LOCALIZATION IN QSAR: THEORY OF LOCAL MODEL DEVELOPMENT, PROGRAMMING AND IMPLEMENTATION OF THE AIQSAR SOFTWARE PACKAGE AND THE APPLICATION OF THE METHODOLOGY TO THE DIFFERENTIAL GENE EXPRESSION PREDICTIONS

Kristijan Vuković

Doctoral Dissertation
Jožef Stefan International Postgraduate School
Ljubljana, Slovenia

Supervisor: Prof. Dr. Emilio Benfenati, Mario Negri Institute of Pharmacological Research, Milan, Italy
Co-Supervisor: Prof. Dr. Marjana Novič, National Institute of Chemistry, Ljubljana, Slovenia

Evaluation Board:
Prof. Dr. Sašo Džeroski, Chair, Jožef Stefan Institute, Ljubljana, Slovenia
Prof. Dr. Janez Mavri, Member, National Institute of Chemistry, Ljubljana, Slovenia
Prof. Dr. Igor Kuzmanovski, Member, Sts. Cyril and Methodius University, Skopje, North Macedonia

MEDNARODNA PODIPLOMSKA ŠOLA JOŽEFA STEFANA
JOŽEF STEFAN INTERNATIONAL POSTGRADUATE SCHOOL



Kristijan Vuković

LOCALIZATION IN QSAR: THEORY OF LOCAL MODEL DEVELOPMENT, PROGRAMMING AND IMPLEMENTATION OF THE AIQSAR SOFTWARE PACKAGE AND THE APPLICATION OF THE METHODOLOGY TO THE DIFFERENTIAL GENE EXPRESSION PREDICTIONS

Doctoral Dissertation

LOKALIZACIJA V QSAR: TEORIJA RAZVOJA LOKALNEGA MODELA, PROGRAMIRANJE IN UVAJANJE PROGRAMSKEGA PAKETA AIQSAR TER UPORABA METODOLOGIJE ZA NAPOVEDOVANJE DIFERENCIALNEGA IZRAŽANJA GENOV

Doktorska disertacija

Supervisor: Prof. Emilio Benfenati

Co-Supervisor: Prof. Marjana Novič

Ljubljana, Slovenia, February 2026

To Noa and Lili

Acknowledgments

I would like to express my sincere gratitude to my supervisors, Emilio and Marjana, for their guidance, support and invaluable expertise throughout my doctoral research. Their mentorship has been instrumental in shaping this work.

This research was supported by the European Union's Marie Skłodowska-Curie Action under grant agreement No. 721975, the "in3" project. I am also grateful to my colleagues from in3, especially Ana and Ivo, for the collaborative environment and helpful discussions.

I want to thank my friends for making this journey so much more enjoyable. To Karlo: thank you for your resolute friendship and for always being there, through every stage of this quest. To Lovro, Tomislav G., Tomislav F., Zlatan and Glorija: thank you for the legendary gaming sessions and the essential breaks that kept me sane.

Finally, my deepest thanks go to my family. To my parents and my brother Viktor: thank you for your enduring love and for lifelong encouragement. Above all, I want to express my profound gratitude to Lucija. Your unwavering patience and support were my greatest strength during the long hours of writing. Thank you for selflessly managing the weight of our family life, thus making this achievement possible.

Abstract

QSAR methods are a cornerstone of contemporary cheminformatics, yet conventional global models remain limited in their ability to provide reliable predictions for complex chemical spaces and endpoints. This dissertation develops, implements and applies a framework for localized QSAR modelling that attempts to address these limitations.

The aiQSAR methodology is established with a complete workflow for QSAR localization that includes molecular descriptor and fingerprint calculation, similarity-based selection of local groups, descriptor filtering and application of multiple machine learning methods for consensus predictions. The Applicability Domain Measure assigns a quantitative reliability score to each target. This approach is validated within a community-wide computational toxicology initiative with external benchmarking and across five acute rat oral toxicity endpoints, demonstrating that aiQSAR methodology achieves predictive performance that is competitive with, and for some challenges superior to, counterpart models.

This methodology is implemented into an open-source software designed to make local QSAR modelling accessible to both expert and non-specialist users. aiQSAR integrates 38 pre-built endpoints through a dual interface with a streamlined Basic mode for routine predictions and an Advanced mode enabling full control over local group selection, similarity parameters, consensus strategies and with a model exporting functionality. Multiple output formats are supported, including concise PDF summaries, detailed text reports suitable for downstream computational analysis and single-compound reports that facilitate expert inspection of local group chemicals and the associated similarity metrics.

The developed methodology is applied within a novel context, for prediction of chemically induced differential gene expression. Transcriptomic profiles of treated samples for a large panel of genes are quantified using the TempO-Seq RNA sequencing platform and associated bioinformatics pipeline. Distinct patterns of differential expression are revealed through exploratory data analysis, indicating that almost all genes exhibit the differential expression following chemical treatment in only a single direction, either upregulation or downregulation, for all modulating chemicals. Finally, aiQSAR modelling results demonstrate that for a large subset of genes, differential expression is predictably encoded in chemical structure and that local QSAR models can capture these relationships.

Taken together, the theoretical development, software implementation and transcriptomics case study demonstrate that localization in QSAR is a robust and versatile strategy. The aiQSAR methodology and software provide an immediately usable platform for chemical risk assessment, drug discovery and mechanistic toxicology, while establishing a foundation for future integration of QSAR with transcriptomics in emerging scientific areas of computational, predictive and developmental biology.

Povzetek

Metode QSAR so temelj sodobne kemijske informatike, vendar konvencionalni globalni modeli ostajajo omejeni v svojih zmožnostih zagotavljanja zanesljivih napovedi za kompleksni prostor kemijskih struktur in končne točke strupenosti. Ta disertacija razvija, implementira in uporablja okvir za lokalizirane QSAR modele, ki skuša odpraviti te omejitve.

Metodologija aiQSAR je vzpostavljena s popolnim delovnim tokom za QSAR lokalizacijo, ki vključuje izračun molekularnih deskriptorjev in prstnih odtisov, izbiro lokalnih skupin na osnovi podobnosti kemijskih struktur, filtriranje deskriptorjev in uporabo več metod strojnega učenja za soglasne napovedi. Mera področja uporabnosti vsaki ciljni molekuli dodeli kvantitativno oceno zanesljivosti. Ta pristop je validiran v okviru pobude na ravni skupnosti za računalniško toksikologijo z zunanjo primerjavo in na petih končnih točkah akutne podganje oralne toksičnosti, pri čemer je prikazano, da metodologija aiQSAR dosega napovedno učinkovitost, ki je konkurenčna in za nekatere izzive boljša od primerjanih modelov.

Metodologija je implementirana v odprtokodno programsko opremo, ki je zasnovana tako, da je lokalno QSAR modeliranje dostopno tako strokovnjakom kot tudi nestrokovnjakom. aiQSAR integrira 38 vnaprej zgrajenih končnih točk preko dvojnega vmesnika s poenostavljenim osnovnim načinom za rutinske napovedi in naprednim načinom, ki omogoča popoln nadzor nad izbiro lokalnih skupin, parametri podobnosti, strategijami soglasja in funkcionalnostjo izvoza (shranjevanja) modela. Podprtih je več oblik rezultatov, vključno s strnjenimi povzetki v obliki PDF, podrobnimi besedilnimi poročili, primernimi za nadaljnjo računalniško analizo, in poročili o eni sami spojini, ki olajšajo strokovno analizo lokalnih skupin kemikalij in povezanih metrik podobnosti.

Razvita metodologija je uporabljena v novem kontekstu, in sicer za napovedovanje kemično inducirane diferencialne ekspresije genov. Transkriptomski profili obdelanih vzorcev za velik niz genov se količinsko določijo s platformo za sekvenciranje RNA TempO-Seq in povezano bioinformacijsko linijo. Raziskovalna analiza podatkov je razkrila izrazite vzorce diferencialne ekspresije, kar kaže, da skoraj vsi geni kažejo diferencialno ekspresijo zaradi kemijske obdelave le v eni smeri, bodisi povečanje bodisi zmanjšanje, pri vseh modulacijskih kemikalijah. Rezultati aiQSAR modeliranja kažejo, da je za večino genov diferencialna ekspresija napovedljivo kodirana v kemijski strukturi in da lahko lokalni QSAR modeli zajamejo te povezave.

Teoretični razvoj, implementacija programske opreme in študija primera transkriptomike skupaj dokazujejo, da je lokalizacija v QSAR robusten in vsestranski pristop. Metodologija aiQSAR in programska oprema nudita takojšnjo uporabo platforme za kemijsko oceno tveganja, odkrivanje zdravil in mehanistično toksikologijo, hkrati pa vzpostavljata temelje za prihodnje povezovanje QSAR s transkriptomiko na nastajajočih znanstvenih področjih računalniške, napovedne in razvojne biologije.

Contents

List of Figures **xvii**

List of Tables **xix**

Abbreviations **xxi**

Glossary **xxiii**

1 Introduction **1**

2 Literature Review **5**

2.1 QSAR 5

2.1.1 Historical Development 5

2.1.1.1 A Note on Nomenclature: SAR vs QSAR vs QSPR 6

2.1.2 Molecular Descriptors 6

2.1.2.1 Dragon Software for Descriptor Calculation 8

2.1.3 Directions of QSAR Development 9

2.1.3.1 Integrated Consensus Predictions 9

2.1.3.2 Workflow Automation 10

2.1.3.3 Introduction of Machine Learning 10

2.1.4 QSAR in a Regulatory Context 11

2.1.4.1 OECD Principles for QSAR Validation 12

2.1.4.2 REACH Regulation 13

2.1.4.3 The ICH Guideline for Pharmaceuticals 13

2.1.5 Applicability Domain Assessment 13

2.2 Machine Learning 15

2.2.1 Lasso and Elastic-Net 16

2.2.2 Boosted Generalized Linear Models 17

2.2.3 Boosted Logistic Regression 18

2.2.4 Random Forests 18

2.2.5 Extremely Randomized Trees 19

2.2.6 Support Vector Machines 20

2.2.7 Sparse Distance Weighted Discrimination 21

2.2.8 Gaussian Processes with Polynomial Kernels 22

2.2.9 Logistic Model Trees 22

2.2.10 Summary of Applied Methods 23

2.2.11 Neural Networks 24

2.3 Cheminformatics 26

2.3.1 Relationship to Computational Chemistry 26

2.3.2 SMILES 27

2.3.2.1 Atoms, Bonds, Branches and Rings 27

2.3.2.2 Stereochemistry and Chirality 28

2.3.2.3 Disconnected Structures ... 30
2.3.2.4 Canonicalization ... 30
2.3.3 Molecular Fingerprints ... 31
2.3.3.1 PubChem Fingerprints ... 31
2.3.3.2 Extended Fingerprints ... 32
2.3.3.3 Other Fingerprint Variants ... 33
2.3.4 Chemical Similarity ... 35
2.3.4.1 Descriptor-Based Approaches ... 35
2.3.4.2 Fingerprint-Based Approaches ... 36
2.3.4.3 Maximum Common Substructure ... 37
2.3.5 Applications ... 38
2.4 Statistics ... 39
2.4.1 Statistics in Cheminformatics ... 39
2.4.2 Classification Performance Metrics ... 39
2.4.2.1 Accuracy ... 40
2.4.2.2 Balanced Accuracy ... 41
2.4.2.3 Matthews Correlation Coefficient ... 43
2.4.2.4 Cohen's Kappa Coefficient ... 44
2.4.3 Regression Performance Metrics ... 45
2.4.4 Statistical Hypothesis Testing ... 46
2.4.4.1 Parametric and Nonparametric Tests ... 46
2.4.4.2 P-Value ... 47
2.4.5 Cross-Validation Approaches ... 47
2.5 Transcriptomics ... 48
2.5.1 Central Dogma of Molecular Biology ... 48
2.5.1.1 Transcription ... 49
2.5.1.2 Translation ... 49
2.5.1.3 Gene Expression Regulation ... 50
2.5.2 DNA Sequencing Technologies ... 51
2.5.2.1 Pyrosequencing ... 51
2.5.2.2 Illumina Sequencing ... 52
2.5.2.3 Single Molecule Real Time Sequencing ... 54
2.5.2.4 Nanopore Sequencing ... 55
2.5.3 RNA Sequencing Approaches ... 56
2.5.3.1 TempO-Seq ... 57

3 Results and Discussion 59
3.1 Local QSAR Modelling ... 59
3.1.1 Motivation ... 59
3.1.1.1 Overview of Local Modelling in QSAR ... 59
3.1.1.2 Theoretical Advantages ... 60
3.1.1.3 Development within Computational Toxicology ... 61
3.1.2 Development of aiQSAR Methodology ... 61
3.1.2.1 aiQSAR Workflow ... 62
3.1.2.1.1 Step 1: Calculating Descriptors ... 62
3.1.2.1.2 Step 2: Calculating Fingerprints ... 62
3.1.2.1.3 Step 3: Calculating Fingerprint Similarities ... 62
3.1.2.1.4 Step 4: Selecting the Local Group ... 64
3.1.2.1.5 Step 5: Filtering Descriptors ... 65
3.1.2.1.6 Step 6: Predicting the Target Compound ... 65
3.1.2.1.7 Step 7: Calculating the Consensus Value ... 65

3.1.2.1.8 Step 8: Applicability Domain Assessment 66
3.1.2.2 Consensus Thresholds for Binary Classification 66
3.1.3 Testing Framework: Rat Acute Oral Toxicity Initiative 68
3.1.3.1 Overview 68
3.1.3.2 Dataset Compilation, Curation and Composition 68
3.1.3.3 Five Endpoints 69
3.1.3.4 Statistical Analysis 70
3.1.3.5 aiQSAR Performance 70
3.1.3.6 Applicability Domain Performance Relationship 71
3.2 aiQSAR Software 72
3.2.1 Implementation 73
3.2.2 UI Structure 74
3.2.2.1 Basic Mode 74
3.2.2.2 Advanced Mode 75
3.2.2.3 Output and File Management 77
3.2.3 Endpoints 77
3.2.3.1 Acute Toxicity 78
3.2.3.2 Tox21 Data Challenge 78
3.2.3.3 VEGA Software Endpoints 79
3.2.3.4 PubChem Assays 79
3.2.3.5 Mario Negri Institute Endpoints 80
3.2.4 Outputs 80
3.2.4.1 Basic PDF Output – Training 80
3.2.4.2 Basic PDF Output – Test 83
3.2.4.3 Detailed TXT Output 85
3.2.4.4 Single-Compound Output 86
3.2.4.5 New Model Export 89
3.3 Predicting Differential Gene Expressions 89
3.3.1 Motivation 89
3.3.2 TempO-Seq and Differential Expression 90
3.3.2.1 Experimental Design and Sample Preparation 90
3.3.2.2 Differential Expression Quantification 90
3.3.3 Exploratory Data Analysis 91
3.3.3.1 Differential Expression Distribution 92
3.3.3.2 Directional Patterns in Differential Expression 93
3.3.3.3 Dataset Processing and Filtering 94
3.3.4 aiQSAR Results 95
3.3.4.1 Performance Evaluation 95
3.3.4.2 Statistical Analysis of Top 10 Genes 96
3.3.4.3 Comparative Dataset Composition Analysis 97
3.3.5 Differential Expression Predictability and Biological Implications 100
3.3.5.1 Biological Function of Top 10 Genes 100
3.3.5.2 Biological Implications of Predictive Performance 102

4 Conclusions **105**

Appendix A Endpoints Statistics **107**
A.1 Acute Tox - EPA 108
A.2 Acute Tox - GHS 110
A.3 Acute Tox - LD_{50} 112
A.4 Acute Tox - NonTox 113

A.5 Acute Tox - VeryTox 115
A.6 IMN - aromatase 117
A.7 IMN - DILI 119
A.8 IMN - HIA 119
A.9 IMN - Pgp 120
A.10 PubChem AID 941 - TLR4 122
A.11 PubChem AID 624156 - ROS 122
A.12 PubChem AID 1159518 - NFkB 123
A.13 Tox21 - AhR 125
A.14 Tox21 - AR 127
A.15 Tox21 - ARE 129
A.16 Tox21 - ARLBD 131
A.17 Tox21 - aromatase 133
A.18 Tox21 - ATAD5 135
A.19 Tox21 - ER 137
A.20 Tox21 - ERLBD 139
A.21 Tox21 - HSE 141
A.22 Tox21 - MMP 143
A.23 Tox21 - p53 145
A.24 Tox21 - PPAR gamma 147
A.25 Vega - Androgen 149
A.26 Vega - Carcinogenicity 149
A.27 Vega - BCF 150
A.28 Vega - Daphnia EC_{50} 151
A.29 Vega - Estrogen 152
A.30 Vega - Mutagenicity 152
A.31 Vega - Fish 153
A.32 Vega - Fathead 154
A.33 Vega - Henry's law 155
A.34 Vega - KM 156
A.35 Vega - LogP 157
A.36 Vega - Thyroid Receptor Alpha 158
A.37 Vega - Thyroid Receptor Beta 160
A.38 Vega - Water solubility 162

Appendix B Differential Expression – Tested Chemicals and Genes 163
B.1 Chemicals 163
B.2 Genes 169

References 179

Bibliography 189

Biography 191

List of Figures

Figure 2.1: QSAR visualization. 6
Figure 2.2: Commonly defined molecular descriptor categories. 7
Figure 2.3: Visualization of the random forest ML algorithm. 19
Figure 2.4: Visualization of the support vector machine ML algorithm. 20
Figure 2.5: Visualization of the neural network. 25
Figure 2.6: Extended fingerprint generation 33
Figure 2.7: Different MCS ring treatment approaches 37
Figure 2.8: Confusion matrix and derived statistics for the binary classification dataset. 40
Figure 2.9: Balanced accuracy statistic limitation. 42
Figure 2.10: Central dogma of molecular biology. 49
Figure 2.11: Genetic code table. 50
Figure 2.12: Pyrosequencing procedure 52
Figure 2.13: Illumina sequencing platform. 53
Figure 2.14: PacBio SMRT sequencing 54
Figure 2.15: Nanopore sequencing visualization 55
Figure 2.16: TempO-Seq protocol. 57
Figure 3.1: Motivation for QSAR localization 60
Figure 3.2: aiQSAR workflow. 63
Figure 3.3: Local group selection. 64
Figure 3.4: Calculating the ADM. 67
Figure 3.5: Various performance measures plotted against the required number of positive predictions for the Ames mutagenicity test dataset. 67
Figure 3.6: aiQSAR methodology performance on acute toxicity endpoints with respect to the ADM score 72
Figure 3.7: The initial aiQSAR application screen, ...73
Figure 3.8: Basic mode interface for standard test set prediction 74
Figure 3.9: Basic mode interface for single chemical prediction 75
Figure 3.10: Advanced mode settings interface 76
Figure 3.11: Advanced mode interface for model export functionality 77
Figure 3.12: Portion of a basic PDF output for a training set regression endpoint. 81
Figure 3.13: Portion of a basic PDF output for a training set binary classification endpoint. 82
Figure 3.14: Portion of a basic PDF output for a training set multi-class classification endpoint. 83
Figure 3.15: Portion of a basic PDF output for a test set of regression endpoint. 84
Figure 3.16: Portion of a basic PDF output for a test set of a multi-class classification endpoint. 85
Figure 3.17: First page of a single-compound output for the logP endpoint 87
Figure 3.18: Second page of a single-compound output for the logP endpoint 88
Figure 3.19: Schematic of the TempO-Seq experiment and the differential expression analysis. 91

Figure 3.20: Differential expression exploratory data analysis. 92
Figure 3.21: Experimental differential expression results. 94
Figure 3.22: A comprehensive visualization of performance metrics for differential gene expression predictions 96
Figure 3.23: Violin plots visually comparing the number of active chemicals that cause the differential gene expression 99

List of Tables

Table 2.1: Blocks of molecular descriptors calculated by Dragon. 8
Table 2.2: OECD principles for the validation of QSAR models. 12
Table 2.3: Applicability domain assessment methods for QSAR modelling. 14
Table 2.4: VEGA ADI categories. 15
Table 2.5: Machine learning methods used in aiQSAR. 23
Table 2.6: Additional SMILES examples. 29
Table 2.7: PubChem fingerprint bit examples 32
Table 2.8: Some commonly used molecular fingerprint variants. 34
Table 3.1: EPA Toxicity Categories 70
Table 3.2: GHS Toxicity Categories 70
Table 3.3: aiQSAR performance on five acute toxicity endpoints 71
Table 3.4: Confusion matrices and statistical performance metrics for the top 10 best predicted genes 98
Table 3.5: Dataset information and statistical test p-values for comparative analysis of the distribution of active chemicals 100
Table 3.6: Biological characteristics of the top 10 genes with highest differential expression predictability. 101

Abbreviations

3Rs	... replacement, reduction and refinement
AD	... applicability domain
ADI	... Applicability Domain Index
ADM	... Applicability Domain Measure
AhR	... aryl hydrocarbon receptor
APS	... adenosine-5'-phosphosulfate
AR	... androgen receptor
ARE	... antioxidant responsive element
ARLBD	... androgen receptor ligand-binding domain
BA	... balanced accuracy
BCF	... bioconcentration factor
BM	... Brunner-Munzel (test)
CPN	... counterpropagating neural networks
DILI	... drug-induced liver injury
DNA	... deoxyribonucleic acid
emPCR	... emulsion PCR
EPA	... Environmental Protection Agency
ER	... estrogen receptor
ERLBD	... estrogen receptor ligand-binding domain
FAIR	... findable, accessible, interoperable and reusable
GHS	... Globally Harmonized System of Classification and Labeling of Chemicals
GLM	... generalized linear models
GP	... gaussian process
HIA	... human intestinal absorption
HSE	... heat shock factor response element
HTS	... high-throughput screening
ICCVAM	... Interagency Coordinating Committee on the Validation of Alternative Methods
ICH	... International Council for Harmonisation
IP3	... inositol-1,4,5-trisphosphate
IRFMN	... Mario Negri Institute for Pharmacological Research
LLR	... local lazy regression
LMT	... logistic model trees
MAE	... mean absolute error
MCC	... Matthews correlation coefficient
MCS	... maximum common substructure
MLE	... maximum likelihood estimation
MLR	... multiple linear regression
MMP	... mitochondrial membrane potential
mRNA	... messenger RNA
MWW	... Mann-Whitney-Wilcoxon (test)

NFκB	... nuclear factor kappa-B
NGS	... next-generation sequencing
NICEATM	... NTP Interagency Center for the Evaluation of Alternative Toxicological Methods
NIH	... National Institutes of Health
NN	... neural network
NTP	... National Toxicology Program
OECD	... Organisation for Economic Cooperation and Development
P-gp	... P-glycoprotein
PacBio	... Pacific Biosciences
PCR	... polymerase chain reaction
PPAR-γ	... peroxisome proliferator-activated receptor gamma
QSAR	... quantitative structure-activity relationship
QSPR	... quantitative structure-property relationship
QSTR	... quantitative structure-toxicity relationship
R^2	... coefficient of determination
REACH	... Registration, Evaluation, Authorisation and Restriction of Chemicals
ReLU	... rectified linear unit
RMSE	... root mean square error
RNA	... ribonucleic acid
RNA-seq	... RNA sequencing
ROS	... reactive oxygen species
rRNA	... ribosomal RNA
SAR	... structure-activity relationship
SDWD	... sparse distance weighted discrimination
SMILES	... Simplified Molecular Input Line Entry System
SMRT	... single-molecule real-time (sequencing)
SVM	... support vector machines
Tanh	... hyperbolic tangent
TempO-Seq	... Templated Oligo assay with Sequencing readout
TEST	... Toxicity Estimation Software Tool
TLR4	... toll-like receptor 4
Tox21	... Toxicology in the 21st Century
tRNA	... transfer RNA
VEGA	... Virtual models for property Evaluation of chemicals within a Global Architecture

Glossary

QSAR

QSAR (Quantitative Structure-Activity Relationship): A computational method that uses mathematical models to correlate chemical structure with various endpoints, such as biological activity, physicochemical properties, or toxicity, enabling prediction of these values for untested compounds.

aiQSAR: A specialized methodology and an open-source software for localized QSAR modelling developed in this thesis.

Local QSAR: An approach that builds specialized predictive models for a target compound based only on a local group of structurally similar training chemicals, rather than a single global model for all compounds.

Global QSAR: Traditional QSAR methods that use a single mathematical model to represent the entire chemical space of a training set.

Applicability domain (AD): The specific chemical, structural, or biological space within which a QSAR model's predictions are considered reliable and valid.

Applicability Domain Measure (ADM): A quantitative applicability domain score between 1 and 5, assigned to each prediction in the aiQSAR software. Indicates the reliability of the associated prediction based on the structural similarity between the target compound and its local group.

Training set: In QSAR context, a dataset of compounds with known endpoint values used to train and build a predictive model, as opposed to the test set used for independent performance assessment.

Test set: Often interchangeably called validation set, or evaluation set. In QSAR context, a dataset of compounds with known endpoint values used for independent performance assessment.

External set: In QSAR context, a dataset of compounds where endpoint values are not known to the person or the team making predictions. Used for external performance assessment. Because it is sometimes called validation or evaluation set (see above), it is important to examine the use-case-specific terminology.

Cheminformatics

Cheminformatics: Computational and information technology methods that enable the systematic representation, analysis and prediction of chemical properties.

Molecular descriptors: Numerical values derived from chemical structure that encode various structural, electronic, or topological features of a molecule, used as input variables for QSAR modelling.

Molecular fingerprints: Binary string representations of a chemical structure.

Hash collision: In algorithmic fingerprints, the situation where the hashing function maps multiple distinct molecular fragments to the same bit position.

SMILES (Simplified Molecular Input Line Entry System): A standard line notation used to represent chemical structures as compact text strings.

Disconnected structures: Sections of a chemical structure that are not connected by a covalent bond, such as salts, mixtures, or complexes. In SMILES notation separated by a period (.) symbol.

Chemical similarity: A quantitative measure of the degree to which two molecules share structural features, often calculated from molecular descriptors or fingerprints.

Read-across: An expert-based technique used to fill data gaps by inferring a target property, often toxicological endpoint, from chemical analogues with known values and from related data sources.

Machine learning and statistics

Machine learning: A field of artificial intelligence focused on building algorithms that can learn mathematical relationships from training data in order to make predictions and generalize to new, unseen data.

Binary classification: A supervised learning task where the goal is to assign observations into one of two mutually exclusive categories.

Multi-class classification: A supervised learning task where observations are assigned to one of three or more distinct categories.

Regression: A supervised learning task where the goal is to predict continuous numerical values.

Cross-validation: A statistical method used to estimate model performance by partitioning data into subsets used iteratively for training and validation. Common variants include k-fold and leave-one-out cross-validations.

Consensus prediction: A prediction strategy that combines outputs from multiple independent models, often using different algorithms, to generate a final prediction, typically improving robustness and reliability.

Transcriptomics

Transcriptomics: The comprehensive study of the transcriptome, the complete set of RNA transcripts produced by the genome under specific conditions, typically focusing on quantifying mRNA levels to understand gene expression patterns.

Transcription: A process by which genetic information stored in DNA is transcribed into newly assembled RNA.

Translation: A process by which proteins are synthesized using mRNA molecules as templates.

Differential gene expression: Changes in gene expression levels between different experimental conditions.

Downregulation: A decrease in gene expression level for a treated sample, in comparison to the control sample. Also called under-expression.

Upregulation: An increase in gene expression level for a treated sample, in comparison to the control sample. Also called over-expression.

TempO-Seq (Templated Oligo assay with Sequencing readout): A targeted RNA sequencing platform that enables high-throughput quantification of transcriptomic profiles.

Chapter 1

Introduction

The quantitative structure-activity relationship (QSAR) has emerged as an essential tool in cheminformatics for the prediction of chemical values across diverse scientific areas ranging from physicochemical characterization and pharmaceutical drug discovery to regulatory toxicology and environmental hazard assessment. The theoretical foundation of QSAR, that molecular structure contains information sufficient to predict various biological, physicochemical and toxicological endpoints, has been extensively scientifically validated. Still, the practical limitations of conventional QSAR approaches remain significant. Traditional QSAR models, employing either classical linear regression or more sophisticated machine learning algorithms, typically construct models across the entire chemical space of a training dataset. This global approach, while conceptually straightforward, often fails to capture localized structure-activity relationships. The challenge these models face becomes particularly acute with complex, often biologically relevant endpoints where multiple mechanistic pathways may lead to similar biochemical outcomes.

The theoretical motivation for local QSAR modelling is compelling. Chemical space is fundamentally heterogeneous, with distinct regions characterized by different mechanistic drivers of activity. At the same time, a molecule's behavior towards a given endpoint often depends on subtle structural variations that may produce dramatically different outcomes. Finally, the concept of structural analogues, that chemically similar compounds should exhibit similar activities, remains one of the most fundamental and widely applied principles in cheminformatics and computational chemistry. Local QSAR strategy leans into these concepts by treating predictions for each target compound as an inherently local problem, constructing individualized models based on structurally similar training set compounds, rather than imposing a single global model across diverse chemical space. This localization approach should theoretically improve predictive performance by allowing the model to adapt to the mechanistic drivers most relevant to the local neighborhood of the target chemical, while simultaneously minimizing the influence of structurally distant compounds that may introduce noise into the modelling process.

Despite the theoretical appeal of local QSAR approaches, their adoption has been limited in both academic research and regulatory practice. This underutilization is likely due to a combination of technical and regulatory hurdles. On a practical level, local QSAR modelling requires at-runtime construction and execution of predictive models targeted to each individual chemical in a dataset. Furthermore, widespread adoption is hindered by issues related to acceptability, which is intrinsically linked to the model's intended purpose, ranging from R&D and prioritization to the formal authorization of a substance. Authorization carries the highest level of requirements, demanding a degree of correctness where predictions ideally mirror experimental results.

Achieving this high level of correctness is challenging given the current state of the art and the aforementioned technical hurdles. Previous local QSAR frameworks, while showing

promise, have typically been limited to simple mathematical approaches and applicable only to a small number of endpoints. Consequently, a comprehensive methodology for the implementation and validation of local QSAR models has remained absent from the literature, impeding the broader application of this approach for scientific research and its adoption within a regulatory context.

This doctoral dissertation addresses these gaps through three integrated sections with distinct objectives: (1) development of a comprehensive theoretical framework for local QSAR methodology, (2) implementation of this methodology into an open-source software, aiQSAR and (3) application of the aiQSAR protocol to a novel challenge within transcriptomics, differential gene expression predictions. Accompanying the presented results is the literature review establishing the historical context, theoretical foundations and the current state of QSAR research, as well as discussing some broader relevant topics, machine learning algorithms, cheminformatics methods and transcriptomics technologies.

The first section develops a comprehensive theoretical framework for local QSAR methodology. This section discusses the advantages of local modelling approach, describing how localization derives endpoint predictions from the region of chemical space most relevant to each target compound. The workflow for implementing local QSAR is presented in detail, encompassing descriptor calculation, fingerprint generation, local group selection based on chemical similarity, descriptor filtering and consensus prediction calculation across different endpoint types. Applicability domain assessment associates each prediction with a numerical score conveying the reliability of obtained results. This section further validates the methodology on five acute rat oral toxicity datasets modelled as a part of external, community-wide initiative with the aim of assessing toxicologically relevant endpoints, thus demonstrating competitive performance and establishing the soundness of the local QSAR method in comparison to conventional QSAR approaches.

The second section addresses the implementation of the aiQSAR software package. This software makes local QSAR methodology accessible to the scientific community. The implementation includes 38 distinct endpoints encompassing results from the acute toxicity modelling initiative, pharmaceutical properties relevant to drug development, environmental hazard endpoints and assay outcomes informing mechanistic toxicity pathways in computational toxicology. The software is divided into a Basic mode for routine predictions, suitable for users without advanced computational expertise and an Advanced mode enabling customized analyses, parameter optimization and development of new models from user-submitted datasets. The software enables a straightforward implementation of the newly developed endpoints, supporting long-term, community-driven expansion. Multiple output formats are available, including simple summary results, detailed textual outputs for downstream computational analysis, as well as single-compound reports enabling visual assessment of local groups and expert-based analysis of attained predictions. aiQSAR aims to be an accessible computational platform suitable for integration into regulatory assessment and industrial chemical discovery workflows.

The third section applies developed QSAR methodology to a novel domain, the prediction of chemically induced differential gene expression. Gene expression profiling is an advanced technique in transcriptomics, capturing the effects propagated through cellular regulatory networks. The integration of aiQSAR with TempO-Seq, a modern transcriptomics platform specifically developed for targeted RNA sequencing across large number of samples, offers an opportunity for development of predictive models to assess chemically-induced transcriptomic responses at scale. The hypothesis guiding this application is that chemical structure encodes information sufficient to predict which genes will be differentially expressed upon chemical exposure. This analysis moves beyond traditional QSAR domains focused on established endpoints such as toxicity classification

or physicochemical properties and demonstrates the applicability of aiQSAR methodology to complex biological responses.

Preceding the results, literature review establishes the theoretical foundations required for understanding this work and the motivation behind the three discussed objectives. The review begins with a brief introduction of QSAR methodology and its historical development from qualitative structure-activity analysis to modern quantitative approaches. This includes description and categorization of molecular descriptors, the numerical representations of chemical structure essential for QSAR modelling. The review also covers directions of contemporary QSAR development, including the shift toward integrated consensus predictions that combine multiple modelling approaches, workflow automation to standardize and improve reproducibility of results and the increasing adoption of sophisticated machine learning methods. The review addresses QSAR in regulatory contexts, including OECD principles for QSAR validation and REACH regulatory framework that explicitly recognizes computational approaches as alternatives to animal testing, establishing the practical importance of robust and well-validated QSAR models. The literature review further examines applicability domain assessment, the aim of which is determining the reliability of QSAR predictions, and various contemporary approaches to this important aspect of model development.

Then, a brief overview of machine learning methods integrated into the aiQSAR software is presented. This includes widely adopted, general-purpose algorithms used across many scientific disciplines and with applicability well beyond computational chemistry such as random forests, boosted generalized linear models, support vector machines, neural networks and others. The cheminformatics section of the literature review covers the computational representation of chemical structure, including a thorough description of SMILES notation used for encoding molecules into simple and computationally elegant form. A systematic treatment of molecular fingerprints, binary vector representations capturing specific structural features in chemicals, addresses multiple fingerprint schemes from predefined dictionary-based approaches to algorithmic hashing methods, with an emphasis on PubChem and Extended fingerprints, two variants implemented in the aiQSAR software. This section also includes brief discussion on the concept of chemical similarity, as well as some frequently used metrics for quantifying the similarity of either descriptor-based, or fingerprint-based molecular representations.

The literature review further discusses a statistical foundation for QSAR model validation and performance assessment. This covers an analysis of common classification performance metrics appropriate for binary and multi-class prediction tasks, including accuracy, balanced accuracy, Matthews correlation coefficient and Cohen's kappa, addressing the distinct statistical properties and appropriate applications of each metric. Regression performance metrics are similarly covered for continuous endpoint predictions. The review addresses statistical hypothesis testing, the interpretation of p-values in the context of large chemical datasets with many associated endpoints and cross-validation approaches essential during QSAR development. Finally, the review ends by covering the basics of transcriptomics, which are needed for the understanding of the novel QSAR application to differential gene expression predictions. This section introduces the central dogma of molecular biology, establishing the relationship between genotype and phenotype; explores modern DNA sequencing technologies, including Illumina, PacBio SMRT sequencing and nanopore platforms; and extends these topics to the RNA sequencing techniques used for transcriptomic profiling, with an emphasis on the TempO-Seq targeted sequencing platform applied in this research.

The integration of the three objectives covered in this dissertation demonstrates not only that the local QSAR approaches are computationally sound and practically feasible, but also that they offer a flexible framework capable of addressing diverse prediction

challenges spanning from established regulatory endpoints to cutting-edge functional genomics applications. By providing an accessible implementation of local QSAR methodology alongside comprehensive validation and novel applications, this dissertation aims to facilitate broader adoption of these approaches within academic research, regulatory agencies and industrial applications. Furthermore, by extending QSAR into the transcriptomics domain, this research anticipates a path for future integration of computational chemistry approaches within developmental and predictive biology. The aiQSAR software and the methodological framework described here provide both immediate practical value for chemical assessment, as well as foundation for future technical development and application to emerging challenges in computational toxicology, drug discovery and transcriptomics.

Chapter 2

Literature Review

2.1 QSAR

Quantitative Structure-Activity Relationship (QSAR) is an important computational method with applications in drug discovery, toxicological assessment, as well as in other segments of computational chemistry. This section provides a review of QSAR theory, historical development and applications relevant to the topics and results presented in this thesis.

2.1.1 Historical Development

The concept of correlating chemical structure to biological activity extends back centuries, but the formalization of QSAR as a scientific discipline emerged in the early 1960s with seminal work by Hansch and Fujita [1]. They developed the first quantitative approaches to relate molecular properties to pharmacological activity, establishing the mathematical framework that would underpin QSAR. Specifically, they showed that activity could be expressed as a mathematical function of physical and chemical properties of molecules:

$$Activity = f(physicochemical\ and\ structural\ properties) \tag{2.1}$$

This fundamental relationship transformed the previously qualitative field of structure-activity analysis into a rigorous, quantitative discipline capable of generating testable predictions.

Over the subsequent decades, QSAR development evolved significantly. Early approaches relied on simple linear regression and molecular properties such as lipophilicity and molecular size [2]. LogP for example, the logarithm of the octanol-water partition coefficient, is defined as the equilibrium ratio of a compound's concentration in a system composed of two solvents that do not mix, most commonly water and an apolar organic phase such as 1-octanol. It provides a practical quantitative measure of differential solubility between aqueous and hydrophobic environments and is widely used as a descriptor of molecular lipophilicity. Compounds with high LogP values are typically more hydrophobic and tend to preferentially partition into hydrophobic biological microenvironments such as lipid bilayers. Low LogP values indicate greater hydrophilicity and favor distribution in aqueous compartments such as blood serum. In contemporary computational workflows, LogP is frequently obtained from *in silico* QSAR models rather than determined experimentally for every compound of interest [3].

Precisely defining "chemical structure" and the functional form linking structure to an endpoint remains challenging, but these relationships are now routinely used in drug discovery and chemical risk assessment because they provide a practical bridge between molecular features and measurable outcomes (Figure 2.1).

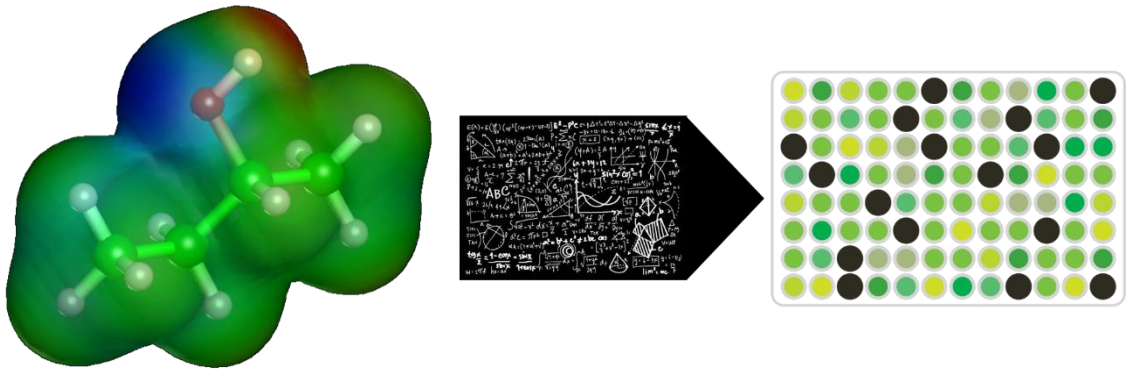


Figure 2.1: QSAR visualization. Mathematical methods (center) are used to translate the chemical structure (left) into measurable and predictable endpoint values (right).

The dramatic expansion of computational chemistry and machine learning in recent decades has transformed QSAR into an essential tool for chemical modelling, hazard prediction and regulatory decision-making. The transition from simple linear models to sophisticated ensemble machine learning methods represents both a technological advancement and a philosophical shift in how structure-activity relationships are conceptualized and exploited [4]. QSAR models today are built using many statistical and machine-learning approaches, trained on large and diverse chemical datasets and described with a wide range of molecular descriptors.

2.1.1.1 A Note on Nomenclature: SAR vs QSAR vs QSPR

SAR (structure-activity relationship) is a qualitative method, expressing structural or heuristic rules without committing to a numerical relationship. SAR typically makes use of rules created by experts to produce models that relate subgroups of the molecule to a biological property. In that sense it can be thought of as a simplified version of QSAR.

The term QSPR (quantitative structure-property relationship) is typically used when a physicochemical property is modeled as the response variable. However, the modelling procedure remains the same. In literature it is occasionally considered a subgroup of QSAR methods [5].

There are other related terms, such as QSTR (quantitative structure-toxicity relationship) for modelling toxicological endpoints. In this thesis, the term QSAR is used as a catch-all label and is considered to encompass all of the described methods.

2.1.2 Molecular Descriptors

Mathematical implementation of QSAR requires the transformation of chemical structure, an inherently qualitative object, into numerical variables, descriptors, suitable for statistical and machine learning analysis. Molecular descriptors are calculated numerical properties derived from chemical structure that encode information about molecular size, shape, polarity, aromaticity, functional group composition and topological features [6]. Descriptors are commonly classified into the following categories (Figure 2.2):

- **Constitutional descriptors** quantify the number and types of atoms, bonds and functional groups present in a molecule.
- **Geometrical descriptors** capture the size and shape features such as molecular surface area and volume, as well as related measures like moments of inertia.

- **Topological descriptors** are computed from the molecular graph (connectivity) and rely only on structural information, without requiring 3D geometry. Examples of topological indices include the Wiener index (the sum of the number of bonds between all nodes in a molecular graph) and the Randić index (derived from the branching segments of a molecule).
- **Physicochemical descriptors** approximate measurable physical properties of molecules, such as water solubility and partitioning behavior.
- **Electrostatic descriptors** represent charge-related properties, including partial atomic charges and features associated with hydrogen-bonding capability.
- **Quantum-chemical descriptors** derive from molecular orbital theory and related electronic structure properties.

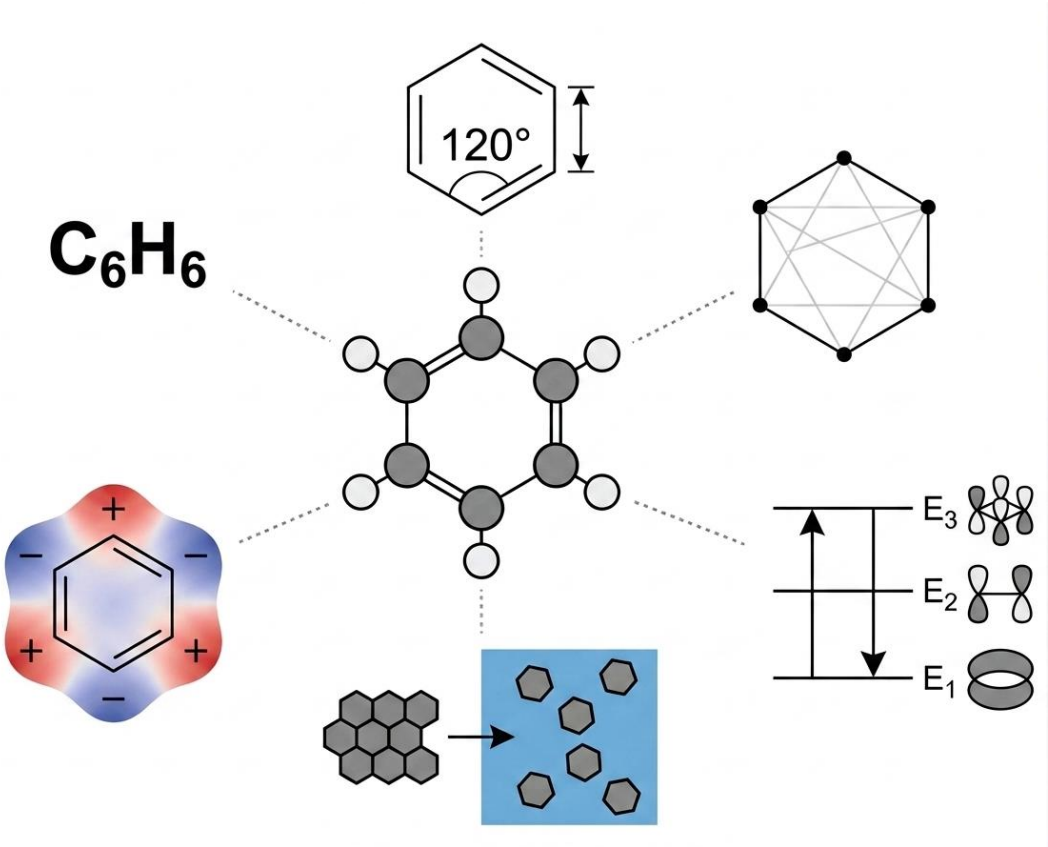


Figure 2.2: Commonly defined molecular descriptor categories. Clockwise from the top: geometrical, topological, quantum-chemical, physicochemical, electrostatic, constitutional.

The selection, quality and breadth of calculated descriptors substantially impact QSAR model performance, but a high-dimensional descriptor spaces also create statistical challenges. With thousands of potential descriptors and relatively limited numbers of training compounds, the risk of overfitting becomes substantial. Consequently, descriptor selection, the process of identifying the subset of descriptors carrying genuine predictive signal while discarding redundant or noisy descriptors, represents a critical step in QSAR model development.

2.1.2.1 Dragon Software for Descriptor Calculation

Dragon is a dedicated software package for the calculation of molecular descriptors [7]. It functions as a stand-alone descriptor engine where user supplies molecular structures and the software computes a large panel of descriptors in a batched manner suitable for large dataset processing. A defining characteristic of Dragon is the breadth of its descriptor space. Currently, Dragon provides 4,885 molecular descriptors, reflecting an intentionally extensive design in which many alternative descriptor families are computed in parallel to support exploratory modelling. This framework further subdivides the above descriptor categories into blocks (Table 2.1).

Table 2.1: Blocks of molecular descriptors calculated by Dragon.

ID	Block Name	Descriptor Count
1	Constitutional descriptors	43
2	Ring descriptors	32
3	Topological indices	75
4	Walk and path counts	46
5	Connectivity indices	37
6	Information indices	48
7	2D matrix-based descriptors	550
8	2D autocorrelations	213
9	Burden eigenvalues	96
10	P_VSA-like descriptors	45
11	ETA indices	23
12	Edge adjacency indices	324
13	Geometrical descriptors	38
14	3D matrix-based descriptors	90
15	3D autocorrelations	80
16	RDF descriptors	210
17	3D-MoRSE descriptors	224
18	WHIM descriptors	114
19	GETAWAY descriptors	273
20	Randic molecular profiles	41
21	Functional group counts	154
22	Atom-centred fragments	115
23	Atom-type E-state indices	170
24	CATS 2D	150
25	2D Atom Pairs	1596
26	3D Atom Pairs	36
27	Charge descriptors	15
28	Molecular properties	20
29	Drug-like indices	27

2.1.3 Directions of QSAR Development

QSAR methodology has evolved substantially over recent decades, with several distinct and complementary directions characterizing contemporary developments. These developments are driven by the need to improve model performance, increase accessibility to non-expert users and enhance the reliability of predictions for regulatory purposes. Three principal directions of QSAR development have emerged as particularly influential: the adoption of integrated approaches in modelling, the systematic automation and standardization of workflows and the migration towards sophisticated machine learning algorithms.

2.1.3.1 Integrated Consensus Predictions

The shift towards an integrated approach in modelling represents a structural change in how predictive models are constructed and utilized. Traditionally, QSAR modelling relied on the selection of an optimal modelling approach for a given endpoint. The integrated approach advocates for the consensus of several predictions, each obtained from a distinct QSAR method, to generate the final model output [8]. The theoretical foundation for consensus modelling rests on the principle that diverse modelling approaches capture different aspects of the structure-activity relationship. Combining individual model predictions through consensus mechanisms can leverage the strengths of different methodological approaches, while potentially mitigating the weaknesses of any single method [9].

This concept has been formally evaluated in multiple studies [10] [11]. Hewitt et al. conducted a comprehensive assessment of consensus regression models compared to single multiple linear regression approaches, examining multiple datasets of varying complexity. Their findings demonstrated that consensus models did not show significant statistical improvements over the best single models, but they did provide enhanced robustness. This conclusion is theoretically sound. Given that consensus models draw their final prediction from multiple individual models, they can never be expected to outperform the single best individual method. However, in a large data sample, the individual best performing method can never be known *a priori*. The demonstrated advantage of the consensus approach is to smooth out the distribution of outcomes and to improve the average performance compared to each individual method on a large sample of endpoints. Zhao et al. developed a hybrid system of multiple QSAR models designed to predict bioconcentration factors (BCF), incorporating both linear and non-linear modelling approaches to capture the complex structure-activity relationships inherent in this endpoint. These developments underscore the maturation of QSAR methodology towards sophisticated multi-model frameworks that acknowledge the intrinsic limitations of any single modelling approach.

A noteworthy practical example of the integrated approach is the structure of a community-wide modelling efforts, most notably the acute toxicity modelling initiative organized by the Interagency Coordinating Committee on the Validation of Alternative Methods (ICCVAM) [12] (see Section 3.1.3). This initiative shifted the prevailing paradigm from selecting a best individual method to explicitly combining predictions from all developed models, with differential weighting schemes. By aggregating predictions from multiple independent research groups employing diverse methodologies, the community-level consensus approach aims to achieve greater predictive reliability and broader acceptance of results from a regulatory standpoint.

2.1.3.2 Workflow Automation

A significant characteristic of contemporary QSAR development is the systematic streamlining and automation of the workflow. This development parallels trends emerging from related fields such as bioinformatics, where the exponential growth of available data has necessitated the development of scalable, automated methodologies [13]. The rationale for automation in QSAR modelling is to standardize the tools and to reduce the time required for model development, thereby enhancing transparency and reproducibility. This development is also significant for regulatory acceptance of QSAR models.

By minimizing user discretion in critical decision points, automation mitigates the influence of subjective choices that can significantly impact model performance and interpretation. To this end, specifically designed workflows have been developed for chemical data retrieval and quality checking prior to modelling applications, addressing one of the frequently overlooked steps in model development [14]. Concurrently, standardized methods for model performance validation have been established and automated, ensuring that models are evaluated consistently [15].

Additional automation efforts in QSAR development span from general platforms for model development, such as ChemSAR [16] and DPubChem [17], a web-based tool integrating QSAR modelling with high-throughput virtual screening capabilities, to specialized tools that aim to address specific frequent steps in QSAR development. The Open Babel chemical informatics toolbox provides standardized, open-source solutions for the automated processing of chemical structures across different software environments [18], while tools for automated data manipulation address class imbalance problems in training datasets [19].

The convergence of these automation efforts shifts the emphasis in QSAR development from specialized models designed by expert practitioners, towards reproducible, standardized procedures that can be executed with minimal intervention. This democratization of QSAR modelling promises to make the methodology more accessible to scientists from diverse backgrounds, but it also necessitates careful consideration for the flexibility needed to address unique characteristics of edge-cases in specific datasets and endpoints where the expert insight remains crucial.

This drive toward automation and standardization aligns seamlessly with the implementation of the FAIR (Findable, Accessible, Interoperable and Reusable) data principles within predictive toxicology [20]. Automated workflows facilitate the systematic capture of metadata, ensuring that QSAR models and their underlying datasets are not only reproducible, but also easily discoverable and interoperable across different computational platforms. By embedding FAIR principles directly into the modelling pipeline, the community moves closer to a rigorous standard of data stewardship where models are preserved for reusability and reproducibility, with the overarching aim of improving their regulatory acceptance.

2.1.3.3 Introduction of Machine Learning

A major development in the field of QSAR is the move from traditional linear methods towards more sophisticated machine learning algorithms. Multiple linear regression (MLR) can be considered a transitional approach between these methodologies, even though there is no clear, fundamental delineation. MLR methods have been foundational to QSAR due to their straightforward mechanistic interpretation where model coefficients directly indicate which molecular descriptors contribute most significantly to the predicted endpoint value. However, empirical evidence increasingly demonstrates that more complex machine learning algorithms achieve substantially better predictive performance [21].

Machine learning approaches derive their performance advantages from the capacity to capture complex structure-activity relationships that linear models cannot effectively represent. These complex and varied relationships are ubiquitous in chemical toxicity and bioactivity, reflecting the intricate biochemical and physicochemical mechanisms underlying molecular interactions. Machine learning algorithms such as random forests, support vector machines, artificial neural networks and gradient boosting methods have been shown to perform significantly better on a wide range of QSAR datasets. Specifically, the Tox21 challenge, a major community initiative for developing predictive models of nuclear receptor and stress response pathways, has demonstrated the superior predictive capability of machine learning approaches across multiple toxicological endpoints [22].

However, the cost of this performance advantage is the loss of mechanistic interpretability. For most machine learning algorithms, extracting mechanistic interpretation is extremely difficult, if not impossible. This problem is often described as the "black box" behavior of machine learning algorithms, wherein the mathematical procedure by which predictions are derived cannot be feasibly reversed or traced to identify the contributing chemical or molecular features. This interpretability problem has initiated a considerable research effort in the broader machine learning and artificial intelligence communities, as the inability to explain model predictions poses significant challenges for regulatory acceptance and scientific understanding [23].

The tension between predictive accuracy and interpretability represents a fundamental challenge in contemporary QSAR development. Regulatory agencies increasingly demand both high predictive accuracy and mechanistic understanding of how models arrive at their predictions. Consequently, while machine learning algorithms are being increasingly adopted for their superior predictive performance, there is concurrently a growing recognition of the need to develop methods and tools that can provide interpretability and explainability of machine learning predictions. Despite these challenges, the superior performance of machine learning algorithms has led to their widespread adoption in QSAR modelling. Contemporary best-practice QSAR modelling frequently employs ensemble methods combining multiple machine learning algorithms through consensus approaches.

2.1.4 QSAR in a Regulatory Context

The regulatory landscape for chemical safety assessment, traditionally reliant on *in vivo* animal studies, is increasingly accepting and in some cases mandating the use of *in silico* methods. This transition is driven by the need to reduce the high costs and time associated with experimental testing, but also by the ethical concerns for animal welfare. The latter is famously encapsulated as the 3Rs principle: Replacement, Reduction and Refinement [24], calling for the transition from animal testing methods, both for ethical reasons, but also due to scientific concerns, which recognize that animal models often do not map adequately to the relevant human endpoints.

Current legislation in the EU, the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH) regulation, explicitly promotes the use of QSAR methods to fill data gaps and minimize animal testing [25]. Similarly, in the pharmaceutical sector, the International Council for Harmonisation (ICH) guideline has established a global standard for using QSAR models as a primary tool for assessing mutagenic impurities [26].

2.1.4.1 OECD Principles for QSAR Validation

For a QSAR model to be accepted for regulatory purposes, it must demonstrate scientific validity and transparency. The global standard for this validation is established by the Organisation for Economic Cooperation and Development (OECD) guidelines. The OECD adopted five principles that every QSAR model should adhere to [27] (Table 2.2).

Table 2.2: OECD principles for the validation of QSAR models.

