# Prediction of Anti-HIV-1 Activity of a Series of Tetrapyrrole Molecules

Rozália Vanyúr, Károly Héberger,* and Judit Jakus

Institute of Chemistry, Chemical Research Center, Hungarian Academy of Sciences, P.O. Box 17,
H-1525 Budapest, Hungary

Received January 25, 2003

Anti-HIV-1 activities of 20 tetrapyrroles (hematoporphyrin derivatives, meso-tetraphenylporphyrins, a chlorin, and a phthalocyanine) were predicted based on their molecular structures using artificial neural networks. The molecular structures were optimized by HyperChem program using MM+ molecular mechanics and conformational search for the global minimum conformer. Eighty-seven theoretical descriptors were calculated for characterization of molecular structures. The network architecture was optimized, and suitable descriptors were selected applying a novel variable selection method. The 3DNET program was used for the calculation of descriptors and for neural network computations. The reliability of models was tested by randomization of biological activity data, leave-one-out, leave-n-out cross-validation, and external validation process. The predictive ability of the artificial neural network was compared to other model building methods, like multiple linear regressions and partial least squares projection to latent structures. For prediction of anti-HIV-1 activity, the artificial neural network gave the best results at cross-validation processes and at external validation as well. We built four nonlinear models with good predictive ability in all validation steps, which can be applied to predict the anti-HIV-1 activity of tetrapyrrole-type compounds in a much better way than with any other three-dimensional quantitative structure−activity relationship methods published to date.

## INTRODUCTION

Porphine-based and related molecules are promising agents for therapy of several diseases. Photodynamic therapy (PDT), for example, uses a combination of a photosensitizing agent and visible light for treatment of many kinds of diseases such as solid tumors, age related macular degeneration, and others.[1] A wide range of tetrapyrrole-type compounds such as porphyrins, chlorins, benzochlorins, bacteriochlorins, pheophorbides, phthalocyanines, and naphthalocyanines has been studied[2] as photosensitizer candidates to date. The photodynamic sterilization of blood and blood products is another possible application for these types of compounds.[3] Several viruses, bacteria, and other parasites can be transmitted by blood or blood products causing unexpected and irreversible life-threatening damage to people in need of other medical treatments. The most frightening transfusion-transmitted disease is the human immunodeficiency virus (HIV-1) infection. The need for an effective prevention and treatment of this disease makes tetrapyrrole-type sensitizers to be important alternative compounds in antiviral drug development. As a result of previous research, it has been shown that the benzoporphyrin derivative monoacid A (BPD-MA)[4] and a chlorin-type compound named CHVD[5] can inactivate HIV-1 virus in blood through photodynamic treatment.

In addition to PDT, it turned out that porphyrins are also able to prevent HIV-1 virus infection on their own, without any light.

One of the possible ways for HIV virus to infect the target cell is its binding to the T cell receptor protein CD4.[6] Several tetrapyrroles can inhibit infection preventing the interaction of viral envelope protein gp120 with CD4[7] by binding to the viral protein. For instance, polyanionic derivatives of porphyrins such as carboxylated[8] and polysulfonated[9] ones had also been shown to have high inhibitory activities.

Because of the wide range of possible applications and the large number of synthesized tetrapyrrole molecules, studies on the relationship between their chemical structures and biological activities appear to gain more and more importance. These analyses are potentially powerful tools, since they can provide specific information based on the structures and parameters of the compounds involved, which in turn may help in suggesting new strategies for the synthesis of novel antiviral drugs of higher efficacy.

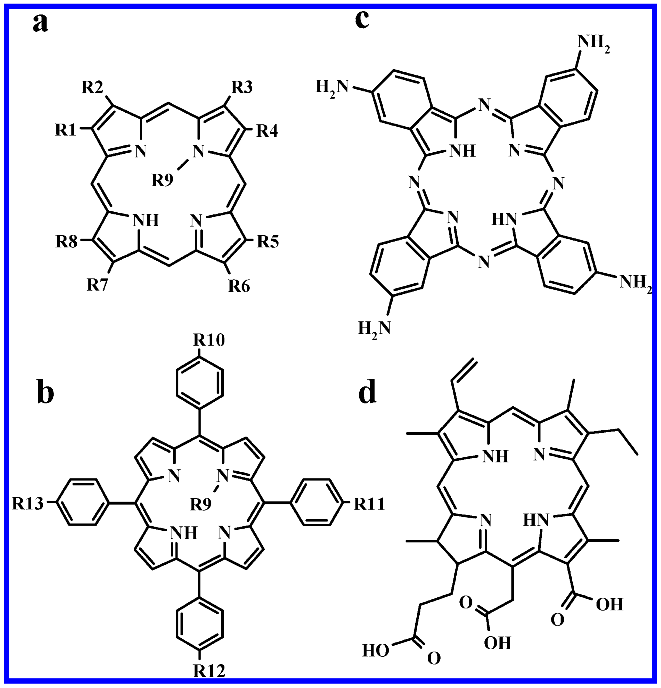
There are several structure−activity relationship studies on photosensitizers and their biodistribution,[2,10] photodynamic activity during PDT,[11−16] and photohaemolysis,[17] but none of these descriptive studies are quantitative and could be used to predict the biological activities of new compounds. Lipophilicity can be used to describe nonspecific biological activities for certain groups of compounds. It can be calculated based on the chemical structures of the given molecules. The first quantitative structure−activity relationship (QSAR) studies on tetrapyrroles used lipophilicity as an independent variable to describe PDT activity.[18] It was reported[18,19] that PDT activities of pyropheophorbide derivatives are a nonlinear function of lipophilicity. These studies used a quasi-mechanistic, one-variable nonlinear activity-lipophilicity relationship model.

Twenty-one porphyrins have been examined as anti-HIV agents[20] in the first QSAR study that applied a three-dimensional (3D) QSAR model to photosensitizers of a wide range of tetrapyrrole compounds. The CoMFA model used

* Corresponding author phone: 36-1-438-4143; fax: 36-1-325-7554; e-mail: heberger@chemres.hu.

10.1021/ci0304627 CCC: $25.00 © 2003 American Chemical Society
Published on Web 09/18/2003

**Chart 1**



in the study, with a leave-one-out cross-validated predictive value of $Q^2 = 0.59$, could be used as a predictive tool for development of further anti-HIV compounds. The analysis has shown that three negatively charged substituents on tetraphenylporphyrins are necessary for a good anti-HIV-1 activity.

In one of our earlier QSAR studies,[21] we compared artificial neural network (ANN) and multiple linear regression (MLR) methods in analyzing a congeneric series of pyropheophorbide derivatives as antitumor PDT agents. From a robust pool of QSAR descriptors, we could select the information-rich ones using neural networks even for a limited, although structurally closely related number of molecules. Our model was able to predict PDT activity of pheophorbide compounds that were similar to the studied ones.

