or via their transformations) correlated to other potential descriptors.

2.1.2. Hybrid Genetic Algorithm. The following is a brief description of the Model Builder (MB). Its tunable operational and control parameters are given in capital italics.

Parallelization. A tunable number of NCONT parallel GA processes are started simultaneously on different CPUs, in order to simulate distinct "continents" in which Darwinian evolution may explore diverging paths. If a new fittest chromosome is found by a process, it is allowed to "migrate" to another of the parallel runs.

Population Initialization. After having built, for each locus *i* of the chromosome, the list of allowed functions $\delta(i)$ to be used in conjunction with D_i, an initial population of NPOP random chromosomes is used as departure point of Darwinian evolution. Due to the obvious interest in models with a minimal number of descriptors, an explicit upper threshold for the allowed number of variables, $\mu = 5$, is set at this point. μ is an adaptive parameter, gradually incremented during the evolution process (see later on). Two possible chromosome initialization schemes are alternatively considered: (a) Random pick: first, a random value for the effective number *n* of selected descriptors is drawn between 1 and μ. All the loci of the new chromosome are set to 0, then a number of *n* loci between 1 and *N* are randomly picked, and for each such locus *i* one of the available non-null options for $\delta(i)$ is randomly chosen. (b) Random ancestor crossovers: if a set of fit chromosomes was provided by previous runs, new ones may be produced by random crossovers of these "ancestors", while ensuring that (b1) the result is original (not among the ancestors) and (b2) it has less than μ selected descriptors.

A tunable probability PANCEST to apply ancestor crossover instead of random picking controls the stochastically alternating use of above-mentioned options.

Fitness Estimation. The fitness (see Figure 1) of chromosomes with more than μ selected descriptors is set to an arbitrary low level. Otherwise, it is calculated as follows: the matrix of the transformed descriptor values (*n* columns) is submitted to leave-a-third-out cross-validated MLR, returning a cross-validated correlation coefficient Q^2 and the predicted activity Y_m^{XV} for each molecule in the training set. Fitness is related to Q^2 amended by a model size-dependent penalty, with a tunable VARPENALTY parameter and $\bar{A}_{\mathrm{TS}}$ being the average TS compound activity.

$$\mathrm{Fit}(C) = Q^2(C) - n \times \mathrm{VARPENALTY}$$

$$Q^2(C) = 1 - \frac{\sum_{m \in \mathrm{TS}} [Y_m^{XV}(C) - A_m]^2}{\sum_{m \in \mathrm{TS}} [\bar{A}_{\mathrm{TS}} - A_m]^2} \qquad (2)$$

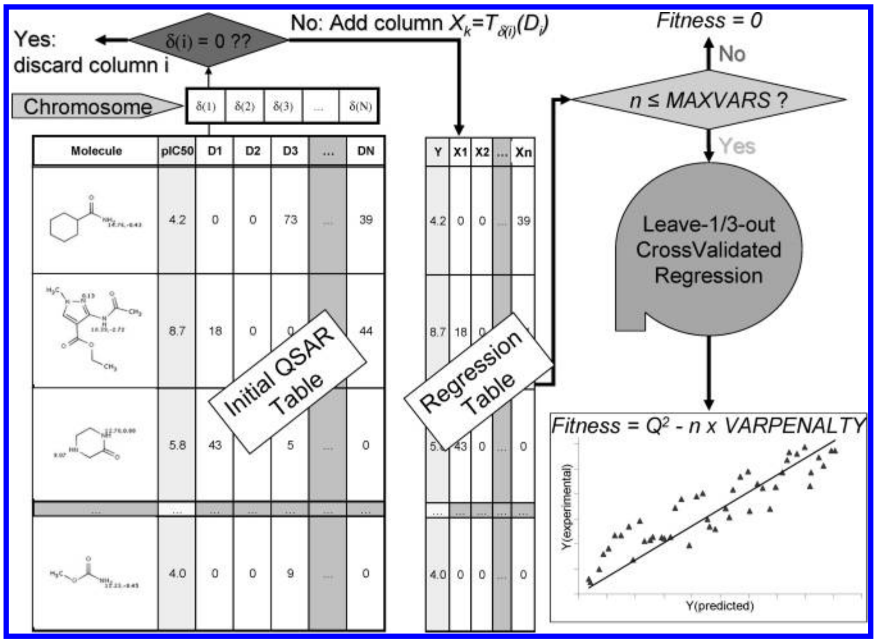

Figure 1. Evaluation of the fitness of a chromosome encoding a general nonlinear QSAR model. The MAXVARS control parameter is referred to as "μ" in the text.

Reproduction. Given a current generation of NPOP chromosomes, Darwinian evolution is simulated by generating a buffer population of offspring from crossovers and mutations. Parent chromosomes are randomly paired. Crossovers are performed at a randomly chosen positions, while checking that offspring differs from parents. Mutations, occurring at a tunable rate MUTFRQ, are random changes of the content of randomly picked loci i: $\delta(i) > 0$ mutate to 0, while $\delta(i) = 0$ will be toggled to one of the allowed transformer function codes.

Selection. The algorithm chooses either the *global* or the *intrafamily* scheme to pick the NPOP chromosomes making up the next generation out of the extended population emerging from reproduction.

In the *global* selection scheme, all the chromosomes of the extended population are sorted by decreasing fitness and then subjected to diversity filtering. The similarity of two chromosomes C and c is defined by the error pattern correlation score $S(C,c)$, where $Err_m(C) = Y_m(C) - A_m$ is the prediction error of the activity A calculated according to the model coded by chromosome C, for the molecule m:

$$S(C,c) = \frac{\sum_{m\in \mathrm{TS}} \mathrm{Err}_m(C)\mathrm{Err}_m(c)}{\sqrt{\sum_{m\in \mathrm{TS}} \mathrm{Err}_m{}^2(C) \times \sum_{m\in \mathrm{TS}} \mathrm{Err}_m{}^2(c)}} \tag{3}$$

If $S(C,c) >$ MAXSIM, the less fit of C and c will be discarded. This selection strategy is favoring convergence and is therefore applied only once every NGLOBAL generations.

If the *intrafamily* selection scheme is chosen, then offspring issued from crossovers will compete against its parents only: if the fittest child is fitter than the best of parents, then both parent chromosomes are replaced by the children. Mutants may only replace the "wild type" if they are fitter than the latter. This selection scheme favors population diversity and is therefore the default procedure.

The fittest NPOP chromosomes passing the similarity filter will form the next generation. If less than NPOP passed, random ones will be added to restore nominal population size.

Deterministic (*Lamarckian) Chromosome Optimization.* Occasional use of "Lamarckian" approaches (back-copying into the chromosome the knowledge about locally fitter configurations, obtained by exploring the neighborhood of a solution) has been shown[7] to enhance Darwinian evolution. The herein used local optimizer tool alternatively discards each of the n selected descriptors and reassesses the fitness of all the models of size $n-1$. If any of the latter is found to be fitter than the parent, it will take its place.

Population Refreshment Strategies. Failure to find a new fittest chromosome during NWAIT successive generations triggers a reinitialization ("apocalypse") of all but the fittest chromosome (elitist strategy). At this point, μ is incremented by 5 unless it already had reached its (tunable) maximum. Larger models are thus gradually being allowed for, as soon as no more progress can be obtained with fewer variables.

MB Termination. Eventually, if a series of NAPOC successive apocalypses did not lead to fitter solutions, a run stops.

Triplicate Runs and MB Success Score (*Meta-Fitness).* The SQS meta-optimization loop (Figure 2) pilots the search for the most appropriate MB operational parameters. The success of the triplicate MB run (3 successive runs with same operational parameters) defined by the quality of the solution set L produced. A large L, richer in fit models, means that the current choice of operational parameters has been judicious. The meta-fitness score, a measure of sampling success, is defined as

$$\mu\mathrm{Fitness} = \log \sum_{c\in L} \exp[-\beta \times \mathrm{Fit}(c)] \tag{4}$$

μFitness is an implicit function of the operational parameters that leads to sampling of the local solution set L. The temperature factor was empirically set to $\beta = 30$.

