Quantitative structure activity relationship studies on some N-benzylacetamide and 3-(phenylamino)propanamide derivatives with anticonvulsant properties

Adedirin Oluwaseye[1], Adamu Uzairu[2], Shallangwa Gideon Adamu[3], Abechi Stephen Eyije[4]

[1]Chemistry Advance Research Center, Sheda Science and Technology Complex,
Abuja, Nigeria,

[2,3, 4]Chemistry Department, Ahmadu Bello University, Zaria, Nigeria.

Abstract: ***The activity of a number of N-benzylacetamide and 3-(phenylamino)propanamide analogues with anticonvulsant properties was described using the quantitative structure activity relationship model by applying it to 80 compounds. The molecular descriptors of the compounds were obtained by quantum chemical calculations combined with molecular modeling calculations. The resulting model has correlation coefficient R of 0.92 meaning it explains up to 92% of the variance in the activity of the compounds that made up the data set. The model was successfully validated internally by leave one out cross validation and y-scrambling test. External validation was done using the Golbraikh and Tropsha criteria and predicted square correlation coefficient for the test set R2pred. Statistical analysis shows that the anticonvulsant activity of the studied compounds depends mainly on the Kier2, RDF50s, AATS4i and VE2_D descriptors.***

Keywords: Epilepsy, QSAR, GFA, Molecular descriptors**.**

1. Introduction

Epilepsy is one of the most common and serious neurological disorder and it is characterized by recurrent aberrant synchronized discharge of a neuronal population termed ‘seizures’[1] which results from a temporary electrical disturbance of the brain due to an imbalance between excitatory and inhibitory neurotransmitters [2]. Seizures result from phasic changes in the firing properties of groups of neurons, usually within a descrete focal point, to an intermittent high-frequency burst-firing mode resulting in synchronized disruptive ‘electrical’ discharges owing to changes in neuronal excitability [1]. Epilepsy is affecting about 50 million individuals worldwide of which about 10.5 million are children [3] and almost 90% of these people are in the developing countries [4]. Untreated epilepsy can lead to impaired intellectual function or death and it is usually accompanied with psychosocial prejudices and other psycho pathological consequences such as loss of self-esteem and poor quality of life [5].

Treatment of epilepsy with chemical agent i.e. pharmacotherapy is one of major approach of treatment for epilepsy and since the introduction of Phenobarbital as an effective anticonvulsant agent a number of antiepileptic drugs (AEDs) have been introduced into the market. The first generation AEDs are Phenytoin, carbamazepine, valproate, Phenobarbital, primidone, ethosuximide and benzo-diazepines. These are capable of preventing seizure generation irrespective of the underlying etiology [6]. Increase in the understanding of the biological causes of epilepsy over the last 30 years have allowed the development of the second generation AEDs including oxcarbazepin, Gabapentin, lamotrigine, felbamate, Tiagabin, Vigabatrin, Topiramate, Levetiracetam, and Zonisamide. These second generation of AEDS targets the cellular mechanism that is responsible for epileptic discharges [6]. Despite the development these AEDs over 30% of people with epilepsy do not have seizure control and others do so only at the expense of significant dose related toxicity and peculiar adverse effects that range in harshness from minimal brain impairment and megaloblastic anemia to death from aplastic anemia or hepatic failure [7].These limitations demand the need for the development of more effective and safer antiepileptic drugs.

The first step for the discovery and development of new molecular candidates with improved anticonvulsant activity and no/lesser neurotoxicity should be concerned with the application of methodologies and techniques that will take care of the time factor, reduce high cost of experimental runs and prevents the serendipitous synthesis in organic chemistry [8]. In this sense, computer-aided drug design methodologies have played an essential role for the discovery of compounds in medicinal chemistry [9,10], pharmaceutical design [11] and drug metabolism [12]. They have helped to improve the process of optimization of the molecular structure with defined purposes [13]. One of such techniques is quantitative-structure activity relationship (QSAR) analysis which have proved to be useful tool for predicting biological activities of compounds by utilizing experimental data and molecular structure [14]. Quantitative structure-activity relationship (QSAR) analysis is an area of computational research which builds models of biological activity using physicochemical properties of a series of compounds. The underlying assumption is that the variations of biological activity within a series of similar structures can be correlated with changes in measured or computed molecular features (molecular descriptors) of the molecules. These molecular descriptors could measure, for example, hydrophobic, steric, and electronic properties which may influence biological activity. The pre-requisite of developing QSAR equations is the availability of a wide range of molecular structures and their complementary activities and the field of AED research is very dynamic with numerous compounds of varying chemical classes reported to posses' anticonvulsant activity. One of such class of compounds is the carboxamides, which include N-benzylacetamides and 3-(phenylamino)propanamides. Reported here is the QSAR study on some new N-benzylacetamides and 3-(phenylamino)propanamides derivitatives which show anticonvulsant activity and relatively low neurotoxicity value compare to reference drug molecule using Genetic function approximation algorithm as the modeling tool. The best model obtained from the study was then used to design in silico new compounds with improved anticonvulsant activity values.

2. Materials and methods

1.1. Dataset

80 derivatives of N-benzylacetamide and 3-(phenylamino)propanamide were chosen from literatures [15,16] which are concerned only with the synthesis of the derivatives and their pharmacological test using similar assay (Maximum electroshock seizure test (MES) on albino mice). The anticonvulsant activity data were expressed as ED_{50} (mg/kg) and recalculated to molar unit for easy comparison between molecules [17], thereafter their Logarithmic values (-log ED_{50}) was calculated thus correlating the data linear to the free energy change [18] and presented in Table 1 with the structure of each molecules as the observed anticonvulsant activity for each molecule. ED_{50} is defined as a measure of the dose quantity that is effective in 50% of the tested animals and TD_{50} is a measure of the dose quantity that presents toxicity in 50% of the tested animals.

1.2. Calculation of molecular descriptors

In order to identify the effect of the molecular structure on the anticonvulsant activity of selected compounds, molecular descriptors which map the structure of compounds into a set of numerical values representing various molecular properties were calculated because only these numerical properties can correlate more directly with the activity [19]. The process started by constructing the 2D structure of each molecule and subsequent conversion to 3D using sketch and view tools in Spartan 14 package [20]. Molecular mechanics and PM3 quantum chemical procedure were used for optimization and energy minimization of the molecules until the root mean square (RMS) gradient value was smaller than 10^{-6} a.u. Furthermore, in order to obtain reliable energetic and accurate data on electronic properties of molecules the single-point energy calculations were performed at the DFT/B3LYP level of theory using the 6-31G** basis set. Some electronic descriptors were obtained from Spartan 14 and the optimized molecules were imported into PaDEL-descriptor [21] software for the calculation of other molecular descriptors.

The total number of calculated descriptors was 1865. Among the calculated descriptors those for which no value was available for all the compounds were disregarded. Also, descriptors of which the value is constant (or near-constant) were excluded. The remaining descriptors were scaled by standardization (auto-scaling) using the equation below:

$$X^{1} = \frac{(X_i - \bar{X}_i)}{\sigma_i}$$

Where 'x^{1}'is the scaled descriptor, '$\bar{x}_i$' is the mean for each column of descriptors 'x' and $\bar{\sigma}_i$ is the standard deviation of each column of descriptor. This gives all the variables in the data set equal importance in the model, thus removing the dependence of the regression coefficient on unit of the descriptors [22].

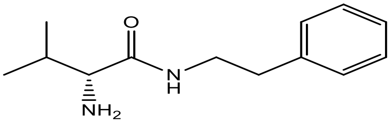
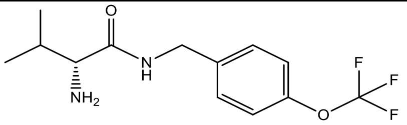
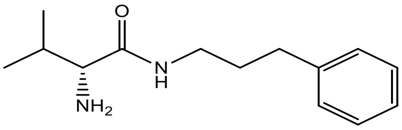
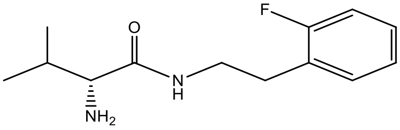
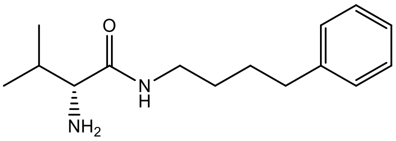
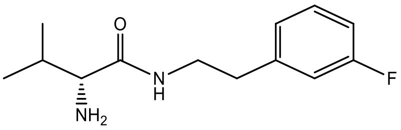
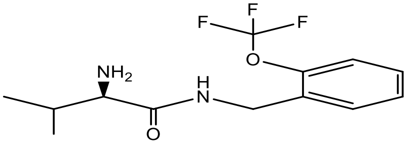
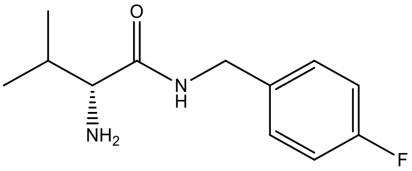
1.3. Data division

Euclidean distance based clustering method available in Datadivision 1.2 software [23] was used to divide the data into training (modeling) set and test (evaluation) set which constitute about 30 percent of the entire data [24]. The training set are used to construct the model and the test set (data set that are not included in constructing the model but cover the whole range of the training set) are used to check the ability of the model to predict external data set [25]. The data division method considers variability in both x and y dimensions to split the data into training set and test set. It start by calculating the Euclidean distance ED_x in the independent variable X-space and ED_y in the dependent variable y-space separately for each pair (p,q) of samples in the data set. The $ED_x(p,q)$ and $ED_y(p,q)$ are divided by their maximum values in the data set in order to assign an equal importance to the distributions of samples in the x and y spaces. The results are then added together to give a normalized x-y distance $ED_{XY}(p,q)$. Then, the selection start by taking the pair for which the $ED_{XY}(p,q)$ distance is largest and with subsequent iteration, the algorithm selects a sample that exhibits the least distance with respect to any sample already selected. The procedure is repeated until the number of sample required is achieved, thus resulting in two sets of sample from the original data set where one is reported as the training set and the other as the test set [26]. Consequently the initial data set in this work was splitted into two subsets: a training subset (N_{train} = 50) and a external validation subset (N_{test} = 30).

1.4. Selection and mapping of descriptor to Activity

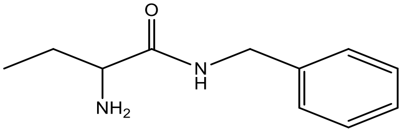
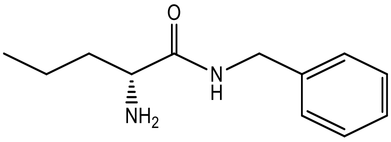
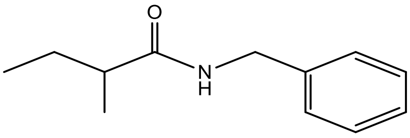
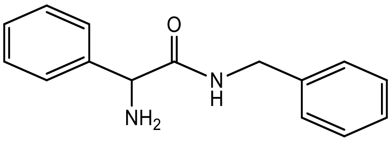
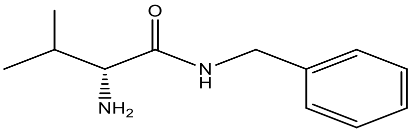
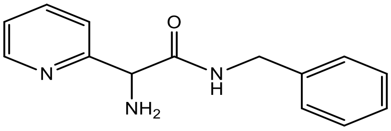
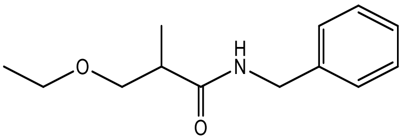
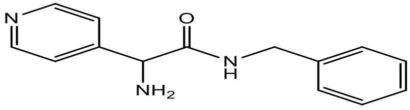
Selection of important descriptors that best explain the anticonvulsant activity of the selected compounds in this work was done from training (modeling) set only [27] using Genetic function approximation (GFA) method [28] available in Material studio 7.0 software. GFA is a combination of Holland's Genetic Algorithm (GA) and Friedman's multivariate adaptive regression splines (MARS) algorithm. It uses a genetic algorithm to perform a search over the space of possible QSAR models and uses certain fitness function (LOF) score obtained via multivariate adaptive regression splines algorithm to estimate the fitness of each model. The GFA algorithm approach has a number of important advantages over other techniques: it builds multiple models rather than a single model; it automatically selects which features are to be used in its basis functions and determines the appropriate number of

Table 1: Selected N-benzylacetamide and 3-(phenylamino)propanamide derivatives molecular structure, anticonvulsant and neurotoxicity values

S/N	Molecular structure	pED$_{50}$	pTD$_{50}$	S/N	Molecular structure	pED$_{50}$	pTD$_{50}$
1	 (R)-2-amino-3-methyl-N-phenethylbutanamide	4.343	3.644	5	 (R)-2-amino-3methyl-N-(4-trifluoromethoxy)benzyl) butanamide	4.271[a]	3.539[b]
2	 (R)-2-amino-3methyl-N-(3-phenylpropyl)butanamide	4.166[a]	3.736	6	 (R)-2-amino-N-(2-fluorophenethyl)-3-methyl butanamide	3.846	3.480
3	 (R)-2-amino-3methyl-N-(4-phenylbutyl)butanamide	4.354	3.815	7	 (R)-2-amino-N-(3-fluorophenethyl)-3-methyl butanamide	3.946[a]	3.551[b]
4	 (R)-2-amino-3methyl-N-(2-trifluoromethoxy) benzyl)butanamide	4.499	3.755	8	 (R)-2-amino-N-(4-fluorobenzyl)-3-methyl butanamide	4.029	3.360

