



## Refinement of Catalyst hypotheses using simplex optimisation

Ulf Norinder

*AstraZeneca R&D Södertälje, SE-151 85 Södertälje, Sweden*

*(Tel: +46 8 553 25057; Fax: +46 8 553 288 77; E-mail: ulf.norinder@astrazeneca.com)*

Received 17 May 1999; Accepted 1 February 2000

**Key words:** ACE, Catalyst, citest, HIVPR, hypothesis, receptor information, simplex optimisation, SQEP, 3D QSAR

### Summary

The program HypoOpt in combination with the MSI program citest has been used to optimise and expand 3D QSAR Catalyst hypotheses using simplex optimisation coupled with cross-validation. Three data sets related to angiotensin converting enzyme inhibition, squalene epoxidase inhibition and HIV protease inhibition were used to investigate the outcome of hypothesis optimisation. Simplex optimisation using leave-one-out cross-validation during the hypothesis refinement resulted in improved models with respect to predictivity of an external test set. Furthermore, the utilisation of the geometry of the active site for the HIV protease inhibitors, represented by Catalyst 'excluded volume' features, resulted in an optimised hypothesis with improved predictivity compared with the corresponding hypothesis derived without receptor information.

### Introduction

Since the advent of 3D QSAR techniques such as the hypothetical active site lattice (*HASL*) method [1], receptor modeling from the three dimensional structure and physicochemical properties of the ligand molecules (*REMOTEDISC*) [2] and Comparative Molecular Field Analysis (*CoMFA*) related methods [3–8] in the late 1980s, a large number of investigations have been described in the literature. One of the latest additions to the field of 3D QSAR is the Catalyst methodology [9]. So far, only a few publications on the use of Catalyst have been published [10–11]. A majority of these publications have used Catalyst as a tool to obtain reasonable starting alignments for further 3D QSAR studies of CoMFA type. However, there are limitations and rules in the Catalyst hypothesis methodology as to how the program will develop a hypothesis. These protocols make the program more or less useful to deal with a particular problem at hand. In this work the possibilities to further develop and optimise Catalyst hypotheses have been investigated.

### Data sets

In this study three different data sets have been used to investigate different aspects of Catalyst hypothesis refinement.

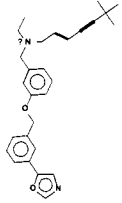
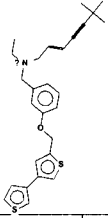
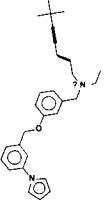
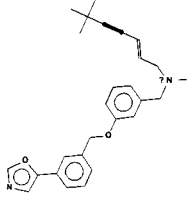
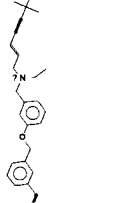
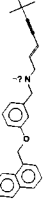
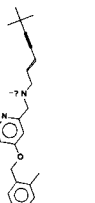
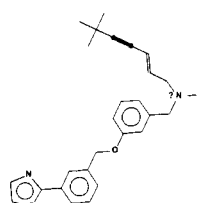
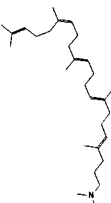
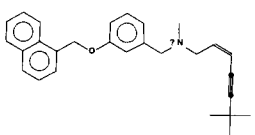
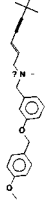

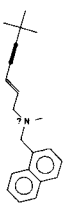
1. The first data set of some Angiotensin Converting Enzyme (ACE) inhibitors [12–13] (Table 1, Scheme 1) was used to probe the impact of simplex refinement, cross-validation and different mathematical formalisms (see below for further explanations) on a merged Catalyst hypothesis. The data set was divided into four equal parts (five compounds each). A cross-validation procedure [14] was applied where each of the four parts was left out once as test set and the corresponding hypothesis was developed with the remaining three parts.
2. The second data set of some squalene epoxidase inhibitors (SQEP) [12, 15] (Table 2, Scheme 2) was investigated for the same purpose as with the ACE data set using the same cross-validatory procedure and, in addition, to also study the effect of different mapping (see Mapping of a compound to a hypothesis section for further details) conditions

| ACE          |             | 20 Compounds |         | Page 1 |
|--------------|-------------|--------------|---------|--------|
|              |             |              |         |        |
| nleu-ala-pro | val-trp     | leu-ala-pro  | ile-tyr |        |
|              |             |              |         |        |
| phe-ala-pro  | arg-ala-pro | phe-pro-pro  | ile-pro |        |
|              |             |              |         |        |
| ala-pro      | ala-val     | glu-ala-pro  | val-pro |        |
|              |             |              |         |        |
| gly-phe      | ala-leu     | ala-gly      | gly-glu |        |
|              |             |              |         |        |
| gly-lys      | pro-pro     | ala-his      | gly-asp |        |

Scheme 1. The ACE data set.

of compounds to the generated hypothesis. In this data set there exists an additional complication in that two of the least active compounds, SDZ-87-469 and terbinafine, do not map to the hypothesis using the 'Fast Fit' superimpositioning scheme, i.e., rigid fit of the compounds, but need 'Best Fit'

superimpositioning, which involves internal optimisation of the compound in torsion angle space as well, to fit the original hypothesis. For this data set there exist two distinct possibilities during the simplex optimisation procedure:

| SQEP  |  | 17 Compounds   |  | Page 1   |
|---|--|--|--|--|
|    |  |   |  |    |
| cpd1205kcal   |  | NB-5985kcal  |  | cpd1215kcal  |
|    |  |   |  |     |
| cpd785kcal  |  | cpd-465kcal  |  | cpd-375kcal  |
|    |  |   |  |   |
| cpd-1125kcal  |  | cpd765kcal   |  | 2ADHS  |
|  |  |  |  |  |
| cpd-Z-195kcal   |  | cpd-125kcal  |  | SDZ-87-4695kcal  |
|  |  |  |  |  |
| terbinafine5kcal  |  |  |  |  |

Scheme 2. The SQEP data set.

