The Use of the GRID Program in the 3-D QSAR Analysis of a Series of Calcium-Channel Agonists

Andrew M. Davis,*,† Nigel P. Gensmantel,† Erik Johansson,‡ and David P. Marriott†

Fisons PLC Research and Development Laboratories, Bakewell Road, Loughborough, Leicestershire, LE11 0RH, United Kingdom

Received June 2, 1993⊗

The use of GRID in the 3-D QSAR analysis of a series of calcium-channel agonists is described. Partial least-squares analysis of GRID maps showing the interaction energy between an alkyl hydroxyl probe and a series of agonists in 3-D space generated a predictive quantitative model of the variation of biological activity. The macroscopic descriptors CLOGP and CMR were included in the analysis, and the importance of appropriate block scaling is highlighted. The discussion highlights the interpretation of the resulting regression maps, and the steric, electrostatic, lipophilic, and hydrogen-bonding preferences of the calcium-channel receptor are identified.

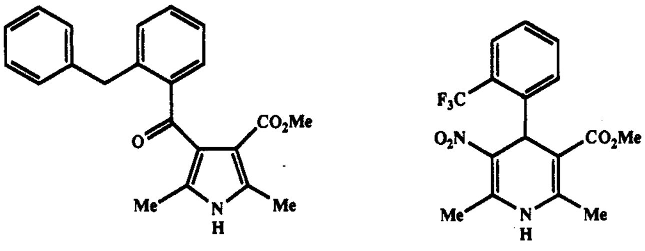
Introduction

In general, descriptors used in QSAR studies only model the magnitude, not the directional preferences of a particular physical property. Traditional QSAR studies have used descriptors based on experimentally derived 1-octanol–water partition coefficients to model the "hydrophobic effect", Hammett substitutent constants to model electronic effects, and a wide range of descriptors, from molecular weights to complex topological indices, to model steric interactions.[1-3] These types of descriptors could generate a data set with 10's of different descriptors. The traditional statistical tool used in such analyses has been multiple linear regression.

In recent years, the growth in importance of computational chemistry approaches has provided a plethora of molecular and atom-based descriptors that can and have been used in QSAR studies. These include descriptors derived from individual atomic partial charges, HOMO/LUMO energies, and nucleophilic/electrophilic superdelocalizabilities, etc.[4-6] Including these types of descriptors, one could easily end up with a data matrix of up to 100 descriptors to analyze. Multivariate statistical techniques had to be adopted with so many descriptors. Using simple multiple linear regression with so many variables can cause severe problems because of chance correlations, collinearity, and multicollinearity.[7] Techniques such as principal components analysis, principal components regression, factor analysis, and partial least-squares analysis, which identify smaller numbers of uncorrelated underlying descriptors that can describe biological activity, have been increasingly applied.[4,8-10]

The traditional and computational chemistry types of descriptors are, in general, scalar properties. However, the CoMFA approach of Cramer, Patterson, and Bunce looked at molecules in 3-D, from the viewpoint of the "receptor", and described the magnitude and directional preferences of electronic and steric interactions.[11,12] The technique measured the interaction energy in terms of steric and electrostatic interactions between a methyl probe bearing unit positive charge at a series of regular grid positions around and through a series of molecules. The molecules were previously overlaid/aligned to occupy the same position in space. This technique generates, typically, many thousands of descriptors (the interaction energies over the series of molecules at particular points in space) and necessitates the use of a multivariate data analysis technique; the method generally used is PLS.[13] One advantage of the CoMFA technique is that the results of the analysis can be mapped back into 3-D space providing a 3-D picture of the forces important in controlling biological activity.

Chart 1



FPL 64176 (13)

BAY K 8644

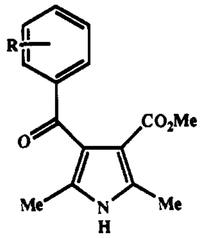
We have used the GRID force field[14-17] to compute the interaction energy between a series of target molecules and a probe atom or group, over a regular 3-D grid both around and through the target molecules. GRID calculates the total energy of interaction, which is the sum of electrostatic, steric, and hydrogen-bonding terms. The probe can be chosen from a wide range of predefined probe molecules. The force field was originally developed to probe the interior of proteins for interaction sites useful for drug design. GRID has been used successfully to predict binding sites of small ligands in proteins,[18] and it has been extended to evaluate properties of small molecules. We have used the RS/1-table-based suite of statistical software to compile and manage the individual grids.[19] We have analyzed data generated by GRID using PLS, as implemented inside the SIMCA multivariate analysis package.[20] The analysis also included the more traditional CLOGP and CMR descriptors. The resulting model was used to identify relationships between the physicochemical properties and biological activity of a set of calcium-channel agonists.

The discovery and the synthesis of methyl 2,5-dimethyl-4-[2-(phenylmethyl)benzoyl]-1*H*-pyrrole-3-carboxylate (13)

† Fisons PLC Research and Development Laboratories.
‡ Umetri AB, Tvistevagen, Umea, Sweden.
⊗ Abstract published in *Advance ACS Abstracts*, February 15, 1994.

0022-2623/94/1837-0963$04.50/0

Table 1. CLOGP, CMR, and Force of Contraction (EC_{50}) Measured Relative to BAY K 8644 for 36 Compounds Used in the QSAR Analysis



compd	R	CLOGP	CMR	relative force (EC_{50})
1	2-Cl	2.63	7.58	0.0943
2	2-CF_3	3.09	7.60	0.27
3	2-OCH_3	2.03	7.70	0.0053
4	2-H	2.18	7.08	0.059
5	2-OCO(2′-OH-C_6H_5)	4.10	10.40	0.34
6	2-CH_3	2.67	7.58	0.14
7	2-F	2.34	7.10	0.0093
8	2,4-Cl_2	3.35	8.07	0.33
9	2-I	3.04	8.39	0.22
10	2-Br	2.78	7.86	0.15
11	2-OCH_2Ph	3.80	10.21	1.13
12	2-Cl, 4-NO_2	2.41	8.30	0.16
13	2-CH_2Ph	4.01	10.06	35.5
14	2-Ph	4.01	9.60	0.174
15	2-SCH_2Ph	4.61	10.87	2.89
16	2-$SOCH_2Ph$	2.41	10.90	0.312
17	2-SO_2CH_2Ph	2.16	10.93	0.021
18	2-CH_2CH_2Ph	4.62	10.52	8.00
19	2-CH_3, 4-CH_3	3.17	8.01	0.0568
20	2-SPh	4.62	10.40	2.57
21	2-SOPh	2.18	10.44	0.34
22	2-NHPh	4.79	9.96	18.91
23	2-CH_2(4′-NO_2Ph)	3.84	10.79	4.31
24	2-CH_2(2′-NO_2-Ph)	3.56	10.78	2.90
25	2-S(4′-NO_2-Ph)	4.46	11.13	1.24
26	2-O(4′-NO_2-Ph)	4.13	10.47	0.96
27	2-CH_2(4′-NH_2-Ph)	2.87	10.43	0.0457
28	2-OSO_2(4′-Me-Ph)	3.06	10.62	0.0072
29	2-OPh	4.21	9.75	2.90
30	2-NH-pyrid-2′-yl	3.94	9.75	7.70
31	2-$CH_2C_6H_{11}$	5.32	10.15	27.6
32	2-NHC_6H_{11}	4.84	10.06	27.6
33	2-Br, 4-F	2.92	7.877	0.220
34	2-CH_2(4′-F-Ph)	4.24	10.08	14.0
35	2-CH_2Ph, 4-F	4.25	10.08	19.0
36	2-CH_2(4′-F-Ph), 4-F	4.40	10.09	19.0

(FPL 64176, Chart 1) and analogs were recently described.[21] These represent the first examples of a new class of calcium-channel activators. The modulation of transmembrane calcium movement is an important area of current pharmacological research with applications in many therapeutic areas. The discovery of FPL 64176 was directly guided by a linear regression model, which showed the importance of lipophilicity and steric size to the observed activity. Thus, this data set provided a good vehicle for our study of the usefulness of GRID and SIMCA to drug design. We hoped the inclusion of more recently synthesized compounds would provide a deeper insight into the physicochemical factors controlling the activation of the calcium channel by this class of compounds.

