Available online at www.sciencedirect.com
SCIENCE DIRECT
Talanta 68 (2005) 54–60
Talanta
www.elsevier.com/locate/talanta

Improving QSAR models for the biological activity of HIV Reverse Transcriptase inhibitors: Aspects of outlier detection and uninformative variable elimination

M. Daszykowski [a,b], I. Stanimirova [a], B. Walczak [a,b,*], F. Daeyaert [c], M.R. de Jonge [c], J. Heeres [c], L.M.H. Koymans [c], P.J. Lewi [c], H.M. Vinkers [c], P.A. Janssen [c,✠], D.L. Massart [a]

[a] *FABI, ChemoAC, Vrije Universiteit Brussel, Laarbeeklaan 103, B-1090 Brussels, Belgium*
[b] *Department of Chemometrics, Institute of Chemistry, The University of Silesia, 9 Szkolna Street, 40-006 Katowice, Poland*
[c] *Center for Molecular Design, Janssen Pharmaceutica N.V., Antwerpsesteenweg 37, B-2350, Vosselaar, Belgium*

Received 9 September 2004; received in revised form 4 March 2005; accepted 26 April 2005
Available online 17 June 2005

Abstract

The goal of this study is to derive a methodology for modeling the biological activity of non-nucleoside HIV Reverse Transcriptase (RT) inhibitors. The difficulties that were encountered during the modeling attempts are discussed, together with their origin and solutions. With the selected multivariate techniques: robust principal component analysis, partial least squares, robust partial least squares and uninformative variable elimination partial least squares, it is possible to explore and to model the contaminated data satisfactory. It is shown that these techniques are versatile and valuable tools in modeling and exploring biochemical data.

Keywords: Robust calibration; Feature selection; Multivariate calibration; QSAR

1. Introduction

There are several proven drug targets for the treatment of HIV infection. One of them is Reverse Transcriptase (RT), which plays a vital role in the HIV life cycle. RT is responsible for transcription of the viral single-stranded RNA into double-stranded DNA. After integration of the double-stranded DNA into the host cell genome, the infected cell can perform the synthesis of peptides necessary for HIV replication. One strategy to stop HIV replication is to inhibit RT with small molecules (inhibitors) belonging to the class of non-nucleotide RT inhibitors. These inhibit the RT enzyme by binding to an allosteric binding site.

Over several years, many publications concerning inhibition of Reverse Transcriptase and modelling biological activity of inhibitors were published. The more recent publications present QSAR (quantitative structure–activity relationships) models relating biological activity of inhibitors and their structure, described by various descriptors [1–4]. Here, our efforts to develop QSAR models of NNRTIs, using the calculated interaction energies between a set of 202 inhibitors and the amino acids lining the NNRTI binding site as descriptors, are reported. The interaction energies are obtained after docking the inhibitors in the NNRTI binding site using a pharmacophore-based docking algorithm [5]. The inhibitors are docked into seven instances of the HIV-RT binding site, obtained from the complexes of the enzyme with seven different NNRTIs of which the structure has been determined using X-ray crystallography. After docking, the interaction energies of the inhibitors with the amino acids in the NNRT binding site are computed using molecular

* Corresponding author. Tel.: +48 32 359 12 46; fax: +48 32 259 99 78.
E-mail address: beata@us.edu.pl (B. Walczak).
✠ Dr. Paul A.J. Janssen, founder of Janssen Pharmaceutica and mentor of the Center for Molecular Design, passed away on November 11, 2003.

0039-9140/$ – see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.talanta.2005.04.071

mechanics [6]. These interactions are splitted into the interactions with the side chain and backbone moieties of the different amino acids, and into the Coulomb, Van der Waals and hydrogen-bond components of these interactions.

The biological activity of the 202 inhibitors to be modeled are the 50% inhibitory concentrations, expressed as pIC_{50} against wild-type HIV and four mutant strains in a cell-based assay [7]. It is assumed that there exist a relation between the interaction energies and the biological activity. The stronger interactions an inhibitor provides, the better its biological activity should be.

The aim of this study is to construct a model for predicting the biological activity of an inhibitor based on the description of its interaction energies. During the modeling, different problems can be encountered, that may require specific approaches or strategies of model improvement. For instance, the outlying observations and quality of the variables are issues that should be taken into account.

