Published in final edited form as:
J Chem Inf Model. 2009 October ; 49(10): 2312–2322. doi:10.1021/ci9002427.

3D-Pharmacophore Mapping Using 4D-QSAR Analysis for the Cytotoxicity of Lamellarins Against Human Hormone-Dependent T47D Breast Cancer Cells

Poonsiri Thipnate[1,2], **Jianzhong Liu**[3,4], **Supa Hannongbua**[1,2,*], and **A. J. Hopfinger**[3,4]

[1] Department of Chemistry, Faculty of Science, Kasetsart University, Chatuchak, Bangkok 10900, Thailand [2] Center of Nanotechnology KU, Kasetsart University, Chatuchak, Bangkok 10900, Thailand [3] College of Pharmacy, MSC09 5360, 1 University of New Mexico, Albuquerque, New Mexico 87131-000, USA [4] The Chem21 Group, Incorporated, 1780 Wilson Drive, Lake Forest, IL 60045

Abstract

4D-QSAR and 3D-pharmacophore models were built and investigated for the cytotoxicity using a training set of 25 lamellarins against human hormone dependent T47D breast cancer cells. Receptor-independent (RI) 4D-QSAR models were first constructed from the exploration of eight possible receptor binding alignments for the entire training set. Since the training set is small (25 compounds), the generality of the 4D-QSAR paradigm was then exploited to devise a strategy to maximize the extraction of binding information from the training set, and to also permit virtual screening of diverse lamellarin chemistry. 4D-QSAR models were sought for only six of the most potent lamellarins of the training set as well as another subset composed of lamellarins with constrained ranges in molecular weight and lipophilicty. This overall modeling strategy has permitted maximizing 3D-pharmacophore information from this small set of structurally complex lamellarins that can be used to drive future analog synthesis and the selection of alternate scaffolds. Overall, it was found that formation of an intermolecular hydrogen bond and hydrophobic interactions for substituents on the E ring most modulate the cytotoxicity against T47D breast cancer cells. Hydrophobic substitutions on the F-ring can also enhance cytotoxic potency. A complementary high throughput virtual screen to the 3D-pharmacophore models, a 4D-fingerprint QSAR model, was constructed using absolute molecular similarity. This 4D-fingerprint virtual high throughput screen permits a larger range of chemistry diversity to be assayed than the 4D-QSAR models. The optimized 4D-QSAR 3D-pharmacophore model has a LOO cross-correlation value of $xv\text{-}r^2 = 0.947$, while the optimized 4D-fingerprint virtual screening model has a value of $xv\text{-}r^2 = 0.719$. This work reveals that it is possible to develop significant QSAR, 3D-pharmacophore and virtual screening models for a small set of lamellarins showing cytotoxic behavior in breast cancer screens that can guide future drug development based upon lamellarin chemistry.

Keywords

lamellarins; breast cancer; 4D-QSAR; 3D-pharmacophore; virtual screen

*Corresponding author. Supa Hannongbua: Tel: +66 2 562 5555; Fax: +66 2 579 3955; fscisph@ku.ac.th.

INTRODUCTION

In 1985 the first four lamellarins, A-D were isolated from the marine prosobranch mollusk, *Lamellaria* sp. and their structures determined by an X-ray crystallographic and ^{1}H-NMR study.[1] A family of more than 30 lamellarins which consist of three structural groups such as an unsaturated D-ring fused (Figure 1(a)), a saturated D-ring fused (Figure 1(b)), and an unfused central pyrrole ring group (Figure 1(c)) have been isolated and investigated in terms of their biological activity profiles.[2,3] These compounds, especially the fused central pyrrole ring lamellarins, have been found to be cytotoxic to a wide range of cancer cell lines. Lamellarins C and U (Tables 1 and 2) demonstrate potent cytotoxicity against 10 human tumor cell lines (A549, HCT-116, LOX IMVI, MALME-3M, MCF-7, MOLT-4, OVCAR-3, PC-3, SF-295, UO-31)[3] lamellarin D shows potent cytotoxic activity against human prostate cancer cells (DU-145, LNCaP) and leukemia cells (K562),[3] and lamellarins I, K, and L exhibit significant cytotoxicity against P388 and A549 cultured cancer cell lines.[3] Additionally, lamellarin I and D have an effective cytotoxic activity against multidrug resistant reversal (MDR) cell lines by inhibiting P-glycoprotein (P-GP) mediated drug efflux.[4,5] Some lamellarins have also been demonstrated to act on cancer cell mitochondria to induce apoptosis.[6,7] Moreover, lamellarin D is an effective stabilizer of human topoisomerase I-DNA covalent complexes, and, thus, capable of stimulating DNA cleavage.[2,8,9] Based on its biological actions, lamellarin D was identified as a novel lead candidate by Bailly and coworkers.[2,3,6–11]

The first reported study on structure activity relationship (SAR) of lamellarin was done by Ishibashi *et al.* in 2002.[12] It was reported that the hydroxyl groups at position C-8 and C-20 were important for cytotoxicity against a HeLa cell line, while the hydroxyl group at C-14 and two methoxy groups at C-13 and C-21 were not essential for activity. The C5-C6 double bond in the D-ring, or planarity of the chromophore, is necessary for activity.[5,8,12] In more recent findings, Chittchang *et al*. not only substantiated the significant contributions of the C5-C6 olefin moiety, as well as the hydroxyl groups at C8 and C20, but also demonstrated the importance of the C7-hydroxy group for the first time.[13] These findings were also substantiated by carrying out three-dimensional quantitative structure-activity relationship (3D-QSAR) analyses this past year.[14]

Treatment of the lamellarins data set is representative of a class of real-world problems in drug discovery; namely how to optimize the extraction of SAR information for, in turn, optimizing lead development efforts from a small number of structurally complex, hard to synthesize compounds that have been tested and observed to exhibit a wide-range of endpoint activity. Modeling such small data sets can be criticized on the basis of too little data to generate reliable, and useful, results to drive lead optimization. Yet doing nothing with the information resident in such data sets obviously contributes nothing to streamlining lead development efforts. A key to resolving this dilemma may reside in the type and level of sophistication of the modeling employed. A high-level modeling approach wherein detailed structural, thermodynamic and electronic information about each complex compound of the data set may negate some of the drawbacks to the small size of the data set. In a sense, ‘quality’ is used to compensate for ‘quantity’.

Multiple complementary applications of the 4D-QSAR paradigm[15] may be a good way to extend our knowledge and understanding the structure-activity relationships of lamellarins using this ‘quality’ for ‘quantity’ argument. The fourth ‘dimension’ of the 4D-QSAR paradigm is ensemble sampling the spatial features of the members of a training set.[15] This sampling process, in turn, allows the construction of optimized dynamic spatial QSAR models, in the form of 3D pharmacophores, which are dependent on conformation, alignment, and pharmacophore-grouping.[16] This method has been proven both useful and reliable for the construction of quantitative 3D pharmacophore models, especially for sets of flexible ligand analogues when the geometry of the corresponding receptor is not known.[15–17]

Complementary to building 4D-QSAR models that embed 3D-pharmacophores is the construction of high-throughput 4D-fingerprint models for virtual screening. The 4D-fingerprints can be derived independent of any molecular alignment, and are based upon an inductive approach to establish 4D molecular similarity measures across any collection of chemical compounds.[18] The 4D-QSAR paradigm has been successfully applied for a variety of chemical classes and biological endpoints including glucose analogs, flavonoid analogs, propofol analogs, the AHPBA and THP inhibitors of HIV-1 protease,[18] human serum albumin (HSA),[19,20] a local lymph node assay (LLNA) data base,[21,22] skin penetration enhancers,[23] and HIV-1 integrase inhibitors.[24]

The search for structure-activity relationships and/or pharmacophores for natural, and synthetic, lamellarins screened for cytotoxic activity against 11 cancer cell lines is on-going. However, the prominent activities observed in a few of the lamellarins screened against human hormone-dependent T47D breast cancer cells seemed to us to be best explored, and the corresponding SAR delineated and exploited, by using the 4D-QSAR methodology for the reasons cited above.

MATERIALS AND METHODS

Lamellarin data set and cytotoxic activity

Twenty six lamellarins were analyzed in this work. The chemical structures of all 26 lamellarins are given in Table 1 and 2. These compounds were synthesized and purified by Ploypradith *et al.*[25] The cytotoxic activity (−logIC_{50}) against human hormone-dependent T47D breast cancer cells have been recently reported and are included as parts of Tables 1 and 2.[13]

Receptor-independent (RI) 4D-QSAR analysis applied to the lamellarin data set

Since the geometry of the receptor is not available in this study, the receptor-independent form of 4D-QSAR analysis, referred to as RI-4D-QSAR, has been employed. The ten operational steps in RI-4D-QSAR have been presented in detail previously,[15] and also given in the 4D-QSAR software version 3.0 *User Guide.*[26] Therefore, these 10 steps of RI-4D-QSAR analysis are only summarized here as follows:

Step 1—An initial 3D structure of each lamellarin was constructed in the neutral form using the HyperChem 7.5 software.[27] Partial atomic charges were computed using the semiempirical AM1 method. Each structure was then minimized with no geometric constraint. These energy-minimized structures were used as the initial structures in the conformational ensemble sampling of step 3.

Step 2—Atoms of each molecule were classified into seven types of interaction pharmacophore elements (IPEs). Each type is represented by different number code from 0 to 6 as defined in Table 3.

Step 3—Molecular dynamics simulations (MDS) was used to sample the conformational states available to each analogue, and to generate its corresponding conformational ensemble profile (CEP). The MDSs were done using the MOLSIM package[28] and MM2 force field.[29,30] The temperature for the MDS is set at 300 K with a simulation sampling time of 40 ps with intervals of 0.001 ps for a total sampling of 40000 conformations of each lamellarin compound. The atomic coordinates of each conformation and its intramolecular energy sampled during the MDS were recorded every 0.02 ps for a total of 2000 “frames”, or steps, in constructing the CEP of each compound.

Step 4—The set of three-ordered atoms in trial alignments are defined in Table 4. In this study eight alignments were explored across the overall lamellarin core structure.

Step 5—Each conformation of a compound from its CEP was aligned in the grid cell lattice using the invariant coordinates of the three-ordered atom alignment. In this study, the size of the cubic grid cells of the lattice are 1 $Å^3$, and the overall grid cell lattice size was chosen to fully enclose each compound of the training set. The normalized occupancy of each grid cell by each IPE atom type over the CEP for a given alignment forms a unique set of QSAR descriptors referred to as grid cell occupancy descriptors, GCODs. The GCOD descriptors were computed, and used as the trial descriptor pool in 4D-QSAR analysis. Non-GCOD descriptors of the training set compounds can also be included in the trial basis set (descriptor pool). In this particular study the logarithm of the 1-octanol/water partition coefficient (log *P*) and the compound’s molecular weight (MW) were selectively added to the trial basis set descriptors in some of the model building studies. The log *P* and MW values the training set compounds are reported in Table 1.

Step 6—A 4D-QSAR analysis generates an enormous number of trial QSAR descriptors, GCODs, because of the large number of grid cells and the seven IPEs. Partial least squares (PLS) regression analysis[31] is used to perform a data reduction analysis between the observed dependent variable measures and the corresponding set of GCOD values.

Step 7—The most highly weighted PLS GCOD descriptors (currently the top 200), generated in step 6, are used to form the trial descriptor pool for genetic algorithm (GA) model optimization. The specific GA currently used in the 4D-QSAR software is modification of the genetic function approximation (GFA).[32] The GFA optimization is initiated using N (currently 300) randomly generated 4D-QSAR models. Mutation probability over the crossover optimization cycle is set at 10%. The smoothing factor, a GFA operations variable, controls the number of independent variables in the QSAR models, is varied in order to determine the optimal number of descriptors for the 4D-QSAR models. The diagnostic measures used to analyze the resultant 4D-QSAR models generated by the GFA include (i) descriptor usage as a function of crossover operation, (ii) linear cross correlation among descriptors and/or dependent variables (biological activity measures), (iii) number of significant and independent 4D-QSAR models, and (iv) indices of model significance including the correlation coefficient, r^2, leave one-out, LOO, cross-validation correlation coefficient, xv-r^2, and Friedman’s lack of fit (LOF).[33] In this particular 4D-QSAR application, the alignment similarity comparisons were limited to models having same number GCODs.

Step 8—Steps 4–7 are repeated until all trial alignments are included in the 4D-QSAR analyses.

Step 9—The inspection and evaluation of the population of models are obtained from the set of trial alignments in this step. The goal of this step is to identify the best and distinct set of 4D-QSAR models which is referred to as the manifold model of the analysis.

Step 10—The “active” conformation of each compound is hypothesized at this step. This conformer is achieved by identifying all conformer states sampled for each compound that are within ΔE of the global minimum energy conformation of the CEP. Currently, ΔE is set at 2 kcal/mol. Each member of the resultant set of energy-filtered conformations is then individually evaluated in the best 4D-QSAR model. The conformation within 2 kcal/mol of the apparent global minimum that predicts the highest activity in the best 4D-QSAR model is defined as the active conformation.

4D-fingerprint virtual screening analysis applied to the lamellarin data set

The theory and corresponding methodology of the universal 4D-fingerprints for constructing the main distance-dependent matrix (MDDM) and computing corresponding eigenvalues for each matrix, using 4D molecular similarity (MS), have been presented in detail in previous work.[18,34] The types of atoms composing a molecule are currently defined as the IPEs shown in Table 3. A unique MDDM is constructed for each of the eight distinct and identical IPE pairs. The elements of the MDDM are defined as following:

$$E_{(\nu,d_{ij})}=e^{(-\nu\langle d_{ij}\rangle}\tag{1}$$

The "universal constant *(ν)*" in eq. 1, which is equal to 0.25,[34] has been selected such that the difference in the sum of eigenvalues for any two arbitrary compounds with the same number, *n*, of a particular IPE type, *m*, is maximized. The term $\langle d_{ij}\rangle$ is average distance between the atom pair ij of IPE type u and v.

$$\langle d_{ij}\rangle=\sum\nolimits_{\mathrm{k}}\mathrm{d}_{ij}(\mathrm{k})\,\mathrm{p}(\mathrm{k})\tag{2}$$

where p(k) refer to the thermodynamic probability of the kth conformer state sampled in the assessment of conformational flexibility, and d_{ij}(k) is the corresponding distance between atom pair *i* and *j* of IPE types *u* and *v* for the kth conformer state. Then, similarity eigenvalues are derived by the diagonalization of the MDDM. For same-term IPE pairs, such as *u* = *v*, the MDDM are square upper/lower triangular. These matrices can be directly diagonalized. The resulting eigenvalues determined from the MDDM are normalized and ranked in numerically descending order in their eigenvector representation. The *n*th normalized eigenvalue for IPE type *m* of a compound $\alpha, \in_{mn}(\alpha)$, can be obtained by scaling the non-normalized eigenvalue $\in_{\mathrm{mn}}'(\alpha)$ relative to the rank of its MDDM.

$$\in_{\mathrm{mn}}(\alpha)=\in_{\mathrm{mn}}(\alpha)'/\mathrm{rank}(\alpha)_m\tag{3}$$

Determination of eigenvalues of the MDDM for *u* ≠ *v*, the so-called cross-terms for IPE pairs that are not the same, requires a different strategy since these matrices may, or may not, be square. In the case of rectangular MDDM (*u* ≠ *v*), the following square MDDM are constructed

$$\mathrm{MDDM}(u,u)=\mathrm{MDDM}(n_u,n_v)\times\mathrm{MDDM}(n_u,n_v)^{\mathrm{T}}\tag{4}$$

$$\mathrm{MDDM}(v,v)=\mathrm{MDDM}(n_v,n_u)\times\mathrm{MDDM}(n_v,n_u)^{\mathrm{T}}\tag{5}$$

For MDDM(*u*,*u*) and MDDM(*v*,*v*) have the same rank and trace, both have the same set of eigenvalues. Hence, for each pair of IPE (*u* ≠ *v*)

$$\in(\alpha)_{u,v}=\{[\in(\alpha)]_{\mathrm{MDDM}(u,u)}\}^{1/2} \tag{6}$$

According to all possible combinations of the eight IPE types, there are 36 possible molecular similarity eigenvectors from the MDDM for each compound *α*. The similarity eigenvectors have been calculated for the set of compounds, the estimation of molecular similarity for a pair of compounds *α* and β begins with a definition for molecular dissimilarity, given by

$$\mathrm{D}_{\alpha\beta}=\sum_{\mathrm{i}}|\in(\alpha)_i-\in(\beta)_i| \tag{7}$$

where $i = i^{\text{th}}$ eigenvalue in the corresponding eigenvetor of a specific IPE pair. Molecular similarity is then defined as

$$S_{\alpha\beta}=(1-D_{\alpha\beta})(1-\phi) \tag{8}$$

where $\boldsymbol{\phi}$ = |rank(α) − rank(β)|/(rank(α) + rank(β). The rank of the matrices is essentially the number of atoms of a specific IPE type present. The $\boldsymbol{\phi}$ term in eq. 8 serves to reincorporate molecular size information. Similar to the measure for dissimilarity, the similarity measure is a value between 1 and 0, where a value closer to 1 refers to compounds that are more similar, and closer to 0 refers to compounds that are more dissimilar.

The descriptor set for *α* consists of all of the eigenvalues of all of the eigenvectors derived from all of the MDDM for compound *α*. In this work, a threshold cutoff value which equal to 0.002 is applied, and those normalized eigenvalues below the threshold value are disregarded.

The maximum number of significant eigenvalues specific to that data set for a particular compound and a particular IPE type, *m*, is determined, $\epsilon_{m,max}$. All the eigenvectors for IPE type, *m*, for each molecule across lamellarin data set are then assigned $\epsilon_{m,\mathrm{max}}$ eigenvalues for IPE type *m*. Eigenvectors that otherwise contain less than $\epsilon_{m,\mathrm{max}}$ elements have the "missing" eigenvalues set to zero.

The total set of descriptors, ϵ_{total}, for a compound in the data set will be the sum of the 36 eigenvalues of $\epsilon_{m,\mathrm{max}}$ length which can be a large number for the data set in this work.

Finally, the sets of 4D-fingerprints across each of the molecules of the training set form the trial descriptor pool to build the 4D fingerprint virtual screens. The building procedure of these virtual screens is identical to that employed in constructing the RI-4D-QSAR models. That is, steps 6 through 9 given above for the RI-4D-QSAR methodology are used.

RESULTS AND DISCUSSION

Receptor-independent (RI)-4D-QSAR analysis

Optimized RI-4D-QSAR models were constructed for each of the eight trial alignments listed in Table 4. Alignments 1, 2, 4, and 7 contain atoms from two rings (A and B), (B and C), (C and D), and (C and F), respectively. Alignment 5 and 6 only contain atoms from ring E and ring F, respectively. Only two alignments, 3 and 8, distribute the three-ordered atoms across three rings namely rings A, B, and C for alignment 3 and rings A, E, and F for alignment 8. The r^2 and xv-r^2 values from the best corresponding five-term RI-4D-QSAR models of each alignment are given in Table 4. Five terms in a model corresponds to the largest model that can be built by allowing at least 5 observations [compounds] per model-term for the training set. The optimized 5-term model represents an initial upper-bound exploration of the type, and corresponding quality, of an RI-4D-QSAR model that can be expected from the structure-activity data set. Alignment 1 yields the poorest fits with r^2 = 0.964 and xv-r^2 = 0.929. The differences among r^2 and xv-r^2 of the remaining alignments are quite small, or the alignment of lamellarin is not significant to the 4D-QSAR model. However, based on the greater r^2 (0.999) and xv-r^2 (0.998), alignment 3 appears to be the best alignment for 4D-QSAR analysis of lamellarin data set.

The optimum number of descriptors in a model is determined by monitoring when xv-r^2 becomes effectively constant, or decreases, with increasing model size. Figure 2 is a plot of the number of descriptor terms in an optimized alignment 3 model versus the corresponding r^2 and xv-r^2. An inspection of Figure 2 reveals that the maximum number of descriptor terms in the RI-4D-QSAR model providing additional fit to the training set data is three. There is no meaningfully enhanced model fitting by including more than three descriptor terms. Thus, the optimized RI-4D-QSAR model for the 25 lamellarins generated from alignment 3 is given by eq. 9. Among top-ten 4D-QSAR models obtained from alignment 3, eq 9 (or model 3) is the best 4D-QSAR model since it has the highest xv-r^2, and all other top-ten models are basically the same as model 3. This commonality to model 3 by the other top-ten models can be inferred from Table 5 by the high cross-correlations of their residuals of fit to those of eq. 9;

$$\begin{aligned} -\log IC_{50} = 5.14 + 16.90GC1\,(-5, 6, 2, np) - 56.33GC2\,(-3, 4, -5, any) \\ +64.62GC3\,(-1, 5, 0, np) \\ n=25,\ r^2=0.971,\ xv - r^2=0.947 \end{aligned} \tag{9}$$

GC*i* (*x*, *y*, *z*, *X*) is the *i*th GCOD descriptor term located at (*x*, *y*, *z*) in the reference grid cell and alignment space, and having the *X* type IPE as defined in Table 3. Figure 3 is a plot of the predicted, using eq. 9, versus actual $-\log IC_{50}$ values. All of the predicted $-\log IC_{50}$ values are within ± 1 log unit of the corresponding observed values, and there are no outliers.

Two GCODS (GC1 and GC3) of eq. 9 correspond to pharmacophore sites of nonpolar atom occupancy, both of which increase potency. These two GCODS both have positive regression coefficients with values of 16.90 and 64.62, respectively. GCOD GC2, having an ‘any’ IPE type, has a negative regression coefficient with value of −56.33. Consequently occupancy of the GC2 site by any type of atom will lead to a decrease in the potency of anti-breast cancer activity of the corresponding lamellarin. From an analysis of eq. 9 it is found that the any IPE type at (−3, 4, −5) has about three times more of a negative effect upon $-\log IC_{50}$ than the positive effect of the nonpolar IPE type at (−5, 6, 2), and about the same, but opposite effect on $-\log IC_{50}$ as the nonpolar IPE type at (−1, 5, 0). None of the best models from GFA model optimization contain GCOD descriptors which deal with specific atom-atom interactions like hydrogen bonding.

In order to further search for pharmacophore sites which are specifically associated with lamellarins exhibiting high cytotoxic activity, an additional RI-4D-QSAR analysis was carried out. The training set of this study was limited to the six lamellarins (D, M, N, X, ε, and Dehydrolamellarin J of Table 2) that have the highest $-\log IC_{50}$ values, and are not redundant in their structural features. The RI-4D-QSAR models were constructed and optimized by using the same methodology and alignment used to build eq. 9. The best RI-4D-QSAR model from this small high activity data set of lamellarins is given by eq. 10.

$$-\log IC_{50}=10.31-50.52GC1\,(-1,1,-6,\text{any})+1.58GC2\,(-1,4,-6,\text{np})\qquad n=6,\ r^2=0.997,\ xv-r^2=0.984 \tag{10}$$

The regression coefficients of the descriptors of eq. 10 suggest placing any type of atom at (−1, 1, −6) has about 30 times more negative effect on $-\log IC_{50}$ than the positive gain by locating a nonpolar atom or group at (−1, 4, −6). Certainly eq. 10 is, or borders upon, being an over-fit model. However, eq. 10 and its 3D-pharmacophore are only used as adjuncts to eq. 9 and its 3D-pharmacophore. That is, eq. 10 is being used to provide a higher-resolution view of the SAR features most characteristic of the high activity lamellarins of the training set. Equation 9 and its 3D-pharmacophore are used outside that context.

