Indian Journal of Chemistry
Vol. 63, March 2024, pp. 315-324
DOI: 10.56042/ijc.v63i3.6966

CP-MLR/PLS directed structure-activity study in modeling of the aggrecanase-1 inhibitory activity of biphenylsulfonamides

Nidhi Shekhawat*[a] & Prithvi Singh[b]

[a] Department of Zoology, Vedanta Post-Graduate Girls' College, Reengus 332 404, India

[b] Department of Chemistry, S. K. Government Post-Graduate College, Sikar 332 001, Rajasthan, India

E-mail: nidhi14sikar@gmail.com

Received 27 November 2023; accepted (revised) 22 February 2024

The inhibition activity of biphenylsulfonamide derivatives on aggrecanase-1 has been determined through quantitative analysis of molecular descriptors. The resulting models account for more than 83% of the variance in observed inhibition activity and have been satisfactorily validated by test-set statistics. Molecular features such as mean square distance (MSD), polarizability weighted lag-1 (GATS1p), and electronegativity weighted lag-5 (GATS5e) have been found to be crucial for receptor site interaction, along with the presence of an H attached to CO (sp3) with no heteroatom X attached at next C (H-046). Higher values of MSD, GATS5e, and H-046, coupled with lower values of GATS1p, improve the compound's activity profile. The partial least-squares (PLS) study reveals a "single window" structure-activity model using the most significant descriptors, with two optimum components explaining 84% of the variance in observed activity values. The applicability domain (AD) study confirms the models' predictability, with all compounds except one outlier within the proposed model's AD. The AD analysis has also identified as one structurally influential compound.

Keywords: Aggrecanase-1 inhibitors, Biphenylsulfonamides, Molecular descriptors, QSAR, Combinatorial protocol in multiple linear regression (CP-MLR) analysis

Osteoarthritis (OA) is a disease that results in the progressive loss of articular cartilage in joints, causing functional impairment, discomfort, and pain. The action of degradative proteolytic enzymes causes the loss of collagen and aggrecan, the two major components of cartilage. Agrecanases are the enzymes responsible for aggrecan degradation, and they represent a family of enzymes with significant therapeutic potential as osteoarthritis targets[1,2].

Aggrecanase-1 (ADAMTS-4)[3] and Aggrecanase-2 (ADAMTS-5) are zinc-containing metalloproteases belonging to the ADAMTS (a disintegrin and metalloprotease possessing thrombospondin domain) family. In osteoarthritis, these enzymes cleave the aggrecan IGD (interglobular domain) at Glu373-Ala374. In a surgically induced OA[4,5] model in mice, ADAMTS-5 (but not ADAMTS-4) is responsible for disease progression. However, questions remain about the relative importance of ADAMTS-4 and ADAMTS-5 in human disease. Furthermore, ADAMTS-4/ADAMTS-5 double-knockout mice are physiologically normal[6] and do not develop osteoarthritis. As a result, inhibiting aggrecanases represents an appealing target for developing new drugs that may slow the progression of OA[1,2].

In previous study, Xiang *et al.*[7] found that replacing the biphenyl P1' group in sulfonamides could enhance aggrecanase-1 activity. Building on this, Hopper *et al.*[8], from the same research group, synthesized a range of biphenylsulfonamides to determine their effectiveness against aggrecanase-1. Their investigation identified a series of ((4-keto)-phenoxy) methyl biphenyl-4-sulfonamide analogs that showed improved potency in aggrecanase-1 inhibition. However, their structure-activity relationship (SAR) study involved a trial-and-error approach in which they modified the substituents at various biphenyl positions of the sulfonamide.

This communication aims to conduct a 2D-quantitative SAR (2D-QSAR) study on the reported compounds. The goal is to provide insights into the molecular features of the compounds that contribute to their inhibition actions towards aggrecanase-1. In this congeneric series, where a relative study is being performed, the 2D descriptors may play a significant role in establishing the relationships between the biological activities of the compounds. The simplicity of calculating descriptors and interpreting them in a physical sense in 2D-QSAR studies helps

to explain the molecular basis of the biological activities of the compounds.

Materials and Methods

The compounds (general structure in Fig. 1), along with their inhibition activity, IC_{50} values were taken from the literature[8], and are listed as pIC_{50} on a molar basis in Table 1.

Validation of the models was achieved through both internal and external approaches. The internal validation process involved the use of leave-one-out (LOO) and leave-five-out (L5O) procedures. External validation was performed using a set of test compounds, which were chosen from the total 30 compounds, while remaining compounds were included in the training-set. To select the test-set compounds, the single linkage hierarchical cluster

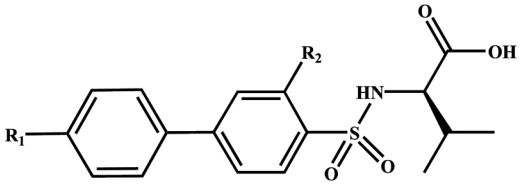
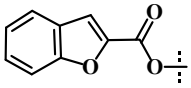
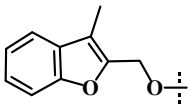
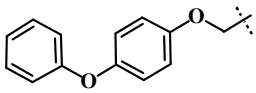
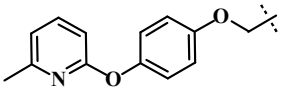
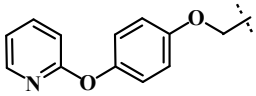
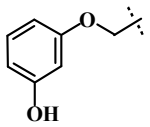
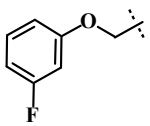
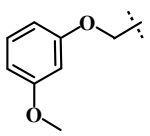
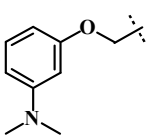