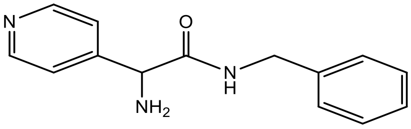
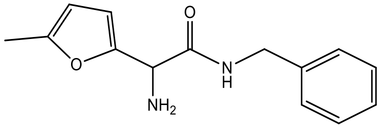
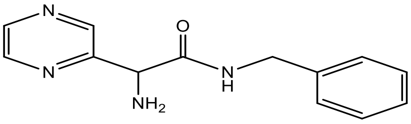
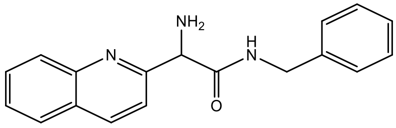
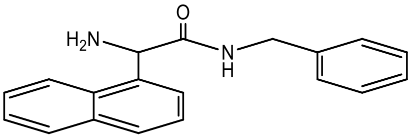
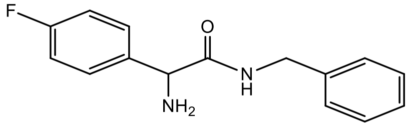
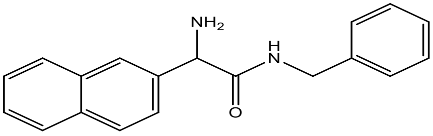
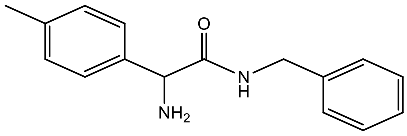
‘a’ represent the anticonvulsant activity values for data used as test set for anticonvulsant models, ‘b’ represent the neurotoxicity values for data used as test set for neurotoxicity models, pED$_{50}$ represent negative logarithm value of the dose quantity that is effective in 50% of the tested animals and pTD$_{50}$ represent negative logarithm value of the dose quantity that present toxicity in 50% of the tested animals

Table 1continued: Selected N-benzylacetamide and 3-(phenylamino)propanamide derivatives molecular structure, anticonvulsant and neurotoxicity values

S/N	Molecular structure	pED$_{50}$	pTD$_{50}$	S/N	Molecular structure	pED$_{50}$	pTD$_{50}$
9	 2-amino-N-benzylbutanamide	4.115	3.469[b]	13	 (R)-2-amino-N-benzylpentanamide	3.383	3.619[b]
10	 N-benzyl-2-methylbutanamide	4.076[a]	3.064[b]	14	 2-amino-N-benzyl-2-phenylacetamide	4.138[a]	3.593
11	 2-amino-N-benzyl-3-methylbutanamide	3.534[a]	3.469	15	 2-amino-N-benzyl-2-(pyridin-2-yl)acetamide	3.624	3.127[b]
12	 N-benzyl-3-ethoxyl-2-methylpropanamide	3.992	3.453	16	 2-amino-N-benzyl-2-(pyridin-4-yl)acetamide	3.903[a]	3.127[b]

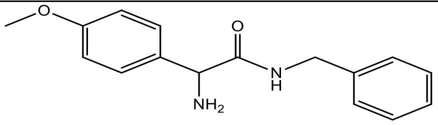
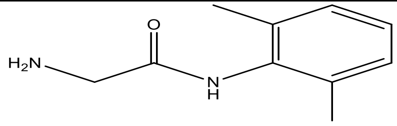
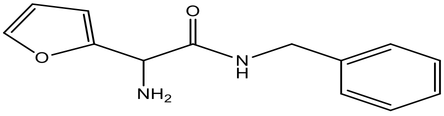
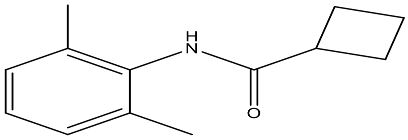
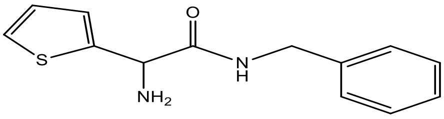
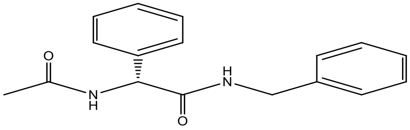
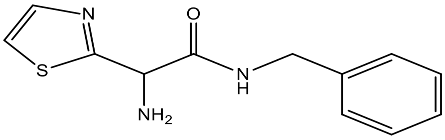
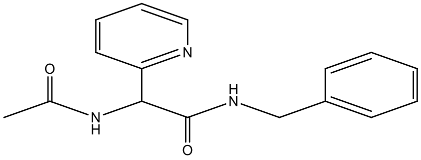
‘a’ represent the anticonvulsant activity values for data used as test set for anticonvulsant models, ‘b’ represent the neurotoxicity values for data used as test set for neurotoxicity models, pED$_{50}$ represent negative logarithm value of the dose quantity that is effective in 50% of the tested animals and pTD$_{50}$ represent negative logarithm value of the dose quantity that present toxicity in 50% of the tested animals

Table 1 continued: Selected N-benzylacetamide and 3-(phenylamino)propanamide derivatives molecular structure, anticonvulsant and neurotoxicity values

S/N	Molecular structure	pED$_{50}$	pTD$_{50}$	S/N	Molecular structure	pED$_{50}$	pTD$_{50}$
17	 2-amino-N-benzyl-2-(pyridin-4-yl)acetamide	3.852	3.180	21	 2 -amino-N-benzyl-2-(5-methylfuran-2yl)acetamide	3.259	3.351^{b}
18	 2-amino-N-benzyl-2-(pyrazin-2-yl)acetamide	3.493	3.692	22	 2-amino-N-benzyl-2-(quinolin-2-yl)acetamide	3.567	3.493
19	 2-amino-N-benzyl-2-(naphtalen-1-yl)acetamide	3.060^{a}	3.692	23	 2-amino-N-benzyl-2-(4-fluorophenyl)acetamide	3.774	3.497^{b}
20	 2-amino-N-benzyl-2-(naphtalen-2-yl)acetamide	3.455^{a}	3.351	24	 2-amino-N-benzyl-2-(p-toly)acetamide	3.810^{a}	3.001

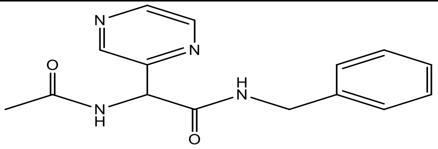
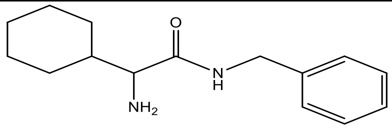
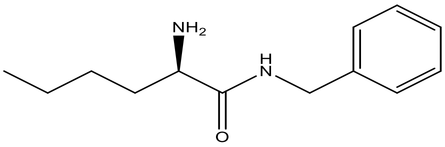
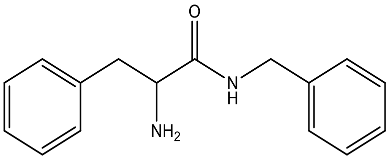
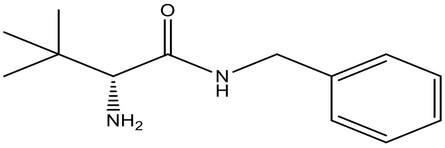
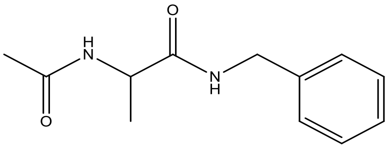
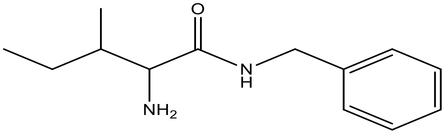
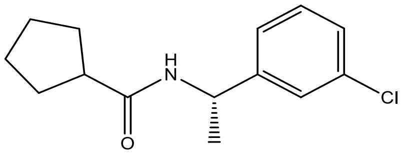
'a' represent the anticonvulsant activity values for data used as test set for anticonvulsant models, 'b' represent the neurotoxicity values for data used as test set for neurotoxicity models, pED$_{50}$ represent negative logarithm value of the dose quantity that is effective in 50% of the tested animals and pTD$_{50}$ represent negative logarithm value of the dose quantity that present toxicity in 50% of the tested animals

Table 1 continued: Selected N-benzylacetamide and 3-(phenylamino)propanamide derivatives molecular structure, anticonvulsant and neurotoxicity values

S/N	Molecular structure	pED$_{50}$	pTD$_{50}$	S/N	Molecular structure	pED$_{50}$	pTD$_{50}$
25	 2-amino-N-benzyl-2-(4-methoxypheny)acetamide	4.036	3.183	29	 2-amino-N-(2,6-dimethylphenyl) actamide	4.168	3.001
26	 2-amino-N-benzyl-2-(furan-2yl)acetamide	3.242^{a}	3.756^{b}	30	 N-(2,6-dimethylphenyl) cyclobutanecarboxamide	3.680	3.548
27	 2-amino-N-benzyl-2-(thiophen-2yl)acetamide	4.411	3.247	31	 (R)-2-acetamido-N-benzyl-2-phenylacetamide	3.944	3.577^{b}
28	 2-amino-N-benzyl-2-(thiazol-2yl)acetamide	4.278	3.337^{b}	32	 2-acetamido-N-benzyl-2-(pyridine-2-yl) acetamide	3.603	3.690^{b}

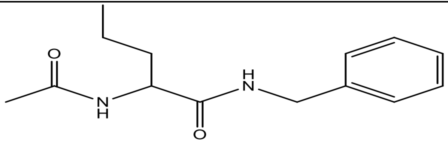
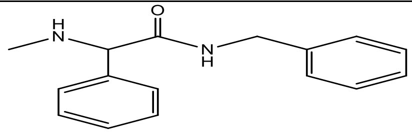
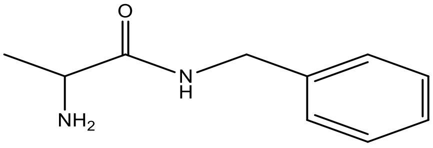
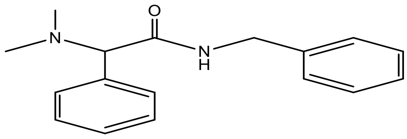
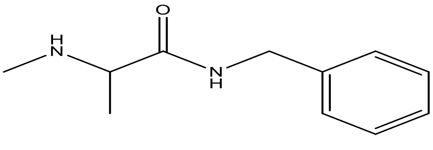
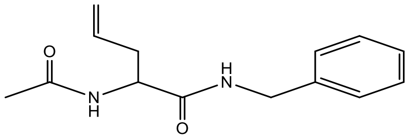
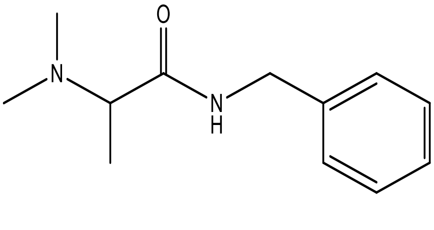
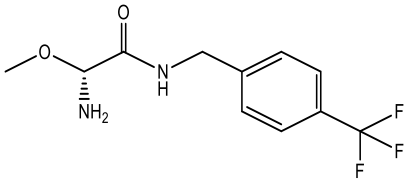
'a' represent the anticonvulsant activity values for data used as test set for anticonvulsant models, 'b' represent the neurotoxicity values for data used as test set for neurotoxicity models, pED$_{50}$ represent negative logarithm value of the dose quantity that is effective in 50% of the tested animals and pTD$_{50}$ represent negative logarithm value of the dose quantity that present toxicity in 50% of the tested animals

Table 1 continued: Selected N-benzylacetamide and 3-(phenylamino)propanamide derivatives molecular structure and activity values

S/N	Molecular structure	pED$_{50}$	pTD$_{50}$	S/N	Molecular structure	pED$_{50}$	pTD$_{50}$
33	 2-acetamido-N-benzyl-2-(pyrazine-2-yl) acetamide	3.256^{a}	3.495	37	 2-amino-N-benzyl-2-cyclohexylacetamide	3.433^{a}	3.191
34	 (R)-2-amino-N-benzylhexanamide	3.996	3.495	38	 2-amino-N-benzyl-3phenylpropanamide	3.559^{a}	3.367^{b}
35	 (R)-2-amino-N-benzyl-3,3-dimethylbutanamide	3.793	3.495^{b}	39	 2-acetamido-N-benzylpropanamide	3.764^{a}	3.472^{b}
36	 2-amino-N-benzyl-3-methylpentanamide	3.439	3.694	40	 N-(1-(3-chlorophenyl)ethyl)cyclopentane carboxamide	3.981^{a}	3.534^{b}

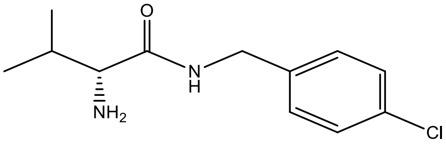
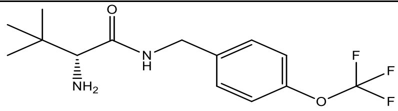
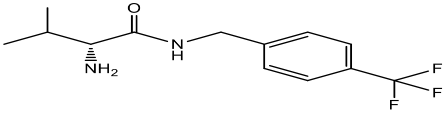
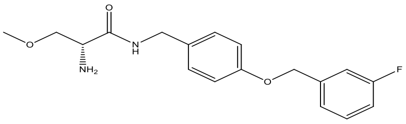
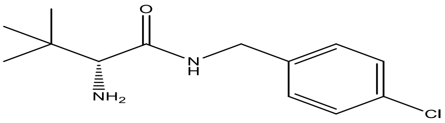
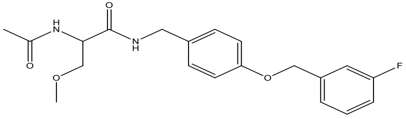
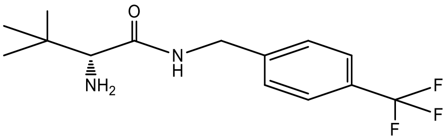
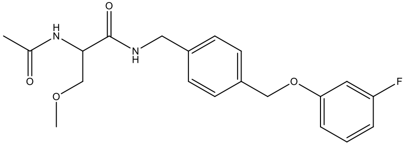
‘^{a}’ represent the anticonvulsant activity values for data used as test set for anticonvulsant models, ‘^{b}’ represent the neurotoxicity values for data used as test set for neurotoxicity models, pED$_{50}$ represent negative logarithm value of the dose quantity that is effective in 50% of the tested animals and pTD$_{50}$ represent negative logarithm value of the dose quantity that present toxicity in 50% of the tested animals

