Basic Principles of (QSAR) Quantitative Structure Activity Relationship and its methods

Rohit Dnyaneshwar Chaudhari

Department of Phamacy, Shastry Institute of pharmacy, Erandol, Jalgaon

***Abstract*: - complementing combinatorial chemistry and high-throughput screening. Virtual sifting and screening of combinatorial libraries have recently gained attention as strategies based on quantitative structure-activity relationship (*QSAR*) examination, a field with established methodology and successful history. These chemoinformatic methods heavily rely on it. We discuss the computational methods used to create QSAR models in this audit. We begin by outlining their suitability for high-throughput screening and recognizing a QSAR show's common plot. Following this, we focus on the methods used to create the three fundamental components of the QSAR demonstration, specifically the methods used to depict the atomic structure of compounds, select instructive descriptors, and anticipate actions. We present both the recently presented QSAR-specific well-established strategies and procedures. Scientists and regulators have turned their attention to developing general validation principles for QSAR models in the context of chemical regulation in response to the recent REACH Policy of the European Union (previously known as the Setubal principles, now the OECD principles). Some fundamentals are briefly discussed in this paper: statistical validation, the Applicability Domain (*AD*), and an unambiguous algorithm an example of a quick check of the applicability domain for MLR models and some concerns regarding the reproducibility of the *QSAR* algorithm are presented. Cross validation, bootstrap, and other well-known statistical methods for external validation are contrasted with common misconceptions and myths regarding popular methods for confirming internal predictivity, particularly for MLR models. There is evidence that only models that have been validated externally after their internal validation can be considered reliable and applicable for both external prediction and regulatory purposes. The differences between the two validating approaches are highlighted.**

***Keywords:* - QSAR, molecular descriptors, feature selection, machine learning.**

INTRODUCTION :- Quantitative structure-activity relationships (QSAR) correlate, within congeneric series of compounds, affinities of ligands to their binding sites, inhibition constants, rate constants, and other biological activities, either with certain structural features (Free Wilson analysis) or with atomic, group or molecular properties, such as lipophilicity, polarizability, electronic and steric properties .[5,6] Since then, QSAR equations have been utilized to describe tens of thousands of biological activities that are contained in a variety of drug series and drug candidates. The physicochemical properties of the ligands 1-3,16 have been successfully correlated, particularly with data on enzyme inhibition. In cen: In some cases, when the proteins' X-ray structures were available, the QSAR regression models' results could be interpreted with the additional information from the three-dimensional (31) structures.[7,8] High throughput screening (HTS) is a recent technological advancement that has had a significant impact on the drug discovery pipeline. It makes it possible to quickly synthesize many small-molecule compounds and evaluate their activity when used in conjunction with combinatorial chemistry.[9,10] The focus has shifted from sifting through large, diverse molecule collections to libraries that are designed more rationally as experience with these technologies has progressed. [11] Virtual filtering and screening have been recognized as complementary to high-throughput screening considering the requirement for knowledge-guided compound screening. [12,13] These methods heavily rely on quantitative structure-activity relationship (QSAR) analysis, which has undergone continuous development ever since Hansch's work [14] at the beginning of the 1960s. Finding a model that makes it possible to link activity and structure within a family of compounds is the primary goal of the QSAR approach. [15,16] Recent discussions in the scientific and regulatory communities have focused a lot on model validation. It was thought to be important to develop a set of internationally accepted principles for QSAR validation, give regulatory bodies a scientific basis for deciding whether QSAR estimates of regulatory endpoints are acceptable, and encourage QSAR models to be accepted by everyone.

Recent discussions in the scientific and regulatory communities have focused a lot on model validation. It was thought to be important to develop a set of internationally accepted principles for QSAR validation, give regulatory bodies a scientific basis for deciding whether QSAR estimates of regulatory endpoints are acceptable, and encourage QSAR models to be accepted by everyone. At an international workshop that was held in Setubal in 2002, a set of principles for determining the validity of QSARs were proposed as the "Setubal Principles." The following data ought to be associated with a QSAR model in order to make it easier to consider it for regulatory purposes: 1) an established endpoint; 2) a straightforward algorithm; 3) a clearly defined field of application; (4) appropriate measures of robustness, predictability, and goodness-of-fit; 5) if possible, a mechanistic interpretation. The consolidation of the previous Principles 5 (internal validation) and 6 (external validation) into a single Principle 4 represented the most significant change to the Setubal principles in the new OECD principles. However, it is important to note that, at the September 2004 OECD meeting, some experts requested that this Principle be reworded as two separate Principles, like the original Setubal version, because the new approach does not sufficiently emphasize the requirement for external validation. Other participants thought that the single Principle was better for allowing regulatory acceptance flexibility. The author, considering her way to deal with QSAR demonstrating .[17]

PRINCIPLES OF QSAR (Quantitative Structure Activity Relationship): -

If there are no nonlinear dependences of transport or binding on physicochemical properties, all QSAR analyses assume that the various structural properties or features of a compound make linear additive contributions to its biological activity. Dedicated investigations, such as the scoring function of the de novo drug design program LUDI (Eq^{n} 1)[18,19,20], demonstrate this straight forward assumption; moreover, the aftereffects of numerous Free Wilson and Hansch investigations support this idea.

$$\Delta G_{binding} \triangleq \Delta G_0 + \Delta G_{hb} + \Delta G_{ionic} + \Delta G_{lipo} \quad \Delta G_{rot} \qquad (1)$$

Reduction in translational and rotational entropy as a whole, ΔG_0 = 5.4 kJ mol^{-1} Ideal hydrogen bond that is neutral, ΔG_{hb} – -4.7 kJ mol^{-1} Contact with liposomes ΔG_{lipo} — -0.17 J mol^{-1} A^{-2} Loss of entropy per ligand's rotatable bond $\Delta G_{binding}$ with constant term .

