# Optimal Division of Molecules Into Training And Test Sets With A New Tool To Predict Pharmacophore In 3D-QSAR

**Tuğba Alp Tokat** ( tugbaalp81@gmail.com )
Erciyes University
**Burçin Türkmenoğlu**
Erzincan Binali Yıldırım University
**Yahya Güzel**
Erciyes University

**Research Article**

**Keywords:** 3D-QSAR, Flavonoid, Splitting, Vector Fingerprint Function, Klopman Index.

**Posted Date:** November 1st, 2021

**DOI:** https://doi.org/10.21203/rs.3.rs-949300/v1

## Abstract

According to the descriptors in the pharmacophore model, dividing molecules into training and test sets serves to create a good model. It is difficult to track the Local Reactive Descriptor (LRD) effect of the pharmacophore at each interaction point in the 3D metric system. A subset of clusters of atoms can correspond to all or part of the pharmacophore structure. In this study, the multidimensional system of the subset was reduced to a one-dimensional index and the Vector Fingerprint Functions (VFF) of the molecules were created. Models were established by dividing molecules with close and similar VFFs into training and test sets. Sub-clusters were examined for all molecules by applying the Genetic Algorithm (GA). The model was predicted using the Leave One Out-Cross Validation (LOO-CV) method and verified with an external test set. The statistical results of the model obtained according to the division in the new method we developed ($Q^2$ = 0.604 and $R^2$ = 0.760 for training-80 and external test-20 sets, respectively) were compared with random and manual division results.


## Introduction

In 3D-QSAR studies, it is necessary to group local descriptors derived from the geometric and electronic values of the atoms in the 3D metric system into common clusters. According to the similarities of descriptors within a set, molecules can be divided into training and test sets in the most appropriate way. However, the similarity must be in all clusters in the VFF's index, not in one cluster. Similarity and dissimilarity in VFF serve this. Similar ones can be divided into training and test sets at the rates of 4/5 and 1/5. The accumulation of molecules with similar descriptors only in the training or test set prevents a good model to be found. Therefore, by dividing such molecules in appropriate ratios into two sets, a model is created in one set and validated in the other. The creation and validation of the model is very meaningful by placing similar molecules in two different sets. Optimum division of molecules with similar activities and descriptors into training and test sets increases the reliability of the model.

To create clustering, the molecules must be arranged to interact with the receptor in the same orientation and geometry. For this reason, the compounds are superimposed 3D compatible against the receptor in a common chemical structure. Atomic stacks in similar coordinates can interact with a corresponding common point of the receptor. Some, but not all, of the clusters formed may play a role in the activity of molecules. To determine the sub-cluster that constitutes the pharmacophore structure, different sub-clusters formed by combination should be considered. One of the subgroups may have a 3D structure that represents all or part of the pharmacophore. Since such a sub-cluster in the 3D coordinate system creates a multidimensional vector field, the resulting pharmacophore interactions cannot be directly observed. To easily analyze the relationship between the descriptor and the activity, the multidimensional systems of the sub-clusters can be reduced to a one-dimensional vector system as an index. The ordering of the sub-cluster as an index can reveal the similarity in the change in the activity of molecules with similar descriptors in the same index. The activity similarities of two molecules with similar descriptors throughout the elements of the sub-cluster can be clearly observed by imitating each other very well along the one-dimensional index. Accordingly, molecules with close activity values can be followed clearly according to their one-dimensional functional changes. Thus, one of the two molecules with the same descriptors can be divided into the training set and the other into the test set, considering both activity similarity and descriptive similarity.

The purpose of reducing the multi-dimensional vector space to one-dimensional index; a) Helping to obtain a better model from spatially complex data and avoiding loading more information, b) Viewing data in a simpler and compact size rather than a high-dimensional area, c) In this way, the main variation types are If it is expressed less than the number (making estimates from these numbers if possible), d) It may be appropriate to carry the most important variations to the model instead of meaningless and unnecessary dimensions.

The "similarity principle" approach, in which similar molecules will likely exhibit similar (physicochemical and/or biological) properties, is often used in cheminformatics. In chemistry, properties close to the defining properties of the leading compounds are an important paradigm as a guide and determinant in the design and synthesis of new analogues [1]. Addressing similarities in clusters is widely used, such as the Dimensionality reduction procedure of Locally Linear

Embedding (LLE) [2]. The multidimensional input area has been transformed into a one-dimensional function and the concept of Dimensional Reduction (DR) has been applied [3]. Principal Component Analysis (PCA) is a standard algorithm commonly used to reduce size [4]. There is a Multi-Dimensional Scaling (MDS) technique such as PCA and more general [5]. There are also many different size reduction algorithms such as PCA, from classic Linear Algebra to non-linear techniques such as Kohonen Self-Organizing Maps (SOM) [6], MDS [7], Stochastic Embedding [8]. These techniques are more widely used to look for a common purpose of similarity and similarity matrices [9]. In previous 3D-QSAR studies, the optimum division in both sets was discussed by the similarities of the descriptors found in the sets in the vector space system rather than in the metric space [3]. In other studies, vector space has been found to have superior performance than metric space [10–11]. Independent methods of many different descriptors are used for molecular similarity, whereas there are other molecular similarity comparisons, including molecular graphic matching approaches [12]. The similarity coefficient of Tanimoto (Jaccard) is widely used to measure molecular similarity [13–14]. However, many other similarity/distance methods have been considered as vector spaces by researchers [15].The most common preference is to select the regulated variables after the sequence given by the descriptor variance of sub-clusters [16]. In particular, the Most Predictive Variable Method (MPVM) has been used successfully to select the optimum variables [17]. Clusters can have infinite dimensions in space, but cluster centers with total M-dimensions for the molecule being studied are discussed [18].To identify the best sub-cluster of molecular descriptors the vector space reduction is performed using the Truncated Singular Value Decomposition (TSVD) [19]. The more compatible the structure, descriptors, and activity values of the molecules, the better the model is built [20].

There are a variety of techniques that can produce training and test sets in QSAR reviews. The simplest of these, the first of the four techniques, is random or the second is manual [21–22]. According to this, molecules are placed in training and test sets using random computer-generated numbers between the numbers of the members of the data set, or by using the manual technique considering the structure and activity values [23]. The other third and fourth methods are automatic and rational algorithms that provide activity balancing. In the third method, in order to differentiate the data values in the examined molecules into sets; Automatic model generation such as Bayesian Neural Networks (BNN) [24], Relational Neural Networks (RelNNs) [25], Gauss Process (GP) [26–27] are used. In the fourth method for rational division of sets, SOM [28], D-optimal designs [29–32], sphere exclusion [32], Directional Sphere Exclusion (DISE) [33], Kennard-Stone (KS) algorithm [34–35], such as different rational algorithms are applied.

In this study, the experimental activity values were compared to the theoretical activity values of 100 flavonoid molecules taken from the literature [36], and the molecules were divided into training and test sets according to their similar VFFs. While the receptor parameter remains constant for each index during the formation of VFF, the electronic values of atoms in the same index can vary from molecule to molecule. The change in the activity of a molecule can be observed graphically in VFF. The increase or decrease in VFFs is because the electronic value of the atom of the molecule in that index is positive/negative and large/small. Since fingerprints VFFs show how molecule activity varies throughout the index, molecules that show similarity in all index elements are optimally distributed to training and test sets without clustering in the same set. By sharing similar activity changes of molecules between training and test sets, the pharmacophore model can be suggested more reliably.

# Vector Fingerprint Function (Vff)

VFF is characteristic of each molecule that reduces the vector space field of the sub-cluster responsible for activity from multiple dimensions to one dimension. As the name suggest, the activity of a molecule varies functionally in each of the sub-cluster elements listed as a one-dimensional index. Increasing or decreasing contributions of atoms are easily seen in interaction points formed throughout the index. At each of these points, while the parameter of the receptor is constant, the molecule's Local Reactivity Descriptors (LRDs) (large/small and positive/negative) contribute differently and lead to different fingerprints in each molecule. Therefore, VFF, which shows interaction at every point, is a good recognition tool that shows the behavior of molecules. In short, the changes of the activities in the interaction points are different for each molecule and

with adding up contributions of the activity at each point, a value close to the experimental activity shown by the molecule can be calculated. The change in activity calculated throughout the index is the fingerprint of the molecule. As can be seen from here, the biological or physicochemical properties of the molecules in the studied series can be easily determined by one-dimensional VFF value. The increase and decrease in the activity of molecules examined against VFF values can be monitored graphically. The effect of each interaction point for the proposed pharmacophore is clearly visible with VFF. Here, similarities and differences in VFF can be easily seen in the form of a molecule's fingerprint.

