

# Mapping Property Distributions of Molecular Surfaces: Algorithm and Evaluation of a Novel 3D Quantitative Structure–Activity Relationship Technique

Nikolaus Stiefl and Knut Baumann\*

Department of Pharmacy and Food Chemistry, University of Wuerzburg, Am Hubland, D 97074 Wuerzburg, Germany

Received October 21, 2002

A novel molecular descriptor called MaP (mapping property distributions of molecular surfaces) is presented. It combines facile computation, translational and rotational invariance, and straightforward interpretability of the computed models. A three-step procedure is used to compute the MaP descriptor. First, an approximation to the molecular surface with equally distributed surface points is computed. Next, molecular properties are projected onto this surface. Finally, the distribution of surface properties is encoded into a translationally and rotationally invariant molecular descriptor that is based on radial distribution functions (distance-dependent count statistics). The calculated descriptor is correlated with biological data through chemometric regression techniques in combination with a variable selection. The latter is used to identify variables that are highly relevant for the model and hence for its interpretation. Three applications of the new descriptor are presented, each representing a different area of 3D-QSAR. For reasons of comparability, the new descriptor was tested on the steroid “benchmark” data set. Furthermore, a highly diverse data set with potentially eye-irritating compounds was studied, and third, a set of flexible structures with a modulating effect on the muscarinic  $M_2$  receptor were studied. Not only were all models highly predictive but interpretation of the back-projected variables into the original molecular space led to biologically and chemically relevant conclusions.

## Introduction

The basis for various structure–activity correlation techniques is the description of chemical structures by means of numbers. During the past decades, a vast number of molecular descriptors have been developed.<sup>1</sup> For instance, computer programs such as Dragon<sup>2</sup> compute up to 1800 descriptors. Such molecular descriptors may have very different complexity but can be classified according to their “dimensionality”. For instance, one-dimensional descriptors (1D) include bulk properties and physicochemical properties, such as  $\log P$  and molecular weight. The 2D descriptors require knowledge of the molecular graph, i.e., the way the different atoms are connected and include, among others, predefined structural fragments,<sup>3</sup> connectivity indices,<sup>4</sup> atom pairs,<sup>5</sup> or the distribution of atomic properties in a mathematical graph.<sup>6</sup> The 3D descriptors are based on the Cartesian coordinates of the molecule and include, among others, scalar descriptors, such as volume and surface area, and multivariate techniques such as CoMFA,<sup>7</sup> CoMSIA,<sup>8</sup> HASL,<sup>9</sup> CoMMA,<sup>10</sup> EVA,<sup>11</sup> WHIM,<sup>12</sup> MS-WHIM,<sup>13</sup> 3D-MoRSE,<sup>14</sup> and GRIND.<sup>15</sup> Many 3D techniques (CoMFA, CoMSIA, GRIND) generate an output that is easily visualized and can be used as an “idea generator” for new drug candidates. This visualization can be described as a back-projection of the model into the original molecular space. Models that allow a direct suggestion of new compounds to synthesize are very helpful in the drug discovery process.

A widespread feature of 3D descriptors is their dependence on the orientation of the molecules in space and toward each other. To be able to compare the molecules, they need to be aligned. The alignment determines to what extent the descriptors differ from one molecule to the next. Consequently, it substantially influences the results of the evaluation. Hence, significant and relevant results can only be expected if the alignment was carried out properly. Often, the need for an alignment limits the application of certain descriptors to homogeneous data sets, and even then the alignment is not always easily performed. As a consequence, different groups started to develop alignment-independent molecular descriptors. The descriptors can be split into two groups: not back-projectable ones (e.g., WHIM, MS-WHIM, EVA, 3D-MoRSE, CoMMA) and back-projectable alignment-independent 3D molecular descriptors (e.g., GRIND). The invention of the GRIND descriptor<sup>15</sup> marked a new era for translationally and rotationally invariant descriptors. GRIND encodes non-covalent binding forces relevant for receptor–ligand interaction and is at the same time easily interpretable because it can be back-projected meaningfully into the original molecular space. The MaP (mapping property distributions of molecular surfaces) descriptor that is presented in this contribution also aims at translational and rotational invariance and easy interpretability. It is similar in spirit to GRIND, though different in important details. First, MaP does not start with a grid-based field of interaction energies around the molecule being encoded (GRIND uses the GRID force field<sup>16</sup>) but with the molecular surface. Second, the variables used

\* To whom correspondence should be addressed. Phone: ++49 931 888-5473. Fax: ++ 49 931 888-5494. E-mail: knut.baumann@mail.uni-wuerzburg.de.

to compute the descriptor are categorical in nature (H-bond donor, H-bond acceptor, hydrophobic, and hydrophilic) rather than continuous (interaction energies). Third, owing to the categorical variables, a different mathematical transformation is used to encode the surface properties. Therefore, different pieces of information of the molecules are used to represent them numerically for structure–activity correlations.

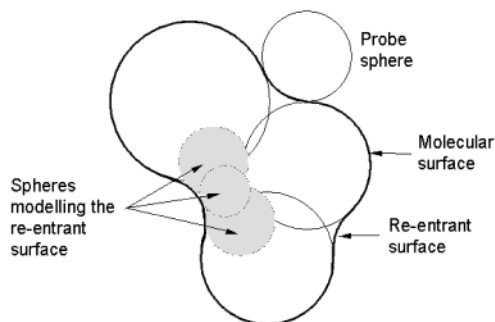
Although translationally and rotationally invariant descriptors (TRI descriptors) do not change when the molecule is rotated or translated, they are sensitive to different conformations of the molecule (as is every truly 3D descriptor; exceptions are 4D descriptors that are 3D in nature and were specifically designed to take conformational flexibility into account<sup>17</sup>). Consequently, MaP needs reasonable conformations of the molecules under study as input. Therefore, it is beneficial to match conformations across a data set where possible rather than using randomly or arbitrarily selected conformers. If a successful alignment rule already exists, MaP can take advantage of this rule. If there is no obvious way to align the molecules, MaP can also do without an alignment rule, as will be exemplified with two heterogeneous data sets.

The article is organized as follows. First, the method is described in detail and then relations to other approaches are discussed. Next, the application of MaP to different data sets will be studied to give an impression of the capabilities of the new descriptor. The example data sets are selected in order to mirror the application to different areas of 3D-QSAR. They include the well-studied steroid data set<sup>18</sup> as a benchmark data set, the eye irritation data set<sup>19</sup> as an example of a very heterogeneous data set, and the M<sub>2</sub> modulator data set<sup>20–26</sup> as an example of some highly flexible structures.

## Methods

**Geometry Optimization.** The geometries of the steroid data set were generated by CORINA<sup>27</sup> and are identical to those that were used in a previous study.<sup>14</sup> Geometries for the eye-irritation data set were obtained in a two-step procedure. First, 2D connection tables of the structures were converted to 3D coordinates using CONCORD.<sup>28</sup> These geometries were then refined with molecular mechanics and the Tripos force field as implemented in Alchemy 2000.<sup>28</sup> Owing to their high flexibility, M<sub>2</sub> modulators were treated in a special way that is described in the Results and Discussion.

**Calculation of the Molecular Surface.** First test runs were carried out with molecular surfaces generated with the well-known Connolly algorithm<sup>29</sup> implemented in SPOCK<sup>30</sup> and the MSMS algorithm.<sup>31</sup> After inspection of the preliminary results, it was found that the number of surface points generated is not equally distributed across the surface but the distribution was found to be dependent on the shape of the molecule. This is due to the fact that the number of surface points in saddle-shaped areas is much higher than on other parts of the surface. Since the MaP algorithm is based on count statistics of surface point pairs, this could lead to inconsistent models. For instance, if the data set consists of differently shaped molecules (which is usually the case), the size and shape information that is implicitly encoded in the absolute number of the counts

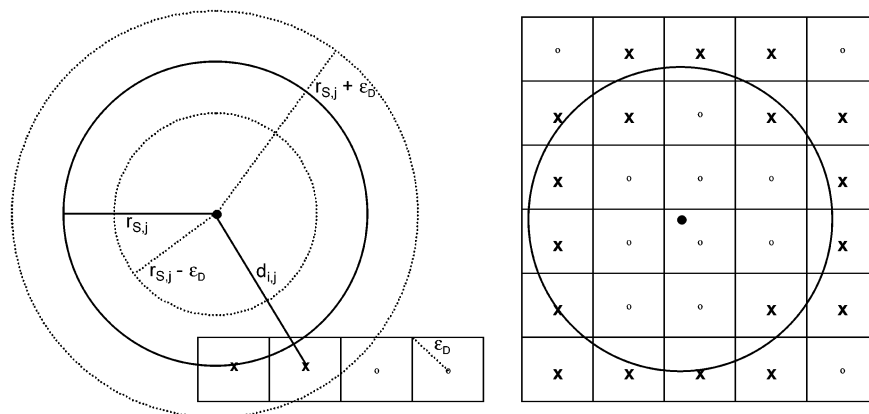


**Figure 1.** Illustration of the molecular and re-entrant surface. The gaps between two or more atoms that are not accessible to the probe sphere rolling on the van der Waals surface are modeled in the GEPOL algorithm with a superimposition of several additional spheres (shown in gray).

may get lost. Moreover, if for some molecules important properties are located in areas of low point density whereas for others the same properties are located in areas with higher point density, the encoded information becomes blurred and the count statistics may no longer be comparable between molecules. Avoiding this potential problem was achieved by implementing the GEPOL<sup>32</sup> algorithm and modifying it to generate a grid-based surface with equally distributed surface points. In the first step two sets of spheres are generated whose surfaces approximate the molecular surface. First, there are spheres around each atom with the respective van der Waals radii. Overlapping the van der Waals spheres of the atoms results in the van der Waals surface. Second, there are spheres to model the differences between the van der Waals surface and the molecular surface. Here, the molecular surface is defined as the contact surface that is accessible by the inward-facing part of a probe sphere (usually a water molecule, represented by a sphere of radius of 1.4 Å) as it rolls on the van der Waals surface of the target molecule. This second set of spheres models the so-called re-entrant surface. The re-entrant surface regions occur where the gaps between two or more atoms are too narrow for the probe to penetrate. In Figure 1 the different sets of spheres and the re-entrant surface are illustrated. Both sets of spheres, the one consisting of the van der Waals spheres around the atoms and the one modeling the re-entrant surface, are treated alike in the algorithm for generating equally distributed surface points, as will be outlined below. After the spheres were generated, the molecules are analyzed to find the atom type exhibiting the largest van der Waals radius and the extremes in the *x*, *y*, and *z* directions. These values are used to define a regular three-dimensional grid around the molecule where the vertexes are derived by adding or subtracting the maximal van der Waals radius to the aforementioned extreme *x*, *y*, and *z* values. This ensures that the grid is large enough to surround the compound. The grid spacing can be varied by the user. Grid points are defined as surface points if they fulfill the following requirement: they need to be on the van der Waals radius of an atom or on the radius of a sphere modeling the re-entrant surface, which is fulfilled if

$$(r_{S,j} - \epsilon_D) < d_{i,j} \leq (r_{S,j} + \epsilon_D)$$

where  $d_{i,j}$  equals the Euclidean distance of the *i*th grid



**Figure 2.** Surface generation for a single atom with the dGEPOL algorithm. Left: the two radii displayed as dotted lines represent the atomic van der Waals radius plus and minus the maximum discretization error  $\epsilon_D$  (illustrated on the right-hand bottom corner). If the grid point lies within these two radii, it is included as a surface point. Right: example surface generated. The  $\times$  signs depict grid points included, whereas unfilled dots do not fulfill the equation stated in the text.

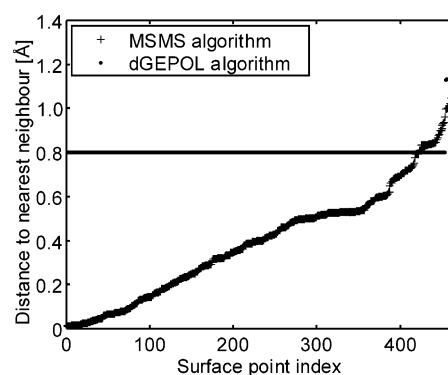
point to the  $j$ th sphere of the molecule,  $r_{S,j}$  is the radius of the  $j$ th sphere, and  $\epsilon_D$  is the maximum discretization error that depends on the grid spacing.  $\epsilon_D$  equals the Euclidean distance of a grid point to the center of a cube composed of eight grid points. It can be expressed in terms of the grid spacing (gsp) as follows:

$$\epsilon_D = \sqrt{\frac{3}{4}} \text{gsp}$$

The condition is checked for all grid points with respect to all spheres in the molecule. Figure 2 illustrates the generation of surface points for a single atom. For the sake of clarity, the dimensionality is reduced to two dimensions.

Unfortunately, the grid-based approach to approximate the surface bears the potential risk that the translational and rotational invariance of the descriptor is lost. This is because the assignment of the lattice points as surface points depends on the distance to the closest sphere (atom). As a consequence of the discretization error  $\epsilon_D$ , this distance may vary depending on the molecule's orientation within the grid box. Moreover, the assignment of the respective surface property to a particular surface point may also change because this assignment is also distance-dependent (see below). Precautions are taken to avoid this undesirable feature of the descriptor. First, the molecule is oriented in the grid box along the principal moments of inertia of the molecule.<sup>1,33</sup> Second, the origin of the grid box and the center of mass of the molecule always coincide. These two precautions render the surface (and thus the descriptor) translationally and rotationally invariant because irrespective of the initial position and orientation of the molecule in space the resulting surface will always be identical. Moreover, the distances between the atoms and the surface points are also always identical. The orientation of the molecule within the grid box used for computing the molecular surface must not be confused with an alignment carried out for descriptors such as CoMFA. The orientation step carried out here rotates the molecules with respect to the coordinate system and not with respect to a reference molecule.

