Published in final edited form as:

J Chem Inf Model. 2006 ; 46(5): 1984–1995. doi:10.1021/ci060132x.

A Novel Automated Lazy Learning QSAR (ALL-QSAR) Approach: Method Development, Applications, and Virtual Screening of Chemical Databases Using Validated ALL-QSAR Models

Shuxing Zhang, **Alexander Golbraikh**, **Scott Oloff**, **Harold Kohn**, and **Alexander Tropsha**[*]

The Laboratory for Molecular Modeling, Division of Medicinal Chemistry and Natural Products, School of Pharmacy, CB # 7360 Beard Hall, University of North Carolina, Chapel Hill, NC 27599, USA

Abstract

A novel Automated Lazy Learning Quantitative Structure-Activity Relationship (ALL-QSAR) modeling approach has been developed based on the lazy learning theory. The activity of a test compound is predicted from locally weighted linear regression model using chemical descriptors and biological activity of the training set compounds most chemically similar to this test compound. The weights with which training set compounds are included in the regression depend on the similarity of those compounds to a test compound. We have applied the ALL-QSAR method to several experimental chemical datasets including 48 anticonvulsant agents with known ED_{50} values, 48 dopamine D_1-receptor antagonists with known competitive binding affinities (K_i), and a *Tetrahymena pyriformis* dataset containing 250 phenolic compounds with toxicity IGC_{50} values. When applied to database screening, models developed for anticonvulsant agents identified several known anticonvulsant compounds that were not only absent in the training set but highly chemically dissimilar to the training set compounds. This initial success indicates that ALL-QSAR can be further exploited as a general tool for accurate bioactivity prediction and database screening in drug design and discovery. Due to its local nature, the ALL-QSAR approach appears to be especially well suited for the development of highly predictive models for the sparse or unevenly distributed datasets.

Introduction

Many QSAR approaches have been developed during the past few decades[1–9]. The major differences between various approaches are due to structural parameters (descriptors) used to characterize molecules and the mathematical approaches used to establish a correlation between descriptor values and biological activity. Most of the modeling techniques assume a linear relationship between molecular descriptors and a target property, which may be an adequate methodology for many datasets. However, the advances in combinatorial chemistry and high throughput screening technologies have resulted in the explosive growth of the amount of structural and biological data. The explosive growth of publicly available (e.g., via the PubChem project[10]) experimental SAR data made the problem of developing robust QSAR models more challenging. This progress has provided an impetus for the development of fast, nonlinear QSAR methods that can capture structure-activity relationships for large and complex data. This laboratory among others has concentrated on the development of automated QSAR approaches with variable selection and stochastic optimization. The examples of methods include k-Nearest Neighbors (*K*NN)[11–15], Simulated Annealing-Partial Least

[*] Corresponding author, School of Pharmacy, Campus Box 7360, 327 Beard Hall, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599-7360., Telephone (919) 966-2955, FAX: (919) 966-0204, Email: Alex_Tropsha@unc.edu.

Squares (SA-PLS)[12], Support Vector Machines (SVM)[16–18], and the Automated Lazy Learning QSAR (ALL-QSAR), which is introduced in this paper.

Lazy learning methods[19–23], also known as instance/memory-based learning, defer processing of the training data until a query needs to be answered. Two typical features of this approach are the storage of training data in memory and relevant data retrieval to answer a specificquery. Locally weighted regression is an important technique that belongs to this class of learning approaches. There are several available programs, such as LOWESS and LOESS (derived from the term “locally weighted scatter plot smooth”), that have become standard mathematical statistical tools included in the S statistical package[24]. Despite their apparent popularity as general purpose statistical applications, lazy learning approaches have been rarely used in the QSAR analysis so far. For instance, Helma recently applied lazy learning concept in the rodent carcinogenicity and Salmonella mutagenicity studies[25]. However, although the name of the approach used by Helma sounds similar to ALL-QSAR, their approach should be actually regarded as modified k nearest neighbor method[25]. Kulkarni et al.[26] used locally linear embedding to reduce the nonlinear dimensions and the reduced set was subsequently modeled with robust support vector regressors. Lazy learning technique was also applied to the discovery of toxicological patterns that capture structural regularities among carcinogenic chemical compounds[20]. There have been other efforts in transductive and semi-supervised learning that are conceptually similar to our method. For example, recently Demiriz and Breneman and colleagues reported on their efforts to employ variations of block-coordinate-descent algorithms to find local solutions in their semi-supervised support vector machines (S3VM) implementation[27 29].

In most current applications of machine learning approaches to QSAR problems, a single global linear or non-linear model is typically developed to fit all of the training set data. A general global model can be developed by minimizing the following target function[22]:

$$C=\sum_{i} L(f(x_i,\beta),y_i) \tag{1}$$

where the y_i are the observed response vector (activity) values corresponding to the input vectors (descriptors) x_i, β is the coefficient vector for the model $\hat{y}_i = f(x_i, \beta)$, and $L(\hat{y}_i, y_i)$ is a general loss function (the prediction errors between $\hat{y}_i$ and y_i)[22].

In general, for large and/or diverse datasets it is practically impossible to establish a single linear or even non-linear relationship between descriptor variables and the target property in a high dimensional descriptor space that would be able to approximate the response variable satisfactorily. This is related to Richard Bellman’s famous “curse of dimensionality” concept[30]. There are two ways to solve this problem[22]: using a more complex global model or fitting a simple model to local patterns instead of the whole region of interest. Herein, we discuss the application of locally weighted linear regression[31–33] to building local linear models based on compounds in the vicinity of a query compound, which can accurately predict its target property. The method does not use a fixed number of nearest neighbors; the predictions are based on the distance-weighted activities of the training set compounds in the vicinity of a test compound. This “vicinity” depends upon the local density of compounds in the training set. Rather than building a single global model for the entire dataset this approach produces multiple local linear models that collectively form a global model, either linear or non-linear.

In addition to developing a predictive training set model, we also define the model applicability domain to avoid excessive extrapolation upon external prediction. This domain is defined as a similarity threshold between the training set compounds and a test set compound [34]. If the similarity is beyond this threshold, the prediction is considered unreliable.

As a proof of concept, the ALL-QSAR method has been applied to several experimental chemical datasets that were previously used to develop QSAR models using alternative approaches. These datasets included 48 anticonvulsant agents with known ED_{50} values that were recently studied in our group[13], 48 dopamine D_1-receptor antagonists with known competitive binding affinities (K_i)[12], and a *Tetrahymena pyriformis* dataset containing 250 phenolic compounds with toxicity IGC_{50} values[35].

The ALL-QSAR models developed for anticonvulsant agents have been applied to the hit identification via chemical database mining. The screening results demonstrate that our approach achieves the efficient detection of known anticonvulsants with chemical structures significantly different from those in the training set as well as novel structures. These initial promising results indicate that the ALL-QSAR method affords robust and externally predictive QSAR models that can be used to discover novel compounds with the desired biological profile. We expect that this novel QSAR approach can be applied to a wide variety of available experimental datasets in combination with the virtual screening using predictive QSAR models leading to the discovery of novel potent biologically active agents.

Methods

Theory of Locally Weighted Linear Regression for QSAR

Most data fitting algorithms are designed to converge to a solution that provides the best fit for the experimental data[22]; so overfilling of the model is a common problem. The ALL-QSAR approach attempts to overcome this problem by generating a series of locally weighted regression models that employ only a small fraction of compounds in the entire dataset, which are chemically similar to the query compound. The core of this approach is a simple assumption that similar compounds have similar biological activities, i.e., physicochemical properties and biological activities of molecules change concurrently with the changes in the chemical structure[36]. To date the formal definition of chemical similarity is still controversial. Similarity between chemical compounds is a fuzzy concept and often perceived intuitively based on expert judgment. As indicated recently by Jaworska et al.[37], “A chemist would describe ‘similar’ compounds in terms of ‘approximately similar backbone and almost the same functional groups’. A synthetic chemist may regard two molecules as similar when their topological descriptions of atoms and connecting bonds contain a sufficiently large number of common features.” There is no absolute measure of chemical similarity and each case should be justified for every specific activity. The general approach to unambiguously define similarity between two compounds is to evaluate the resemblance of their structures using their most informative representations[37,38]. Here we have employed the Euclidean distance in multidimensional descriptor space as the similarity metric. In our approach, the predictions for the test set compounds are made on the fly without having to produce a model *a priori*, which makes the process very fast and eliminates the need to regenerate models frequently as new data becomes available. The derivation of the related equations has been fully discussed by Atkeson and co-workers[22]. Here we give a brief introduction.

For linear regression with one variable and one output for N points,

$$\widehat{y}(x)=\beta_0+\beta_1 x \tag{2}$$

a global model, i.e. coefficients β_0 and β_1 in equation (2) can be found by residual minimization[39]:

$$\varepsilon=\sum_{k=1}^{N}(y_k-\beta_0-\beta_1 x_k)^2 \tag{3}$$

After applying weighting to each point, equation (3) can be replaced by the following equation[22]:

$$\varepsilon=\sum_{k=1}^{N} w_k^2 (y_k - \beta_0 - \beta_1 x_k)^2 \tag{4}$$

where *k* is the compound number. Similarly, as applied to multidimensional descriptor space (i.e. *M* dimensions, assuming N>=M), we will get the following regression[39]:

$$\widehat{\mathrm{y}}(\mathrm{x})=\beta_1 f_1(\mathrm{x})+\beta_2 f_2(\mathrm{x})+\ldots+\beta_{\mathrm{M}} f_{\mathrm{M}}(\mathrm{x}) \tag{5}$$

which can also be written as[22]:

$$\widehat{\mathrm{y}}(\mathrm{x})=\beta^{\mathrm{T}} f(\mathrm{x}) \tag{6}$$

where $\beta^T=(\beta_1, \beta_2,\ldots, \beta_M)$ and $f^T(x) = (f_1(x), f_2(x),\ldots f_M(x))$. After applying the weighting the equation for the residual will be established[22]:

$$\varepsilon=\sum_{\mathrm{k}=1}^{\mathrm{N}} w_{\mathrm{k}}^2 (\mathrm{y}_{\mathrm{k}} - \beta^{\mathrm{T}} f_{\mathrm{k}})^2 \tag{7}$$

Using Cholesky decomposition[22,40], β can be obtained with[39]:

$$\beta=(\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}\mathrm{X}^{\mathrm{T}}\mathrm{y} \tag{8}$$

where X^TX is an M×M matrix and X^Ty is an M-column vector. They can be calculated by the following formulas[22,39]:

$$\mathrm{X}^{\mathrm{T}}\mathrm{X}=\sum_{\mathrm{k}=1}^{\mathrm{N}} w_{\mathrm{k}}^2 f_{\mathrm{k}} f_{\mathrm{k}}^{\mathrm{T}} \tag{9}$$

$$\mathrm{X}^{\mathrm{T}}\mathrm{y}=\sum_{\mathrm{k}=1}^{\mathrm{N}} w_{\mathrm{k}}^2 f_{\mathrm{k}} \mathrm{y}_{\mathrm{k}} \tag{10}$$

Using the local linear regression approach, the activity of each compound is predicted from the activities of the most chemically similar compounds in the training set. All data on compounds' activity and their descriptors are represented in a matrix form. Each compound can be represented as a point in the multidimensional descriptor space. For a test compound, the local linear regression assigns higher weights to compounds of the training set that are closer to the test compound in the descriptor space. The distances between points are calculated using Euclidean metrics in the entire descriptor space. The weighting function (also called a kernel function) is a distance-based Gaussian function[22]:

$$w=\exp(-d^2/2K^2) \tag{11}$$

Here, *w* is the weight of a point in the training set, *d* is the distance between this point and the query, and *K* is a smoothing parameter (also known as kernel width or band width). Other techniques such as quadratic and tricube functions can also be used for weighting, as discussed by Atkeson et al.[22] A Gaussian function described in Eq. (11) decays smoothly as the distance d increases, so it can account for data distributions[22].

