This chapter explains the proposed model to predict biological properties of certain chemicals.

### 3.1 INTRODUCTION

This chapter explains the model for prediction using decision support system for investigation of microbicidal activity of certain chemical compounds. The main objective of this research work is to identify the best suitable QSAR model to predict the biological properties of certain chemicals.

### 3.2 BASIC MODEL

The basic structure of model for studies on computational models to predict the biological activity of certain chemical is as given in figure 3.1

![Figure 3.1 - Basic Structure of Proposed Model](image)

The model takes series of chemical structures as input. A chemical structure includes molecular geometry, electronic structure and crystal structure of molecules. Molecular geometry refers to the spatial arrangement of atoms in a molecule and the chemical bonds that hold the atoms together[86-87].
All chemical substances need to be tested in terms of their toxicological and environmental properties before their use. Toxicology is concerned with the study of the adverse effects of chemicals on living organisms. It also studies the harmful effects of chemical, biological and physical agents in biological systems that establish the extent of damage in living organisms. The relationship between dose and its effects on the exposed organism is of high significance in toxicology. Certain factors are included in toxicity like dosage, route of exposure, the species, age, gender and environment. To perform prediction it is required to use data mining methods like clustering, classification, conceptual clustering, Inductive learning, summarization, regression, case-based learning.

Using QSAR and regression analysis a decision support system is designed and developed to relate a set of structural descriptors of a chemical compound to its biological activity. Advantages of constructing this model are: very fast, often free, reduce the number of animals used in experiments. Figure 3.2 shows an approach to QSAR.
From the above diagram, by using QSAR anyone can remove IN VITRO and IN VIVO process, alternative path comes into existence by selecting QSAR technique.

For proposed model chemical structure is an input. On this input need to apply pre-processing technique to generate homogeneous structure/data which will ready for mining in terms of descriptor selection from existing knowledge base and using it decision can be made in terms of prediction of biological activity of certain chemicals.

Quantitative structure–activity relationship models (QSAR models) are regression or classification models used in the chemical and biological sciences and engineering.

### 3.2.1 Knowledge Base

A knowledge base is a special kind of database for knowledge management that provides the means for the computerized collection, organization and retrieval of knowledge. To predict the biological activity of certain chemical, the model uses information stored in knowledge base. The knowledge base is storage of knowledge in the database about a specific subject of interest.

Using knowledge base and data mining techniques like QSAR and regression analysis, the proposed decision support system investigates microbial activity of certain chemical compounds and makes prediction.

Knowledge base design for specific purpose like input in the form of chemical structure and predictive outcome or model generation through a decision support system of QSAR.

Existing knowledge applied on the mined chemical structures is used for a prediction. By supplying various suitable parameters, multiple predicted
models are generated among which most suitable model is used as an outcome which is new knowledge in the form of predicted model.

### 3.3 Purpose of QSAR

Quantitative structure-activity relationships (QSARs) attempt to quantify the relationship between an aspect of chemical structure and an activity or property imparted by that structure. Chemical structure is often described by descriptors (e.g., electrophilicity, hydrogen bonding, and molecular fragments) or physical-chemical properties (e.g., Log P) which are then used to develop a mathematical correlation between a group of structures and a defined activity or endpoint.

Correlations usually take the form of statistical algorithms developed through a variety of techniques (e.g., univariate, regression, multiple linear regressions, partial least squares analysis).

The QSAR model should meet the requirements of the OECD principles [89]:
A defined endpoint, an unambiguous algorithm, a defined domain of applicability, appropriate measures of goodness-of-fit, robustness and predictively; a mechanistic interpretation if possible.

Testing the predictive performance of a (Q)SAR tool on chemicals that are similar to the pesticide in question and have empirical data available for them can provide another source of information for evaluating the reliability of predictions. For an instance, a starting point for testing the predictive performance of a (Q)SAR tool for a sodium salt of an organic acid would be to generate a prediction for a de-salted acid form of the compound for which empirical data are available [88]. Which chemicals to use would depend on the type and quality of empirical data available for them, the parameter used
to assess similarity (e.g., physical-chemical parameters, structure, metabolism) and the degree of similarity.