Comparing Connolly and MSMS surfaces with those generated by the discretized GEPOL (dGEPOL) algorithm shows that the surfaces are quite similar when a



**Figure 3.** dGEPOL algorithm generates surface points that are equally distributed on the approximation to the molecular surface. It can be seen that in most cases the distance to the nearest-neighboring surface point equals the surface resolution (0.8 Å). In some cases, this distance equals  $(\sqrt{2})(0.8)$  Å, which occurs if the nearest neighbor is not the next horizontal or vertical grid node but the next diagonal grid node. Standard algorithms for computing the molecular surface area such as the Connolly and the MSMS algorithm do not show this feature. Here, this is shown for the MSMS algorithm. It should be noted that the distances to the nearest-neighboring surface point were sorted in both cases.

grid spacing (gsp) of 0.8 Å or lower is used. Hence, a default value of 0.8 Å was used throughout this work. The difference between the dGEPOL, and the MSMS and Connolly surface is simply that the dGEPOL surface consists of equally distributed surface points. To illustrate this difference, the distance to the nearest-neighboring surface point for aldosterone was computed, where the number of surface points was adjusted to be approximately equal. The result is shown in Figure 3. The MSMS surface used a density of 1.4 vertexes/Å<sup>2</sup>. It can be seen that the dGEPOL surface results in the desired distribution of surface points, whereas the distribution for the MSMS surface varies markedly. It should be noted that in the dGEPOL algorithm two possible distances to the nearest neighbor exist: 0.8 Å (horizontal and vertical) and  $(\sqrt{2})(0.8)$  Å (diagonal). Results for other molecules are very similar. Increasing the triangulation density (vertex/Å<sup>2</sup>) of the MSMS algorithm did not improve the distribution. These latter results are not surprising, since the Connolly and MSMS algorithms were not developed to generate equally distributed surface points but to compute the

surface area. Hence, for this particular application, the dGEPOL is better suited than the Connolly and the MSMS algorithms. Note that alternative methods for the calculation of grid-based surfaces are also available.<sup>34</sup>

**Mapping the Molecular Properties.** The basic idea behind the MaP descriptor is that interactions between biologically active compounds and their corresponding receptors can be described by steric features, hydrophobic interactions, and electrostatic interactions. The last is often described by H-bond donor–acceptor interactions. Owing to the fact that interactions form between the surfaces of receptor and ligand, compounds are classified according to properties representing the potential for interaction that can be projected onto the molecular surface. In the current version of MaP, four different properties are used: hydrophobicity (L), hydrophilicity (H), H-bond acceptor (A), and H-bond donor (D). Steric features of the molecules are encoded implicitly through the distribution of distances between surface point pairs, as will become clear soon.

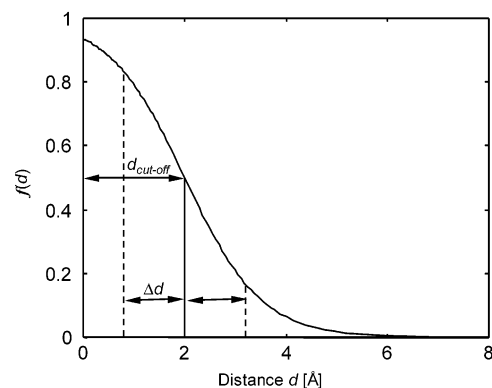
The H-bond acceptor–donor assignment algorithm follows the simple rules implemented in Tripos' SYBYL. A list of electronegative elements (N, O) is used to specify interesting atoms. All hydrogen atoms connected directly to these atoms are defined as donor atoms. All partially negative charged atoms from that list that do not exceed a maximum number of bonds ( $N = 3$ ,  $O = 2$ ) connected to them are defined as acceptor atoms. Since atom charges are used in a qualitative way, Gasteiger–Hückel charges are most often sufficient for this application. Moreover, fluorine may also be specified as an acceptor atom by the user. It should be noted, however, that fluorine acts only rarely as hydrogen acceptor.<sup>35</sup> Since the  $C(sp^3)–F$  is a better hydrogen bond acceptor than  $C(sp^2)–F$ ,<sup>35</sup> only the former is considered as a potential acceptor. The atomic hydrophobicity was assigned on the basis of the fragmental approach by Ghose and co-workers.<sup>36</sup> Mapping of the atomic properties was done according to the following rules. The atom closest to a particular surface point is defined as its base atom. If the base atom is classified as an H-bond acceptor (A) or H-bond donor (D), the surface point is assigned the respective property. Only surface points that are not classified as the latter can be assigned the hydrophobic (L) or hydrophilic (H) attribute. Because there is no physical rationale for a certain distance dependence when mapping the hydrophobicity (hydrophilicity) to a particular surface point, a Fermi-type function  $f(d)$  following Brickmann and co-workers was implemented:<sup>37</sup>

$$f(d_{i,j}) = \frac{1}{\exp[a(d_{i,j} - d_{\text{cutoff}})]} + 1$$

with

$$a = \frac{2}{\Delta d}$$

where  $d_{i,j}$  is the distance of the  $i$ th surface point to the  $j$ th atom,  $2 \Delta d$  defines the range wherein the function decays, and  $d_{\text{cutoff}}$  is some cutoff value that is termed proximity distance. This proximity distance should be larger than the largest van der Waals radius of any



**Figure 4.** Shape of the Fermi-type function with default values used in this study ( $d_{\text{cutoff}} = 2 \text{ \AA}$  and  $\Delta d = 2 \text{ \AA}$ ).

atom in the data set under consideration. The function is smooth and finite over the entire range of definition. Moreover,  $f(d)$  decays from values close to unity to values close to zero in the interval

$$d_{\text{cutoff}} - \Delta d < d < d_{\text{cutoff}} + \Delta d$$

and is therefore not susceptible to overcompensation of local effects by long-range dependencies provided that  $\Delta d$  is chosen appropriately. Figure 4 shows the function with the default values used in this work. The actual hydrophobic potential (HP) assigned to a particular surface point is given as

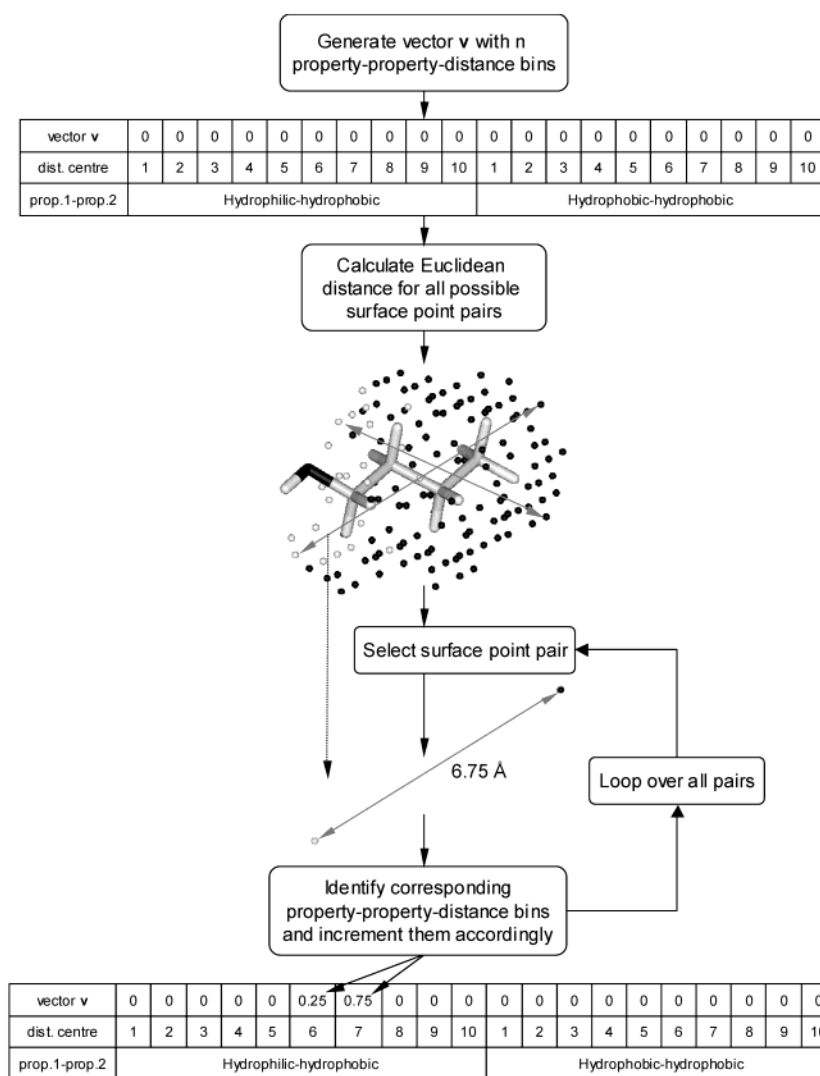
$$\text{HP} = \sum_{i=1}^{n_A} A_i f(d_{i,j})$$

where  $n_A$  is the number of atoms,  $A_i$  is the contribution of the  $i$ th atom to the octanol–water partition coefficient as defined by Ghose and co-workers,<sup>36</sup> and  $d_{i,j}$  is the distance of the  $i$ th atom to the  $j$ th surface point. A figure showing the result of the property mapping onto the molecular surface can be found in Supporting Information. The figure depicts properties mapped onto the surface by the program MOLCAD<sup>37</sup> (properties represented continuously) and the respective categorized surface generated with the described algorithm.

**Calculating the Descriptor.** The MaP descriptor is a vector that is calculated in a three-step binning procedure. The dimension of the MaP vector ( $n$ ) is defined by the number of properties included ( $p$ ) and the number of distance bins ( $c$ ) in the following manner:

$$n = \left( \frac{p(p+1)}{2} \right) c$$

where  $c$  depends on a user-defined resolution (res), which is set to  $1 \text{ \AA}$  by default.  $c$  is obtained as follows: the distance bins range from  $\geq (k-1)(\text{res}) + \text{res}/2 \text{ \AA}$  to  $< (k)(\text{res}) + \text{res}/2 \text{ \AA}$ , where  $k$  is initialized with 1 and is incremented by 1 until the upper boundary is larger than the largest distance between two surface points in the data set ( $d_{\text{max}}$ ). Then the generation of distance bins stops and  $c = k$ . The respective bin centers (bc) are located at  $(k)(\text{res}) \text{ \AA}$ . Next, for each property combination of two surface points (e.g.,  $A \leftrightarrow A$ ,  $A \leftrightarrow D$ ,  $A \leftrightarrow H$ ,  $A \leftrightarrow L$ , ...), a segment of  $c$  vector entries is created, resulting in  $n$  bins where each bin corresponds to one specific property–property–distance combination. Consequently,



**Figure 5.** Schematic description of the algorithm for computing the MaP descriptor. Displayed are the hydrophilic and hydrophobic surface points of butanol (i.e., no acceptor or donor regions are displayed). The midpoint of the distance interval is given in the figure. Fuzzy counts are used for incrementing the matching bins (see text).

the dimension of the MaP descriptor depends on  $p$ ,  $res$ , and  $d_{max}$ . The algorithm for the calculation of the descriptor is displayed schematically in Figure 5 and is described in detail in the following.

**Step 1.** A vector  $\mathbf{v}$  of dimension  $n$  is allocated and initialized with zeros. The vector  $\mathbf{v}$  consists of  $(p(p + 1))/2$  segments, one for each property combination of dimension  $c$ . Each segment is subdivided into  $c$  distance bins with the aforementioned boundaries.

**Step 2.** For each surface point pair, the Euclidean distance ( $d_{i,j}$ ) between the  $i$ th and the  $j$ th point is calculated as

$$d_{i,j} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2}$$

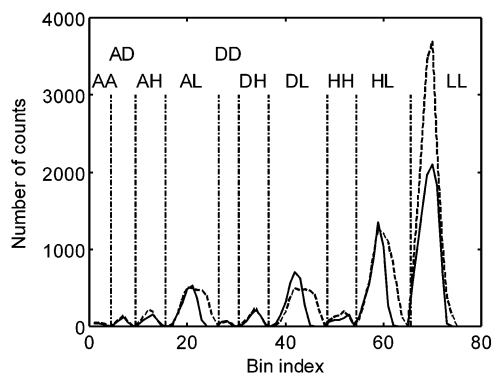
where  $x$ ,  $y$ , and  $z$  are the corresponding coordinates for surface points  $i$  and  $j$ .

**Step 3.** The matching bins of each surface point pair in vector  $\mathbf{v}$  are incremented. Matching means that the segment in  $\mathbf{v}$  corresponding to the property combination of the particular surface point pair is identified first. Within this segment, the two distance bins closest to the actual distance  $d_{i,j}$  are incremented proportionally.