If necessary, ridge regression[41] can be employed to address the problem of insufficient data or singular data matrices. If the number of descriptors M is higher than the number of compounds N, or the X matrix is singular or poorly defined (i.e. $|X^TX|$ is zero or close to zero), then the ridge regression should be used[41]. In this case, equation (8) is replaced by

$$\beta(\lambda)=\left(\mathbf{X}^{\mathbf{T}}\mathbf{X}+\lambda\mathbf{I}\right)^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{y} \tag{12}$$

where λ is a small positive bias, and **I** is the M×M identity matrix. If λ=0, the ridge regression coincides with the standard least squares regression. If N<M, (M−N) additional rows are added to matrix X as follows:

$$\begin{bmatrix} a_{11}+\lambda & a_{12} & \ldots & \ldots & \ldots & \ldots & \ldots & a_{1M} \\ a_{21} & a_{22}+\lambda & a_{23} & \ldots & \ldots & \ldots & \ldots & a_{2M} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ a_{N1} & a_{N2} & \ldots & a_{NN}+\lambda & \ldots & \ldots & \ldots & a_{NM} \\ 0 & 0 & 0 & \ldots & \lambda & \ldots & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \ldots & \ldots & \ldots & \lambda \end{bmatrix}$$

Usually, 0≤ λ ≤1. Introduction of the bias solves the problem of singularity. In case of poorly defined matrices, the regression coefficients become smaller and the system stabilizes. At the same time, the residual sum of squares does not change significantly[41]. Thus, if the number of descriptors is higher than the number of compounds, adding parameter λ is equivalent to adding (M-N) dummy compounds to the dataset. First, the matrix of descriptors is appended by additional rows of zeros to become a square M×M matrix. Then λ values are added to the main diagonal elements of this squared matrix. Activities for dummy compounds are also zeros. After minimizing the locally weighted sum of squared residuals (Eq. 13),

$$C=\sum_{i} w_i^2 (f(x_i,\beta) - y_i)^2 \tag{13}$$

the coefficient (β) for each polynomial term (descriptor) is obtained. For every new query, the nearest neighbor data points and their weights change, so a different local linear model is built. Once the coefficient (β) for each descriptor is determined, the descriptor values for the query compound are entered in the regression model (Eq. 5) to obtain its activity value. Although the local model is linear, the global model can be either linear or non-linear (Figure 1). Kernel width is a very important smoothing parameter since it controls how fast the weighting function decays as the distance increases. The kernel width is optimized during the process of model building and the resulting optimal value is ultimately used in the external prediction.

ALL-QSAR also uses the applicability domain to assess whether the query is sufficiently similar to the training set to make reliable activity prediction. If the query point is outside the applicability domain, the program makes no prediction (*vide infra*).

Chemical Datasets and Molecular Descriptors

In order to demonstrate that ALL-QSAR can produce predictive models, we have applied this method to three experimental datasets studied with alternative QSAR approaches earlier: 48 anticonvulsant agents[13], 48 antagonists of the dopamine D_1 receptor[12], and 250 toxic phenol analogs[35].

The 48 anticonvulsant agents (see Supporting Information I for structures) have been studied previously using k-nearest neighbor (*K*NN) QSAR[13]. They are chemically diverse functionalized amino acid (FAA) structures and their *in vivo* ED_{50} values (mg/kg) were experimentally determined in mice using the Maximal Electroshock-Seizure-induced (MES) assay, a standard test for the anticonvulsant activity. Although the transport or metabolic properties of FAA may influence the measured ED_{50} values, they are not considered at this moment. For QSAR calculations, the compound dose values (mg/kg) were converted to the decimal logarithm of μmol/animal weight (μmol/kg).

The 48 structurally similar D_1 dopamine antagonists (see Supporting Information II) have been intensively studied by the Mailman group to characterize the D_1 dopamine receptor antagonist binding pharmacophore[42,43]. Receptor affinity was assessed by competition for [^{3}H]-SCH23390 binding site with a radioreceptor bioassay in rat striatal membranes[42,43]. All values were expressed as $pK_{i;}$ (K_i in M).

For toxicity modeling, 250 phenols[35] were used and their toxicity values were from a population growth impairment test carried out for the ubiquitous freshwater ciliated protozoan *T. pyriformis* (strain GL-C). The 50% growth inhibition concentration, IGC_{50}, was determined following the protocol previously described by Schultz[44]. The phenols are structurally heterogeneous and they are capable of being metabolized or oxidized to quinones.

MolConnZ descriptors[45–51] were used for the QSAR studies on anticonvulsant agents[13] and D_1 antagonists[12]. The MolConnZ descriptors characterize a wide range of topological and physicochemical properties of molecular structures. These descriptors include molecular connectivity indices, kappa molecular shape indices, electrotopological state indices, graph's radius and diameter, counts of different vertices, counts of paths and edges between different kinds of vertices, and many other descriptors[45 51].

A total of 160 descriptors were calculated for the phenol dataset. These descriptors, such as LogP, pK_a, electrostatic potential, volume, etc., were used to represent the physicochemical, structural and topological properties that were relevant to toxicity. They were computed using several commercially available programs; calculation details are described by Cronin and colleagues[35].

QSAR Modeling with ALL-QSAR

We employed the SE8 (Sphere Exclusion version 8) software developed in our group[34] to generate multiple chemically diverse training and test sets by splitting the original data sets. The algorithm considers each compound as a point in the multidimensional descriptor space. The procedure starts with the calculation of the distance matrix **D** between representative points in the descriptor space. Let D_{min} and D_{max} be the minimum and maximum elements of **D**, respectively. *N* sphere radii are defined by the following formulas, $R_{min}=R_1=D_{min}$, $R_{max}=R_N=D_{max}/4$, $R_i=R_1+(i-1)*(R_N$minus;$R_1)/(N-1)$, where $i=2,\ldots,N-1$. Each sphere radius corresponds to one division into the training and test set. A sphere-exclusion algorithm used in this study consisted of the following steps, (i) Select randomly a compound, (ii) Include it in the training set. (iii) Construct a sphere around this compound, (iv) Select compounds from this sphere and include them alternatively into test and training sets, (v) Exclude all compounds from within this sphere for further consideration, (vi) If no more compounds left, stop. Otherwise let *m* be the number of spheres constructed and *n* be the number of remaining compounds. Let d_{ij} ($i=1,\ldots,m$; $j=1,\ldots,n$) be the distances between the remaining compounds and sphere centers. Select a compound corresponding to the lowest dy value and go to step (ii).

The ALL-QSAR model building workflow is shown in Figure 2. The model building consists of the following steps:

1. Assign a lowest predefined value to the kernel width K, i.e., 0.01.
2. Start with the first compound in the test set and calculate the Euclidean distances between this query and all compounds in the training set. If the distance from the test set compound to its nearest neighbor is higher than some predefined threshold applicability domain (APD), no model is built since the activity prediction in this case

is considered unreliable. Then, process the next compound in the test set in a similar way; when all compounds in the test set are processed, go to step 7. If the compound of the test set is within the applicability domain, go to step 3. The applicability domain was calculated as follows. First, the average of Euclidean distances between all points of the training set was calculated. Then using the distances lower than the average, a new average distance <*d*> as well as the standard deviation σ between these distances was calculated. The applicability domain APD was defined as follows

$$APD=<d>+Z\sigma \qquad (14)$$

where Z is an empirical cutoff value which in this work was equal to 0.5.

3. The weight of every compound in the training set is calculated according to its distance to the query compound (Eq. 11).
4. Calculate coefficients β using Eq. 8 or 12.
5. Predict the target property of the query compound using Eq. 5.
6. Repeat step 2 for the next compound. If the procedure was repeated for all compounds, go to step 7.
7. Calculate the correlation coefficient between the predicted and experimental activity values of the test set compounds.
8. If kernel width is lower than the predefined value, add a predefined increment value and repeat the process starting from step 2.
9. Sort models by the R^2, starting from the highest value, and select the top 10 best models based on the R^2 values.

The ALL-QSAR modeling workflow is presented in Figure 3. First, if the number of compounds is large enough (> 100), compounds (ca. 20% of the total) are randomly excluded to form an external validation set. The remaining compounds are split into multiple training and test sets using the SE8 program as described above. Then ALL-QSAR models are built (Figure 2) and used to predict the test set. For the external validation set, consensus prediction is used, which consists of the averaging the activity of each compound predicted by all acceptable models; this approach is believed to yield more accurate predictions[15,36,52].

Robustness and Predictive Power of Models

The robustness of the models was examined by comparing them to those obtained using randomized activities of the training set (this procedure is commonly referred to as Y-randomization test[36]). Briefly, we repeated the QSAR calculations with the randomized activities of the training sets. We also compared the R^2 values in the process of the iteration procedure for actual and random activities of training sets to see if there is any significant difference. This randomization was repeated five times for each split. Models with high R^2 built with real activities of the training sets are considered acceptable and reliable only if there were no models with high R^2 built with randomized activities of the training sets.

To estimate the predictive power of a QSAR model, the following criteria were used, as discussed elsewhere[34,53], (i) Correlation coefficient R^2 between the predicted and observed activities; (ii) coefficients of determination (predicted versus observed activities R_0^2, and observed versus predicted activities $R_0'^2$); (iii) slopes k and k′ of regression lines (predicted versus observed activities, and observed versus predicted activities) through the origin. We conclude that a QSAR model has an acceptable predictive power if the following conditions are satisfied:

$$R^2>0.7 \qquad (15)$$

$$\frac{(R^2 - R_0^2)}{R^2} < 0.1 \text{ and } 0.85 < \mathrm{k} < 1.15 \tag{16a}$$

or

$$\frac{(R^2 - R'^2_0)}{R^2} < 0.1 \text{ and } 0.85 \le k' \le 1.15 \tag{16b}$$

$$|R_0^2 - R'^2_0| < 0.3 \tag{17}$$

R_0^2 is a quantity characterizing linear regression with the Y-intercept set to zero (i.e., described by Y = kX, where Y and X are actual and predictive activity, respectively) which is different from conventional R^2 for the best fit linear regression (i.e., Y = aX + b). The reason why we introduce R_0^2 and require k (or k′) be close to one is that when one compares actual vs. predicted activity, an exact fit is required, not (just) a linear correlation. Thus, a model with k = 0.9 (i.e., Y = 0.9X) and high R^2 and R_0^2 (e.g., 0.7) does have a high accuracy whereas a model described by Y = 0.8X + 2 and higher R^2 (e.g., 0.9), but low R_0^2 (e.g., 0.5) actually implies poor accuracy. Most authors ignore this caveat and present resulting statistics in the form of the best-fit linear regression between actual and predicted activities. In our opinion, this practice is insufficient, since in fact we are looking for models capable of reproducing the experimental data as opposed to producing predicted activity values that correlate with the experimental data.

Finally, acceptable models (i.e. models satisfying aforementioned conditions) were selected for further validation with the randomly pre-selected compounds (i.e., external validation set). If the prediction for this external dataset was accurate, the model was considered as validated and applicable for database mining or virtual library screening for the computational hit identification.

Ridge Regression Parameter Tuning

Calculations with different values of λ varying between 10^{-15} and 10^{-5} were performed. The value of λ which gave the highest correlation coefficient R^2 between predicted and observed activities of the test set was chosen as the model characteristic.

