QSAR − How Good Is It in Practice? Comparison of Descriptor Sets on an Unbiased Cross Section of Corporate Data Sets

Peter Gedeck,*,† Bernhard Rohde,‡ and Christian Bartels‡

Novartis Institutes for BioMedical Research, Novartis Horsham Research Centre, Wimblehurst Road, Horsham, West Sussex, RH12 5AB, U.K., and Novartis Institutes for BioMedical Research, CH-4002 Basel, Switzerland

Received September 20, 2005

The quality of QSAR (Quantitative Structure−Activity Relationships) predictions depends on a large number of factors including the descriptor set, the statistical method, and the data sets used. Here we study the quality of QSAR predictions mainly as a function of the data set and descriptor type using partial least squares as the statistical modeling method. The study makes use of the fact that we have access to a large number of data sets and to a variety of different QSAR descriptors. The main conclusions are that the quality of the predictions depends both on the data set and the descriptor used. The quality of the predictions correlates positively with the size of the data set and the range of biological activities. There is no clear dependence of the quality of the predictions on the complexity of the data set. All of the descriptors tested produced useful predictions for some of the data sets. None of the descriptors is best for all data sets; it is therefore necessary to test in each individual case, which descriptor produces the best model. In our tests, 2D fragment based descriptors usually performed better than simpler descriptors based on augmented atom types. Possible reasons for these observations are discussed.

1. INTRODUCTION

Over the years, a large number of studies were published that report QSAR (Quantitative Structure−Activity Relationships) models for biological activities.[1] With few exceptions[2−6] these models are usually based only on a small number of data points, as it is not easy for academic groups to access large QSAR data sets and publish studies based on them. Good QSAR models for large data sets developed in the pharmaceutical industry, on the other hand, can often not be published due to the necessary protection of intellectual property. Larger data sets are typically only found in QSPR (Quantitative Structure−Property Relationships) studies,[7] but findings from such studies have only a limited value for QSAR work.

Another problem also linked to the availability of large data collections is that comparisons between different QSAR approaches are generally only based on one or a few data sets.[8] It is usually not clear if the findings of such studies can be applied to new data sets.

There is also quite a difference between data sets used in publications and real-life data sets originating directly from a drug discovery project. Drug discovery data sets are often far from ideal; for example, estimated data (e.g., $> 10\ \mu$M), drift in data quality over time, diversity of a data set or lack thereof, and insufficient data support to study the interaction of different substitution points. It would be interesting to study the impact of all these issues on the quality of QSAR models; this is, however, beyond the scope of this publication. Here, we will explore the effect of data set diversity, of adding estimated data and of different descriptors. We will look at small and large QSAR data sets from various drug discovery projects run by Novartis and compare QSAR models built with a variety of different descriptors using the results for all of these data sets.

We extracted almost 1000 data sets from our data warehouse ranging from 50 to 3000 data points. About half of the data sets contain estimated data; the other half has them pruned away. In total, we looked at 143 000 different compounds with a large structural variation. For each data set, we used nine different descriptor sets to build QSAR models using partial least squares (PLS). To assess the predictive power of the QSAR models, we split each data set into a training set and an independent test set using two approaches. The first one ('interpolation') assigns every other compound sorted by activity to the training set (50−50 experiment). The second approach ('extrapolation') takes the 10% most active and the 10% least active compounds as the test set (10+10 experiment). A total of about 17 000 QSAR models were generated in this study.

2. METHOD

2.1. Data Sets. Our corporate data warehouse was mined for all biological assays reporting IC_{50} values with between 50 and 3000 results. Most data sets contained results for which only a lower bound for the activity (estimated data, e.g., $IC_{50} > 10\ \mu$M) was given. We studied the effect of including estimated data in the QSAR models by generating two separate data sets for each such assay: one containing all data points using the lower bound as the biological activity (e.g., $IC_{50} > 10\ \mu$M becomes $IC_{50} = 10\ \mu$M), the other where estimated data had been removed.

* Corresponding author phone: +44-1403-320; e-mail: peter.gedeck@novartis.com.

† Novartis Institutes for BioMedical Research, Novartis Horsham Research Centre.

‡ Novartis Institutes for BioMedical Research, Basel.

10.1021/ci050413p CCC: $33.50
Published on Web 08/03/2006

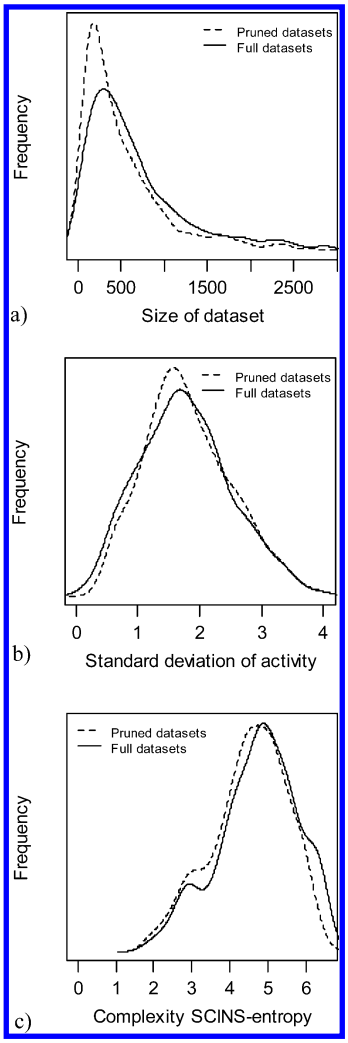


Figure 1. Characterization of data sets. Distribution of (a) data set sizes, (b) standard deviation of activity, and (c) SCINS-entropy of full and pruned data sets.

We studied 944 data sets, 486 of them containing only exact measurements (pruned data sets) and 458 also containing estimated data (full data sets). The differences in numbers are due to the fact that we applied the data set size boundaries after collecting the data for each data set. In total, this study covers slightly more than 570 000 data points.

The Avalon database can organize the samples tested by their so-called main fragment. This is the remainder of the structure after known counterions have been removed and the structure has been neutralized if possible. Stereochemistry is retained during this process. For this study, results originating from samples with the same main fragment were considered repeated measurements and summarized as their median. Note that most descriptors (with the arguable exception of GRIND and HQSAR) used in this study also do not capture stereochemical differences. The final collection of constitutions contained approximately 143 000 different compounds.

The Avalon database reports IC_{50} data in mol/L, e.g., μM or nM. As it is common practice in QSAR studies, we used the negative logarithm of the IC_{50} values divided by 1 mol/L to build the QSAR models. Figure 1 shows the distributions of data set size (1a) and of standard deviation of biological activity (1b) for all data sets. Fifty percent of the data sets had more than 415 rows; the average number of rows is 600. The average of the standard deviation of the logarithm of the biological activity is 1.8.

An additional characteristic of the data set which might impact QSAR model quality is the diversity or complexity of the data set. The measure we used is based on a reduced graph description of the chemical structures in the data set. In the first step, terminal chains are removed to derive the molecular framework as defined by Bemis and Murcko.[9] The framework is then characterized by the number of linkers, rings, ring assemblies, and bridge bonds. In addition, topology of rings and chain lengths are determined. All these characteristics are combined in a string called the SCINS descriptor.[10] The probability distribution $(p_i)_{i=1,n}$ of SCINS descriptors is determined for each data set, and its Shannon entropy is calculated; we call this measure the SCINS-entropy. A data set with compounds of only one molecular framework would have a SCINS-entropy of 0; two equally distributed frameworks have a SCINS-entropy of 1; three equally distributed frameworks have a SCINS-entropy of 1.6; and so on. The distribution of SCINS-entropy of all data sets is shown in Figure 1(c). The average SCINS-entropy is 4.6 corresponding to an equal distribution of 24 different SCINS classes.

2.2. Descriptors. A total of nine different descriptor types were used in this study. They covered a range of types with respect to dimensionality and complexity: pure 2D graph-based descriptors such as atom or fragment based fingerprints or counts or Similog pharmacophore triplets and GRIND, an alignment-free 3D descriptor. The algorithms varied from dictionary based methods (MDL public keys, ALogP atom type counts) to more comprehensive enumerating generators (GRIND, FCFC, HQSAR, and Avalon fingerprints).

2.2.1. GRIND (Almond). It is widely believed that 3D descriptors should provide better descriptions of the binding interactions with the protein. However, most 3D methods suffer from two constraints. First, the correct conformation of a molecule must be used, which may not even be the lowest energy conformation, to compare structurally different compounds; second, the compounds must be properly aligned, a step that is time-consuming and may introduce user bias.

The GRid-INdependent (GRIND) descriptors[11] were developed with the aim to overcome the alignment problem and were therefore selected for this study. However, they still require a suitable conformation. As we were looking at about 143 000 different compounds, we were not able to carry out extensive conformational studies for each of the different assays. Instead, we decided to use Concord 4.0[12,13] to generate a single conformation and use this for the GRIND descriptor calculation. Using the default settings for the GRIND calculations, we then obtained what we refer to as the Almond descriptors. This means GRIND was run with the DRY, O, and N1 probes and a grid spacing of 0.5. On average, this generated a descriptor of length 260. However, when the descriptors of a data set were combined by aligning corresponding blocks and filling missing values with zero, we got a descriptor with an average length of 540.