Table 1 continued: Selected N-benzylacetamide and 3-(phenylamino)propanamide derivatives molecular structure, anticonvulsant and neurotoxicity values

S/N	Molecular structure	pED$_{50}$	pTD$_{50}$	S/N	Molecular structure	pED$_{50}$	pTD$_{50}$
41	 2-acetamido-N-benzylpentanamide	3.743	3.343	45	 N-benzyl-2-(methylamino)-2-phenylacetamide	3.815^{a}	3.571
42	 2-amino-N-benzylpropanamide	3.803^{a}	3.288	46	 N-benzyl-2-(dimethylamino)-2-phenylacetamide	4.140	3.482
43	 N-benzyl-2-(methylamino)propanamide	3.866	3.110^{b}	47	 2-acetamido-N-benzylpent-4-enamide	4.039	3.744
44	 N-benzyl-2-(dimethylamino)propanamide	3.462^{a}	3.571	48	 (R)-2-amino-2-methoxy-N-(4-(trifluoromethyl) benzyl)acetamide	4.292	3.512

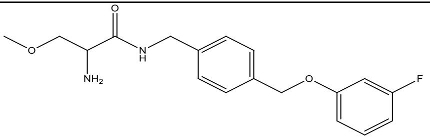
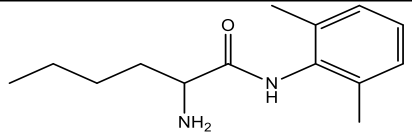
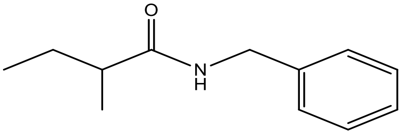
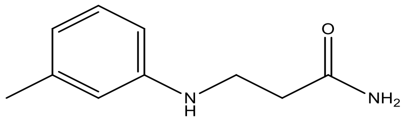
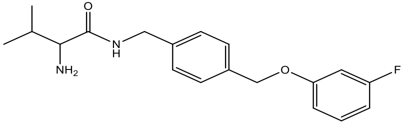
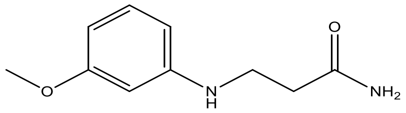
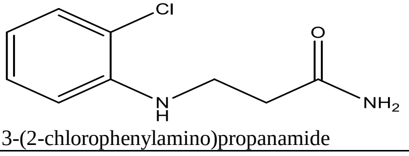
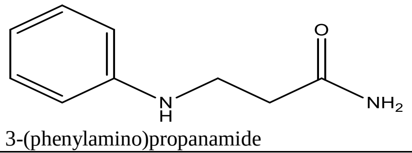
‘^{a}’ represent the anticonvulsant activity values for data used as test set for anticonvulsant models, ‘^{b}’ represent the neurotoxicity values for data used as test set for neurotoxicity models, pED$_{50}$ represent negative logarithm value of the dose quantity that is effective in 50% of the tested animals and pTD$_{50}$ represent negative logarithm value of the dose quantity that present toxicity in 50% of the tested animals

Table 1 continued: Selected N-benzylacetamide and 3-(phenylamino)propanamide derivatives molecular structure, anticonvulsant and neurotoxicity values

S/N	Molecular structure	pED$_{50}$	pTD$_{50}$	S/N	Molecular structure	pED$_{50}$	pTD$_{50}$
49	 (R)-2-amino-N-(4-chlorobenzyl)-3-methylbutanamide	4.080	3.682	53	 (R)-2-amino-3,3-dimethyl-N-(4-(trifluoromethoxy) benzyl)butanamide	4.346	3.758^{b}
50	 (R)-2-amino-3methyl-N-(4-trifluoromethyl)benzyl) butanamide	3.872^{a}	3.482	54	 (R)-2-amino-N-(4-((3 fluorobenzyl)oxy)benzyl)-3-methoxypropanamide	4.460	4.158^{b}
51	 (R)-2-amino-N-(4-chlorobenzyl)-3,3-dimethyl butanamide	4.036	3.336^{b}	55	 2-acetamido-N-(4-((3-fluorobenzyl)oxy)benzyl)-3-methoxypropanamide	4.804	4.573
52	 (R)-2-amino-3,3-dimethyl-N-(4-(trifluoromethyl) benzyl)butanamide	4.008^{a}	3.620^{b}	56	 2-acetamido-N-(4-((3-fluorophenoxy) methyl)benzyl)-3-methoxypropanamide	4.572^{a}	3.859^{b}

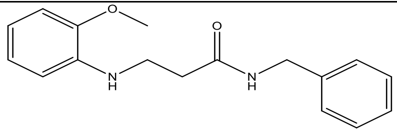
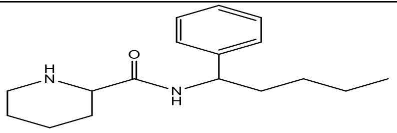
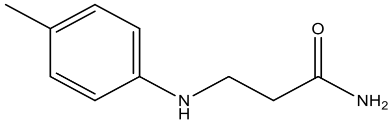
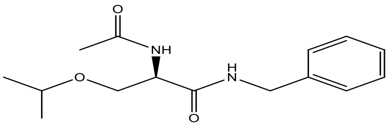
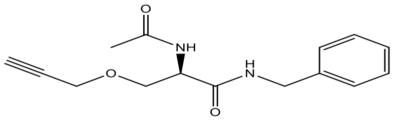
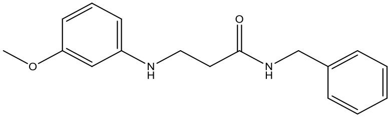
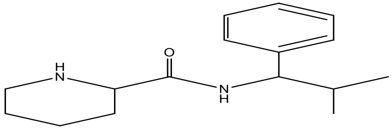
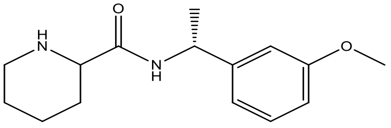
‘^{a}’ represent the anticonvulsant activity values for data used as test set for anticonvulsant models, ‘^{b}’ represent the neurotoxicity values for data used as test set for neurotoxicity models, pED$_{50}$ represent negative logarithm value of the dose quantity that is effective in 50% of the tested animals and pTD$_{50}$ represent negative logarithm value of the dose quantity that present toxicity in 50% of the tested animals

Table 1 continued: Selected N-benzylacetamide and 3-(phenylamino)propanamide derivatives molecular structure, anticonvulsant and neurotoxicity values

S/N	Molecular structure	pED$_{50}$	pTD$_{50}$	S/N	Molecular structure	pED$_{50}$	pTD$_{50}$
57	 2-amino-N-(4-((3-fluorophenoxy)methyl)benzyl)-3-methoxypropanamide	3.533	3.064	61	 2-amino-N-(2,6-dimethylphenyl) hexanamide	4.606	3.643
58	 N-benzyl-2-methylbutanamide	4.440^{a}	3.497	62	 3-(m-tolylamino)propanamide	4.421	3.986
59	 2-amino-N-(4-((3-fluorophenoxy)methyl)benzyl)-3-methylbutanamide	4.542	3.009^{b}	63	 3-(3-methoxyphenylamino) propanamide	4.275	3.590
60	 3-(2-chlorophenylamino)propanamide	4.333^{a}	3.809^{b}	64	 3-(phenylamino)propanamide	4.093^{a}	3.701^{b}

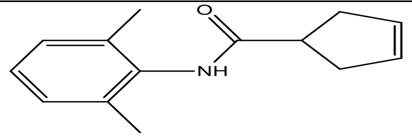
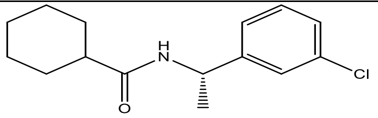
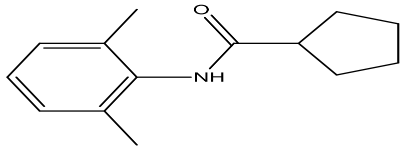
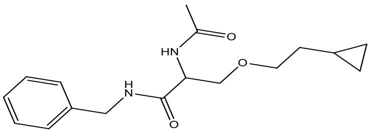
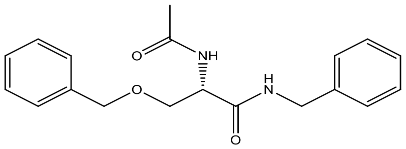
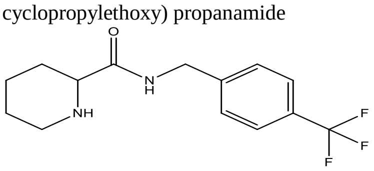
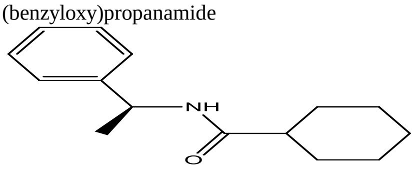
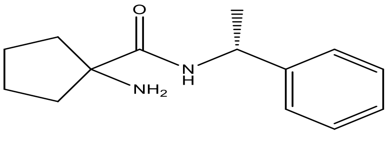
‘^{a}’ represent the anticonvulsant activity values for data used as test set for anticonvulsant models, ‘^{b}’ represent the neurotoxicity values for data used as test set for neurotoxicity models, pED$_{50}$ represent negative logarithm value of the dose quantity that is effective in 50% of the tested animals and pTD$_{50}$ represent negative logarithm value of the dose quantity that present toxicity in 50% of the tested animals

Table 1 continued: Selected N-benzylacetamide and 3-(phenylamino)propanamide derivatives molecular structure, anticonvulsant and neurotoxicity values

S/N	Molecular structure	pED$_{50}$	pTD$_{50}$	S/N	Molecular structure	pED$_{50}$	pTD$_{50}$
65	 N-benzyl-3-((2-methoxyphenyl)amino)propanamide	4.083^{a}	3.391^{b}	69	 N-(1-phenylpentyl)piperidine-2-carboxamide	4.096	3.367
66	 3-(p-tolylamino)propanamide	4.034	3.667	70	 (R)-2-acetamido-N-benzyl-3-isopropoxy propanamide	4.085	3.594
67	 (R)-2-acetamido-N-benzyl-3-(prop-2-yn-1-oxy) propanamide	4.234	3.596	71	 N-benzyl-3-((3-methoxyphenyl)amino) propanamide	3.756	2.856
68	 N-(2-methyl-1-phenylpropyl) piperidine-2-carboxamide	4.073	3.632^{b}	72	 N-((R)-1-(3-methoxylphenyl)ethyl)piperidin-2-carboxamide	3.630^{a}	3.157

‘a’ represent the anticonvulsant activity values for data used as test set for anticonvulsant models, ‘b’ represent the neurotoxicity values for data used as test set for neurotoxicity models, pED$_{50}$ represent negative logarithm value of the dose quantity that is effective in 50% of the tested animals and pTD$_{50}$ represent negative logarithm value of the dose quantity that present toxicity in 50% of the tested animals

Table 1 continued: Selected N-benzylacetamide and 3-(phenylamino)propanamide derivatives molecular structure, anticonvulsant and neurotoxicity values

S/N	Molecular structure	pED$_{50}$	pTD$_{50}$	S/N	Molecular structure	pED$_{50}$	pTD$_{50}$
73	 N-(2,6-dimethylphenyl)cyclopent-3-enecarboxamide	3.593	3.013^{b}	77	 N-(1-(3-chlorophenyl)ethyl)cyclohexane carboxamide	3.169	3.127
74	 N-(2,6dimethylphenyl) cyclopentanecarboxamide	3.545^{a}	3.744	78	 2-acetamido-N-benzyl-3-(2-cyclopropylethoxy) propanamide	3.821	3.554
75	 2-acetamido-N-benzyl-3-(benzyloxy)propanamide	3.708	3.213^{b}	79	 N-(4-(trifluoromethyl)benzyl)piperidin-2-carboxamide	3.834^{a}	3.336
76	 N-(1-phenylethyl)cyclohexanecarboxamide	3.483^{a}	3.003	80	 (R)-1-amino-N-(1-phenylethyl)cyclopentane carboxamide	3.733	3.266

‘a’ represent the anticonvulsant activity values for data used as test set for anticonvulsant models, ‘b’ represent the neurotoxicity values for data used as test set for neurotoxicity models, pED$_{50}$ represent negative logarithm value of the dose quantity that is effective in 50% of the tested animals and pTD$_{50}$ represent negative logarithm value of the dose quantity that present toxicity in 50% of the tested animals

basis functions to be used by testing full-size models rather than incrementally building them; it is better at discovering combinations of basis functions that take advantage of correlations between features; it incorporates the LOF (lack of fit) error measure developed by Friedman that resists overfitting and allows user control over the smoothness of fit; it can use a larger variety of basis functions in construction of its models, for example, splines, Gaussians, or higher-order polynomials; and study of the evolving models provides additional information, not available from standard regression analysis, such as the preferred model length and useful partitions of the data set [28].

In brief, GFA starts by generating an initial “m” population of equation by random choice of molecular descriptors. These equations form the parent generation and in each generation, the fitness of every individual descriptor in the population is evaluated. Roulette wheel parent selection rule is applied where $(\frac{m}{2})$ pairs of parents are chosen randomly from the current “m” population to form a new population of “children equation” by recombination (crossover) and possibly mutation of descriptors. Fitness of each children equation is evaluated and the more fit children equations are used to replace the parent equation for the next iteration step in the algorithm. The process continues and terminate when either a maximum number of generations has been produced, or a satisfactory fitness level has been reached for the population. In this work the equation length was set to range from 5 to 12 terms and a constant in order to comply with the generally known semi-empirical ‘rule of thumb’ and Toplis ratio [29, 30] which say the ratio of descriptor to the number of molecule used to construct a model should not exceed 1:6. The population size was set to 10000, maximum generation was set to 500, number of top equation returned was set to 3, mutation probability was set to 0.1, and scaled LOF smoothness parameter was set to 0.5. The selected combinations of descriptors with no or minimum co-linearity which are map to the anticonvulsant activity producing multiple QSAR models were saved for subsequent studies.

