Comparative Molecular Field Analysis (CoMFA)

Hugo Kubinyi
BASF AG, D-67056 Ludwigshafen, Germany

Abbreviations

3D = three-dimensional; C = molar concentration of a drug; CBG = corticosteroid binding globulin; CoMFA = comparative molecular field analysis; CoMSIA = comparative molecular similarity indices analysis; GOLPE = generating optimal linear PLS estimations; PLS = partial least squares; PRESS = predictive residual sum of squares; RMS = root mean squares; TBG = testosterone binding globulin.

1 INTRODUCTION

Most drugs that are nowadays used in human therapy interact with certain macromolecular biological targets, e.g., with enzymes, receptors, ion channels, and transporters, in some cases even with desoxyribonucleic acid.[1–3] With the exception of some irreversibly reacting enzyme inhibitors (e.g., acetylsalicylic acid and penicillin) and of alkylating agents, most drug actions result from the noncovalent association of a small ligand (the drug) to a specific binding site at the macromolecule. Thus, a precondition for the biological activity of a drug is its high affinity to this binding site. Enzyme inhibitors prevent either the binding of a substrate or the catalytic reaction at the active site; receptor antagonists hinder the binding of an agonist or the adoption of an 'active' receptor conformation. The situation is more difficult with receptor agonists and partial agonists. In addition to their affinity to a certain binding site, they have different intrinsic activities, i.e., different abilities to produce the agonist-mediated biological effect. Nowadays the most reasonable hypothesis is that receptor agonists stabilize the 'active' conformation of a receptor, whereas antagonists stabilize its 'inactive' conformation.[1–4]

The building blocks of all biological macromolecules belong to one group of optically active enantiomers. Thus, all drug receptors (the term 'receptor' is most often generally used for any biological macromolecule which is the binding partner of a drug) are themselves enantiomers, i.e., they possess asymmetrical geometries. Therefore it is not surprising that enantiomers of chiral drugs differ in their biological activities, a fact that is not adequately considered in most quantitative structure–activity relationships (QSAR) studies (cf. Sections 2 and 4 and ***QSAR in Drug Design***).

2 CoMFA METHODOLOGY

2.1 History

Classical QSAR correlates biological activities of drugs with physicochemical properties or indicator variables which encode certain structural features.[4–8] In addition to lipophilicity, polarizability, and electronic properties, steric parameters are also frequently used to describe the different size of substituents. In some cases, indicator variables have been attributed to differentiate racemates and active enantiomers.[5,6] However, in general, QSAR analyses consider neither the 3D structures of drugs nor their chirality.

A binding site at a receptor which 'looks' at a ligand would not see atoms and bonds, as we chemists do. From a far distance, it would 'feel' the electrostatic potential of the molecule (Figure 1a) and, at a closer distance, the relatively hard body of the molecule with its charge distribution pattern at the solvent-accessible surface (Figure 1b).

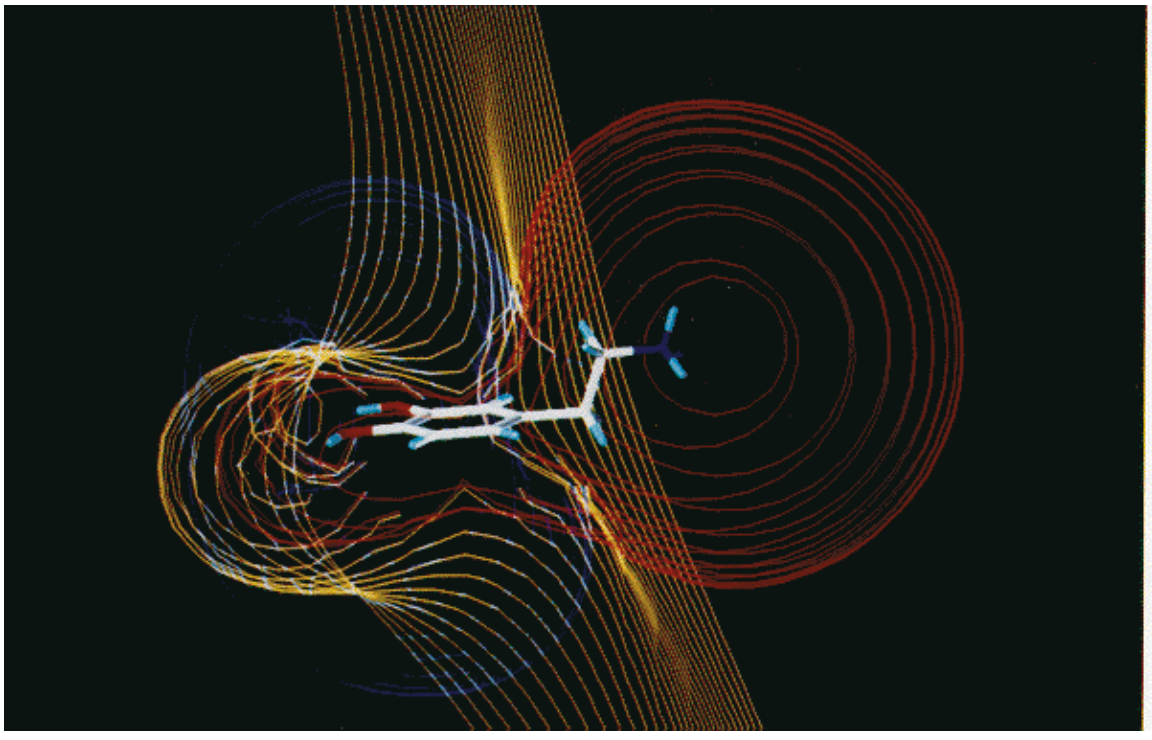
In 1979, Cramer and Milne made a first attempt to compare molecules by aligning them in space and by mapping their molecular fields to a 3D grid.[9] In the following years, this approach was further developed as the DYLOMMS (dynamic lattice-oriented molecular modelling system) method[10] but was not very well accepted by the scientific community. Several important facts had to work together to allow a broader application of this approach:

1. In 1986, Svante Wold proposed the use of partial least squares (PLS) analysis, instead of principal component analysis, to correlate the field values with the biological activities (see ***Partial Least Squares (PLS) in Chemistry***);
2. in 1988, a key publication appeared in the *Journal of the American Chemical Society*[11] and the method was called comparative molecular field analysis (CoMFA) from then on; and
3. appropriate software became commercially available.[12]

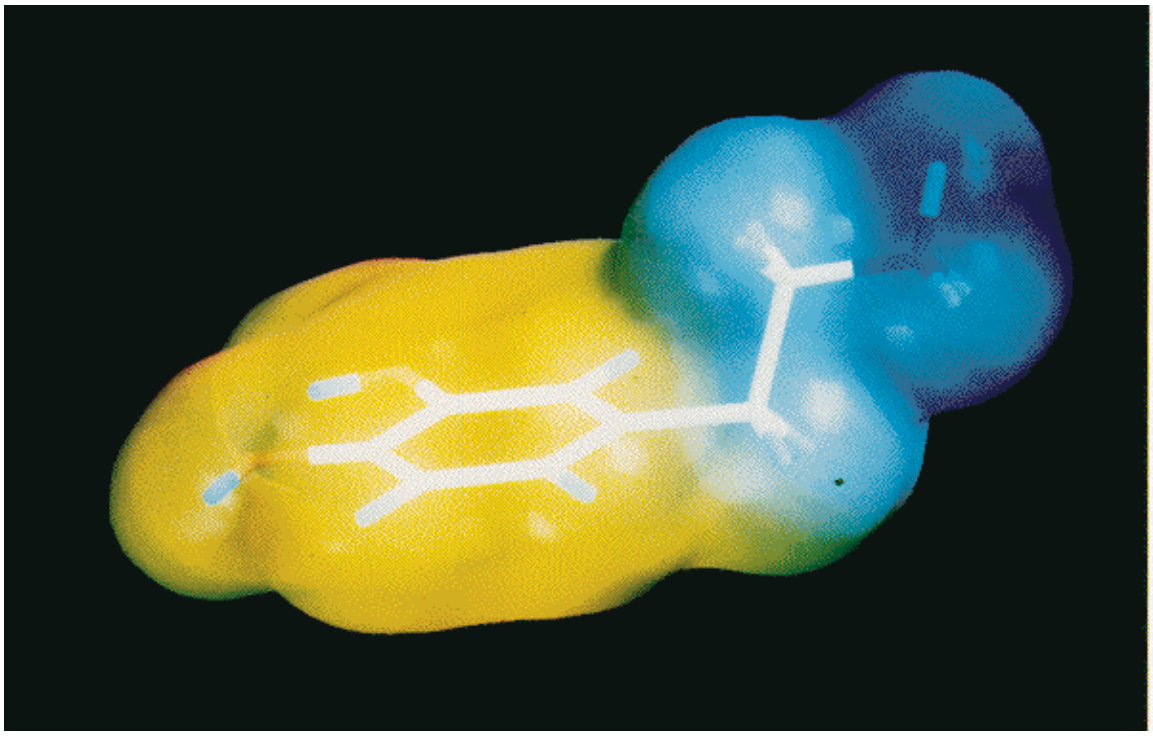
Since 1988, a few hundred publications, several reviews (e.g., Refs. 13–18), and three books[10,19] have appeared on this subject. Despite some major problems in its proper application (cf. Section 7), the method is now generally estimated as a useful tool for deriving 3D QSAR models.

