Turkish Journal of Chemistry
http://journals.tubitak.gov.tr/chem/
Research Article

Turk J Chem
(2018) 42: 1577 – 1597

doi:10.3906/kim-1801-86

4D-QSAR analysis and pharmacophore modeling for alkynylphenoxyacetic acids as CRTh2 (DP2) receptor antagonists

Semiha KÖPRÜ, Emin SARIPINAR*
Department of Chemistry, Faculty of Science, Erciyes University, Kayseri, Turkey

Received: 30.01.2018 • **Accepted/Published Online:** 22.07.2018 • **Final Version:** 06.12.2018

Abstract: In this study, we performed the pharmacophore modeling and 4D-QSAR research of alkynylphenoxyacetic acid analogues as CRTh2 receptor opponent agents by utilizing the electron conformational genetic algorithm method. Quantum chemical calculations and conformational analyses of the compounds were carried out at HF/6-31G* level. Then electron conformational matrices of congruity were prepared for each conformer of each compound, which are represented by electronic and structural properties. As a result of the comparison of the matrices that are called electron conformational submatrices of activity, the pharmacophoric group of the compounds responsible for the activity was found at the determined tolerance intervals. The genetic algorithm and nonlinear least squares regression methods were applied to estimate the conjectural activity and investigate the most reliable molecular identifiers as feature selection from a large parameter pool. The compounds in the dataset were randomly segregated for training (61 compounds) and test sets (25 compounds) for statistical analysis. Validation of the 4D-QSAR model was appraised by the leave-one-out cross-validation technique. For the best model the $r^2_{training}$, r^2_{test}, q^2, q^2_{ext1}, q^2_{ext2}, and q^2_{ext3} values were found to be 0.8580, 0.8571, 0.8105, 0.8282, 0.8145, and 0.8475, respectively.

Key words: Electron conformational genetic algorithm, 4D-QSAR, CRTh2 receptor antagonist, pharmacophore, drug design, alkynylphenoxyacetic acid

1. Introduction

The CRTh2 receptor (also known as DP2), which is an arachidonic acid metabolite and is released from mast cells, seems to act as a central negotiator for the treatment of asthma and other inflammatory illnesses.[1,2] Allergic diseases and inflammation are commonly related to immunoglobulin E (IgE) production and mast cell activation.[3] Mast cells indirectly participate in asthmatic reactions. During an acute asthmatic attack, prostaglandin D2 (PGD2) is released into the lungs by mast cells and bronchoconstriction occurs.[4–6] PGD2 exerts its biological responses by means of two high affinity G protein-coupled receptors, the DP1 receptor and a chemoattractant receptor-homologous molecule represented classically by T-helper 2 cells (CRTh2) receptor. CRTh2 expresses on Th2 cells, eosinophils, and basophils and intercedes in the chemotactic activation of these cells in response to PGD2.[7,8] A large number of articles have recently been published on the identification and optimization of CRTh2 antagonists.[9] In this context, it has been proposed that selective antagonism of the CRTh2 receptor may have a beneficial effect on the allergic physiological response.[10]

Drug design is a composite and laborious process that includes an association of innovative experimental

*Correspondence: emin@erciyes.edu.tr

and computational methods.[11] Quantitative structure-activity relationships (QSARs) are an important part of modern drug design and medicinal chemistry and are used to detect the correlation between the biological activity of a set of molecules and molecular descriptors by means of a statistical or mathematical tool.[12−16] In a QSAR study, molecules are characterized by the presence of molecular descriptors that contain physicochemical and structural features. The main aim of QSAR modeling is to assemble a correlation between the activities of compounds and their physicochemical, topological, and geometrical properties. The selection of potential molecular descriptors from a set of biologically active conformers is the most important step in QSAR model generation to understand the nature of molecular features prior to actual QSAR model building.[17,18]

Many different QSAR methods have been developed over the years and new techniques are still being developed at present. Basically, these approaches differ in their definitions of the principles and levels and the molecular structure. The QSAR modeling varies from one-dimensional (1D) to four-dimensional (4D) levels in terms of the models. One-dimensional models are based solely on the gross formula of one compound and reflect only the composition of one molecule. In addition, it is clearly not possible for "structure-activity" relations to be sufficiently resolved using such approaches.[19] 2D QSAR provides information on the structure of a compound and is based on its structural formula; for that reason, these models represent the topology of the molecule.[20] Three-dimensional QSAR models deliver precise structural information for a conformer only, taking into consideration the molecular composition, topology, and 3D shape. These models are widely used and the choice of conformer is often coincidental. Recently, an optimistic computational approach related to specific biological properties at a defined receptor was the four-dimensional QSAR (4D-QSAR) formalism, which was proposed by Hop?nger et al. in 1997. The main focus of this approach is to combine the pharmacophore, conformational, and alignment freedom into the design of 3D-QSAR models for the training sets of the structure activity datasets, making the fourth "dimension".[21]

The criteria for 4D-QSAR analysis are the need to regard multiple (i) conformations, (ii) alignments, and (iii) infrastructure groups in erecting QSAR models. The aforementioned "QSAR degrees of freedom" are generally kept stable in 3D-QSAR analysis.[11] In this approach, the molecular descriptors are the occupancy frequencies of the different atom types in the cubic grid cells during the molecular dynamics simulation time according to each trial alignment, corresponding to an ensemble averaging of conformational behavior.[16,22,23]

The genetic algorithm is the most applied powerful optimization technique in different areas. Furthermore, this method has been widely used for selecting the most relevant descriptors in QSAR studies.[24−26] The aim of this work is to determine the pharmacophore and to predict the biological activity for alkynylphenoxyacetic acid derivatives using the electron conformational-genetic algorithm (EC-GA) technique. This methodology is preferred for each conformational ensemble profile compound in place of a single conformation.

Examples of successful application of the EC-GA method have been studied in different series by our group to determine the pharmacophore and to predict biological activity as a 4D-QSAR approach for N-morpholino triaminotriazines, benzotriazines, penicillins, 1,4-dihydropyridines, HEPT derivatives, TIBOs, pyrrolo[2,1-c][1,4]benzodiazepines, and finally organometallic Ru(II) complexes, which was a 4D-QSAR study applied to Ru complexes to make a quantitative prediction of activity. The basis and details of this method are given in the literature.[27−34]

2. Results and discussion

A dataset containing 86 compounds consisting of CRTh2 receptor antagonist derivatives was used in the 4D-QSAR study. The structural and experimental data for the alkynylphenoxyacetic acid derivatives used in this study were taken from the literature[35] and are listed in Table 1. The number of conformers and the predicted activities of compounds corresponding to experimental values are also given in Table 1.

Mulliken charges and bond orders/interatomic distances were utilized to generate the electron conformational matrices of congruity (ECMCs) for individual conformations of the entire compound set and placed in diagonal and nondiagonal positions, respectively. Nondiagonal elements are of two types: bond orders for chemically bonded atom pairs and interatomic distances for nonbonded atom pairs. The ECMC of reference compound 85 for the lowest energy conformer is given in Figure 1, which contains the distance and charge values for compounds.