In the present work, we applied the ANN to predict the anti-HIV-1 activity of 20 noncongeneric, structurally different tetrapyrroles (Chart 1 and Table 1) divided into four groups: hematoporphyrins, tetraphenylporphyrins, a chlorin, and a phthalocyanine. We have compared the ANN computation method with other methods such as MLR, partial least squares projection to latent structures (PLS), and the comparative molecular field analysis used by Debnath et al.[20]

## EXPERIMENTAL SECTION

Experimental biological data of compounds were published by Debnath and co-workers.[20] The anti-HIV-1 activity (log $1/EC_{50}$) was expressed as the logarithm of reciprocal concentration ($\mu$M) of porphyrin at which the production of the HIV-1 core protein P24 was reduced to 50% of that detected in HIV-1 infected MT-2 cells. The amount of virus protein was determined by enzyme-linked immunoadsorbent assays. The error of measurement (SD) was the highest for *N*-methyl-protoporphyrin IX ($EC_{50} = 1.030 \pm 0.343$) and the lowest for meso-tetra(4-carboxamidophenyl)-porphine ($EC_{50} = 54.708 \pm 5.518$). The most effective compound in that study was the 5-phenyl-10,15,20-tri(p-carboxyphenyl)-porphine ($EC_{50} = 0.612 \pm 0.056$).

**Table 1.** List of Studied Compounds

| | name of the compound |
|---|---|
| P1 | protoporphyrin IX ($R_1$, $R_3$, $R_5$, and $R_8$: $CH_3$; $R_2$ and $R_4$: $CH{=}CH_2$; $R_6$ and $R_7$: $CH_2{-}CH_2{-}COOH$; $R_9$: H) |
| P2 | *N*-methyl-protoporphyrin IX ($R_1$, $R_3$, $R_5$, $R_8$, and $R_9$: $CH_3$; $R_2$ and $R_4$: $CH{=}CH_2$; $R_6$ and $R_7$: $CH_2{-}CH_2{-}COOH$) |
| P3 | mesoporphyrin IX ($R_2$ and $R_4$: $CH_2{-}CH_3$; $R_1$, $R_3$, $R_5$, and $R_8$: $CH_3$; $R_6$ and $R_7$: $CH_2{-}CH_2{-}COOH$; $R_9$: H) |
| P4 | deuteroporphyrin IX, 2-hydroxyethyl-4-vinyl ($R_1$, $R_3$, $R_5$, and $R_8$: $CH_3$; $R_2$: $HO{-}CH{-}CH_3$; $R_4$: $CH{=}CH_2$; $R_6$ and $R_7$: $CH_2{-}CH_2{-}COOH$; $R_9$: H) |
| P5 | deuteroporphyrin IX, 2-vinyl-4-hydroxymethyl ($R_1$, $R_3$, $R_5$, and $R_8$: $CH_3$; $R_2$: $CH{=}CH_2$; $R_4$: $CH_2{-}OH$; $R_6$ and $R_7$: $CH_2{-}CH_2{-}COOH$; $R_9$: H) |
| P6 | deuteroporphyrin IX, 2,4-bisglycol ($R_1$, $R_3$, $R_5$, and $R_8$: $CH_3$; $R_2$ and $R_4$: $HO{-}CH{-}CH_2{-}OH$; $R_6$ and $R_7$: $CH_2{-}CH_2{-}COOH$; $R_9$: H) |
| P7 | uroporphyrin III ($R_1$, $R_3$, $R_5$, $R_8$: $CH_2{-}COOH$; $R_2$, $R_4$, $R_6$ and $R_7$: $CH_2{-}CH_2{-}COOH$; $R_9$: H) |
| P8 | uroporphyrin I ($R_1$, $R_3$, $R_5$, $R_7$: $CH_2{-}COOH$; $R_2$, $R_4$, $R_6$, and $R_8$: $CH_2{-}CH_2{-}COOH$; $R_9$: H) |
| P9 | pentacarboxyl-porphyrin I ($R_1$, $R_3$, $R_5$: $CH_3$; $R_2$, $R_4$, $R_6$, and $R_8$: $CH_2{-}CH_2{-}COOH$; $R_7$: $CH_2{-}COOH$; $R_9$: H) |
| P10 | hexacarboxyl-porphyrin I ($R_1$, $R_5$: $CH_3$; $R_2$, $R_4$, $R_6$, and $R_8$: $CH_2{-}CH_2{-}COOH$; $R_3$, $R_7$: $CH_2{-}COOH$; $R_9$: H) |
| P11 | heptacarboxyl-porphyrin I ($R_1$, $R_5$, $R_7$: $CH_2{-}COOH$; $R_2$, $R_4$, $R_6$, and $R_8$: $CH_2{-}CH_2{-}COOH$; $R_3$: $CH_3$; $R_9$: H) |
| MP1 | meso-tetraphenylporphyrin ($R_{10}$, $R_{11}$, $R_{12}$, $R_{13}$, and $R_9$: H) |
| MP2 | meso-tetra(4-carboxyphenyl)-porphine ($R_{10}$, $R_{11}$, $R_{12}$, and $R_{13}$: COOH; $R_9$: H) |
| MP3 | *N*-methyl-meso-tetra(4-carboxyphenyl)-porphine ($R_{10}$, $R_{11}$, $R_{12}$, and $R_{13}$: COOH; $R_9$: $CH_3$) |
| MP4 | 5-phenyl-10,15,20-tri(4-carboxyphenyl)-porphine ($R_{10}$, $R_{11}$, and $R_{13}$: COOH; $R_{12}$, $R_9$: H) |
| MP5 | 5-(*p*-chlorophenyl)-10,15,20-tri(p-carboxyphenyl)-porphine ($R_{10}$, $R_{11}$, and $R_{13}$: COOH; $R_{12}$: Cl; $R_9$: H) |
| MP6 | 5-(p-carboxyphenyl)-10,15,20-tri(tolyl)-porphine ($R_{10}$, $R_{11}$, and $R_{13}$: $CH_3$; $R_{12}$: COOH, $R_9$: H) |
| MP7 | meso-tetra(4-carboxamidophenyl)-porphine ($R_{10}$, $R_{11}$, $R_{12}$, and $R_{13}$: $CONH_2$; $R_9$: H) |
| PC1 | 2,9,16,23-tetraamino-phthalocyanine |
| C1 | chlorin $e_6$ |

## THEORETICAL CALCULATIONS

Geometry optimization and conformational search were carried out using $MM^+$ molecular mechanics method by HyperChem for Windows program.[22] All dihedral angles around single bonds of the substituents were randomly varied. A minimum of 256 iterations was completed for each structure with 0.1 kcal/Å convergence criteria for the gradient. Energies of 1024 conformers for every compound were calculated, minimized, and compared. We used the conformers with the smallest energy to represent the 3D structures of the molecules. The calculation of HOMO and LUMO energies of the minimum conformers was carried out by a semiempirical quantum mechanical method: AM1. The two-dimensional (2D) and 3D molecular structures as well as the experimental data on anti-HIV-1 activity reported by Debnath et al.[20] were stored in Isis/Base format.[23]

Theoretical 2D and 3D QSAR descriptors were calculated from molecular data sets (MDL SDF format) using 3DNET program.[24] The descriptors of the presented models are collected in Table 2.