THE GENERAL PLAN OF A QSAR STUDY :-

Can be broken down into three categories: extracting descriptors from molecular structure, selecting those that are informative in the context of the analyzed activity, and finally using the values of the descriptors as independent variables to define a mapping that correlates them with the activity in question. These chemoinformatic methods are used to build QSAR models. These phases are realized by the typical QSAR system, as shown in Fig. 1

214 *Combinatorial Chemistry & High Throughput Screening, 2006, Vol. 9, No. 3* **Dudek et al.**

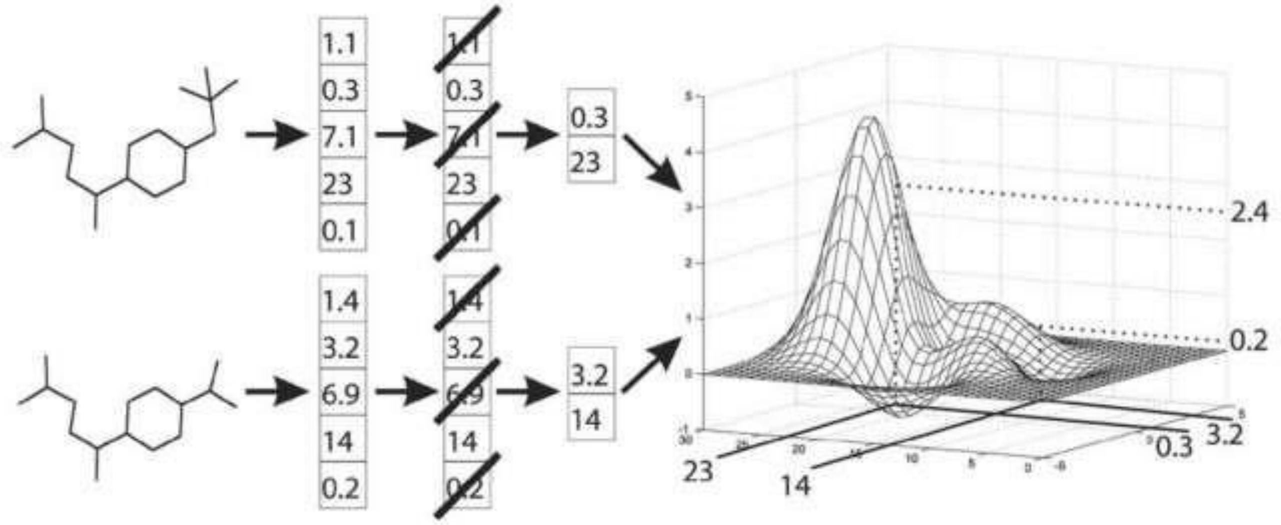


Fig. (1). Main stages of a QSAR study. The molecular structure is encoded using numerical descriptors. The set of descriptors is pruned to select the most informative ones. The activity is derived as a function of the selected descriptors.

- ***Generation of Molecular Descriptors from Structure :-***

Structure-Based Creation of Molecular Descriptors In any case, the construction cannot be straight forwardly utilized for making structure-action mappings because of reasons originating from science and software engineering. In the first place, the synthetic designs do not normally contain in an express structure the data that connects with movement. The structure must be deconstructed in order to extract this data. Different judiciously planned sub-atomic descriptors emphasize different compound properties verifiable in the design of the atom. Only those properties might have a stronger relationship with the activity. These properties include geometric and topological characteristics as well as physicochemical and quantum-chemical ones.

- ***Selection of Relevant Molecular Descriptors :-***

There are numerous applications that can generate hundreds or even thousands of distinct molecular descriptors. In most cases, only a few of them have a significant relationship with the activity. Additionally, many of the descriptors are correlated with one another. Several aspects of QSAR analysis suffer as a result. Some statistical techniques require a significantly greater number of compounds than descriptors. Large datasets would be required for the use of large descriptor sets. Even though other methods can handle datasets with high descriptor-to-compound ratios, they lose accuracy, especially for compounds that are not seen during model preparation. Enormous number of descriptors likewise influences interpretability of the last model.

MOLECULAR DESCRIPTORS :-

The structure of the compound is mapped by molecular descriptors into a set of numerical or binary values that are thought to be important for explaining activity. Based on how much they depend on information about the molecule's 3D orientation and conformation, there are two broad families of descriptors.

- **2D QSAR Descriptors :-**

The broad family of descriptors used in the 2D-QSAR approach all possess the characteristic of being independent of the compound's 3D orientation. These adjectives include straightforward measures of the molecule's constituent parts, its topological and geometrical properties, computations of electrostatic and quantum-chemical descriptors, and advanced fragment-counting techniques.