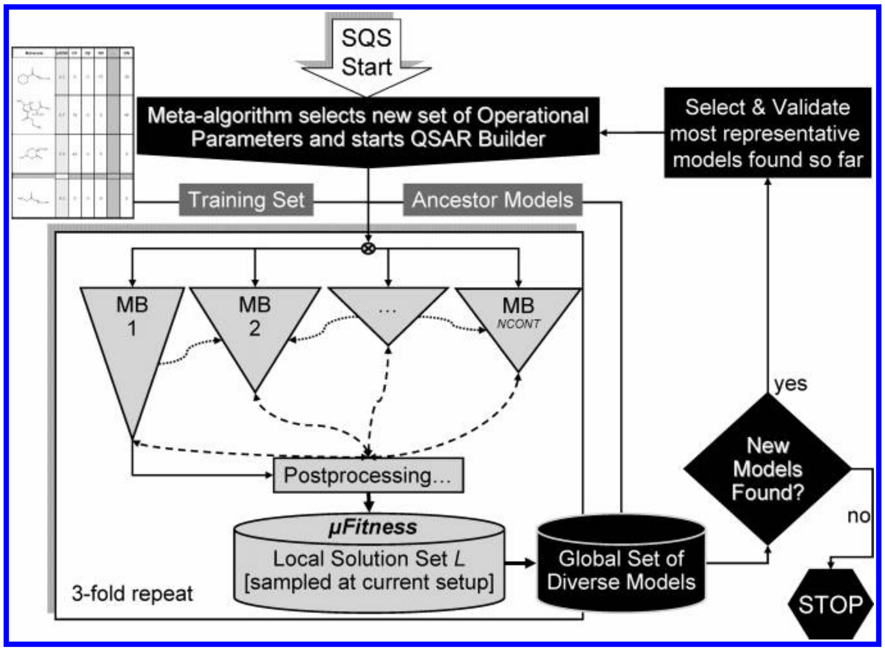
SQS Start
Meta-algorithm selects new set of Operational Parameters and starts QSAR Builder
Select & Validate most representative models found so far
Training Set
Ancestor Models
MB 1
MB 2
...
MB NCONT
yes
New Models Found?
no
Postprocessing...
μFitness
Local Solution Set L [sampled at current setup]
Global Set of Diverse Models
3-fold repeat
STOP

Figure 2. Overview of the stochastic QSAR build-up procedure, featuring the distributed hybrid GA-based QSAR builder, driven by a meta-optimization loop in search of the most appropriate operational parameter setup.

SQS Flowchart. After completion of a triplicate MB run, the local solution set L is merged with the global database of visited models, which is sorted by fitness. Redundant models at MAXSIM > 0.9 are rejected. Meta-optimization continues, according to a basic GA scheme, as long as the latest triplicate runs continued to add new valuable solutions to the global database.

2.1.3. SQS Run Validation and Postprocessing. After each triplicate run, a current set of most representative models is extracted from the up-to-date global set: chromosomes with Q^2 within 0.2 units of the current Q^2 maximum are classified with respect to the number of selected descriptors n, and, for each of these size classes, the top 10 fittest representatives are selected. Each selected model C is subjected to external validation in order to calculate their validation correlation score, on hand of the predicted activity values $Y_m(C)$ for all the m of the validation set, where $\bar{A}_{\mathrm{VS}}$ represents the average activity of VS compounds (for models with prediction errors exceeding the ones of the "null" model $Y_m = \bar{A}_{\mathrm{VS}}$, negative $R^2{}_V$ values are truncated to 0):

$$R^2{}_V(C) = \max\left\{0,1 - \frac{\sum_{m\in\mathrm{VS}} [Y_m(C) - A_m]^2}{\sum_{m\in\mathrm{VS}} [A_m - \bar{A}_{\mathrm{VS}}]^2}\right\} \qquad (5)$$

$R^2{}_V$ truncation is required because very low negative values which, per se, have no quantitative significance beyond the one of a flag for validation failure will skew the average $<R^2{}_V>$ values used to compare the relative proficiency of various sampling strategies (see further on). For example, a QSAR method leading to 4 models of $R^2{}_V = 0.9$ and one of $R^2{}_V = -5.0$ is obviously to be preferred to one producing 5 models of $R^2{}_V = 0.0$, although averages of untruncated $R^2{}_V$ suggest the contrary.

The set of representative models extracted after each triplicate MB run is steadily evolving, as newly visited configurations are added to the global pool of found models. All the models selected as "representative" at any instance of the meta-optimization loop will be considered in the final analysis, if their Q^2 lies within 0.2 units of the latest, absolute Q^2 maximum. Any SQS simulation, defined by (1) the used compound set **C** and the considered TS/VS splitting scheme **S** = 1..5, (2) the used descriptors **D**, and (3) the nonlinearity policy **N** = "L"(linear), "P"(polynomial), or "N"(nonlinear), will be thus associated with its set **R(C,S,D,N)** of representative model chromosomes. However, for most practical purposes, it makes sense to merge the representative sets of the 5 SQS runs corresponding to different splitting schemes. **R(C,D,N),** the merger of all **R(C,S,D,N)** sets, will serve as a representative sample of built models under given "premises" (e.g., compound set, descriptors and nonlinearity policy).

All the SQS simulations were repeated once—let the corresponding duplicate sets of models be denoted **R′**. SQS reproducibility is assessed by comparing the "twin" sets **R** and **R′**. In order to compare results of SQS run under different premises, the merged sets **R*** = **R**U**R′** from both twin runs will be used.

2.1.4. SQS Consensus Models. SQS CM were built, for each splitting scheme **S**, as the plain averages of the equations in the corresponding merged sets **R***. CM coefficients are set to the average of coefficients in each individual equation from **R*(C,S,D,N)**. Only CM concerning linear ISIDA and FPT based models will be discussed here.

2.2. Stepwise Model Build-Up Procedure. The QSPR/MLRA and Variable Selection Suite (VSS) modules of the ISIDA software have been used to build QSAR models using multilinear regression analysis.

2.2.1. Variable Selection. Two different strategies of stepwise variables selection have been used. The first (SR-1) is based on three steps procedure involving filtering, forward stepwise, and backward stepwise stages.

(1) Filtering Stage. The program eliminates variables D_i which have small correlation coefficient with the activity,

$R_{Y,i} < R^0_{Y,i}$ and those highly correlated with other variables D_j ($R_{i,j} > R^0_{i,j}$), which were already selected for the model. In this work, the values $R^0_{Y,i} = 0.001$ and $R^0_{i,j} = 0.99$ were used. Fragments always occurring in the same combination in each compound of the training set (concatenated fragments) are treated as one extended fragment. "Rare" fragments (i.e., found in less than m molecules, here $m \geq 2$) were excluded from the training set.

(2) Forward Stepwise Preselection Stage. This is an iterative procedure, on each step of which the program selects two variables D_i and D_j maximizing the correlation coefficient $R_{Y,ij} = (R^2_{Y,i} + R^2_{Y,j} - 2R_{Y,i}R_{Y,j}R_{ij})/(1 - R^2_{ij})$ between D_i and D_j and dependent variable Y. At the first step ($p = 1$), the modeled activity for each compound is taken as its experimental one $Y_1 = A_m$. At each next step p, as the activity values Y_p were used residuals $Y_p = Y_{p-1} - Y_{\text{calc}}$, where $Y_{\text{calc}} = c_0 + c_iD_i + c_jD_j$ is calculated activity by the two-variables model with selected variables D_i and D_j. This loop is repeated until the number of variables k reaches a user-defined value; in this work, k is set to half of the molecule number in the full set.

(3) Backward Stepwise Selection Stage. The final selection is performed using backward stepwise variable selection procedure based on the t statistic criterion.[12−14] Here, the program eliminates the variables with low $t_i = c_i/s_i$ values, where s_i is standard deviation for the coefficient c_i at the ith variable in the model. First, the program selects the variable with the smallest $t < t_0$, then it performs a new fitting excluding that variable. This procedure is repeated until $t \geq t_0$ for selected variables or if the number of variables reaches the user's defined value. Here, t_0, the tabulated value of Student's criterion is a function of the number of data points, the number of variables, and the significance level.

Selected descriptors are used by ISIDA to build the multilinear correlation equation $Y = c_0 + \Sigma c_iD_i$. The Singular Value Decomposition method[15] (SVD) is used to fit the coefficients. The most robust models are selected at the training stage according to statistical criteria.

The SR-1 calculations were initiated from the initial pools of ISIDA and FPT descriptors. Forty-six (CU), 42 (HEPT), and 36 (TIBO) descriptors of both types have been preselected by the forward stepwise preselection procedure. Initial pools of descriptors for CU, HEPT, and TIBO contained, respectively, 1586, 1114, and 576 ISIDA fragments and 1328, 1347, and 1201 of FPT. Further reduction of the number of variables was performed using backward stepwise variable selection procedure based on t statistic criterion allowing building the QSPR models containing desirable number of variables. Eventually, 5−7 models were selected for each splitting scheme **S**. These models included 10-33, 28-39, and 5-27 ISIDA descriptors and 16-23, 14-24, and 10-23 FPT descriptors for CU, HEPT, and TIBO training sets, respectively. The second strategy of variables selection (SR-2) involves splitting of the initial pool of fragment descriptors into subsets, followed by filtering and backward stepwise selection stages. ISIDA generates 319 subsets of fragment descriptors corresponding either to sequences of particular length from $n_{\min}$ to $n_{\max}$ atoms containing atoms and bonds, atoms only or bonds only, or to augmented atoms. Three or four models per subset possessing reasonable statistical criteria at the training stage have been selected, for each splitting scheme **S**, from the SR-2 calculations. They involve from 9 to 21 descriptors representing the sequences of atoms, bonds, and atoms/bonds containing up to 8 atoms. SR models were subjected to the same external validation procedure using the respective validation sets associated with the considered splitting schemes **S**, with respect to which the validation correlation coefficient R^2_V was calculated according to eq 5.