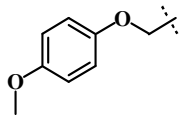
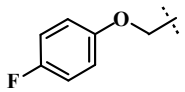
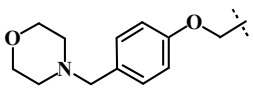
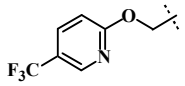
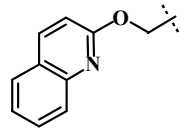
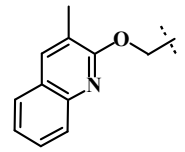
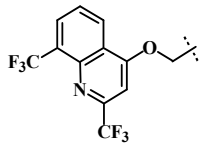
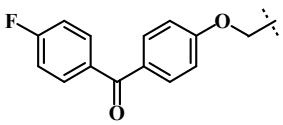
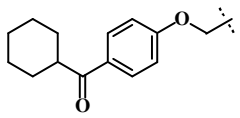
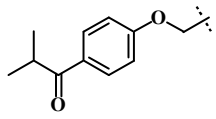
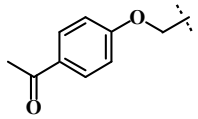
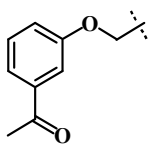
Fig. 1 — General structure of biphenylsulfonamide derivatives.

Table 1 — Important molecular descriptors, observed and modeled aggrecanase-1 inhibition activities of biphenylsulfonamides (Fig. 1 for general structure)

Compd.	R_1	R_2	MSD	GATS5e	GATS1p	H-046	pIC_{50} (M)			
							Obsd.[a]	Calcd.		
								Eq. (5)	LOO	PLS
1	H	H	0.260	1.512	1.208	6	5.52	5.57	5.59	6.02
2		H	0.264	1.344	1.138	6	6.46	6.11	6.06	6.32
3		H	0.267	1.284	1.159	9	6.15	6.25	6.26	6.26
4		H	0.282	1.199	1.135	6	6.40	6.46	6.47	6.23
5		H	0.281	1.301	1.166	6	6.14	6.28	6.29	6.15
6		H	0.282	1.211	1.159	6	6.05	6.24	6.28	6.07
7		H	0.274	1.307	1.153	6	5.96	6.21	6.23	6.18
8		H	0.274	1.187	1.166	6	5.92	5.91	5.91	5.96
9		H	0.274	1.246	1.214	6	5.74	5.54	5.51	5.51
10		H	0.272	1.313	1.231	6	5.31	5.41	5.43	5.44

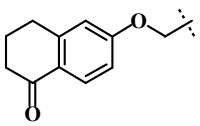
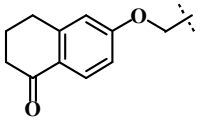
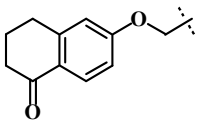
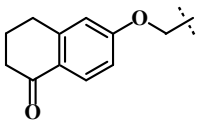
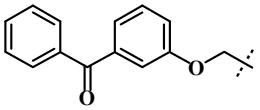
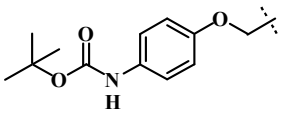
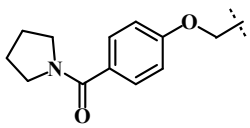
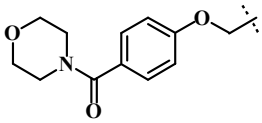
(*Contd.*)

Table 1 — Important molecular descriptors, observed and modeled aggrecanase-1 inhibition activities of biphenylsulfonamides (Fig. 1 for general structure) — (*Contd.*)

Compd.	R_1	R_2	MSD	GATS5e	GATS1p	H-046	pIC_{50} (M)			
							Obsd.[a]	Calcd.		
								Eq. (5)	LOO	PLS
11		H	0.278	1.116	1.214	6	5.80	5.47	5.37	5.47
12[b]		H	0.276	1.043	1.166	6	5.77	5.76	--	5.92
13		H	0.282	1.162	1.241	6	5.22	5.39	5.46	5.20
14[b]		H	0.277	0.998	1.200	6	6.10	5.40	--	5.56
15		H	0.272	1.437	1.132	6	6.40	6.53	6.56	6.48
16		H	0.266	1.413	1.141	9	6.70	6.58	6.56	6.58
17[c]		H	0.237	0.948	1.142	6	6.70	4.72	--	5.03
18[b]		H	0.277	1.144	1.099	6	7.10	6.57	--	6.47
19		H	0.277	1.372	1.172	16	7.05	7.35	7.89	7.23
20		H	0.278	1.251	1.189	12	7.00	6.58	6.44	7.06
21[b]		H	0.278	1.244	1.173	6	6.37	6.04	--	6.14
22[b]		H	0.272	1.243	1.173	6	6.00	5.87	--	6.08

