A Comparison of Methods for Modeling Quantitative Structure−Activity Relationships

Jeffrey J. Sutherland,[†] Lee A. O'Brien,[‡] and Donald F. Weaver*[,§]

Departments of Chemistry and Pathology, Queen's University, Kingston, Ontario K7L 3N6, Canada, and Departments of Medicine (Neurology) and Chemistry and School of Biomedical Engineering, Dalhousie University, Halifax, Nova Scotia B3H 4J3, Canada

Received April 13, 2004

A large number of methods are available for modeling quantitative structure−activity relationships (QSAR). We examine the predictive accuracy of several methods applied to data sets of inhibitors for angiotensin converting enzyme, acetylcholinesterase, benzodiazepine receptor, cyclooxygenase-2, dihydrofolate reductase, glycogen phosphorylase b, thermolysin, and thrombin. Descriptors calculated with CoMFA, CoMSIA, EVA, HQSAR, and traditional 2D and 2.5D descriptors were used for developing models with partial least squares (PLS). In addition, the genetic function approximation algorithm, genetic PLS, and back-propagation neural networks were used for deriving models from 2.5D descriptors (i.e., 2D descriptors and 3D descriptors calculated from CORINA structures and Gasteiger−Marsili charges). Predictive accuracy was assessed using designed test sets. It was found that HQSAR generally performs as well as CoMFA and CoMSIA; other descriptor sets performed less well. When 2.5D descriptors were used, only neural network ensembles were found to be similarly or more predictive than PLS models. In addition, we show that many cross-validation procedures yield similar estimates of the interpolative accuracy of methods. However, the lack of correspondence between cross-validated and test set predictive accuracy for four sets underscores the benefit of using designed test sets.

Introduction

The fundamental premise of quantitative structure−activity relationships (QSAR) is that a macroscopic property of a molecule, such as binding affinity at a receptor, is determined by its molecular structure. QSAR methods attempt to capture the relationship between structural attributes of molecules and their biological activity. Traditionally, QSAR has been applied retrospectively to shed light on the manner by which molecules within a congeneric series modulate activity. However, QSAR methods are increasingly used for making predictions on novel derivatives, either for affinity at a biological receptor or for targets associated with ADMET properties, such as hERG, cytochrome P450, and P-glycoprotein.[1] To be useful for such applications, QSAR models must be capable not only of generalizing within a congeneric series (i.e., interpolate among compounds in the data set) but of correctly predicting activities for compounds outside the chemical space represented by the training set.

A large number of methods have been described in the literature for the modeling of structure−activity relationships. Some methods consider only the connection table of a molecule (i.e., 2D methods), while others consider the physicochemical properties of molecules in their bioactive conformation (i.e., 3D methods).[2,3] 2D methods using "traditional" molecular descriptors, such the χ indices,[4] counts of rotatable bonds, and molecular weight, have been used for hundreds (perhaps thousands) of QSAR analyses. More recently, a method that makes use of molecular holograms defined from the (2D) connection table of molecules has been described.[5,6] Hologram QSAR (HQSAR) encodes the presence or absence of molecular fragments in a manner analogous to the encoding of chemical structures used for similarity and substructure searching of chemical databases. In one sense, HQSAR may be viewed as a revival of the early Free−Wilson approach in which activities are correlated with the presence of various functional groups.[7]

Because ligand−receptor interactions are inherently 3D properties, there has been much effort to develop QSAR methods that exploit 3D properties of molecules. The most widely used 3D QSAR method is comparative molecular field analysis[8] (CoMFA), in which electrostatic and steric potential energies are calculated between a positively charged carbon atom located at each vertex of a rectangular grid and a series of molecules embedded within the grid. CoMFA requires the specification of both conformations and the relative alignments of molecules in the data set. A related method uses molecular potentials smoothed with Gaussian functions, eliminating singularities in the CoMFA steric and electrostatic fields that occur at atomic nuclei.[9] Comparative molecular similarity indices analysis (CoMSIA) has been shown to reduce the sensitivity to small changes in the alignment of compounds or the orientation of the grid. In addition, hydrogen-bonding and hydrophobic fields have been introduced to supplement the steric and electrostatic fields that only capture enthalpic contributions to binding.[10] Other 3D methods

* To whom correspondence should be addressed. Phone: (902) 494-7183. Fax: (902) 494-1310. E-mail: weaver@chem3.chem.dal.ca.

[†] Department of Chemistry, Queen's University.

[‡] Department of Pathology, Queen's University.

[§] Dalhousie University.

10.1021/jm0497141 CCC: $27.50 © 2004 American Chemical Society
Published on Web 09/21/2004

eliminate the need for aligning molecules relative to each other, although molecular conformation must still be specified. QSAR by eigenvalue analysis (EVA) uses a normal mode calculation to simulate the IR spectrum of a molecule, with each descriptor representing the intensity of the spectrum for a small range of frequencies.[11] The EVA descriptor sets for each molecule are submitted to QSAR analysis after a number of preprocessing steps.

Field-based 3D QSAR methods (e.g., CoMFA, CoMSIA), HQSAR, and EVA produce many more descriptors per molecule than the number of molecules in a typical data set. Partial least squares (PLS) is used to reduce the dimensionality of the descriptor set to a small number of orthogonal latent variables correlated with the property being modeled.[12] Even with traditional descriptors, there are often too many descriptors to allow the use of multiple linear regression as in classical QSAR.[13] Statistical methods such as PLS or other methods such as selection of descriptor subsets with genetic algorithms[14] or neural networks[15,16] can be used for developing models.

A number of methods for developing QSAR models have become well-established and are available in commercial software packages. Their accessibility through intuitive user interfaces has widened the community of QSAR practitioners beyond that of computational chemists. However, there are presently no clear guidelines to facilitate the selection of one method over others because there have been few wide-ranging comparisons of various approaches for modeling structure–activity relationships. The steroid data set used for validating CoMFA has become a standard set for comparing QSAR methods, as have others such as the Selwood data set[17] and a set of pyrimidine and triazine dihydrofolate reductase inhibitors.[18,19] The steroid and Selwood data sets are too small for allowing the assessment of predictive accuracy with a large test set. In this work, we compare various methods for encoding molecular structure (CoMFA, CoMSIA, HQSAR, EVA, and traditional descriptors) using eight data sets ranging from 66 to 397 compounds. For traditional descriptors, the genetic function approximation (GFA) algorithm, genetic PLS, and back-propagation neural networks are compared to PLS. The data sets are distributed in electronic format in the Supporting Information, allowing other researchers to compare their methods to the established approaches available in commercial software packages.

Methods

This work was carried out with Cerius2, version 4.8.1 (Accelrys, Inc.; San Diego, CA), operating under IRIX 6.5, and Sybyl, version 6.9 (Tripos Inc.; St. Louis, MO), operating under Red Hat Linux 7.3. Repetitive procedures were automated with Tcl (Cerius2), SPL (Sybyl), and AWK scripts.

(i) QSAR Data Sets. Eight data sets were used for comparing QSAR methods. In addition to tabulations of compounds and literature references, the selection of molecular conformations, alignment rules, and grids used for field-based 3D QSAR are described in the Supporting Information. Representative compounds from each data set are shown in Figure 1.

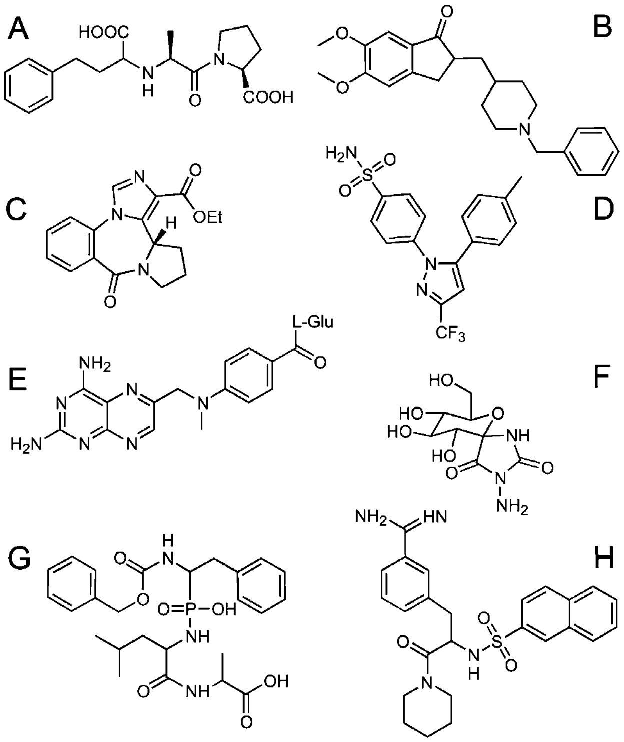

Figure 1. Representative compounds from each QSAR data set: (A) enalaprat (ACE); (B) E2020 (AchE); (C) Ro14-5974 (BZR); (D) celecoxib (COX2); (E) methotrexate (DHFR); (F) glucopyranose spirohydantoin (GPB); (G) ZPLA (THER); (H) naphtho derivative of 4-TAPAP (THR).

(1) A set of 114 angiotensin converting enzyme (ACE) inhibitors has been taken from the work of Depriest et al., which describes their use for CoMFA modeling.[20] Activities are spread over a wide range, with pIC_{50} values ranging from 2.1 to 9.9. (2) A set of 111 acetylcholinesterase (AchE) inhibitors has been assembled from the work of Sugimoto et al., with pIC_{50} values ranging from 4.3 to 9.5. A subset of these compounds has been studied with CoMFA.[21] (3) A set of 163 ligands for the benzodiazepine receptor (BZR) has been assembled from the work of Haefely et al. with pIC_{50} values ranging from 5.5 to 8.9. A subset has been used for validating several QSAR methods.[22] (4) A set of 322 cyclooxygenase-2 (COX2) inhibitors, assembled from work of Seibert et al., have pIC_{50} values that range from 4.0 to 9.0. A subset has been studied using CoMFA.[23] (5) A set of 397 dihydrofolate reductase inhibitors (DHFR) has been assembled from the work of Queener et al., with pIC_{50} values for rat liver enzyme ranging from 3.3 to 9.8. We have recently described CoMSIA models for this series of compounds,[24] and a subset has been used for deriving 2D QSAR models with neural networks.[25] The final three sets have been prepared by Klebe et al. All three have been widely studied with field-based 3D QSAR methods. (6) A set of 66 inhibitors of glycogen phosphorylase b (GPB) have pK_i values ranging from 1.3 to 6.8.[26] (7) A set of 76 thermolysin inhibitors (THER) have pK_i values ranging from 0.5 to 10.2.[9] (8) A set of 88 thrombin inhibitors (THR) have pK_i values ranging from 4.4 to 8.5.[27]

For the BZR, COX2, and DHFR sets, the compounds included in this work were selected from larger collec-