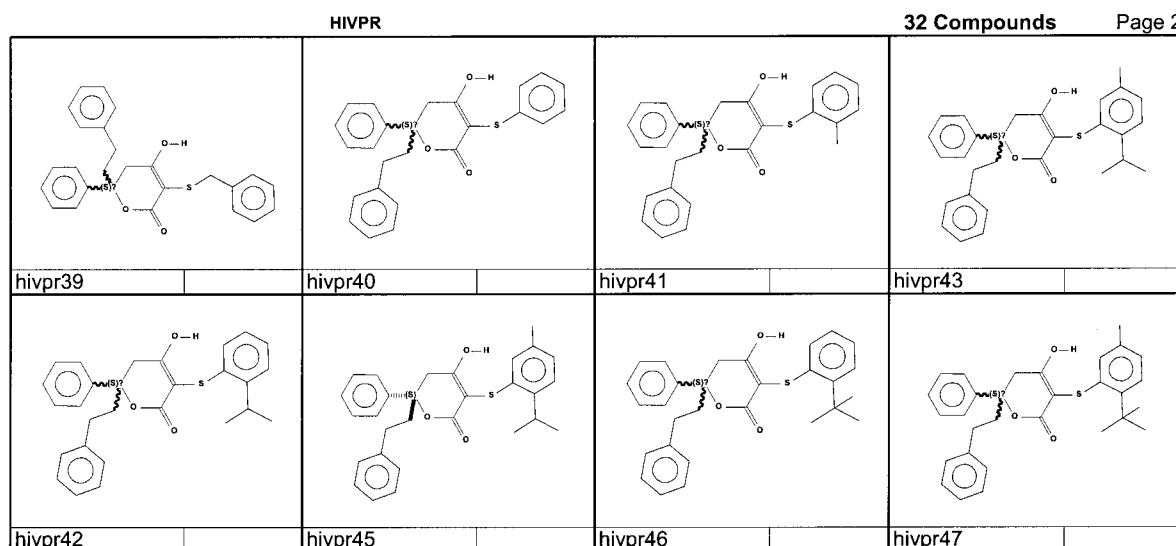
Mode 1. Allow the two compounds SDZ-87-469 and terbinafine with poor activities the possibility not to map to the final hypothesis.

Mode 2. Force SDZ-87-469 and terbinafine to fit to the final hypothesis using the 'Fast Fit' mode.

3. The third data set of some HIV protease inhibitors (HIVPR) [16] (Table 3, Scheme 3) was used to study the effect of 'excluded volumes', i.e., spheres within which compounds cannot have any part present, during hypothesis generation. Presently, Catalyst cannot treat 'excluded volumes' during the actual hypothesis generation al-

|         |         |         |         |
|---------|---------|---------|---------|
|         |         |         |         |
| hivpr2  | hivpr16 | hivpr17 | hivpr18 |
|         |         |         |         |
| hivpr19 | hivpr20 | hivpr21 | hivpr22 |
|         |         |         |         |
| hivpr23 | hivpr24 | hivpr25 | hivpr27 |
|         |         |         |         |
| hivpr28 | hivpr26 | hivpr29 | hivpr30 |
|         |         |         |         |
| hivpr31 | hivpr32 | hivpr33 | hivpr34 |
|         |         |         |         |
| hivpr35 | hivpr36 | hivpr37 | hivpr38 |

Scheme 3. The HIVPR data set.



Scheme 3. (continued).

Table 1. Experimental and predicted activities for the ACE data set

| Compound name | Experimental pIC50 | Cross-validated predicted pIC50 |       |
|---------------|--------------------|---------------------------------|-------|
|               |                    | FESS                            | PRESS |
| nleu-ala-pro  | -2.85              | -5.93                           | -1.72 |
| val-trp       | -3.23              | -5.00                           | -4.76 |
| leu-ala-pro   | -3.36              | -3.84                           | -3.86 |
| ile-tyr       | -3.57              | -5.21                           | -5.21 |
| phe-ala-pro   | -3.62              | -4.45                           | -2.67 |
| arg-ala-pro   | -4.20              | -4.28                           | -4.61 |
| phe-pro-pro   | -4.89              | -4.09                           | -4.09 |
| ile-pro       | -5.18              | -4.98                           | -4.76 |
| ala-pro       | -5.43              | -6.08                           | -6.49 |
| ala-val       | -5.48              | -5.92                           | -6.01 |
| glu-ala-pro   | -5.56              | -4.87                           | -4.86 |
| val-pro       | -5.62              | -4.95                           | -4.90 |
| gly-phe       | -5.65              | -5.94                           | -5.81 |
| ala-leu       | -6.20              | -5.61                           | -5.56 |
| ala-gly       | -6.40              | -7.02                           | -7.10 |
| gly-glu       | -6.73              | -6.17                           | -6.19 |
| gly-lys       | -6.73              | -6.91                           | -7.08 |
| pro-pro       | -6.88              | -5.86                           | -5.76 |
| ala-his       | -6.95              | -7.01                           | -7.06 |
| gly-asp       | -6.96              | -6.89                           | -6.91 |

Table 2. Experimental and predicted activities for the SQEP data set

| Compound name    | Experimental pIC50 | Cross-validated predicted pIC50 |       |                    |
|------------------|--------------------|---------------------------------|-------|--------------------|
|                  |                    | FESS                            | PRESS | PRESS <sup>a</sup> |
| cpd1205kcal      | -0.43              | -4.27                           | -1.43 | -2.39              |
| NB-5985kcal      | -0.64              | -0.77                           | -0.84 | -1.46              |
| cpd1195kcal      | -0.85              | -1.35                           | -1.17 | -2.15              |
| cpd1215kcal      | -1.41              | -1.12                           | -1.02 | -1.61              |
| cpd785kcal       | -1.80              | -2.23                           | -2.23 | -2.56              |
| cpd465kcal       | -2.18              | -1.35                           | -1.40 | -1.75              |
| cpd-E-195kcal    | -2.40              | -2.93                           | -2.81 | -3.49              |
| cpd-375kcal      | -2.43              | -2.19                           | -2.16 | -2.23              |
| cpd-1125kcal     | -2.84              | -4.04                           | -4.78 | -3.29              |
| cpd765kcal       | -3.08              | -2.75                           | -2.35 | -1.78              |
| cpd-345kcal      | -3.23              | -2.09                           | -1.99 | -2.51              |
| 2ADHS            | -3.38              | -3.78                           | -4.98 | -5.67              |
| cpd-Z-195kcal    | -3.49              | -3.31                           | -3.37 | -2.81              |
| cpd-125kcal      | -3.78              | -2.95                           | -2.93 | -2.85              |
| cpd-425kcal      | -3.98              | -3.46                           | -3.43 | -4.48              |
| SDZ-87-4695kcal  | -4.63              | -4.98                           | -4.98 | -4.55              |
| terbinafine5kcal | -4.97              | -4.64                           | -4.78 | -5.74              |

