Journal of Computer-Aided Molecular Design, **16:** 79–93, 2002.
KLUWER/ESCOM

Internally defined distances in 3D-quantitative structure-activity relationships

Christian Th. Klein[a,*], Norbert Kaiblinger[b] & Peter Wolschann[a]
[a]*Institute of Theoretical Chemistry and Molecular Structural Biology, University of Vienna, Währinger Strasse 17, A-1090 Vienna, Austria;* [b]*Institute of Mathematics, University of Vienna, Strudlhofgasse 4, A-1090 Vienna, Austria*

Received 31 October 2001; accepted 18 March 2002

Key words: 3D descriptors, benchmark steroids, internal coordinate system, solvent accessible surface

Summary

A new type of 3D-QSAR descriptors is introduced. For each molecule under consideration an internal coordinate system is defined relative to molecular points, such as positions of atoms in the molecule or centers of mass or certain substructures. From the origin of this system distances to the solvent accessible surface are calculated at defined spherical coordinate angles, θ and φ. The distances represent steric features, while the molecular electrostatic potentials at the intersection points with the surface represent the electrostatic contributions. The approach is called IDA (internal distances analysis). Matrices obtained by varying the spherical coordinate angles by fixed increments are correlated with the biological activity by partial least squares (PLS). The descriptors, tested with the benchmark steroids and an also well characterized benzodiazepine data set, turn out to be highly predictive. Additionally, they share the advantage of grid-based methods that the obtained models can be visualized, and thus be directly used in a rational drug design approach.

Introduction

Biologically active substances interact in most cases with biomolecules, triggering specific molecular mechanisms, such as activation of an enzyme cascade or opening of an ion channel, which finally leads to a certain biological response. Quantitative structure-activity relationships (QSAR) correlate this response with molecular properties of the compounds of interest. Because the response depends on the concentration of the active substance at the site of action and on the strength of interaction with the biological macromolecule, both of these aspects must be modeled quantitatively by QSAR. The most general relationship connecting molecular properties to activity is

$$\log(1/D_0) = A + \log B - \Delta G/RT, \qquad (1)$$

where A is a constant, logB summarizes the transport properties and the ability to cross biological membranes of the compounds considered, and ΔG denotes the free energy of interaction with the biomolecule. If D_0 is the dose of the administered substance to obtain a certain response, then $\log(1/D_0)$ is a quantitative measure of the biological activity: the lower the dose to obtain the response, the more active a compound is.

The first QSAR of type (1) was obtained by Hansch [1], who assumed that ΔG is a function of steric, electronic and hydrophobic properties of the active compounds, which were modeled by appropriate constants; the ability to cross biological membranes by passive transport was described by the octanol water partition coefficient P.

In the meantime, quantitative structure-activity relationships (QSAR) have become widely used tools to generate models of the biological response. QSAR techniques are particularly useful when the explicit structure of the interaction site with the biological (macro)molecule is not known.

[*] To whom correspondence should be addressed. Present address: Boehringer Ingelheim Austria, Dr. Boehringer Gasse 5-11, A-1121 Vienna, E-mail: christian.klein@vie.boehringer-ingelheim.com

In 'classical' QSAR, molecular descriptors of different levels are correlated with the biological activity. One-dimensional descriptors such as indicator variables affirm the presence or absence of an atom or a group in the molecule; two-dimensional descriptors, e.g., topological [2]and connectivity indices [3] – extract information from the way atoms are connected together in the molecule, while three dimensional descriptors like volume, surface or electrostatic potential are calculated from the spatial structure of the molecule.

The more recently developed 3D-QSAR methods generate descriptors exclusively from the three-dimensional structures of the molecules. Among the most pre-eminent 3D-QSAR methods is CoMFA (Comparative Molecular Field Analysis) [4]. The superimposed molecules are placed in a three-dimensional grid and at each grid point a steric and a electrostatic interaction energy with a probe is calculated. The molecular fields obtained in this way are then correlated with the biological activity by partial least squares (PLS) analysis [5]. A critical step in CoMFA (and generally in grid-based methods) is the superposition of the molecules: since the typically used Lennard-Jones ('6–12') steric fields and Coulomb electrostatic fields are 'hard', i.e., they change their values from close to zero to huge numbers within very small distances, the alignment of the molecules sensitively influences the correlation.

A recent grid based method, which introduces the concept of 'mean centered activity', is SOMFA (Self-Organizing Molecular Field Analysis) [6]. The steric contributions are modeled by a binary code (a value of 1 inside and 0 outside the van der Waals envelope); the electronic contributions are obtained by calculating the molecular electrostatic potential at the grid points, rather than evaluating the interaction energy with a probe. The mean centered activity is obtained by subtracting the mean activity of the molecular training set from each molecule's activity. Multiplying the shape or electrostatic value at every grid point for a given molecule by the mean centered activity, the grid points are weighted so that the most active and the least active molecules have higher values than the less interesting ones close to the mean activity. This kind of descriptor filtering increases the predictivity of the obtained models.

To bypass the problems arising from the use of 'hard' potentials in combination with grids, the concept of molecular similarity was introduced to QSAR studies. In one of the approaches [7], [8] N by N molecular similarity calculations (each molecule compared to every other) are performed, and the data matrices obtained are analyzed by PLS and a neural network.

A method which uses spatial similarity or dissimilarity rather than molecular similarity of the whole molecules is CoMSIA [9]. On the basis of an alignment function used in SEAL [10], similarity fields of the molecules (placed in a grid) to different probe atoms are evaluated. CoMSIA fields are based on Gaussian potentials and are consequently much 'softer' than CoMFA fields.

A somewhat different approach is Compass [11]. Sampling points are scattered on a surface 2 Å outside the average van der Waals envelope of the aligned compounds. Steric distances and hydrogen bond donor and acceptor distances are used as descriptors. Steric distances represent the distance from the sampling point to the closest atom, while hydrogen bond distances are measured from the sampling point to the nearest heteroatom with donor or acceptor properties. A neural network to derive linear or nonlinear relationships between the biological potency and the independent variables subsequently analyzes the obtained data.

Within molecular shape analysis (MSA) quantitative measurements of the molecular shape are performed [12], i.e., descriptors like difference volume, common overlap steric volume or non common overlap steric volume are calculated relative to a reference structure.

3D pharmacophoric models [13], [14] have proven to be useful descriptors that can be successfully employed in database searching and prediction of biological activity. Such models consist of geometrical requirements of pharmacophoric features, i.e., of chemical moieties important for activity like hydrophobic groups, hydrogen bond donors and acceptors, or charged/ionizable groups.

Receptor surface models (RSM) [15] generate a surface enclosing the common volume of the most potent ligands. Points on this surface are described by the complement of average partial charge, electrostatic potential or hydrogen bond ability. The molecules under consideration are then docked into the model by keeping the receptor surface fixed, and energy contributions are calculated, on the basis of which 3D-QSAR models are deduced.

To overcome the problem of molecular superposition CoMMA (Comparative Molecular Moment Analysis) [16] calculates descriptors based on 3D structures without reference to a common orientation frame. Descriptors are the moments of inertia (shape), magnitude of dipole and principal quadrupole moment (electrostatics), and additional parameters, which relate shape and charges.

The MS-WHIM (Weighted Holistic Invariant Molecular) approach [17] overcomes the problem of superposition by calculating statistical parameters (eigenvalue proportion, skewness and kurtosis) from a score matrix obtained from weighted principal component analysis (PCA).

A different approach used to generate alignment insensitive 3D descriptors [18] is based on autocorrelation of molecular surface properties such as hydrophobicity or electrostatic potential. Autocorrelation vectors are unique for a given molecular geometry and are rotationally and translationally invariant.