No.	Principle	Description
1	Defined endpoint	QSAR model should be associated with a clearly defined endpoint. The biological, toxicological or physicochemical property that the model is intended to predict must be unambiguously specified, including the experimental test method, units of measurement, exposure conditions and any relevant data handling rules.
2	Unambiguous algorithm	QSAR model should have an unambiguous algorithm. The model description should include the exact procedure used to generate predictions. The mathematical form of the model and all parameter values must be specified so that predictions can be reproduced independently.
3	Defined domain of applicability	Domain of applicability for the QSAR model should describe the chemical and biological space for which the model is expected to provide reliable predictions. The description of the applicability domain should allow an assessor to determine whether a given substance falls within, near the border of, or outside the domain and should also identify any additional limitations of the model.
4	Appropriate statistical metrics	QSAR model should be associated with appropriate measures of goodness-of-fit, robustness and predictivity. The OECD guidance stresses that external validation with an independent data set is preferred. The performance statistics should be reported for data within the applicability domain and interpreted in the context of the intended regulatory use.
5	Mechanistic interpretation	If possible, a QSAR model should be associated with a mechanistic interpretation. Where feasible, the descriptors and mathematical relationships used in the model should be interpretable in terms of known or plausible mechanisms of action and the relevance of the descriptors for the predicted endpoint should be demonstrated. The OECD explicitly recognizes that a mechanistic interpretation is not always achievable and the absence of such an interpretation does not automatically disqualify a model from regulatory use.

These principles have become the benchmark for evaluating QSAR models in regulatory submissions worldwide. They ensure that models are not merely statistically sound, but scientifically defensible and reproducible by independent researchers.

2.1.4.2 REACH Regulation

Under the EU REACH regulation, registrants are required to submit data on the intrinsic properties of substances, which allows the use of QSAR as a valid alternative to standard testing requirements, provided that the model is scientifically valid, the substance falls within the model's applicability domain and the model adheres to the OECD principles listed above [28].

For example, in environmental risk assessment QSARs are routinely used to predict physicochemical properties, bioaccumulation and aquatic toxicity. This is particularly valuable in addressing data gaps where significant, but not exhaustive evidence for chemical safety is present. In such cases, computational methods can significantly reduce the testing burden, with minimal downside. However, regulators remain cautious about accepting QSAR predictions as a complete substitute for experimental data without supporting information, such as read-across justifications or multiple complementary QSAR predictions [29].

2.1.4.3 The ICH Guideline for Pharmaceuticals

Unlike other frameworks where QSAR is an alternative option, ICH M7 guideline recommends *in silico* analysis as a primary and preferred method for hazard identification of impurities in drug substances and drug products [30]. The guideline requires the use of two complementary QSAR methodologies to maximize predictive confidence:

- **Expert rule-based systems**: Models that predict toxicity based on known structural alerts derived from literature and from expert knowledge. These systems identify and flag chemical structures and functional groups known to be associated with mutagenic potential.
- **Statistical-based systems**: Models that use machine learning or statistical regression methods to correlate molecular descriptors with biological activity. These systems leverage large training datasets to identify structure-activity patterns.

The decision algorithm established in ICH M7 combines these two prediction paths to reach a final conclusion regarding the safety of impurities. If both methods predict a negative result for mutagenicity, the impurity is considered non-mutagenic and no further testing is required. If one method predicts positive outcome and the other one negative, an expert review is performed using supporting literature to interpret the data and resolve the conflict [31]. If both systems predict positive, the impurity is considered potentially mutagenic and must be controlled below appropriate thresholds or subjected to further investigation.

This framework has significantly reduced the need for mutagenicity testing of impurities during drug development, representing a landmark achievement in replacing animal testing with computational methods [32]. The ICH M7 guideline has been adopted globally and serves as a model for how QSAR can be integrated into regulatory decision-making [30].

2.1.5 Applicability Domain Assessment

The applicability domain (AD) is defined as the chemical, structural and biological space within which QSAR predictions are considered reliable and valid. It is a part of the model's output, together with the corresponding endpoint predictions. In essence, the AD establishes the boundaries beyond which model predictions cannot be confidently trusted and should not be used for regulatory or operational decision-making. The significance of

AD assessment lies in its ability to distinguish between reliable predictions made through interpolation within the training data space and unreliable predictions resulting from extrapolation beyond the model's training boundaries [33].

Interpolation refers to predictions made for compounds that are structurally and chemically similar to those in the training dataset used to develop the model. In interpolative predictions, the target compound occupies a position within the chemical space defined by the training data. The reliability of interpolative predictions is usually high because the model has been directly trained on similar compounds in the training set and the ranges of molecular descriptor values for the target compound fall within the ranges observed in the training set. Conversely, extrapolation occurs when predictions are made for compounds that fall outside the chemical and structural space covered by the training dataset. Extrapolative predictions are a challenge in QSAR modelling because the model attempts to generalize beyond its direct knowledge. Extrapolative predictions are associated with molecular descriptor values for the target compound that extend beyond the ranges observed in training data and, consequently, reduced reliability and lower accuracy of model predictions.

Although several broad categories of AD approaches can be distinguished (Table 2.3), many contemporary QSAR methods and computational tools use some combination of these approaches for the AD assessment.

Table 2.3: Applicability domain assessment methods for QSAR modelling.

Method	Source for calculation	Description
Global, statistical	Descriptor space	Assesses representativeness of compounds in the descriptor space. Assumes that if a target compound has molecular descriptors within the ranges observed in the training set, it is within the applicability domain.
Local, Similarity-Based	Similarity metrics	Measures structural similarity between a target compound and the training set compounds.
Distance-Based	Descriptor or fingerprint space	Measures the distance between the target compound and its nearest training set neighbors.
Confidence-Based	Machine learning predictions	Assesses prediction consistency between multiple QSAR methods.

Notably, VEGA platform [34] calculates the Applicability Domain Index (ADI), a continuous measure of prediction reliability that accounts for the concordance between the predicted value and experimental values of similar compounds, the exact prediction results for compounds similar to the target and looks for relevant chemical features in both the target and the training set compounds. Based on these criteria, the ADI value is calculated and the reliability of prediction is categorized into one of three tiers (Table 2.4).

Table 2.4: VEGA ADI categories.

Category	Range	Prediction Confidence
High Reliability	ADI > 0.85	The compound is well within the applicability domain. Similar compounds in the training set are predicted accurately.
Moderate Reliability	$0.65 <$ ADI < 0.85	The compound is likely within the domain, but some issues exist.
Low Reliability	ADI < 0.65	The compound is outside the applicability domain. The prediction is potentially unreliable due to low similarity, rare fragments, or conflicting data in the neighborhood.

2.2 Machine Learning

Machine learning is a class of computational algorithms that aim to identify mathematical relationships between features, as the input variables and target, as the output variable, through iterative optimization processes that minimize prediction error on training data [35]. This data driven approach fundamentally differs from classical statistical methods that utilize explicit mathematical assumptions about data relationships. Machine learning algorithms make fewer prior assumptions and instead allow the data itself to reveal the underlying patterns.

The field of machine learning is traditionally partitioned into three primary frameworks, based on the nature of available training data and learning objectives:

- **Supervised learning** methods are used for training data that consist of explicit input-output pairs, the features and the corresponding target values. The algorithm learns to map inputs to outputs by minimizing discrepancies between actual and predicted values.
- **Unsupervised learning** methods address the cases where training data consists of input features without corresponding target values. Unsupervised algorithms aim to identify inherent patterns and clusters within the data. Some examples include clustering, where similar observations are grouped together, dimensionality reduction, projecting high-dimensional data into lower-dimensional spaces and anomaly detection, identifying unusual or outlying samples.
- **Reinforcement learning** describes scenarios where machine learning methods aim to take sequential actions that maximize cumulative reward signals received from the environment.

QSAR modelling falls squarely within supervised learning. Training data consist of compounds paired with known endpoint values and algorithms learn to predict endpoint values for new chemicals. Within this category, however, there are three distinct modelling paradigms, each requiring specialized mathematical framework:

- **Regression** is used for predicting continuous numerical values. Examples include bioconcentration factors, various toxicity endpoints and other dose-response measurements.

- **Binary classification** is used for assigning observations into one of two mutually exclusive categories. Examples include mutagenicity and carcinogenicity assessments, as well as various toxicity classifications.
- **Multi-class classification** is used for classification problems with multiple discrete classes. Examples include mode-of-action endpoints and various toxicity categorization frameworks.

The selection of an appropriate machine learning algorithm depends upon several considerations: the dimensionality of the feature space, the volume and quality of available training data, the computational resources available and the interpretability demands for predictions [36]. In QSAR applications within regulatory toxicology, ensemble methods, which aggregate predictions from multiple diverse algorithms, have demonstrated superior performance compared to individual models, particularly for datasets exhibiting heterogeneous chemical space and imbalanced class distributions [9].

2.2.1 Lasso and Elastic-Net

Linear regression is the simplest regression approach where model assumes the target value y to be a linear combination of input features:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon \tag{2.2}$$

where β_0 through β_p are regression coefficients and ε is the error term. Model fitting proceeds by identifying coefficient values that minimize the sum of squared residuals (differences between predicted and actual values) across the training set.

Generalized linear models (GLM) are a natural extension of classical linear regression toward more sophisticated machine learning frameworks, while maintaining high interpretability, which is of considerable importance in QSAR applications. The classical least squares approach to parameter estimation is replaced by maximum likelihood estimation (MLE), which identifies coefficients maximizing the likelihood of observing the training data. However, when the number of descriptors exceeds the number of training samples, or when multicollinearity exists among descriptors, a common occurrence in QSAR due to correlations among molecular properties, classical MLE becomes unstable and prone to overfitting. Regularization addresses this challenge by incorporating a penalty term into the objective function [37]:

$$L(\beta) = -l(\beta; X, y) + \lambda \cdot P(\beta) \tag{2.3}$$

where $l(\cdot)$ is the log-likelihood, $\lambda \geq 0$ is the regularization parameter controlling the strength of penalty application and $P(\beta)$ is the penalty function.

Two fundamental regularization approaches are widely employed: Lasso (Least Absolute Shrinkage and Selection Operator) and Ridge regression, which differ in their respective penalty formulations. Lasso applies penalty proportional to the sum of absolute coefficient values:

$$P_{\text{Lasso}}(\boldsymbol{\beta}) = \sum_{j=1}^{p} |\beta_j| \tag{2.4}$$

whereas Ridge regression employs a penalty proportional to the sum of squared coefficients:

$$P_{Ridge}(\boldsymbol{\beta}) = \sum_{j=1}^{p} \beta_j^2 \tag{2.5}$$

Elastic-Net synthesizes these two approaches through a weighted combination [38]:

$$P_{\text{ElasticNet}}(\boldsymbol{\beta}) = \alpha \sum_{j=1}^{p} |\beta_j| + (1-\alpha) \sum_{j=1}^{p} \beta_j^2 \tag{2.6}$$

where the mixing parameter $\alpha \in [0,1]$ balances sparsity when α approaches 1, with the stability benefits of Ridge regression when α approaches 0.

2.2.2 Boosted Generalized Linear Models

Boosting methodologies are a family of techniques that iteratively construct sequence of models, each fitted to correct the residuals or misclassifications of previous models. During the training of each iteration in the sequence, the instances poorly predicted by its predecessors are emphasized. Ultimately, all models are combined through weighted aggregation [39].

Boosted generalized linear algorithm proceeds through the following iterative procedure [40]:

Algorithm 2.1: GLM Boosting.

Initialize fitted values: $\hat{\mathrm{f}}^{[0]} = \bar{y}$

for iteration $m = 1,2,\ldots,M$:

Compute negative gradient (pseudo-residuals): $\mathbf{r}^{[m]} = -\frac{\partial \ell(\hat{\mathbf{f}}^{[m-1]})}{\partial \hat{\mathbf{f}}}$

Identify the variable with strongest correlation to residuals:

$$j^* = \arg \max_j |\text{Corr}(\mathbf{r}^{[m]}, \mathbf{x}_j)|$$

Fit univariate GLM to residuals: $\hat{g}_m = \text{GLM}(\mathbf{r}^{[m]} \sim \mathbf{x}_{j^*})$

Update fitted values: $\hat{\mathbf{f}}^{[m]} = \hat{\mathbf{f}}^{[m-1]} + \nu \hat{g}_m$

where $\nu \in \langle 0,1]$ is the learning rate.

end

For regression tasks, the loss function represents errors squared:

$$\ell(f) = \frac{1}{n} \sum_{i=1}^{n} (y_i - f(\mathbf{x}_i))^2 \tag{2.7}$$

with corresponding negative gradient:

$$-\frac{\partial \ell(f)}{\partial f(\boldsymbol{x}_i)} = y_i - f(\boldsymbol{x}_i) \tag{2.8}$$

which directly yields the residuals of ordinary least squares regression.

2.2.3 Boosted Logistic Regression

For classification problems, boosting methodologies adapt to the discrete output space through modified loss functions and probability-based predictions [41]:

$$\ell_{\text{Binomial}}(f) = \frac{1}{n}\sum_{i=1}^{n} \left[\log(1 + e^{-2y_i f(\boldsymbol{x}_i)})\right] \tag{2.9}$$

where the response $y_i \in \{-1,1\}$ denotes binary class membership. The connection to logistic regression emerges through the probabilistic interpretation:

$$P(y = 1|\boldsymbol{x}) = \frac{1}{1 + e^{-f(\boldsymbol{x})}} \tag{2.10}$$

The boosting procedure iteratively refines the classification boundary by focusing on wrongly classified edge-case observations, thereby improving the overall classification accuracy through ensemble aggregation.

2.2.4 Random Forests

Decision trees and their ensemble variants are among the most successful and practically deployed machine learning algorithms in contemporary applications. Tree-based methods possess the intrinsic capacity to capture complex relationships and high-order interactions among features without explicit specification of interaction terms [42].

The fundamental operation underlying tree-based regression and classification involves recursive binary partitioning of the feature space $\mathbb{R}^p$ into disjoint rectangular regions, $R_1, R_2, \ldots, R_K$, with constant predictions within each region. For regression, the prediction in region R_k is the mean response of training observations within that region. For classification, the response is the most frequent class.

The partition is constructed through recursive binary splitting. At each node, a feature j and threshold t are selected to minimize the resulting loss function. For regression with squared-error loss, the objective at each split is:

$$\min_{j,t} \left[\sum_{i:x_{ij} \leq t} (y_i - \overline{y_L})^2 + \sum_{i:x_{ij} > t} (y_i - \overline{y_R})^2 \right] \tag{2.11}$$

where $\overline{y_L}$ and $\overline{y_R}$ are mean responses in left and right child nodes. This greedy algorithm, recursive binary splitting, produces a sequence of increasingly complex trees. Tree complexity is controlled through stopping criteria, usually either the minimum node size or the maximum depth.

A fundamental limitation of individual decision trees is high variance in predictions resulting from instability to small perturbations of the dataset. A small change in training data can fundamentally alter the tree creation sequence, therefore producing a substantially different tree structure, which completely changes the associated predictions. Random Forest is an ensemble approach that aims to address this instability by constructing

multiple independent trees, each trained on a bootstrap sample drawn, with replacement, from the original training set [43] (Figure 2.3).

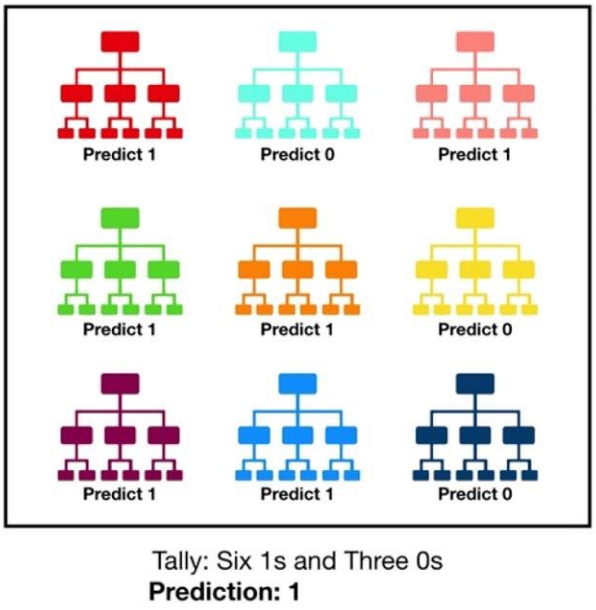


Figure 2.3: Visualization of the random forest ML algorithm. The outputs from individual decision trees are incorporated into the final prediction.

To further enhance ensemble diversity, feature randomness is introduced by selecting a subset of m features, from the overall set p. This feature sampling mechanism, commonly with $m = \sqrt{p}$ for classification and $m = p/3$ for regression, prevents overly emphasizing a small number of strongest features and allows weaker, but complementary features to contribute.

The final random forest prediction aggregates individual tree predictions through:

$$\hat{f}_{\mathrm{RF}}(\boldsymbol{x}) = \frac{1}{B}\sum_{b=1}^{B} \hat{f}_b(\boldsymbol{x}) \tag{2.12}$$

for regression and through majority voting for classification:

$$\hat{y}_{\mathrm{RF}}(\boldsymbol{x}) = \arg\ \max_{\mathrm{k}} \sum_{b=1}^{B} (\hat{f}_b(\boldsymbol{x}) = k) \tag{2.13}$$

2.2.5 Extremely Randomized Trees

Extremely randomized trees, or extra trees modify the random forest algorithm through two principal mechanisms. First, split thresholds are randomly selected for each feature rather than optimized over all possible thresholds. Second, the entire dataset is used for training each tree, rather than bootstrap samples. Randomization is achieved solely through feature and threshold selection. These modifications can reduce computational complexity while potentially achieving comparable or superior predictive performance [44].

The split selection in extra trees is done by selecting a random threshold t_j, from the interval $[x_{j,\min}, x_{j,\max}]$ for each feature j. Then, the impurity reduction, ΔI_j is computed:

$$\Delta I_j = I(\text{parent}) - \frac{n_L}{n} I(\text{left}) - \frac{n_R}{n} I(\text{right}) \tag{2.14}$$

where $I(\cdot)$ is the Gini index for classification and variance for regression, while n_L and n_R are the number of samples in left and right child respectively. The feature and threshold pair yielding maximum ΔI_j is selected in each step.

2.2.6 Support Vector Machines

Support vector machines (SVM) is a supervised machine learning method that aims to find the optimal boundary, called a hyperplane, that separates data points of different classes, by maximizing the distance to the closest point on either side. This distance is called the margin and the SVM is said to look for "margin maximization" [45] (Figure 2.4).

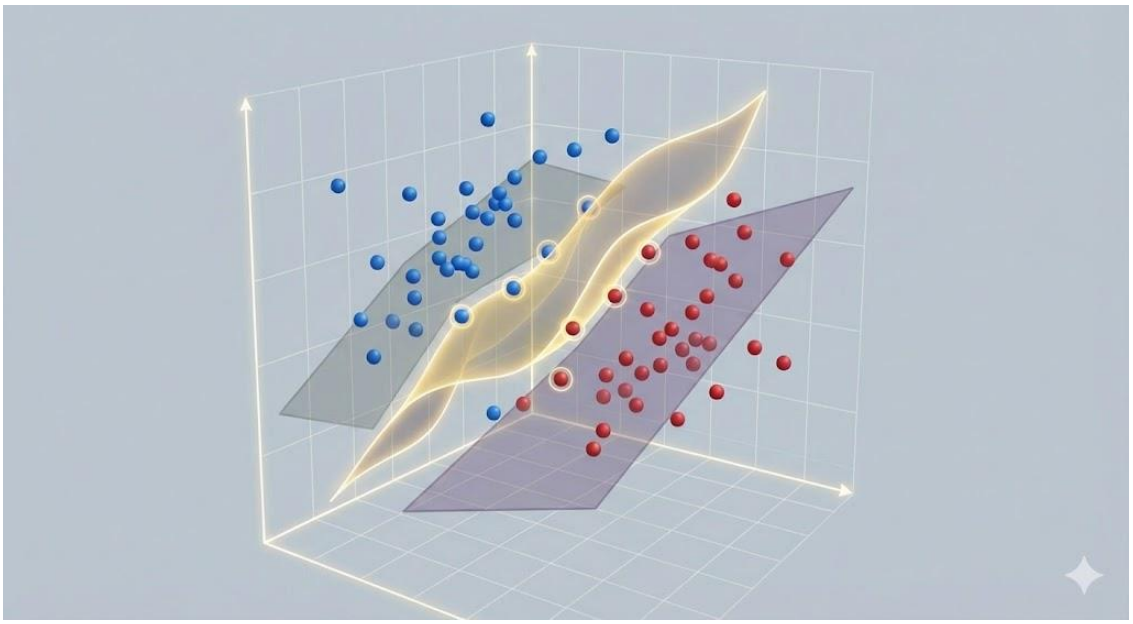


Figure 2.4: Visualization of the support vector machine ML algorithm. The highlighted plane represents the maximum margin separation of points. The curvature of the plane indicates the application of the kernel trick to achieve the separation in a higher-dimensional space.

Formally, given a set of data points, viewed as p-dimensional vectors, with response $y_i \in \{-1,1\}$, SVM seeks to identify a $(p-1)$-dimensional hyperplane defined by the equation:

$$w^T x + b = 0 \tag{2.15}$$

that maximizes the margin to the nearest points, defined as $\frac{2}{\|w\|}$. This equation defines the decision boundary. A new point is classified by simply evaluating which side of the hyperplane it falls on.

In practice, many datasets cannot be separated by a hyperplane due to the insufficient dimensionality of data points. To resolve this, SVM employ a "kernel trick" that implicitly maps the data into a higher-dimensional space where the separator exists, but without ever explicitly computing that higher-dimensional representation. As the name suggests, this is a mathematical trick and the dimensionality of the dataset isn't really changed. The dimensions of the dataset correspond to the features of data points (in the case of QSAR, these are molecular descriptors) and therefore cannot be unjustifiably changed. The kernel creates this higher dimensionality with specific properties, such that it never needs to be defined, only used during the computation. The only value that kernel function computes is the similarity between data points in this higher space, without needing to know the explicit coordinates of the transformation.

The kernel function is defined as:

$$K(x_i, x_j) = \phi(x_i) \cdot \phi(x_j) \tag{2.16}$$

where ϕ is a mapping function to a higher-dimensional space. Commonly used kernel functions are linear, polynomial and radial basis function kernels.

Additionally, SVM may employ a soft margin approach that allows some misclassifications, "margin violations". Essentially, not all training data points of the same class need to be on the same side of the hyperplane. Rather than strictly enforcing that all points are correctly classified, soft margin SVM minimizes:

$$\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \xi_i \tag{2.17}$$

where ξ_i measure the degree of violation for each point and C is a regularization parameter that controls the trade-off between maximizing the margin and minimizing classification errors. Large C penalizes misclassifications heavily while a smaller C allows more violations, making the model more tolerant to noise and outliers.

In the context of QSAR, which frequently deals with imbalanced classes, a class weighting scheme can be neatly used to set the value for the regularization parameter C by making it proportional to class frequencies:

$$C_k = \frac{C \cdot n}{2n_k} \tag{2.18}$$

where n_k is the number of class-k observations in the training set.

2.2.7 Sparse Distance Weighted Discrimination

Distance weighted discrimination is an alternative to SVM for high-dimensional classification. Rather than margin maximization, distance weighted discrimination minimizes the sum of inverse distances from observations to the separating hyperplane [46]:

$$\underset{\boldsymbol{w},b}{min} \sum_{i=1}^{n} \frac{1}{\xi_i} \text{ subject to } y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) \geq 1 \tag{2.19}$$

where ξ_i is the signed distance from observation i to the hyperplane. This objective function, minimizing the sum of reciprocals, rather than the sum of distances themselves, produces more balanced separating hyperplanes when $p >> n$.

Sparse distance weighted discrimination (SDWD) incorporates sparsity constraints through regularization [47]:

$$\min_{\boldsymbol{w},b,\boldsymbol{\xi}} \sum_{i=1}^{n} \frac{1}{\xi_i} + \lambda \|\boldsymbol{w}\|_1 \text{ subject to } y_i(\boldsymbol{w}^T \boldsymbol{x}_i + b) \geq 1 \tag{2.20}$$

This is the same sparsity mechanism used in lasso regression.

2.2.8 Gaussian Processes with Polynomial Kernels

A Gaussian Process (GP) is a collection of random variables such that any finite subset follows a multivariate Gaussian distribution. For regression, a GP is specified by a mean function $m(\mathrm{x})$ and covariance function $k(\mathrm{x}, \mathrm{x}')$:

$$f(x) \sim \mathcal{GP}(m(x), k(x, x')) \tag{2.21}$$

with observations arising from:

$$y_i = f(x_i) + \epsilon_i, \qquad \epsilon_i \sim \mathcal{N}(0, \sigma_n^2) \tag{2.22}$$

where σ_n^2 is the measurement noise variance. Assuming a zero-mean prior, the distribution over the function value at a test point x_*, given training data (X, y) is also Gaussian:

$$p(f_*|x_*, X, y) = \mathcal{N}(\mu_*, \Sigma_*) \tag{2.23}$$

with mean and covariance given as:

$$\mu_* = k_*^T (K + \sigma_n^2 I)^{-1} y \tag{2.24}$$

$$\Sigma_* = K_{**} - k_*^T (K + \sigma_n^2 I)^{-1} k_* \tag{2.25}$$

where K is the training covariance matrix $K_{ij} = k(\mathrm{x}_i, \mathrm{x}_j)$ and k_* is the vector of covariances between the test point and training points.

The polynomial kernel is defined as:

$$k_{\text{poly}}(x, x') = ((x^T x') + c)^d \tag{2.26}$$

where d is the polynomial degree and c is a constant.

2.2.9 Logistic Model Trees

Logistic model tree (LMT) is a classification model with an associated supervised training algorithm that combines logistic regression with decision tree learning [48]. LMT combines the advantages of tree-based partitioning with the probabilistic modelling of logistic regression. The methodology grows a tree structure where each leaf node contains a logistic regression model rather than a simple class label. This hybrid approach performs better on classification problems with complex decision boundaries.

During tree construction, at each node, LMT evaluates two splitting criteria:

1. Gini index reduction associated with binary feature split.
2. Improvement in model fit when fitting logistic regression to the data in the node.

The algorithm selects the split maximizing improvement and then recursively continues with child nodes. The probability of class membership at a leaf node is predicted through:

$$P(y = 1|\boldsymbol{x}) = \frac{1}{1 + e^{-(\beta_0 + \sum_{j=1}^{p} \beta_j x_j)}} \tag{2.27}$$

where coefficients are estimated via maximum likelihood on the subset of training data reaching that leaf.

2.2.10 Summary of Applied Methods

The summary of all machine learning approaches used for the published version of the aiQSAR software (see Section 3.1.2.1) and discussed in previous sections is presented here (Table 2.5).

Table 2.5: Machine learning methods used in aiQSAR.

R Method	Full Name	Endpoint Type	ML Family
glmboost	Boosted Generalized Linear Model	Regression, binary classification	Boosting
gaussprPoly	Gaussian Process with Polynomial Kernel	Regression	Gaussian process
glmnet	Lasso and Elastic-Net Regularized Generalized Linear Models	Regression	Regularized linear model
rf	Random Forest	Regression, binary classification, multi-class classification	Tree-based ensemble
extraTrees	Random Forest by Randomization (Extremely Randomized Trees)	Regression, binary classification, multi-class classification	Tree-based ensemble
LogitBoost	Boosted Logistic Regression	Binary classification, multi-class classification	Boosting, logistic
svmLinear Weights	Linear Support Vector Machines with Class Weights	Binary classification	Support vector machine
LMT	Logistic Model Trees	Binary classification, multi-class classification	Tree-based ensemble, logistic
sdwd	Sparse Distance Weighted Discrimination	Binary classification, multi-class classification	Distance-weighted classification

2.2.11 Neural Networks

Neural network, or neural net (NN) is a computational model inspired by the structure and functions of biological neural networks. Several neural net approaches have been used during the intermediate stages of the aiQSAR methodology development. These approaches had the level of performance comparable to that of other methods. However, due to the computational requirements, NN have not been included in the final, published version of the aiQSAR software.

Most common applications of NNs are based on the error backpropagation training strategy. These are supervised NNs constructed from elementary computational units called artificial neurons, which are mathematical abstractions of biological neurons. Each neuron receives multiple input signals, either from previous layers, or from external data. The signals are combined with different weights and a nonlinear transformation, an activation function, is applied. Mathematically, the operation of a single neuron can be expressed as:

$$y = f\left(\sum_{i=1}^{n} w_i x_i + b\right) \tag{2.28}$$

where x_i is the input signals, w_i are the learnable weights that determine the importance of each input, b is a bias term that allows the neuron to shift its response and f is the activation function that introduces nonlinearity. The weights and biases are collectively referred to as the neuron's parameters.

A neural network is formed by arranging many such neurons in organized layers:

- **Input layer** receives raw data and has one neuron per input feature.
- **Hidden layers** perform feature extraction through nonlinear transformations.
- **Output layer** produces predictions or classifications depending on the task.

Data flows sequentially from the input layer through the hidden layers to the output layer in a process called forward propagation [49].

Activation functions are crucial to neural network functionality because they introduce nonlinearity into the system. Without nonlinearity, stacking multiple layers would produce only linear transformations and the entire network would behave equivalently to a single-layer linear model [50]. Nonlinearity is what enables neural networks to learn and represent complex relationships.

There are many activation functions and a few commonly used ones are listed here. Rectified linear unit (ReLU):

$$f(x) = max(0, x) \tag{2.29}$$

returns the input directly if positive and zero otherwise. Despite its simplicity, ReLU has become the default activation function in most modern neural networks because it is computationally efficient and leads to fast convergence. Sigmoid is defined as:

$$f(x) = \frac{1}{1 + e^{-x}} \tag{2.30}$$

and maps inputs to the (0,1) range. Hyperbolic tangent (Tanh) is defined as:

$$f(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \tag{2.31}$$

and maps inputs to the (-1,1) range. It is similar in behavior to sigmoid but often trains faster because its output is centered around zero.

The power of NN comes from being able learn appropriate weight values from data. The standard learning algorithm used to train neural networks is based on the idea of iteratively moving parameters in the direction of decreasing errors. Forward propagation is the process of computing the network's output given an input. Each neuron applies its weighted combination and activation function in sequence, layer by layer. Backpropagation is a systematic method for computing the gradient of the loss function with respect to each weight in the network. The key insight is that the chain rule from calculus allows efficient computation of these gradients in one reverse pass through the network, starting from the output and working back to the inputs. The weights are updated by the gradient descent optimization procedure:

$$w_{new} = w_{old} - \alpha \frac{\partial L}{\partial w} \tag{2.32}$$

where α is the learning rate, a parameter controlling the magnitude of weight adjustments in each step and L is the loss function. By repeatedly applying this update across all weights, with the data from the training set to calculate errors, the network's parameters gradually adjust to improve predictive performance (Figure 2.5).

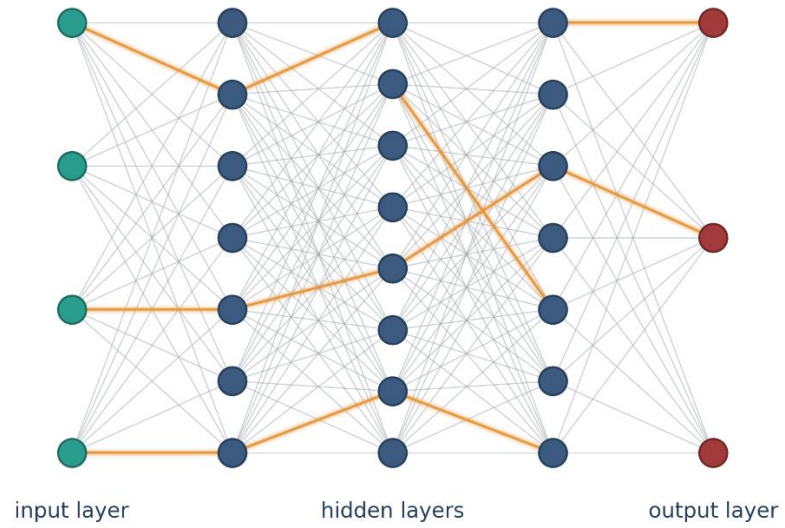


Figure 2.5: Visualization of the neural network. Filaments represent the weights between neurons used in forward propagation. The orange ones represent the changes in weights resulting from backpropagation during model training.

An alternative approach to the NN framework is through counterpropagating neural networks (CPN), a hybrid architecture combining unsupervised and supervised learning [51]. Like a regular NN, the CPN consists of three layers. They are called an input layer, a hidden Kohonen layer and an output Grossberg layer. The Kohonen layer arranges neurons in a typical lattice and learns to organize input patterns in an unsupervised manner. The Grossberg layer associates target output values with winning neurons. Unlike

standard networks where all layers are fully connected, in the CPN structure the input layer connects to the Kohonen layer through competitive weights, which then connect to the Grossberg layer through associative weights. The training procedure occurs in two sequential stages. In Kohonen training phase, the network receives only the input data and learns to organize itself in an unsupervised manner. For each input, the Kohonen layer computes the Euclidean distance to all neurons and identifies a winning neuron that best matches the input. In Grossberg training phase, the Kohonen weights are frozen and only the Grossberg layer is trained. For each input-output pair, the winning neuron from the first phase is identified and the corresponding Grossberg weights are updated.

CPN achieves substantially faster convergence than backpropagation. This is particularly useful for larger datasets and is therefore a good approach in QSAR modelling [52]. However, the CPN is limited by design to the number of neurons in the Kohonen layer. Each neuron represents a discrete class and compounds exciting the same neuron receive identical predictions from the Grossberg layer, potentially sacrificing the resolution in boundary regions. Modern applications frequently adopt hybrid approaches combining CPN's initialization and robustness with backpropagation's prediction refinement, such as the two-stage CP-BPE-ANN methodology, which achieves both the fast training of CPN and the continuous prediction capacity of backpropagation [53].

2.3 Cheminformatics

Cheminformatics is a computational field that enables the systematic representation, analysis and prediction of chemical properties through the application of computational methods. While sometimes treated as a distinct field, cheminformatics is best understood as a specialized subfield of computational chemistry that focuses on the representation, storage and analysis of chemical information at the digital level. By developing standardized computational representations of chemical structures and by deriving numerical descriptors of their properties, cheminformatics enables the transformation of qualitative structural knowledge into quantitative information suitable for mathematical analysis.

The scope of cheminformatics encompasses the development and application of molecular descriptors, the representation and standardization of chemical structures, the calculation and interpretation of chemical similarity measures, the design and optimization of molecular fingerprints, as well as the development and implementation of QSAR models. These tools and methods are used extensively in modern chemical research. Most notably, in drug discovery, computational toxicology, pharmacology and regulatory hazard assessment.

2.3.1 Relationship to Computational Chemistry

Computational chemistry is sometimes defined narrowly to encompass theoretical and computational methods for understanding chemical structure from the physical perspective, including quantum mechanical calculations, molecular dynamics simulations, conformational analysis and other approaches that require explicit consideration of molecular geometry and energies. In that sense, the previously defined scope of cheminformatics would be considered a distinct, yet complementary discipline. The issue arises on the boundaries of these two hypothetically distinct areas. For example, molecular descriptors and fingerprints, used extensively in cheminformatics and central to QSAR, are often derived from quantum chemical calculations or molecular dynamics simulations. In

that sense, most contemporary cheminformatics applications, including QSAR, would fall in-between these fields, essentially combining some elements of both. However, this would leave very little in the strict scientific domain of cheminformatics.

A more practical and currently more common approach is to consider computational chemistry as a broad scientific area, with cheminformatics as a more specialized, information-focused discipline operating on a higher level of abstraction and emphasizing a distinctly informatics perspective. Rather than attempting to derive properties from first principles using quantum mechanics, cheminformatics employs empirical approaches that correlate easily calculated structural features with experimentally measured properties and establishes systematic computational frameworks for organizing, retrieving and analyzing chemical information. This empirical perspective enables cheminformatics to operate efficiently on large datasets comprising anywhere from the hundreds, up to the millions of chemicals [54], thus representing a higher level of abstraction in the computational chemistry framework, dealing with problems that would be prohibitively expensive for lower-level systems. The emphasis in cheminformatics, however, remains on predictive performance, practical utility and information management at scale.

2.3.2 SMILES

Simplified molecular input line entry system (SMILES) is a line notation for entering and representing molecules [55]. This compact, linear representation of chemical structure is particularly useful for cheminformatics applications. It can be processed as plain text without requiring specialized chemical structure editors or binary file formats and can therefore be readily used in large chemical databases. Because of these characteristics, the SMILES notation has become a *de facto* standard in cheminformatics, implemented across virtually all modern software platforms and databases [56].

2.3.2.1 Atoms, Bonds, Branches and Rings

In SMILES notation, atoms are represented by their atomic symbols using standard chemical notation. Hydrogen atoms are generally omitted from the notation. The number of hydrogen atoms attached to each atom is calculated based on standard valence rules. Letter symbols are used for C, N, O, P, S, F, Cl, Br and I, which are assumed to be in their "normal" valence state. Examples:

- `C` - aliphatic carbon with implicit hydrogen atoms
- `N` - aliphatic nitrogen with implicit hydrogen atoms
- `n` - aromatic nitrogen
- `[NH4+]` - ammonium cation
- `[O-]` - oxide anion
- `[Cu+2]` - copper (II) ion

Bonds between atoms are represented by specific symbols or by implicit adjacency. A hyphen (`-`) represents a single bond, an equal sign (`=`) represents a double bond and a hash symbol (`#`) represents a triple bond. Aromatic bonds are represented by lowercase letters. Examples:

- `CC` - single bond (ethane)
- `C=C` - double bond (ethene)
- `C#C` - triple bond (ethyne)
- `c1ccccc1` - aromatic bonds (benzene)

Branches are specified by enclosing the branched structure in parentheses. Parentheses can be nested to represent complex branching patterns. Examples:

- `CC(C)C` - isobutane, a central carbon with three methyl groups
- `CC(C)CO` - isobutyl alcohol, a branched chain with 4 carbons and a terminal hydroxyl group
- `CC(C(=O)O)N` - alanine, a branched structure with a carboxylic acid group
- `C(F)(F)F` - fluoroform, a single carbon bonded to 3 fluorine atoms and 1 hydrogen
- `CC(CCC(=O)N)CN` - a multi-branched structure with multiple functional groups

Ring closures are indicated by breaking one bond in each ring and marking the ring closure with a digit. The same digit appearing on two atoms indicates that those atoms are connected, closing the ring. Each ring closure digit can appear at most twice in a SMILES string. Digits 1 through 9 can be used directly and for 10 or more ring closures, the percentage symbol (`%`) is used before the first digit. Examples:

- `C1CCCCC1` - cyclohexane, a six-membered saturated ring
- `c1ccccc1` - benzene, a six-membered aromatic ring
- `O1CCOCC1` - 1,4-dioxane, six-membered ring with 2 oxygen atoms
- `C1CC2CCC1C2` - norbornane, a bi-cyclic structure

Aromatic atoms in rings are specified using lowercase letters. Aromaticity is always implicit and all adjacent atoms assume aromatic bonding. Examples:

- `c1ccccc1` - benzene
- `c1ccncc1` - pyridine
- `c1ccc(cc1)c2ccccc2` - biphenyl, two benzene rings connected
- `o1cccc1` - furan, a five-membered ring with oxygen
- `n1cccnc1` - pyrimidine, a six-membered ring with 2 nitrogen atoms

These rules are sufficient to unambiguously encode the majority of chemicals. Additional examples of more complex SMILES constructed by combining them are also presented here (Table 2.6).

2.3.2.2 Stereochemistry and Chirality

SMILES notation can be extended to molecules that exist as stereoisomers. Accurate representation of such molecular structures requires the ability to specify the three-dimensional spatial arrangement of atoms and groups.

Table 2.6: Additional SMILES examples.

Chemical Name	Formula	SMILES
Ethanol	C_2H_6O	CCO
Dimethyl ether	C_2H_6O	COC
Acetaldehyde	C_2H_4O	CC=O
Acetic acid	$C_2H_4O_2$	CC(=O)O
Methyl formate	$C_2H_4O_2$	COC=O
Hydrogen cyanide	CHN	C#N
Cyanide anion	CN^-	[C-]#N
Methyl isocyanate	C_2H_3NO	CN=C=O
Propane	C_3H_8	CCC
Isobutane	C_4H_{10}	CC(C)C
Cyclopropane	C_3H_6	C1CC1
Cyclohexane	C_6H_{12}	C1CCCCC1
1,4-Cyclohexadiene	C_6H_8	C1=CCC=CC1
Benzene	C_6H_6	c1ccccc1
Toluene (Methylbenzene)	C_7H_8	Cc1ccccc1
Aniline	C_6H_7N	Nc1ccccc1
Furan	C_4H_4O	o1cccc1
Imidazole	$C_3H_4N_2$	n1c[nH]cc1
Naphthalene	$C_{10}H_8$	c1ccc2ccccc2c1

Double bond notation can (but isn't required to) specify whether substituents are on the same side (cis) or on the opposite (trans) sides of a carbon-carbon double bond. Directional bonds are specified using forward slash (/) for a bond extending above the plane of the paper and backslash (\) for a bond extending below the plane of the paper. The specification is based on the order of atoms in the SMILES string such that, if you traverse the double bond from left to right in SMILES order, the two higher-priority groups are either on the same side (cis) or opposite sides (trans). Examples:

- `F/C=C/C=C/C` - completely specified configuration, trans orientations on both double bonds
- `F/C=C/C=CC` - the first double bond is fully specified as trans, the second is unspecified
- `CC/C=C/C` - trans configuration around the double bond
- `CC/C=C\C` - cis configuration around the double bond

The simplest and most common kind of chirality is tetrahedral, where four neighboring atoms are evenly arranged around a central atom. If all four neighbors are different from each other in any way, mirror images of the structure will not be identical. The two mirror images are known as "enantiomers" and are the only two forms that a tetrahedral center can have. In SMILES, tetrahedral centers may be indicated by a simplified chiral specification following the atomic symbol of the chiral atom, with (`@`) indicating that the following neighbors are listed anticlockwise and (`@@`) indicating that the neighbors are listed clockwise. If a chiral specification is not present for a chiral atom, its chirality is implicitly not specified. Examples:

- `N[C@H](C)C(=O)O` - D-Alanine
- `N[C@@H](C(=O)O)C` - still D-Alanine, with different atom ordering, but also the opposite stereochemistry notation
- `[C@H](F)(Cl)(Br)` - a tetrahedral carbon with four different groups
- `C(F)(Cl)Br` - same molecule, but with chirality implicitly unspecified

The general chiral specification used in SMILES consists of the (`@`) symbol, followed by a two-letter chiral class indicator, followed by a numerical chiral permutation designator. The default class for four connections is tetrahedral, "`TH`". Most chiralities have more than two possible choices and the choices are assigned from a table numerically. The "`@`" and "`@@`" notations used above are shortcuts for the full specifications "`@TH1`" and "`@TH2`". The square-planar class symbol is "`SP`". Possible values are "`@SP1`", "`@SP2`" and "`@SP3`". The trigonal-bipyramidal class symbol is "`TB`". This is the default chiral class for degree five. With 20 possible stereoisomers, the values are "`@TB1`" through "`@TB20`". The octahedral class symbol is "`OH`". This is the default chiral class for degree six. With 30 possible stereoisomers, the values are "`@OH1`" through "`@OH30`".

2.3.2.3 Disconnected Structures

Most common cases of disconnected structures encountered in cheminformatics are salts and mixtures, but there are other examples, such as complexes. All of these are exceptionally challenging for QSAR modelling. The main issue is that the disconnected nature of the chemical structure "breaks" many sophisticated molecular descriptor calculations since many descriptors can't be explicitly calculated for the full, disconnected system. There are three common approaches to dealing with this issue. First, only the largest (usually by molecular weight) substructure can be used in the whole modelling exercise. This is often the preferred method and the one used in aiQSAR. Second, all descriptors that can't be calculated for these systems are simply ignored and not used for model building. Third, all descriptors are calculated for all substructures and the individual decision regarding which to use is done on a case-by-case basis. Regardless, the SMILES notation remains simple, with disconnected segments simply separated by a period (`.`). Examples:

- `[Cu+2].[O-]S(=O)(=O)[O-]` - copper(II) sulfate
- `[Na+].[Cl-]` - sodium chloride

2.3.2.4 Canonicalization

One of the fundamental properties of SMILES is that there are usually many valid SMILES strings that represent the same molecule. For example, ethanol can be written as:

- `CCO` - carbon-carbon-oxygen
- `OCC` - oxygen-carbon-carbon
- `C(O)C` - carbon branched with oxygen and carbon
- `C(C)O` - carbon branched with carbon and oxygen

All of these are equally valid SMILES, differing only in the location of the molecule that is chosen for the string to start. Similarly, in cyclic structures, the ring closure can begin at any atom and in branched structures, the branches can be listed in any order. To address this ambiguity, a canonicalization algorithm can be arranged that assigns a unique

SMILES labeling to almost every molecular structure. There is no industry-wide standard for this, but there is a widely, although not universally accepted method [56], which is also used in aiQSAR.

2.3.3 Molecular Fingerprints

Molecular fingerprints are a representation scheme that enables the encoding of chemical structure into numerical vectors suitable for computational analysis. Unlike molecular descriptors, which encode continuous, complex variables like physicochemical properties, fingerprints are almost exclusively binary vectors where each bit represents the presence of a particular structure, pattern, or other feature within a molecule [57]. Only binary fingerprints are discussed in this section. Mathematically, fingerprint is represented as a binary vector $\mathrm{f}(M) = [f_1, f_2, \ldots, f_n]$, with $f_i \in \{0,1\}$ and with each element representing the presence or absence of a particular structural feature or pattern. The dimensionality n varies greatly depending on the specific fingerprint, with ranges from under one hundred to several thousand bits in length. The process of fingerprint generation involves systematic enumeration of structural features, ranging from simple atomic composition, to complex topological patterns and molecular fragments, followed by a mapping procedure that assigns specific bit to each identified feature or pattern.

A fundamental distinction exists between predefined and algorithmic fingerprints. Predefined fingerprints, sometimes called dictionary-based, employ a fixed "dictionary" of chemically meaningful structural features. Each bit position corresponds to a specific structural pattern defined *a priori*. This approach offers the advantage of chemical interpretability since the presence of any particular bit directly identifies the corresponding structural feature. PubChem is an example of a frequently used predefined fingerprint of length 881 [58]. As the name suggests, this fingerprint is used for similarity assessment when querying the PubChem database [59]. It is also used in the aiQSAR software.

Algorithmic fingerprints, sometimes called path-based, employ hashing functions to map identified structural features to specific bit positions in a binary string of fixed length. However, since the theoretical space of possible molecular fragments vastly exceeds practical fingerprint lengths, multiple distinct fragments must be mapped to identical bit positions. Formally, a hash collision occurs when:

$$h(f_i) = h(f_j) \text{ for } f_i \neq f_j \tag{2.33}$$

Algorithmic fingerprints sacrifice explicit chemical interpretability, but offer instead the superior coverage of chemical space, without requiring predetermined definition of relevant features and, consequently, readily scale to large and diverse chemical datasets. Extended-connectivity fingerprints (Extended) represent a family of algorithmic fingerprints [60]. The 1024-bit Extended variant is used in the aiQSAR software.

2.3.3.1 PubChem Fingerprints

The PubChem fingerprint is a publicly available, predefined structural key system developed by the National Institutes of Health (NIH) as part of the PubChem chemical database. The PubChem fingerprint comprises 881 bits, each corresponding to a specific structural feature. spanning a wide range of atom types, bond types, connectivity patterns, ring systems and aromaticity patterns (Table 2.7).

Table 2.7: PubChem fingerprint bit examples. Specific bit positions were chosen semi-randomly to include all structure types. Descriptions are directly imported from the PubChem documentation [58].

Position	Bit Description
15	>= 2 N
52	>= 1 Ca
168	>= 4 unsaturated non-aromatic carbon-only ring size 5
185	>= 2 any ring size 6
254	>= 1 unsaturated non-aromatic heteroatom-containing ring size 10
421	C=S
441	C(-C)(=C)
520	C=C-C:C
572	N:C:N:C
657	C-N-C:C-C
708	C-C(C)-C-C
766	Sc1c(S)cccc1
800	CC1CC(N)CCC1
873	SC1C(Cl)CCC1
880	BrC1C(Br)CCC1

An advantage of PubChem fingerprints is in their explicit chemical interpretability. Because each bit position corresponds to a specific, predefined structural feature, identification of which bits are active in a particular compound allows for a straightforward chemical interpretation. However, the fixed dictionary of 881 features implies that novel structural patterns not encompassed within the predefined set cannot be effectively analyzed, potentially limiting performance for some non-standard chemical spaces.

2.3.3.2 Extended Fingerprints

Extended fingerprints are a class of algorithmic fingerprints that capture molecular connectivity relationships by considering atomic environments defined to a specific depth. These fingerprints incorporate stereochemical and circular structures in addition to the standard connectivity information. They are generated through a systematic procedure that identifies all linear paths in the molecular structure up to a defined maximum length and encodes information about atoms, bonds and branching patterns encountered along each path.

The fingerprint generation process begins with the assignment of an initial integer identifier to each non-hydrogen atom of the input molecule. This identifier captures some local information about the corresponding atom in such a way that various atom properties are packed into a single integer value using a hash function. Hashing function is a function whose output is a specific value of a fixed length, for any different input of an arbitrary length. After that, a number of iterations are performed to combine the initial atom identifiers with identifiers of neighboring atoms until a specified depth is reached. Each iteration captures larger and larger neighborhoods around each atom, which are then encoded into single integer values using the chosen hashing method and these identifiers are collected into a list. Finally, all generated hashes are represented as a long string of bits, which is then folded to the specified length using the “OR” operator (Figure 2.6) [61].

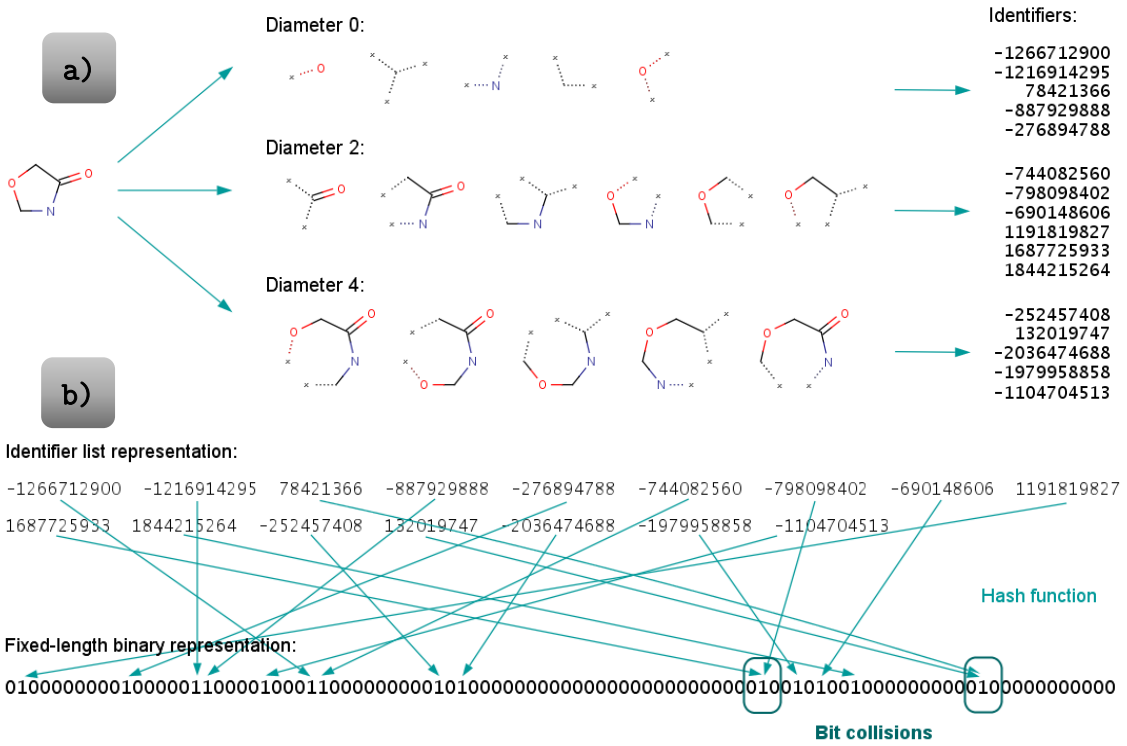


Figure 2.6: Extended fingerprint generation [61]. **a)** Unique identifiers for each molecular connectivity segment are obtained through a hashing function. **b)** All hashes are represented in a string of a fixed length by folding the full string through an "OR" operator. Effectively, ignoring bit collisions.

The theoretical advantage of Extended fingerprints is that they dynamically adapt to any chemical environment by not having a predefined feature set, which naturally limits molecular structures that can be captured. The main limitation of this scheme are bit collisions that occur when the sparse hashing output is folded. In that case, the interpretation is less direct since identifying a bit that the chemicals share doesn't immediately determine which structural features align. Additionally, these fingerprints may emphasize local structural features over larger-scale topological patterns, which could be captured by a broadly defined bit in a predefined fingerprint scheme (for example, PubChem bit 168: ">= 4 unsaturated non-aromatic carbon-only ring size 5").

In the aiQSAR software, Extended fingerprint with depth 6 and 1,024-bit resolution is used to complement the PubChem fingerprint and ensure the representation of both predefined and algorithmic fingerprint categories. While PubChem fingerprints emphasize chemically defined functional groups and scaffolds, extended fingerprints emphasize overall molecular connectivity and topology. Empirical evaluation demonstrated that using both fingerprint types simultaneously improves robustness of the aiQSAR methodology by ensuring that all training set compounds that have the potential of positively contributing to the model construction get included in the first step of the local group selection algorithm (see Section 3.1.2.1.4).

2.3.3.3 Other Fingerprint Variants

Beyond the two fingerprints selected for inclusion in the aiQSAR software, there are many other fingerprint schemes, some of which were tested during the aiQSAR methodology development. Other commonly used fingerprint variants are discussed here (Table 2.8).

Table 2.8: Some commonly used molecular fingerprint variants.

Fingerprint	Type	Bits	Description
MACCS Keys	Dictionary-based	166	Predefined structural keys developed by MDL Information Systems
Topological Torsion	Path-based	1,024 (default)	Linear topological paths of defined length, default 4 bonds
RDKit Fingerprint	Path-based	2,048 (default)	Exhaustive enumeration of all molecular subgraphs, up to 7 atoms by default
2D Pharmacophore Fingerprints	Pharmacophore	1,024 (default)	Identifies pharmacophoric features and from them calculates inter-feature topological distances
3D Pharmacophore Fingerprints	Pharmacophore	1,024 (default)	Extends 2D pharmacophore fingerprints to 3D space by also considering geometric pharmacophoric features
Klekota-Roth	Dictionary-based	4,860	Comprehensive fingerprints encoding a large set of predefined chemical patterns and functional groups
MinHash Fingerprint	Locality-sensitive hashing	2,048 (default)	Uses locality-sensitive hashing of molecular substructures
MAP4	Atom pair-based hashing	2,048 (default)	Multiresolution hashing of atom pair environments to fixed-length bit vectors

The MACCS key system is a predefined, dictionary-based fingerprint developed in the 1980s. It is one of the most historically validated fingerprint approaches. The system comprises 166 predefined structural features, each corresponding to a specific structural pattern that was selected based on developer's extensive experience in pharmaceutical research. The chemical features encompassed by MACCS keys include simple elemental composition, specific functional groups, ring systems with their aromaticity patterns and more complex topological features such as specific substitution patterns and connectivity relationships [62].

Topological torsion fingerprints enumerate all linear paths within a specified maximum length and encode the atom types, bond types and branching patterns for each path. These path descriptions are then hashed into a fixed-length binary vector, similar to Extended fingerprints but without ring closure considerations [63].

The RDKit fingerprint is another path-based implementation, available within the RDKit cheminformatics toolkit. It hashes the exhaustive enumeration of molecular subgraphs up to a specified maximum size, with default value of 7 atoms, including all linear and cyclic subgraph patterns and with aromaticity considered [64].

