Development of Quantitative Structure−Activity Relationships and Classification Models for Anticonvulsant Activity of Hydantoin Analogues

Jeffrey J. Sutherland† and Donald F. Weaver*,‡

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

Received November 18, 2002

Classification and QSAR analysis was performed on a large set of hydantoin derivatives with measured anticonvulsant activity in mice and rats. The classification set comprised 287 hydantoins having maximal electroshock (MES) activity expressed in qualitative form. A subset of 94 hydantoins with MES ED_{50} values was used for QSAR analysis. Numerical descriptors were generated to encode topological, geometric/structural, electronic, and thermodynamic properties of molecules. Analyses were performed with training and test sets of diverse compounds selected using their representation in a principal component space. Cell- and distance metric-based selection methods were employed in this process. For QSAR, a genetic algorithm (GA) was used for selecting subsets of 5−9 descriptors that minimize the rms error on the training sets. The most predictive models have rms errors of 0.86 ($r^2 = 0.64$) and 0.73 ($r^2 = 0.75$) $\ln(1/ED_{50})$ units on the cell- and distance metric-derived test sets, respectively, and showed convergence in the selected descriptors. Classification models were developed using recursive partitioning (RP) and spline-fitting with a GA (SFGA), a novel method we have implemented. The most predictive RP and SFGA models have classification rates of 75% and 80% on the test sets; both methods produced models with similar discriminating features. For QSAR and classification, consensus schemes gave improved predictive accuracy.

INTRODUCTION

Currently available drugs for the treatment of epilepsy are symptomatically effective in only 60−70% of patients.[1] Regrettably, the recently introduced drugs (e.g. tiagabine, lamotrigine, topiramate, vigabatrin, felbamate, gabapentin) do not appear to be significantly superior to older drugs (e.g. diphenylhydantoin, carbamazepine, valproate) in terms of seizure suppression efficacy.[1] Therefore, despite their introduction over 60 years ago, hydantoins remain an important class of antiepileptic drugs. Accordingly, it is of interest to develop models capable of explaining anticonvulsant activity across the spectrum of hydantoin derivatives. Phenytoin, the prototype hydantoin, binds to the neuronal voltage-gated sodium channel (NVSC) protein, eliciting its pharmacological effects by prolonging the channel's bioelectrically inactive state. The structural requirements for hydantoin binding to NVSC remain unknown. A CoMFA study on a limited set of hydantoins with measured binding affinity for NVSC provided qualitative information on the nature of the binding site.[2] However, most anticonvulsants have been and continue to be subjected to the maximal electroshock (MES) seizure test, an experimental model for tonic-clonic seizures. A simple Hansch-type equation has been used to explain MES activity of 11 hydantoins.[3] Another study examined the role of the $^1\chi^{(v)}$ index for explaining the activity of 82 hydantoins, simply identified as MES active or inactive.[4] Here, we outline the development and validation of models capable of explaining MES activity across a large, diverse set of hydantoins.

We have used a two-pronged approach to modeling hydantoin activity. Quantitative structure−activity relationship (QSAR) models have been developed to quantify the level of activity for hydantoins showing some MES activity. Since in vivo activities compiled from a variety of sources were used, the goal was to develop simple QSAR models that could be used to estimate the activity of compounds rapidly and reliably. In addition, classification models were developed to take advantage of the significantly larger set of compounds for which MES activity has been reported in various forms. Novel compounds predicted to be active by the classification models could then be subjected to the QSAR models for refining their predicted activity. Because of difficulties encountered in developing predictive recursive partitioning[5] (RP) models for classification, we have implemented a novel genetic algorithm-based approach to partitioning that is similar in philosophy to that of RP, but replaces its incremental approach to variable selection with a best-subset approach.

EXPERIMENTAL SECTION

Data Sets. All biological activities have been taken from the literature.

(i) QSAR Set. For QSAR modeling, we have assembled a set of 94 hydantoin derivatives with measured MES activities in mice or rats following oral administration. Activities are expressed as the concentration required to

* Corresponding author fax: (902)494-1310; e-mail: weaver@chem3.chem.dal.ca.
† Queen's University.
‡ Dalhousie University.

10.1021/ci025639w CCC: $25.00
Published on Web 03/20/2003

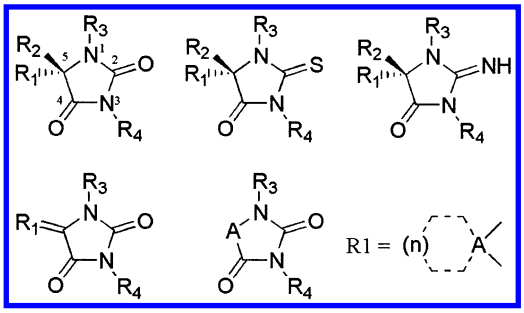

Figure 1. Scaffolds present in the hydantoin data set.

protect 50% of animals from seizures (ED_{50}), in millimoles of compound per kilogram of animal weight (mmol/kg). These were converted to $\ln(1/ED_{50})$ units, which fell into the range [−2.24, 3.43]. The use of data from mice and rats together appears justified from a comparison of 14 compounds where activities in both animals have been reported. Using multiple values reported for some compounds, the uncertainty in activity was estimated to be 0.72 $\ln(1/ED_{50})$ units.

(ii) Classification Set. A set of 287 hydantoin derivatives was assembled for classification analysis. Restrictions on allowed testing procedures were loosened to include hydantoins tested by MES following intraperitoneal administration and hydantoins with results not expressed as ED_{50} values. Classification of compounds as active (1) or inactive (0) was achieved using the following scheme: (1) for activities expressed as ED_{50} values, those with $ED_{50} < 4.23$ mmol/kg (or $ED_{50} < 1000$ mg/kg) were classified as active, (2) for activities expressed as the percent increase in threshold required to induce seizures, any compound with nonzero percent increase was considered active, and (3) literature classification as active/inactive was used directly. This gave 172 active compounds and 115 inactive compounds. The threshold $ED_{50} < 4.23$ mmol/kg was selected bearing in mind the role of classification models as QSAR preprocessors in virtual screening, lying at the low end of activities for hydantoins used in QSAR development. Among the compounds used for classification, there are 147 with measured ED_{50} values and only 17 with $ED_{50} > 2.19$ mmol/kg (or $ED_{50} > 500$ mg/kg). Because of the small number of compounds with quantitative MES data around the selected threshold, the classification models are not sensitive to the selected value. The selection of 1000 mg/kg as a threshold for activity is consistent with three sources that gave both determined and undetermined ED_{50} values (i.e. the approximate value where ED_{50} values are replaced with "inactive" or "$ED_{50} > x$ mg/kg"). The complete set of compounds with activities is given in the Supporting Information. It includes derivatives with aliphatic, aromatic, heterocyclic, ether, ketone, hydroxy, ester, carboxylic acid, and halide substituents on the scaffolds in Figure 1.

Computational Methodology. This work was performed using the molecular modeling package Cerius2 v. 4.6 (http://www.accelrys.com), controlled with Tcl scripts for repetitive tasks. Modeling of hydantoin activities proceeded in the following stages: (1) structure entry and optimization, (2) descriptor generation, (3) training and test set assembly, (4) QSAR model development, and (5) classification model development.

