Development of Linear, Ensemble, and Nonlinear Models for the Prediction and Interpretation of the Biological Activity of a Set of PDGFR Inhibitors

Rajarshi Guha and Peter C. Jurs*

Department of Chemistry, Penn State University, University Park, Pennsylvania 16802

Received May 4, 2004

A QSAR modeling study has been done with a set of 79 piperazyinylquinazoline analogues which exhibit PDGFR inhibition. Linear regression and nonlinear computational neural network models were developed. The regression model was developed with a focus on interpretative ability using a PLS technique. However, it also exhibits a good predictive ability after outlier removal. The nonlinear CNN model had superior predictive ability compared to the linear model with a training set error of 0.22 log(IC_{50}) units ($R^2 = 0.93$) and a prediction set error of 0.32 log(IC_{50}) units ($R^2 = 0.61$). A random forest model was also developed to provide an alternate measure of descriptor importance. This approach ranks descriptors, and its results confirm the importance of specific descriptors as characterized by the PLS technique. In addition the neural network model contains the two most important descriptors indicated by the random forest model.

INTRODUCTION

The investigation of anti-cancer drugs has focused on a number of targets. One of the initial focus areas was compounds that could interfere in DNA synthesis and function. Essentially such compounds stimulated apoptotic pathways. Such a self-destructive approach is limited in terms of efficiacy and selectivity. An alternative approach that has been the target of intense research is the development of compounds that are able to interfere with cellular signal transduction mechanisms. Cell growth is one area in which signal transduction plays a vital role. Essentially, growth factors bind to specific cell surface receptors initiating a cascade of events which lead to activation of genes or other growth mechanisms. An important class of growth receptors are the receptor tyrosine kinases (RTK's). This class of kinase is a member of the family known as protein tyrosine kinases which transmit growth signals via a phosphorylation mechanism.[1] The structures of RTK's consist of three parts—a ligand binding region on the cell membrane, a region spanning the cell membrane, and tyrosine kinase domains within the cell.[2−4] Four main RTK's are known, and platelet derived growth factor receptor (PDGFR) is the RTK that is considered in this work.

A large number of compounds have been investigated as putative PDGFR inhibitors. Examples include 1-phenylbenzimidazoles,[5] arylquinoxalines,[6] piperazinylquinazolines,[4] and various pyrimidine analgoues.[7−9] The mode of action of PDGFR inhibitors is competition with ATP binding at the intracellular kinase domains. Thus the biological activity of prospective inhibitors can be investigated with phosphorylation assays. Much experimental work has been carried out on this family of proteins, and a number of QSAR studies have been carried out as well. Kurup et al.[1] conducted an extensive review of QSAR models for tyrosine kinase inhibitors (including PDGFR). All the models reported were linear in nature and were developed using a limited number of descriptors. Shen et al.[10] developed a series of linear regression models for the set of 1-phenylbenzimidazoles described by Palmer[5] using electronic descriptors and a PLS routine to build the final models.

This study involves the development of a set of linear as well as nonlinear models to predict and interpret the biological activity of a set of piperazinylquinazolines investigated by Pandey.[4] The data set consists of 79 molecules with the biological activity reported as IC_{50} values. Activity values were obtained from a phosphorylation assay with and without human plasma. The original study investigated the structure−activity trend of these molecules experimentally, but no computational models were developed. We note that Khadikar et al.[11] have reported a QSAR study using this data set. However, their study was restricted to linear regression models using topological descriptors only. Furthermore, they restricted themselves to using IC_{50} values from the assay in the absence of human plasma. The models we present concentrate on the biological activity values obtained from the assay in the presence of human plasma. Furthermore we present results from linear regression models as well as nonlinear computational neural network models. We used a wide variety of descriptors rather than restricting ourselves to single classes. Finally, in addition to prediction, the linear model is subjected to a PLS-based interpretation method to explain the structure−activity trend embodied in it.

DATA SET

The data set investigated consisted of 79 molecules that were derivatives of 4-piperazinylquinazolines and were investigated for their ability to inhibit PDGFR phosphorylation.[4] The structures of these molecules have been presented by Pandey et al.[4] The molecules were evaluated for their inhibition of βPDGFR phosphorylation in MG63 cells.[4,12] The assays were carried out both in the presence and absence of human plasma resulting in two sets of IC_{50}values. For the purposes of this study, these were converted to −log(IC_{50}) values. However a number of the measurements made

10.1021/ci049849f CCC: $27.50
Published on Web 09/14/2004

in the absence of plasma were reported as '< 0.004'. Since this indicates that the response was possibly below the limit of detection, these molecules would have to be ignored for the purposes of model building, thus decreasing the size of the data set. Hence we only considered the set of measurements made in the presence of human plasma thus allowing the use of all 79 compounds.

METHODOLOGY

This study used the Automated Data Analysis and Pattern Recognition Toolkit[13,14] (ADAPT) to calculate descriptors and develop QSAR models. The ADAPT methodology allows for the development of linear models (using multiple linear regression) and nonlinear models (using computational neural networks, CNN). In addition a random forest[15] model was built using the R software package.[16] A brief description of the ADAPT methodology is presented, followed by a description of the random forest technique.

The first step was to divide the molecules into three sets—the training, cross-validation, and prediction set (known as QSAR sets). Molecules in the training set were used to develop linear and nonlinear models. The cross-validation set is used in the case of nonlinear models to monitor neural network training. In the case of linear models and random forest models, these training and cross-validation sets are combined. The prediction set is not used during model development but serves to test the predictive ability of the final models. The ADAPT methodology uses an activity binning method to create QSAR sets whereby the molecules are binned based on their activity values and the sets are populated by a weighted selection from the bins. This resulted in a training set containing 57 molecules, a cross-validation set containing 9 molecules, and a prediction set containing 13 molecules.

Next, the 3D molecular structures were entered using Hyperchem[17] and geometry optimized using MOPAC 7.01 with the PM3 Hamiltonian. After structures were optimized molecular descriptors were calculated. The ADAPT package calculates nearly 350 descriptors. The descriptors can be classified into three classes. Geometric descriptors encode structural features of a molecule and require an accurate 3D representation of the molecule. Examples of this class include moment of inertia,[18] molecular surfaces, and volumes.[19] Topological descriptors also encode structural information. However in contrast to geometric descriptors this class of descriptors considers the molecule as a mathematical graph and subsequently calculates topological and graph invariants. Examples include path lengths, distance edge vectors,[20] and connectivity indices.[21−24] The third class of descriptors are electronic descriptors characterizing the electronic environment of a molecule. Examples include HOMO and LUMO energies and electronegativity. In addition to the three classes mentioned above ADAPT also calculates hybrid descriptors which are combinations of the above types of descriptors. Examples include the charged partial surface area descriptors[25] and hydrophobicity descriptors.[26]

A total of 321 descriptors were calculated for these 79 compounds. Many of the raw descriptors contain little information or are correlated with other descriptors. Furthermore, a general rule of thumb is that the final number of descriptors should be such that the ratio of molecules to descriptors is approximately 6, so as to prevent chance correlations. Thus the next step is to reduce the number of descriptors via objective feature selection. This procedure is used to reduce the initial descriptor pool to a more manageable size by using statistical methods which ignore the dependent variable. Objective feature selection in ADAPT involves the use of a correlation test (where if a given pair of descriptors have a Pearson correlation greater than a user specified cutoff, a random member of the pair is deleted) and an identical test (where descriptors which contain a user specified percentage of identical values are deleted). In this study the correlation cutoff and identical cutoff were set to 0.75 and 0.75, respectively. As a result the original descriptor pool was reduced to 41 descriptors. This reduced descriptor pool was used to generate the linear and nonlinear models discussed in this study. For the ensemble model using random forest, the original 321 descriptor pool and the reduced pool were used.