C1	C4	C2	C6	C5	C3	Cl1	O1	C7	C8	O2	O3	H6	C9	C10	C11	C12	C13	C14	C15	C16	S1	O4	O5	C17	C20	C18	C19	C22	
-0.133	2.771	1.390	1.428	2.405	2.403	2.704	3.660	4.422	5.440	6.298	5.682	5.098	2.455	3.544	4.933	7.679	5.865	5.617	6.984	7.185	8.501	8.262	9.692	8.917	10.200	10.787	5.419	5.224	C1
	-0.222	2.410	2.386	1.436	1.434	3.996	2.367	3.342	4.669	5.688	4.932	4.239	3.731	4.873	6.274	9.001	6.888	7.224	8.534	8.256	9.274	8.974	10.634	9.203	10.288	10.530	3.698	7.081	C4
		0.006	2.398	2.787	1.391	4.004	2.400	3.149	4.067	4.967	4.296	3.752	1.020	2.631	4.073	6.853	4.881	4.956	6.294	6.239	7.477	7.288	8.735	7.716	8.977	9.480	4.313	4.833	C2
			-0.185	1.415	2.756	1.035	4.123	4.983	6.206	7.122	6.471	5.802	3.716	4.861	6.275	9.037	7.158	6.974	8.347	8.501	9.753	9.444	10.986	10.083	11.310	11.815	5.668	6.517	C6
				-0.168	2.398	2.711	3.637	4.560	5.899	6.876	6.171	5.461	4.227	5.418	6.860	9.639	7.613	7.688	9.049	8.987	10.112	9.777	11.426	10.225	11.365	11.713	4.944	7.350	C5
					0.440	4.503	0.789	2.394	3.563	4.592	3.815	3.183	2.465	3.551	4.924	7.630	5.523	5.921	7.202	6.884	7.936	7.709	9.281	7.895	9.039	9.346	3.275	5.913	C3
						-0.010	5.870	6.694	7.921	8.796	8.189	7.514	5.155	6.203	7.538	10.217	8.498	8.068	9.423	9.792	11.106	10.766	12.266	11.596	12.848	13.437	7.320	7.421	Cl1
							-0.747	0.750	2.380	3.503	2.582	2.010	2.831	3.602	4.758	7.254	5.120	5.846	6.998	6.412	7.294	7.136	8.674	6.991	8.057	8.214	2.422	6.082	O1
								0.097	0.946	2.373	2.427	2.398	3.235	3.740	4.678	6.952	4.657	5.956	6.960	5.903	6.494	6.151	7.927	6.171	7.087	7.289	0.974	6.511	C7
									0.689	1.879	1.068	1.879	3.641	3.689	4.199	6.044	3.923	5.434	6.223	4.977	5.429	5.299	6.838	4.840	5.759	5.868	2.506	6.216	C8
										-0.581	2.196	2.978	4.381	4.208	4.414	5.827	3.754	5.687	6.268	4.615	4.694	4.433	6.113	4.036	4.774	4.943	3.018	6.698	O2
											-0.708	0.658	3.798	3.764	4.176	5.909	4.121	5.180	5.948	5.063	5.706	5.898	7.013	4.939	5.976	5.923	3.429	5.830	O3
												0.538	3.466	3.661	4.324	6.336	4.556	5.246	6.171	5.602	6.441	6.634	7.734	5.795	6.897	6.861	3.449	5.653	H6
													-0.119	3.063	2.634	5.414	3.512	3.571	4.881	4.843	6.177	6.119	7.381	6.517	7.857	8.465	4.669	3.675	C9
														-0.191	1.058	4.223	2.434	2.478	3.730	3.708	5.151	5.245	6.284	5.603	7.010	7.719	5.242	2.923	C10
															-0.015	2.782	1.411	1.373	2.399	2.399	4.020	4.381	5.005	4.647	6.120	6.960	6.159	2.534	C11
																-0.149	2.407	2.420	1.430	1.398	2.742	3.805	2.937	3.716	5.081	6.136	8.314	3.798	C12
																	-0.133	2.423	2.766	1.375	2.724	3.027	3.871	3.430	4.841	5.799	6.002	3.808	C13
																		0.095	1.455	2.786	4.554	5.200	5.201	5.268	6.778	7.629	7.461	0.960	C14
																			-0.227	2.388	4.025	4.939	4.317	4.860	6.310	7.244	8.422	2.515	C15
																				-0.384	0.977	2.617	2.611	2.807	4.201	5.315	7.180	4.293	C16
																					1.548	1.377	1.470	1.013	2.756	4.125	7.523	6.061	S1
																						-0.844	2.489	2.630	3.191	4.590	7.027	6.668	O4
																							-0.787	2.635	3.202	4.596	8.970	6.666	O5
																								-0.548	0.952	2.512	7.049	6.705	C17
																									-0.346	0.966	7.719	8.228	C20
																										-0.478	7.756	8.992	C18
																											-0.501	8.031	C19
																												-0.501	C22

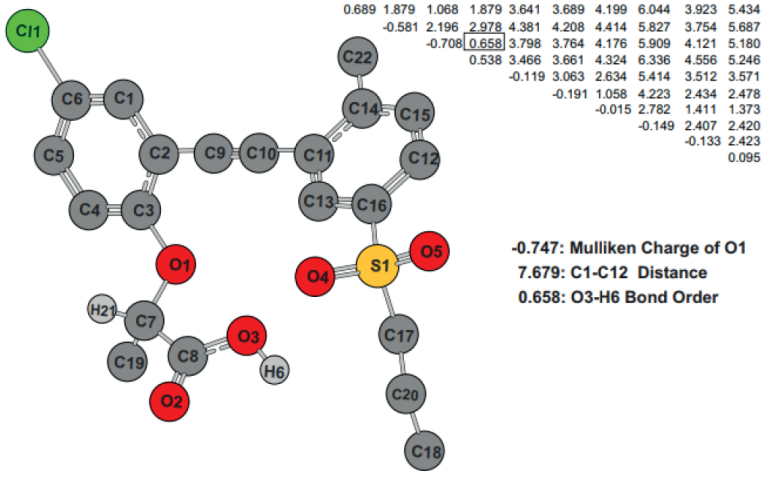


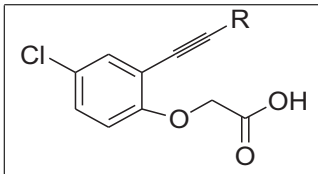
Figure 1. ECMC of template compound 85 for the lowest energy conformer. The diagonal elements show Mulliken charges and the nondiagonal elements indicate the distance for bonded atoms, which is not dependent on the bond order. Carbon-bonded hydrogen atoms are not included in the ECMC for simplicity.

The pharmacophore group contains the minimum atomic group required for the activity. Therefore, 86 molecules are classified as high and low activity. The lowest energy conformation of component 85, which has the highest activity, is taken as a reference.

Alkynylphenoxyacetic acid derivatives were divided into two groups indiscriminately as being of high and low activity for the detection of the pharmacophore. The compounds with A^{exp} of ≥ 7.721 (44 compounds) were categorized as high-activity compounds, while the other compounds (42 compounds) were considered as low-activity compounds. According to Eq. (1) below, the probability parameters of the pharmacophore occurring from these atoms are given as $P_\alpha = 0.8970$, $\alpha_a = 0.6940$. As a result, we found that the pharmacophore group consists of C3, O1, C8, O2, O3, H6, and C22 atoms.

Examination of the pharmacophoric atoms in the molecule indicates that the presence of C3, O1, C8, O2,

Table 1. Chemical structures, experimental activity (A^{exp}), and calculated (A^{calc}) activity values for the alkynylphenoxyacetic acid dataset.



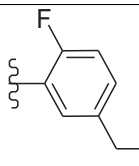
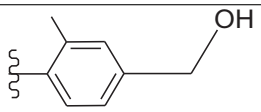
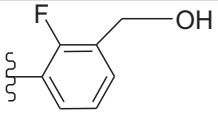
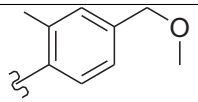
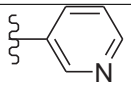
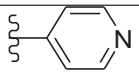
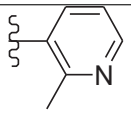
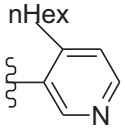
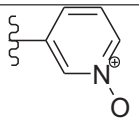
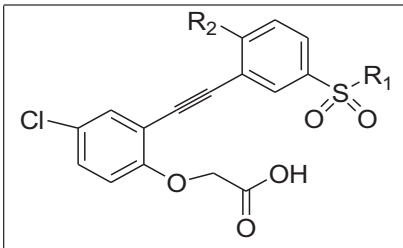
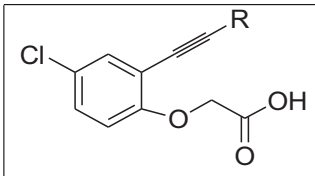
No.[a]	R	CN[b]	$A^{exp(c)}$	$A^{calc(d)}$
1	Phenyl	7	6.9666	7.0693
2	4-Chlorophenyl	5	7.2218	7.1129
3*	2-Methoxyphenyl	5	6.7212	6.9355
4*	3-Trifluorophenyl	6	7.0362	7.3699
5*	2,4-Difluorophenyl	4	6.9957	6.7437
6	5-Chlorothiophene-2-yl	5	6.7375	6.6875
7		24	7.1367	7.3736
8		20	7.5528	7.5666
9		16	7.3872	7.3707
10*		25	7.6576	8.0428
11		5	6.9318	7.2230
12		4	6.2211	6.1639
13		7	6.8447	7.1030
14		20	7.3665	7.4234
15*		5	7.7696	7.5258

Table 1. Continued.



No.[a]	R_1	R_2	CN[b]	A$^{exp(c)}$	A$^{calc(d)}$
16	Me	H	11	7.8239	7.4820
17*	$CH_2CH_2CH_2OH$	H	25	7.5528	7.8179
18	CH_2CH_2OH	H	25	8.0000	8.0006
19	Me	Me	13	8.3010	8.4302
20	Me	F	12	8.4815	8.0985
21	Me	Cl	11	8.4559	8.3839
22*	Me	n-Pr	25	8.0655	8.5468
23	Me	i-Pr	15	8.3372	8.0553
24*	Et	Me	21	8.0655	8.2677
25*	n-Pr	Me	25	8.6990	8.7276
26	i-Pr	Me	20	8.2218	8.2570
27*	i-Bu	Me	20	8.5229	8.6269
28*	CH_2Ph	Me	15	8.3010	8.3498
29	CH_2CH_2OH	Me	20	7.7212	8.0937
30	$CH_2CH_2CH_2OH$	Me	25	8.0969	7.9057
31*	n-Pr	F	20	8.4685	8.4017
32*	i-Pr	F	15	8.3979	8.3196
33	n-Pr	Cl	20	8.6778	8.6764
34	i-Pr	Cl	21	8.5376	8.4283
35	NMe_2	H	16	7.8539	7.4883
36	NMe_2	Me	19	8.3010	8.2786
37	NEt_2	Me	20	8.4685	8.5478
38	NH-t-Bu	Me	20	8.3010	8.4911
39	NHMe	Me	20	8.0000	8.0743
40	NHEt	Me	20	8.3565	8.6213
41	NH-i-Pr	Me	20	8.4685	8.5904
42	NMe-i-Pr	Me	20	8.4202	8.2562
43	Piperidine	Me	24	8.3010	8.5297
44	Morpholine	Me	25	8.4437	8.1639
45	2-Methylpiperidine	Me	20	8.7212	8.7539
46	$NMeCH_2CH_2OMe$	Me	25	8.3872	8.2260
47	$NHCH_2CH_2OMe$	Me	20	8.4437	8.7529
48	N-Methylpiperazine	Me	10	7.6778	7.8350

Table 1. Continued.