Three compounds from among the 20 tetrapyrroles were set aside for external analysis. They were selected from every

**Table 2.** Descriptors and Independent Variables Calculated by the 3DNET Program Package

| names and abbreviations of descriptors | refs |
|---|---|
| globularity (G) | 26 |
| WHIM descriptors for atomic mass (MASS), atomic position (POS), van der Waals surface (VDW), electronegativity (EN), localized charge, atomic polarizability contribution (POL), atomic electrotopological index (ETPI), atomic lipophilicity contribution (LIPO). Direction dependent eigenvalues ($\lambda_1$ has the direction of the largest, $\lambda_2$ has the direction of the second largest, and $\lambda_3$ has the direction of the smallest elongation of the molecule) are coded in Arabic numbers 1, 2, and 3. For example, POL1 is the first component size directional WHIM index weighted by atomic polarizabilities. Direction independent total molecular size descriptors: $T = \lambda_1 + \lambda_2 + \lambda_3$, $A = \lambda_1\lambda_2 + \lambda_1\lambda_3 + \lambda_2\lambda_3$, $V = T + A + \lambda_1\lambda_2\lambda_3$), and the molecular shape descriptor: $K = [\Sigma_m\lvert(\lambda_m/\Sigma_m\lambda_m)-1/3\rvert]/(4/3)$, $m = 1, 2, 3$. | 27 |
| calculated dipole moment ($\mu$) | 28 |
| hydrogen bond descriptors (HDSA1, HDSA2, HASA1, HASA2) | 29 |
| Bodor descriptors for logP (QN, QO, QNO, QTOT) | 30 |
| degree of chemical bond rotational freedom (DF) | 31 |
| double bond equivalent (DBE = 1 × number of rings + 1 × number of double bonds) | |
| minimum (mESP), maximum (MESP), and average (AESP) of molecular electrostatic potential on the van der Waals surface | 32 |
| minimum (mMLP), maximum (MMLP), and average (AMLP) of molecular lipophilicity potential on the van der Waals surface | 32 |
| energy of the highest occupied molecular orbital (HOMO) | 33 |
| difference between the HOMO and LUMO energies (DE) | |
| electrostatic hydrogen bond basicity (ESB) | 33 |
| electrostatic total hydrogen bond basicity (ESTB = sum of absolute magnitude of (−) charges on H atoms)) | |
| electrostatic hydrogen bond acidity (ESA) | 33 |
| electrostatic total hydrogen bond acidity (ESTA = sum of absolute magnitude of (+) charges on H atoms) | |

molecule set like hematoporphyrins, polycarboxylated-hematoporphyrins, and tetraphenylporphyrins to cover the experimental activity range as uniformly as possible and were not used in either the variable selection, or the model building processes.

**Model Building Procedures.** Multiple linear regression (MLR) calculations were carried out using the Statistica software package.[25] Only forward selection could be applied because the number of descriptors (87) was much higher than the number of compounds. All those descriptors that exceeded the 5% significance level were retained in the model.

Partial least squares projection to latent structures (PLS) calculations using multiple regression analysis were performed using the same Statistica software package and the auto-QSAR module of the 3DNET program. This method produces new variables by a linear combination of the original descriptors and uses them to predict the biological activities. The advantage of this method lies in that it can be used for strongly correlated, noisy data with numerous independent variables (e.g. in case of having more descriptors than compounds). We used the 3DNET program[24] for the neural network computations according to details published in an earlier study of ours.[21] The program contains a fully connected, three-layer, feed-forward computational neural network with back-propagation training. In this net, the transfer function of input and output layers is linear, and the hidden layer has neurons with a hyperbolic tangent transfer function. The relative importance of the input descriptors was calculated based on the weights and the descriptors with less than 5% relative importance (or the last one if its relative importance was higher than 5%) were left out. Then, the calculation was performed again without these descriptors. Descriptor combinations with good descriptive properties (using less than six variables) were used for further studies.

The promising descriptor combinations were selected by leave-one-out cross-validation (LOO) procedure. It means that the first sensitizer is left out, and then a model is built using the remaining compounds. This model is used to verify the calculated biological activity of the sensitizer that was left out. Each molecule is left out once and only once. The goodness of prediction is calculated and characterized by the $Q^2$ value

$$Q^2 = 1 - \frac{\sum_{i=1}^{17}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{17}(Y_i - \bar{Y}_i)^2} \qquad (1)$$

where $i$ = number of tested molecules; $Y_i$ and $\hat{Y}_i$ are the experimental and calculated activities of the left out compound ($i$), respectively; and $\bar{Y}_i$ = average of the left-in experimental activities. The latter quantity changes slightly from run to run.

The reliability of models selected by LOO cross-validation method was proved by shuffling the values for dependent variable (activity data), leave-n-out cross-validation, and external validation. Each test was performed on the same validation sets in case of all nine models listed in Table 3.

During the shuffling procedure, biological activity values were reallocated (mixed together), whereas the other descriptor values remained in the same position. If the statistical properties did not change significantly, then the model before shuffling was not better than the one obtained using random numbers as descriptors.

Leave-n-out (LNO) cross-validation, which is similar to the leave-one-out cross-validation, was carried out eight times with four, variously left-out compounds for each model. Every compound was stored in a validation set at least once.

**Table 3.** Predictive Performance of Models on Anti-HIV-1 Activity

| model no. | model type | descriptors in the model[a] | $R^2$ [b,f] (SD)[c] | leave-one-out $Q^2$ [d] | leave-n-out $Q^2$ [e] | external $Q^2$ |
|---|---|---|---|---|---|---|
| I | MLR | AMASS, MASS1, ESTA | 0.8275 (0.4440) | 0.715 | 0.730 | −1.253 |
| II | MLR | $\mu$, POL1, APOL | 0.7898 (0.4901) | 0.650 | 0.390 | −4.234 |
| III | ANN_3N | KMASS, DF, DBE | 0.9541 (0.1576) | 0.770[g] | 0.747 | 0.788 |
| IV | ANN_4N | HOMO, DF, DBE | 0.9982 (0.0281) | 0.846 | 0.879 | 0.616 |
| V | ANN_3N | DF, DBE | 0.9654 (0.1242) | 0.673[g] | 0.584 | 0.773 |
| VI | ANN_4N | DF, DBE | 0.9877 (0.0787) | 0.475 | 0.342 | 0.700 |
| VII | ANN_4N | DF, ESTA | 0.9654 (0.1368) | 0.567[h] | 0.619 | 0.742 |
| VIII | PLS_9C | HDSA2, HASA2, QN, QO, KMASS, KPOS, KVDW, KEN, KPOL, KETPI, KLIPO | 0.9040 (0.2280) | 0.416 | −0.479 | 0.759 |
| IX | PLS_3C | QO, KVDW, KPOL | 0.6766 (0.4183) | 0.417 | 0.339 | 0.563 |

[a] Abbreviations for descriptors are listed in Table 2. [b] Correlation coefficient between the experimental and the calculated activities. [c] Standard deviations of the calculated activities are in brackets. [d] Predictive ability (leave-one-out cross-validated correlation between the experimental and the predicted activities). [e] Average $Q^2$ value of leave-n-out cross-validation (four compounds were left out eight times). [f] For models I and II without intercept with the *y* axis, $R^2$ represents the position of explained variability around the origin. This value cannot be compared to $R^2$ when the intercept is included. [g] $Q^2$ value of leave-one-out cross-validation without using uroporphyrin III for prediction. [h] $Q^2$ value of leave-one-out cross-validation without using 2,9,16,23-tetraamino-phthalocyanine for prediction.