The next stage involves subjective feature selection in which a simulated annealing algorithm[27] or a genetic algorithm[28,29] was employed to search for optimal descriptor subsets to build linear and nonlinear models. In contrast to objective feature selection, subjective feature selection uses the dependent variable to search for the best subsets of descriptors. The search routines mentioned above are stochastic in nature. Simulated annealing is a computational analogy of the annealing process in glasses and metals first described by Metropolis et al.[30] In a traditional annealing process the system starts out at a high temperature in a highly disordered state. The system is then allowed to slowly cool (such that thermal equilibrium is maintained) to a more ordered state of lower energy. In the ideal case, the system would approach the perfectly ordered state at a temperature of 0 K. When this idea is applied to a combinatorial descriptor search problem, the state of the system is represented by a subset of descriptors, and the energy of the system is described by an objective function. The result of the algorithm is that at the end of the *cooling schedule* the system will be in its lowest energy state—that is, the final subset of descriptors will generate the optimum value of the objective function. One of the important aspects of this algorithm is the definition of the temperature in the context of a combinatorial problem. In addition, aspects of the cooling schedule—iterations, initial temperature, and temperature decrement also play an important role. The details of the ADAPT implementation of the simulated annealing algorithm can be found in refs 27 and 31. The genetic algorithm is a combinatorial optimization scheme that is based on biological evolutionary mechanisms. In this method an initial pool of descriptor subsets (termed chromosomes) of a specified length are randomly generated. Next the *fitness* of each member (defined by its chromosome) of the pool is evaluated via an objective function and members ranked on the basis of their fitness. Next, genetic operations are applied: members of the population are allowed to mate. The mating process involves two steps. First, crossover occurs whereby the parts of the chromosome of two mating members are swapped with each other. Second, mutation is allowed to occur whereby a single part of the chromosome (i.e., a single descriptor) is randomly changed to another descriptor. The result of mating is that a new set of individuals are created. This set constitutes the new genera-

tion, and the process is repeated until the maximum observed fitness converges. In terms of combinatorial optimization this procedure leads to descriptor subsets of maximum fitness, that is, subsets which optimize the value of the objective function. The algorithm has many variable features—including the nature of crossover, probabilities of crossover and mutation, and methods to prevent premature convergence. The specific details of the ADAPT implementation of the genetic algorithm are described in ref 29.

As mentioned above, both types of search algorithms require an objective function to be optimized. In ADAPT, the objective function is either a multiple linear regression routine (for linear models) or a three layer, fully connected, feed forward CNN routine (for nonlinear models). In the former case, the value of the objective function is represented by the root-mean-square error (RMSE). Thus the optimization algorithms search for descriptor subsets that lead to linear regression models with the lowest RMSE. An additional constraint is that linear models are only accepted if their *t* statistic is greater than 4.0, so as to avoid statistically irrelevant models. When the objective function is a CNN, the simulated annealing algorithm or genetic algorithm selects descriptor subsets that minimize the value of a RMSE based cost function defined[32,33] as

$$\text{Cost} = \text{TSET}_{\text{RMS}} + 0.5|\text{TSET}_{\text{RMS}} - \text{CVSET}_{\text{RMS}}| \quad (1)$$

Once a set of low cost CNN models was obtained a more rigorous analysis was performed to obtain the optimal CNN parameters. In both optimization methods, a list of top performing descriptor subsets are generated. These subsets are then analyzed in more detail to obtain the final optimal QSAR model.

RANDOM FORESTS

The random forest[15] technique is an example of ensemble learning techniques (which include boosting[34,35] and bagging[36]). More specifically it is an extension of the recursive partitioning[37,38] algorithm. Recursive partitioning (also known as decision trees) is useful in QSAR modeling due to its ability to handle high dimensional data and ignore irrelevant descriptors.[39] However one disadvantage of this algorithm is that it usually leads to poor predictive ability. The random forest technique was developed by Breiman[15] and is essentially an ensemble of decision trees. Mathematically a random forest may be denoted as

$$R = \{T_1(\mathbf{X}), T_2(\mathbf{X}) \cdots T_B(\mathbf{X})\}$$

where $T_i(\mathbf{X})$ is a single decision tree and $\mathbf{X}$ represents a single molecular descriptor vector. A random forest is built in two steps. First from the whole data set of *n* molecules a bootstrap sample is generated. In the second step a decision tree is built using this bootstrap sample. These two steps are repeated until the desired number, *B*, of trees have been generated. The algorithm for the development of a tree differs from the traditional recursive partitioning algorithm in that at each node the split is determined by a set of *m* randomly selected descriptors from the descriptor pool.

The end result of this training procedure is an ensemble of trees. When such an ensemble is applied to a regression problem, the predicted value for a given molecule is the average of predictions made by the individual trees. A more detailed mathematical analysis of random forests is given by ref 15. The main utility of random forests in this study was to investigate whether it would be able to provide any information regarding the importance of specific descriptors.

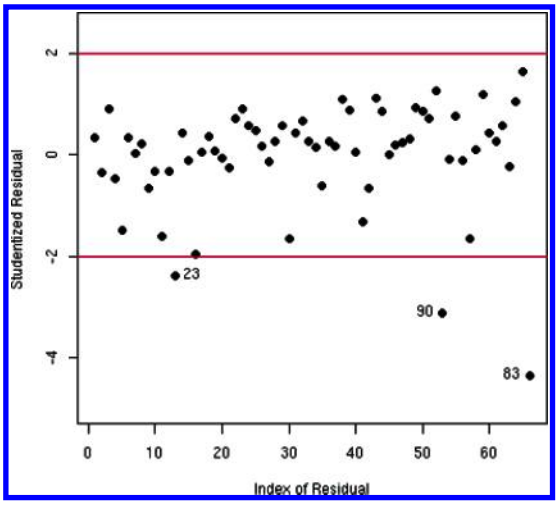

Figure 1. A plot of the studentized residuals from the 3-descriptor linear model with outliers marked.

For this study the R software package was used to develop a random forest model. The R implementation allows the user to specify several parameters. We focused on the number of descriptors randomly selected to split nodes on and the minimum node size (that is, the minimum number of members in a node, below which a node is not split). In general the defaults in the R implementation of the random forest algorithm lead to good models. However, we performed a grid search to find optimal values of the parameters using the tune function from the e1071 package in R.

RESULTS

Linear Models. A series of linear models was developed using the genetic algorithm to search for optimal descriptor subsets. A training set of 68 compounds was used initially. The best model obtained was a 9-descriptor model. However, it exhibited poor regression statistics (no *t*-values were greater than 3.0 and *p* values for the coefficients were on the order of 0.01). Furthermore, none of the models except the 9-descriptor model were validated when investigated using a PLS analysis. A 3-descriptor model with similar statistics but a much lower R^2 and RMSE was also investigated. One aspect of these two models as well as nearly all the models developed using the GA was that three compounds (**23, 83,** and **90**) were consistently flagged as training set outliers. Outliers were detected by plotting studentized residuals versus the compound index for each of the linear models developed. An example of the residual plot for the 3-descriptor model is shown in Figure 1. Apart from the molecules mentioned above, some models usually had one or two other molecules which could be considered as borderline outliers. However since these borderline molecules varied from model to model, we did not consider them further. Since the outliers mentioned above were found in nearly all the models that were generated, we felt it was justified to remove them from the training pool and to reexamine the models. Thus, the training set now contains 65 molecules. One common feature

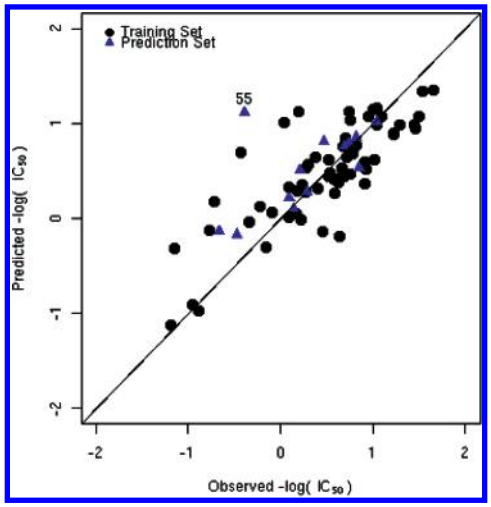

Figure 2. A plot of observed versus predicted $-\log(IC_{50})$ values from the best linear model after training set outliers were removed. The annotated point represents a prediction set outlier.