Table 2. Distribution of the Interaction Energy Ranges between the Hydroxyl Probe and the 36 Compounds at Each GRID Point in Space (which are each columns in the compiled RS/1 table)

range intervals ($E_{max} - E_{min}$ in each column, kcal/mol)	number of columns/GRID points with range in the interval
0 up to 0.1	12818
0.1 up to 0.2	965
0.2 up to 0.3	196
0.3 up to 0.4	94
0.4 up to 0.5	96
0.5 up to 1.0	427
1.0 up to 1.5	193
1.5 up to 2.0	141
2.0 up to 3.0	165
3.0 up to 4.0	66
4.0 up to 5.0	146
5.0 up to 6.0	172
6.0 up to 7.0	81
7.0 up to 8.0	42
8.0 up to 9.0	16
9.0 up to 10.0	7
total	15625

Materials and Methods

Synthesis and biological testing protocols of the calcium-channel agonists have been described elsewhere.[21,22] In brief, the compounds were tested for their ability to increase cardiac contractility using guinea pig atria paced at 1 Hz. The inotropic potency of the compounds was measured as the concentration of drug to increase developed tension to 50% of the isoprenaline maximum in the 1-Hz paced guinea pig atria. The results were reported relative to the EC_{50} of Bay K 8644, Chart 1, the standard calcium-channel agonist, and included in the QSAR analysis as log(relative inotropic potency). The macroscopic descriptors CLOGP and CMR were calculated using MEDCHEM, version 3.54.[23] The compounds studied are shown along with CLOGP, CMR, and their inotropic EC_{50} data in Table 1.

Molecular Alignment. The X-ray conformation observed for FPL 64176 (**13**) was used as the starting point for the construction of the 3-D structures of the 36 compounds. Substituent variations were built in CHEM-X[34] using standard bond lengths and angles. The structures were not fully optimized. Full optimization would have introduced small differences in bond angles, bond lengths, and torsion angles to the fixed parts of the molecules in the test set, which would have introduced "noise" into the GRID analysis. Here, all of the molecules contained a common molecular fragment, the dimethyl-substituted pyrrole ring, known to be important for binding. The structural variation occurred on the phenyl ring at the position ortho to the linking keto group adjoining the pyrrole ring. Initial molecular alignment involved overlaying the pyrrole ring of each structure followed by conformational analysis of the side chain. Here, we fitted all the side chains to the conformation adopted by the benzyl side chain of FPL 64176, since this conformation was a low-energy one for the compound in question. The compound FPL 64176 is one of the most active compounds in the series, and so, the alignment approximates to the active-analog approach. Since it is not possible to deduce the bioactive conformation of FPL 64176, we have selected an arbitrary conformation (the X-ray conformation).

The particular conformation of FPL 64176 chosen is irrevelant, as the frame of reference of the analysis is the molecule and not its absolute position in 3-D space. It is possible, and probably likely, that the "active" conformation of FPL 64176 is not that observed in the X-ray structure, but that does not matter to this analysis. If we had chosen another conformation of the benzyl side chain, since our alignment rules were the same, we believe the resulting statistical analysis would have been virtually identical. Each GRID point in space will become a descriptor variable in the PLS analysis and be represented by a column in the input table. A global rotation of all the varying substituents of the test set would correspond to permuting the order in which the descriptors/columns appear in this table. (This assumes that at the new *x*, *y*, and *z* positions, each GRID point lies in the same position relative to the substituent as previously. It also assumes that the nonrotated parent part of the molecule offers a constant interaction across the set of compounds.) Permuting the columns of a table has no effect upon extraction of principal components or PLS components. In a recent paper, Klebe and Abraham demonstrated for a set of inhibitors of thermolysin and human rhinovirus 14, where protein–ligand complex crystallographic data provided information on the true binding conformation, that alignments based on a theoretical binding conformation gave CoMFA models of equal or superior predictive power compared

Table 3. PLS Regression Models for the Full 36-Compound Data Set[a]

block variances	PLS 1[b]	PLS 2	PLS 3	PLS 4	overall r^2
model 1	r^2 = 0.69				
CLOGP = 1.0					
act = 1.0					
model 2	r^2 = 0.42	n/s	n/s	n/s	0.42
GRID = 1458					
act = 1.0					
model 3	r^2 = 0.42	n/s	n/s	n/s	0.42
GRID = 1458					
CLOGP = 1					
CMR = 1					
act = 1					
model 4	r^2 = 0.60	r^2 = 0.71	r^2 = 0.77 n/s	r^2 = 0.86	0.86
GRID = 1					
CLOGP = 1					
CMR = 1					
act = 1					

[a] n/s, not significant by cross-validation (5% level); PRESS > LIMIT (0.9025). [b] PLS 1, the first PLS component, and PLS 2, the second PLS component, etc.

Table 4. PLS Regression Models for the Full 36-Compound Data Set Showing the Effect of Changing the Relative Scaling of the GRID Block vs the Macroscopic Descriptors[a]

ratio variances (GRID/log P, etc.)	PRESS				
	PLS 1	PLS 2	PLS 3	PLS 4	overall PRESS
10:1	0.6478	0.9912, n/s	0.9869, n/s		0.6478
5:1	0.5963	0.8785	1.1535, n/s	0.9869	0.5160
3:1	0.5146	0.8742	1.0060, n/s	0.8485	0.3817
2:1	0.4554	0.8973	1.0692, n/s	0.8114	0.3315
1:1	0.4268	0.8515	1.2799, n/s	0.7765	0.2821
0.5:1	0.4316	0.8506	0.9846, n/s	0.7895	0.2854
0.33:1	0.4337	0.8595	0.9643, n/s	0.8044	0.2891
0.2:1	0.4349	0.8668	0.9554, n/s	0.8115	0.2923
0.1:1	0.4355	0.8707	0.9511, n/s	0.8144	0.2937

[a] n/s, not significant (5% level); PRESS > LIMIT (0.9025).

to those based on the experimental binding conformation.[24] The active-analog approach assumes that for compounds to be active they should adopt a similar conformation to the most active of the series.[25] We hypothesize that compounds that adopt different binding modes are likely to be observed as outliers in x–y correlation space in this type of analysis. Until a 3-D QSAR analysis becomes available where X-ray data on protein–ligand complexes are also available on all the cases studied, we will not be able to test this hypothesis.

