DOI: 10.17344/acsi.2023.8566 *Acta Chim. Slov.* **2024,** *71,* 170–178

Scientific paper

QSAR Modeling of Sphingomyelin Synthase 2 Inhibitors for Their Potential as Anti-Atherosclerotic Agents

Dejan Petrović,[1,2] Marina Deljanin Ilić,[1,2] Dejan Simonović,[2] Zoran Marčetić,[3] Milovan Stojanović,[1,2] Sanja Stojanović,[2] Nebojša Arsić,[4] Dušan Sokolović[4] and Aleksandar M. Veselinović[5*]

[1] *Faculty of Medicine, University of Niš, Niš, Serbia*

[2] *Institute for Treatment and Rehabilitation, Niška Banja, Serbia*

[3] *Medical faculty, University of Pristina, Kosovska Mitrovica, Serbia*

[4] *Health Center Medveđa, Medveđa, Serbia*

[5] *Faculty of Medicine, University of Niš, Department of Biohemistry, Niš, Serbia*

[6] *Faculty of Medicine, University of Niš, Department of Chemistry, Niš, Serbi*

** Corresponding author: E-mail: aveselinovic@medfak.ni.ac.rs*
Fax: +381 18 4238770; Phone: +381 18 4570029

Received: 11-30-2023

Abstract

Sphingomyelin synthase 2 (SMS2) has emerged as a promising target for atherosclerosis threatment. However, the availability of selective SMS2 inhibitors and their associated pharmacological properties remains limited. This research paper explores various QSAR modeling techniques applied to a range of compounds acting as SMS2 inhibitors. Multiple distinct QSAR modeling methodologies were employed, including conformation-independent, GA-MLR and 3D based QSAR modeling, and their mutual correlations were investigated, Various statistical methods were applied to assess the quality, robustness, and predictive capacity of these developed models, yielding favorable results. Furthermore, molecular fragments derived from SMILES notation descriptors, which account for the observed changes in the evaluated activity, were defined. The methodology presented in this research holds potential for identifying novel agents for atherosclerosis treatment by targeting sphingomyelin synthase 2.

Keywords: Sphingomyelin synthase 2, atherosclerosis, QSAR, Molecular modeling, Drug design

1. Introduction

Sphingomyelin (SM) is a major phospholipid in the circulatory system, and scientific literature indicates that human plasma SM levels are an independent risk factor for coronary heart disease.[1–3] Moreover, in patients with acute coronary syndrome, the measurement of human plasma SM levels can serve as a valuable prognostic tool.[3] Studies have demonstrated that control mice exhibit approximately one-fourth the plasma SM levels compared to apoE KO mice, and this increase in plasma SM levels may be associated with the development of atherosclerosis in these animals.[4,5] Additionally, SM has been shown to have significant effects on the metabolism of apoB-containing lipoproteins, and a deficiency in SM could potentially reduce the atherogenic properties of the mice.[6,7]

Inhibition of serine palmitoyltransferase (SPT), the initial enzyme involved in sphingomyelin (SM) biosynthesis, has been shown to reduce SM levels in mouse models.[8,9] However, it is worth noting that this approach may lead to various off-target side effects because the entire de novo synthesis pathway of sphingolipids can be affected by the inhibition of SPT. As an alternative strategy to lower SM levels, inhibiting sphingomyelin synthase (SMS) is considered.

Scientific literature suggests that the overexpression of sphingomyelin synthase (SMS) promotes the ac-

cumulation of atherogenic lipoproteins and increases the atherogenic potential. Conversely, in a mouse model, the alleviation of atherosclerosis is linked to the reduction of sphingomyelin (SM) accumulation due to SMS deficiency.[10–13] Based on these findings, SMS2 emerges as a potential therapeutic target for atherosclerosis, and the development of future anti-atherosclerotic drugs may be connected to the use of selective SMS2 inhibitors. However, it's important to note that the limited number of reported SMS2 inhibitors is partly attributed to experimental challenges, hindering their exploration as potential anti-atherosclerotic agents.

The process of drug discovery and development is often exceptionally time-consuming, as it relies on various time and resource constraints. To address this challenge, chemoinformatic studies are employed. Chemoinformatics, which involves *in silico* methods, offers a wide range of applications, including the identification of novel lead compounds and the optimization of the pharmacological activity or pharmacokinetic properties of existing chemical compounds with known biological activities.[14,15] Among the various chemoinformatic methods, Quantitative Structure-Activity Relationship (QSAR) has emerged as the most prominent and widely used approach. In contemporary QSAR studies, models are constructed by employing diverse molecular descriptors derived from specific molecule structures, each with its own strengths and limitations. These models are then expressed as mathematical equations that establish a relationship between the biological activities of the studied molecules and their chemical characteristics, as represented by the molecular descriptors.[16–18]

In this research, a variety of in silico methods were employed to identify new compounds with the potential to inhibit sphingomyelin synthase 2 (SMS2). The study developed QSAR models based on the following approaches: conformation-independent molecular descriptors, utilizing both SMILES notation and local graph invariants, in conjunction with the Monte Carlo optimization method; 2D molecular descriptors, with the aid of a genetic algorithm and multiple linear regression; and 3D field contribution. One of the primary objectives of the study was to identify molecular fragments or structural features that lead to SMS2 inhibition effects and to assess the correlations between these different methods. The research successfully identified fragments present in small molecules that are relevant to ligand-receptor interactions, which can potentially be applied in the design and development of anti-atherosclerotic agents.

2. Materials and Methods

In this study, a dataset comprising 51 molecules known to exhibit inhibitory effects on SMS2 was collected from the scientific literature.[19,20] The general chemical structures of these molecules are illustrated in Figure 1. The activities of these molecules, quantified as pIC_{50} values, were used as the dependent variables in the analysis. The SMILES notation for all the molecules used in the study, along with their corresponding pIC_{50} values, is provided in Table S1 within the Supplementary Material. To ensure the robustness of the analysis, the dataset was randomly divided into three sets: a training set consisting of 38 compounds (75%) and a test set comprising 13 compounds (25%). The normality of the activity distribution for all the dataset splits was assessed following the methodology described in a published reference.[21]

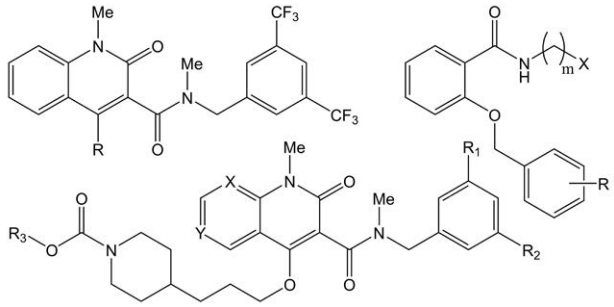

Figure 1. General chemical structures of used molecules for QSAR models development.