Database Mining using All-QSAR models for Anticonvulsant Agents

As illustrated in Figure 4, predictive ALL-QSAR models for anticonvulsant agents were used to screen chemical databases for novel anticonvulsant compounds. For this purpose, 10 best models with the highest predictive power were used. Two available databases have been explored: the Chemical Diversity (ChemDiv) database[54] and the National Cancer Institute (NCI) database[55], including 500,357 and 237,771 chemical structures, respectively. The NCI database was curated as follows: metal-containing compounds as well as compounds with incomplete chemical structures (which cannot be processed by MolConnZ) were excluded leaving 227,522 compounds for database mining. All compounds in the ChemDiv database have been used for the screening.

MolConnZ descriptors were calculated for all compounds in both databases and linearly normalized based on the minimum and maximum values of descriptors in the training set[12]. Euclidean distances between each training set compound and each compound in the database were calculated in the entire descriptor space. Database compounds within the chosen similarity cutoff value Z=0.5 (see Eq. 14) were selected as initial hits. Because of the differences of kernel widths of the models, the 10 hit lists were not identical. Thus, the initial lists were additionally refined by selecting consensus hits, i.e., molecules found in all ten individual hit lists. Then several compounds from the consensus prediction were proposed for biological testing based on their structures and medicinal chemists' experience.

Results and Discussions

Chemical Dataset Preprocessing

Three datasets were used to validate ALL-QSAR approach. For the 48 anticonvulsant agents with known ED_{50} values, 189 MolConnZ descriptors were kept after deleting those with a zero value or zero variance[13]. For the 48 dopamine D_1 receptor antagonists with known pK_i values 341 descriptors were retained for the model development[12]. To build quantitative structure-toxicity relationship (QSTR) models, 160 descriptors used in the original publication[35] were employed for the *Tetrahymena pyriformis* dataset (a set of 250 phenol compounds) with toxicity IGC_{50} values. In addition, 50 phenols were randomly excluded prior to generating training and test sets. They were set aside as an external validation set. This process was performed neither for the D_1 antagonists nor anticonvulsant datasets since both datasets were too small.

Using the SE8 software, 50 splits were made for each dataset to obtain multiple training and test sets. The splitting (as well as the prior random exclusion of 50 phenols from the *Tetrahymena pyriformis* dataset) was repeated three times in order to generate statistically significant models.

ALL-QSAR Model Development and Validation

Figure 5 shows the relationship between the ridge regression parameter (λ) and the corresponding R^2 values for one of the phenol dataset divisions. From the plot, we find that the optimal λ value is close to 10^{-9}. Interestingly, all of the datasets gave similar optimal values of this parameter: with $\lambda \leq 10^{-9}$ (small enough), the values of R^2 are the highest and most stable. Thus, $\lambda = 10^{-9}$ was used in the ALL-QSAR studies of all datasets.

Each of the three datasets was divided into multiple training and test sets, and 100 models were built for each split to optimize the R^2 values by changing kernel width from 0.01 to 1.00 with an increment of 0.01. The top 10 best models were selected for validation and prediction.

ALL-QSAR models for anticonvulsants—The black curve in Figure 6 shows the trajectory of R^2 during its optimization for the split of anticonvulsant dataset with 39 compounds in the training set and 9 compounds in the test set. With kernel width (KW) between 0 and 0.15, no predictive model was obtained. With KW between 0.15 and 0.7, several local maximum peaks have been found; and after 0.7, the prediction power of the models has converged and no further improvement was observed. We selected predictions with KW = 0.40 as our best modeling parameter. The best model obtained for this split has an $R^2 = 0.90$ (Figure 7 and Model 1 in Table 1). For another split including 14 compounds in the test set, R^2 is as high as 0.76 with KW = 0.36 (Figure 8 and Model 8 in Table 1). Statistics for the other best models are presented in Table 1. These models could be used for further prediction and validation, for example, if there is any external or new data.

ALL-QSAR models for D_1 antagonists—Similar procedures were applied to the 48 dopamine D_1 antagonists. On average, around 40 out of 100 models for each split have R^2 values higher than 0.70. In Table 2 the best 10 models for this dataset are presented. One of the best models was developed for 37 compounds in the training set and 11 compounds in the test set (Model 1 in Table 2): the correlation coefficient between predicted and observed activities of the test set was $R^2 = 0.97$ (Figure 9) with kernel width 0.17. For split with 32 and 16 compounds in the training and test set, respectively (Model 4 in Table 2), the correlation coefficient R^2 = 0.87 was obtained for the test set after excluding two compounds Ant08 and NNC01-0127 (see Supporting Information II for the structures) which were outside the applicability domain (Figure 10). The best kernel width for this split was 0.14 and other selected models are included in Table 2. Ant08 has a -Br group which was not present in the training set compounds. We found there were several bromine related descriptors for this compound and so we propose that these descriptors make the compound unique and thus not predicted by our models. NNC01-0127 incudes a seven -member ring with a positive charge on the amine. It seems this feature makes the compound different from other antagonists in the training sets.

ALL-QSAR models for toxic phenols—Compounds in this dataset were tested for IGC_{50} toxicity[35]. Since this dataset is much larger than both the anticonvulsant and D_1 antagonist datasets, a hold-out subset was selected for the external validation in order to assess the real predictive power of the obtained models. As mentioned above, the random selection of 50 phenol analogs as an external validation set was repeated three times and each time the remaining 200 phenols were split into multiple training and test sets. The model development process was performed as described in Figure 3. More than 1500 models were found to have R^2 higher than 0.70. The best 10 models were selected to make predictions of the corresponding external subset (Table 3). The results demonstrate that the models built with ALL-QSAR have high and stable predictive power for the external structures. As indicated in Figure 11 and Model 1 (Table 3), the prediction of 45 phenol analogs of the test set gave R^2 = 0.90 (training set included 155 compounds), the remaining 5 compounds were not included since they were outside the applicability domain. All of the acceptable models had R^2 values ranging between 0.71 and 0.90 for the prediction of 35 to 108 test set compounds, and correspondingly 165 to 92 training set compounds (Tables 3 and 4).

Prediction with the Consensus Method

As discussed above, this approach was only applied to the phenol dataset due to its relatively large size. With the consensus approach, the toxicity for each of the randomly selected 50 phenol compounds in the independent validation set was predicted as the average of the predicted toxicity for each compound based on the 10 best individual models (Tables 3 and 4). The results demonstrate that the consensus prediction is stable with R^2 of 0.86, 0.88 and 0.85 respectively in three repeated experiments (three random selections of the 50 phenols, Table 4). Figure 12 shows one of the consensus predictions of the 50 external compounds in our study. Several compounds were not predicted since they were outside of the applicability domain. These compounds are a series of hydroquinones or phenols substituted in the 2- or 4-position by a nitro group. They are believed to be more toxic and susceptible to oxidation[35].

Robustness and Predictive Power of the Models

In order to evaluate the model robustness, we have performed the Y-randomization test (see in Methods). The grey curve in Figure 6 shows the split (39 in the training set and 9 in the test set) but with activities in the training set randomized. It can be seen that most of the models built with real activities of the training set (black line) have significantly higher R^2 than models built with randomized activities. This Y-randomization process suggests that high R^2 of our best models is not due to a chance correlation or overfitting; i.e. our models are robust and predictive. In order to increase the statistical significance, the Y-randomization test was repeated five times for each split. With D_1 antagonists as an example, the highest R^2 for the random datasets was 0.06 while the lowest R^2 for the real datasets was 0.67. In general, if the relationships between activity and descriptors are not random, the models built with randomized activity of the training sets must have no predictive ability. Indeed, no predictive model built with randomized training set data was found for all of the datasets.

Tables 1–3 demonstrate the predictive power of the best 10 models for each dataset. These models satisfy all of the criteria (15)–(17) of predictive models. As discussed above, these models have high R^2 values. We have found that all these models have very small difference between R^2 and $R_0{}^2$ as well as $R_0{}^2$ and $R_0'^2$. The relative differences between R^2 and $R_0{}^2$ values (Eqs. 16a and b) are less than 0.1. The differences between slopes k and k′ are also small and they range between 0.85 and 1.15, as required by Eqs. 16a and b. Condition (17) is also satisfied. These statistics demonstrate that these models have very high and stable predictive power.

Comparison with Other QSAR Approaches

In order to evaluate the predictive power and accuracy of the ALL-QSAR method, our models were compared with those obtained with other QSAR approaches applied to the same datasets. Although the exactly same splits were not used for different approaches, the results are still comparable since models are developed in the same chemistry space. As is shown in Table 4, earlier calculations[13] on the 48 anticonvulsant agents using the same set of MolConnZ descriptors with variable selection *k*NN method resulted in $R^2 = 0.81$ for five anticonvulsant compounds in the test set, 0.72 for nine compounds in the test set and 0.67 for 10 compounds in the test set. With SA-PLS, the best models obtained had $R^2 = 0.77$ for seven compounds in the test set and 0.67 for eight compounds in the test set. Obviously most of our ALL-QSAR models are better than the previously developed ones, even for larger test sets and correspondingly smaller training sets (Table 1). Similarly for the 48 dopamine D_1 antagonists, one of our acceptable models had $R^2 = 0.87$ (Model 4 in Table 2) for 16 antagonists in the test set while none of *k*NN, CoMFA and SA-PLS was able to generate models with R^2 values higher than 0.70 for test sets with more than 10 compounds[12]. The SVM models for this dataset are comparable to the model developed with the ALL-QSAR method, with R^2 =0.80 for 13 test set compounds[12]. Our QSTR studies for 250 toxic phenols gave even more promising results: 1500 accepted models out of 5000 have R^2 values ranging from 0.71 to 0.90. The highest R^2=0.90 was obtained for one of the test sets with 50 compounds (Model 1 in Table 3), whereas Cronin et al. reported R^2 = 0.66~0.82 for 50 compounds using different PLS approaches[35]. In particular, the consensus predictions of the external validation sets demonstrate that these models are indeed robust and stable with average predictive $R^2 = 0.86$ (the results are summarized in Table 4).

We shall emphasize that ALL-QSAR calculations were at least 20 times faster than similar calculations with either *k*NN or SVM on the Dell OptiPlex GX270 running RedHat Linux 9.0. For example, either *k*NN or SVM took more than two days to build 2000 models for the anticonvulsant dataset while ALL-QSAR only needed about 2 hours to complete the calculations. Similar computational efficiency was observed for the 48 D_1 dopamine antagonists.

Discovery of Novel Anticonvulsants Using QSAR-Based Database Mining

As illustrated in Figure 13, 2000 ALL-QSAR models were developed using MolConnZ descriptors and 10 models with the highest predictive power were used for database mining of novel anticonvulsant agents. Initial hits included a total of 2920 compounds within the applicability domain, which means they are similar to the training set compounds. 399 compounds were consensus hits: 326 of them from ChemDiv and 73 from NCI repositories, respectively. Supporting Information IX lists some of the 73 chemical structures and their predicted anticonvulsant activities. They were also identified previously with the *k*NN approach[13]. Two compounds not found by ALL-QSAR but predicted by *k*NN[13] are also presented (see Supporting Information).

The anticonvulsant activities of these 399 molecules were then predicted using the 10 ALL-QSAR models with applicability domains specific to each model. This procedure resulted in the final selection of 91 compounds, which were found within all individual applicability domains, with high predicted activity (less than 60 mg/kg). The criterion (60 mg/kg) is arbitrary but based on the NIH standard which considers the anticonvulsant activity (MES ED_{50} in mice) promising if it is less than 100 mg/kg. The results of the individual database mining study are described below.

NCI Database Screening

Application of the 10 best individual ALL-QSAR models to the NCI database mining resulted in 73 consensus hits, and their activities were predicted by all 10 models. Only 22 compounds with consistently high predicted anticonvulsant activity values (MES ED_{50} less than 60 mg/kg) were selected. Among these compounds 9 hits were the same as predicted earlier using *k*NN-QSAR method[13] whereas two identified earlier by *k*NN were missing. One of the missing hits was shown to be inactive (ID number: 655432, see Supporting Information IX).