Pharmacophore-based fingerprints represent a hybrid method between predefined and algorithmic approaches. 2D pharmacophore fingerprints encode various predefined pharmacophoric features within a molecule, such as hydrogen bond donor, hydrogen bond acceptor, charge, hydrophobicity and aromaticity. Then, the topological distances between

all pairs of features are calculated, similar to the algorithmic, hash-based method. The resulting fingerprint is potentially more relevant to biological activity than purely topological information. The 3D extension of pharmacophore fingerprints further incorporates spatial geometric information about the pharmacophoric features. Both types are implemented in the RDKit cheminformatics toolkit [64].

The Klekota-Roth fingerprint is an extensive, dictionary-based fingerprint encoding 4,860 predefined substructure patterns. This much larger set of features enables comprehensive coverage of chemical space, enabling detailed chemical interpretation and discrimination of subtle structural differences. However, the large dimensionality results in a very sparse fingerprint where most bits are zero [65].

MinHash fingerprints are an algorithmic variant that applies locality-sensitive hashing concepts to molecular fingerprinting. The idea behind this approach is to approximate compact fixed-length fingerprints with the property that similar molecules end up with high probability of having similar MinHash fingerprints [66]. Full fingerprint is never explicitly generated. In a regular hashing algorithm, there is no guarantee that similar structural patterns will end up having similar hashes. This is due to the nature of the hashing function, as well as to the subsequent folding of the generated sparse string, which ignores collisions. A MinHash algorithm works in a probabilistic manner where the entire fingerprint is never explicitly calculated for every chemical. Instead, the similarity between chemicals in a dataset is rapidly approximated and iteratively refined. This fingerprinting approach is used exclusively in the very large databases where explicit fingerprint calculations, as well as the subsequent similarity comparisons, are not computationally feasible.

MAP4 is another hybrid fingerprint variant that combines predefined atom pairings as structural features with hashing algorithmic step in which all features up to a given depth are compared. Specifically, the circular substructures with radii of 1 and 2 bonds around each atom in an atom-pair are combined with the topological distance separating the two central atoms. These so-called atom pairs are hashed and the resulting set of hashes is MinHashed to form the MAP4 fingerprint [67].

2.3.4 Chemical Similarity

Chemical similarity is an important concept in cheminformatics that quantifies the degree to which two molecules share structural features and properties [68]. This idea forms the theoretical framework for various applications in drug discovery, virtual screening, regulatory hazard assessment and QSAR more broadly. Despite its widespread application, chemical similarity remains a complex topic that requires careful consideration of different approaches and their respective limitations [69].

2.3.4.1 Descriptor-Based Approaches

Distance-based approaches to chemical similarity measure the multidimensional distance between compound representations in descriptor space. When comparing molecular structures based on continuous descriptor vectors, several different metrics can be employed.

The Euclidean distance measures the distance between two points in n-dimensional descriptor space:

$$d_{\text{Euclidean}}(A, B) = \sqrt{\sum_{i=1}^{n} (x_{A,i} - x_{B,i})^2} \tag{2.34}$$

where $x_{A,i}$ and $x_{B,i}$ are the descriptor values for molecules A and B respectively. Euclidean distance is a straightforward and natural metric. However, the high dimensionality typical of descriptor spaces creates challenges for Euclidean distance metric. High-dimensional spaces exhibit the "curse of dimensionality" where distances between most point pairs become approximately equal, reducing discriminative power and giving an increased weighting to the outliers, which contribute disproportionally. The Manhattan distance measures the sum of absolute differences between descriptor vectors:

$$d_{\text{Manhattan}}(A,B) = \sum_{i=1}^{n} |x_{A,i} - x_{B,i}| \tag{2.35}$$

and is less sensitive to extreme values in individual descriptors.

2.3.4.2 Fingerprint-Based Approaches

Molecular fingerprints are the most widely used representation for similarity assessment. Fingerprints encode molecules into high-dimensional binary vectors. For this type of representation, there are various similarity metrics that can be easily evaluated. For two molecules A and B with binary fingerprints f_A and f_B, the following values can be defined:

- N_a = number of 1 bits in fingerprint f_A
- N_b = number of 1 bits in fingerprint f_B
- N_{ab} = number of 1-bit positions for both f_A and f_B (intersection)
- $N_{a\cup b}$ = number of 1-bit positions for at least one of f_A or f_B (union)

The Tanimoto coefficient is one of the most widely used similarity measures in cheminformatics [69]. It calculates the ratio of common features to the total number of features present in either molecule:

$$T_c(A,B) = \frac{N_{ab}}{N_{a\cup b}} = \frac{N_{ab}}{N_a + N_b - N_{ab}} \tag{2.36}$$

The Tanimoto coefficient ranges from 0 to 1. Importantly, it ignores the shared absence of features, bits where both molecules are 0.

There are many other similarity metrics for binary fingerprints and a few commonly used approaches are listed here. The Dice coefficient measures structural similarity as:

$$D_c(A,B) = \frac{2N_{ab}}{N_a + N_b} \tag{2.37}$$

while the Cosine coefficient normalizes the bit vector representations before computing their dot product:

$$C_c(A,B) = \frac{f_A \cdot f_B}{|f_A||f_B|} = \frac{N_{ab}}{\sqrt{N_a N_b}} \tag{2.38}$$

Finally, an interesting approach is the Braun-Blanquet coefficient, which computes the ratio of shared features to the total number of features in the molecule that has more of them:

$$B_c(A,B) = \frac{N_{ab}}{\max(N_a, N_b)} \tag{2.39}$$

2.3.4.3 Maximum Common Substructure

Maximum common substructure (MCS) approach provides an intuitive, graph-theoretic method for assessing molecular similarity. The MCS of two molecules is defined as the largest substructure, called subgraph, common to both molecules [70]. The usual requirement is for atom types, bond types and connectivity to be preserved. However, some requirements may be relaxed, or even further constrained when assessing the MCS. Options include whether to consider bond types, charges and if the disconnected structures should be allowed. Additionally, the algorithm may allow some aspects of rings to be broken when assessing the MCS (Figure 2.7).

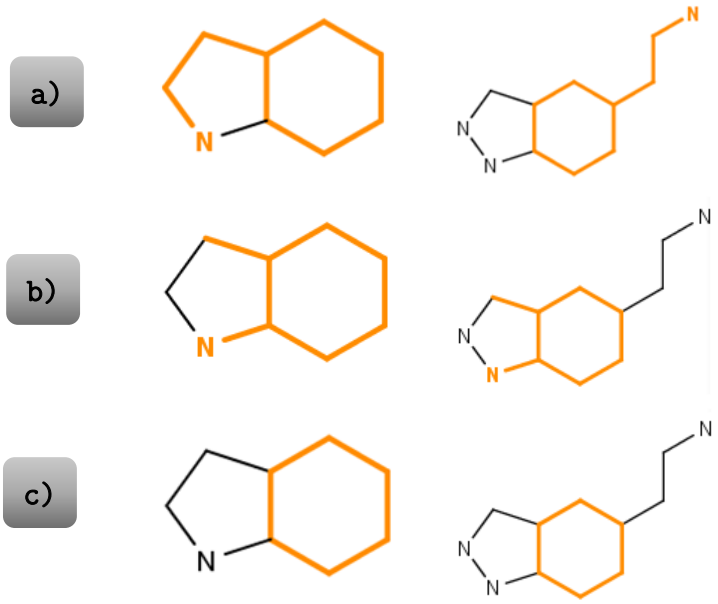


Figure 2.7: Different MCS ring treatment approaches [70]. **a)** Ring is completely ignored. **b)** Two bonds match only if both are in rings or both are in chains. **c)** Rings should not be broken.

The main advantage of the MCS approach is that it is highly intuitive. This method directly corresponds to the simple notion that structurally similar chemicals "look" the same, have the same core segment. Additionally, local structural similarities, where molecules share large substructural motifs, despite global dissimilarity, get naturally captured by the MCS. The substructural motif might correspond to the shared pharmacophoric core of bioactive molecules. In these cases, global fingerprint-, or descriptor-based approaches may get overwhelmed by the many additional molecular segments, unrelated to the activity, effectively "missing the forest for the trees".

However, MCS is computationally expensive, with algorithms typically exhibiting exponential computational complexity with respect to the number of bonds. Additionally, and most importantly, the pharmacophoric core concept may also work in reverse, where

the MCS misses molecular segments and functional groups relevant for chemical activity by focusing on a structurally large, uninformative section of the molecule. For this reason, it's much more common to employ MCS as a complement to other similarity assessment approaches discussed previously, rather than as a stand-alone method.

2.3.5 Applications

Cheminformatics applications span many research areas, both in chemistry, but also more broadly. In pharmaceutical research, cheminformatics supports drug discovery and development. Virtual screening of large chemical libraries helps to identify lead compounds with favorable physicochemical and other relevant properties. Early in the drug discovery pipeline, cheminformatics models are used to predict properties such as protein binding, bioavailability, metabolic stability and toxicity. These predictions guide medicinal chemistry efforts in optimizing drug candidates, before expensive preclinical testing, thus substantially accelerating development and reducing associated costs.

Regulatory agencies and industry employ cheminformatics models for environmental and toxicological hazard assessment. QSAR is used to predict a wide range of toxicologically relevant endpoints including acute toxicity, bioaccumulation potential, mutagenicity and other relevant endpoints. The prohibitive cost and ethical constraints of *in vivo* toxicity testing have created a substantial regulatory demand for computational alternatives. Here, QSAR and other cheminformatic approaches provide rapid, cost-effective predictions that complement *in vitro* and *in vivo* data. For example, the REACH regulation in the EU explicitly recognizes QSAR models as valid alternatives to animal testing under specific conditions, providing a regulatory mandate for their development and application (see Section 2.1.4).

Many regulatory frameworks, including REACH, recognize the potential of read-across approaches to assess chemical similarity and identify chemical analogues. These are expected to share mechanisms of toxicity with structurally similar compounds that have available toxicological data. Cheminformatics methods are used to assess structural similarity and identify the appropriate analogues. Chemical fingerprints, for example, can be rapidly employed for database searching and analogue selection in support of read-across assessments.

QSAR cheminformatics models are extensively used to predict relevant physicochemical properties for novel, untested and yet un-synthesized chemicals. Values such as water solubility and partition coefficients can be calculated with a high degree of certainty and be confidently used as inputs for subsequent experimental or synthetic protocols. This application has expanded substantially with the growth of high-throughput synthesis and combinatorial chemistry, where synthesis rates often outpace the capacity for experimental property characterization.

Furthermore, in chemical synthesis, cheminformatics is also used to design chemical libraries with optimized diversity and with relevant coverage of chemical space. Diversity analysis ensures that candidate compounds explore different regions of chemical space and target different mechanistic hypotheses. Conversely, clustering within the library design can focus compound selection around specific chemical scaffolds or property ranges relevant to the target.

Finally, cheminformatics is essential in chemical database development, curation and standardization where it provides an infrastructure for organizing, systematizing and retrieving chemical information at scale. Large public and proprietary chemical databases ("big data" in chemistry) require these systematic methods for structure encoding, validation, duplicate detection and overall quality control.

2.4 Statistics

The development and validation of QSAR models fundamentally rely on statistical principles and quantitative performance assessment. The choice of appropriate statistical metrics, validation techniques and hypothesis testing procedures directly determines the reliability, reproducibility and regulatory acceptability of predictive models. This section provides a comprehensive review of statistical methodologies commonly applied to evaluate cheminformatics models, with particular emphasis on classification and regression frameworks relevant to local QSAR applications. This establishes the mathematical foundations necessary for understanding the evaluation metrics employed throughout this thesis [71].

2.4.1 Statistics in Cheminformatics

Statistical analysis is used to assess the QSAR performance on all endpoint types. Unlike in many biological disciplines where experimental uncertainty dominates and sample sizes are inherently limited; cheminformatics datasets frequently contain thousands of compounds with associated experimental measurements. This abundance of data creates distinct opportunities, but also some statistical challenges. The primary objective of statistical analysis in cheminformatics is to estimate predictive performance on new, unseen data. The main method of achieving this is by in-depth exploratory data analysis of the training set statistical metrics [72].

Binary classification tasks, such as distinguishing toxic from non-toxic chemicals or predicting the presence of a specific pharmacological activity, represent one of the most common applications of cheminformatics models. These tasks produce predictions that fall into one of two categories. The reliability of binary classification models cannot be adequately summarized by a single metric and different applications may prioritize different aspects of model performance. Regulatory toxicology contexts, for instance, may prioritize the minimization of false negatives, as such errors carry substantial public health consequences. Conversely, high-throughput virtual screening applications may tolerate higher false positive rates in exchange for improved ability to correctly identify as many active compounds as possible since false positives in this context don't pose any significant downside.

Multi-class classification problems extend the binary classification framework to scenarios involving three or more discrete categories. Toxicological endpoints such as the EPA hazard categories, or various mode-of-action classifications represent multi-class problems where compounds must be assigned to one of several distinct categories.

Regression tasks, where models predict continuous values such as lethal dose, LD_{50} or the partition coefficients, LogP, require distinct statistical metrics from classification. Regression models provide point estimates of the predicted property value and their performance is quantified through measurement of the discrepancy between predicted and experimental values. The mathematical framework differs substantially from classification metrics since regression inherently involves measurement of continuous error distributions rather than discrete categorical accuracy.

2.4.2 Classification Performance Metrics

There are many different performance metrics that are applicable to classification endpoints. Additionally, there has been a proliferation of various statistical measures in machine learning and cheminformatics literature in general, which has created some challenges. The selection of which metrics to report often appears arbitrary and inconsistent

across studies, potentially reflecting individual preference or disciplinary convention. However, there are some essential metrics that are ubiquitous. For any classification training set, an indispensable statistical output is the confusion matrix, showing the distribution of counts between predicted and experimental categories. Beyond the confusion matrix, many derived statistics can be calculated (Figure 2.8), with various pros and cons. This section provides detailed discussion of the commonly reported metrics, their mathematical definitions, the appropriate contexts of application and their inherent limitations, with the emphasis on metrics selected for performance assessment and implementation into the aiQSAR software.

		Predicted condition			
	Total population $= P + N$	**Predicted positive**	**Predicted negative**	Informedness, bookmaker informedness (BM) $= TPR + TNR - 1$	Prevalence threshold (PT) $= \frac{\sqrt{TPR \times FPR} - FPR}{TPR - FPR}$
Actual condition	**Real Positive (P)** [a]	**True positive** (TP), hit[b]	**False negative** (FN), miss, underestimation	True positive rate (TPR), recall, sensitivity (SEN), probability of detection, hit rate, power $= \frac{TP}{P} = 1 - FNR$	False negative rate (FNR), miss rate type II error [c] $= \frac{FN}{P} = 1 - TPR$
	Real Negative (N) [d]	**False positive** (FP), false alarm, overestimation	**True negative** (TN), correct rejection[e]	False positive rate (FPR), probability of false alarm, fall-out type I error [f] $= \frac{FP}{N} = 1 - TNR$	True negative rate (TNR), specificity (SPC), selectivity $= \frac{TN}{N} = 1 - FPR$
	Prevalence $= \frac{P}{P + N}$	Positive predictive value (PPV), precision $= \frac{TP}{TP + FP} = 1 - FDR$	False omission rate (FOR) $= \frac{FN}{TN + FN} = 1 - NPV$	Positive likelihood ratio (LR+) $= \frac{TPR}{FPR}$	Negative likelihood ratio (LR−) $= \frac{FNR}{TNR}$
	Accuracy (ACC) $= \frac{TP + TN}{P + N}$	False discovery rate (FDR) $= \frac{FP}{TP + FP} = 1 - PPV$	Negative predictive value (NPV) $= \frac{TN}{TN + FN} = 1 - FOR$	Markedness (MK), deltaP (Δp) $= PPV + NPV - 1$	Diagnostic odds ratio (DOR) $= \frac{LR+}{LR-}$
	Balanced accuracy (BA) $= \frac{TPR + TNR}{2}$	F_1 score $= \frac{2\,PPV \times TPR}{PPV + TPR} = \frac{2\,TP}{2\,TP + FP + FN}$	Fowlkes–Mallows index (FM) $= \sqrt{PPV \times TPR}$	*phi* or Matthews correlation coefficient (MCC) $= \sqrt{TPR \times TNR \times PPV \times NPV} - \sqrt{FNR \times FPR \times FOR \times FDR}$	Threat score (TS), critical success index (CSI), Jaccard index $= \frac{TP}{TP + FN + FP}$

Figure 2.8: Confusion matrix and derived statistics for the binary classification dataset [73]. The figure showcases plethora of various performance measures that can be derived from the matrix.

2.4.2.1 Accuracy

Accuracy is the simplest and most intuitive classification metric that calculates the proportion of all predictions that are correct. For a binary classification task with:

- N total samples
- T_p true positives - correctly predicted positive instances
- T_n true negatives - correctly predicted negative instances
- F_p false positives - negative instances incorrectly predicted as positive
- and F_n false negatives - positive instances incorrectly predicted as negative,

accuracy is defined as:

$$\text{Acc} = \frac{T_p + T_n}{T_p + T_n + F_p + F_n} = \frac{T_p + T_n}{N} \tag{2.40}$$

Despite its simplicity and intuitive interpretation, accuracy suffers from substantial limitations when applied to imbalanced datasets. In a binary classification task where 95% of compounds in the validation set are non-toxic and 5% are toxic, a trivial model that predicts every compound to be non-toxic would achieve 95% accuracy while providing no useful information. This makes accuracy inadequate for most real use cases in cheminformatics. More appropriate metrics are the ones that account for class distribution and value the predictions differently based on the percentage of chemicals in corresponding classes. However, as the simplest and most intuitive performance measure, accuracy is often reported alongside other more appropriate metrics, which is the case in the aiQSAR software as well.

Accuracy can be easily extended to multi-class classification. In that case, Eq. (2.40) would simply include correct classifications from all categories in the numerator.

2.4.2.2 Balanced Accuracy

Balanced accuracy (BA) computes the arithmetic mean of sensitivity for each class independently, thereby providing equal weight to correct predictions in each class regardless of class frequencies. In this context, the simple accuracy has an effect of giving each class the weight proportional to its representation in the dataset. For binary classification, BA is defined as:

$$\text{BA} = \frac{\text{Sensitivity} + \text{Specificity}}{2} = \frac{1}{2}\left(\frac{T_p}{T_p + F_n} + \frac{T_n}{T_n + F_p}\right) \tag{2.41}$$

For the highly imbalanced toxicity example presented previously, a model predicting all compounds as non-toxic would achieve a balanced accuracy of only 50%, correctly reflecting that the model provides no ability to distinguish the minority class [74].

BA directly extends to multi-class scenarios by averaging the sensitivity for each class:

$$\text{BA}_{\text{multi-class}} = \frac{1}{K}\sum_{i=1}^{K} \text{Sensitivity}_i = \frac{1}{K}\sum_{i=1}^{K} \frac{T_p^{(i)}}{T_p^{(i)} + F_n^{(i)}} \tag{2.42}$$

where K is the number of classes and $T_p^{(i)}$ and $F_n^{(i)}$ are true positive and false negative rates for class i, respectively. This way, BA remains unaffected by class imbalances even in the multi-class calculations and provides equal consideration to both the more and the less represented classes.

Despite its advantages over simple accuracy, BA also has significant limitations. The arithmetic averaging approach treats all classes as equally important, which may not reflect the actual consequences of misclassifications in applied contexts. More importantly, BA

implicitly assumes the distribution of classes for the predictor that are experimentally observed and, in cases where class frequencies of a predictor diverge significantly from the observed frequencies, BA can be quite misleading. Considering a task where there are 20% positives, with 80% negatives, a model predicting 5% positives, but correctly classifying all such predictions is very useful and performs significantly better than random. Yet, this example yields BA value of only 0.625, indicating just slightly above average performance (Figure 2.9a). The same BA outcome can be achieved with almost completely random model that has no predictive capacity (Figure 2.9b).

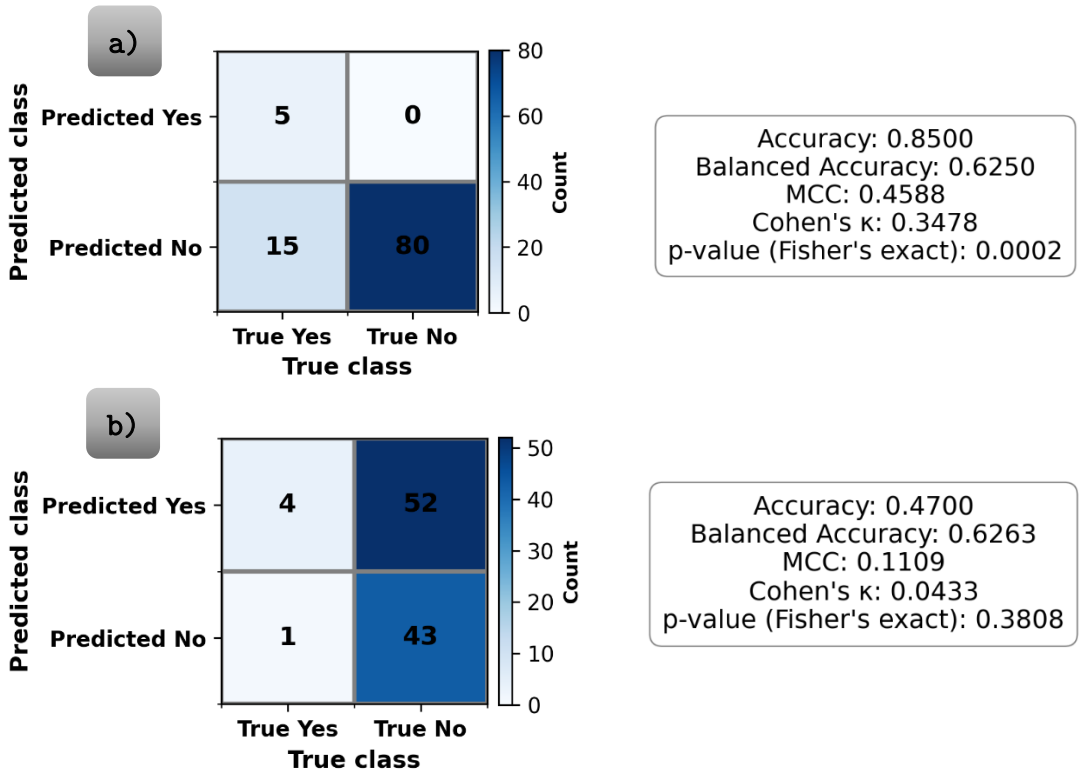


Figure 2.9: Balanced accuracy statistic limitation. Two models achieve approximately the same balanced accuracy. While model **a)** performs very well and has significant predictive capacity, model **b)** is almost no better than random guessing.

The issue in these examples is that the number of predicted values for the two models doesn't match the underlying distribution. This can frequently happen since experimentally there is often no expectation for the new, untested samples to completely match the distribution of training set values. In the QSAR context, example would be a drug discovery task where the activity is quite rare and a randomly constructed training set would have only a small fraction of positive chemicals. Yet, new chemicals, experimentally assessed for some activity-related endpoint would be expected to skew towards positive chemicals, similar to the example in Figure 2.9b, where the model might classify more chemicals as positives than the baseline distribution would indicate. Overall, BA is a good statistical metric when the distribution of predicted values matches the distribution of observed values, regardless of whether the predictions themselves are correct or incorrect. If this is not the case, BA can severely overestimate or underestimate the performance and

more appropriate metrics, the ones that take into account the distribution of predicted values, need to be assessed.

2.4.2.3 Matthews Correlation Coefficient

The Matthews correlation coefficient (MCC) provides a single metric that accounts for all elements of the confusion matrix. Originally developed for assessing the quality of protein structure predictions [75], MCC ranges from -1 for perfect negative correlation, through 0 indicating random predictions, to 1 for fully correct prediction. For binary classification, MCC is defined as:

$$\mathrm{MCC} = \frac{T_p \cdot T_n - F_p \cdot F_n}{\sqrt{(T_p + F_p)(T_p + F_n)(T_n + F_p)(T_n + F_n)}} \tag{2.43}$$

The numerator aims to capture the difference between correct and incorrect predictions, while the denominator normalizes this difference by the geometric mean of the marginal totals. Marginal totals are total positives, total negatives, predicted positives and predicted negatives, the row and column sums of the confusion matrix. Notably, if any of these four totals equal zero, MCC is undefined. The MCC penalizes both false positives and false negatives symmetrically and it accounts for class imbalance.

MCC is closely related to the χ^2 statistic for a 2×2 contingency table:

$$|MCC| = \sqrt{\frac{\chi^2}{n}} \tag{2.44}$$

where n is the total number of observations. MCC is generally regarded as one of the best measures for assessing the binary classification performance [76].

The primary disadvantage of MCC is that it can't be readily extended to a multi-class classification. Since there isn't a direct way to generalize it, several different formulations exist [77]. A commonly applied definition for multi-class MCC is:

$$\mathrm{MCC}_{\text{multi-class}} = \frac{\mathrm{cov}(X, Y)}{\sqrt{\mathrm{var}(X) \cdot \mathrm{var}(Y)}} \tag{2.45}$$

where X is the vector of true class assignments and Y is the vector of predicted class assignments. This formulation is sometimes also called the R_k statistic and can be written in its fully expanded form [78]:

$$R_k = \frac{\sum_{i,j=1}^{K} C_{ij}(n_{i\cdot} n_{\cdot j} - n C_{ij})}{\sqrt{\sum_{i=1}^{K} n_{i\cdot}(n - n_{i\cdot}) \sum_{j=1}^{K} n_{\cdot j}(n - n_{\cdot j})}} \tag{2.46}$$

where:

- C_{ij} is a confusion matrix element for true class i, predicted class j
- $n_{i\cdot}$ is the number of samples that are true class i
- $n_{\cdot j}$ is the number of samples predicted as class j
- n is the total number of samples
- and K is the number of classes.

The ranges of this MCC are no longer between -1 and 1. Instead, the minimum value is between -1 and 0, depending on the true distribution. The maximum value is always 1. The interpretation of the multi-class MCC in this context is much less straightforward. Additionally, other formulations that emphasize different characteristics of the original MCC also exist making this extension to multi-class problems non-trivial and reducing the overall utility of the statistic.

2.4.2.4 Cohen's Kappa Coefficient

Cohen's kappa (κ, or just kappa) is another widely applied metric for assessing classification performance. Kappa measures the degree to which predictions agree beyond what would be expected by chance, with explicit mathematical accounting for class frequencies, so called "marginal distributions" [79]. Kappa is defined as:

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{2.47}$$

where p_o is the observed agreement and p_e is the expected agreement probability assuming random classification. This elegant definition asks the most relevant question when assessing the predictive capacity of a model: How much better is the observed agreement to the one that would be achieved by randomly selecting categories, given the distributions of predicted and observed values? Same as MCC, kappa ranges from -1 for perfect negative correlation, through 0 indicating random predictions, to 1 for fully correct prediction.

For binary classification, the observed agreement is simply accuracy:

$$p_o = \frac{T_p + T_n}{N} \tag{2.48}$$

and the expected agreement, assuming marginal distributions, is:

$$p_e = \frac{(T_p + F_n)(T_p + F_p)}{N^2} + \frac{(T_n + F_p)(T_n + F_n)}{N^2} \tag{2.49}$$

where the first probability term is the random chance agreement for the positive class, computed as the product of the proportion of actual positives and the proportion of predicted positives, divided by N^2 for normalization. The second term is the analogous probability for the negative class. Expanding the expression, kappa can be written explicitly as:

$$\kappa = \frac{2 \times (T_p \times T_n - F_n \times F_p)}{(T_p + F_p) \times (F_p + T_n) + (T_p + F_n) \times (F_n + T_n)} \tag{2.50}$$

A major advantage of the kappa statistic is its direct applicability to the multi-class classification. Extending kappa to a multi-class task is unambiguous, as it can be directly applied in its original definition, Eq. (2.47). The terms are then expanded appropriately:

$$p_o = \frac{\sum_{i=1}^{k} n_{ii}}{N} \tag{2.51}$$

where n_{ii} is the number of samples correctly classified as category i and N is the total number of samples, with:

$$p_e = \frac{1}{N^2} \sum_{i=1}^{k} n_{i+} \times n_{+i} \tag{2.52}$$

where n_{i+} is total count of observed values of category i and n_{+i} is total count of predicted values of category i.

The main consideration when using kappa is its reliance on marginal distributions. Kappa's baseline, p_e, is often described as "the agreement due to chance", which is only partially correct. The full formulation would be "the agreement that would be expected due to chance, given the quantities specified by the marginal totals". As discussed, this is a desirable property and not accounting for marginal distributions is the main limitation of BA as a metric. In most cases the divergence between marginal totals isn't the main consideration and should be anticipated. Therefore, this should not be considered a limitation of kappa, but it is something to be mindful of when using the statistic and to consider, in addition to the kappa value itself, when assessing model performance.

2.4.3 Regression Performance Metrics

Regression endpoints require quantification of prediction errors on a continuous scale. The corresponding metrics are mathematically more complex, but also more straightforward in interpretation and less ambiguous. The most commonly applied regression metrics quantify the magnitude and the distribution of prediction errors and the proportion of variance explained by the model.

The coefficient of determination (R^2, or sometimes r^2) is the proportion of variance in the experimental values explained by the regression model:

$$R^2 = 1 - \frac{\mathrm{SS}_{res}}{\mathrm{SS}_{tot}} = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - y)^2} \tag{2.53}$$

where y_i is the experimental value for compound i, $\hat{y}_i$ is the predicted value and y is the mean of all experimental values. SS_{res} is the residual sum of squares, prediction errors and SS_{tot} is the total sum of squares. R^2 ranges from 0, meaning that the model explains no variance beyond the mean to 1, where model perfectly predicts all experimental values. The R^2 can be adjusted to penalize model complexity by reducing the core value based on the number of descriptors used by the model:

$$R^2_{adj} = 1 - \frac{(1 - R^2)(N - 1)}{N - K - 1} \tag{2.54}$$

where K is the number of descriptors and N is the sample size. This adjusted R^2 can be used for direct comparison of models with differing numbers of descriptors.

Root Mean Square Error (RMSE) quantifies the average magnitude of prediction errors:

$$\mathrm{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2} \tag{2.55}$$

where y_i is the observed value for compound i, $\hat{y}_i$ is the predicted value and N is the sample size. RMSE is very sensitive to outliers. Large prediction errors contribute disproportionately to the value of the statistic due to the squaring of the terms. Extreme predictions can therefore substantially inflate RMSE, potentially misrepresenting overall model performance [80].

Mean Absolute Error (MAE) is the average absolute magnitude of prediction errors and it provides an alternative to RMSE with reduced sensitivity to outliers:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{2.56}$$

MAE provides a direct measure of typical model performance on individual compounds, which makes it more directly interpretable as compared to RMSE, especially for practitioners unfamiliar with statistical concepts. The downside of MAE is that the absolute value function isn't smooth and for potential downstream analyses, the squaring function used by RMSE has some nicer mathematical properties.

2.4.4 Statistical Hypothesis Testing

Statistical hypothesis testing is a method of statistical inference used to assess whether the data provide sufficient evidence to reject a particular hypothesis. Statistical hypothesis tests begin with the formulation of a null hypothesis (H_0), a statement that no effect exists or that two populations possess identical distributions. An alternative hypothesis (usually, H_1) represents the opposite claim. Importantly, the hypothesis test does not directly assess the probability that the null hypothesis is true. Instead, it computes the probability of observing data as extreme, or more extreme than the data which is actually observed, assuming the null hypothesis is true. This probability is the p-value [81].

For example, in comparing the performance of two independent QSAR models, the null hypothesis might state that the predictive performance of model 1 equals the predictive performance of model 2. If this null hypothesis were true and if the validation datasets of the same size were repeatedly sampled from the underlying population, what proportion of samples would produce differences in observed accuracy (or another selected performance metric) as large, or larger than what was actually observed? This proportion, the p-value, quantifies the rarity of the observed result under the null hypothesis.

2.4.4.1 Parametric and Nonparametric Tests

Hypothesis tests are frequently categorized as parametric or nonparametric. Parametric tests, such as Student's t-test and ANOVA, assume specific data distributions, typically normality and they can provide greater statistical power when their distributional assumptions are satisfied. However, when the data violates these assumptions, parametric tests can produce misleading results. Nonparametric tests, such as the Wilcoxon rank-sum test (also called the Mann-Whitney U test) and the Kruskal-Wallis test, make minimal assumptions on data distribution and, most importantly, operate on ranks rather than on raw values.

In comparing the two independent groups, the Wilcoxon rank-sum test addresses the null hypothesis that these two groups have identical distributions, with the alternative that the distributions differ [82]. The Wilcoxon rank-sum test works by ranking all observations from both groups jointly from smallest to largest and then summing the ranks within each group. The test statistic U is computed as:

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1 \tag{2.57}$$

where n_1 and n_2 are sample sizes and R_1 is the sum of ranks for group 1. For large samples, U approximately follows a normal distribution, which enables the calculation of p-values [83]. Wilcoxon rank-sum test assumes identical distribution shapes with possible location shifts. When this assumption is violated, usually because sample variances are substantially unequal, other statistical tests are more appropriate.

An alternative nonparametric test applicable when the null hypothesis concerns equality of population medians rather than equality of distributions is the Brunner-Munzel (BM) test. The BM test addresses the null hypothesis that, for randomly selected values X from group 1 and Y from group 2, the probability of $X > Y$ equals the probability of $Y > X$. Unlike the Wilcoxon rank-sum test, the BM test does not require equal variances [84].

2.4.4.2 P-Value

The p-value is the probability of observing data as extreme or more extreme than what was observed, given that the null hypothesis is true, $P(\text{data}|H_0)$. Importantly, this definition does not represent the probability that the null hypothesis itself is true, $P(H_0|\text{data})$, which is a different statement and this distinction has important implications for interpretation [85].

A p-value threshold such as 0.05 is a somewhat arbitrary conventional choice. The reasoning underlying this value comes from Fisher's historical suggestion that effects producing p-values below 0.05 could be considered statistically significant. However, this threshold should not be interpreted as providing a sharp boundary between true and false effects, nor does it provide the probability of true findings. An effect producing $p = 0.049$ is not meaningfully different from the one producing $p = 0.051$, yet the conventional framework treats these as fundamentally different [86].

The problem of multiple comparisons presents an important consideration. When many statistical tests are conducted, the probability of observing false positive results increases. If 100 independent hypothesis tests are conducted at $\alpha = 0.05$ threshold, the expected number of false positives is 5, even if all null hypotheses are true. In cheminformatics context where statistical tests are conducted for many endpoints, this problem is often relevant. The simplest and most conservative correction is the Bonferroni method, which simply multiplies each p-value by the number of tests. Depending on the application context, less conservative approaches, such as the Benjamini–Hochberg method, which instead limits the expected proportion of false positives, might be more appropriate [87].

2.4.5 Cross-Validation Approaches

Good model validation constitutes one of the most important aspects of QSAR development. Model performance measured on the same data that was used, in any way, to train the model will always underestimate the error on new, unseen data. This is conventionally called “overfitting” and arises because models employ features of the training data that do not generalize.

To assess the model performance, a cross-validation analysis needs to be correctly executed in order to produce an unbiased performance estimate. The most commonly applied approach is a k-fold cross-validation, where data are divided into k approximately equal subsets (folds). The model is trained k times, each time leaving one subset out as a test set and using the remaining k-1 subsets for training. Performance metrics are computed on each held-out subset and then averaged across all k folds. It is important to ensure that no elements of the left-out subset contribute to the model training procedure. This includes any parameter selection, descriptor filtering, or weight defining steps. All of these can

inadvertently provide the model with information related to the test set and therefore contribute to the overestimation of the model performance. Leave-one-out is a special cross-validation case where k equals the number of samples. This statistically provides the most accurate model performance estimate, but is also the most computationally expensive approach.

Even better performance assessment method is to use a fully separate validation set, entirely unrelated to the dataset used for any model building steps and, ideally, tested externally to generate QSAR predictions, calculate the associated performance metrics and compare them to the ones obtained by the regular cross-validation. Finally, an important limitation of any QSAR model is that its predictions should not be trusted for compounds substantially different from those in the training set, regardless of the performance metrics obtained by cross-validation. This similarity can be defined and quantified in terms of the applicability domain (see Section 2.1.5).

2.5 Transcriptomics

The flow of genetic information in biological systems is described by the central dogma of molecular biology, which establishes the directional pathway of information transfer from deoxyribonucleic acid (DNA) through ribonucleic acid (RNA) to protein. This principle provides the foundation for understanding the transcriptomic analysis [88].

2.5.1 Central Dogma of Molecular Biology

The central dogma postulates that biological information contained within the DNA sequence gets transcribed into RNA, which serves as the template for translation into amino acid sequences of proteins (Figure 2.10). The complete central dogma can be expressed as:

$$\text{DNA} \overset{\text{transcription}}{\rightarrow} \text{RNA} \overset{\text{translation}}{\rightarrow} \text{Protein}$$

This unidirectional flow of information is the basis for genetic expression. However, there are some exceptions to this model: reverse transcription allows for RNA to be transcribed back into DNA under certain circumstances.

DNA is the primary source of genetic information in living organisms. It is composed of two complementary antiparallel strands, each consisting of a sugar-phosphate backbone with nitrogenous bases and is arranged in a double helix structure. The four bases are adenine, guanine, cytosine and thymine and the sequence of these bases encodes the instructions for all biological processes within an organism. The human genome contains approximately 20,000 protein-coding genes, distributed across 23 chromosome pairs and separated by vast regions of non-coding DNA that contributes to gene expression regulation [89].

DNA wraps around histone proteins to form nucleosomes, which are further organized into chromatin. The chromatin structure plays a role in regulating which genes are accessible for transcription. The accessibility of genes is controlled through epigenetic modifications, such as DNA methylation and histone acetylation, which do not change the underlying DNA sequence but influence gene expression patterns [90].

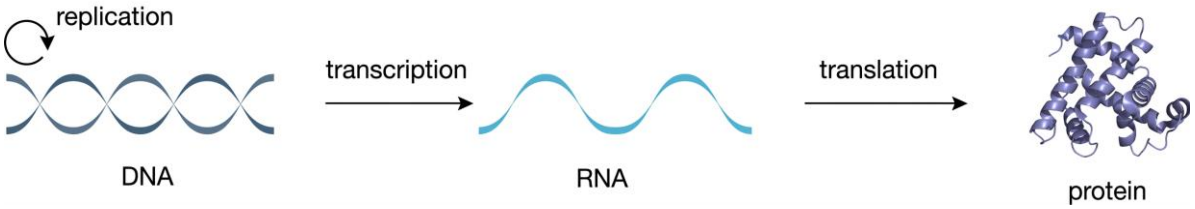


Figure 2.10: Central dogma of molecular biology.

2.5.1.1 Transcription

Transcription is the process by which genetic information stored in DNA gets transcribed into RNA. This process is catalyzed by RNA polymerases, enzymes involved in synthesizing an RNA transcript complementary to the DNA template strand. The general transcription process can be summarized as:

$$\text{DNA} \xrightarrow[\text{RNA Polymerase}]{\text{NTPs}} \text{RNA}$$

However, this is a complex process that involves three distinct RNA polymerases. RNA polymerase I synthesizes ribosomal RNA (rRNA), RNA polymerase II synthesizes messenger RNA (mRNA), as well as many non-coding RNAs and RNA polymerase III synthesizes transfer RNA (tRNA) and other small RNAs.

The primary transcript produced by RNA polymerase II, pre-mRNA, undergoes several post-transcriptional modifications:

- 5' capping, where a 7-methylguanosine cap is added to the 5' end.
- 3' polyadenylation, where a tail of adenine nucleotides is added to the 3' end.
- Splicing, during which introns are removed and exons are ligated together.

The result of these processing steps is mature mRNA, which is exported from the nucleus to the cytoplasm for translation. The transcriptome comprises all RNA molecules present in a cell at a given time. Each cell type maintains a distinct transcriptional profile that determines its molecular identity and functional capabilities through the selective expression of genes.

2.5.1.2 Translation

Translation is the process by which mRNA gets decoded to synthesize proteins. This process occurs on ribosomes, which are composed of ribosomal RNA and ribosomal proteins. Translation involves three main stages: initiation, elongation and termination. The translation process can be summarized as:

$$\text{mRNA} \xrightarrow[\text{Ribosomes + tRNAs}]{\text{Amino Acids}} \text{Protein}$$

During initiation, the ribosome binds to the mRNA at the start codon, with the initiator tRNA carrying methionine. During elongation, successive tRNAs deliver amino acids to the ribosome according to the genetic code, with each three-nucleotide codon specifying a particular amino acid (Figure 2.11). The growing polypeptide chain is progressively extended through peptide bond formation. Termination occurs when the ribosome

encounters a stop codon (UAA, UAG, or UGA), at which point the completed protein is released.

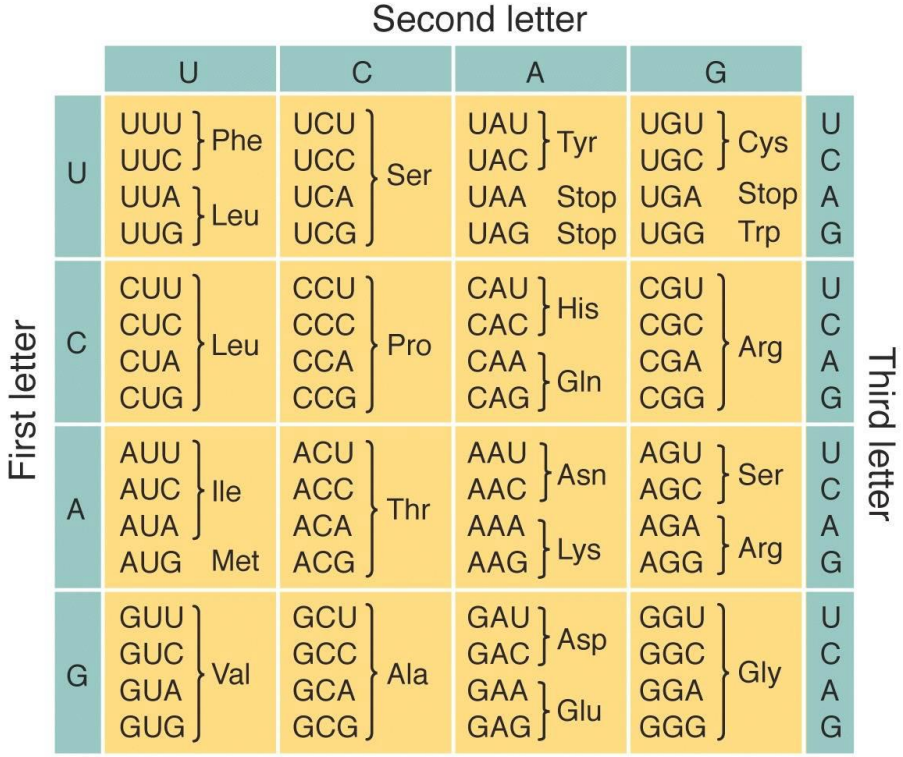


Figure 2.11: Genetic code table.

Post-translational modifications such as phosphorylation, glycosylation, ubiquitination and various others, further alter protein structure and function following translation. These modifications, catalyzed by specialized enzymatic machinery and occurring in specific cellular compartments, are crucial for proper protein folding, localization and activity.

2.5.1.3 Gene Expression Regulation

Any given gene does not produce a single, invariant transcript. Instead, alternative splicing and alternative polyadenylation allow for multiple distinct mRNA molecules to be derived from a single gene. Furthermore, gene expression levels vary dramatically across cell types and conditions. Transcriptomics encompasses the comprehensive study of the transcriptome. It includes the analysis of which genes are active, at what abundance levels they are expressed and how expression patterns change in response to developmental stages or environmental perturbations.

The transcriptome exists in a state of constant flux. Individual RNA molecules have finite lifetimes and are continuously synthesized and degraded. Therefore, the transcriptome represents a temporal snapshot of cellular activity, with composition changing in response to cellular metabolism and external stimuli. The application of transcriptomics to toxicological assessment aims to measure these perturbations in gene expression patterns that are induced by cell exposure to external chemicals. Gene expression measurement at the transcriptomic level requires comprehensive, high-throughput sequencing techniques outlined below.

2.5.2 DNA Sequencing Technologies

DNA sequencing is the process of determining the precise order of nucleotides within a DNA molecule. Sequencing technologies enable this procedure and have been instrumental in recent advances within a wide range of scientific disciplines including biochemical research, medicine and biotechnology. The evolution of sequencing techniques has progressed through multiple generations, each bringing improvements in throughput, speed and cost-effectiveness.

The first practical DNA sequencing method was developed by Frederick Sanger, who received the Nobel Prize in Chemistry in 1980 for this achievement [91]. Sanger sequencing, also known as dideoxy sequencing, uses fluorescently labeled dideoxynucleotides that terminate chain extension when incorporated into a growing DNA strand, creating a population of DNA fragments of varying lengths and each ending at a different position within the original chain. These fragments are separated by capillary electrophoresis, from which the DNA sequence is determined. Sanger sequencing produces highly accurate reads with an error rate of approximately 1 in 10,000 nucleotides. Because of this high accuracy, it's still used even today for clinical applications, variant validation and other use cases that require high fidelity. However, Sanger sequencing is quite limited in throughput and therefore not applicable to large-scale genomic projects where this limitation necessitated the development of higher-throughput alternatives.

The emergence of next-generation sequencing (NGS) technologies fundamentally transformed the landscape of genomic and transcriptomic research by dramatically increasing throughput while at the same time reducing cost per sequenced base. These technologies enabled massively parallel sequencing, allowing for orders of magnitude increase in the amount of DNA material that can be processed. Current NGS platforms can generate hundreds of billions of nucleotides in a single run [92]. The error rate of NGS is higher than Sanger sequencing, ranging from 0.1% to 10% depending on the specific method, but the massive redundancy of coverage compensates for this higher error rate in most applications [93].

2.5.2.1 Pyrosequencing

Pyrosequencing is a DNA sequencing technique that relies on detection of pyrophosphate, $P_2O_7^{4-}$ released during nucleotide incorporation [94]. It is a type of sequencing by synthesis technique, which means that the nucleotide content is determined during the *in vitro* DNA replication process. This is achieved by monitoring pyrophosphate release that accompanies nucleotide incorporation into the growing DNA chain (Figure 2.12).

The process begins by template preparation using the emulsion PCR (emPCR) amplification (Figure 2.12a). A library of target DNA regions is prepared and adaptors containing universal sequences are ligated to their ends. These universal primers enable the amplification of all genetic material with common PCR primers. After ligation, DNA is separated into single strands and captured onto beads under conditions that favor one DNA molecule per bead. Once DNA is captured, emPCR amplification is performed, resulting in multiple copies of DNA fragment from targeted genomic region on each bead. These are deposited into individual PicoTiterPlate wells in which the sequencing is performed [95]. Pyrosequencing is in its essence a bioluminescence method. The release of pyrophosphate during nucleotide incorporation is converted into visible light using series of enzymatic reactions (Figure 2.12b). Different nucleotides, in form of dNTPs are added consecutively to the reaction. If the added nucleotide is complementary to the next unpaired nucleotide in the fragment on the bead, DNA polymerase will incorporate it and thus extend the strand. Simultaneously, pyrophosphate is released. To convert this into

measurable signal, separate beads that have ATP sulfurylase and luciferase attached to them are loaded into wells surrounding the template beads. Additionally, wells are incubated with adenosine-5-phosphosulfate (APS) and luciferin. This produces a series of reactions that end with the light signal:

$$P_2O_7^{4-} + APS \leftrightarrow ATP + SO_4^{2-}$$

$$ATP + luciferin \leftrightarrow luciferyl\ adenylate + P_2O_7^{4-}$$

$$luciferyl\ adenylate + O_2 \rightarrow oxyluciferin + AMP + CO_2 + h\nu$$

Therefore, only wells that contain the bead, which has the nucleotide complementary to the one that is being incorporated will produce this bioluminescence signal, with the intensity proportional to the amount of pyrophosphate available to initiate the reaction. If multiple nucleotides are incorporated, due to consecutive nucleotides in the DNA fragment, more pyrophosphate will be released and the higher intensity measured (Figure 2.12c). This effect imposes accuracy limitations to the method since after certain number of consecutive nucleotides, light intensity differences become indistinguishable and the precise number of nucleotides in that DNA fragment containing the homopolymer can't be determined.

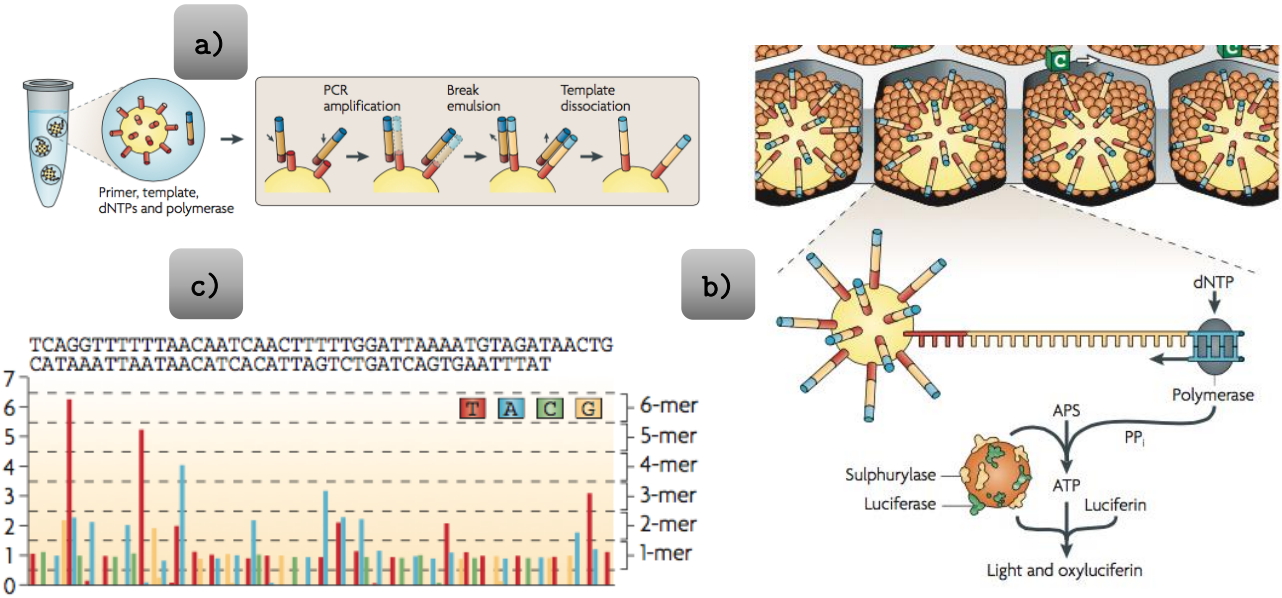


Figure 2.12: Pyrosequencing procedure. **a)** emPCR. **b)** Converting pyrophosphate into the light signal by a series of enzymatic reactions. **c)** Flowgram [96].

2.5.2.2 Illumina Sequencing

Illumina maintains a dominant position in the NGS market and its sequencing by synthesis platform is currently the most widely deployed NGS method globally. The Illumina platform dominates the NGS landscape due to a combination of high throughput, competitive cost per base, short turnaround time and extensive bioinformatics support.

Illumina sequencing approach is based on modified deoxyribonucleotides that can act as reversible terminators during DNA synthesis (Figure 2.13) [97]. First, solid-phase amplification is used to produce copies of targeted DNA fragments. Fragments are ligated with the adaptor sequence and primed to the complementary sequences attached to a glass slide (Figure 2.13a). Then, bridge amplification is performed, where each fragment is connected to the adjacent immobilized primer and copied by the DNA polymerase. Sequencing proceeds through sequential incorporation of modified nucleotides (Figure 2.13b). Modified nucleotides contain terminating groups that prevent polymerase activity, while also incorporating fluorescent dyes. Each nucleotide is labeled with a different dye and can therefore be distinguished by the fluorescent signal. In each sequencing step, all modified nucleotides are added into the reaction mixture. In each step only one nucleotide, complementary to the next un-paired nucleotide on the DNA fragment, is added to each strand on the glass slide (Figure 2.13c). Synthesis can't proceed until the terminating group is removed, but before this, unused nucleotides are washed away and fluorescent signals measured. This ensures that each step corresponds to the extension of targeted DNA fragments by exactly one position, while all four nucleotides are examined simultaneously. Output of this method is a nucleotide sequence of each DNA fragment cluster (Figure 2.13d).

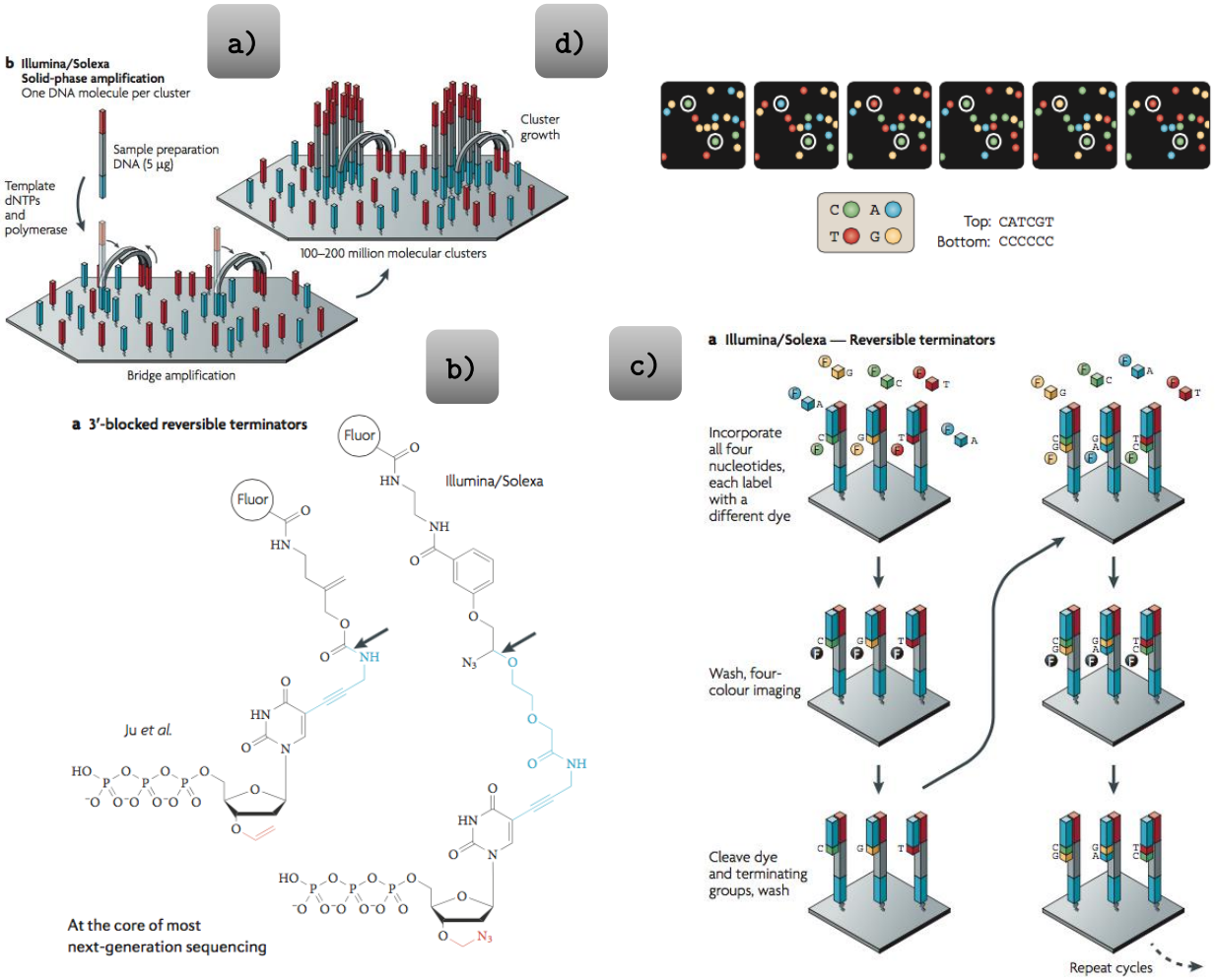


Figure 2.13: Illumina sequencing platform. **a)** Bridge amplification on a glass slide. **b)** Modified nucleotides. **c)** Sequencing procedure. **d)** Output [96].