In a very recent publication [19] an autocorrelation strategy has also been applied, starting from a molecular interaction field (MIF). However, not all the computed terms are summed up; only the largest contribution is stored, while the other products are discarded. In this way, the autocorrelation vector can be transformed back to the original variables, with the great advantage that the obtained model can be visualized, i.e., a virtual receptor site can be obtained.

In the present paper we introduce a conceptually simple and efficient method to generate 3D molecular descriptors from the spatial structure. A coordinate system is chosen with its origin in the interior of the molecule. From this origin, distances to the solvent accessible surface of the molecules at defined spherical coordinate angles, θ and φ, respectively, are calculated. The points obtained on the surface are associated with the corresponding electrostatic potential, $V_{\theta\varphi}$. The approach – which will be called IDA (internal distances analysis) – turns out to be highly predictive. The descriptors are easy to interpret and the obtained models can be visualized as in the case of grid-based approaches.

Methods

Consider the geometry of a molecule defined by its Cartesian coordinates x_i, y_i and z_i. If the atoms of the molecule are described as spheres with radii equal to the corresponding van der Waals radii, then the van der Waals surface of the molecule is the exposed surface resulting from the mutual intersection of these spheres. If a spherical probe rolls over this surface, the solvent accessible surface (SAS) [20] is obtained. The radius of the probe is usually taken as 1.4 Å in SAS calculations.

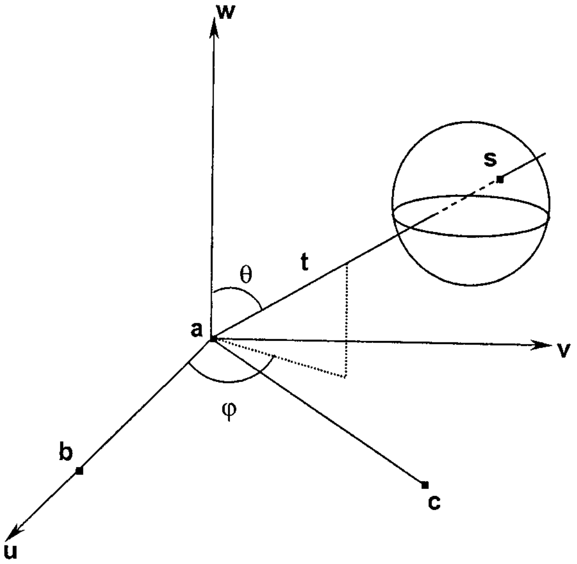

Figure 1. Definition of the internal coordinate system from three molecular points **a**, **b**, and **c**. **t** is the unit vector, **s** is the intersection point of the ray with the surface point of the sphere that constitutes the solvent accessible surface.

Consider now three points, **a**, **b**, **c**, not all in one line, defined by particular molecular features such as positions of atoms or centers of mass of certain substructures. As indicated in Figure 1, an internal coordinate system (**u**, **v**, **w**) with the origin **a** can be defined in the following way:

$$\begin{aligned} \mathbf{u} &= \mathbf{b} - \mathbf{a},\ \mathbf{w} = \mathbf{u} \times \mathbf{h},\ \mathbf{v} = \mathbf{w} \times \mathbf{u} \text{ with} \\ \mathbf{h} &= \mathbf{c} - \mathbf{a} \end{aligned} \tag{2}$$

or after normalization

$$\mathbf{u} = \frac{\mathbf{u}}{|\mathbf{u}|},\ \mathbf{v} = \frac{\mathbf{v}}{|\mathbf{v}|},\ \mathbf{w} = \frac{\mathbf{w}}{|\mathbf{w}|}. \tag{3}$$

Next we change to a spherical coordinate system (r, θ, φ) with origin **a**. The radius r is the distance to the origin, θ is the angle to the **w**-axis and φ is the angle to the **u**-axis in the (**u**, **v**)-plane. By using the vectors **u**, **v** and **w** we can switch between the original coordinates **x** and the spherical coordinates (r, θ, φ):

$$\begin{aligned} \mathbf{x} &= \mathbf{a} + r\mathbf{t} \text{ with} \\ \mathbf{t} &= \mathbf{u}\sin\theta\cos\varphi + \mathbf{v}\sin\theta\sin\varphi + \mathbf{w}\cos\theta, \end{aligned} \tag{4}$$

where **t** is the unit vector in the (**u**, **v**, **w**)-system. With this new coordinate system the molecular descriptors are obtained as follows: any pair of angles

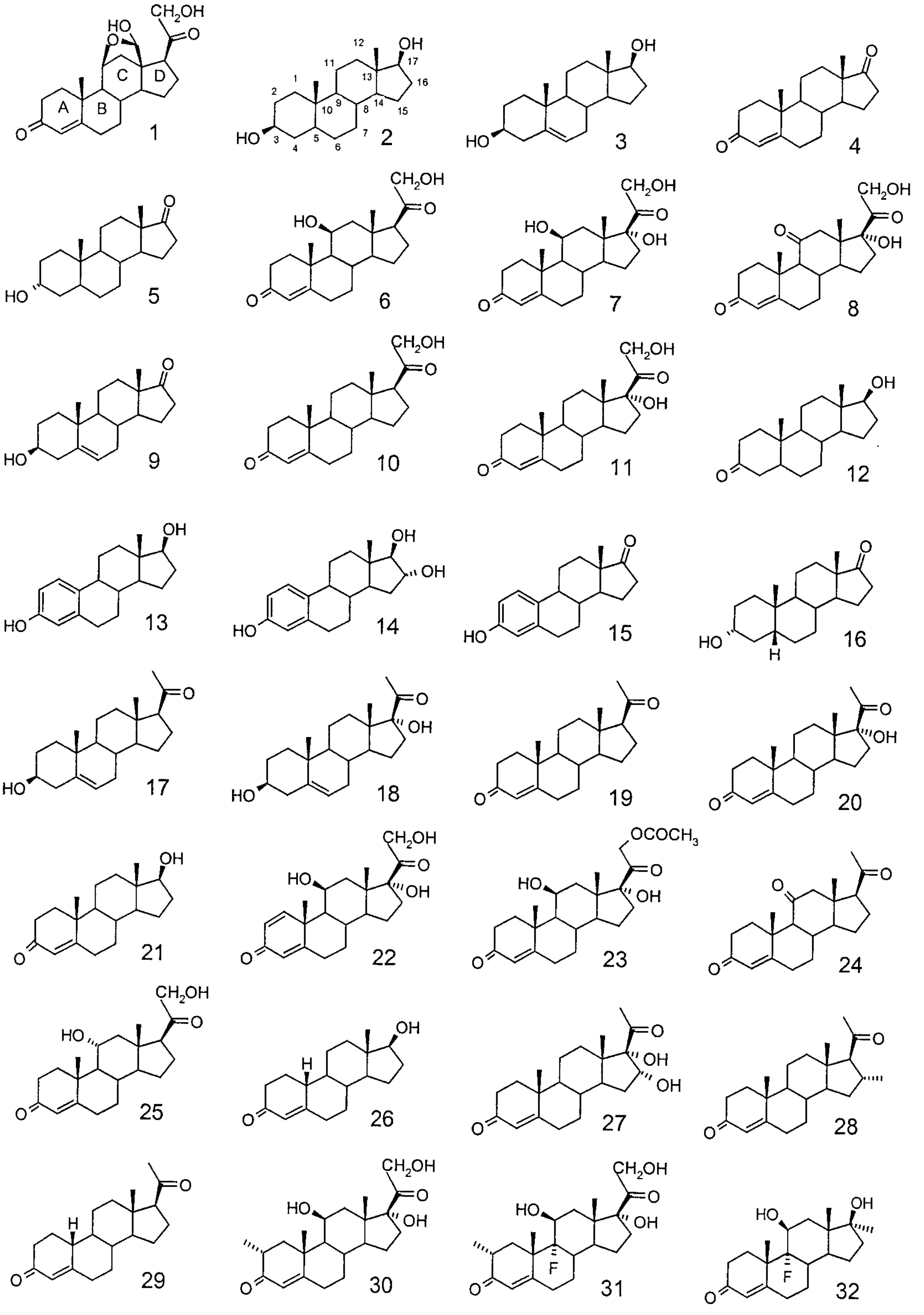

Chart 1. The benchmark steroid set. The IUPAC numbering of the steroid core and the nomenclature of the rings are given.