### 3.4 DATA ANALYSIS METHODS

Different Statistical Methods have been used in QSAR for the extraction of useful information from the data.

![Data Analysis Method Diagram](image)

**Figure 3.3 Data Analysis Method**

[Studies On Computational Model To Predict The Biological Properties Of Certain Chemicals](#)
Data analysis is the process of evaluating data using analytical and logical reasoning to examine each component of the data provided. This form of analysis is just one of the many steps that must be completed when conducting a research experiment.

Data from various sources is gathered, reviewed, and then analysed to form some sort of finding or conclusion. There are a variety of specific data analysis method, some of which include data mining, text analytics, business intelligence, and data visualizations.

Various physicochemical 2D and 3D properties of the molecules shall be computed and correlated with the observed biological activity data through systematic multiparameter statistical regression analysis, to evolve out meaningful mathematical equations of Quantitative Structure Activity Relationships (QSAR).

Partial least square (PLS) is a method for constructing predictive models when the factors are many and highly collinear. Note that the emphasis is on predicting the responses and not necessarily on trying to understand the underlying relationship between the variables.

3.5 DESIGNING OF QSAR MODEL

Chemicals have various effects on organisms to which they are exposed, some of which are desirable and some are undesirable. Nowadays, the number of these chemicals is rapidly increasing, in pace with the fast industrial development [90], that is why it is extremely important to assess their potential toxicological effects on organisms and the environment.

Two main streams have been developed in order to explain the complex relationships between molecules and observed quantities, or endpoints. The first one is related to the search for relationships between molecular structures and physicochemical properties and is called QSPR (Quantitative Structure Property Relationships).
Structure-Property Relationships)[91-93]. The second one, which is the focus of our work, is related to the search for relationships between molecular structures and biological activities and is called QSAR (Quantitative Structure-Activity Relationships).

Since chemical structure was elucidated, the relationship between chemical structure and biological activity has intrigued scientists. It has been recognized that the investigation of QSARs may provide useful tools for obtaining information regarding the effects of chemicals on man and the environment [94-96]. Initially developed to assess the value of drugs, QSARs are now proposed as a method to assess general toxicity.

QSARs are based on the assumption that the structure of a molecule (its geometric, steric and electronic properties) contains the features responsible for its biological activity. For example, as already explained in the previous sections, biological activity can be expressed quantitatively as in chemistry, toxicology and QSAR: an introduction the concentration of a substance required to give a certain biological response. When the information encoded
in the molecular structure is expressed by molecular descriptors in the form of numbers, one can form a quantitative structure-activity relationship between the two [97-99]. By QSAR models, the biological activity of a new or untested chemical can be inferred from the molecular structure of similar compounds whose activities have already been assessed.

QSAR's most general mathematical form is:

It is therefore evident that the three key components required for the development of a QSAR model are:

- Some measure of the activity (in this case toxicity) for a group of chemicals in a biological or environmental system – toxicological endpoint
- A description of the physicochemical properties and/or structure for this group of chemicals
  - molecular descriptors
- A form of statistical relationship to link activity and descriptors

Figure 3.5 - QSAR model visualization: building graphical models that relate biological activity of molecules to their structure

At first sight, the selection of compounds for developing QSAR models may appear to be self-evident. If we are interested in the biological effects of a certain group of chemicals we collect or measure all the compounds of that group that can be found. However, this strategy is not the best way to gather data and it may happen that too many results for the wrong
compounds prevent the establishment of a good QSAR model. The successful construction of QSAR models requires experimental design, in which each compound included corresponds to a design point and the experimental factors that need to be varied in order to create the design are the physicochemical properties that characterize the compounds. The toxicological endpoint can also be an experimental factor and the goal is to develop a model that links the endpoint to the physicochemical descriptors. It is crucial that the design includes compounds that give both high and low values of the endpoint of interest, and if possible, a uniformly-spread range of intermediate values.

