Towards new computational tools for predicting toxicity
TOWARDS NEW COMPUTATIONAL TOOLS FOR PREDICTING TOXICITY

SWAPNIL CHAVAN

LINNAEUS UNIVERSITY PRESS
Abstract


The toxicological screening of the numerous chemicals that we are exposed to requires significant cost and the use of animals. Accordingly, more efficient methods for the evaluation of toxicity are required to reduce cost and the number of animals used. Computational strategies have the potential to reduce both the cost and the use of animal testing in toxicity screening. The ultimate goal of this thesis is to develop computational models for the prediction of toxicological endpoints that can serve as an alternative to animal testing. In Paper I, an attempt was made to construct a global quantitative structure-activity relationship (QSAR) model for the acute toxicity endpoint (LD50 values) using the Munro database that represents a broad chemical landscape. Such a model could be used for acute toxicity screening of chemicals of diverse structures. Paper II focuses on the use of acute toxicity data to support the prediction of chronic toxicity. The results of this study suggest that for related chemicals having acute toxicities within a similar range, their lowest observed effect levels (LOELs) can be used in read-across strategies to fill gaps in chronic toxicity data. In Paper III a k-nearest neighbor (k-NN) classification model was developed to predict human ether-a-go-go related gene (hERG)-derived toxicity. The results suggest that the model has potential for use in identifying compounds with hERG-liabilities, e.g., in drug development.

Keywords
Acute toxicity, chronic toxicity, hERG, k-NN, LD50, LOEL, Munro database, QSAR, read-across, toxicity
Table of Contents

List of publications........................................................................................................... 5
Additional published work outside the scope of this thesis ............................................. 6
Popular science summary ................................................................................................. 9
Abbreviations ................................................................................................................... 11
Chapter 1: Introduction ................................................................................................. 12
  1.1 Biological activity testing ......................................................................................... 12
  1.1.1 Pharmaceutical drug testing ............................................................................... 13
  1.1.2 Environmental toxicant testing .......................................................................... 13
  1.1.3 Food safety testing ............................................................................................. 14
  1.2 Dose-response relationships .................................................................................. 14
  1.2.1 How drugs act ..................................................................................................... 14
  1.2.2 Drug targets ....................................................................................................... 15
  1.2.3 Drug-receptor interaction ................................................................................... 16
  1.2.4 Dose-response curve .......................................................................................... 17
  1.3 Structure-activity relationships ............................................................................. 17
  1.3.1 Introduction of bio-isosteres ............................................................................. 18
  1.3.2 Modification of size and shape ......................................................................... 19
  1.3.3 Introduction of new substituents ....................................................................... 19
  1.4 Quantitative structure-activity relationships ....................................................... 19
  1.4.1 Brief history ....................................................................................................... 19
  1.5 Current state of QSAR and related techniques ..................................................... 20
  1.5.1 Evolution of molecular descriptors ................................................................... 20
  1.5.2 Types of QSARs ............................................................................................... 21
  1.5.3 Latest techniques for developing QSAR ............................................................ 21
  1.5.4 Latest applications of QSARs .......................................................................... 21
  1.5.5 Chemical category approach ............................................................................ 21
  1.5.6 Data gap filling and Read-across ...................................................................... 22
  1.6 Statistical methods and tools ................................................................................ 23
  1.6.1 Multiple linear regression (MLR) ...................................................................... 23
  1.6.2 Principal component analysis (PCA) ................................................................. 23
  1.6.3 k-nearest neighbor (k-NN) .............................................................................. 24
  1.7 Objectives.............................................................................................................. 27
Chapter- 2: Acute toxicity prediction using QSAR technique ....................................... 28
  2.1 Research questions and objectives ....................................................................... 29
  2.2 Summary of work .................................................................................................. 29
  2.2.1 Munro database ............................................................................................... 29
  2.2.2 Methodology ..................................................................................................... 30
  2.2.3 Results and discussion ...................................................................................... 32
  2.3 Conclusion ............................................................................................................ 38
Chapter-3: Acute toxicity - supported chronic toxicity prediction ................................ 39
  3.1 Research questions and objectives ..................................................................... 40
LIST OF PUBLICATIONS

This thesis is based on the following papers, which are referred to in the text by their Roman numerals. All the published papers are reproduced with the permission from the respective publishers.


Additional published work outside the scope of this thesis


To my family
POPULAR SCIENCE SUMMARY

We are exposed to many chemical substances in our everyday life. This exposure stems from a variety of sources such as our homes, workplaces, food we consume, drug therapies we may be taking and through industrial wastes. The need to assure our safety from exposure to chemicals with adverse effects requires information about the toxicity of the numerous substances we are exposed to, however, such evaluations are currently performed through animal testing and are very expensive.

In addition to this, millions of new drugs are being synthesized annually by pharmaceutical industries worldwide, which demands the use of thousands of animals for the purpose of safety and efficacy evaluations during preclinical testing. Furthermore, the global production of chemicals by agricultural, biotechnological, chemical, cement, fertilizer, metal, polymers, textile, etc., industries amounts up to billions of tons. Based on available information, the toxicological evaluations of these chemicals would need about 54 million animals in European Union and about 11.5 million animals in the USA during this decade for testing purposes. Thus, large animal resources, time, efforts and costs will be required to meet the need of toxicological evaluations of all these chemicals.

To save animals and speed up the toxicological evaluations, development of new non-animal-testing methods and models is required. To date, many non-animal-testing methods are available, among which computational methods that can relate structural features of chemical substances with a particular function or biological activity have been shown to be specially promising.

Generally, a structure activity relationship (SAR) model can be developed using the following steps. (1) Collection of biological/toxicological data for a group of chemicals. (2) Description of chemicals in the form of numerical representations i.e. in the form of descriptors. (3) Derivation of a relationship between the descriptors and biological/toxicological endpoints through
application of mathematical/statistical techniques.

In this thesis, relationships between chemical structure and biological activity have been explored using computational tools to develop mathematical models correlating structure and activity with the goal of producing general models for predicting acute toxicity and chronic toxicity, as well as a model for identifying toxicity derived from human Ether-a-go-go-Related Gene (hERG)-derived cardiotoxicity.
## ABBREVIATIONS

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADME</td>
<td>Absorption, Distribution, Metabolism and Excretion</td>
</tr>
<tr>
<td>CAS</td>
<td>Chemical Abstracts Service</td>
</tr>
<tr>
<td>CDK</td>
<td>Chemistry Development Kit</td>
</tr>
<tr>
<td>CV</td>
<td>Cross Validation</td>
</tr>
<tr>
<td>EC50</td>
<td>Effective Concentration 50%</td>
</tr>
<tr>
<td>ECHA</td>
<td>European Chemical Agency</td>
</tr>
<tr>
<td>ER</td>
<td>Error Rate</td>
</tr>
<tr>
<td>FDA</td>
<td>Food and Drug Administration</td>
</tr>
<tr>
<td>GA</td>
<td>Genetic Algorithm</td>
</tr>
<tr>
<td>GHS</td>
<td>Globally Harmonized System</td>
</tr>
<tr>
<td>hERG</td>
<td>Human Ether-à-go-go-Related Gene</td>
</tr>
<tr>
<td>HESS</td>
<td>Hazard Evaluation Support System</td>
</tr>
<tr>
<td>IUPAC</td>
<td>International Union of Pure and Applied Chemistry</td>
</tr>
<tr>
<td>k-NN</td>
<td>k-Nearest Neighbor</td>
</tr>
<tr>
<td>LD50</td>
<td>Lethal Dose 50%</td>
</tr>
<tr>
<td>LOEL</td>
<td>Lowest Observed Effect Level</td>
</tr>
<tr>
<td>MACCS</td>
<td>Molecular ACCess System</td>
</tr>
<tr>
<td>MOA</td>
<td>Mode Of Action</td>
</tr>
<tr>
<td>NEDO</td>
<td>New Energy and Industrial Technology Development Organization</td>
</tr>
<tr>
<td>NER</td>
<td>Non Error Rate</td>
</tr>
<tr>
<td>NOEL</td>
<td>No Observed Effect Level</td>
</tr>
<tr>
<td>OECD</td>
<td>Organisation for Economic Co-operation and Development</td>
</tr>
<tr>
<td>PCA</td>
<td>Principal Component Analysis</td>
</tr>
<tr>
<td>QSAR</td>
<td>Quantitative Structure-Activity Relationship</td>
</tr>
<tr>
<td>RDT</td>
<td>Repeated Dose Toxicity</td>
</tr>
<tr>
<td>REACH</td>
<td>Registration, Evaluation, Authorization and Restriction of Chemicals</td>
</tr>
<tr>
<td>RTECS</td>
<td>Registry of Toxic Effects of Chemical Substances</td>
</tr>
<tr>
<td>SAR</td>
<td>Structure-Activity Relationship</td>
</tr>
<tr>
<td>SMILES</td>
<td>Simplified Molecular-Input Line-Entry System</td>
</tr>
<tr>
<td>TdP</td>
<td>Torsades De Pointes</td>
</tr>
</tbody>
</table>
CHAPTER 1: INTRODUCTION

The biological activities of large numbers of chemicals, such as pharmaceutical drugs, food substances, environmental and agricultural products and industrial chemicals, need to be determined to ensure safety. This is a time consuming and expensive activity that has the potential to be improved by adopting computational techniques. Several \textit{in silico} toxicology and pharmacology tools are being developed for assisting in the profiling of toxic substances in order to investigate their mode of action (MOA), risk assessment and safety studies. The work presented here focuses on different computational approaches for assessing biological/toxicological activity. This thesis reveals quantitative structure-activity relationship (QSAR) studies for the prediction of acute toxicity, chronic toxicity and human Ether-a-go-go-Related Gene (hERG)-derived toxicity.

1.1 Biological activity testing

Humans are exposed to many chemicals and substances throughout their lifespan.\textsuperscript{3} These exposures occur from different sources like the food supply, drug therapy, living environment, workplaces and industrial wastes. Thus there is a need to ensure human safety from all the chemicals which humans are exposed to. To meet this demand, it is necessary to perform a biological evaluation of all these chemicals and substances. The major aim of the biological evaluation of chemicals is to identify their MOA, potency, and to monitor human and environmental safety.

Bioassay is a method that determines the biological activity of a substance by measuring its effect on an organism and comparing it with the activity of an agreed standard. In other words, bioassay is defined as the estimation of the concentration or potency of a substance by measurement of the biological response that it produces.\textsuperscript{4} The main uses of bioassay are:

- to estimate the pharmacological activity of a substance,
- to explore the functions of endogenous mediators, and
to determine the toxicological and unwanted effects of new or undefined substances for monitoring human or environmental safety

1.1.1 Pharmaceutical drug testing
Thousands of new drugs are synthesized yearly by pharmaceutical industries worldwide in order to find breakthroughs for many different human diseases and disorders. All these drugs need to undergo rigorous animal testing before being approved for human trials.\textsuperscript{5,6} Animal testing involves pharmacokinetic and pharmacodynamic profiling of drugs. Pharmacokinetics investigate new drugs' absorption, distribution, metabolism and excretion (ADME) while pharmacodynamics investigate their biological activity (efficacy) and toxicity.\textsuperscript{7}

The efficacy testing of a new drug is investigated by testing its strength in curing the induced illness of interest in the test animal. This preclinical study involves the acute, sub-acute and chronic toxicity testing. The acute toxicity testing is performed to study rapid poisoning, whereas sub-acute toxicity testing investigates whether any toxic metabolite of the drug has been formed over time. The chronic toxicity investigates the toxic effect of a drug over prolonged (up to the total lifespan of the test animal) and repeated exposures.\textsuperscript{8,9} After successfully passing preclinical tests, the drug undergoes human trials. In this step, the drug is administered in a double-blind controlled trial that enables clinical practitioners to determine the effect of the drug and its dose-response relationship.\textsuperscript{10-12}

1.1.2 Environmental toxicant testing
Environmental toxicants are defined as chemical or physical agents released into the general environment that can produce adverse health effects among large numbers of people.\textsuperscript{13} Environmental toxicants affect the respiratory tract, the gastrointestinal tract, dermal tissues and other organs, and thereby cause tissue damage and/or loss of structure and function of vital organs. The intensity of such toxic effects is based upon the doses of toxicants. Generally, these toxicants can cause acute toxicity for short-term exposure, but most often they cause chronic effects due to long-term exposures.