The best descriptor combination selected by internal validation processes (leave-one-out and leave-n-out cross-validation) was applied to the three compounds, which were put aside at the beginning of the analysis. Always the same training and (internal and external) validation sets were used for model comparison. The calculated external $Q^2$ values (eq 1) for these molecules were used to demonstrate the predictive ability of the selected descriptor combinations.

## RESULTS AND DISCUSSION

The relationship between descriptors calculated based on the chemical structures and the anti-HIV-1 activity of 20 different tetrapyrroles has been analyzed. The predictive model building abilities of three methods—(MLR, PLS, and ANN)—were compared.

In Table 3 we listed the best MLR models built in two ways: (i) using standard forward selection (model I) and (ii) using principal component analysis (PCA) to improve the variable selection process (model II).

Model I used the direction dependent eigenvalue, a WHIM descriptor weighted by atomic mass, MASS1, the total molecular size WHIM descriptor for atomic mass, AMASS, and the electrostatic total hydrogen bond acidity, ESTA (cf. Table 2).

In building of model II, the whole input data matrix (i.e. all descriptors and the anti-HIV-1 activities together) has been subjected to principal component analysis. Two principal components were retained in this model. We have selected descriptors that were in proximity to the points of anti-HIV-1 activity (proximity was evaluated by visual inspection in the 2D factor loading space). Three descriptors (calculated dipole moment, $\mu$; WHIM descriptors weighted by atomic polarizability, APOL and POL1) have been chosen by the standard forward stepwise procedure from among the PCA-preselected descriptors (DHL, $\mu$, HDSA2, ESTB, QO, KMASS, KPOS, POL1, POL3, TPOL, VVDW, APOL, KETPI, KLIPO, and mESP [for explanation of descriptors see Table 2]).

The square of correlation coefficient ($R^2$) indicates the goodness of fit (i.e. description), which was fairly high in both cases, showing that the descriptors in both models explain the anti-HIV-1 activity with very good efficiency. The LOO $Q^2$ value in both MLR models (models I and II) indicates acceptable predictive ability.

In Table 3, we also listed the promising ANN models (models III, IV, V, VI, and VII) for predicting log(1/$EC_{50}$) activity of the studied compounds. The goodness of fit was very good in all cases shown by the $R^2$ value being close to 1. During the leave-one-out cross-validation process, the models using three hidden neurons in the network architecture could not predict the activity value of uroporphyrin I (P8 compound in Table 1) independently from the number of variables applied in the models. The LOO calculation resulted in a better predictive ability of these ANN models when P8 compound was used only for training but not for prediction: LOO $Q^2 = 0.549$ for model III and $Q^2 = 0.345$ for model V for all 17 compounds (data not shown), while LOO $Q^2 = 0.770$ for model III and $Q^2 = 0.673$ for model V when P8 compound was left out.

In model IV, both parameters—the goodness of fit ($R^2$) and the predictive ability ($Q^2$)—gave high values. The two-descriptor model (model VI) could describe the anti-HIV activity well (high $R^2$ value), but was not suitable for prediction: $Q^2 = 0.475$. Model VII could not predict the biological activity of the 2,9,16,23-tetraamino-phthalocyanine compound during the LOO cross-validation. Nevertheless, the $Q^2$ value increased close to 0.6 even when the phthalocyanine was left out.

Every ANN model used the degree of chemical bond rotational freedom (DF) as an independent variable. This descriptor could describe the anti-HIV-1 activity of hematoporphyrin derivatives and of the chlorin without any other variable. However, it could not differentiate among the meso-tetraphenylporphyrins (Figure 1). Using complementary descriptors, like the double bond equivalent (DBE) and the energy of the highest molecular orbital (HOMO), resulted in a model, which was suitable to predict the biological activities of tetraphenylporphyrins as well (model IV). The DBE descriptor divided the compounds into four groups. Within these groups the phthalocyanine and tetraphenylporphyrins were divided into various subgroups. This variable was able to make some distinction among the tetraphenylporphyrins (Figure 2) just like the HOMO energy (Figure 3). The molecular shape descriptor (KMASS) weighted by atomic mass can only separate the phthalocyanine from the

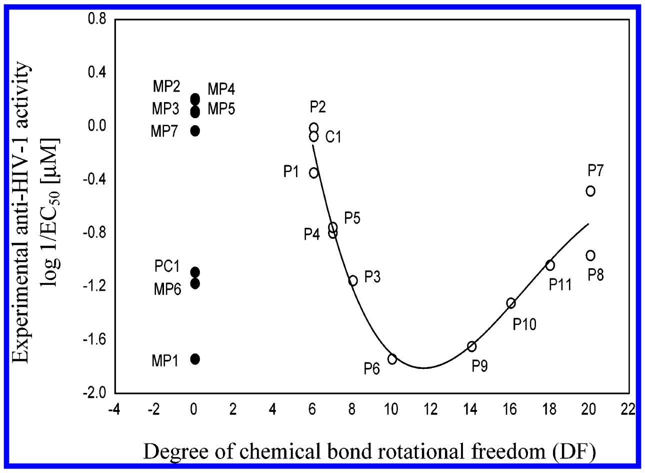


**Figure 1.** Experimental anti-HIV-1 activity versus degree of chemical bond rotational freedom data for tetrapyrroles: open circles − hematoporphyrins and the chlorin; full circles − tetraphenylporphyrins and the phthalocyanine.

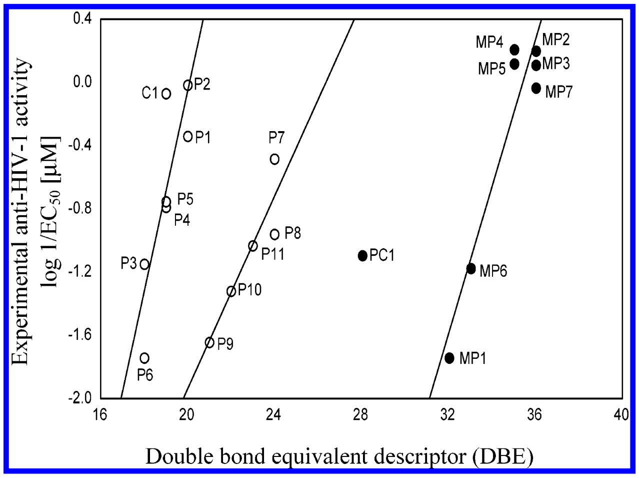


**Figure 2.** Experimental anti-HIV-1 activity versus double bond equivalent data for tetrapyrroles: open circles − hematoporphyrins and the chlorin; full circles − tetraphenylporphyrins and the phthalocyanine.