2.2 Steps of a CoMFA

CoMFAs describe 3D structure–activity relationships in a quantitative manner. For this purpose, a set of molecules is first selected which will be included in the analysis. As a most important precondition, all molecules have to interact with the same kind of receptor (or enzyme, ion channel, transporter) in the same manner, i.e., with identical binding



(a)



(b)

Figure 1 Dopamine (3,4-dihydroxyphenethylamine, Formula **3** of Figure 2, Section 4). (a) Electrostatic potential contour lines, calculated for the neutral molecule (blue lines indicate electropositive regions, yellow lines show neutral regions and red lines indicate electronegative regions); atoms are color-coded (carbon white, hydrogen light blue, nitrogen blue, oxygen red). (b) Solvent-accessible surface of the positively charged form of dopamine, with color coding for the electrostatic surface properties (blue areas show electropositive regions, yellow areas indicate neutral regions; Figure 1(b) is reproduced from Ref. 3 with kind permission of Spektrum Akademischer Verlag, Heidelberg)

sites in the same relative geometry. In the next step, a certain subgroup of molecules is selected which constitutes a training set to derive the CoMFA model. The residual molecules are considered to be a test set which independently proves the validity of the derived model(s) (Section 3). Atomic partial charges are calculated and (several) low energy conformations are generated. A pharmacophore hypothesis is derived to orient the superposition of all individual molecules and to afford a rational and consistent alignment (Section 4).

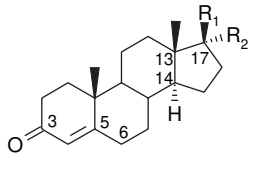
A sufficiently large box is positioned around the molecules and a grid distance is defined. Different atomic probes, e.g., a carbon atom, a positively or negatively charged atom, a hydrogen bond donor or acceptor, or a lipophilic probe, are used to calculate field values in each grid point, i.e., the energy values which the probe would experience in the correponding position of the regular 3D lattice (Section 5). These 'fields' correspond to tables, most often including several thousands of columns, which must be correlated with the binding affinities or with other biological activity values. PLS analysis is the most appropriate method for this purpose (Section 6). Normally cross-validation is used to check the internal predictivity of the derived model.

The result of the analysis corresponds to a regression equation with thousands of coefficients. Most often it is presented as a set of contour maps. These contour maps show favorable and unfavorable steric regions around the molecules as well as favorable and unfavorable regions for electropositive or electronegative substituents in certain positions (Section 6). Predictions for the test set (the compounds not included in the analysis) and for other compounds can be made, either by a qualitative inspection of these contour maps or, in a quantitative manner, by calculating the fields of these molecules and by inserting the grid values into the PLS model (Section 6).

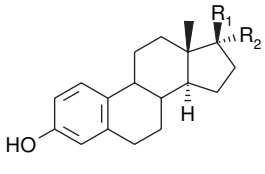
Despite the straightforward definition of CoMFA, there are a number of serious problems and possible pitfalls.[10] Several CoMFA modifications have been described which solve or avoid some of these problems (Section 7).[19] In addition, alternatives to CoMFA were developed, e.g., comparative molecular similarity indices analysis (CoMSIA) (Section 5)[19,20] and other 3D quantitative similarity–activity relationship (QSiAR) methods.[21–24]

2.3 A CoMFA Application

The first application of CoMFA[11] is an illustrative example. It correlates the binding affinities of 21 steroids (e.g., **1** and **2**) to human corticosteroid (CBG) and testosterone binding globulins (TBG). Since steroids are relatively rigid systems (with the exception of the side chains in position 17), the alignment was performed by a rigid body RMS fit of carbon atoms 3, 5, 6, 13, 14, and 17.



1



2

The results, using different options, indicated that the steric field mostly led to an explanation of the variance in the binding data. The fit of the affinity values, expressed by the squared correlation coefficient r^2, was close to 0.9 (0.897 for CBG and 0.873 for TBG). After cross-validation (Section 6), a predictive squared correlation coefficient Q^2 of around 0.6 (0.662 for CBG and 0.555 for TBG) was obtained. Most interesting is the result of the prediction for ten compounds not included in the original CoMFA studies. For the CBG binding data of compounds #22–31 (for structures and numbering see Ref. 11), an $r^2_{\text{pred}} = 0.65$ was given. However, this value must be wrong, owing to a misplacement of compound #7 in Figure 7 of Ref. 11; the correct value is $r^2_{\text{pred}} = 0.31$.[20,24] A possible explanation for this poor test set predictivity is discussed in Section 3.

3 SERIES DESIGN AND TRAINING AND TEST SET SELECTION

The application of statistical methods depends on a proper experimental design, for the training set from which a QSAR model is derived, as well as for the test set, for which biological data shall be predicted (e.g., Refs. 4, 5, 10). In QSAR analyses, this important precondition is most often neglected. No wonder that in such cases problems arise from a biased object selection and from different structural and parameter spaces of the training and test sets.

The training set compounds should span a parameter space in which all data points are more or less equally distributed. The structures and all relevant properties of the test set compounds should not be too far from the test set compounds. To derive statistical models with reasonable experimental effort, an appropriate design scheme should be used to cover the property space with the smallest possible number of objects. Redundancy is minimized by following this recommendation. On the other hand, some redundancy should be included to avoid the possibility that cross-validation (Section 6) is no longer applicable and that single point errors distort the final QSAR model. Especially the latter topic is most often neglected as a possible reason for poor test set prediction. Reasonable results for the test set predictions can only be expected by including sufficient redundancy in the training set compounds.

Most QSAR and 3D QSAR studies are retrospective analyses without an appropriate series design. The consequences are either a poor fit of the training set data or a lack of predictivity for the test set compounds. The poor predictivity of the CBG CoMFA models (compounds **1** and **2**; Section 2.3) is not surprising if one considers that compound #23 is the only one in the whole data set which bears a 21-acetoxy group and that also compound #31, a 9-fluoro-substituted steroid, is outside the structural space of the training set.

The problems of this data set are easily understood if a Free–Wilson analysis is applied.[4–6] The training set compounds (#1–21) can be described by a simple one-parameter regression equation (equation 1; the term 4,5–C=C– indicates the presence or absence of a cycloaliphatic 4,5-double bond in ring A of the steroids).[24] The internal predictivity of this model ($Q^2 = 0.726$; $s_{\text{PRESS}} = 0.630$) and the test set predictivity ($n = 10$; $r^2_{\text{pred}} = 0.477$; $s_{\text{PRESS}} = 0.733$) are even slightly better than the CoMFA result (Section 2.3).

$$\log 1/\text{CBG} = 2.022(\pm 0.52)4,5\text{–C=C–} + 5.186(\pm 0.36)$$

$$(n = 21; r = 0.882; s = 0.568; F = 66.41;$$
$$Q^2 = 0.726; s_{\mathrm{PRESS}} = 0.630) \quad (1)$$

Selection of compounds #1–12 and #23–31 (instead of compounds #1–21) gives a much better presentation of the structural space in the training set and, correspondingly, a significantly better prediction of the binding affinities of the test set compounds (see below). Equation (2) is obtained if compounds #1–12 and #23–31 are used as the training set ($n = 21$).[24]

$$\log 1/\mathrm{CBG} = 1.667(\pm 0.75)4,5\text{-C=C-} + 5.306(\pm 0.65)$$
$$(n = 21; r = 0.731; s = 0.697; F = 21.82;$$
$$Q^2 = 0.454; s_{\mathrm{PRESS}} = 0.754) \quad (2)$$

Despite the worse fit and internal predictivity, as compared with equation (1), the validity of this model is proven by its excellent test set (compounds #13–22) predictivity: $r^2_{\mathrm{pred}} = 0.909$; $s_{\mathrm{PRESS}} = 0.406$). The differences between both models, especially in their test set predictivity, provide striking evidence for the influence of the training and test set selections on the obtained results. Thus, a careful selection of the training set molecules is of utmost importance. A broad variety of structural features should be included in these molecules, in order to allow reliable predictions for the test set compounds.