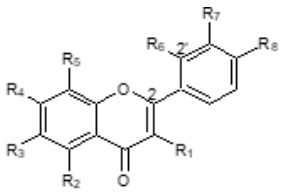
Since the activity change of molecules like each other at each point is functionally visible with VFF, it is possible to divide the molecules into training and test sets safely. VFFs of the molecules in the training set can be divided similarly to the molecules in the test sets. Thus, it may be possible to verify the model proposed in the training set with analogues in the test set. The purpose of preparing VFF is to create a good model by dividing the molecules into two sets in the best way according to the similarities of VFF without the need for larger molecule sets.

Here, for the first time, we will divide the molecules into training and test sets by tracking the activity change in the one-dimensional vector index. Accordingly, we will try to create an optimum model by preventing similar molecules from accumulating in one set, distributing the appropriate number between the two sets. We will discuss the division of the compounds with an automated and rational approach according to the activity similarities originating from VFF, which will be introduced to the literature for the first time. Since we do not have the chance to rewrite the algorithm of other rational or automatic division methods in the homemade Molecular Conformer Electron Topological (MCET) method, we will not be able to compare their performance with that of VFF. We will compare this separation method with only random and manual separation methods to show the differences, innovations, and developments.

## Principle And Method

In Table 1, 100 flavonoid molecule series given the skeletal structure and activities were taken from the literature and the pharmacophore structure of the activities were examined [36].

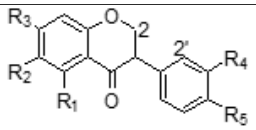
Table 1
Molecular skeletons and activity values of the investigated molecule series.



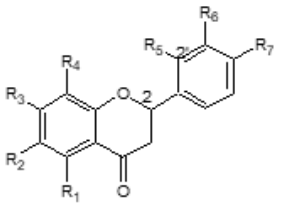

| Number | Nomenclature | R1 | R2 | R3 | R4 | R5 | R6 | R7 | R8 | pIC50 | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | | Obs. | Pred. |
| n01* | Flavone | H | H | H | H | H | H | H | H | 1.800 | 1.801 |
| n02 | 5-Hydroxyflavone | H | OH | H | H | H | H | H | H | 1.810 | 1.780 |
| n03 | 6-Hydroxyflavone | H | H | OH | H | H | H | H | H | 1.800 | 1.778 |
| n04 | 3'-Hydroxyflavone | H | H | H | H | H | H | OH | H | 1.730 | 1.776 |
| n05 | 4'-Hydroxyflavone | H | H | H | H | H | H | H | OH | 1.820 | 1.734 |
| n06 | 7-Hydroxyflavone | H | H | H | OH | H | H | H | H | 1.760 | 1.812 |
| n07 | 2',3'-Dihydroxyflavone | H | H | H | H | H | OH | OH | H | 1.720 | 1.787 |
| n08 | 2',4'-Dihydroxyflavone | H | H | H | H | H | OH | H | OH | 1.720 | 1.738 |
| n09 | 3',4'-Dihydroxyflavone | H | H | H | H | H | H | OH | OH | 1.700 | 1.799 |
| n10* | 2',3'-Dimethoxyflavone | H | H | H | H | H | OCH3 | OCH3 | H | 1.810 | 1.785 |
| n11 | 2',4'-Dimethoxyflavone | H | H | H | H | H | OCH3 | H | OCH3 | 1.690 | 1.757 |
| n12* | 4'-Hydroxy-3'-methoxyflavone | H | H | H | H | H | H | OCH3 | OH | 1.700 | 1.762 |
| n13 | 2',3-Dihydroxyflavone | OH | H | H | H | H | OH | H | H | 1.750 | 1.788 |
| n14 | 3,3'-Dihydroxyflavone | OH | H | H | H | H | H | OH | H | 1.710 | 1.790 |
| n15 | 3,4'-Dihydroxyflavone | OH | H | H | H | H | H | H | OH | 1.650 | 1.751 |
| n16 | 2',3-Dimethoxyflavone | OCH3 | H | H | H | H | OCH3 | H | H | 1.740 | 1.790 |
| n17 | 3,3'-Dimethoxyflavone | OCH3 | H | H | H | H | H | OCH3 | H | 1.710 | 1.805 |
| n18 | 3,4'-Dimethoxyflavone | OCH3 | H | H | H | H | H | H | OCH3 | 1.840 | 1.750 |
| n19 | 3-Hydroxy-2'-methoxyflavone | OH | H | H | H | H | OCH3 | H | H | 1.670 | 1.778 |
| n20* | 3-Hydroxy-4'-methoxyflavone | OH | H | H | H | H | H | H | OCH3 | 1.710 | 1.710 |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| n21 | 3,6-Dihydroxyflavone | OH | H | OH | H | H | H | H | H | 1.590 | 1.796 |
| n22* | 3,6-Dimethoxyflavone | OCH3 | H | OCH3 | H | H | H | H | H | 1.820 | 1.803 |
| n23 | 3-Hydroxy-6-methylflavone | OH | H | CH3 | H | H | H | H | H | 1.820 | 1.790 |
| n24 | 3-Hydroxy-6-methoxyflavone | OH | H | OCH3 | H | H | H | H | H | 1.880 | 1.782 |
| n25 | 3-Hydroxy-7-methoxyflavone | OH | H | H | OCH3 | H | H | H | H | 1.820 | 1.780 |
| n26 | 3’,5-Dihydroxyflavone | H | OH | H | H | H | H | OH | H | 1.800 | 1.783 |
| n27* | 4’,5-Dihydroxyflavone | H | OH | H | H | H | H | H | OH | 1.690 | 1.734 |
| n28 | 4’-Hydroxy-5-methoxyflavone | H | OCH3 | H | H | H | H | H | OH | 1.720 | 1.736 |
| n29* | 3’,5-Dimethoxyflavone | H | OCH3 | H | H | H | H | OCH3 | H | 1.800 | 1.767 |
| n30 | 5,7-Dihydroxyflavone | H | OH | H | OH | H | H | H | H | 1.850 | 1.788 |
| n31 | 5,7-Dimethoxyflavone | H | OCH3 | H | OCH3 | H | H | H | H | 1.830 | 1.759 |
| n32* | 5-Hydroxy-7-methoxyflavone | H | OH | H | OCH3 | H | H | H | H | 1.620 | 1.791 |
| n33* | 2’,6-Dihydroxyflavone | H | H | OH | H | H | OH | H | H | 1.730 | 1.765 |
| n34 | 3’,6-Dihydroxyflavone | H | H | OH | H | H | H | OH | H | 1.670 | 1.763 |
| n35* | 4’,6-Dihydroxyflavone | H | H | OH | H | H | H | H | OH | 1.800 | 1.724 |
| n36* | 2’,6-Dimethoxyflavone | H | H | OCH3 | H | H | OCH3 | H | H | 1.750 | 1.747 |
| n37* | 3’,6-Dimethoxyflavone | H | H | OCH3 | H | H | H | OCH3 | H | 1.660 | 1.758 |
| n38 | 4’,6-Dimethoxyflavone | H | H | OCH3 | H | H | H | H | OCH3 | 1.730 | 1.725 |
| n39 | 6-Hydroxy-2-methoxyflavone | H | H | OH | H | H | OCH3 | H | H | 1.800 | 1.742 |
| n40* | 6-Hydroxy-3’-methoxyflavone | H | H | OH | H | H | H | OCH3 | H | 1.770 | 1.756 |
| n41 | 6-Hydroxy-4’-methoxyflavone | H | H | OH | H | H | H | H | OCH3 | 1.790 | 1.715 |
| n42 | 4’-Hydroxy-6-methoxyflavone | H | H | OCH3 | H | H | H | H | OH | 1.770 | 1.713 |
| n43 | 6-Hydroxy-7-methoxyflavone | H | H | OH | OCH3 | H | H | H | H | 1.840 | 1.780 |
| n44 | 6,7- | H | H | OH | OH | H | H | H | H | 1.670 | 1.798 |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Dihydroxyflavone | | | | | | | | | | |
| n45 | 7,8-Dihydroxyflavone | H | H | H | OH | OH | H | H | H | 1.660 | 1.728 |
| n46 | 6,7-Dimethoxyflavone | H | H | OCH3 | OCH3 | H | H | H | H | 1.750 | 1.779 |
| n47 | 7,8-Dimethoxyflavone | H | H | H | OCH3 | OCH3 | H | H | H | 1.690 | 1.764 |
| n48 | 8-Hydroxy-7-methoxyflavone | H | H | H | OCH3 | OH | H | H | H | 1.770 | 1.676 |
| n49* | 2',7-Dihydroxyflavone | H | H | H | OH | H | OH | H | H | 1.730 | 1.783 |
| n50* | 3',7-Dihydroxyflavone | H | H | H | OH | H | H | OH | H | 1.760 | 1.799 |
| n51 | 4',7-Dihydroxyflavone | H | H | H | OH | H | H | H | OH | 1.800 | 1.756 |
| n52 | 2',7-Dimethoxyflavone | H | H | H | OCH3 | H | OCH3 | H | H | 1.760 | 1.773 |
| n53 | 3',7-Dimethoxyflavone | H | H | H | OCH3 | H | H | OCH3 | H | 1.770 | 1.783 |
| n54 | 4',7-Dimethoxyflavone | H | H | H | OCH3 | H | H | H | OCH3 | 1.820 | 1.764 |
| n55 | 7-Hydroxy-2-methoxyflavone | H | H | H | OH | H | OCH3 | H | H | 1.730 | 1.781 |