<sup>a</sup>All compounds, including SDZ-87-4695 and terbinafine, were forced to map to the hypothesis.

though they may later be added manually. Compound 45 was docked into the active site according to the description in Reference 16 since the X-ray

structure of A-74704 complexed to HIV-1 protease (9HVP) [17] is known (The H<sub>2</sub>O301 overlaps the lactone of 45. The 6-phenyl of 45 overlaps with the phenyl substituent of A-74704 in P1. The o-isopropyl group on the 3-phenylthio group of 45

Table 3. Experimental and predicted activities for the HIVPR data set

| Compound name | Experimental pIC50 | Cross-validated pIC50 |                   |                  |                   |                  |                   |
|---------------|--------------------|-----------------------|-------------------|------------------|-------------------|------------------|-------------------|
|               |                    | Tr. <sup>a</sup>      | Test <sup>a</sup> | Tr. <sup>b</sup> | Test <sup>b</sup> | Tr. <sup>c</sup> | Test <sup>c</sup> |
| hivpr16       | −0.23              | 0.12                  |                   | 0.94             |                   | 0.53             |                   |
| hivpr17       | −0.11              | 0.09                  |                   | 0.93             |                   | 0.05             |                   |
| hivpr20       | 0.38               | 0.41                  |                   | 0.92             |                   | 0.88             |                   |
| hivpr21       | 1.43               | 1.69                  |                   | 0.92             |                   | 1.35             |                   |
| hivpr24       | 0.11               | 0.55                  |                   | 0.92             |                   | 0.36             |                   |
| hivpr26       | 0.32               | 0.23                  |                   | 0.92             |                   | 0.20             |                   |
| hivpr28       | 1.08               | 1.77                  |                   | 0.93             |                   | 0.80             |                   |
| hivpr33       | −0.32              | 0.02                  |                   | 0.97             |                   | 0.37             |                   |
| hivpr34       | −0.61              | 0.22                  |                   | 0.92             |                   | 0.18             |                   |
| hivpr36       | 1.02               | 0.30                  |                   | 0.93             |                   | 0.49             |                   |
| hivpr37       | 2.30               | 1.75                  |                   | 1.00             |                   | 1.23             |                   |
| hivpr38       | 0.59               | 0.29                  |                   | 0.93             |                   | 0.27             |                   |
| hivpr39       | 1.22               | 0.21                  |                   | 0.93             |                   | 1.22             |                   |
| hivpr40       | 0.89               | 1.37                  |                   | 0.86             |                   | 0.78             |                   |
| hivpr41       | 1.14               | 1.38                  |                   | 0.92             |                   | 1.17             |                   |
| hivpr42       | 1.85               | 1.75                  |                   | 0.92             |                   | 1.99             |                   |
| hivpr43       | 2.14               | 1.69                  |                   | 0.92             |                   | 2.59             |                   |
| hivpr46       | 2.44               | 2.52                  |                   | 0.92             |                   | 2.51             |                   |
| hivpr47       | 2.01               | 1.25                  |                   | 0.92             |                   | 0.66             |                   |
| hivpr2        | −0.48              |                       | 0.09              |                  | 0.93              |                  | −0.38             |
| hivpr18       | 0.28               |                       | −0.15             |                  | 0.96              |                  | 0.92              |
| hivpr19       | 0.80               |                       | 0.13              |                  | 0.97              |                  | 0.53              |
| hivpr22       | 1.77               |                       | 2.12              |                  | 0.92              |                  | 2.53              |
| hivpr23       | 2.15               |                       | 2.23              |                  | 0.92              |                  | 3.09              |
| hivpr25       | 0.39               |                       | 0.16              |                  | 0.96              |                  | 0.99              |
| hivpr27       | 0.59               |                       | 0.46              |                  | 0.93              |                  | 0.08              |
| hivpr29       | 0.82               |                       | 0.25              |                  | 0.96              |                  | 1.03              |
| hivpr30       | 1.16               |                       | 0.16              |                  | 0.93              |                  | 0.86              |
| hivpr31       | 1.24               |                       | 0.28              |                  | 0.95              |                  | 1.34              |
| hivpr32       | 0.64               |                       | 0.30              |                  | 0.92              |                  | 0.44              |
| hivpr35       | 1.08               |                       | 0.35              |                  | 0.93              |                  | 0.99              |

Training set (Tr) and test set (Test):

<sup>a</sup>No excluded volumes with citest optimisation.<sup>b</sup>Excluded volumes with no citest optimisation.<sup>c</sup>Excluded volumes with citest optimisation.

occupies P1', while the methyl *para* to the thio group reaches into P2'. These groups overlap with the P1' and P2' substituents of A-74704: see Figure 1). Compound 46 was then aligned onto the corresponding substructure of 45 (Figure 2). The heavy atoms, i.e., excluding hydrogens, of 46 docked in the active site were superimposed onto the corresponding atoms of 46 mapped to the original hypothesis (Figure 3). The positions of the protein residue atoms of the 'active site' (a sphere of 10 Å surrounding compound 46) of 9HVP were used as locations for the Catalyst 'excluded vol-

umes' during simplex refinement of the original hypothesis (see Original hypothesis section for a description of the 'original hypothesis'). Thus 578 'excluded volumes' with fixed locations in Cartesian space and with a radius of 0.5 Å were added to the original hypothesis (see Figure 4). The HIVPR data set was divided into a training set and a test set of 19 compounds and 12 compounds, respectively (see Table 3).