2.2.2. SR Consensus Models. The ISIDA software may generate CM combining the information issued from several individual models obtained in SR-2 calculations.[16,17] The idea is to use simultaneously a set of best models, for which the values of cross-validation correlation coefficient $Q^2 \geq Q^2_{\text{lim}}$, where Q^2_{lim} is a user-defined threshold. Thus, for each compound from the test set, the program computes the activity as an arithmetic mean of values obtained with these models, excluding those leading to outlying values according to Grubbs's test.[18] Our experience shows[10,17,19] that such an ensemble modeling allows one to smooth inaccuracies of individual models. Three CM were prepared for CU, HEPT, and TIBO derivatives ($Q^2_{\text{lim}} = 0.7$, 0.85, and 0.65, respectively) based on 11-32, 11-29, and 22-60 individual models for training sets of splitting schemes **S**.

2.3. Compound Sets. The QSAR modeling has been performed on different types of anti-HIV activity for three families of compounds: cyclic ureas (CU), 1-[2-hydroxyethoxy)methyl]-6-(phenylthio)thymines (HEPT), and tetrahydroimidazobenzodiazepinones (TIBO).

CU Derivatives. The HIV-1 protease inhibition constants K_i for 118 compounds were selected from refs 20−22. Activities $A = \log(1/K_i)$ vary between 5 and 11.

HEPT Derivatives. Effective concentrations of compounds required to achieve 50% protection of MT-4 cells against the cytopathic effect of HIV-1 (EC_{50}) for 93 molecules have been collected.[23−30] Modeling was performed with respect to the activities $A = \log(1/EC_{50})$, which vary from 3.9 to 9.2.

TIBO Derivatives. The concentration required to 50% the HIV-1 reverse transcriptase enzyme inhibition (IC_{50}) for 84 TIBO derivatives has been critically selected.[31−33] Modeling was performed for the $\log(1/IC_{50})$ values which vary from 3.1 to 8.5.

The data sets (2D structures and activities) are also used in ref 34.

2.4. Molecular Descriptors. Three main categories of descriptors have been considered in this work: (1) **ISIDA:** Substructural molecular fragments including sequences of atoms and bonds (from 2 to 15 atoms per sequence) as well as atoms with their closest environment ("augmented atoms") were generated by ISIDA.[8,12,14] (2) **CAX**: Two-point topological pharmacophore fingerprints and various other descriptors provided by ChemAxon's *generatemd* utility (including calculated logP, logD, the Topological Polar Surface Area, and four BCUT descriptors). All these were obtained with default ChemAxon setups. (3) **FPT**: Three-point topological fuzzy pharmacophore triplets were generated according to setup number one discussed in the above-cited publication.

2.5. Specific Statistical Approaches Used To Compare Model Build-Up Results. The steps undertaken to specifically address the key questions formulated in the Introduction will be briefly underlined in the following:

2.5.1. Reproducibility of SQS Simulations. Two different aspects of SQS reproducibility were addressed:

1. Were the model equations rediscovered when repeating the procedure? A Rediscovery Rate (RR) was estimated by reporting the number of equations visited by both instances of the SQS run to the total number of visited models (“visited” refers to any chromosome for which a trace has been kept, irrespective of fitness and validation score).

2. Do repeated, “twin” SQS runs generate models with similar validation behavior? In this sense, the validation behavior (R^2_V) of models from **R** and **R′** is considered to be a random variable of unknown distribution law. Assuming that **R** and **R′** contain respectively $N_{\mathbf{R}}$ and $N_{\mathbf{R'}}$ models distributed around average values $A(\mathbf{R}) = \langle R^2_V \rangle_{\mathbf{R}}$ and $A'(\mathbf{R'}) = \langle R^2_V \rangle_{\mathbf{R'}}$ with variances $s(\mathbf{R})$ and $s'(\mathbf{R'})$ respectively, a statistical criterion t rejecting the “null hypothesis” that the two averages A and A' are identical (all the differences being attributable to sampling fluctuations) can be calculated:

$$t = \frac{|A(R) - A'(R')|}{\sqrt{\frac{s^2(R)}{N_R} + \frac{s'^2(R')}{N_{R'}}}} \qquad (6)$$

Equation 6 is applicable for arbitrary distribution laws,[35] provided that the numbers of instances N_R and $N_{R'}$ of the random variables are much larger than 30. A large t value signals low SQS reproducibility (repeats lead to model sets of significantly different validation propensities).

2.5.2. Comparison of Model Sets Issued from Different SQS Runs. Several of the key questions addressed in introduction require a methodology to compare the validation propensity distributions of models obtained under different premises. In order to decide whether the differences are significant, the typical stochastic shift observed in respective the twin runs will serve as baseline.

Let $\mathbf{R}_P$ and $\mathbf{R}_Q$ be representative model sets obtained under different premises. For example, $\mathbf{R}_P = \mathbf{R}(\mathbf{C},ISIDA,L)$ **U** $\mathbf{R}(\mathbf{C},ISIDA,P)$ and $\mathbf{R}_Q = \mathbf{R}(\mathbf{C},CAX,L)$ **U** $\mathbf{R}(\mathbf{C},CAX,P)$ to compare linear and polynomial models built on hand of set **C** with ISIDA and CAX descriptors, respectively, etc. The merged sets from the twin SQS calculations, $\mathbf{R^*}_P = \mathbf{R}_P\mathrm{U}\mathbf{R'}_P$ and $\mathbf{R^*}_Q = \mathbf{R}_Q\mathrm{U}\mathbf{R'}_Q$ contain models with average validation propensities $A(\mathbf{R^*}_P) = \langle R^2_V \rangle_{\mathbf{R^*}P}$ and $A(\mathbf{R^*}_Q) = \langle R^2_V \rangle_{\mathbf{R^*}Q}$, respectively. According to eq 6 (under consideration of the variances of R^2_V within the two sets $\mathbf{R^*}_P$ and $\mathbf{R^*}_Q$), let t_{P-Q} be the statistical criterion rejecting the hypothesis that the observed shift

$$\Delta_{P-Q} = |A(R^*_P) - A(R^*_Q)| \qquad (7)$$

has no statistical significance (is due to normal fluctuations):

$$t_{P-Q} = \frac{\Delta_{P-Q}}{\sqrt{\frac{s^2(R^*_P)}{N_{R^*_P}} + \frac{s^2(R^*_Q)}{N_{R^*_Q}}}} \qquad (7a)$$

However, only part of Δ_{P-Q} may be ascribed to the differential impact of working premises P and Q on the model building process, as this magnitude is also influenced by the stochastic noise affecting the calculated averages. In the following, it will be assumed that the noise in Δ_{P-Q} can be related to the t factor of the least reproducible simulations, $max(t_{P-P'},t_{Q-Q'})$, where $t_{P-P'}$ and $t_{Q-Q'}$ are measures of sampling reproducibility under premises P and Q, respectively. Under this conservative assumption, the minimal guaranteed shift δ_{P-Q} that can be directly attributed to the change in sampling premises is

$$\delta_{P-Q} = \Delta_{P-Q} - \max(t_{P-P'},t_{Q-Q'}) \times \sqrt{\frac{s^2(R^*_P)}{N_{R^*_P}} + \frac{s^2(R^*_Q)}{N_{R^*_Q}}} \qquad (8)$$

The second term in eq 8 evaluates the random shift attributable to imperfect sampling (s and N representing the validation score variances and the number of models in the merged sets). The above equation amounts to a relative interpretation of t_{P-Q}: the observed shift of average validation propensity Δ_{P-Q} is considered relevant if $t_{P-Q} > \max(t_{P-P'},t_{Q-Q'})$. If $\delta_{PQ} > 0$, one of the premises P and Q is significantly more appropriate for QSAR modeling than the other. Density distribution histograms monitoring the percentage of models in **R*** having R^2_V within each of the ten bins spanning the range 0−1 may provide a detailed illustration of observed differences.

2.5.3. Comparison of the Stepwise and SQS Model Building Strategies. Stepwise regression (SR) leads to a limited set of models that cannot be considered as randomly spread in the problem phase space. Therefore, the statistical treatment envisaged to compare various SQS runs cannot be extended to include models built by the stepwise approach. For this reason, we compared the relative numbers of successfully validating SR and SQS models for which $R^2_V \geq R^2_{V,\mathrm{lim}}$, where $R^2_{V,\mathrm{lim}}$ has been allowed to vary between 0.6 and 0.8.