✓ ***Constitutional Descriptors :-***

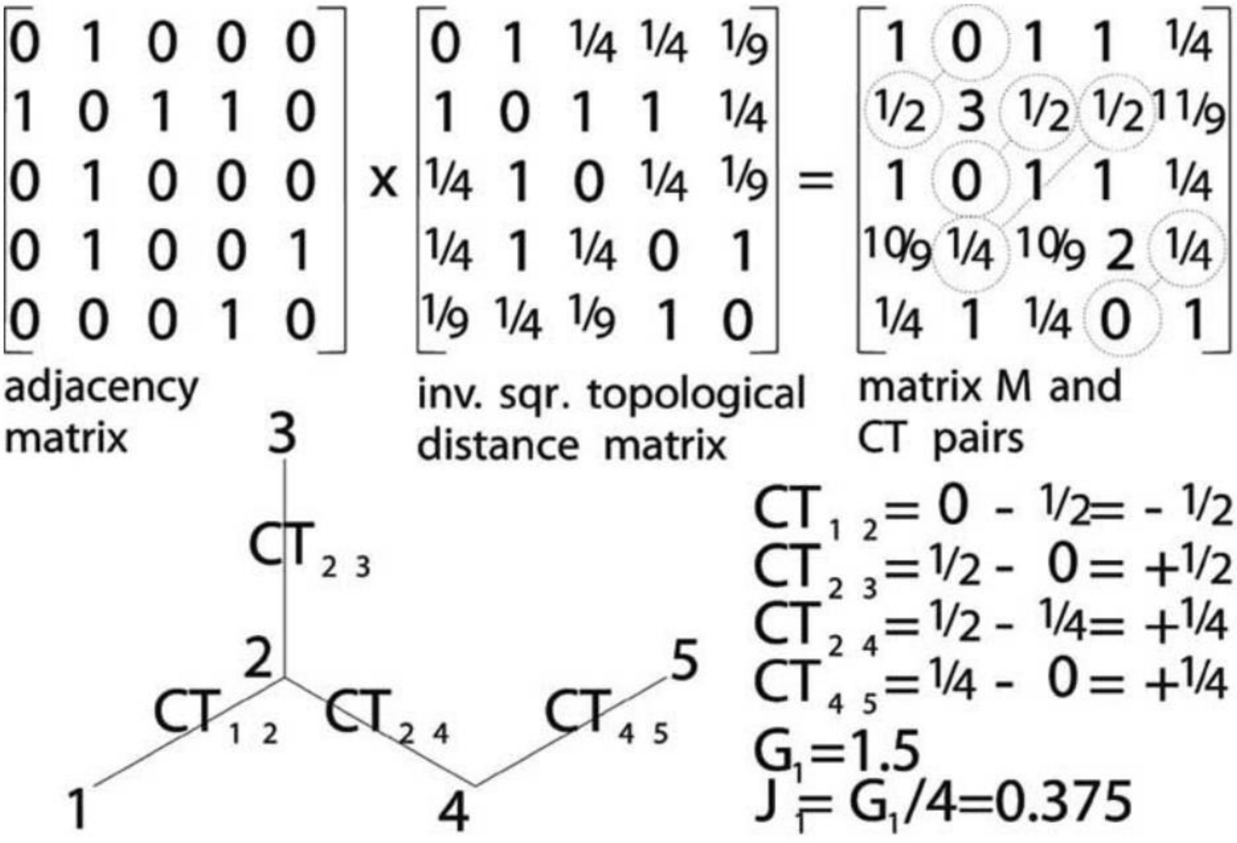
Constitutional descriptors describe the molecule's properties in relation to the elements that make up its structure. The computation of these descriptors is quick and simple. The molecular weight, the total number of atoms in the molecule, and the number of atoms with different identities are all examples of constitutional descriptors. Additionally, several bond-related properties, such as the total number of aromatic rings and single, double, or triple bonds, are utilized**.**

✓ ***Electrostatic and Quantum-Chemical Descriptors :-***

Electrostatic descriptors capture information on electronic nature of the molecule.[21] These include descriptors containing information on atomic net and partial charges.[22] Descriptors for highest negative and positive charges are also informative, as well as molecular polarizability. [23] Partial negatively or positively charged solvent-accessible atomic surface areas have also been used as informative electrostatic descriptors for modeling intermolecular hydrogen bonding.[24] Energies of highest occupied and lowest unoccupied molecular orbital form useful quantum chemical descriptors , in addition to the derivative quantities like absolute hardness [25]

✓ ***Topological Descriptors :***

The structure of the compound is described by the topological descriptors as a graph, with atoms serving as vertices and covalent bonds as edges. The Wiener index [26], which counts the total number of bonds in the shortest paths between all pairs of non-hydrogen atoms, was the foundation upon which many indices quantifying molecular connectivity were established. Randic indices x [27], which are defined as the sum of geometric averages of the edge degrees of atoms within paths of given lengths, Balaban's J index [28], and Shultz index [29] are additional topological descriptors. Topological descriptors like the Galvez topological charge indices or the Kier and Hall indices xv [30]can include information about valence electrons. The initial ones make use of geometric



averages of valence connectivity that follow paths. The latter takes measurements of atoms' topological valences and the net transfer of charges between pairs of atoms separated by a certain number of bonds. The derivation of topological indices is illustrated in Fig.2, we demonstrate how the Galvez indices for a single bond's atom distance can be calculated.

Fig. 2 :- Example of the Galvez first-order topological charge indices G_1 and J_1 for isopentane. The matrix product M_{ij} of atom adjacency matrix and topological distance matrix defined as inverse squares of inter-atom distances, is used to define the charge terms CT_{ij} as $M_{ij} – M_{ij}$. The G_k indices are defined as the algebraic sum of absolute values of the charge terms for pairs of atoms separated by $_k$ bonds. The J_k indices result from normalization of G_k by the number of bonds in the molecule .

- **3D QSAR Descriptors :-**

Compared to the 2D-QSAR approach, the 3D-QSAR method requires significantly more computational power. In most cases, obtaining numerical descriptors for the compound structure requires several steps. The compound's conformation must first be determined using experimental data or molecular mechanics, and then it must be refined by minimizing energy [31,32]. Then, the conformers in dataset must be consistently adjusted in space. Finally, a computational search for various descriptors is carried out in the space with immersed conformer. Additionally, methods that are not dependent on the compound alignment have been developed.

✓ ***Alignment-Dependent 3D QSAR Descriptors :-***

The information about the receptor for the modeled ligand is very important for the group of methods that need to align molecules before descriptors can be calculated. The alignment can be guided by studying the receptor-ligand complexes if such data are available. Otherwise, the structures in space must be superimposed using only computational methods [33,34]. Atom-atom or substructure-substructure mapping are two examples of these approaches.

✓ ***Alignment-Independent 3D QSAR Descriptors :-***

Molecular translation and rotation in space-invariant 3D descriptors are another category. As a result, no compound superposition is required.

