

T.C.
YEDİTEPE UNIVERSITY
INSTITUTE OF HEALTH SCIENCES
DEPARTMENT OF PHARMACEUTICAL TOXICOLOGY

**EVALUATION OF COMEDOGENICITY AND
IRRITANCY POTENTIAL OF SOME COSMETIC
INGREDIENTS USING *IN SILICO* METHODS**

DOCTOR OF PHILOSOPHY THESIS

SEBLA ÖZTAN AKTÜRK, MSc

SUPERVISOR
Prof.Dr.Hande SİPAHİ

İSTANBUL-2023

THESIS APPROVAL FORM

Institute : Yeditepe University Institute of Health Sciences
Programme : Pharmaceutical Toxicology
Title of the Thesis : Evaluation of Comedogenicity and Irritancy Potential of Some Cosmetic Ingredients Using *In Silico* Methods
Owner of the Thesis : Sebla Öztan Aktürk
Examination Date : 15.02.2023

This study has been approved as a Doctorate Thesis in regard to content and quality by the Jury.

	Title, Name-Surname (Institution)
Chair of the Jury:	Prof.Dr.Ahmet Aydın, Yeditepe University
Supervisor:	Prof.Dr.Hande Sipahi, Yeditepe University
Member	Prof.Dr.Melek Türker Saçan, Boğaziçi University
Member	Prof.Dr.Onur Erdem, Sağlık Bilimleri University
Member	Dr.Ebru Türköz Acar, Yeditepe University

APPROVAL

This thesis has been deemed by the jury in accordance with the relevant articles of Yeditepe University Graduate Education and Examinations Regulation and has been approved by Administrative Board of Institute with decision dated and numbered

Prof.Dr.Bayram YILMAZ
Director of Institute of Health Sciences

BEYAN

Bu tezin kendi çalışmam olduğunu, planlanmasından yazımına kadar hiçbir aşamasında etik dışı davranışımın olmadığını, tezdeki bütün bilgileri akademik ve etik kurallar içinde elde ettiğimi, tez çalışmasıyla elde edilmeyen bütün bilgi ve yorumlara kaynak gösterdiğimi ve bu kaynakları kaynaklar listesine aldığımı, tez çalışması ve yazımı sırasında patent ve telif haklarını ihlal edici bir davranışımın olmadığını beyan ederim.

15.02.2023

Sebla ÖZTAN AKTÜRK

DECLARATION

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it contains no material previously published or written by another person nor material which has been accepted for the award of any other degree except where due acknowledgment has been made in the text.

15.02.2023

Sebla ÖZTAN AKTÜRK



DEDICATION

This thesis is dedicated to my precious family, especially my lovely sons,
Emir Kağan and Atahan
&
Likewise, to my mother “Belkıs Öztan” and my mother-in-law “Ayten Aktürk” who
never withheld their support and help from me

ACKNOWLEDGEMENTS

Firstly, I would like to express my sincere appreciation to my advisor Prof. Dr. Hande SİPAHİ as the mother of the ideas for this Ph.D. thesis study, for her invaluable support, motivation, patience, and immense knowledge. Her vision and open-minded guidance helped me in all the time of research and writing of this thesis. Also, I am so grateful to her for believing in me and supporting me during my doctoral education.

Besides my advisor, I would like to send my deepest thanks to Dr. Gulçin TUĞÇU for her limitless guidance on methodology and investigation, continuous support, forward-thinking comments, patience, encouragement, and insight throughout the duration of my thesis.

I am thankful to our Vice-Rector, Head of Pharmaceutical Toxicology Department, Prof. Dr. Ahmet AYDIN for the core knowledge about toxicology that I earned. In all lessons during my toxicology education, I have been there with my all ears and listened to him with complete attention. His core knowledge about toxicology I obtained from him during my doctoral education has lighted my way and been a guide to me in all evaluations in this thesis. Also, I would like to express my deepest appreciation to give me a chance to join his student laboratory team for my experience.

I experienced the pride and honor of belonging as a member of the Yeditepe family for 16 years. Initially, I would like to express my sincere thanks to the Dean of the Faculty of Pharmacy, Prof. Dr. Meriç KÖKSAL AKKOÇ for her precious support during my doctoral education by giving me access to all research possibilities and laboratories. In addition, I am thankful to the staff of the Institute of Health Science for their helpfulness and understanding during my pregnancy in my thesis period.

My special thanks go to the former Dean of the Yeditepe Faculty of Pharmacy, Prof. Dr. Dilek Demir EROL, who provided me an opportunity to join their laboratory team as an assistant by opening the door of my academic journey in 2007.

My biggest thank goes to my dear family; to my dear son, Emir Kağan and Atahan, and also, to my lovely husband Serkan for their understanding and support all the time. The important part of my success belongs to my mother and my mother-in-law. Without their invaluable support, it would not be possible to achieve this thesis study.

TABLE OF CONTENTS

THESIS APPROVAL FORM.....	i
BEYAN.....	ii
DECLARATION	iii
DEDICATION	iv
ACKNOWLEDGEMENTS	v
TABLE OF CONTENTS.....	vi
LIST OF TABLES	ix
LIST OF FIGURES	xi
LIST OF SYMBOLS AND ABBREVIATIONS.....	xiv
ABSTRACT.....	xviii
ÖZET	xix
1. INTRODUCTION AND AIM	1
2. GENERAL INFORMATION	4
2.1 <i>In Silico</i> Toxicology and <i>In Silico</i> Methods.....	4
2.1.1 QSAR and <i>In Silico</i> Modelling	6
2.1.2 History of QSAR.....	7
2.1.3 Toxicological Endpoint.....	8
2.1.4 Molecular Descriptors	9
2.1.4.1 Types and Dimensions of Descriptors	9
2.1.4.2 Software for Calculation of Descriptors.....	12
2.1.5 QSAR Modelling Approaches	13
2.1.6 QSAR Modelling Flowchart	13
2.1.7 Factors Affecting the Predictive Ability of QSAR Models	14
2.1.8 Advantages of QSAR Modelling	16
2.1.9 Application Area of QSAR	17
2.1.10 Adverse Outcome Pathway	18
2.2 Machine Learning	19
2.2.1 History of Machine Learning	21
2.2.2 Classification of Machine Learning	22
2.2.3 Feature Selection.....	28
2.2.4 Machine Learning Models	29
2.2.5 Machine Learning Algorithm.....	32
2.2.6 Evaluation Metrics of Classification Models in ML	36
2.2.7 Application Areas of Machine Learning.....	40
2.3 Comedogenicity	42

2.3.1 Comedogenicity and <i>Acne cosmetica</i>	42
2.3.2 Mechanism of Comedo Formation.....	43
2.3.3 Pathogenesis of Acne	43
2.3.4 Acne and Oxidative Stress	45
2.3.5 Lipid Peroxidation and Acne	49
2.3.6 Chloracne and Hyperkeratinization.....	57
2.4 Skin Irritation	61
2.4.1 Definition of Skin Irritation	61
2.4.2 Mechanism of Skin Irritation	61
3. MATERIAL AND METHODS	63
3.1 Materials.....	63
3.1.1 Dataset.....	63
3.2 Methods.....	81
3.2.1 Data Collection.....	82
3.2.2 Data Curation	83
3.2.3 Calculation of Molecular Descriptors	85
3.2.4 Training/Test Set Division	87
3.2.5 The Selection of Descriptors for the QSAR Classification Model	88
3.2.6 Evaluation Metrics	91
3.2.7 Classifier	92
3.2.8 Model Development.....	93
3.2.9 Model Validation	94
3.2.10 Performance Evaluation of the Classifiers and QSAR Models.....	95
3.2.12 Applicability Domain.....	96
3.2.13 Evaluation of the Relationship Between Comedogenicity and Skin Irritation.....	97
4. RESULTS	99
4.1 Comedogenicity Studies	99
4.1.1 The Training and Test Sets for Comedogenicity.....	99
4.1.2 QSAR Models for Comedogenicity	105
4.1.3 Comparison of Classifiers	113
4.1.4 Comparison of the Classification Models with RF	114
4.1.5 Optimization of the Model for Comedogenicity	115
4.1.6 Predictive Ability of Optimized QSAR Model for Comedogenicity	116
4.1.7 Applicability Domain for Comedogenicity Model	117
4.2 Skin Irritation	120
4.2.1 The Training and Test Sets for Skin Irritation	120
4.2.2 QSAR Models for Skin Irritation	122

4.2.3 Optimization of the Skin Irritation Models	126
4.2.4 Predictive Ability of Optimized QSAR Model for Skin Irritation	127
4.3 The Relationship Between Comedogenicity and Skin Irritation.....	128
4.3.1 Distribution of Dataset	128
4.3.2 The determination of relationship via Chi-square test	128
5. DISCUSSION	131
5.1 Comedogenicity	131
5.1.1 The Division of the Dataset into Training and Test Set	131
5.1.2 The Descriptors in the Comedogenicity Models.....	134
5.1.2 The Selected Descriptors for Comedogenicity Models via Learning-based.....	135
5.1.3 Classifier Performance	141
5.1.4 Model Performance	142
5.1.5 The Predictive Ability of Best QSAR Model for Comedogenicity.....	146
5.1.6 Applicability Domain for Comedogenicity	147
5.2 Skin Irritation	149
5.2.1 Evaluation of Training and Test Set for Skin Irritation.....	149
5.2.2 The Descriptors in the Skin Irritation Models.....	149
5.2.3 The Performance of Skin Irritation Models	152
5.2.4 Skin Irritation Mechanism of Models	156
5.2.5 The Predictive Ability of the Best QSAR Model for Skin Irritation.....	161
5.3 The Relationship Between Comedogenicity and Skin Irritation.....	162
5.3.1 Evaluation of the Dataset	162
5.3.2 Evaluation of the Relationship Between Comedogenicity and Irritation	162
6. CONCLUSION	167
6.1 Comedogenicity	167
6.2 Skin Irritation	169
6.3 The Relationship Between Comedogenicity and Skin Irritation.....	170
7. REFERENCES.....	172
8. APPENDICES	190
9. CURRICULUM VITAE	231

LIST OF TABLES

Table 1. Description of five dimensions of molecular descriptors.....	10
Table 2. Descriptors used in QSAR/QSPR modelling	11
Table 3. Software for calculating molecular descriptors and fingerprints	12
Table 4. 2X2 table of a confusion matrix for binary classification.....	36
Table 5. The scores for kappa	39
Table 6. Free radicals as contributors to oxidative stress	47
Table 7. Composition of skin surface lipids.....	55
Table 8. Adverse effects of squalene peroxidation	56
Table 9. Dioxin TEFs according to World Health Organization (WHO)	59
Table 10. The chemical names, structures, SMILES strings, CAS numbers of the datasets	63
Table 11. Descriptors calculated for the dataset compounds.....	85
Table 12. The comparison of setting parameters of two feature selection methods	88
Table 13. The Xmax and Xmin value of each descriptor (Xi) in the training set.....	96
Table 14. Frequency table as an input for Chi-Square test.	97
Table 15. Interpretation of Cramer's V value	98
Table 16. Intra-cluster variance for three optimal cluster	102
Table 17. Inter-cluster distance for three optimal cluster.....	102
Table 18. The distribution of comedogenicity dataset into training and test set.....	103
Table 19. The selected attributes by WEKA via the correlation-based method.	105
Table 20. The classification models obtained via the correlation-based method.....	106
Table 21. The selected attributes by WEKA via the learning-based method.....	107
Table 22. The classification models obtained via the learning-based method.	108
Table 23. The results of the comparison of two feature (descriptor) selection method	109
Table 24. The names and the description of the selected descriptors for this study.	110
Table 25. The amount of information gain (IGA) of descriptors selected by WEKA	111
Table 26. Correlation information between feature variable and target variable.....	111
Table 27. The performance of the top three classifiers for each model	113
Table 28. The comparison of the classification models with RF for comedogenicity	114
Table 29. Alva model with three descriptors after removing the two irrelevant features.	115
Table 30. The comparison of initial models with optimized models	115
Table 31. The outliers and out of AD for Alva Model optimized with 3 descriptors	117
Table 32. The distribution of skin irritation dataset into training and test set.....	120
Table 33. The selected attributes by WEKA via the learning-based method.....	122
Table 34. The description of the descriptors selected by WEKA for skin irritation	123
Table 35. The amount of information gain (IGA) of descriptors selected by WEKA	124

Table 36. The classification models obtained via the learning-based method for irritation	125
Table 37. The comparison of initial models with optimized models for skin irritation	126
Table 38. The results of Chi-Square test to find the relationship in terms of comedogenicity.	129
Table 39. The results of the Chi-Square test to find the relationship in terms of irritation.....	129
Table 40. The results for the strength of association between comedogenicity and irritation	130



LIST OF FIGURES

Figure 1. The computational tools in silico toxicology, steps and methods	5
Figure 2. QSAR model	6
Figure 3. Short History of QSAR evolution.....	7
Figure 4. Molecular descriptors used in QSAR modelling	9
Figure 5. The representation of five dimensions of molecular descriptors	10
Figure 6. Key steps of QSAR modelling	13
Figure 7. QSAR protocol (Erselvam, 2016).....	14
Figure 8. The descriptor selection process for predictive QSAR model creation	15
Figure 9. Advantages of QSAR model development	16
Figure 10. Application area of QSAR	17
Figure 11. Successful applications of QSAR in drug discovery.	17
Figure 12. Schematic AOP diagram.....	18
Figure 13. AOP diagram of chemically-induced skin allergy showing the different pathways..	18
Figure 14. Comparison between artificial intelligence (AI), machine learning (ML).	19
Figure 15. ML model needs human intervention for manual validation.....	20
Figure 16. Data science, a combination of techniques and theories taken from many fields	20
Figure 17. A short history of AI family.	21
Figure 18. Milestones of ML in its history	21
Figure 19. Machine learning and its distribution	22
Figure 20. Learning diagram for reinforcement learning.....	26
Figure 21. Flowchart of parallel and sequential ensemble learning	27
Figure 22. Flowchart of stacking ensemble learning	27
Figure 23. The comparison of classic ML and DL	28
Figure 24. Flow diagram of feature selection methods).....	28
Figure 25. Outline of machine learning	29
Figure 26. . Linearly separable or not separable data.....	30
Figure 27. Linear classification and regression.....	30
Figure 28. Linear regression versus logistic regression	30
Figure 29. Linear and non-linear ML algorithms	32
Figure 30. Structure of a decision tree	33
Figure 31. An example of the structure of a random forest	35
Figure 32. The area under the ROC curve	38
Figure 33. ROC curve	39
Figure 34. Precision-Recall curve	39
Figure 35. The popular applications of machine learning.....	40

Figure 36. <i>Acne cosmetica</i>	42
Figure 37. Closed comedones	42
Figure 38. initial lesion of acne (microcomedo) and progression of a lesion in acne	43
Figure 39. Pathogenesis of acne	44
Figure 40. The target organ in acne is the pilosebaceous unit composed of a hair follicle.....	45
Figure 41. Various cutaneous disorders are related with oxidative stress	45
Figure 42. Sources of Free radicals (ROS and RNS) and against antioxidant defense system...	48
Figure 43. The pathological role of free-radicals induced-oxidative stress for various diseases	48
Figure 44. The model of progression of acne vulgaris.....	49
Figure 45. Three stages of lipid peroxidation	50
Figure 46. Mechanism of lipid peroxidation.....	50
Figure 47. By product of lipid peroxidation and their results	52
Figure 48. Xenobiotics-induced lipid peroxidation and detoxification by the antioxidant	52
Figure 49. The structure of tert butyl hydroperoxide	53
Figure 50. The structure of carbon tetrachloride.....	53
Figure 51. The structure of quinolinic acid.....	53
Figure 52. The possible cellular response to ROS in lipid membrane	54
Figure 53. The structure of various lipids found in sebum and stratum corneum.....	55
Figure 54. Ex-President Viktor Yushchenko of Ukraine before and after dioxin poisoning)....	57
Figure 55. The pathogenesis of chloracne and hyperpigmentation by dioxins	58
Figure 56. The chemical structure of selected dioxins and dioxin-like compounds	59
Figure 57. Hyperkeratinization induced by dioxins in the orifices of hair follicle	60
Figure 58. The mechanism of skin irritation that induced by surfactants	62
Figure 59. The flowchart of QSAR modeling for comedogenicity prediction	81
Figure 60. General dataset curation workflow	83
Figure 61. The workflow of the correlation-based method.....	89
Figure 62. The workflow of learning-based method.....	90
Figure 63. The QSAR model development process for comedogenicity prediction.....	93
Figure 64. The diagram of k-fold cross-validation with k=10 for ML model).	94
Figure 65. The graph of the optimal number of clusters (k=6) is based on functional group.....	99
Figure 66. The graph of the optimal number of clusters (k=5) is based on all the descriptors .	100
Figure 67. The graph of number of clusters (k=5 and k=7) based on all fingerprints.....	101
Figure 68. The stacked bar chart of ten chemical class distribution in the dataset	104
Figure 69. The stacked bar chart of ten chemical class distribution in the training set.....	104
Figure 70. The stacked bar chart of ten chemical class distribution in the test set	104
Figure 71. Decision Tree View for selected 5 descriptors for Alva Model	112
Figure 72. Decision Tree View for 3 descriptors for optimum Alva Model	112

Figure 73. The stacked bar chart of true/false prediction of QSAR Model AlvaDesc	116
Figure 74. The q-q plot of Alva_S3Knormalized training data (n=92).....	118
Figure 75. The boxplot graph of Alva_S3Knormalized. training data (n=92).....	118
Figure 76. The q-q plot of Alva_ATSC7i normalized training data (n=92)	118
Figure 77. The boxplot graph of Alva_ATSC7inormalized training data (n=92).....	118
Figure 78 The q-q plot of Alva_CATS2D_08_LLnormalized training data (n=92).....	118
Figure 79 The boxplot graph of Alva_CATS2D_08_LLnormalized training data (n=92).....	118
Figure 80. The comparison of the boxplot graph of training and test set to determine the outside of AD for Alva_S3Knormalized descriptor in the Alva model.	119
Figure 81. The comparison of the boxplot graph of training and test set to determine the outside of AD for Alva_ATSC7inormalized descriptor in the Alva model.	119
Figure 82. The comparison of the boxplot graph of training and test set to determine the outside of AD for Alva_CATS2D_08_LL normalized descriptor in the Alva model.....	119
Figure 83. The dataset for skin irritation (n=121).....	121
Figure 84. The training set for skin irritation (n=92)	121
Figure 85. The test set for skin irritation (n=29).....	121
Figure 86. The stacked bar chart of true/false prediction of QSAR ModelMoldfor irritation...127	127
Figure 87. Venn diagram of 121 cosmetic ingredients in the dataset	128
Figure 88. The distribution of 121 cosmetic ingredients in terms of irrit. and comed.....129	129
Figure 89. Adverse outcome pathway for comedogenicity prediction via our QSAR model... 137	137
Figure 90. Cellular damage induced by oxidative stress.....	159
Figure 91. The possible cellular response to ROS in lipid membrane	159
Figure 92. 4-HNE advocate cell survival or induces cell death	160
Figure 93. SC lipid matrix in the skin	163
Figure 94. Alteration of SC lipid content due to exogenous irritants, 1% of SLS	164

LIST OF SYMBOLS AND ABBREVIATIONS

ACC	: Accuracy
AD	: Applicability Domain
AHR	: Aryl Hydrocarbon Receptor
AI	: Artificial Intelligence
ANN	: Artificial Neural Network
ANOVA	: Analysis of Variance Test
AO	: Adverse Outcome
AOPs	: Adverse Outcome Pathways
AUC	: Area Under the ROC Curve
AU-PRC	: AreaUnder the Precision-Recall Curve
C	: Carbon
CART	: Classification and Regression Tree
CAS	: Chemical Abstract Services
CAT	: Catalase
CATS	: Chemical Advanced Template Search
Corr.	: Correlation
CV	: Cross-Validation
CYP1A1	: Cytochrome P450, Family 1, Subfamily A, Polypeptide 1
D	: Dimensional
DL	: Deep Learning
DT	: Decision Trees
EA	: Electron Affinity
ER	: Error Rate
F score	: Harmonic mean of precision and recall
FN	: False Negative
FP	: False Positive
FPR	: False Positive Rate
FR	: Free Radicals
GR	: Glutathione Reductase
GSH	: Glutathione
GSH-Px	: Glutathione Peroxidase

4-HNE	: 4-Hydroxynonenal
H	: Entropy
H ₂ O ₂	: Hydrogen Peroxide
HL	: Hidden Layer
ID	: Identity
IG	: Information Gain
IGA	: Information Gain Amount
IL	: Input Layer
IL-1	: Interleukin-1
IoT	: Internet of Things
IP	: Ionization Potential
IQR	: Interquartile Range
KE	: Key Events
kNN	: k-Nearest Neighbours
L•	: Lipid Radical
LH	: Unsaturated Lipid
Log P	: Logarithm of the Octanol-Water Partition Coefficient
LogΦΔ	: Logarithm of Singlet Oxygen Generation Quantum Yields
LOO•	: Lipid Peroxyl Radical
LOOH	: Lipid Peroxide
LPO	: Lipid Peroxidation
LR	: Logistic Regression
M	: Model
MCC	: Matthews' Correlation Coefficient
MDA	: Malondialdehyde
MIE	: Molecular Initiating Event
MITF	: Microphthalmia-associated transcription factor
ML	: Machine Learning
NB	: Naïve Bayes
NNge	: Non-Nested generalisation exemplars, ML algorithm
NPV	: Negative Predictive Value
NRF2	: Nuclear factor erythroid 2-related factor

O ₂ •	: Superoxide Radicals
¹ O ₂	: Singlet Oxygen
•OH	: Hydroxyl Radicals
OECD	: Organization for Economic Co-operation and Development
OL	: Output Layer
p	: probabilities
PCA	: Principal components analysis
PCBs	: Biphenyls
PCDDs	: Polychlorinated dibenzo-p-dioxins
PCDFs	: Dibenzofurans (PCDFs)
PGE ₂	: Prostaglandin E ₂
PLA ₂	: Phospholipase A ₂
PLS	: Partial Least Squares
PPARs	: Peroxisome Proliferators-Activated Receptors
PPP	: Potential Pharmacophore Points
PPV	: Positive Predictive Value
PUFAs	: Polyunsaturated Fatty Acids
QSAR	: Quantitative Structure-Activity Relationships
QSPR	: Quantitative Structure-Property Relationship
RF	: Random Forest
RFE	: Recursive feature elimination
RL	: Reinforcement Learning
RNS	: Reactive Nitrogen Species
ROS	: Reactive Oxygen Species
SAR	: Structure-Activity Relationships
SARSA	: State-Action-Reward-State-Action
SDF	: Structure Data File
SDS	: Sodium Dodecyl Sulfate
SEN	: Sensitivity
SLS	: Sodium Lauryl Sulfate
SMILES	: Simplified Molecular-Input Line Entry System
SOD	: Superoxide Dismutase

SPC	: Specificity
SSE	: Sum Squared Error
SVM	: Support Vector Machine
TCDD	: 2,3,7,8-tetrachlorodibenzo-p-dioxin
TN	: True Negative
TNF- α	: Tumor necrosis factor alpha
TNR	: True Negative Rate
TP	: True Positive
TYR	: Tyrosinase
TYRP1	: Tyrosinase-related Protein 1
WEKA	: Waikato Environment for Knowledge Analysis

ABSTRACT

Oztan Akturk, S. (2022). Evaluation of Comedogenicity and Irritancy Potential of Some Cosmetic Ingredients Using *In Silico* Methods. Yeditepe University, Institute of Health Science, Department of Pharmaceutical Toxicology, PhD thesis, İstanbul.

Comedogenicity and skin irritation are the common adverse reactions to cosmetic ingredients. While comedogenicity is the potential of a ingredient that induce formation of comedones which are the initial lesion for acne. Skin irritation is the formation of reversible damage to the skin after the application of a chemical for up to 4 hours. Before cosmetic testing on animals was prohibited in 2013, comedogenicity was tested on rabbits ears. However, there has been no full replacement of animal testing on cosmetics. For this reason, in this thesis study, we proposed to develop two different QSAR classification model to predict the comedogenic and skin irritaticy potential of cosmetic ingredients by using different machine learning algorithms and types of molecular descriptors.

Our dataset was obtained from the literature, composed of 121 cosmetic ingredients mainly fatty acids, fatty alcohols and their derivatives and pigments, tested on rabbits for both comedogenicity and irritation. The total of 4837 molecular descriptors were calculated by the means of four different software. Descriptors for models were selected via WEKA by comparing correlation-based and learning-based (via decision tree algorithm) method. Modelling studies were performed with WEKA using various classifiers. The 10-fold cross-validaton (CV) was used to evaluate the performance of model, also predictivity of the model was evaluated on test set. All models were compared using classification accuracy, AUC, MCC and F-score, and then, the best model was chosen by this way. The result of QSAR modelling with random forest classifiers are promising for both comedogenicity and skin irritancy prediction. While for comedogenicity prediction, Alva model with 3 descriptor gave successful results with 86.96% accuracy for the 10-fold CV model and 79.31% accuracy for the test set. Skin irritation model, Mold2 with eight descriptors gave a successful performance with 85.75% accuracy for 10-fold CV model and 82.75% accuracy for the test set. In conclusion, this stuy provided an animal-friendly, rapidly perform and inexpensive two different classification model for the safety evaluation of cosmetics. Further studies are needed to develop their performance and to extend their chemical domain.

Keywords: *in silico*, comedogenic, machine learning, skin irritation, QSAR, cosmetics

ÖZET

Oztan Akturk, S. (2022). Komodojenite ve Cilt İrritasyon Potansiyeli Olan Kozmetik Bileşenlerinin *in Silico* Methodlar ile Değerlendirilmesi. Yeditepe Üniversitesi, Sağlık Bilimleri Enstitüsü, Farmasötik Toksikoloji ABD, Doktora Tezi, İstanbul.

Komodojenite ve deri tahrişi kozmetik bileşenlere karşı görülen yaygın advers reaksiyonlardır. Komodojenite, bir bileşenin akne için ilk lezyon olan komedon oluşumunu indüklemeye potansiyelidir. Deri tahrişi ise bir kimyasalın uygulanmasının ardından deride geri dönüşümlü hasar oluşturmaktır. 2013 yılında Avrupa Komisyonu tarafından kozmetikler üzerinde hayvan testleri yasaklanmadan önce, kozmetiklerin komodojenik potansiyeli tavşanlarda test edilmekteydi. Ancak, hayvan testleri henüz alternatifleriyle tam olarak değiştirilememiştir. Bu nedenle, bu tez çalışmasında, farklı makine öğrenme algoritmaları ve moleküler tanımlayıcı türleri kullanarak kozmetik bileşenlerin komodojenik ve deri tahrişi potansiyelini tahmin etmek için iki farklı QSAR sınıflandırma modeli geliştirmeyi amaçladık.

Verisetimiz literatürde hem komodojenite hem de irritasyon için tavşan kulaklarında test edilmiş, yağ asitleri, yağ alkolleri ve türevleri ve pigmentler gibi, 121 kozmetik içerikten oluşmaktadır. Toplam, 4837 moleküler tanımlayıcı dört farklı yazılım aracılığıyla hesaplandı. Modeller için tanımlayıcılar WEKA’da korelasyon ve öğrenmeye dayalı (karar ağacı algoritması aracılığıyla) iki yöntem karşılaştırılarak seçildi. Modelleme çalışmaları yine WEKA’da çeşitli sınıflandırıcılar test edilerek yapıldı. Model performansı 10 kat çapraz-doğrulama ile ve tahmin yeteneği ise test seti üzerinde değerlendirildi. Tüm modeller, sınıflandırma doğruluğu, AUC, MCC, F skor gibi metrikler açısından karşılaştırıldı ve en iyi model seçildi. Rastgele orman sınıflandırıcısıyla geliştirilen QSAR model sonuçları hem komodojenite hem de deri tahriş tahmini için umut vericidir. Komodojenite tahmini için 3 tanımlayıcı Alva modeli çapraz doğrulama için %86.96 doğruluk ve test seti için %79.31 doğrulukla başarılı sonuç vermişken, Deri tahriş modeli, sekiz tanımlayıcı Mold2 model ise çapraz doğrulama için %85.75 doğruluk, test seti için %82.75 doğrulukla başarılı performans vermiştir. Sonuç olarak, bu çalışma kozmetiklerin güvenlik değerlendirmesi için hayvan dostu, ucuz ve hızlı performans gösteren iki farklı sınıflandırma modeli sağlamıştır. İleri çalışmalarla performanslarının artırılması ve kimyasal alanının genişletilmesi gerekmektedir.

Anahtar kelimeler: *in silico*, komodojenik, makine öğrenmesi, deri tahrişi, QSAR, kozmetik

1. INTRODUCTION AND AIM

In March 2013, cosmetic testing on animals was prohibited by European Commission under the Cosmetic Regulation (1). In terms of the pros and cons of alternative test methods to animal testing, replacement is the best way to avoid animal pain and suffering. Replacing animal experiments is necessary not only for animal welfare but also for saving time and cost and improving the quality of experimental data. But still, full replacement of animal testing with alternative methods has not been yet possible. Therefore, there is a need for alternative testing and predicting strategies for cosmetic research (2).

In silico methods are promising alternative methods to animal models to predict the cosmetic testing results (3). *In silico* means “performed on computer or computer simulation model” (4). The aim of “*in silico*” or computational toxicology is to predict the toxicity of molecules on the basis of computational models. These models can be created by experts (e.g., via structural alerts or read across) or created automatically (e.g. via machine learning techniques). In predictive toxicology, quantitative structure-activity relationships (QSARs) is one of the mostly used computational methods that create relationship between the chemical structure of a molecule and the toxicity endpoint. It is derived from the assumption that similar molecules display similar biological activity. For that reason, finding the significant and relevant molecular descriptors are important issue to build a good and predictive QSAR model (4, 5, 6, 7).

Recently, in predictive toxicology, application of machine learning methods and techniques have been widely used to create better QSAR models by the way of more sophisticated analyses (8, 9). In general, machine learning algorithms learn from the data by the way of mapping and finding the function between the independent and dependent variables. The best function was found by mapping the best from input to output variables. (10). In QSAR modelling, there are numerous machine learning algorithm for classification, in other words, classifiers such as decision trees, random forest, logistic regression, support vector machines (8, 10, 11, 12).

One of the typical adverse reactions to cosmetics is comedogenicity, especially for acne-prone skin (13). Comedogenicity is the potential of cosmetics (may be ingredients or product) to induce formation of comedones (blackheads and whiteheads) which are the initial lesion of acne (13, 14, 15, 16). Acne cosmetica is caused by cosmetic

products containing comedogenic ingredients (17) and firstly reported by French dermatologists in the middle of 1940s. This skin problem is characterized a noticeable acneiform eruption containing mainly of closed comedones with few papulo-pustules on the forehead and temples (15). In 1972, popular American-dermatologists, Kligman and Mills were tested the comedogenic potential of twenty-five face cream on rabbit ears, they found the half of them as mildly comedogenic. The commonly known comedogenic cosmetic ingredients are butyl stearate, lauric acid, lauryl alcohol, oleic acid, oleyl alcohol, lanolin, petrolatum and pigments (16). The comedogenicity of cosmetics had been tested on rabbit ears before cosmetic testing on animal was prohibited by EU Comission under the Cosmetic Regulation in 2013. (1, 15). However, an alternative test method to animal models has not been possible yet for comedogenicity prediction.

The other typical adverse reaction to cosmetics is skin irritation, especially for sensitive skin. Skin irritation is the formation of reversible damage to the skin after the application of a cosmetic ingredient for up to 4 hours (13, 18). In the context of safety assessment of cosmetics, the evaluation of the skin irritancy potential of an cosmetic ingredient is a necessity (19). In the past, skin irritation potential of chemicals had been evaluated widely on the draize rabbit skin. In recent years, *in silico* methods for prediction of skin irritation have been offered an attractive alternative to both animal and *in vitro* tests due to their practicability, low cost and animal welfare (20). In the literature, there are some *in silico* techniques to predict the skin irritation potential of certain class of chemicals. One of initial SAR models for skin irritation was developed by Enslein et al., (1987) separately for aromatic and aliphatic compounds. Due to the fact that used database did not give numerical scores, severity ratings were used in the model (20, 21). Barrat (1995) derived QSARs for skin corrosivity of 41 organic acids and bases and 15 phenols. Skin corrosivity data of chemicals was evaluated associated with their logP, pKa, melting point, and molecular volume. The skin permeability (LogP) and cytotoxicity (pKa) for each chemical class showed good performance to discriminate skin corrosive chemicals from non-corrosive ones (22). Hayashi et al., (1999) built a skin irritation QSAR model for 24 phenols related to reactivity and skin permeability (23). Smith et al., (2000) built a SAR model to distinguish 34 esters from non-irritant esters by the way of using nineteen physiochemical properties illustrating the electronic, steric, transportational features (24). Kodithala et al., (2002) built the chemical-membrane interaction (MI)-QSAR models for 22 hydroxy-organic acids, 13 aliphatic alcohols, and

11 phenols to predict skin irritation (25). They were all linear regression model. Golla et al., (2009) developed one of the first non-linear QSPR skin irritation models based on the Draize rabbit test data and related to skin penetration, reactivity and molecular size. Furthermore, various expert systems based on knowledge were developed to estimate the skin irritation potential of chemicals by using a number of physiochemical properties of molecules or via pattern recognition algorithms (20).

Briefly, in the literature, there have been available SAR/QSAR models developed for the prediction of skin irritation potential of some specific chemicals classes such as phenols, esters, organic acids. However, there has been no SAR/QSAR models to evaluate and predict the comedogenic potential of chemicals.

Therefore, in this thesis study, we aimed;

- 1- to develop the first QSAR classification model to predict comedogenicity
- 2- to develop novel QSAR classification model to predict the skin irritation potential of some cosmetic ingredients with a validated *in silico* model in order to get a better prediction accuracy.

The outcome of the thesis will contribute rationale of cosmetic product design and provide animal-friendly, rapidly perform, and inexpensive models to predict comedogenicity and skin irritation potential of some cosmetic ingredients. Furthermore, they will also potentially be used for safety assessment approaches for cosmetic products.

2. GENERAL INFORMATION

2.1 *In Silico* Toxicology and *In Silico* Methods

In silico means “performed on computer or computer simulation” (4). In recent years, *in silico* toxicology has become increasingly important in the assessment of chemical safety/toxicity and the drug development process (26). The aim of *in silico* toxicology is to predict chemical toxicity via computational modelling (27). Computational methods is used to analyze and predict the toxicity of chemicals (6). The computational tools using *in silico* toxicology are given in the following (Figure 1) (28).

- a) databases used for keeping data belongs to chemicals such as structural, physicochemical properties and their toxicity data
- b) software used for calculating molecular descriptors
- c) simulation tools for mimicing the biology of the systems and molecular dynamics
- d) modelling methods for predicting the toxicity
- e) modelling software for building the model
- f) expert systems for pre-built models
- g) visualization tools

The significant *in silico* modelling methods for toxicity prediction is given in the following (Figure 1) (28).

- 1) Quantitative structure-activity relationships (QSARs)
- 2) Read-across
- 3) Structural alerts
- 4) Dose and time response models,
- 5) Pharmacokinetic and pharmacodynamic models
- 6) Uncertainty factors models

In silico modelling methods include five major steps for developing the prediction models as seen in Figure 1. These are (1) collecting the biological activity data and structural data, (2) calculating molecular descriptors used in the modelling studies (3) developing a prediction model, (4) evaluating of the accuracy of the model, and (5) interpreting the developed model (28).

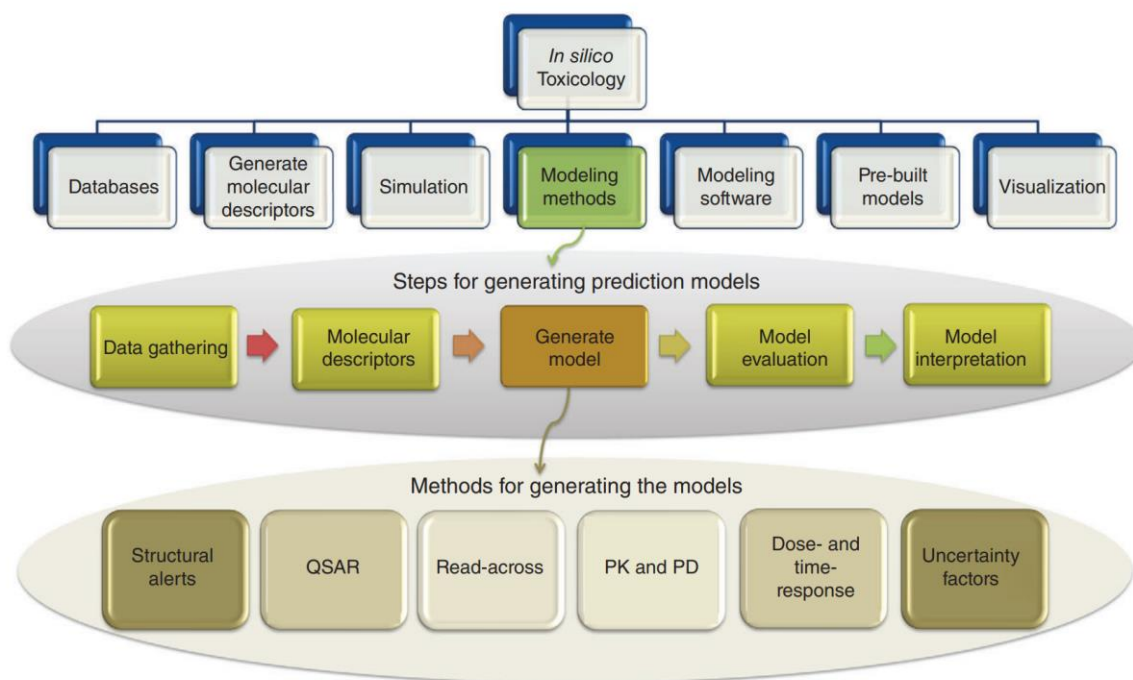


Figure 1. The computational tools in silico toxicology, steps and methods for generating predictive models (28).

The selection of *in silico* method has been including the following five criteria (6):

- 1) Relevant mechanism or toxicological effects
- 2) Model validity:

There are five principles defined by OECD for validation of an *in silico* model

- a. a defined endpoint
- b. an unambiguous algorithm
- c. a defined domain of applicability
- d. appropriate measures of goodness of fit, robustness and predictivity
- e. a mechanistic interpretation, if possible (6).

- 3) Chemical space: *in silico* models will only make predictions for specific classes of chemicals, they are so-called “applicability domain”
- 4) Model combination
- 5) Supporting an expert review

2.1.1 QSAR and *In Silico* Modelling

Quantitative structure-activity relationship (QSAR) is the widely used *in silico*-modelling methods that has been using the molecular descriptors to predict toxicity of chemicals (28). This methodology has been used a mathematical model derived from a training set of compounds. The training set includes the compounds that were found to be positive and negative in a given toxicological endpoint or to induce a continuous response that the model will predict (6).

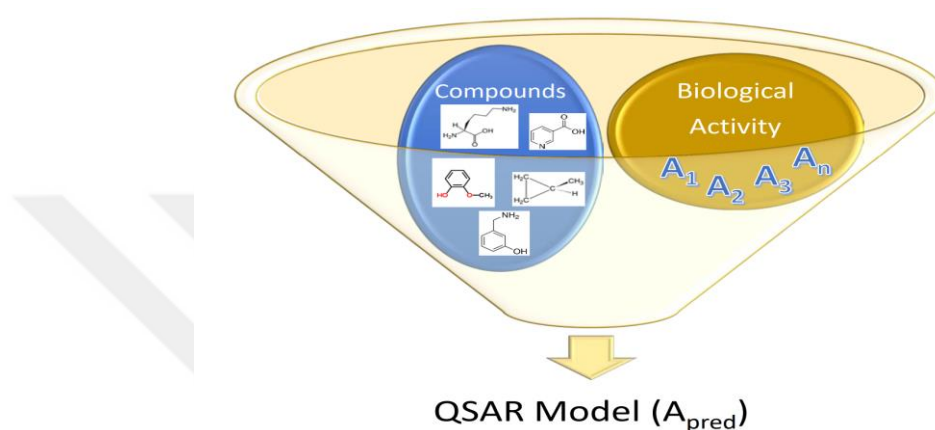


Figure 2. QSAR model (29)

- In a QSAR model (Figure 2);
 $A_{pred} = f(D_1, D_2, \dots, D_n)$, where
 A_{pred} : the toxicological endpoint that will be predicted.
 $D_1, D_2, \dots, D_n$: molecular descriptors (calculated from the chemical structure)
 $A_1, A_2, \dots, A_n$: toxicological endpoint (biological activity) of training compounds
- The SAR (Structure-Activity relationship) is an approach to finding qualitative relationships between the molecular structure of compounds and their biological activity (toxicological endpoints).
- QSAR is a mathematical model that relates a quantitative measure of the structure of molecules and their biological activity (toxicological endpoints) (30).
- QSAR model is based on the principle that structurally similar compounds can be likely to have similar biological activity (7).
- The QSPR (Quantitative Structure-Property Relationship) is found the relationship between the structure of compounds and physical property (31).

2.1.2 History of QSAR

QSAR modelling is born in the toxicology field. Corwin Hansch (with the study in 1962) has been regarded as the father of the modern QSAR by current predictive toxicologists (32). He had proved the presence of a quantitative relationship between biological activity (acute toxicity) and log P (octanol-water partition coefficient, hydrophobicity term) which is the measure of lipophilicity of a compound that encode the hydrophobic interaction of the compound with the cell membrane (31, 32). He had investigated the relationship between the activity of compounds and their hydrophobic (log P), electronic (E), steric (S) and other effects using multiple regression correlation methodology as given in the follows:

$$\text{Biological activity (Y)} = a + b \log P + cE + dS + \dots$$

The major pioneering events for QSAR studies are given in Figure 3 (33).

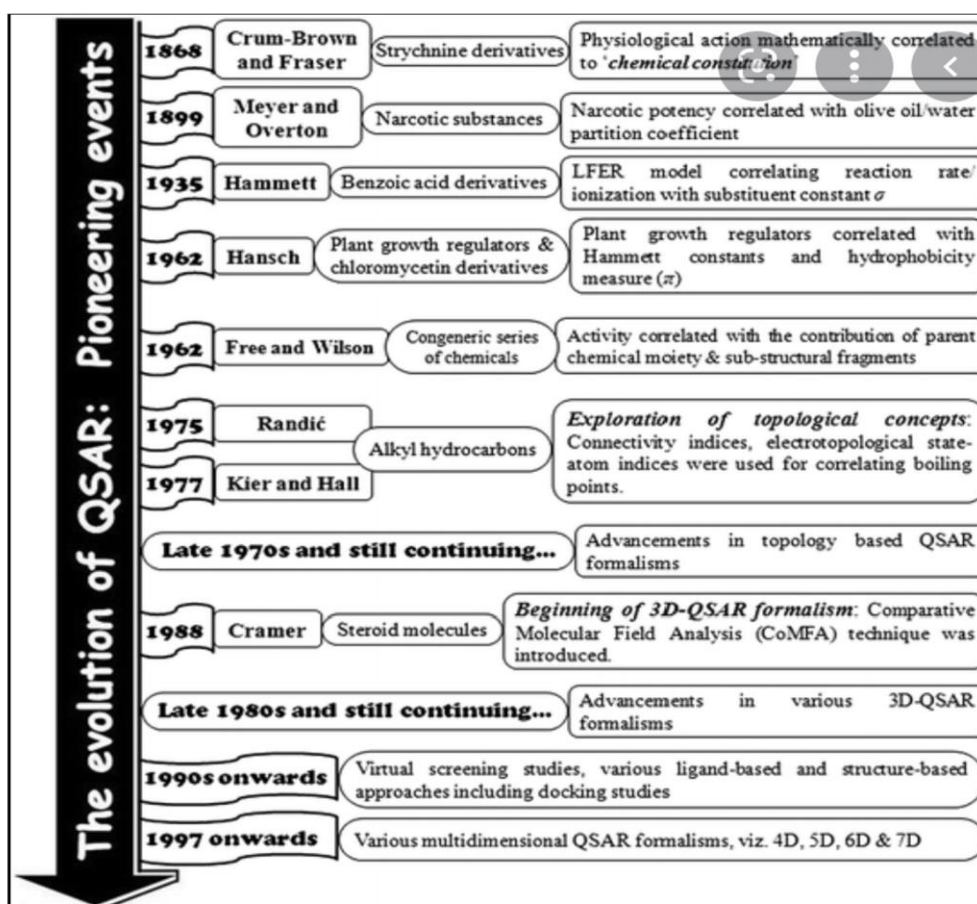


Figure 3. Short History of QSAR evolution (32, 33).

2.1.3 Toxicological Endpoint

In QSAR, endpoint refers to any biological activity or physicochemical property associated to the chemical structure of a molecule that can be measured and modelled (34). In other words, input experimental data is called an endpoint. If this experimental data is related to the toxicity, it is a kind of toxic response that can be modelled by QSAR or QSPR as a dependent variable (31). Exposure to chemicals can cause acute or chronic toxicity to different organs or result in adverse health effects (35). These toxicological endpoints (biological activity) being predicted by a given QSAR model can be listed in the following (6, 35, 36).

- Acute toxicity
- Skin irritation and corrosion
- Eye irritation and corrosion
- Skin sensitization
- Dermal toxicity and percutaneous absorption
- Mutagenicity or genotoxicity
- Carcinogenicity
- Repeated dose toxicity
- Reproductive toxicity
- Developmental toxicity
- Toxicokinetic and metabolism
- Photo-induced toxicity
- Hepatotoxicity
- Renal toxicity
- Neurotoxicity
- Endocrine toxicity
- Immunotoxicity
- Cardiotoxicity
- Ocular toxicity
- Cytotoxicity

2.1.4 Molecular Descriptors

Molecular descriptors are quantitative representations of various molecular properties or chemical structures of compounds (30). Their structural dimensional (D) representations are range from 0D to 4D as shown in Figure 4 (37). There are different category of molecular descriptors used in OSAR modelling. Some of the most used main category of molecular descriptors are given in Fig 4. They are constitutional descriptors, geometrical descriptors, thermodynamic descriptors and electronic descriptors (38). The quality of the QSAR model is dependent on both the type and number of molecular descriptors. The selection of descriptors is very important due to the fact that it affects the predictive performance of a model (30).



Fig 4. Some of Main Category of Molecular Descriptors Used in QSAR Modelling (38)

2.1.4.1 Types and Dimensions of Descriptors

The structure of any molecule can be transformed into 0D, 1D, 2D, 3D and 4D molecular descriptors. The representation of these five dimensions of a molecule is shown in Figure 4. As the dimension of the descriptors increase, the information that provides as well increases, however, the ease of calculation of descriptors decreases. Furthermore, the description of the five dimensions of the molecular descriptor and their examples are given in Table 1. 0D, 1D and 2D molecular descriptors are easy to interpret. Therefore, they are common choice descriptors. Due to the fact that 3D and 4D descriptors have conformational problem, this condition affects their computation and complexity. The selection of the combination of descriptor types depends on the goal of the model that will be developed. For an optimal reliable model, the descriptors could be easily computed and have good prediction capability and mechanistic interpretation (30). Some

of the topological and geometrical descriptors used in QSAR model building according to their chronological order with their description is listed in Table 2 (38).

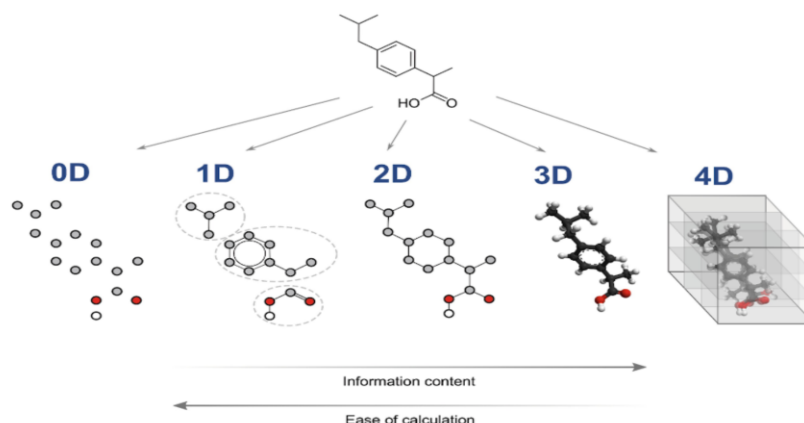


Figure 5. The representation of five dimensions of molecular descriptors (37)

Table 1. Description of five dimensions of molecular descriptors and their examples (39, 40).

Dimension	Description	Examples
0D	Chemical formula derived descriptors. Not provide any information about the molecular structure or atom connections.	- Constitutional descriptors (Molecular weight, number of aromatic rings) - Molecular property descriptors - Count descriptors (count of bond, atom, non-hydrogen atom)
1D	Descriptors derived using the various sub-structural molecular fragments such as functional groups. Not whole topology, as a part of the chemical structure.	- Fingerprints (binary vector, where 1 represents the presence of a structural feature and 0 represents its absence). - Count of fragments - H-bond acceptor/donor - Crippen AlogP98
2D	Descriptors calculated from the 2D molecular graph of the molecule. They provide information on molecular topology-based representation such as atomic composition and connectivity. They are sensitive to structural features of the molecules (size, shape, and symmetry).	- Topological descriptors - Eigenvalue-based descriptors - Connectivity indices - Descriptors containing topological and electronic information - Descriptors encoding property info such as atom pairs, multiples of pharmacophore. 2D Autocorrelation descriptors
3D	Descriptors derived from the 3D structure of the molecule. They encode various spatial as well as geometrical information of molecules. Also, they display the static representation of ligands. It takes a long time to calculate.	3D-MoRSE descriptors - WHIM descriptors - GETAWAY descriptors 3D Autocorrelation descriptors - Quantum-chemical descriptors (HOMO, LUMO energy). - Atomic coordinates, size, steric, surface and volume descriptors - Techniques such as; ComFA (Comparative Molecular Field Analysis) CoMSIA (Comparative molecular similarity index analysis)
4D	Descriptors derived from multiple representations of the ligand molecule using various configurations and orientation	- GRID descriptors - Volfsurf

Table 2. Descriptors used in QSAR/QSPR modelling according to their historical development (38).

Year	Index	Author(s)	Description
<i>Topological (2D) descriptors</i>			
1947	Wiener (<i>W</i>)	Wiener	Defined as the sum of edges in the shortest path in a chemical graph between all pairs of nonhydrogen atoms in a molecule
1971	Hosoya (<i>Z</i>)	Hosoya	The number of sets of nonadjacent bonds in a molecule, useful for building the QSAR and QSPR models describing the physical properties
1972	Zagreb indices	Gutman	The first topological indices used for the total π -energy of conjugated molecules
1975	Connectivity	Randic	The first-order connectivity index is a bond additive molecular structural invariant. Such indices have wide applications in structure–property and structure–activity studies
1975	Higher order connectivity	Kier <i>et al.</i>	Higher order connectivity indices are weight paths, where the higher weight has been assigned to terminal bonds and a lesser weight to the less exposed inner bonds
1980	Autocorrelation indices	Moreau-Broto	First introduced to define a relation between atoms as a function of their spatial separation; have certain advantages for any QSAR/QSPR study: fragment independent; invariant to roto-translation; encode the identity and electronic attributes, including atom types, partial atomic charges, electronegativity, and polarizability
1982	Balaban <i>J</i>	Balaban	Distance connectivity index, or average distance sum connectivity, is also one of the most discriminating molecular descriptors. Its value is independent of molecular size or number of rings
1985	Kappa shape	Kier	Shape of molecules defined in terms of number of atoms and their bonding pattern. Present in various orders (e.g., first, second, third order, etc.); first order index involves a count of single bonds, whereas second order index involves a count of a two bond path, and so on
1993	Hyper-Wiener index <i>WW</i>	Randic	Hyper-Wiener index <i>WW</i> of a chemical tree <i>T</i> is defined as the sum of the products n_1n_2 , over all pairs <i>u</i> , of vertices <i>T</i>
1994	Szeged (<i>Sz</i>)	Gutman	Obtained as a bond additive quantity in which bond contributions are given as the product of the number of atoms closer to each of the two points of each bond
2001	Modified Hosoya index <i>Z*</i>	Randic	Frequency of occurrence of individual CC bonds in the patterns of disjoint bonds are considered
2001	Modified Wiener index <i>W*</i>	Nikolic <i>et al.</i>	Bond contributions determined using the reciprocal of the product of the number of atoms on each side of a bond
2003	Novel Wiener index	Li <i>et al.</i>	Obtained as a bond additive quantity in which bond contributions are given as the product of the number of atoms closer to each of the two points of each bond
2007	Topological Maximum Cross Correlation (TMACC)	Melville <i>et al.</i>	Generated from atomic properties determined by molecular topology. These descriptors are based on concepts derived from autocorrelation descriptors
2010	Augmented Zagreb index (AZI)	Furtula <i>et al.</i>	Based on the atom-bond connectivity (ABC) index, used for obtaining the extreme AZI values for chemical trees; applicable for tight upper and lower bounds of chemical trees
2011	Superaugmented eccentric distance sum connectivity	Gupta <i>et al.</i>	Fourth-generation topological indices (TIs) exhibiting high discriminating power, low degeneracy, and high sensitivity toward both the presence and relative position of heteroatom(s)
<i>Geometrical (3D) descriptors</i>			
1977	3D-Molecular representation of Structures based on Electron diffraction (MoRSE)	Soltzberg and Wilkins	MoRSE descriptors have been shown to have good modeling power for different biological and physicochemical properties and can be used for the simulation of infrared spectra
1994	Weighted Holistic Invariant Molecular (WHIM)	Todeschini <i>et al.</i>	WHIM descriptors are used to capture the relevant 3D information regarding the molecular size, shape, symmetry, and atom distribution; have been used for modeling toxicology and various physicochemical properties. At least ten different WHIM types (e.g., WSIZ, WSHA, WDEN) of descriptors with different molecular features have been developed
1995	3D Autocorrelation	Wagener <i>et al.</i>	These descriptors are calculated by applying the autocorrelation function at distinct points on molecular surface. They are unique for a given geometry, sensitive conformational change, and are invariant to roto-translation
2002	GEometry, Topology, and Atom-Weights Assembly (GETAWAY)	Consonni <i>et al.</i>	GETAWAY descriptors are based on spatial correlation formula that give weight to the atom to count atomic mass, van der Waals volume, and electronegativity along with 3D information. Based on matrix operator and information indices, seven GETAWAY descriptors have been reported so far

2.1.4.2 Software for Calculation of Descriptors

There are various software packages for the calculation of the molecular descriptors used in the QSAR modelling study. These commonly used packages are listed in Table 3. (30, 38, 40).

Table 3. Software for calculating molecular descriptors and fingerprints (38, 40).

Software	Descriptors	Type of Descriptors	Web adress	Free or not
AlvaDesc	5471	0D,1D,2D,3D descriptors and fingerprints (MACCS and Hashed chemical)	https://www.alvascience.com/alvadesc/	Commercial
ACD/labs	-	LogD, logP, pKa values, physiochemical and ADME properties	https://www.acdlabs.com/	Freeware
ADMET predictor	297	Constitutional, functional group counts, topological, E-state, 3D descriptors, molecular patterns, acid-base ionization, empirical estimates of quantum	http://research.chem.psu.edu/pcigroup/adapt.html	Commercial
ALOGPS		log P, log S	www.vcclab.org	Freeware
CDK		Topological, geometrical, electronic, constitutional	https://cdk.github.io/	Freeware
CODESSA	1500	Constitutional, topological, geometrical, charge-related, semi-empirical, thermodynamical	www.codessa-pro.com	Commercial
E-DRAGON	1664	0D,1D,2D,3D descriptors. (Constitutional, functional group and fragment counts, but also several topological and geometrical descriptors)	http://www.vcclab.org/lab/edragon/	Freeware
Mold2	779	0D, 1D and 2D molecular descriptors	https://www.fda.gov/science-research/bioinformatics-tools/mold2	Freeware
MOE	400	2D, 3D descriptors Topological, physical properties, structural keys, etc.	https://www.chemcomp.com/	Commercial
Mordred	1825	1D, 2D descriptors Atom count, Autocorrelations, RingCount, TopologicalIndex, GeometricalIndex, MorSE vb.	https://mordred.phs.osaka-u.ac.jp/	Freeware
PaDEL	1875	663 (1D, 2D descriptors), and 134 (3D descriptors) and 10 types of fingerprints.	www.yapcwsoft.com/dd/padeldescriptor/	Freeware
OpenBabel	-	MOLPRINT2D, 166-bit MACCS, Daylight fingerprint (FP2), structural key fingerprints	http://openbabel.org/	Freeware
QuaSAR-Descriptors	-	2D, 3D descriptors Physical Properties, Subdivided Surface Areas, Atom and Bond Counts, Kier& Connectivity Indices, Pharmacophore Feature Descriptors, MOPAC Descriptors (3D)	http://www.cadaster.eu/sites/cadaster.eu/files/challenge/descriptors.htm	Freeware
TEST (Toxicity Estimation Software Tool)	797	0D,1D,2D descriptors Constitutional, Topological descriptors, Molecular Fragments, Molecular Properties, 2D Autocorrelation descriptors, vb.	https://www.epa.gov/chemical-research/toxicity-estimation-software-tool-test	Freeware

2.1.5 QSAR Modelling Approaches

There are two QSAR modelling approaches; classification model and regression model (Figure 6).

2.1.5.1 Classification model

If binary experimental data (e.g. toxic or not toxic) are introduced to represent the activity or property in the input matrix for the model generation, the QSAR model is referred to as classification model and make a qualitative prediction (30).

2.1.5.2 Regression model

If the considered endpoint is represented by numerical continuous experimental data, the QSAR model is referred to as a regression model and make a quantitative prediction (30).

2.1.6 QSAR Modelling Flowchart

The key steps in any QSAR modelling are summarized in Figure 6. These are data collection, calculation of suitable molecular descriptors via descriptor packages, the selection of descriptor for modelling, model generation, optimization, finally validation both internally and externally and in silico prediction (30).

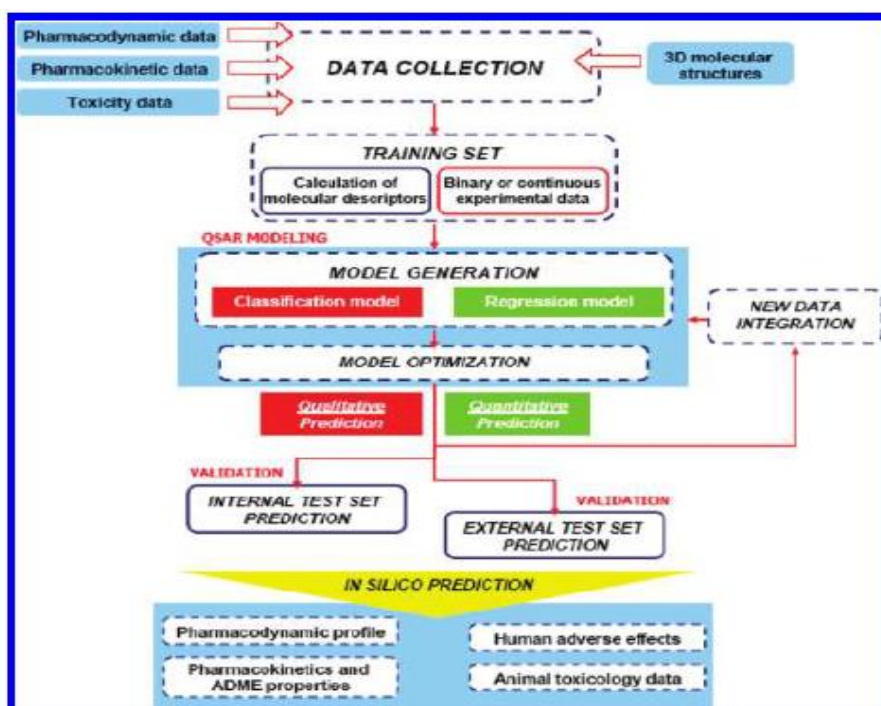


Figure 6. Key steps of QSAR modelling (30).

QSAR models encode the relationship between the descriptors selected for model and the toxicological response. After the model is built and validated, the model can be used to make an *in-silico* prediction (6). All stages of any QSAR modelling protocol are shown in Figure 7 (34, 41).

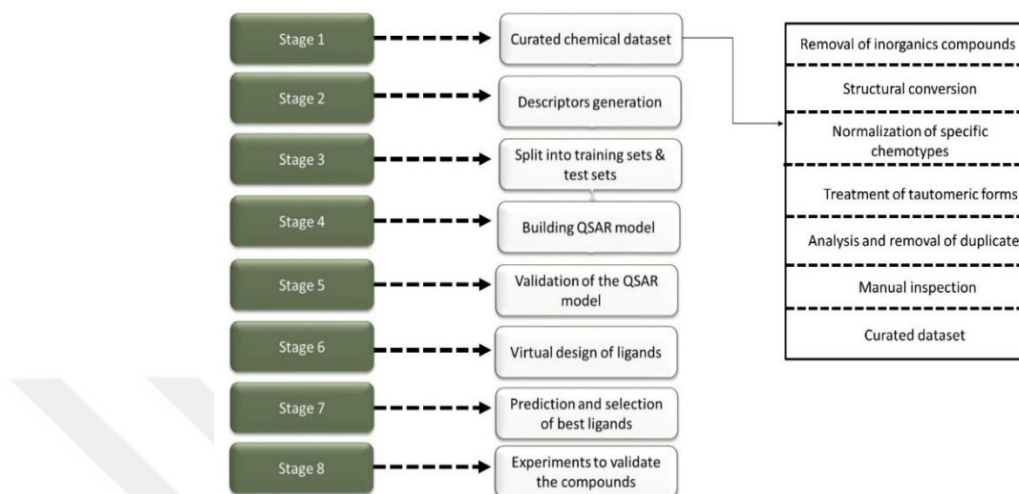


Figure 7. QSAR protocol (41)

The success of any QSAR model varies based on many factors e.g the quality of data, selection of appropriate descriptors, choice of methods for modelling and validation method of the developed model. Validation of QSAR models must be mainly for robustness, prediction performances and applicability domain (AD). Ultimately any QSAR model should be capable of making accurate and reliable predictions (42, 43, 44, 45, 46, 47).

2.1.7 Factors Affecting the Predictive Ability of QSAR Models

There are three main factors that affect the predictivity of QSAR models: They are quality of experimental data, descriptor selection and statistical algorithm (28).

a. Quality of experimental data:

It is important to obtain a dataset from one resource (same laboratory and same experimental design) in terms of the predictivity and the statistical significance of the QSAR model. When the data is obtained from multiple resources, it may affect the quality of data negatively in terms of conditions being different in the application during the experiment. Also, the dataset should be balanced which is the equal number of active and inactive due to the fact that some statistical methods for modelling are unstable to unbalanced datasets (28).

b. Descriptor selection:

The feature selection method is important to choose the best descriptor for a predictive QSAR model. The descriptor selection process continues until a good performance has been gotten for the model (Figure 8). Furthermore, the number of descriptors in a QSAR model should be small as possible for interpretability. Also, descriptors should be independent, uncorrelated, and non-redundant (28).

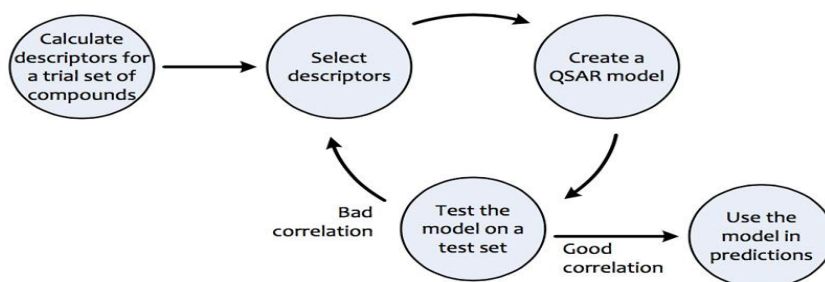


Figure 8. The descriptor selection process for predictive QSAR model creation

c. Statistical algorithm:

The model should be as possible as simple and accurate. The complex algorithm can be given the better performance, however, the interpretability of the model is an important issue.

2.1.8 Advantages of QSAR Modelling

The main advantages of QSAR model development are given in Figure 9 (48).

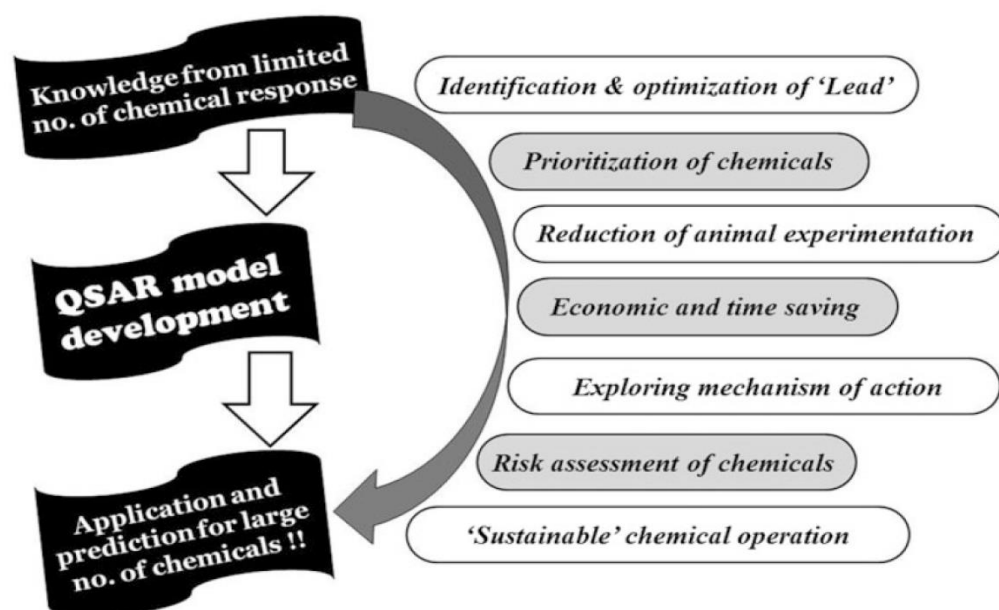


Figure 9. Advantages of QSAR model development (48).

2.1.9 Application Area of QSAR

The popular application area and disciplines of QSARs are drug discovery and lead optimization, material science and predictive toxicology for the purpose of risk assessment, toxicity prediction and regulatory decision (48). All of the application areas of QSAR are given in Figure 10.

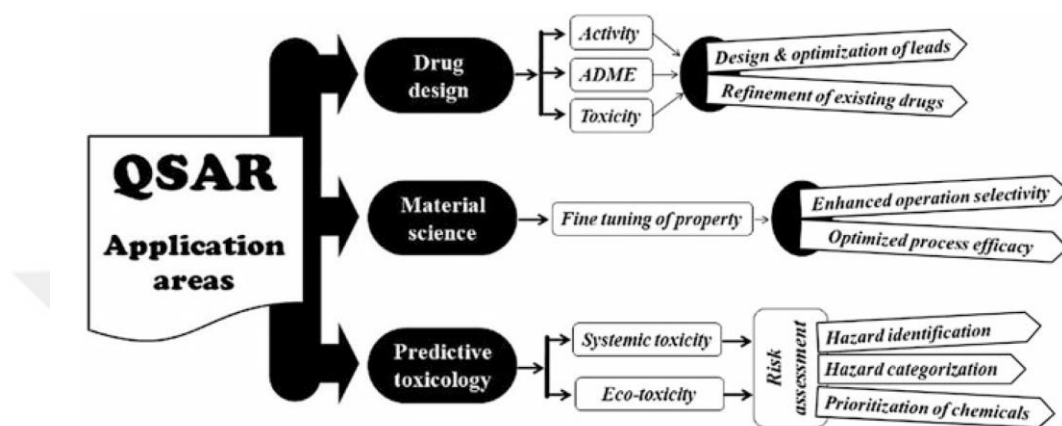


Figure 10. Application area of QSAR (48)

The successful applications of QSAR in drug discovery are given in Figure 11.

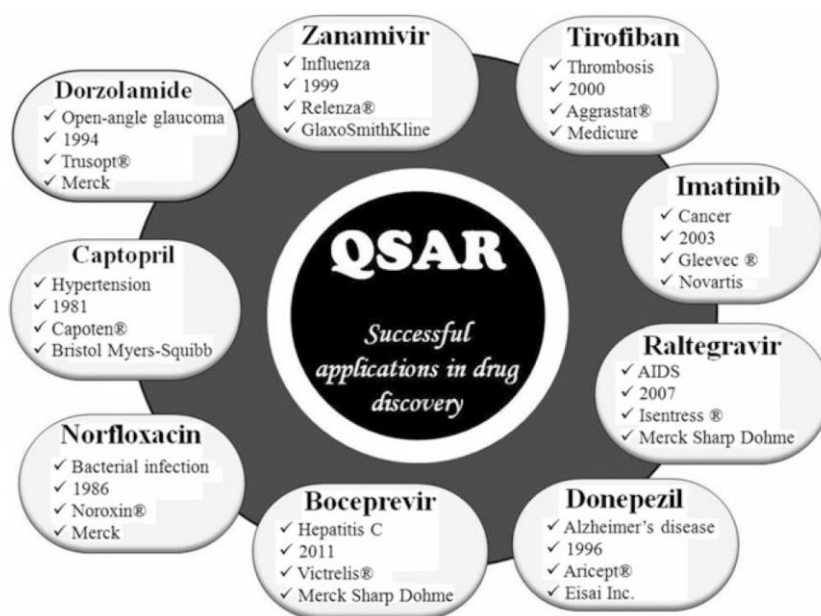


Figure 11. The successful applications of QSAR in drug discovery (48).

2.1.10 Adverse Outcome Pathway

In predictive toxicology, toxicity prediction strategies such as QSAR should be related to key events in Adverse Outcome Pathways (AOPs). An AOP is describes as a series of connected events that starts with a molecular initiating event (MIE) (e.g. inhibition of an enzyme) like as in Figure 12. It progresses by means of a series of key events (KE) that manifest at cellular level, tissue, organ and organism levels. Each AOP should have one MIE and one AO (adverse outcome), but there is no limitation on the number of KEs. One of the main goal of AOPs is to connect available knowledge on mechanisms of toxicity and describe the key events leading to an AO which is a toxicological endpoint (49, 50).



Figure 12. Schematic AOP diagram (49).

*MIE: molecular initiating event, KE: key event, KER: key event relationship, n: number of the AOP

The first well-characterized AOP, describing the chemical potential for causing skin allergy has been published by OECD (2012) (51). (Figure 13). The AOP Knowledge Base is one of the essential components of the OECD program that facilitates scientific collaboration on an international scale to aid both the development and evaluation of AOPs (52). All events in the AOP of skin allergy, that are from chemical structure through molecular initiating events, key events, adverse outcomes, are shown in Figure 13. (52).

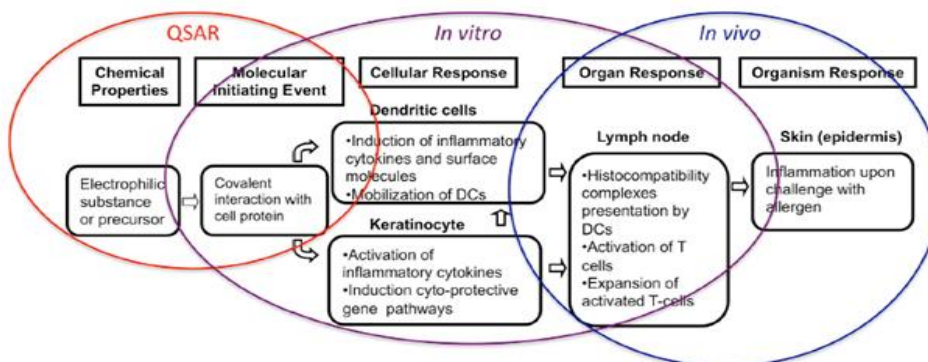


Figure 13. AOP diagram of chemically-induced skin allergy showing the different pathways.

2.2 Machine Learning

Machine Learning (ML) is a subdivision of Artificial Intelligence (AI) due to its learning and decision-making abilities (53). ML is focused on the learning feature of AI by developing an algorithm that best represents the set of data (54). AI is a technique that enables machines to mimic human behaviour. The potential application of AI in our daily lives has been becoming things like forecasting weather, self-driving cars or predicting the amounts of new customers for companies. ML and deep learning (DL) can be defined as the ability of AI to learn a representation of data with multiple levels of abstraction with the help of an algorithm that shows the size of the changes when compared to its previous internal parameters. The relationship between artificial intelligence (AI), machine learning (ML) and deep learning (DL) are shown in Figure 14 (55).

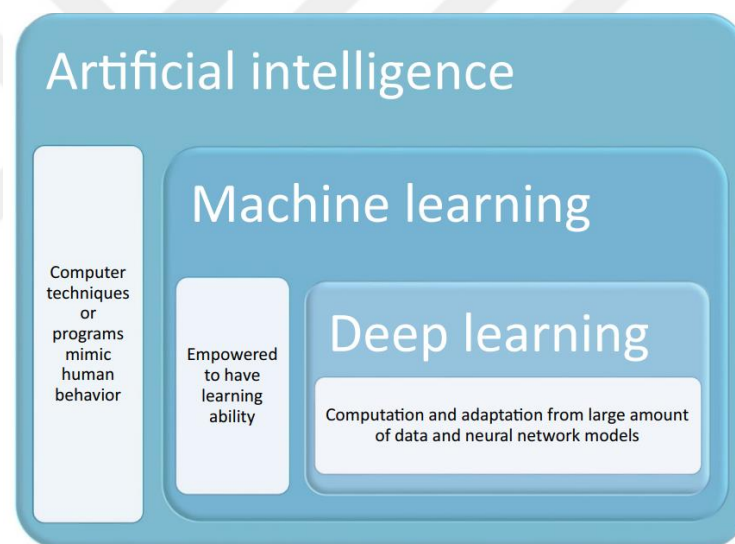


Figure 14. Comparison between artificial intelligence (AI), machine learning (ML) and deep learning (DL) (55).

Although DL is just a subset of ML, the difference between ML and DL can be explained in terms of their capabilities and need for human intervention. Besides, DL has functions in a similar way to ML. Despite the fact that; ML models have progressively become better at performing their specific function as new data they take, they still need human intervention (Figure 15). However, the DL model has no need for human help since its algorithm determine whether or not a prediction is accurate through its own neural network. DL is based on an artificial neural network (ANN) which is an ML

algorithm with a brain-like logical structure used in AI models. If an AI algorithm returns an incorrect prediction, then an engineer has to step in and make adjustments.

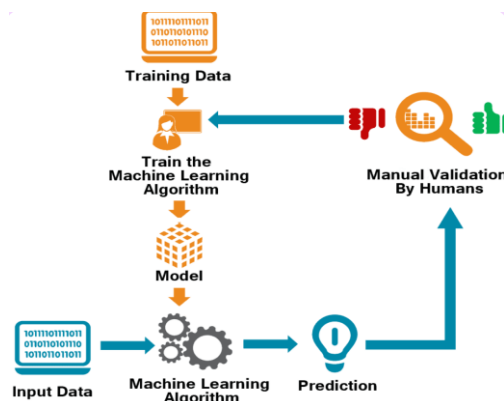


Figure 15. ML model needs human intervention for manual validation

The application of AI technologies in the areas such as epidemiological prediction of cancer can be very useful for public health and safety (55). Today, in predictive toxicology, the ML-based QSAR model has been widely used to describe all forms of predictive modelling (5, 56). The ML-based QSAR modelling has been playing important role in the prediction of toxicity, property and category. Since the 1990s, ML algorithms such as random forest (RFs) and support vector machines (SVMs) have been widely used to build QSAR models for the prediction of toxicity and discovery and designing of new drug molecules (56). Besides, data science uses ML algorithms to build predictive models, extract useful data and predict future patterns and behaviours (Figure 16).

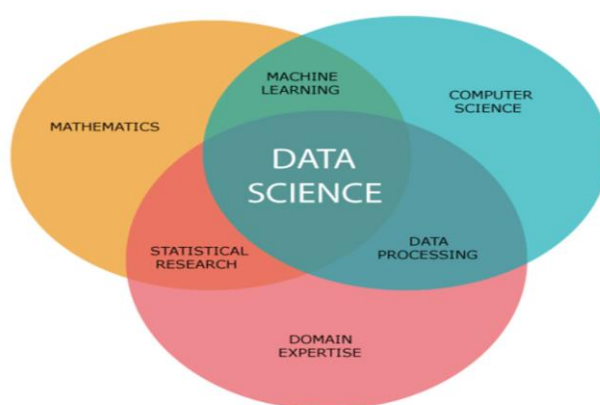


Figure 16. Data science, a combination of techniques and theories taken from many fields

2.2.1 History of Machine Learning

The first definition of artificial intelligence (AI) was established by Alan Turing in the 1950s. He studied how machinery could be used to mimic processes of the human brain. The first publication of AI which entitled Intelligent Machinery belongs to him (57). ML is a sub-field of AI that includes all approaches allowing computers to learn from data without being explicitly programmed. ML techniques have been applied widely since the 1980s due to the fact that ML algorithms evolved as they were exposed to more and more data (58). Deep learning (DL) is a technique under the ML umbrella and AI family that has been applied since the 2010s due to the advancement of computational resources (Figure 17). (58).

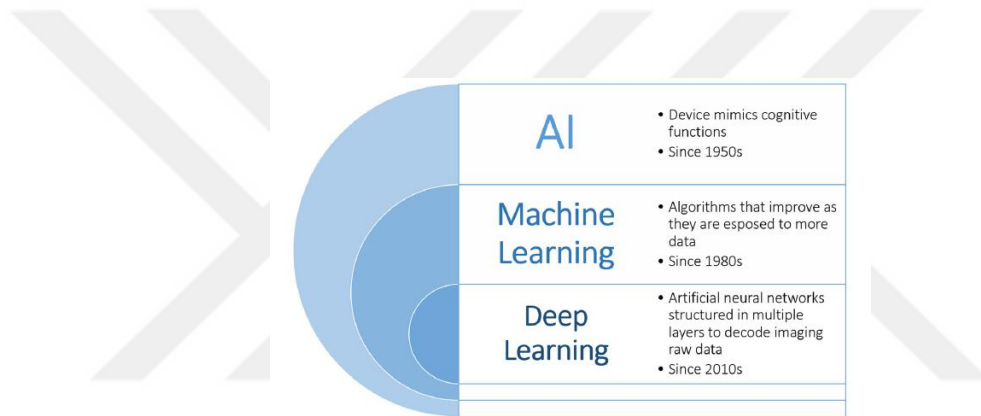


Figure 17. A short history of AI family (58).

Also, milestones of ML in its history is given in Figure 18. (59).

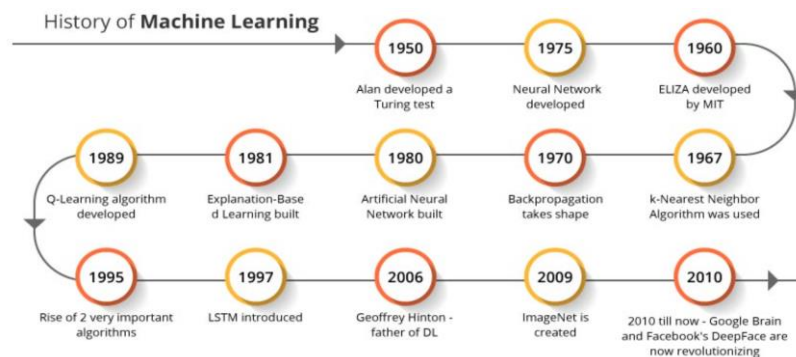


Figure 18. Milestones of ML in its history (59).