Exposure-response relationships can be established for humans, however there already exist many toxicity studies using animals. Therefore, some researchers have recommended extrapolating human toxicities of environmental toxicants from animal bioassay data.\textsuperscript{14} Toxicity testing in animals is dependent upon direct toxicity assessment in whole organisms, where the organisms are exposed to the toxicant(s) of interest, and observed for any sign of adverse health effects. The duration of the exposure depends upon the type of toxicity being examined. The study duration for short-term acute effects can be 96 hours to 14 days, sub-chronic effects from few weeks to few months, and
chronic effects are usually studied for a significant portion of the organism's whole lifespan.\textsuperscript{15}

### 1.1.3 Food safety testing

The food that constitutes our daily diet consists of many directly or indirectly added substances which may or may not be toxic for human health. Among the directly added constituents are chemicals purposefully used in food production, processing and storage that finally end up in the diet, while other chemicals are used for enhancing the taste, colour and flavour of food. Among the indirectly added substances, are the residues of drugs (hormones) and/or feed additives used in animal and milk production, pesticides used for vegetable production, natural toxins and chemo preventives in plants, microbial toxins (bacterial, fungal) in food, and materials used in food processing and packaging. The consequences of these substances are the health problems that can arise, such as allergies, cancer, neurotoxicity, hepatotoxicity, infertility, and acute and chronic poisoning.\textsuperscript{16}

To establish food safety large efforts have been made by toxicologists and epidemiologists. The estimation of the concentration of food additives and adulterants is commonly used to identify their level of daily intake, while the identification of toxins and carcinogenic constituents in food is usually accomplished by the bioassay method and application of the dose-response curve.

### 1.2 Dose-response relationships

The biological activity can be described graphically with respect to the concentration of the drug (or toxicant). The magnitude of such a biological effect can be displayed as a function of the drug concentration in the form of a dose-response curve.\textsuperscript{17,18} Thus, the dose response curve is an important descriptor in understanding drug activity.

#### 1.2.1 How drugs act

According to basic principles of pharmacology, a drug must exert some chemical effect on the cell constituents of a body tissue in order to exhibit a pharmacological/toxicological action. Such a chemical effect can only occur if a drug binds to specific cell constituents.\textsuperscript{19} The binding sites of drugs are known as targets. The pharmacological research is mainly driven by the act of identifying the mechanisms by which a drug associates with its target to exhibit a physiological response. The basic concept of the drug-receptor interaction is described by the “lock and key” mechanism in which a receptor has a complementary binding site for a drug (Figure 1.2.1). If a drug has the correct shape to fit in the receptor binding site, then a key (drug) can open a lock (receptor).
1.2.2 Drug targets

There are several types of body components that serve as drug targets. The term ‘receptor’ is often used to refer to any target molecule with which a drug can interact to elicit pharmacological action.

Figure 1.2.2 Receptors: A) Angiotensin converting enzyme, B) Pho84 phosphate transporter, C) hERG K⁺ channel, D) AT-rich DNA decamer in S. aureus with N,N’-bis[3-(1,4,5,6-tetrahydropyrimidin-2-yl)phenyl]biphenyl-4,4’-dicarboxamide and E) Neuraminidase (in crimson red) on the surface of the influenza virus. Images A, C and D are produced from PDB files 1J37, 1BLS and 4U8A using EduPymol. Image B represents in-house homology model of Pho84, the image has been produced using VMD. Image E is obtained from Virology blog with the kind permission of Prof. Vincent Racaniello.
The dopamine receptors in the brain are receptors for antipsychotic drugs; the $K^+$ channels are receptors for antiarrhythmic drugs (e.g. dofetilide); the angiotensin-converting enzyme is the receptor for vasodilators (e.g. captopril); the proton-coupled phosphate transporter makes major contribution to phosphate transport in *S. cerevisiae*; DNAs are receptors for different compounds (e.g. an AT-rich DNA decamer in *S. aureus* is inhibited by bisamidine compounds) and surface glycoprotein, like neuraminidase, is a receptor for antiviral drugs (e.g. zanamivir) (Figure 1.2.2).

### 1.2.3 Drug-receptor interaction

Most drugs have a high specificity towards binding to a particular receptor. Conversely, a receptor that acts as a drug target often has a high degree of selectivity for a particular drug. This complementary specificity of a receptor binding site towards a ligand, defines the molecular recognition property of a receptor. But any change in a receptor, such as removal of one or more amino acids, can alter the shape of the complementary binding site, and make the receptor inactive for a particular ligand.

**Figure 1.2.3** Distinction between affinity and efficacy. $K_{d1}$ and $K_{d2}$ are rate constants for the binding reaction while $\alpha$ and $\beta$ are rate constant for the receptor activation reaction.

Generally, the binding of a ligand to a receptor may lead to a physiological response through the activation or inactivation of that receptor. The binding of a drug to a receptor is known as 'drug affinity'; while if a receptor elicits a tissue response upon drug binding, it is known as 'drug efficacy' (Figure 1.2.3).
An agonist is a drug that has high affinity as well as high efficacy, and an antagonist has high affinity, but zero efficacy.  

1.2.4 Dose-response curve 

The binding capacity of a drug can be measured directly, but to estimate its biological response, such as contraction or relaxation of muscles, activation or inhibition of enzymes or changes in membrane potential or heart rate, the pharmacologists often need to study the drug with the help of a dose-response curve. A dose-response curve is a simple plot that relates the amount of the drug (or toxicant, pollutant, stressor, radiation) with the biological response. The dose of a drug is often plotted on the X-axis, while its response is plotted on the Y-axis.  

![Dose-response curve](image)

**Figure 1.2.4 Dose-response curve describing the therapeutic effect of a drug.**

Figure 1.2.4 shows a typical dose-response curve, which is plotted on a semi-logarithmic scale and is characteristically sigmoidal in shape. A curve is characterized by a threshold, maximal effect, and sub-maximal effect ($EC_{50}$). The first point that indicates that a response rises above the zero response level is known as the threshold concentration. The desired response of a drug can be seen as above the threshold concentration level. The maximal attainable response of a drug is known as the ceiling effect. The drug concentration that elicits a halfway response between the baseline and the ceiling effect is referred to as the $EC_{50}$ (Effective Concentration 50%).

1.3 Structure-activity relationships 

The biological effect is governed by the structure of a ligand. The relationship between the structure of the ligand and the biological activity is termed as a structure-activity relationship (SAR). The presence of key structural elements,
such as a particular functional group, fragment or sub-structure in a ligand, often results in a certain variation in its biological activity. The analysis of the SAR of a lead compound and its analogs may be used to determine the significance of parts of the structure of the lead compound that result in increased biological activity and reduced unwanted side effects. The SAR can be used to develop a new drug with improved activity and less side effects.

SARs are usually investigated by creating slight modifications to the lead compound to produce new analogs and measuring the influence of these structural changes on their biological activity. These structural changes can be brought about in a number of ways as presented below.

1.3.1 Introduction of bio-isosteres

New analogs to existing lead compounds can be derived by replacing an existing structural moiety of a lead with a new moiety. The choice of the new moiety is often based on the concept of isosteres. Isosteres are the chemical groups that exhibit similarity in some of their chemical or physical properties as a result of having the same number of total or valence electrons while not necessarily having the same number of atoms (Figure 1.3.1). It is expected that being similar in structures, isosteres often exhibit similar pharmacokinetic and pharmacodynamic properties, however isosteric analogs do not show a similar type of activity in every case.

![Figure 1.3.1: Examples of isosteres, where each row represents isosteric groups.](image)

-CH₃  -NH₂  -OH

-CH₂-  -NH-  -O-

![Chemical structures](image)

-CH₃  -NH₂  -OH

-CH₂-  -NH-  -O-

![Chemical structures](image)

-CH₃  -NH₂  -OH

-CH₂-  -NH-  -O-

![Chemical structures](image)
1.3.2 Modification of size and shape
The structural changes in a lead compound can be brought about through the modification of size and shape of the lead. For example, this can be accomplished by altering the chain length or size of the ring, changing the number of double and triple bonds, and adding or removing ring moieties.

1.3.3 Introduction of new substituents
A new substituent can be introduced in order to replace previously existing substituents or to occupy previously unsubstituted positions in a lead. Through the incorporation of a new substituent, the pharmacokinetic and pharmacodynamic properties of the lead compound can be improved. The choice of substituent is mainly driven by the objective of enhancing a certain property of an analog over a lead compound, such as improving solubility, enhancing permeation across cell membranes or reducing the rate of metabolism.

1.4 Quantitative structure–activity relationships
As discussed previously, structural variation influences the biological activity of a particular class of chemicals, which provides a path to discover the structure–activity relationships. Studying these relationships statistically and computationally has paved the path for the establishment of computational predictive techniques; QSAR is one such technique.

QSAR is a mathematical relationship between the biological activity and the physicochemical parameters in the form of an equation. Generally, these parameters are representative of properties, such as lipophilicity, electronic effects, steric effects, and chemical compositions. These properties are defined in the form of numerical data (also known as descriptors) that are obtained experimentally or can be calculated using computer programs. For a set of compounds such properties can be measured or calculated, and can be related with their biological activity by means of mathematical equations using different statistical methods such as regression analysis.29

1.4.1 Brief history
In 1865 Crum-Brown and Fraser proposed that the structural modification in a series of poisons led to significant differences in their actions. Based on this, they postulated that the physiological action of a molecule is a function of its chemical constitution and can be expressed as:

\[
\text{Physiological action} = \text{function (chemical constitution)} \quad (1.4.1)
\]
Hansch et al. in 1962 published a study of the influence of the Hammett constant and hydrophobicity on the structure-activity relationship of plant growth regulators.

\[ \text{Log } 1/C = a(\log P)^2 + b \log P + c\sigma + \ldots + k \]  

Where,
- \( C \) = concentration that produces biological effect
- \( P \) = octanol/water coefficient
- \( a, b, c \) = coefficients
- \( k \) = Y-intercept
- \( \sigma \) = electronic Hammett constant

The Free-Wilson approach proposed that no physicochemical parameters are required for the structure activity relationship; instead the contribution of each structural feature was of interest. The simplest equation for a biological response (BR) follows:

\[ \log \text{BR} = \Sigma \text{(substituent contributions)} + \text{contribution from base molecule} \]  

1.5 Current state of QSAR and related techniques

Today’s QSAR models are much more advanced, as they are augmented with different graphical approaches, high dimensional descriptors, and studied in tandem with advanced mathematical equations and high computing resources.