The 3D-pharmacophores defined by eqs. 9 and 10 are shown in Figures 4(a) and 4(b), respectively. The reference structure superimposed on each of the 3D-pharmacophores in these two figures is the predicted active conformation of the most active compound (lamellarin D) using eq. 9. The red spheres in Figures 4(a) and 4(b) represent those GCOD descriptor terms which have negative regression coefficients. Correspondingly, the blue spheres delineate GCOD descriptors having positive regression coefficients in the corresponding best RI-4D-QSAR equation. From an inspection of Figure 4(a), a red sphere near the R_2 and R_3 groups specifies a pharmacophore site where occupancy by any type of atom, or group, decreases potency since the corresponding regression coefficient −56.33. Two blue spheres are found near R_4 and R_5 suggesting that substitution of nonpolar groups to occupy one, or both, sites is conducive to increasing the cytotoxic activity of the lamellarins.

The 3D-pharmacophore of the high activity model, eq. 10, is represented by one red sphere (GCOD) located around R_1 and R_2, and a blue sphere (GCOD) positioned near R_2 and R_3. The most active compounds of the potent lamellarins seemingly achieve most of their additional $-\log IC_{50}$ potency, as compared to the less potent lamellarins, by not having any atoms or groups at (−1, 1, −6) in contrast to increasing occupancy by nonpolar atoms or groups at the GCOD located at (−1, 4, −6). The 30:1 ratio of not occupying the GCOD at (−1, 1, −6) as compared to having a nonpolar atom or group at (−1, 4, −6) is consistent with the relative binding energy contributions of an intermolecular hydrogen bond involving the OH near (−1, 1, −6) as compared to a hydrophobic binding effect due to the methyl of the methoxy group near (−1, 4, −6) as is shown in Figure 4(b).

Overall, the highly active compounds are seemingly distinguished from one another in eq. 10 by their ability to form an intermolecular hydrogen bond where the hydrogen bond acceptor atom in the receptor is expected to be near (−1, 1, −6). Some additional increase in $-\log IC_{50}$ can also be realized by having a hydrophobic substituent group of the ligand occupying the (−1, 4, −6) site. The two GCODs of eq. 10 may be a higher resolution representation of the single GC2 (−3, 4, −5, any) GCOD found in eq. 9.

In order to evaluate the possible roles of ligand molecular weight, MW on cytotoxic potency, $-\log IC_{50}$, this property were included as part of the trial basis set of descriptors in a GFA model optimization study. Unfortunately, no GFA model optimization could be realized. An inspection of the MW value of the training set compounds revealed that three lamellarins (lam K-triacetate, lam χ-triacetate, and lam U-diacetate) have very high MWs relative to the other training set compounds. These three lamellarins were removed to form a revised training set. Two lamellarins (lam F and K) were defined as a test set. GFA model building and optimization repeated for this 21 compound training set in the same manner as employed in developing eqs. 9 and 10. Ten best models were determined from the GFA optimization, and the residuals of fit cross-correlations between each pair of these models are given in Table 6. All pairs of the top-ten models have residuals of fit highly correlated to one another, with a value of at least

0.70, indicating these 10 models are all very nearly the same model. Therefore, the best of the ten models was selected as the preferred RI-4D-QSAR model for this training set, and is given by eq. 11.

$$\begin{aligned}-\log(\mathrm{IC}_{50}) = {} & 10.31 - 4.77\mathrm{GC1}(-2,1,-6,\mathrm{np}) - 33.91\,\mathrm{GC2}(-3,4,-5,\mathrm{any}) \\ & -8.12\mathrm{GC3}(3,3,2,\mathrm{np}) \\ & \mathrm{n}=12,\ r^2=0.935,\ xv-r^2=0.890\end{aligned} \tag{11}$$

Figure 5 is a plot of the observed versus the predicted −logIC$_{50}$ values determined from using eq. 11. The 3D-pharmacophore embedded in the RI-4D-QSAR model given by eq. 11 is shown in Figure 6 with lamellarin D again the reference compound. All three GCOD descriptors of eq. 11 correspond to pharmacophore sites where an increasing occupancy decreases activity. One pharmacophore site, (−3, 4, 5, any) from eq. 11, is identical to a site from eq. 9, while the pharmacophore site at (−2, 1,−6, np) from eq. 11 is very close to the pharmacophore site of eq. 10 located at (−1, 1,−6, any) as can be seen by comparing Figure 6 to Figures 4(a) and 4(b). The third pharmacophore site of eq. 11 located at (3, 3, 2), which predicts the occupany of nonpolar groups to decrease −logIC$_{50}$, is unique to this model as compared to eqs. 9 and 10. This new GCOD descriptor term of eq. 11 and the decrease in r^2 and xv-r^2 may be an indication of a significant pharmacophore-site dependence on one, or more, of the four lamellarins eliminated from the training set used to build eq. 11 and its corresponding 3D-pharmacophore.

An attempt was made to further explore if log *P* plays a role in the structure-activity relationship of the lamellarin training set by forcing an overfitting in the GFA model building and optimization process. The log *P* descriptor was the only non-GCOD descriptor added to the trial basis set (descriptor pool) at step 5 of 4D-QSAR methodology. Overfit RI-4D-QSAR models were permitted under the same methodology, same alignment, and for all lamellarins in training set as used to develop eq. 9. None of the 10 most significant overfit 4-term or 5-term RI-4D-QSAR models contained a log *P* descriptor term. Therefore, it was concluded that molecular lipophilicity is not a major contributing factor in the specification of the cytotoxic activity for the lamellarins studied in this analysis.

The predicted −logIC$_{50}$ of lamellarin F calculated by using eq. 11 is 5.74. This value are very close to actual −logIC$_{50}$ value of 5.34. The RI-4D-QSAR model obtained by removed out high MW compounds showed a good predict the activity of lamellarin F. Lamellarin K was synthesized and tested after the 4D-QSAR models reported in this paper were constructed. However, lamellarin K has an unexpected high activity [−logIC$_{50}$ = 7.04] for the saturated D-ring series of compounds. Hence, it was thought important to see if this high activity could be predicted by the 4D-QSAR models, or if this saturated D-ring analog had features outside those captured by the models. The predicted −logIC$_{50}$ values of lamellarin K obtained from eqs. 9, 10 and 11 are 5.33, 9.77 and 6.00, respectively. Thus, the 4D-QSAR models developed in this study cannot well-predict the activity of lamellarin K, but their composite set of predicted activities bracket around the observed activity. Moreover, while the individual models did not adequately predict the experimental endpoint, it is to be noted that the average of these three predicted values is 7.03 which is a value very close to the experimental −logIC$_{50}$ value of 7.04.

Lamellarin K has a unique three hydroxyl substituent pattern at R1, R4 and R7. However, other analogs in Table 1 have three hydroxyl substituents, and some analogs without hydroxyl substutuents are more active than those with three hydroxyls, for example, compare lamellarin χ triacetate (5.54) to lamellarin E (5.28) in Table 1. All of the best 4D-QSAR models, eqs. 9, 10 and 11 are rich in GCOD terms involving nonpolar IPE types. Polar and/or hydrogen bonding capabilities from hydroxyl groups are not explicitly present in the descriptor terms of the 4D-QSAR models. All of these observations, in composite, suggest that the unique hydroxyl substituent pattern of lamellarin K make it the ‘magic bullet’ in terms of high inhibition potency relative to the other saturated D-ring analogs of Table 1.

4D-fingerprint virtual screens

4D-fingerprint virtual high throughput screens permit a larger range of chemistry diversity to be assayed more quickly than do RI-4D-QSAR models. In this study 4D-fingerprints models were generated using all 25 of the lamellarins in the training set. Lamellarin K was used as a modest means to validate 4D-fingerprints model as well as RI-4D-QSAR models. Two types of 4D-fingerprints can be constructed: those 4D-fingerprints explicitly dependent upon a particular alignment, and absolute 4D-fingerprints which are alignment independent.[35] Absolute 4D-fingerprints were used in this analysis to maximize the range of lamellarin chemical diversity that could be reasonably screened. That is, a 4D-fingerprint screening model built independent of alignment is more general than its corresponding alignment-dependent screen, but at the cost of being somewhat less significant in its fit to the training set data.

The absolute 4D-fingerprints were derived for each of the 25 training set lamellarins using the modeling methodology given above in the *Methods* section. These 4D-fingerprints formed the trial basis set for model building. No non-4D-fingerprints were added to this trial descriptor pool. Model building and optimization in deriving the 4D-fingerprint QSAR equations, which are the high-throughput virtual screens, was carried in the identical fashion used to build the RI-4D-QSAR models.

Figure 7 is a plot of number of descriptor terms in a 4D-fingerprint model versus r^2 and xv-r^2. The xv-r^2 of the 4D-fingerprints of the 4- and 5-term models are very nearly the same, and xv-r^2 behaves in something of an erratic fashion for models having 5, or more, descriptor terms. The optimized 4-descriptor term virtual screening model appears, on the basis of xv-r^2, to capture maximum fitting to the training set data without overfitting. Thus, the 4-term QSAR model given by eq. 12 was selected as the preferred absolute 4D-fingerprint virtual screen. Equation 12 is the best 4-term model from the top-ten 4-term models derived in the GFA optimization. Table 7 shows the linear cross-correlation matrix of the residual of fit for the top-ten 4-term models. This table reveals that all pairs of models have highly correlated residuals of fit, greater than 0.85, to one another. Thus, eq. 12 represents the best and only distinct fit to the training set data using absolute 4D-fingerprints.

$$\begin{aligned} -\log \mathrm{IC}_{50} = & -7.39 - 452.65 \in_7(\mathrm{any,np}) + 1357.10 \in_{11}(\mathrm{any,hs}) + 9.58 \in_3(\mathrm{p}^+,\mathrm{aro}) \\ & -94.31 \in_2(\mathrm{np,hs}) \\ & n=25,\ r^2=0.831,\ xv-r^2=0.719 \end{aligned} \tag{12}$$

For reference in defining the 4D-fingerprints, $\in_7$(any,np) represents the seventh largest eigenvalue from the MDDM of the IPEs u = (any) and v = (np) molecular similarity vector capturing all pairs of atoms in each lamellarin assigned IPEs of any and nonpolar, respectively.

The relative significance and weight of each 4D-fingerprint descriptor term in eq. 12 was measured in terms of its frequency of use in the GFA model optimization process. The idea in monitoring frequency of use is that the more significant is a descriptor to establishing a fit to the training set data, the more often it will be used in the repetitive GFA optimization process. The frequencies of descriptor usage during GFA optimization are shown in Table 8. An inspection of Table 8 indicates that $\in_{11}$(any,hs) and $\in_7$(any,np) are the first and second important features governing the SAR of lamellarin cytotoxicity potency, respectively. Increased potency of the lamellarins arises from increasing the values of $\in_{11}$(any,hs) and/or $\in_3$(p^+,aro), while a decrease in lamellarin cytotoxicity accompanies an increase in the values of the $\in_7$(any,np) and $\in_2$(np,hs) 4D-fingerprints. Figure 8 is a plot of $-\log IC_{50}$ values predicted using eq. 12 versus the corresponding observed $-\log IC_{50}$ values.

The predicted activity of lamellarin K, the test compound, using eq. 12 is 7.33 which differs from the observed activity of 7.04 by only 0.29 log unit. Additional $-\log IC_{50}$ predictions using eq. 12 were made for a small virtual library of eight lamellarin derivatives, see Table 9, generated by making substitutent changes at R_1-R_5. These results indicate that the 4D-fingerprint model is responsive to predicting $-\log IC_{50}$ values over a wide $-\log IC_{50}$ potency range from nearly inactive values for ML6 to very potent activities for lamellarins ML4, ML5, and ML7. Equation 12 also has captured the SAR that both the number and positioning of –OH on the E-ring is a critical factor to potency. In general, more hydroxyls on the ring are better. But the importance of hydroxyl positioning, particularly at R_3 is dramatically shown for ML5, the most active analog in Table 10 [10.05], as compared ML6, the least active analog [3.03] which differs only from ML5 by having no hydroxyl at R_3.

Comparison of the 4D-fingerprints QSAR virtual screening model to the RI-4D-QSAR models

The RI-4D-QSAR model given by eq. 9 with three descriptor terms is a more significant fit to the training set data (xv-r^2 = 0.947 and r^2 = 0.971) than the four descriptor 4D-fingerprint model given by eq. 12 (xv-r^2 = 0.719 and r^2 = 0.831). Presumably the inclusion of alignment information in eq. 9 provides this boost in the overall fitting quality of this model as compared to eq. 12. But eq. 12 in not being dependent on alignment correspondingly permits a wider range of variations lamellarin chemistry to be considered. Table 10 is the linear correlation matrix of the residuals of fit of eq. 9, the RI-4D-QSAR, to eq. 12, the absolute 4D-fingerprint virtual screen, as well as correlations of both models to the observed $-\log IC_{50}$ cytotoxicity values. The correlation coefficient of 0.797 between the residuals of fit for eqs. 9 and 12 indicates that these two models are basically the same, but eq. 9, owing to inclusion of alignment, fits the training set better, overall, than eq. 12.

Comparison of the predicted inhibition potencies from the 4D-fingerprints QSAR virtual screening model to the *(RI)*-4D-QSAR models was also investigated using ML5 and ML6, the most and the least potent compounds given in Table 9. The predicted $-\log IC_{50}$ values of ML5 obtained from the *(RI)*-4D-QSAR models by eqs. 9 and 11 are 5.39 and 9.20, respectively, and the $-\log IC_{50}$ values of ML6 obtained from the two equations are 7.44 and 10.23, respectively. It was found that there is an agreement in prediction only for ML5 between the *(RI)*-4D-QSAR model (9.20 by eq. 11) and the 4D-fingerprint QSAR model (10.05 by eq. 12).

CONCLUSION

This work puts forth a ‘quality in place of quantity’ strategy to handle small data sets composed of structurally complex, hard to synthesize compounds that can exhibit a wide-range in endpoint activity. A high-level modeling approach providing detailed structural, thermodynamic and electronic information about each complex compound of the data set is used to negate the lack-of-data drawbacks to the small size of the data set. In this study the flexibility, yet high-level of modeling sophistication of the 4D-QSAR paradigm is used to explore different subpopulations of the data set in extracting the maximum SAR information from the data set in terms of a pseudo consensus RI-4D-QSAR model and its corresponding 3D-pharmacophore. The consensus aspect to the RI-4D-QSAR modeling arises from the fact that the same methodology and parameters, including alignment, can be used in any manner across any subpopulations of the data set. As such, all resulting models are not only directly comparable, but to an appreciable extent can be combined to elucidate a high-resolution 3D-pharmacophore. In addition, the 4D-fingerprint formulation of the 4D-QSAR paradigm permits alternate model generation, particularly useful in virtual screening. Still, the 4D-fingerprint models are once again directly comparable to the RI-4D-QSAR models so as to exact additional information from the data set, as well as to evaluate the self-consistency across all the models constructed.

The consensus set of 4D-QSAR models expressed by eqs. 9–12, suggest that the ability to form a ligand-receptor intermolecular hydrogen bond and hydrophobic interactions for substituents on the E ring most modulate the cytotoxicity against T47D breast cancer cells. The optimization of this intermolecular hydrogen bond, and, to a lesser extent, the hydrophobic interactions, are coupled to the alignment freedom of a lamellarin owing, in turn, to other possible substitutions across the molecule and their possible interactions with sites on the receptor.

Hydrophobic substitutions on the F-ring can also enhance cytotoxic potency, but given that the 3D-pharmacophore sites for these interactions arise for the entire data set, and not the restricted high activity data subset, would indicate these are likely minor binding pharmacophore sites. Attempts to force the lipophilicity of the entire lamellarin into a 4D-QSAR model were unsuccessful. Thus, the finding of 3D-pharmacophore sites, where occupancy by nonpolar atoms and/or groups can modulate activity, likely reflect specific interactions at these sites, and not global lipophilic features of the lamellarins.

Lamellarin K, synthesized and tested after the modeling studies reported here were carried out, likely has its very high activity relative to other saturated D-ring analogs because of its unique three hydroxyl group substituent pattern. The average predicted $-\log IC_{50}$ value developed in this study sufficient predicts the activity of lamellarin K. This suggests that in order to get a high-resolution 4D-QSAR model to distinguish some substituent patterns from others for the saturated D-ring lamellarins analogs, more lamellarins data set is required.

The 4D-fingerprint virtual screening model, eq. 12, is highly consistent with the general RI-4D-QSAR model given by eq. 9. Consequently, eq. 12 can be used to rapidly screen prospective compounds without concern for alignment, but with the expectation that the 3D-pharmacophore of eq. 9 will be relevant to helping understand findings from virtual screenings. A good test to evaluate how much SAR information is actually captured in eq. 12 as a virtual screening tool, given it is based upon this relatively small training set of lamellarins would be to make and test ML5 and ML 6 of Table 9. These two compounds are predicted to differ by seven orders of magnitude in $-\log IC_{50}$ values, yet they differ by at their respective R_3 substituents. A large difference in measured $-\log IC_{50}$ values would help to validate eq. 12, while a small difference would suggest that eq. 12 has very limited resolution in correctly explaining small structural differences in the lamellarins.

Acknowledgments

The Thailand Research Fund (RTA5080005) and the Commission on Higher Education (CHE) are gratefully acknowledged for research grant. We are grateful to Faculty of Science and the Graduate School of Kasetsart University, Phetchaburi Rajabhat University, the Chulabhorn Research Institute (CRI), the Chulabhorn Graduate Institute (CGI), and the College of Pharmacy, University of New Mexico for research supporting. The National Center of Excellent in Petroleum, Petrochemical Technology and Advanced Materials under Ministry of Education, KU Research and Development Institute, Laboratory for Computational and Applied Chemistry, NANOTEC Center of Excellence at KU, and the Large Scale Research Laboratory of the National Electronics and Computer Technology are also acknowledged for computing and research facilities. This work was also supported by the National Institutes of Health, USA through the NIH Roadmap for Medical Research, Grant 1 R21 GM075775. Information on Novel Preclinical Tools for Predictive ADME-Toxicology can be found at http://grants.nih.gov/grants/guide/rfa-files/RFA-RM-04-023.html. Links to nine initiatives are found here http://nihroadmap.nih.gov/initiatives.asp. This work was also supported, in part by The Proter & Gamble Company. Resources of the Laboratory of Molecular Modeling and Design at UNM, and at The Chem21 Group, Inc., were used in performing these studies. We also thank the reviewers for the valuable comments.

REFERENCES AND NOTES

1. Andersen RJ, John Faulkner D, Cun-heng H, Van Duyne GD, Clardy J. Metabolites of the marine prosobranch mollusc Lamellaria sp. J Am Chem Soc 1985;107:5492–5495.
2. Bailly C. Lamellarins, from A to Z: A family of anticancer marine pyrrole alkaloids. Curr Med Chem - Anti-Cancer Agents 2004;4:363–378.
3. Fan H, Peng J, Hamann MT, Hu JF. Lamellarins and related pyrrole-derived alkaloids from marine organisms. Chem Rev 2008;108:264–287. [PubMed: 18095718]
4. Vanhuyse M, Kluza J, Tardy C, Otero G, Cuevas C, Bailly C, Lansiaux A. Lamellarin D: A novel pro-apoptotic agent from marine origin insensitive to P-glycoprotein-mediated drug efflux. Cancer Lett 2005;221:165–175. [PubMed: 15808402]
5. Quesada AR, Garcia Gravalos MD, Fernandez Puentes JL. Polyaromatic alkaloids from marine invertebrates as cytotoxic compounds and inhibitors of multidrug resistance caused by P-glycoprotein. Br J Cancer 1996;74:677–682. [PubMed: 8795567]
6. Gallego MA, Ballot C, Kluza J, Hajji N, Martoriati A, Castera L, Cuevas C, Formstecher P, Joseph B, Kroemer G, Bailly C, Marchetti P. Overcoming chemoresistance of non-small cell lung carcinoma through restoration of an AIF-dependent apoptotic pathway. Oncogene 2008;27:1981–1992. [PubMed: 17906690]
7. Kluza J, Gallego MA, Loyens A, Beauvillain JC, Sousa-Faro JMF, Cuevas C, Marchetti P, Bailly C. Cancer cell mitochondria are direct proapoptotic targets for the marine antitumor drug lamellarin D. Cancer Res 2006;66:3177–3187. [PubMed: 16540669]
8. Facompre M, Tardy C, Bal-Mahieu C, Colson P, Perez C, Manzanares I, Cuevas C, Bailly C. Lamellarin D: A Novel Potent Inhibitor of Topoisomerase I. Cancer Res 2003;63:7392–7399. [PubMed: 14612538]
9. Marco E, Laine W, Tardy C, Lansiaux A, Iwao M, Ishibashi F, Bailly C, Gago F. Molecular determinants of topoisomerase I poisoning by lamellarins: Comparison with camptothecin and structure-activity relationships. J Med Chem 2005;48:3796–3807. [PubMed: 15916431]
10. Dias N, Vezin H, Lansiaux A, Bailly C. Topoisomerase inhibitors of marine origin and their potential use as anticancer agents. Top Curr Chem 2005;253:89–108.
11. Tardy C, Facompre M, Laine W, Baldeyrou B, Garci?a-Gravalos D, Francesch A, Mateo C, Pastor A, Jime?nez JA, Manzanares I, Cuevas C, Bailly C. Topoisomerase I-mediated DNA cleavage as a guide to the development of antitumor agents derived from the marine alkaloid lamellarin D: Triester derivatives incorporating amino acid residues. Bioorg Med Chem 2004;12:1697–1712. [PubMed: 15028262]
12. Ishibashi F, Tanabe S, Oda T, Iwao M. Synthesis and structure-activity relationship study of lamellarin derivatives. J Nat Prod 2002;65:500–504. [PubMed: 11975488]
13. Chittchang M, Batsomboon P, Ruchirawat S, Ploypradith P. Cytotoxicities and structure-activity relationships of natural and unnatural lamellarins towards cancer cell lines. ChemMedChem 2009;4:457–65. [PubMed: 19152364]
14. Thipnate P, Chittchang M, Thasana N, Saparpakorn P, Ploypradith P, Hannongbua S. 3D-QSAR analysis for cytotoxicity of lamellarins against human hormone-dependent T47D and hormone-independent MDA-MB-231 breast cancer cells. submitted.
15. Hopfinger AJ, Wang S, Tokarski JS, Jin B, Albuquerque M, Madhav PJ, Duraiswami C. Construction of 3D-QSAR models using the 4D-QSAR analysis formalism. J Am Chem Soc 1997;119:10509–10524.
16. Hopfinger AJ, Reaka A, Venkatarangan P, Duca JS, Wang S. Construction of a Virtual High Throughput Screen by 4D-QSAR Analysis: Application to a Combinatorial Library of Glucose Inhibitors of Glycogen Phosphorylase b. J Chem Inf Comput Sci 1999;39:1151–1160.
17. Krasowski MD, Hong X, Hopfinger AJ, Harrison NL. 4D-QSAR analysis of a set of propofol analogues: *Mapping binding sites* for an anesthetic phenol on the GABA$_A$ receptor. J Med Chem 2002;45:3210–3221. [PubMed: 12109905]
18. Senese CL, Duca J, Pan D, Hopfinger AJ, Tseng YJ. 4D-fingerprints, universal QSAR and QSPR descriptors. J Chem Inf Comput Sci 2004;44:1526–1539. [PubMed: 15446810]

19. Liu J, Yang L, Li Y, Pan D, Hopfinger AJ. Prediction of plasma protein binding of drugs using Kier-Hall valence connectivity indices and 4D-fingerprint molecular similarity analyses. J Comput -Aided Mol Des 2005;19:567–583. [PubMed: 16267692]

20. Liu J, Yang L, Li Y, Pan D, Hopfinger AJ. Constructing plasma protein binding model based on a combination of cluster analysis and 4D-fingerprint molecular similarity analyses. Bioorg Med Chem 2006;14:611–621. [PubMed: 16214346]

21. Li Y, Pan D, Liu J, Kern PS, Gerberick GF, Hopfinger AJ, Tseng YJ. Categorical QSAR models for skin sensitization based upon local lymph node assay classification measures Part 2: 4D-Fingerprint three-State and Two-2-State logistic regression models. Toxicol Sci 2007;99:532–544. [PubMed: 17675333]

22. Li Y, Tseng YJ, Pan D, Liu J, Kern PS, Gerberick GF, Hopfinger AJ. 4D-fingerprint categorical QSAR models for skin sensitization based on the classification of local lymph node assay measures. Chem Res Toxicol 2007;20:114–128. [PubMed: 17226934]

23. Lyer M, Zheng T, Hopfinger AJ, Tseng YJ. QSAR analyses of skin penetration enhancers. J Chem Inf Model 2007;47:1130–1149. [PubMed: 17472334]

24. Iyer M, Hopfinger AJ. Treating chemical diversity in QSAR analysis: Modeling diverse HIV-I integrase inhibitors using 4D fingerprints. J Chem Inf Model 2007;47:1945–1960. [PubMed: 17661457]

25. Ploypradith P, Petchmanee T, Sahakitpichan P, Litvinas ND, Ruchirawat S. Total synthesis of natural and unnatural lamellarins with saturated and unsaturated D-rings. J Org Chem 2006;71:9440–9448. [PubMed: 17137371]

26. 4D-QSAR User's Manual, V., The ChemBats21 Group, Inc., 1780 Wilson Dr., Lake Forest, IL 60045, 2003.