2. Theory

2.1. Partial least squares (PLS)

The goal of partial least squares (PLS) is to construct a linear model between a property of interest, called a dependent variable, **y**, and a set of explanatory variables, **X**. PLS minimizes the sum of squared residuals between observed and predicted **y**, by finding a limited set of orthogonal latent factors, **T**, maximizing the covariance between **X** and **y** [8–10]. The PLS model can be presented as:

$$\mathbf{y} = \mathbf{Tq} + \mathbf{e} = \mathbf{Xb} + \mathbf{e} \tag{1}$$

where **q** contains the regression coefficients associated with the latent PLS factors; **T** and **e** ($m \times 1$) are the vectors of errors, **X** is the data matrix ($m \times n$), and **b** contains the regression coefficients related to the n original variables. The regression coefficients are computed as:

$$\mathbf{b} = \mathbf{W}(\mathbf{P}^{\mathrm{T}}\mathbf{W})^{-1}\mathbf{q} \tag{2}$$

where **W** is the matrix of loadings obtained by maximizing the covariance criterion, and **P** is the product of **X** and **T**.

To ensure a good predictive ability of the PLS model, the calibration set should represent all possible sources of the data variance, and the number of PLS factors should be optimized. In our study, the number of the PLS factors was evaluated by cross-validation [9]. The predictive ability of a model is expressed by the cross-validated root mean square error of prediction (RMSECV):

$$\mathrm{RMSECV} = \sqrt{\frac{\sum_{i=1}^{m}(y_i - \hat{y}_{-i})^2}{m}} \tag{3}$$

where y_i is the experimental value of the dependent variable of the ith object, and $\hat{y}_{-i}$ is the predicted value of the dependent variable of the ith object with the model built without the ith sample.

An alternative to RMSECV is the q^2 statistic:

$$q^2 = 1 - \frac{\sum_{i=1}^{m}(y_i - \hat{y}_{-i})^2}{\sum_{i=1}^{m}(y_i - \bar{y})^2} \tag{4}$$

2.2. Robust partial least squares

PLS as a least squares method is very sensitive to outliers, i.e. objects not following the same model as the majority of the data. The outlying objects in the **y** direction can be relatively easily identified, but the outlying objects in the **X**-space are not necessarily outliers from the calibration model. A distinction between good and bad leverage objects is possible only in the calibration setting. Outliers in calibration cannot be identified by focussing on the residuals from the regression model already influenced by them, but they can be found simply by the residuals from a robust model. In our study, the Evolution Program (EP) [13,14] was used to construct a robust PLS model. The idea of EP relies on an optimization strategy, alike to a Genetic Algorithm [15–17], aiming to find a subset of objects for which the fit and the predictive power of the PLS model is optimal. EP works with a population of solutions, coded as binary strings of length equal to the number of objects in the data set. 1s in a string stand for the model set objects. As an example, let us consider a string with 20 elements: [0 0 0 0 0 1 0 1 1 0 1 0 1 1 0 1 0 1 1 1]. The 1s in positions 6, 8, 9, 11, 13, 14, 16, 18, 19 and 20 means that these objects are used for model construction. This subset of objects is evaluated, based on model fit and predictive ability. For data with outliers, this can be done by calculating the sum of squared residuals for the assumed fraction of non-contaminated data, sorted according to their squared residuals. The highest level of data contamination is 49%, but it can be assumed smaller. The user also declares the highest complexity of the PLS model, $f^{\max}$. For m objects in the data set, given the fraction of contamination p and the maximal complexity of the model $f^{\max}$, each subset of objects selected for the model construction contains m^* objects. The value m^* is within the range:

$$f^{\max} < m^* \ll m(1-p) \tag{5}$$

The model built for the m^* objects is used to predict **y** for all the objects in the data set. The m squared residuals, $(y_i - \hat{y}_i)^2$, are sorted and the sum of the smallest $1-p$ squared residuals is the fitness of the model:

$$\mathrm{Fitness} = \frac{1}{(y_i - \hat{y}_i)^2_{i \in A}} \tag{6}$$

where A is the fraction of the data with the $1-p$ smallest residuals.

This strategy ensures that both the fit and the prediction are taken into account in the model evaluation. For each selected subset of objects, the model complexity is estimated with cross-validation.

The first population of strings is constructed randomly, but the next ones are built based upon the fitness of the strings, or models, in the previous population. The strings are selected for reproduction according to the 'roulette rule' (randomly but with a probability proportional to their fitness). The selected strings are then modified and each string is replaced by a random selection of m^* objects chosen from the $1-p$ objects with the lowest squared residuals in the previous model (specific genetic operator of the EP approach). The algorithm is stopped when a maximal number of generations is reached or when a homogenous population is obtained.