1.5. Statistical quality and model validation

Multiple Linear Regression (MLR) [31] and correlation analyses were carried out on each group of descriptor that constitute a model using Microsoft excel (version 2007) software in order to evaluate the statistical quality of the regression equations at statistical significance level of 0.05 ($P < 0.05$) and evaluate the extent of multicolinearity between descriptors . The evaluated parameters for assessing statistical quality includes square correlation coefficient (R^2), the adjusted squared correlation coefficient (R^2_{adj}) , Standard error of estimation (SEE), Predicted error sum of square (PRESS), the variance ratio F, the t-statistics and p-values for each descriptors. The extent of multi co-linearity between descriptors was evaluated with variance inflation factors VIF of each descriptor estimated from the correlation matrix among the descriptors only (excluding the activity column vector). The VIFs are the diagonal element of the inverse matrix obtained from the correlation matrix [32].

The models obtained were validated internally by calculations of the leave one out cross-validation squared correlation coefficient ($R^2_{CV} = Q^2$) values (33). In addition, the robustness of the proposed models were checked by permutation testing: parallel models were developed based on a fit to randomly reordered Y-data (Y-randomization) [24, 30]. According to the basic approach of Wold and Eriksson [33] all randomization methods consisted of ten randomization runs for any data set size. Externally, the models were validated by calculated predictive squared correlation coefficient R^2_{pred} and other criteria for predictive models proposed by Golbraikh and Tropsha [34]. All computations were performed on a HP core i5-4200U workstation.

1.6. Contribution of selected descriptors

The importance of each descriptor in the models in relation to other descriptor in the same model and each descriptor contribution to that model as a whole was estimated using the mean effect. Mean effect was calculated from the coefficient of each descriptor in a model and their value in the data matrix using the relationship below:

$$MF_j = \frac{\beta_j \sum_{i=1}^{i=n} d_{ij}}{\sum_j^m \beta_j \sum_i^n d_{ij}}$$

Here MF_j is the mean effect of a descriptor j in a model, β_j is the coefficient of the descriptor J in that model and d_{ij} is the of that descriptor of interest in the data matrix for each molecule in the training set, m is the number of descriptor that appear in the model and n is the number of molecules in the training set [35]

1.7. Models applicability domain

Leverage approach that utilizes Williams plot was employed to define the applicability domain of the models reported. Williams plot for a model is a graphical view of leverage values for each molecule in the entire data set versus their standardized cross validated residual obtained by the model [24]. The leverage (h_{ii}) value for each molecule is obtained has the diagonal elements of the hat matrix constructed for both training set and test set while, standardized cross validated residual for each molecule is obtained from the relation below estimated for the training and test set:

$$SDR = \frac{\hat{y} - y}{\sqrt{\frac{\sum_{i=1}^{n}(\hat{y} - y)^2}{n}}}$$

Here ‘y’ is the observe (experimental) activity value for either the training or the test set, $\hat{y}$ is the predicted activity value by the models either for the training or test set and n is the number of molecules either in the training or test set. The hat matrix for the training set was obtained using Microsoft excel through the following consecutive steps.

i. Add a column vector whose elements are only ‘1s’ to the descriptor only data matrix i.e. $\mathbf{X}_{\mathbf{tr}}$ (n$\times$m)

ii. Obtain the transpose of matrix $\mathbf{X}_{\mathbf{tr}}$ (n$\times$m) using the TRANSPOSE functions to obtain a new matrix $\mathbf{X}^{\mathbf{T}}_{\mathbf{tr}}$ (m$\times$ n).

iii. Post-multiply the transpose matrix $\mathbf{X}^{\mathbf{T}}_{\mathbf{tr}}$ with matrix $\mathbf{X}_{\mathbf{tr}}$ (n$\times$m) in step1 using the function =MMULT (array1, array2) i.e $\mathbf{X}^{\mathbf{T}}_{\mathbf{tr}}$(m$\times$ n)$\cdot\mathbf{X}_{\mathbf{tr}}$(n$\times$m) = $\mathbf{A}$(m$\times$m) i.e. $\mathbf{X}^{\mathbf{T}}_{\mathbf{tr}}\mathbf{X}_{\mathbf{tr}}$. Note that this multiplication is not commutative.

iv. Obtain the inverse of matrix A (m$\times$m) using the function =MINVERSE (array1) i.e $\mathbf{A}^{-1}$(m$\times$m)$\equiv$ $(\mathbf{X}^{\mathbf{T}}_{\mathbf{tr}}\mathbf{X}_{\mathbf{tr}})^{-1}$. This matrix $\mathbf{A}^{\mathbf{T}}$ is sometime called ‘the clone’ matrix (Mniovski *et al*, 2007).

v. Post-multiply the matrx $\mathbf{X}_{\mathbf{tr}}$ with ‘the clone’ i.e. $\mathbf{A}^{-1}$(m$\times$m) using the function =MMULT(array1,array2) = i.e. $\mathbf{X}_{\mathbf{tr}}$(n$\times$m) $\cdot$ $\mathbf{A}^{\mathbf{T}}$(m$\times$m) =$\mathbf{B}$(n$\times$m) i.e $\mathbf{X}_{\mathbf{tr}}\mathbf{A}^{-1}$(n$\times$m). Note that this multiplication is not commutative.

vi. Finally obtain the hat matrix ‘$\mathbf{H}_{\mathbf{tr}}$’ for the training set by post multiplication of matrix $\mathbf{B}$(n$\times$m) with the transpose matrix $\mathbf{X}^{\mathbf{T}}_{\mathbf{tr}}$ (m$\times$ n) obtain in step2 i.e. $\mathbf{B}$(n$\times$m)$\cdot$ $\mathbf{X}^{\mathbf{T}}_{\mathbf{tr}}$ (m$\times$ n) = $\mathbf{H}_{\mathbf{tr}}$ (n$\times$n).

Here n is the number of molecule in that make up the training set and m is the number of descriptors that appear in a model. The hat matrix for the training set can be expressed in term of the descriptor matrix($\mathbf{X_{tr}}$) only as $\mathbf{X_{tr}(X^T{}_{tr}X_{tr})^{-1}X^T{}_{tr}}$ and the diagonal elements of this matrix represent the leverages 'h_{ii}' for the molecules that made up the training set. If X_{te} is the descriptor matrix for molecule in the test set arranged in sequence as that of the training set, the hat matrix for test-set 'H_{te}' was obtained using similar procedure with slight medication

i. Add a column vector whose elements are only '1s' to the descriptor only data matrix i.e. $\mathbf{X_{te}}$ (z×m)

ii. Obtain the transpose of matrix $\mathbf{X_{te}}$ (z×m) using the TRANSPOSE functions to obtain a new matrix $\mathbf{X^T{}_{te}}$ (m× z).

iii. Use the clone matrix obtained for the training set i.e. $\mathbf{A^{-1}}$(m×m) ≡($\mathbf{X^T{}_{tr}X_{tr}}$)$^{-1}$ to post multiply the matrix $\mathbf{X_{te}}$ (n×m) i.e $\mathbf{X_{te}}$ (z×m) ·$\mathbf{A^{-1}}$(m×m) = $\mathbf{C}$(z×m). This is an attempt to map the test into the domain of the training set.

iv. Finally obtain the hat matrix '$\mathbf{H_{te}}$' for the test set by post multiplication of matrix $\mathbf{C}$(z×m) with the transpose matrix $\mathbf{X^T{}_{te}}$ (m× z) obtain in step2 i.e. $\mathbf{C}$(z×m)· $\mathbf{X^T{}_{tr}}$ (m× z) = $\mathbf{H_{te}}$ (z×z).

Here z is the number of molecule in that make up the training set and m is the number of descriptors that appear in a model. The hat matrix for the training set can be expressed in term of the descriptor matrix($\mathbf{X_{te}}$) only as $\mathbf{X_{te}(X^T{}_{tr}X_{tr})^{-1}X^T{}_{te}}$ and the diagonal elements of this matrix represent the leverages 'h_{ii}' for the molecules that made up the test set. The leverages obtained for both set are plotted as abscissa against the standardized cross validated residual which is the ordinate to obtain the Williams plot and the cut leverage 'h*' on the x-axis is obtained using the relation below, while the cut off for the standardized cross validated residual on the y-axis has been reported to be 2.5 ≤ y ≤ 3 [36, 37]

$$h^* = \frac{3(m+1)}{n}$$

Here m is the number of descriptors that appear in a model and n is the number of molecule in that make up the training set only.

1.8. In silico generation of compounds

QSAR models have always been used as powerful tools for virtual screening of compounds with a given biological activity [12]. It can also be used as knowledge generators, that is, by interpreting the meaning of the molecular descriptors in the models, it could be possible to create new molecules that in principle would have the given biological activity [39]. Combining the knowledge about a reference molecule in the data set, knowledge about the relative important of the molecular descriptors in the models and structural interpretation of the most important molecular descriptors employed to create the model, effort was made to use the QSAR models reported as a knowledge generators to create new hypothetic molecules that in principle would have improved anticonvulsant activity and lesser neurotoxicity.

3. Result and discussion

1.9. Data set

Single column statistics performed on the training set and test set data reported in Table2 show that the maximum of the test set for both the anticonvulsant and neurotoxicity is less than the maximum and training set. The minimum of test set for anticonvulsant activity is also less than the minimum for its training set, also, the minimum for neurotoxicity are approximately equal for both set. This indicates that the test set is interpolative i.e. derived within the minimum – maximum range of the training set. The mean and the standard deviation of the two data sets provide insight to the relative difference of mean and point distribution of the two set. In these cases for both anticonvulsant and neurotoxicity value the mean and standard deviation of the training set and test set are similar. This shows that the spread in both set are comparable.

Table 2: Descriptive statistics for the anticonvulsant and neurotoxicity value of the data set

	–log (ED_{50})		–log (TD_{50})	
Statistics	Train set	Test set	Train set	Test set
Maximum	4.804	4.572	4.573	4.158
Minimum	3.169	3.060	2.856	3.009
Mean	3.979	3.818	3.477	3.472
σ	0.364	0.369	0.291	0.275

-log (ED50) represent the anticonvulsant activity and –log (TD50) represent the neurotoxicity value and σ is the standard deviation.

1.10.QSAR models

The analysis of the GFA models produced explores the structural and physicochemical contribution of the compounds with anticonvulsant activity and neurotoxicity. A number of molecular descriptors were identified as being correlated with anticonvulsant and neurotoxicity values. Interestingly, there seems to be high overlap of few descriptors among the QSAR equations. Reported in Table 3 is the Top 3 QSAR models for anticonvulsant activities and Table 4 for neurotoxicity with their statistical and validation parameters. The presented models incorporate 4 to 6 descriptors and since the Topliss and Costello rule [31] allows the use of up to 8 descriptors for a training set consisting of 50 compounds and the relationship between adjusted coefficient of determination (R^2_{adj}) and coefficient of determination (R^2) i.e. $R^2_{adj} < R^2$ is true for these models, then they are not over parameterized. The multiple correlation coefficient R of these models ranges from 0.922 to 0.961 explaining over 91% of all variance in the data set for both anticonvulsant activity and neurotoxicity. Supporting the claim, many independent QSAR models can provide useful activity correlation on the same data [28]. Presented in Table 5 and 6 are the model statistics for the anticonvulsant and neurotoxicity models reported in Table 3 and 4 respectively. From Tables 5 and 6, the F test value at $p < 0.05$ significant level ranges from 45.91 to 63.89 for the anticonvulsant models and 115.3 to 134.9 for neurotoxicity models. These values were high therefore; the variation in the activity explained by the collective descriptors is more than could be reasonably attributed to chance. Also, the low standard error of estimation (SEE) for the models ranges from 0.087 to 0.153, suggesting that the equations have good correlation with the data and is statistically significant. The t-statistics for all descriptor were greater than 2 and their corresponding p-values are less than 0.05 at 95% confidence level, suggesting all the descriptors in the regression equations were independent and make significant contribution to the models. The correlation between each descriptor was calculated and used to evaluate variance inflation factor (VIF) for each descriptor. Generally, if VIF is equal to 1, then no inter-correlation exists for each variable; if it falls between 1and 5, the related model is acceptable; if it is larger than 10, the related model is unstable and a recheck is necessary [38]. In the study, VIF values of the all the descriptors were less than five (see Table 5 and 6), indicating that the obtained models have statistical significance, and the descriptors were found to be reasonably orthogonal [39].

Furthermore, the predictive powers of the models assessed using various internal and external validation tests showed that the leave one out internal cross validation coefficient of determination R^2_{CV} (Q^2) ranges from 0.821 to 0.897. Also, modified correlation coefficient $R^2_{m(loo)}$ introduced by Roy et al [42] such that

$$R^2_{m(loo)} = r^2 \times \left(1 - \sqrt{(r^2 - r_0^2)}\right)$$

where r^2 and r^2_0 are the square correlation coefficients between the observed and the (leave-one-out) predicted values of compounds with and without intercept respectively, had values ranging from 0.793 to 0.899 for the models. These values were greater than 0.5, meaning the models are stable. Also, the models R^2 values were greater than their Q^2 values, this indicated that they were not over-fitted [40].

In order to ascertain whether the good results produced by the reported models are not due to chance correlation or structural dependency of the training set, Y-randomization tests were performed. The average values for ten randomization runs gave coefficient of determination $\bar{R}^2_{rand}$ and cross validated coefficient of determination $\bar{Q}^2_{rand}$ values ranging from -0.216 to 0.117 (see Table 5 and 6). These values were smaller than 0.2 indicating absence of chance correlation [40]. Furthermore, a parameter ${}^cR^2_p$ that relate the average randomization runs coefficient of determination $\bar{R}^2_{rand}$ and non-randomized R^2 was calculated for all the models using the relation

$${}^cR^2_p = R \times \sqrt{R^2 - \bar{R}^2_{rand}}$$

Their ${}^cR^2_p$ values ranged from 0.803 to 0.899. These values are more than the threshold which is 0.5 and therefore the models are not obtained by chance [40].

