Chapter 7

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

7.1 Introduction

The investigation of anti-cancer drugs has focussed on a number of targets. One of the initial focus areas was compounds that could interfere in DNA synthesis and function and as a result, stimulate 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 a family known as protein tyrosine kinases which transmit growth signals via a phosphorylation mechansim.[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 chapter.

A large number of compounds have been investigated as putative PDGFR inhibitors. Examples include 1–phenylbenzimidazoles,[5] arylquinoxalines,[6] piperazinylquinazolines[3] and various pyrimidine analogs.[7–9] The mode of action of PDGFR inhibitors is competition with ATP binding at the intra-cellular 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

This work was published as Guha, R.; Jurs, P.C., "The Development of Linear, Ensemble and Non-linear Models for the Prediction and Interpretation of the Biological Activity of a Set of PDGFR Inhibitors", *J. Chem. Inf. Comput. Sci.*, **2004**, *44*, 2179–2189.

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 et al.[3] The dataset consisted of 79 compounds 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 compounds experimentally, but no computational models were developed. We note that Khadikar et al[11] have reported a QSAR study using this dataset. 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 any single class. Finally, in addition to prediction, the linear model was analysed using the PLS interpretation method to explain the structure-activity trends embodied in it.

7.2 Dataset

The dataset consisted of 79 compounds that were derivatives of 4–piperazinylquinazolines and were investigated for their ability to inhibit PDGFR phosphorylation.[3] The structures of these compounds have been presented by Pandey et al.[3] The compounds were evaluated for their inhibition of PDGFR phosphorylation in MG63 cells.[3,12] 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 in the absence of plasma were reported as < 0.004. Since this indicates that the response was possibly below the limit of detection, these compounds would have to be ignored for the purposes of model building, thus decreasing the size of the dataset. Hence we only considered the

set of measurements made in the presence of human plasma thus allowing the use of all 79 compounds.

7.3 Methodology

This study used the ADAPT[13,14] package to calculate descriptors and develop QSAR models. As described previously, the ADAPT methodology allows for the development of linear models and nonlinear models. In addition a random forest[15] model was built using the R software package.[16] A brief overview of the aspects of the methodology specific to this study are dicussed below and further details are presented in Chapter 3.

The first step was to divide the compounds into the three sets - the training, cross-validation and prediction set (known as QSAR sets) using an activity binning method. This resulted in a training set containing 57 compounds, a cross-validation set containing 9 compounds and a prediction set containing 13 compounds. Next, the 3-D 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 resulting in a total of 321 descriptors. This descriptor pool was reduced using objective feature selection. Setting correlation and identical cutoffs of 0.75 and 0.75, respectively, this procedure resulted in a reduced pool of 41 descriptors. This descriptor pool was used to generate the linear and nonlinear models discussed in this chapter.

The next stage was subjective feature selection in which a simulated annealing algorithm[18] or a genetic algorithm[19,20] was employed to search for optimal descriptor subsets to build linear and nonlinear models. For both types of models a number of candidate models were generated. In the case of the linear regression model, the final model selected was the one that had the lowest RMS error. In the case of the CNN models, the set of models for descriptor subsets of a given size were analyzed more rigorously to determine the optimal architecture. The final descriptor subset and architecture selected was the one that had the lowest value of the cost function described in Section 3.4.2.

The random forest technique has been discussed in detail in Section 2.3.1. In this study we did not deviate significantly from the default settings of the random forest implementation in the R software package. 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.

7.4 Results

7.4.1 Linear Models

A series of linear models were 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 7.1. Apart from the compounds mentioned above, some models usually had one or two other compounds which could be considered as borderline outliers. Since these borderline compounds varied from model to model, we did not consider them further. Since the three 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 was reduced to 65 compounds. One common feature of these compounds 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 4-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 7.1, compounds **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 $-\log(\mathrm{IC}_{50})$ values for the 3-descriptor model (with training set outliers removed) is shown in Figure 7.2. The statistics of the model are summarized in Tables 7.1 and 7.2. The ranges of the descriptors used are shown in Table 7.3. 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 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 7.2. 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 this outlier is shown in Figure 7.3. A simple comparison of the structure of this outlier with other structures in the dataset 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[21] 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.[22] 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. For the compounds in this study, tertiary nitrogens were generally members of cycles and all compounds had central pyrimidine and piperazine rings. As a result, larger values of this descriptor indicate the presence of cyclic and non-cyclic side chains containing nitrogen.

The RNHS-3 descriptor is a hydrophobic surface area (HSA) descriptor developed by Stanton et al.[23] It is defined as

$$\frac{\max(SA^{-}) \sum H_i}{\log P} \tag{7.1}$$

where $\max(SA^{-})$ is the surface area of the most hydrophilic atom, H_i are the hydrophilic constants (which are values of Wildman and Crippens[24] atomic hydrophobicity constants that are 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.[25] The original HSA descriptors classified atoms as either hydrophilic or hydrophobic using the atomic hydrophobicity constants of Wildman and Crippen.[24] 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.[25] 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 7.4. 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[26] and the details of the interpretation methodology was presented in Section 3.6. A number of examples of this technique have been reported.[22,26] The PLS analysis was carried out with Minitab[27] 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 7.5. Table 7.6 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 7.4, 7.6 and 7.8) to interpret the structure-activity trend characterized by the model.