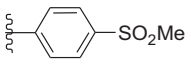
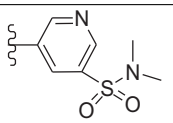
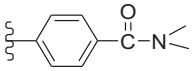
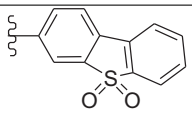
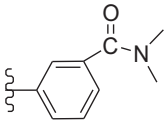
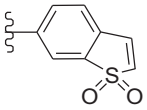
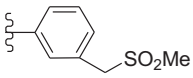
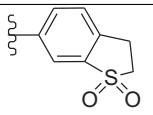
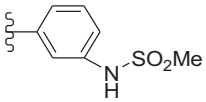
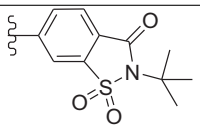
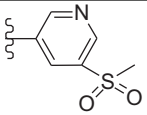
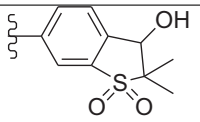
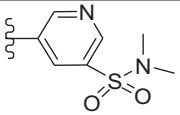
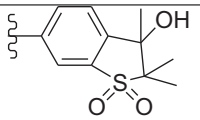
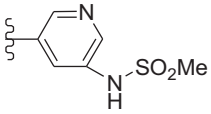
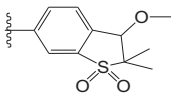
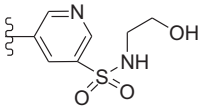
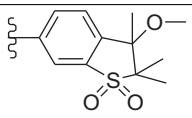
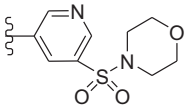
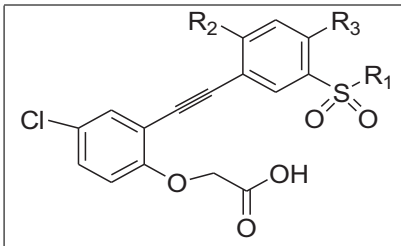
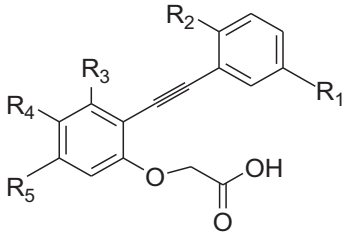
No.[a]	R	CN[b]	A$^{exp(c)}$	A$^{calc(d)}$	No.[a]	R	CN[b]	A$^{exp(c)}$	A$^{calc(d)}$
49		9	7.0655	7.2086	**59***		17	7.3872	7.9430
50		8	7.2596	7.1479	**60***		5	7.9208	7.9445
51		11	7.2840	7.4088	**61**		6	7.6021	7.4524
52*		20	7.3188	7.5752	**62**		8	7.6383	7.5021
53*		20	7.2676	7.4253	**63**		5	7.5086	7.7702
54*		11	7.2218	7.4364	**64**		15	7.4815	7.3576
55*		19	7.4318	7.4745	**65**		19	7.5528	7.4084
56		20	7.7212	7.1742	**66**		11	7.6778	7.7684
57		14	6.5214	6.5918	**67**		16	7.6021	7.6864
58		20	7.3872	7.4862					

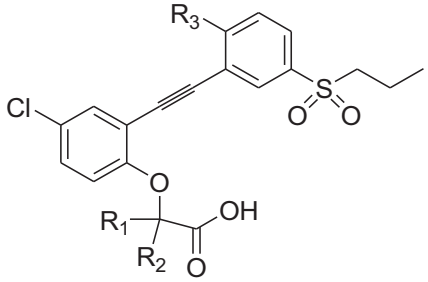
Table 1. Continued.



No.[a]	R_1	R_2	R_3	CN[b]	A$^{exp(c)}$	A$^{calc(d)}$		
68	Me	H	4-Methoxyphenyl	**20**	8.1549	8.4355		
69	Me	H	3-Methoxyphenyl	**20**	8.0969	8.0604		
70	n-Pr	Me	NHCOMe	**20**	8.8861	8.6469		
71*	i-Pr	H	CONMe2	**14**	7.1871	7.5608		
72*	i-Pr	H	CO-Morpholine	**13**	7.0605	7.2370		



No.[a]	R_1	R_2	R_3	R_4	R_5	CN[b]	A$^{exp(c)}$	A$^{calc(d)}$
73	H	Cl	H	H	H	5	6.1267	6.5414
74*	H	Cl	H	Me	H	6	6.5482	6.5599
75	SO2Pr	H	Cl	H	H	20	7.2596	7.5294
76*	SO2Pr	H	Naphthyl	H	H	20	7.3565	7.1783
77*	SO2Pr	H	H	CF_3	H	24	8.0000	7.9138
78	H	Cl	H	CN	H	7	7.6198	7.3701
79	SO2Pr	Me	H	CN	H	20	8.1549	7.8090
80	SO2Me	iPr	H	CN	H	17	7.9208	7.8265
81*	SO2Pr	H	Phenyl	H	H	19	7.7212	7.4786
82	SO2Pr	H	2,4-Dimethyl-thiazol-5-yl	H	H	20	7.2676	7.4637
83	SO2Pr	H	3-Thienyl	H	H	20	7.7447	7.4758
84	SO2Pr	H	2-Thienyl	H	H	20	7.9208	7.4687



No.[a]	R_1	R_2	R_3	CN[b]	A$^{exp(c)}$	A$^{calc(d)}$		
85	Me	H	Me	25	8.9208	8.9208		
86	Et	H	Me	21	7.5229	8.3662		

[a]The molecules that are marked with * are test set compounds.

[b]Number of conformers for each compound.

[c]Experimental activity values of the dataset.

[d]Calculated activity values of the dataset.

O3, H6, and C22 atoms plays an important role in the effectiveness of alkynylphenoxyacetic acid derivatives in the biochemical interaction mechanism. Three of the seven atoms in the pharmacophore group are oxygen atoms O1, O2, and O3 and the highest positive charge is concentrated in the C8 atom, which is part of the carbonyl group. The presence of C8, O2, O3, and H6 atoms in both pharmacophoric group atoms and in 9 important parameters affecting activity indicate the importance of the carboxy acid group in the drug interaction mechanism. The analysis of the pharmacophore group shows that the presence of C3, O1, C8, O2, O3, H6, and C22 atoms is a principal component of activity within the specifics of the drug-receptor biochemical interaction mechanism for alkynylphenoxyacetic acid derivatives.

The O1, O2, and O3 atoms are defined as hydrogen-bond acceptors and H6 is a hydrogen-bond donor. These atoms are heavily influenced by hydrogen bonds, ion-dipole interactions, and electrostatic effects. The O2, C8, O3, and H6 atoms are on a rigid plane. The C3 atom in the aromatic ring and aliphatic C22 atoms are located in hydrophobic positions. The O1, O2, O3, and C22 atoms are negatively charged and the C3, C8, and H6 atoms are positively charged. In high-activity compounds, the highest tolerance value is for the distance between the atoms C22 and H6, which represents the flexibility of this position, and the C8-H6 distance with the lowest tolerance value, which corresponds to a rigid plane. First, the positioning flexibility of all atoms in the overall pharmacophore structure is taken into account in the EC-GA method, and then the relationships between the flexibility and activity of the compounds are used for bioactivity predictions.

The tolerance values in the electron conformational submatrices of activity (ECSAs) for the high- and low-activity molecules are shown in Table 2. The yellow atoms in the molecule represent the pharmacophore group atoms. The matrices in Table 2 correspond to the tolerance matrices for the reference compound, high-activity, low-activity, and all compounds, respectively. From the table, we can see that the tolerance values of high- and low-activity compounds are different from each other.

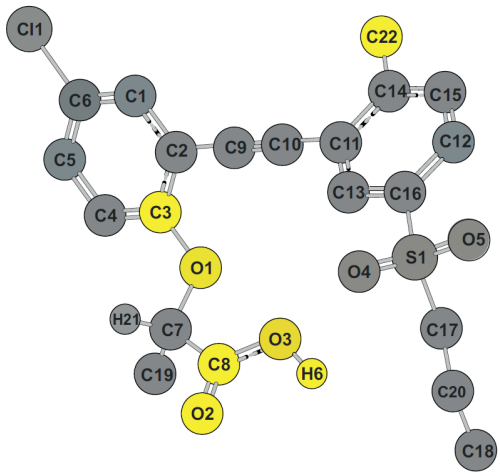
Generally, the tolerance values of high-activity compounds for ECSAs are lower than those of low-activity compounds. For instance, the tolerance value for the distance between atoms H6 and C22 can be given as ± 1.909 for high-activity compounds and ± 5.312 for low-activity compounds.

In Table 2, the first submatrix is the ECSA showing the tolerance values of the reference compound. The second and third matrices show tolerance values for high-activity compounds and low-activity compounds, respectively. Then the maximum tolerance values were calculated for all the conformers of all compounds without any restriction on the tolerance values of the pharmacophore group. The last submatrix shows the tolerance values for 1338 conformers of all 86 compounds.

The most important part of the development of a model is the separation of data into training and test sets. To prove the validity of the QSAR model, 86 molecules were randomly segregated into a training set and a test set. The constructed EC-GA model, which was established using 61 training compounds, was tested with 25 test compounds. The molecules were arbitrarily chosen, then fixed and scanned (between 1 and 15) to select the most suitable subparameter set from a different number of parameters using MATLAB[36] software.