| Number | Nomenclature | R1 | R2 | R3 | R4 | R5 | pIC50 | |
|---|---|---|---|---|---|---|---|---|
| | | | | | | | Obs. | Pred. |
| n56 | 7-Hydroxyisoflavone | H | H | OH | H | H | 1.730 | 1.851 |
| n57* | 4’,7-Dihydroxyisoflavone (daidzein) | H | H | OH | H | OH | 1.840 | 1.785 |
| n58 | 7-Hydroxy-4’-methoxy isoflavone(formononetin) | H | H | OH | H | OCH3 | 1.770 | 1.764 |
| n59 | Formononetin-7-O-glucoside (ononin) | H | H | GLU | H | OCH3 | 1.830 | 1.827 |
| n60 | 4’,7-Di methoxyisoflavone | H | H | OCH3 | H | OCH3 | 1.810 | 1.811 |
| n61 | Daidzein-7-d-glucoside (daidzin) | H | H | GLU | H | OH | 1.770 | 1.754 |
| n62 | 4’,6,7-Trihydroxyisoflavone | H | OH | OH | H | OH | 1.630 | 1.813 |
| n63 | 4’,7-Dihydroxy-6-methoxyisoflavone (glycitein) | H | OCH3 | OH | H | OH | 1.820 | 1.805 |
| n64 | 4’,6,7-Trimethoxyisoflavone | H | OCH3 | OCH3 | H | OCH3 | 1.860 | 1.825 |
| n65 | 4’,5,7-Trihydroxyisoflavone (genistein) | OH | H | OH | H | OH | 1.760 | 1.798 |
| n66* | 5,7-Dihydroxy-4’-methoxyi soflavone (biochnin A) | OH | H | OH | H | OCH3 | 1.760 | 1.774 |
| n67* | 4’,5,7-Trimethoxyisoflavone | OCH3 | H | OCH3 | H | OCH3 | 1.740 | 1.845 |
| n68 | Biochanin A-7-d-glucoside (sissotrin) | OH | H | GLU | H | OCH3 | 1.860 | 1.849 |
| n69 | 4’,7-Dimethoxy-5-hydroxyisoflavone | OH | H | OCH3 | H | OCH3 | 1.870 | 1.823 |
| n70 | Genistein-7-d-glucoside (genistin) | OH | H | GLU | H | OH | 1.800 | 1.750 |
| n71 | 3’,4’,7-Trihydroxyisoflavone | H | H | OH | OH | OH | 1.780 | 1.674 |
| n72 | 6,7-Dimethoxyisoflavone | H | OCH3 | OCH3 | H | H | 1.770 | 1.895 |
| n73 | 3’,4’-Dimethoxy-7-hydroxyisoflavone | H | H | OH | OCH3 | OCH3 | 1.740 | 1.798 |
| n74 | 3’,4’,6,7-Tetramethoxyisoflavone | H | OCH3 | OCH3 | OCH3 | OCH3 | 1.720 | 1.730 |
| n75 | 6-Chloro-7-methylisoflavone | H | Cl | CH3 | H | H | 1.770 | 1.745 |



| Number | Nomenclature | R1 | R2 | R3 | R4 | R5 | R6 | R7 | pIC50 | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | Obs. | Pred. |
| n76* | 2′,6-Di hydroxyflavanone | H | OH | H | H | OH | H | H | 1.670 | 1.602 |
| n77* | 3′,4′,5,7-Tetrahydroxy flavanone (eriodictyol) | OH | H | OH | H | H | OH | OH | 1.170 | 1.237 |
| n78 | Flavanone | H | H | H | H | H | H | H | 1.390 | 1.471 |
| n79 | 2′-Hydroxyflavanone | H | H | H | H | OH | H | H | 1.670 | 1.599 |
| n80* | 4′-Hydroxyflavanone | H | H | H | H | H | H | OH | 1.450 | 1.420 |
| n81 | 6-Hydroxyflavanone | H | OH | H | H | H | H | H | 1.320 | 1.507 |
| n82 | 7-Hydroxyflavanone | H | H | OH | H | H | H | H | 1.450 | 1.606 |
| n83 | 5-Methoxyflavanone | OCH3 | H | H | H | H | H | H | 1.710 | 1.563 |
| n84 | 7-Methoxyflavanone | H | H | OCH3 | H | H | H | H | 1.620 | 1.595 |
| n85 | 2′,3′-Dimethoxyflavanone | H | H | H | H | OCH3 | OCH3 | H | 1.550 | 1.625 |
| n86* | 3′,4′-Dimethoxyflavanone | H | H | H | H | H | OCH3 | OCH3 | 1.510 | 1.687 |
| n87 | 5,7-Dimethoxyflavanone | OCH3 | H | OCH3 | H | H | H | H | 1.530 | 1.597 |
| n88 | 4′-Hydroxy-5,7-Dimethoxyflavanone | OCH3 | H | OCH3 | H | H | H | OH | 1.580 | 1.548 |
| n89 | 4′-Hydroxy-3′-methoxyflavanone | H | H | H | H | H | OCH3 | OH | 1.740 | 1.645 |
| n90 | 5-Hydroxy-7-methoxyflavanone | OH | H | OCH3 | H | H | H | H | 1.740 | 1.596 |
| n91 | 2′,3′,6-Trimethoxyflavanone | H | OCH3 | H | H | OCH3 | OCH3 | H | 1.750 | 1.691 |
| n92 | 2′,4′,6-Trimethoxyflavanone | H | OCH3 | H | H | OCH3 | H | OCH3 | 1.510 | 1.332 |
| n93 | 3′,4′,6-Trimethoxyflavanone | H | OCH3 | H | H | H | OCH3 | OCH3 | 1.650 | 1.550 |
| n94 | 2′,3,7-Trimethoxyflavanone | H | H | OCH3 | H | OCH3 | OCH3 | H | 1.570 | 1.691 |
| n95 | 4′-Methoxyflavanone | H | H | H | H | H | H | OCH3 | 1.320 | 1.420 |
| n96 | 3′,6-Dimethoxyflavanone | H | OCH3 | H | H | H | OCH3 | H | 1.740 | 1.909 |
| n97 | 8-Chloro-4′-fluoroflavanone | H | H | H | Cl | H | H | F | 1.450 | 1.490 |
| n98 | 6-Chloro-4′-methoxyflavanone | H | Cl | H | H | H | H | OCH3 | 1.180 | 1.239 |
| n99 | 8-Chloro-4′-methoxyflavanone | H | H | H | Cl | H | H | OCH3 | 1.450 | 1.389 |
| n100 | 5,7-Dihydroxy-4′-methoxy flavanone(isosakuranetin) | OH | H | OH | H | H | H | OCH3 | 1.350 | 1.612 |

Molecular conformers were determined by using MMFF with Spartan'10 program and quantum chemical calculations were made with Hartree-Fock 6-31G*. For each conformer to be used in MCET method, '*.txt' files produced from Spartan were converted to "Electron Topological Matrix" (ETM) files with ETM-Program (ETM-P) [37–40].

