الجمهورية الجزائرية الديمقراطية الشعبية

BADJI MOKHTAR - ANNABA UNIVERSITY

جامعة باجي مختار-عنابة

Year 2021

Faculty of Sciences

Chemistry Department

THESIS

Submitted in Partial Requirement for the Degree of Doctorate in Sciences

Option: Analytical Chemistry and Environment

Title

Molecular modeling of biologically active molecules using different theoretical approaches

by

Khalid BOUHEDJAR

Jury

President	Mr. Messaoud LIACHA	Professor	University of Annaba
Supervisor	Mr. Abdelmalek KHORIEF NACEREDDINE	Professor	ENSET - Skikda
Co-Supervisor	Mr. Abdelhafid DJEROUROU	Professor	University of Annaba
Examiner	Mr.Salah-Eddine DJILANI	Professor	University of Annaba
Examiner	Mr. Mohamed AbdEsselem DEMS	MRA	CRBt-Constantine
Examiner	Mr.Hacene BENDJEFFAL	MCA	ENSET - Skikda

Acknowledgements

I would like sincerely to thank my supervisor Professor Abdelmalek KHORIEF NACEREDDINE who directed me in this research with generosity, limitless giving, and patience. He worked tirelessly on reading the drafts, making useful suggestions. My other special thanks certainly go to my Co-supervisor Professor Abdelhafid DJEROUROU for his great help and guidance.

Then I would like to express my gratitude to Professor Messaoud LIACHA for accepting to president this hury of thesis. My thanks go also to Professor Salah-Eddine DJILANI, Doctor Mohamed Abd Esselem DEMS and Doctor Hacene BENDJEFFAL for their time and effort in reviewing this work.

I also want to thank Abdelbasset BOUKELIA for Deep learning part, and Professor Giuseppina Gini and Professor Emilio Benfinati for help, hospitality and suggestions during my internship in Milano (Italy).

Finally, I am deeply and forever indebted to my parents, my wife and friends Amine BELAIDI, Saàd MEBREK, Abdelhamid BENKHEMISSA for their support.

إلى أمي وأبي

إلى زوجتي الغالية وأنس

إلى أخوتي وأخواتي وكل العائلة

الملخص:

خلال العشر سنوات الأخيرة، لوحظ تطور سريع في التكنولوجيا الحيوية والكيميائية نذكر على سبيل المثال التدفق العالي للفحص والتركيب التسلسلي، نتج عنه كم هائل من المعطيات حيث أن هذه الأخيرة تستلزم استحداث ودمج طرق تحليلية جديدة ووسائل التعلم الآلي. الهدف الأساسي من العمل المنجز هو تنفيذ نموذج ممرر آلي من العلاقة الكمية بنية – فعالية – فعالية (QSAR-PP) من أجل التنبؤ المباشر بسمية مجموعة كبيرة من المركبات العضوية وذلك عن طريق امتداد العلاقة بين قيم بيانات السمية لعدة كائنات و المؤشرات الجزيئية. قمنا بالتحري لعدد هائل من المعطيات المتعلقة بالسمية الحادة لخمسة كائنات مائية هي

السمك (Fish)، *برغوث الماء* (Daphnia *magna*)، الطحالب (Algae) مأخوذة من المنصة VEGA-Hub و كذلك رباعية الغشاء *Vibriofischeri, Tetrahymenapyriformis* . القسم الثاني من العمل المنجز اقترحنا فيه مقاربة اللغة الطبيعية باستعمال الشبكة العصبونية المتضمنة حيث كان الهدف من هده الطريقة هو تحويل نضام الإدخال المبسط الجزيئي إلى عمود من الحروف المتضمنة من اجل تمثيل معنى المركبات. هذه الأشعة تغدي خوارزميات التعلم الآلي الخاضعة للإشراف مثل الشبكة العصبية للذاكرة طويلة المدى التلافيفية، Support Vector Machine و الشجرة العشوائية لبناء نموذج العلاقة الكمية بنية-فعالية على سمية مجموعات البيانات. النتائج المتحصل عليها من بيانات سمية *Tetrahymena pyriformis* (IGC_{50}) و بيانات السمية الحادة للفئران معبرة بمتوسط الجرعة المميتة للفئران المعالجة (LD_{50}) بينت أن طريقتنا يمكن استعمالها من أجل التنبؤ بنشاطية المركبات الكيميائية بكفاءة عالية.

كلمة مفتاحية:

نماذج العلاقة الكمية بين البنية والفعالية، نماذج العلاقة الكمية بين البنية والفعالية-الفعالية، الارتباط بين الأنواع، السمية المائية، تعلم عميق، التعلم الالي، معالجة اللغة الطبيعية.

Abstract

Over the past decade, rapid development in biological and chemical technologies such as high-throughput screening and parallel synthesis has been significantly increased the amount of data, which requires the creation and the integration of new analytical methods and machine learning techniques. The main aim of the first part of this thesis was to develop an Auto Pass-Pass Quantitative Structure-Activity-Activity Relationship (PP QSAAR) model for direct prediction of the toxicity of a larger set of compounds, applying the extrapolation of the obtained model combing the toxicity values of different species and molecular descriptors. We have investigated a large acute toxicity data set of five aquatic organisms including: fish, *Daphnia magna* and *algae* from the VEGA-Hub, as well as *Tetrahymenapyriformis* and *Vibrio fischeri*. In the second part of our work, we have proposed a natural language processing approach, based on embedding deep neural networks. Our method aims to transform the Simplified Molecular Input Line Entry System format into word embedding vectors to represent the semantics of compounds. These vectors are fed into supervised machine learning algorithms such as convolutional long short-term memory neural network, support vector machine and random forest to build up quantitative structure–activity relationship models on toxicity data sets. The obtained results on toxicity data to the ciliate *Tetrahymenapyriformis*, and acute toxicity rat data (LD_{50}) show that our approach can eventually be used to predict the activities of chemical compounds efficiently.

Keywords: QSAR, QSAAR, inter-species correlation, aquatic toxicity, Deep Learning, machine learning, natural language processing.

Résumé

Au cours de la dernière décennie, le développement rapide des techniques biologiques et chimiques tel que le criblage à haut débit et la synthèse parallèle a considérablement augmenté la quantité des données, ce qui nécessite la création et l'intégration de nouvelles méthodes analytiques et techniques d'apprentissage automatique. L'objectif principal de la première partie de cette thèse était de développer un modèle Auto Pass-Pass de quantification de la relation structure-activité-activité (PP-QSAAR) pour la prédiction directe de la toxicité de grand ensemble de composés. Le modèle développé s'extrapole par la combinaison des valeurs de toxicité des espèces et les descripteurs moléculaires. Nous avons étudié un vaste ensemble de données sur la toxicité aiguë de cinq organismes aquatiques y compris des poissons, des *Algues*, d'un crustacé, ainsi que *Tetrahymenapyriformis* et *Vibriofischeri*. Dans la deuxième partie de notre travail, nous avons proposé une approche de traitement de langage naturel basée sur l'intégration de réseaux de neurones profonds. Notre méthode vise à transformer le format Simplified Molecular Input Line Entry System en vecteurs d'embarquement de mots pour représenter la sémantique des composés. Ces vecteurs sont introduits dans des algorithmes d'apprentissage automatique supervisé tels que le réseau neuronal convolutif à mémoire de long-court terme, la machine à support vecteur et les forêts aléatoires pour construire des modèles QSAR sur des ensembles de données de la toxicité. Les résultats obtenus des IGC_{50} chez *Tetrahymenapyriformis* et les données aiguës chez le rat exprimées en (DL_{50}) montrent que notre approche peut être utilisée pour prédire efficacement les activités de composés chimiques.

Mot clés: QSAR, QSAAR, Correlation Inter-espèces, Toxicité aquatique, Deep Learning, Machine Learning, Traitement du langage naturel.

CONTENTS

ABBREVIATIONS

Abbreviation	Meaning
1-D, 2D, 3D	One-Dimension, Two-Dimensional, Three-Dimension
AD	Applicability domain
ADAM	Adaptive moment estimation
CAS	Chemical abstract
CDK	Chemistry Development Kit
DL	Deep Learning
DM	Daphnia magna
EC_{50}	50% effective concentration
ECHA	European Chemicals Agency
EPA	Environmental protection agency
GAs	Genetic algorithms
IGC_{50}	Concentration that causes 50% growth inhibition
KNIME	Konstanz information miner
LC_{50}	Lethal concentration causing death in 50% test organisms
LD_{50}	Lethal dose, 50%
LMO	Leave many out
LOO	Leave one out
LSTM	long short-term memory
MACCS	Molecular ACCess System
MDL	Molecular Design Limited
ML	Machine Learning
MLR	Multiple linear regression
MOA	Mode of action
NIH	National institute of health
NLP	Natural language processing
OCHEM	Online Chemical Modeling Environment
OECD	Organisation for Economic Co-operation and Development
OLS	Ordinary Least Squares
QMRF	QSAR Model Reporting Format

QAARs	Quantitative Activity–Activity Relationships
QSAAR	Quantitative Structure–Activity–Activity Relationships
QSAR	Quantitative Structure-Activity Relationship
REACH	Registration, Evaluation, Authorisation and Restriction of Chemicals
RF	Random forest method
SMILES	Simplified molecular Input line entry system
SVM	Support vector machine
TP	Tetrahymena pyriformis
VEGA	Virtual models for evaluating the properties of chemicals within a global architecture
VF	Vibrio fischeri

LIST OF TABLES

LIST OF FIGURES

GENERAL INTRODUCTION

The Quantitative Structure Activity Relationship (QSAR) approach is one of the most commonly used methods for the prediction of biological properties to aid the drug discovery process and for hazard and risk assessment of the chemical compounds. It is an adequate alternative way for animal expensive tests and time-consuming experiments. Since the mid-1960s, the QSAR paradigm ('similar compounds have similar activities') remains the foundation of all QSAR models developed so far [1]. The relationships established in the models allow the prediction of the activities for novel compounds based on structural features, which are encoded in a numerical notation called molecular descriptors.

In this thesis, the *in silico* method (QSAR) modeling was applied to fill data gaps of substances lacking experimental data. The first part of this study focuses on the analysis of the acute toxicity of chemicals towards five species of aquatic organisms, namely *fish*, *Daphnia magna*, and *algae* from the VEGA-Hub, as well as *Tetrahymena pyriformis* and *Vibrio fischeri*. The developed models underwent thorough validations according to regulatory recommendations [2], this work was based on inter-species extrapolation, which aims to predict one species from an endpoint in another species when the data is missing [3–10], or to predict the toxicity of compounds for large size (i.e *fish*) from lower organisms (i.e. ciliate) [11]. The hybrid method, known as Quantitative Structure Activity-Activity Relationships (QSAARs), which uses molecular descriptors in addition to the toxicity data and it is considered as a promising approach for predicting toxicity have been introduced [12,13]. This later merges the inter-species and QSAR approaches, based on the relationship between two different biological endpoints [14–16].

In the second part, we propose a novel integrative Deep Learning DL-QSAR, Random Forest-QSAR, and SVM-QSAR approaches based on embedding deep learning for predicting high-quality toxicity models through natural language processing of SMILES notation. The recent rise in popularity of Deep Learning (DL) brought several inventions in QSAR modeling and ideas from various fields of data science. It is appropriate for finding the best statistical model for predicting biological activity. In 2012, when Dahl's team won at the Merck Molecular Activity Challenge public interest was in using DL in QSAR [17]. After that, numerous research groups have used DL models to predict many parameters, including activity, toxicity, solubility, and various other proprieties [18–20].

This thesis was structured into three parts, first, we have begun in the first part with an overview of the European REACH (Registration, Evaluation, Authorisation and Restriction of Chemicals) regulation and OECD (Organisation for Economic Co-operation and Development) validation. Next, in the second part, the mean steps of QSAR starting from data and source, algorithms (Machine learning, Deep Learning), and validation were introduced. The third part contains the application of two projects with the results and discussion section, in which the first one is dedicated to the inter-species extrapolation QSAAR/QAAR, while the second one is an application of Machine Learning and Deep Learning in QSAR modeling.

PART I.STATE OF THE ART

I.1. Toxicity

Toxicology science is the study of the adverse effects of chemicals on living organisms, and the biological properties of any chemical are the effects of its structural characteristics. The toxic dose differs according to the considered specific chemical. "*All things are poison and nothing is without poison; only the dose makes a thing not a poison*" (Paracelsus, 1493). This means that some substances induce death by the concentration of few micrograms per kilo, while others may be quite toxic even if their concentrations are much higher [21].

The increasing number of different chemical compounds may pose a high risk to the environment and can cause adverse humans health effects, chemicals can pass in the body through inhalation, ingestion, or dermal exposure. Assessment of chemical hazards is necessary to ensure human safety. There are three different kinds of experiments to assess the biological activity of the compound, animal testing is *in vivo* experiments, tissue culture cells testing is *in vitro* experiments, and computer simulation is *in silico* experiments.

Toxicological tests should be performed in order to evaluate which of these chemicals are safe and which can potentially contaminate the environment and cause toxicity. Traditional toxicity testing like *in vivo* and *in vitro* have been used for a long time, however, both are expensive and time-consuming. As consequence, they are not sufficient for toxicology to thrive in the era of information. Thus, as a complement to the *in vitro* and *in vivo* methods, computational toxicology is a powerful risk assessment tool for testing the chemicals toxicity; it aims at facilitating efficient simulation and prediction of environmental exposure, hazard, and risk of chemicals through various *in silico* models.

There are multiple *in silico* methodologies that are commonly integrated into the risk assessment process, such as QSAR modeling, which is one of such techniques that allows the development of mathematical correlation (usually statistically) between the chemical features (descriptors) and similar compounds.

I.1.1 Quantify Toxicity

Toxicity is one of the most difficult properties to modeling due for different reasons, for example, toxic effects are depend on multi parameters, such as species, organ, and time. In general, there is two way of experimental tests, *in vitro* test in which mostly they use single cells, and *in vivo* test by using an organisms, such as *fish*, *Daphnia*, etc. Aquatic acute toxicity tests are performed by exposing test organisms to the toxicant and observing their behavior for certain duration, at a predefined time. Every substance has the potential to become a lethal toxicant above certain doses or concentrations.

The common toxicity values predicted in QSAR modeling are expressed as a concentration at which 50% of the test species are killed by the toxic effect of the compound at a given time (e.g. IC_{50}, LC_{50}, EC_{50}, etc.) which can be obtained experimentally using the concentration-response curve as illustrated in Figure 1.

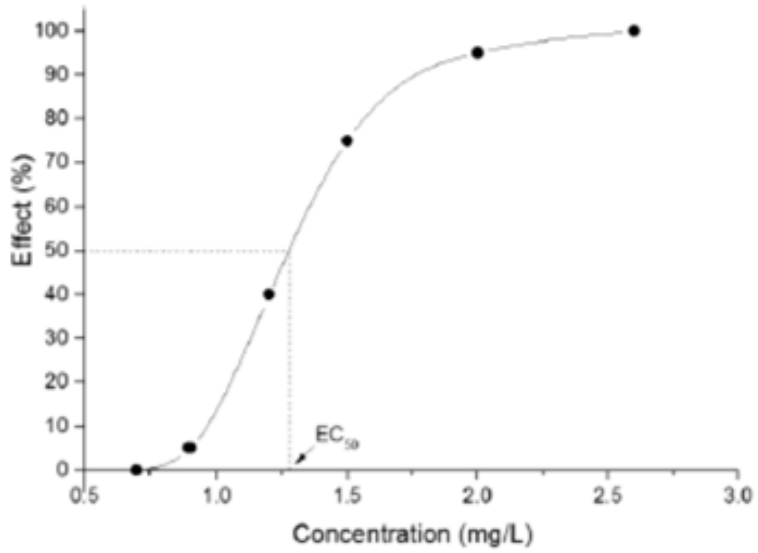


Figure 1.Example of to define EC50 from concentration response curve.

I.2. QSAR and the regulation of chemicals

The assessment of acute toxicity is an important component in the safety evaluation of substances and represents a standard information requirement within several legislative texts on chemicals.

There are many national and international efforts that have been peformed to assess the toxicity of the industrial organic compounds to humans and the environment. The first step in the assessment process is to offer the necessary information of each endpoint to be evaluated. In fact, *in vivo* testing is listed as the option of last resort in the European Union's Registration, Evaluation, Authorisation and restriction of Chemicals (REACH) legislation.

I.2.1. REACH regulation

REACH legislation is an European community regulation for ensure the safe of chemicals regarding the human health and the environment starting from June 2007 [22]. This legislation engaged for controlling of any substances imported or manufactured in the European market larger than 1 ton/year. However, over a decade REACH legislation and Organization for Economic Co-operation Development (OECD) have been encouraged the alternative ways of chemicals testing that do not use laboratory animals for ethical point of view and economic reasons [23]. Several molecular modeling methods that have been developed over the last years for predicting the toxic effects of industrial chemicals [24].

According to REACH requirements, the use of molecular modeling approaches such as quantitative structure-activity relationships (QSARs) are prioritized, to avoid unnecessary testing and reduce animal tests. QSAR can greatly help in early risk assessment as a data gap filling method. The main objectives of QSAR modeling in ecotoxicology are the classification of chemicals data based on mechanism of action and predict the messing data or design of a safe chemical before synthetized "a priori" [25].

I.2.2. OECD validation of QSARs (Five principles)

To assessing the validity of QSARs some guidelines were proposed in 2002 at the international workshop in Setúbal, Portugal as "Setúbal principles"[26]. However, those principles were adjusted in 2004 by the famous *Five principles* which were approved through the OECD in November 2004 by the OECD member countries at the 37th Joint Meeting of the Chemicals Committee and Working Party on Chemicals, Pesticides, and Biotechnology[27].

As given in OECD Guidance Document [28], the validation term is defined as follows:

''. . . the process by which the reliability and relevance of a particular approach, method, process or assessment is established for a defined purpose''.

The OECD principles identify the types of information that are considered useful for the assessment of QSARs in the regulatory purposes. They constitute the basis of a conceptual framework, but they do not in themselves provide criteria for the regulatory acceptance of QSARs. For the validation of QSARs under the REACH directives, the assessment of QSAR model validity should be performed by reference to the OECD principles, which is associated with the following information:

- A defined endpoint: it is necessary to confirm the transparency of the predicted endpoint and the modeled experimental system. A defined endpoint referring to a specific effect on a specific organ under precise conditions.
- An unambiguous algorithm: to ensure transparency in the description of the model's algorithm
- A defined domain of applicability: this principle expresses the need to justify that a given model is being used within the boundary of its limitations when making a given prediction.
- Appropriate measures of goodness-of-fit, robustness, and predictivity: those measures are provided in the two steps of the QSAR model development, internal validation to avoid the over-fitting and external validation for check the predictive ability.
- A mechanistic interpretation, if possible: that means it is not mandatory to provide a mechanistic interpretation of a given model. However, it can add a strong point to the confidence in the model already established on the previous principles.

The QSAR Model Reporting Format (QMRF) and the QSAR Prediction Reporting Format (QPRF) are the appropriate format for documenting the characteristics and validity of the model, and may be used to justify the adequacy of the QSAR prediction.

I.3. Toxicity Data Source

Biological experimental data provide the basis for the development of the quantitative structure activity relationships (QSARs). In the literature, there are several attempts devoted to develop models on the basis of a larger data set named "global QSARs". However, high quality and reliable toxicity data are required to develop a good QSAR model. This is possible only if the data will be obtained from a consistent and reliable protocol performed to the same

standards, and undertaken in the same laboratory and even by the same technicians [29]. In addition, the data obtained from multiple sources and generated by different methods was the subject of large debate and research, in which the problem is how to optimize the integration of data [7–9].

I.3.1. *In-house* Data

The *In-house* toxicity data that obtained from direct measurement is very helpful for developing *in silico* models, particularly in order to developing datasets for QSAR models. Thus, the traceability and the quality assurance can be checked, and the possibility of asking for an information and an experiment protocol from the generating data. Therefore, the use of this kind of data is very credible.

I.3.2. Books and Journal articles

Book and Journal articles are the traditional sources of the chemical and properties data since many decades. There is a variety of book sources that provide listings of data of environmental significance. For example, more than 10000 substances were listed in "Handbook on Physical Properties of Organic Chemicals" [30], with experimental and estimated physicochemical properties. On the other hand, many other books that can be found with toxicological information include the Handbooks of Ecotoxicological Data [31,32]. In addition, Supplementary information in journal articles is commonly being available from the publisher or authors as a source of toxicity information that can provide ready data sets for modeling.

I.3.3. Web Database

With the developments in the throughput screening and virtual screening, the amount of both the experimental bioassay data and computational physical and chemical data are increasing. Therefore, the storing and publishing this vast amount of data in a well organized way is becomes necessary. Thus, several computational algorithms are actively developed to organize and store this huge volume of available information, in the form of databases [33]. In addition, numerous publicly available compound databases that contain a large number of assay results which provide both active and inactive compound records are also available. The ChEMBL[34] and the PubChem[35] are the largest-scale compound and bioactivity databases obtained from the literature. The PubChem is an NIH funded effort that was initiated in 2004

to provide chemically annotated information for free to the scientific community. This database contains more than 157 million chemical records, with more than 1million compounds have been tested in about 3000 bioassays, and more than 500000 active compounds. The ChEMBL contains more than 1.5 million chemicals and nearly 14 million activity measurements for over 11000 targets.
Recently, several modern toxicological databases go much elsewhere simple data retrieval. They including the possibilities to modeling and make the predictive models, for instance:

a. QSAR Toolbox

One of the most important tools for QSAR data toxicity modeling software is OECD QSAR Toolbox that contains many databases and other data which given by various collaborative partners (Organisation for Economic Co-operation and Development (OECD), and the Laboratory of Mathematical Chemistry at the Bourgas University, Bulgaria and the European Chemicals Agency (ECHA). The OECD-QSAR Toolbox is a free software designed to make practical qualitative and quantitative structure–activity relationship based on predictions of toxicity. The Toolbox, provides information of chemicals based on "chemical category" concept in form of structure-searchable, [36,37].

b. OCHEM

The Online Chemical Database (OCHEM) [38] is a platform that consists of 2858801 chemical and biological records for 636 properties, collected from 13098 sources. The platform includes two main systems, namely, the database of properties measured experimentally and the modeling structure [39].

c. VEGA

The VEGA (Virtual models for Evaluating the properties of chemicals within a Global Architecture) platform [40], is a free platform that offers tens of models addressing physicochemical, environmental and ecotoxicological properties. In the VEGA-HUB platform, there are multiple tools, which are dedicated to the exploration and analysis of the properties of chemical substances [41]. The most commonly used ecotoxicity databases developed over several years are summarized in Table 1, in which more descriptions are given in reference[42].

Table 1.List of the most commonly used toxicity and Ecotoxicity Databases.

Databases	Web Accessibility
Danish (Q)SAR Database	http://qsar.food.dtu.dk/
Developmental Toxicity (DevTox)	http://www.devtox.org
Distributed Structure-Searchable Toxicity Database (DSSTox)	http://www.epa.gov/ncct/dsstox/index.html
ECOTOXicology Knowledgebase (ECOTOX)	http://cfpub.epa.gov/ecotox/
European Chemical Substances Information System (ESIS)	https://old.datahub.io/dataset/esis
Extension TOXicologyNETwork (EXTOXNET)	http://extoxnet.orst.edu/ghindex.html
eTox	www.etoxproject.eu/
FraunhoferRepDose	http://www.fraunhofer-repdose.de/
Gene-Tox	http://toxnet.nlm.nih.gov/cgi-bin/sis/htmlgen?GENETOX
Hazard Evaluation Support System (HESS) Attached Database (HESS DB)	https://www.nite.go.jp/en/chem/qsar/hess-e.html
Hazardous Substances Data Bank (HSDB)	https://toxnet.nlm.nih.gov/cgi-bin/sis/htmlgen?HSDB
Integrated Risk Information System (IRIS)	https://www.epa.gov/iris
Japan Existing Chemical Database (JECDB)	http://dra4.nihs.go.jp/mhlw_data/jsp/SearchPageENG.jsp
MDL	http://www.iop.vast.ac.vn/theor/conferences/smp/1st/kaminuma/ChemDraw/toxicity.html
National Toxicology Program (NTP)	http://ntp.niehs.nih.gov/
Organization for Economic Cooperation and evelopment (OECD)	http://www.oecd.org/chemicalsafety/riskassessment/echemportalglobalportaltoinformationonchemicalsubstances.htm
Optimized Strategies for Risk Assessment of Industrial Chemicals Through Integration of Non-test and Test Information (OSIRIS)	www.osiris.ufz.de
Risk Assessment Information System (RAIS)	http://rais.ornl.gov/
Toxicology Testing in the 21st Century (Tox21)	http://www.epa.gov/ncct/Tox21/
ToxCast	https://www.epa.gov/chemical-research/toxcast-ashboard
Toxicology Data Network (TOXNET)	http://toxnet.nlm.nih.gov/
Toxicity Reference Database (ToxRefDB)	http://www.epa.gov/comptox/toxrefdb/
Toxic Substances Control Act Test Submissions (TSCATS)	https://catalog.data.gov/dataset/toxicsubstances-control-act-test-submissions-2-0-tscats-2-0/resource/fbe133b5-d0bd-4c2c-a290-fd4deec4a5b9
US FDA Chemical Evaluation and Risk Estimation System (CERES)	https://www.accessdata.fda.gov/scripts/fdatrack/view/track_project.cfm?program¼cfsan&id¼CFSANOFAS-Chemical-Evaluation-and-Risk-Estimation-System
VITIC	http://www.lhasalimited.org/products/vitic-nexus.htm
WikiPharma	www.wikipharma.org

I.4. Inter-species hybrid QSARs

I.4.1. Inter-species Quantitative Activity–Activity Relationships (QAARs)

In toxicology, alternative species in the risk assessment of chemicals is one of the most used methods. This is for two reasons; the first one is for realize the REACH recommendation, which is for reducing the *in vivo* (animal) test. The second one is consists an alternative of the experimental determination of some species properties that is a more costly and time intensive process, compared to other species. Therefore, inter-species quantitative correlation (QAAR) is a promising field that has received little attention which aims to predict one species from an endpoint in another species when the data is missing. The possible reasons relate to the limitations of the technic in the quality of the used data [6,43–48]. Generally, regression analysis is the simplest method for developing the linear inter-species relationship. The model of this method has the form:

$$C_1 = aC_2 + c \qquad \text{Equation 1}$$

Where,

C_1 is the endpoint of the species to be replaced.