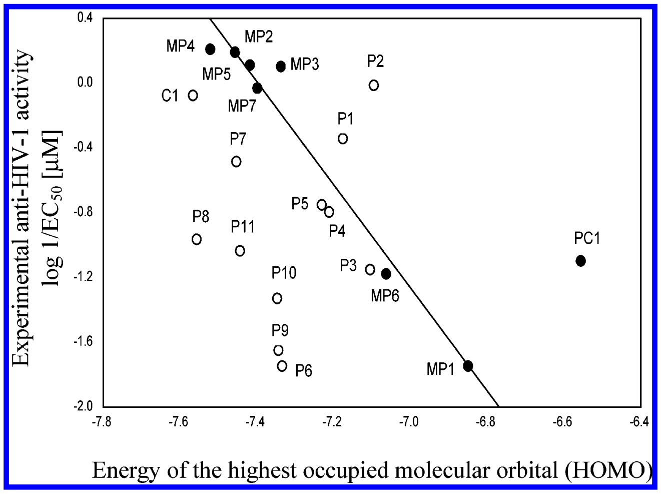


**Figure 3.** Experimental anti-HIV-1 activity versus energy of the highest occupied molecular orbital data for tetrapyrroles: open circles − hematoporphyrins and the chlorin; full circles − tetraphenylporphyrins and the phthalocyanine.

tetraphenylporphyrins. The electrostatic total hydrogen bond acidity (ESTA) versus anti-HIV-1 activity data suggests that there is an ideal electrostatic acidity range (between 0.9 and 1.1) for optimal anti-HIV-1 agents (Figure 4). This descriptor, however, could not be used on its own to predict the biological activity because of the resulting high standard error of description.

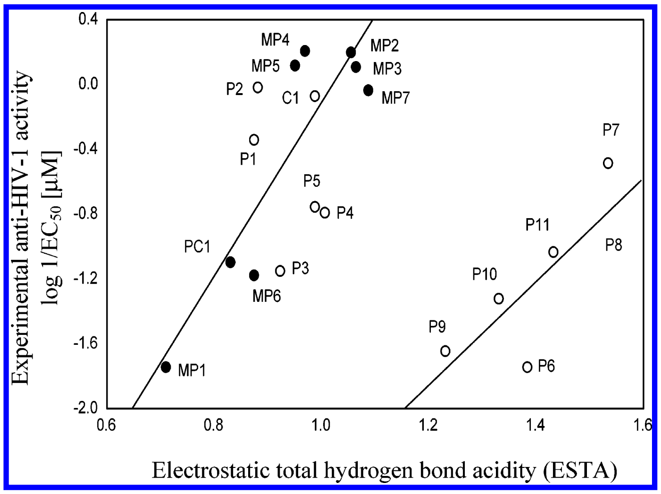


**Figure 4.** Experimental anti-HIV-1 activity versus electrostatic total hydrogen bond acidity data for tetrapyrroles: open circles − hematoporphyrins and the chlorin; full circles − tetraphenylporphyrins and the phthalocyanine.

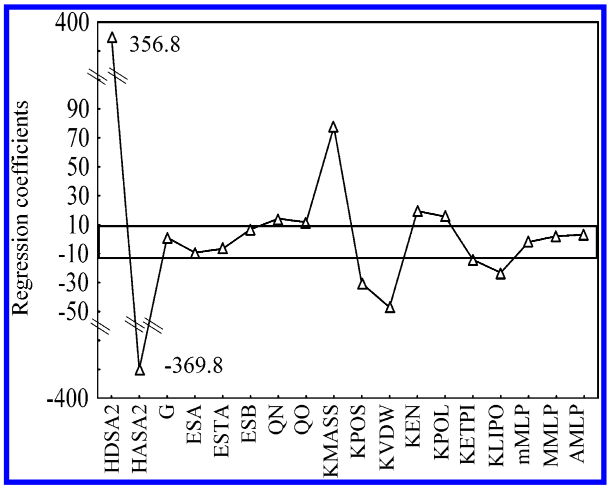


**Figure 5.** Regression coefficients of descriptors in PLS model using nine components of the combination of 18 variables. The lines show the ±10 range. Descriptors with regression coefficient values lower than −10 or higher than 10 were used for building of the PLS model VIII.

Models VIII and IX in Table 3 are PLS models. We selected three PLS components from a combination of 87 descriptors in the first step of the analysis. Based on the regression coefficients (higher than 0.1), 18 descriptors (HDSA2, HASA2, G, ESA, ESTA, ESB, QN, QO, KMASS, KPOS, KVDW, KEN, KPOL, KETPI, KLIPO, mMLP, AMLP, and MMLP; for explanation see Table 2) were retained for the next PLS computation. The new model included nine PLS components from a combination of these 18 descriptors.

In the third step, the number of descriptors was further reduced to 11 (HDSA2, HASA2, QN, QO, KMASS, KPOS, KVDW, KEN, KPOL, KETPI, and KLIPO). These descriptors have regression coefficient values higher than 10 horizontal or lower than −10 in the second model (Figure 5). The new model was built using nine PLS components from the combination of these 11 descriptors (model VIII

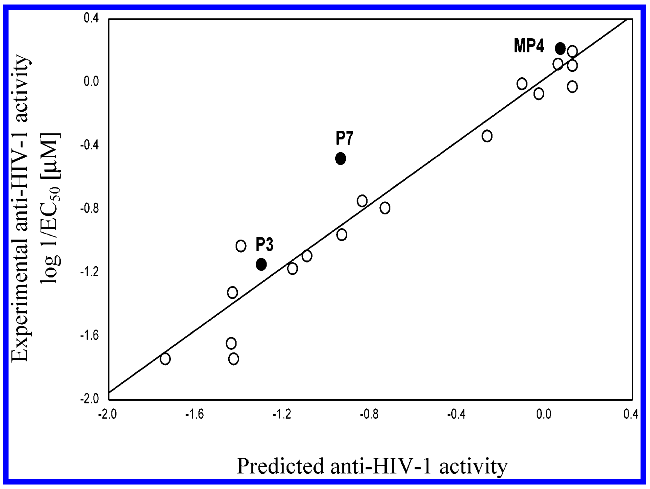


**Figure 6.** Predictive ability of ANN model III on external validation set of compounds used for: open circles − training; full circles − prediction. The external $Q^2$ value of this model was 0.788 explaining 78.8% of the changes in the activity.

in Table 3). This model can describe the structure−anti-HIV-1 activity relationship, but its prediction is not reliable because of the negative $Q^2$ value after leave-n-out cross-validation. The model optimization was continued by the PLS method in an auto-QSAR module of the 3DNET program using leave-one-out $Q^2$ maximization process. The resulting model used 3 out of 11 descriptors of model VIII. Thus, PLS model IX was built using three PLS components by combining QO, KVDW, and KPOL descriptors. This is actually a model equivalent to the MLR calculations. Neither the fit ($R^2 = 0.677$), nor the predicting ability (LOO $Q^2 = 0.417$) of this model was good enough, however.