The subset of objects corresponding to the string with the highest fitness is used for the construction of the robust model and the detection of outliers in the calibration according to the robust scale, proposed by Rousseeuw and Leroy [18]. It should be emphasized that the number of identified outliers can be much smaller than the assumed fraction of data contamination. After removal of the outliers, the final PLS model is constructed and evaluated.

2.3. Uninformative variable elimination partial least squares (UVE-PLS)

A PLS calibration model can be much improved by excluding uninformative variables that have high variance but small covariance with the dependent variable **y**. Model improvement means a decrease of the complexity and/or decrease of RMSECV (increase of predictive ability). In our study, the uninformative variable elimination-partial least squares approach (UVE-PLS), proposed by Centner et al. [19] was used. This multivariate approach uses a cut-off value for the PLS coefficients that is determined by adding irrelevant variables to the original data and evaluating the corresponding PLS coefficients. The data matrix **X** ($m \times n$) is augmented with a matrix **N** ($m \times k$) containing random numbers with very small magnitude (of the order 10^{-10}). The number of new variables, k, ought to be higher than 300. These new random variables do not influence the PLS model. The m vectors of regression coefficients, **b**, are calculated with leave-one-out cross validation and saved into the matrix **B** ($m \times n+k$). Its first n columns are the regression coefficients related to the experimental variables, and the k remaining columns are related to the uninformative variables (see Fig. 1).

The stability of the regression coefficient for the jth variable is then defined as:

$$s_j = \frac{\text{mean}\,\mathbf{B}(:, j)}{\text{std}\,\mathbf{B}(:, j)} \tag{7}$$

where mean **B**(:,j) is the mean value of the m elements of the jth column of **B**, and std **B**(:,j) is the standard deviation of the m elements of the jth column of **B**.

The k noisy variables are irrelevant to model **y**, and thus, to discriminate stable and unstable regression coefficients, a cut-off value is defined as:

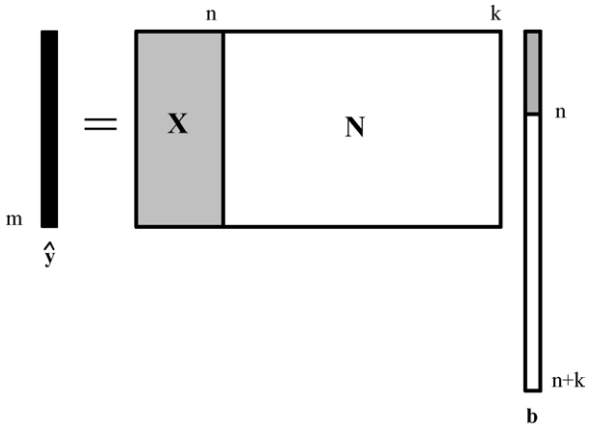

Fig. 1. Graphical representation of the UVE-PLS model.

$$\text{cut-off} = \max\left(\mathbf{s}(n+1:n+k)\right) \tag{8}$$

i.e. the maximal value of the vector $\boldsymbol{s}(n+1{:}n+k)$ containing the stabilities of the regression coefficients associated with the k noise variables.

All of the experimental variables with stability of the regression coefficients below the cut-off value are irrelevant to model **y** and are eliminated from the original data set, because their information content is not higher than the information content of the random variables.

2.4. DATA

The explanatory variables of the HIV data are the calculated interaction energy matrices, obtained by docking 202 non-nucleoside RT inhibitors into seven Reverse Transcriptase crystals, denoted as DTQ, DTT, EET, IKY, RT4, RT5 and TVR, respectively [20–24]. The interaction energy matrices contain the Van der Waals, Coulomb and hydrogen bond interaction energies of the inhibitors with the back bone and side-chain moieties of the individual amino acids comprising the target enzymes. The number and composition of the amino acids in the crystals are not identical, and therefore, the number of interactions (variables) for each table is not the same, but all of the energy tables have 202 rows (NNRTIs). The unfolded energy tables contain in a total of 2304 variables.

The 202 NNRTIs belong to two chemical classes of compounds, known as DATA and DAPY-like inhibitors. The groups are represented by 78 and 124 NNRTIs, respectively. The selected NNRTIs have high inhibitory activities against the wild-type HIV and four mutant strains (181C, 103N, 100I and 188L). The inhibitory activities, expressed as pIC_{50} for each HIV virus vary between 5.5 and 9.5, 4.2 and 9.6, 4.3 and 9.5, 4.3 and 8.9, and 4.3–8.7, respectively.