(i) Structure Entry and Optimization. Molecules were sketched in the S configuration, consistent with the stereochemistry observed for active iminohydantoins.[6] Structures were optimized using the CFF97 force field[7] with no cutoff for van der Waals or electrostatic interactions. The "Boltzmann jump" procedure (i.e. a Monte Carlo conformational search) was used to select optimal torsion angles (using 400 perturbations per jump sequence, followed by energy-minimization). This was repeated 100 times, and the lowest energy conformer was retained as the final structure. For the purpose of calculating descriptors, atomic charges were assigned using the charge equilibration method.[8]

(ii) Descriptor Generation. Numerical descriptors that encode topological, geometric/structural, electronic, and thermodynamic properties were calculated. For 3D descriptors, the low energy conformer identified for each molecule was selected. Initially, 127 descriptors were generated for the QSAR set, and 117 for the classification set (no quantum descriptors or hydantoin ring partial charge descriptors were calculated for the classification set).

Prior to generating models, descriptors were removed by examining each data set separately. The first reduction eliminated descriptors having distributions of values clumped about a few distinct values, since these are not useful for explaining a continuous variation of activity. The second reduction eliminated one descriptor from each pair having a pairwise correlation coefficient r satisfying $|r| > 0.95$, since they contain nearly the same information. The descriptor to be retained was selected using two criteria: (1) it was more easily interpretable (e.g. retaining molecular volume over information-content descriptors[9]), or (2) it was less correlated with the remaining descriptors. Descriptors having $|r| > 0.90$ for three or more descriptor pairs were also eliminated. This reduced the number of descriptors to 57 for the QSAR set and 50 for the classification set. Table 1 lists descriptors used in the models presented here.

(iii) Training and Test Set Assembly. Series design ensures the data set contains molecules that span descriptor space effectively while removing redundancies that can artificially affect model statistics.[23] For the QSAR data set, three methods were used for assigning compounds to training and test sets. In the simplest approach, molecules were ordered by activity, and every third was selected for the test set, with the remaining molecules making up the training set. These are identified as "by-activity" (BA) sets (63 training, 31 test compounds).

The other approaches applied principal component analysis[24] (PCA) for reducing the (autoscaled) descriptor pool to seven principal components (PCs) accounting for 81% of descriptor variance. Training and test sets were assembled by representing molecules in the seven dimensional space spanned by the orthogonal PCs (scaled to unit variance). A cell-based method was used to select 60 molecules within the PC space. The cell divisions were obtained by binning the scores for each PC into equal intervals. Bins were added to each PC, ordered by decreasing range of PC scores, until the number of occupied cells reached 60. The molecule closest to the cell center was selected from each occupied cell. A test set of 20 compounds was formed by removing every third from the list of 60 cell-selected compounds ordered by activity. The remaining 40 form the training set. These are identified as "cell" (C) sets.

Table 1. Descriptors Used in Representative QSAR and Classification Models

topological descriptors	
$^3\chi_p^v$	valence-modified connectivity index of order 3 and "path" subgraph type[10,11]
$^3\kappa_\alpha$	α-modified shape index of order 3[11−13]
J_X	Balaban atom radii-modified index of average-distance sum connectivity[14,15]
geometric/structural descriptors	
R_og	radius of gyration[16]
V_m	molecular volume defined using *vdW* surface of atoms
σ	molecular density, MW/V_m
N_{rot}	no. of rotatable bonds (excluding AT_3 and AT_2 terminal bonds, such as CH_3 and NH_2)
electronic descriptors	
LUMO	AM1 LUMO[17]
Q_1	partial charge on hydantoin ring atom 1 (see Figure 1)
hybrid descriptors	
S_{ssCH2}, S_{aaCH}, etc.	electrotopological state indices[18]
PPSA3, PNSA3, etc.	charged partial surface area (CPSA) descriptors[19]
thermodynamic descriptors	
$ALogP_{98}$	LogP calculated from atoms using 1998 parameters[20,21]
F_{oct}	desolvation energy in 1-octanol calculated with fragment-based method[22]

Table 2. Number of Compounds in the Intersection of QSAR Training and Test Sets[a]

set	C − DM	C − BA	DM − BA
training	27 (40)	28 (40)	28 (40)
test	10 (20)	8 (20)	8 (20)

[a] Values in parentheses indicate the maximum possible size of the intersection; C = cell set, DM = distance metric set, BA = by-activity set.

Another method for selecting compounds in PC space consists of maximizing a Euclidian distance-based metric defined as

$$f = \left[\prod_i \prod_{j \neq i} d_{ij}^2\right]^{\frac{2}{N(N-1)}} \quad (1)$$

where d_{ij} is the distance between molecules i and j in PC space, and N is the number of selected molecules. A set of 60 compounds was selected by optimizing f using a Monte Carlo procedure, described in detail elsewhere.[25] Briefly, an initial subset of compounds is chosen, and a modified Metropolis scheme (where f replaces the energy) is used to accept or reject random changes to the subset of selected molecules. The temperature used in the Metropolis scheme was lowered from 5000 to 10 K in 10% increments (100 000 trial sets at each temperature). A test set of 20 compounds was formed by removing every third from the list of 60 distance metric-selected compounds ordered by activity. The remaining 40 form the training set. These are identified as the "distance metric" (DM) sets. The overlap of sets produced by the three approaches is given in Table 2.

Starting with the complete set of 287 hydantoins, the cell and distance metric methods were each applied to select subsets of diverse compounds that would serve as classification test sets; the C and DM test sets consists of 99 compounds (53 actives, 46 inactives) and 100 compounds (47 actives, 53 inactives), respectively. For RP, the remaining 188 (C) and 189 (DM) compounds were used for the training set. For SFGA, this was further reduced to 54 actives and 54 inactives by applying the cell and distance metric procedures again (see below for descriptions of RP and SFGA). The intersections of the C and DM sets contain 45 compounds (of a maximum 108) for the SFGA training sets and 47 compounds (of a maximum 99) for the test sets.

(iv) QSAR Model Development. The GFA (genetic function approximation) algorithm[26] was used to select 4−9 descriptors, with the length of models fixed during evolution (15 000 crossovers, 750 individuals, 50% mutation probability); models were fit to the C, DM, and BA training sets using multiple linear regression (MLR). Only the model with the largest r^2_{train} value was retained from each population. To investigate the presence of nonlinear effects, models allowing descriptors in quadratic form were also developed. For each set of parameters (number of descriptors, linear or linear and quadratic terms, training set), three GFA runs starting with different random seeds were performed, and the model with the largest r^2_{train} selected.

(v) Classification Model Development. Recursive partitioning (RP) and spline-fitting with a genetic algorithm (SFGA) were used for developing classification structure−activity relationships (CSAR). The RP method categorizes objects by deriving a binary decision tree in which descriptors are used to split the data set into smaller, homogeneous subsets.[5] It has become a popular approach for performing CSAR analyses.[27−30] RP can account for nonlinearities such as threshold effects, interactions between descriptors (cross-terms) or activity dependence on higher order terms.[29] It provides an interpretable model for explaining activities, in contrast to classification methods such as the k-nearest neighbors method (kNN).[31] We have used the CART algorithm[5] to classify the C and DM training sets (gini splitting rule, 40 evenly spaced knots per descriptor). Trees were split until terminal nodes contained 5, 10, 15, or 20 compounds (each was examined independently). Scaled pruning factors of 2−5 were examined systematically (α in ref 5 multiplied by the number of compounds in the training set). The active and inactive classes were given equal weight in determining misclassification costs.