1.5.1 Evolution of molecular descriptors

The descriptors represent structures of molecules and their physicochemical properties. The molecular description of a compound is explained by constitutional descriptors. The 1-D descriptors are derived by counting structural fragments, such as the number of primary and secondary carbon atoms, amine and nitro groups, etc. The more advanced descriptors, like 2-D and 3-D, represent topological, geometrical, sterochemical and other kind of information about compounds. Topological descriptors explain the bonding collection in a molecule. Geometrical descriptors reveal information related to the shape and size of compounds, while electrostatic descriptors explain information about e.g. molecular charge, polarizability and topological polar surface area. Even more complex and computationally expensive descriptors, such as quantum chemical descriptors, describe information related to the electronic structure. Recently 4-D and 5-D descriptors have been developed. These descriptors have considered an ensemble of conformations for each ligand and induced fit phenomena that take place after ligand-receptor interaction, respectively.
1.5.2 Types of QSARs
The QSAR can be classified based on various factors, such as the mathematical technique used (e.g. regression and classification), the property to be investigated (e.g. quantitative structure-property relationship, group-based QSAR, fragment-based QSAR and pharmacophore-similarity-based QSAR,) and type of descriptors used (e.g. 2D-QSAR, 3D-QSAR, 4D-QSAR and multidimensional QSAR).

1.5.3 Latest techniques for developing QSAR
Various techniques are used for the construction of QSARs. Common examples of regression analysis include partial least square (PLS), multiple linear regression (MLR), k-nearest neighbor (k-NN) regression and principal component regression (PCR) (See section 1.6). Common examples of classification methods are the k-nearest neighbor (k-NN) classification, discriminant analysis and classification trees. The most commonly used 3D-QSAR techniques are comparative molecular field analysis (COMFA) and comparative molecular similarity index analysis (COMSIA). In the newly emerging field of multidimensional QSAR, multiple conformations of the ligand, induced fit effects of the ligand-receptor complex and the solvation energy of the ligand-receptor interaction have been incorporated to closely mimic biological processes. Expert systems composed of a group of diverse models combined with different databases (e.g. the JRC QSAR model database) have been recently developed.

1.5.4 Latest applications of QSARs
There are a large number of applications of QSAR models within industry, academia, and governmental and regulatory agencies. A few of the uses are listed below:
- Prediction of chemical properties, e.g. melting point, boiling point, logP.
- Prediction of biological activity, e.g. inhibitory concentration of drug (IC50), EC50, LD50, NOEL, mutagenicity.
- Prediction of ADME properties of new chemical entities in drug discovery.
- Risk management of chemicals by regulatory authorities.
- Prediction of the environmental behaviour of organic pollutants.
- Prediction of the chromatographic retention of a chemical.
- Assessment of the nanotoxicity of nanomaterials.

1.5.5 Chemical category approach
The European Chemical Agency (ECHA) guidance document on non-animal-testing approaches has endorsed techniques, such as chemical grouping, read-across and weight-of-evidence (WOE) along with QSARs, for
extending and elaborating the existing information so as to improve new testing strategies for predicting human hazards and assessing environmental risks.

A chemical category is defined as a group of chemicals whose toxicological or physicochemical properties are likely to be similar because of their structural similarity. The 'similarity' rationale can be described as chemicals that have:

- a common functional group(s) (e.g. epoxide, aldehyde, etc.)
- incremental or decremental changes in structure (e.g. chain length)
- common chemical classes or constituents
- a common precursor or metabolite

Categories of chemicals are often established with the assumption that a series of chemicals with common/similar structural features will exhibit coherent trends in their physico-chemical properties, and thereby, in their toxicological effects and/or environmental fate properties. These so-called coherent trends in their behaviour are generally associated with a common underlying mechanism of action.

1.5.6 Data gap filling and Read-across

Read-across is a technique that predicts endpoint information (i.e. biological activity or chemical property) of a chemical based on data from the same endpoint from another chemical that is similar in some aspects, like a structural similarity and/or similar properties.

![Figure 1.5.6 Data gap filling using read-across, interpolation or extrapolation from one tested chemical to an untested chemical.](image)

Figure 1.5.6 is an example of a data gap filling for the logP property of a chloro-alkane within the defined category of chloro-alkanes. There is an incremental change in the chain length of members which is associated with an incremental change in their logP. The read-across technique can be effectively used here for the data gap filling.

Interpolation is the estimation of an endpoint for a chemical using measured values from other members on both sides of that chemical within the given
category range. Extrapolation refers to the estimation of an endpoint for a chemical that is near or at the boundary of a given category, by using measured values from internal category members.

1.6 Statistical methods and tools

Various multivariate analysis methods are used for establishing structure-function relationships. There are two types of multivariate analysis available: supervised (MLR, k-NN, etc.) and unsupervised (PCA, decision tree, etc.).

1.6.1 Multiple linear regression (MLR)

A regression analysis correlates independent variables (X) with the dependent variable (Y) where the linear relationship is established between ‘m’ number of Xs and the biological activity data (Y) through a linear equation. Such a relationship can be expressed with a multiple-linear equation:

\[ Y = b_0 + b_1X_1 + b_2X_2 + \ldots + b_mX_m \]  

(1.6.1.1)

Where,

- \( b_0 \) = Y intercept
- \( b_1 \) = estimated slope of a regression of Y on \( X_1 \)

The MLR technique is generally used to identify the best set of independent variables, so that a dependent variable can be predicted as accurately as possible. The best independent variables can be identified with the help of statistical approaches such as backward elimination or stepwise selection.\(^{40}\)

The backward elimination approach removes the least significant variables from a model and refits the model. This procedure continues until the ‘stopping criteria’ are met, such as the highest \( r^2 \) or lowest error, whereas the stepwise selection begins without any variables in a model, and then constructs the model adding new variables until no further variables are significant. It is possible to exclude a variable(s) that is less significant than the variable(s) added later.\(^{41}\)

1.6.2 Principal component analysis (PCA)

Principal component analysis is a method for converting possibly correlated variables into a set of uncorrelated variables by means of orthogonal transformation. The basic idea behind PCA is to reduce the dimensionality of data that contains a large number of interrelated variables, while retaining as much variation as possible in the data. This is achieved by transforming the data into a new set of variables known as principal components (PCs). The PCs are uncorrelated, and are in order i.e. the first PC consists of highest information (explained % variance) than the second PC and so on (Figure 1.6.2).
The PCA was undertaken through analysing of score and loading plots. The score plot enables interpretation of relationships among the samples. If the samples are close, then they are similar to each other and vice-versa. On the other hand the loadings plot enables interpretation of the relationships among variables. The samples placed on the right side of the score plot are characterized by having high values of variables projected on the right side of the loadings plot and vice-versa.

![ PCA plots ]

**Figure 1.6.2** The representation of principal components of PCA in three-dimensional space. PC-1 is a vector that best fits the data in the direction of maximum variance, whereas PC-2 projects perpendicularly to PC-1 in the direction of the second maximum variance in the dataset.

### 1.6.3 $k$-nearest neighbor (k-NN)

#### 1.6.3.1 Background

$k$-NN is a non-parametric method used for regression and classification. It is one of the simplest and fastest instance-based learning techniques, and is among the first choices for a classification model development when there is little or no prior information regarding the distribution of the data. The $k$-NN algorithm was originally proposed by Fix and Hodges in 1951.

#### 1.6.3.2 Definitions and theory

A $k$-NN model is defined by a set of samples for which response variables are known. Each sample from a dataset consists of independent and dependent variables that are continuous or categorical in nature. The models constructed for continuous and categorical types of dependent variables are referred to as $k$-NN regression and $k$-NN classification, respectively.

Estimating the outcome for a given unknown sample (query), can be achieved by calculating the distance matrix, and thereby finding $k$-samples that are closest in distance to the query sample. In the $k$-NN regression model, the outcome for a query sample can be found by averaging the outcomes of its $k$-neighbors, while in the $k$-NN classification, by assigning the category of the majority of its $k$-neighbors (Figure 1.6.3).
For a given query point, the \( k \)-NN algorithm makes predictions based on the outcome of the \( k \) neighbors nearest to that point. Thus, to perform such a prediction, a metric for calculating the distance between the query point and training dataset samples needs to be defined. Two well-known choices to measure this distance are Euclidean, and Manhattan.\(^{44}\) The distance matrix calculations by these methods can be performed using following formulas:

\[
\text{Euclidean distance} = \sqrt{\sum_{i=1}^{k}(x_i - y_i)^2} \tag{1.6.3.2.1}
\]

\[
\text{Manhattan distance} = \sum_{i=1}^{k}|x_i - y_i| \tag{1.6.3.2.2}
\]

To determine the optimal \( k \) value (i.e. the optimal number of neighbors), several methods can be used, such as a risk function, empirical rules and cross validation. The \( k \) value that gives the optimal model statistics is chosen for the construction of a model.\(^{42}\)

1.6.3.3 Advantages and disadvantages

The main advantage of the \( k \)-NN method is that it is the simplest of all machine learning techniques. In addition, it is a sophisticated classification technique\(^{45}\) since it is a non-linear, non-parametric approach and it can handle multi-class dataset. The \( k \)-NN is a sensitive method in its applications to distance matrices and scaling techniques. A disadvantage of the \( k \)-NN method is that it can be affected by the local structure of the dataset.
1.6.3.4 Significance

The main significance of this technique is that it is non-parametric, meaning that it does not form any assumptions on the underlying data distribution. This is important, as in most of the cases, training data does not follow the common theoretical assumptions (e.g., linearly separable). Non-parametric algorithms like $k$-NN are very useful in such cases. Another significance is that it is a lazy algorithm, which makes decisions based on the entire training data set i.e. it does not discard any information from the training data.46

1.6.3.5 Applications

The $k$-NN algorithm has been used in a number of fields,47-49 such as for:

- Biological/toxicological activity prediction of bioactive compounds and toxicants.
- Gene expression data studies.
- Protein-protein interaction studies.
- Predicting 3D structures of proteins.
1.7 Objectives

The objective of this thesis was to develop predictive models for different types of toxicities using computational methods. The work involved the construction of QSAR models for acute (LD$_{50}$), repeated dose (LOEL) and hERG-derived toxicities.

In paper I, the objective was to investigate the possibility of establishing a global QSAR model for the acute toxicity endpoint (LD$_{50}$) by applying the $k$-NN technique using the Munro database that represents a broad set of chemicals that are diverse in terms of their structures and MOAs.

Several MOAs contribute towards the acute toxicity endpoint (LD$_{50}$), and similarly, many effects are studied in order to derive the chronic toxicity endpoint (LOEL). Both these endpoints are influenced by multiple toxicity mechanisms, and are expressed in the same unit *i.e.* milligram of chemical per kilogram of bodyweight (mg/kg). Therefore, in paper II, the objective was to determine whether it is possible to predict the chronic toxicity (LOEL) using acute toxicity (LD$_{50}$) data by means of $k$-NN classification and the read-across approach.