C_2 is the endpoint of the alternative species.

a is the regression coefficient.

c is a constant.

I.4.2. Quantitative Structure–Activity–Activity Relationships (QSAARs)

Quantitative structure-activity-activity relationships (QSAARs) is another extrapolating technique that has been used occasionally, that may be to the complexity of their model which is based on a hybrid approach, merges the inter-species and QSAR approaches and based on the relationship between two different biological endpoints [49–52]. QSAARs attempt to improve inter-species relationships (QAARs) by the addition of molecular descriptors that are commonly applied in QSAR analysis. The general form of QSAAR is:

$$C_1 = aC_2 + \text{Descriptor (n)} \ldots + c \qquad \text{Equation 2}$$

Where,

As in the case of QSAR, it should be cautioned not to add too many descriptors to avoid the bias of the equation [52]. Both techniques (QAAR and QSAAR) are used to extrapolate from toxicity data for one endpoint a predicted toxicity at another data, especially for chemicals in which there is a limited toxicity.

PART II.TOOLS AND METHODS

II.1. QSAR Methodologies

Quantitative Structure-Activity Relationships (QSAR) is an intersection between bioinformatics and cheminformatics providing an effective means for exploring and exploiting the relationship between chemical structure and its biological action towards the development of the models [53].

The fundamentally of QSAR methodology is to establish a mathematical equation relationship between the variance in molecular structures encoded by molecular descriptors and the particular activity or property associated with them[54]. The equation relationship (QSAR models Eq.03) can be used to predict data for untested compounds with a lack of experimental data [55].

$$\text{Response} = f(\text{Descriptors/ Features}) \qquad \text{Equation 3}$$

The main objective of QSAR modeling is the development of predictive models based on the data set of compounds, in which this model can be used for the prediction of activities or properties of new compounds, that not involved in the building up of this model[56]. However, prior to using sash model, and in order to make a reliable and predictive model, this last should be checked by internal validation, and assessed in terms of predictive power, according to recently stated OECD principles [57].

Historically, QSAR dates back to 1863 in the thesis of Cros [58], entitled "Action de l'alcool amylique sur l'organisme" which noted a relationship between toxicity of primary aliphatic alcohols and their water solubility. However, the pioneering contributions of Corwin Hansch and his collaborators in the 1960s [59] related to QSAR analysis marked the official borne of this field. After that, the attention in QSARs has increased progressively.

II.1.1. Development of QSAR

In general, the development of any QSAR model is mainly divided into five steps:

i. Dataset collection, structures, and endpoint activity or property.
ii. Descriptors calculation.
iii. Data splitting into training and test set.
iv. Model building (modeling algorithms).
v. A statistical validation.

The pathway of QSAR model development are illustrated in Figure 2 [60,61].

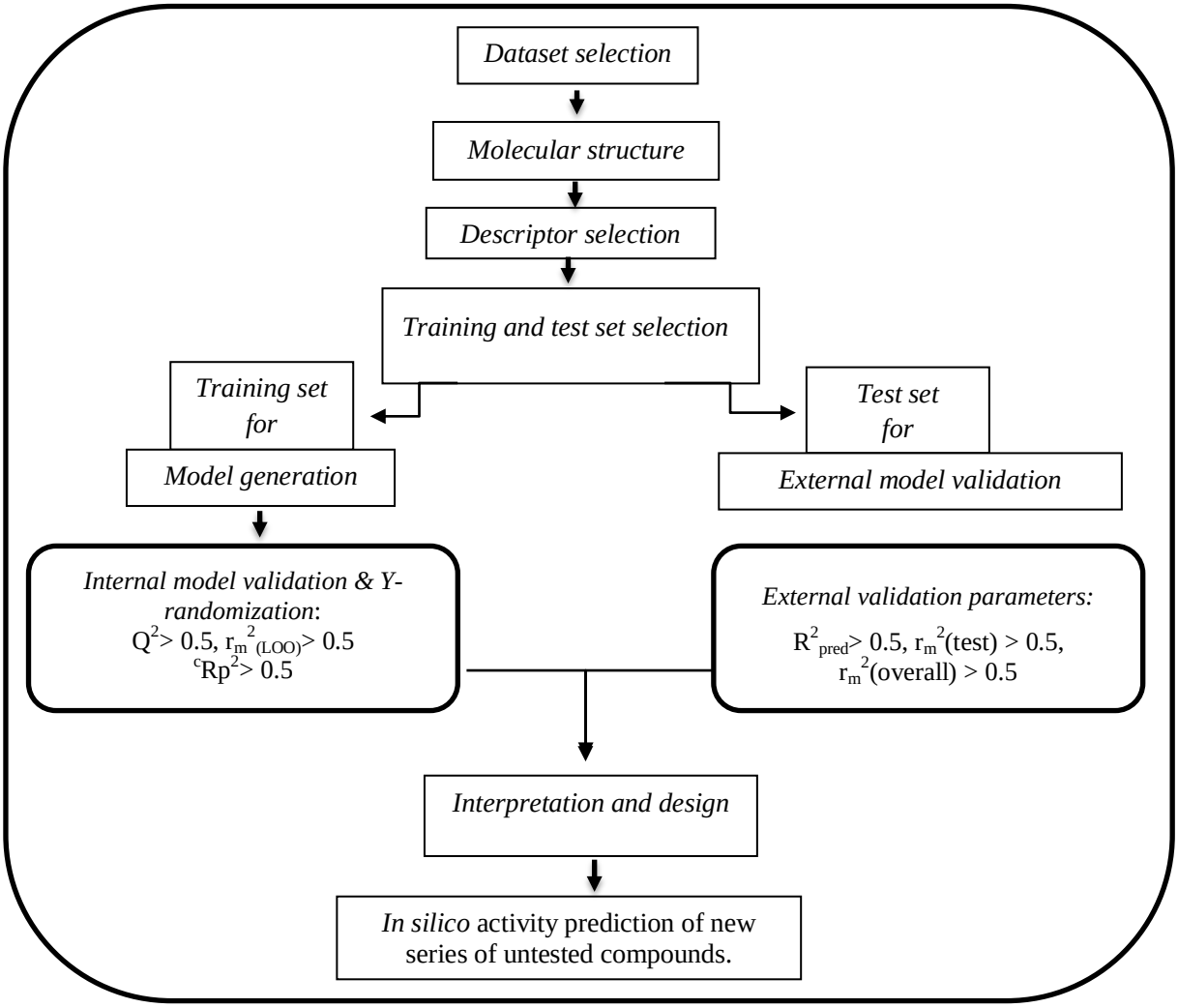


Figure 2. An illustration of workflow concept used in QSAR modeling.

The collection of the molecules with the known experimental biological activity is the first step of any QSAR modeling. This step may be the most delicate part because some conditions needs to be fulfilled, such as, the similarity of the compounds, the same experimental protocol. The data set of structures is considered as the input of molecular descriptors calculation. In some cases, geometric optimization is more important before descriptors calculation.

The next step is the splitting of data into training and test sets using different approaches (clustering, sphere exclusion, and activity ranking). However, random splitting is the most used technique which may be performed several times in order to get an average

representation. The use of a suitable training set is very important since the chemical space and the size of data will affect on the following steps. In the modeling part, the correlation between the biological activity values and descriptors values of the training set is realized using several machine learning include, for example, Random forest, SVM, and neural network.

II.1.2. KNIME platform for QSAR modeling

The KNIME [62] (Konstanz Information Miner) is an open-source workflow technology with a graphical user interface based on collections of node known as ‘‘extensions’’ that allow the data process and its transported *via* connections between those nodes [63]. KNIME provides an easy visual assembly workbench that enables scientists to create and visualize complex workflows easily[64]. In addition, it not limited to the ability to analyze the results, thus, it might comprise several processing and analytical steps, including, statistical analysis and data visualization and data mining on experimentally[65]. Also, it supports a wide range of functionality and has an active cheminformatics/ bioinformatics community.

Due to the huge cost and time necessary in manual procedure analysis of the big data, the improvement of the workflow like KNIME power on chemical structure curation procedures, or for building up workflow models based on several chemistry community nodes, to calculate and predict the physicochemical properties and biological activities is become necessary. The concept used in QSAR/QSPR modeling is represented in Figure 3.

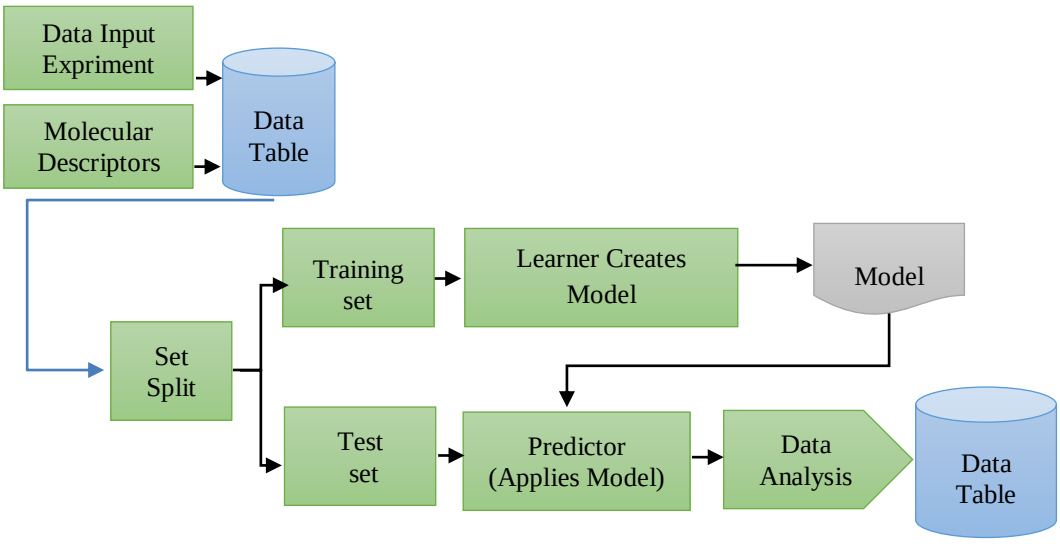


Figure 3.An illustration of the KNIME workflow concept used in QSAR modeling.

II.1.2.1. Chemical data curation workflow

In QSAR modeling, the collection of Chemical structures are often gathered from a variety of public and private databases, which considered among the most important and delicate steps. These databases contain records for thousands of chemicals, biological, and physicochemical properties, in which each chemical generated automatically from SMILES, 2-D, and 3-D structures [66]. There is a possibility that some of the chemical structures were not translated correctly or presence of molecules in salts format. So to avoid these problems, Tropsha *et al*, [67], have integrated several protocols into the standardized chemical data curation strategy. The workflow implemented on KNIME platform by the buildup of a workbench for visual assembly and interactive execution of data pipelines, following the most important steps required in curate a chemical data is summarized in Figure. 4.

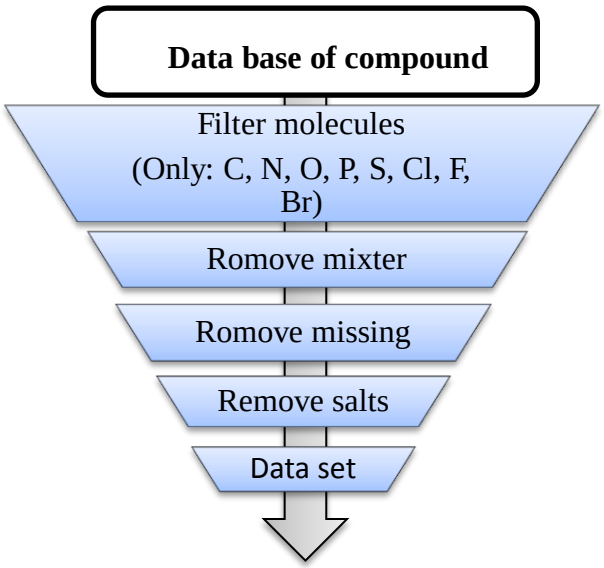


Figure 4. The important steps required to curate a chemical data set on KNIME platform.

There are different cheminformatics nodes specific for various chemical tasks (reading, calculating and writing…etc) are now being added over time, like CDK[68], Indigo [69] RDKit,[70] and ErlWood [71].

By carefully following the instructions of data curation protocols described previously, and using chemical community nodes, we can build up a KNIME workflow (Figure 5) as following:

- Exclude some structures with missing experimental log *P* values,
- Removing hydrogen atoms,

- Removed inorganic chemicals and chemical mixtures,
- Removed chemicals that contained atoms other than C, H, O, N, F, Cl, Br, I, S, P. (Because most software can calculate molecular descriptors only for organic structures [55]),
- Optimizes the geometry of molecules, and aromatized it, if necessary.

The major node descriptions used in data curation are summarized in Table 2.

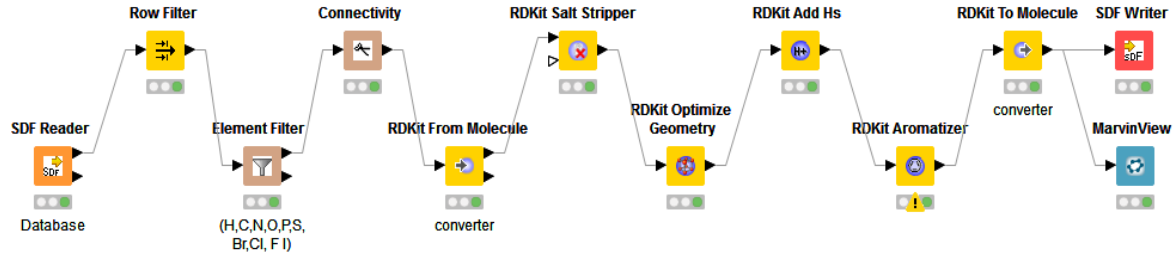


Figure 5.A KNIME workflow illustrating the various steps in data curation strategy used in our studies.

Table 2. Names and descriptions of major nodes used in data curation.

Node names	Node description
Row filter	The node allows row to exclude missing values or to use a range.
Element filter	Keep either molecules that only have types C,H,N,O,P,S, Br, F, I.
Connectivity	This node is used to completely remove the compounds that contain unconnected molecules.
RDKit Salt Stripper	This node is used for removing salts.
RDKit Optimizes geometry	Optimizes the geometry of molecules.
RDKit Add Hs	Adds hydrogens to molecule.
RDKit Aromatizer	Aromatized RDKit Molecules
MarvinView node [72]	is an advanced chemical viewer for single and multiple chemical structures,

II.1.2.2. Model workflow development

There are four main steps that must be taken during the preparation of workflow; data processing (input, output, and converts between various data types), data curation, calculation of descriptors and chemometric methods. In our case, multiple linear regression analysis was performed using KNIME node for data mining, followed by predictive ability (see Figure 6).

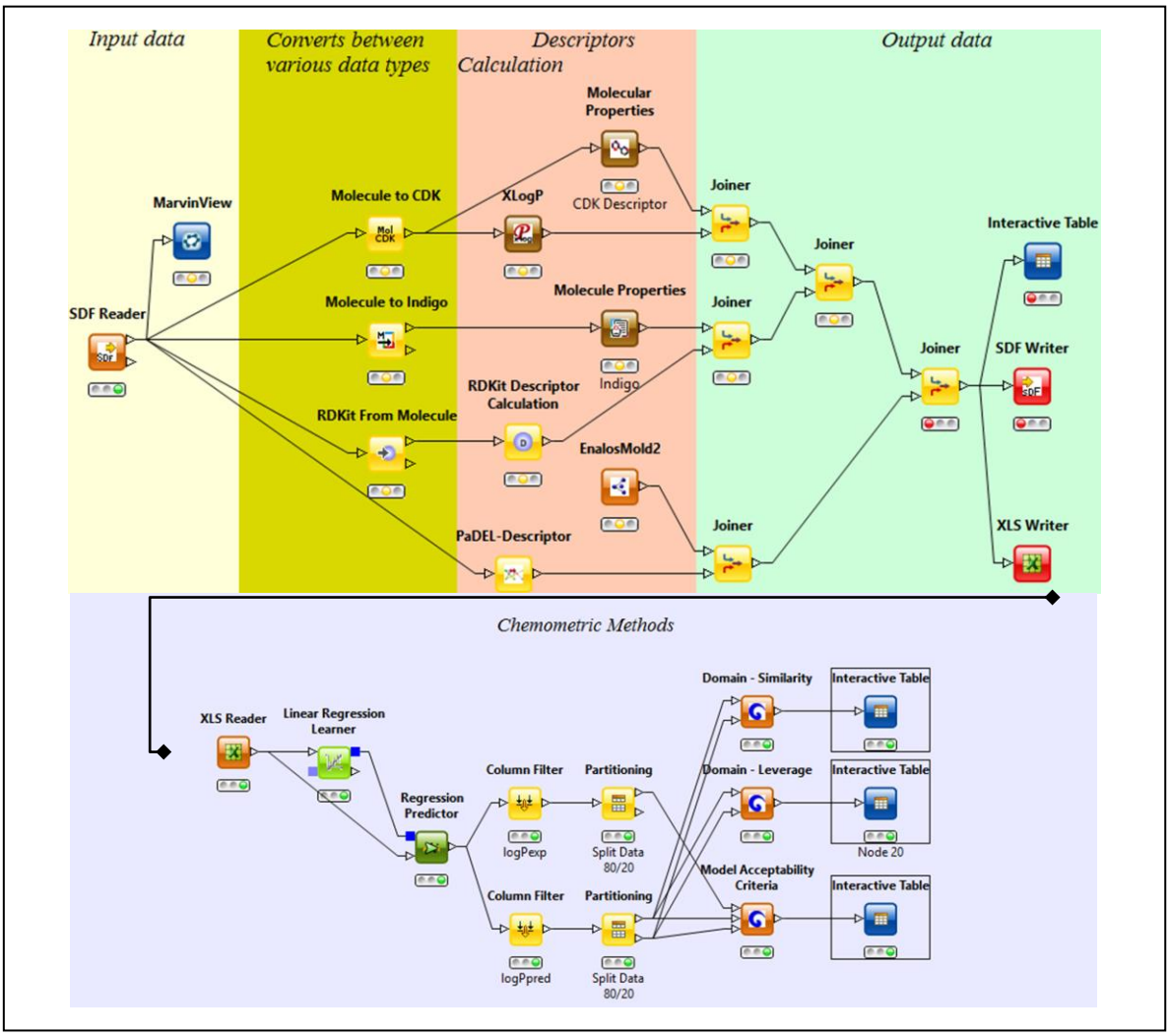


Figure 6. KNIME workflow illustrating the data processing in the Automated QSAR process.

II.2.Representation of molecules

The amount of information stored in molecular databases is increasing constantly, there is a need to generate simplifications of the molecular representation of compound databases to open new approaches for study the chemical space, optimize the storage and enhance the speed of computations. Thus, the chemical compounds can be represented by different methods using diverse rules and criteria depending on the molecule representation [73].

SMILES Notation

SMILES (Simplified Molecular Input Line System) [74] is a system for chemical structure encoding, that is developed in the late 1970s by David Weiniger, which was introduced as a simplified format to represent small molecules in two-dimensions; C representing a carbon atom, N a nitrogen atom, and so forth. Double and triple bonds are immediately apparent by the = and # symbols[75,76].

The four simple rules to apply for SMILES string are:

1. Atoms are represented by atomic symbols.

2. Double bonds are represented by = and triple bonds are represented by #.

3. Branching is represented by parentheses.

4. Ring closures are indicated by pairs of matching numbers.

File formats

1- **Molfile:** Mol was created by MDL, it is the most supported format used in chemoinformatics. In this format, the information of the atoms, atomic bond, connectivity and the coordination of the molecules are including.
2- **SDF file:** SDF is an extension of MDL, in which in this format more information can be added by MDL in the second portion of the files.

Fingerprints

The fingerprint format is a bit-wise string, consists of a sequence of ones and zeros where a one or zero in a specific position indicates the presence or absence of structural fragments

(Figure 7). The advantage of this format is the increasing of calculation speed and reducing storage space. Examples of frequently used fingerprints implemented in specific tools are MACCS and PubChem keys.

- MACCS keys, the Molecular ACCess System, are created by Molecular Design Limited. This descriptor encodes the atoms types, rings, and bond information, generated in a 166 key-bits format [77].
- The PubChem binary substructure keys are developed to be used by the PubChem database in order to perform the searching queries [78]. The length of this string is 881 bits.

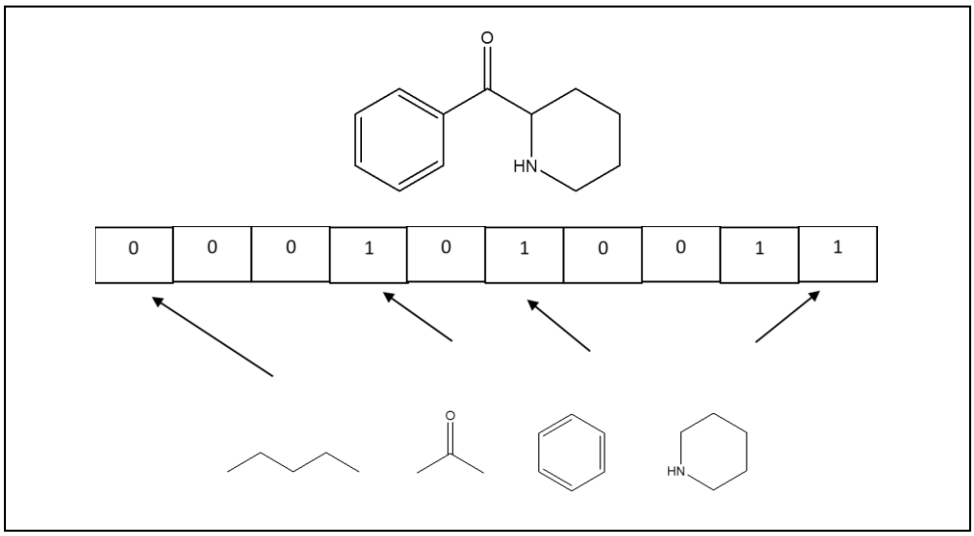


Figure 7.Encoding molecular structures in a bit string

II.3. Molecular descriptors

Molecular descriptors are numerical representations of chemical features that are encoded the information of the chemical structure of a molecule. The numerical values are derived by some algorithm describing a specific aspect of a compound. Today, there are more than 5,000 descriptors, which classified into three categories depending on the method of their computation or determination; 1-D encoding chemical composition, 2-D encoding topology and 3-D descriptors encoding shape and functionality.

- **One dimensional-Structural Keys:** A substructure of a molecule can be considered as a 1D molecular representation. One-dimensional molecular representation consists of molecular fragments such as functional groups, rings, bonds and substituents. If a certain fragment is present in a molecule, a particular bit in the string is set to one, otherwise to zero. Thus, each bit in this array encodes a particular fragment.
- **Two dimensional-Topological Descriptors:** A 2D molecular descriptor contains topological information which describes the bonding of atoms in a molecule by elucidation the type of bonding and the interaction of particular atoms.
- **Three dimensional-Geometric Descriptor:** The 3D descriptors are calculated from a geometrical representation of a molecule. Those descriptors are usually provided more information than the 2D and 1D descriptors.
- **Other descriptors:** Further groups of geometric descriptors widely used in QSAR studies are derived from the molecular surface, molecular volume, and other geometrical properties.