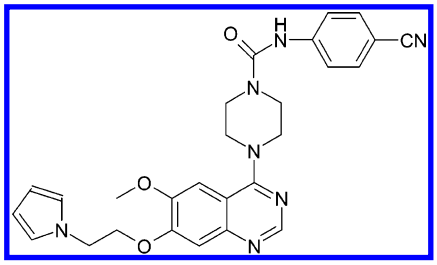

Figure 3. The structure of the prediction set outlier (**55**) from the best linear and nonlinear CNN models.

of these molecules that may justify their removal is that the 6 position on the quinazoline ring in these compounds has an ethoxy group (in the case of **83** and **90**) or a hydrogen (in the case of **23**), whereas the majority of compounds have longer (bulkier) functional groups at this position. Furthermore in the case of **83** and **90**, the 7 position does have a ring moiety at the end of the four-membered chain and thus can be considered relatively bulky. However as will be discussed later, such a feature (bulky groups attached to long chains at the 6 and 7 positions on the quinazoline moiety) is characteristic of compounds with high activity, whereas these compounds have quite low values of activity. This observation is supported to some extent by the fact that in Figure 1, molecules **83** and **90** are significantly more outlying than **23**. The statistics of all the models improved and most models were validated using the PLS technique (including the 9- and 3-descriptor models mentioned previously). Since the aim of a modeling technique is parsimony, we chose to present the results and an interpretation of the 3-descriptor model.

A plot of the observed versus predicted $-\log(IC_{50})$ values for the 3-descriptor model (with training set outliers removed) is shown in Figure 2. The statistics of the model are summarized in Tables 1 and 2. The ranges of the descriptors used are shown in Table 3. The R^2 for the model was 0.65 and the RMSE was 0.38. The value of the *F*-statistic was 37.06 (on 3 and 59 degrees of freedom) compared to a critical value of 2.76 (at the 0.05 significance level) with a *p*-value of 1.4×10^{-13}. Finally the variance inflation factors for all the descriptors was less than 1.6 indicating the absence of collinearities in the model. For the prediction set the R^2 was

Table 1. Regression Statistics for the Best Linear Regression Model[a]

descriptor	beta	standard error	*t*	*P*	VIF
constant	0.50529	0.0499	10.129	1.59×10^{-14}	
MDEN-23	0.13957	0.0516	2.703	8.97×10^{-3}	1.23
RNHS-3	0.23205	0.0501	4.576	2.49×10^{-5}	1.26
SURR-5	−0.43415	0.0529	−8.196	2.56×10^{-11}	1.12

[a] MDEN-23 − molecular distance edge vector between secondary and tertiary nitrogens;[20] RNHS-3 − relative hydrophilic surface area[26] defined as the product of the sum of the hydrophilic constants and surface area of the most hydrophilic atom divided by overall log P; SURR-5 − the ratio of atomic constant weighted hydrophobic (low) surface area to the atomic constant weighted hydrophilic surface area.[26,42]

Table 2. Summary of Overall Statistics for the Best Linear Regression Model

	number of molecules	RMSE	R^2
training set	65	0.38	0.65
prediction set	13	0.47	0.38

Table 3. Ranges of the Descriptors Used in the Best Linear Regression Model

descriptor	maximum	minimum	mean
MDEN-23	7.466	1.784	2.796
RNHS-3	−1.637	−37.726	−4.752
SURR-5	−1.633	−4.423	−3.180

0.38 and the RMSE was 0.47. Though the RMSE is not significantly higher than for the training set, the low value of R^2 is influenced by the prediction set outlier noted in Figure 2. Removal of this compound (**55**) from the prediction set resulted in a R^2 of 0.84 and RMSE of 0.24. The structure of the outlier is shown in Figure 3. A simple comparison of the structure of the outlier with others in the data set does not reveal why it would be predicted poorly. However the PLS analysis of this model described below does shed some light on the behavior of this compound in the linear model.

The three descriptors used in the model were MDEN-23, RNHS-3, and SURR-5. The MDEN-23 descriptor is the molecular distance edge vector[20] between secondary and tertiary nitrogens. The descriptor is defined as the geometric mean of the topological path lengths between secondary and tertiary nitrogens. The original implementation of this descriptor only considered carbons and can be interpreted as characterizing the extension of side chains from the main body of a molecule.[40] The characteristic feature of the compounds in this study is that they all contain a piperazine and pyrimidine substructure. The two substructures are connected via the nitrogen on the piperazine group. As a result the MDEN-23 descriptor captures the linkage between the two rings. Furthermore, a number of compounds have side groups containing secondary and (or) tertiary nitrogens (examples include compounds **5**, **32**, and **54**). The MDEN-23 descriptor thus characterizes the 'nitrogen backbone' of these compounds. Since for the compounds in this study tertiary nitrogens are generally members of cycles and all compounds have central pyrimidine and piperazine rings, larger values of this descriptor indicate the presence of cyclic and noncyclic side chains containing nitrogen.

The RNHS-3 descriptor is a hydrophobic surface area (HSA) descriptor developed by Stanton et al.[26] It is defined as

$$\frac{\max(SA^{-})\sum H_i}{\log P}$$

where max (SA^{-}) is the surface area of all the most hydrophilic atom, H_i are the hydrophilic constants (which are values of Wildman and Crippens[41] atomic hydrophobicity constants less than 0), and log P is the logarithm of the octanol−water partition coefficient. Thus this descriptor is a measure of the relative hydrophilic surface area of a molecule. The presence of this descriptor in the model is not surprising considering that all compounds in the study contain three or more nitrogens along with oxygens in a number of cases.

The SURR-5 descriptor is a modification of the HSA descriptor described by Mattioni.[42] The original HSA descriptors classified atoms as either hydrophilic or hydrophobic using the atomic hydrophobicity constants of Wildman and Crippen.[41] In the modified version hydrophobic atoms are divided into *low hydrophobic* (atoms with hydrophobic constants between 0 and 0.4) and *high hydrophobic* (atoms with hydrophobic constants greater than 0.4). The modification increases the differentiability of the HSA descriptors and has been shown to be effective in structure−activity studies.[42] SURR-5 is defined as the ratio of the atomic constant weighted hydrophobic (low) surface area and the atomic constant weighted hydrophilic surface area. This descriptor thus characterizes the various portions of the molecular surface in terms of hydrophobicity and hydrophilicity. Absolute values greater than one indicate that the molecular surface is mainly hydrophobic, and values less than one indicate that the molecular surface is mainly hydrophilic.

To ensure that the results described above did not arise by chance, randomized runs were carried out. A randomized run consisted of scrambling the dependent variable and building the model using the same descriptors as in the original model. This procedure was repeated 500 times, and the average values of the R^2 and RMSE were calculated for both the training and prediction sets. It is expected that if a true structure−activity relationship is captured by the original model, the randomized models should exhibit lower values of R^2 and higher values of RMSE when compared to the original model. The results from our runs indicate this to be the case. The average value of R^2 and RMSE for the training set was 0.05 and 0.72, respectively. For the prediction set they were 0.08 and 1.04, respectively. The statistics of the randomized runs are summarized in Table 4. It should be noted that in all the runs compound **55** was not removed from the prediction set.

The 3-descriptor, linear model was then subjected to a PLS analysis to provide an interpretation of the structure−activity relationship embodied by the model. This technique has been described by Stanton,[43] and a number of examples of this technique have been reported.[40,43] The PLS analysis was carried out with Minitab[44] using a leave-one-out cross-validation scheme. The results of the PLS analysis indicated that all 3 components were validated, and thus the model was not overfit. A summary of the statistics for the 3 components are shown in Table 5. Table 6 shows the X-weights for the 3 PLS components. The X-weights for a given component indicate the contributions of each descriptor to that component. As can be seen, in each component one descriptor has a very high absolute value and thus is the main contributor to that component. We consider each component separately and use the weights and the score plots (Figures 4, 6, and 8) to interpret the structure−activity trend characterized by the model.

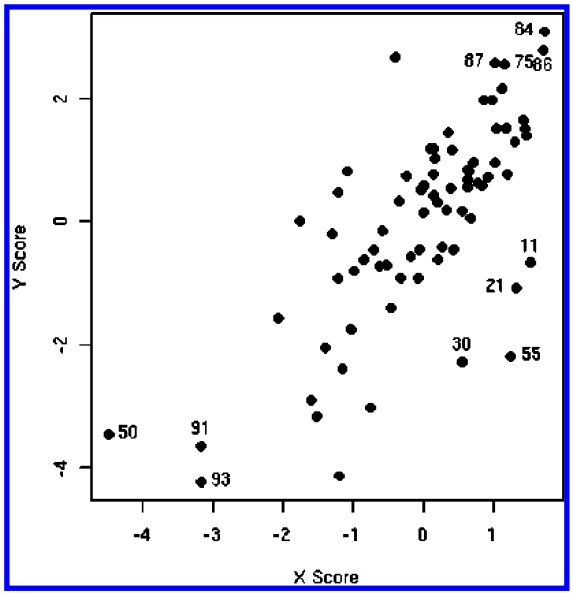

Figure 4. The score plot for PLS component 1.