(θ, φ) defines a ray starting at point **a**. Let $r_{\theta\varphi}$ denote the distance from the origin **a** to the solvent accessible surface (SAS) in this direction. The intersection point **s** with the surface is calculated from the conditions (i) that it should lie on the ray, $\mathbf{s} = \mathbf{a} + r_{\theta\varphi}\,\mathbf{t}$, and (ii) that it should intersect a sphere which constitutes the surface, $|\mathbf{s} - \mathbf{k}_i| = R_i$. Here, $\mathbf{k}_i$ are the atomic coordinates and R_i are the radii of their van der Waals spheres, extended by 1.4 Å to obtain the SAS. Consequently, $r_{\theta\varphi}$ is calculated from the following equation:

$$\left|\mathbf{a} + r_{\theta\varphi}\mathbf{t} - \mathbf{k}_i\right| = R_i. \tag{5}$$

Abbreviating $\mathbf{p}_i = \mathbf{k}_i - \mathbf{a}$, the distance from the origin **a** to the solvent accessible surface results as a solution of Equation (5):

$$r_{\theta\varphi} = \mathbf{t}\cdot\mathbf{p}_i + \sqrt{\mathbf{t}\cdot\mathbf{p}_i - \mathbf{p}_i\cdot\mathbf{p}_i + R_i^2}, \tag{6}$$

where the dot denotes the usual scalar product of vectors. The radii implemented in the IDA source code are those from the SERF [21] program for SAS calculation.

The distances $r_{\theta\varphi}$ are taken as steric contributions in the QSAR analysis. For the electronic contributions, the electrostatic potentials at the corresponding intersection points **s** are considered:

$$V_{\theta\varphi} = \sum_i \frac{q_i}{|\mathbf{s} - \mathbf{k}_i|}. \tag{6}$$

The $\{r_{\theta\varphi}, V_{\theta\varphi}\}$ matrix generated by varying θ ∈ [0°, 180°] and φ ∈ (0°, 360°) by fixed increments is translationally and rotationally invariant, since the (**u**, **v**, **w**) coordinate system is internally defined.

Filtering is applied to reduce redundancy in the data set: if two points are closer to each other than 0.2 Å, one of the points is excluded, as presented in Figure 2 for the situation where the increments of θ and φ are 5 deg.

The data obtained in this way are correlated with the biological activity employing PLS analysis as implemented in the TSAR 3.3 software [22]. The statistical parameters to estimate the quality of a PLS vector i are the correlation coefficient r, the cross-validated correlation coefficient q^2, the predictive sum of squares PRESS, the statistical significance F and the statistical significance according to Ståhle and Wold [23]:

$$S = \sqrt{\frac{\mathrm{PRESS}_i}{\mathrm{SSR}_{i-1}}}, \tag{7}$$

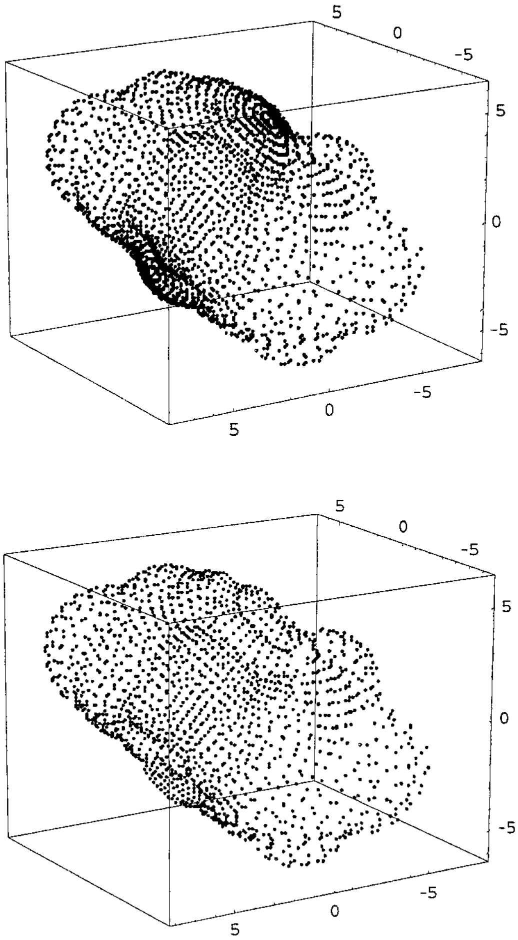

Figure 2. Filtering out points which are closer to each other than 0.2 Å. Using increments of 5° for θ and ϕ, 2592 points on the solvent accessible surface are obtained, which after filtering reduce to 1737.

where SSR_{i-1} is the sum of squares of the residuals of the $(i-1)$th component. According to this criterion the lower S is, the more significant is the PLS vector. S should preferably be less than one. Cross-validation is performed leaving out each compound in turn (LOO).

To compare with other methods, the benchmark set of 31 steroids first studied by Cramer et al. [4] is employed (Chart 1, Table 1); compounds 1–21 are used as training set and compounds 22–31 as a test set.

The structures of the steroid set were downloaded from the Gasteiger group homepage [24]. In this set, some mistakes present in previous versions are fixed [15]. Since the structures are generated with CORINA [25] but not minimized, AM1 [26] minimization was performed within TSAR3.3 to obtain electronic charges.

Table 1. The steroid benchmark set with the corresponding CGB affinities. In many publications the affinities are given as pK values; however log *K*'s are more obvious, since an increased value indicates increased binding affinity. In addition to the 'classical' benchmark set (compounds 1–31) another fluorinated steroid is included.

Compound	Steroid	CGB affinity (log *K*)
1	aldosterone	6.279
2	androstanediol	5.000
3	5-androstenediol	5.000
4	4-androstenedione	5.763
5	androsterone	5.613
6	corticosterone	7.881
7	cortisol	7.881
8	cortisone	6.892
9	dehydroepiandrosterone	5.000
10	11-deoxycorticosterone	7.653
11	11-deoxycortisol	7.881
12	dihydrotestosterone	5.919
13	estradiol	5.000
14	estriol	5.000
15	estrione	5.000
16	etiocholanolone	5.255
17	pregnenolone	5.255
18	17α-hydroxypregnenolone	5.000
19	progesterone	7.380
20	17-hydroxyprogesterone	7.740
21	testosterone	6.724
22	prednisolone	7.512
23	cortisol-21-acetate	7.553
24	4-pregnene-3,11,20-trione	6.779
25	epicorticosterone	7.200
26	19-nortestosterone	6.144
27	16α,17α-dihydroxy-4-pregnene-3,20-dione	6.247
28	16-α-methyl-4-pregnene-3,20-dione	7.120
29	19-norprogesterone	6.817
30	2α-methylcortisol	7.688
31	2α-methyl-9α-fluoro-cortisol	5.797
32	fluoxymesterone	5.000

Two choices of internal coordinate systems are analyzed: (i) **a** (the origin) is the center of mass of the steroid core, **b** and **c** are the centers of mass of the rings C and A, respectively; (ii) **a** (the origin) is the center of mass of ring C, **b** and **c** are the centers of mass of the rings B and A, respectively (see Chart I for ring nomenclature and IUPAC numbering).