2.2.2. ALogP Atom Type Counts. In the ALogP method developed by Ghose et al.,[14] the counts of 120 atom types are used in a regression model to predict log P or molar refractivity. As the descriptor formed by the counts of the

Table 1: Feature Classes Used To Generate Avalon Fingerprints[a]

feature class	description	av no. of bits[b]
ATOM_COUNT	count ranges of certain atom types, bond types, and special atom environments	5.7
ATOM_SYMBOL_PATH	paths of atom bond sequences of varying length and specificity	38.8
AUGMENTED_ATOM	indicators of different single shell atom environments	13.9
AUGMENTED_BOND	combined indicators of bond end environments	4.0
HCOUNT_PAIR	hydrogen presence at bond ends	5.8
HCOUNT_PATH	paths starting at hydrogen bearing atoms	18.3
RING_PATH	paths restricted to ring bonds	4.2
BOND_PATH	paths ignoring atom type	21.1
HCOUNT_CLASS_PATH	similar to HCOUNT_PATH but only distinguishing carbon from heteroatoms	9.5
ATOM_CLASS_PATH	selected paths of carbon/heteroatoms ignoring bond order	13.9
RING_PATTERN	paths indicating uncommon ring features	7.5
RING_SIZE_COUNTS	counts of bonds contained in rings of different sizes	4.5
DEGREE_PATHS	paths indicating graph degree	25.1
CLASS_SPIDERS	graph distance triples to a central atom for special central and leaf atom types	8.9
FEATURE_PAIRS	graph distance pairs for special end point atom types	20.8
ALL_PATTERNS	bits from all feature classes ORed together	157.6

[a] See the Supporting Information for more details. [b] Average number of bits set in the 512 bit wide fingerprints of a 1% sample of the Novartis corporate database when the fingerprints generator is restricted to this feature class.

120 atom types basically encodes the major structural features of druglike compounds, it should be useable in QSAR models.

2.2.3. Avalon Fingerprints. Similar to Daylight fingerprints, Avalon uses a fingerprint generator that enumerates certain paths and feature classes of the molecular graph. Table 1 lists the feature classes used. A more detailed description of those classes is available as Supporting Information.[15] The fingerprint bit positions are hashed from the description of the feature; however, the hash codes for all the path-style features are computed implicitly while they are enumerated.

2.2.4. Pipeline Pilot FCFC Fragment Counts. Pipeline Pilot from SciTegic[16] provides descriptors that are based on an extended-connectivity fragmentation scheme. The descriptors have been described previously.[17] In this case, the fragments are based on atom environments of different size. Once all fragments have been generated for a compound, unique numbers are determined based on the Morgan algorithm.[18] For this step, initial atom codes need to be assigned. Here, the functional role of atoms is encoded by looking at its potential as a hydrogen-bond acceptor, a hydrogen-bond donor, a positively or negatively ionizable group, if it is aromatic, or a halogen. A 32-bit hashing scheme is used to limit the range of the generated numbers. Similar to HQSAR (see below), we decided to use the counts of fragments instead of just occurrences as in normal fingerprints. Within Pipeline Pilot this type of descriptor is called FCFC, with the letters standing for feature (F) based on pharmacophore atom type using connectivity (C) based generation of fragments (F) counts (C).

The size of the generated fragments can be controlled by defining the maximum diameter of a fragment as 2, 4, 6, or more bonds. A maximum diameter of 2 bonds means that for each atom, fragments including the next neighbors are considered; with a maximum diameter of 4, the neighbors and their neighbors are included. In the following text, FCFC2 stands for descriptors generated with fragments up to a diameter of 2 bonds, FCFC4 with fragments up to a diameter of 4 bonds, and so on. FCFC0 would stand for the atoms only. The FCFC4 include all FCFC2 fragments; therefore, with increasing diameter the complexity of the descriptors increases and the following relation holds: FCFC0 $\subseteq$ FCFC2 $\subseteq$ FCFC4 $\subseteq$ FCFC6. In this study we included FCFC2, FCFC4, and FCFC6 descriptors. If we talk about this class of descriptors as a whole, we will use FCFCx.

On average, the FCFC2 fragmentation scheme produces approximately 21 fragments per molecule. This number increases to an average of 40 fragments for FCFC4 and 110 fragments for FCFC6. However, when all the fragments of a data set are combined, the total number increases considerably. On average, a data set has about 170 different fragments for FCFC2, 1360 for FCFC4, and 3500 in the case of FCFC6. With such a large number of columns, the PLS calculations for medium to large data sets would have taken several hours. We therefore reduced the size of the full descriptor matrix for a data set by running a principal component analysis (PCA) in R[19] and keeping a maximum of 200 principal components. As most fragments occur only in a few compounds of a data set, this basically removes the noise from the data set. Due to the large size of the initial descriptor matrix, we generated the correlation matrix used in the PCA from a sample of, at the most, 1000 rows in order to reduce the time for the PCA calculation. Although a PCA would not have been necessary for the FCFC2 descriptors, we decided to treat FCFC2 the same way as FCFC4 and FCFC6. A comparison of results obtained for FCFC2 with and without PCA shows that the effect of this is only small. An alternative to using PCA would have been to fold the descriptors onto a smaller range similar to HQSAR.

2.2.5. Hologram QSAR (HQSAR). In Hologram QSAR[20] a compound is fragmented into all possible linear and branched fragments whose size is within a defined range. The fragments can be further distinguished based on the atoms, bonds, connectivity, attached hydrogens, hydrogen bond donor/acceptor potential, and chirality. Each fragment is converted into a practically unique number by converting its representation as an SLN (Sybyl Line Notation) to an integer using a CRC (cyclic redundancy check) algorithm. To reduce the size of the descriptor, the numbers are further mapped into a molecular hologram of user defined length.

In this study, the default settings of the HQSAR module in Sybyl 6.9 were used. The fragments varied in size from 4 to 7 atoms. Only atom, bond, and connectivity information was considered in distinguishing fragments. Holograms of length 401 were generated.

Table 2: Comparison of the Nine Different Descriptor Sets Used in This Study

descriptor	type	number	dimensionality	atom abstraction
Almond	counts	540 (average)	3D	pharmacophore
ALogP	counts	120	2D	functional (atom environment)
Avalon	binary	512	2D	specific and generalized atom types including hydrogen counts
FCFC2	counts	<200 (170 before PCA)	2D	pharmacophore
FCFC4	counts	200 (1360 before PCA)	2D	pharmacophore
FCFC6	counts	200 (3500 before PCA)	2D	pharmacophore
HQSAR	counts	401	2D	atom type
MDL	binary	166	2D	atom type and functional
Similog	counts	200 (2750 before PCA)	2D	pharmacophore

2.2.6. MDL Public Key Substructural Feature Counts. The MDL public keys[21,22] are a set of 166, mostly substructural features. They are a subset of the full set of 960 keys that were developed for rapid substructural searching of ISIS databases.

2.2.7. Similog Descriptors. The Similog descriptors were developed at Novartis by Floersheim for similarity searching in databases.[23] Similar to the above-described FCFC descriptors, atoms are initially classified according to their pharmacophoric properties (potential hydrogen bond donor, potential hydrogen bond acceptor, bulkiness, and electropositivity). There are 8031 nonredundant triplets of such atom classes with mutual graph distance ranges. The counts of these triplets are used as descriptors of a molecule.

On average, a compound has about 400 such nonzero counts. Once the triplets of the compounds in a data set are combined, this number increases to between 1400 and 4000 with an average of about 2750 different keys per data set. We reduced the number of descriptors using PCA to 200.

2.2.8. Summary. The nine different descriptor sets are summarized in Table 2 for reference.

2.3. Statistical Models. To assess the quality of the statistical models, we split each data set into test and training sets using two different rules. In the first approach, the data were first sorted by activity, and the compounds were alternately assigned to test and training sets.[24] Thus, 50% of the compounds were used for training and 50% for testing. This approach will be referred to below as the 50−50 approach. In the second approach, we again sorted each data set by activity and used the top 10% and the bottom 10% as the test set. In case of equal activities, compounds were assigned randomly. The remaining 80% of the compounds were used for training. The second approach will be referred to below as the 10+10 approach. While the first approach should allow assessing the ability of a model to interpolate, the second approach should give us an idea of how well methods extrapolate.

The partial least squares (PLS) method was used with all descriptors. The PLS calculations were carried out in Sybyl 6.9.[25] The PLS default settings were used, which means that all data and descriptor sets were individually scaled. First using the training set, the SAMPLS method[26] with leave-one-out cross-validation was used to determine an optimal number of latent variables (up to 20) by monitoring the q^2 values of the cross-validation. The number of latent variables that lead to a maximum q^2 value was used to generate a PLS model with the full training set. This model was then applied to the independent test set to determine its quality. In most cases, five or less latent variables were selected (Figure 2 shows the distribution of model sizes). In the case of data sets for which a PCA was required due to the large number of descriptors, models with fewer variables were generated on average.