To further validate the predictive power of the models more explicitly, they were used to predict the activity of the test set data. The test set constitute s set of 30 compounds obtained from the data set by the method employed for dataset division as explained above .The R^2_{pred} which reflect the degree of correlation between the observed and predicted activity data for the test set [40] for the models were calculated using the relation below

$$R^2_{pred} = 1 - \frac{\Sigma\left(Y_{obs(test)} - Y_{pred(test)}\right)^2}{\Sigma\left(Y_{obs(test)} - \bar{Y}_{training}\right)^2}$$

Here, $Y_{obs(test)}$ and $Y_{pred(test)}$ are the observed and predicted activity data for the test set compounds, while $\bar{Y}_{training}$ indicates the mean observed activity of the training set. As can be seen in Table 3 and 4 for the anticonvulsant models, Model1 and Model2 have R^2_{pred} less than the stipulated value 0.5 therefore they are considered to be less predictive but Model 3 has R^2_{pred} value of 0.735 and therefore considered as well predictive [40]. For the neurotoxicity models, Model5 and Model6 have R^2_{pred} less than the stipulated value 0.5 therefore they are considered to be less predictive but Model3 has R^2_{pred} value of 0.509 and therefore considered as well predictive. Golbraikh and Tropsha [34] proposed set of parameters for determining satisfactorily external predictive model were examined including:

a. $Q^2 > 0.5$
b. $R^2_{(test)} > 0.6$
c. $r^2 - r^2_0/r^2 < 0.1$ and $0.85 \leq k \leq 1.15$ or $r^2 - r'^2/r^2 < 0.1$ and $0.85 \leq k' \leq 1.15$
d. $|r^2_0 - r'^2_0| < 0.3$

Where $R^2_{(test)}$ and r^2 represent the same parameter i.e. the square correlation coefficients of the plot of observed versus predicted values for the test set with intercept, r^2_0 is the square correlation coefficients between observed versus predicted values for the test set without intercept i.e. through the origin, r'^2_0 is the reverse of r^2_0 i.e. the square correlation coefficients between the predicted versus observed values for the test set without intercept, k is the slope of the plot of observed versus predicted values for the test set without intercept and k′ is the reverse of k i.e. the slope of the plot of predicted versus observed values for the test set without intercept. For the anticonvulsant models reported in Table 3, Model1 and Model2 have $r^2 - r^2_0/r^2$ equal to 0.300 and 0.395 respectively, while Model3 passed all the Golbraikh and Tropsha criteria. For the neurotoxicity models reported in Table 4, Model5 and Model6 have their r2 < 0.6 and $r^2 - r^2_0/r^2$ equal to 0.329 and 0.353 respectively, while Model4 passed all the Golbraikh and Tropsha criteria. In addition, modified determination coefficient for the test set data, designated $R^2_{m(test)}$ determine the propinquity between observed and predicted activity of the test set using the relation and reported that a model is acceptable if it has $R^2_{m(test)} > 0.5$ [39, 40] For the anticonvulsant models reported in Table 3, Model1 and Model2 have $R^2_{m(test)}$ values equal to 0.473 and 0.436 respectively, while model3 has $R^2_{m(test)}$ $R^2_{m(test)}$ value of 0.648. For the neurotoxicity models reported in Table 4, Model4 and Model5 have $R^2_{m(test)}$ values equal to 0.558 and 0.507 respectively, while model6 has $R^2_{m(test)}$ $R^2_{m(test)}$ value of 0.212. The above analysis shows that Model3 and Model4 are the best model for this subset of data reported by the Euclidean based clustering coupled with GFA

Table3: Anticonvulsant models obtained by genetic function algorithm from Euclidean distance based clustering data division method

Model 1				Model 2				Model 3			
pED_{50} = 3.996(±0.022)				pED_{50} = 3.992(±0.022)				pED50 = 3.948(±0.021)			
-0.107(±0.021)AATS7p				-0.112(±0.021)AATS7p				+0.310(±0.025) AATS4i			
+0.287(±0.038)ATSC4i				+0.316(±0.039)ATSC4i				+0.388(±0.028) Kier2			
+0.163(±0.032)GATS8s				+0.196(±0.031)GATS8s				+0.288(±0.024) VE2_D			
+0.374(±0.047)SpMax4Bhs				+0.488(±0.044)SpMax4Bhs				-0.193(±0.024) RDF50s			
+0.241(±0.031)Kier1				-0.319(±0.028)RDF50s							
-0.294(±0.028)RDF50s				+0.202(±0.025)Surface area							
VALIDATION PARAMETERS											
Internal		External		Internal		External		Internal		External	
R^2	0.885	R^2_{pred}	0.482	R^2	0.873	R^2_{pred}	0.408	R^2	0.850	R^2_{pred}	0.735
R^2_{adj}	0.869	r^2	0.666	R^2_{adj}	0.855	r^2	0.661	R^2_{adj}	0.837	r^2	0.699
F	55.25	r^2_0	0.466	F	49.32	r^2_0	0.399	F	63.89	r^2_0	0.688
Q^2	0.850	r'^2_0	0.665	Q^2	0.823	r'^2_0	0.658	Q^2	0.821	r'^2_0	0.648
SDEP	0.143	$R^2_{m(test)}$	0.473	SDEP	0.159	$R^2_{m(test)}$	0.436	SDEP	0.152	$R^2_{m(test)}$	0.646
SEE	0.135	$R^2_{m(overall)}$	0.631	SEE	0.145	$R^2_{m(overall)}$	0.585	SEE	0.146	$R^2_{m(overall)}$	0.755
PRESS	1.026	$\|r^2_0-r'^2_0\|$	0.199	PRESS	1.265	$\|r^2_0-r'^2_0\|$	0.258	PRESS	1.159	$\|r^2_0-r'^2_0\|$	0.041
LOF	0.084	k	0.969	LOF	0.084	k	0.969	LOF	0.085	k	0.993
$R^2_{m(loo)}$	0.833	$r^2_0- r^2_0/r^2$	0.300	$R^2_{m(loo)}$	0.793	$r^2_0- r^2_0/r^2$	0.395	$R^2_{m(loo)}$	0.806	$r^2- r^2_0/r^2$	0.015
$\bar{R}^2_{rand}$	0.107	k′	1.026	$\bar{R}^2_{rand}$	0.117	k′	1.026	$\bar{R}^2_{rand}$	0.101	k′	1.003
$\bar{Q}^2_{rand}$	-0.209	$r^2-r'^2/r^2$	0.001	$\bar{Q}^2_{rand}$	-0.216	$r^2-r'^2/r^2$	0.005	$\bar{Q}^2_{rand}$	-0.106	$r^2-r'^2/r^2$	0.074
$^cR^2_p$	0.814	R	0.941	$^cR^2_p$	0.807	R	0.934	$^cR^2_p$	0.803	R	0.922

Table4: Neurotoxicity models obtained by genetic function algorithm from Euclidean distance based clustering data division method

Model 4				Model 5				Model 6			
pTD_{50} = 3.479(±0.013)				pTD_{50} = 3.479(±0.012)				pTD_{50} = 3.487(±0.013)			
+0.522(±0.026) TIC5				-0.211(±0.014) ETA_Epsilon_3				+0.4278(±0.023) TIC5			
-0.227(±0.016) nRing				+0.482(±0.024)TIC5				+0.180(±0.013)RotBtFrac			
+0.148(±0.015)VE1_D				+0.125(±0.013)VE1_D				+0.125(±0.015)VE2_D			
-0.269(±0.022)RDF60i				-0.256(±0.021) RDF60i				-0.231(±0.022)RDF60i			
VALIDATION PARAMETERS											
Internal		External		Internal		External		Internal		External	
R^2	0.911	R^2_{pred}	0.509	R^2	0.923	R^2_{pred}	0.375	R^2	0.919	R^2_{pred}	0.362
R^2_{adj}	0.903	r^2	0.623	R^2_{adj}	0.916	r^2	0.561	R^2_{adj}	0.912	r^2	0.286
F	115.3	r^2_0	0.573	F	134.9	r^2_0	0.377	F	128.3	r^2_0	0.185
Q^2	0.890	r'^2_0	0.583	Q^2	0.904	r'^2_0	0.552	Q^2	0.897	r'^2_0	-0.116
SDEP	0.096	$R^2_{m(test)}$	0.558	SDEP	0.094	$R^2_{m(test)}$	0.507	SDEP	0.096	$R^2_{m(test)}$	0.212
SEE	0.091	$R^2_{m(overall)}$	0.729	SEE	0.087	$R^2_{m(overall)}$	0.687	SEE	0.089	$R^2_{m(overall)}$	0.677
PRESS	0.457	$\|r^2_0-r'^2_0\|$	0.010	PRESS	0.424	$\|r^2_0-r'^2_0\|$	0.176	PRESS	0.459	$\|r^2_0-r'^2_0\|$	0.301
LOF	0.095	k	0.981	LOF	0.095	k	0.990	LOF	0.096	k	0.995
$R^2_{m(loo)}$	0.886	$r^2- r^2_0/r^2$	0.079	$R^2_{m(loo)}$	0.899	$r^2- r^2_0/r^2$	0.329	$R^2_{m(loo)}$	0.892	$r^2- r^2_0/r^2$	0.353
$\bar{R}^2_{rand}$	0.086	k′	1.017	$\bar{R}^2_{rand}$	0.101	k′	1.007	$\bar{R}^2_{rand}$	0.053	k′	1.001
$\bar{Q}^2_{rand}$	-0.122	$r^2-r'^2/r^2$	0.064	$\bar{Q}^2_{rand}$	-0.109	$r^2-r'^2/r^2$	0.017	$\bar{Q}^2_{rand}$	-0.171	$r^2-r'^2/r^2$	1.407
$^cR^2_p$	0.874	R	0.955	$^cR^2_p$	0.875	R	0.961	$^cR^2_p$	0.899	R	0.958

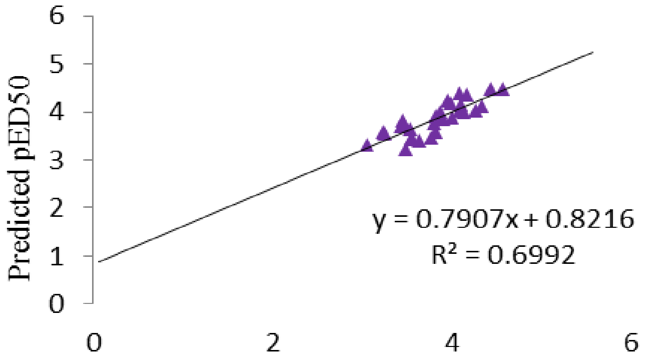
Table5: The t-statistic, p-values, variance inflation factor and mean effect for each descriptor in the models for anticonvulsant activity obtained by Euclidean distance based clustering data division method coupled with genetic function algorithm

	Model 1				Model 2				Model 3			
descriptors	t-stat	p-value	VIF_j	MF_j	t-stat	p-value	VIF_j	MF_j	t-stat	p-value	VIF_j	MF_j
AATS7p	-5.237	4.6E-06	1.232	-0.553	-5.438	2.3E-06	1.252	-0.313				
ATSC4i	7.476	2.6E-09	2.939	0.899	8.716	4.7E-11	2.899	0.434				
AATS4i									12.32	4.3E-16	1.610	0.166
GATS8s	5.066	8.1E-06	1.040	1.911	5.430	2.5E-06	1.050	1.795				
SpMax4Bhs	8.058	3.9E-10	4.836	2.442	11.23	2.3E-14	4.317	1.955				
Kier 1	7.797	9.2E-10	2.593	-1.654								
Kier 2									14.04	4.6E-18	2.247	0.388
RDF 50s	-10.51	1.8E-13	2.009	-2.838	-10.56	1.6E-13	2.074	-1.865	-8.013	3.3E-10	1.640	-0.512
VE2_D									11.80	2.3E-15	1.493	0.959
S.area					7.220	6.2E-09	1.856	-0.826				

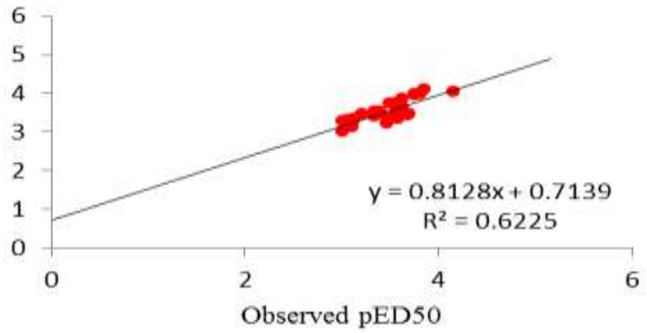
Table 6: The t-statistic, p-values, variance inflation factor and mean effect for each descriptor in the models for Neurotoxicity values obtained by Euclidean distance based clustering data division method coupled with genetic function algorithm

	Model 4				Model 5				Model 6			
descriptors	t-stat	p-value	VIF_j	MF_j	t-stat	p-value	VIF_j	MF_j	t-stat	p-value	VIF_j	MF_j
TIC5	20.08	4.2E-74	2.994	-12.45	20.46	1.9E-24	3.268	3.234	18.57	9.8E-23	3.119	1.795
nRing	-14.24	2.7E-18	1.749	11.41								
ETA_ε_3					-15.18	2.5E-19	1.404	-0.952				
RotBtFrac									14.16	3.5E-18	1.088	-0.373
VE1_D	9.764	1.1E-12	1.247	-5.149	9.682	1.4E-12	1.131	0.665				
VE2_D									8.57	5.1E-11	1.089	0.570
RDF 60i	-12.22	6.8E-16	2.358	7.177	-12.12	9.1E-16	2.851	-1.946	-10.49	1.2E-13	2.984	-0.993

method used in this work for anticonvulsant activity and neurotoxicity values respectively. Thus, it can be suggested that these models can be used for the prediction of the anticonvulsant activities and neurotoxicity values for new compounds in the domain of these models. Observed and predicted values of anticonvulsant activity and neurotoxicity values for the training and test set compounds by Model 3 and 4 are given in Table 7 and graphically represented in Figure 1.