3. Results and discussion

For modeling the multivariate biochemical data, a property of interest (biological activity of inhibitors), **y**, is predicted based on a set of explanatory variables **X**, i.e. calculated interaction energies between an inhibitor and the amino acids in the RT pocket. The multivariate nature of the HIV data and the high correlation among explanatory variables suggest PLS as a pioneer modeling technique. Our data set consists of five biological activities determined for wild type and four mutant strains of the HIV virus and seven tables of energies, one for every crystal structure of the RT pocket (DTQ, DTT, EET, IKY, RT4, RT5 and TVR). Therefore, different PLS models can be built. For brevity, only the PLS models constructed for the mean biological activity using all crystals simultaneously and the biological activity for each HIV type using all crystals simultaneously are discussed. There are other options, for instance, to predict the mean activity using separate crystals. The PLS technique is not affected by an excess of explanatory variables so PLS can be directly applied to unfolded data, i.e., on a matrix **X**u constructed by joining the energy tables belonging to each of the seven binding pockets.

The PLS model for the mean biological activity of inhibitors does not have a good predictive ability. Its RMSECV and q^2 are 0.6910 and 0.3141, respectively. The same problem is faced for models constructed for individual activities. The RMSECVs and q^2 values of the models vary from 0.6040 to 0.8577 and 0.1672 to 0.3421, respectively (see Table 1).

One of the possible reasons of lack of model's fit could be due to a non-linear relationship between **X** and **y**, and/or the presence of outlying observations in **X** and/or **y**. Outliers in **X** (so-called leverage objects) can occur either due to problems with the docking of an inhibitor into certain crystals or due to a unique interaction of an inhibitor with the RT amino acids. Often, verification of the outlying character of an inhibitor in **X**-space can be conducted with histograms for the sum of the interaction energies for each inhibitor in every crystal. For the studied data, the obtained histograms do not indicate high leverages in **X**-space.

It is important to point out that leverages in **X** do not always have a negative effect on the regression model, and that it is necessary to distinguish between 'good' and 'bad' leverages. In our applicationsm, a good leverage would be an inhibitor that provides strong interactions, atypical for other inhibitors in **X**-space. A bad leverage is an inhibitor that is far from the others in the **X**-space as a result of wrong docking. To check the presence of leverages in **X**, robust PCA [25,26] can be applied. This method offers the opportunity to detect leverages based on the so-called robust and orthogonal distances, and constructs a set of robust principal components not affected by them. The leverages are found as those that exceed a cut-off value proposed by Hubert and Engelen [27]. If a sample has high orthogonal and robust distances, it does not fit the PCA model constructed for the majority of the data. Objects with small orthogonal distances but large robust distances fit the PCA model well, even if they are far from the remaining objects in the PCA space. A visual inspection of the score plots of the two first robust PCs obtained for the unfolded data, **X**u, points out that several DAPY-like inhibitors (o), numbers 104, 121, 180, 193 and 195 are far away from the group they should belong to (see Fig. 2a).

This indicates that some of the inhibitors are possibly wrongly docked in crystals. The leverage diagnostic was performed in the space of 11 robust principal components selected, based on the criterion that the sum of their eigenvalues divided by the total sum of eigenvalues should exceed 80% [25]. There are many inhibitors exceeding the cut-off value for both robust and orthogonal distances (see Fig. 2b). In order to be sure that the leverages found are due to improper docking and not due to uniqueness of the inhibitors, the docking of each inhibitor was visually inspected for all of the crystals. The overview of this study and more detailed infor-

Table 1
Overview of the PLS, EP and UVE-PLS models ($\bar{\mathbf{y}}$ represents mean biological activity, and **y**1–**y**5 the individual activities of the wild HIV type and its four mutant strains)

Data	PLS		EP			UVE-PLS		
	f	RMSECV (q^2)	f	RMSECV (q^2)	No. of outliers	f	RMSECV (q^2)	No. of variables
$\bar{\mathbf{y}}$, **X**u	3	0.6910 0.3141	7	0.3618 0.7803	40	4	0.3675 0.7732	151
y1, **X**u	1	0.6040 0.1672	9	0.3077 0.6832	43	6	0.3132 0.6717	83
y2, **X**u	2	0.8577 0.3421	8	0.4075 0.8022	45	5	0.4331 0.8018	73
y3, **X**u	3	0.7773 0.3277	9	0.4061 0.7441	46	6	0.3932 0.7600	116
y4, **X**u	1	0.7628 0.3150	9	0.4002 0.7671	46	4	0.3915 0.7771	163
y5, **X**u	4	0.7678 0.2944	7	0.3986 0.7746	42	6	0.4282 0.7399	122