In each case, models were built based on the training sets and then used to predict the independent test sets. The quality of the predictions was assessed by three statistical measures. The first was the multivariate predictive r^2_{pred} given by eq 1.

$$r^2_{\text{pred}} = 1 - \frac{\sum_{i \in \text{test}} (y_i^{\text{pred}} - y_i^{\text{act}})^2}{\sum_{i \in \text{test}} (y_i^{\text{act}} - \bar{y}^{\text{act}})^2} \quad (1)$$

Here, y_i^{pred} is the predicted activities for the test set, and y_i^{act} is the measured activities of the test set; $\bar{y}^{\text{act}}$ is the average of the measured activities of the test set. The value of r^2_{pred}

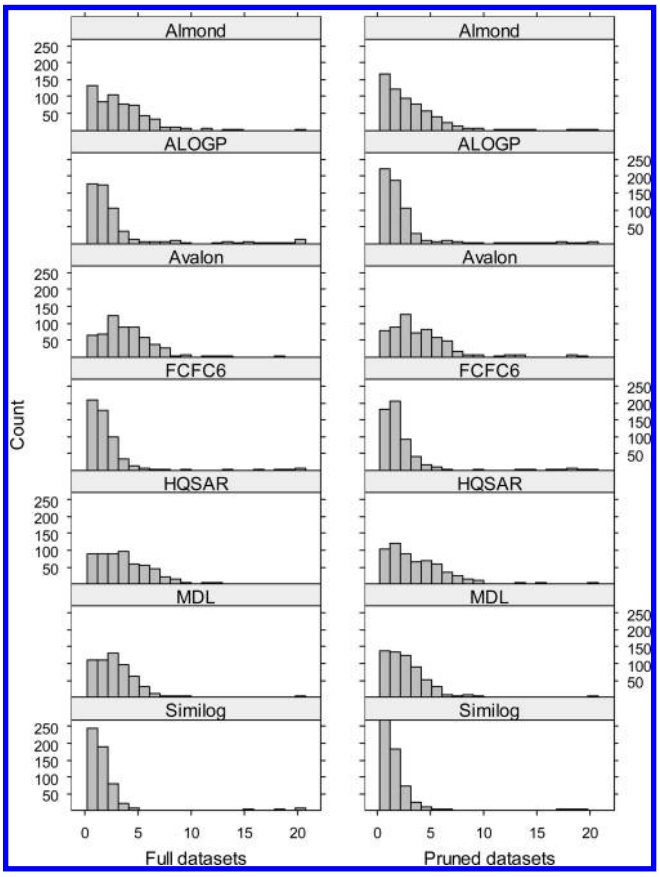


Figure 2. Histogram of the number of latent variables selected by the automatic procedure for the 50−50 experiment. In most cases (85%), the number of latent variables selected was less than 5. The distributions for the 10+10 experiment look very similar. The different behavior of the Similog and FCFC6 models is due to the PCA used for reduction of descriptor numbers prior to PLS.

can range between 1 and $-\infty$. In most graphs, negative values are set to 0 in order to improve the visualization.

The second statistical measure is the squared correlation coefficient between predicted and actual activities given by eq 2.

$$r^2_{\text{corr}} = \frac{(\sum_{i\in \text{test}} (y_i^{\text{pred}} - \bar{y}^{\text{pred}})(y_i^{\text{act}} - \bar{y}^{\text{act}}))^2}{\sum_{i\in \text{test}} (y_i - \bar{y}^{\text{pred}})^2 \sum_{i\in \text{test}} (y_i^{\text{act}} - \bar{y}^{\text{act}})^2} \quad (2)$$

Here, $\bar{y}^{\text{pred}}$ is the average of the predicted values, and the remaining symbols are the same as in eq 1. The squared correlation coefficient r^2_{corr} can vary between 1 and 0. The main difference to r^2_{pred} is that correlation r^2_{corr} measures the association between two variables in general, whereas r^2_{pred} requires the magnitudes of predicted and actual data to be the same. An affine transformation of the predicted values would leave r^2_{corr} unchanged but would cause r^2_{pred} to change.

The root-mean-square error (RMSE) of the prediction is the third statistical measure used. It is given by eq 3.

$$\text{RMSE} = \sqrt{\frac{1}{N-1} \sum_{i\in \textit{test}} (y_i^{\text{pred}} - y_i^{\text{act}})^2} \quad (3)$$

Here, N is the size of the test set.

By combining the nine different descriptors with the 944 different assay results, we obtained a total of 8496 data sets for which we generated statistical models using PLS. In combination with the two different approaches for splitting off independent test sets, we evaluated 16 992 statistical models.

Selection of the test and training sets is an issue in recent literature.[27,28] The data sets of the present study are quite large, so that this is less important here. To address this question and to back up our study, we did two additional sets of experiments for a subset of 120 arbitrarily selected data sets using the Avalon descriptors. First, we did experiments by randomly assigning 50% of the compounds to the training set (random experiment). Second, we repeated the 50−50 experiment but scrambled the measured values by randomly interchanging the measured values of compounds (scrambling experiment). We repeated the random and scrambling experiments 10 times for each of the selected data sets. Figure 3 summarizes the results. As expected, the r^2_{pred} values for a specific data set depend on the test set and training set sample and there vary in the 10 repeats. There are cases where the standard deviation of the r^2_{pred} values can be large; however, in most cases the variation is only small, and the median standard deviation of r^2_{pred} is only 0.05. The scrambling experiments gave on average r^2_{pred} values of −0.06 with a median standard deviation for 10 repeats of 0.03, showing, unsurprisingly, that we cannot make predictions, if we remove the correlation of the structures with the measured activities.

2.4. Computational Details. In most cases, data management, data processing, and data analysis were controlled through protocols in SciTegic's Pipeline Pilot[16] running on a standard desktop PC under Windows XP. The calculations were carried out on a variety of machines. If possible, time-consuming compute jobs were submitted to an 8 CPU IBM xSeries 445 server running Redhat Enterprise Linux AS release 3 (Xeon, 3GHz, 8Gb) through a SOAP service. The Sun Grid Engine (SGE) queuing system[29] was used to manage the computing jobs on this machine.

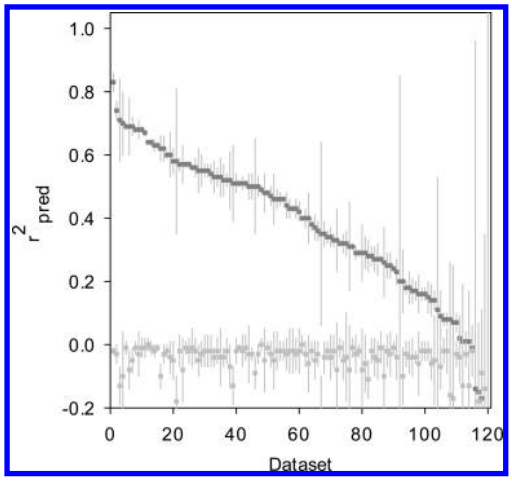


Figure 3. Validation through randomization experiments. One hundred twenty data sets were randomly selected and QSAR models generated using Avalon descriptors. For each data set, two randomization experiments were performed: Ten random test/training set splits (50−50) were generated and QSAR models built. The dark gray squares give the average r^2_{pred} values for each data set. The gray lines are ± standard deviation. The median standard deviation of the r^2_{pred} values for all data sets is 0.05. The light gray squares are based on a y-scrambling experiment. On average, the r^2_{pred} values dropped to −0.06 with a median standard deviation for 10 repeats of 0.03.

Almond/GRIND descriptors were calculated on a SGI Tezro (2 R16000, 700 MHz, 1 Gb), taking about 10 days to complete the full data set using the Almond program as available from Tripos.[25] The required 3D structures were created using Concord[12,13] on the same machine. ALogP, FCFC, and MDL public keys descriptors were calculated within Pipeline Pilot 3.0.6 from SciTegic[16] on a standard desktop PC running Windows XP. The dimensionality reduction using PCA for the FCFC descriptors principal component analysis was run on the IBM server using R version 1.9.0.[19] HQSAR descriptors were calculated within Sybyl 6.9 on the IBM server using the mentioned SOAP services. Avalon descriptors were calculated on a standard PC with a 1.6 GHz Intel Pentium 4 processor running under Windows XP. Similog descriptors were calculated on a SGI Origin 2000 (8 R10000, 250 MHz, 1 Gb) using proprietary software.

The PLS models were generated in Sybyl 6.9 on the IBM server using a SPL script to control the calculation.