4 PHARMACOPHORE HYPOTHESES AND ALIGNMENT

The specific interaction of drugs with proteins depends on a structural complementarity between the ligand and its binding site, in the 3D arrangement of all relevant molecular properties. Pharmacophores are 3D models of such structural features (Figure 2), in the simplest case a 2D three-point pharmacophore.

The strategy applied in Figure 2 is called the 'active analog approach'[4,5,10]; flexible compounds are compared with rigid analogs, in order to determine which geometry of a flexible ligand corresponds to the biologically active conformation. Any rigidization of a flexible drug in a wrong geometry leads to inactive molecules. On the other hand, experience shows that freezing the bioactive conformation of a ligand may lead to superactive and highly selective analogs.[1–3]

As already mentioned (Section 1), enantiomers of chiral compounds differ in their biological effects. This can be easily illustrated by the different odors of the closely related monoterpenes (*R*)- and (*S*)-limonene and (*R*)- and (*S*)-carvone (**9a,b** and **10a,b**; Figure 3), which result from the stereospecific interaction of these compounds with the olfactory G-protein-coupled receptors.

With respect to their relative affinities to a chiral binding site, all different pairs of enantiomers differ more or less in their relative affinities. Eudismic ratios, i.e., the ratio of the affinity or biological activity of the 'more' active analog (the eutomer) to the 'less' active one (the distomer) between 1 and 500 000 have been observed (e.g., Refs. 3, 25, 26).

Whereas the alignment of compounds **1** and **2** (Section 2.3) seems to be obvious, there are more complex situations which demand a detailed analysis of the functional group similarity of the compounds in different orientations, as is e.g., the case for dihydrofolate **11** and methotrexate **12** (Figure 4).[3,5,10,27]

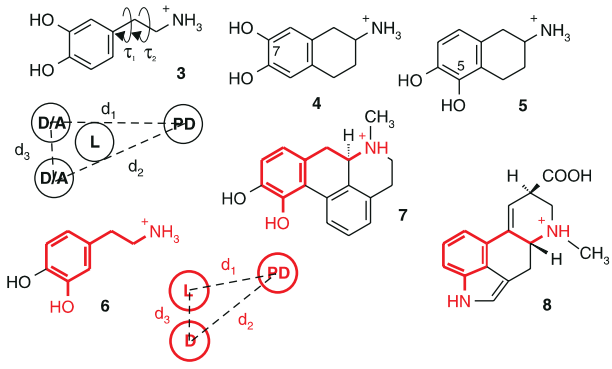


Figure 2 Dopamine **3** is a neurotransmitter with two phenolic hydroxy groups, a phenyl ring and a positively charged ammonium group, separated from the aromatic ring by two carbon atoms. Correspondingly, a pharmacophore model can be defined which contains two donor/accptor groups D/A, a positively charged donor group PD, and a lipophilic area L (model below structure **3**). However, dopamine has two flexible bonds. Rotation around the angles τ_1 or τ_2 leads to other geometries, with different distances d_1 and d_2. To differentiate between the rotamers, i.e., between different conformations, compounds **4** and **5** (racemic mixtures) have been investigated. In vivo, compound **4** is the more active analog. If both compounds are, however, directly injected into dopamine-receptive brain areas of the rat, the 5,6-dihydroxy analog **5** is 100 times more active, according to the dopamine conformation presented in **6**. Accordingly, apomorphine **7** and some analogs and derivatives of lysergic acid **8** are dopamine agonists. Under the assumption of similar binding modes, the pharmacophore geometries shown in red allow a mutual alignment of compounds **5–8**

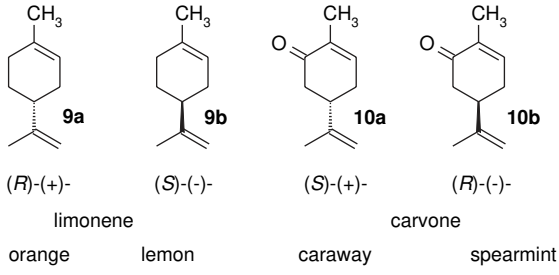


Figure 3 Enantiomers of chiral drugs show different biological properties. The monoterpenes (*R*)- and (*S*)-limonene (**9a**, **b**) and (*R*)- and (*S*)-carvone (**10a**, **b**) interact with the olfactory receptors (which are, like many drug receptors, G-protein-coupled receptors) in a different manner, producing different odors that are indicated under the individual formulas

A common pharmacophore within a series of compounds does not necessarily mean that all compounds need to have identical molecular frames. One of the most important advantages of the CoMFA method results from the fact that molecules with identical pharmacophores but different atom connectivities can be combined in one analysis.

The superposition of all molecules is performed according to the equivalent functional groups that are identified as the pharmacophore, either by hand or by an appropriate field fit.[10,28] A valuable tool for the superposition of molecules within a congeneric series of compounds is the program SEAL.[29] It allows the definition of a 'similarity index' A_{F} (equation 3) between two molecules A and B in any relative orientation to each other. For this purpose, atomic similarity

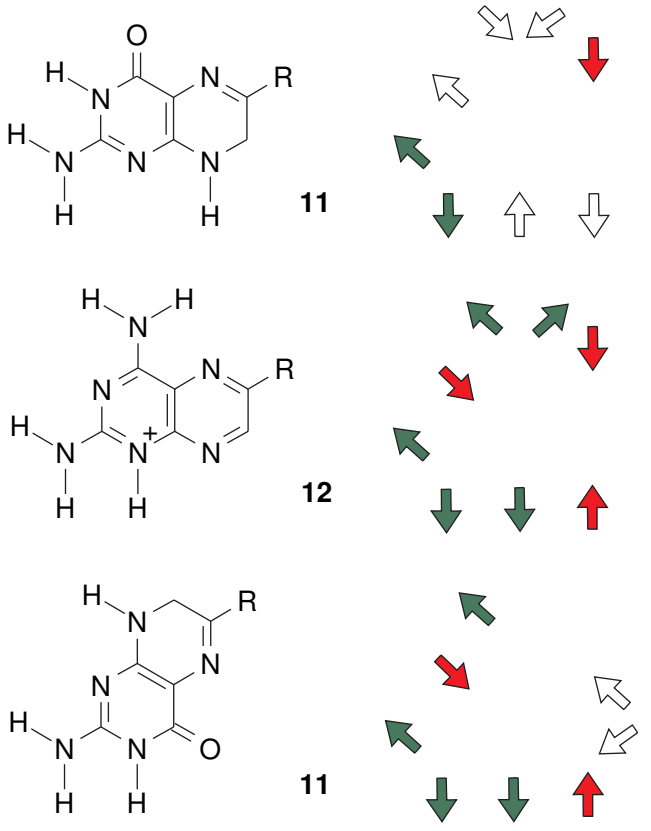


Figure 4 If dihydrofolate **11** (upper orientation) and methotrexate **12** (middle) are superimposed according to the heteroatom positions in the pterine rings (R = *p*-aminobenzoylglutamic acid), they differ in their hydrogen bond donor and acceptor properties. If **11** is rotated around the C–R bond (below), a much better similarity in the hydrogen bond patterns is observed (six accordances instead of three); identical directions of hydrogen bond donors and acceptors are indicated by filled red (hydrogen bond acceptors) and green arrows (hydrogen bond donors). The binding mode of **12** to the enzyme dihydrofolate reductase (DHFR) was predicted from the 3D structure of the DHFR complex with **11** and later confirmed by a protein crystallographic study[27]

values are calculated between each of the m atoms of molecule A to each of the n atoms of molecule B. The sum of these values over all atom pairs of A and B defines the similarity index A_F. In equation (3) r_{ij} is the distance between atoms i and j, α is a user-defined value, w_E and w_S are user-attributed values to give different weights for electrostatic and steric overlap, q_i and q_j are the partial charges at atoms i of molecule A and j of molecule B, and v_i and v_j are arbitrary powers (default = 1) of the van der Waals, radii of atoms i and j. Any other atomic property, e.g., hydrophobicity, might be added to the definition of w_{ij}.[29,30]

$$A_F = -\sum_{i=1}^{m}\sum_{j=1}^{n} w_{ij} e^{-\alpha r_{ij}^2}; \quad w_{ij} = w_E q_i q_j + w_S v_i v_j + \cdots \qquad (3)$$

Equation (3) is a bell-shaped Gaussian function. Because of the exponential distance dependence, the highest 'similarity' is achieved if all atoms and corresponding functionalities of both molecules are 'closest' to each other. Within a certain distance, the similarity value between two atoms increases significantly if both atoms come closer together (cf. Figure 7, Section 5). If both atoms are already close enough, small shifts are well tolerated to achieve also a good superposition of some other atoms. Even automated superpositions, starting from arbitrary positions of both molecules and without defining any pharmacophore hypotheses and orientation rules, can be performed with this program.[30]

5 BOX, GRID SIZE, AND 3D FIELD CALCULATIONS

After superposition of the molecules, a rectangular box is placed around all molecules, keeping a minimum distance of a few Å around the structures. A grid distance (default value = 2.0 Å) is selected to generate points at the intersections of a regular 3D lattice. According to the dimensions of the box and the chosen grid distance, normally a few to several thousand points are generated (Figure 5).