2. 1. QSAR Modeling Utilizing the Monte Carlo Optimization Method

The Monte Carlo optimization method was employed to develop a conformation-independent QSAR model using a hybrid approach that incorporates both molecular graph and SMILES notation-based descriptors. The molecular graph-based descriptors included local graph invariants based on fundamental graph concepts like paths and walks, with their detailed mathematical definitions available in the literature.[22] The optimal topological descriptors from the molecular graph-based approach comprised Morgan extended connectivity indices of increasing orders (EC0), valence shells of range 2 and 3 (s2, s3), path numbers of length 2 and 3 (p2, p3), the count of carbon atom neighbors (Number Of Carbon), and the count of non-carbon atom neighbors (*Number of Non Carbon*). In contrast, SMILES notation-based molecular descriptors offer a mechanistic interpretation, as they are related to molecular fragments. The numerical value of each SMILES notation descriptor for a molecule contributes to the molecule's correlation weight (DCW). This DCW is mathematically defined as the sum of all the defined SMILES descriptor correlation weights (CW), in accordance with Equation 1.

$$\begin{aligned} DCW(T,Nepoch) &= zCW(ATOMPAIR) + \\ &xCW(NOSP) + yCW(BOND) + \\ &tCW(HALO) + rCW(HARD) + \alpha\Sigma CW(S_k) + \\ &\beta\Sigma CW(SS_k) + \gamma\Sigma CW(SSS_k) \end{aligned} \quad (1)$$

In Equation 1, the variables z, x, y, t, α, β and γ denote either the value 1 (indicating "yes") or 0 (indicating "no"). These values determine whether the corresponding SMILES descriptor is utilized in the model's development. The symbol S_k specifies the SMILES atom with one SMILES notation symbol (or two inseparable ones) and is linked to the local descriptors. are additionally constructed as linear combinations of two and three SMILES atoms, represented by the SS_k and SSS_k symbols, respectively. The second category of optimal descriptors in accordance with SMILES notation is the global descriptor, which pertains to the overall characteristics of the studied molecule. The study utilized the following global SMILES notation-based descriptors: ATOMPAIR, HALO, BOND, NOSP and HARD, all defined based on the methodology published in reference.[23] The development of the QSAR model in this study involved a combination of both SMILES notation (both local and global) and local graph invariant descriptors. This approach enabled the calculation of the DCW for the molecules as per Equation 2.

$$DCW(T,N_{epoch}) = \Sigma CW(S_k) + \Sigma CW(SS_k) + \Sigma CW(SSS_k) + \Sigma CW(EC0_k) + \Sigma CW(PT2_k) + \Sigma CW(PT3_k) + \Sigma CW(VS2_k) + \Sigma CW(VS3_k) + \Sigma CW(NNC_k) \quad (2)$$

In addition to the previously defined symbols S_k, SS_k and SSS_k, Equation 2 incorporates the following symbols: The Morgan connectivity index of zero order (the hydrogen-suppressed graph was used in this research) – $EC0_k$, paths of length of 2 and 3 – $PT2_k$ and $PT3_k$, valence shell 2 and 3 – $VS2_k$, and $VS3_k$, and Nearest Neighbors – NNC_k.[22] The molecular descriptors mentioned above were all computed using the CORAL software (CORrelation and Logic), which can be accessed at http://www.insilico.eu/coral. Once an optimal descriptor is identified through the application of the Monte Carlo method, each descriptor is assigned a numerical value known as the correlation weight (CW). The Monte Carlo method accomplishes this by generating suitable random numbers and observing how this fractional number corresponds to a specific property or properties. The CW value is then randomly assigned to the descriptors based on the SMILES notation for each individual Monte Carlo run and for a specified endpoint.

The optimization process of the Monte Carlo method involves performing numerical calculations to determine the correlation weights that yield the maximum correlation coefficient value between the optimal descriptor and a given endpoint. When utilizing this method for creating a QSAR model, it's essential to consider two key parameters. Threshold is a coefficient used to categorize a range of molecular features, which encompass both SMILES-based indices and SMILES-based molecular fragments. These features are derived from SMILES notation and sorted into two categories: a) active ones (in this case, the modeling process involves the correlation weight); and b) rare ones (in this case, the modeling process omits the correlation weight).

The process is executed as follows: If a particular molecular feature (X) extracted from the SMILES notation of molecules in the training set occurs fewer than T times, then the molecule descriptor X is excluded from the model construction. Consequently, the numerical value for this feature (the correlation weight of X, CW(X)) is set to zero, categorizing it as "rare." All other molecular features that occur more frequently are considered "active" and can be employed in the model-building process. N_{epoch}, representing the epoch number in Monte Carlo optimization, is crucial for achieving the highest statistical quality within the training set. When an unlimited number of epochs is employed, the training set attains the maximum correlation coefficient through the mentioned Monte Carlo optimization. However, it's important to note that the maximum correlation coefficient between the endpoint for the external test set and the optimal descriptor is achieved with a specific, finite number of epochs. The calculations favor this specific epoch number, as it offers excellent predictive potential for the obtained model, provided that the number of epochs reaches this value. However, it's worth noting that an increase in the threshold (T) results in a decrease in the correlation coefficient within the training set. Nonetheless, it is important to highlight that there exists a threshold value that maximizes the correlation coefficient of the test set. From a practical perspective, the mentioned threshold is the preferred choice. Furthermore, defining optimal values for both the threshold (T) and the Monte Carlo optimization epoch number (N_{epoch}), is essential for constructing a robust QSAR model. This construction involves the utilization of both SMILES notation and optimal descriptors based on the molecular graph, as outlined in reference.[23]