Advantages of Illumina sequencing are massive parallelization, enabling 100+ million reads per sequencing run and high sequence accuracy compared to other NGS methods, with error rates of approximately 0.1% for early bases and slightly higher for later bases. It has one of the most competitive per basis cost currently on the market, which is especially relevant for large-scale projects, as well as an extensive bioinformatics infrastructure and standardized data analysis workflows developed by the community.

The main limitation is relatively short read length compared to other sequencing platforms, usually ranging between 50 and 300 nucleotides. This complicates genome assembly and makes it difficult to resolve repetitive sequences or complex structural variations. Of course, this also extends to transcript assembly and isoform quantification, particularly for complex genes with multiple overlapping transcripts. Additionally, certain sequence contexts exhibit systematically elevated error rates. Most notably, homopolymeric runs, stretches of identical nucleotides (for example, "AAAAAAA"). In these cases, the same color of light gets emitted in successive cycles, once for each nucleotide of the homopolymer, which degrades the capacity of optical detection.

2.5.2.3 Single Molecule Real Time Sequencing

Pacific Biosciences (PacBio) has developed single-molecule real-time (SMRT) sequencing that observes the incorporation of nucleotides into DNA chain in real-time at the single-molecule level (Figure 2.14) [98].

PacBio SMRT sequencing is done on a chip that contains many small wells (Figure 2.14a). At the bottom of each well, a single stranded DNA fragment is immobilized with an active DNA polymerase molecule synthesizing its complementary strand (Figure 2.14b). Nucleotides that are being added are modified with different fluorescent labels, which produce a specific light signal at the time of their incorporation. Nucleotide sequence is determined from the continuous measurement of fluorescence spectra. Unlike other methods, SMRT doesn't use any signal amplification procedure and consequently requires very sensitive measurement device. PacBio uses zero-mode waveguide detectors [99]. This detection is the main technological advancement that enabled the SMRT procedure.

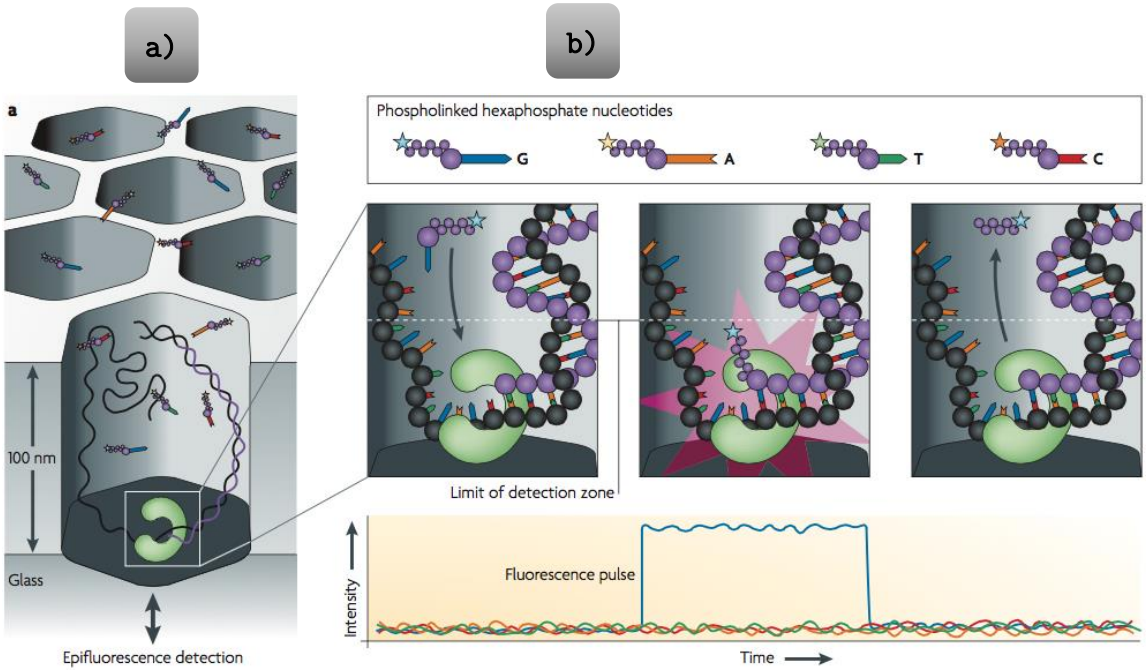


Figure 2.14: PacBio SMRT sequencing. **a)** Chip layout. **b)** Sequencing procedure [96].

An advantage of SMRT is that it's done in real time, which makes the whole procedure a lot faster. There is no need for any reagent exchange or termination stops that are part of other NGS methods. But the primary advantage of SMRT are the read lengths that this approach can generate. PacBio produces read lengths averaging 10,000 to 20,000 nucleotides, with even longer reads of over 100 kilobases achievable with specialized library preparation. This enables a much wider range of applications such as complete isoform characterization, genome assembly with superior coverage and detecting a broad range of structural variants, including those that can't be captured with shorter reads.

The main limitation of SMRT is the error rate. Single nucleotide detection causes base calling accuracy problems since there are no averaging effects that would enhance signal-to-noise ratio. Combined with increased polymerase error rate, due to fluorescent modifications on the nucleotides, this causes a significantly lower accuracy compared to other NGS methods [100]. The error rate of PacBio sequencing is up to 1%. Some protocols enable a ten-fold reduction to approximately 0.1% [101]. However, these are circular approaches where the same DNA molecule is sequenced multiple times, which increases the time and the cost of the procedure. Additionally, the per base cost of SMRT is still substantially higher than for other techniques with shorter reads, which restricts the application to specific problems where the advantages of long reads justify the increased cost.

2.5.2.4 Nanopore Sequencing

Oxford Nanopore Technologies has developed a nanopore-based long-read sequencing technique (Figure 2.15). Individual DNA or RNA strands pass through biological nanopore channels embedded in a lipid bilayer. As a nucleotide transits the nanopore, it obstructs ionic current flow across the pore, which produces a characteristic electrical signal dependent on the nucleotide identity. Sequential nucleotide passage through the nanopore generates a continuous ionic current trace. Machine learning models trained on these current signatures determine the nucleotide sequence [102].

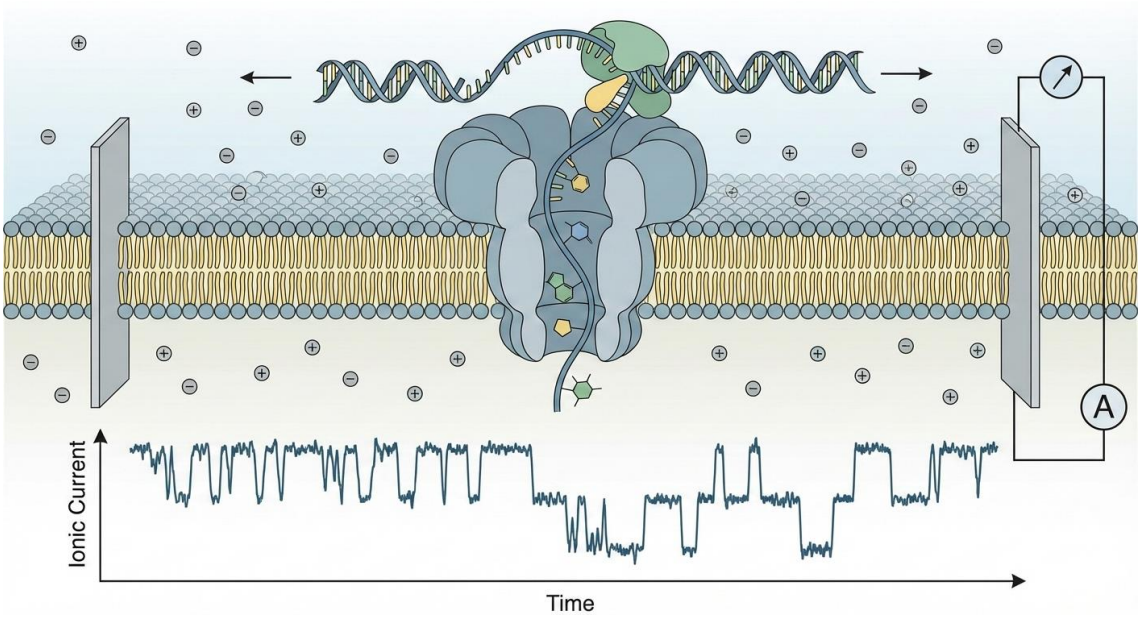


Figure 2.15: Nanopore sequencing visualization.

Main advantages of nanopore sequencing are:

- Ultra-long reads, exceeding 100 kilobases and even exceeding megabase lengths with optimized library preparation protocols.
- Direct RNA sequencing capacity, which eliminates the requirement for cDNA synthesis, the initial step in most other RNA sequencing approaches (discussed in the next Section).
- Direct detection of DNA modifications through altered ionic current signatures.
- Accessibility and portability of instruments. MinION device, frequently used in nanopore sequencing, can be easily transported and even field-deployed [103].

The main limitation is the substantially higher error rate, ranging from 5% to 15% with current machine learning models, although the base accuracy can be improved through an increased sequencing depth [102]. Additionally, nanopore has lower throughput per sequencing run compared to other NGS approaches.

2.5.3 RNA Sequencing Approaches

RNA sequencing (RNA-seq) applies the principles and the technologies of DNA sequencing to the transcriptome, enabling comprehensive measurement of gene expression. However, almost all sequencing methods take the DNA molecule as an input. Therefore, a key step in most RNA-seq workflows is the conversion of the transcriptome into a library of DNA fragments suitable for sequencing. The key challenge in this transformation is maintaining quantitative accuracy, meaning that the relative abundance of sequences in the final library should faithfully reflect the initial distribution of the corresponding transcripts in the sample.

Two distinct RNA-seq approaches are whole-transcriptome and targeted RNA sequencing, the main difference being, as the names suggest, the breadth of transcript coverage. Whole-transcriptome sequencing, also called bulk RNA-seq, aims to comprehensively capture and quantify all RNA molecules in a sample. Conversely, targeted RNA-seq focuses sequencing effort on a predetermined set of genes of interest, enabling either deeper coverage of those genes or equivalent coverage at reduced cost.

The most common approach, even for the bulk RNA-seq, is to primarily focus on mRNA, which is subsequently translated into proteins. In either case, the library preparation starts with extraction of the total RNA from cells or tissue. Then, mRNA may be enriched through poly(A) selection. Mature mRNA molecules possess a poly(A) tail, making them amenable to capture using oligo(dT) beads. This enriches for mRNA, which is magnetically isolated with the beads, while most rRNAs, long non-coding RNAs and circular RNAs get excluded. Alternatively, rRNAs can be selectively removed from total RNA through complementary oligonucleotide hybridization and enzymatic digestion. This approach preserves non-polyadenylated transcripts. Following enrichment or depletion, RNA is fragmented to appropriate sizes, depending on the sequencing method that will be used. This is achieved either through physical shearing or by enzymatic treatment. The RNA fragments are then converted to cDNA through reverse transcription. The standard protocol uses random hexamers or oligo(dT) primers for synthesis of the first strand. Then, the second cDNA strand is synthesized by DNA polymerase, generating double-stranded cDNA copies of the original transcriptome. The cDNA is subsequently subjected to standard DNA library preparation, which depends on the sequencing platform. Most often, cDNA is repaired to ensure blunt ends (where one strand isn't longer than the other) and then ligated to the sequencing adapters. Optionally, the library may be amplified through

PCR. Finally, prepared libraries are loaded onto the sequencer at appropriate concentrations and sequenced according to the instrument specifications [104].

Whole-transcriptome RNA-seq provides comprehensive coverage, but it requires relatively high sequencing depth for genes with low expression and is consequently expensive for large-scale projects. Conversely, the most common targeted RNA-seq approach employs probe-based capture methods. In this approach, oligonucleotide probes complementary to genes of interest are mixed with cDNA leading to hybridization with matching cDNA sequences. Probes contain biotinylation at their end, permitting subsequent capture on streptavidin-coated magnetic beads. Non-bound molecules are washed away, leaving an enriched population of targeted transcripts.

In either case, following the sequencing procedure, raw reads are aligned to a reference genome or transcriptome. The read counts assigned to each gene or transcript constitute the primary quantitative output. However, raw read counts from RNA-seq need to be corrected for some common artefacts. Library size needs to be normalized since different samples end up sequenced to varying depths and therefore have different total read counts. Normalization methods scale read counts by library size. Also, the transcript length differences need to be corrected since longer transcripts naturally accumulate more sequencing reads compared to the shorter transcripts of the same expression level. After these corrections, normalized RNA-seq data provide quantitative measures of expression, suitable for differential expression assessment and other downstream analyses.

2.5.3.1 TempO-Seq

The TempO-Seq (Templated Oligo assay with Sequencing readout) platform, developed by BioSpyder Technologies is an innovative approach to targeted RNA sequencing that differs from traditional cDNA-based methods. TempO-Seq is a ligation-based assay that directly targets RNA molecules, without requiring RNA extraction and reverse transcription into cDNA (Figure 2.16) [105].

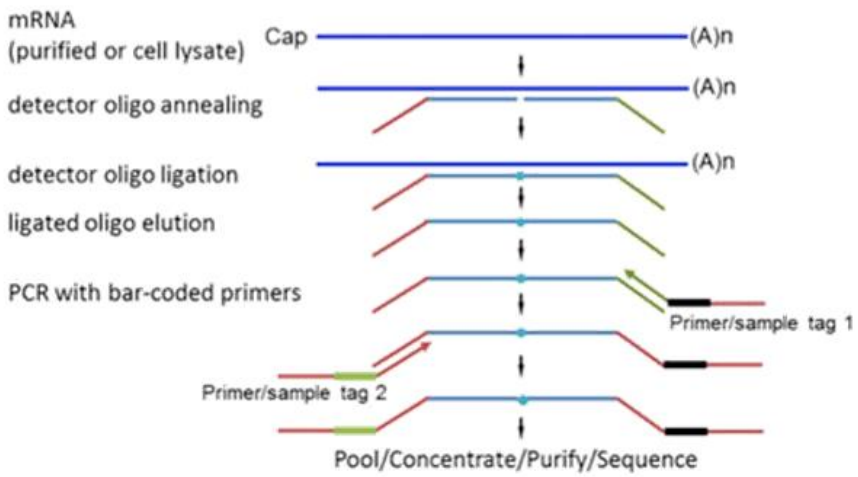


Figure 2.16: TempO-Seq protocol.

The core of the TempO-Seq design are synthetic oligonucleotide probes that have two functional components: The target-specific sequences hybridize to complementary RNA, while a universal primer site ("bar code") is shared across all probes in a single sample. This design enables simultaneous detection of multiple transcripts within a single reaction, while providing specificity through ligation-based hybridization to targeted RNA molecule in each sample. Following hybridization, excess unbound detector oligos are enzymatically removed. Adapters are added through subsequent enzymatic steps, converting the ligated products into templates suitable for PCR amplification. The amplification generates sufficient material for sequencing [106]. Final products are then pooled into a combined library and sequenced using the Illumina NGS platform. The detection of each sequenced fragment indicates a positive signal for the corresponding targeted gene. Finally, since the sequence of each ligated detector probe is known (from bar codes), the bioinformatic processing is used to associate each read count to its original sample [107].

The primary advantage of TempO-Seq is the ability to pool multiple samples into a single sequencing run, which is a significant resource optimization. Additionally, no RNA extraction and reverse transcription is required, which further simplifies the procedure, reducing both time and cost. While targeted, panels are designed to cover all relevant segments of human transcriptome. The most commonly used panel covers 3,257 genes selected for their relevance in toxicology, development and disease. Finally, TempO-Seq demonstrated results comparable to traditional RNA-seq methods [108].

The main limitation of TempO-Seq is naturally its targeted RNA-seq approach. This is further complicated by the design of the system, making it fully incompatible with the whole-transcriptome framework, since this would require specific oligo probe for each target in the transcriptome. Additionally, while TempO-Seq achieved high precision in comparison to traditional RNA-seq approaches, it still lagged in sensitivity, making it unsuitable for detection of very rare transcripts that require deeper sequencing. Finally, since this approach is relatively new, additional studies comparing it with conventional methods are necessary to confirm its performance.

Chapter 3

Results and Discussion

3.1 Local QSAR Modelling

The motivation for this research is to test, analyze and develop local QSAR models. The objective of localization is to improve statistical performance of QSAR models. The inspiration is a theoretical framework for the method that could pinpoint chemicals in the training set most similar to the specific target and use only these chemicals for machine-learning prediction of the endpoint values. This approach could have an advantage in QSAR modelling since mathematical procedure wouldn't be confounded by unrelated chemicals in the dataset.

3.1.1 Motivation

Local QSAR modelling represents a conceptually distinct paradigm from traditional global QSAR approaches. While conventional global models attempt to derive a single set of quantitative relationships applicable across an entire chemical space, local modelling generates compound-specific predictions by selecting and utilizing only the most structurally relevant subset of training data for each target chemical.

3.1.1.1 Overview of Local Modelling in QSAR

Local modelling approaches in QSAR represent an under-explored but conceptually compelling alternative to global model development. One of the earliest and most notable implementations of local modelling principles in computational toxicology is the Toxicity Estimation Software Tool (TEST) [109]. TEST implements a simple QSAR methodology where the toxicity of a target compound is estimated from a basic linear function of molecular descriptors. The linear approach within TEST calculates toxicity by assessing values from around thirty structurally most similar chemicals in the training set, thereby implementing a basic form of local modelling through chemical similarity-based selection. This approach has been widely employed for diverse toxicological endpoints, demonstrating the practical viability of local methods in regulatory contexts.

Beyond TEST, the local lazy regression (LLR) framework and related approaches have been explored in the QSAR literature [110]. These methods construct predictions for target molecules using their local chemical neighborhood rather than considering the entire training dataset at once. Notably, the lazar framework has implemented local modelling approaches for toxicity prediction [111]. The fundamental principle uniting these diverse local approaches is the recognition that chemical space is not uniformly populated and that structure-activity relationships may be locally specific to particular chemical subspaces. By restricting model development to compounds structurally similar to the target, these approaches attempt to identify and exploit local patterns that may be obscured or averaged away in global models.

Despite these examples, local modelling approaches have remained under-explored relative to global QSAR methods. Historical limitations in computational resources, combined with the practical challenges in implementing at-runtime predictions constrained

their adoption. However, with contemporary increases in processing power and the widespread availability of computing infrastructure, these barriers have been substantially removed, enabling reconsideration of local approaches with renewed methodological sophistication.

3.1.1.2 Theoretical Advantages

The fundamental advantage of local modelling lies in addressing inherent heterogeneity of chemical space. In comprehensive chemical databases, compounds span diverse structural classes with varying physicochemical properties and mechanistic pathways influencing any given endpoint. Many toxicological endpoints in particular exhibit multiple mechanistic pathways that depend on chemical structure. For example, acute oral toxicity can arise through distinct mechanisms such as central nervous system depression, respiratory irritation, hepatotoxicity, or nephrotoxicity [112]. Different chemical classes predominantly engage different mechanisms. A global model attempting to capture all these pathways simultaneously may sacrifice accuracy and interpretability in any single pathway and may not accurately reflect the actual structure-activity relationships relevant for specific chemical subgroups or mechanistic categories.

Localization directly addresses this mechanistic heterogeneity by allowing the model to adapt to the local chemical neighborhood of each target compound. By design, a local group of compounds structurally similar to a target is more likely to share similar mechanistic pathways, toxicokinetic profiles and metabolic processing than a random sample from the entire training set. Consequently, the mathematical relationships derived from such a local group more faithfully represent the relevant mechanistic framework operating in that chemical domain (Figure 3.1).

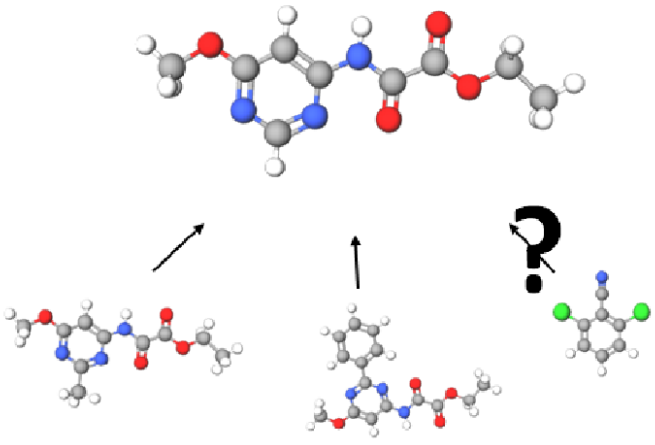


Figure 3.1: Motivation for QSAR localization: How much information related to the target chemical (top) can realistically be expected from a structurally very divergent chemical (bottom right)?

In practical terms, the localization strategy should theoretically yield several quantifiable benefits. First, the local group of similar compounds may follow more consistent and simpler mathematical relationships characterized by lower mechanistic noise, thereby reducing model complexity and improving interpretability. Second, by

excluding mechanistically dissimilar compounds that would otherwise introduce noise to the global model, localization should reduce prediction errors and increase robustness, particularly for chemicals operating through specialized mechanisms. Third, for chemical domains with limited training data or underrepresented structural classes, local models can focus computational resources on the relevant chemical space, potentially alleviating issues of class imbalance or sparse sampling. And fourth, the localization strategy permits transparent assessment of prediction reliability: a chemical whose prediction is based on structurally similar training compounds operating through common mechanisms can be considered more trustworthy than one based on mechanistically distant chemicals or unrelated structures.

3.1.1.3 Development within Computational Toxicology

The practical motivation for developing and refining local QSAR methods is particularly acute within computational toxicology and regulatory safety assessment. Regulatory toxicological databases, such as those compiled for acute oral toxicity, encompass tens of thousands of structurally diverse chemicals. As discussed above, mechanistic heterogeneity renders single global models suboptimal for broad applicability across such diverse chemical spaces.

Local QSAR modelling addresses this heterogeneity by allowing each target compound to be evaluated against an optimized set of mechanistically and structurally related training compounds. Beyond the statistical advantages described, this approach provides substantial benefits from a regulatory and practical perspective. Regulatory agencies and hazard assessors require not only point predictions, but also supporting evidence and justification for those predictions [27]. The local group of structurally and mechanistically similar compounds identified by the local modelling approach provides an interpretable reference set that, while not constituting formal read-across in the regulatory sense, nevertheless provides chemical and biological context grounding the prediction. Regulators can inspect the local group, evaluate the chemical similarity of identified analogues, assess the range of experimental values for those analogues and thereby develop defensible confidence in the derived prediction. This grounding in chemical analogues might help in bridging the gap between black-box machine learning predictions and interpretable, read-across approaches traditionally used in regulatory assessment, particularly under international frameworks such as REACH (see Section 2.1.4). Local QSAR models effectively merge the statistical rigor of machine learning with the chemical intelligibility of read-across, offering regulators transparent reasoning for predictions.

The regulatory and practical value of local modelling is particularly pronounced in hazard classification and risk assessment contexts where decision frameworks and acceptable levels of prediction uncertainty vary substantially across regulatory scenarios. Local models, by providing chemically meaningful applicability domain information alongside predictions, enable regulatory users to make context-appropriate decisions regarding model utilization. A prediction for a compound falling in a well-characterized chemical neighborhood may be accepted with greater confidence than a prediction for an isolated or poorly-characterized chemical structure, even if both emerge from the same mathematical model architecture.

3.1.2 Development of aiQSAR Methodology

Building upon the theoretical framework and historical context outlined above, a comprehensive local QSAR methodology was developed to advance the local modelling approach with modern machine learning and computational infrastructure. The aiQSAR

methodology represents an evolution that has undergone refinements since its initial development. The initial methodology, published in 2019 [113] and employed for the acute toxicity modelling initiative (see Section 3.1.3), established the core framework and demonstrated its utility. Subsequently, based on empirical results and practical considerations for software implementation, several parameters and design choices were refined for the final software version presented here.

3.1.2.1 aiQSAR Workflow

The aiQSAR methodology is structured as an eight-step workflow that operates with an at-runtime execution: for each target chemical in the dataset, the complete model building procedure is executed *ab initio,* specifically for that compound (Figure 3.2). This at-runtime approach, while computationally intensive, provides fundamental advantages by ensuring that model construction is optimized for the local chemical space of each target, eliminating potential issues that arise from static descriptor selection and enabling the software to adapt dynamically to user-specified parameters. The complete workflow is presented here.

Step 1: Calculating Descriptors

Molecular descriptors are calculated for all compounds in the training set and the evaluation set. In the initial methodology these were 3,839 descriptors calculated using the Dragon 7 software, encompassing constitutional indices, topological descriptors and molecular properties. In the final version of the aiQSAR software, descriptor calculation has been refined through implementation of a specialized R library that selects a curated subset of 1938 descriptors from the complete Dragon descriptor space [114].

Step 2: Calculating Fingerprints

Two distinct types of molecular fingerprints are calculated for all compounds using the R package “rcdk” [115]: PubChem fingerprints (structural keys defined by the PubChem database) and Extended fingerprints (calculated using extended sequences considering chemical bonds up to six atoms of depth, folded to 1,024 bits). The dual fingerprint approach provides complementary structural information. PubChem fingerprints encode specific pharmacophoric features and structural patterns, while Extended fingerprints capture broader topological connectivity relationships. This complementary information strategy enhances the robustness of chemical similarity assessment.

Step 3: Calculating Fingerprint Similarities

For each target compound, Tanimoto distances are calculated between the target and all training set compounds using both fingerprint types. The Tanimoto similarity coefficient ranges from 0 (complete dissimilarity) to 1 (identity). It is well-established for fingerprint-based similarity assessment and commonly used in cheminformatics applications (see Section 2.3.4).

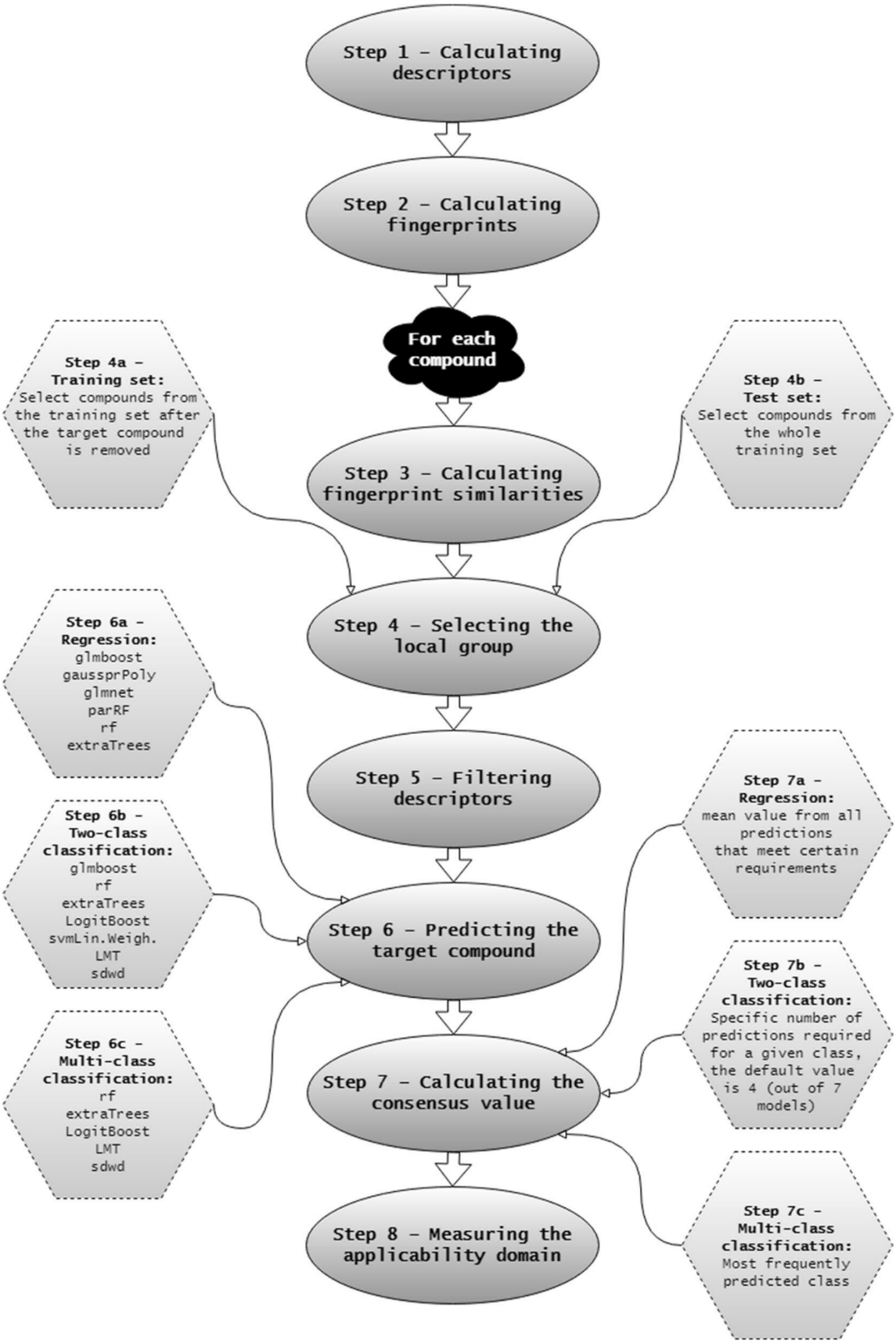


Figure 3.2: aiQSAR workflow.

Step 4: Selecting the Local Group

An algorithm was developed to select an optimal number of training set compounds most similar to the target compound based on the fingerprint distances calculated in Step 3 (Figure 3.3). The initial methodology selected between 20 and 50 compounds while the final software version implements a more flexible approach with a default local group size of 30 compounds. However, the user is able to select any range between 10 and 100 compounds as appropriate for their specific application context. This change addresses practical considerations: smaller local groups may be preferred when working with highly specialized chemical domains where only a few true structural analogues exist, while larger groups may be desirable when modelling broader chemical space.

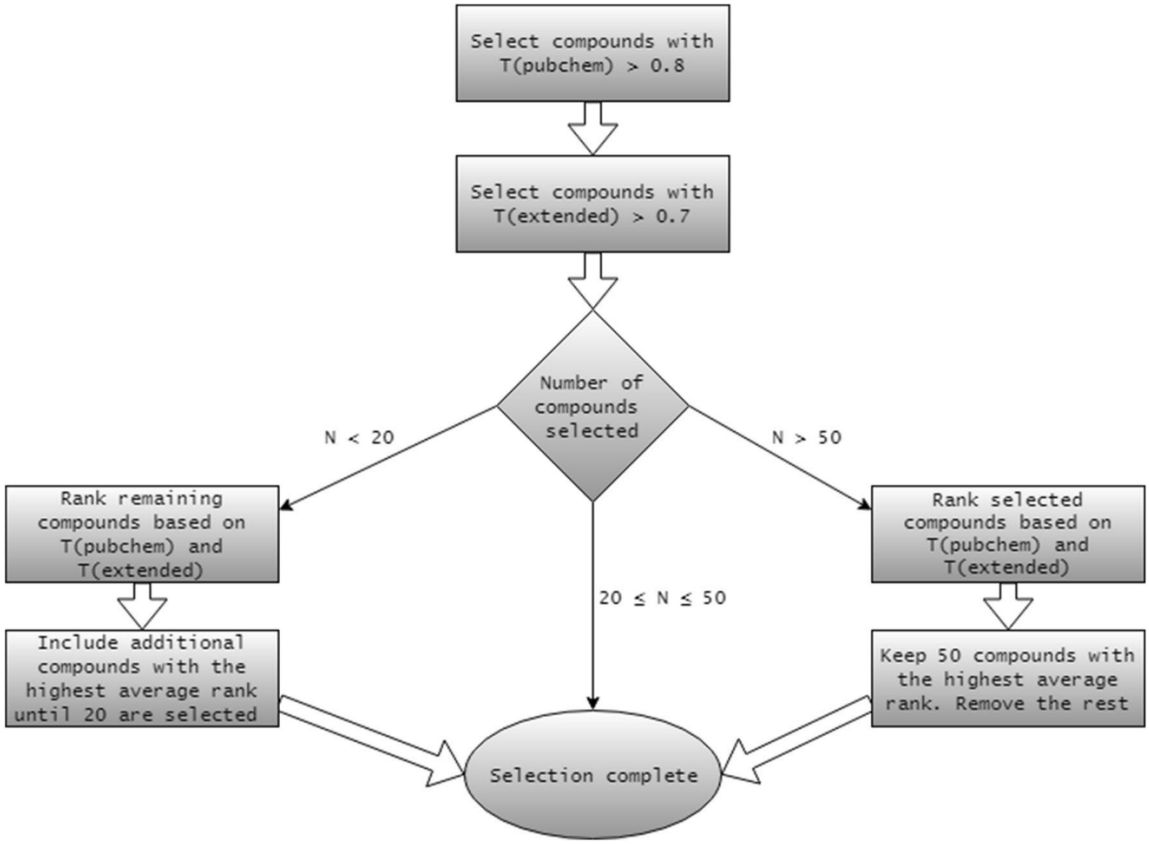


Figure 3.3: Local group selection. T(pubchem) and T(extended) are Tanimoto distances based on PubChem and Extended fingerprints respectively.

The selection algorithm operates through a two-phase process designed to ensure inclusion of all compounds that have the potential of positively contributing towards the calculation. In the first phase, all compounds meeting primary threshold criteria are identified: PubChem similarity ≥ 0.80 and Extended similarity ≥ 0.70. If this initial filtering yields fewer than the lower target for the number of compounds (*e.g.*, fewer than 30), the algorithm proceeds to second phase. If it yields more than the upper target, compounds are ranked by average rank scores from both fingerprint types and the top compounds are retained to reach exactly the upper target number. In the second phase, if necessary, similarity threshold is removed to permit additional candidates based on the

average rank scores from both fingerprints until the lower target for the number of compounds is reached.

Step 5: Filtering Descriptors

Descriptors with missing values for any compound in the local group or the target compound are discarded, ensuring data completeness for all modelling algorithms. Remaining descriptors are filtered to remove zero variance (descriptors with identical values across all compounds in the local group) and near-zero variance descriptors using the "nearZeroVar" function from the R package "caret" [116]. The near-zero variance filter identifies descriptors with extremely low variability that provide minimal discriminatory power. A descriptor that exhibits variance and predictive power in one chemical subspace (for example, number of oxygen atoms in an oxygen-rich domain) may be uninformative or highly correlated with other descriptors in another subspace.

Step 6: Predicting the Target Compound

Multiple machine learning algorithms are applied at-runtime to the local group of compounds to generate independent predictions for the target compound. The specific algorithms employed depend on the endpoint type and are listed below.

- **Regression endpoints**: glmboost (Boosted Generalized Linear Model) [117], gaussprPoly (Gaussian Process with Polynomial Kernel) [118], glmnet (Lasso and Elastic-Net Regularized Generalized Linear Models) [38], rf (Random Forest) [119] and extraTrees (Randomized Forest) [120].
- **Binary classification endpoints**: glmboost, rf, extraTrees, LogitBoost (Boosted Logistic Regression) [121], svmLinearWeights (Linear Support Vector Machines with Class Weights) [122], LMT (Logistic Model Trees) [123] and sdwd (Sparse Distance Weighted Discrimination) [124].
- **Multi-class classification endpoints**: rf, extraTrees, LogitBoost, LMT and sdwd.

The selection of these specific algorithms was informed by prior comprehensive evaluation across diverse datasets and represents a deliberate balance between computational efficiency and predictive performance. Initially, all models available within the "caret" package were tested with default parameters on a small number of datasets. Well performing models were selected for subsequent rounds of testing on a larger number of datasets. Finally, all models that achieved specific performance threshold were extensively evaluated in different combinations until desired balance between computational complexity and predictive performance was achieved.

Step 7: Calculating the Consensus Value

Predictions from all individual algorithms are combined into a single consensus output value using endpoint-specific aggregation procedures described here.

- **Regression consensus**: After all individual algorithm predictions are generated, extreme outlier predictions are identified and excluded from consensus calculation. Outliers are conservatively defined as predictions >10% higher than the maximal training set value or >10% lower than the minimal training set value. This

conservative outlier exclusion procedure prevents individual algorithms that may produce aberrant predictions from unduly influencing the final estimate. The final prediction is calculated as the mean of the remaining predictions, with equal weighting given to all included models.

- **Binary classification consensus**: The final prediction is determined by majority vote across the seven individual model predictions. Each algorithm casts a "vote" for either class. The default cutoff value is 4 out of 7 predictions. This cutoff can be adjusted in the software's advanced mode to suit different decision contexts. For example, in drug discovery screening where large numbers of compounds are being evaluated and researchers can afford false positives, a lower cutoff (*e.g.*, 2 out of 7) may be preferred in order to maximize detection of active compounds. Conversely, in toxicity assessment, where false positives incur significant costs, a more conservative cutoff (*e.g.*, 6 out of 7) may be required for the safety prediction. This flexibility is valuable in regulatory applications, as it permits the same model to serve diverse stakeholder requirements.
- **Multi-class classification consensus:** The final prediction is the most frequently predicted class across the five individual models. In the initial methodology, in case of a tied vote (*e.g.*, two models predict class A and two predict class B), the class that is least represented in the training set was selected, following the principle that underrepresented classes require stronger evidence to conclude membership. In the final version of the aiQSAR software, the outcome with a tied vote is referenced as such, with all tied classes reported to the user. This allows for greater flexibility and gives the user the final say on the prediction outcome.

Step 8: Applicability Domain Assessment

An Applicability Domain Measure (ADM) is calculated for each target compound based on average Tanimoto similarities between the target compound and all compounds in its local group. The calculation procedure is hierarchical and reflects the principle that consistent high similarity across multiple fingerprints indicates higher confidence in the local model's applicability to the target (Figure 3.4). This measure enables transparent assessment of prediction reliability: compounds with higher ADM scores, indicating greater average structural similarity to their local training set neighbors, are expected to have more reliable predictions, while compounds assigned low ADM scores fall into poorly-characterized regions of chemical space and warrant cautious interpretation.

3.1.2.2 Consensus Thresholds for Binary Classification

As mentioned in the previous section, an important feature of aiQSAR's binary classification implementation is flexibility in adjusting consensus prediction thresholds based on decision context. While the default consensus threshold is 4 out of 7 model agreements, users can specify alternative cutoffs. This flexibility permits the calibration of sensitivity-specificity balance to match the relevant regulatory or research requirements. Analysis of consensus thresholds using the Ames mutagenicity dataset, a binary classification endpoint assessing the mutagenic potential of chemical compounds, demonstrates that varying the number of required positive predictions systematically changes performance metrics (Figure 3.5).

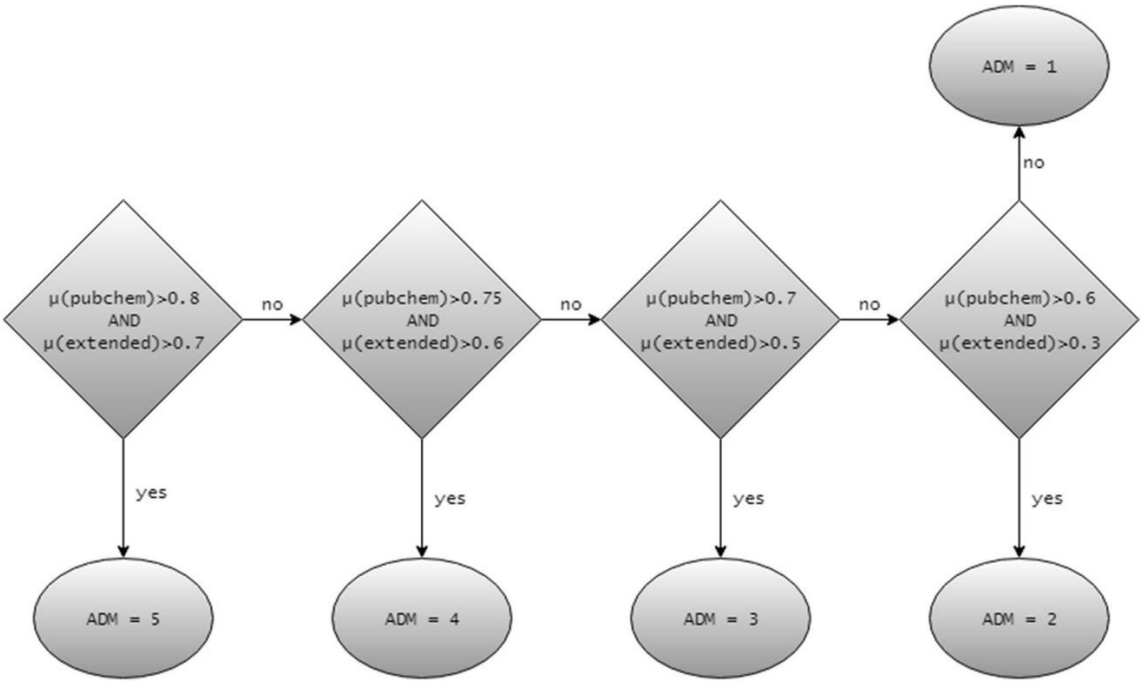


Figure 3.4: Calculating the ADM. μ(pubchem) and μ(extended) are average values of Tanimoto distances between the target compound and compounds in its local group.

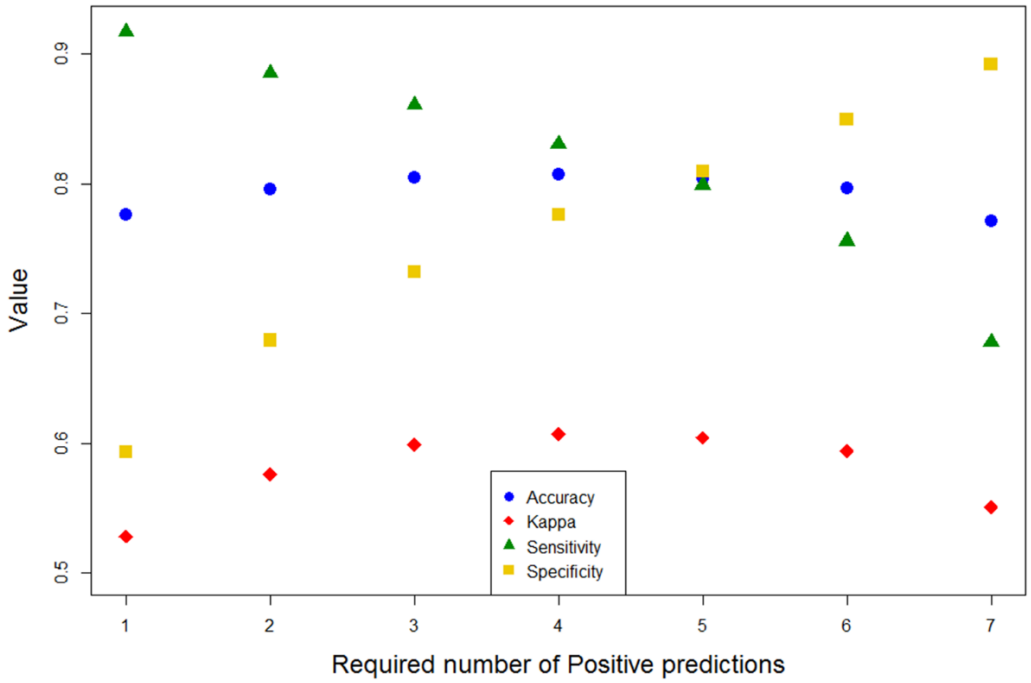


Figure 3.5: Various performance measures plotted against the required number of positive predictions for the Ames mutagenicity test dataset.

The Ames test is implemented as one of aiQSAR's 38 pre-built endpoints in the final software, enabling users to evaluate mutagenic potential alongside toxicity assessments. This adaptive threshold capability represents a substantive advance over fixed-threshold approaches, enabling the same computational model to serve diverse applications through a parameter adjustment.

3.1.3 Testing Framework: Rat Acute Oral Toxicity Initiative

To validate the local QSAR approach and establish its practical utility for regulatory toxicological prediction, the aiQSAR methodology was comprehensively tested during a large-scale collaborative modelling initiative focused on rat acute oral toxicity. This section describes the initiative structure, datasets, endpoints and results [125].

3.1.3.1 Overview

Acute oral toxicity, measured as the median lethal dose (LD_{50}) in rodents is a fundamental toxicological parameter required for regulatory hazard classification and risk management of chemicals. The formal definition of LD_{50} is the dose of a substance needed to kill 50% of test animals within 14 days of oral exposure. Regulatory agencies globally mandate this acute toxicity data for chemical classification and hazard assessment [126].

However, the exclusive reliance on *in vivo* acute toxicity testing is severely constrained by practical, economic and ethical considerations. Individual LD_{50} studies require between 20 and 60 test animals per chemical, incurring significant costs and requiring several weeks for experimental completion. With thousands of new chemicals that are introduced annually, comprehensive *in vivo* testing represents an untenable burden. Moreover, the emerging regulatory frameworks and directives mandate the reduction of animal testing through the adoption of validated alternative methods, which includes computational approaches. The Three Rs principle (Replacement, Reduction and Refinement) guides regulatory policy in multiple jurisdictions, with QSAR and other computational methods emphasized as key tools in this effort [24].

Within this regulatory context, the U.S. Interagency Coordinating Committee on the Validation of Alternative Methods (ICCVAM), with support from the National Toxicology Program (NTP) and Interagency Center for the Evaluation of Alternative Toxicological Methods (NICEATM), organized a large-scale international collaborative initiative to develop and validate predictive *in silico* models for acute oral toxicity. This collaborative project represented a systematic application of the OECD principles for QSAR model validation and was structured to rigorously evaluate submitted models against high-quality experimental data. The initiative assembled 35 international research groups contributing 139 distinct computational models, representing diverse modelling approaches including traditional QSAR, machine learning ensemble methods, deep learning approaches and the local group-based aiQSAR methodology [127].

3.1.3.2 Dataset Compilation, Curation and Composition

The NICEATM collaboration compiled and curated a comprehensive database of rat acute oral LD_{50} data encompassing 11,992 chemicals with QSAR-ready structures. The data collection process involved systematic retrieval from several authoritative sources, as well as experimental testing for a small number of new chemicals.

Rigorous data curation procedures were implemented to ensure dataset integrity and suitability for QSAR model development. Chemical structures were standardized using a comprehensive workflow that identified and systematically removed inorganic compounds, organometallic compounds, mixtures and improperly characterized structures.

Standardization procedures addressed varied representation of chemical salts and counterions, ensured consistent treatment of tautomeric forms and normalized isomeric representations.

Data aggregation addressed the common occurrence of multiple LD_{50} measurements per chemical substance. Individual chemicals in regulatory databases frequently have multiple independent LD_{50} determinations from different testing laboratories or time periods, reflecting the inherent reproducibility limitations of the *in vivo* tests, as well as natural biological variability. To obtain a single representative value per unique chemical structure, a conservative aggregation strategy was implemented: the median of the lowest quartile was computed from discrete LD_{50} values, with limit test data and range interval measurements explicitly excluded.

The resulting curated dataset of 11,992 chemicals was partitioned into training and evaluation sets, with approximately 75%-25% split and using semi-random selection that preserved the distribution of LD_{50} values and toxicity categories across both partitions. Not all chemicals possessed sufficiently precise data for every endpoint, with the maximum dataset size (11,992) representing the complete curated dataset and specific endpoints utilizing subsets sized accordingly.

Furthermore, an additional set of approximately 45,000 "decoy" chemicals was introduced to expand the set of chemicals for which the prediction is requested. This additional set, together with the evaluation set, constituted the test set of overall 48,138 compounds. Within this test set, there were 2,998 compounds hidden from participants with known toxicity values across all endpoints. This structure ensured a rigorous assessment of the initiative results, given that participants didn't know which chemicals ended up being evaluated. Additionally, it provided the community with the most up-to-date predictions of toxicity values for a large number of potentially interesting compounds that didn't have the experimental data available [128].

3.1.3.3 Five Endpoints

Five distinct toxicological endpoints of regulatory significance were modeled within the initiative, spanning both regression, binary classification and multi-class classification endpoints.

1. LD_{50} point estimate (regression): Prediction of discrete median lethal dose values expressed as mg/kg body weight, with values transformed to logarithm of mmol/kg for statistical modelling purposes. This continuous endpoint directly quantifies acute toxicity severity and serves as the basis for regulatory classification and determination of protective measures.

2. "Very toxic" binary classification: Binary identification of "very toxic" chemicals ($LD_{50} \leq 50$mg/kg) versus all others. This threshold corresponds to the highest hazard category under both EPA and globally harmonized systems and represents chemicals of greatest concern for acute toxicity hazards requiring the most stringent protective measures and labeling.

3. "Non-toxic" binary classification: Binary identification of "non-toxic" chemicals ($LD_{50} > 2{,}000$mg/kg) versus all others. This threshold distinguishes chemicals of minimal acute toxicity concern from those requiring active hazard management and protective measures.

4. EPA hazard categories (multi-class classification): Four-category hazard ranking consistent with U.S. EPA regulatory requirements [129] (Table 3.1).

Table 3.1: EPA Toxicity Categories.

Category	LD_{50} Range (mg/kg)	Hazard Level
Category I	≤ 50	Highest toxicity
Category II	$50 < LD_{50} \leq 500$	Moderate toxicity
Category III	$500 < LD_{50} \leq 5{,}000$	Lower toxicity
Category IV	$> 5{,}000$	Minimal concern

5. Globally Harmonized System (GHS) hazard categories (multi-class classification): Five-category hazard ranking (Table 3.2). The GHS classification is mandated internationally and represents the global standard for chemical hazard communication.

Table 3.2: GHS Toxicity Categories.

Category	LD_{50} Range (mg/kg)	GHS Classification
Category 1	≤ 5	Highest acute toxicity
Category 2	$5 < LD_{50} \leq 50$	Very high acute toxicity
Category 3	$50 < LD_{50} \leq 300$	High acute toxicity
Category 4	$300 < LD_{50} \leq 2{,}000$	Moderate acute toxicity
Category 5	$2{,}000 < LD_{50} \leq 5{,}000$	Low acute toxicity concern

These five endpoints span the complete spectrum of practical regulatory requirements and have enabled rigorous evaluation of the aiQSAR methodology across diverse prediction tasks and difficulty levels.

3.1.3.4 Statistical Analysis

Model performance was evaluated using endpoint-appropriate statistical metrics applied to predictions falling within each model's declared applicability domain. For binary and multi-class classification endpoints, balanced accuracy, sensitivity and specificity were calculated from confusion matrices. Balanced accuracy was prioritized for classification endpoints to account for potential class imbalance in the training and evaluation sets. For continuous LD_{50} predictions, RMSE and R^2 were employed as standard regression performance metrics [127]. In addition to performances scored and ranked by NICEATM panel, some more advanced performance metrics were calculated both for the full training and evaluation sets, as well as at different applicability domain levels in order to demonstrate the hypothesized positive correlation between structural similarity towards the local training set and prediction reliability.

3.1.3.5 aiQSAR Performance

Results of the comprehensive evaluation demonstrate that the aiQSAR methodology achieves competitive performance across all five acute oral toxicity endpoints (Table 3.3).

Table 3.3: aiQSAR performance on five acute toxicity endpoints from the NICEATM collaborative modelling initiative.

Endpoint	Train Accuracy	Train Kappa	Evaluation Acc	Evaluation Kappa	NICEATM score rank
Very Tox	0.94	0.58	0.93	0.58	10th out of 30
Non Tox	0.78	0.55	0.78	0.55	9th out of 31
EPA class	0.63	0.42	0.64	0.43	5th out of 23
GHS class	0.63	0.44	0.63	0.45	1st out of 19
Endpoint	**Train R2**	**Train RMSE**	**Evaluation R2**	**Evaluation RMSE**	**NICEATM score rank**
log (LD_{50}, mmol/kg)	0.56	0.60	0.61	0.56	3rd out of 23

For binary classifications, evaluation set accuracies of 0.93 (very-toxic classification) and 0.78 (non-toxic classification) reflected strong discrimination of extreme toxicity categories critical for regulatory hazard assessment. Notably, aiQSAR ranked among top submissions: 1st out of 19 for GHS multi-class classification, 5th out of 23 for EPA multi-class classification, 3rd out of 23 for continuous LD_{50} prediction and within top 10 out of 30-31 for binary classifications. This competitive ranking is particularly noteworthy given that aiQSAR employed no proprietary tools, specialized domain-specific descriptors, or advanced optimizations beyond the core methodology, operating instead with open-source software and a unified automated workflow applicable across diverse chemical domains and endpoints. For LD_{50} point estimate regression, evaluation set R^2 of 0.61 with RMSE of 0.56 log units demonstrated robust transferability from training to evaluation sets.

Additionally, it should be noted that some evaluation set chemicals had acute oral toxicity values available through public online sources not officially part of the NICEATM challenge dataset. Some modelling teams identified and utilized these supplementary data, potentially improving their reported performance metrics. The aiQSAR ranks presented here are based exclusively on performance evaluated by the NICEATM panel and without reliance to any external data sources.

3.1.3.6 Applicability Domain Performance Relationship

A critical validation of the local aiQSAR approach was obtained from systematic analysis of performance metrics plotted against the ADM rank. For all five endpoints, a consistent positive correlation between ADM rank and prediction metrics was observed across both the training and the evaluation sets: compounds with higher ADM scores, indicating greater average structural similarity to their local group, consistently demonstrated superior performance compared to compounds with lower ADM scores (Figure 3.6).