The exact increment for each bin depends on the deviation of  $d_{i,j}$  from the closest distance bin center ( $bc$ ). The increment for the main bin is given as follows:

$$inc = 1 - \frac{abs(bc - d_{i,j})}{res}$$

That means that the maximum increment for the main bin amounts to 1.0 ( $bc = d_{i,j}$ ), whereas the minimum increment equals 0.5 ( $d_{i,j} = res/2$ ). Depending on the sign of  $bc - d_{i,j}$ , the bin above or below the bin centered at  $bc$  is incremented with the remainder of  $1 - inc$ . This concept of fuzzy counts is due to Sheridan and co-workers.<sup>38</sup> A different version was also described by Brown and Martin.<sup>39</sup> Fuzzy counts improve on the standard method where two very close distances around a bin boundary will not contribute to the respective neighboring bin. In the case of MaP, the advantage of using fuzzy counts to compute count statistics is that differences between molecules in terms of the orientation in the grid box (used for the surface computation) are leveled off. Since each molecule in the data set shows different moments of inertia, it will be oriented differ-



**Figure 6.** MaP descriptor for cyclohexanol (dotted line) and hexanol (solid line), which are compounds of the eye-irritation data set. Displayed is the union of nonconstant columns of both compounds. The dashed vertical lines display the property combination boundaries, where A means acceptor, D donor, H hydrophilic, and L hydrophobic. It can easily be seen that the peaks in the hexanol curve are much broader compared to those of cyclohexanol. This is due to the entries in the long-distance bins. Moreover, hexanol shows a far higher number of hydrophobic surface point pairs (LL). Both findings can be explained with the different shape of the two molecules.

ently in the grid box. Using the fuzzy counts helps to minimize the influence of the discretization error. In addition to that, back-projection of selected variables can also be controlled with the help of fuzzy counts. By default, all connections between surface point pairs that fall into the selected bin are shown (i.e.,  $\text{inc} > 0.5$ ). This gives an impression about the size and the location of the involved surface patches. Since these plots sometimes look complex, the user can choose the level of the increment for back-projecting variables. Only surface point pairs exceeding this level are then back-projected onto the molecular surface for the purpose of interpretation. This back-projection resembles the way pharmacophores are mapped onto molecules except that surface points rather than the atoms themselves are used. A variant of this technique was used in the context of QSAR before.<sup>40</sup> For the future, merging single connections pointing in the same direction to hyperboloid-like geometric figures is planned. The absolute number of counts is then reflected by the size of the hyperboloid and a color-coding scheme.

The result of the algorithm outlined above is a set of  $(p(p+1))/2$  radial distribution functions. Put differently, for each property combination of surface point pairs, i.e., for each segment in vector  $\mathbf{v}$ , a distance-dependent histogram is generated. Owing to the equally distributed surface points, this histogram encodes information about the size and shape of the molecule as well as the property distribution along the molecular surface. It may be argued that encoding both property distribution and molecular size at the same time leads to confounded variables. However, depending on the structures under scrutiny, those variables that depend on size and those that encode the arrangement of specific properties are easily distinguished by mapping the most informative variables (see below) onto selected molecules of the data set. In Figure 6 the entire MaP vector for hexanol and cyclohexanol is shown. As an example, variable LL (i.e., connections between hydrophobic surface points) strongly depends on the size and shape in the case of hexanol and cyclohexanol. In the extended hexanol, far more

counts of short and intermediate distances for LL are observed. On the other hand, variable AD (acceptor  $\leftrightarrow$  donor), which is caused by the hydroxyl group, does not depend on size at all. Variables AL (acceptor  $\leftrightarrow$  hydrophobic) and HL (hydrophilic  $\leftrightarrow$  hydrophobic) take an intermediate position. The counts for short and intermediate distances for both molecules are roughly identical, whereas hexanol also shows counts for longer distances because of its shape.

**Relation to Other Approaches.** In the following, we briefly outline the relation of MaP to other approaches, and in particular, we show commonalities and differences between GRIND and MaP. GRIND was the first TRI descriptor that encodes noncovalent binding forces and is at the same time easily interpretable. GRIND uses GRID fields with different probes to characterize the properties of the molecules under study. Typically, the O probe (H-bond acceptor), the N1 probe (H-bond donor), and the DRY probe (hydrophobicity) are employed. Since properties such as H-bond donor-acceptor and hydrophobicity-hydrophilicity are known to mediate the binding of a drug to its receptor,<sup>41</sup> it is natural that GRIND and MaP employ these properties for their calculations. However, the way this set of properties is calculated differs completely. Whereas GRIND uses the GRID force field to obtain continuous interaction energies, MaP uses simple schemes to assign property categories to surface points. Owing to these mathematically different variables (continuous vs categorical), the mathematical transformations of the raw data are different, and as a consequence, the information encoded also differs significantly (see below). Both techniques use a distance-dependent mathematical function. Since distances between two points of a rigid object are invariant to translation and rotation of the object, distance-dependent mathematical functions are often used for TRI descriptors<sup>5,14,38,42-45</sup> and are not limited to GRIND and MaP. GRIND uses a particular form of the autocorrelation technique.<sup>46,47</sup> The so-called maximum auto- and cross-correlation (MACC-2) transform<sup>15</sup> was a landmark invention because it combines the translational and rotational invariance of autocorrelation techniques with easy interpretability. In the MACC-2 transform, only the maximum product of interaction energies per distance bin is considered and enters the molecular descriptor. Since GRIND uses the entire GRID field around the molecules under study (thousands of points), a data reduction step is employed before the MACC-2 transform is computed. This is accomplished by clustering the interaction energy values. The remaining values are then transformed into three autocorrelograms and three cross-correlograms (i.e., all combinations of probes are computed). MaP also computes statistics for all combinations of surface point properties to thoroughly characterize the distribution of surface point properties (categories). Owing to the categorical nature of the MaP raw data, the occurrences of surface point pairs separated by a particular distance are counted. This is similar to the geometric atom pairs (ag) and binding pairs (bg) of Sheridan and co-workers,<sup>38</sup> and the potential-pharmacophore-point (PPP) distances of Brown and Martin<sup>39</sup> that were used for similarity calculations. The key difference is that we extend these atom-based descriptors to molecular surfaces. Atom

pairs and PPP descriptors are based on the idea that the occurrence of particular atom pairs or PPP pairs will be different in active and inactive substances. The same can be said for MaP with respect to surface point pairs. However, there is one important difference. MaP counts reflect the size of surface patches (absolute number of counts) of selected properties. For instance, the acceptor area around a carbonyl group is larger than that of a hydroxyl group. As a result, the counts between two carbonyl groups are larger than those between a carbonyl and a hydroxyl group. If the exchange of a carbonyl for a hydroxyl group affects activity, such a variable will be helpful for describing the differences (e.g., steroids). Consider counts between hydrophobic surface points at short distances, say 1 or 2 Å, as another example. These variables encode the size of the hydrophobic surface area and may be related to drug transport or penetration across certain barriers. Counts between surface points at larger distances describe the relative positions of surface patches with selected properties and their size. GRIND, on the other hand, uses the maximum auto- or cross-correlation (i.e., the largest product of interaction energies) per distance bin. That means that for each distance bin only the strongest interaction pair is encoded. Consequently, apart from their different raw data and mathematical transformations of these data, the information encoded by the two descriptors is rather different. A common feature of both descriptors is how they are back-projected into the original molecular space. These procedures resemble the way pharmacophores are usually mapped onto molecules<sup>40</sup> except that points in space rather than the atoms themselves are used. MaP's back-projection is different in that all connections between surface point pairs determining the count of a certain variable are shown. In this way, often size and shape information related to biological activity is naturally displayed (see, for example, Figures 8 and 10).

Currently the MaP descriptor is restricted to connections between two surface points. This limitation may result in counts of the same variable that stem from different parts of the molecule. Theoretically, if only one part of the counts is relevant for biological activity, this may lead to inconsistencies in the QSAR model. However, this scenario did not occur for the data sets analyzed thus far. One of the possible explanations for this behavior is that MaP variables are to a certain extent redundant (i.e., the same phenomenon can be explained by different variables), of which those that are least confounded with other phenomena are picked by the variable selection procedure described in the next section.

**Chemometric Methods.** For a data set, the MaP vectors of all  $m$  molecules represented by  $n$  variables are stacked into a matrix of dimensions  $m \times n$ . Next, all variables that are constant, i.e., show a variance of zero, are excluded. Principal component regression (PCR) or partial least-squares regression (PLS) is used to compute the regression models. For the theory of PCR and PLS, see Martens and Naes.<sup>48</sup> The number of latent variables (LV) for the full model (no variable selection; see below) was determined as the first local minimum in the LV vs the leave-one-out cross-validated root-mean-squared error of prediction (RMSEP<sub>CV-1</sub>) plot.<sup>49</sup>

RMSEP<sub>CV-1</sub> is defined as

$$\text{RMSEP}_{\text{CV-1}} = \sqrt{\frac{1}{m} \sum_{i=1}^m (y_{i,\text{obs}} - y_{i,\text{pred}})^2}$$

where  $y_{i,\text{obs}}$  is the observed property value of the  $i$ th object that was left out and  $y_{i,\text{pred}}$  is the corresponding predicted property value of this object. From RMSEP<sub>CV-1</sub>, the respective  $R^2_{\text{CV-1}}$  was calculated as

$$R^2_{\text{CV-1}} = 1 - \frac{(\text{RMSEP}_{\text{CV-1}})^2 m}{\sum_{i=1}^m (y_{i,\text{obs}} - \bar{y})^2} = 1 - \frac{\text{PRESS}}{\text{SYY}}$$

where  $\bar{y}$  is the mean of all responses. Moreover, the usual coefficient of determination ( $R^2$ ) and the root-mean-squared error of calibration (RMSEC) are used as figures of merit and were computed as

$$R^2 = 1 - \frac{\sum_{i=1}^m (y_{i,\text{obs}} - \hat{y}_i)^2}{\text{SYY}}$$

and

$$\text{RMSEC} = \sqrt{\frac{\sum_{i=1}^m (y_{i,\text{obs}} - \hat{y}_i)^2}{m - \text{df}}}$$

where  $\hat{y}_i$  is the fitted value and df is the number of degrees of freedom used by the regression model. As a rough approximation to df, the number of latent variables is used here.

To identify the most informative variables (MIV) of the model, a reverse-elimination-method tabu search<sup>50-52</sup> (REM-TS) is employed. REM-TS is a stepwise variable selection method that is guided by the principle of "steepest descent, mildest ascent". In each iteration of the REM-TS procedure, a variable is either added to the model or removed from the model. If there are moves that improve the objective function, the one with the largest improvement is executed (steepest descent). If there are only detrimental moves, the one with the least impairment of the objective function is executed (mildest ascent). Since REM-TS also accepts detrimental moves, it cannot get trapped in local optima. During one iteration, the status of each variable is switched systematically (in → out, out → in) to determine the best move. That means that the search trajectory of REM-TS is deterministic. The management of the search history is done in a way to avoid the situation where one and the same solution is visited more than once (so-called strict TS). If a move would lead back to an already visited solution, it is set tabu and cannot be executed. This is where the name of the search heuristic stems from. The only user-defined parameter for REM-TS is a termination criterion. In this work, the search was terminated during the first descent when the changes in objective function were only small (<3%). This often amounted to a forward selection with an early stopping

rule and does not use the full potential of REM-TS. This restriction was applied to lower the risk of chance correlation. Owing to its greedy search philosophy and its systematic search procedure, REM-TS is extremely efficient and search runs for a structure descriptor such as MaP, i.e., for matrices with up to a few hundred variables, take only a few minutes on a 1 GHz personal computer. The MIV selected by REM-TS are back-projected into the original molecular space for interpretation purposes. In variable selection, differences between PCR and PLS tend to be small because both regression techniques make extensive use of the response information while selecting variables.<sup>53</sup> Since our PCR routines are faster, only PCR was used in combination with variable selection. Note that PCR models with as many latent variables as explanatory variables are identical to the respective multiple linear regression (MLR) models. If the selected variables are largely orthogonal to each other, REM-TS often ends up with the respective MLR model.

The objective function of a variable selection technique is the most crucial part of the entire procedure. If the objective function is chosen inappropriately, the predictive power of the chosen models will suffer. It was shown that the widely used leave-one-out cross-validation (LOO-CV), when used as an objective function in a variable selection procedure, yields statistically inconsistent results. Broadly speaking, it can be said to lead to overfitting of the data.<sup>54,55</sup> This was realized early by Cruciani and Clementi who use a more stringent validation procedure for their variable selection procedure.<sup>56–58</sup> In this work, a leave-multiple-out cross-validation<sup>59</sup> (LMO-CV) procedure was used for variable selection to effectively avoid overfitting.<sup>53</sup> The number of cross-validation runs ( $B$ ) was always set to 3 times the number of objects in the data set ( $B = 3m$ ) to achieve a reasonably low variance of the estimated prediction error. The percentage level of objects left out was set to 50, which was found to be a reasonable default value in earlier studies.<sup>60</sup> Leave-multiple-out cross-validated root-mean-squared errors and the respective coefficients of determination were computed as

$$\text{RMSEP}_{\text{CV}-k} = \sqrt{\frac{1}{B} \sum_{b=1}^B \frac{1}{k} \sum_{i=1}^k (y_{b,i,\text{obs}} - y_{b,i,\text{pred}})^2}$$

where  $B$  is the number of cross-validation runs,  $k$  is the number of objects left out (nearest integer to  $0.5m$ ),  $y_{b,i,\text{obs}}$  is the observed property value of the  $i$ th object in the  $b$ th cross-validation run that was left out,  $y_{b,i,\text{pred}}$  is the corresponding predicted property value of this object, and the subscript  $k$  indicates the number or the percentage of objects left out. From the  $\text{RMSEP}_{\text{CV}-k}$  value, the respective cross-validated squared multiple correlation coefficient  $R^2_{\text{CV}-k}$  was computed as  $R^2_{\text{CV}-1}$  (exchange  $\text{RMSEP}_{\text{CV}-1}$  for  $\text{RMSEP}_{\text{CV}-k}$ ). It must be noted that the estimate of the prediction error obtained by LMO-CV is biased upward.<sup>61</sup> Consequently, results obtained by LMO-CV will always be worse than those obtained by LOO-CV. For the sake of comparability, the results of LOO-CV are also given. For the full models, i.e., no variable selection applied, both figures of merit are also given. Whenever no test set was available for independently validating the variable selection results,

a scrambling test was performed in order to check for the risk of chance correlations.<sup>62,63</sup> The employed permutation test is based on the repetitive randomization of the response vector. In each cycle of the test, the response vector is randomly rearranged, the entire selection procedure (using the same settings as for the original model,  $B = 3m$ ,  $k = 0.5m$ , termination; no marked change during first descent) is carried out on the scrambled data, and  $R^2_{\text{CV}-k}$  is recorded for each cycle. All computations were done from scratch after scrambling the responses, since scrambling of only the finally selected model yields far overoptimistic results.<sup>64</sup> If the majority of the  $R^2_{\text{CV}-k}$  values of the scrambled data sets is much lower than the  $R^2_{\text{CV}-k}$  value of the original data set, it can be concluded that the derived model is relevant. The number of permutations for each test was set to 500, since scrambling and running the entire selection procedure are computationally quite expensive.

