الجمهورية الجزائرية الديمقراطية الشعبية

MINISTRY OF HIGHER EDUCATION AND SCIENTIFIC RESEARCH
UNIVERSITY YAHIA FARES MEDEA
FACULTY OF TECHNOLOGY
Department of Process and Environmental Engineering

Master's Thesis

Submitted in fulfillment of the requirements for the Degree of Master in Pharmaceutical Process Engineering

THEME

QSPR MODELING OF THE SOLUBILITY OF DRUG-LIKE COMPOUNDS IN SUPERCRITICAL CARBON DIOXIDE USING INTELLIGENT COMPUTER TECHNIQUES

Proposed & Supervised by:
Dr. Cherif SI-MOUSSA
Dr. Mabrouk HAMADACHE

Prepared & presented by,
Miss. Imane EULDJI

Promotion 2019/2020

Acknowledgments

First and foremost, I thank Allah (SWT) for having given me the strength and the courage to accomplish this modest work.
And for letting me live to see this thesis.

I'm deeply grateful to my supervisor, ***Dr. Cherif SI-MOUSSA*** *& my co-supervisor* ***Dr. Mabrouk HAMADACHE****:*

for their coaching, availability, and sense of listening and exchange.

and for providing guidance and feedback throughout this project.

A special thanks should be address to ***Pr. Othmane BENKORTBI*** *for the encouragement, and the influence that he could give, not only for me rather for all of his students.*

I would also like to extend my thanks to the members of the jury for having done us the honor and the pleasure of accepting and judging our work.

Also extend my deepest gratitude to all of our teachers for the training they gave us during our university studies.

In the end, I thank all the people who contributed, directly or indirectly, to the production of this thesis.

May the Almighty Allah richly bless all of you.



Dedication

I dedicate my dissertation work to my family and many friends.

A special feeling of gratitude to my loving parents, whose words of encouragement and push for tenacity ring in my ears. And to whom I wish them a long and happy life, may Allah protect them and give me strength so that I can give them back a little bit of their goodness and kindness.

My Brothers Rafik , and Riadh whose have never left my side and are very special.

To my dear grandfather and my dear grandmother I wish them a long life.
To the whole family

To my dear best friends Achwak, and Nassima whose I wish them all the best.

All my friends

And to all my promotion of pharmaceutical process engineering.

And finally, to all whose wish me only the best.

Iméne

Abstract

ملخص

النمذجة QSPR لمدى قابلية المركبات الشبيهة بالأدوية في ثاني أكسيد الكربون الفائق باستخدام تقنيات الكمبيوتر الذكية

تعتبر عملية تحسين قابلية ذوبان الأدوية في ثاني أكسيد الكربون فوق الحرجة واحدة من أقوى الطرق. ومع ذلك , يمكن معالجة العقبات الرئيسية لهذه التقنية في حدود التجارب المملة والمكلفة. الهدف من هذه الدراسة هو اقتراح نموذج تنبئي جديد لربط قابلية المركبات الشبيهة بالأدوية في ثاني أكسيد الكربون فوق الحرجة للذوبان. تم تطبيق نموذج QSPR لأول مرة على إجمالي عدد المركبات الصيدلانية التي تم جمعها133 , وتحويلها إلى تدوين SMILES , ثم تم حساب واصفاتها الجزيئية باستخدام برنامج Padel. تم تقليل الواصفات الجزيئية 1544 التي تم الحصول عليها إلى 11 واصفًا مهمًا: AATS3v و MATS2e و GATS4c و GATS3v و GATS4e و GATS3s و nBondsM و AVP-0 و SHBd و MLogP و MLFER_S. بالإضافة إلى هاتين الخاصيتين المستقلتين المكثفتين (درجة الحرارة والضغط) , فإن هذه الخصائص الـ 13 هي متغيرات الإدخال المستخدمة في تشكيل مجموعة البيانات النهائية لـ 3590 نقطة بيانات تجريبية. في شكل دراسة مقارنة لأربع تقنيات للتعلم الآلي: تم فحص اثنين منDFO-SVR, ANNs, ونموذج ANFIS , وقدرتهم التنبؤية / الارتباطية وجدواهم. جميع النماذج التي تم تطويرها صالحة إحصائيًا مع إجمالي AARD يتراوح بين 0.73% و 2.46% , حيث يُظهر نموذج DFO-SVR أفضل نتائج الأداء وهو الأنسب للتنبؤ بالبيانات المستقبلية ويمكن أن يكون تستخدم للتنبؤ بالذوبان في الأدوية المركبة , خاصة تلك التي لم يتم اختبارها وكذلك المركبات الجديدة.

الكلمات الرئيسية: الذوبانية ,مركبات تشبه الادوية,$SC\text{-}CO_2$, SMILES, QSPR, الواصفات ,SVR, ANNs ,ANFIS .GA, DFO,

Abstract

QSPR MODELING OF THE SOLUBILITY OF DRUG-LIKE COMPOUNDS IN SUPERCRITICAL CARBON DIOXIDE USING INTELLIGENT COMPUTER TECHNIQUES

The solubility enhancement process of drugs in supercritical carbon dioxide is considered as one of the most powerful methods. However, the major obstructs of this technique can be resumed in the limit of using tedious and expensive experiments. The objective of the current study is to propose a new predictive model to correlate the solubility of drug-like compounds in supercritical carbon dioxide. A QSPR Model was first applied to a total number of 133 pharmaceutical compounds that have been collected, converted to the SMILES notation, and then their molecular descriptors were calculated using the software Padel. The obtained 1544 molecular descriptors have been reduced to 11 significant descriptors: AATS3v, MATS2e, GATS4c, GATS3v, GATS4e, GATS3s, nBondsM, AVP-0, SHBd, MLogP, and MLFER_S. Adding to these two independent intensive state properties (temperature and pressures), those 13 features are used as input variables forming the final dataset of 3590 experimental data points. In the form of a comparative study of four machine learning techniques: two ANNs, one DFO-SVR, and one ANFIS model their predictive/correlative capacity and feasibility have been examined. The models developed are all statistically valid with a global AARD between 0.73% and 2.46%, where the DFO-SVR model shows the best performance results and is the best-fit candidate for predicting future data and can be used to predict the solubility of pharmaceutical compounds, particularly for those that have not been tested as well as new compounds.

keywords: Solubility, drug-like compounds, SC-CO_2, QSPR, SMILES, Descriptors, ANNs, SVR, ANFIS, DFO, GA.

Abstract

Résumé

MODÉLISATION QSPR DE LA SOLUBILITÉ DE COMPOSÉS DE TYPE MÉDICAMENT DANS DU DIOXYDE DE CARBONE SUPERCRITIQUE À L'AIDE DE TECHNIQUES INFORMATIQUES INTELLIGENTES

Le processus d'amélioration de la solubilité des médicaments dans le dioxyde de carbone supercritique est considéré comme l'une des méthodes les plus puissantes. Cependant, les principaux obstacles à cette technique peuvent être repris dans la limite des expériences fastidieuses et coûteuses. L'objectif de la présente étude est de proposer un nouveau modèle prédictif pour corréler la solubilité des composés de type médicamenteux dans le dioxyde de carbone supercritique. Un modèle QSPR a d'abord été appliqué à un nombre total de 133 composés pharmaceutiques qui ont été collectés, convertis en notation SMILES, puis leurs descripteurs moléculaires ont été calculés à l'aide du logiciel Padel. Les 1544 descripteurs moléculaires obtenus ont été réduits à 11 descripteurs significatifs: AATS3v, MATS2e, GATS4c, GATS3v, GATS4e, GATS3s, nBondsM, AVP-0, SHBd, MLogP et MLFER_S. En ajoutant à ces deux propriétés d'état intensif indépendantes (température et pressions), ces 13 caractéristiques sont des variables d'entrée utilisées formant l'ensemble de données final de 3590 points de données expérimentales. Sous la forme d'une étude comparative de quatre techniques d'apprentissage automatique: deux ANN, un DFO-SVR et un modèle ANFIS, leur capacité prédictive / corrélative et leur faisabilité ont été examinées. Les modèles développés sont tous statistiquement valables avec un AARD global compris entre 0,73% et 2,46%, où le modèle DFO-SVR montre les meilleurs résultats de performance et est le candidat le mieux adapté pour prédire les données futures et peut être utilisé pour prédire la solubilité des produits pharmaceutiques composés, en particulier pour ceux qui n'ont pas été testés ainsi que de nouveaux composés.

Mots-clés: Solubilité, composés de type médicamenteux, SC-CO_2, QSPR, SMILES, Descripteurs, ANNs, SVR, ANFIS, DFO, GA.

Table of Content

Table of content.....i

List of Figures.....v

List of Tables.....viii

Abbreviations.....x

Nomenclature and symbols xiii

Introduction.....1

Chapter I: Basics of Solubility and the Supercritical Fluid Technology

I.1 Introduction.....3

I.2.Solubility.....3

I.2.1. Definitions.....3

I.2.2. Unit of Solubility.....4

I.2.3. Term Insoluble.....4

I.2.4. Saturated Solutions and Solubility Product.....4

I.2.5. Factors that affect the solubility.....5

I.2.5.1. Solute-Solvent Interactions.....5

I.2.5.2. Temperature.....6

I.2.5.3. Pressure.....7

I.2.5.4. Polarity.....7

I.2.5.5. Molecular Size.....7

I.2.6. Solubility as a Challenge in Drug R&D.....7

I.2.7. The Biopharmaceutics Classification System (BCS).....8

I.2.8. Overcoming Poor Drugs Solubility Challenge Importance.....8

I.3. Supercritical Fluid Technology(SFT).....9

I.3.1. Definitions.....10

I.3.2. Properties of SCFs.....11

I.3.3. Supercritical Fluid Carbon Dioxide(SC-CO_2).....12

I.3.4. Principles of well-known SCF-based Techniques for Pharmaceutical Processing.....13

I.3.4.1. SCFs as solvents.....13

I.3.4.2. SCF as anti-solvent.....13

I.3.4.3. SCF as solute/plasticizer/co-solvent/atomizing or nebulizing agent.....14

Table of Content

I.3.4.4. Miscellaneous.....14
I.3.5. Application of SCFs in Pharmaceutical Industry.....15
I.3.5.1. Solubility Enhancement.....16
I.3.5.2. Nanoparticle Technology.....16
I.3.5.3. Product Sterilization.....16
I.3.5.4. Separation of Enantiomeric Mixtures.....16
I.3.5.5. Fractionation of Essential Oils.....17
I.3.5.6. Purification of Excipients.....17
I.3.5.7. Preparation of Pharmaceutical Powders.....17
I.3.5.8. Complex Formation & Preparation of Inclusion Compounds.....17
I.3.5.9. Drug Delivery Systems.....17
I.4. Conclusion.....18

Chapter II: Chemoinformatics & Programming tools

II.1. Introduction.....19
II.2. Basic concepts.....19
II.2.1. Database.....19
II.2.2. SMILES.....19
II.2.3. Molecular Descriptors.....20
II.2.4. Artificial Intelligence(AI) 20
II.2.5. Machine Learning(ML) 21
II.3. Quantitative structure-Activity/Propriety Relationship.....23
II.3.1. Property Prediction (QSAR/QSPR).....24
II.3.2. Model Construction(QSAR/QSPR).....25
II.4. Artificial neural network (ANN).....27
II.4.1. Artificial neuron vs. Biological neuron.....29
II.4.2. Activation Function.....30
II.4.2.1. Step function (Threshold).....31
II.4.2.2. Linear Function.....31
II.4.2.3. Non-Linear Function.....32
II.4.3. ANN Topologies.....33
II.5. Support Vector Machine (SVM).....35
II.5.1. Basic Idea of Statistical Learning Theory.....35
II.5.2. Learning and Generalization.....36

Table of Content

II.5.3. Support Vector Regression(SVR).....37
II.5.4. Kernel Functions.....40
II.6. Adaptive Neuro-Fuzzy Inference System (ANFIS).....41
II.6.1. Fuzzy systems.....43
II.6.2. Architecture and algorithm.....44
II.7. Genetic algorithm(GA).....48
II.8. Dragonfly Optimization Algorithm.....50
II.9. Conclusion.....55

Chapter III: Solubility modeling of drug-like compounds in Supercritical carbon dioxide

III.1.Introduction.....56

ARTIFICIAL NEURAL NETWORK (ANN) predictive model

A. MLP-ANN "one hidden layer" predictive Model.....56
III.2.Material and methods.....56
III.2.1.Datasets.....56
III.2.2.Descriptor calculation.....58
III.2.3.Feature Selection.....59
III.2.4.Model development.....60
III.2.5.Model validation.....61
III.2.6.Applicability domain.....64
III.3. Results and discussion.....66
III.3.1.Selection of relevant descriptors.....66
III.3.2.ANN-QSPR Modeling.....74
III.3.3.Conclusion.....81
B.MLP- ANN "two hidden layer" predictive Model.....82
III.4.Model development.....82
III.5.Statistical criteria for evaluation of the model.....84
III.6.Results and Discussion.....87
III.6.1.ANN modeling.....87
III.6.2.Conclusion.....98

Table of Content

DFO-SVR predictive model

III.7.Model development.....99
III.8.Results and Discussion.....100
III.8.1.DFO-SVR modeling.....100
III.8.2.Conclusion.....105

ANFIS predictive model

III.9.Model development.....106
III.10.Results and Discussion.....107
III.10.1.ANFIS modeling.....107
III.10.2.Conclusion.....112

Model vs. Model

III.11. Evaluation and performance comparison of the optimized models.....112
III.12.Conclusion.....115

Conclusion.....117
Bibliography.....119
Annex.....131

List of Figures

Figure	Figures Name	Page
Chapter I		
Fig. I.1	Pressure Temperature phase diagram for a CO_2.	9
Fig. I.2	The phases of SCFs.	10
Fig. I.3	Pressure-Temperature Phase Diagram for carbon dioxide.	12
Fig. I.4	Process diagram of main SCF-based techniques.	15
Chapter II		
Fig. II.1	AI is the overall category that includes machine learning and natural language processing.	21
Fig. II.2	Representation of the relationship between AI,ML and DL.	21
Fig. II.3	Flow chart of machine learning and the algorithms used for training the machine.	23
Fig. II.4	The basic approach to QSAR/QSPR/QSTR.	24
Fig. II.5	Schematic representation of QSAR/QSPR/QSTR.	26
Fig. II.6	Working principle of an artificial neuron.	28
Fig. II.7	Example of simple artificial neural network.	29
Fig. II.8	Biological and artificial neuron design.	29
Fig. II.9	Activation function type Threshold.	31
Fig. II.10	Activation function type Linear.	31
Fig. II.11	Activation function type Non-Linear (Sigmoid).	32
Fig. II.12	Activation function type Non-Linear (Piecewise Linear).	32
Fig. II.13	Activation function type Non-Linear (Gaussian).	33
Fig. II.14	Feed-forward (FNN) and recurrent (RNN) topology of an ANN.	33
Fig. II.15	Number of Epochs Vs Complexity.	36
Fig. II.16	Regression with the ε-insensitive tube.	39
Fig. II.17	Effect of kernel transformation.	41
Fig. II.18	The typical structure of a fuzzy system.	43
Fig. II.19	Schematic structure of the TSK fuzzy model.	45
Fig. II.20	The ANFIS architecture with two input parameters x and y.	46

Fig. II.21	A simple flow sheet of GA.	49
Fig. II.22	Dynamic dragonfly swarms.	51
Fig. II.23	Static dragonfly swarms.	51
Fig. II.24	Primitive corrective patterns between individuals in a swarm.	52
Chapter III		
Fig. III.1	Normality distribution plot of mole fraction of solubility data: (a) before transformation.	57
Fig. III.2	Normality distribution plot of mole fraction of solubility data: (b) after conversion (-Ln(y2)).	58
Fig. III.3	The dispersion diagram of the molar solubility fraction, observed vs. predicted for both the training and test set. model BFGS, tanh.	74
Fig. III.4	Williams plot for MLP-ANN (one hidden layer) model.	76
Fig. III.5	The dispersion diagram of the molar solubility fraction, for the validation set, model BFGS, tanh.	77
Fig. III.6	Residuals vs. Observed for the validation Data of the selected 12 descriptors (model 12).	78
Fig. III.7	Residuals vs. Observed for the validation Data of the selected 13 descriptors (model 14b).	78
Fig. III.8	Residuals vs. Observed for the validation Data of the selected 14 descriptors (model 14).	79
Fig. III.9	The dispersion diagram of the molar solubility fraction, observed vs. predicted for the train, test, and global set. Model (A).	90
Fig. III.10	The dispersion diagram of the molar solubility fraction, observed vs. predicted for the train, test, and global set. Model (B).	91
Fig. III.11	The dispersion diagram of the molar solubility fraction, observed vs. predicted for the train, test, and global set. Model (C).	92
Fig. III.12	The relative error vs. the solubility for model (A).	93
Fig. III.13	The relative error vs. the solubility for model (B).	93
Fig. III.14	The relative error vs. the solubility for model (C).	93
Fig. III.15	Bar distribution diagrams of the Absolute Relative Error [%] vs. number of DP in %, model (A).	94
Fig. III.16	Bar distribution diagrams of the Absolute Relative error [%] vs. number of DP in %, model (B).	94

Fig. III.17	Bar distribution diagrams of the Absolute Relative Error [%] vs. number of DP in %, model (C).	94
Fig. III.18	Williams plot for MLP-ANN, two hidden layer, model (A).	95
Fig. III.19	Williams plot for MLP-ANN, two hidden layer, model (B).	95
Fig. III.20	Williams plot for MLP-ANN, two hidden layer, model (C).	96
Fig. III.21	Relative contribution of each input variable, the Garson method, model (A).	97
Fig. III.22	Relative contribution of each input variable, the Garson method, model (B).	97
Fig. III.23	Relative contribution of each input variable, the Garson method, model (C).	98
Fig. III.24	The dispersion diagram of the molar solubility fraction, observed vs. predicted for the train, test, and global set. Model DFO-SVR.	102
Fig. III.25	The relative error vs. the solubility for DFO-SVR model.	103
Fig. III.26	Bar distribution diagrams of the Absolute Relative Error [%] vs. number of DP in %, model DFO-SVR.	103
Fig. III.27	Williams plot for DFO-SVR model.	104
Fig. III.28	Results of the sensitivity analysis performed on the DFO-SVR model.	105
Fig. III.29	The dispersion diagram of the molar solubility fraction, observed vs. predicted for the train, test, validation, and global set, model ANFIS.	109
Fig. III.30	The relative error vs. the solubility for model ANFIS.	110
Fig. III.31	Bar distribution diagrams of the Absolute Relative Error [%] vs. number of DP in %, model ANFIS.	110
Fig. III.32	Williams plot for ANFIS model.	111
Fig. III.33	Results of the sensitivity analysis performed on the ANFIS model.	112
Fig. III.34	Comparison of AARD% global.	113
Fig. III.35	Bar distribution diagrams of the Absolute Relative Error [%] vs. number of DP in%, training set.	113
Fig. III.36	Bar distribution diagrams of the Absolute Relative Error [%] vs. number of DP in %, test set.	114

List of Tables

Tables	Tables Name	Page
Chapter I		
Tab. I.1	Some Properties of Supercritical Fluids vs. Gases and Liquids.	10
Chapter III		
Tab. III.1	Acceptability criteria (A.C) of model.	64
Tab. III.2	Statistical parameters for internal validation.	67
Tab. III.3	Statistical parameters for external validation.	68
Tab. III.4	Statistical of multi-collinearity of 13 descriptors.	69
Tab. III.5	Statistical of multi-collinearity of 14 descriptors.	69
Tab. III.6	Statistical of multi-collinearity of 15 descriptors.	70
Tab. III.7	Statistical of multi-collinearity of 17 descriptors.	70
Tab. III.8	Statistical of multi-collinearity of 13 descriptors(14b).	71
Tab. III.9	Matrices of correlation of the selected features.	72
Tab. III.10	Characteristics of the descriptors selected in the optimal model.	73
Tab. III.11	Validation parameters for the train, test, and global sets for MLP-ANN(one hidden layer) model.	75
Tab. III.12	Different errors for the train, test, and global sets for MLP-ANN(one hidden layer) model.	75
Tab. III.13	Type and contribution of the descriptors selected in the ANN model.	80
Tab. III.14	Representation of the selected Activation function.	83
Tab. III.15	Representation of the three selected MLP-ANNs with two hidden layers models.	88
Tab. III.16	Validation parameters for the train, test, and global sets for the three MLP-ANN (two hidden layer) models.	88
Tab. III.17	Different errors for the train, test, and global sets for the three MLP-ANN (two hidden layer) models.	89
Tab. III.18	Optimized parameters of the DFO-SVR model.	100
Tab. III.19	Validation parameters for the train, test, and global sets for DFO-SVR model.	101
Tab. III.20	Different errors for the train, test, and global sets for DFO-SVR model.	101
Tab. III.21	Details summary of the fuzzy inference system (FIS) learning parameters.	107

List of Tables

Tab. III.22	Validation parameters for the train, test, and global sets for the ANFIS model.	108
Tab. III.23	Different errors for the train, test, and global sets for the ANFIS models.	108
Tab. III.24	Comparison of the differences between the AARD% of the test set and the training set for the selected applied models.	114

Abbreviations

Abbreviation	Signification
A.C	Acceptability Criteria.
AD	Applicability Domain.
AI	Artificial Intelligence.
AKA	Also Known As.
ANFIS	Adaptive Neuro-Fuzzy Inference System.
ANN	Artificial Neural Network.
API	Active Pharmaceutical Ingredients.
BCS	Biopharmaceutics Classification System.
Co_2	Carbon Dioxide.
CPCSP	Continuous Powder Coating Spraying Process.
DELOS	Depressurization Of An Expanded Liquid Organic Solution.
DFO	Dragonfly Optimization.
DFO-SVR	Dragonfly Optimization- Support Vector Regression.
DL	Deep Learning.
EDP	Experimental Data Points.
FIS	Fuzzy Inference System.
GA	Genetic Algorithm.
GAMA	Gas Assisted Melting Atomization.
GA-MLR	Genetic Algorithm-Multiple Linear Regression.
GAS	Gas Anti-Solvent.
GRAS	Generally Regarded As Safe.
i.e.	Exempli Gratia "For Example".
ISCRPE	Improved Supercritical Reverse Phase Evaporation.
IUPAC	International Union Of Pure And Applied Chemistry.
MFs	Membership Functions.
ML	Machine Learning.
MLP	Multi Layer Perceptrons.

Abbreviations

MLP-ANN	Multi Layer Perceptrons-Artificial Neural Networks.
MLR	Multiple Linear Regression.
NB	Nota Bene "Note Well".
OECD	Organization For Economic Co-Operation And Development.
Pc	Critical Pressure.
PGSS	Particles From Gas Saturated Solutions.
PLS	Partial Least-Squares.
PPRGEL	Precipitation By Pressure Reduction Of Gas-Expanded Liquids.
QSAR	Quantitative Structure-Activity Relationship.
QSPR	Quantitative Structure-Property Relationship.
R&D	Research And Development.
RBF-ANN	Radial Basis Function Artificial Neural Networks.
RESAS	Rapid Expansion From Supercritical Solution To Aqueous Solution.
RESOLV	Rapid Expansion Of A Supercritical Solution Into A Liquid Solvent.
RESS	Rapid Expansion Of Supercritical Solutions.
RESS-N	Rapid Expansion From Supercritical Solution With A Non-Solvent.
RESS-SC	Rapid Expansion Of Supercritical Solutions With A Solid Co-Solvent.
SAA	Supercritical-Assisted Atomization.
SAS	Supercritical Anti-Solvent.
SAS-EM	Supercritical Anti-Solvent Precipitation With Enhanced Mass Transfer.
SCF	Supercritical Fluid.
SC-CO_2	Supercritical Fluid Carbon Dioxide.
SCRPE	Supercritical Reverse Phase Evaporation.
SEDS	Solution Enhanced Dispersion By Supercritical Fluids.
SEE-C	Continuous Supercritical Emulsions Extraction.
SFE	Supercritical Fluid Extraction.
SFEE	Supercritical Fluid Extraction Of Emulsion.
SFT	Supercritical Fluid Technology.
SMILES	Simplified Molecular Input Line Entry Specification.
SVC	Support Vector Classification.

Abbreviations

SVM	Support Vector Machine.
SVR	Support Vector Regression.
Tc	Critical Temperature.
TSK model	Takagi-Sugeno-Kang Model.
US-FDA	United States Food And Drug Administration.
VIF	Variation Inflation Factors.
Viz.	used, especially in written English, when you want to give more detail or be more exact about something you have just written.

Nomenclature and Symbol

Symbol	Designation	Unity
w_i	Weight	[-]
b_i	Bias	[-]
Kc	Constant of ionic equilibrium	[-]
Ks	solubility product	[-]
M	Molarity	mol/L
m	Molality	mol/Kg
n	Number of compounds in the data set	[-]
P	Pressure	Mpa
S	Mass per volume	g/m^3
T	Temperature	Kelvin
W	Mass per mass	g/Kg
w	Mass fraction	[-]
y	Mole fraction	[-]
y_2	Molar fraction	mol/mol
Yobs	Observed value of Y.	[-]
Ypred	Predicted Y-value.	[-]
f	fittest function	[-]

Introduction

Introduction

From idea to drug; the development of new drugs is a long and complicated process [1]. This process well-known started with an idea studied in a stage called Research and Development (R&D) and ended with the commercialization of a drug, where it usually takes from 10 to 20 years. However, very few ideas make it that far due generally to the influence of physicochemical parameters. In recent years, researchers have shown an increased interest in the impact of physicochemical and molecule proprieties on both pharmacokinetic and pharmacodynamic processes, as well as drug safety, often in a conflicting way [2]. Where the results obtained of those studies had set on a list of the top five keys of critical compound properties in early compound screening, viz., solubility, dissociation constant, permeability, stability, and lipophilicity; On the top of this list we have the solubility [3].

Solubility, the phenomenon of dissolution of a solute in a solvent to give a homogenous system [3], plays a key role in the formulation development to achieve the bioavailability and therapeutic action of the drug at the target site [4]. This can be explained as a drug with low aqueous solubility that can represent the major problem encountered with the formulation development of new chemical entities as well as for the generic development. It also is known as a drug with a low bioavailability [3, 4].

According to the new studies, almost 70% of the new drug candidates had faced numerous problems in the formulation and development stage because of showing poor aqueous solubility [5]. This issue has changed upside down the whole focus of researchers, from inventing new drugs into facing the low drug solubility challenge of existing drugs. Usually, solid drugs for oral administration. Poor aqueous solubility and bioavailability of Active Pharmaceutical Ingredients (APIs) represent major bottlenecks that the pharmaceutical industry has struggled to overcome over the last decades. Some of the most successful or promising approaches for the enhancement of the bioavailability of APIs include the production of more soluble solid-state forms (e.g. amorphous solid dispersions, cocrystals), inclusions complexes and solid nanoparticles/nanosuspensions. Therefore, there is a need for pharmaceutical companies to successfully overcome engineering and manufacturing hurdles associated with these issues.

Supercritical Fluid (SCF) Technology is one such technique that has become an important tool in the pharmaceutical industry in the last two decades [6]. Thanks to being safe, environmentally friendly and economical; this technique has been suggested as one of the powerful contributions for drug solubility enhancement techniques, nanosizing, and solubilization (particle size reduction). The most widely supercritical fluid used is carbon dioxide (SC-CO_2) thanks to its many advantages; low critical Temperature (T_c = 31.10C) and critical Pressure (P_c = 73.8 bar), non-

toxic, non-flammable, and inexpensive. The low critical temperature of CO_2 makes it attractive for processing heat-labile molecules [7].

With regards to the application of SCF technology, the knowledge of the solubility of drug-like molecules in SC-CO_2 is both fundamental and necessary for the design and operation of the processes based on this technology. Experimental measurements of any property is obviously the best way to access the values for this property, but these are unlikely to be available due to the huge number of compounds, temperature and pressures ranges, and mixing compositions for which data are required. It is therefore clear that the experimental way to determine the solubility of all compounds for all operation conditions is not only expensive and laborious but also impossible. Consequently, the modeling for the estimation of this property for correlation and/or prediction purposes has been and still remains interesting from the academic and industrial point of view.

Reviewing the literature indicates that the straightforward way to model the solubility of solid drug-like molecules in SC-CO_2 are the equations of state. Empirical density-based models are also used for their simplicity. However both approaches are generally quiet diverse and not accurate enough. Techniques that relate Activity, Property, and Toxicity to the molecular structure are now well established modeling tool to predict or correlate such quantities under the acronyms QSAR, QSPR and QSTR respectively. Furthermore, computational intelligence techniques based on machine learning and artificial intelligence based optimization techniques have been successfully applied to extract knowledge from big data. The combination of these two mentioned concepts in chemistry is known as chemoinformatics.

The aim of this work is to apply QSPR concepts with computational intelligence techniques to a database of experimental measurements of solubility of drug-like compounds in SC-CO_2 in order to develop and validate models that are able to predict and/or correlate this property with an acceptable level of accuracy.

This work is made up of three chapters:

❖ The first chapter gives an overview on the parameter solubility and on the supercritical fluids technology.

❖ Then, the second would give a general description of the knowledge necessary that a developer should know to model a property similar to the one we chose in this work, the solubility, with the little information available only on the tools and methods used in this dissertation. The focus will be on QSAR/QSPR concept and machine learning methods.

❖ The third chapter is devoted to presenting the results and discussions.

❖ Finally, this dissertation will end with general concluding remarks and the possible perspectives of this work.

Chapter I:

Basics of Solubility and the Supercritical Fluid Technology

Chapter I: Basics of Solubility and Supercritical Fluid Technology

I.1.Introduction

This chapter is devised into two parts; the main focus of the first part would be about giving a general representation of the solubility as a phase equilibrium parameter, describing its relationship with the bioavailability of drugs. The second part would include an appraisal of the most renowned high-pressure techniques so far to obtain products with better performances, well-known as supercritical fluid technology.

I.2. Solubility

I.2.1. Definitions

So far, there has been little disagreement on the definition of solubility which can be explained with the existence of several definitions that have been proposed:

- The quality of being able to be dissolved, or the degree to which something can be dissolved [8].
- Amount of a substance, called the solute, which dissolves in a unit volume of a liquid substance, called the solvent, to form a saturated solution under specified conditions of temperature and pressure [9].
- It is a chemical property referring to the ability for a given substance, the solute, to dissolve in a solvent [10].
- The maximum quantity of solute that can dissolve in a certain quantity of solvent or quantity of solution at a specified temperature or pressure, in the case of gaseous solutes [11].

In familiar cases, a solute is a solid and a solvent is a liquid, but the solute or solvent might be a gas, liquid, or solid. The solvent can either be a pure substance or a mixture.

I.2.2. Unit of Solubility

The International Union of Pure and Applied Chemistry (IUPAC) define solubility in terms of a proportion of solute to solvent. Allowable units of concentration include:

- Molarity (M, mol/l),
- Molality (m, mol/Kg),
- Mass per volume (S, g/m^3),
- Mass per mass (W, g/Kg),
- Mole fraction (y),
- Mass fraction (w),

I.2.3.Term Insoluble

Solubility varies over an extended range from infinitely soluble such as ethanol in water to poorly soluble such as silver chloride in water [3].

The term insoluble is often applied to poorly soluble compounds [10]; in very few cases is it true that no solute dissolves. Generally, an insoluble solute still dissolves a little. While there is no hard-and-fast limit that defines a substance as insoluble, it is common to apply a threshold where a solute is insoluble if less than 0.1 grams dissolves per 100 milliliters of solvent [12].

We distinguish the substances [13]:

- Called insoluble: S<1g/l,
- Moderately soluble: 1g/l <S<10 g/l, and
- Soluble: S>1g/l.

I.2.4. Saturated Solutions and Solubility Product

For a specific solvent at a fixed temperature, a solute is added gradually in increasing amounts until a stage where no matter how long we wait or how vigorously we stir, some of the solutes remain undissolved. The solution obtained said to be saturated [14].

In case the solute is an electrolyte, its ionization occurs in solution and the degree of dissociation depends on the concentration of dissolved electrolytes at a particular temperature [14].

Consider, in general, the electrolyte of the type $AxBy$ which is dissociated as:

$$AxBy \leftrightarrow xA^{y+} + yB^{x-} \quad \text{I.1}$$

Applying the law of action to the **ionic equilibrium**,

$$Kc = \frac{[A^{y+}]^x[B^{x-}]^y}{[AxBy]} \quad \text{I.2}$$

When the solution is saturated,

$$[AxBy] = K' \text{ (constant)} \quad \text{I.3}$$

Or

$$[A^{y+}]^x[B^{x-}]^y = Kc.[AxBy] = KcK' = Ks\ (constnat) \quad \text{I.4}$$

Ks is termed as the **solubility product**.

Thus, **solubility product** is defined as the product of concentrations of the ions raised to a power equal to the number of times the ions occur in the equation representing the dissociation of the electrolyte at a given temperature when the solution is saturated [13,14].

I.2.5. Factors that affect the solubility

Solubility is a thermodynamic parameter, which can be influenced by the presence of other chemical species in a solution, the phases of the solute and solvent, temperature, pressure, solute particle size, and polarity...while, for the same solvent-solute, the two main affected factors are the temperature and the pressure.

I.2.5.1. Solute-Solvent Interactions

The relation between the solute and solvent is very important in determining solubility. Strong solute-solvent attractions equate to greater solubility while weak solute-solvent attractions equate to lesser solubility [17].

I.2.5.2. Temperature

Temperature changes affect the solubility of solids, liquids, and gases differently. However, those effects are finitely determined only for solids and gases [17].

- **Solids**

The effects of temperature on the solubility of solids differ depending on whether the reaction is endothermic or exothermic. Using Le Châtelier's principle, the effects of temperature in both scenarios can be determined.

First, consider an endothermic reaction, increasing the temperature results in a stress on the reactants side from the additional heat. Le Châtelier's principle predicts that the system shifts toward the product side in order to alleviate this stress. By shifting towards the product side, more of the solid is dissociated when equilibrium is again established, resulting in increased solubility.

Second, consider an exothermic reaction, increasing the temperature results in stress on the product's side from the additional heat. Le Châtelier's principle predicts that the system shifts toward the reactant side in order to alleviate this stress. By shifting towards the reactant's side, less of the solid is dissociated when equilibrium is again established, resulting in decreased solubility.

- **Gases**

In understanding the effects of temperature on the solubility of gases, it is first important to remember that temperature is a measure of the average kinetic energy. As the temperature increases, kinetic energy increases. The greater kinetic energy results in greater molecular motion of the gas particles.

As a result, the gas particles dissolved in the liquid are more likely to escape to the gas phase and the existing gas particles are less likely to be dissolved. The converse is true as well.

The trend is thus as follows: increased temperatures mean lesser solubility and decreased temperatures mean higher solubility.

- **Liquids**

In the case of liquids, there is no defined trends for the effects of temperature on the solubility of liquids.

I.2.5.3. Pressure

For solids and liquids, known as condensed phases, the pressure dependence of solubility is typically weak and is usually neglected in practice. However, the solubility of gases shows significant variability based on pressure.

Typically, gas will increase in solubility with an increase in pressure. This effect can be mathematically described using an equation called Henry's law [16].

Henry's law dictates that when the temperature is constant, the solubility of the gas corresponds to its partial pressure. Consider the following formula of Henry's law[17]:

$$p = k_h c \qquad \text{I.5}$$

Where,

p : The partial pressure of the gas above the liquid,

k_h : Henry's law constant, and

c : The concentration of the gas in the liquid.

I.2.5.4. Polarity

In most cases, solutes dissolve in solvents that have a similar polarity. Chemists use a popular aphorism to describe this feature of solutes and solvents: **"Like dissolves like"**.
Non-polar solutes do not dissolve in polar solvents and the other way round [15].

I.2.5.5. Molecular Size

The larger the molecules of the solute are the larger is their molecular weight and their size. It is more difficult for solvent molecules to surround bigger molecules. If all of the above-mentioned factors are excluded, a general rule can be found that larger particles are generally less soluble. If the pressure and temperature are the same as out of two solutes of the same polarity, the one with smaller particles is usually more soluble [15].

I.2.6. Solubility as a Challenge in Drug R&D

In the last several decades, drug research and development have faced a number of issues related to the pharmacokinetics of new chemical entities. Numerous new drug candidates have been reported to have failed in human clinical phases (mainly phase I) after oral administration due to

ineffectiveness caused by unsatisfactory pharmacokinetics. The same problem is also occurring during preclinical research and development. It has been clearly demonstrated that solubility is one of the major factors affecting the *in vivo* systemic exposure of drug candidates [19].

The number of such drug candidates has increased enormously and almost 70% of the new drug candidates have shown poor aqueous solubility. Since approximately 65% of the human body is made up of water, a drug must have certain water solubility and possess an acceptable bioavailability level. Overall, the poor bioavailability of a drug substance might result thereby leading to insufficient clinical outcomes. Therefore, the poor water solubility of many drugs is one of the major obstacles in the development of highly potent pharmaceutics [20].

I.2.7. The Biopharmaceutics Classification System (BCS)

The BCS is the scientific framework for classifying drug substances based on their aqueous solubility and intestinal permeability. It is a drug development tool that allows estimation of the contributions of three major factors, dissolution, solubility, and intestinal permeability that affect the oral absorption of drugs [3].

According to the BCS, drug substances are classified as follows:

- **Class I:** High permeability and solubility,
- **Class II:** High permeability and low solubility,
- **Class III:** Low permeability and high solubility, and
- **Class IV:** Low permeability and low solubility [4].

Those that belong to classes III and IV are most likely to return to the lead optimization phase for the improvement of their physicochemical properties also provide a number of challenges for formulation scientists working on the oral delivery of drugs [18].

I.2.8. Overcoming Poor Drugs Solubility Challenge Importance

A drug with low aqueous solubility, tend to be eliminated from the gastrointestinal tract before they get the opportunity to fully dissolve and be absorbed into the blood circulation. Also, the reason behind a low bioavailability is poor dose proportionality, which greatly hinders their clinical translations. In such cases, dose augmentation would be necessary to ensure that the drug attains the therapeutic concentration range in blood.

After oral administration, this dose augmentation at times causes topical toxicity in the gastrointestinal tract and such toxicity results in a decline in patient compliance. On the other hand,

consuming a large amount of Active pharmaceutical Ingredients (API) would raise the manufacturing cost of developing and manufacturing the drug product.

In short, these poorly water-soluble drugs shown a number of negative clinical effects including potentially serious issues of inter-patient variability, higher patient costs, inefficient treatment, and more importantly, increased risks of toxicity or even death [3, 4, 20].

Thus far, this issue has changed a large concept when it comes to research. Since it cannot be dismissed, it should be then reduced or even removed. Herewith, various technologies have been developed to overcome this kind of problems such as reducing size, adding co-solvents and a lot more... All have been used for one purpose but with different contributions. We distinguish, Supercritical Fluid Technology.

This technology precipitates micro- as well as nano-sized particles with large surface areas, controlled uniform particle sizes, and smooth surfaces which eventually augment drug bioavailability [27]. AKA, drug solubility.

I.3. Supercritical Fluid Technology (SFT)

Many physicochemical properties can be used; the boiling point, dew point, density and, so on ... All for characterizing a substance or mixture as a solid, liquid or, gas [21]. However, at high enough temperatures the kinetic energy of the molecules is high enough to overcome any intermolecular forces that would condense the sample into the liquid phase. On the other hand, high enough pressures would not allow a sample to stay in the pure gaseous state [22].

This single phase is termed a supercritical fluid (SCF) and is characterized by a critical temperature and pressure (Tc and Pc).

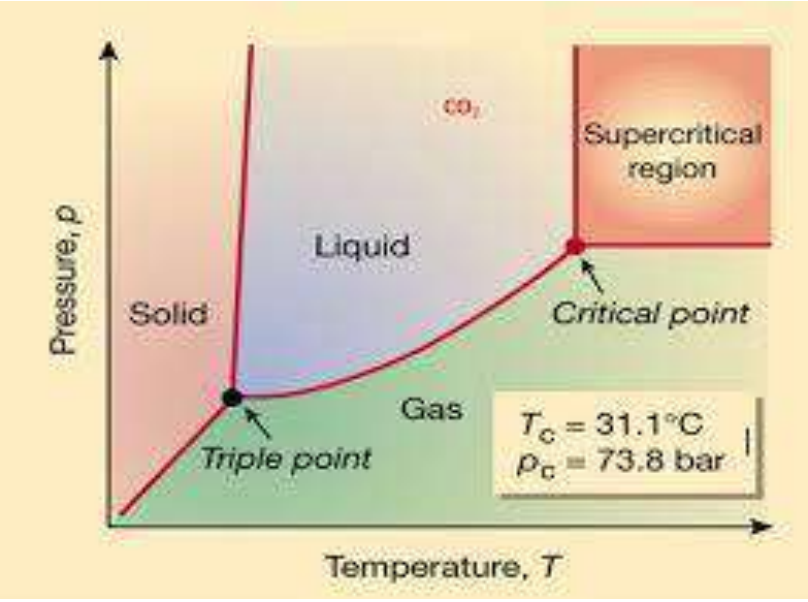


Figure .I.1. Pressure Temperature phase diagram for a CO_2.

The figure **.I.1** highlights the critical point, where a substance does not exist in either the liquid or gas phase. Under those conditions, it is called a "supercritical fluid", and has properties between those of a liquid and a gas.

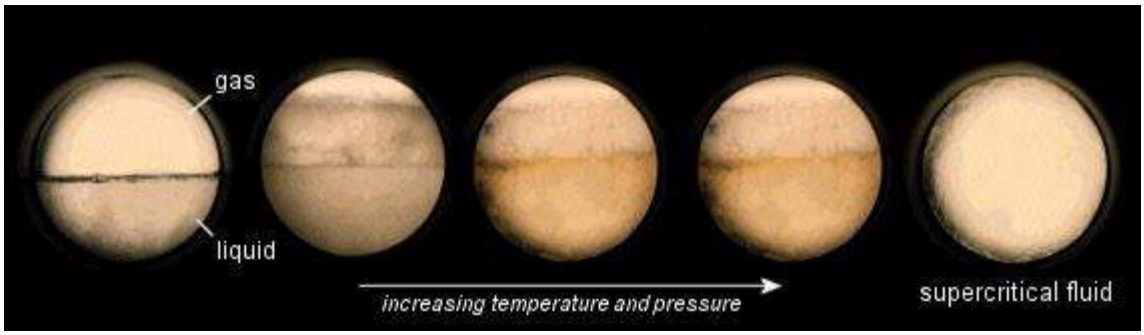


Figure .I.2. The phases of SCFs.

This figure **.I.2** shows how the increase of temperature and the pressure decrease the boundary between the liquid phase and gas phase. Until the moment T=Tc and P=Pc we observe boundary disappears (liquid and gas phases) creating one single-phase "phase supercritical". In fact, a SCF is an intermediate phase between vapor and liquid phases.

I.3.1.Definitions

SCT is a distinct concept that utilizes the solvent properties of supercritical fluids (SCFs) above its critical temperature and pressure conditions [23]. This technology was particularly motivated by environmental concern, for the capability of some SCFs to replacing finally toxic industrial solvent [6]. It is also well-known as a green and novel technology for various processes.

SCFs are substances with both gas- and liquid-like properties [21]. It is gas-like in that it is a compressible fluid that fills its container, and is liquid-like in that it has comparable densities and solvating power [24].

In other words, an SCF can behave as either a liquid or a gas but is actually neither [6]. Supercritical behavior only occurs when the substance is above its critical temperature and pressure [24].

Table .I.1. Some Properties of Supercritical Fluids vs. Gases and Liquids [21].

	Gas	SCFs	Liquid
Density (g/cm^3)	10^{-3}	0.1-1	1
Diffusion coefficient(cm^2/s)	10^{-1}	10^{-3}-10^{-4}	$<10^{-5}$
Viscosity(g/cm.s)	10^{-4}	10^{-3}-10^{-4}	10^{-2}

I.3.2.Properties of SCFs

SCFs have been studied extensively for the past two decades in attempts to gain accurate and detailed knowledge of their fundamental properties. Such knowledge is essential to the utilization and optimization of SFT in materials preparation and processing [25].

Almost all gases can be transformed into SCFs above their specific critical pressure and temperature values. In some cases, extremely high temperatures and pressures are required to attain the SCF state [23]. In addition, the Pc, Tc values increase with the molecular weight or intermolecular hydrogen bonding or polarity [6].

An SCF may be considered macroscopically homogeneous but microscopically inhomogeneous, consisting of clusters of solvent molecules and free volumes [25].

In general terms, SCFs have properties between those of a gas and a liquid [22]; the density of an SCF is comparable to that of liquids, while its diffusivity and viscosity are comparable to those of gases [23]. This Janus-faced nature of SCFs arises from the fact that the gaseous and liquid phases merge together and become indistinguishable at the critical point [25].

In addition, there is no surface tension in a supercritical fluid, as there is no liquid to gas phase boundary. By changing the pressure and temperature of the fluid, the properties can be "tuned" to be more liquid or gas-like [22]. Table I.1 illustrate how some physical properties of supercritical fluids compare to gases and liquids,

The SCF has a unique thermo-physical property. As the pressure is raised, the density of the gas increases without a significant increase in viscosity while the ability of the fluid to dissolve compounds also increases [6].

The most interesting aspects of SCFs are that their physical properties are very sensitive to temperature and pressure and there exists density in-homogeneity in the critical region [25].

Not all properties of SCFs are intermediate between those of gases and liquids; compressibility and heat capacity, for example, are significantly higher near the critical point than they are in liquids or gases [25].

Therefore, one of the most important properties of SCFs is their ability to act as solvents [22], which is altered by careful control of changes in temperature and/or pressure [6].

All supercritical fluids are completely miscible with each other; therefore a single phase for a mixture can be guaranteed if the critical point is exceeded [22].

I.3.3. Supercritical Carbon Dioxide (SC-CO_2)

Among the number of gases tried as SCF, carbon dioxide is considered as the best option for the SFT because of its low critical point (31.3 ºc, 7.4 Mpa), whose critical parameters are simple to be obtained in an industrial apparatus[25]. Attractiveness for heat-sensitive materials, it is inert, leaves no traces behind after the process, as well as being inexpensive, non-inflammable, having GRAS (generally regarded as safe) status, and being easy to recycle or to dispose of [6].

One of the distinguishing properties of supercritical carbon dioxide is its small and linear structure which increases its permeability. Nevertheless, supercritical carbon dioxide has some limitations, such as non-polarity and its capacity to form specific solvent-solute interactions. Therefore, to enhance its solubilizing power, it is highly desired to improve its polarity by adding a suitable co-solvent [74].