- ***The 2D- Versus 3D-QSAR Approach :-***

In drug design, it is generally believed that 3D approaches are superior to 2D ones. However, research shows that this assumption may not always be correct. Due to the dependence of the quality of the outputs on the orientation of the rigidly aligned molecules on the user's terminal, conventional CoMFA's results may, for instance, frequently be non-reproducible [35,36]. Even though some solutions to these alignment issues have been proposed, it is still challenging to precisely align structurally distinct molecules in three dimensions. Additionally, when alignment-independent descriptors are considered, the distinction between 2D and 3D QSAR methods is not always clear. When comparing the BCUT to the WHIM descriptors, this is clear. Both utilize a comparative logarithmic strategy, i.e., settling an eigenproblem for a lattice depicting the compound – the availability lattice in the event of BCUT descriptors and covariance lattice of 3D co-ordinates if there should arise an occurrence of Impulse . A deeper connection exists between 3D-QSAR and the topological approach, a 2D method. It is since a compound's topology frequently determines its geometry. Estrada et al. provided an elegant illustration, who demonstrated that topological indices can predict the dihedral angles of biphenyl as a function of the substituents attached to it [37]. Along the equivalent line, an evidently normally 3D property, chirality, has been anticipated utilizing chiral topological files [38], built by bringing a satisfactory load into the topological lattice for the chiral carbons .

MAPPING THE MOLECULAR STRUCTURE TO ACTIVITY :-

Given the chose descriptors, the last move toward building the QSAR model is to infer the planning between the action and the upsides of the elements. Methods that are straightforward but effective model the activity as a linear function of the descriptors. This method is extended to more intricate relations by non-linear methods. The nature of the activity variable is the basis for yet another significant division among the mapping methods. Predicting a continuous value presents a regression challenge. The classification problem arises when only certain classes of activity need to be predicted, such as partitioning compounds into active and inactive forms. As previously mentioned, the dependent variable is modeled in regression as a function of the descriptors. A decision defines the final model in a classification framework. boundary in the descriptor space that divides the classes. The ways to deal with QSAR planning are portrayed in Fig. 3

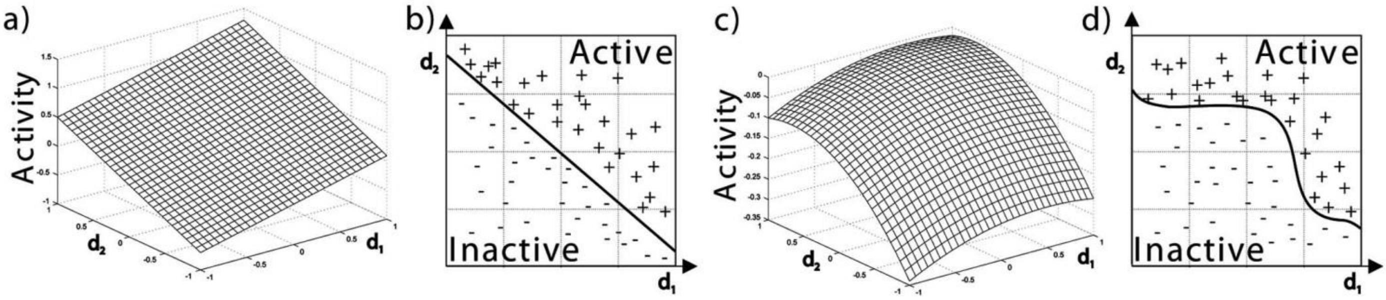


Fig. (**3**). Approaches to QSAR mapping: a) linear regression with activity as a function of two descriptors, d1 and d2, b) binary classification with linear decision boundary between classes of active (+) and inactive (-) compounds, c) non-linear regression, d) non-linear binary classification .

- **Linear Models :-**

Linear models have been the basis of QSAR analysis since its beginning. They predict the activity as a linear function of molecular descriptors. In general, linear models are easily interpretable and sufficiently accurate for small datasets of similar compounds, especially when the descriptors are carefully selected for a given activity .

✓ ***Multiple Linear Regression :-***

Multiple Linear Regression (MLR) models the activity to be predicted as a linear function of all descriptors. Based on the examples from the training set, the coefficients of the function are estimated. These free parameters are chosen to minimize the squares of the errors between the predicted and the actual activity. The main restriction of MLR analysis is the case of large descriptors-to-compounds ratio or multicollinear descriptors in general. This makes the problem ill-conditioned and makes the results unstable. Multiple linear regression is among the most widely used mapping methods in QSAR in last decades. It has been employed in conjunction with genetic description selection for modeling GABAA receptor binding, antimalarial activity, HIV-1 protease inhibition and glycogen phosphorylase inhibition [39], exhibiting lower cross-validation error than partial least squares, both using 4D-QSAR fingerprints. MLR has been applied to models in predictive toxicology [40,41], Caco-2 permeability [42] and aqueous solubility [43]. In prediction of physicochemical properties [44,45] and of COX-2 inhibition [46], MLR proved significantly worse than non-linear support vector machine, yet comparable or only slightly inferior to RBF neural network. However, in studies of logP [47], it proved worse than other models, including multi-layer perceptron and Decision Tree.

✓ ***Partial Least Squares :-***

MLR's issues with multicollinear or over-abundant descriptors can be resolved using partial least squares (PLS) linear regression [48,49]. Even though there are a lot of descriptors, the method assumes that only a small number of latent independent variables control the modeled process. By dividing the input matrix of descriptors into two parts the scores and the loadings—the PLS aims to learn about the latent variables indirectly**.** The scores are orthogonal and, while being able to capture the descriptor information, allow also for good prediction of the activity. The estimation of score vectors is done iteratively. The first one is derived using the first eigenvector of the activity descriptor combined variance-covariance matrix. Next, the descriptor matrix is deflated by subtracting the information explained by the first score vector. The resulting matrix is used in the derivation of the second score vector, which followed by consecutive deflation, closes the iteration loop. In each iteration step, the coefficient relating the score vector to the activity is also determined. PLS has been used successfully with 3D QSAR and HQSAR, such as in a study of estrogen receptor binding and nicotinic acetylcholine receptor binding modeling [50,51]. It has also been used in a study [52] with multiple datasets, including blood-brain barrier permeability, toxicity, P-glycoprotein transport, reversal of multidrug resistance, CDK-2 antagonism, COX-2 inhibition, dopamine binding, and log D. The results were generally superior to those of decision trees, but they were typically inferior to those of ensembles, SVM, and the most recent methods. PLS regression was put to the test in order to predict COX-2 inhibition , but it performed worse than a neural network or a decision tree. PLS, on the other hand, performed better than a neural network in a study of solubility prediction [53]. Melting point and logP prediction models based on PLS have been identified in studies [54]. Finally, PLS models have been developed for BBB permeability [55,56], mutagenicity, toxicity, inhibition of tumor growth and anti-HIV activity [57], and aqueous solubility [58,59].