Substantial differences between the $R^2$ and $Q^2$ values for the models should be a warning about the predictive ability of the models.

**Validation of the Predictive Ability of the Models.** Shuffling of the activity values three times yielded a negative average cross-validated $Q^2$, which is unacceptable for all models. The statistical properties ($Q^2$ and residual error) changed significantly. This proves that there is a definite role of the selected descriptors in describing the biological activity. The extremely bad fit indicates that the models before shuffling are much better than the ones obtained using random numbers as activities.

Based on the leave-one-out cross-validation results, six models (from model I to V, and model VII) can be considered as suitable for prediction of anti-HIV activity ($Q^2$ near or above 0.6).

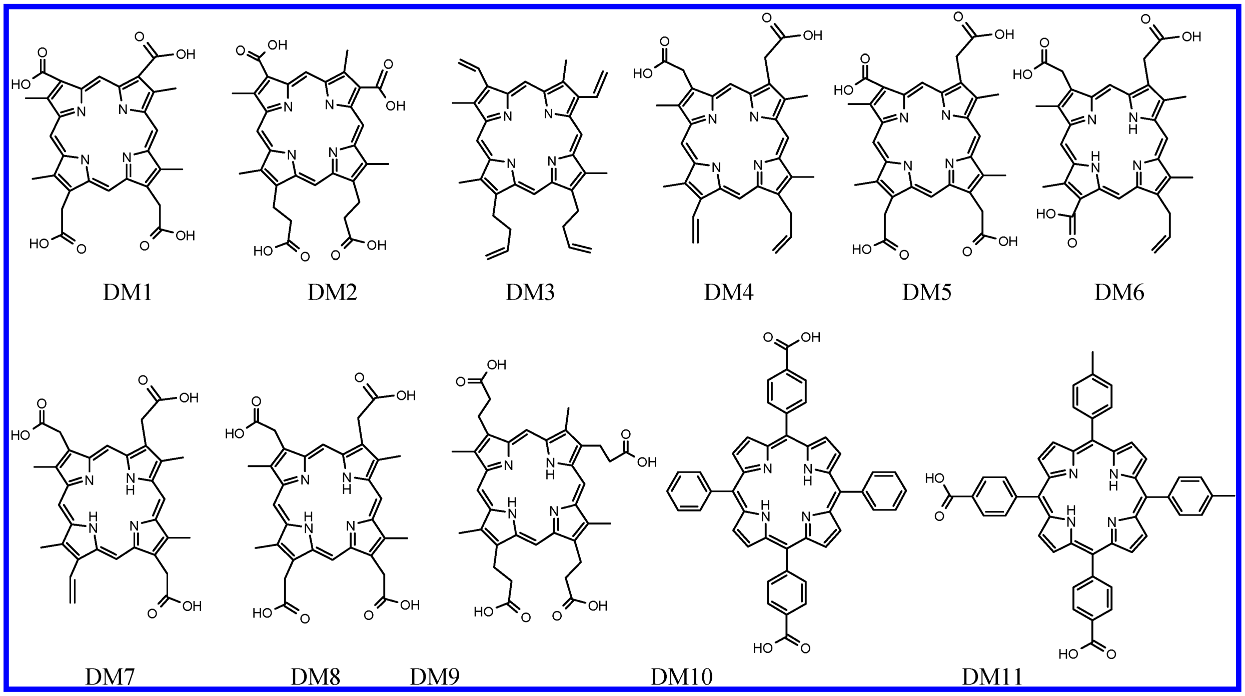
Five models (MLR model I and two three-descriptor ANN models III and IV as well as two two-descriptor models V and VII) were selected with a $Q^2$ near or above 0.6 after the leave-n-out procedure (see Table 3).

Results of the external validation process on 17 chosen training compounds can also be seen in Table 3. Despite good results obtained in leave-one-out cross-validation and leave-n-out procedure, the MLR models could not predict the anti-HIV activity of the three left out compounds. The best prediction was achieved by a three-descriptor ANN model, model III (Figure 6). The external $Q^2$ value of this model was 0.788, explaining 78.8% of the changes in the activity, but all of the ANN models were able to predict the studied activity with a $Q^2 > 0.6$. Because of the small number of compounds in external validation set to predict the anti-HIV-1 activity of tetrapyrrole-based compounds, we recommend using an average of predicted activities calculated by the four ANN models (models III, IV, V, and VII), which show good predictive ability at every validation step.

**Comparison of the Model Building Processes.** Four methods as CoMFA, MLR, PLS, and ANN were compared on building promising models for prediction of anti-HIV-1 activity of tetrapyrrole compounds. The models built by MLR could not predict the biological activities of compounds in the external validation set, while PLS models gave bad results during the internal validation processes. The most reliable method tested in our study was ANN. The model built by CoMFA is a useful one, but mainly for descriptive purposes, although it has been used as a predictive tool for further anti-HIV-1 drug development of porphyrin-type compounds by Debnath et al.[20] Structural features describing the studied activity of tetrapyrroles in Debnath's paper are (i) electrostatics, (ii) placement of charges, and (iii) steric factors. In our study, descriptors in the promising ANN models characterized similar properties: ESTA is a descriptor used in QSAR representing electrostatics, HOMO and ESTA are related to charge, while DF and DBE contains steric factors in a condensed form. Molecules in the data set differ from each other in the number of carboxyl groups, which are responsible for the different electrostatic properties and charge density. The length of carboxylated side chains varies the steric properties. Our models perform much better from the point of view of prediction than the model of Debnath et al. Entirely different sets of molecular descriptors may give good correlations describing or predicting activities for a series of compounds. Table 3, for instance, shows several examples of different QSAR methods giving extremely high correlation coefficients ($R^2$) using different sets of descriptors to explain anti-HIV activities. Not all of these methods (and/or models) could be used for prediction because their $Q^2$ values were not satisfactory, that is, the requirements for a good prediction are much stricter than for description. The reliability and validity of the calculations given in the paper of Debnath and co-workers was tested by a leave-one-out cross-validation method. In our manuscript, we used this and two more up-to-date validation methods. Table 3 shows that leave-n-out and external validations discarded several models that gave good $R^2$ and $Q^2$ in the first steps but could not be used for prediction. Thus, the reliability and validity of the proposed models in this work is much higher than of those presented earlier. ANN calculations seem to be suitable for comparison of tetrapyrrole compounds and for a wide variety of biological activities, introducing a better and a more efficient way not only to describe but also to predict the biological activities of porphyrin-type molecules. An additional advantage of using ANN for QSAR calculations is that it is cost-effective and available for everybody (certain versions can be downloaded from the net), unlike the very expensive CoMFA program used by Debnath et al. Moreover, CoMFA contains a PLS step, which is essentially a linear method in the sense that the variables are combined in a linear way, whereas ANN can handle many different complicated nonlinear relationships.

