Journal of Computer-Aided Molecular Design **17:** 347–365, 2003.

Evaluation of extended parameter sets for the 3D-QSAR technique MaP: Implications for interpretability and model quality exemplified by antimalarially active naphthylisoquinoline alkaloids

Nikolaus Stiefl[1], Gerhard Bringmann[2], Christian Rummey[2] & Knut Baumann[1,*]
[1]*Department of Pharmacy and Food Chemistry and* [2]*Institute of Organic Chemistry, University of Würzburg, Am Hubland, D-97074 Würzburg, Germany*

MS received 20 January 2003; accepted in revised form 5 June 2003

Key words: antimalarials, MaP, naphthylisoquinoline alkaloids, parameter set validation, 3D-QSAR, alignment-independent

Summary

The 3D-QSAR technique MaP (Mapping Property distributions of molecular surfaces) characterises biologically active compounds in terms of the distribution of their surface properties (H-bond donor, H-bond acceptor, hydrophilic, weakly hydrophobic, strongly hydrophobic). The MaP descriptor is alignment-independent and yields chemically intuitive models. In this study, the impact of different operational parameters on the interpretability and model quality was investigated. Based on a set of antimalarially active naphtylisoquinoline alkaloids the effect of hydrophobicity assignment as well as the differentiation of H-bond propensity was evaluated according to a full factorial design. It turns out, that including different categories for H-bond donor strength significantly improved interpretability, reduced model complexity, and made possible the derivation of a novel pharmacophore hypothesis for this dataset. Further analysis of the factorial design reveals, that MaP models are robust to parameter changes and generate consistent models for different parameter settings.

Introduction

Chemical interpretability of a three dimensional quantitative structure activity relationship (3D-QSAR) is one of the major features desired for the design of novel drugs. Hence, apart from model quality, ease of interpretation and chemical meaning can be used to characterise the performance of novel molecular descriptors. Since the interpretation of graphical results is often favoured over mere numerical output, a variety of 3D-QSAR techniques with a graphical output have evolved. These techniques can be split into two major groups. The first group comprises alignment-dependent molecular descriptors such as CoMFA [1], CoMSIA [2], and GRID/PLS [3,4], whereas the other group includes translationally and rotationally invariant (TRI) molecular descriptors such as GRIND [5], or MaP [6]. Even though the aforementioned TRI descriptors (often referred to as "alignment-independent" descriptors) are still conformationally sensitive, one advantage is that a superposition of the chosen conformers is not necessary.

MaP is similar to GRIND in that it also aims at translational and rotational independence, and easy interpretability. However, MaP is different in the way molecular properties are encoded and how they are subsequently numerically represented. First, MaP does not start with a molecular interaction field (GRIND uses the GRID force field [7]), but with the molecular surface. Second, the variables used 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. This specific type of transformation

*Corresponding author. E-mail: knut.baumann@mail.uni-wuerzburg.de

allows the implicit description molecular shape which to this extent is not the case for the MACC-2 transform used by GRIND [8]. In essence, MaP and GRIND use different pieces of information of the molecules to represent them numerically for structure-activity correlations.

Even though inclusion of different surface properties into MaP is straightforward, only the most basic ones were included thus far (H-bond donor, H-bond acceptor, hydrophobicity, and hydrophilicity). Therefore, a systematic evaluation of an extension of MaP's parameter set was performed in this study. It incorporates different H-bond donor and H-bond acceptor strengths, and the variation of the fermi-type function parameters which control the mapping of the hydrophobic potential onto the surface (see below). These parameter variations were evaluated with respect to interpretability and model quality.

In a previous paper [9], a comparison of different alignment techniques prior to CoMSIA for a set of structurally related antimalarially active naphthylisoquinoline alkaloids [10,11,12] (see Figure 1 and Table 1) was performed. One of the main structural features of this type of compounds is the restricted conformational flexibility along the biaryl axis for most of its representatives. This rotationally hindered, but torsionally mostly flexible bond allows a facilitated decision about possible active conformations of the compounds. The alignment procedure for this dataset was very difficult. The difficulties with the alignment arose, because three reasonable alignment modes were found (see Figure 2) of which one was chosen for the study. However, it is not clear whether the chosen alignment represents the actual binding mode since no additional information is available thus far that supports a particular choice. Moreover, it is also not clear if all compounds show the same binding mode. In such cases it was recommended by Bravi and coworkers [13] to use alignment-independent descriptors. Hence, in order to evaluate an alignment-independent descriptor, MaP was applied to this dataset.

In terms of structural elements relevant for biological activity, two main features were identified by the CoMSIA model. First, an unsubstituted nitrogen atom in the isoquinoline ring system increases antimalarial activity (see Figure 1 and Table 1). Second, an oxygen functionality in position 6 of the isoquinoline moiety decreases antimalarial activity. These characteristics motivated the extentions of MaP's parameter set described below.

Information about the antimalarial mode of action of the studied compounds includes knowledge about sensitivity in different erythrocytic stages [14, 15], activity against erythrocytic [14] and exoerythrocytic [16] forms. A remarkable feature of this class of compounds is their antiplasmodial activity *in vivo*. For example, application of dioncophylline C (**3a**) to malaria infected mice resulted in a complete cure of the infection [17]. This *in vivo* activity makes the presented naphthylisoquinoline alkaloids promising lead compounds for novel drugs in the fight against malaria.

The paper is organised as follows. First, the basic algorithm for computing the MaP descriptor is briefly presented. Next, the operational parameters of MaP are outlined and the consequences of their modification are discussed. Subsequently, three parameters of the MaP procedure are varied according to a full factorial design and the resulting models are compared. Finally, the relative merits of the parameter extension are discussed and the results of the best model are summarised.

Methods

Theory of the MaP descriptor

The basic idea behind the MaP descriptor is that receptor-ligand interactions are mainly based on sterics, electrostatics, and hydrophobicity. Except for covalent binding modes, these interactions form between the surfaces of receptor and ligand. Consequently, the MaP descriptor characterises the molecular surface of a potentially biological active compound in terms of the distribution of its surface properties. The latter are mapped onto the surface based on the underlying atoms of the molecule. Currently, electrostatics are represented by H-bond donor and H-bond acceptor surface patches, whereas hydrophobicity mapping is based on localised atomic hydrophobicity values. Moreover, MaP encodes the steric features of the molecules implicitly through the distribution of distances between surface point pairs. The resulting categorised surface points are transformed into a distance dependent count statistics similar to pharmacophoric point pairs used in database screening [18]. These are subjected to a variable selection procedure. The selected variables are back-projected into the original molecular space for interpretation.

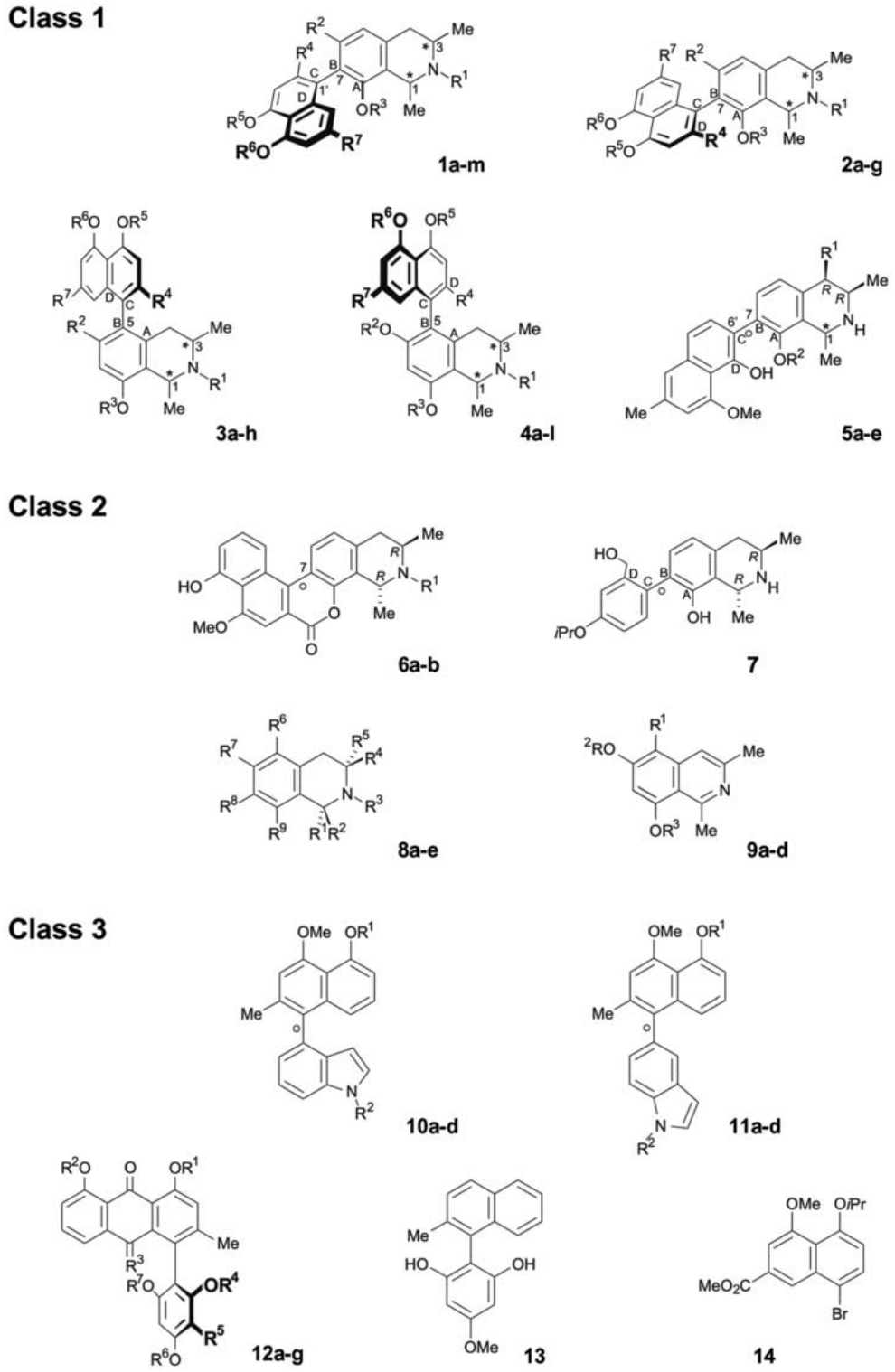

Figure 1. Compounds used in this study. The enumeration scheme encodes group membership (number) and consecutive compound identifiers (number + character). Atoms marked with an asterisk (*) are stereogenic centres. Axes that are freely rotating at room temperature are marked with an 'o'. Atoms that were used to define the torsional angle along the biaryl axis are marked as A, B, C, and D, respectively. Altogether, three classes were identified. The largest class is composed of 39 naphthylisoquinoline alkaloids (groups 1–5). The second class comprises synthetic precursors of the naphthylisoquinolines (group 6), the phenylisoquinoline **7**, and the much smaller isoquinolines (groups 8 and 9). Compounds belonging to the third class were omitted from the study since no evidence of a similar mode of action for these compounds is available thus far (see Methods).