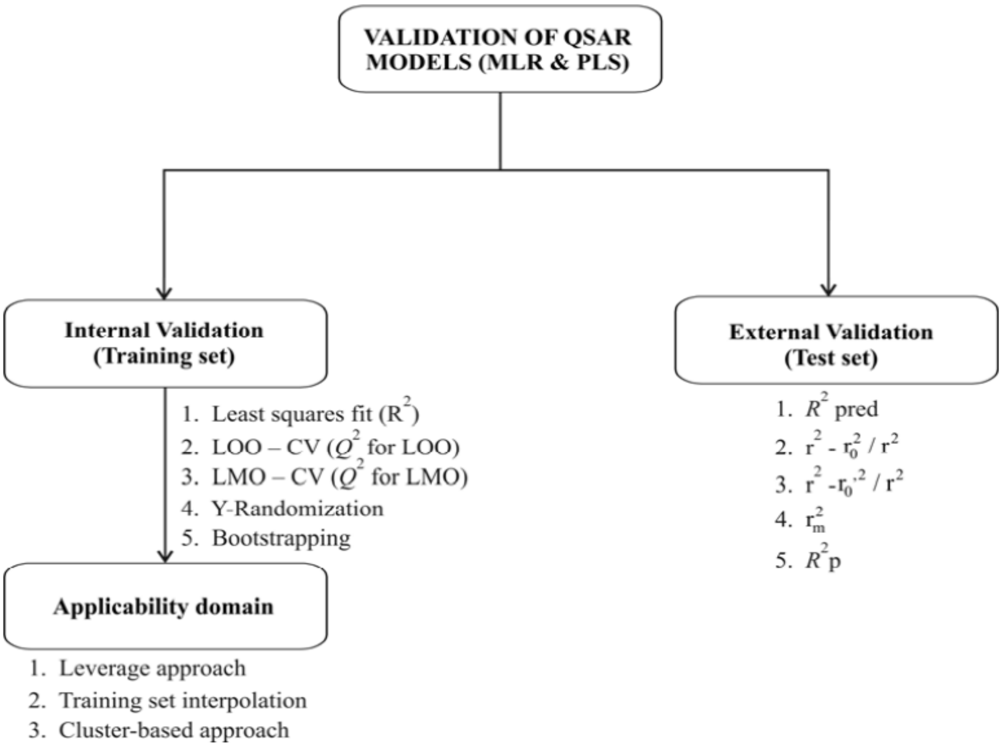
IMPORTANCE OF VALIDATION OF QSAR MODELS

The prediction of the activities of untested chemicals is one of the many uses for the QSAR models.[60] Utilization of structure-activity relationship methods is crucial to the success of drug discovery efforts .In QSAR studies, numerous methods, algorithms, and techniques have been discovered and implemented over time. [61] The selection of the set of descriptors that best describes the most important structural and physicochemical characteristics associated with activity presents the primary obstacle in QSAR. An important first step in the QSAR modeling process is the effective selection of descriptors or variables as well as the quality of biological data. [62] The statistical significance and predictive power of QSAR models determine their application. The capacity of a particular QSAR model to accurately predict unknown chemicals is critical to regulatory decisions, justifications, and application. [63,64] If the QSAR model that was developed is not validated, it may result in erroneous predictions of biological activity. Therefore, the most crucial aspect of QSAR research, following model development, is validation of QSAR models. [65] Clearly, QSAR has matured, but it still has a long way to go. In QSAR modeling, estimating predictions' accuracy is a crucial issue25. Validation of QSAR models has only received significant attention in this decade. [66,67,68,69]

VALIDATION METHODS FOR QSAR MODELS

Validation techniques are required to validate a model's predictive ability using unobserved data and to assist in determining an equation's appropriate complexity. The QSAR model can be validated by either using the data that were used to create the model (an internal method) or using a different set of data (an external method) (Fig. 4). Cross validation (Q2) and least squares fitting (R2) [70,71] , bootstrapping and scrambling (Y-Randomization), adjusted R2 adj), chi-squared test (2), and rootmean-squared error (RMSE) [72,73] are internal approaches to model validation. External validation, such as testing the QSAR model on a test set of compounds, is the best approach. These are statistical techniques that are used to make sure that the models that are made are accurate and reliable (a "good model"). Confirming the model as a "good model" is crucial because a poor model can cause more harm than good.