RESULTS AND DISCUSSION

3.1. Individual Results. There are clear differences in the performance of the various descriptor sets. Table 3 collates the results for each experiment split by data set type. Figure 4 visualizes the same results graphically. In Figure 5, the distribution of r^2_{pred} is compared for the full and the pruned data sets. Including estimated data improves, in general, the quality of the predictions. The improvement of r^2_{pred} for the 50−50 experiment is on average 0.06, which is quite

Table 3: Performance of the Different Descriptor Sets[a]

descriptor	50−50 experiment		10+10 experiment	
	full data sets (458)	pruned data sets (486)	full data sets (458)	pruned data sets (486)
Almond	5% (24)	4% (17)	4% (18)	5% (23)
ALogP	12% (54)	9% (43)	7% (32)	8% (37)
Avalon	29% (135)	18% (87)	17% (80)	17% (81)
FCFC2	16% (72)	10% (46)	6% (27)	6% (28)
FCFC4	19% (87)	9% (45)	7% (30)	7% (33)
FCFC6	23% (103)	12% (55)	9% (42)	7% (32)
HQSAR	27% (121)	16% (76)	17% (77)	16% (79)
MDL	18% (83)	11% (55)	9% (40)	8% (41)
Similog	15% (69)	6% (30)	4% (17)	4% (18)
best descriptor	35% (161)	24% (119)	23% (105)	22% (109)

[a] The table lists the percentage of good models assuming a cutoff value of $r^2_{pred} > 0.5$ for good models. The number of good models is given in brackets.

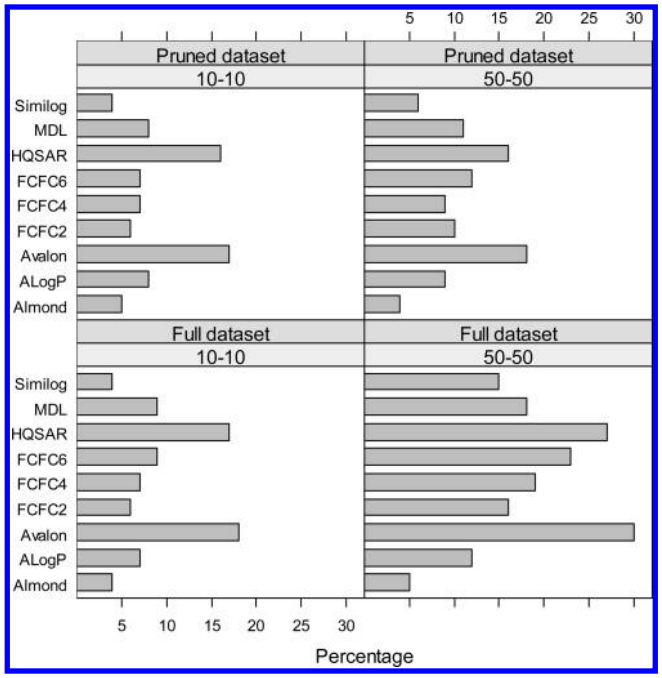


Figure 4. Graphical representation of Table 3, showing the percentage of good models ($r^2_{pred} > 0.5$) by descriptor, data set type, and experiment.

considerable. Using a *t*-test it can be shown that this improvement is significant ($p<0.005$) for the 50−50 experiment. In the 10+10 experiment we see no general shift (results not shown).

This can be explained by noting that the addition of the estimated data may affect the overall quality of the predictions in two different ways and in opposite directions. First, increasing the size of the training data set is expected to increase the quality of the model and to improve the precision of the predictions. Second, estimated data are difficult to predict since estimates are lower bounds, and even exact estimates of the value will have a residual. [30] Thus adding estimates to the test set is expected to decrease r^2_{pred}.

The relative importance of these two opposing effects is different for the 50−50 and the 10+10 experiment, since the estimates are distributed differently in the test and training sets in these two experiments. In the 50−50 experiment, the test and the training sets contain the same fraction of estimates. In the 10+10 experiment, in all but one case, the 10% least actives contain all the estimated data points, and

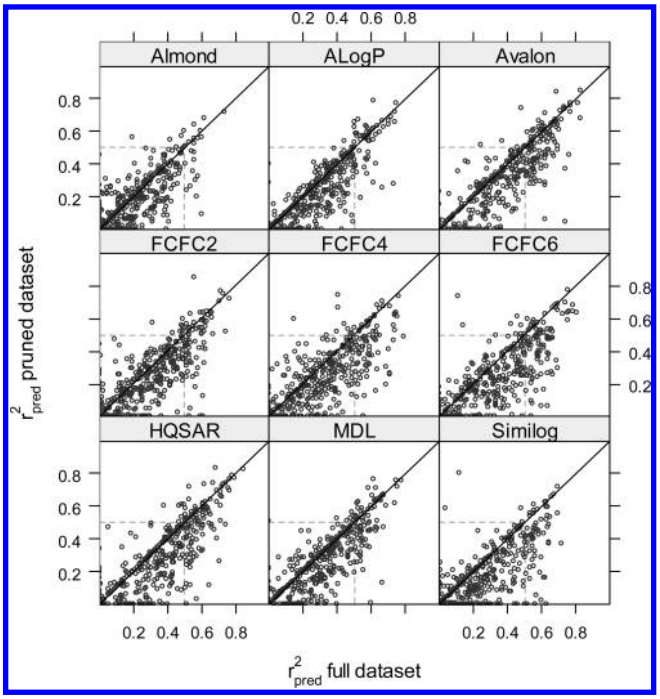


Figure 5. Comparison of r^2_{pred} obtained with full and pruned data sets in the 50−50 experiment. Negative r^2_{pred} values are set to 0 to improve the visualization. Including estimated data generally improves the quality of the prediction. The dotted line indicates the cutoff value of 0.5 used to identify good models.

on average 5% of the test set data points are estimated. Based on this irregular distribution of the estimates, the second effect that decreases the quality of the predictions upon addition of the estimates is expected to be larger in the 10+10 experiment than in the 50−50 experiment and seems to have counterbalanced the advantage of adding additional data to the training set.

Avalon and HQSAR descriptors generated the largest number of good models. The 3D method Almond produced the smallest number of good models. This is disappointing but not surprising considering that we generated the descriptors from a single, automatically selected conformation. Although GRIND descriptors are believed to be relatively robust with respect to conformational changes, having more realistic, data set dependent conformations would certainly be beneficial.

The good performance of Avalon and HQSAR relative to other fragment based approaches such as FCFCx, MDL public keys, and ALogP atom type counts could be due to the fact that Avalon and HQSAR use fragment sets that are biased toward the features contained in the data set. In the case of HQSAR it is done automatically for an individual data set. For Avalon, in contrast, this was achieved by tuning the fingerprint generator during the development of the Avalon data warehouse to improve screen-out for generic medicinal chemistry queries. To some extent, the FCFCx descriptors achieve the same goal as HQSAR; however, the large number of descriptors and consequently the large sparseness of the descriptor matrix might introduce too much noise for the statistical method. Although the ALogP atom type counts have been shown to produce good property models (ALogP and MR),[31,32] it might be that they encode structural features that are not sufficient to describe binding of ligands to proteins.

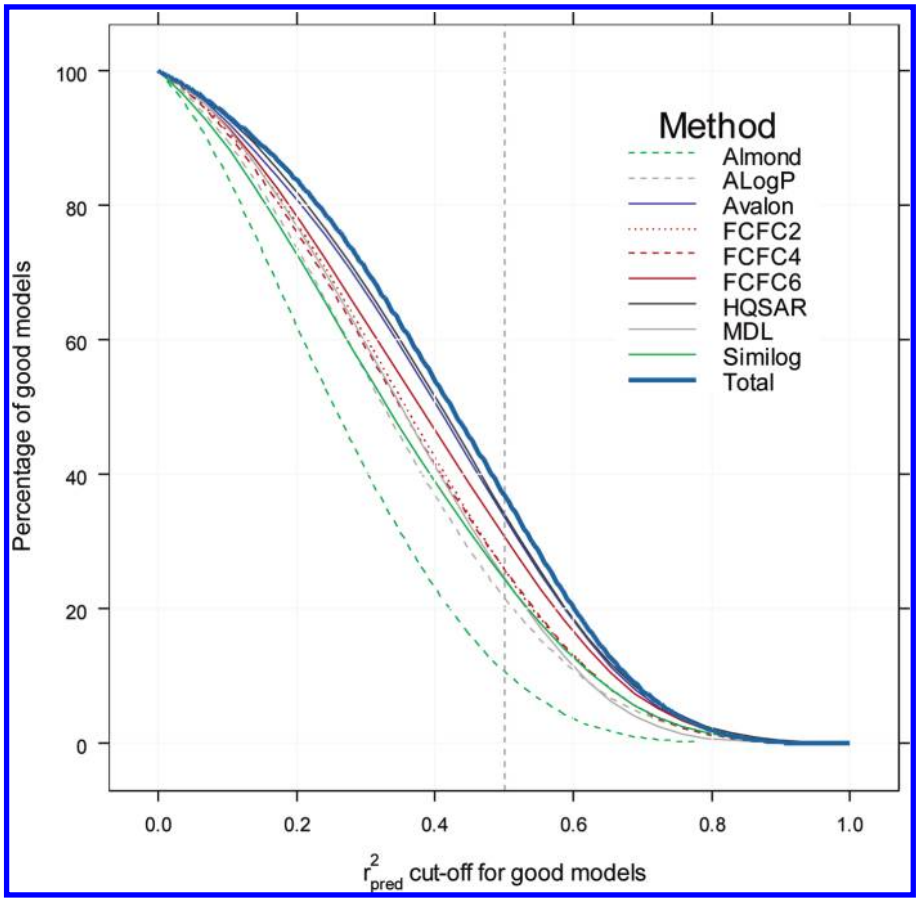


Figure 6. Cumulative histogram showing the percentage of good models as a function of r^2_{pred} cutoff for the different methods for the full data sets. The curve labeled 'Total' shows the percentage of good models if the best descriptor set is used for the respective data set. The dashed vertical line indicates the used cutoff of $r^2_{pred} = 0.5$. For the generation of this graph, kernel density distribution functions were used.