The response data, which are measures of the biological activity of compounds and represent the output variables in the QSAR models, can be measured directly by the investigators or collected from the literature. Knowledge of the precision and range of these data is of high importance. Some measurements have a natural range, but others may cover many orders of magnitude which may be deceptive. It is therefore dangerous to take these data at face value. Examination of their distribution can be very useful because it can indicate where a certain type of processing is required. The precision is another property of interest because the model should have a standard error no better than the measurement errors. This is because of the fact that it should not be possible to calculate something more precisely that it can be measured. A standard error that is better (less) than the experimental one is a good indication that the model has been over fitted, which means that it fits the training data set well, but cannot generalize to other sets, which is the purpose for fitting a model.

The descriptor data, which capture information about the chemical structure of compounds and represent the input variables in the QSAR models, can be obtained from a variety of sources. In the early period of QSAR modelling, the choice of the descriptors was limited because
they were generally tabulated physicochemical properties. Nowadays, there are over 3000 different molecular descriptors and it is common to use many more descriptor variables than there are compounds in the set when building a QSAR model. This leads to the need for dimension reduction, variable elimination and variable selection, which are different techniques for reducing the complexity of a problem in order to be able to recognize useful and informative patterns in the data [100]. Dimension reduction is the process of reducing the number of random variables under consideration and is usually performed by a mathematical procedure called Principal Component Analysis (PCA) in which new variables called principal components are created from linear combinations of the original variables. Variable elimination is the process by which unhelpful or unnecessary variables are removed from a data set [101]. Common procedures for variable elimination are Corchop and unsupervised forward selection. Even after eliminating unnecessary variables from a data set, there may still be many variables to choose from when building a model. In this case variable selection is used, whose aim is to choose descriptors that will be useful in some sort of mathematical model and will lead to a model that will generalize to other unseen compounds. There are many diverse procedures for variable selection and some are built in to the process of model building, such as the forward stepping multiple regression.
In this context, it must be mentioned that one of the major problems in QSAR modelling is the availability of high quality experimental data for building the models. The input data must be both accurate and precise in order to develop a meaningful model. Any developed QSAR model is statistically as valid as the data that led to its development [32].

In addition to this, a problem related to molecular descriptors is their reproducibility: experimental values can differ greatly even when referred to the same compound. As an illustration, several approaches have been developed for the theoretical calculation of the partition coefficient (log P), but in these calculations it is not uncommon to have differences of several orders of magnitude[102-103]. In modern QSAR approaches, it is common to use a wide set of theoretical molecular descriptors of different kinds which take into account the various features of the chemical structure. There are many software packages that calculate wide sets
different theoretical descriptors. The greatest advantage of theoretical descriptors is the fact that they can be calculated homogeneously by defined software for all chemicals, including those not yet synthesized but represented by a hypothesized chemical structure, and therefore they are reproducible.

A variety of methods for building QSAR models exists. These methods are called pattern recognition methods because their aim is to devise algorithms that could learn to distinguish patterns in a data set. They can be classified as supervised (for example, Multiple Linear Regression, Discriminant Analysis, Partial Least Squares, Classification and Regression Trees, Neural Networks, etc.) or unsupervised (for example, Principal Component Analysis, Cluster Analysis, k-Nearest Neighbours, Nonlinear Mapping, etc.), where supervision refers to the use of the response data which are being modelled. Unsupervised learning makes no use of the response, meaning that the algorithms seek to recognize patterns in the descriptor data only. The advantage of unsupervised learning is the lower likelihood of chance effects, due to the fact that the algorithm is not trying to fit a model. On the other hand, supervised learning does use the response data and care needs to be taken to avoid chance effects [104]. Another significant difference between supervised and unsupervised learning methods is the ratio of compounds (p) to variables (n) in a data set. Some supervised learning techniques may not work due to failure to invert a matrix, while others may give a false, apparently correct, classification. Even though this is not a problem for unsupervised methods, the presence of extra variables that have no useful information may obscure meaningful patterns.