2.2.2 Classification of Machine Learning

In ML, there are five types of learning methods used for solving different tasks. They are (1) supervised learning, (2) unsupervised learning and (3) Ensemble learning, (4) reinforcement learning, (5) deep learning (54, 58). All types of machine learning methods and their distributions are shown in Figure 19. (10).

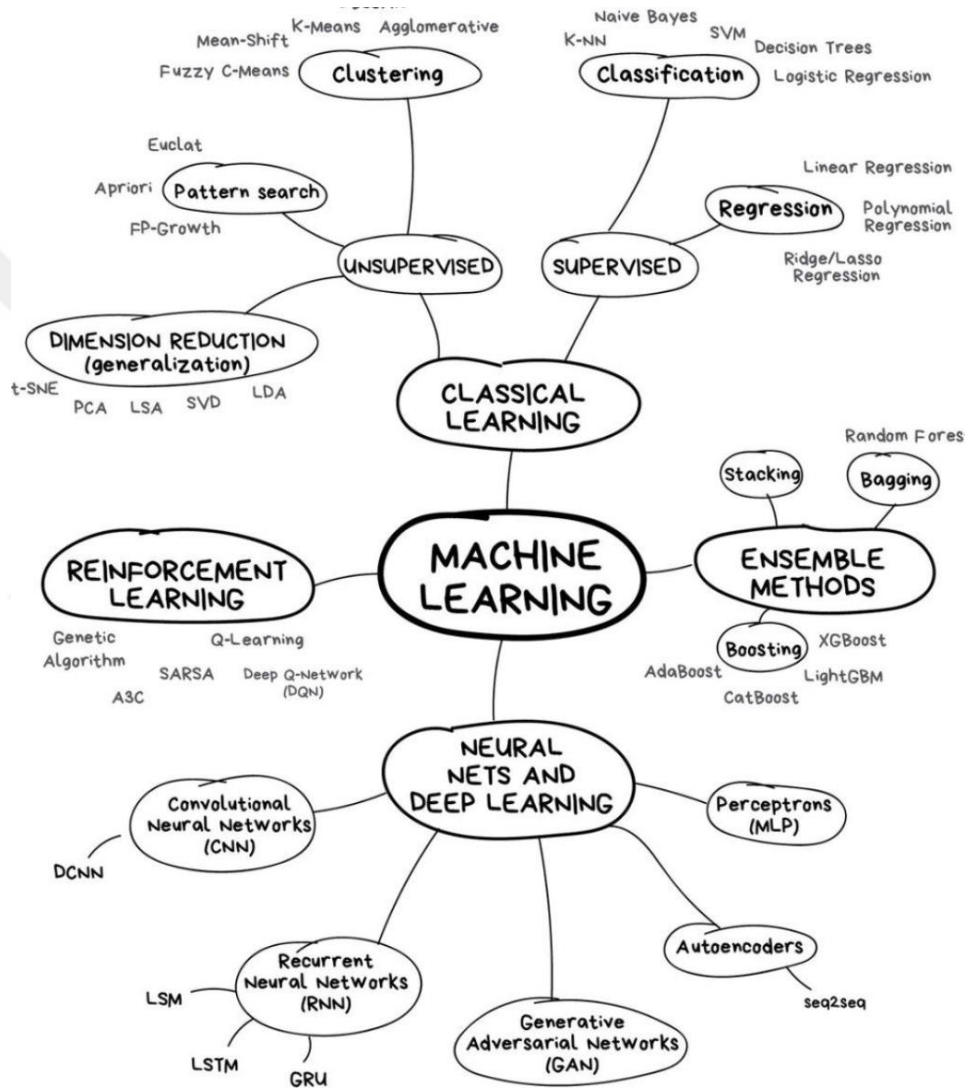


Figure 19. Machine learning and its distribution

2.2.2.1 Supervised Learning

Supervised learning is a learning method based on supervision of both input and output which maps input data into output data to develop predictive models using the algorithm $Y = f(x)$, in other notation, output = f (input). Computer learns from labelled data by supervised learning. It uses classification and regression techniques to develop ML models (53, 54). The key points of the ML model developed under supervision are a selection of features, selection of algorithm and the quality of training set (56). Supervised learning is a type of task-driven approach to predict a category or continuous numeric features. With the help of supervised learning, two types of ML models can be built. They are the classification model and regression model (10, 53).

2.2.2.1.A Classification model

The classification model predicts the different class labels in the data (10). If output variable or experimental data is categorical such as toxic-non-toxic, active-inactive, true-false, classification algorithms are used to build an ML classification model. In classification, the supervised task is to separate data into categories (10, 53). In ML, the most common and popular classification algorithms are K-Nearest Neighbours (kNN), Naïve Bayes (NB), Decision Tree (DT), Support Vector Machines (SVM), Random Forest (RF) and Logistic Regression (LR) (53). The classification model can be divided into three sub-categories which are binary classification, multiclass classification, multi-label classification (10).

- **Binary classification:** The model, if classification tasks have two class labels such as” toxic and not toxic” or “true and false” or “yes and no”
- **Multiclass classification:** The model, if classification tasks have more than two classes such as classifying a set of images of fruits; apple, orange, pear
- **Multi-label classification:** The model, if one input data belongs to more than one class, like a person who speaks two languages such as English and French (10).

2.2.2.1.B Regression model

The regression model predicts a continuous quantity (10). When output variable or experimental data is continuous numerical, regression algorithms are used to build an ML regression model (10, 53). The regression model is used in numerous fields such as prediction of drug response modelling, trend analysis, cost estimation, financial forecasting and marketing. Some of the common regression algorithms are simple or multiple linear regression, polynomial regression, LASSO regression and ridge regression (10).

2.2.2.2 Unsupervised Learning

Unsupervised learning is a learning method based on only input data in order to group and interpret the data. Computer learns from unlabeled data without the need for human intervention in unsupervised learning (10, 53). Unsupervised learning is a type of data-driven approach for discovering the data structure. It can be divided into groups such as clustering, pattern/structure recognition, dimension reduction and feature learning (10).

2.2.2.2.A Clustering

Clustering, an unsupervised learning method, is a kind of data analysis technique for identifying and grouping data according to their similarities, such in some sense more similar to each other than objects in other groups. In ML and data science, the most popular clustering algorithms are k-means clustering, mean-shift clustering, agglomerative clustering (10).

2.2.2.2.B Pattern search

Association rule learning is a rule-based ML approach in order to find the relationship between variables by using the “IF-THEN” statements. In data mining, logic dependent, frequent-pattern based and tree-based are some of the associations of the rule-based ML methods. The most popular association rule learning algorithms are Apriori, ECLAT and FP-Growth (10).

2.2.2.2.C Dimension reduction

The dimension reduction is an unsupervised-learning technique for better interpretations, avoiding overfitting and lowering computational cost in order to simplify the model by reduction of dimension. “Feature selection” and “feature extraction” can be used as dimension reduction techniques. While feature selection keeps a subset of the original features, feature extraction creates brand new ones (10).

- **Feature selection:** it is also known as attribute selection or variable selection for a dataset. It is the process of choosing the optimal features (variables or predictors) for building a ML model. It reduces the complexity of the model by eliminating the irrelevant features and minimizing the overfitting problem by simplifying and generalizing the model. Also, it increases the accuracy of the model. Therefore, feature selection is considered as one of the important points that affect the effectiveness and efficiency of the ML model (10). Some of the popular feature selection techniques are Pearson's correlation coefficient, Chi-squared test, analysis of variance (ANOVA) test and recursive feature elimination (RFE) (10).
- **Feature extraction:** It reduces the number of features in a dataset by producing new ones from the existing ones and then removing the original features. It provides a better understanding of the data, increases the prediction accuracy of the model, reduces the training time and computational cost. Principal components analysis (PCA) is the popular feature extraction technique to extract a lower-dimensional space creating new brand components from the available features in a dataset (10).

2.2.2.3 Reinforcement Learning

Reinforcement learning (RL) is an ML technique based on interacting with the environment. The four elements of RL are agent, environment, rewards and policy. (Figure 20). In RL, an agent is trained on a reward and punishment mechanism. In other words, in an interactive environment, the agent is learned by trial and error from its actions and experiences using as inputs. In RL, the problem is solved using a Markov Decision Process (MDP) that is related to sequentially making decisions. There are two approaches in RL; model-based and model-free techniques. The most important difference between these approaches is the policy network which is necessary for model-based RL. The popular RL algorithms are SARSA (State–Action–Reward–State–Action), Q-learning, Deep Q Network. In RL, agents such as robots or computers learn and make a decision for playing chess and controlling their motion. The application area of RL is robot motion control, aircraft control, driverless car, supply chain logistics, control theory, game theory, manufacturing, multi-agent systems (10).

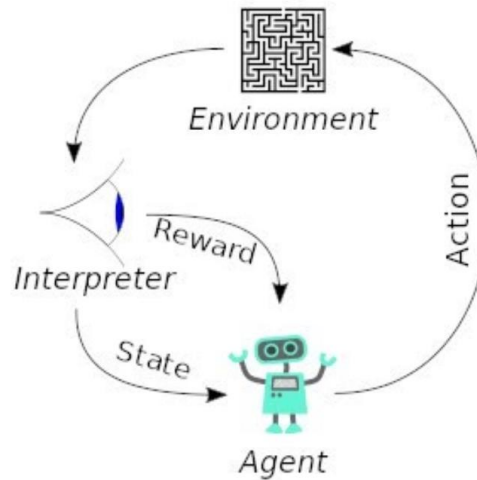


Figure 20. Learning diagram for reinforcement learning

2.2.2.4 Ensemble Learning

In ML, it is a kind of hybrid learning model by combining multiple models to obtain better predictive performance (10, 60). The ensemble learning model is divided into three categories. They are;

- **Boosting:** It is a sequential ensemble learning method used for reducing bias. (Figure 21). Adaboost and XGBoost are the popular boosting algorithms (10, 61)
- **Bagging:** It is a parallel ensemble learning method used for decreasing the variance (Figure 21). Also, bagging is known as bootstrap aggregation. It is applied to a high-variance ML algorithm, typically decision trees. Random Forest is one of the most popular bagging ensemble learning algorithms (10, 61).
- **Stacking:** It is heterogeneous learning that combines parallel and meta-learner (Figure 22). The meta-learner is an algorithm that combines the base classifier. (61). Also, it is a kind of integration strategy that integrates low level-learner and high level-learner such as meta-classifier (62).

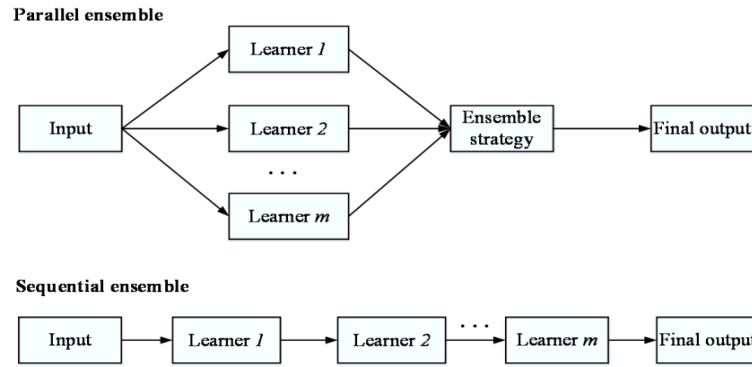


Figure 21. Flowchart of parallel and sequential ensemble learning (62).

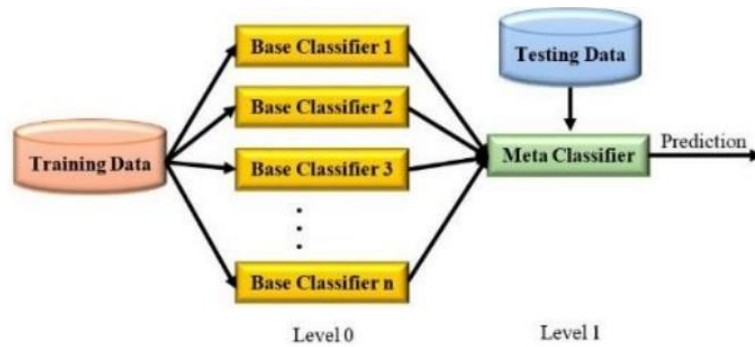


Figure 22. Flowchart of stacking ensemble learning (63).

2.2.2.5 Deep Learning and Neural Networks

Deep Learning (DL) is based on artificial neural network (ANN) models inspired by brain modeling studies. ANN is an ML algorithm inspired by brain-like neural networks with a brain-like logical structure used in AI models (54, 57). Each ANN contains nodes (analogous to cell bodies) that communicate with other nodes via connections (analogous to axons and dendrites). Much in the way synapses between neurons are strengthened when their neurons have correlated outputs in a biological neural network, connections between nodes in an ANN are weighted based on their ability to provide an appropriate outcome (54). When classic ML compared with DL approaches applied to a classification task (Figure 23). Both of them use an ANN organised in different layers (IL :input layer, HL: hidden layer, OL: output layer). The DL approach avoid the design of special use of feature extractor by using a deep neural network that represents complex features as a composition of simpler ones (58).

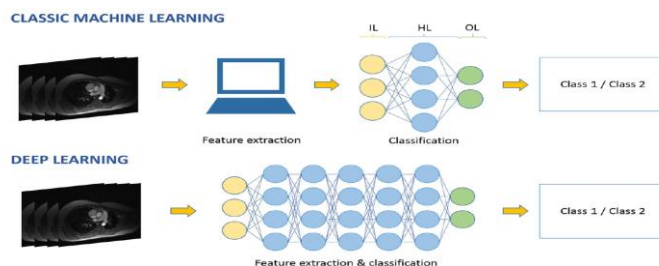


Figure 23. The comparison of classic ML and DL (58)

2.2.3 Feature Selection

In ML, the feature selection process is applied to find the best descriptor subset in order to build a predictive QSAR model (30). There are three kinds of feature selection approaches. They are filter, wrapper and embedded approaches. The advantages and disadvantages of these feature selection approaches are given in Figure 24 (38). Furthermore, feature selection is known as attribute selection, variable selection, or variable subset selection in ML.

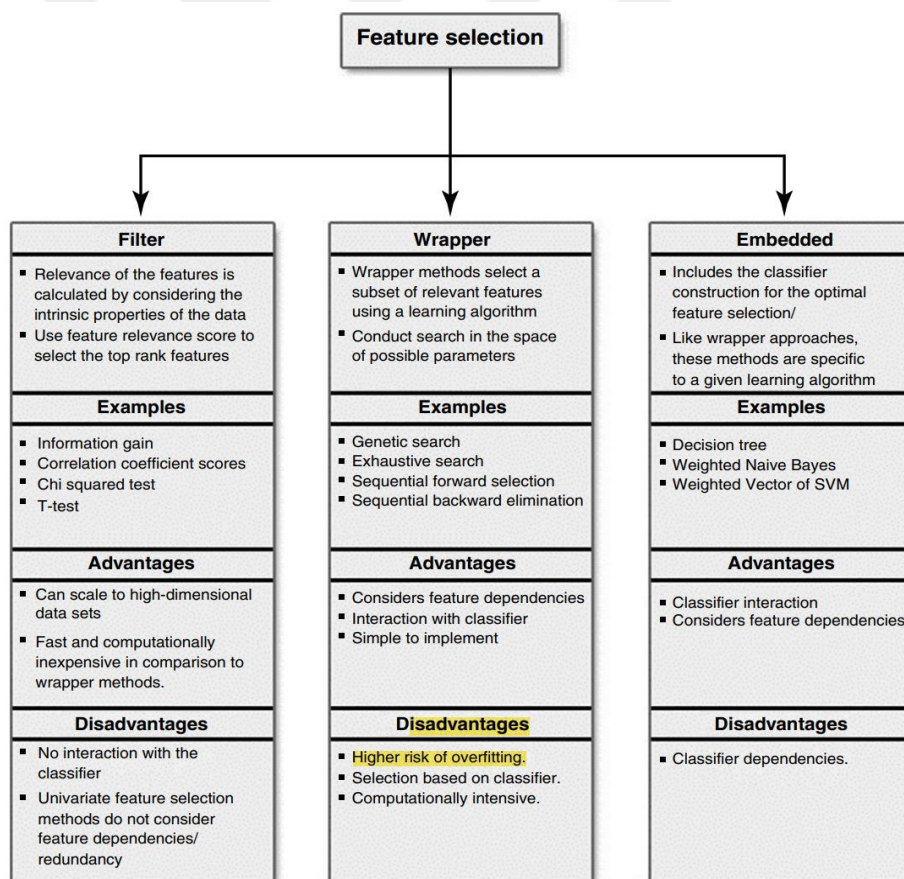


Figure 24. Flow diagram of features selection methods with examples and their advantages and disadvantages (38).

2.2.4 Machine Learning Models

The aim of ML models is to find out the patterns from the input data and generalize these input data and then apply the new data to the trained ML model to predict (8). ML models can be used for drug discovery and predictive toxicology. For example, in vivo toxicological effects of drugs, cosmetics, foods, consumer products, and agrochemicals can be predicted in silico via ML models (8).

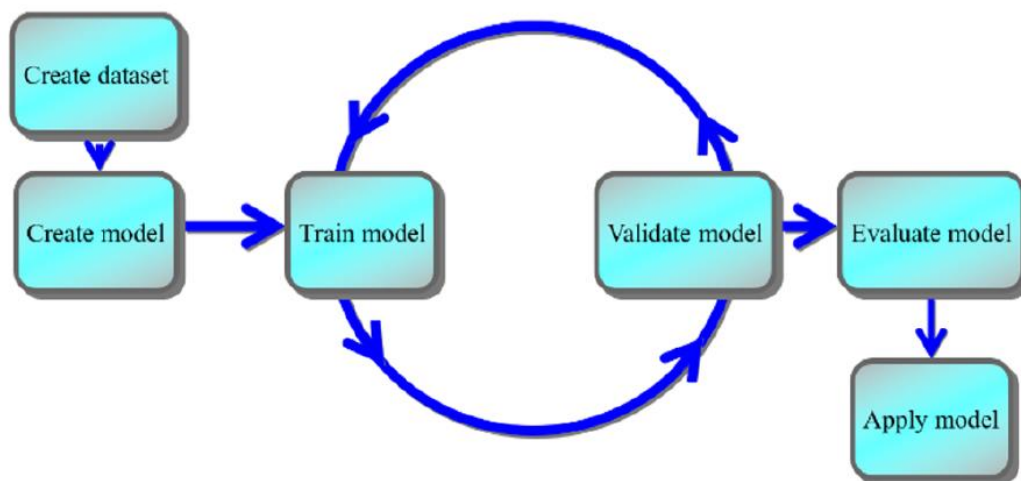


Figure 25. Outline of machine learning (8).

Machine learning contains several steps that are shown in Figure 25. These steps are collecting a dataset, building a model, training the model, performing the model validation, repeating training and the model validation until an appropriate predictive model is obtained, and then finally testing the model on new data (8).

In data science, it is important to find whether the data is linear or non-linear in order to select the proper algorithm to train a good model. For a classification model (to predict class), drawing a 2-dimensional scatter plot can be used to determine the relationship between input and output data. For a regression model (to predict numerical value), drawing a 2-dimensional scatter plot and then applying simple linear regression (with the least square error method) can be used to determine the relationship between input and output data (64). An ML model learns and recognizes the patterns from the input dataset and then makes decisions for the new dataset (7). The ML model can be linear or non-linear, it may depend on data and selected ML algorithms.

2.2.4.1 Linear Model

A linear model describes linear relationships between input data (n independent variables) and output data (dependent variable) and it makes a prediction by using the linear function of input features. In classification, linear data means linearly separable data. (Figure 26). If data are not linearly separable, a linear classification can not completely distinguish two classes (64, 65). For instance, logistic regression (Figure 28) is a linear classification ML algorithm that can be used to build a linear classification model (66).

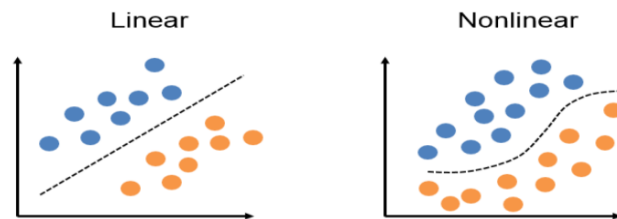


Figure 26. . Linearly separable or not separable data

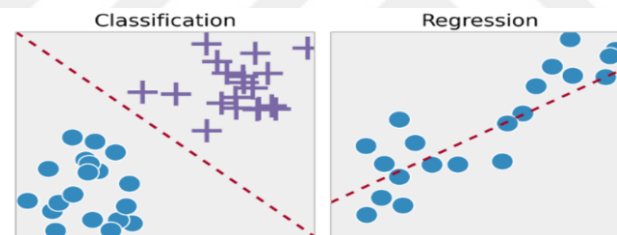


Figure 27. Linear classification and regression

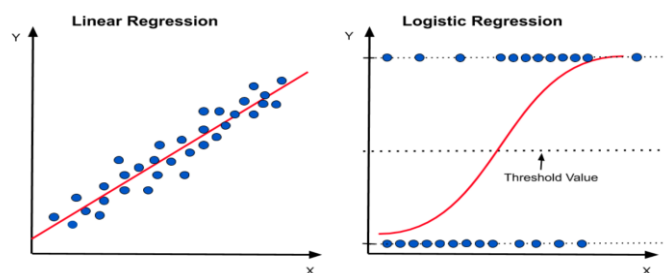


Figure 28. Linear regression versus logistic regression

In regression (Figure 27), for predicting the numerical value, firstly, data is visualized by using a scatter plot and then simple linear regression is applied to the dataset, afterward, the least squared error is checked. The dataset is evaluated how fits the regression model by calculating the R-squared value. If this value is closer to 1, the dataset can be considered linear (65).

2.2.4.2 Non-Linear Model

A non-linear model describes non-linear relationships between input data and output data and it makes a prediction by using the non-linear function of input features (70). In classification, a non-linear dataset means a linearly non-separable dataset (64). The non-linear classifier can be used to separate data (instances) when the data are not linearly separable. For example, kNN or SVM is a non-linear classification algorithm that can be used to build a non-linear classification model (66).

While non-linear classifiers that are kNN and kernel, make predictions using the concept of distance or similarity to training instances, nonlinear classifiers that are decision trees and neural networks (multi-layer perception) make predictions based on features instead of distance or similarity (65).

The brain is a complex, non-linear and parallel computer. It has the capability of pattern recognition, perception, learning and memorizing and generalizing and motor control much faster than any computer and many others. Therefore, an artificial neural network (ANN), a non-linear parallel ML algorithm, is inspired by the structure of brain-like neural networks (57).

In recent years, ML modeling methods provide various novel non-linear algorithms for both classification and regression models. They represent alternative tools to traditional linear techniques such as well-known Partial Least Squares (PLS) analysis (30). Furthermore, non-linear feature selection techniques such as decision trees, support vector machines, and ANN are powerful methods in order to accurately describe the relationship existing between the structural features and desired biological activity or property (30).

2.2.5 Machine Learning Algorithm

An algorithm is the set of rules to obtain the expected output data from the given input data. ML algorithm is learned from the training data and then makes a prediction or decision via the model built (7). ML algorithms for building linear and non-linear classification or regression models are given in Figure 29 (66).

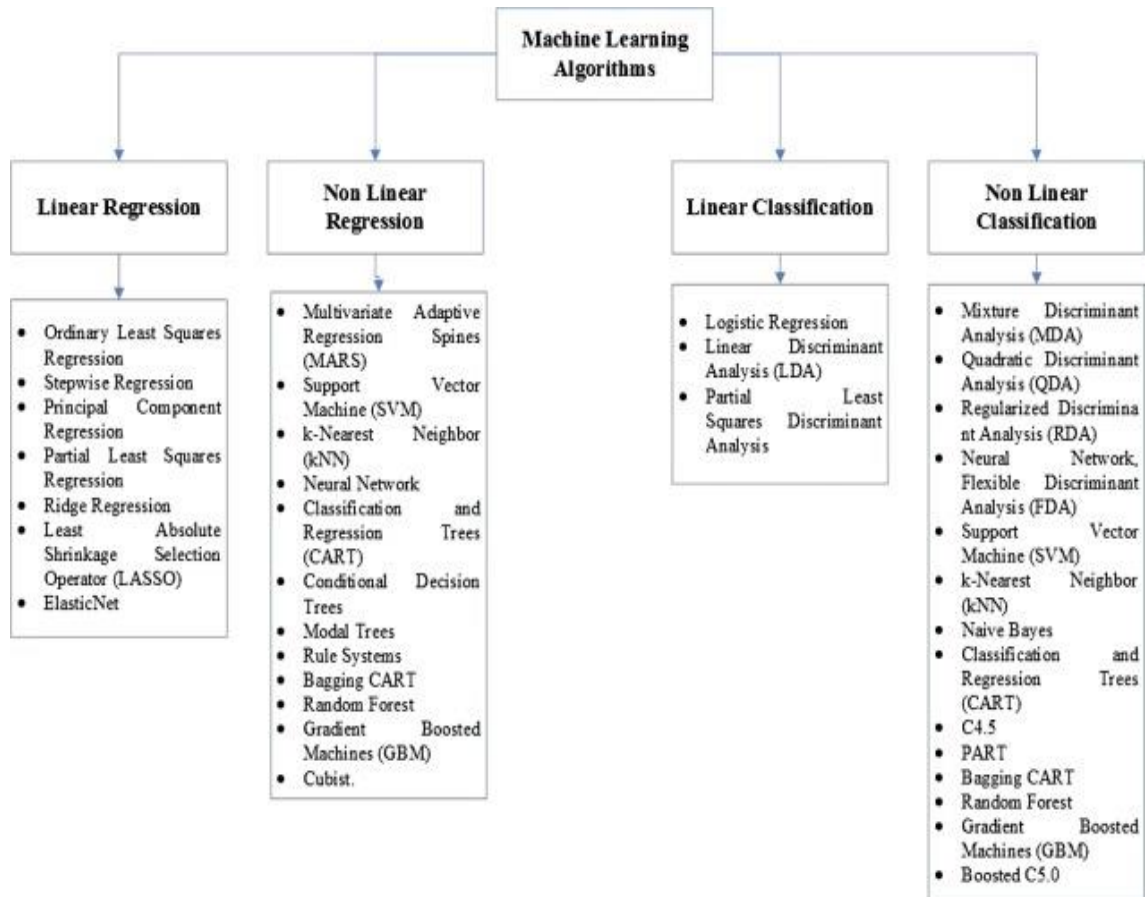


Figure 29. Linear and non-linear ML algorithms (66).

A classification algorithm, known as a classifier, is a machine learning algorithm that categorizes the input data into different classes such as active and inactive (8, 67). The most commonly used classification algorithms are decision trees, random forest, k-neighbors (k-NNs), Naïve Bayes and they are explained in the followings;

2.2.5.1 Decision Tree

A decision tree is a non-linear ML algorithm that can be used in both classification and regression models (10, 66). A decision tree can be used to visualize and support the decision-making process (68). The widely used algorithms for building a decision tree are ID3, C4.5, and CART (Classification and Regression Tree) (10). While building a decision tree, each feature is evaluated by decision tree algorithms, the most important features are selected by the algorithm in a top-down approach. For splitting the tree, entropy, information gain, gain ratio or Gini index can be used as a measure of the attribute or feature selection. The mathematical formula of them is given below (10, 69, 71). The decision tree consists of a root node, decision nodes, and leaf nodes as shown in Figure 30. A **root node** is the topmost decision on the tree where the first-time is split based on the best predictor (attribute) of the training set (10, 69).

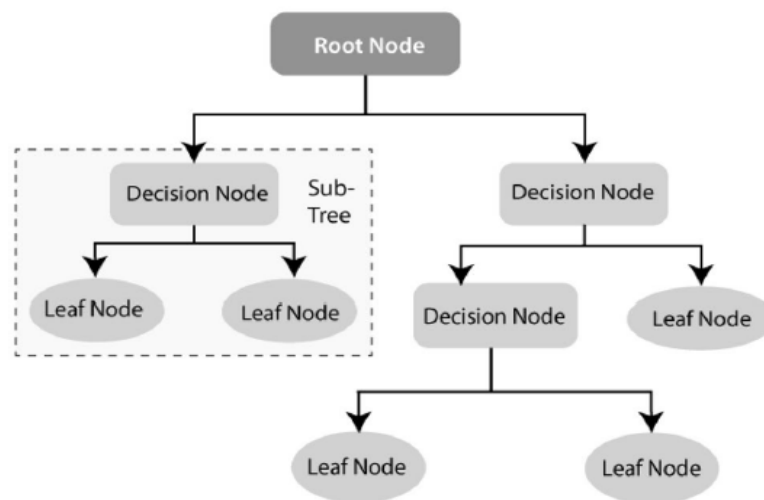


Figure 30. Structure of a decision tree (10).

The decision nodes are the intermediate steps of the tree and they are used to split the tree based on the different values of the features. **The leaf nodes** are the stopping criteria or endpoints of the tree and they represent the prediction of the model (69). The decision tree classifies the instances by controlling the attribute or feature defined by each node which is from root to leaf nodes corresponding to the attribute value (10). The training data is used for choosing the most suitable features in the decision trees. Generally, a decision tree is built from a top to bottom approach, where the feature having maximum information gain that selected first in the classification (71). The entropy of a

dataset is used to measure the impurity of a dataset or uncertainty present and it is used in the calculations. If there is a zero entropy, it means zero impurity and there are one class and totally pure (10, 69, 72). In order to select the splitting attributes, entropy and information gain are used. When training data is a subset $S = s_1, s_2 \dots$ of already classified samples. Each sample $S_i = x_1, x_2 \dots$ is a vector where $x_1, x_2 \dots$ represent attributes or features of the sample. The given event probabilities $p_1, p_2, \dots, p_s$, where $p_i = 1$. (71). The formula of entropy is defined as;

$$\text{Entropy} = H(p_1, p_2, \dots, p_s) = \sum -(p_i \cdot \log_2 p_i) \quad (71)$$

$$\text{Entropy: } H(X) = -\sum_{i=1}^N p(X_i) \log_2 p(X_i) \quad (10, 72).$$

$N = \text{number of distanct class values}$

Information gain (IG) is calculated by finding the differences between the entropies of the original dataset and the weighted sum of the entropies from entropies for each of the subdivided dataset (71). The mathematical formula of IG and also IG of a feature was given in below (69, 71).

$$\text{Information Gain (feature}_d\text{)} = \text{Entropy (Dataset)} - \text{Entropy (feature}_d\text{)} \quad (69).$$

$$\text{IG (D, S)} = H(D) - \sum P(D_i) \cdot H(D_i) \quad (71)$$

Some of the examples for the application of decision tree are user behavior analysis and cybersecurity analytics (10).

2.2.5.2 Random Forest

Random Forest (RF) is an ensemble learning method which composed of a large number of decision trees. RF classifier is a well-known non-linear, ensemble learning algorithm that can be used in both classification and regression models (10, 66). RF uses the “parallel ensemble learning” which connects multiple decision trees (or multiple learners) in parallel, in the training process as displayed in Figure 31. It uses the majority voting or averages of the final results. In this way, it reduces the overfitting problem and increases the prediction accuracy when compared to a single decision tree. RF integrates the random feature selection and bagging (bootstrap aggregation), in order to build a set of decision trees with a controlled variation. RF has been used in many application areas in ML and data science (10, 62).

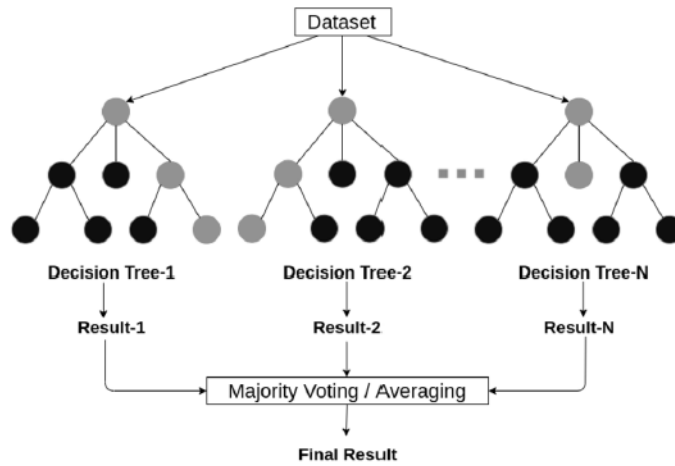


Figure 31. An example of the structure of a random forest

2.2.5.3 Naive Bayes

The Naive Bayes (NB) classifier is based on the Bayes theorem as given below with the hypothesis of freedom between each couple of features (73). It is found the probability of A (hypothesis) happening, given that B (evidence) has occurred. It is used for binary and multi-class classification in various field such as text classification, spam filtering and etc. The small amount of data is enough to predict the necessary parameters (10).

$$\text{Bayes Theorem: } P(A|B) = \frac{P(B|A) P(A)}{P(B)} \quad (73).$$

2.2.5.4 K-Nearest Neighbors

K-Nearest Neighbors (KNN) classifier is a non-linear, lazy learning or instance-based learning algorithm. It can be used for both classification and regression models. It uses data and classifies new data points based on similarity measures such as Euclidean distance. Classification is calculated from the majority vote of the k nearest neighbors of each point. The most important thing for KNN is to choose the optimal number of neighbors (10).

2.2.6 Evaluation Metrics of Classification Models in ML

Evaluation of the performance of the classification model is an important task in machine learning. There are several metrics in order to evaluate the performance of a classification model. They are confusion matrix, classification accuracy (ACC), error rate (ER), sensitivity, specificity, Matthews' correlation coefficient, F score, positive predictive value, negative predictive value, false-positive rate, area under the ROC curve, area under the precision-recall curve, cohen's kappa (67, 74).

2.2.6.1 Confusion Matrix

In classification, a confusion matrix is used to summarize the prediction results of the classification model in order to evaluate the performance of the model. It can be applied to both binary and multi-class classification. The 2X2 table of a confusion matrix for binary classification is shown in Table 4 (74).

Table 4. 2X2 table of a confusion matrix for binary classification (74).

		Predicted value	
		A (positive class)	B (negative class)
Actual value	A (positive class)	True positive (TP)	False negative (FN)
	B (negative class)	False positive (FP)	True negative (TN)

The key terms of the confusion matrix are given in the following;

- TP (True positive): the number of correctly predicted positive class
- TN (True negative): the number of correctly predicted negative class
- FP (False positive): the number of incorrectly predicted as positive but actually negative.
- FN (False negative): the number of incorrectly predicted as negative but actually positive.

All evaluation metrics are obtained from the confusion matrix by calculating the formula for each metric.

2.2.6.2 Classification Accuracy

Classification accuracy (ACC) represents the ratio between the number of correct predictions and the number of all predictions in the dataset (74).

$$ACC (\%) = \frac{TP+TN}{TP+TN+FP+FN}$$

2.2.6.3 Error Rate

Error rate (ER), known as classification error rate, is the percentage of incorrectly predicted classifications (67).

$$ER (\%) = 1 - ACC$$

2.2.6.4 Sensitivity

Sensitivity (SEN) measures the ratio of correctly predicted compounds belonging to the positive class. Recall (REC), sensitivity (SEN) and True Positive Rate (TPR) are the same and they have the same formula as follows (74).

$$SEN = REC = TPR = \frac{TP}{(TP + FN)}$$

2.2.6.5 Specificity

Specificity (SPC) measures the ratio of correctly predicted compounds belonging to the negative class. Specificity (SPC) and True Negative Rate (TNR) are the same and they have the same formula as follows (74).

$$SPC = \frac{TN}{(TN + FP)}$$

2.2.6.6 Matthews' correlation coefficient

Matthews' correlation coefficient (MCC) is calculated from all four values of the confusion matrix by considering data imbalance. It is regarded to be one of the best metrics for binary classification performance. A value of 1 indicates the perfect prediction (67, 74, 75).

$$MCC = \frac{TP.TN-FP.FN}{\sqrt{(TP+FP).(TP+FN).(TN+FP).(TN+FN)}}$$

2.2.2.7 F score

F score is the harmonic mean of precision and recall. The worst value: 0 and the best value:1). (67).

$$F1 \text{ score} = \frac{2.TP}{2.TP+FP+FN} = 2. \frac{precision.recall}{precision+recall}$$

2.2.6. 8 Positive Predictive Value

A positive predictive value (PPV) is the ratio of true positive predictions among all positive predictions. Positive predictive value (PPV) and Precision (PREC) are the same and they have the same formula as follows (67, 74).

$$PPV = PREC = \frac{TP}{(TP + FP)}$$

2.2.6.9 Negative Predictive Value

A negative predictive value (NPV) is the ratio of true negative predictions among all negative predictions (74).

$$NPV = \frac{TN}{(TN + FN)}$$

2.2.6.10 False Positive Rate

The False Positive Rate (FPR) is equal to 1 minus Specificity (SPC) (67). False positive rate = 1-specificity

$$FPR = 1 - SPC = \frac{FP}{(TN + FP)}$$

2.2.6.11 Area under the ROC curve

The area under the ROC curve (AUC) (Figure 32) measures the quality of the classifier and high AUC values are considered an efficient model. The AUC value ranges between 0 and 1. Also, it is used for classifier comparison. (TPR =SEN, FPR=1-SPC) (76).

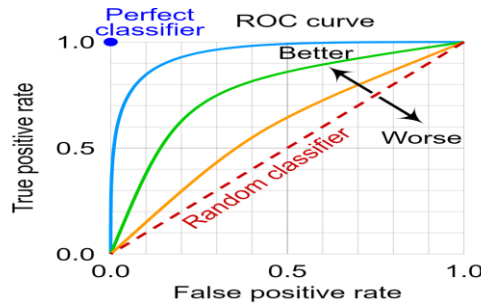


Figure 32. The area under the ROC curve

2.2.6.12 Area under precision-recall curve

The area under the Precision-Recall curve (AU-PRC) (Figure 34) is more informative than AUC with imbalanced datasets (67).

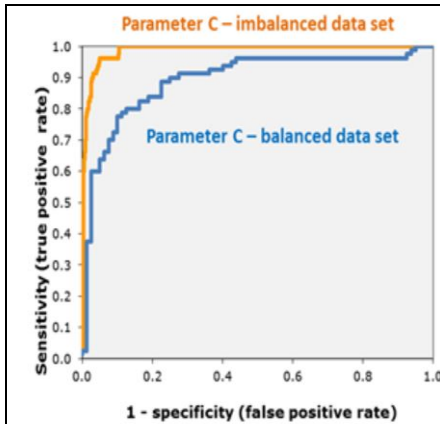


Figure 33. ROC curve

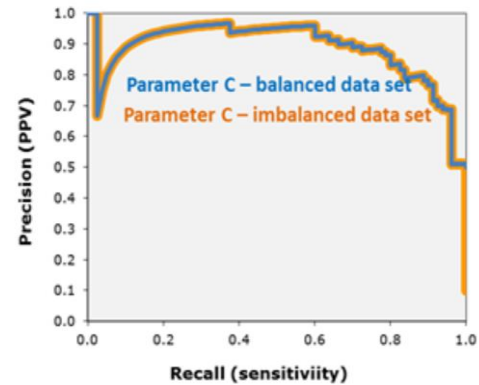


Figure 34. Precision-Recall curve

2.2.6.13 Kappa

Kappa (Cohen's kappa), measures the agreement of prediction with the true class – 1.0 signifies perfect agreement The scores for kappa is given in Table 5 (77).

Table 5. The scores for kappa

Kappa statistics	Strength of agreement
<0.00	poor
0.00-0.20	slight
0.21-0.40	fair
0.41-0.60	moderate
0.61-0.80	substantial
0.81-1.00	almost perfect

2.2.7 Application Areas of Machine Learning

Machine learning has become very popular in our daily lives and used in various applications due to its learning ability from the past and making intelligent decisions. The most popular area of ML technology in the following (Figure 35) (10).



Figure 35. The popular applications of machine learning

a. Predictive analytics and intelligent decision-making :

The major application area of ML is intelligent decision-making by using data-driven predictive analytics. It is based on capturing and exploiting the relationship between explanatory and predicted variables from past events in order to predict the unknown outcome. For making decisions, the accurate predictions gave insight to industries, businesses, and various organizations including governments, financial agencies, e-commerce companies, sales and marketing, transportation, social networking, and telecommunications. For example; identifying criminals or detecting credit card fraud (10).

b. Image, speech, and pattern recognition:

Image recognition: Medical (labelling as cancerous cell or not in x-ray), Social media (Facebook; face detection and tagging suggestions)

Speech recognition: Google online assistant, Siri, Alexa, Cortana (typically uses sound and linguistic model by using ML methods)

Pattern recognition: describe as automated recognition of patterns and regularities in the data. It is used in image analysis, classification (Gmail; spam mail detection; spam or not), clustering, feature selection, sequence labelling (10)

c. E-commerce and product recommendations:

The most well-known application of ML is product recommendation. It is used in analyzing the consumer's purchasing histories and making a customized suggestions for their next purchase based on their behaviour and preferences. E-commerce companies, online retailers such as Amazon easily manage inventory, prevent out of

stock situations, optimize logistics and warehouses by using predictive modelling based on ML techniques (10).

d. Healthcare:

ML can help to solve diagnostic and prognostic problems, such as disease prediction, medical knowledge extraction, patient management and detecting regularities in patient data (10).

e. Cybersecurity and threat intelligence:

In order to protect networks, hardware, systems and data from digital attack. For example, the clustering method of ML can be used to determine the cyber-abnormalities and policy violations (10).

f. Traffic prediction and transportation:

For smart cities, traffic and transportation prediction is important. It is based on the travel history by the means of various routes.

g. Internet of things (IoT) and smart cities:

Internet of Things (IoT) is the area of Industry 4.0, which turns every object into smart objects with a connection to the internet by allowing them to transmit data and automated tasks without any need for human interaction. The application area of ML is for smart cities, smart governance, smart home and many other fields such as in education, communication, transportation, agriculture, healthcare, retail (10).

h. Sustainable agriculture:

The sustainable agriculture supply chain is important to increase agricultural productivity by reducing the negative effects of the environment. It is based on information and technologies where knowledge is transferred from various phases of sustainable agriculture, such as pre-production, production, in processing and distribution. The predictions in each phase are summarized in the following (10);

- **pre-production phase:** the prediction of crop yield, soil properties, irrigation etc),
- **production phase:** weather prediction, disease and weed detection, soil nutrient management)
- **in processing phase:** for demand estimation, production planning, etc.
- **distribution phase:** the inventory management, consumer analysis, etc.

2.3 Comedogenicity

2.3.1 Comedogenicity and *Acne cosmetica*

While *comedogenicity* is a common adverse reaction to cosmetics that induces comedones or clogged pores (13, 15, 16). *Acne cosmetica* is caused by cosmetic products containing comedogenic substances (17). It was characterized as a recognizable acneiform eruption consisting mainly of closed comedones with occasional papule-pustules on the parts of the face such as the forehead, temples, and cheeks as shown in Figure 36 and Figure 37. It was first reported by French dermatologists in the mid-forties. They reported that hair pomade and brilliantine caused flareups on the upper face regions containing the forehead and temple. These flareups were attributed to impurities in cosmetic products (15). Plewig et al .,(1970) reported “*pomade acne*” caused by the long-term use of pomade (scalp creams and oils) (78). Kligman and Mill (1972) reported half of 25 facial cosmetic creams tested in their study were found to be mildly comedogenic when tested in rabbit ears. Common ingredients of these cosmetics, namely, butyl stearate, lauryl alcohol, oleic acid, lanolin, and petrolatum were also comedogenic (16).



Figure 36. *Acne cosmetica*



Figure 37. Closed comedones

2.3.2 Mechanism of Comedo Formation

“Comedo” or “microcomedo” is the initial lesion of acne. It is produced by the hyper-proliferation of keratinocytes (known as hyperkeratinization) and sebaceous gland hyperplasia with seborrhea (known as hypersecretion of sebum) in hair follicles. As a result, sebum, fatty acids, and keratin accumulate in the hair follicles (pilosebaceous unit) and cause the formation of comedones as shown in Figure 38. (79). “Comedo” is a non-inflammatory acne lesion. There are two types of comedo which are closed comedo (blackheads) and open comedo (whiteheads). If comedo becomes inflamed by *P.acnes* bacteria, inflammatory acne lesions such as papule, pustule or nodule occurs. (79).

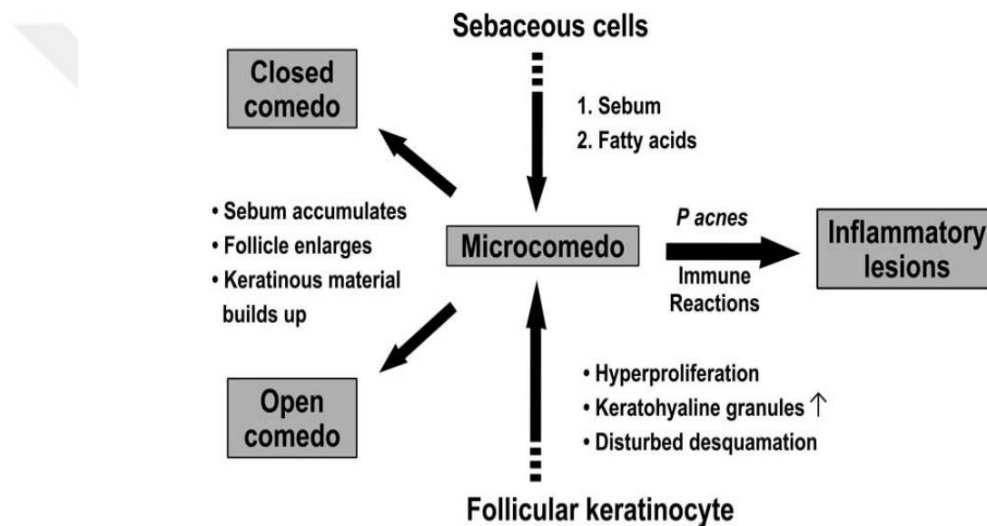


Figure 38. Initial lesion of acne (microcomedo) and progression of a lesion in acne (79).

2.3.3 Pathogenesis of Acne

Acne is a skin problem affecting 80% of adolescents and young adults from 11 to 30 years old. The target organ in acne is the pilosebaceous unit that composed of a hair follicle with a sebaceous gland (Figure 40). The distribution of acne is primarily on the face, chest and back which areas with the greatest number of pilosebaceous glands (79).

Acne lesions begin with microcomedo that is not visible to the naked eye (Figure 38). When sebum and keratin accumulate and follicle enlarges, “microcomedo” turns into “comedo” that is visible to the naked eye. The sebaceous glands start to become bigger with androgenic stimulus at 7 to 8 years of age, with the result of increased sebum production. Androgens drive changes in both sebocytes and follicular keratinocytes that

lead to comedo which is a non-inflammatory acne lesion. Sebum secretion increases with the influence of androgens, the major sebotropic hormones (79). The degree of comedonal acne was found positively correlated with the level of adrenal androgen dehydroepiandrosterone sulfate (DHEA-S) in the prepubertal girls (14). On the other hand, comedogenesis occurs when abnormal follicular keratinization (due to increased keratinocytes proliferation) accumulates in the sebaceous follicle. The four main factors given in the following that influence the development of acne are summarized in Figure 39. (79).

- 1) increased sebum production due to the sebaceous gland hyperplasia with seborrhea
- 2) follicular hyperkeratinization (comedogenesis)
- 3) abnormal growth of *Propionibacterium acnes* in the follicle
- 4) inflammation and immune response (results in inflammatory acne lesion)

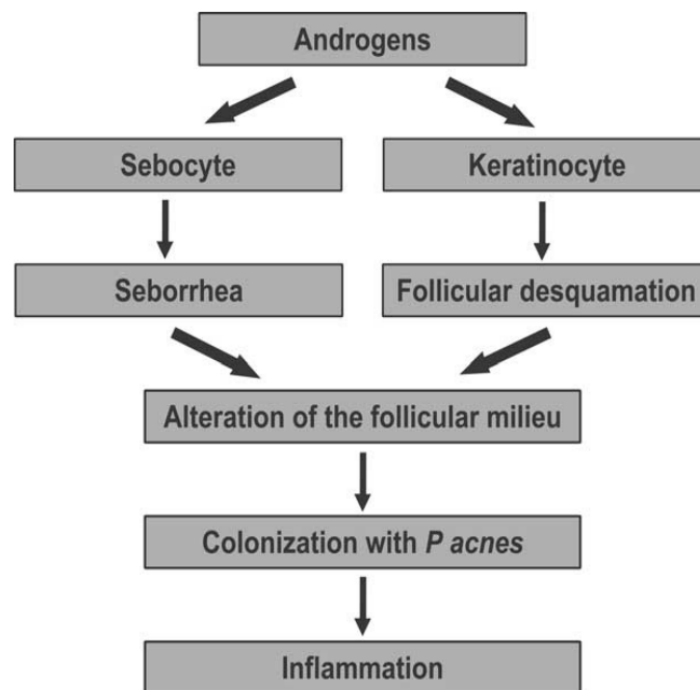


Figure 39. Pathogenesis of acne (79).

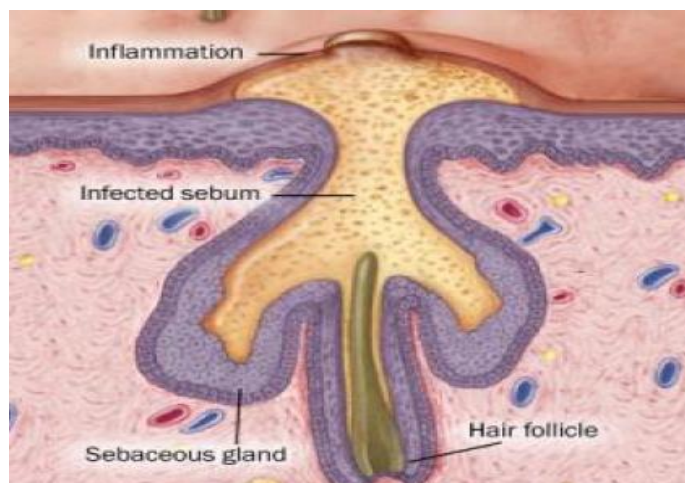


Figure 40. The target organ in acne is the pilosebaceous unit composed of a hair follicle with a sebaceous gland (14).

2.3.4 Acne and Oxidative Stress

Oxidative stress and inflammation are involved in the pathogenesis of acne (80). Furthermore, various dermatological disorders are related to oxidative stress with the aberrant antioxidant system as shown in Figure 41. In skin homeostasis, the balance between the antioxidant-defense system and oxidative stress is maintained. If this redox balance in the skin is disrupted, various dermatological disorders such as acne, psoriasis, skin cancer, vitiligo, rosacea occurs (82, 83).

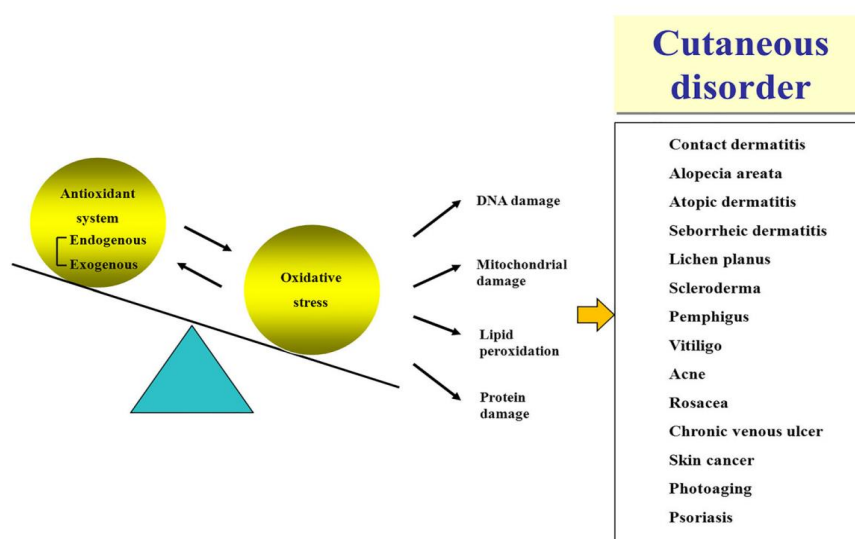


Figure 41. Various cutaneous disorders are related with oxidative stress and aberrant antioxidant system (82).

Under normal physiological conditions, reactive oxygen species (ROS) production due to environmental and cell-generated is regulated and ROS participates in both pathogen defense and cellular signaling. However, inadequate ROS detoxification or ROS overproduction generates oxidative stress, resulting in DNA damage, mitochondrial damage, lipid peroxidation, and protein damage (82, 83). If skin chronically is exposed to the ROS-induced oxidative stress that is generated from both endogenous and exogenous sources, it results in various damage in the skin (80).

The endogenous source of ROS in the skin is autooxidation and enzymatic oxidation. ROS and lipid peroxidation is blamed for the initiation and progression of acne in the pilosebaceous unit. Because lipid peroxidation of fatty acids and unsaturated triterpenes such as squalene in sebum produces both intra and extracellular ROS. These powerful oxidants have capable of comedogenic potential (80).

2.3.4.1 Oxidative stress

Oxidative stress is an imbalance between the production and detoxification of reactive oxygen reactive species (ROS) and the accumulation of ROS in cells and tissues that leads to cell and tissue damage (84). The common ROS are hydroxyl radicals ($\bullet\text{OH}$), superoxide radicals ($\text{O}_2\bullet$), singlet oxygen ($^1\text{O}_2$) and hydrogen peroxide (H_2O_2) (Table 6).

Free radicals (FR) are oxidants that can be organic or inorganic chemicals with one or more unpaired electron(s). They show by an additional point ($\bullet$) and they are highly reactive and short-lived (85). The free radical, both ROS and RNS (reactive nitrogen species) are considered as contributors to oxidative stress. They are given in Table 6 (85).

Table 6. Free radicals as contributors to oxidative stress (85).

Name	Formula	Characteristics
Hyperoxide/superoxide	$\cdot\text{O}_2^-$	Highly unstable, signaling function, synaptic plasticity
Hydrogen peroxide	H_2O_2	Cell toxicity, signaling function, generation of other ROS
Hydroxyl radical	$\cdot\text{OH}$	Free radical, highly unstable, very reactive agent
Alkoxyl radical	$\text{RO}\cdot$	Free radical, reaction product of lipids
Peroxyl radical	$\text{ROO}\cdot$	Free radical, reaction product of lipids
Hypochlorite anion	OCl^-	Reactive oxygen species, reactive chlorine species, enzymatically generated by myeloperoxidase
Singlet oxygen	$^1\text{O}_2$	Induced/excited oxygen molecule, radical and nonradical form
Ozone	O_3	Environmental toxin
Nitric oxide	$\cdot\text{NO}$	Environmental toxin, endogenous signal molecule
Peroxynitrite	ONOO^-	Highly reactive reaction intermediate of $\cdot\text{O}_2$ and $\cdot\text{NO}$
Nitrogen dioxide	$\cdot\text{NO}_2$	Highly reactive radical, environmental toxin
Nitrogen oxides	NO_x	Environmental toxins, including NO and $\cdot\text{NO}_2$, derived from the combustion process

2.3.4.2 Sources of oxidative stress

The free radicals (ROS and RNS) are produced from both endogenous sources and exogenous sources (85). ROS are normally produced as by-product during the oxygen metabolism (84). They have function physiological signaling, host defense mechanism (85) and involvement in inflammatory response (86). Some of the endogenous sources of free radicals are mitochondria, peroxisomes, endoplasmic reticulum, lipoxygenases. Some of the exogenous sources of free radicals are xenobiotics, drugs, environmental pollutants, heavy metals, transition metals, environmental toxins, pesticides, alcohol, radiation (84, 85, 87) as shown given in Figure 42 (85).

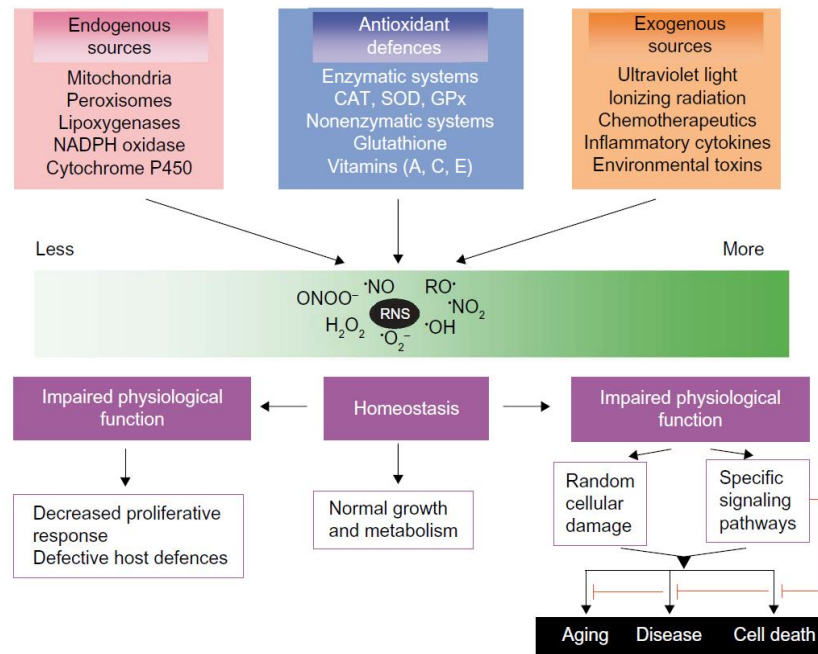


Figure 42. Sources of Free radicals (ROS and RNS) and against antioxidant defense system (85).

Free radicals give damage various biological molecules such as lipids, proteins and nucleic acids, thereby changing the normal redox status of body leading to increased oxidative stress (87). The free radical-induced oxidative stress is responsible for the pathogenesis of various diseases such as cardiovascular diseases, Diabetes Mellitus neurodegenerative disease (such as Alzheimer's Disease (AD), Multiple Sclerosis (MS), Parkinson's Disease (PD), asthma, liver damage, cancer, rheumatoid arthritis, cataract and aging (87).

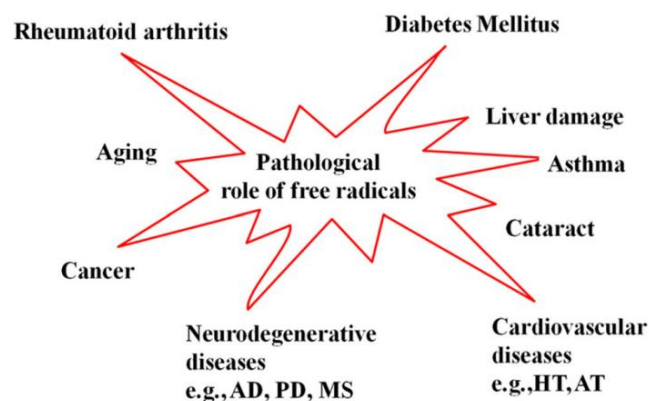


Figure 43. The pathological role of free-radicals induced-oxidative stress for various diseases (87)

2.3.5 Lipid Peroxidation and Acne

ROS and lipid peroxidation are considered to play a major role in the initiation and progression of acne (80). In acne, sebum lipid composition is changed because of the sebum oxidation via lipid peroxidation of fatty acids in the sebum. Powerful oxidants generated by lipid peroxidation are suspected promoting of hyperkeratinization and thereby comedogenesis (80, 82, 88, 89). Various studies reported that lipid peroxide (LOOH) caused by oxidation of sebum is related to the progression of acne vulgaris (88). Tochio et al., (2009) hypothesized the stages of progression model of acne as shown in Figure 44.

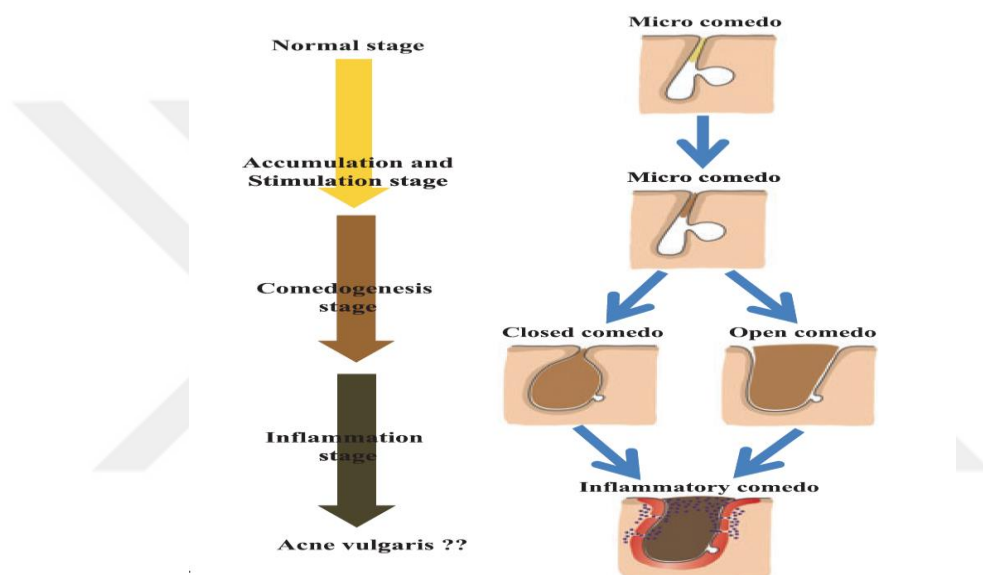


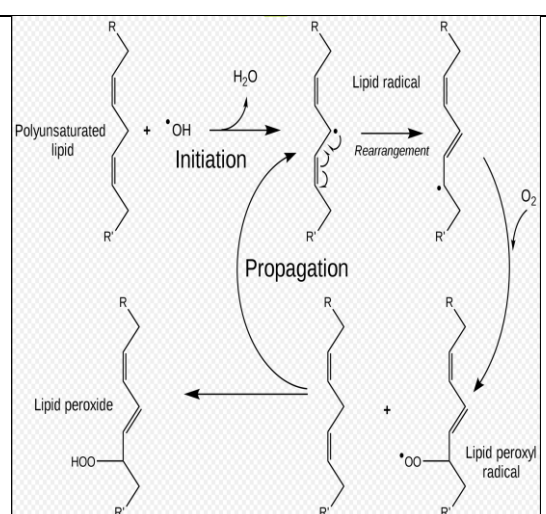
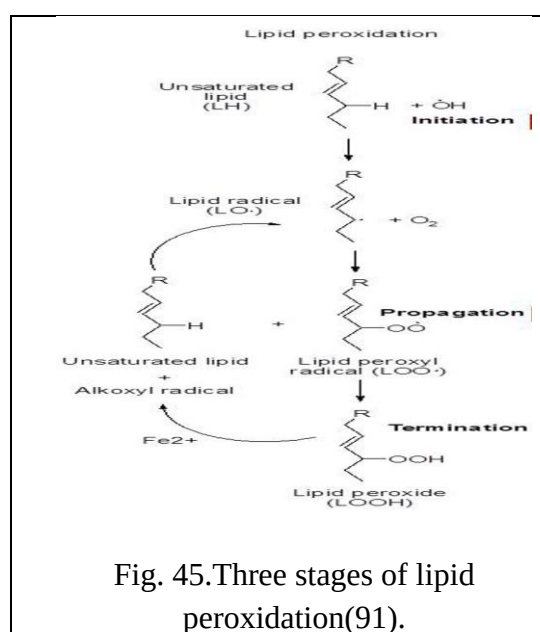
Figure 44. The model of progression of acne vulgaris (88).

- **In normal stage:** the amount of LOOH in the comedo is low.
- **Accumulation and stimulation stage:** the amount of LOOH begins to increase and induces the upregulation of IL-1 α (IL-1 α).
- **Comedogenesis stage:** proliferation and hyperkeratinization are induced by upregulation of IL-1 α . Accumulated sebum has obstructed the follicle and been oxidized.
- **Inflammation stage:** excess amount of LOOH is involved in inflammatory changes of acne.

2.3.5.1 Lipid peroxidation

Lipid peroxidation is an the attack of the oxidants (ROS or free radicals) on lipids containing *carbon-carbon double bond(s)*, particularly *polyunsaturated fatty acids* that involve hydrogen abstraction from a carbon atom. It carries on with oxygen addition and results in lipid peroxy radicals and hydroperoxides (90, 91). Lipid peroxidation is initiated mainly by $\text{OH}\cdot$ radical, goes on in the form of a chain reaction and the integrity of the membrane is irreversibly impaired (92). Lipid peroxidation occurs in cell membranes. The membrane lipids are predominantly susceptible to oxidative attacks of $\text{OH}\cdot$ due to the presence of polyunsaturated fatty acids (PUFAs) in membrane phospholipids. Polyunsaturated fatty acids in the skin can serve as excellent substrates for peroxidation because they contain active ****bis-allylic methylene groups**. The carbon-hydrogen bonds on these activated methylene parts have lower bond dissociation energies, making these hydrogen atoms more easily abstracted in radical reactions (80, 91, 92, 93).

The process of lipid peroxidation occurs in three stages; initiation, propagation, and termination (Figure 45). *The initiation stage of lipid peroxidation* begins with *hydrogen atom abstraction*. Various species can abstract the first hydrogen atom and include the radicals: hydroxyl, alkoxyl, peroxy, and possibly $\text{OH}\cdot$. Especially, polyunsaturated fatty acids (LH) in the membrane lipids are susceptible to peroxidation because abstraction from a methylene group of a hydrogen atom, which contains only one electron, leaves at the back an unpaired electron on the carbon (Figure 46).



**Footnote: the structure of Bis-allylic methylene groups: $\Sigma -\text{CH} = \text{H} - \text{CH}_2 - \text{CH} = \text{CH} -$

As shown in Figure 46, the initial reaction of $\text{OH}\cdot$ with polyunsaturated lipid (LH) generates a lipid radical ($\text{L}\cdot$) (80, 91, 92, 94).

In the *propagation stage*, lipid radical ($\text{L}\cdot$) reacts with molecular oxygen (O_2) and produces lipid peroxyl radical ($\text{LOO}\cdot$) which is unstable. Therefore, chain reactions occur, it reacts with another polyunsaturated lipid (LH) because lipid radicals induce the formation of various lipid radicals (91, 93, 94). Lipid peroxidation gives damage to lipids. As well, the formation of lipid peroxidation products leads to the spreading of free radical reactions (93).

In the **termination stage**, lipid peroxide (LOOH) (91) occurs (Figure 45) that modified the proteins as seen in Figure 47 (93). In the early stage of lipid peroxidation (during the propagation stage), lipid hydroperoxide occurs as the major primary product of lipid peroxidation (90). They are non-radical intermediates derived from unsaturated fatty acids such as phospholipids, glycolipids, cholesterol and its ester (Figure 47). The formation of lipid hydroperoxide occurs via enzymatic and non-enzymatic reactions involving ROS which are in charge of toxic effects by way of various tissue damage in the body (93).

In the late stage of lipid peroxidation, lipid peroxides break down to form aldehydes such as MDA (malondialdehyde) and 4-HNE (4-hydroxynonenal) which are the biomarker of lipid peroxidation. They can make adducts with protein and DNA such as MDA-protein adducts and MDA-DNA adducts that leading cancer (Figure 47) or 4-HNE-protein adduct leading to damage protein and induces autophagy-lysosomal pathway (dysfunctional proteins) (90, 93).

Antioxidant defense systems via enzymatic and non-enzymatic mechanisms, detoxify the ROS (91, 93, 94). Inadequate ROS detoxification can result in damage such as protein damage, DNA damage, mitochondrial damage, lipid damage due to the by-product of lipid peroxidation (Figure 48) (82). As well, antioxidant enzymes produced by the body are superoxide dismutase (SOD), catalase (CAT), glutathione peroxidase (GSH-Px), glutathione reductase (GR). They neutralize many types of free radicals (Figure 48) (93). Non-enzymatic antioxidant molecules are vitamin E (tocopherol), glutathione (GSH), vitamin A (retinol), vitamin C (ascorbic acid), and micronutrients (iron, copper, zinc, selenium, manganese) which serve as enzymatic cofactors for detoxification process (Figure 48) (93).

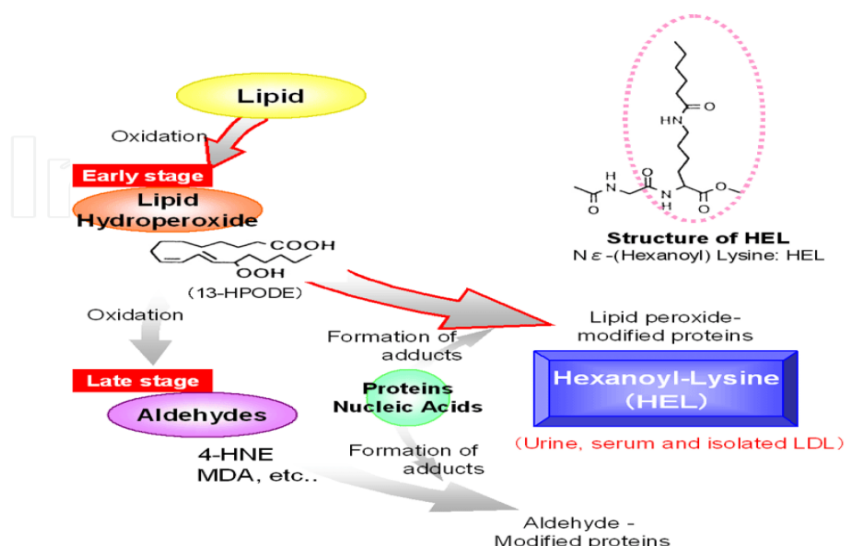


Figure 47. By product of lipid peroxidation and their results (93).

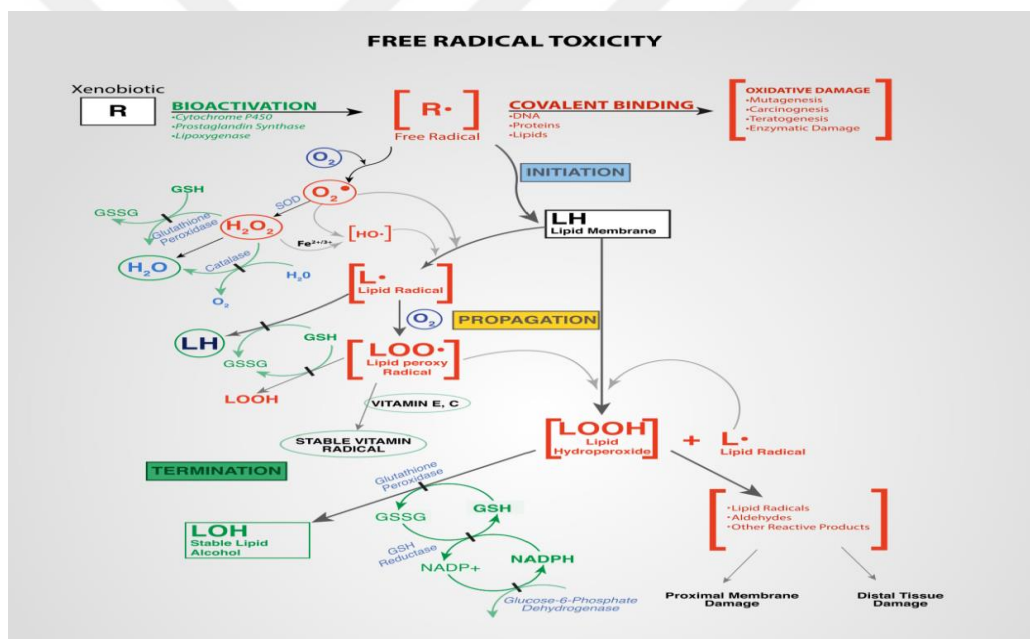


Figure 48. Xenobiotics-induced lipid peroxidation and detoxification by the antioxidant defense system.

The most important mechanism for the formation of free radicals from chemicals is the bioactivation of xenobiotics by microsomal cytochrome P-450 system. This system produces free radicals by adding an electron (reduction) to the molecules (as in carbon tetrachloride) or by removing an electron (oxidation) from the target molecule (95).

2.3.5.2 Chemicals induce lipid peroxidation

The common chemicals are known as inducing lipid peroxidation and used in vivo mammalian model for lipid peroxidation are given in the followings (90).

- a. **Tert butyl hydroperoxide:** It is an organic oxidant including a tertiary butyl group with the formula $(\text{CH}_3)_3\text{COOH}$ (Figure 49). It is a hydroperoxide used as a bleaching agent, prooxidizing and initiating agent for polymerization. It is a strong free radical source, therefore, it has been used to induce lipid peroxidation for in vivo models.

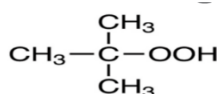


Figure 49. The structure of tert butyl hydroperoxide

- b. **Carbon tetrachloride:** It is a toxic organic compound with the formula CCl_4 . It has four chlorine atoms joined to a central carbon atom by single covalent bonds with tetrahedral configurations (Figure 50). It is non-polar due to its symmetric geometry. It is carcinogenic and used as a solvent and pesticides. It has been used in vivo model in order to induce lipid peroxidation.

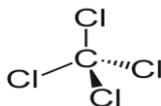


Figure 50. The structure of carbon tetrachloride

- c. **Quinolinic acid:** It is a potent neurotoxic compound, known as pyridine-2,3-dicarboxylic acid, also biosynthetic precursor to Niacin (Figure 51). It is the metabolite of the kynurenine pathway. It is normally found in the human brain and cerebrospinal fluid at the nanomolar level and is often involved in the pathogenesis of various neurological disorders. It has been used in vivo models in order to induce lipid peroxidation mediated via hydroxy radicals.

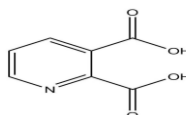


Figure 51. The structure of quinolinic acid

- d. **Transition metal ions:** They are redox-active metal and pro-oxidant impact. Therefore, they induce and initiate lipid peroxidation by generating oxygen

radicals, mainly hydroxy radical by the means of Fenton's and Haber-Weiss reactions. The redox-active transition metals that induce lipid peroxidation for in vivo models are iron, copper, chromium, nickel, manganese, cadmium and vanadium (90).

2.3.5.3 Response of cells to ROS

The possible response of cells against ROS is dependent on the level of lipid peroxidation (LPO) mediated stress and ROS-induced signal transduction. The cellular response can be occur as follow in Figure 52. 1) survival, no effect due to redox balance 2) survival, antioxidant induction at low LPO 3) apoptotic death, at moderate LPO 4) necrotic death, at high LPO (96).

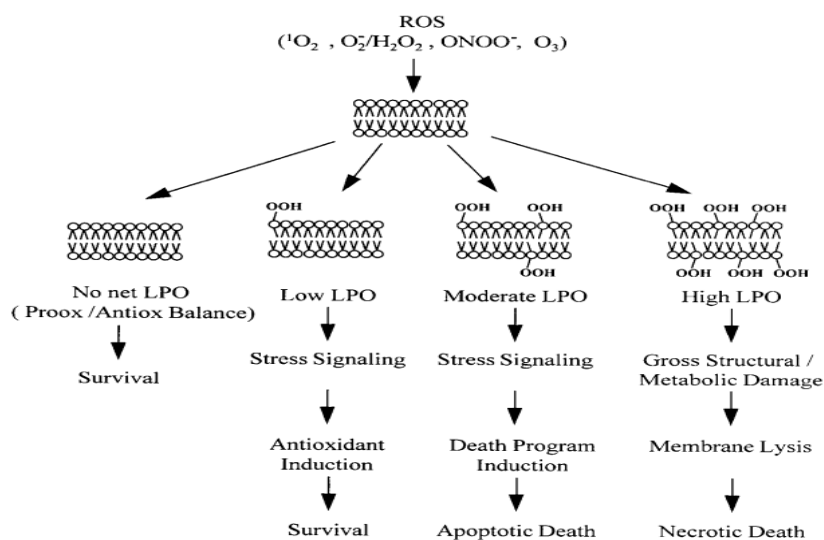


Figure 52. The possible cellular response to ROS in lipid membrane (96)

ROS have important physiological functions for the regulation of the intracellular signalling pathways. ROS can regulate the various response from growth factor stimulation to generation of inflammation (97). Oxidants (or ROS) can induce enzyme transducers such as protein kinase C (PKC), mitogen-activated protein kinases (MAPKs). Oxidant derived-species of LPO such as LOOH are early intermediates in the signaling pathway that may mimic the effects of certain growth factor (e.g. tumor necrosis factor- α) and this has stimulated them as possible second messengers (96).

2.3.5.4 Skin surface lipids

A lipid layer covers the surface of the skin. The majority of epidermal surface lipids are sebaceous origin produced by sebaceous glands called human sebum. Lipids produced by the epidermal cells (keratinocytes origin) fill the spaces between the cells, like mortar or cement. While sebum is a mixture of fatty acids, squalene, triglycerides, wax esters, cholesterol and its ester, epidermal lipids are a mixture of fatty acids, ceramides and cholesterol. The structure of lipids found in sebum and stratum corneum are shown in Figure 53 (98).

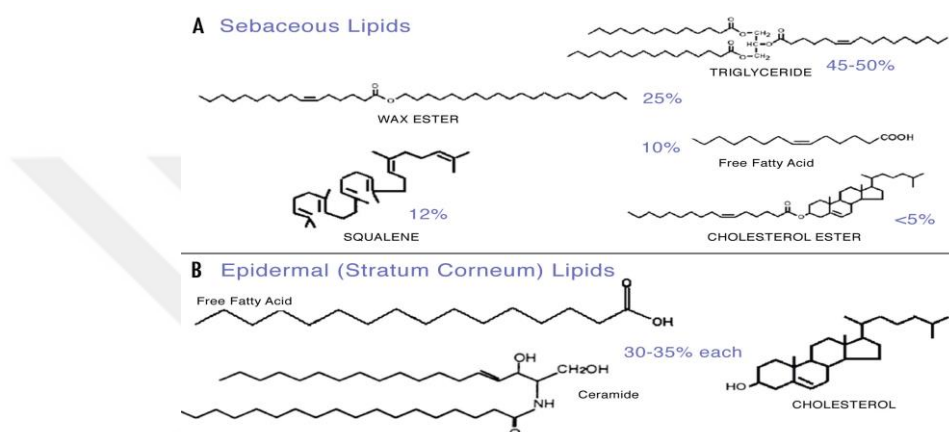


Figure 53. The structure of various lipids found in sebum and stratum corneum

The composition of sebum and epidermal lipids are given in Table 7 (99).

Table 7. Composition of skin surface lipids (92).

Lipids	Sebum (%)	Epidermal lipids (%)
Triglycerides	30-50	30-35
Free fatty acids	15-30	8-16
Wax esters	26-30	-
Squalene	12-20	-
Cholesterol Esters	3-6	15-20
Cholesterol	1.5-2.5	20-25

2.3.5.5 Alteration of sebum composition and acne

The changes in the composition of skin lipids are important in the development of acne (100). In acne patients, the level and the characteristics of sebum have been found changed (80, 89). Increased sebum production, alteration of lipid composition of sebum,

changes in the oxidant/antioxidant ratio characteristic of skin surface lipids are the main things in the pathogenesis of acne (100).