(*Contd.*)

Table 1 — Important molecular descriptors, observed and modeled aggrecanase-1 inhibition activities of biphenylsulfonamides (Fig. 1 for general structure) — (*Contd.*)

Compd.	R_1	R_2	MSD	GATS5e	GATS1p	H-046	pIC_{50} (M) Obsd.[a]	pIC_{50} (M) Calcd. Eq. (5)	pIC_{50} (M) Calcd. LOO	pIC_{50} (M) Calcd. PLS
23[b]		H	0.270	1.405	1.146	10	6.64	6.75	--	6.48
24		F	0.261	1.355	1.145	10	6.74	6.43	6.38	6.22
25		CF_3	0.239	1.091	1.152	10	5.49	5.34	5.24	5.42
26		OCF_3	0.236	1.162	1.184	10	4.70	5.05	5.35	5.16
27		H	0.267	1.375	1.100	6	6.64	6.60	6.59	6.62
28		H	0.282	1.598	1.241	6	6.23	6.02	5.85	5.93
29		H	0.277	1.410	1.197	6	5.72	6.02	6.06	5.81
30[b]		H	0.277	1.223	1.229	6	5.28	5.45	--	5.41

[a]IC_{50} values are taken from reference[8], [b]Test-set compound, [c]Outlier compound

procedure in SYSTAT[9] was utilized. This process created a cluster tree based on pIC_{50} values and Euclidean distances, from which roughly 25% of the total compounds were chosen to maximize the distance between them. The normalized Euclidean distances were computed through root-mean-squared distances in SYSTAT, and the single linkage clustering method was utilized to generate longer clusters that allowed for object selection at different intervals based on the distance between the two closest members.

Molecular Descriptors

The compounds being studied were drawn using the standard procedure in ChemDraw[10] and then transferred to DRAGON software[11] to compute descriptors for 0D-, 1D-, and 2D-classes. Table 2 provides definitions and scopes of these descriptor classes, which were used to address the structural features in this study.

To develop the QSAR models, the combinatorial protocol in multiple linear regression (CP-MLR) computational procedure[12] was utilized. The computed descriptors were then scaled[13] so that their values ranged from 0 to 1, ensuring that none of the descriptors dominated each other like in the case of pre-scaled descriptors with larger or smaller values. This scaling ensured that all descriptors had an equal influence in the QSAR models.

Table 2 — Descriptor classes used for the analysis of aggrecanase-1 inhibition activity

Descriptor class (acronyms)	Definition and scope
Constitutional (CONST)	Dimensionless or 0D descriptors; independent from molecular connectivity and conformations.
Topological (TOPO)	2D descriptor from molecular graphs and independent conformations.
Molecular walk counts (MWC)	2D descriptors representing self-returning walk counts of different lengths.
Modified Burden eigenvalues (BCUT)	2D descriptors representing positive and negative eigenvalues of the adjacency matrix, weights the diagonal elements and atoms.
Galvez topological charge indices (GLVZ)	2D descriptors representing the first 10 eigenvalues of corrected adjacency matrix.
2D-autocorrelations (2DAUTO)	Molecular descriptors calculated from the molecular graphs by summing the products of atom weights of the terminal atoms of all the paths of the considered path length (the lag).
Functional groups (FUNC)	Molecular descriptors based on the counting of the chemical functional groups.
Atom-centred fragments (ACF)	Molecular descriptors based on the counting of 120 atom-centred fragments, as defined by Ghose-Crippen.
Empirical (EMP)	1D descriptors represent the counts of non-single bonds, hydrophilic groups and ratio of the number of aromatic bonds and total bonds in an H-depleted molecule.
Properties (PROP)	1D descriptors representing molecular properties of a molecule.

Model Development

The CP-MLR procedure is a commonly used method for developing models in QSAR studies. Our previous publications[14-18] provided detailed discussions on the procedural aspects and implementation of this method. The software includes four built-in filters that streamline the variable selection process, resulting in a unique solution. Filter-1 selects variables with inter-parameter correlations up to a specific upper limit (≤ 0.79). Filter-2 controls variable entry to the regression equation using t-values of coefficients (≥ 2.0). Filter-3 ensures the comparability of equations with a different number of variables based on the square root of the adjusted multiple correlation coefficient of the regression equation, r-bar. Filter-4 estimates equation consistency with cross-validated Q^2, using LOO cross-validation as the default option ($0.3 \leq Q^2 \leq 1.0$). To improve the information content and explanatory power of descriptors, the threshold of filter-3 is successively incremented with an increasing number of descriptors per equation, with the r-bar value of the preceding optimum model serving as the new threshold for the next generation.

To validate each CP-MLR model, a randomization test[19,20] was conducted to detect any chance correlations. This test involved multiple rounds of randomizing the activity to determine if any correlations were linked to the model. The process was repeated 100 times, and the number of scrambled activity models with regression statistics equal to or better than the original activity model was calculated to express the percentage chance correlation of the model under scrutiny. To ensure the superiority of the original model, a statistical index, r^2_{randY}(s.d.), was computed. This index represents the mean random squared multiple correlation coefficient of the regressions in the activity (Y) randomization study, along with its standard deviation from 100 simulations. To assess the goodness of fit of the models, the multiple correlation coefficient (r), the standard deviation (s), and the F-ratio between the variances of calculated and observed activities (F) were examined. The internal validation was ascertained through the cross-validated index, Q^2, from leave-one-out (Q^2_{LOO}) and leave-five-out (Q^2_{L5O}) procedures. A value greater than 0.5 of the Q^2 index suggests a reasonably robust model.