The most heavily weighted descriptor in PLS component one is SURR-5. As can be seen, its weight is significantly higher than the other two descriptors and thus plays an important role. Figure 7.4 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 7.5. 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 SURR-5 descriptor, this indicates that active compounds are characterized by a large hydrophobic surface area. This is consistent with the fact that the cell based assay used by Pandey et al.[3] 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 under-predict any compounds as shown by the empty upper left corner. However, compounds **11**, **21**, **30** and **55** are over-predicted 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 $-\log(\mathrm{IC}_{50})$ 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 7.5 it is clear that the piperazinylquinazoline backbone is common to both active and inactive structures. The active structures shown (as well as in nearly 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 can be understood by considering the molecular surfaces. Figures 7.9, 7.10 and 7.11 show molecular surfaces for compounds **75**, **93** and **50** colored by hydrophobicity values, drawn using PyMOL.[28] 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 trimethyl amine 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 that 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 7.6 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.[3] 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 dataset, as an indicator of nitrogen containing rings separated by long paths. This would imply that compounds 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 7.7. It is evident that the active compounds have bulky nitrogen containing side groups on the phenoxy ring. In the case of the compounds shown here it is an indole group. In the case of the inactive compounds these are absent. This confirms the observations made by Matsuno[29] and Pandey[3] 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 under-predicts a large number of active compounds which were correctly predicted by component 1. However, component 2 corrects for the over-prediction of some of the compounds from component 1. As can be seen from the score plot in Figure 7.6, compounds **11**, **21**, **30** and **55** are now shifted towards the lower left. Thus, this component compensates for the over-prediction 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 7.5 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 7.8), this component does not predict any low activity compounds. Furthermore the under-predicted compounds (**93** and **91**) have already been correctly predicted as inactive by component 1 and the over-predicted 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 compounds **47** and **96** as active whereas they were under-predicted by component 2.

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 compounds. And as has been shown, large values of this descriptor correlate with higher activities.

As noted before, this dataset had been studied by Khadikar[11] who developed a set of linear regression models. However their methodology differed significantly in that they used the compounds with reported activities in the absence of human plasma. As a result this restricted the size of the dataset. Furthermore the linear models were developed after removing 10 compounds from the already reduced dataset. Finally, their models were developed using a stepwise linear regression technique which is not necessarily an efficient way to search for optimal descriptor subsets.[30,31] 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 dataset, we believe that the lower value of R^2 for our model does not detract from its main utility as an interpretive model. Furthermore, the descriptors present in our best linear model allow a clear interpretation of the structure-activity trend which confirms observations made by Pandey et al.[3] The topological descriptors present in the model described by Khadikar do not lend themselves to a detailed interpretation.

7.4.2 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 was searched for subsets that lead to CNN models with low values of the cost function described in Section 3.4.2. Once a number of suitable subsets were found, the number of hidden layer neurons were 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 7.7. A comparison of the statistics in Tables 7.7 and 7.2 clearly indicate the improved performance of the nonlinear CNN model compared to the linear model. The seven descriptors present in the model are N5CH,[32–34] WTPT-3,[35] WTPT-4,[35] FLEX-4, RNHS- 3,[23] SURR-5[23] 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 5^{th} order chains which are defined as a sequence of 5 atoms containing a ring. This definition thus includes 5-membered rings, 4-membered rings with a methyl side chain and a 3-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[32–34] 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 APAVG descriptor is based on atom pairs as defined by Carhart et al.[36] The atom pair method 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 7.12 shows a plot of the predicted versus observed $-\log(\mathrm{IC}_{50})$ 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.

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 rebuilt. The procedure was repeated 100 times and the averages of the RMSE and R^2 values are reported in Table 7.8. 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.

7.4.3 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 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 and was described above.

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, a node size of 5, and 13 descriptors were used at each split point. The model was built using all the compounds in the dataset and the entire 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. The procedure by which descriptor importances are obtained from a random forest model has been described in Section 2.3.1. Figure 7.13 shows a plot of descriptor importance (only the 10 most important descriptors are shown, ranked in decreasing order of importance).

It is clear that SURR-5 is deemed to be the most important descriptor. Interestingly, RNHS-3 and MDEN-23 are ranked relatively low. 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 7.13 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, 41-descriptor, 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 $\log P$ values of the compounds. The $\log P$ values were calculated using a fragment based approach developed by Mattioni for the HSA descriptors mentioned earlier. A scatter plot of $\log P$ versus SURR-5 for the dataset showed no distinct correlations ($R^2 = 0.17$). We also made scatter plots of $\log P$ 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 $\log P$ value of the compounds.

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 3 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.

7.5 Conclusions

The results presented in this chapter indicate that the linear regression and CNN models developed during this study, 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 supporting 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 generally not amenable to interpretation.[22] It exhibited good statistics for both training and prediction. Furthermore it also contained the top two descriptors, as identified by the random forest, including SURR-5, once again underlying the importance of this descriptor to the structure-activity relationship.

An interesting extension to this work would be to develop a 3-D QSAR model using CoMFA[37,38] which would allow a more detailed view of the specific interactions that

are described by our 2-D 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 2-D methodology we employ cannot provide a direct view of the binding between these compounds and the PDGF receptor, and hence inhibitory activity. This implies that any conclusions made on the basis of our models are oriented towards the activity value rather than activity mechanism (via binding features). A 3-D method such as CoMFA would allow for a more direct understanding of the interactions of the compounds 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 the crystal structure of PDGFR. The crystal structures of tyrosine kinase receptors related to the PDGF receptor have been reported[39,40] though we are not aware of crystal structures of the PDGF receptor specifically. Using 3-D structures based on homology modeling would possibly allow the initial development of a binding model for this receptor and the compounds described here.

In summary this work resulted in the development of 2-D 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.

Table 7.1. The regression statistics for the best linear regression model.

Descriptor	β	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	-8.19	2.56×10^{-11}	1.12

MDEN-23 - molecular distance edge vector between secondary and tertiary nitrogens;[21] RNHS-3 - relative hydrophilic surface area[23] 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[23, 25]

Table 7.2. A 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 7.3. 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

Table 7.4. The 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 7.5. A 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 7.6. The 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

Table 7.7. The 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

Table 7.8. Summary of the statistics for the training, cross-validation and prediction sets from randomized runs using the best CNN model*.