If the data were split into a training set (66%) and a test set (33%), Kennard–Stone's CADEX algorithm<sup>65,66</sup> on the structure descriptor data was used. The CADEX algorithm, which results in a balanced and representative split, was applied here because it performed well in training set selection in another study.<sup>65</sup> We note, however, that a balanced rather than a random split into training and test sets may underestimate the true prediction error.<sup>67</sup>

**Parameter Settings.** The default parameters of the entire procedure are set as follows. All available properties were included in the MaP generation (i.e., A, D, H, L). The default grid spacing (gsp) was set to 0.8 Å. This led to a sensible number of surface points while describing the molecular surface reasonably well. The proximity distance  $d_{\text{cutoff}}$  and  $\Delta d$  of the Fermi-type function were set to 2 and 2 Å, respectively. That way, local hydrophobic effects were not overcompensated by the rest of the molecule and a distinction between aromatic ring systems and aliphatic side chains was possible. The cutoff value for the specification of hydrophobicity–hydrophilicity of a surface point was set to zero; i.e., surface points with a negative value are defined as hydrophilic and vice versa. If a distinction between strongly hydrophobic (abbreviated Ls) and weakly hydrophobic side chains (abbreviated Lw) was necessary, an additional cutoff value was set to 0.12. The MaP descriptor also includes the specification of the acceptor and donor strength, but this feature was not used in this study, since no significant differences in model quality were achieved. The resolution (res) for the radial distribution functions (distance-dependent count statistics) was always set to 1.0 Å. With the aforementioned parameter settings, the MaP descriptor is quickly computed and the dependence of computation time and molecule size is negligible for druglike compounds. Approximately 15 s on a 1 GHz personal computer were needed for the largest molecule in this study ( $\text{C}_{44}\text{H}_{58}\text{N}_6\text{O}_2$ ).

## Results and Discussion

**Steroid Data Set.** To be able to compare the MaP descriptor to other molecular descriptors, we first studied the “steroid data set” introduced by Cramer and co-workers.<sup>7</sup> This data set consists of 31 steroids for which the binding affinity to the corticosteroid-binding



**Table 1.** Series of 31 Steroids Binding to the Corticosteroid-Binding Globulin

no.	compd name	log <i>K</i> (CBG affinity)
1	aldosterone <sup>a</sup>	6.279
2	androstanediol	5.000
3	androstenediol	5.000
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	deoxycorticosterone	7.653
11	deoxycortisol	7.881
12	dihydrotestosterone	5.919
13	estradiol	5.000
14	estriol	5.000
15	estrone	5.000
16	etiocolanolone	5.255
17	pregnenolone	5.255
18	17-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	16 $\alpha$ ,17-dihydroxy-4-pregnene-3,20-dione	6.247
28	16 $\alpha$ -methyl-4-pregnene-3,20-dione	7.120
29	19-norprogesterone	6.817
30	11 $\beta$ ,17,21-trihydroxy-2 $\alpha$ -methyl-4-pregnene-3,20-dione	7.688
31	11 $\beta$ ,17,21-trihydroxy-2 $\alpha$ -methyl-9 $\alpha$ -fluoro-4-pregnene-3,20-dione <sup>a</sup>	5.797

<sup>a</sup> Outlier, removed from analysis.

globulin (CBG) was measured. Various groups used this data set to compare the quality of their 3D-QSAR methodologies. Hence, this data set has become one of the most often discussed ones and can be seen as a benchmark data set for novel molecular descriptors.<sup>18</sup> Even though this data set is not the ideal 3D benchmark data set,<sup>18</sup> it was used for the sake of comparability.

The names of the structures and the corresponding biological activities are listed in Table 1. The 3D structures were provided by Gasteiger.<sup>14</sup> No further geometry optimization was carried out. The MaP procedure was applied to the structures, with the default parameters. The largest distance between two surface point pairs was 17.5 Å ( $\Rightarrow c = 17$ ). Hence, 170 variables were obtained for each structure and a matrix of 31  $\times$  170 was obtained for the entire data set. A total of 168 out of the 170 variables showed nonconstant variance

and were retained. Preliminary analysis (variable selection on the entire set of 31 structures) revealed that compound **1** and compound **31** are outliers that were excluded from further analysis. Compound **31** was identified as an outlier in many previous studies.<sup>18</sup> Since compound **31** is the only compound in the data set that does have a fluorine atom attached to it, this finding becomes understandable. On the other hand, compound **1** (aldosterone) is the only molecule with a cyclic hemiacetal. That means that both compounds are outliers in structural space.

PCR with leave-one-out cross-validation (LOO-CV) for selecting the number of latent variables (LV) was employed to build a model for the remaining 29 compounds. The model obtained shows an optimal dimensionality of 5 and a leave-one-out cross-validated squared multiple correlation coefficient  $R^2_{CV-1}$  of 0.53. The dimensionality of the best model obtained with the LMO-CV for selecting the number of principal components is also 5 with a  $R^2_{CV-50\%}$  value of 0.38 (see Table 2). This model was obtained with a leave-50%-out cross-validation where the data set was randomly split 87 times (3*m*) into construction and validation data. Results for PLS are slightly better than those obtained by PCR. Since both models are not satisfactory and for the purposes of interpretation, REM-TS in combination with PCR was run to select the most informative variables (MIV). The objective function for simultaneously selecting variables and the optimal number of latent variables was LMO-CV with 50% of the data left out. REM-TS ended up with a four-parameter model and an  $R^2_{CV-50\%}$  value of 0.84. The following variables were selected: AL<sub>12</sub>, DH<sub>14</sub>, DH<sub>1</sub>, and AD<sub>6</sub>. They are coded as property–property–distance triplets. For instance, DH<sub>14</sub> counts all surface point pairs from donor regions to hydrophilic regions separated by a distance of 13.5–14.5 Å. That means that the subscript indicates the center of the distance interval. The selected variables resulted in:

$$\hat{y} = 6.4102 + 0.0032 \text{ AL}_{12} - 0.0106 \text{ DH}_{14} + 0.0420 \text{ DH}_1 - 0.0039 \text{ AD}_6$$

$$R^2_{CV-50\%} = 0.84; \text{ RMSEP}_{CV-50\%} = 0.43; R^2 = 0.91; \text{ RMSEC} = 0.36; m = 29; \text{ LV} = 4$$

It can be seen that variables 1 (AL<sub>12</sub>) and 3 (DH<sub>1</sub>) are both positively correlated to the dependent variables, which means that high scores for these variables result in high predicted biological activity. Closer inspection of the selected variables reveals that variables AL<sub>12</sub> and DH<sub>14</sub> as well as variables DH<sub>1</sub> and AD<sub>6</sub> need to be interpreted as an ensemble because they describe

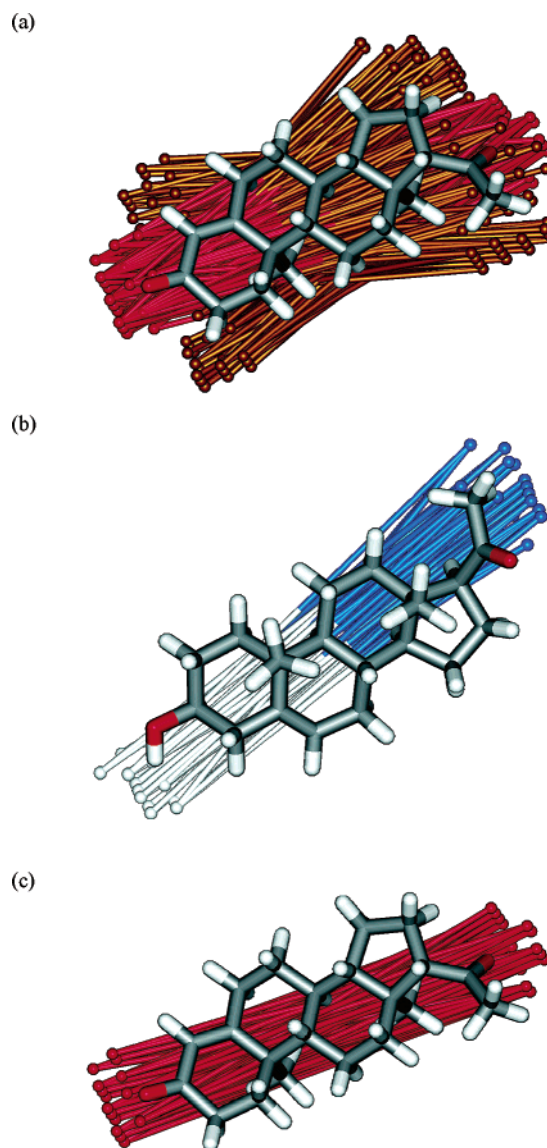
**Table 2.** Results for the Steroid Data Set

RT <sup>a</sup>	RMSEP <sub>CV-1</sub> <sup>b</sup>	$R^2_{CV-1}$ <sup>c</sup>	RMSEP <sub>CV-50%</sub> <sup>d</sup>	$R^2_{CV-50\%}$ <sup>d</sup>	RMSEC <sup>e</sup>	$R^2$ <sup>f</sup>	RMSEP <sub>Test</sub> <sup>g</sup>	$R^2_{\text{Test}}$ <sup>g</sup>	<i>m/m</i> <sub>Test</sub> <sup>h</sup>	<i>n</i> <sup>i</sup>	LV <sup>j</sup>
PCR	0.75	0.53	0.86	0.38	0.68	0.70			29	168	5
PLS	0.66	0.63	0.85	0.40	0.57	0.78			29	168	4
PCR-VS <sup>k</sup>	0.38	0.88	0.43	0.84	0.36	0.91			29	4	4
PCR-VS <sup>k</sup>	0.39	0.89	0.51	0.81	0.38	0.92	0.44	0.81	20/9	4	4
GRIND-PLS <sup>l</sup>		0.76		0.75 <sup>m</sup>		0.83			29	69	2
GRIND-PLS <sup>l</sup>		0.64		0.64 <sup>m</sup>		0.82	0.26	0.93	20/9	63	2

<sup>a</sup> RT: regression technique. <sup>b</sup> RMSEP<sub>CV-1</sub>: leave-one-out cross-validated root-mean-squared error of prediction. <sup>c</sup>  $R^2_{CV-1}$ : leave-one-out cross-validated coefficient of determination. <sup>d</sup> Same as in footnotes b and c for leave-50%-out cross-validation. <sup>e</sup> RMSEC: root-mean-squared error of calibration. <sup>f</sup>  $R^2$ : coefficient of determination. <sup>g</sup> Same as in footnotes b and c for test set prediction. <sup>h</sup> *m*: number of objects. <sup>i</sup> *n*: number of variables. <sup>j</sup> LV: number of latent variables. <sup>k</sup> VS: variable selection with tabu search. <sup>l</sup> Data taken from ref 15. <sup>m</sup> Data were obtained by a repetitive 5-fold cross-validation, which is similar to a leave-20%-out cross-validation.

several features relevant for biological activity. Variables describing several pieces of information are preferentially selected by the variable selection procedure owing to the tough validation criteria (leave-50%-out). As a result, models are highly compact but sometimes harder to interpret. For interpretational purposes, the selected MIV were back-projected into the original molecular space (Figure 7). Variable one ( $AL_{12}$ ) encodes the existence of a side chain with a carbonyl group in position 17 in combination with a carbonyl function in position 3 (e.g., progesterone). It can be seen that there are two V-shaped clusters for this variable. One starts at the carbonyl function in position 3, and the other starts at the carbonyl in position 20. Hence, this variable encodes the number of acceptor groups and their relative orientation to hydrophobic surface patches. This combination of functional groups describes the main components for high biological activity and increases the scores for this variable. Additionally, a decrease of the score for this variable is found when the methyl group in position 10 (less hydrophobic surface area) is missing (19-norprogesterone). Variable  $DH_{14}$  is identified as a punishment term (negative sign in the equation) for compounds that possess the mentioned side chain in position 17 but miss the carbonyl function in position 3 (e.g., pregnenolone). Variables three ( $DH_1$ ) and four ( $AD_6$ ) are mainly “fine-tuning” variables of the main structural requirements described so far. They describe the presence of additional hydroxyl groups ( $DH_1$ ) and their geometrical arrangement relative to the side chain at position 17 ( $AD_6$ ). If a compound is missing either of them, its biological activity will be reduced compared to the respective compound bearing that group (e.g., progesterone  $\leftrightarrow$  17 $\alpha$ -hydroxyprogesterone).