Charges. The GRID-defined atomic charges are assigned on the basis of atom types. They are insensitive to changes in structure in small molecules, e.g., changing a substituent on a phenyl ring does not change the charges on the ring atoms. Therefore, the GRID charges were replaced with MNDO/PM3 Mulliken charges, calculated using MOPAC 5.0 running on a Convex C220 minisupercomputer. MOPAC charges would give more representative charges for small molecules. Previously, we have used Gasteiger charges, but the use of MOPAC-derived charges should give a better representation of inductive and mesomeric electronic effects in the molecules under study.

Probes. An alkyl hydroxyl probe was selected as the probe molecule, as this would provide information on electrostatic interactions and hydrogen-bond donation and accepting ability. It also has a size, therefore generating steric information. We decided that the nature of the probe was unimportant as long as it could interact via all mechanisms. It is possible that a probe also bearing a formal charge would put a more appropriate emphasis upon electrostatic interactions, and GRID affords the possibility of defining custom probes if necessary.

During the GRID calculations, the bulk dielectric was set to 4.0, representing the estimated dielectric of the active site of a receptor. In the preliminary work, we used a bulk dielectric of 80.0, but we decided using the more realistic lower value would give a better representation of hydrophobic effects. Setting the dielectric to 4.0 would also increase the contribution of the electrostatic term and provide a good compromise between electrostatic and steric terms. If the dielectric had been set lower than 4.0, then the electrostatic term would start to dominate.

The interaction energies between the set of test molecules and the hydroxyl probe were measured at 1-Å spacings over a 25-Å cubic GRID generating 15 625 points for each molecule. The GRID spacing should be as small as is practicable to use. In CoMFA work, GRID spacings of 2 Å are often used. As long as the increased redundancy in the data set can be adequately removed, smaller grid spacing gives less sparse, more informative regression maps.

Map Data Preparation. The GRID maps were compiled into a table in RS/1, each column representing a point in space and each row a compound in the test set, generating a 36-row × 15625-column table. The x, y, and z coordinates of the GRID points were written as the column titles of this table. The column titles provide the key to collapsing the dimensionality of the GRID block, to removing redundant information, and to regenerating the original GRID later in the analysis for the display of results. The negative energy values generally ranged from 0 to −9 kcal/mol but the positive values from 0 to 50.0 kcal/mol (the cutoff value set by GRID). As extraction of PLS components is scale-dependant, this would unduly bias the analysis to the steric terms.[26] We therefore scaled all positive energies by 12.5 so they would only cover the range 0–4.0 kcal/mol.

To analyze the information content of the RS/1 map data table, a table was constructed showing the distribution of column/GRID point ranges, Table 2. Analysis of this distribution table demonstrated that the compiled GRID map data table contained many GRID points/columns at which the probe showed little or no variation in interaction energy across the set of test compounds. This was because: (a) a very large grid was used; therefore, many GRID points were so far away from all the molecules that the interaction energy between the probe and all molecules was 0 or nearly 0 kcal/mol; (b) the common parts of the molecule provide a constant interaction with the probe; and (c) as part of the molecular volume is common to the whole set, there are regions of space where the probe is inside the van der Waals surface of the whole set, so the interaction energies were constant at 4.0 kcal/mol (after scaling).

Inclusion of these redundant columns would grossly affect the chance of extracting a useful PLS model. A table was constructed in RS/1 that was a subset of the 15 625 master table that contained columns/GRID points where the range of energy values ($E_{max} - E_{min}$) was greater than 0.2 kcal/mol, generating a 1842-column table. This cutoff was arbitrary, and we could have equally used a higher cutoff, e.g., at a 0.3 or 0.4 kcal/mol range, without losing too much x-block information. Thus, only around 10% of the data contained any potentially useful information. The whole molecule descriptors CLOGP and CMR were added to this table with the activity data to generate a 1845-column × 36-row data table for analysis.

Statistical Analysis. The PLS routine implemented in SIMCA, version 4.4, was used. In the version available to us, up

Table 5. PLS Regression Models for the Full 36-Compound Test Set Showing the Effect of Removing Redundant x-Descriptors Using VINFM[a]

	PRESS and $\sum r^2$				
VINFM cutoff	PLS 1	PLS 2	PLS 3	PLS 4	PLS 5
1845 cols with VINFM ≥ 0.0	0.4268 r^2 = 0.601	0.8515 r^2 = 0.714	1.2799, n/s r^2 = 0.786	0.777 r^2 = 0.86	1.1593, n/s
795 cols with VINFM ≥ 0.2	0.4261 r^2 = 60.2	0.8512 r^2 = 0.715	1.287, n/s r^2 = 0.788	0.788 r^2 = 0.858	1.096, n/s
511 cols with VINFM ≥ 0.4	0.4252 r^2 = 0.602	0.8469 r^2 = 0.718	1.2915, n/s r^2 = 0.791	0.7895 r^2 = 0.857	1.0162, n/s
391 cols with VINFM ≥ 0.6	0.4242 r^2 = 0.603	0.8472 r^2 = 0.718	1.2888, n/s r^2 = 0.791	0.7774 r^2 = 0.858	0.9490, n/s
309 cols with VINFM ≥ 0.8	0.4246 r^2 = 0.602	0.8442 r^2 = 0.717	1.2865, n/s r^2 = 0.793	0.7913 r^2 = 0.854	0.9336, n/s
205 cols with VINFM ≥ 1.0	0.4217 r^2 = 0.605	0.8510 r^2 = 0.719	1.2935 r^2 = 0.78	0.8560 r^2 = 0.831	0.9322, n/s

[a] In each case, the block variances were scaled to 1.0. n/s, not significant (5% level); PRESS > LIMIT (0.9025).

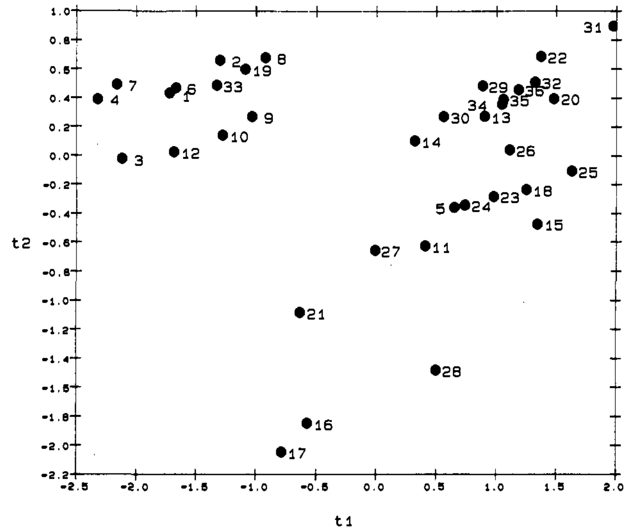


Figure 1. Plot of the x-scores on PLS component 1 vs component 2, $t1$ vs $t2$ for the 36-compound block-scaled model. This illustrates how the compounds are spread over the x-space of the first two components most important in describing biological activity.