27. *HyperChem Program Release 7.5 for Windows*; Hypercube, I.

28. Doherty, D. C. M. U. S. G., The ChemBats21 Group, Inc., 1780 Wilson Dr., Lake Forest, IL 60045, 1997.

29. Allinger NL. Conformational analysis. 130. MM2. A hydrocarbon force field utilizing V1 and V2 torsional terms. J Am Chem Soc 1997;99:8127–8134.

30. Hopfinger AJP, RA. Molecular mechanics force-field parametrization precedures. J Comput Chem 1984;5:486–492.

31. Glen WGD, WJ, Scott DR. Principal components analysis and partial least squares. Tetrahedron Comput Methods 1989;2:349–354.

32. Rogers D, Hopfinger AJ. Application of genetic function approximation to quantitative structure-activity relationships and quantitative structure-property relationships. J Chem Inf Comput Sci 1994;34:854–866.

33. Friedman, J. M. a. r. s. T. R. N. L. f. C. S., Department of Statistics, Stanford University, Standford, CA, November 1988 (revised August 1990).

34. Duca JS, Hopfinger AJ. Estimation of Molecular Similarity Based on 4D-QSAR Analysis: Formalism and Validation. J Chem Inf Comput Sci 2001;41:1367–1387. [PubMed: 11604039]

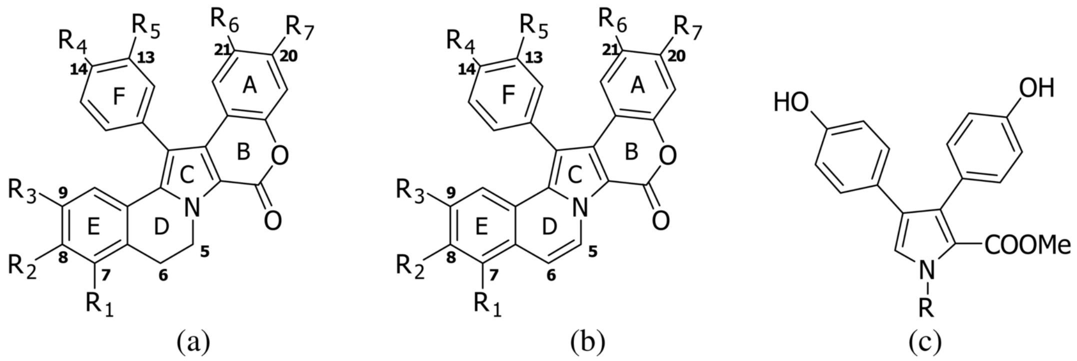


Figure 1.

The three scaffold groups forming the training set of lamellarins.

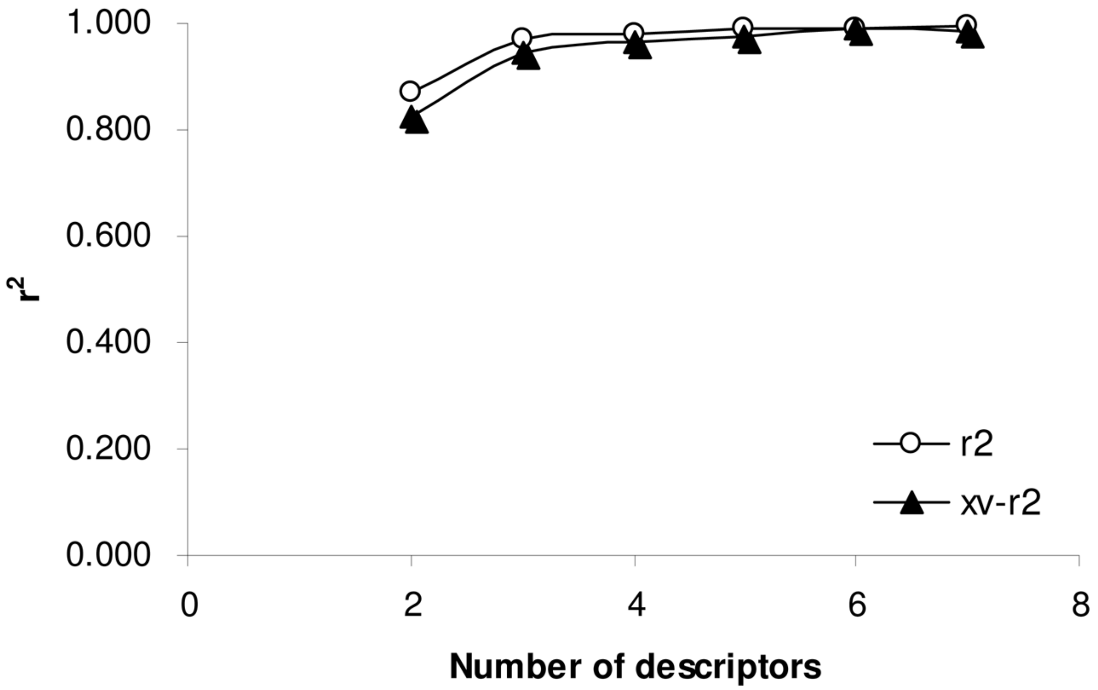


Figure 2.
Plot of the number of RI-4D-QSAR model descriptor terms versus r^2, and xv-r^2 for the complete training set.

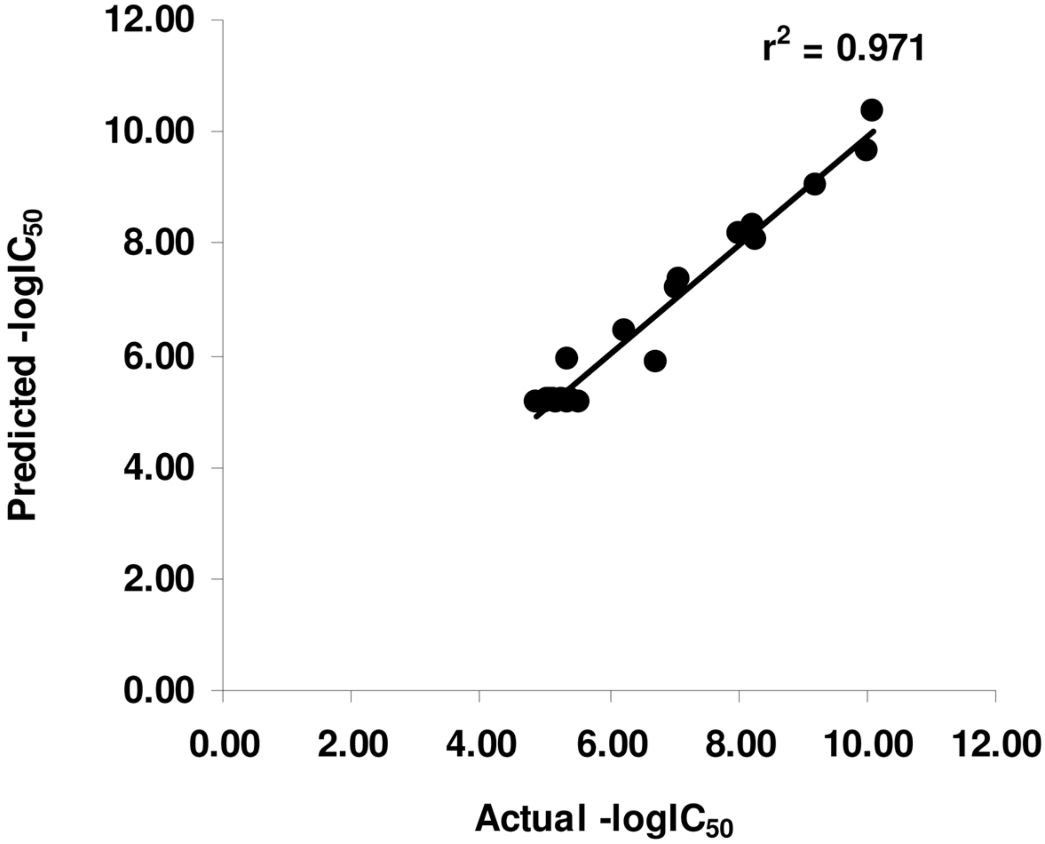


Figure 3.

Predicted $-\log IC_{50}$ values, using the RI-4D-QSAR model for the 25 lamellarins data set, versus the observed $-\log IC_{50}$ values.

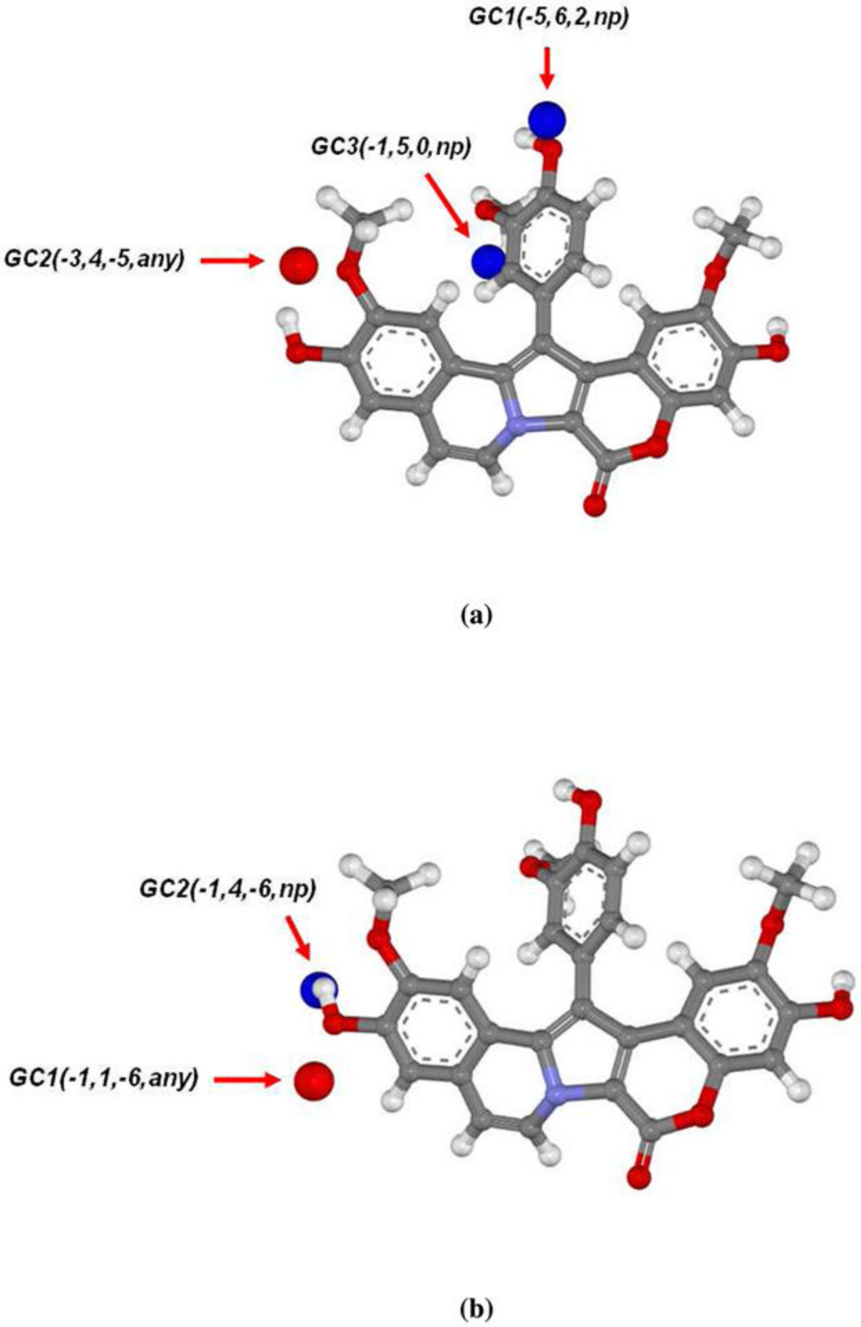


Figure 4.

The 3D-pharmacophores from (a) eq. 9 based upon the full training set of 25 lamellarins, and (b) from the 6 high activity compounds of the lamellarin training set. The 3D-pharmacophores are shown relative to the predicted active conformation of the most active compound (lamellarin D). The red spheres refer to pharmacophore sites having negative regression coefficients in the 4D-QSAR equation, and blue spheres refer to pharmacophore sites having positive regression coefficients.

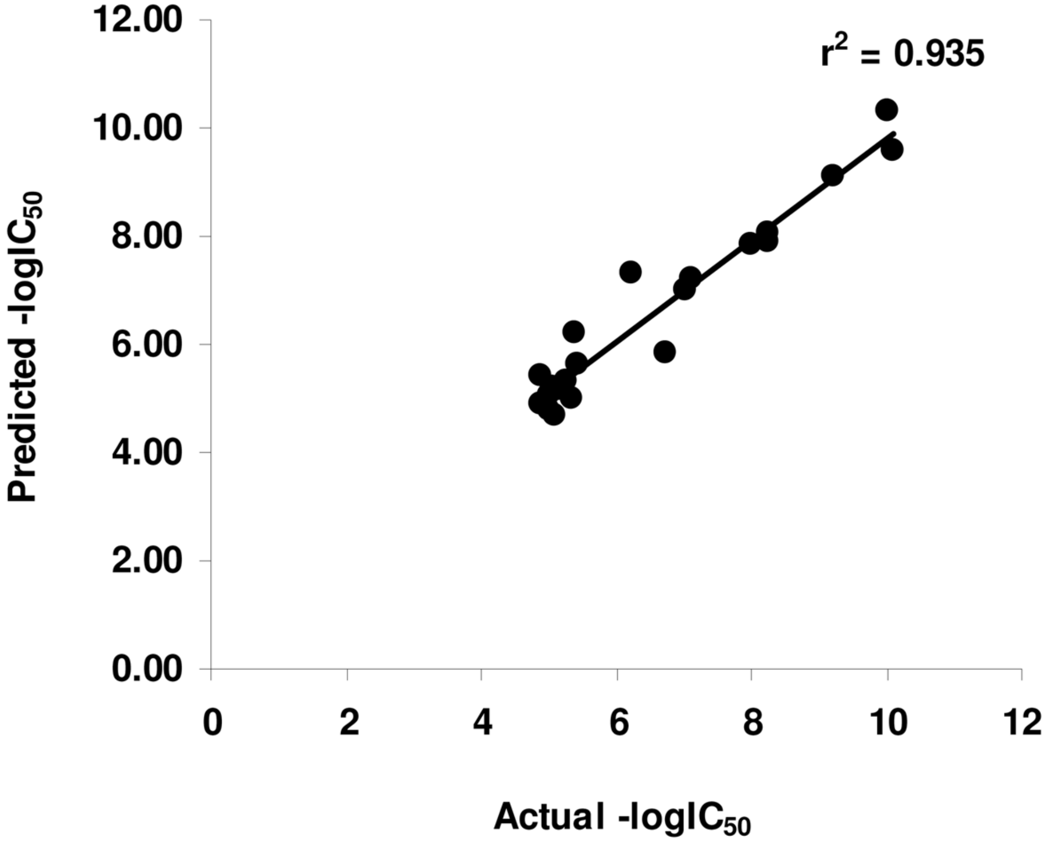


Figure 5.

Predicted −logIC_{50} values, using the RI-4D-QSAR model for the 21 lamellarins data set, versus the observed −logIC_{50} values.

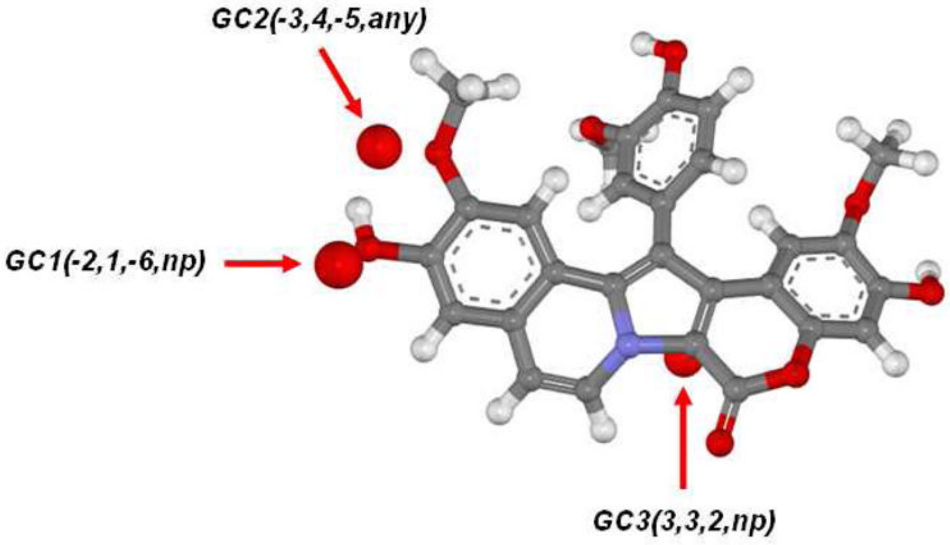


Figure 6.

The 3D-pharmacophores from eq. 11 based upon the 21 lamellarins training set. The 3D-pharmacophores are shown relative to the predicted active conformation of the most active compound (lamellarin D). The red spheres refer to pharmacophore sites having negative regression coefficients in the 4D-QSAR equation, and blue spheres refer to pharmacophore sites having positive regression coefficients.

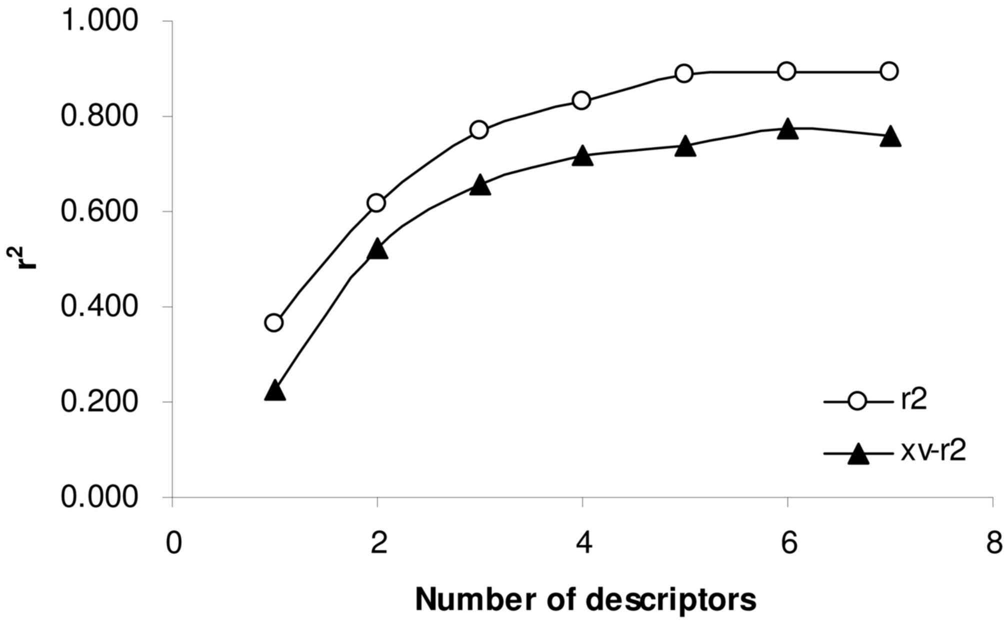


Figure 7.
Plot of the number of 4D-fingerprint model descriptor terms versus r^2, and *xv*-r^2 for the complete training set.

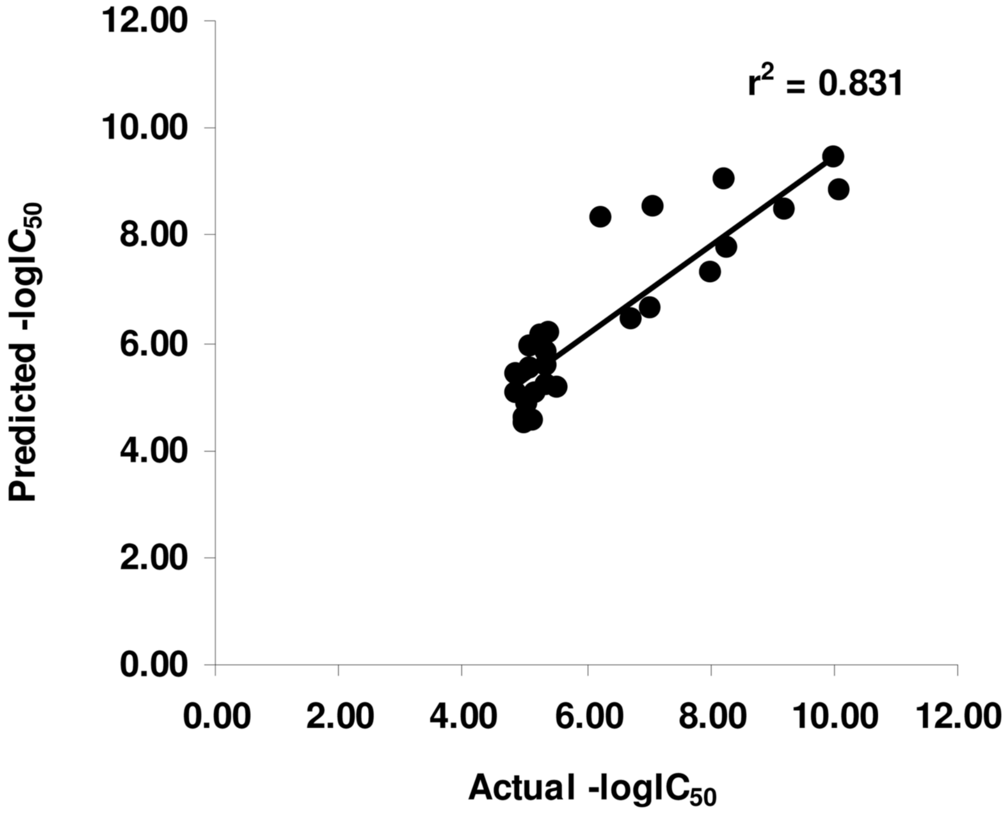
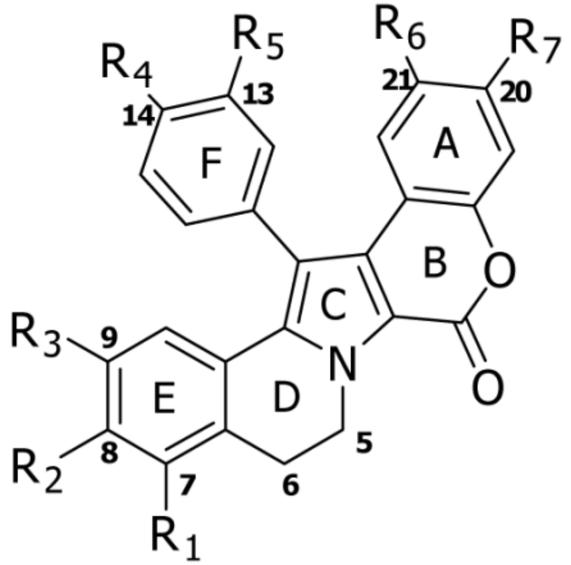


Figure 8.