**Prediction of the anti-HIV-1 Activity of Designed Compounds.** Our models suggest that compounds with DF $\leq 6$ and DBE $\approx 20$ must have the highest anti-HIV-1 activity

**Chart 2**



among the hematoporphyrin derivatives. Based on this assumption, new hypothetical compounds (nine hematoporphyrins and two tetraphenylporphyrins, Chart 2) were designed as examples of expected biological activities. Two of the hematoporphyrins (DM8 and DM9) were designed to have low biological activities, that is, with DF higher than 6. Six molecules were designed to be promising virus inhibitors with DF = 6 value, and in one case an extrapolation was performed with DF = 4 (DM1). The original data set included tetraphenylporphyrins with zero, one, three, and four carboxyl groups. We designed two compounds with two carboxyl groups. An average of predicted activities has been calculated based on the best ANN models (model III, IV, V, and VII) that showed good predictive ability at every validation step. Results are shown in Figure 7. The biological activities of the two designed tetraphenylporphyrins (DM10 and DM11) were predicted to be between those of compounds with one (MP6) and three (MP4 and MP5) carboxyl groups. From the obtained results it seems that three carboxyl groups are necessary for a molecule to show high anti-HIV-1 activity. We found the same calculated effect with the designed hematoporphyrin derivatives. Molecules with zero (DM3) and two carboxyl groups (DM4) were predicted to have lower activities among the six promising compounds (DM2−DM7). The predicted activities of the remaining four molecules with three or four carboxyl groups (DM2, DM5, DM6, and DM7) resulted to have higher activity than that of protoporphyrin IX (MP1), which was the starting compound used for molecular design with an activity of −0.34.

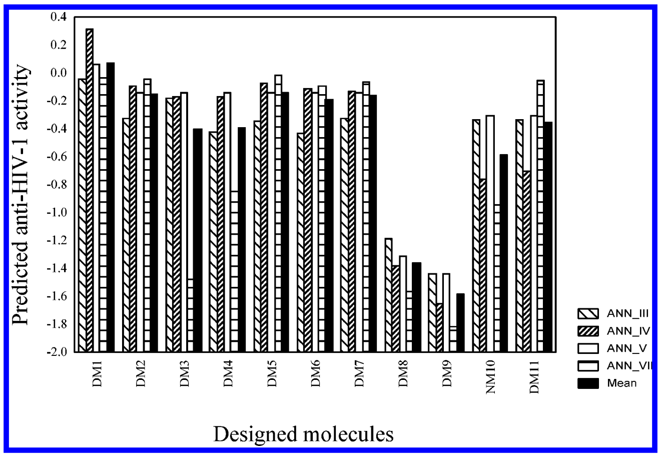


**Figure 7.** Predicted anti-HIV-1 activity values of the designed molecules calculated by the best four ANN models. The average of predicted activities (full column) is recommended to be taken into account.

Although, we only have experimental activity data for natural porphyrins with DF ≥ 6 values, compounds with lower DF values may have higher anti-HIV-1 activity, as seen on the curve in Figure 2. In our prediction, the designed compound DM1 with DF = 4 was found to have the highest activity (0.073).

## CONCLUSIONS

It is relatively easy to calculate the above-mentioned descriptors using computational methodologies for any new molecule to be tested. The anti-HIV-1 activity of new tetrapyrrole molecules can then be predicted based on suitable models presented in this study. These models can be used for the preselection of promising compounds (without their chemical synthesis), which can inhibit viral infection without testing them in complicated biological assays. Because of the widely diverse structures in the training set, many hundreds of compounds designed with our limitations (DF ≤ 6 and DBE ≈ 20) can be tested this way before synthesizing the promising ones.

All model-building methods (MLR, ANN, PLS, and CoMFA) are suitable to build models, which can describe the relationship between chemical structures and anti-HIV-1 activities of the given compounds reasonably well.

Despite good results in leave-one-out and leave-n-out cross-validation procedures, the best MLR model could not properly predict the anti-HIV activity of the three left-out compounds during external validation. This proves that leave-one-out and leave-n-out cross-validation methods without external validation can lead to models with poor or erroneous predictive ability.

Similarly, the best PLS model can be used for description, but it is not suitable for prediction because of the unacceptable result in leave-n-out cross-validation.

ANN proved to be a suitable method to build models that predicted the anti-HIV-1 activity for a noncongeneric series of tetrapyrroles even for a limited number of compounds with clustering tendency. The best ANN models revealed that the degree of chemical bond rotational freedom, DF, is a very important parameter and must be taken into consideration. Combining it with descriptors such as the electrostatic total hydrogen bond acidity, ESTA, and/or double bond equivalent, DBE, plus a WHIM descriptor weighted by atomic mass, KMASS, can result in models with high predictive ability for further design of tetrapyrrole-based compounds in anti-HIV-1 drug development.

Several compounds designed with different anti-HIV-1 activities demonstrate the usefulness and power of the QSAR methodology.

In our case, ANN was superior to other methods such as MLR, PLS, and CoMFA in predicting anti-HIV-1 activity. Therefore, we recommend using an average of predicted activities calculated by ANN models (model III, IV, V, and VII), which show good predictive ability at every validation step.

## ACKNOWLEDGMENT

The authors express thanks to the National Scientific Research Foundation of Hungary for financial help (OTKA T034986).

## REFERENCES AND NOTES

(1) Pandey, R. K. Recent advances in photodynamic therapy. *J. Porphyrins Phthalocyanines* **2000**, *4*, 368−373.

(2) Boyle, R. W.; Dolphyn D. Structure and biodistribution relationships of photodynamic therapy. *Photochem. Photobiol.* **1996**, *64*, 469−485.

(3) Matthews, J. L.; Sogandares-Bernal, F.; Judy, M.; Gulliya, K.; Newman, J.; Chanh, T.; Marengo-Rowe, A. Inactivation of viruses with photoactive compounds. *Blood Cells* **1992**, *18*, 75−88.

(4) North, J.; Coombs, R.; Levy, J. Photodynamic inactivation of free and cell-associated HIV-1 using the photosensitizer, benzoporphyrin derivative. *J. Acquir. Immune Defic. Syndr.* **1994**, *7*, 891−898.

(5) Grandadam, M.; Ingrand, D.; Huraux, J.; Aveline, B.; Delgado, O.; Vever-Bizet, K.; Brault, D. Photodynamic inactivation of cell-free HIV strains by a red-absorbing chlorin-type photosensitizer. *J. Photochem. Photobiol. B* **1995**, *31*, 171−177.

(6) Pollard, S. R.; Meier, W.; Chow, P.; Rosa, J. J.; Wiley, D. C. CD-4 binding regions of human immunodeficiency virus envelope glycoprotein gp120 defined by proteolytic digestion. *Proc. Natl. Acad. Sci. U.S.A.* **1991**, *88*, 11320−11324.