Monte Carlo method simulations are carried out using iterative algorithms to uncover the distribution of an unknown probabilistic entity. In the Monte Carlo optimization process, the epoch number is still a part of the equation for a specific target function within the training set. The initial step involves setting the CW (SA) for each SMILES SA attribute, with all CW values commencing at $1\pm0.01\times Rnd$ (where Rnd is a random value generator with a range between 0 and 1). The usual sequential order of attribute numbers is replaced with a random sequence. The subsequent step involves evaluating the initial value of the target function and making further adjustments to the correlation weights. After this, the relevant steps must be reiterated in the Monte Carlo optimization process for all the non-rare attributes, as specified in references.[23,24] The linear regression approach is used to compute the QSAR model (utilizing the training set) as indicated in Equation 3. This is achieved when the numerical data regarding the correlation weights are derived from the model, leading to favorable statistical results for the test set. In this specific study, the search for the optimal combination of T and

N_{epoch} was carried out within the ranges of 1–5 for T and 0–50 for N_{epoch}.

$$Ac = C_0 + C_1 \times DCW(T,N_{epoch}) \quad (3)$$

2. 2. QSAR Modeling Using Genetic Algorithm in Conjunction with Multiple Linear Regression

In this section, 2D descriptors were calculated using PaDEL.[25] Descriptors with low variance were eliminated from the initial descriptor pool, and further reduction of descriptors was conducted based on filtering using high pairwise correlation coefficients. The QSARINS program (QSAR-INSUBRIA) available at www.qsar.it was employed for various descriptor reductions and for the development of QSAR models.[26,27] After reducing the number of descriptors, they were scaled, and suitable QSAR models were created using the genetic algorithm (GA) optimization method, following the same molecule splitting approach as used in conformation-independent modeling.[28,29] Within the QSARINS program, the genetic algorithm (GA) is combined with multiple linear regression (MLR) as the fitness evaluator.[30,31] For the development of QSAR models, the following parameters were adjusted according to the total number of features in the model: the number of variables in GA optimization was set to 4, the number of GA iterations (generations per size) was set to 500, the population size (the number of models on which GA evolves) was set to 10, and random mutations for generating a diverse pool of descriptors (mutation rate) were set at a 20% mutation rate.

2. 3. 3D Field-based QSAR Model

Before creating the 3D-based QSAR model, geometry optimization was performed on all the molecules using the MMFF94 force field, utilizing Marvin sketch software (Marvin 6.1.0, 2013, ChemAxon). The split that yielded the highest r^2 for the conformation-independent model was employed to divide the molecules into the training and test sets for QSAR model development. The following parameters were utilized in model construction: a maximum of 6 PLS (Partial Least Squares) factors, steric and electrostatic force fields limited to 30.0 kcal/mol, a grid spacing of 1.0 Å with a 3.0 Å extension beyond the training set limits, and elimination of all variables with a standard deviation less than 0.01. The primary software used for developing the 3D field-based QSAR model was Schrodinger Maestro Version 11.5.011.

2. 4. Validation of the Developed QSAR Models

Various validation metrics were employed to assess the quality of the developed conformation-independent and 2D-based QSAR models. These metrics included the determination of the squared correlation coefficient. (r^2), the root-mean-squared-error (RMSE), leave-one-out and leave-many-out cross-validation coefficients, the F-value, the mean absolute error (MAE), and y-scrambling, as referenced.[32–35] To further validate the developed QSAR models, the following statistical metrics were employed: R_m^2 and MAE-based metrics, the correlation coefficient (CCC), and the index of the ideality of correlation (IIC), as described.[36] The applicability domain (AD) is a pivotal aspect of any QSAR model and must be established before utilizing the model.[37,38] In this study, a literature-derived AD method was employed to define applicability domains for conformation-independent QSAR models.[39] It is essential to define the applicability domain (AD) for prediction purposes before making use of any QSAR model. Furthermore, establishing the applicability domain (AD) is an essential and integral component of a pertinent, sturdy, trustworthy, and valid QSAR model. In this study, the AD for the developed QSAR models was determined by examining the "statistical defects" of conformation-independent molecular descriptors, specifically *d*(A), which had been previously employed in the construction of QSAR models.[23,24,36] These calculations were carried out using the CORAL software, following the procedures outlined in Equation 4.

$$d(A) = \frac{|P(A_{train}) - P(A_{test})|}{N(A_{train}) - N(A_{trest})} \quad (4)$$

In the equation above, $P(A_{train})$ and $P(A_{calib})$ denote the probabilities of a conformation-independent attribute or descriptor (A) in the training and test sets, respectively. Meanwhile, N(Atrain) and N(Acalib) represent the frequency of occurrence of a conformation-independent attribute or descriptor (A) in the training set and the test set, respectively. The statistical SMILES defect (D) is the cumulative sum of the defects, d(A), of all the attributes found in the SMILES notation of the molecules. It is computed according to Equation 5.

$$D = defect(SMILES) = \sum_{k=1}^{NA} d(A) \quad (5)$$

A molecule is labeled as an outlier if it falls outside the defined applicability domain (AD), which happens when its D exceeds 2 times Dav, where Dav represents the average D calculated for the relevant set (whether it's the training or test set) in which the molecule is located. The AD for the GA-MLR QSAR models was established using a distance-based approach, and the outliers were detected using the Williams plot, which plots standardized residuals against leverages.

3. Results and Discussion

Table 1 provides the numerical values of all the metrics utilized to assess the quality of the developed confor-

Table 1. The statistical quality of the developed conformational-independent QSAR models for sphingomyelin synthase 2 inhibition