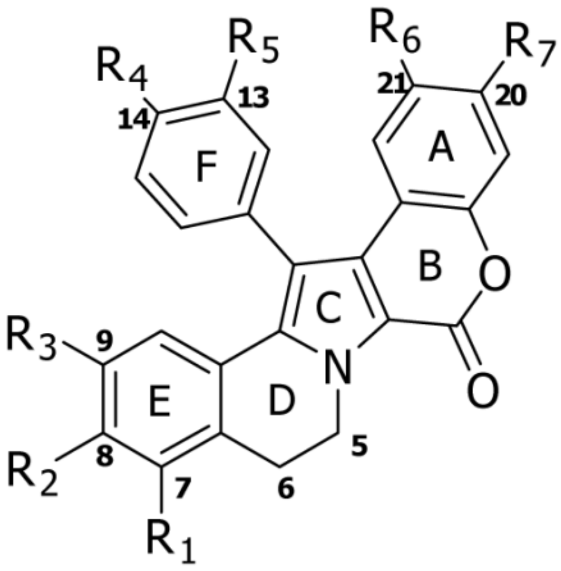
Predicted −$\log IC_{50}$ values, using the 4D-fingerprint model for the 25 lamellarins data set, versus the observed −$\log IC_{50}$ values.

Table 1

Chemical structures and cytotoxic activities, $-\log IC_{50}$, of lamellarins with a saturated D-ring



Lamellarin	MW	*Log P**	Substituent group							$-$Log IC_{50}
			R_1	R_2	R_3	R_4	R_5	R_6	R_7	
C	545.6	3.4	OMe	OMe	OMe	OH	OMe	OMe	OH	5.11
E	531.5	3.4	OH	OMe	OMe	OMe	OH	OMe	OH	5.28
F	545.2	3.4	OH	OMe	OMe	OMe	OMe	OMe	OH	5.34
G	501.5	3.6	H	OH	OMe	OMe	OH	OH	OMe	5.07
I	559.6	3.4	OMe	OMe	OMe	OMe	OMe	OMe	OH	5.02
J	515.5	3.7	H	OH	OMe	OMe	OMe	OMe	OH	4.89
K**	531.51	3.7	OH	OMe	OMe	OH	OMe	OMe	OH	7.04
L	501.5	3.6	H	OH	OMe	OMe	OH	OMe	OH	5.36
T	545.6	3.4	OMe	OMe	OMe	OMe	OH	OMe	OH	4.88
U	515.5	3.7	H	OMe	OMe	OMe	OH	OMe	OH	4.99
Y	501.1	3.6	H	OMe	OH	OMe	OH	OMe	OH	5.14



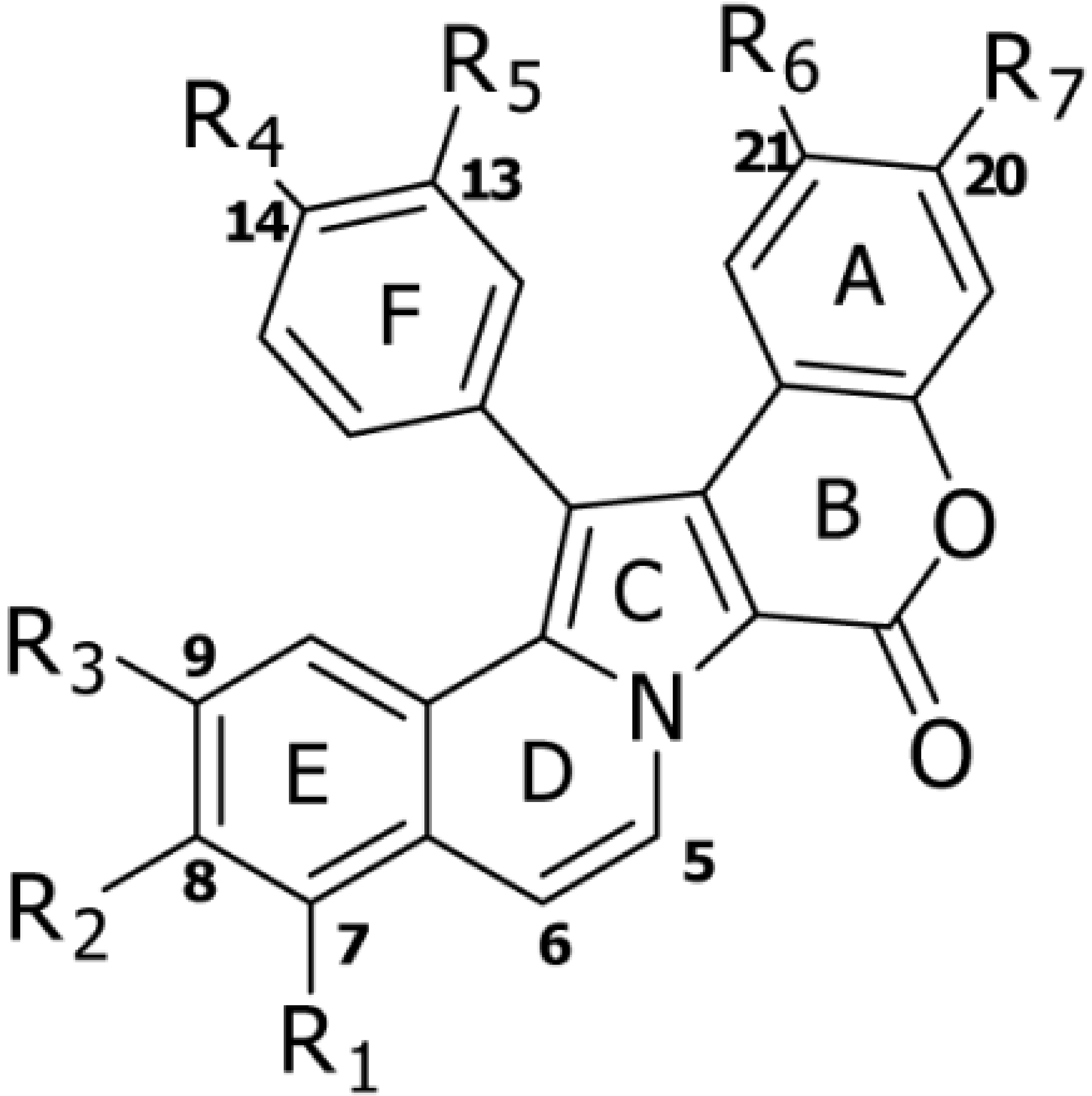
			Substituent group							
Lamellarin	**MW**	***Log P*** [*]	**R_1**	**R_2**	**R_3**	**R_4**	**R_5**	**R_6**	**R_7**	**−Log IC_{50}**
χ	501.5	3.6	H	OH	OMe	OH	OMe	OMe	OH	5.42
K triacetate	657.2	2.7	OAc	OMe	OMe	OAc	OMe	OMe	OAc	5.18
U diacetate	599.2	3.2	H	OMe	OMe	OMe	OAc	OMe	OAc	5.10
χ triacetate	627.2	3.0	H	OAc	OMe	OAc	OMe	OMe	OAc	5.54

[*] Calculated by *CS ChemDraw Ultra* version 5.0 (CambridgeSoftCorporation, Cambridge, MA, USA)

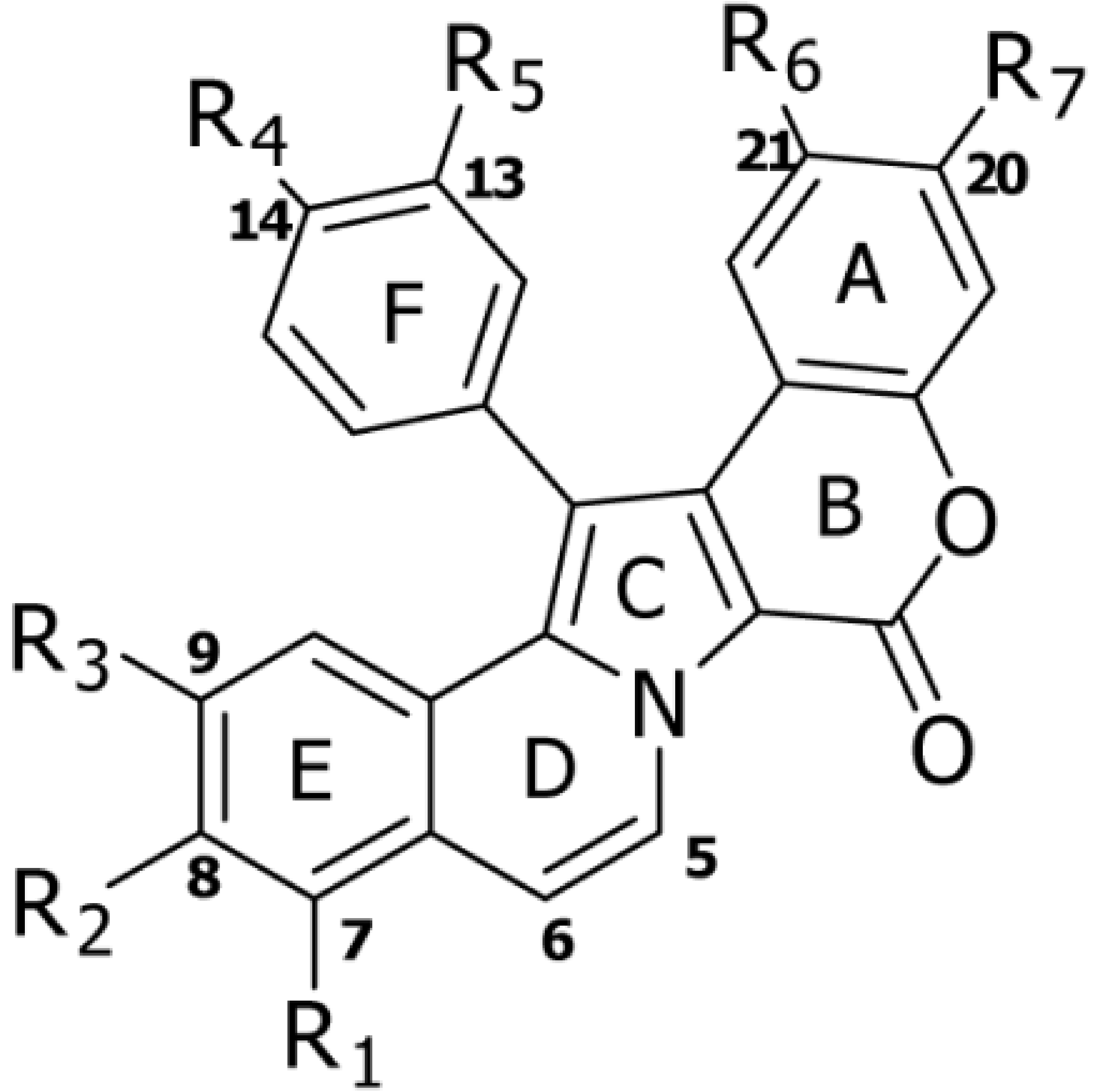
[**] This compound was not included in the original 25 lamellarin compounds training set used to build the 4D-QSAR and 3D-pharmacophore models. However, it shows unexpected high activity for the saturated D-ring series of compounds and thus investigated as part of this study.

Table 2

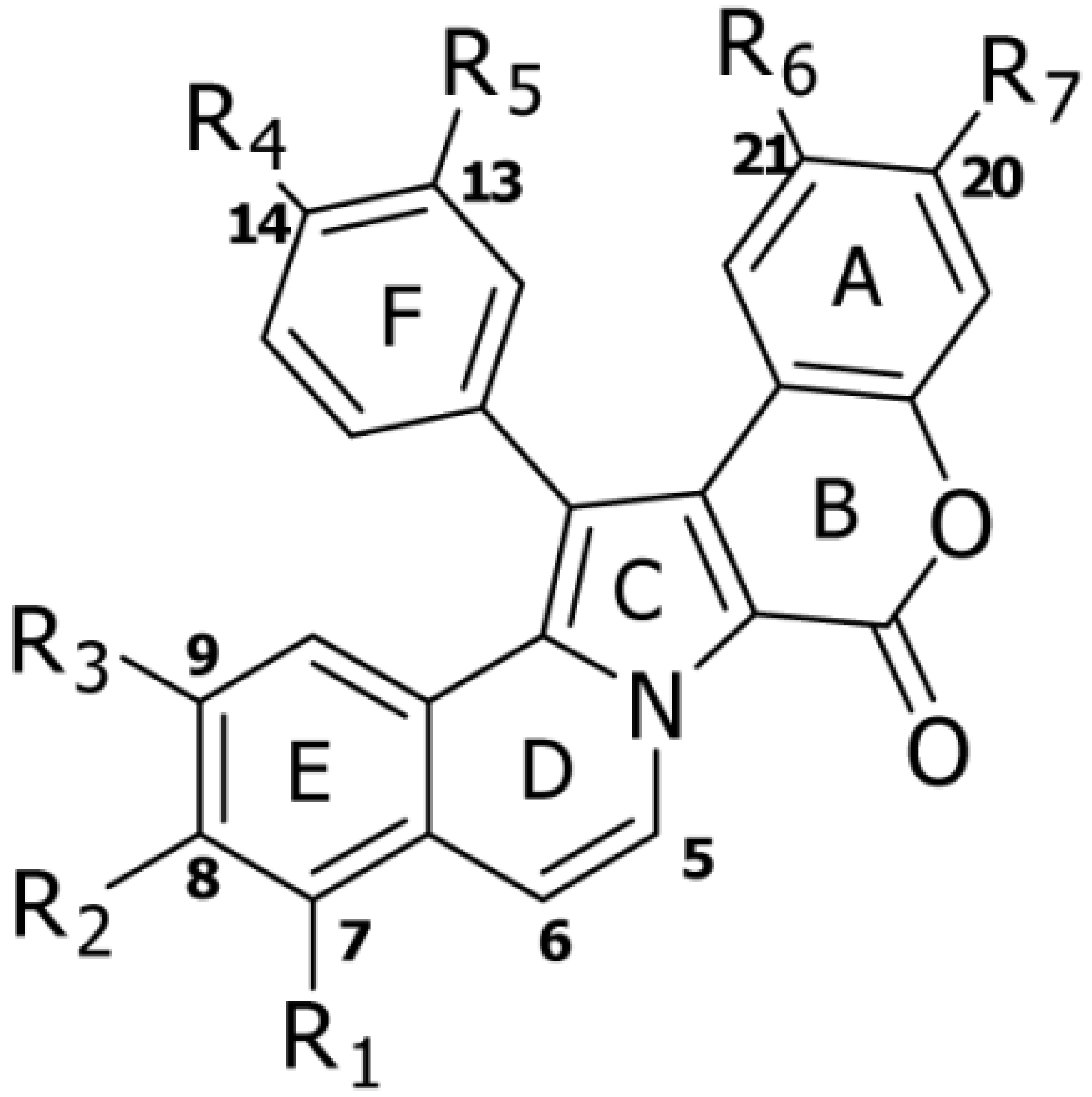
Chemical structures and cytotoxic activities, $-\log IC_{50}$, of lamellarins with an unsaturated D-ring



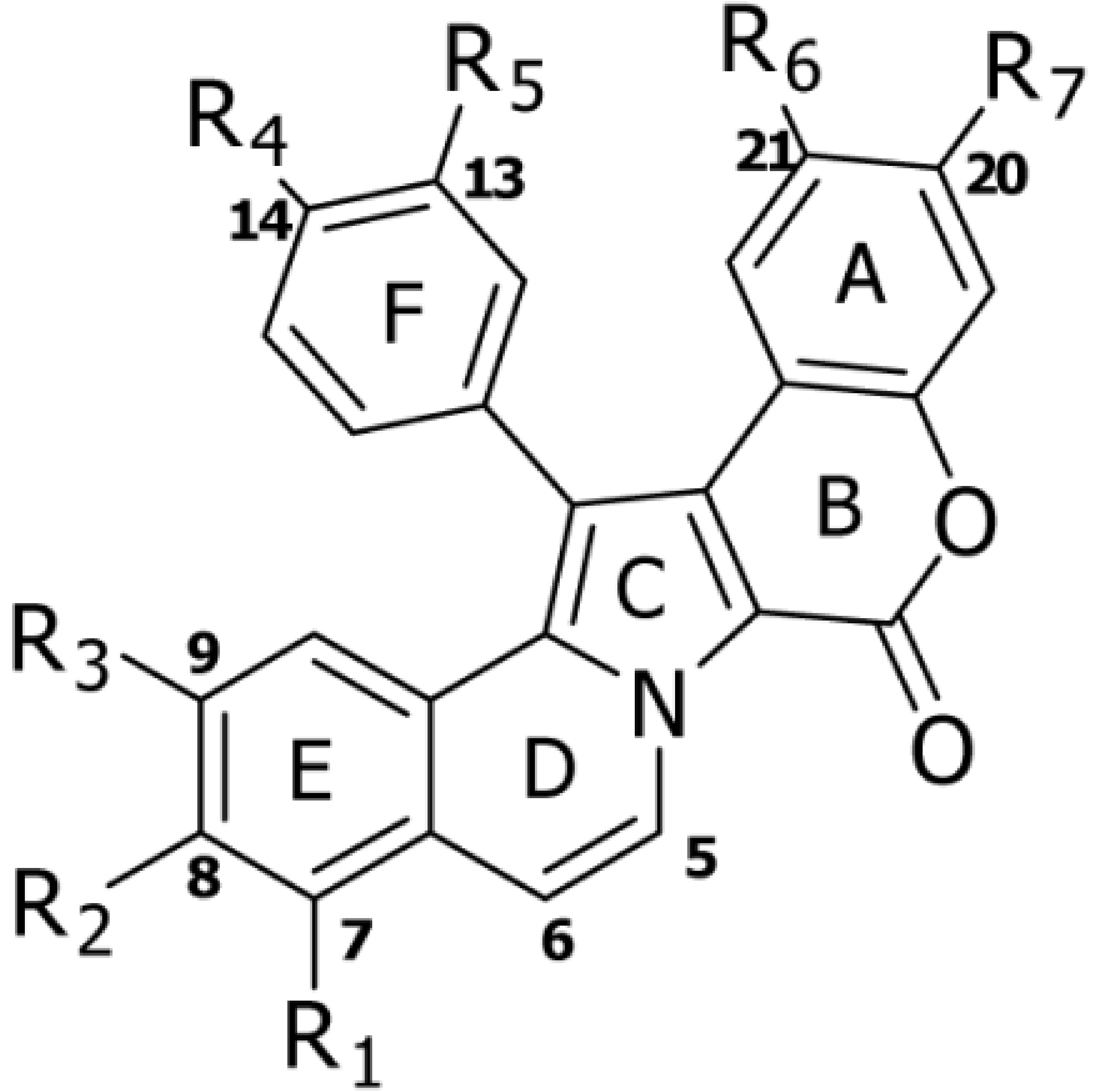
Lamellarin	MW	*Log P**	Substituent group							−Log IC_{50}
			R_1	R_2	R_3	R_4	R_5	R_6	R_7	
B	543.5	3.3	OMe	OMe	OMe	OH	OMe	OMe	OH	6.74



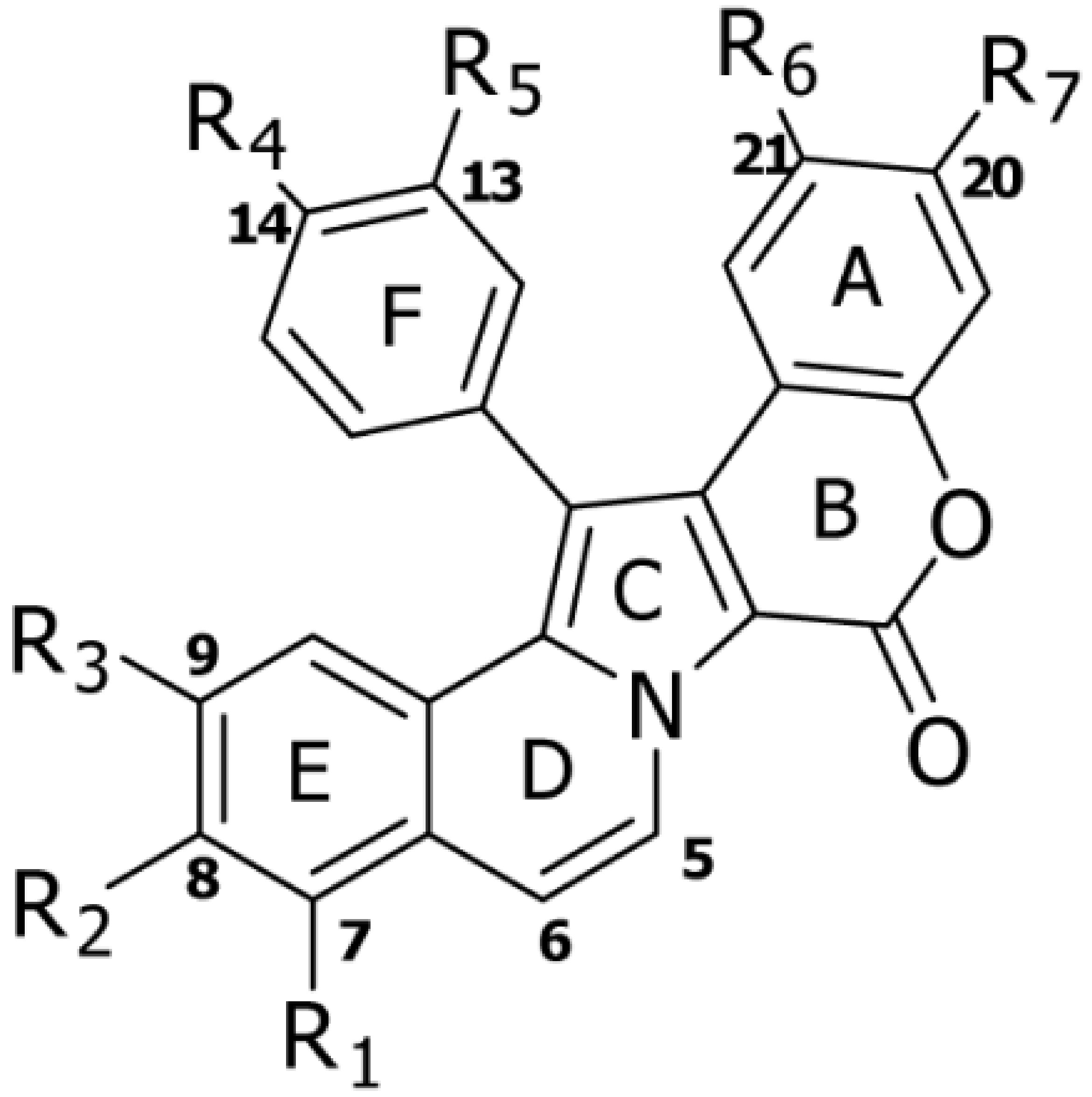
			Substituent group							
Lamellarin	**MW**	***Log P****	**R_1**	**R_2**	**R_3**	**R_4**	**R_5**	**R_6**	**R_7**	**−Log IC_{50}**
D	499.5	3.6	H	OH	OMe	OH	OMe	OMe	OH	10.10