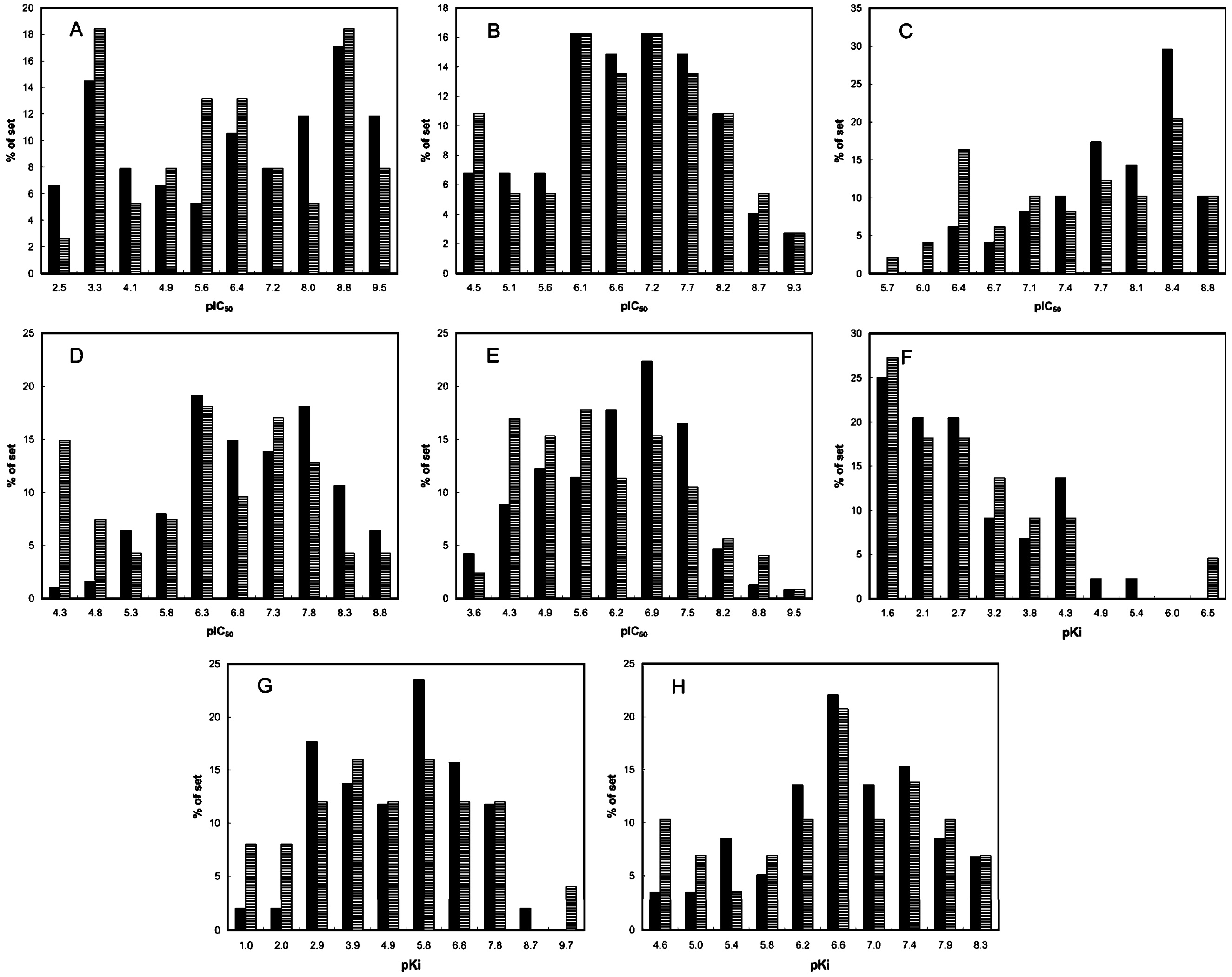

Figure 2. Distribution of activities for training sets (solid bars) and test sets (hashed bars): (A) ACE; (B) ACHE; (C) BZR; (D) COX2; (E) DHFR; (F) GPB; (G) THER; (H) THR.

tions that contained many highly redundant compounds. Using 2D (structural) fingerprints with the Tanimoto coefficient[28] (T_c) to calculate the pairwise similarity of compounds, subsets were selected using the sphere-exclusion algorithm[29] implemented in Cerius2. This gave subsets of compounds for which all pairs have $T_c < 0.975$. The excluded compounds were not used in any model development or evaluation and are not included in the Supporting Information. Within the same three sets, a number of compounds are reported with indeterminate activities (e.g., IC$_{50}$ > 10 μM); these were assigned to an inactive set used to verify if models can correctly identify inactives.

The compounds in each data set were divided between training and test sets. Approximately 33% were selected by "cherry picking" with a maximum dissimilarity algorithm[30,31] and assigned to the test set, with the remaining compounds assigned to the training set. The sets were structured this way to maximize the diversity of the test set and to examine the predictive accuracy of methods when extrapolating outside the training set. The maximum dissimilarity algorithm (the MaxMin function in Cerius2) maximizes the minimum squared distance from each compound to all other compounds in the selected subset, with pairwise distances determined using $1 - T_c$. The optimization uses a Monte Carlo procedure[31] that we have coupled to a simulated annealing protocol implemented in Tcl (up to 100 000 trial sets per pseudotemperature, which is lowered in 10% increments from 5000 to 10 K). The MaxMin function was optimized under restraint such that the selected compounds have a distribution of activities similar to that of the complete set. Without restraints, the selection procedure tended to yield training sets depleted of low- and high-activity compounds. The penalty function was obtained by assigning compounds to 10 evenly separated bins covering the range of pIC$_{50}$ or pK_i values. This gave reasonably similar distributions of activities for training and test set compounds (Figure 2). Since the variation of molecular properties such as molecular weights and number of rotatable bonds are related to the variation in T_c values, such properties are distributed over a wider range for the diversity-enhanced test set. The composition of the sets is summarized in Table 1.

(ii) CoMFA and CoMSIA. CoMFA[8] and CoMSIA[9] field-based descriptors were calculated with Sybyl using default parameters. A lattice with 2 Å grid spacing and extending at least 4 Å in each direction beyond the aligned molecules was used. For CoMSIA, separate models were derived using four combinations of fields: (1) steric and electrostatic, (2) steric, electrostatic, and

Table 1. Description of QSAR Data Sets

	ACE	AchE	BZR	COX2	DHFR	GPB	THER	THR
train[a]	76	74	98	188	237	44	51	59
test	38	37	49	94	124	22	25	29
inact			16	40	36			
train $\langle T_c \rangle$[b]	0.46	0.56	0.51	0.48	0.47	0.56	0.67	0.75
test $\langle T_c \rangle$	0.36	0.51	0.48	0.44	0.43	0.50	0.51	0.69
test−train[d] $\langle T_c \rangle$	0.40	0.53	0.49	0.45	0.44	0.51	0.56	0.71
train $\min\{T_c\}$[c]	0.90	0.93	0.92	0.92	0.91	0.90	0.94	0.94
test $\min\{T_c\}$	0.68	0.78	0.76	0.79	0.75	0.72	0.76	0.84
test−train $\min\{T_c\}$	0.82	0.88	0.86	0.87	0.86	0.92	0.82	0.91

[a] Number of training, test, inactive set compounds. [b] Average pairwise value of T_c for the set. [c] Average value of T_c calculated over pairs of most similar compounds. [d] Calculated using test set−training set pairs.

hydrophobic, (3) steric, electrostatic, and hydrogen-bonding, and (4) steric, electrostatic, hydrophobic, and hydrogen-bonding. The "minimum σ" value for removing descriptors with low variance was set to 2.0 for CoMFA and 1.0 for CoMSIA. Block scaling (CoMFA standard scaling) was applied to descriptors prior to QSAR analysis.

Net formal charges were determined by deprotonating carboxylic acids and phosphates and protonating non-aryl basic amines (except NH_2 groups that coordinate Zn in the ACE set), and scaled MNDO ESP-fit partial charges[32] were calculated with MOPAC 6.0 using atomic coordinates obtained by energy-minimizing the aligned molecules with the MMFF94S force field and MAXIMIN2 routine in Sybyl (200 steps, other parameters default). For the THER set, Gasteiger−Marsili charges[33] as implemented in Sybyl were used.

(iii) EVA. For EVA descriptor calculations[11] using Sybyl, all ionizable groups were neutralized. Initial structures were generated from SMILES strings using the CORINA program[34] available as a web-server at http://www2.chemie.uni-erlangen.de/software/ corina/free_struct.html. Normal mode calculations were performed with the AM1 Hamiltonian using parameters specified in the EVA_AM1.par file. Normal mode frequencies between 200 and 4000 cm^{-1} were used for defining EVA profiles. Several values of the resolution factor (σ) were used for representing "spectrum" absorptions: 1, 2, 4, 6, 8 10, 14, 18, 22, 26, 30. The interval width at which the profile is sampled (L) was set to $\sigma/2$. This is well below the maximum suggested values.[35] No filtering or scaling of descriptors was applied prior to QSAR analysis.

(iv) HQSAR. For HQSAR descriptor calculations,[5,6] all ionizable groups were neutralized. Holograms were generated with Sybyl using default path lengths for fragments (i.e., 4−7). The generation of molecular fragments is effected by distinguishing atom types (A), bond types (B), connectivity (Co), and chirality (C). In addition, hydrogens (H) can be considered for defining fragments. Guided by the study of Seel et al.,[36] four combinations of these parameters were considered: {ABCo}, {ABCoH}, {ABCoC}, and {ABCoHC}. Ideally, a particular fragment occurring in a given molecule would be represented by one position in the string of integers (hologram) that encodes the frequency at which each fragment occurs. Because different molecules will generate different fragments and different string lengths, a hashing procedure is used to give integer strings of fixed length. Seel et al. have found that this hashing procedure is detrimental for PLS modeling and recommend that unhashed fragment strings be used when possible, or at least long hashed strings for which model statistics are reasonably insensitive to hologram length. To minimize the chance of fragment collisions due to hashing (i.e., two or more fragments contributing to the same integer in the string), we use holograms of length 4999, much longer than the values suggested by the Sybyl interface. This results in sparse strings, with typical counts of descriptors having nonzero values over all training set compounds, being 1500−3500 for holograms excluding hydrogen and 3000−4500 for holograms including hydrogen in fragment generation. No filtering or scaling of descriptors was applied prior to QSAR analysis.