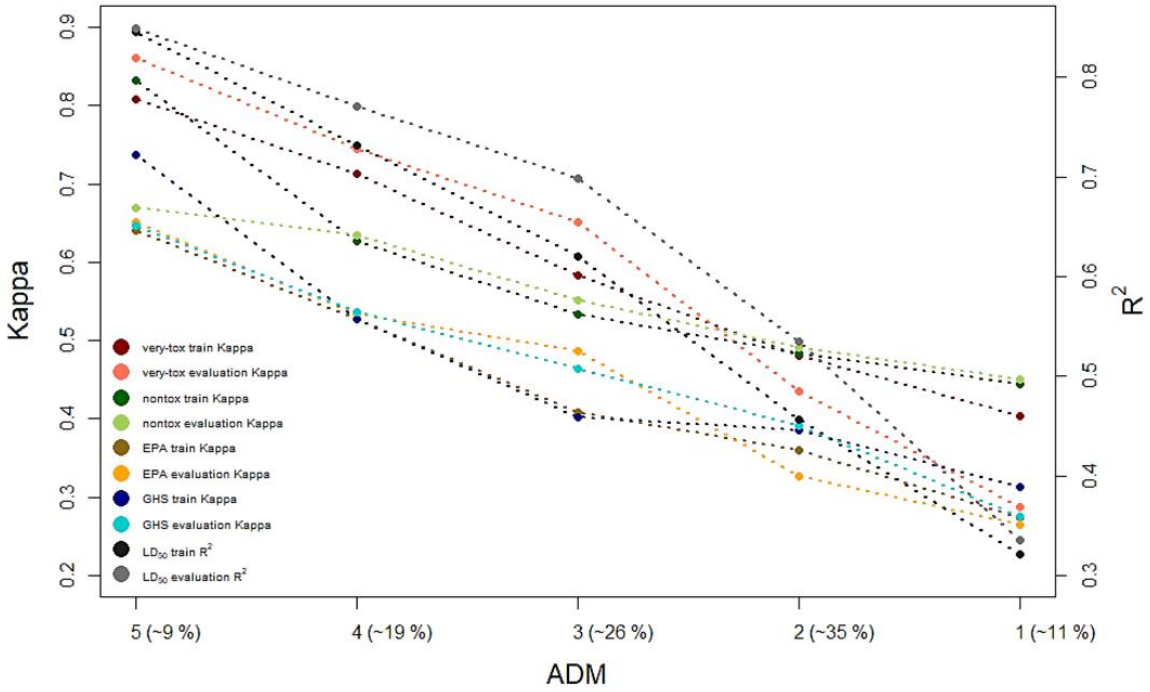


Figure 3.6: aiQSAR methodology performance on acute toxicity endpoints with respect to the ADM score. Left axis plots kappa as a measure for classification performance. Right axis plots R^2 as a measure for regression performance. Numbers in parentheses give percentages of compounds in each ADM group. These numbers are approximate since the final number of compounds differed slightly between the five endpoints.

This strong performance correlation validates the fundamental hypothesis underlying aiQSAR methodology: Chemical prediction reliability is determined by structural analogy to the local group. For example, in EPA category classification, evaluation set kappa improved systematically from approximately 0.3 at ADM=1 to 0.65 at ADM=5. This relationship enables transparent communication of prediction confidence by assessing whether the predicted chemical falls within a well characterized region of chemical space (high ADM) or a poorly defined one (low ADM). The decision-making regulatory context can be calibrated accordingly. This ADM gradient, providing continuous information about relative prediction reliability, is more informative than a simple binary accept/reject classification that was required for the NICEATM initiative.

3.2 aiQSAR Software

The aiQSAR software is a fully integrated computational platform designed to implement the local QSAR modelling paradigm at scale. This section provides a comprehensive description of the software. The development, architecture and operational features of aiQSAR are presented, along with detailed description of the 38 integrated endpoints and their underlying datasets.

3.2.1 Implementation

The aiQSAR software has been designed as a user-accessible computational platform that bridges the gap between complex machine learning methodology and practical usability. The objective was delivering sophisticated local modelling tool accessible to users without extensive programming expertise, while simultaneously providing advanced parameterization options for more technical users engaged in method development and model refinement. The software is implemented using the R programming language and the Shiny web application framework, enabling platform independence whilst ensuring compatibility with the underlying statistical and machine learning libraries that constitute the computational backbone.

The software interface employs a hierarchical approach to user interaction, providing a streamlined Basic mode for standard predictive applications and a comprehensive Advanced mode for custom model development and method refinement (Figure 3.7). This dual-interface framework aims to achieve the software's dual objectives of accessibility and flexibility. The Basic mode presents users with an intuitive graphical interface requiring only minimal input specification: selection of an endpoint, selection of input datasets and selection of desired output types. No configuration of algorithmic parameters is required in Basic mode, making it suitable for regular users without deep technical knowledge of the underlying methods. Conversely, the Advanced mode provides access to the complete suite of methodological parameters governing local model construction, similarity thresholds, local group sizing and consensus strategies. This layered design ensures that power-users and model developers can examine, verify and optimize each aspect of the computational procedure while maintaining quick and easy access for simpler applications.

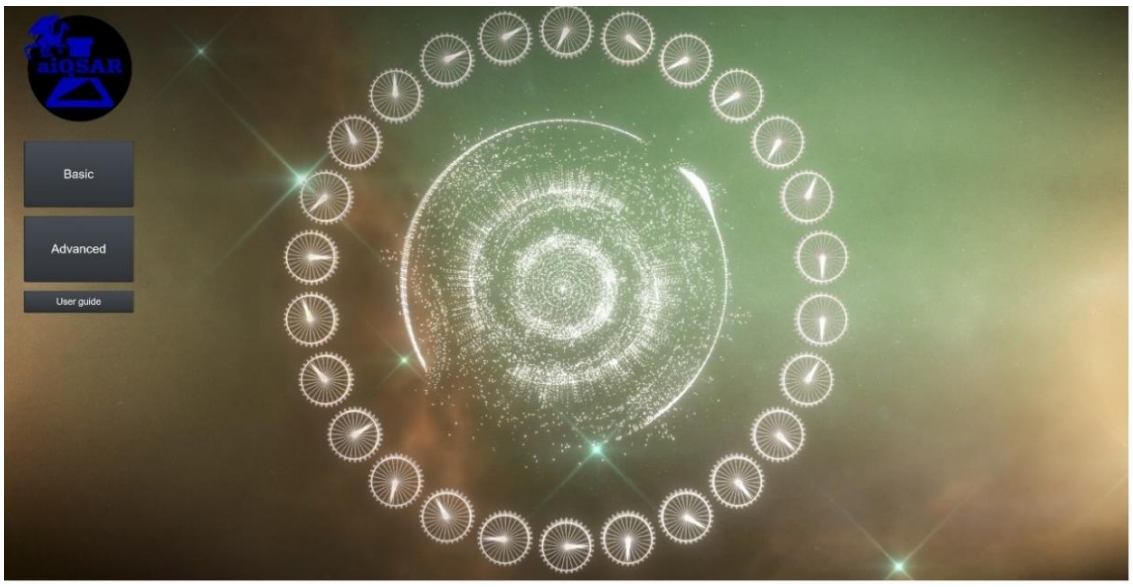


Figure 3.7: The initial aiQSAR application screen, as encountered upon software launch, displaying the aiQSAR interface with simple navigation menu and links to Basic and Advanced modes.

The software operates exclusively on the Windows operating system and requires no formal installation procedure. Users simply download and extract the compressed archive and execute the application directly [130]. This accessible deployment approach was a

deliberate design choice to minimize barriers to software adoption among researchers lacking extensive computational expertise and infrastructure.

3.2.2 UI Structure

As discussed, the aiQSAR software user interface was designed to accommodate diverse users with varying sophistication levels and application requirements. The following section details the structure, functionality and operational characteristics of both Basic and Advanced modes.

3.2.2.1 Basic Mode

The Basic mode interface provides an interaction pathway optimized for simplicity. Within this mode, users must specify four categories of information: the target endpoint for prediction, the datasets (training and/or test set), the specific tasks to be performed and the desired output formats (Figure 3.8).

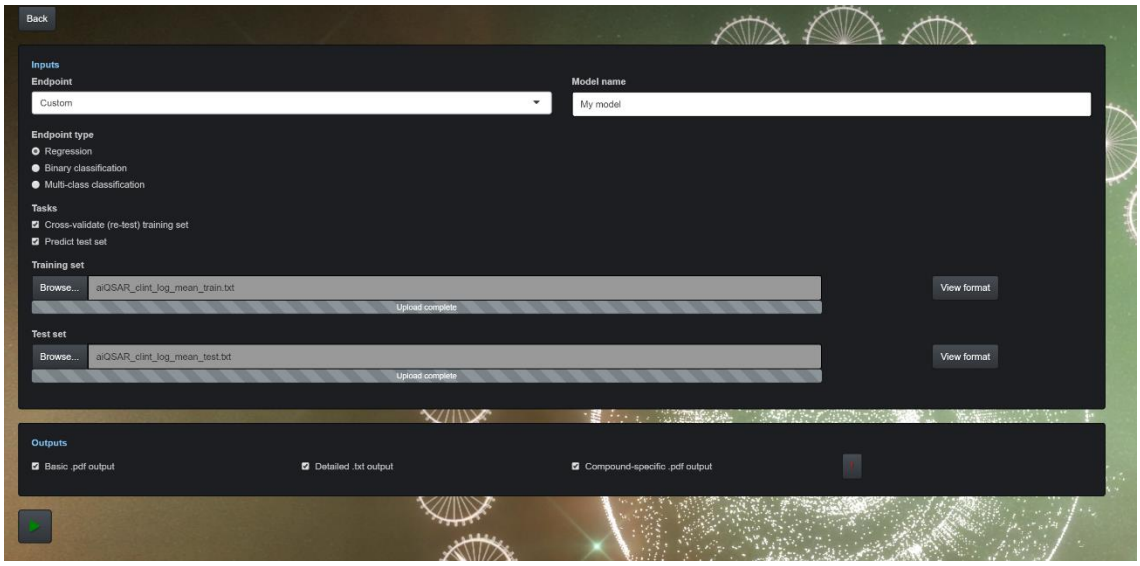


Figure 3.8: Basic mode interface for standard test set prediction, showing input panels for endpoint selection, training and test set file specifications, task selection and output format options.

Endpoint selection: Users may choose from 38 pre-built endpoints encompassing acute toxicity, Tox21 assays, VEGA-derived models, PubChem assays and additional specialized endpoints developed at Mario Negri Institute (see Section 3.2.3). Alternatively, users may specify a custom endpoint, requiring provision of a new training dataset and explicit declaration of the endpoint type (regression, binary classification, or multi-class classification with up to 8 distinct classes).

Dataset specification: For training sets, each row should consist of a chemical compound defined by SMILES and experimental endpoint value, separated by one or more blank spaces. For test sets, each row should specify a single SMILES. Automated validation provides explicit error message if structural requirements are not satisfied.

Task selection: Task specification determines whether the software performs prediction on the test set and/or cross-validation (re-testing) of the training set. It is most

often the case that predicting the test set is the only task, especially for the pre-built models. It would make sense for the user to request cross-validation of the training set as well for new, custom dataset where user doesn't intend to create and export a new model (which is possible in the Advanced mode, see Section 3.2.4.5).

Output format selection: Three output options are provided: basic PDF output, detailed text output and compound-specific PDF output (see Section 3.2.4).

Another available option is the prediction of a single chemical for any of the pre-built or user-created models. In this case, no file inputs are necessary. Instead, user should simply provide the SMILES for the chemical of interest and run a shortened version of the computational procedure (Figure 3.9). In this case the only generated output is the compound-specific PDF output.

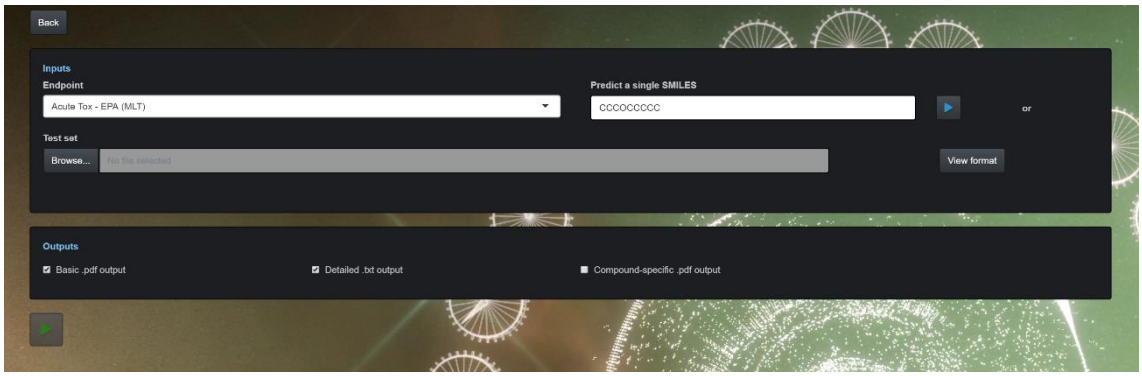


Figure 3.9: Basic mode interface for single chemical prediction, with SMILES input field and run button, eliminating the need for test set files when predicting individual compounds.

3.2.2.2 Advanced Mode

The Advanced mode interface expands Basic mode functionality with parametric controls governing the underlying aiQSAR algorithm, enabling experienced users to fine-tune model performance for specific use cases. The Advanced mode settings panel provides granular control over the core parameters (Figure 3.10).

SMILES canonicalization: A binary toggle controls whether all submitted SMILES strings are canonicalized according to CDK standards prior to processing.

Local model size parameters: The interface provides slider controls governing minimum and maximum numbers of compounds selected for local model construction. The user can select any range between 10 and 100 compounds.

Fingerprint similarity thresholds: Separate sliders control cutoff values for PubChem and Extended fingerprint similarities employed in the initial screening phase of local group selection. Default values are established at 0.80 for PubChem similarity and 0.70 for Extended similarity. These thresholds determine the stringency of initial candidate compound selection prior to secondary refinement procedures (see Section 3.1.2.1.4). Reducing the cutoff value for a given fingerprint type will increase the weight of that fingerprint in local group selection step by "locking-in" more compounds sorted in accordance to that fingerprint type in the first step of the selection procedure.

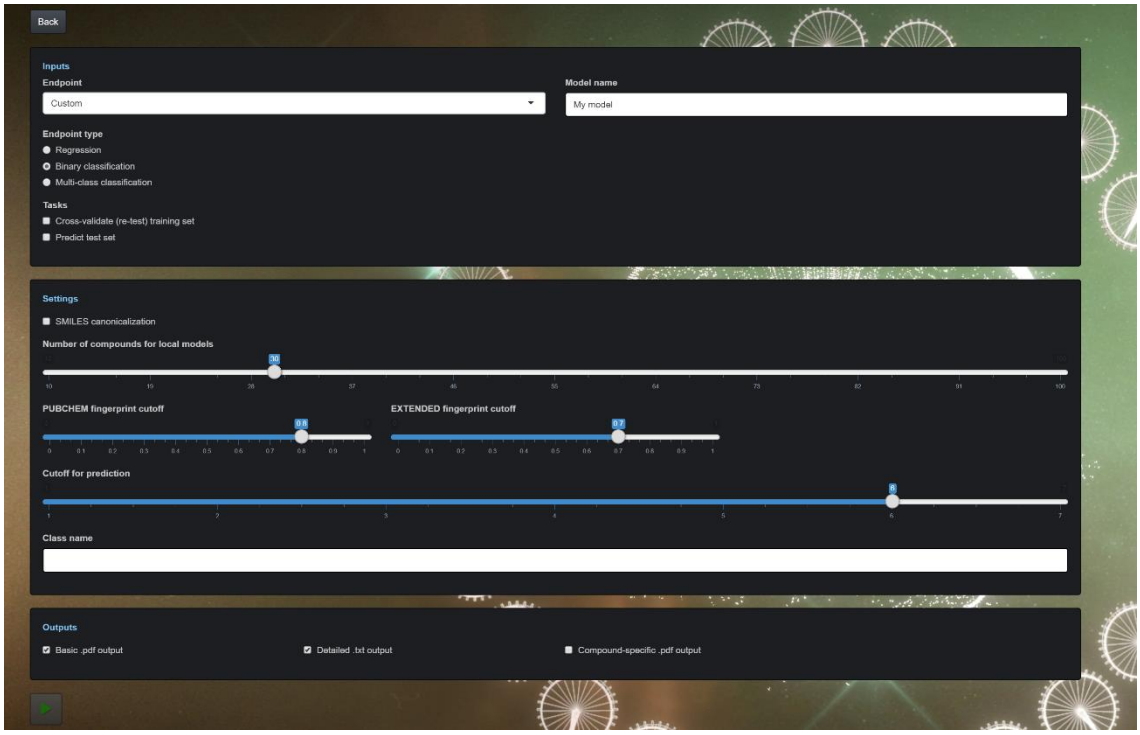


Figure 3.10: Advanced mode settings interface, revealing the full range of configurable parameters including the input formatting options, fingerprint cutoffs, local group selection criteria and consensus building options.

Consensus method specification (regression endpoints only): Enables selection between mean (default) and median aggregation methods for combining predictions from individual regression models.

Prediction cutoff specification (binary classification only): An advanced setting enables modification of the default requirement for consensus classification predictions. The default requirement is four out of seven individual model predictions for a given class. Importantly, when the default cutoff is utilized, missing predictions from individual models are automatically excluded from the consensus calculation, requiring only 50% of available models to predict the designated class. Modification of this parameter changes the consensus calculation to require the specified number of predictions regardless of missing values, thereby implementing potentially stricter consensus requirements in rare cases when some of the individual methods don't provide a prediction.

Export new model functionality: The Advanced mode provides an additional output option for custom endpoint modelling, "Export as a new model" (Figure 3.11). When this option is selected alongside cross-validation of the training set, the trained models are packaged into an .rData format file that can be permanently integrated into the software's model library by copying the file into the software's endpoints subfolder. This functionality enables community-driven model development and sharing of new or refined models. Once integrated into the library, new models are available under the drop-down menu in both Basic and Advanced modes (see Section 3.2.4.5).

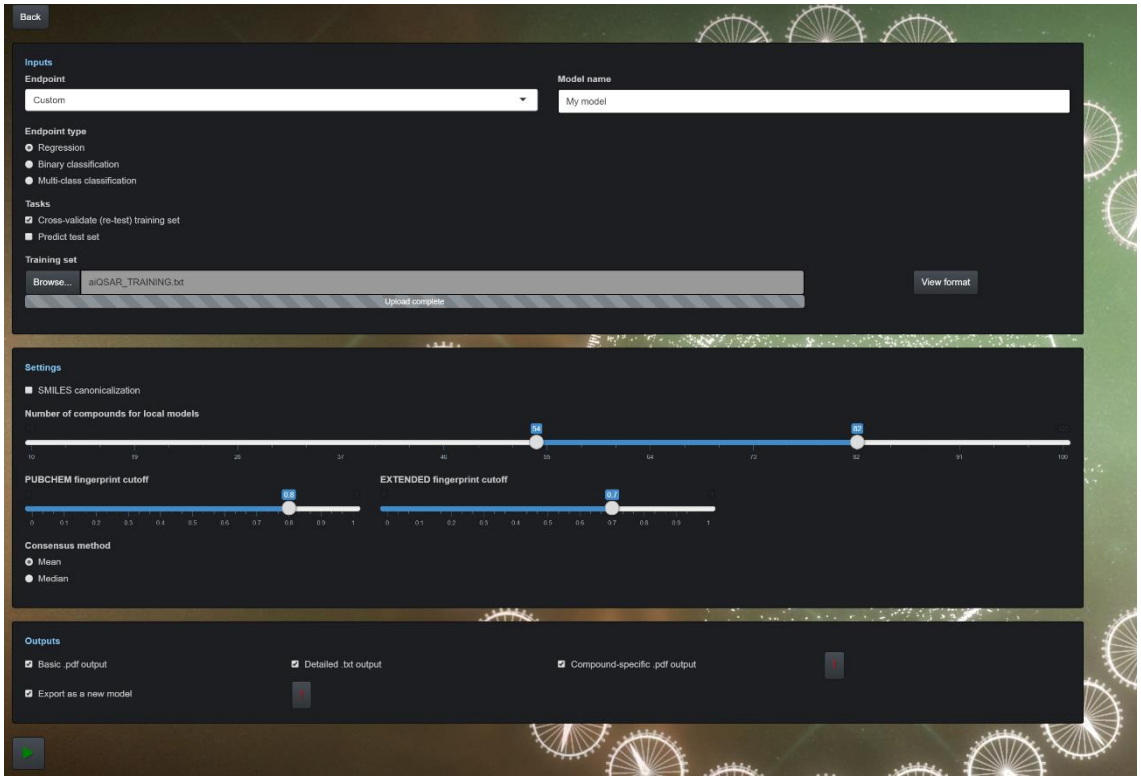


Figure 3.11: Advanced mode interface for model export functionality, showing options for saving trained models, including warnings about proper file naming conventions and instructions for subsequent model deployment.

3.2.2.3 Output and File Management

All aiQSAR outputs are stored in a shiny output directory created within the installation directory at first execution. Within this directory, organized subdirectories are automatically created for each analysis where required. Notably, compound-specific PDF outputs are stored in dedicated subdirectories with one output file per compound. The software implements automatic file naming and serialization protocols to prevent filename conflicts during multiple sequential analyses, appending numerical suffixes when required. Exported models are stored as .rData files suitable for sharing and subsequent use.

The organized file management approach ensures that users can locate specific results without confusion and facilitates integration of aiQSAR results into larger computational workflows. Additionally, the preservation of all output files serves as complete documentation of the computational analysis, essential for regulatory applications where traceability and reproducibility are required.

3.2.3 Endpoints

The aiQSAR software implements 38 pre-trained predictive models spanning multiple toxicological, biological and physicochemical endpoints. These endpoints can be organized into five categories based on the origin of their datasets. The comprehensive endpoint descriptions are presented here.

3.2.3.1 Acute Toxicity

The acute toxicity endpoint suite comprises five distinct models predicting rat oral acute toxicity. All models derive from the large-scale NICEATM initiative (see Section 3.1.3). This collaborative effort compiled and curated the most comprehensive publicly available dataset of rat acute oral LD_{50} values, encompassing 8,994 training set compounds and 2,998 external validation compounds [125]. For the final, implemented model, all training and validation set compounds were combined (for each endpoint separately). The five acute toxicity endpoints are defined as follows:

1. LD_{50} point estimates (regression)
2. Very toxic (VeryTox, binary classification)
3. Nontoxic (NonTox, binary classification)
4. EPA classification (multi-class, 4 categories) (Table 3.1)
5. GHS classification (multi-class, 5 categories) (Table 3.2)

3.2.3.2 Tox21 Data Challenge

The Tox21 (Toxicology in the 21st Century) initiative was a collaborative effort among the U.S. Environmental Protection Agency, National Institutes of Health and National Toxicology Program to provide publicly accessible in vitro screening data for thousands of environmental chemicals and pharmaceuticals [131]. The aiQSAR software implements 12 binary classification models from the Tox21 data challenge, predicting activity for the following assay endpoints:

1. Aryl hydrocarbon receptor (AhR): predicting activation of the aryl hydrocarbon receptor, a ligand-activated transcription factor involved in xenobiotic metabolism and immune response.
2. Androgen receptor (AR) full length: predicting full-length androgen receptor binding and activation, encompassing both ligand-binding domain and N-terminal domain interactions.
3. Androgen receptor ligand-binding domain (ARLBD): predicting binding specifically to the androgen receptor ligand-binding domain.
4. Antioxidant responsive element (ARE): predicting activation of the antioxidant responsive element, a regulatory sequence responding to oxidative stress.
5. Aromatase: predicting inhibition of cytochrome P450 aromatase (CYP19A1), the enzyme catalyzing estrogen biosynthesis from androgens.
6. ATAD5: predicting activation of a genotoxicity-responsive luciferase reporter in human embryonic kidney cells expressing ATAD5, identifying genotoxic compounds.
7. Estrogen receptor alpha (ER) full length: predicting estrogen receptor alpha binding and activation via the complete receptor structure.
8. Estrogen receptor alpha ligand-binding domain (ERLBD): predicting binding specifically to the estrogen receptor ligand-binding domain.
9. Heat shock factor response element (HSE): predicting activation of heat shock response, indicative of proteotoxic stress.
10. Mitochondrial membrane potential (MMP): predicting disruption of mitochondrial membrane potential, indicating mitochondrial toxicity.
11. p53: predicting agonistic activation of p53, the critical tumor suppressor protein.
12. Peroxisome proliferator-activated receptor gamma (PPAR-γ,): predicting activation of PPAR-γ, a nuclear receptor involved in metabolic regulation.

For all Tox21 endpoints, the binary classification of 1 is active, meaning that the compound demonstrates assay-specific activity above threshold and classification 0 is inactive, compound below activity threshold.

3.2.3.3 VEGA Software Endpoints

The VEGA (Virtual models for property Evaluation of chemicals within a Global Architecture) software provides publicly available QSAR models for large number of diverse endpoints [34]. The aiQSAR implementation incorporates 14 models derived from the VEGA software datasets:

1. Androgen receptor (binary classification): predicting androgen receptor mediated effects.
2. Bioconcentration factor (BCF, regression): predicting bioaccumulation potential, representing the degree of bioconcentration in aquatic organisms.
3. Carcinogenicity (binary classification): predicting carcinogenic hazard.
4. Daphnia EC_{50} (regression): predicting acute toxicity to daphnia magna, expressed as 48-hour median effect concentration.
5. Estrogen receptor (binary classification): predicting estrogen receptor-mediated effects.
6. Fathead minnow LC_{50} (regression): predicting acute toxicity to fathead minnow, expressed as 96-hour lethal concentration.
7. Fish acute toxicity (regression): predicting acute toxicity to multiple fish species.
8. Henry's law constant (regression): predicting air-water partitioning.
9. Half-Life/K_M (regression): predicting biological half-life.
10. LogP (regression): predicting octanol-water partition coefficient, a fundamental descriptor of lipophilicity.
11. Mutagenicity (binary classification): predicting mutagenic potential.
12. Thyroid receptor alpha (multi-class): predicting thyroid receptor alpha binding effects.
13. Thyroid receptor beta (multi-class): predicting thyroid receptor beta binding effects.
14. Water solubility (regression): predicting water solubility.

3.2.3.4 PubChem Assays

The PubChem bioassay database provides access to over 22 million compounds and comprehensive high-throughput screening (HTS) data from diverse biological assays [59]. The aiQSAR software implements three binary classification models derived from selected PubChem bioassays:

1. TLR4 inhibition: predicting inhibition of toll-like receptor 4 (TLR4) from the PubChem assay AID 941. TLR4 is a critical component of innate immune response. Inhibition indicates potential immunosuppressive or anti-inflammatory activity.
2. ROS induction: predicting induction of reactive oxygen species (ROS) from the PubChem assay AID 624156. ROS induction indicates oxidative stress potential and represents a key molecular initiating event in adverse outcome pathways.
3. NFκB activation: predicting activation of nuclear factor kappa-B (NFκB) from the PubChem assay AID 1159518. NFκB is a critical transcription factor in inflammatory and immune responses. Activation indicates immunostimulatory or pro-inflammatory potential.

All PubChem endpoints are binary classification with 1 indicating that chemical is active in the respective screening assay and 0 that it is inactive.

3.2.3.5 Mario Negri Institute Endpoints

In addition to these categories of published endpoint models, the aiQSAR software incorporates four additional endpoints developed at the Mario Negri Institute for Pharmacological Research (IRFMN) from the more recently published or, at the time of software release, yet unpublished research:

1. P-glycoprotein (multi-class): predicting P-glycoprotein (P-gp) interaction. P-gp represents the primary transporter affecting drug bioavailability and distribution. The multi-class output distinguishes between substrates (compounds actively transported), inhibitors (compounds blocking transport of endogenous or co-administered substrates) and non-active compounds. This classification is an important step in predicting drug-drug interactions and bioavailability [132].
2. Human intestinal absorption (binary classification): predicting the percentage human intestinal absorption (%HIA). The binary classification distinguishes high absorption (HIA $> 30\%$) and poor absorption (HIA $\leq 30\%$). This value is relevant in predicting oral drug bioavailability and gastrointestinal exposure to environmental contaminants [133].
3. Aromatase effect (multi-class): predicting aromatase activity. The model distinguishes agonists (facilitating estrogen synthesis), antagonists (inhibiting estrogen synthesis) and inactive compounds.
4. Drug-induced liver injury (binary classification): predicting the potential for drug compounds to induce hepatotoxicity *in vivo*. This endpoint distinguishes compounds with documented DILI liability from those lacking such evidence, providing potential identification of hepatotoxicity risk.

3.2.4 Outputs

The aiQSAR software generates results in four distinct output formats, each optimized for specific downstream applications and analytical requirements. All outputs are stored in organized subdirectories within the shiny output directory, with systematic file naming conventions enabling clear identification and archival of results.

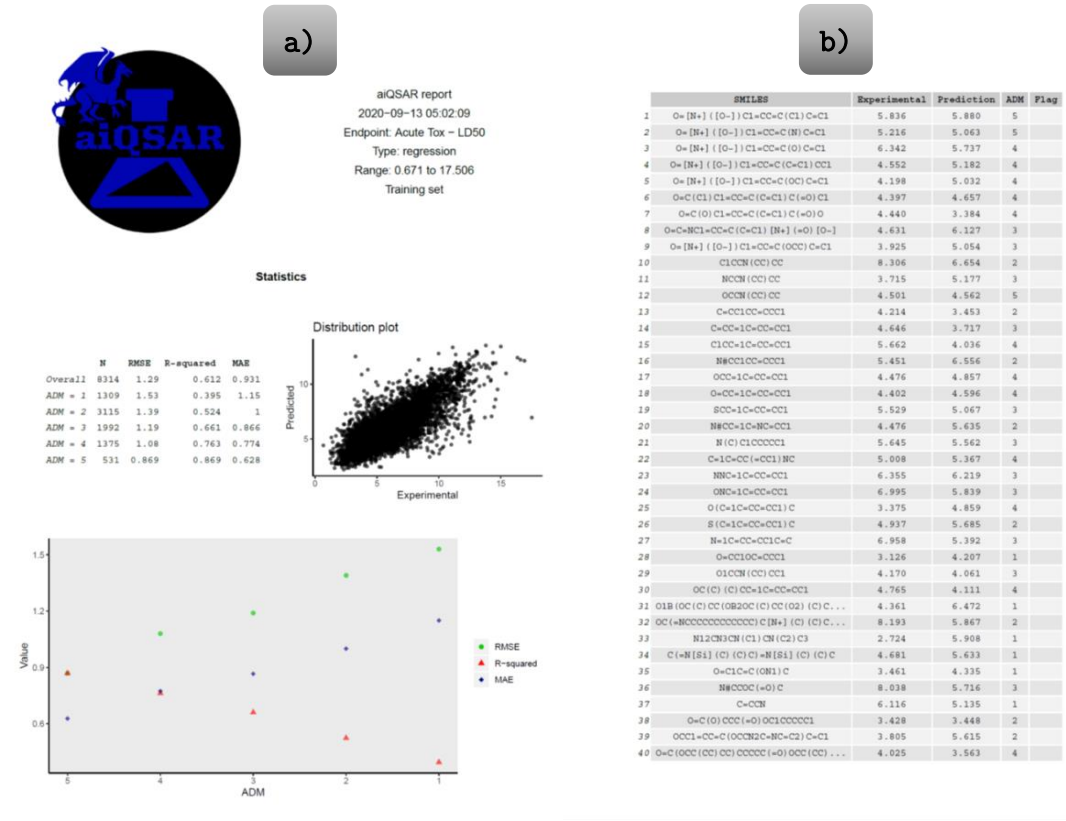
3.2.4.1 Basic PDF Output – Training

The basic PDF output is the standard reporting format, providing a concise summary of model performance, predictions and Applicability Domain Measures in a single portable document. The structure and content of basic output vary according to endpoint type, reflecting the distinct statistical measures appropriate for regression, binary classification and multi-class classification predictions respectively. This is the most common output type used to assess the performance and the quality of newly generated models. The summaries from the training set basic outputs for all 38 pre-implemented endpoints are presented in the Appendix A.

For regression endpoints, the basic training set output (Figure 3.12) contains:

- A header section specifying the endpoint, analysis timestamp, endpoint type and value range.
- Overall statistics table presenting sample size, RMSE, MAE and R^2 metrics.
- Scatter plot of experimental versus predicted values.

- Plot of performance metrics from the previous point plotted against the ADM category.
- A comprehensive predictions table listing all compounds with SMILES, experimental values, predicted values, ADM ranks and data quality flags.



	SMILES	Experimental	Prediction	ADM	Flag
1	O=[N+]([O-])C1=CC=C(C1)C=C1	5.836	5.880	5	
2	O=[N+]([O-])C1=CC=C(N)C=C1	5.216	5.063	5	
3	O=[N+]([O-])C1=CC=C(O)C=C1	6.342	5.737	4	
4	O=[N+]([O-])C1=CC=C(C=C1)CC1	4.552	5.182	4	
5	O=[N+]([O-])C1=CC=C(OC)C=C1	4.198	5.032	4	
6	O=C(C1)C1=CC=C(C=C1)C(=O)C1	4.397	4.657	4	
7	O=C(O)C1=CC=C(C=C1)C(=O)O	4.440	3.384	4	
8	O=C=NC1=CC=C(C=C1)[N+](=O)[O-]	4.631	6.127	3	
9	O=[N+]([O-])C1=CC=C(OCC)C=C1	3.925	5.054	3	
10	C1CCN(CC)CC	8.306	6.654	2	
11	NCCN(CC)CC	3.715	5.177	3	
12	OCCN(CC)CC	4.501	4.562	5	
13	C=CC1CC=CCC1	4.214	3.453	2	
14	C=CC=1C=CC=CC1	4.646	3.717	3	
15	C1CC=1C=CC=CC1	5.662	4.036	4	
16	N#CC1CC=CCC1	5.451	6.556	2	
17	OCC=1C=CC=CC1	4.476	4.857	4	
18	O=CC=1C=CC=CC1	4.402	4.596	4	
19	SCC=1C=CC=CC1	5.529	5.067	3	
20	N#CC=1C=NC=CC1	4.476	5.635	2	
21	N(C)C1CCCCC1	5.645	5.562	3	
22	C=1C=CC(=CC1)NC	5.008	5.367	4	
23	NNC=1C=CC=CC1	6.355	6.219	3	
24	ONC=1C=CC=CC1	6.995	5.839	3	
25	O(C=1C=CC=CC1)C	3.375	4.859	4	
26	S(C=1C=CC=CC1)C	4.937	5.685	2	
27	N=1C=CC=CC1C=C	6.958	5.392	3	
28	O=CC1OC=CCC1	3.126	4.207	1	
29	O1CCN(CC)CC1	4.170	4.061	3	
30	OC(C)(C)CC=1C=CC=CC1	4.765	4.111	4	
31	O1B(OC(C)CC(OB2OC(C)CC(O2)(C)C...	4.361	6.472	1	
32	OC(=NCCCCCCCCCCCC)C[N+](C)(C)C...	8.193	5.867	2	
33	N12CN3CN(C1)CN(C2)C3	2.724	5.908	1	
34	C(=N[Si](C)(C)C)=N[Si](C)(C)C	4.681	5.633	1	
35	O=C1C=C(ON1)C	3.461	4.335	1	
36	N#CCOC(=O)C	8.038	5.716	3	
37	C=CCN	6.116	5.135	1	
38	O=C(O)CCC(=O)OC1CCCCC1	3.428	3.448	2	
39	OCC1=CC=C(OCCN2C=NC=C2)C=C1	3.805	5.615	2	
40	O=C(OCC(CC)CC)CCCCC(=O)OCC(CC)...	4.025	3.563	4	

Figure 3.12: Portion of a basic PDF output for a training set regression endpoint. **a)** The header, overall statistics table, scatter plot of experimental versus predicted values and performance metrics across ADM categories. **b)** Portion of the predictions table.

For binary classification endpoints, the basic training set output (Figure 3.13) contains:

- A header section specifying the endpoint, analysis timestamp, endpoint type and value range.
- Confusion matrix for the full dataset.
- Overall statistics table presenting sample size, accuracy, balanced accuracy, kappa, MCC and the p-value.
- Plot of performance metrics from the previous point with respect to the ADM category.
- Confusion matrices and overall statistics for each ADM category.
- Plot of performance metrics from the previous point with respect to the cutoff of required predictions for a given class.
- Confusion matrices and overall statistics for each cutoff of required predictions for a given class.
- A comprehensive predictions table listing all compounds with SMILES, experimental values, predicted values, ADM ranks and data quality flags.

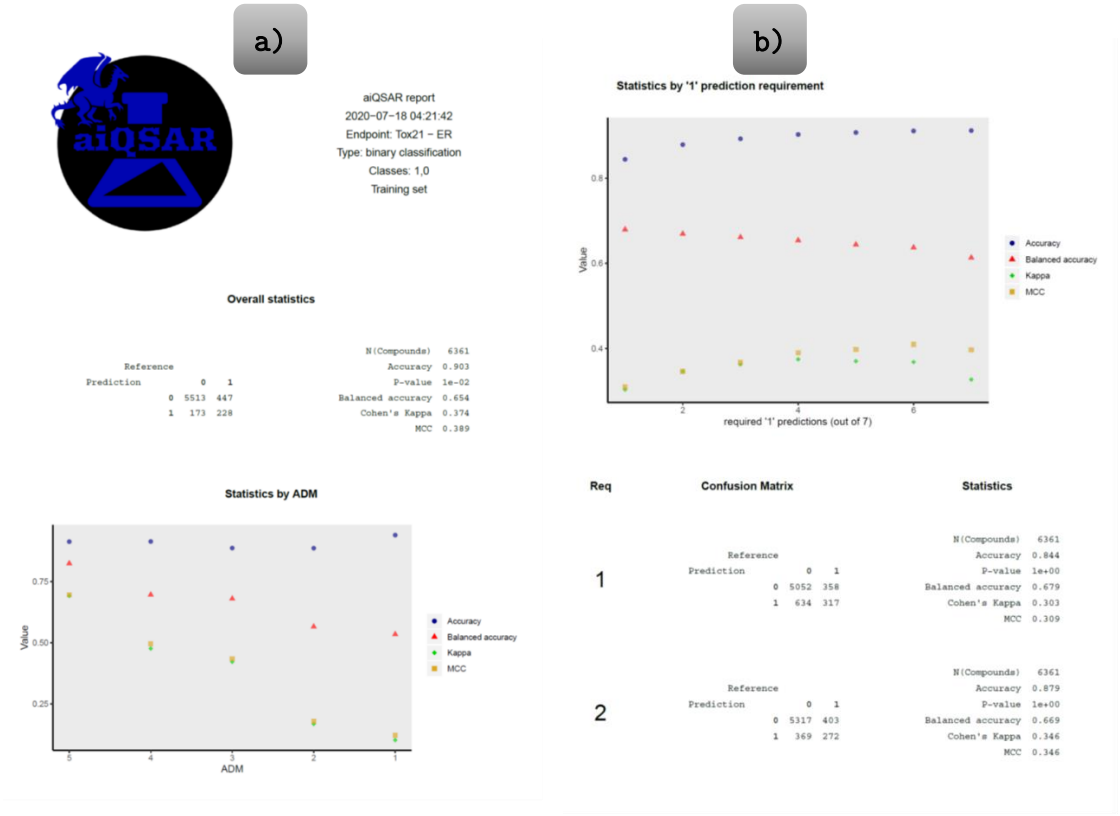


Figure 3.13: Portion of a basic PDF output for a training set binary classification endpoint. **a)** The header, confusion matrix, overall statistics and plot of performance metrics with respect to the ADM categories. **b)** Plot of performance metrics with respect to the number of required predictions for a given class (in this case "1", meaning positive) and the portion of confusion matrices and overall statistics for each cutoff of required predictions for a given class.

For multi-class classification endpoints, the basic training set output (Figure 3.14) contains:

- A header section specifying the endpoint, analysis timestamp, endpoint type and value range.
- Confusion matrix for the full dataset.
- Overall statistics table presenting sample size, accuracy, average balanced accuracy, Cohen's kappa and the p-value.
- Plot of performance metrics from the previous point with respect to the ADM category.
- Confusion matrices and overall statistics for each ADM category.
- A comprehensive predictions table listing all compounds with SMILES, experimental values, predicted values, ADM ranks and data quality flags.

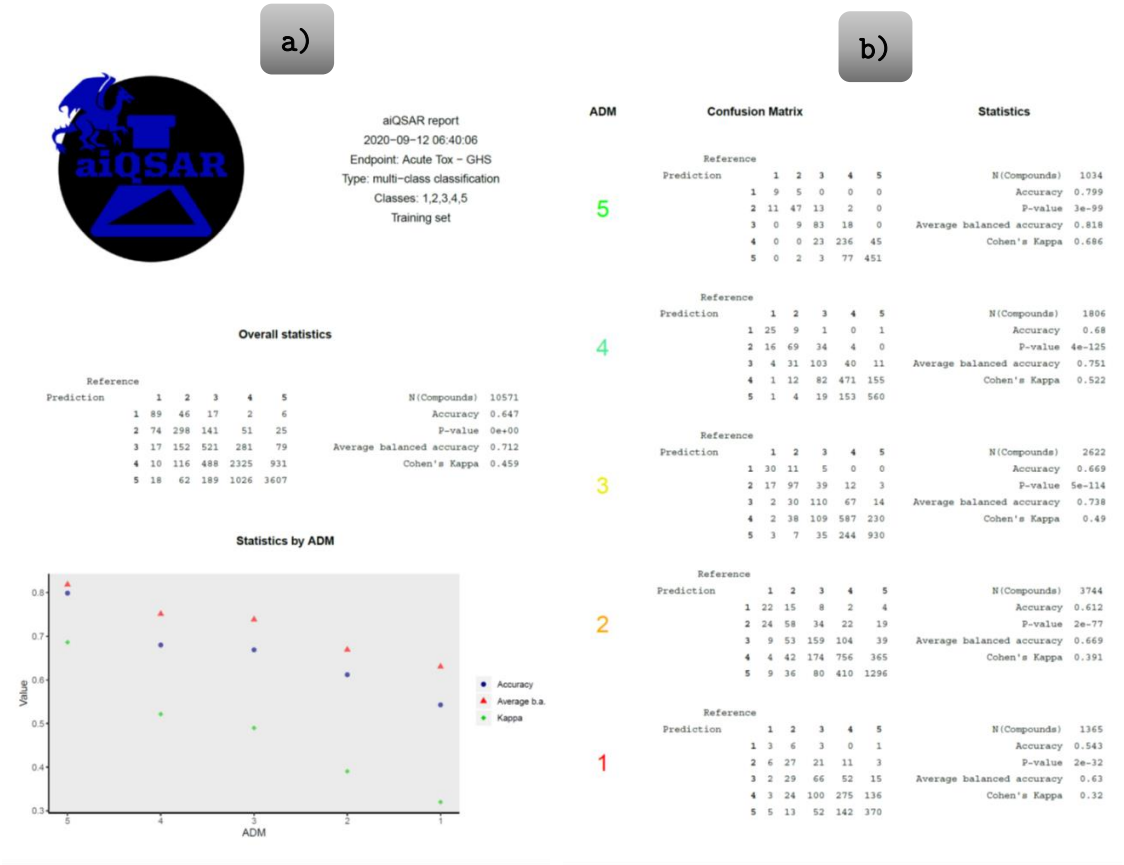
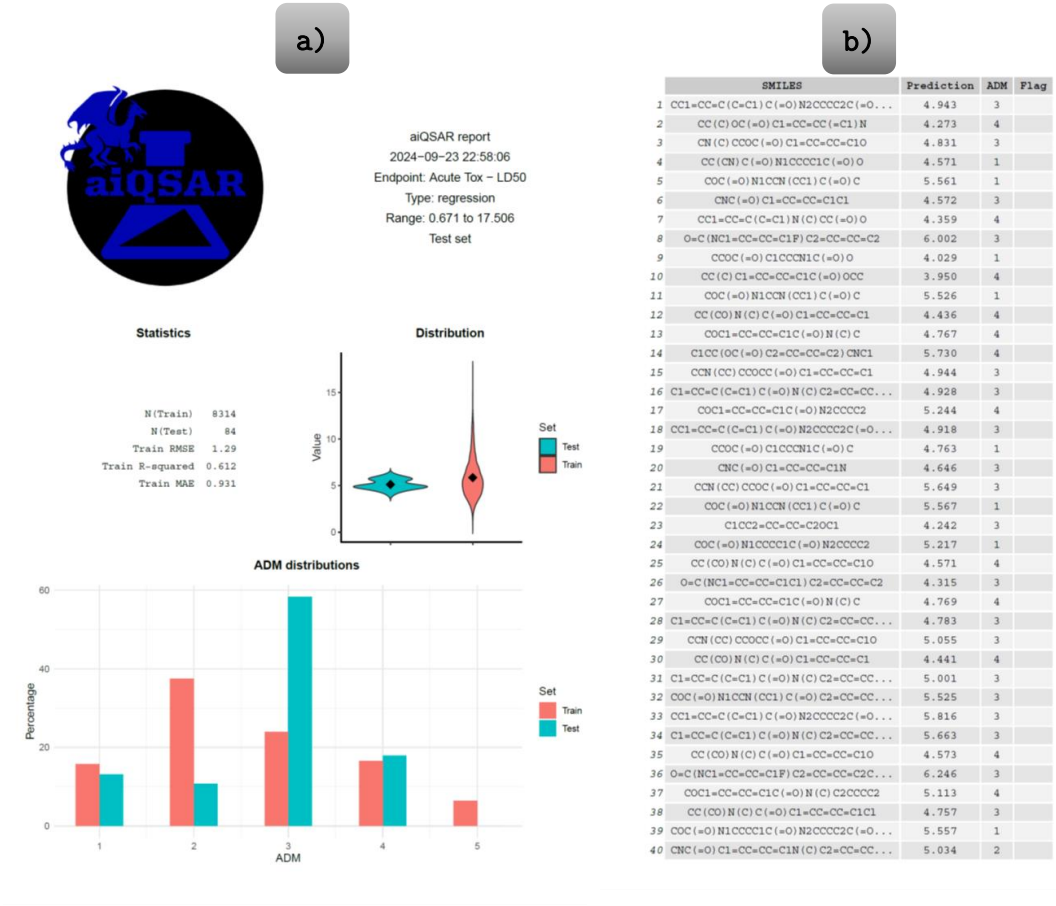


Figure 3.14: Portion of a basic PDF output for a training set multi-class classification endpoint. **a)** The header, confusion matrix, overall statistics and plot of performance metrics with respect to the ADM categories. **b)** Confusion matrices and overall statistics for each ADM category.

3.2.4.2 Basic PDF Output – Test

For regression endpoints, the Basic test set output (Figure 3.15) contains:

- A header section specifying the endpoint, analysis timestamp, endpoint type and value range.
- Summary statistics table presenting sample size, RMSE, MAE and R^2 metrics from the training set (if available).
- Violin plots comparing the training set and the predicted test set value distributions.
- Bar plot comparing the training set and the test set ADM rank distribution.
- A comprehensive predictions table listing all compounds with SMILES, predicted values, ADM ranks and data quality flags.



	SMILES	Prediction	ADM	Flag
1	CC1=CC=C(C=C1)C(=O)N2CCCC2C(=O...	4.943	3	
2	CC(C)OC(=O)C1=CC=CC(=C1)N	4.273	4	
3	CN(C)CCOC(=O)C1=CC=CC=C1O	4.831	3	
4	CC(CN)C(=O)N1CCCC1C(=O)O	4.571	1	
5	COC(=O)N1CCN(CC1)C(=O)C	5.561	1	
6	CNC(=O)C1=CC=CC=C1C1	4.572	3	
7	CC1=CC=C(C=C1)N(C)CC(=O)O	4.359	4	
8	O=C(NC1=CC=CC=C1F)C2=CC=CC=C2	6.002	3	
9	CCOC(=O)C1CCCN1C(=O)O	4.029	1	
10	CC(C)C1=CC=CC=C1C(=O)OCC	3.950	4	
11	COC(=O)N1CCN(CC1)C(=O)C	5.526	1	
12	CC(CO)N(C)C(=O)C1=CC=CC=C1	4.436	4	
13	COC1=CC=CC=C1C(=O)N(C)C	4.767	4	
14	C1CC(OC(=O)C2=CC=CC=C2)CNC1	5.730	4	
15	CCN(CC)CCOCC(=O)C1=CC=CC=C1	4.944	3	
16	C1=CC=C(C=C1)C(=O)N(C)C2=CC=CC...	4.928	3	
17	COC1=CC=CC=C1C(=O)N2CCCC2	5.244	4	
18	CC1=CC=C(C=C1)C(=O)N2CCCC2C(=O...	4.918	3	
19	CCOC(=O)C1CCCN1C(=O)C	4.763	1	
20	CNC(=O)C1=CC=CC=C1N	4.646	3	
21	CCN(CC)CCOC(=O)C1=CC=CC=C1	5.649	3	
22	COC(=O)N1CCN(CC1)C(=O)C	5.567	1	
23	C1CC2=CC=CC=C2OC1	4.242	3	
24	COC(=O)N1CCCC1C(=O)N2CCCC2	5.217	1	
25	CC(CO)N(C)C(=O)C1=CC=CC=C1O	4.571	4	
26	O=C(NC1=CC=CC=C1C1)C2=CC=CC=C2	4.315	3	
27	COC1=CC=CC=C1C(=O)N(C)C	4.769	4	
28	C1=CC=C(C=C1)C(=O)N(C)C2=CC=CC...	4.783	3	
29	CCN(CC)CCOCC(=O)C1=CC=CC=C1O	5.055	3	
30	CC(CO)N(C)C(=O)C1=CC=CC=C1	4.441	4	
31	C1=CC=C(C=C1)C(=O)N(C)C2=CC=CC...	5.001	3	
32	COC(=O)N1CCN(CC1)C(=O)C2=CC=CC...	5.525	3	
33	CC1=CC=C(C=C1)C(=O)N2CCCC2C(=O...	5.816	3	
34	C1=CC=C(C=C1)C(=O)N(C)C2=CC=CC...	5.663	3	
35	CC(CO)N(C)C(=O)C1=CC=CC=C1O	4.573	4	
36	O=C(NC1=CC=CC=C1F)C2=CC=CC=C2C...	6.246	3	
37	COC1=CC=CC=C1C(=O)N(C)C2CCCC2	5.113	4	
38	CC(CO)N(C)C(=O)C1=CC=CC=C1C1	4.757	3	
39	COC(=O)N1CCCC1C(=O)N2CCCC2C(=O...	5.557	1	
40	CNC(=O)C1=CC=CC=C1N(C)C2=CC=CC...	5.034	2	

Figure 3.15: Portion of a basic PDF output for a test set of regression endpoint. **a)** The header, summary statistics table, violin plots comparing the training and the test set value distributions and a bar plot comparing the distribution of ADM ranks. **b)** Portion of the predictions table.

Binary classification and multi-class classification endpoints have very similar basic PDF test set outputs (Figure 3.16) containing:

- A header section specifying the endpoint, analysis timestamp, endpoint type and value range.
- Summary statistics table presenting sample size, accuracy, (average) balanced accuracy, kappa and the p-value.
- Pie charts comparing the training and the predicted test set value distributions.
- Bar plot comparing the training set and the test set ADM rank distribution.
- A comprehensive predictions table listing all compounds with SMILES, predicted values, ADM ranks and data quality flags.

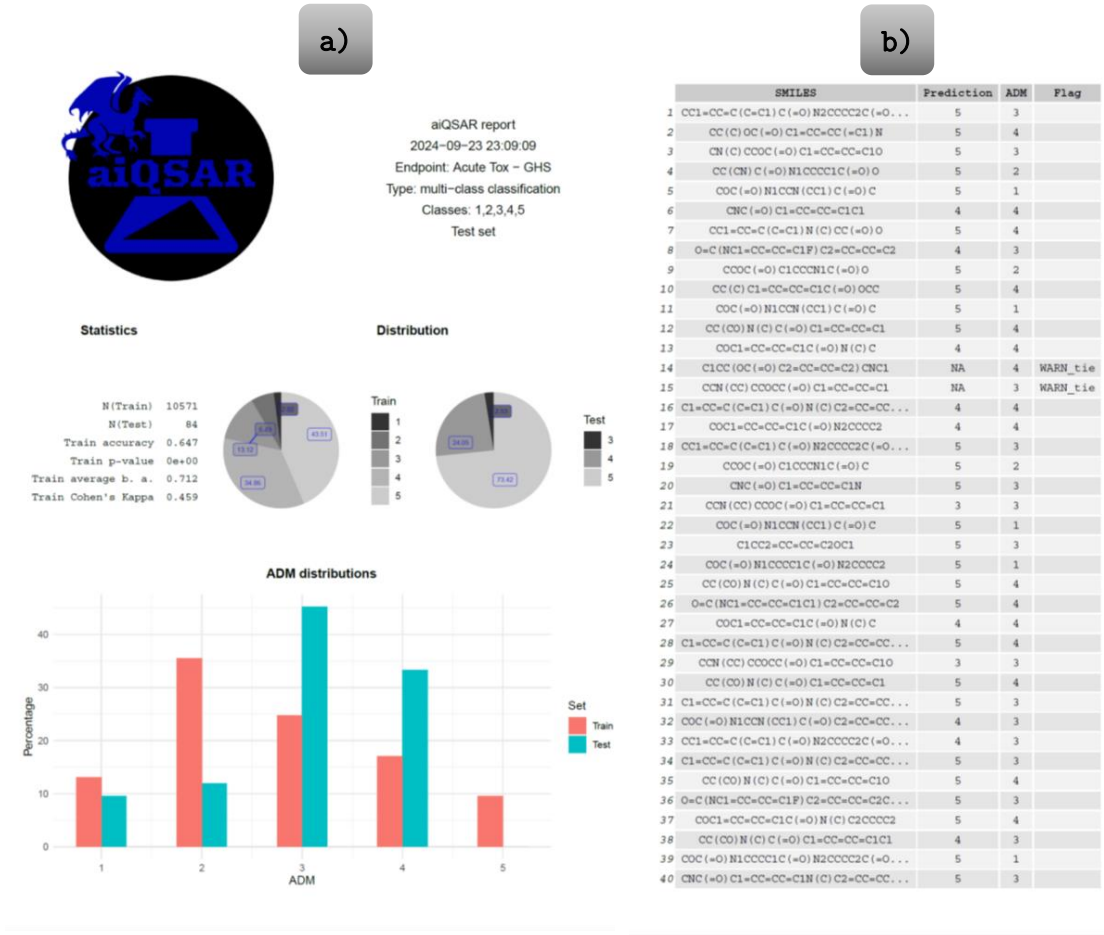


Figure 3.16: Portion of a basic PDF output for a test set of a multi-class classification endpoint. **a)** The header, summary statistics table, pie charts comparing the training and the test set value distributions and a bar plot comparing the distribution of ADM ranks. **b)** Portion of the predictions table.

When training set models for new endpoints were not explicitly evaluated, the training set sections of the test set basic output remain blank. In this case, the predictions are generated purely from the pre-trained model during the same run. This is possible in a situation where user provides a training set, but only requests the test set predictions (likely because they have obtained the training set performance in a previous run, or at a different computing location).

3.2.4.3 Detailed TXT Output

The detailed text output provides extended tabular results optimized for downstream computational analysis and integration into external data management systems. This output is generated as a tab-separated text file suitable for import into spreadsheet applications or data analysis platforms.

The detailed output contains all information presented in basic outputs, with supplementary detail on intermediate computational steps. Specifically, for each target compound, detailed output provides: SMILES representation, experimental endpoint value (if available), predicted endpoint value, ADM rank, exact average Tanimoto fingerprint similarities (PubChem and Extended) between the target compound and compounds in its local group, predictions from each individual machine learning method and all quality control flags.

With detailed output, user can inspect individual method results and local group composition to gain a deeper understanding of aiQSAR predictions and to conduct further custom analyses.

3.2.4.4 Single-Compound Output

When requested, the software generates individual PDF documents for each compound analyzed, whether from the training set, test set, or single-SMILES queries. Each single-compound PDF provides detailed information specific to that chemical structure and its prediction.

For each target compound, a single-compound PDF output presents (Figure 3.17, Figure 3.18):

- The endpoint information.
- A 2D molecular structure visualization of the target compound.
- The predicted endpoint value with the range of individual predictions, or the distribution of predictions, for regression or classification endpoints respectively.
- The ADM rank.
- 2D structures of all local group compounds, ranked by the decreasing similarity to the target.
- For each local group compound, the experimental endpoint value with PubChem and Extended Tanimoto fingerprint similarities to the target.

This output type provides intuitive chemical insight into model predictions. Users examining single-compound outputs can visually assess local group quality, judge whether selected analogues are chemically reasonable, evaluate the range of experimental values within the local group and use this information to gain further confidence in the reliability of predictions beyond simple scalar value of the ADM category. This transparency is particularly valuable for regulatory applications where model predictions must be justified to external stakeholders.