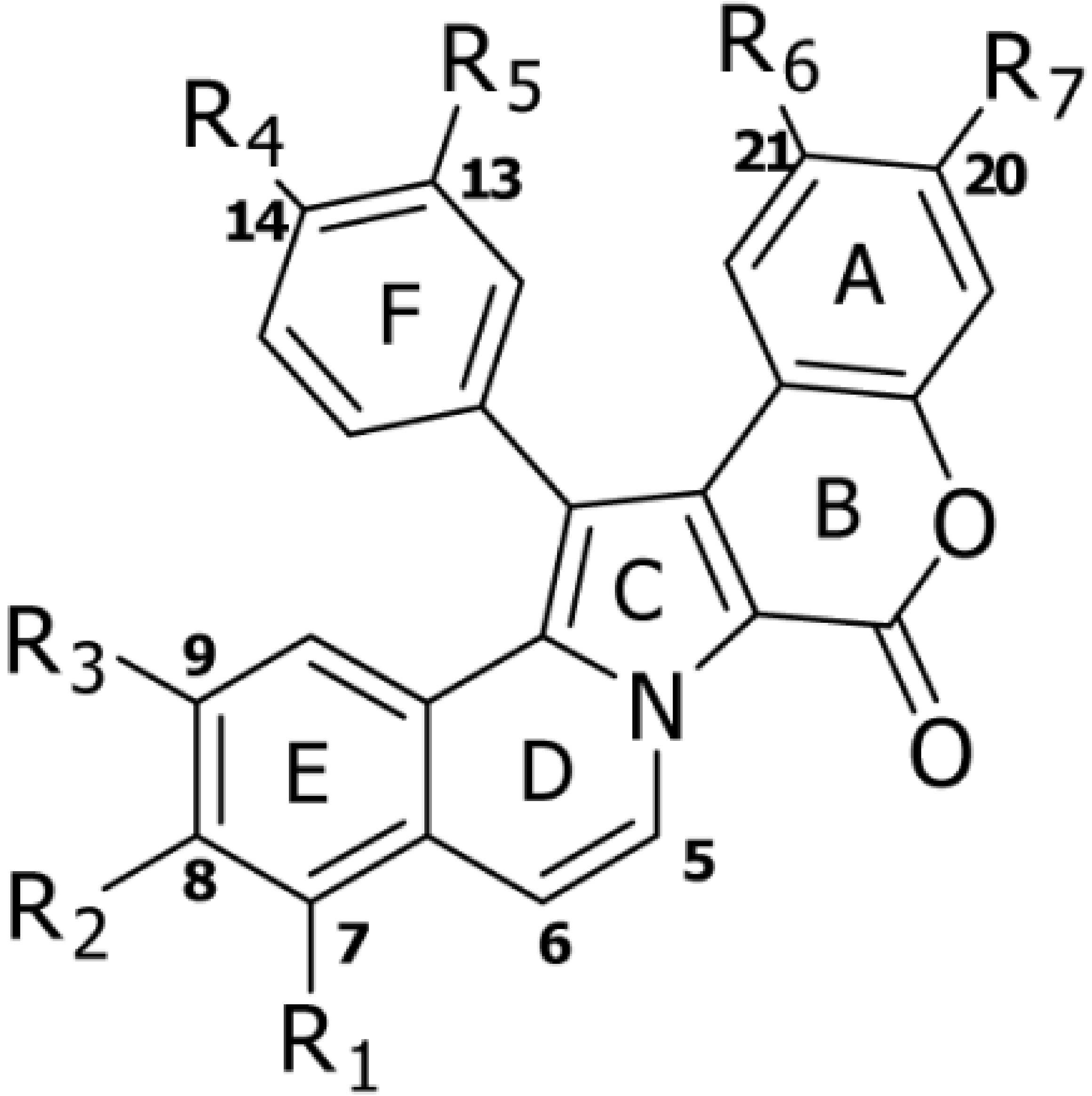
Lamellarin	MW	*Log P**	Substituent group							−Log IC_{50}
			R_1	R_2	R_3	R_4	R_5	R_6	R_7	
M	529.5	3.3	OH	OMe	OMe	OH	OMe	OMe	OH	8.02



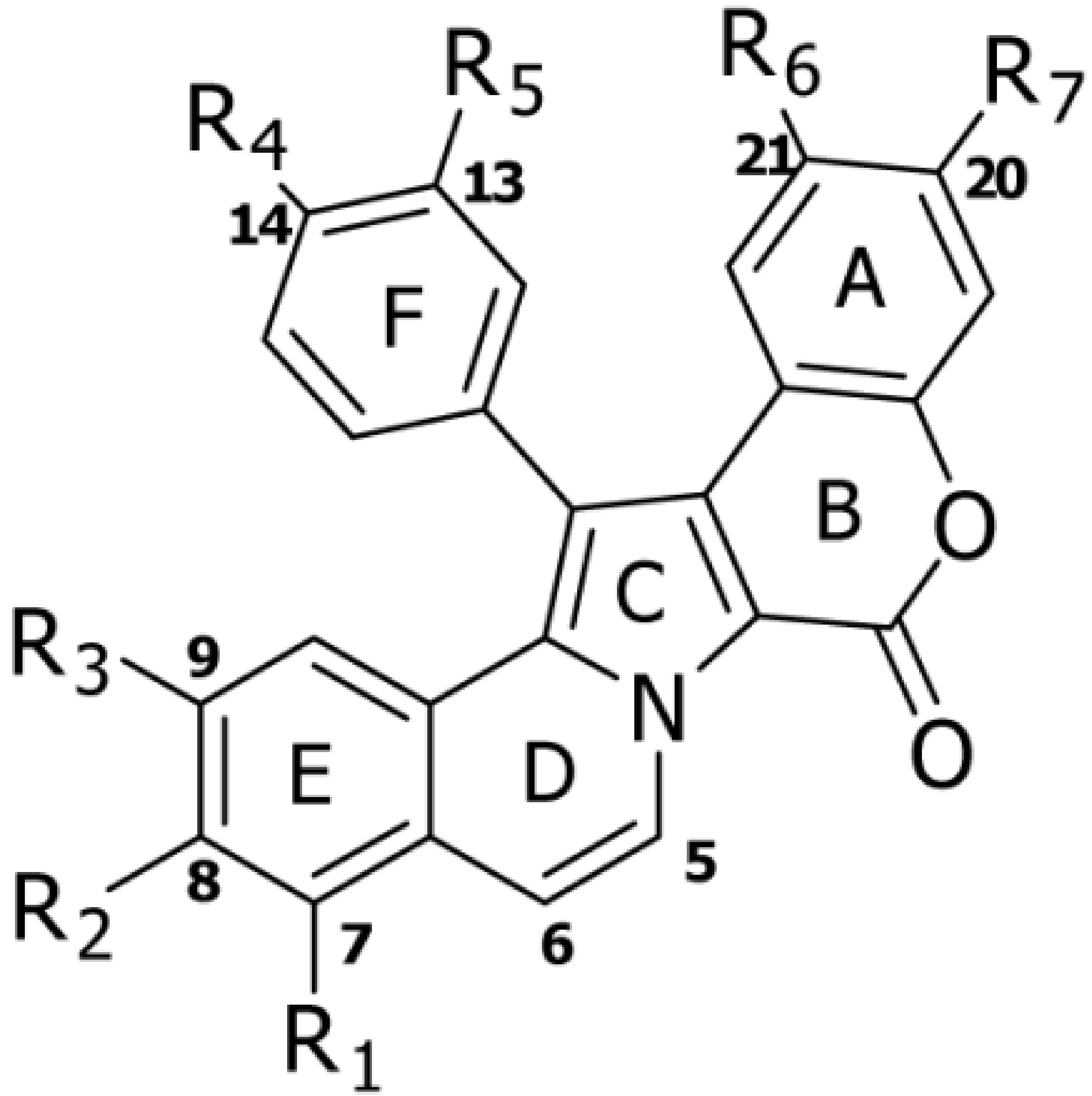
			Substituent group							
Lamellarin	**MW**	***Log P*** [*]	**R_1**	**R_2**	**R_3**	**R_4**	**R_5**	**R_6**	**R_7**	**$-$Log IC_{50}**
N	499.5	3.6	H	OH	OMe	OMe	OH	OMe	OH	9.22



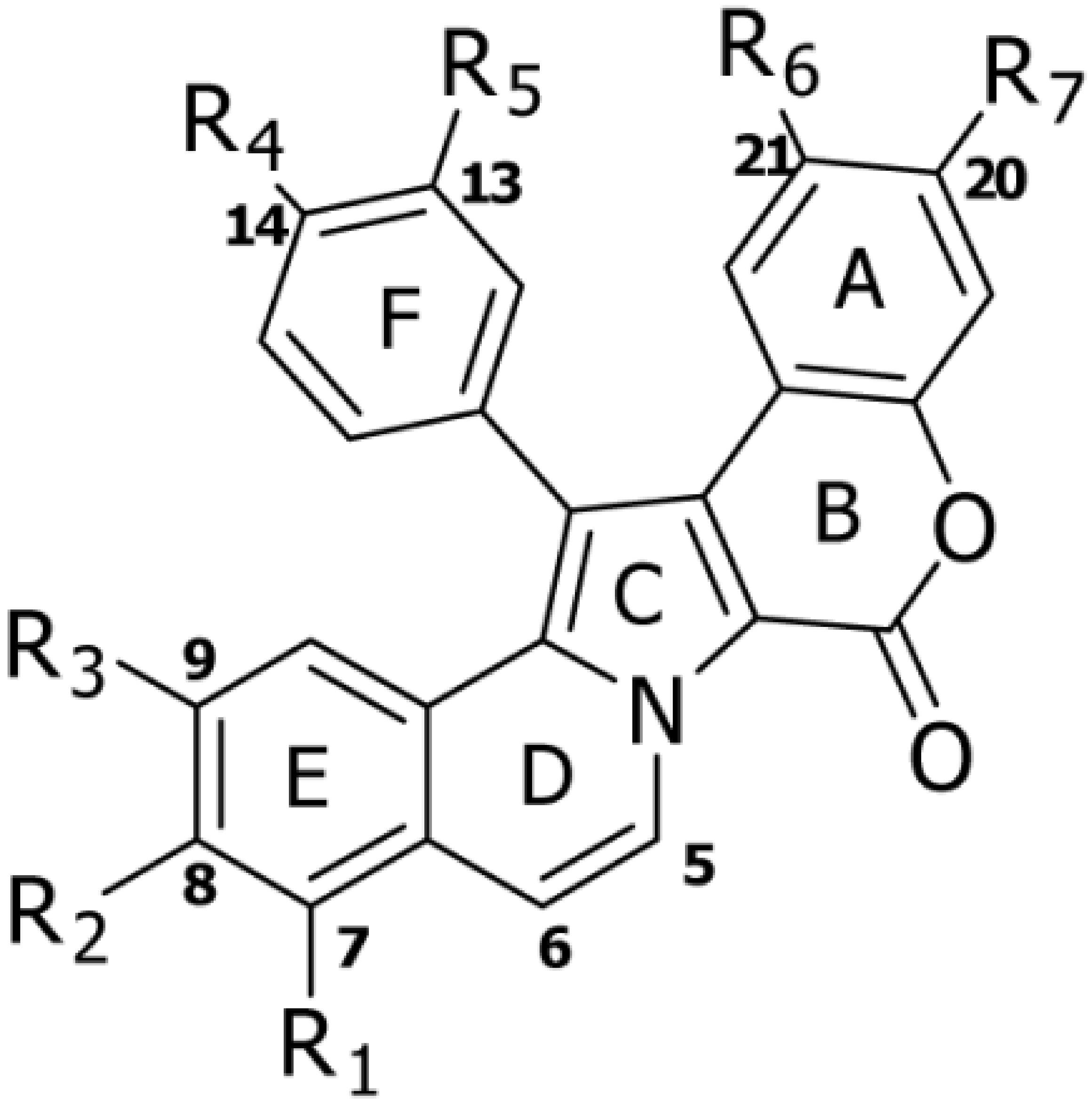
Lamellarin	MW	*Log P*[*]	Substituent group							−Log IC_{50}
			R_1	R_2	R_3	R_4	R_5	R_6	R_7	
W	543.5	3.3	OMe	OMe	OMe	OMe	OH	OMe	OH	5.37



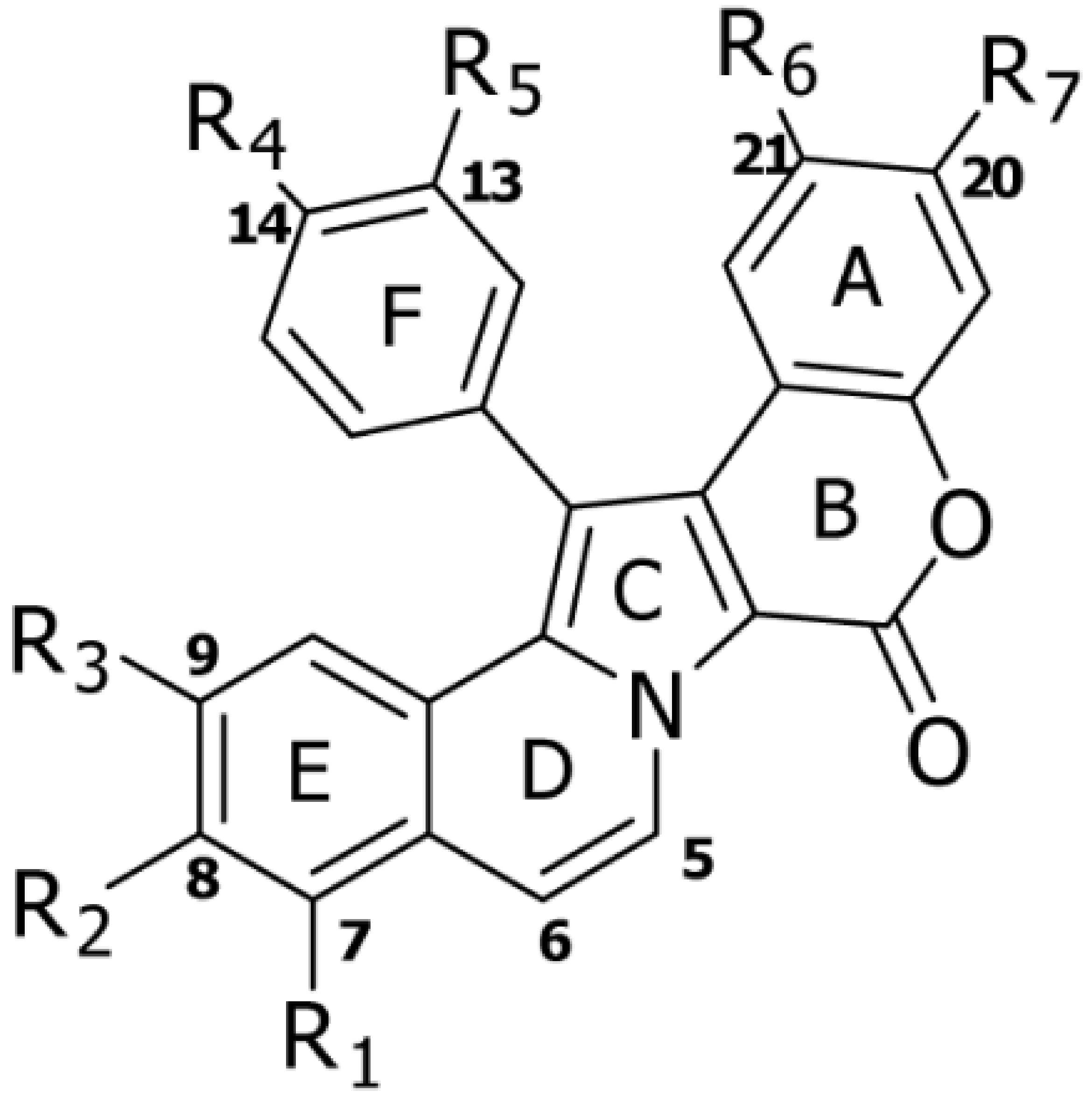
Lamellarin	MW	*Log P**	Substituent group							−Log IC_{50}
			R_1	R_2	R_3	R_4	R_5	R_6	R_7	
α	513.5	3.6	H	OMe	OMe	OMe	OH	OMe	OH	6.23



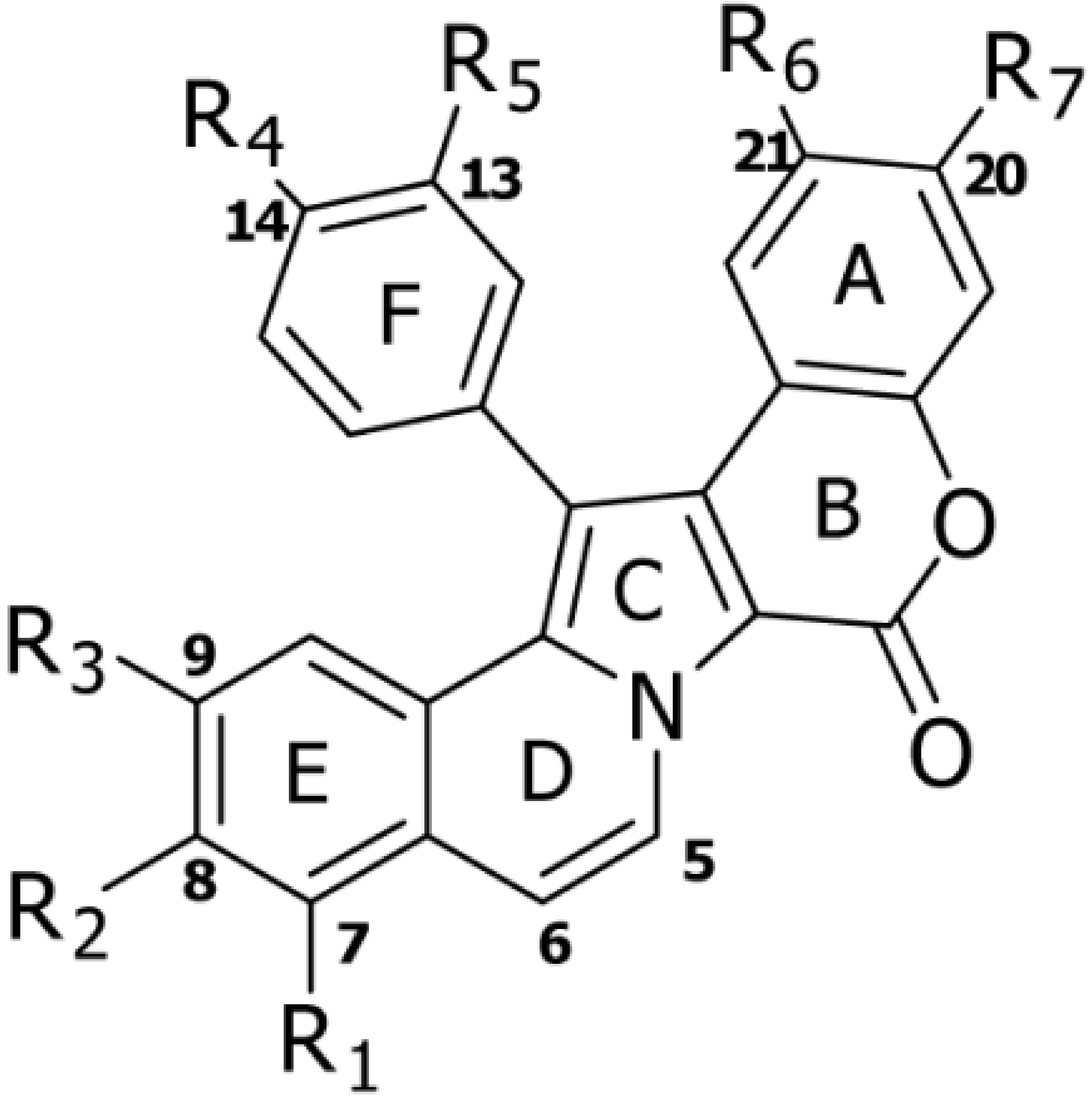
			Substituent group							
Lamellarin	**MW**	***Log P****	**R_1**	**R_2**	**R_3**	**R_4**	**R_5**	**R_6**	**R_7**	**$-$Log IC_{50}**
X	529.5	3.3	OH	OMe	OMe	OMe	OH	OMe	OH	8.25