(a)



(b)

Figure1. Graphs showing the observed versus predicted activity of test set compounds for (a) anticonvulsant by Model 3 and (b) neurotoxicity by Model 4

1.11.Models applicability domain

Applicability domain (AD) is the physic-chemical, structural and biological space on which the training set of the model has been developed and for which the model can make reliable prediction [41]. It ensure that the model is used to predict only those compounds that are similar, in term of a given distance measure, to training set compounds. The AD for the created models that is defined using Williams plots are given in Figure 2. This allows a graphical detection of both outliers and influential compounds in the models [42. The influence of compounds on the models was detected when the leverage value is greater than the warning leverage (in these cases h* = 0.3) [43]. The plot shows that the leverage values for the compounds are less than the warning value of 0.3 except for molecule 39 for neurotoxicity model. Also from Figure 2, it is obvious that there are no outlier compounds with standard residual > 3δ or < -3 δ for both training set and test sets. All result confirms that Model3 and Model4 are valid and can be utilized to predict the anticonvulsant activity and neurotoxicity values respectively for compounds in the AD of the models.

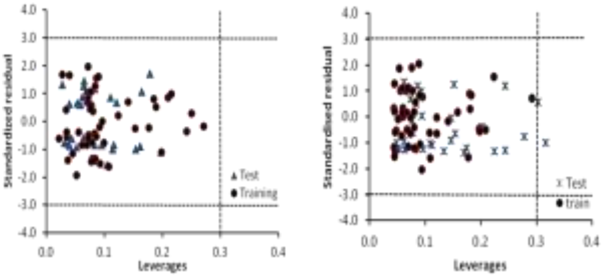
2. Interpretation of descriptors for anticonvulsant activity

Table7. Observed and predicted values of anticonvulsant activity and neurotoxicity values for test set compounds

	Anticonvulsant				Neurotoxicity		
Sr. No	Y^a_{obs}	Y^a_{pred}	Res	Sr.No	Y^b_{obs}	Y^b_{pred}	Res
2	4.16	4.35	-0.18	5	3.53	3.74	-0.20
5	4.27	4.02	0.24	7	3.55	3.66	-0.11
7	3.94	4.22	-0.27	9	3.47	3.22	0.25
10	4.07	4.03	0.05	10	3.06	3.31	-0.24
11	3.53	3.64	-0.11	13	3.62	3.42	0.19
14	4.14	3.97	0.17	15	3.12	3.36	-0.23
16	3.90	3.82	0.08	16	3.13	3.32	-0.19
19	3.06	3.30	-0.24	21	3.35	3.52	-0.17
20	3.46	3.80	-0.35	23	3.49	3.74	-0.25
24	3.81	3.57	0.23	26	3.75	3.94	-0.19
26	3.24	3.57	-0.33	28	3.33	3.40	-0.07
33	3.26	3.54	-0.29	31	3.57	3.35	0.22
37	3.43	3.70	-0.26	32	3.69	3.45	0.23
38	3.56	3.42	0.14	35	3.49	3.30	0.18
39	3.76	3.46	0.30	38	3.36	3.51	-0.14
40	3.98	4.17	-0.19	39	3.47	3.37	0.09
42	3.80	3.76	0.04	40	3.53	3.41	0.12
44	3.46	3.71	-0.25	43	3.11	3.14	-0.0
45	3.82	3.89	-0.07	51	3.34	3.52	-0.18
50	3.87	3.98	-0.11	52	3.62	3.85	-0.23
52	4.01	3.86	0.15	53	3.76	3.99	-0.23
56	4.57	4.48	0.09	54	4.16	4.05	0.11
58	4.44	4.46	-0.02	56	3.86	4.11	-0.25
60	4.33	4.12	0.22	59	3.01	3.27	-0.26
64	4.09	4.13	-0.04	60	3.81	3.99	-0.19
65	4.08	4.37	-0.28	64	3.70	3.48	0.22
72	3.63	3.39	0.24	65	3.39	3.51	-0.12
74	3.55	3.45	0.09	68	3.63	3.64	-0.01
76	3.48	3.21	0.27	73	3.01	3.01	0.01
79	3.83	3.83	0.01	75	3.21	3.46	-0.24

Note: the serial number correspond the position occupied by the test set data in Table 1. Y^a_{obs} and Y^a_{pred} are the observed and predicted anticonvulsant activity values. Y^b_{obs} and Y^b_{pred} are the observed and predicted neurotoxicity value. Res is the difference between observed and predicted neurotoxicity value

The selected model for anticonvulsant activity is Model 3 (see Table 3) and contained the following descriptors AATS4i, Kier2, VE2_D and RDF50s. The mean effect (MF) values for each descriptor are reported in Table 5. Generally QSAR equation is a result of mutual effect of different descriptors on the activity. The overall activity of molecule is determined by increase, decrease or mutual difference in the value of these descriptors. Assessment of these descriptors may be a starting point for designing anticonvulsant molecule with improved activities and to gain useful chemical insight into the anticonvulsant activity for the compounds used in training the model, interpretation to the descriptor in the model is provided below.



(a) (b)

Figure2. Williams plot for (a) anticonvulsant (Model3) and (b) neurotoxicity (Model4)

The first descriptor in the model is AATS4i belonging to the family of autocorrelation descriptors that are based on autocorrelation function using a particular physico-chemical properties as the weighting scheme. It is a member of the most known spatial autocorrelation of a topological structure (ATS) descriptors defined on a molecular graph which are based on graph invariant describing how the property considered is distributed along the topological structure of molecule [47]. Usually, they assume an additive scheme and thus correspond to a decomposition of the square molecular property consider in different atomic contributions [13]. The average of spatial autocorrelation of a topological structure (AATS) descriptors are obtained from ATS by dividing each term from by the corresponding number of contributions, thus avoiding any dependence on molecular size [13]. Generally, this class of descriptors expresses how numerical values of the autocorrelation function at intervals equal to the lag are correlated. AATS4i is specifically described as the average autocorrelation of topological structure (Broto-Moreau) -lag 4 weighted by first ionization potential [13]. The first ionization potential of pairs of atoms that are four bonds apart (i.e. path length of four) is the weighting scheme used, thus, changing the type of atoms or group that are four path length apart changes the AATS4i values. It was reported that ATS_4 value increase as the ionization energy of substituted atom reduces [13]. Using training set only (see Table 1and 8) it was observed that molecules containing phenyl ring carrying oxygen and halogen have a higher value of this descriptor. AATS4i has the lowest absolute mean effect (MF) value of 0.166 indicating it has the least influence in determining the anticonvulsant activity of the molecules used to construct the model in relation to other descriptor [14]. Considering the regression coefficients in the equation and the sign of the MF, higher numerical value of this descriptor reduces ED_{50} value thereby increasing the anticonvulsant activity of the molecule [35].

The second descriptor in the model is Kier2 defined as the second order shape attribute (second order kappa shape index) [44]. It belongs to the family of topological indexes, based on molecular connectivity approach that give the shape attribute of the molecule. These indexes were derived based on the assumption that the shape of a molecule is a function of the number of atoms and their adjacency relationships(bonding scheme) and the use path count designated as ‘^{m}P’ was used as information source to capture the information about different structure possessing different shape, . The counts of each order of path length can be viewed as describing individual attributes of shape, each a part of the manifold of attributes into which shape may be dissected. In order to transform ‘^{m}P’ into an index that carries information for any number of atoms in the molecule a particular shape attribute ‘$^{m}P_i$’ having an intermediate relationship between two extreme shapes was defined. It is generally true that the shape of each molecule in an isomeric series is different and so, the extreme shapes must be common to subsets of molecules of any number of atoms. Therefore, the extremes selected for any order of attribute, m, are the maximum, $^{m}P_{max}$, and minimum, $^{m}P_{mim}$, counts of paths in the molecular graphs of molecules with a common atom count. Therefore, a shape attribute of a particular order, m, for a particular molecule, i, is:

$^{m}P_{max} \geq {}^{m}P_i \geq {}^{m}P_{min}$

Note that the number of atoms is the same for all three structures. This set of numerical relationships was transformed into a single number for each attribute by examining each order ‘m’ and deriving an algorithm to encode the shape information. Specifically for second order shape attribute-defined by the count of two-bond path ‘$^{2}P_i$’ and related to the shape extremes represented by $^{2}P_{max}$ and $^{2}P_{min}$ a star graph in which all atoms but one are adjacent to a central atom was used and the numerical value of $^{2}P_{max}$ for any count of atom A is $^{2}P_{max}$ = (A-1)(A-2)/2 and the numerical value of ^{2}Pmin for any count of atom A is $^{2}P_{min}$ = A-2. An algorithm which yields a numerical index for any molecule with A atoms and with $^{2}P_i$ was derived (embracing the information for the extremes in the equation above) using the product ratios of $^{2}P_i$ to $^{2}P_{max}$ and $^{2}P_i$ to $^{2}P_{min}$ such that it give an index of shape of order two (using $^{2}\kappa$ (kappa) as the index symbol) [44]:

$^{2}\kappa = 2\ (^{2}P_{max}\ {}^{2}P_{min})/(^{2}P_i)^2$

The scaling factor of 2 in the numerator makes the value ^{2}K = A-1 for all linear graphs, where A-1 is the number of graph edges of skeletal bonds for acyclic molecules. This equation can be expressed in terms of the count of atoms A:

$^{2}\kappa = (A - 1)\ (A - 2)^2/\ (^{2}Pi)^2$

$^{2}\kappa$ encode information about branching and relative spatial density of molecules i.e. it gives information relating to the degree of star graph or linear graph-likeness of a molecule. With increased branching or cyclicity the value of this descriptor decreases. Its value increases with increase in the covalent radii of the molecule - as reported for butyl halide with bromobutane having the highest $^{2}\kappa$ – because the changes parallel a shape that is less star-like [44]. Using training set only (Table 1and 8) and comparing molecules 1,3, 4, 6, 8, 9 and 12, it was observed that molecule 3 has the highest value of $^{2}\kappa$ descriptor. This may be due to increase in linear graph-likeness. Molecule 12 has lower value of this descriptor despite increase in linearity compare to molecule 9, however this decrease may be attributed to the addition of O(sp^3) with lower covalent radius(0.74) compare to C(sp^3) with a higher covalent radius (0.77). Another example was observed in molecule 15 and 17 whose $^{2}\kappa$ descriptor value are similar and higher when compare to molecule 18, this may be attributed to the presence of additional N(sp^2) in the pyrazine ring present in molecule 18 that imparted further reduction in covalent radii compare to the pyridine ring in molecules 15 and 17. Descriptors values for the molecules that constitute the training set can be found in Table S1 in the supplementary file. For the QSAR model Kier 2 has absolute mean effect (MF) value of 0.388 which is greater than that of AATS4i but lesser than the remaining two descriptors. This indicates Kier 2 has second least influence in determining the anticonvulsant activity of molecules used to construct the model in relation to other descriptor. Considering the regression coefficient in the equation and the sign of the MF, higher numerical value of this descriptor reduces ED_{50} value thereby increasing the anticonvulsant activity of the molecules.

The third descriptor in the model is VE2_D defined as the average coefficient sum of the last eigenvector from topological distance matrix (VE2_D) [13]. It is among the eigenvalue based descriptor (spectral indices) derived from graph theoretical matrix. specifically obtained from the coefficients of the eigenvector associated with the unique negative eigenvalue of the distance matrix which were used as - local vertex invariants (LOVIs), able to provide discrimination among graph vertices; higher values correspond to vertices of lower degree, those farther from the center or from a vertex of high degree. This makes them qualify as index of branching with lower value corresponding to increase branching and higher value corresponding to less branching as evident in the substitution of H atom on the phenyl ring for series of N,N-dimethyl-α-bromo-phenetylamines with bulkier group or more electron with drawing group like CH3, F, Br est.[13] Based on the sum of these LOVIs, the VED indices were proposed as molecular descriptors [13]:

VE2_D = $\frac{VE1\_D}{A}$ and VE_D1 = $\sum_{i=1}^{A} l_i A$

Where A is the number of molecular graph vertices and $l_i A$ are the coefficients (i.e., loadings) of the eigenvector associated with the largest negative eigenvalue (i.e. Ath eigenvalue of the decreasing eigenvalue sequence). Using training set only molecule with less branching on β-carbon of molecule and molecule with bulkier O(sp^3) and O(sp^2) in their skeleton had higher value of this descriptors. From Table 5, VE2_D has the highest absolute mean effect (MF) value of 0.959 indicating that it has the highest influence in determining the anticonvulsant activity of molecules used to construct the model in relation to other descriptor. Considering the regression vector and the sign of the MF, higher numerical value of this descriptor reduces ED_{50} value thereby increasing the anticonvulsant activity of the molecules.

Radial distribution function descriptor (RDF) was the last descriptor in the model. it’s a 3D structure descriptors based on the distance distribution in the geometrical representation of molecules [13]. They can provide information about steric hindrance or structure/activity properties of a molecule, distribution of interatomic distances in the entire molecule, bond distances, ring types, planar and nonplanar systems, and atom types depending on the property of the molecule or atom included in the function as the weighting scheme [13]. Formally, the radial distribution function of an ensemble of “A” atoms can be interpreted as the probability distribution of finding an atom in a spherical volume of radius R and the general form of the radial distribution function is represented by the equation below

$$g(R) = f \sum_{i=1}^{A-1} \sum_{j=i+1}^{A} w_i w_j e^{-\beta(R - r_{ij})^2}$$

Where f is a scaling factor, w characteristic atomic properties of the atoms i and j, r_{ij} the interatomic distance between the ith and jth atom, and A the number of atoms. The exponential term contains the distance r_{ij} between the atoms i and j and the smoothing parameterβ, which defines the probability distribution of the individual interatomic distances and can be interpreted as a temperature factor that defines the movement of atoms [13]. The function plots shape (*i.e*., distance between two atoms) on the x-axis, the respective property coefficient on the y-axis thereby separating geometry from property distribution. With $w_i w_j$ =1 this function is a representation of the overall shape of the molecule based on the frequencies of all atom pair distances within each radial distance step [50].