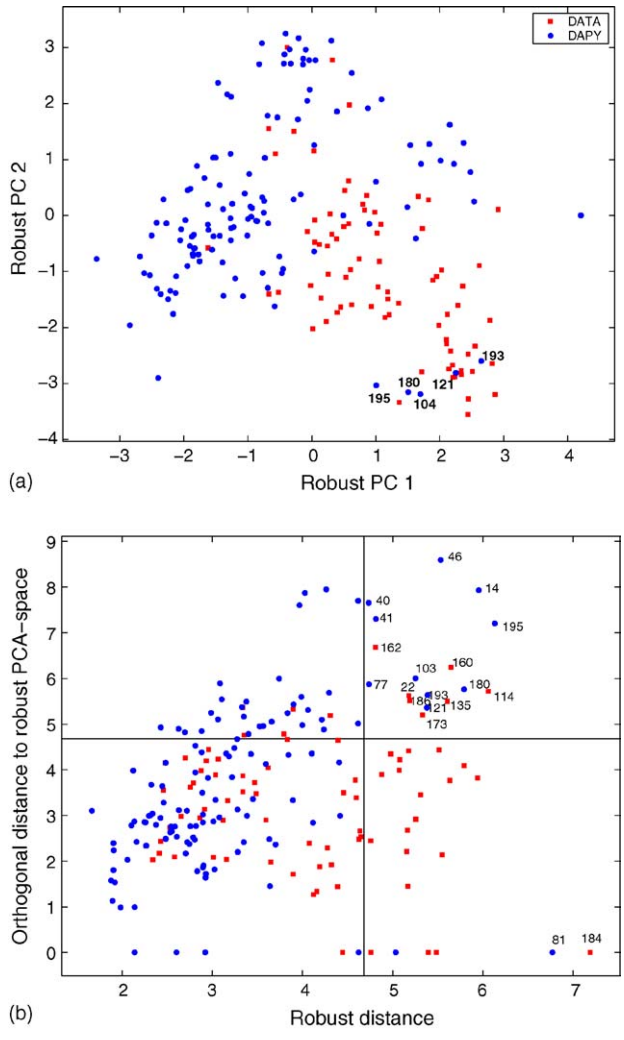

Fig. 2. Robust PCA: (a) projection of inhibitors on the plane defined by the first two robust principal components; (b) identification of leverage objects based on orthogonal and robust distances.

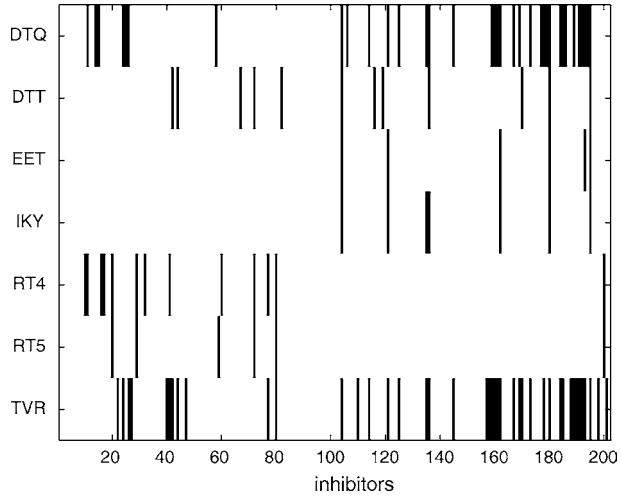

Fig. 3. Summary of wrongly docked molecules (denoted as black lines) in seven studied RT crystals (DTQ, DTT, EET, IKY, RT4, RT5 and TVR).

mation about wrongly docked inhibitors is given in Fig. 3 and Table 2, respectively.

As shown in Fig. 3, there are difficulties in the dockings of several inhibitors. For instance, far location of DAPY-like inhibitors (numbers 104, 121, 180, 193 and 195) is caused by wrong docking in three to five crystals. Docking is a very difficult, crucial and sophisticated step. It requires an exhaustive search of the best orientation of an inhibitor in the RT pocket via an optimization procedure [5]. The landscape of the potential solutions is dominated by sub-optimal solutions, and thus, an optimization procedure can be easily trapped in one of them. This can eventually result in a wrong orientation of the inhibitor in an RT pocket. Finding wrongly docked inhibitors just based on the set of their interaction energies is not always possible. Depending on the level of 'outlyingness', the inhibitor still can establish interactions with amino acids in RT, but to a smaller degree. Some of the outliers are relatively easy to find out. For instance, DATA or DAPY-like inhibitors can be discriminated by the specific interactions they provide. If the inhibitor belongs to DAPY-like inhibitors and it differs much with the interaction energies that are typical for DAPY-like compounds, it can be considered as an outlier.