(v) 2D and 2.5D. "Traditional" descriptors were calculated using Cerius2. The states of ionizable groups were those used for field-based 3D QSAR. For strictly 2D descriptors, we used the "Combichem" defaults in Cerius2 (e.g., χ indices,[4] counts of rotatable bonds, or molecular weight, etc.) and E-state indices[37] (both sums of indices and counts for each atom type). In addition, we have calculated whole-molecule 3D descriptors such as molecular volume and charged partial surface area (CPSA) descriptors.[38] These are calculated using Gasteiger−Marsili charges[33] implemented in Cerius2 (the Polygraph set) and the CORINA structures[34] generated from SMILES strings. Because the charges and structures are determined with a straightforward and unambiguous approach, we refer to these as 2.5D descriptors. Some descriptors were removed by examining each training set separately. The first reduction eliminated descriptors having the same value for more than 90% of compounds. The second reduction eliminated one descriptor from each pair having a pairwise correlation coefficient r satisfying $|r| > 0.95$, retaining 2D descriptors over 2.5D descriptors and simple descriptors (e.g., molecular weight) over complex descriptors (e.g., information-content descriptors[39]). This reduced the initial set of 189 descriptors to 31−44 2D descriptors, or 56−75 2.5D descriptors. The descriptors were autoscaled (mean-centered and divided by the standard deviation) prior to QSAR analysis.

(vi) Model Derivations. All partial least squares (PLS) analyses were performed in Sybyl using default settings (except scaling/filtering of descriptors described above). For 2.5D descriptors, additional models were developed with the genetic function approximation (GFA) algorithm,[14] genetic PLS,[40] and back-propagation feed-forward neural networks[15,16] implemented in Cerius2.

For GFA analyses, descriptor subsets are selected with a genetic algorithm and fit using multiple linear regression (MLR). Models containing only linear terms

(GFA-l) and models with nonlinear terms (GFA-nl) were developed. Three combinations of nonlinear terms were considered for GFA-nl: (1) linear and quadratic terms, (2) linear and spline terms, and (3) linear, quadratic, and spline terms. Fixed-length models were generated, containing 2−13 terms (in addition to the regression constant). For GPLS analyses, descriptors are selected with a genetic algorithm, but PLS is used for fitting models instead of MLR. Each combination of the following parameters was examined systematically: 9, 14, 19, 24 descriptors (or 10, 15, 20, 25 terms including the constant) and 1−8 PLS components. For GFA and GPLS, 300 individuals, 10 000 crossover operations, a 10% mutation rate, and 50% spline knot shift rate were used. Except for the spline shift rate, these nondefault parameters are used by one developer of the GFA algorithm.[41] In addition to the best-ranked model from each population, we considered ensemble models in which the average prediction from the 200 fittest individuals is used for compiling test set statistics.

Back-propagation feed-forward neural network (NN) models were developed using the descriptors selected by GFA-l applied to the complete training set (i.e., not cross-validated models; descriptors are listed in the Supporting Information). Because of the large number of adjustable connections in NN models, it is a common practice to use a variable selection method prior to network training.[25,42] Networks have the architecture: input nodes = number of descriptors, x hidden layer nodes, 1 output node. The number of hidden layer nodes was varied between 2 and 1 less than the number of input nodes. The networks were trained using BFGS minimization of connection weights with default parameters. A 10% holdout sample randomly selected from the training set was used to decide when to stop training the network. Initial connection weights were randomly assigned to values between −0.5 and 0.5. Training was halted after 4000 epochs or if the rms error of prediction on the holdout set had not decreased during the previous 300 epochs. The first model was always generated with the random number generator seed 1 969 530 170, and the "update seed" option was used for subsequent networks to allow reproducibility of results. In addition to a single network, an ensemble of 10 networks from which predictions were averaged was used for compiling test set statistics.

(vii) Assessment of Predictive Accuracy. All combinations of parameters (e.g., each combination of fields for CoMSIA) and model complexity (e.g., number of PLS components, GFA terms, etc.) were examined systematically by cross-validation. For PLS models, several procedures were used. "Leave-one-out" (LOO) CV was performed with the SAMPLS routine.[43] In addition, "leave-$^1/_{10}$-out" (L10%O), "leave-$^1/_5$-out" (L20%O), and "leave-$^1/_3$-out" (L33%O) were performed. For L10%O CV, training set compounds are divided into 10 groups. Each group is excluded in turn and predicted from models fit using the other nine groups. This was repeated five times, giving estimates of predictive accuracy calculated from 50 models. For L20%O and L33%O, 10 and 20 cycles were performed, yielding 50 and 60 models, respectively. Because of the high computational cost of repeating descriptor selection for each CV model, we use only the L10%O procedure for the GFA, GPLS, and NN methods. The optimal combination of parameters and model complexity was chosen as that which minimizes the value of s_{PRESS}. s_{PRESS} is calculated by dividing the sum of squared prediction residuals by $N - A - 1$ rather than $N - 1$ as in the standard error; inclusion of the number A of PLS components, GFA terms, or NN hidden layer nodes has the effect of penalizing larger models. Occasionally, s_{PRESS} reaches a first minimum, rises, then falls again with increasing complexity; the first minimum was selected as the optimal complexity.[44] It is noted that values of s_{cv} reported in the Results are not s_{PRESS} but the standard error of cross-validation predictions (i.e., normalized by $N - 1$).

For the combination of parameters and model complexity deemed most predictive from cross-validation, final models were developed from the full training set and used to make predictions for the test set. Some workers use the average activity from the training set to calculate r^2_{test}, while others use the average from the test set. In the present work, we use the average training set activity. Because of the design procedure used to assemble test sets, differences between the average activity calculated from the training and test sets are small.

For the BZR, COX2, and DHFR sets, several inactive compounds were used to assess if QSAR models identify them as low-activity compounds. Compounds in the inactive set were deemed correctly classified by models if the predicted activity was less than the average training set activity (pIC_{50} values of 7.89, 6.98, and 6.23 for the BZR, COX2, and DHFR sets, respectively). While these thresholds are arbitrary, they were selected on the basis that compounds predicted to have less than average activity are likely to be of little interest for follow-up synthesis and screening. The upper bound for activity reported in the literature (e.g., pIC_{50} = 6, for $IC_{50} > 1\ \mu$M) is lower than the selected threshold for all compounds.

Results

Statistical analyses for seven descriptor sets and five model-building methods are outlined as follows. First, we describe PLS models using descriptors from CoMFA, CoMSIA with steric and electrostatic fields (CoMSIA basic), CoMSIA with additional fields (CoMSIA extra), EVA, HQSAR, and traditional 2D and 2.5D encoding of molecular structures. Second, we describe additional models derived using 2.5D descriptors and GFA with linear terms only (GFA-l), GFA with nonlinear terms (GFA-nl), GPLS, and NN. For brevity, a reference to "2.5D" implies the use of PLS, while "2.5D-GFA-l" indicates that the model was obtained using the GFA algorithm with linear terms.

(i) Cross-Validated Predictive Accuracy. For all methods, the optimal combination of parameters for calculating descriptors and model complexity was determined using cross-validation (CV). For CoMFA, CoMSIA basic, 2D, and 2.5D, we have not varied parameters for calculating descriptors. For CoMSIA extra, we examined the use of hydrogen-bonding fields, hydrophobic fields, or both in addition to steric and electrostatic fields. For EVA, we examined several sets of $\{\sigma,L\}$ values (Methods). For HQSAR, we considered

Table 2. Cross-Validation Statistics for PLS Analyses

	CoMFA	CoMSIA basic	CoMSIA extra	EVA	HQSAR	2D	2.5D
			ACE				
param[a]	3	3	2/pho	4/18	4/HC	3	4
q^2_{LOO}	0.68	0.65	0.66	0.70	0.72	0.68	0.72
$q^2_{\mathrm{LOO,no\ out}}$[b]	0.76 (3)	0.71 (2)	0.72 (2)	0.77 (3)	0.79 (2)	0.74 (3)	0.76 (3)
$q^2_{\mathrm{L10\%O}}$	0.69	0.66	0.67	0.70	0.72	0.68	0.72
$s_{\mathrm{cv,L10\%O}}$	1.32	1.38	1.36	1.29	1.24	1.32	1.24
			AchE				
param[a]	5	6	4/all	4/2	5/H	1	1
q^2_{LOO}	0.52	0.48	0.49	0.42	0.34	0.32	0.31
$q^2_{\mathrm{LOO,no\ out}}$[b]	0.62 (3)	0.58 (3)	0.56 (2)	0.49 (2)	0.49 (3)	0.35 (3)	0.35 (3)
$q^2_{\mathrm{L10\%O}}$	0.52	0.45	0.46	0.41	0.33	0.32	0.30
$s_{\mathrm{cv,L10\%O}}$	0.84	0.90	0.89	0.94	1.00	1.01	1.02
			BZR				
param[a]	3	3	3/pho	2/22	4/C	3	3
q^2_{LOO}	0.32	0.41	0.45	0.40	0.42	0.36	0.35
$q^2_{\mathrm{LOO,no\ out}}$[b]	0.50 (4)	0.49 (4)	0.53 (4)	0.46 (3)	0.50 (5)	0.42 (3)	0.44 (3)
$q^2_{\mathrm{L10\%O}}$	0.32	0.40	0.45	0.39	0.39	0.37	0.34
$s_{\mathrm{cv,L10\%O}}$	0.55	0.52	0.49	0.52	0.52	0.53	0.54
			COX2				
param[a]	5	6	4/all	5/14	7/-	7	7
q^2_{LOO}	0.49	0.43	0.57	0.45	0.50	0.49	0.55
$q^2_{\mathrm{LOO,no\ out}}$[b]	0.59 (9)	0.59 (9)	0.68 (9)	0.63 (13)	0.60 (9)	0.61 (9)	0.64 (9)
$q^2_{\mathrm{L10\%O}}$	0.48	0.43	0.56	0.45	0.49	0.48	0.52
$s_{\mathrm{cv,L10\%O}}$	0.74	0.77	0.68	0.76	0.73	0.74	0.71
			DHFR				
param[a]	5	5	4/all	9/18	6/HC	6	6
q^2_{LOO}	0.65	0.63	0.65	0.64	0.69	0.51	0.53
$q^2_{\mathrm{LOO,no\ out}}$[b]	0.74 (10)	0.73 (11)	0.76 (12)	0.76 (12)	0.78 (13)	0.63 (11)	0.64 (11)
$q^2_{\mathrm{L10\%O}}$	0.65	0.64	0.65	0.64	0.69	0.51	0.52
$s_{\mathrm{cv,L10\%O}}$	0.76	0.76	0.76	0.76	0.71	0.89	0.88
			GPB				
param[a]	4	4	4/hyd	3/14	2/-	2	3
q^2_{LOO}	0.42	0.43	0.61	0.58	0.66	0.31	0.46
$q^2_{\mathrm{LOO,no\ out}}$[b]	0.51 (2)	0.40 (1)	0.67 (1)	0.68 (2)	0.66 (0)	0.39 (2)	0.57 (2)
$q^2_{\mathrm{L10\%O}}$	0.47	0.36	0.62	0.56	0.66	0.27	0.42
$s_{\mathrm{cv,L10\%O}}$	0.79	0.86	0.66	0.71	0.63	0.92	0.82
			THER				
param[a]	4	6	3/hyd	4/10	4/-	4	5
q^2_{LOO}	0.52	0.54	0.51	0.48	0.49	0.62	0.66
$q^2_{\mathrm{LOO,no\ out}}$[b]	0.54 (1)	0.54 (1)	0.58 (2)	0.57 (2)	0.59 (2)	0.73 (4)	0.68 (2)
$q^2_{\mathrm{L10\%O}}$	0.49	0.49	0.50	0.43	0.47	0.62	0.65
$s_{\mathrm{cv,L10\%O}}$	1.36	1.36	1.35	1.44	1.38	1.17	1.12
			THR				
param[a]	4	5	4/all	4/6	6/H	6	4
q^2_{LOO}	0.59	0.62	0.72	0.47	0.50	0.62	0.52
$q^2_{\mathrm{LOO,no\ out}}$[b]	0.66 (4)	0.72 (3)	0.81 (4)	0.56 (4)	0.61 (3)	0.73 (3)	0.64 (3)
$q^2_{\mathrm{L10\%O}}$	0.50	0.52	0.66	0.45	0.40	0.55	0.45
$s_{\mathrm{cv,L10\%O}}$	0.68	0.67	0.56	0.71	0.74	0.65	0.71