Table 1. Biological activity, substitutents, and conformations of the stereogenic centres of the compounds included in the analysis (Templates are displayed in Figure 1).

Compound identifier	Compound name	$-\log(IC_{50})$	SCC[a]	R^1	R^2	R^3	R^4	R^5	R^6	R^7	SP[b]
1a	Dioncopeltine A[c]	1.880	1R/3R	H	H	H	CH_2OH	Me	H	H	543
1b	Ancistrocladisine	−0.236	3S[d]	–	OMe	Me	Me	Me	Me	H	576
1c	*Cis*-1,2-Dihydroancistrocladisine	−0.237	1R/3S	H	OMe	Me	Me	Me	Me	H	584
1d	*Trans*-1,2-Dihydroancistrocladisine	−0.265	1S/3S	H	OMe	Me	Me	Me	Me	H	584
1e	Ancistrocongoline D	−0.687	1R/3R	H	OH	Me	Me	Me	Me	H	563
1f	Ancistrogriffine A	0.729	1S/3S	H	OH	Me	H	Me	Me	Me	576
1g	Ancistrobertsonine D	−0.639	1R/3S	H	OH	Me	Me	Me	Me	H	563
1h	*N*-Benzyldioncopeltine A	0.299	1R/3R	Bn	H	H	CH_2OH	Me	H	H	650
1i	*N*-5′-*O*-Dibenzyldioncopeltine A	−0.001	1R/3R	Bn	H	Bn	CH_2OH	Me	H	H	761
1j	Habropetaline A	1.896	1R/3R	H	H	H	CH_2OH	Me	Me	H	562
1k	Dioncophylline A	0.419	1R/3R	H	H	H	Me	Me	Me	H	549
1l	5′-*O*-Demethyldioncophylline A	0.381	1R/3R	H	H	H	Me	Me	H	H	526
1m	4-*O*-5-*O*-Didemethyldioncophylline A	−0.343	1R/3R	H	H	H	Me	H	H	H	504
2a	7-*epi*-Dioncopeltine A	0.538	1R/3R	H	H	H	CH_2OH	Me	H	H	530
2b	Ancistrogriffine C	−0.138	1S/3S	H	OMe	H	H	H	Me	Me	565
2c	*N*-Benzyl-7-*epi*-dioncopeltine A	0.257	1R/3R	Bn	H	H	CH_2OH	Me	H	H	657
2d	7-*epi*-Dioncophylline A[c]	0.516	1R/3R	H	H	H	Me	Me	Me	H	566
2e	*N*-Methyl-7-*epi*-dioncophylline A	−0.172	1R/3R	Me	H	H	Me	Me	Me	H	580
2f	Dioncophylline D[e]	0.797	1R/3R	H	H	H	H	H	Me	Me	543
2g	–[c]	−0.369	1S/3S	Me	OMe	Me	H	Me	Me	Me	616
3a	Dioncophylline C[c]	1.782	1R/3R	H	H	H	Me	Me	H	H	511
3b	Hamateine	−0.261	–[f]	–	H	Me	Me	Me	Me	H	559
3c	Ancistrocongoline C[c,d]	−0.853	1R/3R	Me	OH	Me	H	Me	Me	Me	601
3d	Ancistrolikokine B	0.226	1S/3R	H	OH	Me	H	Me	H	Me	544
3e	Ancistrolikokine C	−0.356	1R/3R	Me	OH	Me	H	Me	H	Me	583
3f	Ancistrobertsonine B	−0.868	1R/3S	Me	OMe	Me	Me	Me	Me	H	608
3g	*N*-Methyldioncophylline C	0.453	1R/3R	Me	H	H	Me	Me	H	H	535
3h	Korupensamine B	0.936	1R/3R	H	OH	H	H	H	Me	Me	528
4a	Ancistrocladine	−0.542	1S/3S	H	H	Me	Me	Me	Me	H	593
4b	*N*-Methylancistrocladine	−0.329	1S/3S	Me	H	Me	Me	Me	Me	H	607
4c	Ancistrocongoline A	0.300	1R/3R	Me	H	H	H	H	Me	Me	556
4d	Ancistrocongoline B	0.423	1R/3R	H	Me	H	H	Me	Me	Me	580
4e	Ancistroealaine A	−0.401	3S[c]	–	Me	Me	H	Me	Me	Me	615
4f	Ancistroealaine B	−0.102	1S/3S	H	Me	Me	H	H	Me	Me	584
4g	Ancistrolikokine A	0.464	1R/3R	Me	Me	H	H	H	Me	Me	566
4h	Ancistrolikokine D	−0.305	3S[d]	–	H	Me	H	Me	H	Me	562
4i	Ancistrobertsonine A	−1.051	1S/3S	Me	H	Me	H	Me	Me	Me	605
4j	Ancistrobertsonine C	−0.347	1R/3S	Me	Me	Me	H	Me	Me	Me	635
4k	Ancistrocladeine	−0.277	–[f]	–	H	Me	Me	Me	Me	H	593
4l	Korupensamine A[c]	0.738	1R/3R	H	H	H	H	H	Me	Me	527
5a	Dioncophyllinol B[c]	1.074	1R	OH	H						544
5b	1-*epi*-Dioncophylline B	0.370	1S	H	H						547
5c	8-*O*-Methyl-1-*epi*-dioncophylline B	1.431	1S	H	Me						570
5d	Dioncophylline B	0.636	1R	H	H						552
5e	8-*O*-Methyldioncophyllinol B	−0.221	1R	OH	Me						569
6a	Dioncolactone A[c]	−0.085	H								524
6b	*N*-Benzyldioncolactone A	−0.392	Bn								634
7	–	1.454									531
8a	–	0.363	Me	Bn	H	H	H	H	OCH_2OCH_3		498
8b	–	0.047	Me	Bn	H	H	H	Br	OCH_2OCH_3		514
8c	–	−0.652	H	Bn	Me	H	OMe	H	OMe		481
8d	–	−0.321	H	H	Me	Br	OMe	H	OMe		392
8e	–	−0.036	H	Bn	Me	Br	OMe	H	OMe		492
9a	–	−0.101	H	H	H						318
9b	–	−0.647	Br	Bn	Bn						609
9c	–	−0.857	Ph	H	H						410
9d	–	−0.617	H	Bn	Bn						587

[a]SCC: stereogenic centre conformation. [b]SP: number of surface points. [c]Structure used as template of the group (see Methods). [d]Double bond between C_1 and N_2. [e]Biaryl axis is freely rotating at room temperature. [f]Isoquinoline part fully unsaturated. Bn: Benzyl. Ph: Phenyl.

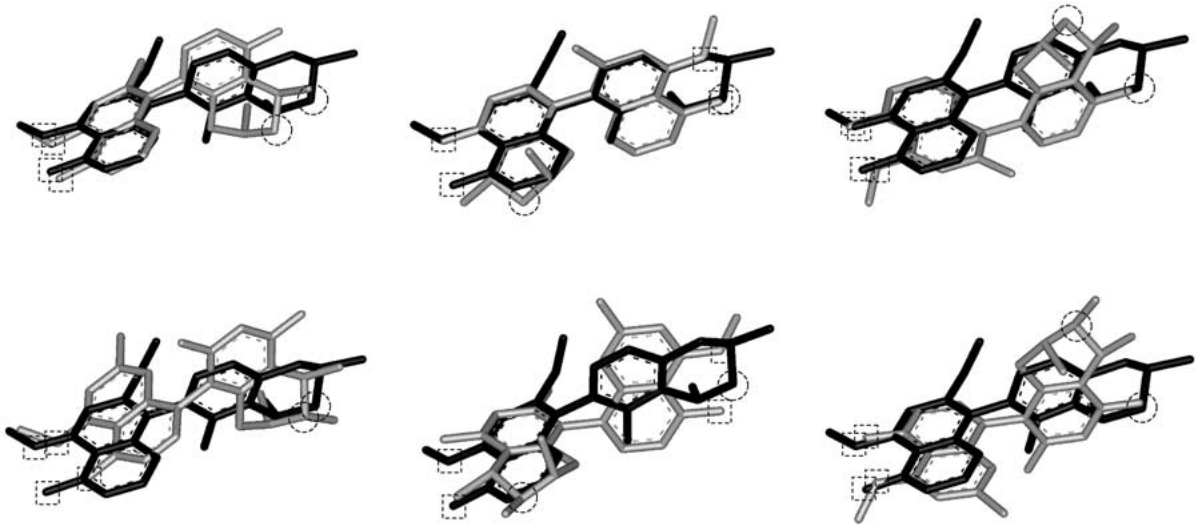

Figure 2. Three different alignment modes (left/middle/right) were found in reference 9 for dioncophylline C (**3a**, top half, light) with dioncopeltine A (**1a**, top half, dark) and ancistrocongoline A (**4c**, lower half, light) with dioncopeltine A (**1a**, lower half, dark). Dashed lines highlight the isoquinoline-nitrogen (circle) and the oxygen atoms attached to the naphthalene moiety (squares). The left alignment was chosen for the CoMSIA study, however, no information about the binding mode for the studied compounds is available thus far.

Calculating the molecular surface. Since the MaP algorithm is based on the count statistics of surface point pairs, a surface of equally distributed points is needed. Unfortunately, well known surface calculation algorithms like the Connolly [19] and MSMS [20] algorithm were not designed to produce such a surface. Inspection of surfaces calculated with these algorithms revealed a shape dependent distribution of surface points. Therefore, the GEPOL algorithm [21] was modified as to generate a grid-based surface with equally distributed surface points (dGEPOL). It should be pointed out, however, that other algorithms which generate a surface with equally distributed surface points may also be used (e.g. [22]).