	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

* The architecture used was 7–3–1

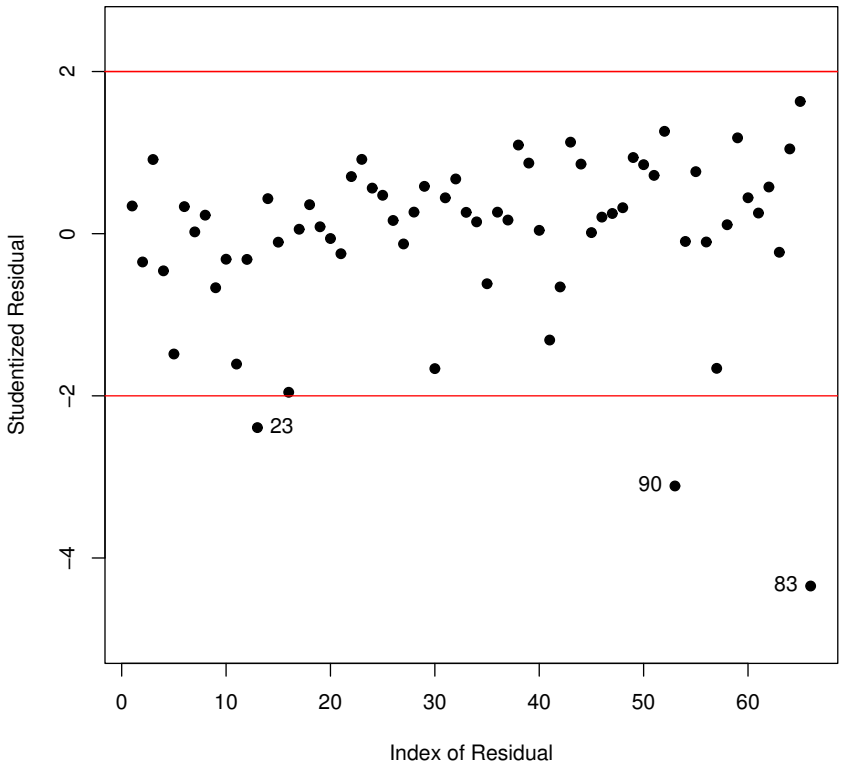


Fig. 7.1. A plot of the studentized residuals from the 3-descriptor linear model with outliers marked.

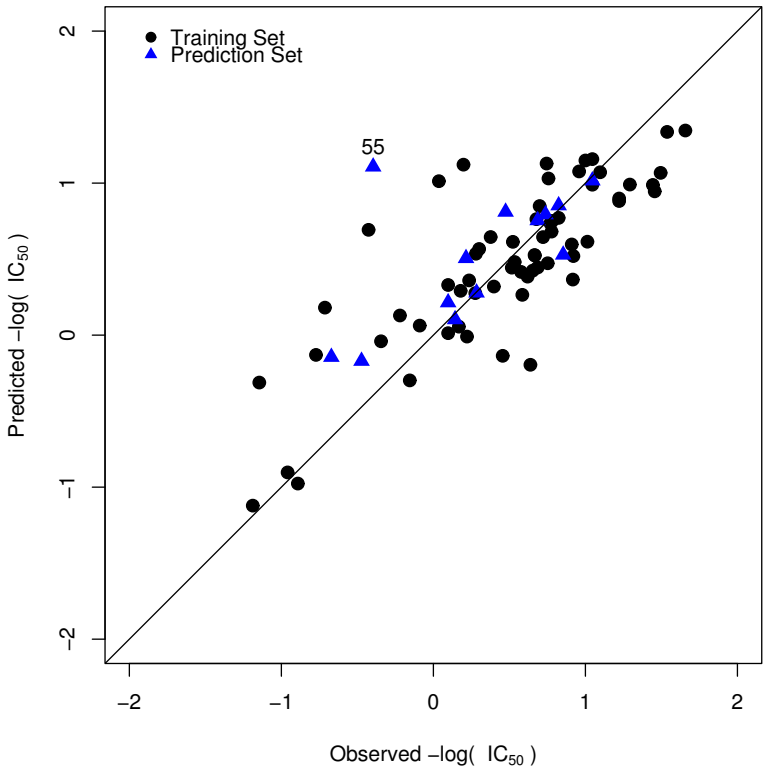


Fig. 7.2. A plot of observed versus predicted $-\log(\mathrm{IC}_{50})$ values from the best linear model after training set outliers were removed. The annotated point represents a prediction set outlier.

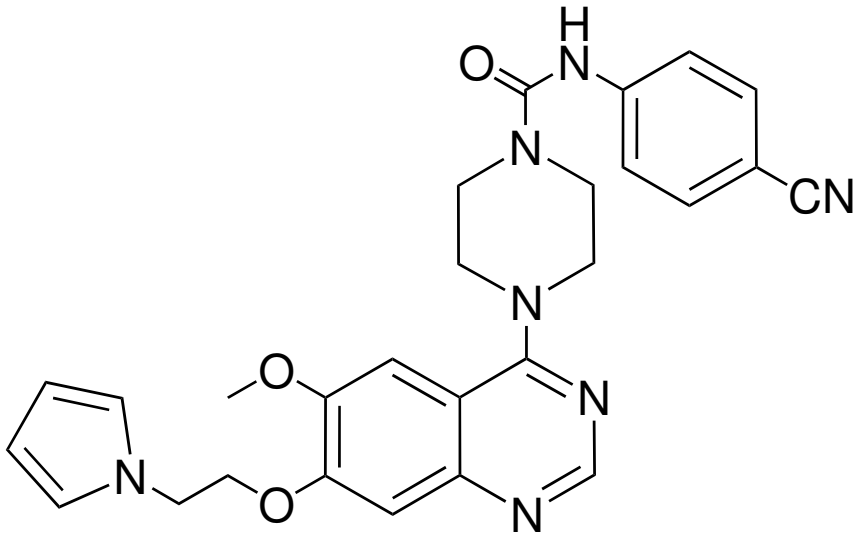


Fig. 7.3. The structure of the prediction set outlier (**55**) from the best linear and nonlinear CNN models.