In paper III, the objective was to investigate the possibility of developing a QSAR model by applying the $k$-NN approach to the structurally diverse set of Fenichel and Ochem database compounds that exhibits a unique MOA for hERG K$^+$ channel blockade.
CHAPTER- 2: ACUTE TOXICITY PREDICTION USING QSAR TECHNIQUE

Millions of new chemical substances are synthesized annually in the drug discovery processes of the world’s many pharmaceutical companies. The preclinical testing of these substances demands the use of thousands of animals for the purpose of safety and efficacy evaluations. Moreover, the global production of chemicals by e.g. agricultural, fertilizer, chemical, cement, metal, biotechnological, textile and polymer industries amounts to billions of tons. According to a study at University of Leicester, UK, toxicological evaluation of these chemicals would need additionally about 12 million animals for testing purposes.\textsuperscript{50,51} Furthermore, 99\% of the chemicals that exist in the market today are not subjected to safety evaluation testing, and testing of these chemicals would need a large animal resources, time, efforts and cost.\textsuperscript{50}

To save animals and speed up the toxicological evaluation, there is a need for the development of new non-testing methods and models. Several non-testing methods are available, and among them computational methods are the quickest and cheapest. The QSAR technique is one of the successful computational method that has been applied extensively. Moreover, the REACH annex XI has also envisaged the use of QSAR methods for testing chemicals. Therefore, better QSAR models are needed to evaluate the safety of chemicals that have diverse structures and come from a variety of sources (such as chemical, pharmaceutical, agricultural, polymer and food industries). Thus, we have made an attempt to develop a global QSAR model using the Munro database that consists of a variety of chemicals from pharmaceutical, agricultural, food and environmental fields, so that this model can contribute towards the screening of acute toxicity of unknown compounds of diverse structures and origins.
2.1 Research questions and objectives

The Munro database\textsuperscript{52} represents a wide chemical landscape, \textit{i.e.}, it consists of a variety of pharmaceuticals, agricultural and industrial chemicals, substances used in food production, and chemicals that have an impact on the environment. In this respect, Stocherro \textit{et al.}\textsuperscript{53} have referred to this database as ‘a world of chemicals’. Accordingly, this dataset is of particular interest for computer modelers due to its diverse content. A few computational approaches have been applied on the Munro database, such as the establishment of a toxicological threshold of concern (TTC) using a decision tree approach,\textsuperscript{52} and the application of Principal Component Analysis (PCA), Orthogonal Bidirectional Projections to Latent Structures-Discriminant Analysis (O2PLS-DA) and the clustering approach, for the purpose of classifying the Munro database chemicals according to toxicity.\textsuperscript{53}

The above-mentioned studies incorporated sub-chronic toxicity endpoint data (NOEL) for the classification of the Munro database, while there are no published studies using acute toxicity endpoint data (like LD\textsubscript{50}) on this database. Since the database consists of diverse chemicals, a computational model developed with such a dataset could serve as a better screening tool for predicting toxic potentials of unknown compounds of varied sizes and shapes. Moreover, the benefit of using acute toxicity values over chronic toxicity values is that they are obtained in a short time duration (1–14 days), which can contribute to cost efficiency, and with less cumbersome experiments in comparison to those used for the sub-chronic and chronic toxicity values, which generally take from 28 days to 2 years of study, and involve huge amounts of money and require significant effort.

To construct a model that can be used for the quick screening of unknown compounds of diverse structures and origins in ascertaining their acute toxicity by application of the Globally Harmonized System\textsuperscript{54} of chemical classification, we have made an effort to develop a \textit{k}-NN classification model for the acute toxicity (LD\textsubscript{50}) endpoint using Munro database chemicals, and this constitutes the main objective of this study (\textbf{Paper I}).

2.2 Summary of work

2.2.1 Munro database

The Munro dataset\textsuperscript{52} contains a total of 613 chemicals; all these chemicals were examined and authenticated based on the correct structure, the correct IUPAC name and the correct CAS registry number (RN). The SMILES notations for each structure were carefully checked in ChemSpider,\textsuperscript{55} SigmaAldrich\textsuperscript{56} and PubChem.\textsuperscript{57} The SMILES for cis/trans isomers and R/S enantiomers were carefully inspected and only canonical SMILES were taken
into consideration. Salts and mixtures were removed from the original dataset, as were also duplicate records and records missing either structure or CAS registry number. The chemicals containing diazo or guanidine functionalities were discarded because of the presence of resonance structures that result in different SMILES notations for the same chemical. There were 469 chemicals that had correct CAS, IUPAC name and SMILES notations. The LD$_{50}$ values for all these sorted chemicals were searched in the Toxnet$^{58}$ and RTECS$^{59}$ webservers. Records with more than one value were discarded. Finally, LD$_{50}$ values for 441 chemicals were retrieved from the Toxnet and the RTECS webservers.

2.2.2 Methodology

2.2.2.1 Descriptor calculations:

Two-dimensional Dragon molecular descriptors were used for the model construction. A total of 3668 descriptors were computed for each of the 441 chemicals using the Dragon 6 software.$^{60}$ A filtering of the descriptors was achieved by discarding descriptors with one or more missing values, as well as constant or near constant values. Subsequently, the descriptors with a pair correlation larger than 95% were excluded. A reduced pool of the descriptor set was rendered with 1106 descriptors.

2.2.2.2 Descriptor filtering and outlier detection:

Principal Component Analysis (PCA) was employed to filter descriptors in order to remove irrelevant ones. The PCA study was performed on all 441 chemicals using 1106 descriptors. The data was auto-scaled before the PCA study. Ten principal components were undertaken for the PCA study. To sort the relevant descriptors a loading score threshold of 0.06 was used, which filtered 460 descriptors. These descriptors were considered for further studies and other irrelevant descriptors were removed. In addition, five chemicals were identified as potential outliers in the PCA score plot, and thus they were excluded from further studies. The final dataset considered for the model development comprised of 436 chemicals along with their 460 descriptors.

2.2.2.3 Model development:

2.2.2.3.1 Classification scheme:

The quantitative toxicological response was classified in a qualitative class on the basis of the Globally Harmonized System.$^{54}$ Three classes were formed: Class I: LD$_{50}$ ≤ 300 mg/kg/day (highly toxic); Class II: 300 < LD$_{50}$ ≤ 2000 mg/kg/day (intermediate toxic); and Class III: LD$_{50}$ > 2000 mg/kg/day (low to non-toxic).
2.2.2.3.2 Classification model:

A classification model is the mathematical relationship between a set of descriptors and response variables. The \( k \)-NN classification method was employed to discover the appropriate relationship between molecular structures and the toxicity of chemicals. The \( k \)-NN algorithm is based on the \( k \)-nearest neighbor classification rule as described by Hart et al.\(^6\) In this algorithm, each query chemical is classified according to the classes of its closest neighbors. The closest neighbors are identified on the basis of a distance matrix. Several methods of distance calculations between chemicals on the basis of binary data exist to date.\(^6\) We have selected the ‘Euclidean’ distance method for calculation of the distance matrices. While applying the \( k \)-NN algorithm, the optimal value of \( k \) needs to be determined. We have used the cross validation method for this purpose. A series of \( k \) values were assigned (from \( k=1 \) to 10), and based on the highest NER (non-error rate) and lowest class error, an optimal \( k \) value was identified.

2.2.2.3.3 Descriptor selection by Genetic Algorithms:

Genetic algorithms (GAs) were employed for identifying significant descriptors and for removing non-significant descriptors. GAs are performed on a random population of chromosomes that consist of molecular descriptors. Optimization of a defined fitness function is achieved by simulating the evolutionary process, which gives new chromosomes by combinations with the chromosomes of the initial population by means of genetic operations, such as crossover and mutation. The GAs were optimized on the basis of the non-error rate (NER) which represents the predictive power of the model in correctly classifying chemicals.\(^5,6\)

In order to extract all 1106 descriptors for all 436 chemicals, accomplish the PCA study and to conduct descriptor selection by Genetic Algorithms, we have employed the “ga_toolbox” and “pca” Matlab modules developed at Milano Chemometrics and QSAR Research Group, University of Milano-Bicocca, Milan, Italy.\(^5\)

2.2.2.3.4 Model validation:

The 436 Munro dataset chemicals were randomly split, keeping 80% of the chemicals from every class in the training set and the remaining 20% in the test set (Table 2.2.2.3.4). To select the molecular descriptors and to construct the classification models, the training set was used. The predictive power of this model was evaluated by employing the test set chemicals.
Table 2.2.2.3.4. Classifications of the 436 chemicals in the training and test sets prior to the $k$-NN model construction.

<table>
<thead>
<tr>
<th></th>
<th>Class 1</th>
<th>Class 2</th>
<th>Class 3</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training set</td>
<td>82</td>
<td>136</td>
<td>129</td>
<td>347</td>
</tr>
<tr>
<td>Test set</td>
<td>21</td>
<td>35</td>
<td>33</td>
<td>89</td>
</tr>
<tr>
<td>Total</td>
<td>103</td>
<td>171</td>
<td>162</td>
<td>436</td>
</tr>
</tbody>
</table>

The training set of 347 chemicals with 460 descriptors was assigned for variable selection (by the GA) coupled with $k$-NN classification. The internal validation of models was assessed by 5-fold cross validation; four groups were used for testing the class membership of the omitted group, where the class of the majority of $k$-neighbors was assigned to the member of the omitted group. The best model, constructed on the training set that had a higher NER$_{tr}$ and the lowest class error, was subjected for external validation. The test set comprised of 89 chemicals was used for external validation. Finally, performance of the best model was assessed by means of parameters such as non-error rate (NER), sensitivity, specificity, precision and error rate (ER).

To classify chemicals into the training and test sets, as well as to perform the GA-coupled $k$-NN classification, the “classification toolbox” Matlab module developed at Milano Chemometrics and QSAR research group was used.

2.2.3 Results and discussion
2.2.3.1 Genetic Algorithm outcome:

The GA strategy was applied in order to select significant descriptors, which later were used for the construction of the $k$-NN model. The best $k$-NN model obtained from the GA strategy consisted of 25 descriptors, and was associated with a NER$_{tr}$ of 0.67 and NER$_{te}$ on the training set of 0.66 (see Table 2.2.3.1.2). The $k$ selection with 5-fold cross validation provided an optimal $k$ value of 1, i.e., only the closest chemical was used to predict the class of each target chemical. The 25 descriptors that were used in the $k$-NN classification are shown in Table 2.2.3.1.1.

Table 2.2.3.1.1 Overview of the 25 descriptors derived by the genetic algorithm coupled with $k$-NN classification.