Table 4. Average Statistics for the Training and Prediction Set Predictions Made by 500 Randomized Models

	R^2		RMSE	
	mean	std. deviation	mean	std. deviation
training set	0.05	0.04	0.72	0.03
prediction set	0.08	0.11	1.04	0.12

Table 5. Summary of the Statistics from the PLS Analysis of the Best 3-Descriptor Linear Model

component	X variance	error SS	R^2	PRESS	Q^2
1	0.51	14.80	0.52	16.67	0.45
2	0.78	12.11	0.60	13.43	0.56
3	1.00	12.07	0.61	13.27	0.56

Table 6. Weights for the 3 Validated Components from the PLS Analysis of the 3-Descriptor Linear Model

descriptor	component 1	component 2	component 3
MDEN-23	−0.16	0.93	0.30
RNHS-3	0.55	−0.17	0.81
SURR-5	−0.82	−0.29	0.48

The most heavily weighted descriptor in the first component is SURR-5. As can be seen its weight is significantly higher than the other two descriptors and thus plays an important role. Figure 4 shows the score plot for the first PLS component. Points in the upper right and lower left are correctly predicted as active and inactive compounds, respectively. The structures of some representative active and inactive compounds for this component are compared in Figure 5. Compounds **75**, **84**, **86,** and **87** are regarded as active, and they are characterized by high absolute values of the SURR-5 descriptor. From the description of the

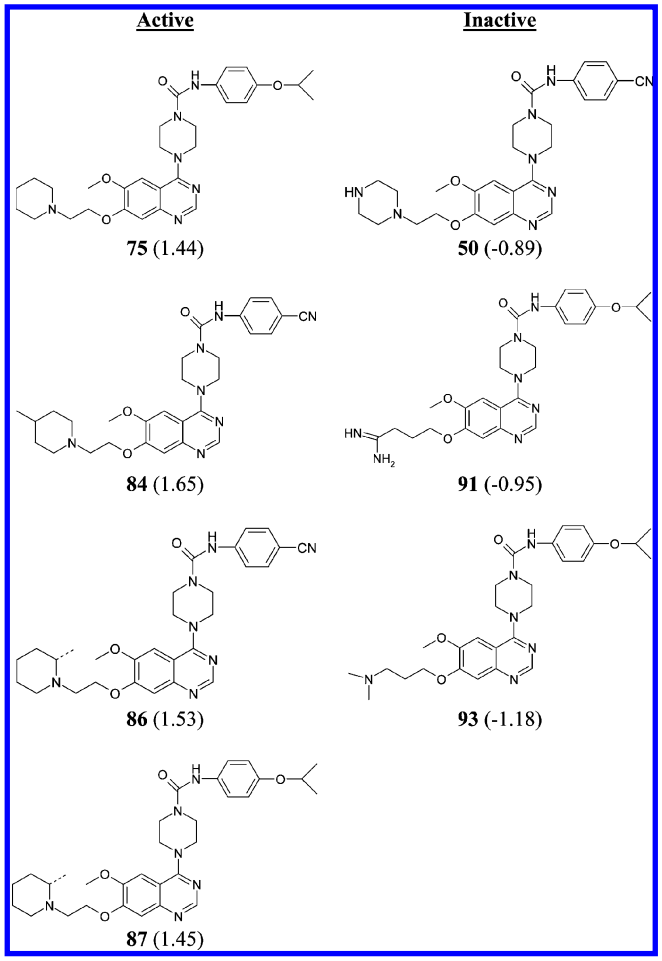

Figure 5. A comparison of the structures of the active and inactive compounds predicted by component 1 from the 3-component PLS model. Activity values in $-\log(IC_{50})$ units are provided within brackets.

SURR-5 descriptor this indicates that active molecules are characterized by a large hydrophobic surface area. This is consistent with the fact that the cell based assay used by Pandey et al.[4] reports the activity of the compounds against the kinase target modulated by their ability to pass through the cell membrane. Clearly, compounds with a higher proportion of hydrophobic surface area would have a better ability to enter the cell. Component 1 does not underpredict any compounds as shown by the empty upper left corner. However, compounds **11**, **21**, **30,** and **55** are overpredicted by this component. An interesting point to note is that compound **55** which was a significant outlier in the linear model (and is also an outlier in the nonlinear CNN model) has a high absolute value of the SURR-5 descriptor but has a low observed activity (-0.39 $-\log(IC_{50})$ units). As a result this compound does not follow the general structure−activity trend for the SURR-5 descriptor. As will be shown in the results for the random forest, the SURR-5 descriptor is a very significant descriptor. Since **55** does not follow the trend for this descriptor, this explains to some extent its position as an outlier. Compounds **50**, **91,** and **93** are predicted correctly as inactive and are characterized by low absolute values of the SURR-5 descriptor. Considering the structures shown in Figure 5 it is clear that the piperaznylquinazoline backbone is common to both active and inactive structures. However the active structures shown (as well as in nearly

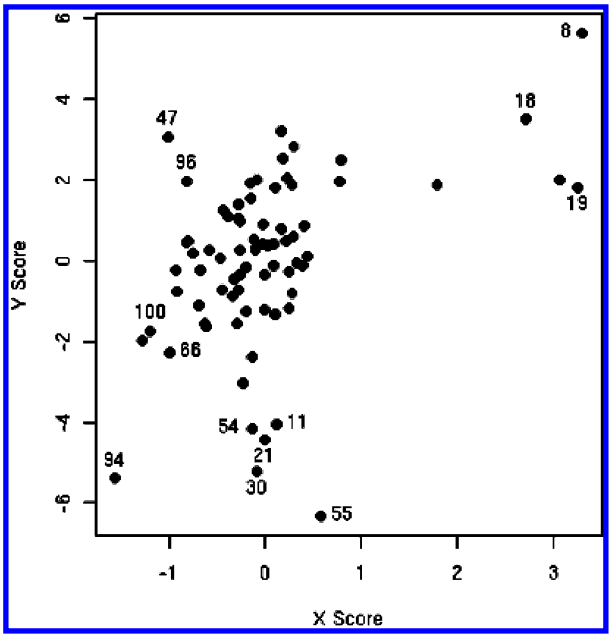

Figure 6. The score plot for PLS component 2.

all the active compounds for this component) all have a bulky hydrophobic group linked to the 7 position on the quinazoline ring. However compound **50** has a piperazine ring linked to the 7 position but exhibits a low activity. This may be understood by considering molecular surfaces. Figures 9−11 show molecular surfaces for compounds **75**, **50**, and **93** colored by hydrophobicity values drawn using PyMOL.[45] Blue regions indicate areas of high hydrophilicity, and red regions indicate areas of high hydrophobicity. The bulky piperidine group in **75** is largely hydrophobic compared to the trimethylamine group in **93** which has a distinct hydrophilic center. In light of these observations, the surface of **50** shows that the amide center on the piperazine ring creates a large hydrophilic center and thus is similar in this respect to **93**. One would thus expect activity would be improved by having bulky groups without hydrophilic centers connected to the 6 or 7 position on the quinazoline ring.