Structure canonicalisation. One of the major drawbacks of all grid-based QSAR approaches is that for different positions of the molecule within the grid as well as for different grid spacings, slightly varying results may be obtained. In case of MaP this bears the risk that translational and rotational invariance is partially lost. Therefore, different precautions are taken to render the surface -and with it the molecular descriptor- translationally and rotationally invariant. First, the centre of mass of the molecule and the centre of the grid box always coincide. Second, the molecules are aligned within the grid according to their principle moments of inertia. This is achieved by solving the eigenvector problem of the mass weighted covariance matrix of the molecular coordinates. The eigenvector matrix is used as rotation matrix for the original molecular coordinates [23]. In order to retain a right-handed coordinate system of the molecules the determinant of the eigenvector matrix needs to be positive [24]. It is crucial to emphasise that this step represents a rotation of a single molecule with respect to the coordinate system. It must not be confused with an alignment with respect to a template molecule as it is necessary for 3D-QSAR techniques such as CoMFA and CoMSIA.

Mapping the molecular properties. The currently employed surface properties are H-bond acceptor (A), H-bond donor (D), hydrophilic (H), weakly hydrophobic (Lw), and strongly hydrophobic (Ls). Assignment of these properties is based on the characteristics of the atoms found in the molecule. Atomic attribute assignment is following two different frameworks. All H-bond donor/acceptor atoms are identified according to the simple rules implemented in Tripos' SYBYL [25]. Additionally, localised hydrophobicity values are assigned to each atom based on the fragmental approach by Ghose [26]. Mapping of the atomic properties onto the discretised surface follows two rules. (1) The atom closest to a particular surface point is defined as its base atom. If the base atom is classified as H-bond acceptor or H-bond donor the surface point is assigned the respective property

[27]. (2) Only surface points that are not classified as the latter can be assigned the hydrophobic or hydrophilic attribute. As 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 [28] was implemented as follows:

$$f(d_{i,j}) = \frac{1}{\exp((2/\Delta d) \cdot (d_{i,j} - d_{cut-off})) + 1}$$

where $d_{i,j}$ is the distance of the *i*th surface point to the *j*th atom, $2{\cdot}\Delta d$ defines the range wherein the function decays, and $d_{cut-off}$ is some cut-off value which is termed *proximity distance*. This proximity distance should be larger than the largest van der Waals radius of any atom in the dataset under consideration. In order to avoid overcompensation of local effects by long-range dependencies, Δd needs to be chosen appropriately. The actual hydrophobic potential (*HP*) assigned to a particular surface point is given as follows:

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

where n_A is the number of atoms, A_i is the localised hydrophobicity value of the *i*th atom (atomic contributions to logP according to Ghose [26]), and $d_{i,j}$ is the distance of the *i*th surface point to the *j*th atom.

Calculating the descriptor. The calculation of the MaP descriptor can be summarised in a three-step binning procedure and is displayed schematically in Figure 3map_calculation.

Step 1. A vector **v** of dimension n is allocated and initialised with zeros. The vector **v** consists of $(p \cdot (p+1))/2$ segments. Each segment describes a unique property combination where p equals the number of properties included in the study (e.g. $p = 5$ if the five standard surface properties A, D, H, Lw, and Ls are used). These segments are subdivided into c distance bins, where c depends on a user-defined resolution (*res, res* = 1 Å by default), and the maximum distance between surface points. The resulting dimension n of the Map vector **v** is therefore defined by

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

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 follows:

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

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 **v** are incremented. Matching means that the segment in **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 based on the concept of fuzzy counts [29,30]. The exact increment for each bin depends on the distance of $d_{i,j}$ to the closest distance bin centre (*bc*). The increment for the main bin is given as follows:

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

That means that the maximum increment for the main bin amounts to 1.0 ($bc = d_{i,j}$), whereas the minimum increment equals to 0.5 ($d_{i,j} = res/2$). Depending on the sign of $bc - d_{i,j}$, the bin above or below the bin centred at *bc* is incremented with the remainder of $1 - inc$.

The result of the algorithm outlined above is a set of $(p \cdot (p+1))/2$ distance dependent count statistics. Put differently, for each property-combination of surface point pairs, i.e. for each segment in vector **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. For ease of interpretation each variable is assigned an abbreviation describing the respective property-distance combination. For example, the variable encoding a hydrogen-bond donor surface patch (D) in a distance of 6 Å to a hydrophilic surface patch (H) will be written as DH_6 throughout the manuscript. The other property-distance combinations are abbreviated accordingly.

Chemometric methods

For a dataset of m molecules the MaP-vectors of dimension n are stacked to a matrix of dimensions $m \times n$. Next, all variables which are constant (variance equal to zero) are excluded. In this study principal component regression (PCR) is used to correlate the mean-centred descriptor matrix with biological activity values, but partial least square regression (PLS) could have been used as well. For the theory of PCR and PLS, see Martens and Naes [31]. PCR was preferred over PLS since differences in predictive ability tend to be small in combination with variable selection

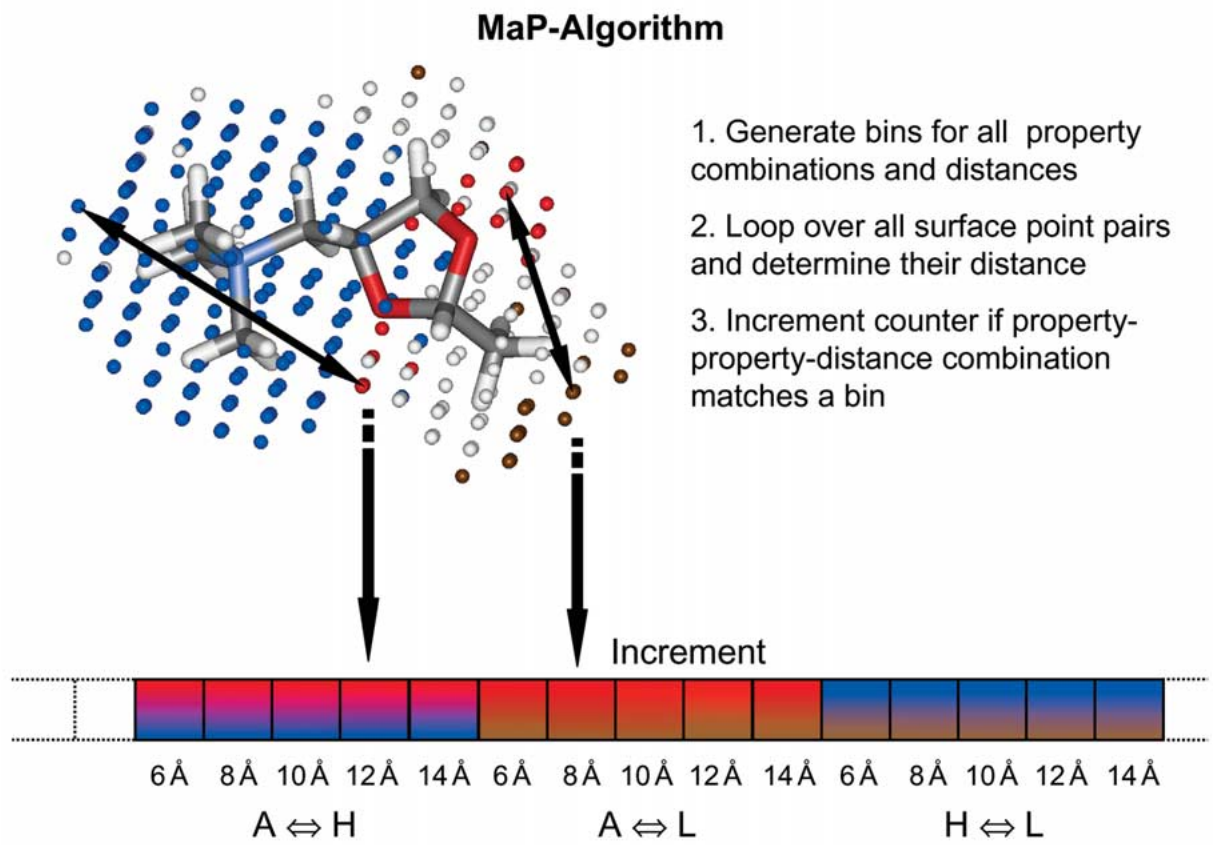

Figure 3. Scheme of the three-step calculation procedure of the MaP descriptor. Here, hydrophilic surface points (H) are coloured blue, H-bond acceptor surface points (A) are coloured red and hydrophobic surface points (L) are coloured brown. The midpoint of the distance interval is given in the Figure. Fuzzy counts are used for incrementing the matching bins (see text).

(see below) [32], and because our PCR algorithms run faster. The data were not scaled. Since not all variables of the descriptor are related to the property under study, the MaP descriptor is combined with a variable selection technique to identify the most informative variables (MIV). Stepwise regression [33] was used for variable selection in this study. Stepwise regression is a greedy search algorithm that either adds a variable to the model or removes a variable from the current model. At each step of the search the move which results in the largest improvement of the objective function is executed. The search ends at the latest in the first local optimum. To avoid that only weakly significant variables are introduced in the model, the search was stopped when changes in the objective function were only small ($< 3\%$). Since stepwise regression can easily be cast in the framework of tabu search our previously published reverse-elimination-method tabu search (REM-TS) [32] with the aforementioned stopping rule was used. With this termination criterion tabu search resembles stepwise regression. The early stopping rule was used to lower the risk of chance correlation in variable selection. Extended searching using the full capabilities of tabu search, which is a global optimisation strategy such as genetic algorithms, will also find different variable combinations with comparable or better values of the objective function (internal model quality). However, it was found in several other cases that variable combinations selected at an early stage of the search are more intuitive to interpret and that the respective models are more robust in terms of test set predictivity [34].

Since the objective function of a variable selection technique is the most crucial part of the entire procedure the stringent leave-multiple-out cross-validation (LMO-CV) procedure [35] was used for variable selection to effectively avoid overfitting [32]. The number of cross-validation runs (B) was always set to three times the number of objects in the dataset

($B = 3 \cdot m$) 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 [36,37]. Leave-multiple-out cross-validated root mean squared errors and the respective coefficients of determination were computed as follows:

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

where B is the number of cross-validation runs, k the number of objects left out (nearest integer to $0.5 \cdot m$), $y_{b,i,obs}$ is the observed property value of the ith object in the bth cross-validation run that was left out, $y_{b,i,pred}$ is the corresponding predicted property value of this object, and the subscript -k indicates the number, or the percentage of objects left out. The acronym *RMSEP* used in this contribution is also referred to as *SDEP*. From the $RMSEP_{CV-k}$ value the respective cross-validated squared multiple correlation coefficient R^2_{CV-k} can be computed as follows:

$$R^2_{CV-k} = 1 - \frac{((RMSEP_{CV-k})^2 \cdot m)}{\sum_{i=1}^{m} (y_{i,obs} - \bar{y})^2}$$

where $y_{i,obs}$ is the observed property value of the ith object that was left out and $\bar{y}$ is the mean of all property values. The denominator of the equation is also termed *SYY*. In case of the usual leave-one-out cross-validation ($k = 1$, $B = m$), the corresponding R^2-value is R^2_{CV-1} which is often referred to as q^2 in the QSAR literature. If $k \gg 1$ the estimate of the prediction error obtained by LMO-CV ($RMSEP_{CV-k}$) is biased upwards [38]. 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.