Fig. 4


VALIDATION OF QSAR MODELS (MLR & PLS)
Internal Validation (Training set)
1. Least squares fit (R2)
2. LOO – CV (Q2 for LOO)
3. LMO – CV (Q2 for LMO)
4. Y-Randomization
5. Bootstrapping
External Validation (Test set)
1. R2 pred
2. r2 - r02 / r2
3. r2 -r0'2 / r2
4. rm2
5. R2p
Applicability domain
1. Leverage approach
2. Training set interpolation
3. Cluster-based approach

INTERNAL VALIDATION

Least Squares Fit :-

Least squares fitting is the most common internal method for validating the model. The R2 (squared correlation coefficient) is used to compare the predicted and experimental activities using this validation method, which is like linear regression. The robust straight-line fit is a better way to figure out R2 because it gives less weight to data points that are farther from the center of the data (that is, data points that are a certain standard deviation from the model) in the calculation. An option in contrast to this technique is the evacuation of exceptions (compounds from the preparation set) from the dataset in an end favor to streamline the QSAR model and is just substantial if severe factual principles are followed. The fact that the number of descriptors in the QSAR model is acceptable when the difference between the R2 adj value and the R2 adj value is less than 0.3 If the difference is greater than 0.3, the number of descriptors cannot be considered acceptable.

$$R^2 = \left[\frac{N\Sigma XY - (\Sigma x)\,(\Sigma Y)}{\sqrt{([N\Sigma X^2 - (\Sigma X)^2]\,[N\Sigma Y^2 - (\Sigma Y)^2])}} \right]^2$$

Fit of the Model :-

Attack of the QSAR models still up in the air by the techniques of chi-squared (□2) what is more, root-mean squared mistake (RMSE). These methods are used to determine whether the model has the predictive quality shown by the R2 value. The utilization of RMSE shows the blunder between the mean of the trial esteems and anticipated exercises. The difference between the predicted and experimental bioactivities can be seen in the chi squared value:

$$\chi 2 = \sum_{i=1}^{n} \left(\frac{(y_i - \hat{y}_i)^2}{\hat{y}_i} \right)$$

$$\text{RMSE} = \text{Sqrt} \left(\sum_{i=1}^{n} \frac{(\hat{y}_i - y_m)^2}{n - 1} \right)$$

Where, y and are a compound's predicted and experimental bioactivity in the training set, ym is the mean of the experimental bioactivities, and n is the number of molecules in the data set being studied.

Even though the model has a large R2 value (0.7), the model's poor ability to accurately predict bioactivities is reflected in its large chi-square or RMSE values (0.5 and 1.0, respectively). The chi and RMSE should be low (0.5 and 0.3, respectively, for a good predictive model). In addition to assisting in the creation of models, these methods of error checking are particularly helpful in the creation and validation of models for nonlinear data sets, such as those produced by Artificial Neural Networks (ANN).[74] However, excellent R2 values, RMSE values, and 2 values alone are not sufficient indicators of the validity of the model. Accordingly, elective boundaries should be given to demonstrate the prescient capacity of models. In principle, there are two reasonable validation strategies that could be considered: one that is based on prediction and the other that is based on how predictor variables fit rearranged response variables.

Cross-validation :-

Cross-validation, also known as CV, Q2, q2, or jack-knifing, is a common technique for internally validating a QSAR model. CV process rehashes the relapse commonly on subsets of information. In most cases, the missing molecule is excluded once (only) at a time, and the predicted values of the missing molecule are used to calculate the R. Occasionally, more than one molecule is left out at a time (leave many out, or LMO). A lot of the time, CV is used to figure out how big a model can be for a given set of data. For a QSAR equation, the cross-validated R2 is typically lower than the overall R2. It is utilized as a diagnostic instrument to evaluate an equation's predictive power.

CV was used to determine whether a model had been overfitted and measure its predictive power. Over-fitting occurs when a predictive model accurately describes the relationship between predictors and response but fails to accurately predict new compounds. When the R2 value from the original model is significantly larger than the Q2 value (difference between R2 and Q2 should not be greater than 0.3), overfitting of the model is typically suspected. [75] Values of the CV are thought to be more indicative of the model's ability to predict. Therefore, in the case of R2, CV is regarded as a measure of prediction goodness but does not fit. The CV procedure begins with the removal of one or more compounds from the training set, which serves as a temporary test set. Using the descriptors from the original model, a CV model is built from the remaining data points and tested on the removed molecules to see if it can correctly predict the bioactivities. The process of removing a molecule and creating and validating the model against the individual molecules is carried out for the entire training set in the leave-one-out (LOO) method of CV. After everything is done, the mean of all the Q2 values is taken and reported. An enhanced training set of the compounds (data points) used to figure out R2 is the data used to get Q2. It is thought that the method of removing just one molecule from the training set is inconsistent.

Bootstrapping :-

Bootstrapping is yet another method of internal validation in which samples from the data set are chosen at random. Sub samples of the data are repeatedly analyzed in the simplest form of bootstrapping, as opposed to subsets of the data. Each sub-sample is a randomly selected sample with replacements drawn from the entire sample. A typical bootstrap validation generates K groups of size n by randomly selecting n objects repeatedly from the initial data set. While some of these items may never be selected, others may be included in the same random sample multiple times. The target properties for the excluded samples are predicted using the model constructed from n randomly selected objects. A high average Q2 during the bootstrap validation demonstrates the robustness of the model.

EXTERNAL VALIDATION

According to several authors, the only way to estimate a QSAR model's true predictive power is to compare the predicted and observed activities of an external test set of (enough) compounds that were not used in the model's development. [76,77,78,79] How can we select the training and test set in external validation? Roy et al. In one of their articles, they discussed clearly how we can solve this problem. [80] To gauge the prescient force of a QSAR model, Golbraikh and Tropsha suggested utilization of the accompanying factual qualities of the test set . [81]

CONCLUSION

In order to comprehend the model's reliability for predicting a novel compound that is not included in the data set, validation of QSAR models is an extremely crucial component. If we consider a total of one thousand QSAR models that have been reported, only 50 to 60 of them are predictive, but it is uncertain whether all of the conditions and validation parameters discussed in this article have been adhered to by these 60 models. We believe that both internal and external validation strategies are important, and to check the model's robustness, one should use all available validation strategies. Only few All of the validation characteristics outlined in this article65,66 were observed in reported QSAR models. In conclusion, the chemical space of the training and test sets must be analyzed in addition to the recommendations for the validation of QSAR models by various scientists and researchers; In terms of congeneric character and structural similarity, real outliers must be identified and eliminated. Even in that case, making predictions using QSAR models is still a risky process. To select reliable predictive QSAR models, we still require appropriate validation methods.[82,83]