<table>
<thead>
<tr>
<th>Entry</th>
<th>Name</th>
<th>Description</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>MATS1e</td>
<td>Moran autocorrelation of lag 1 weighted by Sanderson electronegativity</td>
<td>2D autocorrelations</td>
</tr>
<tr>
<td>2</td>
<td>SpMAD_B(s)</td>
<td>Spectral mean absolute</td>
<td>2D matrix-based</td>
</tr>
<tr>
<td></td>
<td>Description</td>
<td>Formula/Definition</td>
<td>Type</td>
</tr>
<tr>
<td>---</td>
<td>---------------------------------------------------------------------------</td>
<td>----------------------------------------------------------------------------------</td>
<td>--------------------------------</td>
</tr>
<tr>
<td>3</td>
<td>SpPosA_B(p)</td>
<td>Normalized spectral positive sum from Burden matrix weighted by I-State</td>
<td>2D matrix-based descriptors</td>
</tr>
<tr>
<td>4</td>
<td>MATS1v</td>
<td>Moran autocorrelation of lag 1 weighted by van der Waals volume</td>
<td>2D autocorrelations</td>
</tr>
<tr>
<td>5</td>
<td>Mi</td>
<td>Mean first ionization potential (scaled on Carbon atom)</td>
<td>Constitutional indices</td>
</tr>
<tr>
<td>6</td>
<td>AAC</td>
<td>Mean information index on atomic composition</td>
<td>Information indices</td>
</tr>
<tr>
<td>7</td>
<td>SpMAD_B(m)</td>
<td>Spectral mean absolute deviation from Burden matrix weighted by mass</td>
<td>2D matrix-based descriptors</td>
</tr>
<tr>
<td>8</td>
<td>GATS1p</td>
<td>Geary autocorrelation of lag 1 weighted by polarizability</td>
<td>2D autocorrelations</td>
</tr>
<tr>
<td>9</td>
<td>C-026</td>
<td>R--CX--R</td>
<td>Atom-centred fragments</td>
</tr>
<tr>
<td>10</td>
<td>SIC0</td>
<td>Structural Information Content index (neighborhood symmetry of 0-order)</td>
<td>Information indices</td>
</tr>
<tr>
<td>11</td>
<td>nDB</td>
<td>Number of double bonds</td>
<td>Constitutional indices</td>
</tr>
<tr>
<td>12</td>
<td>SIC1</td>
<td>Structural Information Content index (neighborhood symmetry of 1-order)</td>
<td>Information indices</td>
</tr>
<tr>
<td>13</td>
<td>ATS6e</td>
<td>Broto-Moreau autocorrelation of lag 6 (log function) weighted by Sanderson electronegativity</td>
<td>2D autocorrelations</td>
</tr>
<tr>
<td>14</td>
<td>P_VSA_MlR_3</td>
<td>P_VSA-like on Molar Refractivity, bin 3</td>
<td>P_VSA-like descriptors</td>
</tr>
<tr>
<td>15</td>
<td>DLS_02</td>
<td>Modified drug-like score from Oprea et al., (6 rules)</td>
<td>Drug-like indices</td>
</tr>
<tr>
<td>16</td>
<td>nCL</td>
<td>Number of Chlorine atoms</td>
<td>Constitutional indices</td>
</tr>
<tr>
<td>17</td>
<td>J_Dz(Z)</td>
<td>Balaban-like index from Barysz</td>
<td>2D matrix-based descriptors</td>
</tr>
<tr>
<td>18</td>
<td>SM6_B(s)</td>
<td>Spectral moment of order 6 from Burden matrix weighted by I-State</td>
<td>2D autocorrelations</td>
</tr>
<tr>
<td>19</td>
<td>GATS1v</td>
<td>Geary autocorrelation of lag 1 weighted by van der Waals volume</td>
<td>2D autocorrelations</td>
</tr>
<tr>
<td>20</td>
<td>JGI4</td>
<td>Mean topological charge index of order 4</td>
<td>2D autocorrelations</td>
</tr>
<tr>
<td>21</td>
<td>P_VSA_i-4</td>
<td>P_VSA-like on ionization</td>
<td>P_VSA-like</td>
</tr>
</tbody>
</table>
194 of 347 training set chemicals were correctly predicted by the \( k \)-NN classification model. The sensitivity describes the model's ability to correctly identify the class for a chemical. For training set prediction, the \( k \)-NN classification cross-validated model displayed sensitivities of 0.54, 0.49 and 0.65 for classes I, II and III, respectively (Table 2.2.3.1.2). Specificity characterizes the ability of the particular class to reject the chemicals of all other classes. The \( k \)-NN classification cross-validated model displays high specificity values for all 3 classes, i.e., the model can predict highly toxic chemicals (class I) with a specificity rate of 0.81, and 0.76 and 0.78 for classes II and III, respectively. For the external validation, the model has shown sensitivities of 0.39, 0.35 and 0.55 for classes I, II and III, respectively, and corresponding specificities of 0.73, 0.68 and 0.74. The model correctly identified classes of 38 out of 89 chemicals from the external set.

Table 2.2.3.1.2. Classification parameters of the \( k \)-NN classification model.

<table>
<thead>
<tr>
<th>NER</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Class</td>
<td>Class</td>
</tr>
<tr>
<td></td>
<td>I</td>
<td>II</td>
</tr>
<tr>
<td>Fitting</td>
<td>0.66</td>
<td>0.53</td>
</tr>
<tr>
<td>CV</td>
<td>0.67</td>
<td>0.54</td>
</tr>
<tr>
<td>External</td>
<td>0.57</td>
<td>0.39</td>
</tr>
</tbody>
</table>

2.2.3.2 Analysis based on significant descriptors:

To study how structures were related with the toxicity classes, we did a Principal Component Analysis (PCA) study on both the training set and test set chemicals using the 25 descriptors as mentioned in the Table 2.2.3.1.1.

2.2.3.2.1 Score plot:

In the score plot, structurally similar chemicals occurred closer to each other with respect to the first two principal components, while chemicals that differed structurally were found further away from each other. Figure 2.2.3.2.1a shows how chemicals are distributed based on their structural similarities and differences. Variance associated with the first principal component (PC-1) was 20% and for the second principal component (PC-2) 16%.
We analyzed the score and loading plots to study how clusters in the score plot are characterized by chemicals with similar structures, as well as how the selected molecular descriptors encode the correct information to visualize and separate these structural clusters. Further analysis revealed that most of the aliphatic alcohols were observed at the left-most side of the score plot, while halogenated benzenes were observed on the right-most side along PC-1. The halogenated alkenes and halogenated heterocyclic structures were observed projecting downward along with the PC-2. The chemicals containing phosphate and sulphate functionalities were present along with PC-2 in the upward direction. For the training set chemicals, a series of specific characteristic structural sets are shown grouped with marked areas in Figure 2.2.3.2.1a.

Comparison of the score plot of the test set (Figure 2.2.3.2.1b) to that of the training set (Figure 2.2.3.2.1a) revealed that many similar functional groups formed clusters in similar manners.
2.2.3.2.2 Loadings plot:

The PCA Loadings plot (Figure 2.2.3.2.2) was studied to investigate the significance of molecular descriptors in determining each principal component and thus the role of descriptors for the separation of clusters identified in the score plot. The chemicals occurring on the right side of the score plot (with positive scores on PC-1) are characterized by high values of descriptors present on the right side of the loadings plot (positive values on PC-1) and vice-versa. The plot is divided in two regions; the outer ellipse indicates 100% explained variance, while the inner ellipse indicates that only 50% variance in the data could be explained. The first two PCs have explained 36% variance and the outer 8 descriptors are those with the highest loadings weight on these PCs.

Additionally, the loadings plot was analyzed with respect to the score plot, and the characteristics related to the distribution of chemicals in the PCA score plot were noted. It was found that the chemicals in the green marked sectors represent chemicals with a halogenated alkene moiety, and are predominantly class I chemicals. It was also observed that acyclic halogenated alkenes were present along PC-2, followed by more complex halogenated structures, such as halogen substituted aromatic and heterocyclic structures. The descriptor nCl that signifies number of chlorine atoms, has a high weight for chemicals in the green marked sectors. Similarly, the descriptor C-026 (atom centered
fragments) and the MATS-1v (Moran autocorrelation of lag 1 weighted by van der Waals volume) have high weight for all chemicals in the red marked sector which was dominated by class-1 chemicals (in the score plot for the training set). Likewise, descriptors GATS1p (Geary autocorrelation of lag 1 weighted by polarizability) and GATS1v (Geary autocorrelation of lag 1 weighted by van der Waals volume) have a high weight for chemicals in the black marked sectors, where class III chemicals are dominant (i.e., acids, esters and alcohols). Moreover, descriptors Mi (Mean first ionization potential (scaled on the carbon atom)) and nDB (number of double Bonds) show high weight for chemicals in the upper-left section; this clarifies the increase of unsaturation along PC-1 towards the right direction, and the class III chemicals are dominant in this region. The presence of phosphate and sulphate functionality increased in the upward direction along PC2. This is because of high weight of the descriptors B01[S-P] (Presence/absence of S - P at topological distance 1), B03[C-S] (Presence/absence of C—S at topological distance 3) and P-117 (X3-P = X (phosphate)).

![Figure 2.2.3.2.2 Loadings plot showing the influence of the selected descriptors on PC-1 and PC-2.](image)

**Figure 2.2.3.2.2 Loadings plot showing the influence of the selected descriptors on PC-1 and PC-2.**

2.2.3.3 Toxicological significance:

While developing the \(k\)-NN classification model for LD\(_{50}\) values incorporating physicochemical property-based two-dimensional descriptors, we illustrated in section 2.2.3.2 how structural information encoded in molecular descriptors is related to the classes of toxicity. We found that in some cases, chemicals sharing similar structural features were representing the same class of toxicity, while in a few other cases, structurally similar chemicals were from different classes of toxicity. The reasons for this could be e.g. the role of pharmacokinetic and pharmacodynamic parameters of chemicals,
erroneous data of LD$_{50}$ experiments, inter-laboratory variations in LD$_{50}$ values and implementation of outdated OECD toxicological guidelines for conducting LD$_{50}$ experiments. In addition, it has been said elsewhere that LD$_{50}$ values are imprecise values and not biological constants.$^{58}$ Literature has also documented that

... the numeric LD$_{50}$ per se is not equivalent to acute toxicity. It should always be remembered that lethality is only one of many reference points used to characterize acute toxicity.$^{58}$

Also, in an LD$_{50}$ experiment 50% mortality of test population occurs, which indicates that the death is a collective representation of several mechanisms of toxic actions including on-target, off-target and non-specific effects. Keeping this in mind, the structurally similar chemicals may exhibit a similar mechanism of action(s) that results in similar toxic potency, while in other cases, structurally similar chemicals could show dissimilar off-target, non-specific effects and varied pharmacokinetic properties that could lead to differences in their toxic potentials.

### 2.3 Conclusion

The main goal of this study (Paper I) was to construct a preliminary QSAR model that can be used for screening the toxic potential of unknown chemicals. The Munro database that consists of diverse structures was employed for the model calibration. This is the first study that has undertaken the objective of classification of the Munro database chemicals on the basis of LD$_{50}$ data using the GHS classification system of chemicals. The most relevant molecular descriptors were identified by means of the Genetic Algorithm and later used for the construction of the $k$-NN classification model. The further PCA study confirmed the importance of the selected descriptors in the clustering of chemicals based on structural features, as well as highlighted their potential use for establishing a GHS scheme for classifying Munro database chemicals. We believe that this QSAR-based model can be useful for predicting acute toxicity, and this approach may contribute towards a reduction in the use of animals in toxicity studies.
CHAPTER-3: ACUTE TOXICITY - SUPPORTED CHRONIC TOXICITY PREDICTION

Chronic toxicity refers to the ability of toxic substances to incite harm to an organism upon repeated exposures over a longer duration of time. The purpose of repeated dose toxicity (RDT) studies is to identify the toxicological profile of a toxic substance after repeated administration. Chronic toxicity experiments mainly focus on the identification of organ toxicity and dose-response relationships, and support the No Observed Adverse Effect Level (NOAEL) calculation that is needed before the clinical evaluation of an investigational substance begins. Thus, chronic toxicity data should be part of the safety assessment information that helps in carrying out human clinical trials.

Hence, toxicological screening is an important aspect for the safety assessment of investigational new drugs and of existing potential harmful substances of industrial and/or environmental origin. Additionally, it is essential to establish RDT studies for many such substances that may cause harm to human beings. To achieve this goal, extensive animal testing, huge cost and much time will be required. The 3R principle (reduce, refine and replace) has been endorsed by an ethical framework for the purpose of reducing animal testing and to save time and cost. Accordingly, several animal tests have been replaced by in vitro and in silico testing methods. In the case of RDT however, the usage of in vitro methods has not yet been validated. Moreover, the RDT endpoint represents a multitude of biological effects occurring in various organs at different time intervals. This poses a challenge for QSAR experts trying to develop in silico models for this endpoint, and because of this, there are very few computational models that exist to date. Therefore, in this study (Paper II) we have made an attempt to construct a new QSAR model for the prediction of RDT.
3.1 Research questions and objectives

The acute toxicity is defined as adverse effects occurring upon oral or dermal administration of a single dose of a chemical, or multiple doses given within 24 hours, or upon an inhalation exposure of 4 hours.74 The mechanism of toxicity for any chemical is generally based on several paradigms: on-target effects, off-target effects, immunotoxic actions and idiosyncratic toxic effects.75 Hence, the acute toxicity endpoint (LD₅₀) is a measure of several mechanisms of toxic actions including off-target and non-specific effects.