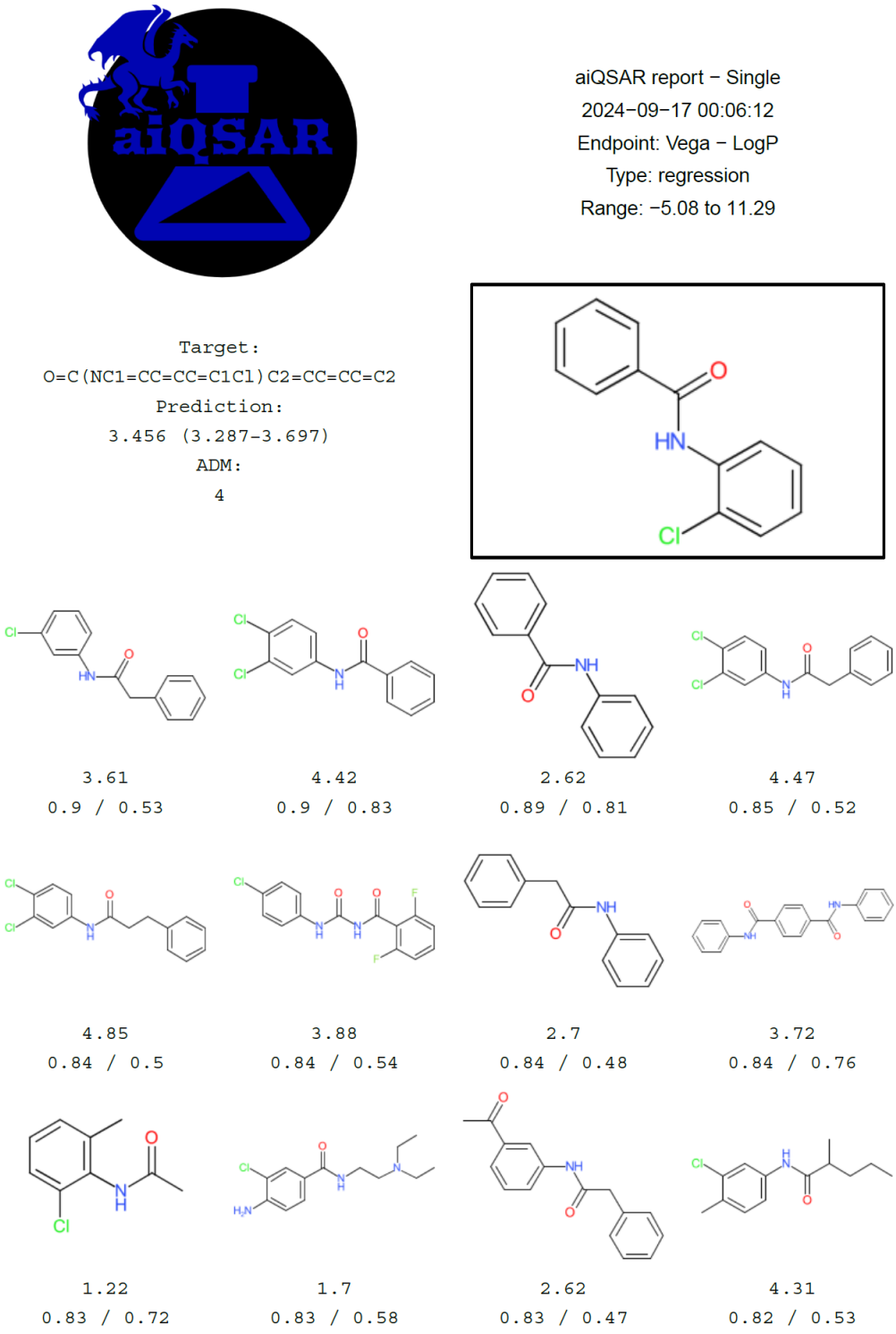


Figure 3.17: First page of a single-compound output for the logP endpoint of a randomly selected chemical with default value (30) for the local group size parameter. This page showcases the 12 most similar local group chemicals.

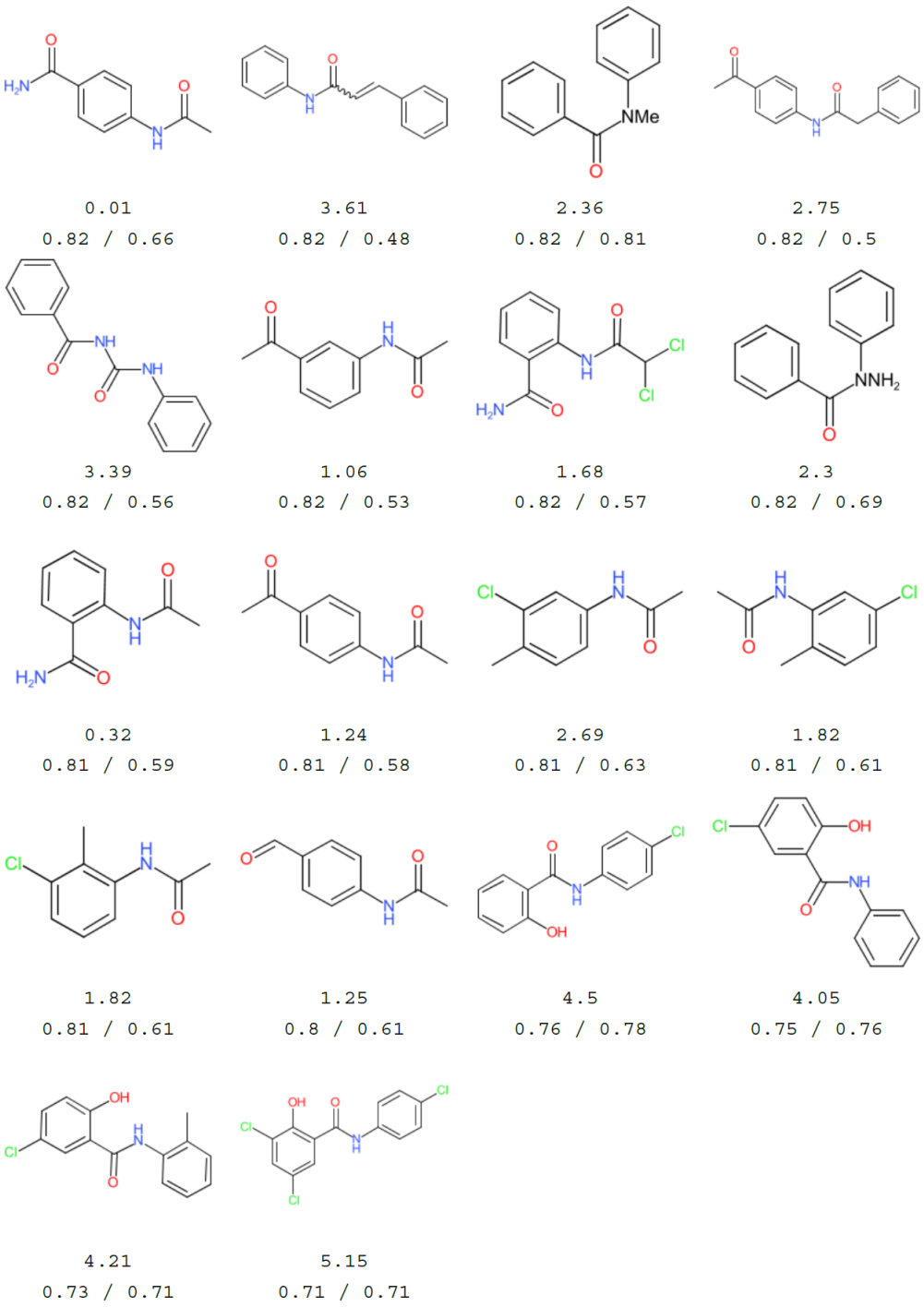


Figure 3.18: Second page of a single-compound output for the logP endpoint of a randomly selected chemical with default value (30) for the local group size parameter. This page showcases the remaining 18 local group chemicals.

3.2.4.5 New Model Export

For researchers developing custom QSAR models, the software provides an "Export as a new model" functionality that generates a persistent model file in ".rData" binary format. This output option is available exclusively in Advanced mode when "custom" endpoint is selected and "Cross-validate (re-test) training set" is the only task specified. Exporting the training set as a new model will create a "model_name.rData" file in the "…/shiny/output" directory. This ".rData" file contains all necessary information for the trained model: the descriptor matrix, calculated fingerprints and machine learning model objects for all five (regression), seven (binary classification), or five (multi-class classification) individual methods.

To use an exported model in future aiQSAR sessions, users must: (1) exit aiQSAR completely; (2) copy the ".rData" file from the output directory to the "endpoints" subfolder in the aiQSAR installation directory; and (3) restart aiQSAR. The new endpoint will subsequently appear in the endpoint selection list, available for predictions on new test sets and single SMILES.

Users should not manually rename the ".rData" file at any time, including the automatically generated extension in parentheses! The naming convention is essential for aiQSAR's internal file recognition system. The newly created models can be shared externally: Anyone can utilize the custom model by copying the received ".rData" file into their "endpoints" subfolder. This architecture facilitates collaborative model development and enables rapid dissemination of optimized models within the scientific community.

3.3 Predicting Differential Gene Expressions

The human genome contains approximately 20,000 protein-coding genes that give rise to a transcriptome of significantly greater complexity due to alternative splicing and post-translational modifications [134]. Gene expression, however, is not uniform. The dynamic regulation of gene expression is a fundamental biological process and understanding how chemical exposures alter gene expression patterns is crucial for predictive toxicology, pharmacogenomics and the development of mechanism-of-action frameworks for environmental and pharmaceutical compounds.

The following sections present the technological foundation for measuring the differential gene expression, the methodological approach for applying QSAR to gene expression predictions, comprehensive performance assessment across 478 genes and mechanistic insights into which biological processes are predictable from chemical structure.

3.3.1 Motivation

Traditional whole-transcriptome sequencing provides comprehensive characterization of gene expression changes but remains prohibitively expensive for large-scale screening. This cost barrier has necessitated the development of targeted transcriptomics approaches that focus sequencing efforts on a predefined gene panel while maintaining statistical power for differential expression analysis. The TempO-Seq platform represents a significant technological advancement, enabling simultaneous processing of thousands of biological samples in a single sequencing run through sample barcoding, while restricting analysis to a curated library of 3,257 scientifically relevant genes (see Section 2.5.3.1).

The integration of computational chemistry with functional genomics offers an opportunity to predict how novel chemicals will affect transcriptomic responses. The

underlying hypothesis is that chemical structure, as quantified through molecular fingerprints and chemical descriptors, contains sufficient information to predict which genes will be significantly altered by chemical exposure and in which direction. This work demonstrates how chemoinformatic predictions can complement and potentially reduce the need for expensive transcriptomics experiments, establishing a bridge between chemical structure and biological response space. Rather than relying on expensive and time-consuming experimental screening of every compound across every potential gene, QSAR predictions offer a cost-effective and rapid alternative for prioritizing which compounds require further experimental investigation. This is particularly important given millions of chemicals that are registered globally and detailed transcriptomic data being available for only a tiny fraction.

3.3.2 TempO-Seq and Differential Expression

The TempO-Seq platform is a targeted RNA-sequencing technique designed to reduce the cost and complexity of large-scale gene expression profiling, while maintaining statistical robustness. The 3,257-gene panel was curated to represent genes involved in diverse biological processes relevant to toxicological assessment, including stress response pathways, metabolic processes, apoptosis, inflammation and cell cycle regulation. This expert-curated panel represents approximately 15% of the protein-coding genome.

3.3.2.1 Experimental Design and Sample Preparation

In the experimental workflow, the renal proximal tubule-like cells (derived from human kidney cells and commonly employed as a model of renal toxicity) were exposed to 220 distinct chemicals at a concentration of 10 μM. This is standard concentration, commonly selected as a reference for toxicology testing. It balances the sensitivity in detecting true biological responses, with physiological relevance (avoiding overwhelming non-specific stress responses at extremely high concentrations). After 24 hours of treatment, total RNA was extracted from treated and control cells and subjected to the TempO-Seq procedure.

Each chemical treatment was represented by three independent biological replicates, meaning that three separate cell culture experiments were performed for each compound. Control cells were represented by six replicates to provide a highly stable baseline for comparison. This design structure of employing more control replicates than treatment replicates is a standard approach in transcriptomics that leverages the greater stability of baseline expression levels in comparison to the treated variants.

Following the experimental procedure, the full TempO-Seq workflow was applied to all cell samples, together with the associated bioinformatics pipeline. This procedure generated a gene count table as an output (see Section 2.5.3). Overall, the final gene count matrix consisted of 220 chemicals (columns) $\times$ 3,257 genes (rows), plus 6 additional columns for the control samples.

3.3.2.2 Differential Expression Quantification

Differential expression analysis aims to identify genes whose expression levels differ significantly between treatment and control conditions (Figure 3.19), accounting for both the magnitude of change and the statistical confidence in that change. This requires integration of two complementary metrics: fold-change (magnitude) and p-value (confidence).

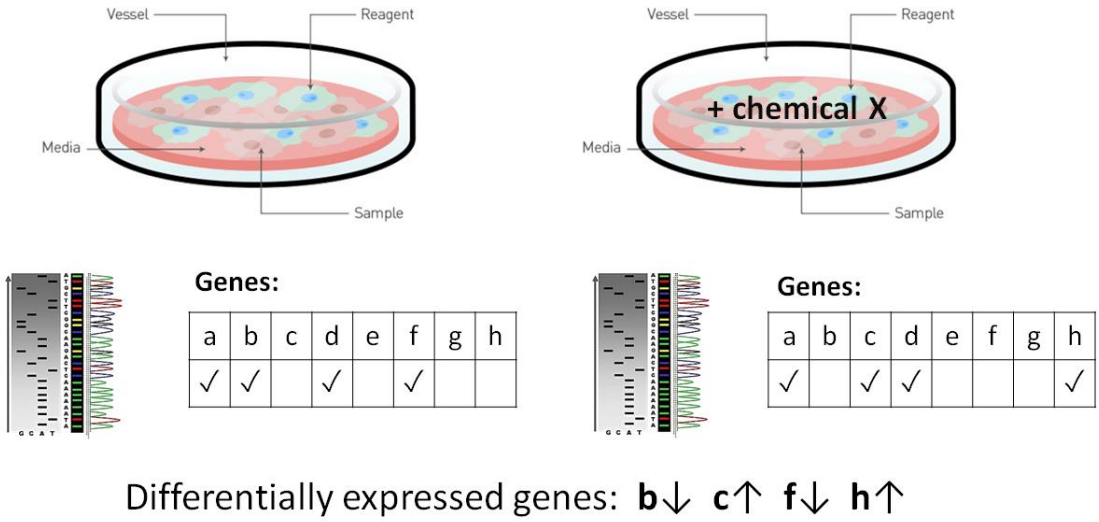


Figure 3.19: Schematic of the TempO-Seq experiment and the differential expression analysis. After treating the sample with the chemical X, genes 'c' and 'h' are observed to be upregulated (over-expressed), genes 'b' and 'f' are downregulated (under-expressed) and the remaining genes exhibit no differential expression.

Gene expression differences between treated and untreated cells are quantified using the $\log_2$-fold-change (ltfc), defined as:

$$ltfc = \log_2\left(\frac{\text{Mean}_{\text{treated}}}{\text{Mean}_{\text{control}}}\right) \tag{3.1}$$

The adjusted p-value (p_{adj}) is the statistical probability of the observed difference being significant after accounting for multiple hypothesis testing. With a large number of independent statistical tests, there is a substantial probability of observing extreme fold-changes by chance alone. Additionally, different gene lengths contribute to the baseline variation of the expected counts in the experimental sequencing procedure. To control for this multiple testing problem, the Benjamini-Hochberg false discovery rate correction is used, which is a default statistical correction in the TempO-Seq computational analysis pipeline [87].

A standard significance threshold of $p_{adj} \leq 0.05$ is employed, below which differential expression is considered statistically significant. This threshold corresponds to an expected false discovery rate of 5%, meaning that among genes called significant, approximately 95% are true differential expression events.

3.3.3 Exploratory Data Analysis

The TempO-Seq analysis observed differential expression values for 3,257 genes across 220 chemical treatments. This resulted in a matrix of approximately 700,000 potential gene-chemical combinations, of which the vast majority exhibited no statistically significant differential expression. Specifically, around 20,000 statistically significant events were detected, leaving approximately 97% of all gene-chemical pairs with no significant change ($p_{adj} > 0.05$). This reflects the selective nature of gene regulation where each gene responds to a limited subset of chemicals rather than responding broadly across the entire chemical library space.

3.3.3.1 Differential Expression Distribution

The distribution of gene responsiveness across the chemical library provides insight into the breadth of chemical effects and the selectivity of gene regulation (Figure 3.20). The strongly skewed distribution reveals important properties of gene regulation, with most genes responding to a relatively narrow subset of compounds rather than serving as universal stress response markers. This selectivity likely reflects the specialized regulatory mechanisms controlling each gene. The small subset of genes occupying the right tail of the distribution (>50 chemicals) likely represent general cellular stress response genes that are activated by diverse mechanisms of toxicity. These genes might essentially act as "canaries" signaling that the cell has experienced stress, regardless of the specific chemical trigger involved.

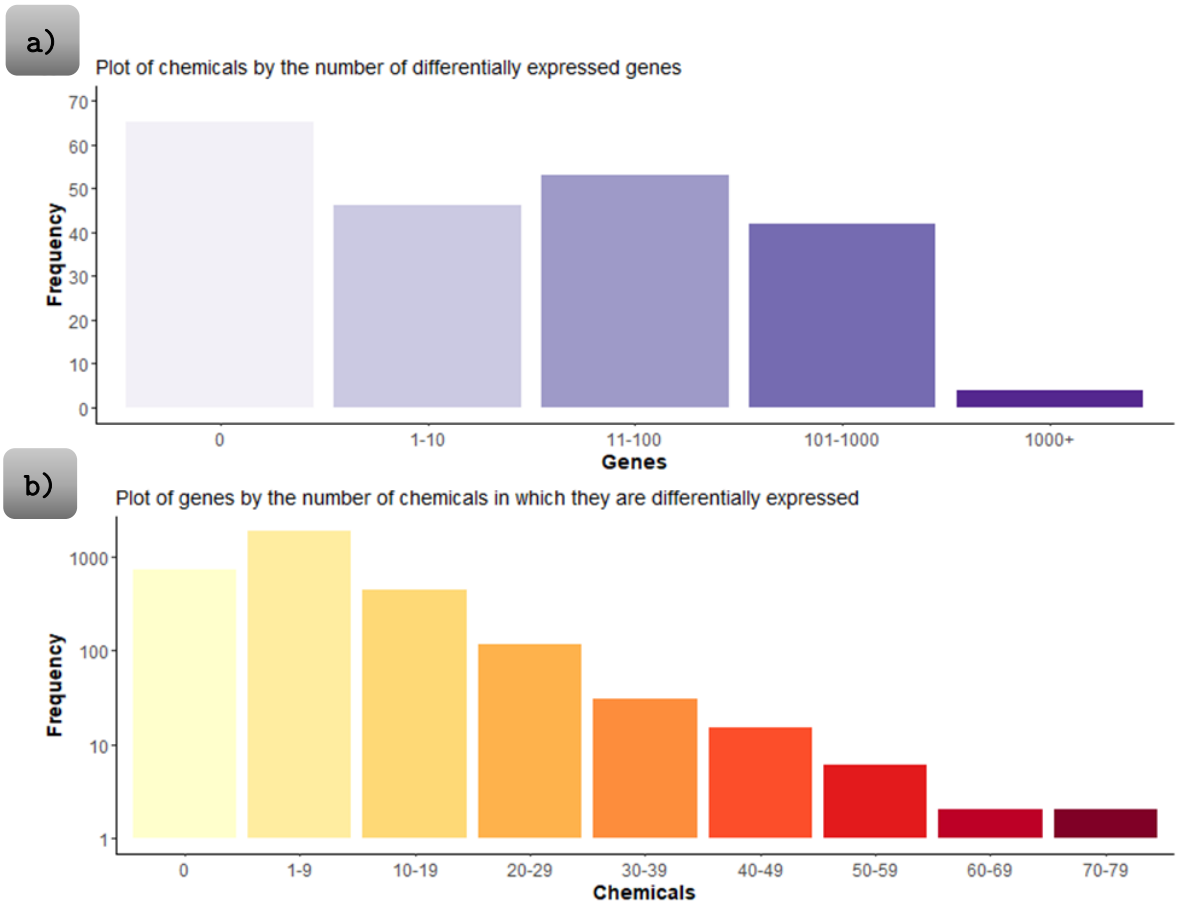


Figure 3.20: Differential expression exploratory data analysis. **a)** Distribution of genes by chemical response frequency. **b)** Distribution of chemicals by gene response frequency.

This distribution pattern has important implications for QSAR modelling: genes on the left side of the distribution (selective, few chemicals) may be highly predictable because their regulation is tightly coupled to specific structural features. Conversely, genes on the right side (general stress response, many chemicals) may be less predictable because their activation can be triggered through multiple distinct structural pathways and mechanisms. This hypothesis is extensively tested later (see Section 3.3.4.3).

Furthermore, this skewed distribution is consistent with studies of chemical-induced gene expression changes in diverse cell types and endpoints. For example, estrogen receptor activation [135] and acute liver toxicity responses [136] show similarly skewed distribution of differential gene expression indicating a fundamental biological principle rather than an artifact of the TempO-Seq experimental design.

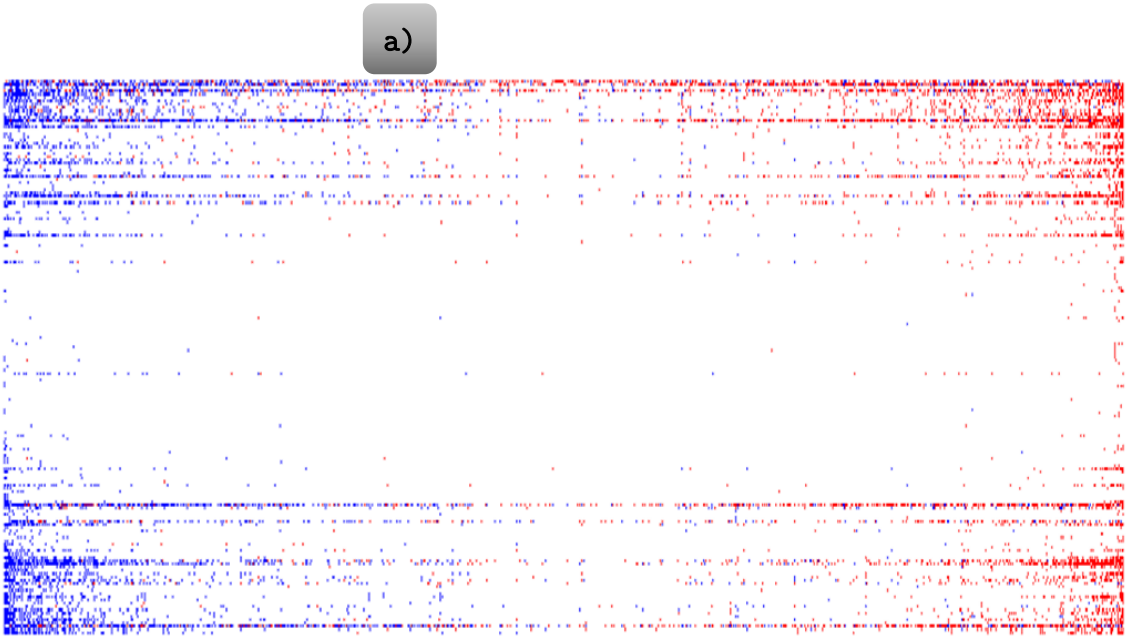
3.3.3.2 Directional Patterns in Differential Expression

An important empirical observation emerged from the exploratory data analysis: differential expression exhibited marked directional asymmetry rather than random, bidirectional changes. For the vast majority of genes, differential expression, when it occurred, proceeded predominantly in a single direction in response to a wide variety of different chemicals. Only a small fraction of genes demonstrated mixed expression patterns, showing upregulation in response to some chemicals and downregulation in response to others. This separation is striking and non-random, suggesting fundamental constraints on gene regulatory logic. Understanding this asymmetry provides both mechanistic insights and practical modelling implications (Figure 3.21).

In Figure 3.21a, the left side of the matrix corresponds to genes that are most downregulated and this section is expected to be mostly blue. Same with the right-side for the upregulated genes. However, the middle part is strikingly empty, meaning that there are almost no genes that are downregulated by some chemicals and upregulated by an approximately equal number of other chemicals (which would add up close to zero). This section is therefore populated exclusively by genes that are not affected in either direction and remains mostly white. In terms of the chemicals (rows), the situation is different: The top section corresponds to the chemicals that cause the upregulation of most genes. However, in this section there are a lot of blue positions as well, meaning that even the chemicals that upregulate the largest number of genes, downregulate quite a few genes as well. The same is true for the bottom section of the plot: This section corresponds to the chemicals that downregulate the most genes, but these chemicals also upregulate a significant number of genes. This separation of genes into the downregulated and the upregulated ones is observed most clearly for the genes that are differentially expressed by the largest number of chemicals. Looking at the top 10 such genes in Figure 3.21b, it is evident that they are all either almost exclusively upregulated or exclusively downregulated.

Biologically, this means that basically all genes fall into the category of those that can either be upregulated, or downregulated by the exposure to a single chemical. For the extreme cases this is quite expected: If the level of a gene expression in a cell is very low, it's difficult to reduce it even further, at least to the point that would be statistically significant. Conversely, if the expression is very high, increasing it further, to the statistically significant point, doesn't seem to happen frequently. However, these results indicate this to be the case for almost all genes and, as a consequence of this biological behavior, there are almost no genes that can be both significantly upregulated by some chemicals and significantly downregulated by others.

Binary classification is a preferred modelling approach for this type of biologically complex set of endpoints and, given the described bias of expression directionality, should be even more favored compared to the continuous predictions of fold-change magnitudes. By leveraging the intrinsic directional structure of cellular responses, the binary classification modelling approach aligns with underlying biological organization and is used in subsequent analyses.



b)

↓		↑
1	144	75
1	144	75
76	144	0
1	156	63
59	161	0
0	162	58
53	164	3
53	165	2
55	165	0
0	165	55

Figure 3.21: Experimental differential expression results. **a)** Complete differential expression heatmap with genes in columns and chemicals in rows. Each position corresponds to a gene either upregulated (red), downregulated (blue), or unaffected (white) by a given chemical. The matrix is sorted from left to right and from bottom to top by the ascending sum of values where downregulation adds -1 and upregulation adds 1. **b)** Table of the genes that are differentially expressed by the largest number of chemicals. Columns from left to right: Number of chemicals that downregulate, do not affect and upregulate the gene.

3.3.3.3 Dataset Processing and Filtering

To identify genes amenable to QSAR modelling, the dataset was filtered to include only genes where at least 11 chemical treatments cause statistically significant differential expression ($p_{adj} \leq 0.05$) in either direction (upregulation or downregulation). This threshold was selected to return the sufficient number of data points for statistical analysis while at the same time avoiding severe class imbalance in binary classification endpoints, ensuring that at least 5% of the dataset for each gene corresponds to the minority class.

This filtering procedure retained 478 genes (14.7% of the total panel). For each of the 478 genes, the complete aiQSAR workflow was executed as an independent modelling exercise. Furthermore, 10 out of 220 tested chemicals were considered duplicates for the purpose of QSAR modelling since they were the equivalent chemical compounds with different counterions or in different mixtures. These compounds were used in all analyses and for the gene selection procedure, but were merged for the final aiQSAR modelling exercise.

3.3.4 aiQSAR Results

The application of aiQSAR methodology to the TempO-Seq differential gene expression dataset yielded predictions for all 478 genes across the 210 unique chemical treatments. To comprehensively evaluate whether model predictions exceeded random chance expectations, multiple statistical measures were computed for each gene independently and analyzed both individually and across aggregated performance distributions. The results provide quantitative evidence that the local QSAR approach is capable of capturing structure-activity relationships governing differential gene expression, establishing that chemical structure encodes substantial predictive information regarding transcriptional responses.

Binary classification was employed for all tested genes due to the complexity of the endpoint and because preliminary exploratory analysis revealed marked directional asymmetry in differential expression. For the vast majority of genes, expression changes proceed predominantly in a single direction (upregulation or downregulation) in response to diverse chemicals, with bidirectional responses (upregulation by some chemicals, downregulation by others) remaining rare. All chemicals were therefore classified as either active, in case they cause the differential expression in any direction (with the requirement $p_{adj} \leq 0.05$) or inactive, if they cause no differential expression of the targeted gene.

3.3.4.1 Performance Evaluation

Final calculation generated approximately 100,000 models (478 x 210), each consisting of seven individual machine learning methods, as employed in the aiQSAR methodology (see Section 3.1.2.1) with the default majority vote (4 out of 7) required for the prediction of a given outcome. All performance metrics were calculated, treating each gene as a separate, independent endpoint (Figure 3.22). The full list of 210 tested chemicals, as well as the list of 478 targeted genes, with the corresponding dataset information and complete endpoint performance metrics are presented in the Appendix B.

All four metrics display pronounced distributions shifted substantially toward higher positive values, indicating that the majority of genes achieve performance substantially exceeding random expectation. Accuracy shows a narrow distribution centered around 0.85-0.90. Balanced accuracy, which weights sensitivity and specificity equally and is therefore more robust to class imbalance, shows a somewhat broader distribution centered around 0.55. Cohen's kappa, the primary metric accounting for chance agreement, displays a distribution with median value substantially greater than zero, indicating that predictions systematically exceed random guessing. Matthews Correlation Coefficient, which integrates all elements of the confusion matrix into a single correlation coefficient, similarly shows a distribution shifted toward positive values. The concordance across all four metrics provides robust evidence that genuine predictive signal is present.

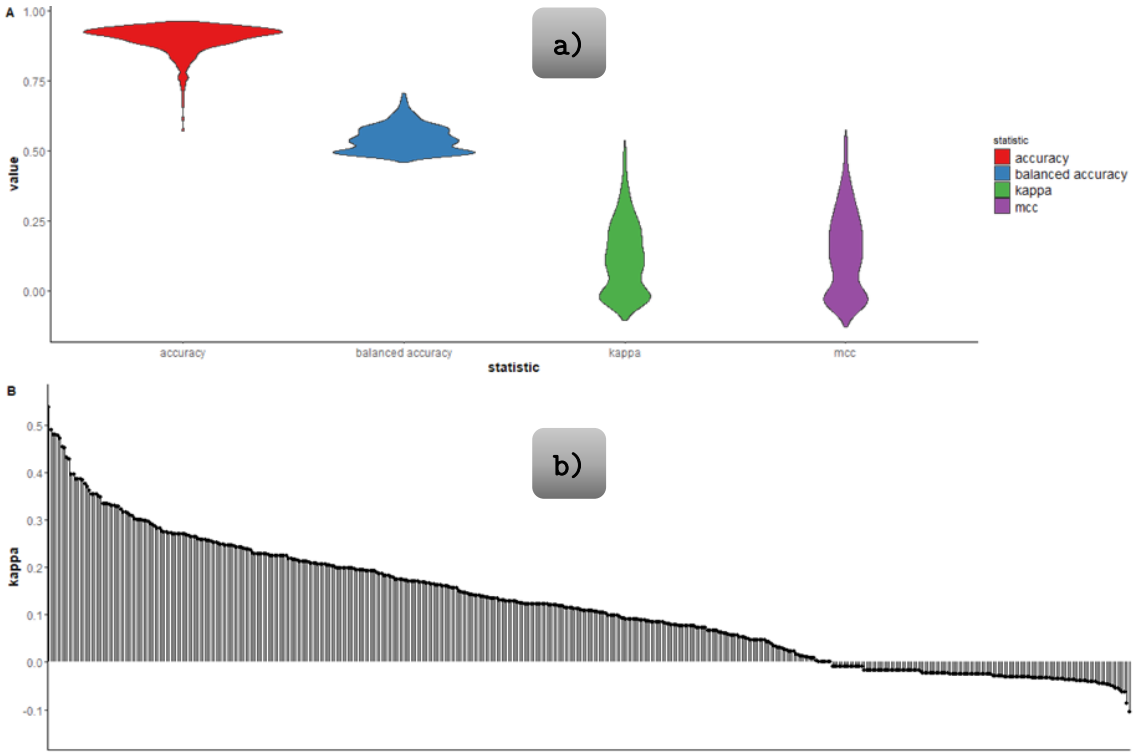


Figure 3.22: A comprehensive visualization of performance metrics for differential gene expression predictions of all 478 genes. **a)** Violin plots displaying the distribution of four classification metrics (accuracy, balanced accuracy, kappa and MCC). **b)** Detailed distribution of kappa values across the whole gene set.

The full distribution of kappa values is distinctly asymmetric, with substantial density concentrated in the positive region and minimal density in the negative region. If chemical structure carried no predictive information regarding differential gene expression and all predictions were indistinguishable from random chance, kappa values would be expected to cluster tightly around zero, following approximately a normal distribution, with equal proportions of positive and negative values. The observed pronounced leftward skew toward positive values constitutes direct empirical evidence that aiQSAR models systematically capture genuine structure-activity relationships and can be confidently used for predictions of differential expression outcomes on a substantial subset of genes analyzed by the TempO-Seq experiment.

3.3.4.2 Statistical Analysis of Top 10 Genes

Even though the overall performance of the aiQSAR methodology shows a significant predictive capacity, this isn't evenly distributed across the targeted genes. Given the lowest level of the kappa values distribution at approximately -0.1, for which there is clearly no predictive capacity, similar cutoff can be hypothesized on the positive side. Using this arbitrarily selected cutoff, the genes can be divided into those that are amenable to the aiQSAR modelling and those that are not. Interestingly, selecting the kappa > 0.105 as a requirement (the opposite of the lowest negative kappa value) gives 243 genes, or just over half of the dataset, that can be well-predicted. The similar selection can be made with the MCC statistic and it yields 239 genes (requirement: MCC > 0.127, opposite of the lowest

negative MCC value). This very crude and tentative analysis suggests that approximately half of the genes can be predicted well, or at least better-than-random using the described method.

Of course, this crude selection is just for exploratory purposes and shouldn't be relied upon for further use. Any subsequent, downstream analyses should calibrate the cutoff parameters to align with the specific requirements of the respective research topics. Additionally, some genes are very well suited for the aiQSAR predictive modelling. The top 10 genes, as ranked by the kappa performance and corresponding nicely to the cutoff requirement of kappa > 0.4, can be robustly predicted and analyzed with high confidence. Detailed statistical performance for these genes is presented here (Table 3.4), with the statistical summary for all 478 modelled genes provided in the Appendix (B.2).

The performance metrics for these top 10 genes indicate robust predictive capacity of the aiQSAR methodology. This is especially the case given the significant class imbalances of the differential gene expression endpoints. For this reason, high accuracy values aren't indicative of good model performance. As is the case in many QSAR applications, other statistical measures need to be assessed (see Section 2.4.2). In this case, confluence between balanced accuracy, kappa and MCC gives high confidence in reliability of results. An interesting observation for the top performing genes is that all of them seem to have relatively small number of active chemicals, even compared to the generally low level of activity in the full dataset. In the next section this peculiarity is explored further, while the biological description of the top 10 best predicted genes for chemically induced differential expression is assessed later (see Section 3.3.5.1).

3.3.4.3 Comparative Dataset Composition Analysis

The best predictive performance is achieved for genes that, on initial inspection, appear to be associated with a very small number of active chemicals. Mathematically, this might be unintuitive since a better class balance could hypothetically be associated with better statistical performance. Biologically however, this shouldn't necessarily be surprising: Genes associated with relatively high number of chemicals that cause differential expression might represent general cellular stress responses that are activated by diverse mechanisms of toxicity. In that case, these genes would essentially act as "canaries" signaling that the cell has experienced some type of stress, regardless of the specific pathway involved. Conversely, genes with differential expression triggered by a relatively low number of chemicals may be more predictable because their regulation would be tightly coupled to particular structural features, thus triggering a specific toxicity pathway.

To test whether there truly is a difference in the number of active chemicals for the genes whose differential expression is the most predictable, a more robust statistical test is required. In the first step, the distribution of the number of active chemicals is compared between the top 10 genes and the remainder of the dataset (Figure 3.23). Visual assessment indicates that there seems to be a difference between these distributions. For the rigorous statistical test, the appropriate null hypothesis is that a typical, median value for the number of active chemicals between the group of top 10 genes and the remainder of the dataset is the same. Notably, the distribution of this number of active chemicals isn't relevant since there is no expectation that this value should follow any type of distribution for either the top 10 genes or for the full dataset. Additionally, the continuous model between the number of active chemicals and the kappa statistic isn't the appropriate statistical measure either: Given a large number of genes for which there clearly isn't any statistically significant predictive capacity (approximately half of the dataset), the relationship between these variables for the full dataset isn't expected to have any statistically significant correlation.

Table 3.4: Confusion matrices and statistical performance metrics for the top 10 best predicted genes, measured by the kappa statistic, for the differential gene expression predictions by the aiQSAR methodology. In confusion matrices, the chemicals are classified as 'Yes' if they cause the differential expression of the corresponding gene and 'No' if they don't.

Gene	Confusion matrix	Accuracy	Balanced Acc.	Kappa	MCC
ZNF395	REF. / PRED. Yes No / Yes 5 1 / No 7 197	0.962	0.706	**0.538**	0.573
LAMP3	REF. / PRED. Yes No / Yes 7 2 / No 11 190	0.938	0.689	**0.489**	0.523
NRG1	REF. / PRED. Yes No / Yes 5 1 / No 9 195	0.952	0.676	**0.479**	0.527
TNFRSF10B	REF. / PRED. Yes No / Yes 5 1 / No 9 195	0.952	0.676	**0.479**	0.527
ITPKC	REF. / PRED. Yes No / Yes 5 2 / No 8 195	0.952	0.687	**0.477**	0.503
NEAT1	REF. / PRED. Yes No / Yes 6 3 / No 9 192	0.943	0.692	**0.472**	0.489
EDEM1	REF. / PRED. Yes No / Yes 5 1 / No 10 194	0.948	0.664	**0.454**	0.507
RPA3	REF. / PRED. Yes No / Yes 4 1 / No 8 197	0.957	0.664	**0.452**	0.5
PLAU	REF. / PRED. Yes No / Yes 6 2 / No 12 190	0.933	0.661	**0.432**	0.472
ANXA1	REF. / PRED. Yes No / Yes 4 0 / No 10 196	0.952	0.643	**0.427**	0.521

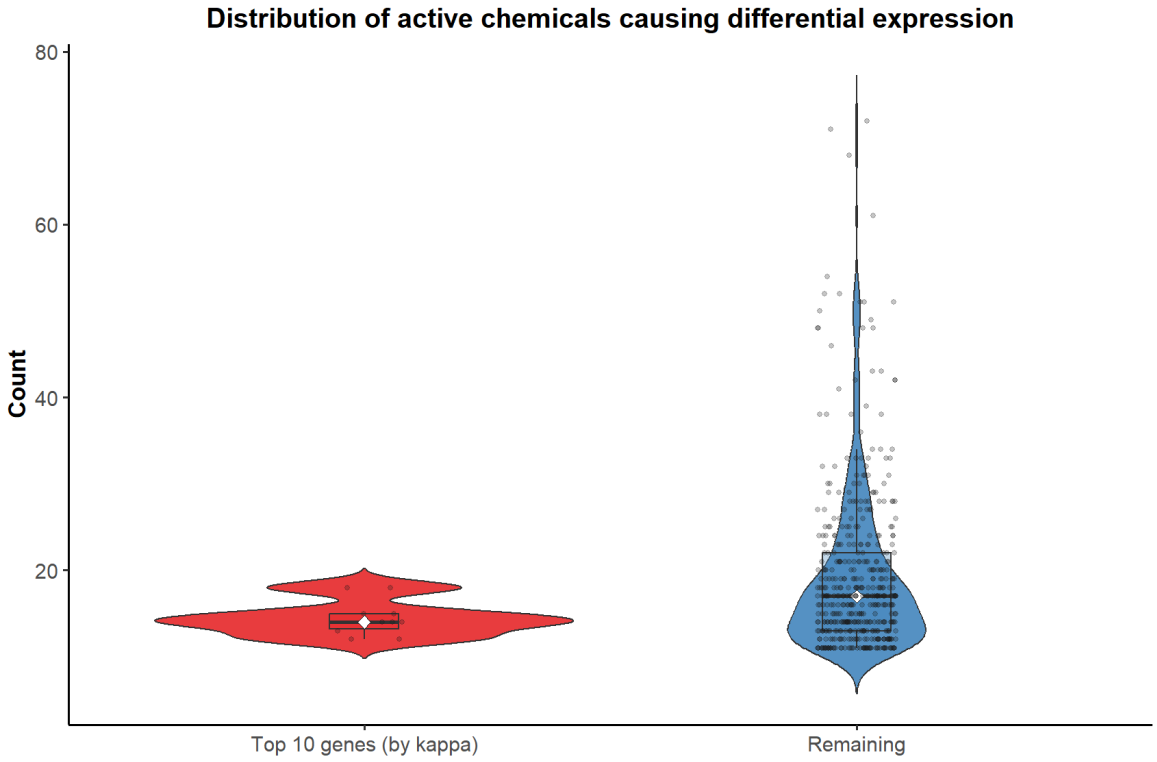


Figure 3.23: Violin plots visually comparing the number of active chemicals that cause the differential gene expression between the top 10 best predicted genes (as measured by the kappa statistic) and the remainder of the dataset.

Wilcoxon rank-sum, or Mann-Whitney-Wilcoxon (MWW) test, a nonparametric statistical test of the null hypothesis that randomly selected values X and Y from two populations have the same distribution, is the most commonly used test for ordinal comparisons. With p(MWW) = 0.112, the result of this test is not significant enough to reject the null hypothesis. It's notable that the two groups have very different spreads, which likely decreases the statistical power of the MWW test. A more mathematically complex statistical test that doesn't assume equal variances and should therefore be more appropriate in this case is Brunner-Munzel (BM) test, a nonparametric test of the null hypothesis that, for randomly selected values X and Y from two populations, the probability of X being greater than Y is equal to the probability of Y being greater than X. With p(BM) = 0.015, the null hypothesis can be rejected and the alternative hypothesis, that the probability of a value indicating the number of active chemicals from the set of top 10 genes is not equal to the value indicating the number of active chemicals from the set of remaining genes, can be accepted. Specifically, given the distribution of median values for the number of active chemicals, for randomly selected values from the top 10 genes set (X) and from the remaining set (Y), probability of X being greater than Y is lower than the probability of Y being greater than X.

The statistical significance is further assessed by two additional evaluations, for top 5 and for top 20 genes (Table 3.5). These additional results convincingly demonstrate that the group of best predicted genes has a lower median value for the number of active chemicals causing the differential expression compared to the group of remaining genes from the TempO-Seq experiment.

Table 3.5: Dataset information and statistical test p-values for comparative analysis of the distribution of active chemicals between the 5, 10 and 20 genes with the most accurate differential expression predictions, as measured by the kappa statistic, and the remaining genes.

#Genes	mean(top)	mean(rest)	median(top)	median(rest)	p(MWW)	p(BM)
5	14.2	19.4	14	17	0.192	0.082
10	14.5	19.4	14	17	0.112	**0.015**
20	15.3	19.5	14	17	**0.034**	**0.019**

3.3.5 Differential Expression Predictability and Biological Implications

The comprehensive evaluation of the aiQSAR methodology applied to differential gene expression predictions requires the assessment of both quantitative predictive performance and the qualitative biological interpretation of results. This section integrates the statistical findings presented with corresponding biological analysis of the genes, thereby establishing the scientific foundations for extending local QSAR methodology beyond traditional chemical endpoints to the domain of transcriptomics. This departure from conventional QSAR applications requires evaluation not only of statistical performance metrics, but also assessment of the mechanistic relevance and biological interpretability of model results. This integrative approach addresses both methodological and domain-specific questions: Does the local QSAR framework adequately capture structure-activity relationships governing differential gene expression? Do genes demonstrating high predictive performance possess common biological characteristics? And can differential expression predictions inform mechanistic understanding of chemical-induced changes in cellular gene regulation?

3.3.5.1 Biological Function of Top 10 Genes

The 10 genes demonstrating highest predictive performance (Table 3.4) originate from different chromosomal locations and encode proteins of diverse molecular classes. The identification of these genes provides an opportunity to examine whether superior predictability correlates with identifiable biological characteristics. Comprehensive biological information for the top 10 genes is presented here (Table 3.6), together with a functional analysis and predictability assessment of each gene [137].

ZNF395 (kappa = 0.538) encodes a zinc finger transcription factor involved in gene-specific transcriptional regulation. Zinc finger proteins are a large family of transcription factors whose DNA-binding specificity and transcriptional activity can be modulated by chemical signaling. This might render their expression particularly sensitive to structure-specific chemical inputs.

LAMP3 (kappa = 0.489) functions as a membrane protein in lysosomal and autophagosomal compartments, with established roles in immune cell maturation and antigen presentation. Activation of autophagy and lysosomal degradation pathways would directly upregulate LAMP3 expression, contributing to specific patterns in differential regulation.

Table 3.6: Biological characteristics of the top 10 genes with highest differential expression predictability.

Gene	Full Name	Protein Type	Primary Function
ZNF395	Zinc Finger Protein 395	Transcription factor	Gene regulation, zinc finger-mediated DNA binding
LAMP3	Lysosomal-Associated Membrane Protein 3	Transmembrane protein	Lysosomal trafficking, immune cell activation
NRG1	Neuregulin 1	Growth factor	ErbB signaling, cell proliferation and differentiation
TNFRSF10B	TNF Receptor Superfamily Member 10B	Death receptor	Extrinsic apoptosis initiation, TRAIL signaling
ITPKC	Inositol Trisphosphate 3-Kinase C	Kinase	IP3 metabolism, intracellular calcium regulation
NEAT1	Nuclear Enriched Abundant Transcript 1	Long non-coding RNA	Paraspeckle assembly, transcriptional regulation
ANXA1	Annexin A1	Calcium-binding protein	Anti-inflammatory signaling, membrane trafficking, exocytosis
EDEM1	ER Degradation Enhancing Alpha-Mannosidase-Like Protein 1	Mannosidase	ER quality control, misfolded protein degradation
RPA3	Replication Protein A3	Single-stranded DNA-binding protein	DNA replication, DNA damage recognition and repair
PLAU	Plasminogen Activator, Urokinase	Serine protease	Fibrinolysis, plasminogen activation, cell migration

NRG1 (kappa = 0.479) functions as an ErbB receptor ligand involved in cell proliferation, differentiation and survival signaling. As a secreted or membrane-bound growth factor, NRG1 expression is tightly regulated by signaling cascades that may be preferentially activated by particular chemical structures.

TNFRSF10B (kappa = 0.479) initiates extrinsic apoptosis upon engagement by TNF-related apoptosis-inducing ligand (TRAIL). The expression of pro-apoptotic receptors is typically held tightly suppressed in normal cells to prevent unwanted apoptosis, but becomes rapidly upregulated in response to specific cellular stress signals. TNFRSF10B expression changes reflect the activation of stress-specific transcription factors, which likely respond predictably to particular chemical structures.

ITPKC (kappa = 0.477) catalyzes an important step in inositol-1,4,5-trisphosphate (IP3) metabolism, thereby modulating intracellular calcium signaling. IP3-mediated calcium mobilization serves as a universal second messenger system responsive to diverse extracellular stimuli, but the components of this pathway are typically expressed at levels proportional to cellular calcium signaling demands. Chemical exposure that alters cellular

calcium homeostasis would contribute to the increased demand for calcium signaling components.

NEAT1 (kappa = 0.472) encodes a long non-coding RNA that nucleates the assembly of paraspeckles, nuclear bodies involved in the sequestration of some mRNA and regulation of gene expression [138].

EDEM1 (kappa = 0.454) functions as a core component of the ER-associated degradation (ERAD) pathway, selectively eliminating misfolded proteins from the endoplasmic reticulum. Chemical exposure that triggers protein misfolding activates the unfolded protein response, leading to upregulation of ERAD components including EDEM1. This activation of proteotoxic pathways could be triggered by specific chemical structures, which would contribute to the robust predictability of EDEM1 expression.

RPA3 (kappa = 0.452) is an important component of the replication protein A complex, which protects single-stranded DNA and coordinates DNA replication and DNA damage recognition. Chemicals that induce DNA damage trigger upregulation of this DNA repair machinery.

PLAU (kappa = 0.432) encodes a serine protease that converts inactive plasminogen into active plasmin, thereby initiating fibrinolysis and cell migration. PLAU expression is tightly controlled by inflammatory mediators and growth factors and its upregulation in response to chemical exposure may reflect activation of specific inflammatory or tissue remodeling pathways preferentially triggered by particular chemical structures.

ANXA1 (kappa = 0.427) functions as a potent endogenous anti-inflammatory mediator and regulator of membrane dynamics and exocytosis. As a marker of cellular stress and inflammation, ANXA1 expression should be tightly coupled to inflammatory signaling cascades. The activation of these cascades, which might be triggered by specific chemical classes, would correspondingly increase the ANXA1 expression.

Interestingly, all of these genes, with the notable exception of NEAT1, encode for proteins rather than regulatory RNAs, suggesting that protein-level responses may be more tightly coupled to chemical structure than regulatory RNA responses.

3.3.5.2 Biological Implications of Predictive Performance

The presented analysis demonstrates the capacity of the aiQSAR methodology to capture genuine structure-activity relationships governing differential gene expression, with approximately 50% of tested genes achieving performance exceeding random expectation.

Additionally, the presence of a clear subset of genes with very strong performance demonstrates that the differential expression for certain genes is particularly tightly coupled to chemical structure, likely reflecting genes whose regulatory mechanisms are sensitive to specific molecular features.

The identified directionality asymmetry, where virtually all genes exhibiting differential expression do so predominantly in a single direction, as opposed to showing bidirectional responses, reflects fundamental constraints on cellular regulatory logic and carries relevant implications for understanding gene regulation in response to chemical exposure.

Another particularly interesting observation, tested with statistical significance, is that genes with the highest predictive performance are those responding to relatively fewer chemical treatments. A tentative biological explanation, that these genes might represent targets of specific molecular mechanisms engaging defined toxicity pathways, in contrast to general stress response genes activated by diverse toxicological mechanisms, needs to be assessed further with a larger sample size.

Future development of this methodology might extend to other transcriptomic contexts. The diversity of chemical effects and the large number of transcriptional programs operating in response to chemical exposure suggest that training datasets encompassing

larger chemical libraries, longer exposure durations and, most importantly, multiple cell types would likely improve the predictive performance and generalizability of differential expression models. Additionally, integration of transcriptomic data with other phenotypic endpoints such as morphological changes, might enable multi-modal QSAR approaches capable of predicting the full spectrum of chemically induced cellular responses.

Finally, the aiQSAR framework for predicting the cellular genetic differential expression can, in its current form, be applied to a wide range of biochemically relevant research topics where the differential expression of one, or many, genes is of scientific interest. Clear example would be the engineering of new, or modifications of existing cell lines, where the intermediate developmental steps often involve looking for differential expressions of relevant genes between sequential stages of cellular maturation.

Chapter 4

Conclusions

This dissertation demonstrated a local QSAR methodology, implemented into an open-source software aiQSAR, as a robust, practical and generalizable framework for predicting chemical properties and biological activities across a broad range of endpoints. Through comprehensive methodological development, software implementation and multi-faceted validation, this work establishes aiQSAR as a viable and accessible alternative to conventional global QSAR approaches, with clear potential for broader adoption in research, industry and within the regulatory context.

The validation of the aiQSAR methodology on acute oral toxicity endpoints, analyzed as part of a community-wide initiative, demonstrates that this approach achieves competitive performance. The observed consistency across five challenges, ranging from regression point estimates of chemical toxicity where aiQSAR achieved performance in the top 15% of submissions, scored by an external panel, to five category toxicity ranking consistent with globally harmonized classification systems where it achieved top performance out of the nineteen submitted models, establishes the methodological robustness of this approach across varying endpoint complexity and classification granularity. Importantly, the demonstrated relationship between Applicability Domain Measure and prediction reliability validates the theoretical underpinning of local QSAR: predictions made for compounds with high structural similarity to local training groups consistently exhibit superior performance compared to predictions for compounds with lower average similarity to their corresponding structural neighbors within the training set. This relationship between chemical similarity and predictive performance fundamentally validates the core hypothesis motivating local QSAR approaches and demonstrates that the additional computational complexity of constructing individualized models is justified by measurable improvements in predictive performance and interpretability.

The successful implementation of the aiQSAR software represents a significant advance in accessibility and practical utility of local QSAR methods. The software package, incorporating 38 distinct endpoints and supporting both routine predictions through a simplified user interface, as well as customizable model development interface, provides the computational infrastructure necessary to move local QSAR from a methodological concept to a practical tool suitable for integration into industrial and regulatory workflows. aiQSAR supports regression, binary classification and multi-class classification endpoints and offers a comprehensive suite of different output formats. Basic output provides a summary of modelling results with the most common performance statistics and readily interpretable visualizations, while a detailed textual output enables downstream computational analysis of local group composition, predictions from individual machine learning methods and more granular performance metrics, both the implemented ones, but also others that can be easily computed from this output format. Finally, a compound-specific output enables the visual identification of the local training compounds associated with each prediction, enabling the software to serve diverse user groups with varying levels of computational expertise and methodological sophistication. The modular software design, supporting straightforward incorporation of new endpoints, positions the software for long-term utility, extensibility and community-driven development, while the quality assurance

infrastructure, including automated applicability domain assessment and quality control flags, ensures that predictions meet rigorous standards appropriate for regulatory and industrial contexts.

The application of the aiQSAR methodology to differential gene expression predictions demonstrates the generalizability of this approach beyond traditional cheminformatic endpoints. With approximately 50% of tested genes achieving adequate predictive performance and a clear subset of genes exhibiting particularly robust performance, this analysis strongly indicates that chemical structure contains substantial information relevant to predicting transcriptomic responses. The exploratory data analysis of the differential expression dataset reveals an interesting and notable pattern that almost no genes demonstrate a significant proportion of chemicals triggering both upregulation and downregulation. This directional asymmetry, where genes exhibit almost exclusively unidirectional responses to chemical exposure, likely reflects fundamental constraints on cellular regulatory logic and reveals mechanistic principles governing gene regulation in response to chemical stress. Another statistically significant finding is the relationship between differential expression predictability and the number of chemical treatments activating each gene. Genes demonstrating the highest predictive performance are those responding to relatively fewer chemical treatments, suggesting that these genes might represent targets of specific molecular mechanisms engaging defined toxicity pathways, as opposed to general stress response genes activated by diverse toxicological triggers.

Despite the demonstrated successes in differential expression prediction, important limitations of this research need to be acknowledged. The differential expression analysis, while demonstrating proof-of-concept, was conducted on a single cell type exposed to a limited set of chemicals. Future extension to larger training datasets encompassing more diverse chemical libraries and, most importantly, multiple different cell types would be required to validate these results. This expansion would also likely improve predictive performance, as well as broaden the applicability of the resulting differential expression models. Additionally, the integration of transcriptomic data with complementary phenotypic endpoints through multi-modal QSAR approaches would be an opportunity to develop more comprehensive predictive frameworks capturing the full spectrum of chemically-induced cellular responses.