		Training set							Test set						
		r^2	CCC	IIC	q^2	s	MAE	F	r^2	CCC	IIC	q^2	s	MAE	F
Split 1	1 run	0.9087	0.9522	0.8579	0.8959	0.363	0.308	358	0.8907	0.9364	0.9438	0.8453	0.368	0.324	90
	2 run	0.8742	0.9329	0.8415	0.8597	0.426	0.346	250	0.8711	0.9175	0.9333	0.8071	0.395	0.306	74
	3 run	0.8934	0.9437	0.7652	0.8776	0.392	0.32	302	0.8763	0.9292	0.9361	0.8133	0.395	0.309	78
	Av	0.8921	0.9429	0.8215	0.8777	0.394	0.325	303	0.8794	0.9277	0.9377	0.8219	0.386	0.313	81
Split 2	1 run	0.9098	0.9528	0.8584	0.8958	0.341	0.277	363	0.9496	0.9415	0.9744	0.9315	0.401	0.311	207
	2 run	0.9152	0.9557	0.8581	0.9026	0.331	0.273	388	0.9433	0.9518	0.9712	0.9151	0.372	0.313	183
	3 run	0.9092	0.9525	0.8582	0.8949	0.342	0.272	361	0.9427	0.9396	0.9708	0.9196	0.408	0.338	181
	Av	0.9114	0.9537	0.8582	0.8978	0.338	0.274	371	0.9452	0.9443	0.9721	0.9221	0.394	0.321	190
Split 3	1 run	0.9203	0.9585	0.7766	0.9084	0.337	0.272	416	0.8612	0.9223	0.928	0.8226	0.431	0.303	68
	2 run	0.8927	0.9433	0.7649	0.8797	0.391	0.296	300	0.8526	0.9213	0.9232	0.8076	0.433	0.304	64
	3 run	0.8706	0.9308	0.8398	0.8526	0.429	0.342	242	0.8622	0.9246	0.9284	0.8173	0.411	0.277	69
	Av	0.8945	0.9442	0.7938	0.8802	0.386	0.303	319	0.8587	0.9227	0.9265	0.8185	0.425	0.295	67

r^2 – Correlation coefficient; CCC – Concordance correlation coefficient; IIC – Index of ideality of correlation; q^2 – Cross-validated correlation coefficient; s – Standard error of estimation; MAE – Mean absolute error; F – Fischer ratio; Av – Average value for statistical parameters obtained from three independent Monte Carlo optimization runs

mation-independent QSAR models created through the Monte Carlo optimization method. The results indicate that the Monte Carlo optimization method yielded QSAR models with strong predictive capabilities and satisfactory reproducibility. Based on the applied metrics, the most favorable QSAR model was achieved with the second split, featuring a T value of 4 and an N_{epoch} of 15. No outliers were identified, as the methodology applied for the applicability domain (AD) indicated that all molecules fell within the defined AD. Figure 2 illustrates a graphical representation of the best-performing QSAR model (the one with the highest obtained r^2 value) for all three splits in the best Monte Carlo optimization run. The concordance correlation coefficient (CCC) was employed to validate the QSAR models obtained, particularly with respect to their reproducibility. The results indicated that all the models exhibited high reproducibility. Additionally, the results for the MAE-based metric were noted as "GOOD," further confirming the validity of the developed QSAR model.

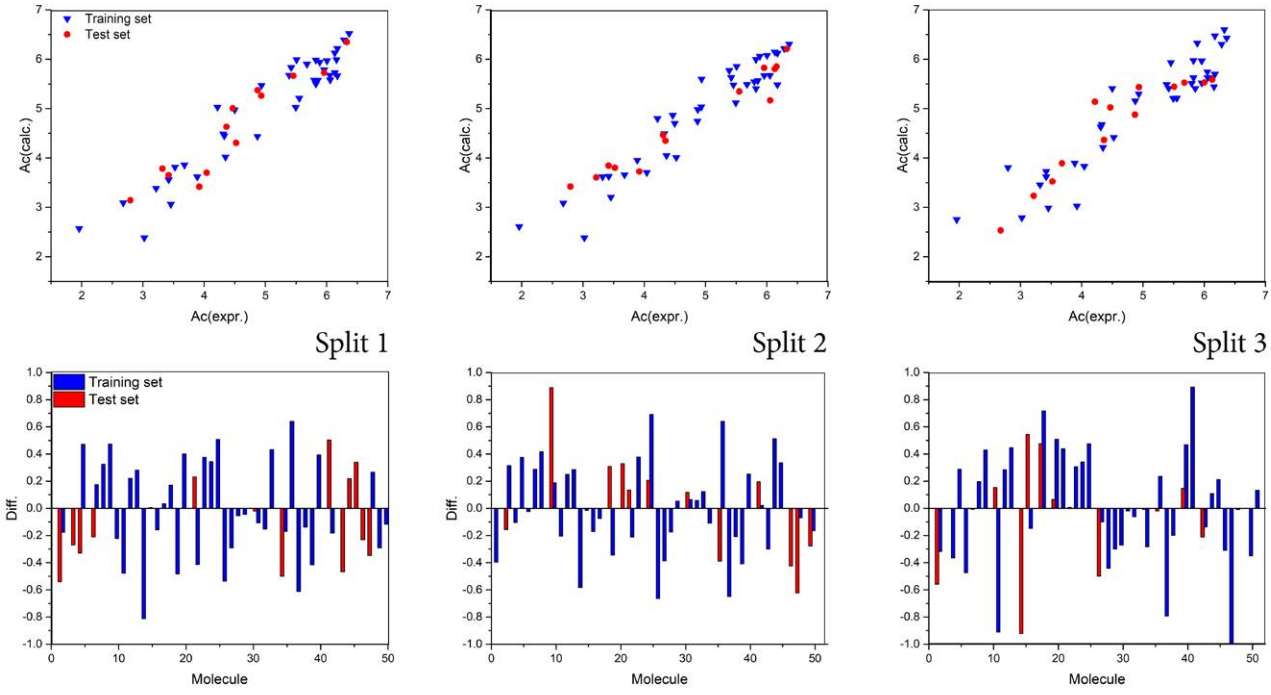

Figure 2. Above) Graphical presentation of the best Monte Carlo optimization runs (the highest value for r^2) for the developed QSAR models; Below) Diff. – Difference between experimental and calculated values for pIC_{50}.

The robustness of the developed QSAR models was assessed using Y-randomization, where Y values were shuffled in 1000 trials for ten separate runs. The outcomes, as presented in Table S2, suggest that the developed QSAR models do not rely on accidental correlations. The final assessment of the quality of the developed QSAR models was conducted using the calculated index of the ideality of correlation (IIC), and the results strongly suggest that the developed QSAR models possess a high predictive potential.