When the sebum content of acne patients was compared to normal skin, decreased level of linoleic acid as fatty acid, decreased level of vitamin E as a major sebum antioxidant, high level of squalene peroxide, significantly different ratio between saturated and unsaturated fatty acids were found. Also, the presence of lipoperoxidases in the sebum of acne patients indicates the lipid peroxidation of lipids, especially peroxidation of squalene (99, 100). Human sebum transports lipophilic antioxidants such as vitamin E or humectants such as glycerol in order to protect skin from lipid peroxidation and proper barrier function respectively (89, 98). Lipid peroxidation of fatty acids and squalene produce both intracellular and extracellular ROS. The production of ROS can change the composition of sebum (80). The oxidation of sebum lipids induce alteration of keratinocyte differentiation and induce IL-1 secretion, leading to follicular hyperkeratinization that induce the formation of comedone and progression of acne (99, 101).

The role of lipid peroxidation by-products in sebum, particularly squalene peroxides as the main products are important for the onset and development of acne. The adverse effects of squalene peroxidation products, especially monohydroperoxide have been evaluated with both invitro and in vivo experiment and given in Table 8.

Table 8. Adverse effects of squalene peroxidation (99).

Adverse effect	Acne
Comedogenicity	+
Sebaceous gland hyperplasia	+
Hyperkeratosis	+
Release of inflammatory mediators from keratinocytes	+

By product of lipid peroxidation change the sebum composition and impact the keratinocytes proliferation and differentiation. Also, they are capable of activation of peroxisome proliferators-activated receptors (PPARs), which are the nuclear transcription factors involved in the controlling of lipid metabolism and inflammation (99).

2.3.6 Chloracne and Hyperkeratinization

Chloracne is an acne-like skin condition such as comedones, papules, pustules. It is caused by exposure to halogenated aromatic hydrocarbons compounds. Dioxin and dioxin-like compounds are known as the most significant chloracnegen (102, 103). Exposure routes of chloracnegen could be via direct skin contact, ingestion, and inhalation. Exposure to dioxins can be occupational, accidental, or even intentional poisoning as in the case of the Ukrainian ex-president, Viktor Yushchenko. Chloracne has been a typical occurrence of dioxin poisoning with TCDD as shown in Figure 54. (102, 103, 104).



Figure 54. Ex-President Viktor Yushchenko of Ukraine before and after dioxin poisoning with 2,3,7,8-TCDD (104).

Dioxins induce chloracne and hyperpigmentation (102). In dioxin toxicity, **oxidative stress** and **hyperkeratinization** are responsible for chloracne. (105, 106). In the formation of chloracne, dioxins show their toxicity via **aryl hydrocarbon receptor (AHR)** by the means of the **production of ROS** in **keratinocytes and sebocytes** (106). A scheme of pathogenesis of chloracne is given in Figure 55. Firstly, dioxins bind the AHR and induce the production of xenobiotic-metabolizing enzyme CYP1A1 which leads to an excessive amount of ROS generation in sebocytes and keratocytes that result in **hyperkeratinization and chloracne** (102). Also, dioxins activate the AHR and upregulate the expression of melanogenic genes including MITF, TYR, TYRP1, and TYPR2. Thus, melanogenesis is increased in **melanocytes** and results in hyperpigmentation (102).

Furue and Tsuji suggested (2019) cinnamaldehyde and cinnamon-containing herbal drugs for chloracne for their dual effect. They are potent inhibitors of AHR-mediated CYP1A1 expression. Also, they activate NRF2 and upregulate gene expression

of antioxidant enzymes for neutralizing ROS. Thus, they may inhibit the formation of chloracne (102).

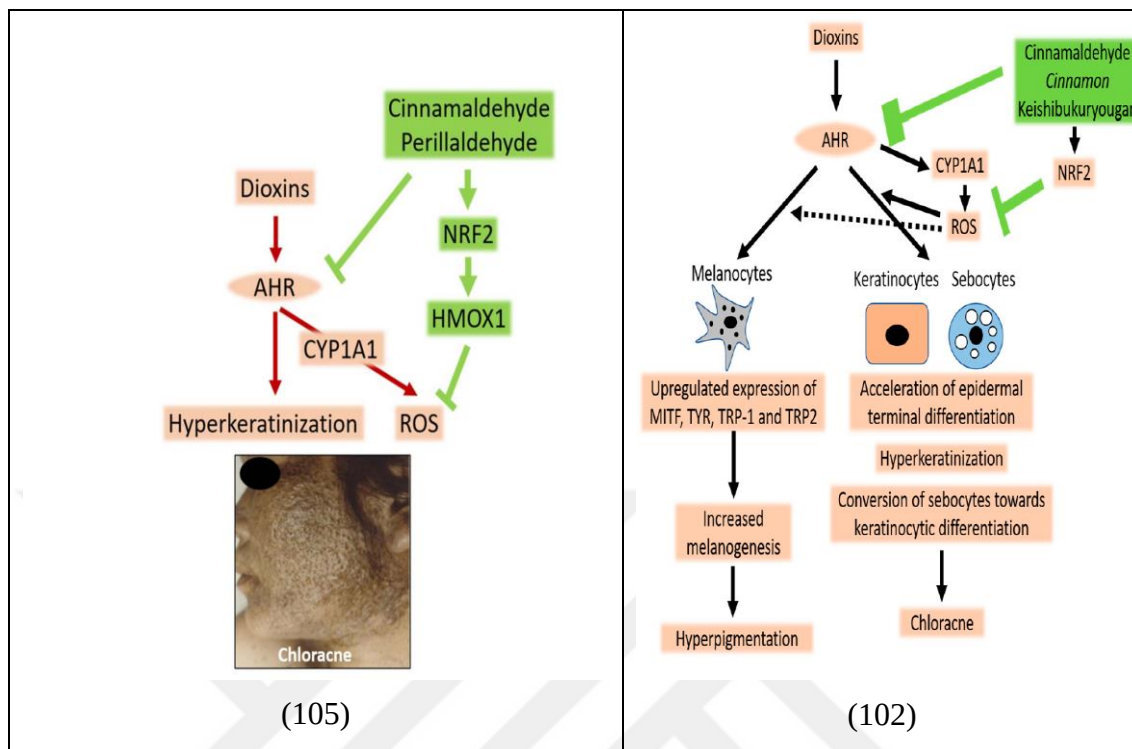


Figure 55. The pathogenesis of chloracne and hyperpigmentation by dioxins (102, 105)

* Abbreviations: MITF: microphthalmia-associated transcription factor, TYR: tyrosinase, TYRP1 and TYRP2: tyrosinase-related protein 1 and 2, NRF2: nuclear factor erythroid 2-related factor

In fact, dioxins are chemically stable but tolerant to CYP1A1 activity (106). It is known that CYP1 enzyme is induced by the AH receptor (107). Therefore, dioxins bind AHR and induce oxidative stress by inducing CYP1A1 which metabolizes the ligand and induces ROS generation during the metabolic process. In brief, dioxin-induced oxidative stress is responsible for dioxin-mediated toxicity (106).

Dioxins are environmental contaminants produced by industrial processes such as in a process of burning, chlorine bleaching of paper and manufacturing of some pesticides, herbicides and fungicides. Dioxin and dioxin-like compounds consist of a large group of chemicals which includes polychlorinated dibenzo-p-dioxins (PCDDs), dibenzofurans (PCDFs), biphenyls (PCBs) and structurally related chemicals. The structure of these compounds is given in Figure 56 (104).

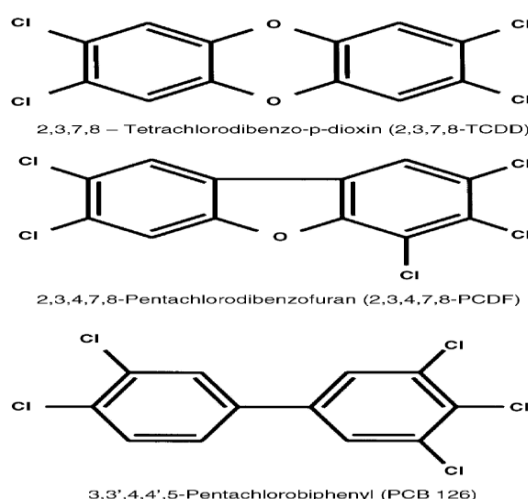


Figure 56. The chemical structure of selected dioxins and dioxin-like compounds

Dioxins are composed of two benzene rings attached by two oxygen atoms and include four to eight chlorine atoms. There are total up to 75 dioxins and analogs. TCDD (2,3,7,8-tetrachlorodibenzo-p-dioxin) is the most toxic dioxin. TCDD exposure was from Vietnam war, as contaminant of Agent Orange herbicide, from and industrial accident in Seveso, Italy. PCDF exposure from the mass poisoning in Taiwan, is known as the Yucheng illness (102, 104). The toxic potential of dioxins and dioxin-like compounds are given in Table 9. Dioxin toxic equivalency factors (TEFs) based on their toxic relative potency compared to the most toxic dioxin, TCDD (104).

Table 9. Dioxin TEFs according to World Health Organization (WHO)

		WHO TEF
Dioxins	2,3,7,8-Tetra-CDD	1
	1,2,3,7,8-Penta-CDD	1
	1,2,3,4,7,8-Hexa-CDD	0.1
	1,2,3,6,7,8-Hexa-CDD	0.1
	1,2,3,7,8,9-Hexa-CDD	0.1
	1,2,3,4,6,7,8-Hepta-CDD	0.01
	OCDD	0.0001
Dibenzofurans	2,3,7,8-Tetra-CDF	0.1
	1,2,3,7,8-Penta-CDF	0.05
	2,3,4,7,8-Penta-CDF	0.5
	1,2,3,4,7,8-Hexa-CDF	0.1
	1,2,3,6,7,8-Hexa-CDF	0.1
	1,2,3,7,8,9-Hexa-CDF	0.1
	2,3,4,6,7,8-Hexa-CDF	0.1
	1,2,3,4,6,7,8-Hepta-CDF	0.01
	1,2,3,4,7,8,9-Hepta-CDF	0.01
	OCDF	0.0001
Coplanar PCBs	3,3',4,4'-TCB (77)	0.0001
	3,4,4',5-TCB (81)	0.0001
	3,3',4,4',5-PeCB (126)	0.1
	3,3',4,4',5,5'-HxCB (169)	0.01
Mono-ortho-PCBs	2,3,3',4,4'-PeCB (105)	0.0001
	2,3,4,4',5-PeCB (114)	0.0005
	2,3',4,4',5-PeCB (118)	0.0001
	2',3,4,4',5-PeCB (123)	0.0001
	2,3,3',4,4',5-HxCB (156)	0.0005
	2,3,3',4,4',5'-HxCB (157)	0.0005
	2,3',4,4',5,5'-HxCB (167)	0.00001
	2,3,3',4,4',5,5'-HpCB (189)	0.0001

The skin is one of the most target organs for dioxins because of the being AH receptor ligands. The AHR is plentifully expressed in skin cells such as sebocytes, melanocytes and keratinocytes including epidermal keratinocytes (102, 106). Dioxins activate AHR excessively that induces epidermal hyperkeratinization via keratinocyte differentiation depending on oxidative stress and then keratin plug formation occurs in the orifices of hair follicles and sebaceous gland as shown in Figure 57. Also, hyperpigmentation is induced by dioxins in the skin (106).

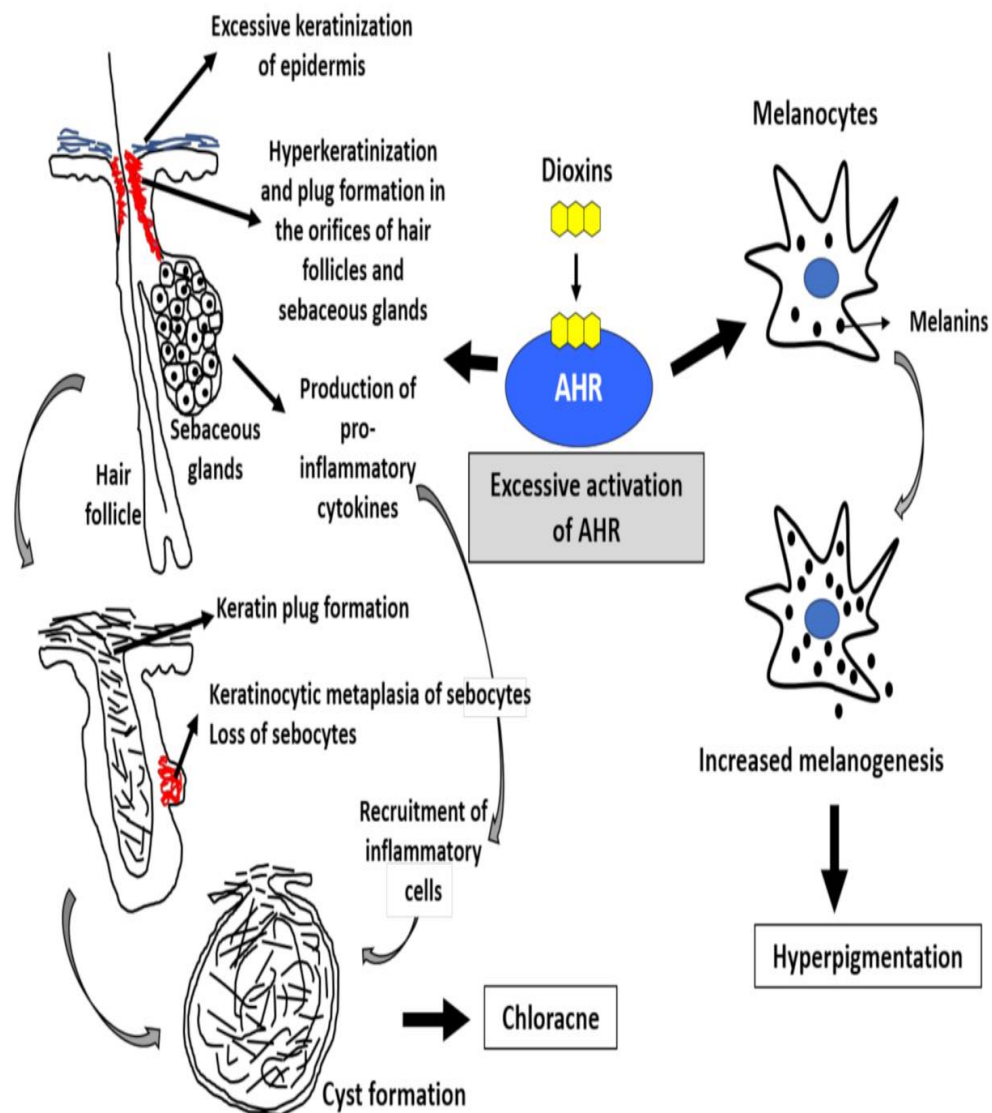


Figure 57. Hyperkeratinization induced by dioxins in the orifices of hair follicle and sebaceous glands. Hyperpigmentation induced by dioxins in melanocytes in the skin. (106).

2.4 Skin Irritation

2.4.1 Definition of Skin Irritation

Skin irritation is one of the most common adverse reactions to chemicals among skin problems. Skin irritation is characterized by the presence of symptoms of erythema, oedema, desquamation and dryness of the skin, itching, and pain. The characteristic sign of irritation is inflammatory reactions and altered homeostasis. Skin irritation is initiated directly via the inflammatory effect on the skin. Therefore, it is defined as a local inflammatory reaction that seems shortly after stimulation and fades away during a few days (108). A large number of chemicals from different chemical classes can lead to chemically induced skin irritation. Indeed, every chemical is an irritant to some degree, even water has possible irritant potential after long-term exposure (108).

2.4.2 Mechanism of Skin Irritation

In general, different chemical classes and chemical structures of compounds have irritant potential. Therefore, there can be different pathways involved in skin irritation. Generally, it is accepted that chemicals initiate and modulate skin irritation by two distinct pathways. Firstly, they gave damage to the barrier function of the stratum corneum, and secondly, irritants directly affect the cells of the skin. These two pathways lead to irritation either alone or in combination (108).

The skin has a “**lipid cement**” composition that consists of various lipids such as ceramides, free fatty acids, free cholesterol, and cholesterol esters. This lipid composition acts as an effective barrier function against endogenous and exogenous factors. It is shown that concentration and exposure time of irritants affect the response of the skin that is strongly related to the barrier capacity of the stratum corneum layer of the skin. It is known that irritants enter the stratum corneum by leading to delipidation and protein denaturation. Delipidation is the process in which the lipid balance of the skin is disrupted and results in the change of lipid cement composition and loss of the barrier function of the skin (108).

The best-described mechanism of skin irritation is the skin irritation that is triggered by surfactants (Figure 58). The surfactants, e.g. sodium dodecyl sulfate (SDS), sodium lauryl sulfate (SLS), are used as a positive control for skin irritation tests. Because they disrupt the skin cell membrane and begin the inflammatory cascade as the main switch of releasing the IL-1alpha which is the proinflammatory cytokine found in the cytoplasm of keratinocytes of

the skin epidermal layer. Figure 58 shows the skin irritation pathway in which surfactants initiate the release of the IL-1 α and as well, leading to the induction of secondary mediators such as IL-6, IL-8, TNF- α , PLA₂, and PGE₂ which are the molecular response, followed by morphological effects and typical symptoms of skin irritation (108).

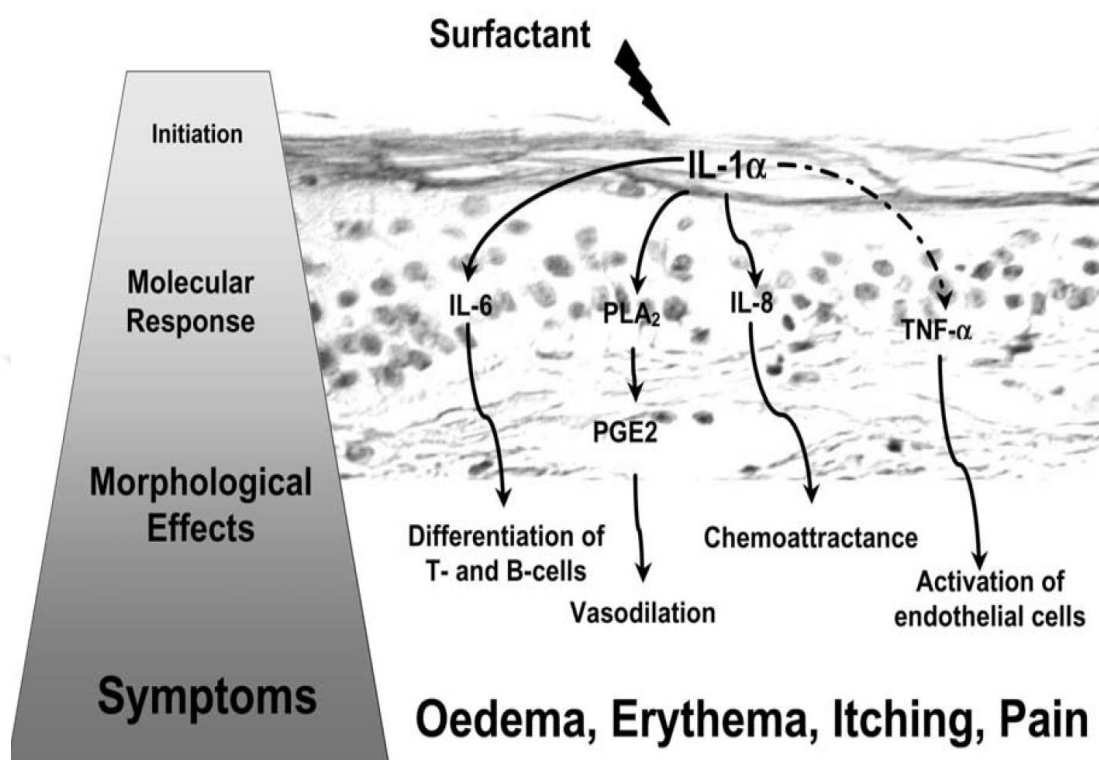


Figure 58. The mechanism of skin irritation that induced by surfactants (108).

Physico-chemical properties of molecules increase the efficacy of cosmetic products but also, can adversely affect the skin. Molecules having lipid-solving properties like surfactants and emulsifiers due to their hydrophobic tail and a hydrophilic head can interact with the lipids of the skin. Therefore, they gave damage to the lipid barrier function of the skin membrane and initiate inflammation which leads to skin irritation. In this context, aromatic and aliphatic hydrocarbons and their derivatives can lead to skin irritation. Also, inorganic chemicals like salts, acids, and bases can trigger skin irritation.

In contrast to this skin irritation mechanism, irritants without membrane-damaging properties probably do not initiate the inflammation alone by the release of IL-1 α .

3. MATERIAL AND METHODS

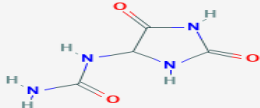
3.1 Materials

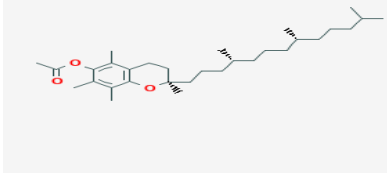
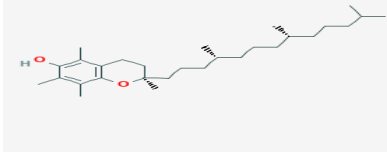
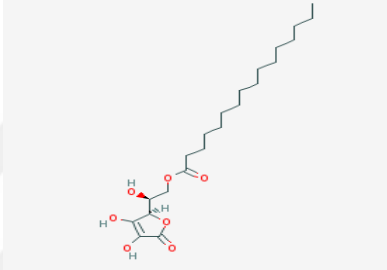
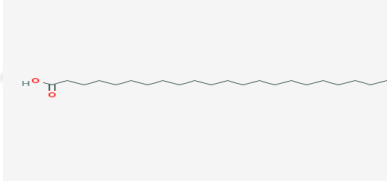
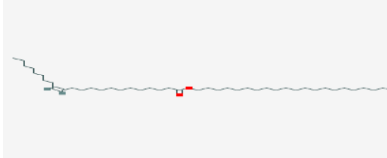
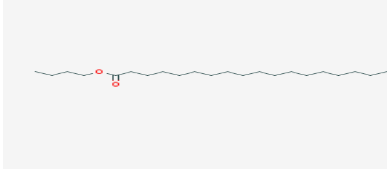
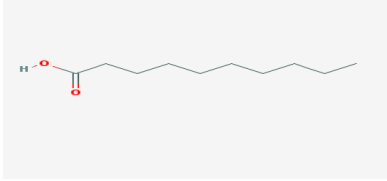
3.1.1 Dataset

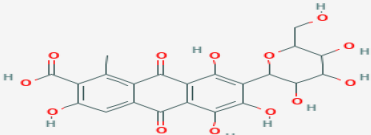
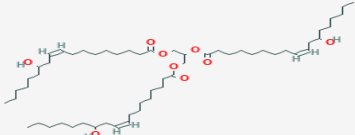
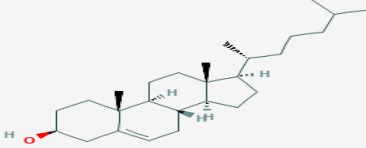
The dataset was composed of 125 cosmetic ingredients which were retrieved from the comedogenicity and irritancy study of a popular dermatologist, Dr. James Fulton (1989) in the literature which had been performed on rabbits in the Acne Research Institute (15). The chemical names, structures, formulas, SMILES (Simplified molecular-input line entry system) strings, and CAS (chemical abstract services) numbers of the dataset were collected via various databases from the internet and listed in Table 10. The compounds in this dataset consist of ten chemical classes which are fatty acid and derivatives (n=33), alcohols-fatty alcohols-sugars and their derivatives (n=43), pigments (n=13), silicones (n=2), sterols (n=3), vitamins (n=6), preservatives & additives (n=7) miscellaneous (n=11), thickeners (n=1) and oils (n=2).

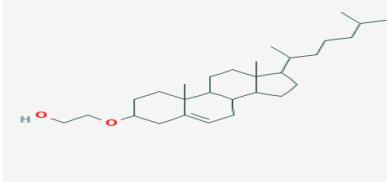
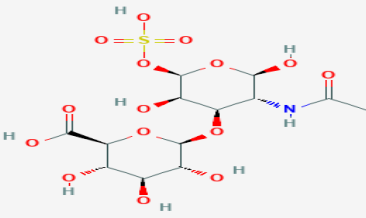
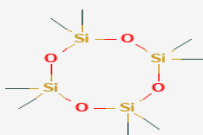
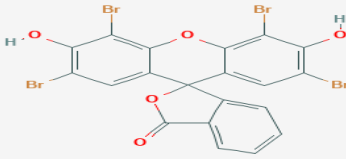
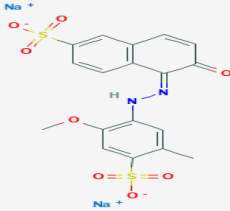
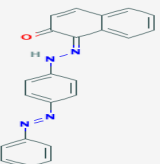
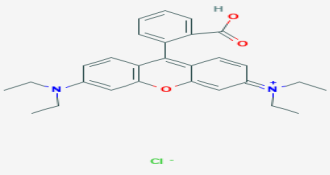
*n=number of compound in each chemical class

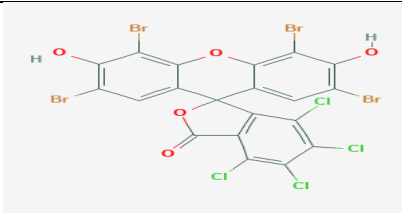
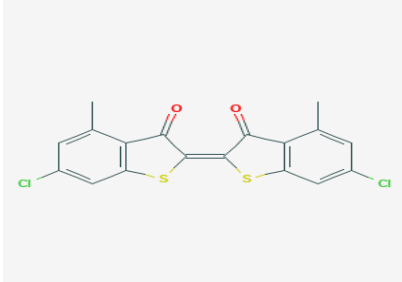
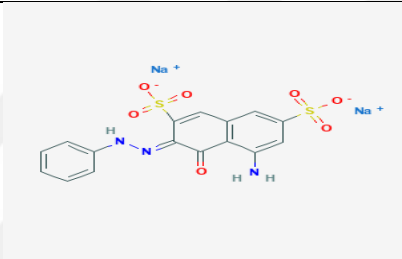
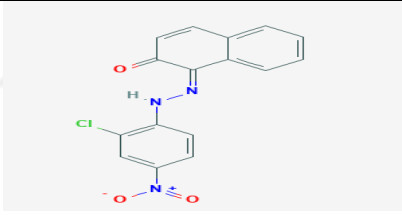
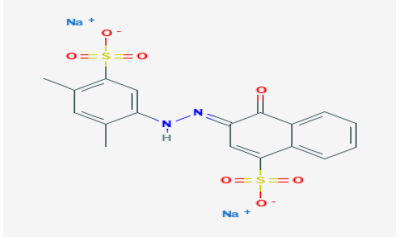
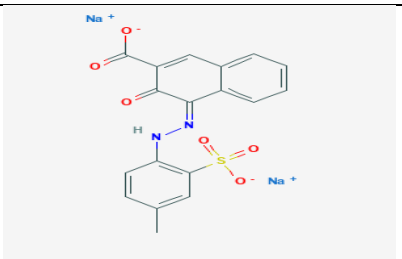
Table 10. The chemical names, structures, SMILES strings, CAS numbers of the datasets

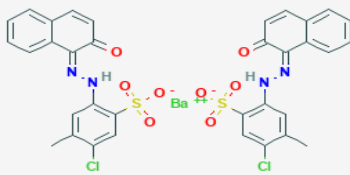
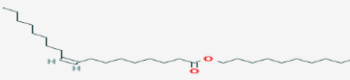
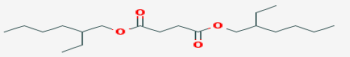
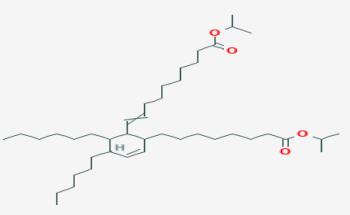
ID	Chemical Name	Structure	CAS	SMILES
1	1-3 butanediol (butylene glycol)		107-88-0	CC(CCO)O
2	Acetone		67-64-1	CC(=O)C
3	Allantoin		97-59-6	C1(C(=O)NC(=O)N1)NC(=O)N

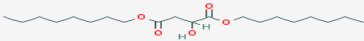
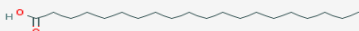
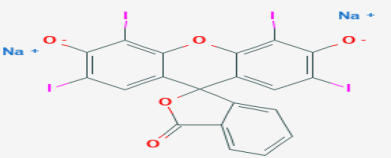
ID	Chemical Name	Structure	CAS	SMILES
4	Alpha-tocopherol acetate (Vitamin E acetate)		58-95-7	<chem>CC1=C2C(=C(C(=C1)OC(=O)C)C)CCC(O2)CCCC(C)CCC(C)CCCC(C)C</chem>
5	Alpha-tocopherol (Vitamine E, Alpha-Tocopherol)		59-02-9	<chem>CC1=C(C(=C2CCC(OC2=C1C)(C)CCCC(C)CCCC(C)CCCC(C)C)C)O</chem>
6	Ascorbyl palmitate		137-66-6	<chem>CCCCCCCCCCCCCCCC(=O)OC[C@H]1C([C@@H](O1)C(=O)O)O</chem>
7	Behenic acid		112-85-6	<chem>CCCCCCCCCCCCCCCCCCCCCCCCCCCC(=O)O</chem>
8	Behenyl erucate		18312-32-8	<chem>CCCCCCCCCCCCCCCCCCCCCCCCCCCCOC(=O)CCCCCCCCCCCCCCCC=CCCCCCCCC</chem>
9	Butyl stearate		123-95-5	<chem>CCCCCCCCCCCCCCCCCCCC(=O)OCCCC</chem>
10	Capric acid (Decanoic acid)		334-48-5	<chem>CCCCCCCCCCCC(=O)O</chem>

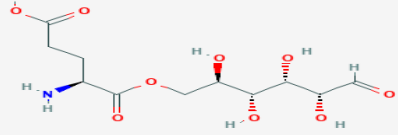
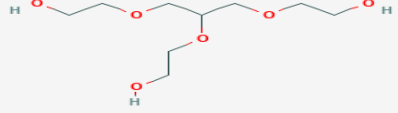
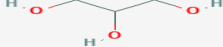
ID	Chemical Name	Structure	CAS	SMILES
11	Caprylic acid (Octanoic acid)		124-07-2	<chem>CCCCCCCC(=O)O</chem>
12	Carmine (Carminic acid)		1260-17-9	<chem>CC1=C2C(=CC(=C1C(=O)O)O)C(=O)C3=C(C2=O)C(=C(C(=C3O)O)C4C(C(C(C(O4)CO)O)O))O</chem>
13	Castor oil		8001-79-4	<chem>CCCCCCC(CC=CCCCCCCC(=O)OCC(COC(=O)CCCCCCCC=CCCC(CCCCCC)O)OC(=O)CCCCCCCC=CCCC(CCCCCC)O)O</chem>
14	Cetyl acetate		629-70-9	<chem>CCCCCCCCCCCCCCCCCOC(=O)C</chem>
15	Cetyl alcohol		29354-98-1	<chem>CCCCCCCCCCCCCCCCCO</chem>
16	Cetyl palmitate		540-10-3	<chem>CCCCCCCCCCCCCCCCCOC(=O)CCCCCCCCCCCCCCCC</chem>
17	Cholesterol		57-88-5	<chem>CC(C)CCCC(C)C1CCC2C1(CCC3C2CC=C4C3(CCC(C4)O)C)C</chem>

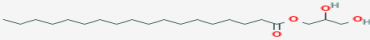
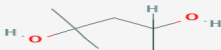
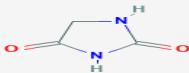
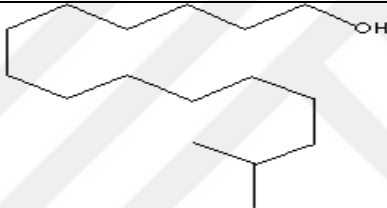
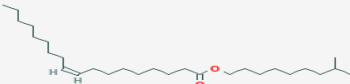
ID	Chemical Name	Structure	CAS	SMILES
18	Choleth 24		27321-96-6	<chem>CC(C)CCCC(C)C1C CC2C1(CCC3C2CC =C4C3(CCC(C4)OC CO)C)C</chem>
19	Chondroitin sulfate		9088-44-2	<chem>CC(=O)NC1C(C(C(OC1O)OS(=O)(=O) O)O)OC2C(C(C(C(O2)C(=O)O)O)O)O</chem>
20	Cyclomethicone (Cyclomethicone 4)		556-67-2	<chem>C[Si]1(O[Si](O[Si](O[Si](O1)(C)C)(C)C)C)C</chem>
21	D&C red 21 (Bromeosin; C.I. 45380:2; C.I; Eosin acid)		15086-94-9	<chem>C1=CC=C2C(=C1)C (=O)OC23C4=CC(= C(C(=C4OC5=C(C(=C(C=C35)Br)O)Br)Br)O)Br</chem>
22	D&C Red 40 (Allura Red)		25956-17-6	<chem>CC1=CC(=C(C=C1S (=O)(=O)[O-]])OC)N=NC2=C(C= CC3=C2C=CC(=C3) S(=O)(=O)[O-]])O.[Na+].[Na+]</chem>
23	D&C red17 (Sudan III)		85-86-9	<chem>OC1=CC=C2C=CC =CC2=C1/N=N/C3 =CC=C(/N=N/C4= CC=CC=C4)C=C3</chem>
24	D&C red19 (Rhodamine B)		81-88-9	<chem>CCN(CC)C1=CC2= C(C=C1)C(=C3C=C C(=[N+](CC)CC)C= C3O2)C4=CC=CC= C4C(=O)O.[Cl-]</chem>

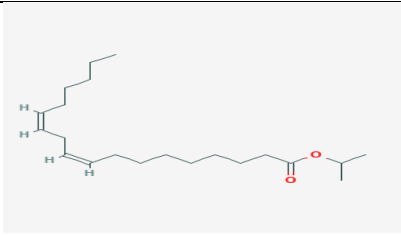
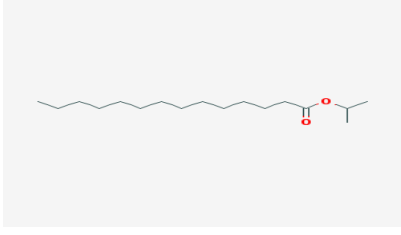
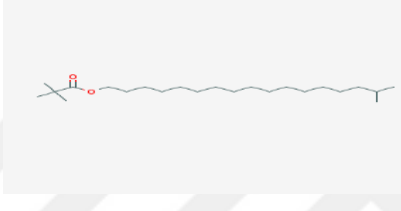
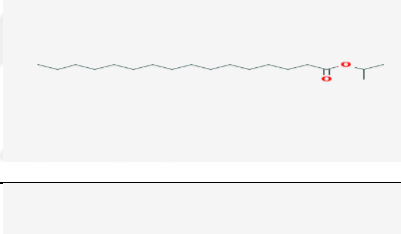
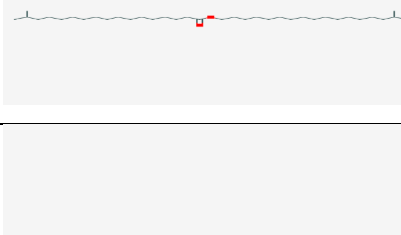
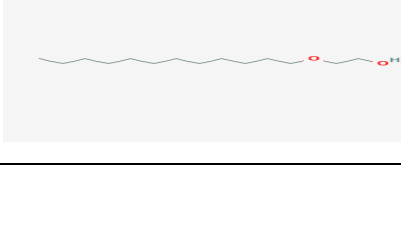
ID	Chemical Name	Structure	CAS	SMILES
25	D&C red27 (Acid Phloxine PB)		13473-26-2	<chem>C1=C2C(=C(C(=C1Br)O)Br)OC3=C(C(=C(C=C3C24C5=C(C(=C(C(=C5Cl)Cl)C)Cl)C(=O)O4)Br)O)Br</chem>
26	D&C red30 (Permanent Pink)		2379-74-0	<chem>CC1=CC(=CC2=C1C(=O)C(=C3C(=O)C4=C(S3)C=C(C=C4C)Cl)S2)Cl</chem>
27	D&C red33 (Azofuchsin)		3567-66-6	<chem>C1=CC=C(C=C1)N/N=C/2\C(=CC3=C(C(=CC(=C3C2=O)N)S(=O)(=O)[O-])S(=O)(=O)[O-].[Na+].[Na+]</chem>
28	D&C red36 (Permansa Red)		2814-77-9	<chem>C1=CC=C\2C(=C1)C=CC(=O)/C2=N\NC3=C(C=C(C=C3)[N+](=O)[O-])Cl</chem>
29	D&C red4 (Ponceau SX Lake)		4548-53-2	<chem>CC1=CC(=C(C(=C1N=NC2=C(C3=CC=CC=C3C(=C2)S(=O)(=O)[O-])O)S(=O)(=O)[O-])C.[Na+].[Na+]</chem>
30	D&C red6 (Brilliant Carmine 6B)		5281-04-9	<chem>CC1=CC(=C(C(=C1)N/N=C\2/C3=CC=CC=C3C(=C2=O)C(=O)[O-])S(=O)(=O)[O-].[Na+].[Na+]</chem>

ID	Chemical Name	Structure	CAS	SMILES
31	D&C red9 (Brilliant red)		5160-02-1	<chem>CC1=CC(=C(C=C1C)S(=O)(=O)[O-]))N/N=C\2/C3=CC=CC=C3C=CC2=O.CC1=CC(=C(C=C1C)S(=O)(=O)[O-]))N/N=C\2/C3=CC=CC=C3C=CC2=O.[Ba+2]</chem>
32	Decyl oleate		3687-46-5	<chem>CCCCCCCCCCCCOC(=O)CCCCCCCC=CCCCCCCC</chem>
33	Di (2 ethylhexyl) succinate		2915-57-3	<chem>CCCCC(CC)COC(=O)CCC(=O)OCC(C)CCCC</chem>
34	Diethylene Glycol Monoethyl Ether		111-90-0	<chem>CCOCCOCCO</chem>
35	Diisopropyl dimerate		308065-68-1	<chem>CCCCCCC1C=CC(C(C1CCCCC))C=CCCCCCCC(=O)OC(C)C)CCCCCCCC(=O)OC(C)C</chem>
36	Dimethicone		9016-00-6	<chem>C[Si]1(O[Si](O[Si](O[Si](O1)(C)C)(C)C)(C)C)C</chem>

ID	Chemical Name	Structure	CAS	SMILES
37	Diisopropyl adipate		6938-94-9	<chem>CC(C)OC(=O)CCCCC(=O)OC(C)C</chem>
38	Dioctyl malate		15763-02-7	<chem>CCCCCCCCCOC(=O)CC(C(=O)OCCCCCCC)O</chem>
39	Dioctyl succinate		14491-66-8	<chem>CCCCCCCCCOC(=O)CCC(=O)OCCCCCCC</chem>
40	Eicosanoic acid		506-30-9	<chem>CCCCCCCCCCCCCCCCCCCCCCCC(=O)O</chem>
41	Erythrosine (D&C red 3)		16423-68-0	<chem>C1=CC=C2C(=C1)C(=O)OC23C4=CC(=C(C(=C4OC5=C(C(=C(C=C35)I)[O-])I)[O-])I.[Na+].[Na+])</chem>
42	EthanolSD40		64-17-5	<chem>CCO</chem>
43	Ether (ethyl ether)		60-29-7	<chem>CCOCC</chem>

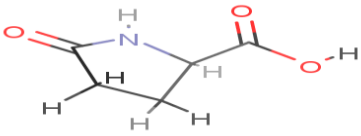
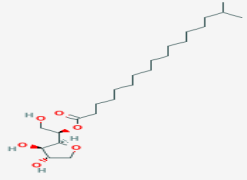
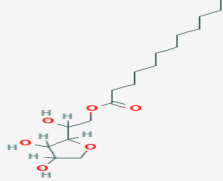
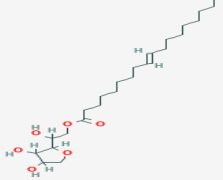
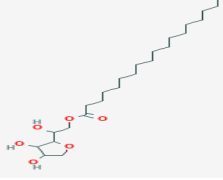
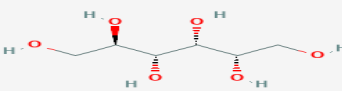
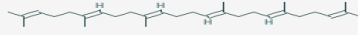
ID	Chemical Name	Structure	CAS	SMILES
44	Ethylene glycol monoethyl ether		110-80-5	CCOCCO
45	Ethylene glycol monostearate (Glycol stearate)		111-60-4	CCCCCCCCCCCCCCCCCCCC(=O)OCCO
46	Ethylhexyl palmitate		29806-73-3	CCCCCCCCCCCCCCCCCCCC(=O)OCC(CC)CCCC
47	Ethylhexyl pelargonate		59587-44-9	CCCCCCCC(=O)OCC(CC)CCCC
48	Glucose glutamate		59279-63-9	C(CC(=O)O)[C@@H](C(=O)OC[C@H]([C@H]([C@@H]([C@H](C(=O)O)O)O)O)N
49	Glycereth-26		31694-55-0	OCCOCC(COCCO)OCCO
50	Glycerin		56-81-5	C(C(CO)O)O

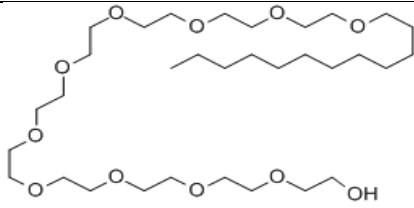
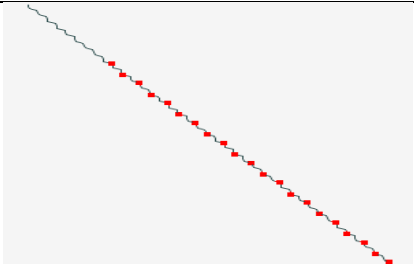
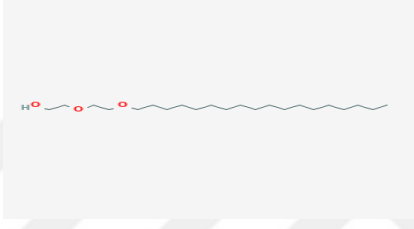
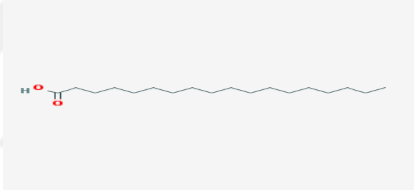
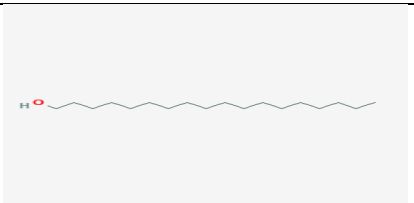
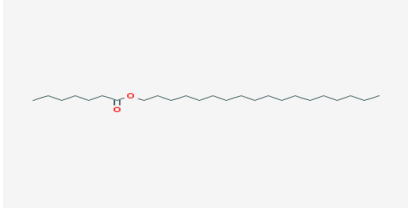
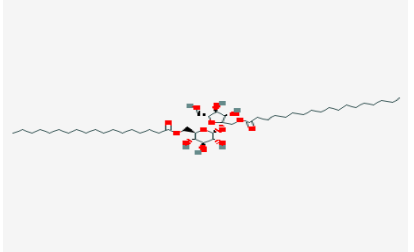
ID	Chemical Name	Structure	CAS	SMILES
51	Glyceryl stearate NSE		123-94-4	<chem>CCCCCCCCCCCCCCCCCCCC(=O)OCC(O)CO</chem>
52	Hexylene glycol		107-41-5	<chem>CC(CCC(C)(C)O)O</chem>
53	Hydantoin		461-72-3	<chem>C1C(=O)NC(=O)N1</chem>
54	Isocetyl alcohol		36311-34-9	<chem>CC(C)CCCCCCCCCCCCO</chem>
55	Isodecyl oleate		59231-34-4	<chem>CCCCCCCC=CCCCCCCC(=O)OCCCCCCCC(C)C</chem>
56	Isopropyl alcohol		67-63-0	<chem>CC(C)O</chem>
57	Isopropyl isostearate		31478-84-9	<chem>CC(C)CCCCCCCCCCCCCCCC(=O)OC(C)C</chem>

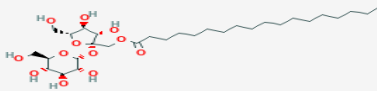
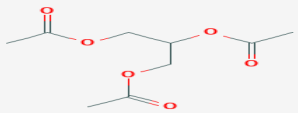
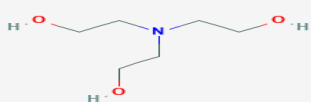
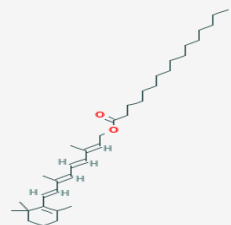
ID	Chemical Name	Structure	CAS	SMILES
58	Isopropyl linolate		22882-95-7	<chem>CCCCC=CCC=CCC CCCCC(=O)OC(C) C</chem>
59	Isopropyl myristate		110-27-0	<chem>CCCCCCCCCCCCC C(=O)OC(C)C</chem>
60	Isostearyl neopentanoate		58958-60-4	<chem>CC(C)CCCCCCCCC CCCCCOC(=O)C(C)(C)C</chem>
61	Isopropyl palmitate		142-91-6	<chem>CCCCCCCCCCCCC CCC(=O)OC(C)C</chem>
62	Isostearyl isostearate		41669-30-1	<chem>CC(C)CCCCCCCCC CCCCCOC(=O)CC CCCCCCCCCCCCC(C)C</chem>
63	Laureth 23		68439-50-9	<chem>CCCCCCCCCCCCCO CCOCCOCCOCCO CCOCCOCCOCCO CCOCCOCCOCCO CCOCCOCCOCCO CCOCCOCCOCCO CCOCCOCCO</chem>
64	Laureth 4 (polidocanol)		4536-30-5	<chem>CCCCCCCCCCCCCO CCO</chem>

ID	Chemical Name	Structure	CAS	SMILES
65	Lauric acid (Dodecanoic acid)		143-07-7	<chem>CCCCCCCCCCCC(=O)O</chem>
66	Lithium stearate		4485-12-5	<chem>[Li+].CCCCCCCCCCCCCCCC(=O)[O-]</chem>
67	Methyl paraben		99-76-3	<chem>COC(=O)C1=CC=C(C=C1)O</chem>
68	Myristic acid (tetradecanoic acid)		544-63-8	<chem>CCCCCCCCCCCCCCCC(=O)O</chem>
69	Myristyl alcohol		112-72-1	<chem>CCCCCCCCCCCCCCCCO</chem>
70	Myristyl lactate		1323-03-1	<chem>CCCCCCCCCCCCCCCCOC(=O)C(C)O</chem>
71	Myristyl myristate		3234-85-3	<chem>CCCCCCCCCCCCCCCCOC(=O)CCCCCCCCCCCC</chem>
72	m-xylene		108-38-3	<chem>CC1=CC(=CC=C1)C</chem>
73	o-xylene		95-47-6	<chem>CC1=CC=CC=C1C</chem>

ID	Chemical Name	Structure	CAS	SMILES
98	PPG 2 myristyl ether propionate		74775-06-7	<chem>CCCCCCCCCCCCCCCCOCC(C)OCC(=O)CC</chem>
99	Propyl paraben		94-13-3	<chem>CCCC(=O)OC1=CC=C(O)C=C1</chem>
100	Propylene glycol		57-55-6	<chem>CC(CO)O</chem>
101	Propylene glycol (PG) diperlargonate		41395-83-9	<chem>CCCCCCCCCCCC(=O)OCC(C)OC(=O)CCCCCCCC</chem>
102	PVP		9003-39-8	<chem>C=CN1CCCC1=O</chem>
103	p_xylene		106-42-3	<chem>CC1=CC=C(C=C1)C</chem>
104	Sodium hyaluronate		9067-32-7	<chem>CC(=O)NC1C(C(C(OC1O)CO)O)OC2C(C(C(C(O2)C(=O)O)OC3C(C(C(C(O3)CO)O)OC4C(C(C(C(O4)C(=O)O)O)O)O)NC(=O)C)O)O.[Na+]</chem>
105	Sodium pca (cactus)		54571-67-4	<chem>C1CC(=O)NC1C(=O)[O-].[Na+]</chem>

ID	Chemical Name	Structure	CAS	SMILES
106	Sodium PCA (pubchem,sdf)		54571-67-4	<chem>[H]OC(=O)[C@]1([H])N([H])C(=O)C([H])([H])C1([H])[H]</chem>
107	Sorbitan isostearate		54392-26-6	<chem>CC(C)CCCCCCCCCCCCC(=O)O[C@H](CO)[C@@H]1[C@@H]([C@H](CO)O)O</chem>
108	Sorbitan laurate		5959-89-7	<chem>CCCCCCCCCCCC(=O)OCC(C1C(CO1)O)O</chem>
109	Sorbitan oleate (Span 80)		1338-43-8	<chem>CCCCCCCC/C=C/C(CCCCCC(=O)OCC(C1C(CO1)O)O)O</chem>
110	Sorbitan stearate (Sorbitan monostearate, Span 60)		1338-41-6	<chem>CCCCCCCCCCCCCCCC(=O)OCC(C1C(CO1)O)O</chem>
111	Sorbitol		50-70-4	<chem>C([C@H]([C@H]([C@@H]([C@H](CO)O)O)O)O)O</chem>
112	Squalene		111-02-4	<chem>CC(=CCC/C(=C/CC/C(=C/CC/C(=C/C/C(=C/C(C)C)\C)\C)/C)/C</chem>
113	Stearath-20		2136-72-3	<chem>CCCCCCCCCCCCCCCCCO</chem>

ID	Chemical Name	Structure	CAS	SMILES
114	Steareth 10 (Brij® S10)		13149-86-5	CCCCCCCCCCCCC CCCCOCCOCCOC COCCOCCOCCOC COCCOCCOCCO
115	Steareth-100		9005-00-9	CCCCCCCCCCCCC CCCCOCCOCCOC COCCOCCOCCOC COCCOCCOCCOC COCCOCCOCCOC COCCOCCOCCOC COCCO
116	Steareth-2		16057-43-5	CCCCCCCCCCCCC CCCCOCCOCCO
117	Stearic acid_		57-11-4	CCCCCCCCCCCCC CCCC(=O)O
118	Stearyl alcohol		112-92-5	CCCCCCCCCCCCC CCCCCO
119	Stearyl heptanoate		66009-41-4	CCCCCCCCCCCCC CCCCCOC(=O)CCC CCC
120	Sucrose distearate		135260-52-5	CCCCCCCCCCCCC CCCCC(=O)OCC1C (C(C(C(O1)OC2(C(C(C(O2)CO)O)O)C OC(=O)CCCCCCCC CCCCCCCCC)O)O) O

ID	Chemical Name	Structure	CAS	SMILES
121	Sucrose stearate		25168-73-4	<chem>CCCCCCCCCCCCCCCC(=O)OCC1(C(C(C(O1)CO)O)O)OC2C(C(C(C(O2)CO)O)O)O</chem>
122	Triacetin		102-76-1	<chem>CC(=O)OCC(COC(=O)C)OC(=O)C</chem>
123	Tridecyl neopentanoate		106436-39-9	<chem>CCCCCCCCCCCCCCCCOC(=O)C(C)(C)C</chem>
124	Triethanolamine		102-71-6	<chem>C(CO)N(CCO)CCO</chem>
125	Vitamin A palmitate		79-81-2	<chem>CCCCCCCCCCCCCCCCCCC(=O)OCC=C(C)C=CC=C(C)C=CC1=C(CCCCC1(C)C)C</chem>

3.2 Methods

The key steps of QSAR modeling are data collection, calculation of molecular descriptors, model development, and validation (30). In this study, the flowchart of the QSAR modeling for comedogenicity prediction was given in Figure 59. Various software tools were selected among those user-friendly, free-license, and widely recognized by academicians, authorities, and the industry during this study.

Flowchart of the QSAR Modelling

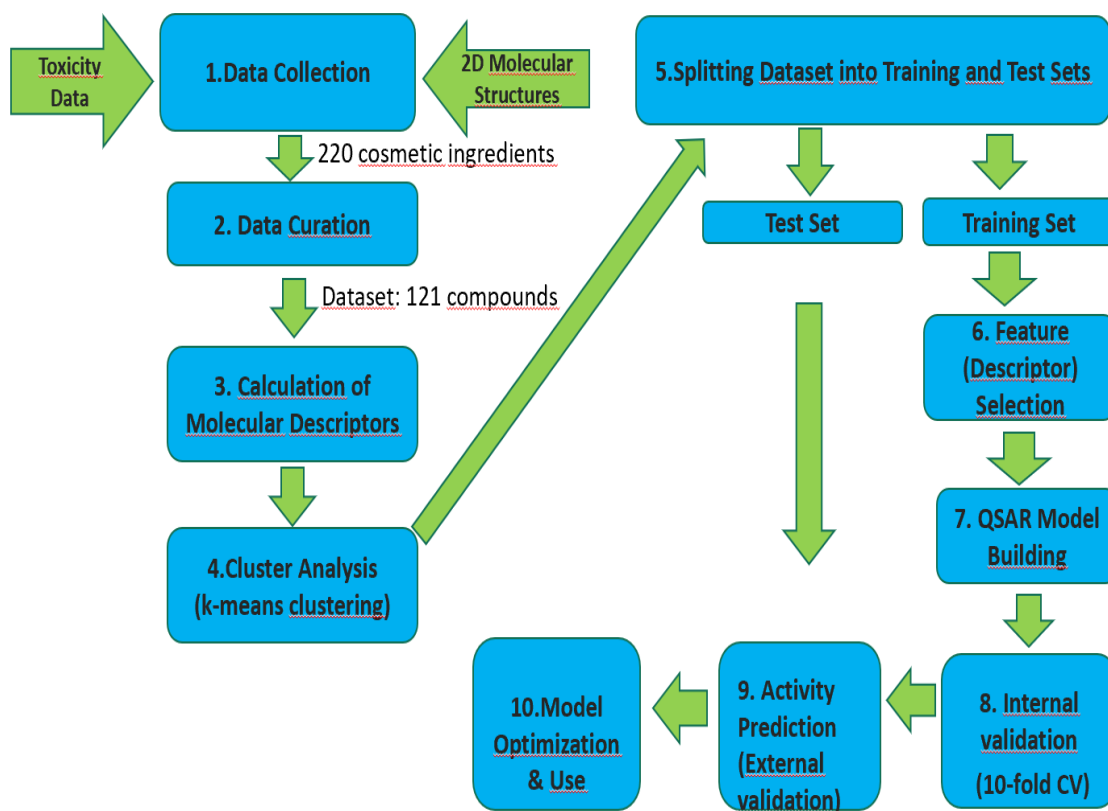


Figure 59. The flowchart of QSAR modeling for comedogenicity prediction

3.2.1 Data Collection

3.2.1.1 Collection of comedogenicity and irritation data

In this study, *in vivo* comedogenic and irritant activity data of a group of cosmetic ingredients were obtained from the same laboratory and the same experimental design from the study of a popular dermatologist on acne, James Fulton (1989) in the literature. (15). In the *in vivo* study, the comedogenic activity of the cosmetic ingredients had been tested on the New Zealand albino rabbit ears. For testing, ingredients had been mixed in propylene glycol at a 9 to 1 dilution. The opposite ear of each animal had been used as an untreated control. Also, isopropyl myristate had been applied as a positive control. For comedogenic activity, follicular keratosis had been evaluated both macroscopically (visually) and microscopically by measuring the width of the follicular keratosis using a micrometer. The comedogenic activity of chemicals had been graded on a scale of 0 to 5. (15). We considered comedogenicity grade of 0 as inactive, comedogenicity grade of 1, 2, 3, 4, and 5 as active. As well, irritation activity of the same group of cosmetic ingredients had been tested on rabbit ears. For the irritation test, irritancy had been induced by the repeated application of chemicals on the surface epidermis in the rabbit ears. The irritant activity of chemicals had been evaluated on a similar scale of 0 to 5 (15).

3.2.1.2 Collection of molecular structures

First of all, the chemical names, structures, formulas, SMILES strings and CAS numbers of the cosmetic ingredients retrieved from the study of Fulton were collected (15) and listed in Table 10. Secondly, the two-dimensional (2D) molecular structures of chemicals were imported as 2D SDF (structure data file) format from PubChem Database (<https://pubchem.ncbi.nlm.nih.gov/>) (109), Structure Databases; ChemSpider (<http://www.chemspider.com/>) (110), Mol-Instincts Chemical Property Databases (<https://www.molinstincts.com/>) (111), and ChemicalBook-chemical search engine (<https://www.chemicalbook.com/>) (112). For those structures, not available were drawn using the Marvin Sketch ChemAxon tool (<https://docs.chemaxon.com/>) (113). Some of the chemical structures were converted from SMILES to SDF file formats using CACTUS (<https://cactus.nci.nih.gov/translate/>) (114). In order to validate the structures of the chemicals downloaded from the databases, the structures were checked by their individual CAS numbers using the websites <https://webbook.nist.gov/chemistry/cas-ser/> (115) and <https://www.lookchem.com/> (116). Unavailable molecular structures were drawn using Marvin Sketch v.21.2.0 (<https://www.chemaxon.com>) (113). Also, Marvin

(v.21.2.0) was used for displaying and characterizing the molecular structures (<https://www.chemaxon.com>) (117).

3.2.2 Data Curation

Data curation was made to standardize the chemicals' onset of molecular modeling. The general data curation steps are shown in Figure 60 (118).

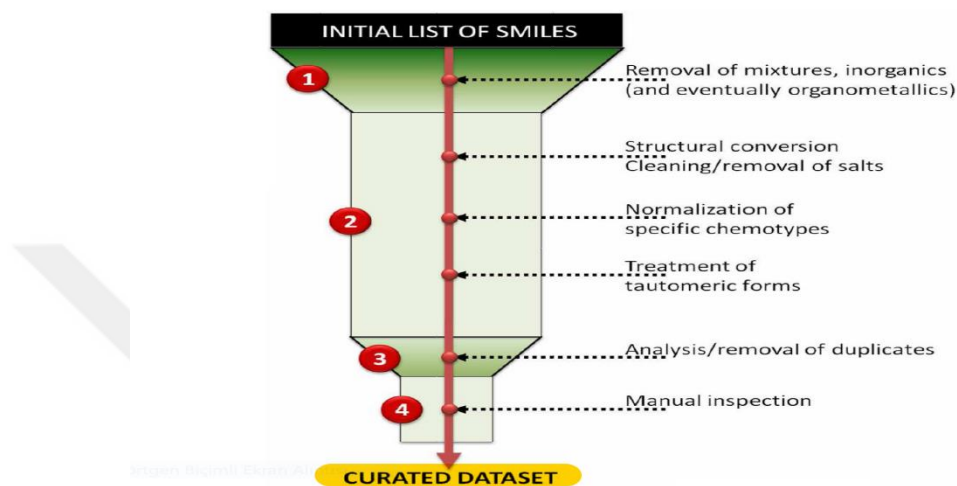


Figure 60. General dataset curation workflow

After data collection was completed by obtaining 2D SDF format, structure standardization was made by data curation. In the data curation process, initially, all mixtures and inorganic compounds (including organometallics) were removed due to the fact that the majority of molecular descriptors can be computed for organic compounds only. For example, the following compounds have been removed: natural waxes and their derivatives (such as lanolin, acetylated lanolin, beeswax, carnauba wax), natural oils (such as almond oil, olive oil), inorganic compounds (such as bentonite, kaolin, talc, magnesium aluminum silicate, precipitated sulfur, water-soluble sulfur, zinc oxide, ultramarine violet), herbs (such as chamomile extracts), mixtures (such mixture of xylene isomers). Moreover, some compounds were removed due to the fact that their SMILES strings could not be identified definitely due to incoherent names or they corresponded to a mixture of compounds (118).

The second step of data curation was structural conversion and removal of salts. Various programs could make this structural conversion, e.g., ChemAxon Marvin Sketch (v.19.3) (113, 117, 119) and Open BabelGUI (v.2.4.1) (120). ChemAxon Marvin Sketch

was used to convert a library of SMILES and obtained structures for possible errors such as manual drawing errors or conversion errors from 2D structures into SMILES strings. Moreover, removal of salts prior to QSAR was made although properties of salts form of compounds could be very different from those of the corresponding neutral molecules. Similar to inorganics, salts were not treated by most of the descriptor-calculating software, and their existence could produce various errors in the calculation of descriptors. Marvin Sketch ChemAxon Tool and OpenBabel were used to identify salts, remove counterions and neutralize the remaining organic compounds by attaching hydrogen.

In the third step, structural duplicates were checked in the datasets. If identical structures are found, duplicates were removed, and finally, manual inspection was completed (118). After these data curation steps, 121 compounds remained as a QSAR-ready structure from the original dataset containing 220 cosmetic ingredients which were retrieved from the study of Fulton (1989) in the literature (15).

The list of 121 compounds used in this QSAR study as dataset were given in the Appendix 8.1 Table S1, presented in the form of ID (Identity) number, chemical class no, formula, CAS number and SMILES string.

3.2.3 Calculation of Molecular Descriptors

Molecular descriptors are the most important parts of QSAR modeling since they are the mathematical representation of chemicals and they have been used as input data while building QSAR models (40). A total of 4837 (0D, 1D, 2D) molecular descriptors were calculated using four software, namely Mold2 v.2.0.0 (121), AlvaDesc v1.0.14 (122), T.E.S.T v.4.2.1 (123) and PaDEL v.2.21 (124). All calculated descriptors that form the descriptor pool consist of 777 Mold2, 1819 AlvaDesc, 797 T.E.S.T, and 1444 PaDEL descriptors. The type and number of molecular descriptors calculated for each software package were given in Table 11. As calculated descriptors had different ranges, all descriptors were normalized using WEKA (Waikato Environment for Knowledge Analysis) machine learning software before the selection of descriptors and modelling. The descriptors used in each model were normalized in WEKA by selecting unsupervised attribute.normalize from the filtering pane (81). In the normalization, the values of calculated descriptors are rescaled to a consistent range from 0 to 1. Thus, the descriptor values have become comparable and each descriptor has an equal chance of contributing to the overall analysis (125, 126).

Table 11. Descriptors calculated for the dataset compounds

Software	Descriptor class	Descriptor type	Number of descriptors
AlvaDesc	0D,1D,2D	Constitutional indices, RDF descriptors Ring descriptors, Topological indices, Walk and path counts, Connectivity indices, Randic molecular profiles, Information indices, Functional group counts, 2D matrix-based descriptors, Atom-centered fragments, 2D autocorrelations, Atom-type E-state indices, Burden eigenvalues, Pharmacophore descriptors, P_VSA-like descriptors, 2D atom pairs, ETA indices, Edge adjacency indices, Charge descriptors, Molecular properties, Drug-like indices, CATS 2D descriptors	1819
MOLD2	0D,1D, 2D	1D- counts for atoms chemical-physical property 2D -counts for atoms, counts for bonds, counts for functional groups, chemical-physical property, structural features, 2D autocorrelation, Balaban index, connectivity index, detour index, distance (topological) index, Eigen value-based descriptors, information content, Kier index, molecular walk counts, Schultz index, topological charge index, Wiener index, Zagreb index	777
T.E.S.T.	0D,1D,2D	Constitutional descriptors, Topological descriptors, Chi Connectivity Indices, Kappa Shape Indices, Electrotopological State Indices, Molecular Fragments, 2D Molecular Properties, Information Indices, Molecular distance edge vector, Burden eigenvalue descriptors,	797

		Walk and Path counts 2D Autocorrelation descriptors, Molecular properties (such as the octanol-water partition coefficient).	
PADEL	2D	Acidic group count, ALOG, APol, Aromatic atoms count, Aromatic bonds count, Atom count, Autocorrelation, Barysz matrix, Basic group count, BCUT, Bond count, BPol, Burden modified eigenvalues, Carbon types, Chi chain, Chi cluster, Chi path cluster, Chi path, Constitutional, Crippen logP and MR, Detour matrix, Eccentric connectivity index, Atom type electrotopological state, Extended topochemical atom, FMF Descriptor, Fragment complexity, H-bond acceptor count, H-bond donor count, Hybridization ratio, Information content, Kappa shape indices, Largest chain, Largest Pi system, Longest aliphatic chain, Mannhold LogP, McGowan volume, Molecular distance edge, Molecular linear free energy relation, Path counts, Petitjean number, Ring count, Rotatable bonds count, Rule of five, Topological, Topological charge, Topological distance matrix, Topological polar surface area, Van der Waals volume, Vertex adjacency information (magnitude), Walk counts, Weight, Weighted path, Wiener numbers, XLogP, Zagreb index	1444
		TOTAL (Modelling)	4837
Fingerprints descriptors of PADEL	2D	FP (1024), ExtFP (1024), Estate-FP (79), GraphFP (1024), MACCSFP (166), Pubchem FP (880), SubFP (30)	4227

After clustering, the molecules in each cluster were manually distributed to the training and test set by keeping the ratio approximately, 75%-25% in terms of active and inactive class distribution. The distribution of the 121 chemicals into training and test sets and their class distribution as active (comedogenic) and inactive (non-comedogenic) are given in Appendix 8.2_Table S2A and TableS2C. As well, summary of table distribution for comedogenicity are given in Table 18. Furthermore, their distribution for skin irritation are given in Table 32.

3.2.4 Training/Test Set Division

Homogeneity and representativeness are important for the division of training and test set in order to develop a predictive OSAR model. The selection of the training set is crucial and all chemical classes in the dataset should be represented in the training set (127). Therefore, our dataset was divided into clusters for the selection of compounds into the training and test set.

3.2.4.1 Clustering

Clustering algorithms enable us to obtain homogeneous clusters of molecules. In this way, similar training and test set would be possible. K-means clustering algorithm was shown to be an effective method in the division of dataset into the training and test set (128). Therefore, in this study, k-means clustering was chosen for the data clustering and performed using IBM SPSS v.25 (129). In the clustering studies, three different types of variables which are all of the descriptors (N=1819 descriptors of Alva model), the functional group descriptors (N=154 descriptors of Alva), all of the fingerprints (N=4504 descriptors of PaDEL (FP(n=1024), ExtFP(n=1024), EstateF(n=79), GraphFP (n=1024), MACCSFP (n=166), Pubchem FP (n=880), SubFP (n=30)) (124) were tested separately as input variables with the clustering number of 10, 9, 8, 7, 6, 5, 4, 3, 2. The Elbow method was used to determine the optimal number of clusters (k) for each cluster based on all descriptors, the functional group descriptors and all fingerprints respectively (130). After the optimal value of cluster (k) was chosen for each of them, the quality of the three optimal clusters was compared by measuring the intra-cluster distance and inter-cluster distance (131,132). And then, the best optimal cluster (k) was determined for the selection of compound to obtain a balanced training and test set.

3.2.4.2 Splitting dataset into training and test set

After the clustering study, the comedogenicity and irritation dataset was split into training and test set by manually selecting the compounds from each cluster determined according to obtained best optimum cluster number (k). In this selection, %75 of the molecules in each cluster were distributed to the training and the remaining 25% of them to the test set by considering balanced between the active and inactive class distribution ratio.

3.2.5 The Selection of Descriptors for the QSAR Classification Model

Feature (attribute) selection is the most important process in order to build a predictive and interpretable ML-based QSAR model (30). In the preliminary studies, as a feature selection method, correlation-based (via *CfsSubsetEval*) and learning-based (via *WrapperSubsetEval* with J48 classifier) attribute selection methods of WEKA v.3.8.4 (81, 237) were tested on four descriptor software packages (AlvaDesc, Mold2, TEST and PADEL) and then they were compared to each other on the comedogenicity dataset (N=121) in terms of the number of descriptors being selected automatically by WEKA machine learning software. The descriptor selection process mainly consists of two parts, attribute evaluator and search method (strategy). The setting parameters of the correlation-based and learning-based methods for the descriptors selection of the QSAR model were given in Table 12 (81, 237).

Table 12. The comparison of setting parameters of two feature selection methods

	Correlation-based	Learning-based
Attribute Evaluator	<i>CfsSubSetEvaluator</i>	<i>WrapperSubSetEvaluator</i>
Classification Algorithm	-	The decision tree J48 classifier (C4.5 algorithm)
Search Strategy	Best First (Forward direction)	Best First (Forward direction)
SubSet Evaluation	Local predictive ability	Classification accuracy
Accuracy Estimation	-	5-fold-cross validation
Learning scheme	-	Weka. classifiers. trees.j48
Scheme options	-	C0.25-M2

In ML, there is a range of search strategies such as random-hill climbing or heuristic such as forward selection or backward elimination, in order to remove or add features. *Best-First* is a heuristic search algorithm that is used to select top-ranking descriptors (38). In the forward direction, selection begins with an empty set of features, and features are gradually incorporated into the growing subset. On the contrary, the backward direction starts with all the features and gradually eliminates the least ranking ones (38, 133).

3.2.5.1 Correlation-based feature selection method

The correlation-based feature selection method is a kind of filter approach (Figure 61). In the filter approach, the suitability of the features is calculated by considering the intrinsic properties of data. It uses the suitability score of features to select the top-rank features (38). In the correlation-based method in WEKA, the *CfsSubSet* evaluator is chosen as an attribute evaluator (81). It evaluates the worth of a subset of attributes by considering the individual predictive ability of each feature along with the degree of redundancy between them. The subset of features that are highly correlated with the class but with low intercorrelations with each other is preferred (133). Attribute selection involves searching through all possible combinations of attributes in the data to find which subset of attributes works best for prediction (134). In the search process, the search algorithm evaluates the merit of feature subsets. As a search strategy, **Best First (Forward Direction)** starts with an empty set of features and adds features one at a time. In each iteration, the feature that causes the largest increase of the evaluation function is added. The search stops when adding new features does not cause any improvement of the evaluation function (133, 135). As a summary of this descriptor selection process, the workflow of the correlation-based method was given in Figure 61. (133, 136). In this study, correlation based-feature selection was automatically performed by WEKA, according to the following workflow of the correlation-based method in the preliminary studies.

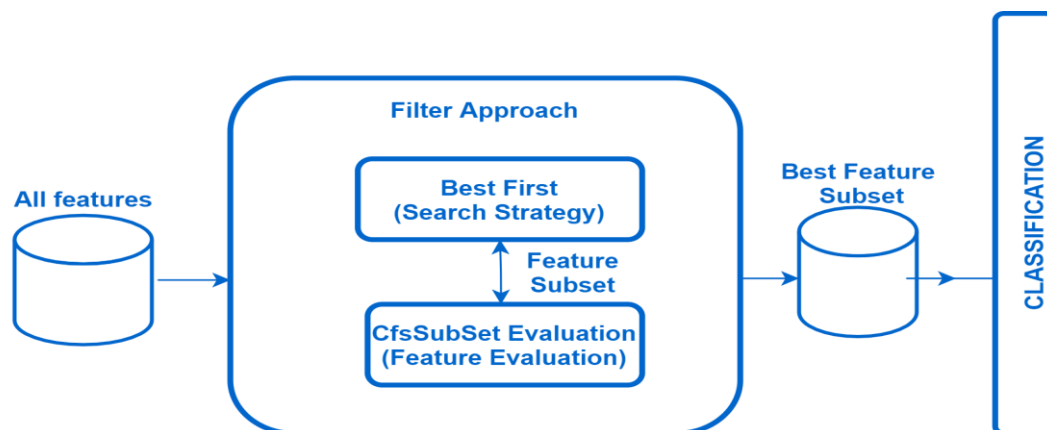


Figure 61. The workflow of the correlation-based method (133, 136)

3.2.5.2 Learning-based feature selection method

The learning-based feature selection method is a kind of wrapper approach since it selects a subset of suitable features using a learning algorithm (Figure 62) (38). In the learning-based method, *WrapperSubSetEvaluator* is chosen as an attribute evaluator with the decision tree, J48 classifier (C4.5 algorithm) as the classification algorithm (81). It selects a feature subset using a learning algorithm as part of the evaluation function (135). The decision tree classifier (J48) evaluates the feature subsets by employing cross-validation to estimate the classification accuracy of the learning scheme (81). In this study, the most significant descriptors (attributes) among the available descriptors were selected via learning-based method using ***WrapperSubsetEval* with J48 classifier** as attribute evaluator and ***Best-First (forward direction)*** as search method. In the attribute evaluation, the J48 classifier was used for building a decision tree in WEKA by the means of the C4.5 algorithm that chooses the attribute with the highest normalized information gain to make the decision (71, 137, 138). As a search strategy, *Best-First(Forward Direction)* started with an empty set of features and added features one at a time. In each iteration, the feature that causes the largest increase in the evaluation function was added. The search stopped when adding new features did not cause any improvement of the evaluation function (133, 135). Briefly, in this study, the descriptor subset selection was automatically performed by WEKA according to the following workflow of the learning-based method in Figure 62 (133, 136).

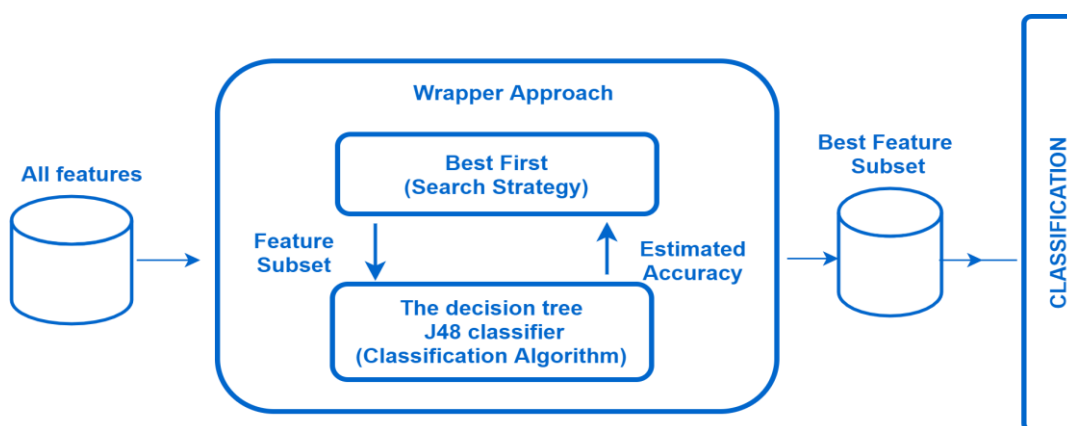


Figure 62. The workflow of learning-based method (133, 136)

The best 2D descriptor subsets were selected separately from the training set for Mold2, AlvaDesc, T.E.S.T, and PaDEL packages. The selected descriptors were then used as input variables for developing the QSAR models.

3.2.6 Evaluation Metrics

All evaluation metrics shown below were used for the evaluation of the performance of the classifiers and the classification models (67, 73, 75, 125). The formula of all these metrics are given in Section 2.2.6.

- Classification accuracy (ACC) is used to measure the overall performance of models for the prediction accuracy of the comedogenic potential. It is the ratio of the correct predictions within all predictions made (75, 125).
- Error rate (ER), known as classification error rate, is the percentage of incorrect classifications.
- Area under the ROC curve (AUC), quantifies the ability to correctly rank compounds according to their comedogenic potential and is also used for classifier comparison.
- Area under the Precision-Recall curve (AU-PRC) is more informative than AUC with imbalanced datasets (67).
- Matthews' correlation (MCC) is a correlation coefficient calculated from all four values of the confusion matrix by considering data imbalance. It is regarded to be one of the best metrics for binary classification performance. A value of 1 indicates the perfect prediction (67, 75).
- Positive predictive value (PPV) is the ratio of true positive predictions among all positive predictions.
- Negative predictive value (NPV) is the ratio of true negative predictions among all negative predictions.
- Sensitivity (SEN) measures the ratio of correctly identified comedogenic molecules.
- Specificity (SPC) measures the ratio of correctly predicted non-comedogenic molecules
- F score is the harmonic mean of precision and recall (67).
- Kappa (Cohen's kappa), measures the agreement of prediction with the true class – 1.0 signifies perfect agreement (77).

3.2.7 Classifier

In data science, a classifier is a machine learning algorithm that categorizes the input data into different class labels (such as toxic or not) based on the set of features (8, 67). The commonly used evaluation metrics to evaluate the performance of the classifiers are ACC, AUC and MCC (67, 74). WEKA is a collection of machine learning algorithms for classification and regression models (7). In this study, the performance of the various classifiers was tested in WEKA (81). Among them, KStar, RF, and NNge were chosen as the top three classifiers in the preliminary study.

KStar is one of the classification algorithms of the lazy learning method. KStar is similar to the K-nearest neighbor (K-NN) algorithm. It uses an entropy-based-distance function and is defined as a method of cluster analysis that targets the partition of n observation into k clusters where each observation belongs to the cluster with the nearest mean (139, 140).

Random Forest (RF) is one of the most popular, tree-based classifiers for classification and regression models. RF uses the ensemble learning method for building a forest of random trees and putting a large number of decision trees together into a model. (141) Since the features are randomly selected in each decision split that improves the prediction power. It is capable of high prediction accuracy and better generalization ability due to the fact that ensemble strategies and random sampling (8, 142, 143).

Non-Nested generalization exemplars (NNge) is a kind of classification algorithm of rule-based classifiers. It is a nearest-neighbor-like algorithm using non-nested generalized exemplars which are hyperrectangles. Therefore, it classifies using the collection of if-then rules. NNge forms a generalization each time a new example is inserted into the database, by joining it to its nearest neighbor of the same class (144, 145).

3.2.8 Model Development

In the model development studies, WEKA v. 3.8.4 was used for building ML-based 2D QSAR classification models for comedogenicity prediction. WEKA is a user-friendly, freely available machine learning software and consists of various ML algorithms (81). Classification is a supervised ML learning technique that builds a classification model to predict a class label, such as whether a chemical is toxic or not, or whether a certain picture is a car or not (10, 53). In this study, we aimed to develop a 2D QSAR classification model by using different machine learning algorithms (classifiers) and types of molecular descriptors. For this purpose, firstly, various classifiers were tested with our QSAR ready dataset, and the top three classifiers (KStar, RF, NNge) were chosen for the modeling study. Secondly, a total of 12 models were developed by using the top three classifiers (KStar, RF, NNge) with the selected descriptors for four descriptor packages (AlvaDesc, Mold2, TEST, and PaDEL). Only training set chemicals were used to train the model and to select the molecular descriptors. Thirdly, internal and external validation of developed models were performed. Finally, validated QSAR models were evaluated mainly by the means of classification accuracy, error rate, AUC, MCC, and other metrics, and then all models were compared. The best model and classifier were chosen accordingly. The test set chemicals were used to evaluate the predictive ability of the trained models. The QSAR model development process for comedogenicity prediction is summarized in Figure 63.