In conclusion, this dissertation establishes local QSAR as a mature, validated and practical approach to chemical prediction, demonstrating that the newly developed and local aiQSAR methodology delivers competitive predictive performance while offering advantages in interpretability and mechanistic understanding. The successful implementation of the aiQSAR software provides an accessible, user-friendly platform suitable for research and regulatory use, with extensibility supporting long-term development and community engagement [139]. The successful application to differential gene expression prediction demonstrates the flexibility and generalizability of this methodology, opening pathways for future development and application to emerging challenges in computational toxicology, drug discovery and systems biology. By providing both theoretical foundation and practical implementation, this work aims to facilitate broader adoption of local QSAR methods and to contribute towards the ongoing scientific developments in cheminformatics.

Appendix A

Endpoints Statistics

In this appendix, all performance statistics and the associated metrics for the 38 aiQSAR endpoints are provided. The format of the appendix is the list of endpoints with the data extracted from their accompanying Basic PDF output files.

For regression endpoints, the first page of the output is provided. This page gives the full summary of the endpoint. For multi-class classification endpoints, the first two pages are provided. These give the full summary of the endpoint. For binary classification endpoints, the first two pages are provided, but these give only a part of the summary. Specifically, the distribution of performance metrics with respect to the cutoff for the number of predictions required for a given class is not provided here. Additionally, for some classification endpoints, due to a small number of chemicals in certain ADM categories, only the main statistics are provided, without further subdivision into the respective ADM categories. All remaining information, as well as the full outputs for all endpoints, including the exact predictions for all compounds in the endpoint datasets, can be accessed in the installation directory of the aiQSAR software, as previously discussed in the thesis (see Section 3.2.4.1).

A.1 Acute Tox - EPA

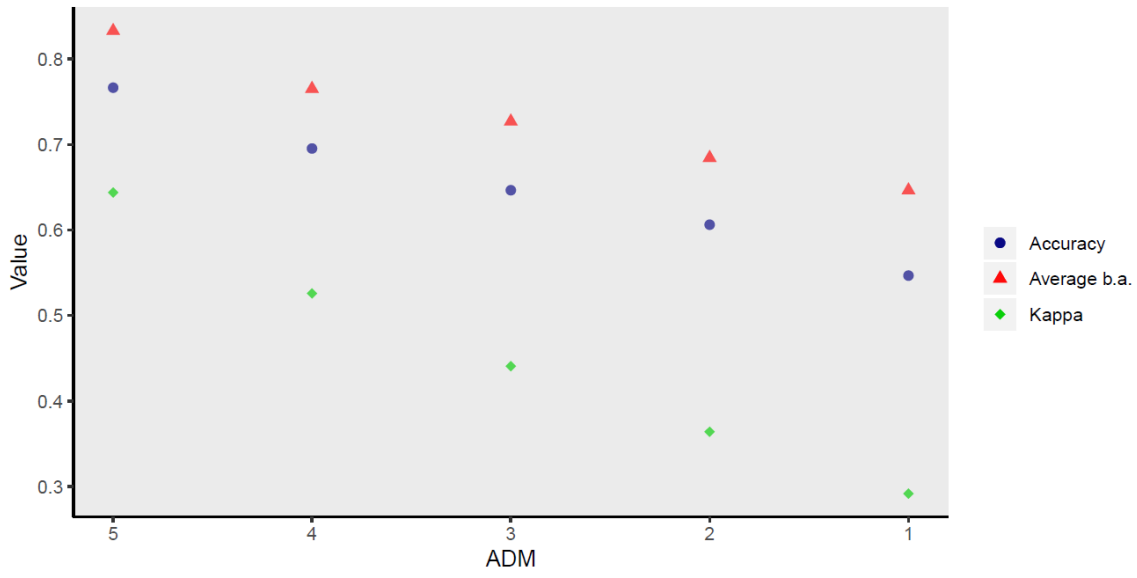
aiQSAR report
2020−09−11 11:15:44
Endpoint: Acute Tox − EPA
Type: multi−class classification
Classes: 1,2,3,4
Training set

Overall statistics

Prediction \ Reference	1	2	3	4
1	532	194	48	15
2	228	1139	598	69
3	122	903	4051	948
4	24	66	592	1023

N(Compounds)	10552
Accuracy	0.639
P-value	3e-179
Average balanced accuracy	0.719
Cohen's Kappa	0.431

Statistics by ADM



ADM	Confusion Matrix	Statistics

5

Reference				
Prediction	1	2	3	4
1	71	11	1	0
2	13	131	27	1
3	1	49	397	70
4	0	2	59	169

N(Compounds)	1002
Accuracy	0.766
P-value	2e-75
Average balanced accuracy	0.833
Cohen's Kappa	0.644

4

Reference				
Prediction	1	2	3	4
1	122	45	5	0
2	44	233	87	9
3	11	132	733	130
4	3	3	87	181

N(Compounds)	1825
Accuracy	0.695
P-value	2e-64
Average balanced accuracy	0.765
Cohen's Kappa	0.526

3

Reference				
Prediction	1	2	3	4
1	163	52	9	2
2	48	266	145	13
3	36	208	1023	244
4	2	15	152	242

N(Compounds)	2620
Accuracy	0.647
P-value	5e-47
Average balanced accuracy	0.727
Cohen's Kappa	0.441

2

Reference				
Prediction	1	2	3	4
1	133	57	28	8
2	74	358	227	31
3	52	337	1443	393
4	15	29	224	337

N(Compounds)	3746
Accuracy	0.606
P-value	1e-30
Average balanced accuracy	0.684
Cohen's Kappa	0.364

1

Reference				
Prediction	1	2	3	4
1	43	29	5	5
2	49	151	112	15
3	22	177	455	111
4	4	17	70	94

N(Compounds)	1359
Accuracy	0.547
P-value	2e-08
Average balanced accuracy	0.646
Cohen's Kappa	0.292

A.2 Acute Tox - GHS

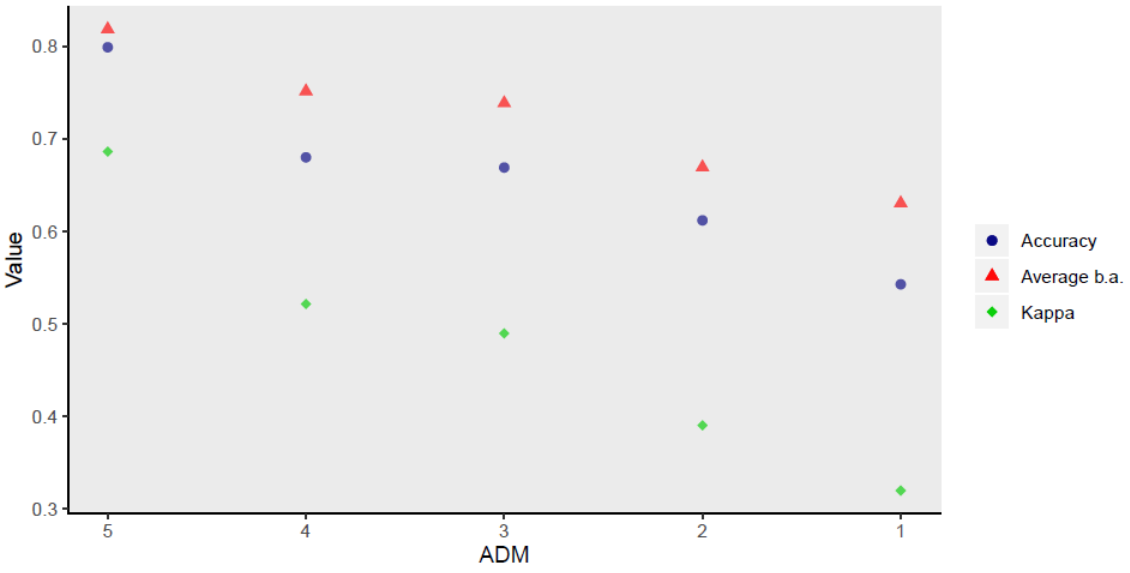
aiQSAR report
2020−09−12 06:40:06
Endpoint: Acute Tox − GHS
Type: multi−class classification
Classes: 1,2,3,4,5
Training set

Overall statistics

Prediction \ Reference	1	2	3	4	5
1	89	46	17	2	6
2	74	298	141	51	25
3	17	152	521	281	79
4	10	116	488	2325	931
5	18	62	189	1026	3607

N(Compounds)	10571
Accuracy	0.647
P-value	0e+00
Average balanced accuracy	0.712
Cohen's Kappa	0.459

Statistics by ADM



ADM	Confusion Matrix	Statistics

5

Prediction \ Reference	1	2	3	4	5
1	9	5	0	0	0
2	11	47	13	2	0
3	0	9	83	18	0
4	0	0	23	236	45
5	0	2	3	77	451

N(Compounds)	1034
Accuracy	0.799
P-value	3e-99
Average balanced accuracy	0.818
Cohen's Kappa	0.686

4

Prediction \ Reference	1	2	3	4	5
1	25	9	1	0	1
2	16	69	34	4	0
3	4	31	103	40	11
4	1	12	82	471	155
5	1	4	19	153	560

N(Compounds)	1806
Accuracy	0.68
P-value	4e-125
Average balanced accuracy	0.751
Cohen's Kappa	0.522

3

Prediction \ Reference	1	2	3	4	5
1	30	11	5	0	0
2	17	97	39	12	3
3	2	30	110	67	14
4	2	38	109	587	230
5	3	7	35	244	930

N(Compounds)	2622
Accuracy	0.669
P-value	5e-114
Average balanced accuracy	0.738
Cohen's Kappa	0.49

2

Prediction \ Reference	1	2	3	4	5
1	22	15	8	2	4
2	24	58	34	22	19
3	9	53	159	104	39
4	4	42	174	756	365
5	9	36	80	410	1296

N(Compounds)	3744
Accuracy	0.612
P-value	2e-77
Average balanced accuracy	0.669
Cohen's Kappa	0.391

1

Prediction \ Reference	1	2	3	4	5
1	3	6	3	0	1
2	6	27	21	11	3
3	2	29	66	52	15
4	3	24	100	275	136
5	5	13	52	142	370

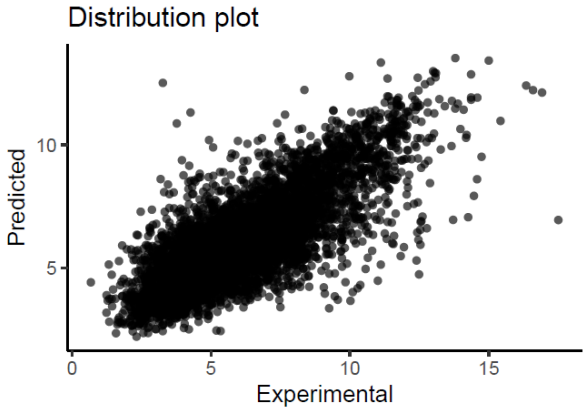
N(Compounds)	1365
Accuracy	0.543
P-value	2e-32
Average balanced accuracy	0.63
Cohen's Kappa	0.32

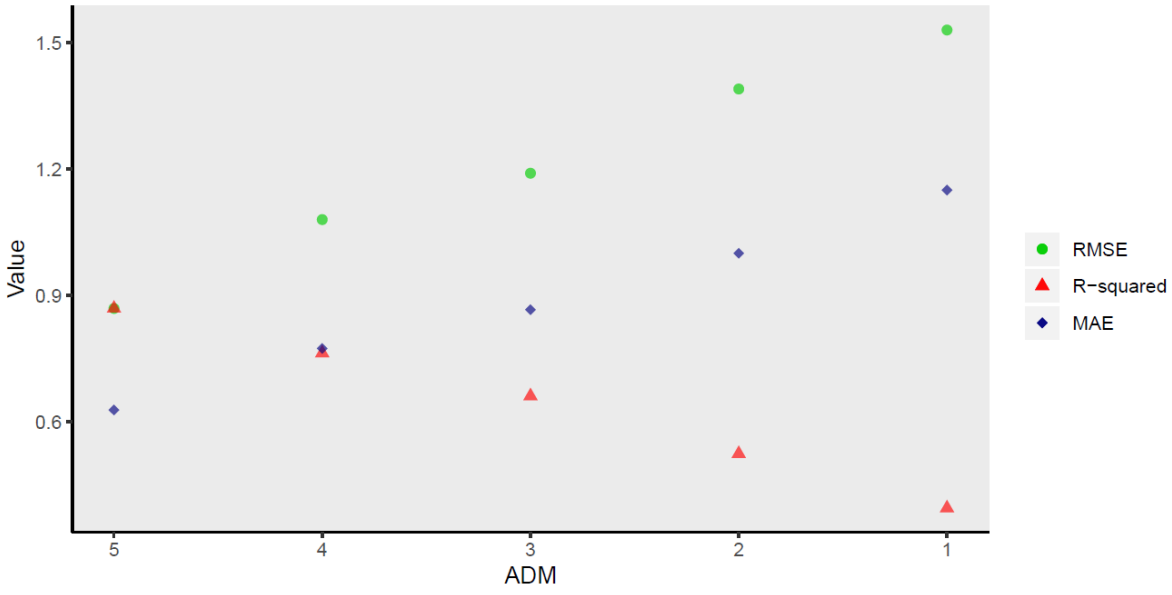
A.3 Acute Tox - LD_{50}

aiQSAR report
2020−09−13 05:02:09
Endpoint: Acute Tox − LD50
Type: regression
Range: 0.671 to 17.506
Training set

Statistics

	N	RMSE	R-squared	MAE
Overall	8314	1.29	0.612	0.931
ADM = 1	1309	1.53	0.395	1.15
ADM = 2	3115	1.39	0.524	1
ADM = 3	1992	1.19	0.661	0.866
ADM = 4	1375	1.08	0.763	0.774
ADM = 5	531	0.869	0.869	0.628





A.4 Acute Tox - NonTox

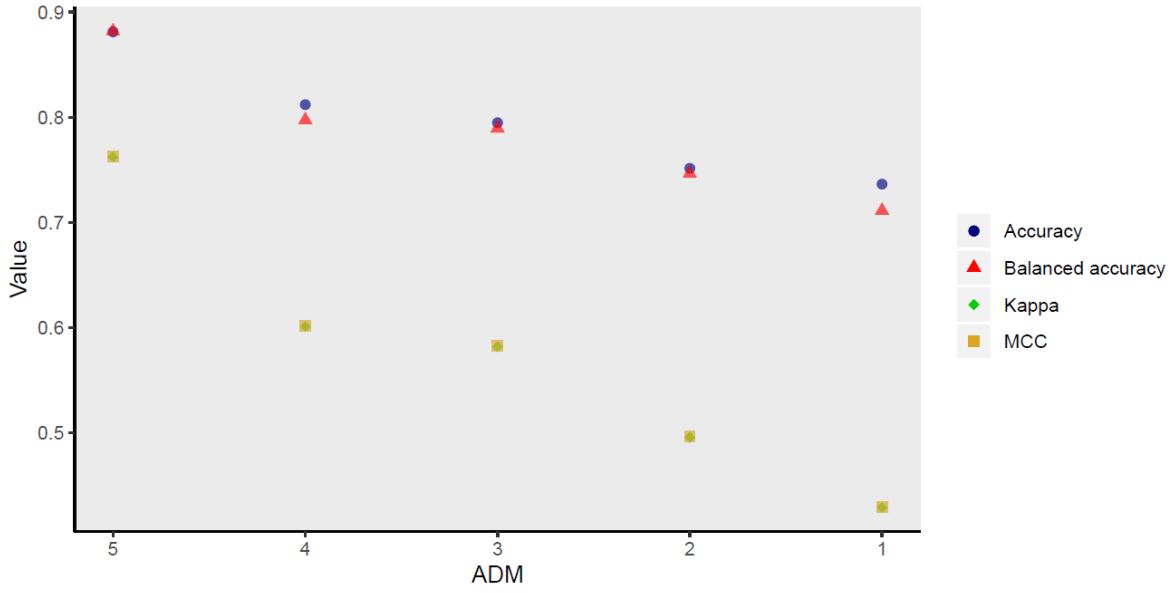
aiQSAR report
2020−09−09 21:51:38
Endpoint: Acute Tox − NonTox
Type: binary classification
Classes: TRUE,FALSE
Training set

Overall statistics

Reference				
Prediction		**FALSE**	**TRUE**	
	FALSE	5186	1311	
	TRUE	1076	3434	

N(Compounds)	11007
Accuracy	0.783
P-value	0e+00
Balanced accuracy	0.776
Cohen's Kappa	0.555
MCC	0.556

Statistics by ADM



ADM	Confusion Matrix	Statistics
5	Reference Prediction FALSE TRUE FALSE 488 53 TRUE 73 447	N(Compounds) 1061 Accuracy 0.881 P-value 4e-134 Balanced accuracy 0.882 Cohen's Kappa 0.762 MCC 0.763
4	Reference Prediction FALSE TRUE FALSE 998 199 TRUE 157 540	N(Compounds) 1894 Accuracy 0.812 P-value 2e-80 Balanced accuracy 0.797 Cohen's Kappa 0.601 MCC 0.602
3	Reference Prediction FALSE TRUE FALSE 1266 308 TRUE 249 892	N(Compounds) 2715 Accuracy 0.795 P-value 4e-148 Balanced accuracy 0.789 Cohen's Kappa 0.582 MCC 0.583
2	Reference Prediction FALSE TRUE FALSE 1702 541 TRUE 427 1225	N(Compounds) 3895 Accuracy 0.751 P-value 3e-154 Balanced accuracy 0.747 Cohen's Kappa 0.496 MCC 0.497
1	Reference Prediction FALSE TRUE FALSE 732 210 TRUE 170 330	N(Compounds) 1442 Accuracy 0.736 P-value 3e-19 Balanced accuracy 0.711 Cohen's Kappa 0.429 MCC 0.43

A.5 Acute Tox - VeryTox

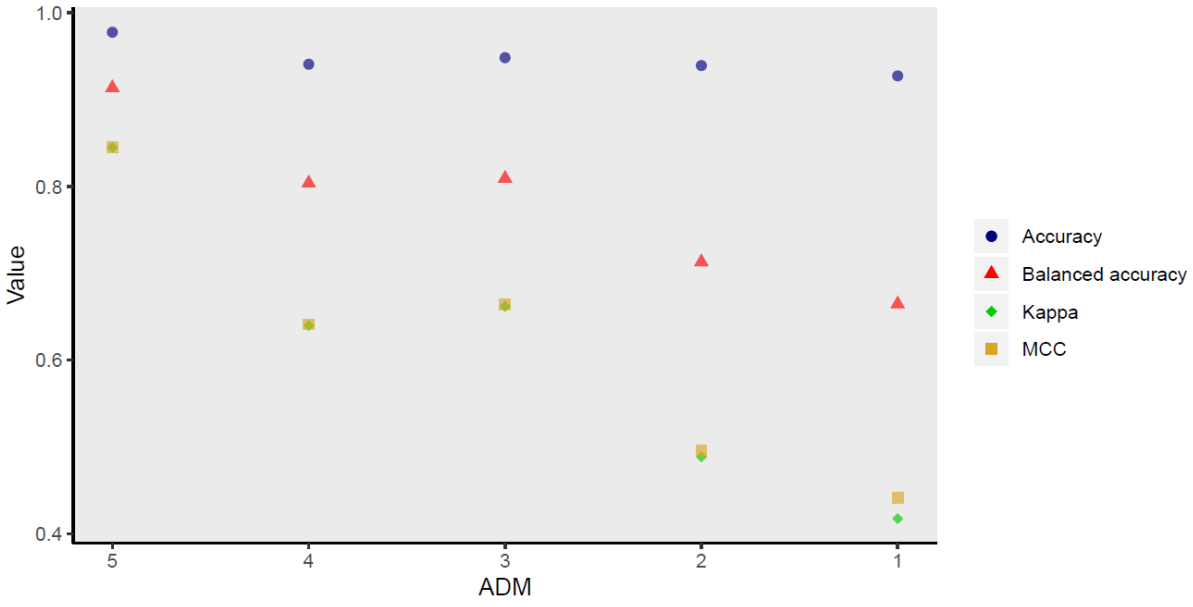
aiQSAR report
2020−09−10 16:16:39
Endpoint: Acute Tox − VeryTox
Type: binary classification
Classes: TRUE,FALSE
Training set

Overall statistics

Prediction \ Reference	FALSE	TRUE
FALSE	9951	411
TRUE	212	520

N(Compounds)	11094
Accuracy	0.944
P-value	4e-29
Balanced accuracy	0.769
Cohen's Kappa	0.595
MCC	0.601

Statistics by ADM



ADM	Confusion Matrix			Statistics	
5				N(Compounds)	1068
		Reference		Accuracy	0.978
	Prediction	**FALSE**	**TRUE**	P-value	4e-16
	FALSE	972	14	Balanced accuracy	0.914
	TRUE	10	72	Cohen's Kappa	0.845
				MCC	0.845
4				N(Compounds)	1939
		Reference		Accuracy	0.941
	Prediction	**FALSE**	**TRUE**	P-value	3e-09
	FALSE	1706	68	Balanced accuracy	0.804
	TRUE	47	118	Cohen's Kappa	0.64
				MCC	0.641
3				N(Compounds)	2742
		Reference		Accuracy	0.948
	Prediction	**FALSE**	**TRUE**	P-value	3e-14
	FALSE	2442	89	Balanced accuracy	0.809
	TRUE	53	158	Cohen's Kappa	0.662
				MCC	0.664
2				N(Compounds)	3915
		Reference		Accuracy	0.939
	Prediction	**FALSE**	**TRUE**	P-value	9e-04
	FALSE	3548	159	Balanced accuracy	0.713
	TRUE	79	129	Cohen's Kappa	0.489
				MCC	0.496
1				N(Compounds)	1430
		Reference		Accuracy	0.927
	Prediction	**FALSE**	**TRUE**	P-value	3e-02
	FALSE	1283	81	Balanced accuracy	0.665
	TRUE	23	43	Cohen's Kappa	0.418
				MCC	0.441

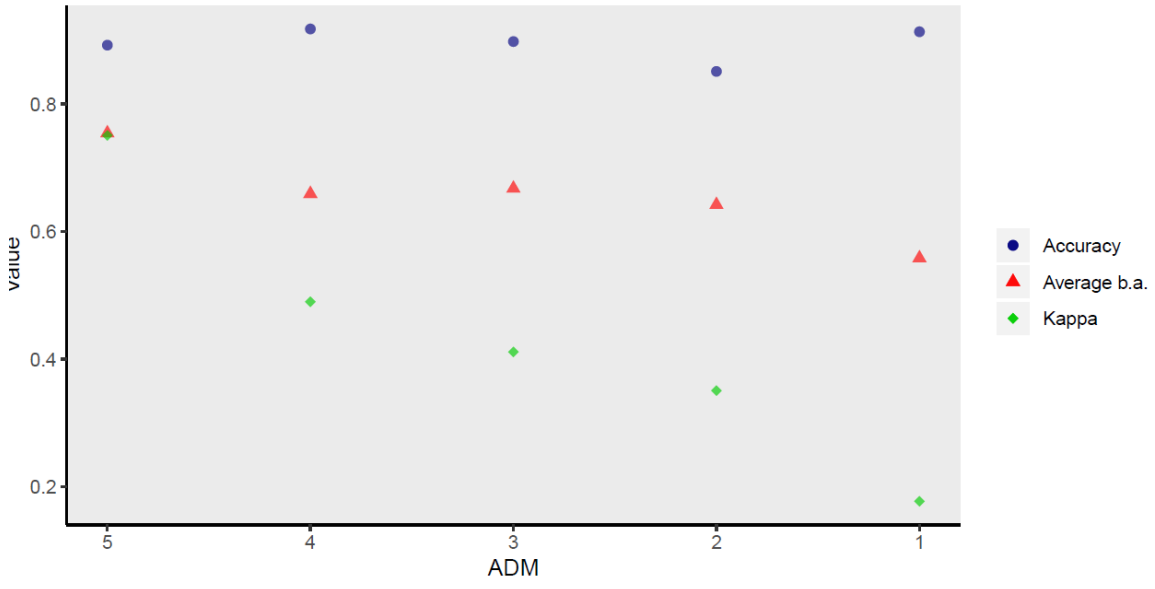
A.6 IMN - aromatase

aiQSAR report
2020−09−21 21:56:41
Endpoint: IMN − aromatase
Type: multi−class classification
Classes: agonist,antagonist,inactive
Training set

Overall statistics

Reference					
Prediction	**agonist**	**antagonist**	**inactive**	N(Compounds)	2804
agonist	110	7	55	Accuracy	0.884
antagonist	3	37	21	P-value	4e−05
inactive	147	92	2332	Average balanced accuracy	0.671
				Cohen's Kappa	0.437

Statistics by ADM



ADM	Confusion Matrix				Statistics	
5	Reference				N(Compounds)	195
	Prediction	agonist	antagonist	inactive	Accuracy	0.892
	agonist	49	1	8	P-value	1e-11
	antagonist	0	0	1	Average balanced accuracy	0.754
	inactive	9	2	125	Cohen's Kappa	0.751
4	Reference				N(Compounds)	328
	Prediction	agonist	antagonist	inactive	Accuracy	0.918
	agonist	14	2	4	P-value	1e-01
	antagonist	0	0	3	Average balanced accuracy	0.659
	inactive	12	6	287	Cohen's Kappa	0.49
3	Reference				N(Compounds)	539
	Prediction	agonist	antagonist	inactive	Accuracy	0.898
	agonist	12	1	10	P-value	1e-01
	antagonist	2	9	1	Average balanced accuracy	0.668
	inactive	29	12	463	Cohen's Kappa	0.411
2	Reference				N(Compounds)	1142
	Prediction	agonist	antagonist	inactive	Accuracy	0.851
	agonist	33	2	29	P-value	1e-01
	antagonist	1	24	13	Average balanced accuracy	0.642
	inactive	72	53	915	Cohen's Kappa	0.351
1	Reference				N(Compounds)	600
	Prediction	agonist	antagonist	inactive	Accuracy	0.913
	agonist	2	1	4	P-value	6e-01
	antagonist	0	4	3	Average balanced accuracy	0.559
	inactive	25	19	542	Cohen's Kappa	0.177

A.7 IMN - DILI



aiQSAR report
2020−09−21 06:40:32
Endpoint: IMN − DILI
Type: binary classification
Classes: 0,1
Training set

Overall statistics

Reference		
Prediction	0	1
0	861	402
1	376	839

N(Compounds)	2478
Accuracy	0.686
P-value	7e-78
Balanced accuracy	0.686
Cohen's Kappa	0.372
MCC	0.372

A.8 IMN - HIA



aiQSAR report
2020−09−19 15:41:03
Endpoint: IMN − HIA
Type: binary classification
Classes: no,yes
Training set

Overall statistics

Reference		
Prediction	no	yes
no	67	34
yes	81	1119

N(Compounds)	1301
Accuracy	0.912
P-value	2e-03
Balanced accuracy	0.712
Cohen's Kappa	0.491
MCC	0.502

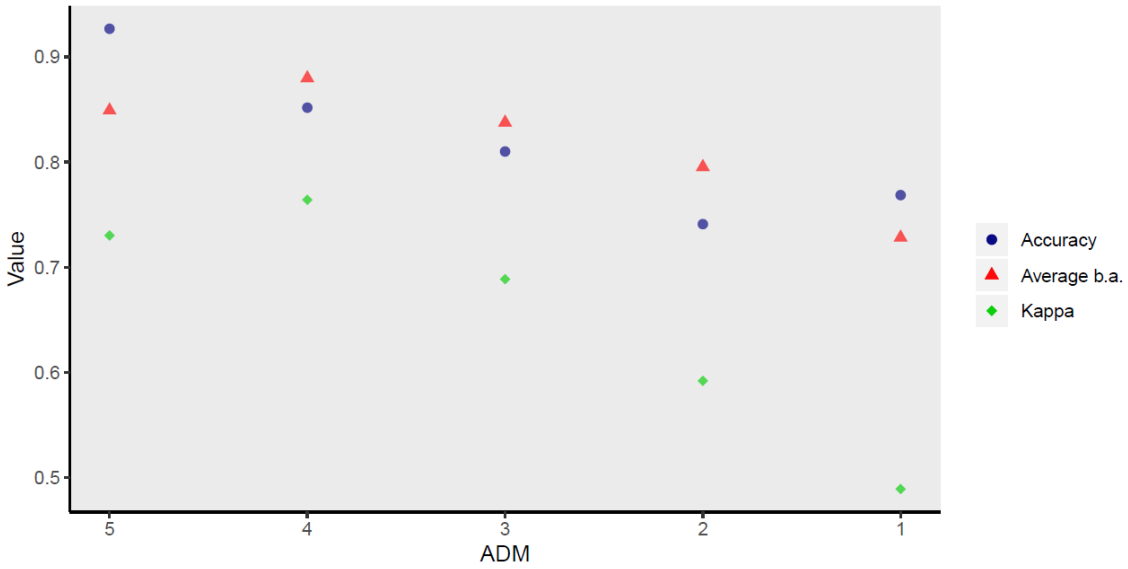
A.9 IMN - Pgp

aiQSAR report
2020−09−18 04:39:11
Endpoint: IMN − Pgp
Type: multi−class classification
Classes: inactive,inhibitor,substrate
Training set

Overall statistics

Reference					
Prediction	**inactive**	**inhibitor**	**substrate**		
inactive	650	102	74	N(Compounds)	2414
inhibitor	126	1000	53	Accuracy	0.814
substrate	58	35	316	P-value	5e-266
				Average balanced accuracy	0.846
				Cohen's Kappa	0.701

Statistics by ADM



ADM	Confusion Matrix				Statistics	
5	Reference				N(Compounds)	519
	Prediction	inactive	inhibitor	substrate	Accuracy	0.927
	inactive	35	10	5	P-value	1e-09
	inhibitor	16	425	2	Average balanced accuracy	0.849
	substrate	5	0	21	Cohen's Kappa	0.73
4	Reference				N(Compounds)	344
	Prediction	inactive	inhibitor	substrate	Accuracy	0.852
	inactive	56	17	6	P-value	3e-45
	inhibitor	16	147	5	Average balanced accuracy	0.88
	substrate	3	4	90	Cohen's Kappa	0.764
3	Reference				N(Compounds)	516
	Prediction	inactive	inhibitor	substrate	Accuracy	0.81
	inactive	120	18	17	P-value	3e-47
	inhibitor	26	232	12	Average balanced accuracy	0.838
	substrate	15	10	66	Cohen's Kappa	0.689
2	Reference				N(Compounds)	780
	Prediction	inactive	inhibitor	substrate	Accuracy	0.741
	inactive	284	42	30	P-value	8e-54
	inhibitor	59	182	30	Average balanced accuracy	0.795
	substrate	23	18	112	Cohen's Kappa	0.592
1	Reference				N(Compounds)	255
	Prediction	inactive	inhibitor	substrate	Accuracy	0.769
	inactive	155	15	16	P-value	3e-03
	inhibitor	9	14	4	Average balanced accuracy	0.728
	substrate	12	3	27	Cohen's Kappa	0.489

A.10 PubChem AID 941 - TLR4



aiQSAR report
2020−09−01 12:51:43
Endpoint: PubChem AID 941 − TLR4
Type: binary classification
Classes: Inactive,Active
Training set

Overall statistics

Reference		
Prediction	**Active**	**Inactive**
Active	121	73
Inactive	54	100

N(Compounds)	348
Accuracy	0.635
P-value	5e−07
Balanced accuracy	0.635
Cohen's Kappa	0.27
MCC	0.271

A.11 PubChem AID 624156 - ROS



aiQSAR report
2020−09−04 05:03:38
Endpoint: PubChem AID 624156 − ROS
Type: binary classification
Classes: Active,Inactive
Training set

Overall statistics

Reference		
Prediction	**Active**	**Inactive**
Active	177	87
Inactive	321	8436

N(Compounds)	9021
Accuracy	0.955
P-value	1e−05
Balanced accuracy	0.673
Cohen's Kappa	0.443
MCC	0.468

A.12 PubChem AID 1159518 - NFkB

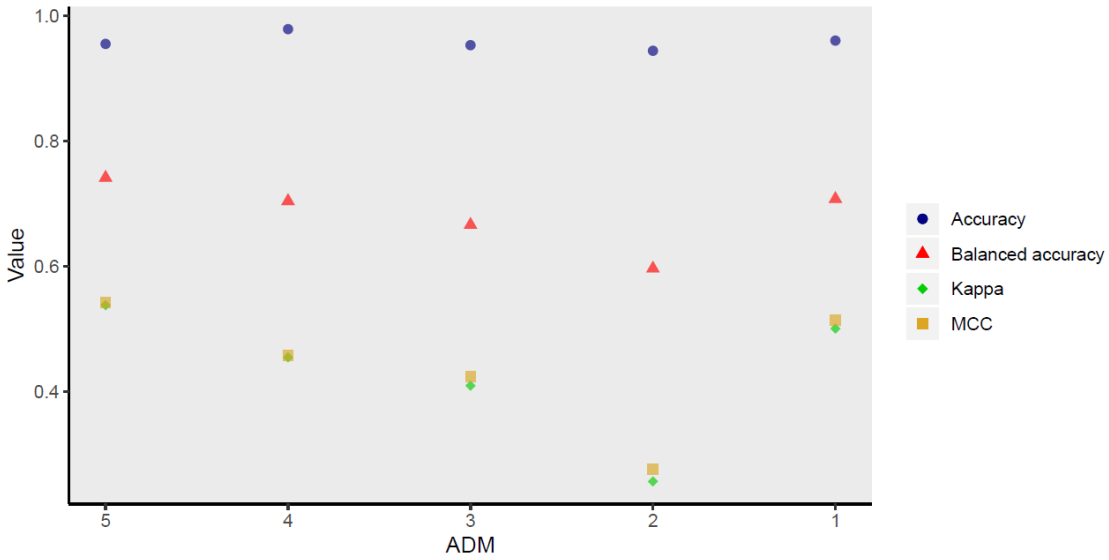
aiQSAR report
2020−09−06 10:06:33
Endpoint: PubChem AID 1159518 − NFkB
Type: binary classification
Classes: yes,no
Training set

Overall statistics

	Reference	
Prediction	**no**	**yes**
no	6752	230
yes	91	114

N(Compounds)	7187
Accuracy	0.955
P-value	1e-01
Balanced accuracy	0.659
Cohen's Kappa	0.394
MCC	0.408

Statistics by ADM



ADM	Confusion Matrix			Statistics	
				N(Compounds)	804
		Reference		Accuracy	0.955
5	Prediction	no	yes	P-value	7e-02
	no	745	23	Balanced accuracy	0.741
	yes	13	23	Cohen's Kappa	0.538
				MCC	0.542
				N(Compounds)	1085
		Reference		Accuracy	0.979
4	Prediction	no	yes	P-value	5e-01
	no	1052	14	Balanced accuracy	0.704
	yes	9	10	Cohen's Kappa	0.454
				MCC	0.458
				N(Compounds)	1517
		Reference		Accuracy	0.953
3	Prediction	no	yes	P-value	2e-01
	no	1419	51	Balanced accuracy	0.666
	yes	20	27	Cohen's Kappa	0.409
				MCC	0.423
				N(Compounds)	2565
		Reference		Accuracy	0.944
2	Prediction	no	yes	P-value	8e-01
	no	2394	107	Balanced accuracy	0.596
	yes	36	28	Cohen's Kappa	0.256
				MCC	0.276
				N(Compounds)	1216
		Reference		Accuracy	0.961
1	Prediction	no	yes	P-value	5e-02
	no	1142	35	Balanced accuracy	0.707
	yes	13	26	Cohen's Kappa	0.5
				MCC	0.514

A.13 Tox21 - AhR

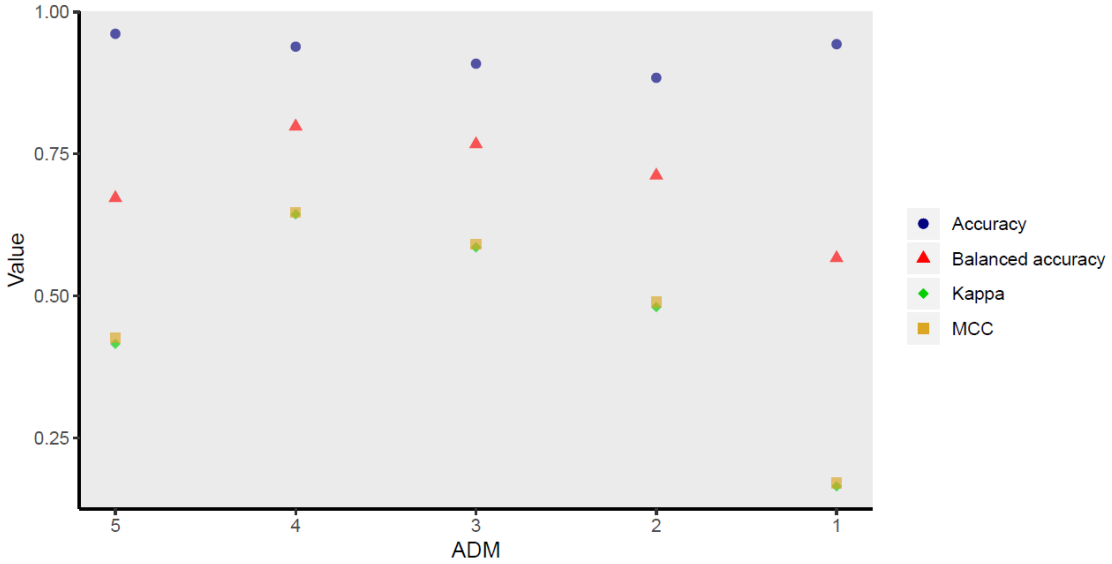
aiQSAR report
2020−07−10 01:30:58
Endpoint: Tox21 − AhR
Type: binary classification
Classes: 1,0
Training set

Overall statistics

	Reference	
Prediction	0	1
0	5892	391
1	192	375

N(Compounds)	6850
Accuracy	0.915
P-value	2e-13
Balanced accuracy	0.729
Cohen's Kappa	0.517
MCC	0.524

Statistics by ADM



ADM	Confusion Matrix	Statistics
5	Reference Prediction 0 1 0 637 18 1 8 10	N(Compounds) 673 Accuracy 0.961 P-value 4e-01 Balanced accuracy 0.672 Cohen's Kappa 0.416 MCC 0.427
4	Reference Prediction 0 1 0 871 39 1 22 64	N(Compounds) 996 Accuracy 0.939 P-value 2e-06 Balanced accuracy 0.798 Cohen's Kappa 0.644 MCC 0.647
3	Reference Prediction 0 1 0 1181 86 1 44 114	N(Compounds) 1425 Accuracy 0.909 P-value 1e-08 Balanced accuracy 0.767 Cohen's Kappa 0.586 MCC 0.591
2	Reference Prediction 0 1 0 2099 204 1 95 179	N(Compounds) 2577 Accuracy 0.884 P-value 9e-07 Balanced accuracy 0.712 Cohen's Kappa 0.481 MCC 0.489
1	Reference Prediction 0 1 0 1104 44 1 23 8	N(Compounds) 1179 Accuracy 0.943 P-value 1e+00 Balanced accuracy 0.567 Cohen's Kappa 0.165 MCC 0.171

A.14 Tox21 - AR

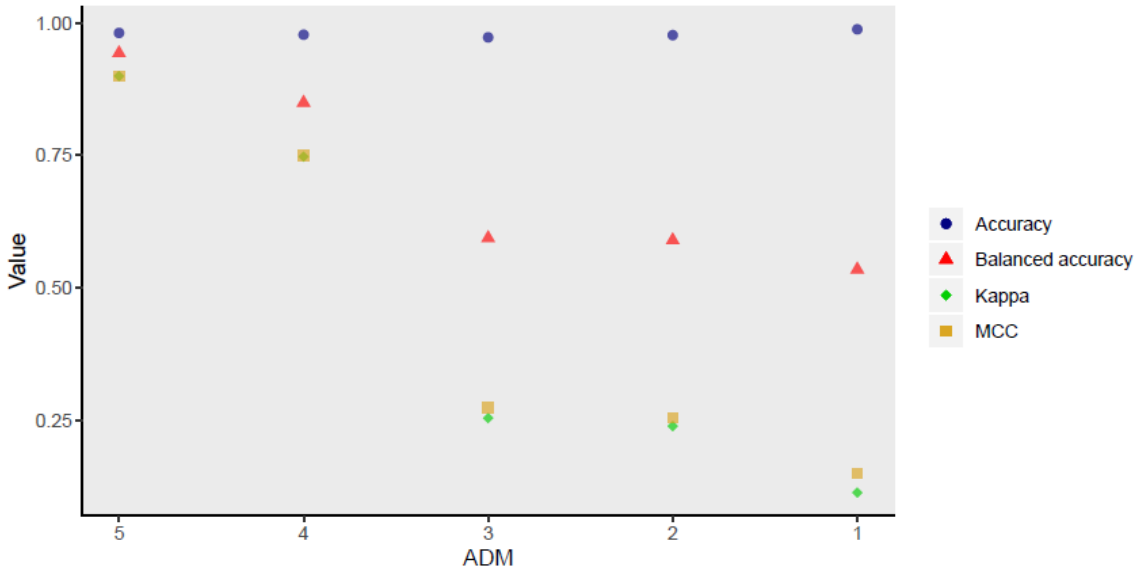
aiQSAR report
2020-07-10 13:18:28
Endpoint: Tox21 - AR
Type: binary classification
Classes: 1,0
Training set

Overall statistics

	N(Compounds)	7602
	Accuracy	0.978
	P-value	5e-10
	Balanced accuracy	0.763
	Cohen's Kappa	0.612
	MCC	0.621

Reference		
Prediction	0	1
0	7299	120
1	46	137

Statistics by ADM



ADM	Confusion Matrix			Statistics	
5		Reference		N(Compounds)	786
	Prediction	0	1	Accuracy	0.981
	0	695	9	P-value	3e-22
	1	6	76	Balanced accuracy	0.943
				Cohen's Kappa	0.9
				MCC	0.9
4		Reference		N(Compounds)	1158
	Prediction	0	1	Accuracy	0.978
	0	1091	17	P-value	1e-06
	1	9	41	Balanced accuracy	0.849
				Cohen's Kappa	0.748
				MCC	0.75
3		Reference		N(Compounds)	1601
	Prediction	0	1	Accuracy	0.973
	0	1549	33	P-value	7e-01
	1	11	8	Balanced accuracy	0.594
				Cohen's Kappa	0.255
				MCC	0.274
2		Reference		N(Compounds)	2833
	Prediction	0	1	Accuracy	0.977
	0	2756	48	P-value	8e-01
	1	18	11	Balanced accuracy	0.59
				Cohen's Kappa	0.24
				MCC	0.255
1		Reference		N(Compounds)	1224
	Prediction	0	1	Accuracy	0.988
	0	1208	13	P-value	7e-01
	1	2	1	Balanced accuracy	0.535
				Cohen's Kappa	0.114
				MCC	0.15

A.15 Tox21 - ARE

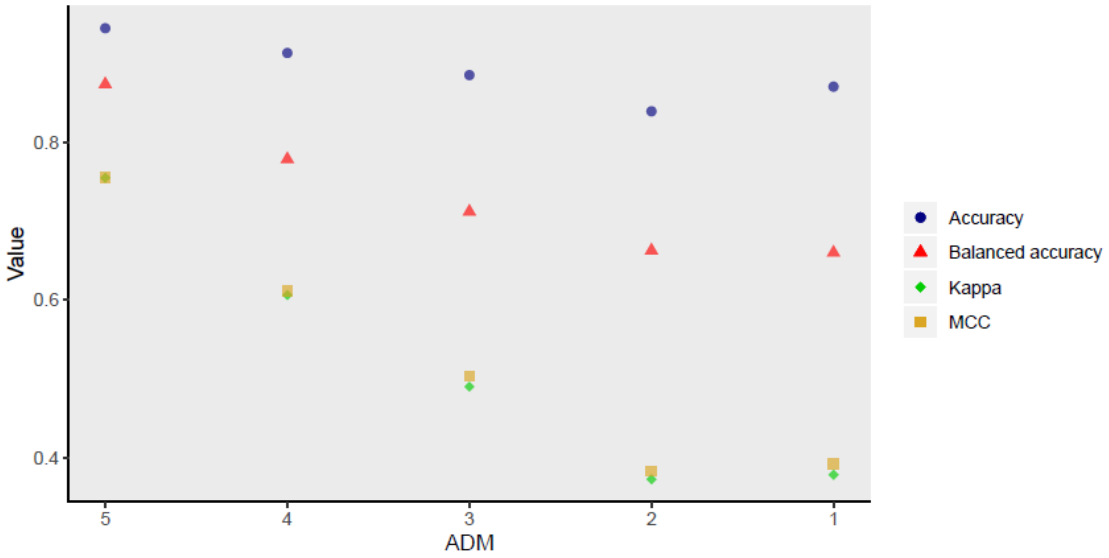
aiQSAR report
2020-07-11 23:07:00
Endpoint: Tox21 - ARE
Type: binary classification
Classes: 1,0
Training set

Overall statistics

	Reference	
Prediction	0	1
0	4878	523
1	239	438

N(Compounds)	6078
Accuracy	0.875
P-value	3e-13
Balanced accuracy	0.705
Cohen's Kappa	0.465
MCC	0.474

Statistics by ADM



ADM	Confusion Matrix	Statistics
5	Reference Prediction 0 1 0 480 17 1 15 59	N(Compounds) 571 Accuracy 0.944 P-value 1e-09 Balanced accuracy 0.873 Cohen's Kappa 0.754 MCC 0.755
4	Reference Prediction 0 1 0 778 54 1 28 78	N(Compounds) 938 Accuracy 0.913 P-value 4e-07 Balanced accuracy 0.778 Cohen's Kappa 0.606 MCC 0.611
3	Reference Prediction 0 1 0 944 97 1 37 83	N(Compounds) 1161 Accuracy 0.885 P-value 7e-05 Balanced accuracy 0.712 Cohen's Kappa 0.49 MCC 0.503
2	Reference Prediction 0 1 0 1746 250 1 116 158	N(Compounds) 2270 Accuracy 0.839 P-value 1e-02 Balanced accuracy 0.662 Cohen's Kappa 0.373 MCC 0.383
1	Reference Prediction 0 1 0 930 105 1 43 60	N(Compounds) 1138 Accuracy 0.87 P-value 8e-02 Balanced accuracy 0.66 Cohen's Kappa 0.378 MCC 0.392

A.16 Tox21 - ARLBD

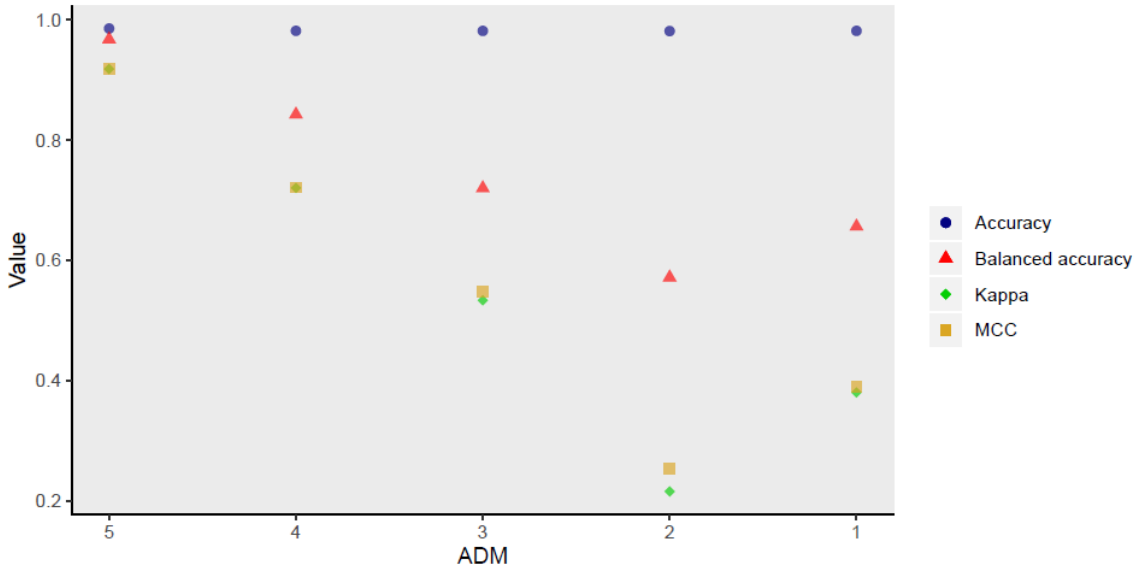
aiQSAR report
2020−07−10 23:50:56
Endpoint: Tox21 − ARLBD
Type: binary classification
Classes: 1,0
Training set

Overall statistics

	Reference	
Prediction	0	1
0	6803	92
1	37	126

N(Compounds)	7058
Accuracy	0.982
P-value	3e−11
Balanced accuracy	0.786
Cohen's Kappa	0.652
MCC	0.66

Statistics by ADM



ADM	Confusion Matrix	Statistics
5	Reference Prediction 0 1 0 671 4 1 7 69	N(Compounds) 751 Accuracy 0.985 P-value 1e-20 Balanced accuracy 0.967 Cohen's Kappa 0.918 MCC 0.918
4	Reference Prediction 0 1 0 1034 12 1 8 27	N(Compounds) 1081 Accuracy 0.981 P-value 5e-04 Balanced accuracy 0.842 Cohen's Kappa 0.72 MCC 0.721
3	Reference Prediction 0 1 0 1408 20 1 7 16	N(Compounds) 1451 Accuracy 0.981 P-value 7e-02 Balanced accuracy 0.72 Cohen's Kappa 0.533 MCC 0.547
2	Reference Prediction 0 1 0 2533 41 1 8 7	N(Compounds) 2589 Accuracy 0.981 P-value 6e-01 Balanced accuracy 0.571 Cohen's Kappa 0.215 MCC 0.254
1	Reference Prediction 0 1 0 1157 15 1 7 7	N(Compounds) 1186 Accuracy 0.981 P-value 6e-01 Balanced accuracy 0.656 Cohen's Kappa 0.38 MCC 0.39

A.17 Tox21 - aromatase

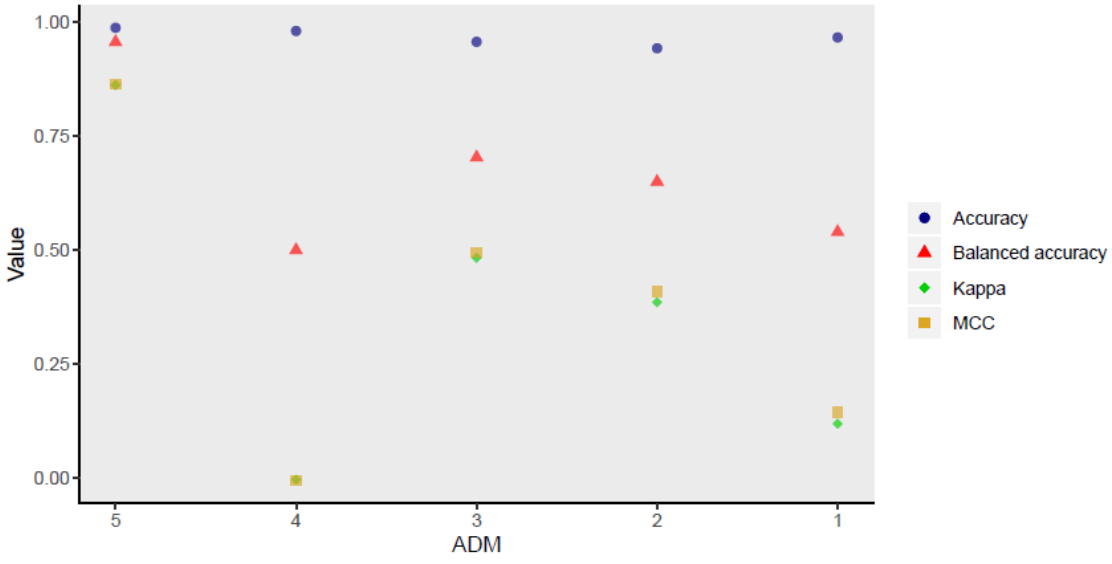
aiQSAR report
2020-07-12 09:59:19
Endpoint: Tox21 - aromatase
Type: binary classification
Classes: 1,0
Training set

Overall statistics

Reference			
Prediction		0	1
	0	5710	196
	1	61	102

N(Compounds)	6069
Accuracy	0.958
P-value	7e-03
Balanced accuracy	0.666
Cohen's Kappa	0.422
MCC	0.443

Statistics by ADM



ADM	Confusion Matrix			Statistics	
				N(Compounds)	503
		Reference		Accuracy	0.986
5	Prediction	0	1	P-value	2e-05
	0	473	2	Balanced accuracy	0.955
	1	5	23	Cohen's Kappa	0.861
				MCC	0.862
				N(Compounds)	864
		Reference		Accuracy	0.979
4	Prediction	0	1	P-value	7e-01
	0	846	16	Balanced accuracy	0.499
	1	2	0	Cohen's Kappa	-0.00413
				MCC	-0.00662
				N(Compounds)	1232
		Reference		Accuracy	0.955
3	Prediction	0	1	P-value	7e-02
	0	1149	39	Balanced accuracy	0.702
	1	16	28	Cohen's Kappa	0.482
				MCC	0.494
				N(Compounds)	2331
		Reference		Accuracy	0.941
2	Prediction	0	1	P-value	8e-02
	0	2146	106	Balanced accuracy	0.649
	1	31	48	Cohen's Kappa	0.384
				MCC	0.408
				N(Compounds)	1139
		Reference		Accuracy	0.965
1	Prediction	0	1	P-value	8e-01
	0	1096	33	Balanced accuracy	0.538
	1	7	3	Cohen's Kappa	0.118
				MCC	0.144

A.18 Tox21 - ATAD5

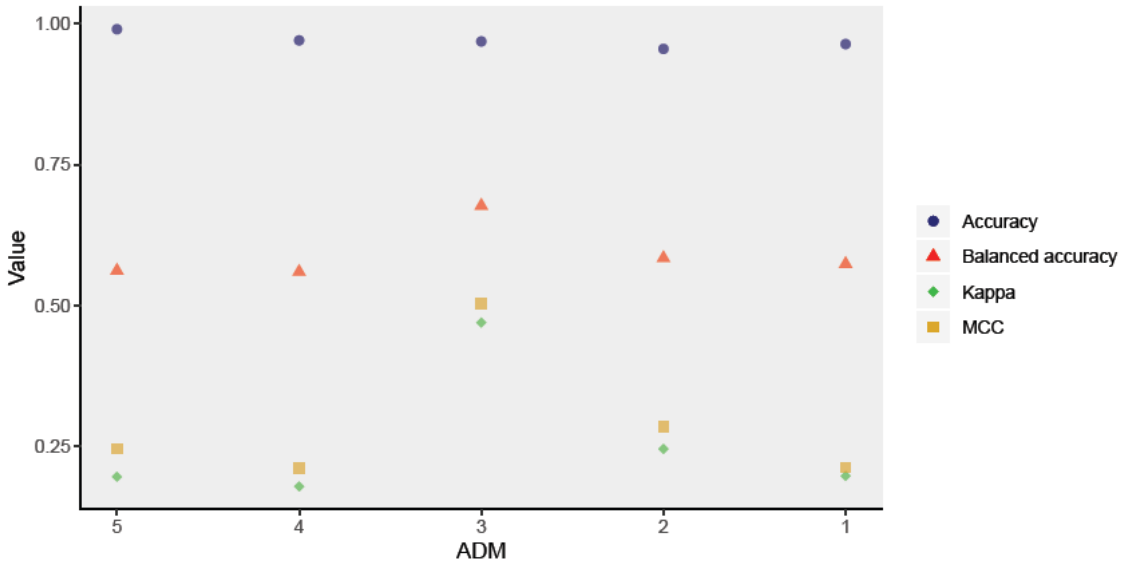
aiQSAR report
2020-07-14 06:37:45
Endpoint: Tox21 – ATAD5
Type: binary classification
Classes: 1,0
Training set