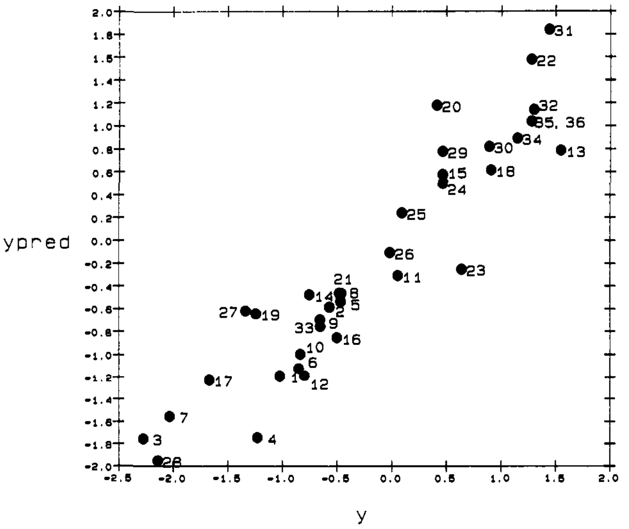


Figure 2. Plot of predicted $\log_{10}$ relative force vs observed $\log_{10}$ relative force for the 36-compound block-scaled model.

to a 60-case × 5600-variable matrix could be analyzed. After each component had been extracted, the significance of that component to the model and the overall model significance were checked by cross-validation. Using as default seven groups, which with 36 cases approximated to leave-five-out approach, the significance was tested with the PRESS statistic. The prediction error sum of squares (PRESS) is the squared differences between observed and predicted values when the objects i are kept out of the model for each y variable m:

$$\text{PRESS} = \sum_{im}(\hat{y}_{im} - y_{im})^2$$

For each PLS component, the PRESS/SS was calculated, where SS is the residual sum of squares of the previous dimension, and the (PRESS/SS)m was calculated for each y variable. When the PRESS/SS (total or for any dimension) is smaller than a significance LIMIT (5% level), the tested dimension is considered significant.

The use of cross-validation to test the model significance has many advantages over using distribution-based tests of model significance such as F-tests. Cross-validation always tests the model in prediction, and as we want to use the model to guide the design of new compounds, then this is preferable. Also, the use of cross-validation does not impose any assumptions upon the distribution of errors in the model, which may not be valid with this type of data.[27] (Most statistical tests of significance assume errors follow a normal distribution.)

Data Block Scaling. The effect of changing the variance of the GRID block of descriptors relative to the macroscopic descriptors was examined. The block variance of the GRID block is calculated by summing the individual column variances for all the m columns for the i cases in the GRID block:

$$\text{block var} = \sum\nolimits_1^m \sum\nolimits_1^i (\bar{x} - x_i)^2/(i - 1)$$

Block scaling to unity was achieved by dividing the interaction energy for every case in every column by the total block standard deviation:

$$\text{scaled values} = x_{i,m}/\text{block SD}$$

for i cases and m columns in the block.

Here, the block variance for the 1842 GRID columns was 1458, so to scale the block variance to unity, each interaction energy for all cases in all columns in the block was divided by √1458, i.e., 38.18. The scaling of the x-blocks is easily achieved in SIMCA by defining columns in the block and assigning a scaling "weight" to that block of 1/38.18, i.e., 0.062.

The PLS weights, which show the importance of the original variables to each PLS component, and the regression coefficients, which show the importance of the original variables to the complete multicomponent model, were extracted from the SIMCA output and used to recompile maps in RS/1. The points in space that were not included in the PLS analysis had their weights/regression coefficients set to zero. The map tables were output in a format readable by the molecular modeling program CHEM-X and displayed on an Evans and Sutherland PS300 graphics terminal supporting stereo.

Results

PLS analysis is sensitive to the scaling of the x-block descriptors. Because the units of the GRID columns are

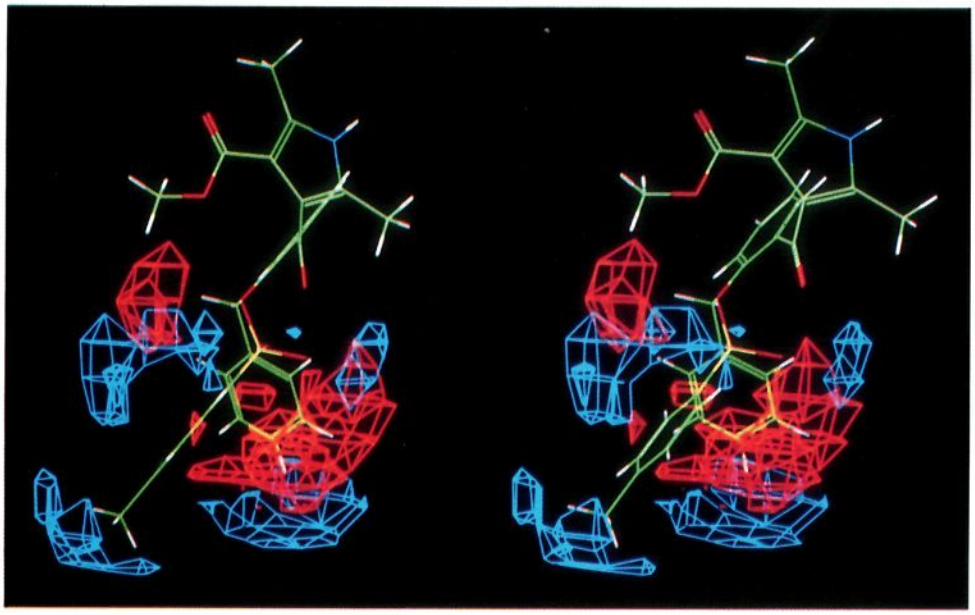


Figure 3. Negative regression coefficients (blue) and positive regression coefficients (orange) superimposed on the most active compound (**13**) and least active compound (**28**) of the 36-compound block-scaled model.

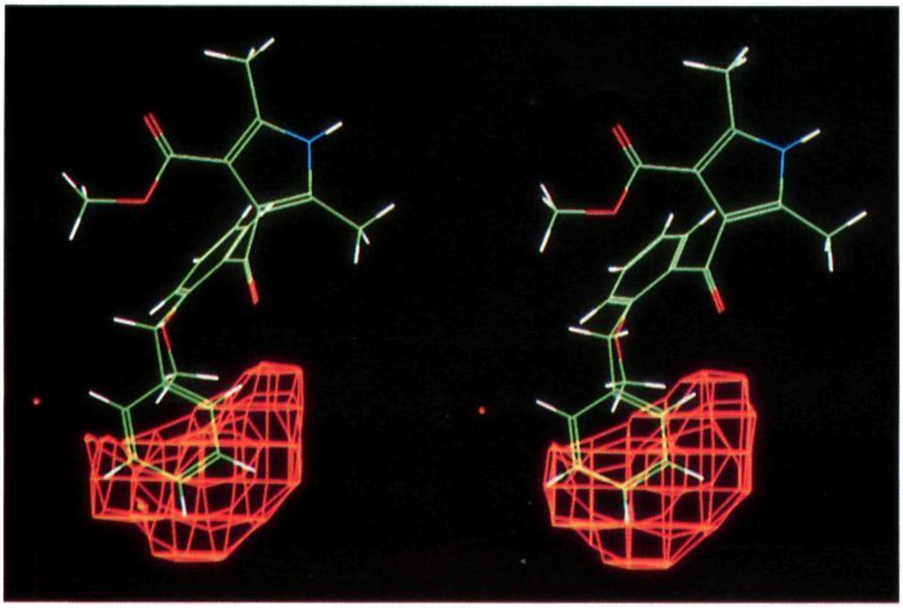


Figure 4. Negative PLS weights (blue) and positive PLS weights (orange) of the GRID points onto PLS 1 superimposed onto the most positively influential (high *t*1 vs *u*1) compound (**13**) and most negatively influential (low *t*1 vs *u*1) compound (**3**) of the 36-compound block-scaled model. The macroscopic descriptor CLOGP weights heavily, also, on this component. The component is dominated by positive weights (at the contouring level shown no negative weights are displayed), regions in space where it is favorable to place lipophilicity.

identical, i.e., kcal/mol, the GRID columns were not autoscaled. Autoscaling, which sets each column's variance to unity, would put undue weight on columns containing little variation in interaction energy over the test set of compounds. But inclusion of one whole-molecule descriptor such as CLOGP along with 100's or 1000's of columns of GRID information requires careful attention to scaling. Table 3 shows the effect of changing the relative scaling of the variance of the GRID block to CLOGP and CMR column variances.