Todeschini and Consonni [79] in the textbook Molecular Descriptors gave a comprehensive and detailed overview of all kinds of molecular descriptors for Chemoinformatics. In this regards, there is various software for calculating the molecular descriptors, namely, CODESSA, DRAGON, and many other freely available software like PaDel and CDK. Table 3 summaries the most know software used in QSAR modeling.

Table 3. Available Software and Web Services used in QSAR modeling.

Software name	Full name	Web Accessibility
CORALSEA	Freeware to build quantitative structure–property/activity relationships (QSPR/QSAR)	http://www.insilico.eu/coral/CORALSEA.html
BlueDesc	Open-source descriptor calculator	http://www.ra.cs.uni-tuebingen.de/software/bluedesc/welcome_e.html
E-DRAGON	Online descriptor calculator	http://146.107.217.178/lab/edragon/
Mold2	Free descriptors generator software	http://www.fda.gov/ScienceResearch/BioinformaticsTools/Mold2/default.htm
CODESSA	COmprehensiveDEscriptors for Structural and Statistical Analysis	http://www.codessa-pro.com
PaDEL-Descriptor		http://www.moleculardescriptors.eu/resources/resources.htm
OChEM	Online Chemical Modeling Environment is a web-based platform for QSAR modeling	http://www.ochem.eu
DRAGON		http://www.talete.mi.it/products/dragon_description.htm
CDK	Chemistry Development Kit	
MOE		www.chemcomp.com
alvaDesc		https://chm.kode-solutions.net/products_alvadesc.php
MOLCONN-Z		www.edusoft-lc.com/molconn
MOLD2		www.fda.gov

II.4.Statistical and modeling methods in QSAR

II.4.1. Validation and metrics of the model

In general, there is no role for the number of compounds is required to develop a meaningful relationship, however, it is widely accepted that for generating every descriptor in a QSAR, a five until ten compounds are required [80]. Technically, much more compounds are essential to obtain statistically robust QSARs. On the other hand, with the increasing number of computed molecular descriptors using various software and in the case when a lack of a sufficient number of dataset, the risk of chance correlations increases considerably with the number of tested variables and with the number of variables included in the final model in comparison with the number of compounds that used for generating the model [81]. Therefore, the validation of the developed models is required.

The validation approaches seek to check the reliability, the acceptability and the predictivity of the developed model [82], which plays a crucial role in defining the applicability of the model for the prediction of designed molecules [83]. The validation of different quality metrics can be categorized into two classes. The first class is internal validation like leave-one-out and leave-five-out cross-validation procedures and Y-randomization. The second class is external validation based on the test set compounds. Both techniques have been extensively used by diverse groups of researchers for assessing the predictive ability of the developed model [60].

II.4.1.1. Internal validation

The internal validation is based on the molecules used in the QSAR model development [84,85].The cross-validation technique mainly involves internal validation, where a sample of n observations is partitioned into calibration and validation subsets. The calibration subset is used to construct a model, while the validation subset is used to test how well the model predicts the new data that not used in the calibration procedure. The aim of the internal validation is to evaluate the model robustness by perturbing the training set. The models are required to be internally cross-validated during the modeling process.

a. Cross-Validation

Cross-validation is the most common technique used for checking the reliability of statistical models. In this technique, a certain number of sets are created by eliminating one compound from the set LOO (leave-one-out) or a small group LMO (leave-many-out) of compounds. For each set, a new QSAR model is then rebuilton the remaining set and the compounds that have not been used in model development can be used for evaluation of the model predicting based on the resulting equation [86,87].The metrics used for the judgment of the LOO and LMO are the predicted residual sum of squares (PRESS) and cross-validated Q^2 for the model[88]. These metrics are calculating according to the following equations:

$$\text{PRESS} = \sum(Y_{\text{obs}}\text{-}Y_{\text{pred}})^2 \qquad \text{Equation 4}$$

$$Q^2 = 1\text{-} \frac{\text{PRESS}}{\sum(Y_{\text{obs(train)}} - \bar{Y}_{training})^2} \qquad \text{Equation 5}$$

b. Y-Randomization (scrambling)test

This technique is also known as Y-scrambling and it is considered in an internal validation test, which is used for check the robustness of the models[84,85,89].In this test, validation is performed by permuting the response values (Y) with respect to the X matrix which it kept unaltered. The procedure is repeated several times and the new models are expected to have low R^2 and Q^2 values. ${}^{c}R_P^2$ is a parameter used for this purpose witch computed from the following equation [90]:

$${}^{c}R_P^2 = \text{R X}\sqrt{R^2 - R_r^2} \qquad \text{Equation 6}$$

The degree of variation in the values of the squared mean correlation coefficient of the randomized model (R_P^2) and squared correlation coefficient of the nonrandom model (R^2) is reflected in the value of the ${}^{c}R_P^2$[91].

c. Bootstrapping

Bootstrapping [92] is another validation technique based on the random splitting of the data set several times into training and test sets. However, contrary to cross validation, in this validation technique, LOO and LMO exclusion of the compounds may be excluded once, or

several times, as well as never. The developed model is used to predict the remaining excluded compounds and the average Q^2value is also calculated.

II.4.1.2. External Validation Set

The mean aim of internal validation is for verifying internal predictivity, and for check the stability of the models, however, this kind of validation cannot achieve true external validation[93]. Golbraikh and Tropsha [89] have shown that the predictive power of QSAR models can be claimed only if the model was successfully applied for the prediction of the external test set compounds (the compounds that not used in the model development). Since the past decade, a huge debate has been discussed on the applying of internal or external validation to ensure the predictivity of the models [94].

To reflect the quality of predictions, various metrics based on Golbraikh and Trospsha's criteria[84,85] have been proposed [95]like Q^2_{F1},Q^2_{F2}, Q^2_{F3},and r^2_mas detailed by the following equations:

$$Q^2_{F1} = 1 - \frac{\sum(Y_{obs(test)} - Y_{pred(test)})^2}{\sum(Y_{obs(test)} - \bar{Y}_{training})^2} \qquad \text{Equation 7}$$

Q^2_{F2} proposed by Schuurmann *et al*. [96] , given by:

$$Q^2_{F2} = 1 - \frac{\sum(Y_{obs(test)} - Y_{pred(test)})^2}{\sum(Y_{obs(test)} - \bar{Y}_{test})^2} \qquad \text{Equation 8}$$

$$Q^2_{F3} = 1 - \frac{\sum(Y_{obs(test)} - Y_{pred(test)})^2/n_{test}}{\sum(Y_{obs(train)} - \bar{Y}_{train})^2/n_{test}} \qquad \text{Equation 9}$$

$$r^2_m = r^2 . (1-\sqrt{(r^2 + r^2_0)}) \qquad \text{Equation 10}$$

Model Acceptability Golbraikh and Tropsha's criteria

In order to check the predictive ability of any QSAR model, Golbraikh and Tropsha proposed a set of statistical criteria as follow:

$q^2 > 0.5$

$R^2 > 0.6$

$\frac{(R^2 - R_0^2)}{R^2} <$ 0.1 and $0.85 \leq K \leq 1.15$

$\frac{(R^2 - R`_0^2)}{R^2} <$ 0.1 and $0.85 \leq K` \leq 1.15$

$|R^2 - R`_0^2| <0.3$

R^2 : Correlation coefficient between the predicted and observed activities.

q^2: External cross validation.

R_0^2: Coefficient of determination (determined the predicted versus observed activities).

$R`_0^2$: Coefficient of determination: observed versus predicted activities.

k = slope: predicted versus observed activities regression lines through the origin.

k' = slope: observed versus predicted activities regression lines through the origin.

The model is acceptable if the leave-one-out cross-validated R^2 (q^2) values were greater than 0.5 for the training sets and the correlation coefficient R^2 values were greater than 0.6 for the test sets.

II.4.2. Applicability domain of models (AD)

According to the OECD principle 3 "*The applicability domain (AD) is a theoretical region in chemical space, defined by the model descriptors and modeled response, and thus by the nature of the chemicals in the training set, as represented in each model by specific molecular descriptors"*, the application of the developed model is for making possible the prediction of the new compounds within a specific domain. As a result, if a test set is found outside AD for a particular model, then that model is excluded [97]. The AD can be defined as either a priori regardless of model descriptors or it can be assessed a posteriori on the basis of the molecular descriptors of the training set. In the literature, a variety of approaches was proposed, in which a part of them are based on the interpolation space used by model descriptor space, while the other part are based on the response space of the training set molecules [98–100]. The most important of those approaches are the followings:

1. Ranges in the descriptor space
2. Geometrical methods
3. Distance-based methods
4. Probability density distribution
5. Range of the response variable

The most AD technique used in QSAR modeling based on the distance-based method is the leverage. It determined by two cut-offs, in which the first is the diagonal values hat matrix (h). If the compounds in the training set have h greater than the critical hat values h* (h*= 3p' /n, where p' is the number of the model variables plus one, and n is the number for training compounds), this means that these are structurally very influential in determining model parameters [94]. The second one considers outliers detected based on standardized residuals of compounds when the cross-validated standardized residuals are more than 2.5 standard deviation units. The Williams plot is the easy common graph for visualizing ~~of~~ the leverage methodology.

The leverage approach has also been applied to investigate the degree of inter species extrapolation for the predictions of compounds. The Insubria graph can visualize the interpolated and the extrapolated predictions from species to species [101]. The Enalos Nodes can be used to define the applicability domain [102–104]. The Enalos Domain-Similarity node is based on the Euclidean distances [93] and the Enalos Domain-Leverages node is based on the leverages [105]. The applicability domain KNIME workflow is available also in our previous work [106].

II.4.3. Feature Selection

Currently, there are over 5000 molecular descriptors can be calculated by means of dedicated software within few seconds. However, only a certain number of descriptors are correlated to the response. As a result, there are a limited number of descriptors that could be used in QSAR model generation. Hence, to avoid over fitting and allow the model interpretation, the feature selection techniques is the best answer for those problems. Thereby, a different algorithms for variable selection have been proposed in literature, such as Genetic Algorithms (GAs) [107–109], Ant Colony Optimization (ACO), All Subset Models (ASM) and Sequential Search (SS).

a. Genetic Algorithms (GAs)

GAs are widely used in QSAR modeling to find the optimal subsets of descriptors [110], in which the idea behind this algorithm is based on the natural inspiration evolutionary to optimize the searching methods [111]. Thus, the gene is corresponding to a descriptor and a sequence of genes or the chromosome is corresponding to a model. First, the absence and

presence of the variable are represented as a binary victor and the chromosomes population is randomly initialized. Then, the second step is the building of the model according to the predefined fitness function (for example Q^2 or R^2). The next step is the reproduction stem for the creation of the child population from the parent's witch are randomly selected. After, basing on their scaled fitness scores, the crossover children are produced from the parents by crossover and mutation children through the mutation. These two operations crossover and mutation processes are repeated until a stop criterion is satisfied. Figure 8 provides a scheme for the algorithm of Gas [112].

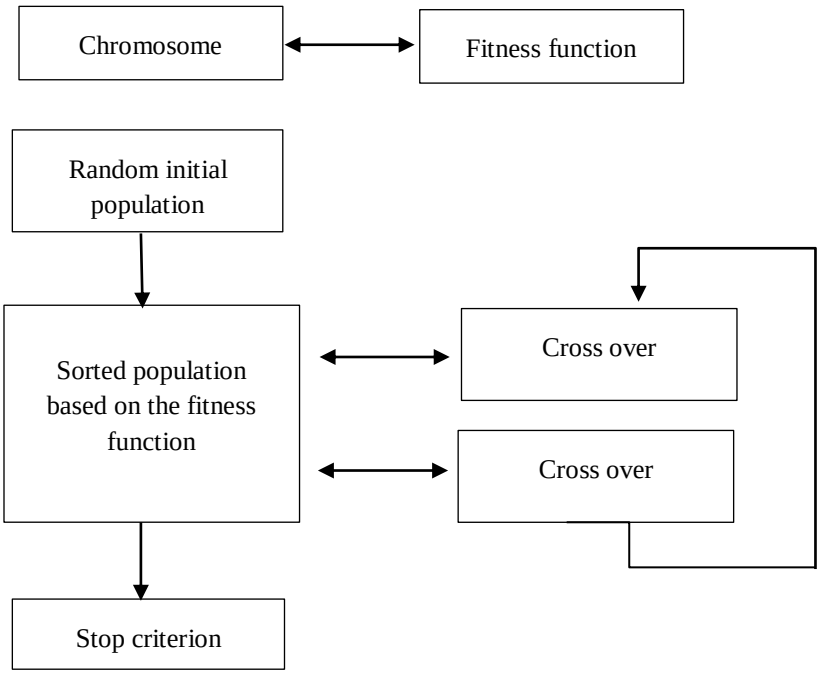


Figure 8. Genetic function workflow

b. Stepwise methods (SW)

Stepwise regression method is another feature selection, which is considered among the most known subset selection methods used in QSAR. Stepwise regression is based on two different strategies, namely, Forward Selection (FS) and Backward Elimination (BE).

c. All Subset Models (ASM)

The All Subset Models (ASM) is the simplest method of selection, however, this technique is very computationally expensive. It consists of the generation of all the possible combinations of the *p* variables in the size from 1 to *p*. The best subset can be reached in a reasonable time, but, for a large numbers of variables, it takes a long unknown time.

d. Sequential Search (SS)

Sequential Search (SS) [113] is another simple method aimed to find the optimal subsets of variables for a specified model size. In this method, each variable is replaced at a same time with all the remaining variables until the best model is obtained.

e. Ant Colony Optimization (ACO)

Ant Colony Optimization (ACO) [114,115] was inspired from the colonies of ants, who look for the shortest path connecting their nest and the source of food by pheromone deposition, in which they deposit pheromone from the home to a food source. Subsequent ants will generally choose paths with more pheromone and after many trials, they will converge on an optimal path.

II.5. Machine Learning methods

Computational modeling including ML techniques play an increasingly important role in chemoinformatics, especially QSAR modeling. Modern QSAR analysis takes advantage of a number of advanced Machine ML. The role of ML usually is the extraction of the most important feature selection by exploring the descriptors combinations. Several ML methods are already used in various aspects of QSAR modeling including Random Forest, SVM and Deep Learning [116].

II.5.1. Support Vector Machines (SVM)

SVM is a machine learning technique that was originally designed to solve classification problems by using nonlinear kernel functions to map data into high-dimensional space through finding an optimally separating hyperplane [117]. After that, SVM has been generalized for application to continuous values. In QSAR modeling, several works based on

SVM have been published [118]. This method is used to compare between deep learning and SVM approaches.

II.5.2. Random Forest (RF)

The Random Forest technique is one of the best methods used in QSAR in terms of accuracy of prediction in comparison with the other ensemble learning models [119]. In random forest methodology, each recursive partitioning "tree" trained on subsampled subsets of compounds characterizes a consensus nonlinear model derived from a large number of single models. The major advantages of this method are quite resistant to over fitting and time-consuming [120].

II.5.3. Deep Learning (DL)

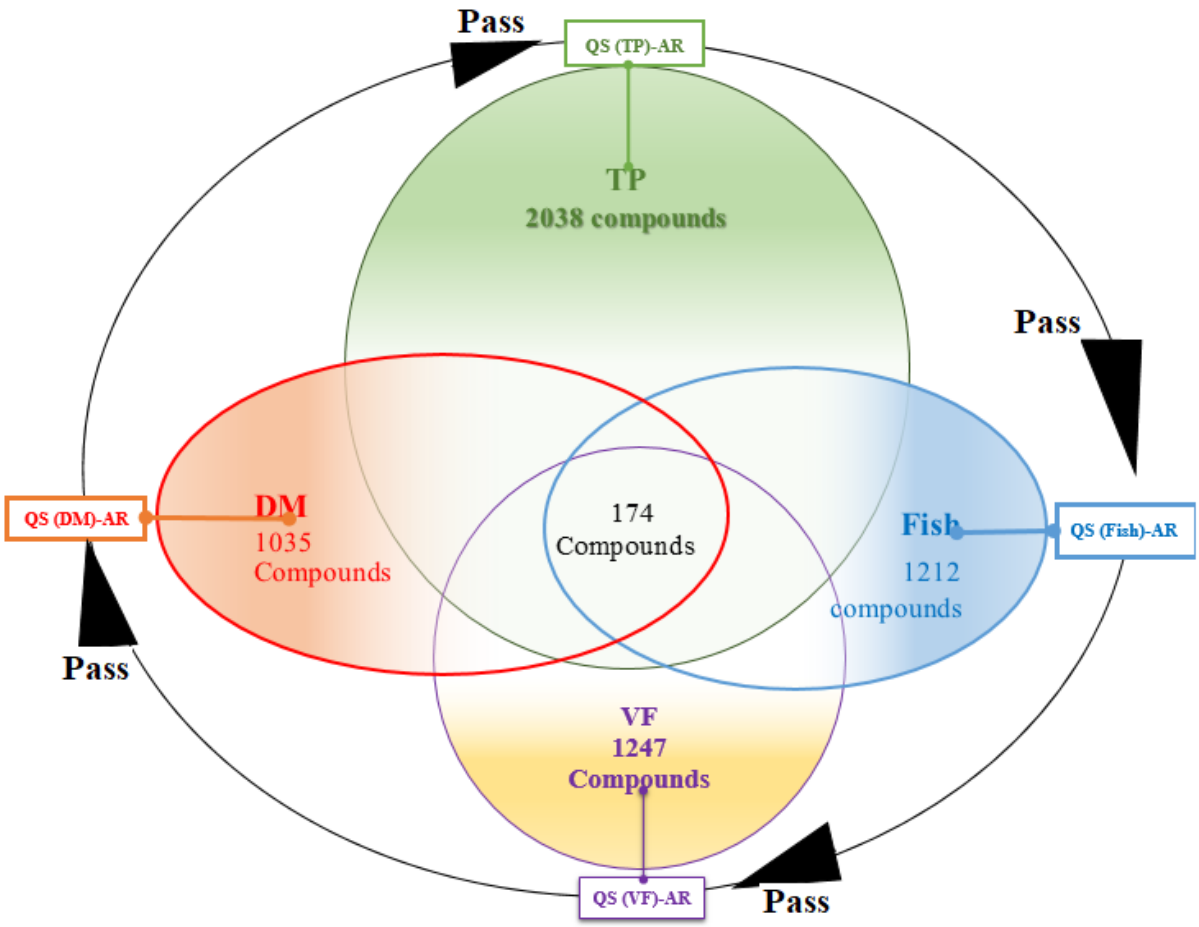
Deep learning algorithms become the most exciting research area in machine learning. It is an automatic general-purpose learning procedure, which has been commonly adopted in many fields [121]. Currently, an increasingly number of approaches are applied in QSAR, based on deep architecture, form molecular feature extraction. Some of these approaches are based on images of molecules, another some on traditional molecular descriptors and less commonly on their SMILES strings. For instance, Panteleev *et al*. reviewed the recent developments in machine learning application in drug discovery notably, new deep learning-based [122]. Uesawa, *et al* used a deep learning approach by incorporating 360° images of molecular conformations for extracting the feature representation learning in QSAR analysis [123]. Heo *et al*. have used deep-learning-based QSAR models basing on molecular descriptors to predict the qualitative and the quantitative effects of endocrine disrupting chemicals (EDCs), in which their results were compared with MLR and SVM methods [124]. Ghasemi *et al.* performed a study, in which a deep belief network were used to investigate the model performance on the Kaggle data set [125]. Fernandez *et al*. have developed a new deep learning tool that can automatically extract and learn toxicity-related structural features of chemical compounds from their graphic images for predicting 12 biological endpoints described in the Tox21 challenge [126]. Zhang *et al*. compared the performance of the predictive of DNN algorithms models and random forest model for Human ether-a-go-go related gene (hERG) activity, in which the structural and physicochemical properties were used in this study. [127]. Benfenati *et al*. used deep learning algorithms to find a good accuracy of the Ames test for mutagenicity, only from the SMILES notation of the chemicals

[128]. Hirohara *et al.* used a convolutional neural network for TOX 21 dataset based on SMILES representation of compounds [129].

The public attention to the use of deep learning in QSAR was in 2012 when Dahl's team won the Merck Molecular Activity Challenge by applying a deep learning approach on the Kaggle dataset (www.kaggle.com) using a large number of 2D topological descriptors to capture complex statistical patterns [17]. In 2014, another deep learning models were also developed to win the entry in NIH Tox21 Data Challenge. Those DL models showed a best predictive performance [130]. Since that time, numerous research groups have used with success the deep learning models to predict many parameters, including activity, toxicity, solubility, and various other proprieties[18–20].

The ANN and DL that use larger numbers of hidden layers of nonlinear processing units, in which each layer is used to predict the biological activity of compounds. Typically, we can find four popular NN architectures which used in DL, deep neural network (DNN) [131], convolution CNN [132], recurrent neural network (RNN) [133], and autoencoder (AE) [134].

III.1. QSAAR and QAAR Modeling: application in aquatic toxicity



III.1.1. Introduction

In the environmental risk assessment, there are several processes for determining the potential benefits and the side effects of exposure to chemicals. However, the development of the toxicological concept, such as high throughput screening has lead to a growth in the number of assays. These assays have provided an explosion in the number of pathway-related data available, which could be used to develop new computational models. *In silico* approaches, the toxicity prediction from structure have been used for many decades because they provide a faster alternative to otherwise time-consuming laboratory testing methods. QSAR is the most successful technique that can be used for chemical risk assessment for the protection of human and environmental health, which makes them interesting to regulators, especially, in the absence of experimental data.

The Inter-species Quantitative Structure–Activity Relationships[52,135] is a particular QSAR technique based on the extrapolation of data toxicity against spices to those for another species.

When the toxicity values of defined chemicals for one endpoint correlate well with the values for another endpoint, the chemicals can be expected to have similar modes of action with respect to both endpoints. In contrast, if the toxicity values of defined chemicals for one endpoint are not correlated with the values for another endpoint, the chemicals can be expected to be different mechanistic categories with respect to the two endpoints. Thereby, these chemicals may show species-specific toxicities.

In this work, huge datasets have been used to evaluate the aquatic toxicity against four species, namely, *fish*, *D. magna*, Algae *T. pyriformis* and *V.fischeri*. The models were developed based on QSAR modeling using two approaches, which are QSAAR (Quantitative Structure-Activity-Activity Relationship) and QAAR (Quantitative Activity-Activity Relationship). A new workflow named auto-pass-pass has been proposed as a tool to fill data gaps in environmental risk assessment using the KNIME platform under the REACH regulation.

III.1.2. Data and Curation

Duplicate starting molecules were identified by canonical SMILES and merged into a single example with all observed values. The open-source KNIME platform

(www.knime.com) [136] was used for checking and plotting data and pattern matching between different resources using CAS code like, the Enalos + CIR node [137] (Figure 09).

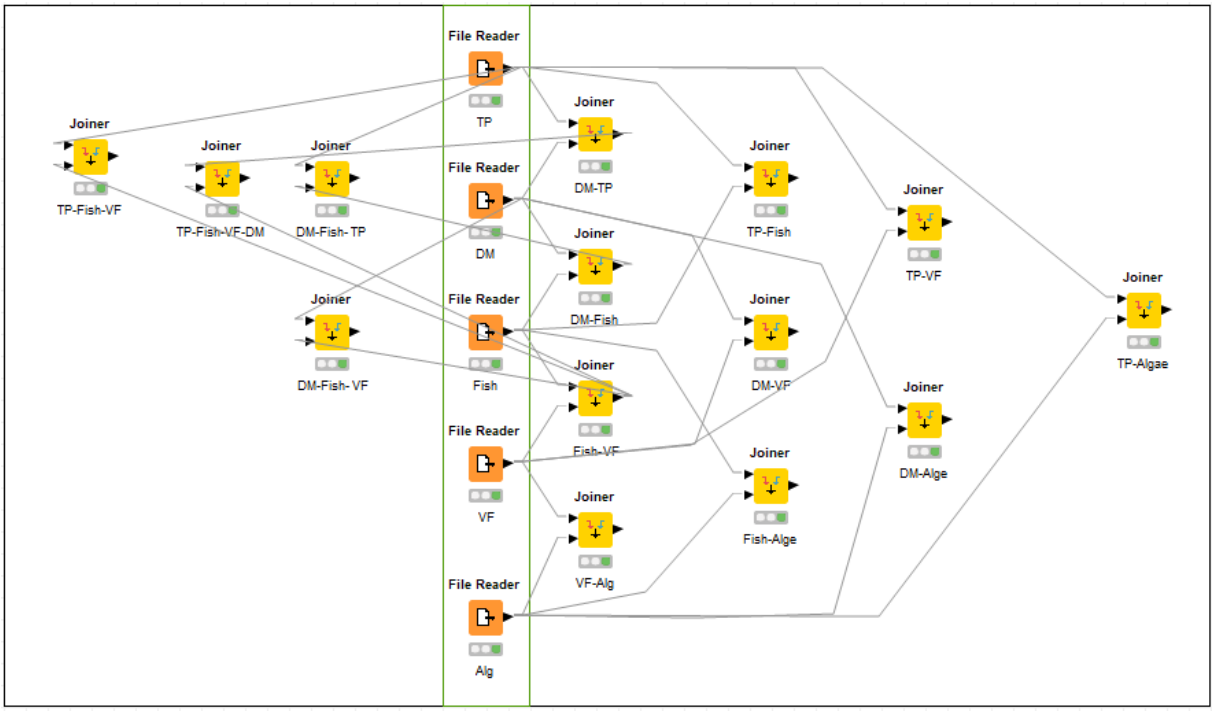


Figure 9.KNIME workflow used for data collection and curation.