A distance of 2 Å seems to be an extremely wide grid distance if one considers that even a few tenths of an Å are responsible for the difference between van der Waals, attraction and strong repulsion between two atoms. However, this distance is dictated by the exponential increase of the computational effort if smaller grid distances are chosen.

Either before the generation of the different conformations of the molecules, or at the latest now, atomic partial charges are determined for all analogs, preferably by a semiempirical method, such as AM1, PM3, or MNDO. These charges are used to calculate, separately for every molecule, the electrostatic field values in all grid points, using a charged atom as a probe. In addition, steric fields are calculated, using a neutral atom. The electrostatic and steric field values in the individual grid points are based on the coulomb potential function (equation 4; E_C = coulomb interaction energy, q_i = partial charge of atom i of the molecule, q_j = charge of the probe atom, D = dielectric constant, r_{ij} = distance between atom i of the molecule and the grid point j, where the probe atom is located) and the Lennard-Jones potential function (equation 5; E_{vdW} = van der Waals, interaction energy, r_{ij} = distance between atom i of the molecule and the grid point j where the probe atom is located; A_{ij} and C_{ij} are constants that depend on the van der

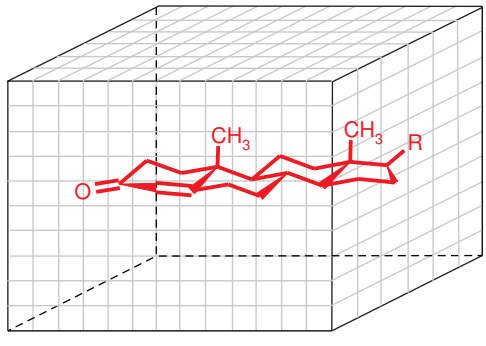


Figure 5 A steroid molecule (cf. formula **1**, Section 2.3) in a box with a regular grid. For better presentation, only the grid lines at the surface of the box and only one molecule are shown instead of a superposition of all molecules; despite the fact that this box is much smaller than in most CoMFA studies, it already includes $14 \times 11 \times 7 = 1078$ grid points

Waals, radii of the corresponding atoms), respectively.[10,11]

$$E_{\mathrm{C}} = \sum_{i=1}^{n} \frac{q_i q_j}{D r_{ij}} \tag{4}$$

$$E_{\mathrm{vdW}} = \sum_{i=1}^{n} (A_{ij} r_{ij}^{-12} - C_{ij} r_{ij}^{-6}) \tag{5}$$

In close proximity to the surface of the atoms both potentials have very steep slopes. They approach infinite values if the atom positions of two molecules overlap. To avoid this, arbitrary cut-offs are defined and all larger positive (or negative) values are set to these cut-off values (Figure 6).

In addition or alternatively to these fields, hydrophobic fields, calculated e.g., by the program HINT,[10,31] or GRID fields[10,32] can be used. Arbitrary weights may be attributed to the different fields. An appropriate scaling of all variables has to be performed if additional properties, e.g., the lipophilicity parameter log P ($P = n$-octanol/water partition coefficient),[4–8,33] are included, to give a comparable weight to the individual fields and the single parameter(s).

A recently developed CoMFA version, CoMSIA,[19,20] calculates SEAL similarity fields. In this modification, probes are used to calculate their 'similarity indices' to the investigated molecules in the different grid points. These fields are then correlated with the biological activity values, as in CoMFA. The most important advantage of the SEAL fields is their 'smooth' nature (Figure 7). The slopes of the underlying Gaussian functions are not as steep as the Coulomb and Lennard-Jones potentials; therefore, no cut-off values need to be defined. Even in the case of overlapping atoms, values within a reasonable range result from these functions.

Naturally, there are many grid points with only minor variation in the field values, e.g., the steric potential within the common overlap volume of all molecules or the steric and electrostatic potentials far outside the molecules. To eliminate such grid points, a 'minimum sigma' condition is defined, i.e., all points are eliminated which have a lower variance in their field values than defined by this minimum sigma option.[10]

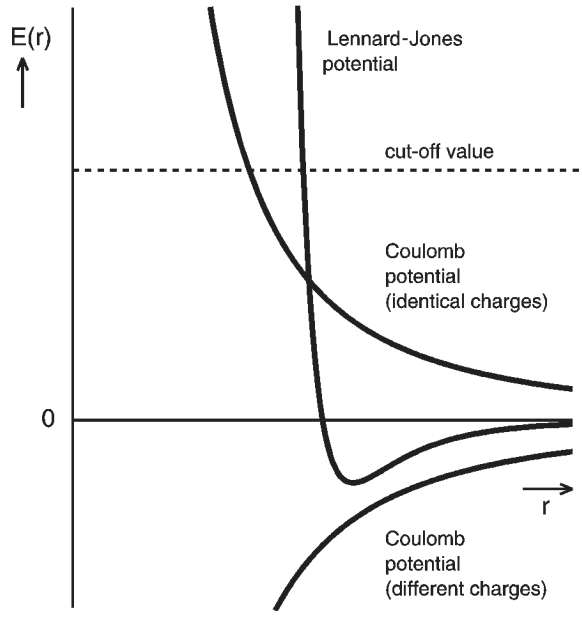


Figure 6 Electrostatic and steric fields in CoMFA studies are calculated from Coulomb and Lennard-Jones potentials, respectively. Because of the steep slopes of these functions, cut-off values define the upper limits (and lower limits of the coulomb potential; not shown in the diagram) of individual grid values (redrawn from Ref. 3 with kind permission from Spektrum Akademischer Verlag, Heidelberg)

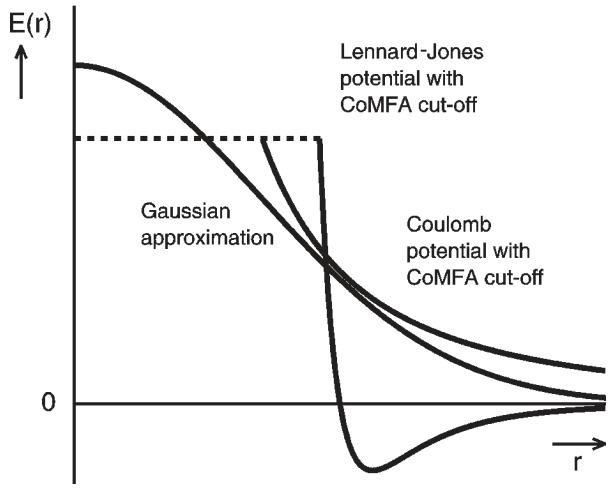


Figure 7 The bell-shaped Gaussian functions of SEAL fields are good approximations of Lennard-Jones and Coulomb potentials that are limited by cut-off values, with the advantage that they are 'smooth' functions (redrawn from Ref. 3 with kind permission from Spektrum Akademischer Verlag, Heidelberg)

6 DERIVATION AND VALIDATION OF 3D QSAR MODELS

Because of the enormous number of x variables that are generated in the field calculations, regression analysis cannot be applied. In the very beginning of 3D QSAR studies, principal components were derived from the **X** block (i.e., the table of field values) and then correlated with the biological activity values.[10] In 1986, Svante Wold proposed to use PLS analysis. PLS analysis[5,10,34–37] resembles principal component regression analysis in its derivation of vectors from the **Y** and the **X** blocks. However, there is a fundamental difference: in PLS analysis, the orientation of the so-called ***u*** and ***t*** vectors does not exactly correspond to the orientation of the principal components. They are slightly skewed within their confidence hyperboxes, in order to achieve a maximum intercorrelation (Figure 8).