Other figures of merit used in this study are the usual coefficient of determination (R^2) and the root mean squared error of calibration (*RMSEC*; also known as *s*), which were computed as follows

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

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

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

In order to check for the risk of chance correlations [39,40], validation of the variable selection results was performed in two ways. First, a permutation test based on the repetitive randomisation of the response vector was employed. In each cycle of the test the response vector is randomly rearranged, the entire selection procedure (using the following settings: $B = 3 \cdot m$, $k = 0.5 \cdot m$, termination criterion: change of < 3 % of the objective function) is carried out on the scrambled data, and R^2_{CV-k} is recorded for each cycle. All computations were done from scratch after scrambling the responses since scrambling only the finally selected model yields far overoptimistic results [37]. If the majority of the R^2_{CV-k} values of the scrambled datasets is lower than the R^2_{CV-k} value of the original dataset it is concluded that the derived model is relevant. The second method to validate the derived models was the prediction of the biological activity for independent test sets. Here, two different types of test set splits were used. On the one hand, the test set originally published was employed (referred to as TS_{orig}, see Table 2). Additionally, the smaller naphthalene-free isoquinolines and the synthetic precursors, that could not be aligned in the previous publication [9], were used as a test set here (referred to as TS_{pre}, see Table 2). TS_{orig} is rather small ($m_{Test} = 5$). As a consequence, random fluctuations of the statistical figures of merit for test set prediction are rather large [41]. Hence, the training set compounds ($m = 38$) and the original test set (TS_{orig}) were merged ($m = 43$) and subsequently split into a new training set (66%, $m = 28$) and a new test set (33%, $m_{Test} = 15$) using Kennard-Stone's CADEX algorithm on the descriptor data [42,43]. This test set is referred to as TS_{CADEX} (see Table 2). The CADEX algorithm maximises the minimal Euclidean distances between already selected objects and the remaining objects. The selection starts with the two most distant objects using the Euclidean distance. For each of the remaining objects the shortest Euclidean distance to the already selected objects is computed and stored in a distance list. Next, the object with maximum distance in the distance list is selected. This procedure is repeated until enough objects are selected. The CADEX algorithm results in a balanced and representative split of the data. It was applied since it had performed well in training set selection in other studies [42,44].

Table 2. Compounds included in the different training and test sets used in this study. Compounds are indicated by their identifier. (Templates are displayed in Figure 1).

TS^{a}_{orig}		TS^{b}_{pre}		TS^{c}_{CADEX}	
Training set	Test set	Training set	Test set	Training set	Test set
1a–1d, 1g–1m	1e, 1f[d]	Same as TS_{orig}	6a[d], 6b	1c, 1e, 1i–1j, 1m	1a–1b, 1d, 1g–1h 1k–1l
2a, 2c–2f	2b, 2g		7	2a, 2c–2g	2b
3a–3b, 3d–3h	3c		8a–e	3b, 3d–3h	3a, 3c
4a–4c, 4e–4l	4d		9a–d	4b–4c, 4e–4h, 4j–4k	4a, 4d, 4i, 4l
5a–5b, 5c[d], 5d–5e				5a–5b, 5d	5e

[a]Original test set used in the CoMSIA study [9].
[b]Test set comprises naphthalene-free isoquinolines and the synthetic precursors.
[c]Dataset split obtained using the CADEX algorithm [42,43].
[d]Outlier, removed from analysis.

Test set representativity was assessed using a statistical test for test set outliers in **X**-space [45]. In order to detect test set outliers, the Mahalanobis distance between each test set object and the centroid of the training set is computed. Since PCR was used as regression technique, the Mahalanobis distance was computed for the optimal principal component space of the respective model. To achieve this, a principal component analysis (PCA) was carried out prior to computing the Mahalanobis distance. Next, training and test set objects were projected into the optimal PC-space as follows:

$$\mathbf{T} = (\mathbf{X} - \mathbf{1}_m \cdot \bar{\mathbf{x}}) \cdot \mathbf{V}_{PC}$$

$$\mathbf{t}_i = (\mathbf{x}_i - \bar{\mathbf{x}}) \cdot \mathbf{V}_{PC}$$

where **T** is the scores matrix of the training set objects in PC-space, **X** is the matrix of the training data descriptor variables, and **V** is the Eigenvector matrix obtained from the PCA of the mean-centred predictor variables of the training set. The subscript *PC* indicates how many principal components were used, $\mathbf{1}_m$ is a column vector of ones, $\bar{\mathbf{x}}$ is the row vector of column means of **X**, $\mathbf{x}_i$ is the row vector of predictor variables of the *i*th test object, and $\mathbf{t}_i$ is the score vector of the *i*th test set object in *PC*-space. Mahalanobis distance (*MD*) was then computed as

$$MD_i^2 = \mathbf{t}_i \cdot \left(\frac{\mathbf{T}^{\mathrm{T}} \cdot \mathbf{T}}{m-1} \right)^{-1} \cdot \mathbf{t}_i^{\mathrm{T}}$$

and compared to the 90th percentile of a χ^2 distribution with *PC* degrees of freedom (the superscript 'T' refers to the transpose operator). The test is only an approximation since it assumes that the mean and the covariance matrix are known (see [46]).

Table 3. Default parameters and extensions of the MaP parameter set.

Parameter	Value
Default:	
Grid spacing	0.8 Å
Proximity distance $d_{cut-off}$	2 Å
Δd	2 Å
Hydrophilic/hydrophobic cut-off	0
Weak hydrophobic/strong hydrophobic cut-off	0.12
Resolution (*res*) of the radial distribution function	1 Å
Extension:	
Δd	1.5 Å
Weak hydrophobic/strong hydrophobic cut-off	0.13
H-bond donor: Partial charge on heteroatom	0.18
H-bond acceptor: atom type of heteroatom	N/O

Parameter settings

Table 3 displays the default parameter settings of the MaP procedure and the extensions of the parameter set discussed in the following. (1) In order to identify the influence of long-range dependencies of the fermi-type function, parameter Δd was set to 1.5 Å. This change necessitates an adaption of the weak hydrophobicity/strong hydrophobicity (Lw/Ls) cut-off to 0.13. Variation of Δd to a lower value emphasises local effects of hydrophobicity and reduces the impact of long-range effects. Hence, variability of the actual surface property increases. (2) To distinguish between different H-bond donor/acceptor strengths, additional cut-off values were incorporated into the parameter set. For H-bond donors the partial charge of the hydrogen atom connected to the respective het-

Table 4. Model settings applied.

Model number	Δd[a]	H-bond donor differentiation	H-bond acceptor differentiation	Total number of descriptor variables
1	−	−	−	300
2	+	−	−	300
3	−	+	−	420
4	+	+	−	420
5	−	−	+	420
6	+	−	+	420
7	−	+	+	560
8	+	+	+	560

[a]For Δd '−' equals to 2 Å and '+' equals to 1.5 Å.

eroatom is employed [47]. The cut-off applied was set to 0.18 after inspection of the AM1 atomic partial charge distribution of hydrogens in H-bond donors published by Ghafourian and Dearden [47]. To distinguish between different H-bond acceptors strengths, the atom type (N/O) of the relevant atom is used [48,49]. Even though the latter differentiation seems to be insufficient since H-bond acceptor strength is highly dependent on the substitution pattern [49], it is useful for a qualitative assessment. Essentially, nitrogen atoms are assigned a strong hydrogen-bond acceptor attribute, whereas oxygen atoms are classified as weak hydrogen-bond acceptors. Additional information on oxygen H-bond acceptor strength is included in the size of the corresponding surface patches. For instance, owing to its larger H-bond acceptor surface patch size, the carbonyl oxygen shows an increased probability to interact with an hydrogen-bond donor group [48].

As has been stated before, the goal of this study was to investigate the impact of the following operational parameters of the MaP procedure: hydrophobicity mapping (Δd), differentiation of the H-bond donor strength, and differentiation of the H-bond acceptor strength. Therefore, the three parameters were varied according to a full factorial design which is shown in Table 4 . The resulting models were assessed according to their statistical quality and their interpretability. It should be noted, that the total number of variables for the descriptor stays constant for varying Δd but increases if H-bond donor or acceptor strength is differentiated. For each differentiation added, the parameter p in Equation (1) increases by one resulting in a larger number of descriptor variables (see also Table 4 for the total number of descriptor variables calculated).

Geometry optimisation of the dataset

The dataset under scrutiny is shown in Figure 1 and Table 1. It can be subdivided into three major classes. The largest class (class 1) is composed of 39 naphthylisoquinoline alkaloids with varying coupling positions (groups 1–5). The second class comprises compounds containing an isoquinoline moiety. It includes synthetic precursors of the naphthylisoquinolines (group 6), the phenylisoquinoline **7** and the much smaller isoquinolines (groups 8–9). Compounds belonging to the third class lack the isoquinoline moiety and can be subdivided into nitrogen containing and nitrogen free compounds. The former are synthetic naphthylindenes (groups 10–11) and the latter incorporate knipholone and derivatives thereof (group 12) as well as the phenylnaphthalene **13** and the bromonaphtheate **14**.

It was already noted [9], that the compounds belonging to class 3 need to be treated with caution when included in the same QSAR model with classes 1 and 2. This is owing to the fact that no evidence of a similar mode of action for these compounds is given so far. Even though these compounds can be included into a model without loss of model quality [9], this must not be confused with a potentially similar binding mode but is more likely to be a result of the low variance in biological activity. Hence, in this contribution the third class of compounds is omitted from the QSAR study.

All structures were built according to the following scheme. First, a template molecule for each group of compounds was selected (see Table 1 and Figure 1). For the much smaller isoquinolines, the isoquinoline moiety of *N*-benzyl-7-*epi*-dioncopeltine A (**2c**, template for group 8) and hamateine (**3b**, template for group 9) were chosen owing to their high similarity of the respective substructure. Because of previous energy calculations [9], the torsional angle along the bond connecting the two ring systems of the axially chiral compounds was set to 90° (atoms that were used to define this angle are marked in Figure 1). Next, all template structures were minimised using the Tripos force field and Gasteiger-Hückel charges as implemented in SYBYL 6.8 [25]. For all template molecules, the flexible hydroxyl and methoxy groups were subjected to a conformational search in order to identify low energy conformers. Since similar flexible groups of the resulting template conformers were

pointing into similar directions in space no additional re-orientation of these groups was carried out. Starting from the obtained template molecules, the other compounds were built and minimised according to a standard protocol which was as follows: all constant structural features of compounds with the same template were kept fixed and the varying parts were added and subsequently geometry optimised. No further user interaction was involved. It is worth to note, that after geometry optimisation similar groups oriented similar in space. After all compounds were geometry optimised, the full dataset was canonicalised, AM1 charges were assigned, and the MaP descriptor was calculated.