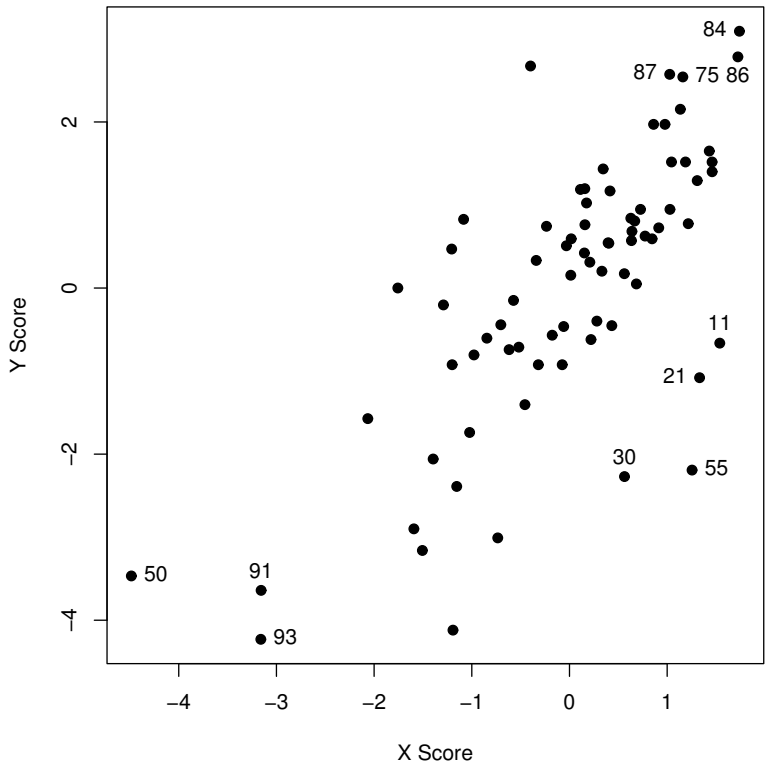


Fig. 7.4. The score plot for PLS component 1.

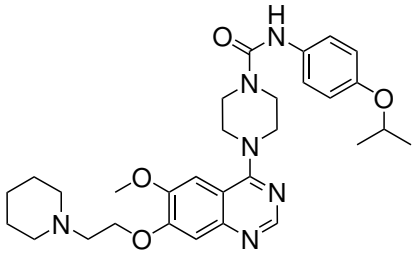
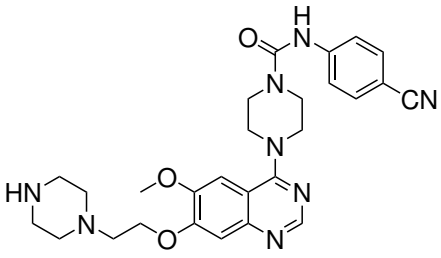
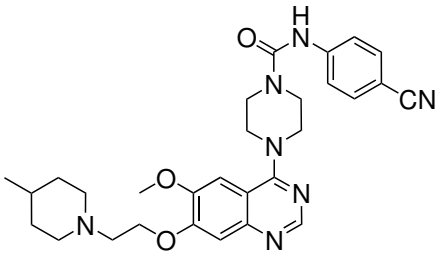
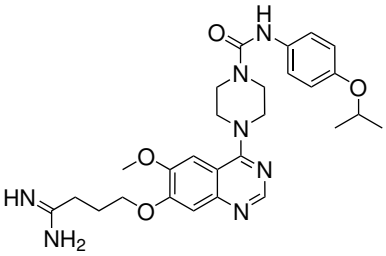
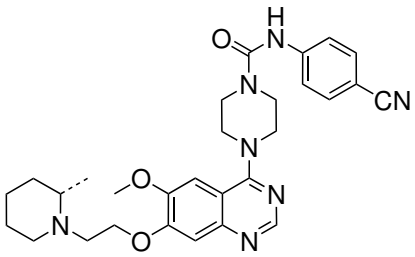
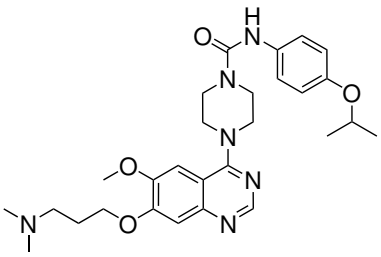
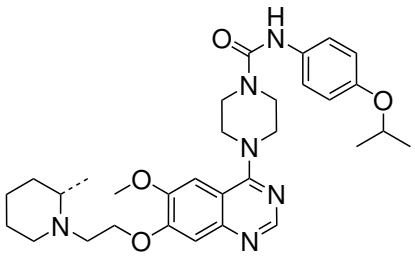
Active	Inactive
	
75 (1.44)	**50** (-0.89)
	
84 (1.65)	**91** (-0.95)
	
86 (1.53)	**93** (-1.18)
	
87 (1.45)	

Fig. 7.5. A comparison of the structures of the active and inactive compounds predicted by component 1 from the 3-component PLS model. Activity values in $-\log(\mathrm{IC}_{50})$ units are provided within brackets.

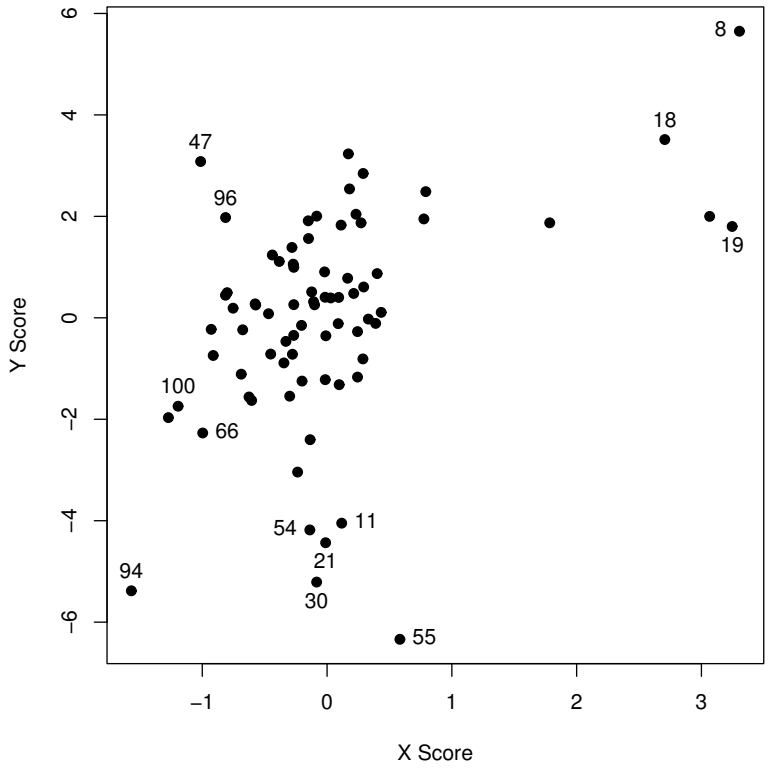


Fig. 7.6. The score plot for PLS component 2.