The external validation or predictive power of the derived model is based on test-set compounds. The statistical index r^2_{Test}, representing the squared correlation coefficient between the observed and predicted data of the test-set, was also computed for this purpose. A value greater than 0.5 of r^2_{Test} suggests that the model obtained from training-set has reliable predictive power.

Partial Least-Squares Analysis

The PLS or partial least-squares linear regression technique[21-23] is a useful solution for addressing issues with multi-collinearity or an over-abundance of descriptors in MLR. This method involves projecting the information in the descriptor matrix X onto a small number of latent variables, or PLS components, which are linear combinations of the original variables. The matrix Y is used concurrently to estimate the most relevant latent variables in X to predict Y variables. Before PLS progression, the descriptors are preprocessed by auto-scaling, using weights based on their standard deviation, and the

data is mean-centered. Scaling is essential due to the varying orders of magnitude in descriptor values.

To determine the optimal number of LVs, cross-validation was used. This process involved withholding samples from calibration and using them for prediction, repeating until each sample was withheld once. The predicted values of the left-out samples were then compared to the observed values using a predicted residual sum of squares or PRESS. The PRESS obtained in the cross-validation was calculated each time and a new LV was added to the model.

Applicability Domain

QSAR models are valuable for accurately predicting new analogs in a series. However, it's important to note that a model is only valid within its training domain, and new compounds must be assessed to determine whether they belong to the domain before applying the model. To determine the applicability domain[24,25], leverage values are calculated for each compound. The Williams plot, which shows standardized residuals plotted against leverage values, can be used to identify response outliers (Y-outliers) and structurally influential compounds (X-outliers) in the model. The AD is established within a squared area between ± β×(s.d.) and a leverage threshold h*. The threshold h* is typically set at 3(k + 1)/n (where n is the number of compounds in the analysis and k is the number of independent descriptors of the model), while β is usually between 2 to 3. It's important to note that prediction accuracy is uncertain for compounds with high leverage values (h > h*), while compounds with lower leverage values are likely to have a higher probability of accurately predicted values.

Results and Discussion

A set of 475 descriptors, ranging from 0D-2D DRAGON classes, were computed for 30 compounds listed in Table 1. Based on their activity profiles, seven compounds (Compounds 12, 14, 18, 21, 22, 23, and 30; Table 1) were chosen for the test-set through SYSTAT, while the rest were used for the training-set. The computed descriptors were then scaled and subjected to CP-MLR analysis to obtain significant models. During preliminary analysis, it was observed that compound 17 (Table 1) displayed unusual behavior compared to other analogs in the series. The 2,8-bis-trifluoromethyl substituents in the quinoline ring of this compound could impede its orientation and make it difficult to produce the desired effects while binding at the receptor sites. Thus, this compound was eliminated from the data set to develop models of statistical significance. Models were derived in succession using one, two, three, and four descriptors, but only the models with four descriptors remained statistically significant. A total of 14 such models, sharing 15 descriptors were obtained, but only the five most significant are documented through Equations (1)-(5).

pIC_{50} = 0.838(0.230)MSD + 0.713(0.239)GATS5e – 1.408(0.224)GATS1p
+ 0.967(0.206)C–003 + 5.795; n = 22, r = 0.913, s = 0.272, F(4,17) = 21.321,
r^2_{randY}(s.d.) = 0.402(0.135), Q^2_{LOO} = 0.699, Q^2_{L5O} = 0.683, r^2_{Test} = 0.598 ... (1)

pIC_{50} = 1.399(0.237)MSD – 1.243(0.212)GATS1v + 0.615(0.239)GATS5e
+ 1.464(0.236)H–046 + 5.103; n = 22, r = 0.914, s = 0.271, F(4,17) = 21.487,
r^2_{randY}(s.d.) = 0.424(0.125), Q^2_{LOO} = 0.585, Q^2_{L5O} = 0.697, r^2_{Test} = 0.566 ... (2)

pIC_{50} = –1.097(0.284)GGI7 – 1.530(0.227)GATS1p + 0.865(0.241)C–001
+ 0.876(0.209)C–003 + 7.038; n = 22, r = 0.914, s = 0.271, F(4,17) = 21.505,
r^2_{randY}(s.d.) = 0.410(0.132), Q^2_{LOO} = 0.582, Q^2_{L5O} = 0.597, r^2_{Test} = 0.669 ... (3)

pIC_{50} = 1.286(0.238)MSD – 1.515(0.229)GATS1p + 0.682(0.244)C–001
+ 0.970(0.238)H–046 + 5.671; n = 22, r = 0.914, s = 0.269, F(4,17) = 21.775,
r^2_{randY}(s.d.) = 0.417(0.127), Q^2_{LOO} = 0.715, Q^2_{L5O} = 0.733, r^2_{Test} = 0.838 ... (4)

pIC_{50} = 1.339(0.222)MSD + 0.724(0.226)GATS5e – 1.343(0.212)GATS1p
+ 1.142(0.220)H–046 + 5.295; n = 22, r = 0.923, s = 0.257, F(4,17) = 24.402,
r^2_{randY}(s.d.) = 0.406(0.134), Q^2_{LOO} = 0.659, Q^2_{L5O} = 0.635, r^2_{Test} = 0.580 ... (5)

The class, brief description, average regression coefficient, and total incidences for 15 identified descriptors are given in Table 3.