3. RESULTS AND DISCUSSIONS

3.1. Reproducibility of SQS Runs. The ∼2 × 10^4 locally fit and diverse chromosomes typically encountered in the global database after SQS completion represent only about 0.1% of the estimated 2 × 10^7 effectively visited states (from typically 600 MB runs—featuring ∼10 meta-generations × 10 triplicate runs/meta-generation, involving two or three parallel executions of the GA sampler—each processing ∼300 generations, with ∼100 chromosomes/generation). 99.6% of chromosomes were therefore either not fit enough (not even according to early, less constraining fitness standards) or redundant—it is thus irrelevant to include them in the calculation of Rediscovery Rates (RR). The probability that any model chromosome visited and considered for storage by a SQS run will be again encountered upon restarting a sampling process, under identical premises, is of 6% at best and may be as low as 0.03%. A SQS run visits a very limited fraction of the phase space volume of the problem, and the larger the volume, the less the chances to revisit any given chromosome. This is observed upon increasing of the descriptor set size N: for linear models, at $N = 217$, the CAX descriptor space offers by far the smallest sampling volume and scores the highest RRs (between 4% for CU and 6% for TIBO compounds), followed by ISIDA (1%, CU to 3.4%, TIBO) and eventually by FPT (0.5%, CU to 2.7%, TIBO). Phase space volume increases when enabling nonlinearity (with FPT) also triggers a RR decrease

Table 2. Average Validation Score Differences Observed When Repeating the SQS Simulations under Specified Premises[a]

set: CU	ISIDA	CAX	FPT
L	0.017 (3.80)	0.022 (5.79)	**0.053 (13.37)**
P	**0.068 (14.83)**	0.012 (3.75)	0.008 (2.43)
N	N/A	N/A	0.012 (4.68)
set: HEPT	ISIDA	CAX	FPT
L	0.016 (2.52)	*0.034 (7.19)*	0.008 (1.60)
P	0.024 (5.09)	0.021 (3.95)	0.022 (4.60)
N	N/A	N/A	*0.032 (7.28)*
set: TIBO	ISIDA	CAX	FPT
L	0.006 (1.20)	0.016 (2.31)	**0.062 (8.71)**
P	0.015 (3.08)	0.021 (2.59)	0.007 (1.00)
N	N/A	N/A	***0.107 (16.89)***

[a] Compound set − associated with individual tables, nonlinearity policy − in rows, used descriptors − in columns). Calculated t factors are given in parentheses. Coding concerns the size of observed shifts: $|A(\mathbf{R}) - A'(\mathbf{R}')| < 0.025$: no fill; <0.05: italics; <0.075: boldface; >0.1: italics and boldface.

(0.2%, CU to 2.2%, TIBO for polynomial, but only 0.04%, all sets, for fully nonlinear models).

SQS runs therefore appear to barely "scratch the surface" of the phase space to explore to the point of questioning their usefulness altogether. It is therefore not surprising that the high t values reported for the large majority of SQS premises monitored in Table 2 clearly support the hypotheses that observed average validation propensity shifts cannot be ascribed to fluctuations expectable on behalf of two random walks exploring a common phase space zone. Observed shifts are deemed relevant—for example, at $t = 1.64$ there is only a 10% chance that observed shifts are due to fluctuations, while at $t = 3.29$ this chance drops to 0.1%. Statistical significance notwithstanding, shifts are, in general, quite small on an absolute scale: repeated SQS runs return in 15 out of 21 cases $<R^2_V>$ values within 0.025. Shifts do not appear to be related to the total phase space volume, except for the fact that CAX descriptors show, again, the best stability. All the FPT-based nonlinear model sets for CU showed an excellent reproducibility of the average R^2_V values in spite of lowest RRs. While nonlinearity seems to enhance stability of CU models, the HEPT set is characterized by overall low average shifts.

By contrast, the sampling of FPT-based TIBO models in general and of nonlinear models in particular shows significant reproducibility problems. An in-depth analysis of FPT-based nonlinear TIBO models has evidenced a surprisingly strong dependence of the validation propensities on the TS/VS splitting schemes. Only splitting schemes #3 and #5 lead to model sets with very high—and highly reproducible—average validation propensities—for split #2, $A(\mathbf{R}) = 0.52$, $A(\mathbf{R}') = 0.61$, for split #5, $A(\mathbf{R}) = 0.66$, $A(\mathbf{R}') = 0.67$, while for all other splits $0.12 < A < 0.32$. TIBO average validation propensity values are thus very sensitive to the sizes of representative model sets. These sizes may actually vary by a factor of 2, as they basically depend on how quickly the meta-optimization termination criterion is fulfilled. Insofar validation is not strongly splitting scheme-dependent, doubling the size of the solution pool for a specific splitting scheme upon SQS repeat will not impact on the average. This is not the case with TIBO—specifically, the repeat of the TIBO run with the well validating splitting scheme #5 lead to 1193 solutions, compared to only 626 found initially, which is more than sufficient in order to bias the second twin run toward significantly higher global averages.

The splitting scheme dependence of the TIBO model validation propensity is strongest with the pharmacophore triplets. This probably signals the existence of a small subset (<<1/5 of the 73 compounds) featuring a specific but relevant triplet. Validation failures may arise if splitting schemes group most of these examples in VS, so that learning does not have enough examples at hand to pick the key triplet. Switching from triplets to pharmacophore pairs is a radical solution to avoid sparsely populated descriptor matrices. CAX two-point pharmacophore models are expectedly more robust with respect to splitting scheme choice and therefore more reproducible. Unfortunately, this is not a solution to the QSAR problem: data compression upon resolution decrease causes various specific pharmacophore signatures to lose their identities when merged into coarser categories. This may well enhance reproducibility, in the sense of having models reproducibly *failing* to validate. Although fluctuation-prone, FPT TIBO models are as successful in validation tests as their CAX counterparts (see later on).

As a general conclusion, in most situations a single SQS run—or perhaps even fewer MB repeats—may produce a set of models that is representative in terms of validation propensities. Repeated SQS runs will typically discover novel equations but—in most cases—hardly any with radically improved validation behaviors. It may thus be concluded that these QSAR problem phase spaces feature many different attraction pools with roughly the same validation propensities. Several repeats are however mandatory if a complete mapping of the problem space is envisaged, and in view of measured RRs, the cost for enumerating all the properly cross-validating models may easily become prohibitive if $N > 200$.

3.2. Effect of Nonlinearity and Descriptor Choice on Model Validation Propensities. A nearly systematic increase from linear to polynomial to nonlinear models was expectedly witnessed for both Q^2 values and associated training set correlation coefficients (results not shown). The benchmarking study in Table 3 however confirms that overfitting does not occur: no significant validation propensity loss was observed in spite of phase space volume

Table 3. Benchmark of the Relative Average Validation Propensities of Models with Respect to the Nonlinearity Policy and Descriptor Choice[a]

		CU	HEPT	TIBO
N vs L	best	0.677(=)	0.581 (N)	0.420 (=)
	Δ	0.024	0.093	0.051
	δ	–	0.069	–
	t	10.11	28.48	10.65
N vs P	best	0.677 (N)	0.581 (N)	0.420 (=)
	Δ	0.049	0.076	0.080
	δ	0.039	0.051	–
	t	22.91	22.36	16.66
P vs L	best	0.653 (=)	0.505 (=)	0.369 (=)
	Δ	0.025	0.017	0.029
	δ	–	0.003	–
	t	9.48	4.92	5.72
FPT vs ISIDA	best	0.639 (FPT-1)	0.627 (ISIDA)	0.403 (ISIDA)
	Δ	0.071	0.131t	0.049
	δ	0.040	0.126	0.032
	t	33.26	50.60	15.88
FPT vs CAX	best	0.639 (FPT-1)	0.548 (CAX)	0.355 (=)
	Δ	0.048	0.052	0.009
	δ	0.035	0.032	–
	t	26.30	20.62	2.32
ISIDA vs CAX	best	0.591 (=)	0.627 (ISIDA)	0.403 (ISIDA)
	Δ	0.023	0.079	0.057
	δ	–	0.059	0.046
	t	10.91	30.61	17.36

[a] Top half, the pairwise comparisons of the three explored nonlinearity premises (N,P,L), based on FPT descriptors, report the average R^2_V of the winning nonlinearity policy (shown in parentheses, = if no statistically relevant differences exist) the observed average shift Δ – eq 7, the minimal guaranteed average shift δ – eq 8, and the t factor of the observed shift – eq 7a. Bottom half, benchmarking of the validation propensities of the joined linear and polynomial model pools obtained with the three different classes of descriptors.

increase. Nonlinear models actually outperform their linear counterparts. HEPT nonlinear models are significantly better validators. With CU and TIBO, they also fare better, but shifts fall short from reaching statistical significance. Nonlinear approaches furthermore outperform polynomial models, being significantly better in two out of three cases. Polynomial and linear models have no significantly differing behaviors. The tendencies underlined by the statistical studies can be intuitively illustrated by density distribution histograms with respect to the validation scores of representative models: in Figure 3, HEPT/FPT-based nonlinear models (grid fill pattern) preferentially accumulate at the top end of the R^2_V scale (X axis, middle plot), unlike the equivalent CU models (upper plot), In the lower plot, it can be seen that the relative proportion of models failing to validate appears to be lower within the nonlinear TIBO models.