In QSAR analysis, it is quite complex and time-consuming to search for the most suitable subset among a large number of parameters, because there are many possible combinations in order to establish a correct, robust, and valid model. Generally, selecting a proper subset of descriptors from a large descriptor pool is very hard in the QSAR modeling process. Therefore, we developed the GA procedure written in a MATLAB environment for the selection of the most important descriptors. The genetic algorithm codes in this study were written in MATLAB and were run on parallel (multicore and multiprocessor) computers. The GA forms random chromosomal subclusters and covers input parameters for the QSAR model.

Table 2. (a) The pharmacophore group ECSAs of the reference molecule (85) for alkynylphenoxyacetic acid derivatives; (b) tolerance matrix of ECSAs for 27 compounds with high activity; (c) tolerance matrix of ECSAs for 8 compounds with low activity; (d) tolerance values for 1338 conformers of 86 compounds.



a) ECSA of reference compound (85)							
C3	O1	C8	O2	O3	H6	C22	Pha atoms
0.440	0.789	3.563	4.592	3.815	3.183	5.913	C3
	–0.747	2.380	3.503	2.582	2.010	6.082	O1
		0.689	1.879	1.068	1.879	6.216	C8
			–0.581	2.196	2.978	6.698	O2
				–0.708	0.658	5.830	O3
					0.538	5.653	H6
						–0.501	C22
b) Tolerance values for 44 compounds with high activity							
C3	O1	C8	O2	O3	H6	C22	Pha Atoms
±0.081	±0.043	±0.096	±0.117	±0.099	±0.079	±1.588	C3
	±0.020	±0.018	±0.015	±0.028	±0.035	±1.426	O1
		±0.045	±0.076	±0.020	±0.006	±0.956	C8
			±0.017	±0.006	±0.007	±1.492	O2
				±0.007	±0.011	±1.552	O3
					±0.008	±1.909	H6
						±0.342	C22
c) Tolerance values for 42 compounds with low activity							
C3	O1	C8	O2	O3	H6	C22	Pha atoms
±0.091	±0.078	±0.346	±1.222	±0.761	±2.184	±3.896	C3
	±0.047	±0.048	±0.782	±0.984	±2.286	±3.203	O1
		±0.047	±0.071	±0.043	±0.042	±3.954	C8
			±0.013	±0.041	±0.659	±4.526	O2
				±0.037	±0.061	±4.256	O3
					±0.023	±5.312	H6
						±2.086	C22
d) Tolerance values for 1338 conformers of 86 compounds							
C3	O1	C8	O2	O3	H6	C22	Pha atoms
±0.101	±0.128	±0.419	±1.237	±0.928	±2.287	±3.902	C3
	±0.066	±0.084	±0.835	±1.035	±2.289	±3.203	O1
		±0.053	±0.836	±0.061	±0.057	±4.678	C8
			±0.830	±0.042	±0.677	±5.372	O2
				±0.039	±0.065	±4.762	O3
					±0.027	±5.705	H6
						±2.242	C22

The lsqnonlin function in the MATLAB[36] toolbox is used to obtain the κ_j values of the relevant model parameters, numerically solving the system of differential equations for the best subset of variables. GA steps are considered with the training set. The process is initiated with the random generation of the first population containing a certain number of parents. Model parameters (kappa indices κj) are encoded as integers with chromosomes. The lsqnonlin function is used to calculate the κj values of the model parameters. In this study, we studied the GA with the following parameters: number of production, 300; population size, 300; number of repetitions, 100; cross-section, 85%; mutation rate, 1.5%.

Since the optimal number of parameters for the model is not known initially, a number of calculations were made to examine the relationship between the selected parameters and the model's estimated power (q^2). Against the number of parameters r^2 (for the training and test set), q^2, q^2_{ext1}, q^2_{ext2}, q^2_{ext3}, and concordance correlation coefficient (CCC) graphs were plotted, as shown in Figure 2.

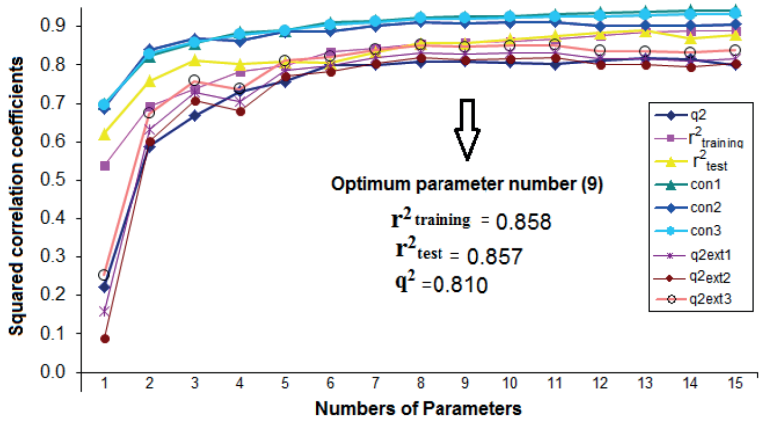


Figure 2. The graph of the optimal number of descriptors and $r^2_{training}$, r^2_{test}, q^2, q^2_{ext1}, q^2_{ext2}, q^2_{ext3}, con1, con2, and con3.

Figure 2 shows that as the number of parameters increases, the r^2 and q^2 values increase to 9 descriptors. After 9 descriptors, the model reaches stability and there is no increase in model performance with a higher number of identifiers. Generally, the ratio of the number of model compounds to the number of parameters should not exceed 1:5, which is called the potential overfitting risk.[37] In fact, as can be seen in Figure 2, although the graph shows that it is within acceptable limits for 5 parameters, 9 parameters are determined as optimum points according to the 5:1 rule.

When all conformers are included in the account, the r^2 value of 0.858 with 0.052 standard error value for the training set with the 9 most suitable parameters is satisfactory. Similarly, the r^2 and standard error values obtained for the test set are 0.857 and 0.074, respectively.

The optimum 9 molecular identifiers that were chosen with the GA and κj values used in biological activity calculation for alkynylphenoxyacetic acid derivatives are displayed in Table 3. When the descriptors in Table 3 are examined, it is shown that geometric parameters are most effective in terms of molecular elasticity on the activity of the reference molecule, as shown in Figure 3. The van der Waals (vdW) atomic radius is one of the important criteria used for determining interactions between ligand and bioreceptor. Distance parameters, which are defined by molecular conformation, have proven to be useful in the development of QSAR models and in predicting important properties of active conformation. $a^{(2)}$, $a^{(4)}$, and $a^{(5)}$ parameters are related to the distance + vdW atomic radius for some atoms (C22, C18, O2, C6). The $a^{(1)}$ and $a^{(2)}$ parameters are related

to the C22-C18 distance and the van der Waals radius. The C18 atom designates the most remote atom bound to the C22 atom in the R3 position, other than the hydrogen atoms bonded to the carbon atoms. The $a^{(3)}$, $a^{(4)}$, and $a^{(5)}$ parameters, which contain pharmacophore atoms, are the orthogonal distance from the C22 atom to the C1-O1-O3 plane, the orthogonal distance from the O2 atom to the C1-H6-C12 plane (Å) + vdW radius of the O2 atom, and the orthogonal distance from the O2 atom to the C1-H6-C12 plane (Å) + vdW radius of the O2 atom, respectively.

Table 3. κ_j values and the optimal nine molecular parameters used to calculate activity for alkynylphenoxyacetic acid derivatives.

$a_{ni}^{(j)}$	Molecular parameters	κ_j value
$a^{(1)}$	C22-C18 distance (Å)	0.0216
$a^{(2)}$	C22-C18 distance (Å)+ vdW radius	0.0564
$a^{(3)}$	Orthogonal distance from the C22 atom to the C1-O1-O3 plane (Å)	0.0243
$a^{(4)}$	Orthogonal distance from the O2 atom to the C1-H6-C12 plane (Å) + vdW radius of O2 atom	–0.0139
$a^{(5)}$	Orthogonal distance from the C6 atom to the C1-O1-C12 plane (Å) + vdW radius of C6 atom	–0.0713
$a^{(6)}$	The angle of between Cl1-O3-C18 atoms	0.0108
$a^{(7)}$	The angle between the O3-H6-C12 plane and the O2-O4 lines	0.2155
$a^{(8)}$	The angle between the C6-O3-H6 plane and the Cl1-C18 lines	0.1040
$a^{(9)}$	The dihedral angle between the O2-C5-C22-C18 plane	–0.0109

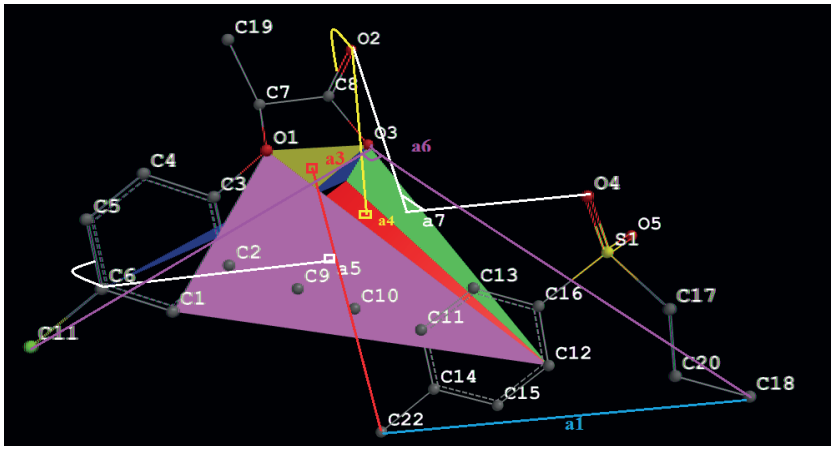


Figure 3. Plot of $a^{(1)}$, $a^{(3)}$, $a^{(4)}$, $a^{(5)}$, $a^{(6)}$, and $a^{(7)}$ parameters. $a^{(1)}$ is the C22-C18 distance (in blue), $a^{(3)}$ is the orthogonal distance from the C22 atom to the C1-O1-O3 plane (in violet red), and $a^{(4)}$ is the orthogonal distance from the O2 atom to the C1-H6-C12 plane (Å) and the vdW radius of the O2 atom (in yellow). Similarly, $a^{(5)}$ (in white) indicates the orthogonal distance from the C6 atom to the C1-O1-C12 plane (Å) and vdW radius of the C6 atom. $a^{(6)}$ is the angle between the Cl1-O3-C18 atoms (radian) (in pink) and $a^{(7)}$ is the angle between the O3-H6-C12 plane and the O2-O4 lines (radian) (in white).