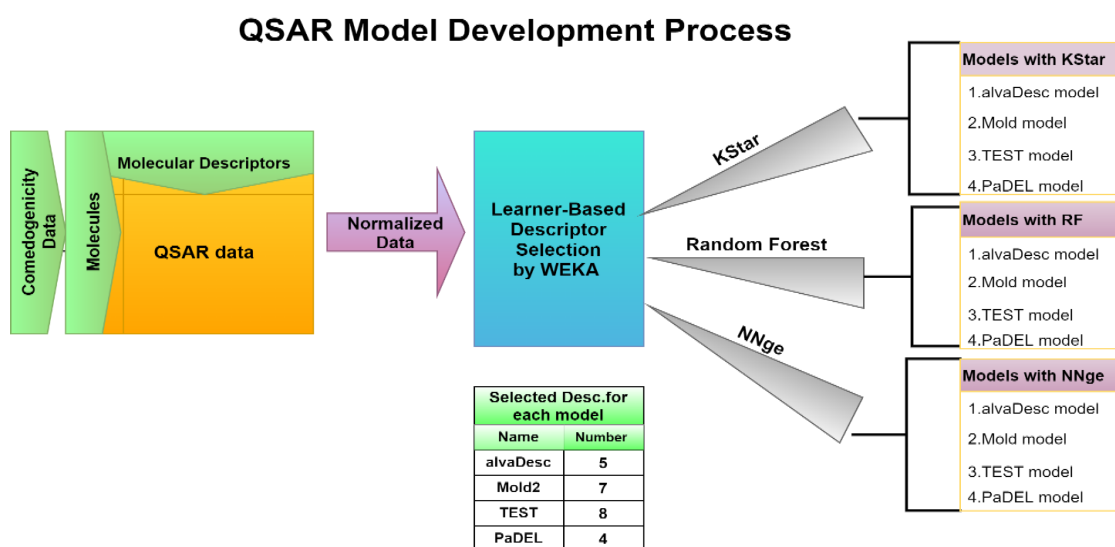


Figure 63. The QSAR model development process for comedogenicity prediction.

3.2.9 Model Validation

According to the OECD guideline, validation of the QSAR model can be performed by internal and external validation. In this study, the robustness of the developed ML-based QSAR models was evaluated by using 10-fold cross-validation as internal validation. (8, 34,146). In 10-fold cross-validation (Figure 64), data used for model development (training set) is split into “ten-parts (10% each)”, and then, the iteratively, last one part (10%) is used for testing, the other remaining nine parts are used as training data. In each iteration, one part for testing (blue) is shifted to the left. This process is repeated 10 times with a different testing test for evaluation each time. In this way, a total of 10 different ML models are trained and evaluated ($E_1, E_2, E_3, \dots, E_{10}$). In the end, the final performance is evaluated as the average of all 10 iterations (E). For example, the final accuracy of the cross-validated model is the average accuracy of ten iterations (81, 147, 148). The performance of the classification models was verified by comparing the cross-validated error rate (31, 34). As external validation, the predictive ability of the developed QSAR models was evaluated on the test set by applying the test set chemicals which never used in the model development (34). The predictive ability of classification models was verified by comparing the test error rate. The first condition for model validity is that the Topliss ratio should be a ratio of at least 5 to 1 between the number of chemicals in the training set and the descriptors used in the model (34, 126, 149). As well, in this study, the Topliss ratio was calculated for each developed QSAR model to verify model validity.

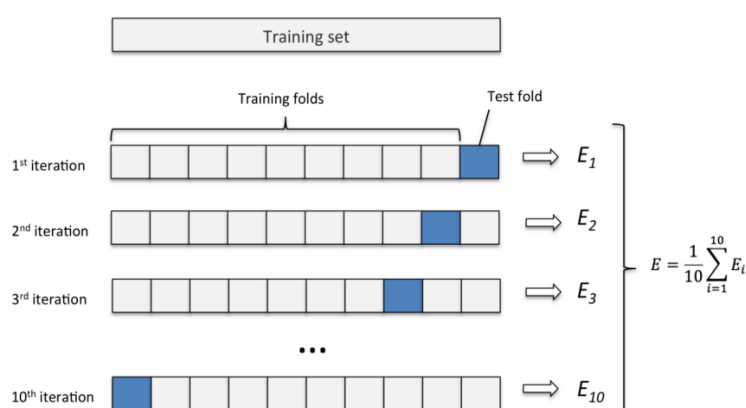


Figure 64. The diagram of k-fold cross-validation with k=10 for ML model as internal validation (147).

3.2.10 Performance Evaluation of the Classifiers and QSAR Models

The cross-validated results were used for a proper performance evaluation (146). The performance of the various classifiers was evaluated by comparing the cross-validated results of the commonly used evaluations metrics, which are ACC, AUC, and MCC, and the best classifier was chosen accordingly (34, 67, 74). The goodness-of-fit of a classification QSAR model can be assessed in terms of its Cooper statistics that are accuracy, PPV, NPV, SEN, and SPC (31). The performance of the classification models was evaluated by comparing the cross-validated results and Cooper statistics (34), and the best model was chosen accordingly. The predictive ability of the classification models was evaluated by comparing the results for the test set and Cooper statistics for the test set (31).

The classification model can be characterized as “excellent” if $AUC = 0.9-1.0$; as “good”, if $AUC = 0.8-0.9$; as “fair”, if $AUC = 0.7-0.8$; as “poor”, if $AUC = 0.6-0.7$ and as “fail”, if $AUC = 0.5-0.6$. (125). The best model was characterized according to the AUC results of cross-validated models.

3.2.11 Optimization of the Model and Activity Prediction

After the successful models were determined, the best QSAR model was selected and optimized. In the optimization process, the best model was optimized by removing irrelevant descriptors from the model according to the IGA score. After optimization, the predictive ability of the QSAR model chosen best was evaluated on the test set for each of ten chemical classes.

3.2.12 Applicability Domain

The applicability domain (AD) is necessary to make a reliable prediction, determine which types of chemicals and toxicity endpoints can be applied to the model (28). The chemical structure space defined the AD in classification. Also, chemical structure space is expressed with molecular descriptors. Therefore, the AD can be defined in terms of molecular descriptors. The reliable prediction can be predicted within the molecular descriptor space covered by the training set compounds. Therefore, the border of the AD can be found by determining the outliers (150). In this study, we applied the boxplot method by the means of SPSS in order to detect outliers in the training set (151). For each descriptor (X_i) in the Alva model with 3 descriptors, the maximum ($X_{\max}$) and minimum ($X_{\min}$) values of the descriptors in the training set were found (Table 13). After that, the dataset was normalized to the range (0 to 1) according to the $X_{\max}$ and $X_{\min}$ of the value of the training set using the formula below (152). The boxplot of the training set was obtained using SPSS for each normalized descriptor ($X_{\text{normalized}}$). The extreme outliers (*) in the boxplots were detected for each descriptor. The extreme outliers were represented by (*) those who were 3 times less or greater than interquartile range (IQR)=H (151). In order to determine the outside of AD, test set compounds normalized according to the $X_{\max}$ and $X_{\min}$ value of the training set for each descriptor in the model, and then, the boxplot of the test set was obtained using SPSS (129). The extreme outliers (*) in the boxplot of the test set were determined as outside of AD. Furthermore, we determined the chemical classes that composing of our training to describe AD via these chemical classes.

$$X_{\text{normalized}} = \frac{X - X_{\min}}{X_{\max} - X_{\min}} \quad (152)$$

Table 13. The $X_{\max}$ and $X_{\min}$ value of each descriptor (X_i) in the training set used for calculating the normalized form of descriptors

Training Set	ALVA_S3K (X1)	ALVA_ATSC7i (X2)	ALVA_CATS2D_08_LL (X3)
Xmax	80.078	4.233	47.000
Xmin	0.825	0.003	0.003
Test Set	ALVA_S3K	ALVA_ATSC7i	ALVA_CATS2D_08_LL
Xmax	65.319	5.694	54.000
Xmin	2.974	0.000	0.003

3.2.13 Evaluation of the Relationship Between Comedogenicity and Skin Irritation

In data analysis, if there are two categorical variables in the dataset, the Chi-Square test is used to determine the relationship between these two categorical variables (153). Therefore, in this study, two categorical variables, comedogenicity and irritation were described as whether a cosmetic ingredient is comedogenic “Yes” or non-comedogenic “No”, as well, as whether a cosmetic ingredient is irritant “active” or not irritant “inactive”. And then, to evaluate the relationship between the comedogenic and irritant potential of 121 cosmetic ingredients, the Chi-Square test (X^2) was performed using IBM SPSS statistics v.28. (129). A frequency table was used as input for the Chi-Square test in SPSS. Therefore, a 2X2 frequency table was created as shown in Table 14. It was an effective way of finding dependence or lack of it between the two categorical features in the dataset.

- Null Hypothesis (H_0): The two variables comedogenicity and irritation were independent of each other. There was no association between the two variables.
- Alternative Hypothesis (H_1): The two variables comedogenicity and irritation were associated with each other.
- For the interpretation of the results: If the** p-value was less than 0.05, it was rejected the null hypothesis (H_0) that comedogenicity was not associated with the irritation and the alternative hypothesis (H_1) was accepted.

** p-value: The significance value (p-value) is the probability that a result occurred by chance. The p-value is compared to a predetermined significance level to find whether a test is statistically significant (129). If a p-value is found less than the significance level (such as 0.05 or 0.001), the test is accepted as statistically significant (129).

Table 14. Frequency table as an input for Chi-Square test.

		Irritation		
		Active	Inactive	Total
Comedogenicity	Comedogenic	41	33	74
	Non-comedogenic	3	44	47
	Total	44	77	121

Cramer's V measures how strongly two categorical variables are associated with each other. It is a number between 0 and 1. For interpretation of association according to the value of Cramers V (Table 15) was given in the following (129).

Table 15. Interpretation of Cramer's V value (154).

Value of Cramer's V	Interpretation of association
0-0.1	No or negligible relationship
0.1-0.2	Weak association.
0.2-0.4	Moderate association.
0.4-0.6	Relatively strong association
0.6-0.8	Strong association.
0.8-1.0	Very strong association.

4. RESULTS

4.1 Comedogenicity Studies

4.1.1 The Training and Test Sets for Comedogenicity

4.1.1.A Clustering Studies

In the clustering studies, three different types of clusters were obtained based on all of the molecular descriptors (0D, 1D, 2D) in AlvaDesc software (n=1819), the functional group descriptors of AlvaDesc software (n=154), and all of the molecular fingerprints in PADEL software (n=4504). The optimal number of clusters (k) were found from the “elbow of the graphs” shown in Figure 65, Figure 66, Figure 67 for three cluster types. In the Elbow method, the sum squared error (SSE) per cluster was calculated from ANOVA results until the maximum number of clusters had been determined. In the graph, the plotting the SSE as a function of the number of clusters, and by comparing the difference SSE of each cluster, the optimal cluster number was shown by the most extreme difference forming the angle of the elbow. In Figure 65, k=6 was the elbow of the graph of cluster based on functional group descriptors that show the optimal cluster number.

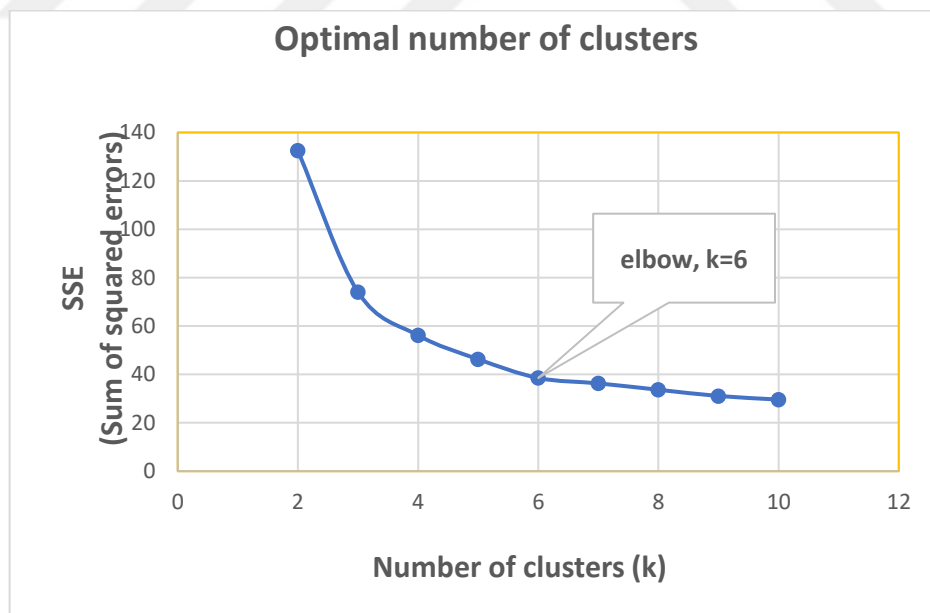


Figure 65. The graph of the optimal number of clusters (k=6) based on functional group descriptors

k	SSE
2	132,452
3	74,043
4	56,170
5	46,294
6	38,529
7	36,306
8	33,677
9	31,091
10	29,619

In Figure 66, k=5 was the elbow of the graph of cluster based on all the descriptors (for the AlvaDesc model) that show the optimal cluster number.

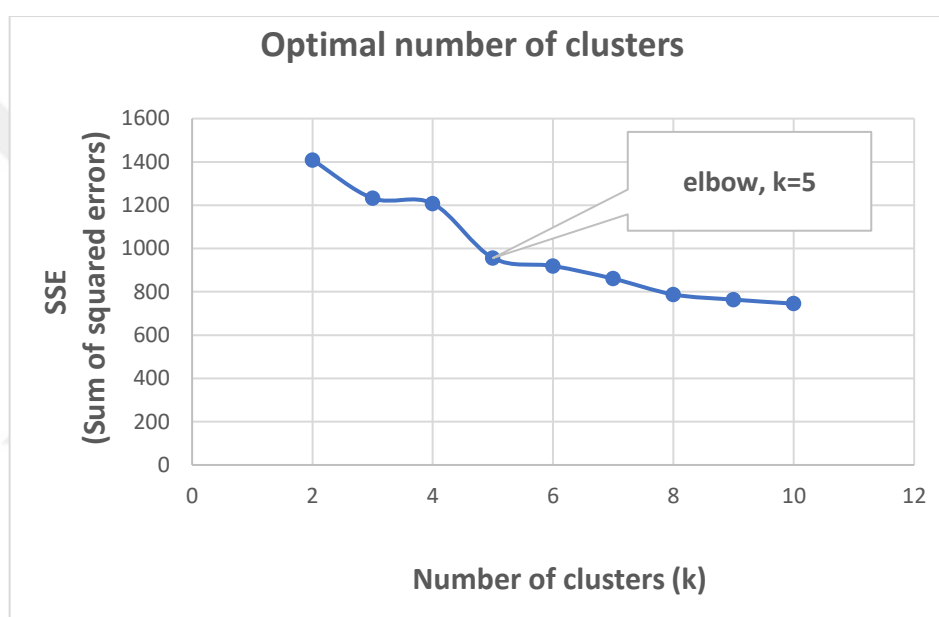


Figure 66. The graph of the optimal number of clusters (k=5) is based on all the descriptors in Alva software packages.

K	SSE
2	1408,9
3	1232,1
4	1207,3
5	955,91
6	919,22
7	860,39
8	787,54
9	763,7
10	745,07

In Figure 67, $k=5$ and $k=7$ was the elbow of the graph of cluster based on all fingerprints that show the optimal cluster number.

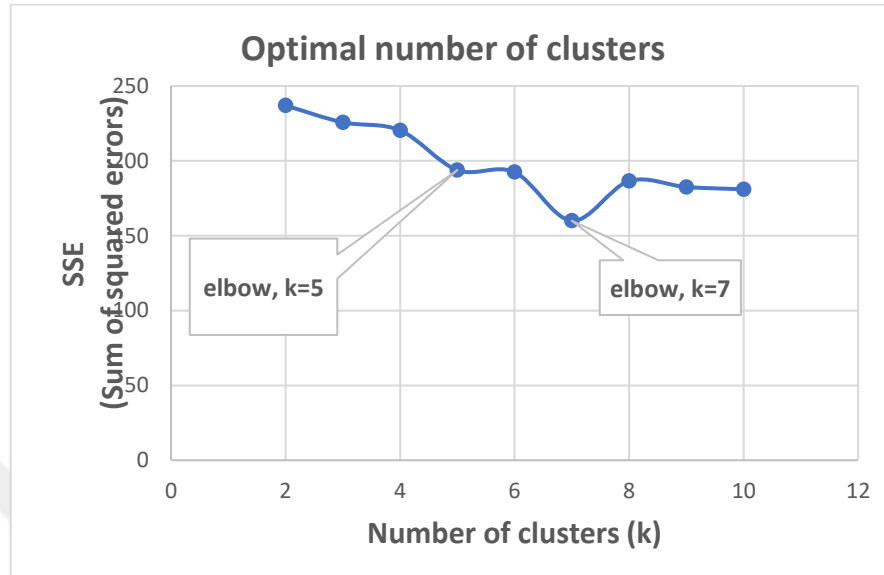


Figure 67. The graph of the optimal number of clusters ($k=5$ and $k=7$) based on all fingerprints

k	SSE
2	236,983
3	225,749
4	220,371
5	193,834
6	192,449
7	159,985
8	186,512
9	182,538
10	180,982

The quality of the three different optimal clusters was compared by measuring the intra-cluster variance and inter-cluster distance. The results of intra-cluster variance for three optimal clusters were shown in Table 16.

Table 16. Intra-cluster variance for three optimal cluster

Cluster-based on functional group descriptors k=6			Cluster-based on all descriptors (Alva model) k=5			Cluster-based on all fingerprints k=5		
Cluster no	Variance	Cluster size (n)	Cluster no	Variance	Cluster size (n)	Cluster no	Variance	Cluster size (n)
1	8.18	8	1	118.74	12	1	9.83	87
2	0.00	10	2	83.57	23	2	11.20	7
3	3.39	34	3	40.54	4	3	21.87	13
4	12.08	8	4	88.03	12	4	13.62	16
5	8.58	52	5	122.38	74	5	0.00	2
6	10.47	13						
Weighted average of intra-cluster variance		6.88		108.97			11.49	

Additionally, the results of inter-cluster distance for three optimal clusters were shown in Table 17.

Table 17. Inter-cluster distance for three optimal cluster

Functional group	min.	10.93
	max	12401.39
	average inter-cluster distance	4146.33
All descriptors	min.	45.232
	max	87.04
	average inter-cluster distance	63.10
All fingerprints	min.	13.78
	max	28.23
	average inter-cluster distance	22.99

Good clustering minimizes intra-cluster variance and maximizes inter-cluster distance. When comparing the results of intra-cluster variance, the optimal cluster (k=6) based on functional group descriptors gave the minimum variance. Moreover, comparing the results of average inter-cluster distance, the optimal cluster (k=6) based on functional group descriptors gave the maximum distance. Therefore, cluster (k=6) based on functional group descriptors was chosen as the best optimal cluster for the division of the dataset into training and test set. The cluster members of the comedogenicity study were given in the Appendix 8.2_Table S2B

4.1.1.B Division of the dataset into training and test set for comedogenicity

After clustering, the molecules in each cluster were manually distributed to the training and test set by keeping the ratio approximately, 75%-25% in terms of active and inactive class distribution. The distribution of the 121 chemicals into training and test sets and their class distribution as active (comedogenic) and inactive (non-comedogenic) are given in Table 18.

Table 18. The distribution of comedogenicity dataset into training and test set

	Active (Comedogenic)	Inactive (Non-comedogenic)	Total	Active/Inactive
Training set	55	37	92	1.49
Test set	19	10	29	1.9
Total	74	47	121	

Training and test with their comedogenic activity data (active or inactive) used for modeling are listed in Appendix 8.2_Table S2A and Table S2C. They were given as comedogenic activity and their grade (A) and training and test set (C) for comedogenicity. Furthermore, the distribution of active/inactive compounds in the ten chemical classes was shown as the stacked bar charts for the dataset, training, and test set. The stacked bar chart of the distribution of ten chemical classes in the comedogenic dataset, training, and test set is given in Figure 68, Figure 69 and Figure 70 respectively.

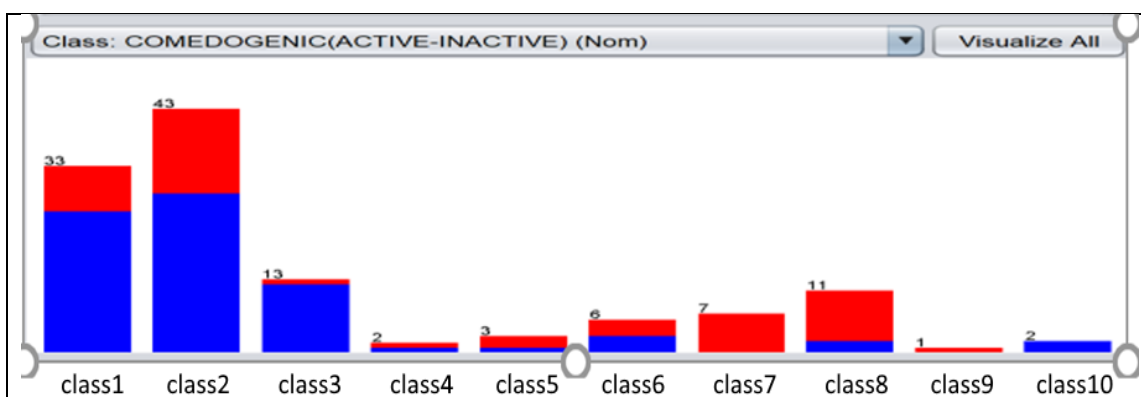


Figure 68. DATASET (n=121) The stacked bar chart of ten chemical class distribution in the dataset in terms of comedogenic active/inactive compounds (Blue= Comedogenic (Active), Red= Non-Comedogenic (Inactive))

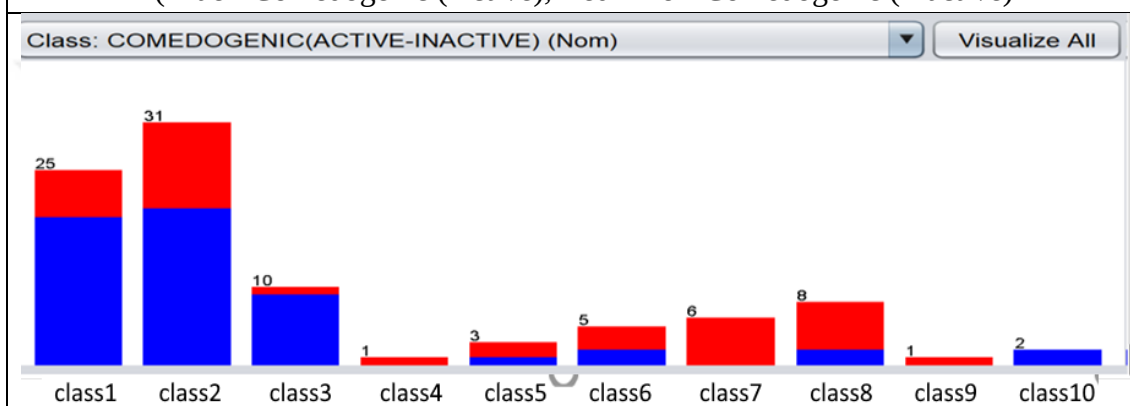


Figure 69. TRAINING SET (n=92) .The stacked bar chart of ten chemical class distribution in the training set in terms of active/inactive compounds (Blue= Comedogenic (Active), Red= Non-Comedogenic (Inactive))

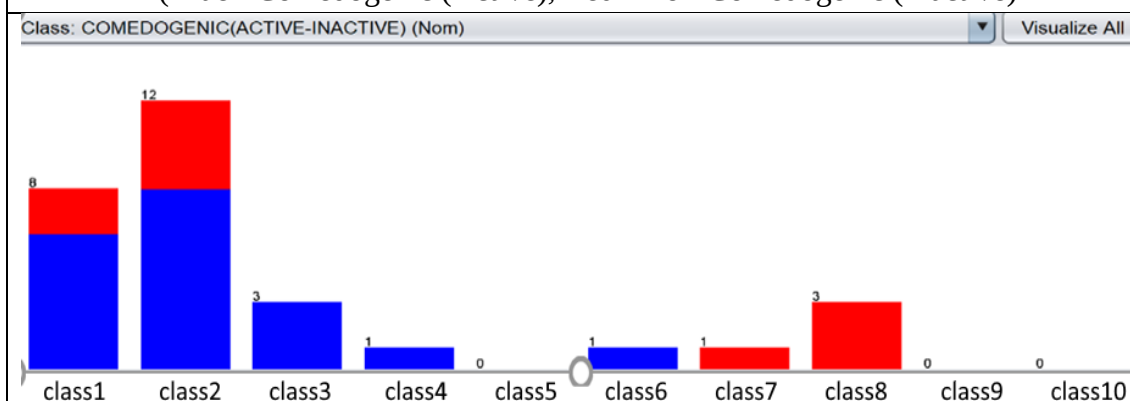


Figure 70. TEST SET (n=29). The stacked bar chart of ten chemical class distribution in the test set in terms of active/inactive compounds (Blue= Comedogenic (Active), Red= Non-Comedogenic (Inactive))

class1: fatty acid and derivatives, class2: alcohols-fatty alcohols-sugars and their derivatives, class3: pigments, class4: silicones, class5: sterols, class6: vitamins, class7: preservatives & additives, class8: miscellaneous, class9: thickeners, class10: oils.

4.1.2 QSAR Models for Comedogenicity

In the preliminary study, the correlation-based feature selection was compared to learning-based feature selection in terms of the number of selected descriptors and model performance. The results for QSAR model via both two method were given below for comedogenicity.

4.1.2.1 Results of Correlation-based Feature Selection Method

4.1.2.1.A The selected descriptors for comedogenicity models

The number of selected attributes for each comedogenicity model built with AlvaDesc, Mold2, TEST, and PADEL descriptor packages via correlation-based method were given respectively in Table 19.

Table 19. The selected attributes by WEKA via the correlation-based method.

AlvaDesc	Mold2	TEST	PADEL
16	17	10	26
ALVA_H%	MLD2_D123	TST_xvp9	PDL_ATS2p
ALVA_NNRS	MLD2_D186	TST_SsssCH_acnt	PDL_AATS2m
ALVA_BAC	MLD2_D187	TST_icycem	PDL_AATS3m
ALVA_SpMaxA_D/Dt	MLD2_D199	TST_MDEC22	PDL_AATS1s
ALVA_ATSC4i	MLD2_D211	TST_BEHe3	PDL_AATSC2c
ALVA_MATS1e	MLD2_D215	TST_BAC	PDL_AATSC1e
ALVA_GATS4m	MLD2_D276	TST_Ms	PDL_GATS5c
ALVA_GATS5e	MLD2_D289	TST_nDB	PDL_GATS6c
ALVA_SpMax3_Bh(e)	MLD2_D429	TST_MATS7v	PDL_VE2_Dzv
ALVA_SpMin4_Bh(i)	MLD2_D477	TST_-C(=O)-[nitrogen aliphatic attach]	PDL_BCUTw-1l
ALVA_P_VSA_MR_7	MLD2_D482		PDL_BCUTc-1h
ALVA_Eta_alpha_A	MLD2_D531		PDL_nBondsD
ALVA_Eta_C_A	MLD2_D561		PDL_SpMin4_Bhi
ALVA_SpMAD_EA(dm)	MLD2_D597		PDL_SpMax4_Bhs
ALVA_NsssCH	MLD2_D625		PDL_VCH-5
ALVA_SHED_DA	MLD2_D722		PDL_ASP-5
	MLD2_D775		PDL_ASP-6
			PDL_VP-7
			PDL_SHBd
			PDL_SsssCH
			PDL_meanl
			PDL_hmin
			PDL_ETA_EtaP
			PDL_ZMIC4
			PDL_VE3_D

4.1.2.1.B Classification models for comedogenicity

The classification models obtained via the correlation-based method for AlvaDesc, Mold2, TEST, and PADEL descriptor packages were given respectively in Table 20.

Table 20. The classification models obtained via the correlation-based method

Model	Classifier	Data	Number of selected descriptors	*Train/Test	ACC (%)	MCC	AUC	F score
AlvaDesc	RF	Original	16	Train	100%	1,00	1,00	1,00
				CV (10-fold)	82,61%	0,64	0,89	0,83
				Test	72,41%	0,46	0,80	0,73
AlvaDesc	RF	Normalized	16	Train	100%	1,00	1,00	1,00
				CV (10-fold)	83,70%	0,66	0,88	0,84
				Test	86,20%	0,70	0,89	0,86
Mold2	RF	Original	17	Train	100%	1,00	1,00	1,00
				CV (10-fold)	84,78%	0,68	0,89	0,85
				Test	75,86%	0,45	0,82	0,76
Mold2	RF	Normalized	17	Train	100%	1,00	1,00	1,00
				CV (10-fold)	85,87%	0,70	0,88	0,86
				Test	51,72%	0,17	0,79	0,51
TEST	RF	Original	10	Train	100%	1,00	1,00	1,00
				CV (10-fold)	85,87%	0,70	0,90	0,86
				Test	72,41%	0,42	0,77	0,73
TEST	RF	Normalized	10	Train	100%	1,00	1,00	1,00
				CV (10-fold)	86,96%	0,73	0,90	0,87
				Test	75,86%	0,45	0,76	0,76
PADEL	RF	Original	26	Train	100%	1,00	1,00	1,00
				CV (10-fold)	85,87%	0,71	0,91	0,86
				Test	65,51%	0,32	0,79	0,66
PADEL	RF	Normalized	25	Train	100%	1,00	1,00	1,00
				CV (10-fold)	83,70%	0,66	0,90	0,84
				Test	72,41%	0,42	0,78	0,73

Abbreviation: RF: Random Forest, CV: cross-validation, ACC: accuracy, MCC: Matthews 'correlation coefficient, AUC: area under the ROC curve,

*Train:92/Test:29, Setting: Default Parameters

**Topliss ratio = Nchemicals of training set/Ndescriptors in the model

4.1.2.2 Results of Learning-based Feature Selection Method

4.1.2.2.A The selected descriptors for comedogenicity models

The number of selected attributes for each comedogenicity model built with AlvaDesc, Mold2, TEST and PADEL descriptor packages via learning-based method were given respectively in Table 21.

Table 21. The selected attributes by WEKA via the learning-based method

AlvaDesc	Mold2	TEST	PADEL
5	7	8	4
ALVA_S3K	MLD2_D130	TST_xvp9	PDL_ATS2m
ALVA_piPC01	MLD2_D338	TST_xvc3	PDL_AATS3m
ALVA_ATSC7i	MLD2_D489	TST_Gmin	PDL_SpMax1_Bhi
ALVA_MATS6p	MLD2_D590	TST_Qsv	PDL_SIC2
ALVA_CATS2D_08_LL	MLD2_D606	TST_ic	
	MLD2_D679	TST_ssi	
	MLD2_D715	TST_piPC04	
		TST_piPC05	

4.1.2.2.B The classification models for comedogenicity

The classification models obtained via learning-based method for AlvaDesc, Mold2, TEST and PADEL descriptor packages were given respectively in Table 22.

Table 22. The classification models obtained via the learning-based method.

Model	Classifier	Data	Number of selected descriptors	*Train/Test	ACC (%)	MCC	AUC	F score
AlvaDesc	RF	Original	5	Train	100%	1,00	1,00	1,00
				CV (10-fold)	83,70%	0,66	0,86	0,84
				Test	58,60%	0,08	0,68	0,59
AlvaDesc	RF	Normalized	5	Train	100%	1,00	1,00	1,00
				CV (10-fold)	84,80%	0,68	0,87	0,85
				Test	79,31%	0,53	0,81	0,78
Mold2	RF	Original	7	Train	100%	1,00	1,00	1,00
				CV (10-fold)	84,78%	0,68	0,89	0,85
				Test	62,10%	0,22	0,72	0,63
Mold2	RF	Normalized	7	Train	100%	1,00	1,00	1,00
				CV (10-fold)	85,87%	0,71	0,89	0,86
				Test	75,86%	0,44	0,76	0,75
TEST	RF	Original	8	Train	100%	1,00	1,00	1,00
				CV (10-fold)	86,96%	0,73	0,89	0,87
				Test	75,86%	0,45	0,86	0,76
TEST	RF	Normalized	8	Train	100%	1,00	1,00	1,00
				CV (10-fold)	86,96%	0,73	0,89	0,87
				Test	69,00%	0,37	0,77	0,70
PADEL	RF	Original	4	Train	100%	1,00	1,00	1,00
				CV (10-fold)	83,70%	0,66	0,88	0,84
				Test	75,86%	0,48	0,88	0,76
PADEL	RF	Normalized	4	Train	100%	1,00	1,00	1,00
				CV (10-fold)	81,52%	0,61	0,88	0,82
				Test	79,31%	0,57	0,88	0,80

Abbreviation: RF: Random Forest, CV: cross-validation, ACC: accuracy, MCC: Matthews
'correlation coefficient, AUC: area under the ROC curve,

*Train:92/Test:29, Setting: Default Parameters

**Topliss ratio = Nchemicals of training set/Ndescriptors in the model

4.1.2.3 Comparison of Correlation-based and Learning-based Method

When the two descriptor selection method, correlation-based method was compared to the learning-based method for comedogenicity, the learning-based method was found successful in terms of the number of reduced descriptors selected and the percent of reduced descriptors and also for the model accuracy results (**Table 20** and **Table 22**). The comparison results for each model were given in Table 23.

Table 23. The results of the comparison of two feature (descriptor) selection method

Model	Descriptor Selection Method	Number of selected descriptors	*Toplis ratio
AlvaDesc	Correlation-based	16	5.75
	Learning-based	5	18.4
	The number of reduced descriptors	11	
	The percent of reduced descriptors	~70%	
Mold2	Correlation-based	17	5.41
	Learning-based	7	13.14
	The number of reduced descriptors	10	
	The percent of reduced descriptors	~60%	
TEST	Correlation-based	10	9.2
	Learning-based	8	11.5
	The number of reduced descriptors	2	
	The percent of reduced descriptors	20%	
PADEL	Correlation-based	26	3.54
	Learning-based	4	23
	The number of reduced descriptors	22	
	The percent of reduced descriptors	85%	

** Topliss ratio = Nchemicals of training set/Ndescriptors in the model

4.1.2.4 The Significant Descriptors Selected After the Comparison of Two Method

In the comedogenicity studies, the most significant descriptors were selected via the learning-based method. They were selected from among the number of calculated descriptors for each software due to the fact that it was found successful according to the method comparison results. The names and the description of the selected descriptors by WEKA for each software were given in Table 24. The original value and normalized value of all selected descriptor were given in Appendix 8.3_TableS3.1 and TableS3.2.

Table 24. The names and the description of the selected descriptors for this study.

Software	Number of calculated descriptors	Selected Descriptor	Description	Block
AlvaDesc	1819	ALVA_S3K	3-path Kier alpha-modified shape index	*Topological indices
		ALVA_piPC01	Molecular multiple path count of order 1	Walk and path counts
		ALVA_ATSC7i	Centred Broto-Moreau autocorrelation of lag 7 weighted by ionization potential	*2D autocorrelations
		ALVA_MATS6p	Moran autocorrelation of lag 6 weighted by polarizability	2D autocorrelations
		ALVA_CATS2D_08_LL	CATS2D Lipophilic-Lipophilic at lag 08	*Pharmacophore descriptors
Mold2	777	MLD2_D130	Number of halogen atoms in each molecule	
		MLD2_D338	Distance+detour path on ring index of order 9 (circuits)	
		MLD2_D489	Moran topological structure autocorrelation length-3 weighted by atomic van der Waals volumes	
		MLD2_D590	Highest eigenvalue from Burden matrix weighted by polarizabilities order-3	
		MLD2_D606	Number of secondary C-sp2	
		MLD2_D679	Number of group ethers (aliphatic)	
		MLD2_D715	Number of group CH2R2	
TEST	797	TST_xvp9	Valence 9th order path chi index	Chi Connectivity Indices-Valence
		TST_xvc3	Valence 3rd order cluster chi index	Electrotopological State Indices-Maximum and Minimum Values
		TST_Gmin	Minimum E-State value in molecule	2-D Molecular Properties
		TST_Qsv	Average of Qs and Qv	Information Indices-Connectivity Based
		TST_ic	Information content	Information Indices-Connectivity Based
		TST_ssi	Standardized information content	
		TST_piPC04	Molecular multiple path count of order 04	Walk and Path Counts
		TST_piPC05	Molecular multiple path count of order 05	Walk and Path Counts
PaDEL	1444	PDL_ATS2m	Broto-Moreau autocorrelation - lag 2 / weighted by mass	Autocorrelation
		PDL_AATS3m	Average Broto-Moreau autocorrelation - lag 3 / weighted by mass	Autocorrelation
		PDL_SpMax1_Bhi	Largest absolute eigenvalue of Burden modified matrix - n 1 / weighted by relative first ionization potential	Burden modified eigenvalues
		PDL_SIC2	Structural information content index (neighborhood symmetry of 2-order)	Information content

In the learning-based method, the decision tree J48 classifier (C4.5 algorithm) used information gain to choose the best attribute. Therefore, the amount of information gain obtained from selected attributes by WEKA was given in Table 25 for all models.

Table 25. The amount of information gain (IGA) of descriptors selected by WEKA

Alvadesc Model		Mold2 Model		TEST Model		PADEL Model	
Selected Attribute	IGA	Selected Attribute	IGA	Selected Attribute	IGA	Selected Attribute	IGA
ALVA_ATSC7i	0.336	MLD2_D130	0	TST_xvp9	0.389	PDL_ATS2m	0.308
ALVA_CATS2D_08_LL	0.314	MLD2_D338	0	TST_xvc3	0	PDL_AATS3m	0.305
ALVA_S3K	0.151	MLD2_D489	0	TST_Gmin	0	PDL_SpMax1_Bhi	0
ALVA_piPC01	0	MLD2_D590	0	TST_Qsv	0.123	PDL_SIC2	0
ALVA_MATS6p	0	MLD2_D606	0	TST_ic	0		
		MLD2_D679	0	TST_ssi	0		
		MLD2_D715	0,166	TST_piPC04	0		
				TST_piPC05	0		
IGA range for all calculated AlvaDesc attributes		IGA range for all Mold2 attributes		IGA range for all TEST attributes		IGA range for all calculated PADEL attributes	
	0.34-0		0,31-0		0.42-0		0.35-0

For the selected descriptors via learning-based method, the correlation (corr.) information between feature and target variable was obtained by the means of “CorrelationAttributeEval+Ranker” which was the attribute evaluation in WEKA. The correlation information with target for the selected descriptors via learning-based method are given in Table 26.

Table 26. Correlation (Corr.) information between feature variable and target variable

Alvadesc Model		Mold2 Model		TEST Model		PADEL Model	
Selected Attribute	Corr.	Selected Attribute	Corr.	Selected Attribute	Corr.	Selected Attribute	Corr.
ALVA_ATSC7i	0,026	MLD2_D130	0,157	TST_xvp9	0,075	PDL_ATS2m	0,005
ALVA_CATS2D_08_LL	0,07	MLD2_D338	0,009	TST_xvc3	0,141	PDL_AATS3m	0,019
ALVA_S3K	0,185	MLD2_D489	0,25	TST_Gmin	0,029	PDL_SpMax1_Bhi	0,046
ALVA_piPC01	0,153	MLD2_D590	0,285	TST_Qsv	0,269	PDL_SIC2	0,307
ALVA_MATS6p	0,052	MLD2_D606	0,149	TST_ic	0,148		
		MLD2_D679	0,008	TST_ssi	0,087		
		MLD2_D715	0,125	TST_piPC04	0,039		
				TST_piPC05	0,021		
Corr.. range for all calculated desp		Corr. range for all calculated		Corr. range for all calculated		Corr. range for all	
	0.45-0		0,41-0		0.44-0		0.45-0

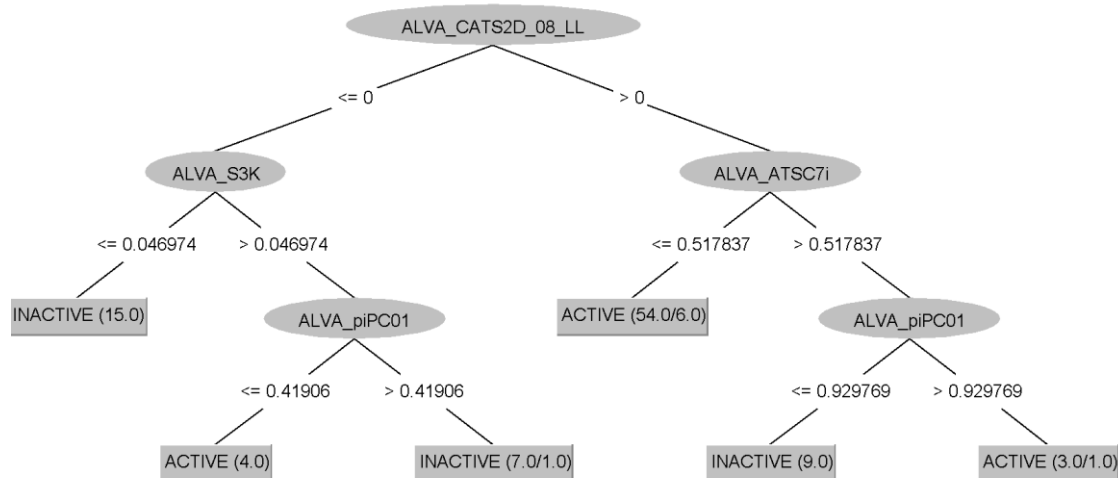


Figure 71. Decision Tree View for selected 5 descriptors for Alva Model (Normalized Training 92)

The decision trees for Alva Model with 5 and 3 descriptors are given in Figure 71 and Figure 72.

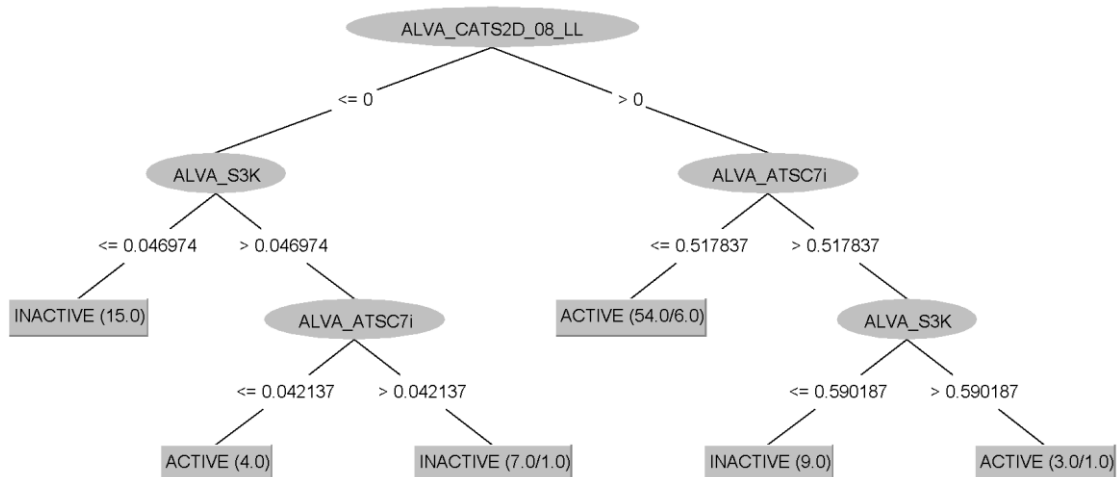


Figure 72. Decision Tree View for 3 descriptors for optimum Alva Model (Normalized Training 92)

4.1.3 Comparison of Classifiers

The performance of the various classifiers used for comedogenicity models in WEKA was tested with our QSAR-ready dataset. Among them, KStar, RF, and NNge were chosen as the top three classifiers in the preliminary tests. And then, a total of 12 models were developed by using these classifiers with four different descriptor packages (AlvaDesc, Mold2, TEST, and PaDEL). The performance evaluation results of the top three classifiers for each model are shown in Table 27.

Table 27. The performance of the top three classifiers for each model

		AlvaDesc Models			Mold2 Models			TEST Models			PaDEL Models		
		ACC (%)	AUC	MCC	ACC (%)	AUC	MCC	ACC (%)	AUC	MCC	ACC (%)	AUC	MCC
KStar	Train	100%	1.00	1.00	100%	1.00	1.00	100%	1.00	1.00	100%	1.00	1.00
	CV (10-fold)	80.44%	0.82	0.59	83.69%	0.87	0.66	75.00%	0.79	0.48	79.35%	0.80	0.56
	Test	75.86%	0.79	0.45	65.51%	0.76	0.27	65.52%	0.74	0.31	72.41%	0.79	0.36
RF	Train	100%	1.00	1.00	100%	1.00	1.00	100%	1.00	1.00	100%	1.00	1.00
	CV (10-fold)	84.78%	0.87	0.68	85.87%	0.89	0.71	86.96%	0.89	0.73	81.52%	0.88	0.61
	Test	79.31%	0.81	0.53	75.86%	0.76	0.44	68.96%	0.77	0.37	79.31%	0.88	0.57
NNge	Train	100%	1.00	1.00	100%	1.00	1.00	100%	1.00	1.00	100%	1.00	1.00
	CV (10-fold)	78.26%	0.78	0.55	80.43%	0.80	0.59	76.10%	0.75	0.50	76.10%	0.75	0.50
	Test	72.41%	0.72	0.42	41.38%	0.51	0.01	68.96%	0.76	0.53	65.52%	0.66	0.31

*All results are given in " weighted average".

Abbreviations: CV: cross-validation, RF: random forest, NNge: non-nested generalized exemplars, ACC: accuracy, AUC: area under the ROC curve, MCC: Matthews' correlation coefficient

For proper evaluation, cross-validated results of the model were used. When comparing the top three classifiers with the commonly used evaluation metrics (ACC, AUC, MCC), all models with RF gave the best results with all evaluation metrics for both the cross-validated models and the test sets. Therefore, RF was chosen as the best classifier in our modelling study.

4.1.4 Comparison of the Classification Models with RF

Since RF was chosen as the best classifier, comedogenicity modelling studies continued with the classification models built using RF. The performance of the classification models with RF was compared by evaluating the results of ACC, ER, AUC, AU-PRC, MCC, PPV, NPV, SEN, SPC, F score, and kappa statistics (Table 28).

Table 28. The comparison of the classification models with RF for comedogenicity

Model	Classifier	Number of descriptors	Train/Test	ACC (%)	ER	AUC	AU-PRC	MCC	PPV	NPV	SEN	SPC	F score	Kappa
AlvaDesc	RF	5	Train	100%	0.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
			CV (10-fold)	84.78%	0.15	0.87	0.86	0.68	0.87	0.81	0.85	0.84	0.85	0.68
			Test	79.31%	0.21	0.81	0.85	0.53	0.78	0.83	0.79	0.65	0.78	0.49
Mold2	RF	7	Train	100%	0.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
			CV (10-fold)	85.87%	0.14	0.89	0.89	0.71	0.88	0.83	0.86	0.84	0.86	0.70
			Test	75.86%	0.24	0.76	0.80	0.44	0.77	0.71	0.76	0.64	0.75	0.42
TEST	RF	8	Train	100%	0.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
			CV (10-fold)	86.96%	0.13	0.89	0.87	0.73	0.86	0.88	0.87	0.84	0.87	0.72
			Test	68.97%	0.31	0.77	0.77	0.37	0.81	0.54	0.69	0.70	0.70	0.36
PaDEL	RF	4	Train	100%	0.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
			CV (10-fold)	81.52%	0.18	0.88	0.87	0.61	0.84	0.78	0.82	0.80	0.82	0.61
			Test	79.31%	0.21	0.88	0.91	0.57	0.88	0.67	0.79	0.80	0.80	0.56

*All results are given in weighted average

CV: cross-validation, ACC: accuracy, ER: error rate, AUC: area under the ROC curve, AU-PRC: area under the Precision-Recall curve, MCC: Matthews' correlation coefficient, PPV: positive predictive value, NPV: negative predictive value, SEN: sensitivity, SPC: specificity.

When comparing all of the classification models, Mold2 and AlvaDesc with RF gave successful results with 85.87% and 84.78% accuracy for the cross-validated models and 75.86% and 79.31% accuracy for the test set. On the other hand, the TEST model showed a good performance for the cross-validated model (86.96 % accuracy), but poor performance for the evaluation on the test set (68.97 % accuracy). Likewise, the PaDEL model did not perform as well as other models for the cross-validated model (81.52%), but performed well for the evaluation on the test set (79.31% accuracy). Furthermore, experimental values for both training set and test set and predicted values for four models (M) for comedogenicity that are $M1_{pred_alvaDesc}$, $M2_{pred_Mold2}$, $M3_{pred_TEST}$, $M4_{pred_PaDEL}$ are given in Appendix 8.4_TableS4.

4.1.5 Optimization of the Model for Comedogenicity

In the Alva model, we have just found the two features, Alva_piPC01 (IGA=0) and Alva_MATS6p (IGA=0), that had no score of information gain, they were considered as irrelevant descriptors. When we removed the irrelevant descriptors from the model, while the accuracy of the test set remained the same (79.31%), the accuracy of the cross-validated model was increased from 84.78% to 86.96%. Thereby, overfitting was reduced and the number of descriptors in the Alva model was reduced from 5 to 3 descriptors (Table 29). As a result, we have improved the performance of the Alva model by removing two irrelevant features. For other models, when we removed the features that had no score of information gain (IGA=0) from the model, overfitting was reduced for each of them. However, both the performance of the model and the predictive ability of the model was reduced for the Mold2, TEST, and PADEL model.

Table 29. Alva model with three descriptors after removing the two irrelevant features.

Descriptors	*IGA	Alva model with 5 descriptors (ATSC7i, CATS2D_08_LL, S3K, piPC01, MATS6p)				Alva model with 3 descriptors (ATSC7i, CATS2D_08_LL, S3K)			
			ACC	AUC	MCC		ACC	AUC	MCC
ALVA_ATSC7i	0.336								
ALVA_CATS2D_08_LL	0.314								
ALVA_S3K	0.151	Train	100%	1.00	1.00	Train	100%	1.00	1.00
ALVA_piPC01	0	CV10-fold)	84.78%	0.87	0.68	CV10-fold)	86.96%	0.85	0.73
ALVA_MATS6p	0	Test	79.31%	0.81	0.53	Test	79.31%	0.75	0.53

*IGA=information gain amount

The comparison of initial models and optimized models was given in Table 30. Therefore, as a result of the comparison of initial and optimized models (Table 30), the Alva model with three descriptors was chosen as the best model in the study.

Table 30. The comparison of initial models with optimized models

		Initial Models			
The number of descriptors	Model	Alva	Mold2	TEST	PADEL
		5desp	7desp	8desp	4desp
Accuracy	Train	100%	100%	100%	100%
	CV-10	84.78%	85.87%	86.95%	83.7%
	Test	79.31%	75.86%	68.96%	75.9%
		Optimized Models			
(The number of descriptors (IGA>0))	Model	Alva	Mold2	TEST	PADEL
		3desp	1desp	2desp	2desp
Accuracy	Train	100%	100%	100%	100%
	CV-10	86.96%	72.83%	78.26%	72.32%
	Test	79.31%	68.97%	62.1%	68.96%

4.1.6 Predictive Ability of Optimized QSAR Model for Comedogenicity

After the optimization of the comedogenicity models, the Alva model with three descriptors was chosen as the best model in the study. Therefore, the predictive ability of the best QSAR model which is the Alva model with 3 descriptors, was evaluated on the test set (N=29) for each of ten chemical classes. The results of true/false prediction of QSAR model _{optimized AlvaDesc} on the test set for each of ten chemical classes was given in Figure 73. (Total= 29 Prediction [23 True/ 6 False])

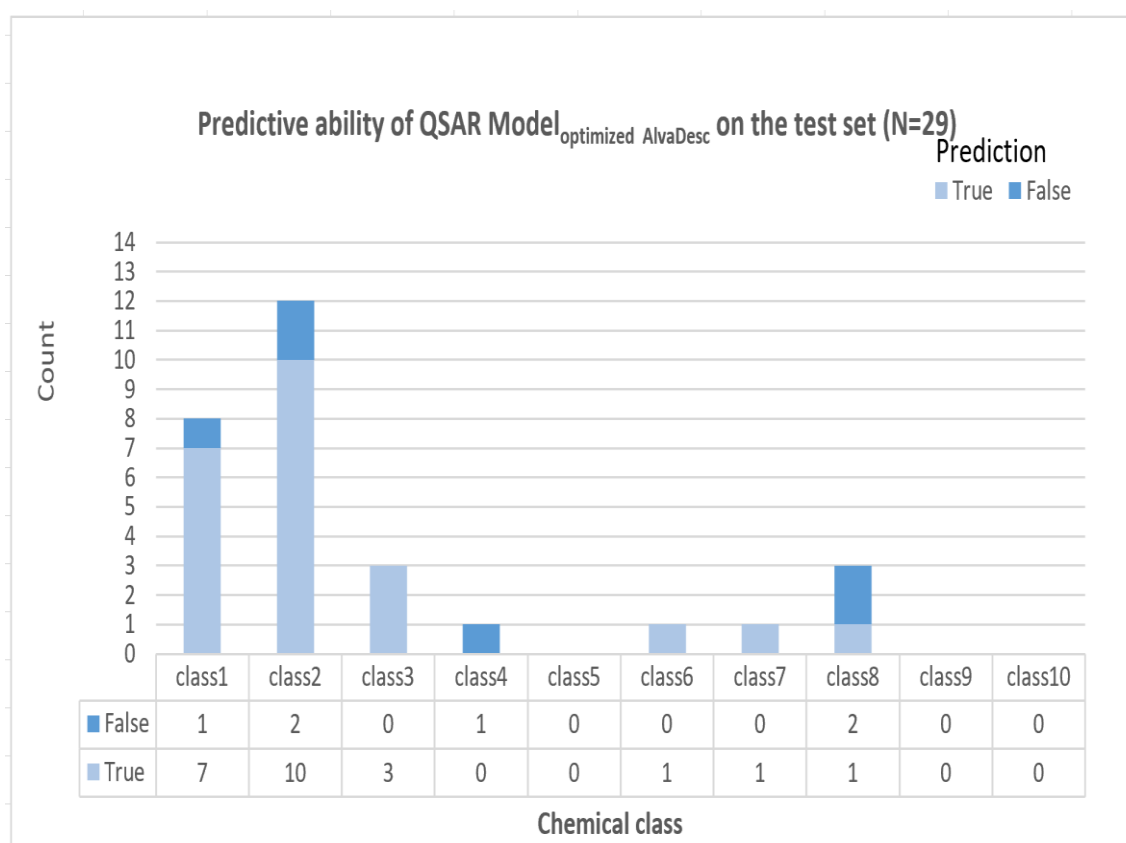


Figure 73. The stacked bar chart of true/false prediction of QSAR Model optimized AlvaDesc on the test set (N=29) for each of ten chemical classes.

class1: fatty acid and derivatives, class2: alcohols-fatty alcohols-sugars and their derivatives, class3: pigments, class4: silicones, class5: sterols, class6: vitamins, class7: preservatives & additives, class8: miscellaneous, class9: thickeners, class10: oils

4.1.7 Applicability Domain for Comedogenicity Model

The border of AD was determined only for Alva Model with three descriptors which is the best model in the study. According to the boxplot method, three compounds in the training set were found as outliers (ID 63, ID115, and ID78) and two compounds in the test set were found as out of AD (ID92 and ID35). The name of the outliers and out of AD were given in Table 31.

Table 31. The outliers and out of AD for Alva Model optimized with 3 descriptors

Outliers (3 compounds)	Out of AD (2 compounds)
ID63: Laureth-23 (Figure 74 &Figure 75)	ID92: Pentaerythrical tetraisostearate (Figure 80 & Figure 82)
ID115: Steareth-100 (Figure 74 &Figure 75)	ID35: Diisopropyl dimerate (Figure 82)
ID78: Octyldodecylstearoyl stearate (Figure 79 & Figure 82)	

The normal Q-Q plot of training data of $S3K_{normalized}$, $ATSC7i_{normalized}$, and $CAT2D_08-LL_{normalized}$ descriptor was given in Figure 74, Figure 76, Figure 78 respectively. Also, the boxplot graph of $S3K_{normalized}$, $ATSC7i_{normalized}$, $CAT2D_08-LL_{normalized}$ descriptor was given in Figure 75, Figure 77 and Figure 79 respectively. In each figure, extreme values were represented by (*) those who were 3 times greater than the interquartile range (IQR) in the boxplots.

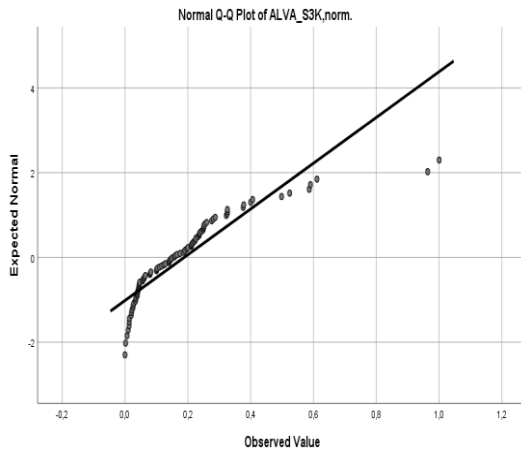


Figure 74. The q-q plot of Alva_S3Knormalized training data (n=92)

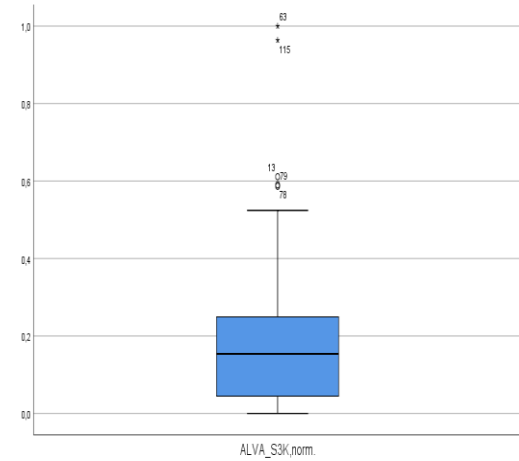


Figure 75. The boxplot graph of Alva_S3Knormalized. training data (n=92)

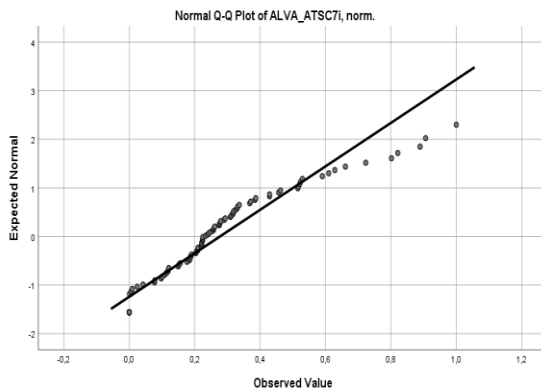


Figure 76. The q-q plot of Alva_ATSC7i normalized training data (n=92)

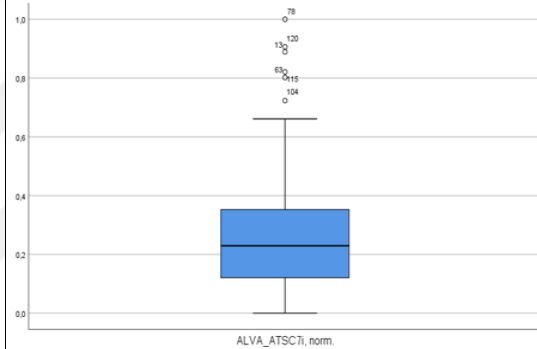


Figure 77. The boxplot graph of Alva_ATSC7inormalized training data (n=92)

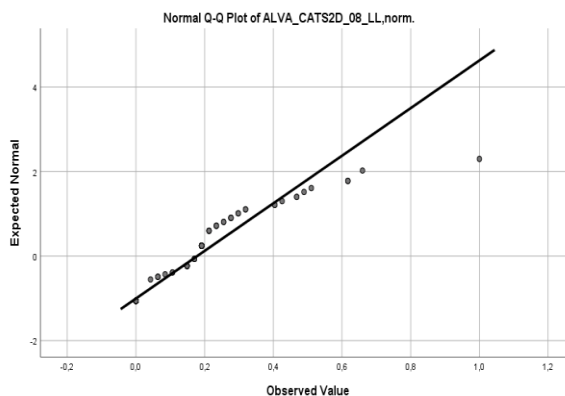


Figure 78 The q-q plot of Alva_CATS2D_08_LLnormalized training data (n=92).

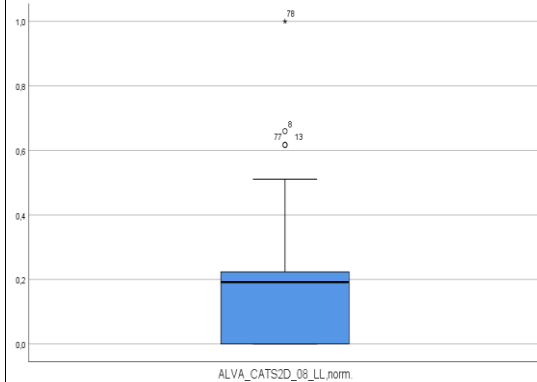


Figure 79 The boxplot graph of Alva_CATS2D_08_LLnormalized training data (n=92)

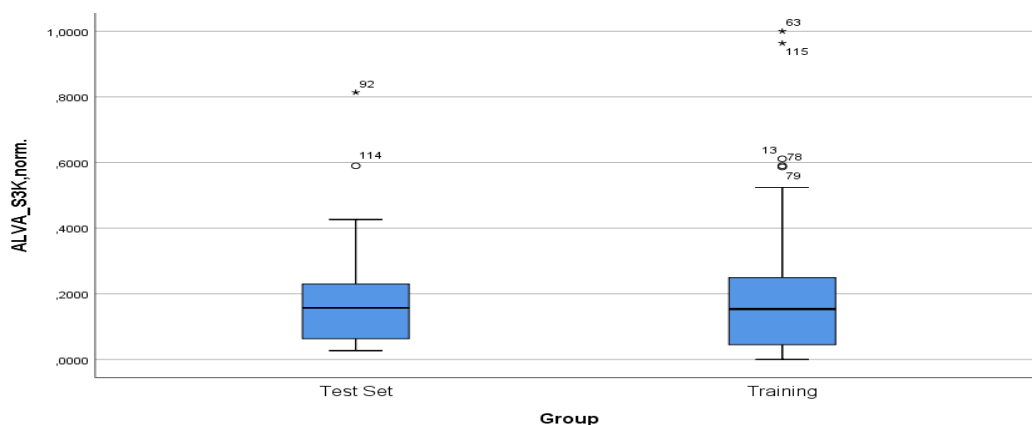


Figure 80. The comparison of the boxplot graph of training and test set to determine the outside of AD for Alva_S3Knormalized descriptor in the Alva model.

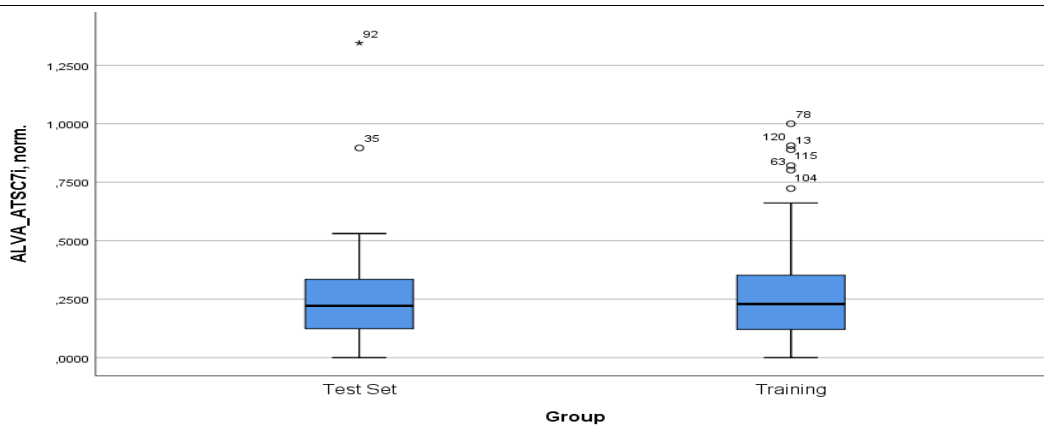


Figure 81. The comparison of the boxplot graph of training and test set to determine the outside of AD for Alva_ATSC7inormalized descriptor in the Alva model.

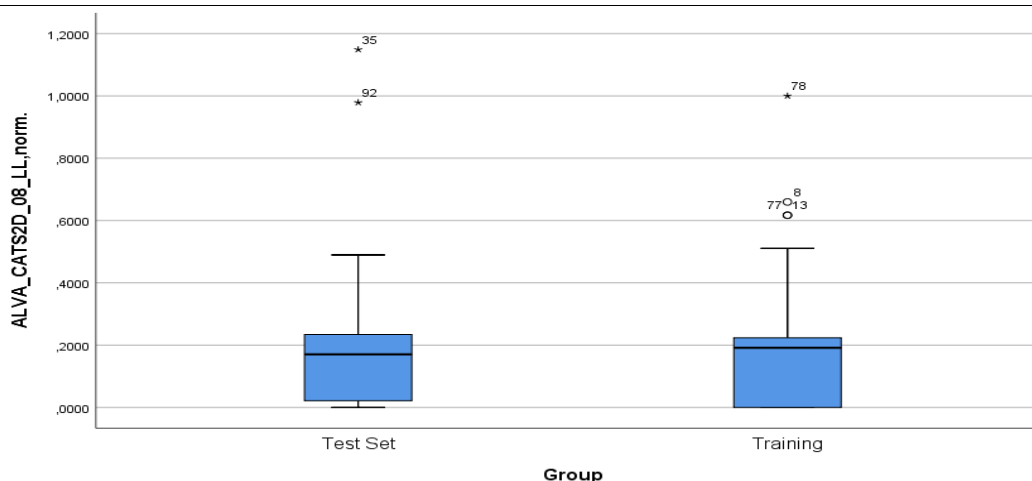


Figure 82. The comparison of the boxplot graph of training and test set to determine the outside of AD for Alva_CATS2D_08_LL normalized descriptor in the Alva model.

4.2 Skin Irritation

4.2.1 The Training and Test Sets for Skin Irritation

4.2.1.A Clustering studies

Since the chemicals are the same for both comedogenicity and skin irritation studies, the clustering studies for both of them are the same. Therefore, as it was in the clustering studies for comedogenicity, the cluster (k=6) based on functional group descriptors was again chosen as the best optimal cluster for the division of the skin irritation dataset into training and test set. The cluster members of the skin irritation study were given in Appendix 8.5_Table S5B.

4.2.1.B Division of the skin irritation dataset into training and test set

After clustering, the molecules in each cluster were manually distributed to the training and test set by keeping approximately 75%-25% of the ratio to divide the dataset. Also, active and inactive class distribution for skin irritation was tried to keep balanced for both training and test set. The distribution of the 121 chemicals into training and test sets for skin irritation and their class distribution as active (irritant) and inactive (non-irritant) are given in Table 32.

Table 32. The distribution of skin irritation dataset into training and test set

	Active (Irritant)	Inactive (Non-irritant)	Total	Active/Inactive
Training set	32	60	92	0,53
Test set	12	17	29	0,71
Total	44	77	121	

After the division of the dataset, the training and test set used for skin irritation modeling are listed in Appendix 8.5_Table S5. These chemicals in the dataset are shown as a list of irritant activities (active or inactive) and their irritation grade (Table S5.A) and also shown as a list of training and test set (Table S5.C) for skin irritation. Furthermore, the distribution of active/inactive compounds for skin irritation in the ten chemical classes are shown as the stacked bar charts for the skin irritation dataset, training, and test set. The stacked bar chart of the distribution of ten chemical classes in the skin irritation dataset, training, and test set are given in Figures 80, Figure 81 and Figure 82 respectively.

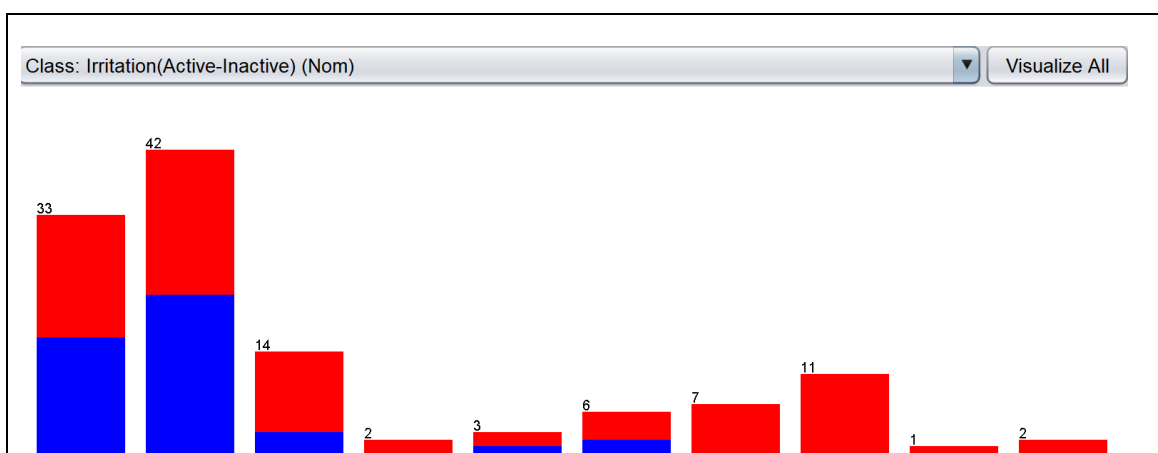


Figure 83. The dataset for skin irritation (n=121)
Distrubution of ten chemical classs in the dataset. (Blue= Irritant Red=Non-irritant)

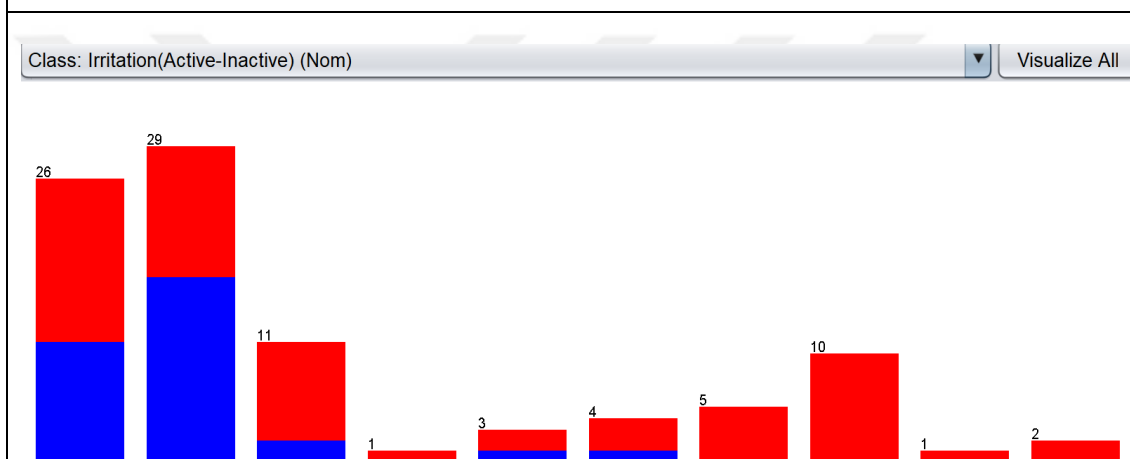


Figure 84. The training set for skin irritation (n=92)
Distrubution of ten chemical classs in the training set (Blue= Irritant, Red=Non-irritant)

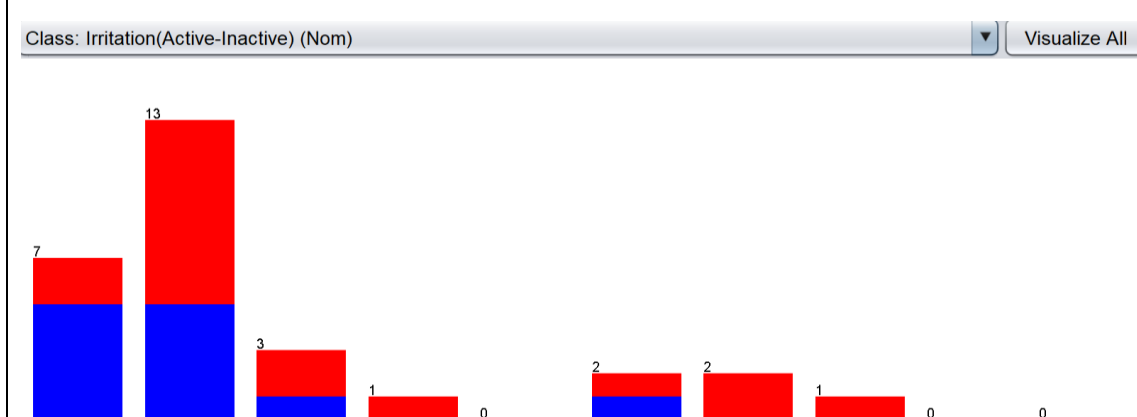


Figure 85. The test set for skin irritation (n=29)
Distrubution of ten chemical classs in the test set (Blue= Irritant, Red=Non-irritant)

4.2.2 QSAR Models for Skin Irritation

In the skin irritation studies, the significant descriptors for QSAR models were selected via the learning-based method since it was found the best successful feature selection method in the comedogenicity studies.

4.2.2.A The selected descriptors for skin irritation models

The number of attributes selected via the learning-based method for each skin irritation model built with AlvaDesc, Mold2, TEST, and PADEL descriptor packages were given respectively in Table 33. The original value and normalized value of all selected descriptors were given in Appendix 8.6_TableS6.1 and TableS6.2.

Table 33. The selected attributes by WEKA via the learning-based method

AlvaDesc	Mold2	TEST	PADEL
2	8	5	4
ALVA_Mi	MLD2_D004	TST_xch9	PDL_naAromAto m
ALVA_ATSC7e	MLD2_D196	TST_xvc4	PDL_AATS3e
	**MLD2_D223	TST_BELe1	PDL_AATS0i
	**MLD2_D294	TST_BEIp1	PDL_SdssC
	**MLD2_D489	TST_piPC04	
	**MLD2_D600		
	**MLD2_D606		
	**MLD2_D777		

** The Mold2 descriptors which were used in the Normalized Mold2 model.

4.2.2.B The description of descriptors selected via the learning-based method

For the skin irritation models, the names and the description of the descriptors selected by WEKA software for various descriptor packages (AlvaDesc, Mold2, TEST, PaDEL) were given in Table 34.

Table 34. The description of the descriptors selected by WEKA for skin irritation

Software	Number of calculated descriptors	Selected Descriptor	Description	Block
AlvaDesc	1819	ALVA_Mi	mean first ionization potential (scaled on Carbon atom)	Constitutional descriptor
		ALVA_ATSC7e	Centred Broto-Moreau autocorrelation of lag 7 weighted by Sanderson electronegativity	2D autocorrelation
Mold2	777	MLD2_D004	Number of 05-membered rings	
		MLD2_D196	Sum absolute electron topological negative variation	
		MLD2_D223	Average valence vertex connectivity order-5 Index	
		MLD2_D294	Bond information content order-1 index	
		MLD2_D489	Moran topological structure autocorrelation length-3 weighted by atomic van der Waals volumes	
		MLD2_D600	Number of ring secondary C-sp3	
		MLD2_D606	Number of secondary C-sp2	
		MLD2_D777	Molecular regression coefficients surface LogP index	
TEST	797	TST_xch9	Simple 9th order chain chi index	Chi Connectivity Indices-Simple
		TST_xvc4	Valence 4th order cluster chi index	Chi Connectivity Indices-Valence
		TST_BELe1	Lowest eigenvalue n. 1 of Burden matrix / weighted by atomic Sanderson electronegativities	Burden Eigenvalue Descriptors
		TST_BELe1	Lowest eigenvalue n. 1 of Burden matrix / weighted by atomic polarizabilities	Burden Eigenvalue Descriptors
		TST_piPC04	Molecular multiple path count of order 04	Walk and Path Counts
PaDEL	1444	PDL_naAromAtom	Number of aromatic atoms	AromaticAtomsCountDescriptor
		PDL_AATS3e	Average Broto-Moreau autocorrelation - lag 3 / weighted by Sanderson electronegativities	AutocorrelationDescriptor
		PDL_AATS0i	Average Broto-Moreau autocorrelation - lag 0 / weighted by first ionization potential	AutocorrelationDescriptor
		PDL_SdssC	Sum of atom-type E-State: =C<	ElectrotopologicalStateAtomTypeDescriptor

In the learning-based feature selection method, the decision tree, J48 classifier (C4.5 algorithm) used information gain to choose the best attribute (descriptor). Therefore, the amount of information gain obtained from selected attributes by WEKA was given in Table 35 for all models.

Table 35. The amount of information gain (IGA) of descriptors selected by WEKA for all models

Alvadesc Model		Mold2 Model		TEST Model		PADEL Model	
Selected Attribute	IGA	Selected Attribute	IGA	Selected Attribute	IGA	Selected Attribute	IGA
ALVA_Mi	0,326	MLD2_D004	0	TST_xch9	0	PDL_naAromAtom	0
ALVA_ATSC7e	0	MLD2_D196	0	TST_xvc4	0	PDL_AATS3e	0,183
		MLD2_D223	0,223	TST_BELe1	0,13	PDL_AATS0i	0,326
		MLD2_D294	0,13	TST_BEUp1	0.28	PDL_SdssC	0
		MLD2_D489	0	TST_piPC04	0		
		MLD2_D600	0				
		MLD2_D606	0				
		MLD2_D777	0.117				

4.2.2.C Classification models for skin irritation

The classification models obtained via the learning-based method for AlvaDesc, Mold2, TEST and PADEL descriptor packages were given respectively in Table 36.

Table 36. The classification models obtained via the learning-based method for irritation

Model	Classifier	Data	Number of selected descriptors	*Train/Test	ACC (%)	F score	MCC	AUC
AlvaDesc	RF	Original	2	Train	97,83%	0,98	0,95	0,998
				CV (10-fold)	76,09%	0,76	0,47	0,82
				Test	62,07%	0,61	0,19	0,68
AlvaDesc	RF	Normalized	2	Train	97,83%	0,98	0,95	0,998
				CV (10-fold)	76,09%	0,76	0,47	0,82
				Test	58,62%	0,49	0,05	0,45
Mold2	RF	Original	8	Train	100%	1,00	1,00	1,00
				CV (10-fold)	85,87%	0,86	0,68	0,88
				Test	82,75%	0,83	0,64	0,85
**Mold2	RF	Normalized	6	Train	100%	1,00	1,00	1,00
				CV (10-fold)	85,87%	0,86	0,68	0,86
				Test	62,07%	0,60	0,18	0,73
TEST	RF	Original	5	Train	100%	1,00	1,00	1,00
				CV (10-fold)	86,96%	0,87	0,72	0,92
				Test	72,41%	0,73	0,45	0,81
TEST	RF	Normalized	5	Train	100%	1,00	1,00	1,00
				CV (10-fold)	86,96%	0,87	0,72	0,92
				Test	65,51%	0,62	0,26	0,71
PADEL	RF	Original	4	Train	100%	1,00	1,00	1,00
				CV (10-fold)	79,35%	0,79	0,53	0,82
				Test	65,52%	0,65	0,27	0,65
PADEL	RF	Normalized	4	Train	100%	1,00	1,00	1,00
				CV (10-fold)	81,52%	0,81	0,58	0,82
				Test	68,97%	0,69	0,39	0,67

Abbreviation: RF: Random Forest, CV: cross-validation, ACC: accuracy, MCC: Matthews' correlation coefficient, AUC: area under the ROC curve, *Train:92/Test:29, Setting: Default Parameters

** The Mold2 descriptors which were used in the Normalized Mold2 model.

Furthermore, experimental values for both training set and test set and predicted values for four models (M) for irritation that are $M1_{pred_alvaDesc.}$, $M2_{pred_Mold2}$, $M3_{pred_TEST}$, $M4_{pred_PaDEL}$ are given in Appendix 8.7_TableS7.