Inclusion of CLOGP and CMR with the 1842 columns of GRID information without block scaling, Table 3, model 3, has no effect upon the model obtained when compared to the model extracted from just the GRID information, Table 3, model 2. Although CLOGP alone describes 69% of the *y*-block variation, without block scaling, the variable does not contribute significantly to the model. But when the GRID block variance is scaled to give the total variance of all columns of 1.0, the same as the CLOGP column, the complete model now explains 86% of the activity data in four PLS components, Table 3, model 4. This shows that in this data set where lipophilicity is known to be important in controlling the observed inotropic potency, the best PLS model can only be identified after appropriate scaling. A similar approach has recently been used by Silipo[28] and McFarland[29] for the inclusion of macroscopic descriptors with CoMFA data. Kim has demonstrated that in some cases lipophilic effects can be parameterized directly from

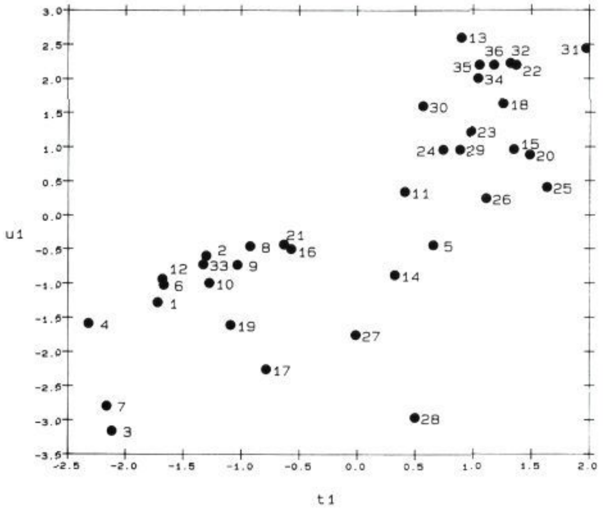


Figure 5. Plot of (x/y)-scores on PLS 1, $t1$ vs $u1$ for the 36-compound block-scaled model.

the molecular field of CoMFA.[30] But for this data set where lipophilicity is known to be important in controlling biological activity, the best model can only be extracted by explicitly including the macroscopic descriptors CLOGP and CMR with the GRID data with appropriate block scaling.

The questions now arise, how does one know what the relative scaling between the blocks should be and how does this effect the model extracted? Table 4 shows how the model is affected by altering the relative scaling between the GRID and CLOGP x-blocks. The choice of 1:1 for the relative scaling of the GRID block to the variance of the macroscopic descriptors (which were each scaled to unit variance) was arbitrary. But it did allow both the macroscopic descriptors and the GRID information to contribute to the model. The importance of scaling down the variance of the GRID block drastically, from 1458:1 as in the raw data to 10:1, shows steady improvement in the quality of the models. Below 2:1, the model is optimal. Once the variance ratio is approximately below 2:1, a stable model results.

Recently, a number of papers have described the improvements in model predictability by reducing the dimensionality of the x-block, i.e., by removing redundant data that contributes little to the x–y correlation.[31–33] SIMCA computes a diagnostic called the variable influence, which shows the influence on y of every term in the model. This is computed for each component and cumulatively for the whole PLS model. For a PLS dimension, the variable influence is given by the squared PLS weight of that term multiplied by the percent explained sum of squares of that PLS dimension. The cumulative variable information is the sum of VINFM over all dimensions.

GRID points that are not important to y can be identified by extracting PLS components without cross-validation (generally extracting one or two more than are identified by cross-validation) and reading the cumulative VINFM data generated. GRID points with very small cumulative VINFM values are of little importance in controlling y and do not model x, whereas GRID points with large cumulative VINFM values are most important in the PLS model. The points in space with very low VINFM values can be removed from the model development process without affecting the fitting to y. Table 5 shows the results of using VINFM to cut the dimensionality of the model from 1842 columns. As can be observed, the model is virtually identical once the dimensionality has been reduced from 1842 down to 205 columns. This demonstrates that the model contains many GRID points that, although they contain x-block information, contain little information useful in the x–y correlation. The PRESS statistics for the first four components, which were significant in the full model, are slightly improved in the reduced models. The PRESS for the fifth component markedly decreases as the dimensionality is reduced but in this work still does not reach a level of significance where it could be included. Thus, strong signals in the data are little affected by the noise reduction, but weaker signals may become significant as the dimensionality is reduced.

Discussion

Since the mechanism of biological action might differ between different types of compounds, it is difficult to construct model QSAR's that apply to structurally diverse compounds. One way to ensure that the compounds are structurally homogeneous is to plot the first few PLS or

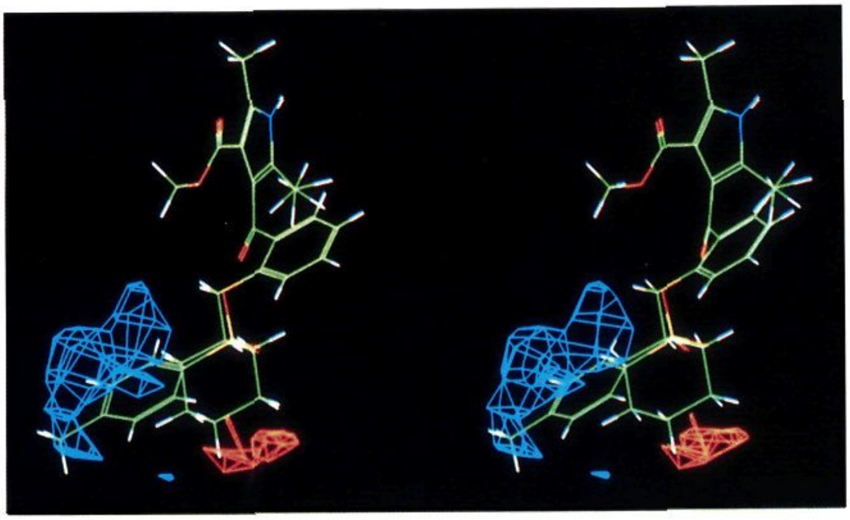


Figure 6. Negative PLS weights (blue) and positive PLS weights (orange) of the GRID points on PLS 2 superimposed onto compound **31** (high $t2$ vs $u2$) and compound **28** (low $t2$ vs $u2$) of the 36-compound block-scaled model. The model identifies unfavorable interactions for large side chains.