It could be argued that a cutoff of $r^2_{pred} = 0.5$ is too strict a criterion to identify good models. Figure 6 gives the percentage of good models as a function of the cutoff value. A not too unreasonable cutoff value of $r^2_{pred} = 0.4$ would result in 55% of all models being classified as good. Also, the graph shows that choosing a different cutoff value would have changed only the absolute values but not the qualitative interpretation. The choice of a cutoff value is debatable and depends strongly on how the model is used. If accurate predictions are required, then a larger cutoff value is certainly justified; if however the model is used to prioritize compounds in a combinatorial library, then even models with a low r^2_{pred} value can become useful.

3.2. Influence of Data Set Characteristics. It is to be expected that model quality is dependent on the characteristics of the data set used. We therefore analyzed the data sets by size, spread of biological activity, and structural diversity of the compounds. Ultimately, one would hope that this would yield guidelines on minimum requirements for deriving good models.

Figures 7−9 show model quality r^2_{pred} as a function of data set size, spread of biological activity in the data set, and complexity of the data set. The results exhibit trends, indicated by the locally weighted LOESS[33] regression lines, which show the expected behavior. Increasing the data set size or increasing the range of biological activities makes it more likely to obtain better models. A standard deviation of the pIC_{50} values of at least 1.0 seems to be required to get good models. This might be necessary in order to compensate

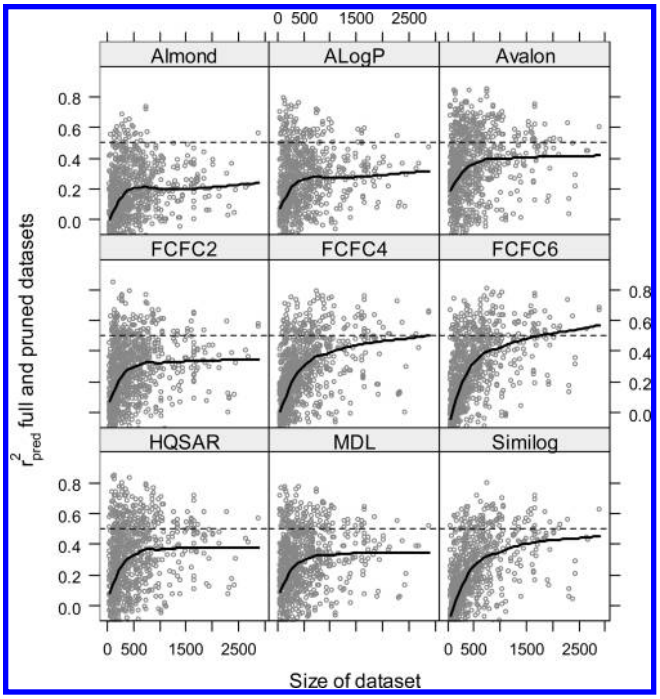


Figure 7. Dependence of r^2_{pred} on data set size. The thick black line is a locally weighted LOESS regression line; the dotted black line indicates the r^2_{pred} cutoff value for good models. There is some correlation between data set size and quality of the models.

the influence of experimental errors on model quality. Also, the correlation of model quality r^2_{pred} with the range of

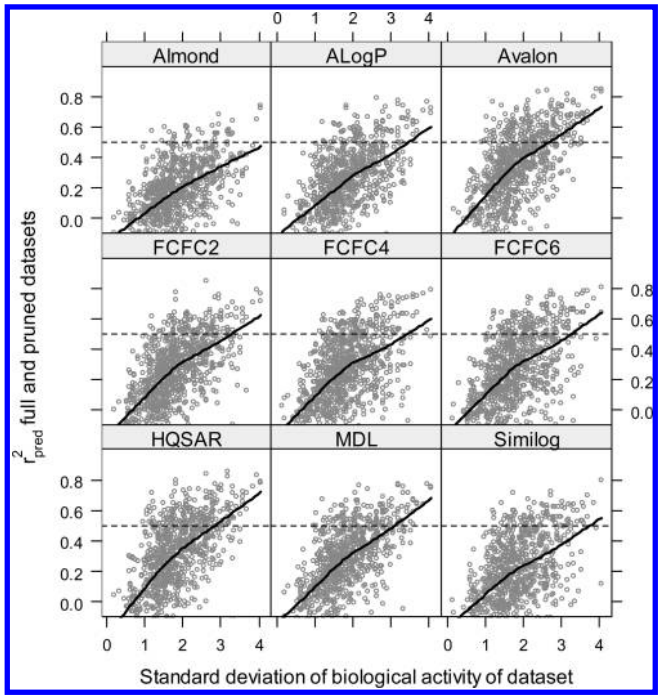


Figure 8. Dependence of r^2_{pred} on spread of the biological activity of the data sets. The thick black line is a locally weighted LOESS regression line; the dotted gray line indicates the r^2_{pred} cutoff value for good models. There is a correlation between spread of biological activity and quality of the models.

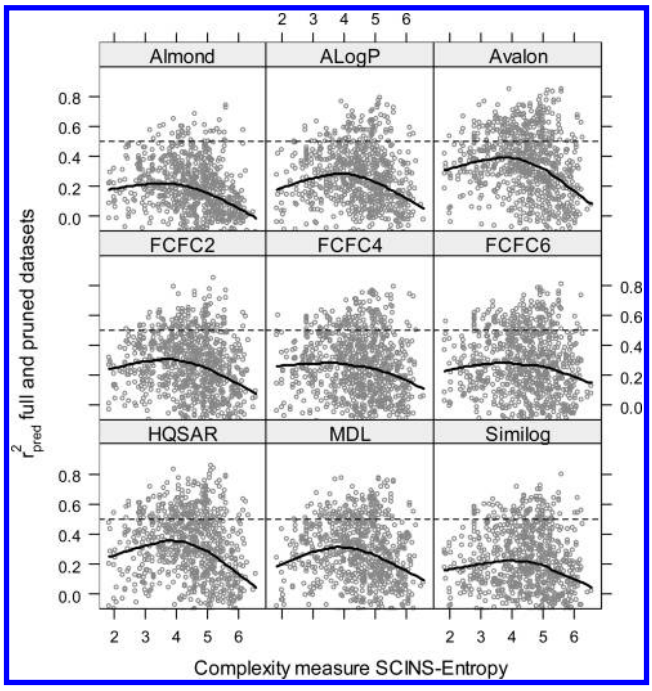


Figure 9. Dependence of r^2_{pred} on data set complexity SCINS-entropy. The thick black line is a locally weighted LOESS regression line; the dotted gray line indicates the r^2_{pred} cutoff value for good models.

biological activities is explained at least partially by the direct dependence of the quality r^2_{pred} on the variance of the biological activity as expressed in eq 1.

In Figure 10, the RMSE of the model is shown as a function of the standard deviation of the biological activity. The distance of the points from the diagonal is a function of r^2_{pred} and indicates the quality of the model. Points further