4.2.3 Optimization of the Skin Irritation Models

In the optimization studies for skin irritation models, as performed in the optimization studies for comedogenicity models, we removed the features that had no score of information gain (IGA=0) from the models, and then, the performance of the models (via ten-fold cross-validation) and the predictive ability of the models (via test set) were evaluated by classification accuracy for the Alva, Mold2, TEST, and PADEL model. The comparison of the initial models and optimized models was given in Table 37. According to the results of the model comparison, the initial model of Mold2 with eight descriptors was chosen as the best model in the skin irritation study.

Table 37. The comparison of initial models with optimized models for skin irritation

		Initial Models			
Model		Alva	Mold2	TEST	PADEL
The number of descriptors		2desp	8desp	5desp	4desp
Accuracy	Train	97,83 %	100%	100%	100%
	CV-10	76,09 %	85.87%	86.96%	79,35%
	Test	62,07 %	82,75%	72,41%	65,52%
		Optimized Models			
Model		Alva	Mold2	TEST	PADEL
(The number of descriptors (IGA>0))		1desp	3desp	2desp	2desp
Accuracy	Train	95,65%	100%	100%	100%
	CV-10	67,39%	86,96%	70,65%	70,65%
	Test	62,07%	72,41%	65,75%	65,52%

4.2.4 Predictive Ability of Optimized QSAR Model for Skin Irritation

After the optimization of the models, the initial model of Mold2 with eight descriptors was chosen as the best model in the skin irritation study. Therefore, the predictive ability of the best QSAR model for skin irritation, which was the initial model of Mold2 with eight descriptors, was evaluated on the test set (N=29) for each of the ten chemical classes. The results of the true/false prediction of the QSAR model_{initial Mold2} on the test set for each of the ten chemical classes were given in Figure 86. (Total= 29 Prediction [24 True/ 5False])

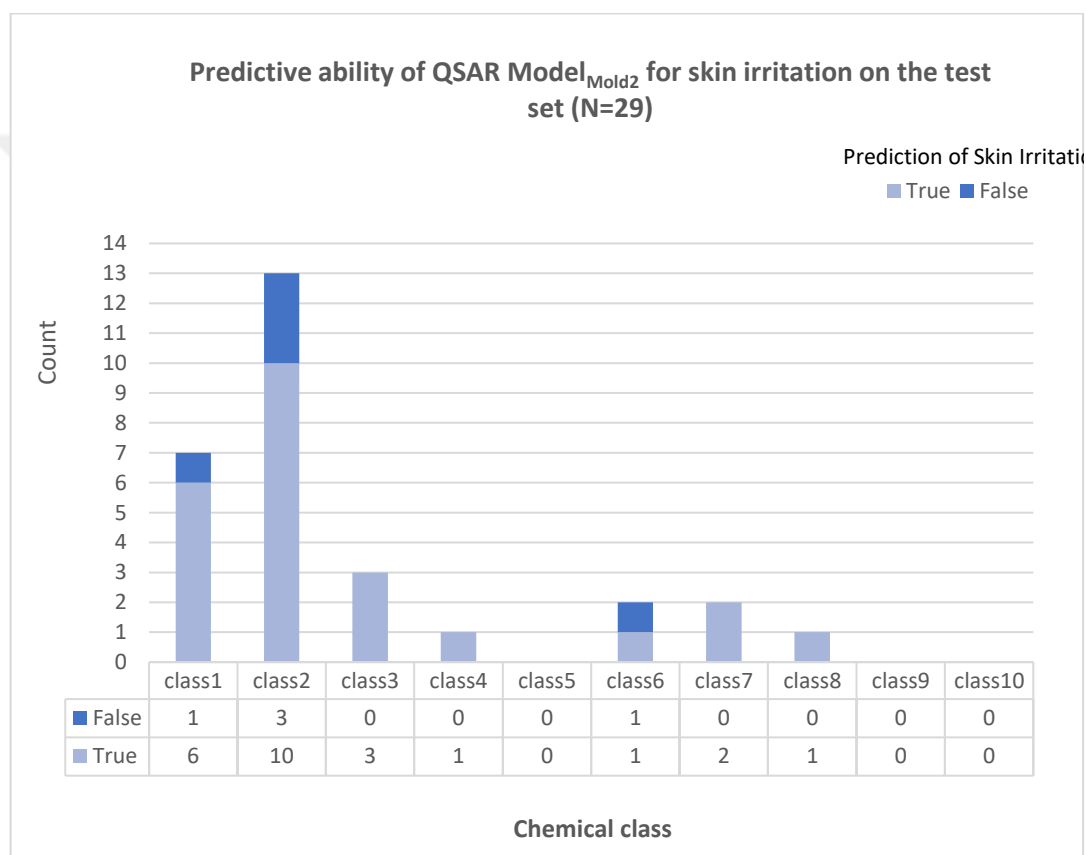


Figure 86. The stacked bar chart of true/false prediction of QSAR Model Mold2 for skin irritation on the test set (N=29) for each of ten chemical classes

class1: fatty acid and derivatives, class2: alcohols-fatty alcohols-sugars and their derivatives, class3: pigments, class4: silicones, class5: sterols, class6: vitamins, class7: preservatives & additives, class8: miscellaneous, class9: thickeners, class10: oils

4.3 The Relationship Between Comedogenicity and Skin Irritation

4.3.1 Distribution of Dataset

A total of 121 cosmetic ingredients in our dataset were shown with a Venn diagram (Figure 87) to visually represent the relationship between comedogenic and irritant cosmetic ingredients. Thus, the similarities and differences between the two groups were clarified and became interpretable.

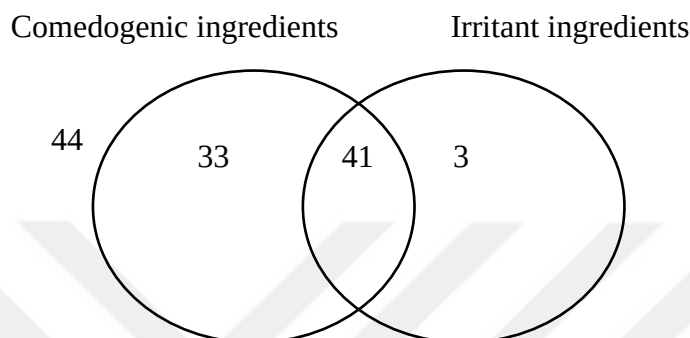


Figure 87. Venn diagram of 121 cosmetic ingredients in the dataset

The distribution of a total of 121 cosmetic ingredients has been shown in the follows;

- While 61.2% of our total 121 substances are comedogenic, 36.4% are irritants. Therefore, our dataset is composed of mostly comedogenic substances and a few of them is irritant (N=121)
- While 55.4 % of 74 comedogenic is irritant, 6.4% of 47 non-comedogenic is irritant (N=121)
- While 93.2% of 44 irritant is comedogenic, 42.9% of 77-non-irritant is comedogenic (N=121)

4.3.2 The determination of relationship via Chi-square test

A chi-square test was used to determine whether there was a significant relationship between the proportion of comedogenicity which are irritant and non-irritant cosmetic ingredients. Results indicated that 93.2% of irritants were comedogenic, whereas 42.9% of non-irritants were comedogenic (Table 39). Since the **p-value was found less than 0.001, the null hypothesis was rejected. Therefore, it was found statistically significant associations between comedogenicity and irritation at the significance level ($p=0.000000018$, $p<0.001$) as shown in Table 38 and Table 39.

Table 38. The results of the Chi-Square test to find the relationship between two variables, in terms of comedogenicity.

Irritation	Comedogenicity		Total	p-value
	Comedogenic	Non-Comedogenic		
	N (%)			
Active	41(55.4%)	3(6.4 %)	44(36.4 %)	0.000000018**
Inactive	33 (44.6%)	44(93.6%)	77(63.6%)	
Total	74 (100.0)	47 (100.0)	121 (100.0)	

Chi-Square test-, Fisher's Exact Test ** $p < 0.001$

Table 39. The results of the Chi-Square test to find the relationship between two variables in terms of irritation

Comedogenicity	Irritation		Total	p-value
	Active	Inactive		
	N (%)			
Comedogenic	41(93.2%)	33(42.9%)	74 (61.2%)	0.000000018**
Non-comedogenic	3(6.8%)	44(57.1%)	47(38.8%)	
Total	44 (100.0)	77 (100.0)	121 (100.0)	

Chi-Square test-, Fisher's Exact Test ** $p < 0.001$

The distribution of comedogenicity in the irritant and non-irritant cosmetic ingredients was shown in Figure 88 (N=121).

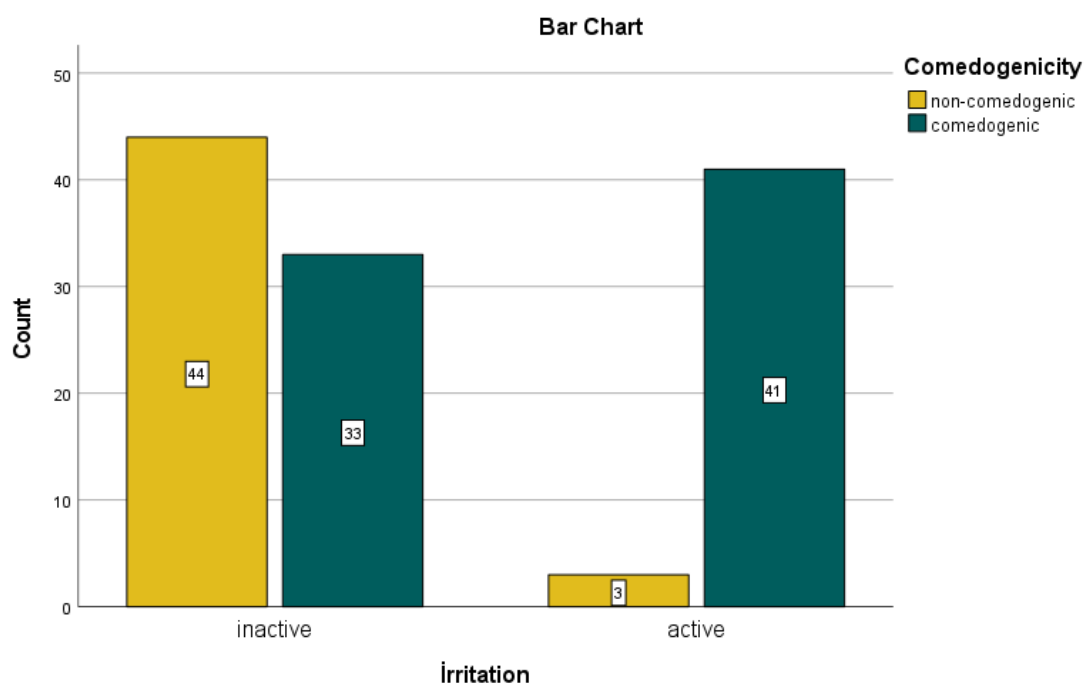


Figure 88. The distribution of 121 cosmetic ingredients in terms of irritation and comedogenicity.

The Cramér's Value and Phi from chi-square analyses is given in Table 40.

Table 40. The results for the strength of association between comedogenicity and irritation from the Chi-Square test.

	Value	Significance
Cramer's V	0,497	p<0,001
Phi	0,497	p<0,001

According to the interpretation table of Cramer's value, there was found relatively strong association between comedogenicity and irritation.



5. DISCUSSION

5.1 Comedogenicity

5.1.1 The Division of the Dataset into Training and Test Set

5.1.1.A Clustering Studies

All QSAR theories assume that similar compounds have similar biological activities. Therefore, for a predictive QSAR model, the points of the test set must be similar to the points of the training set in terms of molecular representativeness in the descriptor space (155). Besides, data clustering is the process of identifying groups in the dataset, known as clusters, based on some similarity measures (156, 157). Leonard and Roy (2006) proposed that k-means clustering-based division of training and test set as a reliable method for QSAR modeling. Furthermore, Leonard and Roy (2006) reported that all chemical classes should be represented in the training set. Otherwise, there is an apparent risk that small clusters with few members will not be represented in the final training set (155). In the report (2005) which was published by European Chemicals Bureau for the computational toxicology action, classification of chemicals into-similarity-based chemical categories could be both useful for building QSAR models and for defining chemical category proposals of AD in the regulatory application (158). The definition of a chemical similarity measure depends on functional groups or common substructures of molecular features under analysis (158). Therefore, in this study, the main goal of clustering studies was to obtain homogeneous and representative clusters based on chemical structural similarity measures. These similarity measures were functional groups, fingerprints, and all types of descriptors in the model (155, 158, 159). For this reason, in the clustering studies, three different types of clusters were obtained, based on all of the descriptors (n=1819), the functional group descriptors (n=154), and all of the fingerprint (n=4504). In order to find optimum clusters, the elbow method was used. When comparing the optimal clusters in Figure 65, the most dramatic decrease in SSE that forms an angle of elbow is k=6, it is a good curve for the cluster based on functional groups. In Figure 66, SSE decreases by k=3, but suddenly on k=4, it increases and decreases once again, elbow is formed on k=5 for the cluster based on all descriptors in the Alva model. In Figure 67, the graph for cluster-based on all fingerprints shows the irregular pattern that SSE decreases by k=5 and forming an angle of the elbow on k=5, it increases on k=6 and decreases once again, forming an angle of the elbow on k=7.

Distance is a kind of similarity/dissimilarity index to quantify similarity such as Euclidean distance (158). Effective clustering maximizes intra-cluster similarities and minimizes inter-cluster similarities (160). K-means clustering is a centroid-based clustering (160), that is capable of identifying clusters by repeated comparison of input patterns until the stable clusters obtained based on the clustering criteria (157). In the k-means clustering, inter-cluster distance is the distance between two clusters and is defined as the distance between their centroids. This term is used for the separation of the clusters. Intra-cluster distance is the distance between all pairs of points in the cluster that are used for the compactness of the clusters (156).

The quality of the optimal clusters was evaluated by measuring the intra-cluster variance and inter-cluster distance (156). While a homogenous cluster has high intra-cluster similarity and low intra-cluster variance, a heterogeneous cluster has low intra-cluster similarity and high intra-cluster variance. According to the results of intra-cluster variance (Table 16), since cluster-based on the functional group had the minimum intra-cluster variance, it is the most homogenous and most compact cluster, on the contrary, since cluster based on all the descriptors had the maximum intra-cluster variance, it is the most heterogeneous cluster. According to the results of the inter-cluster distance (**Table 17**), since the cluster based on the functional group had the maximum inter-cluster distance, it is the best well-separated cluster. On the contrary, since the cluster based on fingerprints had the minimum inter-cluster distance, it is the bad separated cluster. O'Boyle and Sayle reported (2016) that some fingerprints are not suitable for measuring structural similarities such as MACCs keys and Daylight-type-path-based fingerprints. Huang et al., (2019) reported that it is useful to understand of the functional group of drugs leading to specific properties for interpretable fingerprints linked to sub-structures' relevance to the molecular properties of drugs (161).

In organic chemistry, functional groups are specific groups of atoms bonded in a certain arrangement that provides a compound characteristic physical and chemical properties. Besides, a functional group within a compound is responsible for the characteristic chemical reactions of that compound. Furthermore, the same functional group behaves in a similar way by undergoing similar reactions, regardless of the compound of which part it is. Any molecule can contain more than one functional group. At a simple level, molecules with a certain functional group can be predicted to react in a similar way with other substances. Additionally, functional groups are often used to

“functionalize a compound affording it different physical and chemical properties than it would have in its original form. Functional groups can also be used to distinguish similar compounds from each other (162). As a traditional approach in organic chemistry, the classification of organic molecules by their functional groups is that it also classifies their chemical behaviour (163). Also, in QSAR studies, similarity-based classification of chemicals has been recommended for defining the chemical domain for AD in both research and regulatory application (158). Therefore, in this study, the goal of k-means clustering based on the functional group of the dataset was to provide chemical class homogeneity and representativeness while the training and test set division. When the quality of the clusters was evaluated, the optimum cluster, $k=6$ based on the functional groups gave the best homogenous and chemical class representative cluster. The distribution of the optimum number of cluster membership which is $k=6$ cluster for our dataset is given in Appendix 8.2_ Table S2B.

5.1.1.B Evaluation of Training and Test Set for Comedogenicity

After clustering, compounds selected from each cluster were distributed approximately in the ratio of 3:1 among training and test set, regarding the active-inactive ratio was considered. Thereby, the balanced training and test set were obtained. The stacked bar chart of chemical class distribution in terms of active/inactive compounds for the dataset, training and test set were shown in Figure 68, Figure 69, Figure 70 respectively. When comparing the chemical class distribution of training and test set in terms of representativeness (Figure 69) and (Figure 70), it was similar to the dataset (Figure 68). This condition showed that the clusters used for the division of the dataset into the training and test set were homogeneous and represented the chemical class distribution of the dataset. When comparing the distribution of ten chemical class and their active and inactive ratio, in the training (Figure 69) and test set (Figure 70), especially chemical class 1 (fatty acid and derivatives), chemical class 2 (alcohols-fatty alcohols-sugars and their derivatives), chemical class 3 (pigments) were the best representative groups of the dataset in both training and test set. Evaluation metrics behave differently under balanced and imbalanced datasets. Therefore, the class distribution of data should be analyzed to choose appropriate metrics for meaningful performance evaluations (67). When we evaluated the class distribution of the dataset by analyzing the ratio of the active to inactive, we found this ratio equals to 1.49 for the training set whereas it found equals to 1.9 for the test set (Table 18). This ratio was close to each other for both the training and

the test set. Therefore, the splitting of the training and the test set was considered similar to the dataset. Class imbalance represents the difference in the numbers of active and inactive instances, usually with the inactive outnumbering the active (67). When we evaluated the training and test set in terms of class imbalance, since this difference is less for the training than the test set, the training set can be considered more balanced than the test set.

5.1.2 The Descriptors in the Comedogenicity Models

5.1.2.A Comparison of Feature Selection Method for QSAR Model

In machine learning, feature selection is the most important process in order to select the most relevant and significant feature (164). While the correlation-based method (CfsSubSet) evaluated attributes subset that is highly correlated with the class but with low intercorrelations each other, the learning-based method (Wrapper with the J48 classifier) uses the decision tree classifier (J48) to evaluate attribute subset and employs cross-validation to estimate the classification accuracy of the learning scheme for the evaluation of each subset (81). When we compared the correlation-based method to the learning-based on four different descriptor software packages (AlvaDesc, Mold2, TEST and PADEL) by the means of developed models via the random forest. (Table 23). The learning-based feature selection led to simpler and more interpretable models than correlation-based by reducing the number of descriptors by the range of 20-85%, while models' fit and predictivity were almost unchanged (Table 20 versus Table 22). The number of reduced descriptors for each descriptor software were given in respectively. These were from 16 to 5 descriptors (70% reduced) for AlvaDesc, from 17 to 7 descriptors (60% reduced) for Mold2, from 10 to 8 descriptors (20% reduced) for TEST, and from 26 to 4 (85% reduced) for PADEL (Table 23). As a result, since the fewer descriptors provide the better classification and more interpretable model, the clear winner of the comparison is the learning-based feature selection method.

5.1.2.B The Selected Descriptors for Comedogenicity Models via Learning-based

Adhatrao et al., (2013) reported that “in a decision tree, at each node of the tree, J48 classifier (C4.5) chooses one attribute of the data that most effectively splits data set of samples *S* into subsets that can be one class or the other (71). It is the normalized information gain (difference in entropy) that results from choosing an attribute for splitting the data. The attribute factor with the highest normalized information gain is considered to make the decision. The C4.5 algorithm then continues on the smaller sub-lists having the next highest normalized information gain. In the decision tree, information gain measures the amount of information obtained from the feature about the target variable. IGA is also used for the best feature selection. Entropy is related to the level of uncertainty. IGA is calculated from the difference of entropy before and after splitting of the decision tree. If data contains a homogenous subset of observations or if all the observations belong to one class, the entropy of the dataset becomes zero (71, 165, 166).

The amount of information gain (IGA) from all variables (a total of 1882) for Alva descriptors was ranged between 0.346-0. Alva_ATSC7i (IGA=0.336) and Alva_CATS2D_08_LL (IGA=0.314) were the two successful descriptors with the highest amount of information gain that selected from the best feature subsets by WEKA. Alva_S3K is the third descriptor in the Alva model with the IGA=0.151 (Table 27). On the other hand, Alva_ATSC7i (0.026) and Alva_CATS2D_08_LL (0.070), and Alva_S3K (0.185) had found low correlations with the activity (Table 26). Because J48 used in the learning-based feature selection is the non-linear classifier, a direct linear correlation is not anticipated between the activity and the variables in such a non-linear model (167, 168).

The CATS 2D (Chemical Advanced Template Search) descriptors are pharmacophore descriptors that have been widely used in similarity search (169, 170, 235). They are a particular case of autocorrelation descriptors, where the atom-type definition is related to the concept of potential pharmacophore points (PPP). Pharmacophore is defined as the actual features that are responsible for the molecular behavior of the molecules (169, 171, 235). CATS2D_08_LL represents the presence of lipophilic atom pairs at a topological distance of 8 bonds that may be relevant in describing molecular biological activity or feature causing molecular behavior (172, 235).

Our dataset consists of mostly lipophilic substances such as fatty acids, fatty alcohol, and their derivatives. Therefore, this descriptor may be probably pointed out the fatty acids and fatty alcohols and their derivatives. Kligman et al., (1970) reported that among the saturated fatty acids comedogenic activity began at C (carbon)10, reached a peak at C12 and C14, lessened at C16 and C18, and was absent at C20. In our dataset including the fatty acid range of C8-C22 (Appendix 8.2_TableS2A), while behenic acid (C22) was inactive as comedogenic, capric acid (C10) was active as comedogenic (173). CATS2D_08_LL, the selected descriptor with learning-based feature selection by WEKA, can be related to the lipophilic pharmacophoric points that describe this different molecular behavior or biological activity. This kind of lipophilic pharmacophoric point may determine the comedogenic potential of lipid or lipidic substances.

In recent studies on acne show the role of oxidative stress in the formation of acne. The comedogenesis is the initial stage of acne. Lipid peroxidation is considered to play a major role in the initiation and progression of acne. Also, oxidants generated by lipid peroxidation are suspected promoting of hyperkeratinization and thereby comedogenesis (80, 82, 88). Bhat et al., (2017) reported that comedo formation is driven by increased levels of androgens, changes in sebum lipid composition, *P. acnes* overgrowth, and local cytokine milieu. Also, Bhat et al., (2017) reported the role of sebum in acne that hypersecretion of sebum, alteration of the lipid composition of sebum, and the changes in oxidant/antioxidant ratio and the characteristic of the skin surface lipids are the main events in acne pathogenesis. The composition of skin lipids is important and the normal composition of sebum produced by sebaceous glands consists of a group of oils including triglycerides, fatty acid, breakdown products, squalene, wax esters, cholesterol, and its esters. (100, 173). In the skin of acne patients, this composition was found changed with the decreased level of linoleic acid (polyunsaturated essential fatty acid) and the presence of lipoperoxides due to the peroxidation of squalene (100).

Previous studies reported that sebum oxidation is the initial phase of comedogenesis (88). In lipid peroxidation, skin lipids that the polyunsaturated fatty acids are predominantly susceptible because abstraction from a methylene (-CH₂-) group of a hydrogen atom, which contains only one electron, leaves at the back an unpaired electron on the carbon. The initial reaction of •OH with polyunsaturated fatty acids produces a lipid radical (L•), which in turn reacts with molecular oxygen to form a lipid

hydroperoxide (LOOH) (91). In this study, the selected descriptor for the Alva model, CATS2D_08_LL, maybe mimic the polyunsaturated fatty acids due to its lipophilic pharmacophoric points thereby showing similar molecular behavior. In acne, endogenous sources of oxidative stress components such as reactive oxygen species (ROS) in the skin initiate lipid peroxidation (80). If so, exogenous sources of ROS such as cosmetics, xenobiotics, drugs, pollutants (172) in the skin can initiate lipid peroxidation. Therefore, cosmetics, when applied to the skin, can change the composition of the skin increasing oily substance content and causing some kind of reaction such as lipid peroxidation, as shown in Figure 89). In the light of our literature review and results obtained from the QSAR model optimized AlvaDesc model in this study, we have suggested AOP for comedogenicity prediction (Figure 89).

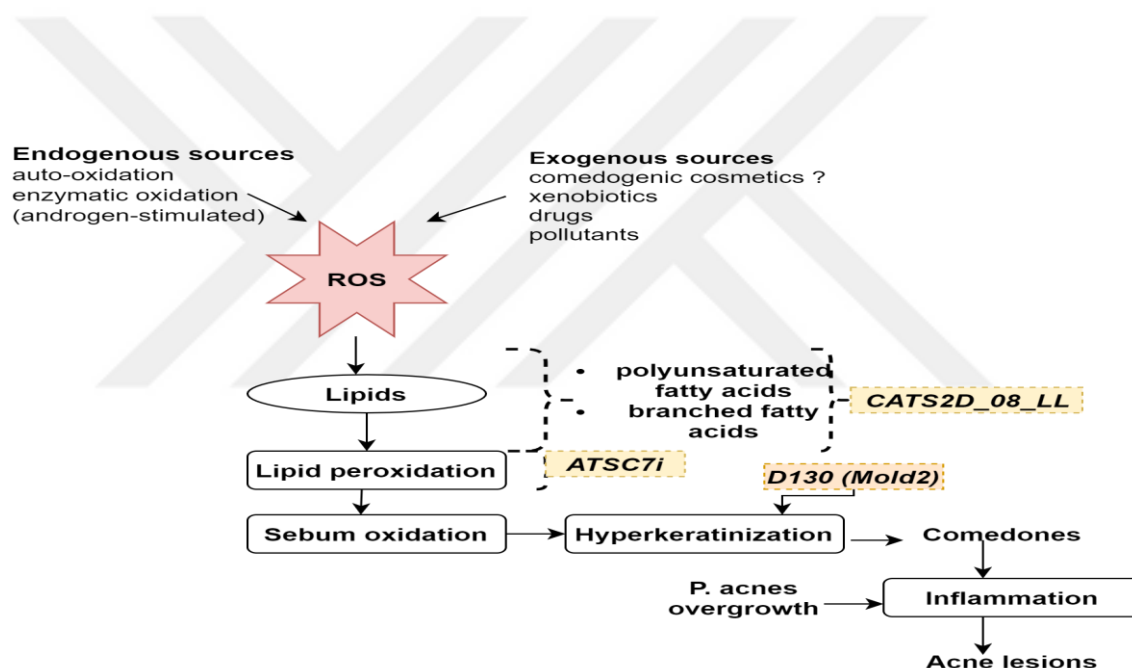


Figure 89. Adverse outcome pathway of our QSAR Alva Model for comedogenicity prediction

Lipid peroxidation and polyunsaturated fatty acids in skin lipids are considered responsible in the initiation of comedogenesis. The significant descriptors in our QSAR model supported by the literature obtained via learning-based methods are ATSC7i (ROS source for lipid peroxidation) and lipids (CATS2D_08_LL) for AlvaDesc and D130 for Mold2 (halogen atoms for hyperkeratinization related to the halogen acne).

Alva_ATSC7i, the second successful descriptor in the Alva model, is an autocorrelation descriptor that is defined as centered Broto-Moreau autocorrelation of lag

7 weighted by ionization potential. Ionization potential describing as the amount of energy required to remove an electron from an atom or molecule (175). Buglak et al., (2021) reported that the ATSC7i descriptor is directly proportional to the logarithm of singlet oxygen generation quantum yields ($\text{Log}\Phi\Delta$) and also reported that structures with high numbers of nitrogen, oxygen, and halogen atoms have high ATSC7i values due to higher ionization potential of such atoms (176). Also, Zhang et al., (2021) reported that they found evidence that singlet oxygen is produced through the Russell mechanism during the self-reaction of lipid peroxyl radicals generated via iron-dependent lipid peroxidation. The generated singlet oxygen accelerated the oxidative damage of lipid membranes by propagating lipid peroxidation. As a result, ATSC7i descriptor is proportional to singlet oxygen and related to lipid peroxidation (177). Therefore, the ATSC7i descriptor in our Alva model can represent ROS source for lipid peroxidation. Furthermore, Baumgartner et al., (2002) reported that experimental studies indicating that reactivity in a homolog series of radical or molecular reactions strongly correlated with the ionization potential (IP) of the group of chemicals that is electron donating in transition state and with the electron affinity (EA) of the electron withdrawing species (178).

Alva_S3K, the third descriptor in the Alva model, defining as 3-path Kier alpha-modified shape index. It is a topological index that has been used for the biological activity or properties of molecules which are correlated with the chemical structure of molecules (179). Especially, the shape of the molecule is important for molecular interactions (180). Also, 3-path Kier alpha-modified shape index that has been shown to contribute negatively to the anti-mycobacterial activity of the branched molecules in comparison with non-branching states. That negative effect on anti-mycobacterial activity can trigger acne due to the fact that bacterial colonization has been implicated in the pathology of acne (80, 181).

In the Mold2 model, D130 represents the number of halogen atoms in each molecule. Halogen acne is defined as an acneiform eruption caused by iodides, bromides, and fluorides (182, 183). We consider that there is a relationship between the descriptor of D130 and halogen acne due to the fact that both of them point to halogen. Pigments containing halogen (Br, Cl, I) were indeed classified as comedogenic in this study. This relationship may explain why halogen-containing pigments are comedogenic. Because “comedo” is a type of acne and characterized by formation of open comedones

(whiteheads) and closed comedones (blackheads) but no sign of any inflammatory papules or pustules (14).

Furthermore, the Mold2_D130 descriptor, the selection of halogen atoms as a descriptor in the comedogenicity model by WEKA indicates that halogens may also be comedogenic. There is a concept called halogen acne in the literature. We know that comedones are the initial lesions of acne. There is hyperkeratinization in the toxicity mechanism of halogens. Comedogenicity also means the potential of a substance to form comedones by hyperkeratinization mechanism. In the light of all this information in the literature, since a comedogenic substance causes hyperkeratinization and similarly, halogen atoms also cause hyperkeratinization too, resulting in the conclusion that halogens can be comedogenic.

In this context, we have explained more clearly by adding mechanism information about halogen toxicity and comedogenicity while halogen toxicity, such as the exposure to the dioxin-like halogenated compounds, is characterized by hyperkeratinization of the ducts of the sebaceous gland (184), the comedo formation is the result of follicular hyperkeratinization which is one of the interplaying factors in the etiopathogenesis of acne (100). Moreover, recent studies claim that the oxidants generated by lipid peroxidation are suspected promoting of hyperkeratinization (80). As a result, literature information supports the possibility of a relationship between the Mold2_D130 descriptor that relates to the halogen atom and comedogenicity in terms of hyperkeratinization. However, although the Mold2_D130 descriptor was selected as the important attribute by WEKA in the learning-based feature selection, the amount of information gain found zero for Mold2_D130. This situation may point to uncertainty and may be due to the inadequate information obtained from training data (71, 166). Moreover, in the literature, the toxicity mechanism of 78 polyhalogenated compounds was modeled by using autocorrelation descriptor by the means of the hydrophobicity potential on the van der Waals surface in terms of aromatic hydrocarbon receptor affinity (185). Therefore, there may be a relationship between the Mold2_D489 which was an autocorrelation descriptor, with length-3 by weighted atomic van der Waals volumes and MLD130 in terms of halogen toxicity.

When we evaluated the all model, the autocorrelation descriptors were found as common descriptors selected by WEKA for the three of four descriptor packages. They were Alva_ATSC7i, Alva_MATS6p, MLD2_D489, PDL_ATS2m, PDL_AATS3m. The

autocorrelation descriptors are based on molecular graphs and define a relation between atoms as a function of their spatial separation. They encode the identity and electronic attributes including atom types, atomic charges, ionization potential, electronegativity, polarizability and Van der Waals volume (38). The role of oxidative stress in the formation of acne has been discussed recently (80, 82, 88, 100). Due to the fact that the autocorrelation descriptors were found common in three of four QSAR models. Therefore, we have looked for the relationship between oxidative stress and the autocorrelation descriptors. Among these descriptors, Alva_ATSC7i was the most significant descriptor in terms of model success and with the highest amount of information gain score (IGA= 0.336). Alva_ATSC7i is defined as Centred Broto-Moreau autocorrelation of lag 7 weighted by ionization potential. In the literature, our findings of oxidative stress on acne were mainly related to lipid peroxidation that has a major role in the initiation of comedogenesis and the progression of acne. Buglak et al., (2021) reported that singlet oxygen generation highly correlated with ATSC7i descriptor. Zhang et al., (2021) reported that the generated singlet oxygen accelerated the oxidative damage of lipid membranes by propagating lipid peroxidation. Therefore, we can conclude that there is a correlated relationship between the ATSC7i and lipid peroxidation (176, 177).

The other autocorrelation descriptors were explained as follows; Alva_MATS6p is defined as Moran autocorrelation of lag 6 weighted by polarizability (IGA=0). MLD2_D489 is defined as Moran topological structure autocorrelation length-3 by weighted atomic van der Waals volumes. (IGA=0). PDL_ATS2m, Broto-Moreau autocorrelation - lag 2 / weighted by mass (IGA=0.308), PDL_AATS3m, Average Broto-Moreau autocorrelation - lag 3 / weighted by mass (IGA=0.305). When we investigated the relationship between the descriptors we found and acne, we obtained data showing that only three of them (ATSC7i, MATS6p and ATS2m) were related by acting on oxidative stress and tyrosinase. It was found in this study that the other descriptor, MLD2_D489, might determine the adverse effects on the skin, but since the mechanism is not known, it is estimated that they act by affecting the oxidative stress or tyrosinase enzyme, which is important in the development of acne (186). In terms of the relationship between the polarity and oxidative stress, the polarity of the solvent increases the contribution of two other mechanisms, that is, single-electron transfer, followed by proton transfer and sequential proton loss electron transfer (187). When we looked at the relationship between the hydroxyl radicals ($\bullet\text{OH}$), oxidative stress and descriptors,

ATS2m was used as a descriptor in a QSPR-MLR (multiple linear regression) model for the reaction with hydroxy radicals ($\bullet\text{OH}$) in the description of the various organic micropollutants including solvents, pharmaceutical and pesticides in a UV/ hydrogen peroxide process (188).

5.1.3 Classifier Performance

Classifiers learn the mapping from the input variables to the target variable (146). In the comedogenicity studies, in order to select the best mapping classifier for our input variables to the target variable of comedogenicity, various classifiers in WEKA were tested in the preliminary study. Among them, KStar, RF, and NNge were chosen as the top three classifiers to be used for comedogenicity modelling studies. In the literature, RF was generally chosen as the classifier for QSAR modelling study since it had a good prediction performance (125, 189). Varsou et al., (2017) and Gregorutti et al. (2017) were reported that RF had a good discriminative performance with a high accuracy prediction for classification model (125, 189). Similarly, in our study, comparing the top three classifiers (Table 27), RF showed the best predictive performance with the accuracy range of 86.96-81.52%, whereas Kstar gave a performance with an accuracy range of 83.69-75.00% and NNge gave a performance with the accuracy range of 80.43-76.10% for cross-validated four models. Also, RF gave the best predictive ability with the accuracy range of 79.3-68.96% whereas Kstar showed the accuracy range of 75.86-65.51% and NNge showed the accuracy range of 72.41-41.48% for four models evaluated on the test sets. The AUC value measures the quality of the classifier, a high AUC value is considered as an efficient model (31, 34, 67). Similarly, among the top three classifiers, RF was characterized as the best classifier due to the AUC range of 0.89-0.87 whereas Kstar was characterized as good to fair classifier due to the AUC range of 0.87-0.79 and NNge was characterized as fair due to the AUC range of 0.80-0.75 for cross-validated four models. Moreover, RF gave the best performance results with MCC range of 0.73-0.61 while Kstar with MCC range of 0.66-0.48 and NNge with MCC range of 0.59-0.50 for the cross-validated four models. As a result, RF was chosen as the best classifier among the top three classifiers in terms of common evaluation metrics that are accuracy, area under the ROC curve and MCC.

5.1.4 Model Performance

In the comedogenicity studies, since RF was chosen as the best classifier, the performance of the four different classification models built via RF were compared each other (Table 28). Cross-validated results were used for the proper performance evaluation of these models (146). Varsou et al., (2017), reported that overall success rate is measured by the means of classification accuracy (125). Similarly, in this study, ACC was mainly used to measure the overall success rate of the models for the prediction of the comedogenic potential. When comparing all the classification models (Table 28), Mold2 and AlvaDesc with RF gave the successful performance with 85.87% and 84.87% accuracy for the cross-validated models. Also, they gave successful predictive ability with 75.86% and 79.31% accuracy for the test set respectively. On the other hand, TEST model showed a good performance for the cross-validate model (86.96 % accuracy), but poor predictive ability for the evaluation on the test set (68.97 % accuracy). Likewise, PaDEL model did not perform good performance as well as other models performed for cross-validate model (81.52%), but performed good predictive ability for the evaluation on the test set (79.31% accuracy).

Gramatica (2013) reported that AUC is an index of the goodness of the classification model and a perfect model has an area of 1.0 while if AUC equals 0.5, the classifier has no discriminative power (31). The AUC results of cross-validated two successful models were as follows: 0.89 for Mold2, 0.87 for AlvaDesc. Accordingly (Section 3.2.10), the predictive performance of our two cross-validated models could be characterized as good. The AUC results on test set for two successful models were 0.76 for Mold2 and 0.81 for AlvaDesc. Accordingly, predictive ability of Mold2 model is characterized as fair, while predictive ability of AlvaDesc model is characterized as good. Our two successful models meet the first condition for model validity as mentioned in Section 3.2.9 (Model validation) with the Topliss ratio 13.1 (92 chemicals/7 descriptors) for Mold2 and 18.4 (92 chemicals/ 5 descriptors) for AlvaDesc that their Topliss ratio is greater than five (Table 23).

Saito and Rehmsmeier (2015) reported that AU-PRC (area under the Precision-Recall curve) is more informative than AUC (area under the ROC curve) while evaluating binary classification models on imbalanced dataset. Because AU-PRC score is changed between balanced and imbalanced dataset but AUC score is not changed (67). Therefore,

in order to evaluate whether the dataset is balanced or imbalanced, results of AU-PRC and AUC of all models with RF compared. The results of cross-validated AU-PRC were 0.86 for AlvaDesc, 0.89 for Mold2, 0.87 for TEST, 0.87 for PaDEL while the results of cross-validated AUC were 0.87 for AlvaDesc, 0.89 for Mold2, 0.89 for TEST, 0.88 for PaDEL. As a result, the cross-validated results for AU-PRC were similar to the cross-validated results for AUC. These similarities showed that our models build on balanced dataset.

In the modelling study, the training set is used to fit the models; cross-validation is used to estimate prediction error for model selection and the test set is used to assess the generalization error of the final chosen model (190). When all classification model for comedogenicity built via RF compared with each other (Table 28), TEST model showed the lowest prediction error (0.13) but highest generalization error (0.31). PaDEL model showed the highest prediction error (0.18) and lowest generalization error (0.21) like AlvaDesc model. As a result, AlvaDesc model (0.15, 0.21) and Mold2 model (0.14, 0.24) were selected as successful two models with the optimum prediction error rate and generalization rate respectively. Among the two successful models, Mold2 model was the best predictive and fitted model due to the minimum cross-validation error (0.14). On the other hand, training error was found zero for all of four models. If the learning and fitting the training data too much, the models do not generalize the data and leading to overfitting (190). The indication of overfitting is that where the test error or cross-validation error is much greater than the training error (191). In our study, training more data can provide better generalization by improving the performance of the models. However, due to fact that limited comedogenicity data is available in the literature, increasing the training data is not possible for now. Also, new data adding should be relevant. There are some strategies could be taken to mitigate overfitting in our study. One of them is removing the irrelevant and the redundant features. If too much irrelevant and redundant feature is present, the redundant feature directly leading to overfitting (164).

In the comedogenic model optimization study (Section 3.2.11 and Section 4.1.5), in the Alva model, we have just found the two features, Alva_piPC01 (IGA=0) and Alva_MATS6p (IGA=0), that had no score of information gain, they were considered as irrelevant descriptors. When we removed the irrelevant descriptors from the model, while the accuracy of the test set remained the same (79.31%), the accuracy of the cross-

validated model was increased from 84.78% to 86.96%. Thereby, overfitting was reduced and the number of descriptors in Alva model was reduced from 5 to 3 descriptors. (Table 29). As a result, we have improved the performance of Alva model by removing two irrelevant features. For other models, when we removed the features that having no score of information gain (IGA=0) from the model, overfitting was reduced for each of them. However, both the performance of the model and the predictive ability of model was reduced for Mold2, TEST and PADEL model (Table 30). The accuracy of cross-validated model was reduced from 85.87% to 72.83 for Mold2, from 86.95% to 78.26% for TEST, from 83.7% to 72.32% for PADEL. The accuracy of the predictive ability of the model was reduced from 75.86% to 68.97 for Mold2, from 68.96% to 62.1% for TEST and from 75.9% to 68.96% for PADEL. Secondly, the wrapper method selects the subset of relevant features using a learning algorithm however, one of the disadvantages of the wrapper method is the risk of overfitting (38). Another feature selection method can be tested to reduce overfitting. Additionally, decision trees, J48 (C4.5 algorithm), used as classifier in the wrapper method is the non-parametric machine learning algorithm that has been more risk of overfitting the training data, may be sometimes, it is hard to explain why specific predictions are made. In here, pruning of decision trees can be used after it has learned in order to remove some of the details picked up and also training much more data could be require to estimate the mapping function to solve the overfitting (192).

In machine learning model, “train” represented the training process which ML algorithm is learned from the training set to build a ML model by finding a pattern that maps the input data attributes to the target and it outputs an ML model which captures these patterns (146). In our ML model, there is a non-linear relationship between the activity and the variables. In such a non-linear ML model, the training error is not indicative error of the ML model when applied to new data. For the good estimation of error rate of the model, error of K-fold cross-validation could be used due to the fact that its error close to the error of the model on new data (146). For example, in terms of accuracy, it rigorously employs the training set to evaluate what accuracy will be on new data. In other words, K-fold CV error is a very good estimate for how the ML model will perform on the new data which allow to forecast the error of the model and to select the best model (146). In briefly, for the proper estimation of the accuracy of the ML model, k-fold cross-validation results of ML model could be considered as the accurate predictive performance of ML models. In this study, 100% or 1.0 value of evaluation metrics for

the training set depends on the ability of ML algorithm to capture intricate non-linear relationships between input data and the associated output (193). When ML algorithm, random forest (RF), was replaced with another ML algorithm such as J48, Bayes.net, Logistic model trees (LMT), training accuracy was decreased below the 100 percent. For example, training accuracy; 91.30% (J48), 85.87% (Bayes.Net), 88.04% (LMT). Also, AUC and MCC of training set were decreased below 1.0 as seen respectively training AUC; 0.92(J48), 0.88(Bayes.Net), 0.91(LMT). Furthermore, 10-fold CV accuracy of model were decreased as seen in the following; 84.78% (J48), 80.44% (Bayes.Net), 83.70% (LMT).

Top three ML algorithm in the study for Alva model are RF, K-Star and Nnge that gave the 100% accuracy for training set. This condition may be related to their good generalization ability. Because, random forest known as successful in various domain, since it has good classification and generalization ability (194). 100% accuracy or 1.0 value of other metrics for the training set means that training error is zero. However, as we mentioned above, the training error is not indicative error of the ML model when applied to new data. Cross-validation error is the very good estimate to forecast the error of the model and to select the best model (146). Therefore, in this study, while comparing of the performance of the model, 10-fold CV results were used to select the best model.

Yildirim (2020) reported that F score is more useful than accuracy if there is a class imbalance problem because it takes into account false positive and false negatives predictions (195). Güneş et al., (2021) was used F-score as a primary evaluation metric for their prediction model since the dataset was imbalanced (196). In our study, F score and accuracy results were approximately similar for all model. These results confirmed that there was not a class imbalance problem for our dataset. Chicco and Jurman (2020) reported that when there is class imbalance, MCC produces a more informative and truthful score in evaluating binary classification model than accuracy and F1 score (74). MCC is regarded to be one of the best metrics for binary classification performance since it is calculated all four values of the confusion matrix by considering data imbalance (67, 75). When comparing all the models by the means of results of CV-MCC (Table 28) , since the difference of MCC results were high between internal and external performance, TEST model was considered unsuccessful. On the other hand, PaDEL model did not perform as well as other models for MCC-CV (0.61) but performed good for the evaluation on the test set (MCC-test:0.57). While Mold2 model was considered more

successful than AlvaDesc model in terms of internal performance (MCC-CV: 0.71(Mold2)>0.68(AlvaDesc)). AlvaDesc model was considered more successful than Mold2 model in terms of external performance (MCC-CV: 0.53 (AlvaDesc) >0.44 (Mold2)).

The internal performance of classification models for comedogenicity was assessed by comparing CV-results and Cooper statistics-CV and best model chosen accordingly (34). For AlvaDesc model, Cooper statistics-CV were 0.87(PPV), 0.81(NPV), 0.85(SEN), 0.84(SPC) whereas 0.88(PPV), 0.83(NPV), 0.86(SEN), 0.84(SPC) for Mold2 model. When two models were compared, Cooper-CV results were very close but Mold2 model gave slightly better results than AlvaDesc CV-results. The external performance of classification models was assessed by comparing the Cooper statistics-test set results. For Alva Desc model, Copper statistics-test were 0.78 (PPV), 0.83 (NPV), 0.79(SEN), 0.65(SPC), whereas they were 0.77(PPV), 0.71(NPV), 0.76(SEN), 0.64 (SPC) for Mold2 model. When two models were compared, the predictive power of AlvaDesc model was better than the predictive power of Mold2 model.

5.1.5 The Predictive Ability of Best QSAR Model for Comedogenicity

After optimization of AlvaDesc model, the predictive ability of the Alva_{optimized} model which chosen as best QSAR model in the study, was evaluated on the test set (N=29) for each of ten chemical classes. The stacked bar chart of true/false prediction of QSAR Model_{optimized Alva Desc} on the test set was given in Figure 73. When the predictive ability of the model_{optimized Alva Desc} (with 3 descriptors) was evaluated on the test set for each of ten chemical classes, for class1, while 7 of the 8 predictions were true, 1 of the 8 predictions was false. False prediction was Tridecyl neopentanoate (ID 123). For class2 (alcohols-fatty alcohols-sugars& their derivatives), while 10 of the 12 predictions were true, 2 of the 12 predictions were false. False predictions were Polysorbate 20 (ID 95) and Triacetin (ID 122). For class3 (pigments), 3 of 3 predictions were true. For class4, class6, class7, one prediction was made for each of them. For both class 6 and class 7, prediction was true. However, for class 4, prediction for dimethicone (ID 36) was false. Class5, class9 and class10, there was no prediction due to the fact that there was no chemical represented these classes in the test set. Lastly, for class 8, while 1 of 3 prediction was true, 2 of 3 prediction were false. They were diethylene glycol monoethyl ether (ID 34)

and Octyl salicylate (ID 76). Overall, these prediction results confirmed that chemical class 1 (fatty acid and derivatives), chemical class 2 (alcohols-fatty alcohols-sugars and their derivatives), chemical class 3 (pigments) were the chemical domain of the Alva model with the highest prediction rate.

5.1.6 Applicability Domain for Comedogenicity

The AD of model determines the limits of model for reliable prediction (28). The AD of QSAR models include the descriptor domain, the chemical domain, which is also called chemical space or structural domain, and the mechanistic domain (197). The chemical domain of our QSAR model mainly consist of fatty acids, fatty alcohols and their derivatives and pigments. These chemical classes in the dataset limited the AD of the model for reliable prediction. The model in this study was built only for fatty acids and their derivatives (class 1) such as fatty acid esters and alcohols, fatty alcohols, organic alcohols, sugars and their derivatives (class2) and pigments (class3) to predict the comedogenicity.

In this study, after optimization of the models, Alva model with 3 descriptors were chosen as the best model that consist of Alva_S3K, Alva_ATSC7i and Alva_CATS2D_08_LL. (Table 29). In the context of the descriptor domain of our QSAR model, for each descriptor in the Alva optimized model, Q-Q plot and the boxplot graph was drawn for the training data. It was reported that Q-Q plot is used to compare the distrubution of data to a theoretical distribution such as the normal distribution. The boxplot graph show that how the values in the data were spread out and a standardized way of displaying the distribution of the data through their quartiles. Also, it can be used for the identification of outliers (198). Therefore, for each descriptor Q-Q plot and the boxplot were compared and outlier for training set was determined by the means of boxplot. For the training data of Alva_S3Knormalized, while Q-Q plot (Figure 74) were displaying the two points which were outside the normal distribution. The boxplot of (Figure 75) showing the same two extreme outliers* (ID63=1, and ID115=0.9635) which represented by (*) in the training set. For the training data of Alva_ATSC7inormalized, both Q-Q plot (Figure 76) and the boxplot of (Figure 77) showed some extent outlier but no extreme outliers. For the training data of Alva_CATS2D_08_LLnormalized, while both Q-Q plot (Figure 78) were displaying one point which was outside the normal distribution. The boxplot of CATS2D_08_LLnormalized, (Figure 79) show the same extreme outlier (*) (ID

78=1). As a result, 3 outliers (ID 63*, ID115*, ID78*) were found in the training set of Alva model. Due to fact that normalized value of the test set were calculated by using the $X_{\max}$ and $X_{\min}$ value of training set compounds, test set and training set were compared on the same scale. In Figure 80, Figure 81, Figure 82, the training and the test set were evaluated on the same domain (range) for each descriptor. Therefore, extreme values (*) of the test set showed the outside of the AD. ID92* was found as extreme values (*) of the test set in Figure 80, Figure 81, Figure 82 for each descriptor. Normalized value of ID92 were detected as $ID92_{Alva_S3K}=0.814$ (Figure 80), $ID92_{Alva_ATSC71}=1.345$ (Figure 81) $ID92_{Alva_CATS2D_08_LL}=0.979$ (Figure 82). Also, ID35* was detected as extreme value in the test set for only in Figure 82. As a result, ID92* and ID35* were found as outside of AD.



5.2 Skin Irritation

5.2.1 Evaluation of Training and Test Set for Skin Irritation

Homogeneity and representativeness are important for the division of training and test sets in order to develop a predictive OSAR model. The selection of the training set is crucial and all chemical classes in the dataset should be represented in the training set (127). Therefore, our dataset was divided into clusters for the selection of compounds into the training and test set. In this skin irritation study, the compounds selected from each cluster (n=1,2,3,4,5,6) were distributed approximately in the ratio of 3:1 among the training (n=92) and test set (n=29), the ratio of active and inactive for skin irritation was tried to keep balanced and obtained 0.53 for training and 0.71 for the test set (**Table 32**). For skin irritation, the stack bar chart of chemical class distribution in terms of active/inactive compounds was shown in **Figure 83**, **Figure 84** and **Figure 85**. When comparing the chemical class distribution of training and test set in terms of representativeness (**Figure 84**) and (**Figure 85**), it was similar to the dataset (**Figure 83**). This condition showed that the clusters used for the division of the dataset into the training and test set were homogeneous and represented the chemical class distribution of the dataset. When comparing the distribution of ten chemical class and their active and inactive ratio in the training (**Figure 84**) and test set (**Figure 85**), especially chemical class 1 (fatty acid and derivatives), chemical class 2 (alcohols-fatty alcohols-sugars and their derivatives), chemical class 3 (pigments) were the best representative groups of the dataset in both training and test set.

5.2.2 The Descriptors in the Skin Irritation Models

All descriptors used in the skin irritation models (Mold2, Alva, PaDEL and TEST) were selected by WEKA via the learning-based method since it was found the best successful feature selection method in the comedogenicity studies. According to the literature, chemical reactivity, skin permeability, and molecular size are the important determinants of skin irritation (20). The log P (the logarithm of the octanol/water partition coefficients) is the most important lipophilicity parameter related to the skin permeability in order to predict the skin irritation potential of chemicals (199, 200). Moreover, chemicals first have to penetrate the skin via interacting with the lipid membrane of the skin in order to initiate inflammation which leads to skin irritation (201, 108, 202).

For the skin irritation models we built in this study, Log P index (MLD2_777 (IGA=0,117), first ionization potential (Alva_Mi (IGA=0,326)) and (PDL_AATS0i (IGA= 0,326)) and electronegativity (TST_BELe1 (IGA=0,13)) and PDL_AATS3e (IGA= 0,183)) are the most significant descriptors with the high information gain amount (IGA) which were supported from the literature associated with skin irritation determinants for QSAR modeling.

In the Mold2 skin irritation model, the MLD_777 descriptor represents the molecular regression coefficient surface LogP index. In general, the LogP is the measure of lipophilicity (203). For all QSAR models developed for skin irritation, Log P was related to the skin permeability or penetration of chemicals via interaction with the skin lipid membrane (25, 201, 108). Hayashi et al., (1999) built QSAR skin irritation models for phenols. For their models, LogP was used as a descriptor for skin permeability (23). Moreover, Kodithala et al., (2002) reported that chemical-membrane interaction (MI) is the main event in the toxicity of skin irritation, and also, the predictive QSAR model for skin irritation should be built with the chemical-membrane interaction descriptors. In their QSAR model for phenol and alcohol, the Log P descriptor was used for skin permeability as membrane interaction descriptors (25). Similarly, our Mold2 model contains the Log P as the interaction descriptors with lipid membrane for the skin irritation, thereby it has met the condition for the predictive QSAR model. Furthermore, the Mold2 model has been found as the most successful QSAR model for skin irritation in this study.

2D-descriptors are derived from the molecular structure as a graph of a chemical. They contain information about atomic composition and connectivity of atoms in the molecules such as in the autocorrelation descriptors. They encode property information as in the atom pairs (40). 2D-autocorrelation descriptors are topological descriptors (236), calculated by an algorithm such as ATS (Moreau-Broto), MATS (Moran) from lag 1 to lag 8 for four different weighting schemes. They are shown as lag (topological distance between pairs of atoms) and weight that can be e (Sanderson electronegativity), i (ionization potential), c (charges), v (Van der Waals volume) and m (relative atomic mass). As a result, 2D-autocorrelation descriptors are calculated by adding together the products of atom weight of the terminal atoms of all the paths of the considered lag (204). They encode the relative position of atoms by calculating the division between atom pairs in terms of the number of bonds (205).

In the TEST and PADEL models, electronegativity is one of the important descriptors associated with reactivity for both skin irritation models (26). These descriptors are TST_BELe1 (type of Burden Eigenvalue descriptor) representing the lowest eigenvalue n. 1 of Burden matrix/weighted by atomic Sanderson electronegativities, and, PDL_AATS3e (type of autocorrelation descriptor) representing Average Broto-Moreau autocorrelation - lag 3/weighted by Sanderson electronegativities. Electronegativity can be used to make predictions about the chemical behavior of atoms and elements (206). Also, electronegativity is defined as the ability of an atom to attract electrons related to chemical reactivity (26). Chemical reactivity is dependent on high electronegativity. Elements such as halogens (e.g. F, Cl, Br) are highly reactive due to their high electronegativity (207). Therefore, both TEST and PADEL models provide the descriptors, one of the important determinants of skin irritation, which is chemical reactivity in terms of electronegativity. Furthermore, Hayashi et al., (1999) built the QSAR skin irritation models for a group of phenols regarding chemical reactivity via LUMO (Lowest Unoccupied Molecular Orbital) value and skin permeability via logP (23, 199, 208). Similarly, the electronegativity descriptor of both our models which are PADEL and TEST supports the literature information about QSAR models on skin irritation in terms of chemical reactivity.

In the Alva and PADEL models, the first ionization potential (IP1) is one of the other important descriptors selected by WEKA that can be correlated with chemical reactivity. These descriptors found with the same IGA (0.326) are Alva_Mi (mean first ionization potential that scaled on Carbon atom) and PDL_AATS3e (Average Broto-Moreau autocorrelation - lag 0 / weighted by first ionization potential). The first ionization potential is defined as the amount of energy required to remove one valence electron from a neutral atom (209). When an electron is removed from the atom, the atom is called to be ionized. Furthermore, the first ionization potential is an atomic property that shows the outermost electronic configuration, and besides, IP1 provides information about the binding energy of valence electrons and the degree of relativistic stabilization (209).

Zhan et al., (2003) indicated that there are some linear relationships between the calculated HOMO (the highest occupied molecular orbital energies) and calculated ionization potential (IP) and also between the calculated LUMO (the lowest unoccupied molecular orbital) energies and electron affinity. Therefore, by using these linear

correlations, the calculated HOMO and LUMO energies can be used to predict the ionization potential and electron affinity (206). Besides, both Hayashi et al., (1999) and Kodithala et al., (2002) developed the QSAR model to predict the skin irritation potential of phenols and alcohols (aliphatic and alicyclic). For their QSAR skin irritation model, the calculated HOMO and LUMO energy were used as descriptors for chemical reactivity (23, 25). Since there are linear correlations between the calculated HOMO and LUMO energies and the ionization potential of chemicals, ionization potential can be used as descriptors instead of them for skin irritation models. Therefore, for Alva and PaDEL model, the first ionization potential of molecules can be related to chemical reactivity for skin irritation models.

5.2.3 The Performance of Skin Irritation Models

In the skin irritation study, the performance of the four different classification models built via RF was compared to each other in terms of original (unnormalized) and normalized data (**Table 36**). In general, the models with original data were given better results than the models with normalized data for skin irritation studies, except for the PaDEL model. Moreover, among the all models with original data, the Mold2 model with 8 descriptors was the best QSAR model for skin irritation. The classification accuracy of the Mold2 model with RF was given the best performance with 85,87% accuracy for the cross-validated model also, and the predictive ability of the Mold2 model on the test was 82,75 % accuracy. Furthermore, the Mold2 model was a higher F-score, MCC, and AUC than other models with 0.86, 0.68, and 0.88 respectively for the cross-validated model, also and, the predictive ability of the Mold2 model on the test set with 0.83, 0.64, and 0.85 respectively.

According to the literature on skin irritation/corrosion for the QSAR models, there are only available QSAR models developed for some specific chemical classes such as phenols, alcohols, esters, organic acids, and bases. Also, most of these QSAR models have been developed from small datasets (208). Similarly, our Mold2 QSAR model for skin irritation was developed for fatty acids and their derivatives, fatty alcohols, sugar and their derivatives and pigments, also developed from the small dataset 92/28, training and test set. The underlying cellular and molecular mechanism of skin irritation/corrosion are still poorly understood (208). In the regulatory context, while skin corrosion is irreversible damage of the skin, which is called visible necrosis through the epidermis

and into the dermis, skin irritation is the reversible damage of the skin after the application of a chemical for up to 4 hours, based on the *in vivo* animal test (208). Besides, the proposed mechanistic interpretation for the QSAR skin irritation model has been discussed by Walker et al., (2004) and Rorije & Hulzebos (2005) that the most organic chemicals first have to penetrate before being reactive to the skin (202, 210). In our skin irritation study, when we evaluated the skin irritation mechanism of the best QSAR model, the Mold2 model with eight descriptors, the surface LogP index (MLD2_D777) descriptor has been explained similarly to the literature that skin penetration mechanism is related to the lipophilicity of molecules.

There are various SAR/QSAR models for skin irritation. Based on the rabbit data, they indicated that irritants have higher logP and dipole moments but lower molecular volume than non-irritants. Furthermore, their authors proposed that irritants show enhanced permeability and higher reactivity (24). Hayashi et al., (1999) built a skin irritation QSAR model for 24 phenols related to reactivity (positive LUMO value) and skin permeability (LogP). Their skin irritation model is a linear regression model ($r^2=0.67$) in which the skin irritation mechanism of phenols suggests the reactivity of phenols with skin proteins and amino acids (23, 199, 208). Phenols with positive LUMO values were considered as electron donors by reacting with skin amino acids in both the epidermis and dermis. As a result, they result in skin irritation. On the other hand, phenols with negative LUMO values were considered electron donors by only reacting with skin amino acids in the epidermis, resulting in relatively low irritation (199). The reactivity difference among phenols was found as dependent on both their electron releasing or electron attracting substituent and also whether the LUMO energy is positive or negative (23, 199).

Smith et al., (2000) built a SAR model for 42 esters to discriminate irritants from non-irritants. In the classification modeling studies, regression analysis and linear discriminant analysis were used. Their study indicates that physicochemical features of esters have a contribution to their irritant effects which are chemical partitioning into the epidermis and intermolecular reactions are important (24). Kodithala et al., (2002) built the chemical-membrane interaction (MI)-QSAR models for 22 hydroxy-organic acids ($r^2=0.54-0.86$), 13 aliphatic alcohols ($r^2=0.95$), and 11 phenols ($r^2=0.52$) to predict skin irritation. They are linear regression models ($r^2:0.52-0.95$) that the prediction

performance has been increased with 1st) effective concentration of chemicals in the skin lipid membrane (phospholipid-rich region of cellular membrane) 2nd) the binding capacity of chemicals in the skin lipid membrane (phospholipid-rich region of cellular membrane) 3th) the reactivity of the compounds with skin protein via HOMO and LUMO of the molecules (25, 199).

In general, literature SAR/QSAR models (208) were linear models developed with linear regression analysis (20, 199), our QSAR models were non-linear models developed with a non-linear algorithm, namely random forest. While the linear models were evaluated using correlation coefficient (r^2), our non-linear SAR models were evaluated using classification accuracy. In the literature, while linear regression found the correlation, in other words, the linear relationship between the structure/property and the descriptors (20), the non-linear algorithms found the non-linear relationship between the structure/property and the descriptors in our models. Golla et al., (2009) developed one of the first non-linear QSPR skin irritation models based on the Draize rabbit test data, for 186 various classes of chemicals (20). The optimal set of descriptors was chosen with the artificial neural network analysis (ANN). The prediction potential of their QSPR model was found successful with an r^2 of 0,78. In terms of mechanistic interpretation of the model, their descriptors displayed a good correlation with skin penetration, reactivity, and molecular size (20).

In the literature for skin irritation, there are commercial computerized models for organic chemicals. They are TOPKAT, MultiCASE, MolCode, DEREK Nexus, HazardExpert, ACD/Percepta. Also, BfR rule-based which is free programs including skin irritation are OECD QSAR Toolbox, ToxTree, Toxmatch, and ToxPredict (208). These computerized models are the EXPERT SYSTEMS based on the use of knowledge to determine the skin irritation potential of a chemical. They find the skin irritation either by relating to available physicochemical properties of chemicals or via pattern recognition algorithms (20). In a similar way, our QSAR models are machine learning (ML) models which can be computerized with more data in the future. Also, ML-based models can be re-developed by adding and removing new descriptors.

The pKa is one of the important molecular properties that affect the skin irritation potential of chemicals, especially for organic acids and bases (24, 211). The pKa is described as the negative base-10 logarithm of the acid dissociation constant (Ka) in the solution of a chemical. Berner et al., (1988) reported that the relationship between skin irritation and pKa was found to be linearly correlated for $pK_a \leq 4$. When pKa values ≤ 4 for acids and $pK_a > 8$ for bases, skin irritation was rapidly increased and becomes significant, and also, skin irritation indicators (e.g erythema and edema) appeared (199, 212, 213). Furthermore, for organic acids and bases, the severity of skin irritation was found to correlate with the strength of the acid or base, since it was measured by their pKa value (24, 212). The pKa is the measure of cytotoxicity for the skin corrosivity model and also, the pKa is highly predictive for acids and bases in terms of acute skin irritation (22, 214).

Barrat (1995) derived QSARs for skin corrosivity of 20 organic acids, 21 organic bases, and 15 phenols by the means of principal component analysis. In data analysis, skin corrosivity data of chemicals was evaluated associated with their log P, pKa, melting point, and molecular volume. The skin permeability (Log P) and cytotoxicity (pKa) for each chemical class showed good performance to discriminate skin corrosive chemicals from non-corrosive ones (22).

Whittle et al., (1996) derived QSAR to determine the skin corrosivity potential of some fatty acids by the means of principal component analysis with the dataset of 27 organic acids (215). In that study, the corrosive potential of some fatty acids, which are propanoic acid (C3), butanoic acid (C4), hexanoic acid (C6), octanoic acid (C8), decanoic acid (C10) and dodecanoic acid (C12), were studied in vitro skin corrosivity test using both rat skin and human skin. All the fatty acids with alkyl chain lengths from C3 up to and including C8 which are propanoic acid (C3), butanoic acid (C4), hexanoic acid (C6), octanoic acid (C8) were found as corrosive to rat skin. The corrosivity mechanism of organic acids was considered a function of the ability of the chemical to penetrate the skin (LogP) and its cytotoxicity (pKa) (215). The variables that affect skin permeability are LogP, melting point and molecular volume. Additionally, cytotoxicity is the reactivity parameter measured via the pKa value of the fatty acids for skin corrosivity. The organic acids with lower LogP, higher melting points, and larger molecular volumes were related to the lower skin permeability. For fatty acids, the increase in melting point with an increasing alkyl chain of 10 carbon and longer are found non-corrosive to rabbit skin,

however, fatty acids with C3 up to C8 were found corrosive to rabbit skin (215). Similarly for our study, the dataset that we trained our QSAR models contains the fatty acids which are caprylic acid (C8-octanoic acid, active), lauric acid (C12-dodecanoic acid, active), myristic acid (C14-tetradecanoic acid, inactive), stearic acid (C18, inactive) for skin irritation. When the predictive ability of our best QSAR Mold2 model was evaluated on the test set. The two fatty acids included in the test set were both predicted correctly. They were capric acid (C10-actual: active, predicted: active) and palmitic acid (C16-actual: inactive, predicted: inactive). As a result, the predictive ability of the Mold2 model gave good performance for both two fatty acids, one is corrosive and another one is non-corrosive.

5.2.4 Skin Irritation Mechanism of Models

According to the literature on QSAR for skin irritation, skin permeability and reactivity are the most important parameters in terms of the mechanism of the skin corrosivity/irritation (20, 24, 23, 25, 199, 208, 211). Similarly, we found the Log P index (MLD2_D777) with 0.117 of IGA for skin permeability and the first ionization potential (Alva_Mi and PDL_AATS0i) with 0.326 of IGA for reactivity as being the most important descriptors to explain the skin irritation mechanism of our study. Therefore, our descriptors are consistent with the literature on skin irritation.

Eskes (2016) reported that the underlying mechanisms of skin irritation are still poorly understood (208). Welss et al., (2004) reported that chemicals initiate and modulate skin irritation by two distinct pathways. Firstly, they gave damage to the barrier function of the stratum corneum, and secondly, irritants directly affect the cells of the skin. These two pathways can lead to irritation either alone or in combination (108). Fluhr et al., (2008) reported that the main pathological mechanisms of skin irritation include skin barrier disruption, induction of a cytokine cascade and involvement of the oxidative stress network that results in inflammatory reaction (216).

In the context of the relationship between skin irritation and ROS, there are some studies; It is found that at a cellular level, irritants induced oxidative stress via significant modification in the physiology of keratinocytes. Keratinocytes naturally produce ROS as a second messenger and regulate the protein kinase level and activity. However, at the non-physiological increased level of ROS can lead to cellular damage via oxidizing membrane lipids and proteins that result in cytotoxicity and modified gene expression

Also, the dose-dependent irritant response of tributyltin (TBT), a potent skin irritant (217), was found when applied to the ear of Balb/c mice that may be related to ROS (108). Additionally, Shen et al., (2008) investigated the relationship between trichloroethylene (TCE)-induced skin irritation and oxidative stress on the BALB/c hairless mice skin, they found the acute TCE-irritation dependent on concentration, the increased MDA levels, and inhibited SOD activities of the mice skin (218). Some studies with ROS-generated systems such as intra-dermally injected ROS proved the involvement of ROS in skin irritation (108). Moreover, for the verification of the skin irritant-induced oxidative stress, firstly, skin irritation was induced by an anionic surfactant (lauryl sarcosine %1) via generation of superoxide anion ($\bullet\text{O}_2^-$) and then ROS-scavenging vitamin E (alpha-tocopherol acetate%30) was applied to the rat skin. As a result, skin irritation was significantly reduced in terms of erythema score. Also, vitamin E protected the cultured keratinocytes from cell damage by the means of reduction of superoxide anion levels (108, 219). ROS includes superoxide anion radical ($\bullet\text{O}_2^-$), hydroxy radical ($\text{HO}\bullet$), peroxy radical ($\text{ROO}\bullet$) and also non-radical species, such as hydrogen peroxide (H_2O_2), lipid peroxides, hypochlorous acid (HOCl) and singlet oxygen ($^1\text{O}_2$) (220). Therefore, it may be concluded that anionic surfactants as well-known potent irritants (221) may be the sources of ROS for the skin. Mizutani et al., (2016) stated that another anionic surfactant, sodium lauryl sulfate (SLS) stimulates the production of ROS through the interaction with skin cell membranes (225). Also, Effendy and Maibach, (1995) reported that SLS interacts with the skin causing damage to the barrier properties of the stratum corneum (SC); both disrupting of lipid and protein structure of SC (221). On the other hand, the skin has a lipid membrane, that acts as a lipid barrier to the environment and chemicals. This barrier function of the skin depends on the lipid composition of skin called "SC lipid matrix" in the outermost layer of skin, the stratum corneum (SC). The composition of the SC lipid matrix is dominated by three lipid classes which are fatty acids, cholesterol, and ceramides (222, 223). Skin irritants (e.g. 1% SLS) caused the change in lipid content and organization of SC, which disrupts the barrier function of SC (224).

In our skin irritation study, we found the first ionization potential with the highest information gain (IGA) is 0.326 which indicates the reactivity for both two models (Alva and PADEL). In our comedogenicity studies, ionization potential was found related to ROS (reactive oxygen species) producing potential of substances. Moreover, Mizutani et

al., (2016) reported that a well-known skin irritant, SLS stimulates the production of ROS via interaction with the skin lipid membrane (225). Additionally, SLS disrupts the lipid and protein structure of the skin lipid barrier (224). Therefore, in summary, due to the ROS-producing potential of skin irritants, they may attack the skin lipid membrane and reacts via lipid peroxidation. Also, by-products of lipid peroxidation, which are lipid peroxides in the early stage, and aldehydes such as MDA (malondialdehyde) and 4-HNE (4-hydroxynonenal) in the late stage, may give damage to the protein in the skin lipid membrane since they can make adducts with protein or modify protein. As a result, the formation of lipid peroxide-modified protein or aldehyde-modified protein such as MDA-protein adducts and 4-HNE-protein adducts leads to damage to the protein structure of the skin lipid membrane (90, 93).

Our QSAR models for skin irritation were developed mainly to predict the irritation potential of fatty acids and their derivatives, fatty alcohols, sugars, and their derivatives due to the chemical domain of our training set. Fatty acids are carboxylic acids, in other words, organic acids with long aliphatic tails, either saturated or unsaturated. All the fatty acids with alkyl chain lengths from C3 to C8 were found corrosive to rat skin (215). Therefore, caprylic acid (C8-octanoic acid) and capric acid (C10) that belongs to our dataset are skin corrosive. According to the regulatory context, skin corrosion is defined, based on the in vivo animal test, as “the production of irreversible damage of the skin; expressly, “visible necrosis” through the epidermis and into the dermis, following the application of test chemicals for up to 4 hours” (18).

Necrosis is cell death by the way of injury or disease, particularly in a localized area of a tissue organ (226). Recent studies have shown the role of lipid peroxidation in cell death both necrosis and apoptosis (Figure 90) (90, 227, 228). One of the lipid peroxide degradation products and also a biomarker of lipid peroxidation, 4-HNE induces apoptosis or necrosis depending on the 4-HNE concentration (90) (Figure 92). ROS induce oxidative damage in cells lipid membrane via lipid peroxidation, when oxidative stress exceeds the antioxidant capacity (Figure 90). ROS induced oxidative stress can result in cellular damage (Figure 91 and Figure 92).

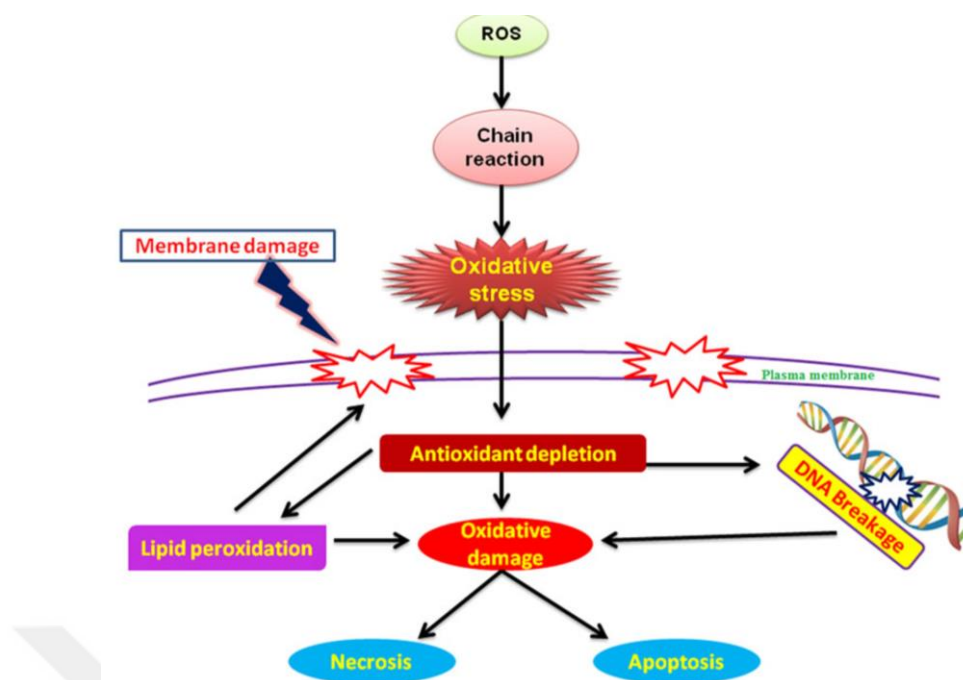


Figure 90. Cellular damage induced by oxidative stress (228)

While a moderate level of lipid peroxidation induces the apoptotic death, a high level of lipid peroxidation induces necrotic death (90, 96, 228) as shown in Figure 90, Figure 91 and Figure 92.

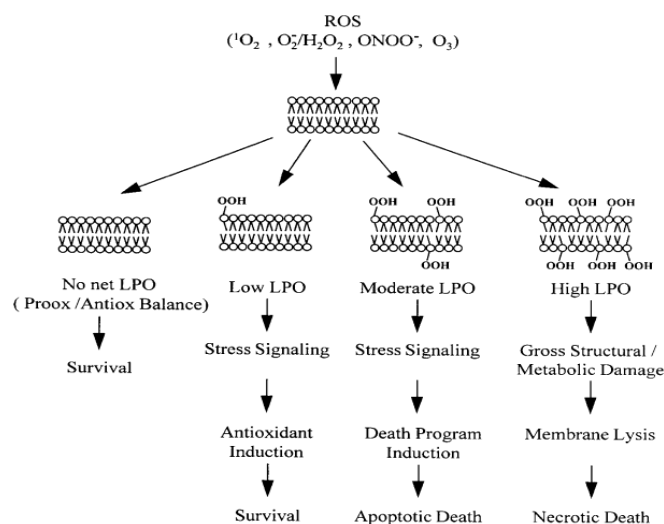


Figure 91. The possible cellular response to ROS in lipid membrane (96)

4-HNE is a highly reactive end product of lipid peroxidation. It is known that 4-HNE is a potent regulator of the various cell process. These are oxidative stress signalling, cell proliferation, transformation, and cell death (229). It can make protein adducts and

modifiy the cell signalling. The accumulation of 4-HNE is found in various oxidative-stress-related disease such as metabolic syndrome (229).

The lipid peroxidation end product of 4-HNE-Induced Apoptosis and Necrosis are displayed in Figure 91. Moreover, 4-HNE can advocate cell survival or induce cell death depending on cell type, damage/repair capacities and cellular metabolic condition (90).

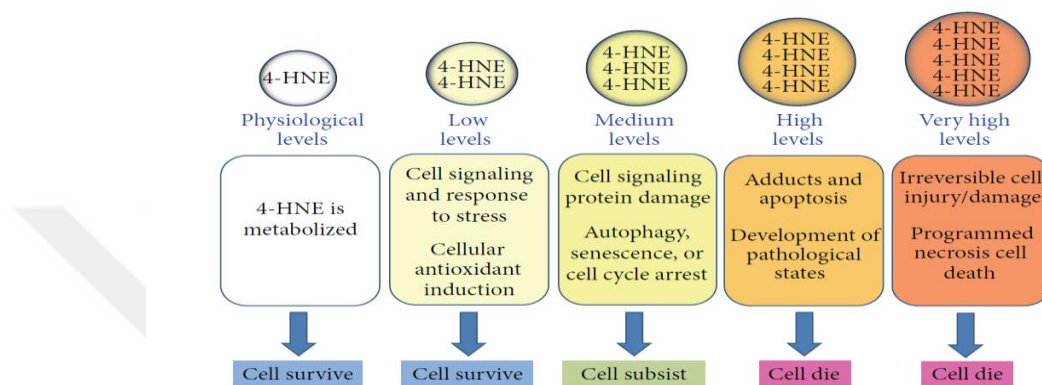


Figure 92. 4-HNE advocate cell survival or induces cell death (90)

At low levels of 4-HNE serves as signaling molecules, stimulating gene expression, increasing the cellular antioxidant capacity and employing of the adaptive response; at medium levels of 4-HNE result in organelle and protein damage leading to induction of autophagy, senescence or cell cycle arrest; at high and very high levels of 4-HNE advocate adducts formation and apoptosis or necrotic cell death (90).

In summary, recent studies indicate that necrosis is the cell death promoted by the high level of 4-HNE which is the lipid peroxidation end product due to the fact that ROS-induced oxidative damage in the cellular lipid membrane. Therefore, we can conclude that skin corrosivity may be related to chemical-induced oxidative damage via lipid peroxidation. Furthermore, skin corrosive our dataset contains, caprylic acid (C8) and capric acid(C10) may show their corrosive effects depending on their ROS producing potential.

5.2.5 The Predictive Ability of the Best QSAR Model for Skin Irritation

The predictive ability of the Mold2 skin irritation model, which was chosen as the best QSAR model for skin irritation, was evaluated on the test set (N=29) for each of the ten chemical classes. The stacked bar chart of true/false prediction of QSAR Model Mold2 on the test set was given in **Figure 86**. When the predictive ability of the modelMold2 (with 8 descriptors) was evaluated on the test set for each of ten chemical classes, for class1 (fatty acids and their derivatives), while 6 of the 7 predictions were true, 1 of the 7 predictions was false. False prediction was isostearyl isostearate (ID62). For class 2 (alcohols-fatty alcohols-sugars& their derivatives), while 10 of the 13 predictions were true, 3 of the 13 predictions were false. The false predictions were ethylene glycol monostearate (ID45), glyceryl stearate (ID51), and hexylene glycol (ID52). For class 3 (pigments), 3 of 3 predictions were true. For class 4 (silicones) and class 8 (miscellaneous), one prediction made was true. For class5, class9 and class10, there was no prediction due to the fact that there was no chemical representing these classes in the test set. For class6, while 1 of the 2 predictions was true, 1 of the 2 predictions was false. The false prediction was alpha-tocopherol (ID5). Overall, these prediction results confirmed that chemical class1 (fatty acids and their derivatives), chemical class2 (alcohols-fatty alcohols-sugar and their derivatives), and chemical class 3 (pigments) were the chemical domain of the QSAR Model Mold2 for skin irritation with the highest prediction rate.