The mathematical formulations for the top-performing QSAR models, as determined by the test set r^2 values for all the splits, are provided in Equations 6–8.

$$\text{Split 1: } pIC_{50} = -1.1653(\pm 0.0716) + 0.0290(\pm 0.0003)\times DCW(3,7) \quad (6)$$

$$\text{Split 2: } pIC_{50} = -1.6041(\pm 0.0810) + 0.0400(\pm 0.0005)\times DCW(4,15) \quad (7)$$

$$\text{Split 3: } pIC_{50} = -1.8678(\pm 0.0950) + 0.0359(\pm 0.0005)\times DCW(2,7) \quad (8)$$

The equations (Eq. 6–8) show that for split 1, the preferred values for T and N_{epoch} are 3 and 7, respectively. For split 2, the preferred values are 4 for T and 15 for N_{epoch}, while for split 3, the preferred values are 2 for T and 7 for N_{epoch}. Equation 9 represents the mathematical equation that characterizes the developed QSAR models generated through GA-MLR modeling for all the splits. A graphical representation of this equation is provided in the supplementary material. The numerical values for all the calculated statistical parameters suggest that the developed QSAR models exhibit satisfactory predictive potential and robustness in terms of prediction. The statistical parameters used for the fitting criteria were as follows:

R^2 : 0.9543; R^2_{adj}: 0.9472; R^2-R^2_{adj} : 0.0071; LOF : 0.1176; Kxx : 0.5102; ΔK : 0.0554; RMSE : 0.2526; MAE : 0.1882; RSS : 2.4253; CCC tr: 0.9766; s: 0.2753; F: 134 The statistical parameters used for internal validation criteria were as follows: Q^2_{loo} : 0.9341; R^2-Q^2_{loo} : 0.0202; RMSE: 0.3035; MAE : 0.2257; PRESS : 3.5003; CCC : 0.9664; Q^2_{LMO} : 0.9234. The statistical parameters used for external validation criteria were as follows: RMSE: 0.6506; MAE: 0.5620; PRESS: 5.5020; R^2 : 0.6316; CCC : 0.7916; $r^2m_{aver.}$: 0.6047; Δr^2m : 0.0353. The model development included the consideration of the following molecular descriptor: Eta_D_beta_A, which represents the ETA average measure of electronic features; C-040 – Atom-centred fragments R-C(=X)-X / R-C#X / X=C=X; SsssCH – Sum of sssCH E-states; SaaN – Sum of aaN E-states; MLogP – Mannhold LogP.

$$pIC_{50} = 3.4370 + 3.3715\times Eta\_D\_beta\_A - 1.4345\times C\text{-}040 + 1.5324\times SsssCH + 0.2257\times SaaN + 0.4852\times MLogP \quad (9)$$

The 3D QSAR model exhibited a test set correlation coefficient of 0.9392, with a standard deviation of 0.2967. Additionally, the training set correlation coefficient for the 3D QSAR model was 0.6843, with a standard deviation of 0.2824. These values collectively indicate that the model demonstrates good predictability. The results derived from the 3D QSAR model provide the following Gaussian field fraction contributions: 0.4240 for steric interactions, 0.0825 for electrostatic interactions, 0.2815 for hydrophobic interactions, 0.1971 for hydrogen bond acceptor inter-

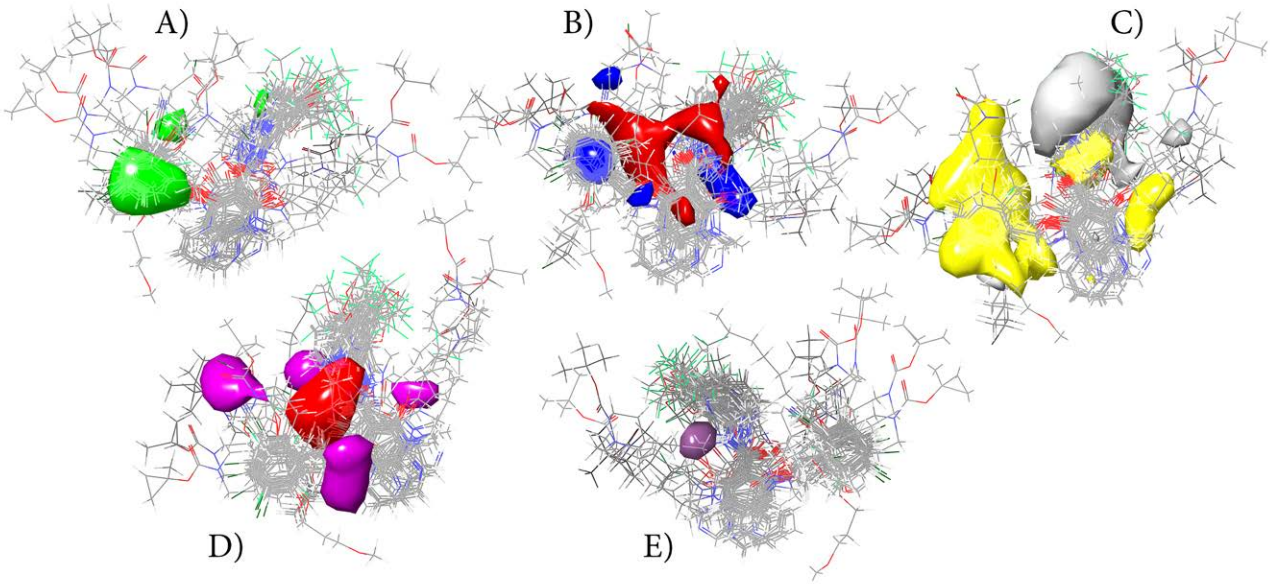

Figure 3. 3D QSAR model fields (fields are shown as surfaces). A) Steric – favourable regions (green); B) Hydrophobic – favoured (yellow) and disfavoured (white); C) Electrostatic – favoured electropositive (blue) and disfavoured electronegative (red); D) Hydrogen bond acceptor – favoured (red) and disfavoured (magenta); E) Hydrogen bond donor – favoured (purple) and disfavoured (cyan).