Biological data

Antiplasmodial activity was determined employing a modification of the [^{3}H]-hypoxanthine incorporation assay [50]. The testing was performed using the K1 strain, which is resistant to chloroquine and pyrethamine. In essence, infected human red blood cells are exposed to serial drug dilutions in microtiter plates for 48 hours at 37 °C in a gas mixture with reduced levels of oxygen and increased levels of CO_2. Then [^{3}H]-hypoxanthine is added to each well and after further incubation for 24 hours the wells are harvested on glass fibre filters and counted in a liquid scintillation counter. From the resulting sigmoidal curve the IC_{50}-values are calculated in ng/ml and then converted into molar units [nmol/ml]. The latter are transformed into $[-\log(IC_{50})]$ values for the QSAR studies.

Results

First, all models were screened for possible outliers by inspection of leave-one-out cross-validated residuals of the training dataset for all eight models. For the training dataset one compound was identified as an outlier (8-*O*-methyl-1-*epi*-dioncophylline B (**5c**), see Table 1) due to its notably larger residuals in all models. The compound was removed from the training set and all models were recalculated. Next, the test set (TS_{orig}, Table 2) was screened for test set outliers in **X**-space using the Mahalanobis distance [45,46]. None of the six test set objects was identified as an outlier in **X**-space in any of the eight models given in Table 5. Despite the fact that all compounds seemed well embedded in training data space, ancistrogriffine A (**1f**) was predicted too low by approximately one log unit in all models (outlier in **Y**-space) and was removed from the test set. Although the Mahalanobis distance did not reveal any unusual structural features for compound **1f**, it is indeed a structural singleton as will be detailed in the discussion. The detection of compounds **5c** and **1f** as outliers was not unexpected since both of them were already identified as outliers in the CoMSIA model [9]. Additionally, in the CoMSIA model two more compounds of the training data needed to be removed. Owing to these additional outliers and owing to the fact that the composition of the training data was different for the published CoMSIA model (inclusion of non-nitrogen containing compounds), a direct comparison of the statistical figures of merit of the two techniques is hard to perform and thus was not pursued.

For all variations of the parameter set (see below) a more or less pronounced influence on test set prediction (TS_{orig}) was found. However, this may not necessarily be a consequence of significant differences in test set prediction but may simply be a result of statistical variations. The relative standard deviation in $RMSEP_{Test}$ can be computed as follows [41]:

$$rel\,sdv(RMSEP) = \sqrt{\frac{1}{2 \cdot m}} \qquad (2)$$

where m is the number of compounds that were used to estimate *RMSEP*. For a test dataset of this size ($m_{Test} = 5$) it is about 32%. This carries over to R^2_{Test}-values ranging from 0.40 to 0.84 for an $RMSEP_{Test}$ of 0.40 ± 0.13 in the case considered here ($SYY_{Test} = 2.30$). Owing to this very high relative standard deviation, R^2_{Test} is not the ideal parameter to compare the performance of the different parameter settings. Hence, $R^2_{CV-50\%}$ was used as an indicator for model quality.

However, it should be noted that the $R^2_{CV-50\%}$-values themselves are slightly overoptimistic since $R^2_{CV-50\%}$ is used as the objective function for variable selection. This phenomenon is called selection bias in the statistical literature [51]. Yet, selection bias does not matter here because all models compared were generated with identical settings for the objective function.

Analysis of parameter changes

Inspection of the different models built according to Table 4 showed no significant influence of any of the varied parameters on the $R^2_{CV-50\%}$ (see Table 5). This is supported by the fact that the two most significant variables (MIV) of all models are essentially identical in their chemical meaning.

Table 5. Results for the full factorial design analysis.

Model	$RMSEP_{CV-1}$ [a]	R^2_{CV-1} [b]	$RMSEP_{CV-50\%}$ [c]	$R^2_{CV-50\%}$ [c]	*RMSEC* [d]	R^2 [e]	$RMSEP_{Test}$ [f]	R^2_{Test} [f]	m/m_{Test} [g]	n [h]	n_{sel} [i]	*PC* [j]
1	0.41	0.64	0.44	0.59	0.37	0.75	0.39	0.68	38/5	269	6	5
2	0.37	0.72	0.41	0.65	0.34	0.79	0.59	0.24	38/5	271	4	4
3	0.41	0.65	0.42	0.63	0.39	0.71	0.42	0.62	38/5	352	2	2
4	0.38	0.70	0.41	0.65	0.35	0.77	0.33	0.76	38/5	354	3	3
5	0.39	0.67	0.43	0.61	0.34	0.79	0.31	0.79	38/5	349	5	5
6	0.35	0.74	0.38	0.69	0.32	0.82	0.61	0.19	38/5	350	5	5
7	0.38	0.69	0.40	0.67	0.36	0.75	0.42	0.62	38/5	435	3	2
8	0.38	0.70	0.39	0.68	0.35	0.76	0.43	0.60	38/5	436	4	2

[a] $RMSEP_{CV-1}$: leave-one-out cross-validated root mean squared error of prediction. [b] R^2_{CV-1}: leave-one-out cross-validated coefficient of determination. [c] Same as [a] and [b] for leave-50%-out cross-validation. [d] *RMSEC*: root mean squared error of calibration. [e] R^2: coefficient of determination. [f] Same as [a] and [b] for test set prediction of TS_{orig}. [g] m: number of objects. [h] n: number of variables with non-zero variance. n^i_{sel}: number of variables selected by the search algorithm. [j] *PC*: number of principal components.

However, differentiating hydrogen-bond donor strength results in smaller models (i.e. less selected variables) with fewer principal components which is considered an advantage. Moreover, models missing this hydrogen-bond donor differentiation are harder to interpret and show in some cases a reduced external predictivity (models 2 and 6, Table 5) indicating the importance of this differentiation for the dataset under study.

Criteria for model selection and analysis of model 4

For QSAR methods employing a variable selection procedure the object to selected variable ratio should be at least 6 in order to reduce the risk of chance correlations [52]. The latter risk is directly related to the number of available variables (the larger the pool of candidate variables, the larger the risk of chance correlation) [39]. Consequently, choosing a model for closer investigation from the set of models with varying MaP parameters was done along the following lines: (1) Model interpretability should be chemically meaningful. (2) The number of MIVs should be small. (3) The less parameters are added to the original parameter set the better it is, since more parameters result in more variables (i.e. a smaller p is preferred).

According to the guidelines stated above, model 4 (Table 5) was chosen for statistical evaluation and for closer investigation of the dataset because of its good statistical figures of merit, its few variables, and its very good interpretability (see Discussion).

Permutation test of the data. A permutation test was carried out to assess the risk of chance correlation. The results of the permutation test for model 4 are summarised as follows: The 50th- and 95th-percentile of the distribution of $R^2_{CV-50\%}$-values (referred to as $R^2_{CV-50\%,PT}$) for 500 permutations of the response vector are 0.05 and 0.32, respectively. That means that 95% of the data show a $R^2_{CV-50\%,PT}$ of less than 0.32. The maximal $R^2_{CV-50\%,PT}$ was 0.57. Hence, the probability of chance correlation is very low. Results of the permutation test are also displayed in Figure 4. Here, the evolution of the percentiles (including the minimum and maximum) depending on the number of permutations is shown. It can be seen that after 200 permutations the 25th-, 50th-, and 75th-percentiles are stable and that no $R^2_{CV-50\%,PT}$ is equal to or higher than the original $R^2_{CV-50\%}$.

Table 6. Actual vs. predicted -log(IC_{50}) values for TS_{orig} found for model 4.

Compound identifier	Actual $-\log(IC_{50})$	Predicted $-\log(IC_{50})$
1e	−0.6870	−0.3748
2b	−0.1380	0.2711
2g	−0.3690	−0.4196
3c	−0.8530	−0.4124
4d	0.4230	0.7149

Test set prediction, and new dataset split. Test set prediction of the original test set (TS_{orig}, $m_{Test} = 5$) was good ($R^2_{Test} = 0.76$, see also Figure 5 and Table 6). In addition to the original test set, biological activities of some synthetic precursors and isoquinolines (TS_{pre}, $m_{Test} = 12$) were predicted.

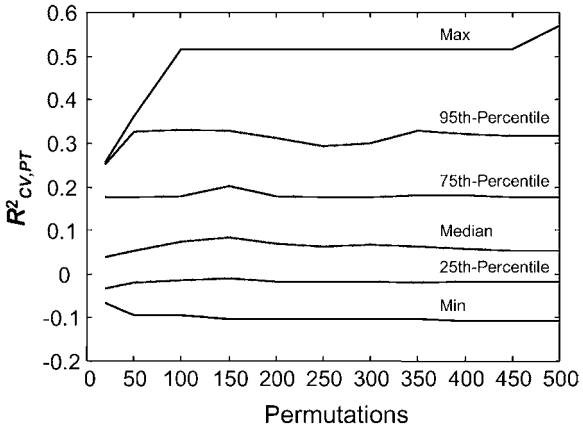

Figure 4. Results of the permutation test plotted against the number of permutations. It can be seen that the percentiles are quite stable after 200 permutations. Fewer permutations should not be carried out. Although percentiles change only slightly after 200 permutations more permutations are advantageous with respect to the power of the test [54].

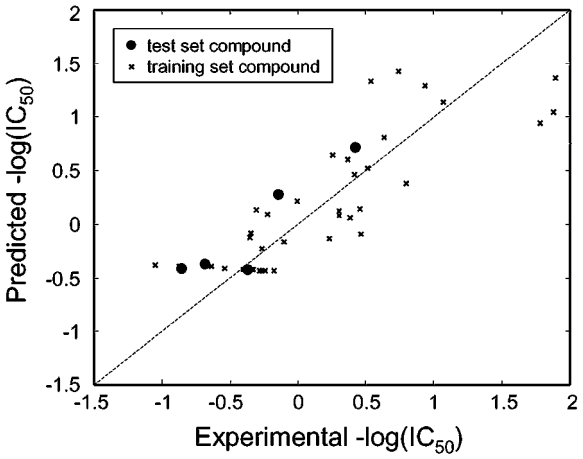

Figure 5. Experimental vs. calculated biological activity scatter plot obtained for model 4 and TS_{orig}.