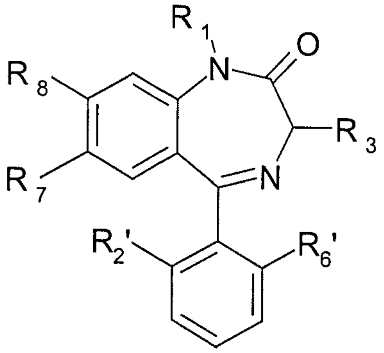

Figure 3. General structure of the benzodiazepines used in this study.

For the *test set* (compounds 22–31) the standard deviation of errors of prediction is calculated, which is a measure of the external predictivity of a model:

$$\mathrm{SDEP} = \sqrt{\frac{\sum_{i=1}^{n} (y_{i,\mathrm{pred}} - y_{i,\mathrm{act}})^2}{n}}, \qquad (8)$$

where y_{act} is the actual, y_{pred} is the predicted value of the biological activity. In a similar fashion the standard deviation of errors of calculation (SDEC) is evaluated for the *training set* by replacing y_{pred} by y_{calc}, which is the value of calculated activity from the PLS regression equation. The SDEC thus reflects the internal predictivity (i.e., the predictivity within the training set) of the model.

A second data set – also extensively studied by various QSAR methods – consisting of 57 benzodiazepines is also employed [27–29]. Their general structure is presented in Figure 3, the biological activities and structural details in Table 2.

Results and discussion

Benchmark Steroids

In Table 3 several 'best' models, judged by their standard deviations of errors of prediction (SDEP) are presented, obtained with different increments (I_θ and I_φ) for θ and φ.

The SDEP is listed for the training set 22–31 and 22–30, since compound 31 is an outlier and the behavior of this compound will be analyzed in more detail. The SDEP for the complete test set will further be referred to as SDEP_{10}, and SDEP_9 for the test set excluding steroid 31. Those models that have the lowest SDEP_9 are considered best.

Table 2. The benzodiazepine data set used in this study.

Name	R_7	R_1	R_2'	R_6'	R_3	R_8	log IC_{50}
Clonazepam	NO_2	H	Cl	H	H	H	0.255
Delorazepam	Cl	H	Cl	H	H	H	0.255
Diazepam	Cl	Me	H	H	H	H	0.908
Flunitrazepam	NO_2	Me	F	H	H	H	0.508
Halazepam	Cl	CH_2CF_3	H	H	H	H	1.964
Lorazepam	Cl	H	Cl	H	OH	H	0.544
Meclonazepam	NO_2	H	Cl	H	Me	H	0.079
Nitrazepam	NO_2	H	H	H	H	H	1.000
Nordazepam	Cl	H	H	H	H	H	0.973
Oxazepam	Cl	H	H	H	OH	H	1.255
Ro05-2904	CF_3	H	H	H	H	H	1.114
Ro05-2921	H	H	H	H	H	H	2.544
Ro05-3061	F	H	H	H	H	H	1.602
Ro05-3072	NH_2	H	H	H	H	H	2.587
Ro05-3367	Cl	H	F	H	H	H	0.301
Ro05-3418	NH_2	Me	H	H	H	H	2.663
Ro05-3590	NO_2	H	CF_3	H	H	H	0.544
Ro05-4082	NO_2	Me	Cl	H	H	H	0.342
Ro05-4336	H	H	F	H	H	H	1.322
Ro05-4435	NO_2	H	F	H	H	H	0.176
Ro05-4520	H	Me	F	H	H	H	1.146
Ro05-4528	CN	Me	H	H	H	H	2.850
Ro05-4608	H	Me	Cl	H	H	H	0.580
Ro05-4619	NH_2	H	Cl	H	H	H	1.875
Ro05-4865	F	Me	H	H	H	H	1.230
Ro05-6820	F	H	F	H	H	H	0.869
Ro05-6822	F	Me	F	H	H	H	0.708
Ro06-7263	Cl	Cl	H	H	Me	H	1.690
Ro06-9098	NO_2	CH_2OCH_3	H	H	H	H	2.633
Ro07-2750	Cl	$(CH_2)_2OH$	F	H	H	H	1.389
Ro07-3953	Cl	H	F	F	H	H	0.204
Ro07-4065	Cl	Me	F	F	H	H	0.613
Ro07-4419	H	H	F	F	H	H	1.279
Ro07-5193	Cl	H	Cl	F	H	H	0.477
Ro07-5220	Cl	Me	Cl	Cl	H	H	0.740
Ro07-6198	H	H	F	F	H	Cl	1.447
Ro07-9957	I	Me	F	H	H	H	0.462
Ro07-4878	Cl	H	F	H	Me	H	0.544
Ro07-6896	NO_2	Me	F	H	Me	H	0.845
Ro13-3780	Br	Me	F	F	H	H	0.380
Ro14-3074	N_3	H	F	H	H	H	0.724
Ro20-1310	Cl	$C(CH_3)_3$	H	H	H	H	2.792
Ro20-1815	NH_2	Me	F	H	H	H	1.813
Ro20-2533	Et	H	H	H	H	H	1.566
Ro20-2541	CN	Me	F	H	H	H	1.477
Ro20-3053	COMe	H	F	H	H	H	1.255
Ro20-5397	CHO	H	H	H	H	H	1.633

Table 2. Continued

Name	R_7	R_1	R_2'	R_6'	R_3	R_8	log IC_{50}
Ro20-5747	$CH=CH_2$	H	H	H	H	H	1.380
Ro20-7078	Cl	H	F	H	Cl	H	0.724
Ro20-7736	NHOH	Me	F	H	H	H	1.982
Ro20-8065	Cl	H	F	H	H	Cl	0.556
Ro20-8552	Me	H	F	H	H	Cl	1.146
Ro20-8895	H	H	F	H	H	Me	1.279
Ro22-3294	Cl	H	Cl	Cl	H	H	0.845
Ro22-4683	NO_2	$C(CH_3)_3$	F	H	H	H	2.477
Ro22-6762	Cl	Me	H	H	H	Cl	1.602
Temazepam	Cl	Me	H	H	OH	H	1.204

In all instances low dimensional models, usually with 1 or 2 components, have the best external predictivity. Since according to some authors [30] the F-value of a model should exceed the percentage point *at least four times,* all of the models shown are statistically significant at the 99% level ($F_{0.01,1,19} = 8.18$, $F_{0.01,2,18} = 6.01$).

In all cases a randomization of the target properties leads to complete loss of predictive ability, the q^2-values dropping below zero.

In Table 4 the models obtained from the data sets with 400 and 1112 descriptors are analyzed in more detail: the statistical and predictivity parameters are given for PLS vectors of different dimensions.

A higher value of q^2 does not necessarily lead to a better external predictivity: for the 400 descriptors case, q^2 reaches values up to 0.919, but the one component model with a q^2 of 0.693 has the best predictivity judged by SDEP. A similar situation can be observed for the 1112 descriptors case (and all other cases).