According to the bottom half of Table 3, validation propensities—of merged linear and polynomial model sets—in different descriptor spaces are clearly uncorrelated with descriptor set size, or else CAX-based models should have been top performers. Descriptor spaces of dimension $N \leq 1600$ are thus not a prohibitively large "haystack" for SQS to find well validating models. The observed validation propensity differences are thus the expression of the different nature of the chemical information encoded by the descriptors. ISIDA fragment descriptors are outperforming FPT on the HEPT and TIBO sets, while FPT is the most successful in building CU models. This could have been expected, as the latter family of cyclic ureas is built around a single common and large scaffold and is therefore less diverse in terms of represented substructures (out of the $N = 1586$ populated ISIDA fragments, many are common substructures of the cyclic urea scaffold itself). FPT descriptors, encoding specific information about the pharmacophore "ornaments" around the scaffold, are thus better suited for CU QSAR model buildup. CU is furthermore the only set for which CAX descriptors are not outperformed by ISIDA, this underlining the CU preference for a pharmacophore-oriented approach. In this context, the pharmacophore triplets also outperform the CAX pharmacophore pairs, but this trend may vanish (TIBO) or reverse (HEPT) with compound sets of increased internal topological diversity. The validation propensity distribution histograms of Figure 4 clearly illustrate the above-mentioned QSAR-ability differences for HEPT (ISIDA>CAX>FPT) and CU (FPT>ISIDA>CAX). Histograms of winning ISIDA and respectively FPT descriptors witness a clear global shift (lower participations at low R^2_V compensated by higher ones at the high end). With TIBO, however, the situation is less clear-cut: although FPT-based models have an extremely high rate of complete validation failure (30% have $R^2_V<0.1$), they also have the best rate of strong validation success (at $R^2_V \geq 0.7$). ISIDA-based models are failsafe but not outstanding—rarely completely failing to validate but rarely scoring excellent validation scores—and nevertheless better in terms of average validation propensities.

3.3. Relative Performance of Stochastic QSAR Sampling and Stepwise Regression. Relative validation success rates serve as comparators of SR vs SQS performance. They express the probability that a model picked at random out of the pool of equations provided by the method validates beyond a specified $R^2_{V,\mathrm{lim}}$ criterion. The multiple successful models produced by SQS may be useless if "drowned" in a much larger collection of bad validators.

Table 4 shows that for CU fragment-based approaches the density of well validating SQS models is systematically twice as big as with SR, but both approaches do produce valuable

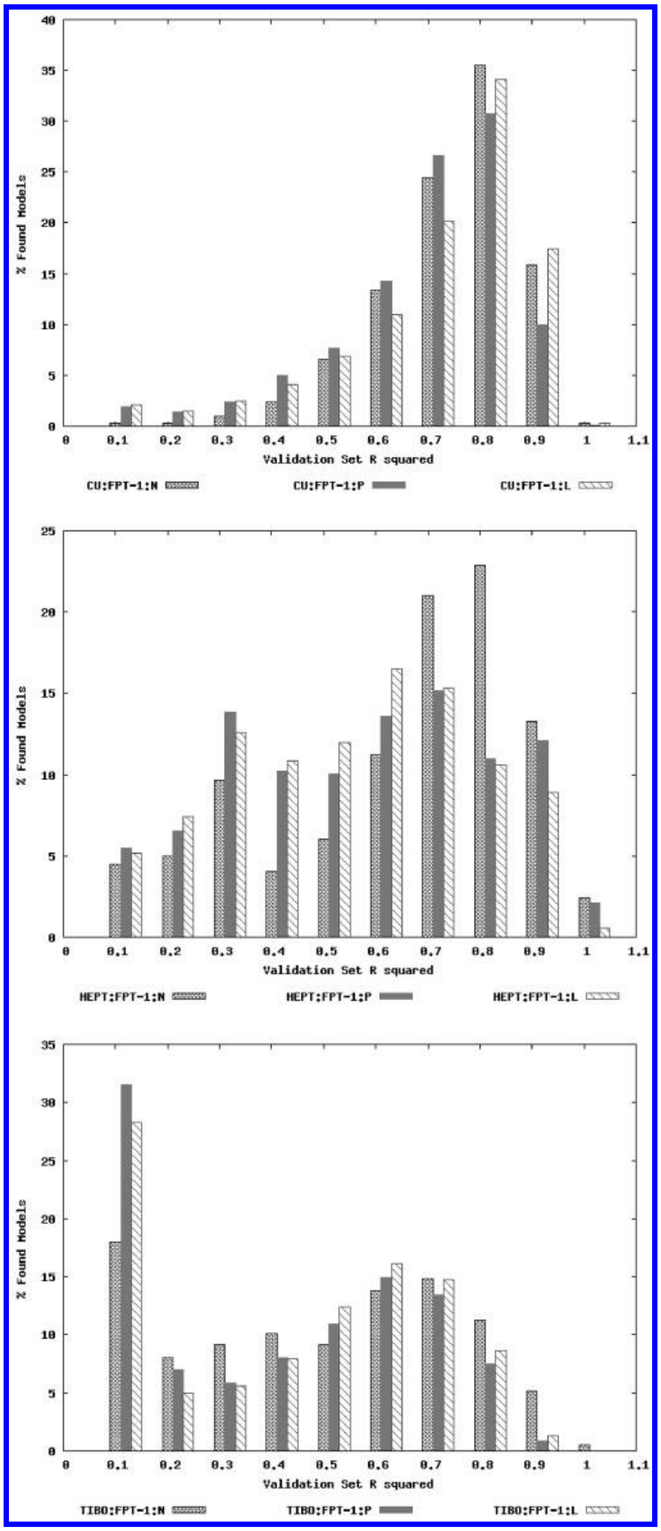

Figure 3. Comparative density distribution histograms of representative nonlinear (grid filling), polynomial (solid gray), and linear (hashed) FPT based models, representing on Y the percentage of models of ***R*** having validation correlation coefficients within each of the 10 bins listed on X (label represents upper bin threshold).

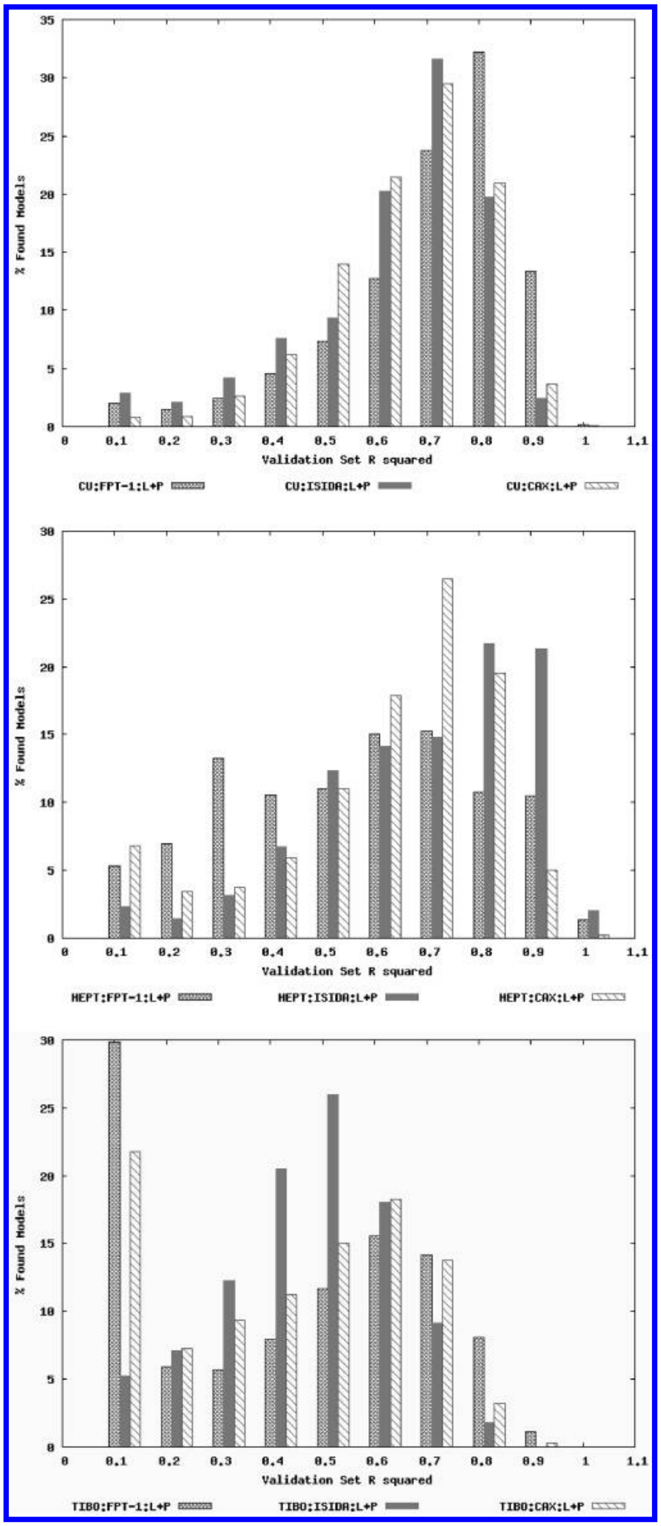

Figure 4. Comparative density distribution histograms of merged linear and polynomial SQS model sets obtained with FPT (grid filling), ISIDA (solid gray), and CAX (hashed) descriptors (check Figure 3 for additional information).