Since these test data differ from the structures of the training data, a test for test set prediction outliers using the Mahalanobis distance [45] was carried out before prediction. Compound **6a** (dioncolactone A) was identified as a test set outlier and was removed. The remaining 11 test set coumpounds of TS_{pre} were then predicted and the result was comparatively good with an R^2_{Test} of 0.67 (see Table 7).

Test set prediction is the gold standard in validating QSAR models, and is of particular importance if variable selection is applied for obtaining the model [53]. It was already mentioned that statistical figures of merit for test set prediction show a large random fluctuation for small datasets (see Equation 2). Hence, to further validate the MaP model the training dataset ($m = 38$) and TS_{orig} ($m_{Test} = 5$) were merged and subsequently split into a training set of reduced size ($m = 28$) and an enlarged test set ($m_{Test} = 15$, TS_{CADEX}) using the CADEX algorithm. The MIVs for this new split of the data were determined using the same selection procedure as before (PCR, stepwise regression, leave-50%-out CV, early stopping rule). The selected variables for the CADEX split ($DwDs_7$, ADs_{11}, DsH_{14}; $m = 28$) were very similar to the variables selected for the original training set ($DwDs_7$, ADs_{10}, and DsH_{15}; $m = 38$). This supports the relevance of the selected variables and underlines the robustness of MaP models. Finally, the 15 compounds of TS_{CADEX} were predicted which resulted in a R^2_{Test} of 0.70 indicating a good predictive ability of the chosen model (see Table 7). Finally, permutation testing for the training data of the CADEX split indicated a low probability of chance correlation since none of the 500 permutated datasets showed an equal or higher $R^2_{CV-50\%,PT}$ than the real $R^2_{CV-50\%}$ (50th-percentile = 0.09, 95th-percentile = 0.45, and maximum $R^2_{CV-50\%,PT} = 0.68$).

Discussion

One of the disadvantages of the inclusion of additional parameters into the MaP parameter set is the inflation of the number of total variables. These additional variables increase the risk of chance correlation in the variable selection procedure. Nevertheless, if a preliminary analysis indicates that a specific property might be crucial for biological activity, additional parameters should be included. Otherwise the user should go with the more conservative default settings. In case of this dataset, the model obtained with the default settings (model 1, Table 5) revealed all important properties for biological activities. However, introducing additional parameters improved model size as well as model interpretability as will be detailed below.

Models without further H-bonding differentiation. Models with no further differentiation of H-bonding properties (models 1,2) always included variable DH_6 with a positive regression coefficient in the equation. DH_6 describes an H-bond donor surface patch in a distance of 6 Å to a hydrophilic surface patch. Even though this variable already pinpoints one of the important features for biological activity, its interpretation is rather difficult since it implicitly encodes

Table 7. Extended test set results for Model 4 (for symbols and abbreviations see Table 5).

Test set	$RMSEP_{CV-1}$	R^2_{CV-1}	$RMSEP_{CV-50\%}$	$R^2_{CV-50\%}$	*RMSEC*	R^2	$RMSEP_{Test}$	R^2_{Test}	m/m_{Test}	*N*	n_{sel}	*PC*
TS_{pre}[a]	0.38	0.70	0.41	0.65	0.35	0.77	0.41	0.67	38/11	354	3	3
TS_{CADEX}[b]	0.28	0.78	0.33	0.70	0.25	0.85	0.46	0.70	28/15	364	3	3

[a] TS_{pre} comprises the synthetic precursors and the isoquinolines.

[b] The CADEX split was applied to all 43 compounds of the original training and test dataset.

two characteristics of biologically active compounds. Firstly, it highlights the spatial arrangement of potential H-bond donor groups and secondly, it encodes the distribution of hydrophilic surface patches. Put differently, DH_6 is increased if the following two combinations of functional groups are found in a molecule. (1) Oxygen atoms in positions 4′ and 5′ of the naphthyl moiety, in combination with a hydroxyl group in position 8 and a hydroxymethyl group in position 2′, respectively. Here, the surface area between the oxygen atoms of the substituted naphthyl moiety shows a hydrophilic character (H), whereas the hydroxyl groups act as H-bond donors (D). (2) An unsubstituted nitrogen in the isoquinoline moiety, in combination with a hydroxyl group in 8 and a hydroxymethyl group in position 2′, respectively. Here, the surface of the saturated ring system close to the nitrogen atom of the isoquinoline ring system is partly hydrophilic (H) and the hydroxyl groups act as H-bond donor groups (D). If the nitrogen atom is substituted, the size of the hydrophilic surface patch drastically decreases (Figure 6) and with it the biological activity of the molecules. Both combinations occur at a distance of 6 Å (Figure 7).

Unfortunately, identification of these features needed a careful analysis since two phenomena are confounded in this variable. This is owing to the fact, that *N*-substitution is not only encoded by the size of the hydrophilic surface mentioned above, but also by an H-bond donor surface patch close to the hydrogen atom of the unsubstituted isoquinoline nitrogen. The latter in combination with the hydrophilic part (H) of the hydroxyl or hydroxymethyl groups in positions 8, 4′, and 2′, respectively, results in additional counts in DH_6.

Models with H-bond donor differentiation. Interpretation of the models for which weak and strong hydrogen bond donors are differentiated, was highly simplified. All models employing this extended parameter set (models 3, 4, 7, and 8) identify two variables ($DwDs_7$ and ADs_{10}) which are chemically similar to the ones identified by the models discussed above (default parameters: DH_6 and AD_{11}). However, interpretation of $DwDs_7$ is much more straightforward than for DH_6 since confounding is reduced.

Comparing $DwDs_7$ and DH_6 showed that the distance of the two variables differs by 1 Å. This is a consequence of the different direction in which the H-bond donor surface area and the hydrophilic surface area of the unsubstituted nitrogen atom are pointing. The QSAR equation found for the model chosen for further investigation owing to its overall good properties (model 4) is as follows:

$$\begin{aligned} -\log(IC_{50}) &= 0.1844 + 0.0092 \cdot DwDs_7 \\ &\quad + 0.0072 \cdot ADs_{10} \\ &\quad - 0.0354 \cdot DsH_{15} \end{aligned}$$

$$\begin{aligned} R^2_{CV-50\%} &= 0.65, \quad RMSEP_{CV-50\%} = 0.41, \\ R^2 &= 0.77, \quad R^2_{Test} = 0.76, \\ m/m_{Test} &= 38/5, \quad PC = 3 \end{aligned}$$

where $-\log(IC_{50})$ is the estimated biological activity, m is the number of objects of the training data, m_{Test} is the number of test objects (TS_{orig}), PC is the number of principal components and the data are mean-centred. When using PLS instead of PCR as regression technique for variable selection the same MIVs and the same number of latent variables were identified. In both cases the number of latent variables equals the number of selected variables. Consequently, the mathematical models for PCR and PLS, resemble that of multiple linear regression (MLR) in this case. Therefore, all figures of merit not depending on cross-validation are identical for PCR and PLS. In addition to that cross-validated figures of merit were equal within rounding error precision.

In the derived QSAR equation high biological activity is mainly described by $DwDs_7$ and ADs_{10}. The importance of the single variables can be described by the fraction of the variance of the responses they explain. For model 4, $DwDs_7$ explains 51% of

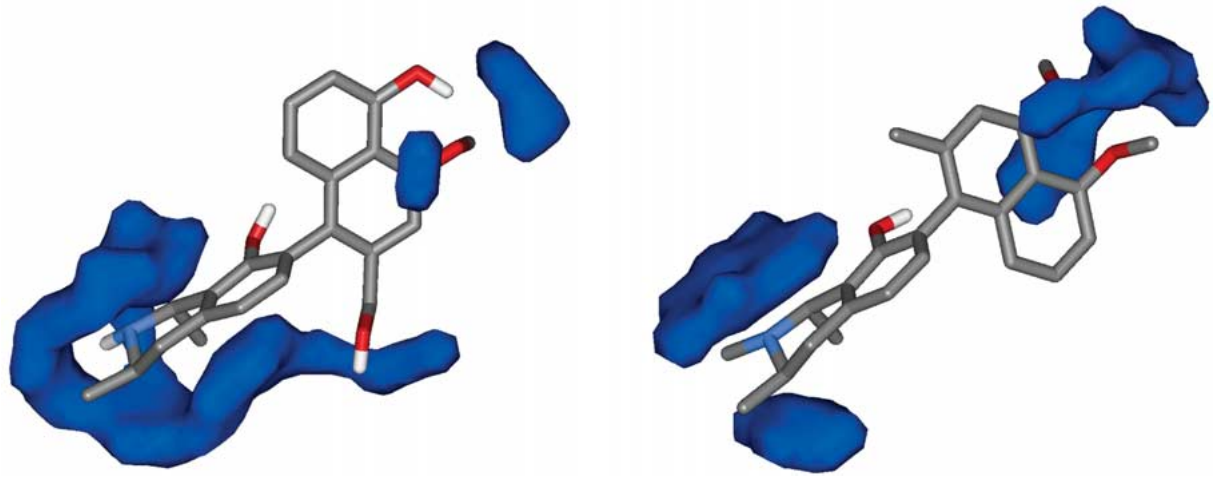

Figure 6. Hydrophilic surface areas for the *N*-unsubstituted dioncopeltine A (**1a**, left) and the *N*-substituted *N*-methyl-7-*epi*-dioncophylline A (**2e**, right). For *N*-substituted compounds, the hydrophilic surface area of the isoquinoline moiety is drastically decreased, which in turn results in lower predicted biological activity.

the variance of the data, adding ADs_{10} explains an additional 12%, and adding DsH_{15} explains two more percent. Hence $DwDs_7$ and ADs_{10} account for the bulk of the variance, whereas DsH_{15} is responsible for some fine-tuning regarding the overall shape of the compounds (position of the biaryl axis).

$DwDs_7$ encodes the spatial arrangement of the unsubstituted nitrogen atom in position 2 of the isoquinoline moiety (Dw) with respect to the strong H-bond donor groups (Ds) in positions 8 (hydroxyl) and 2′ (hydroxymethyl), respectively (see Figure 7). It can easily be seen that nitrogen substitution is encoded implicitly. If the nitrogen atom is substituted, the weak H-bond donor surface patch (Dw) is missing. As a result, the variable $DwDs_7$ cannot be found in *N*-substituted compounds, which in turn renders their predicted biological activity very low. For unsubstituted compounds, the absence of any of the two hydroxyl groups results in a reduced value for $DwDs_7$, indicating that both moieties are necessary for enhanced activity.