In Table 4, the statistical significance is given in terms of S. Being calculated from the predictive sum of squares (PRESS, Equation 7), it reflects better the external predictivity of the models than F, which only takes into account the correlation of the training set. For the 10 component model of the 400 descriptors case, F is as high as 249.25; however, it has a significantly lower predictivity in terms of SDEP than the best (one component) model with F equal to 64.33.

On the other hand, one can clearly recognize – even by visual inspection – the correlation between q^2 and the standard deviations of errors of calculation, SDEC. This correlation [taking values from all models and all components up to max(q^2)] is indeed high, with $r = 0.95$, $F = 506.2$ and $s = 0.02$.

The number of optimal components a model must have to give the best predictivity certainly depends on the data set, i.e., on the correlation between the independent variables and on the noise in the data set. It is known that including too many PLS components into a model may only fit the noise in the data set to the noise in the target property. This situation is well illustrated in Table 4: while the external predictivity (SDEP) reaches an optimum in the models of low dimension, the fraction of explained variance (r^2) increases up to 0.996 in the 400 descriptors case upon including further components, and the *SDEC* decreases to values as small as 0.07.

Table 3 shows that good predictions are obtained with small and large numbers of descriptors. However, it appears that the models derived from large descriptor sets are more reliable, as they are less sensitive to the choice of the internal coordinate systems. In the 148 descriptors case, for example, the differences of $SDEP_{10}/SDEP_9$ are rather large for the two different internal coordinate systems. In contrast, for the 3474 and 3398 descriptors sets, respectively, no significant distinction between the external predictivities can be observed for the two different coordinate systems. It is to be noted that the differences in the number of descriptors in the cases where I_φ and I_θ are (10, 10), (10, 5) and (5, 5), respectively, stem from filtering out points closer to each other than 0.2 Å (in the other cases no points closer than 0.2 Å exist).

With IDA, 2α-methyl-9α-fluoro-cortisol (steroid 31) is consistently predicted closer to the experimental value than with most other 3D-QSAR methods. An exception is the RSM method which predicts steroid 31

Table 3. Best predictivity of models as a function of number of descriptors **NDesc**.

Model	I_φ	I_θ	Ndesc	Noc	r	F	q^2	AEP31	$SDEP_{10}/SDEP_9$
1[a]	40	20	148	1	0.879	64.70	0.699	0.730	0.432/0.384
2[b]			148	1	0.898	75.53	0.743	0.869	0.572/0.529
3[a]	30	20	196	1	0.877	63.25	0.696	0.634	0.426/0.396
4[b]			196	1	0.839	45.51	0.625	0.263	0.406/0.418
5[a]	30	15	268	1	0.876	62.89	0.690	0.879	0.417/0.327
6[b]			268	1	0.844	46.97	0.634	0.442	0.411/0.408
7[a]	20	15	400	1	0.878	64.33	0.693	0.791	0.401/0.330
8[b]			400	2	0.939	67.92	0.712	1.169	0.539/0.415
9[a]	20	10	616	1	0.873	61.17	0.678	0.698	0.407/0.360
10[b]			616	2	0.938	66.63	0.694	0.835	0.462/0.400
11[a]	10	10	1122	1	0.880	65.10	0.698	0.724	0.405/0.352
12[b]			1152	1	0.846	47.78	0.637	0.359	0.394/0.397
13[a]	10	5	2192	1	0.881	65.82	0.702	0.793	0.417/0.351
14[b]			2222	1	0.853	50.85	0.651	0.462	0.398/0.390
15[a]	5	5	3474	1	0.896	77.54	0.747	1.089	0.486/0.359
16[b]			3398	1	0.886	69.78	0.719	1.137	0.485/0.343

$\mathbf{I}_\varphi$ – angle increment for φ; $\mathbf{I}_\theta$ – angle increment for θ; **r** – correlation coefficient; $\mathbf{q}^2$ – cross-validated correlation coefficient; **F** – statistical significance; **AEP31** – absolute error of prediction of steroid 31; **Noc**-number of components; $\mathbf{SDEP}_{10}$ – standard deviation of errors of prediction evaluated for the test set: cpds. 22–31; $\mathbf{SDEP}_9$ – standard deviation of errors of prediction evaluated for the test set:cpds 22–30.

[a]The internal coordinate system is defined as follows: **a** (origin) is the center of mass of ring C, **b** is the center of mass of ring B and **c** is the center of mass of ring A.

[b]The internal coordinate system is defined as follows: **a** (the origin) is the center of mass of the steroid core, **b** is the center of mass of ring C and **c** is the center of mass of ring A.

close to the experimental value with $\log K$ of 6.049, corresponding to a residual of 0.252 [31]. The best prediction of steroid 31 made by IDA (Table 3, model 4) is comparable to this result ($\log K$ of 6.06, residual of 0.263). In Table 5 the absolute error of prediction of steroid 31, AEP31, together with $SDEP_{10}$ and $SDEP_9$ is presented for different methods. All of the methods compared are discussed in the introduction. The CoMFA results were obtained by Bravi et al. [17] using fractional factorial design (FFD) [32] to select only those variables with highest explanatory power.

Because 2α-methyl-9α-fluoro-cortisol is an outlier for most methods its prediction cannot be taken as a measure of the quality of a model, since it has to be established, whether the experimental or the predicted value is wrong. Therefore, the models with the best external predictivity are considered to be those which have the lowest $SDEP_9$. However, with IDA $SDEP_9$ is also lower in many instances, showing that it compares well to the other 3D-QSAR methods. With the RSM method, steroid 23 is predicted to be a large outlier, with a residual of −3.469. Therefore, $SDEP_{10}$ is much larger than for the other methods, although it has a good overall predictivity for the rest of the test set. Thus, in Table 5 the $SDEP_9$ for the RSM method is calculated for the test set excluding steroid 23.

To make sure that the better predictions of IDA are not artifacts stemming from the choice of the atomic radii (SERF: $r_C = 1.7$ Å, $r_O = 1.42$ Å, $r_F = 1.45$ Å, $r_H = 1.1$ Å), descriptors using the radii of Motoc and Marshall [33] ($r_C = 1.53$ Å, $r_O = 1.36$ Å, $r_F = 1.30$ Å, $r_H = 1.08$ Å) were also calculated. The results are qualitatively the same. For most of the models, both $SDEP_9$ and $SDEP_{10}$ are lower than for other methods, with steroid 31 being predicted throughout closer to the experimental value.

The reasons for the problems with steroid 31 have been interpreted in different ways. Because it is the only compound with a fluorine atom, some authors [15, 34] suggest that this atom might be responsible for the worse prediction. Other authors claim [6] that fluorination per se is not a problem, but rather the differences in the experimental techniques used to derive the training data and those to derive the test data [35]. They used 35-fluoxymesterone, fluorinated at the same position as 2α-methyl-9α-fluoro-cortisol,

Table 4. Statistical parameters, standard deviations of errors of predictions (**SDEP**), standard deviations of errors of calculations (**SDEC**), obtained from data sets with number of descriptors, **Ndesc**, of 400 and 1112, respectively.