models at any $R^2_{V,\mathrm{lim}}$. For HEPT fragment-based approaches, the percentage of robust models as well as its trend as a function of $R^2_{V,\mathrm{lim}}$ are on the whole quite similar. SQS models are, at the strictest threshold of $R^2_{V,\mathrm{lim}} = 0.8$, at least as likely—and sometimes (HEPT/FPT) significantly more likely—to validate than SR models. The percentages of models exceeding a validation score of 0.6 are also quite satisfactory.

The situation with respect to the TIBO set is clearly the most intriguing: with ISIDA descriptors, the SQS tool is seen to consistently generate many well validating models that are unfortunately outnumbered by the validation failures. The SR approach, by contrast, picks a set of models with excellent validation propensities. When using FPT descriptors, the model family proposed by the SR tool is a family of nonvalidating equations. SQS, by contrast, discovers a

Table 4. Number of Models (and the Percentage They Represent out of the Entire Pool of Found Equations, %) for Which $R^2_V \geq R^2_{V,\mathrm{lim}}$[a]

		CU		HEPT		TIBO	
$R^2_{V,\mathrm{lim}}$		SR	SQS	SR	SQS	SR	SQS
ISIDA	0.6	51 (33%)	3381 (64%)	124 (78%)	2886 (55%)	74 (31%)	538 (13%)
	0.7	12 (8%)	1624 (31%)	105 (65%)	2152 (41%)	25 (10%)	102 (2%)
	0.8	1 (1%)	207 (4%)	50 (31%)	1220 (23%)	9 (4%)	2 (0.05%)
FPT	0.6	13 (52%)	5971 (72%)	12 (40%)	3254 (36%)	3 (10%)	1437 (25%)
	0.7	10 (40%)	4278 (52%)	7 (23%)	1847 (20%)	0	579 (10%)
	0.8	5 (20%)	1492 (18%)	0	880 (10%)	0	79 (1%)

[a] SR models involving ISIDA descriptors include both the ones issued from SR-1 and SR-2 strategies, whereas those based on the FPT descriptors were all obtained with SR-1. Situations in which the percentages of successful validators are at least twice as large as those seen by the alternative procedure were shaded (7 in favor of SQS, 3 in favor of SR).

Table 5. Validation (R^2_V) Correlation Coefficients and Fraction of Worse Performing Individual Models (*%W*) for SQS and Respectively SR (in Parentheses) Consensus Equations for Each Splitting Scheme **S** and ISIDA Descriptors[a]

	CU		HEPT		TIBO	
S	R^2_V	%W	R^2_V	%W	R^2_V	%W
1	0.42 (0.46)	73 (63)	0.82 (0.81)	96 (88)	0.59 (0.57)	86 (65)
2	0.69 (0.65)	74 (83)	**0.56 (0.70)**	75 (79)	**0.44 (0.67)**	90 (100)
3	*0.80 (0.59)*	90 (64)	0.83 (0.87)	86 (91)	**0.60 (0.74)**	72 (80)
4	*0.76 (0.63)*	87 (83)	0.94 (0.90)	98 (82)	0.46 (0.55)	77 (90)
5	0.78 (0.70)	96 (100)	**0.51 (0.85)**	75 (94)	*0.63 (0.44)*	90 (48)

[a] Cells with either of the SQS or SR consensus equation outperforming the other by more than 0.1 R^2_V units are highlighted: italics when SQS wins, boldface when SR is better (3 in favor of SQS and 4 in favor of SR). "Nonredundancy" is indeed desirable, but a fail-safe definition for nonredundancy is difficult to give.

significant number of successful validators which, furthermore, are less "diluted" by nonvalidators. With the reserve that more studies on different sets should be run in order to reinforce the herein drawn conclusions, it can be said that the SQS approach offers the best guarantees to discover well validating models, if they exist. If SQS were to be used as a generator of numerous equations, only in order to pick one out at random and use it instead of a SR model, then the advantage of SQS is the virtual certainty that well validating models will be *present* in the initial pool of choices. However, SR also lived up to this expectation in 5 challenges out of 6, failing only under the TIBO/FPT premises. In light of this, the advantage of SQS is small and overshadowed by the higher computer cost. The disadvantage of SQS is that the well validating models, although potentially numerous, might yet represent only a small fraction of the entire solution pool. In such cases, picking a single SQS model from the large equation pool may prove a riskier approach than selecting any of the related SR models. However, SQS was only once affected by this caveat in six runs.

3.4. Consensus Strategy. The alternative to random picking single models among the representative ones is the use of consensus equations (Table 5). The R^2_V values of linear SQS and SR (in parentheses) CM built for each of 5 splitting schemes of the 3 compound sets with ISIDA descriptors. Individual models with validation scores below the consensus approach were accounted for as percentages of worse performers *%W*. Typically, SQS CM validate better than 80% of single equations. With certain splitting schemes, they may score low R^2_V values but nevertheless outperform most of individual models. Plain average consensus modeling remains a valid strategy even in the context of the many diverse equations from the representative SQS sets.

SR and SQS performances are again quite similar, in spite of the more sophisticated build-up procedure of SR CM, using outlier detection:[18] a count of situations in which either of the methods outperforms the other by more than 0.1 R^2_V score units reveals 4 in favor of SR, 3 in favor of SQS, and 8 draws.

3.5. Beyond Statistics: What Can Be Learned from Problem Space Mapping? If we take the ultimate goal of QSAR studies to be the discovery of at least one well-validating equation, then SQS does not appear be worth the excess computer effort (days of work on X86 biprocessor workstations, rather than hours). Both methods perform quite well—but then, *all* the QSAR models that were ever published do perform well in terms of validation tests, though few were reported to discover actives in actual virtual screening. In light of the insights gained from this study, this is not surprising. In a QSAR problem space with few well validating models, SR would be the ideal modeling tool—in as far as it succeeds to discover them. Such sets may exist, but none was encountered here—these feature *several broad zones populated by models of comparable validation propensity*. In there, SR equations have no special status—they are just typical models among many thousands. Similar results might be obtained with less sophisticated stochastic samplers[36] also relying on upfront descriptor candidate filtering. That any two well validating models may nevertheless have different application ranges and return diverging predictions when applied to external compounds is a fact as obvious as the "Kubinyi paradox".[16] A *Gedanken experiment* suffices to explain why: consider a mixed set of pharmacophore and electronic effect descriptors on a series in which a substituted phenyl ring is a key feature. It was found that -OH, -OR, -SR, -NR_2, or -NHR substituted compounds tend to be active, while -halogen, -H, and -alkyl substituted ones are not. There are two alternative explanations: "hydrogen bond acceptor required" (pharmacophore descriptor entering the model) or "electron-enriched phenyl required" (electronic effect descriptor chosen). They are of comparable predictive power and validation propensity: since the phenyls carrying an acceptor are—as far as this set goes—electron-enriched, pharmacophore and electronic effect descriptors are strongly correlated. SR discards one of the two and comes up with one single alternative. Therefore, correct prediction of the say carbonyl-substituted analog is a matter of luck: only the pharmacophore model returns "active" (−CR=O is an acceptor). As both training and test sets fail to include such an example—for good reason, perhaps: electron-withdrawing effects may cause synthesis problems—SR does not have any means to issue a warning about this intrinsic degeneracy of the chemical information in the training set. By contrast, SQS may correctly enumerate both alternatives. There is still no way to know which is mechanistically correct, but having predictions carried out with both of these apparently indiscernible models may at last evidence the inherent limitations of the training set and suggest how to enrich it.