A typical RDF descriptor is denoted by RDFSω where S is the interatomic distance which ranges $1.0 \leq s \leq 15.5$ in units of 0.5 and ω is the weights which could be designated unweighted (u) or any measurable molecular property. Specifically, RDF50s define as the radial distribution function at 5.0 interatomic distance weighted by relative I-state (electronic inductive effect) [21] is the fourth descriptor in the selected QSAR model. This suggest the occurrence of a linear relationship between anticonvulsant activity and the 3D molecular distribution of relative inductive effect of atoms or group of atoms in the molecules calculated at radius of 5.0 Å from the geometrical centers of each molecule. The presence of electronegative atom in a chain of carbon atom withdraw electron towards itself causing positive charge to be relayed among the other atoms in the chain. This electron-withdrawing inductive effect is also called negative (-) I-effect. However some groups, such are alkyl group, are less electron-withdrawing than hydrogen and are therefore considered as electron releasing and tend to give electron whenever they are attached as a substituent, leading to induction effect known as positive (+) I-effect. Generally, relative inductive effects have been experimentally measured with reference to hydrogen, in decreasing order of -I effect or increasing order of +I effect, as follows:

–NH_3^+ > –NO_2 > –SO_2R > –CN > –SO_3H > –CHO > –CO > –COOH > –COCl> –$CONH_2$ > –F > –Cl > –Br > –I > –OR > -OH > –NH_2 > –C_6H_5 > –CH=CH_2 > –H

The strength of inductive effect is also dependent on the distance between the substituent group and the main group that react; the greater the distance, the weaker the effect. Partial charges on the atoms in a molecule influences RDF descriptor greatly leading to descriptor with either characteristic positive/negative peak distribution [45]. Using the training set molecule (see Table 1 and 8), higher value of RDF for molecule 75 may be attributed to additional carbonyl group on its chain. In the model RDFs had the second highest absolute mean effect MF value of 0.512, indicating that it has the highest influence in determining the anticonvulsant activity of molecules used to construct the model in relation to other descriptor. The negative sign of its mean effect value indicted that higher numerical value of this descriptor increases ED$_{50}$ value, thereby reducing the anticonvulsant activity of the molecules.

3. Interpretation of descriptors for Neurotoxicity value

The selected model for neurotoxicity studies is Model 4 (see Table 4) and it contained the following descriptors TIC5, nRing, VE1_D and RDF60i. The mean effect MF values for the descriptor incorporated into the model are presented in Table 6. The first descriptor in the model is TIC5 define as total information content index of neighborhood symmetry of order 5 (13). It is a member of topological information indices of a graph based on neighbor degree and edge multiplicity. This family of descriptor is calculated from a hydrogen included molecular graph by partitioning graph vertices into equivalence classes. Two vertices v_i and v_j are said to be equivalent if they belong to the same chemical element, have same vertex degree (same neighbor), mth order path starting from vi (^mP_i) correspond to mth order path starting from v_j (^mP_j) and have same conventional bond order of the edges in the path. Specifically, total information content of order m descriptors is obtained by multiplying the mth order neighborhood information content (IC$_m$) with the number of graph vertices A where IC$_m$ is obtained from Shannon's enthropy:

$$IC_m = -\sum_{g=1}^{G} \frac{A_g}{A} \cdot \log_2 \frac{A_g}{A} \quad \text{and} \quad TIC_m = A \cdot IC_m$$

Where the summation goes over the G equivalence classes, A_g is the cardinality of the gth equivalence class. This class of descriptor is interpreted as a measure of structural complexity per vertex [13]. It was observed that increase in the linearity

Table 8: Model 3 descriptors values

S.No	pED50	AATS4i	Kier2	VE2_D	RDF50s
1	4.343	0.529	-0.250	-0.641	-1.231
3	4.354	0.472	0.810	-0.574	-0.723
4	4.499	1.162	0.074	0.439	0.398
6	3.846	1.655	-0.692	-0.908	-0.543
8	4.029	0.814	-0.814	-0.716	-1.214
9	4.115	0.284	0.263	-0.075	-0.381
12	3.992	0.626	-0.281	-0.534	-0.292
13	3.383	-0.318	-0.118	0.103	0.924
15	3.624	-2.471	0.301	1.664	1.417
17	3.852	-1.142	0.301	0.724	1.270
18	3.493	-0.367	-0.027	-0.078	0.758
21	3.259	-1.005	-0.027	-0.078	0.809
22	3.567	-0.726	-0.602	0.091	-0.887
23	3.774	-0.786	-1.753	2.576	-0.221
25	4.036	-1.481	0.951	1.215	0.342
27	4.411	-1.025	0.951	1.215	0.425
28	4.278	-0.721	0.951	1.215	0.499
29	4.168	1.675	-1.075	0.247	-0.566
30	3.680	0.353	-0.250	-0.591	-0.443
31	3.944	0.266	-0.118	0.103	-0.092
32	3.603	-0.189	-0.250	-1.049	-1.162
34	3.996	1.057	-1.335	-0.331	-1.250
35	3.793	0.074	-0.814	-0.716	-1.005
36	3.439	-0.020	-0.765	-0.747	-0.481
41	3.743	-1.326	0.378	0.665	0.410
43	3.866	-0.517	0.810	-0.063	0.860
46	4.140	2.435	-0.500	0.203	1.962
47	4.039	1.055	-0.692	-0.908	-1.102
48	4.292	2.557	-0.389	-0.466	1.443
49	4.080	2.726	-0.580	-1.068	1.651
51	4.036	1.710	-0.145	-0.606	0.045
53	4.346	0.018	2.025	-0.739	0.941
54	4.460	-0.605	3.102	-0.490	1.087
55	4.804	-0.582	3.102	-0.490	0.348
57	3.533	0.165	-0.814	-0.716	-0.922
59	4.542	-0.179	-1.335	2.871	-1.388
61	4.606	-0.061	-0.194	2.328	-0.616
62	4.421	-0.260	-1.335	2.401	-1.458
63	4.275	-0.936	-0.814	2.164	-1.463
66	4.034	0.086	-1.373	1.752	-1.651
67	4.234	-0.361	1.906	-0.740	0.935
68	4.073	0.082	-0.027	-0.093	0.520
69	4.096	-0.414	0.885	-0.492	0.817
70	4.085	-1.316	1.400	-0.565	0.278
71	3.756	-0.880	-0.027	-0.767	-0.445
73	3.593	-0.272	-1.409	0.336	-0.612
75	3.708	-0.631	2.485	-0.776	3.957
77	3.169	-0.607	-0.500	-0.720	-0127
78	3.821	-0.332	1.454	-0.391	2.070
80	3.733	0.567	-1.278	0.032	0.282

Note: the serial number correspond the position occupied by the training set data in Table 1 of molecules increases the value of this descriptor e.g. comparing molecule 3 with higher value of this descriptor to molecule 1 (see Table 1 and 9). TIC_5 has mean effect MF value of -12.4 which is the highest absolute MF value indicating that it has the highest influence determining the neurotoxicity of the molecules in relation to other descriptor. The negative value of MF shows that higher numerical value of this descriptor, decreases pTD_{50} value thereby increasing the TD_{50} value indicating higher value of this descriptor reduces the neurotoxicity of the molecules. The values of this descriptor can be increased by increasing the number of atoms that are 4-bond away from a given vertex.

The second descriptor in the model is nRings defined as the number of ring present in a molecule [21]. It was observed that molecule with the same type and number of ring have the same value for this descriptor e.g. molecule 1, 2, 3 est (see Table 1 and 9). However molecule 18 has higher value of this descriptor compare to molecule 17 with the same number of ring. This observation may be attributed to the presence of additional $N(sp^2)$ on molecule 18 which reduces the size of the ring. nRings has mean effect MF value of 11.41 which is the second highest absolute MF value indicating that it has second highest influence determining the neurotoxicity of the molecules in relation to other descriptor. The positive value of MF shows that higher numerical value of this descriptor, increases pTD_{50} value thereby decreasing the TD_{50} value indicating higher value of this descriptor increases the neurotoxicity of the molecules.

The third descriptor in the model is VE1_D defined as the coefficient sum of the last eigenvector from topological distance matrix [13]. It is among the eigenvalue based descriptor (spectral indices) derived from graph theoretical matrix. specifically obtained from the coefficients of the eigenvector associated with the unique negative eigenvalue of the distance matrix which were used as - local vertex invariants (LOVIs), able to provide discrimination among graph vertices; higher values correspond to vertices of lower degree, those farther from the center or from a vertex of high degree. This makes them qualify as index of branching with lower value corresponding to increase branching and higher value corresponding to less branching as evident in the substitution of H atom on the phenyl ring for series of N,N-dimethyl-α-bromo-phenetylamines with bulkier group or more electron with drawing group like CH_3, F, Br est. Based on the sum of these LOVIs, the VED indices were proposed as molecular descriptors [13]:

VE_D1 = $\sum_{i=1}^{A} l_i A$

Where A is the number of molecular graph vertices and $l_i A$ are the coefficients (i.e., loadings) of the eigenvector associated with the largest negative eigenvalue (i.e. A^{th} eigenvalue of the decreasing eigenvalue sequence). Using training set data it was observed that increasing the molecular weight of the compound either by inclusion of additional $-CH_2-$ or heteroatoms into the chain and increased branching increases the value of this descriptor e.g. the order of increasing VE1_D value for the first four molecule was molecule 1 < 2 < 3 < molecule 4. Similar for molecule 45 < 46 < 47(see Table 1 and 9). VE1_D has mean effect (MF) value of -5.149 which is the least absolute MF value indicating that it has the least influence in determining the anticonvulsant activity of molecules used to construct the model in relation to other descriptor. The negative sign of the MF shows that higher numerical value of this descriptor reduces pTD_{50} i.e. increasing the TD_{50} indicating that higher values of this descriptor reduces the neurotoxicity of the molecules.

Table 9: Model 4 descriptors values

S.No	pTD50	TIC_5	nRing	VE1_D	RDF60i
1	3.644	-0.208	1	-0.746	-0.266
2	3.736	0.264	1	-0.658	0.274
3	3.815	0.806	1	-0.601	0.988
4	3.755	0.024	1	0.774	0.384
6	3.480	-0.149	1	-0.602	-0.249
8	3.360	-0.729	1	-1.019	-0.943
11	3.469	-0.729	1	-0.881	-0.501
12	3.453	-0.167	1	-0.164	0.547
14	3.593	-0.219	2	0.480	0.169
17	3.180	-0.403	2	0.181	-0.532
18	3.692	1.158	2	2.710	1.241
19	3.692	1.099	3	1.382	-0.253
20	3.351	-0.161	3	0.277	-0.303
22	3.493	-0.139	3	0.041	-0.960
24	3.001	0.561	2	-0.413	1.397
25	3.183	-0.593	2	0.086	-0.737
27	3.247	-0.781	2	0.086	-0.669
29	3.001	-0.920	2	0.964	-0.214
30	3.548	0.699	2	1.923	1.721
33	3.495	0.050	2	-0.164	0.745
34	3.495	-0.615	1	0.166	-0.341
36	3.694	-0.331	1	0.481	-0.822
37	3.191	0.327	1	-0.582	0.923
41	3.343	-0.615	1	0.166	-0.334
42	3.288	-1.003	1	-0.879	-0.644
44	3.571	0.305	1	0.947	-0.121
45	3.571	0.523	2	1.775	0.577
46	3.482	0.168	2	-0.011	0.932
47	3.744	-0.874	1	0.296	-1.817
48	3.512	-0.729	1	-1.020	-0.977
49	3.682	-0.290	1	-0.433	-0.751
50	3.482	-0.165	1	-1.160	-0.874
55	4.573	2.923	2	-0.117	1.043
57	3.064	-0.933	2	-0.880	-0.371
58	3.497	0.701	1	1.383	-0.089
61	3.643	-1.418	1	1.694	-1.401
62	3.986	0.024	1	-0.254	-0.604
63	3.590	0.958	1	-0.463	0.816
66	3.667	0.561	1	-0.739	0.526
67	3.596	1.067	1	0.023	0.891
69	3.367	-1.819	2	1.694	-1.586
70	3.594	1.158	1	-0.477	0.436
71	2.856	-0.792	2	-0.676	0.979
72	3.157	0.246	2	-1.279	0.044
74	3.744	1.503	2	1.924	1.753
76	3.003	0.089	2	-1.279	0.301
77	3.127	-0.213	2	0.181	-0.169
78	3.554	1.820	2	-0.193	2.514
79	3.336	0.187	2	-0.883	0.011
80	3.266	-0.108	2	0.023	-0.318

the serial number correspond the position occupied by the training set data in Table 1

The last descriptor in the model is RDF 60i which is radial distribution function descriptor at 6.0 interatomic distance weighted by first ionization potential [21]. This suggest the occurrence of a linear relationship between neurotoxicity values and the 3D molecular distribution of first ionization potential of atoms in the molecules calculated at radius of 6.0 Å from the geometrical centers of each molecule. Using the training set molecule (see Table 1 and 9), molecules containing one phenyl ring to which atoms like F, CL is directly attached has lower value of this descriptor e.g. molecule 76 had RDF60i value > that of molecule 77 > that of molecule 79. Therefore, addition of atoms with reduced first ionization potential gives lower value of this descriptor. Ionization potential is the amount of energy required to remove the most loosely bond electron, valence electron of an isolated gaseous atom to form a cation [44]. RDF60i has mean effect (MF) value of 7.177 which is the third absolute MF value indicating that it had the third influence in determining the anticonvulsant activity of molecules used to construct the model in relation to other descriptor. The positive sign of the MF showed that higher numerical value of this descriptor increases pTD_{50} i.e. reducing the TD_{50} indicating that higher values of this descriptor increases the neurotoxicity of the molecules.