In Figure 7 projections of important variables on the highly active progesterone and weakly active pregnenolone are shown. Each line connecting two points represents an increment for that variable during computation of the descriptor, i.e., the more lines, the higher the value for the corresponding compound. It can easily be seen that the H-bond acceptor properties in positions 3, 20, and 21 and the H-bond donor properties in positions 11 and 17 play an important role in binding affinity. Combining variables  $AL_{12}$  and  $AD_6$  (shown separately in parts a and b of Figure 7) defines a triangle of pharmacophoric elements for glucocorticoid activity. However, instead of variable  $AL_{12}$ , we had expected an AA variable because the H-bond acceptor property of the carbonyl in position 3 is more important than its surrounding hydrophobicity. Inspecting the correlation of variable  $AL_{12}$  to other variables of the long-distance AA type (12–16 Å) revealed that it is highly correlated to variable  $AA_{14}$  ( $r = 0.93$ ). However, exchanging the two variables increases the objective function of the variable selection procedure slightly. Hence, variable selection will not exchange the variables. Doing so manually resulted in a more straightforward back-projection and easier interpretability. Since variable selection is guided by statistical criteria and not by criteria related to human perception, this selection is reasonable in terms of the chosen objective function but unfortunate in terms of interpretability. A model built with the two variables exchanged yielded



**Figure 7.** Progesterone (a, c) and pregnenolone (b) displayed with varying back-projections of property–property–distance triplets. Each line connecting two points is equivalent to an increment of at least 0.8 (fuzzy count) in vector  $\mathbf{v}$  for the corresponding variable. (a)  $AL_{12}$ . Red lines: ending on acceptor areas (A). Brown lines: ending on hydrophobic areas. The most important variable in the model is  $AL_{12}$ . High values for this variable are accompanied with a carbonyl function in position 3 and a side chain with a carbonyl function in position 17. This is the minimum requirement for high activity. (b)  $DH_{14}$ . Blue lines: ending on hydrophilic areas (H). White lines: ending on donor areas (D). Shown is the 3-hydroxy derivative of the highly active compound in (a). In addition to the lower count for variable  $AL_{12}$  (3-carbonyl  $\leftrightarrow$  3-hydroxyl), variable  $DH_{14}$  is increased (a carbonyl is lacking the H-bond donor), which results in a much lower predicted activity for this compound (negative sign of the regression coefficient). (c)  $AA_{14}$ . Red lines: ending on acceptor areas (A). The variable  $AA_{14}$  is highly correlated to  $AL_{12}$  and encodes the same information. However, pharmacophore identification is more straightforward and easier to comprehend using this variable. Model quality does not change significantly.

no significant difference in test set prediction ( $R^2_{\text{Test}} = 0.78$  vs  $R^2_{\text{Test}} = 0.80$ ) but resulted in more straightforward back-projection (Figure 7c). This highlights that visual inspection of the models helps to identify pharmacophoric patterns contained in the data and, on the

other hand, helps to reveal numerical artifacts of the modeling procedure. Since no test set was used for independent validation of the model, the aforementioned permutation test was run to further validate the equation given above. The results are summarized as quantiles of the distribution of  $R^2_{CV-50\%}$  values resulting from 500 variable selections on the scrambled data (scrambling from scratch), which are referred to as  $R^2_{CV-50\%,PT}$ . The median  $R^2_{CV-50\%,PT}$  was 0.12, the 95% quantile was 0.39, and the maximum value was 0.68. Since all  $R^2_{CV-50\%,PT}$  values are smaller than the real  $R^2_{CV-50\%}$ , it can be concluded that the probability of chance correlation is quite low ( $p < 0.002$ ).

Apart from using all 31 (29) molecules, the usual split into training and test data was used. The training set included compounds **2–21**, and the test set compounds are compounds **22–30**. Briefly, comparing the models for all steroids ( $m = 29$ ) with the subset of only 20 steroids shows that except for AD<sub>6</sub> the variables selected are equal for both models. This interchange of one variable (AD<sub>6</sub> ↔ HH<sub>7</sub>) is due to the altered composition of the dataset that is used for variable selection. Nevertheless, the chemical meaning of the selected variables is the same for both models, which highlights the importance of interpretability of MaP. A review of various 3D-QSAR methods applied to this data set by Coats<sup>18</sup> showed that the average  $R^2_{Test}$  obtained so far is about 0.76. Test set prediction of the obtained model ( $R^2_{Test} = 0.81$ ; see Table 2) is hence deemed satisfactory. Compared to GRIND,<sup>15</sup> MaP performs better on the internal figures of merit, whereas GRIND performs better on test set prediction ( $R^2_{Test} = 0.93$ ). It should be noted, however, that the figures of merit (internal vs external) for MaP are more balanced than those of GRIND.

**Prediction of Eye Irritation of Organic Chemicals.** Recently, Hopfinger and co-workers<sup>19</sup> presented a novel technique to predict the extent of eye irritation triggered by various chemical substances. The eye-irritation potential of the compounds was evaluated using the Draize in vivo rabbit eye irritation test.<sup>68</sup> The response data collected in this test are a combination of weighted scores for eye irritation of the cornea, conjunctiva, and iris of albino rabbit eyes graded after distinct time periods. Since this method involves live animal testing, the demand for alternative in vitro or computational models to reduce the number of animal tests has rapidly increased over the past decades (see ref 19 and references therein). The data set used for this application was established by the European Center for Ecotoxicology and Toxicology of Chemicals (ECETOC) as a "standard" data set for chemicals whose Draize rabbit eye irritation potential was measured according to OECD Guideline 405.<sup>69</sup> The potencies measured with the Draize test were adjusted as proposed by Hopfinger.<sup>19</sup> There, the molar adjusted eye scores (MES) used as dependent variables were calculated using

$$MES = \frac{MAS}{\text{molarity}}$$

with

$$\text{molarity} = \frac{(\rho)(1000)}{M_r}$$

**Table 3.** Series of 38 Compounds of the ECETOC Data Set Included in the Analysis<sup>a</sup>

no.	compd name	MES
hydrocarbons		
<b>1*</b>	3-methylhexane	0.10
<b>2*</b>	2-methylpentane	0.26
<b>3</b>	methylcyclopentane	0.41
<b>4</b>	1,9-decadiene	0.37
<b>5</b>	dodecane	0.45
<b>6*</b>	1,5-hexadiene	0.55
<b>7</b>	cis-cyclooctene	0.43
<b>8</b>	1,5-dimethylcyclooctadiene	0.44
aromatics		
<b>9</b>	4-bromophentole	0.19
<b>10</b>	2,4-difluoronitrobenzene	0.40
<b>11</b>	3-ethyltoluene	0.32
<b>12</b>	4-fluoroaniline	6.62
<b>13*</b>	xylene	1.10
<b>14*</b>	toluene	0.96
<b>15*</b>	styrene	0.77
<b>16*</b>	1-methylpropylbenzene	0.31
<b>17</b>	1,3-disopropylbenzene	0.38
ketones		
<b>18</b>	methyl amyl ketone	2.26
<b>19</b>	methyl isobutyl ketone	0.59
<b>20</b>	methyl ethyl ketone	4.48
<b>21</b>	acetone	4.83
alcohols		
<b>22*</b>	n-butanol	5.47
<b>23</b>	isobutanol	6.44
<b>24*</b>	2-propanol	2.34
<b>25*</b>	propylene glycol	0.10
<b>26</b>	2-ethyl-1-hexanol	7.82
<b>27</b>	glycerol	0.12
<b>28</b>	hexanol	8.13
<b>29</b>	butyl cellsolve	8.99
<b>30*</b>	cyclohexanol	8.29
acetates		
<b>31*</b>	ethyl acetate	1.47
<b>32*</b>	methyl acetate	3.14
<b>33</b>	methyltrimethyl acetate	0.36
<b>34</b>	ethyltrimethyl acetate	0.63
<b>35</b>	cellosolve acetate	2.03
<b>36</b>	n-butyl acetate	0.99
<b>37</b>	ethyl 2-methylacetoacetate	2.55
acids		
<b>38</b>	2,2-dimethylbutanoic acid	5.59

<sup>a</sup> Members of the test set are marked with an asterisk (\*).

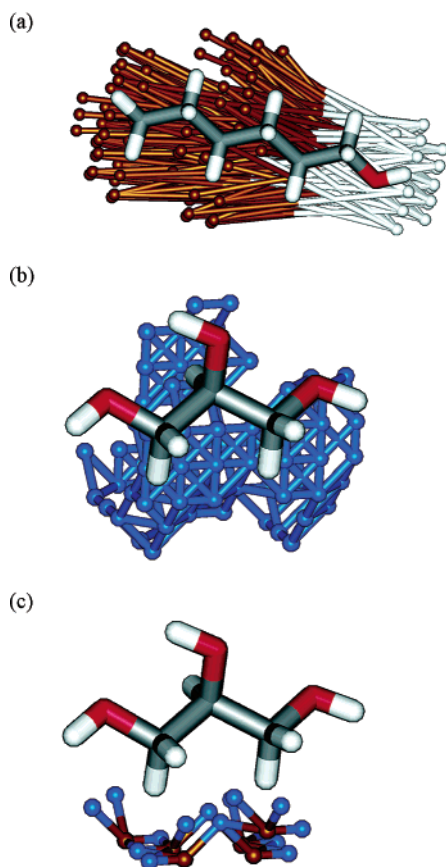
where MAS is the maximum average score of the Draize eye test,  $\rho$  is the density, and  $M_r$  is the relative molecular mass of the test chemical. This adjustment was made because activities used in QSAR studies are normally expressed as molar concentrations producing a fixed response. The compounds and the corresponding MES included in the data set are given in Table 3. The set of compounds used in the test series is structurally highly diverse. Owing to this high diversity, alignment of the included compounds is difficult if not impossible. Hence, commonly applied 3D-QSAR techniques that require an alignment step tend to fail. Since MaP does not need this alignment step, it can easily be applied to this set of 38 compounds.

The MaP procedure was applied to the data set with the default parameters ( $d_{\max} = 19.5 \text{ \AA} \Rightarrow c = 19$ ), and 117 variables (out of 190) with nonconstant variance were retained for analysis. PCR with LOO-CV for selecting the number of LVs led to a reasonable model for the whole data set ( $R^2_{CV-1} = 0.51$ ; see Table 4). However, model performance decreases considerably when using the more stringent leave-50%-out CV

**Table 4.** Results for the Eye-Irritation Data Set<sup>a</sup>

RT	RMSEP <sub>CV-1</sub>	R <sup>2</sup> <sub>CV-1</sub>	RMSEP <sub>CV-50%</sub>	R <sup>2</sup> <sub>CV-50%</sub>	RMSEC	R <sup>2</sup>	RMSEP <sub>Test</sub>	R <sup>2</sup> <sub>Test</sub>	m/m <sub>Test</sub>	n	LV
PCR	1.93	0.51	2.27	0.325	1.76	0.67			38	117	6
PLS	1.77	0.59	2.21	0.36	1.45	0.76			38	117	4
PCR-VS	1.34	0.76	1.51	0.70	1.24	0.82			38	4	2
PCR-VS	1.25	0.82	1.47	0.75	1.11	0.87	1.40	0.67	25/13	4	2
MI-QSAR		0.65				0.71			36	3	3
MI-QSAR <sup>b</sup>		0.73				0.78			38	5	5

<sup>a</sup> For definition of symbols and abbreviations, see Table 2. <sup>b</sup> This model also includes parabolic terms.



**Figure 8.** Back-projection of two MIVs for the eye irritation data set. Each line connecting two points is equivalent to an increment of at least 0.8 (fuzzy count) in vector  $\mathbf{v}$  for the corresponding variable. (a) DL<sub>7</sub> and DL<sub>10</sub> (superimposed). Brown lines: ending on hydrophobic areas (L). White lines: ending on donor areas (D). Hexanol is displayed with variables DL<sub>7</sub> and DL<sub>10</sub>. These variables are large for detergent-like compounds (e.g., aliphatic alcohols). Since DL<sub>7</sub> and DL<sub>10</sub> are positively correlated to the molar eye score, large values of these variables represent a high eye-irritating potential. (b) HH<sub>1</sub> and HL<sub>1</sub>. Brown lines: ending on hydrophobic areas (L). Blue lines: ending on hydrophilic areas (H). Glycerol shown with one variable that encodes the size of hydrophilic surface patches (HH<sub>1</sub>) and one that describes the distribution of hydrophilic surface patches relative to hydrophobic surface patches (HL<sub>1</sub>). If this ratio becomes unbalanced (e.g., too hydrophilic compounds), as is the case here, the eye-irritating potential decreases.