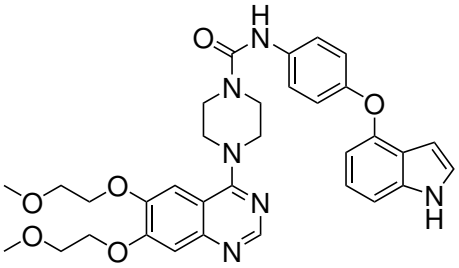
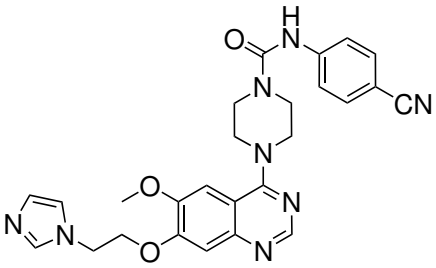
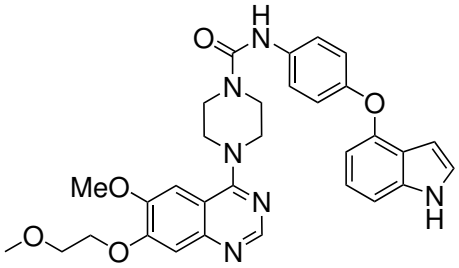
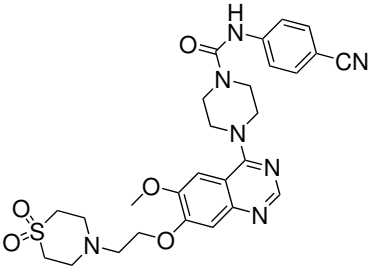
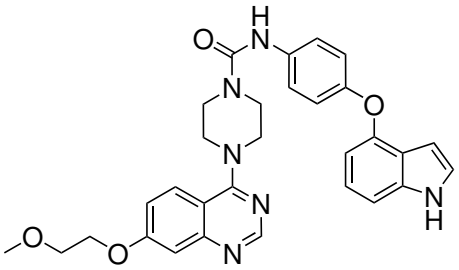
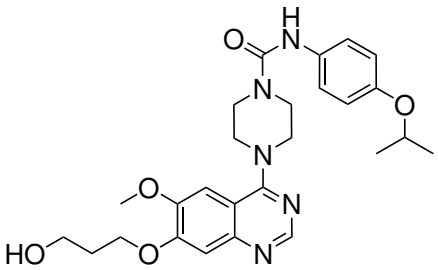
Active	**Inactive**
	
8 (1.49)	**54** (-0.71)
	
18 (0.77)	**66** (-0.47)
	
19 (0.77)	**94** (-1.14)

Fig. 7.7. A comparison of the structures of the active and inactive compounds predicted by component 2 from the 3 component PLS model. Activity values in $-\log(\mathrm{IC}_{50})$ units are provided within brackets.

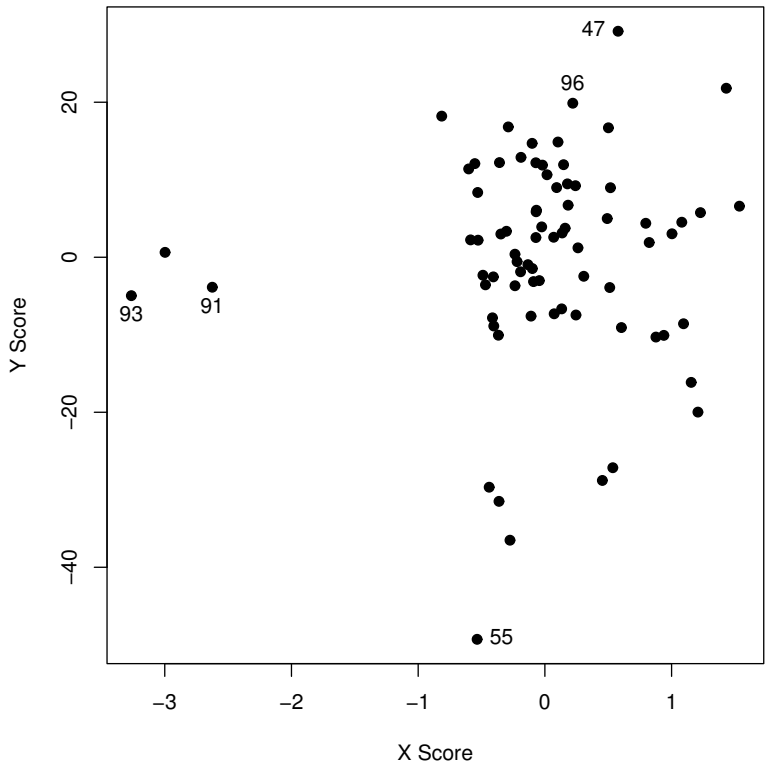


Fig. 7.8. The score plot for PLS component 3.