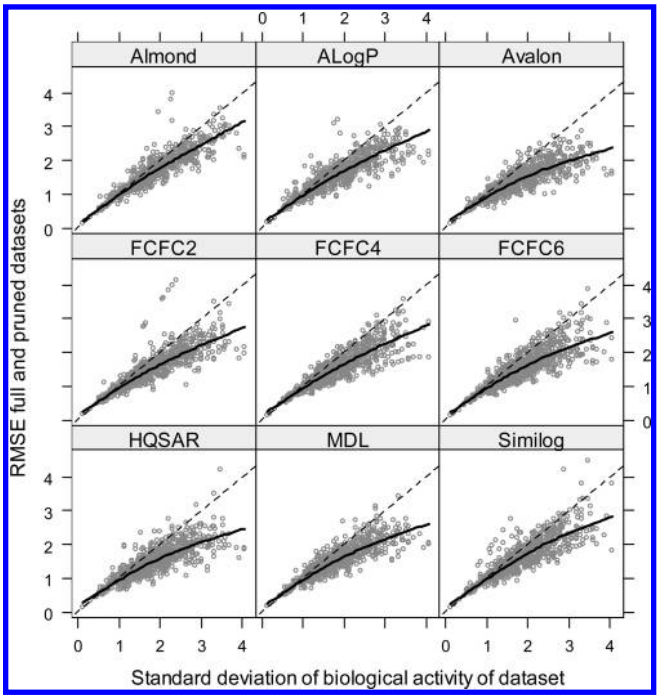


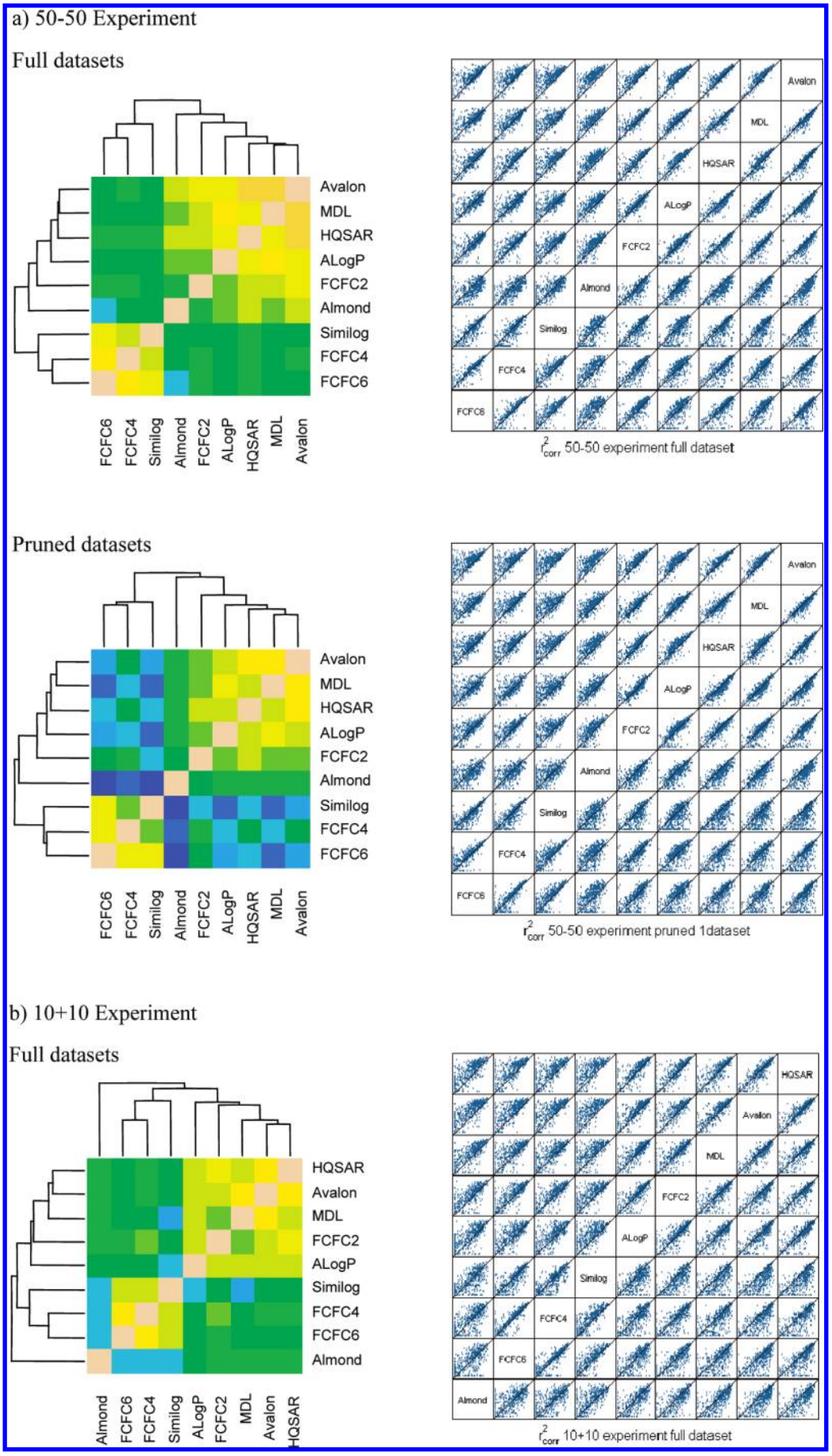
Figure 10. Dependence of RMSE on standard deviation of the biological activity of the data sets. The thick black line is a locally weighted LOESS regression line; the dotted gray line indicates the line of equality.

below the diagonal characterize good models that provide additional information compared to a simple model that assumes that the activities are distributed independently of the structures. This graph basically reiterates the results from Figure 8. A standard deviation of the biological activity (pIC50) of at least 1.0 is required to obtain good models. The curvature of the LOESS regression line indicates the increased probability of achieving good models for data sets with a larger activity range.

The variation of model quality with SCINS-entropy is less obvious (Figure 9). The data seem to suggest small and large values of the SCINS-entropy leading to a decrease of model quality. A small SCINS-entropy indicates a data set of low complexity, which would probably benefit from R-group descriptor based QSAR. Large SCINS-entropy, in contrast, means a data set of high structural diversity. Such data sets could be validation sets from HTS or results of profiling activities. Interestingly, most of the complex data sets also have only a small spread of activity, which would support this interpretation.

3.3. Comparing Descriptors. The r^2_{corr} values obtained for the different descriptor sets are compared in Figure 11 in a scatterplot matrix. It is clear from these graphs that the correlation of the different r^2_{corr} values between different descriptors is quite high; it ranges between 0.51 and 0.91. A more condensed comparison of the correlation is shown in the heat maps.

The descriptors fall into three distinct sets: (1) the three high-dimensional descriptors FCFC4, FCFC6, and Similog [All three descriptors use a comparable pharmacophoric atom typing, which may explain that the methods perform similarly.]; (2) Almond, the only 3D descriptor in our study, forms basically a group of its own; and (3) the remaining descriptors: ALogP, Avalon, FCFC2, HQSAR, and MDL.


a) 50-50 Experiment
Full datasets
Avalon
MDL
HQSAR
ALogP
FCFC2
Almond
Similog
FCFC4
FCFC6
FCFC6 FCFC4 Similog Almond FCFC2 ALogP HQSAR MDL Avalon
r^2_{corr} 50-50 experiment full dataset
Pruned datasets
r^2_{corr} 50-50 experiment pruned 1dataset
b) 10+10 Experiment
Full datasets
Almond FCFC6 FCFC4 Similog ALogP FCFC2 MDL Avalon HQSAR
r^2_{corr} 10+10 experiment full dataset

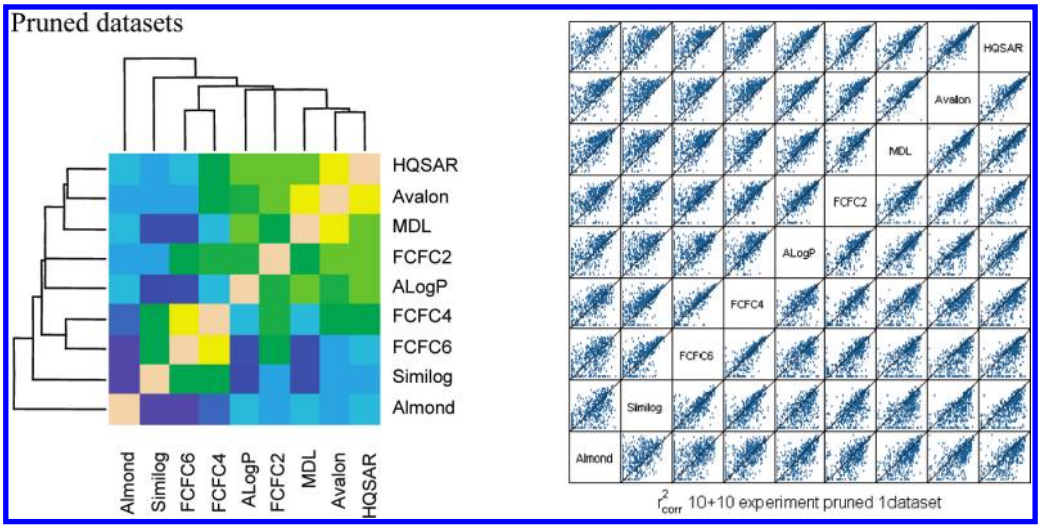


Figure 11. Pairwise comparison of r^2_{corr} results as scatterplot matrices (right) and visualization of correlation matrices as heat maps (left). In each case, the correlation matrix between r^2_{corr} values for different descriptors was determined, and the correlation coefficients used to obtain a hierarchical clustering (single-linkage clustering) were indicated by the dendrogram shown in the heat maps. The hierarchical clustering was also used to order the descriptors in the scatterplot matrix. Note that the order of descriptors in the various graphs can therefore be different. The colors correspond to correlation coefficients that range from 0.5 (blue) over 0.7 (green), 0.8 (yellow), to 0.9 (orange).

Almond and Similog are both descriptors that capture longer ranging relative arrangements of pharmacophoric features and therefore differ most from fragment based 2D descriptors which encode more local features. This difference is also reflected in the heat maps.