In the early days of QSAR, aggressive pruning of the set of initial descriptors was a technical constraint, and eliminating correlated terms was the less worse of arbitrary choices forced upon the user. Descriptor correlations are often training/test set specific (even with thousands of compounds[37]). Moreover, the small difference between two large and correlated terms may nevertheless "hide" a genuine independent variable—a well-known example being the free energy, not related to either enthalpy or entropy although the latter two are often correlated.[38] In spite of such sources of potential pitfalls, the paradigm of the nonredundant descriptor set became enshrined in QSAR building protocols, although the technical bottleneck at its origin has long since vanished. In our opinion, the argument that SR approaches are superior to stochastic sampling because they return few models and thus avoid the question of which one to actually use in drug design is fallacious. An extensive mapping of the QSAR problem space may be the key to reducing the overall failure rate of QSAR-driven virtual screening, but classical validation tests cannot shed light on how to best use the wealth of SQS-generated information. This work proves that, with hundreds of compounds and thousands of candidate descriptors, such a complete mapping would be feasible in a matter of weeks using state-of-the-art desktop PCs. Further effort will be dedicated to understand how to best use problem space maps in virtual screening.

4. CONCLUSIONS

The SQS model build-up procedure reported and tested in this paper successfully discovered multiple, successfully validating models under various working premises. This section sums up the observations with respect to the key questions in the Introduction and ends with a general debate of the insights gained due to this work.

1. Reproducibility. Given the huge problem space volumes it is meant to sample, SQS will never list all possible QSAR models nor rediscover the same if repeated. This notwithstanding, the model sets it actually produces display comparable validation performances. Sets of SQS equations close to optimal cross-validated Q^2 were found to be rich in successfully validating models.

2. Dependence on Phase Space Volume: Nonlinearity Policy and Descriptor Choice. Average validation propensities of SQS models are independent of problem space volumes: neither the introduction of preset nonlinear transformations nor moving from smaller to larger descriptor sets triggered overfitting artifacts. On the contrary, the introduction of nonlinearity actually had a positive impact. This proves that appropriate model building—pressure to minimize the number of entering variables and systematic cross-validation as part of the model fitness estimation—may avoid overfitting.

3. Stochastic vs Stepwise Model Building: Comparison of Individual Models. While SR may occasionally fail to produce any validating models when SQS succeeds (this has been observed once in six cases covered by the study), the latter approach may occasionally "hide" the numerous, valuable equations within an even larger set of nonvalidators (seen once in six cases, as well). Pharmacophore descriptors are more likely to cause SR failure with the current sets.

4. Consensus Strategy. Consensus SQS models were found to display an extremely robust behavior, virtually always showing better characteristics than 70–90% of individual models. The scale-up from tens of related SR models to 10^4 randomly picked SQS equations does not impede on the validity of the CM paradigm. SQS CM are however not outstandingly better, nor worse, than SR consensus equations as far as the validation exercise may tell, but they might have decisive advantages when used in virtual screening of large databases.

Finding more models does not automatically imply finding much better models—in the sense of standard intrafamily validation tests. Such (necessary, but hardly sufficient) tests are not the ultimate QSAR validity criterion and cannot tell which is the mechanistically sound equation out of many "apparently" equivalent forms (equivalent as far as training and validation sets go but not necessarily throughout the space of druglike compounds). The benefits of a global analysis of QSAR problem space must therefore be addressed from the more general point of view of actual utility in drug design. SQS-driven model sampling might be used to get an idea on the degree of degeneracy of the chemical information in the training set and on the novel compounds to add to training in order to lift some of these degeneracies.

Abbreviations: **QSAR** − quantitative structure−activity relationships, **TS/VS** − training/validation set, **FPT** − fuzzy pharmacophore triplets, **ISIDA** − fragment descriptors, **CAX** − ChemAxon descriptors, **CM** − consensus models, **MLR** − multilinear regression, **SR** − stepwise regression, **GA** − genetic algorithm, **MB** − Model Builder: distributed, GA-based QSAR model generator, **SQS** − stochastic QSAR sampler, managing the parameter control and centralizing the output of repeated MB runs, **RR** − rediscovery rate

Supporting Information Available: Data sets, splitting schemes, and molecular descriptors plus an information file describing the exact contents of each data file—everything under Unix text format. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Lucic, B.; Nadramija, D.; Basic, I.; Trinajstic, N. Towards generating simpler QSAR models: Nonlinear multivariate regression versus several neural network ensembles and related methods. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1094−1102.

(2) Milicevic, A.; Nikolic, S.; Trinajstic, N. Toxicity of aliphatic ethers: A comparative study. *Mol. Diversity* **2006**, *10*, 95−99.

(3) Adam, M. Integrating research and development: the emergence of rational drug design in the pharmaceutical industry. *Stud. Hist. Philos. Biol. Biomed. Sci.* **2005**, *36*, 513−37.

(4) Hawkins, D. M.; Basak, S. C.; Mills, D. Assessing model fit by cross-validation. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 579−586.

(5) Baumann, K. Cross-validation as the objective function for variable selection techniques. *TrAC, Trends Anal. Chem.* **2003**, *22*, 395−406.

(6) Bonachera, F.; Parent, B.; Barbosa, F.; Froloff, N.; Horvath, D. Fuzzy Tricentric Pharmacophore Fingerprints. 1 - Topological Fuzzy Pharmacophore Triplets and adapted Molecular Similarity Scoring Schemes. *J. Chem. Inf. Model.* **2006**, published on Web 10.21.2006.

(7) Parent, B.; Kökösy, A.; Horvath, D. Optimized Evolutionary Strategies in Conformational Sampling. *Soft Computing* **2007**, *11*, 63−79.

(8) Solov'ev, V. P.; Varnek, A.; Wipff, G. Modeling of Ion Complexation and Extraction Using Substructural Molecular Fragments. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 847−858.

(9) http://www.chemaxon.com/jchem/index.html?content=doc/user/Screen.html (accessed Feb 2, 2006).

(10) Kubinyi, H.; Hamprecht, F. A.; Mietzner, T. Three-dimensional quantitative similarity-activity relationships (3D QSiAR) from SEAL similarity matrices. *J. Med. Chem.* **1998**, *41*, 2553−2564.

(11) Baurin, N.; Mozziconacci, J.-C.; Arnoult, E.; Chavatte, P.; Marot, C.; Morin-Allory, L. 2D QSAR Consensus Prediction for High-Throughput Virtual Screening. An Application to COX-2 Inhibition Modeling and Screening of the NCI Database. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 276−285.

(12) Varnek, A.; Fourches, D.; Solov'ev, V. P.; Baulin, V. E.; Turanov, A. N.; Karandashev, V. K.; Fara, D.; Katritzky, A. R. "In Silico" Design of New Uranyl Extractants Based on Phosphoryl-Containing Podands: QSPR Studies, Generation and Screening of Virtual Combinatorial Library, and Experimental Tests. *J. Chem. Inf. Comput. Sci.* **2005**, *44*, 1365−1382.

(13) Katritzky, A. R.; Kuanar, M.; Fara, D. C.; Karelson, M.; Acree, W. E., Jr.; Solov'ev, V. P.; Varnek, A. QSAR modeling of blood:air and tissue:air partition coefficients using theoretical descriptors. *Bioorg. Med. Chem.* **2005**, *13*, 6450−6463.

(14) Varnek, A.; Fourches, D.; Hoonakker, F.; Solov'ev V. P. Substructural fragments: an universal language to encode reactions, molecular and supramolecular structures. *J. Comput.-Aided Mol. Des.* **2005**, *19*, 693−703.

(15) Golub, G. H.; Reinsch, C. Singular value decompositions and least squares solutions. *Numer. Math.* **1970**, *14*, 403−420.

(16) Katritzky, A. R.; Kuanar, M.; Slavov, S.; Dobchev, D. A.; Fara, D. C.; Karelson, M.; Acree, W. E., Jr.; Solov'ev, V. P.; Varnek, A. Correlation of blood-brain penetration using structural descriptors. *Bioorg. Med. Chem.* **2006**, , *14*, Jul 15, 4888−4917.

(17) Tetko, I. V.; Solov'ev, V. P.; Antonov, A. V.; Yao, X.; Doucet, J. P.; Fan, B.; Hoonakker, F.; Fourches, D.; Jost, P.; Lachiche, N.; Varnek, A. Benchmarking of Linear and Nonlinear Approaches for Quantitative Structure-Property Relationship Studies of Metal Complexation with Ionophores. *J. Chem. Inf. Model.* **2006**, *46*, 808−819.

(18) Grubbs, F. Procedures for Detecting Outlying Observations in Samples. *Technometrics* **1969**, *11*, 1−21.

(19) Varnek, A.; Wipff, G.; Solov'ev, V. P.; Solotnov, A. F. Assessment of the Macrocyclic Effect for the Complexation of Crown-Ethers with Alkali Cations Using the Substructural Molecular Fragments Method. *J. Chem. Inf. Comput. Sci*. **2002**, *42*, 812.