[a] Number of PLS components that minimizes s_{PRESS}. For CoMSIA extra, pho, hyd, and all indicate the use of hydro**pho**bic fields, **hyd**rogen-bonding fields, and **all** types, in addition to steric and electrostatic fields. For HQSAR, C and H indicate the use of chirality and hydrogens for defining holograms. [b] q^2 excluding the indicated number of outliers.

the combinations {ABCo}, {ABCoC}, {ABCoH}, and {ABCoHC} (Methods). Model complexity refers to the number of PLS or GPLS components, the number of GFA or GPLS terms, and the number of NN hidden-layer nodes. For PLS models, all parameter or model complexity combinations were examined systematically using both leave-one-out (LOO) and leave-$^1/_{10}$-out (L10%O) CV.

The optimal combination of descriptor parameters and number of PLS components identified using LOO and L10%O CV were identical in nearly all cases. For CoMSIA extra on the AchE set, LOO CV minimizes s_{PRESS} at five components ($q^2 = 0.51$) while L10%O CV minimizes s_{PRESS} at four components. Both identify the use of all CoMSIA fields as most predictive. For CoMSIA basic on the DHFR set, LOO CV minimizes s_{PRESS} at six components ($q^2 = 0.64$) while L10%O CV minimizes s_{PRESS} at five components. For 2D on the THER set, L10%O CV minimizes s_{PRESS} at five components ($q^2 = 0.64$) while LOO CV minimizes s_{PRESS} at four components. Finally, for 2.5D on the THER set, LOO CV minimizes s_{PRESS} at six components ($q^2 = 0.67$) while L10%O CV minimizes s_{PRESS} at five components. The smaller number of components is used for all subsequent analyses.

Statistics are reported in Table 2 for the most predictive combination of parameters and model complexity. To assess the effect of outliers on the value of q^2_{LOO}, compounds with residuals more than 2 standard deviations from the average residual were identified and excluded from the calculation. Small adjustments (most by less than 0.05) were made to the threshold such that compounds with residuals near 2σ are included or excluded depending on the number and identity of

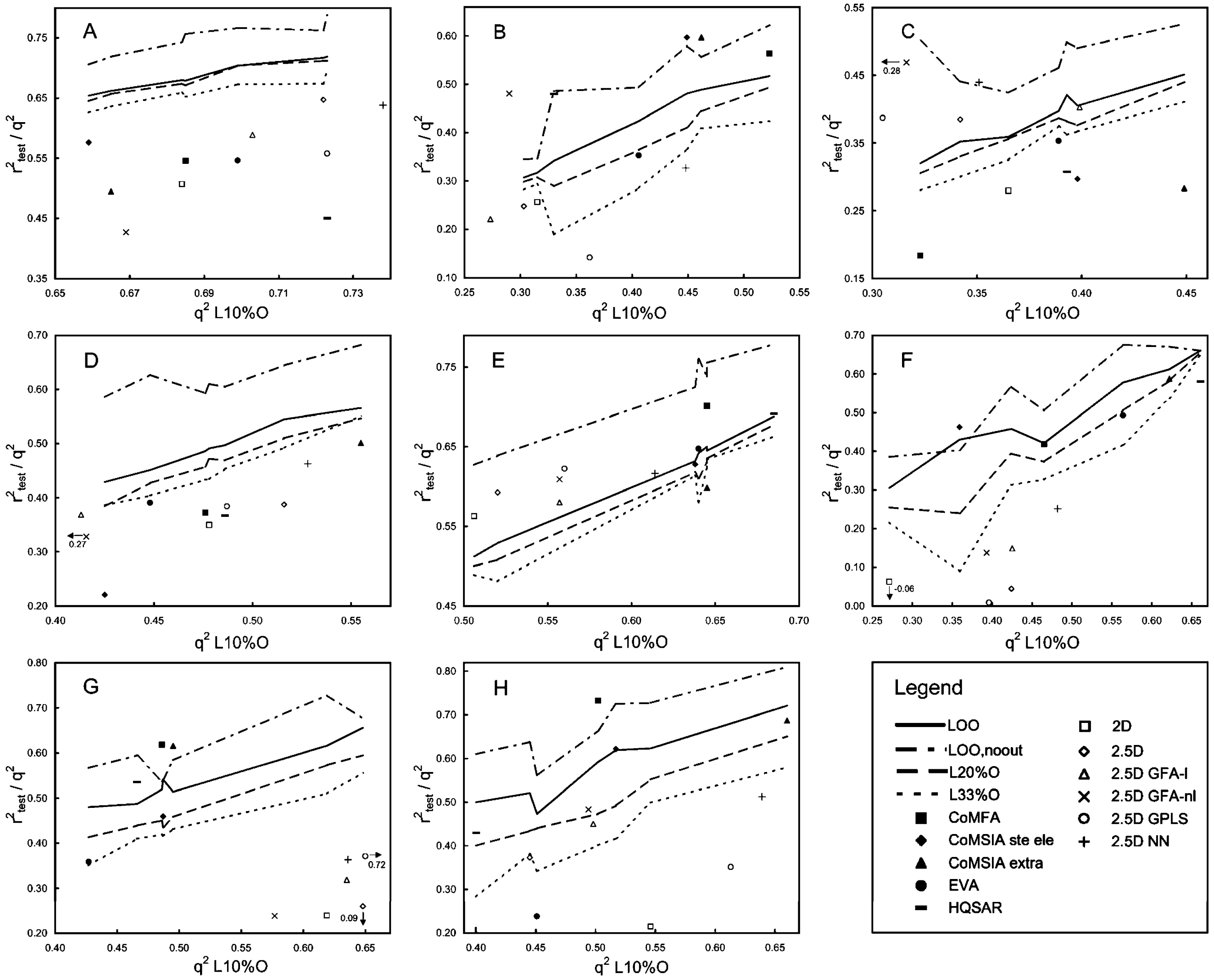

Figure 3. Comparison of correlation coefficients from L10%O CV with those from LOO, LOO without outliers, L20%O, and L33%O CV (lines). Comparisons of $q^2_{\mathrm{L10\%O}}$ and $r^2_{\mathrm{test,no\ out}}$ are plotted with various symbols. For the GPB set, r^2_{test} is shown instead of $r^2_{\mathrm{test,no\ out}}$. (A) ACE; (B) ACHE; (C) BZR; (D) COX2; (E) DHFR; (F) GPB; (G) THER; (H) THR.

outliers for related methods. For example, a small adjustment might be made to the threshold for CoMSIA basic such that a CoMFA outlier is also defined as a CoMSIA basic outlier. Thresholds used for defining outliers are indicated in the Supporting Information, in addition to lists of (test set) outliers for each combination of method and data set.

For leave-several-out CV, it is possible to use various group sizes. We have repeated PLS analyses using the optimal parameters/complexity with leave-$^1/_5$-out (L20%O) CV and leave-$^1/_3$-out (L33%O) CV. The values of q^2 obtained using these larger group sizes are compared to those from L10%O CV in Figure 3. Tables indicating values of L20%O and L33%O are including in the Supporting Information.

For the ACE set, all methods have similar values of q^2. The field-based methods perform substantially better for the AchE set. Before outlier removal, all 2D methods perform poorly. However, the value of q^2 for HQSAR improves substantially after removal of three outliers (most of the difference is accounted for by 2-3M and 2-12 that have LOO residuals of ~3). For the BZR set, the use of hydrophobic fields in CoMSIA yields the best results, with HQSAR only slightly less predictive. It is noted for the BZR set that s_{CV} values fall in a narrow range. As such, q^2 values are sensitive to small differences among methods. For the COX2 set, CoMSIA with all fields is somewhat more predictive than 2.5D and HQSAR; all methods perform reasonably well with differences further reduced after outlier removal. Except for 2D and 2.5D, all methods give high q^2 values for the DHFR set, with HQSAR performing best. For the GPB set, HQSAR performs best followed by CoMSIA with hydrogen-bonding fields. CoMFA and CoMSIA basic perform no better than 2.5D. For the THER set, 2.5D performs best while HQSAR is comparable to field-based methods. CoMSIA with all fields performs best for the THR set; CoMFA and CoMSIA basic perform no better than 2D.

Using 2.5D descriptors and the other model-building methods, only L10%O CV was performed because of the high computational cost of repeated GFA and GPLS runs for each combination of parameters and complexity. For GFA-nl, models were developed using linear and quadratic terms, linear and spline terms, or linear, quadratic, and spline terms simultaneously. Statistics are shown only for the most predictive combination (Table 3).

Table 3. Cross-Validation Statistics for Various Model-Building Methods Using 2.5D Descriptors

	PLS	GFA-l	GFA-nl	GPLS	NN
			ACE		
param[a]	4	7	2/lqs	9/1	7/2
$q^2_{L10\%O}$	0.72	0.70	0.67	0.72	0.74
$s_{cv,L10\%O}$	1.24	1.28	1.36	1.24	1.20
			AchE		
param[a]	1	5	5/ls	14/1	5/2
$q^2_{L10\%O}$	0.30	0.27	0.29	0.36	0.45
$s_{cv,L10\%O}$	1.02	1.04	1.03	0.97	0.90
			BZR		
param[a]	3	8	7/lqs	19/1	8/4
$q^2_{L10\%O}$	0.34	0.40	0.26	0.31	0.35
$s_{cv,L10\%O}$	0.54	0.51	0.57	0.55	0.54
			COX2		
param[a]	7	5	3/lqs	24/3	5/2
$q^2_{L10\%O}$	0.52	0.41	0.27	0.49	0.53
$s_{cv,L10\%O}$	0.71	0.79	0.88	0.73	0.70
			DHFR		
param[a]	6	10	8/ls	24/7	10/6
$q^2_{L10\%O}$	0.52	0.56	0.56	0.56	0.61
$s_{cv,L10\%O}$	0.88	0.85	0.85	0.84	0.79
			GPB		
param[a]	3	3	3/lq	24/2	3/2
$q^2_{L10\%O}$	0.42	0.43	0.39	0.40	0.48
$s_{cv,L10\%O}$	0.82	0.82	0.84	0.84	0.78
			THER		
param[a]	5	5	5/lq	9/5	5/3
$q^2_{L10\%O}$	0.65	0.64	0.58	0.72	0.64
$s_{cv,L10\%O}$	1.12	1.14	1.23	0.99	1.20
			THR		
param[a]	4	8	3/lqs	9/3	8/2
$q^2_{L10\%O}$	0.45	0.50	0.49	0.61	0.64
$s_{cv,L10\%O}$	0.71	0.68	0.68	0.60	0.58

[a] Number of GFA terms or PLS components. For GFA-nl, the nature of allowed terms is indicated: **l**inear, **q**uadratic, and **s**plines. For GPLS, the number of terms is followed by the number of PLS components. For NN, the number of input and hidden-layer nodes is given.