One of the consensus hits and four additionally designed analogs were chosen previously for synthesis and experimental testing[13]. Briefly, the compound (C4) (Table 5) initially selected had a fully-substituted N-terminal amide group, a functional group usually not providing anticonvulsant activity[13]. Another important feature of this compound was the terminal carbobenzyloxy (Cbz) group, which was not present in the training set (Table 5)[13]. Several analogous compounds (C1–C3, C5) of C4 were designed *de novo*. In total, five compounds were synthesized and submitted to the NIH's Anticonvulsant Screening Program (ASP) for the MES test, which was also used for the training set compounds. Four compounds (C1, C3–C5) have been confirmed active and one (C2) was inactive, which is consistent with our ALL-QSAR prediction, while *k*NN gave a false positive prediction for (C2)[13]. The chemical structures of all compounds and the available testing results at this time are shown in Table 5.

The anticonvulsant activities for C1–C5 demonstrate the ability of ALL-QSAR and our mining approach to identify molecules from large databases with chemical substructures different from those observed in the FAA training set. It shows that our ALL-QSAR models built for anticonvulsant agents have high predictive power which affords the correct identification of experimental hits similar to those identified earlier with *k*NN[13].

ChemDiv Database Screening

Originally 326 compounds were identified from 500,357 compounds in the ChemDiv database. Based on their activity profile ($ED_{50} < 60$ mg/kg) 69 structures were selected. Importantly and interestingly, the compounds have very high structural diversity. Based on the structures, they have been grouped into several families. It turns out that one of the families, as is shown in Figure 14, is related to Dimmock's semicarbazone analogues which are potent anticonvulsant agents[56,57]. These anticonvulsant pharmacophores do not exist in the training set compounds and the hits were not ever identified by any of our previous studies. Our groups are now in the process of planning the organic synthesis and experimental evaluation of the most promising hit and their derivatives and the results of our experimental studies will be reported in the forthcoming publications.

Conclusions

We have developed a novel QSAR modeling approach, Automated Lazy Learning (ALL)-QSAR. This method represents a query-oriented algorithm that uses locally weighted linear regression for model development. We have demonstrated that ALL-QSAR is a computationally efficient and effective algorithm to develop statistically robust and predictive models. The advantages of the ALL-QSAR method are as follows. 1) The method can be used to develop non-linear models for large and diverse datasets. ALL-QSAR models combine the simplicity of linear models with the complexity of global non-linear models by describing the latter models as a set of simple locally weighted linear models. 2) No fixed number of nearest neighbors is used; the predictions are made using distance-based weighting and depend on the data density. 3) It is computationally efficient. The ALL-QSAR method has been successfully tested on several experimental datasets, and we will continue to explore it as a general approach to build predictive QSAR models. More importantly, these models have been successfully applied to computational hit identification for anticonvulsant agents via large chemical database mining. The screening results demonstrate that our approach identified molecules with chemical substructures highly dissimilar to those observed in the training set. The predicted high anticonvulsant activity of some hits was confirmed by experiments. These promising results suggest that ALL-QSAR can be exploited as a general tool for the design and discovery of novel, potent biologically active compounds.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

The authors express their gratitude to the COMP Division of the American Chemical Society and Chemical Computing Group (CCG) for their grant to S.Z. allowing him to present the preliminary results of this study at the 228th ACS National meeting in Philadelphia in 2004. The study was supported in part by the NIH research grant GM066940 and the NIH Center Planning grant P20HG003898. We thank Drs. T. Netzeva and M. Cronin for providing the phenol dataset.

References

1. Dietrich SW, Dreyer ND, Hansch C, Bentley DL. Confidence-Interval Estimators for Parameters Associated with Quantitative Structure-Activity-Relationships. J Med Chem 1980;23:1201–1205. [PubMed: 7452669]
2. Hadjipavloulitina D, Hansch C. Quantitative Structure-Activity-Relationships of the Benzodiazepines - A Review and Reevaluation. Chem Rev 1994;94:1483–1505.
3. Hansch C, Muir RM, Fujita T, Maloney PP, Geiger E, Streich M. The Correlation of Biological Activity of Plant Growth Regulators and Chloromycetin Derivatives with Hammett Constants and Partition Coefficients. J Am Chem Soc 1963;85:2817–2824.
4. Hansch C, Kurup A, Garg R, Gao H. Chem-bioinformatics and QSAR: A review of QSAR lacking positive hydrophobic terms. Chem Rev 2001;101:619–672. [PubMed: 11712499]
5. Hansch C, Leo A, Mekapati SB, Kurup A. Qsar and Adme. Bioorg Med Chem 2004;12:3391–3400. [PubMed: 15158808]
6. Klein TE, Huang C, Ferrin TE, Langridge R, Hansch C. Computer-Assisted Drug Receptor Mapping Analysis. ACS Symposium Series 1986;306:147–158.
7. Kubinyi H. Quantitative Relationships Between Chemical-Structure and Biological-Activity. Chemie in Unserer Zeit 1986;20:191–202.
8. Kubinyi H. QSAR and 3D QSAR in drug design.1 methodology. Drug Discovery Today 1997;2:457–467.
9. Kurup A, Mekapati SB, Garg R, Hansch C. HIV-1 protease inhibitors: A comparative QSAR analysis. Curr Med Chem 2003;10:1679–1688. [PubMed: 12871116]
10. PubChem Project. 2006. http://pubchem.ncbi.nlm.nih.gov/
11. Zheng WF, Tropsha A. Novel variable selection quantitative structure-property relationship approach based on the k-nearest-neighbor principle. J Chem Inf Comput Sci 2000;40:185–194. [PubMed: 10661566]

12. Oloff S, Mailman RB, Tropsha A. Application of validated QSAR models of D1 dopaminergic antagonists for database mining. J Med Chem 2005;48:7322–7332. [PubMed: 16279792]
13. Shen M, Beguin C, Golbraikh A, Stables JP, Kohn H, Tropsha A. Application of predictive QSAR models to database mining: Identification and experimental validation of novel anticonvulsant compounds. J Med Chem 2004;47:2356–2364. [PubMed: 15084134]
14. Oloff S, Zhang S, Sukumar N, Breneman C, Tropsha A. Chemometric analysis of ligand receptor complementarity: identifying Complementary Ligands Based on Receptor Information (CoLiBRI). J Chem Inf Model 2006;46:844–851. [PubMed: 16563016]
15. Zhang S, Golbraikh A, Tropsha A. Development of quantitative structure-binding affinity relationship models based on novel geometrical chemical descriptors of the protein-ligand interfaces. J Med Chem 2006;49:2713–2724. [PubMed: 16640331]
16. Xue CX, Zhang RS, Liu HX, Yao XJ, Liu MC, Hu ZD, Fan BT. An accurate QSPR study of O-H bond dissociation energy in substituted phenols based on support vector machines. J Chem Inf Comput Sci 2004;44:669–677. [PubMed: 15032549]
17. Yao XJ, Panaye A, Doucet JP, Zhang RS, Chen HF, Liu MC, Hu ZD, Fan BT. Comparative study of QSAR/QSPR correlations using support vector machines, radial basis function neural networks, and multiple linear regression. J Chem Inf Comput Sci 2004;44:1257–1266. [PubMed: 15272833]
18. Kovatcheva A, Golbraikh A, Oloff S, Xiao YD, Zheng WF, Wolschann P, Buchbauer G, Tropsha A. Combinatorial QSAR of ambergris fragrance compounds. J Chem Inf Comput Sci 2004;44:582–595. [PubMed: 15032539]
19. Aha DW. Lazy learning. Artif Intell Rev 1997;11:7–10.
20. Armengol E, Plaza E. Discovery of toxicological patterns with lazy learning. Knowledge-Based Intellignet Information and Engineering Systems, Pt 2, Proceedings 2003;2774:919–926.
21. Armengol E, Plaza E. Relational case-based reasoning for carcinogenic activity prediction. Artif Intell Rev 2003;20:121–141.
22. Atkson CG, Moore AW, Schaal S. Locally weighted learning. Artif Intell Rev 1997;11:11–73.
23. Wettschereck D, Aha DW, Mohri T. A review and empirical evaluation of feature weighting methods for a class of lazy learning algorithms. Artif Intell Rev 1997;11:273–314.
24. Cleveland WS. Lowess - A Program for Smoothing Scatterplots by Robust Locally Weighted Regression. American Statistician 1981;35:54.
25. Helma C. Lazy structure-activity relationships (lazar) for the prediction of rodent carcinogenicity and Salmonella mutagenicity. Mol Divers 2006:1–12. [PubMed: 17123027]Online First
26. Kumar R, Kulkarni A, Jayaraman VK, Kulkarni BD. Structure-Activity Relationships using Locally Linear Embedding Assisted by Support Vector and Lazy Learning Regressors. Internet Electron J Mol Des 2004;3:118–133.
27. Breneman, C. Bioinformatics Workshop. Rensselaer Polytechnic Institute; New York, NY: Drug-Design through Semi Supervised Learning; p. 999
28. Demiriz, A.; Bennett, KP.; Embrechts, MJ. Semi-Supervised Clustering Using Genetic Algorithms. In: Dagli, CH., editor. Intelligent Engineering Systems through Artificial Neural Networks. ASME Press; New York: 1999. p. 809-814.
29. Demiriz, A.; Bennet, K. Optimization approaches to semisupervised learning. In: Ferris, MC.; Mangasarian, OL.; Pang, J-S., editors. Applications and algorithms of complementarity. Kluwer Academic Publishers; Boston: 2000. p. 1-19.
30. Bellman, R. Adaptive Control Processes: A Guided Tour. Princeton University Press; New Jersey, NJ: 1961.
31. Atkeson, CG.; Reinkensmeyer, DJ. Using associative content-addressable memories to control robots. 1. Austin; Texas: 1988. p. 792-797.
32. Atkeson, CG. Memory-based approaches to approximating continous functions. Casdagli and Eubank; 1992. p. 503-521.
33. Atkeson CG, Moore AW, Schaal S. Locally weighted learning for control. Artif Intell Rev 1997;11:75–113.

34. Golbraikh A, Shen M, Xiao ZY, Xiao YD, Lee KH, Tropsha A. Rational selection of training and test sets for the development of validated QSAR models. J Comput Aided Mol Des 2003;17:241–253. [PubMed: 13677490]

35. Cronin MTD, Aptula AO, Duffy JC, Netzeva TI, Rowe PH, Valkova IV, Schultz TW. Comparative assessment of methods to develop QSARs for the prediction of the toxicity of phenols to Tetrahymena pyriformis. Chemosphere 2002;49:1201–1221. [PubMed: 12489717]

36. Tropsha A, Gramatica P, Gombar VK. The importance of being earnest: Validation is the absolute essential for successful application and interpretation of QSPR models. QSAR Comb Sci 2003;22:69–77.

37. Nikolova N, Jaworska J. Approaches to Measure Chemical Similarity - a Review. QSAR Comb Sci 2003;22:1006–1026.

38. Willett P. Chemoinformatics - similarity and diversity in chemical libraries. Curr Opin Biotechnol 2000;11:85–88. [PubMed: 10679335]

39. Rencher, AC. Methods of Multivariate Analysis. John Wiley & Sons.; New York, NY.: 2002. p. 1-738.

40. Press, WH.; Flannery, BP.; Teukolsky, SA.; Vetterling, WT. Numerical Recipes in C: The Art of Scientific Computing. Cambridge University Press; New York, NY: 1992. p. 1-1020.