The parameters $a^{(6)}$, $a^{(7)}$, $a^{(8)}$, and $a^{(9)}$ are the angular identifiers and are given in radians. The hydrogen atoms bound to the carbon atoms are not shown in Figure 3 in order to give a clearer representation.

Efficient methods for predicting biological activity should consider the entire range of probable molecular conformations and determine the activity of each conformer. All the conformers interact with a bioreceptor independently.

The activity is exponentially related to the S function ($A \sim e^{-S}$).[38] The positive sign, the product of kappa values and parameters, indicates the antipharmacophore shielding group (APS); negative values indicate the auxiliary group (AG). The parameters $a^{(1)}$, $a^{(2)}$, $a^{(3)}$, $a^{(6)}$, $a^{(7)}$, and $a^{(8)}$ are APS groups and $a^{(4)}$ and $a^{(5)}$ are AG groups. $a^{(9)}$ belongs to both AG and APS groups because it has positive and negative values in different conformers of the same compound.

In this study, to determine the best validity model, to measure the stability of the model, and to test the validity of the results, internal and external validation procedures were performed by the leave-one-out cross-validation (LOO-CV) technique. In addition, the results of external verification and the CCC also give information about the performance of the model. The E-statistic procedure was performed to analyze the single effect of each parameter on the activity values. When an identifier is subtracted, too much reduction in performance indicates that the identifier has a large effect on the model. The E-statistical results show the effect of 9 molecular parameters on model performance. As a result, ignoring the concerned descriptor, the difference in the model performance was investigated on the E, $r^2_{training}$, $se_{training}$, r^2_{test}, se_{test}, q^2, q^2_{ext1}, q^2_{ext2}, q^2_{ext3}, con1, con2, and con3 values of the model developed by the EC-GA method for the alkynylphenoxyacetic acid series that are shown in Table 4.

Table 4. E-statistical results for alkynylphenoxyacetic acid derivatives show how $r^2_{training}$, $se_{training}$, r^2_{test}, se_{test}, q^2, q^2_{ext1}, q^2_{ext2}, q^2_{ext3}, con1, con2, and con3 values are affected by each descriptor.

Parameter	E	r^2_{tr}	se_{tr}	r^2_{test}	se_{test}	q^2	q^2_{ext1}	q^2_{ext2}	q^2_{ext3}	con1	con2	con3
$a^{(j)}_{ni}$												
a(1)	0.855	0.836	0.053	0.855	0.079	0.778	0.811	0.796	0.832	0.914	0.904	0.912
a(2)	0.334	0.627	0.079	0.680	0.118	0.432	0.490	0.449	0.547	0.790	0.772	0.785
a(3)	0.840	0.838	0.052	0.847	0.081	0.774	0.835	0.822	0.853	0.915	0.914	0.916
a(4)	0.990	0.858	0.049	0.862	0.077	0.809	0.843	0.830	0.861	0.926	0.917	0.924
a(5)	0.647	0.786	0.06	0.806	0.091	0.707	0.794	0.777	0.817	0.885	0.884	0.885
a(6)	0.965	0.851	0.050	0.830	0.085	0.804	0.820	0.806	0.840	0.920	0.900	0.916
a(7)	0.444	0.722	0.068	0.672	0.174	0.573	0.555	0.519	0.605	0.849	0.776	0.830
a(8)	0.779	0.828	0.054	0.789	0.095	0.757	0.728	0.706	0.759	0.909	0.866	0.898
a(9)	0.995	0.853	0.050	0.848	0.081	0.809	0.829	0.816	0.848	0.922	0.907	0.919

When Table 4 is examined, the parameter with the most effect on the activity is $a^{(2)}$, which has the smallest E value (0.334) and therefore the lowest q^2 (0.432) value. In addition, the lowest $r^2_{training}$ values in the table belong to parameter $a^{(2)}$, indicating a decrease in model performance and statistical values. Therefore, the $a^{(2)}$ parameter, which represents the C22-C18 distance (Å) + vdW radius, is the most significant parameter for this model.

The $a^{(9)}$ parameter, which corresponds to the dihedral angle between the O2-C5-C22-C18 plane (radian), with the highest E value (0.995) has the least contribution to model performance. If q^2 is neglected when $a^{(9)}$ is neglected, this parameter means that the effect on the model and activity can be neglected without causing

a loss in model performance.

The second parameter with the highest modulus contribution is $a^{(7)}$ (the angle between the O3-H6-C12 plane and the O2-O4 lines (radian)) with the E value of 0.444 and q^2 value of 0.573. As a result, the optimal 9 parameters revealed by the developed EC-GA model showed how important the descriptors were in the biological activity of the alkynylphenoxyacetic acid derivatives.

The calculations with biological activity and statistical analysis were first examined in all conformers and secondly on each of the compounds only on the basis of the lowest energy conformer. The values of q^2, $r^2_{training}$, r^2_{test}, q^2_{ext1}, q^2_{ext2}, q^2_{ext3}, con1, con2, and con3 for all the conformers of each compound were calculated as 0.8105, 0.8580, 0.8571, 0.8282, 0.8145, 0.8475, 0.9258, 0.9092, and 0.9219, respectively. The values of q^2, $r^2_{training}$, r^2_{test}, q^2_{ext1}, q^2_{ext2}, q^2_{ext3}, con1, con2, and con3 for the single conformer were calculated as 0.7261, 0.8167, 0.8082, 0.7725, 0.7545, 0.7982, 0.9032, 0.8891, and 0.9003, respectively. It was shown by the established models and the developed EC-GA method that the results obtained when using multiple conformers in the study are much better than when using a single conformer. In the EC-GA model developed in this study, we obtained better statistical results by considering all the conformers of all compounds.

Accordingly, we found that the results obtained when using all conformers of the compounds were higher than when using single conformers. The statistical analysis results obtained with the optimum 9 parameters for single and all conformers values are shown in Figure 4.

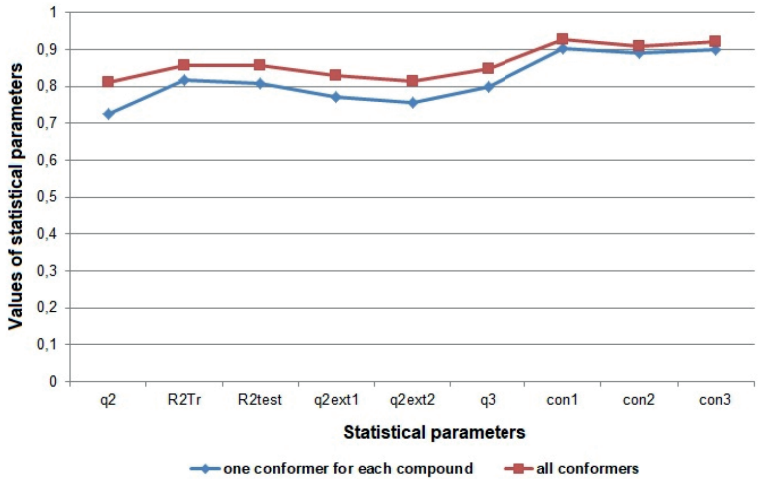


Figure 4. Comparison of statistical results of biological activity values for alkynylphenoxyacetic acid derivatives for the lowest energy conformer and all conformers of each compound with the optimum 9 parameters.