The most heavily weighted descriptor in PLS component 2 is MDEN-23. Figure 6 shows the score plot for the second PLS component. Compounds predicted correctly as active (**8**, **18,** and **19**) exhibit very high values of this descriptor, whereas compounds predicted correctly as inactive (**54, 66, 94,** and **100**) exhibit smaller values. Large values of this descriptor are characterized by a larger number of longer paths between secondary and tertiary nitrogens. This may be indirectly interpreted as a count of nitrogens. Pandey et al.[4] mention that in several cases removing basic groups (such as secondary amines in this case) greatly reduces potency. Thus, larger numbers of secondary nitrogens would enhance the activity of potential inhibitors. Another aspect of this descriptor that has been described previously is that it may be interpreted, in the case of the current data set, as an indicator of nitrogen containing rings separated by long paths. This would imply that molecules with large cyclic side chains connected to the backbone via long chains would exhibit higher values of this descriptor. The structures of some of the active and inactive compounds are shown in Figure 7. It is evident that the active compounds have bulky nitrogen containing side groups on the phenoxy ring. In the case of

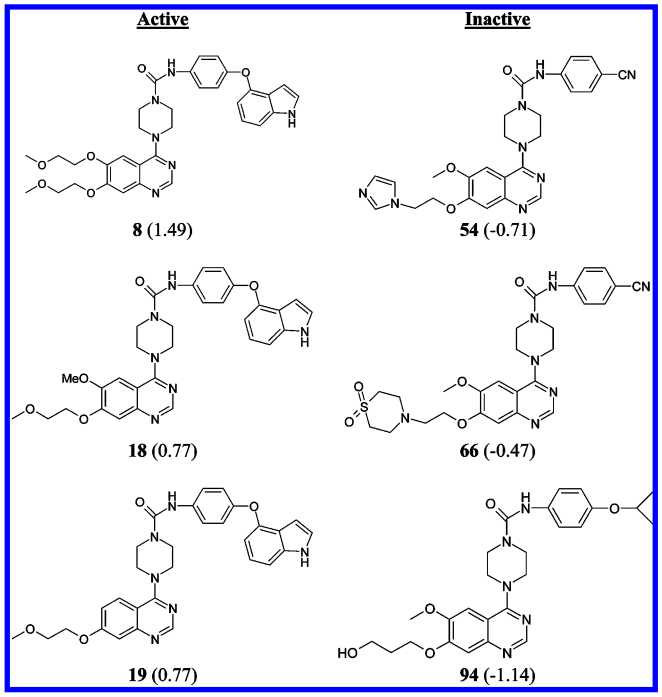

Figure 7. A comparison of the structures of the active and inactive compounds predicted by component 2 from the 3 component PLS model. Activity values in $-\log(IC_{50})$ units are provided within brackets.

the molecules shown here it is an indole group. In the case of the inactive molecules these are absent. This confirms the observations made by Matsuno[46] and Pandey[4] that bulky hydrophobic side groups along with electron donating centers enhance activity. However **54** does appear to be anomalous in that it does contain a relatively hydrophobic side group (attached to the quinazoline ring) yet is inactive.

Once again the importance of the SURR-5 descriptor is evident as the second component underpredicts a large number of active compounds which were correctly predicted by component 1. However, component 2 corrects for the overprediction of some of the compounds from component 1. As can be seen from the score plot in Figure 6, compounds **11**, **21**, **30**, and **55** are now shifted toward the lower left. Thus, this component compensates for the overprediction of these compounds by component 1 by taking into account bulky hydrophobic groups attached to the phenyl ring. It should be noted that though **55** is predicted relatively better in this component than the previous one, it is still midway between the two lower quadrants. However, it does follow the trend for the MDEN-23 descriptor (i.e., lower values indicate lower activities) better than for the SURR-5 descriptor.

Finally, we consider PLS component 3. Table 5 shows that the increase in R^2 gained by adding component 3 to the model is only 0.01. Thus, it is expected that this component will not be able to explain any significant structure−activity trend described by the most heavily weighted descriptor (RNHS-3). As can be seen from the score plot (Figure 8), this component does not predict any low activity compounds. Furthermore the underpredicted compounds (**93** and **91**) have already been correctly predicted as inactive by component 1, and the overpredicted compounds in the lower right corner were also correctly predicted as moderately inactive by components 1 and 2. However this component does contribute to the structure−activity relationship to some extent by correctly predicting molecules **47** and **96** as active, whereas they were underpredicted by component 2.

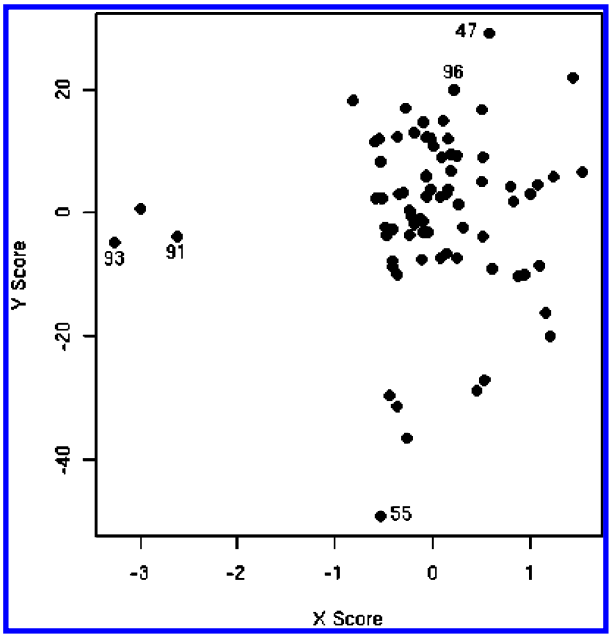

Figure 8. The score plot for PLS component 3.

Combining the two main trends discussed in this section, we see that there is a competition between a requirement for bulky hydrophobic side groups and higher numbers of nitrogens (which create hydrophilic centers). The fact that component 1 explains the majority of the structure−activity trend implies that the latter requirement plays a stronger role. Thus it may be expected that compounds with a piperazinylquinazoline backbone would exhibit increased activity by having bulky hydrophobic nitrogen containing groups attached to the phenyl moiety as well as at the quinazoline moiety. Furthermore, bulk may be increased at the quinazoline moiety by attaching side groups at both the 6 and 7 positions. This would imply that the groups would have to be bonded by relatively long paths to the 6 and 7 positions to avoid steric hindrance. Assuming that the linker groups contain nitrogen, this would result in larger values of the MDEN-23 descriptor for those molecules. And as has been shown, large values of this descriptor correlate with higher activities.

As noted before, this data set had been studied by Khadikar[11] who developed a set of linear regression models. However their methodology differed significantly in that they used the molecules with reported activities in the absence of human plasma. As a result this restricted the size of the data set. Furthermore the linear models were developed after removing 10 molecules from the already reduced data set. Finally, the models were developed using a stepwise linear regression technique which is not necessarily an efficient way to search for optimal descriptor subsets. The best linear model reported in this work exhibits a lower value of R^2 than the corresponding 3-descriptor model reported by Khadikar. However considering the fact that this statistic is well-known to be misleading and the fact that we used a larger data set, we believe that the lower value of R^2 for our model does not detract from its main utility, i.e., as an interpretive model. Furthermore, the descriptors present in

Table 7. Statistics for the Best Nonlinear CNN Model

	number of molecules	RMSE	R^2
training set	57	0.22	0.94
cross-validation set	9	0.21	0.90
prediction set	13	0.32	0.61

our best linear model allow a clear interpretation of the structure−activity trend which confirms observations made by Pandey et al.[4] The topological descriptors present in the model described by Khadikar do not lend themselves well to a detailed interpretation.

Nonlinear CNN Models. Nonlinear CNN models were developed by using the CNN routine as the objective function for the genetic algorithm. The full training set of 57 compounds was used. For a given CNN architecture the descriptor space is thus searched for subsets that lead to CNN models with low values of the cost function given by Equation 1. Once a number of suitable subsets are found, the number of hidden layer neurons is varied to determine the optimal CNN architecture. This procedure resulted in a 7−3−1 CNN model. The statistics of the model are given in Table 7. A comparison of the statistics in Tables 7 and 2 clearly indicate the improved performance of the nonlinear CNN model compared to the linear model. The seven descriptors present in the model are N5CH,[21,23,24] WTPT-3,[47] WTPT-4,[47] FLEX-4, RNHS-3,[26] SURR-5,[26] and APAVG. It should be noted that two of the descriptors (RNHS-3 and SURR-5) are also present in the best linear model. N5CH is the number of 5th order chains which are defined as a sequence of 5 atoms containing a ring. This definition thus includes five-membered rings, four-membered rings with a methyl side chain, and a three-membered ring with an ethyl side chain. The WTPT descriptors are based on Randic's molecular ID and are termed weighted path descriptors. They combine features of connectivity indices[21,23,24] and path counts and are independent of molecular geometry. WTPT-3 considers all weighted paths starting from any heteroatom, and WTPT-4 considers weighted paths starting only from oxygen atoms. The FLEX-4 descriptor characterizes conformational flexibility. More specifically this descriptor evaluates the fractional mass of the rotatable atoms. RNHS-3 and SURR-5 have been described previously. Finally, the AP-AVG descriptor is based on atom pairs as defined by Carhart et al.[48] The atom pair method take describes molecular features by considering pairs of atoms together with the path between them. As a result a given molecule will have a set of atom pair strings which contain the start and end atom types and the path length between them. These atom pair strings can be hashed to give a 32 bit number which have been used as a similarity measure. APAVG is defined as the average of the atom pair hash values.