From Table 3, it emerges that no method is consistently better than PLS. GFA with linear terms is only substantially more predictive for the BZR and THR sets but not better than 2D-PLS for THR and substantially worse for COX2. The inclusion of nonlinear terms in GFA does not yield higher q^2 values, although smaller models with similar predictive accuracy are obtained for the ACE, DHFR, and THR sets. GPLS performs better (AchE, THER, THR) or similarly compared to PLS. It may be useful in situations where "noisy" variables are deleterious for PLS modeling. For NN models, higher q^2 values are obtained for most sets. However, it must be noted that the full training set was used in the GFA-l analysis from which the input descriptors were selected for NN model derivation. As such, the compounds left out in each CV test set have an influence on NN models. The q^2 values for NN models are always similar (COX2) or lower (all other sets) than those obtained if multiple linear regression is used with the same subset of descriptors (i.e., what the GFA developers refer to as regression-only or partial cross-validation[40]).

(ii) Test Set Predictive Accuracy. Having identified the optimal parameters and complexity by cross-validation, final models were developed using the complete training sets. Their predictive accuracy was assessed using the designed test sets. Because of its quadratic dependence on residuals, the value of r^2_{test} can be significantly affected by a few outliers. For this reason, we rely primarily on values of r^2_{test} after outlier removal to assess the general predictive accuracy of methods while keeping in mind the relative number of outliers. However, for the GPB set we use values of r^2_{test} before removal of the outlier **57** (glucopyranose spirohydantoin; cf. Figure 1) because it is the most interesting compound in the series. Since the same test sets are used for all comparisons, r^2 values increase monotonically with decreasing standard errors of prediction, and identical conclusions would be reached by giving primary consideration to the latter. Statistics for PLS applied with various descriptor sets and various model-building methods applied with 2.5D descriptors are summarized in Tables 4 and 5. As for cross-validation, small adjustments were made to thresholds for defining outliers such that the number and identity of outliers are consistent among similar QSAR methods.

Except for HQSAR, all methods perform well on the ACE test set. For the AchE test set, the three field-based methods perform similarly, followed by HQSAR. EVA, 2D, and 2.5D perform poorly. For the BZR test set, CoMFA performs very poorly; 2.5D performs best but not well enough to be considered useful. As noted for the cross-validation results, the narrow range of s_{test} values (excluding CoMFA) must be kept in mind. Also, field-based QSAR methods and EVA perform inadequately in the classification of inactive compounds. For the COX2 test set, CoMSIA with all fields is most predictive, and other methods excluding CoMSIA basic perform similarly. Only CoMSIA with all fields and 2.5D perform acceptably for classification of inactive compounds. HQSAR and CoMFA perform similarly on the DHFR test set; 2D and 2.5D perform worst but are nonetheless reasonably predictive. The field-based and HQSAR models classify more than 90% of inactives correctly; other methods perform reasonably well. For the GPB test set, HQSAR and CoMSIA with hydrogen-bond fields are most predictive; 2D and 2.5D are useless. The decrease in r^2_{test} values upon removal of **57** arises because its deviation from the average activity is greater than the prediction residuals. For the THER set, field-based methods perform somewhat better than HQSAR and EVA; 2D and 2.5D perform poorly. Field-based QSAR methods perform well for the THR test set, with CoMFA performing best; EVA and 2D produce useless models.

In addition to PLS, other model-building methods were applied with 2.5D descriptors. For the ACE test set, PLS performs better than any other method. For the AchE test set, GFA-nl performs substantially better, and NN-ens performs somewhat better than PLS. For the BZR test set, GFA-nl and NN-ens perform better than PLS. It may appear that GFA-l is more predictive than PLS; however, this is probably fortuitous because the ensemble model performs only slightly better than PLS (lowering the outlier threshold for GFA-l-ens to obtain three outliers leaves the value of $r^2_{test,no\ out}$ unchanged). Only NN and NN-ens perform substantially better than PLS for the COX2 test set. For the DHFR test set, all methods produce predictive models having similar accuracy, and all methods perform poorly for the GPB test set. For the THER set, all methods perform better than PLS although no model can be

Table 4. Training and Test Set Statistics for PLS Analyses

	CoMFA	CoMSIA basic	CoMSIA extra	EVA	HQSAR	2D	2.5D
ACE							
r^2_{train}	0.80	0.76	0.73	0.84	0.84	0.76	0.82
s_{train}	1.04	1.15	1.22	0.93	0.95	1.15	1.00
r^2_{test}	0.49	0.52	0.49	0.36	0.30	0.47	0.51
s_{test}	1.54	1.48	1.53	1.72	1.80	1.57	1.50
$r^2_{\text{test,no out}}$[a]	0.55 (1)	0.58 (1)	0.49 (0)	0.55 (2)	0.45 (2)	0.51 (1)	0.65 (2)
$s_{\text{test,no out}}$	1.47	1.41	1.53	1.44	1.64	1.52	1.31
AchE							
r^2_{train}	0.88	0.86	0.86	0.96	0.72	0.40	0.38
s_{train}	0.41	0.45	0.45	0.23	0.64	0.94	0.95
r^2_{test}	0.47	0.44	0.44	0.28	0.37[c]	0.16	0.16
s_{test}	0.95	0.98	0.98	1.11	1.01[c]	1.20	1.20
$r^2_{\text{test,no out}}$[a]	0.56 (1)	0.60 (1)	0.60 (1)	0.35 (1)	0.48 (2)	0.26 (1)	0.25 (2)
$s_{\text{test,no out}}$	0.87	0.81	0.81	1.05	0.92	1.09	1.04
BZR							
r^2_{train}	0.61	0.62	0.62	0.51	0.64	0.51	0.52
s_{train}	0.41	0.41	0.41	0.47	0.40	0.46	0.46
r^2_{test}	0.00	0.08	0.12	0.16	0.17	0.14	0.20
s_{test}	0.97	0.93	0.91	0.89	0.88	0.90	0.87
$r^2_{\text{test,no out}}$[a]	0.18 (3)	0.30 (3)	0.28 (3)	0.35 (3)	0.31 (2)	0.28 (3)	0.38 (3)
$s_{\text{test,no out}}$	0.81	0.75	0.75	0.72	0.74	0.76	0.71
% class inact[b]	69	63	63	63	75	88	88
COX2							
r^2_{train}	0.70	0.69	0.69	0.68	0.70	0.62	0.68
s_{train}	0.56	0.56	0.57	0.58	0.55	0.63	0.58
r^2_{test}	0.29	0.03	0.37	0.17	0.27	0.25	0.27
s_{test}	1.24	1.44	1.17	1.33	1.26	1.27	1.25
$r^2_{\text{test,no out}}$[a]	0.37 (5)	0.22 (5)	0.50 (4)	0.39 (5)	0.37 (5)	0.35 (5)	0.39 (5)
$s_{\text{test,no out}}$	1.09	1.20	0.99	1.08	1.07	1.11	1.11
% class inact[b]	65	63	70	63	53	58	70
DHFR							
r^2_{train}	0.79	0.76	0.75	0.81	0.81	0.61	0.65
s_{train}	0.59	0.62	0.63	0.55	0.55	0.79	0.75
r^2_{test}	0.59	0.52	0.53	0.57	0.63	0.47	0.49
s_{test}	0.89	0.96	0.95	0.90	0.84	1.00	0.99
$r^2_{\text{test,no out}}$[a]	0.70 (6)	0.63 (6)	0.60 (6)	0.65 (6)	0.69 (6)	0.56 (5)	0.59 (6)
$s_{\text{test,no out}}$	0.73	0.81	0.84	0.82	0.75	0.88	0.85
% class inact[b]	92	92	97	83	92	75	81
GPB							
r^2_{train}	0.84	0.78	0.92	0.89	0.77	0.55	0.70
s_{train}	0.43	0.50	0.30	0.36	0.52	0.72	0.59
r^2_{test}	0.42	0.46	0.59	0.49	0.58	−0.06	0.04
s_{test}	0.94	0.90	0.79	0.88	0.80	1.27	1.20
$r^2_{\text{test,no out}}$[a]	0.37 (1)	0.34 (1)	0.37 (1)	0.34 (1)	0.34 (1)	−0.06 (0)	0.04 (0)
$s_{\text{test,no out}}$	0.70	0.82	0.70	0.72	0.72	1.27	1.20
THER							
r^2_{train}	0.85	0.85	0.77	0.86	0.81	0.79	0.85
s_{train}	0.73	0.73	0.91	0.72	0.82	0.86	0.73
r^2_{test}	0.54	0.36	0.53	0.36	0.53	0.14	0.07
s_{test}	1.59	1.87	1.60	1.87	1.59	2.16	2.24
$r^2_{\text{test,no out}}$[a]	0.62 (1)	0.46 (1)	0.62 (2)	0.36 (0)	0.54 (1)	0.24 (1)	0.09 (1)
$s_{\text{test,no out}}$	1.34	1.60	1.33	1.87	1.48	1.90	2.07
THR							
r^2_{train}	0.86	0.88	0.89	0.83	0.87	0.79	0.75
s_{train}	0.36	0.34	0.32	0.39	0.35	0.43	0.47
r^2_{test}	0.63	0.55	0.63	0.11	−0.25	0.04	0.28
s_{test}	0.70	0.76	0.69	1.08	1.27	1.12	0.96
$r^2_{\text{test,no out}}$[a]	0.73 (1)	0.62 (1)	0.69 (1)	0.24 (1)	0.43 (3)	0.21 (1)	0.37 (1)
$s_{\text{test,no out}}$	0.56	0.66	0.60	0.96	0.83	1.01	0.87

[a] Number of outliers is indicated in parentheses. [b] Percentage of inactives correctly classified. [c] AchE inhibitor 2-36 is excluded because its predicted activity exceeds its measured activity by more than 12 pIC_{50} units. It is also excluded from the calculation of the average and standard deviation of test set residuals, although is it counted among the two residuals listed.

considered predictive. For the THR set, NN and NN-ens are most predictive, with GFA-l and GFA-nl exceeding the predictive accuracy of PLS only after the removal of outliers. Excluding GFA-nl and NN-ens applied to the DHFR set, PLS does as well as any other method for classifying inactives.