2.1. Conclusions

In this study, we developed a 4D-QSAR model with different combinations in the MATLAB program using the EC-GA method for alkynylphenoxyacetic acids as CRTh2 receptor opponent agents. The model results obtained by the EC-GA method show that the validity coefficient of the model is high. Since there is no QSAR study in the literature related to the working compound series, no comparison can be made. In this 4D-QSAR study using the EC-GA method within the scope, the EC method and genetic algorithm optimization technique were used together and ligand-based computer-aided drug design was done at the same time. The EC-GA method consists of a combination of the genetic algorithm and EC methods. The EC-GA method and EMRE program

developed by our group have been successful in previous 4D-QSAR studies. The EC-GA method has been described as an effective method to apply to complex nonlinear relationships between conformers of compounds and biological activities using the genetic algorithm and activity formula of Eq. (3). The following results were obtained from this study:

1. The remarkable point in this study is to consider both the pharmacophore definition and the conformational flexibility of all compounds depending on the Boltzmann distribution in predicting bioactivity. For this purpose the EMRE software package, which was developed in a comprehensive way to be used for pharmacophore identification, molecular parameter, and activity calculation, was successfully applied to alkynylphenoxyacetic acids. The designated pharmacophore group atoms are C3, O1, C8, O2, O3, H6, and C22.
2. Determining sufficient identifiers in nonlinear QSAR studies is difficult because there are no absolute rules for governing this choice. For each compound, 522 identifiers were prepared and calculated with EMRE software, and then selected with the best set of genetic algorithms for the calculated identifiers. The constructed EC-GA model shows coherent descriptor contributions to activity with 61 compounds for the training set and 25 compounds for the test set. The regression coefficients for the training and test sets are 0.8580 and 0.8571; the cross-validation coefficient, q^2, is 0.8105; and the q^2_{ext1}, q^2_{ext2}, and q^2_{ext3} values are 0.8282, 0.8145, and 0.8475, respectively.

As a result, with the obtained EC-GA model, the goodness of fit between the experimental and calculated events is over 0.8000 for the dataset. The internal and external validations emphasize the robustness of the model for the training and test sets, the stability of selected identifiers, and therefore the stability of the estimates. The results of this study could be useful in the planning of more potent CRTh2 receptor antagonists.

3. Experimental

3.1. Compounds

In this paper we performed 4D-QSAR studies of 86 compounds for CRTh2 receptor antagonists. In this sense, determination of the pharmacophore and estimation of biological activity were obtained by the EC-GA method that was developed by our group.

The structures of alkynylphenoxyacetic acid derivatives and the K_i (nM) values were obtained from the literature.[35] The K_i (nM) values were converted into molar units and then to a negative logarithmic scale (pK_i). These pK_i values are shown in Table 1 as experimental activities (A^{exp}).

Construction of the 3D structures of compounds was carried out using Spartan 10 software,[39] which has been employed for compounds for conformational analysis and quantum chemical calculations at the HF/6-31G* level, in water medium. Even though the more complicated basis sets give more accurate results, they expend a great deal of computation time. In the case of a large number of compounds and conformations, the required computation time increases due to much larger basis sets. Accordingly, we have considered the basis set HF/6-31G*, which is faster and sufficiently small without compromising the required level of accuracy.

3.2. Computational details and QSAR analysis

Quantitative structure-activity relationship modeling generally involves three modules: first, the gathering or, if possible, designing of a training set of chemicals; second, selecting descriptors that can properly relate chemical

structure to biological activity; and third, performing statistical methods that correlate changes in structure with changes in biological activity.[40] The EC-GA method involves a systematic procedure with six stages, which are as follows: the first stage includes conformational analysis and quantum chemistry calculations of the compounds in bioactivity prediction; the second stage is the formation of the ECMCs for each conformer of the compounds; the third stage of the EC-GA method is the identification of the pharmacophore and determination of the ECSAs using the ECMCs by means of EMRE software; the fourth stage is the determination and preparation of molecular descriptors; the fifth stage consists of an application of a statistical approach for QSAR model development; and the last stage is model validation. In this step, the LOO-CV procedure is implemented for both the robustness and predictive ability of the developed QSAR model.

In the EC-GA method, following the conformational analysis for each molecule classified according to Boltzmann distribution, the Boltzmann distribution of conformers under 1/10,000 are eliminated and the residue conformers are sorted according to the lowest energy level and registered.

All the electronic features of each of the conformations were characterized with the ECMC, which includes electronic and geometric parameters. The ECMC is a square matrix that is symmetrical according to the diagonal elements and also nondiagonal; the diagonal elements are components of this matrix. In this manner, the Mulliken charges of atoms are the diagonal components of the matrix that define the electronic properties of the atoms; the nondiagonal components are of two kinds. In the first case, when the atoms are chemically bonded to each other, the element indicates the degree of bonding. In the second case, for atoms that do not bond, this element gives the distance between atoms. Thus, it is assumed that all conformers are represented electronically and geometrically.[41,42]

In the ECMC, each matrix is made up of small subunits that represent the atomic properties of atoms or the properties between atoms. The score of lines in the matrices corresponds to the number of atoms in the molecule (n). Assuming that the matrices are formed from vertical strata, each layer corresponds to a different conformation of the compound. For this reason, the conformer number of each compound is the same as the number of layers and each layer will have the properties of molecular atoms. In this way, a similar matrix is formed for each compound in the dataset.

The diagonal matrix elements (charges) of the reference compound (compound 85) and other compounds are compared at the specified tolerance interval. The load of each atom in the reference molecule is compared to that of each atom of the other molecules. Thus, all submatrices of $n \times n$ of both matrices must be compared. The difference between the values of the diagonal elements of the two matrices is the matrix element that does not exceed the tolerance value, and those exceeding the tolerance value are ignored.

As a result, we designed an algorithm that can prevent unnecessary comparisons by using the EMRE program, so that the size of the matrix is reduced by a predetermined increase from the minimum tolerance value to the maximum tolerance value to reveal common matrix elements in the matrix.

The concept of pharmacophore is often expressed as a group of properties that include combinations of the chemical, structural, and physical properties of the compound, which must provide a precise biological activity.

In the final part of comparison process we gain a lot of submatrices, which are called ECSAs, and these represent the pharmacophore. The limits of tolerance values change step by step. As a result of the comparison, the expressions of P_α and α_a were used from the literature to determine the best of the possible submatrices that represent the likelihood of the pharmacophore.[43] P_α and α_a are shown as follows:

$$P_\alpha = (x1 + 1)/(x1 + x3 + 2), \tag{1}$$

$$\alpha_a = (x1 \times x4 - x2 \times x3)/(y1 \times y2 \times y3 \times y4)1/2. \quad (2)$$

x1 and x2 are the numbers of high-activity compounds with and without pharmacophores. x3 and x4 are the numbers of low-activity compounds with and without pharmacophores. y1 and y2 are the numbers of high-activity and low-activity compounds, respectively.

In this study, for 86 alkynylphenoxyacetic acid derivatives, 1338 ECMCs were appraised to superpose with the EMRE software. After forming the ECMCs, the studied molecules were categorized as high-activity and low-activity compounds. To find the pharmacophore, compound 85, which had the highest activity value and was considered as the reference, was contrasted with other ECMCs in which the lowest energy conformers of the compounds were compatible with the determined tolerances.

First, the lowest energy conformers of the compounds are taken into consideration in the comparison process. If the lowest energy conformers of high-activity compounds do not contain ECSAs, the higher energy conformers are examined for ECSAs. Then, for the best ECSA, the highest tolerance values for all compounds and all their conformers are determined without any tolerance constraints. At the end of this process, the pharmacophore found for alkynylphenoxyacetic acid derivatives by the EC-GA method stands for the 3D arrangement of the constructional and chemical compounds and is shown in the results and discussion section (Table 2).

Although the pharmacophore is one of the most important conditions for activity, it is not enough by itself. It cannot express why different molecules with identical pharmacophores exhibit different activities. The pharmacophore contains a group of atoms necessary for the activity, but other groups (Out-of-Pha groups, OOP) that act on the activity may also be present. APS and AG groups can be defined as atomic groups acting on the activity other than the pharmacophore. APS groups act in a way that reduces activity by steric hindrance or screening while interacting with the bioreceptor, while AG groups contribute to the activity. AG and APS groups are separated by positive or negative effects on biological efficiency. Subsequently, in this work we investigated which groups increased or decreased the biological activity, apart from the pharmacophore.

Conformational analysis is a problem that is often seen to be very complicated in computer-aided drug design, and drug molecules with mostly organic structures have a large number of conformers. In QSAR studies, either the lowest energy conformer is commonly used or the active conformer is used. For this reason, the general formula of activity is solved by adding the Boltzmann distribution of all conformers and the Boltzmann distribution of each conformer as a function of temperature and energy. The general formula of biological activity was developed by Bersuker et al., in the following equation:[43]

$$A_n = A_l \frac{\sum_{i=1}^{m_l} e^{-E_{li}/kT}}{\sum_{i=1}^{m_n} e^{-E_{ni}/kT}} \frac{\sum_{i=1}^{m_n} \delta_{ni}[Pha] e^{-S_{ni}} e^{-E_{ni}/kT}}{\sum_{i=1}^{m_l} \delta_{li}[Pha] e^{-S_{li}} e^{-E_{li}/kT}}, \quad (3)$$

where A_n activity is for the nth compound and the A_l activity value corresponds to the reference compound. δ is a kind of Dirac δ function, which depends on the pharmacophore present. This function takes 1 if the pharmacophore is present and 0 if not.

E_{li} and E_{ni} are the relative energies of the ith conformation of the reference compound in kcal mol^{-1} and the ith conformation of the nth compound (in kcal mol^{-1}), respectively. R(kcal mol^{-1} K^{-1}) is the gas constant and T is the temperature in K.

Activity is exponentially related to the S function.[38] The numerical values of the parameters and the values of κj are taken into consideration in determining the AG and APS groups. If the product of the numerical value of the relevant parameter and the value of κj is positive, this parameter indicates the activity increasing AG; if negative, it indicates the APS group.