Figure 12 shows a plot of the predicted versus observed $-\log(IC_{50})$ values from the CNN model. It is encouraging to see that the performance of the nonlinear model was very good on the training set as shown the RMSE and R^2 values. The plot is also substantially less scattered than the corresponding plot for the linear model. As noted on the plot, there are two possible prediction set outliers. When compound **55** was removed from the prediction set, and the remaining compounds were processed by the model, the R^2 value for the prediction set rose to 0.72 and the RMSE decreased to 0.27. Due to the lack of interpretability of CNN models any explanation as to why this compound is an outlier is not feasible.

Figure 9. Molecular surface plot of **75**, colored by hydrophobicity values (blue is most hydrophilic and red is most hydrophobic).

Figure 10. Molecular surface plot of **93**, colored by hydrophobicity values (blue is most hydrophilic and red is most hydrophobic).

Figure 11. Molecular surface plot of **50**, colored by hydrophobicity values (blue is most hydrophilic and red is most hydrophobic).

As in the case of the linear model, the nonlinear CNN model was also tested for random correlations. As before, the dependent variable was scrambled, and the CNN model was rebuilt. The procedure was repeated 100 times, and the averages of the RMSE and R^2 values are reported in Table 8. As can be seen the average RMSE is more than triple that of the original runs. The average values of R^2 are also very poor. These results indicate that chance played very little role in the performance of the CNN model.

Random Forest Model. The linear and nonlinear models presented so far have two descriptors in common, RNHS-3 and SURR-5. We also note that using the genetic algorithm resulted in a large number of linear and nonlinear models which contained these descriptors. SURR-5 was present in

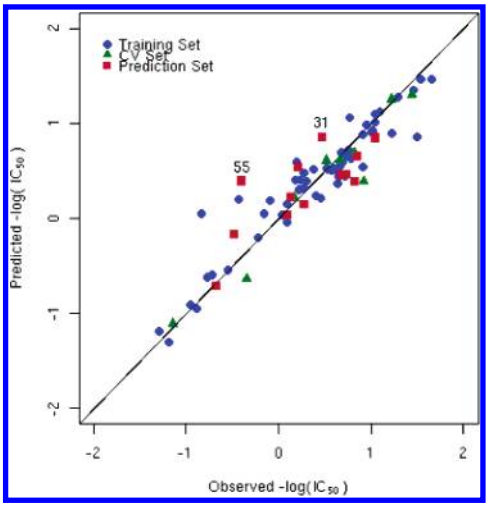

Figure 12. A plot of the observed versus predicted $-\log(IC_{50})$ values for the best nonlinear CNN model. The annotated points are possible prediction set outliers.

Table 8. Summary of the Statistics for the Training, Cross-Validation, and Prediction Sets from Randomized Runs Using the Best CNN Model[a]

	R^2		RMSE	
	mean	std. deviation	mean	std. deviation
training set	0.10	0.19	0.71	0.11
cross-validation set	0.10	0.23	0.96	0.14
prediction set	0.01	0.10	1.11	0.14

[a] The architecture used was 7−3−1.

more than 90% of the models evaluated. Clearly, this descriptor must be information rich. The role played by this descriptor in the linear model has been analyzed using PLS. The role of the descriptors in the nonlinear CNN model are not amenable to analysis due to the black box nature of the CNN methodology.

We built a random forest model to investigate whether it would provide any further information regarding the importance of descriptors, specifically SURR-5. As mentioned previously random forest parameters were tuned using a grid search, and the final forest was built with 500 trees and a node size of 5 and 13 descriptors were used at each split point. The model was built using all the molecules in the data set and the reduced pool of 41 descriptors. The predictive ability of this model was not significantly better than the linear regression or nonlinear CNN models. However our main focus was on the importance ascribed to specific descriptors by the random forest model. Figure 13 shows a plot of descriptor importance (only the 10 most important descriptors are shown, ranked in decreasing order of importance).

The method by which a random forest measures variable importance has been discussed in the literature.[37,39] Essentially, for each tree in the forest out-of-bag (OOB) predictions are made, and for each OOB set the descriptors are individually scrambled and predictions are made. After the model has been built, the mean squared errors (MSE) are calculated for the initial OOB prediction and the OOB predictions for the set with the scrambled descriptor. The difference between these two values for each descriptor provides an estimate of the importance for that descriptor. It is clear that SURR-5 is deemed to be the most important

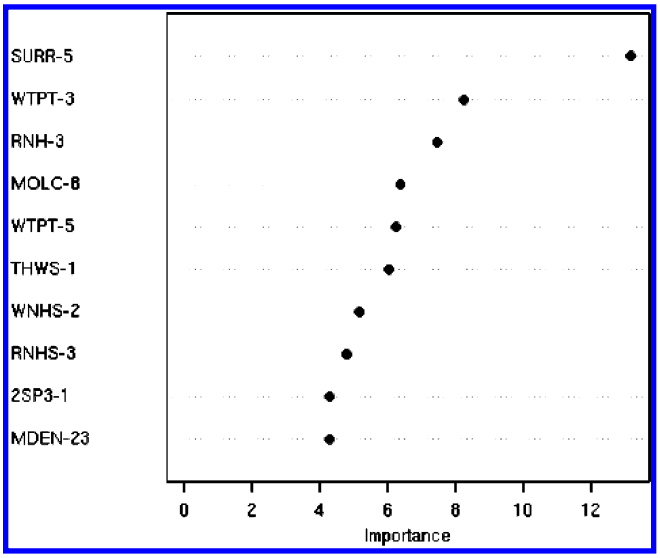

Figure 13. A variable importance plot generated from the random forest model built using the reduced descriptor pool with no molecules excluded from the training or prediction set.* (*SURR-5 − the ratio of atomic constant weighted hydrophobic (low) surface area to the atomic constant weighted hydrophilic surface area;[26,42] WTPT-3 − sum of path lengths starting from heteroatoms;[47] RNH-3 − sum of hydrophilic constants divided by the value of log P;[42] MOLC-8 − path-cluster of length 4 molecular connectivity index;[22,53] WTPT-5 − sum of path lengths starting from nitrogen;[47] THWS-1 − total hydrophobic weighted surface area[42] defined as the sum of the product of atomic log *P* values and hydrophobic atom surface areas; WNHS-2 − surface weighted hydrophilic surface area[42] defined as the product of the hydrophilic surface area multiplied by the total molecular surface area divided by 1000; RNHS-3 − relative hydrophilic surface area[26] defined as the product of the sum of the hydrophilic constants and surface area of the most hydrophilic atom divided by overall log P; 2SP3−1 − the number of sp^3 carbons bound to two other carbons; MDEN-23 − molecular distance edge vector between secondary and tertiary nitrogens[20]).

descriptor. Interestingly, RNHS-3 and MDEN-23 are relatively low ranked. Furthermore, the PLS analysis indicated that for the linear regression model, MDEN-23 was able to account for more of the structure−activity trend compared to RNHS-3. From Figure 12 it is clear that the increase in MSE is not very large in going from MDEN-23 to RNHS-3. At the same time it should be noted that the algorithms underlying PLS and random forests are substantially different. Most importantly, the random forest is working with the whole reduced pool (41 descriptors), and thus it is able to compare and contrast more descriptors than considered in the PLS analysis. Thus a relationship detected by a PLS analysis will not necessarily show up in a random forest. However it is encouraging that the most important descriptor from the random forest model describes the majority of the structure−activity trend in the PLS analysis. We also note that the CNN model contains the two most important descriptors as identified by the random forest. Furthermore the remaining descriptors in the CNN model are present in the top 20 descriptors as measured by the random forest. This is not surprising as the CNN model is built by allowing the GA to search for the best 7-descriptor subset from the whole reduced pool. Once again, a direct correspondence between descriptors is not expected due to the different algorithms underlying the respective models.