We have collected a total of 5881compounds that present good coverage of a wide range of the chemical space. The datasets are cleaned in order to retain only measurements taken under similar experimental conditions as required by the OECD guideline of the given endpoint.

Tetrahymenapyriformis IGC$_{50}$ (TP)

Several previous studies have used *T. pyriformis*to develop QSAR linear models for toxicity evaluation. The ciliate *T. pyriformis* remains an excellent primary source of information, in terms of size, molecular diversity, and quality because it was developed in a single laboratory [138]. Toxicity data [log (1/IGC$_{50}$)] were compiled from the literature [139][138,140–147], in which the dataset containing experimental acute toxicity data of 2038 compounds for *T. pyriformis.* The toxicity is expressed as the concentration that causes 50% growth inhibition (IGC$_{50}$) after very short times 40 h or 48 h.

Fish 50% lethal concentration

The data set contains fish aquatic toxicity of 1212 compounds were taken from the VEGA online platform (http://www.vegahub.eu/) [148]. The lethal concentration for a 50%

kill of the sample was determined statistically as LC_{50} (mol/l, 96 h). The data refer to fresh water, obtained according to OECD 203 on these species, fathead minnow (*Pimephalespromelas*) (most abundant values), guppy (*Poeciliareticulata*), rainbow trout (*Oncorhynchusmykiss*) and medaka (*Oryziaslatipes*), from these sources: ECOTOX (US EPA) [149], the Ministry of Environment in Japan [150], papers of Su *et al.* [151] Gomez-Ganau *et al*. [152].

50% effective concentration (EC_{50}) to Daphnia magna (DM)

The *D. magna* are freshwater organisms that have been used intensively for the last three decades for assessing the effects of chemicals in regulatory testing or for measuring the toxicity of water samples. The toxicity data for 1035 compounds in *D. magna* were taken from the VEGA-hub online platform (http://www.vegahub.eu/) [148] (830 compounds) and from paper of Li *et al*. [139,153] (205 compounds). Note that these data are expressed as EC_{50}, which is the toxicity reported as the 50% effective concentration in 48 h and merged these with the lethal concentration (LC50).

50% bioluminescence inhibiting concentration to Vibrio fischeri (VF)

The dataset for this activity contained 1247 compounds causing 50% inhibition of bioluminescence after 15 or 30 min exposure to *V. fischeri*. All compounds were taken from papers of Li *et al*. [139,153].

72-hours algae growth inhibition (algae) with Pseudokirchneriellasubcapitata

The *Algae* have an important role in the aquatic ecosystem. It is a major food source for higher plankton feeding organisms. The objective of the growth inhibition test on Algae is to determine the effect of a substance or a sample on the growth parameters of freshwater microalgae. The toxicity data for 359 compounds in algae were taken from the VEGA-hub and the toxicity was reported as EC_{50} (50% effective concentration in 72 h).

Assignments Modes of action (MOA)

Various structure-based classification schemes have been developed for classification of chemicals based on the mode of toxic action (MOA) [154]. We have used the OECD QSAR Toolbox (https://www.qsartoolbox.org/home) to group the chemicals based on Verhaar classification. The MOAs of 3661 compounds were classified according to this software into

class 1 (narcosis or baseline toxicity), class 2 (less inert toxicity), class 3 (unspecific reactivity mechanism), class 4 (specific reactivity mechanism) and class 5 (not possible to classify).

III.1.3. Explorative analysis

The inter-species correlation was investigated in order to check the relation between the species. The available pair interspaces correlation with the colored compounds according to the corresponding mode of action are illustrated in Figure10. We notice that nearly 50% of compounds for each species have shared some mode of action which is the unspecific reactivity. Table 04 recapitulates the number of available compounds in each MOA class.

Table 4. The number of compounds distributed in each mode of action class.

Mode of action Class		Number of available compounds				
	TP-DM	TP-Fish	DM-Fish	VF-TP	VF-Fish	VF-DM
Class 1	054 (17%)	095 (18%)	136 (22%)	088 (16%)	080 (23%)	071 (17%)
Class 2	025 (08%)	039 (08%)	028 (05%)	038 (07%)	031 (09%)	029 (07%)
Class 3	160 (52%)	281 (54%)	241 (40%)	314 (57%)	159 (45%)	137 (32%)
Class 4	007 (02%)	013 (03%)	049 (08%)	010 (02%)	018 (05%)	125 (29%)
Class 5	064 (21%)	090 (17%)	154 (25%)	104 (19%)	067 (19%)	067 (16%)

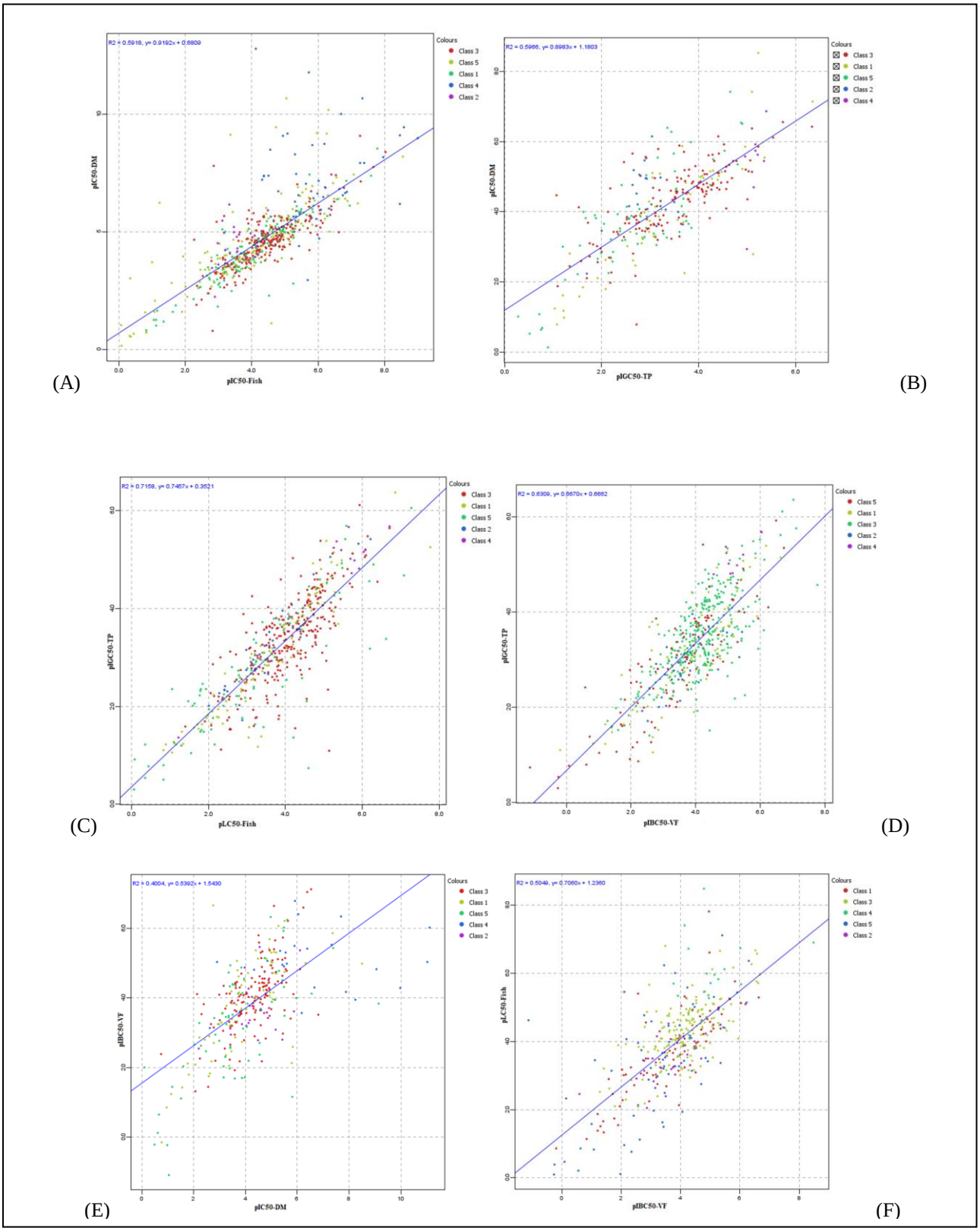


Figure 10. Graphs of the interspecies correlation according to their MOA classes. (A) Plot of DM against Fish toxicity for 608 organic chemicals. (B) Plot of *TP* against Fish toxicity for 518 organic chemicals. (C) Plot of *TP* against DM toxicity for 310 organic chemicals. (D) Plot of *TP* against VF toxicity for 570 organic chemicals. (F) Plot of Fish against VF toxicity for 355 organic chemicals.

Molecular descriptor calculation

A total set of 3839 molecular descriptors was calculated using DRAGON software (version 7.0) to describe each compound chemical diversity using SMILES strings output files from the KNIME node. These descriptors include various quantum chemical, constitutional, topological, geometric and electrostatic descriptors. Constant, near constant, and highly correlated descriptors were treated. The meaning of these molecular descriptors and the calculation procedure are summarized in the DRAGON software and explained in detail, with related literature references, in the Handbook of Molecular Descriptors by Todeschini and Consonni [155].

III.1.4. Results and discussion

III.1.4.1. Inter-species correlation of toxicity

In this part, we explore the inter-species correlation of the four species. Table 5 reports the inter-species correlations between toxicities with the number of compounds.

Table 5.Inter-species correlation measures of toxicity for four species for all common compounds.

Species (A) – Species(B)	Number of common compounds	r	R^2
TP-DM	310	0.77	0.60
fish-DM	608	0.77	0.59
VF-DM	329	0.63	0.40
Alg-DM	299	0.67	0.45
TP-fish	518	0.85	0.72
VF-fish	355	0.71	0.50
Alg-fish	269	0.75	0.56
TP-Alg	144	0.74	0.55
VF-Alg	125	0.56	0.31
TP-VF	552	0.79	0.63

r: Pearson correlation coefficient, R^2: coefficients of determination

Form Table 5, we can notice that for the common compounds, there is a good relationships for toxicity between the pair *TP-DM*, fish-*DM* and *TP*-fish with coefficients of determination values, that are from 0.6 to 0.7. However, there is a weak relation between algae and the other species with, coefficients of determination less than 0.56 although all species toxicity positively contributed to each other. Similar patterns were found for relationships in the

literature [156][3]. With the Pearson correlation coefficient r, some species gave higher mean values (see Table 6).

Table 6. The Pearson correlation coefficient r for the different species.

	TP	Fish	DM	Algae	VF
TP		0.85	0.77	0.74	0.79
Fish	0.85		0,77	0.75	0.71
DM	0.77	0.77		0.67	0.63
Algae	0.74	0.75	0.67		0.56
VF	0.79	0.71	0.63	0.56	
Mean	0.79	0.77	0.71	0.68	0.67

We most notably remark is the high value of TP species (0.79), that may be probably related to the fact that these data are much more homogeneous than the other data collection because they mostly come from a single laboratory. This is a clear indication that the noise in the data makes the statistical analysis evaluation less robust. Also, the fish is the species with the second highest values, that may be related to the fact that the average number of compounds for the correlations with fish is 437, which is the largest number. Thus, the larger number of observations may reduce the influence of noise.

We also examined the regression for the subset of compounds with data for three and four species. Tables 7 and 8 reports the results for these subsets, with data of three and four species, respectively. We notice that the regression values are higher than those reported in Table 5. This could be due to the fact that the substances in these tests are more common than the others and have more experiments. Therefore, this is another factor that could reduce the uncertainty.

Table 7. Correlation of available three-fold activities with some common compounds.

Number of common compounds		**Species (A) – Species(B)**		
		TP-fish	**TP-DM**	**DM-fish**
251	**r**	0.84	0.78	0.81
	R^2	0.69	0.60	0.66

r: Pearson correlation coefficient R^2: coefficients of determination

Table 8. Correlation of available four-fold activities with the common some compounds.

Number of common compounds	Species	TP	DM	fish	TP	DM	Fish
		r			R^2		
174	**DM**	0.77	1		0.59	1	
	Fish	0.85	0.79	1	0.72	0.62	1
	VF	0.84	0.74	0.74	0.71	0.55	0.54

r: Pearson correlation coefficient, R2: coefficients of determination

Furthermore, we have studied the possibility of improvement of the statistical performance by the greater homogeneity of the substances population. For doing this, we have considered the substances with the same MOA. Table 9 recapitulates the number of available compounds in each MOA class, in which Class 3 of Verhaar is the only class characterized by the largest unspecific reactivity. The Coefficients of determination the number of compounds for each MOA class are given in Table 10.

Table 9. Numbers of compounds in each MOA class.

Mode of action Class	Number of compounds					
	TP-DM	**TP-fish**	**DM-fish**	**VF-TP**	**VF-fish**	**VF-DM**
Class 1	054 (17%)	095 (18%)	136 (22%)	088(16%)	080 (23%)	071 (17%)
Class 2	025 (08%)	039 (08%)	028 (05%)	038(07%)	031 (09%)	029 (07%)
Class 3	160 (52%)	281 (54%)	241 (40%)	314(57%)	159 (45%)	137 (32%)
Class 4	007 (02%)	013 (03%)	049 (08%)	010(02%)	018 (05%)	125 (29%)
Class 5	064 (21%)	090 (17%)	154 (25%)	104(19%)	067 (19%)	067 (16%)

As indicated in Tables 9 and 10, there are a good inter-species correlation between the two species TP and fish for all MOAs except class 3; that may be due to the large number of compounds in this class (45%). In addition, Class 1 has the best correlation for all inter-species pairs.

Table 10. Coefficients of determination the number of compounds for each MOA class.

Mode of action Class	R^2 of compounds					
	TP-DM	TP-fish	DM-fish	VF-TP	VF-fish	VF-DM
Class 1	0.73	0.86	0.83	0.70	0.62	0.51
Class 2	0.34	0.88	0.32	0.42	0.34	0.12
Class 3	0.53	0.55	0.60	0.50	0.23	0.38
Class 4	0.22	0.89	0.03	0.60	0.40	0.07
Class 5	0.69	0.70	0.61	0.75	0.02	0.43
Global classes	0.60	0.72	0.60	0.63	0.50	0.40

III.1.4.2. QSAAR and QSAR analyses

i. Results on *T.*pyriformis IGC50 (TP)

In this part, we attempted to build a QSAR model for TP using only theoretical descriptors. Thus, the molecular descriptors were selected by means of the genetic algorithm implemented in QSARINS software and the calibrated models were then validated using the test set. The statistical qualities of the three best models are reported in Tables 11.

Table 11. Best models obtained for IGC_{50} towards *T.periformis*.

Number of compounds	Dependent variable	Descriptors	R^2	Q^2	Q^2_{ext}	ΔK_{xy}	ΔK_x
2038	TP	nDB ATSC5p MATS2e MATS1i GATS1p Hy ALOGP	0.68	0.68	0.66	9 %	37.23%
		nDB PW4 ATSC5p MATS2e GATS3m GATS1v ALOGP	0.67	0.67	0.64	36.74%	32.44%
		nDB MATS2e MATS1i GATS1p P_VSA_m_4 Hy ALOGP	0.67	0.67	0.66	38.65%	34.39%

The performance of the calibrated TP models is generally satisfactory and the performance in fitting, cross-validation and on the external test set is comparable. The relatively high number of descriptors in these models can be due to the fact that such a big dataset may cover a wide range of structurally diverse chemicals. Thus, a high number of descriptors are required to explain the most part of the variance.

Then, we have developed models using the linear regression with the addition of the other species toxicity as independent variables, together with molecular descriptors. Note that Algae were excluded from the Structure-Activity-Activity study, because of the weak inter-species relationships between this species and the other ones. Increasing the number of descriptors in QSAAR models is not very effective to improve their statistical quality, so we have used only three descriptors in the TP model, in which Table 12 shows their statistical parameters.

Table 12. Statistical parameters of T. pyriformis QSAAR, QSAR and inter-species models for the common endpoint.

Dependent variable	Model code	Number of compounds	Descriptors	R^2	Q^2	Q^2_{ext}
$pIGC_{50}$-TP		310	**pIC_{50}-DM SM1_B(m) MLOGP**	**0.80**	**0.79**	**0.73**
			MLOGP	0.63	0.62	0.51
			pIC_{50}-DM	0.56	0.55	0.61
	a	**$pIGC_{50}$-TP** = -1.04 + (0.36) **pIC_{50}**-DM + (0.93) SM1_B (m) + (0.26) MLOGP				
		518	**pLC_{50}-fish SM1_B(m) MLOGP**	**0.84**	**0.84**	**0.78**
			pLC_{50}-Fish	0.74	0.74	0.67
			ALOGP	0.60	0.59	0.68
	b	**$pIGC_{50}$-TP** = -1.13 + (0.46) **pLC_{50}**-fish + (0.91) SM1_B (m) + (0.17) MLOGP				
		552	**$pIBC_{50}$ –VF SM1_Dz(e) ALOGP**	**0.76**	**0.75**	**0.75**
			$pIBC_{50}$ –VF ALOGP	0.71	0.70	0.69
			$pIBC_{50}$ -VF	0.63	0.63	0.68
	c	**$pIGC_{50}$-TP** = 0.58 + (0.42) **$pIBC_{50}$** -VF + (1.06) SM1_Dz(e) + (0.33) ALOGP				

The results in Table 12 show that the inclusion of molecular descriptors in the inter-species relationships can significantly improve the inter-species relationships.

All TP QSAAR models gave a good statistical performance in both internal and external validation (see Table 12). We can improve the prediction of the substance of interest using experimental data from a second species and chemical descriptors. This method gives values higher than that in the case of models with only molecular descriptors. This integrated

strategy uses both the experimental values from another test related to the endpoint of interest and the molecular descriptors.

The TP QSAR models with chemical descriptors have moderate statistical quality, with r^2 values of about 0–66-0.68 in all sets of compounds (Table 12). Conversely, the statistical quality becomes good when the activity-activity relationship is also added, with values ranging from 0.73 to 0.84 (see Table 12).

For the TP species in models (a) and (e), both descriptors (SM1_B(m)), which is the spectral moment of order 1 from the Burden matrix weighted by mass and (MLOGP), involved in the equations, gave the best models. However, the results from the equation with fish values (b) are quite higher than with the model using DM values (a), considering the r^2 (0.84 and 0.80, respectively).

ii. Results on fish ($LC_{50\%}$) lethal

For the first step, the QSAR models of *fish* have been built using only theoretical descriptors, in which, the molecular descriptors were selected employing a genetic algorithm implemented in QSAINS software and the calibrated models were then validated using the test set. The statistical qualities of the three best models are reported in Table 13.

Table 13. The statistical qualities of the three best QSAR models obtained for *fish*.

Number of compounds	Dependent variable	Descriptors	R^2	Q^2	Q^2_{ext}	ΔK_{xy}	ΔK_x
1212	Fish	ARR SM1_B(p) EE_B(i) SM3_B(s) ATS2v GATS1i MLOGP	0.59	0.58	0.59	55.44	54.87
		ARR SM1_B(p) SM3_B(i) SM4_B(s) ATS2v GATS1i MLOGP	0.59	0.58	0.59	55.03	54.42
		ARR SM1_B(p) EE_B(i) SM3_B(s) ATS2v GATS1i BLTA96	0.59	0.58	0.59	55.44	54.87

To investigate the differences between toxicities in different aquatic organisms and to select a significant descriptor to improve the inter-species relationships, a genetic algorithm was performed among the toxicity and the calculated descriptors.

We have added some descriptors to the inter-species models in order to develop QSAARs models. The Statistical parameters of fish QSAAR and QSAR together with the inter-species models are reported in Table 14. We can see in Table 14 that the improvement of the QSAAR models of the *fish* has a good statistical performance in internal and external validation.

Table 14. Statistical parameters of fish QSAAR and QSAR together with the inter-species models.

Dependent variable	Model code	Number of compounds	Descriptors	R^2	Q^2	Q^2_{ext}
pLC_{50}-fish		518	$pIGC_{50}$-TP SM14_AEA(ri) SAdon	**0.77**	**0.76**	**0.61**
			$pIGC_{50}$-TP	0.75	0.75	0.60
			ALOGP	0.51	0.51	0.59
	g	**pLC_{50}-fish = $pIGC_{50}$-TP**				
		608	pIC_{50}-DMMATS1s ALOGP	**0.68**	**0.67**	**0.74**
			pIC_{50}-DM ALOGP	0.67	0.66	0.74
			pIC_{50}-DM	0.58	0.57	0.65
	h	**pLC_{50}-fish** = 1.39 + (0.50) **pIC_{50}-DM** + 0.29				
		355	$pIBC_{50}$ –VF IDETSp PosA_B(p)	**0.66**	**0.65**	**0.61**
			$pIBC_{50}$ -VF	0.50	0.49	0.60
	i	**pLC_{50}-fish** = -3.98 + (0.46) **$pIBC_{50}$ -VF** + (4.62) **SpPosA_B(p)** + (0.01) IDET				

The *D. magna* and *V. fischeri* toxicities were less correlated with *fish* toxicity (R^2 = 0.68 and 0.66, respectively), compared to *T. pyriformis*. In addition, the best model of *fish* toxicity prediction is that by using the QSAAR model of equation (g).

iii. Results on *D. magna* (*DM*)

The best three results obtained with the different regression methods for *D.magna* using only theoretical descriptors are collated in Table 15. The descriptors were selected through a genetic algorithm implemented in QSAINS software and the calibrated models were then validated using the test set.

Table 15. Best QSAR models obtained for *D. magnas.*

Number of compounds	Dependent variable	Descriptors	R^2	Q^2	Q^2_{ext}	ΔK_x	ΔK_{xy}
1035	DM	ATSC1p P_VSA_v_3 P_VSA_s_2 O-058 SsOH CATS2D_08_LL MLOGP2	0.60	0.59	0.49	40.33%	41.81%
		ATSC1p P_VSA_v_3 P_VSA_s_2 O-058 CATS2 D_00_DA CATS2D_08_LL MLOGP2	0.60	0.58	0.49	40.53%	41.99%
		ATSC8m ATSC1p P_VSA_v_3 P_VSA_i_3 CATS2D_08_LL MLOGP2 LLS_01	0.59	58.51	0.49	47.04%	47.53%

We have added some descriptors to the inter-species models in order to develop QSAARs models. The statistical parameters of *D. magna* QSAAR and QSAR together with the inter-species models are collected in Table 16. From Table 16, we can notice that the improvement of the QSAAR on the *fish* models leads to a good statistical performance in both internal and external validations.

Table 16. Statistical parameters of *D. magna* QSAAR and QSAR together with the inter-species models.