In a chronic toxicity study, the smallest dose of a test substance given to an organism for a period of 28 days, 91 days or 2 years that causes any detectable effect is known as the Lowest Observed Effect Level (LOEL). While deriving LOEL values, many effects such as acetylcholine esterase inhibition, inflammation, hypertrophy, necrosis, blood chemical examinations and histopathological findings are studied for a test chemical. Currently no generally accepted alternative methods are available for replacing repeated dose \textit{in vivo} testing. As a variety of endpoints are examined, it is necessary to develop a unified method based on alternative approaches with complementary endpoints for RDT testing.76

Since acute toxicity data (LD₅₀) can be used when setting dose levels for repeated dose studies,77 also LD₅₀ and LOEL can be influenced by multiple toxicological mechanisms.75,78 Furthermore, LD₅₀ and LOEL values are expressed in the same unit of mg/kg/day. Therefore, if lethal doses (LD₅₀) of a test chemical and its structurally similar analogs are in the same range or within an order of magnitude, then we may be able to predict the LOEL of a test chemical using the LOEL of its structurally similar analogs.

To explore this idea, we developed a \textit{k}-NN classification model using two LD₅₀ based classes as a response variable. We have used PaDEL fingerprints79 to construct the \textit{k}-NN model, through which \textit{k}-neighbors for each of the chemicals were identified. All the training and test set chemicals were assumed as ‘queries’ and their corresponding \textit{k}-neighbors as ‘analogs’. Accordingly, a query together with its \textit{k}-analogs was considered as a single category. In this way, individual categories were formed for each query from the training and the test sets. If the \textit{k}-NN classification model correctly predicts the class of a query, its category should be considered as qualified for the read-across studies. Subsequently, its LOEL will be calculated by taking the arithmetic mean of the LOELs of its \textit{k}-analogs within that category, which establishes the main objective of this study.
3.2 Summary of work

3.2.1 RDT NEDO database
In 2007-2010 the New Energy and Industrial Technology Development Organization (NEDO) implemented a database of chemicals for the RDT endpoint in the development of the Hazard Evaluation Support System (HRESS) integrated platform, which was later incorporated in the OECD QSAR toolbox version 2.2.290-92 279 substances were obtained from the RDT NEDO database through the OECD QSAR toolbox 2.2. The database was screened for “Examination item-LOEL”, “Effect-Total effect”, “Organism-Rat”, “Gender-Male”, “strain-Crj:CD(SD)” and “route-oral (gavage)”. We sorted out 249 chemicals, and all were authenticated with respect to structure, IUPAC name and CAS registry number (RN). Duplicate records, salts, mixtures and a bulky chemical were discarded, resulting in a dataset of 224 chemicals. Acute toxicity (LD₅₀) values (organism-rat, route-oral) for 134 of the 224 chemicals were collected from the Toxnet web server.58 16 of these 134 chemicals were found to have LOEL values larger than their LD₅₀ values and were discarded. The final dataset thus consisted of 118 chemicals. The LOEL values for these were derived from assays of varying duration, such as 28, 42, 44, 46, 49, 56, 90, 91 and 98 days; we have included data from all assays for completeness. These 118 chemicals were then classified into one of two classes (toxic and non-harmful) based on the threshold of the LD₅₀ values as per the Globally Harmonized System54 Table 3.2.1.1.

Table 3.2.1.1. Classifications of the 118 chemicals in the training and test sets prior to k-NN model construction.

<table>
<thead>
<tr>
<th>Description</th>
<th>LD₅₀ (mg/kg/day)</th>
<th>Number of entries</th>
<th>Training set entries</th>
<th>Test set entries</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class 1</td>
<td>Highly toxic,</td>
<td>70</td>
<td>56</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>toxic and harmful</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Class 2</td>
<td>&gt; 2000</td>
<td>48</td>
<td>38</td>
<td>10</td>
</tr>
</tbody>
</table>

3.2.2 Methodology

3.2.2.1 Fingerprint calculations:
A total of eight types of fingerprints were employed for the construction of the k-NN classification models. PaDEL software was used for the calculation of eight types of fingerprints for the training and test set chemicals.79 These fingerprints include: Estate (length-79), CDK (length-1024), Extended CDK
(length-1024), CDK Graph (length-1024), Pubchem (length-881), MACCS (length-166), Substructural (length-307) and Klekotha-Roth (length-4860) fingerprints. Each of the eight types of fingerprints was used separately to construct a classification model.

3.2.2.2 Classification model development:

A k-NN classification method was employed for prediction of toxicity classes of chemicals. The ‘Jaccard-Tanimoto’ distance method was used for calculation of distance matrices. The 5-fold cross validation was used for the determination of an optimal k value (for more details refer section 2.2.2.3.2).

The k-NN method provided a final output for all eight types of fingerprints. All these models were then validated using the external test set. The test set of 24 chemicals which was not considered for the model calibration, was employed for the external validation of each model. Several validation parameters were studied to evaluate the optimal model; these parameters were NER, sensitivity, specificity and class error.\(^{42}\)

3.2.2.3 Model selection and read-across:

The parameters for the internal and external validations were studied to identify the optimal model, which was further used for the read-across studies. We have considered all training and test set chemicals as ‘queries’. By applying the k-NN approach, k-neighbors were identified for each query and called ‘analogs’. The particular query with its corresponding k-analogs was considered as a single category. To predict the LOEL of a query in a particular category, we took the arithmetic mean of the LOELs of all its k-neighbors.

3.2.3 Results and discussion:

3.2.3.1 k-NN classification:

Using the k-NN method, we constructed eight classification models for the respective fingerprint types. Among them, the Estate fingerprint based k-NN model showed the best model statistics (Table 3.2.3.1).

<table>
<thead>
<tr>
<th></th>
<th>NER</th>
<th>K</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Class 1</td>
<td>Class 2</td>
</tr>
<tr>
<td>Fitting</td>
<td>0.75</td>
<td>3</td>
<td>0.73</td>
<td>0.76</td>
</tr>
<tr>
<td>CV</td>
<td>0.74</td>
<td>3</td>
<td>0.77</td>
<td>0.71</td>
</tr>
<tr>
<td>External</td>
<td>0.81</td>
<td>3</td>
<td>0.71</td>
<td>0.90</td>
</tr>
</tbody>
</table>
3.2.3.2 Read-across for LOEL prediction:

The LOEL predictions of all the training and test set queries are summarized in Table 3.2.3.2. The ratio of actual and predicted LOEL for each query was calculated, and is referred to as the fold difference (Fold_diff). Our approach has correctly predicted LOELs of 54 out of 94 training set chemicals and 14 out of 24 test set chemicals to within a 10 fold difference.

**Table 3.2.3.2.** Training and test set query categorization (qualified and non-qualified) based on $k$-NN model-based predicted class, and further divided based upon an order of magnitude difference.

<table>
<thead>
<tr>
<th>Fold_diff</th>
<th>Training set queries</th>
<th></th>
<th></th>
<th>Test set queries</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Qualified category</td>
<td>Non-qualified category</td>
<td>Total</td>
<td>Qualified category</td>
<td>Non-qualified category</td>
<td>Total</td>
</tr>
<tr>
<td>&lt;10</td>
<td>54</td>
<td>17</td>
<td>71</td>
<td>14</td>
<td>3</td>
<td>17</td>
</tr>
<tr>
<td>10-100</td>
<td>12</td>
<td>4</td>
<td>16</td>
<td>4</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>&gt;100</td>
<td>4</td>
<td>3</td>
<td>7</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Total</td>
<td>70</td>
<td>24</td>
<td>94</td>
<td>19</td>
<td>5</td>
<td>24</td>
</tr>
</tbody>
</table>

# order of magnitude, fold differences (Fold_diff) <10, 10-100 and >100.

Overall, the Estate fingerprint based $k$-NN model predicted the correct classes of 89 of 118 queries combined from the training and the test sets. Moreover, our results showed that if the LD$_{50}$ class of query was predicted correctly by the classification method, then it is more likely that its LOEL would be predicted to within an order of magnitude. Our study well establishes that 68 of 89 (76%) queries (of training and test sets) from the qualified type of category were found to have their LOEL predictions with a fold difference of less than 10.

Our model has failed to predict the correct class of 24 of 94 training set chemicals and 5 of 24 test set chemicals. This could be because of the presence of low quality acute toxicity data (LD$_{50}$) or the limitations of the fingerprints to identify structurally more similar analogs for those queries. Moreover, our study led to the correct prediction of the LOELs of 16 of 70 training set and 5 of 19 test set chemicals from the qualified type of categories with a more than 10-fold difference in comparison to their actual LOELs. The major reason for this is that these queries had the lowest LOEL in their particular categories. Hence, their predicted LOELs were calculated using members of other sides (*i.e.*, the higher LOEL side) in their respective categories, which resulted in values that were too large.
3.3 Conclusion

A recent report from ECHA has highlighted the potential of the “read-across” method to fill toxicological information data gaps. At present there is no existing rule or criteria for the acceptance or elimination of analogs from a category that are needed for read-across studies. Also, there are no rules for the validation of a category. Since LD₅₀ data can be used in setting the dose levels for chronic toxicity studies, and both the endpoints are influenced by multiple mechanisms including off-target and non-specific effects. This motivated us to explore a new approach for supporting the acceptance of a category for the execution of read-across, i.e., if the classification model could correctly predict the class of a query (toxic or non-harmful, based on LD₅₀ values) by means of a k-NN approach; then such a correctly predicted query and its corresponding k-analogs can be used to perform a read-across study in order to predict its LOEL. The toxicity classes (class 1 or 2) of 89 of 118 queries were predicted correctly by our model and 68 of these were found to have their LOEL predictions with a fold difference of less than 10, in comparison to their actual LOELs.

The toxicological significance of this study is supported by the notable relationship found between different mechanisms of acute (LD₅₀) and chronic toxicity (LOEL), such as acute kidney toxicity which is well explained by chronic effects like urine volume, chloride levels, creatinine levels and serum protein levels. In the same manner, acute toxicity in the liver is elucidated by chronic effects such as levels of serum indicators and liver hypertrophy, the mitochondrial toxicity is explained by hypothermia, the locomotor activity is explained by choline esterase levels, and so on. In addition to this, the Estate fingerprint based model has identified structurally similar k-analogs for many of the queries. As the notion suggests, chemicals similar in molecular structure often have similar modes of actions, and thus exhibit similar properties. The observation of structures of members in those categories (i.e., k-analogs with their corresponding query chemicals) revealed that they could affect similar targets for exhibiting their specific (on-target) and non-specific (off-target) modes of actions. In some cases our model has failed to identify structurally similar k-analogs for the queries. As category members in those cases differ structurally, they do not follow similar modes of actions, and thus LOEL predictions cannot be performed in those cases.