( $R^2_{CV-50\%} = 0.33$ ). Results for PLS are slightly better. Since the model using all available variables is not fully satisfactory and since not all variables encode information about the eye-irritating potential of the compounds, the reverse-elimination-method tabu search (REM-TS) was run for variable selection. The objective function for simultaneously selecting variables and the optimal number of LVs was again the default LMO-CV with 50% of the data left out. The final equation extracted from

**Table 5.** Chemical Structures and Biological Activity of Muscarinic M<sub>2</sub> Receptor Modulators (Group 1)<sup>a</sup>

no.	R1	R2	R3	R4	R5	R6	R7	X	n	pEC <sub>50</sub>
1*	Me	Me	Me	Me	Me	H	H	C	6	6.490
7*	Me	Me	Me	Me	Me	H	H	C	6	5.842
11*	Me	Me	Me	Me	H	H	H	C	3	5.420
12	Me	Me	Me	Me	H	H	H	C	4	5.570
13*	Me	Me	Me	Me	H	H	H	C	5	5.857
14*	Me	Me	Me	Me	H	H	H	C	7	6.409
15*	Me	Me	Me	Me	H	H	H	C	8	6.244
16	Me	Me	Me	Me	H	H	H	C	10	6.276
34	H	H	C <sub>3</sub> H <sub>6</sub> Ph	C <sub>3</sub> H <sub>6</sub> Ph	H	H	H	C	6	6.004
35	H	H	C <sub>2</sub> H <sub>4</sub> CN	C <sub>2</sub> H <sub>4</sub> CN	H	H	H	C	6	6.131
36	H	H	cyc-C <sub>6</sub> H <sub>11</sub>	cyc-C <sub>6</sub> H <sub>11</sub>	H	H	H	C	6	7.347
37	H	H	cyc-C <sub>6</sub> H <sub>11</sub>	cyc-C <sub>6</sub> H <sub>11</sub>	Me	H	Me	C	6	7.222
38	Me	Me	Me	Me	H	H	H	N	6	5.630
39	Me	Me	Me	Me	Cl	Cl	H	C	6	6.940
40*	Me	Me	Me	Me	F	F	H	C	6	6.340

<sup>a</sup> Members of the test set are marked with an asterisk (\*).

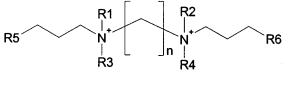
the data includes four MIVs and is given as

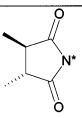
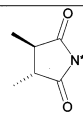
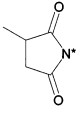
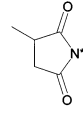
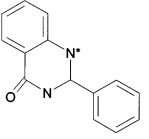
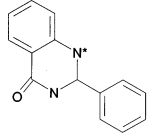
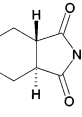
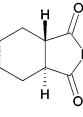
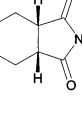
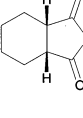
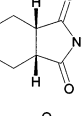
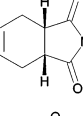
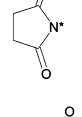
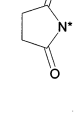
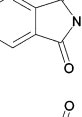
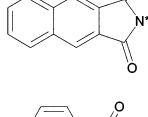
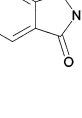
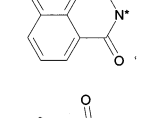
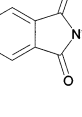
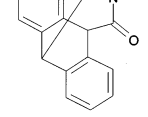
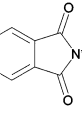
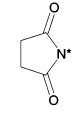
$$\hat{y} = 2.39 + 0.00738 \text{ DL}_7 + 0.00645 \text{ HH}_1 + 0.00680 \text{ DL}_{10} + 0.00373 \text{ HL}_1$$

$$R^2_{CV-50\%} = 0.70; \text{ RMSEP}_{CV-50\%} = 1.51; R^2 = 0.82; \text{ RMSEC} = 1.24, m = 38, \text{ LV} = 2$$

In contrast to Hopfinger, who deemed two compounds as outliers, we were able to use the full data set (see Table 4). Additionally, a scrambling procedure with 500 permutations was applied to the data given above. No apparent evidence for chance correlation was found. All  $R^2_{CV-50\%}$  values of the scrambling test were found to be far lower than those of the real model (median  $R^2_{CV-50\%,PT} = -0.01$ ; 95% quantile = 0.21;  $\max(R^2_{CV-50\%,PT}) = 0.33$ ). To further validate the model, the dataset was split into a representative training set (25 objects) and a test set (13 objects). Test set prediction was found to be good ( $R^2_{Test} = 0.67$ ), which is another indicator that the derived models are relevant.

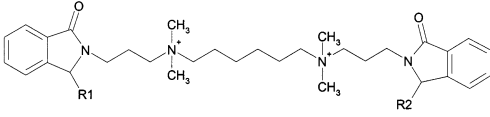
The selected MIVs were back-projected into the original molecular space and led to various conclusions about the structure–toxicity relationships of highly eye-irritating compounds. Figure 8 shows the back-projection of the four variables for selected compounds. Particularly, variables 1 (DL<sub>7</sub>) and 3 (DL<sub>10</sub>) are very specific and can easily be interpreted. These variables encode an H-bond donor moiety that is separated by a certain distance from lipophilic parts of the molecule. Compounds showing this characteristic give rise to large scores for DL<sub>7</sub> and DL<sub>10</sub> and are thus deemed potentially









**Table 6.** Chemical Structures and Biological Activity of Muscarinic M<sub>2</sub> Receptor Modulators (Group 2)<sup>a</sup>


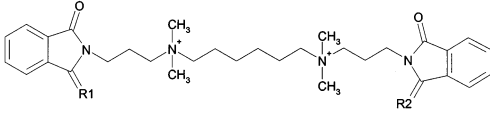
No.	R1	R2	R3	R4	n	R5	R6	pEC <sub>50</sub>
5	Me	Me	Me	Me	7			4.917
6*	Me	Me	Me	Me	7			4.495
9	Me	Me	Me	Me	6			6.125
19	Me	Me	Me	Me	7			6.097
20	Me	Me	Me	Me	7			5.347
21*	Me	Me	Me	Me	7			5.658
22	Me	Me	Me	Me	7			3.821
41*	Me	Me	Me	Me	6			6.830
42	Me	Me	Me	Me	6			7.200
43	Me	Me	Me	Me	6			6.400
44	Me	Me	Me	Me	6			5.010







<sup>a</sup> Members of the test set are marked with an asterisk (\*).

strong eye irritants. Variables 2 (HH<sub>1</sub>) and 4 (HL<sub>1</sub>) must be seen as an ensemble that needs combined interpretation. The latter set of variables describes the size of the hydrophilic surface area (HH<sub>1</sub>) responsible for solvation processes, as well as where the hydrophilic surface

**Table 7.** Chemical Structures and Biological Activity of Muscarinic M<sub>2</sub> Receptor Modulators (Group 3)<sup>a</sup>


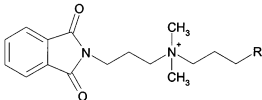
No.	R1	R2	pEC <sub>50</sub>
23	H	H	5.700
24	OH	OH	5.220
25*	OMe	OMe	5.170
26	OEt	OEt	5.640
27*			5.800
28			5.900
29			6.100
33			6.000

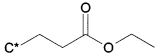
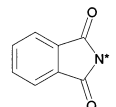
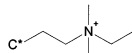
<sup>a</sup> Members of the test set are marked with an asterisk (\*).**Table 8.** Chemical Structures and Biological Activity of Muscarinic M<sub>2</sub> Receptor Modulators (Group 4)<sup>a</sup>


No.	R1	Substitution pattern for R1	R2	Substitution pattern for R2	pEC <sub>50</sub>
30		E		E	7.150
31		Z		Z	7.07
32*		E		Z	7.18

<sup>a</sup> Members of the test set are marked with an asterisk (\*).

patches are embedded in the molecular surface (HL<sub>1</sub>: encodes neighboring hydrophilic and lipophilic areas). Larger hydrophilic surface areas are evoked by functional groups incorporating oxygen or nitrogen atoms. Consequently, compounds having that characteristic have high values for HH<sub>1</sub>. Depending on the type of functional group, the score for variable HH<sub>1</sub> differs. High scores for HH<sub>1</sub> are found if the heteroatom is part of a carbonyl or an ether group because then the hydrophilic surface area is much larger as for hydroxyl or amine groups. Variable HL<sub>1</sub> can be described as a complementary variable to HH<sub>1</sub>. HL<sub>1</sub> mainly modulates the behavior of compounds having vicinal donor groups (i.e., hydroxyl or amine). In contrast to compounds with non-neighboring hydrophilic groups, compounds with vicinal donor groups show a large contiguous hydrophilic surface area (large score for HH<sub>1</sub>) with only a

**Table 9.** Chemical Structures and Biological Activity of Muscarinic M<sub>2</sub> Receptor Modulators (Group 5)<sup>a</sup>


No.	R1	pEC <sub>50</sub>
2*	C <sub>3</sub> H <sub>7</sub> OH	4.620
3*	C <sub>3</sub> H <sub>7</sub>	4.268
4		3.955
17		5.055
18		5.000

<sup>a</sup> Members of the test set are marked with an asterisk (\*).

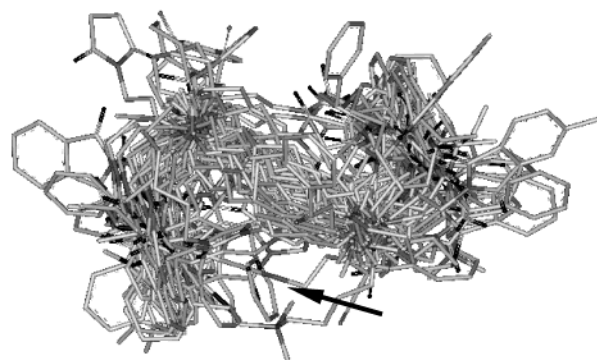
**Table 10.** Chemical Structure and Biological Activity of Additional Muscarinic M<sub>2</sub> Receptor Modulators Used in This Study

no.	structure	pEC <sub>50</sub>
8	methoctramine	5.523
10	verapamil	4.509

small contact area to hydrophobic surface patches (small score for HL<sub>1</sub>). An example is glycerol, which has a very low eye-irritating potential (HL<sub>1</sub> score very low).

In summary, hydrophilic compounds (represented through their H-bond donor capacity) with hydrophobic areas at a distance of about 7 and 10 Å will be predicted to show a high molar eye score. The characteristics described are usually very well fulfilled by detergent-like compounds such as aliphatic alcohols, which are known to disturb membrane structures.<sup>70</sup> If the hydrophilicity of a compound becomes too high compared to its overall size (ratio of HH<sub>1</sub> to HL<sub>1</sub>), the molar eye score decreases as well.

**Modulators of the Muscarinic M<sub>2</sub> Receptor.** The third data set is one not previously studied for QSAR. It consists of a set of allosteric modulators of the muscarinic M<sub>2</sub> receptor that are currently under study by Holzgrabe and co-workers. The structures were extracted from the literature<sup>20–26</sup> and are displayed in Tables 5–10. The pEC<sub>50</sub> values given in the tables describe the potency of a compound to retard the dissociation of the radioligand [*3H*]-*N*-methylscopolamine from the muscarinic M<sub>2</sub> receptor. The potencies were obtained from in vitro assay of a suspension of myocardial membranes prepared from porcine hearts. Since the pEC<sub>50</sub> values differ depending on the medium

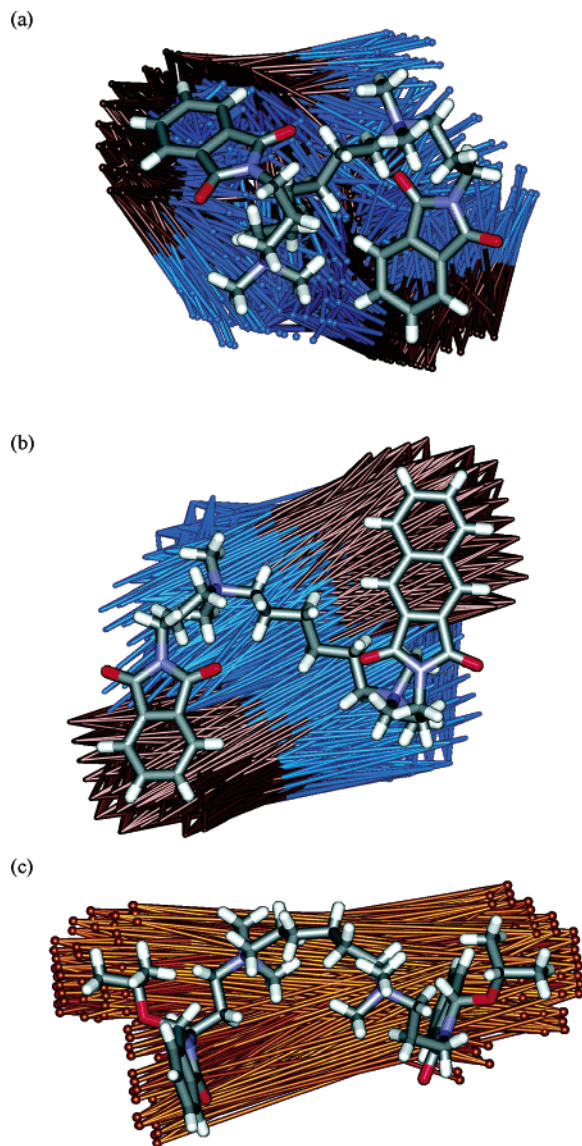
**Figure 9.** Compounds of the M<sub>2</sub> muscarinic modulator data set. The overlay shows that many compounds do have the proposed S shape, but some fulfill the hypothetical binding conformation with a different conformation (see arrow).