Natural, Mulliken and Electrostatic atomic charges, Fukui indices such as f +, f- and coefficients in HOMO / LUMO orbitals, etc. electronic values of atoms can be regarded as LRDs of molecules. These descriptors provide detailed analysis to understand the 3D interaction of the molecule with the receptor. The Fukui index and Frontier Orbital approach, using atomic

coefficients in Frontier orbitals, are related to hydrogen bond interactions that are H-donor/H-acceptor, and covalent interactions. On the other hand, Natural, Mulliken and Electrostatic atomic charges are calculated according to the energy values of all orbitals occupied and their atomic coefficients and are associated with columbic interactions. Since the "Klopman Index" relates to both ionic and covalent interactions using atomic charges and atomic coefficients it is more realistic than LRD descriptors mentioned above, and it was used for the first time in a study submitted by us for L-R interactions. According to these characteristics, the Klopman Index used in the interaction energy between L-R is a very comprehensive and powerful descriptor. In MCET [38, 40–44], we developed a new algorithm for KI that considers the total value of both Coulombic and covalent interactions.

The molecules must match according to a template for LRDs to be optimally clustered in the 3D metric system and brought into a receptor-compatible geometry. For this purpose, a common core structure at all molecules, which is part of the pharmacophore, can form the beginning of the matching. The core structure is a geometric and electronic structure formed by the combination of 3, 4 or 5 atoms in the selected template. At least one of them must be functional (X: O, N or S at C-X and C=X) atom. Its core structure is not only common to all molecules electronically and geometrically, but it must also ensure that the remaining atoms matching in the maximum amount. With this structure, the beginning of the pharmacophore is formed, and then the pharmacophore can be completed by adding only useful ones from the remaining clusters [45].

both the clustering of other atoms and the determination of the pharmacophore structure. The more realistic the core structure chosen as the common area of the pharmacophore, the better the remaining atoms of the molecules will overlap with those of the template. For this purpose, the x, y, and z values of the first three atoms in the common core structure of molecules is placed in Cartesian coordinates as (0, 0, 0-origin)1, (x, 0, 0)2 and (x, y, 0)3. Thus, the first three atoms of all conformers are in a common plane and the coordinates of the other atoms are rearranged accordingly. The atomic coordinates of molecules can be categorized in the same set of other molecules consistent with the cube volumetric tolerance ($d\tau = dx * dy * dz$; where $dx = dy = dz$). Unlike atoms in the core structure, for the other atoms, the electronic values in the same cluster do not need to be within a certain limit. Molecules containing atoms in a cluster will contribute to the interaction with the receptor according to their positive / negative and small / large electronics values.

According to the core structure, the conformer with the maximum number of atoms matching the template is chosen to represent its molecule. However, it is possible that different conformer structures are representative for each of the core structures derived from the template. The second critical step is to select the molecular conformation that matches the template at the maximum number of atoms and the most compatible with respect to the receptor. Initially, the molecules align with the core atoms of the template, while the conformation with the highest number of superimposed atoms will represent all conformers in its molecule. The most compatible structure of the molecules according to the receptor is involved in the interaction. When talking about the skeleton of a molecule, this selected and compatible conformation should come to mind. By preventing unsuitable structures from representing the molecule, perfect clustering of atomic mechanisms is ensured. At a given volume tolerance scale, the atoms of the conformers can optimally superimpose, depending on their corresponding positions. Of course, the best clustering occurs when all molecules can be superimposed with the template structure with the highest number of atoms. The clustering of superimposed atoms provides a 3D similarity between molecules, and some of these clusters may have similarities that can interact with the receptor. With an arrangement resulting from the similarity in the 3D coordinate system, the ligands are directed against the receptor in a harmonious manner by guiding in the same way.

As a result of alignment and superimposition, molecules clustered in the same vector space and their atoms are determined as vector elements in two separate sequences. Furthermore, the coordinates of all atoms in the template framework may not carry the entire structure of the receptor and the pharmacophore. Although the template frame is used as a reference, it may not have enough representative power for a mature clustering. In addition to the template, atoms of different skeletal molecules give rise to different space coordinate values. When constructing a 3D-QSAR model, the number of clusters that can meet all the interaction points of the receptor side is achieved by means of multiple and different structure samples of the molecular samples. numerous and different structures of molecules may be required for enough clustering. This leads to

reference molecules that provide new cluster centers. The idea that most and least active molecules have atoms that significantly alter activity may mean that their atomic coordinates are worth referencing. Some of these molecules are reference molecules that provide number-rich clustering at different coordinates in addition to the atomic coordinates of the template. The selection of reference molecules with different skeletal structures allows the production of more diverse clusters. If the coordinate values of the atoms in the reference molecules are like those in the template, it does not create a new field, but if it is different, there may be the beginning of a new cluster field. However, when a cluster resource does not have enough atoms within its tolerance limits, it cannot lead to clustering [38]. Since an atom from each molecule can be placed in a cluster, clusters that do not contain enough atoms are neglected. Although enough is relative, it can be determined as the ratio of the number of atoms in the cluster to the number of molecules (e.g. 1/3, 2/5, 1/4).

To keep the total number of M clusters manageable, the number of atoms in the cluster is increased or decreased by increasing or decreasing the tolerance value, dτ. Clustering of atoms by their positions is to consider their reactivity depending on where they are located. In ligand-based approaches, the receptor's interaction points can be determined by safely clustering the ligand atoms. Atoms in the same cluster can interact by corresponding a point on the receptor side. For volume limits that exclude some atoms and contain others, the length limits are changed by the tolerance. As the limits of the cube volume change, the numbers of atoms in the clusters change. Whether there is a molecule atom in a cluster will be very effective in creating or validating the model. For this reason, the common volume measurement used for all clusters of the model under study can be decided after testing different cube volumes. The edge length of a cube that creates volume for cluster can be taken close to the bond length (about 1 Å) of the H-atom. When different tolerance values are added to the edge length, clusters of different volume can be derived.

Clusters in the total number M form M-dimensional vector spaces with different atoms in each set. It can interact with the ligand receptor in an m-portion of the geometric surface consisting of these clusters. This contact can be considered as the geometrical structure of the m-dimensional sub-cluster, the structure of the receptor responsible for activity. Clusters that do not provide enough improvement to prevent unwanted background noise are ignored in the sub-cluster and an effective sub-cluster is created accordingly. Once the total number of clusters has been determined, sub-clusters are created by applying a combination. A stochastic control is performed with GA for each sub-cluster examined. During the creation of the sub-cluster, a new set is created by adding or subtracting field using GA with acceptance or rejection. For this, a filter is made with minimum errors in all compounds, and a sub-cluster is created by rejecting insufficient variables with GA. The most suitable sub-cluster to construct a pharmacophore model has the best statistical result compared to another sub-cluster.

As can be seen from many previous applications, GA has been able to find the most appropriate solutions to the problems with an equally wide range of research areas [46]. In this context, even the lowest solution level produced by GA cannot be proven to include the most suitable solution. Optimization problems are solved by checking the accuracy of parameters to save/restore parameter sets to ensure that they are applied effectively to many problems [47]. Some results produced by GA mean that it confirms an optimal solution that is consistent across samples. This does not mean that the problem is found with an analytical solution, but it can be of great help in non-polynomial problems [48]. As with other programs, there are many uncertainties in MCET program that cannot be avoided due to the algorithm GA is applied to. Therefore, it is more appropriate to talk with GA about a compatible and practical solution rather than an analytical exact solution.

A 3D QSAR model using LRDs is required to demonstrate the 3D interaction between L-R. The creation of a good model is possible with the correctly selected LRD type. Different models are formed from different types of LRDs of atoms corresponding to the sub-cluster [43]. The parameter values of the receptor are calculated as adjustable constant according to LRD values of the ligand atoms corresponding to the sub-cluster. Using the sub-cluster model, the activity of each molecule is calculated using LRD values of the atoms on the ligand side and the parameters of the receptor side.

The activity of the molecule that contains atom in a cluster will vary depending on the electronic value of the atom and the corresponding parameter of the receptor. However, given the large number of clusters, only a certain sub-cluster corresponding to the interaction points of the receptor will make sense. Since the elements of the sub-cluster are

multidimensional vector spaces, the contribution of each is difficult to follow. Therefore, it is useful to place the sub-cluster in a consecutive index along an axis to show the activity change of each molecule depending on the atom it contains. Since the discussed sub-cluster items are listed on the axis with the same index number, the activity change of all molecules will take an individual shape. Changes in the activity of the molecule at each point along the axis create a VFF. It is now both safe and easy to divide molecules into training and test sets based on VFF similarity.