The F-values in all of the above-mentioned equations remained significant at a 99% level [$F_{4,17}(0.01)$ = 4.669]. The r^2 value accounted for over 83.36% of the variance in observed pIC_{50} values. The data within the parentheses, representing standard errors of regression coefficients, are significant at

Table 3 — Identified descriptors[a] along with their physical meaning, average regression coefficient, and incidence[b], in modeling the aggrecanase-1 inhibition activity

S. No.	Descriptor	Descriptor class	Physical meaning	Average regression coefficient (incidence)
1	MSD	TOPO	Mean square distance index (Balaban).	1.273 (7)
2	LP1	TOPO	Lovasz-Pelikan index (leading eigenvalue).	–0.842 (1)
3	MWC09	MWC	Molecular walk count of order 09.	–1.158 (1)
4	GGI7	GLVZ	Topological charge index of order 07.	–1.096 (1)
5	JGI6	GLVZ	Mean topological charge index of order 06.	–0.925 (2)
6	ATS5m	2DAUTO	Broto-Moreau autocorrelation of a topological structure-lag 5/ weighted by atomic masses.	–0.981 (1)
7	ATS8m	2DAUTO	Broto-Moreau autocorrelation of a topological structure-lag 8/ weighted by atomic masses.	–1.081 (1)
8	GATS1v	2DAUTO	Geary autocorrelation-lag 1/weighted by atomic van der Waals volumes.	–1.243 (1)
9	GATS5e	2DAUTO	Geary autocorrelation-lag 5/weighted by atomic Sanderson electronegativitites.	0.684 (3)
10	GATS1p	2DAUTO	Geary autocorrelation-lag 1/weighted by atomic polarizabilities.	–1.471 (13)
11	nCs	FUNC	Number of total secondary C(sp^3).	–0.866 (1)
12	nRORPh	FUNC	Number of ethers (aromatic).	–0.503 (1)
13	C-001	ACF	Corresponds to CH3R/ CH4.	0.874 (9)
14	C-003	ACF	Corresponds to CHR3.	0.881 (8)
15	H-046	ACF	Corresponds to H attached to C0(sp^3) and no X is attached to next C.	1.137 (6)

[a]The descriptors have been identified from the models, emerged from CP-MLR protocol with a training-set of 22 compounds for the inhibition of aggrecanase-1, [b]The average regression coefficient of the descriptor corresponding to all models and the total number of its incidence. The arithmetic sign of the coefficient represents the actual sign of the regression coefficient in the models

more than a 95% level. To measure the internal robustness of the models, indices Q^2_{LOO} and Q^2_{L5O} (> 0.5) were used, while the r^2_{Test} (> 0.5) was used to justify their external validation. None of these models showed any chance correlation in the activity randomization study. The direction of a specified descriptor's influence in the above models was indicated by the sign of the regression coefficient. For instance, a positive regression coefficient produced an incremental effect, while a negative regression coefficient produced a detrimental effect on aggrecanase-1 inhibition activity.

According to Table 3, the 2DAUTO and ACF classes had the highest participation in 14 models. The 2DAUTO descriptors utilize physicochemical properties like atomic masses, atomic van der Waals volumes, atomic Sanderson electronegativities, and atomic polarizabilities to calculate spatial autocorrelations on a hydrogen-depleted molecular graph. These descriptors, including ATSwk and GATSwk, describe the distribution of specific properties along a topological molecular structure through the path k. ACF descriptors, on the other hand, represent important structural moieties in the compounds, such as C-001, C-003, and H-046, which correspond to CH3R/CH4, CHR3, and H attached to C0(sp3) with no heteroatom, respectively, and X is attached to next C. The TOPO class includes descriptors like MSD and LP1, which respectively represent the mean square distance index (Balaban) and Lovasz-Pelikan index. GLVZ class descriptors, GGI7 and JGI6, represent the topological charge index of order 7 and the mean topological charge index of order 6, respectively. The MWC class descriptor, MWC09, accounts for the molecular walk count of order 09. Lastly, the FUNC class descriptors, nCs and nRORPh, denote the number of total secondary C(sp3) and the number of aromatic ethers, respectively. Table 2 explains the abbreviations used for the various classes.

The most significant Equation (5) was used to calculate the pIC_{50} values for both the training-set and test-set compounds. The calculated values are listed in Table 1 for comparison with the observed values. The graphical display in Fig. 2 shows the variation of observed *versus* calculated pIC_{50}s, except for outlier compound 17. A systematic behavior is observed for the training- and test-set compounds, indicating a good fit for the model.

In the LOO (leave-one-out) procedure, modified data-sets are created by removing one compound from the training-set so that each observation is removed only once. A model is then developed for each reduced data-set and used to predict the response

values of the deleted observations. The predicted pIC_{50} values obtained this way closely agree with the observed values (Table 1).