SAMPLS is a modification of PLS analysis. In SAMPLS, the PLS vectors, also called latent variables, are derived from the $n \times n$ covariance matrix.[38] Whereas SAMPLS has no major advantages, as compared with ordinary PLS analysis, it operates a few to several orders of magnitude faster in cross-validation runs (see below), owing to a much smaller number of arithmetic operations. SAMPLS is only one example of so-called kernel algorithms; other modifications, being applicable to data sets with several different ***y*** vectors, have been described (e.g., Refs. 19, 39–42).

As in regression analysis, in PLS analysis the correlation coefficient r also increases with the number of extracted vectors. Dependent on the number of components, often perfect

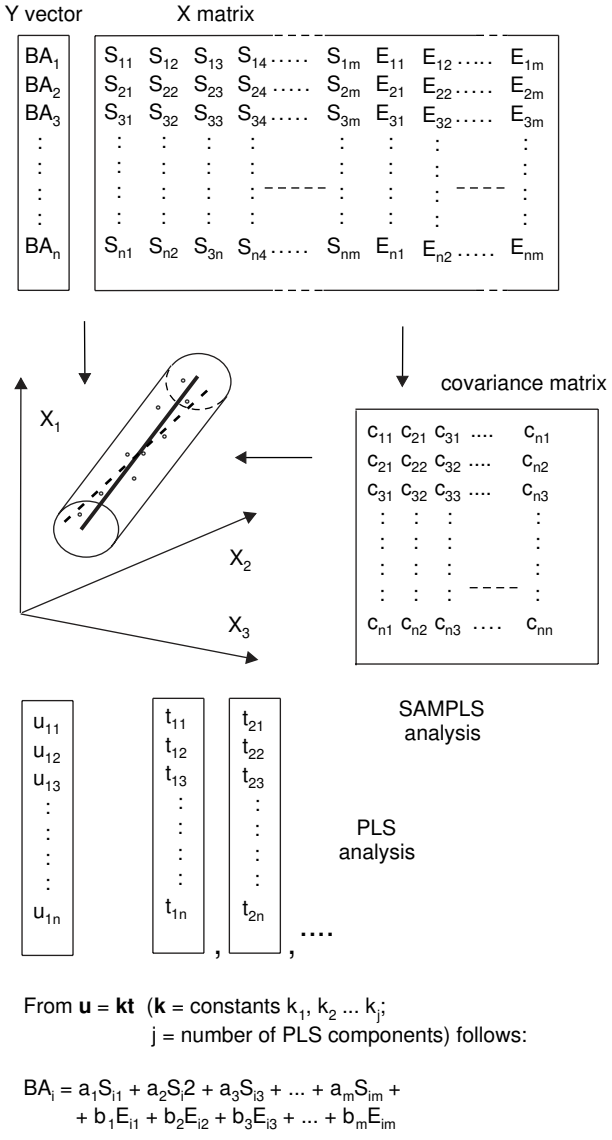


Figure 8 PLS analysis derives vectors $\boldsymbol{u}$ and $\boldsymbol{t}$ from the **Y** block (or $\boldsymbol{y}$ vector; BA_i = logarithms of relative affinities or other biological activities) and the **X** block (S_{ij} = steric field variable of molecule i in the grid point j; E_{ij} = electrostatic field variable of molecule i in the grid point j) that are related to principal components. These 'latent variables' are skewed within their confidence hyperboxes to achieve a maximum intercorrelation (diagram). SAMPLS is a PLS modification which first derives the covariance matrix of the **X** block and then the PLS result from this covariance matrix. Especially in cross-validation (see below), SAMPLS analysis is much faster than ordinary PLS analysis

correlations are obtained in PLS analyses, owing to the large number of x variables. Correspondingly, the goodness of fit is no criterion for the validity of a PLS model. The significance of additional PLS vectors is determined by cross-validation.[10,35–37,43] In the most common leave-one-out cross-validation, one object (i.e., one biological activity value) is eliminated from the training set and a PLS model is derived from the residual compounds. This model is used to predict the biological activity value of the compound which was not included in the model. The same procedure is repeated after elimination of another object until all objects have been eliminated once. The sum of the squared differences, $\text{PRESS} = \Sigma(y_{\text{pred}} - y_{\text{obs}})^2$, between these 'outside-predictions' and the observed y values is a measure for the internal predictivity of the PLS model. For larger data sets, an alternative to the leave-one-out technique is recommended to yield more stable PLS models. Several objects are eliminated from the data set at a time, randomly or in a systematic manner, and the excluded objects are predicted by the corresponding model.[10,43]

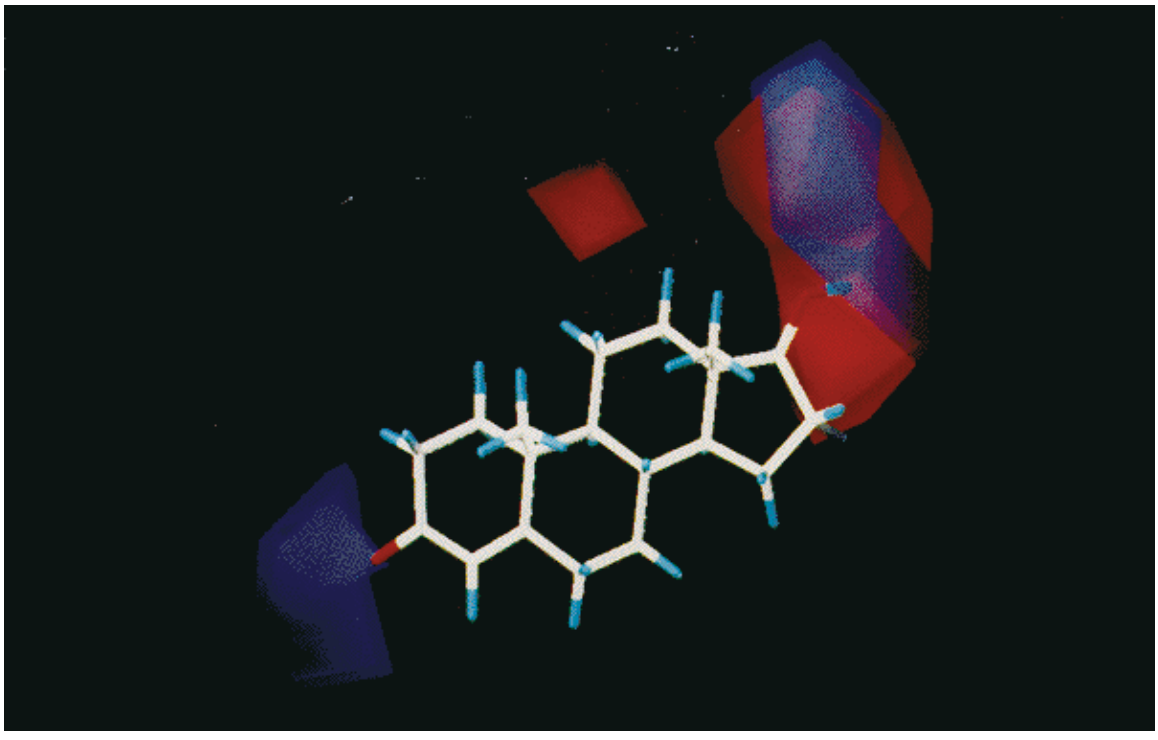
In cross-validation, a Q^2 (r^2_{PRESS}) value is defined like r^2 in regression and PLS analysis, using PRESS instead of the unexplained variance $\Sigma(y_{\text{calc}} - y_{\text{obs}})^2$. Cross-validated Q^2 values are always smaller than the r^2 values, including all objects (r^2_{FIT}). As long as only significant PLS vectors are derived, Q^2 increases, whereas decreasing Q^2 values indicate overprediction. In severe cases of overprediction PRESS may become even larger than the overall variance of the y values; then negative Q^2 values are obtained, indicating that the predictions from the model are worse than 'no model', i.e., taking the y_{mean} values as 'predictions'.[10,35,43] The significance of leave-one-out cross-validation results has to be commented on: in highly redundant data sets, where all or at least most objects have close neighbors in multidimensional parameter space, this procedure gives a much too optimistic result.