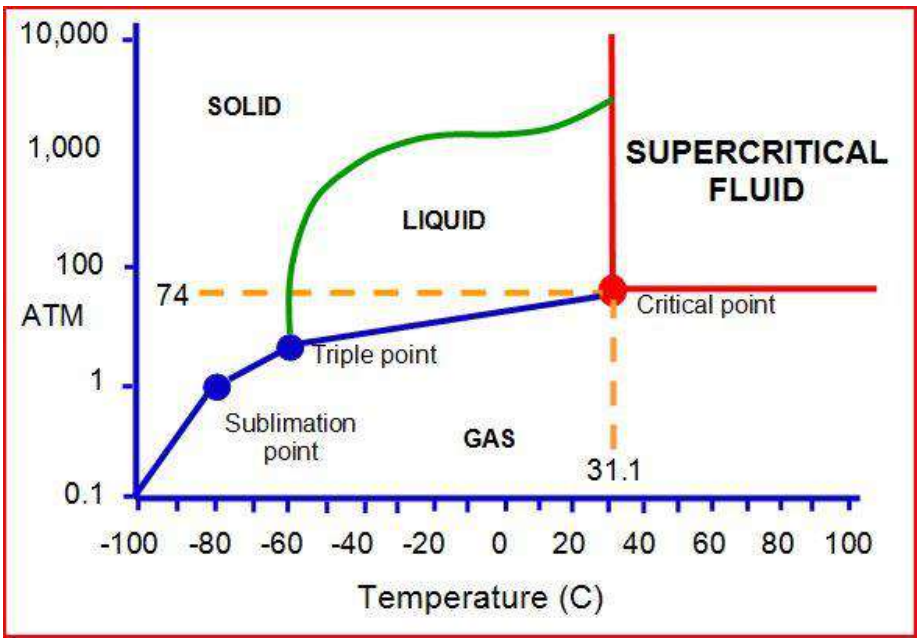


Figure .I.3. Pressure-Temperature Phase Diagram for carbon dioxide.

This diagram indicates the four regions: solid, liquid, gas and supercritical fluid region of CO_2.

Another benefit of using SC-CO_2 is that carbon dioxide is gaseous under ambient conditions of pressure and temperature, thus enabling the end-product to be separated spontaneously from CO_2 simply by depressurizing. This means that SC-CO_2 processes require a

reduced number of unit operations compared to conventional methods which need several successive steps of drying and separating in order to recover a solvent-free end-product [46].

Furthermore, one is more likely to accurately control product characteristics using SCFs because of a large number of operating parameters and conditions (pressure, temperature, nature of the phases in presence, compositions, flow rates, introduction modes, mixing conditions,…) which can be varied in order to obtain an end-product with specific properties. Polymorphic screening, for example, may be more successful with SCFs than with conventional methods [46].

Moreover, other SCFs include water and solvents such as acetone, CO2/ethanol (EtOH) mixture, chlorodifluoromethane, diethyl ether, nitrous oxide, propane, and trifluoromethane operated at their respective supercritical conditions [27].

I.3.4. Principles of well-known SCF-based Techniques for Pharmaceutical Processing

Down below a short description of the principles and common modifications of the most used SCF-based processes is shown [75]:

I.3.4.1. SCFs as solvents

The Rapid Expansion of Supercritical Solutions (RESS) explores the significant modification of the solvent power of SCFs with changes in pressure. It operates in a batch mode following two steps: First, dissolution of the solute in SC-CO_2; Second, sudden depressurization of the solution through a heated nozzle/capillary, as shown in **Figure .I.4**.

The solvent power of SC-CO_2 is reduced, leading to the super-saturation of the solution and instantaneous precipitation of the solute in small particles. The main advantages of the RESS process are its simplicity and the absence of organic solvents.

Some of these limitations have been addressed by other similar processes such as RESOLV (Rapid Expansion of a Supercritical Solution into a Liquid Solvent), RESAS (Rapid Expansion from Supercritical Solution to Aqueous Solution), RESS-N (Rapid Expansion from Supercritical Solution with a Non-Solvent), and RESS-SC (Rapid Expansion of Supercritical Solutions with a Solid Co-solvent).

I.3.4.2. SCF as anti-solvent

Anti-solvent methods are based on the high solubility and diffusivity of SCFs in organic solvents. The solute is usually dissolved in an organic solvent and then the solution is put in contact with the SC-CO_2. The diffusion of SC-CO_2 into the liquid phase decreases the solvation power of

the organic solvent, leading to the precipitation of the solute. A final washing step with pure SC-CO_2 needs to be carried out to remove residual solvent from the powder.

Therefore, the basic requirements to allow precipitation via anti-solvent techniques are: complete miscibility between solvent and anti-solvent (CO_2) at the operating conditions and the mixture solvent/anti-solvent must not solubilize the solute.

Different anti-solvent processes follow the same principles but with distinct ways of mixing the organic solution and the supercritical fluid. Like the Gas Anti-Solvent (GAS) process, the Supercritical Anti-Solvent (SAS) (**Figure .I.4.**). This process is also known as the Aerosol Solvent Extraction System (ASES) or Precipitation by Compressed Anti-solvent, The Solution Enhanced Dispersion by Supercritical Fluids (SEDS), and the Supercritical Anti-Solvent Precipitation with Enhanced Mass Transfer (SAS-EM).

I.3.4.3. SCF as solute/plasticizer/co-solvent/atomizing or nebulizing agent

Supercritical or dense CO_2 can play other roles in particle generation including co-solute, co-solvent, dispersing agent, and plasticizer. In the Particles from Gas Saturated Solutions (PGSS), SC-CO_2 is dissolved in a substance (or mixture of substances) at high pressure, acting as a plasticizer. The viscosity and melting point (or glass transition temperature) of the material is decreased, allowing it to be sprayed into ambient conditions.

Upon depressurization, the gas evaporates, and the system is cooled drown (Joule-Thomson effect) below the melting point of the material, leading to the solidification of the droplets and formation of fine particles. PGSS can be operated in batch mode or continuously feeding the materials into the precipitator through a static mixer (**Figure .I.4.)**.

Variations of PGSS are: Continuous Powder Coating Spraying Process (CPCSP) applied for the manufacture of powder coatings, and Gas Assisted Melting Atomization (GAMA) which uses air to improve the atomization.

In other processes, organic solvents and SC-CO_2 are used to generate expanded liquid solutions: the Depressurization of an Expanded Liquid Organic Solution (DELOS), the Precipitation by Pressure Reduction of Gas-expanded Liquids (PPRGEL), and the Supercritical-Assisted Atomization (SAA).

I.3.4.4. Miscellaneous

In the Supercritical Fluid Extraction of Emulsion (SFEE) and Continuous Supercritical Emulsions Extraction (SEE-C), the solute is dissolved in an organic solvent, and then an emulsion in water is prepared. The emulsion is atomized in counter current with SC-CO_2 so that the organic phase is extracted generating a suspension in water.

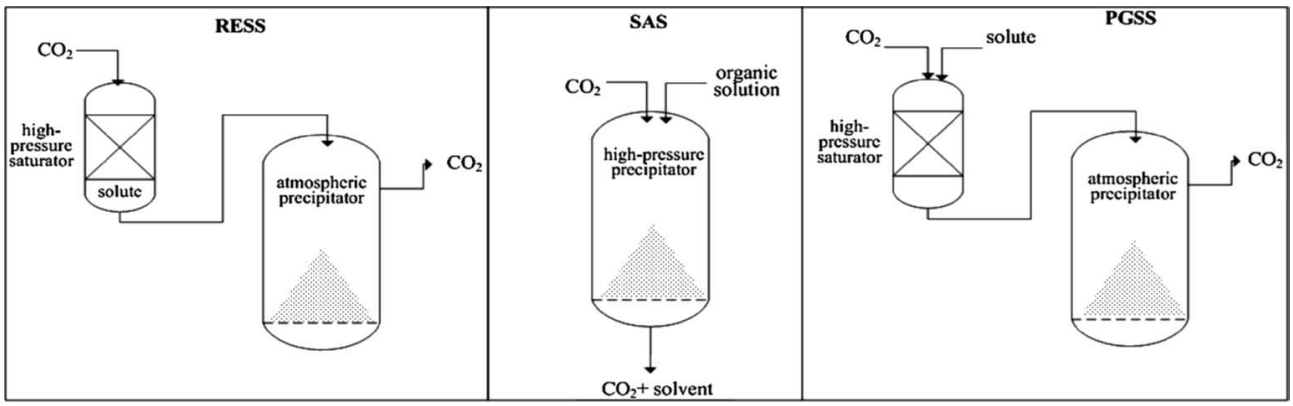


Figure .I.4. Process diagram of main SCF-based techniques.

SFEE/SEE is said to offer more flexibility to tune particle characteristics, enhanced reduction in particle size compared to SAS, and the ability to avoid particle agglomeration due to the presence of water, however, drying of the product needs to be carried out afterward.

In the Supercritical Reverse Phase Evaporation (SCRPE), SC-CO_2 is mixed with an organic solution containing a phospholipid and then an aqueous solution is introduced in the resulting homogeneous mixture to generate liposomes in a batch mode. Later, the same group developed the Improved Supercritical Reverse Phase Evaporation (ISCRPE), a completely solvent-free method in which the liposomes are formed by first mixing the water solution and phospholipid with the posterior introduction of SC-CO_2.

More details can be found elsewhere [6,23,28,75].

I.3.5.Application of SCFs In Pharmaceutical Industry

In the last few decades, the introduction of SCT in the pharmaceutical industry has seen a remarkable increase due to its variance properties; safety, environment friendly, and economical. This allows it also to replace some of the conventional processes such as solvent, foods, pharmaceuticals, fragrances, flavors, and pesticide extraction, biological and polymerization reactions, distillation, manufacturing nanoparticles, and wiped film evaporation [26].

Almost 98% of applications have been developed using CO_2 (the most used) as the SCF because of its low critical temperature and pressure [23].This, can be seen as a real advantage since a large number of pharmaceuticals are thermo-sensitive. As well as recognized as safe by the United States Food and Drug Administration (US-FDA) in the pharmaceutical processing [27].

Without forgetting to mention that those applications are applicable using server SCF processes that have been earlier mentioned. Below some of the interesting applications of SCF in pharmaceutical areas [28]:

I.3.5.1. Solubility Enhancement

SC-CO_2 has been used for the micronization of drug particles with particle size in the sub-micron or nano levels, thereby enhancing their solubility in water or other solvents. When it combined with other technologies such as composite or coacervate formation, and special types of equipment such as spray drying, freeze-drying, and specialized nozzles, the advantages are increased several folds and the difficulties in scaling up of this method are diminished. Co-solvents can be added in this process, which leads to increased solubility of the desired solute in SC-CO_2 because of the formation of hydrogen bonds. The increased solubility of poorly soluble drugs, using this approach, leads to the enhancement of dissolution rate and hence the bioavailability of the drug can be increased significantly.

I.3.5.2. Nanoparticle Technology

The environmentally safe SCFs have replaced the use of toxic organic solvents, which are generally used in the conventional techniques for nanoparticle formation such as solvent extraction, spontaneous emulsification or solvent diffusion, and organic phase separation methods. The major problem associated with the formation of nanoparticle composites using conventional agitators is to obtain a homogenous dispersion of nanoparticles in polymeric particles. If the mixture of polymeric particles and nanoparticles is taken in an extraction vessel containing SC-CO_2, and high shear stress of 1 Pa or more is applied in the vessel, then a highly dispersed nanoparticle-polymer system can be obtained.

I.3.5.3. Product Sterilization

SFT has found its place even in the process of sterilization. The methods conventionally used for sterilization, viz. Radiation sterilization, ethylene oxide sterilization, moist heat sterilization, sterilization using electron beam have been found to adversely affect the properties of biomedical and pharmaceutical substances. So, sterilization using supercritical CO_2 is used to inactivate the bacterial endospores, eliminate medical contaminants, and is also used in terminal packaging.

I.3.5.4. Separation of Enantiomeric Mixtures

The chiral nature of optically active substances is of great concern in the pharmaceutical industry as these drugs are metabolized differently in the body to produce different therapeutic activities. It might be possible that one of the enantiomers of a particular drug is responsible for the pharmacological activity while the other enantiomer of the same drug produces toxicity. To achieve

the resolution of a racemic mixture of a pharmaceutical drug and use the appropriate and desirable enantiomer, SFT has emerged as a novel method in recent years for this.

I.3.5.5. Fractionation of Essential Oils

Volatile and essential oils are prone to thermal degradation if their extraction is carried out by methods such as steam distillation. SFT has, therefore, gained momentum in the process of fractionation and extraction of essential and volatile oils. This is one of the areas where supercritical fluid extraction (SFE) has lent its major contribution.

I.3.5.6. Purification of Excipients

SCFs has been proved to be able to effectively remove the low molecular weight impurities, present in pharmaceutical excipients. This can be made possible only when the impurities present in the substance migrate rapidly to the particle surface, which is a prerequisite for SFE.

I.3.5.7. Preparation of Pharmaceutical Powders

The inherent properties of SC-CO_2 in SFT can be used for the processing of macromolecules such as proteins, peptides, and nucleic acids. The method has advantages over conventional lyophilization methods. Advantages include the incorporation of delicate biologicals without loss of biological activity and control over the morphology of the powder particles. Several studies have shown that diverse types of particles can be obtained when processing aqueous protein solutions into powders by using SCFs [23].

I.3.5.8. Complex Formation & Preparation of Inclusion Compounds

Supercritical fluid processes have been widely used in the production of complexes and inclusion compounds of steroids and other drugs. Complexes of poorly water-soluble drugs such as naproxen and ibuprofen were prepared with β-cyclodextrin using SC-CO_2. Owing to the complex formation, the incidence of gastrointestinal side effects were reduced, whereas, the wettability and hence the dissolution rate of the drugs were enhanced.

I.3.5.9. Drug Delivery Systems

Drug delivery systems such as solid-liquid nanoparticles are important tools to introduce drugs at the preselected targets in therapeutic doses into the body. Delivery of a drug is one of the most challenging research areas in pharmaceutical sciences. With a combination of drugs and

innovative delivery systems such as lipid nanocarriers, drug efficiency and safety have been improved significantly. There are various techniques available to produce drug-loaded solid lipid nanoparticles. Environmentally friendly supercritical CO_2 is the most popular and inexpensive solvent which has been used for the preparation of nanodrugs and nanocarriers in drug delivery system with supercritical fluid technology.

I.4.Conclusion

A drug with an unacceptable aqueous solubility, Also Known As bad bioavailability, represents one of the unwanted issues that researchers are looking for overcome these days. While it is suggested that the supercritical fluid technology could give a remarkable contribution as it has been excessively used, the needs for solubility data of drugs in SCFs is a prerequisite. Despite the fact that the experimental data of drug solubility in SC-CO_2 is now available, it remains however insufficient as is not possible, due to cost and tediousness, to measure the solubility of all potential drug-like molecules at all operating conditions of pressure and temperature. Here comes the necessity of models that can correlate and/or predict the solubility of drug-like compounds in SC-CO_2 at various operating conditions of pressure and temperature.

Luckily, the addition of informatics to this area, or in other words, machine learning methods to chemistry, combined with QSAR/QSPR had created a new discipline called chemoinformatics as a powerful tool computer drug discovery. We will introduce it in the next upcoming chapters.

Chapter II: Chemoinformatics & Programming tools

Chapter II: Chemoinformatics & Programming tools

II.1. Introduction

Since the day professor Karl Steinbush (German computer scientist) coined the word Informatik by publishing a paper called Informatik: Automatische Informationsverarbeitung ("Informatics: Automatic Information Processing"), researchers have observed the entry of computer science in numerous areas, with different ways such as artificial intelligence, data science, big data, data mining and so on.

With every interaction, it could simplify and optimize in a low period of time questions have been known as insolvable. The Same is with the chemistry domain. AKA, chemoinformatics; The use of computer and informational techniques, applied to a range of problems in the field of chemistry.

"Chemoinformatics is the mixing of those information resources to transform data into information and information into knowledge for the intended purpose of making better decisions faster in the area of drug lead identification and optimization".

F.K. Brown, 1998

As it is familiar, explaining with details all of its concepts in few lines keeps being impossible. Herewith, this chapter will only consider giving general views on informatics-bases, models, and algorithms related to the theme.

II.2. Basic concepts

II.2.1. Database

Database, also called electronic database (datasets), is a collection of information that is organized so that it can be easily accessed, searched, managed, and updated. It is usually controlled by a database management system (DBMS) [29,30].

II.2.2. SMILES

SMILES, refers to the Simplified Molecular Input Line Entry Specification, which is a chemical notation language specifically designed for modern chemical information processing [31]. SMILES is based on the principles of molecular graph theory [32].

II.2.3. Molecular Descriptors

A molecular descriptor is defined as the final result of a logical and mathematical procedure, which transforms chemical information encoded within a symbolic representation of a molecule into a useful number or the result of some standardized experiment [33].

II.2.4. Artificial Intelligence (AI)

What is artificial intelligence? over the last several years, This question has been debated and until this moment no formal definition exists.

Rich and Knight (1991) define AI as *"The study of how to make computers do things which, at the moment, people do better".*

B. Buchanan and E. Shordiffe (1983) *"Artificial Intelligence is a branch of computer science dealing with symbolic, non-algorithmic methods of problem-solving".*

A. Barr and E. Feigenbaum (1981) have proposed the following definition *"Artificial Intelligence is the part of computer science concerned with designing intelligent computer systems, that is, systems that exhibit characteristics we associate with intelligence in human behavior".*

Marvin Lee Minsky, who is considered as one of the founding fathers of AI, defines it as follows: *"the science of making machines do things that would require intelligence if done by men. It requires high-level mental processes such as: perceptual learning, memory and critical thinking".*

No matter how it defined, Knowledge-based or Artificial Intelligence techniques are used increasingly as alternatives to more classical techniques to model environmental systems with the recognition of its potential.

We review some of the techniques covered: Machine learning, Data mining, case-based reasoning, rule-based systems, artificial neural networks, fuzzy models, genetic algorithms, cellular automata, multi-agent systems, swarm intelligence, reinforcement learning, hybrid systems..., and so on [39]. AI mimics human perception, learning, and reasoning to solve complex problems.

NB:

Most AI concepts/techniques are Intertwined one in another. Depending on the use of those techniques in different disciplines it can be meant different things.

II.2.5. Machine Learning (ML)

ML is an area of computer science and a subset of AI that provides computers with the capability to think and learn without being programmed explicitly. ML is used in various computational tasks and its root objective is to train the machine with the help of data provided [41].

However, the value of Machine Learning is that it allows you to continue learning from data and predict the future [40], while the main focus is to make computers learn from past experiences.

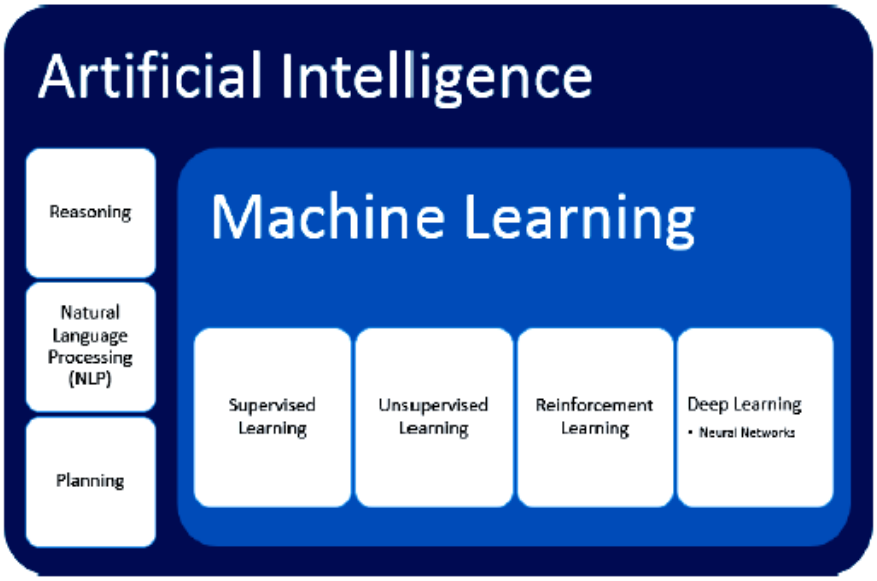


Figure .II.1. AI is the overall category that includes machine learning and natural language processing [40].

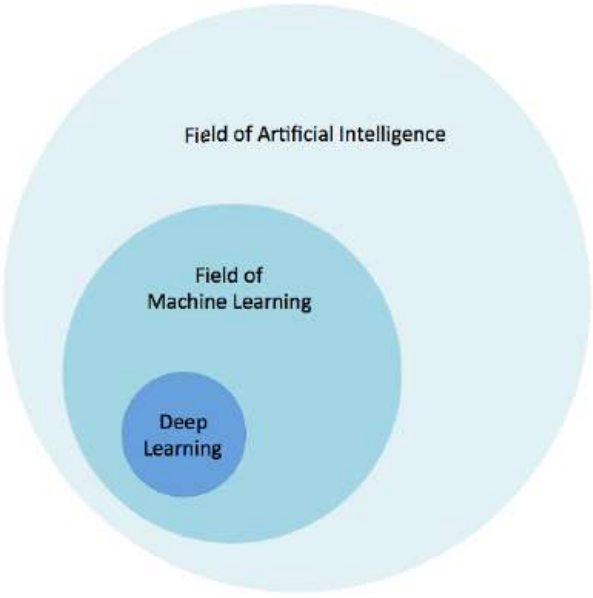


Figure .II.2. Representation of the relationship between AI, ML and DL.

As it can been seen, the field of AI is broad and has been around for a long time. Deep Learning (DL) is a subset of the field of machine learning, which is a subfield of AI.

The machine is trained on some datasets, and then, the algorithms are applied, so that the machine can make predictions and learn, respectively, on the given datasets. A model is programmed, procedures are operated upon it, and inputs are taken to obtain data-driven predictions instead of firm static program instructions. Each instance in the dataset that is used by the machine learning algorithms is depicted by features belonging to the same set. The features can be categorical, binary, or continuous [42].

We can use either supervised learning-based models or unsupervised learning models for prediction. The basic idea behind supervised learning-based models is to predict a target variable. Supervised learning is also referred to as predictive modeling. Classification is a popular predictive modeling algorithm while dealing with categorical variables. Another type of supervised learning is regression where we predict continuous outcomes. This procedure determines the computational methods and incorporates the patterns in large data. In descriptive modeling or unsupervised learning-based approaches, a model is always constructed through clusters of the data [41].

The primary objective of machine learning is that once the algorithm learns what to do with the given data, it performs its task automatically.

Machine learning and data mining often employ the same methods and overlap significantly. They can be roughly distinguished as follows:

- Machine learning, focuses on prediction, based on known properties learned from the training data.
- Data mining focuses on the discovery of (previously) unknown properties in the data. This is the analysis step of Knowledge Discovery in Databases [47].

Furthermore, detail information can be found elsewhere [41,42, 47].

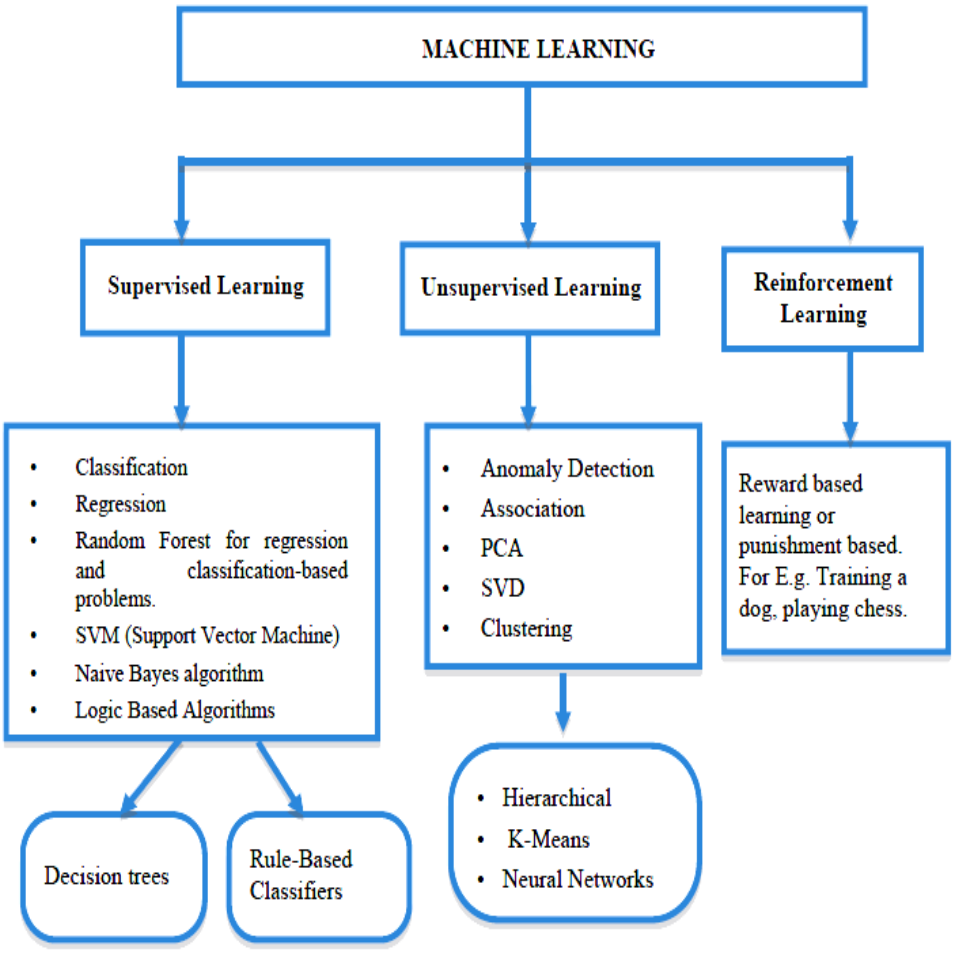


Figure .II.3. Flow chart of machine learning and the algorithms used for training the machine.

ML, represents a well-known family of algorithmic techniques based on a solid statistical theory and able to handle various experimental datasets.

II.3. Quantitative structure-Activity/Propriety Relationship

QSAR are mathematical models that rely on the basic principle of chemistry that states that the biological activity of any ligand or compound is associated with the arrangement of atoms forming the molecular structure. In other words, structurally related molecules possess similar biological activities [34].

This Structural information is denoted in the form of descriptors/features which are numerical values associated with the chemical constitution of a molecule structure ranging from atom counts to topological surface area [36].

Where the response is represented as a function of these molecular descriptors as depicted in **Eq.II.1**.

$$\text{Response (Biological activity or property)} = f\ \text{(Molecular Descriptors...)} \qquad \textbf{II.1}$$

The concept of QSAR has typically been used for drug discovery and development and has gained wide application for correlating molecular information with not only biological activities but also with other physicochemical properties, which has therefore been termed Quantitative Structure-Property Relationship (QSPR) [37] and toxicity Quantitative Structure-Toxicity Relationship (QSTR).

QSPR is a mathematical correlation identified between the physical response of a molecule and its structural information. QSPR approaches proved their ability in predicting various physical properties of a molecule from its structural information [36]. Such as melting point, boiling point, stability, dielectric constant, reactivity, diffusion coefficient, thermodynamic properties, hydrophobicity, and of course the solubility, our chosen property [34].

NB:

The study presented in this thesis aims to represent the solubility (the response) as a function of selected molecular descriptors with the addition of two independent intensive state variables the temperature and the pressure as depicted in **Eq.II.2**.

$$\text{Response (Solubility)} = f\ \text{(Temperature, Pressure ,Selected Descriptors)} \qquad \textbf{II.2}$$

II.3.1. Property Prediction (QSAR/QSPR)[38]

The relationships between many chemicals and, particularly, biological data of compounds and their structure are too complex to be directly predicted on first principles. In such cases, an indirect two-step process has to be invoked.

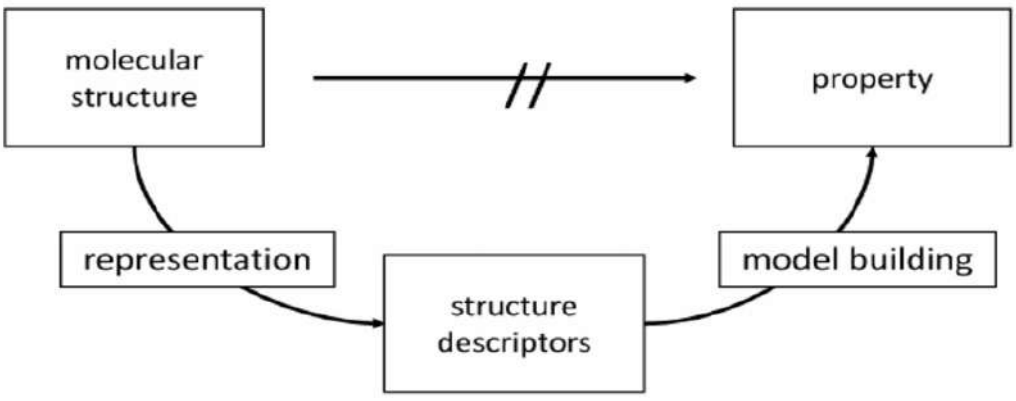


Figure .II.4. The basic approach to QSAR/QSPR/QSTR.

In the first step, a molecular structure is represented by structure descriptors. In the second step, a dataset of structures as represented by their descriptors and their associated properties is submitted to data analysis and model building method.

II.3.2. Model Construction (QSAR/QSPR)[34]

The quality of the QSAR/QSPR model can be directly related to:

- The data used for its construction. Hence, prior to model development, it is necessary to gain insights into the data.
- A thorough understanding of the problem and influencing factors assists in discerning meaningful relationships.
- Relevant background information about the system under study either biological or chemical is to be collected via literature search.
- Datasets for model construction need to be chosen carefully as poor and inconsistent data would lead to a corrupt model.
- Also, several other factors like the division of data into training and test datasets, molecular descriptors, statistical methods for model development that influence the quality of the model.

The schematic representation of the QSAR/QSPR model construction is given in **Figure.II.5.** A brief description of the steps involved in the QSAR/QSPR model generation is discussed in the following chapter.

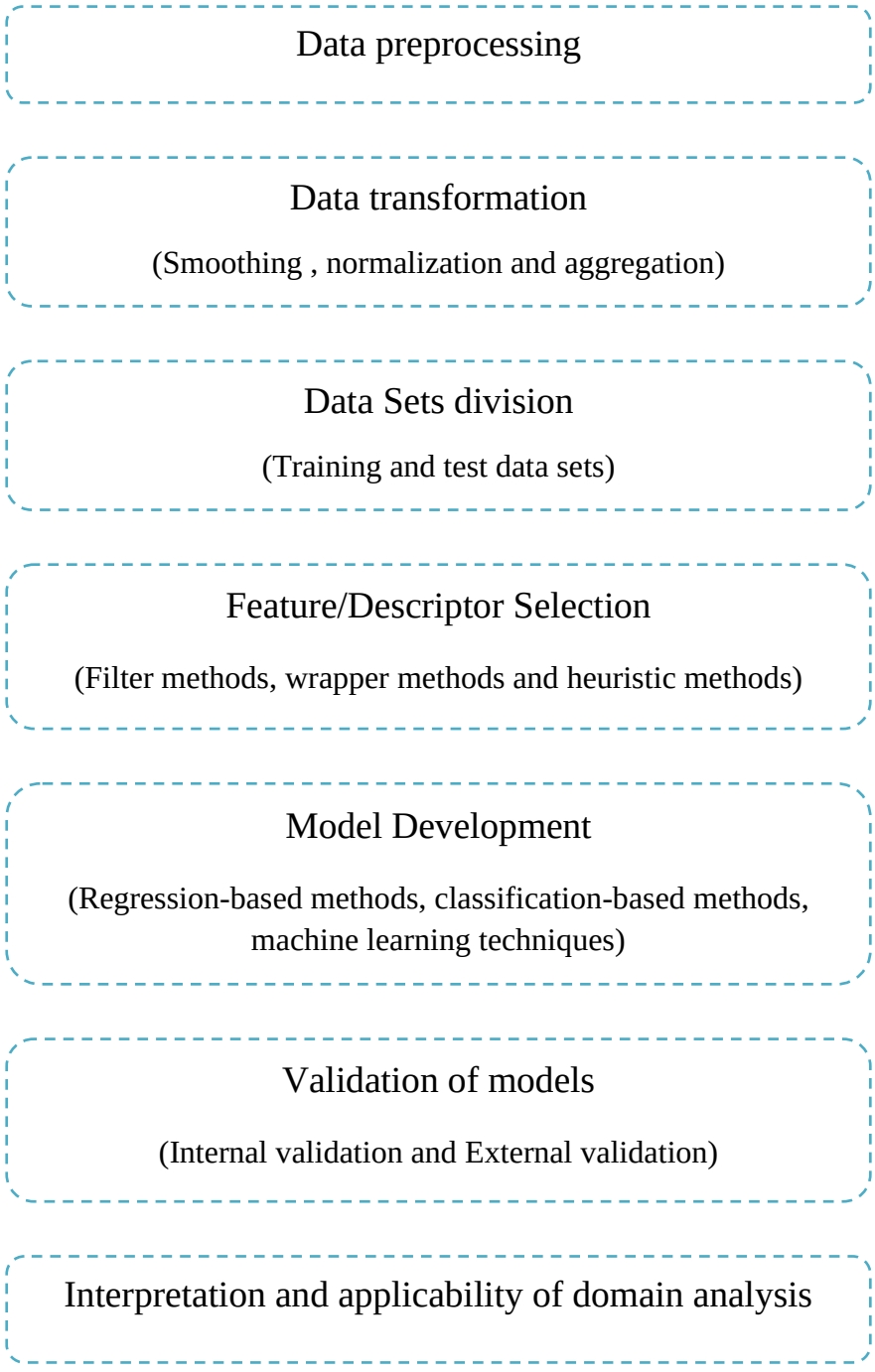


Figure .II.5. Schematic representation of QSAR/QSPR/QSTR [34].

The main challenge in QSAR is to select the group of descriptors that describe the most critical structural and physicochemical features associated with the activity. Quality of biological data and effective descriptor or variable selection is an initial essential part of the QSAR modeling

process. The application of QSAR models depends on the statistical significance and predictive ability of these models [50].

For the regulatory acceptance of QSAR models, five guidelines are agreed by the Organization for Economic Co-operation and Development (OECD), and these cover the following criteria: (i) a defined endpoint, (ii) an unambiguous algorithm, (iii) the domain of applicability, (iv) appropriate measures of goodness of fit, robustness, and predictivity of the developed model, and (v) a mechanistic interpretation, if possible [51].

These guidelines are now referred to as OECD Principles for the validation of QSARs. As a result, different groups of researchers have shown their keen interest in the development of more appropriate validation metrics for precise and predictive QSAR model development [51].

II.4. Artificial neural network (ANN) [55,56,57,38]

An artificial neural network, often just called a neural network, is a computational model that is inspired by the structure and function of biological neural networks in the brain. The Basic building block of every ANN is the artificial neuron, AKA "nodes", which is a simple mathematical model (function). Such a model has three simple sets of rules: multiplication, summation, and activation.

At the entrance of artificial neuron, the inputs are weighted that means that every input value is multiplied with individual weight (w_i) . In the middle section of an artificial neuron is a sum function that sums all weighted inputs and bias (b_i). At the exit of the artificial neuron, the sum of previously weighted inputs and bias is passing trough activation function that is also called the transfer function. Its schematic representation can be found in **Figure .II.6.**

This type of model consists of a number of artificial neurons (i.e., nonlinear processing units) that are connected to each other via synaptic weights (or simply just weights). As it can "learn" a task by adjusting weights. There are supervised and unsupervised models.

A supervised model requires a "teacher" or desired (ideal) output to learn a task. An unsupervised model does not require a "teacher", but it leans a task based on a cost function associated with the task. Another type, reinforcement learning which is also is a ML technique that sets parameters of an ANN, where data is usually not given, but generated by interactions with the environment.

No matter which method has been used, the task of learning is to set the values of weight and biases on the basis of learning data to minimize the chosen cost function.

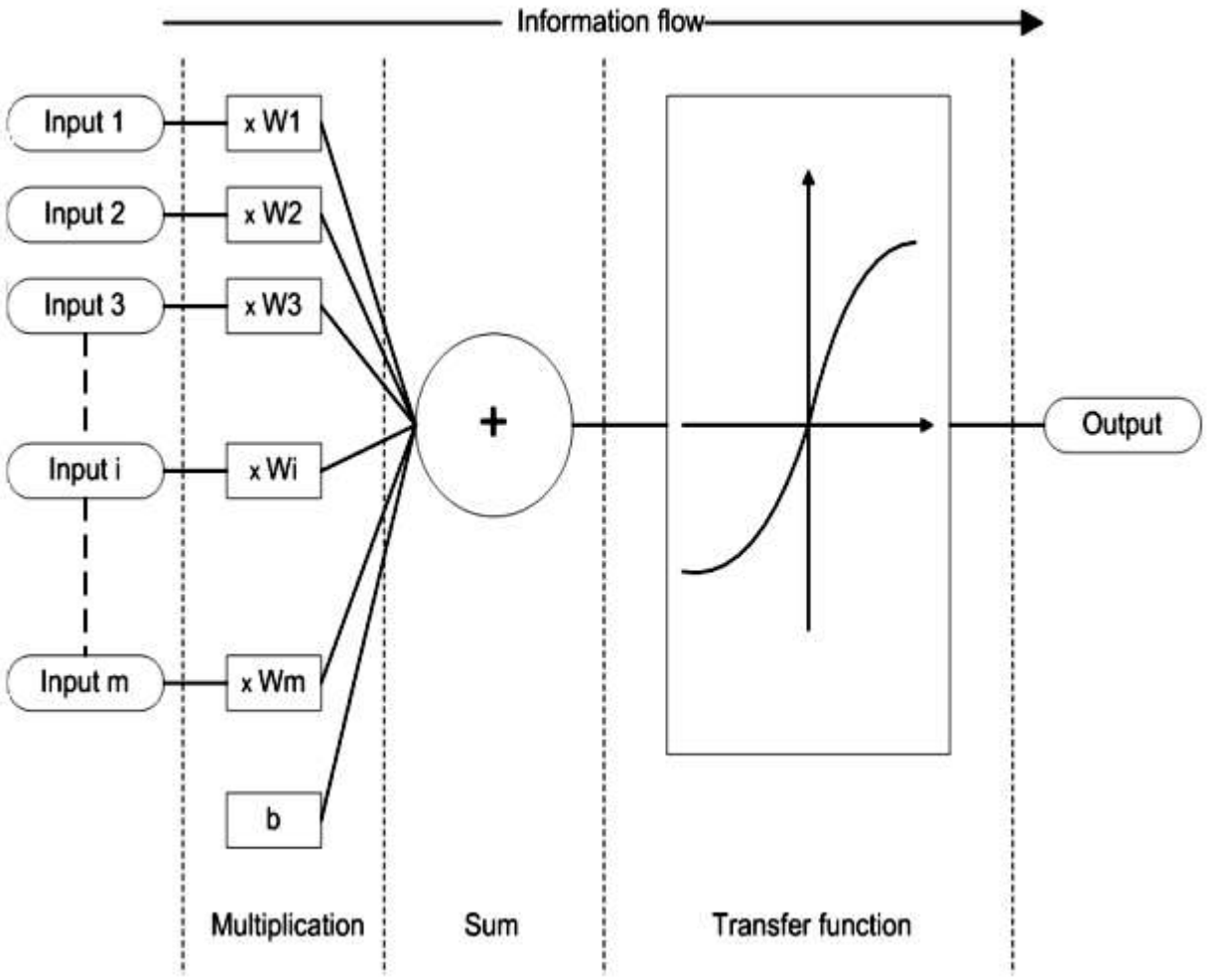


Figure .II.6. Working principle of an artificial neuron.

ANNs have been successfully used in various applications such as biological, medical, industrial, control engendering, software engineering, environmental, economical, and social applications. The high versatility of ANNs comes from its high capability and learning function. It has been theoretically proved that an artificial neural network can approximate any continuous mapping by arbitrary precision. Desired continuous mapping or the desired task is acquired in an artificial neural network by learning.

Although the working principles and simple set of rules of an artificial neuron look like nothing special the full potential and calculation power of these models come to life when we start to interconnect them into ANNs (**Figure .II.7.**). These ANNs use the simple fact that complexity can grow out of merely a few basic and simple rules.

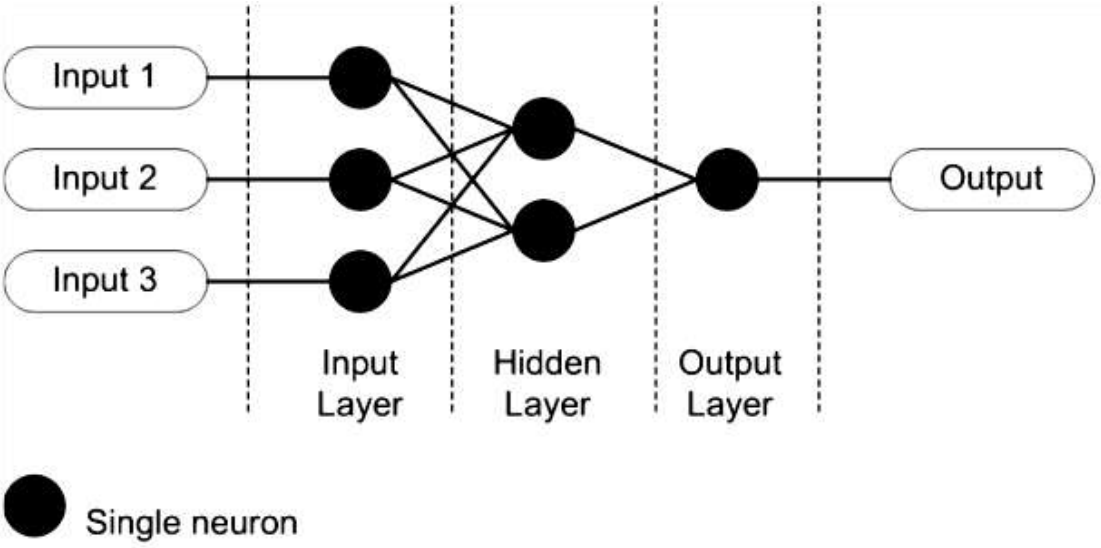


Figure .II.7. Example of simple artificial neural network.

While observing this figure we have to mention that for easier handling and mathematical describing of an ANN we group individual neurons into layers. We can see That any network has an input layer, an output layer, and a different number of hidden layers. A neuron is linked to all neurons in the next layer.

II.4.1. Artificial neuron Vs. Biological neuron

An artificial neuron is a basic building block of every ANN. Its design and functionalities are derived from observation of a biological neuron that is the basic building block of biological neural networks (systems) which includes the brain, spinal cord, and peripheral ganglia. Similarities in design and functionalities can be seen in **Figure .II.8.**

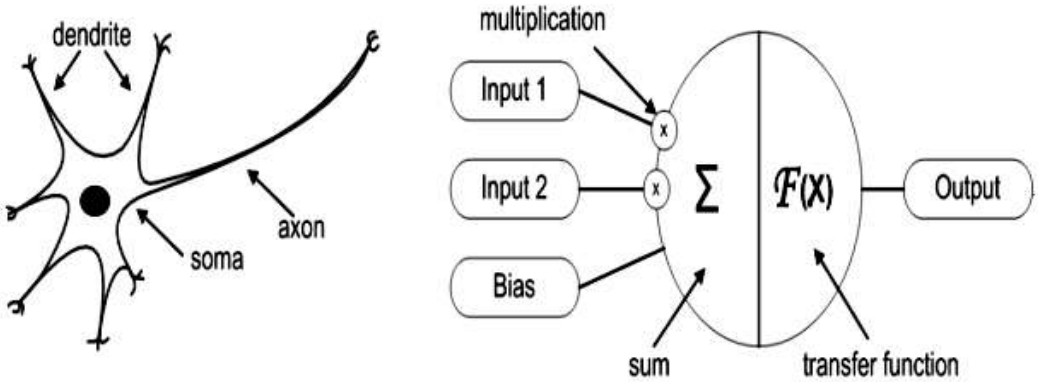


Figure .II.8. Biological and artificial neuron design.

The left side of a figure represents a biological neuron with its soma, dendrites, and axon. Where the right side of a figure represents an artificial neuron with its inputs, weights, transfer function, bias, and outputs.

In the case of biological neuron information comes into the neuron via dendrite, soma processes the information and passes it on via axon. while the case of the artificial neuron the information comes into the body of an artificial neuron via inputs that are weighted. The body of an artificial neuron then sums the weighted inputs, bias, and "processes" the sum with an activation function. At the end, an artificial neuron passes the processed information via output(s). The benefit of artificial neuron model simplicity can be seen in its mathematical description below:

$$y(k) = F(\sum_{i=0}^{m} w_i(k) \cdot x_i(k) + b) \qquad \text{II.3}$$

Where,

$x_i(k)$ Input value in discrete time k where i goes from 0 to m,

$w_i(k)$ Weight value in discrete time k where i goes from 0 to m,

b Bias,

F A transfer function, and

$y_i(k)$ Output value in discrete time k.

As seen from a model of an artificial neuron and its **Eq.II.3.** the major unknown variable of our model is its transfer function.

II.4.2. Activation Function

The transfer function transferred the input signals to output signals, defines the properties of the artificial neurons and can be any mathematical function. They can be chosen on the basis of the problem that artificial neuron needs to solve, and in most cases, we choose it from the following set of functions: Step function, Linear function, and Non-linear (most used Sigmoid) function.

II.4.2.1.Step function (Threshold)

If this transfer function is used in artificial neuron we call this artificial neuron perceptron. Perceptrons are used for solving classification problems and as such, it can be most commonly found in the last layer of artificial neural networks.

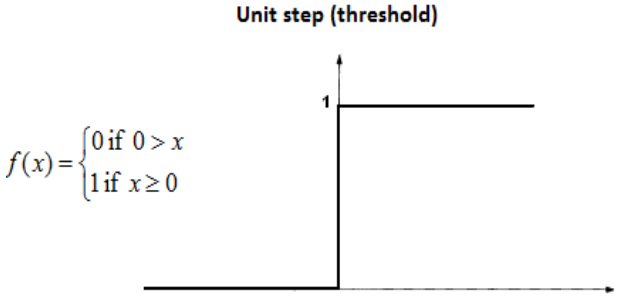


Figure .II.9. Activation function type Threshold.

II.4.2.2. Linear Function

In the case of a linear transfer function, an artificial neuron is doing simple linear transformation over the sum of weighted inputs and bias. Such an artificial neuron is in contrast to perceptron most commonly used in the input layer of artificial neural networks. the most used is the linear transfer function (purelin).

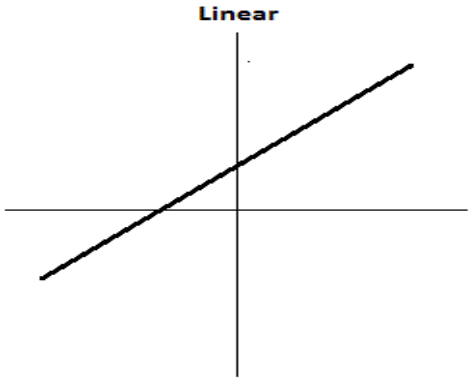


Figure .II.10. Activation function type Linear.

II.4.2.3.Non-Linear Function

When we use non-linear function the most commend use are represented below:

The **Sigmoid** functions are mostly used as activation functions in many practical applications of these networks. The most commonly used transfer functions are Hyperbolic Tangent Sigmoid function (tansig), and Log- Sigmoid transfer function (logsig).