The above discussion indicates the relative importance of the SURR-5 descriptor in both linear and nonlinear models. Since SURR-5 describes the hydrophobicity of a surface we investigated its relation to the logP values of the molecules. The logP values were calculated using a fragment based approach developed by Mattioni for the HSA descriptors mentioned earlier. A scatter plot of logP versus SURR-5 for the data set showed no distinct correlations ($R^2 = 0.17$). We also made scatter plots of logP versus the other descriptors, and none of them showed any correlations (R^2 ranging from 0.01 to 0.20) except in the case of RNHS-3. However this is to be expected as the functional form of this descriptor includes the logP value of the molecule.

We also investigated whether the most important descriptors from the random forest model would lead to good linear or CNN models. We evaluated a regression model and carried out a PLS analysis using the top three descriptors, but the RMSE and R^2 were poorer than those reported for the best linear model. Even though the PLS analysis validated all three descriptors, the total R^2 explained was less than for the best model. The descriptors were also used in CNN models. Three architectures were investigated, 3−2−1, 3−3−1, and 3−4−1. However none of the models performed significantly better than the reported model.

CONCLUSION

The results presented in this work indicate that the regression and CNN models developed exhibit interpretability as well as predictive ability. Though the linear model was developed mainly for purposes of structure−activity interpretation, removal of one prediction set outlier improved its predictive ability drastically. The application of a PLS analysis allows for the interpretation of the structure−activity trends embodied in the model. The interpretation clearly indicates the importance of the hydrophobic surface area descriptor, SURR-5. This is also confirmed by the random forest model which provides a measure of descriptor importance. The model ranked SURR-5 as the most important descriptor. However the other descriptors in the linear are also relatively important with respect to the whole descriptor pool. The main conclusions from the PLS interpretation indicate bulky hydrophobic groups and nitrogen centers increase activity. These observations have been made experimentally, thus confirming our theoretical model. As noted before these two trends compete against each other. However, the PLS and random forest results also indicate the relatively more important role of hydrophobic groups. The CNN model was developed primarily for predictive ability as such models are not amenable to interpretation.[40] It exhibits good statistics for both training and prediction. Furthermore it also contains the top two descriptors as identified by the random forest, including SURR-5, once again underlying its importance to the structure−activity relationship. An interesting extension to this work would be to develop a 3D QSAR model using CoMFA[49,50] which would allow a more detailed view of the specific interactions that are described by our 2D models. The predictions described in the preceding sections are based on the correlation of molecular descriptors to experimental activity and thus may be considered relatively abstract. That is, the 2D methodology we employ cannot provide a direct view of the binding between these molecules and the PDGF receptor and hence inhibitory activity. This implies that any conclusions made on the basis of our models are oriented toward the activity value rather than activity mechanism (via binding features). A 3D method such as CoMFA would allow for a more direct understanding of the interactions of the molecules considered here with the PDGF receptor. In addition, a CoMFA model would allow for the prediction of binding energies. Combined with a systematic modification of the side groups at the 6 and 7 positions in silico, this would allow not only confirmation of the experimental data described here but could also be used as a stepping stone to the synthesis of more potent inhibitors. The fundamental requirement for such a study would be the crystal structure of PDGFR. The crystal structures of tyrosine kinase receptors related to the PDGF receptor have been reported[51,52] though we are not aware of crystal structures of the PDGF receptor specifically. Using 3D structures based on homology modeling would possibly allow the initial development of a binding model for this receptor and the molecules described here. However, as our group concentrates on the 2D aspects of QSAR modeling we did not have the tools or expertise to carry out a 3D QSAR investigation.

In summary this work resulted in the development of 2D QSAR models which are able to provide a detailed interpretation of the structure−activity relationship for the PDGFR inhibitors studied as well as a predictive model which could conceivably be used as a screening tool for analogous compounds.

REFERENCES AND NOTES

(1) Kurup, A.; Garg, R.; Hansch, C. Comparative QSAR Study of Tyrosine Kinase Inhibitors. *Chem. Rev.* **2001**, *101*, 2573−2600.

(2) Schlessinger, J.; Ullrich, A. Growth Factor Signaling By Receptor Tyrosine Kinases. *Neuron* **1992**, *9*, 383.

(3) Iida, H.; Seifert, R.; Alpers, C. E.; Gronwald, R. G.; Philips, P. E.; Pritzl, P.; et al. Platelet derived growth factor (PDGF) and PDGF Receptor (PDGFR) Are Induced In Mesangial Proliferative Nephritis In The Rat. *Proc. Natl. Acad. Sci.* **1995**, *88*, 6560−6564.

(4) Pandey, A.; Volkots, D. L.; Seroogy, J. M.; Rose, J. W.; Yu, J.-C.; Lambing, J. L.; et al. Identification of Orally Active, Potent, and Selective 4-Piperazinylquinazolines as Antagonists of the Platelet-Derived Growth Factor Receptor Tyrosine Kinase Family. *J. Med. Chem.* **2002**, *45*, 3772−3793.

(5) Palmer, B. D.; Kraker, A. J.; Hartl, B. G.; Panopoulos, A. D.; Panek, R. L.; Batley, B. L.; et al. Structure−Activity Relationships for 5-Substituted 1-Phenylbenzimidazoles as Selective Inhibitors of the Platelet-Derived Growth Factor Receptor. *J. Med. Chem.* **1999**, *42*, 2373−2382.

(6) Kubo, K.; Shimizu, T.; Ohyama, S.; Murooka, H.; Nishitoba, T.; Kato, S.; et al. A novel series of 4-phenoxyquinoxazolines: potent and highly selective inhibitors of PDGF receptor autophophorylation. *Bioorg. Med. Chem. Lett.* **1997**, *7*, 2935−2940.

(7) Boschelli, D. H.; Wu, Z.; Klutchko, S. R.; Showalter, H. D. H.; Hamby, J. M.; Lu, G. H.; et al. Synthesis and Tyrosine Kinase Inhibitory Activity of a Series of 2-Amino-8H-pyrido[2,3-d]pyrimidines: Identification of Potent, Selective Platelet-Derived Growth Factor Receptor Tyrosine Kinase Inhibitors. *J. Med. Chem.* **1998**, *41*, 4365−4377.

(8) Klutchko, S. R.; Hamby, J. M.; Boschelli, D. H.; Wu, Z.; Kraker, A. J.; Amar, A. M.; et al. 2-Substituted Aminopyrido[2,3-d]pyrimidin-7(8H)-ones. Structure−Activity Relationships Against Selected Tyrosine Kinases and in Vitro and in Vivo Anticancer Activity. *J. Med. Chem.* **1998**, *41*, 3276−3292.

(9) Kraker, A. J.; Hartl, B. G.; Amar, A. M.; Barvian, M. R.; Showalter, H. D. H.; Moore, C. W. Biochemical and cellular effects of c-Src kinase-selective pyrido[2,3-d]pyrimidine tyrosine kinase inhibitors. *Biochem. Pharmacol.* **2000**, *60*, 885−898.

(10) Shen, Q.; Lu, Q.-Z.; Jiang, J.-H.; Shen, G.-L.; Yu, R.-Q. Quantitative structure−activity relationships (QSAR): studies of inhibitors of tyrosine kinase. *Eur. J. Pharm. Sci.* **2003**, *20*, 63−71.