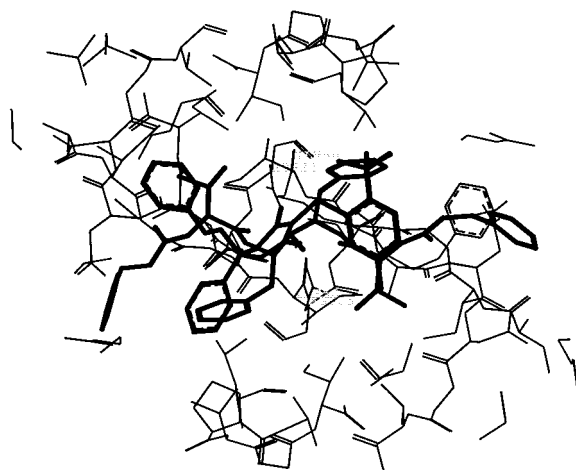


Figure 1. Superimpositioning of compound 45 from the HIVPR data set and A-74704.

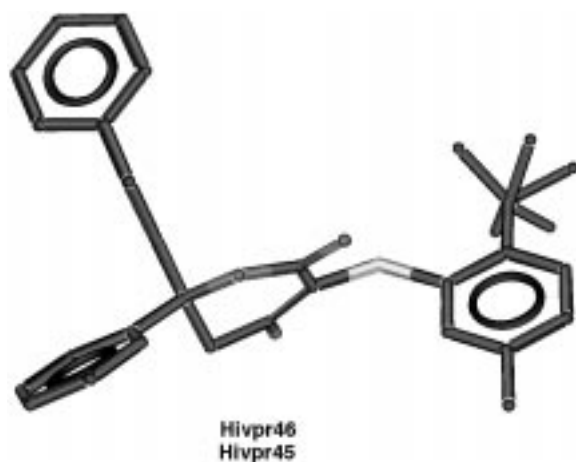


Figure 2. Superimpositioning of compounds 45 and 46 from the HIVPR data set.

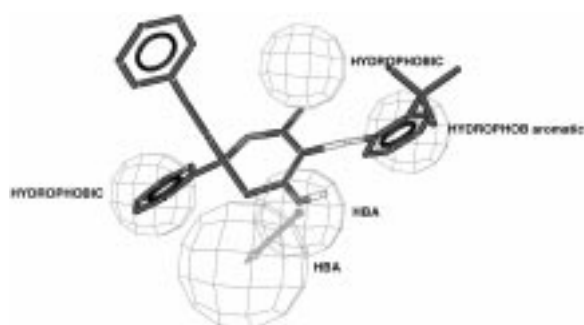


Figure 3. Compound 46 from the HIVPR data set mapped onto the original (starting) hypothesis.

## Methods of calculation

### Conformational analysis

The conformational models provided with the Catalyst software were used for the ACE and SQEP data sets without further alteration [12, 13, 15]. The creation of conformational models for every compound of the HIVPR data set was performed using 'Best' conformer generation. An energy range of 20 kcal mol<sup>-1</sup> above the global minimum was set as a cut-off.

### Fitting of compounds to the hypothesis

The actual fitting of each compound to a hypothesis was performed by the MSI software 'citest' [9]. This is a stand alone program which uses the same mapping procedure as is used in Catalyst. The citest program reports back, among various other statistics, how well the compound matches the proposed hypothesis in terms of a series of regular Catalyst 'Fit' scores. An advantage of citest is the possibility to specify a maximum number of features in a proposed hypothesis that do not need to be mapped by the compound in question (the omit option).

### Simplex optimisation

The Catalyst hypotheses were refined using simplex optimisation [18] included into the program HypoOpt [19]. HypoOpt generates the hypothesis, runs the 'citest' program and evaluates the outcome from citest in terms of the minimisation function presently used, i.e., the FESS or PRESS function (see below).

All features in the Catalyst hypothesis have equal weights (set to 1.0). A relationship between log(activities) and the corresponding Fit-values for all compounds in the training set after mapping of each compound to the hypothesis is computed using linear regression.

In this study two different minimisation functions have been employed. The value for the simplex minimisation function is then computed using either of two schemes:

1. The sum of squares between the calculated (fitted) activity  $\{-\log(\text{act}_{\text{fit}})\}$  and the experimental activity  $\{-\log(\text{act}_{\text{exp}})\}$  for all compounds (FESS) of the training set, where all compounds were used in the linear regression analysis:

$$\text{FESS} = \sum \{-\log[\text{act}_{\text{fit}}(i)] + \log[\text{act}_{\text{exp}}(i)]\}^2 \quad (1)$$