Consider an example of a series molecule consisting of a series of high-dimensional clusters.

**X** = {x (n, m) ∈ RN, M, n = 1, 2, ..., N; m = 3, 4, ... M.

Where n and m are the molecular number and the sub-cluster number of the pharmacophore, while N and M are given as the total number of the molecules and clusters, and molecular activity should start after m = 3, since the interaction point of the pharmacophore will occur with at least 3 atoms. The size of the rows and columns of the independent variable matrix X is the number of molecules (N) and sets (M), respectively. On the other hand, the x-independent variable is the electronic value of the atom in the m-cluster of the n-molecule in the X matrix [49].

Dimensionality Reduction (DR) is performed for the first time by using m-number scenarios which are effective from M-number total clusters.

A transform **X**∈RN, M (Dimensionality Reduction with GA) → **X**∈RN, m, m<=M. The calculation of activities using scenario-effected clusters and their corresponding parameters is given by Eq. (1).

**y** = f (**X**, **w**) (1)

Here, both **y** and **w**, are N and m-dimensional vector values as the activities of the molecules and the parameters of the receptor side, respectively. In Eq. (2), the **yn** activity can be calculated using the **xn** vector variable and **w** parameters with **mn** (mn <=m) clusters for the nth molecule.

**yn** = f(**xn**, **w**) (2)

A second DR consists of transforming the **xn** vector variable with the mn clusters in the three dimensions for the nth molecule into the **zn** vector variable in one-dimension.

Another transform (DR with VFF) is **xn**∈Rm → **zn**∈R:

**zn** = g(**xn**, **w**), **z**∈R, (3)

The function **zn** in Eq. (3) shows the fingerprint for n-molecule.

According to the **z**-vector values in the one-dimensional z-axis, the **y** activities in the y-axis can be calculated as in Eq. (4).

**y** = f(**z**) (4)

The x-independent variable values and w-parameters in the multi-dimensional vector space are used on the m indices along the z axis. The **x** and **w** arguments in multi-dimensional sub-clusters are placed in indexes on the one-dimensional z-axis. Accordingly, plotting dependent variable activity along the y-axis is an important simplification. The values of **z** are also not plotted, since **z**, easily, appears as positive or negative with increasing or decreasing activity in each index. The difference between **y** and **z** is that the values of z are the positive or negative value for the corresponding index, while the values of y represent the total change in activity until that index due to a positive or negative change. In short, the values of **y** are the sum of the values of **z**. The values of the y-axis values corresponding to each index along the axis form VFF. Molecules with similar changes in their VFF can easily replace each other. For the same zn element of any two molecules examined for

similarity, the differences in the y-axis in the partial least square method were evaluated by the total deviation and two molecules with the best similarity were placed in the training and test sets.

The activity of the molecule varies in proportion to the electronic amount of the atom settled in any element of the sub-cluster. There is no change in the activity of molecules without atoms in any index of the sub-cluster. Here, m is the total number of interactions on the receptor side, which is the interaction point of the pharmacophore. The interaction number of the ligands can be mn, maximum m (mn <= m), and if a molecule contains an atom in the entire sub-cluster, it becomes mn = m. Therefore, molecules can have the same or different number of mn. The formation of different VFFs for each molecule is not only due to the difference in clusters, but also because of the difference in LRD value in each cluster. Therefore, VFFs of each compound are determined using LRD values of the atoms occupying the sub-cluster. Molecules with different activities due to different VFFs may have different LRDs within the same element, even if they do not contain atoms in different elements of the same sub-cluster. The difference between VFFs is due to both having different sub-clusters and different LRD values corresponding to the same element index. For molecules that contain atoms in different elements of the same pharmacophore sub-cluster, two different sub-clusters can occur. While the pharmacophore sub-cluster belongs to the receptor, two different sub-cluster belong to the molecules. Accordingly, even if the two molecules have the same sub-cluster, they may have different VFFs due to different LRD values. VFF change resulting from the multiplication of LRD value of the Ligand side in the filled sub-clusters by the parameters on the receptor side is considered as the changes of the activity. The sub-clusters of molecules with similar activities and the similarities of LRD values of the atoms in each cluster give rise to similar VFF. The similarity or difference in VFFs of the two compounds is calculated by using the atoms present and absent in each cluster and LRD values in the atoms. This similarity or difference leads to different divisions in training and test sets, and therefore different models. It is pointless to use two VFFs that are very similar to each other in training or test sets. However, placing the molecules with two similar VFF, separately, into sets helps identify and validate the model. For structure-activity similarity, the molecules whose activity is close to each other are grouped together by ordering the molecules according to their activities. Considering that the molecules in the training and test sets are approximately 4/5 and 1/5, the molecules are divided into groups of 5. The maximum similarity between structure-activity is due to the similarity in VFFs used as arguments in the model. If LRD values of the two molecules in a sub-cluster were the same (this is almost impossible), the function curves of VFFs would be exactly overlapping and would necessarily have the same activity change at each interaction point. Due to VFF function curve being very close to each other, each of the two molecules can be placed in training and test sets. It is a good approach to estimate and verify the model created by this arrangement.

LRD value in the sub-cluster can contribute positively or negatively and larger or smaller to molecular activity. Due to LRD similarities in the same indices of the sub-cluster, the activities of the molecules are similar at these points. VFFs of the two molecules, whose LRD similarity is very close at each point of the sub-cluster, are activity changes. This means that the index numbers in the sub-cluster and LRD values of atoms involved in the activity of a molecule divided into the test set are like at least one of the molecules in the training set. On the other hand, despite the same clusters, different VFFs are formed from different LRDs. VFFs of the two molecules can only be like the same clusters and close LRD values. The better this similarity, the stronger the probability that one of the two compounds will be included in training and the other in the test set. Thus, both the same sub-cluster and similar LRD at each point generate similar VFF values, indicating the degree of similarity of the molecules, and two similar molecules can be divided into sets accordingly. Here we present an important innovation in the best and most consistent splitting algorithm by comparing the models obtained with VFFs. For samples in the same series of molecular activity data, VFF-dependent splitting performance results in different parameters and models. The model selection will be based on the prediction ($Q^2$) and validation ($R^2$) forces of the training and test sets using the independent variable LRDs in the sub-cluster. The original training and the displacement of several samples in the (external) test sets indicate the emergence of different models; The separation of both sets cannot be left to chance, random or manual algorithm. Therefore, it will be tried to show the examples in the training and test sets in a model that can explain the interaction between L-R optimally by splitting it according to VFF descriptor.

To reduce the effect of parameters due to the variation of the original training and (external) test sets, advanced models with LOO-CV (although the number of molecules in the training set are small) are presented. The proposed parameters of the model in the training set were checked in the external test set. The number of samples in the data set was used more efficiently without the need for a separate validation set by the LOO-CV method. As a result, the training and testing sets of VFFs were both trained and externally judged with a coherent distinction.

## Result And Discussion

Since the geometric structure of each conformer remains constant, the non-diagonal values given in length as Å remain the same. However, the electronic data of the atoms at the diagonal location are given as variable in rows on the matrix. Diagonal values are optionally marked and used as atomic descriptors in Figure 1.

Since the template is chosen from a simple and low atomic structure and only one conformation cannot contain all interaction points, it is necessary to create new reference positions to increase the number of proposed interaction points. In addition to the atomic positions of the template, the atomic positions of the most and least active molecules can be used as reference points for clustering. It is possible to find the positions that increase the activity from the most active those and decrease it from the least active those. Therefore, the positions of the interaction points can be created from the coordinate values of the atoms of some selected molecules. These coordinate values are the values after the first three atoms of the core structure are drawn to the coordinate center in all molecules. For the interaction points of the proposed pharmacophore; a) atomic numbers of the conformer in the reference

According to the tolerance scale (e.g. 0.3 Å), atoms with similar positions were placed in the same cluster. With respect to the atoms in each cluster, it leads to two vectors of the number sequences of molecules and atoms. Thus, a multidimensional vector space was created from clusters with different positions in the 3D system and the vector elements in them. Except for the cluster elements that make up the core structure, in other clusters it is unlikely that all molecules will have one atom. Generally, clusters do not contain the atoms of all molecules. Therefore, depending on the number of atoms associated with a threshold, a cluster may exist or be ignored, and the total number of clusters increases or decreases relative to a relative threshold value. If the number of atoms settled in the cluster reaches approximately 2/5 of the total molecule, the number of sub-clusters has been increased by one. Among these clusters, the total number of members of the sub-cluster predicted by GA was found to be m = 9 in this study and their coordinate values are given in Table 2.