The highest correlations are obtained for the pairs Avalon/MDL and Avalon/HQSAR. Avalon and MDL, the two binary fingerprints, are essentially substructure based descriptor sets with MDL encoding less information than Avalon. This causes the MDL descriptor based models to be consistently worse than the Avalon models. Avalon and HQSAR have a more or less similar performance.

In the FCFCx descriptor family, FCFC4 and FCFC6 behave very similarly. The added complexity when going from a maximum fragment diameter of 4−6 seems not to improve models. This could be due to the principal component analysis that was required prior to the PLS model building. The PCA will remove rare fragments and therefore especially large fragments. In contrast, FCFC2 performs quite differently. This is mainly due to small data sets (<500 data points) for which good models are obtained for FCFC2 but not for FCFC4 and FCFC6. The added complex fragments seem to lead to overfitting of the training set and, therefore, poor performance of the models on the test set.

Overall, it is possible to get good models for 30−40% of all data sets using an automated, noninteractive protocol. Even though some descriptors perform considerably better than others, no single descriptor is capable of generating all possible good models. For one data set, even the least well performing GRIND/Almond descriptor gives the only good model. Only very few data sets—less than five—yield good models with all descriptors. For 20% of the data sets, good models could be obtained with four or more descriptor sets.

3.4. Varying the Complexity of Structural Descriptors. Most descriptor sets used in this study are based on fragmentations of the molecular structure. AlogP and the FCFCx based descriptors seem to perform equally well achieving good models for about 10% of the data sets, with the exception of the 50−50 experiment with the full data sets (Table 3). The two descriptor sets differ in the type of how they treat individual atoms. ALogP basically has only fragments of individual atoms but differentiates atom types by local environment—discriminating for example the oxygen in an aldehyde and an amide group. FCFCx differentiates atoms by their pharmacophore binding properties. The apparent simplification compared to ALogP is counterbalanced by the creation of multiatom fragments, which essentially will allow differentiating functional groups. As already mentioned, the FCFCx descriptors perform a lot better in the 50−50 experiment with the full data sets. AlogP descriptors only encode the properties of a very small atomic environment, and it would be impossible to infer much of the global molecular structure from the AlogP descriptors. This is not the case for the FCFCx descriptors; the larger fragments are capable of capturing the relative arrangement of binding features that are further apart. The increase in complexity when going from FCFC2 to FCFC4 seems to cause a slight improvement of the models in general. A much smaller variation is observed between the results for FCFC4 and FCFC6, which is most likely due to the initial noise reduction using principal component analysis (PCA).

With the default settings used, the complexity of the HQSAR fragments is probably somewhere between FCFC4 and FCFC6. Fragments up to a size of 7 atoms correspond to considering some of the second next neighbors (as in FCFC4) and by adding connectivity information of the atoms extends, to some extent, to the third next neighbors (as in FCFC6). The effect of folding into a vector with 401 elements is comparable to the principal component analysis but will retain the influence of some fragments only occurring in few molecules (removed as noise in PCA).

The Avalon fingerprints and the MDL public keys were both developed to allow efficient database queries. Interestingly, the r^2_{corr} values for both methods are usually highly correlated, which can easily be seen in the hierarchical clustering dendrograms of Figure 11. This high correlation indicates that both descriptors encode similar structural

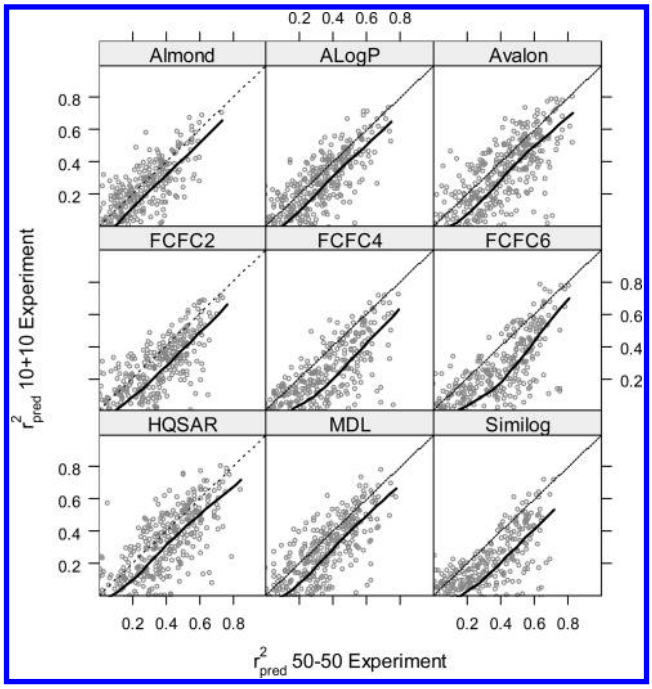


Figure 12. Comparison of r^2_{pred} values obtained for the 50−50 and the 10+10 experiment for the full data sets. The dotted line is the line of equality; the thick black line is a locally weighted LOESS regression.

features in the data sets. However, the MDL public keys are too simple to capture the structural information required for consistently good QSAR models.

3.5. Comparing Inter- and Extrapolation. It is informative to compare the results for the 50−50 experiment with the 10+10 experiment. Figure 12 compares the r^2_{pred} values of both cases on the full data sets. Irrespective of the descriptor set used, the r^2_{pred} values obtained in the 50−50 experiment are on average higher than those in the 10+10 experiment. The shift seems to be somewhat larger for the FCFC4, FCFC6, and Similog descriptors. As these are the high-dimensional descriptor sets, this could indicate that their models concentrate too much on details in the training set and are less suitable for extrapolating. Similar but less pronounced results are obtained with the pruned data sets.

3.6. Comparing r^2_{corr} Correlation and r^2_{pred} as Indicators of Model Quality. In addition to r^2_{pred}, we also calculated r^2_{corr} between the predicted and actual activities for the test set compounds. In Figure 13, we compare the results for the two experiments, the 50−50 and the 10+10 split of the full data sets; the results for the pruned data sets look qualitatively the same. For the 50−50 experiment there is basically no difference between the two measures, which means that the QSAR models predict not only activities in the right order but also of the right magnitude. This is different for the 10+10 experiment, where r^2_{corr} is on average about 0.2 higher than r^2_{pred}. To understand this difference, it is important to see that correlation r^2_{corr} measures the association between two variables in general, whereas r^2_{pred} requires the magnitudes of predicted and actual data to be the same. An affine transformation of the predicted values would leave r^2_{corr} unchanged but would cause r^2_{pred} to change. In the 50−50 experiment, the activities are not only predicted more or less in the right order but also are of the right magnitude. Therefore, r^2_{pred} and r^2_{corr} values are basically the same for good models. Predicting activities correctly in the 10+10 experiment would in many cases require knowledge about structural features not present in the training set. Activities (pIC_{50}) are therefore usually underestimated to some extent for the more active compounds and overestimated for the less active compounds. This causes r^2_{pred} to be lower than r^2_{corr}. A high r^2_{corr} value, however, indicates that the model is capable of distinguishing compounds with high and low activity.

4. SUMMARY AND OUTLOOK

The performance of different QSAR methods was compared using 944 data sets with a total of 143 000 compounds. The compounds were described by nine different types of descriptors and statistical models built using PLS. No further attempts were made to improve the results, e.g., by using nonlinear statistical models such as neural networks or by clustering the compounds. In particular, we did not remove any outliers. Typical lead optimization data sets contain mainly compounds belonging to a small number of chemical

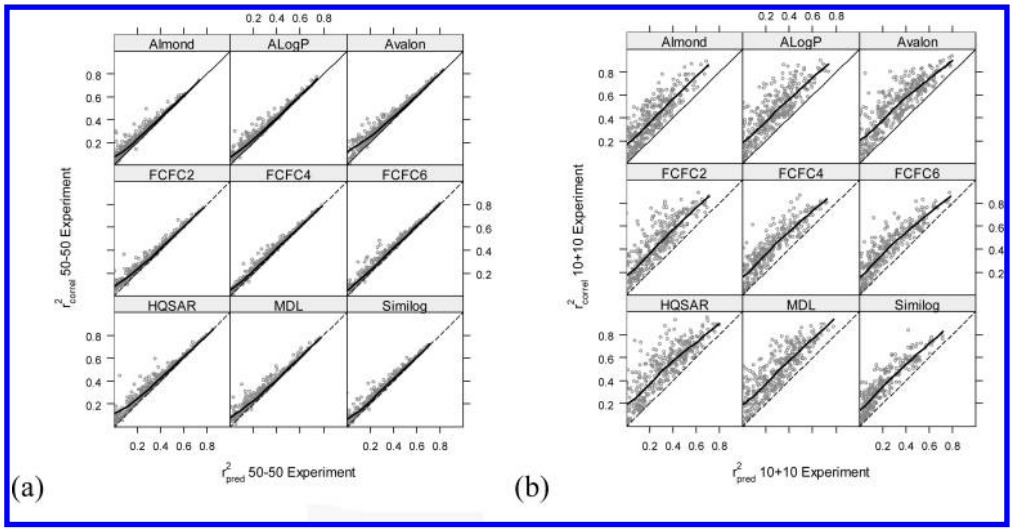


Figure 13. Comparison of r^2_{pred} and r^2_{corr} for (a) the 50−50 experiment and (b) the 10+10 experiment using the full data sets. The black line is the line of equality; the thick black line is a locally weighted LOESS regression. The results for the pruned data sets are qualitatively the same.

classes and a set of structurally very diverse compounds from hit finding activities or competitor patents. These singletons are in general very difficult to predict. Similar results were reported by Olah et al. in a study based on smaller, published structure−activity data sets.[35] Their conclusions are consistent with our findings.