(20) Wilkerson, W. W.; Akamike, E.; Cheatham, W. W.; Hollis, A. Y.; Collins, R. D.; DeLucca, I.; Lam, P. Y.; Ru, Y. HIV Protease Inhibitory Bis-benzamide Cyclic Ureas: A Quantitative Structure-Activity Relationship Analysis. *J. Med. Chem.* **1996**, *39*, 4299−4312.

(21) Wilkerson, W. W.; Dax, S.; Cheatham, W. W. Nonsymmetrically Substituted Cyclic Urea HIV Protease Inhibitors. *J. Med. Chem.* **1997**, *40*, 4079−4088.

(22) Lam, P. Y.; Ru, Y.; Jadhav, P. K.; Aldrich, P. E.; DeLucca, G. V.; Eyermann, C. J.; Chang, C. H.; Emmett, G.; Holler, E. R.; Daneker, W. F.; Li, L.; Confalone, P. N.; McHugh, R. J.; Han, Q.; Li, R.; Markwalder, J. A.; Seitz, S. P.; Sharpe, T. R.; Bacheler, L. T.; Rayner, M. M.; Klabe, R. M.; Shum, L.; Winslow, D. L.; Kornhauser, D. M.; Hodge, C. N. Cyclic HIV protease inhibitors: synthesis, conformational analysis, P2/P2′ structure-activity relationship, and molecular recognition of cyclic ureas. *J. Med. Chem.* **1996**, *39*, 14−3525.

(23) Miyasaka, T.; Tanaka, H.; Baba, M.; Hayakawa, H.; Walker, R. T.; Balzarini, J.; De Clercq, E. A Novel Lead for Specific Anti-HIV-1 Agents: 1-[(2-Hydroxyethoxy)methyl]-6-(phenylthio)thymine. *J. Med. Chem.* **1989**, *32*, 2507−2509.

(24) Tanaka, H.; Baba, M.; Hayakawa, H.; Haraguchi, K.; Miyasaka, T.; Ubasawa, M.; Takashima, H.; Sekiya, K.; Nitta, I.; Walker, R. T.; De Clercq, E. Lithiation of uracil nucleosides and its application to the synthesis of a new class of anti-HIV-1 acyclonucleosides. *Nucleosides Nucleotides* **1991**, *10*, 397−400.

(25) Tanaka, H.; Baba, M.; Saito, S.; Miyasaka, T.; Takashima, H.; Sekiya, K.; Ubasawa, M.; Nitta, I.; Walker, R. T.; Nakashima, H.; De Clercq, E. Specific anti-HIV-1 acyclonucleosides which cannot be phosphorylated: synthesis of some deoxy analogs of 1-[(2-hydroxyethoxy)-methyl]-6-(phenylthio)thymine. *J. Med. Chem.* **1991**, *34*, 1508−1511.

(26) Tanaka, H.; Baba, M.; Ubasawa, M.; Takashima, H.; Sekiya, K.; Nitta, I.; Shigeta, S.; Walker, R. T.; De Clercq, E.; Miyasaka, T. Synthesis and anti-HIV activity of 2-, 3-, and 4-substituted analogs of 1-[(2-hydroxyethoxy)methyl]-6-(phenylthio)thymine (HEPT). *J. Med. Chem.* **1991**, *34*, 1394−1399.

(27) Tanaka, H.; Baba, M.; Hayakawa, H.; Sakamaki, T.; Miyasaka, T.; Ubawasa, M.; Takashima, H.; Sekiya, K.; Nitta, I.; Shigeta, S.; Walker, R. T.; Balzarini, J.; De Clercq, E. A New Class of HIV-1 Specific 6-Substituted Acyclouridine Derivatives: Synthesis and Anti-HIV-1 Activity of 5- or 6-Substituted Analogs of 1-[(2-Hydroxyethoxy)-methyl]-6-(phenylthio)thymine (HEPT). *J. Med. Chem.* **1991**, *34*, 349−357.

(28) Tanaka, H.; Takashima, H.; Ubasawa, M.; Sekiya, K.; Nitta, I.; Baba, M.; Shigeta, S.; Walker, R. T.; De Clercq, E.; Miyasaka, T. Synthesis and antiviral activity of deoxy analogs of 1-[(2-hydroxyethoxy)methyl]-6-(phenylthio)thymine (HEPT) as potent and selective anti-HIV-1 agents. *J. Med. Chem.* **1992**, *35*, 4713−4719.

(29) Tanaka, H.; Takashima, H.; Ubasawa, M.; Sekiya, K.; Nitta, I.; Baba, M.; Shigeta, S.; Walker, R. T.; De Clercq, E.; Miyasaka, T. Structure-activity relationships of 1-[(2-hydroxyethoxy)methyl]-6-(phenylthio)-thymine analogs: effect of substitutions at the C-6 phenyl ring and at the C-5 position on anti-HIV-1 activity. *J. Med. Chem.* **1992**, *35*, 337−345.

(30) Tanaka, H.; Takashima, H.; Ubasawa, M.; Sekiya, K.; Inouye, N.; Baba, M.; Shigeta, S.; Walker, R. T.; De Clercq, E.; Miyasaka, T. Synthesis and Antiviral Activity of 6-Benzyl Analogs of 1-[(2-Hydroxyethoxy)methyl]-5-(phenylthio)thymine (HEPT) as Potent and Selective Anti-HIV-1 Agents. *J. Med. Chem.* **1995**, *38*, 2860−2865.

(31) Kukla, M. J.; Breslin, H. J.; Pauwels, R.; Fedde, C. L.; Miranda, M.; Scott, M. K.; Sherrill, R. G.; Raeymaekers, A.; van Gelder, J.; Andries, K.; Moens, L. J.; Janssen, M. A. C.; Janssen, P. A. J. Synthesis and anti-HIV-1 activity of 4,5,6,7-tetrahydro-5-methylimidazo[4,5,1-jk]-[1,4]benzodiazepin- 2(1H)-one (TIBO) derivatives. *J. Med. Chem.* **1991**, *34*, 746−751.

(32) Ho, W.; Kukla, M. J.; Breslin, H. J.; Ludovici, D. W.; Grous, P. P.; Diamond, C. J.; Miranda, M.; Rodgers, J. D.; Ho, C. Y.; De Clercq, E.; Pauwels, R.; Andries, K.; Janssen, M. A. C.; Janssen, P. A. J. Synthesis and anti-HIV-1 activity of 4,5,6,7-tetrahydro-5-methylimidazo-[4,5,1-jk][1,4]benzodiazepin- 2(1H)-one (TlBO) derivatives. 4. *J. Med. Chem.* **1995**, *38*, 794−802.

(33) Breslin, H. J.; Kukla, M. J.; Ludovici, D. W.; Mohrbacher, R.; Ho, W.; Miranda, M.; Rodgers, J. D.; Hitchens, T. K.; Leo, G.; Gauthier, D. A.; Ho, C. Y.; Scott, M. K.; De Clercq, E.; Pauwels, R.; Andries, K.; Janssen, M. A. C.; Janssen, P. A. J. Synthesis and anti-HIV-1 activity of 4,5,6,7-tetrahydro-5-methylimidazo [4,5,1-jk][1,4]-benzodiazepin-2(1H)-one (TIBO) derivatives. 3. *J. Med. Chem.* **1995**, *38*, 771−793.

(34) Solov'ev, V. P.; Varnek, A. Anti-HIV activity of HEPT, TIBO, and cyclic urea derivatives: structure-property studies, focused combinatorial library generation, and hits selection using substructural molecular fragments method. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1703−19.

(35) Grais, B. L'interprétation des sondages aléatoires : problèmes d'estimation et de comparaison. In *Méthodes Statistiques*, 3rd ed.; Dunod Eds.; Dunod: Paris, France, 1992; pp 288−296.

(36) Rolland, C.; Gozalbes, R.; Nicolai, E.; Paugam, M. F.; Coussy, L.; Barbosa, F.; Horvath, D.; Revah, F. G-protein-coupled receptor affinity prediction based on the use of a profiling dataset: QSAR design, synthesis, and experimental validation. *J. Med. Chem*. **2005**, *48*, 6563−74.

(37) Horvath, D.; Mao, B.; Gozalbes, R.; Barbosa, F.; Rogalski, S. L. Strengths and Limitations of Pharmacophore-Based Virtual Screening. In *Cheminformatics in Drug Discovery*, 1st ed.; Oprea, T. I., Ed.; WILEY-VCH Verlag GmbH & Co. KGaA: Weinheim, Germany, 2004; pp 117−137.

(38) Ford, D. M. Enthalpy-Entropy Compensation is Not a General Feature of Weak Association. *J. Am. Chem. Soc.* **2005**, *127*, 16167−16170.

CI600476R