In multivariate analyses such as QSAR, it is well-known that incremental approaches for variable selection often find suboptimal solutions when applied to large sets of variables (e.g. stepwise MLR). One recently described approach attempts to overcome this limitation in RP by using simulated annealing for descriptor selection.[32] Another that combines kNN and RP has been found to outperform both methods.[33] Genetic algorithms (GAs) have proven robust at finding predictive subsets of 1D and 2D descriptors for classification.[34] We have applied spline-fitting with a genetic algorithm (SFGA) for developing classification models, where the GFA algorithm[26] is used to select combinations of descriptor splines. As for RP, splines partition the data set into compounds having similar features and can account for nonlinear behavior. The spline function $<f(x) - a>$ equals zero if the value of $f(x) - a$ is negative; otherwise, it equals $f(x) - a$. Activities returned by a SFGA model are continuous

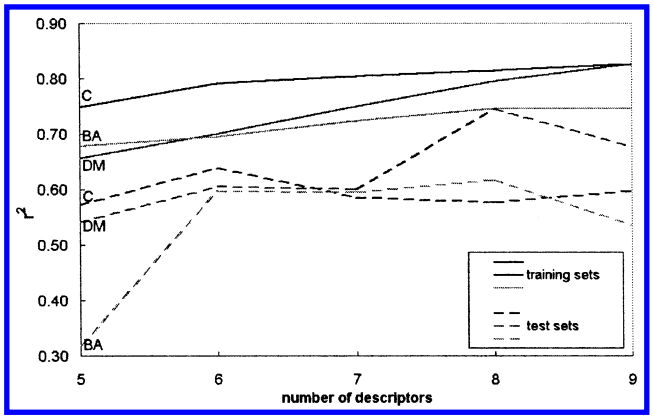

Figure 2. Linear model correlation coefficient (r^2) for C, BA, and DM training sets (solid lines) and test sets (dashed lines).

in nature. They are converted to discrete values by designating as active (1) those with predicted activities ≥ 0.5 and inactive (0) those with predicted activities < 0.5. The number of crossovers, population size, descriptor mutation probability, and knot shift probability were set to 15 000 operations, 750 individuals, 50% and 50%, respectively. Fixed length models having 4−8 descriptor splines were evolved separately. Only the most predictive model (for the training set) was retained from each population.

RESULTS AND DISCUSSION

(i) QSAR Models. The models discussed in detail include 3D descriptors. Model labels used below are interpreted as "c", "d", or "b" for models derived from C, DM, or BA sets, respectively, followed by the number of descriptors and "q" if quadratic descriptors were allowed:

$$\text{c6: } -0.648 + 0.807\,N_{rot} - 2.62\,{}^3k_\alpha - 6.66\,Q_1 + 0.229\,S_{aaCH} + 0.152\,F_{oct} - 1.39\,\text{LUMO}$$

$$\text{d6: } -5.06 + 0.689\,N_{rot} - 2.38\,{}^3k_\alpha - 6.98\,Q_1 - 8.25\,\text{RPSA} + 0.0407\,\text{WNSA1} + 24.9\,\text{RPCG}$$

$$\text{d8: } 15.6 + 0.952\,N_{rot} - 3.65\,{}^3k_\alpha - 9.86\,Q_1 - 0.250\,\text{PNSA3} + 129\,\text{FNSA3} + 0.662\,S_{ssCH2} - 1.79\,S_{dssC} - 0.333\,\text{PPSA3}$$

$$\text{b6: } 2.50 + 0.337\,N_{rot} + 1.48\,{}^3\chi_p{}^v - 0.967\,S_{dssC} + 19.8\,\text{FNSA1} - 19.8\,\text{RPSA} - 0.0416\,V_m$$

$$\text{c6q: } -0.00319 + 0.774\,N_{rot} - 2.54\,{}^3k_\alpha + 6.58\,Q_1{}^2 + 0.222\,S_{aaCH} - 0.00521\,F_{oct}{}^2 - 1.42\,\text{LUMO}$$

The predictive accuracy of models (r^2_{test}) was used to select the optimal number of descriptors (Figure 2). For the C and BA sets, model c6 and b6 are of optimal complexity. For the DM set, model d8 appears to be sufficiently more predictive than d6 to justify its greater complexity (using r^2_{test} or s_{PRESS}). Models c6, d6, d8, and b6 are selected as representatives of their respective series (henceforth called the representative models). Because d6 is smaller, reasonably predictive and performs differently than d8 (see below), it is included in the set of representative models. When r^2 is compiled over all nontraining set compounds (i.e. including the 34 compounds left out by series design in addition to the test set), the values of 0.52, 0.53, and 0.49 are obtained for c6, d6, and d8, respectively. The values calculated using the test set of 20 compounds are preferred, since they eliminate some of the redundancy present in the data set. When quadratic descriptors are allowed, DM and BA set models perform better than linear models for five descriptors only. Quadratic C set models are slightly more predictive than their linear counterparts; c6q is of optimal complexity.

Scatter plots of predicted and experimental activities show that deviations from experimental values are mostly bound by the estimated uncertainty (Figure 3). A thorough analysis shows compound **91** is the only universal outlier; it is poorly fit for d6, d8 and b6 and is poorly predicted by c6. It is closely related to **92** ($(C_2H_5)_2NCH_2CH_2-$ instead of $(CH_3)_2NCH_2CH_2-$ at R_3), which is well fit by all representative models. Given their structural similarity, their large activity difference of 1.84 $\ln(1/ED_{50})$ units reported by independent groups could indicate that the experimental value for **91** is in error. The predictive accuracy of models does not depend strongly on the substituent category of a compound (i.e. hydantoins with aliphatic substituents, aromatic substituents, etc.).

Cross-validation (CV) is often used to assess model predictive accuracy and the complexity justified by the data, especially when the size of the data set precludes the designation of separate test sets. Here, we have performed "leave-several-out" cross-validation (LSO) by removing 1/3rd of 94 compounds in turn (3×10 times) and using them as a prediction set for models fit with the remaining 2/3rds of the data set (repeating descriptor selection). A less robust method is "leave-one-out" cross-validation (LOO) where only one compound is removed at a time. As expected for large data sets, the q^2_{LOO} values are consistently higher than their LSO equivalents (Figure 4). Quadratic models perform better for small models only. The values of q^2_{LOO} and q^2_{LSO} increase monotonically for linear models, failing to identify the optimal number of descriptors. (The same holds if the corresponding s_{PRESS} values are used.)

To assess if models are the result of chance correlations, the activities of training and test set compounds were scrambled and refit 10 times (repeating descriptor selection). The following was obtained for the representative models: model ($r^2_{train,average}$, $r^2_{train,max}$); c6 (0.47, 0.60), d6 (0.44, 0.60), d8 (0.53, 0.61), b6 (0.30, 0.38). r^2_{test} was negative for all trials. The larger values for r^2_{train} on the C and DM sets compared to the BA set result from the smaller number of training set compounds in those sets.