41. Draper, NR.; Smith, H. Applied Regression Analysis. John Wiley; New York, NY: 1981. p. 1-709.

42. Wyrick SD, Myers AM, Booth RG, Kula NS, Baldessarini RJ, Mailman RB. Synthesis of [N-(Ch3)-H-3]-Trans-(1R,3S)-(-)-1-Phenyl-3-N,N-Dimethylamino-1,2,3,4-Tetrahydronaphthalene (H-2-Pat). Journal of Labelled Compounds & Radiopharmaceuticals 1994;34:131–134.

43. Minor DL, Wyrick SD, Charifson PS, Watts VJ, Nichols DE, Mailman RB. Synthesis and Molecular Modeling of 1-Phenyl-1,2,3,4-Tetrahydroisoquinolines and Related 5,6,8,9-Tetrahydro-13Bh-Dibenzo[A,H]Quinolizines As D-1 Dopamine Antagonists. J Med Chem 1994;37:4317–4328. [PubMed: 7996543]

44. Schultz TW, Bearden AP, Jaworska JS. A novel QSAR approach for estimating toxicity of phenols. SAR QSAR Environ Res 1996;5:99–112. [PubMed: 8751817]

45. MolConnZ. MolConnZ Version 405. Quincy, MA: Hall Associates Consulting; 2002.

46. Hall LH, Mohney B, Kier LB. The Electrotopological State - An Atom Index for Qsar. Quant Struct -Act Relat 1991;10:43–51.

47. Hall LH, Kier LB. 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.

48. Hall LH, Kier LB, Brown BB. Molecular Similarity feed on Novel Atom-Type Electrotopological State Indexes. J Chem Inf Comput Sci 1995;35:1074–1080.

49. Hall LH, Kier LB. Issues in representation of molecular structure - The development of molecular connectivity. J Mol Graph Model 2001;20:4–18. [PubMed: 11760002]

50. Kier LB, Murray WJ, Hall LH. Molecular connectivity. 4. Relationships to biological activities. J Med Chem 1975;18:1272–1274. [PubMed: 1238571]

51. Kier, LB.; Hall, LH. Molecular Connectivity in Chemistry and Drug Research. Academic Press; New York: 1976.

52. Perez C, Pastor M, Ortiz AR, Gago F. Comparative binding energy analysis of HIV-1 protease inhibitors: incorporation of solvent effects and validation as a powerful tool in receptor-based drug design. J Med Chem 1998;41:836–852. [PubMed: 9526559]

53. Golbraikh A, Tropsha A. Beware of q(2)! J Mol Graph Model 2002;20:269–276. [PubMed: 11858635]

54. ChemDiv. 2005. http://www.chemdiv.com

55. NCI. 2004. http://dtp.nci.nih.gov/docs/3d_database/structural_information/smiles_strings.html

56. Dimmock JR, Puthucode RN, Smith JM, Hetherington M, Quail JW, Pugazhenthi U, Lechler T, Stables JP. (Aryloxy)aryl semicarbazones and related compounds: A novel class of anticonvulsant agents possessing high activity in the maximal electroshock screen. J Med Chem 1996;39:3984–3997. [PubMed: 8831764]

57. Dimmock JR, Vashishtha SC, Stables JP. Anticonvulsant properties of various acetylhydrazones, oxamoylhydrazones and semicarbazones derived from aromatic and unsaturated carbonyl compounds. Eur J Med Chem 2000;35:241–248. [PubMed: 10758285]

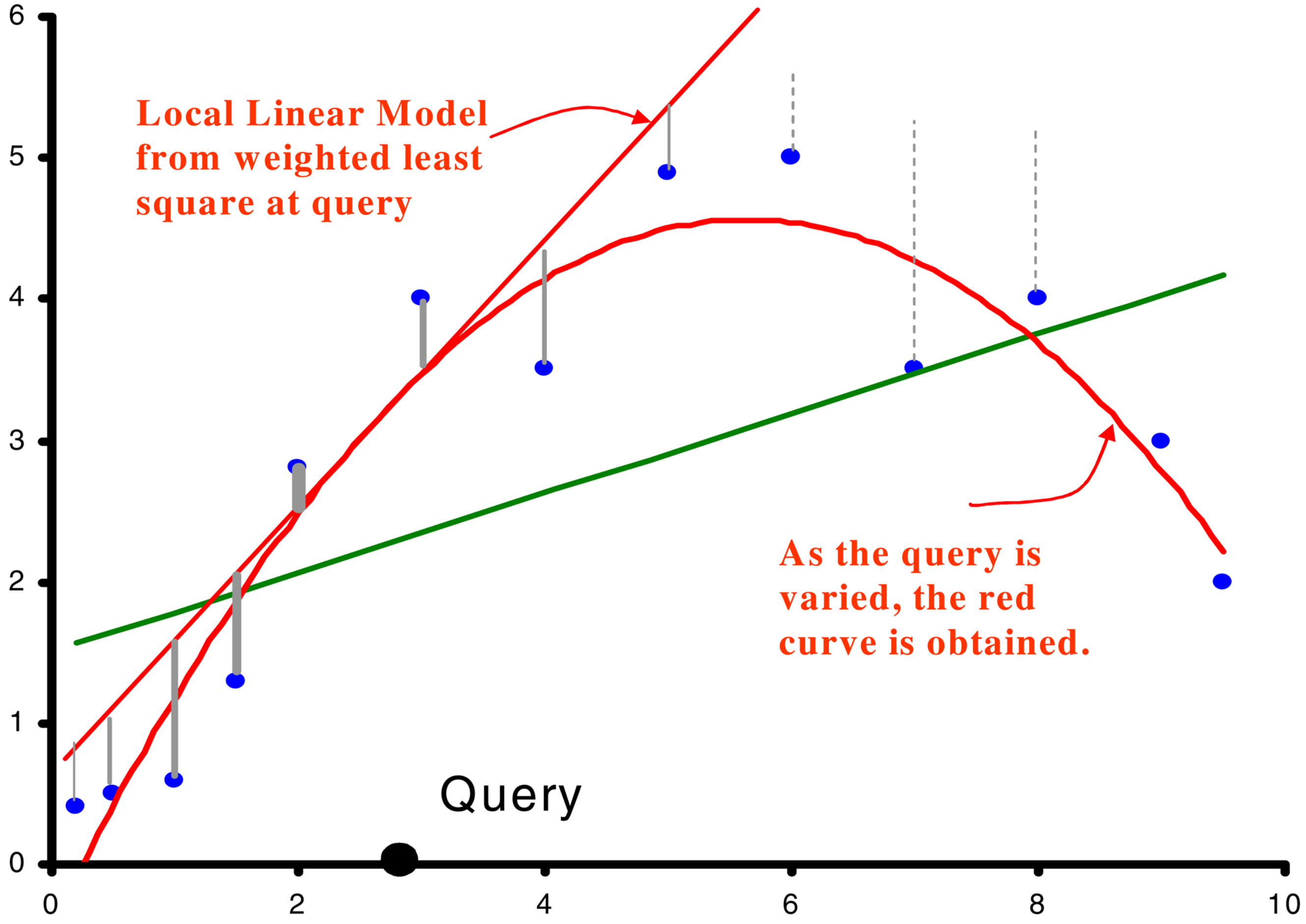


Figure 1.

Locally weighted regression. The Figure highlights the difference between the global linear regression and the locally weighted linear regression. The green line is the global linear regression and the red straight line is the weighted linear regression, where the thickness of gray lines indicates the strength of the weight. The red curve line is the final function obtained after combining local linear regressions for all the points.

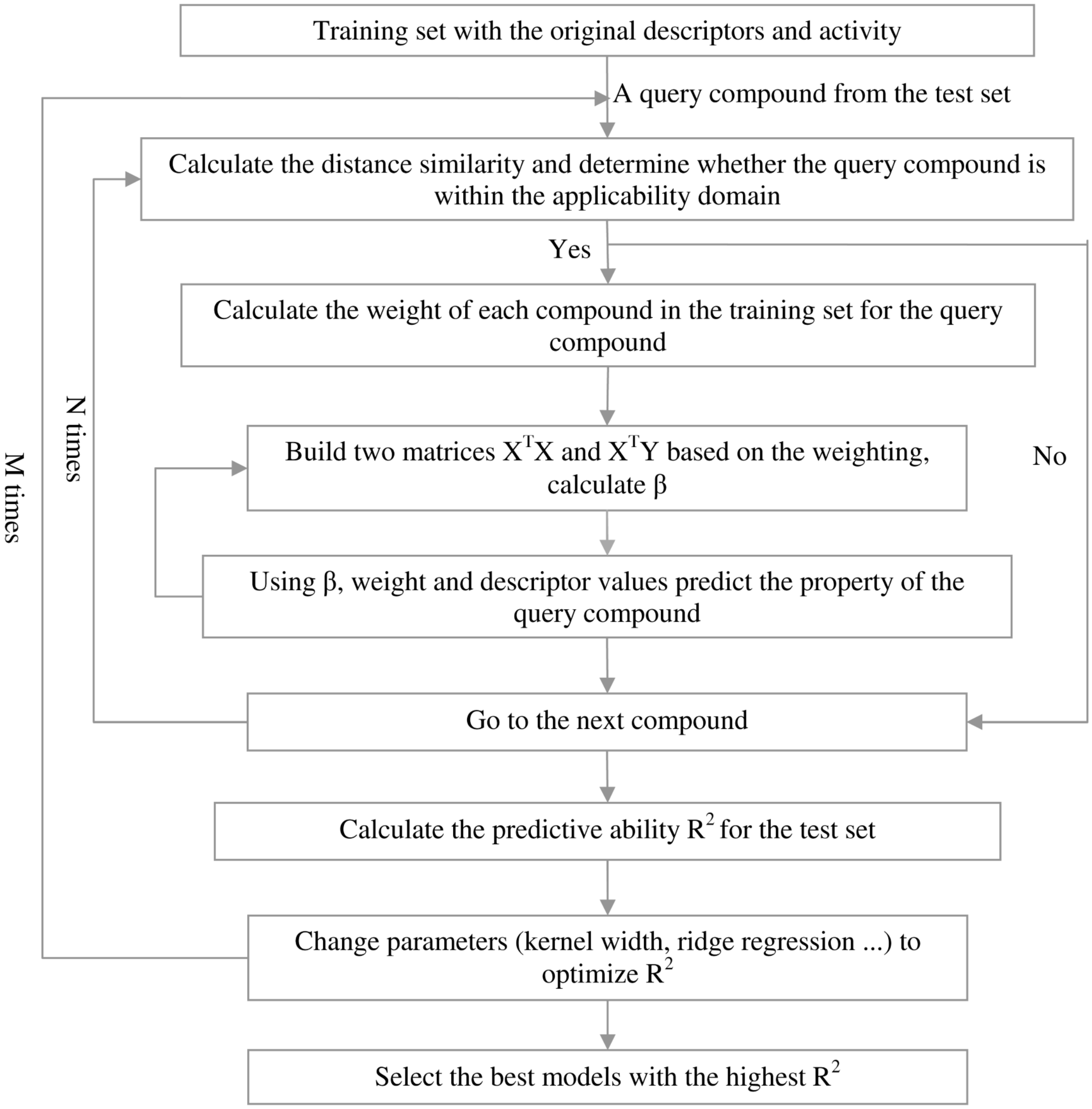

Training set with the original descriptors and activity
A query compound from the test set
Calculate the distance similarity and determine whether the query compound is within the applicability domain
Yes
Calculate the weight of each compound in the training set for the query compound
Build two matrices X^TX and X^TY based on the weighting, calculate β
Using β, weight and descriptor values predict the property of the query compound
Go to the next compound
No
N times
M times
Calculate the predictive ability R^2 for the test set
Change parameters (kernel width, ridge regression ...) to optimize R^2
Select the best models with the highest R^2

Figure 2.
Flowchart of the ALL-QSAR method.