The removal of highly correlated descriptors is not necessary for PLS analysis, since descriptors are reduced to a series of uncorrelated latent variables. In this work, we use the reduced descriptor set for all analyses (e.g., all pairs of descriptors satisfy $|r| \leq 0.95$). The retention of highly correlated 2.5D descriptors for PLS gave models having essentially the same predictive accuracy. Differences in r^2_{test} when retaining all descriptors range from −0.02 (DHFR) to 0.04 (AchE).

Table 5. Training and Test Set Statistics for Various Model-Building Methods Using 2.5D Descriptors

	PLS	GFA-l	GFA-l-ens	GFA-nl	GFA-nl-ens	GPLS	GPLS-ens	NN	NN-ens
ACE									
r^2_{train}	0.82	0.83	0.83	0.77	0.78	0.78	0.78	0.78	0.84
s_{train}	1.00	0.96	0.97	1.13	1.10	1.11	1.10	1.11	0.93
r^2_{test}	0.51	0.49	0.50	0.39	0.32	0.45	0.43	0.39	0.51
s_{test}	1.50	1.53	1.51	1.68	1.77	1.60	1.62	1.68	1.51
$r^2_{\mathrm{test,no\ out}}$[a]	0.65 (2)	0.61 (2)	0.59 (2)	0.45 (1)	0.43 (2)	0.59 (2)	0.56 (2)	0.44 (1)	0.64 (2)
$s_{\mathrm{test,no\ out}}$	1.31	1.37	1.37	1.61	1.61	1.41	1.46	1.58	1.27
AchE									
r^2_{train}	0.38	0.60	0.60	0.68	0.71	0.58	0.58	0.68	0.63
s_{train}	0.95	0.77	0.77	0.69	0.66	0.79	0.79	0.69	0.74
r^2_{test}	0.16	0.16	0.22	0.29	0.40	0.13	0.14	−0.04	0.21
s_{test}	1.20	1.20	1.15	1.10	1.01	1.22	1.21	1.34	1.16
$r^2_{\mathrm{test,no\ out}}$[a]	0.25 (2)	0.28 (2)	0.22 (0)	0.48 (2)	0.48 (2)	0.13 (0)	0.14 (0)	0.15 (1)	0.33 (2)
$s_{\mathrm{test,no\ out}}$	1.04	1.10	1.15	0.88	0.88	1.22	1.21	1.18	1.00
BZR									
r^2_{train}	0.52	0.63	0.62	0.64	0.64	0.52	0.52	0.62	0.66
s_{train}	0.46	0.40	0.41	0.40	0.40	0.46	0.46	0.41	0.39
r^2_{test}	0.20	0.22	0.21	0.20	0.20	0.20	0.18	0.39	0.34
s_{test}	0.87	0.86	0.86	0.87	0.87	0.87	0.88	0.76	0.79
$r^2_{\mathrm{test,no\ out}}$[a]	0.38 (3)	0.49 (3)	0.40 (2)	0.46 (3)	0.47 (3)	0.41 (4)	0.39 (4)	0.46 (2)	0.44 (3)
$s_{\mathrm{test,no\ out}}$	0.71	0.67	0.76	0.69	0.68	0.69	0.70	0.65	0.67
% class inact[b]	88	88	81	81	81	88	88	81	81
COX2									
r^2_{train}	0.68	0.58	0.59	0.54	0.54	0.64	0.64	0.65	0.65
s_{train}	0.58	0.66	0.66	0.69	0.69	0.61	0.61	0.60	0.60
r^2_{test}	0.27	0.28	0.25	0.12	0.13	0.28	0.26	0.31	0.32
s_{test}	1.25	1.24	1.27	1.37	1.37	1.24	1.26	1.22	1.21
$r^2_{\mathrm{test,no\ out}}$[a]	0.39 (5)	0.41 (6)	0.37 (5)	0.32 (5)	0.33 (5)	0.39 (5)	0.38 (5)	0.46 (5)	0.46 (5)
$s_{\mathrm{test,no\ out}}$	1.11	1.09	1.07	1.12	1.11	1.10	1.11	0.99	0.99
% class inact[b]	70	63	63	60	60	65	55	70	70
DHFR									
r^2_{train}	0.65	0.68	0.68	0.65	0.65	0.65	0.73	0.78	0.79
s_{train}	0.75	0.72	0.72	0.75	0.75	0.75	0.66	0.59	0.59
r^2_{test}	0.49	0.46	0.48	0.50	0.53	0.49	0.53	0.42	0.54
s_{test}	0.99	1.01	1.00	0.98	0.95	0.99	0.94	1.05	0.94
$r^2_{\mathrm{test,no\ out}}$[a]	0.59 (6)	0.58 (6)	0.58 (6)	0.60 (7)	0.61 (6)	0.59 (6)	0.62 (6)	0.56 (6)	0.62 (6)
$s_{\mathrm{test,no\ out}}$	0.85	0.87	0.86	0.83	0.83	0.85	0.82	0.91	0.83
% class inact[b]	81	78	81	86	86	81	92	83	92
GPB									
r^2_{train}	0.70	0.67	0.64	0.69	0.67	0.81	0.81	0.73	0.74
s_{train}	0.59	0.62	0.65	0.60	0.62	0.47	0.47	0.56	0.55
r^2_{test}	0.04	−0.02	0.15	−0.08	0.14	0.04	0.01	0.28	0.25
s_{test}	1.20	1.25	1.14	1.28	1.14	1.21	1.23	1.05	1.07
$r^2_{\mathrm{test,no\ out}}$[a]	0.04 (0)	−0.61 (1)	−0.28 (1)	−0.74 (1)	0.35 (1)	0.04 (0)	0.01 (0)	−0.14 (1)	−0.10 (1)
$s_{\mathrm{test,no\ out}}$	1.20	1.12	1.00	1.16	1.01	1.21	1.23	0.94	0.93
THER									
r^2_{train}	0.85	0.82	0.82	0.81	0.82	0.88	0.88	0.83	0.86
s_{train}	0.73	0.80	0.80	0.82	0.80	0.66	0.67	0.79	0.71
r^2_{test}	0.07	0.20	0.30	0.16	0.21	0.33	0.33	0.16	0.19
s_{test}	2.24	2.08	1.95	2.13	2.07	1.91	1.90	2.13	2.10
$r^2_{\mathrm{test,no\ out}}$[a]	0.09 (1)	0.38 (1)	0.32 (1)	0.21 (1)	0.24 (1)	0.36 (1)	0.37 (1)	0.35 (1)	0.36 (1)
$s_{\mathrm{test,no\ out}}$	2.07	1.82	1.79	1.93	1.90	1.75	1.72	1.86	1.84
THR									
r^2_{train}	0.75	0.84	0.82	0.71	0.70	0.79	0.79	0.86	0.84
s_{train}	0.47	0.39	0.40	0.52	0.53	0.43	0.44	0.36	0.38
r^2_{test}	0.28	0.13	0.27	0.11	0.12	0.11	0.16	0.26	0.23
s_{test}	0.96	1.06	0.97	1.07	1.07	1.08	1.04	0.98	1.00
$r^2_{\mathrm{test,no\ out}}$[a]	0.40 (2)[c]	0.46 (2)	0.45 (2)	0.47 (3)	0.48 (3)	0.30 (2)	0.35 (2)	0.51 (2)	0.51 (2)
$s_{\mathrm{test,no\ out}}$	0.86	0.81	0.82	0.75	0.74	0.93	0.89	0.75	0.77

[a] Number of outliers is indicated in parentheses. [b] Percentage of inactives correctly classified. [c] For 2.5D PLS on THR, a lower threshold was used than in Table 4 in order that outlier **10** excluded by most other methods is excluded for PLS.

Discussion

The predictive accuracy of several QSAR methods was examined using eight data sets. Physicochemical properties of compounds were encoded using three field-based methods (CoMFA, CoMSIA with steric and electrostatic fields, and CoMSIA with additional fields), EVA, HQSAR, and traditional descriptors (2D and 2.5D). QSAR methods were developed using partial least squares (PLS). For 2.5D descriptors, additional models were developed with the genetic function approximation (GFA) algorithm (linear terms only or linear and nonlinear terms), genetic PLS, and feed-forward back-propagation neural networks; both single models and ensembles of models were considered.

The predictive accuracy of models was examined from two perspectives. First, cross-validation (CV) was used to assess the ability of models to generalize. This can be regarded as a measure of accuracy for interpolation. Second, the accuracy of methods in making predictions for test set compounds was used to assess their ability to extrapolate. Because of the design procedure employed for assigning compounds to training and test sets, the latter is enriched in structural outliers. To be useful in prioritizing the synthesis of novel derivatives, models must have reasonable extrapolative accuracy in addition to interpolative accuracy. For the purpose of comparing the various QSAR methods, we define a "useful" model as one that gives q^2 or r^2_{test} roughly equal to or greater than 0.50 after outlier removal.

From cross-validation of PLS models, it emerges that all methods perform reasonably well for most sets; all methods produce models that are useful for interpolation for the ACE, COX2, DHFR, THER, and THR sets. Some trends can be noted.

(1) Except for the BZR set for which CoMSIA basic (i.e., with steric and electrostatic fields) performs substantially better and COX2 for which CoMFA performs better, both approaches for encoding steric and electrostatic fields give similar results. The differences for the BZR and COX2 sets disappear when outliers are excluded.

(2) Additional fields in CoMSIA yield substantially more predictive models for the COX2, GPB, and THR sets. For the other sets, q^2 values are similar to those for CoMFA and CoMSIA basic.

(3) HQSAR performs as well as or better than field-based methods for all sets except AchE and THR. HQSAR often exceeds the predictive accuracy of 2.5D by a large margin and is only substantially less predictive for the THER set. EVA and HQSAR give similar CV statistics for all sets (after outlier removal).

(4) 2.5D performs as well as or better than 2D except for THR. For the THR set, "noisy" variables adversely affect PLS modeling; all 2D variables are present in the 2.5D set.

A comparison of predictive accuracy on the designed test sets for PLS models leads to the following observations.

(1) Field-based QSAR methods are the most predictive. CoMFA produces models useful for extrapolation from five sets, and CoMSIA with additional fields produces useful models from seven sets.

(2) CoMFA performs either similarly (three sets) or better (four sets) than CoMSIA with steric and electrostatic fields only. It performs worse for BZR, but the small difference in s_{test} values must be kept in mind.