The standard deviation of predictions, s_{PRESS}, is calculated from PRESS, the sum of the squared errors of these predictions, considering the number of degrees of freedom. SDEP (the standard deviation of the error of predictions)[10] corresponds to s_{PRESS} but the number of degrees of freedom is not considered in the calculation of the SDEP value. The smallest s_{PRESS} or SDEP value should be taken as the criterion for the optimum number of components. Alternatively, an increase of the Q^2 value by a certain percentage, e.g., 5%, may be defined as the criterion to accept a further PLS component. As long as only significant components are extracted in the PLS analysis, PRESS, SDEP and s_{PRESS} will decrease; if too many components are derived, overprediction results and PRESS, SDEP and s_{PRESS} increase.

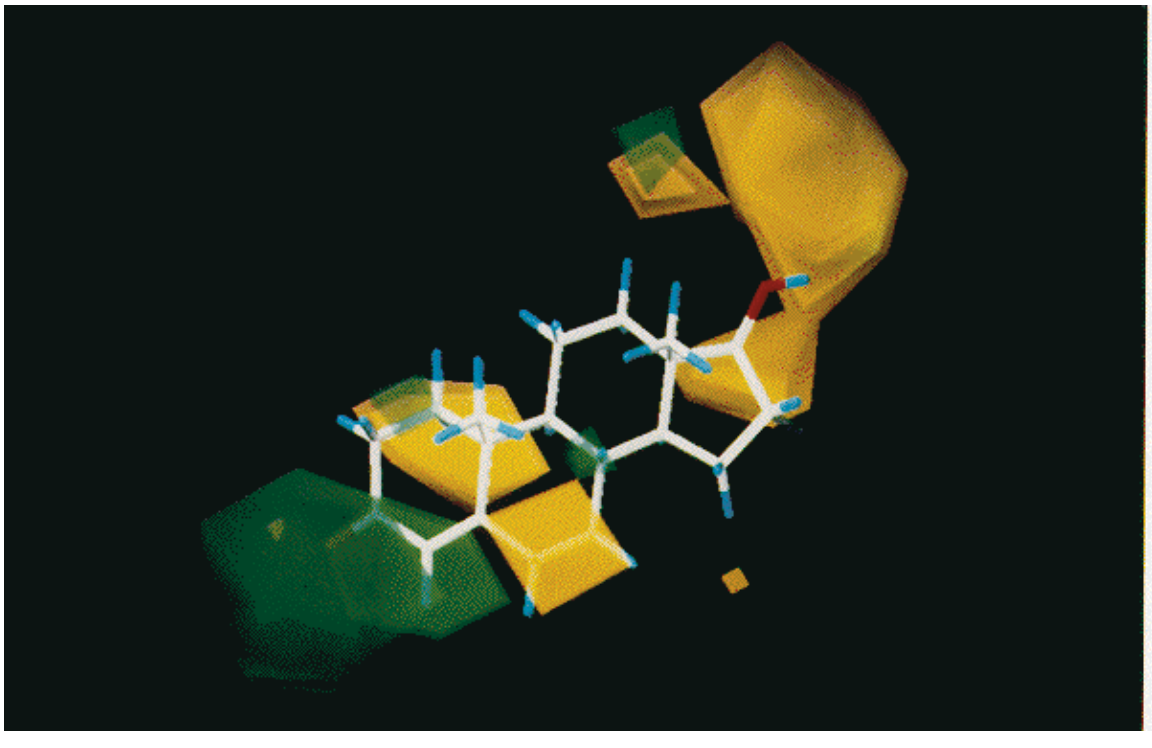
Bootstrapping is a procedure in which n random selections out of the original set of n objects are performed several times to simulate different samplings from a larger set of objects. In each run some objects may not be included in the PLS analysis, whereas some others might be included more than once. Confidence intervals for each term can be estimated from such a procedure, giving an independent measure of the stability of the PLS model.[10,36,43]

A rigorous alternative to cross-validation and bootstrapping demands a repetitive scrambling of the y values. Only if the results from a PLS model, using the original order of the y values, is significantly better than the results from the 'scrambled' models, using randomly ordered y values in the PLS analysis, can one be sure that a relationship indeed exists between the biological data and the **X** block.

Although the PLS method is claimed to be a robust modeling technique, experience shows that too many noise variables, i.e., variables that do not contribute to prediction, obscure the result. For prediction such additional variables are most often useless or even detrimental. The problem of perturbation of CoMFA models by many irrelevant grid points has been approached by developing a variable selection procedure, called GOLPE (generating optimal linear PLS estimations).[10,36,37] In GOLPE, first a D-optimal design is used



(a)



(b)

Figure 9 3D contour maps around testosterone (**1**, Section 2.3; $R_1 = OH$, $R_2 = H$) as the result of a CoMFA analysis of the TBG affinities of different steroids.[11] (a) The color codings indicate regions where electronegative substituents enhance (blue) or reduce (red) the binding affinity. (b) Regions where substitution enhances (green) or reduces (yellow) the binding affinity (reproduced from Ref. 13 with kind permission from VCH, Weinheim)

to preselect nonredundant variables. A subset of variables, having a higher degree of orthogonality in multidimensional parameter space, is selected by this procedure. Then a fractional factorial design is used to run PLS analyses with different combinations of these variables. The predictive abilities of the models are checked by a cross-validation procedure. The influence of every variable can be estimated from a comparison of the PRESS values of the models including this variable and those not containing it. Explanatory variables, i.e., grid points that significantly contribute to prediction, are kept, all others are eliminated.

The results of a PLS analysis can be transformed to regression coefficients of the **X** block variables that are used for the calculation and prediction of biological activity values. Because of the large number of regression coefficients, a direct interpretation of the corresponding equation is impossible. An appropriate way to visualize the results is the generation of contour maps which show the volumes of regions that are larger or smaller than certain user-defined positive or negative values (Figures 9a and 9b).

7 SOME PRACTICAL PROBLEMS

Because of the complexity of the CoMFA method, many different problems arise in running the analyses and in interpreting the results obtained (see e.g., Refs. 10, 11, 15–19).

The search for the 'bioactive' conformation and the common pharmacophore constitutes a serious problem. It is one of the most important sources of wrong conclusions and errors in all CoMFA studies. The risk of deriving irrelevant geometries can only be reduced by a consideration of rigid analogs. Even then, the alignment poses problems because there are many cases of different binding modes of seemingly closely related analogs. The more X-ray structures of ligand–protein complexes are determined, the larger becomes the number of examples where such unexpected binding modes are observed.[3,5,10,44]

Another possible source of error is the mutual alignment of all molecules. Because the training set contains active and inactive molecules, too much weight might be attributed to the 3D structures of the inactive compounds in the mutual alignment. CoMFA models have been refined by a re-alignment of the training set molecules, guided by the first, preliminary CoMFA results; molecules for which too low activities are calculated, are reoriented to achieve an 'improved' (but more subjective) alignment.[45] The definition of pairwise similarity indices of molecules, derived from 3D fields, would allow pairwise alignments, without any consideration of all other molecules;[24] the disadvantage of this approach is the loss of the 3D information for predictions.

Of course, any risks in the generation of the conformations and in the alignment could be avoided by looking at the 3D structures of ligand–protein complexes which are derived from X-ray crystallographic or multidimensional NMR studies.[46,47] The advantage of such a procedure is obvious but the question arises: is CoMFA really the adequate tool for such cases?

The problems of inadequate training and test set selections have already been discussed in detail in Section 3. The same problem is also observed in cross-validation. In well-designed training sets, where a small number of objects is selected to explore the parameter space with a minimum number of experiments, cross-validation fails: the eliminated objects cannot be predicted by models which are derived from objects that do not contain all structural features of the excluded objects.

As already discussed, the functions that are used in CoMFA studies create relatively 'hard' fields (cf. Figure 6, Section 5). Especially the variables of the steric fields sometimes show only values close to zero (no atoms around) or at the cut-off value (inside the molecules). Correspondingly, contour maps are most often fragmented and difficult to interpret, especially if a variable selection procedure has been applied in the analysis.[48]

The systematic investigation of the dependence of CoMFA results on the box orientation has led to a CoMFA modification which partitions the box into regular volumes.[49] Instead of single variable selection, a regional selection is performed: only regions significantly contributing to fit and internal predictivity are selected for the further analysis. Single and domain variable selection procedures have received critical comment.[50] More promising seems to be a newly described GOLPE-guided region selection procedure, where irregular 'relevant' regions are selected.[51]

Contour maps from CoMSIA studies, using the much 'smoother' Gaussian SEAL fields (cf. Figure 7, Section 5), seem to produce also smoother, more coherent fields that are easier to interpret.[20]

8 CoMFA APPLICATIONS IN DRUG DESIGN

There are now a few hundred practical applications of CoMFA in drug design. Most applications are in the field of ligand–protein interactions, describing affinity or inhibition constants. In addition, CoMFA has been used to correlate steric and electronic parameters.[10] Less appropriate seems the application of CoMFA to in vivo data, even if lipophilicity is considered as an additional parameter. As most CoMFA applications in drug design have been comprehensively reviewed in three books[10,19] and in some reviews,[17,18] Table 1 gives only an overview of some typical applications that have been reported in the last few years.