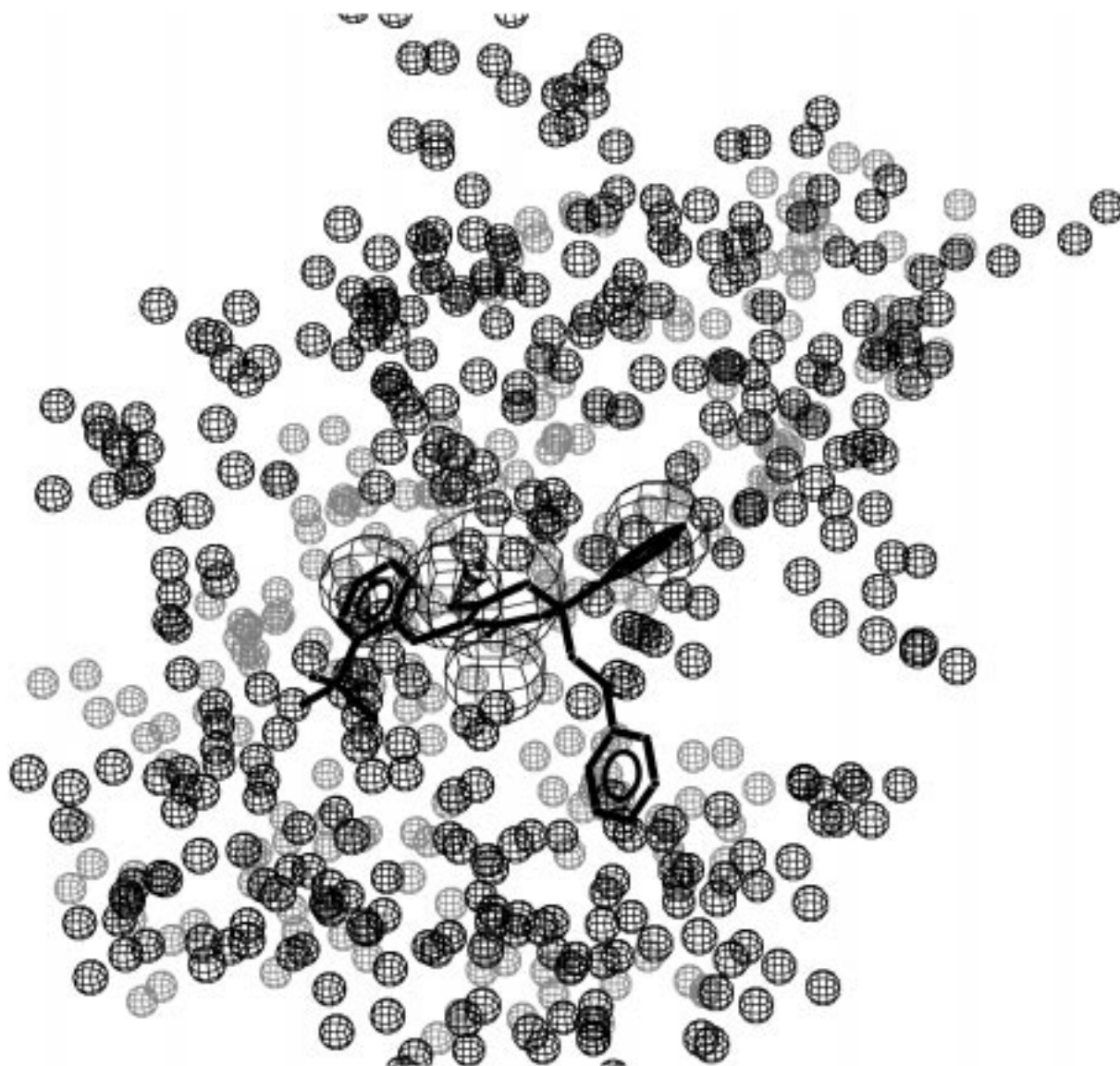


Figure 4. Addition of active site 'excluded volumes' to the HIVPR original hypothesis.

2. The sum of squares between the predicted activity  $\{-\log(\text{act}_{\text{pred}})\}$  and the experimental activity  $\{-\log(\text{act}_{\text{exp}})\}$  for all compounds (PRESS) of the training set. The predicted activity of each compound is computed from a linear regression model based on the remaining training set compounds in a leave-one-out (LOO) cross-validation (CV) manner [14]:

$$\text{PRESS} = \sum \{-\log[\text{act}_{\text{pred}}(i)] + \log[\text{act}_{\text{exp}}(i)]\}^2 \quad (2)$$

Provisions were also taken so that the slope of the derived regression model could not assume negative

values. In these cases a high value (set to 1000) was used as outcome from the simplex evaluation to force the minimisation routine away from such areas on the function surface. The maximum number of simplex evaluations was set to 200. The convergence criterion (tolerance) was defined as:

$$\text{tolerance} = 2.0 * |(Y(\text{HIGH}) - Y(\text{LOW}))| / (|Y(\text{HIGH})| + |Y(\text{LOW})|) \quad (3)$$

$$+ |Y(\text{LOW})| \quad (4)$$

where Y(HIGH) and Y(LOW) were the highest and lowest retained property values, e.g., FESS or PRESS



values depending on which formalism was used, for the training set, respectively. The minimisation was stopped if the tolerance reached values below 0.05. The variables to be optimised were the positions of the features in a hypothesis. Thus, all hypothesis features were translated in the x, y, and z directions of a Cartesian coordinate system. The features consisting of two points, e.g., hydrogen bond donor and acceptor functions as well as aromatic functions, were also rotated around the three axes (x, y and z) of the coordinate system in internal space keeping the internal distance between the two points fixed (pre-set from the start to the default intra-feature distance). The starting set-up for the simplex optimisation (N variables to be optimised need N+1 starting simplex points) was created using a fractional factorial design [20] protocol where each feature was moved 0.5 Å and each two-point feature was rotated 10 deg with respect to the original hypothesis (see below for a definition). The original hypothesis was used as the first starting point.

#### *Original hypothesis*

In principle any reasonable hypothesis can be used as starting point for the optimisation, although the choice of original hypothesis will most likely influence the outcome of the simplex optimisation procedure. In this work the original hypotheses of the three investigated data sets are the results from hypothesis generation using the Catalyst software. The original hypotheses for the ACE (ace1.6–1.7.chm) and SQEP (Hypo7functions.chm) data sets were taken from the MSI applications supplied as part of Catalyst [13] while the corresponding hypothesis for the HIVPR data set (Hivpr.5.chm) was generated using Catalyst version 3.0 without inclusion of excluded volumes. The original hypotheses are listed in Table 4.

#### *Mapping of a compound to a hypothesis*

When a compound is mapped to a hypothesis the Catalyst software (or citest) searches for the optimum set of 3D alignments between features in the hypothesis and matching features in the molecule (the Fit). A number of previously calculated conformers within a specified energy threshold are used as possible candidates for each mapping. Up to one hundred different mappings are presented ranked in decreasing order of quality based upon the computed Fit value (score).

## **Results and discussion**

### *ACE data set*

The CV investigation (four validation groups) of the ACE data set using the FESS function resulted in four models with significantly improved internal statistics compared with the corresponding results from the non-optimised hypothesis (see Table 5). More importantly, the external predictive abilities of the CV test sets were not equally impressive. Compounds nleu-ala-pro, val-trp and ile-tyr were severely underestimated by the respective models with deviations of 3.08, 1.77, 1.64, respectively. However, the two latter compounds are underestimated in all the developed hypotheses including the original one from Sprague [13]. An interesting change was noted in the CV study when the function to be minimised was changed from FESS to PRESS. Compounds val-trp and ile-tyr were still poorly predicted but the activity of the most active compound nleu-ala-pro was now better (over-)estimated (1.53, 1.64, −1.13, respectively). Also the rest of the compounds were predicted with higher accuracy compared to the outcome of the FESS CV investigation. The root mean squared error (RMSE) for the predicted activities of the compounds decreased from 1.02 in the FESS study to 0.82 in the PRESS investigation. The corresponding mean RMSE values for the training sets were 0.32 and 0.36, respectively. Again this demonstrates the well known balance between the fit of a model and its predictive ability. This means that the FESS based model is overfitted while the PRESS based model is more balanced with respect to fit and predictive ability. The corresponding investigations using the Catalyst software (version 3.0 or 3.1) are presently not possible because the program only supports a maximum use of five features during hypothesis generation. Thus, merged features, such as the one used in this study, can only be regressed, i.e., the Fit-values correlated to the activities using linear regression analysis, but not optimised. The original and optimised hypotheses are depicted in Figure 5.