The degree of independence of descriptors within models was examined by removing each in turn and regressing it against the others. For c6 and d6, ${}^3\kappa_\alpha$ and N_{rot} have high r^2 values (0.91−0.92) that arise from their high pairwise correlation. For b6, V_m has $r^2 = 0.95$ due mostly to its high pairwise correlation with ${}^3\chi_p{}^v$ and N_{rot}. For d8, all but Q_1 have substantial multicorrelation with the others ($r^2 = 0.82-0.97$). Despite this, removal of one of these descriptors reduces the accuracy of models significantly. For that which causes the smallest r^2_{train} reduction, r^2_{train}, and r^2_{test} are reduced to 0.62 and 0.55 for c6, 0.58 and 0.54 for b6, 0.73 and 0.42 for d8, and 0.60 and 0.41 for b6. Interestingly, when FNSA3 is removed from d8 (97% of its variance is explained by the

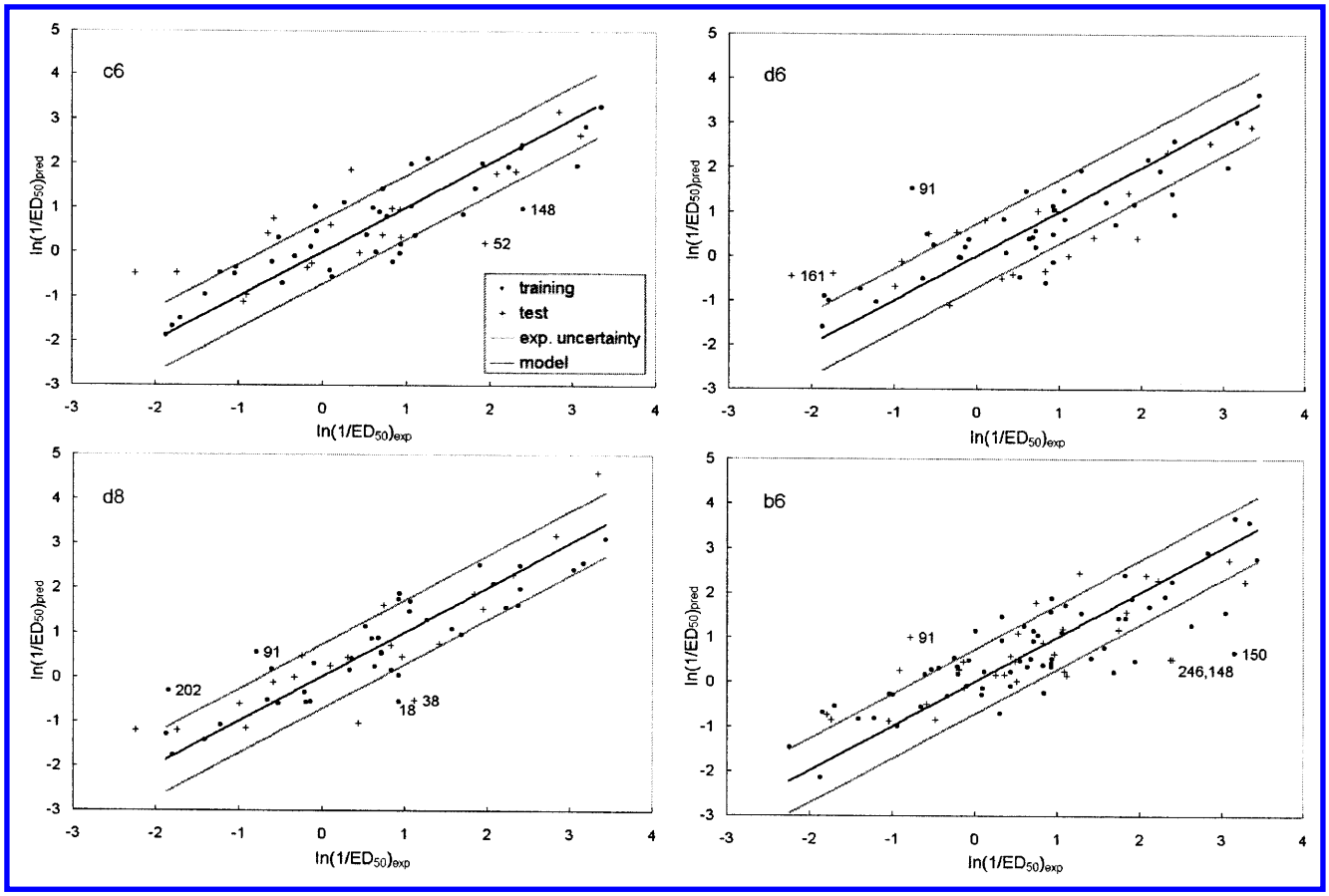

Figure 3. Scatter plots of predicted and measured activities. The black line indicates the perfect one-to-one correlation, the gray lines indicates the estimated error range of 0.72 ln(1/ED_{50}) units. Outliers are labeled.

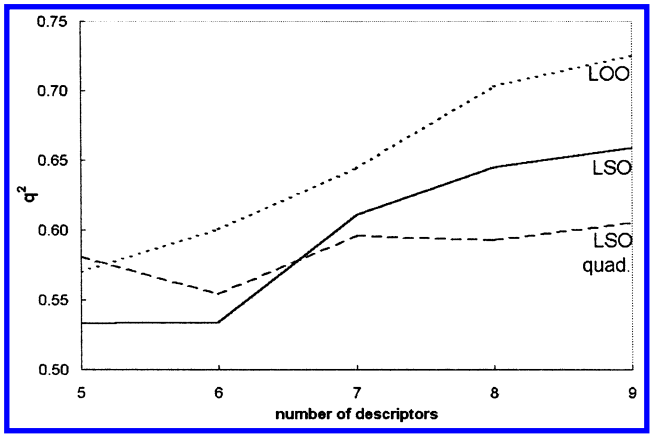

Figure 4. Correlation coefficient (q^2) for leave-one-out (LOO) and leave-several-out (LSO) cross-validation. "quad." indicates that quadratic descriptors were allowed.