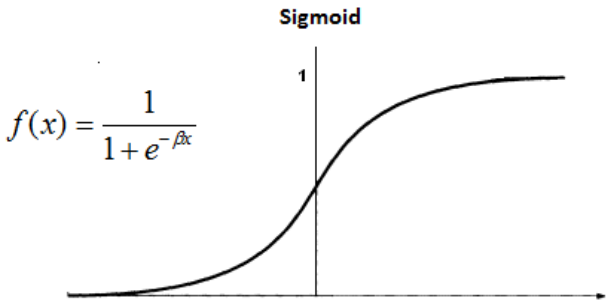


Figure .II.11. Activation function type Non-Linear (Sigmoid).

The **Piecewise Linear** function, the output is proportional to the total weighted output.

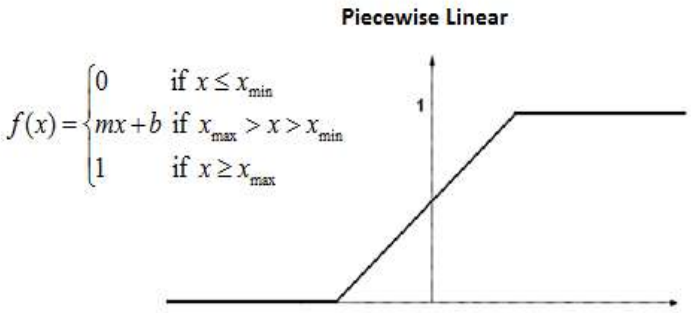


Figure .II.12. Activation function type Non-Linear (Piecewise Linear).

The **Gaussian** functions are bell-shaped curves that are continuous. The node output (high/low) is interpreted in terms of class membership (1/0), depending on how close the net input is to a chosen value of average.

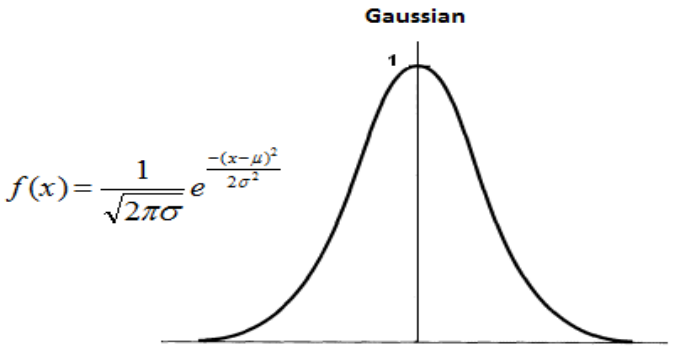


Figure .II.13. Activation function type Non-Linear (Gaussian).

In addition to other activation functions such as Bipolar, log, Sin, and Tanh...

II.4.3. ANN Topologies

The combination of two or more artificial neurons forms an artificial neural network. The way that individual artificial neurons are interconnected is called topology, architecture, a graph of an artificial neural network, or algorithm that is much famous...The fact that interconnection can be done in numerous ways, is the result of numerous possible topologies that are divided into two basic classes. **Figure .II.14**.

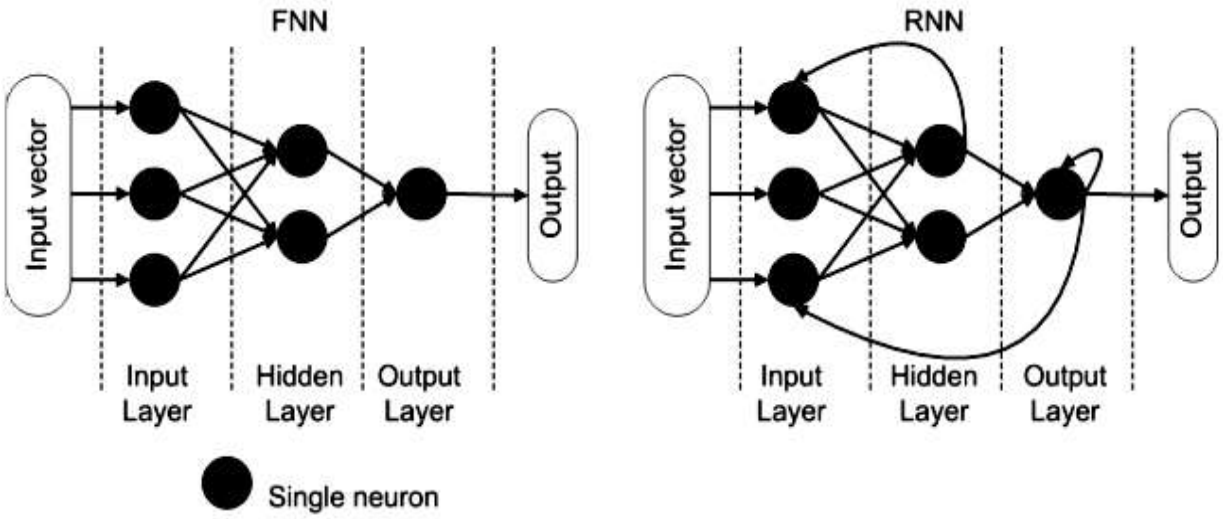


Figure .II.14. Feed-forward (FNN) and recurrent (RNN) topology of an ANN.

The left side of the figure represents simple feed-forward topology (acyclic graph) where information flows from inputs to outputs in only one direction, and the right side of the figure represent simple recurrent topology (semi-cyclic graph) where some of the information flows not only in one direction from input to output but also in opposite direction.

A **feed-forward network** is a non-recurrent network that contains inputs, outputs, and hidden layers; the signals can only travel in one direction. Input data is passed onto a layer of processing elements where it performs calculations. Each processing element makes its computation based upon a weighted sum of its inputs. The new calculated values then become the new input values that feed the next layer. This process continues until it has gone through all the layers and determines the output.

A **feed-back network** (e.g., recurrent neural network or RNN) has feed-back paths meaning they can have signals traveling in both directions using loops. All possible connections between neurons are allowed. Since loops are present in this type of network, it becomes a non-linear dynamic system that changes continuously until it reaches a state of equilibrium.

The most commonly and widely used model in many executable applications is the feed-forward neural network (FF network). It was the first and simplest kind of artificial neural network design.

Common types of artificial neural networks are Multi-layer Perceptrons Artificial Neural Networks(MLP-ANN), Hopfield artificial neural networks, Radial Basis Function Artificial Neural Networks (RBF-ANN). Because, there are no cycles or loops in this network, only in one direction, the information is moved to forward from the input nodes, through the hidden nodes to the output nodes.

Those two networks provides the use of various training algorithms (functions). They belong to one of the following five classes:

- Descent of the gradient,
- Descent of the conjugate gradients,
- Quasi-Newton,
- Levenberg-Marquardt, and
- Regularization.

ANNs comprise of an optimal number of neurons in their hidden layers. If this value is more than the number of neurons in the network, the network is not trained good and low accurate values are predicted. Also, if it is less than the number of neurons, quick convergence and over-fitting occur.

For more details on ANNs[56].

II.5. Support Vector Machine (SVM)

Support Vector machine is currently a hot topic in the statistical learning area. In the late 1990s, the traditional neural network approaches suffered severe difficulties with generalization and producing models which are the main reason for the foundation of support vector machines. It was developed in 1995 by Vladimir Vapnik and soon gained popularity due to many attractive features [58].

In statistical learning, SVMs have supervised learning methods with associated learning algorithms that analyze the dataset. It is first been introduced as a method for solving classification problems(SVC). However, due to many attractive features, it is recently extended to the area of regression analysis. Known as Support Vector Regression (SVR). The regression version of SVM is the widespread form of application of SVM [58,62].

SVMs belong to a family of generalized linear classifiers. In another term, SVM is a classification and regression prediction tool that uses ML theory to maximize predictive accuracy while automatically avoiding over-fit to the data [59].

SVMs can be defined as systems that use hypothesis space of linear functions in a high dimensional feature space, trained with a learning algorithm from optimization theory that implements a learning bias derived from statistical learning theory. It is also being used for many applications, such as handwriting analysis, face analysis, and so forth, especially for pattern classification and regression-based applications [59].

The advantage of this method is more convenient than the other models in the training phase and also has high efficiency. This algorithm designates the best separating line to classify the data with a more safety margin. In this method, vectors are elected as a selection criterion, and these vectors are applied to the best boundaries and categories for data. These vectors are termed as Support Vector [60].

In a nutshell, we identify at least three reasons for the success of SVMs: Their ability to learn well with only a very small number of free parameters, their robustness against several types of model violations and outliers, and last but not least their computational efficiency compared with several other methods [61].

II.5.1. Basic Idea of Statistical Learning Theory

Statistical learning theory is a framework for ML areas related to the fields of statistics and functional analysis. The goal for statistical learning is finding the fittest function, representing f the systematic information of the response from its predictors, which will be used for prediction and inference.

To estimate this unknown function, the approach used in statistical learning is called training data, whose name comes from the usage that trains our method on how to estimate f. Every point in the training data is an input-output pair, where the input maps an output. The statistical leaning problem consists of inferring the function that maps between the input and the output, such that the learned function can be used to predict the output from future input.

Statistical learning theory is widely used in regression modeling and data classification. In the regression problem, the responses are quantitative values which take a continuous range of values, meanwhile, the responses in classification problem are qualitative values which are elements from a discrete set of labels. In the statistical learning area, the approach to find the function f are different in many aspects, for example, the choice of the loss function, however, they also share some similarities [58].

II.5.2. Learning and Generalization

Early ML algorithms aimed to learn representations of simple functions. Hence, the goal of learning was to output a hypothesis that performed the correct classification (or regression) of the training data and early learning algorithms were designed to find such an accurate fit to the data.

The ability of a hypothesis to correctly classify data, not into the training set, is known as its generalization. SVM performs better in terms of not over-generalization when the neural networks might end up over-generalizing easily.

Another thing to observe is to find where to make the best trade-off in trading complexity with the number of epochs; the illustration brings to light more information about this **Figure.II.15.** [59].

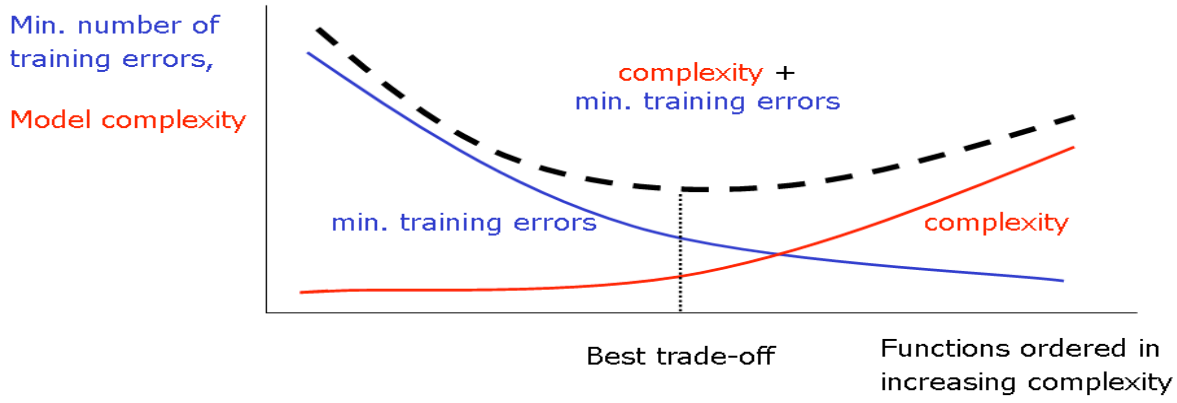


Figure .II.15. Number of Epochs Vs Complexity.

II.5.3. Support Vector Regression (SVR)

Support vector regression (SVR), the extended version of SVM, has become a paramount computational tool due to its effective applications in regression and prediction problems. SVR was developed for the prediction of regression functions, and it is based on the implicit nonlinear conversion of the data into a higher-dimensional feature space [62].

The fundamentals of SVR are briefly reported here considering a regression function *F* which is estimated based on the training data in the form of :

$$F = \{(x_i, y_i) \quad i = 1,2,3, \dots, n \qquad \text{II.4}$$

where,

x_i , y_i The input and output sets, respectively, and

n The overall number of the dataset.

SVR predicts the target values by utilizing the regression function as follows:

$$F = w^T \varphi(x) + b \qquad \text{II.5}$$

where,

F The predicted output,

w The weight vector,

b The bias, and

$\varphi(x)$ The high-dimensional input vector.

Flatness in **Eq.II.5** means that one seeks small *w*. For this, it is required to minimize the Euclidean norm, i.e. $\|w\|^2$.The coefficients *w* and *b* were computed by minimizing the risk function, as given below.

$$R(F) = \frac{1}{2} \|w\|^2 + C \frac{1}{n} \sum L_\varepsilon(y_i, F_i) \qquad \text{II.6}$$

where,

$$L_{\varepsilon}(y_i, F_i) = \begin{cases} 0, & if\ |y_i - F_i| \leq \varepsilon \\ |y_i - F_i| - \varepsilon, & overwise \end{cases} \quad \text{II.7}$$

The penalty parameter *C* is used to identify the trade-off between function intricacy and losses *"represents the balance between the complexity of the model and the approximate error "*. As can be seen in **Figure .II.16**, the parameter ε is the loss function which represents the range between the actual values and the regression function.

This range can be seen as a tube around the regression function. All points located on the exterior of the tube are considered as training errors. $L_{\varepsilon}(y_i, F_i)$ is called the ε- insensitive loss function.

Eq.II.6. can be changed to the following constrained form.

$$minimize\ \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) \quad \text{II.8}$$

$$Subject\ to \begin{cases} w \cdot \varphi(x) + b - y_i \leq \varepsilon + \xi_i \\ y_i - (w \cdot \varphi(x) + b) \leq \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \geq 0, \quad i = 1,2,3,\dots,n \end{cases} \quad \text{II.9}$$

Where,

ξ_i & ξ_i^* Two positive slack variables, which assume nonzero values on the exterior of the ε-tube and zero inside **Figure .II.16**.

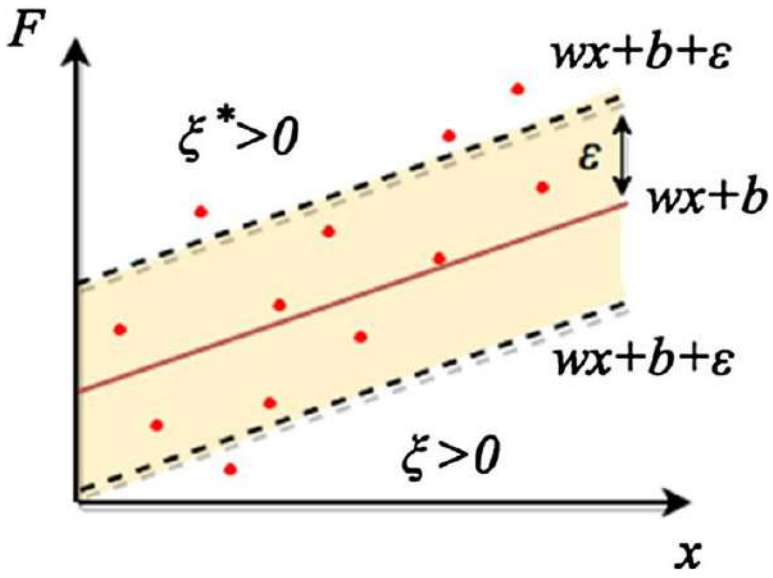


Figure .II.16. Regression with the ε-insensitive tube.

The optimization problem in **Eq.II.8.** can be solved more easily in its dual formulation. Therefore, a standard dualization technique using Lagrangian multipliers has been employed. Using the Lagrangian multipliers, this problem can be written in the dual formulation as follows:

$$Maximize\left\{-\frac{1}{2}\sum_{i,j=1}^{n}(\alpha_i,\alpha_i^*)\left(\alpha_j,\alpha_j^*\right)(\varphi(x_i)\cdot\varphi\left(x_j\right))-\varepsilon\sum_{i=1}^{n}(\alpha_i,\alpha_i^*)+\sum_{i=1}^{n}y_i(\alpha_i,\alpha_i^*)\right\} \quad \text{II.10}$$

$$Subject\ to\begin{cases}\sum_{i=1}^{n}(\alpha_i-\alpha_i^*)=0\\ 0\le\alpha_i\le C \quad i=1,2,3,\dots,n\\ 0\le\alpha_i^*\le C \quad i=1,2,3,\dots,n\end{cases} \quad \text{II.11}$$

where,

α_i, α_i^* nonlinear Lagrangian multipliers.

The SVR function can be obtained by solving the dual maximization problem in **Eq.II.10.** as follows:

$$F(x,\alpha_i,\alpha_i^*)=\sum_{i=1}^{n}(\alpha_i-\alpha_i^*)(\ \varphi(x_i)\cdot\varphi\left(x_j\right))+b \quad \text{II.12}$$

The vector inner-product $(\varphi(x_i) \cdot \varphi(x_j))$ represents the mapping function from the input space to feature space, and it can be replaced by a kernel function $k(x_i, x_j)$. Hence, **Eq.II.12.** becomes:

$$F(x, \alpha_i, \alpha_i^*) = \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) k(x_i, x_j) + b \qquad \text{II.13}$$

II.5.4. Kernel Functions

The basic idea with nonlinear SVMs(SVC, SCR) is to map training data into a higher dimensional feature space via some mapping *φ(x)* and construct a separating hyper-plane with maximum margin in the input space. As shown in **Figure .II.17.** the linear decision function in the feature space corresponds to a non-linear decision boundary in the original input space. By use of a kernel function, $k(x_i, x_j) = (\varphi(x_i) \cdot \varphi(x_j))$, it is possible to compute the separating hyper-plane without explicitly carrying out the mapping into feature space. To ensure that a kernel function actually corresponds to some feature space, it must be symmetric, i.e. , $k(x_i, x_j) =, k(x_j, x_i)$.

Typically, kernels need to satisfy Mercer's Theorem, which states that the kernel matrix, must be positive semi-definite, i.e. it has no non-negative eigenvalues.

$$K = (k(x_i, x_j))_{i,j=1}^{n} \qquad \text{II.14}$$

Typical choice for kernels are:

❖ **Linear Kernel**

$$K(x_i, x_j) = \varphi(x_i) \cdot \varphi(x_j) \qquad \text{II.15}$$

❖ **Polynomial Kernel**

$$K(x_i, x_j) = (xy + 1)^P \qquad \text{II.16}$$

❖ **Radial Basis Kernel Function (RBF)**

$$K(x_i, x_j) = e^{-\|x-y\|^2/2\sigma^2} \qquad \text{II.17}$$

❖ **Sigmoid Kernel**

$$K(x_i, x_j) = \tanh(kxy + \delta) \qquad \text{II.18}$$

❖ **Gaussian function**

$$k(x_i, x_j) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \qquad \text{II.19}$$

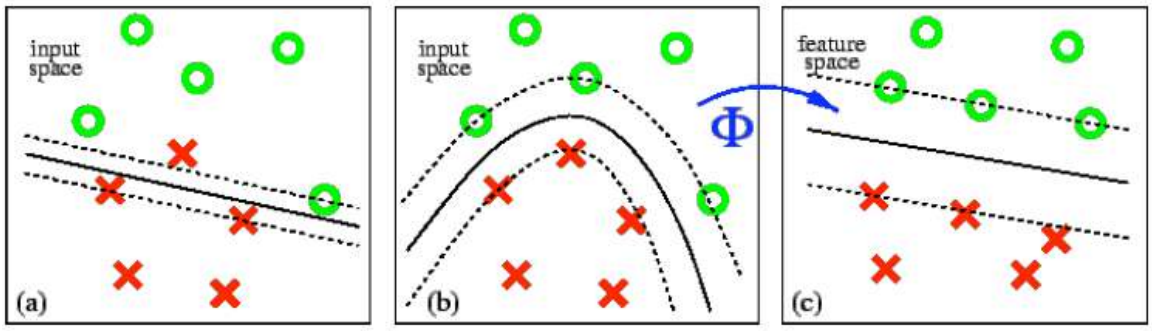


Figure .II.17. Effect of kernel transformation.

The data is not linearly separable in the input space (a), and the data can only be separated by a non-linear surface (b). This non-linear surface in the input space corresponds to a linear surface in a kernel-mapped feature space.

Each kernel corresponds to some feature space and because no explicit mapping to this feature space occurs, optimal linear separators can be found efficiently in feature spaces with millions of dimensions [63].

II.6. Adaptive Neuro-Fuzzy Inference System (ANFIS)

Adaptive Neuro-Fuzzy Inference System (ANFIS) is a combination of fuzzy systems and neural networks. In ANFIS, a neural network is used to decide the parameters of the fuzzy system. The necessity for manual optimization of the fuzzy system parameters is mainly removed by ANFIS.

The system parameters are adjusted automatically with the use of the Neural network. e.g.… adjusting the membership functions bounds, leading to improved performance without operator intervention.

The neuro-fuzzy system with the learning capability of neural networks and with the benefits of the rule-base fuzzy system can improve the performance considerably and can provide a mechanism to incorporate past observations into the classification, or regression, process.

In the neural networks, the training basically builds the system. However, using a neuro-fuzzy scheme, the system is built by fuzzy logic definitions and is then refined using neural network training algorithms.

Both neural networks and fuzzy logic are universal estimators. They can estimate any function to any prescribed accuracy if sufficient hidden neurons and fuzzy rules are available. Back-propagation and Gradient descent algorithms are used to adjust the parameters of membership functions (fuzzy sets) and the weights of defuzzification (neural networks) for fuzzy neural networks. For updating the parameters, ANFIS applies these two techniques.

The ANFIS is a FIS (Fuzzy Inference System) implemented in the framework of an adaptive fuzzy neural network. In ANFIS, the knowledge representation of a FIS is combined with the learning power of ANNs.

The basic objective of ANFIS is to integrate the features of fuzzy systems and neural networks. The advantage of fuzzy is that prior knowledge is denoted into a set of constraints to decrease the optimization research space. The adaptation of back propagation to a structured network so as to automate fuzzy control parametric tuning is utilized from Neural Network.

ANFIS employs gradient descent algorithm to fine-tune premise parameters which define membership functions. For consequent parameters that describe the coefficients of each equation, the Least squares method is used by ANFIS to identify them.

This approach combines the gradient descent algorithm and the least-squares method, thus called hybrid learning method. To achieve good generalization of unseen data, the size of the training data set must be as equivalent at least as the number of adjustable parameters in ANFIS [66].

The neurons in ANFIS have different structures:

• The Membership function is described by parameterized soft trapezoids (Generalized Bell Functions).

• Functions are linear regressions and multiplication with w, that is, normalized weights ω, and Output (Algebraic Sum).

• The Normalization is by Sum and arithmetic division.

• The rules are differentiable T-norm usually product [66].

II.6.1. Fuzzy systems

The term "system" is usually understood as a set of interacting components with a well-defined structure and organized as an intricate whole that can be distinguished from the "external" environment. A system communicates with the environment through so-called inputs and outputs.

Fuzzy systems are structures based on fuzzy techniques oriented towards information processing, where the usage of classical sets theory and binary logic is impossible or difficult. In the literature, terms such as fuzzy system, fuzzy model, a system based on fuzzy rules, fuzzy controller, or fuzzy associative memory are used interchangeably depending on the application type. Their main characteristic involves symbolic knowledge representation in a form of fuzzy conditional (if-then) rules.

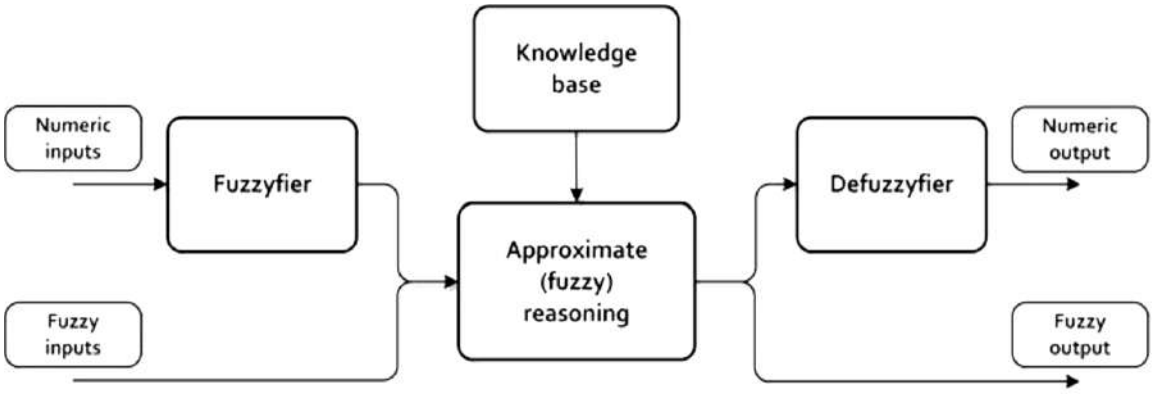


Figure .II.18. The typical structure of a fuzzy system.

The typical structure of a fuzzy system as depicted in **Figure .II.18.** consists of four functional blocks: The fuzzifier, the fuzzy inference engine, the knowledge base, and the defuzzifier. Both linguistic values (defined by fuzzy sets) and crisp (numerical) data can be used as inputs for a fuzzy system. If crisp data are applied, then the inference process is preceded by fuzzification, which assigns the appropriate fuzzy set to the nonfuzzy input.

The values of input variables are mapped into linguistic values of the output variable by means of the appropriate method of approximate reasoning (inference engine) using expert knowledge, which is represented as a collection of fuzzy conditional rules (knowledge base). In addition to the linguistic values, the numerical data may be required as the fuzzy system output. In such cases, defuzzification methods are used, which assign the representative crisp data to the resultant output fuzzy set.

Practical applications of fuzzy systems include problems for which the complete mathematical description is unavailable, or where the usage of the precise (nonfuzzy) model is uneconomical or highly inconvenient. The ability to process inaccurate information makes a fuzzy system an excellent tool, for example, for control processes, system identification, decision support, and signal and image processing. More mathematical details can be found in [65].

Nevertheless, the main problem with fuzzy logic is that there is no systematic procedure for the design of a fuzzy controller. On the other hand, a neural network has the ability to learn from the environment (input-output pairs), self-organize its structure, and adapt to it in an interactive manner [67]. This is a simple reason behind the existence of ANFIS. A system that is able to methodologies self-organize network structure and to adapt parameters of the fuzzy system for predicting different kinds of properties.

II.6.2. Architecture and algorithm

ANFIS is a type of neural network focused on the Takagi-Sugeno fuzzy inference system. It is an AI technique currently using hydrological processes. ANFIS is firstly introduced by Jang (1993) and it is based on the first-order Sugeno fuzzy model. ANFIS commonly uses either back-propagation or a combination of back-propagation and least square estimation for the prediction of a Membership Functions (MFs) parameter. The most significant purpose of integrating neural networks with fuzzy systems is to exercise learning ability of neural network while the learning capability is a benefit in terms of a fuzzy system. On the other hand, there are more benefits for a combined system from the aspect of a neural network [68].

Takagi-Sugeno type fuzzy inference system is used in ANFIS where every rule's output can be a constant term or can be a linear combination of input variables addition to a constant term. The weighted average of every rule's output is the final output [68].

An adaptive network is a multilayer feed-forward network. This network has five layers. The input and output layers in this network represent the descriptors and the response, respectively.

The hidden layer contains the node functions as the MFs and the rules, in which each node performs a particular function. In this system, "the square nodes" (adaptive nodes) have parameters; While "the circle nodes" (fixed nodes) have no parameters (see **Figure .II.20.**). For example, in a Sugeno fuzzy model with two inputs, *x* and *y*, and one output, a common rule set with two fuzzy" if-then" rules is as follows [69]:

Rule I: if $x=A_1$ *and* B_1, then

$$f_1 = p_1x + q_1y + r_1 \qquad \text{II.20}$$

Rule II: if $x=A_2$ *and* $y=B_2$, then

$$f_2 = p_2x + q_2y + r_2 \qquad \text{II.21}$$

where,

f	An output parameter (response),
p, *q*,& *r*	Are linear parameters, and
A& *B*	Are nonlinear parameters.

Then adaptive neural network and its operationally identical to the FIS are presented in **Figure.II.19.** and **Figure II.20.** respectively.

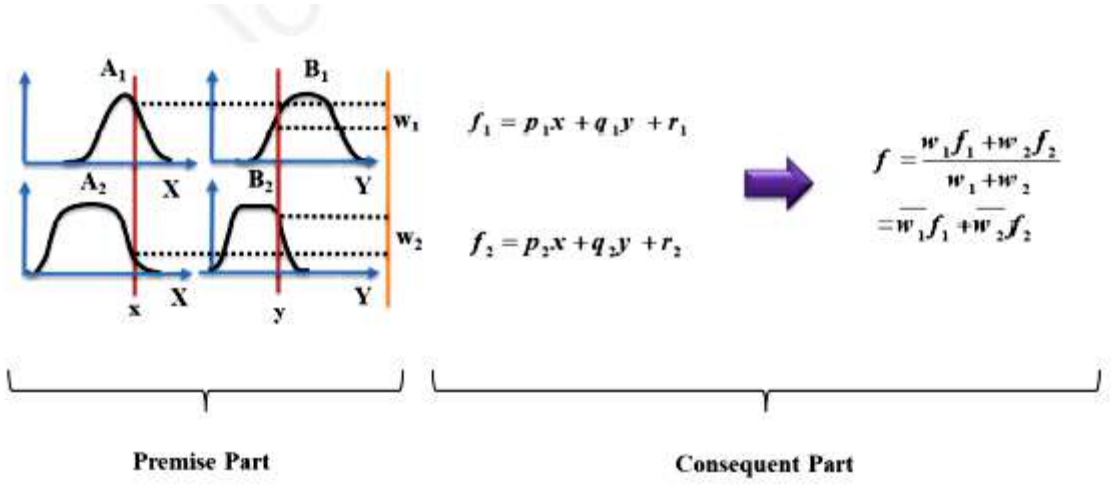


Figure .II.19. Schematic structure of the TSK fuzzy model [70].

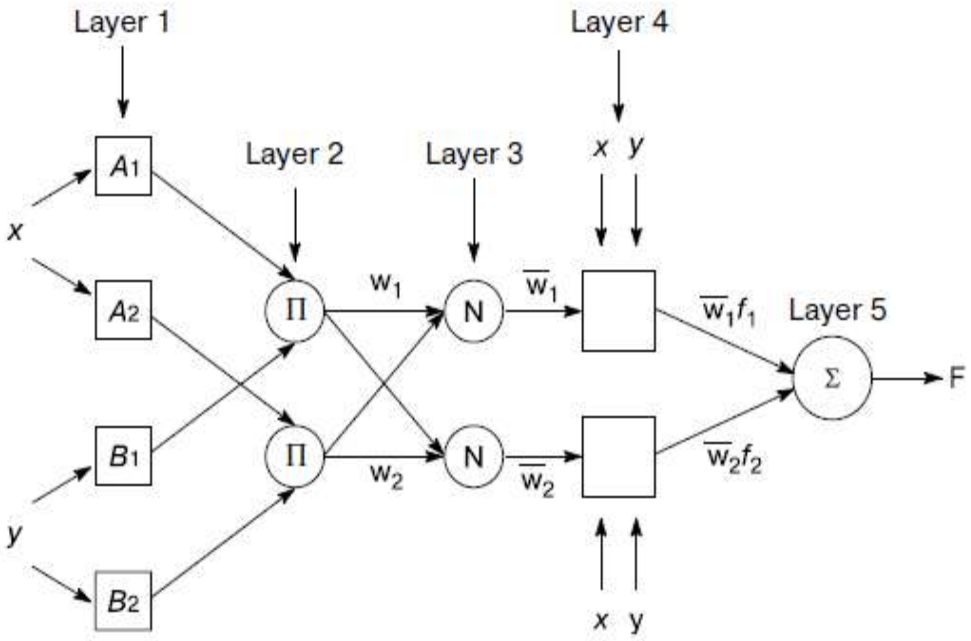


Figure .II.20. The ANFIS architecture with two input parameters *x* and *y*.

The mechanism for the Sugeno model and the corresponding equivalent ANFIS are shown in **Figure .II.20.** Design of five layers of the system is described below.

Layer 1. input nodes. Each node of this layer generates membership grades to which they belong to each of the appropriate fuzzy sets using MFs. It's "a square node" with a node function.

$$\begin{cases} O_i^1 = \mu_{Ai}(x) \\ O_i^2 = \mu_{Bi}(y) \end{cases} \qquad \text{II.22}$$

where,

$i = 1, 2$,

x, y The crisp inputs to node i, and

A_i &B_i The linguistic label (small, large, *etc*.) associated with the node function. μ_{Ai} ,μ_{Bi} respectively.

Due to smoothness and concise notation, the Gaussian and bell-shaped MFs are increasingly popular for specifying fuzzy sets. The bell-shaped MFs have one more parameter than the Gaussian MFs, so a non-fuzzy set can be approached when the free parameter is tuned.

In other words, O_i^1 is the membership function of A_i. Usually, the node function $\mu_{Ai}(x)$ (same for O_i^2) has a bell-shaped (Gaussian) distribution; However, any continuous and piecewise differentiable functions, such as commonly used trapezoidal or triangular shaped MFs, are also qualified candidates for the node functions in the **layer *1***. These MFs have parameters, which referred to as the premise parameters.

Layer 2. rule nodes. In the second layer, the AND operator is applied to obtain one output that represents the result of the antecedent for that rule, i.e., firing strength. Firing strength means the degrees to which the antecedent part of a fuzzy rule is satisfied and it shapes the output function for the rule. Hence the outputs O^2_i of this layer are the products of the corresponding degrees from ***Layer 1***.

Every node in the ***layer 2*** is "a circle node" labeled Π (see **Figure .II.20.**), which multiplies the incoming signals and sends the product out. For instance,

$$O_i^2 = w_i = \mu_{A_i}(x) \cdot \mu_{B_i}(y) \qquad \text{II.23}$$

where,

i = 1, 2. Each node output represents the firing strength of a rule.

Layer 3. average nodes. In the third layer, Every node in the layer *3* is "a circle node" labeled N; The i^{th} node calculates the ratio of the i^{th} rule firing strength to the sum of the firing strengths of all rules Consequently, w_i is taken as the normalized firing strength.

$$O_i^3 = \bar{w}_i = \frac{w_i}{w_1 + w_2} \qquad \text{II.24}$$

where,

i=1,2.

Layer 4. consequent nodes. The node function of the fourth layer computes the contribution of each i^{th} rules toward the total output and the function defined as:

$$O_i^4 = \bar{w}_i f_i = \bar{w}_i (p_i x + q_i y + r_i) \qquad \text{II.25}$$

where,

$\bar{w}_i$ The output of the ***layer 3***, and

$\{p_i, q_i, r_i\}$ The parameter set.

The parameters in the ***layer 4*** are referred to as the consequent parameters. Every node *i* in the ***layer 4*** is "a square node" with a node function.

Layer 5. Output nodes. The single-node computes the overall output by summing all the incoming signals. Accordingly, the defuzzification process transforms each rules fuzzy results into a crisp output in this layer a single node in the layer *5* is "a circle node" labeled Σ computes the overall output O_i^5 as the summation of all incoming signals.

$$O_i^5 = overall\ output = \sum_i \bar{w}_i f_i = \frac{\sum_i w_i f_i}{\sum_i w_i} \qquad \text{II.26}$$

Thus, in the present work we constructed the transparent and efficient ANFIS system, which is functionally equivalent to the Sugeno fuzzy model.

II.7. Genetic algorithm (GA)

GA is a search technique used in computing to find true or approximate solutions to optimization and search problems. It categorized as global search heuristics. A particular class of evolutionary algorithms that use techniques inspired by evolutionary biologics such as inheritance, mutation, selection, and crossover (also called recombination). Implemented as a computer simulation in which a population of abstract representations (called chromosomes, the genotype , or the genome) of candidate solutions (called individuals, creatures, or phenotypes) to an optimization problem evolves toward better solutions.

Evolution usually starts from a population of randomly generated individuals and happens in generations. In each generation, the fitness of every individual in the population is evaluated, multiple individuals are selected from the current population (based on their fitness), and modified (recombined and possibly mutated) to form a new population.

The new population is then used in the next iteration of the algorithm. Commonly, the algorithm terminates when either a maximum number of generations has been produced, or a satisfactory fitness level has been reached for the population [71].

GA is an evolutionary and global search algorithm and is quite a strong optimization approach for complicated search domain. John Holland established this technique in early 1970 by inspiring from natural genetic and Darwin's evolutionary hypotheses. The stages of the algorithm are shortly defined as following:

Reproduction: in this stage, a newborn group of population is produced by choosing the most suitable solution in the present population. In this section complicated contest with specific probability is applicable.

Crossover: the main phase of the genetic algorithm is crossover. Crossover, AKA recombination, is a genetic algorithm combines the genetic data of two parents to produce new children in the generating process.

Mutation: mutation is a genetic operator employed to retain genetic verity from one offspring of chromosomes to the following generation. It is comparable to biological mutation. The mutation changes one or more gene values in a chromosome. The **Figure .II.21.** illustrates a simplified flow sheet of GA [70].

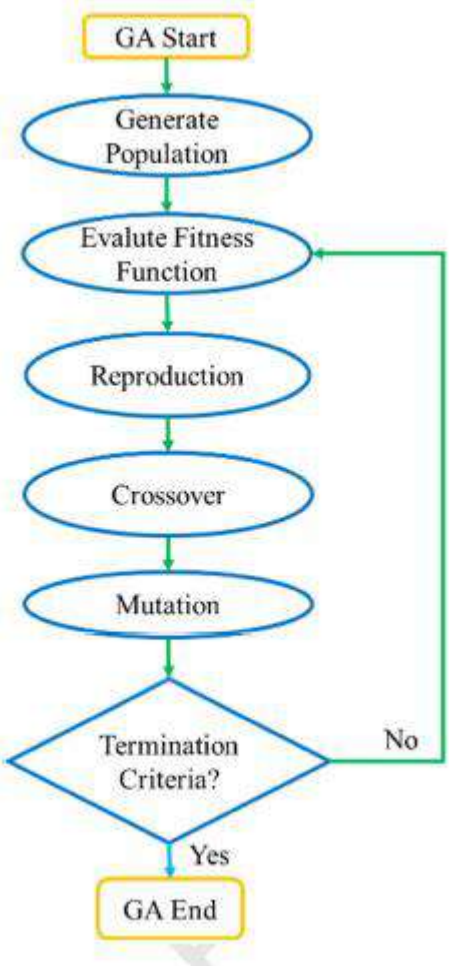


Figure .II.21. A simple flow sheet of GA [70].

The initial population $G(0)$ is generated randomly. Thereafter $G(t)$ produces $G(t + 1)$ through selection and reproduction. A proportion of the population is selected to breed and produce new chromosomes. Selection is according to the fitness of individual solutions, i.e. proximity to a perfect solution, most often by roulette selection and deterministic sampling. Roulette selection randomly selects a parent with a probability calculated from the fitness f_i of each individual by [73]:

$$F_i = \frac{f_i}{\sum_i f_i} \qquad \text{II.27}$$

Deterministic sampling assigns a value $C_i = \text{RND}(nF_i) + 1$ to organism i of the n organisms in the population, where RND rounds its argument to the nearest integer. Each organism is selected as a parent C_i times.

Below Some important points about Genetic Algorithms [72]:

1. Based on the idea of the continued existence of the fittest, another population is framed that comprises of the fittest guidelines in the existing population and children's values of these principles too.
2. The classification accuracy is used to find the fitness of the rule.
3. Mutation and crossover are two genetic operators and mutation are utilized to construct children or offspring.
4. In the crossover, the substring from a pair of guidelines is exchanged to create a fresh set of rules/guidelines.
5. In mutation, arbitrarily pointed bits in a string of a rule are reversed.
6. Parallelization is very easy to achieve in such algorithms and can be used for classification as well as for optimization problems.

II.8. Dragonfly Optimization Algorithm

The dragonfly optimization (DFO) algorithm is a recently established new and efficient swarm intelligence optimization technique proposed by Mirjalili [64]. It was inspired by the dynamics of dragonflies in nature. Dragonflies are carnivorous insects that catch and eat a wide variety of small insects, from gnats and mosquitoes to wasps and butterflies [62].

Generally, dragonfly swarms are both dynamic and static in the natural world. Dynamic swarms, or migratory swarms, form as large groups (hundreds of thousands of dragonflies) and fly in a single direction for long distances as shown in **Figure .II.22.** During the process of static swarms, in which the dragonflies hunt prey, they fly in small groups frequently over a well-determined small area and much closer to the land as shown in **Figure. II.23.** [64].

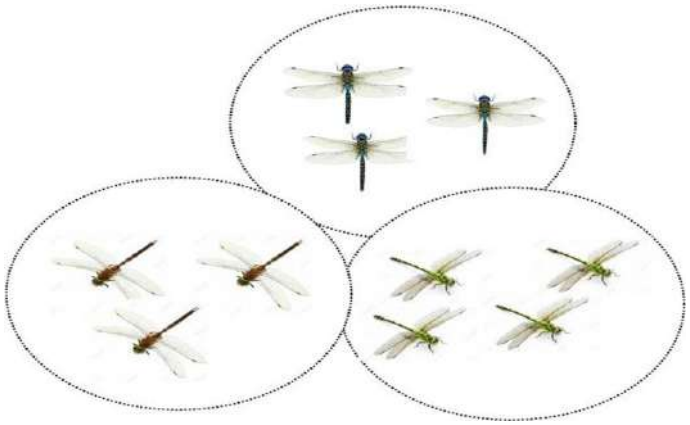


Figure .II.22. Dynamic dragonfly swarms.



Figure .II.23. Static dragonfly swarms.

Naturally, the instinct of each individual in the swarm imposes to attract to the nurturing sources and distract outward enemies. From these two conducts, the DFO algorithm is inspired. The position updating of each individual in the swarm is represented in **Figure .II.24.** and mathematically explained as follows [64].

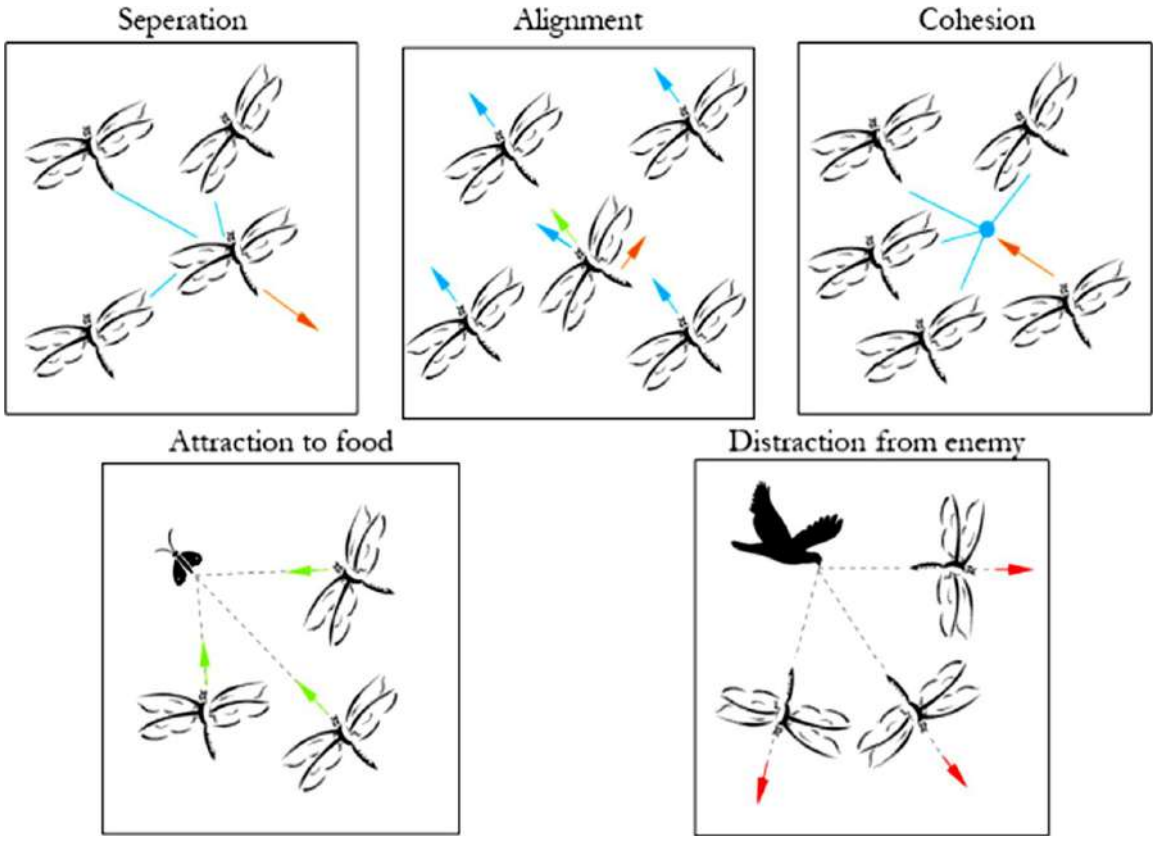


Figure .II.24. Primitive corrective patterns between individuals in a swarm [64].

Separation (S) the aim of this step is to eschew the collision of individuals with their neighbors in the static swarm. This separation is expressed by the following equation.

$$S_i = \sum_{j=1}^{n} X - X_j \qquad \text{II.28}$$

where,

X,$X\,j$ The positions of the current individual and the j^{th} neighboring individual, respectively, and

n The number of neighboring individuals.

Alignment (A) the purpose of this step is to match the velocity of each individual with the other. The alignment is given by

$$A_i = \frac{\sum_{j=1}^{n} V_j}{n} - X \qquad \text{II.29}$$

where,

Vj is the velocity of neighboring individual j .

Cohesion (C) refers to the movement of individuals towards the center of the swarm's group.

$$C_i = \frac{\sum_{j=1}^{n} X_j}{n} - X \qquad \text{II.30}$$

Attraction towards the food (F) All individuals tend to move towards the food.

$$F_i = X^+ - X \qquad \text{II.31}$$

where,

X^+ shows the position of the food source.

Distraction outwards an enemy (E) is calculated as follows:

$$E_i = X^- - X \qquad \text{II.32}$$

where,

X^- shows the position of the enemy.

The position of each dragonfly is updated based on step vector ΔX, which is calculated as follows.

$$\Delta X_i = (sS_i + aA_i + cC_i + fF_i + eE_i) + w\Delta X_t \qquad \text{II.33}$$

where,

s, *a*, & *c*	Represent, respectively, the separation, alignment and cohesion weights,
f & *e*	The food and the enemy factors, respectively,
w	The inertia weight, and
t	The iteration counter.

The updated position vector is calculated as follows.

$$X_{t+1} = X_t + \Delta X_{t+1} \qquad \text{II.34}$$

In the case of no neighboring solutions founded, dragonflies fly around the search space using a random walk, or Lévy flight, to ameliorate their randomness, stochasticity, and exploration. In this case, the dragonflies update their position based on the following equation [62]:

$$X_{t+1} = X_t + Lévy(d) \cdot X_t \qquad \text{II.35}$$

where,

t The current iteration, and

d The dimension of the search space.