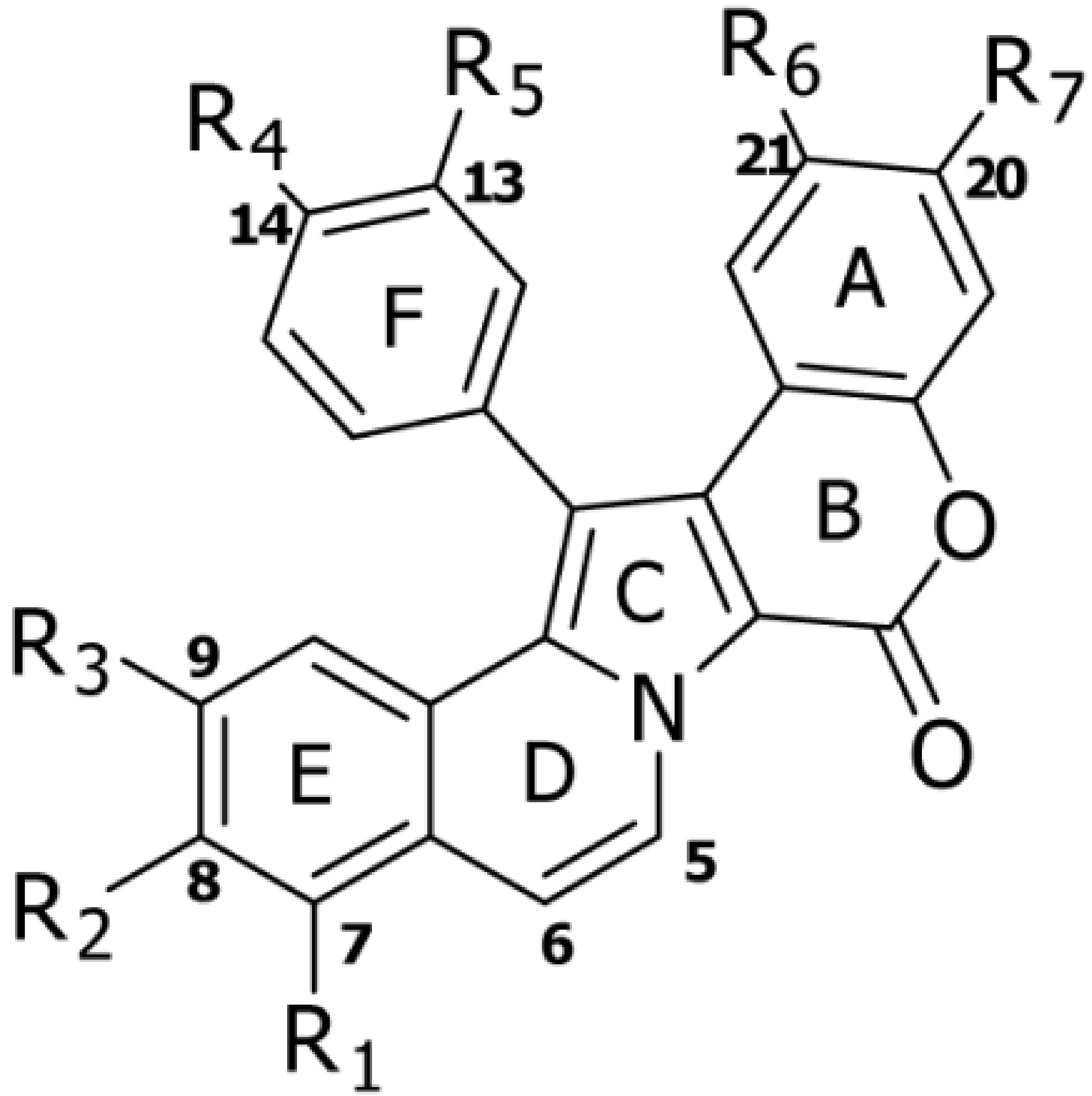
Lamellarin	MW	*Log P*[*]	Substituent group							−Log IC_{50}
			R_1	R_2	R_3	R_4	R_5	R_6	R_7	
ε	543.5	3.3	OH	OMe	OMe	OMe	OMe	OMe	OH	8.26



			Substituent group							
Lamellarin	MW	*Log P*[*]	R_1	R_2	R_3	R_4	R_5	R_6	R_7	$-\text{Log IC}_{50}$
ζ	557.6	3.4	OMe	OMe	OMe	OMe	OMe	OMe	OH	7.05



			Substituent group							
Lamellarin	MW	*Log P**	R_1	R_2	R_3	R_4	R_5	R_6	R_7	−Log IC_{50}
Dehydro. J	513.5	3.6	H	OH	OMe	OMe	OMe	OMe	OH	10.01



			Substituent group							
Lamellarin	**MW**	***Log P****	**R_1**	**R_2**	**R_3**	**R_4**	**R_5**	**R_6**	**R_7**	**$-\text{Log IC}_{50}$**
Dehydro. Y	499.5	3.6	H	OMe	OH	OMe	OH	OMe	OH	7.10

Calculated by *CS ChemDraw Ultra* version 5.0 (CambridgeSoftCorporation, Cambridge, MA, USA)

Table 3

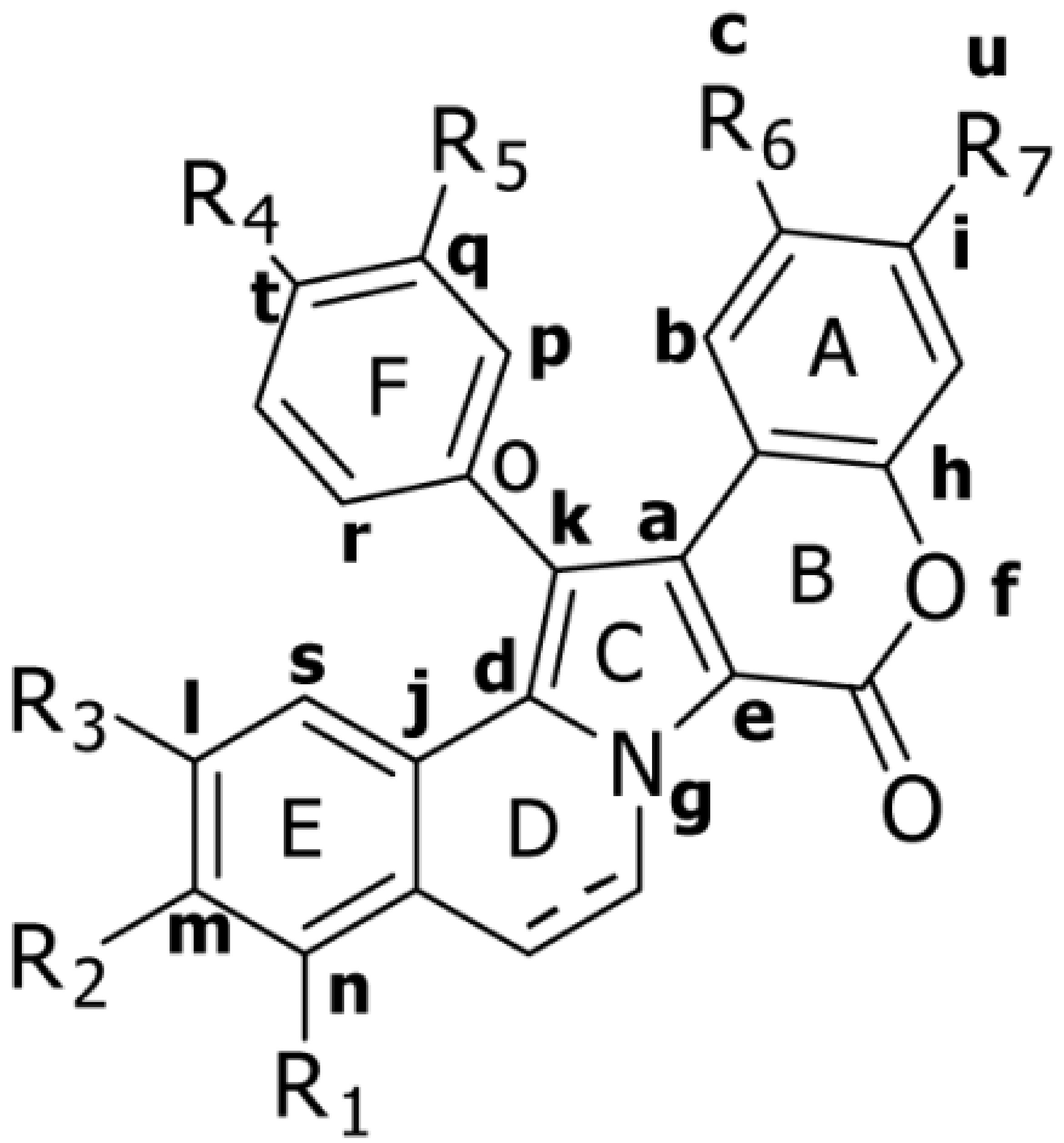
The set of Interaction Pharmacophore Elements (IPEs) used in the RI-4D-QSAR and 4D-fingerprint QSAR Analyses

IPE description	Symbol	Number code
all atoms in the molecule	any	0
nonpolar atoms	np	1
polar atoms of positive partial charge	p^+	2
polar atoms of negative partial charge	p^-	3
hydrogen bond acceptor atoms	hba	4
hydrogen bond donor atoms	hbd	5
aromatic atoms	aro	6
non-hydrogen atoms[a]	hs	7

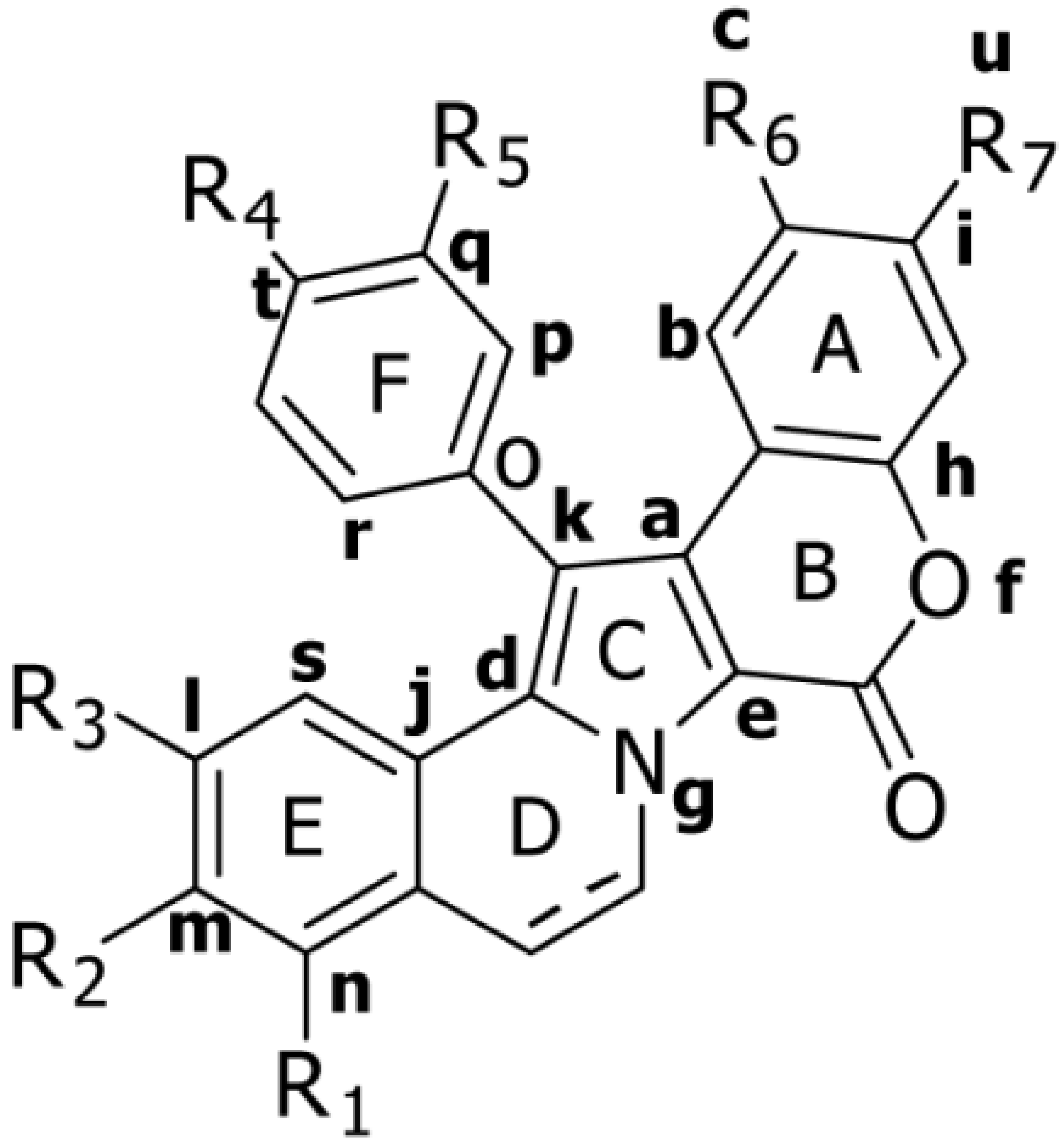
[a] hydrogen-suppressed use only in 4D-fingerprint QSAR analysis

Table 4

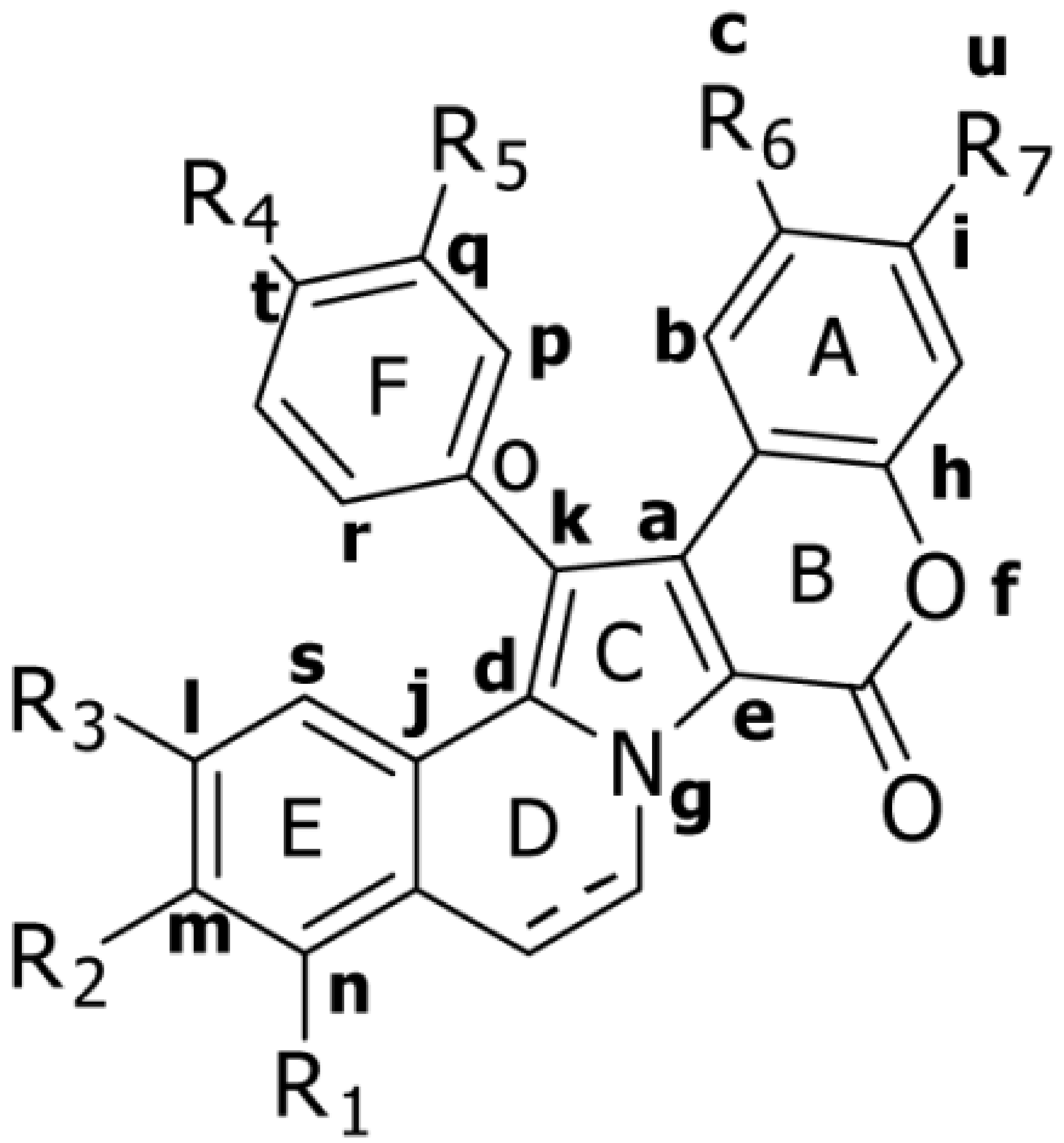
Set of trial alignment used in constructing the best five-term RI-4D-QSAR models.



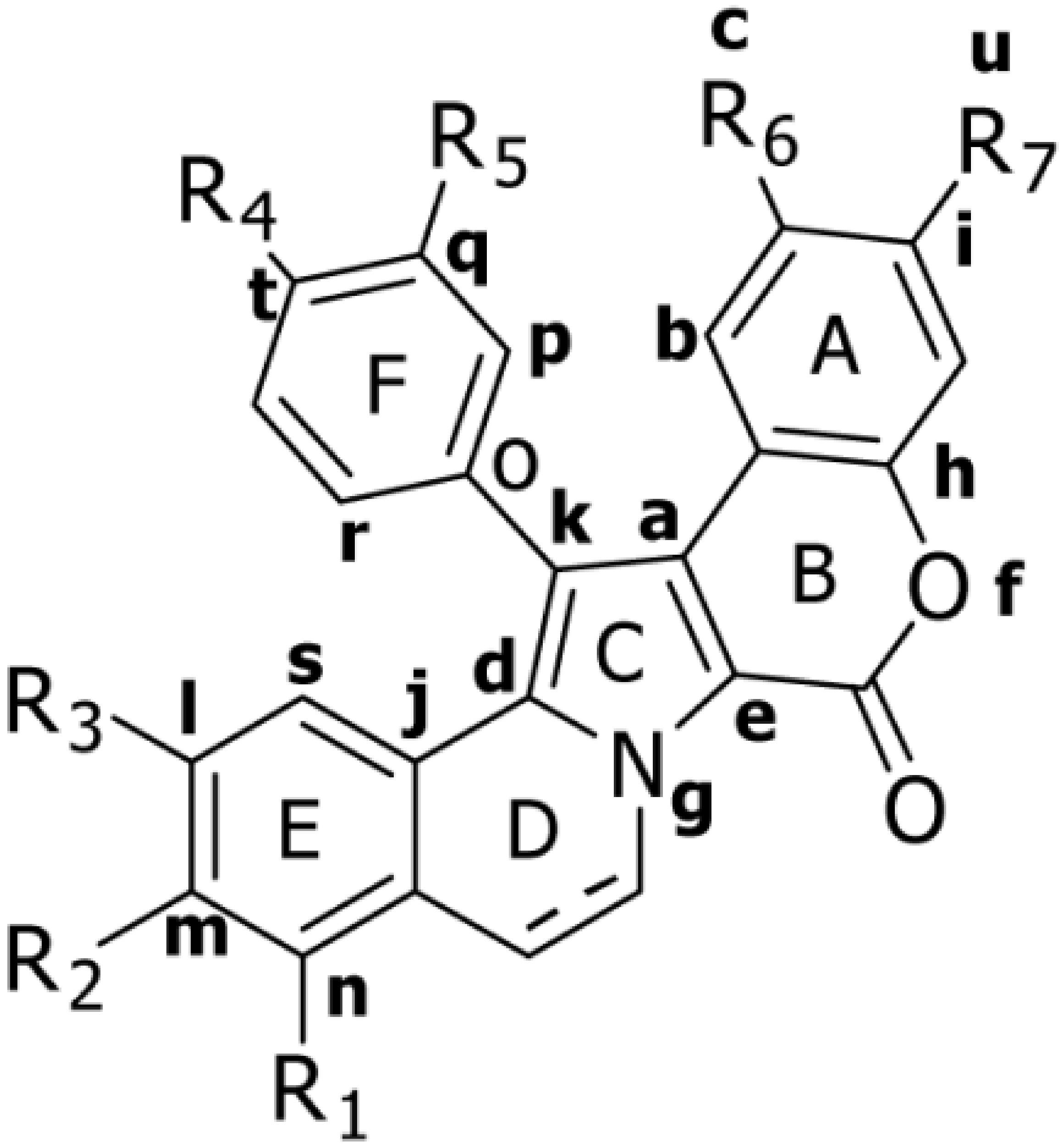
Alignment	First atom	Second atom	Third atom	r^2	$xv\text{-}r^2$
1	a	b	c	0.964	0.929



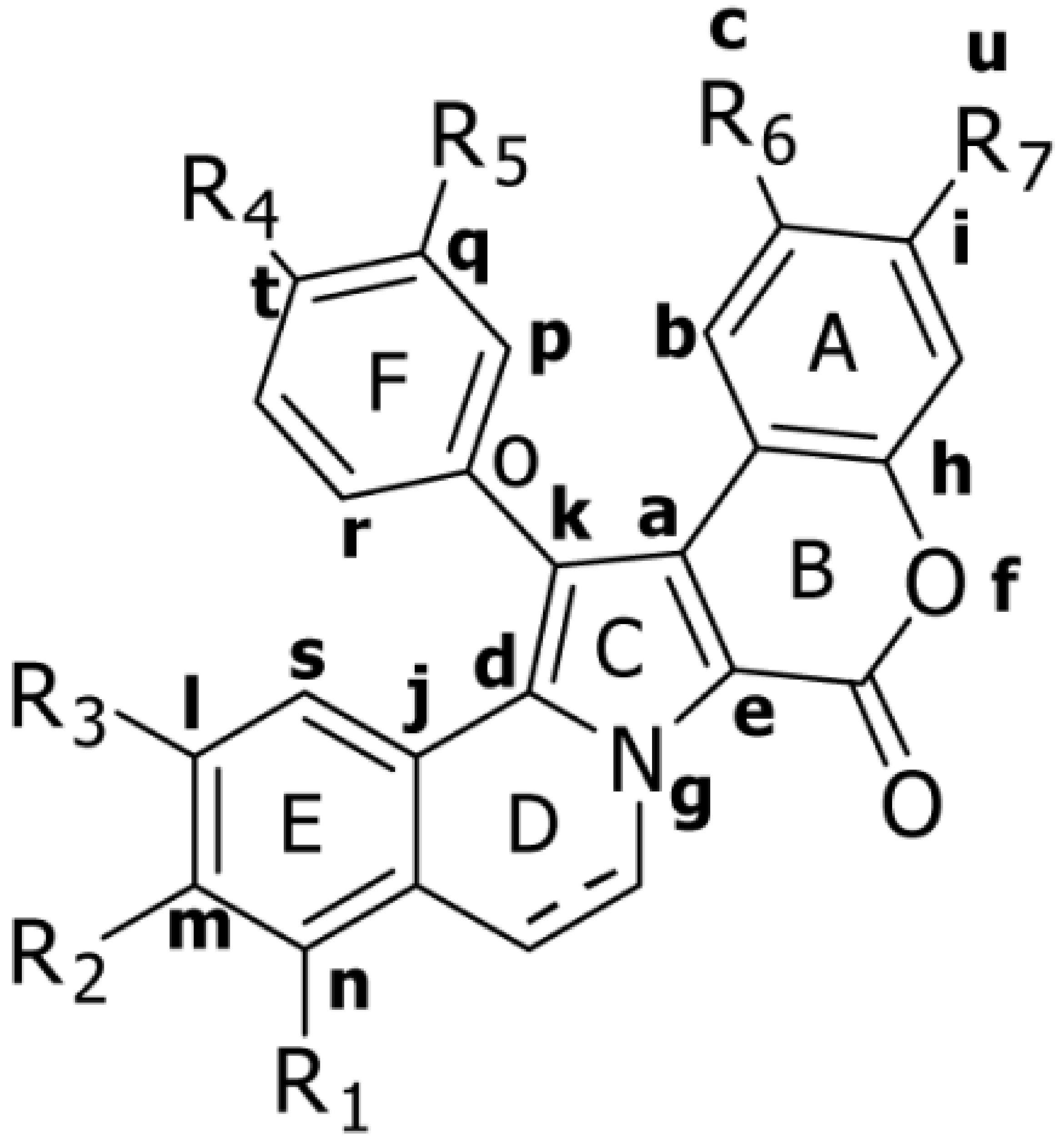
Alignment	First atom	Second atom	Third atom	r^2	$xv\text{-}r^2$
2	d	e	f	0.996	0.992