Table 2. The example of DCW(4,15) calculation

SMILES notation:
CN(C(=O)c1c(OCCCC2CCCCC2)c2ccccc2n(c1=O)C)Cc1cc(cc(c1)C(F)(F)F)C(F)(F)F
DCW = 113.53444
pIC_{50}(calc.) = 3.7003

SA(CW)	CW	SA(CW)	CW	SA(CW)	CW	SA(CW)	CW
10011001000	−0.9	2...n...(...	0.7295	c...c...(...	0.4785	N...(...C...	0.1565
(...(.....	−0.5175	BOND10000	2.2636	C...C.....	0.4516	n...(...c...	0.0245
(..........	−0.2838	C...(.....	−0.5648	c...c.....	0.0705	N.....	−0.7298
(...C...(...	−0.9619	c...(.....	0.1721	c...C.....	0.1702	n.....	0.0083
(...F...(...	0.3981	C...(...=...	−0.8874	C...c...1...	0.311	n...2.....	−0.2463
++++F---B2==	0.9554	C...(...1...	0.4141	c...c...1...	0.0902	n...2...c...	−0.8469
++++F---N===	2.0554	c...(...2...	−4.0999	C...C...2...	−0.5719	N...C.....	0.238
++++F---O===	2.0323	C...(...C...	0.0284	c...c...2...	−1.4034	Nmax.1.....	2.2076
++++N---B2==	2.413	c...(...c...	0.4344	C...C...C...	−2.8767	NOSP110000	6.3387
++++N---O===	3.2561	c...(...O...	0.1071	c...c...c...	0.1149	O...(.....	−0.9913
++++O---B2==	−1.8357	C.....	0.0043	C...N...(...	−0.9921	O...(...C...	0.675
=...(.....	0.667	c.....	0.0275	C...O...(...	−0.7001	O.....	0.1213
=.....	0.4955	c...1...(...	0.2445	Cmax.2.....	−1.702	O...=...(...	−0.6294
=...1.....	0.4017	c...1.....	0.192	F...(...(...	−0.8584	O...=.....	−0.8534
=...O...(...	0.4016	c...1...=...	−0.8758	F...(.....	0.1322	O...=...1...	0.0153
1...(.....	0.1372	c...1...c...	0.3067	F...(...C...	−0.8886	O...C.....	0.3938
1.....	0.3516	C...2...(...	−0.7831	F...(...F...	−0.7542	O...C...C...	0.3251
1...c...(...	0.0674	C...2.....	−0.8551	F.....	0.0899	Omax.3.....	6.8713
2...(.....	−4.0883	c...2.....	−2.6633	HALO100000	−0.7419	Smax.0.....	4.2841
2.....	−0.7638	c...2...c...	−1.7803	N...(.....	−0.5548		
2...c...(...	0.296	c...C...(...	−0.0875	n...(.....	0.6401		

actions, and 0.0149 for hydrogen bond donor interactions. These results suggest that steric interactions, followed by hydrophobic interactions, exert the most significant influence on the studied activity, particularly with regard to the increase in the size of substituent groups. In contrast, electrostatic and hydrogen bond donor interactions have the least impact. The surfaces representing the fields obtained for the developed 3D QSAR model are depicted in Figure 3.

One of the main objectives of this research was to identify the molecular fragments defined as optimal descriptors in the SMILES notation that have both positive and negative impacts on the studied activity, as referenced.[23,24,40–43] The comprehensive list of calculated molecular descriptors, based on both the SMILES notation and the molecular graph, can be found in Table S3 (Supplementary material). For clarity, an example of the calculation for the molecule's summarized correlation weight (DCW) and the studied activity (pIC50) is provided in Table 2, with the molecular graph-based descriptors omitted to facilitate interpretation. Additionally, a graphical representation of the molecular fragments for the same molecule is presented in Figure 4.

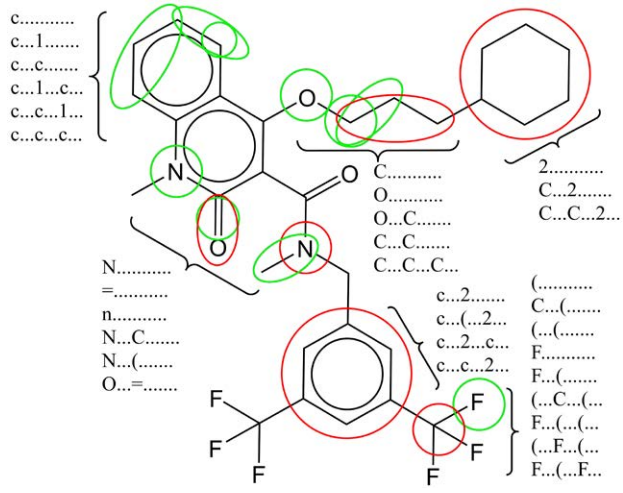

Figure 4. Molecular fragments contribution to sphingomyelin synthase 2 inhibition (green – increase, red – decrease).

Based on the results obtained from QSAR modeling, the SMILES notation reveals the following molecular fragments that influence pIC_{50} activity: "C....." – carbon atom or a methyl group; "O....." – oxygen atom or hydroxyl group; "C...C....." – representing two connected carbon atoms or an ethyl group; "c.....", "c...1.....", "c...c.....", "c...1...c...", "c...c...1...", and "c...c...c..." – one aromatic carbon atom, two or three linear combinations of aromatic carbon atoms; "O...C..." – referring to a methoxy group or two connected carbon and oxygen atoms; "c...1...(...", "c...(...1...", "c...(...C...", "c...C...(...", "c...1...C..." – linked to the addition of at least one methyl group to benzene, resulting in branching; "(...(.....", "(.....", "(...C...(..." SMILES

notation fragment associated with molecular branching: “F.....”, “c...F.....”, “F...c...1...” SMILES notation fragments associated with the addition of a fluorine atom to the benzene ring. "N....."”— representing a nitrogen atom with a negative impact on studied activity, but N....."— denoting a nitrogen atom involved in molecular branching has a positive impact. Similar to the aromatic carbon, the aromatic nitrogen atom, indicated by the "n....." molecular descriptor, also exerts a positive influence on the studied activity. "N...C..."— the primary amine group contributes positively, while secondary and tertiary amines, indicated with branching as "C...N...", have a negative impact. "=....." – a double bond exerts a positive influence, but the double bond with the oxygen atom, represented as "O..=.," negatively affects the studied activity. The presence of one ring, whether aromatic or aliphatic, positively impacts the studied activity. This molecular feature is defined by the following molecular descriptors: “1.....”, “c...1.....”, “c...c...1...”, "C...(...1..."). Nevertheless, a further increase in the number of rings, whether aromatic or aliphatic, has a negative impact on the studied activity: “c...2.....”, "c...(...2..."), “c...2...c...”, “c...c...2...”, “2.....”, “C...2.....”, “C...C...2...”. Molecular branching as a feature and molecular branching with involved carbon atoms defined as “(.....”, “(...(.....”, “C...(.....”, "(...C...(..."')" have a negative impact on the studied activity. Both fluorine atoms ("F." and molecular branching involving fluorine atoms “F...(...(...”, “(...F...(...” and “F...(...F..” positively affect the studied activity.