ADs_{10} additionally highlights the importance of an unsubstituted hydroxyl function in position 8 of the isoquinoline ring system (strong H-bond donor) as well as an H-bond acceptor region in position 4' of the naphthyl system. If the latter can also act as H-bond donor group (i.e. a hydroxyl function) a triangle of potentially pharmacophoric points including the nitrogen atom of the isoquinoline ring is formed (Figure 7) and the predicted biological activity is further enhanced.

Combination of variables $DwDs_7$ and ADs_{10} results in a pharmacophore hypothesis consisting of three important surface area properties in a triangular arrangement. The hypothesis incorporates a weak H-bond donor surface area in a distance of 6–8 Å to a strong H-bond donor surface area ($DwDs_7$), and a strong H-bond donor surface area at a distance of 10–12 Å to an H-bond acceptor surface (ADs_{10}). The third side of the triangle constitutes a weak H-bond donor surface area (from **Dw**Ds_7) in a distance of 12–14 Å to an H-bond acceptor surface area (from **A**Ds_{10}). These findings are displayed in Figure 8. It is important to note that this hypothesis can be fulfilled by compounds of the structurally different groups. However, since structurally different groups fulfil the hypothesis with different substructures, counts in the respective variables differ. More active compounds of groups 1 and 2 show higher counts for variable ADs_{10} than the more active compounds of groups 3 and 4. This between-group difference is accounted for by variable DsH_{15} (negative sign), for which the respective compounds of groups 1 and 2 show higher counts. Parts of the hypothesis, like the unsubstituted nitrogen atom, were already identified as relevant for high biological activity by the CoMSIA model [9]. This agreement supports the relevance of the model outlined above.

Models with H-bond acceptor differentiation. When comparing models with and without additional H-bond acceptor differentiation (Table 5, models 5,6 vs. models 1,2) no obvious advantages with respect to $R^2_{CV-50\%}$, or model complexity can be identified.

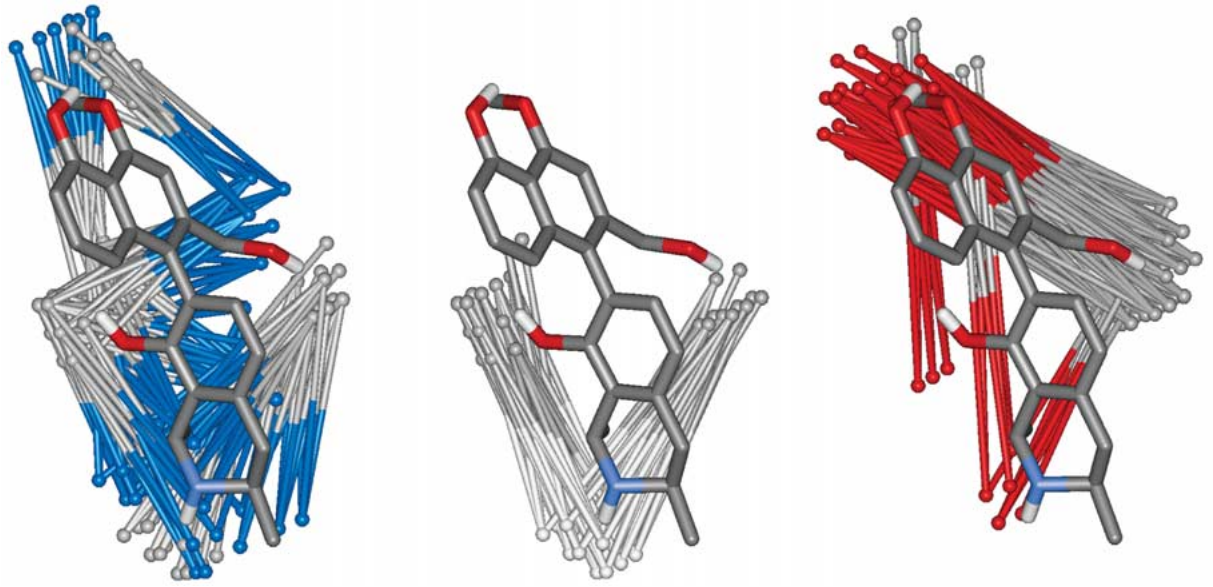

Figure 7. Variable DH_6 (left), $DwDs_7$ (middle), and ADs_{10} (right) back-projected into the original molecular space of dioncopeltine A (**1a**). DH_6 describes essentially the same feature as $DwDs_7$ but it is much harder to interpret.

Incorporating both, H-bond donor and H-bond acceptor differentiation (models 7,8), resulted in no added-value when compared to H-bond donor differentiation alone (models 3,4). Hence, with respect to the criteria for model selection mentioned before, H-bond acceptor differentiation was not effective for this dataset.

Test set prediction and new dataset split. Prediction of the original test set (TS_{orig}, $m_{Test} = 5$) was good ($R^2_{Test} = 0.76$) after one outlier was removed (outlier in **Y**-space). Analysis of the residuals exhibited that the outlier ancistrogriffine A (**1f**) was predicted too low by an amount of approximately one unit in $-\log(IC_{50})$. Revisiting the test set predictions of the other models in Table 5 confirmed that the biological activity of this compound was always underestimated by the test set predictions. This indicates that this structure differs from all other compounds in the same group. When scrutinising ancistrogriffine A (**1f**), it became apparent that it is the only compound in this group with a methyl substituent in position 7′ and an unsubstituted 4′ position. Although the model is not able to correctly predict this compound, visual inspection revealed reasons of this different behaviour. In this case, a highly active outlying compound helped to identify new features potentially responsible for increased biological activity.

Owing to the structural difference of the synthetic precursors and the isoquinolines (TS_{pre}) to the training data, prediction of their activities is rather difficult. One may suspect that the binding mode is different. If, however, a general pharmacophoric pattern is responsible for binding of the compounds, prediction of structurally different compounds should be possible. For the studied naphthylisoquinolines, this seems to be the case since predictions of the structurally different test compounds of TS_{pre} were comparatively good. Dioncolactone A (**6a**) was removed prior to prediction since it was identified as a test set outlier (see Results). Predicting this compound despite its structural differences results in a predicted biological activity that is too low. This can be explained as follows: the lacking chemical stability of the lactone substructure in the biological assay in combination with its configurational (torsional) flexibility along the biaryl axis might be a reason for its measured activity. If this strained lactone structure was hydrolysed (e.g. by esterases or by water) a hydroxyl function in position 8 and a free carboxylate group at C-2′ become available. These substructures in combination with an unsubstituted nitrogen atom in the isoquinoline moiety are part of the pharmacophoric pattern identified by the aforementioned model. This might explain the biological activity observed and the wrong prediction based on the lactone ring.

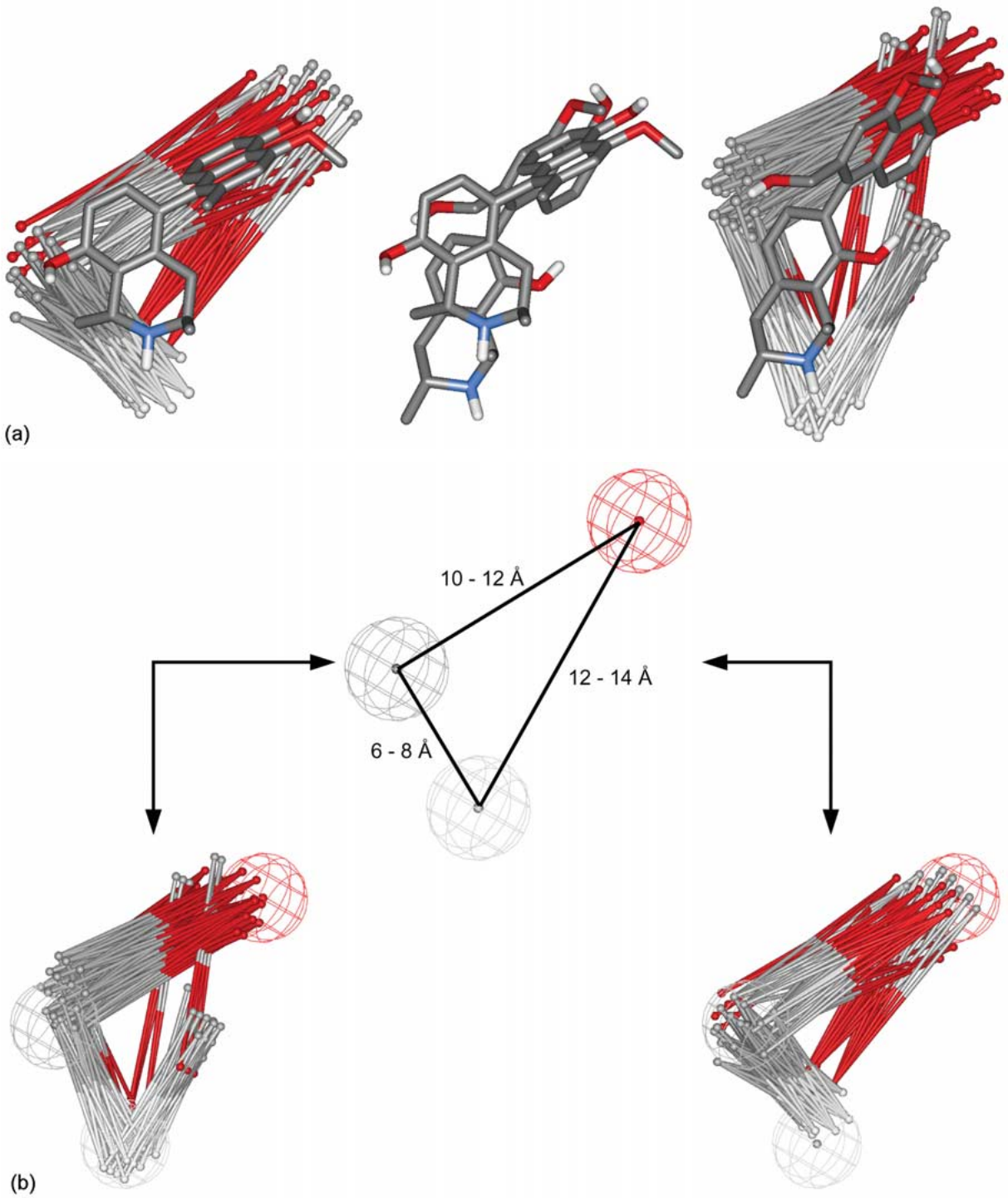

Figure 8. **(a)** Rigid body fit of dioncophylline C (**3a**, left) and dioncopeltine A (**1a**, right). The structures were aligned using the hydrogen atoms connected to the isoquinoline nitrogen, the hydrogen atoms of the hydroxymethyl group in position 2′, and the hydroxyl group in position 8, and the oxygen atoms in position 4′. It can be easily seen that the atom based alignment of the highly active molecules is rather bad (centre). Nevertheless, back-projecting the selected variables of both compounds highlights the similarity between them. Note that the compounds displayed have a completely different connection pattern along the biaryl axis. **(b)** Combination of the first two selected variables in a pharmacophore model. Since MaP is based on single conformations of the molecules, the distances found between the different surface patches are approximations.