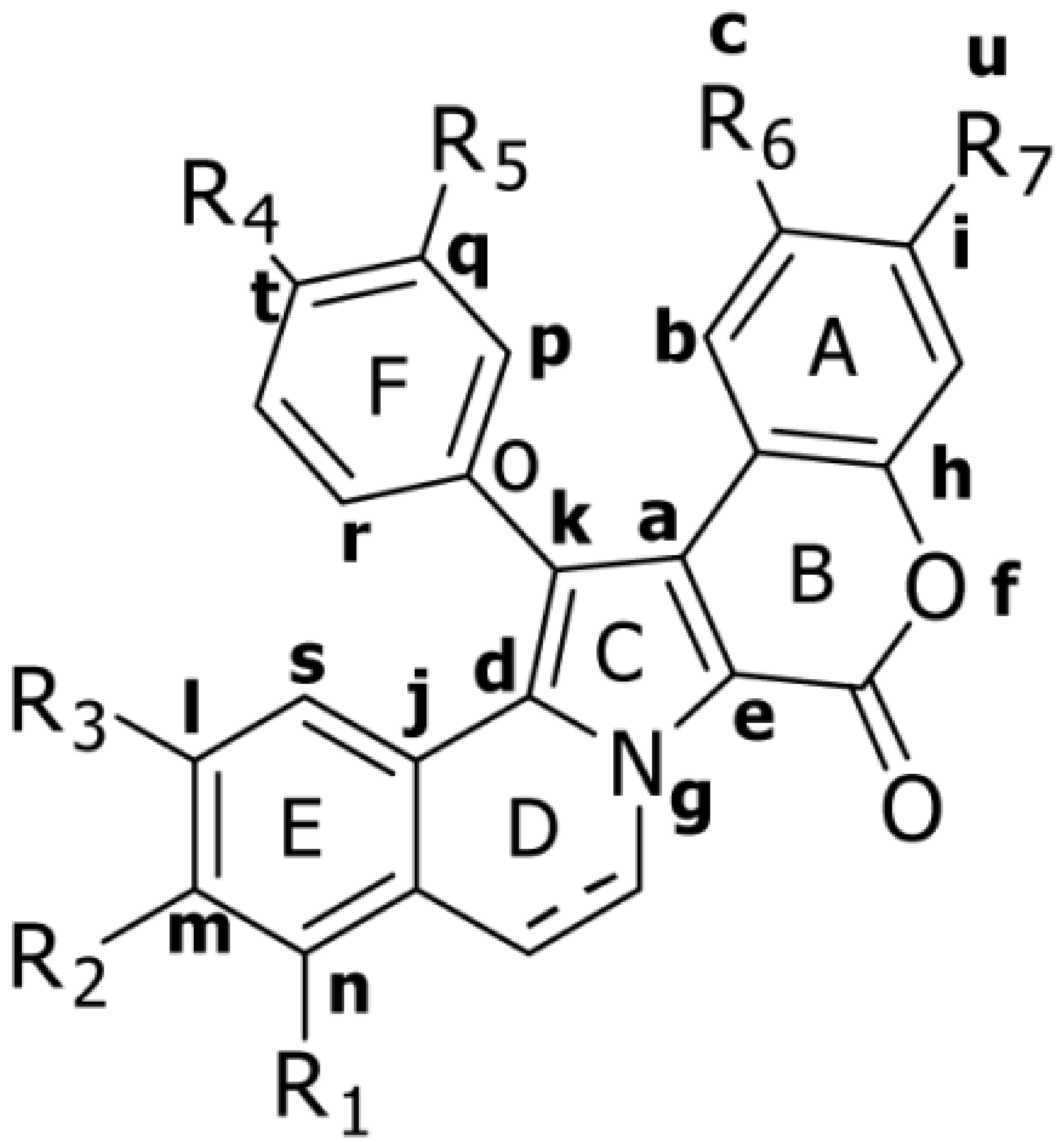
Alignment	First atom	Second atom	Third atom	r^2	$xv\text{-}r^2$
3	g	h	i	0.999	0.998



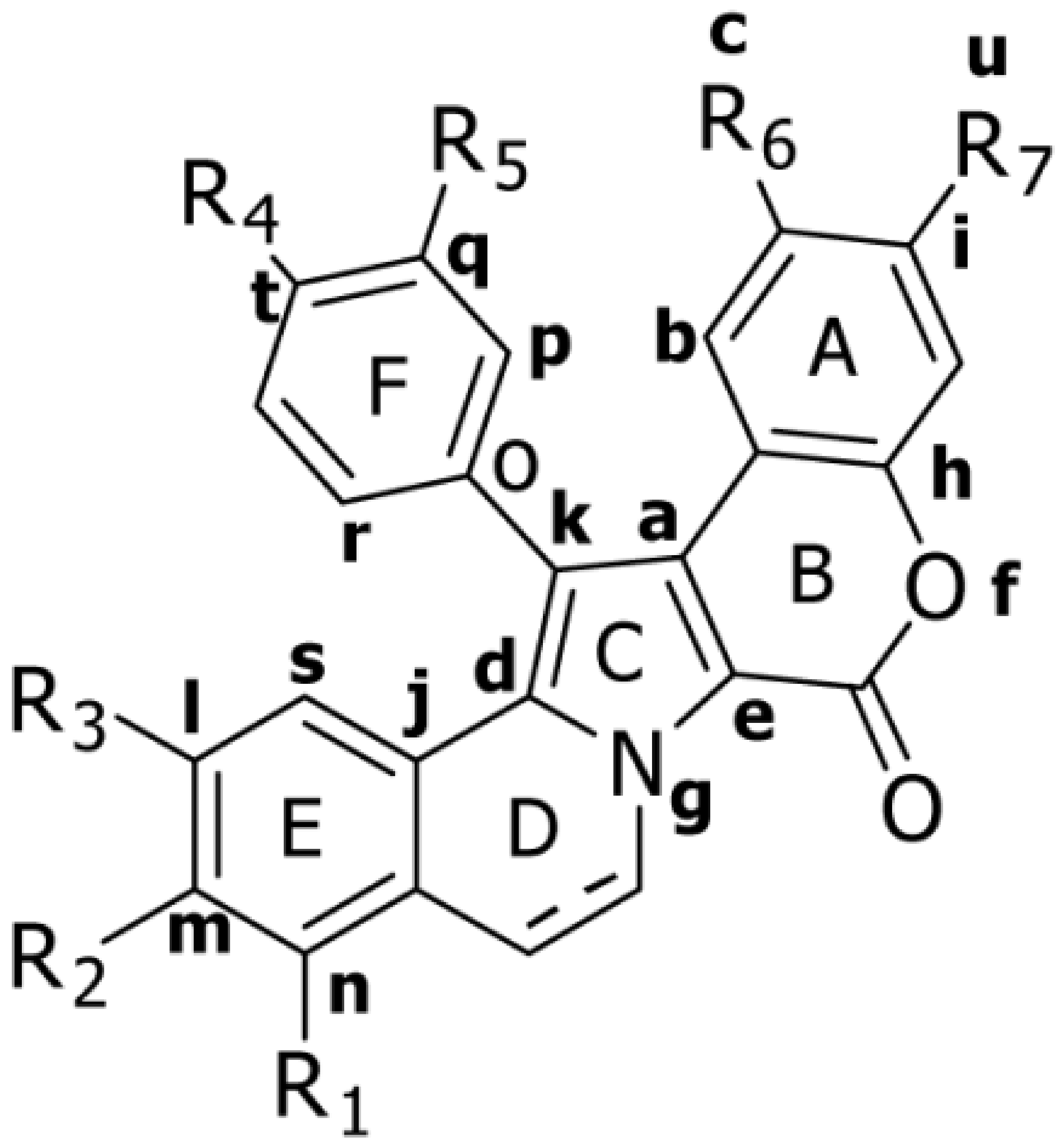
Alignment	First atom	Second atom	Third atom	r^2	$xv\text{-}r^2$
4	j	d	k	0.997	0.995



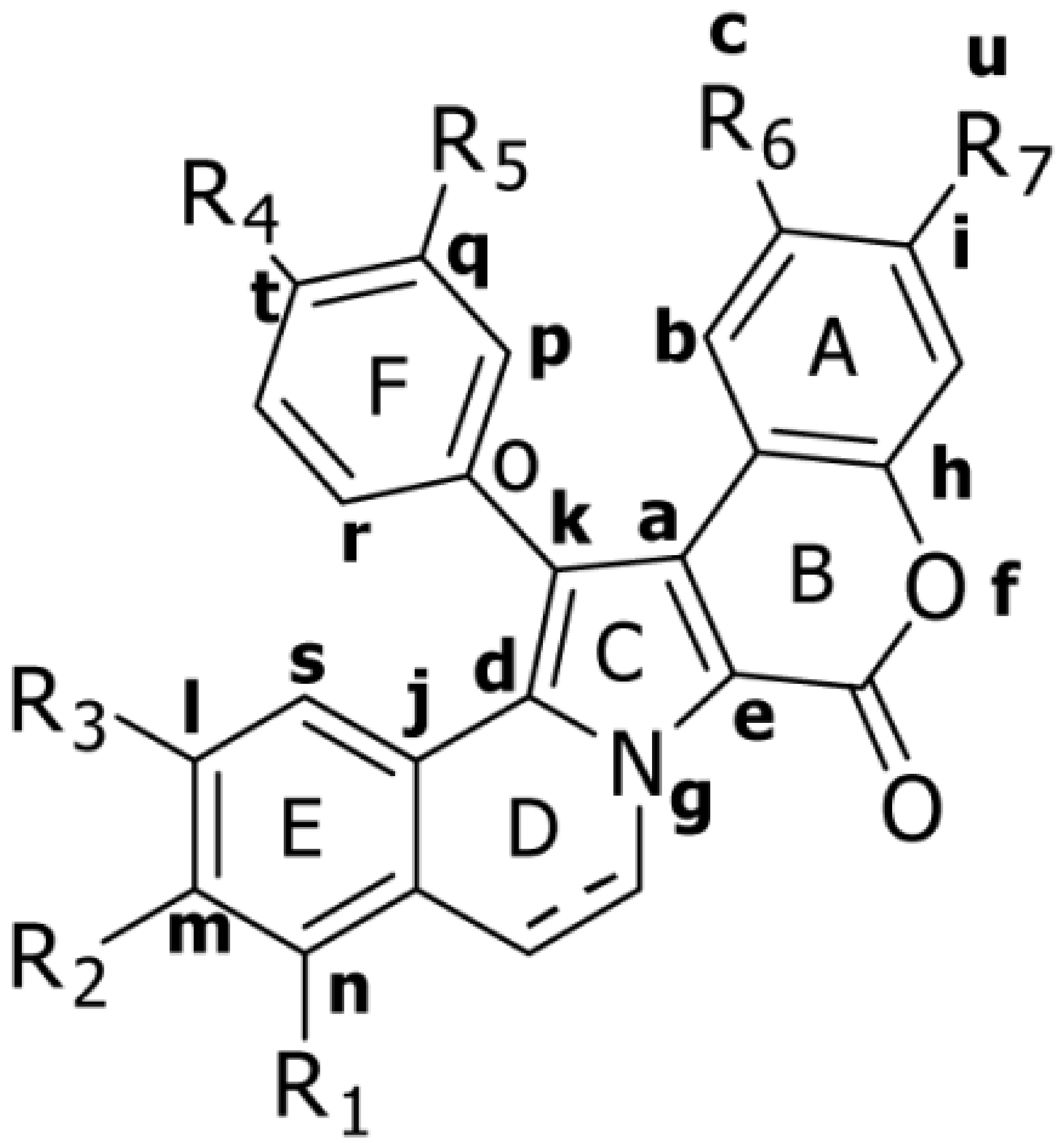
Alignment	First atom	Second atom	Third atom	r^2	$xv\text{-}r^2$
5	l	m	n	0.995	0.984



Alignment	First atom	Second atom	Third atom	r^2	$xv\text{-}r^2$
6	o	p	q	0.997	0.993



Alignment	First atom	Second atom	Third atom	r^2	$xv\text{-}r^2$
7	k	o	r	0.999	0.997



Alignment	First atom	Second atom	Third atom	r^2	$xv\text{-}r^2$
8	s	t	u	0.999	0.995

Table 5

The cross-correlation matrix for the top-ten models of the 25 lamellarins training set

Model no.	1	2	3	4	5	6	7	8	9	10
1	1.000									
2	0.518	1.000								
3	0.760	0.816	1.000							
4	0.578	0.872	0.903	1.000						
5	0.624	0.640	0.806	0.858	1.000					
6	0.525	0.405	0.635	0.631	0.683	1.000				
7	0.601	0.665	0.823	0.887	0.962	0.638	1.000			
8	0.618	0.689	0.837	0.902	0.987	0.654	0.979	1.000		
9	0.618	0.689	0.837	0.902	0.987	0.654	0.979	1.000	1.000	
10	0.608	0.674	0.828	0.892	0.963	0.644	0.998	0.981	0.981	1.000

Table 6

The cross-correlation matrix for the top-ten models of 21 lamellarins training set

Model no.	1	2	3	4	5	6	7	8	9	10
1	1.00									
2	0.91	1.00								
3	0.83	0.97	1.00							
4	0.93	0.93	0.90	1.00						
5	0.92	0.99	0.97	0.93	1.00					
6	0.83	0.97	0.99	0.90	0.97	1.00				
7	0.93	0.93	0.90	0.99	0.93	0.90	1.00			
8	0.94	0.79	0.72	0.93	0.80	0.73	0.93	1.00		
9	0.91	0.98	0.95	0.91	0.98	0.95	0.91	0.80	1.00	
10	0.84	0.97	0.99	0.90	0.97	0.99	0.90	0.73	0.96	1.00

Table 7

The linear cross-correlation matrix of the top-ten models from the four descriptor term 4D-fingerprint models

Model no.	1	2	3	4	5	6	7	8	9	10
1	1.00									
2	0.93	1.00								
3	0.96	0.90	1.00							
4	0.93	0.94	0.93	1.00						
5	0.89	0.94	0.90	0.96	1.00					
6	0.91	0.86	0.94	0.88	0.85	1.00				
7	0.95	0.90	0.99	0.93	0.91	0.94	1.00			
8	0.92	0.91	0.92	0.91	0.89	0.92	0.93	1.00		
9	0.90	0.92	0.90	0.90	0.97	0.87	0.91	0.89	1.00	
10	0.91	0.92	0.92	0.90	0.96	0.89	0.94	0.90	0.98	1.00

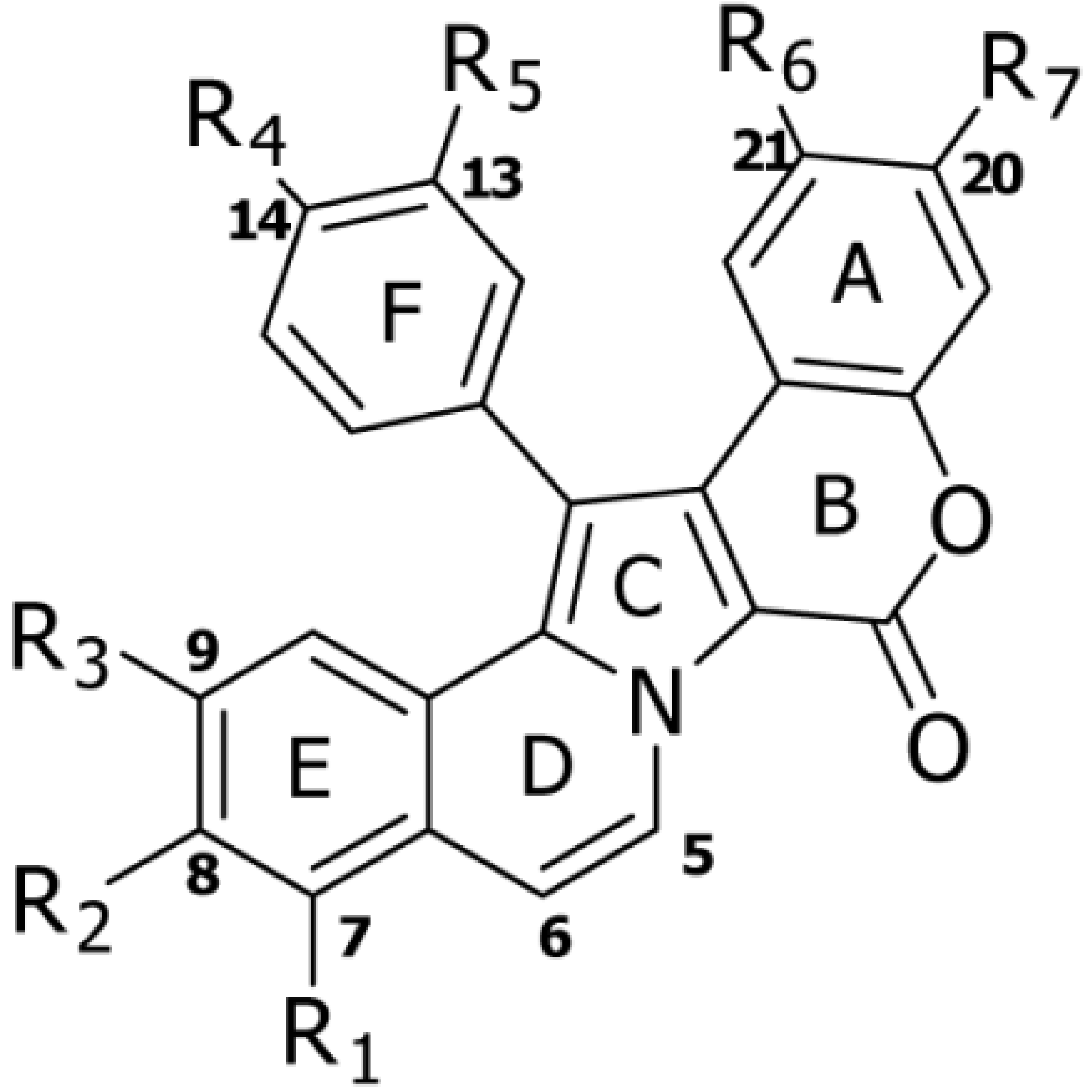
Table 8

The frequency of use and corresponding significance ranking of each descriptor term in 4D-fingerprint virtual screening model

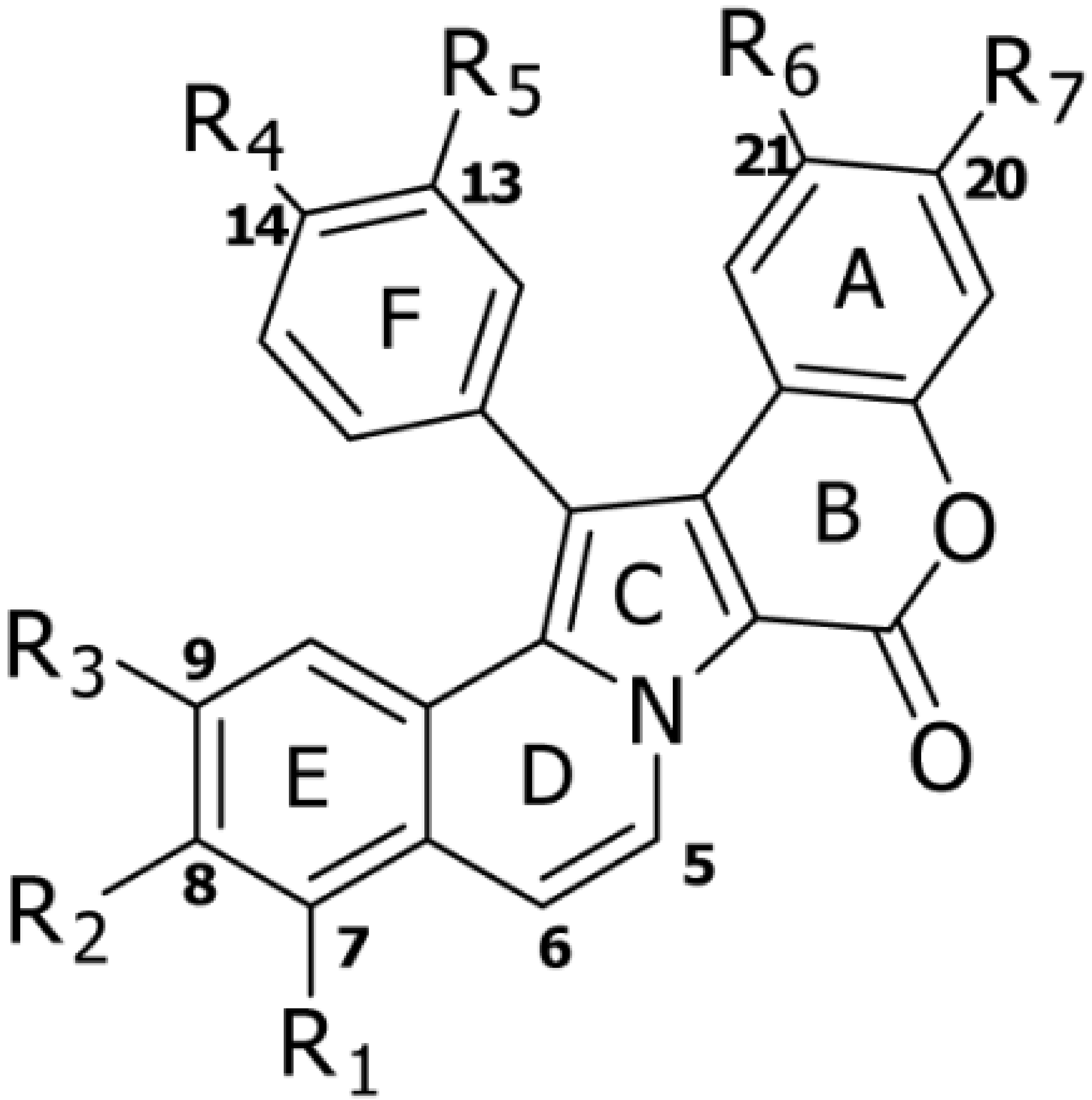
	$\in_7$(any, np)	**$\in_{11}$(any,hs)**	**$\in_3$(p+,aro)**	**$\in_2$(np,hs)**
Frequency	124	128	51	17
Ranking	2	1	5	11

Table 9

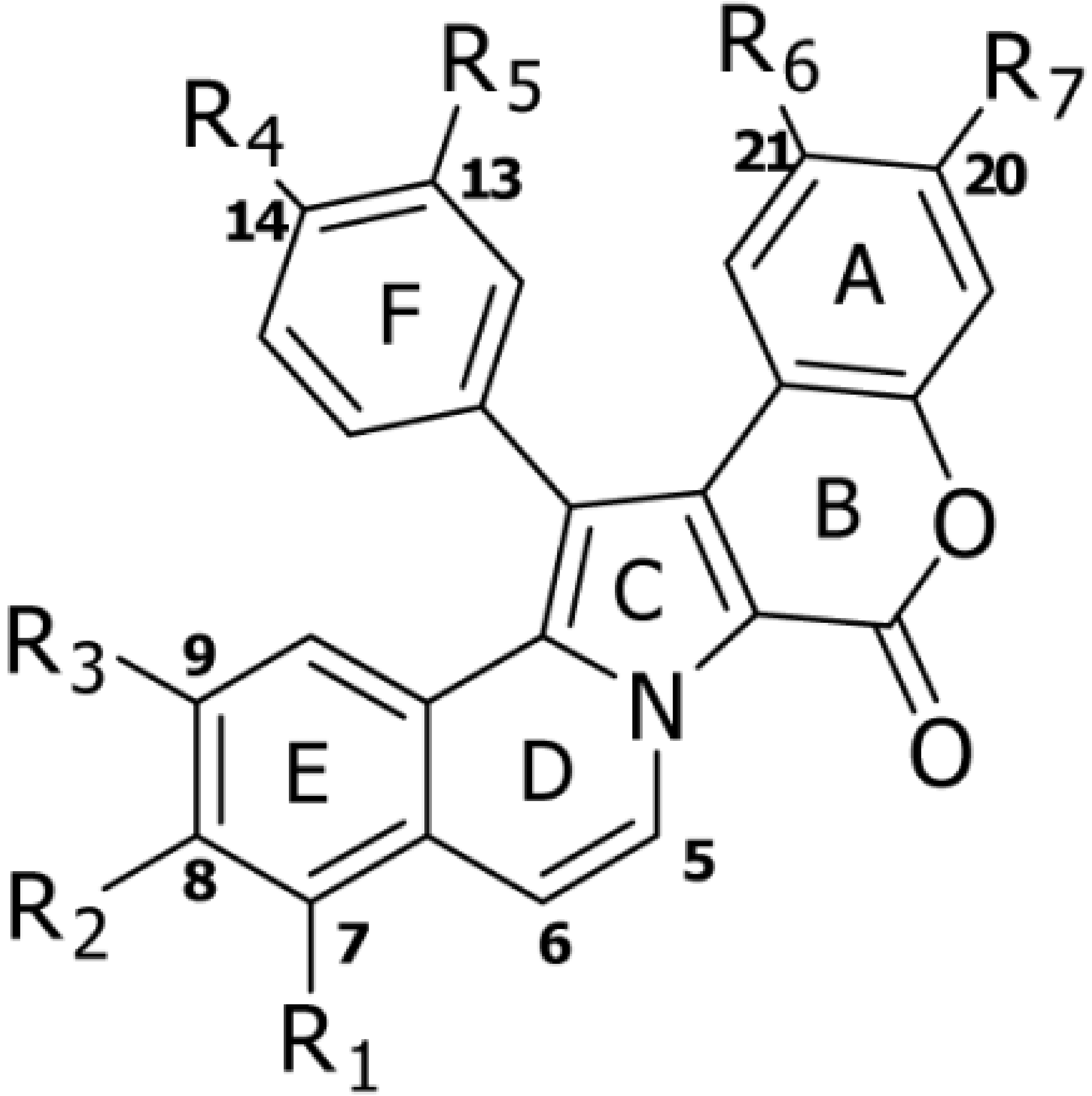
A virtual library of lamellarins built around substitutent variations at R_1-R_5 and the corresponding predicted −logIC_{50} obtained by using equation 12.