### *SQEP data set*

The SQEP hypothesis is also the result of a merger between hypotheses. This hypothesis consists of seven features and, as was the case for the ACE hypothesis, cannot presently be optimised with Catalyst. The SQEP data set was divided into four CV groups and the predictivity was evaluated using both the FESS and

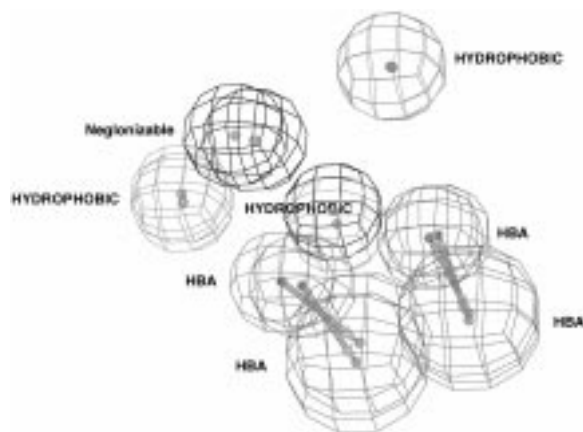


Figure 5. Original and optimised hypotheses for the ACE data set.

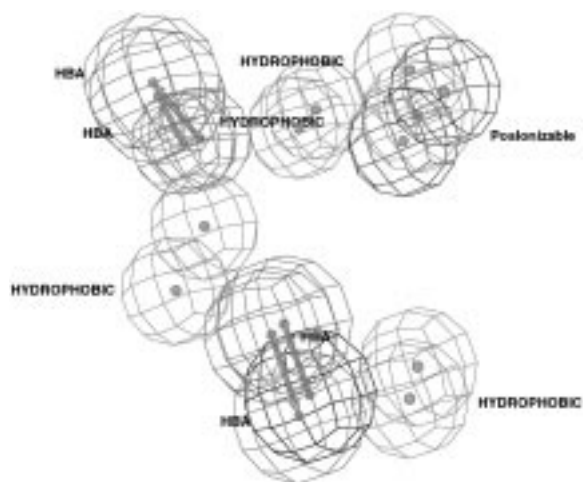


Figure 6. Optimised hypotheses using mode 1 and 2, respectively, for the SQEP data set.

the PRESS minimization functions. Again, as was the case for the ACE data set, the external predictivity of the test set for the CV models (mode 1; using 'Fast Fit' superimpositioning) was better when using the PRESS function during the simplex optimisation compared with the corresponding models derived with the FESS function. Apart from the most active compound, which is underestimated in both schemes, although less by the PRESS function based models, the predictive abilities of the models developed using the PRESS function are clearly superior to the corresponding models using the FESS function (see Table 5). The corresponding CV models, where SDZ-87-469 and terbinafine were forced to fit the hypothesis (mode 2), using 'Fast Fit' superimpositioning, showed somewhat decreased predictive ability although the overall performance of the models was good. Interesting to note

is that by optimising the hypothesis one can, in this case, obtain a model (Table 5: SQEP press, fitall) with equal predictive ability to which all compounds fit compared with the first model (Table 5: SQEP press) to which SDZ-87-469 and terbinafine did not fit. The fact that the two models have comparable statistics is probably more of a coincidence but, more interestingly, it points out the possibility to have a consistent procedure of fit in a model if one so desires. The two models are not dramatically different although a certain difference of the locations of the features can be noted (Figure 6).

#### *HIVPR data set with inclusion of excluded volumes*

The special emphasis of this investigation compared with the previous two examples was the use of 'excluded volumes' during the actual hypothesis generation. The PRESS minimisation function was used. The use of 'excluded volumes' did not improve the training set model but a slight improvement was found for the predictive ability of the test set. However, to be noted is the fact that it is not enough to merely position the 'excluded volumes' of an active site or surface onto the presently used hypothesis. This may result, as seen in this example (Table 5), in a non-existing training set model ( $r^2 = 0.00$ , Figure 7) compared to the corresponding model with refinement ( $r^2 = 0.64$ , Figure 8). This behaviour results from the superimpositioning procedure where unfavourable interactions (bumps) occur when the atoms of the active site are used as positions for the 'excluded' volumes. Thus, in order to obtain a good model having 'excluded volumes' located close to the compounds one needs to optimise the hypothesis to relax the bumps that exist. This situation differs from the one where (a few) 'excluded volumes' are placed further/far away from the compounds in order to eliminate some unreasonable modes of alignment of one or more of the compounds under investigation. In this latter case the 'excluded volumes' do not interact with the compounds as closely as in the case where the 'excluded volumes' are used as an active site surface for closer volume and alignment constraint purposes. Although the use of 'excluded volumes' in this example did not improve the statistics and predictive ability of the model drastically compared to the hypothesis without 'excluded volumes', the former (expanded) one still has significant advantages when performing 3D database searches. The 'expanded' hypothesis may act as a useful filter to limit the number of hits and to

Table 4. Catalyst features contained in the original hypotheses

| Data set | Hypothesis name    | Features <sup>a</sup>                      |
|----------|--------------------|--|
| ACE      | ace1.6-1.7.chm     | 2 HBA, 3 HYDROPHOBIC, 1 NegIonizable       |
| SQEP     | Hypo7functions.chm | 2 HBA, 1 PosIonizable, 4 HYDROPHOBIC       |
| HIVPR    | Hivpr.5.chm        | 1 HBA, 1 HYDROPHOB_aromatic, 2 HYDROPHOBIC |

<sup>a</sup>HBA = hydrogen bond acceptor. NegIonizable = negative ionizable group, e.g., a carboxylic acid. PosIonizable = positive ionizable group, e.g., an amine. HYDROPHOBIC = hydrophobic center. HYDROPHOB\_aromatic = hydrophobic center consisting of only aromatic atoms.