Equation (5) suggests that compounds exhibiting higher values of the MSD, GATS5e, and H-046 descriptors, and lower values of the GATS1p descriptor, are more effective in inhibiting aggrecanase-1. These descriptors indicate that a compound's mean square distance, polarizability weighted lag-1, and electronegativity weighted lag-5 are crucial for proper interaction with receptor sites. Moreover, the compound should be attached to an H on C0(sp3) without any heteroatom X attached to the next C. These requirements have been observed in highly potent compounds, 18-20 and 24; Table 1 of the present series.

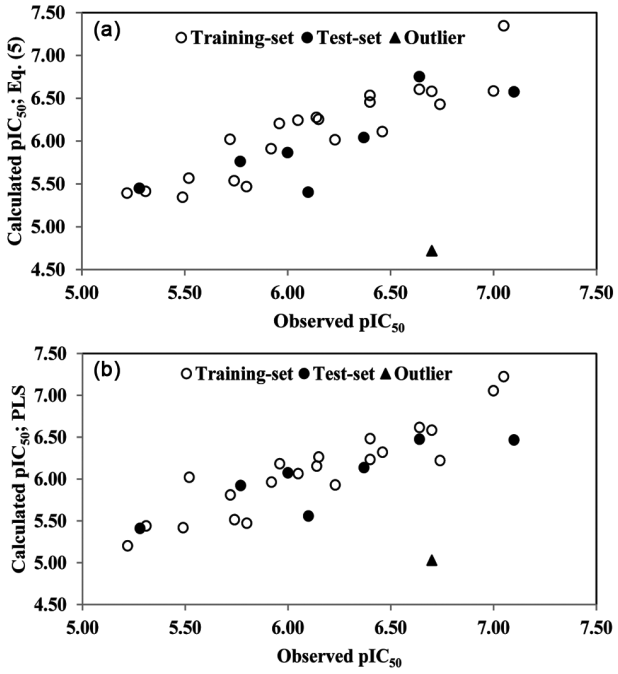


Fig. 2 — Plot of observed *versus* caculated aggrecanase-1 inhibition activity (pIC_{50}) values for biphenylsulfonamides using Equation (5) and PLS.

The 15 descriptors identified through CP-MLR (as outlined in Table 3) were subjected to PLS analysis. The aim was to develop a comprehensive "single window" structure-activity model and gauge the effectiveness of these descriptors in elucidating the inhibition actions of biphenylsulfonamide derivatives. Additionally, the analysis facilitates a comparison of the relative significance among these descriptors. Through the normalized regression coefficients, fraction contributions were obtained, which permits a comparison within the modeled activity.

During the PLS analysis, the descriptors were standardized to have a zero mean and unit standard deviation. This was done to ensure that each descriptor contributed equally to the analysis. Cross-validation revealed that two components were optimal for the documented descriptors, explaining 84.46% of the variance in the observed pIC_{50}s. Table 4 presents

Table 4 — PLS and MLR-like PLS equations from the descriptors of CP-MLR identified models for aggrecanase-1 inhibition activity

A: PLS equation

PLS components	PLS coefficient (s. e.)[a]
Component-1	0.213 (0.025)
Component-2	–0.173 (0.030)
Constant	6.061

B: PLS regression statistics

Symbol	Estimate
n	22
r	0.919
s	0.249
F	51.457
Q^2_{LOO}	0.742
Q^2_{L5O}	0.750
r^2_{Test}	0.628

C: MLR-like PLS equation

S. No.	Descriptor	MLR-like coefficient[b]	F.C. (order)[b]	S. No.	Descriptor	MLR-like coefficient[b]	F.C. (order)[b]
1	MSD	0.326	0.071 (7)	9	GATS5e	0.359	0.072 (6)
2	LP1	–0.158	–0.032 (10)	10	GATS1p	–0.785	–0.169 (1)
3	MWC09	–0.125	–0.018 (13)	11	Cs	–0.132	–0.020 (12)
4	GGI7	–0.195	–0.033 (9)	12	nRORPh	–0.120	–0.028 (11)
5	JGI6	–0.224	–0.051 (8)	13	C-001	0.459	0.096 (5)
6	ATS5m	0.031	0.005 (14)	14	C-003	0.629	0.146 (2)
7	ATS8m	–0.025	–0.004 (15)	15	H-046	0.647	0.136 (3)
8	GATS1v	–0.524	–0.120 (4)		Constant	6.449	

[a]Regression coefficient of PLS factor and its standard error, [b]Coefficients of MLR-like PLS equation in terms of descriptors for their original values; F.C. is fraction contribution of regression coefficient, computed from the normalized regression coefficients obtained from the autoscaled (zero mean and unit standard deviation) data

the PLS equation, regression statistics, MLR-like PLS coefficients for 15 descriptors, and their fractional contributions to activity.

The calculated pIC_{50} values remained in close agreement with the observed ones (Table 1), while the plots showing variation of observed *versus* calculated pIC_{50} is given in Fig. 2.

Fig. 3 displays the contribution of normalized regression coefficients of the descriptors to aggrecanase-1 inhibition activity. The Fig. also shows different orders that indicate the level of significance of the descriptors. The orders are also listed in Table 4.