The outliers in **y** are the result of the experimental error in determining the biological activity. The biological activity of inhibitors is studied in cell-based assays [7]. One source of error in **y**-space is a cell penetration effect. Before an inhibitor

Table 2
Overview of the quality of docking in seven crystals

Crystal	Outliers	No. of outliers
DTQ	11 14 15 24 25 26 58 104 106 114 121 125 135 136 145 159 160 161 162 167 169 173 177 178 179 180 184 185 186 189 191 192 193 194 195	35
DTT	42 44 67 72 82 104 116 119 136 170 180 195	12
EET	104 121 162 180 193 195	6
IKY	104 121 135 136 162 180 195	7
RT4	10 11 16 17 20 29 32 41 60 72 77 80 200	13
RT5	20 29 59 72 80 200	6
TVR	22 24 26 27 40 41 42 44 47 77 80 104 110 114 121 125 135 136 145 157 158 159 160 161 162 167 169 170 173 178 180 184 185 188 189 190 191 192 193 195 198 201	42

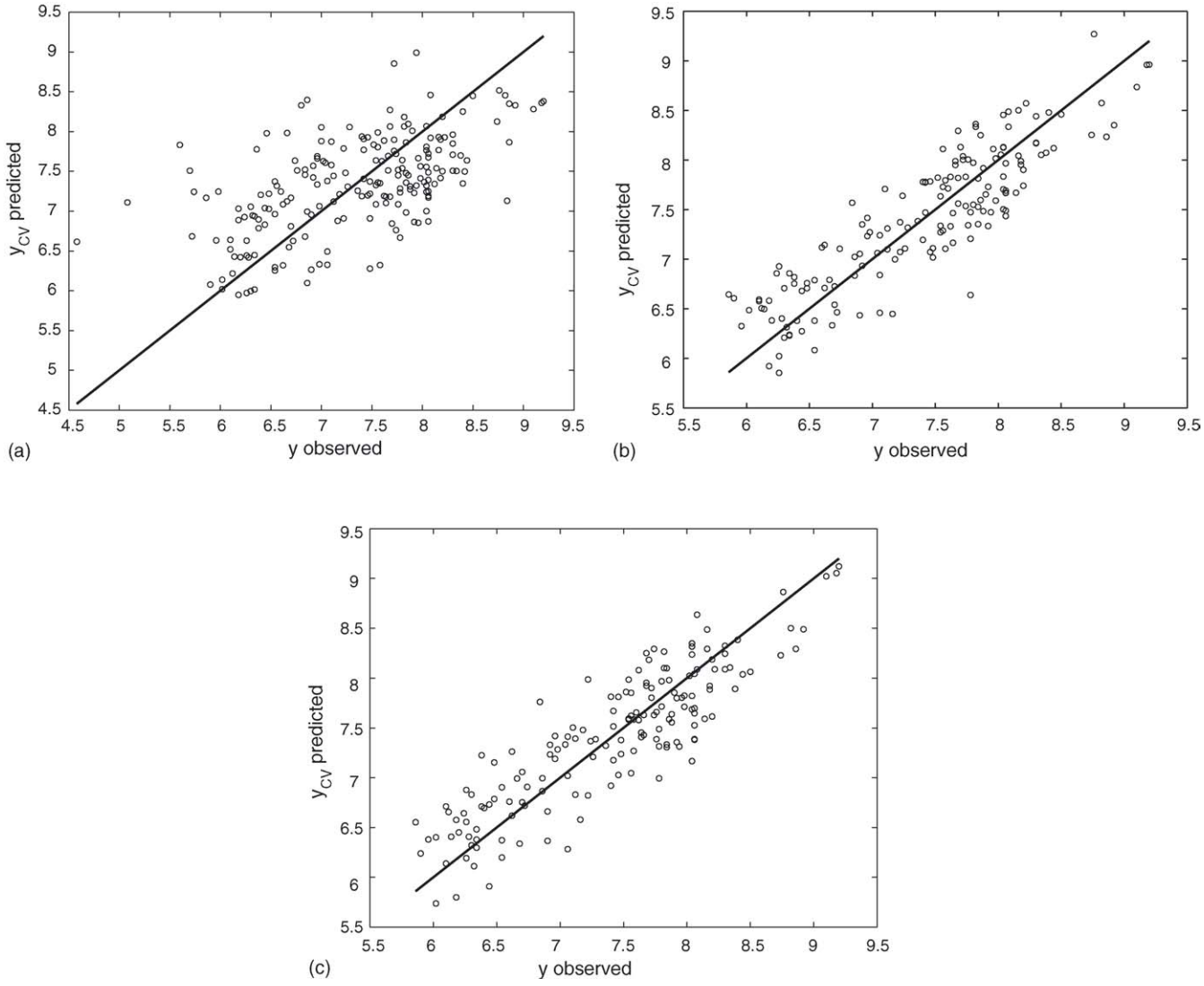

Fig. 4. Models for the unfolded energy tables and mean biological activity: (a) initial PLS model: mean activity observed (**y**) vs. mean activity predicted based on leave-20-out cross-validation (y_{CV}); (b) EP model (robust PLS) with assumed faction of data contamination 25%: mean activity observed (**y**) vs. mean activity predicted based on leave-20-out cross-validation (y_{CV}); (c) UVE-PLS model after removing the outliers and the uninformative explanatory variables: mean activity observed (**y**) vs. mean activity predicted based on leave-20-out cross-validation (y_{CV}).

reaches RT, it has to penetrate the cell membrane to enter the cell, and at this stage, some of the inhibitors encounter difficulties. For such inhibitors, the observed activity is smaller than expected from the calculated interaction with the RT enzyme. To eliminate the cell penetration effect, a loss of activity (a contrast), i.e., the difference between activity for the wild type, and activity for a mutant strain can be used. For these models, the RMSECVs and q^2 values vary from 0.5139 to 0.6500 and 0.2375 to 0.3303, respectively (see Table 3). Although, working with contrasts instead of original activities eliminates one problem, the PLS models for four contrasts and the unfolded **X**u shows no improvement, taking into account a smaller variability of the contrasts compared to original activity.

Being aware of outliers presence in the data, the robust PLS model for mean activity and unfolded **X**u was constructed. The expected data contamination was set to 25%. With the robust PLS approach, the data contamination can be assumed to be high without a risk of throwing out too many inhibitors. The final robust PLS models give evidence

Table 3
Results of PLS models for four biological activity contrasts **c**1–**c**4

Data	f	RMSECV (q^2)
c1, **X**u	2	0.6500
		0.3157
c2, **X**u	3	0.5139
		0.3303
c3, **X**u	3	0.5283
		0.2375
c4, **X**u	1	0.6042
		0.2525