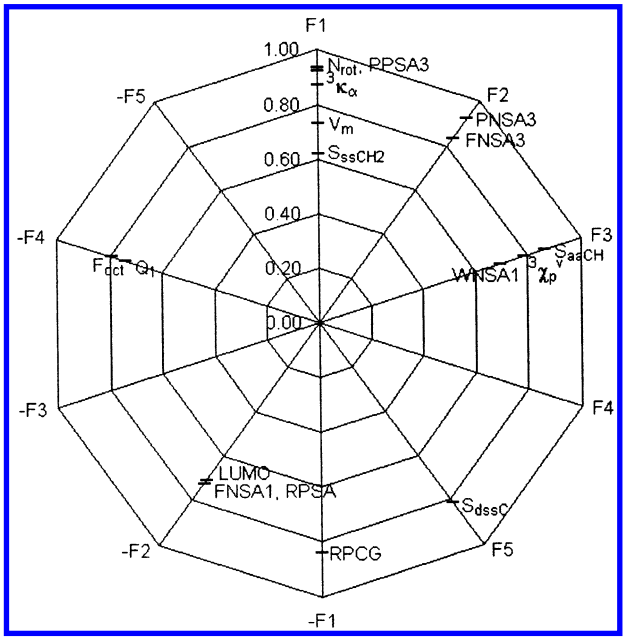

Figure 5. Projection of descriptors against the factor that accounts for the largest fraction of their variance. The descriptor variance explained by each factor is obtained by squaring the loading (i.e. the coordinate on the factor axis).

other descriptors), it has a disastrous effect on the model ($r^2_{train} = 0.37$, $r^2_{test} = -0.48$). Backward-stepwise regression did not find subsets of descriptors with a smaller decrease in model accuracy. It appears that despite redundant information in some descriptors, they contain information necessary for maintaining the predictive accuracy of models.

In addition to examining the statistical validity of models, it is important to assess if multiple models have a similar physical basis. In other words, the descriptors present in different models should show convergence. However, one should not expect to obtain nearly identical equations when using a pool of correlated descriptors. Factor analysis[24] (FA) was used to determine the nature of common underlying factors (latent variables) among the models. FA was carried out using the complete QSAR set and the descriptors of representative models. The use of five factors accounts for 86% of the variance of the original descriptors. Figure 5 groups descriptors according to the (VARIMAX-rotated) factor that accounts for most of their variance. When

examined in terms of the underlying factors, the four models each contain four independent factors, some represented by more than one descriptor. Those descriptors belonging to the same factor do not always have a concerted effect (e.g. N_{rot} and ${}^3\kappa_\alpha$). By replacing each descriptor with the appropriate factor in the normalized-coefficient QSAR equations and combining factors appearing more than once, the following emerges (normalizing QSAR equation coefficients removes their dependence on the numerical magnitude of descriptors):

$$\text{c6: } -\text{F1} + \text{F2} + \text{F3} - \text{F4} = \text{activity}$$

$$\text{d6: } -\text{F1} + \text{F2} + \text{F3} + \text{F4} = \text{activity}$$

$$\text{d8: } -\text{F1} + \text{F2} + \text{F4} - \text{F5} = \text{activity}$$

$$\text{b6: } -\text{F1} + \text{F2} + \text{F3} - \text{F5} = \text{activity}$$

The notation "- Fx + Fy = activity" indicates that as Fx decreases and Fy increases, activity increases. F1 can be associated with molecular shape. N_{rot}, V_m, and ${}^3\kappa_\alpha$ increase with elongation of molecular shape, all corresponding to increasing F1. F1 reveals unexpected relationships; descriptors such as PPSA3 and RPCG would usually be thought of as electronic descriptors. The effect of molecular shape on activity is consistent with that observed in other studies.[3,4] F2 represents the influence of charge on activity. LUMO becomes more negative as electrophilicity increases (or as excess negative charge decreases), FNSA1 and RPSA decrease as negatively charged surface area decreases, and PNSA3 and FNSA3 become less negative as the negatively charged surface area decreases, all corresponding to increasing F2. F3 accounts for 70% of the variance of the sum of atomic polarizabilities P_a.[22,35] The correlation coefficient between P_a and S_{aaCH}, WNSA1, ${}^3\chi_p{}^v$ are 0.77, 0.81, and 0.89, respectively; this association is intuitively reasonable given the definition of these descriptors. Thus, an increase in F3 corresponds to an increase in polarizability. Taken together, the statements about F1, F2, and F3 can be expressed as "to increase activity, decrease elongation of shape, decrease surface charge and increase polarizability". Less can be said about F4: F_{oct} only appears in C set models, and only half of Q_1 variance is explained by F4. F5 is represented by S_{dssC} alone. S_{dssC} is highly correlated with S_{dO} ($r = -0.86$); it accounts for the electrotopological state of (hydantoin ring) carbonyl groups. Disagreement (or agreement) should be considered within the approximation of using factors that represent only part of descriptor variance. The scheme is useful for examining similarities between models and rationalizing the role of more complex descriptors through their relation to simpler ones.

In addition to comparing the descriptors of QSAR models, it is useful to examine their variability in predicting activities. An intuitive comparison emerges from a loadings plot derived by applying PCA to model residuals (Figure 6). From this, it emerges that model d6 is a sort of "intermediate" model among the representative set.

When multiple models are available, averaging the predictions of individual models has proven useful for increasing predictive accuracy.[36] Here, a consensus prediction for the activity of a compound is calculated by averaging its predicted activities from four individual models (c6, d6, d8, b6) (Table 4). Using the arithmetic mean for calculating consensus predictions is arbitrary; other choices are possible. Unfortunately, there is no correlation between the variance of the four predictions and the actual error of prediction; this would be useful for assessing the reliability of true predictions.

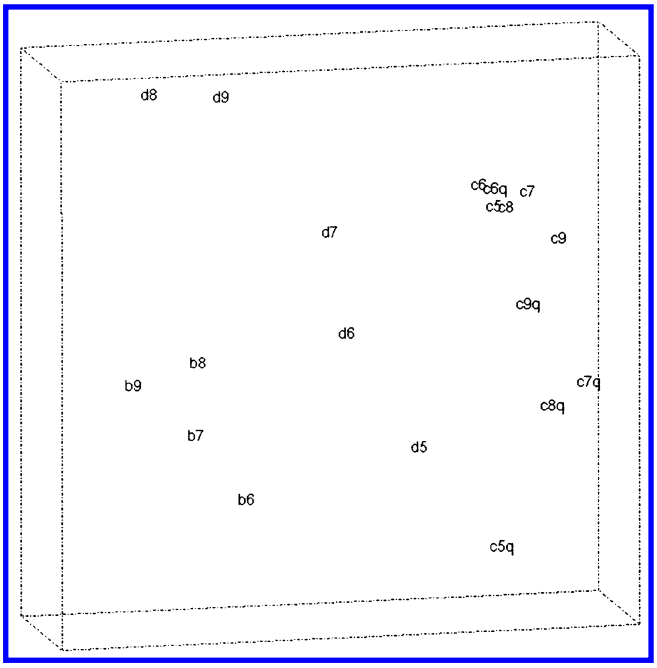

Figure 6. Loadings associated with the first three principal components (PCs), obtained from PCA on model residuals; three PCs account for 86% of residual variance. Model proximity in the loading space corresponds to similarity of residual profiles.

Table 3. Statistics of Representative QSAR Models[a]

model	r^2_{train}	r^2_{test}	rms error train	rms error test
c6	0.79	0.64	0.63	0.86
d6	0.70	0.61	0.74	0.91
d8	0.80	0.75	0.61	0.73
b6	0.70	0.60	0.75	0.82
c6q	0.79	0.68	0.63	0.81

[a] c6, d6, d8, c6q: 40 training compounds, 20 test compounds; b6: 63 training compounds, 31 test compounds.