Table 2
Molecule, conformer, and atomic numbers of the atoms where the interaction points are determined, coordinate values of the atom positions and the calculated "Kappa" and "Ksi" parameter values of the corresponding receptor.

| Mol_Conf No | Atom No | x, y, z Cartesian coord. | | | Position (m=) | Kappa-(κ) values | Xi-(ξ) values |
|---|---|---|---|---|---|---|---|
| | | x | y | Z | | | |
| n01_03 | O1 | 01* | 01* | 01* | 1 | -0.205 | 1.957 |
| n01_03 | O2 | 4.057 | 02* | 02* | 2 | -0.046 | 12.457 |
| n01_03 | C2 | 0.665 | 1.177 | 03* | 3 | -0.257 | -4.672 |
| n77_02 | C6 | 0.498 | 3.555 | -0.051 | 4 | 0.012 | -17.426 |
| n77_02 | C5 | 1.883 | 3.670 | -0.025 | 5 | -0.010 | -0.417 |
| n77_02 | C1 | -0.133 | 2.315 | -0.035 | 6 | 0.048 | 7.428 |
| n77_02 | C7 | 0.700 | -1.124 | 2.031 | 7 | -1.305 | -0.341 |
| n77_02 | O5 | 2.885 | -1.685 | 4.925 | 8 | -0.343 | 0.048 |
| n77_02 | C14 | 1.785 | -1.415 | 4.161 | 9 | 0.061 | 2.813 |
| *Coordinate values where x, y, or z are 'zero'. | | | | | | | |

Once the coordinates of a compound are arranged with respect to the core structure, the clustering of the atoms in the compound occurs readily. The x, y, z-coordinates of the first three atoms of the core structure (01,01,01; x2,02,02; x3, y3,03) are drawn to provide the common order in the coordinates of the compounds. According to this, atom-1 (O1) is in the coordinate center and origin, atom-2 (O2) is on the axis of x, and atom-3 (C2) is in the xy-plane. According to these three atoms, other atoms were subjected to common geometric arrangement with internal coordinate values.

By sorting activities (e.g. from large to small), grouping of molecules with similar activity can be easy. Since the ratio of molecules in the test set to those in the training set is approximately 1/4, a group of 5 molecules is created to divide 1 into the test set and the other 4 into the training set. The activities of the two molecules can mimic each other as much as the similarities or differences of LRDs corresponding to the same index of interaction points. Therefore, the similarities of molecular descriptors in a group can be followed from the sum of squares of differences of LRD values at each interaction point. In other words, the smaller the difference between LRD values of the two molecules at each point, the smaller the sum of the Partial Small Square (PLS) values. Since molecules may not contain atoms in the same element of the sub-cluster studied, or may contain atoms with different LRDs, each molecule may have different VFFs, and therefore, differences arise in the activity of each molecule. VFFs of a selected example 5 molecule group are graphically shown in Figure 2. The similarity seen in the graph can be determined by the smallest total value calculated using PLS. Thus, considering the similarities at each point, two most similar molecules at all points are divided into training and test sets. The most appropriate division of the molecules into both clusters is understood by the predicted and validated activities being very close to the observed ones.

In this article, the automated and rational VFF method we have just developed has been compared with random and manual modeling techniques for the flavonoid series. Comparing VFF's other automated or rational approaches can only be within MCET method. The paradigm of MCET, which operates according to LRD like the Klopman index, is quite different from any method. Therefore, it does not make sense to compare the automatic or rational approach used in a study with the applied VFF in only MCET. On the other hand, since we could not write any automated or rational methods in MCET for now, we were able to compare VFF only with manual and random methods. However, dealing with the details of molecular interaction at every point of the pharmacophore shows that the method we developed is a very safe and rational approach.

Table 3

$Q^2$ and $R^2$ results for different splitting methods used

| Method | $Q^2$ | $R^2$ |
|---|---|---|
| **VFF[a]** | **0.604** | **0.760** |
| Manual | 0.566 | 0.683 |
| Random | 0.627 | 0.510 |
| *a The best result is shown in bold* | | |

The statistical numbers of the models from VFF (shown in bold) and the other two approaches are given in Table 3. The fact that the model resulting from the split with VFF compared to other approaches has high statistical results shows that VFF is safer. Since 100 molecules have very different skeletons and scattered conformation structures, it has not been easy to choose the pharmacophore. After calculating the parameters of the receptor side of the interaction between L-R, the activity changes in VFF form containing one-dimensional index elements is graphically shown. The user can understand how the interaction at each point of the proposed model changes with the graph in VFF. Therefore, for researchers, the 3D pharmacophore proposed by using VFF has been more attractive than expected.

A model obtained with the LOO-CV applied in the training set of molecules was proposed with the value of $Q^2$ = 0.604 and was confirmed with $R^2$ = 0.760 in the strategically excluded test set. The reasons why $R^2$ is greater than $Q^2$ are.

a) The number of molecules in the test set is considerably less than the training set,

b) The molecules in the training set are divided sufficiently similarly to the molecules in the test set,
c) The independent variables used in the model are small and well chosen.

When the $Q^2$ and $R^2$ values from which the experimental activity values obtained from the literature [36] are compared with the values found because of the MCET method, differences are observed. The reason for this is that we use the 4D-QSAR MCET method while using the hologram quantitative structure-activity relationships method in the literature [36]. Since the calculations are made according to different parameters in two methods, no comparison can be made.

# Conclusion

In order to calculate the activity of the flavonoid series derivatives and divide them into training and test sets, three different division methods were used in the MCET method and the models created accordingly were compared. It has been found that a new application, which we define as VFF, which is among the splitting methods, has the best performance.

In the 3D metric system, different subsets formed between clustered atoms formed as a result of alignment and superimpose were examined and their effects on activity were investigated by GA. The contribution of each element of the subset to the activity was observed as a fingerprint of the relevant molecule. A subset that gives the best results shows the geometric and electronic properties of Pha. The vectorial index of the interaction points of a subset that is thought to have the best result was given and the activity change in each index is plotted graphically for 5 molecules chosen as samples 4 in the training set and 1 in the test set. As seen in the graph, how much the activity of a molecule increases or decreases for each interaction point of the subset can be shown for the first time in this study. How the subset contributes to the activity of any molecule at each point of a one-dimensional vector index has formed the molecule's fingerprint. According to the similarities in VFF, the molecules were safely divided into training and test sets, have been preventing the accumulation of two separate molecules that have representative or very similar to each other in one set.

# Declarations

**Funding** This work was financially supported by Erciyes University Scientific Research Projects (BAP) of Turkey (Grant no. FDK-2018-8187).

**Conflicts of interest** There is no conflict of interest for all participating authors. On behalf of all authors, I declare that there is no conflict of interest.

**Author contribution** The authors of the current manuscript Tuğba Alp Tokat, Burçin Türkmenoğlu and Yahya Güzel contributed equally to this work. All authors read and approved the final manuscript.