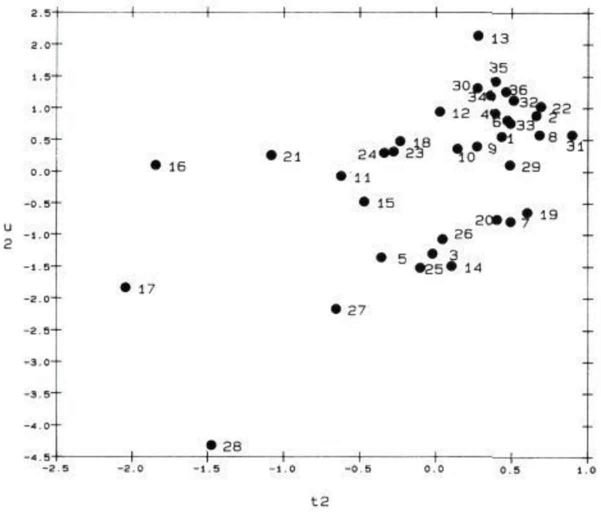


Figure 7. Plot of (x/y)-scores on PLS 2, $t2$ vs $u2$ for the 36-compound block-scaled model.

PCA score dimensions of the molecular descriptors. These plots, commonly denoted $t1$ vs $t2$ for the first two dimensions, should be free of groupings. In Figure 1, we have plotted $t1$ vs $t2$, the scores of the 36 compounds on the first two components of the four-component block-scaled model. It does show two groupings and suggests two different types of compounds that should be treated separately. A close examination of the compounds in the two groups reveals no apparent difference other than their size. The grouping therefore comes from small hydrophilic compounds (upper left corner) and large compounds. We believe that the grouping comes from a lack of "medium-sized compounds" and that here all the compounds can be treated as structurally homogeneous.

The four-PLS-component model with block scaling of the 36-compound set, Table 3, model 4, describes 86% of the variation in $\log_{10}$ force observed; a plot of y vs y-predicted is shown in Figure 2. We can estimate if the model is overfitted, too many PLS components, or underfitted, too few PLS components, by comparing the square of the measurement error to the residual variation around the model, $(y - y_{\text{pred}})^2$. In this case, assuming a measurement error of 2-fold (0.303 in $\log_{10}$ units) and the $(y - y_{\text{pred}})^2$ of 0.14, we find that the squared measurement error, 0.0908, is slightly less than the $(y - y_{\text{pred}})^2$. This indicates that no significant under- or overfit is present.

The contribution of the GRID data to model 4 is shown in Figure 3, which shows the regression coefficients for each GRID point mapped back into 3-D space. This is displayed with the structures of compound **13** and **28**, representatives of the most active and least active compounds studied, respectively. The overall regression coefficient map represents a composite picture of the four-PLS components extracted for the 36-compound model. A number of discrete regions in space are mapped out by positive and negative regression coefficient contours. What do the negative and positive regions mean in terms of the interaction energies at those positions? A negative regression coefficient/PLS weight indicates that at that position in space as the interaction energy between the probe and the series of molecules gets more negative the compounds become more active. This could be due to a favorable electronic/hydrogen-bonding interaction with the receptor being identified or an unfavorable steric interaction. Conversely, a positive regression coefficient/PLS weight shows that as the interaction energy across the set becomes more positive the observed binding energy becomes higher. This could be due to an unfavorable electrostatic interaction with the receptor, or a region in space where it is favorable to place steric bulk being identified.

A four-component PLS model indicates that four underlying statistical/physical properties have been identified as important in describing y. The overall GRID regression map, though, shows many discrete mapped regions. In the simplest case for a four-component PLS model, one could expect four positive and their respective four negative mapped regions, but due to collinearity in the data set, more are often observed. The question is, how many of these mapped regions offer independent useful information and which are they? This problem is twice as complicated in CoMFA, as the electrostatic and

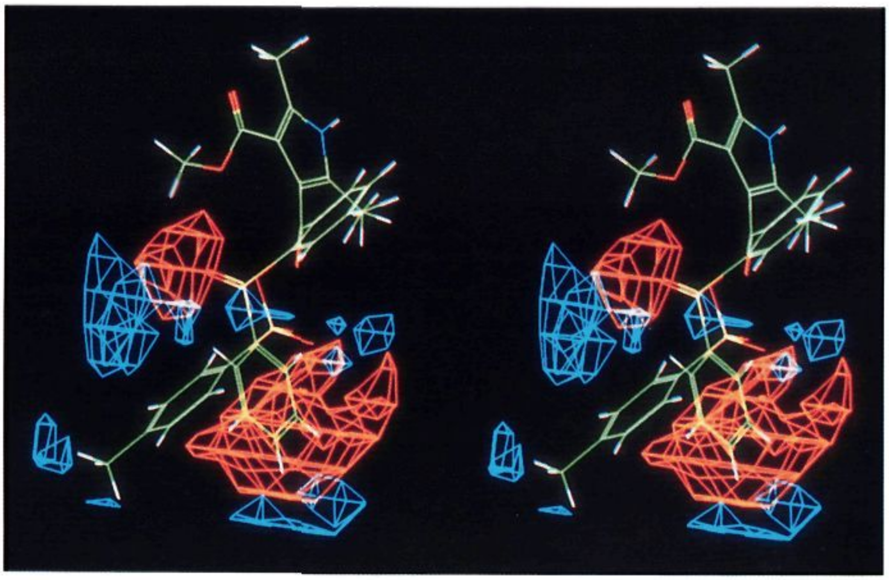


Figure 8. Negative weights (blue) and positive weights (orange) of the GRID points onto PLS 3 superimposed onto compound **21** (high $t3$ vs $u3$) and compound **28** (low $t3$ vs $u3$) of the 36-compound block-scaled model. This identifies that benzyl substituents and their isosteres are more active than phenethyl isosteres.

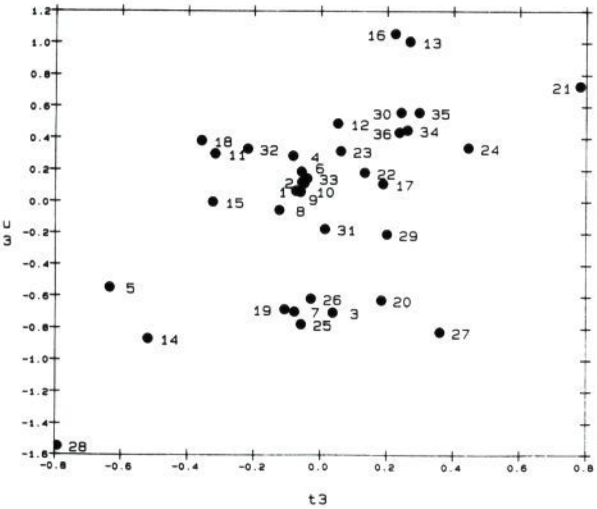


Figure 9. Plot of the (x/y)-scores on PLS 3, $t3$ vs $u3$ for the 36-compound block-scaled model.

steric information is shown on two separate maps. To answer this question, we inspected the individual PLS-component weighting maps. Each PLS component identifies a separate "underlying statistical physical property" that is important in determining biological activity. GRID points with high PLS weights are important in defining that component, i.e., are highly collinear with that component and therefore contain similar information. Therefore, all mapped regions that weight onto a single PLS component should have a single statistical/physical interpretation. The interpretation of the PLS weighting maps was aided by examining them with a plot of the scores (t's) of the compounds on the PLS x-component vs the scores of the y/y's on the PLS y-component (u's). Compounds that appear at the positive and negative ends of the t/u axis are those whose GRID fields are most important in defining that component. The weightings maps were displayed over the structures of the two compounds with the most negative t/u and the most positive t/u values, compounds that are most influential in defining that component.