Table 4. Consensus QSAR Model r^2 Statistics

set	no. of compds	c6	d6	d8	b6	cons.
all	94	0.64	0.60	0.62	0.66	0.72
all nontraining[a]	15	0.72[b]	0.63[b]	0.65[b]	0.77[b]	0.80

[a] The "all nontraining" set contains compounds not selected by series design in addition to test set compounds. They are included to increase the number of compounds in the intersection. [b] Calculated over the subset of compounds in the intersection.

Models containing only 2D descriptors were also developed. The predictive accuracy of (optimal length) models is lower: model (r^2_{train}, r^2_{test}); c6−2D (0.77, 0.55), d6−2D (0.68, 0.37), b6−2D (0.62, 0.51). Projection of descriptors into the space of latent variables (from factor analysis) shows that the 2D models lack a description of the surface charge effect (F2). The lower predictive accuracy of 2D models is not limited to certain substituent categories. Models c6-2D and b6-2D are of reasonable quality and could be used instead of the 3D models if structural optimization of a large set of analogues is too costly (they are provided in the Supporting Information).

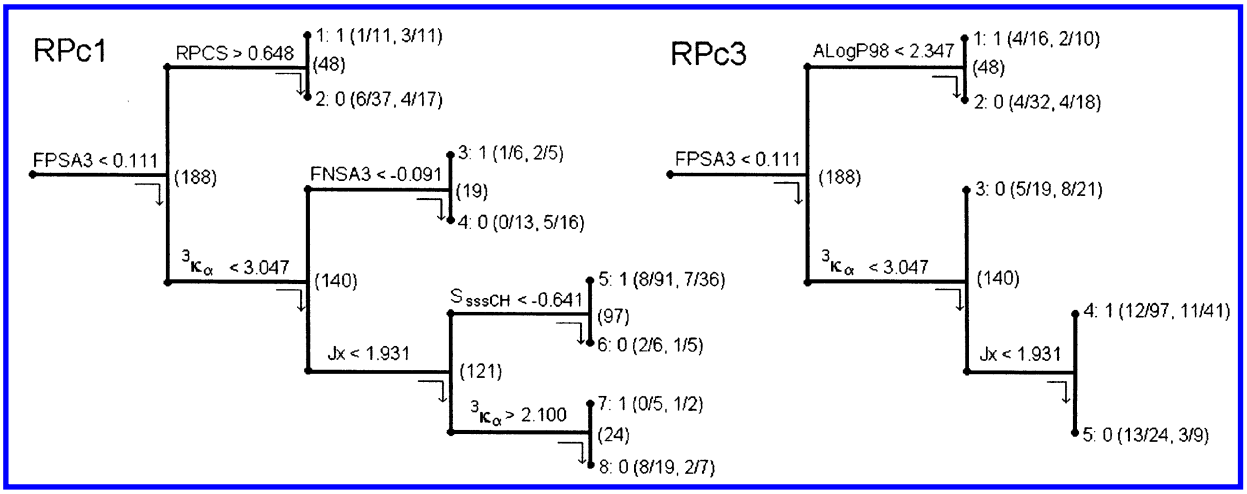

Figure 7. RPc1 and RPc3 classification models. The tree is traversed by taking the lower branch if the inequality is satisfied and taking the upper branch if it is not satisfied. Terminal nodes labels [node number: assigned activity (misclassified/training compounds at node, misclassified/test compounds at node)].

Table 5. Conformation Dependence of 3D Descriptors

	av conformer variance (fraction of QSAR set variance)		av conformer variance (fraction of QSAR set variance)
V_m	0.02 ± 0.05	FNSA3	0.16 ± 0.07
PPSA3	0.22 ± 0.15	RPSA	0.12 ± 0.05
PNSA3	0.20 ± 0.10	WNSA1	0.18 ± 0.11
FNSA1	0.16 ± 0.06	LUMO	0.26 ± 0.22

In contrast to 2D descriptors, the use of 3D descriptors introduces a dependence of QSAR predictions on the selected conformers. For each hydantoin in the QSAR set, 500 conformers were generated by randomly adjusting dihedral angles and retaining those within 5 kcal/mol of the global energy minimum. The average variation in descriptor values is mostly limited to less than 20% of the set variance (Table 5). This represents an upper limit for uncertainty in descriptor values. If compounds are energy-minimized after perturbing torsion angles, the number of distinct conformers is substantially reduced.

(ii) Classification Model Development. Model labels used below, such as RPc1 means **RP** model from **C** set, number **1**; Sc5 means **S**FGA model from **C** set with **5** splines. Models discussed in detail appear below and in Figure 7:

$$\text{Sc5: } 0.82 + 0.81 \langle J_X - 1.87\rangle - 0.24 \langle 3.01 - \text{AlogP}_{98}\rangle - 0.34 \langle R_o g - 3.00\rangle - 34.28 \langle \text{FPSA3} - 0.107\rangle - 3.93 \langle \rho - 1.26\rangle$$

$$\text{Sd6: } 0.86 - 0.89 \langle 1.90 - J_X\rangle - 0.33 \langle 2.55 - \text{AlogP}_{98}\rangle - 1.96 \langle 2.92 - R_o g\rangle - 0.06 \langle -13.88 - F_{oct}\rangle + 0.87 \langle 1.93 - {}^3k_a\rangle - 0.19 \langle S_{aasC} - 0.81\rangle$$

For RP models, RPc1 is the most predictive (Figure 8). It was derived with a minimum of five samples per node using a pruning factor of 2.0. RPc3 was derived with a minimum of 15 samples per node using a pruning factor of 2.0 (3.0 and 4.0 gave the same tree). Despite being less predictive, it is appealing in its simplicity. RP models produced from the DM training set are unbalanced and deep; most models had test set classification rates for actives below 50%, with only 2 at roughly 60%. Even stringent parameters (15 samples per node; pruning factor of 4.0) gave a tree with six levels of nodes. Using a greedy splitting rule produced models of poorer quality. DM set RP models are not presented here; discussion of RP models below refers to C set models.

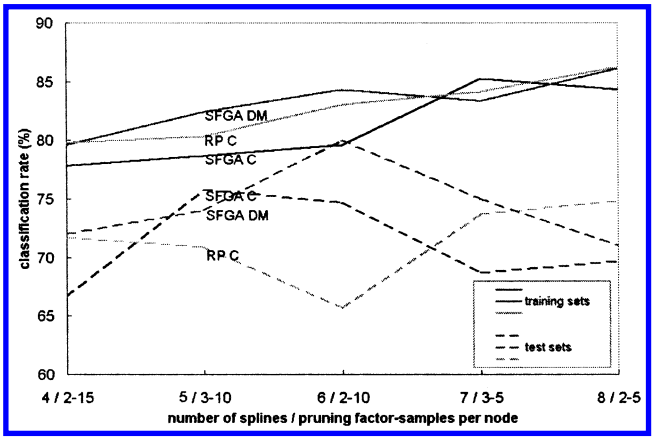

Figure 8. Training set (solid lines) and test set (dashed lines) classification rates for C set RP models and C and DM set SFGA models. Only the five most predictive RP models are included, ordered by increasing number of terminal nodes. The "number of splines" labels correspond to SFGA models, "pruning factor-samples per node" labels correspond to RP models.