When identifying biological activity, the importance of a descriptor is inversely proportional to its order. After analysis, GATS1p, C-003, and H-046 descriptors were found to be the most influential ones in modeling aggrecanase-1 inhibition activity. The other significant descriptors, in decreasing order of importance, are GATS1v, C-001, GATS5e, MSD, JGI6, and GGI7 (Table 4, S. Nos. 8, 13, 9, 1, 5, and 4). Descriptors that positively contribute to the activity should be enhanced with higher values to further improve it. On the other hand, descriptors with negative contributions should be avoided as they lower the activity. Thus, lower or more negative values of such descriptors may actually increase the activity of a compound.

The models resulting from all 30 compounds in the series were analyzed for their applicability domain (AD). The AD was represented on the Williams plot, which illustrates the relationship between standardized residuals and leverage (h_i) values. To derive the corresponding model based on the entire data-set, the most significant descriptors from Equation (5) were used. The resulting model is presented in Equation (6), and the standardized residuals and leverage values calculated in conjunction with it were used to determine the ADs.

$$pIC_{50} = 0.900(0.287)MSD + 0.195(0.333)GATS5e - 1.521(0.279)GATS1p + 0.816(0.314)H–046 + 6.025; n = 30, r = 0.790, s = 0.389, F(4,25) = 10.410, r^2_{randY}(s.d.) = 0.355(0.117), Q^2_{LOO} = 0.232, Q^2_{L5O} = 0.368 \quad ... (6)$$

The limits of normal values for the standardized residuals (response or Y-outliers) were set as ±2.5×s.d. while the leverage threshold was h*. The graphical representation for this model is given in Fig. 4.

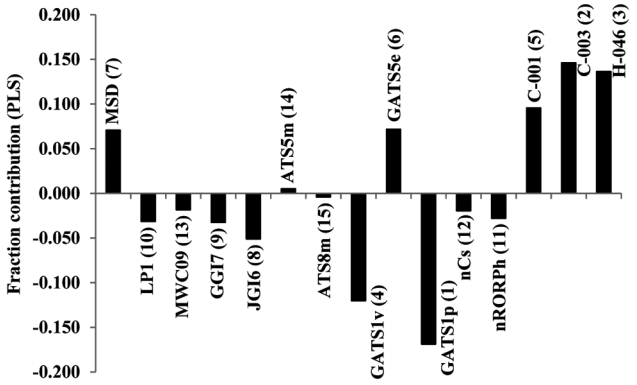


Fig. 3 — Plot between fraction contribution of MLR-like PLS coefficients (normalized) and 15 identified descriptors (Table 4) associated with aggrecanase-1 inhibition activity values of biphenylsulfonamides.

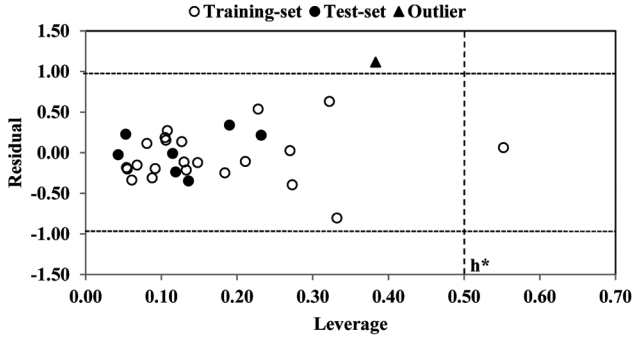


Fig. 4 — Williams plot based on whole data-set for aggrecanase-1 inhibition activity values of biphenylsulfonamides listed in Table 1 (h* value is 0.50 and residual limits are ±2.5×s.d.)

From this Fig. compound **19**, with larger leverage value (= 0.552), was identified as the most influential analog of present series. Similarly compound **17**, with a larger residual (= ±1.116), appeared as an obvious outlier, which on elimination, resulted in a significant correlation Equation (7)

$$pIC_{50} = 1.457(0.226)MSD + 0.588(0.244)GATS5e - 1.476(0.195)GATS1p + 1.057(0.224)H–046 + 5.305; n = 29, r = 0.905, s = 0.271, F(4,24) = 27.043, r^2_{randY}(s.d.) = 0.378(0.126), Q^2_{LOO} = 0.685, Q^2_{L5O} = 0.655 \quad ... (7)$$

Compared to Equation (6), Equation (7) shows a marked improvement in statistical parameters. This indicates that the descriptors utilized are suitable for evaluating the inhibition activity of biphenylsulfonamide derivatives on aggrecanase-1. Furthermore, all other data-points were found to be within the acceptable range of this model, demonstrating its capability to assess both the training and test compounds. The model

boasts high-quality parameters, maintains excellent fitting power, and can evaluate external data.

Conclusion

The use of QSAR models has yielded valuable insight into the inhibitory activity of biphenylsulfonamide derivatives against aggrecanase-1. Notably, a significant model has identified certain features, including mean square distance (MSD), polarizability weighted lag-1 (GATS1p), and electronegativity weighted lag-5 (GATS5e), as crucial in ensuring proper interaction with receptor sites. Structural component, such as H attached to C0(sp3) with no heteroatom X attached at next C (H-046), has also been found to be desirable. Ultimately, the study concluded that compounds with higher values of MSD, GATS5e, and H-046, and lower values of GATS1p, are more effective in inhibiting aggrecanase-1. These requirements were observed in several potent compounds within the present series.