(7) Neurath, A. R.; Strick, N.; Haberfield, P.; Jiang, S. Rapid prescreening for antiviral agents against HIV-1 based on their inhibitory activity in side-directed immunoassays. II. Porphyrins reacting with the V3 loop of gp120. *Antiviral Chem. Chemother.* **1992**, *3*, 55−63.

(8) Neurath, A. R.; Strick, N.; Debnath, A. K. Structural requirements for and consequences of an antiviral porphyrin binding to the V3 loop of the human immunodeficiency virus (HIV-1) envelope glycoprotein gp120. *J. Mol. Rec.* **1995**, *8*, 345−357.

(9) Song, R.; Witrouw, M.; Schols, D.; Robert, A.; Balzarini, J.; De Clercq, E.; Bernadou, J.; Meunier, B. Anti-HIV activities of anionic metalloporphyrins and related compounds. *Antivir. Chem. Chemother.* **1997**, *8*, 85−97.

(10) Nakajima, S.; Hayashi, H.; Omote, Y.; Yamazaki, Y.; Hirata, S.; Maeda, T.; Kubo, Y.; Takemura, T.; Kakiuchi, Y.; Shindo, Y.; Koshimizu, K.; Sakata, I. The tumour-localizing properties of porphyrin derivatives. *J. Photochem. Photobiol. B* **1990**, *7*, 189−198.

(11) Roeder, B.; Naether, D.; Lewald, T.; Braune, M.; Nowak, C.; Freyer, W. Photophysical properties and photodynamic activity in vivo of some tetrapyrroles. *Biophys. Chem.* **1990**, *35*, 303−312.

(12) Margaron, P.; Grégoire, M. J.; Ščasnár, V.; Ali, H.; van Lier, J. E. Structure−photodynamic activity relationships of a series of 4-substituted zinc phthalocyanines. *Photochem. Photobiol.* **1996**, *63*, 217−223.

(13) He, J.; Larkin, H. E.; Li, Y.; Rihter, B. D.; Zaidi, S. I. A.; Rodgers, M. A. J.; Mukhtar, H.; Kenney, M. E.; Oleinick, N. L. The synthesis, photophysical and photobiological properties and in vitro structure−activity relationships of a set of silicon phthalocyanine PDT photosensitizers. *Photochem. Photobiol.* **1997**, *65*, 581−586.

(14) Anderson, C. Y.; Freye, K.; Tubesing, K. A.; Li, Y. S.; Kenney, M. E.; Mukhtar, H.; Elmets, C. A. A comparative analysis of silicon phthalocyanine photosensitizers for in vivo photodynamic therapy of RIF-1 tumors in C3H mice. *Photochem. Photobiol.* **1998**, *67*, 332−336.

(15) Ball, D. J.; Mayhew, S.; Wood, S. R.; Griffiths, J.; Vernon, D. I.; Brown, S. B. A comparative study of the cellular uptake and photodynamic efficacy of three novel zinc phthalocyanines of differing charge. *Photochem. Photobiol.* **1999**, *69*, 390−396.

(16) Lilge, L.; Wilson, C. B. Photodynamic therapy of intracranial tissues: a preclinical comparative study of four different photosensitizers. *J. Clin. Laser Med. Surg.* **1998**, *16*, 81−91.

(17) De Paolis, A.; Chandra, S.; Charalambides, A. A.; Bonnett, R.; Magnus, I. A. The effect on photohaemolysis of variation in the structure of the porphyrin photosensitizer. *Biochem. J.* **1985**, *226*, 757−766.

(18) Henderson, B. W.; Bellnier, D. A.; Greco, W. R.; Sharma, A.; Pandey, R. K.; Vaughan, L. A.; Weishaupt, K. R.; Dougherty, T. J. An in vivo quantitative structure−activity relationship for a congeneric series of pyropheophorbide derivatives as photosensitizers for photodynamic therapy. *Cancer Res.* **1997**, *57*, 4000−4007.

(19) Potter, W. R.; Henderson, B. W.; Bellnier, D. A.; Pandey, R. K.; Vaughan, L. A.; Weishaupt, K. R.; Dougherty, T. Parabolic quantitative structure−activity relationships and photodynamic therapy: application of a three-compartment model with clearance to the in vivo quantitative structure−activity relationships of a congeneric series of pyropheophorbide derivatives used as photosensitizers for photodynamic therapy. *Photochem. Photobiol.* **1999**, *70*, 781−788.

(20) Debnath, A. K.; Jiang, S.; Strick, N.; Lin, K.; Haberfield, P.; Neurath, A. R. Three-dimensional structure−activity analysis of a series of porphyrin derivatives with anti-HIV-1 activity targeted to the V3 loop of the gp120 envelope glycoprotein of the human immunodeficiency virus type 1. *J. Med. Chem.* **1994**, *37*, 1099−1108.

(21) Vanyúr, R.; Héberger, K.; Kövesdi, I.; Jakus J. Prediction of tumoricidal activity and accumulation of photosensitizers in photodynamic therapy using multiple linear regression and artificial neural networks. *Photochem. Photobiol.* **2002**, *75*, 471−478.

(22) HyperChem Release 6.03. HyperCube Inc., Canada, 1999.

(23) IsisBase 2.3. Molecular Design Limited Information Systems AG, Gewerbestrasse 12, CH-4123 Allschwil 2, Switzerland, 1996.

(24) 3DNET version 1.0. ViChem Ltd., Hermann Otto St. 15, 1022 Budapest, Hungary, 1998.

(25) Statistica 6.0 software package. Statsoft Tulsa, Oklahoma, USA, 1999.

(26) Meyer, A. Y. The size of molecules. *J. Chem. Soc. Rev.* **1986**, *15*, 449−474.

(27) Todeschini, R.; Gramatica, P. New 3D molecular descriptors: The WHIM theory and QSAR applications. *Perspect. Drug Discov. Des.* **1998**, 9/10/11, 355−380.

(28) Mortier, W. J.; van Genechten, K.; Gasteiger, J. Electronegativity equalization-application and parametrization. *J. Am. Chem. Soc.* **1985**, *107*, 829−835.

(29) Katritzky, A. R.; Lobanov, V. S.; Karelson, M. Normal boiling points for organic compounds: correlation and prediction by a quantitative structure−property relationship. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 28−41.

(30) Bodor, N.; Huang, M. J.; Harget, A. Neural network studies. Part 3. Prediction of partition coefficients. *J. Mol. Struct. (THEOCHEM)* **1994**, *309*, 259−266.

(31) Andrews, P. R.; Craik, D. J.; Martin, J. L. Functional-group contributions to drug receptor interactions. *J. Med. Chem.* **1984**, *27*, 1648−1657.

(32) Gaillard, P.; Carrupt, P. A.; Testa, B.; Boudon, A. Molecular lipophilicity potential, a tool in 3D QSAR-method and applications. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 83−86.

(33) Cronce, D. T.; Famini, G. R.; De Soto, J. A.; Wilson, L. Y. Using theoretical descriptors in quantitative structure−property relationships − some distribution equilibria. *J. Chem. Soc., Perkin Trans.* **1998**, *6*, 1293−1301.

CI0304627