The nature of the response data that they are capable of handling is another important feature of modelling methods. In this context, there are two types of methods: methods that deal with classified responses (for example, mutagen / not mutagen, toxic / slightly toxic / non-
toxic) and methods that handle continuous data (the response is a potency of an end-point)[105]. For the modelling of categories, a wide range of classification methods exists, including: Discriminant Analysis, k-Nearest Neighbours (KNN), Classification and Regression Trees (CART), Support Vector Machine, etc. For the modelling of continuous data, the most widely used method is Multiple Regression Analysis (MRA), a simple approach that leads to a result that is easy to understand. MRA is a powerful means for establishing a correlation between independent variables (molecular descriptors) and a dependent variable (biological activity). In addition, Artificial Neural Networks can be used for modelling both classified and continuous data.

After the model is developed, regardless of the type, it is of crucial importance to assess its performance by validating its predictive application. Most statistics packages generate a variety of statistical quantities for the common modelling approaches which will enable a judgement of significance and will give some guidance on whether the model may have arisen by chance[106]. This is based on the assumption that the data conform to some statistical distribution, usually multivariate normal. Unfortunately, this only indicates how well the model fits the data within the modelling assumptions and does not really give any information on how well the model might work. The best fit models are not the best ones for prediction. As a consequence, the only way to know how well a model may work is to try it out.
Chapter 3: Proposed Model

Figure 3.7 - Depiction of the recursive process for developing QSAR models

One common approach is the Leave-One-Out Cross Validation (LOO or CV), which involves leaving out one compound, fitting the model to the remainder of the set, making a prediction for the left out compound and repeating the process for each of the compounds in the set. A variety of statistics can be generated using this procedure, for example LOO R2 (called Q2) and a predictive residual sum of squares (PRESS). The disadvantage of LOO is that only a small part of the data set is omitted and if outliers occur in pairs or groups they will not be identified [108]. A better approach is to leave out some larger portion of the set (10% or 20%) and to repeat this a number of times. This allows the generation of a set of predicted values for the compounds so that estimates may be made of the likely errors in prediction. The disadvantage of this approach is that it is computationally intensive and suffers from a combinatorial explosion as the sample size is increased.

However, none of these procedures allows us to judge whether a relationship is real or it has happened by chance. One way to check for chance effects is to scramble the response values and then try to build models using the scrambled data. This can be repeated a number of times and some fit statistics, such as R2, can be tabulated for the resulting...
Chapter-3 : Proposed Model

models. If the R2 value for the model of the unscrambled response is higher than the R2 value for the scrambled sets, it is reasonable to assume that the model is not a chance fit.

Figure 3.8-QSAR model validation by using a training set and a test set

Obviously, the best test of a model is to present it with unseen data, either by holding back some of the original data to form a test set or by synthesizing or testing some more compounds once the model has been built. Only a stable and predictive model can be considered a reliable model and can be usefully interpreted for its mechanistic meaning.

There is an argument that, if the main aim of QSAR modelling is simply prediction, the attention should be focused on model quality and it is not necessary to try to interpret models. Another argument is that it is dangerous to attempt to interpret models, since correlation does not imply causality. Regarding the interpretability of QSAR models, Livingstone states: “The need for interpretability depends on the application, since a validated mathematical model relating a target property to chemical features may, in some cases, be all that is necessary, though it is obviously desirable to attempt some explanation
of the “mechanism” in chemical terms, but it is often not necessary”[108]. On this basis, we can differentiate predictive QSARs, where the focus is best prediction quality, from descriptive QSARs, where the focus is descriptor interpretability.

To summarize all that was previously mentioned, the ideal QSAR model should:

1. Consider an adequate number of molecules for sufficient statistical representation,
2. Have a wide range of quantified end-point potency (for example, several orders of magnitude) for regression models or adequate distribution of molecules in each class (for example, active and inactive) for classification models,
3. Be applicable for reliable predictions of new chemicals (validation and applicability domain) and
4. Allow to obtain mechanistic information about the modelled end-point.

3.6 SUMMARY

This chapter explains the developed model for prediction using decision support system for investigation of microbicidal activity of certain chemical compounds. Also QSAR model visualization, the process of QSAR modeling for protecting the biological activity of novel compounds, depiction of the recursive process for developing QSAR models and QSAR model validation by using a training set and a test set are given systematically.