(3) The use of additional CoMSIA fields produces models that are similarly (four sets) or much more predictive (four sets) than CoMSIA with steric and electrostatic fields. When compared to CoMFA, CoMSIA with additional fields is more predictive for three sets and only substantially less predictive for the DHFR set. For the latter, the CoMSIA model is nonetheless very predictive. Thus, CoMSIA with additional fields appears to be preferable to CoMFA because of the simpler contours and the lower sensitivity of model statistics to the orientation and coarseness of the grid and to small changes in the alignment.[27]

(4) HQSAR compares favorably to field-based QSAR methods for five sets. It produces models useful for extrapolation from five sets. In addition, the COX2 and THR models with $r^2_{\text{test,no out}}$ values of 0.37 and 0.43 are not useless. It is noted that HQSAR occasionally produces very inflated activity predictions (e.g., 2-36 from the AchE set has a predicted pIC_{50} value of 17.2, despite having $T_c = 0.92$ compared to the most similar training set compound). Outlier analysis allows for these situations to be detected.

(5) 2.5D descriptors yield more predictive models than strictly 2D descriptors. Because structures are unambiguously determined by CORINA, 3D descriptors such as volume, charged surface area descriptors, etc. should be included in cases where traditional descriptors are used. However, models useful for extrapolation are obtained for the ACE and DHFR sets only. Notably, the superior performance on the THER set of 2.5D over HQSAR, observed for cross-validation, is not observed for the test set.

(6) EVA produces models that compare favorably with field-based methods for five sets. However, EVA is only substantially more predictive than HQSAR for the ACE set, while HQSAR is substantially more predictive than EVA for four sets. Considering these observations, the costly normal mode calculations, and difficult interpretation of EVA models, HQSAR is generally preferable for QSAR modeling in which an alternative to field-based methods is sought.

From a comparison of various model-building methods using 2.5D descriptors, it emerges that only neural network ensembles are worthy of consideration as alternatives to PLS. More predictive models are obtained for the COX2, THR, BZR, AchE, and THER sets, with the first two to three sets yielding models that are sufficiently predictive to be considered useful (i.e., approaching $r^2_{\text{test,no out}} = 0.5$) where the PLS models were not. For the remaining three sets, ensemble NN models perform similarly to PLS. Comparing single and ensemble NN models, the latter perform better, often by a large margin. This is consistent with the results of Agrafiotis et al.,[42] although the difference between the two methods is larger in our work. This may be related to differences in neural network architectures. Excluding the suspicious results for the BZR set, GFA with linear terms (GFA-l) significantly exceeds PLS performance only for the THER set and is somewhat more predictive for the THR set. Only the THR model is sufficiently predictive to be considered useful. Thus, the significantly longer run times for GFA (several minutes per run, thus several hours for CV), coupled with the sensitivity of models to the training compounds and the random number generator seed, may not seem worthwhile. However, the GFA-l models are never significantly worse than the PLS models. (While we have not relied on the "lack-of-fit" function[14] to automatically select the optimal number of descriptors during evolution, it is worth mentioning that there is an approximate correspondence between the optimal model size from evolution and the model size determined by examining each separately with cross-validation; the former gives mostly smaller models.) When nonlinear terms are included (GFA-nl), significantly more predictive models are obtained for the AchE and BZR sets. In both cases, the models are sufficiently predictive to be considered useful where the PLS or GFA-l models are not. However, the GFA-nl models are substantially worse than GFA-l or PLS for several sets. The better performance on the AchE and BZR sets could not have been anticipated from their low q^2 values. GPLS produces a more predictive model than PLS only for the THER set but still not sufficiently predictive to be useful. The use of GFA and GPLS ensembles generally does not provide more predictive models than the first-ranked model in the ensemble, although the average performance of the ensemble may help identify suspicious results (e.g., GFA-l on BZR).

For closely related methods, such as CoMSIA and CoMFA or various model-building approaches applied to 2.5D descriptors, the same compounds tend to be universal outliers. The attributes of outliers generally conform to expectation: many CoMSIA/CoMFA outliers have functional groups occupying regions of the alignment space not present in training compounds, and outliers for 2D methods tend to have divergent structures. In addition, a significant number of compounds are outliers for all methods. We suspect that the occasional spurious predictions for HQSAR result from the hashing procedure used to obtain integer strings of fixed length.

There have been a number of comparisons of cross-validation (CV) and test sets for determining the predictive accuracy of QSAR models.[45−47] Golbraikh et al.[45,46] advocate the use of test sets, claiming that CV is unreliable for estimating predictive accuracy. Hawkins et al. have criticized their work on the basis of the small test sets used for drawing comparisons.[47] In contrast, Hawkins et al. suggest that cross-validation is equally effective as large test sets and that leaving out the test set compounds is a waste of valuable information. A comparison of q^2 values calculated with the LOO or LSO procedures using various group sizes reveals a good correspondence among the various approaches (Figure 3). The LOO coefficient, often described as less robust or giving results that are too optimistic, is perhaps the best CV procedure for PLS modeling because it allows the application of the fast SAMPLS algorithm.[43] For the largest sets (ACE, BZR, COX2, and DHFR), only small decreases in q^2 are observed when using larger CV groups, with a fairly uniform decrease among methods. Except for CoMFA and 2.5D-PLS on the BZR set, excluding outliers does not alter trends among methods. In contrast, there is greater variation in q^2 for the smaller sets. While this may seem counterintuitive, it is more likely that a poor distribution of compounds will be achieved for small groups of compounds selected at random. For leave-$^1/_5$-out CV, the probability of selecting the 20 most active molecules from a set of 100 is 2×10^{-21}, compared to 2×10^{-4} for selecting the 4 highest activity molecules from a set of 20. The same argument would suggest that using small external test sets assembled by random selection is not recommended; this is corroborated by the work of Hawkins et al. Because CV with large groups is equivalent to repetitive model derivation with substantial test sets, it appears that the LOO procedure gives a good estimate of the ability of models to generalize.

However, comparing the values of $q^2_{\mathrm{L10\%O}}$ and $r^2_{\mathrm{pred,no\ out}}$ reveals a different trend. There is a reasonable correspondence between CV and test set results for the ACE, COX2, DHFR, and GPB sets, and methods with the highest CV correlation coefficients also have the highest test set coefficients for the AchE set. For the BZR, THER, and THR sets, there is no correspondence of coefficients. We note, however, that the THER and THR test sets are small and may not accurately represent the full extent of chemical space for the series. Some important distinctions between the work of Hawkins et al.[47] and the present study should be noted. In the former study, only one data set was examined with one descriptor set (traditional descriptors) and one model-building method (ridge regression); the test sets were assembled by random selection. We suggest that an external test set of 20 or more compounds designed to verify the extrapolative accuracy of models is useful in QSAR modeling (if sufficient samples are available). As is usually done when assessing predictive accuracy with CV, the final model should be derived using all training and test set compounds after having identified the most reliable QSAR approach.

In addition to predictive accuracy, there are other issues that are relevant to the selection of a QSAR method over others. One important consideration is the "interpretability" of the model, or the insights given as to what might represent promising modifications of existing compounds. In this respect, the most predictive methods (field-based 3D QSAR and HQSAR) give models that are more easily interpreted than EVA or 2D and 2.5D. The latter can be particularly difficult to interpret because they are whole-molecule descriptors. They provide no information on what aspects of molecules should be modified to enhance activity.

Conclusion

We have examined the predictive accuracy of a large number of QSAR methods applied to eight data sets. Field-based 3D QSAR methods were found to be the most predictive for extrapolating outside the training set. Unfortunately, these methods are not very tractable for virtual screening of large collections of congeneric compounds because of the manual labor involved in aligning structures. To a large extent, the recently described topomer-CoMFA method[48] addresses this limitation. We have found that HQSAR produces models with predictive accuracy similar to that of field-based methods in many cases. Therefore, it represents the most promising of the approaches we have examined for virtual screening applications.

While there is some variation over the data sets, the trends that we have discerned in this assessment of QSAR methods are sufficiently robust to aid QSAR practitioners in selecting a method over alternative approaches. Because of the widespread application of QSAR methods in concert with the synthesis and screening of analogue series, our observations will facilitate the analysis of QSAR predictions by the medicinal chemist. A better correspondence between expected and actual prediction accuracy will enhance the usefulness of QSAR models for all involved in the lead optimization process.

No doubt, some readers will interpret our results differently; the provision of complete statistics for all methods allows one to decide for themselves what technique might be most suitable for their particular needs. Also, the methods we have examined are only those available in commercial software packages. There are many approaches that have been described in the literature that appear very predictive while overcoming limitations of established methods (e.g., CoMMA,[49] MS-WHIM,[50] 4D-QSAR,[51] Quasar,[52] PharmPrint,[53] and GRIND[54] to name a few). It is fair to say, however, that many research groups have their favorite one or two data sets that they employ to validate methods they develop. As is evident from this work, a method that is predictive for a particular data set (especially for "easy" sets such as the ACE and DHFR sets) may be no better than existing approaches for a typical QSAR data set. Reasonably, most researchers want to develop methods, not assemble data sets. In distributing the data sets considered in this work as an expanded benchmark for QSAR methods, it will be possible to achieve more objective comparisons among methods.

Acknowledgment. We thank H. Gohlke for providing us with the GPB and THER sets, M. Böhm for the THR set, and G. Marshall for the ACE set. D.F.W. acknowledges support from the Natural Sciences and Engineering Research Council (NSERC), the Canadian Institutes of Health Research (CIHR), and the Canada Research Chairs Program. J.J.S. acknowledges support from a CIHR doctoral research award.

Supporting Information Available: Tabulations of compounds and their associated literature references, $q^2_{\mathrm{L20\%O}}$ and $q^2_{\mathrm{L33\%O}}$ values for PLS analyses, thresholds used for defining outliers and lists of outliers, descriptions of 3D QSAR grids and alignment procedures, 2.5D descriptors used for NN modeling, data sets in electronic format (both in MDL SD format and as Sybyl mol2 databases to preserve charges), and computer-readable ASCII tables of 2.5D descriptors. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Ekins, S. Predicting undesirable drug interactions with promiscuous proteins in silico. *Drug Discovery Today* **2004**, *9*, 276−285.

(2) Stone, M.; Jonathan, P. Statistical thinking and technique for QSAR and related studies. 1. General theory. *J. Chemom.* **1993**, *7*, 455−475.

(3) Oprea, T. I.; Waller, C. L. Theoretical and practical aspects of three-dimensional quantitative structure−activity relationships. In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Boyd, D. B., Eds.; Wiley-VCH: New York, 1997; pp 127−182.

(4) Hall, L. H.; Kier, L. B. The molecular connectivity chi indexes and kappa shape indexes in structure−property modeling. In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Boyd, D. B., Eds.; Wiley-VCH: New York, 1991; pp 367−422.

(5) Lowis, D. R. HQSAR: A new, highly predictive QSAR technique. http://www.tripos.com/sciTech/inSilicoDisc/media/LITCTR/HQSAR_AP. PDF.