REFERENCE :-

1. Kubinyi, H. (1997). QSAR and 3D QSAR in drug design Part 1: methodology. *Drug discovery today*, *2*(11), 457-467.
2. Dudek, A. Z., Arodz, T., & Gálvez, J. (2006). Computational methods in developing quantitative structure-activity relationships (QSAR): a review. *Combinatorial chemistry & high throughput screening*, *9*(3), 213-228.
3. Gramatica, P. (2007). Principles of QSAR models validation: internal and external. *QSAR & combinatorial science*, *26*(5), 694-701.
4. Veerasamy, R., Rajak, H., Jain, A., Sivadasan, S., Varghese, C. P., & Agrawal, R. K. (2011). Validation of QSAR models-strategies and importance. *Int. J. Drug Des. Discov*, *3*, 511-519.
5. Ramsden, C.A., ed. (1990) Quantitative Dntg Deszgn (Comprehenszve Medicinal Cbemistr39 (Vol. 4), Fergamon Press
6. van de Waterbeemd, H., Testa, B. and Folkers, G., eds (1997) Compute~= assisled Lead Finding and Optimization (Proceedings of the IF ~' Earopean Symposium on Quantitative Structure-Activity Relationships: Lausanne, 1996), Verlag Helvetica Chimica Acta and VCH
7. Ramsden, C.A., ed. (1990) Quantitative Dntg Deszgn (Comprehenszve Medicinal Cbemistr39 (Vol. 4), Fergamon Press
8. Hansch, C. and Leo, A. (1995) Exploring QSAR. Fund~onentals and Applications in Chemist O' and 3ioloRv, American Chemical Society
9. Bleicher, K.H.; Boehm, H.-J.; Mueller, K.; Alanine, A.I. Nat. Rev. Drug Discov., 2003, 2, 369-378.
10. Gershell, L.J.; Atkins, J.H. Nat. Rev. Drug Discov., 2003, 2, 321- 327.
11. Goodnow, R.; Guba, W.; Haap, W. Comb. Chem. High Throughput Screen., 2003, 6, 649-660.
12. Shoichet, B.K. Nature, 2004, 432, 862-865.
13. Stahura, F.L.; Bajorath, J. Comb. Chem. High Throughput Screen., 2004, 7, 259-269.
14. Hansch, C.; Fujita, T. J. Am. Chem. Soc., 1964, 86, 1616-1626.
15. Bajorath, J. Nat. Rev. Drug Discov., 2002, 1, 882-894.
16. Pirard, B. Comb. Chem. High Throughput Screen., 2004, 7, 271- 280
17. J. Jaworska, M. Comber, C. Auer, C. J. van Leeuwen, Environ. Health Perspect. 2003, 111, 1358 – 136
18. Kubinyi, H., ed. (1993) 3D QSAR in Drug Design. Theory, Metbods and Applications, ESCOM Science Publishers
19. B6hm, H-J. (1994)J. Comput.-AidedMol. Design 8, 243-256
20. B6hm, H-J. and Klebe, G. (1996) Angeu'. Chem. 108, 2750-2778, Angew. Chem., Int. Eel. Engl. 35, 2588-2614
21. Mulliken, R.S. J. Phys. Chem., 1955, 23, 1833-1840.
22. Cammarata, A. J. Med. Chem., 1967, 10, 525-552.
23. Stanton, D.T.; Egolf, L.M.; Jurs, P.C.; Hicks, M.G. J. Chem. Info. Comput. Sci., 1992, 32, 306-316.
24. Klopman, G. J. Am. Chem. Soc., 1968, 90, 223-234.
25. Zhou, Z.; Parr, R.G. J. Am. Chem. Soc., 1990, 112, 5720-5724.
26. Wiener, H. J. Am. Chem. Soc., 1947, 69, 17-20.
27. Randic, M. J. Am. Chem. Soc., 1975, 97, 6609-6615
28. Balaban, A.T. Chem. Phys. Lett., 1982, 89, 399-404.
29. Schultz, H.P. J. Chem. Inf. Comput. Sci., 1989, 29, 227-222.
30. Kier, L.B.; Hall, L.H. J. Pharm. Sci., 1981, 70, 583-589
31. Guner, O.F. Curr. Top. Med. Chem., 2002, 2, 1321-1332.
32. Akamatsu, M. Curr. Top. Med. Chem., 2002, 2, 1381-1394.
33. Lemmen, C.; Lengauer, T. J. Comput.-Aided Mol. Des., 2000, 14, 215-232.
34. Dove, S.; Buschauer, A. Quant. Struct.-Act. Relat., 1999, 18, 329- 341
35. Cho, S.J.; Tropsha, A. J. Med. Chem., 1995, 38, 1060-1066.
36. Cho, S.J.; Tropsha, A.; Suffness, M.; Cheng, Y.C.; Lee, K.H. J. Med. Chem., 1996, 39, 1383-1395.
37. Estrada, E.; Molina, E.; Perdomo-Lopez, J. J. Chem. Inf. Comput. Sci., 2001, 41, 1015-1021.
38. de Julian-Ortiz, J.V.; de Gregorio Alapont, C.; Rios-Santamarina, I., Garcia-Domenech, R.; Galvez, J. J. Mol. Graphics Modell., 1998, 16, 14-18.