they are measured in, only data from Mg<sub>2</sub>HPO<sub>4</sub> buffer test systems were used. Owing to their high flexibility, the conformational freedom of the structures needed to be constrained. This was achieved by employing a previously published pharmacophore model that was derived from surface properties and shape information of the compounds using Kohonen neural networks.<sup>71</sup> This model consists of two positively charged atoms and two aromatic groups in a well-defined geometric configuration. Alcuronium, a rigid allosteric M<sub>2</sub> receptor modulator of high potency, was used as a template molecule for the localization of the pharmacophore model. The atoms defining the charged groups were the two quaternary nitrogen atoms, whereas the hydrophobic points were fixed at the centers of mass of atoms 4, 5, 6, and 7 of the indolindione ring. All tolerances for mapping the pharmacophoric points to alcuronium were set to 1.0 Å. Mapping of the compounds was performed in two steps. First, structures were fully protonated at basic nitrogens using an automated script coded with the DAYLIGHT Programmer's Toolkit.<sup>72</sup> Second, the UNITY<sup>73</sup> module available in SYBYL<sup>74</sup> was employed for conformationally flexible searching of the compound database for the best fit of the ligands to the specified pharmacophore model. All settings were set to default except for the search type which was set to "flexible", the timeout which was set to 120 s, and the Lipinski parameter which was disabled. That way, similar conformations for the compounds were obtained, yet the data are not aligned in the usual sense. It should be recalled that MaP is translationally and rotationally invariant but depends on the conformations of the structures under study. The molecules in group 5 (Table 9) are only one-half of the lead structure (W84, compound 7). Because these compounds cannot meet the requirements of the pharmacophore model and consequently cannot meet the requirements of the search query, they were used in the same conformation as the most similar "complete" (ditopic) structure. These

**Table 11.** Results for the M<sub>2</sub> Modulator Data Set<sup>a</sup>

RT	RMSEP <sub>CV-1</sub>	R <sup>2</sup> <sub>CV-1</sub>	RMSEP <sub>CV-50%</sub>	R <sup>2</sup> <sub>CV-50%</sub>	RMSEC	R <sup>2</sup>	RMSEP <sub>Test</sub>	R <sup>2</sup> <sub>Test</sub>	m/m <sub>Test</sub>	n	LV
PCR	0.60	0.55	0.63	0.49	0.57	0.63			44	333	3
PLS	0.57	0.57	0.64	0.49	0.51	0.69			44	333	2
PCR-VS	0.39	0.81	0.42	0.78	0.36	0.86			44	4	4
PCR-VS	0.41	0.80	0.45	0.76	0.38	0.86	0.47	0.68	29/15	5	4

<sup>a</sup> For definition of symbols and abbreviations, see Table 2.



**Figure 10.** Back-projection of three MIVs of the muscarinic  $M_2$  modulator dataset. Each line connecting two points is equivalent to an increment of at least 0.8 (fuzzy count) in vector  $\mathbf{v}$  for the corresponding variable. (a)  $HLs_6$ . Dark-brown lines: ending on strongly hydrophobic areas (Ls). Blue lines: ending on hydrophilic areas (H). The projection of  $HLs_6$  shows that this property–property–distance triplet is obtained by starting on an aromatic ring and ending on both positively charged atoms. Hence, this variable reflects the requirements of the proposed pharmacophore, and it encodes information about the overall conformation of the molecules (S shape). Since compounds in group 5 consist of only one aromatic ring and one positively charged atom, the values for  $HLs_6$  are smaller. (b)  $HLs_{14}$ . Dark-brown lines: ending on strongly hydrophobic areas (Ls). Blue lines: ending on hydrophilic areas (H). This variable yields high values if the aromatic ring system is large and flat (right-hand side). If the lateral ring system is smaller (left-hand side) or unsaturated, the number of counts is drastically reduced, which results in a much lower inhibitory activity. (c)  $LwLw_{18}$ . Brown lines: ending on weakly hydrophobic areas (Lw). If one of the carbonyl groups is substituted by large hydrophobic groups, this variable yields high values. Since the sign of the regression coefficient is negative, this results in decreased predicted activity.

conformations were obtained using SYBYL's multifit procedure, which actually corresponds to an alignment

step. The entire compound database consisting of 44 molecules is shown in Figure 9 after the pharmacophore mapping was carried out.

The MaP descriptor was computed with default parameters with the exception of one parameter. An additional hydrophobic category was introduced to differentiate between strongly hydrophobic (Ls) and weakly hydrophobic (Lw) parts of the molecules ( $\Rightarrow p = 5$ ). The respective cutoff value for the hydrophobic potential was set to 0.12. That means that a hydrophobic potential of  $<0$  is classified as hydrophilic, a hydrophobic potential of  $0-0.12$  is classified as weakly hydrophobic (Ls), and values greater than 0.12 are classified as strongly hydrophobic (Ls). The maximum distance ( $d_{max}$ ) for this data set is  $29.5 \text{ \AA}$  ( $c = 29$ ). This results in a  $44 \times 435$  matrix of which 103 constant columns were excluded. PCR with three LVs yielded a  $R^2_{CV-1}$  of 0.55 and a  $R^2_{CV-50\%}$  of 0.49 (see Table 11). Again, PLS gave similar results ( $R^2_{CV-1} = 0.57$ ;  $R^2_{CV-50\%} = 0.49$ ) with two latent variables. Probing the data with variable selection yielded better models. Leave-50%-out cross-validation with 132 ( $3m$ ) random splits in construction and validation data sets was used to simultaneously select variables and the optimum number of latent variables. The best equation found during the search is

$$\hat{y} = 5.8064 + 0.000202 ALs_9 + 0.000540 HLs_6 - \\ 0.000374 LsLs_{10} + 0.000280 HLs_{14} - \\ 0.000254 LwLw_{18}$$

$$R^2_{CV-50\%} = 0.80; \quad RMSEP_{CV-50\%} = 0.43; \quad R^2 = 0.84; \\ RMSEC = 0.37; \quad m = 44; \quad LV = 5$$

Once more, the probability of chance correlation is low ( $p < 0.002$ ) because no variable selection run on the scrambled responses yielded a larger  $R^2_{CV-50\%,PT}$  value than the real  $R^2_{CV-50\%}$  (median  $R^2_{CV-50\%,PT} = 0.10$ ; 95% quantile = 0.33;  $\max(R^2_{CV-50\%,PT}) = 0.57$ ). Moreover, the dataset was split into a representative training set (29 objects) and a test set (15 objects). Test set prediction performed well ( $R^2_{Test} = 0.68$ ) and supports the finding that the derived models are relevant.

Interpretation of the selected variables reflects the pharmacophore model used for the conformational restriction of the structures. Again, the selected variables need to be evaluated as an ensemble. Variables 1 ( $ALs_9$ ), 2 ( $HLs_6$ ), and 4 ( $HLs_{14}$ ) describe the geometric arrangement of the aromatic side chain with respect to the positively charged nitrogen as follows. Variables 2 and 4 directly encode the distance of the hydrophilic nitrogen to the lateral aromatic groups, whereas  $ALs_9$  encodes information about the type of the aromatic group, favoring the existence of acceptor atoms in the direct neighborhood of the aromatic system as, for instance, in the phthalimido moiety. Additionally, variable 1 ( $ALs_9$ ) encodes information about the size and shape of the aromatic ring system (large and flat ring systems show high counts for this variable). Variable 3 ( $LsLs_{10}$ ) can be described as a punishment (negative sign) term for compounds with large hydrophobic systems lying outside the common backbone (e.g., compound **34** or compound **43**). The "ditopic" nature of the compounds is identified implicitly because compounds

that are reduced in size will have a lower or no value for the mentioned variables. In Figure 10 the back-projection of different MIV are shown for compounds **7** (Figure 10a), **41** (Figure 10b), and **27** (Figure 10c), of which compound **7** is the lead compound of the series.

Since this compound class is under active research, new compounds are currently added and will be incorporated into a more detailed model. Preliminary results showed that test set predictions for new compounds are quite reliable with the presented model. These findings will be part of an upcoming publication.

## Conclusion

A novel translationally and rotationally invariant molecular descriptor was presented. The theoretical basis for the MaP descriptor are the so-called radial distribution functions (distance-dependent count statistics), which have the advantage of encoding molecular shape as well as surface property distributions in a single vector. The single MaP variables encode the size (absolute number of counts) and the orientation (distance) of surface patches with selected properties. The relative merits of the MaP descriptor are easy implementation, translational and rotational invariance, fast computation, good model quality, and simple model interpretability by back-projection of the descriptor into the original molecular space. Hence, two major requirements for an efficient 3D-QSAR methodology are met, namely, translational and rotational invariance and good interpretability. Therefore, the MaP descriptor not only represents yet another molecular descriptor but allows the computational as well as the medicinal chemist to intuitively build and understand structure-activity relationships of the compounds under study.

In terms of fit and predictive power, the models obtained using the new descriptor compared very well to other 3D-QSAR techniques. This statement holds for the steroid benchmark data set, and equally good results were obtained for the eye-irritation data set, which is difficult to model owing to its high diversity. Finally, a new model for allosteric modulators of the muscarinic M<sub>2</sub> receptor was established that consists of highly flexible structures. Most often, the MaP descriptor is used in conjunction with variable selection because typically not all variables are related to the property under study and because interpretation is based on a few highly predictive variables. Great care must be taken to thoroughly validate models coming from variable selection routines because they are generally overoptimistic and susceptible to chance correlations. To prevent both, a stringent leave-multiple-out cross-validation, instead of the usual leave-one-out CV, was used, the final models were additionally validated with a permutation test rerunning the entire selection procedure (scrambling from scratch), and the datasets were split into independent training and test sets for external test set prediction.

MaP is sensitive to different conformations of molecules and thus needs sensible input conformations. For a data set as the M<sub>2</sub> receptor modulators, sensible conformations were obtained with the help of a pharmacologically active, rigid template molecule. The crucial role of the fourth dimension of QSAR (conforma-

tional flexibility) was pointed out by Hopfinger.<sup>17</sup> Since the mathematical foundation of the MaP descriptor is extremely easy, an extension to the fourth dimension is conceivable and is currently under investigation in our research group.

**Acknowledgment.** We greatly acknowledge the insightful comments of and fruitful discussions with Iain McLay, GSK, Stevenage, U.K. We also thank Francis Atkinson, GSK, Stevenage, U.K., for contributing to the initial processing of the M<sub>2</sub> modulator data set. Financial support of the Fonds der Chemischen Industrie, Frankfurt/Main, Germany, is gratefully acknowledged. Moreover, two anonymous referees are acknowledged, whose comments helped to improve the presentation of the material.

**Supporting Information Available:** Figure showing the property mapping (H-bond donor/acceptor, hydrophobic/hydrophilic) onto the molecular surface. This material is available free of charge via the Internet at <http://pubs.acs.org>.

## References

- (1) Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; Wiley-VCH: Weinheim, Germany, 2000.
- (2) Todeschini, R.; Consonni, V.; Pavan, M. *Dragon. Software for the Calculation of Molecular Descriptors*, version 1.11; <http://www.disat.unimib.it/chm/Dragon.htm>.
- (3) Adamson, G. W.; Lynch, M. F.; Town, W. G. Analysis of Structural Characteristics of Chemical Compounds in a Large Computer-Based File. Part 2. Atom Centred Fragments. *Perkin Trans. 2* **1970**, 3702–3706.
- (4) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Chemistry and Drug Research*; Academic Press: New York, 1976.
- (5) Carhart, R. E.; Smith, D. H.; Venkataraghavan, R. Atom Pairs as Molecular Features in Structure-Activity Studies: Definition and Applications. *J. Chem. Inf. Comput. Sci.* **1985**, 25, 64–73.
- (6) Baumann, K. An Alignment-Independent Versatile Structure Descriptor for QSAR and QSPR Based on the Distribution of Molecular Features. *J. Chem. Inf. Comput. Sci.* **2002**, 42, 26–35.
- (7) Cramer, R. D.; Patterson, D. E.; Bunce, J. D. Comparative Molecular Field Analysis (CoMFA). 1. Effect of Shape on Binding of Steroids to Carrier Proteins. *J. Am. Chem. Soc.* **1988**, 110, 5959–5967.
- (8) Klebe, G.; Abraham, U.; Mietzner, T. Molecular Similarity Indices in a Comparative Analysis (CoMSIA) of Drug Molecules To Correlate and Predict Their Biological Activity. *J. Med. Chem.* **1994**, 37, 4130–4146.
- (9) Doweiko, A. M. The Hypothetical Active Site Lattice. An approach to Modeling Active Sites from Data on Inhibitor Molecules. *J. Med. Chem.* **1988**, 31, 1396–1406.
- (10) Silverman, B. D.; Platt, D. E. Comparative Molecular Moment Analysis (CoMMA): 3D QSAR without Molecular Superposition. *J. Med. Chem.* **1996**, 39, 2129–2140.
- (11) Ferguson, A. M.; Heritage, T.; Jonathon, P.; Pack, S. E.; Philips, L.; Rogan, J.; Snaith, P. J. EVA: A New Theoretically Based Molecular Descriptor for Use in QSAR/QSPR Analysis. *J. Comput.-Aided Mol. Des.* **1997**, 11, 143–152.
- (12) Todeschini, R.; Gramatica, P. 3D-Modelling and Prediction by WHIM Descriptors. Part 5. Theory Development and Chemical Meaning of WHIM Descriptors. *Quant. Struct.-Act. Relat.* **1997**, 16, 113–119.
- (13) Gancia, E.; Bravi, G.; Mascagni, P.; Zaliani, A. Global 3D-QSAR Methods: MS-WHIM and Autocorrelation. *J. Comput.-Aided Mol. Des.* **2000**, 14, 293–306.
- (14) Schur, J. H.; Selzer, P.; Gasteiger, J. The Coding of the Three-Dimensional Structure of Molecules by Molecular Transforms and Its Application to Structure-Spectra Correlations and Studies of Biological Activity. *J. Chem. Inf. Comput. Sci.* **1996**, 36, 334–344.
- (15) Pastor, M.; Cruciani, G.; McLay, I.; Pickett, S.; Clementi, S. Grid-Independent Descriptors (GRIND): A Novel Class of Alignment-Independent Three-Dimensional Molecular Descriptors. *J. Med. Chem.* **2000**, 43, 3233–3243.
- (16) Goodford, P. J. A Computational Procedure for Determining Energetically Favorable Binding Sites on Biologically Important Macromolecules. *J. Med. Chem.* **1985**, 28, 849–857.
- (17) Hopfinger, A. J.; Wang, S.; Tokarski, J. S.; Jin, B.; Albuquerque, M.; Madhav, P. J.; Duraiswami, C. Construction of 3D-QSAR