The APS and AG groups can be defined as atomic or atomic groups acting on activity other than the pharmacophore. APS groups act on reducing activity by steric hindrance or screening while interacting with the bioreceptor, while AG groups contribute to the activity. The effect of the AG and the APS groups on biological activity is defined with the equation of the S function:

$$S_{n_i} = \sum_{j=1}^{N} \kappa_j a_{ni}^{(j)}, \tag{4}$$

where N is the number of chosen AG parameters; the κ_j constant is a variable coefficient obtained as a result of the analysis and is different for each parameter. $a_{ni}^{(j)}$ is a parameter representing the jth type of property in the jth conformer of the nth compound.

In the EC-GA method, it is possible to calculate the κ_j variable constants with the help of this equation. To obtain the κ_j values of each parameter, the lsqnonlin function in MATLAB is used. For the κ_j values corresponding to each parameter, the previously unresolved nonlinear exponential function, Eq. (3), is solved by means of the nonlinear least squares method (lsqnonlin) and the genetic algorithm optimization method in MATLAB. The numbers κ_j =1, 2,..., N procured hereby describe the weights of the $a_{ni}^{(j)}$ parameters in the total APS/AG effect.

In general, in QSAR studies there are a large number of identifiers, and since some of them are importantly related to activity, it is important to select the appropriate descriptors that best represent the structural changes and information to achieve a reliable QSAR model.[44] In the EC-GA method, a genetic algorithm optimization technique was used to determine the best descriptors affecting the biological activity of the drug molecule and to arrive at the conclusion in a short time. Also, it is possible to prepare thousands of parameters to define molecules and conformers, and to choose the parameters that are most suitable for both the training and test set.

In previous studies, different experiments were carried out by changing the κ_j adduct, population number, and κ_j limits and many different results were obtained. In this way, a model is created using each subset of parameters and their corresponding κ_j values. The model with the highest regression coefficient r^2 and lowest standard error (se) is the best.

Within the scope of the present report, 522 different molecular identifiers, which include topological and geometrical parameters considering the pharmacophore group, were constituted and these parameters were used for each conformer of 86 compounds during calculation in EMRE software.[27]

When the genetic algorithm is used for descriptor selection, it represents the set of parameters corresponding to a possible solution for the problem and selected from the parameter pool. The GA codes were written in MATLAB by the authors. The κ_j values were calculated by means of the lsqnonlin function using MATLAB software by solving the mathematical differential equations in order to obtain the best subset of descriptors.

Prediction capacity is the most important feature of a QSAR model.[45,46] After constructing a model with the genetic algorithm optimization technique, one of the compounds is tested with the LOO-CV method,

which is used to accurately approximate the validity of the model and is often applied to assure the statistical stability of the QSAR model.[47]

QSAR analysis aims to continue the progress of approved models to accurately and precisely predict the biological activities of compounds that may be potentially involved in drug discovery.[48] For this purpose, internal validation and external validation techniques were used for model validation. Internal validation is an ideal technique to assess the quality and fitness of the model and was accomplished by the leave-one-out (LOO) method. The cross-validated coefficient, q^2, is calculated according to the following formulas:

$$q^2_{ext1} = 1 - \frac{\sum_{n=1}^{N} \left| A^{\exp}_{n_{test}} - A^{pred}_{n_{test}} \right|^2}{\sum_{n=1}^{N} \left| A^{\exp}_{n_{test}} - \overline{A}^{\exp}_{n_{training}} \right|^2}, \tag{5}$$

$$q^2_{ext2} = 1 - \frac{\sum_{n=1}^{N} \left| A^{\exp}_{n_{test}} - A^{pred}_{n_{test}} \right|^2}{\sum_{n=1}^{N} \left| A^{\exp}_{n_{test}} - \overline{A}^{\exp}_{n_{test}} \right|^2}, \tag{6}$$

where N shows the total number of molecules in the training set. $\overline{A}^{\exp}_{n_{training}}$ and $\overline{A}^{\exp}_{n_{test}}$ are the averages of the experimental activity of the training and test sets. SSY is the sum of the squares of the deflections of the experimental activities from $\overline{A}^{\exp}_{n}$.

However, it should be emphasized that the actual performance of a model could only be revealed using an external verification set. For this reason, the performance of the model was further appraised by external validation.[49] The true predictive power must be estimated by comparing the predicted and observed activities of an external test set of compounds that were not used in the model development. A model that is externally predictive should also be robust. In this validation all of the molecules are separated into two parts, namely the training set and the test set. The training set is used to create a model, and this model is used to predict the biological activities of the test set compounds.[50]

The difference between the measured and estimated values reveals the model's performance. The external explained variance[51] is calculated as follows:

$$q^2_{ext3} = 1 - \frac{\frac{\left[\sum_{n=1}^{N_{test}} \left| A^{\exp}_{n_{test}} - A^{pred}_{n_{test}} \right|^2\right]}{N_{test}}}{\frac{\left[\sum_{n=1}^{N_{training}} \left| A^{\exp}_{n_{training}} - \overline{A}^{\exp}_{n_{training}} \right|^2\right]}{N_{training}}}, \tag{7}$$

where $N_{training}$ and N_{test} are the number of test and training molecules, respectively, and $A^{\exp}_{n\ test}$ and $A^{pred}_{n_{test}}$ are the experimental and predicted activities of the nth molecule in the test set while $A^{\exp}_{n\ training}$ is the experimental activity of the nth compound in the training set. $\overline{A}^{\exp}_{n\ training}$ represents the mean experimental activities of the molecules in the training set.

In addition to external validation, the CCC has been suggested as a different validation criterion and was applied to huge sets by Chirico and Gramatica.[52] The CCC measures precision and at the same time

correctness; for this reason, any deviation of the regression line from the incompatibility line gives a value of CCC less than 1. As a result, this coefficient could be used efficiently for the external validation of QSAR models.[52]

The CCC confirms the perfect sensitivity and accuracy of the model with a value greater than 0.85[53] and this parameter is calculated to check the model reliability with the following equations:

$$CCC = \underset{training}{\overset{\wedge}{\rho}} = \frac{2\sum_{l=1}^{n_{training}} \left(A_l^{pred} - \overline{A}^{pred}\right)\left(A_l^{\exp} - \overline{A}^{\exp}\right)}{\sum_{l=1}^{n_{training}} \left(A_l^{pred} - \overline{A}^{pred}\right)^2 + \sum_{l=1}^{n_{training}} \left(A_l^{\exp} - \overline{A}^{\exp}\right)^2 + n_{training}\left(A_l^{\exp} - \overline{A}^{\exp}\right)^2}, \tag{8}$$

$$CCC = \underset{test}{\overset{\wedge}{\rho}} = \frac{2\sum_{l=1}^{n_{test}} \left(A_l^{pred} - \overline{A}^{pred}\right)\left(A_l^{\exp} - \overline{A}^{\exp}\right)}{\sum_{l=1}^{n_{test}} \left(A_l^{pred} - \overline{A}^{pred}\right)^2 + \sum_{l=1}^{n_{test}} \left(A_l^{\exp} - \overline{A}^{\exp}\right)^2 + n_{test}\left(A_l^{pred} - \overline{A}^{\exp}\right)^2}, \tag{9}$$

$$CCC = \underset{all}{\overset{\wedge}{\rho}} = \frac{2\sum_{l=1}^{n_{all}} \left(A_l^{pred} - \overline{A}^{pred}\right)\left(A_l^{\exp} - \overline{A}^{\exp}\right)}{\sum_{l=1}^{n_{all}} \left(A_l^{pred} - \overline{A}^{pred}\right)^2 + \sum_{l=1}^{n_{all}} \left(A_l^{\exp} - \overline{A}^{\exp}\right)^2 + n_{all}\left(A_l^{pred} - \overline{A}^{\exp}\right)^2}, \tag{10}$$

where $A_l^{\exp}$ and A_l^{pred} refer to the experimental and predicted values of activity, respectively. In the same way, $\overline{A}^{\exp}$ and $\overline{A}^{pred}$ refer to the averages of the experimental and predicted activities. The use of the training and test sets in the formula increased the reliability of the model.

After determining the best model with the genetic algorithm and the best parameters forming this model, the E-statistic method is used to determine which of the parameters in the parameter pool is likely to contribute more to biological activity. The E-statistic formula is given as follows:

$$E = \frac{PRESS_N}{PRESS_{N-1}}. \tag{11}$$

The predictive residual sum of squares (PRESS) is used as the fitness function. Each set of subparameters in the population is sent to the lsqnonlin function to calculate the κ_j values.

The value of the fitness function for each subparameter set is calculated. The PRESS is described as follows:

$$PRESS_N = \sum_{n=1}^{N} \left|A_n^{\exp} - A_n^{pred}\right|^2, \tag{12}$$

where $A_n^{\exp}$ and A_n^{pred} define the experimental and the predicted activities of the nth compound in the LOO-CV process. In this case, N = 9 parameters were used to assess the most effective identifiers in the developed model.

The EC-GA method consists of two sections: the identification of the pharmacophore group after quantum chemical calculations and the estimation of bioactivity by parameterizing the groups/properties affecting the activity. It is possible to quantify activity as long as reliable experimental data are available with this method. The EC-GA is a 4D-QSAR technique that can be used to estimate the pharmacophore and biologic activity. Herein, we reported 4D-QSAR studies for alkynylphenoxyacetic acid analogues and we performed pharmacophore group characterization and biological activity prediction.