Dependent variable	Model code	Number of compounds	Descriptors	R^2	Q^2	Q^2_{ext}
pIC_{50}-DM		608	**pLC_{50}-fishX3vMATS1s**	**0.70**	**0.69**	**0.70**
			pLC_{50}-fishX3v	0.67	0.66	0.50
			pLC_{50}-fish	0.61	0.61	0.37
			ALOGP	0.32	0.31	0.26
	d	**pIC_{50}-DM** = 0.63 + (0.73) **pLC_{50}**-fish + (0.28) X3v + (-1.68) MATS1s				
		310	**$pIGC_{50}$-TP X0A MATS1s**	**0.67**	**0.66**	**0.63**
			$pIGC_{50}$-TP	0.61	0.60	0.54
	e	**pIC_{50}-DM** = 4.56 + (0.84) **$pIGC_{50}$-TP** + (-5.00) X0A + (-1.82) MATS1s				
		329	**$pIBC_{50}$ -VFSpPosA_B(p) ATSC8e**	**0.61**	**0.60**	**0.34**
	f	**pIC_{50}-DM** = -3.19 + (0.58) **$pIBC_{50}$** -VF + (4.00) SpPosA_B(p) + (5.61) ATSC8e				

The *fish* and *T. pyriformis* toxicities were correlated with *D, magna* toxicity (R^2 = 0.70 and 0.67, respectively), however, *V. fischeri*was less correlated with R^2 equal to 0.66. For the toxicity against *D.magna* the equation (d) is the best choice for predicting the missing values.

iv. Results on V. fischeri (VF)

The best three results obtained with the different regression methods for QSAAR model against *VF* using only theoretical descriptors are collated in Table 17. The descriptors were selected through a genetic algorithm implemented in QSAINS software and the calibrated models were then validated using the test set.

Table 17. The three best QSAAR models obtained for *V. fischeri*.

Number of compounds	Dependent variable	Descriptors	R^2	Q^2	Q^2_{ext}	ΔK_{xy}	ΔK_x
1247	VF	MpnDB X2A ATSC4p JGI3 Eig09_AEA(ri) MLOGP	0.56	0.50	0.46	30.92	34.16
		nDB X2A JGI3 Eta_alpha_A Eig09_EA VvdwZAZ BLTF96	0.51	0.50	0.44	26.49	30.6
		nDB X2A JGI3 Eta_alpha_A SM03_AEA(dm) VvdwZAZ BLTF96	0.51	0.50	0.44	26.49	30.6

As in the previous cases, we have added some descriptors to the inter-species models to develop QSAARs models. The obtained Statistical parameters of *V. fischeri* QSAAR and QSAR together with the inter-species models are given in Table 18. We can notice from Table 18 that the improvement of the QSAAR of the *V.fischeri* models gave a good statistical performance in both internal and external validations.

Table 18. Statistical parameters of *V. fischeri* QSAAR and QSAR together with the inter-species models.

$pIBC_{50}$-VF		570	pIGC50-TP Mp SpMax4_Bh(p)	**0.73**	**0.73**	**0.75**
			$pIGC_{50}$-TPATS2p	0.70	0.70	0.69
			$pIGC_{50}$-TP	0.63	0.63	0.60
	j	**$pIBC_{50}$-VF** = -1.75 + **pIGC50**TP (0.66) + (2.27) MP + (0.85) SpMax4_Bh(p)				

The toxicity against *TP* is in a good correlation with *V. fischeri* toxicity (R^2 = 0.73), and thereby equation (j) can be used for predicting the missing values of *V. fischeri* toxicity.

III.1.4.3. Comparison with literature models

Many studies have focused on the relationship between the toxicities of common organic compounds and two or three species. The study by Cronin *et al.* (1991) [11] is one of the earliest works that highlighted the possibility of using inter-species relationships in aquatic toxicity. Since then many studies on QAAR have reported a good relationship between different aquatic species like *T.pyriformis, D.magna, V.fischeri* and *fish*. Table 19 summarize some relevant papers on inter-species relationships.

Kahn *et al*. (2007) [27] attempted to use inter-species data as an independents variable together with molecular descriptors in QSAARs model for improving the QAAR. From the list of existing work, the number of compounds used in all QSAAR models was limited compared to our study. This may be because of a lack of data at that time, or the limitations of MOA. The only work with a large number of compounds was by in Furuhama *et al.* (2016) [157], who used the chronic *D. magna* for predicting acute *D. magna*.

Li *et al*. (2015, 2018) [139,153] used a large pool of data on the acute toxicity of organic pollutants to investigate the species-specific thresholds, but not for QSAARs modeling. In our study, drawing on the existing data, especially in VEGAHUB and that of Li *et al.* (2015, 2018), we were able to modeled several QSAR, QAAR, and QSAAR. The best models were implemented in one global Auto *Pass-Pass* workflow in order to predict the toxicity of the new compounds against the four aquatic species and to fill some of the data gaps for chemicals with no experimental data.

Table 19.Quantitative Activity-Activity Relationships (QAARs) and Quantitative Structure Activity-Activity Relationships (QSAARs) models in the literature.

Species	**No.**	**QAAR**	**QSAAR**		
dependent variable-independent variable)			**Descriptors**	**R^2**	**Reference**
fish-*D.magna* fish-*T.pyriformis*	46 74	R^2 = 0.75 R^2 = 0.96			Cronin *et al.* (1991) [11]
fish *(p.romelas)-T.pyriformis* fish*(p.romelas)*-algae *(C.vulgraris)* fish*(p.romelas)-V-fischeri*	91	r = 0.93 r = 0.76 r = 0.89			Cronin *et al.* (2004) [158]
fish (Oryziaslatipes)-DM algae (Pseudokirchneriellasubcapitata)-DM	366 339	R^2 = 066 R^2 = 0.54			Furuhama*et al.* (2015) [159]

fish *(P.promelas)-T.pyriformis* fish *(P.promelas)-D.magna* fish *(P.promelas)- V.fischeri* fish *(P.promelas)*- algae *(Scenedesmusobliquue)*	38 53 56 19	r = 0.75 r = 0.87 r = 0.91 r = 0.72			Zhang *et al.* (2010) [3]
fish*(Onchorynchusmykiss)-DM*	40	R^2 = 87			Cassani*et al.* (2013) [160]
algae *(C.vulgraris)-T.pyriformis*	31	R^2 = 0.75			Tugcu*et al.* (2017) [7]
<u>fish-fish</u> *Guppy-Rainbow trout* *Guppy- Fathead minnow* *Guppy- Medaka* *Fathead minnow- Rainbow trout* *Medaka- Fathead minnow*	 40 133 43 39 58	 R^2 = 0.97 R^2 = 0.99 R^2 = 188 R^2 = 0.96 R^2 = 0.92			Su *et al.* (2014) [161]
fish-*T.pyriformis*	364	R^2 = 0.75	P^C_{avg}, $\#N_{rel}$, HACA2	0.82	Kahn *et al.* (2007) [12]
DM-fish*(P.promelas)* DM- fish *(O.mykiss)* fish-fish *(P.promelas-O.mykiss)*	44 51 36	R^2 = 0.80 R^2 = 0.83 R^2 = 0.85	ATS4s GATS1e AATS7p	0.88 0.88 0.95	Sangion and Gramatica (2016) [15]
fish-*DM*	55	R^2 = 0.47	log S0, smom3	0.69	Furuhama*et al.* (2019) [14]
Chronic*D.magna*-Acute *D.magna*	299	R^2 = 0.54	log D_{ph8}	0.81	Furuhama*et al.* (2016) [157]
fish *(Pimephalespromelas)-D.magna*	77	R^2_{adj} = 0.54	MW	0.61*	Furuhama*et al.* (2018) [16]
DM- fish	77	R^2=0.62	S_dO	0.70	Kar and Roy (2010) [162]
fish-*DM* fish-*T.pyriformis* fish-*V.fischeri* DM-*T.pyriformis* DM-*V.fischeri* TP-*V.fischeri*	50 10 26 5 18 5	R^2 = 0.38 R^2 = 0.72 R^2 = 0.18 R^2 = 0.82 R^2 = 0.26 R^2 = 0.94			Ling *et al.* (2019) [163]
DM-fish *T. pyriformis*-fish *T. pyriformis-D.magna* *V. fischeri*-fish *V. fischeri-DM*	467 288 182 304 294	r= 0.84 r= 0.85 r= 0.77 r= 0.73 r= 0.67			Li *et al.* (2015) [139]
fish-*D.magna* fish-*T.pyriformis*	467 478	R^2 = 0.72 R^2 = 0.72			Li *et al.* (2018) [153]

fish-*V.fischeri*	304	$R^2 = 0.54$			
DM-*T.pyriformis*	287	$R^2 = 0.63$			
DM-*V.fischeri*	294	$R^2 = 0.45$			
TP-*V.fischeri*	556	$R^2 = 0.63$			

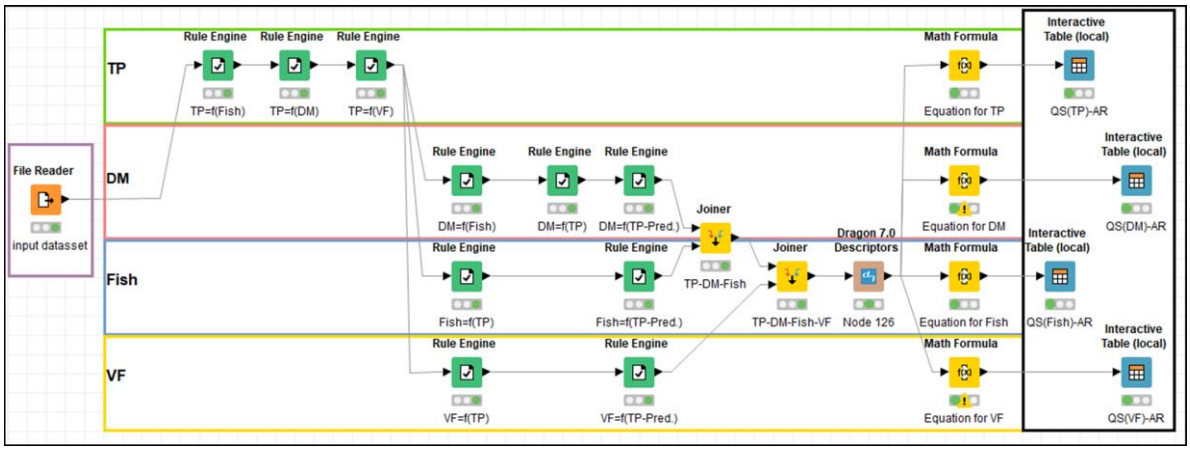
*r: Pearson correlation coefficient, R^2: coefficients of determination, *R^2adjusted, No.: Number of compounds.

III.1.4.4. Auto-pass-pass, a new approach to fill data gaps in environmental risk assessment

The QSAAR models for aquatic toxicity have good statistical parameters, internally robust and stable, with higher quality than the inter-species -based and descriptor-based models. The algorithms in Tables 20 and 21 were used to predict 6637 missing toxicity data, as follows:

- Fish-based in the model (b) was the best for predicting toxicity data against *TP* species.
- If the fish data are not available, then *DM*-based in model the (a) can be used for predicting the *TP* missing values.
- The *VF* species is the third model, which can be used for predicting the *TP* species values with the equation of model (c).
- Fish-based model (d) was the best model for predicting toxicity against *DM* species.
- If the DM toxicity data is not available, TP-based model (e) is the best alternative for calculating fish data.
- In case of missing toxicity data against fish, model (g) which is based on *TP* species is the best for doing this, if this not possible, *MD*-based from the model (h) can be used to predict missing values.
- For the *VF* species, the *TP* model (j) is the best for predicting missing values.

For this *Pass-Pass* proposed new idea, we designed a KNIME workflow of inter-species QAAR and QSAAR, to automate the predictions of all activities, as indicated in Figure 11.



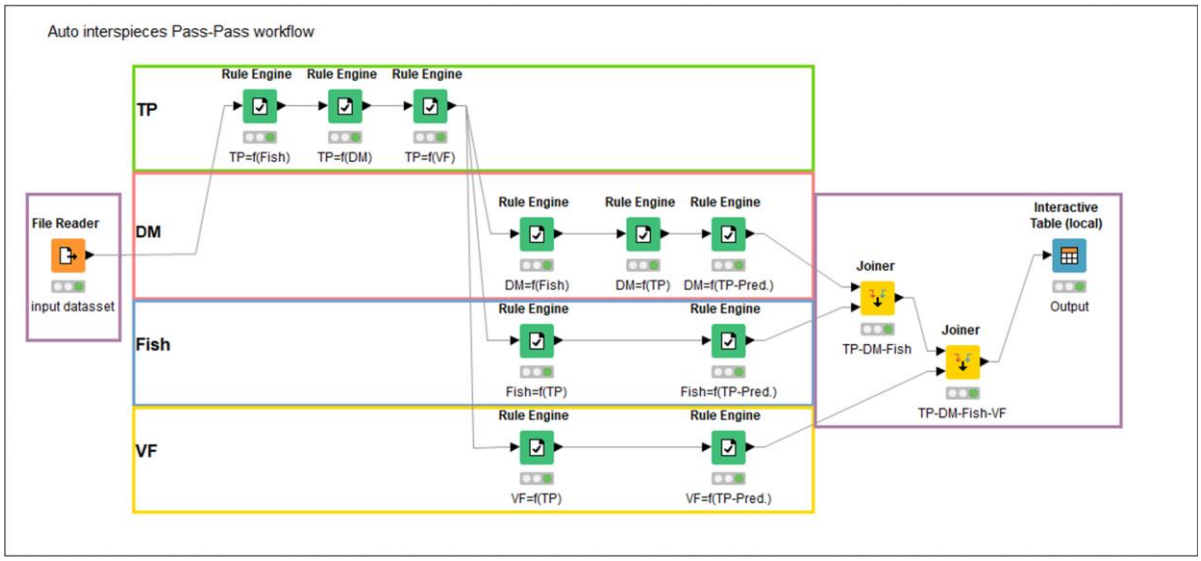


Figure 11. KNIME workflow of Auto interspecies QAAR Pass-Pass and QSAAR Pass-Pass for predicting all activities.

Table 20.Auto inter-species Pass-Pass algorithms in rule engine nodes for predicting of all activities.

fish-prediction	**VF-prediction**
MISSING pLC_{50}$-fish$ =>$ pIGC$_{50}$-TP $ **TRUE** =>pLC_{50}$-fish$ **--- Append Column:** fish-prediction **MISSING**$fish-pred$ =>$TP-pred$ **TRUE** =>$fish-pred$ **--- Replace Column:** fish-pred	**MISSING** $pIBC$_{50}$-VF$=>$ pIGC$_{50}$-TP $ **TRUE=>**$ pIBC$_{50}$-VF$ **--- Append Column:**VF-prediction **MISSING** $VF-pred$ =>$TP-*pred*$ **TRUE** =>$VF-pred$ **--- Replace Column:** VF-pred
TP-prediction	**DM-prediction**
MISSING $pIGC$_{50}$-TP$ => pLC_{50}$-fish$ **TRUE=>**$pIGC50-TP$ **--- Append Column:** TP-pred **MISSING**$TP-pred$ =>pIC_{50}$-DM$ **TRUE=>**$TP-pred$ **--- Replace Column:** TP-pred **MISSING** $TP-prediction$ =>pVF **TRUE** =>$TP-pred$ **--- Replace Column:** TP-pred	**MISSING**pIC_{50}$-DM$=>pLC_{50}$-fish$ **TRUE=>**$pIC50-DM$ **--- Append Column:** DM-pred **MISSING** $DM-prediction$ => $ pIGC50-TP$ **TRUE** =>$DM-pred$ **--- Replace Column:** DM-pred **MISSING** $DM-prediction$ =>$TP-pred$ **TRUE** =>$DM-pred$ **--- Replace Column:** DM-pred

Table 21.Auto QSAAR Pass-Pass algorithms used for prediction of all activities.

fish-prediction	**VF-prediction**
MISSING $pFish$=>$ pIGC$_{50}$-TP $ **TRUE** =>pLC_{50}$-Fish$ **--- Append Column:** fish-pred **MISSING**$pFish-pred$ =>pTP-pred $ **TRUE** =>$pFish-pred$ **--- Replace Column:** fish-pred	**MISSING** $pIBC$_{50}$-VF$=>$-1.75 + 0.66 * pTP + 2.27*MP + 0.85*$SpMax4_Bh(p)$ **TRUE=>**$ pVF$ **--- Append Column:**VF-pred **MISSING**$VF-pred$=>$TP-pred$ **TRUE** =>$VF-pred$ **--- Replace Column:** VF-pred
TP-prediction	**DM-prediction**
MISSINGpTP=>$-1.13 + 0.46 * $pfish$ + 0.91* $SM1_B (m)$ + 0.17* $MLOGP$ **TRUE=>**pTP	**MISSING** pDM=> 0.63 + 0.73* $pfish$ + 0.28* $X3v$ + -1.68* $MATS1s$ **TRUE=>**pDM

--- Append Column: TP-pred **MISSING**$pTP-pred$=>-1.04 + (0.36) *pDM + 0.93*$SM1_B (m)$ + 0.26 *$MLOGP$ **TRUE=>**$pTP-pred$ **--- Replace Column:** pTP-pred **MISSING**$TP-pred$=> 0.58 + 0.42 * pVF + 1.06 * $SM1_Dz(e)$ + 0.33 * $ALOGP$ **TRUE =>**$TP-pred$ **--- Replace Column:** TP-pred	**--- Append Column:** DM-pred **MISSING**$DM-pred$ =>4.56 + 0.84 pTP + -5.00* $X0A$ -1.82*$MATS1s$ **TRUE =>**$DM-pred$ **--- Replace Column:** pDM-pred **MISSING** $DM-pred$ =>$TP-pred$ **TRUE =>**$DM-pred$ **--- Replace Column:** DM-pred

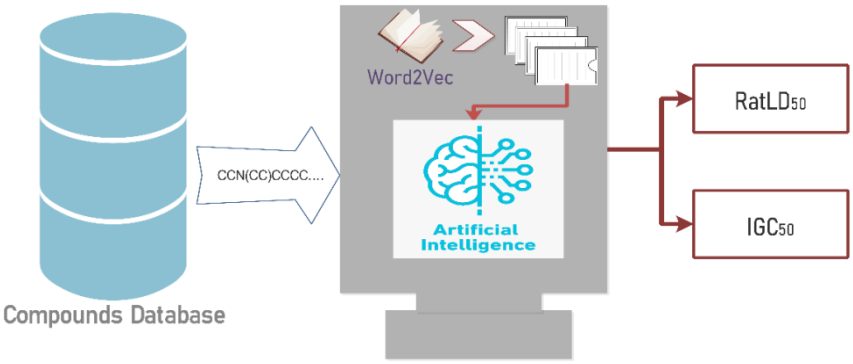
The *Auto-Pass-Pass* KNIME workflow is available at the VEGAHUB web site [164].

III.1.5. Conclusion

Our aim was to propose the models based on Quantitative Structure-Activity-Activity Relationships (QSAAR), which allows the extrapolation of toxicity between *TP*, fish, *VF* and *DM*. The current study provides a fills some data gaps for 1603 compounds against *TP* species, 2605 compounds for *DM* species, 2429 for fish and 2396 for VF species.

These models have been developed and validated based on OECD principles. There is observed a poor inter-species relationship between algae and the other species possibly due to the different toxic mechanisms of action between the aquatic organisms. The proposed aquatic Auto *Pass-Pass* models can also be useful for predicting large amounts of chemicals without experimental values, employing the experimental values from a second species.

III.2. Application of Machine Learning and Deep Learning in QSAR modeling



III.2.1. Introduction

Quantitative structure activity relationship (QSAR) approach is one of the most commonly used methods for the prediction of biological properties in order to facilitate the drug discovery process. It is an adequate alternative way for expensive and time-consuming ecotoxicological experiments.

Over the past few decades, several statistical model algorithms such as multiple linear regression (MLR) and partial least (PLS) have been widely used in QSAR modeling. Commonly, the number of selected descriptors should be low in comparison to the number of data points to avoid over fitting. For this reason, numerous feature selection methods have been used to reduce the number of molecular descriptors [165–167]. There are many feature selection methods, in which the most used ones are stepwise regression and genetic algorithms[168–170]. However, with the amount increasing of biological data, these statistical model algorithms can helped only by feature selection, considering that they were not designed for such problems.

During the last years, more sophisticated machine learning algorithms such as support vector machine (SVM) and random forest (RF) have been proposed to deal with this problem. However, these methods also need pre-processing filter feature selection techniques for selecting subsets of molecular descriptors [171]. Since the amount of biological experimental data (The number of endpoints) is increasing ever, a huge number of descriptor subsets combinations have been explored by feature selection methods which require high computational effort.

Deep learning algorithms become the most exciting research area in machine learning. It is an automatic general-purpose learning procedure, which has been commonly adopted in many fields [121].

In this part, we propose a novel integrative DL-QSAR, RF-QSAR and SVM-QSAR approaches based on embedding deep learning for predicting high-quality toxicity models through natural language processing of SMILES notation. Our approach comprises two steps; the first is the data preprocessing phase, where, we build a corpus of compound substructures with n-length to transform these compounds in word embedding representation. The second is the training model, in which we integrated multiple machine learning algorithms namely,

random forest, support vector machine and convolutional long short-term memory network in order to predict the proprieties or activities of several datasets.

This work is divided into three major sections. Firstly, we describe the proposed approach with the aforementioned datasets. Secondly, we discuss and compare the results obtained by the use of machine learning algorithms based on benchmark datasets. Finally, we extract the main conclusions of this work and draw the future work.

Data sets

To evaluate the performance of our approach two toxicity data sets were used in this work which are:

- **Rat LD_{50} data:** acute oral toxicity, which is expressed as median lethal dose LD_{50} that is one of the most important toxicological endpoints to be assessed in drug discovery. LD_{50} is the dose of a chemical that kills half of the treated rats, in which the values were expressed as (pLD_{50}mol/kg). The data set used in this study containing 7314 compounds reported by several works [172–174].
- **Tetrahymena*pyriformis*IGC_{50} database:** this endpoint is one of the most commonly used data set in QSAR modeling, for the evaluation of the compounds aqueous toxicity. According to (Svetnik *et al*., 2003), the IGC_{50} is the largest amount of aqueous toxicity information, which is tested in a single laboratory by a single reliable and robust method (Cronin *et al*. 2002). The toxicity is expressed as the 50% growth inhibitory concentration ($pIGC_{50}$mol/L) of the T. pyriformis organism after 40 hours. The data was obtained from QSAR Toolbox software (http://oasis-lmc.org) and the Wu and Wei work [175].

III.2.2. The proposed approach

Natural language processing is a computational algorithm for the automatic analysis and representation of human language. NLP-based systems have been successfully applied in many fields of applications such as Google's powerful search engine and Amazon's voice assistant (Alexa). The recent development of deep learning architectures has given NLP algorithms a renaissance. In chemo-informatics, the SMILES format is a single-line text-based molecular notation format, in which it is wildly used for chemical descriptor calculation, because it is particularly suitable for high-speed machine processing. In chemo-informatics, SMILES-based is similar to the natural language since it was introduced as a single-line text representation of a unique molecule in the form of strings over a fixed alphabet (Bjerrum, 2017).

In this work, a novel approach for QSAR modeling is proposed, in which we adopt the word embedding approach to predict the activity or the propriety of compounds by using:

a) Data Processing

The data processing step consists of three main steps. The first one is the build of a corpus containing all substructures possibilities of molecules compounds. In the second one, using the generated corpus, we train a Word2vec model on this corpus. Finally, in the last step, we integrate the generated word2vec model to predict the numerical representation of smiles compounds. Therefore, the neural network modules are trained on an input of SMILES for extracting the maximum important features (encoder module). Then, the extracted features are encoded into a vector format which can be operated as a continuous representation. A graphical representation of all those steps is depicted in Figure 12.