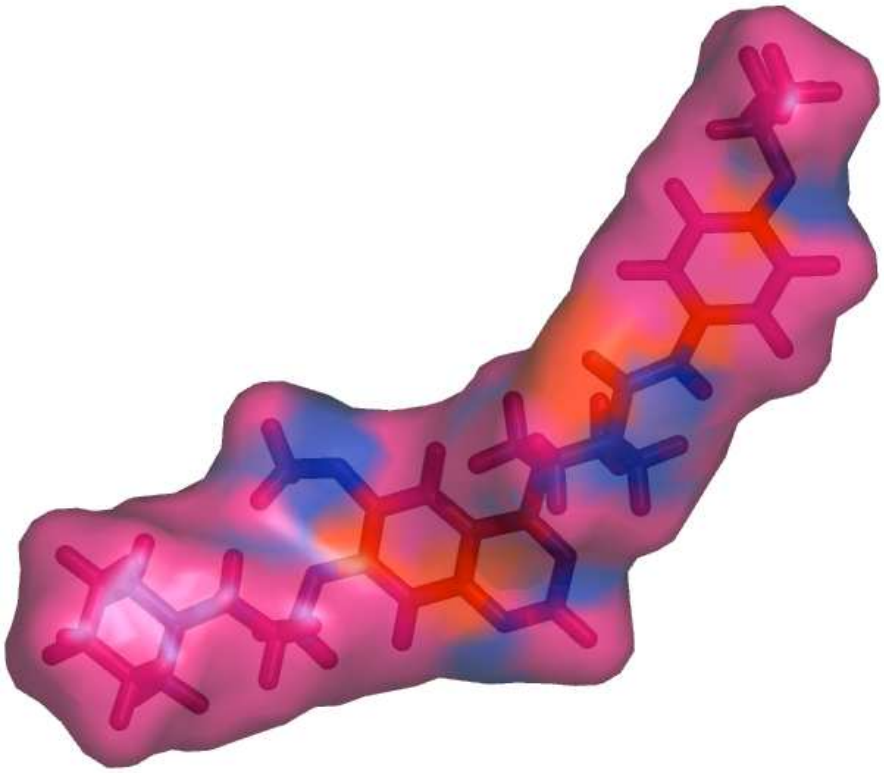


Fig. 7.9. Molecular surface plot of **75**, colored by hydrophobicity values (blue is most hydrophilic and red is most hydrophobic).

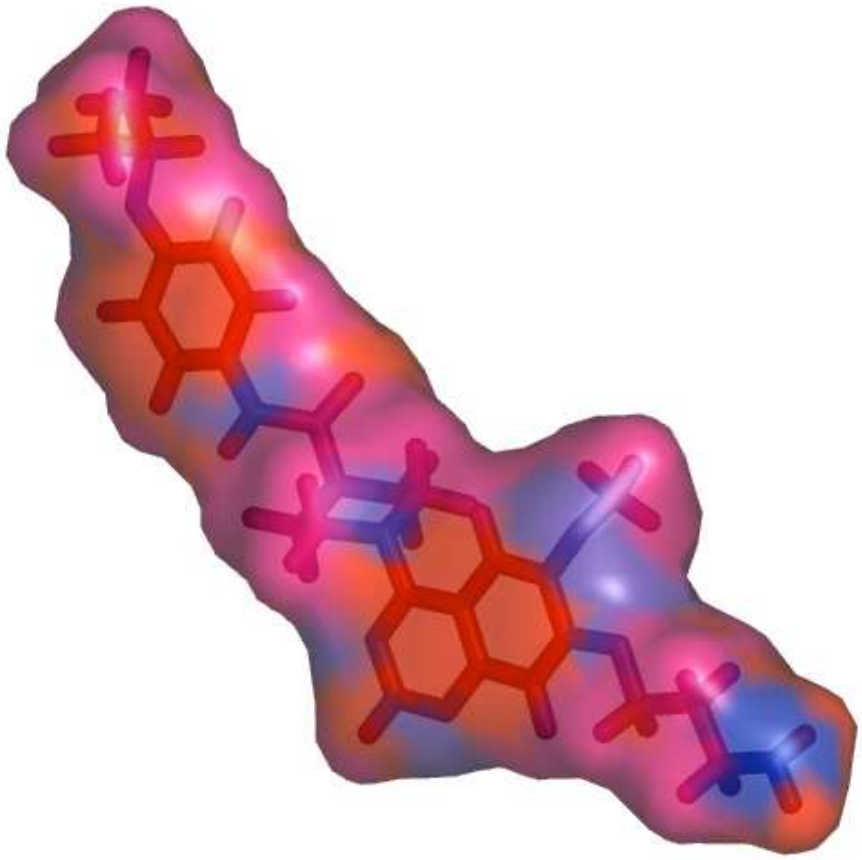


Fig. 7.10. Molecular surface plot of **93**, colored by hydrophobicity values (blue is most hydrophilic and red is most hydrophobic).

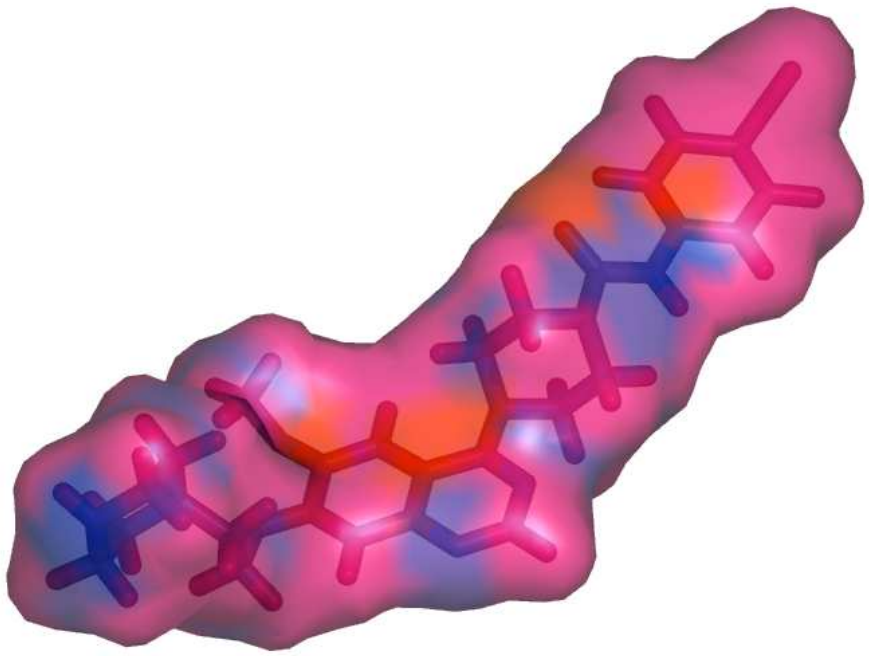


Fig. 7.11. Molecular surface plot of **50**, colored by hydrophobicity values (blue is most hydrophilic and red is most hydrophobic).