Acknowledgments

This work was supported by the Research Fund of Erciyes University under grant number FBD-12-4015. This work was partially supported by TÜBİTAK.

References

1. Monneret, G.; Cossette, C.; Gravel, S.; Rokach, J.; Powell, W. S. *J. Pharmacol. Exp. Ther.* **2003**, *304*, 349-355.
2. Zaghdane, H.; Boyd, M.; Colucci, J.; Simard, D.; Berthelette C.; Leblanc, Y.; Wanga, Z.; Houle, R.; Lévesque, J. F.; Molinaro, C. et al. *Bioorg. Med. Chem. Lett.* **2011**, *21*, 3471-3474.
3. Spik, I.; Brenuchon, C.; Angeli, V.; Staumont, D.; Fleury, S.; Capron, M.; Trottein, F.; Dombrowicz, D. *J. Immunol.* **2005**, *174*, 3703-3708.
4. Hart, P. H. *Immunol. Cell Biol.* **2001**, *79*, 149-153.
5. Murray, J. J.; Tonnel, A. B.; Brash, A. R.; Roberts, L. J.; Gosset, P.; Workman, R.; Capron A.; Oates, J. A. *N. Engl. J. Med.* **1986**, *315*, 800-804.
6. Hardy, C. C.; Robinson, C.; Tattersfield, A. E.; Holgate, S. T. *N. Engl. J. Med.* **1984**, *311*, 209-213.
7. Babu, S.; Madhavan, T. *J. Chosun Natural Sci.* **2015**, *8*, 1-2.
8. Sandham, D. A.; Aldcroft, C.; Baettig, U.; Barker, L.; Beer, D.; Bhalay, G.; Brown, Z.; Dubois, G.; Budd, D.; Bidlake, L. et al. *Bioorg. Med. Chem. Lett.* **2007**, *17*, 4347-4350.
9. Ulven, T.; Kostenis, E. *Expert Opin. Ther. Pat.* **2010**, *20*, 1505-1530.
10. Babu, S.; Sohn, H.; Madhavan, T. *Comput. Biol. Chem.* **2015**, *56*, 109-121.
11. Patel, H. M.; Noolvi, M. N.; Sharma, P.; Jaiswal, V.; Bansal, S.; Lohan, S.; Kumar S. S.; Abbot, V.; Dhiman, S.; Bhardwaj, V. *Med. Chem. Res.* **2014**, *23*, 4991-5007.
12. Hansch, C.; Hoekman, D.; Gao, H. *Chem. Rev.* **1996**, *96*, 1045-1075.
13. Devilliers, J. *Neural Networks in QSAR and Drug Design*; Academic Press: San Diego, CA, USA, 1996.
14. Borosy, A. P.; Keseru, K.; Matyus, P. *Chemom. Intell. Lab. Syst.* **2000**, *54*, 107-122.
15. Sohi, R. S.; Ghasemi, J. B. *Med. Chem. Res.* **2013**, *22*, 1587-1596.
16. Andrade, C. H.; Pasqualoto, K. F. M.; Ferreira, E. I.; Hopfinger, A. J. *Molecules* **2010**, *15*, 3281-3329.
17. Damale, M. G.; Harke, S. N.; Khan, F. A. K.; Shinde, D. B.; Sangshetti, J. N. *Mini-Rev. Med. Chem.* **2014**, *14*, 35-55.
18. Baert, B.; Deconinck, E.; Gele, M. V.; Slodicka, M.; Stoppie, P.; Bod'e, S.; Slegers, G.; Heyden, V. Y.; Lambert, J.; Beetens, J. et al. *Bioorg. Med. Chem.* **2007**, *15*, 6943-6955.
19. Doucet, J. P.; Panaye, A. *QSAR in Environmental & Health Sciences*; CRC Press: New York, NY, USA, 2009.
20. Seel, M.; Turner, D. B.; Wilett, P. *QSAR* **1999**, *18*, 245-252.
21. Santos-Filhoa, O. A.; Hopfinger, A. J. *Quant. Struct.-Act. Relat.* **2002**, *21*, 369-381.
22. Albuquerque, M.; Brito, M.; Cunha, E.; Alencastro, R.; Antunes, O.; Castro, H.; Rodrigues, C. *Curr. Methods Med. Chem. Biol. Phys.* **2007**, *1*, 91-100.
23. Albuquerque, M. G.; Hopfinger, A. J.; Barreiro, E. J.; de Alencastro, R. B. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 925-938.
24. Hunger, J.; Huttner, G. *J. Comput. Chem.* **2009**, *20*, 455-471.
25. Goodarzi, M.; Saeys, W.; Deeb, O.; Pieters S.; Heyden Y. V. *Chem. Biol. Drug Des.* **2013**, *82*, 685-696.
26. Hasegawa, K. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 112-120.

27. Sarıpınar, E.; Geçen, N.; Şahin, K.; Yanmaz, E. *Eur. J. Med. Chem.* **2010**, *45*, 4157-4168.
28. Şahin, K.; Sarıpınar, E.; Yanmaz, E.; Geçen, N. *SAR QSAR Env. Res.* **2011**, *22*, 1-22.
29. Yanmaz, E.; Sarıpınar, E.; Şahin, K.; Geçen, N.; Çopur, F. *Bioorganic Med. Chem.* **2011**, *19*, 2199-2210.
30. Geçen, N.; Sarıpınar, E.; Yanmaz, E.; Şahin, K. *J. Mol. Model.* **2012**, *18*, 65-82.
31. Akyüz, L.; Sarıpınar, E.; Kaya, E.; Yanmaz, E. *SAR QSAR Environ. Res.* **2012**, *23*, 409-433.
32. Akyüz, L.; Sarıpınar, E. *J. Enzyme Inhib. Med. Chem.* **2013**, *28*, 776-791.
33. Özalp, A.; Yavuz, S. Ç.; Sabancı, N.; Çopur, F.; Kökbudak, Z.; Sarıpınar, E. *SAR QSAR Environ. Res.* **2016**, *27*, 317-342.
34. Yavuz, S. Ç.; Sabancı, N.; Sarıpınar, E. *Curr. Comput.-Aid. Drug* **2018**, *13*, 3-18.
35. Crosignani, S.; Pretre, A.; Catherine, L. J.; Fraboulet, G.; Seenisamy, J.; Augustine, J. K.; Missotten, M.; Humbert, Y.; Cleva, C.; Abla, N. et al. *J. Med. Chem.* **2011**, *54*, 7299-7317.
36. MathWorks, Inc. *MATLAB Version 7.0*; MathWorks: Natick, MA, USA, 2004.
37. Topliss, J. G.; Edwards, R. P. *J. Med. Chem.* **1979**, *22*, 1238-1244.
38. Bersuker, I. B.; Bahceci, S.; Boggs, J. E.; Pearlman, R. S. *J. Comput. Aided Mol. Des.* **1999**, *13*, 419-434.
39. Wavefunction, Inc. *Spartan'10*; Wavefunction: Irvine, CA, USA, 2011.
40. Perkins, R.; Fang, H.; Tong, W.; Welsh, W. J. *Environ. Toxicol. Chem.* **2003**, *22*, 1666-1679.
41. Saripinar, E.; Guzel, Y.; Patat, S.; Yildirim, I.; Akcamur, Y.; Dimoglo, A. S. *Arzneimittelforschung.* **1996**, *46*, 824-828.
42. Dimoglo, A.S.; Shvets, N.M.; Tetko, I.V.; Livingstone, D. J. *Quant. Struct. Act. Relat.* **2001**, *20*, 31-45.
43. Bersuker, I. B. *Curr. Pharm. Des.* **2003**, *9*, 1575-1606.
44. Myint, K. Z.; Xie, X. Q. *Int. J. Mol. Sci.* **2010**, *11*, 3846-3866.
45. Consonni, V.; Ballabio, D.; Todeschini, R. *J. Chemometr.* **2010**, *24*, 194-201.
46. Golbraikh, A.; Tropsha, A. *J. Comput.-Aid. Mol. Des.* **2002**, *16*, 357-369.
47. Golbraikh, A.; Wang, X. S.; Zhu, H.; Tropsha, A. In: Leszczynski, J., Ed. *Handbook of Computational Chemistry*; Springer: Berlin, Germany, 2017, pp. 2303-2340.
48. Sun, M.; Zheng, Y.; Wei, H. Chen, J.; Ji, M. *J. Enzym. Inhib. Med. Ch.* **2009**, *24*, 1109-1116.
49. Kovatcheva, A.; Golbraikh, A.; Oloff, S.; Xiao, V.; Zheng, W.; Wolschann, P.; Buchbauer, G.; Tropsha, A. *J. Chem. Inf. Compu. Sci.* **2004**, *44*, 582-595.
50. Gramatica, P. *QSAR Comb. Sci.* **2007**, *26*, 694-701.
51. Haas, E. M. D.; Eikelboom, T.; Bouwman, T. *SAR QSAR Environ. Res.* **2011**, *22*, 545-559.
52. Chirico, N.; Gramatica, P. *J. Chem. Inf. Model.* **2011**, *51*, 2320-2335.
53. Larue, R. T. H. M.; Voorde, L. V. D.; Timmeren, J. E. V.; Leijenaar, R. T. H.; Berbée, M.; Sosef, M. N.; Schreurs, W. M. J.; Elmpt, W. V.; Lambin, P. *Radiother. Oncol.* **2017**, *125*, 147-153.