of smaller contamination in the data than assumed. For the robust PLS model using mean activity, 40 inhibitors are found as outliers in the calibration model, which corresponds to 19.80% of the data contamination. When robust PLS models are constructed using individual activities, in the extreme case, 46 inhibitors were found as outliers in the calibration model. This corresponds to 22.77% of the data contamination.

The robust approach allows an improvement of the initial PLS models. Now, the RMSECV and the q^2 values equal to 0.3618 and 0.7803, respectively. This is 0.3292 gain in terms of RMSECV and 0.4662 in terms of q^2, as compared to classical PLS (see Table 1). Without doubt, the robust model describes the majority of the data well.

A model of biochemical data, as the one studied here, having q^2 value above 0.70 (see Eq. (4)) is considered as a good one. The model can be further improved by variable selection. To achieve this, a UVE-PLS model was constructed after removing outliers. 331 explanatory variables were detected as informative variables by the UVE-PLS method. The RMSECV and the q^2 values for the model are equal to 0.3675 and 0.7732, respectively. Compared to the EP model for mean activity, the UVE-PLS model has lower complexity (see Table 1 and Fig. 4). As an example, the initial model, EP and UVE-PLS models, constructed for mean activity, are presented in Fig. 4.

The overall results of the modeling of the individual activities and contrasts for the unfolded energy tables are presented in Tables 2 and 3.

4. Conclusions

Different problems concerning biochemical data quality can highly influence the estimation of a regression model. The lack of good predictive ability of the linear PLS model can be due to: (1) a non-linear relationship, (2) insufficient information in **X** to describe **y**, (3) existence of outliers in the calibration, and (4) presence of many uninformative explanatory variables with high variance, and small covariance with the dependent variable **y**. In the case of an unsatisfactory initial model, all these aspects should be checked while constructing the PLS model.

It was demonstrated, based on an example of the HIV data, that the robust regression approaches are well suited for applications using biochemical data. By the use of robust approaches, it is possible without any risk, to construct a model followed by the majority of the data, regardless of outlying observations. Robust PCA helps to detect outlying observations in **X**-space, which can give an overall impression about the quality of docking but does not inform about outliers in regression. The regression outliers can be efficiently detected as outliers from a robust PLS model, constructed by the EP approach. Additional improvement of the model can be achieved by uninformative feature elimination by means of UVE-PLS.