According to the PLS analysis, it was determined that there exists a structure-activity model with a "single window" that can be derived from 15 descriptors. This analysis also allowed for a comparison of the importance of these descriptors. The resulting model, which utilized two optimal components, accounted for 84.46% of the variance in the observed activity values of the compounds. Utilizing the strategies outlined in the CP-MLR and PLS analyses could prove beneficial in discovering new analogs for future investigation.

Based on the applicability domain (AD) analysis, it was found that the suggested models have good predictability. All the compounds, except for one obvious outlier, were evaluated correctly and were within the AD of the proposed model. Additionally, this study identified one structurally influential compound.

Acknowledgement

The authors appreciate the help from their colleagues, and one of the authors, NS is thankful to her institution for providing necessary facilities to complete this work.

References

1 Liu R -Q & Trzaskos J M, *Curr Med Chem*, 4 (2005) 251.
2 Summer E U, Qvist P & Tankó L B, *Drug Dev Res*, 68 (2007) 1.
3 Tortorella M D, Burn T C, Pratta M A, Abbaszade I, Hollis J M, Liu R, Rosenfeld S A, Copeland R A, Decicco C P, Wynn R, Rockwell A, Yang F, Duke J L, Solomon K, George H, Bruckner R, Nagase H, Itoh Y, Ellis D M, Ross H, Wiswall B H, Murphy K, Hillman M C, Hollis G F, Newton R C, Magolda R L, Trzaskos J M & Arner E C, *Science*, 284 (1999) 1664.
4 Glasson S S, Askew R, Sheppard B, Carito B A, Blanchet T, Ma H -L, Flannery C R, Kanki K, Wang E, Peluso D, Yang Z, Majumdar M K & Morris E A, *Arthritis Rheum*, 50 (2004) 2547.
5 Glasson S S, Askew R, Sheppard B, Carito B A, Blanchet T, Ma H-L, Flannery C R, Peluso D, Kanki K, Yang Z, Majumdar M K & Morris E A, *Nature*, 434 (2005) 644.
6 Majumdar M K, Askew R, Schelling S, Stedman N, Blanchet T, Hopkins B, Morris E A & Glasson S S, *Arthritis Rheum*, 56 (2007) 3670.
7 Xiang J S, Hu Y, Rush T S, Thomason J R, Ipek M, Sum P -S, Abrous L, Sabatini J J, Georgiadis K, Reifenberg E, Majumdar M, Morris E A, Tam S, *Bioorg Med Chem Lett*, 16 (2006) 311.
8 Hopper D W, Vera M D, How D, Sabatini J, Xiang J S, Ipek M, Thomason J, Hu Y, Feyfant E, Wang Q, Georgiadis K E, Reifenberg E, Sheldon R T, Keohan C C, Majumdar M K, Morris E A, Skotnicki J, Sum P-E, *Bioorg Med Chem Lett*, 19 (2009), 2487-91.
9 SYSTAT, Version 7.0, SPSS Inc, 444 North Michigan Avenue, Chicago IL, 60611.http://www.spss.com.
10 Chemdraw ultra 6.0 and Chem3D ultra, Cambridge Soft Corporation, Cambridge, USA. http://www.camsoft.com.
11 DRAGON software version 3.0-2003 by Todeschini R, Consonni V, Mauri A & Pavan M, Milano, Italy. https://disat.unimib.it/chm/Dragon.htm
12 Prabhakar Y S, *QSAR Comb Sci*, 22 (2003) 583.
13 Golbraikh A & Tropsha A, *J Mol Grap Mod*, 20 (2002) 269.
14 Sharma B K, Pilania P, Singh P & Prabhakar Y S, *SAR QSAR Environ Res*, 21 (2010) 169.
15 Sharma B K, Singh P, Sarbhai K & Prabhakar Y S, *SAR QSAR Environ Res*, 21 (2010) 369.
16 Sharma B K, Singh P, Shekhawat M, Sarbhai K & Prabhakar Y S, *SAR QSAR Environ Res*, 22 (2011) 365.
17 Sharma B K, Singh P, Pilania P, Shekhawat M & Prabhakar Y S, *J Enzy Inhibn & Med Chem*, 27 (2012) 249.
18 Sharma B K, Singh P & Prabhakar Y S, *Brit J Pharm Res*, 3 (2013) 697.
19 So S–S & Karplus M, *J Med Chem*, 40 (1997) 4347.
20 Prabhakar Y S, Solomon V R, Rawal R, Gupta M K & Katti S B, *QSAR Comb Sci*, 23 (2004) 234.
21 Wold S, *Technometry*, 20 (1978) 397.
22 Kettaneh N, Berglund A E & Wold S, *Comput Stat Data Anal*, 48 (2005) 69. http://dx.doi.org/10.1016/j.csda.2003.11.027
23 Stahle L & Wold S, *Prog Med Chem*, 25 (1988) 291.
24 Gramatica P, *QSAR Comb Sci*, 26 (2007) 694.
25 Eriksson L, Jaworska J, Worth A P, Cronin M T D, McDowell R M & Gramatica P, *Env Health Persp*, 111 (2003) 1361.