The original test set (TS_{orig}) was acquired in a natural way (first compounds that became available after the training set was established) rather than designed. As a consequence, it was comparatively small and unbalanced. In an effort to thoroughly validate the MaP model for the naphthylisoquinoline alkaloids a second split into training and test data was computed (CADEX split, see Chemometric Methods). The figures of merit, internal as well as external, for the CADEX split are comparable to those of model 4 (Table 5). Once more, exchanging PCR for PLS in the variable selection routine resulted in identical models and identical figures of merit. Most importantly, the selected variables differ only in terms of the distances between the identified surface properties (difference: 1 Å for variables 2 and 3) and not in the properties themselves (see Results). This consistency in MIVs is a very strong indicator for the robustness and the relevance of the model. It should be recalled that the variable selection procedure for the CADEX split is applied to a training dataset of only 28 structures. This means that fewer compounds are also sufficient to identify the pharmacophoric pattern in the dataset under scrutiny. Nonetheless, the compounds in the test set (TS_{CADEX}) are well predicted, which in turn means that these structures show the pharmacophoric pattern as well. Consequently, the pharmacophore hypothesis postulated based on model 4 (Figure 8) seems very likely even though the distances separating the surface properties vary within a certain range.

Conclusion

A systematic evaluation of an extension of the MaP parameter set was performed. In addition to the variation of a parameter responsible for the mapping of the local hydrophobicity/hydrophilicity, a differentiation of hydrogen bonding strengths was incorporated. All parameter settings were analysed quantitatively (model quality) as well as qualitatively (interpretability). Internal statistical figures of merit were robust with respect to the changes of the parameter set. Test set prediction and interpretability indicated the importance of an H-bond donor strength differentiation for this dataset.

The dataset under scrutiny comprises structurally related antimalarially active naphthylisoquinoline alkaloids. A CoMSIA study in combination with different alignment techniques was already published for this dataset. In particular, the alignment posed significant problems [9] which did not occur in this study owing to the translational and rotational invariance of the MaP procedure.

Structural features deemed important for biological activity in the CoMSIA study were also highlighted in the MaP model (unsubstituted nitrogen atom in the isoquinoline ring system). Additionally, some more important structural features and their geometric arrangement could be identified. This led to a triangular-shaped pharmacophore hypothesis (Figure 8) that includes the unsubstituted nitrogen in position 2, the unsubstituted hydroxymethyl function in $2'$ and the unsubstituted hydroxyl functions in 8, and $4'$. Moreover, the analysis of an outlier (compound **1f**) found in the original test dataset allowed identification of an additional methyl substructure as potentially relevant for biological activity.

In summary, MaP models proved to be robust. No significant changes of internal model quality were observed for minor changes of the parameter set. This indicates that relevant features of potential ligands are sufficiently described in either setting. Differentiation of H-bond donor strength highly simplified model interpretation though. However, since extension of the parameter set is accompanied by an increased number of variables, a thorough validation of the resulting models is necessary.

Acknowledgements

Financial support of the Fonds der Chemischen Industrie, Frankfurt/Main, Germany is gratefully acknowledged.

References

1. Cramer, R.D., Patterson, D.E., Bunce, J.D., J. Am. Chem. Soc., 110 (1988) 5959.
2. Klebe, G., Abraham, U., Mietzner, T., J. Med. Chem., 37 (1994) 4130.
3. Kim, K.H., Martin, Y.C., J. Org. Chem., 56 (1991) 2723.
4. Pastor M, Cruciani, G., Watson, K.A., J. Med. Chem., 40 (1997) 4089.
5. Pastor, M., Cruciani, G., McLay, I., Pickett, S., Clementi, S., J. Med. Chem., 43 (2000) 3233.
6. Stiefl, N., Baumann, K., J. Med. Chem., 46 (2003) 1390.
7. Goodford, P.J., J. Med. Chem., 28 (1985) 849.
8. Fontaine, F., Pastor, M., Sanz, F., In M. Ford, D. Livingstone (Eds.), *Proceedings of the 14th European Symposium on QSAR*, Blackwell Science, Oxford, UK. In Press.
9. Bringmann, G., Rummey C., J. Chem. Inf. Comput. Sci., 43 (2003) 304.

10. Bringmann, G., Pokorny, F., In Cordell, G.A. (Ed.), The Alkaloids. Vol. 46, Academic Press, New York, 1995, pp. 127-271.
11. Bringmann, G., Feineis, D., Act. Chim. Thérapeut., 26 (2000) 151.
12. Bringmann, G., In Vial, H., Fairlamb, A., Ridley, R. (Eds.), *Guidelines and Issue for the Discovery and Drug Development Against Tropical Diseases*, World Health Organisation, Geneva, 2002, in print.
13. Bravi, G., Gancia, E., Green, D.V.S., Hann, M.M., *Methods and Principles in Medicinal Chemistry*, 10 (Virtual Screening for Bioactive Molecules) (2000) 81.
14. François, G., Chimanuka, B., Timperman, G., Holenz, J., Plaizier-Vercammen, J., Aké Assi, L., Bringmann, G., Parasitol. Res., 85 (1999) 935.
15. Chimanuka, B., François, G., Timperman, G., Heyden, Y.V., Holenz, J., Plaizier-Vercammen, J., Bringmann, G., Parasitol. Res., 87 (2001) 795.
16. François, G., Timperman, G., Steenackers, T., Aké Assi, L., Holenz, J., Bringmann, G., Parasitol. Res., 83 (1997) 673.
17. François, G., Timperman, G., Eling, W., Aké Assi, L., Holenz, J., Bringmann, G., Antimicrob. Agents Chemother., 41 (1997) 2533.
18. Carhart, R.E., Smith, D.H., Venkataraghavan, R., J. Chem. Inf. Comput. Sci., 25 (1985) 64.
19. Connolly, M.L., J. Appl. Cryst., 16 (1983) 548.
20. Sanner, M.F., Spehner, J.-C., Olson, A.J., Biopolymers, 38 (1996) 305.
21. Pascual-Ahuir, J.L., Silla, E., J. Comp. Chem., 11 (1990) 1047.
22. Rocchia, W., Sridharan, S., Nicholls, A., Alexov, E., Chiabrera, A., Honig, B., J. Comp. Chem., 23 (2002) 128.
23. Todeschini, R., Consonni, V., *Handbook of Molecular Descriptors*, Wiley-VCH, Weinheim, 2000.
24. Arnold, H., In Hahn, T. (Ed.), *International Tables for Crystallography. Volume A: Space-Group Symmetry*, D. Reidel Publishing Company, Dordrecht 1983, pp. 70-75.
25. SYBYL® 6.8 Tripos Inc., 1699 South Hanley Rd., St. Louis, Missouri, 63144, USA, 2000.
26. Ghose, A.K., Viswanadhan, V.N., Wendoloski, J.J., J. Phys. Chem. A, 102 (1998) 3762.
27. Hahn, M., J. Med. Chem., 38, 2080.
28. Heiden, W., Moeckel, G., Brickmann, J., J. Comput.-Aided Mol. Des., 7 (1993), 503.
29. Sheridan, R.P., Miller, M.D., Underwood D.J., Kearsley, S.K., J. Chem. Inf. Comput. Sci., 36, (1996) 128.
30. Brown, R.D., Martin, Y.C., J. Chem. Inf. Comput. Sci., 36 (1996) 572.
31. Martens H., Naes T., Multivariate Calibration, John Wiley bibitem Sons, Chichester, UK, 1989.
32. Baumann, K., Albert, H., von Korff, M., J. Chemom., 16 (2002) 339.
33. Draper, N.R., Smith, H., *Applied Regression Analysis*, John Wiley & Sons, New York, USA, 1981.
34. Baumann, K. Habilitation-Thesis. Universität Würzburg, Germany, 2003.
35. Geisser S., J. Amer. Statist. Assoc., 70 (1975) 320.
36. Baumann, K., von Korff, M., Albert, H., J. Chemom., 16 (2002) 351.
37. Baumann, K., Quant. Struct.-Act. Relat., 21 (2002) 507.
38. Burman, P.A., Biometrika, 76 (1989) 503.
39. Topliss, J.G., Edwards, R.P., J. Med. Chem., 22 (1979) 1238.
40. Klopman, G., Kalos, A.N., J. Comput. Chem., 6 (1985) 492.
41. Faber, N.M., Chemom. Intell. Lab. Sys., 49 (1999) 79.
42. Wu, W., Walczak, B., Massart, D.L ., Heuerding, S., Erni, F., Last, I.R., Prebble, K.A., Chemom. Intell. Lab. Sys., 33 (1996) 35.
43. Kennard, R.W., Stone, L.A., Technometrics, 11 (1969) 137.
44. de Groot, P.J., Postma, G.J., Melssen, W.J., Buydens, L.M.C., Anal. Chim. Acta 392 (1999) 67-75.
45. Jouan-Rimbaud, D., Bouveresse, E., Massart, D.L., de Noord, O.E.,Anal. Chim. Acta 388 (1999) 283-301.
46. De Maesschalk, R., Jouan-Rimbaud, D., Massart, D.L., Chemom. Intell. Lab. Sys., 50 (2000) 1.
47. Ghafourian, T., Dearden, J.C., J. Pharm. Pharmacol., 52 (2000) 603.
48. Abraham, M.H., Chem. Soc. Rev., 22 (1993) 73.
49. Raevsky, O.A., J. Phys. Org. Chem., 10 (1997) 405.
50. Bringmann, G., Messer, K., Wohlfarth, M., Kraus, J., Dumbuya, K., Rückert, M., Anal. Chem., 71 (1999) 2678.
51. Zucchini W., J. Math. Psychol., 44 (2000) 41.
52. Unger, S.H., Hansch, C., J. Med. Chem., 16 (1973) 745.
53. Golbraikh, A., Tropsha, A., J. Mol. Graphics Modell. 20 (2002) 269-276.
54. Manly B.F.J., *Randomization, Bootstrap and Monte Carlo Methods in Biology (2nd Ed.)*, Chapman&Hall/CRC, Boca Raton, 2001.