(6) Heritage, T. W.; Lowis, D. R. Molecular hologram QSAR. *Rational Drug Design: Novel Methodology and Practical Applications*; Oxford University Press: New York, 1999.

(7) Free, S. M.; Wilson, J. W. A mathematical contribution to structure−activity studies. *J. Med. Chem.* **1964**, *7*, 395−399.

(8) Cramer, R. D.; Patterson, D. E.; Bunce, J. D. Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steriods to carrier proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959−5967.

(9) Klebe, G.; Abraham, U.; Mietzner, T. Molecular similarity indexes in a comparative analysis (CoMSIA) of drug molecules to correlate and predict their biological-activity. *J. Med. Chem.* **1994**, *37*, 4130−4146.

(10) Klebe, G.; Abraham, U. Comparative molecular similarity index analysis (CoMSIA) to study hydrogen-bonding properties and to score combinatorial libraries. *J. Comput.-Aided Mol. Des.* **1999**, *13*, 1−10.

(11) Ferguson, A. M.; Heritage, T.; Jonathon, P.; Pack, S. E.; Phillips, L.; et al. EVA: A new theoretically based molecular descriptor for use in QSAR/QSPR analysis. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 143−152.

(12) Wold, S.; Sjostrom, M.; Eriksson, L. PLS-regression: A basic tool of chemometrics. *Chemom. Intell. Lab. Syst.* **2001**, *58*, 109−130.

(13) Hansch, C.; Fujita, T. $\rho-\sigma-\pi$ analysis. A method for the correlation of biological activity and chemical structure. *J. Med. Chem.* **1964**, *86*, 1616−1626.

(14) Rogers, D.; 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.

(15) King, R. D.; Hirst, J. D.; Sternberg, M. J. New approaches to QSAR: Neural networks and machine learning. *Perspect. Drug Discovery Des.* **1993**, *1*, 279−290.

(16) Peterson, K. L. Artificial neural networks and their use in chemistry. In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Boyd, D. B., Eds.; Wiley-VCH: New York, 2000; pp 53−140.

(17) Selwood, D. L.; Livingston, D. J.; Comley, J. C.; O'Dowd, A. B.; Hudson, A. T.; et al. Structure−activity relationships of antifilarial antimycin analogs: A multivariate pattern recognition study. *J. Med. Chem.* **1990**, *33*, 136−142.

(18) Hansch, C.; Li, R.; Blaney, J. M.; Langridge, R. Comparison of the inhibition of *Escherichia coli* and *Lactobacillus casei* dihydrofolate reductase by 2,4-diamino-5-(substituted-benzyl)-pyrimidines: Quantitative structure−activity relationships, X-ray crystallography, and computer graphics in structure−activity analysis. *J. Med. Chem.* **1982**, *25*, 777−784.

(19) Hansch, C.; Hathaway, B. A.; Guo, Z.; Selassie, C. D.; Dietrich, S. W.; et al. Crystallography, quantitative structure−activity relationships (QSAR), and molecular graphics in a comparative analysis of the inhibition of dihydrofolate reductase from chicken liver and lactobacillus casei by 4,6-diamino-1,2dihydro-2,2-dimethyl-1-(substituted-phenyl)-*s*-triazines. *J. Med. Chem.* **1984**, *27*, 129−143.

(20) Depriest, S. A.; Mayer, D.; Naylor, C. B.; Marshall, G. R. 3D-QSAR of angiotensin-converting enzyme and thermolysin inhibitors. A comparison of CoMFA models based on deduced and experimentally determined active-site geometries. *J. Am. Chem. Soc.* **1993**, *115*, 5372−5384.

(21) Golbraikh, A.; Bernard, P.; Chretien, J. R. Validation of protein-based alignment in 3D quantitative structure−activity relationships with CoMFA models. *Eur. J. Med. Chem.* **2000**, *35*, 123−136.

(22) Maddalena, D. J.; Johnston, G. A. R. Prediction of receptor properties and binding-affinity of ligands to benzodiazepine/GABA(A) receptors using artificial neural networks. *J. Med. Chem.* **1995**, *38*, 715−724.

(23) Chavatte, P.; Yous, S.; Marot, C.; Baurin, N.; Lesieur, D. Three-dimensional quantitative structure−activity relationships of cyclo-oxygenase-2 (COX-2) inhibitors: A comparative molecular field analysis. *J. Med. Chem.* **2001**, *44*, 3223−3230.

(24) Sutherland, J. J.; Weaver, D. F. Three-dimensional quantitative structure−activity and structure−selectivity relationships of dihydrofolate reductase inhibitors. *J. Comput.-Aided Mol. Des.*, in press.

(25) Mattioni, B. E.; Jurs, P. C. Prediction of dihydrofolate reductase inhibition and selectivity using computational neural networks and linear discriminant analysis. *J. Mol. Graphics Modell.* **2003**, *21*, 391−419.

(26) Gohlke, H.; Klebe, G. Drugscore meets CoMFA: Adaptation of fields for molecular comparison (AFMoC) or how to tailor knowledge-based pair-potentials to a particular protein. *J. Med. Chem.* **2002**, *45*, 4153−4170.

(27) Bohm, M.; Sturzebecher, J.; Klebe, G. Three-dimensional quantitative structure−activity relationship analyses using comparative molecular field analysis and comparative molecular similarity indices analysis to elucidate selectivity differences of inhibitors binding to trypsin, thrombin, and factor Xa. *J. Med. Chem.* **1999**, *42*, 458−477.

(28) Willett, P.; Winterman, V. A comparison of some measures for the determination of intermolecular structural similarity measures of intermolecular structural similarity. *Quant. Struct.−Act. Relat.* **1986**, *5*, 18−25.

(29) Clark, R. D. Optisim: An extended dissimilarity selection method for finding diverse representative subsets. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 1181−1188.

(30) Lajiness, M.; Johnson, M. A.; Maggiora, G. M. Implementing drug screening programs using molecular similarity methods. *QSAR: Quantitative Structure−activity Relationships in Drug Design*; Alan R. Liss Inc.: New York, 1989; pp 173−176.

(31) Hassan, M.; Bielawski, J. P.; Hempel, J. C.; Waldman, M. Optimization and visualization of molecular diversity of combinatorial libraries. *Mol. Diversity* **1996**, *2*, 64−74.

(32) Besler, B. H.; Merz, K. M.; Kollman, P. A. Atomic charges derived from semiempirical methods. *J. Comput. Chem.* **1990**, *11*, 431−439.

(33) Gasteiger, J.; Marsili, M. Iterative partial equalization of orbital electronegativity—a rapid access to atomic charges. *Tetrahedron* **1980**, *36*, 3219−3288.

(34) Gasteiger, J.; Rudolph, C.; Sadowski, J. Automatic generation of 3D-atomic coordinates for organic molecules. *Tetrahedron Comput. Methodol.* **1990**, *3*, 537−547.

(35) Turner, D. B.; Willett, P.; Ferguson, A. M.; Heritage, T. Evaluation of a novel infrared range vibration-based descriptor (EVA) for QSAR studies. 1. General application. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 409−422.

(36) Seel, M.; Turner, D. B.; Willett, P. Effect of parameter variations on the effectiveness of HQSAR analyses. *Quant. Struct.−Act. Relat.* **1999**, *18*, 245−252.

(37) Hall, L. H.; Kier, L. B. Electrotopological state indexes for atom types—a novel combination of electronic, topological, and valence state information. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 1039−1045.

(38) Stanton, D. T.; Jurs, P. C. Development and use of charged partial surface-area structural descriptors in computer-assisted quantitative structure property relationship studies. *Anal. Chem.* **1990**, *62*, 2323−2329.

(39) Bonchev, D.; Mekenyan, O.; Trinajstic, N. Isomer discrimination by topological information approach. *J. Comput. Chem.* **1981**, *2*, 127−148.

(40) Dunn, W. J.; Rogers, D. Genetic partial least squares in QSAR. In *Genetic Algorithms in Molecular Modeling*; DeVillers, J., Ed.; Academic Press: London, 1996; pp 109−130.

(41) Krasowski, M. D.; Hong, X. A.; Hopfinger, A. J.; Harrison, N. L. 4D-QSAR analysis of a set of propofol analogues: Mapping binding sites for an anesthetic phenol on the GABA(A) receptor. *J. Med. Chem.* **2002**, *45*, 3210−3221.

(42) Agrafiotis, D. K.; Cedeno, W.; Lobanov, V. S. On the use of neural network ensembles in QSAR and QSPR. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 903−911.

(43) Bush, B. L.; Nachbar, R. B. Sample-distance partial least-squares—PLS optimized for many variables, with application to CoMFA. *J. Comput.-Aided Mol. Des.* **1993**, *7*, 587−619.

(44) For HQSAR on the THR set with hologram definitions including hydrogens, s_{PRESS} decreased continuously with increasing number of components. We selected six components because the value of s_{PRESS} was minimized at six components for holograms excluding hydrogens. The optimal number of components is generally the same for all sets of holograms.

(45) Golbraikh, A.; Tropsha, A. Beware of q2! *J. Mol. Graphics Modell.* **2002**, *20*, 269−276.

(46) Golbraikh, A.; Shen, M.; Xiao, Z. Y.; Xiao, Y. D.; Lee, K. H.; et al. Rational selection of training and test sets for the development of validated QSAR models. *J. Comput.-Aided Mol. Des.* **2003**, *17*, 241−253.

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

(48) Cramer, R. D. Topomer CoMFA: A design methodology for rapid lead optimization. *J. Med. Chem.* **2003**, *46*, 374−388.

(49) Silverman, B. D.; Platt, D. E. Comparative molecular moment analysis (CoMMA): 3D-QSAR without molecular superposition. *J. Med. Chem.* **1996**, *39*, 2129−2140.

(50) Bravi, G.; Gancia, E.; Mascagni, P.; Pegna, M.; Todeschini, R.; et al. MS WHIM, new 3D theoretical descriptors derived from molecular surface properties: A comparative 3D QSAR study in a series of steroids. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 79−92.

(51) Hopfinger, A. J.; Wang, S.; Tokarski, J. S.; Jin, B. Q.; Albuquerque, M.; et al. Construction of 3D-QSAR models using the 4D-QSAR analysis formalism. *J. Am. Chem. Soc.* **1997**, *119*, 10509−10524.

(52) Vedani, A.; Dobler, M.; Zbinden, P. Quasi-atomistic receptor surface models: A bridge between 3D QSAR and receptor modeling. *J. Am. Chem. Soc.* **1998**, *120*, 4471−4477.

(53) McGregor, M. J.; Muskal, S. M. Pharmacophore fingerprinting. 1. Application to QSAR and focused library design. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 569−574.

(54) Pastor, M.; Cruciani, G.; McLay, I.; Pickett, S.; Clementi, S. Grid-independent descriptors (GRIND): A novel class of alignment-independent three-dimensional molecular descriptors. *J. Med. Chem.* **2000**, *43*, 3233−3243.

JM0497141