NDesc		PLS models of dimension									
		1	2	3	4	5	6	7	8	9	10
	S	0.554	1.039	1.547	1.652	2.029	2.699	3.176	3.03	3.257	4.125
	PRESS	6.146	4.923	3.655	2.892	3.354	3.457	2.823	2.169	1.731	1.617
	q^2	0.693	0.754	0.817	0.855	0.832	0.827	0.859	0.891	0.913	0.919
	r^2	0.772	0.924	0.947	0.959	0.976	0.986	0.988	0.992	0.995	0.996
400[a]	SDEC	0.545	0.317	0.264	0.231	0.177	0.136	0.124	0.103	0.079	0.07
	$SDEP_{10}/SDEP_9$	0.401/0.330	0.634/0.443	0.799/0.545	0.807/0.500	0.854/0.472	0.977/0.537	0.996/0.553	0.996/0.541	1.018/0.593	1.036/0.598
	AEP31	0.7912	1.514	1.929	2.066	2.299	2.64	2.676	2.7	2.685	2.697
	AEP32	1.314	1.882	2.186	2.263	2.389	2.577	2.66	2.734	2.901	2.931
	S	0.529	1.036	1.535	1.847	2.045	2.337	2.402	2.482	2.849	3.221
	PRESS	5.614	4.596	4.051	3.948	4.009	3.539	2.753	2.367	2.369	2.426
	q^2	0.719	0.77	0.797	0.803	0.799	0.823	0.862	0.882	0.882	0.878
	r^2	0.786	0.914	0.942	0.952	0.967	0.976	0.981	0.985	0.988	0.993
1112[b]	SDEC	0.531	0.336	0.276	0.251	0.206	0.177	0.158	0.138	0.124	0.096
	$SDEP_{10}/SDEP_9$	0.485/0.343	0.676/0.448	0.825/0.539	0.917/0.574	0.998/0.573	1.008/0.553	1.008/0.551	1.030/0.614	1.087/0.697	1.111/0.763
	AEP31	1.135	1.665	2.047	2.335	2.647	2.723	2.726	2.685	2.728	2.692
	AEP32	1.612	2.045	2.316	2.452	2.555	2.603	2.598	2.545	2.629	2.754

AEP31 and **AEP32** are the absolute errors of prediction of steroids 31 and 32, respectively. **S** is the statistical significance according to Ståhle and Wold [23]. All other abbreviations have the same meaning as in Table 3.

[a] The internal coordinate system is defined as follows: **a** (the origin) is the center of mass of ring C, **b** is the center of mass of ring B and **c** is the center of mass of ring A.

[b] The internal coordinate system is defined as follows: **a** (the origin) is the center of mass of the steroid core, **b** is the center of mass of ring C and **c** is the center of mass of ring A.

Table 5. Predictivity of different 3D-QSAR methods compared to IDA models.

	CoMFA[a]	Similarity matrix analysis[a]	Compass[b]	MS-WHIM[a]	SOMFA[a]	RSM[c]	IDA[d]	IDA[e]	IDA[f]	IDA[g]
$SDEP_9$	0.356	0.385	0.339	0.411	0.367	0.373[h]	0.418	0.330	0.390	0.343
$SDEP_{10}$	0.716	0.640	0.705	0.662	0.584	1.153	0.406	0.401	0.398	0.485
AEP31	1.996	1.660	1.982	1.694	1.441	0.252	0.263	0.791	0.462	1.137

[a] Reference 6.
[b] Reference 11.
[c] Reference 31.
[d] Table 3, model 4, 198 descriptors.
[e] Model 7, 400 descriptors.
[f] Model 14, 2222 descriptors.
[g] Model 16, 3398 descriptors.
[h] For RSM the $SDEP_9$ is calculated excluding steroid 23, which is a large outlier (residual −3.469).

to conclude that fluorination per se does not cause problems.

To study the effect of the fluorine atom, we have also used fluoxymesterone and compared the IDA predictions of 2α-methyl-9α-fluoro-cortisol and fluoxymesterone to the experimental values for different models of different size and components. We assumed that the models of optimal size (number of descriptors) and dimension (number of components) are those which have the best predictivity for compounds 22–30 (the lowest $SDEP_9$), i.e., which best describe the rest of the test. Our results, summarized in Figure 4, can be outlined as follows:

(i) lowest absolute errors of predictions (AEPs) are obtained for both fluorinated steroids with one component PLS models, regardless of whether these models have the best overall predictivity. However, both steroids are outliers;

(ii) with models of components up to 5, 2α-methyl-9α-fluoro-cortisol is generally better predicted than fluoxymesterone;

(iii) with models of higher dimensions, the AEPs reach a plateau for both steroids, with 2α-methyl-9α-fluoro-cortisol being slightly better predicted than fluoxymesterone;

(iv) increasing the number of components in the PLS models has approximately the same impact on the AEPs for both fluorinated steroids (approximately parallel curves in Figure 4);

(v) including fluoxymesterone in the training set gives substantially better predictions of the activity of 2α-methyl-9α-fluoro-cortisol for all models with more than one component; however, the AEPs still remain larger than for the rest of the test set;

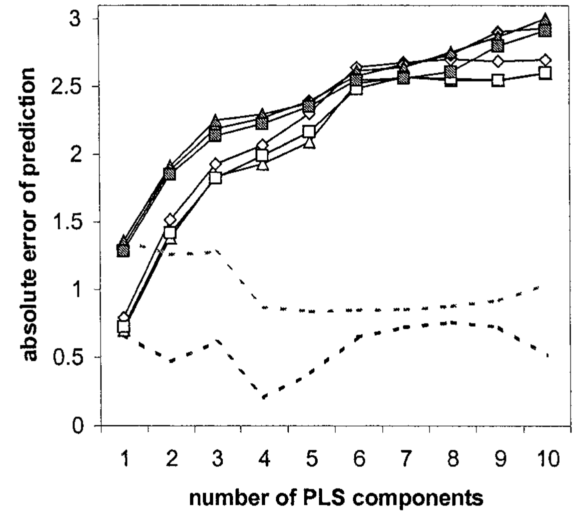

Figure 4. Absolute errors of predictions (AEPs) for the two fluorinated compounds: 2α-methyl-9α-fluoro-cortisol (unfilled) and fluoxymesterone (filled with light gray). PLS models up to dimension 10 are shown for 400 (rhombs), 616 (triangles) and 1112 (squares) descriptors. The black dotted line shows the AEPs for 2α-methyl-9α-fluoro-cortisol if fluoxymesterone is included in the training set; the gray dotted line shows the AEPs for fluoxymesterone if 2α-methyl-9α-fluoro-cortisol is included in the data set, both for the 1112 descriptors case.

(vi) including 2α-methyl-9α-fluoro-cortisol in the training set gives substantially better predictions of the activity of fluoxymesterone for all models with more than one component; however, the AEPs still remain larger than for the rest of the test set.

Finally, IDA shares the advantage of grid based methods that the models can be visualized, thus identifying features important for the biological activity. Figures 5 and 6 show the steric and electrostatic fea-

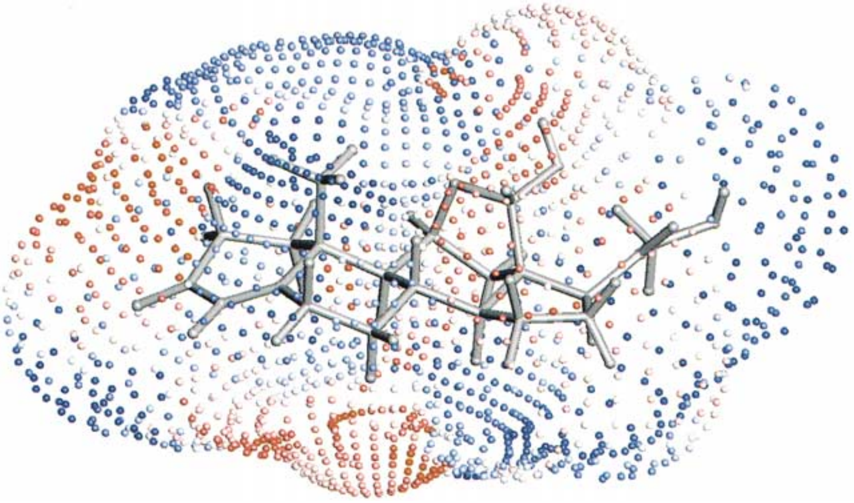

Figure 5. Sterically favorable and unfavorable regions of the steroids. Blue indicates regions (surface points) where steric bulk enhances CGB activity, red where steric bulk is unfavorable. The darker the colors, the more pronounced are the effects, white points being indifferent. Steroid 1 is included as frame of reference.