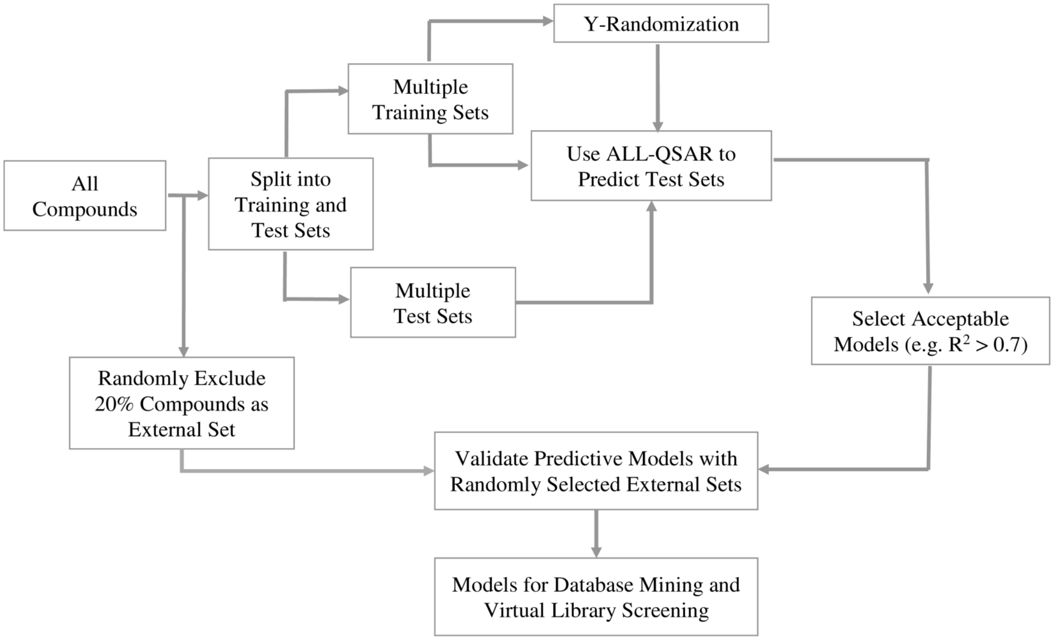

Y-Randomization
Multiple Training Sets
All Compounds
Split into Training and Test Sets
Use ALL-QSAR to Predict Test Sets
Multiple Test Sets
Select Acceptable Models (e.g. R^2 > 0.7)
Randomly Exclude 20% Compounds as External Set
Validate Predictive Models with Randomly Selected External Sets
Models for Database Mining and Virtual Library Screening

Figure 3.

The ALL-QSAR statistical modeling workflow.

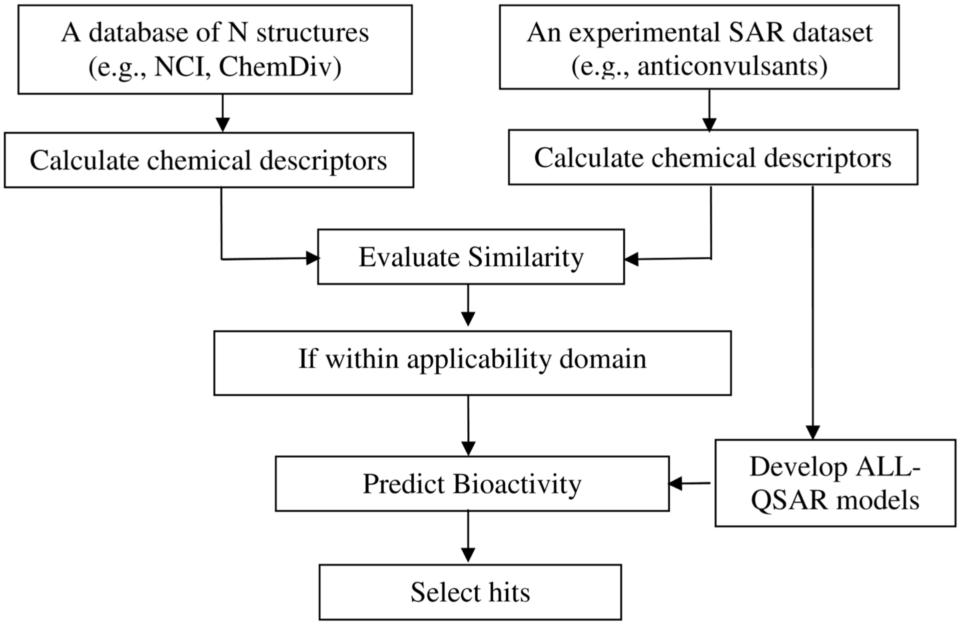

A database of N structures (e.g., NCI, ChemDiv)
An experimental SAR dataset (e.g., anticonvulsants)
Calculate chemical descriptors
Calculate chemical descriptors
Evaluate Similarity
If within applicability domain
Predict Bioactivity
Develop ALL-QSAR models
Select hits

Figure 4.
Flowchart of database mining that employs predictive ALL-QSAR models.

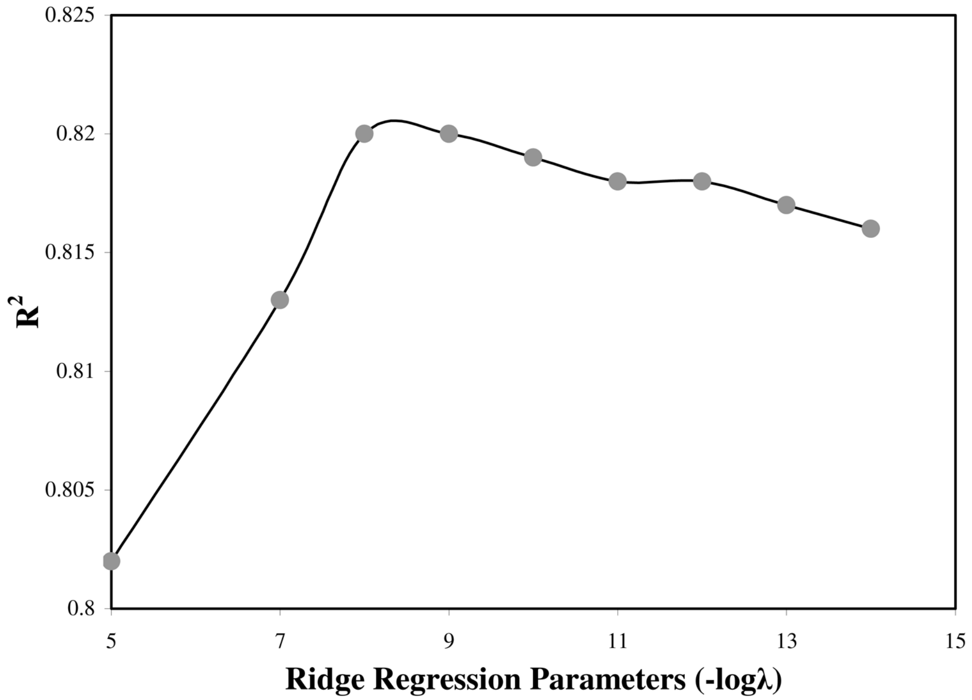


Figure 5.
The correlation between the ridge regression parameter (λ) and the R^2 for one of the Phenol test sets.

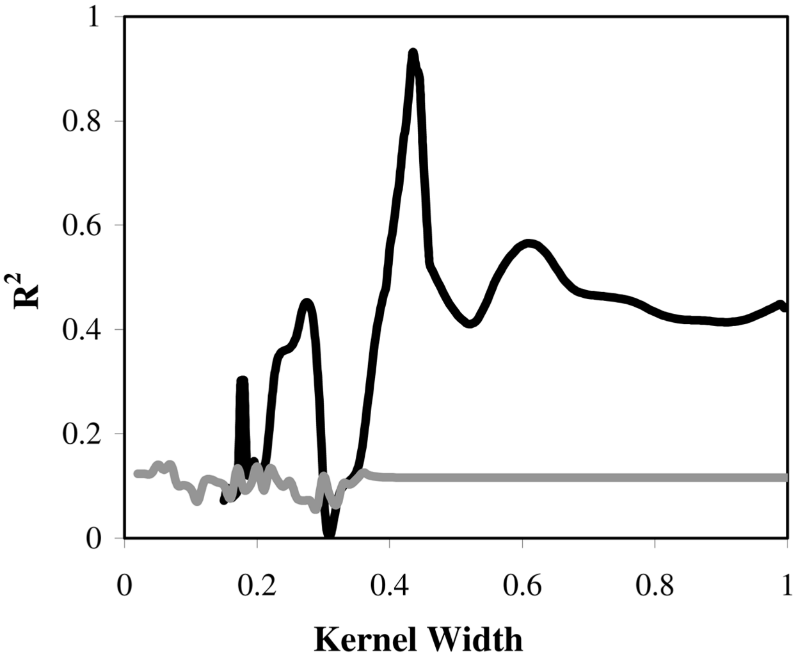


Figure 6.
R^2 trajectory with respect to the kernel width during the model development for 39 anticonvulsant agents in the training set and 9 compounds in the test set. Iterations are shown for the real dataset (black) and the dataset with activity randomized (gray).

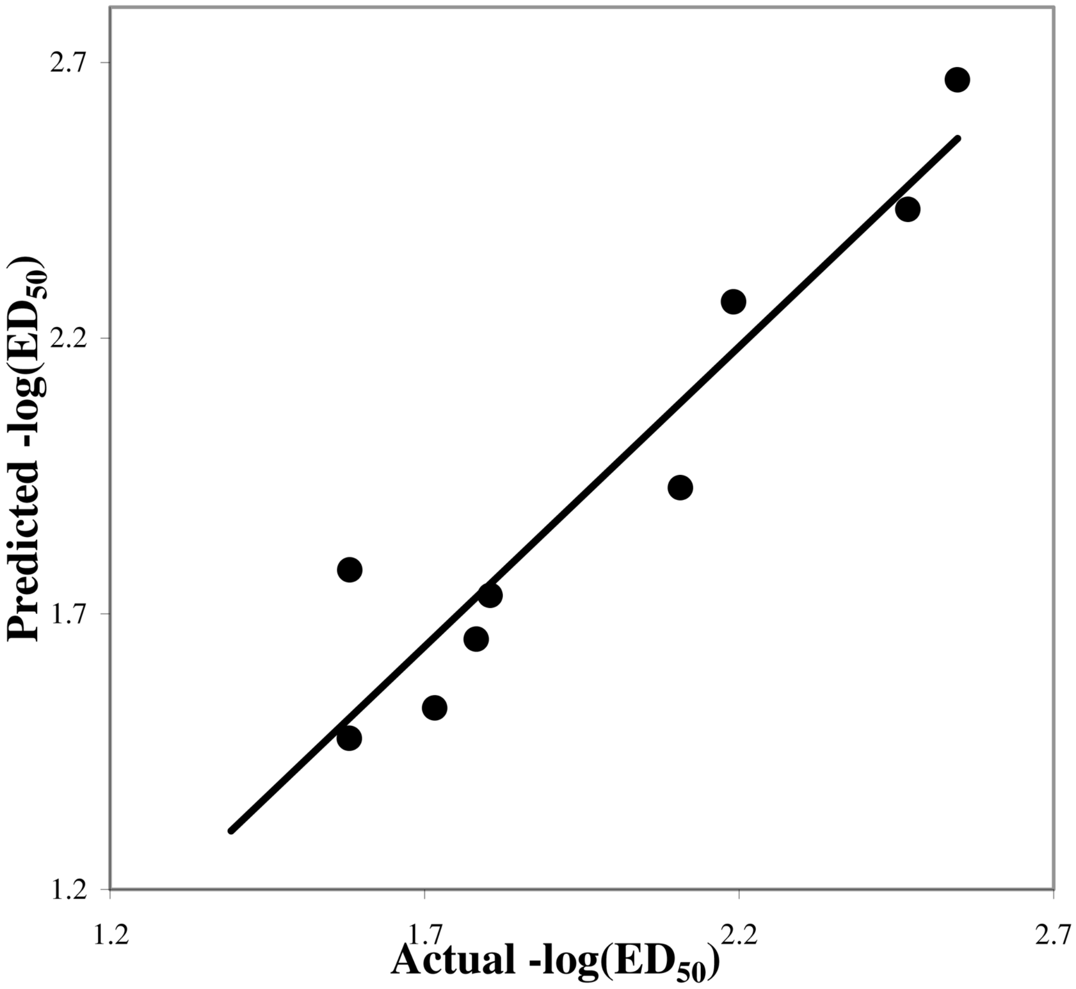


Figure 7.
Activity prediction with ALL-QSAR models for 9 anticonvulsants in the test set. R^2 = 0.90 (Model 1 in Table 1).