**Data Availability** Availability of data and material The data generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

**Code availability** Not applicable.

**Conflict of interest** The authors declare no competing interests.

# References

1. Chitre TS, Asgaonkar KD, Patil SM, Kumar S, Khedkar VM, Garud DR (2017) QSAR, docking studies of 1,3-thiazinan-3-yl isonicotinamide derivatives for antitubercular activity. Comput Biol Chem 68: 211-218

2. Saul LK, Roweis ST (2003) Think globally, fit locally: Unsupervised learning of low dimensional manifolds. J Mach Learn Res 4: 119-155. https://doi.org/10.1162/153244304322972667

3. Kausar S, Falcao AO (2019) Analysis and Comparison of Vector Space and Metric Space Representations in QSAR Modeling. Molecules. *24* (9): 1698. https://doi.org/10.3390/molecules24091698

4. Jolliffe IT (2002) Choosing a Subset of Principal Components or Variables. In: Principal component analysis, 2nd edn. Springer, New York, pp 111-149. https://doi.org/10.1002/0470013192.bsa501

5. Wickelmaier F (2003) An Introduction to MDS, In: Sound Quality Research Unit at Alaborg University, Denmark, pp 1-26.

6. Somervuo P, Kohonen T (1999) Self-organizing maps and learning vector quantization for feature sequences. Neural Process Lett 10 (2): 151-159. https://doi.org/10.1023/A:1018741720065

7. Agrafiotis DK, Lobanov VS (2001) Multidimensional scaling of combinatorial libraries without explicit enumeration. J Comput Chem 22 (14): 1712-1722. https://doi.org/10.1002/jcc.1126

8. Agrafiotis DK (2003) Stochastic proximity embedding. J Comput Chem 24 (10): 1215-1221

9. Low Y, Sedykh A, Fourches D, Golbraikh A, Whelan M, Rusyn I, Tropsha A (2013) Integrative

Chemical-Biological Read-Across Approach for Chemical Hazard Classification. Chem Res Toxicol 26 (8): 1199-1208

10. Gasteiger J (2003) Handbook of chemoinformatics: from data to knowledge. Wiley-VCH, Weinheim

11. Hendrickson JB (1991) Concepts and Applications of Molecular Similarity. Science 252 (5009): 1189

12. Barnard JM (1993) Substructure Searching Methods - Old and New. J Chem Inf Comp Sci 33 (4): 532-538. https://doi.org/10.1021/ci00014a001

13. Flower DR (1998) On the properties of bit string-based measures of chemical similarity. J Chem Inf Comp Sci 38 (3): 379-386. https://doi.org/10.1021/ci970437z

14. Bajusz D, Racz A, Heberger K (2015) Why is Tanimoto index an appropriate choice for fingerprint based similarity calculations?. J Cheminformatics 7(20)

15. Tversky A (1977) Features of Similarity. Psychol Rev 84 (4): 327–352. https://doi.org/10.1037/0033-295X.84.4.327

16. Jadrich RB, Lindquist BA, Pineros WD, Banerjee D, Truskett TM (2018) Unsupervised machine learning for detection of phase transitions in off-lattice systems. II. Applications The Journal of Chemical Physics 149 (194110). https://doi.org/10.1063/1.5049850

17. Cuadras CM, Arenas C (1990) Distance Based Regression-Model for Prediction with Mixed Data. Commun Stat Theory 19 (6): 2261-2279. https://doi.org/10.1080/03610929008830319

18. Coates A, Lee H, Ng AY (2011) An analysis of single-layer networks in unsupervised feature learning. Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, PMLR 15:215-223

19. Cullum JK, Willoughby RA (1985) Real rectangular matrices. In: Lanczos Algorithms for Large Symmetric Eigenvalue Computations (ed). Vol. II. Brikhauser, Boston, pp 273-359

20. Puzyn T, Mostrag-Szlichtyng A, Gajewicz A, Skrzynski M, Worth AP (2011) Investigating the influence of data splitting on the predictive ability of QSAR/QSPR models. Struct Chem 22 (4): 795-804. https://doi.org/10.1007/s11224-011-9757-4

21. Myint KZ, Xie X Q (2010) Recent Advances in Fragment-Based QSAR and Multi-Dimensional QSAR Methods. Int J Mol Sci 11 (10): 3846-3866

22. Racz A, Bajusz D, Heberger K (2015) Consistency of QSAR models: Correct split of training and test sets, ranking of models and performance parameters. Sar Qsar Environ Res 26 (7-9): 683-700

23. Arthur DE, Uzairu A, Mamza P, Abechi SE, Shallangwa GJ (2020) Activity and toxicity modelling of some NCI selected compounds against leukemia P388ADR cell line using genetic algorithm-multiple linear regressions King Saud Univ Sci 32 (1): 324-331. https://doi.org/10.1016/j.jksus.2018.05.023

24. Burden FR, Ford MG, Whitley DC, Winkler DA (2000) Use of automatic relevance determination in QSAR studies using Bayesian neural networks. J Chem Inf Comp Sci 40 (6): 1423-1430

25. Burden FR (1998) Holographic neural networks as nonlinear discriminants for chemical applications. J Chem Inf Comp Sci 38 (1): 47-53

26. Obrezanova O, Csanyi G, Gola JMR, Segall MD (2007) Gaussian processes: A method for automatic QSAR Modeling of ADME properties. J Chem Inf Model 47 (5): 1847-1857

27. Schwaighofer A, Schroeter T, Mika S, Laub J, ter Laak A, Sulzle D, Ganzer U, Heinrich N, Muller KR (2007) Accurate solubility prediction with error bars for electrolytes: A machine learning approach. J Chem Inf Model 47 (2): 407-424

28. Kohonen T (1990) The Self-Organizing Map. IEEE 78 (9): 1464-1480

29. Chiorboli C, Piazza R, Carassiti V, Passerini L, Tosato ML (1993) Application of Chemometrics to the Screening of Hazardous Substances Part II. Advances in the multivariate characterization and reactivity Modeling of Haloalkanes. Chemom Intell Lab Syst 19 (3): 331-336

30. Baroni M, Clementi S, Cruciani G, Kettaneh-Wold N, Wold S (1993) D-Optimal Designs in Qsar. Quant Struct-Act Rel 12: 225-231

31. Gramatica P, Pilutti P, Papa E (2004) Validated QSAR prediction of OH tropospheric degradation of VOCs: Splitting into training-test sets and consensus modeling. J Chem Inf Comp Sci 44 (5): 1794-1802

32. Hudson BD, Hyde RM, Rahr E, Wood J (1996) Parameter based methods for compound selection from chemical databases. Quant Struct-Act Rel 15 (4): 285-289. https://doi.org/10.1002/qsar.19960150402

33. Gobbi A, Lee ML (2003) DISE: Directed Sphere Exclusion. J Chem Inf Comp Sci 43 (1): 317-323

34. Kennard RW, Stone LA (1969) Computer Aided Design of Experiments. Technometrics 11: 137-148. https://doi.org/10.1080/00401706.1969.10490666

35. Snarey M, Terrett NK, Willett P, Wilton DJ (1997) Comparison of algorithms for dissimilarity-based compound selection. J Mol Graph Model 15 (6): 372-385

36. Cho M, Yoon H, Park M, Kim YH, Lim Y (2014) Flavonoids promoting HaCaT migration: I. Hologram quantitative structure-activity relationships. Phytomedicine 21 (4):560-569

37. Türkmenoğlu B, Yılmaz H, Su EM, Alp Tokat T, Güzel Y (2017) 4D-QSAR Study of Flavonoid Derivatives with MCET Method. International Journal of Chemistry and Technology 1: 14-23. https://doi.org/10.32571/ijct.338920

38. Türkmenoğlu B, Güzel Y (2018) Molecular docking and 4D-QSAR studies of metastatic cancer inhibitör thiazoles. Computational Biology and Chemistry 76: 327-337

39. Su EM, Turkmenoglu B, Guzel Y (2016) 3D Biostructure Visualisation Using 4D-QSAR Model for Substitute Ureas Binding at the Raf-1 Kinase Receptor Site. International Journal of Innovative Studies in Sciences and Engineering Technology 2 (12): 67-75

40. Yilmaz H, Boz M, Turkmenoglu B, Guzel Y (2014) Pharmacophore and Functional Group Identification of 4,4 '-dihydroxydiphenylmethane as Bisphenol-A (BSA) Derivative. Trop J Pharm Res 13 (1): 117-126. http://dx.doi.org/10.4314/tjpr.v13i1.17

41. Turkmenoglu B, Guzel Y, Su EM, Kizilcan DS (2020) Investigation of inhibitory activity of monoamine oxidase A with 4D-QSAR using Fukui indices identifier. Mater Today Commun 25. https://doi.org/10.1016/j.mtcomm.2020.101583

42. Tokat TA, Turkmenoglu B, Guzel Y, Kizilcan DS (2019) Investigation of 3D pharmacophore of Nbenzyl benzamide molecules of melanogenesis inhibitors using a new descriptor Klopman index: uncertainties in model. Journal of Molecular Modeling 25 (8). https://doi.org/10.1007/s00894-019-4120-6

43. Kizilcan DS, Turkmenoglu B, Guzel Y (2020) The use of the Klopman index as a new descriptor for pharmacophore analysis on strong aromatase inhibitor flavonoids against estrogen-dependent breast cancer. Struct Chem 31 (4): 1339-1351. https://doi.org/10.1007/s11224-020-01498-9