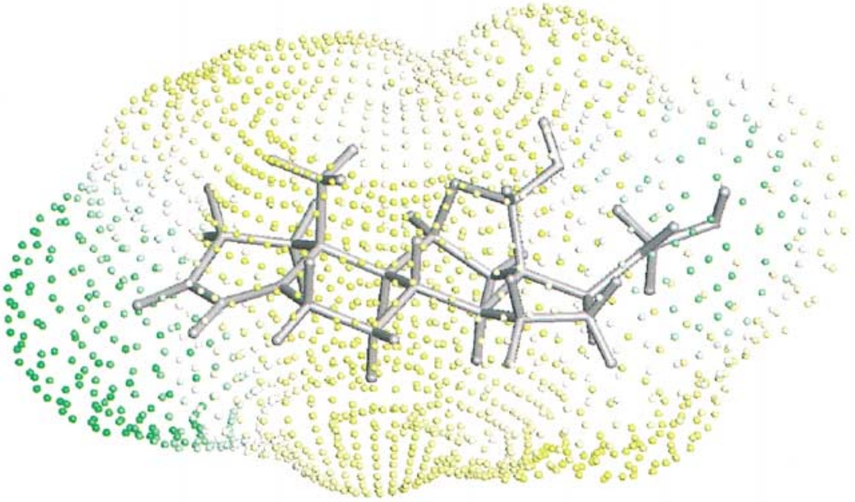

Figure 6. Green indicates regions (surface points) where an increased negative molecular electrostatic potential (MEP) is favorable or a positive MEP is unfavorable; yellow indicates reagions where a negative MEP is unfavorable or a positive MEP is favorable for CGB affinity. The more dark the colors, the more pronounced are the effects, white points being indifferent. Steroid 1 is included as frame of reference.

tures of a model obtained from the 3398 descriptors set.

The steric model is consistent with models obtained by other methods. As with CoMFA, a favorable effect is found in the vicinity of the C17 side chain and below the D-ring of the steroid core. The positive steric interaction of the C17 side chain and of the 19-methyl group is in agreement with the CoMSIA result. It is noticeable that the steric picture obtained by IDA is very similar to that obtained by SOMFA, indicating – like the latter – a favorable region below the D-ring and the C3-substituent, and an unfavorable zone above the C-ring and the C3-substituent, as well as the previously mentioned features. However, the areas with positive and negative contribution are more clearly delimited in the IDA model.

The electrostatic picture obtained by IDA also shows high concordance with the SOMFA results. In the vicinity of the C3 and C17 substituents, increased negative MEPs are beneficial for CGB binding, the effect being more pronounced at C3. Up to this point the IDA and SOMFA results are in perfect agree-

Table 6. Comparison of IDA results with other studies on the benzodiazepine data set.

Study	Methods	Training		Cross-validated	
		r^2	SE	q^2	SE
Maddalena and Johnston[27]	substituent constants, NN	0.880	0.245	0.803	0.321
So and Karplus[28]	substituent constsants, NN	0.941	–	0.882	–
Winkler and Burden[29]	HQSAR (size 4–7), PLS	0.965	0.156	0.701	0.460
Winkler and Burden[29]	HQSAR (size 3–10), PLS	0.982	0.119	0.784	0.411
This study	IDA, Ndesc = 196, PLS	0.997	0.056	0.669	0.569
This study	IDA, NDesc = 268, PLS	0.990	0.098	0.693	0.549
This study	IDA, NDesc = 400,PLS	0.979	0.143	0.787	0.457
This study	IDA, NDesc = 532, PLS	0.979	0.141	0.702	0.544
This study	IDA, NDesc = 616, PLS	0.989	0.106	0.743	0.502
This study	IDA, NDesc = 1102,PLS	0.992	0.087	0.778	0.456

Table 7. Comparison of results obtained with HQSAR and IDA, employing a 52 benzodiazepines training set and a 5 compounds test set. The best HQSAR models (size 3–10 and 4–7, Ref. 31) and the best IDA models (with different number of descriptors, **Ndesc**) are shown. **Noc** is the number of PLS components. The standard deviations of errors of calculation and prediction are calculated for the training set (**SDEC**$_{\text{int}}$ and **SDEP**$_{\text{int}}$, respectively) and reflect the internal predictivity. For the 5 compounds test set, the standard deviation of errors of prediction (**SDEP**$_{\text{ext}}$) is given.

			Training set			Test set
Model	Noc	r^2	$SDEC_{int}$	q^2	$SDEP_{int}$	$SDEP_{ext}$
HQSAR (4–7)	10	0.950	0.171	0.685	0.431	0.470
HSQAR (3–10)	16	0.988	0.091	0.708	0.449	0.523
IDA, NDesc = 268	13	0.936	0.204	0.532	0.511	0.453
IDA, NDesc = 400	13	0.954	0.148	0.661	0.476	0.443

Table 8. IDA models derived from a training set of 37 benzodiazepines, with different numbers of descriptors (**Ndesc**). **RSS** is the residual sum of squares, **PRESS** the predictive sum of squares. **SDEP**$_{\text{int}}$, **SDEC**$_{\text{int}}$ and **SDEP**$_{\text{ext}}$ have the same meaning as in Table 7. **SDEP**$_{\text{ext}}$ is calculated for the test set compounds (2, 3, 7, 8, 12, 15, 19, 23, 26, 27, 29, 31, 34, 35, 37, 39, 40, 43, 49, 52), reflecting the external predictive ability.

	Ndesc = 400	NDesc = 616
RSS	0.585	0.743
PRESS	11.206	10.189
q^2	0.688	0.717
r^2	0.984	0.979
$SDEP_{int}$	0.550	0.524
$SDEC_{int}$	0.091	0.103
$SDEP_{ext}$	0.382	0.424

ment. However, while IDA indicates that below *and* above the steroid rings a negative MEP is unfavorable, SOMFA locates this effect only below (i.e., at the opposite side of the 19-methyl group) the steroid ring. Moreover, with SOMFA the unfavorable effect of a negative MEP extends to the C17 substituent (where it is very pronounced), so that it is difficult to decide which effect (positive or negative MEP) will be dominant. It thus appears that the electrostatic picture obtained by IDA fits better to the steroid structures. It can indeed be observed that the most active steroids (corticosterone, cortisol and deoxycortisol) have a carbonyl group at the C3 position and a hydroxy- *and* hydoxyacetyl- group at C17.

Benzodiazepines

Table 6 shows the results of IDA for different sizes of descriptor sets, compared to other methods.

Maddalena and Johnston [27] and So and Karplus [28] both use substituent constants as descriptors which are correlated with the biological activity employing neural networks. Their models have the highest q^2-values of 0.803 and 0.882, respectively. Unfortunately, these authors do not use test sets for the determination of the external predictivity of their models.

The best results for the whole benzodiazepine set obtained with IDA are comparable to the best HQSAR results of Winkler and Burden [29]. However, using the same test set, viz.: clonazepam, diazepam, Ro-20-3110, Ro-05-2921 and Ro-14-3074, a slightly better external predictivity is observed for IDA, as shown in Table 7.