Levy flight is given by:

$$Lévy(x) = 0.01 \cdot \frac{r_1 \cdot \sigma}{|r_2|^{\frac{1}{\beta}}} \qquad \text{II.36}$$

where,

r_1 & r_2 The random numbers in the range of [0,1],

β A constant, and

σ Is given by

$$\sigma = \left(\frac{\Gamma(1+\beta) \cdot \sin\left(\frac{\pi\beta}{2}\right)}{\Gamma\left(\frac{1+\beta}{2}\right) \cdot \beta \cdot 2^{\left(\frac{\beta-1}{2}\right)}} \right)^{1/\beta} \qquad \text{II.37}$$

where,

$$\Gamma(x) = (x-1)! \qquad \text{II.38}$$

More details can be found in [64].

II.9. Conclusion

The aim of this chapter is to give a general view of some basic concepts related to the theme and the selected algorithms for modeling our property "the solubility of drug-like compounds in supercritical carbon dioxide". Notice that, the existence of several other not mentioned types of algorithms and in every single algorithm different other not mentioned subtypes. On the other hand, based on the use of each one it can be called or better refer to differently. This latterly gives this domain, informatics, its power, and diversity.

As described above, every programming tool, technique, model ...has its own power that is only seen if it has been correctly used with respect to its limit of operation, and depending on the used data every model can show different results. i.e. When it comes to which method is better to choose to represent any property or phenomena, there is no right answer.

However, for a fixed dataset "try and see" is one of the successful methods to compare those algorithms and give a critical decision on which algorithm can represent the experiments results much better than the rest. This concept has been realized in the upcoming chapter.

Chapter III:

Modeling the Solubility of drug-like compounds in Supercritical carbon dioxide

chapter III: Modeling the Solubility of drug-like compounds in Supercritical Carbon dioxide

III.1.Introduction

In this chapter, we will drop the light on the details and steps used to modeling the solubility of drug-like molecules in SC-CO_2. The approach adopted in this work is based on the use of QSPR descriptors together with two intensive state properties (Operating Temperature and Pressure) as input values and several powerful algorithms that we have already described in the proceeding chapter.

ARTIFICIAL NEURAL NETWORK (ANN) predictive model

For this phase, two ANN sub models have been developed using two different software packages.

A. MLP-ANN "one hidden layer" predictive Model

III.2.Material and methods

III.2.1.Datasets

The collection of the database has been and always keeps being the heart of modeling. For this, it is necessary to select experimental data with consideration to its quality. However, the higher the quality of experimental data is the more we develop high-excellent models. As it is well-known, using unreliable data results in unreliable models.

The mole fraction (y_2) of drug-like compounds solubility in SC-CO_2 has been collected from various literatures with a total of 133 observations. Which were divided using DatasetDivisionGUI v1.2 (https://drive.google.com/file/d/0B3L9EmMFAulZZjJwaWNkc0xZZlE/view) where the Kennard stone method has been used.

Employing appropriate data splitting techniques improves the model prediction. The various approaches available for the division of dataset include k-means clustering, based on X response, based on Y Response, random selection, statistical molecular design, sphere exclusion, Kennard-Stone selection, Kohonen’s self-organizing map selection, and extrapolation-oriented test set selection [43].

Then, the data were divided into 80% of Compounds for the training set and 20% for the test set. The training set is used to formulate the QSAR model, while the test set is used to evaluate its predictability and accuracy. Dataset is generally divided in such a way that both the sets occupy the entire descriptor space.

This previous database has been used without the presence of state variables (T,P) in order to select the best acceptable-fit descriptors. Although finding for all the drug-like compounds a solubility fraction at a fixed temperature and pressure is technically impossible. In the opinion of that, every experimenter has the ultimate freedom and a different vision for experience conditions.

That's why fixing an interval for both the temperature and the pressure can be seen as a necessary solution, an interval of [320-333K] for the temperature and [19-21.5 MPa] for the pressure, where the majority of values have been taken at a T= 328K and a P= 20Mpa.

However, to make sure of having a linear distribution, the output property was converted to the corresponding pLn property "-Ln(y2)" .The normality distribution was checked using statistical tests. The normality distribution plots are presented in **Figure .III.1.** and **.III.2.**

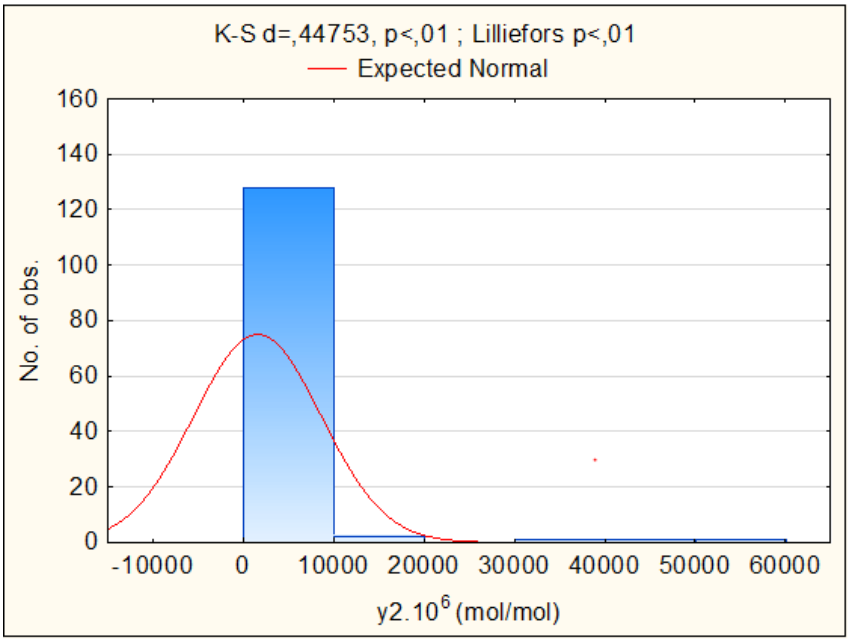


Figure .III.1. Normality distribution plot of mole fraction of solubility data:

(a) before transformation.

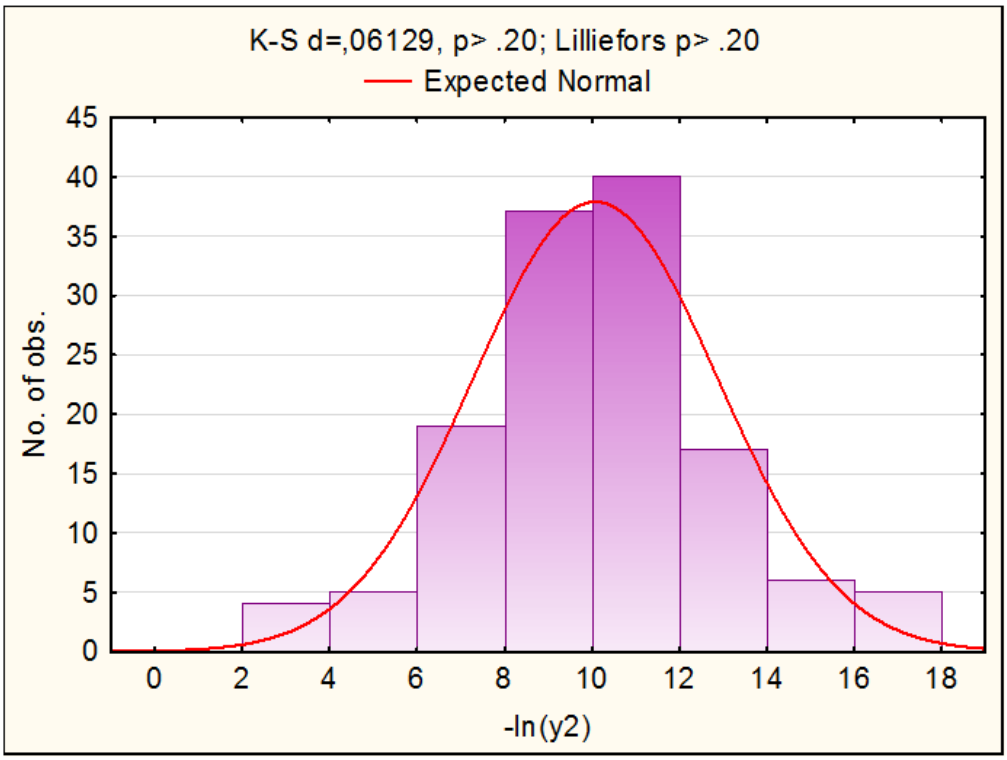


Figure .III.2. Normality distribution plot of mole fraction of solubility data: (b) after conversion (-Ln(y2)).

Those 133 drug-like compounds with experimental data were responsible for collecting a total of 3590 experimental data points (EDP) to form the final database and were divided into two main sets.

The first set with 3490 EDP was retained for QSPR modeling and was randomly split into a training set (n = 2862) and a testing set (n = 628). The second one, included 100 EDP that had been hidden from the QSAR/QSPR model. i.e. was left for the prediction of solid drugs solubility fraction based on the developed mathematical formula.

III.2.2.Descriptor calculation

All the molecular descriptors considered in this study were calculated after conversion of molecule formulas into SMILES notation using an online computer software program called PaDEL (http://www.scbdd.com/padel_desc/index/). The SMILES notations of molecules have been obtained with the aid of Chembio draw software. 1544 molecular descriptors were obtained.

III.2.3.Feature Selection

This step is based on reducing the database horizontally (reduce the number of features). Among the larger number of calculated descriptors, only a few are chosen to define the model. Feature selection after descriptor calculation removes collinearity between the descriptor pairs. The selection of the most appropriate features is done using filter and wrapper methods [44].

Filter methods involve filtering out descriptors, thereby reducing the pool size of descriptors based on inter-variable correlations. Accordingly, molecular descriptors that show inter-correlation are removed retaining only one descriptor from a pair [45]. Descriptors with the lowest variance are also removed [34].

Filter methods use techniques like chi-square analysis, Shannon entropy, odds ratio, GSS coefficient, correlation-based feature selection, Fisher Score, Kolmogorov-Smirnov statistics, and principal component analysis. Distance-based methods like Euclidean distance measures are also grouped under filter methods [34].

The wrapper method uses regression-based approaches to select descriptors. In general, wrapper methods involve more computational power and perform better than filter methods. Recursive feature elimination, variable selection, and modeling based on the prediction, k nearest-neighbor, backward elimination, forward selection, genetic algorithm, Bayesian regularized neural network, factor analysis, and combinatorial protocol are some of the commonly used wrapper methods [34]. Hybrid methods are also being used that combine both filter and wrapper methods for selecting feature [44]. Heuristic methods based on multiple linear regressions are also used in selecting descriptors for QSAR models [34].

Now, returning back to our database for the objective of avoiding the phenomenon of overfitting, and elimination of the insignificant descriptors. The number of descriptors has been reduced into 124 descriptors as a result of the use of several methods.

Starting by removing features having zeroes values using "by default" *V-WSP algorithm* (https://drive.google.com/file/d/0B1EV9j502nhYSXRiQnJFVEk0TlE/view).

Then, invariant descriptors, namely those with absent values or with nearly the same values were manually removed. Also, any descriptor that had identical values for 75% of the samples and any descriptors with a relative standard deviation <0.05 were removed. Finally, half of the descriptors showing an absolute value of the Pearson correlation coefficient >0.95 were also removed.

III.2.4. Model development

QSPR models are developed using one or more statistical model building tools, which may be broadly categorized into regression- and classification-based approaches. Machine learning methods, which use artificial intelligence, may also be useful for predictive model development [35]. Note that the latter may also be used for regression and classification problems.

QSAR modeling AKA, obtain the correlation between features and response can be done for both linear and non-linear properties. Interpretation of the model requires knowledge about the behavior of features and the response [34,36].

Since selecting a suitable mathematical tool for identifying linear or nonlinear behavior is not a trivial task. We distinguish, Multiple Linear Regression (MLR) and Partial Least-Squares (PLS) as the most efficient methods for predicting linear behavior, whereas neural networks and support vectors for predicting nonlinear behaviour [36].

Among several models that have been proposed to predict drug-like compounds solubility and due to the consideration of the intrinsically non-linear nature of QSPR, it is advisable to use flexible methods to model any non-linear relationships, Multilayer perceptrons (MLP) has been chosen.

MLP, are feed-forward multilayer networks that provide flexible frameworks for non-linear function estimation. It consists of an input and an output layer with one or more hidden layers of nonlinearly-activating nodes or sigmoid nodes. This is determined by the weight matrix and it is necessary to adjust the weights of the network [48, 47].

The back-propagation employs gradient fall to minimize the squared error between the network output values and desired values for those outputs. The weights adjusted by an iterative process of the repetitive presence of attributes. Small changes in the weight to get the desired values are done by the process called training the net and is done by the training set (learning rule) [47].

The model considered in this work consists of one input layer (temperature, pressure, and descriptors), one hidden layer and one output layer (solubility). Each layer contains a various number of neurons, which are connected linearly by weights with each other neighboring layers. ANN calculations were performed by the STATISTICA software (STATISTICA 10.0, Tulsa; StatSoft, Inc, OK, USA.). This is for studying the structure-property relationship of various drug-like compounds. A set of various numbers of features was used as input parameters of the network.

III.2.5. Model validation

Acquiring a good quality QSAR/QSPR model depends on many factors, such as the quality of data, the choice of descriptors, variable selection, statistical methods, and validation. However, it is now being accepted that validation is the most crucial concept for the development and application of any QSAR/QSPR model. The validation process confirms the reliability of the developed model along with the acceptability of each step during model development such as assessing the quality of input data, dataset diversity, predictability on external set, the domain of applicability and mechanistic interpretability in addition to statistical assessment [50, 51].

"Validation is one of those words...that is constantly used and seldom defined".

A. R. Feinstein, 1996.

The meaning of this term refers to strategies used to check the reliability of the developed models for their possible application on a new set of data, and confidence of prediction can thus be judged [49].

Based on that, after developing QSAR/QSPR models, it is important to validate any fitted models to check that it is plausible that its predictions will carry over to fresh data not used in the model fitting exercise [49]. Hence, a QSAR/QSPR model can lend to the false prediction of response if the developed model is not validated correctly [50].

Validation methods are needed to establish the predictiveness of a model on unseen data and to help determine the complexity of an equation that the amount of data justifies [50]. Depending on the used datasets, validation can be divided into two types. An **internal method**, case of using data that created the model (training sets) and an **external method**, case of using a separate dataset (test sets). The best method of validating a model is an external method.

Several parameters can be used to check the performance of a QSAR/QSPR model. The chosen ones and its formula are given down below [52]:

$$RMSE = \sqrt{\frac{\sum(\mathrm{Yobs}-\mathrm{Ypred})^2}{\mathrm{n}}} \qquad \text{III.1}$$

$$R = r = \frac{\sum_{i=1}^{n}\left(y_i^{obs}-\bar{y}^{obs}\right)\cdot\left(y_i^{pred}-\bar{y}^{pred}\right)}{\sqrt{\sum_{i=1}^{n}\left(y_i^{obs}-\bar{y}^{obs}\right)^2}\cdot\sqrt{\sum_{i=1}^{n}\left(y_i^{pred}-\bar{y}^{pred}\right)^2}} \qquad \text{III.2}$$

$$R^2 = \mathrm{r}^2 = 1 - \frac{\Sigma(\mathrm{Y_{obs}} - \mathrm{Y_{pred}})^2}{\Sigma(\mathrm{Y_{obs}} - \bar{\mathrm{Y}}_{\mathrm{obs}})^2} \quad \text{III.3}$$

$$Q^2 = 1 - \frac{\Sigma(\mathrm{Y_{obs(training)}} - \mathrm{Y_{pred(training)}})^2}{\Sigma(\mathrm{Y_{obs(training)}} - \bar{\mathrm{Y}}_{\mathrm{obs(training)}})^2} \quad \text{III.4}$$

$$Q^2_{pred} = \mathrm{R}^2_{\mathrm{p}} = 1 - \frac{\Sigma(\mathrm{Y_{obs(test)}} - \mathrm{Y_{pred(test)}})^2}{\Sigma(\mathrm{Y_{obs(test)}} - \bar{\mathrm{Y}}_{\mathrm{obs(training)}})^2} \quad \text{III.5}$$

$$k = \frac{\Sigma(\mathrm{Y_{obs}} \cdot \mathrm{Y_{pred}})}{\Sigma(\mathrm{Y_{pred}})^2} \quad \text{III.6}$$

$$k' = \frac{\Sigma(\mathrm{Y_{obs}} \cdot \mathrm{Y_{pred}})}{\Sigma(\mathrm{Y_{obs}})^2} \quad \text{III.7}$$

$$r_0^2 = 1 - \frac{\Sigma(\mathrm{Y_{obs}} - \mathrm{kY_{pred}})^2}{\Sigma(\mathrm{Y_{obs}} - \bar{\mathrm{Y}}_{\mathrm{obs}})^2} \quad \text{III.8}$$

$$r'^2_0 = 1 - \frac{\Sigma(\mathrm{Y_{pred}} - \mathrm{k'Y_{obs}})^2}{\Sigma(\mathrm{Y_{pred}} - \bar{\mathrm{Y}}_{\mathrm{pred}})^2} \quad \text{III.9}$$

$$r_m^2 = \mathrm{r}^2\left(1 - \sqrt{\mathrm{r}^2 - \mathrm{r}_0^2}\right) \quad \text{III.10}$$

$$r'^2_m = \mathrm{r}^2\left(1 - \sqrt{\mathrm{r}^2 - \mathrm{r}'^2_0}\right) \quad \text{III.11}$$

$$\bar{r}_m^2 = \frac{(r_m^2 + r'^2_m)}{2} \quad \text{III.12}$$

$$\Delta r_m^2 = \left[r_m^2 - r'^2_m\right] \quad \text{III.13}$$

$$\bar{Y}_{obs} = \frac{\sum Y_{iobs}}{n} \quad \text{III.14}$$

$$\bar{Y}_{pred} = \frac{\sum Y_{ipred}}{n} \quad \text{III.15}$$

$$A = (r_0^2 - r'^2_0) \quad \text{III.16}$$

$$M = \frac{(r_0^2 - r'^2_0)}{r_2} \quad \text{III.17}$$

$$N = \frac{(r_2 - r_0^2)}{r_2} \quad \text{III.18}$$

$$PRESS = \sum_{i=1}^{n} (Y_{i\,obs} - Y_{i\,pred})^2 \quad \text{III.19}$$

$$SSY = \sum_{i=1}^{n} Y_{i\,pred}{}^2 \quad \text{III.20}$$

where,

r Correlation coefficient between the observed and predicted value of compounds with interception,

R^2 Squared correlation coefficient between the observed and predicted value of compounds with intercept in case of linear regression. where its well-known by Coefficient of determination,

$RMSE$ Root-Mean Squared Error,

Q^2 Cross-validated correlation coefficient.(also notated CV,q^2),

R_p^2 The degree of correlation between the observed and predicted activity data,

Y_{obs} Observed (experimental) value of Y,

Y_{pred} Predicted (calculated) Y-value of training set, test set or validation set,

n Number of compounds in the data set (training, test...),

$\bar{Y}_{obs}$ Average of Y_{obs} ,

$\bar{Y}_{pred}$ Average of Y_{pred},

r_0^2 Squared correlation coefficient between the observed and predicted value of compounds without intercept,

r'^2_0 It bears the same meaning as r_0^2 , but uses the reversed axes,

k & k' slopes of the corresponding regression lines,

PRESS Predictive Residual Sum of the Squares, and

SSY Sum of the Squares of the response values of the total observations.

These are statistical methodologies used to ensure the model created is sound and unbiased ("good model"). However, that only can assure by comparing them to its acceptability criteria values, which will be represented in the form of a **Table.III.1.**

Table .III.1 Acceptability criteria (A.C) of model [50].

Parameter	M	N	R^2	Q^2	Q^2_{pred}	K	k'	$\bar{r}^2_m$	Δr^2_m	A
A.C value	<0.1		>0.6	>0.5	>0.6	0.85<k/k<1.15'		>0.5	<0.2	<0.3

In addition to :

- R^2 -Q^2 indicate if the model is Overfitting or not. It should not be more than 0.3.
- PRESS/SSY smaller than 0.4, indicates that the QSAR model is reasonable. When a value of this ratio is smaller than 0.1, this indicates an excellent model, and will prove that the developed models predict better than chance. More detail can be found elsewhere [50,52,53].

III.2.6. Applicability domain

The use of modeling to predict unknowing property can be seen as a meaningful step, where a programmer contributes with that knowledge to reduce many deficits. But, like any method, rule, function... a model also has its own limits. Even the most comprehensive and validated models cannot predict reliably properties for all existing compounds, in another words any model has an applicability domain (AD).

The applicability domain is a theoretical region in the space defined by the descriptors of the model and the modeled response [52]. It enables to understand the feasibility of activity or response prediction by the constructed QSAR model for a given set of compounds. So, the QSAR model prediction for a set of compounds is only reliable if the chemical space or applicability domain of those compound falls within the applicability domain of the compounds used for training the model [34].

In the development of a QSAR/QSPR model, the domain of applicability of the molecules plays a critical role in estimating the uncertainty in the prediction of a particular molecule based on how similar it is to the compounds used to build the model [53].

Thus the prediction of a modeled response using QSAR/QSPR is valid only if the compound being predicted falls within the domain of applicability of the model "inside the chemical space covered by the training set" since it is impossible to predict an entire universe of chemicals using a single QSAR/QSPR model [52,53].

Application domain assessment is done through probability density distribution, geometrical methods, distance-based methods, ranges in descriptor space when descriptor space is used, and the range of response variable when modeled response space is used [34].

To visualize the applicability of the domain of our developed models William's plot was used. It is a plot of standardized residuals versus leverage values, Hat diagonal values (hi), and can be used for an immediate and simple graphical detection of both the response outliers i.e., compounds with standardized residuals greater than two standard deviation units (Y outliers) and structurally influential chemicals (X outliers) in the model, i.e., hi>h*.

In this plot, the applicability of domain is established inside a squared area within ± 3 standard deviations and a threshold leverage value (h*). The leverage (h) value of a compound in the original independent variable space is defined as follows [54]:

$$hi = x_i(X^T X)^{-1} x_i \qquad \text{III.21}$$

where,

x_i Descriptor-row vector of the query compound, and

X k x n matrix containing the k descriptor values for each one of the n training compounds.

The warning leverage h* is, generally,

$$h^* = \frac{3(p+1)}{n} \qquad \text{III.22}$$

where,

n Total number of samples in the training set, and

p Number of descriptors involved in the correlation.

When h value of a molecule is lower than h*, the probability of accordance between calculated and experimental values is as high as that of the molecules in the training set compounds. Prediction must be considered unreliable for compounds with a high leverage value (hi>h*). Compounds with hi>h* hardly affect the goodness of fit of the developed model, but these compounds may not be an outlier because of low residual.

It must be noted that compounds with high value of leverage and good fitting in the develop model can stabilize the model. On the other hand, compounds with bad fitting in developed model may be outliers. Thus the leverage and the standardized residual should be used simultaneously for the description of the applicability of domain of the expanded model.

III.3. Results and discussion

III.3.1.Selection of relevant descriptors

Despite the fact that the number of descriptors has been reduced from 1544 into 124 descriptors, " total number of EDP/number of features is > 5, condition viable ", for a QSPR model it keeps seen as a large number. With the aim of reducing more the number to the smallest significant number able to predict the chosen property with an acceptable error, the genetic algorithm method was applied (more specify, GA-MLR), the selection programming tool is GeneticAlgorithm_v4.1 (https://sites.google.com/site/dtclabgatool/).

The algorithm that we had mentioned earlier takes as inputs the training set file (80% of 133 databases) with some statistic values; The variance cut-off which we had it taken by-default (0.001) and the inter-correlation cut-off. Various inter-correlation have been tested 0.75, 0.8, 0.85, and 0.9 were the comparison shows a negligible difference between them, so an inter-correlation of 0.9 has been then taken. For the rest, inputs were taken by-default.

For the output, one of the benefits of GA is the opportunity to fix the number of features (Equation Length), the selection of the appropriate number of descriptors was carried out over an interval of 3 to 17 descriptors.

Those 14 models were examining their statistical performance after adding both the temperature and the pressure data. As a reminder, a total value of 3590 experimental data points (n) was reached with various features number (m) from 5 to 19. "Five features reference to three descriptors, one Temperature, and one pressure for each EDP".

The statistical results of 14 models are represented in the **Table .III.2.** (training set) and **Table .III.3.** (test set) down below.

Table .III.2. Statistical parameters for internal validation.

Training set						
Number of descriptors	R	R^2	RMSE	Q^2	$\bar{r}_m^2$	Δr_m^2
5	0.9395	0.8827	0.8640	0.8827	0.8290	0.1067
6	0.9690	0.9389	0.6235	0.9389	0.9091	0.0588
7	0.9726	0.9460	0.5865	0.9460	0.9202	0.0511
8	0.9829	0.9661	0.4644	0.9661	0.9498	0.0322
9	0.9851	0.9704	0.4340	0.9704	0.9546	0.0291
10	0.9869	0.9739	0.4076	0.9739	0.9615	0.0246
11	0.9882	0.9765	0.3867	0.9765	0.9649	0.0225
12	0.9892	0.9784	0.3706	0.9784	0.9670	0.0214
13	0.9909	0.9819	0.3391	0.9819	0.9732	0.0174
14	0.9926	0.9852	0.3069	0.9852	0.9779	0.0143
14b	0.9905	0.9811	0.3477	0.9810	0.9716	0.0176
15	0.9906	0.9814	0.3443	0.9814	0.9724	0.0177
17	0.9928	0.9856	0.3004	0.9856	0.9785	0.0137
19	0.9939	0.9878	0.2791	0.9878	0.9821	0.0112

This table shows that results above conform to the acceptability criteria. Noted that the more R^2 is closer to 1 the more the model is accurate. Also, the RMSE is used to assess the predictive ability and accuracy of the model. However, models with small RMSE values yield more highly reliable predictions.

Test set											
Number of descriptors	R	R^2	RMSE	Q^2_{pred}	k	k'	$\bar{r}^2_m$	Δr^2_m	A	M	N
5	0.9400	0.8836	0.9050	0.9823	1.0004	0.9921	0.8312	0.1019	0.0137	0.0155	3.0252e-06
6	0.9705	0.9418	0.6402	0.9911	0.9990	0.9972	0.9146	0.0471	0.0029	0.0031	1.5313e-05
7	0.9723	0.9453	0.6202	0.9917	0.9976	0.9988	0.9156	0.0419	0.0030	0.0031	9.1147e-05
8	0.9854	0.9709	0.4528	0.9956	0.9969	1.0012	0.9525	0.0133	8.0907e-04	8.3335e-04	1.5003e-04
9	0.9874	0.9750	0.4199	0.9962	1.0002	0.9982	0.9604	0.0273	8.3578e-04	8.5726e-04	8.5205e-07
10	0.9875	0.9750	0.4194	0.9962	0.9980	1.0004	0.9572	0.0203	8.8372e-04	9.0638e-04	6.3654e-05
11	0.9891	0.9783	0.3906	0.9967	0.9983	1.0003	0.9648	0.0139	4.7995e-04	4.9058e-04	4.5639e-05
12	0.9897	0.9795	0.3800	0.9969	0.9990	0.9997	0.9672	0.0169	4.6364e-04	4.7335e-04	1.5397e-05
13	0.9904	0.9808	0.3673	0.9971	0.9986	1.0002	0.9716	0.0073	2.0366e-04	2.0764e-04	3.2485e-05
14	0.9924	0.9849	0.3264	0.9977	1.0008	0.9983	0.9742	0.0155	3.5782e-04	3.6332e-04	8.9339e-06
14b	0.9907	0.9815	0.3614	0.9972	0.9980	1.0008	0.9687	0.0103	3.9284e-04	4.0026e-04	6.1540e-05
15	0.9899	0.9799	0.3757	0.9969	0.9990	0.9997	0.9698	0.0128	2.9839e-04	3.0449e-04	1.4772e-05
17	0.9901	0.9801	0.3856	0.9967	0.9992	0.9994	0.9635	0.0275	9.6305e-04	9.8264e-04	8.3535e-06
19	0.9931	0.9863	0.3105	0.9979	0.9991	1.0000	0.9778	0.0100	1.9958e-04	2.0235e-04	1.2634e-05

Table .III.3. Statistical parameters for external validation.

Some remarks that can be resumed from the tables are given:

- The value of RMSE, Δr_m^2 , A, M, and N gives a negative correlation with the number of descriptors; The more the number of descriptors increases, more those values decreases.
- The values of k and k' are technically constant (≈1) which means that the number of descriptors has no influence on those two parameters.AKA, no correlation.
- For the rest R, R^2, Q^2, Q_{pred}^2, and $\bar{r}_m^2$ the number of descriptors with them follows a positive correlation; The more the number of descriptors increases, more those values parameters increase.

As a general guess, models with a combination of 13 to 17 features are more candidates for better prediction. In order to make a critical decision, more statistic parameters were introduced. Concerning the notation **14b** more detail will be given later.

Multi-collinearity between the descriptors of the selected models 13, 14, 15, and 17 were detected by calculating their variation inflation factors (VIF) using XLSTAT software. A model is called acceptable If the VIF falls into the range of 1–5. The results are represented from **Table .III.4.** to **. III.7.**

Table .III.4. Statistical of multi-collinearity of 13 descriptors.

Descriptors	**T(k)**	**P(Mpa)**	**AATS4v**	**GATS2m**	**GATS1e**	**GATS3s**	**SpMin6_Bhe**
Tolerance	0.969	0.952	0.250	0.687	0.580	0.721	0.254
VIF	1.032	1.051	3.993	1.455	1.723	1.386	3.939
Descriptors	**AVP-0**	**ETA_dEpsilon_B**		**ETA_Shape_Y**	**ETA_Beta**	**FMF**	**MLFER_S**
Tolerance	0.394	0.233		0.699	0.179	0.542	0.315
VIF	2.539	4.297		1.430	**5.599**	1.846	3.179

Table .III.5. Statistical of multi-collinearity of 14 descriptors.

Descriptors	**T(k)**	**P(Mpa)**	**AATS3v**	**MATS2e**	**GATS4c**	**GATS3v**	**GATS4e**
Tolerance	0.971	0.963	0.368	0.822	0.350	0.767	0.315
VIF	1.030	1.038	2.719	1.217	2.858	1.303	3.173
Descriptors	**GATS3s**	**nBondsM**	**AVP-0**	**SHBd**	**ETA_dEpsilon_B**	**MLogP**	**MLFER_S**
Tolerance	0.741	0.246	0.425	0.531	0.184	0.380	0.404
VIF	1.350	4.059	2.354	1.884	**5.449**	2.630	2.474

Table .III.6. Statistical of multi-collinearity of 15 descriptors.

Descriptors	**T(k)**	**P(Mpa)**	**AATS4v**	**ATSC2m**	**AATSC2v**	**GATS3v**	**GATS1e**	**GATS3s**
Tolerance	0.967	0.949	0.118	0.318	0.290	0.290	0.466	0.630
VIF	1.034	1.053	**8.452**	3.144	3.452	3.425	2.309	1.588
Descriptors	**AVP-0**	**Mp**	**ETA_dEpsilon_B**		**ETA_BetaP**	**MLogP**	**MLFER_E**	**JGT**
Tolerance	0.145	0.106	0.092		0.070	0.353	0.293	0.202
VIF	**6.878**	**9.390**	**10.897**		**14.317**	2.831	3.416	4.953

Table .III.7. Statistical of multi-collinearity of 17 descriptors.

Descriptors	**T(k)**	**P(Mpa)**	**AATS5v**	**ATSC7c**	**AATSC0p**	**GATS4e**	**GATS2s**	**SpMax3_bhm**	**ASP-7**
Tolerance	0.979	0.953	0.200	0.903	0.709	0.733	0.730	0.089	0.378
VIF	1.022	1.050	**5.009**	1.108	1.410	1.364	1.370	**11.286**	2.648
Descriptors	**nssCH2**	**ETA_dEpsilon_D**		**SIC1**	**CIC0**	**ZMIC5**	**piPC9**	**N5HeteroRing**	**JGl3**
Tolerance	0.280	0.471		0.257	0.124	0.310	0.111	0.670	0.584
VIF	3.576	2.125		3.894	**8.061**	3.230	**9.031**	1.492	1.713

Notice that all the selected models have at least one VIF value bigger than 5, that is to say, none of them is acceptable. Also, the more the number of descriptors rises the more the number of out of range VIF values grows. On the other hand, reducing the number of descriptors enhances more the cost errors.

Between those two constraints, various ideas have been tested. Where the best-fit combination of descriptors, 14b, which also checks both of constraints its VIF values are given down below.

Table .III.8. Statistical of multi-collinearity of 13 descriptors (14b).

Descriptors	**T(k)**	**P(Mpa)**	**AATS3v**	**MATS2e**	**GATS4c**	**GATS3v**	**GATS4e**
Tolerance	0.983	0.972	0.532	0.822	0.349	0.769	0.316
VIF	1.017	1.029	1.881	1.216	2.861	1.300	3.164
Descriptors	**GATS3s**	**nBondsM**		**AVP-0**	**SHBd**	**MLogP**	**MLFER_S**
Tolerance	0.741	0.401		0.471	0.568	0.427	0.490
VIF	1.349	2.494		2.125	1.761	2.341	2.041

All VIF values of the descriptors are lower than 3.164, indicating that the obtained model has statistical significance, and the descriptors were found to be reasonably orthogonal. In order to get more details on the correlation between the selected descriptors, the correlation matrix has been established using the XLSTAT software. The result is shown in **Table .III.9.**

The value of the correlation coefficient of each pair of selected descriptors was ≤ 0.590 with one exception value equals to 0.758. This result proves that the selected descriptors were independent.

Therefore, the selection descriptors are as follows: AATS3v, MATS2e, GATS4c, GATS3v, GATS4e, GATS3s, nBondsM, AVP-0, SHBd, MLogP, and MLFER_S. With the addition of T(k) and P(Mpa).

More details on the characterization of the selected descriptors can be found in **Table .III.10.**

Table .III.9. Matrices of correlation of the selected features.

	T(k)	P(Mpa)	AATS3v	MATS2e	GATS4c	GATS3v	GATS4e	GATS3s	nBondsM	AVP-0	SHBd	MLogP	MLFER_S	pLn Property
T(k)	**1**	0.004	0.040	-0.063	-0.083	-0.035	-0.057	0.023	0.022	0.037	-0.060	0.034	0.002	-0.039
P(Mpa)	0.004	**1**	0.061	-0.018	-0.013	-0.047	-0.031	-0.111	0.076	0.003	-0.023	0.096	0.102	-0.225
AATS3v	0.040	0.061	**1**	0.016	0.023	-0.266	0.046	-0.128	0.528	-0.321	-0.138	-0.061	0.365	0.217
MATS2e	-0.063	-0.018	0.016	**1**	0.080	0.148	0.100	-0.283	-0.058	-0.202	0.175	-0.209	-0.094	-0.062
GATS4c	-0.083	-0.013	0.023	0.080	**1**	-0.022	**0.758**	0.172	-0.015	0.013	-0.080	0.168	0.057	-0.202
GATS3v	-0.035	-0.047	-0.266	0.148	-0.022	**1**	0.098	-0.056	-0.426	0.139	-0.073	-0.109	-0.278	-0.146
GATS4e	-0.057	-0.031	0.046	0.100	**0.758**	0.098	**1**	0.168	-0.122	-0.112	-0.029	-0.384	-0.199	-0.126
GATS3s	0.023	-0.111	-0.128	-0.283	0.172	-0.056	0.168	**1**	0.017	0.210	-0.015	-0.099	-0.023	-0.242
nBondsM	0.022	0.076	0.528	-0.058	-0.015	-0.426	-0.122	0.017	**1**	-0.313	0.003	0.252	0.590	0.008
AVP-0	0.037	0.003	-0.321	-0.202	0.013	0.139	-0.112	0.210	-0.313	**1**	-0.456	0.433	-0.046	-0.216
SHBd	-0.060	-0.023	-0.138	0.175	-0.080	-0.073	-0.029	-0.015	0.003	-0.456	**1**	-0.407	0.016	0.180
MLogP	0.034	0.096	-0.061	-0.209	0.168	-0.109	-0.384	-0.099	0.252	0.433	-0.407	**1**	0.377	0.039
MLFER_S	0.002	0.102	0.365	-0.094	0.057	-0.278	-0.199	-0.023	0.590	-0.046	0.016	0.377	**1**	0.250
pLn Property	-0.039	-0.225	0.217	-0.062	-0.202	-0.146	-0.126	-0.242	0.008	-0.216	0.180	0.039	0.250	**1**

Table .III.10. Characteristics of the descriptors selected in the optimal model.

Descriptors Name	Type	Class	Description "Physical-chemical meaning"	VIF
AATS3v	2D	Auto correlation	Average Broto-Moreau autocorrelation - lag 3 / weighted by van der Waals volumes	1.881
MATS2e	2D	Auto correlations	Moran autocorrelation of lag 2 weighted by Sanderson electro-negativity	1.216
GATS4c	2D	Auto correlations	Geary autocorrelation - lag 4 / weighted by charges	2.861
GATS3v	2D	Auto correlations	Geary autocorrelation of lag 3 weighted by van der Waals volume	1.300
GATS4e	2D	Auto correlations	Geary autocorrelation of lag 4 weighted by Sanderson electro-negativity	3.164
GATS3s	2D	Auto correlations	Geary autocorrelation of lag 3 weighted by I-state	1.349
nBondsM	2D	Bond Count	Total number of bonds that have bond order greater than one (aromatic bonds have bond order 1.5).	2.494
AVP-0	2D	Chi Path	Average valence path, order 0	2.125
SHBd	2D	Electro-topological State Atom	Sum of E-States for (strong) hydrogen bond donors	1.761
MLogP	2D	Mannhold LogP	Moriguchi octanol-water partition coeff. (logP)	2.341
MLFER_S	2D	Molecular linear free energy relation " MLFER"	Combined dipolarity/polarizability	2.041

III.3.2. ANN-QSPR Modeling

For this study, with the use of the selected 13th early descript descriptors, a set of 3490 EDP has been taken to build this MLP model (one input, hidden, and output layer). Where it has been divided into a random sample of 82%/18% for both the training set and the test set respectively.

The net was trained using the training function BFGS quasi-Newton methods algorithm by-default. The tanh (hyperbolic tangent) function was chosen as a transfer function for the hidden layer and an identity function for the output layer. The number of hidden neurons was optimized by trial and error procedure in the training process. One output neuron was used to represent the observed solubility fraction(y_2).

To optimize the number of nodes in the hidden layer, several calculations were performed with different numbers of hidden nodes (1–30). This network consisted of 13 inputs and one output for y_2. Therefore, the MLP-ANN has the following architecture {13-20-1}.

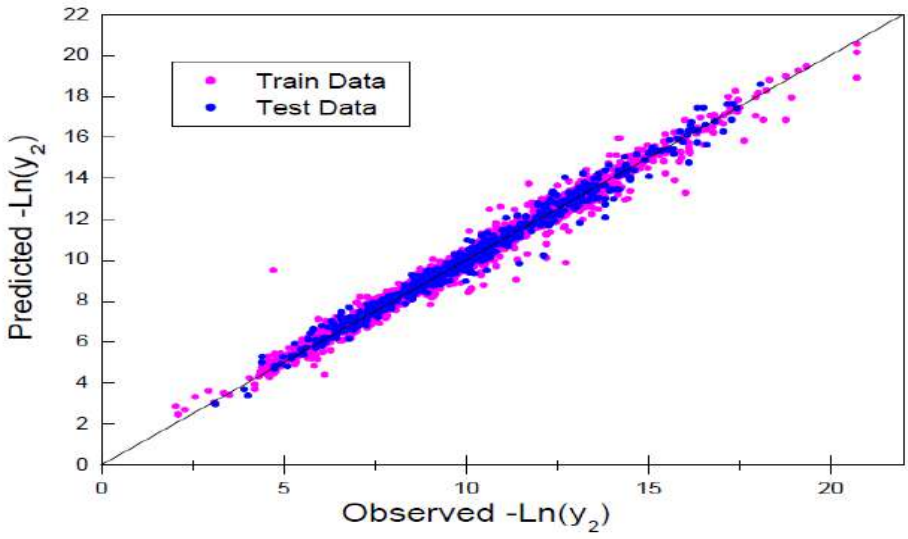


Figure .III.3. The dispersion diagram of the molar solubility fraction, observed vs. predicted for both the training and test sets. Model BFGS, tanh.

Figure .III.3 shows the regression line of the model equation, i.e. predicted vs. observed results for the training, and test set highlighted by different colors. A close correlation between the values predicted by the ANN model and the observed values of the mole solubility fraction was found with the exception of some out-case points, we can say accepted outrange.

For more performance judgment to validate this model, we offer the following statistics:

Table .III.11 Validation parameters for the train, test, and global sets for MLP-ANN(one hidden layer) model.

Statistical values	Train	Test	Global
Q^2_{F1}	0.9810	0.9816	0.9811
Q^2_{F2}	0.9811	0.9815	0.9811
Q^2_{F3}	0.9810	0.9795	0.9807
Q^2_{CCC}	0.9904	0.9906	0.9905
R^2	0.9811	0.9815	0.9811
R	0.9905	0.9907	0.9905
k	1.0001	0.9980	0.9997
k'	0.9987	1.0008	0.9991
R_0^2	1.0000	0.9999	1.0000
R'_0^2	1.0000	1.0000	1.0000
m	-0.0194	-0.0188	-0.0193
n	-0.0193	-0.0193	-0.0189
R_m^2	0.8458	0.8481	0.8481
Af	1.0244	1.0263	1.0248
Bf	1.0007	1.0030	1.0011

Table .III.12. Different errors for the train, test, and global sets for MLP-ANN (one hidden layer) model.

Error	Train	Test	Global
AARD%	2.4284	2.6087	2.4609
RMSE	0.3477	0.3614	0.3502
MSQ	0.1209	0.1306	0.1227

The performance remarks that can be drawn from the Tables are given down below:

The non-linear ANN model gave good results with higher correlation coefficients R, as well as better robustness (Q^2), and the coefficients of determination R^2. In addition to the classical validation parameters, that had shown some great results such as the bias factor *Bf* (*Bf* =0.9 – 1.05) indicate that this model is a good one. Also, the error parameters did show too acceptable results.

The calculation of the coefficient of determination minus the cross-validated correlation coefficient is lower than 0.3 "R^2-Q^2= 1E-3 < 0.3". Indicating non overfitting for this model.

The calculated global PRESS value is lower than the SSY. "PRESS/SSY=82.0382/69136.8009=0.0012 <0.1". This proves that the developed models is an excellent model and predict better than chance.

These results indicated that the ANN not only performed well in model development, but also had excellent prediction and this fact suggested that a non-linear correlation between the solubility fraction and the relevant descriptors.

For the classical validation parameters, more detail information, formula in given elsewhere (see **III.5.**).

❖Applicability domain

The applicability domain of the model was analyzed using a Williams plot (**Figure .III.4.**), where the vertical line is the critical leverage value (h*), and the horizontal lines ±3 are the cut off value for Y_{space}.

Figure .III.4 shows a plot of standardized residual (Y-axis) vs. leverage values (X-axis) referred to as the Williams plot.

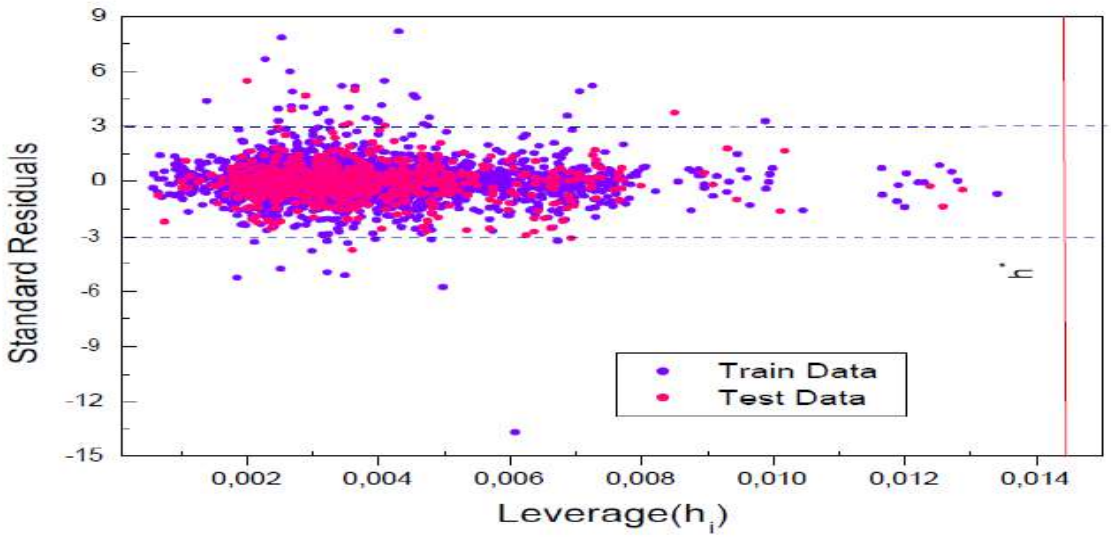


Figure .III.4. Williams plot for MLP-ANN (one hidden layer) model.

In **Figure .III.4**, we can observe that none of the drug-like compounds in the whole dataset have leverage higher than the warning h* value of 0.0147. On the other hand, 2819 EDT belongs to the train set are between the horizontal lines (limit of ± 3) which means that an average of 98.5% of the total training set is covered. Out of the 628 test set, only 10 EDP shows a distribution out of the two horizontal lines (98.41% belong in the AD of the model).

It should be noted that 98.48% of the whole dataset was covered by the model (only 1.35% outranges). Thus, these results show that the MLP-ANN model complies with the third principle of the OECD. Accordingly, the model developed in this study provides excellent predictions for 133 drug-like compounds (3490 EDP). It can be used to predict the molar solubility fraction of those compounds, particularly for those that have not been tested as well as new drugs.

❖ Hidden dataset

The hidden dataset from the model, that included 100 EDP, for the same selected number of features was used for the exploitation of the models. i.e. prediction of the solubility fraction of a not used dataset, known as the validation set. The correlation is shown in **Figure .III.5.**

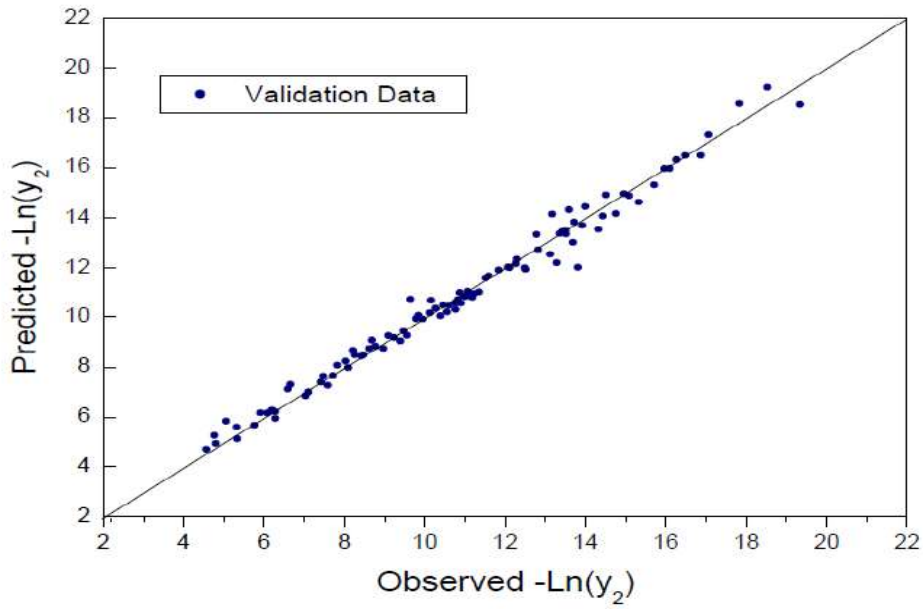


Figure .III.5. The dispersion diagram of the molar solubility fraction, for the validation set, model BFGS, tanh.