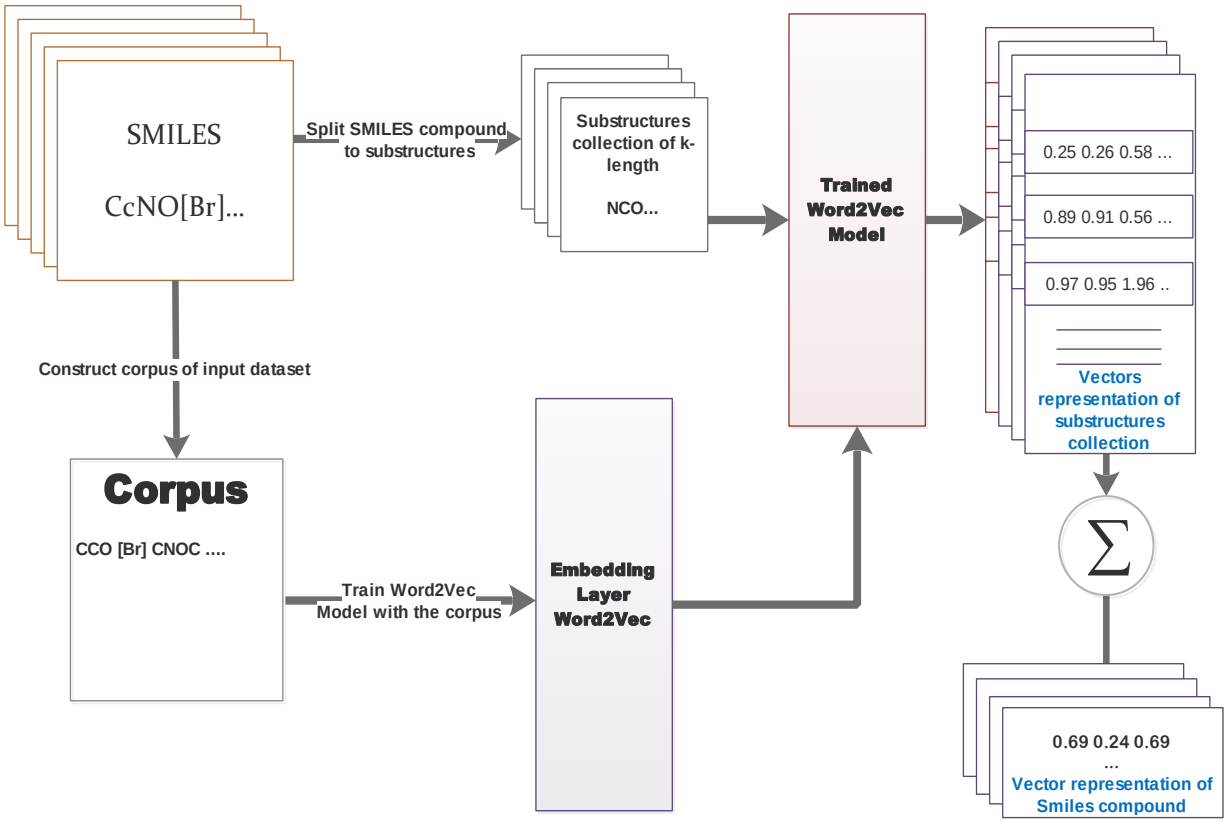


Figure 12. The proposed model for building the vectors representation of SMILES fragments starting from building the corpus to the SMILES fragments embeddings representation.

The proposed method for SMILES molecule data processing is performed as the following steps:

1. **Build the corpus of dataset**

In this step, we try to construct the corpus that contains all substructures possibilities of molecules for a given SMILES dataset. The corpus is defined as follows:

$$Corpus(SMILES_{dataset}) = C_1, C_2, C_3, \dots, C_n \qquad \text{Equation 11}$$

Where, C and n are the substructures and the number of atoms concatenation possibilities, respectively. The high complexity dataset which contains around 700000 compounds has been taken from EPA (Environmental Protection Agency) CompTox Dashboard (https://comptox.epa.gov/dashboard/), which was used to build the corpus. The diversity compounds allow the building of a very large corpus which includes all words combinations from the three datasets. This generated corpus will assess the word2vec model to predict the

optimum vector representation of chemical compounds. Indeed, the generated model will distinct between the embedding representations of molecules with a high precision.

2. Train the embedding model

We have trained a *word2vec* model to transform the molecules SMILES to vectors representation. This word embedding algorithm is one of the best algorithms for natural language processing so far. The *word2vec* model is a two-layer neural networks that aims to deal with words for predicting the vector representations. In this subject, words that share similar contexts are represented by close numerical vectors. We have used the generated corpus from the previous step to train the embedding model in order to obtain a trained transformer noted *Vec(x)*, where, *X* is the generated corpus. This transformer maps a word *x* of Corpus into a continuous vector space of size d.

$$Vec : Corpus \rightarrow \mathbb{R}^d \qquad \text{Equation 12}$$

$$X \rightarrow Vec\,(x)$$

3. Transform SMILES into numerical vector using trained embedding model

In this step, we predict the numerical representation vector of SMILES using the trained proposed embedding model. Thus, in order to be able to predict the vectorization of the SMILES molecule, the latter is split into substructures of compounds. Then, we use equation (1) to calculate the numerical representation vector of the whole SMILES according to the following equation:

$$Vec(SMILE_{molecule}) = \sum Vec\,(X1);\ Vec\,(X2);\ \ldots;\ Vec(Xi)) \qquad \text{Equation 13}$$

Where, SMILES molecule = concatenation (X1, X2....Xi) and i is the number of concatenations of length L in the SMILES molecule. The whole data preprocessing is depicted in Figure 12.

b) Training the Convolutional LSTM Model

The collected data over successive periods of time such as molecule compounds are characterized as a time series. In this case, LSTM is an interesting approach to deal with this type of data. In this kind of deep learning architecture, the model passes from the previous hidden state to the next step of the sequence, in which, the data order is extremely important.

On the other side, convolutional neural networks are the best deep learning models to extract the feature pattern from data which are represented as matrices such as images. In this subject, the convolutional layer aims to extract the features map from the vector representation of compounds.

To predict the toxicity or the propriety of compounds, a convolutional long short-term memory model (ConvLSTM) was used. This last is an extension of the popular long short-term memory (LSTM) RNN. In this model, the fully-connected nodes of the LSTM module are replaced by convolutional gates and thereby making it capable of encoding spatio-temporal features of the SMILES vector representation. The proposed architecture of the model is illustrated in Figure 13.

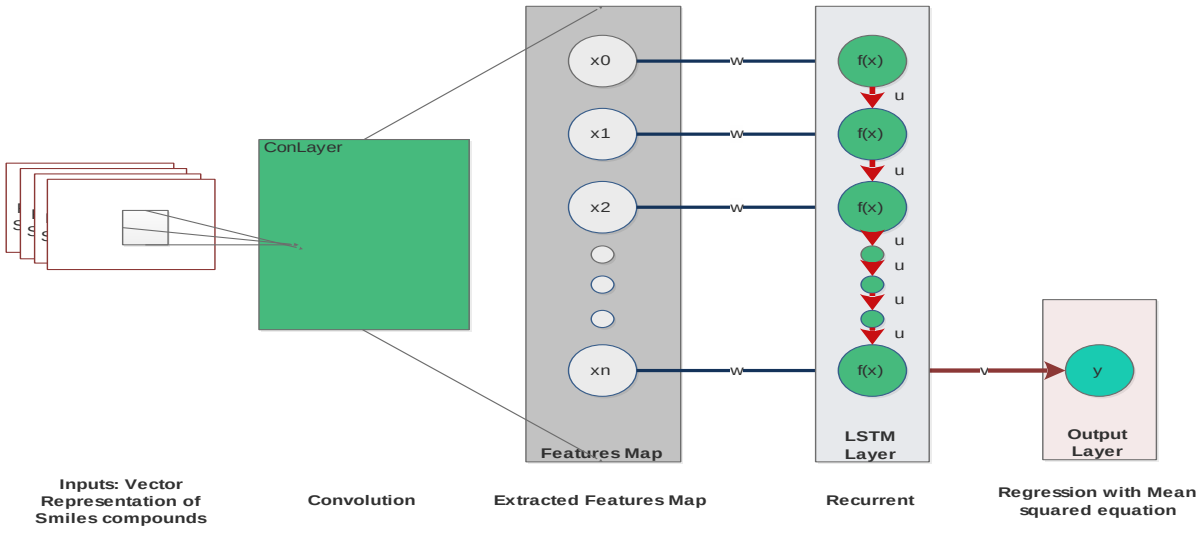


Figure 13. The proposed deep learning model to predict the activity of molecules.

Convolution steps

The extraction of the features pattern from the predicted numerical vectors representation of SMILES fragments is performed using convolutional layers, which contains a set of filters whose parameters need to be learned from the input vectors to obtain a features map.

Rectification

Rectified linear units (ReLu) function was used in the input vectors due to its unsaturation and the high gradient if the nodes of layers are activated. The ReLu function is defined as follows:

$$ReLu(x) = max\ (0;\ x) \qquad \text{Equation 14}$$

LSTM

LSTM layer has a chain structure, with a different structure of repeating module. Instead of having a single neural network layer, there are many interactions in a very special way. After extracting the features map from the vector representation of compounds using a convolutional layer, we have integrated a LSTM layer to deal with this features map as a chain. After this step, LSTM can add information or remove any useless information to predict the activity of these compounds.

Dropout

Dropout layers randomly zeroes the inputs to the next vertex layer during the training with a determining probability of 0.5. This regularizes the network and prevents the over-fitting.

Loss Function

We have used the *MSE* function to measure the divergence between the probabilities distributions corresponding to the assignment of SMILES fragment to a toxicity value. The loss function is given as the following equation:

$$MSE = \frac{1}{N}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \qquad \text{Equation 15}$$

Where, y_i is the experimentally observed response and $\hat{y}_i$ is the theoretically predicted one, i until n is the number of output nodes.

Learning Optimization Policy

To improve the learning of the proposed model, we integrated the ADAM (adaptive moment estimation) optimization algorithm for its computational performances [176]. It aims mainly to adjust the learning settings during the training of the neural network (Figure 14). All processes of the training model are given in Table 22.

Table 22. Configuration details of the proposed deep learning model.

Layer (type)	**Output Shape**	**Param #**
conv_lst_m2d_2 (ConvLSTM2D)	(None, 31, 7, 128)	264704
max_pooling2d_2 (MaxPooling2)	(None, 15, 3, 128)	0
dropout_2 (Dropout)	(None, 15, 3, 128)	0
flatten_2 (Flatten)	(None, 5760)	0
dense_1 (Dense)	(None, 16)	92176
dropout_3 (Dropout)	(None, 16)	0
dense_2 (Dense)	(None, 1)	17
Total params: 356,897 Trainableparams: 356,897 Non-trainableparams: 0		

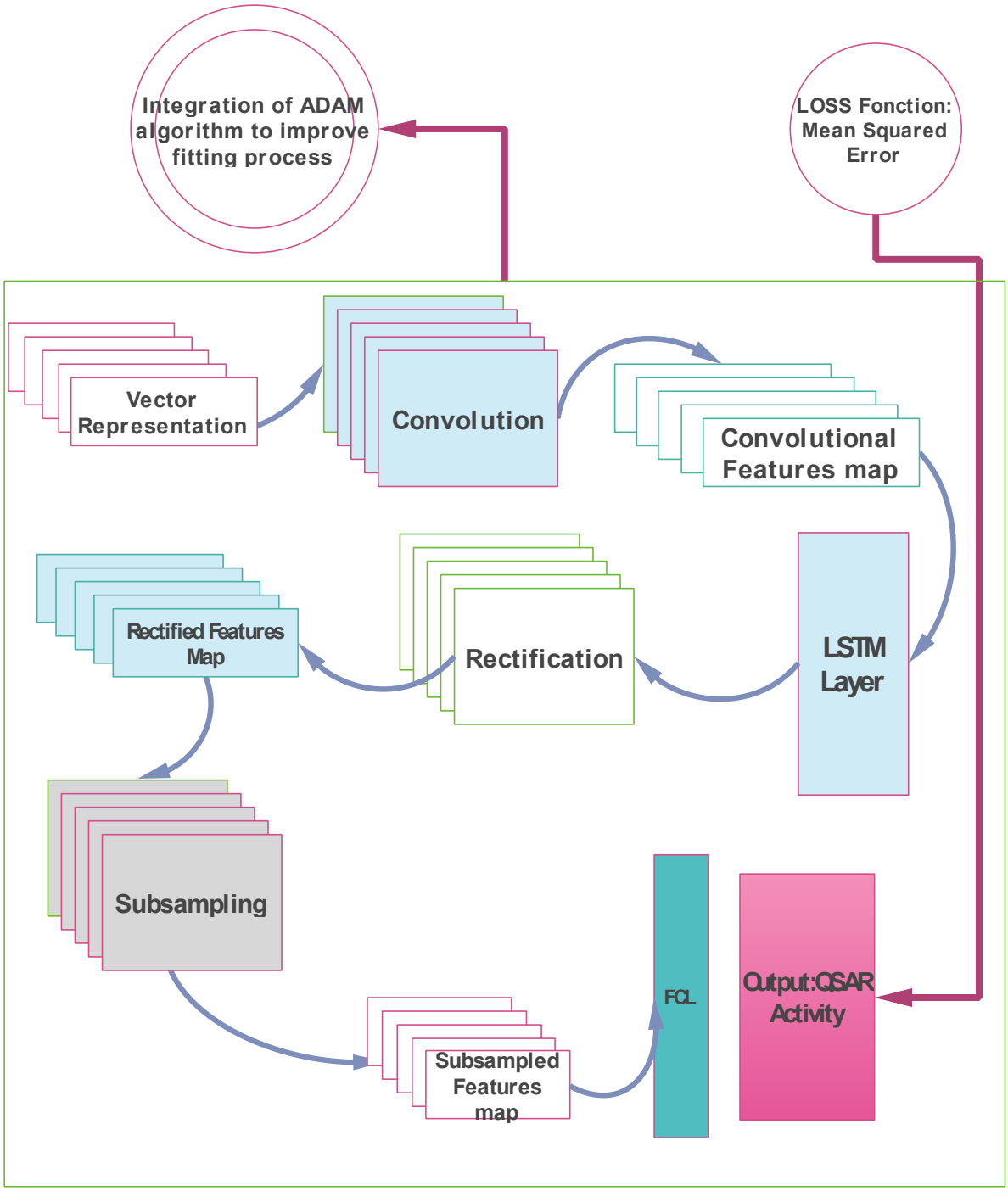


Figure 14.Workflow of Training model steps.

III.2.3. Results and discussion

With the IGC_{50} database of 1792 compounds and rat LD_{50} data of 7314 compounds, the predictive performance of three machine learning models was evaluated by using 80% of the data as a training set and 20% of the data as a test set. The derived compound vectors resulted from the trained embedding model were used as features in order to predict the toxicity values in the three machine learning models.

Word embedding deep learning is one of the most used methods in the NLP field, due to its efficacy to predict numerical vectors that represent word features. In our approach, we have applied a word embedding processing using word2vec to predict SMILES features using a generated corpus that contains all the concatenation possibilities of 1–4 substructures from data set including over 700000 compounds. Training the word2vec model with this corpus was able to generate a prediction function in a continuous space due to the huge number of words in the corpus. While using a small corpus size, the generated word2vec model will predict a word embedding representation in a discrete space, which will influence the performances of machine learning regressors. Indeed, the obtained results showed that by using our approach, we have got a better performance of our machine learning models.

a) Results on *T. pyriformis* (*TP*)

First, in order to define the length L of SMILES substructures, several trials of ConvLSTM modeling were to be examined using different values in order to establish the configuration that gives the best performances in terms of evaluation criteria. Table 23 shows how the change in statistical quality with respect to the changes of word size from 01 to 04.

Table 23.Statistical quality parameters of ConvLSTM model for different L values.

Metric	Length of sub structures (L)			
	L = 01	**L = 02**	**L = 03**	**L = 04**
$R^2_{training}$	0.76	0.85	0.97	/
R^2_{test}	0.62	0.54	0.84	0.00
MAE_{test}	0.43	0.49	0.30	0.78
$RMSE_{test}$	0.60	0.66	0.49	0.98

From Table 23, we can notice that the length of words that give the best performance is L equal to 03, which was selected from the development of the other machine learning models (RF and SVM).

The performance parameters of our approach models are shown in Table 24. In QSAR modeling the external validation is the effective way of quantifying the true predictability of a model performance [177]. Consequently, we have checked the model's quality according to good predictive potency. The ConvLSTM model has R^2 test and mean absolute errors (MAE) values of 0.84 and, 0.3, respectively. Those values are considerably higher than the corresponding values of the RF model (0.63 and 0.42, respectively) and SVM model (0.63, and 0.42, respectively). Consequently, the last two models are almost the same in terms of overall predictions.

The obtained results highlight the potential of the ConvLSTM model for the prediction of IGC_{50}. In contrast, deep network convolutions are very effective for extracting the relevant input features, contrary to SVM, which, depend to the big representation of compounds in the model and to tree based in RF mode.

Table 24.Statistical quality parameters of RF, SVM and ConvLSTM models on T. pyriformis IGC50 training, and test sets using our approach.

Machine learningtechnique	**Training set**	**Test set**		
	R^2	**R^2**	**RMSE**	**MAE**
ConvLSTM	0.97	0.84	0.49	0.30
RF	0.95	0.63	0.60	0.42
SVM	0.86	0.66	0.61	0.38

Comparison with literature models

In the past decades, different QSAR models, descriptors and machine learning have been used for modeling the toxicity of diverse organic compounds to *Tetrahymenapyriformis*. In general, for huge or heterogeneous data sets, it is almost impossible to develop a universal linear model or even nonlinear between descriptors and the target property. For that purpose, numerous efforts have been tried to develop a regression models based on the mechanism of toxicity, which can find witch more detail at QSAR Toolbox web site (http://oasis-lmc.org). For instance, Roy and Ghosh have built a QSTR models taking the toxicity IGC_{50} of a set of 384 aromatic compounds to *T. pyriformis* with extended topochemical atom (ETA) indices.

The authors showed that the best model was obtained from PLS analysis [178]. Another study by Toropov *et al.* have applied a correlations balance of SMILES-based optimal descriptors of 250 phenols to *Tetrahymenapyriformis*, where, they proposed to develop a specific model for different mechanisms of toxicity to *Tetrahymenapyriformis* [179]. To date, there are only a few studies that use a huge amount of heterogeneous data set for modeling the toxicity compounds to *Tetrahymenapyriformis*. With the help of new application of deep learning in drug discovery, two studies [175,180] have been used to approve our approach in the same huge data sets of IGC_{50} (1792 compounds) which are shown in Table 25.

Table 25. Comparison between the prediction results for the T. Pyriformis IGC50 test set.

	R^2	RMSE	MAE
TEST [174]			
Hierarchical	0.719	0.539	0.358
FDA	0.747	0.489	0.337
Group contribution	0.682	0.575	0.411
Nearest neighbor	0.600	0.638	0.451
TEST consensus	0.764	0.475	0.332
ESTDs			
RF	0.625	0.603	0.428
GBDT	0.705	0.538	0.374
ST-DNN	0.708	0.537	0.374
MT-DNN	0.723	0.517	0.378
Consensus	0.745	0.496	0.356
TopTox (Wu & Wei)			
RF	0.736	0.510	0.368
GBDT	0.787	0.455	0.316
ST-DNN	0.749	0.506	0.339
MT-DNN	0.770	0.472	0.331
Consensus	0.802	0.438	0.305
ST-hybrid (Abdul)			
ST-hybrid	0.810	/	/
Our approach			
ConvLSTM	**0.84**	**0.49**	**0.30**
RF	0.63	0.60	0.42
SVM	0.66	0.61	0.38

Toxicity Estimation Software Tool (TEST)
Element Specific Topological Descriptor (ESTD)
Gradient Boosting Decision Tree (GBDT)
Single-Task DNN (ST-DNN)
Multi-Task DNN (ST-DNN)

The results of Table 25 remarkably indicate that our model using ConvLSTM has good statistical performances in external validation than the other models.

b) Results on Rat LD_{50} data

Using a similar methodology and length of words for IGC50 and the ConvLSTM, the performance of our approach models in $RatLD_{50}$ is given in Table 26. The R^2 for test and training sets are 0.91 and 0.69, respectively. On the other hand, the root mean squared errors (RMSE) and mean absolute errors (MAE) are 0.59 and 0.43, respectively.

Table 26. Statistical quality parameters of RF, SVM and ConvLSTM models on Rat LD50 data, training and test sets using our approach.

Machine learningtechnique	Training set	Test set		
	R^2	R^2	MAE	RMSE
ConvLSTM	0.91	0.69	0.43	0.59

Comparison with literature models

Recently, several QSAR models have been developed to predict acute oral toxicity using multiple machine learning techniques. However, similar to IGC_{50} for this endpoint, it is quite difficult to develop a general model with reliable prediction accuracy for the overall data set. For instance, Zhu *et al.* built a combinatorial QSAR models for predicting Rat LD_{50} of 7385 compounds, in order to allow the comparison between their results and the commercial TOPKAT (Toxicity Prediction by Komputer Assisted Technology) [181].

Hou *et al.* applied the relevance vector machine (RVM) technique for building the regression models in order to predict the oral acute toxicity in rate of 7314 diverse chemicals. The obtained models were compared with those built using other six machine learning approaches, which are, counting k-nearest-neighbor regression, RF, SVM, local approximate Gaussian process, multilayer perceptron ensemble and eXtreme gradient boosting. The best model achieved R^2ext ranged from 0.57 to 0.66 [182].

Wu and Wei constructed a prediction models using the element-specific topological descriptor (ESTD) integrated with a variety of advanced machine learning algorithms including two deep neural networks (DNNs), two ensemble methods random forest (RF) and gradient boosting decision tree (GBDT). The main purpose was to construct topological learning strategies for quantitative toxicity analysis and prediction [183]. Table 27 summarizes the

results obtained in that paper together with our results using the same data set. From an analysis of the results mentioned in Table 27, we can conclude that our approach with ConvLSTM gives the best predictive performance than the other approaches (R^2 = 0.69) for the test set.

Table 27. Comparison between the prediction parameters for the Rat LD50 test set.

	R^2	RMSE	MAE
TEST [174]			
Hierarchical	0.58	0.65	0.46
FDA	0.56	0.66	0.48
Group contribution	0.56	0.66	0.48
TEST consensus	0.63	0.59	0.43
ESTDs			
RF	0.59	0.62	0.47
GBDT	0.60	0.61	0.45
ST-DNN	0.60	0.61	0.45
MT-DNN	0.61	0.60	0.44
Consensus	0.63	0.59	0.43
TopTox(Wu & Wei)			
RF	0.62	0.60	0.45
GBDT	0.63	0.59	0.44
ST-DNN	0.61	0.60	0.44
MT-DNN	0.63	0.59	0.43
Consensus	0.65	0.57	0.42
Our approach			
ConvLSTM	**0.69**	**0.59**	**0.43**

Toxicity Estimation Software Tool (TEST)
Element Specific Topological Descriptor (ESTD)
Gradient Boosting Decision Tree (GBDT)
Single-Task DNN (ST-DNN)
Multi-Task DNN (ST-DNN)

III.2.4. Conclusion

In this study, we have developed QSAR models using SMILES notations based only on textual representation without computing any of the molecule descriptors.

The following points are the advantages of our models:

i. The descriptor-based in QSAR modeling is generally costly in time especially when we use massive data.
ii. Our proposed approaches are an alternative method that can be fit in any machine learning algorithms.
iii. Two toxicity data sets were used in the present study.
iv. Experimental results show that the application of ConvLSTM in word embedding vectors of SMILES compounds improves the best predictive performance for IGC_{50} and $RatLD_{50}$ set, not only better than SVM and RF but better than literature approaches.
v. These promising results suggest that our approach can be applicable for predicting any physicochemical, biological, or pharmacological properties of interest.
vi. Because this significantly speeds up the training and lowers the memory requirements, irrespective of the size of the data.

GENERAL CONCLUSION

In order to protect both human health and environment, the European regulation on chemicals (REACH) places emphasis on the decrease of controlled toxicity testing, thus fostering the development of alternative methods, such as statistical methods based on existing data. Most of the time, these tests are carried out according to the Organization for Economic Cooperation and Development (OECD) test guidelines. In this context, several quantitative structure- activity relationships (QSAR) methods that relate the molecular descriptors of chemicals with their toxicity have been developed.

In the first part of this thesis, the state related to QSAR analysis for toxicity was presented together with the theories that support the proposed approach.

The first aim of this work was to build models based on Quantitative Structure-Activity-Activity Relationships (QSAAR), which allows the extrapolation of toxicity between different species.

These models have been developed and validated based on OECD principles. The proposed aquatic Auto *Pass-Pass* models can also be useful for predicting large amounts of chemicals without experimental values, employing the experimental values from a second species.

In the second application of this thesis, we have developed QSAR models using SMILES notations based only on the textual representation without computing any of the molecule descriptors.

The descriptor-based in QSAR modeling is generally costly in time especially when we use massive data. Our proposed approaches are an alternative method that can be fit in any machine learning algorithms.

Tow toxicity data sets were used in the present study. The experimental results show that the application of ConvLSTM in word embedding vectors of SMILES compounds improves the best predictive performance for $IGC_{50}$and $RatLD_{50}$ set, not only better than SVM and RF but better than that of the literature approaches.

These promising results suggest that our approach can be applicable for predicting any physicochemical, biological, or pharmacological properties of interest, because this significantly speeds up the training and lowers the memory requirements, wherever the size of the data.

REFERENCES

[1] J. Heppner, Book Review, Insecta Mundi. 2 (1988) 283–284.

[2] European Commission Environment Directorate General, Guidance Document on the Validation of (Quantitative) Structure-Activity Relationship [(Q)SAR] Models, OECD, 2014. doi:10.1787/9789264085442-en.

[3] X.J. Zhang, H.W. Qin, L.M. Su, W.C. Qin, M.Y. Zou, L.X. Sheng, Y.H. Zhao, M.H. Abraham, Interspecies correlations of toxicity to eight aquatic organisms: Theoretical considerations, Sci. Total Environ. 408 (2010) 4549–4555. doi:10.1016/j.scitotenv.2010.07.022.