For SFGA, models Sc5 and Sd6 are the most predictive of their respective series (Figure 8). There is a convergence of selected descriptors among classification models. In particular, model Sc5 contains all the descriptors appearing in RPc3 except for ${}^3\kappa_\alpha$, and 57% of its variance is explained by ρ and $R_o g$. Sd6 shares three descriptors with each of Sc5, RPc1, and RPc3 (note that the J_X spline in Sd6 partitions the data in the same way as that in Sc5, with large J_X values corresponding to activity; the $R_o g$ splines are opposite). The descriptors that also appear in the representative QSAR models (F_{oct}, ${}^3\kappa_\alpha$) or others that have a similar meaning to those from the QSAR models (e.g. FPSA3—indicating increasing charge on surface, and Sc5 $R_o g$) also affect activity in the same way. The role of individual descriptors among models may vary, depending on the presence or absence of correlated descriptors; disagreements (or agreements) in their roles should be viewed in that context.

For the most predictive model of each type, classification rates as a function of substituent category was examined

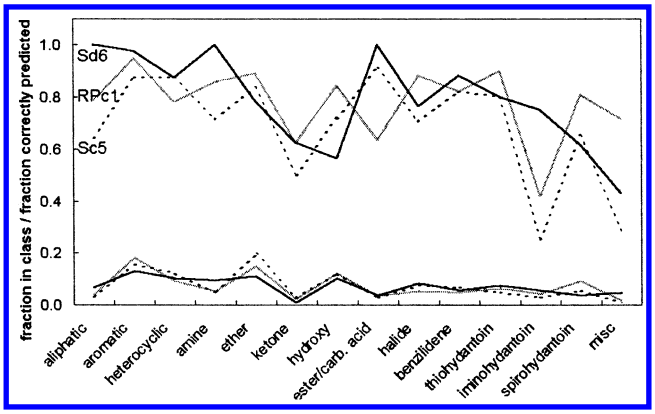

Figure 9. Classification rates of RPc1, Sc5, and Sd6 models across substituent categories, compiled over the complete set of hydantoins (set of upper curves). Fraction of training set compounds in each category used to derive the models (set of lower curves).

Table 6. Classification Model Statistics[a]

	classification rate (%)					
model	train (1)	train (0)	test (1)	test (0)	train (1+0)	test (1+0)
RPc1	87	86	77	72	86	75
RPc3	82	77	72	72	80	72
Sc5	82	76	77	74	79	76
Sd6	83	85	79	81	84	80

[a] RPc1, RPc3: 188 training compounds, 99 test compounds; Sc5: 108 training compounds, 99 test compounds; Sd6: 108 training compounds, 100 test compounds.

Table 7. Comparison of Classification Rates for Consensus and Best Single Models

		best single model		consensus model		no. of predictions/ intersection of sets	
scheme	models used	test (1)	test (0)	test (1)	test (0)	test (1)	test (0)
1	RPc1, RPc3	77	72	80	79	44/53	34/46
2	Sc5, Sc6	77	74	83	79	47/53	33/46
3	Sd5, Sd6, Sd7	79	81	84	81	38/47	42/53
4[a]	RPc1, Sc5, Sd6	84[b]	80[b]	89	94	27/38	18/30

[a] The intersection of the C test set and DM nontraining set was used to increase the number of compounds in the intersection. [b] Calculated over the subset of compounds in the intersection. For schemes 1−3, the intersection is equal to the full test set.

(Figure 9). Classification rates for the test set mirror those for the complete set, except for RPc1 where the test set rate is substantially lower for hydantoins with hydroxy and ester/carboxylic acid substituents. Model classification rate variation across categories is mostly explained by the variation in training set composition. It can be seen that a substantially higher classification rate can be achieved if knowledge of a compound's substituent category is used to select the model for prediction. Using this approach gives an overall classification rate of 89% using the complete set of hydantoins.

As for the QSAR models, the use of consensus schemes is shown to improve classification rates (Table 7). A "unanimous agreement" scheme has been implemented whereby multiple models must agree on the activity class of compounds for an assignment to be made. Consensus schemes 1−3 show improved classification rates even when combining models derived by varying parameters for a given method and training set. Instead of using a unanimous consensus scheme, one can perform consensus analysis by majority (i.e. 2 actives, 1 inactive = active). This did not give improved classification rates, since RPc1 and Sc5 are similar and tend to vote together against Sd6.

CONCLUSIONS

In summary, a number of robust QSAR models have been derived using various subsets of the 94 compound QSAR set. While the models do not contain the same descriptors, the fundamental properties they describe are very similar. As demonstrated with the consensus scheme, they are most useful when used together. The inclusion of quadratic descriptors improved model quality only slightly for one training set and was detrimental for the others. This suggests that nonlinear effects are not prominent in the structure−activity relationship. The descriptor pool we have employed represents a wide variety of descriptor families. However, it is not all-inclusive, and more robust models may include descriptors we have not considered. Adding more descriptors has to be weighed against the increased risk of obtaining chance correlations.

A frequent criticism of QSAR models using conventional descriptors is the difficulty of extracting precise information from models that, by itself, can guide series optimization. These QSAR models must typically be applied to a virtual library of untested derivatives. In contrast, 3D-QSAR approaches such as CoMFA produce contour maps that can be used directly for guiding synthesis. While it is clear that the conclusions from CoMFA are more intuitive than those from conventional QSAR, it is generally accepted that the latter produce more predictive models. The hydantoin data set used here contains in vivo activities from independent researchers using nonuniform testing protocols. It is clear that systematic deviations exist in this data set (e.g. benchmark compounds have consistently higher activities in one source[37] and consistently lower activities in another[38]). For data sets containing in vivo activities from multiple sources, low complexity models are arguably more useful despite their more difficult interpretation.

Predictive classification models identifying similar discriminators have been derived using different methods and training sets. The simple, interpretable models explain activity across several classes of hydantoins. In contrast to SFGA, the commonly used RP method failed to produce useful models from the DM set. The incremental approach for selecting splits makes it more sensitive to the composition of the training set. To our knowledge, the SFGA method implemented in this work represents the first use of such a scheme for classifying compounds according to activity. It produced predictive models from both the C and DM training sets. Notably, models produced from each set are similar despite their small intersection (42%). Compared to RP, substantially smaller sets of compounds can be used for deriving models, making SFGA appropriate at the early stages of a screening project. The popularity of RP stems in part from its ability to handle large libraries of compounds. Despite the greater computational cost of SFGA (in this work ca. 8 min per run compared to 20 s for RP on a R10K 225 MHz processor), the possibility of selecting diverse, nonredundant subsets of compounds for deriving models should make it applicable to larger problems. Once the model is developed, classifying new compounds is substantially faster

than the Cerius2 implementation of CART (0.1 s per 1000 compounds for SFGA vs 20 s for CART).

ACKNOWLEDGMENT

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 CIHR.