The **Figure .III.5** proved that this model can represent the predicted and the observed solubility fraction of validation set with a close positive correlation.AKA, the model is valid for prediction future data.

In order to quantitatively visualize the distribution of the residuals (Y_{obs} - Y_{pred}) vs. the experimental values (observed) of the validation set. We adopted a comparison between this model (13 descriptors) with two others models (models with12 and 14 descriptors).

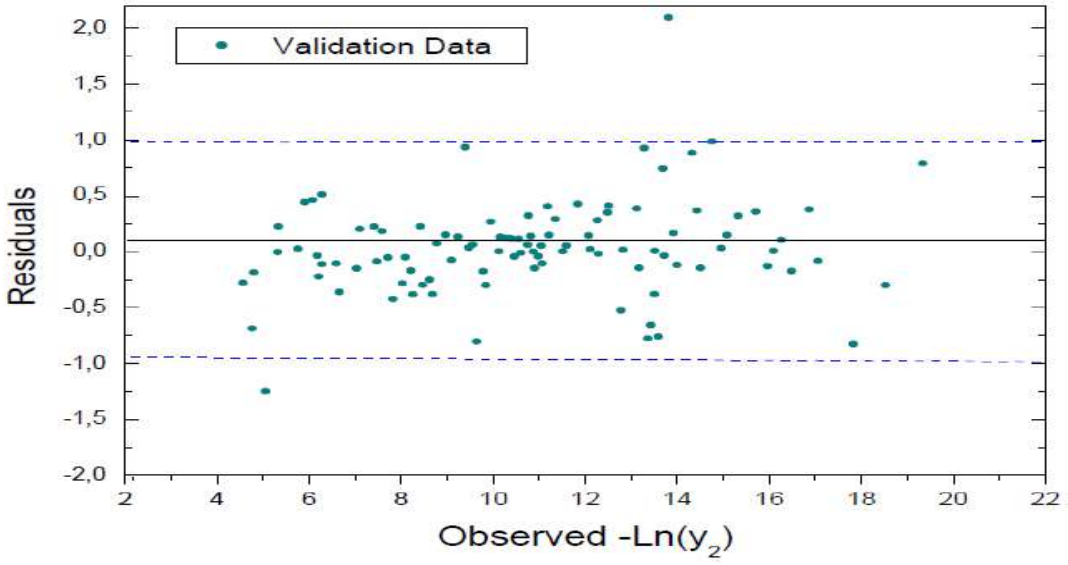


Figure .III.6. Residuals vs. Observed for the validation Data of the selected 12 descriptors (model 12).

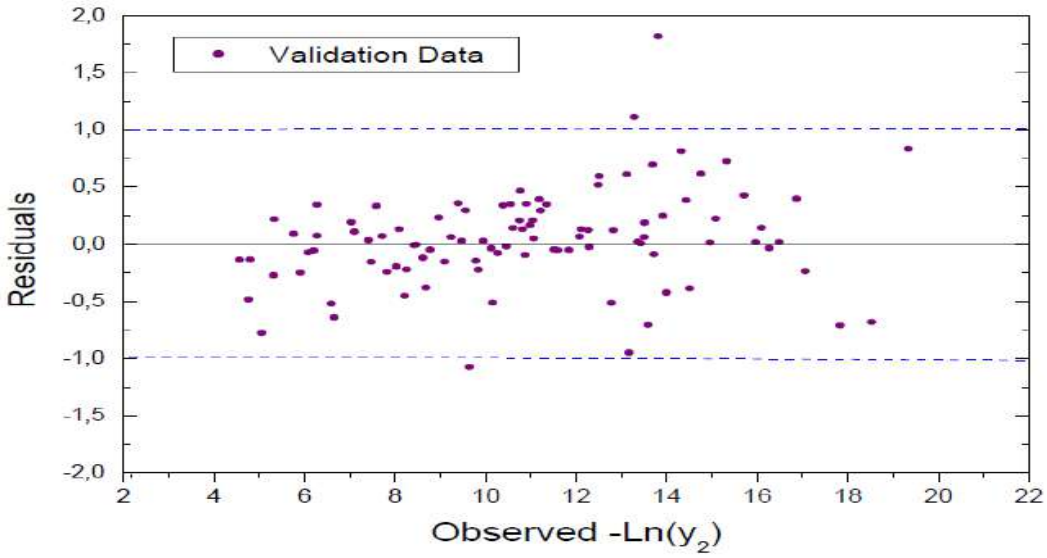


Figure .III.7. Residuals vs. Observed for the validation Data of the selected 13 descriptors (model 14b).

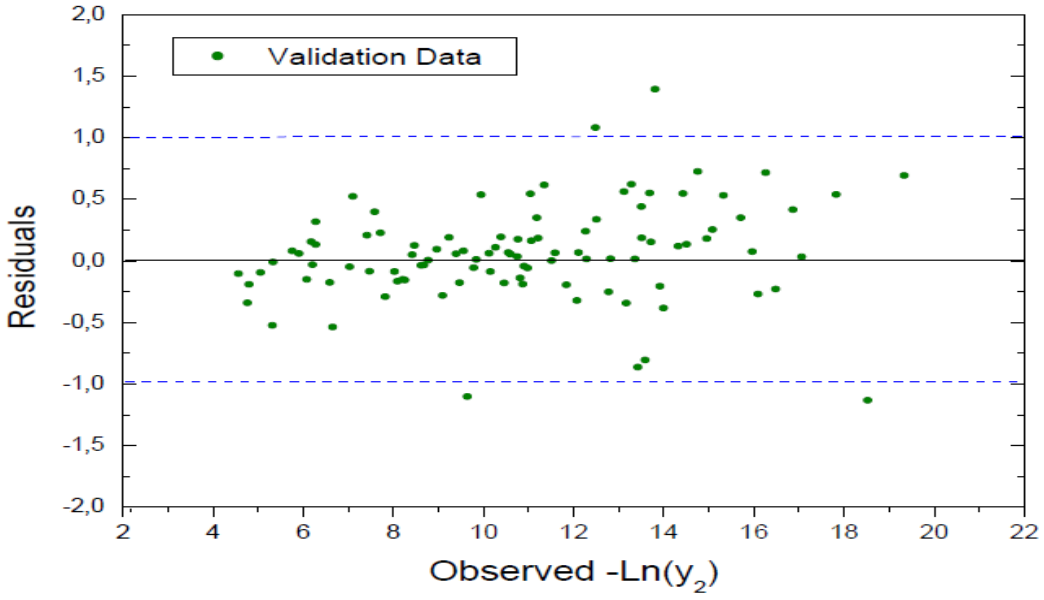


Figure .III.8. Residuals vs. Observed for the validation Data of the selected 14 descriptors (model 14).

From **Figure .III.6** to **.III.8.**

♦ All three models show a residuals value close to zero and lower than the tow bonds ±2 where model 12 shows one out-case (residuals >2). And for the bonds ±1 four out-cases for model 14, and three for both models 12 and 13.

♦ The distribution of data is quite similar for all three models, it did not show a proportional and systematic error, because the distribution of the residuals on both sides of zero was random. Where we can classify them base on the distribution more normalize (closer to the zero line): model 14-13 than 12.

♦ Depending on those two remarks we can over-performed the model 13 on the two others.

❖ Descriptors contribution and interpretation

After the selection of the molecule descriptors, 11 pertinent descriptors with two intensive state variables (pressure and temperature) were used to develop the ANN model, which were nBondsM, GATS4c, AVP-0, MLFER_S, AATS3v, SHBd, MLogP, GATS4e, GATS3s, GATS3v, MATS2e, P and T. The contribution and description of these descriptors is shown in **Table .III.13.**

Table .III.13. Type and contribution of the descriptors selected in the ANN model.

	Descriptor type	Contribution (%)
nBondsM	Bond count	15.19
GATS4c	2D autocorrelations	14.77
AVP-0	2D PaDEL Chi Path Descriptor	14.39
MLFER_S	Molecular linear free energy relation	8.56
AATS3v	2D autocorrelations	7.63
SHBd	Electro-topological state atom type	7.49
MLogP	Molecular properties	7.29
GATS4e	2D autocorrelations	7.19
GATS3s	2D autocorrelations	6.78
GATS3v	2D autocorrelations	6.17
MATS2e	2D autocorrelations	3.67
P	---	0.66
T	---	0.20

To evaluate the scope of each descriptor for the prediction of solid drugs solubility, the relative contributions of the 13 descriptors to the ANN model were determined. The relative contribution level of the descriptors was found to be in the following order: nBondsM, GATS4c, AVP-0, MLFER_S, AATS3v, SHBd, MLogP, GATS4e, GATS3s, GATS3v, MATS2e, P, and T (**Table.III.13.**). In this case, we can see that the variables that have the highest relative contribution are successively nBondsM, GATS4c, and AVP-0.

The 2D descriptor “nBondsM” is defined as the dimensionless total number of bonds that have bond order greater than one (for example aromatic bonds have bond order 1.5) and is encoded to an unsaturation. It referred to as the number of non-kekulized structures. AVP-0 is topological descriptors that give information regarding the connectivity of various atoms in the molecule. This descriptor is able to take into account the presence of heteroatoms in a molecule, as well as double and triple bonds. MLFER_S descriptor is a descriptor for combined dipolarity/polarizability. It is one of the molecular linear free energy relation. SHBd describe the electropological state (E-state)

in different aspects. SHBd represents the sum of E-states for (strong) hydrogen bond donors. An increase in the value of this descriptor means that the molecule is more and more hydrophobic.

The molecular descriptor MLOGP (which is square root of MLOGP2) is considered as a measure of lipophilicity of a molecule. GATS4c, AATS3v, GATS4e, are a 2D auto correlation molecular descriptors which is known respectively as the geary autocorrelation-lag 4/weighted by charges, autocorrelation of lag 3 weighted by vdw volume, geary autocorrelation of lag 4 weighted by atomic Sanderson electro-negativity. GATS3s is 2D auto correlation molecular descriptors which describes how the considered property (atomic masses, polarizability, charges and electro-negativities) is distributed along a topological molecular structure. On the other hand, GATS3v and MATS2e are defined as Geary autocorrelation of lag 3 weighted by van der Waals volume and Moran autocorrelation of lag 2 weighted by Sanderson electro-negativity respectively.

In summary, in this model, it is concluded that the lipophilicity, electro-topology and electronic properties (polarizability, electro-negativity, atomic masses, van der Waals volume, unsaturation, and charges) of a molecule are of major importance in the solid drugs solubility, whereas temperature and pressure does not appear to be a major factor.

The mathematical formula for predicting the molar solubility fraction of drug-like compounds obtained using the ANN approach is given in **Eq.III.25.**

$$\begin{aligned} Ln(y_2) = -(&19.5282 - 9.8738 \cdot 10^{-4} \cdot T(k) - 8.8405 \cdot 10^{-2} \cdot p(Mpa) + \\ &0.0258 \cdot AATS3v - 2.2207 \cdot MATS2e - 4.2687 \cdot GATS4c - \\ &3.2794 \cdot GATS3v + 2.3749 \cdot GATS4e - 2.4007 \cdot GATS3s - \\ &0.2381 \cdot nBondsM - 13.0342 \cdot AVP0 + 0.8605 \cdot SHBd + \\ &0.6915 \cdot MLogP + 1.3707 \cdot MLFER_S \,) \end{aligned} \qquad \text{III.25}$$

Sources and ranges of solubility data of drug-like compounds in supercritical carbon dioxide with the selected feature that had been used to predict this model and the next upcoming model are presented elsewhere (see Annex **Table.1.**).

III.3.3.Conclusion

The aim of this study is to develop an ANN-QSPR model that gives a representation to the molar solubility fraction of drug-like compounds based on their molecular structure and two other intensive state variables (the temperature and the pressure), where we covered the most possible,

from the collection the database, calculation of it descriptors with the use of PaDEL-Descriptor software to modeling and validation of the model.

Thus far, the best model was obtained by the BFGS quasi-Newton algorithm with {13-20-1} network architecture. This built MLP-ANN model was subjected to and approved in several statistical tests currently recommended for the validation of QSAR, which increases her reliability in predicting. All the validations indicate that the built QSAR model was robust and satisfactory.

In conclusion, the model developed in this study meets all of the OECD principles for QSAR validation and can be used to predict the solubility of drug-like compounds, particularly for those that have not been tested as well as new compounds.

B.MLP- ANN "two hidden layer" predictive Model

While building an ANN model there is a main base that should be respected first of all. Represent as follows: the number of nodes of the input layer * the number of nodes of the first hidden layer + the number of nodes of the first hidden layer * the number of the second hidden layer+...+ number of nodes of the n^{th} hidden layer * the number of nodes of the output layer should be lower or equal to the data points used to build that model.

Whilst, for predicting the solubility it is common to use an MLP-ANN prediction model with one input, hidden, and output layer, since it is sufficient to give a good performance. " *Many results seem to confirm that one hidden layer may be enough for most forecasting problems*". On the other hand, the collected data of this study is quite large (13X3590).

Through this, we offer the use of two hidden layers instead of one to break the usually, see its power. Technically speaking, to be one of the fewest developers that built an MLP-ANN with two hidden layers, predictive models, for the molar solubility fraction of drug-like compounds in SC-CO_2.

III.4.Model development

From various ANNs models, the feed-forward algorithm MLP which contains a single input layer, two hidden layers, and one output layer have been chosen. The program has been writing using the programming language MATLAB® R2019b, one of the most successful programming language for solving engineering issues.

A database of 3590 EDP with the selected number of the feature of the previous QSPR model has been directly taken as the input value for this ANN model and for the upcoming models.

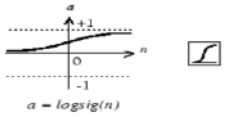
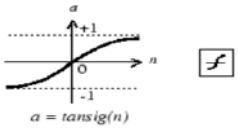
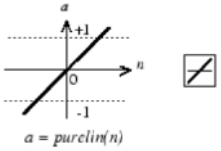
The dataset has been divided using MATLAB® ***dividerand,*** one of the available MATLAB® division functions, into three random sets as shown 70% for the training set and 15% for both the test and validation set.

The programming language MATLAB® provides a remarkable number of training functions for training the network. Such as trainlm, trainscg, traincgf, traincgb, trainbfg, trainnrp, trainxd, and traincgp, also many different kinds of activation functions like hardim, purelin, satlin, logsig, tansig, poslin, tribas, radbas....and so on.

For this study, two types of training functions trainlm "Levenberg-Marquardt back-propagation", and trainbr "Bayesian regularization back-propagation" have been used in collaboration with applying each time one of those two chosen activation function (logsig, and tansig) for the hidden layers and purelin for the output layer, a total number of eight (08) models. The choices of those functions have been done after performing some random tests and checking some online researchers. More information about the activation functions is given in **Table .III.14.**

The choice between those models, select the best network architecture, has significant importance. So, every single model has been developed at the same conditions. i.e. For each model, we had developed a MATLAB® program, where we vary the number of neurons in the hidden layers (13-19 for both hidden layers), the training function, and the activation function of the hidden layers (tansig, and logsig), with a number of 25-30 iterations. In order to select the top three models. Then, we carried out a comparison of their performance. The exploitation of the models is assessed using statistical and graphical tools.

Table .III.14. Representation of the selected Activation function.

AF_ Name	AF_ Description	AF_ Formula	AF_ Graph and symbol
logsig	Log-sigmoid transfer function	$a = \frac{1}{(1+e^{-n})}$	
tansig	Hyperbolic tangent sigmoid transfer function	$a = \frac{2}{(1+e^{-2n})-1}$	
purelin	Linear transfer function	$a = n$	

III.5.Statistical criteria for evaluation of the model

The evaluation of the performance of a regression model, also known as validation, is the most important phase of any project.

The performance evaluation of this model is done with the use of some classical validation parameters. Note that the similarities between this stage with the QSPR model validation in the previous section (see **.III.2.5.**). Both of them share some comparable parameters, the following: r, R^2, k, k', R_m^2, RMSE, and they differ in some others. Their formulas are given down below [76, 77]:

♦ Errors parameters

$$AARD\% = 100 \cdot \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i^{obs}-y_i^{pred}}{y_i^{obs}}\right| \quad \text{III.23}$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(y_i^{obs} - y_i^{pred}\right)^2 \quad \text{III.24}$$

$$RE\% = 100 \cdot \left|\frac{y_i^{obs}-y_i^{pred}}{y_i^{pred}}\right| \quad \text{III.25}$$

where,

AARD Average absolute relative deviation,

MSE Mean squared error, and

RE Relative error

♦ Different expressions for the calculation of Q^2 (Cross-validated)

$$Q_{F1}^2 = 1 - \frac{\sum_{i=1}^{n_{out}}\left(y^{obs}-y^{pred}\right)^2}{\sum_{i=1}^{n_{out}}\left(y^{obs}-\bar{y}_{TR}\right)^2} \quad \text{III.26}$$

$$Q_{F2}^2 = 1 - \frac{\sum_{i=1}^{n_{out}} (y^{obs} - y^{pred})^2}{\sum_{i=1}^{n_{out}} (y^{obs} - \bar{y}_{OUT})^2} \quad \text{III.27}$$

$$Q_{F3}^2 = 1 - \frac{\sum_{i=1}^{n_{out}} (y^{obs} - y^{pred})^2 / n_{OUT}}{\sum_{i=1}^{n_{TR}} (y^{obs} - \bar{y}_{TR})^2 / n_{TR}} \quad \text{III.28}$$

where,

n_{TR} & n_{OUT} The number of training and test objects respectively,

$\bar{y}_{TR}$ Average value of the training set experimental responses, and

$\bar{y}_{OUT}$ Average value of experimental responses of external test objects.

It can be also noted that, for classical cross-validation techniques (e.g. leave-one out), where each training object is used as test object only once, $Q_{F1}^2 = Q_{F2}^2 = Q_{F3}^2$, being $n_{OUT} = n_{TR}$ and $\bar{y}_{OUT} = \bar{y}_{TR}$.

In the last years, other metrics for evaluating prediction ability of external sets were proposed. The concordance correlation coefficient (*CCC*) was proposed to assess the agreement of two trials of an assay or instrument in terms of method validity and repeatability.

Later, the concordance correlation coefficient *CCC* was introduced to evaluate the prediction ability of regression models. In particular, the concordance correlation coefficient, here referred to as Q_{CCC}^2, is defined as:

$$Q_{CCC}^2 = \frac{2 \cdot \sum_{i=1}^{n_{OUT}} (y^{obs} - \bar{y}_{EXP}) \cdot (y^{pred} - \bar{y}_{PRED})}{\sum_{i=1}^{n_{OUT}} (y^{obs} - \bar{y}_{EXP})^2 + \sum_{i=1}^{n_{OUT}} (y^{pred} - \bar{y}_{PRED})^2 + n_{OUT} \cdot (\bar{y}_{EXP} - \bar{y}_{PRED})^2} \quad \text{III.29}$$

where,

$\bar{y}_{EXP}$&$\bar{y}_{PRED}$ The average of the experimental and predicted responses of the external set, respectively.

It can be noted that no information about the training set is used.

♦ Bias factor *Bf* and precision factor *Af*

Good model *Bf* = 0.9 - 1.05,
Acceptable model *Bf* = 0.7 - 0.9,
Use with caution *Bf* = 1.06 - 1.15, and
Unacceptable model *Bf* = *Bf* <0.7 & *Bf* >1.15.

Ross (1996) proposed two different indices for determining the performance of the model. The bias factor (*Bf*) is a measure of the overall agreement between the predicted values and the observed values. It will indicate whether, and to what extent, the forecasts are above or below the equivalence line.

Equal weighting is given to over-prediction, where the predicted value is greater than the observed value, and to under-prediction. The *Bf* is shown in **Eq.III.30.**

$$Bf = 10^{\left(\frac{\sum_{i=1}^{n} log\frac{y^{obs}}{y^{pred}}}{n}\right)} \qquad \text{III.30}$$

Perfect agreement between the observed values and the predicted values would give a *Bf* of 1. A value greater than 1 would indicate that, on average, the predictions were larger than the observed values and would represent an intrinsically safe model; For example a *Bf* of 1.2 would indicate that the forecast was 20% larger than the observed values.

A value less than 1 would indicate an intrinsically safe model where, on average, the forecasts were shorter than those observed. *Bf* is now widely used as a measure of model performance and is a common criterion for comparing different models. In the calculation of *Bf*, the predictions on (+) and under (-) can be cancelled. They therefore do not show absolute errors in the model. This can be done using the precision factor (*Af*). This is based on an equation similar to **Eq.III.30** but ignores whether the difference between the predicted value and the observed value is positive or negative:

$$Af = 10^{\left(\frac{\sum_{i=1}^{n}\left|log\frac{y^{obs}}{y^{pred}}\right|}{n}\right)} \quad \text{III.31}$$

All the differences are now positive, the value of *Af* will always be equal to (if the agreement is perfect) or greater than one. An accuracy of 1.5 means, on average, the predicted value is 50% different (smaller or larger) from the observed value.

Furthermore, the criteria suggested by researchers were used to ensure more validity of the ANN model. These statistics are given as follows:

$$R_o^2 = 1 - \frac{\sum\left(y_i^{pred} - k y_i^{pred}\right)^2}{\sum\left(y_i^{pred} - \bar{y}^{pred}\right)^2} \quad \text{III.32}$$

$$R'^2_o = 1 - \frac{\sum\left(y_i^{obs} - k' y_i^{obs}\right)^2}{\sum\left(y_i^{obs} - \bar{y}^{obs}\right)^2} \quad \text{III.33}$$

$$m = \frac{R^2 - R_o^2}{R^2} \leq 0.1 \quad \text{III.34}$$

$$n = \frac{R^2 - R'^2_o}{R^2} \leq 0.1 \quad \text{III.35}$$

III.6.Results and Discussion

III.6.1.ANN modeling

Among the previous eight (08) MLP-ANN built models (one input layer, two hidden layers, and one output layer) three models were optimized based on their Average absolute relative error (AARD) as objective function. Then, their performances were evaluated and compared one to another in the prepuce of choosing the best-fit model, architecture, that's it is able to predict better the molar solubility fraction of drug-like compounds, particularly for those that have not been tested as well as new drugs.

The chosen three models are represented in form of **Table .III.15.** Their statistical results is given in the **Table .III.16.** for the validation parameters, and **Table .III.17.** for the errors parameters.

Table .III.15. Representation of the three selected MLP-ANNs with two hidden layers models.

Model	Training function	Activation function (1st hidden layer)	Activation function (2nd hidden layer)	Activation function (output layer)	Architecture
(A)	trainbr	logsig	logsig	purelin	{13-19-17-1}
(B)	trainbr	tansig	logsig	purelin	{13-18-14-1}
(C)	trainbr	logsig	tansig	purelin	{13-19-16-1}

Table .III.16 Validation parameters for the train, test, and global sets for the three MLP-ANN(two hidden layer) models.

Architecture	trainbr: logsig-logsig {13-19-17-1}			trainbr: tansig-logsig {13-18-14-1}			Trainbr: logsig-tansig {13-19-16-1}		
Statistical values	Train	Test	Global	Train	Test	Global	Train	Test	Global
Q^2_{F1}	0.9964	0.9921	0.9958	0.9957	0.9932	0.9953	0.9962	0.9935	0.9958
Q^2_{F2}	0.9965	0.9921	0.9958	0.9957	0.9932	0.9953	0.9962	0.9934	0.9958
Q^2_{F3}	0.9964	0.9926	0.9964	0.9957	0.9929	0.9953	0.9962	0.9935	0.9958
Q^2_{CCC}	0.9982	0.9961	0.9979	0.9978	0.9966	0.9976	0.9981	0.9967	0.9979
R^2	0.9964	0.9921	0.9958	0.9957	0.9932	0.9953	0.9962	0.9934	0.9958
R	0.9982	0.9961	0.9979	0.9978	0.9966	0.9976	0.9981	0.9967	0.9979
k	1.0000	0.9993	0.9999	1.0000	1.0014	1.0002	1.0000	0.9989	0.9999
k'	0.9998	1.0002	0.9999	0.9997	0.9982	0.9995	0.9997	1.0007	0.9999
R_0^2	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
R'_0^2	1.0000	1.0000	1.0000	1.0000	0.9999	1.0000	1.0000	1.0000	1.0000
m	-0.0036	-0.0080	-0.0042	-0.0043	-0.0069	-0.0047	-0.0038	-0.0066	-0.0042
n	-0.0036	-0.0080	-0.0042	-0.0043	-0.0068	-0.0047	-0.0038	-0.0066	-0.0042
R_m^2	0.9370	0.9038	0.9315	0.9304	0.9112	0.9271	0.9350	0.9129	0.9313
Af	1.0098	1.0140	1.0105	1.0109	1.0147	1.0114	1.0101	1.0139	1.0106
Bf	1.0002	1.0001	1.0002	1.0002	0.9988	1.0000	1.0001	1.0012	1.0003

Table .III.17. Different errors for the train, test, and global sets for the three MLP-ANN (two hidden layer) models.

Architecture	trainbr: logsig-logsig {13-19-17-1}			Trainbr: tansig-logsig {13-18-14-1}			Trainbr: logsig-tansig {13-19-16-1}		
Error	Train	Test	Global	Train	Test	Global	Train	Test	Global
AARD%	0.9867	1.3910	**1.0474**	1.0888	1.4528	1.1435	1.0093	1.3844	1.0656
RMSE	0.1548	0.2227	**0.1667**	0.1688	0.2174	0.1770	0.1588	0.2088	0.1672
MSE	0.0240	0.0496	**0.0278**	0.0285	0.0473	0.0313	0.0252	0.0436	0.0280

In summary, the results above are in total agreement with the acceptability criteria. Where those models give some good results such as the coefficient of determination (R^2) is above 1, the cross-validation parameters are >0.5, the bias factor ≈1.

For the **Table .III.17** we can notice some difference and signifies the over-surpassing of the model (A) on the two others as a first judgment.

The graphical evaluation is carried out using a scatter diagram (for the training, test, and global sets), the ER% diagram, and the ER% distribution diagram.

Figure .III.9 to **.III.11** show the regression line of the model equation, i.e. predicted vs. observed results for the training, test, global set for the three topologies. A close correlation between the values predicted by the ANN models and the observed values of the mole solubility fraction was found, a tight dispersion of the points on the first bisector. However, we can notice the existence of few out-case points.

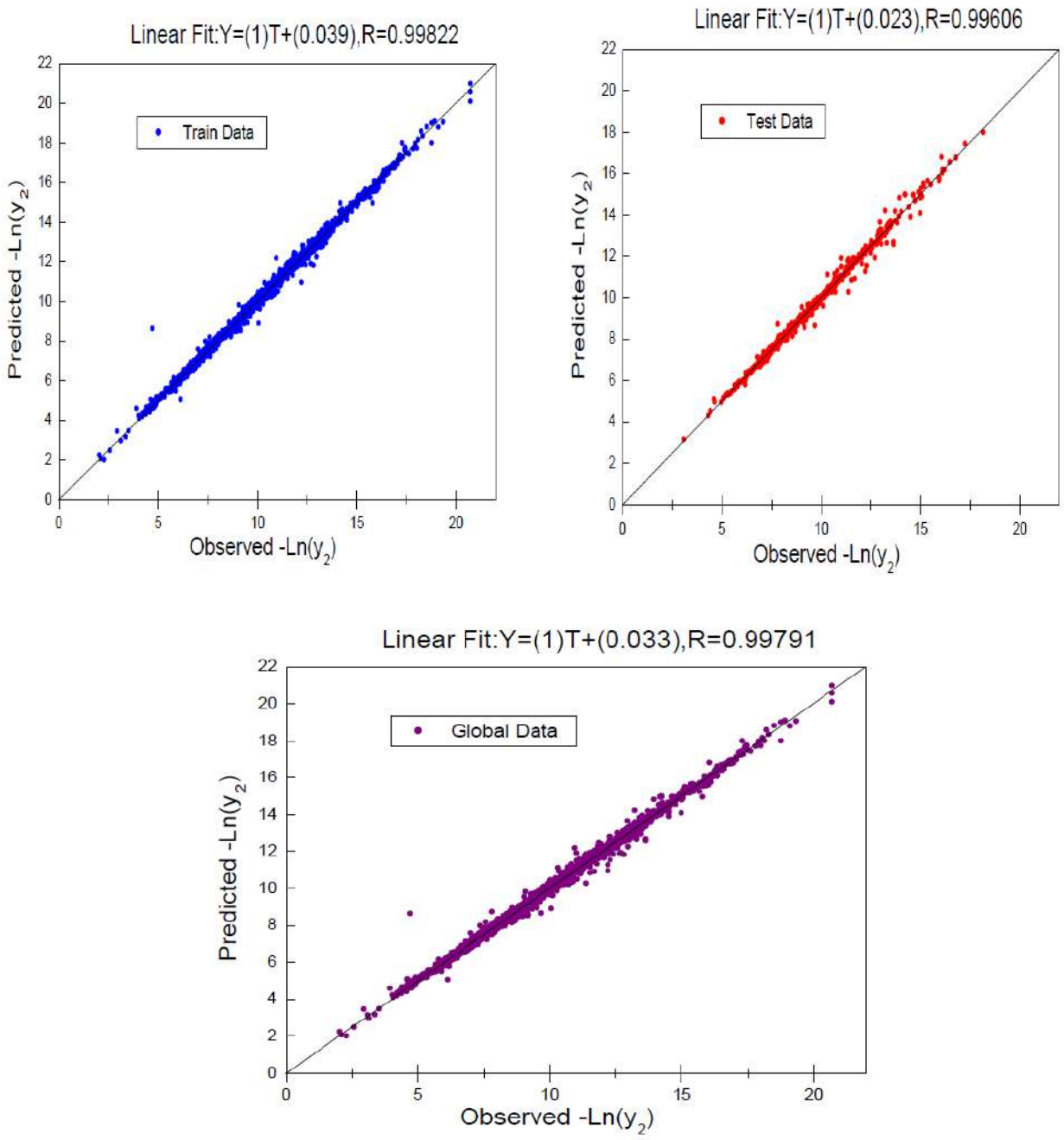


Figure .III.9 The dispersion diagram of the molar solubility fraction, observed vs. predicted for the train, test, and global set. Model (A).

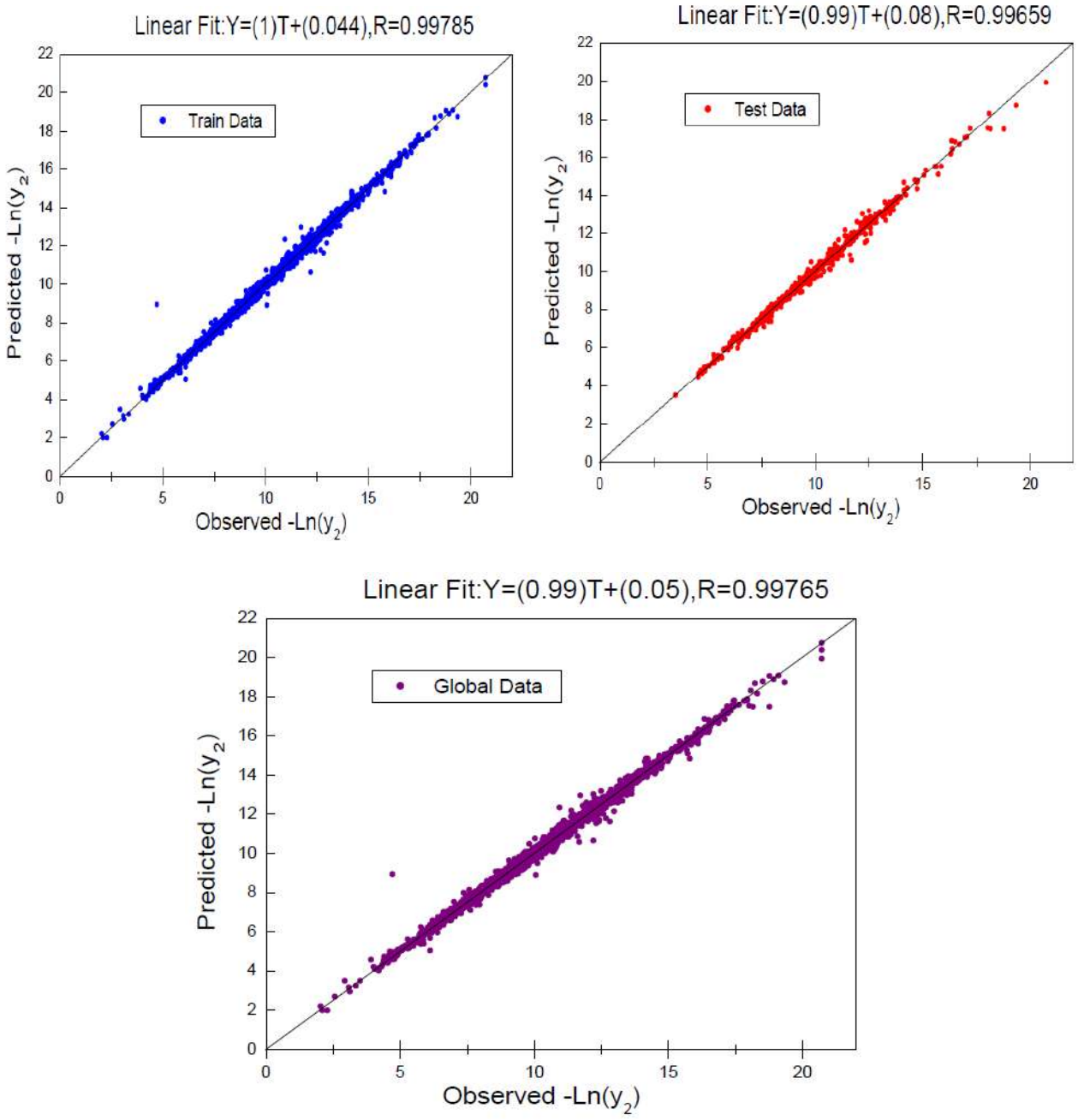


Figure .III.10. The dispersion diagram of the molar solubility fraction, observed vs. predicted for the train, test, and global set. Model (B).

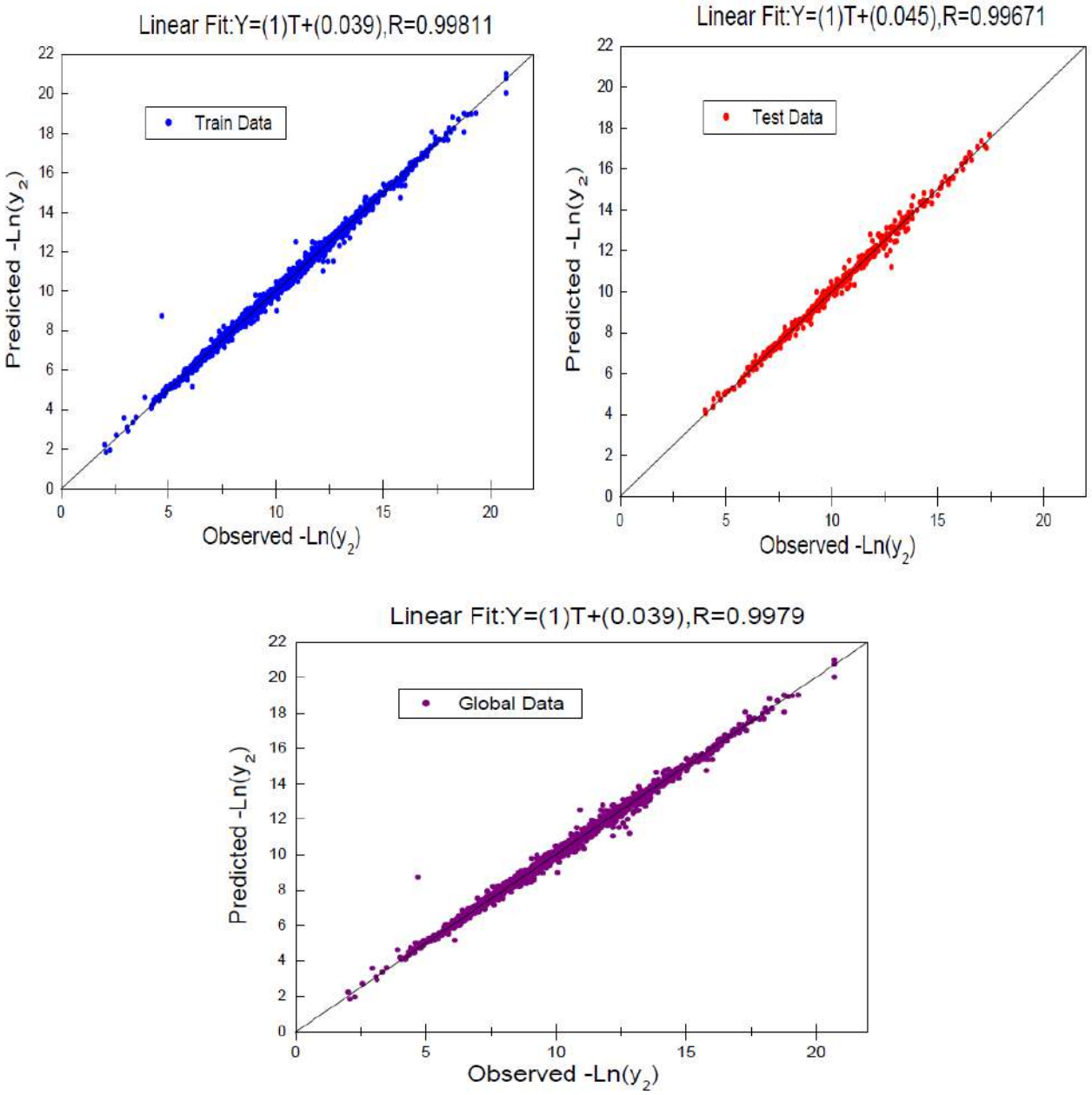


Figure .III.11. The dispersion diagram of the molar solubility fraction, observed vs. predicted for the train, test, and global set. Model (C).

In order to quantitatively visualize the distribution of relative errors, two types of diagrams have been used. The first diagram shows the relative error as a function of the observation and the second shows the distribution of this error.

Figure .III.12. to **.III.14** show the relative error vs. the solubility for those three models respectively. For the three models the majority of the points are located around the zero line. They show similar distributions.

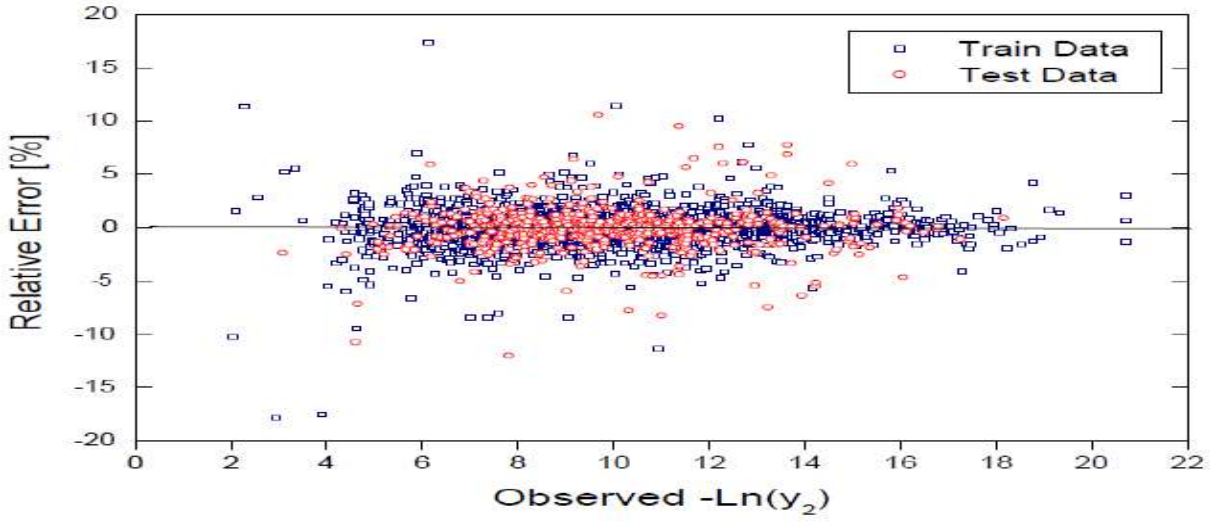


Figure .III.12 The relative error vs. the solubility for model (A).

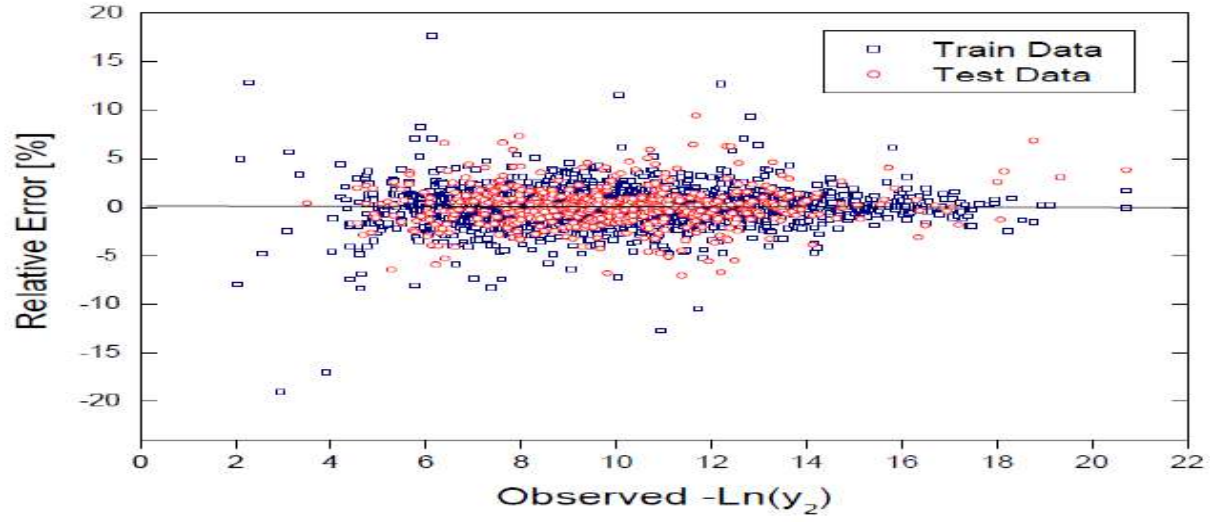


Figure .III.13 The relative error vs. the solubility for model (B).

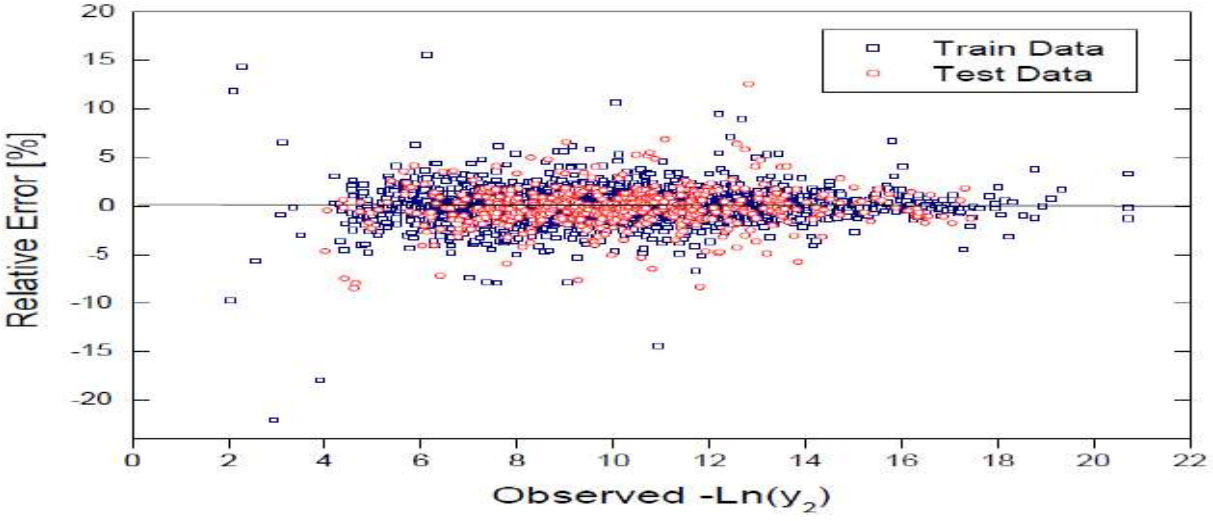


Figure .III.14 The relative error vs. the solubility for model (C).

Those distributions would be more visible in the following bar diagrams. **Figure .III.15** to **.III.17.**

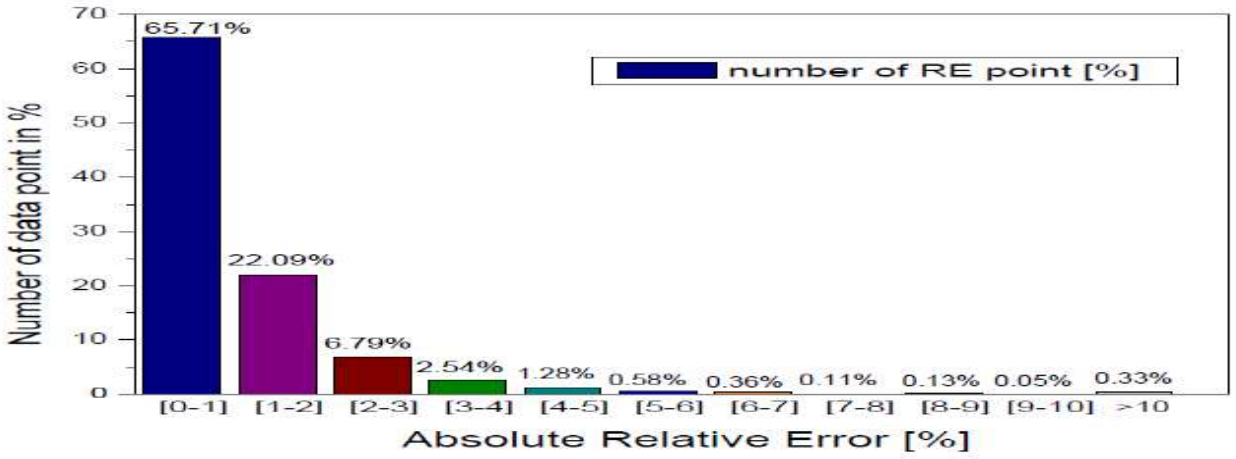


Figure .III.15. Bar distribution diagrams of the Absolute Relative Error [%] vs. number of DP in %, model (A).