Table 5. Statistics for the HypoOpt refined Catalyst hypotheses of the three data sets<sup>a</sup>

| Data set | Method | RMSE           |      |        |      |                          | Comment                                    |
|----------|--------|----------------|------|--------|------|--------------------------|--|
|          |        | R <sup>2</sup> | S    | F      | Tr.  | Test                     |  |
| ACE      | FESS   | 0.93           | 0.36 | 180.80 | 0.33 |                          |  |
|          |        | 0.95           | 0.32 | 246.71 | 0.30 |                          |  |
|          |        | 0.95           | 0.32 | 251.67 | 0.30 |                          |  |
|          |        | 0.94           | 0.36 | 197.00 | 0.34 | 1.02 (0.77) <sup>b</sup> |  |
|          | PRESS  | 0.88           | 0.48 | 95.25  | 0.44 |                          |  |
|          |        | 0.95           | 0.34 | 225.51 | 0.31 |                          |  |
|          |        | 0.94           | 0.37 | 194.83 | 0.34 |                          |  |
|          |        | 0.92           | 0.40 | 158.42 | 0.37 | 0.82 (0.80) <sup>b</sup> |  |
|          | SQEP   | 0.89           | 0.43 | 81.85  | 0.39 |                          |  |
|          |        | 0.98           | 0.20 | 549.47 | 0.19 |                          |  |
|          |        | 0.93           | 0.39 | 139.99 | 0.36 |                          |  |
|          |        | 0.95           | 0.32 | 210.78 | 0.30 | 1.10 (0.60) <sup>c</sup> |  |
| HIVPR    | PRESS  | 0.85           | 0.51 | 56.65  | 0.46 |                          |  |
|          |        | 0.97           | 0.24 | 386.38 | 0.22 |                          |  |
|          |        | 0.91           | 0.44 | 111.45 | 0.40 |                          |  |
|          |        | 0.91           | 0.42 | 117.82 | 0.39 | 0.84 (0.52) <sup>c</sup> |  |
|          | PRESS  | 0.81           | 0.58 | 41.30  | 0.53 |                          | fitall                                     |
|          |        | 0.93           | 0.38 | 147.23 | 0.35 |                          | fitall                                     |
|          |        | 0.84           | 0.58 | 59.41  | 0.53 |                          | fitall                                     |
|          |        | 0.87           | 0.51 | 76.82  | 0.47 | 1.03                     | fitall                                     |
|          | PRESS  | 0.67           | 0.56 | 34.13  | 0.53 | 0.66                     | No excluded volumes<br>Citest optimisation |
|          |        | 0.00           | 0.97 | 0.01   | 0.91 | 0.67                     | Excluded volumes<br>No citest optimisation |
|          |        | 0.64           | 0.58 | 30.25  | 0.55 | 0.48                     | Excluded volumes<br>Citest optimisation    |

<sup>a</sup>R<sup>2</sup>: ordinary correlation coefficient; S: standard deviation; F: ordinary F-value; Root mean squared error (RMSE) for the training set (Tr) models and the overall RMSE for the test set (Test). See text for more details.

<sup>b</sup>Compound nleu-ala-pro excluded.

<sup>c</sup>Compound cpd1205kcal excluded.

identify structures that may actually fit into the active site of the target receptor or enzyme. How the 'excluded volumes' are positioned and the exact number of 'excluded volumes' used depends on the target

investigated, the dynamics of the target, the training set employed and the overall objectives of the investigation.

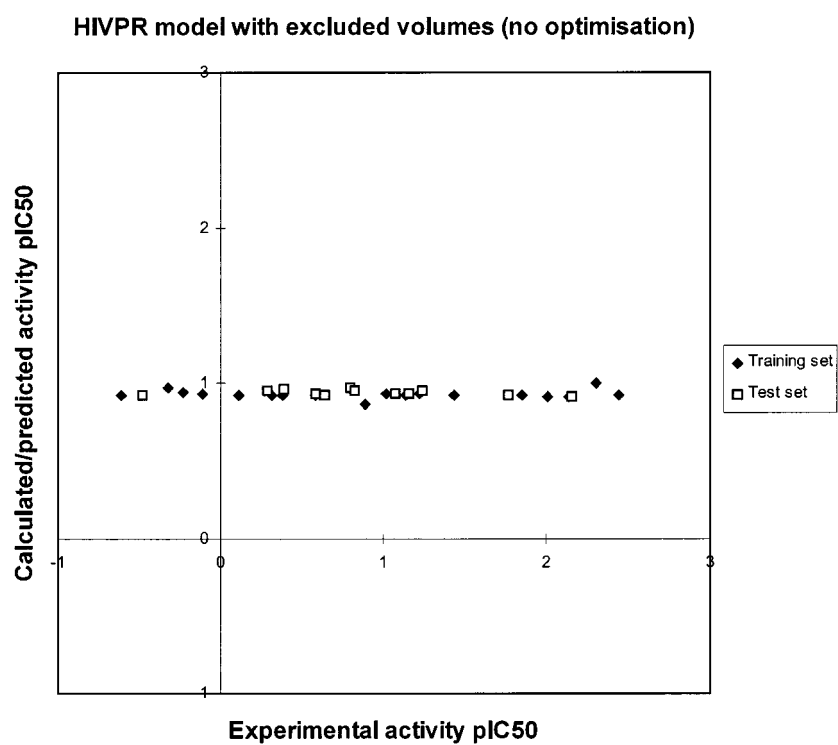


Figure 7. HIVPR model with excluded volumes and without citest optimisation.