Supporting Information Available: A table of all 287 hydantoins, including $\ln(1/ED_{50})$ and classification activities, set membership and literature references. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Perucca, E. The New Generation of Antiepileptic Drugs: Advantages and Disadvantages. *Br. J. Clin. Pharmacol.* **1996**, *42*, 531−543.
(2) Brown, M. L.; Zha, C. C.; Van Dyke, C. C.; Brown, G. B.; Brouillette, W. J. Comparative Molecular Field Analysis of Hydantoin Binding to the Neuronal Voltage-Dependent Sodium Channel. *J. Med. Chem.* **1999**, *42*, 1537−1545.
(3) Hadjipavlou-Litina, D. Review, Reevaluation, and New Results in Quantitative Structure−Activity Studies of Anticonvulsants. *Med. Res. Rev.* **1998**, *18*, 91−119.
(4) Dang, P.; Madan, A. K. Structure−Activity Study on Anticonvulsant (Thio) Hydantoins Using Molecular Connectivity Indexes. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1162−1166.
(5) Breiman, L.; Friedman, J. H.; Olshen, R. A.; Stone, C. J. *Classification and Regression Trees*; Wadsworth: Belmont, CA, 1984.
(6) Sun, Z. Y.; Kwon, C. H.; Wurpel, J. N. D. Structure−Activity Study of 5-Substituted 1-Carbobenzoxy-2-Iminohydantoins as Potential Anticonvulsant Agents. *J. Med. Chem.* **1994**, *37*, 2841−2845.
(7) Maple, J. R.; Dinur, U.; Hagler, A. T. Derivation of Force-Fields for Molecular Mechanics and Dynamics from Ab initio Energy Surfaces. *Proc. Natl. Acad. Sci. U.S.A.* **1988**, *85*, 5350−5354.
(8) Rappe, A. K.; Goddard, W. A. Charge Equilibration for Molecular-Dynamics Simulations. *J. Phys. Chem.* **1991**, *95*, 3358−3363.
(9) Bonchev, D. *Information Theoretic Indices for Characterization of Chemical Structures*; Research Studies Press Ltd.: New York, 1983.
(10) Kier, L. B.; Hall, L. H. *Molecular Connectivity Indices in Chemistry and Drug Research*; Academic Press: New York, 1976.
(11) 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.; 1991; pp 367−422.
(12) Kier, L. B. Shape Indexes of Orders 1 and 3 from Molecular Graphs. *Quant. Struct.-Act. Relat.* **1986**, *5*, 1−7.
(13) Kier, L. B. Distinguishing Atom Differences in a Molecular Graph Shape Index. *Quant. Struct.-Act. Relat.* **1986**, *5*, 7−12.
(14) Balaban, A. T. Highly Discriminating Distance-Based Topological Index. *Chem. Phys. Lett.* **1982**, *89*, 399−404.
(15) Balaban, A. T.; Joshi, N.; Kier, L. B.; Hall, L. H. Correlations between Chemical Structure and Normal Boiling Points of Halogenated Alkanes C1−C4. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 233−237.
(16) McQuarrie, D. A. *Statistical Thermodynamics*; University Science Books: Mill Valley, 1973.
(17) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. The Development and Use of Quantum-Mechanical Molecular-Models. 76. AM1 − a New General-Purpose Quantum-Mechanical Molecular-Model. *J. Am. Chem. Soc.* **1985**, *107*, 3902−3909.
(18) 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.
(19) 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.
(20) Ghose, A. K.; Crippen, G. M. Atomic Physicochemical Parameters for 3-Dimensional Structure-Directed Quantitative Structure−Activity-Relationships. 1. Partition-Coefficients as a Measure of Hydrophobicity. *J. Comput. Chem.* **1986**, *7*, 565−577.
(21) Ghose, A. K.; Viswanadhan, V. N.; Wendoloski, J. J. Prediction of Hydrophobic (Lipophilic) Properties of Small Organic Molecules Using Fragmental Methods: An Analysis of AlogP and ClogP Methods. *J. Phys. Chem. A* **1998**, *102*, 3762−3772.
(22) Hopfinger, A. J. *Conformational Properties of Macromolecules*; Academic Press: New York, 1973.
(23) Wold, S.; Eriksson, L. Statistical Validation of QSAR Results. In *Chemometric Methods in Molecular Design*; van de Waterbeemd, H., Ed.; VCH: New York, 1995; pp 309−317.
(24) Everitt, B. S.; Dunn, G. *Applied Multivariate Analysis*; Oxford University Press: New York, 1992.
(25) Hassan, M.; Bielawski, J. P.; Hempel, J. C.; Waldman, M. Optimization and Visualization of Molecular Diversity of Combinatorial Libraries. *Molecular Diversity* **1996**, *2*, 64−74.
(26) 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.
(27) Bursi, R.; de Gooyer, M. E.; Grootenhuis, A.; Jacobs, P. L.; van der Louw, J. et al. (Q) SAR Study on the Metabolic Stability of Steroidal Androgens. *J. Mol. Graph.* **2001**, *19*, 552−556.
(28) Chen, X.; Rusinko, A.; Young, S. S. Recursive Partitioning Analysis of a Large Structure−Activity Data Set Using Three-Dimensional Descriptors. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 1054−1062.
(29) Hawkins, D. M.; Young, S. S.; Rusinko, A. Analysis of a Large Structure−Activity Data Set Using Recursive Partitioning. *Quant. Struct.-Act. Relat.* **1997**, *16*, 296−302.
(30) Rusinko, A.; Farmen, M. W.; Lambert, C. G.; Brown, P. L.; Young, S. S. Analysis of a Large Structure/Biological Activity Data Set Using Recursive Partitioning. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1017−1026.
(31) van de Waterbeemd, H. Discriminant Analysis for Activity Prediction. In *Chemometric Methods in Molecular Design*; van de Waterbeemd, H., Ed.; VCH: New York, 1995; pp 283−293.
(32) Blower, P.; Fligner, M.; Verducci, J.; Bjoraker, J. On Combining Recursive Partitioning and Simulated Annealing to Detect Groups of Biologically Active Compounds. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 393−404.
(33) Miller, D. W. Results of a New Classification Algorithm Combining K Nearest Neighbors and Recursive Partitioning. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 168−175.
(34) Gao, H.; Lajiness, M. S.; Van Drie, J. Enhancement of Binary QSAR Analysis by a Ga−Based Variable Selection Method. *J. Mol. Graph.* **2002**, *20*, 259−268.
(35) Gasteiger, J.; Hutchings, M. G. New Empirical-Models of Substituent Polarizability and Their Application to Stabilization Effects in Positively Charged Species. *Tetrahedron Lett.* **1983**, *24*, 2537−2540.
(36) Manallack, D. T.; Pitt, W. R.; Gancia, E.; Montana, J. G.; Livingstone, D. J. et al. Selecting Screening Candidates for Kinase and G Protein-Coupled Receptor Targets Using Neural Networks. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1256−1262.
(37) Vida, J. A.; Odea, M. H.; Samour, C. M.; Reinhard, J. F. Anticonvulsants. 5. Derivatives of 5-Ethyl-5-Phenylhydantoin and 5,5-Diphenylhydantoin. *J. Med. Chem.* **1975**, *18*, 383−385.
(38) Nakamura, K.; Ohashi, K.; Nakatsuj. K.; Hirooka, T.; Fujimoto, K. et al. Anticonvulsant Activity of 3-Ethoxycarbonyl-5,5-Diphenylhydantoin (P-6127) in Animals. *Arch. Int. Pharmacodyn. Ther.* **1965**, *156*, 261−270.

CI025639W