44. Guzel Y, Aslan E, Turkmenoglu B, Su EM (2018) 4D-QSAR Studies Using a New Descriptor of the Klopman Index: Antibacterial Activities of Sulfone Derivatives Containing 1, 3, 4-Oxadiazole Moiety Based on MCET Model. Curr Comput-Aid Drug 14 (3): 207-220. https://doi.org/10.2174/1573409914666180514093543

45. Griffiths M, Niblett SP, Wales DJ (2017) Optimal Alignment of Structures for Finite and Periodic Systems. J Chem Theory Comput 13 (10): 4914-4931

46. Holland JH (1973) Genetic Algorithms and the Optimal Allocation of Trials. SIAM Journal on Computing 2: 88-105. https://doi.org/10.1137/0202009

47. Tóth Z (2003) A graphical user interface for evolutionary algorithms. Acta Cybernetica 16 (2): 337-365

48. Goldberg DE, Holland JH (1988) Genetic Algorithms and Machine Learning. Mach Learn 3: 95-99. https://doi.org/10.1023/A:1022602019183

49. Baskin II (2016) Dimensionality Reduction in Chemoinformatics. Generative 1 Topographic Mapping. In: International School-Seminar on Computer-Aided Molecular Designed. Kazan, Russia

# Figures

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **-0.618** | **-0.736** | **0.498** | **0.507** | **-0.466** | **-0.422** | **0.542** | **-0.307** | **0.529** | **-0.767** | **0.517** | **n_chrge** |
| -0.701 | -0.642 | 0.52 | 0.518 | -0.373 | -0.259 | 0.526 | -0.289 | 0.517 | -0.793 | 0.499 | m_chrge |
| -0.482 | -0.669 | 0.755 | 0.776 | -0.845 | -0.786 | 0.476 | -0.391 | 0.474 | -0.688 | 0.475 | e_chrge |
| -0.185 | 0.005 | 0.028 | 0.001 | 0.008 | 0.012 | 0.001 | 0.259 | 0.001 | 0.001 | 0.001 | P_Fkui |
| -0.216 | 0.147 | 0.117 | 0.193 | 0.006 | 0.051 | 0.001 | 0.007 | 0.001 | 0.003 | 0.001 | N_Fkui |
| 0.027 | 0.118 | 0.093 | 0.155 | 0.004 | 0.041 | 0.001 | 0.006 | 0.001 | 0.001 | 0.001 | HOMO |
| 0.052 | 0.004 | 0.022 | 0.001 | 0.006 | 0.01 | 0.001 | 0.207 | 0.001 | 0.001 | 0.001 | LUMO |
| 0.684 | 0.027 | 0.562 | 0.15 | -0.305 | 0.332 | 0.156 | -0.169 | 0.121 | 1.207 | 0.231 | I_Index |
| **O1** | **O2** | **C2** | **C6** | **C5** | **C3** | **H2** | **C10** | **H4** | **O5** | **H5** | |
| -0.618 | 4.051 | 1.345 | 3.591 | 4.125 | 2.401 | 4.762 | 4.989 | 5.447 | 5.951 | 5.989 | **O1** |
| | -0.736 | 3.607 | 5.027 | 4.263 | 2.357 | 1.849 | 6.037 | 6.66 | 5.334 | 4.769 | **O2** |
| | | 0.498 | 2.378 | 2.78 | 1.41 | 3.813 | 5.352 | 4.283 | 6.123 | 6.113 | **C2** |
| | | | 0.507 | 1.389 | 2.783 | 4.276 | 6.77 | 1.908 | 7.61 | 7.665 | **C6** |
| | | | | -0.466 | 2.429 | 3.13 | 6.95 | 2.435 | 7.36 | 7.278 | **C5** |
| | | | | | -0.422 | 2.419 | 5.565 | 4.581 | 5.784 | 5.587 | **C3** |
| | | | | | | 4.564 | 6.006 | 3.146 | 7.05 | 7.149 | **C1** |
| | | | | | | 4.775 | 4.315 | 6.632 | 4.965 | 4.923 | **C8** |
| | | | | | | 3.875 | 5.207 | 6.935 | 5.197 | 4.873 | **C9** |
| | | | | | | 2.384 | 5.469 | 6.02 | 5.245 | 4.861 | **C12** |
| | | | | | | 1.891 | 6.389 | 3.804 | 6.477 | 6.263 | **C4** |
| | | | | | | 5.131 | 2.79 | 6.915 | 3.669 | 3.835 | **C7** |
| | | | | | | 5.857 | 1.382 | 8.216 | 1.366 | 1.923 | **C14** |
| | | | | | | 4.944 | 2.405 | 7.595 | 2.409 | 2.457 | **C13** |
| | | | | | | 6.177 | 2.407 | 6.918 | 4.132 | 4.577 | **C11** |
| | | | | | | 6.92 | 1.384 | 7.593 | 3.617 | 4.305 | **C15** |
| | | | | | | 5.481 | 7.632 | 0.954 | 8.668 | 8.803 | **O4** |
| | | | | | | 0.959 | 7.004 | 4.676 | 6.652 | 6.291 | **O3** |
| | | | | | | 0.542 | 6.791 | 5.543 | 6.201 | 5.733 | **H2** |
| | | | | | | | -0.307 | 8.221 | 2.35 | 3.157 | **C10** |
| | | | | | | | | 0.529 | 9.111 | 9.213 | **H4** |
| | | | | | | | | | -0.767 | 0.952 | **O5** |
| | | | | | | | | | | 0.517 | **H5** |

C5's Natural charge =-0.466
C2-O1 bond length=1.345 Å
C10-O2 distance=6.037 Å

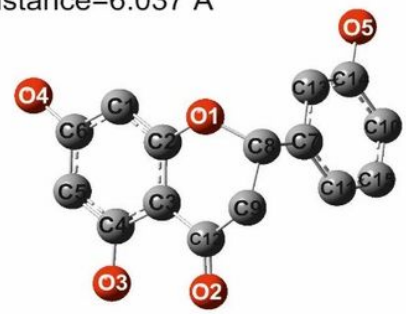



**Figure 1**

ETM matrix; the distances and bonds between atoms are given with the Å value in place of non-diagonal elements, while the electronic values of the atoms (here, atomic charges) are given diagonally.

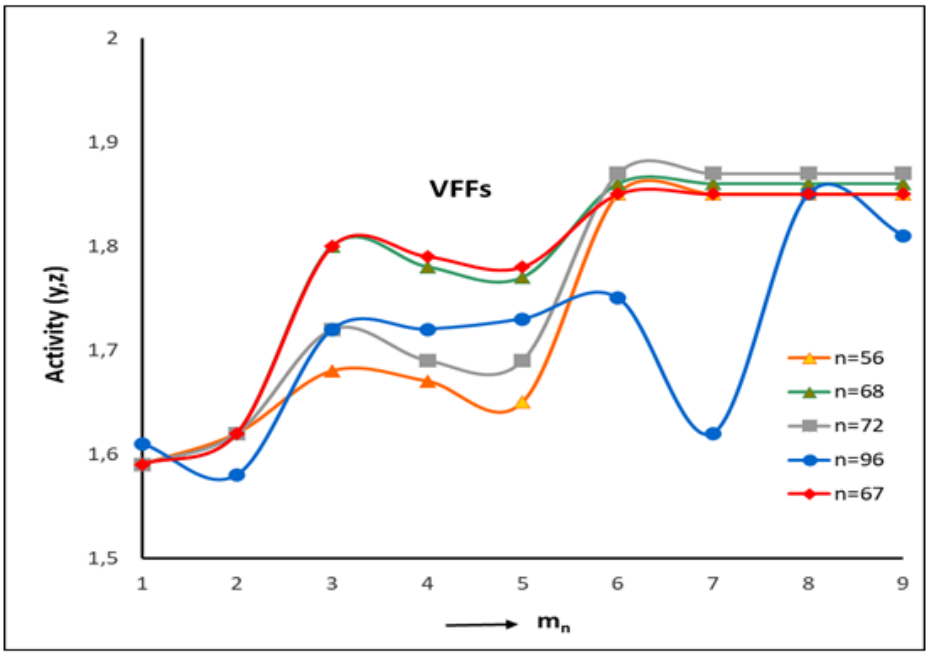



**Figure 2**

Similarities and differences of VFF of the 5 selected molecules as a sample.