[4] X. Wang, B. Fan, M. Fan, S. Belanger, J. Li, J. Chen, X. Gao, Z. Liu, Development and use of interspecies correlation estimation models in China for potential application in water quality criteria, Chemosphere. 240 (2020). doi:10.1016/j.chemosphere.2019.124848.

[5] V. Aruoja, M. Sihtmäe, H.C. Dubourguier, A. Kahru, Toxicity of 58 substituted anilines and phenols to algae Pseudokirchneriella subcapitata and bacteria Vibrio fischeri: Comparison with published data and QSARs, Chemosphere. 84 (2011) 1310–1320. doi:10.1016/j.chemosphere.2011.05.023.

[6] X. Wang, C. Sun, Y. Wang, L. Wang, Quantitative structure-activity relationships for the inhibition toxicity to root elongation of Cucumis sativus of selected phenols and interspecies correlation with Tetrahymena pyriformis, Chemosphere. 46 (2002) 153–161. doi:10.1016/S0045-6535(01)00133-3.

[7] G. Tugcu, M.D. Ertürk, M.T. Saçan, On the aquatic toxicity of substituted phenols to Chlorella vulgaris: QSTR with an extended novel data set and interspecies models, J. Hazard. Mater. 339 (2017) 122–130. doi:10.1016/j.jhazmat.2017.06.027.

[8] S. Cassani, S. Kovarich, E. Papa, P.P. Roy, L. van der Wal, P. Gramatica, Daphnia and fish toxicity of (benzo)triazoles: Validated QSAR models, and interspecies quantitative activity-activity modelling, J. Hazard. Mater. 258–259 (2013) 50–60. doi:10.1016/j.jhazmat.2013.04.025.

[9] L.Y. Fan, D. Zhu, Y. Yang, Y. Huang, S.N. Zhang, L.C. Yan, S. Wang, Y.H. Zhao, Comparison of modes of action among different trophic levels of aquatic organisms for pesticides and medications based on interspecies correlations and excess toxicity: Theoretical consideration, Ecotoxicol. Environ. Saf. 177 (2019) 25–31. doi:10.1016/j.ecoenv.2019.03.111.

[10] S. Kar, K. Roy, First report on interspecies quantitative correlation of ecotoxicity of pharmaceuticals, Chemosphere. 81 (2010) 738–747. doi:10.1016/j.chemosphere.2010.07.019.

[11] M.T.D. Cronin, J.C. Dearden, A.J. Dobbs, QSAR studies of comparative toxicity in aquatic organisms, Sci. Total Environ. 109–110 (1991) 431–439. doi:10.1016/0048-9697(91)90198-N.

[12] I. Kahn, U. Maran, E. Benfenati, T.I. Netzeva, T.W. Schultz, M.T.D. Cronin, Comparative quantitative structure-activity-activity relationships for toxicity to

Tetrahymena pyriformis and Pimephales prometas, ATLA Altern. to Lab. Anim. 35 (2007) 15–24. doi:10.1177/026119290703500112.

[13] J. Devillers, H. Devillers, Prediction of acute mammalian toxicity from QSARs and interspecies correlations, SAR QSAR Environ. Res. 20 (2009) 467–500. doi:10.1080/10629360903278651.

[14] A. Furuhama, T.I. Hayashi, H. Yamamoto, Development of QSAAR and QAAR models for predicting fish early-life stage toxicity with a focus on industrial chemicals, SAR QSAR Environ. Res. 30 (2019) 825–846. doi:10.1080/1062936X.2019.1669707.

[15] A. Sangion, P. Gramatica, Ecotoxicity interspecies QAAR models from Daphnia toxicity of pharmaceuticals and personal care products, SAR QSAR Environ. Res. 27 (2016) 781–798. doi:10.1080/1062936X.2016.1233139.

[16] A. Furuhama, T.I. Hayashi, H. Yamamoto, Development of models to predict fish early-life stage toxicity from acute Daphnia magna toxicity$, SAR QSAR Environ. Res. 29 (2018) 725–742. doi:10.1080/1062936X.2018.1513423.

[17] J. Ma, R.P. Sheridan, A. Liaw, G.E. Dahl, V. Svetnik, Deep neural nets as a method for quantitative structure-activity relationships, J. Chem. Inf. Model. 55 (2015) 263–274. doi:10.1021/ci500747n.

[18] Y. Xu, Z. Dai, F. Chen, S. Gao, J. Pei, L. Lai, Deep Learning for Drug-Induced Liver Injury, J. Chem. Inf. Model. 55 (2015) 2085–2093. doi:10.1021/acs.jcim.5b00238.

[19] S. Kearnes, K. McCloskey, M. Berndl, V. Pande, P. Riley, Molecular graph convolutions: moving beyond fingerprints, J. Comput. Aided. Mol. Des. 30 (2016) 595–608. doi:10.1007/s10822-016-9938-8.

[20] T.B. Hughes, N. Le Dang, G.P. Miller, S.J. Swamidass, Modeling reactivity to biological macromolecules with a deep multitask network, ACS Cent. Sci. 2 (2016) 529–537. doi:10.1021/acscentsci.6b00162.

[21] L. Cardamone, C. Engineering, M. Gocieva, P. Milano, M. Mancusi, B. Engineering, P. Torino, R. Padovani, B. Montrucchio, Chemistry, Toxicology and QSAR : an introduction, Politécnico de Milano, Instituto di Ricerche Farmacologiche "Mario Negri," 2010.

[22] REACH - Environment - European - Commission, (n.d.). http://ec.europa.eu/environment/chemicals/reach/reach_intro.htm (accessed Apr 21, 2013).

[23] J.C. Dearden, The History and Development of Quantitative Structure-Activity Relationships (QSARs), Int. J. Quant. Struct. Relationships. 1 (2016) 1–44. doi:10.4018/IJQSPR.2016010101.

[24] A.B. Raies, V.B. Bajic, In silico toxicology: computational methods for the prediction of chemical toxicity, Wiley Interdiscip. Rev. Comput. Mol. Sci. 6 (2016) 147–172. doi:10.1002/wcms.1240.

[25] P. Gramatica, S. Cassani, A. Sangion, Aquatic ecotoxicity of personal care products: QSAR models and ranking for prioritization and safer alternatives' design, Green Chem. 18 (2016) 4393–4406. doi:10.1039/C5GC02818C.

[26] J.S. Jaworska, M. Comber, C. Auer, C.J. Van Leeuwen, Summary of a workshop on regulatory acceptance of (Q)SARs for human health and environmental endpoints, Environ. Health Perspect. 111 (2003) 1358–1360. doi:10.1289/ehp.5757.

[27] OECD, OECD PRINCIPLES FOR THE VALIDATION, FOR REGULATORY PURPOSES, OF (QUANTITATIVE) STRUCTURE-ACTIVITY RELATIONSHIP MODELS These principles were agreed by OECD member countries at the 37, Biotechnology. (2004) 3–4.

[28] Q.E. Group., The report from the expert group on (Quantitative) Structure-Activity Relationships [(Q)SARs] on the principles for the validation of (Q)SARs., Organ. Econ. CO-OPERATION Dev. Paris. 49 (2004) 206-. http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?doclanguage=en&cote=env/jm/mono(2004)24.

[29] M.T.D. Cronin, T.W. Schultz, Pitfalls in QSAR, J. Mol. Struct. THEOCHEM. 622 (2003) 39–51. doi:10.1016/S0166-1280(02)00616-4.

[30] W.M.M. Philip H. Howard, Handbook of physical properties of organic chemicals, Lewis Publishers, Boca Raton, FL, 1997. doi:10.5860/choice.35-0285.

[31] K.L.E. Kaiser, J. Devillers, Ecotoxicity of Chemicals to Photobacterium Phosphoreum, Gordon and Breach Science Publishers, 1994. https://books.google.dz/books?id=JwizsCuwtmsC.

[32] J. Devillers, J.M. Exbrayat, Ecotoxicity of Chemicals to Amphibians, Gordon and Breach Science Publishers, 1992. https://books.google.dz/books?id=U9UQeEWGPXYC.

[33] M.Y. Galperin, The Molecular Biology Database Collection: 2007 update, Nucleic Acids Res. 35 (2007) D3–D4. doi:10.1093/nar/gkl1008.

[34] A. Gaulton, L.J. Bellis, A.P. Bento, J. Chambers, M. Davies, A. Hersey, Y. Light, S. McGlinchey, D. Michalovich, B. Al-Lazikani, J.P. Overington, ChEMBL: A large-scale bioactivity database for drug discovery, Nucleic Acids Res. 40 (2012) 1100–1107. doi:10.1093/nar/gkr777.

[35] Y. Wang, J. Xiao, T.O. Suzek, J. Zhang, J. Wang, Z. Zhou, L. Han, K. Karapetyan, S. Dracheva, B.A. Shoemaker, E. Bolton, A. Gindulyte, S.H. Bryant, PubChem's BioAssay database, Nucleic Acids Res. 40 (2012). doi:10.1093/nar/gkr1132.

[36] R. Diderichs, Chapter 16. Tools for Category Formation and Read-Across: Overview of the OECD (Q)SAR Application Toolbox, in: n.d.: pp. 385–407. doi:10.1039/9781849732093-00385.

[37] T.W. Schultz, R. Diderich, C.D. Kuseva, O.G. Mekenyan, The OECD QSAR Toolbox Starts Its Second Decade, in: 2018: pp. 55–77. doi:10.1007/978-1-4939-7899-1_2.

[38] OCHEM – The Online Chemical Database, (n.d.). https://ochem.eu/home/show.do. Accessed 29 Apr 2019.

[39] R.P.C. Barros, N.F. Sousa, L. Scotti, M.T. Scotti, Use of Machine Learning and Classical QSAR Methods in Computational Ecotoxicology, in: 2020: pp. 151–175. doi:10.1007/978-1-0716-0150-1_7.

[40] VEGA-QSAR, Virtual models for Evaluating the properties of chemicals within a Global Architecture, (n.d.). http://www.vega-qsar.eu/.

[41] E. Benfenati, A. Lombardo, VEGAHUB for Ecotoxicological QSAR Modeling, in: Methods Pharmacol. Toxicol., 2020: pp. 759–787. doi:10.1007/978-1-0716-0150-1_30.

[42] S. Ghosh, S. Kar, J. Leszczynski, Ecotoxicity Databases for QSAR Modeling, in: Methods Pharmacol. Toxicol., 2020: pp. 709–758. doi:10.1007/978-1-0716-0150-1_29.

[43] S. Raimondo, M.G. Barron, Application of Interspecies Correlation Estimation (ICE) models and QSAR in estimating species sensitivity to pesticides, SAR QSAR Environ. Res. 31 (2020) 1–18. doi:10.1080/1062936X.2019.1686716.

[44] X. Wang, B. Fan, M. Fan, S. Belanger, J. Li, J. Chen, X. Gao, Z. Liu, Development and use of interspecies correlation estimation models in China for potential application in water quality criteria, Chemosphere. 240 (2020). doi:10.1016/j.chemosphere.2019.124848.

[45] G. Tugcu, M.D. Ertürk, M.T. Saçan, On the aquatic toxicity of substituted phenols to Chlorella vulgaris: QSTR with an extended novel data set and interspecies models, J. Hazard. Mater. 339 (2017) 122–130. doi:10.1016/j.jhazmat.2017.06.027.

[46] L.Y. Fan, D. Zhu, Y. Yang, Y. Huang, S.N. Zhang, L.C. Yan, S. Wang, Y.H. Zhao, Comparison of modes of action among different trophic levels of aquatic organisms for pesticides and medications based on interspecies correlations and excess toxicity: Theoretical consideration, Ecotoxicol. Environ. Saf. 177 (2019) 25–31. doi:10.1016/j.ecoenv.2019.03.111.

[47] S. Cassani, S. Kovarich, E. Papa, P.P. Roy, L. van der Wal, P. Gramatica, Daphnia and fish toxicity of (benzo)triazoles: Validated QSAR models, and interspecies quantitative activity–activity modelling, J. Hazard. Mater. 258–259 (2013) 50–60. doi:10.1016/j.jhazmat.2013.04.025.

[48] V. Aruoja, M. Sihtmäe, H.C. Dubourguier, A. Kahru, Toxicity of 58 substituted anilines and phenols to algae Pseudokirchneriella subcapitata and bacteria Vibrio fischeri: Comparison with published data and QSARs, Chemosphere. 84 (2011) 1310–1320. doi:10.1016/j.chemosphere.2011.05.023.

[49] A. Furuhama, T.I. Hayashi, H. Yamamoto, Development of QSAAR and QAAR models for predicting fish early-life stage toxicity with a focus on industrial chemicals, SAR QSAR Environ. Res. 30 (2019) 825–846. doi:10.1080/1062936X.2019.1669707.

[50] A. Sangion, P. Gramatica, Ecotoxicity interspecies QAAR models from Daphnia toxicity of pharmaceuticals and personal care products, SAR QSAR Environ. Res. 27 (2016) 781–798. doi:10.1080/1062936X.2016.1233139.

[51] A. Furuhama, T.I. Hayashi, H. Yamamoto, Development of models to predict fish early-life stage toxicity from acute Daphnia magna toxicity $, SAR QSAR Environ. Res. 29 (2018) 725–742. doi:10.1080/1062936X.2018.1513423.

[52] M.T.D. Cronin, Chapter 18. Biological Read-Across: Mechanistically-Based Species–Species and Endpoint–Endpoint Extrapolations, in: n.d.: pp. 446–477. doi:10.1039/9781849732093-00446.

[53] P. Gramatica, S. Cassani, P.P. Roy, S. Kovarich, C.W. Yap, E. Papa, QSAR modeling is not "Push a button and find a correlation": A case study of toxicity of (Benzo-)triazoles on Algae, Mol. Inform. 31 (2012) 817–835. doi:10.1002/minf.201200075.

[54] P. Ambure, R.B. Aher, A. Gajewicz, T. Puzyn, K. Roy, "NanoBRIDGES" software: Open access tools to perform QSAR and nano-QSAR modeling, Chemom. Intell. Lab. Syst. 147 (2015) 1–13. doi:10.1016/j.chemolab.2015.07.007.

[55] A. Tropsha, Best practices for QSAR model development, validation, and exploitation, Mol. Inform. 29 (2010) 476–488. doi:10.1002/minf.201000061.

[56] E. Papa, F. Villa, P. Gramatica, Statistically validated QSARs, based on theoretical descriptors, for modeling aquatic toxicity of organic chemicals in Pimephales promelas (fathead minnow), J. Chem. Inf. Model. 45 (2005) 1256–1266. doi:10.1021/ci0502121.

[57] OECD, Guidance document on the validation of (quantitative) structure-activity relationship [(Q)SAR] models, (2007). http://www.oecd.org/chemicalsafety/testing/oecd-guidelines-testing-chemicals-related-documents.htm.

[58] A.F.A. Cros, Action de l'alcool amylique sur l'organisme, Faculté de médecine de Strasbourg, 1863. https://books.google.dz/books?id=da6QmgEACAAJ.

[59] C. HANSCH, P.P. MALONEY, T. FUJITA, R.M. MUIR, Correlation of Biological Activity of Phenoxyacetic Acids with Hammett Substituent Constants and Partition Coefficients, Nature. 194 (1962) 178–180. doi:10.1038/194178b0.

[60] K. Roy, I. Mitra, On Various Metrics Used for Validation of Predictive QSAR Models with Applications in Virtual Screening and Focused Library Design, Comb. Chem. High Throughput Screen. 14 (2011) 450–474. doi:10.2174/138620711795767893.

[61] A. Tropsha, Best Practices for QSAR Model Development, Validation, and Exploitation, Mol. Inform. 29 (2010) 476–488. doi:10.1002/minf.201000061.

[62] KNIME Analytics Platform: Professional open-source software, (n.d.). https://www.knime.org.

[63] M.R. Berthold, N. Cebron, F. Dill, T.R. Gabriel, T. Kötter, T. Meinl, P. Ohl, K. Thiel, B. Wiswedel, KNIME - The Konstanz Information Miner, SIGKDD Explor. 11 (2009) 26–31. doi:10.1145/1656274.1656280.

[64] W.A. Warr, Scientific workflow systems: Pipeline Pilot and KNIME, J. Comput. Aided. Mol. Des. 26 (2012) 801–804. doi:10.1007/s10822-012-9577-7.

[65] C. Chichester, D. Digles, R. Siebes, A. Loizou, P. Groth, L. Harland, Drug discovery FAQs: Workflows for answering multidomain drug discovery questions, Drug Discov. Today. 20 (2015) 399–405. doi:10.1016/j.drudis.2014.11.006.

[66] D. Young, T. Martin, R. Venkatapathy, P. Harten, Are the chemical structures in your QSAR correct?, QSAR Comb. Sci. 27 (2008) 1337–1345. doi:10.1002/qsar.200810084.

[67] D. Fourches, E. Muratov, a. Tropsha, Trust but verify: on the importance of chemical structure curation in chemoinformatics and QSAR modeling research, J. Chem. Inf.

Model. 50 (2010) 1189–1204.

[68] S. Beisken, T. Meinl, B. Wiswedel, L.F. De Figueiredo, M. Berthold, KNIME-CDK : Workflow-driven cheminformatics, (2013) 2–5.

[69] GGA Software Services LLC, Indigo Nodes for KNIME, (n.d.). http://ggasoftware.com/opensource/%250Aindigo.

[70] G. Landrum, RDKit Nodes for KNIME, (n.d.). https://tech.knime.org/community/cdk.

[71] Erl Wood Cheminformatics nodes for KNIME, (n.d.). https://tech.knime.org/community/erlwood.

[72] INFOCOM Releases ChemAxon Ltd, (n.d.). https://www.chemaxon.com (accessed July 12, 2012).

[73] A. Mauri, V. Consonni, R. Todeschini, Molecular Descriptors, in: Handb. Comput. Chem., Springer International Publishing, Cham, 2017: pp. 2065–2093. doi:10.1007/978-3-319-27282-5_51.

[74] D. Weininger, SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules, J. Chem. Inf. Model. 28 (1988) 31–36. doi:10.1021/ci00057a005.

[75] W.J. Wiswesser, Historic development of chemical notations, J. Chem. Inf. Comput. Sci. 25 (1985) 258–263. doi:10.1021/ci00047a023.

[76] A.M. Moore, A Line-Formula Chemical Notation., J. Am. Chem. Soc. 77 (1955) 2032–2032. doi:10.1021/ja01612a112.

[77] J.L. Durant, B.A. Leland, D.R. Henry, J.G. Nourse, Reoptimization of MDL Keys for Use in Drug Discovery, J. Chem. Inf. Comput. Sci. 42 (2002) 1273–1280. doi:10.1021/ci010132r.

[78] E. Fernández-de Gortari, C.R. García-Jacas, K. Martinez-Mayorga, J.L. Medina-Franco, Database fingerprint (DFP): an approach to represent molecular databases, J. Cheminform. 9 (2017) 9. doi:10.1186/s13321-017-0195-1.

[79] R. Todeschini, V. Consonni, Molecular Descriptors for Chemoinformatics, Wiley, 2009. doi:10.1002/9783527628766.

[80] T.W. Schultz, T.I. Netzeva, M.T.D. Cronin, Selection of data sets for qsars: Analyses of tetrahymena toxicity from aromatic compounds, SAR QSAR Environ. Res. 14 (2003) 59–81. doi:10.1080/1062936021000058782.

[81] J.G. Topliss, R.J. Costello, Chance correlations in structure-activity studies using multiple regression analysis, J. Med. Chem. 15 (1972) 1066–1068. doi:10.1021/jm00280a017.

[82] M. Balls, B.J. Blaauboer, J.H. Fentem, L. Bruner, R.D. Combes, B. Ekwall, R.J. Fielder, A. Guillouzo, R.W. Lewis, D.P. Lovell, C.A. Reinhardt, G. Repetto, D. Sladowski, H. Spielmann, F. Zucco, Practical Aspects of the Validation of Toxicity Test Procedures, Altern. to Lab. Anim. 23 (1995) 129–146. doi:10.1177/026119299502300116.

[83] K. Roy, On some aspects of validation of predictive quantitative structure–activity relationship models, Expert Opin. Drug Discov. 2 (2007) 1567–1577. doi:10.1517/17460441.2.12.1567.

[84] M. Shen, C. Béguin, A. Golbraikh, J.P. Stables, H. Kohn, A. Tropsha, Application of Predictive QSAR Models to Database Mining: Identification and Experimental Validation of Novel Anticonvulsant Compounds, J. Med. Chem. 47 (2004) 2356–2364. doi:10.1021/jm030584q.

[85] A. Tropsha, P. Gramatica, V. Gombar, The Importance of Being Earnest: Validation is the Absolute Essential for Successful Application and Interpretation of QSPR Models, QSAR Comb. Sci. 22 (2003) 69–77. doi:10.1002/qsar.200390007.

[86] D.W. Osten, Selection of optimal regression models via cross-validation, J. Chemom. 2 (1988) 39–48. doi:10.1002/cem.1180020106.

[87] B. Efron, Estimating the Error Rate of a Prediction Rule: Improvement on Cross-Validation, J. Am. Stat. Assoc. 78 (1983) 316. doi:10.2307/2288636.

[88] A. Ghose, V. Viswanadhan, Tools for Designing Diverse, Druglike, Cost-Effective Combinatorial Libraries, CRC Press, 2001. doi:10.1201/9781482270761-16.

[89] A. Golbraikh, A. Tropsha, Beware of q2!, J. Mol. Graph. Model. 20 (2002) 269–276. doi:10.1016/S1093-3263(01)00123-1.

[90] I. Mitra, A. Saha, K. Roy, Exploring quantitative structure–activity relationship studies of antioxidant phenolic compounds obtained from traditional Chinese medicinal plants, Mol. Simul. 36 (2010) 1067–1079. doi:10.1080/08927022.2010.503326.

[91] K. Roy, S. Kar, R.N. Das, Validation of QSAR Models, in: Underst. Basics QSAR Appl. Pharm. Sci. Risk Assess., Elsevier, 2015: pp. 231–289. doi:10.1016/B978-0-12-801505-6.00007-7.

[92] R. Wehrens, H. Putter, L.M.. Buydens, The bootstrap: a tutorial, Chemom. Intell. Lab. Syst. 54 (2000) 35–52. doi:10.1016/S0169-7439(00)00102-7.

[93] A. Tropsha, A. Golbraikh, Predictive QSAR Modeling Workflow, Model Applicability Domains, and Virtual Screening, Curr. Pharm. Des. 13 (2007) 3494–3504. doi:10.2174/138161207782794257.

[94] P. Gramatica, Principles of QSAR models validation: internal and external, QSAR Comb. Sci. 26 (2007) 694–701. doi:10.1002/qsar.200610151.

[95] K. Roy, P. Ambure, S. Kar, P.K. Ojha, Is it possible to improve the quality of predictions from an "intelligent" use of multiple QSAR/QSPR/QSTR models?, J. Chemom. 32 (2018) e2992. doi:10.1002/cem.2992.

[96] K. Roy, S. Kar, R.N. Das, Understanding the Basics of QSAR for Applications in Pharmaceutical Sciences and Risk Assessment, Elsevier Science, 2015. https://books.google.dz/books?id=bkFOBQAAQBAJ.

[97] K. Roy, S. Kar, P. Ambure, On a simple approach for determining applicability domain of QSAR models, Chemom. Intell. Lab. Syst. 145 (2015) 22–29. doi:10.1016/j.chemolab.2015.04.013.

[98] K. Roy, ed., Quantitative Structure-Activity Relationships in Drug Design, Predictive Toxicology, and Risk Assessment, IGI Global, 2015. doi:10.4018/978-1-4666-8136-1.

[99] T.I. Netzeva, A.P. Worth, T. Aldenberg, R. Benigni, M.T.D. Cronin, P. Gramatica, J.S. Jaworska, S. Kahn, G. Klopman, C.A. Marchant, G. Myatt, N. Nikolova-Jeliazkova, G.Y. Patlewicz, R. Perkins, D.W. Roberts, T.W. Schultz, D.T. Stanton, J.J.M. van de Sandt, W. Tong, G. Veith, C. Yang, Current Status of Methods for Defining the Applicability Domain of (Quantitative) Structure-Activity Relationships, Altern. to Lab. Anim. 33 (2005) 155–173. doi:10.1177/026119290503300209.

[100] J. Jaworska, N. Nikolova-Jeliazkova, T. Aldenberg, QSAR Applicability Domain Estimation by Projection of the Training Set in Descriptor Space: A Review, Altern. to Lab. Anim. 33 (2005) 445–459. doi:10.1177/026119290503300508.