While the best HQSAR model (size 4–7) leads to a SDEP for the test set of 0.470, the 13 components PLS model of the 400 descriptors set of IDA leads to a SDEP of 0.443. For IDA, the models with the best external predictivity (lowest SDEP) and highest q^2 are shown. We have often observed that – as in the benchmark case – the best predictions are not necessarily made with the models which have the highest q^2 values [36]. This situation is once more emphasized in Table 7: The 4–7 size HQSAR model with a q^2 of 0.685 has a better predictivity (SDEP = 0.470) than the 3–10 size model with a q^2 of 0.708 and a SDEP of 0.523. In a recent paper the relationships between LOO q^2 and the external predictivity is analyzed in detail for different data sets [37]. The results confirm the lack of correlation between q^2 and predictive ability for test sets. Although a low q^2 for the training set can serve as an indicator for low predictive power of a model, the opposite is not necessarily true. The authors conclude that external validation, i.e., using a test set, is the only way to establish the predictivity of a QSAR model. The reason for this situation might be the fact that during cross-validation the models ‘see’ all compounds from the training set, and thus q^2 rather reflects the internal than the external predictivity of a model.

Because a test set of 5 compounds is rather small, it was checked whether the predictive ability of the models still remains good for a larger test set: from the 57 benzodiazepines, 20 compounds (2, 3, 7, 8, 12, 15, 19, 23, 26, 27, 29, 31, 34, 35, 37, 39, 40, 43, 49, 52) were randomly picked out and used for external prediction. The results are summarized in Table 8.

The SDEPs are calculated during the cross-validation process ($\mathrm{SDEP_{int}}$) and for the external test set ($\mathrm{SDEP_{ext}}$). It can be seen that the predictions of the 20 compound test set are even better than those for the 5 compounds test set. The best prediction with a $\mathrm{SDEP_{ext}}$ of 0.382 is made with the 400 descriptors case.

Conclusions

The IDA method presented here is an efficient tool for 3D structure-activity relationships. Even with a small number of descriptors – as compared to grid based methods – highly predictive models are obtained. However, deducing models from large descriptor sets is less sensitive to the choice of the internal coordinate system.

For the benchmark steroids, IDA gives results closer to the experimental values than other frequently used 3D-QSAR methods; in the case of the benzodiazepines, the external predictivity appears to be of at least the same quality as with other methods.

The descriptors, distances from an origin and electrostatic potential, have a physical meaning, which is easy to interpret. Visualization of the models yields clear pictures of beneficial and unfavorable contributions to biological activity.

Acknowledgements

N.K. thanks the Austrian Science Fund FWF for support under grant P14485. The authors thank Trevor Newton and Gerhard Ecker for reading the manuscript carefully and for helpful comments.

References

1. Hansch, C.A., Accounts Chem. Res., 2 (1996) 232.
2. Balaban, A.T., Chem. Phys. Lett., 89 (1982) 399.
3. Hall, L.H. and Kier, L.B., In Lipkowitz, K.B. and Boyd, D.B., (eds.) Reviews in Computational Chemistry Vol. 2., VCH Publishers, New York, NY, 1991, pp. 367–422.
4. Cramer, R.D.III, Patterson, D.E. and Bunce, J.D., J. Am. Chem. Soc., 110 (1988) 5959.
5. Dunn, W.J., Wold, S., Edlund, U. and Hellberg, S., Quant. Struct.-Act. Relat., 3 (1984) 131.
6. Robinson, D.D., Winn, P.J., Lyne, P.D. and Richards, W.G, J. Med. Chem., 42 (1999) 573.
7. Good, A.C., So, S.-S. and Richards, W.G, J. Med. Chem., 36 (1993) 433.
8. Good, A.C., Peterson, S.J. and Richards, W.G., J. Med. Chem., 36 (1993) 2929.
9. Klebe, G., Abraham, U. and Mietzner, T., J. Med. Chem., 37 (1994) 4130.

10. Kearsley, S.K. and Smith, G.M., Tetrahedron Comput. Methodol., 3 (1990) 615.
11. Jain, A.N., Koile, K. and Chapman, D., J. Med. Chem., 37 (1994) 2315.
12. Hopfinger, A.J., J. Am. Chem. Soc., 102 (1980) 7196.
13. Green, J., Kahn, S., Savoj, H., Sprague, P, and Teig, S., J. Chem. Inf. Comput. Sci., 34 (1994) 1297.
14. Barnum, D., Greene, J., Smellie, A., and Sprague, P., J. Chem. Inf. Comput. Sci., 36, (1996) 563.
15. Hahn, M. and Rogers, D., J. Med. Chem., 38 (1995) 2080.
16. Silverman, B.D. and Platt, D.E., J. Med. Chem., 39 (1996) 2129.
17. Bravi, G., Gancia, E., Mascagni, P., Pegna, M., Todeschini, R. and Zaliani, A.J, J. Comput. Aid. Mol. Des., 11 (1997) 79.
18. Wagener, M., Sadowski, J. and Gasteiger, J., J. Am. Chem. Soc., 117 (1995) 7769.
19. Pastor, M., Cruciani, G., McLay, I., Pickett, S. and Clementi, S., J. Med. Chem., 43 (2000) 3233.
20. Lee, B. and Richards, F.M., J. Mol. Biol., 55 (1971) 379.
21. Flowers, D.R., J. Mol. Graphics Mod., 15 (1997) 238.
22. TSAR 3.3, Oxford Molecular Ltd., The Medawar Centre, Oxford Science Park, Oxford, 2000.
23. Ståhle, L. and Wold, S., In Ellis, G.P. and West, G.B. (eds.), Progress in Medicinal Chemistry, Vol. 25, Elvesier Scientific Publishers, Amsterdam, 1988, pp. 292–338.
24. Dataset of 31 Steroids Binding to the Corticosteroid Binding Globulin Receptor. http://www2.ccc.uni-erlangen.de/services/steroids/index.html
25. CORINA Molecular Networks, GmbH Computerchemie, Langenmarckplatz 1, Erlangen, 1997.
26. Dewar, M.J.S., Zoebisch, E.G., Healy, E.F. and Stewart, J.J.P., J. Am. Chem. Soc., 107 (1985) 3902.
27. Maddalena, D. and Johnston, G.A.R., J. Med. Chem., 38 (1995) 715.
28. So, S.-S and Karplus, M., J. Med. Chem., 39 (1996) 5246.
29. Winkler, D.A. and Burden, F.R., Quant. Struct.-Act. Relat., 17 (1998) 224.
30. Drapper, N.R. and Smith,H., Applied Regression Analysis, J. Wiley & Sons, New York, NY, 1981, p. 93.
31. Hahn, M. and Rogers, D., J. Med. Chem., 38 (1995) 2091.
32. Baroni, M., Constantino, G., Cruciani, G., Riganelli, D., Valigi, R. and Clementi, S., Quant. Struct.-Act. Relat., 12 (1993) 9.
33. Motoc I. and Marshall, G.R., Chem. Phys. Lett., 116 (1985) 415.
34. Kubinyi, H, Hamprecht, F.A. and Mietzner, T., J. Med. Chem., 41 (1998) 2553.
35. Coats, E.A., In Kubinyi, H., Folkers, G. and Martin, Y.C. (eds), 3D QSAR in Drug Design., Vol 3, Recent Advances, Kluwer/ESCOM, Dordrecht, 1998, pp. 199–213.
36. Klein, C. T., Viernstein, H. and Wolschann, P., Sci. Pharm., 68 (2000) 15.
37. Golbraikh, A. and Tropsha, A., J. Mol. Graph. Mod., 20 (2002) 269.