5.3 The Relationship Between Comedogenicity and Skin Irritation

5.3.1 Evaluation of the Dataset

Our dataset is mainly composed of mostly comedogenics and fewer of them are irritants (61.2% and %36.4) This condition shows that our data is more sufficient to develop an accurate comedogenicity model than an accurate irritation model. While more than half of the comedogenic substances (55.4%) were also found to be irritants. Very few of the non-comedogenic substances were found to be irritants (6.4%). It can be concluded that comedogenicity and irritation can be found together. Since the comedogenic potential of a substance gives damage to the lipid barrier of SC or skin through lipid peroxidation, it creates a suitable environment for the penetration of the substance to show its irritation potential by initiating inflammation. Almost all irritant substances, 93.2% of irritants were found to be comedogenic, while 43% of non-irritants were found to be comedogenic. This ratio can indicate that the possibility of an irritant being a comedogenic is high at 93.2 percent however, the possibility of a non-irritant being a comedogenic is lower at 43.2 percent. Although not all comedogenic substances cause inflammation, some of them can result in inflammation. This may be proportional to the amount of lipid peroxide occurred. The higher lipid peroxidation may be more likely for the substance is being both comedogenic and irritant. Furthermore, the fact that a substance being irritant can be an indication that there will be inflammation in the environment.

5.3.2 Evaluation of the Relationship Between Comedogenicity and Irritation

In this study, we evaluated the relationship between the comedogenic and irritant potential of 121 cosmetic ingredients using chi-square test to determine whether there was a significant association between them. The results show that there was statistically significant associations between the comedogenicity and irritation at the significance level ($p < 0.001$).

As a toxic response, comedogenicity is initiated hyperkeratinization via lipid peroxidation, whereas skin irritation is initiated inflammation (80, 82, 88, 108). According to the literature, both of these toxic responses occur as a result of the changes in the skin lipid composition in skin lipid membrane which acts as a lipid barrier (100, 108).

In skin irritation, it is known that irritants enter the stratum corneum by leading to delipidation which is the process in which the lipid balance of the skin lipid membrane is disrupted and results in the change of lipid cement composition and loss of the barrier function of the skin (108). In comedogenicity which is the initial lesion of acne, the lipid composition of the skin membrane is changed via lipid peroxidation of fatty acids in the sebum (80, 82, 88, 89, 98). Furthermore, for lipid peroxidation, skin membrane lipids are predominantly susceptible to oxidative attacks of $\text{HO}\bullet$ due to the presence of polyunsaturated fatty acids in membrane phospholipids. Also, polyunsaturated fatty acids in the skin serve as excellent substrates for peroxidation (91, 93).

On the other hand, skin acts a barrier for environmental effects. For this skin barrier function, the lipid composition of skin called SC lipid matrix (Figure 94) in the outermost layer of skin, the stratum corneum (SC) has crucial importance. In various skin diseases, the SC lipid barrier is affected such as atopic dermatitis (222). The barrier function of the skin depends on the composition of skin lipids. The composition of the SC lipid matrix is mainly composed of three lipid classes which are fatty acids, cholesterol, and ceramides (222).

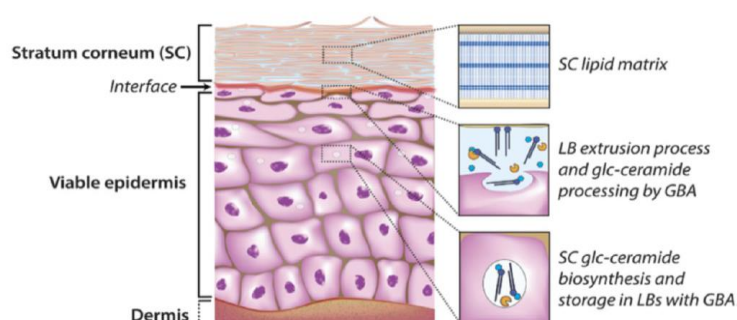


Figure 93. SC lipid matrix in the skin (223).

In our study, our dataset mainly composed of fatty acids and their esters & derivatives and also, fatty alcohol and their derivatives. Therefore, our dataset had similar lipid structures to the SC lipid matrix due to being that fatty acids are one of the main components of SC. As a result, this condition provided our study an advantage to mimic and model the similar mechanism. According to the results shown in **Table 39**, 93.2% of irritants were found comedogenic, whereas 55.4 % of comedogenic substances were found irritant. Therefore, we can conclude that while the probability of an irritant being comedogenic is as high as 93.2 percent, the probability of a comedogenic substance being irritant is lower and found 55.4 percent. Furthermore, the Cramer's value= 0.497, which indicates the strength of association, suggested a relatively strong association between

comedogenicity and irritation (**Table 40**). This relatively strong association may be explained by the interaction of both irritants and comedogenic substances with the skin lipid membrane or the lipid matrix of the stratum corneum (108). These interactions may be explained via lipid peroxidation in the skin lipid membrane.

It is well-known that skin irritants such as surfactants (SLS) disrupt both lipid and protein structure of SC of the skin (221). Also, it is shown that skin irritants (e.g. %1 SLS) caused the changes in the lipid content and the lipid organization of the SC lipid matrix (Figure 95) (224). Furthermore, chemical reactivity (pKa, electronegativity), skin membrane permeability (LogP) and chemical-membrane interaction are the main important skin irritation determinants for modeling (20, 25, 215). Additionally, it is reported that SLS stimulates the production of ROS via the interaction with skin cell membranes (225). Lastly but most importantly, irritants without membrane-damaging properties do not initiate the inflammatory cascade alone by the release of IL-alpha for skin irritation (108). In the light of this information, we may conclude that ROS produced by skin irritants may attack the skin lipid membrane, and skin lipids may react with ROS via lipid peroxidation, thereby, the lipid structure of the skin is disrupted due to being skin lipid peroxidized that results in changes in lipid composition of SC lipid matrix and skin permeability. Moreover, by-products of lipid peroxidation such as MDA or 4-HNE may be modified the protein structure of the skin.

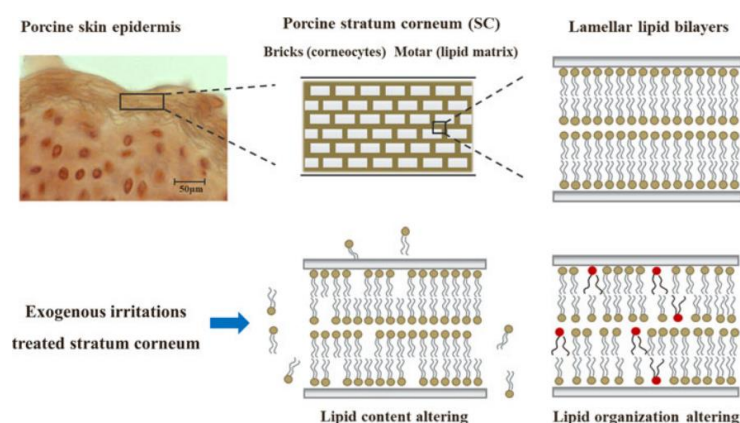


Figure 94. Alteration of SC lipid content due to exogenous irritants, 1% of sodium lauryl sulfate (224).

On the other hand, comedogenic substances can be considered suspicious for oxidative attacks via lipid peroxidation to skin lipid membrane due to containing of polyunsaturated fatty acids (80, 88, 91, 93). As a result, as the lipid balance or lipid

composition of skin membrane is changed by lipid peroxidation, comedogenic substance can disrupt the barrier function of skin for irritant substances.

Secondly, lipid peroxide (LOOH) caused by the oxidation of lipid in the skin lipid membrane via lipid peroxidation may be stimulated both comedogenicity and irritation. According to the literature on lipid peroxide, various studies reported that lipid peroxide caused by the sebum oxidation is related to comedogenicity and the progression of acne. Because excess amount of lipid peroxide was found involved in the inflammatory changes of acne (88). Squalene is a polyunsaturated sebaceous lipid, precursor of steroids, found in skin surface lipids (99, 230). It is reported that while squalene is too small to be detected as comedogenic, squalene oxidation products that is squalene peroxide has been demonstrated as highly comedogenic on rabbit ear skin. Also, the level of squalene peroxide was determined to correlate positively with the size of comedones (99, 231). Furthermore, squalene peroxide-induced comedones showed noticeable hyperkeratinization (hyperkeratosis of epithelium) and noticeable hyperplasia in the pilosebaceous unit. Moreover, squalene peroxide displayed keratinocytes cytotoxicity in vitro skin culture test (99). In addition to this, Ottovani et al., (2006) reported that the involvement of lipid peroxide, particularly squalene peroxide has inflammatory effect on acne (230). Also, by product of lipid peroxidation are able to induce production of pro-inflammatory cytokines and activation of peroxisome proliferators-activated receptor (PPARs) which involved in the control of lipid metabolism and inflammation (99, 100).

In terms of skin irritation, the primary peroxidation product of human sebum squalene monohydroperoxidase has been found a potent primary skin irritant and also potent comedogenic (232). Tanaka and Hayakawa (1986) investigated the effect of lipid peroxide on skin irritation by testing UV-irradiated cosmetic cream containing high levels of lipid peroxide on guinea pig and human skin. In the histological examination of guinea pigs, they found intraepidermal edema, marked spongiosis, hyperkeratosis, dyskeratosis in the epidermis and small cell infiltration in the dermis. For the human test, positive reactions were observed in only 2 patients among 50 patients (233). This condition may be explained by the oxidant/antioxidant balance of the skin cells between two groups. Lipid peroxide participates in redox reaction so that peroxidative injury can be worsen or prevented dependent on the redox status of cell. (96). Especially, glutathione peroxidase (GPx) enzymes are key regulators of lipid peroxides in cells. Inactivation of GPx cause accumulation of lipid peroxide and result in cell death and tissue damage (227).

Lipid peroxides have a crucial role in the regulation and progression of inflammation due to their high reactivity and ability of generate toxic secondary messengers that result in multiple pathologies and cell death. (227). Also, lipid peroxide can exert cytotoxic and genotoxic effect (234)

The biosynthesis of lipid peroxide occurs in two ways. These are via enzymes or non-enzymatic lipid peroxidation. The non-enzymatic lipid peroxidation is mediated by carbon-and-oxygen-centered radical (with Fenton-type chemistry; redox reaction with iron). It consists of three steps; initiation, propagation and termination. In the initiation steps, a radical compound was produced from non-radical molecules by abstracting hydrogen. Highly reactive radicals both hydroxy and peroxy radicals generated in Fenton reactions are capable of abstracting a hydrogen atom from polyunsaturated lipid in the cell membrane. Thus generating a lipid peroxide (LOOH). Radical compounds cause to new radicals that are considered a chain reaction, called propagation step. Antioxidant molecules can terminate the radical propagation by donating their electron called termination step (227).

6. CONCLUSION

6.1 Comedogenicity

In the comedogenicity studies, we compared two feature selection method for significant descriptor selection for the models and then we developed total twelve different ML-based QSAR models using three classifiers (KStar, RF, NNge) by the means of the selected descriptors for four descriptor packages (AlvaDesc, Mold2, TEST, and PaDEL) via learning-based method to predict comedogenic potential of some cosmetic ingredients. The main results for the comedogenicity studies can be listed as follows:

- The learning-based feature selection method led to simpler and more interpretable models than correlation-based method by reducing the number of descriptors by the range of 20-85%, especially 70% reduction of descriptors for Alva model, 60% reduction of descriptors for Mold2 model, while models's fit and predictivity almost unchanged. Also, learning-based method provided us to screen wide range of descriptors sets calculated from different software.
- Random Forest gave the best results as a classifier in the modelling studies for comedogenicity prediction.
- Mold2 and AlvaDesc model with RF gave the successful results with 85.87% and 84.78% accuracy for the cross-validated models and 75.86% and 79.31% accuracy for the test set.
- Our two successful models meet the first condition for model validity with the Topliss ratio 13.1 for Mold2 and 18.4 for AlvaDesc (with 5 descriptors),
- The AUC results of cross-validated two models are as follows: 0.89 for Mold2 and 0.87 for AlvaDesc. Accordingly (Section 3.2.10), our two cross-validated models can be characterized as good.
- The AUC results on test set for two models are 0.76 for Mold2 and 0.81 for AlvaDesc. Accordingly (Section 3.2.10), the predictive ability of Mold2 model is characterized as fair, while predictive ability of AlvaDesc model is characterized as good.
- After the optimization of the comedogenicity models, Alva with three descriptor was chosen as the best model in the comedogenicity study. The performance of the model was improved by removing two irrelevant features from the model, while the accuracy of the test set remained the same (79.31%), the accuracy of the cross-validated model

was increased from 84.78% to 86.96%. Thereby, overfitting was reduced and the number of descriptors in Alva model was reduced from 5 to 3 descriptors.

- Topliss ratio (model validity ratio) was increased from 18.4 to 30.66 for the best model (Alvadesc with 3 descriptors) after optimization study.
- In this study, CATS2D08_LL (lipophilicity-related pharmacophore descriptors) and ATSC7i (ionization potential related-autocorrelation descriptors) are the most significant descriptors with highest information gain amount (IGA) which are 0.314 and 0.336 respectively for the best QSAR model.
- The chemical domain of best model mainly consist of fatty acids and their derivatives such as fatty acid esters and alcohols, fatty alcohols, sugars and their derivatives and pigments to predict the comedogenicity. These chemical classess in the dataset limited the AD of the model for reliable prediction.
- The descriptor domain of best QSAR model is in the range of 0.003-47.00 for Alva_CATS2D_08_LL descriptors, the range of 0.003-4.23 for Alva_ATSC7i and the range of 0.825-80.08 for Alva_S3K descriptors.
- While, the three compounds were found as outliers which are ID63 (Laureth-23), ID115 (Stearath-100) and ID78 (Octyldodecylstearoyl stearate), two compounds were found as out of AD which were ID92 (Pentaerythrical tetraisostearate) and ID35 (Diisopropyl dimerate) for Alva Model optimized with 3 descriptors.

In the comedogenicity studies, we concluded that comedogenicity is a ROS-induced hyperkeratinization. As a best QSAR model for comedogenicity, Alva model with 3 descriptor gives successful results with 86.96 % prediction accuracy for the cross-validated models and 79.31% prediction accuracy for the test set. For our best QSAR model, the lipophilicity-related molecular behaviour (CATS2D08 LL) and ionization potential (ATSC7i) due to ROS producing potential of substances are important descriptors for comedogenicity prediction via lipid peroxidation.

6.2 Skin Irritation

In the skin irritation studies, we developed four different non-linear classification models via Random Forest algorithm by the means of four descriptor packages (AlvaDesc, Mold2, TEST, and PaDEL) using the learning-based descriptor selection method. As the best skin irritation model, the QSAR Model Mold2 with eight descriptors gave a successful performance with 85.75% prediction accuracy for 10-fold cross-validation and 82.75% prediction accuracy for the test set. For our best skin irritation model, the surface LogP index (MLD2_D777, IGA= 0,117) was the most important descriptor related to skin permeability. Since chemicals first have to penetrate the skin via interacting with skin lipid membrane to initiate inflammation that leading to skin irritation. Therefore, the LogP index is one of the most important descriptors found in this study. Besides log P index, first ionization potential for both two models (AlvaDesc and PaDEL) with the same and highest information gain (IGA=0,326 for both Alva_Mi and PDL_AATS0i) and electronegativity for both two models (TEST and PADEL, IGA= 0,13 and IGA=0,183, TST_BELe1, and PDL_AATS3e) were the most significant descriptors which were supported from the literature associated with reactivity as the skin irritation determinants for QSAR modeling. In the literature, the previous skin irritation models were developed only for some specific chemical classes such as phenols, alcohols, esters, organic acids, and bases. Similarly, the chemical domain of our QSAR Model Mold2 for skin irritation covers the fatty acids and their derivatives, fatty alcohols and their derivatives, and pigments depending on our training set.

According to the literature, skin permeability and reactivity are the most important parameters in terms of the mechanism of skin corrosivity/irritation. Similarly, we found the LogP index related to skin permeability and the first ionization potential related reactivity as being the most important descriptors to explain the skin irritation mechanism of our study. Therefore, our descriptors are consistent with the literature on skin irritation.

On the other hand, the underlying mechanisms of skin irritation are still poorly understood. Ionization potential was found in both comedogenicity and irritation studies as the most important descriptor with the highest information gain. While ionization potential was correlated with the logarithm of the ROS-producing potential of substances in the comedogenicity studies, it was associated with reactivity in the skin irritation studies. Furthermore, in the literature, there are some studies about skin irritants-induced

oxidative stress. Moreover, well-known skin irritants, particularly, anionic surfactants was found as stimulating the production of ROS via the interaction with skin cell lipid membranes. Besides, they disrupt both lipid and protein structure of SC of the skin as like occurs in the lipid peroxidation. In briefly, according to the conclusion we obtained from this skin irritation study and literature review about skin irritant induced-ROS, they brings to mind the question that “Is skin irritation/corrosion a ROS-induced inflammation via lipid peroxidation?”

6.3 The Relationship Between Comedogenicity and Skin Irritation

In terms of the relationship between comedogenicity and skin irritation, there was found a statistically significant association between them ($p < 0.001$). This relatively strong association may be explained by the interaction of both irritants and comedogenic substances with the skin lipid membrane via lipid peroxidation. According to the results obtained from our dataset, while the probability of an irritant being comedogenic is as high as 93.2 percent, the probability of a comedogenic substance being irritant is lower and found 55.4 percent.

Ionization potential was found in both comedogenicity and irritation studies as the most important descriptor with the highest information gain. Ionization potential describing as the amount of energy required to remove an electron from an atom or a molecule. The removal of an electron or a hydrogen atom from target molecules means the oxidation of target molecules. Particularly, during the metabolization process of molecules by CYP P450 enzyme system, this system produces free radicals from chemicals. Accordingly, the ionization potential of molecules may be related to their ROS-producing potential for both comedogenicity and skin irritation.

The composition of skin surface lipids has been changed in both comedogenicity and skin irritation. This alteration in SC lipid content can be due to the oxidation of lipid via lipid peroxidation which results in the disruption of lipid in the skin membrane. Also, lipid peroxide, by-product of lipid peroxidation, depending on the level of lipid peroxide can be modified the protein in skin lipid membrane. Moreover, a high level of lipid peroxide can result in oxidative damage such as cell death (apoptotic or necrotic) depending on the oxidant/anti-oxidant balance of the skin cell.

Lipid peroxides can stimulate both comedogenicity and irritation. On the rabbit ear test, while squalene peroxide-induced comedones has been showed noticeable hyperkeratinization and noticeable hyperplasia in the pilosebaceous unit, squalene monohydroperoxidase has been found as potent primary skin irritant on the rabbits. In vitro test on human keratinocytes cell lines (HaCAT), studies displayed that squalene peroxide had a proliferative effect on human keratinocytes which suggests that lipid peroxide initiates the hyperkeratinization. Also, squalene peroxidase induces both the upregulation of PPAR- α at mRNA level which involved in the inflammatory cascade and production of pro-inflammatory cytokines (IL-1, IL-6, IL-8).

In conclusion, this thesis study has provided an animal-friendly, rapidly perform and inexpensive two different classification model to predict the comedogenic and skin irritation potential of some cosmetic ingredients for the rationale of cosmetic product design. Also, by explaining the possible mechanism of both models has been made contribution to literature. Furthermore, both models can be potentially used for the safety assessment of cosmetic products for fatty acids and their derivatives.

7. REFERENCES

1. European (EU) Commission, Press Release (2013). Full EU ban on animal testing for cosmetics enters into force, Brussels, March 11, 2013.
2. Nakamura M, Haarmann-Stemmann T, Krutmann J, Morita A (2018). Alternative test models for skin ageing research. *Exp Dermatol.*2018;27(5):495-500. doi: 10.1111/exd.13519. Epub 2018 Apr 10. PMID: 29478289.
3. Madden JC, Enoch SJ, Paini A, Cronin MTD (2020). A Review of In Silico Tools as Alternatives to Animal Testing: Principles, Resources and Applications. *Altern. Lab. Anim.* 2020; 48(4):146-172. doi:10.1177/0261192920965977.
4. Raunio H. (2011). In silico toxicology–non-testing methods. *Front. Pharmacol* 2011; 2:1-8.
5. Hemmerich J and Ecker GF. (2020). In silico toxicology: From structure–activity relationships towards deep learning and adverse outcome pathways. <https://doi.org/10.1002/WCMS.1475>.
6. Myatt GJ, Ahlberg E, Akahori Y, et al. (2018). In silico toxicology protocols. *Regul Toxicol Pharmacol.*2018;96:1-17. <https://doi.org/10.1016/j.yrtph.2018.04.014>.
7. Balekundri U, Sajjan SS, Madagi SB (2015). Two-dimensional quantitative structure activity relationship models for 5alpha-reductase type 2 inhibitors. *J. Pharm. Investig.*2015;45,:293–299. <https://doi.org/10.1007/s40005-015-0173-4>.
8. Wang MWH, Goodman JM, Allen THE (2021). Machine Learning in Predictive Toxicology: Recent Applications and Future Directions for Classification Models. *Chem. Res. Toxicol.* 2021;34 (2): 217-239. DOI: 10.1021/acs.chemrestox.0c00316
9. Schneider G and Downs GM (2003). Machine learning methods in QSAR modelling. *QSAR & Comb. Sci* 2003;22(5):485-486.
10. Sarker IH. (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. *Sn Compt. Sci.*2021; 2:160. <https://doi.org/10.1007/s42979-021-00592-x>.
11. Hernández DCT (2015) "An Experimental Study of K* Algorithm". *Int. J. Inf. Eng. Electron. Bus.* 2015;7(2):14-19. DOI: 10.5815/ijieeb.2015.02.03.
12. Dada et al. (2019) Non-Nested Generalisation (NNge) Algorithm for Efficient and Early Detection of Diabetes, Information Technology and Intelligent Transportation Systems: *Proceedings of the 3rd ITITS.*2018:314:233 Xi'an, China, Sept. 15-16.

13. Scheman A (2000) Adverse reactions to cosmetic ingredients. *Dermatol Clin* 2000; 4:685-698.
14. Kamra M and Diwan A. (2017). Acne: current perspective. *J. Appl. Pharm. Res.* 2017; 5(3), 01-07. <https://doi.org/10.18231/2348-0335.2017.0001>
15. Fulton JE (1989) Comedogenicity and irritancy of commonly used ingredients in skin care products. *J.Soc.Cosmet. Chem.*1989;40:321-333.
16. Kligman AM and Mills OH (1972). "Acne Cosmetica". *Arch Dermatol.* 1972;106(6):843–850. doi:10.1001/archderm.1972.0162015002901.
17. Strauss JS (1991). Acne and Rosacea. In: *Orkin M, Maibach HI, Dahl MV (eds.) Dermatology (1st ed)*, 1991. By Appleton&Lange. Prentice Hall Publishing. p335.
18. OECD Guideline, 439. (2013). OECD Guidelines for the testing of chemical. In Vitro Skin Irritation: Reconstructed Human Epidermis Test Method. Adopted: 26 July 2013.
19. Macfarlane M, Jones P, Goebel C, Dufour E, Rowland J et al. (2009). A tiered approach to the use of alternatives to animal testing for the safety assessment of cosmetics: Skin irritation. *Regul Toxicol Pharmacol.* 2009; 54:188–1.
20. Golla S, Madihally S, Robinson RL, Gasem K. (2009). Quantitative structure-property relationships modelling of skin irritation. *Toxicol In Vitro* 2009;23: 176-184.
21. Enslein K, Borgstedt HH, Blake BW(1987).Prediction of rabbit skin irritation severity by structure–activity relationships. *In Vitro Toxicol.*1987;1:129–147.
22. Barrat. (1995). Quantitative structure-activity relationships for skin corrosivity of organic acids, bases and phenols. *Toxicol. Lett.*1995;75: 169-176.
23. Hayashi M, Nakamura Y, Higashi K, Kato H, Kishida F, Kaneko H. (1999). A quantitative structure-activity relationship study of the skin irritation potential of phenols. *Toxicol In Vitro.*1999;13(6):915-22. doi: 10.1016/s0887-2333(99)00077-6.
24. Smith JS, Macina OT, Sussman NB, Luster MI, Karol MH. (2000). A robust structure-activity relationship (SAR) model for esters that cause skin irritation in humans. *Toxicol Sci.* 2000;55(1):215-22. DOI: 10.1093/toxsci/55.1.215.
25. Kodithala K, Hopfinger AJ, Thompson ED, Robinson MK. (2002). Prediction of skin irritation from organic chemicals using membrane-interaction QSAR analysis. *Toxicol Sci.* 2002;66(2):336-46. doi: 10.1093/toxsci/66.2.336.
26. Parthasarathi R and Dhawan A. (2018). Chapter 5: In Silico Approaches for Predictive Toxicology. In: Dhawan A et al. (Eds). *In Vitro Toxicol.*2018:91-109.

27. Rim KT. (2020). In silico prediction of toxicity and its applications for chemicals at work. *Toxicol.Environ.HealthSci.*2020;12, 191–202. <https://doi.org/10.1007/s13530-020-00056-4>
28. Raies AB and Bajic VB (2016) In silico toxicology: computational methods for the prediction of chemical toxicity. *WIREs Comput Mol Sci* 2016;6:147–172. doi: 10.1002/wcms.1240.
29. Pradeep P. (2017). QSAR: An overview. Oak Ridge Institute for science and education research participant for computational toxicology. US EPA, 2017.
30. Michielan L and Moro S. (2010). Pharmaceutical perspectives of non-linear QSAR strategies. *J. Chem. Inf. Model.* 50: 961–978. <https://doi.org/10.1021/ci100072z>.
31. Gramatica P. (2013). On the Development and Validation of QSAR Models. In: Reisfeld B., Mayeno A. (eds) *Methods in Molecular Biology. Comput. Toxicol.* 2013;21: 930:519-520. Humana Press, Totowa, NJ. https://doi.org/10.1007/978-1-62703-059-5_21.
32. Gramatica P.(2011). A short history of QSAR Evolution. Research Unit in Environmental Chemistry and Ecotoxicology, DBSF, Insubria University, Varese, Italy, 2011.
33. Jhanwar B, Sharma V, Singla R, Shrivastava B. (2011). QSAR - Hansch Analysis and Related Approaches in Drug Design. *Pharmacologyonline* 2011;1:306-344.
34. OECD (2007) Guidance document on the validation of (quantitative) structure-activity relationship [(Q)SAR] models. In: OECD series on testing and assessment. OECD Publishing, Paris, pp 1–154. <https://doi.org/10.1787/20777876>.
35. Wallace AD. (2012). Chapter Four - Toxic Endpoints in the Study of Human Exposure to Environmental Chemicals. Ed: Ernest Hodgson, *Prog Mol Biol Transl Sci.*2012: 112; 89-115. <https://doi.org/10.1016/B978-0-12-415813-9.00004-0>
36. Westmoreland C and Holmes AM. (2009). Assuring consumer safety without animals: Applications for tissue engineering. *Organogenesis* 2009;5(2):67-72. doi: 10.4161/org.5.2.9128.
37. Grisoni F, Consonni V, Todeschini R. (2018) Impact of Molecular Descriptors on Computational Models. In: Brown J.(eds). *Methods in Molecular Biology. Computational Chemogenomics.*2018;1825.Humana Press, NY. https://doi.org/10.1007/978-1-4939-8639-2_5
38. Khan AU. (2016). Descriptors and their selection methods in QSAR analysis: paradigm for drug design. *Drug Discov Today.*2016;21(8):1291-1302. doi: 10.1016/j.drudis.2016.06.013.

39. Kar S and Leszczynski J. (2019). Exploration of Computational Approaches to Predict the Toxicity of Chemical Mixtures. *Toxics* 2019; 7(1):15. <https://doi.org/10.3390/toxics7010015>.
40. Mauri A. (2020) AlvaDesc: A Tool to Calculate and Analyze Molecular Descriptors and Fingerprints. In: Roy K. (eds) *Ecotoxicological QSARs. Methods in Pharmacology and Toxicology*. New York. https://doi.org/10.1007/978-1-0716-0150-1_32.
41. Panneerselvam, Suresh (2016): QSAR protocol. <https://dx.doi.org/10.6084/m9.figshare.3113542.v1>.
42. Dearden JC. "In silico prediction of drug toxicity". *J. Comput. Aided Mol. Des.* 2003; 17: (2–4): 119–27.
43. Wold S and Eriksson L. "Statistical validation of QSAR results". In Waterbeemd, Han van de. *Chemometric methods in molecular design*. Weinheim: VCH. pp. 309, 1995
44. Tropsha A, Gramatica P, Gombar VJ. (2003). "The Importance of Being Earnest: Validation is the Absolute Essential for Successful Application and Interpretation of QSPR Models". *QSAR&Comb. Sci.* 2003; 22: 69–77.
45. Gramatica P. (2007). "Principles of QSAR models validation: internal and external". *QSAR&Comb. Sci.* 2007; 26: 694–701.
46. Chirico N, Gramatica P. (2012). "Real external predictivity of QSAR models. Part 2. New intercomparable thresholds for different validation criteria and the need for scatter plot inspection". *J. Chem Inf Model.* 2012; 52: (8): 2044–58.
47. Roy K. "On some aspects of validation of predictive quantitative structure-activity relationship models". *Expert Opin. Drug Dis.* 2007; 2 (12): 1567–77.
48. Roy K, Kar S, Das RN. (2015). A Primer on QSAR/QSPR modelling: Fundamentals and Concepts. *SpringerBriefs in Molecular Science* 2015; Springer, Cham., <https://doi.org/10.1007/978-3-319-17281-1>.
49. Leist M, Ghallab A, Graepel R et al. Adverse outcome pathways: opportunities, limitations and open questions. *Arch Toxicol.* 2017; 91: 3477–3505
50. Hall L. (2013). Speaker calls for crowdsourcing to advance predictive toxicology. at NIEHS. 2013.
51. Vinken M. The adverse outcome pathway concept: a pragmatic tool in toxicology. *Toxicol.* 2013; 312: 158–65. doi: 10.1016/j.tox.2013.08.011.
52. Langley G, Austin, CP, Balapure AK et al. Lessons from Toxicology: Developing a 21st-Century Paradigm for Medical Research. *Environ. Health Perspect.* 2015; 123: 11

53. Priya V. and Prabu S. (2022). Application of AI in Healthcare. In: Venkatesh C, Rengarajan N., Ponmurugan P, Balamurugan S. (Eds.). Smart Systems for Industrial Applications. John Wiley & Sons. January 2022.
54. Choi RY, Coyner AS, Kalpathy-Cramer J, Chiang MF et al. (2020). Introduction to Machine Learning, Neural Networks, and Deep Learning. *Transl Vis. Sci. Technol.* 2020;9(2):14. doi:10.1167/tvst.9.2.14.
55. Yu C and Helwig EJ. (2021). The role of AI technology in prediction, diagnosis and treatment of colorectal cancer. *Artif.Intell.Rev.*2021;55:323-343. <https://doi.org/10.1007/s10462-021-10034-y>.
56. Mao J, Akhtar J, Zhang X, Sun L et al. (2021). Comprehensive strategies of machine-learning-based quantitative structure-activity relationship models. *iScience* 2021;24(9):103052. <https://doi.org/10.1016/j.isci.2021.103052>.
57. Engelbrecht AP. (2007). *Computational Intelligence: An Introduction*, 2nd Ed. A.P. Engelbrecht. John Wiley & Sons, Ltd, 2007.
58. Pesapane F, Codari M, Sardanelli F. (2018). Artificial intelligence in medical imaging: threat or opportunity? Radiologists again at the forefront of innovation in medicine. *Eur Radiol Exp* 2018;2:35. <https://doi.org/10.1186/s41747-018-0061-6>
59. Bharadwaj G. (2019) History of Machine Learning - How did it Evolve? <https://www.zeolearn.com/magazine/what-is-machine-learning>(accessed June 2022).
60. Rokach L. (2010). Ensemble-based classifiers. *Artif Intell Rev* 2010;33:1–39. <https://doi-org.lproxy.yeditepe.edu.tr/10.1007/s10462-009-9124-7>.
61. Kunapuli G. Ensemble Methods for Machine Learning. MEAP. Manning publication, 2021.
62. Li Y and Chen WA. Comparative Performance Assessment of Ensemble Learning for Credit Scoring. *Mathematics*. 2020;8(10):1756. <https://doi.org/10.3390/math8101756>
63. Chanamarn N, Tamee K, Sittidech P. (2016). Stacking Technique for Academic Achievement Prediction. Int. Workshop on Smart Info-Media Systems in Asia (SISA 2016), Sep. 14–17, 2016.
64. Kumar A. (2020) How to Know if Data is Linear or Non-linear. <https://vitalflux.com/how-know-data-linear-non-linear/>
65. Paul M, 2017. Non-linear classification. Info-4604. Applied Machine Learning. University of Colorado. Lecture note, 2017.

66. Basha SM, Rajput DS (2019) Chapter 9 - Survey on Evaluating the Performance of Machine Learning Algorithms: Past Contributions and Future Roadmap, Ed: Sangaiah AK, Deep Learning and Parallel Computing Environment for Bioengineering Systems, Academic Press, 2019. p153-164. <https://doi.org/10.1016/B978-0-12-816718-2.00016-6>.
67. Saito T and Rehmsmeier M (2015) The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLoS ONE* 2015;10(3): e0118432. <https://doi.org/10.1371/journal.pone.0118432>
68. Abad-Grau MM, Tajtakova M, Aranda DA (2009). Machine learning methods for the market segmentation of the performing arts audiences. *Int. J. Business Environment*, 2009;2(3).
69. Sclangenhaut T, 2022, <https://python-course.eu/machine-learning/decision-trees-in-python.php> (accessed April 1, 2022).
70. Samek W, Montavon G, Lapuschkin S, Anders CJ, Müller KR.(2021) "Explaining Deep Neural Networks and Beyond: A Review of Methods and Applications," in *Proceedings. IEEE* 2021;109(3):247-278. doi: 10.1109/JPROC.2021.3060483.
71. Adhatrao K, Gaykar A, Dhawan A, Jha R, Honrao V. (2013). Predicting Students' Performance Using ID3 And C4.5 Classification Algorithms. *Int. j. Data Min. Knowl. manag. Process* 2013;3(5). DOI: 10.5121/ijdkp.2013.3504.
72. Prakash S. 2020, <https://www.snippetnuggets.com/interviewQuestions/machineLearning/basics/2020-machine-learning-entropy-information-gain-decision-tree.html>. (accessed April 1, 2022).
73. Zhang H, Mao J, Qi HZ, Ding L (2020) In silico prediction of drug-induced developmental toxicity by using machine learning approaches. *Mol Divers*. 2020;24, 1281–1290. doi: 10.1007/s11030-019-09991-y.
74. Chicco D, Jurman G (2020) The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*.2020;21:6. <https://doi.org/10.1186/s12864-019-6413-7>.
75. Wilm A, Stork C, Bauer C, Schepky A, Kühnl J, Kirchmair J (2019). Skin Doctor: Machine Learning Models for Skin Sensitization Prediction that Provide Estimates and Indicators of Prediction Reliability. *Int. J. Mol. Sci.*2019; 20(19):4833. <https://doi.org/10.3390/ijms20194833>.
76. Gupta VK, Rana PS. (2019). Toxicity prediction of small drug molecules of aryl hydrocarbon receptor using a proposed ensemble model. *Turk J Elec Eng&Comp Sci*. 2019;27:2833-2849. doi:10.3906/elk-1809-9.
77. Viera AJ and Garrett JM (2005). Understanding interobserver agreement: the kappa statistic. *Family medicine*. 2005;37(5):360-363.

78. Plewig G, Fulton JE, Kligman AM (1970). Pomade Acne. *Arch Dermatol.* 1970;101(5):580–584. doi:10.1001/archderm.1970.04000050084011.
79. Gollnick H, Cunliffe W an et al. (2003). Management of Acne: A Report From a Global Alliance to Improve Outcomes in Acne. *J.Am. Acad. Dermatol.*2003;49:1-37. <https://doi.org/10.1067/mjd.2003.618>.
80. Mills OH, Criscito MC, Schlesinger TD et al. (2016). Addressing Free Radical Oxidation in Acne Vulgaris.” *J Clin Aesthet Dermatol* 2016;9(1):25-30.
81. Witten IE, Frank E, Hall MA (2016). The WEKA Workbench. Online Appendix for “Data mining: practical machine learning tools and techniques with Java implementations”, 2016. <https://doi.org/10.1145/507338.507355>
82. Baek J and Lee MG. (2016) Oxidative stress and antioxidant strategies in dermatology, *RedoxReport*;2016;21(4):164-169. <https://doi.org/10.1179/1351000215Y.0000000015>.
83. Wagener FADTG, Carels CE, Lundvig DMS (2013). Targeting the Redox Balance in Inflammatory Skin Conditions. *Int. J. Mol. Sci.* 2013;14:9126-9167. <https://doi.org/10.3390/ijms14059126>.
84. Pizzino G, Irrera N, Cucinotta M, et al. (2017). Oxidative Stress: Harms and Benefits for Human Health. *Oxid. Med Cell Longev.*2017;8416763. doi:10.1155/2017/8416763.
85. Domej W, Oetl K, Renner W(2014). Oxidative stress and free radicals in COPD-implications and relevance for treatment. *Int J Chron Obstruct Pulmon Dis.* 2014; 17:9:1207-24. doi: 10.2147/COPD.S51226.
86. Chelombitko MA. (2018). Role of Reactive Oxygen Species in Inflammation: A Minireview. *Moscow Univ. Biol.Sci.Bull.*2018;73:199–202. <https://doi.org/10.3103/S009639251804003X>.
87. Phaniendra A, Jestadi DB, Periyasamy L. (2015). Free radicals: properties, sources, targets, and their implication in various diseases. *Indian J Clin Biochem.* 2015;30(1):11-26. doi: 10.1007/s12291-014-0446-0.
88. Tochio T, Tanaka H, Nakata S, Ikeno H. (2009). Accumulation of lipid peroxide in the content of comedones may be involved in the progression of comedogenesis and inflammatory changes in comedones. *J. Cosmet. Dermatol.*2009;82:152-158. <https://doi.org/10.1111/j.1473-2165.2009.00437.x>
89. Pappas A, Johnsen S. et al. (2009).Sebum analysis of individuals with and without acne. *Dermatoendocrinol.* 2009;1(3):157-161. doi:10.4161/derm.1.3.8473.
90. Ayala A, Munoz MF, Arguelles S. (2014). Lipid Peroxidation: Production, Metabolism, and Signaling Mechanisms of Malondialdehyde and 4-Hydroxy-2-Nonenal. *Oxid. Med. Cell. Longev.* 2014; 31. <http://dx.doi.org/10.1155/2014/360438>

91. Catala A and Diaz M (2017). Ed. Lipid peroxidation on Cell Membranes. In published: *Frontiers in Physiology. Impact of lipid Peroxidation on the physiology and Pathophysiology of cell membranes*, 2017. <https://www.frontiersin.org/research-topics/1269/impact-of-lipid-peroxidation-on-the-physiology-and-pathophysiology-of-cell-membranes>.
92. Aydın A, Sayal, A., Işimer, A. Serbest Radikaller ve Antioksidan Savunma Sistemi. Ankara: Gülhane Askeri Tıp Akademisi Basımevi. 2001.
93. El-Aal M and Hassan HA (2012). In (Ed.), *Lipid Peroxidation. Lipid peroxidation end-products as a key of oxidative stress: Effect of Antioxidant on Their Production and Transfer of Free Radicals*, IntechOpen.2012. <https://doi.org/10.5772/45944>
94. Gianazza E, Brioschi M, Fernandez AM, et al. (2021). Lipid Peroxidation in Atherosclerotic Cardiovascular Diseases. *Antioxid Redox Signal*.2021;49-98.<http://doi.org/10.1089/ars.2019.7955>.
95. Halliwell B, Gutteridge JMC. *Free Radicals in Biology and Medicine*. Fifth Edi. New York: Oxford University Press Inc. 2015.
96. Girotti AW. (1998). Lipid hydroperoxide generation, turnover, and effector action in biological systems. *J Lipid Res*.1998;39(8):1529-42.
97. Finkel T. Signal transduction by reactive oxygen species. *J Cell Biol*.2011;194(1):7-15. doi: 10.1083/jcb.201102095.
98. Pappas A. Epidermal surface lipids. *Dermatoendocrinol*. 2009;1(2):72-6. doi: 10.4161/derm.1.2.7811.
99. Ottaviani M, Camera E, Picardo M. (2010). Lipid mediators in acne. *Mediators Inflamm*. 2010;858176. doi: 10.1155/2010/858176.
100. Bhat YJ, Latief I, Hassan I. (2017). Update on etiopathogenesis and treatment of Acne. *Indian J Dermatol Venereol Leprol*.2017;83:298-306. DOI: 10.4103/0378-6323.199581.
101. Zouboulis CC, Jourdan E, Picardo M.(2014). Acne is an inflammatory disease and alterations of sebum composition initiate acne lesions. *J Eur Acad Dermatol Venereol*. 2014;28:527–532.
102. Furue M, Tsuji G. (2019). Chloracne and Hyperpigmentation Caused by Exposure to Hazardous Aryl Hydrocarbon Receptor Ligands. *Int. J. Environ. Res. Public Health* 2019;16:4864. <https://doi.org/10.3390/ijerph16234864>
103. Schlessinger DI, Robinson CA, Schlessinger J. (2021). Chloracne. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing;2022: <https://www.ncbi.nlm.nih.gov/books/NBK459189/>.
104. Schecter A, Birnbaum L, Ryan JJ, Constable JD. (2006). Dioxins: an overview. *Environ Res*. 2006;101(3):419-28. doi: 10.1016/j.envres.2005.12.003.

105. Furue M, Fuyuno Y, Mitoma C, Uchi H, Tsuji G. (2018). Therapeutic Agents with AHR Inhibiting and NRF2 Activating Activity for Managing Chloracne. *Antioxidants* 2018;7:90. <https://doi.org/10.3390/antiox7070090>.
106. Furue M, Ishii Y, Tsukimori K, Tsuji G. (2021). Aryl Hydrocarbon Receptor and Dioxin-Related Health Hazards—Lessons from Yusho. *Int. J. Mol. Sci.* 2021; 22: 708. <https://doi.org/10.3390/ijms22020708>.
107. Nebert DW, Dalton TP, Okey AB, Gonzalez FJ. (2004). Role of aryl hydrocarbon receptor-mediated induction of the CYP1 enzymes in environmental toxicity and cancer. *J Biol Chem.* 2004;279(23):23847-50. doi: 10.1074/jbc.R400004200.
108. Welss T, Basketter DA, Schröder KR. (2004). In vitro skin irritation: facts and future. State of the art review of mechanisms and models. *Toxicol In Vitro.*2004;18(3):231-43. doi: 10.1016/j.tiv.2003.09.009
109. PubChem Database; <https://pubchem.ncbi.nlm.nih.gov/> (accessed June 10, 2021).
110. ChemSpider, <http://www.chemspider.com>, (accessed June 10, 2021).
111. Mol-Instincts Chemical Property Databases; <https://www.molinstincts.com>, (accessed June 10, 2021).
112. ChemicalBook, <https://www.chemicalbook.com>, (accessed June 10, 2021).
113. MarvinSketch v.21.2.0, ChemAxon, (<https://www.chemaxon.com>), (accessed June 10, 2021).
114. CACTUS, <https://cactus.nci.nih.gov/translate>, (accessed June 10, 2021)
115. Webbook, <https://webbook.nist.gov/chemistry/cas-ser/>, (accessed June 10, 2021).
116. LookChem, <https://www.lookchem.com/> (accessed June 10, 2021).
117. MarvinView v. 21.2.0, ChemAxon, (<https://www.chemaxon.com>), (accessed June 10, 2021).
118. Fourches D, Muratov E, Tropsha A (2010). “Trust, but verify: On the importance of chemical structure curation in cheminformatics and QSAR modelling research” *J Chem Inf Model.* 2010; 50(7):1189-1204. <https://doi.org/10.1021/ci100176x>.
119. ChemAxon (2021). Standardizer. V21.2.0 Budapest, Hungary. Available at: <https://www.chemaxon.com>, (accessed June 10, 2021).
120. O’Boyle et al., (2011). Open Babel: An open chemical toolbox. *Journal of Cheminformatics*, 3:33–46.
121. Hong H, Xie O, Ge W, Qian F et al. (2008). Mold2, Molecular Descriptors from 2D Structures for Chemoinformatics and Toxicoinformatics. *J. Chem. Inf. Model.*2008; 48:1337–1344. DOI: 10.1021/ci800038f. <https://doi.org/10.1021/ci800038f>.

122. Alvascience-Srl (2019). alvaDesc (software for molecular descriptor calculation). v1.0.14, Italy. Available at: <https://www.alvascience.com/>, (Accessed June 10, 2021).
123. T. E. S. T. v.4.2.1, (2016). T. Martin, P. Harten, R. Venkatapathy, D. Young, U. S. EPA/National Risk Management Research Laboratory/ Sustainable Technology Division, Cincinnati, OH, 2016.
124. Yap CW (2011). PaDEL-descriptor: an opensource software to calculate molecular descriptors and fingerprints. *J.Comput.Chem.*2011;32:1466–1474. <https://doi.org/10.1002/jcc.21707>
125. Varsou DD, Melagraki G, Sarimveis H, Afantitis A. (2017) MouseTox: An online toxicity assessment tool for small molecules through Enalos Cloud platform. *Food Chem. Toxicol.* 2017;110, 2017: 83–93. <https://doi.org/10.1016/j.fct.2017.09.058>.
126. Cherkasov A, Muratov EN, Fourches D, Varnek A et al. (2014). QSAR modelling: Where have you been? Where are you going? *J.Med.Chem.*2014;57(12):4977–5010. <https://doi.org/10.1021/jm4004285>.
127. Eriksson L, Jaworska J, Worth AP, Cronin MT, et al (2003) Methods for reliability and uncertainty assessment and for applicability evaluations of classification-and regression-based QSARs. *Environ. Health. Perspect.* 2003; 111(10):1361-1375. doi:10.1289/ehp.5758.
128. Tugcu G, Sipahi H. (2019).QSPR modelling of in vitro degradation half-life of acyl glucuronides.*Xenobiotica*2019; 49:9, 1007-1014. DOI: 10.1080/00498254.2018.1527049
129. IBM Corp. Released (2017). IBM SPSS Statistics for Windows, Version 25.0. Armonk, NY: IBM Corp, 2017.
130. Umargono E, Suseno JE et al. (2019). K-Means Clustering Optimization using the Elbow Method and Early Centroid Determination Based-on Mean and Median. *ISSTEC 2019: 234-240* :In Proceedings of the Int Conf. on Information System and Technology. DOI: 10.5220/0009908402340240.
131. Santos TL and Zarate LE (2015). Categorical data clustering: What similarity measure to recommend? T.R.L. *Exp. Systems with App.*2015;42:1247–1260. <http://dx.doi.org/10.1016/j.eswa.2014.09.012>.
132. Owen S, Anil R, Dunning T, Friedman (2011). Mahout in Action.; Chap:10: Evaluating and improving clustering quality. ISBN 9781935182689 . <https://livebook.manning.com/book/mahout-in-action/chapter-10/44>
133. Hall, M.A. (1999) Correlation-Based Feature Selection for Machine Learning. PhD Thesis, University of Waikato, Hamilton, New Zealand.
134. Witten E. Frank (2000). Data Mining, Practical Machine Learning Tools and Techniques with Java Implementation, Morgan Kaufmann Publishers, 2000.
135. Naqvi G. (2012). A Hybrid Filter-Wrapper Approach for Feature Selection. (Master's Thesis). The department of Technology, Örebro University, Sweden, 2012.

136. Bouaguel W. (2015). On Feature Selection Methods for Credit Scoring (Doctoral dissertation), Université de Tunis, Institut Supérieur de Gestion, Ecole Doctorale Sciences de Gestion Larodec, Tunus, Retrieved from https://www.researchgate.net/publication/271513838_On_Feature_Selection_for_Credit_Scoring.
137. Quinlan R. (1993). C4.5: Programs for Machine Learning. *Machine Learning* 1994; 16:235–240. Morgan Kaufmann Publishers, San Mateo, CA. <https://doi.org/10.1007/BF00993309>.
138. Kohavi R and John GH. (1997). Wrappers for feature subset selection. *Artif. Intell.*1997;(1-2):273-324. [https://doi.org/10.1016/S0004-3702\(97\)00043-X](https://doi.org/10.1016/S0004-3702(97)00043-X).
139. Vijayarani M. (2013) Comparative Analysis of Bayes and Lazy Classification Algorithms. *IJARCCCE* 2013;2:8. ISSN (Online): 2278-1021.
140. Cleary JG, Trigg LE (1995). K*:An Instance-based Learner Using an Entropic Distance Measure. *In:12th Int. Conf. on Mach. Learn.*1995:108-114, <https://doi.org/10.1016/B978-1-55860-377-6.50022-0>
141. Grigorev A. *Machine Learning Bookcamp: Decision trees and ensemble learning*. Manning publication. 2021.
142. Breiman L. (2001).Random Forests. *J.Mach.Learn.*2001;45:5–32 <https://doi.org/10.1023/A:1010933404324>.
143. Ali J, Khan R, Nasir A, Imran M (2012). Random forests and decision trees. *Int J Comput Sci.*2012; 9: 272–278.
144. Martin B. (1995). Instance-Based learning: Nearest Neighbor with Generalization. Hamilton, New Zealand.
145. Sylvain R. (2002) "Nearest Neighbor with Generalization", Unpublished, University of Canterbury, Christchurch, New Zealand, 2002.
146. Brink H. (2016). Real-World Machine Learning. Chapter 4. Model evaluation and optimization, Manning Pub., <https://livebook.manning.com/book/real-world-machine-learning/chapter-4>
147. Ashfaq JM, (2019). Introduction to Support Vector Machines and Kernel Methods. Course note.
148. Moorthy NSHN, Kumar S, Poongavanam SV (2017). Classification of carcinogenic and mutagenic properties using machine learning method. *Compt.Toxicol.* 2017;3: 33-43. <https://doi.org/10.1016/j.comtox.2017.07.002>.
149. Khan K, Benfenati E, Roy K. (2019). Consensus QSAR modeling of toxicity of pharmaceuticals to different aquatic organisms: Ranking and prioritization of the DrugBank database compounds. *Ecotoxicol Environ Saf.*2019;168:287-297. doi: 10.1016/j.ecoenv.2018.10.060.

150. Mathea M, Klingspohn W, Baumann K (2016). Chemoinformatic Classification Methods and their Applicability Domain. *Mol.Inf.*2016;35:160-180. <https://doi.org/10.1002/minf.201501019>.
151. Shoaib T. (2016.). SPSS-Training Data Screening Identification of Potential Outliers. DOI: 10.13140/RG.2.2.12693.37609.
152. Olasupo SB, Uzairu A, Shallangwa G, Uba S. (2020). Quantitative Structure-Activity Relationship (QSAR) Studies and Molecular docking Simulation of Norepinephrine Transporter (NET) Inhibitors as Anti-psychotic Therapeutic Agents. *JOTCSA*. 2020;7(1):179–96. DOI: <https://doi.org/10.18596/jotcsa.577259>.
153. McHugh M. (2013). The Chi-square test of independence. *Biochem Med* 2013; 23(2): 143–149. doi: 10.11613/BM.2013.018.
154. Lee DK, (2016). Alternatives to P-value: confidence interval and effect size. *Korean J Anesthesiol*. 2016;69(6):555-562.
155. Leonard JT and Roy K. (2006). On Selection of Training and Test Sets for the Development of Predictive QSAR models. *QSAR Comb Sci* 2006;25;3:235-251 DOI: 10.1002/QSAR.200510161
156. Omran M. Andries P. Engelbrecht, Salman A. (2007). An overview of clustering methods. *Intelligent Data Analysis*, 11;6: 583-605. DOI: 10.3233/IDA-2007-11602.
157. Ezenkwu CP, Ozuomba S, Kalu C. (2015). Application of K-Means Algorithm for Efficient Customer Segmentation: A Strategy for Targeted Customer Services. *Int. J. Adv.Research in Artif. Intell.*2015;4:10.
158. Saliner AG, Patlewicz G, Worth AP. (2005). A Similarity Based Approach for Chemical Category Classification. European Commission (EC). Italy. EUR 21867 EN. <http://europa.eu.int/>
159. O’Boyle NM, Sayle RA (2016). Comparing structural fingerprints using a literature-based similarity benchmark. *JCheminform*2016; 8:36. <https://doi.org/10.1186/s13321-016-0148-0>.
160. Soler J, Tence F, Gaubert et al (2013). Data Clustering and Similarity. Association for the *Proceedings of the 26th Int. Florida Artif. Intell.Research Society Conference*. Advancement of Artificial Intelligence (www.aaai.org), 2013.
161. Huang K, Xiao C, Glass L, Sun J (2019) Explainable substructure partition fingerprint for protein, drug, and more. *NeurIPS Learning Meaningful Representation of Life Workshop*, 2019.
162. Textbook of Boundless Chemistry (2013). Functional group names, properties and reactions. <https://books.apple.com/gb/book/chemistry/id684002835>. (E-book)
163. Robert JD and Caserio MC. (1997) Basic Principles of Organic Chemistry, (2nd Edition) W. A. Benjamin, Inc. Menlo Park, CA. ISBN 0-8053-8329-8. https://chem.libretexts.org/Bookshelves/Organic_Chemistry/Book%3A_Basic_Principles_of_Organic_Chemistry_

164. Demisse GB, Tadesse T, Bayissa (2017). Data mining attribute selection approach for drought modelling: A case study for greater horn of Africa. *Int. J. Data Min. Knowl. Manag.* 2017;7(4):1-16. DOI: 10.5121/ijdkp.2017.7401.
165. Rokach and Maimon (2014). Data Mining with Decision Trees.Theory and Applications. 2nd Edition.2014. <https://doi.org/10.1142/9097>
166. Jantan H (2010). Human Talent Prediction in HRM using C4.5 Classification *Int. J. Comput. Sci. Eng.* 2010; 2(8): 2526-2534.
167. Nie et al. (2015). Application of J48 Decision Tree Classifier in Emotion Recognition Based on Chaos Characteristics. Proceedings of the 2015 Int. Conf.on Automation, Mechanical Control and Comput. Eng. <https://doi.org/10.2991/amcce-15.2015.330>.
168. Schober P, Boer C, Schwarte LA. (2018). Correlation Coefficients: Appropriate Use and Interpretation. *Anesth.Analg.*2018;126(5):1763-1768. doi:10.1213/ANE.0000000000002864.
169. Nedyalkova M and Simeonov V. (2021). Partitioning Pattern of Natural Products Based on Molecular Properties Descriptors Representing Drug-Likeness. *Symmetry* 2021;13:546. <https://doi.org/10.3390/sym13040546>.
170. Schneider G, Neidhart W, Giller T, Schmid G. (1999). Scaffold-Hopping by Topological Pharmacophore Search: A Contribution to Virtual Screening. *Angew. Chem.* 199;111, 3068 -3070.
171. Pradeepkiran JA, Reddy AP, Reddy PH (2019). Pharmacophore-based models for therapeutic drugs against phosphorylated tau in Alzheimer's disease. *Drug Discov Today*2019;4(2): 616–623. doi:10.1016/j.drudis.2018.11.005.
172. Rojas C, Todeschini R, Ballabio D, Mauri A, Consonni V, Tripaldi P,Grisoni F. (2017). A QSTR-Based Expert System to Predict Sweetness of Molecules. *Frontiers in Chemistry*: 2017;5 <https://www.frontiersin.org/article/10.3389/fchem.2017.00053>
173. Kligman AM, Wheatley VR, Mills OH (1970). Comedogenicity of human sebum. *Arch Dermatol.*1970;102(3):267-275. doi:10.1001/archderm.1970.04000090029005.
174. Kohen and Nyska (2002). Oxidation of Biological Systems: Oxidative Stress Phenomena, Antioxidants, Redox Reactions, and Methods for Their Quantification. *Toxicol. Pathol*; 2002;30(6): 620–650. DOI: 10.1080/0192623029016672 4.
175. Hanson DM, Harvey E, Sweeney R, Zielinski TJ (2021). Ionization Potentials. Retrieved from <https://chem.libretexts.org/@go/page/64789>.<https://doi.org/10.1016/B978-0-444-62645-5.00001-8>.
176. Buglak AA, Charisiadis A, Sheehan A et al. (2021). Quantitative Structure-Property Relationship Modelling for the Prediction of Singlet Oxygen Generation by Heavy-atom-free BODIPY Photosensitizers. *Chem. Eur. J.*2021;27:38. <https://doi.org/10.1002/chem.202100922>

177. Zhang X, Wu L, Zhen W et al. (2021) Generation of singlet oxygen via iron-dependent lipid peroxidation and its role in Ferroptosis. *Fundam. Res.*2021;2(1):66-73. <https://doi.org/10.1016/j.fmre.2021.07.008>
178. Baumgartner MT, Taccaone R, Teruel M, Lane SI. (2002). Theoretical study of the relative reactivity of chloroethenes with atmospheric oxidants (OH, NO₃, O(3P), Cl(2P) and Br(2P)). *Phys. Chem. Chem. Phys.*2002;4:1028-1032. DOI: 10.1039/b108856b.
179. Basak SC, Niemi GJ, Regal RR (1987). Topological indices: their nature, mutual relatedness, and applications. *Math.Model.*1987;(8):300-305. [https://doi.org/10.1016/0270-0255\(87\)90594-X](https://doi.org/10.1016/0270-0255(87)90594-X).
180. Wilson JA, Bender A, Kaya T, Clemons PA (2009). *J. Chem. Inf. Model.* 2009; 49(10): 2231-2241. DOI: 10.1021/ci900190z
181. Gupta RA, Kaskhedikar SG. (2013). Synthesis, antitubercular activity, and QSAR analysis of substituted nitroaryl analogs: chalcone, pyrazole, isoxazole, and pyrimidines. *Med Chem Res* 2013;22:3863–3880. <https://doi.org/10.1007/s00044-012-0385-3>.
182. Freedberg IM, Eisen AZ, Wolff K, Frank K, (2003). Fitzpatrick's Dermatology in General Medicine. (6th ed.). p 684. McGraw-Hill, 2003.
183. Bolognia JL, Jorizzo JL, Bolognia JL (2007) Dermatology: 2-Volume Set. St. Louis: Mosby.
184. Matthews J and Ahmed S (2013). Adverse human health effects of dioxin-like compounds. In: Chap.1-AHR- and ER-Mediated Toxicology and Chemoprevention. [Adv.Mol.Toxicol.2013;7:1-38](https://doi.org/10.1016/B978-0-444-62645-5.00001-8).<https://doi.org/10.1016/B978-0-444-62645-5.00001-8>.
185. Wagener M, Sadowski J, Gasteiger J (1995). Autocorrelation of Molecular Surface Properties for Modeling Corticosteroid Binding Globulin and Cytosolic Ah Receptor Activity by Neural Networks. *J. Am. Chem. Soc.* 1995; 117: 29: 7769-7775. <https://doi.org/10.1021/ja00134a023>
186. Casanola GM, Ponce YM, Hasan MT. (2006). The Dragon Method in the Computational Identification of Novel Tyrosinase Inhibitors. Results Supported by Experimental Assays. Conference: Proceedings of ECSOC-9, The 9th Int. Electronic Conf. on Synthetic Organic Chemistry, 2006.
187. Farrokhnia M. (2020). Density Functional Theory Studies on the Antioxidant Mechanism and Electronic Properties of Some Bioactive Marine Meroterpenoids: Sargahydroquionic Acid and Sargachromanol. *ACS Omega* 2020;5(32):20382-20390. DOI: 10.1021/acsomega.0c02354.
188. Nolte and Ragas (2017). A review of quantitative structure–property relationships for the fate of ionizable organic chemicals in water matrices and identification of knowledge gaps. *Environ.Sci.*2017;19:221-246. <https://doi.org/10.1039/C7EM00034K>

189. Gregorutti B, Michel B, Saint-Pierre P (2017). Correlation and variable importance in random forests. *Stat Comput* 2017;27:659–678. <https://doi.org/10.1007/s11222-016-9646-1>.
190. Hastie T, Tibshirani R, Friedman J (2017). The Elements of Statistical Learning; Data Mining, Inference, and Prediction. 2nd edition, Springer Series in Statistics., 2017, pp 222-364.
191. Mayank Tripathi, 13 June 2020, Underfitting and Overfitting in Machine Learning (datascience.foundation), <https://datascience.foundation/sciencewhitepaper/underfitting-and-overfitting-in-machine-learning>, (accessed June 10, 2021).
192. Wu X, Kumar V, Quinlan JR et al. (2008). Top 10 algorithms in data mining. *Knowl Inf Syst* 2008;14:1–37. <https://doi.org/10.1007/s10115-007-0114-2>
193. Luna JJ, Grisoni F, Schneider G (2020). Drug discovery with explainable artificial intelligence. *Nat.Mach.Intell* 2020;2, 573–584. <https://doi.org/10.1038/s42256-020-00236-4>
194. Li HB, Wang W, Ding HW, Dong J (2010). "Trees Weighting Random Forest Method for Classifying High-Dimensional Noisy Data". *IEEE* 2010;160-163. doi: 10.1109/ICEBE.2010.99.
195. Soner Yıldırım, Dec 15, 2020. F-score: 10 Must-Know Classification Metrics for Machine Learning | by Soner Yıldırım | Towards Data Science, Writing about Data Science, AI, ML, DL, Python, SQL, Stats, Math | [linkedin.com/in/soneryildirim/](https://towardsdatascience.com/10-must-know-classification-metrics-for-machine-learning-2ce3a4ad256e), <https://towardsdatascience.com/10-must-know-classification-metrics-for-machine-learning-2ce3a4ad256e>, (accessed June 10, 2021).
196. Güneş SS, Yeşil Ç, Gurdal EE, Korkmaz EE, Yarım M, Aydın A, Sipahi H. (2021). Primum non nocere: In silico prediction of adverse drug reactions of antidepressant drugs. *Comput. Toxicol.* 2021;18;100165. <https://doi.org/10.1016/j.comtox.2021.100165>.
197. Kühne R, Ebert RU, Schüürmann G. (2009). Chemical domain of QSAR models from atom-centered fragments. *J Chem Inf Model.*2009;49(12):2660-9. doi: 10.1021/ci900313u.
198. Ramachandran KM and Tsokos CP (2015). 1.5.2 Box Plots. In: Descriptive statistics. in Mathematical Statistics with Applications in R (2nd Ed.) <https://doi.org/10.1016/B978-0-12-417113-8.00001-1>
199. Saliner AG, Patlewicz G, Worth, AP (2008). A review of QSAR models for skin and eye irritation and corrosion. *QSAR&Comb.Sci.*2008;27:49–59. <https://doi.org/10.1002/qsar.200710103>.
200. Lee CK, Uchida T, Kitagawa K et al. (1994). Skin Permeability of Various Drugs with Different Lipophilicity. *J. Pharm. Sci.*1994;83:4;562-565. <https://doi.org/10.1002/jps.2600830424>.

201. Seweryn A. (2018). Interactions between surfactants and the skin – Theory and practice. *Adv. Colloid Interf. Sci.* 2018;256:242-255. <https://doi.org/10.1016/j.cis.2018.04.002>.
202. Rorije E and Hulzebos E (2005). Skin irritation QSAR technical report: Evaluation of (Q)SARs for the prediction of Skin Irritation/Corrosion Potential. National Institute of Public Health and Env. Expert Centre for Substances, Netherlands, 2005.
203. Worth AP, Bassan A, Gallegos A et al. (2015). The Characterisation of (Quantitative) Structure-Activity Relationships: Preliminary Guidance. Institute for Health and Consumer Protect.Toxicol. and Chem. Substances Unit, EU Chemicals Bureau, Italy, 2015.
204. www.rsc.org. Electronic Supplementary Material (ESI) for Molecular BioSystems. © The Royal Society of Chemistry 2016. <http://www.rsc.org/suppdata/c6/mb/c6mb00457a/c6mb00457a1.pdf>
205. Sliwoski G, Mendenhall J, Meiler J. Autocorrelation descriptor improvements for QSAR: 2DA_Sign and 3DA_Sign. *J Comput Aided Mol Des* 2016;30, 209–217.
206. Zhan C, Nichols JA, Dixon DA. (2003). Ionization Potential, Electron Affinity, Electronegativity, Hardness, and Electron Excitation Energy: Molecular Properties from Density Functional Theory Orbital Energies. *J. Phys. Chem. A* 2003; 107: 4184-4195. DOI: 10.1021/jp0225774.
207. Textbook of Boundless Chemistry (2013). Nonmetallic Elements, Halogens. <https://books.apple.com/gb/book/chemistry/id684002835>. (E-book)
208. Eskes C. (2016). Application of alternative methods in the regulatory assessment of chemical safety related to human skin and irritation, SeCAM Services & Consultation on Alternative Methods, Switzerland
209. Sato T, Asai M, Borschevsky A et al. (2015). Measurement of the first ionization potential of lawrencium, element 103. *Nature* 2015;520, 209–211. <https://doi.org/10.1038/nature14342>.
210. Walker, J.D., Gerner, I., Hulzebos, E. & Schlegel, K.(2004). (Q)SARs for predicting skin irritation and corrosion: mechanisms, transparency and applicability of predictions. *QSAR & Comb. Sci.* 2004; 23, 721–725.
211. Maier LE, Maibach HI, Malley MO. (2010). Chapter 23 - Irritant Dermatitis, Editor(s): Robert Krieger, Hayes' Handbook of Pesticide Toxicology (3rd Ed.), Academic Press, 2010, pp 647-659, <https://doi.org/10.1016/B978-0-12-374367-1.00023-9>.
212. Berner, B., Wilson, D.R., Guy, R.H. et al. (1988). The Relationship of pK a and Acute Skin Irritation in Man. *Pharm.Res.*1988;5;660–663. <https://doi.org/10.1023/A:1015931105660>.
213. ECOTOX (1995). Skin irritation and corrosion, primary skin irritation index. *European Center for Ecotoxicology, Report 66, Belgium*, 1995.

214. Berner B, Wilson DR, Steffens, Mazzenga GC, Hinz R, and et al, (1990). The relationship between pKa and skin irritation for a series of basic penetrants in man. *Fundam Appl Toxicol.*1990;15(4):760-766. [https://doi.org/10.1016/0272-0590\(90\)90192-M](https://doi.org/10.1016/0272-0590(90)90192-M).
215. Whittle E, Barrat EM, Carter JA, Basketter DA, Chamberlain M. Skin corrosivity potential of fatty acids: In vitro rat and human skin testing and QSAR studies. *Toxicol. in Vitro* 1996;10(1): 95-100. doi: 10.1016/0887-2333(95)00096-8.
216. Fluhr JW, Darlenski R, Angelova-Fischer I, Tsankov N, Basketter D. (2008). Skin Irritation and Sensitization: Mechanisms and New Approaches for Risk Assessment. *Skin Pharmacol Physiol.* 2008;21:124-135. doi: 10.1159/000131077.
217. Corsini M, Viviani B, Marinovich M, Corrado (1997). Role of Mitochondria and Calcium Ions in Tributyltin-Induced Gene Regulatory Pathways. *Toxicol. and App. Pharmacol.*1997;145:(1):74-81. <https://doi.org/10.1006/taap.1996.8100>.
218. Shen T, Zhu OX, Yang S, Wu CH, Zhang HF et al. (2008). Trichloroethylene induced cutaneous irritation in BALB/c hairless mice: Histopathological changes and oxidative damage. *Toxicol.*2008;248:(2-3):113-120, <https://doi.org/10.1016/j.tox.2008.03.016>.
219. Shimizu T, Aioi A, Horiguchi T, Kuriyama K. (1995). Effect of vitamin E on keratinocyte-modulation induced by lauroylsarcosine. *Jpn J Pharmacol.* 1995;67(4):291-5. doi: 10.1254/jjp.67.291.
220. Kruk J and Duchnik E (2014). Oxidative Stress and Skin Diseases: Possible Role of Physical Activity. *Asian Pac. J. Cancer Prev.* 2014; 15(2):561-568. <https://doi.org/10.7314/apjcp.2014.15.2.561>
221. Effendy I and Maibach HI. (1995). Surfactants and experimental irritant contact dermatitis. *Contact Dermatitis.*1995;33(4):217-225. DOI: 10.1111/j.1600-0536.1995.tb00470.x.
222. Agner T (ed): (2016). Skin Barrier Function: Barrier Function in Healthy Subjects and Atopic Dermatitis Patients. *Curr Probl Dermatol.* 2016;49:8-26. doi: 10.1159/000441540.
223. Smeden J, Dijkhoff IM, Helder RWJ, et al., (2017). In situ visualization of glucocerebrosidase in human skin tissue: zymography versus activity-based probe labeling. *J Lipid Res.* 2017;58(12):2299-2309. doi: 10.1194/jlr.M079376.
224. Liu Y and Lunter DJ (2020). Selective and sensitive spectral signals on confocal Raman spectroscopy for detection of ex vivo skin lipid properties. *Translational Biophotonics.*2020;2(3). DOI: 10.1002/tbio.202000003.
225. Mizutani T, Mori R, Hirayama M, Sagawa Y, Shimizu K, Okano Y, Masaki H. (2016). Sodium Lauryl Sulfate Stimulates the Generation of Reactive Oxygen Species through Interactions with Cell Membranes. *J Oleo Sci.* 2016;1;65(12):993-1001. doi: 10.5650/jos.ess16074

226. Necrosis definition https://joint-research-centre.ec.europa.eu/eu-reference-laboratory-alternatives-animal-testing/alternative-methods-toxicity-testing/validated-test-methods/skin-corrosion_en
227. Gaschler MM and Stockwell BR. Lipid peroxidation in cell death. (2017). *Biochem Biophys Res Commun.* 2017;482(3):419-425. doi: 10.1016/j.bbrc.2016.10.086.
228. Finosh T and Jayabalan M. (2013) Reactive oxygen species-Control and management using amphiphilic biosynthetic hydrogels for cardiac applications. *Adv Biosci Biotechnol.*2013;4:1134-1146. doi: 10.4236/abb.2013.412150
229. Dalleau S, Baradat M, Guéraud F, Huc L. Cell death and diseases related to oxidative stress: 4-hydroxynonenal (HNE) in the balance. *CellDeath. Differ.*2013;20(12):1615-30. doi: 10.1038/cdd.2013.138.
230. Ottaviani M, Alestas T, Flori E, et al. (2006). Peroxidated squalene induces the production of inflammatory mediators in HaCaT keratinocytes: a possible role in acne vulgaris. *J Invest Dermatol.* 2006;126(11):2430-7. doi: 10.1038/sj.jid.5700434.
231. Motoyoshi K. Enhanced comedo formation in rabbit ear skin by squalene and oleic acid peroxides. *Br. J. Dermatol.*1983;109(2):191-198. doi:10.1111/j.1365-2133.1983.tb07080.x.
232. Chiba K, Yoshizawa K, Makino I, Kawakami K, Onoue M (2000). Comedogenicity of squalene monohydroperoxide in the skin after topical application. *J Toxicol Sci.*2000;25:77–83.
233. Tanaka T and Hayakawa R. (1986). Lipid Peroxides in Cosmetic Products and Their Effect to Irritate the Skin. *J Clin Biochem Nutr.* 1986;1;3: 201-207.
234. Niki E, Yoshida Y, Saito Y, Noguchi N. (2005). Lipid peroxidation: Mechanisms, inhibition, and biological effects. *Biochem. Biophys. Res. Commun.* 2005;338(1): 668-676. <https://doi.org/10.1016/j.bbrc.2005.08.072>.
235. Grisoni F, Merk D, Byrne R et al. (2018). Scaffold-Hopping from Synthetic Drugs by Holistic Molecular Representation. *Sci Rep* 2018;8(1):16469. <https://doi.org/10.1038/s41598-018-34677-0>.
236. Hollas B. (2003). An Analysis of the Autocorrelation Descriptor for Molecules. *J. Math. Chem.*2003;33: 91–101. <https://doi.org/10.1023/A:1023247831238>.
237. Weka, J48-decision tree (2016). Weka: Decision Trees-J48. Machine Learning for LanguageTechnology, http://stp.lingfil.uu.se/~santinim/ml/2016/Lect_03/Lab02_DecisionTrees.pdf, (accessed June 10, 2021)
238. Functional groups | Carbon | Biology (article) | Khan Academy <https://www.khanacademy.org/science/ap-biology/chemistry-of-life/elements-of-life/a/functional-groups#:~:text=Functional%20groups%20are%20chemical%20motifs,molecule%20they%20are%20found%20in.&text=Functional%20groups%20can%20be%20classified,their%20charge%20and%20polarity%20characteristics>

8. APPENDICES

Appendix 8.1_TableS1.Dataset (6 pages)

Appendix 8.2_TableS2(5 pages)A. Comedogenic Activity and their grade

B. Cluster Membership

C. Training and Test Set for Comedogenicity Model

D. Irritation and their grade

Appendix 8.3_TableS3.1Selected Descriptors with original value (4 pages)

Appendix 8.3_TablesS3.2 Selected Descriptors for modelling with normalized value (5 pages)

Appendix 8.4_TableS4.Mpredicted values for comedogenicity (6 pages)

Appendix 8.5_TableS5(5 pages)A. Irritation and their grade

B. Cluster Membership

C. Training and Test Set for Irritation Model

D. Comedogenic Activity and their grade

Appendix 8.6_TableS6.Selected Descriptors with original value (4 pages)

Appendix8.7_TableS7.Mpredicted values for irritation (5 pages)

Appendix 8.1_TableS1.Dataset (6 pages)

ID	Chemical Name	Chemical Class no	MW	Formula	CAS	SMILES
1	1-3 butanediol (Butylene glycol)	2	90,14	C4 H10 O2	107-88-0	<chem>OCCC(O)C</chem>
2	Acetone	8	58,09	C3 H6 O	67-64-1	<chem>O=C(C)C</chem>
3	Allantoin	7	158,14	C4 H6 N4 O3	97-59-6	<chem>O=C(N)NC1NC(=O)NC1=O</chem>
4	Alpha tocopherol acetate	6	472,83	C31 H52 O3	58-95-7	<chem>O=C(OC=1C(=C(C=2OC(C)(CCC2C1C)CCCC(C)CCCC(C)CCCC(C)C)C)C</chem>
5	Alpha tocopherol	6	430,79	C29 H50 O2	59-02-9	<chem>OC=1C(=C(C=2OC(C)(CCC2C1C)CCCC(C)CCCC(C)CCCC(C)C)C</chem>
6	Ascorbyl palmitate	6	414,6	C22 H38 O7	137-66-6	<chem>O=C1OC(C(O)=C1O)C(O)COC(=O)CCCCCCCCCCCCCCC</chem>
7	Behenic acid	1	340,66	C22 H44 O2	112-85-6	<chem>O=C(O)CCCCCCCCCCCCCCCCCCCC</chem>
8	Behenyl erucate	1	647,3	C44 H86 O2	18312-32-8	<chem>O=C(OCCCCCCCCCCCCCCCCCCCC)CCCCCCCCCCCC=CCCCCCCC</chem>
9	Butyl stearate	1	340,66	C22 H44 O2	123-95-5	<chem>O=C(OCCCC)CCCCCCCCCCCCCCCC</chem>
10	Capric acid	1	172,3	C10 H20 O2	334-48-5	<chem>O=C(O)CCCCCCCC</chem>
11	Caprylic acid (Octanoic acid)	1	144,24	C8 H16 O2	124-07-2	<chem>O=C(O)CCCCCCC</chem>
12	Carmine	3	492,42	C22 H20 O13	1260-17-9	<chem>O=C(O)C1=C(O)C=C2C(=O)C3=C(O)C(O)=C(C(O)=C3C(=O)C2=C1C)C4OC(CO)C(O)C(O)C4O</chem>
13	Castor oil	10	933,61	C57 H104 O9	8001-79-4	<chem>O=C(OCC(OC(=O)CCCCCCCC=CCC(O)CCCCC)COC(=O)CCCCCCCC=CCC(O)CCCCC)CCCCCCCC=CCC(O)CCCCC</chem>
14	Cetyl acetate	1	284,54	C18 H36 O2	629-70-9	<chem>O=C(OCCCCCCCCCCCCCCCC)C</chem>
15	Cetyl alcohol	2	242,5	C16 H34 O	29354-98-1	<chem>OCCCCCCCCCCCCCCCC</chem>
16	Cetyl palmitate	1	480,96	C32 H64 O2	540-10-3	<chem>O=C(OCCCCCCCCCCCCCCCC)CCCCCCCCCCCCCCCC</chem>
17	Cholesterol	5	386,73	C27 H46 O	57-88-5	<chem>OC1CC2=CCC3C(CCC4(C)C(CCC34)C(C)CCCC(C)C)C2(C)CC1</chem>
18	Cholet 24	5	430,79	C29 H50 O2	27321-96-6	<chem>OCCOC1CC2=CCC3C(CCC4(C)C(CCC34)C(C)CCCC(C)C)C2(C)CC1</chem>
19	Chondroitin sulfate	7	463,42	C13 H21 N O15 S	9088-44-2	<chem>O=C(O)C1OC(OC2C(O)C(OC(O)C2NC(=O)C)OS(=O)(=O)O)C(O)C(O)C1O</chem>
20	Cyclomethicone	4	296,68	C8 H24 O4 Si4	556-67-2	<chem>O1[Si](O[Si](O[Si](O[Si]1(C)C)(C)C)(C)C</chem>
21	D&C red 21	3	647,88	C20 H8 O5 Br4	15086-94-9	<chem>O=C1OC2(C=3C=CC=CC13)C4=CC(Br)=C(O)C(Br)=C4OC5=C(Br)C(O)=C(Br)C=C52</chem>