References

[1] S. Balaji, C. Karthikeyan, N.S. Hari Narayana Moorthy, P. Trivedi, Bioorg. Med. Chem. Lett. 14 (2004) 6089–6094.
[2] R. Kunal, J.T. Leonard, Bioorg. Med. Chem. 12 (2004) 745–754.
[3] P. Pungpo, S. Hannongbua, J. Mol. Graph. Modelling 18 (2000) 581–590.
[4] M.R. de Jonge, L.M.H. Koymans, H.M. Vinkers, F.F.D. Daeyaert, J. Heeres, P. Lewi, J. Med. Chem., in press.
[5] F. Daeyaert, M. de Jonge, J. Heeres, L. Koymans, P. Lewi, M.H. Vinkers, P.A.J. Janssen, Proteins 54 (2004) 526–533.
[6] A.R. Leach, Molecular Modelling, Principles and Applications, Longman, London, 1996.
[7] K. Hertogs, M.P. de Béthune, V. Miller, T. Ivens, P. Schel, A. Van Cauwenberge, C. Van Den Eynde, V. van Gerwen, H. Azijn, M. van Houtte, F. Peeters, S. Staszewski, M. Conant, S. Bloor, S. Kemp, B. Larder, R. Pauwels, Antimicrob. Agents Chemother. 42 (1998) 269–276.
[8] H. Martens, T. Næs, Mutivariate Calibration, John Wiley & Sons, Chichester, UK, 1989.
[9] T. Næs, T. Isaksson, T. Fearn, T. Davies, Multivariate Calibration and Classification, NIR Publications, Chichester, UK, 2002.
[10] S. Wold, M. Sjöström, L. Erriksson, Chemom. Intell. Lab. Sys. 58 (2001) 109–130.
[13] B. Walczak, Chemom. Intell. Lab. Sys. 28 (1995) 259–272.
[14] B. Walczak, Chemom. Intell. Lab. Sys. 29 (1995) 63–73.
[15] C.B. Lucasius, G. Kateman, Chemom. Intell. Lab. Sys. 19 (1993) 1–33.
[16] C.B. Lucasius, G. Kateman, Chemom. Intell. Lab. Sys. 25 (1994) 99–145.
[17] D. Brynn Hibbert, Chemom. Intell. Lab. Sys. 19 (1993) 277–293.
[18] P.J. Rousseeuw, A.M. Leroy, Robust Regression and Outlier Detection, John Willey & Sons, New York, 1987.
[19] V. Centner, D.L. Massart, O.E. de Noord, S. de Jong, B. Vandeginste, C. Sterna, Anal. Chem. 68 (1996) 3851–3858.
[20] J. Ren, J. Diprose, J. Warren, R.M. Esnouf, L.E. Bird, S. Ikemizu, M. Slater, J. Milton, J. Balzarini, D.I. Stuart, D.K. Stammers, J. Biol. Chem. 275 (2000) 5633–5639.
[21] M. Hogberg, C. Sahlberg, P. Engelhardt, R. Noreen, J. Kangasmetsa, N.G. Johansson, B. Oberg, L. Vrang, H. Zhang, B.L. Sahlberg, T. Unge, S. Lovgren, K. Fridborg, K. Backbro, J. Med. Chem. 42 (1999) 4150–4160.
[22] J. Lindberg, S. Sigurdsson, S. Lowgren, H.O. Andersson, C. Sahlberg, R. Noreen, K. Fridborg, H. Zhang, T. Unge, Eur. J. Biochem. 269 (2002) 1670–1677.
[23] J. Ren, R.M. Esnouf, A.L. Hopkins, J. Warren, J. Balzarini, D.I. Stuart, D.K. Stammers, Biochemistry 37 (1998) 14394–14403.
[24] K. Das, J. Ding, Y. Hsiou, A.D. Clark Jr., H. Moereels, L. Koymans, K. Andries, R. Pauwels, P.A. Janssen, P.L. Boyer, P. Clark, R.H. Smith Jr., M.B. Kroeger Smith, C.J. Michejda, S.H. Hughes, E. Arnold, J. Mol. Biol. 264 (1996) 1085–1100.
[25] M. Hubert, P.J. Rousseeuw, S. Verboven, Chemom. Intell. Lab. Sys. 60 (2002) 101–111.
[26] I. Stanimirova, B. Walczak, D.L. Massart, V. Simeonov, Chemom. Intell. Lab. Sys. 71 (2004) 83–95.
[27] M. Hubert, S. Engelen, Robust, Bioinformatics 20 (2004) 1728–1736.