Thus, the present tests evaluated a set of QSAR methods with respect to the consistency of their performance with a large number of drug discovery data sets. The fragment based Avalon and the HQSAR descriptors produced the largest number of good models for the data sets. The other descriptors generated a smaller number of good models but complemented the results of the models based on the HQSAR or Avalon descriptors in some cases.

Interestingly, the computationally most involved model based on the Almond descriptors, which uses the three-dimensional structure of the compound for the prediction, performed poorest. This might be due to the fact that the description of the three-dimensional structure of compounds is difficult, in general (e.g., see ref 34), and that the protocol to generate the structures in the present study was inappropriate, in particular. Thus, the information on an inadequate three-dimensional structure might have been encoded into the Almond descriptors used for the prediction.

A pairwise comparison of the model performance measures r^2_{corr} allowed identifying similarities of the descriptors. The used descriptors could clearly be separated into three different classes. This classification could be rationalized by the type of information encoded in the descriptors. It will be interesting to add additional descriptors derived using conceptually different approaches, e.g., electrotopological indices, and include these in the analysis.

There exist a large number of options to try to improve the results obtained with the methods tested. This includes combined use of different types of descriptors to select for each data set the descriptor that performs best, fine-tuning of the parameters of each of the methods, use of nonlinear statistical models to fit the data, preprocessing of the input data to identify outliers, clustering the compounds, or better sampling of the three-dimensional structure used with the three-dimensional descriptors. For all these potential improvements, the inexpensive methods using the HQSAR or Avalon descriptors and a linear statistical model serve as a performance baseline difficult to outperform.

We encourage anyone interested in trying his or her favorite QSAR approach with our data sets to contact us.

ACKNOWLEDGMENT

Thomas Müller is gratefully acknowledged for initiating this study by proposing to generate QSAR models based on Avalon fingerprints using all data available in our internal database. Tripos is gratefully acknowledged for making Almond available during a beta test release of Sybyl 7.0.

Supporting Information Available: A more detailed description of the feature classes used to generate Avalon fingerprints. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Livingstone, D. J. The characterization of chemical structures using molecular properties. A survey. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 195−209.

(2) Mattioni, B. E.; Jurs, P. C. Development of quantitative structure−activity relationships for a set of carbonic anhydrase inhibitors. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 94−102.

(3) Kauffman, G. W.; Jurs, P. C. QSAR and *k*-nearest neighbor classification analysis of selective cyclooxygenase-2 inhibitors using topologically based numerical descriptors. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1553−1560.

(4) Xiao, Z.; Xiao, Y.-D.; Feng, J.; Golbraikh, A.; Tropsha, A.; Lee, K.-H. Antitumor agents. 213. Modeling of epipodophyllotoxin derivatives using variable selection *k* nearest neighbour QSAR method. *J. Med. Chem.* **2002**, *45*, 2294−2309.

(5) Burden, F. R.; Winkler, D. A. Robust QSAR models using Bayesian regularized neural networks. *J. Med. Chem.* **1999**, *42*, 3183−3187.

(6) Sheridan, R. P.; Feuston, B. P.; Meiorov, V. N.; Kearsley, S. K. Similarity to molecules in the training set is a good discriminator for prediction accuracy in QSAR. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1912−1928.

(7) Katritzky, A. R.; Fara, D. C.; Petrukhin, R. O.; Tatham, D. B.; Maran, U.; Lomaka, A.; Karelson, M. The present utility and future potential for medicinal chemistry of QSAR/QSPR with whole molecule descriptors. *Curr. Top. Med. Chem.* **2002**, *2*, 1333−1356.

(8) Sutherland, J. J.; O'Brien, L. A.; Weaver D. F. A comparison of methods for modeling quantitative structure−activity relationships. *J. Med. Chem.* **2004**, *47*, 5541−5554.

(9) Bemis, G. W.; Murcko, M. A. The properties of known drugs. 1. Molecular frameworks. *J. Med. Chem.* **1996**, *39*, 2887−2893.

(10) The internally developed SCINS descriptor encodes number of chains, rings, ring bonds, counts of ring system types, and chain lengths in a unique string.

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

(12) Pearlman, R. S. Rapid generation of high quality approximate 3D molecular structures. *Chem. Des. Aut. News* **1987**, *2*, 1−6.

(13) Pearlman, R. S. 3D molecular structures: generation and use in 3D searching. In *3D QSAR in Drug Design*; Kubinyi, H., Ed.; ESCOM: Leiden, 1993; pp 41−79.

(14) Ghose, A. K.; Viswanadhan, V. N.; Wendoloski, J. J. Prediction of hydrophobic (lipophilic) properties of small organic molecules using fragment methods: An analysis of AlogP and CLogP methods. *J. Phys. Chem. A* **1998**, *102*, 3762−3772.

(15) Lone-pair heterocycles were fingerprinted with localized bonds and as aromatic if they matched the Hückel rule to code both feature sets and use them for substructure screening. This style of fingerprinting was designed to allow screening for substructures with either interpretation of aromaticity while not requiring too many fingerprint bits to support.

(16) *Pipeline Pilot 3.0.6*; Scitegic, Inc.: 9665 Chesapeake Dr., Suite 401, San Diego, CA 92123, U.S.A., 2003.

(17) Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A. Comparison of topological descriptors for similarity-based virtual screening using multiple bioactive reference structures. *Org. Biomol. Chem.* **2004**, *2*, 3256−3266.

(18) Morgan, H. L. The generation of a unique machine description for chemical structures − a technique developed at Chemical Abstracts Service. *J. Chem. Doc.* **1965**, *5*, 107−113.

(19) R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria, 2004. http://www.r-project.org/ (accessed Jun 5, 2006).

(20) Tong, W.; Lowis, D. R.; Perkins, R.; Chen, Y.; Welsh, W. J.; Goddette, D. W.; Heritage, T. W.; Sheehan, D. M. Evaluation of quantitative structure−activity relationship methods for large-scale prediction of chemicals binding to the estrogen receptor. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 669−677.

(21) The MDL or ISIS public keys are already used in MACCS-II and are also known as MACCS keys. ISIS/Base and MACCS are both products of MDL Information Systems, Inc., San Leandro, CA, http://www.mdli.com/ (accessed Jun 5, 2006).

(22) Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. Reoptimization of MDL keys for use in drug discovery. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1273−1280.

(23) Schuffenhauer, A.; Floersheim, P.; Acklin, P.; Jacoby, E. Similarity metrics for ligands reflecting the similarity of the target proteins. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 391−405.

(24) This can be considered a pseudorandom division because the activities are scattered by measurement errors. The method has the advantage that the activity distributions of corresponding training and test sets are very similar.
(25) Sybyl 6.9 is available from Tripos Inc., St. Louis, MO.
(26) Bush, B. L.; Nachbar, R. B. Sample-distance Partial Least Squares: PLS optimized for many variables, with application to CoMFA. *J. Comput.-Aided Mol. Des.* **1993**, *7*, 587−619.
(27) Golbraikh, A.; Tropsha, A. Predictive QSAR modeling based on diversity sampling of experimental datasets for the training and test set selection. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 357−369.
(28) Hawkins, D. M.; Basak, S. C.; Mills, D. Assessing model fit by cross-validation. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 579−586.
(29) http://gridengine.sunsource.net/ (accessed Jun 5, 2006).
(30) A possibility to consider this would be to introduce the experimental cutoff value as an upper bound of the predicted activities.
(31) Wildman, S. A.; Crippen, G. M. Prediction of physicochemical parameters by atomic contributions. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 868−873.
(32) Ghose, A. K.; Crippen, G. M. Atomic physicochemical parameters for three-dimensional structure-directed quantitative structure−activity relationships. I. Partition coefficients as a measure of hydrophobicity. *J. Comput. Chem.* **1986**, *4*, 565−577.
(33) Cleveland, W. S.; Devlin, S. J. Locally weighted regression: An approach to regression analysis by local fitting. *J. Am. Stat. Assoc.* **1988**, *83*, 596−610.
(34) Bartels, C.; Stote, R. H.; Karplus, M. Characterization of flexible molecules in solution: the RGDW peptide. *J. Mol. Biol.* **1998**, *284*, 1641−1660.
(35) Olah, M.; Bologa, C.; Oprea, T. An automated PLS search for biologically relevant QSAR descriptors. *J. Comput.-Aided Mol. Des.* **2004**, *18*, 437−449.

CI050413P