4. Conclusion

The primary objective of this study was to create reliable QSAR models that demonstrate strong predictability, assessed using a range of statistical parameters, for the inhibition of sphingomyelin synthase 2. The Monte Carlo optimization method was employed to compute conformation-independent QSAR models. These models were built using optimal descriptors derived from both a local graph and SMILES notation invariants. A QSAR model was constructed using a genetic algorithm in conjunction with multiple linear regression, utilizing an extensive set of 2D molecule descriptors. The assessment of the robustness and predictive capability of these developed QSAR models was achieved through the application of various statistical techniques. The numerical values derived to validate the developed QSAR models demonstrate their high applicability. A field-based contribution approach was employed to establish the 3D QSAR model, and the results obtained revealed that the steric and hydrophobic parameters had the most significant impact on the inhibition activity. Molecular fragments, employed as SMILES notation fragments in QSAR modeling, with both positive and negative effects on sphingomyelin synthase 2 inhibition were identified through the Monte Carlo optimization method. The methodology outlined in this study can be adapted to discover novel therapeutics for the treatment of atherosclerosis by targeting the inhibition of sphingomyelin synthase 2.

Funding: This work is supported by the Ministry of Education and Science, the Republic of Serbia and the Faculty of Medicine, University of Niš, Republic of Serbia (project No. 70). The authors would like to thank the Ministry of Education, Science and Technological Development of Republic of Serbia (Grant No: 451-03-47/2023-01/200113) for financial support.

Data Availability Statement: Data is contained within the article and Supplementary Materials.

Conflicts of Interest: The authors declare that there are no conflicts of interest in this study.

5. References

1. A. Nilsson, R. D. Duan, *J. Lipid. Res.* **2006**, *47*, 154–171. **DOI:** 10.1194/jlr.M500357-JLR200
2. X. C. Jiang, F. Paultre, T. A. Pearson, R.G. Reed, C. K. Francis, M. Lin, L. Berglund, A. R. Tall, *Arterioscler. Thromb. Vasc. Biol.* **2000**, *20*, 2614–2618. **DOI:** 10.1161/01.ATV.20.12.2614
3. A. Schlitt, S. Blankenberg, D. Yan, H. von Gizycki, M. Buerke, K. Werdan, C. Bickel, K. J. Lackner, J. Meyer, H. J. Rupprecht, X. C. Jiang, *Nutr. Metab. (Lond).* **2006**, *3*, 5. **DOI:** 10.1186/1743-7075-3-5
4. Ts. Jeong, S. L. Schissel, I. Tabas, H. J. Pownall, A. R. Tall, X. Jiang, *J. Clin. Invest.* **1998**, *101*, 905–912. **DOI:** 10.1172/JCI870
5. A. S. Plump, J. D. Smith, T. Hayek, K. Aalto-Setälä, A. Walsh, J. G. Verstuyft, E.M. Rubin, J. L. Breslow, *Cell*, **1992**, *71*, 343–353. **DOI:** 10.1016/0092-8674(92)90362-G
6. J. L. Rodriguez, G. C. Ghiselli, D. Torreggiani, C. R. Sirtori, *Atherosclerosis*, **1976**, *23*, 73–83. **DOI:** 10.1016/0021-9150(76)90119-2
7. Y. Fan, F. Shi, J. Liu, J. Dong, H. H. Bui, D. A. Peake, M. S. Kuo, G. Cao, X. C. Jiang, *Arterioscler. Thromb. Vasc. Biol.* **2010**, *30*, 2114–2120. **DOI:** 10.1161/ATVBAHA.110.213363
8. M. R. Hojjati, Z. Li, H. Zhou, S. Tang, C. Huan, E. Ooi, S. Lu, X. C. Jiang, *J. Biol. Chem.* **2005**, *280*, 10284–10289. **DOI:** 10.1074/jbc.M412348200
9. T. S. Park, R. L. Panek, S. B. Mueller, J. C. Hanselman, W. S. Rosebury, A. W. Robertson, E. K. Kindt, R. Homan, S. K. Karathanasis, M. D. Rekhter, *Circulation* **2004**, *110*, 3465–3471. **DOI:** 10.1161/01.CIR.0000148370.60535.22
10. J. Liu, C. Huan, M. Chakraborty, H. Zhang, D. Lu, M. S. Kuo, G. Cao, X. C. Jiang, *Circ. Res.* **2009**, *105*, 295–303. **DOI:** 10.1161/CIRCRESAHA.109.194613
11. M. Chakraborty, C. Lou, C. Huan, M. S. Kuo, T. S. Park, G. Cao, X. C. Jiang, *J. Clin. Invest.* **2013**, *123*, 1784–1797. **DOI:** 10.1172/JCI60415
12. J. Dong, J. Liu, B. Lou, Z. Li, X. Ye, M. Wu, X. C. Jiang, *J. Lipid. Res.* **2006**, *47*, 1307–1314. **DOI:** 10.1194/jlr.M600040-JLR200
13. Z. Li, Y. Fan, J. Liu, Y. Li, C. Huan, H. H. Bui, M.S . Kuo, T. S. Park, G. Cao, X. C. Jiang, *Arterioscler. Thromb. Vasc. Biol.* **2012**, *32*, 1577–1584. **DOI:** 10.7312/li--16274-033