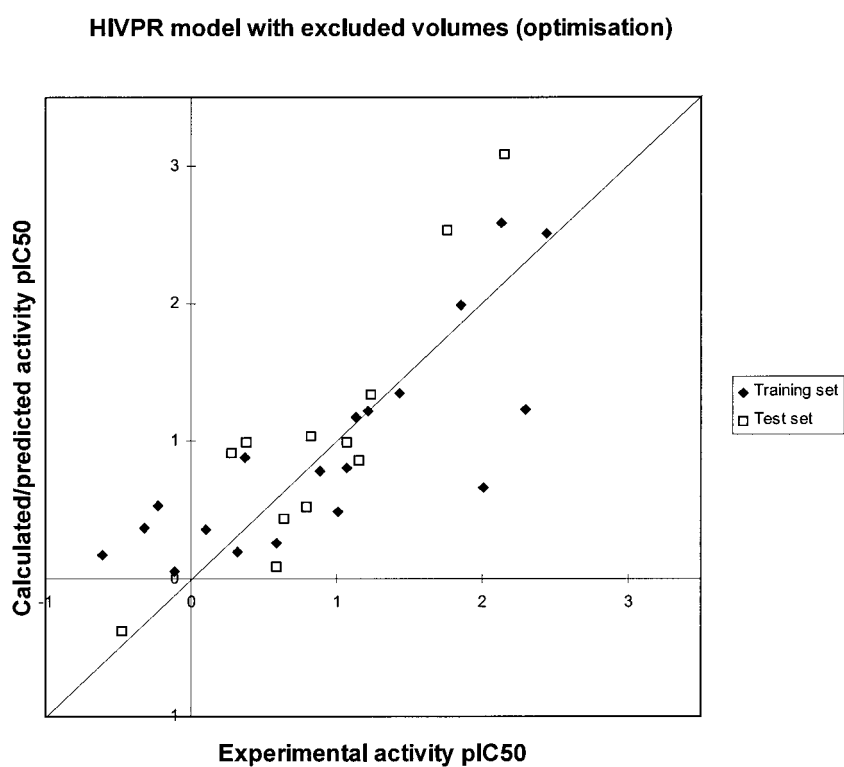


Figure 8. HIVPR model with excluded volumes and with citest optimisation.

## Conclusions

The three investigated data sets in this work have pointed at various interesting possibilities for refining Catalyst hypotheses using simplex optimisation with the possible inclusion of 'excluded volumes' during the refinement procedure. Thus, merged Catalyst hypotheses may be optimised and forced to use the same mapping procedure throughout the data set such that all compounds, even the ones with very low activities, actually map, at least partially, to a hypothesis. Cross-validation of the model through the minimisation function in the simplex optimiser can easily be accommodated into the refinement procedure. The cross-validation based minimisation function (PRESS results) performs at least as well as the corresponding function based on fit (FESS results) rather than internal predictive ability of the training set and most of the times better than the latter function when assessing the derived models by predictive ability of an external test set.

## Future expansions

The present protocols have not attempted to investigate the use of a differentiated and optimised weighting scheme for the features of a hypothesis, nor the optimisation of the radii of these features. One may also envisage that the positions of the 'excluded volumes' could be allowed to move, perhaps within certain restricted domains, during the refinement procedure to allow for a dynamic behaviour of the receptor. Hypothesis refinements using a flexible (Flex Fit) mapping of the compounds, allowing internal optimisation of the structures, is also a possibility that may prove productive in order to obtain a statistically better and more predictive hypothesis.

## References

1. Doweyko, A.M., *J. Med. Chem.*, 31 (1988) 1396.
2. Ghose, A., Crippen, G., Revankar, G., McKernan, P., Smee, D. and Robbins, R., *J. Med. Chem.*, 32 (1989) 746.
3. Cramer, R.D., Patterson, D.E. and Bunce, J.D., *J. Am. Chem. Soc.*, 110 (1988) 5959.
4. Norinder, U., *J. Comput.-Aided Mol. Design*, 7 (1993) 671.
5. Floersheim, P., Nozulak, J. and Weber, J., In Wermuth, C.G. (Ed.) *Trends in QSAR and Molecular Modelling 92* (Proceedings of the 9th European Symposium on Structure-Activity Relationships: QSAR and Molecular Modelling), ESCOM, Leiden, 1993, pp. 227–232.
6. Kubinyi, H. (Ed.) *3D QSAR in Drug Design: Theory, Methods and Applications*, ESCOM, Leiden, The Netherlands, 1993.
7. Klebe, G., Abraham, U. and Mietzner, T., *J. Med. Chem.*, 37 (1994) 4130.
8. Silverman, B.D. and Platt, D.E., *J. Med. Chem.*, 39 (1996) 2129.
9. Molecular Simulations Inc., San Diego, CA, USA.
10. [http://www.msi.com/science/online/references/rdd\\_msi\\_pubs.html](http://www.msi.com/science/online/references/rdd_msi_pubs.html)
11. [http://www.msi.com/science/online/references/catalyst\\_pubs.html](http://www.msi.com/science/online/references/catalyst_pubs.html)
12. Supplied as Application Note data as part of the Catalyst software.
13. Sprague, P.W., Building a Hypothesis for Angiotensin Converting Enzyme Inhibition, MSI application note.
14. Wold, S., *Technometrics*, 20 (1979) 379.
15. Hoffmann, R.D. and Sprague, P.W., Building a Hypothesis for Competitive Inhibition of Rat Liver Squalene Epoxidase, MSI application note.
16. Tummino, P.J., Prasad, J.V.N.V., Ferguson, D., Nouhan, C., Graham, N., Domagala, J.M., Ellsworth, E., Gajda, C., Hagen, S.E., Lunney, E.A., Para, K.S., Tait, B.D., Pavlovsky, A., Erickson, J.W., Gracheck, S., McQuade, T.J. and Hupe, D.J., *Bioorg. Med. Chem.*, 4 (1996) 1401.
17. Erickson, J.W., Neidhart, D.J., VanDrie, J., Kempf, D.J., Wang, X.C., Norbeck, D.W., Plattner, J.J., Rittenhouse, J.W., Turon, M., Wideburg, N., Kohlbrenner, W.E., Simmer, R., Helfrich, R., Paul, D.A. and Knigge, M., *Science*, 249 (1990) 527.
18. Press, W.H., Teukolsky, S.A., Vetterling, W.T. and Flannery, B.P., *Numerical Recipes in FORTRAN: The art of Scientific Computing*, 2nd ed., Cambridge University Press, Cambridge, MA, 1992, pp. 423–436.
19. Norinder, U., in-house software.
20. Box, G.E.P., Hunter, W.G. and Hunter J.S., *Statistics for Experimenters*, Wiley, New York, NY, 1978.