39. Senese, C.L.; Duca, J.; Pan, D.; Hopfinger, A.J.; Tseng, Y.J. J. Chem. Inf. Comput. Sci., 2004, 44, 1526-1539.
40. Trohalaki, S.; Pachter, R.; Geiss, K.; Frazier, J. J. Chem. Inf. Comput. Sci., 2004, 44, 1186-1192.
41. Roy, K.; Ghosh, G. J. Chem. Inf. Comput. Sci., 2004, 44, 559-567.
42. Hou, T.J.; Zhang, W.; Xia, K.; Qiao, X.B.; Xu, X.J. J. Chem. Inf. Comput. Sci., 2004, 44, 1585-1600.
43. Hou, T.J.; Xia, K.; Zhang, W.; Xu, X.J. J. Chem. Inf. Comput. Sci., 2004, 44, 266-275.
44. Xue, C.X.; Zhang, R.S.; Liu, H.X.; Yao, X.J.; Liu, M.C.; Hu, Z.D.; Fan, B.T. J. Chem. Inf. Comput. Sci., 2004, 44, 669-677
45. Xue, C.X.; Zhang, R.S.; Liu, M.C.; Hu, Z.D.; Fan, B.T. J. Chem. Inf. Comput. Sci., 2004, 44, 950-957
46. Yao, X.J.; Panaye, A.; Doucet, J.P.; Zhang, R.S.; Chen, H.F.; Liu, M.C.; Hu, Z.D.; Fan, B.T. J. Chem. Inf. Comput. Sci., 2004, 44, 1257-1266.
47. Tino, P.; Nabney, I.T.; Williams, B.S.; Losel, J.; Sun, Y. J. Chem. Inf. Comput. Sci., 2004, 44, 1647-1653.
48. Wold, S.; Ruhe, A.; Wold, H.; Dunn, W. SIAM J. Sci. Stat. Comput., 1984, 5, 735-743.
49. Phatak, A.; de Jong, S. J. Chemom., 1997, 11, 311-338
50. Zhang, H.; Li, H.; Liu, C. J. Chem. Inf. Model., 2005, 45, 440-448. [109]
51. Waller, C.L. J. Chem. Inf. Comput. Sci., 2004, 44, 758-765. [110]
52. Svetnik, V.; Wang, T.; Tong, C.; Liaw, A.; Sheridan, R.P.; Song, Q. J. Chem. Inf. Model., 2005, 45, 786-799.
53. Clark, M. J. Chem. Inf. Model., 2005, 45, 30-38.
54. Catana, C.; Gao, H.; Orrenius, C.; Stouten, P.F.W. J. Chem. Inf. Model., 2005, 45, 170-176.
55. Sun, H. J. Chem. Inf. Comput. Sci., 2004, 44, 748-757.
56. Adenot, M.; Lahana, R. J. Chem. Inf. Comput. Sci., 2004, 44, 239- 248.
57. Feng, J.; Lurati, L.; Ouyang, H.; Robinson, T.; Wang, Y.; Yuan, S.; Young, S.S. J. Chem. Inf. Comput. Sci., 2003, 43, 1463-1470.
58. Catana, C.; Gao, H.; Orrenius, C.; Stouten, P.F.W. J. Chem. Inf. Model., 2005, 45, 170-176.
59. Sun, H. J. Chem. Inf. Comput. Sci., 2004, 44, 748-757.
60. Kolossov, E.; Stanforth, R. SAR and QSAR Environ. Res. 2007, 18, 89-100.
61. Roy, P.P.; Roy, K. QSAR Comb. Sci. 2008, 27, 302-313.
62. Walker, J.D.; Jawrska, J.; Comber, J.H.I.; Schultz, T.W.; Dearden, J.C. Environ. Toxicol. Chem. 2003, 22, 1653- 1665.
63. He, L.; Jurs, P.C. J. Mol. Graphics Mod. 2005, 23, 503- 523.
64. Roy, P.P.; Leonard, J.T.; Roy, K. Chemom. Intell. Lab. Sys. 2008, 90, 31-42.
65. Tong, W.; Hong, H.; Xie, Q.; Shi, L.; Fang, H.; Perkins, R. Curr. Comput. Aided Drug Des. 2005, 1, 195-205
66. Golbraikh, A.; Tropsha, A. J. Mol. Graphics Mod. 2002, 20, 269-276.
67. Aptula, A.O.; Jeliazkova, N.G.; Schultz, T.W.; Cronin, M.T.D. QSAR Comb.Sci. 2005, 24, 385-390.
68. He, L.; Jurs, P.C. J. Mol. Graph. Mod. 2005, 23, 503-523
69. Worth, A.P.; Leeuwen, C.J.; Hartung, T. SAR QSAR Environ. Res. 15 (2004) 331-343
70. A.R. Leach. Molecular modeling: Principles and applications. Pearson Education Ltd. Harlow, England, 2001.
71. Besal, E. J. Math. Chem. 2001, 29, 191-195.
72. Wold, S.; Ericksson, L. Partial least squares projections to latent structures (PLS) in chemistry. In *Encyclopedia of computationalchemistry*, Ragu & Schleyer, P. (ed.), John Wiley & Sons, Chichester, 1998, Vol. 3, 2006–2021.
73. Yasri, A.; Hartsough, D. J. Chem. Inf. Comput. Sci. 2001, 41, 1218-1227.
74. Fan, Y.; Shi, L.M.; Kohn, K. W.; Pommier, Y.; Weinstein, J.N. J. Med. Chem. 2001, 44, 3254-3263
75. 2001
76. Kubinyi, H.; Hamprecht, F.A.; Mietzner, T. J. Med. Chem. 1998, 41, 2553-2564
77. Guha, R.; Jurs, P.C. J. Chem. Inf. Model. 2005, 45, 65-73.
78. Novellino, E.; Fattorusso, C.; Greco, G. Pharm. Acta Helv. 1995, 70, 149-154.
79. Zefirov, N.S.; Palyulin, V.A. J. Chem. Inf.Comput. Sci. 2001, 41, 1022-1027.
80. Roy, P.P.; Roy, K. QSAR Comb. Sci. 2008, 27, 302-313
81. Golbraikh, A.; Tropsha, A. J. Mol. Graphics Mod. 2002, 20, 269-276.
82. Ravichandran, V.; Shalini, S.; Sundram, K.M.; Dhanaraj, S.A. Eurp. J. Med. Chem. 2010, 45, 2791-2797.
83. Roy, P.P.; Paul, S.; Indrani, M.; Roy, K. Molecules. 2009, 14, 1660-1701.