4. Conclusion

The anticonvulsant activity and neurotoxicity value of N-benzylacetamide and 3-(phenylamino)propanamide derivatives had been quantitatively analyzed in terms of molecular descriptors. The statistically validated QSAR models provided rationales to explain the anticonvulsant and neurotoxicity activity of these derivatives. The descriptors identified through GFA analysis have highlighted the role of the average autocorrelation of topological structure lag 4 weighted by first ionization potential (AATS4i), second order shape attribute (Kier 2), average coefficient sum of the last Eigen vector from topologicl distance matrix (VED-2) and radial distribution function at 5.0 inter-atomic distance weighted by relative I-state in affecting the anticonvulsant activity of the studied group of compound. For a compound belonging to the group of the studied compound to more potent higher values of AATS4i, Kier 2 and VED_2 and lower value of RDF50s are conducive. Also, total information content index of the neighbourhood symmetry of 5-order (TIC5), number of ring (nRing), coefficient sum of the last Eigen vector from topologicl distance matrix (VED-1) and radial distribution function at 6.0 inter-atomic distance weighted by first ionization potential were descriptor found to influence the neurotoxicity of this group of compound. And for any member of the group to be less toxic, lower value of TIC5, VE1_D and higher value of nRing and RDF60i is required. Applicability domain analysis revealed that the suggested models had acceptable predictability with the entire molecules that constituted the training dataset remaining within the applicability domain of the proposed models and the entire test dataset were evaluated correctly except for one molecule (see Figure 2).

References

[1] A.C. Errington, T. Stohr, G. Lees, "Voltage Gated ion Channels: Target or Anticonvulsant Drugs," Current Topics in Medicinal Chemistry, 5, pp. 15-30, 2005.

[2] D. Schmidt, W. Loscher, Epilepsia, 46, pp. 858 – 877, 2005.

[3] Guerrini, R. (2006). Epilepsy in children. Seminar at Department of Child Neurology and Psychiatry, University of Pisa and IRCCS Foundazione Stella Maris, 367:499-524.

[4] R. Thirumurugan, D. Sriram, A. Saxena, J. Stables, P. Yogeeswari, "2,4-Dimethoxyphenylsemicarbazones with anticonvulsant activity against three animal models of seizures: synthesis and pharmacological evaluation," Bioorganic and Medicinal Chemistry, 14, pp. 3106 – 3112, 2006.

[5] A.V. Shindikar, F. Khan, C.L. Viswanathan, "Design, synthesis and in vivo anticonvulsant screening in mice of Novel phenylacetamides," European Journal of Medicinal Chemistry, 41, pp. 786 – 792, 2006.

[6] G. Avanzini, G. Franceschetti, "Prospects for novel antiepileptic drugs,".Drugs, 4(7), pp. 805-814, 2003.

[7] G. Kramer, "Epilepsy in the elderly: some clinical and pharmacotherapeutic aspects," Epilepsia, 42, pp. 55-59, 2001

[8] Kubinyi, H. *QSAR: Hansch analysis and related approaches*, VCH Publishers: Weinheim, New York, Basel, Cambridge, Tokyo, 1993.

[9] A. Speck-Planche, M.N. Cordeiro, "Computer-aided drug design methodologies toward the design of anti-hepatitis C agents," Current Topics in Medicinal Chemistry, *12*, pp. 802-813, 2012.

[10] P. Riera-Fernandez, R. Martin-Romalde, F.J. Prado-Prado, M. Escobar, C.R. Munteanu, R. Concu, A. Duardo-Sanchez, H. Gonzalez-Diaz, "From QSAR models of drugs to complex networks: state-of-art review and introduction of new Markovspectral moments indices, Current Topics in Medicinal Chemistry, 12, pp. 927- 960, 2012.

[11] A. Speck-Planche, M.T. Scotti, V. de Paulo-Emerenciano, "Current pharmaceutical design of antituberculosis drugs: future perspectives," Current Pharmaceutical Discovery Journal, 16, pp. 2656-2665, 2010.

[12] H. Gonzalez-Diaz, "QSAR and complex networks in pharmaceutical design, microbiology, parasitology, toxicology, cancer, and neuro sciences," Current Pharmaceutical Descovery Journal, 16, pp. 2598-2600, 2010.

[13] Todeschini, R.; Consonni, V. *Molecular Descriptors for Chemoin formatics*, WILEY-VCH Verlag GmbH & Co. KGaA: Weinheim,

[14] E. Pourbasheer, S. Riahi, M.R. Ganjali, P. Norouzi, "Application of genetic algorithm-support vector machine (GA–SVM) for prediction of BK-channels activity," European Journal of Medicinal Chemistry, 44, pp. 5023–5028, 2009.

[15] Idris, A.Y.(2008). Synthesis and anticonvulsant studies of 3-anilinopropanamides and their n-benzyl derivatives. Unpublish thesis at Department of Pharmaceutical and Medicinal chemistry, Faculty of Pharmaceutical Sciences Ahmadu Bello University, Zaria.

[16] A.M. King, Synthesis and pharmacological evaluation of primary amino acid derivatives (PAADs): novel neurological agents for the treatment of epilepsy and neuropathic pain. Thesis submitted to Division of Medicinal Chemistry and Natural Products University of North Carolina at Chapel Hill, 2010. Retrieved 5th May 2016from https://cdr.lib.unc.edu/...

[17] B.K. Sharma, M. Shekhawat, P. Singh, "A QSAR study on pyrazole and triazole derivatives as selective canine COX-2 inhibitors," International Journal of Research in Ayurveda and Pharmacy, 2(1), pp.186-197, 2011

[18] M.N. Noolvi, H.M. Patel, V. Bhardwaj, "2D qsar studies on a series of 4-anilino quinazoline derivatives as tyrosine kinase (egfr) inhibitor: an approach to design anticancer agents," Digest Journal of Nonmaterial and Biostructures, 5(2), pp. 387 – 401, 2010.

[19] A.Z. Dudek, T. Arodzb, T, Gálvezc, "Computational methods in developing quantitative structure-activity relationships (QSAR): a review,"Combinatorial Chemistry and High Throughput Screening, 9, pp. 213-228, 2009.

[20] Y. Shao, L.F. Molnar, Y. Jung, J. Kussmann, C. Ochsenfeld, S.T. Brown, A.T.B. Gilbert, L.V. Slipchenko, S.V. Levchenko, D.P. O'Neill, R.A. DiStasio Jr., R.C. Lochan, T. Wang, G.J.O. Beran, N.A. Besley, J.M. Herbert, C.Y. Lin, T. Van Voorhis, S.H. Chien, A. Sodt, R.P. Steele, V.A. Rassolov, P.E. Maslen, P.P. Korambath, R.D. Adamson, B. Austin, J. Baker, E.F.C. Byrd, H. Dachsel, R.J. Doerksen, A. Dreuw, B.D. Dunietz, A.D. Dutoi, T.R. Furlani, S.R. Gwaltney, A. Heyden, S. Hirata, C-P. Hsu, G. Kedziora, R.Z. Khalliulin, P. Klunzinger, A.M. Lee, M.S. Lee, W.Z. Liang, I. Lotan, N. Nair, B. Peters, E.I. Proynov, P.A. Pieniazek, Y.M. Rhee, J. Ritchie, E. Rosta, C.D. Sherrill, A.C. Simmonett, J.E. Subotnik, H.L. Woodcock III, W. Zhang, A.T. Bell, A.K. Chakraborty, D.M. Chipman, F.J. Keil, A.Warshel, W.J. Hehre, H.F. Schaefer, J. Kong, A.I. Krylov, P.M.W. Gill and M. Head-Gordon, "Advances in methods and algorithms in modern quantum chemistry program package,"Phys. Chem. Chem. Phys., **8** , pp. 3172, 2006.

[21] C.w. Yap, "PaDEL-Descriptor: An open source software to calculate molecular descriptors and fingerprints, " Journal of Computational Chemistry, **32** (7), pp. 1466-1474, 2011.

[22] P. Singh, "Quantitative structure-activity relationship study of substituted-[1,2,4] oxadiazoles as S1P$_1$ agonists. Journal of current chemical and pharmaceutical sciences, "3(1), pp. 64-67, 2013.

[23] P. Ambure, R.B. Aher, A. Gajewicz, W.R.K. Pyzyt, "Nano BRIDGES" software: open acess tools to perform QSAR and nano-QSAR modeling. Chemometrics and Intelligent Laboratory Systems, 147(15), pp.1-13, 2015.

[24] P. Gramatica, "Principles of QSAR models validation: internal and external," QSAR and Combnatorial Sciences, 26, pp.694, 2007.

[25] Tropsha and A. Golbraith, "Predictive quantitative structure and activities relationship modelling data preparation and general modelling workflow: In handbook of chemoinformatics algorithms, 2010,"1st Edition, mathematical and computational biology series, chapman and Hall/CRC books, Edited by N.E.Britton, Xihong lin, Hersbel M. Sfaer, Mona Singh and Anna Tramonato. Page 173-232.

[26] R.K.H. Galvao, M.C.U. Araujo, G.E. José, M.J.C.Pontes, E.C. Silva, T.C.B. Saldanha, "A method for calibration and validation subset partitioning," Talanta, 67, pp. 736-740, 2005.

[27] K. Varmuza, P. Filzmoser, "Introduction to multivariate statistical analysis in chemometrics" Boca Raton, FL, USA: CRC Press, 2009.

[28] D. Rogers, A.J. Hopfinger, "Application of genetic function approximation to quantitative structure-activity relationships and quantitative structure-property relationships," Journal of Chemical Informatics and Computer Science, 34, pp. 854-866, 1994.

[29] J.G. Topliss, R.J. Costello "Chance correlations in structure– activity studies using multiple regression analysis," Journal of Medicinal Chemistry, 15, pp. 1066–1068, 1972.

[30]A . Tropsha, P. Gramatica, V.K. Gombar "The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR models," QSAR and Combinatorial Science, 22, pp. 69–77, 2003.

[31] P.D. Allison, "Multiple regression: a primer," (1999) Pine Forge Press, London

[32] A.B. David, K. Edwin, E.W. Roy(2004).Regression diagnostics: indentifying influential data and source of collinearity. 1st Edition, Wiley-interscience, page 85-93

[33] A.Golbraikh, A.Tropsha, "Beware of q2!,". Journal Molecular Graphics and Modeling, 20, pp. 269–276, 2002.

[34] H. Wold, M, Sjoestroem, L. Eriksson, "PLS-regression: a basic tool of chemometrics," Chemom Intelligent Laboratory System, 58, pp. 109-130, 2001.

[35] A. Habibi-Yangjeh, M. Danandeh-Jenagharad, "Application of a genetic algorithm and an artificial neural network for global prediction of the toxicity of phenols to Tetrahymena pyriformis," Monatsh Chemistry, 140, pp. 1279–1288, 2009.

[36] N. Minovski, S. Zuperl, V. Drgan, M. Novi, "Assessment of applicability domain for multivariate counter-propagation artificial neural network predictive models by minimum Euclidean distance space analysis: A case study," Analytica Chimica Acta, 759, pp. 28–42, 2013.

[37] T.I. Netzeva, A.P. Worth, T. Aldenberg, R. Benigni, M.T.D Cronin, P. Gramatica, J.S. Jaworska, S. Kahn, S. Klopman, C.A. Marchant, G. Myatt, N. Nikolova-Jeliazkova, G.Y. Patlewicz, R. Perkins, D.W. Roberts, T.W. Schultz, D.T. Stanton, J.J.M. Sandt, W. Tong, G. Veith, C. Yang, "Current status of methods for defining the applicability domain of (quantitative) structure–activity relationships," Altern. Lab. Anim, 33, pp. 155–173, 2005.

[38] P Gramatical, S. Cassani,, P.P Roy, S. Kovarich, C.W. Yap, E. Papa, "QSAR modeling is not 'push as button and find a correlation': a case study of toxicity of (benzo-)triazoles on algae, " *Molecular Informatics*, 31, pp. 817-835, 2012.

[39] M. Jaiswal, P.V. Khadikar, A. Scozzafava, C.T. Supuran, "Carbonic anhydrase inhibitors: the first QSAR study on inhibition of tumor-associated isoenzyme IX with aromatic and heterocyclic sulfonamides," Bioorganic and Medicinal Chemistry Letters, 14, pp. 3283–3290, 2004.

[40] A. Tropsha, "Best practices for QSAR model development, validation and exploitation," Molecular Informatics, 29, pp. 476–488, 2010.

[41] K .Roy, C. Patrim, M. Indrani, P.K. Ojha, K Supratik, R.N. Das, "Some Case Studies on Application of ''$r_m{}^2$ '' metrics for judging quality of quantitative structure–activity relationship predictions: Emphasis on scaling of response data,"Journal of Computational Chemistry, 34, pp. 1071–1082, 2013

[42] L.H. Hall, and L.B. Kier, in Reviews in Computational Chemistry,Vol. 2 Ed. By K.B. Lipkowitz and D.B. Boyd

(VCH Publishers, New York, 1991; pp. 367–422

[43] L.M. Stock, "The origin of the inductive effect," Journal of Chemical Education , 49 (6), pp. 400, 1972.

[44] M.C. Hemmer, (2007). Radial distribution functions in computational chemistry-theory and application. PhD dissertation submitted to Den Naturwissenschaftlichen Fakultäten der Friendlish-Alexander-Universität Erlangen-Nürnberg, retrieved from www.google.com on 12th may 2016.

Author Profile

< >

First Author/ corresponding author:
Name: Adedirin Oluwaseye
Education: M.Sc (Physical chemistry)
Ocupation: Research Fellow II at Sheda Science and Technology complex, FCT, Nigeria. Also, a student at Ahmadu Bello University, Zaria.

Second Author:
Name: Adamu Uzairu
Education: PhD (Physical and Theoretical Chemistry
Ocupation: Professor of Chemistry at Ahmadu Bello university Zaria**.**

Third Author
Name: Shallangwa Gideon Adamu
Education: PhD (Physical/Inorganc chemistry)
Ocupation: Lecturer (Chemistry) at Ahmadu Bello University Zaria**.**

Fourth Author
Name: Abechi Stephen Eyije
Education: PhD (Physical and Chemistry)
Ocupation: Lecturer (Chemistry) at Ahmadu Bello University Zaria**.**