(11) Khadikar, P. V.; Shrivastava, A.; Agrawal, V. K.; Srivastava, S. Topological Designing of 4-Perazinylquinazolines as Anatagonists of PDGFR Tyrosine Kinase Family. *Bioorg. Med. Chem. Lett.* **2003**, *13*, 3009−3014.
(12) Lokker, N. A.; O'Hare, J. P.; Barsoumian, A.; Tomlinson, J. E.; Ramakrishnan, V.; Fretto, L. J.; Giese, N. A. Functional Importance of the Platelet Derived Growth Factor Receptor Extra-Cellular Immunoglobin Like Domains: Identification of PDGF Binding Site and Neutralizing Monoclonal Antibodies. *J. Biol. Chem.* **1997**, *272*, 33037−33044.
(13) Jurs, P. C.; Chou, J. T.; Yuan, M. Studies of Chemical Structure Biological Activity Relations Using Patter Recognition. In *Computer Assisted Drug Design*; Olsen, E. C., Christoffersen, R. E., Eds.; American Chemical Society: Washington, DC, 1979.
(14) Stuper, A. J.; Brugger, W. E.; Jurs, P. C. *Computer Assisted Studies of Chemical Structure and Biological Function*; Wiley: New York, 1979.
(15) Breiman, L. Random Forests. *Machine Learning* **2001**, *45*, 5−32.
(16) *R: A language and environment for statistical computing*, v. 1.8.1; R Development Core Team; R Foundation for Statistical Computing: Vienna, Austria.
(17) *Hyperchem*, v. 6.01; Hypercube Inc.: Gainesville, FL.
(18) Goldstein, H. *Classical Mechanics*; Addison-Wesley: Reading, MA, 1950.
(19) Pearlman, R. S. Molecular Surface Areas and Volumes and Their Use in Structure/Activity Relationships. In *Physical Chemical Properties of Drugs*; Yalkowsky, S. H., Sinkula, A. A., Valvani, S. C., Eds.; Marcel Dekker: New York, 1980.
(20) Liu, S.; Cao, C.; Li, Z. Approach to estimation and prediction for normal boiling point (nbp) of alkanes based on a novel molecular distance edge (mde) vector, lambda. *J. Chem. Inf. Comput Sci.* **1998**, *38*, 387−394.
(21) Kier, L. B.; Hall, L. H. Molecular connectivity VII: Specific treatment to heteroatoms. *J. Pharm. Sci.* **1976**, *65*, 1806−1809.
(22) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Chemistry and Drug Research*; Academic Press: New York, 1976.
(23) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Structure Activity Analysis*; John Wiley & Sons: 1986.
(24) Kier, L. B.; Hall, L. H.; Murray, W. J. Molecular connectivity I: Relationship to local anesthesia. *J. Pharm. Sci.* **1975**, *64*,
(25) Stanton, D. T.; Jurs, P. C. Development and Use of Charged Partial Surface Area Structural Descriptors in Computer Assissted Quantitative Structure Property Relationship Studies. *Anal. Chem.* **1990**, *62*, 2323−2329.
(26) Stanton, D. T.; Mattioni, B. E.; Knittel, J. J.; Jurs, P. C. Development and Use of Hydrophobic Surface Area (HSA) Descriptors for Computer Assisted Quantitative Structure−Activity and Structure−Property Relationships. *J. Chem. Inf. Comput. Sci.* **2004**, ASAP article.
(27) Sutter, J. M.; Dixon, S. L.; Jurs, P. C. Automated Descriptor Selection For Quantitative Structure−Activity Relationships Using Generalized Simulated Annealing. *J. Chem. Inf. Comput Sci.* **1995**, *35*, 77−84.
(28) Goldberg, D. E. *Genetic Algorithms in Search Optimization & Machine Learning*; Addison-Wesley: Reading, MA, 2000.
(29) Wessel, M. D. Computer Assisted Development of Quantitative Structure − Property Relationships and Design of Feature Selection Routines. Ph.D., Chemistry, Pennsylvania State University, University Park, 1997.
(30) Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M.; Teller, A. H.; Teller, E. Equation of State Calculations by Fast Computing Machines. *J. Chem. Phys.* **1953**, *21*, 1087−1092.
(31) Sutter, J. M. Computer Aided Development of Quantitative Structure Activity Relationships and Analysis of Data from an Artificial Nose. Ph.D., Chemistry, Pennsylvania State University, University Park, 1997.
(32) Lu, X.; Ball, J. W.; Dixon, S. L.; Jurs, P. C. Quantitative Structure−Activity Relationships for Toxicity of Phenols Using Regression Analysis and Computational Neural Networks. *Environ. Toxicol. Chem.* **1994**, *13*, 841−851.
(33) Wessel, M. D.; Jurs, P. C. Prediction of Reduced Ion Mobility Constants from Structural Information Using Multiple Linear Regression Analysis and Computational Neural Networks. *Anal. Chem.* **1994**, *66*, 2480−2487.
(34) Schapire, R. E. In *A brief introduction to boosting*, Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence, 1999; A brief introduction to boosting, boosting.
(35) Freund, Y.; Schapire, R. E. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. System Sci.* **1997**, *55*, 119−139.
(36) Breiman, L. Bagging Predictors. *Machine Learning* **1996**, *24*, 123−140.
(37) Breiman, L.; Friedman, J. H.; Olshen, R. A.; Stone, C. J. *Classification and Regression Trees*; Wadsworth: 1984.
(38) Hawkins, D. M.; Young, S. S.; Rusinko, A. I. Analysis of Large Structure−Activity Data Set Using Recursive Partitioning. *Quant. Struct.-Act. Relat.* **1997**, *16*, 296−302.
(39) Svetnik, V.; Liaw, A.; Tong, C.; Culberson, J. C.; Sheridan, R.; Feuston, B. P. Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1947−1958.
(40) Guha, R.; Jurs, P. C. The Development of QSAR Models To Predict and Interpret the Biological Activity of Artemisinin Analogues. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1440−1449.
(41) Wildman, S. A.; Crippen, G. M. Prediction of Physicochemical Parameters by Atomic Contributions. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 868−873.
(42) Mattioni, B. E. The Development of Quantitative Structure−Activity Relationship Models for Physical Property and Biological Activity Prediction of Organic Compounds. Ph.D., Chemistry, Pennsylvania State University, University Park, 2003.
(43) Stanton, D. T. On The Physical Interpretation of QSAR Models. *J. Chem. Inf. Comput Sci.* **2003**, *43*, 1423−1433.
(44) *Minitab*, v. 14; Minitab Minitab Inc.: State College, PA.
(45) *The PyMOL Molecular Graphics System*, v. 0.95; DeLano, W. L. DeLano Scientific: San Carlos, CA.
(46) Matsuno, K.; Ichimura, M.; Nakajima, T.; Tahara, K.; Fujiwara, S.; Kase, H.; Giese, N. A.; Pandey, A.; Scarborough, R. M.; Yu, J.-C.; Lokker, N. A.; Irie, J.; Tsukuda, E.; Oda, S.; Nomoto, Y. Potent and Selective Inhibitors of PDGFR Phosphorylation. I. Synthesis and structure−activity relationship of a new class of quinazoline derivatives. *J. Med. Chem.* **2002**, *45*, 3057−3066.
(47) Randic, M. On molecular identification numbers. *J. Chem. Inf. Comput. Sci.* **1984**, *24*, 164−175.
(48) Carhart, R. E.; Smith, D. H.; Venkataraghavan, R. Atom Pairs as Molecular Features in Structure−Activity Studies: Definition and Applications. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 64−73.
(49) Cramer, R. D., III; Patterson, D. E.; Bunce, J. D. Comparative molecular field analysis (comfa). i. effect of shape on binding of steroids to carrier protiens. *J. Am. Chem. Soc.* **1988**, *110*, 5959−5967.
(50) Cramer III, R. D.; Patterson, D. E.; Bunce, J. D.; Frank, I. E. Crossvalidation, bootstrapping and partial least squares compared with multiple regression in conventional qsar studies. *Quant. Struct-Act. Relat. Pharmacol., Chem. Biol.* **1988**, *7*, 18−25.
(51) McTigue, M. A.; Wickersham, J. A.; Pinko, C.; Showalter, R. E.; Parast, C. V.; Tempczyk-Russell, A.; Gehring, M. R.; Mroczkowski, B.; Kan, C. C.; Villafranca, J. E.; Appelt, K. Crystal structure of the kinase domain of human vascular endothelial growth factor receptor 2: a key enzyme in angiogenesis. *Structure* **1999**, *7*, 319−330.
(52) Mohammadi, M.; Froum, S.; Hamby, J. M.; Schroeder, M. C.; Panek, R. L.; Lu, G. H.; Eliseenkova, A. V.; Green, D.; Schlessinger, J.; Hubbard, S. R. Crystal structure of an angiogenesis inhibitor bound to the FGF receptor tyrosine kinase domain. *EMBO J.* **1998**, *17*, 5896−5904.
(53) Balaban, A. T. Higly discriminating distance based topological index. *Chem. Phys. Lett.* **1982**, *89*, 399−404.

CI049849F