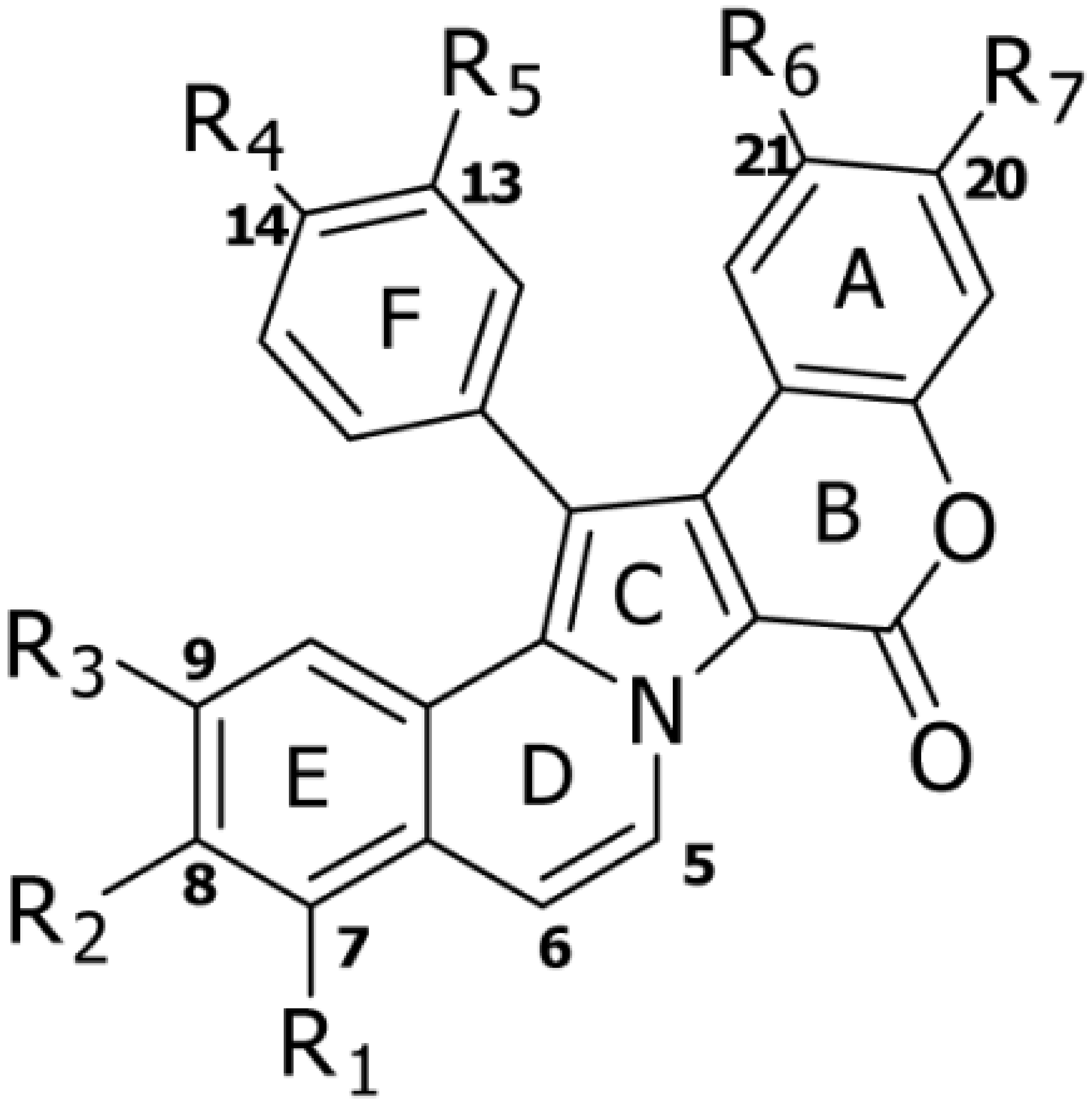
		Substituent group						
Lamellarin	Predicted −logIC_{50}	R_1	R_2	R_3	R_4	R_5	R_6	R_7
ML1	6.43	OH	OH	OMe	OH	OMe	OMe	OH



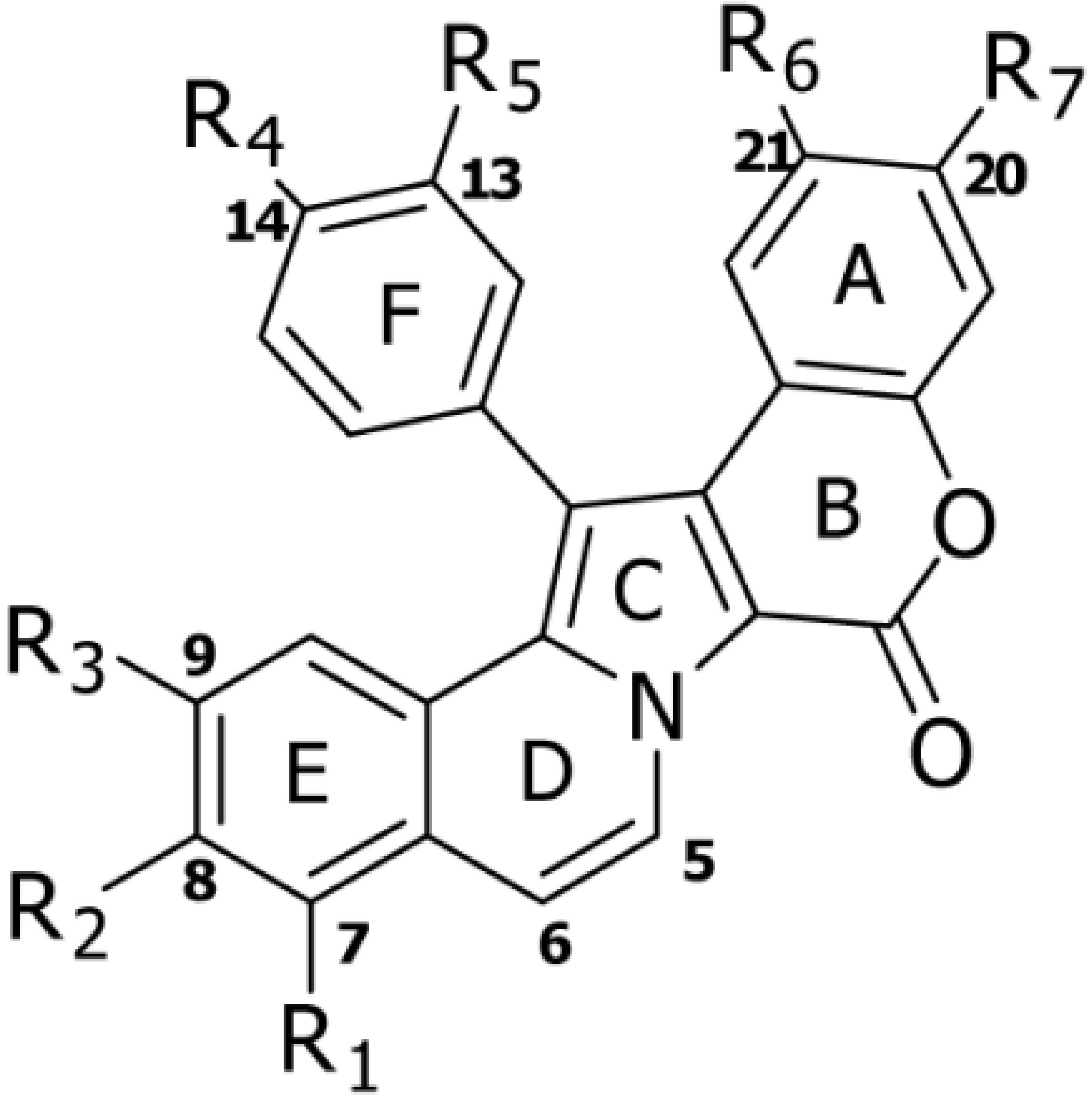
		Substituent group						
Lamellarin	Predicted −logIC_{50}	R_1	R_2	R_3	R_4	R_5	R_6	R_7
ML2	6.22	OMe	OH	OMe	OH	OMe	OMe	OH



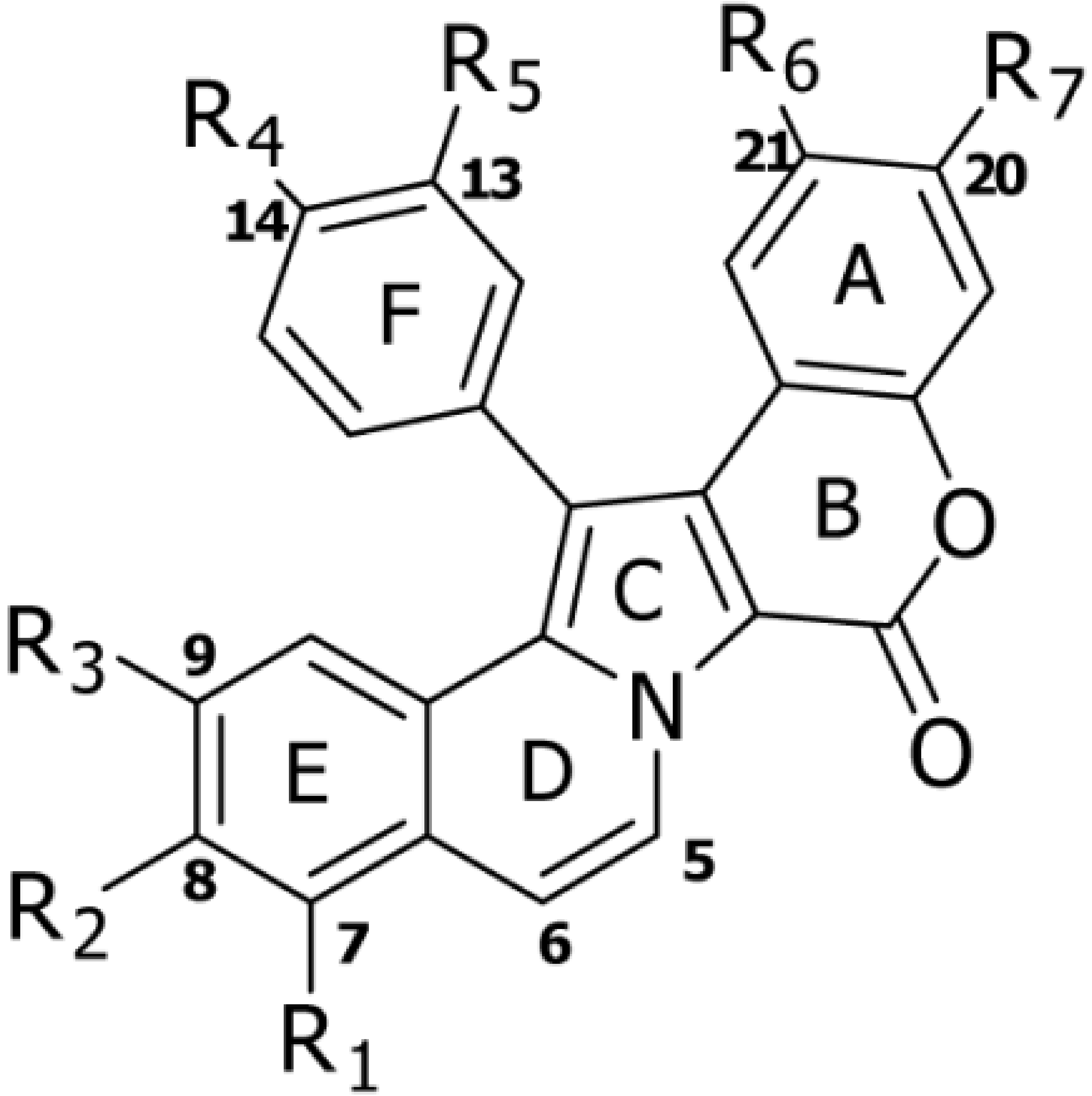
Lamellarin	Predicted −logIC$_{50}$	Substituent group						
		R$_1$	R$_2$	R$_3$	R$_4$	R$_5$	R$_6$	R$_7$
ML3	7.93	H	OMe	OMe	OH	OMe	OMe	OH



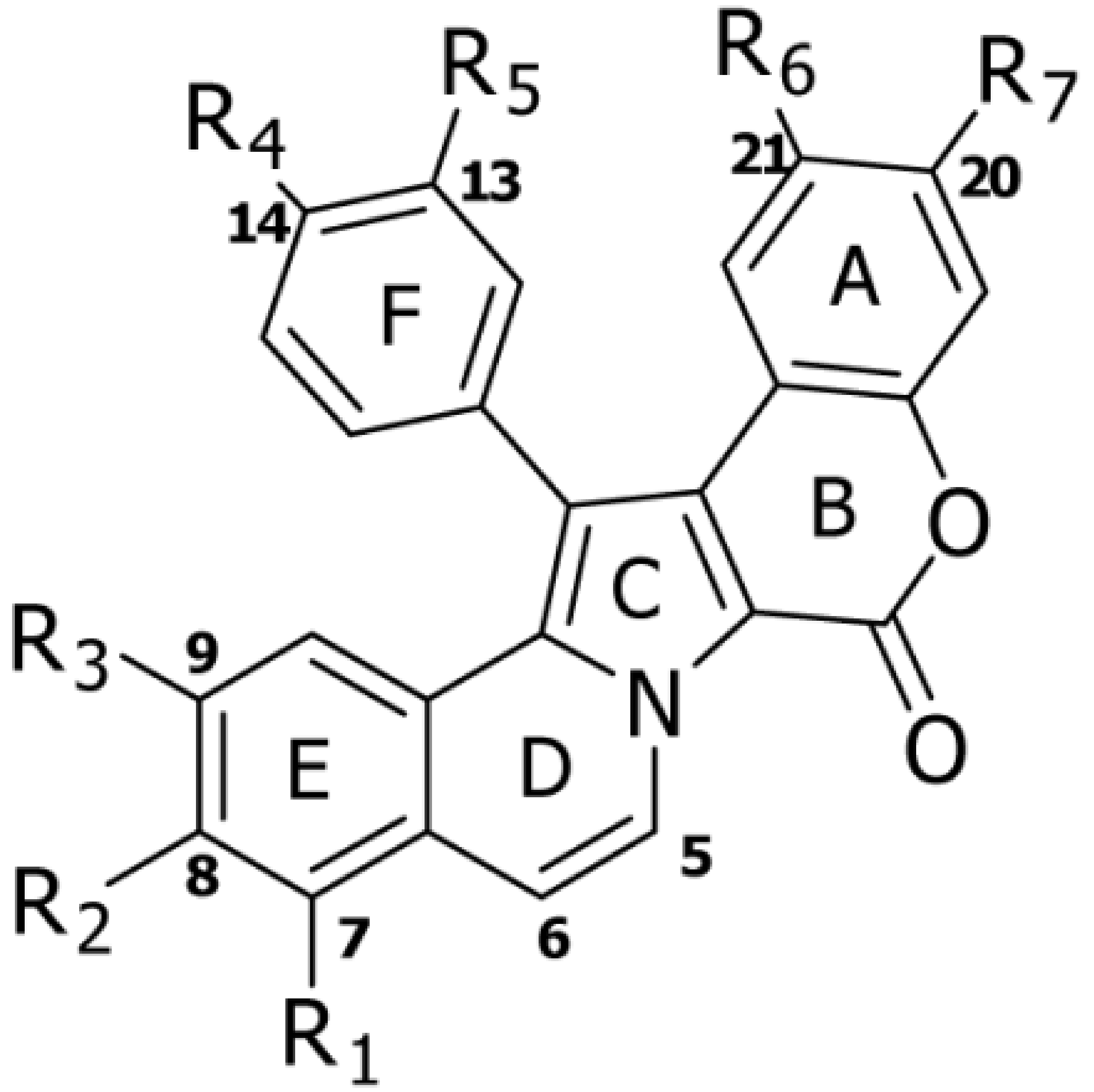
		Substituent group						
Lamellarin	**Predicted −logIC$_{50}$**	**R$_1$**	**R$_2$**	**R$_3$**	**R$_4$**	**R$_5$**	**R$_6$**	**R$_7$**
ML4	8.74	H	H	OMe	OH	OMe	OMe	OH



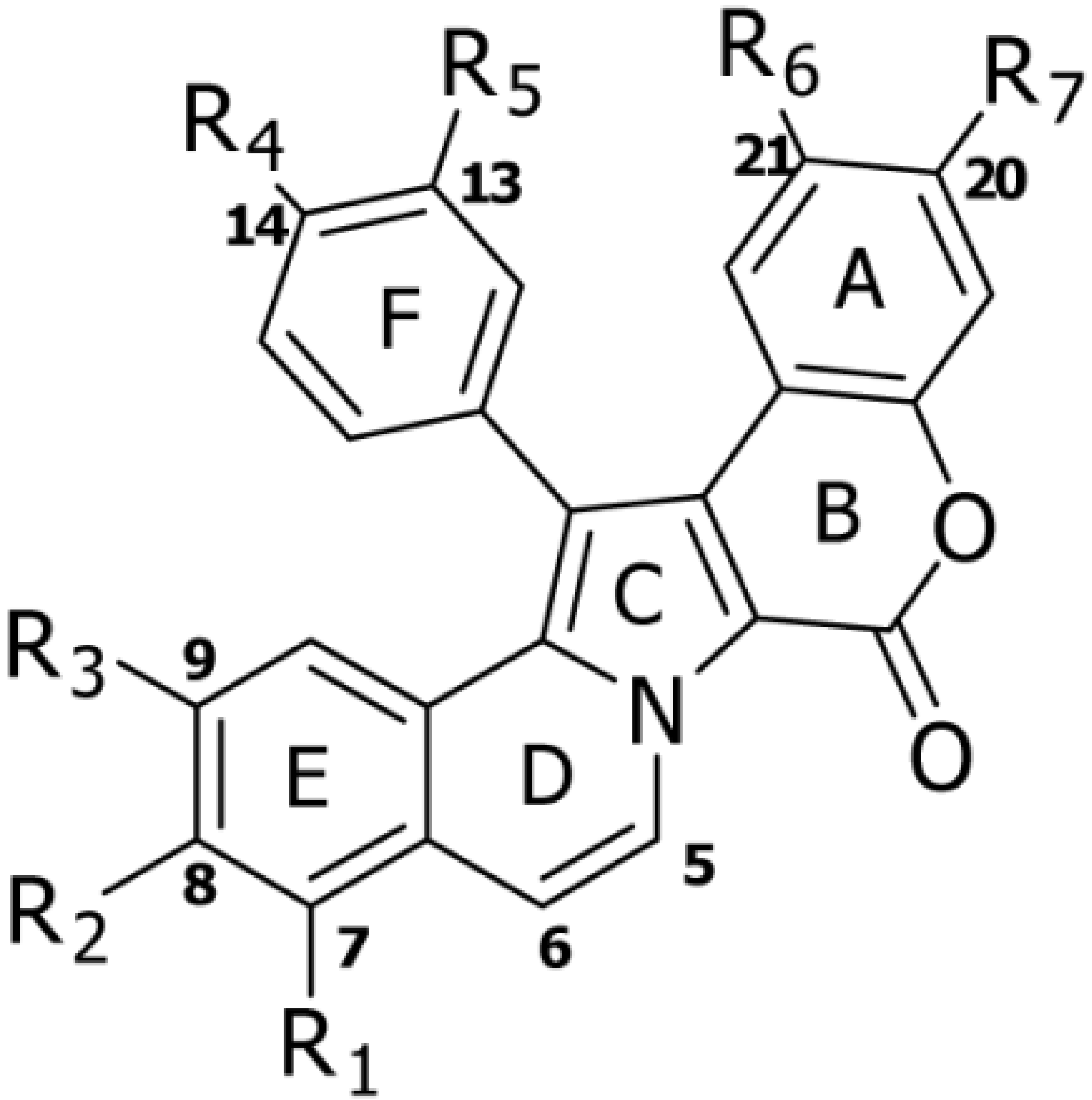
		Substituent group						
Lamellarin	Predicted −logIC$_{50}$	R$_1$	R$_2$	R$_3$	R$_4$	R$_5$	R$_6$	R$_7$
ML5	10.05	H	OH	OH	OH	OMe	OMe	OH



Lamellarin	Predicted −logIC$_{50}$	Substituent group						
		R$_1$	R$_2$	R$_3$	R$_4$	R$_5$	R$_6$	R$_7$
ML6	3.03	H	OH	H	OH	OMe	OMe	OH



Lamellarin	Predicted $-\log IC_{50}$	Substituent group						
		R_1	R_2	R_3	R_4	R_5	R_6	R_7
ML7	8.54	H	OH	OMe	H	OMe	OMe	OH



Lamellarin	Predicted −logIC$_{50}$	Substituent group						
		R$_1$	R$_2$	R$_3$	R$_4$	R$_5$	R$_6$	R$_7$
ML8	7.21	H	OH	OMe	OH	H	OMe	OH

Table 10

The linear cross-correlation matrix of the predicted −logIC_{50} values of the RI-4D-QSAR model (1), the 4D-fingerprint model (2), and the observed cytotoxicity −logIC_{50} values (3)

	1	**2**	**3**
1	1		
2	0.797	1	
3	0.972	0.823	1