22	D&C Red 40 (Allura Red)	3	496,4	C18 H14 N2 O8 S2 Na2	25956-17-6	[Na+].[Na+].O=S(=O)([O-])C1=CC=C2C(C=CC(O)=C2N=NC=3C=C(C(=CC3OC)S(=O)(=O)[O-])C)=C1
23	D&C red17	3	352,42	C22 H16 N4 O	85-86-9	O=C1C=CC=2C=CC=CC2C1=NNC3=CC=C(N=NC=4C=CC=CC4)C=C3
24	D&C red19	3	479,0	C28 H31 N2 O3 Cl	81-88-9	CCN(CC)C1=CC2=C(C=C1)C(=C3C=CC(=[N+](CC)CC)C=C3O2)C4=CC=CC=C4C(=O)O.[Cl-]
25	D&C red27	3	785,64	C20 H4 O5 Br4 Cl4	13473-26-2	O=C1OC2(C3=CC(Br)=C(O)C(Br)=C3OC4=C(Br)C(O)=C(Br)C=C42)C=5C(Cl)=C(Cl)C(Cl)=C(Cl)C15
26	D&C red30	3	393,32	C18 H10 O2 Cl2 S2	2379-74-0	O=C1C(SC2=CC(Cl)=CC(=C12)C)=C3SC4=CC(Cl)=CC(=C4C3=O)C
27	D&C red33	3	467,4	C16 H11 N3 O7 S2Na2	3567-66-6	C1=CC=C(C=C1)N=NC2=C(C3=C(C=C(C3C=C2S(=O)(=O)[O-])S(=O)(=O)[O-])N)O.[Na+].[Na+]
28	D&C red36	3	327,74	C16 H10 N3 O3 Cl	2814-77-9	O=C1C=CC=2C=CC=CC2C1=NNC3=CC=C(C=C3Cl)[N+](=O)[O-]
29	D&C red4	3	436,5	C18 H16 N2 O7 S2	4548-53-2	CC1=CC(=C(C=C1N=NC2=C(C3=CC=CC=C3C(=C2)S(=O)(=O)[O-])O)S(=O)(=O)[O-])C.[Na+].[Na+]
30	D&C red6	3	430,3	C18H12N2Na 2O6S	5858-81-1	CC1=CC(=C(C=C1)N=NC2=C(C(=CC3=CC=CC=C32)C(=O)O)[O-])S(=O)(=O)[O-].[Na+].[Na+]
31	D&C red9	3	888,9	Ba(C17H12Cl N2O4S)2	5160-02-1	CC1=CC(=C(C=C1Cl)S(=O)(=O)[O-])N=NC2=C(C=CC3=CC=CC=C32)O.CC1=CC(=C(C=C1Cl)S(=O)(=O)[O-])N=NC2=C(C=CC3=CC=CC=C32)O.[Ba+2]
32	Decyl oleate	1	422,82	C28 H54 O2	3687-46-5	CCCCCCCCCOC(=O)CCCCCCCC=CCCCCCCC
33	Di (2 ethylhexyl) succinate	1	342,58	C20 H38 O4	2915-57-3	O=C(OCC(CC)CCCC)CCC(=O)OCC(CC)CCCC
34	Diethylene Glycol Monoethyl Ether	8	134,2	C6 H14 O3	111-90-0	CCOCCOCCO
35	Diisopropyl dimerate	1	645,18	C42 H76 O4	308065-68-1	O=C(OC(C)C)CCCCCCCC=CC1C(C=CC(CCCCC)C1CCCC)CCCCCCCC(=O)OC(C)C
36	Dimethicone	4	236,59	C8 H24 O2 Si3	107-51-7	O([Si](O[Si](C)(C)C)(C)C)[Si](C)(C)C
37	Diisopropyl adipate	1	230,34	C12 H22 O4	6938-94-9	O=C(OC(C)C)CCCC(=O)OC(C)C
38	Diocetyl malate	1	358,58	C20 H38 O5	15763-02-7	O=C(OCCCCCCCC)CC(O)C(=O)OCCCCCCCC
39	Diocetyl succinate	1	342,58	C20 H38 O4	14491-66-8	O=C(OCCCCCCCC)CCC(=O)OCCCCCCCC
40	Eiosanoic acid	1	312,6	C20 H40 O2	506-30-9	O=C(O)CCCCCCCCCCCCCCCC
41	Erythrosine(D&C red 3)	3	879,9	C20H6I4Na2O 5	16423-68-0	C1=CC=C2C(=C1)C(=O)OC23C4=CC(=C(C(=C4OC5=C(C(=C(C=C35)I)[O-])I)[O-])I.[Na+].[Na+]
42	EthanolSD40	2	46,08	C2 H6 O	64-17-5	OCC
43	Ether ya da ethyl ether	8	74,14	C4 H10 O	60-29-7	O(CC)CC

[illegible]

	Chemical class name	Chem. class no	Molecules in each class		
	Fatty acid and derivatives	1	33		
	Alcohols-fatty alcohols-sugars and their derivatives	2	43		
	Pigments	3	13		
	Silicones	4	2		
	Sterols	5	3		
	Vitamins	6	6		
	Preservatives and additives	7	7		
	Miscellaneous	8	11		
	Thickeners	9	1		
	Oils	10	2		
	Total		121		

Appendix 8.2_TableS2 (5 pages) A Comedogenic Activity and Their Grade

B. Cluster Membership

C. Training and Test Set for Comedogenicity Model

D. Irritation and Their Grade

*Fatty acids with C8-22 with their comedogenicity grade

ID	Chemical Name	Chemical Class	Cluster members	Training / Test Set for comedogenicity modelling	Formula	Comedogenicity Grade (0-5) (Fulton, 1989)	Comedogenic (ACTIVE-INACTIVE)	Irritancy Grade (0-5) (Fulton, 1989)	Irritation (ACTIVE-INACTIVE)
1	1-3 butanediol (Butylene glycol)	2	cluster 3	Training Set	C4 H10 O2	1	ACTIVE	0	INACTIVE
2	Acetone	8	cluster 3	Training Set	C3 H6 O	0	INACTIVE	0	INACTIVE
3	Allantoin	7	cluster 3	Training Set	C4 H6 N4 O3	0	INACTIVE	0	INACTIVE
4	Alpha tocopherol acetate	6	cluster 5	Training Set	C31 H52 O3	0	INACTIVE	0	INACTIVE
5	Alpha tocopherol	6	cluster 5	Test Set	C29 H50 O2	3	ACTIVE	0-3*	ACTIVE
6	Ascorbyl palmitate	6	cluster 5	Training Set	C22 H38 O7	2	ACTIVE	0	INACTIVE
7	*Behenic acid (C22)	1	cluster 5	Training Set	C22 H44 O2	0	INACTIVE	0	INACTIVE
8	Behenyl erucate	1	cluster 1	Training Set	C44 H86 O2	0	INACTIVE	0	INACTIVE
9	Butyl stearate	1	cluster 5	Training Set	C22 H44 O2	3	ACTIVE	0	INACTIVE
10	*Capric acid (C10)	1	cluster 5	Test Set	C10 H20 O2	2	ACTIVE	2	ACTIVE
11	*Caprylic acid (Octanoic acid) (C8)	1	cluster 3	Training Set	C8 H16 O2	1	ACTIVE	3	ACTIVE
12	Carmine	3	cluster 6	Training Set	C22 H20 O13	0	INACTIVE	0	INACTIVE
13	Castor oil	10	cluster 2	Training Set	C57 H104 O9	1	ACTIVE	0	INACTIVE
14	Cetyl acetate	1	cluster 5	Training Set	C18 H36 O2	4	ACTIVE	2	ACTIVE
15	Cetyl alcohol	2	cluster 5	Training Set	C16 H34 O	2	ACTIVE	2	ACTIVE
16	Cetyl palmitate	1	cluster 1	Training Set	C32 H64 O2	0	INACTIVE	0	INACTIVE
17	Cholesterol	5	cluster 5	Training Set	C27 H46 O	0	INACTIVE	0	INACTIVE
18	Choleth 24	5	cluster 5	Training Set	C29 H50 O2	0	INACTIVE	0	INACTIVE

19	Chondroitin sulfate	7	cluster 4	Training Set	C13 H21 N O15 S	0	INACTIVE	0	INACTIVE
20	Cyclomethicone	4	cluster 3	Training Set	C8 H24 O4 Si4	0	INACTIVE	0	INACTIVE
21	D&C red 21	3	cluster 6	Training Set	C20 H8 O5 Br4	2	ACTIVE	0	INACTIVE
22	D&C Red 40 (Allura Red)	3	cluster 2	Training Set	C18 H14 N2 O8 S2 Na2	2	ACTIVE	2	ACTIVE
23	D&C red17	3	cluster 6	Training Set	C22 H16 N4 O	3	ACTIVE	0	INACTIVE
24	D&C red19	3	cluster 6	Training Set	C28 H31 N2 O3 Cl	2	ACTIVE	0	INACTIVE
25	D&C red27	3	cluster 6	Test Set	C20 H4 O5 Br4 Cl4	2	ACTIVE	0	INACTIVE
26	D&C red30	3	cluster 6	Training Set	C18 H10 O2 Cl2 S2	3	ACTIVE	0	INACTIVE
27	D&C red33	3	cluster 6	Training Set	C16 H11 N3 O7 S2Na2	1	ACTIVE	2	ACTIVE
28	D&C red36	3	cluster 6	Test Set	C16 H10 N3 O3 Cl	3	ACTIVE	0	INACTIVE
29	D&C red4	3	cluster 6	Training Set	C18 H16 N2 O7 S2	2	ACTIVE	1	ACTIVE
30	D&C red6	3	cluster 6	Test Set	C18H12N2Na2O6S	1	ACTIVE	0	INACTIVE
31	D&C red9	3	cluster 6	Training Set	Ba(C17H12ClN2O4S)2	1	ACTIVE	0	INACTIVE
32	Decyl oleate	1	cluster 1	Training Set	C28 H54 O2	3	ACTIVE	0	INACTIVE
33	Di (2 ethylhexyl) succinate	1	cluster 5	Training Set	C20 H38 O4	2	ACTIVE	0	INACTIVE
34	Diethylene Glycol Monoethyl Ether	8	cluster 3	Test Set	C6 H14 O3	0	INACTIVE	0	INACTIVE
35	Diisopropyl dimerate	1	cluster 1	Test Set	C42 H76 O4	0	INACTIVE	0	INACTIVE
36	Dimethicone	4	cluster 3	Test Set	C8 H24 O2 Si3	1	ACTIVE	0	INACTIVE
37	Diisopropyl adipate	1	cluster 3	Training Set	C12 H22 O4	0	INACTIVE	0	INACTIVE
38	Dioctyl malate	1	cluster 5	Training Set	C20 H38 O5	3	ACTIVE	1	ACTIVE
39	Dioctyl succinate	1	cluster 5	Test Set	C20 H38 O4	3	ACTIVE	2	ACTIVE
40	Eiosanoic acid	1	cluster 5	Training Set	C20 H40 O2	2	ACTIVE	0	INACTIVE
41	Erythrosine(D&C red 3)	3	cluster 6	Training Set	C20H6I4Na2O5	3	ACTIVE	0	INACTIVE
42	EthanolSD40	2	cluster 3	Training Set	C2 H6 O	0	INACTIVE	0	INACTIVE
43	Ether ya da ethyl ether	8	cluster 3	Training Set	C4 H10 O	0	INACTIVE	0	INACTIVE
44	Ethylene glycol monoethyl ether	8	cluster 3	Test Set	C4 H10 O2	0	INACTIVE	0	INACTIVE
45	Ethylene glycol monostearate	2	cluster 5	Training Set	C20 H40 O3	0	INACTIVE	0	INACTIVE
46	Ethylhexyl palmitate	1	cluster 5	Training Set	C24 H48 O2	4	ACTIVE	0	INACTIVE

47	Ethylhexyl pelargonate	1	cluster 5	Training Set	C17 H34 O2	2	ACTIVE	3	ACTIVE
48	Glucose glutamate	2	cluster 4	Training Set	C11 H19 N O9	0	INACTIVE	0	INACTIVE
49	Glycereth-26	2	cluster 3	Training Set	C9 H20 O6	0	INACTIVE	0	INACTIVE
50	Glycerin	2	cluster 3	Training Set	C3 H8 O3	0	INACTIVE	0	INACTIVE
51	Glyceryl stearate NSE	2	cluster 5	Training Set	C21 H42 O4	1	ACTIVE	0	INACTIVE
52	Hexylene glycol	2	cluster 3	Training Set	C6 H14 O2	2	ACTIVE	0-1	ACTIVE
53	Hydantoin	7	cluster 3	Training Set	C3 H4 N2 O2	0	INACTIVE	0	INACTIVE
54	Isocetyl alcohol	2	cluster 5	Training Set	C16 H34 O	4	ACTIVE	4	ACTIVE
55	Isoodecyl oleate	1	cluster 5	Test Set	C28 H54 O2	3	ACTIVE	1-2	ACTIVE
56	Isopropyl alcohol	2	cluster 3	Test Set	C3 H8 O	0	INACTIVE	0	INACTIVE
57	Isopropyl isostearate	1	cluster 5	Training Set	C21 H42 O2	5	ACTIVE	0	INACTIVE
58	Isopropyl linoleate	1	cluster 5	Test Set	C21 H38 O2	4	ACTIVE	2	ACTIVE
59	Isopropyl myristate	1	cluster 5	Training Set	C17 H34 O2	5	ACTIVE	3	ACTIVE
60	Isopropyl neopentanoate	1	cluster 5	Training Set	C23 H46 O2	3	ACTIVE	3	ACTIVE
61	Isopropyl palmitate	1	cluster 5	Training Set	C19 H38 O2	4	ACTIVE	1	ACTIVE
62	Isostearyl isostearate	1	cluster 1	Test Set	C36 H72 O2	4	ACTIVE	1	ACTIVE
63	Laureth 23 (Polyethylene (23) lauryl ether)	2	cluster 2	Training Set	C58 H118 O24	3	ACTIVE	0	INACTIVE
64	Laureth 4	2	cluster 5	Test Set	C14 H30 O2	5	ACTIVE	4	ACTIVE
65	*Lauric acid (Dodecanoic acid) (C12)	1	cluster 5	Training Set	C12 H24 O2	4	ACTIVE	1	ACTIVE
66	Lithium stearate	8	cluster 2	Training Set	C18 H35 O2 Li	1	ACTIVE	0	INACTIVE
67	Methyl paraben	7	cluster 3	Training Set	C8 H8 O3	0	INACTIVE	0	INACTIVE
68	*Myristic acid (Tetradecanoic acid) (C14)	1	cluster 5	Training Set	C14 H28 O2	3	ACTIVE	0	INACTIVE
69	Myristyl alcohol	2	cluster 5	Training Set	C14 H30 O	2	ACTIVE	4	ACTIVE
70	Myristyl lactate	1	cluster 5	Test Set	C17 H34 O3	4	ACTIVE	2	ACTIVE
71	Myristyl myristate	1	cluster 1	Training Set	C28 H56 O2	5	ACTIVE	2	ACTIVE
74	Octyl dimethyl PABA	8	cluster 3	Training Set	C17 H27 N O2	0	INACTIVE	0	INACTIVE
75	Octyl methoxycinnamate	8	cluster 3	Training Set	C18 H26 O3	0	INACTIVE	0	INACTIVE

76	Octyl salicylate	8	cluster 3	Test Set	C15 H22 O3	0	INACTIVE	0	INACTIVE
77	Octyldodecyl stearate	1	cluster 1	Training Set	C38 H76 O2	0	INACTIVE	0	INACTIVE
78	Octyldodecyl stearyl stearate	1	cluster 2	Training Set	C56 H110 O4	0	INACTIVE	0	INACTIVE
79	Oleth-10	2	cluster 4	Training Set	C38 H76 O11	2	ACTIVE	1	ACTIVE
80	Oleth-20	2	cluster 5	Training Set	C20 H40 O2	1	ACTIVE	0	INACTIVE
81	Oleth-3	2	cluster 5	Training Set	C24 H48 O4	5	ACTIVE	2	ACTIVE
82	Oleth-3 phosphate	2	cluster 5	Training Set	C24 H49 O7 P	2	ACTIVE	2	ACTIVE
83	Oleth-5	2	cluster 5	Training Set	C28 H56 O6	3	ACTIVE	2	ACTIVE
84	Oleth alcohol	2	cluster 5	Training Set	C18 H36 O	4	ACTIVE	2	ACTIVE
85	Oxybenzone	8	cluster 6	Training Set	C14 H12 O3	0	INACTIVE	0	INACTIVE
86	*Palmitic acid (Hexadecanoic acid)(C16)	1	cluster 5	Training Set	C16 H32 O2	2	ACTIVE	0	INACTIVE
87	Panthenol	6	cluster 3	Training Set	C9 H19 N O4	0	INACTIVE	0	INACTIVE
88	Papain	6	cluster 3	Training Set	C9 H14 N4 O3	0	INACTIVE	0	INACTIVE
89	PEG 400	2	cluster 3	Training Set	C18 H38 O10	1	ACTIVE	0	INACTIVE
90	PEG 200 (600) dilaurate	2	cluster 1	Training Set	C28 H52 O5	3	ACTIVE	2	ACTIVE
91	PEG 8 stearate	2	cluster 4	Training Set	C34 H68 O10	3	ACTIVE	1	ACTIVE
92	Pentaerythritol tetra isostearate	2	cluster 2	Test Set	C77 H148 O8	2	ACTIVE	0	INACTIVE
93	Phenoxyethyl paraben (4-hydroxybenzoic acid)	7	cluster 3	Training Set	C7 H6 O3	0	INACTIVE	0	INACTIVE
94	Phytantriol	5	cluster 5	Training Set	C20 H42 O3	2	ACTIVE	2	ACTIVE
95	Polysorbate 20	2	cluster 4	Test Set	C26 H50 O10	0	INACTIVE	0	INACTIVE
96	Polysorbate 80	2	cluster 4	Training Set	C32 H60 O10	0	INACTIVE	0	INACTIVE
97	PPG 10 cetyl ether	2	cluster 5	Test Set	C19 H40 O2	3	ACTIVE	1	ACTIVE
98	PPG 2 myristyl ether propionate	2	cluster 5	Training Set	C23 H46 O4	3	ACTIVE	2	ACTIVE
99	Propyl paraben	7	cluster 3	Test Set	C10 H12 O3	0	INACTIVE	0	INACTIVE
100	Propylene glycol	2	cluster 3	Training Set	C3 H8 O2	0	INACTIVE	0	INACTIVE
101	Propylene glycol (PG) diperlargonate	2	cluster 5	Test Set	C21 H40 O4	2	ACTIVE	2	ACTIVE
102	PVP	9	cluster 3	Training Set	C6 H9 N O	0	INACTIVE	0	INACTIVE

104	Sodium hyaluronate	7	cluster 2	Training Set	C28H42N2O23Na2	0	INACTIVE	0	INACTIVE
106	Sodium PCA (without sodium)	8	cluster 3	Training Set	C5 H7 N O3	0	INACTIVE	0	INACTIVE
107	Sorbitan isostearate	2	cluster 5	Test Set	C24 H46 O6	2	ACTIVE	0	INACTIVE
108	Sorbitan laurate	2	cluster 5	Training Set	C18 H34 O6	2	ACTIVE	1-2	ACTIVE
109	Sorbitan oleate	2	cluster 5	Test Set	C24 H44 O6	3	ACTIVE	0	INACTIVE
110	Sorbitan stearate	2	cluster 5	Training Set	C24 H46 O6	0	INACTIVE	1	ACTIVE
111	Sorbitol	2	cluster 3	Test Set	C6 H14 O6	0	INACTIVE	0	INACTIVE
112	Squalene	10	cluster 5	Training Set	C30 H50	1	ACTIVE	0	INACTIVE
113	Stearath-20	2	cluster 5	Training Set	C20 H42 O2	2	ACTIVE	1	ACTIVE
114	Steareth 10	2	cluster 4	Test Set	C38 H78 O11	4	ACTIVE	3	ACTIVE
115	Steareth-100	2	cluster 2	Training Set	C58 H118 O21	0	INACTIVE	0	INACTIVE
116	Steareth-2	2	cluster 5	Test Set	C22 H46 O3	2	ACTIVE	2	ACTIVE
117	*Stearic acid (C18)	1	cluster 5	Training Set	C18 H36 O2	3	ACTIVE	0	INACTIVE
118	Stearyl alcohol	2	cluster 5	Training Set	C18 H38 O	2	ACTIVE	2	ACTIVE
119	Stearyl heptanoate	1	cluster 5	Training Set	C25 H50 O2	4	ACTIVE	0	INACTIVE
120	Sucrose distearate	2	cluster 2	Training Set	C48 H90 O13	0	INACTIVE	2	ACTIVE
121	Sucrose stearate	2	cluster 4	Training Set	C30 H56 O12	0	INACTIVE	0	INACTIVE
122	Triacetin	2	cluster 3	Test Set	C9 H14 O6	0	INACTIVE	0	INACTIVE
123	Tridecyl neopentanoate	1	cluster 5	Test Set	C18 H36 O2	0	INACTIVE	3	ACTIVE
124	Triethanolamine	8	cluster 3	Training Set	C6 H15 N O3	2	ACTIVE	0	INACTIVE
125	Vitamin A palmitate	6	cluster 5	Training Set	C36 H60 O2	3	ACTIVE	1-3*	ACTIVE

Appendix 8.3_TableS3.1Selected Descriptors with original value (4 pages)

ID	Class	Train/Test	ALVA_Descriptors					MOLD2_Descriptors (MLD2)							TEST_Descriptors (TST)								PADEL_Descriptors	
			S3K	piPC01	ATSC7i	MATS6p	CATS2D_08_LL	D130	D338	D489	D590	D606	D679	D715	xvp9	xvc3	Gmin	Qsv	ic	ssi	piPC04	piPC05	ATS2m	AATS3m
1	2	Training	5,26	1,79	0,00	0,39	0,00	0	0,00	0,25	0,956	0	0	1	0,00	0,18	-0,35	0,67	0,00	1,00	1,10	0,00	1075,87	23,84
2	8	Training	3,05	1,61	0,00	0,00	0,00	0	0,00	0,00	0,982	0	0	0	0,00	0,20	0,17	0,76	2,00	0,75	0,00	0,00	607,33	14,12
3	7	Training	1,89	2,71	0,00	0,01	0,00	0	0,00	0,47	2,635	0	0	0	0,00	0,24	-1,04	0,40	0,00	1,00	3,47	3,50	2976,97	71,23
4	6	Training	8,77	3,69	2,58	-0,06	20,00	0	159,96	0,49	3,166	0	0	11	0,63	2,28	-0,26	1,27	2,00	0,99	5,03	5,53	8618,92	42,15
5	6	Test	7,60	3,58	2,25	0,00	20,00	0	145,39	0,45	3,153	0	0	11	0,57	2,20	-0,07	1,37	2,00	0,99	4,92	5,38	7807,10	40,51
6	6	Training	11,02	3,50	1,18	0,05	7,00	0	0,00	0,29	3,02	0	2	14	0,33	0,39	-1,42	0,71	0,00	1,00	4,16	4,22	6801,43	41,51
7	1	Training	22,68	3,22	1,24	0,00	13,00	0	0,00	0,11	2,971	0	0	20	0,44	0,06	-0,65	1,14	0,00	1,00	3,09	3,04	4565,94	22,65
8	1	Training	42,38	3,87	2,80	0,00	31,00	0	0,00	0,15	3,046	2	1	38	0,88	0,06	0,03	1,38	0,00	1,00	3,91	3,91	8810,83	23,61
9	1	Training	20,64	3,22	1,23	-0,02	12,00	0	0,00	0,27	2,93	0	1	18	0,37	0,06	0,00	1,25	0,00	1,00	3,18	3,14	4599,70	24,39
10	1	Test	10,75	2,56	0,33	-0,01	1,00	0	0,00	0,27	2,612	0	0	8	0,06	0,06	-0,66	0,86	0,00	1,00	2,30	2,20	2241,44	23,71
11	1	Training	8,79	2,40	0,18	-0,01	0,00	0	0,00	0,35	2,381	0	0	6	0,00	0,06	-0,67	0,79	0,00	1,00	2,08	1,95	1854,02	24,25
12	3	Training	3,47	3,87	1,38	-0,02	0,00	0	222,00	0,39	3,611	0	1	0	0,44	1,32	-2,03	0,45	0,00	1,00	5,80	6,37	10396,79	101,36
13	10	Training	49,26	4,28	3,76	0,00	29,00	0	0,00	0,30	3,392	6	3	39	0,79	0,68	-0,84	0,96	44,00	0,89	4,55	4,62	13848,64	30,92
14	1	Training	18,70	3,04	0,94	0,00	8,00	0	0,00	0,18	2,925	0	1	14	0,29	0,08	-0,15	1,19	0,00	1,00	2,89	2,83	3824,87	23,66
15	2	Training	15,96	2,83	0,79	0,00	7,00	0	0,00	0,04	2,788	0	0	14	0,25	0,00	0,37	1,35	0,00	1,00	2,64	2,56	3013,04	20,78
16	1	Training	30,64	3,56	1,93	0,00	19,00	0	0,00	0,19	3,018	0	1	28	0,61	0,06	0,03	1,35	0,00	1,00	3,53	3,50	6536,79	23,55
17	5	Training	3,56	3,50	2,21	0,09	22,00	0	148,49	0,23	3,063	0	0	11	2,02	2,68	-0,08	1,42	2,00	0,99	4,68	4,99	8055,80	39,38
18	5	Training	4,42	3,58	2,50	0,09	22,00	0	178,39	0,23	3,068	0	1	11	2,16	2,67	0,15	1,34	2,00	0,99	4,73	5,04	8699,36	40,12
19	7	Training	5,97	3,58	1,09	0,04	0,00	0	0,00	0,47	3,355	0	3	0	0,24	1,27	-5,19	0,41	2,00	0,99	4,43	4,70	9877,92	86,50
20	4	Training	5,63	2,83	0,51	0,01	0,00	0	0,00	0,44	2,493	0	0	0	0,00	6,90	-2,11	1,38	40,00	0,38	3,71	3,22	8534,44	104,76
21	3	Training	2,60	3,78	0,66	0,20	7,00	4	167,92	0,68	4,068	0	1	0	0,76	1,82	-1,33	0,67	18,00	0,87	5,81	6,46	15073,60	266,38
22	3	Training	4,45	3,84	0,95	-0,04	9,00	0	126,62	0,56	4,218	0	0	0	0,24	1,70	-4,72	0,49	9,51	0,94	5,73	6,34	9597,65	116,23
23	3	Training	3,95	3,78	0,89	0,02	10,00	0	113,51	0,48	3,88	0	0	0	0,17	0,50	-0,12	0,71	8,00	0,94	5,48	6,03	6543,49	74,20
24	3	Training	4,33	3,87	1,82	0,04	12,00	0	208,03	0,30	3,923	0	0	0	0,56	0,89	-0,94	0,87	30,00	0,82	5,72	6,34	8573,96	67,51
25	3	Test	3,25	3,87	0,52	0,15	11,00	8	201,20	0,85	4,222	0	1	0	1,04	2,37	-1,72	0,65	18,00	0,89	5,95	6,61	18383,07	354,20
26	3	Training	2,48	3,61	0,51	0,08	10,00	2	0,00	0,66	3,851	0	0	0	0,93	1,55	-0,10	0,72	24,00	0,78	5,56	6,06	9478,86	156,96
27	3	Training	4,11	3,78	1,03	0,20	8,00	0	111,93	0,55	4,245	0	0	0	0,38	1,63	-4,90	0,46	8,00	0,94	5,51	6,09	9154,40	109,99
28	3	Test	3,10	3,61	0,53	0,23	10,00	1	88,02	0,41	3,565	0	0	0	0,13	0,62	-0,53	0,55	2,00	0,98	5,28	5,82	6351,54	102,45
29	3	Training	4,20	3,81	0,95	0,07	10,00	0	116,90	0,53	4,243	0	0	0	0,25	1,81	-4,63	0,51	4,00	0,97	5,55	6,11	9424,28	107,22
30	3	Test	3,77	3,75	0,85	0,20	11,00	0	102,56	0,42	3,858	0	0	0	0,20	1,21	-4,53	0,51	2,00	0,98	5,47	6,01	7914,67	100,45
31	3	Training	3,55	3,68	0,75	-0,05	11,00	1	93,39	0,49	3,986	0	0	0	0,18	1,24	-4,51	0,57	2,00	0,98	5,40	5,95	7825,31	108,13
32	1	Training	26,38	3,47	1,57	0,00	15,00	0	0,00	0,24	2,986	2	1	22	0,38	0,06	0,01	1,27	0,00	1,00	3,53	3,53	5711,49	24,73
33	1	Training	13,29	3,26	1,17	-0,12	4,00	0	0,00	0,49	3,06	0	2	10	0,11	0,53	-0,29	0,98	24,00	0,78	3,43	3,64	5101,21	34,19
34	8	Test	7,88	2,20	0,13	0,26	0,00	0	0,00	0,05	2,177	0	2	0	0,00	0,00	0,09	0,79	0,00	1,00	1,79	1,61	1588,25	33,13
35	1	Test	22,88	3,93	3,79	0,00	54,00	0	0,00	0,29	3,272	2	2	24	1,32	1,11	-0,05	1,18	4,00	0,98	4,55	4,72	10142,58	31,10

36	4	Test	9,78	2,56	0,56	-0,01	0,00	0	0,00	0,22	2,217	0	0	0	0,00	8,40	-1,85	1,76	21,51	0,55	3,00	1,95	5468,81	62,67
37	1	Training	10,61	2,89	0,50	0,00	0,00	0	0,00	0,86	2,723	0	2	4	0,05	0,59	-0,20	0,81	20,00	0,69	3,14	2,64	3639,30	36,62
38	1	Training	16,45	3,30	0,99	-0,03	2,00	0	0,00	0,41	2,953	0	2	13	0,15	0,20	-1,42	0,85	0,00	1,00	3,43	3,47	5246,41	33,61
39	1	Test	18,49	3,26	0,96	-0,02	4,00	0	0,00	0,49	2,95	0	2	14	0,16	0,12	-0,29	0,94	24,00	0,78	3,30	3,40	4859,08	29,69
40	1	Training	20,69	3,14	1,09	0,00	11,00	0	0,00	0,13	2,95	0	0	18	0,37	0,06	-0,65	1,11	0,00	1,00	3,00	2,94	4178,53	22,73
41	3	Training	2,84	3,78	0,95	0,14	7,00	4	167,92	0,87	5,67	0	1	0	0,89	2,13	-1,22	0,72	18,00	0,87	5,81	6,46	19589,78	367,74
42	2	Training	1,96	1,10	0,00	0,00	0,00	0	0,00	0,00	0,47	0	0	0	0,00	0,00	0,25	0,76	0,00	1,00	0,00	0,00	301,12	5,72
43	8	Training	3,96	1,61	0,00	0,39	0,00	0	0,00	0,12	0,932	0	1	0	0,00	0,00	0,84	1,35	4,00	0,66	0,69	0,00	722,30	18,58
44	8	Test	5,26	1,79	0,01	0,22	0,00	0	0,00	0,10	1,326	0	1	0	0,00	0,00	0,13	0,78	0,00	1,00	1,10	0,69	944,69	27,04
45	2	Training	19,55	3,18	1,10	-0,01	9,00	0	0,00	0,24	2,929	0	1	16	0,34	0,06	-0,17	1,03	0,00	1,00	3,14	3,09	4434,67	25,67
46	1	Training	19,24	3,30	1,42	-0,04	14,00	0	0,00	0,27	3,012	0	1	18	0,34	0,26	0,02	1,29	0,00	1,00	3,33	3,37	5108,18	25,93
47	1	Training	12,42	3,00	0,94	-0,02	7,00	0	0,00	0,44	2,883	0	1	12	0,13	0,18	0,01	1,20	0,00	1,00	3,14	3,09	3796,10	27,72
48	2	Training	7,35	3,18	0,46	0,03	0,00	0	0,00	0,67	3,113	1	1	2	0,04	0,59	-1,95	0,43	0,00	1,00	3,40	3,43	5162,89	63,06
49	2	Training	10,06	2,71	0,41	0,07	0,00	0	0,00	0,29	2,627	0	3	0	0,02	0,12	-0,28	0,60	10,00	0,83	2,71	2,71	3035,23	44,51
50	2	Training	2,88	1,79	0,00	0,36	0,00	0	0,00	0,34	1,278	0	0	0	0,00	0,13	-0,95	0,45	4,00	0,74	0,69	0,00	1104,55	40,70
51	2	Training	19,73	3,26	1,18	-0,02	9,00	0	0,00	0,27	2,939	0	1	16	0,36	0,19	-0,95	0,93	0,00	1,00	3,22	3,26	5015,71	28,11
52	2	Training	7,13	2,08	0,00	0,42	0,00	0	0,00	0,49	1,136	0	0	1	0,00	1,08	-0,73	0,88	2,00	0,92	1,95	0,00	1914,25	26,66
53	7	Training	0,83	2,30	0,00	0,00	0,00	0	0,00	0,96	1,459	0	0	0	0,00	0,12	-0,40	0,41	0,00	1,00	2,89	2,56	1787,56	52,01
54	2	Training	15,96	2,83	0,79	0,00	7,00	0	0,00	0,06	2,819	0	0	12	0,24	0,41	0,37	1,39	2,00	0,97	2,64	2,56	3134,11	21,26
55	1	Test	26,38	3,47	1,57	0,00	14,00	0	0,00	0,25	3,034	2	1	20	0,36	0,47	0,01	1,29	2,00	0,99	3,53	3,53	5832,56	25,03
56	2	Test	3,34	1,39	0,00	0,00	0,00	0	0,00	0,00	0,8	0	0	0	0,00	0,26	-0,17	0,95	2,00	0,75	0,00	0,00	659,78	9,56
57	1	Training	19,59	3,18	1,17	0,00	11,00	0	0,00	0,30	2,967	0	1	14	0,34	0,70	-0,03	1,28	4,00	0,96	3,22	3,00	4692,00	24,98
58	1	Test	19,07	3,26	1,06	0,00	11,00	0	0,00	0,34	3,186	4	1	12	0,20	0,29	-0,05	1,15	2,00	0,98	3,53	3,47	4470,02	27,44
59	1	Training	15,58	3,00	0,86	0,00	7,00	0	0,00	0,36	2,85	0	1	12	0,23	0,29	-0,04	1,20	2,00	0,98	3,04	2,77	3796,10	25,31
60	1	Training	18,23	3,26	1,32	0,00	13,00	0	0,00	0,63	2,992	0	1	14	0,37	1,70	-0,37	1,33	6,75	0,94	3,14	3,09	5277,67	27,74
61	1	Training	17,59	3,09	1,01	0,00	9,00	0	0,00	0,32	2,898	0	1	14	0,29	0,29	-0,03	1,23	2,00	0,98	3,14	2,89	4183,52	24,91
62	1	Test	34,64	3,66	2,24	0,00	23,00	0	0,00	0,18	3,052	0	1	28	0,71	0,88	0,03	1,41	4,00	0,98	3,64	3,61	7553,75	23,79
63	2	Training	80,08	4,41	3,39	0,19	3,00	0	0,00	0,02	2,935	0	23	10	0,52	0,00	0,02	0,86	0,00	1,00	4,37	4,36	17040,10	38,69
64	2	Test	15,00	2,77	0,65	-0,01	3,00	0	0,00	0,11	2,716	0	1	10	0,17	0,00	0,15	1,17	0,00	1,00	2,56	2,48	2881,77	23,43
65	1	Training	12,73	2,71	0,48	0,00	3,00	0	0,00	0,22	2,746	0	0	10	0,12	0,06	-0,66	0,93	0,00	1,00	2,48	2,40	2628,86	23,37
66	8	Training	18,54	3,04	0,93	0,00	9,00	0	0,00	0,14	2,923	0	0	16	0,31	0,06	-0,90	1,02	2,00	0,98	2,89	2,83	3779,00	22,95
67	7	Training	1,87	2,77	0,10	-0,13	0,00	0	0,00	0,83	2,607	0	1	0	0,00	0,20	-0,40	0,57	4,00	0,89	4,12	4,59	2483,39	75,83
68	1	Training	14,72	2,83	0,63	0,00	5,00	0	0,00	0,19	2,829	0	0	12	0,19	0,06	-0,66	0,98	0,00	1,00	2,64	2,56	3016,27	23,14
69	2	Training	13,96	2,71	0,64	0,00	5,00	0	0,00	0,05	2,715	0	0	12	0,19	0,00	0,37	1,32	0,00	1,00	2,48	2,40	2625,63	20,61
70	1	Test	14,86	3,04	0,87	0,00	6,00	0	0,00	0,42	2,897	0	1	12	0,24	0,18	-0,99	1,00	0,00	1,00	2,89	2,83	4018,49	28,61
71	1	Training	26,64	3,43	1,63	0,00	15,00	0	0,00	0,21	3,003	0	1	24	0,48	0,06	0,02	1,31	0,00	1,00	3,40	3,37	5761,95	23,81
74	8	Training	5,78	3,22	0,81	0,02	7,00	0	0,00	0,43	3,068	0	1	4	0,08	0,63	-0,22	1,03	6,00	0,93	4,40	4,84	4467,94	44,37
75	8	Training	7,21	3,30	0,78	-0,02	10,00	0	0,00	0,36	3,112	2	1	4	0,08	0,42	-0,28	0,92	4,00	0,96	4,42	4,92	4524,84	43,59
76	8	Test	5,83	3,14	0,62	0,00	5,00	0	0,00	0,34	2,948	0	1	6	0,09	0,18	-0,44	0,79	0,00	1,00	4,32	4,64	3883,23	42,13
77	1	Training	33,03	3,71	2,66	0,00	29,00	0	0,00	0,18	3,027	0	1	32	0,85	0,26	0,05	1,40	0,00	1,00	3,76	3,81	7820,10	24,36

78	1	Training	47,60	4,13	4,23	0,00	47,00	0	0,00	0,23	3,107	0	2	47	1,29	0,44	0,03	1,30	0,00	1,00	4,26	4,30	12118,59	25,86
79	2	Training	47,31	3,91	2,18	0,12	9,00	0	0,00	0,05	2,934	2	10	14	0,41	0,00	0,03	0,96	0,00	1,00	3,91	3,91	9785,61	33,29
80	2	Training	20,72	3,14	1,05	0,00	9,00	0	0,00	0,11	2,926	2	1	14	0,26	0,00	0,15	1,21	0,00	1,00	3,14	3,14	3993,56	24,17
81	2	Training	26,63	3,37	1,31	0,04	9,00	0	0,00	0,09	2,929	2	3	14	0,30	0,00	0,07	1,11	0,00	1,00	3,37	3,37	5280,68	27,31
82	2	Training	30,78	3,53	1,40	0,04	9,00	0	0,00	0,17	2,981	2	3	14	0,33	0,42	-4,40	0,87	2,00	0,99	3,53	3,53	7238,85	31,57
83	2	Training	32,54	3,56	1,56	0,07	9,00	0	0,00	0,07	2,93	2	5	14	0,33	0,00	0,05	1,04	0,00	1,00	3,56	3,56	6567,80	29,60
84	2	Training	17,70	3,00	0,89	0,00	9,00	0	0,00	0,08	2,918	2	0	14	0,21	0,00	0,36	1,30	0,00	1,00	3,00	3,00	3350,00	22,10
85	8	Training	2,41	3,26	0,43	-0,07	0,00	0	0,00	0,39	3,234	0	0	0	0,06	0,34	-0,20	0,65	4,00	0,94	4,91	5,40	4006,99	76,87
86	1	Training	16,70	2,94	0,79	0,00	7,00	0	0,00	0,16	2,885	0	0	14	0,25	0,06	-0,65	1,03	0,00	1,00	2,77	2,71	3403,69	22,97
87	6	Training	4,55	2,71	0,33	-0,13	0,00	0	0,00	0,77	2,642	0	0	1	0,00	1,17	-1,23	0,63	2,00	0,96	2,94	2,56	3217,27	48,88
88	6	Training	4,35	3,07	0,33	-0,10	0,00	0	0,00	0,66	3,057	0	0	2	0,02	0,33	-1,08	0,54	0,00	1,00	4,10	3,99	3771,19	52,55
89	2	Training	26,65	3,33	0,87	0,25	0,00	0	0,00	0,01	2,839	0	8	0	0,10	0,00	0,03	0,72	28,00	0,79	3,22	3,18	5671,99	41,81
90	2	Training	12,15	3,61	1,64	0,00	13,00	0	0,00	0,34	3,02	0	1	22	0,60	0,78	-0,98	0,88	32,00	0,81	3,91	3,76	7628,31	29,06
91	2	Training	40,32	3,81	1,96	0,11	9,00	0	0,00	0,13	2,932	0	8	16	0,45	0,06	-0,13	0,91	0,00	1,00	3,78	3,76	8939,59	32,78
92	2	Test	65,32	4,49	5,69	-0,01	46,00	0	0,00	0,35	3,237	0	4	56	1,58	2,58	-1,21	1,19	176,00	0,68	4,63	4,80	18041,51	28,43
93	7	Training	1,65	2,71	0,03	0,12	0,00	0	0,00	0,43	2,607	0	0	0	0,00	0,20	-0,99	0,47	4,00	0,88	4,07	4,54	2299,81	71,80
94	5	Training	11,97	3,14	1,35	0,09	9,00	0	0,00	0,19	3,008	0	0	9	0,24	1,66	-1,17	1,11	2,00	0,98	3,18	3,04	5211,75	31,56
95	2	Test	16,03	3,64	1,83	0,02	3,00	0	0,00	0,30	2,982	0	6	10	0,34	0,41	-0,55	0,76	0,00	1,00	4,09	4,19	8290,87	42,02
96	2	Training	21,43	3,81	2,22	0,01	9,00	0	0,00	0,30	3,253	2	6	15	0,43	0,40	-0,58	0,81	0,00	1,00	4,23	4,34	9462,45	38,35
97	2	Test	19,92	3,04	1,03	-0,01	8,00	0	0,00	0,16	2,847	0	1	15	0,32	0,00	0,25	1,26	0,00	1,00	2,89	2,83	3850,31	22,90
98	2	Training	20,11	3,33	1,35	-0,03	8,00	0	0,00	0,47	2,993	0	3	13	0,29	0,39	-0,20	1,14	0,00	1,00	3,37	3,30	5635,59	30,35
99	7	Test	2,97	2,89	0,29	-0,14	0,00	0	0,00	0,70	2,667	0	1	1	0,02	0,20	-0,35	0,65	4,00	0,92	4,18	4,65	2914,69	57,08
100	2	Training	3,92	1,61	0,00	0,00	0,00	0	0,00	0,62	0,876	0	0	0	0,00	0,18	-0,56	0,60	0,00	1,00	0,00	0,00	882,17	27,08
101	2	Test	17,88	3,30	1,10	0,00	4,00	0	0,00	0,50	2,931	0	2	14	0,18	0,28	-0,36	0,98	0,00	1,00	3,43	3,40	5217,73	30,70
102	9	Training	0,98	2,40	0,00	0,68	0,00	0	0,00	0,66	2,198	1	0	2	0,00	0,16	0,21	0,73	0,00	1,00	3,00	2,40	1846,11	38,54
104	7	Training	9,47	4,11	3,06	0,04	0,00	0	0,00	0,39	3,393	0	7	0	0,56	1,90	-2,30	0,46	0,00	1,00	5,06	5,37	16261,56	78,02
106	8	Training	1,34	2,48	0,00	0,59	0,00	0	0,00	0,95	2,035	0	0	2	0,00	0,23	-0,94	0,44	0,00	1,00	3,00	3,04	2311,87	48,95
107	2	Test	12,83	3,47	1,42	0,00	9,00	0	0,00	0,25	3,052	0	2	14	0,39	0,83	-1,14	0,86	2,00	0,99	3,87	3,87	6999,95	35,40
108	2	Training	8,96	3,26	0,89	0,08	3,00	0	0,00	0,33	2,98	0	2	10	0,22	0,43	-1,14	0,74	0,00	1,00	3,64	3,61	5716,63	38,89
109	2	Test	13,30	3,50	1,29	0,06	9,00	0	0,00	0,28	3,231	2	2	14	0,31	0,43	-1,15	0,82	0,00	1,00	3,87	3,87	6828,42	36,14
110	2	Training	13,50	3,47	1,34	0,06	9,00	0	0,00	0,26	3,028	0	2	16	0,41	0,43	-1,14	0,84	0,00	1,00	3,78	3,76	6878,88	34,51
111	2	Test	3,36	2,48	0,07	-0,03	0,00	0	0,00	0,19	2,172	0	0	0	0,00	0,38	-1,67	0,43	12,00	0,72	2,56	2,20	2847,67	58,74
112	10	Training	20,52	3,58	1,82	0,15	24,00	0	0,00	0,21	3,527	6	0	10	0,22	1,39	1,17	1,54	34,00	0,77	4,04	4,23	5859,66	33,25
113	2	Training	20,97	3,09	1,11	0,00	9,00	0	0,00	0,08	2,865	0	1	16	0,36	0,00	0,15	1,28	0,00	1,00	2,94	2,89	4044,02	22,94
114	2	Test	47,57	3,89	2,23	0,11	9,00	0	0,00	0,03	2,933	0	10	16	0,51	0,00	0,03	0,98	0,00	1,00	3,83	3,81	9836,07	32,27
115	2	Training	77,18	4,37	3,48	0,16	9,00	0	0,00	0,02	2,939	0	20	16	0,66	0,00	0,02	0,91	0,00	1,00	4,33	4,32	16271,67	36,16
116	2	Test	23,88	3,22	1,24	0,02	9,00	0	0,00	0,07	2,881	0	2	16	0,39	0,00	0,10	1,20	0,00	1,00	3,09	3,04	4687,58	24,64
117	1	Training	18,70	3,04	0,94	0,00	9,00	0	0,00	0,14	2,923	0	0	16	0,31	0,06	-0,65	1,07	0,00	1,00	2,89	2,83	3791,11	22,83
118	2	Training	17,96	2,94	0,95	0,00	9,00	0	0,00	0,04	2,84	0	0	16	0,31	0,00	0,37	1,38	0,00	1,00	2,77	2,71	3400,46	20,90
119	1	Training	23,60	3,33	1,40	0,00	14,00	0	0,00	0,24	2,959	0	1	21	0,41	0,06	0,01	1,28	0,00	1,00	3,30	3,26	5180,83	24,07

120	2	Training	23,09	4,17	3,84	0,01	19,00	0	0,00	0,24	3,225	0	4	32	1,05	1,24	-2,85	0,80	0,00	1,00	4,92	5,18	15129,46	43,49
121	2	Training	11,94	3,81	2,24	0,02	9,00	0	0,00	0,26	3,226	0	4	16	0,55	1,00	-2,24	0,68	0,00	1,00	4,47	4,65	11025,83	48,36
122	2	Test	9,22	2,89	0,33	0,21	1,00	0	0,00	0,67	3,098	0	3	0	0,00	0,37	-0,75	0,56	10,00	0,83	3,00	3,14	3540,02	58,48
123	1	Test	13,42	3,04	0,94	0,00	8,00	0	0,00	0,78	2,882	0	1	11	0,22	1,29	-0,36	1,25	4,75	0,94	2,89	2,83	4188,06	29,09
124	8	Training	5,38	2,30	0,04	0,14	0,00	0	0,00	0,05	2,394	0	0	0	0,00	0,16	0,07	0,62	14,26	0,57	2,30	1,95	1883,18	35,54
125	6	Training	17,54	3,81	2,19	0,13	23,00	0	0,00	0,35	3,482	6	1	17	0,41	1,60	-0,07	1,26	2,00	0,99	4,64	4,85	8171,09	34,03

Appendix 8.3_Tables S3.2 Selected Descriptors for modelling with normalized value (5 pages)

			ALVA_Descriptors					MOLD2_Descriptors (MLD2)							TEST_Descriptors (TST)							PADEL_Descriptors (PDL)				
Train. No	ID	Class	S3K	piPC01	ATSC7i	MATS6p	CATS2 D_08_LL	D130	D338	D489	D590	D606	D679	D715	xvp9	xvc3	Gmin	Qsv	ic	ssi	piPC04	piPC05	ATS2m	ATS3m	Sp Max 1_B hi	SI 2
1	1	2	0,06	0,21	0,00	0,64	0,00	0,00	0,00	0,26	0,09	0,00	0,00	0,02	0,00	0,03	0,84	0,32	0,05	0,60	0,00	0,00	0,04	0,05	0,24	0,02
2	2	8	0,03	0,15	0,00	0,16	0,00	0,00	0,00	0,00	0,10	0,00	0,00	0,00	0,00	0,00	0,86	0,32	0,00	1,00	0,00	0,00	0,02	0,02	0,16	0,02
3	3	7	0,01	0,49	0,00	0,18	0,00	0,00	0,00	0,49	0,42	0,00	0,00	0,00	0,00	0,03	0,73	0,17	0,00	1,00	0,00	0,00	0,14	0,18	0,50	0,02
4	4	6	0,10	0,78	0,61	0,08	0,43	0,00	0,72	0,51	0,52	0,00	0,00	0,23	0,00	0,00	0,95	0,83	0,09	0,45	0,12	0,00	0,43	0,10	0,69	0,02
5	6	6	0,13	0,72	0,28	0,22	0,15	0,00	0,00	0,30	0,49	0,00	0,09	0,30	0,00	0,02	0,67	0,04	0,09	0,59	0,12	0,00	0,34	0,10	0,48	0,02
6	7	1	0,28	0,64	0,29	0,16	0,28	0,00	0,00	0,12	0,48	0,00	0,00	0,43	0,00	0,03	0,76	0,23	0,00	1,00	0,19	0,00	0,22	0,05	0,41	0,02
7	8	1	0,52	0,84	0,66	0,16	0,66	0,00	0,00	0,16	0,50	0,33	0,04	0,81	0,00	0,16	0,70	0,42	0,05	0,87	0,33	0,00	0,44	0,05	0,42	0,02
8	9	1	0,25	0,64	0,29	0,14	0,26	0,00	0,00	0,28	0,47	0,00	0,04	0,38	0,00	0,01	0,71	0,34	0,00	1,00	0,36	0,30	0,22	0,05	0,41	0,02
9	11	1	0,10	0,39	0,04	0,15	0,00	0,00	0,00	0,37	0,37	0,00	0,00	0,13	0,00	0,02	0,83	0,19	0,32	0,31	0,40	0,30	0,08	0,05	0,35	0,02
10	12	3	0,03	0,84	0,33	0,14	0,00	0,00	1,00	0,41	0,60	0,00	0,04	0,00	0,06	0,01	0,71	0,46	0,00	1,00	0,43	0,37	0,52	0,26	0,77	0,02
11	13	10	0,61	0,96	0,89	0,17	0,62	0,00	0,00	0,31	0,56	1,00	0,13	0,83	0,09	0,00	0,87	0,81	0,00	1,00	0,43	0,37	0,70	0,07	0,42	0,02
12	14	1	0,23	0,59	0,22	0,16	0,17	0,00	0,00	0,19	0,47	0,00	0,04	0,30	0,12	0,00	0,87	0,84	0,00	1,00	0,45	0,40	0,18	0,05	0,40	0,02
13	15	2	0,19	0,52	0,19	0,16	0,15	0,00	0,00	0,04	0,45	0,00	0,00	0,30	0,11	0,06	0,87	0,87	0,05	0,95	0,45	0,40	0,14	0,04	0,40	0,02
14	16	1	0,38	0,74	0,46	0,16	0,40	0,00	0,00	0,19	0,49	0,00	0,04	0,60	0,09	0,01	0,71	0,51	0,00	1,00	0,45	0,40	0,32	0,05	0,41	0,02
15	17	5	0,03	0,72	0,52	0,27	0,47	0,00	0,67	0,24	0,50	0,00	0,00	0,23	0,01	0,02	0,77	0,17	0,23	0,73	0,47	0,42	0,40	0,09	0,74	0,02
16	18	5	0,05	0,75	0,59	0,28	0,47	0,00	0,80	0,24	0,50	0,00	0,04	0,23	0,12	0,01	0,71	0,55	0,00	1,00	0,48	0,42	0,44	0,10	0,74	0,02
17	19	7	0,06	0,75	0,26	0,21	0,00	0,00	0,00	0,48	0,55	0,00	0,13	0,00	0,15	0,00	0,87	0,86	0,00	1,00	0,48	0,42	0,50	0,22	0,53	0,02
18	20	4	0,06	0,52	0,12	0,18	0,00	0,00	0,00	0,45	0,39	0,00	0,00	0,00	0,13	0,01	0,79	0,69	0,00	1,00	0,50	0,44	0,43	0,27	0,32	0,02
19	21	3	0,02	0,81	0,16	0,41	0,15	1,00	0,76	0,71	0,69	0,00	0,04	0,00	0,00	0,02	0,75	0,00	0,00	1,00	0,50	0,40	0,77	0,72	1,00	0,02
20	22	3	0,05	0,83	0,22	0,11	0,19	0,00	0,57	0,59	0,72	0,00	0,00	0,00	0,14	0,01	0,67	0,55	0,05	0,96	0,50	0,44	0,48	0,31	0,80	0,02
21	23	3	0,04	0,81	0,21	0,19	0,21	0,00	0,51	0,50	0,66	0,00	0,00	0,00	0,14	0,01	0,71	0,59	0,00	1,00	0,50	0,44	0,32	0,19	0,67	0,02
22	24	3	0,04	0,84	0,43	0,21	0,26	0,00	0,94	0,31	0,66	0,00	0,00	0,00	0,00	0,17	0,62	0,20	0,05	0,94	0,51	0,40	0,43	0,17	0,75	0,02
23	26	3	0,02	0,76	0,12	0,26	0,21	0,50	0,00	0,68	0,65	0,00	0,00	0,00	0,17	0,00	0,84	0,77	0,00	1,00	0,51	0,45	0,48	0,42	0,70	0,02
24	27	3	0,04	0,81	0,24	0,40	0,17	0,00	0,50	0,57	0,73	0,00	0,00	0,00	0,17	0,01	0,71	0,62	0,00	1,00	0,52	0,46	0,46	0,29	0,77	0,02
25	29	3	0,04	0,82	0,22	0,25	0,21	0,00	0,53	0,56	0,73	0,00	0,00	0,00	0,10	0,00	0,87	0,79	0,00	1,00	0,52	0,46	0,47	0,28	0,74	0,02
26	31	3	0,03	0,78	0,18	0,10	0,23	0,25	0,42	0,50	0,68	0,00	0,00	0,00	0,00	0,02	0,85	0,28	0,00	1,00	0,52	0,37	0,39	0,28	0,71	0,02

27	32	1	0,32	0,72	0,37	0,16	0,32	0,00	0,00	0,25	0,48	0,33	0,04	0,47	0,00	0,03	0,67	0,03	0,00	1,00	0,52	0,47	0,28	0,05	0,41	0,00
28	33	1	0,16	0,65	0,28	0,01	0,09	0,00	0,00	0,51	0,50	0,00	0,09	0,21	0,10	0,04	0,81	0,70	0,05	0,96	0,52	0,43	0,25	0,08	0,41	0,00
29	37	1	0,12	0,54	0,12	0,16	0,00	0,00	0,00	0,89	0,43	0,00	0,09	0,09	0,20	0,01	0,71	0,65	0,00	1,00	0,53	0,47	0,17	0,09	0,36	0,00
30	38	1	0,20	0,66	0,23	0,13	0,04	0,00	0,00	0,42	0,48	0,00	0,09	0,28	0,02	0,09	0,78	0,36	0,45	0,50	0,54	0,41	0,26	0,08	0,38	0,00
31	40	1	0,25	0,62	0,26	0,16	0,23	0,00	0,00	0,13	0,48	0,00	0,00	0,38	0,16	0,01	0,79	0,56	0,00	1,00	0,54	0,48	0,20	0,05	0,41	0,00
32	41	3	0,03	0,81	0,22	0,33	0,15	1,00	0,76	0,90	1,00	0,00	0,04	0,00	0,06	0,03	0,82	0,70	0,00	1,00	0,54	0,48	1,00	1,00	1,00	0,00
33	42	2	0,01	0,00	0,00	0,16	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,17	0,25	0,76	0,82	0,15	0,91	0,54	0,48	0,00	0,00	0,00	0,00
34	43	8	0,04	0,15	0,00	0,65	0,00	0,00	0,00	0,12	0,09	0,00	0,04	0,00	0,13	0,04	0,81	0,73	0,05	0,97	0,54	0,45	0,02	0,04	0,13	0,00
35	45	2	0,24	0,63	0,26	0,15	0,19	0,00	0,00	0,25	0,47	0,00	0,04	0,34	0,12	0,00	0,84	0,71	0,00	1,00	0,54	0,49	0,21	0,06	0,41	0,00
36	46	1	0,23	0,66	0,34	0,11	0,30	0,00	0,00	0,28	0,49	0,00	0,04	0,38	0,17	0,01	0,82	0,74	0,00	1,00	0,55	0,49	0,25	0,06	0,41	0,00
37	47	1	0,15	0,57	0,22	0,14	0,15	0,00	0,00	0,45	0,46	0,00	0,04	0,26	0,11	0,24	0,63	0,63	0,05	0,97	0,55	0,47	0,18	0,06	0,40	0,00
38	48	2	0,08	0,63	0,11	0,20	0,00	0,00	0,00	0,69	0,51	0,17	0,04	0,04	0,17	0,03	0,67	0,47	0,00	1,00	0,55	0,50	0,25	0,16	0,40	0,00
39	49	2	0,12	0,49	0,10	0,25	0,00	0,00	0,00	0,30	0,41	0,00	0,13	0,00	0,16	0,10	0,81	0,77	0,09	0,94	0,55	0,46	0,14	0,11	0,30	0,00
40	50	2	0,03	0,21	0,00	0,61	0,00	0,00	0,00	0,35	0,16	0,00	0,00	0,00	0,05	0,00	0,82	0,28	0,64	0,67	0,55	0,49	0,04	0,10	0,19	0,00
41	51	2	0,24	0,65	0,28	0,13	0,19	0,00	0,00	0,28	0,47	0,00	0,04	0,34	0,19	0,01	0,82	0,78	0,00	1,00	0,57	0,50	0,24	0,06	0,41	0,00
42	52	2	0,08	0,30	0,00	0,67	0,00	0,00	0,00	0,51	0,13	0,00	0,00	0,02	0,16	0,04	0,82	0,79	0,00	1,00	0,57	0,52	0,08	0,06	0,35	0,00
43	53	7	0,00	0,36	0,00	0,16	0,00	0,00	0,00	1,00	0,19	0,00	0,00	0,00	0,14	0,00	0,83	0,62	0,00	1,00	0,58	0,52	0,08	0,13	0,43	1,00
44	54	2	0,19	0,52	0,19	0,16	0,15	0,00	0,00	0,06	0,45	0,00	0,00	0,26	0,13	0,06	0,78	0,65	0,00	1,00	0,58	0,51	0,15	0,04	0,40	0,00
45	57	1	0,24	0,63	0,28	0,16	0,23	0,00	0,00	0,31	0,48	0,00	0,04	0,30	0,02	0,09	0,51	0,02	0,00	1,00	0,58	0,53	0,23	0,05	0,41	0,00
46	59	1	0,19	0,57	0,20	0,16	0,15	0,00	0,00	0,37	0,46	0,00	0,04	0,26	0,22	0,01	0,82	0,80	0,00	1,00	0,58	0,52	0,18	0,05	0,40	0,00
47	60	1	0,22	0,65	0,31	0,16	0,28	0,00	0,00	0,66	0,49	0,00	0,04	0,30	0,05	0,08	0,77	0,51	0,55	0,65	0,59	0,56	0,26	0,06	0,41	0,00
48	61	1	0,21	0,60	0,24	0,16	0,19	0,00	0,00	0,33	0,47	0,00	0,04	0,30	0,07	0,03	0,59	0,39	0,00	1,00	0,59	0,54	0,20	0,05	0,40	0,00
49	63	2	1,00	1,00	0,80	0,40	0,06	0,00	0,00	0,02	0,47	0,00	1,00	0,21	0,00	0,03	0,65	0,00	0,00	1,00	0,60	0,54	0,87	0,09	0,39	0,00
50	65	1	0,15	0,49	0,11	0,16	0,06	0,00	0,00	0,23	0,44	0,00	0,00	0,21	0,28	0,01	0,82	0,83	0,00	1,00	0,61	0,54	0,12	0,05	0,39	0,00
51	66	8	0,22	0,59	0,22	0,16	0,19	0,00	0,00	0,15	0,47	0,00	0,00	0,34	0,18	0,01	0,82	0,76	0,00	1,00	0,61	0,55	0,18	0,05	0,41	0,00
52	67	7	0,01	0,51	0,02	0,00	0,00	0,00	0,00	0,86	0,41	0,00	0,04	0,00	0,15	0,06	0,13	0,41	0,05	0,98	0,61	0,55	0,11	0,19	0,55	0,00
53	68	1	0,18	0,52	0,15	0,16	0,11	0,00	0,00	0,19	0,45	0,00	0,00	0,26	0,15	0,00	0,82	0,56	0,00	1,00	0,61	0,55	0,14	0,05	0,39	0,00
54	69	2	0,17	0,49	0,15	0,16	0,11	0,00	0,00	0,05	0,43	0,00	0,00	0,26	0,10	0,06	0,64	0,30	0,00	1,00	0,63	0,56	0,12	0,04	0,39	0,00
55	71	1	0,33	0,71	0,38	0,16	0,32	0,00	0,00	0,22	0,49	0,00	0,04	0,51	0,00	1,00	0,48	0,86	0,91	0,00	0,64	0,50	0,28	0,05	0,40	0,00
56	74	8	0,06	0,64	0,19	0,19	0,15	0,00	0,00	0,45	0,50	0,00	0,04	0,09	0,39	0,04	0,82	0,88	0,00	1,00	0,65	0,59	0,22	0,11	0,59	0,00
57	75	8	0,08	0,66	0,18	0,14	0,21	0,00	0,00	0,37	0,51	0,33	0,04	0,09	0,21	0,01	0,80	0,45	0,00	1,00	0,65	0,58	0,22	0,10	0,56	0,00
58	77	1	0,41	0,79	0,63	0,16	0,62	0,00	0,00	0,18	0,49	0,00	0,04	0,68	0,19	0,06	0,64	0,39	0,00	1,00	0,65	0,58	0,39	0,05	0,44	0,00
59	78	1	0,59	0,92	1,00	0,16	1,00	0,00	0,00	0,24	0,51	0,00	0,09	1,00	0,41	0,01	0,82	0,86	0,00	1,00	0,67	0,61	0,61	0,06	0,45	0,00
60	79	2	0,59	0,85	0,52	0,31	0,19	0,00	0,00	0,05	0,47	0,33	0,43	0,30	0,19	0,00	0,82	0,49	0,00	1,00	0,67	0,61	0,49	0,08	0,41	0,00
61	80	2	0,25	0,62	0,25	0,16	0,19	0,00	0,00	0,11	0,47	0,33	0,04	0,30	0,28	0,11	0,66	0,42	0,73	0,69	0,67	0,58	0,19	0,05	0,41	0,00
62	81	2	0,33	0,69	0,31	0,21	0,19	0,00	0,00	0,09	0,47	0,33	0,13	0,30	0,10	0,20	1,00	1,00	0,77	0,63	0,70	0,66	0,26	0,06	0,41	0,00
63	82	2	0,38	0,73	0,33	0,21	0,19	0,00	0,00	0,17	0,48	0,33	0,13	0,30	0,00	0,03	0,66	0,06	0,09	0,81	0,70	0,70	0,36	0,07	0,41	0,00

64	83	2	0,40	0,74	0,37	0,25	0,19	0,00	0,00	0,07	0,47	0,33	0,22	0,30	0,01	0,05	0,65	0,12	0,00	1,00	0,71	0,62	0,32	0,07	0,41	0,00
65	84	2	0,21	0,57	0,21	0,16	0,19	0,00	0,00	0,08	0,47	0,33	0,00	0,30	0,00	0,03	0,75	0,15	0,09	0,83	0,71	0,71	0,16	0,05	0,41	0,00
66	85	8	0,02	0,65	0,10	0,07	0,00	0,00	0,00	0,40	0,53	0,00	0,00	0,00	0,15	0,06	0,59	0,27	0,00	1,00	0,72	0,65	0,19	0,20	0,63	0,00
67	86	1	0,20	0,56	0,19	0,16	0,15	0,00	0,00	0,17	0,46	0,00	0,00	0,30	0,20	0,06	0,72	0,36	0,00	1,00	0,73	0,67	0,16	0,05	0,40	0,00
68	87	6	0,05	0,49	0,08	0,00	0,00	0,00	0,00	0,80	0,42	0,00	0,00	0,02	0,60	0,06	0,82	0,79	0,00	1,00	0,73	0,67	0,15	0,12	0,44	0,00
69	88	6	0,04	0,60	0,08	0,04	0,00	0,00	0,00	0,69	0,50	0,00	0,00	0,04	0,31	0,00	0,82	0,44	0,00	1,00	0,74	0,67	0,18	0,13	0,61	0,00
70	89	2	0,33	0,68	0,21	0,47	0,00	0,00	0,00	0,01	0,46	0,00	0,35	0,00	0,24	0,00	0,82	0,40	0,00	1,00	0,75	0,67	0,28	0,10	0,26	0,00
71	90	2	0,14	0,76	0,39	0,16	0,28	0,00	0,00	0,35	0,49	0,00	0,04	0,47	0,04	0,09	0,78	0,55	0,14	0,89	0,76	0,75	0,38	0,06	0,63	0,00
72	91	2	0,50	0,82	0,46	0,29	0,19	0,00	0,00	0,13	0,47	0,00	0,35	0,34	0,04	0,06	0,77	0,46	0,09	0,93	0,76	0,76	0,45	0,07	0,41	0,00
73	93	7	0,01	0,49	0,01	0,30	0,00	0,00	0,00	0,44	0,41	0,00	0,00	0,00	0,11	0,18	0,00	0,00	0,05	0,98	0,76	0,73	0,10	0,18	0,55	0,00
74	94	5	0,14	0,62	0,32	0,28	0,19	0,00	0,00	0,20	0,49	0,00	0,00	0,19	0,26	0,14	0,46	0,24	0,00	1,00	0,77	0,72	0,25	0,07	0,46	0,00
75	96	2	0,26	0,82	0,52	0,17	0,19	0,00	0,00	0,31	0,54	0,33	0,26	0,32	0,37	0,10	0,68	0,49	1,00	0,82	0,78	0,72	0,47	0,09	0,48	0,00
76	98	2	0,24	0,68	0,32	0,12	0,17	0,00	0,00	0,48	0,49	0,00	0,13	0,28	0,19	0,23	0,80	0,75	0,05	0,98	0,80	0,75	0,28	0,07	0,40	0,00
77	100	2	0,04	0,15	0,00	0,16	0,00	0,00	0,00	0,65	0,08	0,00	0,00	0,00	0,94	0,39	0,80	0,90	0,05	0,98	0,81	0,77	0,03	0,06	0,17	0,00
78	102	9	0,00	0,39	0,00	1,00	0,00	0,00	0,00	0,69	0,33	0,17	0,00	0,04	1,00	0,39	0,84	0,82	0,05	0,98	0,81	0,78	0,08	0,09	0,49	0,00
79	104	7	0,11	0,91	0,72	0,21	0,00	0,00	0,00	0,40	0,56	0,00	0,30	0,00	0,03	0,05	0,78	0,22	0,09	0,91	0,84	0,83	0,83	0,20	0,55	0,00
80	106	8	0,01	0,42	0,00	0,89	0,00	0,00	0,00	0,99	0,30	0,00	0,00	0,04	0,49	0,18	0,37	0,35	0,00	1,00	0,85	0,80	0,10	0,12	0,49	0,00
81	108	2	0,10	0,65	0,21	0,26	0,06	0,00	0,00	0,34	0,48	0,00	0,09	0,21	0,29	0,33	0,77	0,76	0,05	0,98	0,87	0,86	0,28	0,09	0,46	0,00
82	110	2	0,16	0,72	0,32	0,23	0,19	0,00	0,00	0,27	0,49	0,00	0,09	0,34	0,26	0,27	0,45	0,05	0,00	1,00	0,87	0,83	0,34	0,08	0,46	0,00
83	112	10	0,25	0,75	0,43	0,34	0,51	0,00	0,00	0,21	0,59	1,00	0,00	0,21	0,08	0,18	0,11	0,15	0,05	0,97	0,93	0,92	0,29	0,08	0,48	0,00
84	113	2	0,25	0,60	0,26	0,16	0,19	0,00	0,00	0,08	0,46	0,00	0,04	0,34	0,08	0,07	0,80	0,27	0,18	0,90	0,94	0,93	0,19	0,05	0,40	0,00
85	115	2	0,96	0,99	0,82	0,36	0,19	0,00	0,00	0,02	0,47	0,00	0,87	0,34	0,17	0,24	0,05	0,05	0,18	0,90	0,95	0,94	0,83	0,08	0,40	0,00
86	117	1	0,23	0,59	0,22	0,16	0,19	0,00	0,00	0,15	0,47	0,00	0,00	0,34	0,12	0,26	0,09	0,10	0,09	0,95	0,95	0,94	0,18	0,05	0,40	0,00
87	118	2	0,22	0,56	0,22	0,16	0,19	0,00	0,00	0,04	0,46	0,00	0,00	0,34	0,43	0,23	0,80	0,28	0,55	0,65	0,96	0,94	0,16	0,04	0,40	0,00
88	119	1	0,29	0,68	0,33	0,16	0,30	0,00	0,00	0,25	0,48	0,00	0,04	0,45	0,26	0,13	0,67	0,41	0,68	0,71	0,98	0,98	0,25	0,05	0,41	0,00
89	120	2	0,28	0,93	0,91	0,17	0,40	0,00	0,00	0,25	0,53	0,00	0,17	0,68	0,11	0,25	0,07	0,07	0,22	0,90	0,98	0,98	0,77	0,10	0,60	0,00
90	121	2	0,14	0,82	0,53	0,19	0,19	0,00	0,00	0,27	0,53	0,00	0,17	0,34	0,20	0,19	0,50	0,04	0,00	1,00	1,00	0,99	0,56	0,12	0,52	0,00
91	124	8	0,06	0,36	0,01	0,34	0,00	0,00	0,00	0,05	0,37	0,00	0,00	0,00	0,35	0,26	0,61	0,24	0,41	0,80	1,00	1,00	0,08	0,08	0,36	0,00

92	125	6	0,21	0,82	0,52	0,32	0,49	0,00	0,00	0,36	0,58	1,00	0,04	0,36	0,41	0,31	0,62	0,28	0,41	0,80	1,00	1,00	0,41	0,08	0,60	0,00
			ALVA_Descriptors					MOLD2_Descriptors (MLD2)							TEST_Descriptors (TST)								PADEL_Descriptors (PDL)			
Test	ID	Class	S3K	piPC01	ATSC7i	MATS6p	CATS2 D_08_L	D130	D338	D489	D590	D606	D679	D715	xvp9	xvc3	Gmin	Qsv	ic	ssi	piPC 04	piPC 05	AT S2 m	AATS3m	Sp M ax 1_ B hi	Si
1	5	6	0,07	0,71	0,39	0,35	0,37	0,00	0,72	0,53	0,69	0,00	0,00	0,20	0,36	0,26	0,93	0,71	0,01	0,97	0,83	0,81	0,40	0,09	0,60	0,40
2	10	1	0,12	0,38	0,06	0,33	0,02	0,00	0,00	0,32	0,53	0,00	0,00	0,14	0,04	0,01	0,81	0,32	0,00	1,00	0,39	0,33	0,09	0,04	0,20	0,40
3	25	3	0,00	0,80	0,09	0,72	0,20	1,00	1,00	1,00	1,00	0,00	0,10	0,00	0,66	0,28	0,59	0,16	0,10	0,76	1,00	1,00	1,00	1,00	1,00	0,90
4	28	3	0,00	0,72	0,09	0,94	0,19	0,13	0,44	0,48	0,81	0,00	0,00	0,00	0,08	0,07	0,84	0,09	0,01	0,96	0,89	0,88	0,32	0,27	0,60	0,90
5	30	3	0,01	0,76	0,15	0,86	0,20	0,00	0,51	0,49	0,89	0,00	0,00	0,00	0,13	0,14	0,00	0,06	0,01	0,97	0,92	0,91	0,41	0,26	0,60	1,00
6	34	8	0,08	0,26	0,02	1,00	0,00	0,00	0,00	0,06	0,40	0,00	0,20	0,00	0,00	0,00	0,97	0,27	0,00	1,00	0,30	0,24	0,05	0,07	0,00	0,40
7	35	1	0,32	0,82	0,67	0,34	1,00	0,00	0,00	0,34	0,72	0,50	0,20	0,43	0,84	0,13	0,94	0,56	0,02	0,96	0,77	0,71	0,54	0,06	0,55	0,00
8	36	4	0,11	0,38	0,10	0,33	0,00	0,00	0,00	0,26	0,41	0,00	0,00	0,00	0,00	1,00	0,56	1,00	0,12	0,00	0,50	0,29	0,27	0,15	0,17	0,00
9	39	1	0,25	0,60	0,17	0,29	0,07	0,00	0,00	0,57	0,63	0,00	0,20	0,25	0,10	0,01	0,89	0,38	0,14	0,51	0,55	0,51	0,24	0,06	0,26	0,00
10	44	8	0,04	0,13	0,00	0,91	0,00	0,00	0,00	0,11	0,15	0,00	0,10	0,00	0,00	0,00	0,98	0,26	0,00	1,00	0,18	0,10	0,02	0,05	0,00	0,00
11	55	1	0,38	0,67	0,28	0,35	0,26	0,00	0,00	0,29	0,65	0,50	0,10	0,36	0,23	0,06	0,95	0,64	0,01	0,97	0,59	0,53	0,29	0,04	0,31	0,00
12	56	2	0,01	0,00	0,00	0,35	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,03	0,91	0,39	0,01	0,44	0,00	0,00	0,00	0,00	0,00	0,00
13	58	1	0,26	0,60	0,19	0,36	0,20	0,00	0,00	0,40	0,70	1,00	0,10	0,21	0,13	0,04	0,94	0,54	0,01	0,96	0,59	0,52	0,21	0,05	0,31	0,00
14	62	1	0,51	0,73	0,39	0,35	0,43	0,00	0,00	0,21	0,66	0,00	0,10	0,50	0,45	0,10	0,95	0,73	0,02	0,96	0,61	0,55	0,39	0,04	0,31	0,00
15	64	2	0,19	0,45	0,11	0,33	0,06	0,00	0,00	0,13	0,56	0,00	0,10	0,18	0,11	0,00	0,98	0,56	0,00	1,00	0,43	0,38	0,13	0,04	0,28	0,00
16	70	1	0,19	0,53	0,15	0,34	0,11	0,00	0,00	0,49	0,61	0,00	0,10	0,21	0,15	0,02	0,74	0,43	0,00	1,00	0,49	0,43	0,19	0,06	0,29	0,00
17	76	8	0,05	0,56	0,11	0,34	0,09	0,00	0,00	0,39	0,63	0,00	0,10	0,11	0,06	0,02	0,86	0,27	0,00	1,00	0,73	0,70	0,18	0,09	0,47	0,00
18	92	2	1,00	1,00	1,00	0,31	0,85	0,00	0,00	0,41	0,71	0,00	0,40	1,00	1,00	0,31	0,70	0,57	1,00	0,28	0,78	0,73	0,98	0,05	0,36	0,00
19	95	2	0,21	0,73	0,32	0,39	0,06	0,00	0,00	0,35	0,64	0,00	0,60	0,18	0,21	0,05	0,83	0,25	0,00	1,00	0,69	0,63	0,43	0,09	0,39	0,00
20	97	2	0,27	0,53	0,18	0,32	0,15	0,00	0,00	0,19	0,60	0,00	0,10	0,27	0,20	0,00	1,00	0,62	0,00	1,00	0,49	0,43	0,18	0,04	0,29	0,00
21	99	7	0,00	0,48	0,05	0,00	0,00	0,00	0,00	0,82	0,55	0,00	0,10	0,02	0,01	0,02	0,88	0,16	0,02	0,81	0,70	0,70	0,13	0,14	0,47	0,00
22	101	2	0,24	0,62	0,19	0,36	0,07	0,00	0,00	0,58	0,62	0,00	0,20	0,25	0,12	0,03	0,87	0,41	0,00	1,00	0,58	0,51	0,26	0,06	0,28	0,00

23	107	2	0,16	0,67	0,25	0,36	0,17	0,00	0,00	0,30	0,66	0,00	0,20	0,25	0,25	0,10	0,71	0,32	0,01	0,97	0,65	0,59	0,36	0,07	0,37	0,
24	109	2	0,17	0,68	0,23	0,51	0,17	0,00	0,00	0,33	0,71	0,50	0,20	0,25	0,19	0,05	0,71	0,29	0,00	1,00	0,65	0,59	0,35	0,08	0,36	0,
25	111	2	0,01	0,35	0,01	0,27	0,00	0,00	0,00	0,23	0,40	0,00	0,00	0,00	0,00	0,05	0,60	0,00	0,07	0,38	0,43	0,33		0,12	0,14	0,
26	114	2	0,72	0,81	0,39	0,64	0,17	0,00	0,00	0,04	0,62	0,00	1,00	0,29	0,32	0,00	0,95	0,41	0,00	1,00	0,64		0,58	0,52	0,07	0,
27	116	2	0,34	0,59	0,22	0,40	0,17	0,00	0,00	0,08	0,61	0,00	0,20	0,29	0,24	0,00	0,97	0,58	0,00	1,00	0,52	0,46	0,23	0,04	0,30	0,
28	122	2	0,10	0,48	0,06	0,88	0,02	0,00	0,00	0,78	0,67	0,00	0,30	0,00	0,00	0,04	0,79	0,09	0,06	0,62	0,50	0,47	0,16	0,14	0,19	0,
29	123	1	0,17	0,53	0,16	0,34	0,15	0,00	0,00	0,91	0,61	0,00	0,10	0,20	0,14	0,15	0,87	0,61	0,03	0,88	0,49	0,43	0,20	0,06	0,30	0,

Appendix 8.4_TableS4.Mpredicted values for comedogenicity (6 pages)

ID	Chemical Name	Class	Training/Test comedogenicity	Experimental value (ACTIVE-INACTIVE)	M1pred_alvaDesc_optimized (with 3 descriptor)	Error	M1pred_alvaDesc_first (with 5descriptor)	Error	M2pred_Mold2	Error	M3pred_TEST	Error	M4pred_PADEL	Error
1	1-3 butanediol (Butylene glycol)	2	Training Set	ACTIVE	1:ACTIVE		1:ACTIVE		1:ACTIVE		1:ACTIVE		1:ACTIVE	
2	Acetone	8	Training Set	INACTIVE	2:INACTIVE		2:INACTIVE		2:INACTIVE		2:INACTIVE		2:INACTIVE	
3	Allantoin	7	Training Set	INACTIVE	2:INACTIVE		2:INACTIVE		2:INACTIVE		2:INACTIVE		2:INACTIVE	
4	Alpha tocopherol acetate	6	Training Set	INACTIVE	2:INACTIVE		2:INACTIVE		2:INACTIVE		2:INACTIVE		2:INACTIVE	
6	Ascorbyl palmitate	6	Training Set	ACTIVE	1:ACTIVE		1:ACTIVE		1:ACTIVE		1:ACTIVE		1:ACTIVE	
7	Behenic acid	1	Training Set	INACTIVE	2:INACTIVE		2:INACTIVE		2:INACTIVE		2:INACTIVE		2:INACTIVE	
8	Behenyl erucate	1	Training Set	INACTIVE	2:INACTIVE		2:INACTIVE		2:INACTIVE		2:INACTIVE		2:INACTIVE	
9	Butyl stearate	1	Training Set	ACTIVE	1:ACTIVE		1:ACTIVE		1:ACTIVE		1:ACTIVE		1:ACTIVE	
11	Caprylic acid (Octanoic acid)	1	Training Set	ACTIVE	1:ACTIVE		1:ACTIVE		1:ACTIVE		1:ACTIVE		1:ACTIVE	
12	Carmine	3	Training Set	INACTIVE	2:INACTIVE		2:INACTIVE		2:INACTIVE		2:INACTIVE		2:INACTIVE	
13	Castor oil	10	Training Set	ACTIVE	1:ACTIVE		1:ACTIVE		1:ACTIVE		1:ACTIVE		1:ACTIVE	
14	Cetyl acetate	1	Training Set	ACTIVE	1:ACTIVE		1:ACTIVE		1:ACTIVE		1:ACTIVE		1:ACTIVE	
15	Cetyl alcohol	2	Training Set	ACTIVE	1:ACTIVE		1:ACTIVE		1:ACTIVE		1:ACTIVE		1:ACTIVE	
16	Cetyl palmitate	1	Training Set	INACTIVE	2:INACTIVE		2:INACTIVE		2:INACTIVE		2:INACTIVE		2:INACTIVE	
17	Cholesterol	5	Training Set	INACTIVE	2:INACTIVE		2:INACTIVE		2:INACTIVE		2:INACTIVE		2:INACTIVE	
18	Choleth 24	5	Training Set	INACTIVE	2:INACTIVE		2:INACTIVE		2:INACTIVE		2:INACTIVE		2:INACTIVE	
19	Chondroitin sulfate	7	Training Set	INACTIVE	2:INACTIVE		2:INACTIVE		2:INACTIVE		2:INACTIVE		2:INACTIVE	
20	Cyclomethicone	4	Training Set	INACTIVE	2:INACTIVE		2:INACTIVE		2:INACTIVE		2:INACTIVE		2:INACTIVE	