14. S. Ekins, J. Mestres, B. Testa, *Br. J. Pharmacol.* **2007**, *152*, 9–20. **DOI:** 10.1038/sj.bjp.0707305
15. J. Tabeshpour, A. Sahebkar, M.R. Zirak, M. Zeinali, M. Hashemzaei, S. Rakhshani, S. Rakhshani, *Curr. Pharm. Design*. **2018**, 24, 3014–3019. **DOI:** 10.2174/1381612824666180903123423
16. C. Nantasenamat, C. Isarankura-Na-Ayudhya, V. T. Naenna, A. Prachayasittikul, *EXCLI J.* **2009**, *8*, 74–88.
17. P. Liu, W. Long, *Int. J. Mol. Sci.* **2009**, *10*, 1978–1998. **DOI:** 10.3390/ijms10051978
18. M. Pérez González, C. Terán, L. Saíaz-Urra, M. Teijeira, *Curr. Top. Med. Chem.* **2008**, *8*, 1606–1627. **DOI:** 10.2174/156802608786786552
19. Y. Li, T. Huang, B. Lou, et al., *Eur. J. Med. Chem.* **2019**, *163*, 864–882. **DOI:** 10.1016/j.ejmech.2018.12.028
20. T. Yukawa, T. Nakahata, R. Okamoto, et al., *Bioorg. Med. Chem.* **2020**, *28*, 115376. **DOI:** 10.1016/j.bmc.2020.115376
21. P. K. Ojha, K. Roy, *Chemometr. Intell. Lab.* **2011**, *109*, 146–161. **DOI:** 10.1016/j.chemolab.2011.08.007
22. A. A. Toropov, P. Duchowicz, E. A. Castro, *Int. J. Mol. Sci.* **2003**, *4*, 272–283. **DOI:** 10.3390/i4050272
23. A. M. Veselinović, J. B. Veselinović, J. V. Živković, G. M. Nikolić, *Curr. Top. Med. Chem.* **2015**, *15*, 1768–1779.
24. M. Zivković, M. Zlatanović, N. Zlatanović, M. Golubović, A. M. Veselinović, *Mini-Rev. Med. Chem.* **2020**, *20*, 1389–1402. **DOI:** 10.2174/1389557520666200212111428
25. C. W. Yap, *J. Comput. Chem.* **2011**, *32*, 1466–1474. **DOI:** 10.1002/jcc.21707
26. P. Gramatica, S. Cassani, N. Chirico, *J. Comput. Chem.* **2014**, *35*, 1036–1044. **DOI:** 10.1002/jcc.23576
27. P. Gramatica, N. Chirico, E. Papa, S. Cassani, S. Kovarich, *J. Comput. Chem.* **2013**, *34*, 2121–2132. **DOI:** 10.1002/jcc.23361
28. P. Johnson, L. Vandewater, W. Wilson, and et al., *BMC Bioinformatics* **2014**, *15*, S11. **DOI:** 10.1186/1471-2105-15-S16-S11
29. N. Sukumar, G. Prabhu, P. Saha, In: *Applications of Metaheuristics in Process Engineering*, ed. J. Valadi and P. Siarry, Springer, Cham, **2014**, 315–324. **DOI:** 10.1007/978-3-319-06508-3_13
30. B. Hemmateenejad, R. Miri, M. Akhond, M. Shamsipur, *Chemom. Intell. Lab. Syst.* **2002**, *64*, 91–99. **DOI:** 10.1016/S0169-7439(02)00068-0
31. E. Setiawan, K. Wijaya, M. Mudasir, *J. Appl. Pharm. Sci.* **2021**, *11*, 022–027.
32. A. Golbraikh, A. Tropsha, *J. Mol. Graph. Model.* **2002**, *20*, 269–276. **DOI:** 10.1016/S1093-3263(01)00123-1
33. P. P. Roy, J. T. Leonard, K. Roy, *Chemometr. Intell. Lab.* **2008**, *90*, 31–42. **DOI:** 10.1016/j.chemolab.2007.07.004
34. P. K. Ojha, I. Mitra, R. N. Das, K. Roy, *Chemometr. Intell. Lab.* **2011**, *107*, 194–205. **DOI:** 10.1016/j.chemolab.2011.03.011
35. K. Roy, R. N. Das, P. Ambure, R. B. Aher, *Chemometr. Intell. Lab*. **2016**, *152*, 18–33. **DOI:** 10.1016/j.chemolab.2016.01.008
36. A. P. Toropova, A. A. Toropov, *Sci. Total Environ.* **2017**, *586*, 466–472. **DOI:** 10.1016/j.scitotenv.2017.01.198
37. D. Gadaleta, G. F. Mangiatordi, M. Catto, A., Carotti, O. Nicolotti, *IJQSPR* **2016**, *1*, 45–63. **DOI:** 10.4018/IJQSPR.2016010102
38. P. Gramatica, *QSAR Comb. Sci.* **2007**, *26*, 694–701. **DOI:** 10.1002/qsar.200610151
39. A. A. Toropov, A. P. Toropova, A. Lombardo, A. Roncaglioni, E. Benfenati, G. Gini, *Eur. J. Med. Chem.* **2011**, *46*, 1400–1403. **DOI:** 10.1016/j.ejmech.2011.01.018
40. A. Antović, R. Karadžić, J. V. Živković, A. M. Veselinović. *Acta Chim. Slov.* **2023**, *70*, 634–641. **DOI:** 10.17344/acsi.2023.8465
41. N. Nikolić, T. Kostić, M. Golubović, T. Nikolić, M. Marinković, V. Perić, S. Mladenović, A. M. Veselinović. *Acta Chim. Slov.* **2023**, *70*, 318–326. **DOI:** 10.17344/acsi.2023.8081
42. S. Ahmadi, S. Lotfi, S. Afshari, P. Kumar, E Ghasemi, *SAR QSAR Environ. Res.* **2021**, *32*, 1013–1031. **DOI:** 10.1080/1062936X.2021.2003429

Povzetek

Sfingomielin sintaza 2 (SMS2) se je izkazala kot obetavna trača cza zdravljenje ateroskleroze. Kljub temu pa je dostopnost selektivnih zaviralcev SMS2 in njihove povezane farmakološke lastnosti omejena. Ta članek raziskuje različne tehnike modeliranja, osnovane na kvantitativnem razmerju med strukturo in delovanjem (QSAR), ki so bile uporabljene na različnih spojinah, ki delujejo kot inhibitorji SMS2. Uporabili smo različne metodologije modeliranja QSAR, vključno s konformacijsko neodvisnim modeliranjem, GA-MLR in 3D modeliranjem QSAR, proučili pa smo tudi korelacije med njimi. Za oceno kakovosti, robustnosti in napovedne sposobnosti napravljenih modelov smo uporabili različne statistične metode, pri čemer smo dosegli dobre rezultate. Poleg tega smo določili molekularne fragmente, pridobljene iz SMILES notacije deskriptorjev, ki upoštevajo opažene spremembe v ocenjeni aktivnosti. Metodologija, predstavljena v tej raziskavi, ima potencial za identifikacijo novih učinkovin za zdravljenje ateroskleroze z usmerjanjem na SMS2.