PLS 1, the first PLS component, for model 4 describes 61% of the variance in biological activity, 84% of the variance of the CLOGP descriptor, and 67% of CMR weight onto this component. Figure 4 shows the weights of each GRID point that also load upon PLS 1, which was interpreted in conjunction with a plot of $t1$ vs $u1$, Figure 5. The GRID regions mapped therefore show points in space where it is favorable to place a bulky lipophilic substituent. PLS 1 is dominated by regions of positive coefficients. As can be seen from Figure 4, the benzyl side chain of compound **13**, a representative which scores on the high positive $t1/u1$ axis, fills this region of positive contours. This represents regions in space from which the –OH probe is repelled, correlating with high biological activity. On the contrary, the methoxy side chain of compound **3**, a representative of a compound scoring low negative on the $t1/u1$ axis, does not enter this volume in space.

Figure 6 shows the weights of GRID points onto PLS 2, the second PLS component, and Figure 7 shows a plot of $t2$ vs $u2$. The remaining 11% of CLOGP and 30% of CMR load onto this component, the CMR term with a negative weighting. The side chain of compound **28**, scoring low negative on the $t2/u2$ axis, fills a region of negative contours, which dominate this component. Many of the smaller substituents score on the positive end of this component. The component shows that too large a substituent can be detrimental to activity.

Figure 8 shows the weighting of GRID points onto PLS 3, and Figure 9 shows a plot of $t3$ vs $u3$. This shows that benzyl substituents and their isosteres have favorable positive contours around the region of space they occupy, while the region of the aromatic ring of phenyl and phenethyl isoteres is filled with negative contours, which is unfavorable for this steric interaction.

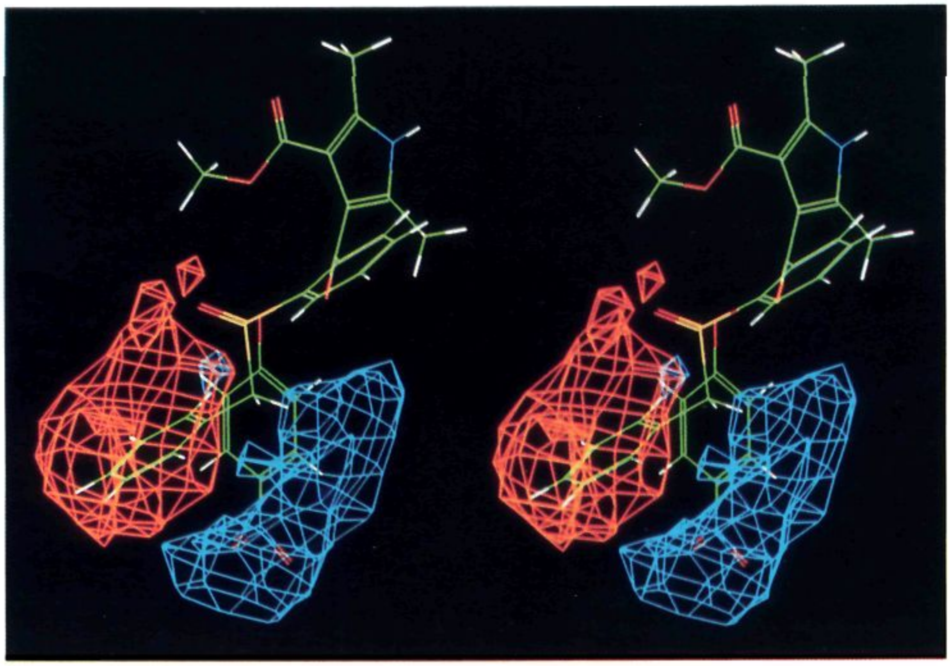


Figure 10. Negative weights (blue) and positive weights (orange) of the GRID points onto PLS 4 superimposed onto compound **16** (high $t4$ vs $u4$) and compound **26** (low $t4$ vs $u4$). Inspection of this map together with Figure 11 suggests that para substitution on benzyl substituents could have a small unfavorable effect upon activity.

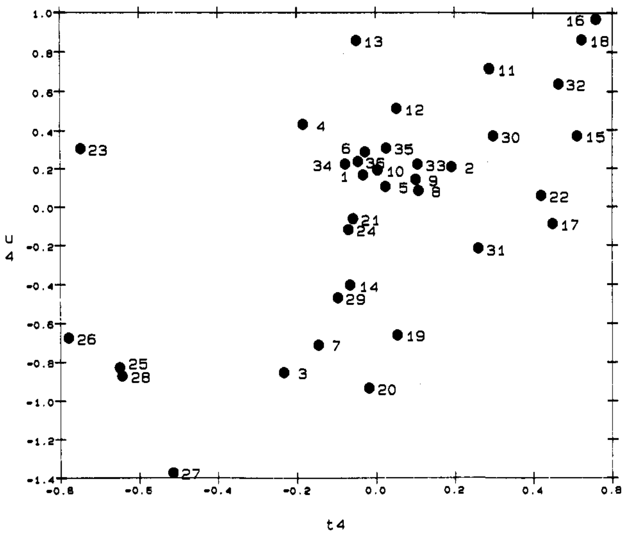


Figure 11. Plot of the (*x*/*y*)-scores on PLS 4, *t*4 vs *u*4 for the 36-compound block-scaled model.

Figure 10 shows the weightings of GRID points onto PLS 4, which explains an extra 7.3% of the *y* variance remaining. Inspection of this with Figure 11, which shows a plot of *t*4 vs *u*4, shows that benzyl and phenethyl isosteres containing para substituents appear to have a small unfavorable effect upon biological activity. This component could also be rebalancing some residual variation introduced by the fitting of the previous components.

The overall regression model therefore shows that lipophilic benzyl substituents are the most active compounds and small hydrophilic substituents are the least active. Of the larger substituents, benzyl substituents and their isosteres are more active than phenethyl substituents and their isosteres once log *P* has been accounted for. Also, para substituents on benzyl or phenethyl compounds have a small unfavorable effect upon the biological activity once their lipophilicity has been accounted for.

Conclusion

We have developed a system for 3-D QSAR on the basis of the programs' GRID, RS/1, and SIMCA. This system offered the flexibility to investigate the effects of data preprocessing on the statistical analysis, and we have shown the importance of removing redundant variables that contain no information and the importance of variable scaling. We have demonstrated how interpretation of the resulting PLS model can be aided not only by examining the overall regression contour maps but also by examining the individual PLS weighting maps. These were used with plots of *t*'s vs *u*'s which show the inner correlation of the extracted PLS *x*-components (*t*'s) and PLS *y*-components (*u*'s). The interpretation via weighting maps would be certainly recommended in any 3-D QSAR analysis where more than one field source is used, for instance, in CoMFA work where the steric and electrostatic fields are treated separately.

Acknowledgment. We wish to thank Dr. Darren Flower for writing programs used in this work and all the staff at Umetri AB for many helpful discussions.

References

(1) Hansch, C.; Fujita, T. σ-π-p-Analysis: A Method for the Correlation of Biological Activity and Chemical Structure. *J. Am. Chem. Soc.* **1964**, *86*, 1616–1626.

(2) Verloop, A.; Hoogenstraaten, W.; Tipker, J. Development and Application of New Steric Substituent Parameters in Drug Design. In *Drug Design*; Ariens, E. J.; Ed.; Academic Press: New York, 1976; Vol. VII.