[101] P. Gramatica, S. Cassani, P.P. Roy, S. Kovarich, C.W. Yap, E. Papa, QSAR Modeling is not "Push a Button and Find a Correlation": A Case Study of Toxicity of (Benzo-)triazoles on Algae, Mol. Inform. 31 (2012) 817–835. doi:10.1002/minf.201200075.

[102] G. Melagraki, A. Afantitis, H. Sarimveis, O. Igglessi-Markopoulou, P.A. Koutentis, G. Kollias, In silico exploration for identifying structure-activity relationship of MEK inhibition and oral bioavailability for isothiazole derivatives, Chem. Biol. Drug Des. 76 (2010) 397–406. doi:10.1111/j.1747-0285.2010.01029.x.

[103] A. Afantitis, G. Melagraki, P.A. Koutentis, H. Sarimveis, G. Kollias, Ligand - Based virtual screening procedure for the prediction and the identification of novel ??-amyloid aggregation inhibitors using Kohonen maps and Counterpropagation Artificial Neural Networks, Eur. J. Med. Chem. 46 (2011) 497–508. doi:10.1016/j.ejmech.2010.11.029.

[104] A. Afantitis, G. Melagraki, H. Sarimveis, P.A. Koutentis, O. Igglessi-Markopoulou, G. Kollias, A combined LS-SVM & MLR QSAR workflow for predicting the inhibition of CXCR3 receptor by quinazolinone analogs, Mol. Divers. 14 (2010) 225–235. doi:10.1007/s11030-009-9163-7.

[105] A. Golbraikh, A. Tropsha, Beware of q^2!, J. Mol. Graph. Model. 20 (2002) 269–276. doi:10.1016/S1093-3263(01)00123-1.

[106] K. Bouhedjar, A.K. Nacereddine, H. Ghorab, A. Djerourou, QSPR Modeling For Critical Temperatures Of Organic Compounds Using Hybrid Optimal Descriptors, Int. J. Quant. Struct. Relationships. 4 (2019) 15–26. doi:10.4018/IJQSPR.2019100102.

[107] D.E. Goldberg, G. David Edward, D.E.G. Goldberg, V.A.P.H.D.E. Goldberg, Genetic Algorithms in Search, Optimization, and Machine Learning, Addison-Wesley Publishing Company, 1989. https://books.google.dz/books?id=2IIJAAAACAAJ.

[108] R. Leardi, R. Boggia, M. Terrile, Genetic algorithms as a strategy for feature selection, J. Chemom. 6 (1992) 267–281. doi:10.1002/cem.1180060506.

[109] R. Leardi, Genetic algorithms in chemometrics and chemistry: a review, J. Chemom. 15 (2001) 559–569. doi:10.1002/cem.651.

[110] V. Venkatraman, A.R. Dalby, Z.R. Yang, Evaluation of Mutual Information and Genetic Programming for Feature Selection in QSAR, J. Chem. Inf. Comput. Sci. 44 (2004) 1686–1692. doi:10.1021/ci049933v.

[111] S. Forrest, Genetic algorithms: principles of natural selection applied to computation, Science (80-.). 261 (1993) 872–878. doi:10.1126/science.8346439.

[112] M. Cassotti, F. Grisoni, T6_moleculardescriptors_variable_selection, Mol. Descriptors. (n.d.) 1–11. https://www.math.arizona.edu/~hzhang/.

[113] A. Miller, Subset Selection in Regression, CRC Press, 2002. https://books.google.dz/books?id=7p59iir822sC.

[114] M. Goodarzi, M.P. Freitas, R. Jensen, Ant colony optimization as a feature selection method in the QSAR modeling of anti-HIV-1 activities of 3-(3,5-dimethylbenzyl)uracil derivatives using MLR, PLS and SVM regressions, Chemom. Intell. Lab. Syst. 98 (2009) 123–129. doi:10.1016/j.chemolab.2009.05.005.

[115] M. Dorigo, C. Blum, Ant colony optimization theory: A survey, Theor. Comput. Sci. 344 (2005) 243–278. doi:10.1016/j.tcs.2005.05.020.

[116] G. Gini, F. Zanoli, Machine Learning and Deep Learning Methods in Ecotoxicological QSAR Modeling, in: 2020: pp. 111–149. doi:10.1007/978-1-0716-0150-1_6.

[117] W.S. Noble, What is a support vector machine?, Nat. Biotechnol. 24 (2006) 1565–1567. doi:10.1038/nbt1206-1565.

[118] H.X. Liu, R.S. Zhang, X.J. Yao, M.C. Liu, Z.D. Hu, B.T. Fan, QSAR study of ethyl 2-[(3-methyl-2,5-dioxo(3-pyrrolinyl))amino]-4-(trifluoromethyl) pyrimidine-5-carboxylate: An inhibitor of Ap-1 and NF-/κB mediated gene expression based on support vector machines, J. Chem. Inf. Comput. Sci. 43 (2003) 1288–1296. doi:10.1021/ci0340355.

[119] B. Chen, R.P. Sheridan, V. Hornak, J.H. Voigt, Comparison of random forest and Pipeline Pilot Naïve Bayes in prospective QSAR predictions, J. Chem. Inf. Model. 52 (2012) 792–803. doi:10.1021/ci200615h.

[120] V. Svetnik, A. Liaw, C. Tong, J. Christopher Culberson, R.P. Sheridan, B.P. Feuston, Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling, J. Chem. Inf. Comput. Sci. 43 (2003) 1947–1958. doi:10.1021/ci034160g.

[121] Y. Bengio, A. Courville, P. Vincent, Representation Learning: A Review and New Perspectives, IEEE Trans. Pattern Anal. Mach. Intell. 35 (2013) 1798–1828. doi:10.1109/TPAMI.2013.50.

[122] J. Panteleev, H. Gao, L. Jia, Recent applications of machine learning in medicinal chemistry, Bioorganic Med. Chem. Lett. 28 (2018) 2807–2815. doi:10.1016/j.bmcl.2018.06.046.

[123] Y. Uesawa, Quantitative structure–activity relationship analysis using deep learning based on a novel molecular image input technique, Bioorganic Med. Chem. Lett. 28 (2018) 3400–3403. doi:10.1016/j.bmcl.2018.08.032.

[124] S.K. Heo, U. Safder, C.K. Yoo, Deep learning driven QSAR model for environmental toxicology: Effects of endocrine disrupting chemicals on human health, Environ. Pollut. 253 (2019) 29–38. doi:10.1016/j.envpol.2019.06.081.

[125] F. Ghasemi, A. Mehridehnavi, A. Fassihi, H. Pérez-sánchez, Deep neural network in QSAR studies using deep belief network, Appl. Soft Comput. J. 62 (2018) 251–258. doi:10.1016/j.asoc.2017.09.040.

[126] M. Fernandez, F. Ban, G. Woo, M. Hsing, T. Yamazaki, E. Leblanc, P.S. Rennie, W.J. Welch, A. Cherkasov, Toxic Colors: The Use of Deep Learning for Predicting Toxicity of Compounds Merely from Their Graphic Images, J. Chem. Inf. Model. 58 (2018) 1533–1543. doi:10.1021/acs.jcim.8b00338.

[127] Y. Zhang, J. Zhao, Y. Wang, Y. Fan, L. Zhu, Y. Yang, X. Chen, T. Lu, Y. Chen, H. Liu, Prediction of hERG K+ channel blockage using deep neural networks, Chem. Biol. Drug Des. (2019) 0–3. doi:10.1111/cbdd.13600.

[128] E. Benfenati, A. Golbamaki, G. Raitano, A. Roncaglioni, S. Manganelli, F. Lemke, U. Norinder, E. Lo Piparo, M. Honma, A. Manganaro, G. Gini, A large comparison of integrated SAR/QSAR models of the Ames test for mutagenicity$, SAR QSAR Environ. Res. 29 (2018) 591–611. doi:10.1080/1062936X.2018.1497702.

[129] M. Hirohara, Y. Saito, Y. Koda, K. Sato, Y. Sakakibara, Convolutional neural network based on SMILES representation of compounds for detecting chemical motif, BMC Bioinformatics. 19 (2018). doi:10.1186/s12859-018-2523-5.

[130] A. Mayr, G. Klambauer, T. Unterthiner, S. Hochreiter, DeepTox: Toxicity prediction using deep learning, Front. Environ. Sci. 3 (2016). doi:10.3389/fenvs.2015.00080.

[131] A.Y. Ng, H. Lee, R. Grosse, R. Ranganath, Unsupervised learning of hierarchical representations with convolutional deep belief networks, Commun. ACM. 54 (2011) 95–103. doi:10.1145/2001269.

[132] C. Szegedy, Wei Liu, Yangqing Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: 2015 IEEE Conf. Comput. Vis. Pattern Recognit., IEEE, 2015: pp. 1–9. doi:10.1109/CVPR.2015.7298594.

[133] S. Fernández, A. Graves, J. Schmidhuber, An application of recurrent neural networks to discriminative keyword spotting, Lect. Notes Comput. Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics). 4669 LNCS (2007) 220–229. doi:10.1007/978-3-540-74695-9_23.

[134] Y. Bengio, Learning deep architectures for AI, 2009. doi:10.1561/2200000006.

[135] M.T.D. Cronin, T.I. Netzeva, J.C. Dearden, R. Edwards, A.D.P. Worgan, Assessment and Modeling of the Toxicity of Organic Chemicals to Chlorella vulgaris : Development of a Novel Database, Chem. Res. Toxicol. 17 (2004) 545–554. doi:10.1021/tx0342518.

[136] M.R. Berthold, N. Cebron, F. Dill, T.R. Gabriel, T. Kötter, T. Meinl, P. Ohl, K. Thiel, B. Wiswedel, KNIME - the Konstanz Information Miner: Version 2.0 and Beyond, SIGKDD Explor. Newsl. 11 (2009) 26–31. doi:10.1145/1656274.1656280.

[137] D.D. Varsou, S. Nikolakopoulos, A. Tsoumanis, G. Melagraki, A. Afantitis, ENALOS+ KNIME nodes: New cheminformatics tools for drug discovery, Methods Mol. Biol. 1824 (2018) 113–138. doi:10.1007/978-1-4939-8630-9_7.

[138] J.A. Castillo-Garit, Y. Marrero-Ponce, J. Escobar, F. Torrens, R. Rotondo, A novel approach to predict aquatic toxicity from molecular structure, Chemosphere. 73 (2008) 415–427. doi:10.1016/j.chemosphere.2008.05.024.

[139] J.J. Li, X.H. Wang, Y. Wang, Y. Wen, W.C. Qin, L.M. Su, Y.H. Zhao, Discrimination of excess toxicity from narcotic effect: Influence of species sensitivity and bioconcentration on the classification of modes of action, Chemosphere. 120 (2015) 660–673. doi:10.1016/j.chemosphere.2014.10.013.

[140] M.T.. Cronin, A.O. Aptula, J.C. Duffy, T.I. Netzeva, P.H. Rowe, I. V Valkova, T. Wayne Schultz, Comparative assessment of methods to develop QSARs for the prediction of the toxicity of phenols to Tetrahymena pyriformis, Chemosphere. 49 (2002) 1201–1221. doi:10.1016/S0045-6535(02)00508-8.

[141] M. Pérez González, H. González Díaz, M.A. Cabrera, R. Molina Ruiz, A novel approach to predict a toxicological property of aromatic compounds in the Tetrahymena pyriformis, Bioorganic Med. Chem. 12 (2004) 735–744. doi:10.1016/j.bmc.2003.11.028.

[142] M.T.D. Cronin, T.W. Schultz, Structure-toxicity relationships for three mechanisms of action of toxicity to Vibrio fischeri, Ecotoxicol. Environ. Saf. 39 (1998) 65–69. doi:10.1006/eesa.1997.1618.

[143] S. Dimitrov, Y. Koleva, T.W. Schultz, J.D. Walker, O. Mekenyan, Interspecies quantitative structure-activity relationship model for aldehydes: Aquatic toxicity, Environ. Toxicol. Chem. 23 (2004) 463–470. doi:10.1897/02-579.

[144] A.R. Katritzky, P. Oliferenko, A. Oliferenko, A. Lomaka, M. Karelson, Nitrobenzene toxicity: QSAR correlations and mechanistic interpretations, J. Phys. Org. Chem. 16 (2003) 811–817. doi:10.1002/poc.643.

[145] F. Schramm, A. Müller, H. Hammer, A. Paschke, G. Schüürmann, Epoxide and thiirane toxicity in vitro with the ciliates tetrahymena pyriformis: Structural alerts indicating excess toxicity, Environ. Sci. Technol. 45 (2011) 5812–5819. doi:10.1021/es200081n.

[146] T.W. Schultz, Tetratox: Tetrahymena pyriformis population growth impairment endpoint - A surrogate for fish lethality, Toxicol. Methods. 7 (1997) 289–309. doi:10.1080/105172397243079.

[147] T.W. Schultz, T.I. Netzeva, D.W. Roberts, M.T.D. Cronin, Structure-toxicity relationships for the effects to Tetrahymena pyriformis of aliphatic, carbonyl-containing, α,β-unsaturated chemicals, Chem. Res. Toxicol. 18 (2005) 330–341. doi:10.1021/tx049833j.

[148] (IRFMN), Mario Negri Institute for Pharmacological Research, (n.d.). http://www.vega-qsar.eu.

[149] US EPA, US EPA, ECOTOX. Available at http://cfpub.epa.gov/ecotox/, United States, 2017, (n.d.).

[150] M. of the E. in Japan, Results of Eco-toxicity tests of chemicals conducted by Ministry of the Environment in Japan (- March 2016), (2016) 1–33.

[151] L.M. Su, X. Liu, Y. Wang, J.J. Li, X.H. Wang, L.X. Sheng, Y.H. Zhao, The discrimination of excess toxicity from baseline effect: Effect of bioconcentration, Sci. Total Environ. 484 (2014) 137–145. doi:10.1016/j.scitotenv.2014.03.040.

[152] S. Gómez-Ganau, M. Marzo, R. Gozalbes, E. Benfenati, Computational Approaches to Evaluate Ecotoxicity of Biocides: Cases from the Project COMBASE, in: 2020: pp. 387–404. doi:10.1007/978-1-0716-0150-1_17.

[153] J.J. Li, X.J. Zhang, Y. Yang, T. Huang, C. Li, L. Su, Y.H. Zhao, M.T.D. Cronin, Development of thresholds of excess toxicity for environmental species and their application to identification of modes of acute toxic action, Sci. Total Environ. 616–617 (2018) 491–499. doi:10.1016/j.scitotenv.2017.10.308.

[154] A. Kienzler, M.G. Barron, S.E. Belanger, A. Beasley, M.R. Embry, Mode of Action (MOA) Assignment Classifications for Ecotoxicology: An Evaluation of Approaches, Environ. Sci. Technol. 51 (2017) 10203–10211. doi:10.1021/acs.est.7b02337.

[155] R. Todeschini, V. Consonni, Handbook of Molecular Descriptors, Wiley, 2000. doi:10.1002/9783527613106.

[156] I. Lessigiarska, A.P. Worth, B. Sokull-Klüttgen, S. Jeram, J.C. Dearden, T.I. Netzeva, M.T.D. Cronin, QSAR investigation of a large data set for fish, algae and Daphnia toxicity, SAR QSAR Environ. Res. 15 (2004) 413–431. doi:10.1080/10629360412331297416.

[157] A. Furuhama, T.I. Hayashi, N. Tatarazako, Acute to chronic estimation of Daphnia magna toxicity within the QSAAR framework, SAR QSAR Environ. Res. 27 (2016) 833–850. doi:10.1080/1062936X.2016.1243151.

[158] M.T.D. Cronin, T.I. Netzeva, J.C. Dearden, R. Edwards, A.D.P. Worgan, Assessment and Modeling of the Toxicity of Organic Chemicals to Chlorella vulgaris: Development of a Novel Database, Chem. Res. Toxicol. 17 (2004) 545–554. doi:10.1021/tx0342518.

[159] A. Furuhama, K. Hasunuma, Y. Aoki, Interspecies quantitative structure–activity–activity relationships (QSAARs) for prediction of acute aquatic toxicity of aromatic amines and phenols, SAR QSAR Environ. Res. 26 (2015) 301–323. doi:10.1080/1062936X.2015.1032347.

[160] S. Cassani, S. Kovarich, E. Papa, P.P. Roy, L. van der Wal, P. Gramatica, Daphnia and fish toxicity of (benzo)triazoles: Validated QSAR models, and interspecies quantitative activity–activity modelling, J. Hazard. Mater. 258–259 (2013) 50–60. doi:10.1016/j.jhazmat.2013.04.025.

[161] L.M. Su, X. Liu, Y. Wang, J.J. Li, X.H. Wang, L.X. Sheng, Y.H. Zhao, The discrimination of excess toxicity from baseline effect: Effect of bioconcentration, Sci. Total Environ. 484 (2014) 137–145. doi:10.1016/j.scitotenv.2014.03.040.

[162] S. Kar, K. Roy, First report on interspecies quantitative correlation of ecotoxicity of pharmaceuticals, Chemosphere. 81 (2010) 738–747. doi:10.1016/j.chemosphere.2010.07.019.

[163] L.Y. Fan, D. Zhu, Y. Yang, Y. Huang, S.N. Zhang, L.C. Yan, S. Wang, Y.H. Zhao, Comparison of modes of action among different trophic levels of aquatic organisms for

pesticides and medications based on interspecies correlations and excess toxicity: Theoretical consideration, Ecotoxicol. Environ. Saf. 177 (2019) 25–31. doi:10.1016/j.ecoenv.2019.03.111.

[164] E. Benfenati, A. Lombardo, VEGAHUB for Ecotoxicological QSAR Modeling, in: 2020: pp. 759–787. doi:10.1007/978-1-0716-0150-1_30.

[165] M. Eklund, U. Norinder, S. Boyer, L. Carlsson, Choosing feature selection and learning algorithms in QSAR, J. Chem. Inf. Model. 54 (2014) 837–843. doi:10.1021/ci400573c.

[166] P.M. Khan, K. Roy, Current approaches for choosing feature selection and learning algorithms in quantitative structure–activity relationships (QSAR), Expert Opin. Drug Discov. 13 (2018) 1075–1089. doi:10.1080/17460441.2018.1542428.

[167] I. Ponzoni, V. Sebastián-Pérez, C. Requena-Triguero, C. Roca, M.J. Martínez, F. Cravero, M.F. Díaz, J.A. Páez, R.G. Arrayás, J. Adrio, N.E. Campillo, Hybridizing Feature Selection and Feature Learning Approaches in QSAR Modeling for Drug Discovery /631/114/2248 /631/154/309 /639/638/563/606 /119/118 article, Sci. Rep. 7 (2017) 1–19. doi:10.1038/s41598-017-02114-3.

[168] S.J. Cho, M.A. Hermsmeier, Genetic algorithm guided selection: Variable selection and subset selection, J. Chem. Inf. Comput. Sci. 42 (2002) 927–936. doi:10.1021/ci010247v.

[169] J.T. Leonard, K. Roy, QSAR by LFER model of HIV protease inhibitor mannitol derivatives using FA-MLR, PCRA, and PLS techniques, Bioorganic Med. Chem. 14 (2006) 1039–1046. doi:10.1016/j.bmc.2005.09.022.

[170] P.K. Ojha, K. Roy, Comparative QSARs for antimalarial endochins: Importance of descriptor-thinning and noise reduction prior to feature selection, Chemom. Intell. Lab. Syst. 109 (2011) 146–161. doi:10.1016/j.chemolab.2011.08.007.

[171] D. Newby, A.A. Freitas, T. Ghafourian, Pre-processing feature selection for improved C&RT models for oral absorption, J. Chem. Inf. Model. 53 (2013) 2730–2742. doi:10.1021/ci400378j.

[172] H. Zhu, T.M. Martin, L. Ye, A. Sedykh, D.M. Young, A. Tropsha, Quantitative structure-activity relationship modeling of rat acute toxicity by oral exposure, Chem. Res. Toxicol. 22 (2009) 1913–1921. doi:10.1021/tx900189p.

[173] T. Lei, Y. Li, Y. Song, D. Li, H. Sun, T. Hou, ADMET evaluation in drug discovery: 15. Accurate prediction of rat oral acute toxicity using relevance vector machine and consensus modeling, J. Cheminform. 8 (2016) 1–19. doi:10.1186/s13321-016-0117-7.

[174] T. Martin, User's Guide for T.E.S.T. (version 4.2), United States Environ. Prot. Agency. (2013) 63. https://www.epa.gov/chemical-research/toxicity-estimation-software-tool-test.

[175] K. Wu, G.W. Wei, Quantitative Toxicity Prediction Using Topology Based Multitask Deep Neural Networks, J. Chem. Inf. Model. 58 (2018) 520–531. doi:10.1021/acs.jcim.7b00558.

[176] D.P. Kingma, J. Ba, Adam: A Method for Stochastic Optimization, (2014) 1–15. http://arxiv.org/abs/1412.6980.

[177] A. Tropsha, P. Gramatica, V.K. Gombar, The importance of being earnest: Validation is the absolute essential for successful application and interpretation of QSPR models, QSAR Comb. Sci. 22 (2003) 69–77. doi:10.1002/qsar.200390007.

[178] K. Roy, G. Ghosh, QSTR with extended topochemical atom (ETA) indices. 12. QSAR for the toxicity of diverse aromatic compounds to Tetrahymena pyriformis using chemometric tools, Chemosphere. 77 (2009) 999–1009. doi:10.1016/j.chemosphere.2009.07.072.

[179] A.A. Toropov, A.P. Toropova, E. Benfenati, A. Manganaro, QSAR modelling of the toxicity to Tetrahymena pyriformis by balance of correlations, Mol. Divers. 14 (2010) 821–827. doi:10.1007/s11030-009-9186-0.

[180] A. Karim, A. Mishra, M.A.H. Newton, A. Sattar, Efficient Toxicity Prediction via Simple Features Using Shallow Neural Networks and Decision Trees, ACS Omega. 4 (2019) 1874–1888. doi:10.1021/acsomega.8b03173.

[181] H. Zhu, T.M. Martin, L. Ye, A. Sedykh, D.M. Young, A. Tropsha, Quantitative structure-activity relationship modeling of rat acute toxicity by oral exposure, Chem. Res. Toxicol. 22 (2009) 1913–1921. doi:10.1021/tx900189p.

[182] T. Lei, Y. Li, Y. Song, D. Li, H. Sun, T. Hou, ADMET evaluation in drug discovery: 15. Accurate prediction of rat oral acute toxicity using relevance vector machine and consensus modeling, J. Cheminform. 8 (2016) 6. doi:10.1186/s13321-016-0117-7.

[183] K. Wu, G.-W. Wei, Quantitative Toxicity Prediction Using Topology Based Multitask Deep Neural Networks, J. Chem. Inf. Model. 58 (2018) 520–531. doi:10.1021/acs.jcim.7b00558.

LIST OF PUBLICATIONS

List of papers published during the doctoral period

- **Khalid Bouhedjar**, Abdelmalek Khorief Nacereddine, Hamida Ghorab, Abdelhafid Djerourou, **QSPR Modeling For Critical Temperatures Of Organic Compounds Using Hybrid Optimal Descriptors,** Journal: International Journal of Quantitative Structure-Property Relationships, 4(4):15-26, (**2019)**DOI: 10.4018/IJQSPR.2019100102.

- Emile Roussel[1], Viet-Khoa Tran-Nguyen[1], **Khalid. Bouhedjar**[1], Mohamed AbdEsselem Dems, Amine, Belaidi, B. Matougui, B. Peres, A. Azioune, O. Renaudet, P. Falson, A. Boumendjel, **Optimization of the chromone scaffold through QSAR and docking studies: Identification of potent inhibitors of ABCG2,** European Journal of Medicinal Chemistry 184 (2019) 111772, https://doi.org/10.1016/j.ejmech.2019.111772
  1: Equally contributed to this work and are co-first authors

- **Khalid Bouhedjar**, Emilio Benfenati , Abdelmalek Khorief Nacereddine, **Modelling Quantitative Structure Activity–Activity Relationships (QSAARs): auto-pass-pass, a new approach to fill data gaps in environmental risk assessment under the REACH regulation,** SAR and QSAR in environmental research SAR and QSAR in Environmental Research Volume 31, (2020) - Issue 10 https://doi.org/10.1080/1062936X.2020.1810770

- **Khalid Bouhedjar**, Abdelbasset Boukelia, Abdelmalek Khorief Nacereddine , Anouar Boucheham, Amine Belaidi, Abdelhafid Djerourou, **A Natural Language Processing Approach based on Embedding Deep Learning from Heterogeneous Compounds for QSAR Modeling.** Volume 96, Issue3 September (2020) 961-972, Chemical Biology& Drug Design**,** DOI: 10.1111/cbdd.13742