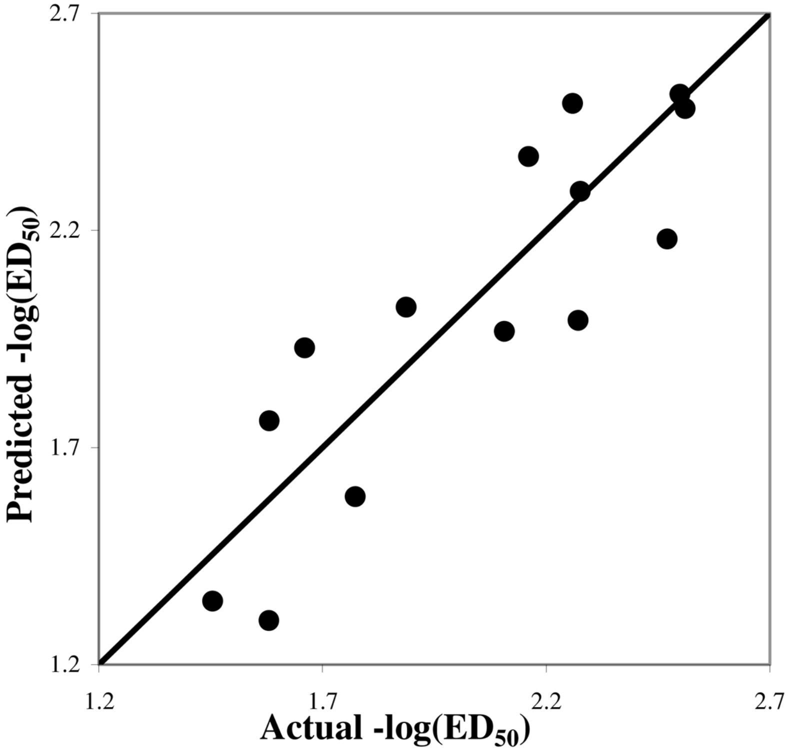


Figure 8.

Activity prediction with ALL-QSAR models for 14 anticonvulsants in the test set. R^2 = 0.76 (Model 8 in Table 1).

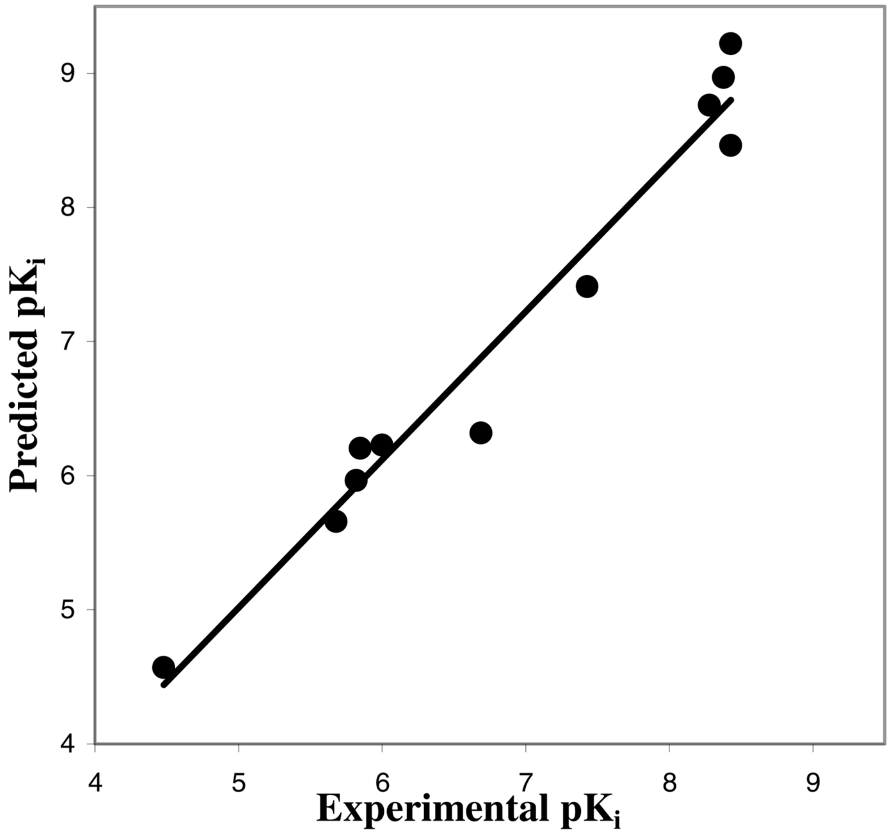


Figure 9.
Correlation between experimental and predicted pK_i for 11 D_1 antagonists in the test set. Training set included 37 compounds. $R^2 = 0.97$ (Model 1 in Table 2)

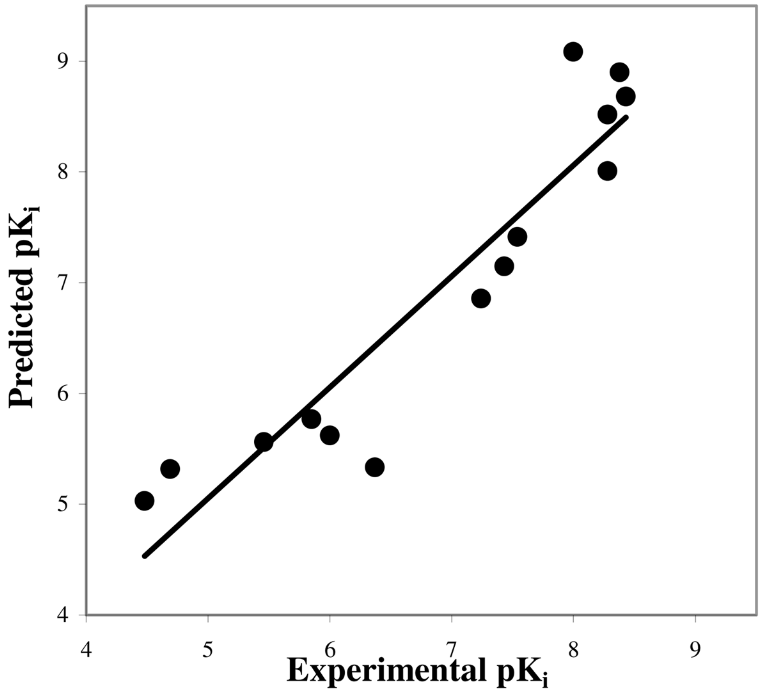


Figure 10.
Correlation between experimental and predicted pKi for 14 D1 antagonists in the test set. Training set included 32 compounds. R2 = 0.87 (Model 4 in Table 2). Two compounds, Ant08 and NNC01-0127, are outside of the applicability domain and not shown in the plot.

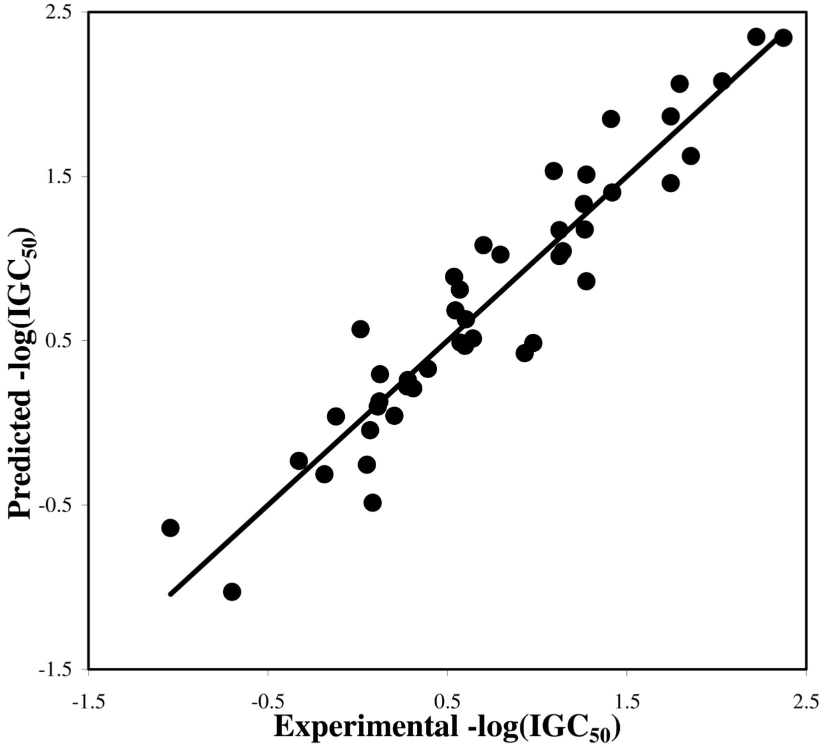


Figure 11.
The best ALL-QSAR model with 150 phenols in the training set: $R^2 = 0.90$ for the prediction of 50 compounds in the test set (Model 1 in Table 3).