9 CONCLUSIONS

CoMFA is a powerful 3D QSAR method which has already shown its practical value in many cases. The results obtained depend a lot on the care that is taken in the definition of the 3D pharmacophore and in the alignment of the molecules. Soft fields seem to be better than hard fields, especially for the interpretation of the contour maps. Variable selection seems unnecessary if such fields are used. On the other hand, the new GOLPE-guided regional selection[51] looks more promising than other variable selection procedures. A general problem is the external predictivity of QSAR models. The better the training set is described, the worse most often is the prediction of the test set compounds (compare, e.g., equations 1 and 2[24] and Ref. 50), a fact which obviously largely depends on the training and test set selections.

Further CoMFA modifications, where $n \times n$ similarity index matrices are calculated from molecular fields and these matrices are correlated with biological data by PLS

Table 1 Overview on some Typical CoMFA Applications in Drug Design, Abstracted in the years 1993–1996

Biological System	Compounds	Source
Enzymes		
Acetylcholinesterase inhibition	*N*-Benzylpiperidines	225/1996
Angiotensin-converting enzyme (ACE) inhibition	Various ACE inhibitors	62/1994 64/1994
Aromatase inhibition	Fadrozole analogs	410/1996
α-Chymotrypsin binding	*N*-Acyl amino acid esters	305/1995
Cytochrome P450 2A5 binding	Coumarines	144/1996
Dihydrofolate reductase inhibition	*N*-Phenyltriazines	70/1995 222/1996
Dipeptidyl peptidase IV inhibition	Dipeptide analogs	338/1996
Glycogen phosphorylase b inhibition	Glucose analogs	306/1995
HIV integrase inhibition	Flavones	543/1995
HIV protease inhibition	Transition state inhibitors	361/1994 219/1995
HIV protease inhibition	Statine derivatives	224/1996
Lanosterol-14α-demethylase inhibition	Azoles	411/1996
Monoaminoxidase A, B inhibition	Indoles	492/1996
Monoaminoxidase B inhibition	Indenopyridazines	143/1996
Papain binding	Phenyl hippurates	364/1994 491/1994
Phenethanolamine *N*-methyltransferase	Benzylamines and cyclic analogs	502/1994
Renin inhibition	Transition state inhibitors chloromethylketones	326/1993
Thermitase inhibition	Peptide methyl- and chloromethylketones	226/1996
Thermolysin inhibition	Transition state inhibitors	326/1993 62/1994 64/1994 304/1995
Topoisomerase II inhibition	Podophyllotoxin analogs	491/1996
Receptors		
5-HT_1 (serotonin) receptor binding	Tetrahydropyridinylindoles	215/1994
5-HT_{1A} (serotonin) receptor binding	Benzodioxanes and -furanes	308/1995
5-HT_{1A} (serotonin) receptor binding	Various serotonin analogs	327/1993 223/1996
5-HT_3 (serotonin) receptor binding	Arylpiperazines	41/1996
α_1-Adrenergic antagonism	Prazosin analogs	365/1994 495/1994
AII (angiotensin) receptor antagonism	Biphenyltetrazoles	397/1995
Androgen receptor binding	Steroids	405/1996
BZ (benzodiazepine) receptor binding	Benzodiazepines	214/1994 219/1994
BZ (benzodiazepine) receptor binding	Benzothiazepinone analogs	311/1995
BZ (benzodiazepine) receptor binding	β-Carbolines	328/1993
CCK A (cholecystokinin) receptor antagonists	Benzodiazepines	500/1994
D_1 (dopamine) receptor antagonists	Tetrahydroisoquinoline analogs	398/1995
D_{1A} (dopamine) receptor binding	Structurally diverse dopamine receptor agonists	337/1996
Estrogen receptor binding	Halogenated estradiol analogs	309/1995
Estrogen receptor binding	Polychlorinated hydroxybiphenyls	339/1996
ET_A (endothelin receptor) antagonism	Arylsulfonamides	483/1995
Gonadotropin hormone release inhibition	Somatostatin analogs	544/1995
NK_1 (neurokinin) receptor antagonism	Tachykinin fragment analogs	142/1996
N-Methyl-D-aspartate (NMDA) receptor binding	Quinoline- and pyridine-carboxylic acids	469/1993

Table 1 *(continued)*

Biological System	Compounds	Source
Purinoceptor antagonism	ATP analogs	409/1996
Morphine σ_1 receptor binding	*N*-Normetazocine analogs	482/1995
Morphine σ_3 receptor binding	Phenylaminotetralin analogs	312/1995
Soluble and membrane transporters		
Corticosteroid and testosterone binding globulins	Steroids	304/1995 481/1995
Dopamine transporter binding	Tropane carboxylic acid esters	496/1994 310/1995
Ion channels		
Calcium channel agonism	Bay K 8644 analogs	499/1994 68/1995
Chloride influx inhibition	1-Oxo-isoindoles	217/1995
Miscellaneous activities		
Ames test, mutagenic activity	lactones	72/1995 493/1996
Anti-HIV-1 activity	Quinolines	40/1996
Anti-HIV gp 120 activity	Porphyrin derivatives	74/1995
Anti-plasmodium (antimalarial) activity	Artemisinine analogs	363/1994
Antitumor activity, in mice	Thioxanthenones	73/1995
Antitumor activity, in vitro (L1210, HCT-8)	Pyrazoloacridines	216/1994
Cytosolic Ah (dioxine) receptor stimulation	Halogenated dibenzodioxins, -furans, and biphenyls	187/1993
Genotoxicity, in vitro	Nitrofurans	467/1993
Human leukemia cell differentiation	Alkylamides	468/1993
Human rhinovirus inhibition	Substituted aryl-oxazolines	326/1993
Inhibition of protein biosynthesis	Cephalotoxine esters	481/1995
Microtubule (tubulin) binding	Taxol analogs	71/1995

The sources refer to the Abstracts Section of the journal *Quantitative Structure–Activity Relationships*; the number of the abstract of e.g., 225/1996 is 225 and the year of its publication is 1996 (years 1993–1996 correspond to *Quant. Struct.–Act. Relat.*, volumes 12–15)

analysis[5,21–23] or by variable selection–regression analysis,[24] have not been discussed here. Although they are 3D QSAR approaches, no contour maps can be derived. The same restriction applies to CoMFA studies when only the SAMPLS version of PLS analysis is used.

Recommendations for CoMFA studies and 3D QSAR publications have been defined.[10,52] These recommendations should help to avoid the most common errors and pitfalls and should ease the reproduction of CoMFA results by other scientists; in a short version they are summarized below.

1. The selection of starting geometries should be rationalized.
2. Methods of geometry optimization should be documented.
3. Charges used in CoMFA and their calculation method should be defined.
4. Alignment criteria and all options (box, grid size, etc.) should be given.
5. Scaling and weighting of fields should be documented.
6. Cross-validation runs should be performed for every analysis.
7. Statistical data for fit and internal predictivity should be given.
8. The number of (significant) PLS components must be presented.
9. Typical problems in cross-validation should be considered.
10. Removed outliers should be mentioned and discussed.
11. Variable selection procedures should be used whenever appropriate.
12. Contour maps of the final model should be provided or at least discussed.
13. Origins of biological data should be documented.
14. Standard errors of biological data should be given.
15. A table with all observed vs. predicted values should be provided.
16. Coordinates of the molecules in the used alignments should be available.
17. Predictions of biological activity values depend on the training set.

10 NOTES

Reviews and books have been cited in most cases in addition to or instead of the original references, in order to provide the corresponding results in their context to related work in the same field.

The journal *Quantitative Structure–Activity Relationships* publishes, in addition to original contributions, every year about 500–600 detailed reviews on scientific papers in the fields of QSAR, 3D QSAR and molecular modeling (compare Table 1).