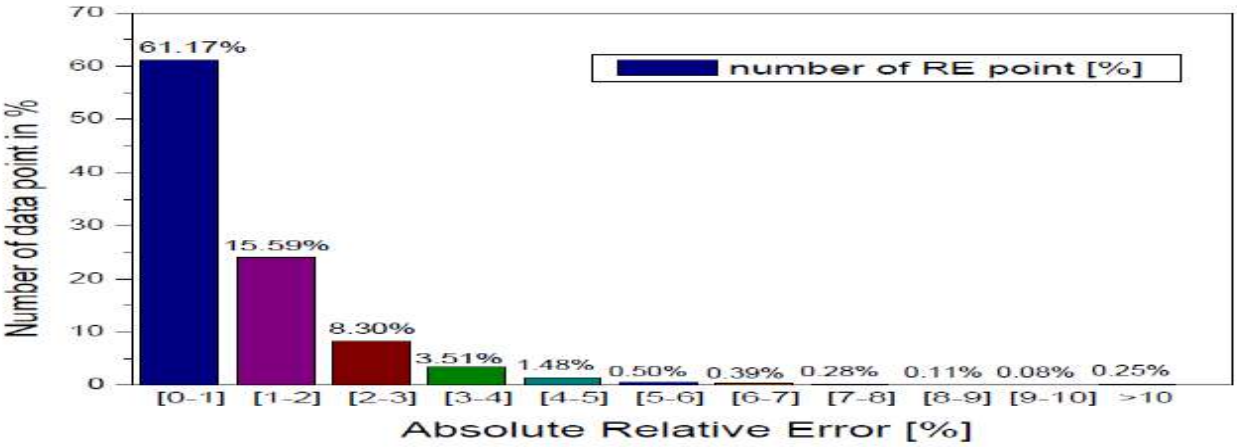


Figure .III.16. Bar distribution diagrams of the Absolute Relative Error [%] vs. number of DP in %, model (B).

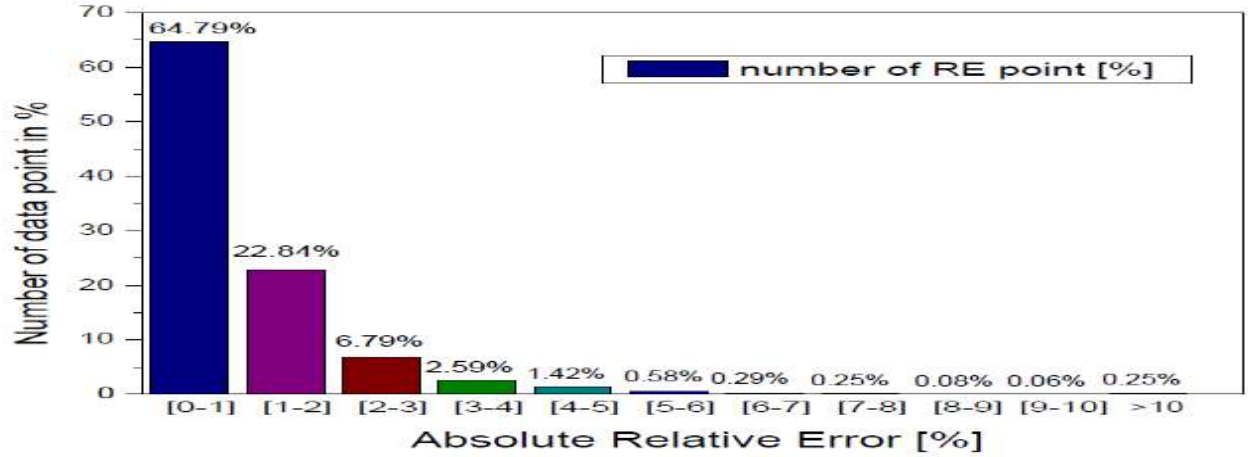


Figure .III.17. Bar distribution diagrams of the Absolute Relative Error [%] vs. number of DP in %, model (C).

❖ Applicability domain

The applicability domain of the three topologies were analyzed using a Williams plot (**Figure .III.18** to **.III.20**), where the vertical line is the critical leverage value (h*) equal 0.0138, and the horizontal lines ±3.

In those Figures, we can observe that none of the drug-like compounds in the whole dataset have leverage higher than the warning h*. Most of the dataset belongs to the AD area, is between the horizontal lines (limit of ± 3), which means that an average of 98.49%, 98.66%, and 98.38% of the whole database is covered for the models (A), (B), and (C) respectively.

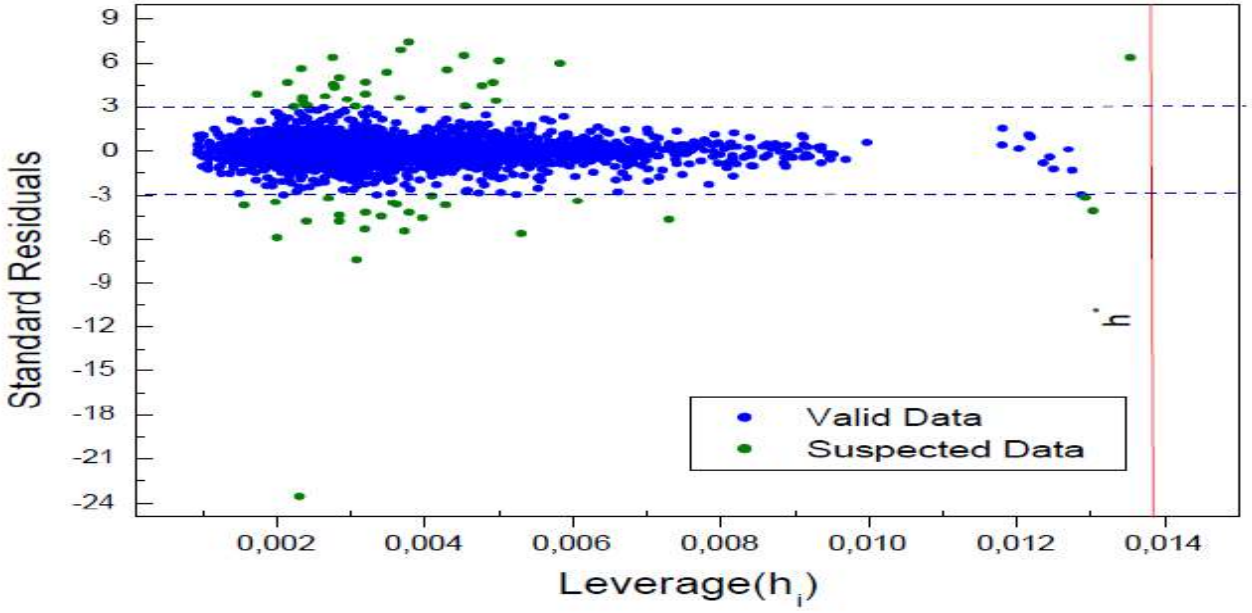


Figure .III.18. Williams plot for MLP-ANN, two hidden layer, model (A).

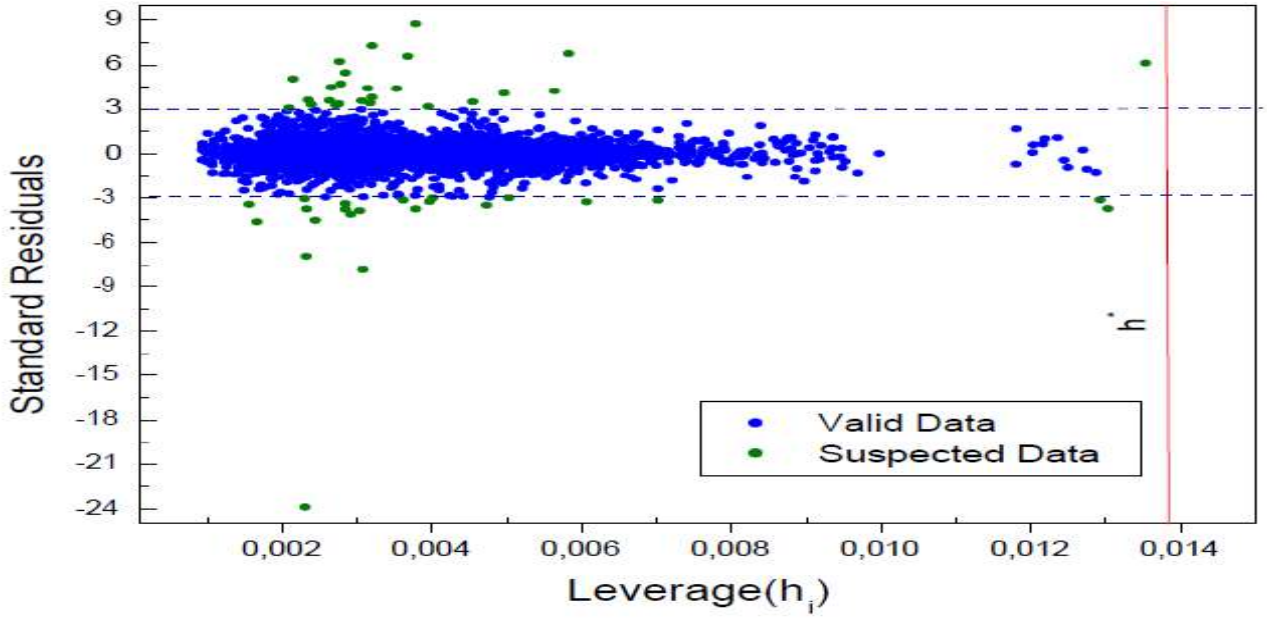


Figure .III.19. Williams plot for MLP-ANN, two hidden layer, model (B).

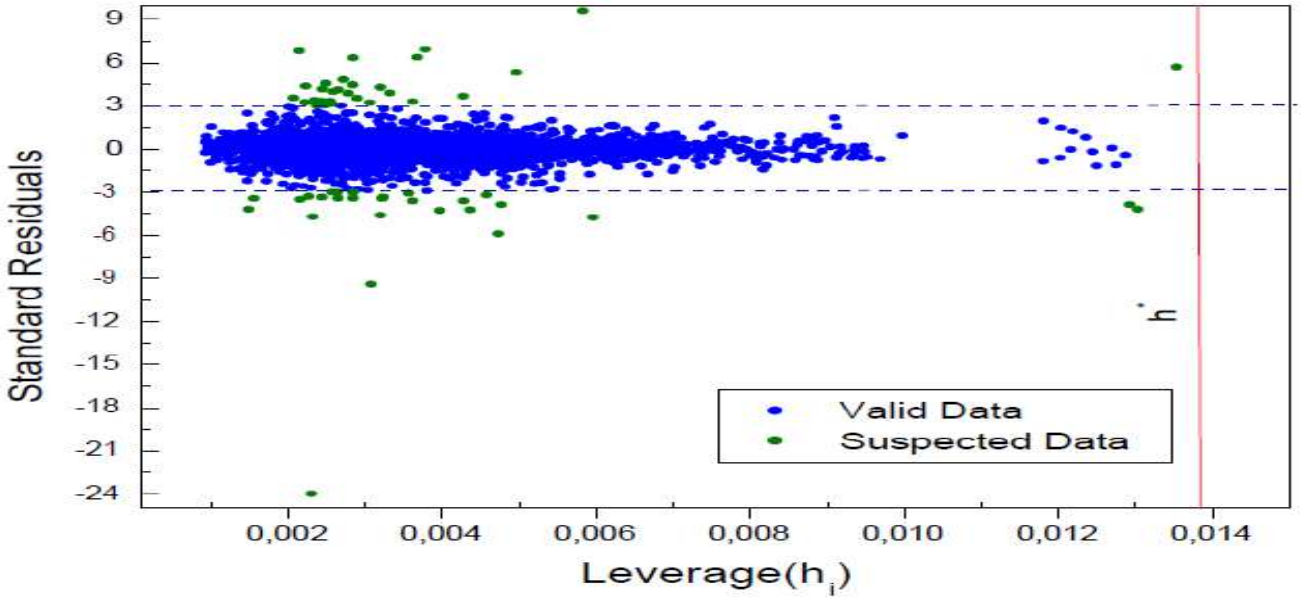


Figure .III.20. Williams plot for MLP-ANN, two hidden layer, model (C).

❖ Sensitivity analysis

ANNs are powerful tools able to predict and analyze easily non-linear relationships. A large number of authors have underlined the interest of using ANNs instead of linear statistical models. In fact, starting from input variables, ANNs have the capacity to predict the output variable but the mechanisms that occur within the network are often ignored. ANNs are often considered as black boxes. Various authors have explored this problem and proposed algorithms to illustrate the role of variables in ANN models. AKA, the contribution of features.

Knowing what contribution each variable makes it of prime importance. It is this explanatory aspect of ANNs that we studied here. These methods were used to determine the influence of each input variable and its contribution to the output. They are not, therefore, pruning methods but procedures to estimate the relative contribution of each input variable.

Seven different methods that allow contribution analysis can be used. While this study will adopt the use of the Garson method to calculate the relative importance of each input variable for the three models [78].

Figure .III.21 depicts sensitivity analysis for model (A) and shows that the P(Mpa) has the highest contribution (15.64%) where the lowest contribution was shared by T(K) (4.25%), GATS3v (4.31%), and MLogP (4.54%). The rest of the inputs have an importance between 6.51% and 9.45%.

Figure .III.22 depicts sensitivity analysis for model (B) and shows that AVP-0 has the highest contribution (10.32%) followed by MLogP (9.98%) with a small difference. The lowest contribution was shared by T(K) (3.09%), and GATS4s (3.95%), where the rest of the inputs have an importance between 6.32% to 9.51%.

Figure .III.23 depicts sensitivity analysis for model (C) and shows that P(Mpa) has the highest contribution (19.54%), where the lowest contribution was shared by MATS2c (3.89%), GATS4s (3.86%), and SHBd (3.37%) .The rest of the inputs have an importance between 4.32% to 10.11%.

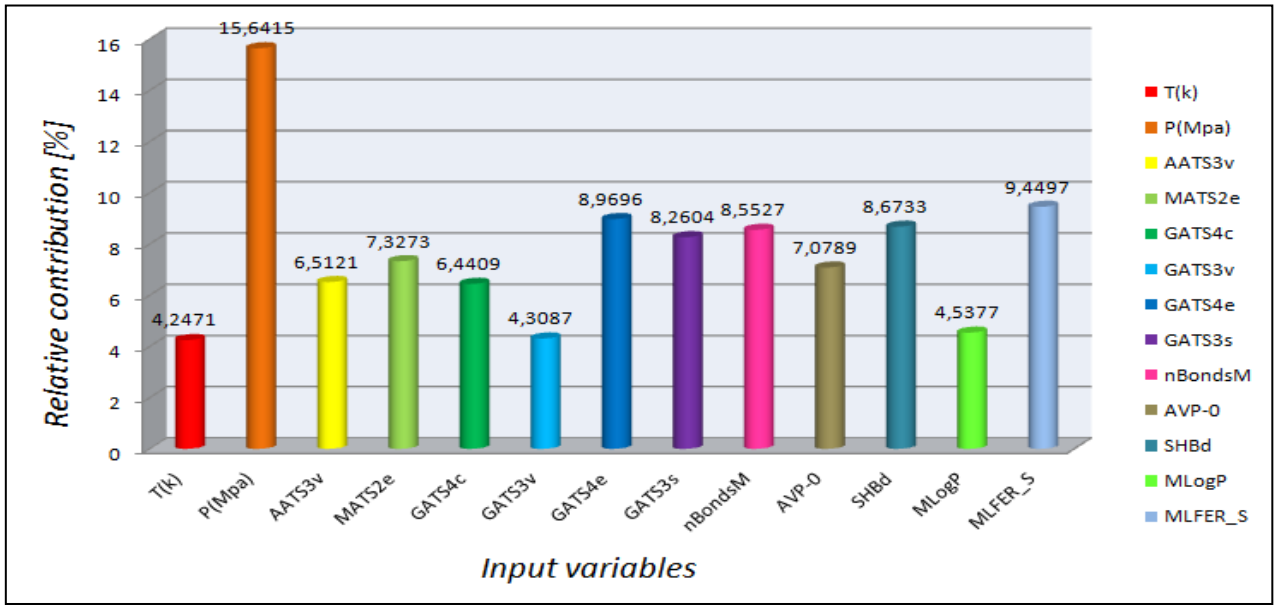


Figure .III.21 Relative contribution of each input variable, the Garson method, model (A).

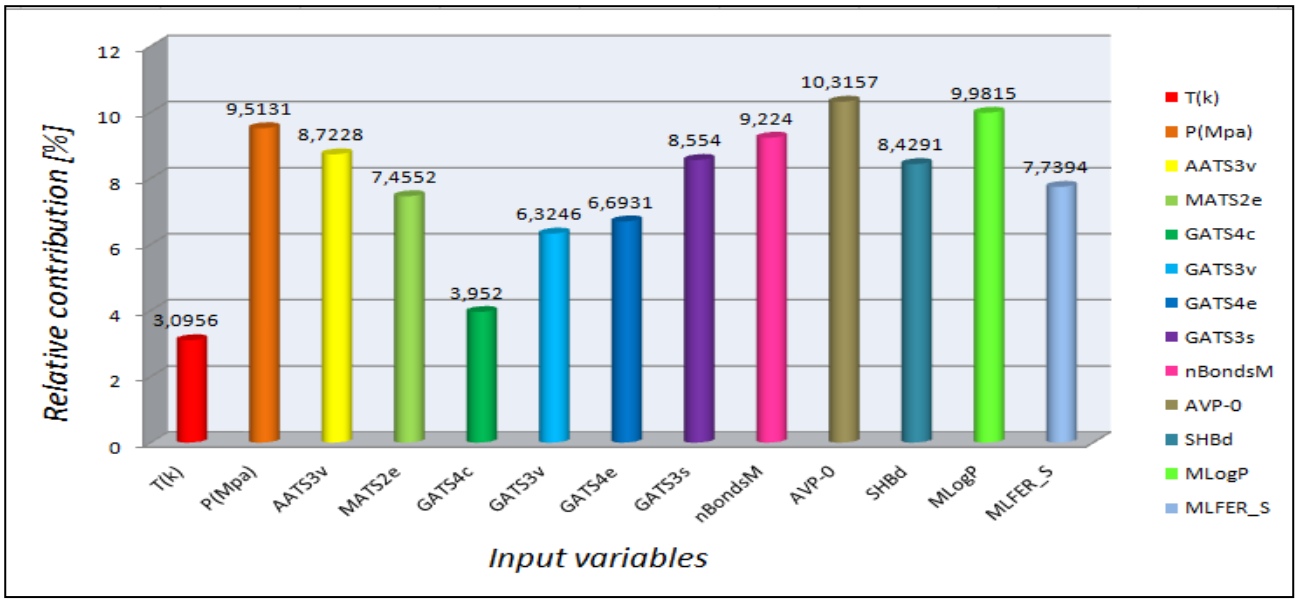


Figure .III.22 Relative contribution of each input variable, the Garson method, model (B).

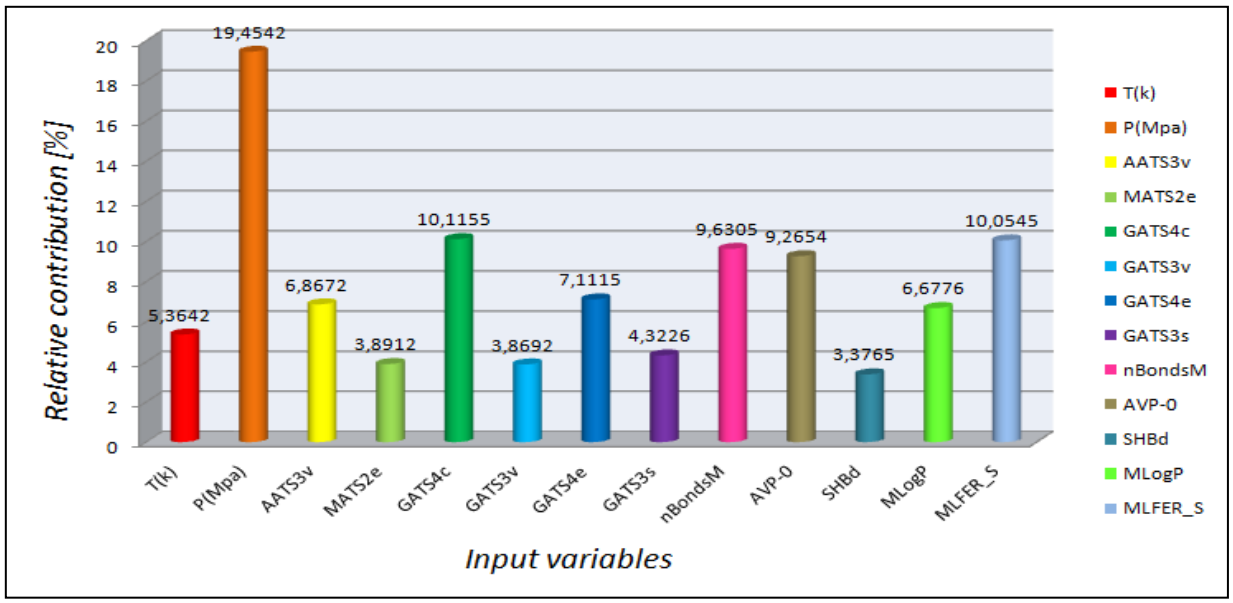


Figure .III.23 Relative contribution of each input variable, the Garson method, model (C).

III.6.2.Conclusion

The study reported in this section was designed to determine the effect of using two hidden layers on the performance of MLP-ANN. On the other hand, showing its influence on the predicted property (solubility fraction), and be one of the few contributors to this type of model, for such kind of properties.

Note that the number of the input nodes (13) * the number of 1st hidden layer nodes (19) + the number of 1st hidden layer nodes (19)* the number of 2nd hidden layer nodes (17) + the number of 2nd hidden layer nodes (17)* the number of the output nodes (1), "equal 587", is still six times smaller than the total number of the dataset (3590) proving the acceptability of this study.

The statistical and graphics results were quite similar resulting in the inability to giving a strict decision about the best-fit model representative predicted model able to be used to predict the solubility of drug-like compounds, particularly for those that have not been tested as well as new compounds. Except that, the model (A) that was obtained by trainbr "Bayesian regularization back-propagation" training function, logsig "Log-sigmoid transfer function" activation function for both hidden layers having the architecture of {13-19-17-1}, is the most candidate. Because it is the model with the fewest errors based on the importance of building models with the smallest cost error overall.

DFO-SVR predictive model

SVR is a robust tool to solve nonlinear regression problems; there are no general guidelines to define their parameters, which is an obstacle to the widespread use of SVR in manufacturing applications and scientific research. Therefore, the main concern for researchers is to reveal convenient parameter values for a given dataset in SVR that can ensure high accuracy.

Inappropriate SVR parameters values cause overfitting or under-fitting problems, and various parameter settings can also lead to considerable differences in performance. Traditionally, the grid search algorithm and gradient descent algorithm are the most commonly applied algorithms to select SVM parameters. Convergence to local minima point, computational complexity, and height computational time requirement are the major drawbacks of these conventional methods. With the development of meta-heuristic optimization algorithms, some of them have been adopted to determine the SVR parameters, such as genetic algorithm (GA) and particle swarm optimization (PSO). However, the performance of these methods is imperfect; The GA has some disadvantages, for instance, the premature convergence and the poor aptitude of local search. Another problem is related to the difficulty of choosing GA operators such as population size, selection method, crossover rate, and mutation rate, which have a significant impact on the convergence to the optimum solution. In a like manner, the effectiveness of the PSO is influenced by the particle's multiple parameters.

Dragonfly optimization (DFO) is a novel optimization technique proposed by Mirjalili [64]. It is inspired by the static and dynamic demeanour of a swarm of dragonflies. Compared to the other well-known optimization techniques, DFO has some advantages, such as a simple concept and easy implementation. The robustness of this new optimization method is tested and validated by authors using many standard benchmark functions. The results show that the DFO algorithm outperforms existing well-known algorithms such as GA and PSO [62].

In this part, a hybrid DFO–SVR model is proposed for predicting the target output.

III.7.Model development

In the proposed model, the DFO algorithm is deployed to seek out the optimal values of SVR parameters; the constant C (box-Constraint), ε the epsilon, and σ the parameter of the kernel function (Kernel-Scale) of the model. Based on experiences of empirical modeling in the field of SVR, the appropriate parameter combination of a SVR model is critical to any improvement of forecasting accuracy.

Therefore, as mentioned above, such a hybrid algorithm can be used to solve the parameter combination problem for an SVR model. i.e. obtain a combination of parameters that improves the forecasting performance of the proposed SVR model.

Nevertheless, the parameters of the kernel function are important for the accuracy of the prediction, so different kernel functions have an important influence on the performance of the support vector machine. In this experiment, the RBF, Gaussian, and polynomial kernel functions were tested to select the wide adaptability function; the one with the better statistics performance.

The developed model is carried out by the ***fitrsvm*** function of MATLAB® R2019b. Where we have chosen the Cross-validation "holdout" method which we have set to 0.3.AKA, 70% of the data is used for learning, and 30% for validation.

The efficiency of the developed model is evaluated with the use of the same statistics tools and graphs that has been used for the ANN model.

III.8.Results and Discussion

III.8.1.DFO-SVR modeling

The DFO algorithm feeds the SVR initially with a random combination of hyper-parameters within their ranges. During a number of five iterations, the steps beginning with the division of data and up until the development of the SVR model are repeated and the minimum value of AARD obtained is saved as best. Then, the DFO generates a new population of hyper-parameters for the SVR algorithm, and the same set of steps are repeated in order to achieve a new best AARD and that is for 200 trials, amongst which the minimum AARD corresponds to the resultant optimal DFO-SVR model.

Among server trails the best-fit DFO-SVR predictive model, that's it is able to predict better the molar solubility fraction of drug-like compounds, particularly for those that have not been tested as well as new drugs is the following:

The optimized parameters of the DFO-SVR models are presented in **Table .III.18.**

Table .III.18 Optimized parameters of the DFO-SVR model.

C	ε	σ	Kernel function
200	0.0797	1.7150	Gaussian
Quantity of support vectors: 2429			

Tables .III.19 and .**III.20** respectively summarize the calculation of the different statistical criteria and the different errors of the optimized DFO-SVR model. It is clear that all the statistical criteria verify the conditions of acceptability. However, there is a notable difference between the training set error and the test set.

Table .III.19 Validation parameters for the train, test, and global sets for DFO-SVR model.

Statistical values	Train	Test	Global
Q^2_{F1}	0.9974	0.9868	0.9945
Q^2_{F2}	0.9974	0.9867	0.9945
Q^2_{F3}	0.9974	0.9885	0.9947
Q^2_{CCC}	0.9987	0.9933	0.9972
R^2	0.9974	0.9867	0.9945
R	0.9987	0.9934	0.9973
k	1.0000	1.0012	1.0004
k'	0.9998	0.9981	0.9993
R_0^2	1.0000	1.0000	1.0000
R'_0^2	1.0000	0.9999	1.0000
m	-0.0026	-0.0134	-0.0055
n	-0.0026	-0.0134	-0.0055
R_m^2	0.9464	0.8733	0.9208
Af	1.0039	1.0151	1.0073
Bf	1.0002	0.9996	1.001

Table .III.20 Different errors for the train, test, and global sets for DFO-SVR model.

Error	Train	Test	Global
AARD%	0.4032	1.4873	0.7284
RMSE	0.1347	0.2823	0.1913
MSQ	0.0181	0.0797	0.0366

The graphical evaluation is carried out using a scatter diagram (for the training, test, and global sets), the ER% diagram, and the ER% distribution diagram.

Figure .III.24 shows the regression line of the model equation, i.e. predicted vs. observed results for the training, test, global set for the model. A close correlation between the values predicted by the DFO-SVR model and the observed values of the mole solubility fraction was found, a tight dispersion of the points on the first bisector. However, we can notice the existence of few out-case points.

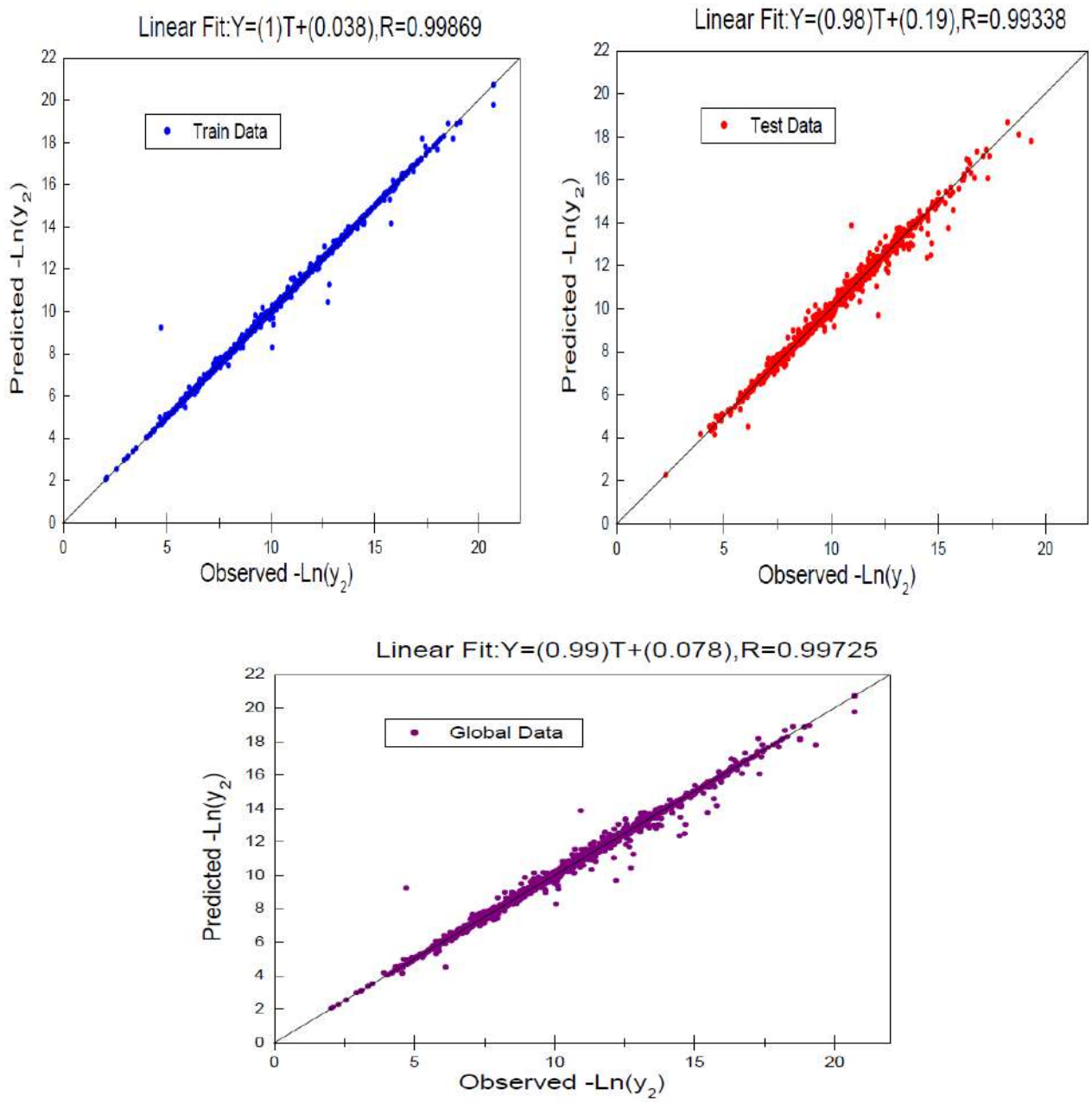


Figure .III.24. The dispersion diagram of the molar solubility fraction, observed vs. predicted for the train, test, and global set. Model DFO-SVR.

In order to quantitatively visualize the distribution of relative errors, two types of diagrams have been used. The first diagram shows the relative error as a function of the observation and the second shows the distribution of this error.

Figure .III.25 show the relative error vs. the solubility for the selected DFO-SVR model. The majority of the data points are located around the zero line spatially for the train set (blue color). Concerning distributions is wholly well. Those distributions would be more visible in the following bar diagram. **Figure .III.26.**

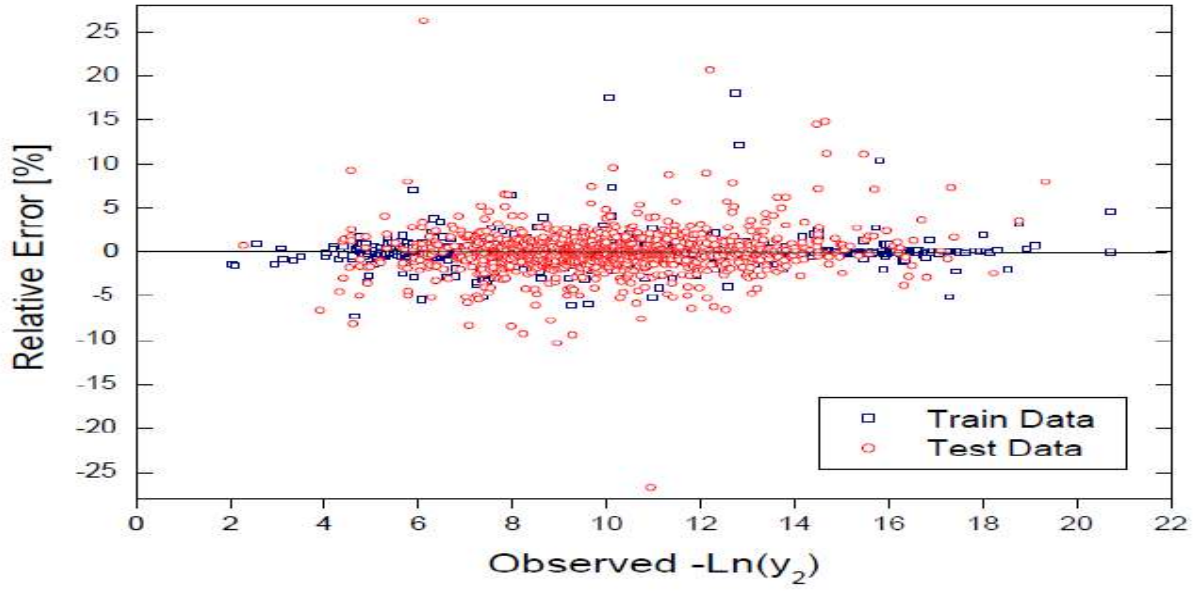


Figure .III.25. The relative error vs. the solubility for DFO-SVR model.

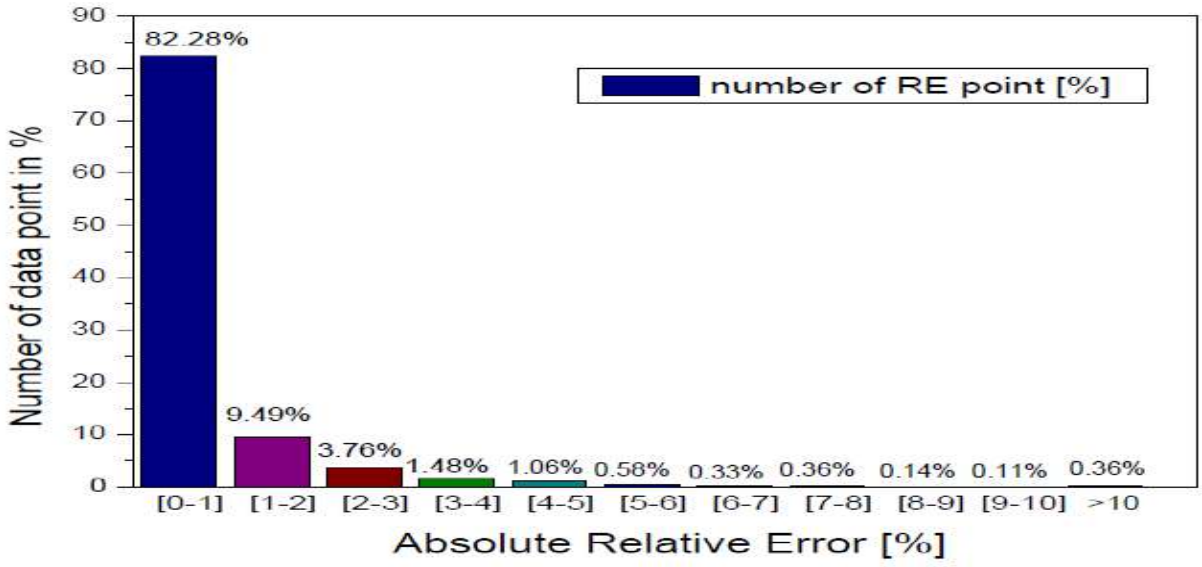


Figure .III.26 Bar distribution diagrams of the Absolute Relative Error [%] vs. number of DP in %, model DFO-SVR.

❖ Applicability domain

The applicability domain of this model is analyzed using a Williams plot (**Figure .III.27.**), where the vertical line is the critical leverage value (h*) equal 0.0167, and the horizontal lines ±3. In **Figure .III.27**, we can observe that none of the drug-like compounds in the whole dataset have leverage higher than the warning h*. Most of the dataset belongs to the AD area, is between the horizontal lines (limit of ± 3), which means that an average of 98.52% of the whole database is covered.

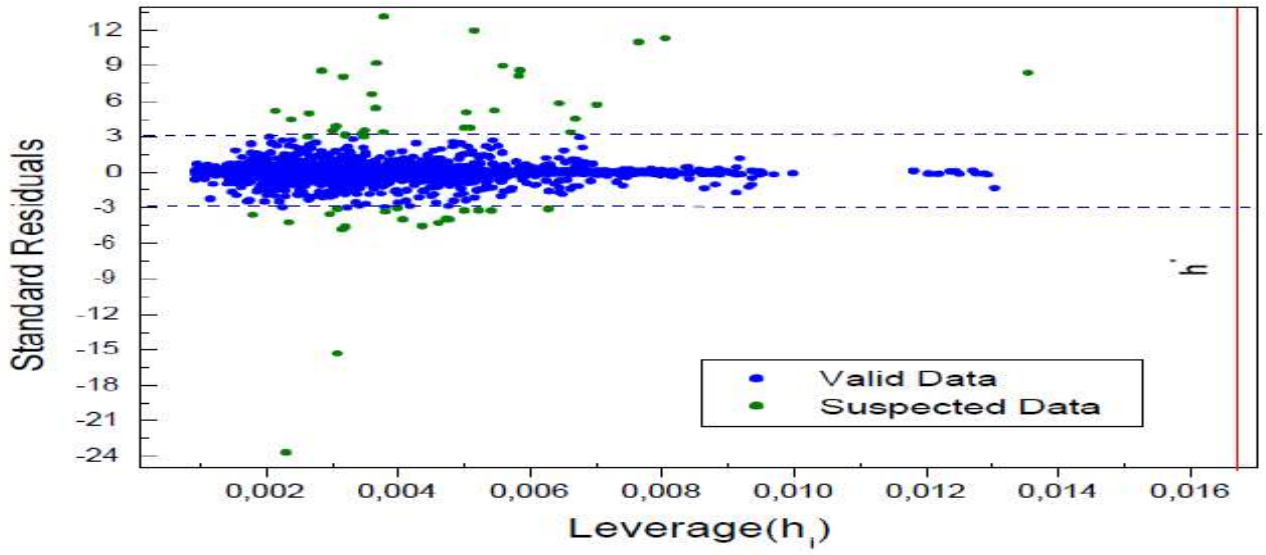


Figure .III.27. Williams plot for DFO-SVR model.

❖Sensitivity Analysis

The proposed model can potentially provide good dependency between the inputs and the output. In order to investigate how the inputs affect the output, a sensitivity analysis was undertaken. The most effective input can be recognized by the relevance factor *r*, which is in the range of −1 to +1 and is stated by the following equation [79]:

$$r_k = \frac{\sum_{i=1}^{n}(X_{k,i}-\bar{X}_k)(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_{k,i}-\bar{X}_k)^2 \sum_{i}^{n}(Y_i-\bar{Y})^2}} \qquad \text{III.36}$$

where,

$X_{k,i}$ The i^{th} importing parameter,

Y_i The i^{th} exporting value,

$\bar{X}_k$ The average value of the k^{th} input,

$\bar{Y}$ The average value of exporting parameter, and

n The number of sets.

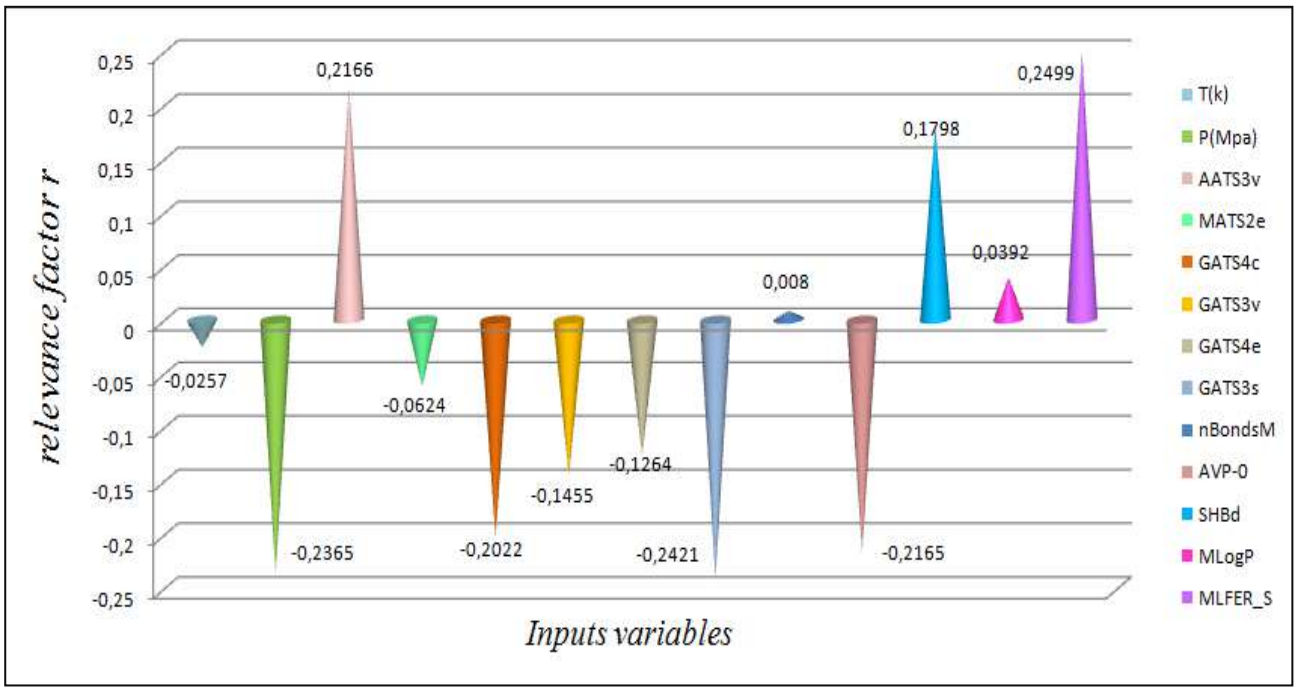


Figure .III.28. Results of the sensitivity analysis performed on the DFO-SVR model.

The absolute value of the relevance factor has a direct relation to the output. As can be seen in **Figure .III.28**, the solubility shows a straight dependency on the descriptors named: AATS3v, nBondsM, SHBd, MLogP, and MLFER_S, and an opposite dependency on the temperature and the pressure, and the rest of descriptors.

In addition, MLFER_S, and pressure are the most relevant input variables with a relevance factor of +0.25 and -0.24 respectively. Against, the temperature and nBondsM are the least relevant with a relevance factor of -0.03 and + 0.01 respectively.

III.8.2.Conclusion

The SVR learning algorithm, associated with the DFO algorithm for optimization of its hyper-parameters was proposed for developing a predicted model for the molar solubility fraction based on 13 selected descriptors from the previous QSPR modeling and the selection of gaussian as the kernel function. The developed model shows a good correlative and predictive accuracy.

These results demonstrate the improved performance of SVM for regression when the hyper-parameters are optimized using the dragonfly algorithm.

ANFIS predictive model

Among all selected algorithms of this work, ANFIS provides a special property, that a memory consuming tool and it is highly dependent on the number of parameters. In other words, the more the number of parameters (input, output features) is large, the more the ANFIS model consumes a large quantity of GB. Unfortunately, this study has the same issue.

In the simplest explanation, if we provide an ANFIS model with two membership functions on each of the 13 inputs (number of our features), then the ANFIS will generate $2^{13} = 8192$ fuzzy rules, and too many fuzzy rules will result in very long training time that the processor won't be able to handle, i.e. an error of lack of memory.

In this case, the GA is proposed to reduce the number of input features.

III.9.Model development

The genetic algorithm is the same used programming tool that has been early mentioned (see **.III.3.**). It takes as inputs the 13 selected descriptors of the QSPR model, with some statistic values where all have been taken by-default. For the output, the selection of the appropriate number of descriptors was reduced to 7 descriptors.

The obtained dataset of seven (7) features was randomly divided into three sets "training", "test" and "validation" according to a ratio of 70%, 15%, and 15% were used to build an ANFIS model. Although there are several methods of developing the fuzzy inference system (FIS), in this study, the Takagi Sugeno FIS system is developed by a grid partitioning method.

The ***genfis*** function was chosen as the processing function in the MATLAB® program, which uses the grid partition to develop the FIS TSK system. For the Membership Functions types, we select two random (***gbellmf***, and ***gaussmf***) MFs for the inputs, and the linear MF for the output.

Training the initial FIS is essential for better performance. In this study, the initial FIS was formed according to a predefined method in the MATLAB® fuzzy logic toolbox, called the hybrid method. The efficiency of the developed model is evaluated with the use of the same statistics tools and graphs that has been used for the previous models.

III.10.Results and Discussion

III.10.1.ANFIS modeling

The various features were selected with the use of GA and then test in the objective of obtaining the best combination able to cover the lack of both input number that's could give the lowest AARD error and the lack of memory of the simulator.

The best-fit combination is: T(k), P(Mpa), AATS3v, GATS4c, nBondsM, AVP-0, and MLFER_S. **Table .III.21** presents the detailed parameters for the main selected FIS training algorithm.

Table .III.21 Detail summary of the fuzzy inference system (FIS) learning parameters.

Parameters	Description\Value
Name	ANFIS
Fuzzy structure	Sugeno-type FIS
Initial FIS	Genfis
Input Membership Function Type	Generalized bell curve membership function (gbellmf)
Output Membership Function Type	Linear
Number of Membership Functions	2
Number of inputs	7
Number of output	1
Optimisation method	Hybrid
Number of nodes	294
Epoch Number	100
Initial Step Size	0.01
Step Size Decrease Rate	0.90
Step Size Increase Rate	1.10
Number of fuzzy rules	128

Tables .III.22 and .**III.23** respectively summarize the calculation of the different statistical criteria and the different errors of the optimized ANFIS model. It is clear that all the statistical criteria verify the conditions of acceptability. Also, the parameters of the error show some good results.

Table .III.22 Validation parameters for the train, test, and global sets for the ANFIS model.