Given the positive results thus far, our approach has successfully demonstrated the applicability of a read-across – k-NN coupled strategy for the prediction of RDT (LOEL) using acute toxicity (LD₅₀) based classes. This approach should provide researchers with a tool to fill data gaps and allow the prediction of sub-chronic or chronic toxicity. This study should benefit computational toxicologists, pharmacologists and risk assessors in forming preliminary categories for a tier 1 read-across study for predicting toxicological endpoints.
Ultimately, this novel read-across – k-NN coupled strategy should contribute to a reduction in the number of animals used for chronic toxicity testing.
CHAPTER-4: HERG TOXICITY PREDICTION USING K-NN CLASSIFICATION METHOD

In previous chapters, we have discussed drug promiscuity i.e. on-target, off-target and non-specific effects of compounds for the exhibition of toxic actions. Similarly, drug targets can also be promiscuous. One such example is the hERG (human Ether-a-go-go-Related Gene) voltage-gated potassium channel. The hERG channel can be affected by several drugs of various therapeutic classes, such as antipsychotic, antihistaminic, antiserotonin, antimuscarinic, α1-blockers, β-blockers, antimalarial, antibiotic and antifungal. The hERG potassium channel is engaged in the conduction of rapid delayed rectifier K+ current that contributes towards maintaining cardiac activity. Inhibition of the hERG channel leads to a fatal disorder called long QT interval, and quite often results in life-threatening ventricular arrhythmias, also known as ‘Torsades de Pointes’. Hence, the hERG has become an important ‘anti-target’, which should be avoided during development of a new drug.

To screen the hERG blocking potential of drugs, several in vitro assays are available to date, such as rubidium-flux assays, radioligand binding assays, in vitro electrophysiology measurements, and fluorescence-based assays. All drug candidates are regularly tested for hERG liability by in vitro assays, but these studies are often costly and labour intensive. Therefore, validated in silico models are of prime importance in this framework. With this study (Paper III), we have made an attempt to develop a predictive and well characterized in silico model for quick screening of hERG inhibitors.
4.1 Research questions and objectives

An anti-target is a receptor that, when affected by a drug candidate, shows side effects.\textsuperscript{85} Pharmaceutical companies need to make sure that new drugs do not exhibit any activity towards anti-targets, most of which are discovered accidentally. The hERG channel is one such well-known anti-target. The blockade of the hERG channel has resulted in withdrawal of several marketed drugs, such as cisapride, terfenadine, astemizole, sertindole, grepafloxacin, etc.\textsuperscript{86} Importantly, this target is inhibited by drugs from several therapeutic classes perhaps due to the receptors larger cavity enables binding of drug candidates of varying shapes and sizes. Moreover, due to the presence of a wider cavity and the open-close nature of the channel, several binding modes of drugs are possible inside the hERG cavity. This makes the target ‘drug-promiscuous’.\textsuperscript{87} To date, there is no crystal structure available for this target; therefore \textit{in silico} study is one viable alternative for predicting the binding potentials of drug candidates for this target.

Here, we aimed to construct the QSAR classification model for screening the hERG actives and inactives (\textit{Paper III}).

4.2 Summary of work

4.2.1 Setting up the database

The 172 Ikr (‘rapid’ delayed rectifier current) channel blockers along with IC\textsubscript{50} data were retrieved from the web servers OCHEM\textsuperscript{88} and Fenichel.\textsuperscript{89} This dataset of 172 compounds was comprised of diverse structures belonging to different therapeutic classes. All compounds were authenticated based on the correct structure and correct IUPAC name. Subsequently, their SMILES notations were verified using ChemSpider.\textsuperscript{55} SigmaAldrich\textsuperscript{56} and PubChem.\textsuperscript{57} The final dataset of 172 compounds was used for model construction. For external validation, a PubChem dataset\textsuperscript{80} of 1953 compounds was chosen. After removal of salts and mixtures the final validation set (test set) was comprised of 1795 compounds.

4.2.2 Methodology

4.2.2.1 Fingerprint calculations:

Eight types of PaDEL fingerprints were calculated for the training and test set compounds using PaDEL software:\textsuperscript{79} CDK, Extended CDK, CDK Graph, Estate, MACCS, Pubchem, Sub-structure and Sub-structure count fingerprints. Each of the fingerprints was used separately to develop a classification model.
4.2.2.2 Class assignment:

The training set compounds were divided into two classes (*i.e.*, actives and inactives) using an IC<sub>50</sub> threshold of 5μM (see Table 4.2.2.2). Similarly, test set compounds were divided into hERG active and inactive classes based on a threshold of 20% inhibition.

**Table 4.2.2.2.** Classification of training and test set compounds.

<table>
<thead>
<tr>
<th></th>
<th>Class 1 (hERG actives)</th>
<th>Class 2 (hERG inactives)</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training</td>
<td>93</td>
<td>79</td>
<td>172</td>
</tr>
<tr>
<td>Test</td>
<td>221</td>
<td>1574</td>
<td>1795</td>
</tr>
</tbody>
</table>

4.2.2.3 Software and Modules:

The Matlab module classification_toolbox developed at the Milano Chemometrics and QSAR Research Group, University of Milan, Italy was used for the construction of a *k*-NN classification model.<sup>65</sup>

4.2.2.4 Model development:

The *k*-nearest neighbor (*k*-NN) classification method was used to build a classification model.<sup>45</sup> A series of *k* values (from 1 to 10) were assigned to develop the model and an optimal (on the basis of lowest class error) *k* value was identified using 5-fold cross validation. Afterwards, all these models were subjected to external validation using the PubChem dataset of 1795 compounds. The performance of the classification model was assessed on the basis of different statistical parameters, *i.e.* non-error rate (NER), sensitivity, specificity, precision and error rate (ER). All the models were analyzed and compared with respect to these statistical parameters.
### 4.2.3 Results and discussion

4.2.3.1 Eight types of $k$-NN classification models:

All classification models were studied, and their statistical parameters are described in Table 4.2.3.1.

<table>
<thead>
<tr>
<th>Entry</th>
<th>Fingerprint</th>
<th>NER</th>
<th>k</th>
<th>Sensitivity Class 1</th>
<th>Class 2</th>
<th>Specificity Class 1</th>
<th>Class 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>CDK</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Fitting</td>
<td>0.68</td>
<td>1</td>
<td>0.72</td>
<td>0.65</td>
<td>0.65</td>
<td>0.72</td>
</tr>
<tr>
<td></td>
<td>CV</td>
<td>0.66</td>
<td>1</td>
<td>0.72</td>
<td>0.61</td>
<td>0.61</td>
<td>0.72</td>
</tr>
<tr>
<td></td>
<td>External</td>
<td>0.54</td>
<td>1</td>
<td>0.52</td>
<td>0.57</td>
<td>0.57</td>
<td>0.52</td>
</tr>
<tr>
<td>2</td>
<td>Estate</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Fitting</td>
<td>0.68</td>
<td>1</td>
<td>0.73</td>
<td>0.62</td>
<td>0.62</td>
<td>0.73</td>
</tr>
<tr>
<td></td>
<td>CV</td>
<td>0.66</td>
<td>1</td>
<td>0.72</td>
<td>0.61</td>
<td>0.61</td>
<td>0.72</td>
</tr>
<tr>
<td></td>
<td>External</td>
<td>0.53</td>
<td>1</td>
<td>0.49</td>
<td>0.57</td>
<td>0.57</td>
<td>0.49</td>
</tr>
<tr>
<td>3</td>
<td>Extended CDK</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Fitting</td>
<td>0.67</td>
<td>1</td>
<td>0.70</td>
<td>0.63</td>
<td>0.63</td>
<td>0.70</td>
</tr>
<tr>
<td>4</td>
<td>CDK Graph</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Fitting</td>
<td>0.65</td>
<td>1</td>
<td>0.70</td>
<td>0.61</td>
<td>0.61</td>
<td>0.70</td>
</tr>
<tr>
<td></td>
<td>CV</td>
<td>0.56</td>
<td>1</td>
<td>0.56</td>
<td>0.57</td>
<td>0.57</td>
<td>0.56</td>
</tr>
<tr>
<td></td>
<td>External</td>
<td>0.64</td>
<td>1</td>
<td>0.69</td>
<td>0.59</td>
<td>0.59</td>
<td>0.69</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.64</td>
<td>1</td>
<td>0.70</td>
<td>0.58</td>
<td>0.58</td>
<td>0.70</td>
</tr>
<tr>
<td>5</td>
<td>MACCS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Fitting</td>
<td>0.68</td>
<td>6</td>
<td>0.76</td>
<td>0.59</td>
<td>0.59</td>
<td>0.76</td>
</tr>
<tr>
<td></td>
<td>CV</td>
<td>0.67</td>
<td>6</td>
<td>0.76</td>
<td>0.57</td>
<td>0.57</td>
<td>0.76</td>
</tr>
<tr>
<td></td>
<td>External</td>
<td>0.55</td>
<td>6</td>
<td>0.54</td>
<td>0.55</td>
<td>0.55</td>
<td>0.54</td>
</tr>
<tr>
<td>6</td>
<td>PubChem</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Fitting</td>
<td>0.60</td>
<td>3</td>
<td>0.69</td>
<td>0.52</td>
<td>0.52</td>
<td>0.69</td>
</tr>
<tr>
<td></td>
<td>CV</td>
<td>0.60</td>
<td>3</td>
<td>0.71</td>
<td>0.49</td>
<td>0.49</td>
<td>0.71</td>
</tr>
<tr>
<td></td>
<td>External</td>
<td>0.57</td>
<td>3</td>
<td>0.62</td>
<td>0.52</td>
<td>0.52</td>
<td>0.62</td>
</tr>
<tr>
<td>7</td>
<td>Sub-structure</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Fitting</td>
<td>0.68</td>
<td>1</td>
<td>0.70</td>
<td>0.67</td>
<td>0.67</td>
<td>0.70</td>
</tr>
<tr>
<td>8</td>
<td>Sub-structure count</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>CV</td>
<td>0.67</td>
<td>1</td>
<td>0.69</td>
<td>0.66</td>
<td>0.66</td>
<td>0.69</td>
</tr>
<tr>
<td></td>
<td>External</td>
<td>0.67</td>
<td>1</td>
<td>0.74</td>
<td>0.61</td>
<td>0.61</td>
<td>0.74</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.68</td>
<td>1</td>
<td>0.72</td>
<td>0.65</td>
<td>0.65</td>
<td>0.72</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.58</td>
<td>1</td>
<td>0.61</td>
<td>0.56</td>
<td>0.56</td>
<td>0.61</td>
</tr>
</tbody>
</table>
4.2.3.2. Consensus $k$-NN classification model:

To further improve the predictive power of these models we developed a series of consensus models. The statistical parameters of these models are presented in Table 4.2.3.2.