A discussion forum for QSAR researchers is the WWW home page of *The QSAR and Modelling Society* at **http://www.pharma.ethz.ch/qsar**. Not only names and e-mail addresses of QSAR colleagues but also tips and tricks, information on recent books and software, and links to related topics can be found there.

11 RELATED ARTICLES

Chemometrics: Multivariate View on Chemical Problems; Linear Free Energy Relationships (LFER); Partial Least Squares (PLS) in Chemistry; QSAR in Drug Design.

12 REFERENCES

1. M. E. Wolff, (ed.), 'Burger's Medicinal Chemistry', **Vol. I**: *Principles and Practice*, Wiley, New York, 5th edn., 1995.
2. C. G. Wermuth, (ed.), 'The Practice of Medicinal Chemistry', Academic Press, London, 1996.
3. H.-J. Böhm, G. Klebe, and H. Kubinyi, 'Wirkstoffdesign', Spektrum Akademischer, Heidelberg, 1996.
4. C. A. Ramsden, (ed.), 'Quantitative Drug Design', Vol. 4, of 'Comprehensive Medicinal Chemistry', eds. C. Hansch, P. G. Sammes, and J. B. Taylor, Pergamon, Oxford, 1990.
5. H. Kubinyi, 'QSAR: Hansch Analysis and Related Approaches', VCH, Weinheim, 1993.
6. H. Kubinyi, in 'Burger's Medicinal Chemistry', Vol. I, ed. M. E. Wolff, Wiley, New York, 5th edn., 1995, pp. 497–571.
7. C. Hansch and A. Leo, 'Exploring QSAR. Fundamentals and Applications in Chemistry and Biology', American Chemical Society, Washington, DC, 1995.
8. H. van de Waterbeemd, (ed.), 'Structure–Property Correlations in Drug Research', Academic Press, Austin, TX, 1996.
9. R. D. Cramer, III and M. Milne, *Abstracts of Papers of the Am. Chem. Soc. M.*, April 1979, Computer Chemistry Section, no. 44.
10. H. Kubinyi, (ed.), '3D QSAR in Drug Design. Theory, Methods and Applications', ESCOM, Leiden, 1993.
11. R. D. Cramer, III, D. E. Patterson, and J. D. Bunce, *J. Am. Chem. Soc.*, 1988, **110**, 5959–5967.
12. SYBYL/QSAR, Molecular Modelling Software, Tripos Inc., 1699 S. Hanley Road, St. Louis, MO 63944, USA.
13. H. Kubinyi, *Chemie in unserer Zeit*, 1994, **23**, 281–290.
14. S. M. Green and G. R. Marshall, *Trends Pharm. Sci.*, 1995, **16**, 285–291.
15. K. H. Kim, in 'Molecular Similarity in Drug Design', ed. P. M. Dean, Chapman & Hall, New York, 1995, pp 291–331.
16. Y. C. Martin and C. T. Lin, in 'The Practice of Medicinal Chemistry', ed. C. G. Wermuth, Academic Press, London, 1996, pp. 459–483.
17. C. J. Blankley, In 'Structure–Property Correlations in Drug Research', ed. van de H. Waterbeemd, Academic Press, Austin, TX, 1996, pp. 111–177.
18. Y. C. Martin, K.-H Kim, and C. T. Lin, in 'Advances in Quantitative Structure–Property Relationships', Vol. I; ed. M. Charton, JAI Press, Greenwich, CT, 1996, pp. 1–52.
19. H. Kubinyi, G. Folkers, and Y. C. Martin, (eds.), '3D QSAR in Drug Design', Vols. 2 and 3, Kluwer, Dordrecht, 1998.
20. G. Klebe, U. Abraham, and T. Mietzner, *J. Med. Chem.*, 1994, **37**, 4130–4146.
21. A. C. Good, S.-S. So, and W. G. Richards, *J. Med. Chem.*, 1993, **36**, 433–438.
22. A. C. Good, S. J. Peterson, and W. G. Richards, *J. Med. Chem.*, 1993, **36**, 2929–2937.
23. A. C. Good, in 'Molecular Similarity in Drug Design', ed. P. M. Dean, Chapman & Hall, New York, 1995, pp. 24–56.
24. H. Kubinyi, in 'Computer-Assisted Lead Finding and Optimization' Proc. 11th European Symp. on Quantitative Structure–Activity Relationships, Lausanne, 1996, eds. van de H. Waterbeemd, B. Testa, and G. Folkers, Verlag Helvetica Chimica Acta and VCH: Basel, Weinheim, 1997; pp 7–28.
25. B. Holmstedt, H. Frank, and B. Testa, 'Chirality and Biological Activity', New York, 1990.
26. E. J. Ariëns, *Trends Pharmacol. Sci.*, 1993, **14**, 68–75.
27. C. Bystroff, S. J. Oatley, and S. J. Kraut, *Biochemistry*, 1990, **29**, 3263–3277.
28. M. Clark, R. D. Cramer, III, D. M. Jones, D. E. Patterson, and P. E. Simeroth, *Tetrahedron Comput. Methodol.*, 1990, **3**, 47–59.
29. S. K. Kearsley and G. M. Smith, *Tetrahedron Comp. Methodol.*, 1990, **3**, 615–633.
30. G. Klebe, T. Mietzner, and F. Weber, *J. Comput.-Aided Mol. Design*, 1994, **8**, 751–778.
31. F. C. Wireko, G. E. Kellog, and D. J. Abraham, *J. Med. Chem.*, 1991, **34**, 758–767.
32. P. Goodford, *J. Chemometrics*, 1996, **10**, 107–117.
33. V. Pliska, B. Testa, and H. van de Waterbeemd, (eds.), 'Lipophilicity in Drug Action and Toxicology', VCH, Weinheim, 1996.
34. R. D. Cramer, III, *Persp. Drug Discov. Design*, 1993, **1**, 269–278.
35. M. Clark and R. D. Cramer, III, *Quant. Struct.–Act. Relat.*, 1993, **12**, 137–145.
36. H. van de Waterbeemd, (ed.), 'Chemometric Methods in Molecular Design', VCH, Weinheim, 1995.
37. H. van de Waterbeemd, (ed.), 'Advanced Computer-Assisted Techniques in Drug Discovery', VCH, Weinheim, 1995.
38. B. L. Bush and R. B. Nachbar, Jr., *J. Comput.-Aided Mol. Design*, 1993, **7**, 587–619.
39. F. Lindgren, P. Geladi, and S. Wold, *J. Chemometrics*, 1993, **7**, 45–59.
40. S. Rännar, F. Lindgren, P. Geladi, and S. Wold, *J. Chemometrics*, 1994, **8**, 111–125.
41. S. De Jong and C. J. F. Ter Braak, *J. Chemometrics*, 1994, **8**, 169–174.
42. S. Rännar, P. Geladi, F. Lindgren, and S. Wold, *J. Chemometrics*, 1995, **9**, 459–470.
43. R. D. Cramer, III, J. D. Bunce, D. E. Patterson, and I. E. Frank, *Quant. Struct.–Act. Relat.*, 1988, **7**, 18–25; erratum 1988, **7**, 91.
44. E. F. Meyer, I. Botos, L. Scapozza, and D. Zhang, *Persp. Drug Discov. Design*, 1995, **3**, 168–195.
45. R. T. Kroemer and P. Hecht, *J. Comput.-Aided Mol. Design*, 1995, **9**, 396–406.
46. S. A. DePriest, D. Mayer, C. B. Naylor, and G. R. Marshall, *J. Am. Chem. Soc.*, 1993, **115**, 5372–5384.
47. G. Klebe and U. J. Abraham, *J. Med. Chem.*, 1993, **36**, 70–80.
48. G. Greco, E. Novellino, M. Pellecchia, C. Silipo, and A. Vittoria, *J. Comput.-Aided Mol. Design*, 1994, **8**, 97–112.
49. S. J. Cho and A. Tropsha, *J. Med. Chem.*, 1995, **38**, 1060–1066.
50. U. Norinder, *J. Chemometrics*, 1996, **10**, 95–105.
51. G. Cruciani, M. Pastor, and C. Clementi, in 'Computer-Assisted Lead Finding and Optimization', Proc. 11th European Symposium on Quantitative Structure–Activity Relationships, Lausanne, 1996, eds. H. van de Waterbeemd, B. Testa, and G. Folkers, Verlag Helvetica Chimica Acta and VCH, Basel, Weinheim, 1997, pp 379–395.
52. U. Thibaut, G. Folkers, G. Klebe, H. Kubinyi, A. Merz, and D. Rognan, *Quant. Struct.–Act. Relat.*, 1994, **13**, 1–3.