- Models Using the 4D-QSAR Analysis Formalism. *J. Am. Chem. Soc.* **1997**, *119*, 10509–10524.
- (18) Coats, E. A. The CoMFA Steroids as a Benchmark Dataset for Development of 3D QSAR Methods. *Perspect. Drug Discovery Des.* **1998**, *12–14*, 199–213.
- (19) Kulkarni, A.; Hopfinger, A. J.; Osborne, R.; Bruner, L. H.; Thompson, E. D. Prediction of Eye Irritation from Organic Chemicals Using Membrane-Interaction QSAR Analysis. *Toxicol. Sci.* **2001**, *59*, 335–345.
- (20) Tränkle, C.; Kostenis, E.; Burgmer, U.; Mohr, K. Search for Lead Structures To Develop New Allosteric Modulators of Muscarinic Receptors. *J. Pharmacol. Exp. Ther.* **1996**, *279*, 926–933.
- (21) Staudt, M.; Tränkle, C.; Mohr, K.; Holzgrabe, U. Contribution of Lateral Substituents in Heptane-Bisammonium Derivatives to the Allosteric Stabilization of Antagonist Binding to M<sub>2</sub>-Receptors. *Life Sci.* **1998**, *62*, 423–429.
- (22) Mohr, K.; Holzgrabe, U. Allosteric Modulators of Ligand Binding to Muscarinic Acetylcholine Receptors. *Drug Discovery Today* **1998**, *3*, 214–222.
- (23) Nassif-Makki, T.; Tränkle, C.; Zlotos, D.; Bejeuhr, G.; Cambareri, A.; Pfletschinger, C.; Kostenis, E.; Mohr, K.; Holzgrabe, U. Bisquarternary Ligands of the Common Allosteric Site of M<sub>2</sub> Acetylcholine Receptors: Search for the Minimum Essential Distances between the Pharmacophoric Elements. *J. Med. Chem.* **1999**, *42*, 849–858.
- (24) Bender, W.; Staudt, M.; Tränkle, C.; Mohr, K.; Holzgrabe, U. Probing the Size of a Hydrophobic Binding Pocket within the Allosteric Site of Muscarinic Acetylcholine M<sub>2</sub>-Receptors. *Life Sci.* **2000**, *66*, 1675–1682.
- (25) Botero Cid, H. M.; Tränkle, C.; Baumann, K.; Pick, R.; Mies-Klöffass, E.; Kostenis, E.; Mohr, K.; Holzgrabe, U. Structure-Activity Relationships in a Series of Bisquarternary Bisphthalimidine Derivatives Modulating the Muscarinic M<sub>2</sub>-Receptor Allosterically. *J. Med. Chem.* **2000**, *43*, 2155–2164.
- (26) Pick, R. Allosterische Modulatoren des muscarinischen Acetylcholinrezeptors—Synthese Bis-Tertiärer Analoga des Leitmoleküls W84 (Allosteric modulators of the muscarinic acetylcholine receptor—Synthesis of bis-tertiary analogs of the lead structure W84). Dissertation, Universität Würzburg, Germany, 2000.
- (27) Sadowski, J.; Gasteiger, J. From Atoms and Bonds to Three-Dimensional Atomic Coordinates: Automatic Model Builders. *Chem. Rev.* **1993**, *93*, 2567–2581.
- (28) *Alchemy 2000*, version 2.05; Tripos Inc. (1699 South Hanley Rd): St. Louis, MO 63144, 1998.
- (29) Connolly, M. L.; Analytical Molecular Surface Calculation. *J. Appl. Crystallogr.* **1983**, *16*, 548–558.
- (30) Christopher, J. A. Computer-assisted modelling and analysis of biomolecular structures. Dissertation, Texas A&M University, College Station, TX, 1998.
- (31) Sanner, M. F.; Spohner, J.-C.; Olson, A. J. Reduced surface: an efficient way to compute molecular surfaces. *Biopolymers* **1996**, *38*, 305–320.
- (32) Pascual-Ahuir, J. L.; Silla, E. GEPOL: An Improved Description of Molecular Surfaces. I. Building the Spherical Surface Set. *J. Comput. Chem.* **1990**, *11*, 1047–1060.
- (33) Rohrbough, R. H.; Jurs, P. C. Molecular Shape and the Prediction of High-Performance Liquid Chromatographic Retention Indexes of Polycyclic Aromatic Hydrocarbons. *Anal. Chem.* **1985**, *59*, 1048–1054.
- (34) Rocchia, W.; Sridharan, S.; Nicholls, A.; Alexov, E.; Chiabrera, A.; Honig, B. Rapid grid-based construction of the molecular surface and the use of induced surface charge to calculate reaction field energies: applications to the molecular systems and geometric objects. *J. Comput. Chem.* **2002**, *23*, 128–137.
- (35) Howard, J. A. K.; Hoy, V. J.; O'Hagan, D.; Smith, G. T. How Good Is Fluorine as a Hydrogen Bond Acceptor? *Tetrahedron* **1996**, *52*, 12613–12622.
- (36) Ghose, A. K.; Viswanadhan, V. N.; Wendoloski, J. J. Prediction of Hydrophobic (Lipophilic) Properties of Small Organic Molecules Using Fragmental Methods: An Analysis of ALOGP and CLOGP Methods. *J. Phys. Chem. A* **1998**, *102*, 3762–3772.
- (37) Heiden, W.; Moeckel, G.; Brickmann, J. A New Approach to Analysis and Display of Local Lipophilicity/Hydrophilicity Mapped on Molecular Surfaces. *J. Comput.-Aided Mol. Des.* **1993**, *7*, 503–514.
- (38) Sheridan, R. P.; Miller, M. D.; Underwood, D. J.; Kearsley, S. K. Chemical Similarity Using Geometric Atom Pair Descriptors. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 128–136.
- (39) Brown, R. D.; Martin, Y. C. Use of Structure-Activity Data To Compare Structure-Based Clustering Methods and Descriptors for Use in Compound Selection. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 572–584.
- (40) McGregor, M. J.; Muskal, S. M. Pharmacophore Fingerprinting. 1. Application to QSAR and Focused Library Design. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 569–574.
- (41) Davis, A. M.; Teague, S. J. Hydrogen Bonding, Hydrophobic Interactions, and Failure of the Rigid Receptor Hypothesis. *Angew. Chem., Int. Ed.* **1999**, *38*, 736–749.
- (42) Soltzberg, L. J.; Wilkins, C. L. Molecular transforms: a potential tool for structure-activity studies. *J. Am. Chem. Soc.* **1977**, *99*, 439–443.
- (43) Zupan, J.; Novic, M. General type of a uniform and reversible representation of chemical structures. *Anal. Chim. Acta* **1997**, *348* 409–418.
- (44) Baumann, K. Uniform-length molecular descriptors for quantitative structure-property (QSPR), quantitative structure-activity (QSAR), classification studies and similarity searching. *TrAC, Trends Anal. Chem.* **1999**, *18*, 36–46.
- (45) Mason, J. S.; Good, A. C.; Martin, E. J. 3-D Pharmacophores in Drug Discovery. *Curr. Pharm. Des.* **2001**, *7*, 567–597.
- (46) Moreau, G.; Broto, P. The Autocorrelation of a Topological Structure: A New Molecular Descriptor. *Nouv. J. Chim.* **1980**, *4*, 357–358.
- (47) Broto, P.; Moreau, G.; Vandyke, C. Molecular Structures: Perception, Autocorrelation Descriptor and SAR Studies—Perception of Molecules: Topological Structure and 3-Dimensional Structure. *Eur. J. Med. Chem.* **1984**, *19*, 61–65.
- (48) Martens, H.; Naes, T. *Multivariate Calibration*; John Wiley & Sons: Chichester, U.K., 1989.
- (49) Osten, D. W. Selection of optimal regression models via cross-validation. *J. Chemom.* **1988**, *2*, 39–48.
- (50) Glover, F. Tabu Search—Part I. *ORSA J. Comput.* **1989**, *1*, 190–206.
- (51) Glover, F. Tabu Search—Part II. *ORSA J. Comput.* **1990**, *2*, 4–32.
- (52) Voss, S. *Intelligent Search* (Habilitationsschrift); Technische Hochschule Darmstadt: Germany, 1993.
- (53) Baumann, K.; Albert, H.; von Korff, M. A Systematic Evaluation of the Benefits and Hazards of Variable Selection in Latent Variable Regression. Part I: Search Algorithm, Theory and Simulations. *J. Chemom.* **2002**, *16*, 339–350.
- (54) Shao, J. Linear Model Selection by Cross-Validation. *J. Am. Stat. Assoc.* **1993**, *88*, 486–494.
- (55) Zhang, P. Model Selection via Multifold Cross Validation. *Ann. Stat.* **1993**, *21*, 299–313.
- (56) Cruciani, G.; Baroni, M.; Clementi, S.; Costantino, G.; Riganelli, D.; Skagerberg, B. Predictive Ability of Regression Models. Part I: Standard Deviation of Prediction Errors (SDEP). *J. Chemom.* **1992**, *6*, 335–346.
- (57) Baroni, M.; Costantino, G.; Cruciani, G.; Riganelli, D.; Valigi, R.; Clementi, S. Generating Optimal Linear PLS Estimations (GOLPE): An Advanced Chemometric Tool for Handling 3D-QSAR Problems. *Quant. Struct.-Act. Relat.* **1993**, *12*, 9–20.
- (58) Cruciani, G.; Clementi, S. GOLPE: Philosophy and Applications in 3D QSAR. In *Advanced Computer Assisted Techniques in Drug Discovery*; van de Waterbeemd, H., Ed.; VCH Weinheim: Weinheim, Germany, 1994; pp 61–89.
- (59) Geisser, S. The Predictive Sample Reuse Method with Applications. *J. Am. Stat. Assoc.* **1975**, *70*, 320–328.
- (60) Baumann, K.; von Korff, M.; Albert, H. A Systematic Evaluation of the Benefits and Hazards of Variable Selection in Latent Variable Regression. Part II: Practical Applications. *J. Chemom.* **2002**, *16*, 351–360.
- (61) Burman, P. A Comparative Study of Ordinary Cross-Validation,  $\nu$ -fold Cross-Validation and the Repeated Learning-Testing Methods. *Biometrika* **1989**, *76*, 503–514.
- (62) Topliss, J. G.; Edwards, R. P. Chance Factors in Studies of Quantitative Structure-Activity Relationships. *J. Med. Chem.* **1979**, *22*, 1238–1244.
- (63) Klopman, G.; Kalos, A. N. Causality in Structure-Activity Studies. *J. Comput. Chem.* **1985**, *6*, 492–506.
- (64) Baumann, K. Distance profiles (DiP): A translationally and rotationally invariant 3D structure descriptor capturing steric properties of molecules. *Quant. Struct.-Act. Relat.* **2002**, *21*, 507–519.
- (65) Wu, W.; Walczak, B.; Massart, D. L.; Heuerding, S.; Erni, F.; Last, I. R.; Prebble, K. A. Artificial neural networks in classification of NIR spectral data: Design of the training set. *Chemom. Intell. Lab. Syst.* **1996**, *33*, 35–46.
- (66) Kennard, R. W.; Stone, L. A. Computer aided design of experiments. *Technometrics* **1969**, *11*, 137–148.
- (67) Roecker, E. B. Prediction-error and its estimation for subset-selected models. *Technometrics* **1991**, *33*, 459–468.
- (68) Draize, J. H.; Woodard, G.; Calvery, H. O. Methods for the Study of Irritation and Toxicity of Substances Applied to the skin and Mucous Membranes. *J. Pharmacol. Exp. Ther.* **1944**, *82*, 377–390.
- (69) *Acute Eye Irritation/Corrosion OECD Guideline for Testing of Chemicals*, No. 405; OECD: Paris, 1987.

- (70) McKarns, S. C.; Hansch, C.; Caldwell, W. S.; Morgan, W. T.; Moore, S. K.; Doolittle, D. J. Correlation between Hydrophobicity of Short-Chain Aliphatic Alcohols and Their Ability To Alter Plasma Membrane Integrity. *Fundam. Appl. Toxicol.* **1997**, *36*, 62–70.
- (71) Holzgrabe, U.; Wagener, M.; Gasteiger, J. Comparison of Structurally Different Allosteric Modulators of Muscarinic Receptors by Self-Organizing Neural Networks. *J. Mol. Graphics* **1996**, *14*, 185–193.
- (72) *DAYLIGHT Programmer's Toolkit*; Daylight Chemical Information Systems Inc. (#360 Mission Viejo): Los Altos, CA 92691, 2000.
- (73) *UNITY*, version 4.2.1; Tripos Inc. (1699 South Hanley Rd): St. Louis, MO 63144, 2000.
- (74) *SYBYL*, version 6.7.1; Tripos Inc. (1699 South Hanley Rd): St. Louis, MO 63144, 2000.

JM021077W