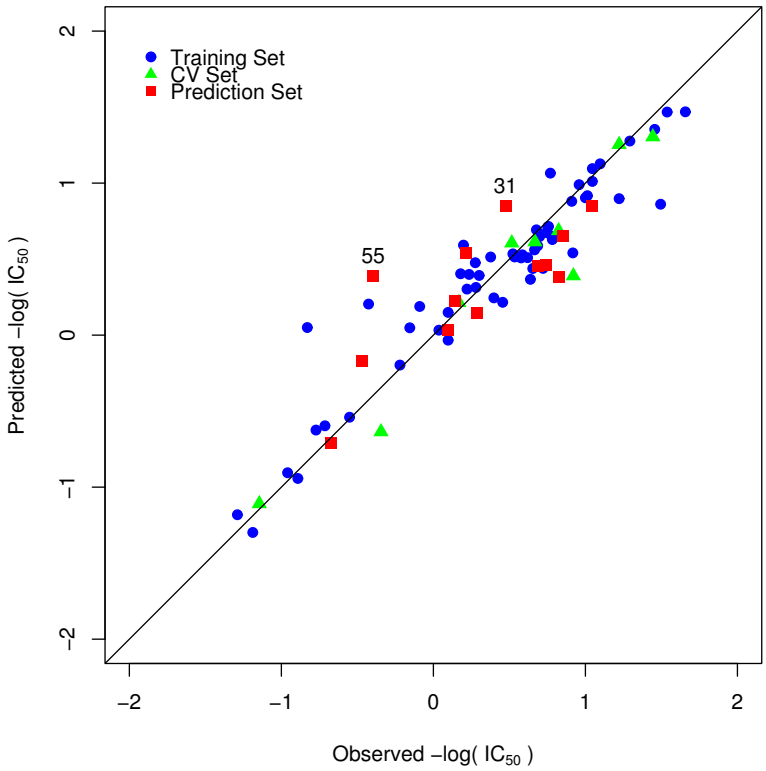
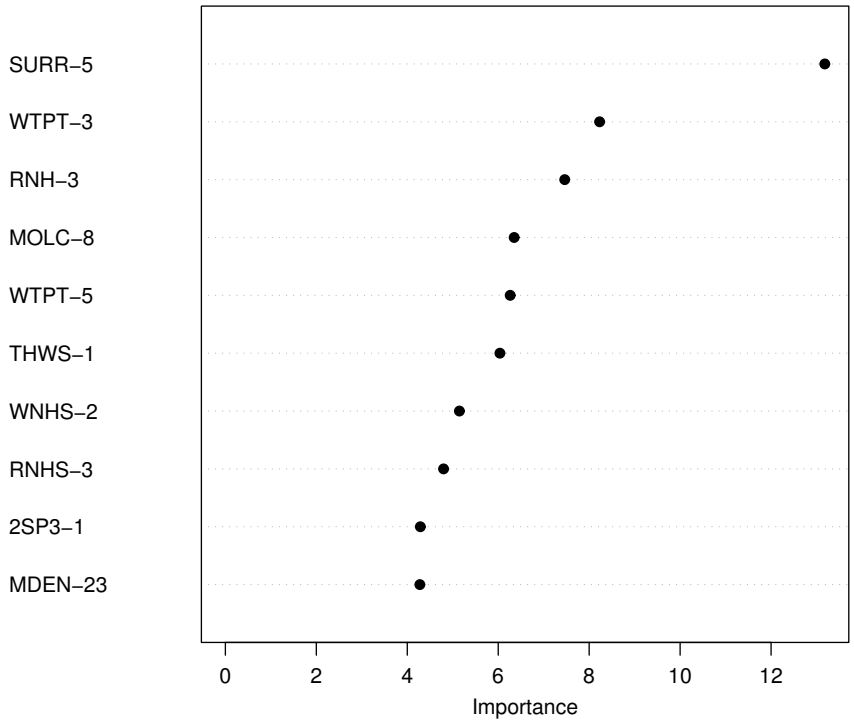


Fig. 7.12. A plot of the observed versus predicted $-\log(\mathrm{IC}_{50})$ values for the best nonlinear CNN model. The annotated points are possible prediction set outliers.

Fig. 7.13. A variable importance plot generated from the random forest model built using the reduced descriptor pool with no compounds 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;[23,25] WTPT-3 - sum of path lengths starting from heteroatoms;[35] RNH-3 - sum of hydrophilic constants divided by the value of log P;[25] MOLC-8 - path-cluster of length 4 molecular connectivity index;[41] WTPT-5 - sum of path lengths starting from nitrogen;[35] THWS-1 - total hydrophobic weighted surface area[25] defined as the sum of the product of atomic log P values and hydrophobic atom surface areas; WNHS-2 - surface weighted hydrophilic surface area[25] 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[23] 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[21]

References

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

[2] Iida, H.; Seifert, R.; Alpers, C.; Gronwald, R.; Philips, P.; Pritzl, P.; Gordon, K.; Gown, A.; Ross, R.; Bowen-Pupe, D. 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.

[3] Pandey, A.; Volkots, D. L.; Seroogy, J. M.; Rose, J. W.; Yu, J.-C.; Lambing, J. L.; Hutchaleelaha, A.; Hollenbach, S. J.; Abe, K.; Giese, N. A.; Scarborough, R. M. 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.