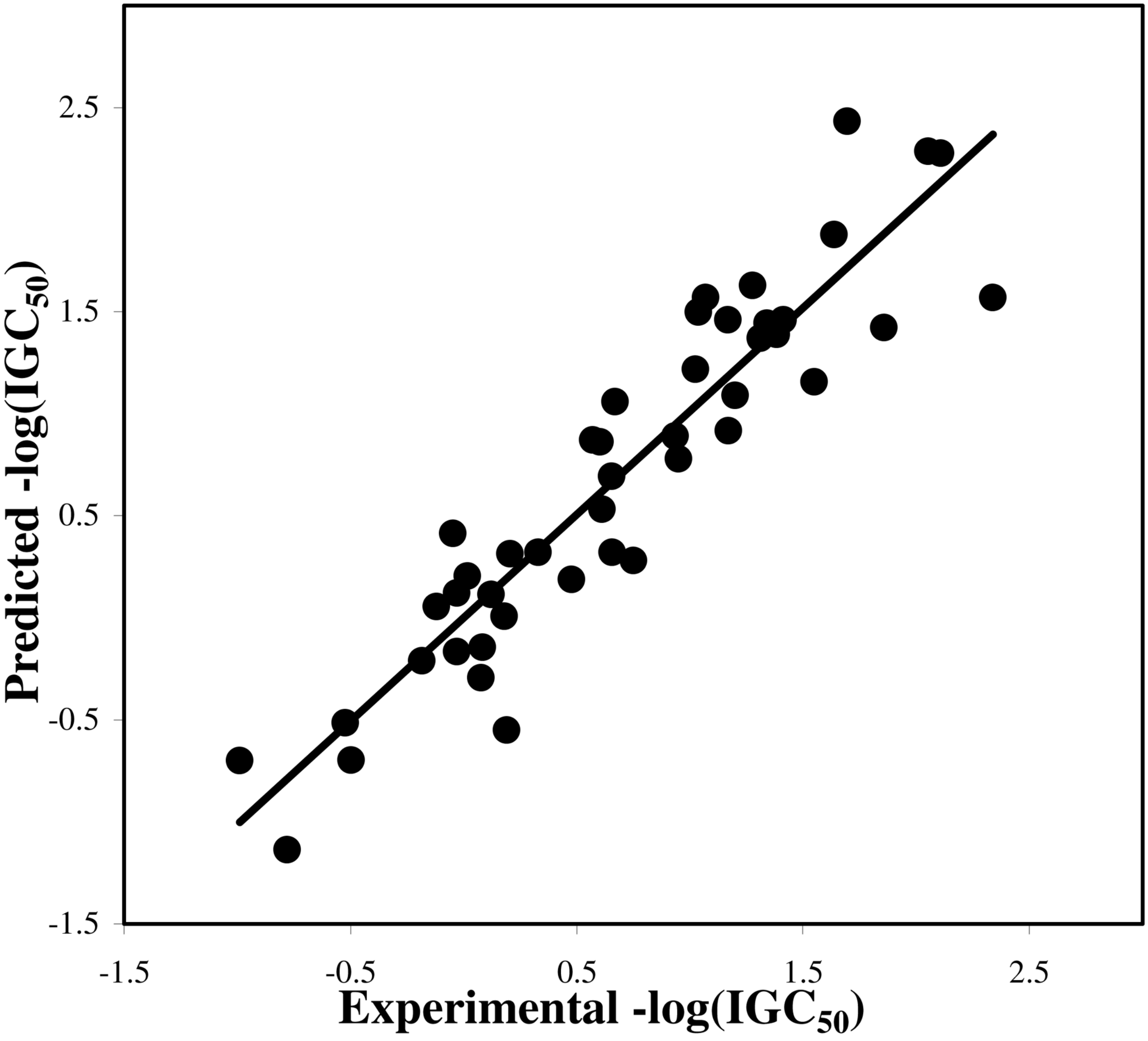


Figure 12.

The consensus prediction of 50 external toxic phenol compounds with the 10 best ALL-QSAR models affords high accuracy of prediction with $R^2 = 0.86$ (Table 3 and 4).

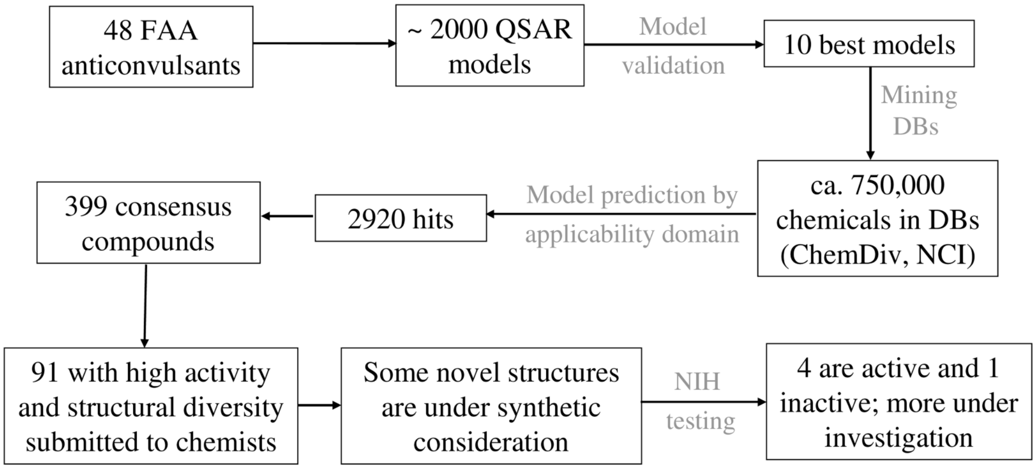

48 FAA anticonvulsants
~ 2000 QSAR models
Model validation
10 best models
Mining DBs
ca. 750,000 chemicals in DBs (ChemDiv, NCI)
Model prediction by applicability domain
2920 hits
399 consensus compounds
91 with high activity and structural diversity submitted to chemists
Some novel structures are under synthetic consideration
NIH testing
4 are active and 1 inactive; more under investigation

Figure 13.
Workflow for the identification of novel anticonvulsant agents using consensus database mining.

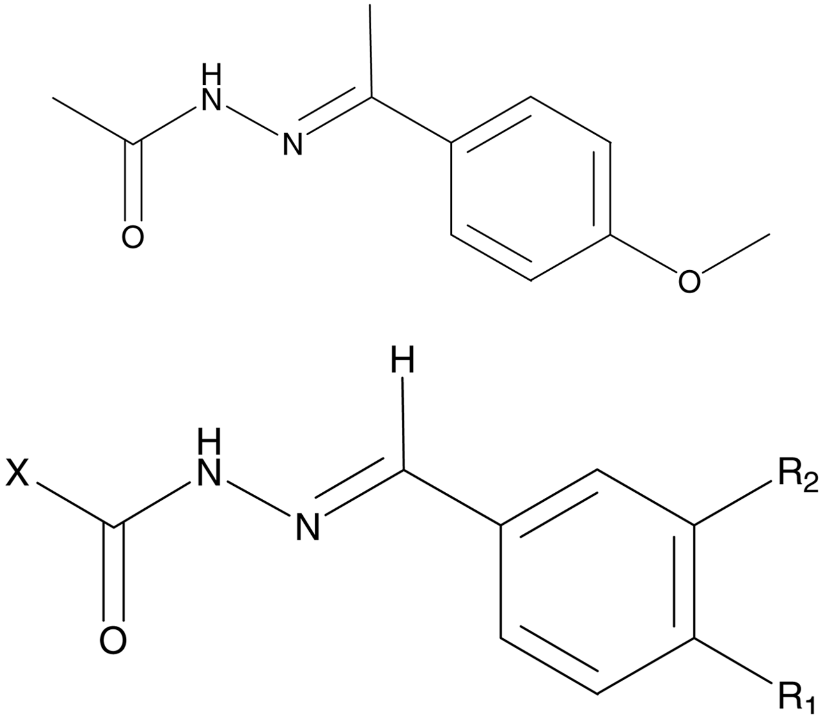


Figure 14.

One of the structures identified in virtual screening (top) and Dimmock’s semicarbazone scaffold (bottom)[56].

Table 1

The best 10 models for the anticonvulsant dataset

Model No.	Splits (Training/Test)	KW	R^2	R_0^2	$R_0'^2$	k	k′	$(R^2-R_0^2)/R^2$
1	39/9	0.40	0.90	0.89	0.86	0.99	1.01	0.011
2	38/10	0.41	0.88	0.86	0.85	0.97	0.99	0.023
3	40/8	0.43	0.86	0.85	0.84	0.95	0.93	0.012
4	39/9	0.41	0.86	0.84	0.83	0.94	0.99	0.023
5	37/11	0.34	0.83	0.82	0.83	0.97	0.92	0.012
6	38/10	0.45	0.81	0.85	0.81	1.10	0.93	0.049
7	36/12	0.37	0.80	0.79	0.78	0.97	1.05	0.013
8	34/14	0.36	0.76	0.76	0.71	0.99	1.00	0.000
9	32/16	0.39	0.75	0.74	0.75	0.91	1.04	0.013
10	32/16	0.39	0.71	0.70	0.68	0.93	1.09	0.014

Table 2

The best 10 models for the pi antagonist dataset

Model No.	Splits (Training/Test)	KW	R^2	R_0^2	$R_0'^2$	k	k′	$(R^2-R_0^2)/R^2$
1	37/11	0.17	0.97	0.96	0.95	1.03	0.97	0.010
2	39/9	0.14	0.95	0.93	0.92	1.08	0.95	0.021
3	37/11	0.13	0.91	0.90	0.91	0.98	1.02	0.011
4	32/16	0.14	0.87	0.87	0.85	1.01	0.98	0.000
5	38/10	0.11	0.86	0.86	0.84	0.95	0.98	0.000
6	40/8	0.16	0.86	0.83	0.81	1.10	0.95	0.035
7	36/12	0.13	0.82	0.81	0.79	1.02	0.93	0.012
8	34/14	0.14	0.82	0.75	0.80	1.13	0.90	0.085
9	34/14	0.15	0.77	0.72	0.74	1.02	0.98	0.065
10	30/18	0.14	0.70	0.70	0.68	1.01	0.95	0.000

Table 3

The best 10 models for the toxic phenol dataset

Model No.	Splits (Training/Test)	KW	R^2	R_0^2	$R_0'^2$	k	k′	$(R^2-R_0^2)/R^2$
1	150/50	0.51	0.90	0.90	0.89	1.00	0.94	0.000
2	164/36	0.55	0.88	0.87	0.88	1.03	0.98	0.011
3	149/51	0.46	0.83	0.81	0.82	0.98	0.99	0.024
4	132/68	0.57	0.83	0.83	0.83	0.97	0.92	0.000
5	165/35	0.48	0.81	0.79	0.75	1.08	0.95	0.025
6	151/49	0.51	0.80	0.79	0.79	0.98	0.99	0.013
7	154/46	0.46	0.77	0.73	0.76	0.96	1.01	0.052
8	146/54	0.50	0.77	0.70	0.76	1.05	0.93	0.091
9	134/66	0.55	0.73	0.69	0.71	1.15	0.94	0.055
10	141/59	0.51	0.71	0.65	0.69	0.87	1.16	0.085
Consensus External Prediction	200/50	N/A	0.86	0.85	0.83	1.01	0.90	0.011

Table 4

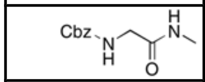
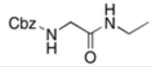
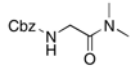
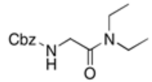
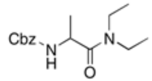
Comparison of ALL-QSAR to other approaches for three chemical datasets.

Methods	Training Set Size	Test Set Size	R^2	Consensus R^2 for External (10 Best Models)
48 Anticonvulsan it Agents				
*k*NN[13]	39	9	0.72	-
kNN[13]	38	10	0.67	-
SA-PLS[13]	40	8	<0.67	-
ALL-QSAR	39	9	0.90	-
ALL-QSAR	34	14	0.76	-
48 D1 Antagonist C Compounds				
kNN[12]	40	8	0.76	-
SVM[12]	35	13	0.80	-
SA-PLS[12]	40	8	0.63	-
CoMFA[12]	40	8	0.45	-
ALL-QSAR	36	12	0.81	-
Cronin’s 250 Phenol Compounds				
Cronin et al.[35]	200	50	0.66–0.82	-
*k*NN[a]	207	43	0.79	-
ALL-QSAR	97–160	40–103	0.71–0.90	0.88
ALL-QSAR	100–165	35–100	0.70–0.87	0.86
ALL-QSAR	92–164	36–108	0.68–0.83	0.85

[a] Golbraikh, A. unpublished data.

Table 5

The results of anticonvulsant activity testing from the Anticonvulsant Screening Project at the National Institutes of Health. Partly adopted from Shen *et al*[13].

			Mice (ip)			Rats (po)	
Structure	**ID**	**MP**	**MES$_{exp}$, ED$_{50}$ (mg/kg)**	**MES$_{kNN}$, ED$_{50}$ (mg/kg)**	**MES$_{ALL-QSAR}$ ED$_{50}$ (mg/kg**	**MES, ED$_{50}$ (mg/kg**	**Tox, TD$_{50}$ (mg/kg**
	C1	105–106	>30, <100	64.3	65.53	<30	>100 (ip)
	C2	99–100	>100, <300	35.0	103.07	52 [1.0]	>500
	C3	57–58	43 [0.25] (41–46)	41.3	33.35	<30	>30
	C4	40–41	52 [0.25] (51–53)	17.0	32.23	~30	>30
	C5	oil	74 [0.25] (69–79)	27.7	9.00	<30	>30