Statistical values	Train	Test	Global
Q^2_{F1}	0.9869	0.9759	0.9836
Q^2_{F2}	0.9869	0.9759	0.9836
Q^2_{F3}	0.9869	0.9759	0.9836
Q^2_{CCC}	0.9879	0.9934	0.9917
R^2	0.9869	0.9759	0.9836
R	0.9934	0.9879	0.9918
K	1.0002	1.0003	1.0002
k'	0.9990	0.9982	0.9987
R_0^2	1.0000	1.0000	1.0000
R'_0^2	1.0000	0.9999	1.0000
m	-0.0133	-0.0247	-0.0167
n	-0.0133	-0.0247	-0.0167
R_m^2	0.8739	0.8242	0.8576
Af	1.0196	1.0283	1.0222
Bf	1.0006	0.9987	1.0000

Table .III.23 Different errors for the train, test, and global sets for the ANFIS model.

Error	Train	Test	Global
AARD%	1.9528	2.7414	2.1896
RMSE	0.2955	0.4011	0.3307
MSQ	0.0873	0.1608	0.1094

The graphical evaluation is carried out the same way as the previous models. i.e. using a scatter diagram (for the training, test, and global sets), the ER% diagram, and the ER% distribution diagram.

Figure .III.29 shows the regression line of the model equation, i.e. predicted vs. observed results for the training, test, validation, and global set for the model. A good correlation between the values predicted by the ANFIS model and the observed values of the mole solubility fraction was found, a tight dispersion of the points on the first bisector. However, we can notice the existence of few out-case points.

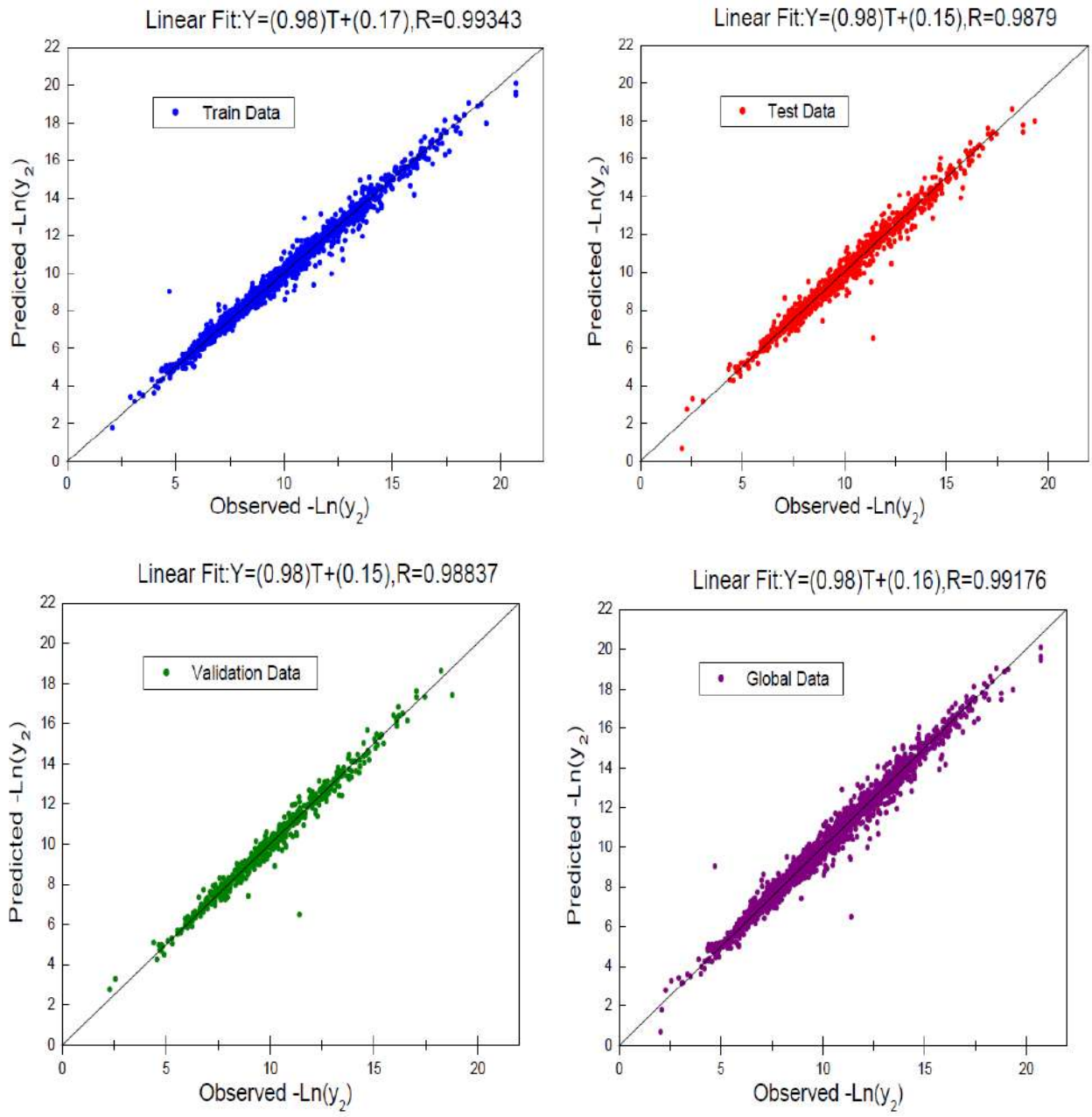


Figure .III.29. The dispersion diagram of the molar solubility fraction, observed vs. predicted for the train, test, validation, and global set, model ANFIS.

In order to quantitatively visualize the distribution of relative errors, two types of diagrams have been used. The first diagram shows the relative error as a function of the observation and the second shows the distribution of this error.

Figure .III.30 shows the relative error vs. the solubility for the selected ANFIS model. The majority of the data points are located around the zero line. Those distributions would be more visible in the following bar diagram. **Figure .III.31.**

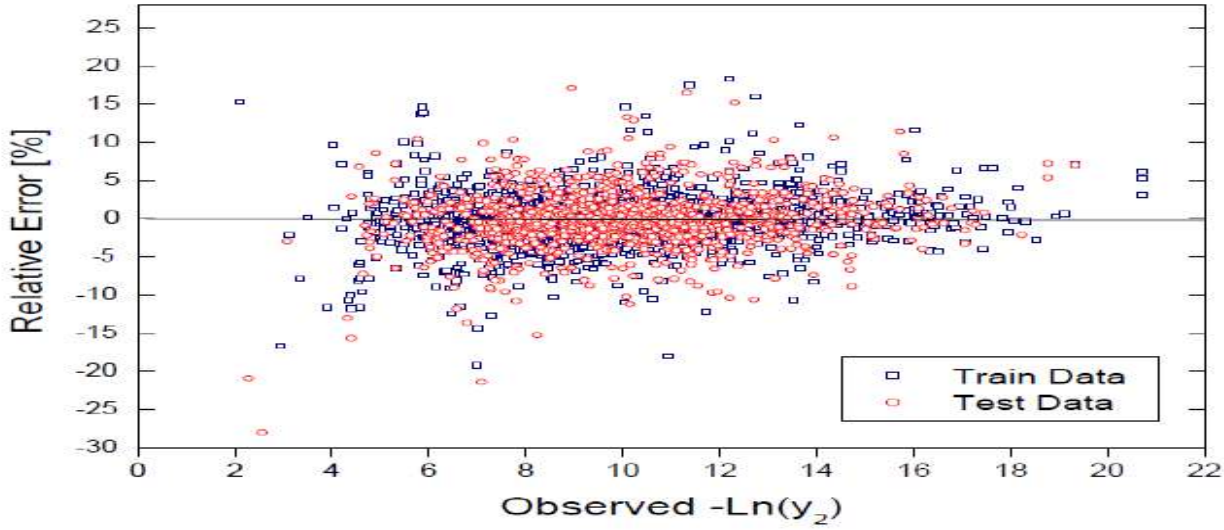


Figure .III.30. The relative error vs. the solubility for model ANFIS.

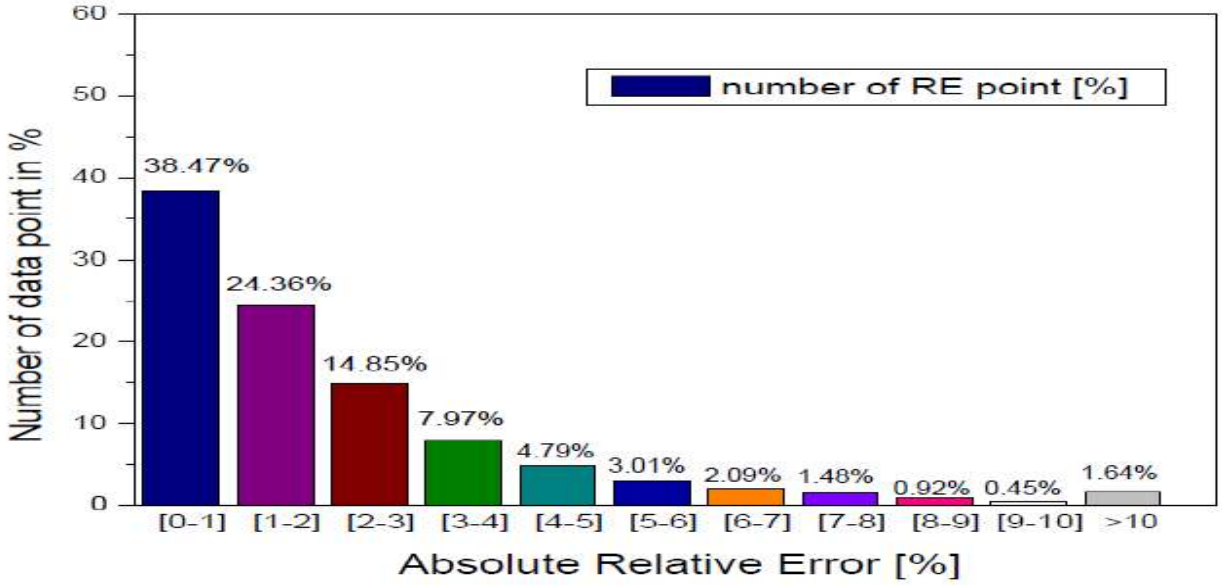


Figure .III.31. Bar distribution diagrams of the Absolute Relative Error [%] vs. number of DP in %, model ANFIS.

❖ Applicability domain

The applicability domain of this model is analyzed using a Williams plot (**Figure III.32.**), where the vertical line is the critical leverage value (h*) equal 0.0096, and the horizontal lines ±3. In **Figure .III.32**, we can observe that none of the drug-like compounds in the whole dataset have leverage higher than the warning h*. Most of the dataset belongs to the AD area, which is between the horizontal lines (limit of ± 3) responded to an average of 98.25% of the whole dataset.

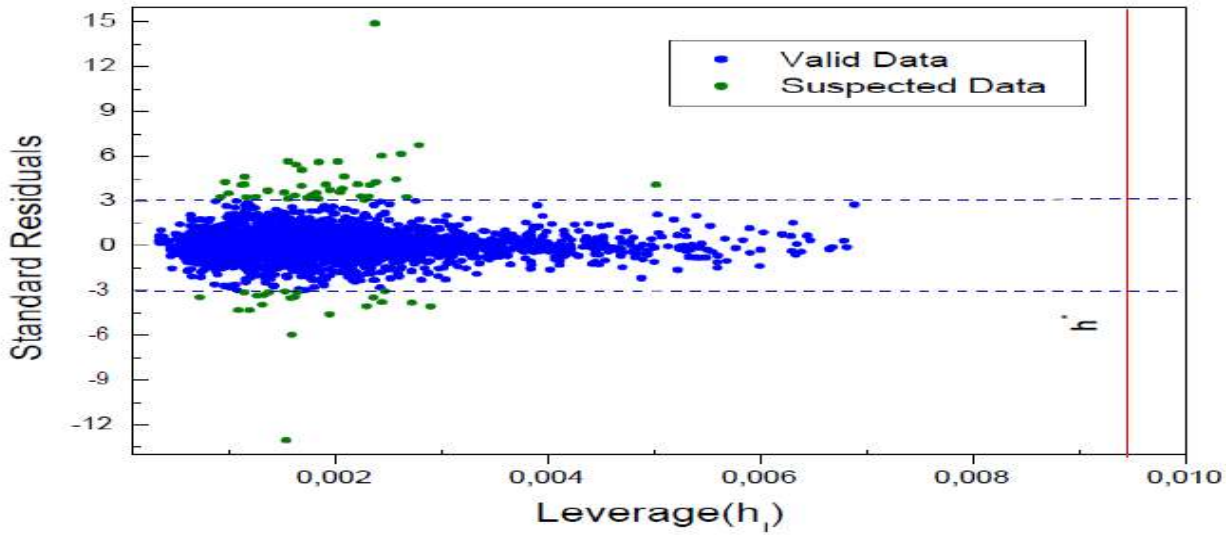


Figure .III.32. Williams plot for ANFIS model.

❖Sensitivity analysis

The proposed model can potentially provide good dependency between the inputs and the output. In order to investigate how the inputs affect the output, a sensitivity analysis was undertaken. The most effective input can be recognized by the relevance factor *r*, which is in the range of −1 to +1. Also, the same applied method of the previous algorithm (DFO-SVR).

As can be seen in **Figure .III.33**, the solubility shows a straight dependency on the descriptors named: AATS3v, nBondsM, and MLFER_S, and an opposite dependency on the temperature, the pressure, and the rest of descriptors (GATS4c, and AVP-0).

In addition, MLFER_S and pressure are the most relevant input variables with a relevance factor of +0.25 and -0.24 respectively against the temperature and nBondsM are the least relevant with a relevance factor of -0.03 and + 0.008 respectively.

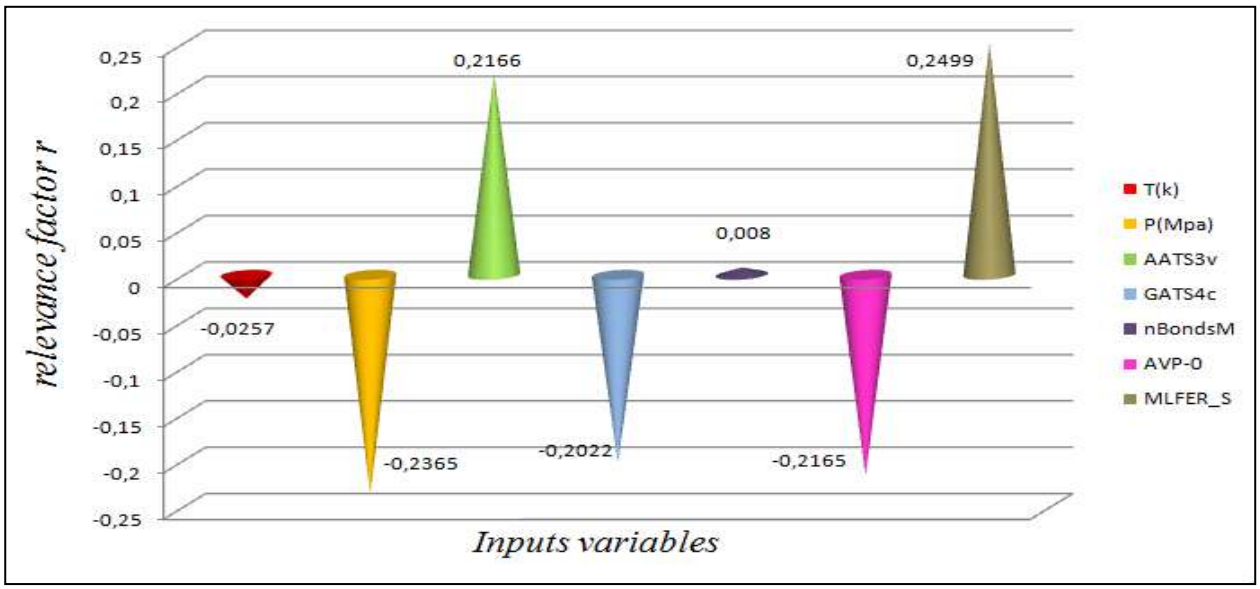


Figure .III.33. Results of the sensitivity analysis performed on the ANFIS model.

III.10.2.Conclusion

The Adaptive Neuro-Fuzzy Inference System was applied to build a model able to predict the target (solubility) base on a limited number of features (7) due to a lack of memory issues. Those were determined by using the GA.

This building model has only taken about half the total number of the best-fit selected descriptors of the QSAR model. Despite that, it could show and retch to results that can be called valid. This last, proved and shows the power of the fuzzy interface system modeling.

Model vs. Model

III.11. Evaluation and performance comparison of the optimized models

Computer science provides a large number of algorithms able to be used in any domain. Depending on the application, those algorithms can be seen as powerful tools as well as the opposite. However, the major obstacle is to clarify which of the views can be given to the algorithm. The concept "try and see" is one of the successful methods to compare those algorithms and give a critical decision on which algorithm can represent the target property better and which one is highly suggested to be not used. This notion can be only applied for fixed data used to precept a specified property or phenomena as long as the developer respect the capacity of each algorithm. Viz. models must be correctly used with respect to its limit of operation.

In order to visualize the results of modeling by machine learning methods, all considered acceptable, we will end this chapter with a summary of the comparison between those models in the aim to finally decide which of them is the most candidate to replacing the experiential trails.

Figure .III.34 shows a comparison of the AARD% obtained by the selected applied models.

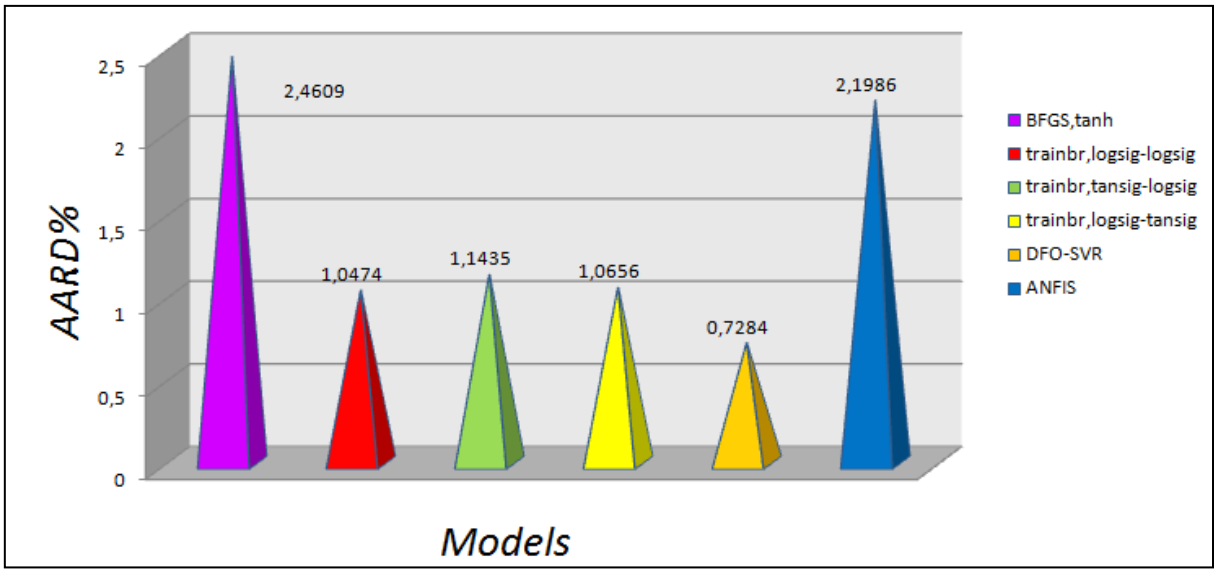


Figure .III.34. Comparison of AARD% global.

The comparison of the generalization capacity and the tendency of over-learning of the different models is illustrated by the distribution figures of the ER of the training and test sets in **Figure .III.35** and **.III.36** respectively.

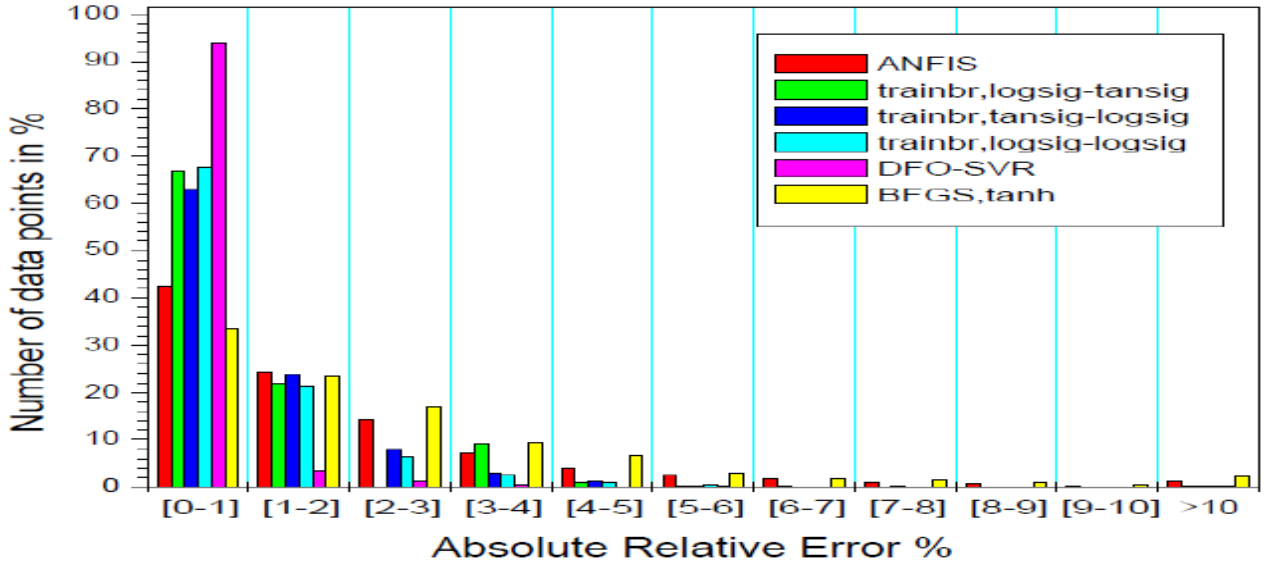


Figure .III.35. Bar distribution diagrams of the Absolute Relative Error [%] vs. number of DP in %, training set.

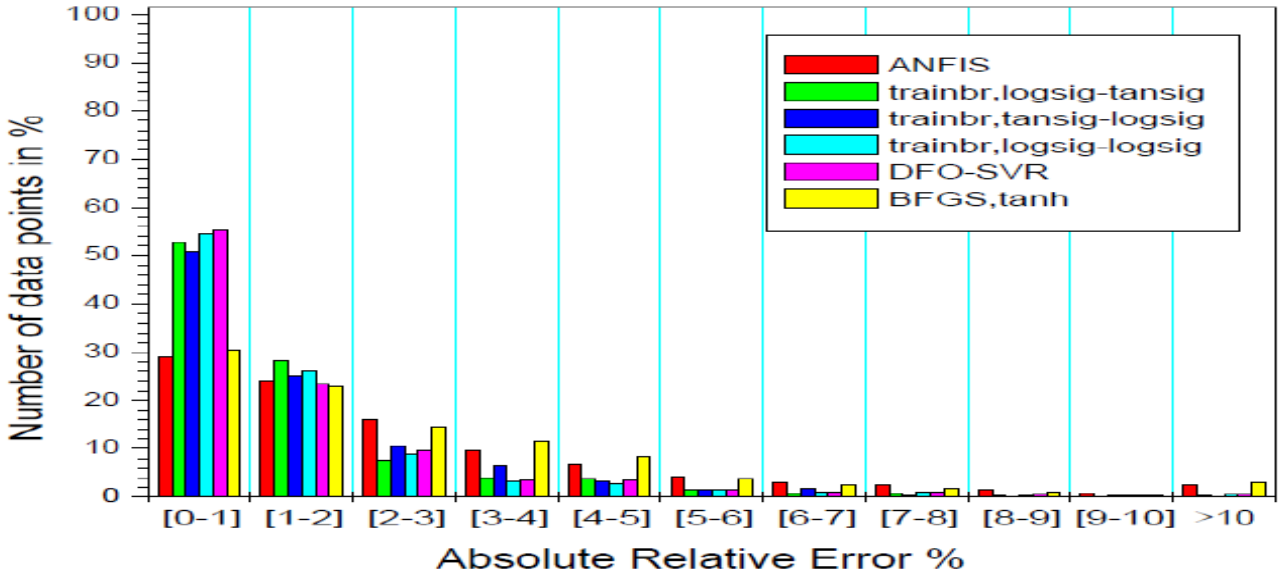


Figure .III.36. Bar distribution diagrams of the Absolute Relative Error [%] vs. number of DP in %, test set.

Table .III.24 shows a comparison of the differences between the AARD% of the test set and the training set.

Table .III.24. Comparison of the differences between the AARD% of the test set and the training set for the selected applied models.

	Train	**Test**	**Δ(test-train)**
ANN-BFGS,tanh	2.4284	2.6087	0.1803
ANN-trainbr,logsig-logsig	0.9867	1.3910	0.4043
ANN-trainbr,tansig-logsig	1.0888	1.4528	0.3640
ANN-trainbr,logsig-tansig	1.0093	1.3844	0.3751
DFO-SVR	0.4032	1.4873	1.0841
ANFIS	1.9528	2.7414	0.7886

The points of the comparison that has been obtained from these results are summarized as follows:

Figure .III.34.

- The model BFGS, tanh has the lowest generalization capacity while the DFO-SVR model has the highest one.
- The use of two hidden layers had decreased the AARD% into about half the result obtained while using one hidden layer. The three trainbr model shows quite similar results.
- The ANFIS model managed to arrive into performance comparable to the model BFGS, tanh despite the fact that the model has used only the half number of descriptors.

Figure .III.35. & Figure .III.36.

- The model DFO-SVR represents the model with the best distribution of the ER% for both the training and test sets.

Table .III.24.

- The BFGS, tanh model show the smallest deviations (0.1803) compared to the other methods.
- The biggest deviations were seen from the model DFO-SVR (1.0841). Notice that this deviation has been seen several times from developers while manipulation with SVM. The reason behind it was not defined. However, we can assume that it is not an over-fitting (R^2 -Q^2 <0.3).

In conclusion, the model DFO-SVR is valid and the best candidate for predicting future data and can be used to predict the solubility of drug-like compounds, particularly for those that have not been tested as well as new compounds.

III.12.Conclusion

The main goal of the current study was to determine a model for predicting the molar solubility fraction of drug-like compounds in supercritical carbon dioxide based on their molecule structures, that have been acquired while applying the QSPR modeling, and based on two independent intensive state variables, the temperature, and the pressure.

For that, the molar solubility fraction (y_2) of a total number of 133 drug-like compounds from various literature have been collected in the aim to form the final database of 13 selected features (AATS3v, MATS2e, GATS4c, GATS3v, GATS4e, GATS3s, nBondsM, AVP-0, SHBd, MLogP, and MLFER_S with the addition of T(k) and P(Mpa)), and 3590 experimental data points that have been used to build models based on four types of algorithms: two ANN model "MLP with one, and two hidden layers", a hybrid DFO-SVR, and ANFIS.

The models developed are all statistically valid with a global AARD between 0.73% and 2.46%. However, the models obtained by the DFO-SVR method have more precise correlation performances and significantly better generalization and interpolation capacities. The analysis of the data revealed the existence of a certain number of suspicious data and made it possible to define the domain of applicability of the models developed.

In summary, all the models are valid. While the best-fit candidate for predicting future data and can be used to predict the solubility of drug-like compounds, particularly for those that have not been tested as well as new compounds, is the model DFO-SVR.

Conclusion

Conclusion

This project was undertaken to highlight the solubility, as the number one influence thermodynamic property, in the formulation development to achieve the bioavailability and therapeutic action of the drug at the target site. As well as indicating one of the most renowned high-pressure techniques so far to overcome such kind of issues well-known as the supercritical fluid technology, the most used is the supercritical carbon dioxide.

In this thesis for the aim of avoiding the defects of traditional thermodynamic dissolution forecasting methods, and limit the use of tedious and expensive experiments we had developed a machine learning model able to predict the molar solubility fraction of drug-like compounds in supercritical carbon dioxide based on their molecule structures, and based on two independent intensive state variables, the temperature, and the pressure.

A QSPR model was first applied to select the best-fit combination able to represent the target property which ends up meeting all of the OECD principles for QSAR validation. The molar solubility fraction (y_2) of a total number of 133 drug-like compounds from various literature have been collected to form the final database of 13 selected features (AATS3v, MATS2e, GATS4c, GATS3v, GATS4e, GATS3s, nBondsM, AVP-0, SHBd, MLogP, and MLFER_S with the addition of T(k) and P(Mpa)), and 3590 experimental data points that have been used to build and select the best representative model, in the form of a comparative study between three computational intelligence techniques: MLP-ANNs with one and two hidden layers, ANFIS and DFO-SVR.

As a summary, the results of all the models were statistically valid with a global AARD between 0.73% and 2.46%. While the best-fit candidate for predicting future data and can be used to predict the solubility of drug-like compounds, particularly for those that have not been tested as well as new compounds, is the model DFO-SVR.

Returning to the issue posed at the beginning of this study, the DFO-SVR model may not be a total replacement to experiences, based on the fact of the impossibility to build an ideal model able to predict exactly all and always with the same quality the experimenter's trails. However, it is now possible to state that the DFO-SVR model gives the necessary performance to predict, better than any model before, the target property. Furthermore, it could be used as a statistical selecting tool to the experiences domain. In the simplest explanation, we propose the following:

Conclusion

For one compound, the experimenter usually does an average of 25 experience points. In the aim to validate the results the experimenter must at least repeat the trails three (3) times for every single experiment point. The total number of trails would be about 75 experience points. That is to say if we eliminated all the other obstacles and difficulties.

On the other hand, that's is the case of using the DFO-SVR model as a selecting tool, a caplet lines of coding is sufficient to write a function able to select, from a various number of random temperature and pressures for the same selected compound, the top five (5) candidate. The total number of experience points would be 15. Which is five times less.

Indeed, this realization project was a very fruitful, beneficial, and enriching experience. It allowed us to exploit the different theoretical concepts of chemoinformatics. It also allowed us to take the first step into the practical world of machine learning developers.

This project we put directly in contact with powerful programming algorithms such as ANNs, SVR, ANFIS, GA, DFO,... one of the powerful programming language MatLab, as well as various software specialized in this field like STATISTICA, and Chembio draw software.

To this end, we hope that we will improve our work in the future. As perspectives we propose:

- ✓ Improving the model accuracy.
- ✓ Modeling the solubility by hybrid algorithms.
- ✓ Modeling the solubility by the programming language Python.
- ✓ Modify the QSAR/QSPR model construction such a selecting different types of descriptors.
- ✓ Deep dive by specifying for each model the data type that can its predicts better (small, medium, or large).

Bibliography

Bibliography

[1]. University of Copenhagen - Study at UCPH From Idea to Drug,. https://studies.ku.dk/masters/pharmaceutical-sciences/from-idea-to-drug/.

[2]. Vallianatou, T.; Giaginis, C.; Tsantili-Kakoulidou, A., The impact of physicochemical and molecular properties in drug design: navigation in the "drug-like" chemical space. *Advances in experimental medicine and biology* 2015, *822*, 187-94.

[3]. Vimalson, D. C., Techniques to enhance solubility of hydrophobic drugs: an overview. *Asian Journal of Pharmaceutics (AJP): Free full text articles from Asian J Pharm* 2016, *10* (2).

[4]. Savjani, K. T.; Gajjar, A. K.; Savjani, J. K., Drug solubility: importance and enhancement techniques. *ISRN pharmaceutics* 2012, *2012*.

[5]. Kakran, M.; Li, L.; Muller, R. H., Overcoming the challenge of poor drug solubility. *Pharm Eng* 2012, *32* (7-8), 1-7.

[6]. Parhi, R.; Suresh, P.; Ghan, R., Supercritical Fluid Technology: A Review. *Journal of Advanced Pharmaceutical Science And Technology* 2013, *1* (1), 13-36.

[7]. Thakkar, F. M. V.; Soni, T.; Gohel, M.; Gandhi, T., Supercritical fluid technology: a promising approach to enhance the drug solubility. *Journal of pharmaceutical sciences and research* 2009, *1* (4), 1.

[8]. Dictionnaire Cambridge Advanced Learner's & Thesaurus © Cambridge University Press Solubility Signification Definition dans le Dictionary Anglais de Cambridge. https://dictionary.cambridge.org/fr/dictionnaire/anglais/solubility.

[9]. Business Dictionary what is solubility ? definition and meaning. http://www.businessdictionary.com/definition/solubility.html.

[10]. ScienceDaily Solubility,. https://www.sciencedaily.com/terms/solubility.htm.

[11]. FLORIDA STATE UNIVERSITY Department of Chemistry & Biochemistry,Solubility,. https://www.chem.fsu.edu/chemlab/chm1046course/solubility.html.

[12]. Anne Marie Helmenstine Solubility Definition in Chemistry,. https://www.thoughtco.com/definition-of-solubility-604649.

[13]. BAUD Corine La solubilite - theorie,. https://edu.ge.ch/chavanne/cours/cours/baudc/2os-2019-2020/la-solubilite/theorie/view.

[14]. IDC TECHNOLOGIES, Solubility Product,. http://www.idc-online.com/technical_references/pdfs/chemical_engineering/Solubility_product.pdf.

[15]. Solubility of Things, Factors affecting solubility,. https://www.solubilityofthings.com/basics/factors_affecting_solubility.php.

[16]. Lumencandela Factors Affecting Solubility,. https://courses.lumenlearning.com/boundless-chemistry/chapter/factors-affecting-solubility/.

[17]. CHEMISTRY, Solubility and Factors Affecting Solubility,. https://chem.libretexts.org/Bookshelves/Physical_and_Theoretical_Chemistry_Textbook_Maps/Supplemental_Modules_(Physical_and_Theoretical_Chemistry)/Equilibria/Solubilty/Solubility_and_Factors_Affecting_Solubility.

[18]. Jouyban, A., *Handbook of solubility data for pharmaceuticals*. Crc Press: 2009.

[19]. Tihanyi, K. K.; Vastag, M., *Solubility, delivery and ADME problems of drugs and drug candidates*. Bentham Science Publishers: 2011.

[20]. Kakran, M.; Li, L.; Müller, R. H., Overcoming the challenge of poor drug solubility. *Pharm Eng* 2012, *32* (7-8), 1-7.

[21]. Bright, F. V.; McNally, M. E. P. *Supercritical fluid technology-theoretical and applied approaches to analytical chemistry*; Washington, DC (United States); American Chemical Society: 1992.

[22]. lumencandela, Supercritical Fluids,. https://courses.lumenlearning.com/introchem/chapter/supercritical-fluids/.

[23]. Deshpande, P. B.; Kumar, G. A.; Kumar, A. R.; Shavi, G. V.; Karthik, A.; Reddy, M. S.; Udupa, N., Supercritical fluid technology: concepts and pharmaceutical applications. *PDA journal of pharmaceutical science and technology* 2011, *65* (3), 333-344.

[24]. Hitchen, S.; Dean, J., Properties of supercritical fluids. In *Applications of Supercritical Fluids in Industrial Analysis*, Springer: 1993; pp 1-11.

[25]. Hamidreza, A. B.; Reza, B. S.; Mahdi, C. S.; Mahdi, D. Y.; Fatemeh, E. F.; Hassan, F. H., Nanotechnology and supercritical fluids. *Journal of Fundamental and Applied Sciences* 2016, *8* (2), 839-859.

[26]. Vaferi, B.; Karimi, M.; Azizi, M.; Esmaeili, H., Comparison between the artificial neural network, SAFT and PRSV approach in obtaining the solubility of solid aromatic compounds in supercritical carbon dioxide. *The Journal of Supercritical Fluids* 2013, *77*, 44-51.

[27]. Kankala, R. K.; Zhang, Y. S.; Wang, S. B.; Lee, C. H.; Chen, A. Z., Supercritical fluid technology: an emphasis on drug delivery and related biomedical applications. *Advanced healthcare materials* 2017, *6* (16), 1700433.

[28]. Girotra, P.; Singh, S. K.; Nagpal, K., Supercritical fluid technology: a promising approach in pharmaceutical research. *Pharmaceutical development and technology* 2013, *18* (1), 22-38.

[29]. TechTarget database (DB),. https://searchsqlserver.techtarget.com/definition/database.

[30]. ENCYCLOPOPADIA BRITANNICA, Database,. https://www.britannica.com/technology/database.

[31]. Weininger, D., SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of chemical information and computer sciences* 1988, *28* (1), 31-36.

[32]. Leach, A. R.; Gillet, V. J., Representation and manipulation of 2D molecular structures. *An Introduction to Chemoinformatics* 2007, 1-25.

[33]. BMC Mordred: a molecular descriptor calculator,. https://jcheminf.biomedcentral.com/articles/10.1186/s13321-018-0258-y.

[34]. Clarancia, S.; Dhanjal, J. K.; Malik, V.; Radhakrishnan, N.; Mannu, J.; Sundar, D., Quantitative Structure-Activity Relationship (QSAR): Modeling Approaches to Biological Applications. 2018.

[35]. Roy, K.; Kar, S.; Das, R. N., Statistical methods in QSAR/QSPR. In *A primer on QSAR/QSPR modeling*, Springer: 2015; pp 37-59.

[36]. Chinta, S.; Rengaswamy, R., Machine Learning Derived Quantitative Structure Property Relationship (QSPR) to predict drug solubility in binary solvent systems. *Industrial & Engineering Chemistry Research* 2019, *58* (8), 3082-3092.

[37]. Muhammad, U.; Uzairu, A.; Ebuka Arthur, D., Review on: Quantitative structure activity relationship (QSAR) modeling. *J. Anal. Pharm. Res* 2018, *7*.

[38]. Gasteiger, J., Chemoinformatics: Achievements and challenges, a personal view. *Molecules* 2016, *21* (2), 151.

[39]. Chen, S. H.; Jakeman, A. J.; Norton, J. P., Artificial intelligence techniques: an introduction to their use for modelling environmental systems. *Mathematics and computers in simulation* 2008, *78* (2-3), 379-400.

[40]. Mueller, J. P.; Massaron, L., *Machine learning for dummies*. John Wiley & Sons: 2016.

[41]. Tewari, K.; Vandita, S.; Jain, S., Predictive Analysis of Absenteeism in MNCS Using Machine Learning Algorithm. In *Proceedings of ICRIC 2019*, Springer: 2020; pp 3-14.

[42]. Dhall, D.; Kaur, R.; Juneja, M., Machine Learning: A Review of the Algorithms and Its Applications. In *Proceedings of ICRIC 2019*, Springer: 2020; pp 47-63.

[43]. Roy, K.; Roy, P.; Leonard, J., On some aspects of validation of predictive QSAR models. *Chemistry Central Journal* 2008, *2* (S1), P9.

[44].Goodarzi,M.,Dejaegher,B.,Heyden,Y.V.,2012.FeatureselectionmethodsinQSARstudies.Journal ofAOACInternational95,636–651.

[45]. Roy, K., Kar, S., Das, R.N., 2015a. A Primer on QSAR/QSPR Modeling: Fundamental Concepts. Springer.

[46]. Badens, E.; Masmoudi, Y.; Mouahid, A.; Crampon, C., Current situation and perspectives in drug formulation by using supercritical fluid technology. *The Journal of Supercritical Fluids* 2018, *134*, 274-283.

[47]. academia, Introduction to Machine Learning The Wikipedia Guide,. https://www.academia.edu/41157657/Introduction_to_Machine_Learning_The_Wikipedia_Guide.

[48]. González-Arjona, D.; López-Pérez, G.; González, A. G., Non-linear QSAR modeling by using multilayer perceptron feedforward neural networks trained by back-propagation. *Talanta* 2002, *56* (1), 79-90.

[49]. Roy, K.; Mandal, A. S., Development of linear and nonlinear predictive QSAR models and their external validation using molecular similarity principle for anti-HIV indolyl aryl sulfones. *Journal of enzyme inhibition and medicinal chemistry* 2008, *23* (6), 980-995.

[50]. Veerasamy, R.; Rajak, H.; Jain, A.; Sivadasan, S.; Varghese, C. P.; Agrawal, R. K., Validation of QSAR models-strategies and importance. *Int. J. Drug Des. Discov* 2011, *2* (3), 511-519.

[51]. Roy, K.; Das, R. N.; Ambure, P.; Aher, R. B., Be aware of error measures. Further studies on validation of predictive QSAR models. *Chemometrics and Intelligent Laboratory Systems* 2016, *152*, 18-33.

[52]. Hamadache, M.; Hanini, S.; Benkortbi, O.; Amrane, A.; Khaouane, L.; Moussa, C. S., Artificial neural network-based equation to predict the toxicity of herbicides on rats. *Chemometrics and Intelligent Laboratory Systems* 2016, *154*, 7-15.

[53]. Roy, K.; Mitra, I., On various metrics used for validation of predictive QSAR models with applications in virtual screening and focused library design. *Combinatorial chemistry & high throughput screening* 2011, *14* (6), 450-474.

[54]. Mansourian, M.; Saghaie, L.; Fassihi, A.; Madadkar-Sobhani, A.; Mahnam, K., Linear and nonlinear QSAR modeling of 1, 3, 8-substituted-9-deazaxanthines as potential selective A 2B AR antagonists. *Medicinal Chemistry Research* 2013, *22* (10), 4549-4567.

[55]. Baghban, A.; Ahmadi, M. A.; Shahraki, B. H., Prediction carbon dioxide solubility in presence of various ionic liquids using computational intelligence approaches. *The Journal of Supercritical Fluids* 2015, *98*, 50-64.

[56]. Suzuki, K., *Artificial neural networks: methodological advances and biomedical applications*. BoD–Books on Demand: 2011.

[57]. Dr. Saed Sayad, Artificial Neural Network,. https://www.saedsayad.com/artificial_neural_network.htm.

[58]. Hu, X., Support Vector Machine and Its Application to Regression and Classification. 2017.

[59]. Jakkula, V., Tutorial on support vector machine (svm). *School of EECS, Washington State University* 2006, *37*.

[60]. Mokhtarzad, M.; Eskandari, F.; Vanjani, N. J.; Arabasadi, A., Drought forecasting by ANN, ANFIS, and SVM and comparison of the models. *Environmental Earth Sciences* 2017, *76* (21), 729.

[61]. Steinwart, I.; Christmann, A., *Support vector machines*. Springer Science & Business Media: 2008.

[62]. Amroune, M.; Bouktir, T.; Musirin, I., Power system voltage stability assessment using a hybrid approach combining dragonfly optimization algorithm and support vector regression. *Arabian Journal for Science and Engineering* 2018, *43* (6), 3023-3036.

[63]. Ertekin, S., Learning in extreme conditions: Online and active learning with massive, imbalanced and noisy data. 2009.

[64].Mirjalili, S.: Dragonfly algorithm: a new meta-heuristic optimization technique for solving single-objective, discrete, and multi-objective problems. Neural Comput. Appl. 27(4), 1053–1073 (2015).

[65]. Czabanski, R.; Jezewski, M.; Leski, J., Introduction to fuzzy systems. In *Theory and Applications of Ordered Fuzzy Numbers*, Springer, Cham: 2017; pp 23-43.

[66]. shodhganga Chapter 5 Adaptive Neuro Fuzzy Inference System,. http://shodhganga.inflibnet.ac.in/bitstream/10603/93079/11/11_chapter%205.pdf.

[67]. Chang, F.-J.; Chang, Y.-T., Adaptive neuro-fuzzy inference system for prediction of water level in reservoir. *Advances in water resources* 2006, *29* (1), 1-10.

68]. Sonmez, A. Y.; Kale, S.; Ozdemir, R. C.; Kadak, A. E., An Adaptive Neuro-Fuzzy Inference System (ANFIS) to Predict of Cadmium (Cd) Concentrations in the Filyos River, Turkey. *Turkish Journal of Fisheries and Aquatic Sciences* 2018, *18* (12), 1333-1343.

[69]. Zarei, K.; Taheri, F., Prediction of solubility of some statin drugs in supercritical carbon dioxide using classification and regression tree analysis and adaptive neuro-fuzzy inference systems. *Russian Chemical Bulletin* 2016, *65* (4), 1131-1138.

[70]. Shojaeian, A.; Asadizadeh, M., Prediction of surface tension of the binary mixtures containing ionic liquid using heuristic approaches; an input parameters investigation. *Journal of Molecular Liquids* 2020, *298*, 111976.

[71]. Davis, T.E., Principe, J.C.: A Markov chain framework for the simple genetic algorithm. Evol. Comput. 1(3), 269–288 (1993).

[72]. Paul, Y.; Kumar, N., A Comparative Study of Famous Classification Techniques and Data Mining Tools. In *Proceedings of ICRIC 2019*, Springer: 2020; pp 627-644.

[73] B.P. Buckeles, F.E. Petry, Genetic Algorithms, IEEE Computer Society Press, Los Alamitos, CA, 1992.

[74]. Rad, H. B.; Sabet, J. K.; Varaminian, F., STUDY OF SOLUBILITY IN SUPERCRITICAL FLUIDS: THERMODYNAMIC CONCEPTS AND MEASUREMENT METHODS-A REVIEW. *Brazilian Journal of Chemical Engineering* 2019, *36* (4), 1367-1392.

[75]. Matos, R. L.; de Melo, S. A. V.; Albuquerque, E. C. C.; Foster, N. R., Dense CO2 technology: Overview of recent applications for drug processing/formulation/delivery. *Chemical Engineering and Processing-Process Intensification* 2019.

[76]. Soleimani, R.; Dehaghani, A. H. S.; Shoushtari, N. A.; Yaghoubi, P.; Bahadori, A., Toward an intelligent approach for predicting surface tension of binary mixtures containing ionic liquids. *Korean Journal of Chemical Engineering* 2018, *35* (7), 1556-1569.

[77]. Todeschini, R.; Ballabio, D.; Grisoni, F., Beware of unreliable Q 2! A comparative study of regression metrics for predictivity assessment of QSAR models. *Journal of chemical information and modeling* 2016, *56* (10), 1905-1913.

[78]. Gevrey M., Dimopoulos I. et Lek S., Review and comparison of methods to study the contribution of variables in artificial neural network models, Ecological Modelling,2003.

[79]. Baghban J.A., Shafiee M., Ahmadi M.H. et Chau K.W., Developing an ANFIS-based swarm concept model for estimating the relative viscosity of nanofluids, Engineering Applications of Computational Fluid Mechanics, 2018.

[80]. Chen, Y.-M.; Chen, Y.-P., Measurements for the solid solubilities of antipyrine, 4-aminoantipyrine and 4-dimethylaminoantipyrine in supercritical carbon dioxide. *Fluid phase equilibria* 2009, *282* (2), 82-87.

[81]. J.-S. Jin, X. Fan, H. Zhang, Y.-W. Wang, Z.-T. Zhang, J. Chem. Eng.Data 59 (2014) 2095.

[82]. Suleiman, D.; Estévez, L. A.; Pulido, J. C.; García, J. E.; Mojica, C., Solubility of anti-inflammatory, anti-cancer, and anti-HIV drugs in supercritical carbon dioxide. *Journal of Chemical & Engineering Data* 2005, *50* (4), 1234-1241.

[83]. Lamba, N.; Narayan, R. C.; Modak, J.; Madras, G., Solubilities of 10-undecenoic acid and geraniol in supercritical carbon dioxide. *The Journal of Supercritical Fluids* 2016, *107*, 384-391.