[4] Schlessinger, J.; Ullrich, A. Growth Factor Signaling By Receptor Tyrosine Kinases. *Neuron* **1992,** *9,* 383–391.

[5] Palmer, B.; Kraker, A.; HArtl, B.; Panopoulos, A.; Panek, R.; Batley, B.; Lu, G.; Trumo-Kallmeyer, S.; Showalter, H.; Denny, W. 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.; Kobayashi, Y.; Yagi, M.; Isoe, T.; Nakamura, K.; Osawa, T.; Izawa, T. 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. *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. *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.; Hartl, B.; Amar, A.; Barvian, M.; Showalter, H.; Moore, C. Biochemical and Cellular Effects of c-Src Kinase-Selective Pyrido [2,3-d] Pyrimidine Tyrosine Kinase Inhibitors. *Biochem. Pharamcol.* **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.; O'Hare, J.; Barsoumian, A.; Tomlinson, J.; Ramakrishnan, V.; Fretto, L.; Giese, N. 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.; Chou, J.; Yuan, M. Studies of Chemical Structure Biological Activity Relations Using Pattern Recognition. In *Computer Assisted Drug Design*; Olsen, E.; Christoffersen, R., Eds.; American Chemical Society: Washington D.C., 1979.

[14] Stuper, A.; Brugger, W.; Jurs, P. *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 Development Core Team, "R: A Language and Environment For Statistical Computing", R Foundation for Statistical Computing, Vienna, Austria, 2004 ISBN 3-900051-07-0.

[17] Hypercube Inc., "Hyperchem", 2001.

[18] Sutter, J.; Dixon, S.; Jurs, P. Automated Descriptor Selection For Quantitative Structure-Activity Relationships Using Generalized Simulated Annealing. *J. Chem. Inf. Comput. Sci.* **1995,** *35,* 77–84.

[19] Goldberg, D. *Genetic Algorithms in Search Optimization & Machine Learning;* Addison Wesley: Reading, MA, 2000.

[20] Wessel, M. *Computer Assisted Development of Quantitative Structure-Property Relationships and Design of Feature Selection Routines,* PhD thesis, Pennsylvania State University, 1997.

[21] 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, λ. *J. Chem. Inf. Comput. Sci.* **1998,** *38,* 387–394.

[22] Guha, R.; Jurs, P. The Development of QSAR Models to Predict and Interpret the Biological Activity of Artemisinin Analogues. *J. Chem. Inf. Comput. Sci.* **2004,** *44,* 1440–1449.

[23] Stanton, D.; Mattioni, B. E.; Knittel, J.; Jurs, P. 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,** *44,* 1010–1023.

[24] Wildman, S.; Crippen, G. Prediction of Physicochemical Parameters by Atomic Contributions. *J. Chem. Inf. Comput. Sci.* **1999,** *39,* 868–873.

[25] Mattioni, B. E. *The Development of Quantitative Structure-Activity Relationship Models for Physical Property and Biological Activity Prediction of Organic Compounds,* PhD thesis, Pennsylvania State University, 2003.

[26] Stanton, D. On The Physical Interpretation of QSAR Models. *J. Chem. Inf. Comput. Sci.* **2003,** *43,* 1423–1433.

[27] Minitab, "Minitab", 2003.

[28] DeLano, W. "The PyMOL Molecular Graphics System", 2002.

[29] Matsuno, K.; Ichimura, M.; Nakajima, T.; Tahara, K.; Fujiwara, S.; Kase, H.; Giese, N.; Pandey, A.; Scarborough, R. M.; Yu, J.-C.; Lokker, N.; 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.

[30] Derksen, S.; Keselman, H. J. Backward, Forward and Stepwise Automated Subset Selection Algorithms: Frequency of Obtaining Authentic and Noise Variables. *British Journal of Mathematical and Statistical Psychology* **1992,** *45,* 265–282.

[31] Mantel, N. Why Stepdown Procedures in Variable Selection. *Technometrics* **1970,** *12,* 621–625.

[32] Kier, L.; Hall, L. Molecular Connectivity VII: Specific Treatment to Heteroatoms. *J. Pharm. Sci.* **1976,** *65,* 1806–1809.

[33] Kier, L.; Hall, L. *Molecular Connectivity in Structure Activity Analysis.;* John Wiley & Sons: Hertfordshire, England, 1986.

[34] Kier, L.; Hall, L.; Murray, W. Molecular Connectivity I: Relationship to Local Anesthesia.. *J. Pharm. Sci.* **1975,** *64,*.

[35] Randic, M. On Molecular Idenitification Numbers. *J. Chem. Inf. Comput. Sci.* **1984,** *24,* 164–175.

[36] Carhart, R.; Smith, D.; Venkataraghavan, R. Atom Pairs as Molecular Features in Structure-Activity Studies: Definition and Applications. *J. Chem. Inf. Comput. Sci.* **1985,** *25,* 64–73.

[37] Cramer III, R.; Patterson, D.; Bunce, J. Comparative Molecular Field Analysis (CoMFA). I. Effect of Shape on Binding of Steroids to Carrier Protiens. *J. Am. Chem. Soc.* **1988,** *110,* 5959–5967.

[38] Cramer III, R.; Patterson, D.; Bunce, J.; Frank, I. 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.

[39] McTigue, M.; Wickersham, J.; Pinko, C.; Showalter, R.; Parast, C.; Tempczyk-Russell, A.; Gehring, M.; Mroczkowski, B.; Kan, C.; Villafranca, J.; 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.

[40] Mohammadi, M.; Froum, S.; Hamby, J.; Schroeder, M.; Panek, R.; Lu, G.; Eliseenkova, A.; Green, D.; Schlessinger, J.; Hubbard, S. Crystal Structure of an Angiogenesis Inhibitor Bound to the FGF Receptor Tyrosine Kinase Domain. *EMBO J.* **1998,** *17,* 5896–5904.

[41] Balaban, A. Higly Discriminating Distance Based Topological Index. *Chem. Phys. Lett.* **1982,** *89,* 399–404.