(3) Balaban, A. T.; Chiriac, A.; Motoc, I.; Simon, Z. *Steric Fit in Quantitative Structure-Activity Relations*: Springer-Verlag: Berlin, 1980.

(4) Ford, M. G.; Greenwood, R.; Turner, C. H. The Structure-Activity Relationships of Pyrethroid Insecticides. 1. A Novel Approach Based on the Use of Multivariate QSAR and Computational Chemistry, *Pestic. Sci.* **1989**, *27*, 305–326.

(5) Ford, M. G.; Livingstone, D. J. Multivariate Techniques for Parameter Selection and Data Analysis Exemplified by a Study of Pyrethroid Neurotoxicity. *Quant. Struct.-Act. Relat.* **1990**, *9*, 107–114.

(6) Boel, M. Theoretical Investigation on Steroid Structure and QSAR. In *Molecular Structure and Biological Activity of Steroids*; Boel, M., Duax, W. L., Eds.; CRC Press: Boca Raton, FL, 1992.

(7) Topliss, J. G.; Edwards, R. P. Chance Factors in Studies of Quantitative Structure-Activity Relationships. *J. Med. Chem.* **1979**, *22*, 1238–1244.

(8) Jonsson, J.; Eriksson, L.; Sjostrom, M.; Wold, S. A Strategy for Ranking Environmentally Occurring Chemicals. *Chemom. Intell. Lab. Syst.* **1989**, *5*, 169–186.

(9) Eriksson, L.; Jonsson, J.; Sjostrom, M.; Wold, S. A Strategy for Ranking Environmentally Occurring Chemicals. Part ii. An Illustration with Two Data Sets of Chlorinated Aliphatics and Aliphatic Alcohols. *Chemom. Intell. Lab. Syst.* **1989**, *7*, 131–141.

(10) Eriksson, L.; Jonsson, J.; Hellberg, S.; et al. Multivariate Quantitative Structure-Activity Relationships for Halogenated Aliphatics. *Environ. Toxicol. Chem.* **1990**, *9*, 1339–1351.

(11) Cramer, R. D.; Patterson, D. E.; Bunce, J. D. Comparitive Molecular Field Analysis (CoMFA). 1. Effect of Shape on Binding of Steroids to Carrier Proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959–5967.

(12) Cramer, R. D.; Patterson, D. E.; Bunce, J. D. Recent Advances in Comparative Molecular Field Analysis (CoMFA). *Prog. Clin. Biol. Res.* **1989**, *291*, 161–165.

(13) Wold, S. Partial Least Squares Analysis. In *3-D QSAR in Drug Design, Theory, Methods and Application*; Kubinyi, H., Ed.; ESCOM Science Publishers: Leiden, Holland, 1993.

(14) Goodford, P. J. A Computational Procedure for Determining Energetically Favorable Binding Sites on Biologically Important Macromolecules. *J. Med. Chem.* **1985**, *28*, 849–857.

(15) Boobyer, D. N. A.; Goodford, P. J.; McWhinnie, P. M.; Wade, R. C. New Hydrogen Bond Potentials for Use in Determining Energetically Favorable Binding Sites on Molecules of Known Structure. *J. Med. Chem.* **1989**, *32*, 1083–1094.

(16) Wade, R. C.; Clark, K. J.; Goodford, P. J. Further Development of Hydrogen Bond Functions for Use in Determining Energetically Favorable Binding Sites on Molecules of Known Structure. 1. Ligand Probe Groups with the Ability to Form Two Hydrogen Bonds. *J. Med. Chem.* **1993**, *36*, 140–147.

(17) Wade, R. C.; Goodford, P. J. Further Development of Hydrogen Bond Functions for Use in Determining Energetically Favorable Binding Sites on Molecules of Known Structure. 2. Ligand Probe Groups with the Ability to Form More Than One Hydrogen Bond. *J. Med. Chem.* **1993**, *36*, 148–156.

(18) Itzstein, M.; Yang, W. W.; Kok, G. B.; et al. Rational Design of Potent Sialidase-based Inhibitors of Influenza Virus Replication. *Nature* **1993**, *363*, 418.

(19) RS/1; BBN Software Products, 10 Fawcett St., Cambridge, MA 02238.

(20) SIMCA 4.4; developed and distributed by Umetri AB, Umea, Sweden.

(21) Baxter, A. J. G.; Dixon, J.; Ince, F.; Manners, C. N.; Teague, S. J. Discovery and Synthesis of Methyl 2,5-dimethyl-4-[2-(phenylmethyl)benzoyl]-1*H*-pyrrole-3-carboxylate (FPL 64176) and Analogues: the First Examples of a New Class of Calcium Channel Activator. *J. Med. Chem.* **1993**, *36*, 2739–2744.

(22) Kennedy, R. H.; Seifen, E. Stimulation Frequency Alters the Inotropic Response of Atrial Muscle to Bay K-8644. *Eur. J. Pharmacol.* **1985**, *107*, 209–214.

(23) MEDCHEM, version 3.54; Daylight CIS: USA, 1993.

(24) Klebe, G.; Abraham, U. On the Prediction of Binding Properties of Drug Molecules by Comparative Molecular Field Analysis. *J. Med. Chem.* **1993**, *36*, 70–80.

(25) Dammkoeler, R. A.; Karasek, S. F.; Shands, E. F. B.; Marshall, G. R. Computer-Aided Drug-Design: the Active-Analog Approach. *J. Comput.-Aided Mol. Des.* **1989**, *3*, 3–21.

(26) Cocchi, M.; Johansson, E. Amino Acids Characterization by GRID and Multivariate Data Analysis. *Quant. Struct.-Act. Relat.* **1993**, *12*, 1–8.

(27) Wold, S. Cross-Validatory Estimation of the Number of Components in Factor and Principal Components Models. *Technometrics* **1978**, *20*, 397–404.

(28) Greco, G.; Novellino, E.; Silipo, C.; Vittoria, A. Study of Benzodiazepines Receptor Sites Using a Combined QSAR CoMFA Approach. *Quant. Struct.-Act. Relat.* **1992**, *11*, 461–477.

(29) McFarland, J. W. Comparitive Molecular Field Analysis of Anticoccidial Triazines. *J. Med. Chem.* **1992**, *35*, 2543–2550.

(30) Kim, K. H. A Novel Method of Describing Hydrophobic Effects Directly from 3-D Structures in 3D-Quantitative Structure-Activity Relationships. *Med. Chem. Res.* **1991**, *1*, 259–264.

(31) Cruciani, G.; Baroni, M.; Clementi, S.; Constantino, G.; Riganelli, D.; Skagerberg, B. Prediction Ability of Regression Models, Part 1. The SDEP parameter. *J. Chemom.* **1992**, *6*, 335–346.

(32) Baroni, M.; Clementi, S.; Cruciani, G.; Constantino, G.; Riganelli, D.; Oberrauch, E. Prediction Ability of Regression Models, Part 2. Selection of the Best Predictive PLS Model. *J. Chemom.* **1992**, *6*, 347–356.

(33) Allen, M. S.; LaLoggia, A. J.; Dorn, L. J.; et al. Predictive Binding of Beta-Carboline Inverse Agonists and Antagonists via the CoMFA/GOLPE Approach. *J. Med. Chem.* **1992**, *35*, 4001–4010.

(34) CHEM-X, developed and distributed by Chemical Design Ltd., Chipping Norton, Oxon., U.K.