[84]. Rosa, A. B.; Marceneiro, S.; Braga, M. E.; Dias, A. M.; de Sousa, H. C., Solubility of all-trans retinoic acid in supercritical carbon dioxide. *The Journal of Supercritical Fluids* 2015, *98*, 70-78.

[85]. Sodeifian, G.; Sajadian, S. A.; Razmimanesh, F., Solubility of an antiarrhythmic drug (amiodarone hydrochloride) in supercritical carbon dioxide: experimental and modeling. *Fluid phase equilibria* 2017, *450*, 149-159.

[86]. M. Hojjati, A. Vatanara, Y. Yamini, M. Moradi, A.R. Najafabadi, J.Supercrit. Fluids 50 (2009) 203.

[87]. Sodeifian, G.; Sajadian, S. A.; Ardestani, N. S., Determination of solubility of Aprepitant (an antiemetic drug for chemotherapy) in supercritical carbon dioxide: empirical and thermodynamic models. *The Journal of Supercritical Fluids* 2017, *128*, 102-111.

[88]. Z. Huang, W.D. Lu, S. Kawi, Y.C. Chiew, J. Chem. Eng. Data 49 (2004) 1323.

[89]. Yamini, Y.; Hassan, J.; Haghgo, S., Solubilities of some nitrogen-containing drugs in supercritical carbon dioxide. *Journal of Chemical & Engineering Data* 2001, *46* (2), 451-455.

[90]. Hojjati, M.; Yamini, Y.; Khajeh, M.; Vatanara, A., Solubility of some statin drugs in supercritical carbon dioxide and representing the solute solubility data with several density-based correlations. *The Journal of Supercritical Fluids* 2007, *41* (2), 187-194.

[91]. Sodeifian, G.; Razmimanesh, F.; Ardestani, N. S.; Sajadian, S. A., Experimental data and thermodynamic modeling of solubility of Azathioprine, as an immunosuppressive and anti-cancer drug, in supercritical carbon dioxide. *Journal of Molecular Liquids* 2020, *299*, 112179.

[92]. Sparks, D. L.; Hernandez, R.; Estévez, L. A.; Meyer, N.; French, T., Solubility of azelaic acid in supercritical carbon dioxide. *Journal of Chemical & Engineering Data* 2007, *52* (4), 1246-1249.

[93]. Garmroodi, A.; Hassan, J.; Yamini, Y., Solubilities of the drugs benzocaine, metronidazole benzoate, and naproxen in supercritical carbon dioxide. *Journal of Chemical & Engineering Data* 2004, *49* (3), 709-712.

[94]. Asghari-Khiavi, M.; Yamini, Y., Solubility of the drugs bisacodyl, methimazole, methylparaben, and iodoquinol in supercritical carbon dioxide. *Journal of Chemical & Engineering Data* 2003, *48* (1), 61-65.

[95].A. Vatanara, A.R. Najafabadi, M. Khajeh, Y. Yamini, J. Supercrit.

[96]. Fluids 33 (2005) 21.M. Johannsen, G. Brunner, Fluid Phase Equilib. 95 (1994) 215.

[97]. Yamini, Y.; Hojjati, M.; Kalantarian, P.; Moradi, M.; Esrafili, A.; Vatanara, A., Solubility of capecitabine and docetaxel in supercritical carbon dioxide: Data and the best correlation. *Thermochimica acta* 2012, *549*, 95-101.

[98]. Khamda, M.; Hosseini, M. H.; Rezaee, M., Measurement and correlation solubility of cefixime trihydrate and oxymetholone in supercritical carbon dioxide (CO2). *The Journal of Supercritical Fluids* 2013, *73*, 130-137.

[99]. Ongkasin, K.; Sauceau, M.; Masmoudi, Y.; Fages, J.; Badens, E., Solubility of cefuroxime axetil in supercritical CO2: Measurement and modeling. *The Journal of Supercritical Fluids* 2019, *152*, 104498.

[100]. C.-C. Tsai, H.-M. Lin, M.-J. Lee, J. Supercrit. Fluids 95 (2014) 17.

[101]. A.Z. Hezave, A. Mowla, F. Esmaeilzadeh, J. Supercrit. Fluids 58(2011) 198.

[102]. Li, B.; Guo, W.; Ramsey, E. D., Solubility Measurements of Chloramphenicol in Supercritical Fluid CO2 Using Static Solubility Apparatus Interfaced with Online Supercritical Fluid Chromatography. *Journal of Chemical & Engineering Data* 2019.

[103]. C.-A. Lee, M. Tang, S.-L. Ho, Y.-P. Chen, J. Supercrit. Fluids 85 (2014)11.

[104]. M. Lashkarbolooki, A.Z. Hezave, Y. Rahnama, H. Rajaei, F. Esmaeilzadeh, J. Supercrit. Fluids 84 (2013) 29.

[105]. Manna, L.; Banchero, M., Solubility of tolbutamide and chlorpropamide in supercritical carbon dioxide. *Journal of Chemical & Engineering Data* 2018, *63* (5), 1745-1751.

[106]. Paula, J. T.; Sousa, I. M.; Foglio, M. A.; Cabral, F. A., Solubility of protocatechuic acid, sinapic acid and chrysin in supercritical carbon dioxide. *The Journal of Supercritical Fluids* 2016, *112*, 89-94.

[107]. Zhang, C.; Jia, D.; Sun, Y.; Zhang, L. In *Measurement and Correlation of the Solubility of Cinildipine in Supercritical CO2*, IOP Conference Series: Earth and Environmental Science, IOP Publishing: 2018; p 012041.

[108]. C.-A. Lee, M. Tang, Y.-P. Chen, Fluid Phase Equilib. 367 (2014) 182.

[109]. Shi, K.; Feng, L.; He, L.; Li, H., Solubility determination and correlation of gatifloxacin, enrofloxacin, and ciprofloxacin in supercritical CO2. *Journal of Chemical & Engineering Data* 2017, *62* (12), 4235-4243.

[110]. C.I. Park, M.S. Shin, H. Kim, J. Chem. Thermodyn. 41 (2009) 30.

[111]. H. Asiabi, Y. Yamini, M. Tayyebi, M. Moradi, A. Vatanara,K. Keshmiri, J. Supercrit. Fluids 78 (2013) 28.

[112]. Li, H.; Jia, D.; Zhu, Q.; Shen, B., Determination, correlation and prediction of the solubilities of niflumic acid, clofenamic acid and tolfenamic acid in supercritical CO2. *Fluid phase equilibria* 2015, *392*, 95-103.

[113]. Yamini, Y.; Moradi, M., Measurement and correlation of antifungal drugs solubility in pure supercritical CO2 using semiempirical models. *The Journal of Chemical Thermodynamics* 2011, *43* (7), 1091-1096.

[114]. Hosseini, M. H.; Alizadeh, N.; Khanchi, A. R., Solubility analysis of clozapine and lamotrigine in supercritical carbon dioxide using static system. *The Journal of Supercritical Fluids* 2010, *52* (1), 30-35.

[115]. J.R. Dean, M. Kane, S. Khundker, C. Dowle, R.L. Tranter, P. Jones,
Analyst 120(8) 2153.

[116]. M. Lashkarbolooki, A.Z. Hezave, Y. Rahnama, R. Ozlati, H. Rajaei,F. Esmaeilzadeh, J. Supercrit. Fluids 84 (2013) 13.

[117]. M. Asghari-Khiavi, Y. Yamini, M.A. Farajzadeh, J. Supercrit. Fluids 30 (2004) 111.

[118]. R.B. Chim, M.B.C. De Matos, M.E.M. Braga, A.M.A. Dias, H.C. De Sousa, J. Chem. Eng. Data 57 (2012) 3756.

[119]. C. Tang, Y.-X. Guan, S.-J. Yao, Z.-Q. Zhu, J. Chem. Eng. Data 59 (11)(2014) 3359.

[120]. Zeinolabedini Hezave, A.; Esmaeilzadeh, F., Solubility measurement of diclofenac acid in the supercritical CO2. *Journal of Chemical & Engineering Data* 2012, *57* (6), 1659-1664.

[121]. Coimbra, P.; Fernandes, D.; Gil, M. H.; Sousa, H. C. d., Solubility of diflunisal in supercritical carbon dioxide. *Journal of Chemical & Engineering Data* 2008, *53* (8), 1990-1995.

[122]. Y. Yamini, P. Kalantarian, M. Hojjati, A. Esrafily, M. Moradi, A. Vatanara, I. Harrian, J. Chem. Eng. Data 55 (2010) 1056.

[123]. Sodeifian, G.; Detakhsheshpour, R.; Sajadian, S. A., Experimental study and thermodynamic modeling of Esomeprazole (proton-pump inhibitor drug for stomach acid reduction) solubility in supercritical carbon dioxide. *The Journal of Supercritical Fluids* 2019, *154*, 104606.

[124]. Zha, X.; Han, S.; Wang, W.; Jiao, Z., Experimental measurement and correlation of solubility of ethosuximide in supercritical carbon dioxide. *The Journal of Chemical Thermodynamics* 2019, *131*, 104-110.

[125]. Hezave, A. Z.; Rajaei, H.; Lashkarbolooki, M.; Esmaeilzadeh, F., Analyzing the solubility of fluoxetine hydrochloride in supercritical carbon dioxide. *The Journal of Supercritical Fluids* 2013, *73*, 57-62.

[126]. Duarte, A. R. C.; Coimbra, P.; de Sousa, H. C.; Duarte, C. M., Solubility of flurbiprofen in supercritical carbon dioxide. *Journal of Chemical & Engineering Data* 2004, *49* (3), 449-452.

[127]. S.A. Shojaee, A.Z. Hezave, S. Aftab, M. Lashkarbolooki, F. Esmaeilzadeh, J. Supercrit. Fluids 78 (2013) 1.

[128]. Xiang, S.-T.; Chen, B.-Q.; Kankala, R. K.; Wang, S.-B.; Chen, A.-Z., Solubility measurement and RESOLV-assisted nanonization of gambogic acid in supercritical carbon dioxide for cancer therapy. *The Journal of Supercritical Fluids* 2019, *150*, 147-155.

[129]. Chen, Y.-M.; Lin, P.-C.; Tang, M.; Chen, Y.-P., Solid solubility of antilipemic agents and micronization of gemfibrozil in supercritical carbon dioxide. *The Journal of Supercritical Fluids* 2010, *52* (2), 175-182.

[130]. Jara-Morante, E.; Suleiman, D.; Estévez, L. A., Solubilities of imipramine HCl in supercritical carbon dioxide. *Industrial & Engineering Chemistry Research* 2003, *42* (8), 1821-1823.

[131]. Wang, B.-C.; Su, C.-S., Solid solubility measurement of ipriflavone in supercritical carbon dioxide and microparticle production through the rapid expansion of supercritical solutions process. *Journal of CO2 Utilization* 2020, *37*, 285-294.

[132]. Heryanto, R.; Abdullah, E. C.; Hasan, M., Solubility of isoniazid in supercritical carbon dioxide. *Journal of Chemical & Engineering Data* 2010, *55* (6), 2306-2309.

[133]. Sabegh, M. A.; Rajaei, H.; Esmaeilzadeh, F.; Lashkarbolooki, M., Solubility of ketoprofen in supercritical carbon dioxide. *The Journal of Supercritical Fluids* 2012, *72*, 191-197.

[134]. Sodeifian, G.; Ardestani, N. S.; Sajadian, S. A.; Panah, H. S., Measurement, correlation and thermodynamic modeling of the solubility of Ketotifen fumarate (KTF) in supercritical carbon dioxide: evaluation of PCP-SAFT equation of state. *Fluid phase equilibria* 2018, *458*, 102-114.

[135]. Wang, Y.; Meng, T.; Jia, D.; Sun, Y.; Li, N., Solubility of Nifedipine and Lacidipine in Supercritical CO 2: Measurement and Correlation. *Journal of Solution Chemistry* 2017, *46* (1), 70-88.

[136]. Sodeifian, G.; Sajadian, S. A.; Derakhsheshpour, R., Experimental measurement and thermodynamic modeling of Lansoprazole solubility in supercritical carbon dioxide: Application of SAFT-VR EoS. *Fluid phase equilibria* 2020, *507*, 112422.

[137]. Y. Yamini, M. Tayyebi, M. Moradi, A. Vatanara, Thermochim. Acta
569 (2013) 48.

[138]. Sodeifian, G.; Razmimanesh, F.; Sajadian, S. A.; Panah, H. S., Solubility measurement of an antihistamine drug (Loratadine) in supercritical carbon dioxide: Assessment of qCPA and PCP-SAFT equations of state. *Fluid phase equilibria* 2018, *472*, 147-159.

[139]. K.-W. Cheng, M. Tang, Y.-P. Chen, Fluid Phase Equilib. 201 (2002) 79.7

[140]. Hezave, A. Z.; Khademi, M. H.; Esmaeilzadeh, F., Measurement and modeling of mefenamic acid solubility in supercritical carbon dioxide. *Fluid phase equilibria* 2012, *313*, 140-147.

[141]. de Sousa, H. C.; Costa, M. S.; Coimbra, P.; Matias, A. A.; Duarte, C. M., Experimental determination and correlation of meloxicam sodium salt solubility in supercritical carbon dioxide. *The Journal of Supercritical Fluids* 2012, *63*, 40-45.

[142]. Yamini, Y.; Arab, J.; Asghari-khiavi, M., Solubilities of phenazopyridine, propranolol, and methimazole in supercritical carbon dioxide. *Journal of pharmaceutical and biomedical analysis* 2003, *32* (1), 181-187.

[143]. C.-S. Su, Y.-P. Chen, J. Supercrit. Fluids 43 (2008) 438.

[144]. Kotnik, P.; Škerget, M.; Knez, Z. e., Solubility of nicotinic acid and nicotinamide in carbon dioxide at T=(313.15 to 373.15) K and p=(5 to 30) MPa: experimental data and correlation. *Journal of Chemical & Engineering Data* 2011, *56* (2), 338-343.

[145]. Medina, I.; Bueno, J. L., Solubilities of zopiclone and nimodipine in supercritical carbon dioxide. *Journal of Chemical & Engineering Data* 2001, *46* (5), 1211-1214.

[146]. S. Zhan, L. Cui, Q. Zhao, L. Chen, J. Wang, S. Chen, S. Ding, J. Solution Chem. 44 (1) (2015) 1.

[147]. Chim, R.; Marceneiro, S.; Braga, M. E.; Dias, A. M.; de Sousa, H. C., Solubility of norfloxacin and ofloxacin in supercritical carbon dioxide. *Fluid phase equilibria* 2012, *331*, 6-11.

[148]. Tang, Z.; Jin, J.-s.; Zhang, Z.-t.; Liu, H.-t., New experimental data and modeling of the solubility of compounds in supercritical carbon dioxide. *Industrial & Engineering Chemistry Research* 2012, *51* (15), 5515-5526.

[149]. Sodeifian, G.; Hazaveie, S. M.; Sajadian, S. A.; Razmimanesh, F., Experimental investigation and modeling of the solubility of oxcarbazepine (an anticonvulsant agent) in supercritical carbon dioxide. *Fluid phase equilibria* 2019, *493*, 160-173.

[150]. Jiao, Z.; Cheng, J.; Guo, M.; Han, S., Measurement and Correlation of Paeonol Solubility in Supercritical Carbon Dioxide. *Journal of Chemical & Engineering Data* 2019, *64* (10), 4424-4429.

[151]. C. Garlapati, G. Madras, Thermochim. Acta 500 (2010) 123.

[152]. M. Ko, V. Shah, P.R. Bienkowski, H.D. Cochran, J. Supercrit. Fluids 4(1991) 32.

[153]. H. Rajaei, A.Z. Hezave, M. Lashkarbolooki, F. Esmaeilzadeh, J.Supercrit. Fluids 75 (2013) 181.

[154].R.D. Weinstein, J.J. Gribbin, K.R. Muske, J. Chem. Eng. Data 50 (2005) 226.

[155]. S.A. Shojaee, H. Rajaei, A.Z. Hezave, M. Lashkarbolooki, F. Esmaeilzadeh, J. Supercrit. Fluids 80 (2013) 38.

[156]. Chen, C.-T.; Lee, C.-A.; Tang, M.; Chen, Y.-P., Experimental investigation for the solubility and micronization of pyridin-4-amine in supercritical carbon dioxide. *Journal of CO2 Utilization* 2017, *18*, 173-180.

[157]. J. Garcia-Gonzalez, M.J. Molina, F. Rodrıguez, F. Mirada, J. Chem.Eng. Data 46 (2001) 918.

[158]. Li, B.; Guo, W.; Ramsey, E. D., Determining the solubility of nifedipine and quinine in supercritical fluid carbon dioxide using continuously stirred static solubility apparatus interfaced

with online supercritical fluid chromatography. *Journal of Chemical & Engineering Data* 2017, *62* (4), 1530-1537.

[159]. Sodeifian, G.; Hazaveie, S. M.; Sajadian, S. A.; Saadati Ardestani, N., Determination of the Solubility of the Repaglinide Drug in Supercritical Carbon Dioxide: Experimental Data and Thermodynamic Modeling. *Journal of Chemical & Engineering Data* 2019, *64* (12), 5338-5348.

[160]. Reverchon, E.; Donsi, G.; Gorgoglione, D., Salicylic acid solubilization in supercritical CO2 and its micronization by RESS. *The Journal of Supercritical Fluids* 1993, *6* (4), 241-248.

[161]. Sodeifian, G.; Sajadian, S. A., Experimental measurement of solubilities of sertraline hydrochloride in supercriticalcarbon dioxide with/without menthol: Data correlation. *The Journal of Supercritical Fluids* 2019, *149*, 79-87.

[162]. Sodeifian, G.; Razmimanesh, F.; Sajadian, S. A.; Hazaveie, S. M., Experimental data and thermodynamic modeling of solubility of Sorafenib tosylate, as an anti-cancer drug, in supercritical carbon dioxide: Evaluation of Wong-Sandler mixing rule. *The Journal of Chemical Thermodynamics* 2020, *142*, 105998.

[163] A.Z. Hezave, S. Shahnazar, H. Rajaei, M. Lashkarbolooki, F. Esmaeilzadeh, Fluid Phase Equilib. 355 (2013) 130.

[164]. Hezave, A. Z.; Aftab, S.; Esmaeilzadeh, F., Solubility of sulindac in the supercritical carbon dioxide: experimental and modeling approach. *The Journal of Supercritical Fluids* 2012, *68*, 39-44.

[165]. Sodeifian, G.; Razmimanesh, F.; Sajadian, S. A., Prediction of solubility of sunitinib malate (an anti-cancer drug) in supercritical carbon dioxide (SC–CO2): experimental correlations and thermodynamic modeling. *Journal of Molecular Liquids* 2020, *297*, 111740.

[166]. Sun, Y.; Li, S.; Quan, C., Solubility of ferulic acid and tetramethylpyrazine in supercritical carbon dioxide. *Journal of Chemical & Engineering Data* 2005, *50* (4), 1125-1128.

[167]. Han, S.; Wang, W.; Jiao, Z.; Wei, X., Solubility of Vitamin E acetate in supercritical carbon dioxide: Measurement and correlation. *Journal of Chemical & Engineering Data* 2017, *62* (11), 3854-3860.

[168]. Ciou, J.-M.; Wang, B.-C.; Su, C.-S.; Liu, J.-J.; Sheu, M.-T., Measurement of solid solubility of warfarin in supercritical carbon dioxide and recrystallization study using supercritical antisolvent process. *Advanced Powder Technology* 2018, *29* (3), 479-487.

Annex

Annex

Table.1. Sources and ranges of Molar solubility fraction data of solid drug-like compounds in supercritical carbon dioxide. Used for building the models.

No.	Nom	T(k)	P(Mpa)	AATS3v	MATS2e	GATS4c	GATS3v	GATS4e	GATS3s	nBondsM	AVP-0	SHBd	MLogP	MLFER_S	-Ln y_2	Ref.
1	4-aminoantipyrine	308.2-328.2	10.03-22.04	208.4654	0.1136	1.4180	0.9713	1.4709	0.7436	8	0.5844	0.5395	2.23	1.124	7.9294-12.2060	[80]
2	4-aminosalicylic acid	308-328	11.0-21	207.6990	0.1805	0.6322	1.2279	0.8292	0.9395	7	0.5101	2.0166	1.79	1.284	10.2348-12.5627	[81]
3	5-fluorouracil	313-328	8-20	184.3560	0.0791	0.9297	1.0804	1.5854	1.1499	3	0.4746	1.4681	1.35	0.495	12.1573-13.5531	[82]
4	10-undecenoic acid	308-333	10-18	114.8072	0.2571	0.6702	0.9895	0.7065	0.7811	2	0.6382	0.7291	2.45	0.548	4.0513-7.2644	[83]
5	All-trans retinoic acid	308.2-328.2	10.2-31.2	182.2340	0.2772	0.5449	1.3259	0.3840	0.6936	6	0.6564	0.7100	3.44	0.983	10.9026-15.4762	[84]
6	Amiodarone hydrochloride	313.2-342.2	12-30	201.2790	-0.0381	1.1632	0.7574	1.2130	0.8687	17	0.7308	0	3.55	2.451	6.8958-10.5926	[85]
7	Anastrozole	308-348	12.2-35.5	211.0515	0.1176	0.8892	1.3940	0.9990	0.7343	13	0.6059	0	2.78	2.518	7.8756-12.5345	[86]
8	Aprepitant	308.15-338.15	12-33	206.2470	0.2919	0.8433	1.1277	0.9294	0.5845	14	0.5178	1.2461	2.45	1.684	9.4756-12.3114	[87]
9	Aspirin	308.15-328.15	12-25	208.2891	0.2894	0.8260	0.9011	0.9965	1.0858	8	0.5370	0.7462	2.01	1.341	7.9662-9.6724	[88]
10	Atropine	308-348	12.2-35.5	165.4072	0.0454	0.7620	1.0509	0.9386	0.7612	7	0.5925	0.4974	2.89	1.698	6.3949-9.7211	[89]
11	Atrovastatin	308-348	12.16-35.46	191.5763	0.0813	0.8363	0.9608	0.7306	0.9913	27	0.5640	1.5960	4.21	3.845	6.5389-13.8155	[90]
12	Azathioprine	313-338	12-27	262.0937	0.0496	0.9303	1.1366	0.8745	0.5750	16	0.5425	0.5242	1.35	2.294	10.9086-12.8222	[91]
13	Azelaic acid	313.15-333.15	10-30	118.5903	0.2997	0.6753	1.0151	0.7861	0.8502	2	0.5892	1.4901	2.01	0.889	11.5009-14.6830	[92]
14	Benzocaine	308-348	12.2-35.5	177.3190	0.1807	1.3478	0.9647	1.6374	0.8585	7	0.5758	0.5452	2.12	1.418	4.4129-8.2171	[93]
15	Bisacodyl	308-348	12.2-35.5	215.9504	0.2447	0.7089	0.9187	0.6682	1.2092	20	0.5587	0	3.33	2.353	7.4473-11.6182	[94]
16	Budesonide	338-358	21.3-38.5	191.9097	0.0289	0.8834	1.0587	0.9264	0.7307	4	0.6023	0.4640	3.55	2.443	10.4447-12.0355	[95]
17	Caffeine	313-353	19.9-34.9	232.0621	0.1562	1.1040	1.2438	1.5655	0.9426	7	0.5844	0	1.68	0.996	6.7855-8.1701	[96]
18	Capecitabine	308-348	15.2-35.4	159.9700	0.1123	1.2248	0.9445	1.0837	0.8026	4	0.5605	1.6171	2.01	1.522	8.7479-12.8223	[97]
19	Carbamazepine	308-348	12.2-35.5	234.2513	0.4052	1.0654	0.8578	0.9715	0.8385	14	0.5392	0.5977	2.78	1.449	9.2722-12.7169	[89]
20	Cefixime trihydrate	308-328	18.3-33.5	229.7306	0.1078	0.9696	0.9675	0.8859	0.9275	12	0.5578	2.3614	1.68	4.226	15.0128-15.6481	[98]
21	Cefuroxime axetil	308-328	8-25	214.9035	0.1220	1.1170	0.9133	0.9549	1.0724	12	0.5616	1.1698	2.01	3.944	11.3960-15.3296	[99]
22	Celecoxib	323.3-343.2	15-30	228.6141	0.4423	0.7527	1.0920	0.8058	0.6159	21	0.5280	0.5598	2.34	2.439	8.7916-11.0942	[100]
23	Cetirizine	308.15-338.15	16-40	182.0835	-0.0552	0.6153	1.0493	0.4804	0.7265	13	0.5931	0.7278	3.11	2.218	5.3144-11.4641	[101]
24	Chloramphenicol	308-338	10-26	196.2615	0.0171	1.1274	0.9920	1.0764	0.8024	8	0.5717	1.3194	1.68	2.107	11.5229-12.4045	[102]

25	Chlormezanone	308.2-328.2	11.91-23.99	207.1075	0.1159	1.2947	1.0358	1.3552	1.1984	9	0.6011	0	2.01	2.799	7.5147-12.2060	[103]
26	Chlorpheniramine maleate	308-338	20-40	188.5535	-0.1532	0.7767	1.1291	0.7621	1.1446	12	0.6388	0	2.89	1.494	7.7611-11.0811	[104]
27	Chlorpropamide	313.15-353.15	10-30	181.1530	0.3666	1.1390	0.9322	1.2696	0.9861	9	0.5996	0.9782	1.79	2.033	9.5361-12.9869	[105]
28	Chrysin	313-333	20-40	233.3240	-0.0366	0.5951	0.9708	0.6932	1.0668	14	0.5173	1.1750	2.67	1.975	15.9959-17.8329	[106]
29	Cinildipine	313-333	12-36	214.7053	0.0962	1.0382	1.0221	1.0327	0.6622	18	0.5612	0.3670	3.44	2.796	10.1291-13.6332	[107]
30	Cinnarizine	308.15-328.15	13.97-24	182.4832	-0.2551	0.9535	1.0520	1.0177	0.9352	19	0.5829	0	4.1	2.112	8.4925-12.7508	[108]
31	Ciprofloxacin	313-333	17-36	199.1704	-0.0040	0.6660	0.7763	0.7512	0.6746	9	0.5453	0.9428	2.56	2.315	15.4831-17.4461	[109]
32	Climbazole	313.2-333.2	10.55-39.89	192.2391	0.0452	1.4104	1.3250	1.5748	1.1387	12	0.6231	0	2.56	1.846	5.3226-7.3858	[110]
33	Clobetasol propionate	308-348	12.2-35.5	205.2235	0.0599	1.0832	1.0154	1.2446	0.9404	5	0.6211	0.0631	3.44	2.316	12.5627-16.1181	[111]
34	Clofenamic acid	313-333	12-36	215.9826	0.1819	0.8208	1.0687	0.8842	0.6788	13	0.5651	1.0603	2.45	1.913	10.2477-13.3837	[112]
35	Clotrimazole	308-348	12.2-35.5	236.5251	-0.0031	0.9540	0.8858	1.0874	0.8180	23	0.5737	0	3.55	2.081	9.1464-15.4249	[113]
36	Clozapine	318-348	12.16-35.46	204.9515	-0.1001	1.0422	0.9839	0.9216	0.9495	13	0.6019	0.3060	2.89	2.681	10.0802-13.6331	[114]
37	Codeine	308-348	12.2-35.5	212.2692	-0.1339	0.9069	1.0213	0.9318	0.7427	7	0.5886	0.3164	3	2.156	7.1346-11.5129	[89]
38	Cortisone acetate	308.15-373.15	8.24-22.65	195.8706	0.0818	0.8991	1.1542	0.8232	0.8992	5	0.6018	0.0819	3.33	2.520	12.3948-16.0227	[115]
39	Cyproheptadine	308-338	16-40	212.7801	-0.2504	0.4639	0.9685	0.5417	0.8336	14	0.5931	0	3.66	1.631	5.7796-10.3039	[116]
40	Cyproterone acetate	308-348	12.2-35.5	209.6532	0.1300	1.0215	1.0128	0.9251	0.8609	5	0.6325	0	3.55	2.127	8.2509-11.2505	[117]
41	Dexamethasone	308.15-328.15	10-35.07	203.3863	-0.0081	0.9111	1.1082	1.1472	0.8336	4	0.5933	0.6436	3.22	2.182	12.7823-13.6416	[118.119]
42	Diazepam	308-348	12-35.07	238.9165	0.0302	1.2376	0.9297	1.1518	0.9749	14	0.5881	0	2.78	2.109	7.4490-10.1519	[89]
43	Diclofenac acid	308.15-338.15	12-40	224.1334	0.1507	1.0423	1.0117	0.8957	0.9816	13	0.5985	0.9706	2.45	2.100	6.2247-10.6627	[120]
44	Diflunisal	308.2-328.2	9.1-24.6	238.7639	0.0631	0.7155	1.0607	0.8642	0.9856	14	0.5012	1.3249	2.34	1.394	11.7271-14.4243	[121]
45	Docetaxel	308-348	15.2-35.4	205.6952	0.1210	1.2119	1.1162	1.1127	0.9009	18	0.5842	0.2923	4.54	3.951	7.2616-10.2045	[97]
46	Dutasteride	308-348	12.2-35.5	195.6823	0.4128	0.6627	1.1882	0.5938	0.4119	9	0.5567	0.6906	3.33	2.185	8.7359-16.1181	[122]
47	Enrofloxacin	313-333	12-36	196.8204	0.0006	0.6230	0.8051	0.6901	0.6612	9	0.5670	0.6212	2.78	2.502	12.0919-17.6322	[109]
48	Esomeprazole	308.2-338.2	12-27	232.2006	-0.0790	1.3285	1.0686	1.3089	1.1303	17	0.6042	0.4044	2.56	3.406	7.0020-11.4176	[123]
49	Ethosuximide	313.15-328.15	9.37-14.8	161.7181	0.26178	0.8615	1.3069	1.0784	1.0598	2	0.6230	0.5480	1.9	0.695	4.6122-5.6694	[124]
50	Exemestane	308-348	12.2-35.5	197.4053	0.0384	0.8557	1.1435	0.8794	0.7824	5	0.6146	0	3.44	1.586	6.2787-11.2818	[86]
51	Finasteride	308-348	12.2-35.5	175.3167	0.1983	0.8732	1.2519	0.7794	0.7831	3	0.6518	0.3783	3.55	1.942	5.7063-7.9983	[122]
52	Fluoxetine hydrochloride	308.15-338.15	16-40	182.7174	0.3698	0.9534	1.0135	1.1184	0.4801	12	0.5559	0.3369	2.78	1.236	7.1160-10.5384	[125]
53	Flurbiprofen	303-323	8.9-24.5	210.3439	0.1535	0.9121	1.0190	1.1564	0.8865	14	0.5516	0.6615	2.78	1.388	8.5332-10.9989	[126]

54	Flutamide	308-348	12.2-35.5	197.4672	0.4344	1.2309	1.1323	1.1207	0.6545	8	0.5392	0.5290	1.9	1.863	7.5824-12.2894	[122]
55	Fluvastatin	308-348	12.16-35.46	192.8681	0.0598	0.6283	0.9066	0.6090	1.0013	19	0.5664	1.4356	3.44	2.486	7.4169-12.2060	[90]
56	Gabapentin	308-338	16-40	141.7961	0.2009	0.8215	1.2575	0.5609	0.6744	1	0.6152	1.0533	2.12	0.829	4.9117-9.3190	[127]
57	Gambogic acid	308.15-328.15	10-30	214.8228	0.0998	1.0269	1.0774	1.1301	0.8507	14	0.6086	0.6876	4.76	2.935	10.6967-13.3269	[128]
58	Gatifloxacin	313-333	12-36	203.8424	-0.0339	0.8266	0.8102	0.7465	0.7000	9	0.5662	0.8703	2.67	2.313	13.3424-16.0598	[109]
59	Gemfibrozil	308.2-328.2	10.01-22.02	173.1881	0.1698	1.2968	1.3523	1.7007	0.8193	7	0.6453	0.6041	2.78	0.992	5.4750-12.7371	[129]
60	Ibuprofen	308-318	8-13	169.2564	0.2781	0.8134	1.1929	0.9692	0.7951	7	0.6351	0.6548	2.67	0.832	5.7257-11.3732	[82]
61	Imipramine hcl	313-323	30-50	188.3197	-0.1823	1.0557	1.0470	1.2625	1.0154	12	0.6213	0	3.33	1.254	11.5229-12.1863	[130]
62	Ipriflavone	308.2-328.2	10-20	209.4584	-0.0323	0.8750	0.8758	0.9364	1.0340	14	0.5713	0	3.11	1.921	8.6797-12.1291	[131]
63	Isoniazid	308-313	13-18.5	188.1176	-0.0365	1.2133	1.0729	1.7202	0.8700	7	0.5242	1.3234	1.68	1.447	12.0154-14.4504	[132]
64	Ketoconazole	308-348	12.2-35.5	196.7504	-0.0590	1.1118	0.9759	1.0464	1.1941	18	0.5969	0	3.22	3.481	8.6535-14.5087	[113]
65	Ketoprofen	308.15-338.15	16-40	208.2778	0.19322	1.0397	1.0351	1.4200	0.8396	14	0.5545	0.6399	2.89	1.768	7.2475-10.7199	[133]
66	Ketotifen fumarate (KTF)	308.2-338.2	12-30	218.9510	-0.0975	0.5679	0.9644	0.6782	1.0090	13	0.6172	0	3.22	1.902	6.8339-10.7663	[134]
67	Lacidipine	313-333	12-36	201.0299	0.2281	1.0778	0.9691	0.9463	0.9087	12	0.6183	0.3455	3.55	2.323	9.7689-13.2277	[135]
68	Lamotrigine	318-348	12.16-35.46	261.4388	0.1705	0.8410	0.7891	1.1977	0.7752	13	0.5935	1.2079	1.68	2.856	12.0406-15.0194	[114]
69	Lansoprazole	308.2-338.2	12-27	214.9262	0.2991	1.0793	0.9955	0.7079	0.8571	17	0.5436	0.4243	2.23	3.139	7.2143-11.3306	[136]
70	Letrozole	308-348	12.2-35.5	232.1377	0.0055	0.7349	1.1631	1.0866	0.5422	19	0.5266	0	2.78	2.455	9.3967-13.8155	[86]
71	Levonorgestrel	308-338	12.2-35.5	180.7864	0.0366	0.9999	1.0594	0.9834	0.7736	3	0.6166	0.5620	3.55	1.391	12.7508-14.7318	[137]
72	Loratadine	308.15-338.15	12-27	208.6064	0.2117	1.2010	0.9043	1.2514	0.6893	14	0.6021	0	3.33	1.680	6.6442-12.3114	[138]
73	Lovastatin	308-348	12.16-35.46	160.7995	0.1706	0.9191	1.0663	1.0480	0.8986	4	0.6266	0.3972	3.55	2.210	9.0793-11.0429	[90]
74	Mandelic acid	308.15-328.15	10.1-23.06	183.1658	0.1404	1.1050	1.1004	1.1027	0.6947	7	0.5242	1.3896	2.01	1.055	5.8417-10.5197	[139]
75	Medroxyprogesterone acetate	308-348	12.2-35.5	191.7921	0.1513	0.9654	1.1190	0.8588	0.8466	4	0.6417	0	3.66	1.783	4.7105-11.0429	[117]
76	Mefenamic acid	308.15-338.15	16-40	213.7418	0.2630	0.9038	1.1078	0.9024	0.6543	13	0.5776	1.0147	2.78	1.770	5.1260-9.3954	[140]
77	Megestrol acetate	308-338	12.2-35.5	199.9641	0.1541	0.9915	1.1091	0.8883	0.8852	5	0.6343	0	3.66	1.898	9.3485-12.7508	[137]
78	Meloxicam sodium salt	303-323	14.9-25.5	252.8421	0.2290	1.4026	1.1312	1.4267	0.9330	15	0.5583	0.4982	2.01	2.973	11.2692-12.3316	[141]
79	Metaxalone	308.2-328.2	11.95-24	172.6197	0.2069	0.8954	1.2285	0.9023	0.9941	7	0.5905	0.5441	2.34	1.292	8.8183-14.3604	[103]
80	Methimazole	308-348	12.2-35.5	183.7796	0.0817	0.6811	1.2079	1.1281	1.0643	2	0.6895	0.4563	1.57	1.609	8.5690-12.1291	[142]
81	Methocarbamol	308.2-328.2	11.99-22.05	162.5395	-0.0777	1.2056	1.0348	1.2422	1.2814	7	0.5739	1.1066	2.23	2.348	7.6227-10.4833	[103]

82	Methylparaben	308-348	12.2-35.5	206.4829	0.0594	1.1821	1.1733	1.6055	0.9390	7	0.5521	0.6907	2.01	1.282	6.7146-9.1338	[94]
83	Metronidazole benzoate	308-348	12.2-35.5	196.6955	0.1749	1.2717	0.9401	1.2640	0.6216	13	0.5352	0	2.12	1.870	5.3927-9.5670	[93]
84	Nabumetone	308.2-328.2	10-22	200.5738	-0.0591	0.7723	1.0898	0.7202	1.4450	12	0.5996	0	2.89	1.527	5.9219-10.1442	[143]
85	Naproxen	308-348	12.2-35.5	216.5803	0.1069	0.8528	1.0591	1.0120	0.8845	12	0.5767	0.6465	2.67	1.398	8.5172-11.5129	[93]
86	Nicotinic acid	313.15-373.15	4.5-30.2	195.7393	0.1870	0.8557	1.1989	1.3406	0.8178	7	0.5124	0.8023	1.79	1.107	11.4727-14.9867	[144]
87	Nifedipine	313-333	12-36	244.3145	0.1565	0.9267	1.0163	1.0799	0.6539	11	0.5654	0.4034	2.45	2.085	9.5599-12.9826	[135]
88	Niflumic acid	313-333	12-36	222.9240	0.4391	0.7966	1.1310	1.0907	0.6016	13	0.4989	1.2355	2.12	2.075	10.7729-13.3269	[112]
89	Nimodipine	313-333	10-25	208.5713	0.1133	1.0184	0.9741	0.9504	0.7040	11	0.5844	0.3696	2.78	2.231	10.0707-14.3263	[145]
90	Nitrendipine	308-328	8-20	227.3893	0.1785	1.0325	0.9440	1.0861	0.6581	11	0.5708	0.3883	2.56	2.085	9.6690-13.8155	[146]
91	Norfloxacin	313.15-323.15	9.94-30.33	197.5117	-0.0077	0.6611	0.8374	0.7473	0.6801	9	0.5566	0.9682	2.45	2.204	10.6209-13.6167	[147]
92	Ofloxacin	318.15-328.15	10.11-30.33	205.1238	-0.0528	0.7616	0.8510	0.7801	0.6670	9	0.5484	0.9501	2.45	2.352	13.5228-14.7571	[147]
93	O-nitrobenzoic acid	308-328	10-21	222.9693	0.2922	1.0443	1.0917	1.2823	0.7149	8	0.4816	0.7762	1.68	1.124	10.2429-13.3968	[148]
94	Oxcarbazepine	313-333	12-27	235.6584	0.2519	0.9417	0.8755	0.9880	0.9602	14	0.5351	0.6060	2.67	1.820	10.5289-16.0228	[149]
95	Oxymetholone	308-328	12.1-30.5	182.5790	0.0218	0.7743	1.2362	0.7991	0.7801	2	0.6394	0.6042	3.44	1.525	15.7193-17.9507	[98]
96	Paclitaxel	310.2-329	14.09-34.69	212.3053	0.0987	1.2584	1.0523	1.1885	0.9297	25	0.5682	0.1712	4.98	5.288	11.9718-13.0928	[82]
97	Paeonol	313.2-333.2	8.6-14.3	215.1309	-0.0969	0.7256	1.0566	0.7984	1.3742	7	0.5829	0.6174	2.12	1.101	4.3275-5.5517	[150]
98	Penicillin G	313.15-333.15	10-35	199.3197	0.2017	1.1924	1.2814	1.0114	0.9273	9	0.5943	0.9122	2.45	2.761	9.6676-12.3804	[151]
99	Penicillin V	314.85-334.85	7.99-28.05	198.3949	0.1103	1.2455	1.2529	1.0672	0.9342	9	0.5865	0.9655	2.34	2.844	7.4594-9.8173	[152]
100	Pentoxifylline	308.15-328.15	11.98-24	186.6443	0.1796	1.1029	1.0849	1.2261	1.0568	8	0.5959	0	2.12	1.467	6.5929-10.4143	[108]
101	Phenazopyridine	308-348	12.2-35.5	191.5390	0.2091	0.9277	1.0141	1.3192	0.9890	13	0.5336	1.1849	2.12	2.298	8.5067-12.3339	[143]
102	Phenylephrine hydrochloride	308.15-338.15	16-40	162.8124	-0.1167	1.0426	1.1078	1.0227	1.0123	6	0.5823	1.5194	2.12	1.230	5.8464-9.2004	[153]
103	Pindolol	298-318	8-27.5	156.2058	-0.1023	1.0287	1.0030	1.0614	1.1802	10	0.6006	1.1534	2.56	1.474	8.4037-10.5197	[154]
104	Piracetam	308.15-328.15	12-24	137.1858	0.1394	1.1730	1.1344	1.2270	1.1751	2	0.5669	0.6283	1.68	1.690	10.1965-11.8006	[108]
105	Piroxicam	308.15-338.15	16-40	239.6693	0.2040	1.3062	0.9891	1.4006	0.9136	16	0.5346	0.9707	2.23	3.344	7.5772-11.3559	[155]
106	Prednisolone	308.15-373.15	8.24-22.65	195.3180	-0.0136	0.7970	1.1443	0.8253	0.7194	4	0.5939	0.7067	3.22	2.228	14.1329-20.7233	[115]
107	Prednisone	308.15-373.15	8.24-22.65	201.6644	0.0017	0.8339	1.1362	0.8480	0.7567	5	0.5894	0.6249	3.22	2.416	13.9286-20.7233	[115]
108	Propranolol	308-348	12.2-35.46	165.2954	-0.0852	1.1230	0.9722	1.1398	1.1646	11	0.6034	0.7016	2.89	1.353	6.0339-10.2375	[142]
109	Protocatechuic acid	313-333	20-40	210.4681	0.1170	1.2791	1.2192	1.4321	0.8480	7	0.4983	2.1877	1.79	1.630	14.1398-15.6994	[106]

110	Pyridin-4-amine	308.2-328.2	10-22	160.3554	-0.1414	0.6104	1.2343	0.5937	1.2442	6	0.5477	0.5606	1.79	1.184	8.4828-10.7757	[156]
111	Pyrocatechol	308.15-338.15	12-40	173.2791	-0.0210	1.5527	1.1495	1.6125	0.8387	6	0.5254	1.4441	1.9	1.273	5.5367-7.3233	[157]
112	Quinine	323.15-343.15	12.5-27.5	187.7316	-0.1334	1.0275	1.0420	0.9547	0.7903	12	0.5857	0.1959	3.22	1.916	9.8955-11.3306	[158]
113	Repaglinide	308-338	12-27	171.4509	0.1245	0.7644	0.9959	0.7364	0.7723	14	0.6052	0.8312	3.77	2.676	9.2594-12.7508	[159]
114	Rosuvastatin	308-348	12.16-35.46	190.1097	0.1763	0.8297	1.0194	0.8149	1.1091	17	0.5760	1.4526	2.67	3.247	8.3183-12.7169	[90]
115	Salicylamide	308.2-328.2	10-22	196.9071	0.1322	0.8218	1.0996	1.0555	1.0942	7	0.5242	1.3200	1.9	1.403	8.4684-11.6813	[143]
116	Salicylic acid	313-333	10-35	202.4417	0.1948	0.8128	1.1015	1.0632	0.8830	7	0.5112	1.4675	1.9	0.888	7.3248-9.3667	[160]
117	Sertraline hydrochloride	308-338	12-30	209.0308	-0.0491	0.9301	0.9282	0.7247	0.8609	12	0.6439	0.2174	3	1.515	9.2829-12.1476	[161]
118	Simvastatin	308-348	12.16-35.46	164.5879	0.1765	0.9780	1.1563	1.1741	0.8803	4	0.6365	0.3912	3.66	2.17	7.5332-13.1223	[90]
119	Sinapic acid	313-333	20-20	224.6376	-0.0717	0.8698	1.0916	0.8907	0.8611	8	0.5580	1.3686	2.12	1.816	14.1664-18.7773	[106]
120	Sorafenib tosylate	308-338	12-27	224.3504	0.3286	1.0790	0.9818	1.0898	0.7270	20	0.5379	1.6217	2.56	3.495	11.2842-14.2011	[162]
121	Spironolactone	308-338	16-40	192.5641	0.1583	0.7561	1.1411	0.7154	0.7552	4	0.6416	0	3.55	2.707	5.2963-9.6803	[163]
122	Sulindac	308.15-338.15	16-40	237.5520	0.1160	0.6604	1.0190	0.6298	1.0402	16	0.5733	0.6175	3.11	3.076	4.7456-10.2181	[164]
123	Sunitinib malate	308-338	12-27	194.3161	0.0706	1.0411	0.9237	0.8549	0.9051	14	0.5958	1.1794	3.11	3.020	9.3659-12.2061	[165]
124	Tetramethylpyrazine	318-338	10-30	198.2682	0.2432	1.6602	1.0242	1.9067	1.9052	6	0.6894	0	2.12	0.725	2.0326-6.1193	[166]
125	Theobromine	313-353	19.3-34.5	225.0545	0.1779	0.9163	1.2139	1.1020	1.0652	7	0.5565	0.6470	1.57	1.013	9.9632-11.6407	[96]
126	Theophylline	313-353	19.9-34.9	226.2795	0.1779	1.1128	1.1859	1.7682	0.9489	7	0.5565	0.5649	1.57	1.027	10.3129-11.4837	[96]
127	Thymidine	308-328	10-30	146.6929	0.3309	0.8811	1.1562	1.3720	1.0501	2	0.5667	1.1470	1.57	0.347	11.7361-13.6332	[82]
128	Tolbutamide	313.15-353.15	10-30	167.1545	0.4476	1.0151	1.0224	1.0788	0.9488	9	0.5981	0.9488	2.12	1.876	7.8116-11.0061	[104]
129	Tolfenamic acid	313-333	12-36	225.3529	0.2001	0.9252	1.0254	1.0827	0.6526	13	0.5850	1.0134	2.56	1.889	10.9184-13.3393	[112]
130	Triclocarban	313.2-333.2	10.93-38.96	230.8210	0.1677	1.1382	1.0555	1.0755	1.0864	13	0.6237	0.9365	2.23	2.187	7.0470-9.3157	[110]
131	Vitamin e acetate	308.15-328.15	8.41-14.19	159.6518	0.2274	1.1469	1.1360	1.1251	0.9920	7	0.6980	0	4.54	1.003	7.2279-8.8817	[167]
132	Warfarin	308.2-328.2	10-18	218.4587	0.1384	1.0262	0.9961	1.0387	1.0450	15	0.5501	0.3817	3.11	2.927	12.1309-15.7261	[168]
133	Zopiclone	313-333	10-25	205.2551	0.0835	1.0419	1.0251	1.1723	0.7797	14	0.5679	0	2.23	2.794	10.7290-13.4100	[145]