21	D&C red 21	3	Training Set	ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE
22	D&C Red 40 (Allura Red)	3	Training Set	ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE
23	D&C red17	3	Training Set	ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE
24	D&C red19	3	Training Set	ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE
26	D&C red30	3	Training Set	ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE
27	D&C red33	3	Training Set	ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE
29	D&C red4	3	Training Set	ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE
31	D&C red9	3	Training Set	ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE
32	Decyl oleate	1	Training Set	ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE
33	Di (2 ethylhexyl) succinate	1	Training Set	ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE
37	Diisopropyl adipate	1	Training Set	INACTIVE	2:INACTIVE	2:INACTIVE	2:INACTIVE	2:INACTIVE	2:INACTIVE
38	Dioctyl malate	1	Training Set	ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE
40	Eiosanoic acid	1	Training Set	ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE
41	Erythrosine(D&C red 3)	3	Training Set	ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE
42	EthanolSD40	2	Training Set	INACTIVE	2:INACTIVE	2:INACTIVE	2:INACTIVE	2:INACTIVE	2:INACTIVE
43	Ether ya da ethyl ether Ethylene glycol	8	Training Set	INACTIVE	2:INACTIVE	2:INACTIVE	2:INACTIVE	2:INACTIVE	2:INACTIVE
45	monostearate	2	Training Set	INACTIVE	2:INACTIVE	2:INACTIVE	2:INACTIVE	2:INACTIVE	2:INACTIVE
46	Ethylhexyl palmitate	1	Training Set	ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE
47	Ethylhexyl pelargonate	1	Training Set	ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE
48	Glucose glutamate	2	Training Set	INACTIVE	2:INACTIVE	2:INACTIVE	2:INACTIVE	2:INACTIVE	2:INACTIVE
49	Glycereth-26	2	Training Set	INACTIVE	2:INACTIVE	2:INACTIVE	2:INACTIVE	2:INACTIVE	2:INACTIVE
50	Glycerin	2	Training Set	INACTIVE	2:INACTIVE	2:INACTIVE	2:INACTIVE	2:INACTIVE	2:INACTIVE
51	Glyceryl stearate NSE	2	Training Set	ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE
52	Hexylene glycol	2	Training Set	ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE

53	Hydantoin	7	Training Set	INACTIVE	2:INACTIVE	2:INACTIVE	2:INACTIVE	2:INACTIVE	2:INACTIVE
54	Isocetyl alcohol	2	Training Set	ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE
57	Isopropyl isostearate	1	Training Set	ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE
59	Isopropyl myristate	1	Training Set	ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE
60	Isopropyl neopentanoate	1	Training Set	ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE
61	Isopropyl palmitate	1	Training Set	ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE
63	Laureth 23 (Polyethylene (23) lauryl ether)	2	Training Set	ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE
65	Lauric acid (Dodecanoic acid)	1	Training Set	ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE
66	Lithium stearate	8	Training Set	ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE
67	Methyl paraben	7	Training Set	INACTIVE	2:INACTIVE	2:INACTIVE	2:INACTIVE	2:INACTIVE	2:INACTIVE
68	Myristic acid (Tetradecanoic acid)	1	Training Set	ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE
69	Myristyl alcohol	2	Training Set	ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE
71	Myristyl myristate	1	Training Set	ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE
74	Octyl dimethyl PABA	8	Training Set	INACTIVE	2:INACTIVE	2:INACTIVE	2:INACTIVE	2:INACTIVE	2:INACTIVE
75	Octyl methoxycinnamate	8	Training Set	INACTIVE	2:INACTIVE	2:INACTIVE	2:INACTIVE	2:INACTIVE	2:INACTIVE
77	Octyldodecyl stearate	1	Training Set	INACTIVE	2:INACTIVE	2:INACTIVE	2:INACTIVE	2:INACTIVE	2:INACTIVE
78	Octyldodecyl stearoyl stearate	1	Training Set	INACTIVE	2:INACTIVE	2:INACTIVE	2:INACTIVE	2:INACTIVE	2:INACTIVE
79	Oleth-10	2	Training Set	ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE
80	Oleth-20	2	Training Set	ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE
81	Oleth-3	2	Training Set	ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE

82	Oleth-3 phosphate	2	Training Set	ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE
83	Oleth-5	2	Training Set	ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE
84	Oleyl alcohol	2	Training Set	ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE
85	Oxybenzone	8	Training Set	INACTIVE	2:INACTIVE	2:INACTIVE	2:INACTIVE	2:INACTIVE	2:INACTIVE
86	Palmitic acid (Hexadecanoic acid)	1	Training Set	ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE
87	Panthenol	6	Training Set	INACTIVE	2:INACTIVE	2:INACTIVE	2:INACTIVE	2:INACTIVE	2:INACTIVE
88	Papain	6	Training Set	INACTIVE	2:INACTIVE	2:INACTIVE	2:INACTIVE	2:INACTIVE	2:INACTIVE
89	PEG 400	2	Training Set	ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE
90	PEG 200 (600) dilaurate	2	Training Set	ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE
91	PEG 8 stearate	2	Training Set	ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE
93	Phenoxyethyl paraben (4-hydroxybenzoic acid)	7	Training Set	INACTIVE	2:INACTIVE	2:INACTIVE	2:INACTIVE	2:INACTIVE	2:INACTIVE
94	Phytantriol	5	Training Set	ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE
96	Polysorbate 80	2	Training Set	INACTIVE	2:INACTIVE	2:INACTIVE	2:INACTIVE	2:INACTIVE	2:INACTIVE
98	PPG 2 myristyl ether propionate	2	Training Set	ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE
100	Propylene glycol	2	Training Set	INACTIVE	2:INACTIVE	2:INACTIVE	2:INACTIVE	2:INACTIVE	2:INACTIVE
102	PVP	9	Training Set	INACTIVE	2:INACTIVE	2:INACTIVE	2:INACTIVE	2:INACTIVE	2:INACTIVE
104	Sodium hyaluronate	7	Training Set	INACTIVE	2:INACTIVE	2:INACTIVE	2:INACTIVE	2:INACTIVE	2:INACTIVE
106	Sodium PCA (without sodium)	8	Training Set	INACTIVE	2:INACTIVE	2:INACTIVE	2:INACTIVE	2:INACTIVE	2:INACTIVE
108	Sorbitan laurate	2	Training Set	ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE
110	Sorbitan stearate	2	Training Set	INACTIVE	2:INACTIVE	2:INACTIVE	2:INACTIVE	2:INACTIVE	2:INACTIVE
112	Squalene	10	Training Set	ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE
113	Stearath-20	2	Training Set	ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE
115	Steareth-100	2	Training Set	INACTIVE	2:INACTIVE	2:INACTIVE	2:INACTIVE	2:INACTIVE	2:INACTIVE
117	Stearic acid	1	Training Set	ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE
118	Stearyl alcohol	2	Training Set	ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE
119	Stearyl heptanoate	1	Training Set	ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE

120	Sucrose distearate	2	Training Set	INACTIVE	2:INACTIVE	2:INACTIVE	2:INACTIVE	2:INACTIVE	2:INACTIVE
121	Sucrose stearate	2	Training Set	INACTIVE	2:INACTIVE	2:INACTIVE	2:INACTIVE	2:INACTIVE	2:INACTIVE
124	Triethanolamine	8	Training Set	ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE
125	Vitamin A palmitate	6	Training Set	ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE

ID	Chemical Name	Class	Training/Test for comedogenicity	Experimental value (ACTIVE-INACTIVE)	M1 _{pred_optimized} Alva (with 3 descriptor)	Error	M1 _{pred_Alva} (with 5 descriptor)	Error	M2 _{pred_Mold2}	Error	M3 _{pred_TEST}	Error	M4 _{pred_PADEL}	Error
5	Alpha tocopherol	6	Test Set	ACTIVE	1:ACTIVE		1:ACTIVE		1:ACTIVE		2:INACTIVE	+	2:INACTIVE	+
10	Capric acid	1	Test Set	ACTIVE	1:ACTIVE		1:ACTIVE		2:INACTIVE	+	1:ACTIVE		1:ACTIVE	
25	D&C red27	3	Test Set	ACTIVE	1:ACTIVE		1:ACTIVE		1:ACTIVE		1:ACTIVE		1:ACTIVE	
28	D&C red36	3	Test Set	ACTIVE	1:ACTIVE		2:INACTIVE	+	1:ACTIVE		1:ACTIVE		2:INACTIVE	+
30	D&C red6	3	Test Set	ACTIVE	1:ACTIVE		1:ACTIVE		1:ACTIVE		1:ACTIVE		2:INACTIVE	+
34	Diethylene Glycol													
34	Monoethyl Ether	8	Test Set	INACTIVE	1:ACTIVE	+	1:ACTIVE	+	2:INACTIVE		2:INACTIVE		2:INACTIVE	
35	Diisopropyl dimerate	1	Test Set	INACTIVE	2:INACTIVE		2:INACTIVE		1:ACTIVE	+	2:INACTIVE		1:ACTIVE	+
36	Dimethicone	4	Test Set	ACTIVE	2:INACTIVE	+	1:ACTIVE		2:INACTIVE	+	2:INACTIVE	+	1:ACTIVE	
39	Dioctyl succinate	1	Test Set	ACTIVE	1:ACTIVE		1:ACTIVE		1:ACTIVE		1:ACTIVE		1:ACTIVE	
44	Ethylene glycol													
44	monoethyl ether	8	Test Set	INACTIVE	2:INACTIVE		2:INACTIVE		2:INACTIVE		2:INACTIVE		2:INACTIVE	
55	Isodecyl oleate	1	Test Set	ACTIVE	1:ACTIVE		1:ACTIVE		1:ACTIVE		1:ACTIVE		1:ACTIVE	
56	Isopropyl alcohol	2	Test Set	INACTIVE	2:INACTIVE		2:INACTIVE		2:INACTIVE		2:INACTIVE		2:INACTIVE	
58	Isopropyl linoleate	1	Test Set	ACTIVE	1:ACTIVE		1:ACTIVE		1:ACTIVE		1:ACTIVE		1:ACTIVE	
62	Isostearyl isostearate	1	Test Set	ACTIVE	1:ACTIVE		1:ACTIVE		1:ACTIVE		2:INACTIVE	+	1:ACTIVE	
64	Laureth 4	2	Test Set	ACTIVE	1:ACTIVE		1:ACTIVE		1:ACTIVE		1:ACTIVE		1:ACTIVE	
70	Myristyl lactate	1	Test Set	ACTIVE	1:ACTIVE		1:ACTIVE		1:ACTIVE		1:ACTIVE		1:ACTIVE	

76	Octyl salicylate	8	Test Set	INACTIVE	1:ACTIVE	+	1:ACTIVE	+	1:ACTIVE	+	1:ACTIVE	+	2:INACTIVE
92	Pentaerythritol tetra isostearate	2	Test Set	ACTIVE	1:ACTIVE		1:ACTIVE		1:ACTIVE		1:ACTIVE		2:INACTIVE +
95	Polysorbate 20	2	Test Set	INACTIVE	1:ACTIVE	+	1:ACTIVE	+	1:ACTIVE	+	1:ACTIVE	+	2:INACTIVE
97	PPG 10 cetyl ether	2	Test Set	ACTIVE	1:ACTIVE		1:ACTIVE		1:ACTIVE		1:ACTIVE		1:ACTIVE
99	Propyl paraben	7	Test Set	INACTIVE	2:INACTIVE		2:INACTIVE		2:INACTIVE		2:INACTIVE		2:INACTIVE
101	Propylene glycol (PG) diperlargonate	2	Test Set	ACTIVE	1:ACTIVE		1:ACTIVE		1:ACTIVE		1:ACTIVE		1:ACTIVE
107	Sorbitan isostearate	2	Test Set	ACTIVE	1:ACTIVE		1:ACTIVE		1:ACTIVE		2:INACTIVE	+	1:ACTIVE
109	Sorbitan oleate	2	Test Set	ACTIVE	1:ACTIVE		1:ACTIVE		1:ACTIVE		2:INACTIVE	+	1:ACTIVE
111	Sorbitol	2	Test Set	INACTIVE	2:INACTIVE		2:INACTIVE		2:INACTIVE		2:INACTIVE		2:INACTIVE
114	Steareth 10	2	Test Set	ACTIVE	1:ACTIVE		1:ACTIVE		1:ACTIVE		2:INACTIVE	+	1:ACTIVE
116	Steareth-2	2	Test Set	ACTIVE	1:ACTIVE		1:ACTIVE		1:ACTIVE		1:ACTIVE		1:ACTIVE
122	Triacetin	2	Test Set	INACTIVE	1:ACTIVE	+	1:ACTIVE	+	1:ACTIVE	+	2:INACTIVE		2:INACTIVE
123	Tridecyl neopentanoate	1	Test Set	INACTIVE	1:ACTIVE	+	1:ACTIVE	+	1:ACTIVE	+	1:ACTIVE	+	1:ACTIVE +

Appendix 8.5_TableS5 (5 pages) A. Irritation and Their Grade

B. Cluster Membership

C. Training and Test Set for Irritation Model

D. Comedogenic Activity and Their Grade

***Fatty acids with C8-22 with their comedogenicity grade**

ID	Chemical Name	Chemical Class	Cluster members	Training / Test Set for comedogenicity modelling	Training / Test Set for irritation modelling	Formula	Comedo. Grade (0-5) (Fulton, 1989)	Comedogenic (ACTIVE-INACTIVE)	Irritancy Grade (0-5) (Fulton, 1989)	Irritation (ACTIVE-INACTIVE)
1	1-3 butanediol (Butylene glycol)	2	cluster 3	Training Set	Training Set	C4 H10 O2	1	ACTIVE	0	INACTIVE
2	Acetone	8	cluster 3	Training Set	Training Set	C3 H6 O	0	INACTIVE	0	INACTIVE
3	Allantoin	7	cluster 3	Training Set	Training Set	C4 H6 N4 O3	0	INACTIVE	0	INACTIVE
4	Alpha tocopherol acetate	6	cluster 5	Training Set	Test Set	C31 H52 O3	0	INACTIVE	0	INACTIVE
5	Alpha tocopherol	6	cluster 5	Test Set	Test Set	C29 H50 O2	3	ACTIVE	0-3*	ACTIVE
6	Ascorbyl palmitate	6	cluster 5	Training Set	Training Set	C22 H38 O7	2	ACTIVE	0	INACTIVE
7	*Behenic acid (C22)	1	cluster 5	Training Set	Training Set	C22 H44 O2	0	INACTIVE	0	INACTIVE
8	Behenyl erucate	1	cluster 1	Training Set	Training Set	C44 H86 O2	0	INACTIVE	0	INACTIVE
9	Butyl stearate	1	cluster 5	Training Set	Training Set	C22 H44 O2	3	ACTIVE	0	INACTIVE
10	*Capric acid (C10)	1	cluster 5	Test Set	Test Set	C10 H20 O2	2	ACTIVE	2	ACTIVE
11	*Caprylic acid (Octanoic acid) (C8)	1	cluster 3	Training Set	Training Set	C8 H16 O2	1	ACTIVE	3	ACTIVE
12	Carmine	3	cluster 6	Training Set	Training Set	C22 H20 O13	0	INACTIVE	0	INACTIVE
13	Castor oil	10	cluster 2	Training Set	Training Set	C57 H104 O9	1	ACTIVE	0	INACTIVE
14	Cetyl acetate	1	cluster 5	Training Set	Training Set	C18 H36 O2	4	ACTIVE	2	ACTIVE
15	Cetyl alcohol	2	cluster 5	Training Set	Training Set	C16 H34 O	2	ACTIVE	2	ACTIVE
16	Cetyl palmitate	1	cluster 1	Training Set	Training Set	C32 H64 O2	0	INACTIVE	0	INACTIVE
17	Cholesterol	5	cluster 5	Training Set	Training Set	C27 H46 O	0	INACTIVE	0	INACTIVE
18	Choleth 24	5	cluster 5	Training Set	Training Set	C29 H50 O2	0	INACTIVE	0	INACTIVE

19	Chondroitin sulfate	7	cluster 4	Training Set	Training Set	C13 H21 N O15 S	0	INACTIVE	0	INACTIVE
20	Cyclomethicone	4	cluster 3	Training Set	Training Set	C8 H24 O4 Si4	0	INACTIVE	0	INACTIVE
21	D&C red 21	3	cluster 6	Training Set	Training Set	C20 H8 O5 Br4	2	ACTIVE	0	INACTIVE
22	D&C Red 40 (Allura Red)	3	cluster 2	Training Set	Training Set	C18 H14 N2 O8 S2 Na2	2	ACTIVE	2	ACTIVE
23	D&C red17	3	cluster 6	Training Set	Training Set	C22 H16 N4 O	3	ACTIVE	0	INACTIVE
24	D&C red19	3	cluster 6	Training Set	Training Set	C28 H31 N2 O3 Cl	2	ACTIVE	0	INACTIVE
25	D&C red27	3	cluster 6	Test Set	Test Set	C20 H4 O5 Br4 Cl4	2	ACTIVE	0	INACTIVE
26	D&C red30	3	cluster 6	Training Set	Training Set	C18 H10 O2 Cl2 S2	3	ACTIVE	0	INACTIVE
27	D&C red33	3	cluster 6	Training Set	Training Set	C16 H11 N3 O7 S2Na2	1	ACTIVE	2	ACTIVE
28	D&C red36	3	cluster 6	Test Set	Test Set	C16 H10 N3 O3 Cl	3	ACTIVE	0	INACTIVE
29	D&C red4	3	cluster 6	Training Set	Test Set	C18 H16 N2 O7 S2	2	ACTIVE	1	ACTIVE
30	D&C red6	3	cluster 6	Test Set	Training Set	C18H12N2Na2O6S	1	ACTIVE	0	INACTIVE
31	D&C red9	3	cluster 6	Training Set	Training Set	Ba(C17H12ClN2O4S)2	1	ACTIVE	0	INACTIVE
32	Decyl oleate	1	cluster 1	Training Set	Training Set	C28 H54 O2	3	ACTIVE	0	INACTIVE
33	Di (2 ethylhexyl) succinate	1	cluster 5	Training Set	Training Set	C20 H38 O4	2	ACTIVE	0	INACTIVE
34	Diethylene Glycol Monoethyl Ether	8	cluster 3	Test Set	Training Set	C6 H14 O3	0	INACTIVE	0	INACTIVE
35	Diisopropyl dimerate	1	cluster 1	Test Set	Training Set	C42 H76 O4	0	INACTIVE	0	INACTIVE
36	Dimethicone	4	cluster 3	Test Set	Test Set	C8 H24 O2 Si3	1	ACTIVE	0	INACTIVE
37	Diisopropyl adipate	1	cluster 3	Training Set	Training Set	C12 H22 O4	0	INACTIVE	0	INACTIVE
38	Diethyl malate	1	cluster 5	Training Set	Training Set	C20 H38 O5	3	ACTIVE	1	ACTIVE
39	Diethyl succinate	1	cluster 5	Test Set	Test Set	C20 H38 O4	3	ACTIVE	2	ACTIVE
40	Eiosanoic acid	1	cluster 5	Training Set	Training Set	C20 H40 O2	2	ACTIVE	0	INACTIVE
41	Erythrosine(D&C red 3)	3	cluster 6	Training Set	Training Set	C20H6I4Na2O5	3	ACTIVE	0	INACTIVE
42	EthanolSD40	2	cluster 3	Training Set	Training Set	C2 H6 O	0	INACTIVE	0	INACTIVE
43	Ether ya da ethyl ether	8	cluster 3	Training Set	Test Set	C4 H10 O	0	INACTIVE	0	INACTIVE
44	Ethylene glycol monoethyl ether	8	cluster 3	Test Set	Training Set	C4 H10 O2	0	INACTIVE	0	INACTIVE

45	Ethylene glycol monostearate	2	cluster 5	Training Set	Test Set	C20 H40 O3	0	INACTIVE	0	INACTIVE
46	Ethylhexyl palmitate	1	cluster 5	Training Set	Training Set	C24 H48 O2	4	ACTIVE	0	INACTIVE
47	Ethylhexyl pelargonate	1	cluster 5	Training Set	Training Set	C17 H34 O2	2	ACTIVE	3	ACTIVE
48	Glucose glutamate	2	cluster 4	Training Set	Training Set	C11 H19 N O9	0	INACTIVE	0	INACTIVE
49	Glycereth-26	2	cluster 3	Training Set	Training Set	C9 H20 O6	0	INACTIVE	0	INACTIVE
50	Glycerin	2	cluster 3	Training Set	Training Set	C3 H8 O3	0	INACTIVE	0	INACTIVE
51	Glyceryl stearate NSE	2	cluster 5	Training Set	Test Set	C21 H42 O4	1	ACTIVE	0	INACTIVE
52	Hexylene glycol	2	cluster 3	Training Set	Test Set	C6 H14 O2	2	ACTIVE	0-1	ACTIVE
53	Hydantoin	7	cluster 3	Training Set	Test Set	C3 H4 N2 O2	0	INACTIVE	0	INACTIVE
54	Isocetyl alcohol	2	cluster 5	Training Set	Training Set	C16 H34 O	4	ACTIVE	4	ACTIVE
55	Isodecyl oleate	1	cluster 5	Test Set	Training Set	C28 H54 O2	3	ACTIVE	1-2	ACTIVE
56	Isopropyl alcohol	2	cluster 3	Test Set	Test Set	C3 H8 O	0	INACTIVE	0	INACTIVE
57	Isopropyl isostearate	1	cluster 5	Training Set	Training Set	C21 H42 O2	5	ACTIVE	0	INACTIVE
58	Isopropyl linoleate	1	cluster 5	Test Set	Training Set	C21 H38 O2	4	ACTIVE	2	ACTIVE
59	Isopropyl myristate	1	cluster 5	Training Set	Training Set	C17 H34 O2	5	ACTIVE	3	ACTIVE
60	Isopropyl neopentanoate	1	cluster 5	Training Set	Test Set	C23 H46 O2	3	ACTIVE	3	ACTIVE
61	Isopropyl palmitate	1	cluster 5	Training Set	Training Set	C19 H38 O2	4	ACTIVE	1	ACTIVE
62	Isostearyl isostearate	1	cluster 1	Test Set	Test Set	C36 H72 O2	4	ACTIVE	1	ACTIVE
63	Laureth 23 (Polyethylene (23) lauryl ether)	2	cluster 2	Training Set	Training Set	C58 H118 O24	3	ACTIVE	0	INACTIVE
64	Laureth 4	2	cluster 5	Test Set	Training Set	C14 H30 O2	5	ACTIVE	4	ACTIVE
65	*Lauric acid (Dodecanoic acid) (C12)	1	cluster 5	Training Set	Training Set	C12 H24 O2	4	ACTIVE	1	ACTIVE
66	Lithium stearate	8	cluster 2	Training Set	Training Set	C18 H35 O2 Li	1	ACTIVE	0	INACTIVE
67	Methyl paraben	7	cluster 3	Training Set	Training Set	C8 H8 O3	0	INACTIVE	0	INACTIVE
68	*Myristic acid (Tetradecanoic acid) (C14)	1	cluster 5	Training Set	Training Set	C14 H28 O2	3	ACTIVE	0	INACTIVE
69	Myristyl alcohol	2	cluster 5	Training Set	Training Set	C14 H30 O	2	ACTIVE	4	ACTIVE
70	Myristyl lactate	1	cluster 5	Test Set	Test Set	C17 H34 O3	4	ACTIVE	2	ACTIVE
71	Myristyl myristate	1	cluster 1	Training Set	Training Set	C28 H56 O2	5	ACTIVE	2	ACTIVE

74	Octyl dimethyl PABA	8	cluster 3	Training Set	Training Set	C17 H27 N O2	0	INACTIVE	0	INACTIVE
75	Octyl methoxycinnamate	8	cluster 3	Training Set	Training Set	C18 H26 O3	0	INACTIVE	0	INACTIVE
76	Octyl salicylate	8	cluster 3	Test Set	Training Set	C15 H22 O3	0	INACTIVE	0	INACTIVE
77	Octyldodecyl stearate	1	cluster 1	Training Set	Test Set	C38 H76 O2	0	INACTIVE	0	INACTIVE
78	Octyldodecyl stearyl stearate	1	cluster 2	Training Set	Training Set	C56 H110 O4	0	INACTIVE	0	INACTIVE
79	Oleth-10	2	cluster 4	Training Set	Test Set	C38 H76 O11	2	ACTIVE	1	ACTIVE
80	Oleth-20	2	cluster 5	Training Set	Training Set	C20 H40 O2	1	ACTIVE	0	INACTIVE
81	Oleth-3	2	cluster 5	Training Set	Training Set	C24 H48 O4	5	ACTIVE	2	ACTIVE
82	Oleth-3 phosphate	2	cluster 5	Training Set	Training Set	C24 H49 O7 P	2	ACTIVE	2	ACTIVE
83	Oleth-5	2	cluster 5	Training Set	Test Set	C28 H56 O6	3	ACTIVE	2	ACTIVE
84	Oleth alcohol	2	cluster 5	Training Set	Training Set	C18 H36 O	4	ACTIVE	2	ACTIVE
85	Oxybenzone	8	cluster 6	Training Set	Training Set	C14 H12 O3	0	INACTIVE	0	INACTIVE
86	*Palmitic acid (Hexadecanoic acid)(C16)	1	cluster 5	Training Set	Test Set	C16 H32 O2	2	ACTIVE	0	INACTIVE
87	Panthenol	6	cluster 3	Training Set	Training Set	C9 H19 N O4	0	INACTIVE	0	INACTIVE
88	Papain	6	cluster 3	Training Set	Training Set	C9 H14 N4 O3	0	INACTIVE	0	INACTIVE
89	PEG 400	2	cluster 3	Training Set	Training Set	C18 H38 O10	1	ACTIVE	0	INACTIVE
90	PEG 200 (600) dilaurate	2	cluster 1	Training Set	Training Set	C28 H52 O5	3	ACTIVE	2	ACTIVE
91	PEG 8 stearate	2	cluster 4	Training Set	Training Set	C34 H68 O10	3	ACTIVE	1	ACTIVE
92	Pentaerythritol tetra isostearate	2	cluster 2	Test Set	Test Set	C77 H148 O8	2	ACTIVE	0	INACTIVE
93	Phenoxyethyl paraben (4-hydroxybenzoic acid)	7	cluster 3	Training Set	Training Set	C7 H6 O3	0	INACTIVE	0	INACTIVE
94	Phytantriol	5	cluster 5	Training Set	Training Set	C20 H42 O3	2	ACTIVE	2	ACTIVE
95	Polysorbate 20	2	cluster 4	Test Set	Training Set	C26 H50 O10	0	INACTIVE	0	INACTIVE
96	Polysorbate 80	2	cluster 4	Training Set	Test Set	C32 H60 O10	0	INACTIVE	0	INACTIVE
97	PPG 10 cetyl ether	2	cluster 5	Test Set	Training Set	C19 H40 O2	3	ACTIVE	1	ACTIVE
98	PPG 2 myristyl ether propionate	2	cluster 5	Training Set	Training Set	C23 H46 O4	3	ACTIVE	2	ACTIVE
99	Propyl paraben	7	cluster 3	Test Set	Test Set	C10 H12 O3	0	INACTIVE	0	INACTIVE
100	Propylene glycol	2	cluster 3	Training Set	Training Set	C3 H8 O2	0	INACTIVE	0	INACTIVE

101	Propylene glycol (PG) diperlargonate	2	cluster 5	Test Set	Test Set	C21 H40 O4	2	ACTIVE	2	ACTIVE
102	PVP	9	cluster 3	Training Set	Training Set	C6 H9 N O	0	INACTIVE	0	INACTIVE
104	Sodium hyaluronate	7	cluster 2	Training Set	Training Set	C28H42N2O23Na2	0	INACTIVE	0	INACTIVE
106	Sodium PCA (without sodium)	8	cluster 3	Training Set	Training Set	C5 H7 N O3	0	INACTIVE	0	INACTIVE
107	Sorbitan isostearate	2	cluster 5	Test Set	Training Set	C24 H46 O6	2	ACTIVE	0	INACTIVE
108	Sorbitan laurate	2	cluster 5	Training Set	Training Set	C18 H34 O6	2	ACTIVE	1-2	ACTIVE
109	Sorbitan oleate	2	cluster 5	Test Set	Test Set	C24 H44 O6	3	ACTIVE	0	INACTIVE
110	Sorbitan stearate	2	cluster 5	Training Set	Training Set	C24 H46 O6	0	INACTIVE	1	ACTIVE
111	Sorbitol	2	cluster 3	Test Set	Test Set	C6 H14 O6	0	INACTIVE	0	INACTIVE
112	Squalene	10	cluster 5	Training Set	Training Set	C30 H50	1	ACTIVE	0	INACTIVE
113	Stearath-20	2	cluster 5	Training Set	Training Set	C20 H42 O2	2	ACTIVE	1	ACTIVE
114	Steareth 10	2	cluster 4	Test Set	Training Set	C38 H78 O11	4	ACTIVE	3	ACTIVE
115	Steareth-100	2	cluster 2	Training Set	Test Set	C58 H118 O21	0	INACTIVE	0	INACTIVE
116	Steareth-2	2	cluster 5	Test Set	Test Set	C22 H46 O3	2	ACTIVE	2	ACTIVE
117	*Stearic acid (C18)	1	cluster 5	Training Set	Training Set	C18 H36 O2	3	ACTIVE	0	INACTIVE
118	Stearyl alcohol	2	cluster 5	Training Set	Training Set	C18 H38 O	2	ACTIVE	2	ACTIVE
119	Stearyl heptanoate	1	cluster 5	Training Set	Training Set	C25 H50 O2	4	ACTIVE	0	INACTIVE
120	Sucrose distearate	2	cluster 2	Training Set	Training Set	C48 H90 O13	0	INACTIVE	2	ACTIVE
121	Sucrose stearate	2	cluster 4	Training Set	Training Set	C30 H56 O12	0	INACTIVE	0	INACTIVE
122	Triacetin	2	cluster 3	Test Set	Training Set	C9 H14 O6	0	INACTIVE	0	INACTIVE
123	Tridecyl neopentanoate	1	cluster 5	Test Set	Training Set	C18 H36 O2	0	INACTIVE	3	ACTIVE
124	Triethanolamine	8	cluster 3	Training Set	Training Set	C6 H15 N O3	2	ACTIVE	0	INACTIVE
125	Vitamin A palmitate	6	cluster 5	Training Set	Training Set	C36 H60 O2	3	ACTIVE	1-3*	ACTIVE

Appendix 8.6_TableS6. Selected Descriptors for Skin Irritation with original value (4 pages)

ID	Class	Irritation (Train/ Test)	Alva Descriptors		PADEL Descriptors (PDL)				MOLD Descriptors (MLD2)								TEST Descriptors (TST)				
			Mi	ATSC7e	AATS0i	naAromAtom	AATS3e	SdssC	D004	D196	D223	D294	D489	D600	D606	D777	xch9	xvc4	BELe1	BEIp1	piPC04
1	2	Training	1,16	0,00	170,45	0,00	7,66	0,00	0	8,94	0	3,45	0,25	0	0	-0,08	0,00	0,00	1,55	1,76	1,10
2	8	Training	1,15	0,00	167,53	0,00	8,29	0,17	0	4,89	0	1,57	0	0	0	0,20	0,00	0,00	1,43	1,65	0,00
3	7	Training	1,18	0,10	177,53	0,00	8,76	-2,13	1	21,90	0,02	3,97	0,47	0	0	-1,54	0,00	0,00	1,44	1,75	3,47
4	6	Test	1,13	0,54	163,98	6,00	7,37	-0,26	0	23,59	0,04	5,26	0,49	2	0	6,51	0,00	0,10	1,96	2,05	5,03
5	6	Test	1,13	0,29	164,12	6,00	7,30	0,00	0	18,87	0,05	5,16	0,45	2	0	6,24	0,00	0,10	1,96	2,05	4,92
6	6	Training	1,14	0,48	165,89	0,00	7,55	-3,21	1	44,15	0,05	4,45	0,29	0	0	1,52	0,00	0,00	1,72	1,86	4,16
7	1	Training	1,14	0,19	166,13	0,00	7,08	-0,65	0	12,44	0,12	3,23	0,11	0	0	5,68	0,00	0,00	1,73	1,87	3,09
8	1	Training	1,14	0,37	165,55	0,00	7,06	0,03	0	16,78	0,11	4,20	0,15	0	2	9,67	0,00	0,00	1,74	1,87	3,91
9	1	Training	1,14	0,22	166,13	0,00	7,13	0,00	0	12,97	0,10	3,97	0,27	0	0	5,68	0,00	0,00	1,72	1,87	3,18
10	1	Test	1,14	0,12	166,79	0,00	7,20	-0,66	0	11,38	0,11	4,23	0,27	0	0	2,68	0,00	0,00	1,69	1,84	2,30
11	1	Training	1,14	0,10	167,07	0,00	7,27	-0,67	0	10,90	0,10	4,06	0,35	0	0	2,06	0,00	0,00	1,67	1,82	2,08
12	3	Training	1,12	2,48	161,79	12,00	8,40	-3,88	0	111,09	0,02	5,62	0,39	0	0	-1,67	0,00	0,00	2,02	2,09	5,80
13	10	Training	1,14	1,28	165,46	0,00	7,30	-1,04	0	74,23	0,07	5,22	0,30	0	6	7,51	0,00	0,00	1,73	1,87	4,55
14	1	Training	1,14	0,17	166,25	0,00	7,16	-0,15	0	10,48	0,11	3,97	0,18	0	0	4,78	0,00	0,00	1,72	1,86	2,89
15	2	Training	1,14	0,11	166,69	0,00	7,03	0,00	0	5,84	0,13	3,75	0,04	0	0	4,51	0,00	0,00	1,72	1,86	2,64
16	1	Training	1,14	0,30	165,95	0,00	7,08	0,03	0	14,68	0,11	3,81	0,19	0	0	7,68	0,00	0,00	1,72	1,87	3,53
17	5	Training	1,13	0,24	163,72	0,00	7,19	1,62	1	14,69	0,05	5,48	0,23	8	0	7,30	0,01	0,22	1,99	2,07	4,68
18	5	Training	1,13	0,32	164,12	0,00	7,24	1,70	1	17,78	0,05	5,63	0,23	8	0	6,82	0,01	0,22	1,99	2,07	4,73
19	7	Training	1,15	2,94	169,25	0,00	8,79	-2,50	0	95,47	0,02	5,50	0,47	0	0	-5,57	0,00	0,04	1,73	1,97	4,43
20	4	Training	1,12	0,04	161,50	0,00	8,14	0,00	0	26,96	0,45	1,57	0,44	0	0	-0,64	0,00	1,00	1,75	1,45	3,71
21	3	Training	1,08	0,89	148,67	20,00	8,33	-0,33	1	39,16	0,02	4,62	0,68	0	0	6,00	0,02	0,03	1,99	2,05	5,81
22	3	Training	1,11	1,01	158,91	16,00	8,14	0,00	0	65,73	0,03	5,23	0,56	0	0	-1,59	0,00	0,08	2,05	2,10	5,73
23	3	Training	1,11	0,13	157,64	18,00	7,68	0,29	0	18,01	0,02	4,99	0,48	0	0	3,53	0,00	0,00	1,94	2,02	5,48
24	3	Training	1,12	0,67	160,34	12,00	7,57	2,70	0	26,07	0,02	5,12	0,30	0	0	4,52	0,00	0,00	2,02	2,09	5,72
25	3	Test	1,07	1,28	146,86	20,00	8,78	-0,19	1	42,25	0,03	4,56	0,85	0	0	7,86	0,01	0,03	1,99	2,04	5,95
26	3	Training	1,08	0,19	148,63	12,00	7,97	2,54	2	26,25	0,04	3,81	0,66	0	0	4,94	0,03	0,00	1,86	1,90	5,56
27	3	Training	1,12	0,91	160,47	12,00	8,16	-2,14	0	67,70	0,03	5,02	0,55	0	0	-1,33	0,00	0,09	1,99	2,05	5,51
28	3	Test	1,11	0,41	158,67	12,00	7,98	0,17	0	26,03	0,02	5,13	0,41	0	0	3,12	0,00	0,00	1,94	2,02	5,28
29	3	Test	1,12	0,88	159,47	12,00	8,04	-1,13	0	64,98	0,03	5,15	0,53	0	0	-0,39	0,00	0,09	1,97	2,04	5,55
30	3	Training	1,11	0,53	158,87	12,00	8,00	-2,78	0	53,42	0,02	5,14	0,42	0	0	1,23	0,00	0,05	1,96	2,03	5,47

31	3	Training	1,11	0,50	157,87	12,00	7,98	-0,06	0	38,68	0,03	5,02	0,49	0	0	2,00	0,00	0,05	1,95	2,02	5,40
32	1	Training	1,14	0,27	165,56	0,00	7,11	0,01	0	15,77	0,10	4,87	0,24	0	2	6,81	0,00	0,00	1,73	1,86	3,53
33	1	Training	1,14	0,53	166,20	0,00	7,36	-0,58	0	27,30	0,06	4,26	0,49	0	0	4,30	0,00	0,00	1,71	1,87	3,43
34	8	Training	1,15	0,05	169,82	0,00	7,92	0,00	0	10,28	0,05	3,88	0,05	0	0	-0,13	0,00	0,00	1,47	1,75	1,79
35	1	Training	1,14	0,49	164,93	0,00	7,23	-0,10	0	32,44	0,07	5,42	0,29	0	2	8,84	0,00	0,00	1,92	2,00	4,55
36	4	Test	1,12	0,04	162,77	0,00	7,83	0,00	0	16,70	0,56	2,18	0,22	0	0	1,39	0,00	1,47	1,74	1,47	3,00
37	1	Training	1,14	0,53	166,62	0,00	7,81	-0,40	0	22,42	0,05	3,12	0,86	0	0	2,31	0,00	0,00	1,64	1,83	3,14
38	1	Training	1,14	0,49	166,51	0,00	7,42	-1,28	0	32,51	0,06	4,68	0,41	0	0	3,51	0,00	0,00	1,67	1,84	3,43
39	1	Test	1,14	0,39	166,20	0,00	7,32	-0,58	0	25,43	0,07	4,29	0,49	0	0	4,30	0,00	0,00	1,67	1,84	3,30
40	1	Training	1,14	0,18	166,19	0,00	7,09	-0,65	0	12,35	0,12	3,37	0,13	0	0	5,24	0,00	0,00	1,73	1,87	3,00
41	3	Training	1,07	0,41	145,39	20,00	8,09	-0,33	1	39,24	0,03	4,62	0,87	0	0	6,40	0,02	0,03	1,99	2,04	5,81
42	3	Training	1,16	0,00	172,06	0,00	7,44	0,00	0	3,14	0	2,42	0	0	0	-0,17	0,00	0,00	1,33	1,59	0,00
43	8	Test	1,15	0,00	169,45	0,00	7,54	0,00	0	2,67	0	2,07	0,12	0	0	0,80	0,00	0,00	1,43	1,69	0,69
44	8	Training	1,16	0,01	170,45	0,00	7,81	0,00	0	6,61	0,07	3,33	0,10	0	0	-0,08	0,00	0,00	1,43	1,71	1,10
45	2	Test	1,14	0,23	166,49	0,00	7,19	-0,17	0	16,67	0,10	3,80	0,24	0	0	4,40	0,00	0,00	1,72	1,87	3,14
46	1	Training	1,14	0,29	166,08	0,00	7,13	0,02	0	14,93	0,09	4,48	0,27	0	0	6,10	0,00	0,00	1,72	1,87	3,33
47	1	Training	1,14	0,23	166,29	0,00	7,24	0,01	0	14,55	0,08	4,67	0,44	0	0	4,54	0,00	0,00	1,69	1,86	3,14
48	2	Training	1,15	1,04	169,71	0,00	8,36	-2,13	0	57,32	0,03	5,02	0,67	0	1	-2,60	0,00	0,00	1,64	1,86	3,40
49	2	Training	1,15	0,70	170,06	0,00	8,14	0,00	0	24,88	0,04	3,91	0,29	0	0	-1,49	0,00	0,00	1,55	1,82	2,71
50	2	Training	1,16	0,00	172,58	0,00	8,19	0,00	0	12,03	0	2,81	0,34	0	0	-1,37	0,00	0,00	1,47	1,73	0,69
51	2	Test	1,14	0,30	166,73	0,00	7,28	-0,28	0	23,44	0,09	4,02	0,27	0	0	3,81	0,00	0,00	1,72	1,87	3,22
52	2	Test	1,15	0,00	169,11	0,00	7,81	0,00	0	11,51	0	3,36	0,49	0	0	0,71	0,00	0,16	1,65	1,83	1,95
53	7	Test	1,17	0,00	173,95	0,00	8,64	-0,66	1	12,14	0,02	3,28	0,96	1	0	-1,20	0,00	0,00	1,32	1,59	2,89
54	2	Training	1,14	0,11	166,69	0,00	7,04	0,00	0	6,55	0,12	3,98	0,06	0	0	4,51	0,00	0,00	1,72	1,87	2,64
55	1	Training	1,14	0,27	165,56	0,00	7,12	0,01	0	16,41	0,09	5,20	0,25	0	2	6,81	0,00	0,00	1,73	1,86	3,53
56	2	Test	1,16	0,00	170,43	0,00	7,66	0,00	0	4,11	0	2,13	0	0	0	0,35	0,00	0,00	1,47	1,69	0,00
57	1	Training	1,14	0,18	166,16	0,00	7,20	-0,03	0	13,26	0,10	4,08	0,30	0	0	5,46	0,00	0,00	1,73	1,87	3,22
58	1	Training	1,14	0,17	164,93	0,00	7,23	-0,05	0	14,76	0,08	4,98	0,34	0	4	5,25	0,00	0,00	1,74	1,86	3,53
59	1	Training	1,14	0,16	166,29	0,00	7,23	-0,04	0	12,31	0,10	4,01	0,36	0	0	4,54	0,00	0,00	1,71	1,86	3,04
60	1	Test	1,14	0,20	166,10	0,00	7,17	-0,08	0	14,06	0,10	4,03	0,63	0	0	5,89	0,00	0,25	1,73	1,87	3,14
61	1	Training	1,14	0,17	166,22	0,00	7,20	-0,03	0	12,44	0,10	3,85	0,32	0	0	5,01	0,00	0,00	1,72	1,86	3,14
62	1	Test	1,14	0,32	165,91	0,00	7,08	0,03	0	16,35	0,11	3,94	0,18	0	0	8,41	0,00	0,00	1,73	1,87	3,64
63	2	Training	1,15	0,75	168,13	0,00	7,87	0,00	0	97,71	0,06	3,06	0,02	0	0	-4,59	0,00	0,00	1,71	1,85	4,37
64	2	Training	1,14	0,14	167,25	0,00	7,18	0,00	0	8,94	0,11	4,46	0,11	0	0	3,12	0,00	0,00	1,71	1,85	2,56
65	1	Training	1,14	0,13	166,59	0,00	7,16	-0,66	0	11,70	0,11	4,12	0,22	0	0	3,25	0,00	0,00	1,70	1,85	2,48
66	8	Training	1,14	0,17	165,91	0,00	7,09	-0,90	0	12,24	0,12	3,54	0,14	0	0	4,78	0,00	0,00	1,72	1,86	2,89
67	7	Training	1,12	0,11	160,53	6,00	7,96	-0,40	0	15,27	0,02	3,58	0,83	0	0	1,47	0,00	0,00	1,84	1,93	4,12
68	1	Training	1,14	0,14	166,45	0,00	7,14	-0,66	0	11,93	0,11	3,92	0,19	0	0	3,79	0,00	0,00	1,71	1,86	2,64
69	2	Training	1,14	0,09	166,85	0,00	7,03	0,00	0	5,76	0,13	3,94	0,05	0	0	4,00	0,00	0,00	1,71	1,86	2,48
70	1	Test	1,14	0,21	166,65	0,00	7,28	-0,51	0	17,14	0,10	4,38	0,42	0	0	3,70	0,00	0,00	1,71	1,86	2,89
71	1	Training	1,14	0,28	166,01	0,00	7,09	0,02	0	14,44	0,11	4,03	0,21	0	0	6,91	0,00	0,00	1,72	1,86	3,40

74	8	Training	1,13	0,22	164,48	6,00	7,34	-0,22	0	14,51	0,04	4,60	0,43	0	0	3,94	0,00	0,00	1,87	1,96	4,40
75	8	Training	1,13	0,29	162,69	6,00	7,38	-0,28	0	16,83	0,04	4,81	0,36	0	2	4,10	0,00	0,00	1,87	1,95	4,42
76	8	Training	1,13	0,18	163,16	6,00	7,38	-0,44	0	19,81	0,04	4,85	0,34	0	0	3,45	0,00	0,00	1,84	1,94	4,32
77	1	Test	1,14	0,41	165,89	0,00	7,07	0,05	0	17,16	0,10	4,15	0,18	0	0	8,76	0,00	0,00	1,75	1,89	3,76
78	1	Training	1,14	0,75	165,78	0,00	7,12	0,07	0	34,90	0,09	4,71	0,23	0	0	10,71	0,00	0,00	1,75	1,90	4,26
79	2	Test	1,14	0,42	167,30	0,00	7,58	0,00	0	46,39	0,06	4,37	0,05	0	2	1,05	0,00	0,00	1,73	1,86	3,91
80	2	Training	1,14	0,17	166,19	0,00	7,14	0,00	0	10,72	0,10	4,91	0,11	0	2	4,42	0,00	0,00	1,73	1,86	3,14
81	2	Training	1,14	0,25	166,59	0,00	7,29	0,00	0	18,32	0,08	5,14	0,09	0	2	3,65	0,00	0,00	1,73	1,86	3,37
82	2	Training	1,14	0,41	166,82	0,00	7,39	0,00	0	32,58	0,08	5,28	0,17	0	2	2,44	0,00	0,04	1,73	1,86	3,53
83	2	Test	1,14	0,31	166,87	0,00	7,40	0,00	0	26,23	0,08	4,95	0,07	0	2	2,90	0,00	0,00	1,73	1,86	3,56
84	2	Training	1,14	0,11	165,91	0,00	7,04	0,00	0	7,38	0,11	4,53	0,08	0	2	4,84	0,00	0,00	1,73	1,86	3,00
85	8	Training	1,11	0,12	156,91	12,00	7,79	-0,20	0	20,57	0,02	4,32	0,39	0	0	3,18	0,00	0,00	1,91	1,99	4,91
86	1	Test	1,14	0,15	166,34	0,00	7,12	-0,65	0	12,10	0,12	3,72	0,16	0	0	4,29	0,00	0,00	1,72	1,86	2,77
87	6	Training	1,15	0,35	169,93	0,00	7,74	-0,52	0	26,31	0,04	4,27	0,77	0	0	-0,46	0,00	0,20	1,70	1,88	2,94
88	6	Training	1,16	0,51	171,04	5,00	7,93	-1,45	1	26,48	0,03	4,64	0,66	0	0	-1,46	0,00	0,00	1,66	1,86	4,10
89	2	Training	1,15	0,36	169,15	0,00	8,09	0,00	0	38,42	0,04	2,69	0,01	0	0	-2,07	0,00	0,00	1,51	1,79	3,22
90	2	Training	1,14	0,41	165,80	0,00	7,31	0,00	0	35,95	0,08	4,63	0,34	4	0	3,98	0,00	0,07	1,87	2,02	3,91
91	2	Training	1,14	0,45	167,32	0,00	7,55	-0,13	0	45,16	0,07	4,09	0,13	0	0	1,87	0,00	0,00	1,72	1,87	3,78
92	2	Test	1,14	1,30	165,73	0,00	7,18	-1,30	0	76,04	0,08	3,97	0,35	0	0	12,01	0,00	0,13	1,73	1,90	4,63
93	7	Training	1,12	0,08	159,59	6,00	8,01	-0,99	0	16,74	0,02	3,50	0,43	0	0	1,13	0,00	0,00	1,84	1,93	4,07
94	5	Training	1,14	0,27	167,06	0,00	7,32	0,00	0	23,57	0,09	4,90	0,19	0	0	3,73	0,00	0,09	1,77	1,90	3,18
95	2	Training	1,15	1,35	167,41	0,00	7,77	-0,21	1	54,77	0,04	5,31	0,30	1	0	-0,06	0,00	0,00	1,70	1,86	4,09
96	2	Test	1,14	1,49	166,74	0,00	7,62	-0,22	1	56,49	0,05	5,68	0,30	1	2	1,41	0,00	0,00	1,73	1,86	4,23
97	2	Training	1,14	0,17	166,83	0,00	7,12	0,00	0	9,63	0,11	4,22	0,16	0	0	4,34	0,00	0,00	1,72	1,86	2,89
98	2	Training	1,14	0,37	166,63	0,00	7,40	-0,18	0	20,96	0,08	4,96	0,47	0	0	4,25	0,00	0,00	1,71	1,86	3,37
99	7	Test	1,12	0,10	161,73	6,00	7,74	-0,35	0	17,14	0,03	3,81	0,70	0	0	2,09	0,00	0,00	1,84	1,93	4,18
100	2	Training	1,16	0,00	171,59	0,00	7,95	0,00	0	8,01	0	3,18	0,62	0	0	-0,53	0,00	0,00	1,47	1,71	0,00
101	2	Test	1,14	0,47	166,17	0,00	7,36	-0,37	0	26,47	0,07	4,51	0,50	0	0	4,52	0,00	0,00	1,68	1,85	3,43
102	9	Training	1,14	0,00	165,98	0,00	7,61	0,21	1	8,32	0,04	3,62	0,66	3	1	0,47	0,00	0,00	1,49	1,71	3,00
104	7	Training	1,15	6,07	168,81	0,00	8,64	-5,28	0	171,78	0,02	6,08	0,39	0	0	-7,05	0,00	0,00	1,76	1,99	5,06
106	8	Training	1,15	0,00	168,50	0,00	8,13	-1,11	1	17,66	0,03	3,75	0,95	2	0	-0,74	0,00	0,00	1,50	1,71	3,00
107	2	Training	1,14	0,59	166,61	0,00	7,48	-0,39	1	39,43	0,06	4,76	0,25	1	0	2,51	0,00	0,00	1,73	1,87	3,87
108	2	Training	1,14	0,43	166,93	0,00	7,63	-0,34	1	37,49	0,05	4,95	0,33	1	0	1,18	0,00	0,00	1,70	1,85	3,64
109	2	Test	1,14	0,47	166,11	0,00	7,50	-0,35	1	39,52	0,06	5,28	0,28	1	2	2,40	0,00	0,00	1,73	1,86	3,87
110	2	Training	1,14	0,48	166,61	0,00	7,47	-0,34	1	38,37	0,06	4,44	0,26	1	0	2,51	0,00	0,00	1,72	1,87	3,78
111	2	Test	1,16	0,46	171,63	0,00	8,36	0,00	0	32,35	0,03	3,55	0,19	0	0	-2,50	0,00	0,00	1,63	1,85	2,56
112	10	Training	1,13	0,14	163,12	0,00	7,12	8,97	0	8,40	0,06	4,03	0,21	0	6	7,93	0,00	0,00	1,80	1,90	4,04
113	2	Training	1,14	0,18	166,77	0,00	7,12	0,00	0	9,23	0,11	3,92	0,08	0	0	4,57	0,00	0,00	1,72	1,87	2,94
114	2	Training	1,15	0,44	167,57	0,00	7,56	0,00	0	44,90	0,07	3,90	0,03	0	0	1,19	0,00	0,00	1,72	1,87	3,83
115	2	Test	1,15	0,57	167,86	0,00	7,74	0,00	0	85,85	0,06	3,30	0,02	0	0	-2,57	0,00	0,00	1,72	1,87	4,33
116	2	Test	1,14	0,23	166,93	0,00	7,20	0,00	0	12,96	0,10	4,19	0,07	0	0	4,17	0,00	0,00	1,72	1,87	3,09

117	1	Training	1,14	0,17	166,25	0,00	7,10	-0,65	0	12,24	0,12	3,54	0,14	0	0	4,78	0,00	0,00	1,72	1,87	2,89
118	2	Training	1,14	0,12	166,57	0,00	7,02	0,00	0	5,91	0,13	3,57	0,04	0	0	4,99	0,00	0,00	1,72	1,87	2,77
119	1	Training	1,14	0,26	166,06	0,00	7,11	0,01	0	13,65	0,10	4,09	0,24	0	0	6,31	0,00	0,00	1,72	1,87	3,30
120	2	Training	1,14	2,50	166,49	0,00	7,57	-1,72	1	113,63	0,05	4,76	0,24	0	0	2,81	0,00	0,07	1,77	1,98	4,92
121	2	Training	1,14	2,04	167,19	0,00	7,81	-0,58	1	83,78	0,04	5,17	0,26	0	0	0,10	0,00	0,03	1,72	1,93	4,47
122	2	Training	1,14	0,57	166,99	0,00	8,78	-1,51	0	28,32	0,03	3,54	0,67	0	0	0,59	0,00	0,00	1,54	1,81	3,00
123	1	Training	1,14	0,17	166,25	0,00	7,21	-0,08	0	13,08	0,09	4,07	0,78	0	0	4,78	0,00	0,25	1,71	1,86	2,89
124	8	Training	1,16	0,19	172,08	0,00	7,68	0,00	0	14,91	0,05	2,64	0,05	0	0	-0,94	0,00	0,00	1,59	1,84	2,30
125	6	Training	1,13	0,31	163,58	0,00	7,17	5,30	0	18,41	0,06	5,06	0,35	3	6	8,47	0,00	0,18	1,90	1,98	4,64

Appendix 8.7_TableS7.Mpredicted values for irritation (5 pages)

ID	Chemical name	Class	Train/Test for skin irritation	Experimental value (Active- Inactive)	M ₁ pred._Alva (with 2 descp)	Error	M ₂ pred_Mold2 (with 8 descp)	Error	M ₃ pred_TEST (5 descp)	Error	M ₄ pred_PaDEL (4 descp)	Error
1	1-3 butanediol (Butylene glycol)	2	Training	INACTIVE	2:INACTIVE		2:INACTIVE		2:INACTIVE		2:INACTIVE	
2	Acetone	8	Training	INACTIVE	2:INACTIVE		2:INACTIVE		2:INACTIVE		2:INACTIVE	
3	Allantoin	7	Training	INACTIVE	2:INACTIVE		2:INACTIVE		2:INACTIVE		2:INACTIVE	
6	Ascorbyl palmitate	6	Training	INACTIVE	2:INACTIVE		2:INACTIVE		2:INACTIVE		2:INACTIVE	
7	*Behenic acid (C22)	1	Training	INACTIVE	2:INACTIVE		2:INACTIVE		2:INACTIVE		2:INACTIVE	
8	Behenyl erucate	1	Training	INACTIVE	2:INACTIVE		2:INACTIVE		2:INACTIVE		2:INACTIVE	
9	Butyl stearate	1	Training	INACTIVE	2:INACTIVE		2:INACTIVE		2:INACTIVE		2:INACTIVE	
11	*Caprylic acid (Octanoic acid) (C8)	1	Training	ACTIVE	1:ACTIVE		1:ACTIVE		1:ACTIVE		1:ACTIVE	
12	Carmine	3	Training	INACTIVE	2:INACTIVE		2:INACTIVE		2:INACTIVE		2:INACTIVE	
13	Castor oil	10	Training	INACTIVE	2:INACTIVE		2:INACTIVE		2:INACTIVE		2:INACTIVE	
14	Cetyl acetate	1	Training	ACTIVE	1:ACTIVE		1:ACTIVE		1:ACTIVE		1:ACTIVE	
15	Cetyl alcohol	2	Training	ACTIVE	1:ACTIVE		1:ACTIVE		1:ACTIVE		1:ACTIVE	
16	Cetyl palmitate	1	Training	INACTIVE	2:INACTIVE		2:INACTIVE		2:INACTIVE		2:INACTIVE	
17	Cholesterol	5	Training	INACTIVE	2:INACTIVE		2:INACTIVE		2:INACTIVE		2:INACTIVE	
18	Cholet 24	5	Training	INACTIVE	2:INACTIVE		2:INACTIVE		2:INACTIVE		2:INACTIVE	
19	Chondroitin sulfate	7	Training	INACTIVE	2:INACTIVE		2:INACTIVE		2:INACTIVE		2:INACTIVE	
20	Cyclomethicone	4	Training	INACTIVE	2:INACTIVE		2:INACTIVE		2:INACTIVE		2:INACTIVE	
21	D&C red 21	3	Training	INACTIVE	2:INACTIVE		2:INACTIVE		2:INACTIVE		2:INACTIVE	
22	D&C Red 40 (Allura Red)	3	Training	ACTIVE	1:ACTIVE		1:ACTIVE		1:ACTIVE		1:ACTIVE	
23	D&C red17	3	Training	INACTIVE	2:INACTIVE		2:INACTIVE		2:INACTIVE		2:INACTIVE	
24	D&C red19	3	Training	INACTIVE	2:INACTIVE		2:INACTIVE		2:INACTIVE		2:INACTIVE	
26	D&C red30	3	Training	INACTIVE	2:INACTIVE		2:INACTIVE		2:INACTIVE		2:INACTIVE	
27	D&C red33	3	Training	ACTIVE	1:ACTIVE		1:ACTIVE		1:ACTIVE		1:ACTIVE	
30	D&C red6	3	Training	INACTIVE	2:INACTIVE		2:INACTIVE		2:INACTIVE		2:INACTIVE	

31	D&C red9	3	Training	INACTIVE	2:INACTIVE		2:INACTIVE	2:INACTIVE	2:INACTIVE
32	Decyl oleate	1	Training	INACTIVE	1:ACTIVE	+	2:INACTIVE	2:INACTIVE	2:INACTIVE
33	Di (2 ethylhexyl) succinate	1	Training	INACTIVE	2:INACTIVE		2:INACTIVE	2:INACTIVE	2:INACTIVE
34	Diethylene Glycol Monoethyl Ether	8	Training	INACTIVE	2:INACTIVE		2:INACTIVE	2:INACTIVE	2:INACTIVE
35	Diisopropyl dimerate	1	Training	INACTIVE	2:INACTIVE		2:INACTIVE	2:INACTIVE	2:INACTIVE
37	Diisopropyl adipate	1	Training	INACTIVE	2:INACTIVE		2:INACTIVE	2:INACTIVE	2:INACTIVE
38	Diethyl malate	1	Training	ACTIVE	1:ACTIVE		1:ACTIVE	1:ACTIVE	1:ACTIVE
40	Eiosanoic acid	1	Training	INACTIVE	2:INACTIVE		2:INACTIVE	2:INACTIVE	2:INACTIVE
41	Erythrosine(D&C red 3)	3	Training	INACTIVE	2:INACTIVE		2:INACTIVE	2:INACTIVE	2:INACTIVE
42	EthanolSD40	3	Training	INACTIVE	2:INACTIVE		2:INACTIVE	2:INACTIVE	2:INACTIVE
44	Ethylene glycol monoethyl ether	8	Training	INACTIVE	2:INACTIVE		2:INACTIVE	2:INACTIVE	2:INACTIVE
46	Ethylhexyl palmitate	1	Training	INACTIVE	2:INACTIVE		2:INACTIVE	2:INACTIVE	2:INACTIVE
47	Ethylhexyl pelargonate	1	Training	ACTIVE	1:ACTIVE		1:ACTIVE	1:ACTIVE	1:ACTIVE
48	Glucose glutamate	2	Training	INACTIVE	2:INACTIVE		2:INACTIVE	2:INACTIVE	2:INACTIVE
49	Glycereth-26	2	Training	INACTIVE	2:INACTIVE		2:INACTIVE	2:INACTIVE	2:INACTIVE
50	Glycerin	2	Training	INACTIVE	2:INACTIVE		2:INACTIVE	2:INACTIVE	2:INACTIVE
54	Isocetyl alcohol	2	Training	ACTIVE	1:ACTIVE		1:ACTIVE	1:ACTIVE	1:ACTIVE
55	Isodecyl oleate	1	Training	ACTIVE	1:ACTIVE		1:ACTIVE	1:ACTIVE	1:ACTIVE
57	Isopropyl isostearate	1	Training	INACTIVE	2:INACTIVE		2:INACTIVE	2:INACTIVE	2:INACTIVE
58	Isopropyl linoleate	1	Training	ACTIVE	1:ACTIVE		1:ACTIVE	1:ACTIVE	1:ACTIVE
59	Isopropyl myristate	1	Training	ACTIVE	1:ACTIVE		1:ACTIVE	1:ACTIVE	1:ACTIVE
61	Isopropyl palmitate	1	Training	ACTIVE	1:ACTIVE		1:ACTIVE	1:ACTIVE	1:ACTIVE
	Laureth 23								
63	(Polyethylene (23) lauryl ether)	2	Training	INACTIVE	2:INACTIVE		2:INACTIVE	2:INACTIVE	2:INACTIVE
64	Laureth 4	2	Training	ACTIVE	1:ACTIVE		1:ACTIVE	1:ACTIVE	1:ACTIVE
65	*Lauric acid (Dodecanoic acid) (C12)	1	Training	ACTIVE	1:ACTIVE		1:ACTIVE	1:ACTIVE	1:ACTIVE
66	Lithium stearate	8	Training	INACTIVE	2:INACTIVE		2:INACTIVE	2:INACTIVE	2:INACTIVE
67	Methyl paraben	7	Training	INACTIVE	2:INACTIVE		2:INACTIVE	2:INACTIVE	2:INACTIVE
	*Myristic acid (Tetradecanoic acid) (C14)								
68		1	Training	INACTIVE	2:INACTIVE		2:INACTIVE	2:INACTIVE	2:INACTIVE
69	Myristyl alcohol	2	Training	ACTIVE	1:ACTIVE		1:ACTIVE	1:ACTIVE	1:ACTIVE
71	Myristyl myristate	1	Training	ACTIVE	1:ACTIVE		1:ACTIVE	1:ACTIVE	1:ACTIVE

74	Octyl dimethyl PABA	8	Training	INACTIVE	2:INACTIVE	2:INACTIVE	2:INACTIVE	2:INACTIVE
75	Octyl methoxycinnamate	8	Training	INACTIVE	2:INACTIVE	2:INACTIVE	2:INACTIVE	2:INACTIVE
76	Octyl salicylate	8	Training	INACTIVE	2:INACTIVE	2:INACTIVE	2:INACTIVE	2:INACTIVE
78	Octyldodecyl stearyl stearate	1	Training	INACTIVE	2:INACTIVE	2:INACTIVE	2:INACTIVE	2:INACTIVE
80	Oleth-20	2	Training	INACTIVE	2:INACTIVE	2:INACTIVE	2:INACTIVE	2:INACTIVE
81	Oleth-3	2	Training	ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE
82	Oleth-3 phosphate	2	Training	ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE
84	Oleyl alcohol	2	Training	ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE
85	Oxybenzone	8	Training	INACTIVE	2:INACTIVE	2:INACTIVE	2:INACTIVE	2:INACTIVE
87	Panthenol	6	Training	INACTIVE	2:INACTIVE	2:INACTIVE	2:INACTIVE	2:INACTIVE
88	Papain	6	Training	INACTIVE	2:INACTIVE	2:INACTIVE	2:INACTIVE	2:INACTIVE
89	PEG 400	2	Training	INACTIVE	2:INACTIVE	2:INACTIVE	2:INACTIVE	2:INACTIVE
90	PEG 200 (600) dilaurate	2	Training	ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE
91	PEG 8 stearate	2	Training	ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE
93	Phenoxyethyl paraben	7	Training	INACTIVE	2:INACTIVE	2:INACTIVE	2:INACTIVE	2:INACTIVE
94	Phytantriol	5	Training	ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE
95	Polysorbate 20	2	Training	INACTIVE	2:INACTIVE	2:INACTIVE	2:INACTIVE	2:INACTIVE
97	PPG 10 cetyl ether	2	Training	ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE
98	PPG 2 myristyl ether propionate	2	Training	ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE
100	Propylene glycol	2	Training	INACTIVE	2:INACTIVE	2:INACTIVE	2:INACTIVE	2:INACTIVE
102	PVP	9	Training	INACTIVE	2:INACTIVE	2:INACTIVE	2:INACTIVE	2:INACTIVE
104	Sodium hyaluronate	7	Training	INACTIVE	2:INACTIVE	2:INACTIVE	2:INACTIVE	2:INACTIVE
106	Sodium PCA (without sodium)	8	Training	INACTIVE	2:INACTIVE	2:INACTIVE	2:INACTIVE	2:INACTIVE
107	Sorbitan isostearate	2	Training	INACTIVE	2:INACTIVE	2:INACTIVE	2:INACTIVE	2:INACTIVE
108	Sorbitan laurate	2	Training	ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE
110	Sorbitan stearate	2	Training	ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE
112	Squalene	10	Training	INACTIVE	2:INACTIVE	2:INACTIVE	2:INACTIVE	2:INACTIVE
113	Stearath-20	2	Training	ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE
114	Steareth 10	2	Training	ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE
117	*Stearic acid (C18)	1	Training	INACTIVE	1:ACTIVE	+	2:INACTIVE	2:INACTIVE
118	Stearyl alcohol	2	Training	ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE
119	Stearyl heptanoate	1	Training	INACTIVE	2:INACTIVE	2:INACTIVE	2:INACTIVE	2:INACTIVE

120	Sucrose distearate	2	Training	ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE
121	Sucrose stearate	2	Training	INACTIVE	2:INACTIVE	2:INACTIVE	2:INACTIVE	2:INACTIVE
122	Triacetin	2	Training	INACTIVE	2:INACTIVE	2:INACTIVE	2:INACTIVE	2:INACTIVE
123	Tridecyl neopentanoate	1	Training	ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE
124	Triethanolamine	8	Training	INACTIVE	2:INACTIVE	2:INACTIVE	2:INACTIVE	2:INACTIVE
125	Vitamin A palmitate	6	Training	ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE

ID	Chemical name	Class	Train/Test for skin irritation	Experimental value (Active- Inactive)	M ₁ pred._Alva (with 2 descp)	Error	M ₂ pred_Mold2 (with 8 descp)	Error	M ₃ pred_TEST (5 descp)	Error	M ₄ pred_PaDEL (4 descp)	Error
4	Alpha tocopherol acetate	6	Test	INACTIVE	2:INACTIVE		2:INACTIVE		1:ACTIVE	+	2:INACTIVE	
5	Alpha tocopherol	6	Test	ACTIVE	2:INACTIVE	+	2:INACTIVE	+	1:ACTIVE		1:ACTIVE	
10	*Capric acid (C10)	1	Test	ACTIVE	1:ACTIVE		1:ACTIVE		1:ACTIVE		1:ACTIVE	
25	D&C red27	3	Test	INACTIVE	2:INACTIVE		2:INACTIVE		2:INACTIVE		2:INACTIVE	
28	D&C red36	3	Test	INACTIVE	2:INACTIVE		2:INACTIVE		2:INACTIVE		2:INACTIVE	
29	D&C red4	3	Test	ACTIVE	2:INACTIVE	+	1:ACTIVE		1:ACTIVE		2:INACTIVE	+
36	Dimethicone	4	Test	INACTIVE	2:INACTIVE		2:INACTIVE		2:INACTIVE		2:INACTIVE	
39	Diethyl succinate	1	Test	ACTIVE	2:INACTIVE	+	1:ACTIVE		1:ACTIVE		2:INACTIVE	+
43	Ether ya da ethyl ether	8	Test	INACTIVE	2:INACTIVE		2:INACTIVE		2:INACTIVE		1:ACTIVE	+
45	Ethylene glycol monostearate	2	Test	INACTIVE	1:ACTIVE	+	1:ACTIVE	+	1:ACTIVE	+	1:ACTIVE	+
51	Glyceryl stearate NSE	2	Test	INACTIVE	1:ACTIVE	+	1:ACTIVE	+	2:INACTIVE		1:ACTIVE	+
52	Hexylene glycol	2	Test	ACTIVE	2:INACTIVE	+	2:INACTIVE	+	2:INACTIVE	+	2:INACTIVE	+
53	Hydantoin	7	Test	INACTIVE	2:INACTIVE		2:INACTIVE		2:INACTIVE		2:INACTIVE	
56	Isopropyl alcohol	2	Test	INACTIVE	2:INACTIVE		2:INACTIVE		2:INACTIVE		2:INACTIVE	
60	Isopropyl neopentanoate	1	Test	ACTIVE	2:INACTIVE	+	1:ACTIVE		2:INACTIVE	+	2:INACTIVE	+
62	Isostearyl isostearate	1	Test	ACTIVE	2:INACTIVE	+	2:INACTIVE	+	2:INACTIVE	+	2:INACTIVE	+
70	Myristyl lactate	1	Test	ACTIVE	1:ACTIVE		1:ACTIVE		1:ACTIVE		1:ACTIVE	
77	Octyldodecyl stearate	1	Test	INACTIVE	1:ACTIVE	+	2:INACTIVE		2:INACTIVE		2:INACTIVE	
79	Oleth-10	2	Test	ACTIVE	1:ACTIVE		1:ACTIVE		1:ACTIVE		1:ACTIVE	
83	Oleth-5	2	Test	ACTIVE	1:ACTIVE		1:ACTIVE		1:ACTIVE		1:ACTIVE	
86	*Palmitic acid (Hexadecanoic acid)(C16)	1	Test	INACTIVE	1:ACTIVE	+	2:INACTIVE		1:ACTIVE	+	2:INACTIVE	

92	Pentaerythritol tetra isostearate	2	Test	INACTIVE	2:INACTIVE	2:INACTIVE	1:ACTIVE	+	2:INACTIVE	
96	Polysorbate 80	2	Test	INACTIVE	2:INACTIVE	2:INACTIVE	2:INACTIVE		1:ACTIVE	+
99	Propyl paraben	7	Test	INACTIVE	2:INACTIVE	2:INACTIVE	2:INACTIVE		2:INACTIVE	
101	Propylene glycol (PG) diperlargonate	2	Test	ACTIVE	2:INACTIVE	+	1:ACTIVE	1:ACTIVE	2:INACTIVE	+
109	Sorbitan oleate	2	Test	INACTIVE	2:INACTIVE	2:INACTIVE	1:ACTIVE	+	2:INACTIVE	
111	Sorbitol	2	Test	INACTIVE	2:INACTIVE	2:INACTIVE	2:INACTIVE		2:INACTIVE	
115	Steareth-100	2	Test	INACTIVE	2:INACTIVE	2:INACTIVE	2:INACTIVE		2:INACTIVE	
116	Steareth-2	2	Test	ACTIVE	1:ACTIVE	1:ACTIVE	1:ACTIVE		1:ACTIVE	

9. CURRICULUM VITAE

Kişisel Bilgiler

Adı	SEBLA	Soyadı	ÖZTAN AKTÜRK
Doğum Yeri	-	Doğum Tarihi	-
Uyruğu		TC Kimlik No	-
E-mail		Tel	-

Öğrenim Durumu

Derece	Alan	Mezun Olduğu Kurumun Adı	Mezuniyet Yılı
Doktora	Farmasötik Toksikoloji	Yeditepe Üniversitesi	2023
Yüksek Lisans	Kozmetoloji	Yeditepe Üniversitesi	2013
Lisans	Kimya Mühendisliği	Yıldız Teknik Üniversitesi	2005

Başarılmış birden fazla sınav varsa (KPDS, ÜDS, TOEFL; EELTS vs), tüm sonuçlar yazılmalıdır

Language Skill	Yabancı Dil Sınav Notu (#)
English	ÜDS: 65

Job Experience (Sondan geçmişe doğru sıralayın)

Görevi	Kurum	Süre (Yıl - Yıl)
Kimyevi Madde Deposu Sorumlusu	Yeditepe Üniversitesi	2005-2021
Laboratuvar Asistanlığı (Eczacılık Fak. Araştırma ve Öğrenci)	Yeditepe Üniversitesi	2007-2017

Bilgisayar Bilgisi

Program	Kullanma becerisi
Microsoft Office Programları (Word, Excel, Powerpoint)	Çok iyi
IBM SPSS Statistics	İyi
WEKA makine öğrenmesi ve Veri madenciliği	Çok iyi
Descriptor Hesaplama (PadeL, Mold2, TEST)	İyi
MATLAB/ CHEMCAD/ MATHCAD	İyi/İyi/İyi

*Çok iyi, iyi, orta, zayıf olarak değerlendirin

Bilimsel Çalışmaları**SCI, SSCI, AHCI indekslerine giren dergilerde yayınlanan makaleler**

- Oztan Akturk S., Tugcu G., Sipahi H. Development of a QSAR model to predict comedogenic potential of some cosmetic ingredients. *Computational Toxicology*, Volume 21, February 2022, 100207.

Uluslararası bilimsel toplantılarda sunulan ve bildiri kitabında (Proceedings) basılan bildiriler

- Oztan Akturk S., Tugcu G., Sipahi H. Importance of Descriptor Selection in Classification Models: A Case Study for Comedogenicity. The 10th Annual Meeting of the American Society for Cellular and Computational Toxicology (ASCCT): Pratical Applications of New Tools in Toxicology, Virtual, October 12-14, 2021, Virtual.
- Oztan Akturk S., Tugcu G., Sipahi H. Development of a QSAR model to predict comedogenic potential of some cosmetic ingredients. 19th International Workshop on (Quantitative) Structure-Activity Relationships in Enviromental and Health Sciences, June 7-9, 2021, Virtual.
- Oztan Akturk S., Yagan Uzuner Y., Zindancı I., Turkoglu Z., Kocaturk Goncu E., Kavala M. Effects of Herbal Anti-Acne Creams on Quality of Life. The 22nd European Academy of Dermatology and Venereology (EADV) Congress, October 2-6, 2013, İstanbul, Türkiye.
- Yagan Uzuner Y., Oztan Akturk S., Zindancı I., Turkoglu Z., Kocaturk Goncu E., Kavala M. Herbal Creams For Acne Prevention and Treatment. FIP (International Pharmaceutical Fedaration) World Congress, 31 Aug-5 Sept, 2013, Dublin, Ireland.
- Oztan Aktürk S, Yagan Uzuner Y. Determination of Gallic Acid in Herbal Cosmetic Cream Formulations by HPLC. APGI Skin and Formulation, 4th Symposium, 4-5 June 2012, Lyon, France.

Ulusal Hakemli konferans/sempozyumların bildiri kitaplarında yer alan yayınlar

- Oztan Akturk S., Yagan Uzuner Y., Zindancı I., Turkoglu Z., Kocaturk Goncu E., Kavala M. Bitkisel İçerikli, Anti-Akne Krem Formülasyonlarının %5 Benzoil Peroksit (BP) ile Karşılaştırılması. 24.Ulusal Dermatoloji Kongresi, 9-13 Ekim, 2012, Gaziantep. (Oral-Tartışmalı Poster)
- Yagan Uzuner Y., Zindancı I., Turkoglu Z., Oztan Akturk S. Akneye Karşı Ürün Projesi. Uluslararası Katılımlı 2. Arge Proje Pazarı, İstanbul, 18-19 Eylül, 2012, İstanbul.
- Oztan Akturk S, Yagan Uzuner Y. Organik ve Doğal Kozmetik Formülasyonlarda Bitkisel Aktiflerin Testleri Yapılırken Karşılaşılan Zorluklar. 2.Kozmetik Kimyası Üretimi, Standardizasyonu Kongresi, 17-19 Şubat 2012, Antalya.

Diğer (Görev Aldığı Projeler/Sertifikaları/Ödülleri)

Ödüller
Yagan Uzuner Y, Zindancı I, Turkoglu Z, Oztan Akturk S. Akneye Karşı Ürün Projesi ile Kozmetik Kategorisinde Birincilik Ödülü. Uluslararası Katılımlı 2. Arge Proje Pazarı, İstanbul Kimyevi Maddeler ve Mamulleri İhracatçıları Birliği (İMMİB), 18-19 Eylül, 2012, İstanbul.
Yüksek Lisans Tezi
Sebla Öztan Aktürk. Designing of Anti-Acne Emulsions Containing Herbal Extracts and Assessing Their Efficacy and Unwanted Effects In Volunteers With Mild to Moderate Acne. Yeditepe University, Institute of Health Sciences, The Department of Cosmetology, İstanbul, 2013.

Kurslar
- Deney Hayvanları Kursu, Yeditepe Tıp Fak. YÜDETAM, İstanbul, 2017. - İyi Klinik Uygulamalar (İKU) Temel Eğitim Kursu, Quintiles, İstanbul, 2016 (TC. Sağlık Bakanlığı, Türkiye İlaç ve Tıbbi Cihaz Kurumu onaylı, Yeditepe Eczacılık-Quintiles işbirliğiyle) - Klinik Araştırmalar Eğitim Programı, Pfizer İlaç, (Sağlık Bakanlığı Onaylı), İstanbul, 2009. - İlk Yardım Kursu, Yeditepe Tıp Fak., 2018. - Satın Alma ve Tedarik Yönetim Kursu, Yeditepe Sürekli Eğitim Merkezi, 2009. - Cosmetic Training Course, Yeditepe Faculty of Pharmacy, İstanbul, 2008. - ISO 9001:2000 İç Denetçi Eğitimi, Ntepe Danışmanlık, 2007.
Sertifikalar
6th International BAU Drug Design Congress, 13-15 Aralık, 2018. 1st International Cosmetology and Cosmetic Gynecology Congress, April 28-30, 2011. 9th ICoS (International Cosmetic Symposium), İstanbul, 2011. 2nd HPCI (Home and Personal Care Ingredients Congress), İstanbul, 2011. 5.Uludağ Dermatoloji Günleri, 26 Şubat-01 Mart 2009. - Etkili İletişim, Problem Çözme, Takım Çalışması, Zaman ve Stres Yönetimi, Diksiyon ve Beden Dili, Motivasyon Eğitimleri, Yeditepe Sürekli Eğitim Merkezi, İstanbul, 2010. 3th International Meeting on Pharmacy and Pharmaceutical Sciences, Marmara University, İstanbul, 2010. - Pharmaeconomics in Today and Future, organized by PEPRIC, Yeditepe Faculty of Pharmacy, İstanbul, 2009. - Nanoteknoloji Eğitim Seminerleri, Yeditepe Bilim ve Teknoloji Topluluğu, 2007 1st International Research Conference on Business Strategies, Yeditepe Üniversitesi, Business Administration Department, 2008.
Üyelikler
TÜKAD (Türk Kozmetik Araştırmacıları Derneği) (2009-2012)