**Table 4.2.3.2.** The statistical parameters of the Consensus models.

<table>
<thead>
<tr>
<th>Model</th>
<th>Dataset</th>
<th>$Q^b$</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Precision</th>
<th>G-mean$^c$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Training</td>
<td>0.73</td>
<td>0.77</td>
<td>0.68</td>
<td>0.74</td>
<td>0.73</td>
</tr>
<tr>
<td></td>
<td>Validation</td>
<td>0.58</td>
<td>0.59</td>
<td>0.58</td>
<td>0.17</td>
<td>0.59</td>
</tr>
<tr>
<td>2</td>
<td>Training</td>
<td>0.70</td>
<td>0.78</td>
<td>0.61</td>
<td>0.70</td>
<td>0.69</td>
</tr>
<tr>
<td></td>
<td>Validation</td>
<td>0.55</td>
<td>0.63</td>
<td>0.54</td>
<td>0.16</td>
<td>0.59</td>
</tr>
<tr>
<td>3</td>
<td>Training</td>
<td>0.69</td>
<td>0.76</td>
<td>0.61</td>
<td>0.70</td>
<td>0.68</td>
</tr>
<tr>
<td></td>
<td>Validation</td>
<td>0.56</td>
<td>0.61</td>
<td>0.55</td>
<td>0.16</td>
<td>0.58</td>
</tr>
<tr>
<td>4</td>
<td>Training</td>
<td>0.70</td>
<td>0.80</td>
<td>0.59</td>
<td>0.70</td>
<td>0.69</td>
</tr>
<tr>
<td></td>
<td>Validation</td>
<td>0.55</td>
<td>0.58</td>
<td>0.54</td>
<td>0.15</td>
<td>0.56</td>
</tr>
<tr>
<td>5</td>
<td>Training</td>
<td>0.72</td>
<td>0.78</td>
<td>0.63</td>
<td>0.72</td>
<td>0.70</td>
</tr>
<tr>
<td></td>
<td>Validation</td>
<td>0.57</td>
<td>0.60</td>
<td>0.56</td>
<td>0.16</td>
<td>0.58</td>
</tr>
<tr>
<td>6</td>
<td>Training</td>
<td>0.72</td>
<td>0.78</td>
<td>0.65</td>
<td>0.72</td>
<td>0.71</td>
</tr>
<tr>
<td></td>
<td>Validation</td>
<td>0.57</td>
<td>0.59</td>
<td>0.57</td>
<td>0.16</td>
<td>0.58</td>
</tr>
</tbody>
</table>

$^a$ Model 1 = Substructure (SS) + Substructure count (SSC) + Extended CDK (ECDK), 2 = PubChem (PC) + SSC + ECDK, 3 = PC + SSC + SS, 4 = PC + SSC + MACCS, 5 = PC + SSC + ECDK + SC + MACCS, 6 = PC + SSC + ECDK + SS + MACCS + CDK + CDK Graph. $^b$ Overall accuracy of prediction. $^c$ $\sqrt{\text{Sensitivity} \times \text{Specificity}}$

The consensus model 2 shows 70% success (121 of 172) in predicting the correct class of the training set compounds. The model correctly classified 73 of 93 (78%) compounds from class 1 and 48 of 79 (61%) compounds from class 2. A G-mean of 69% was observed for the training set. In the case of the test set, 140 of 221 compounds from class 1, and 851 of 1574 compounds from class 2, were predicted correctly. The G-mean associated with the test set prediction was 59%.

4.2.3.3. Comparison of consensus model with previously published models:

The external validation demonstrates true predictability of a QSAR model, and to compare such a predictability, it is necessary that the external validation should have performed on the same dataset. Therefore, we have compared our model with previously published models, those are externally validated with the PubChem dataset (Table 4.2.3.3).
Table 4.2.3.3. Comparison of consensus model with other models externally validated with PubChem database compounds.

<table>
<thead>
<tr>
<th>Model</th>
<th>Method</th>
<th>Training set</th>
<th>Test set</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No of compounds</td>
<td>Sensitivity</td>
<td>Specificity</td>
</tr>
<tr>
<td>Our study</td>
<td>k-NN</td>
<td>172</td>
<td>78 (73/93)</td>
</tr>
<tr>
<td>Su et al. 2012&lt;sup&gt;91&lt;/sup&gt;</td>
<td>SVM</td>
<td>546</td>
<td>90 (188/210)</td>
</tr>
<tr>
<td>Wang et al. 2012&lt;sup&gt;92&lt;/sup&gt;</td>
<td>NB</td>
<td>719</td>
<td>89 (247/279)</td>
</tr>
<tr>
<td>Su et al. 2010&lt;sup&gt;93&lt;/sup&gt;</td>
<td>PLS transformed into binary QSAR</td>
<td>250</td>
<td>R²=0.58</td>
</tr>
<tr>
<td>Li et al. 2008&lt;sup&gt;94&lt;/sup&gt;</td>
<td>SVM</td>
<td>495</td>
<td>55 (83/152)</td>
</tr>
</tbody>
</table>
Three of the four previously published models demonstrate lower overall sensitivities than our model. As it is of more interest to identify the potent hERG blockers (class 1) than that of hERG inactive compounds (class 2), thus our model has better performed in predicting hERG active compounds than other models. Importantly, all the other models have used 3D and 4D descriptors that are computationally expensive to calculate. Moreover, these models have used different methods of descriptor selection, which makes this task more cumbersome. While, in our study we have used 2D descriptors that are fast and easy to calculate and we have not incorporated any complicated method of descriptor selection. Thus, in comparison to the other models, our model has advantages of being fast, simple and efficient in predicting the hERG toxic compounds.

We then attempted to further validate this model by using substances withdrawn on account of QT prolongation. The model correctly predicted the hERG IC50-based classification of 11 of the 15 substances (73%) found in the WITHDRAWN database95 that had not been used in model development. The fact that these substances fall into both classes, five and six in classes 1 and 2, respectively, indicates that the model may be useful for differentiating between hERG-derived QT prolongation and other QT prolongation mechanisms. The interpretation of the QT prolongation endpoint is itself a major challenge as mechanisms other than hERG activity can also underlie QT prolongation.96–98

A general reflection upon examining the hERG active compounds predicted by our model was the prevalence of aromatic and basic functionalities in these compounds. These features have previously been identified as essential components in a pharmacophore for central nervous system activity,99,100 an issue which we believe should be considered in future model development. A further issue derived from this observation is the possibility that this may be indicative of a common evolutionary origin for the hERG voltage dependent K⁺ ion channel and CNS receptors101,102.

4.3 Conclusion

In conclusion, the consensus model constructed using the Extended CDK, PubChem and Sub-structure count fingerprint based models, shows potential as a tool for screening for potential hERG-toxicity. As our model showed comparable performance with models employing more complicated descriptors in the validation with PubChem datasets, it is hoped that it may serve medicinal chemists in filtering drug candidates for hERG toxicity.
CHAPTER-5: CONCLUDING REMARKS AND FUTURE OUTLOOK

This thesis reports on a series of studies where multivariate methods were employed for modelling toxicological endpoints. Collectively, the studies presented demonstrate the potential of computational methods for establishing quantitative structure-activity relationships among diverse chemicals from different databases, such as the Munro, RDT NEDO, Fenichel and Ochem, with different toxicological endpoints such as LD$_{50}$, LOEL, and IC$_{50}$ (hERG).

In **Paper I**, the possibility of developing a global QSAR model for acute toxicity was explored by using the Munro database that contains structures of varying chemical functionalities and molecular volumes. A genetic algorithm was used to identify relevant molecular descriptors that were later employed for the construction of the $k$-NN classification model. The developed model has shown satisfactory predictive ability, and a PCA study confirmed the importance of the significant descriptors in classifying Munro database chemicals using the GHS scheme of chemical classification. Future work aimed at improving the predictive power of this acute toxicity model could include incorporating 3-D descriptors into model construction, though at the cost of time and computing power.

In **Paper II**, acute toxicity values were used to support the prediction of chronic toxicity. The $k$-NN classification model was constructed using LD$_{50}$ based classes as response variables. A chemical, for which the $k$-NN model correctly predicts its class, then has its $k$-analogs used for the prediction of its LOEL value by means of a read-across approach. The results confirmed that this approach better predicts chronic toxicity values, though expanding the number of compounds used should lead to an even better model for chronic toxicity prediction.
In **Paper III**, the \( k \)-NN classification approach was used to predict the hERG-derived toxicity. Results demonstrate that our model performed comparably with models employing more complicated descriptors upon validation with external datasets. Exploring the use of the QT prolongation endpoint instead of hERG toxicity remains of interest for future endeavours. The performance of this model suggests that it could be used as a tool for supporting the screening of the hERG liability of new drugs and other compounds.

In summary, this thesis has explored the use of computational strategies for predicting different types of toxicity with the ultimate goals of reducing the costs and use of animals in the screening of chemical compounds that we are exposed to, and even to assist in the development of new drugs.
ACKNOWLEDGEMENTS

My thesis wouldn’t have been possible without the help of many people, and I am very thankful to all of you. Here, I would like to express my sincere thanks and gratitude to the following persons who had contributed, helped and supported me during PhD studies.

Prof. Ian A. Nicholls, PhD supervisor, for his expert advice and help in designing, conducting and documenting my PhD research studies.

Prof. Tomas Öberg, x-PhD supervisor, for giving me the chance to be a Marie-Curie fellow and guiding me at the start of this PhD project.

Dr. Ran Friedman, Dr. Björn C. G. Karlsson and Dr. Annika Rosengren for their co-supervision in this PhD project.

Prof. Roberto Todeschini for providing me valuable training on descriptor selection methods.

Dr. Davide Ballabao and Dr. Viviana Consonni for their guidance and support during my training at University of Milano.

Dr. Igor Tetko, for all his support, encouragement and guidance throughout the ECO project.

Dr. Eva Schlosser, for her support and help during the ECO project.

Gustaf Olsson, Kerstin Golker and Emma Ahlstrand, who shared the office with me, for their precious help, guidance, Swedish translations and support during all these years of my PhD.

Susan Canali and Jesper Wiklander for proof-reading this thesis.

Stefan Hagberg, University technical assistant, without whom my life wouldn’t have been that easy at this department.

Pernilla, Subramanian, Louise, Siamak, Kathir, Natasha, Javed, Younes and Afridi for being such cheerful lab mates.

Matteo Cassetti and Franchesca Grisoni, teammates at University of Bocca-milano, for their assistance and support during my secondment.

Ahmed Abdelaziz, Robert Körner, Sergii Novotarskyi and Yurii Sushko, teammates at eADMET, for their assistance and technical support during my industrial training.
Isabel, Michael, Rajesh, Faizan, Aleksandra, Ian, Tobias, Stefan B., Sasha, Elina and all other ECO project colleagues for their valuable support and help.

Håkan, Anna S., Dieter, Karthik, Saroj, Sarfraz, Antoine, Stella, Katarzyna, Mireia, Olga, Yu, Isaya, Gaofeng, Mikael, Stina, Camilla, Tomas N., Bhaskar, Kerstin S., Nina, Shan, Ashik, Carlo, Nelson, Mourice and Olga B., my university colleagues, without whom I would have been very lonely.

My parents, my two brothers, and my friends Gahinath, Manoj, Dilip, Amit, Shailesh, Piyush, Vaibhav, Ajit, Kumar, Devendra, Smriti, Vishal, Dharmesh and Bharti for their encouragement and support during all these years away from home.

Simon A., Annika, Sebastian P, Vishnu, Roman, Sebastian S, Rene, Florian, Emma G. and Christian W., my Kalmar friends, who cheered and stood with me in my good and bad times.

Last but not least, I am grateful to the Environmental Chemoinformatics EU FP-7 project (grant number 238701) and Linnaeus University, Sweden, for funding my PhD research.
REFERENCES

10. V. Kakkar, J. Spindler, P. Flute, T. Corrigan, D. Fossard, R. Crellin, S. Wessler and E. Yin, Efficacy of low doses of heparin in prevention of deep-vein


52. I. C. Munro, R. A. Ford, E. Kennepohl and J. Sprenger, Correlation of structural class with no-observed-effect levels: a proposal for establishing a threshold of concern, *Food and Chemical Toxicology*, 1996, 34, 829-867.
54. Globally harmonized system of classification and labelling of chemicals (GHS), *United Nations*, 2011,

62