Overall statistics

Reference		
Prediction	0	1
0	7082	211
1	47	56

N(Compounds)	7396
Accuracy	0.965
P-value	3e-01
Balanced accuracy	0.602
Cohen's Kappa	0.288
MCC	0.323

Statistics by ADM



ADM	Confusion Matrix	Statistics
5	Reference Prediction 0 1 0 773 7 1 1 1	N(Compounds) 782 Accuracy 0.99 P-value 6e-01 Balanced accuracy 0.562 Cohen's Kappa 0.197 MCC 0.246
4	Reference Prediction 0 1 0 1098 28 1 6 4	N(Compounds) 1136 Accuracy 0.97 P-value 7e-01 Balanced accuracy 0.56 Cohen's Kappa 0.179 MCC 0.212
3	Reference Prediction 0 1 0 1470 41 1 8 23	N(Compounds) 1542 Accuracy 0.968 P-value 3e-02 Balanced accuracy 0.677 Cohen's Kappa 0.47 MCC 0.503
2	Reference Prediction 0 1 0 2585 103 1 20 22	N(Compounds) 2730 Accuracy 0.955 P-value 5e-01 Balanced accuracy 0.584 Cohen's Kappa 0.246 MCC 0.286
1	Reference Prediction 0 1 0 1156 32 1 12 6	N(Compounds) 1206 Accuracy 0.964 P-value 9e-01 Balanced accuracy 0.574 Cohen's Kappa 0.198 MCC 0.213

A.19 Tox21 - ER

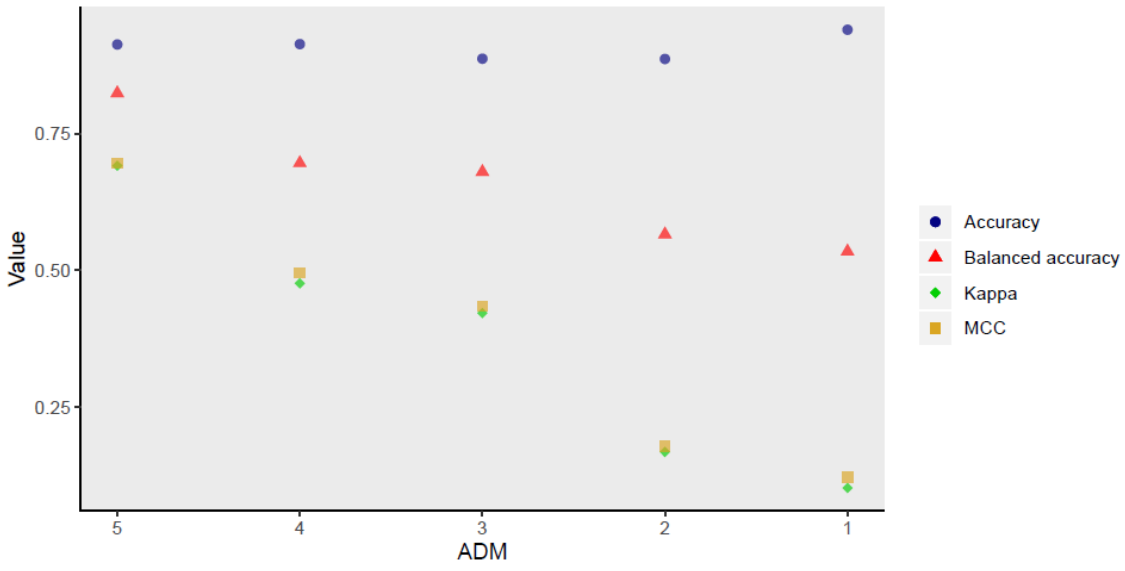
aiQSAR report
2020−07−18 04:21:42
Endpoint: Tox21 − ER
Type: binary classification
Classes: 1,0
Training set

Overall statistics

Reference		
Prediction	0	1
0	5513	447
1	173	228

N(Compounds)	6361
Accuracy	0.903
P-value	1e-02
Balanced accuracy	0.654
Cohen's Kappa	0.374
MCC	0.389

Statistics by ADM



ADM	Confusion Matrix	Statistics
5	Reference Prediction 0 1 0 442 33 1 16 71	N(Compounds) 562 Accuracy 0.913 P-value 6e-11 Balanced accuracy 0.824 Cohen's Kappa 0.691 MCC 0.695
4	Reference Prediction 0 1 0 811 62 1 19 44	N(Compounds) 936 Accuracy 0.913 P-value 5e-03 Balanced accuracy 0.696 Cohen's Kappa 0.477 MCC 0.496
3	Reference Prediction 0 1 0 1091 104 1 44 69	N(Compounds) 1308 Accuracy 0.887 P-value 2e-02 Balanced accuracy 0.68 Cohen's Kappa 0.422 MCC 0.434
2	Reference Prediction 0 1 0 2082 192 1 80 39	N(Compounds) 2393 Accuracy 0.886 P-value 1e+00 Balanced accuracy 0.566 Cohen's Kappa 0.168 MCC 0.179
1	Reference Prediction 0 1 0 1087 56 1 14 5	N(Compounds) 1162 Accuracy 0.94 P-value 9e-01 Balanced accuracy 0.535 Cohen's Kappa 0.103 MCC 0.122

A.20 Tox21 - ERLBD

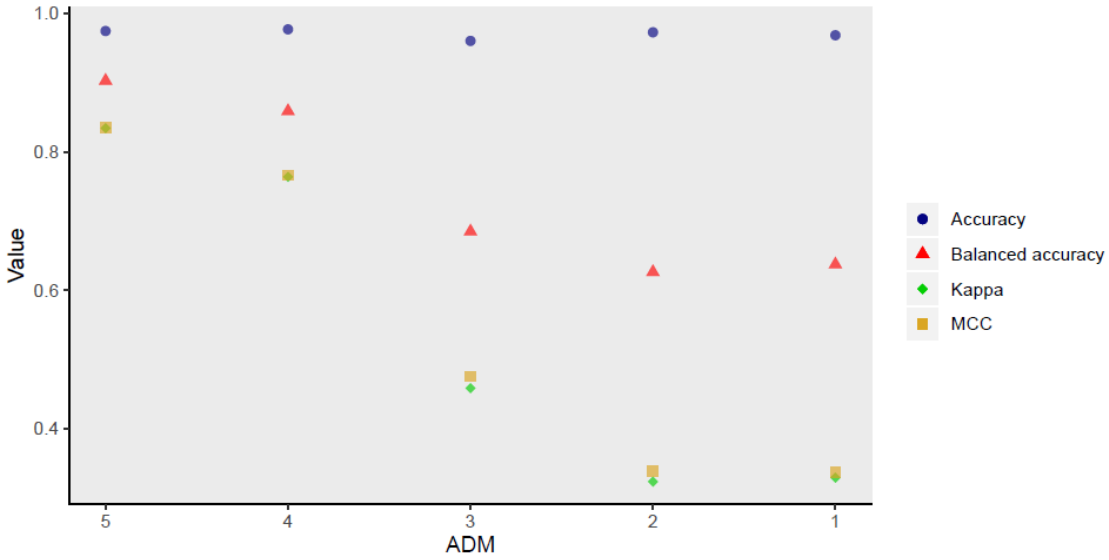
aiQSAR report
2020-07-18 23:19:35
Endpoint: Tox21 - ERLBD
Type: binary classification
Classes: 1,0
Training set

Overall statistics

	Reference	
Prediction	0	1
0	6893	152
1	65	154

N(Compounds)	7264
Accuracy	0.97
P-value	3e-08
Balanced accuracy	0.747
Cohen's Kappa	0.572
MCC	0.58

Statistics by ADM



ADM	Confusion Matrix	Statistics
5	Reference Prediction 0 1 0 669 12 1 7 53	N(Compounds) 741 Accuracy 0.974 P-value 4e-12 Balanced accuracy 0.903 Cohen's Kappa 0.834 MCC 0.835
4	Reference Prediction 0 1 0 1049 17 1 9 45	N(Compounds) 1120 Accuracy 0.977 P-value 1e-07 Balanced accuracy 0.859 Cohen's Kappa 0.764 MCC 0.766
3	Reference Prediction 0 1 0 1392 44 1 15 27	N(Compounds) 1478 Accuracy 0.96 P-value 8e-02 Balanced accuracy 0.685 Cohen's Kappa 0.459 MCC 0.476
2	Reference Prediction 0 1 0 2634 54 1 21 19	N(Compounds) 2728 Accuracy 0.973 P-value 6e-01 Balanced accuracy 0.626 Cohen's Kappa 0.323 MCC 0.339
1	Reference Prediction 0 1 0 1149 25 1 13 10	N(Compounds) 1197 Accuracy 0.968 P-value 7e-01 Balanced accuracy 0.637 Cohen's Kappa 0.329 MCC 0.337

A.21 Tox21 - HSE

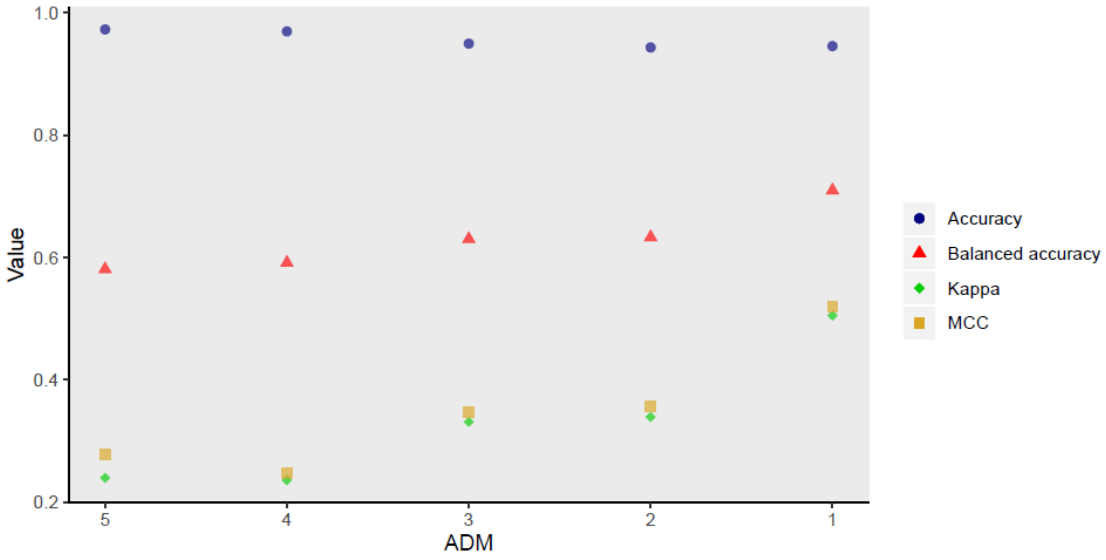
aiQSAR report
2020−08−29 04:21:30
Endpoint: Tox21 − HSE
Type: binary classification
Classes: 1,0
Training set

Overall statistics

	Reference	
Prediction	0	1
0	6335	238
1	86	105

N(Compounds)	6764
Accuracy	0.952
P-value	2e-01
Balanced accuracy	0.646
Cohen's Kappa	0.37
MCC	0.388

Statistics by ADM



ADM	Confusion Matrix	Statistics
5	Reference Prediction 0 1 0 650 15 1 3 3	N(Compounds) 671 Accuracy 0.973 P-value 6e-01 Balanced accuracy 0.581 Cohen's Kappa 0.24 MCC 0.278
4	Reference Prediction 0 1 0 960 21 1 9 5	N(Compounds) 995 Accuracy 0.97 P-value 8e-01 Balanced accuracy 0.592 Cohen's Kappa 0.236 MCC 0.248
3	Reference Prediction 0 1 0 1290 50 1 19 19	N(Compounds) 1378 Accuracy 0.95 P-value 5e-01 Balanced accuracy 0.63 Cohen's Kappa 0.331 MCC 0.347
2	Reference Prediction 0 1 0 2336 104 1 38 41	N(Compounds) 2519 Accuracy 0.944 P-value 4e-01 Balanced accuracy 0.633 Cohen's Kappa 0.339 MCC 0.356
1	Reference Prediction 0 1 0 1099 48 1 17 37	N(Compounds) 1201 Accuracy 0.946 P-value 1e-02 Balanced accuracy 0.71 Cohen's Kappa 0.505 MCC 0.52

A.22 Tox21 - MMP

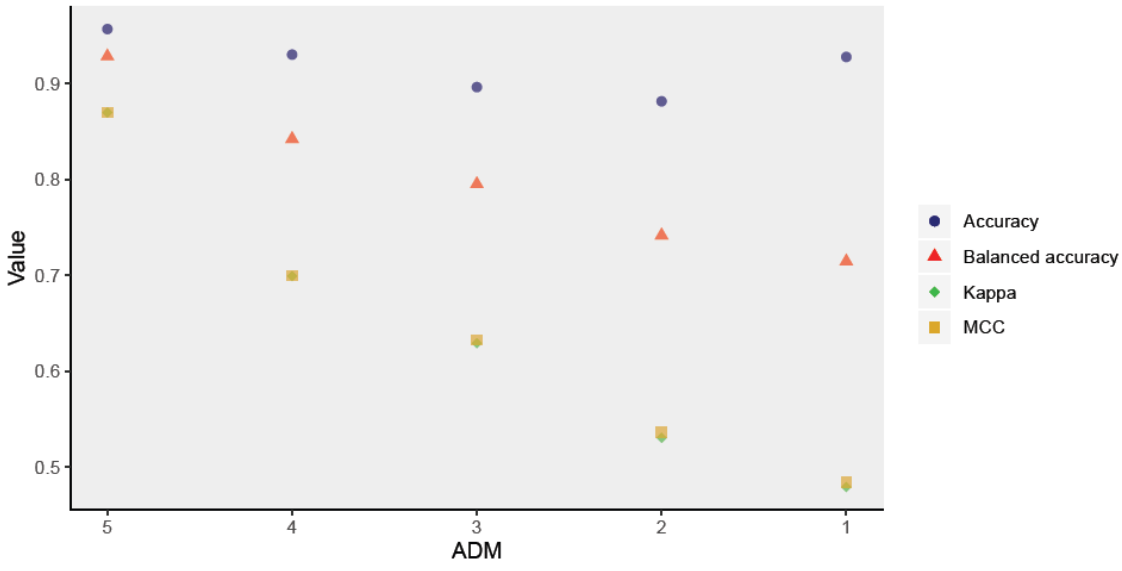
aiQSAR report
2020−08−29 21:20:05
Endpoint: Tox21 − MMP
Type: binary classification
Classes: 1,0
Training set

Overall statistics

Reference		
Prediction	0	1
0	4923	355
1	205	577

N(Compounds)	6060
Accuracy	0.908
P-value	2e-45
Balanced accuracy	0.79
Cohen's Kappa	0.62
MCC	0.623

Statistics by ADM



ADM	Confusion Matrix	Statistics
5	Reference Prediction 0 1 0 448 15 1 10 109	N(Compounds) 582 Accuracy 0.957 P-value 1e-31 Balanced accuracy 0.929 Cohen's Kappa 0.87 MCC 0.87
4	Reference Prediction 0 1 0 741 34 1 28 88	N(Compounds) 891 Accuracy 0.93 P-value 1e-10 Balanced accuracy 0.842 Cohen's Kappa 0.699 MCC 0.7
3	Reference Prediction 0 1 0 898 76 1 43 133	N(Compounds) 1150 Accuracy 0.897 P-value 1e-13 Balanced accuracy 0.795 Cohen's Kappa 0.629 MCC 0.633
2	Reference Prediction 0 1 0 1824 178 1 94 203	N(Compounds) 2299 Accuracy 0.882 P-value 1e-10 Balanced accuracy 0.742 Cohen's Kappa 0.531 MCC 0.536
1	Reference Prediction 0 1 0 1012 52 1 30 44	N(Compounds) 1138 Accuracy 0.928 P-value 7e-02 Balanced accuracy 0.715 Cohen's Kappa 0.479 MCC 0.484

A.23 Tox21 - p53

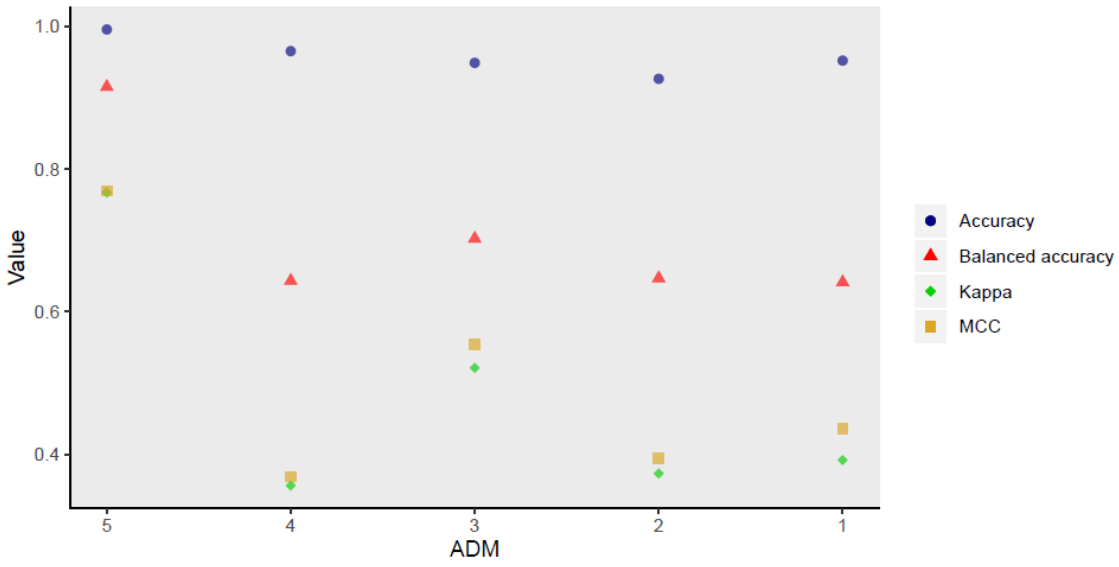
aiQSAR report
2020-08-31 23:48:58
Endpoint: Tox21 – p53
Type: binary classification
Classes: 1,0
Training set

Overall statistics

Reference		
Prediction	0	1
0	6564	287
1	79	148

N(Compounds)	7078
Accuracy	0.948
P-value	3e-04
Balanced accuracy	0.664
Cohen's Kappa	0.423
MCC	0.448

Statistics by ADM



ADM	Confusion Matrix	Statistics
5	Reference Prediction 0 1 0 697 1 1 2 5	N(Compounds) 705 Accuracy 0.996 P-value 1e-01 Balanced accuracy 0.915 Cohen's Kappa 0.767 MCC 0.769
4	Reference Prediction 0 1 0 1021 26 1 11 11	N(Compounds) 1069 Accuracy 0.965 P-value 5e-01 Balanced accuracy 0.643 Cohen's Kappa 0.356 MCC 0.369
3	Reference Prediction 0 1 0 1350 64 1 11 45	N(Compounds) 1470 Accuracy 0.949 P-value 2e-04 Balanced accuracy 0.702 Cohen's Kappa 0.521 MCC 0.554
2	Reference Prediction 0 1 0 2383 147 1 47 67	N(Compounds) 2644 Accuracy 0.927 P-value 8e-02 Balanced accuracy 0.647 Cohen's Kappa 0.373 MCC 0.394
1	Reference Prediction 0 1 0 1113 49 1 8 20	N(Compounds) 1190 Accuracy 0.952 P-value 7e-02 Balanced accuracy 0.641 Cohen's Kappa 0.392 MCC 0.436

A.24 Tox21 - PPAR gamma

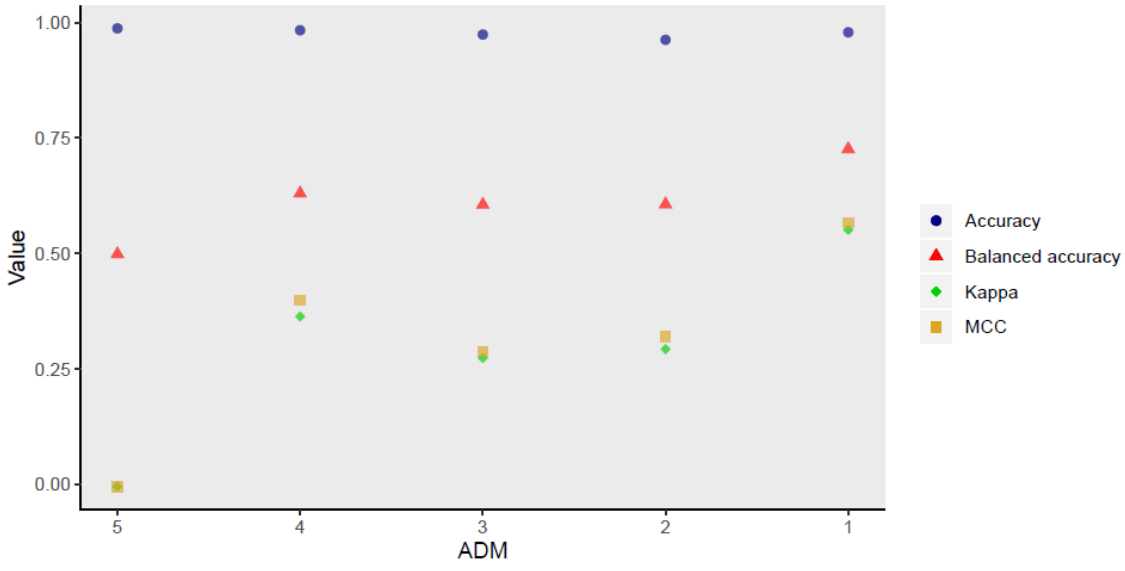
aiQSAR report
2020−09−01 10:26:04
Endpoint: Tox21 − PPAR gamma
Type: binary classification
Classes: 1,0
Training set

Overall statistics

Reference		
Prediction	0	1
0	6542	139
1	41	49

N(Compounds)	6771
Accuracy	0.973
P-value	3e-01
Balanced accuracy	0.627
Cohen's Kappa	0.341
MCC	0.365

Statistics by ADM



ADM	Confusion Matrix	Statistics
5	Reference Prediction 0 1 0 700 7 1 2 0	N(Compounds) 709 Accuracy 0.987 P-value 8e-01 Balanced accuracy 0.499 Cohen's Kappa -0.00441 MCC -0.00531
4	Reference Prediction 0 1 0 999 14 1 3 5	N(Compounds) 1021 Accuracy 0.983 P-value 4e-01 Balanced accuracy 0.63 Cohen's Kappa 0.363 MCC 0.399
3	Reference Prediction 0 1 0 1301 25 1 10 7	N(Compounds) 1343 Accuracy 0.974 P-value 7e-01 Balanced accuracy 0.606 Cohen's Kappa 0.274 MCC 0.288
2	Reference Prediction 0 1 0 2400 74 1 20 21	N(Compounds) 2515 Accuracy 0.963 P-value 5e-01 Balanced accuracy 0.606 Cohen's Kappa 0.293 MCC 0.32
1	Reference Prediction 0 1 0 1142 19 1 6 16	N(Compounds) 1183 Accuracy 0.979 P-value 5e-02 Balanced accuracy 0.726 Cohen's Kappa 0.551 MCC 0.567

A.25 Vega - Androgen



aiQSAR report
2020−07−14 15:58:55
Endpoint: Vega − Androgen
Type: binary classification
Classes: Active,NON−active
Training set

Overall statistics

Reference		
Prediction	**Active**	**NON-active**
Active	99	49
NON-active	98	1418

N(Compounds)	1664
Accuracy	0.912
P-value	5e-05
Balanced accuracy	0.735
Cohen's Kappa	0.526
MCC	0.532

A.26 Vega - Carcinogenicity



aiQSAR report
2020−07−15 15:03:13
Endpoint: Vega − Carcinogenicity
Type: binary classification
Classes: NON−Carcinogen,Carcinogen
Training set

Overall statistics

Reference		
Prediction	**Carcinogen**	**NON-Carcinogen**
Carcinogen	538	233
NON-Carcinogen	243	526

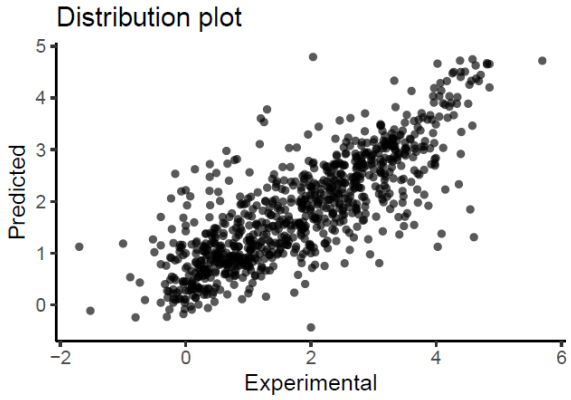
N(Compounds)	1540
Accuracy	0.691
P-value	2e-48
Balanced accuracy	0.691
Cohen's Kappa	0.382
MCC	0.382

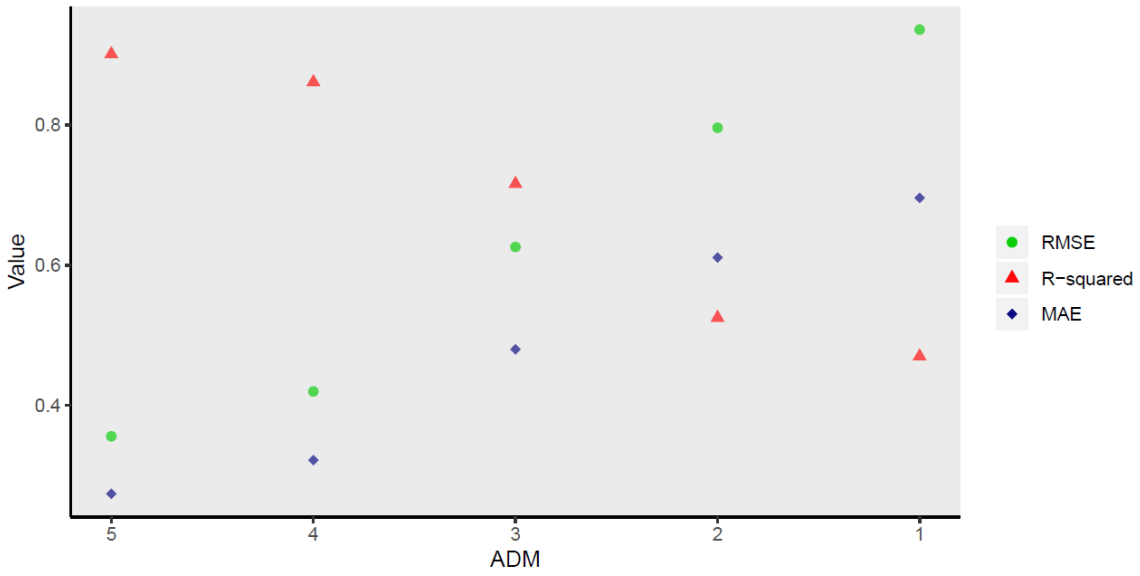
A.27 Vega - BCF

aiQSAR report
2020−07−14 18:22:54
Endpoint: Vega − BCF
Type: regression
Range: −1.7 to 5.694
Training set

Statistics

	N	RMSE	R-squared	MAE
Overall	860	0.781	0.646	0.571
ADM = 1	335	0.936	0.47	0.696
ADM = 2	258	0.796	0.525	0.611
ADM = 3	107	0.626	0.716	0.48
ADM = 4	97	0.42	0.861	0.322
ADM = 5	63	0.356	0.901	0.274



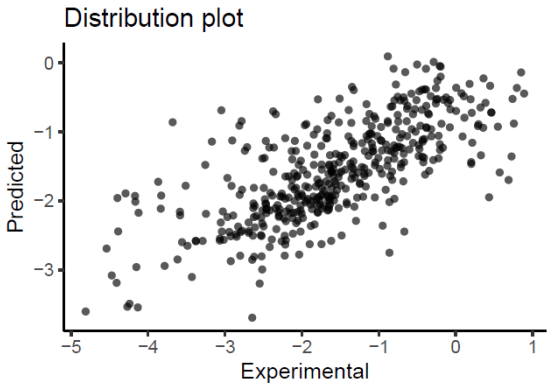


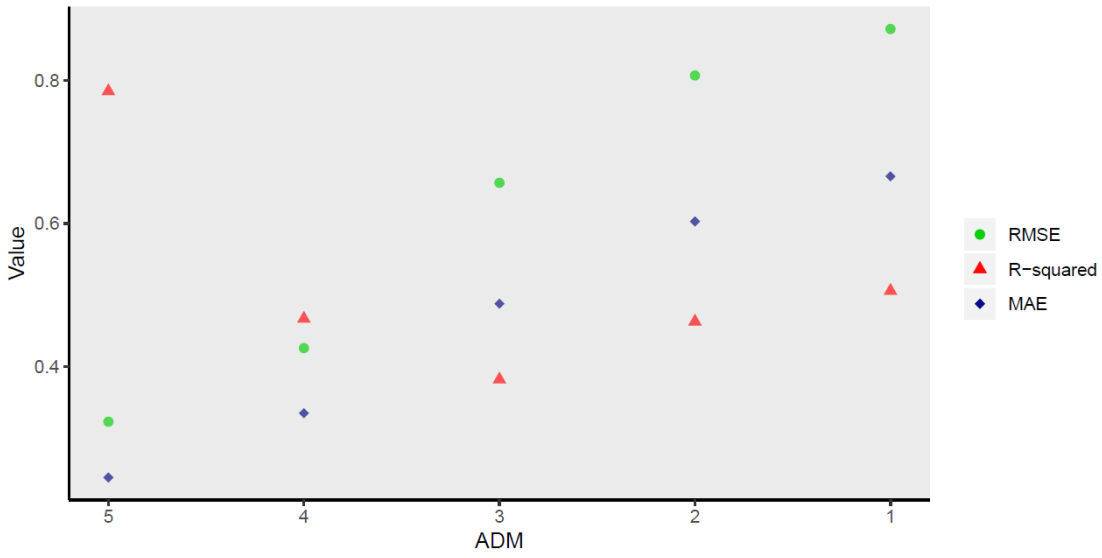
A.28 Vega - Daphnia EC_{50}

aiQSAR report
2020−07−15 16:10:05
Endpoint: Vega − Daphnia EC50
Type: regression
Range: −4.8115 to 0.8894
Training set

Statistics

	N	RMSE	R-squared	MAE
Overall	445	0.777	0.503	0.577
ADM = 1	213	0.872	0.506	0.666
ADM = 2	104	0.807	0.463	0.603
ADM = 3	67	0.657	0.382	0.488
ADM = 4	48	0.426	0.467	0.335
ADM = 5	13	0.323	0.785	0.245





A.29 Vega - Estrogen



aiQSAR report
2020−07−16 14:45:31
Endpoint: Vega − Estrogen
Type: binary classification
Classes: Active,NON−active
Training set

Overall statistics

Reference		
Prediction	**Active**	**NON-active**
Active	57	19
NON-active	32	1421

N(Compounds)	1529
Accuracy	0.967
P-value	5e−06
Balanced accuracy	0.814
Cohen's Kappa	0.673
MCC	0.676

A.30 Vega - Mutagenicity



aiQSAR report
2020−07−23 05:57:07
Endpoint: Vega − Mutagenicity
Type: binary classification
Classes: NON−Mutagenic,Mutagenic
Training set

Overall statistics

Reference		
Prediction	**Mutagenic**	**NON-Mutagenic**
Mutagenic	2690	562
NON-Mutagenic	559	1951

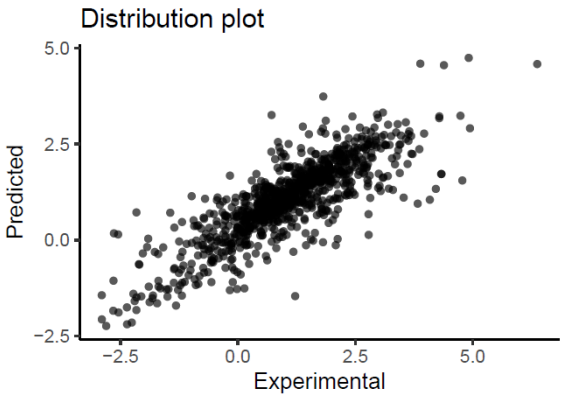
N(Compounds)	5762
Accuracy	0.805
P-value	0e+00
Balanced accuracy	0.802
Cohen's Kappa	0.604
MCC	0.604

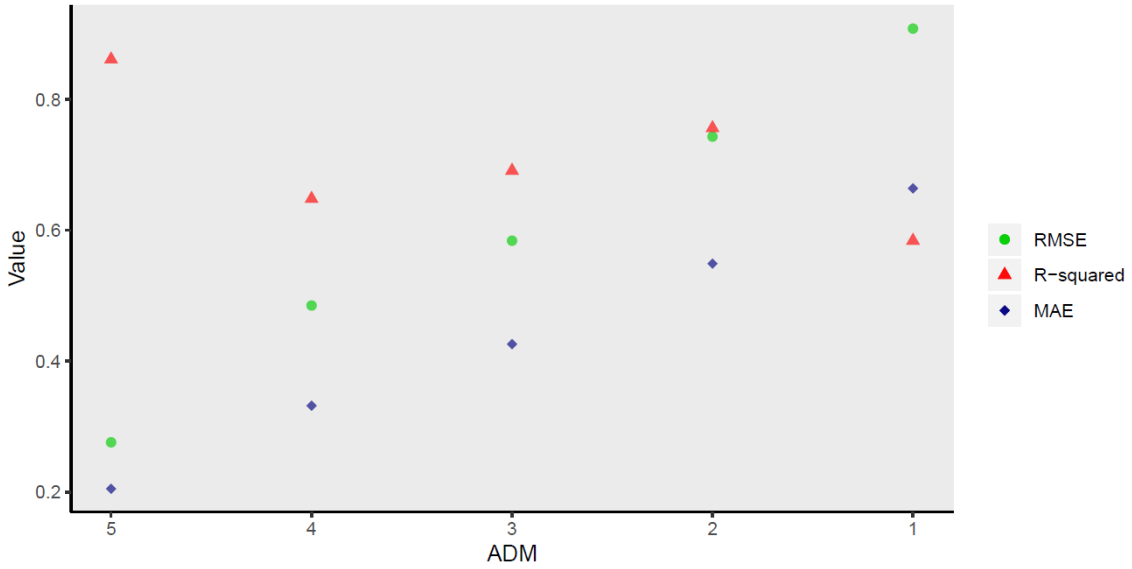
A.31 Vega - Fish

aiQSAR report
2020−07−16 15:38:49
Endpoint: Vega − Fish
Type: regression
Range: −2.9 to 6.376
Training set

Statistics

	N	RMSE	R-squared	MAE
Overall	945	0.728	0.683	0.513
ADM = 1	316	0.908	0.584	0.664
ADM = 2	259	0.743	0.756	0.549
ADM = 3	184	0.584	0.691	0.426
ADM = 4	127	0.485	0.648	0.332
ADM = 5	59	0.276	0.861	0.205



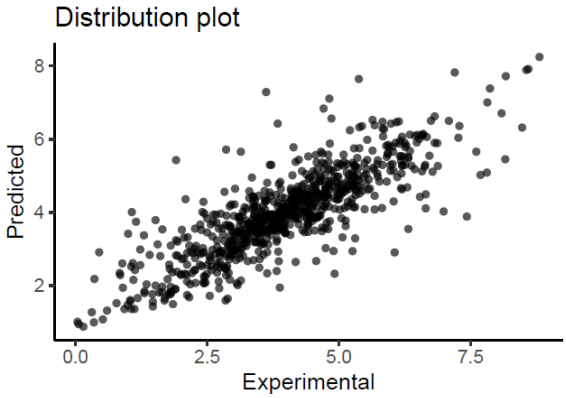


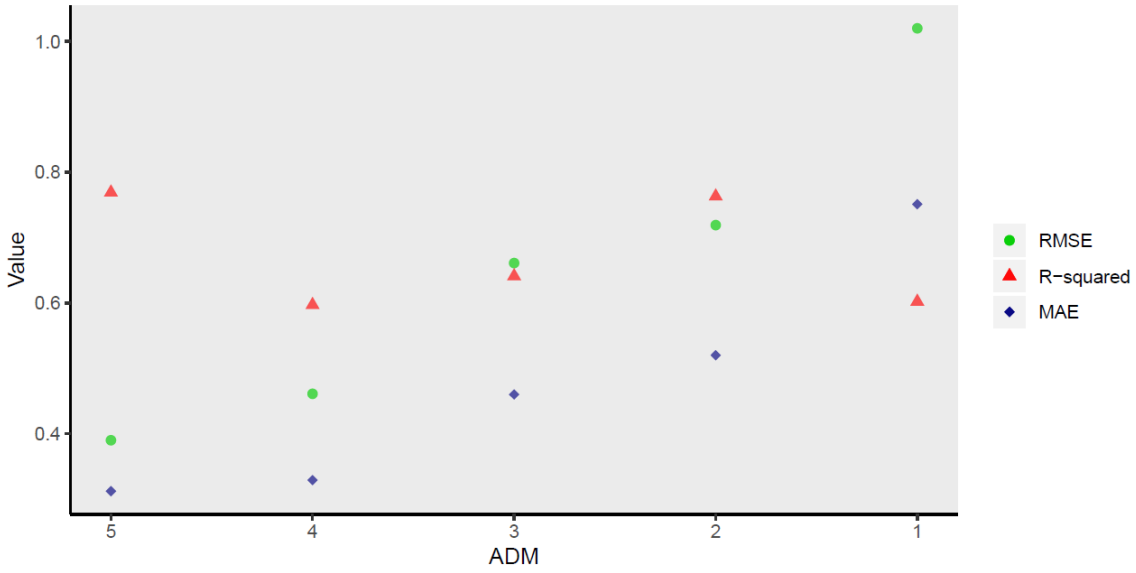
A.32 Vega - Fathead

aiQSAR report
2020−07−15 18:53:05
Endpoint: Vega − Fathead
Type: regression
Range: 0.04 to 8.81
Training set

Statistics

	N	RMSE	R-squared	MAE
Overall	797	0.824	0.679	0.584
ADM = 1	321	1.02	0.602	0.751
ADM = 2	265	0.719	0.763	0.52
ADM = 3	134	0.661	0.641	0.46
ADM = 4	47	0.461	0.597	0.329
ADM = 5	30	0.39	0.769	0.312



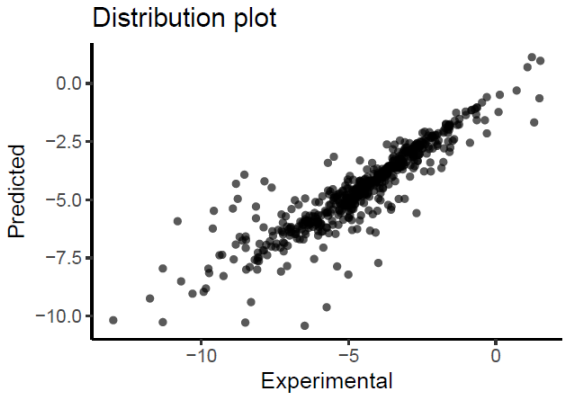


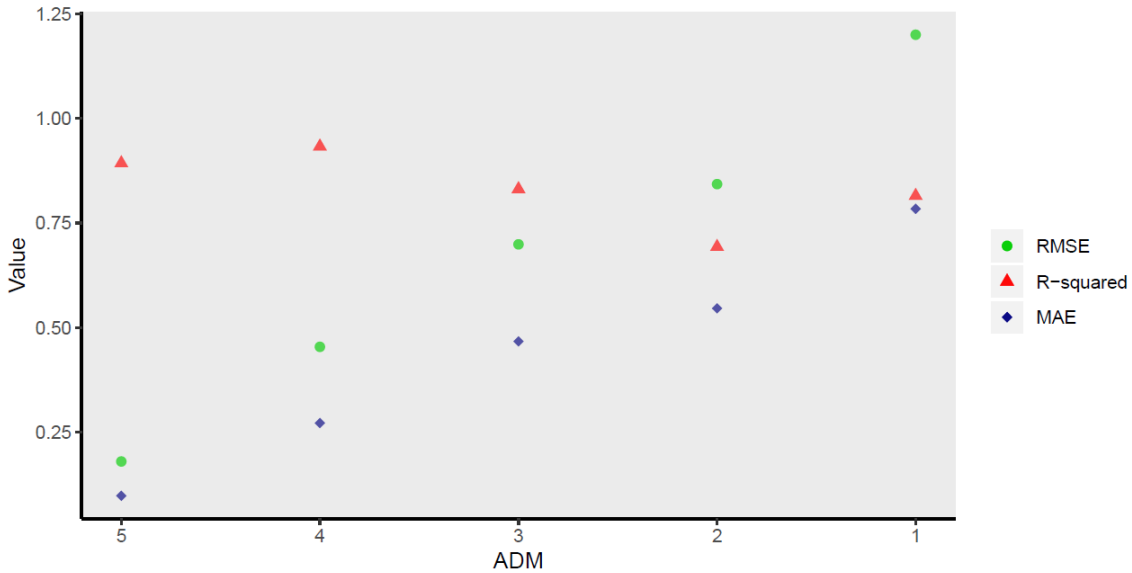
A.33 Vega - Henry's law

aiQSAR report
2020−07−21 12:34:43
Endpoint: Vega − Henry's law
Type: regression
Range: −12.9914 to 1.5185139
Training set

Statistics

	N	RMSE	R-squared	MAE
Overall	589	0.921	0.814	0.547
ADM = 1	247	1.2	0.815	0.784
ADM = 2	140	0.843	0.693	0.546
ADM = 3	66	0.699	0.831	0.467
ADM = 4	46	0.454	0.933	0.272
ADM = 5	90	0.18	0.893	0.098



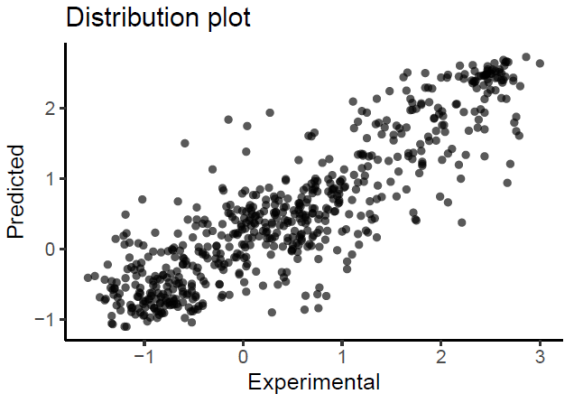


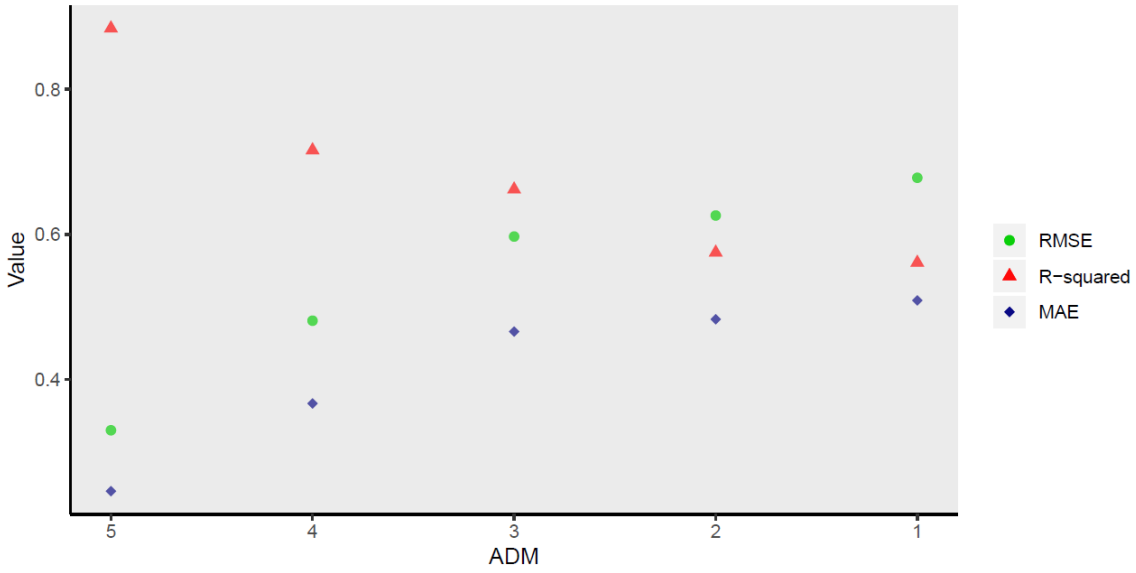
A.34 Vega - KM

aiQSAR report
2020−07−16 17:52:23
Endpoint: Vega − KM
Type: regression
Range: −1.57 to 3
Training set

Statistics

	N	RMSE	R-squared	MAE
Overall	611	0.559	0.77	0.415
ADM = 1	135	0.678	0.561	0.509
ADM = 2	138	0.626	0.575	0.483
ADM = 3	98	0.597	0.662	0.466
ADM = 4	112	0.481	0.716	0.367
ADM = 5	128	0.33	0.884	0.246



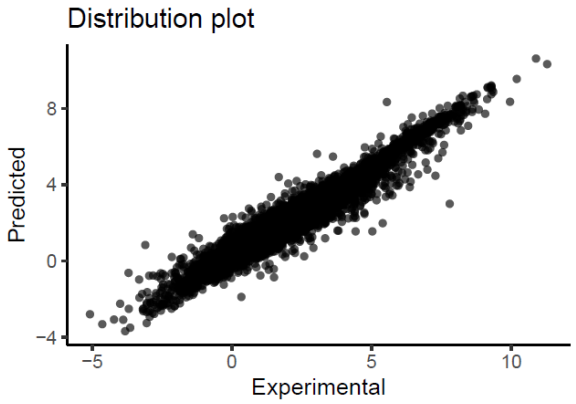


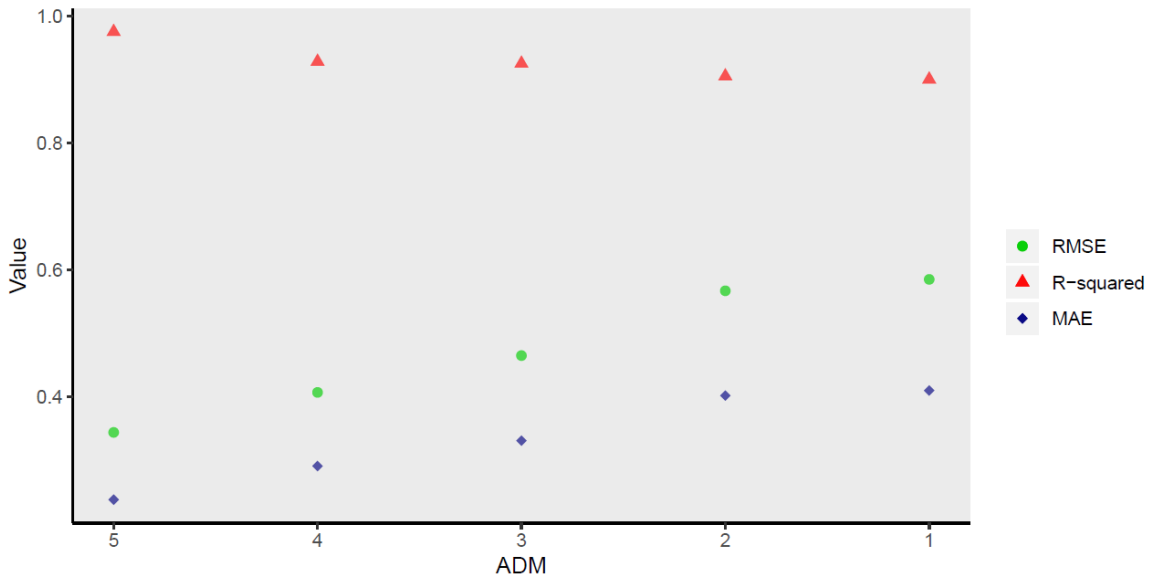
A.35 Vega - LogP

aiQSAR report
2020−07−22 08:12:16
Endpoint: Vega − LogP
Type: regression
Range: −5.08 to 11.29
Training set

Statistics

	N	RMSE	R-squared	MAE
Overall	9961	0.465	0.937	0.323
ADM = 1	823	0.585	0.9	0.41
ADM = 2	2233	0.567	0.905	0.402
ADM = 3	2233	0.465	0.925	0.331
ADM = 4	2504	0.407	0.928	0.291
ADM = 5	2168	0.344	0.975	0.238





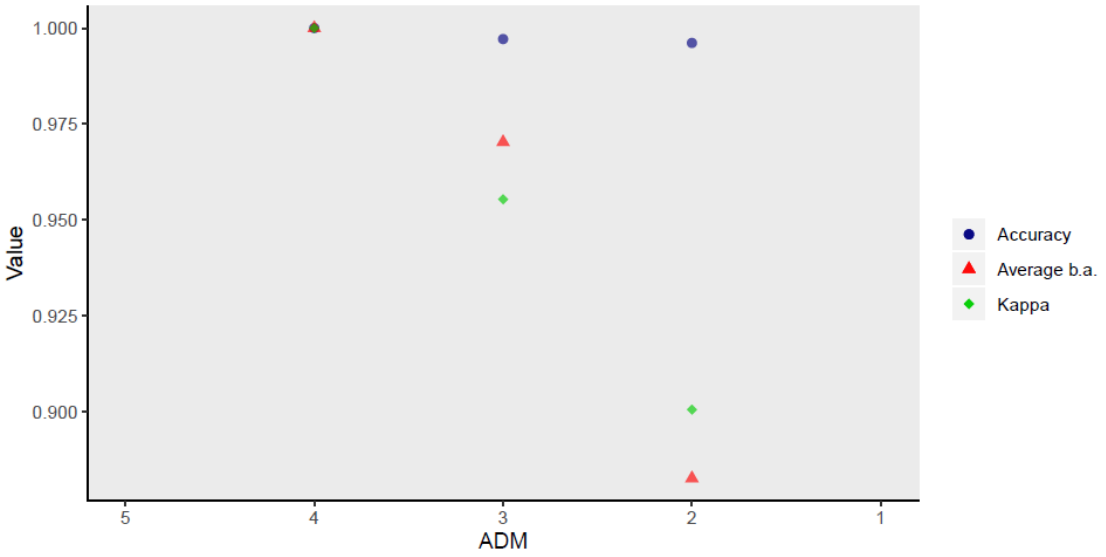
A.36 Vega - Thyroid Receptor Alpha

aiQSAR report
2020-07-25 17:05:42
Endpoint: Vega - Thyroid Receptor Alpha
Type: multi-class classification
Classes: Agonist,Antagonist,Inactive
Training set

Overall statistics

Reference					
Prediction	**Agonist**	**Antagonist**	**Inactive**	N(Compounds)	5462
Agonist	93	0	3	Accuracy	0.998
Antagonist	2	23	0	P-value	1e-39
Inactive	3	3	5335	Average balanced accuracy	0.964
				Cohen's Kappa	0.954

Statistics by ADM



ADM	Confusion Matrix	Statistics
5	NA	NA

4

Reference					
Prediction	**Agonist**	**Antagonist**	**Inactive**		
Agonist	15	0	0		
Antagonist	0	8	0		
Inactive	0	0	831		

N(Compounds)	854
Accuracy	1
P-value	7e-11
Average balanced accuracy	1
Cohen's Kappa	1

3

Reference					
Prediction	**Agonist**	**Antagonist**	**Inactive**		
Agonist	20	0	1		
Antagonist	0	13	0		
Inactive	1	1	1016		

N(Compounds)	1052
Accuracy	0.997
P-value	3e-12
Average balanced accuracy	0.97
Cohen's Kappa	0.955

2

Reference					
Prediction	**Agonist**	**Antagonist**	**Inactive**		
Agonist	34	0	2		
Antagonist	2	2	0		
Inactive	2	2	2026		

N(Compounds)	2070
Accuracy	0.996
P-value	1e-10
Average balanced accuracy	0.883
Cohen's Kappa	0.901

1	NA	NA

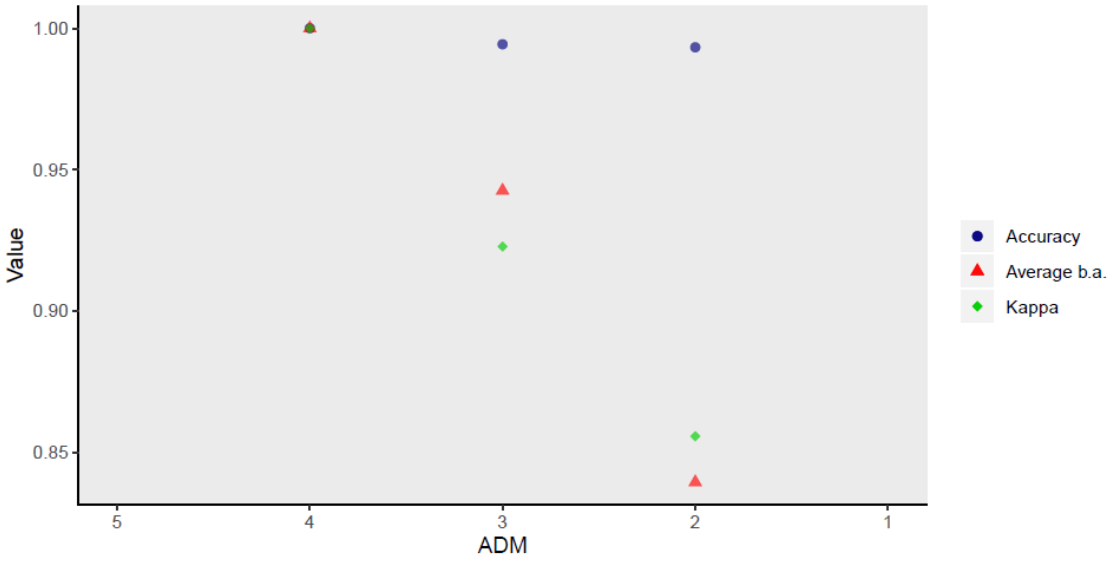
A.37 Vega - Thyroid Receptor Beta

aiQSAR report
2020−07−25 20:15:02
Endpoint: Vega − Thyroid Receptor Beta
Type: multi−class classification
Classes: Agonist,Antagonist,Inactive
Training set

Overall statistics

Reference					
Prediction	**Agonist**	**Antagonist**	**Inactive**	N(Compounds)	5473
Agonist	98	2	1	Accuracy	0.996
Antagonist	3	25	2	P-value	1e-36
Inactive	5	9	5328	Average balanced accuracy	0.92
				Cohen's Kappa	0.918

Statistics by ADM



ADM	Confusion Matrix	Statistics
5	NA	NA

4

Reference					
Prediction	**Agonist**	**Antagonist**	**Inactive**	N(Compounds)	849
Agonist	14	0	0	Accuracy	1
Antagonist	0	8	0	P-value	2e-10
Inactive	0	0	827	Average balanced accuracy	1
				Cohen's Kappa	1

3

Reference					
Prediction	**Agonist**	**Antagonist**	**Inactive**	N(Compounds)	1061
Agonist	23	1	0	Accuracy	0.994
Antagonist	1	13	1	P-value	7e-12
Inactive	1	2	1019	Average balanced accuracy	0.943
				Cohen's Kappa	0.923

2

Reference					
Prediction	**Agonist**	**Antagonist**	**Inactive**	N(Compounds)	2078
Agonist	37	1	0	Accuracy	0.993
Antagonist	2	4	1	P-value	7e-11
Inactive	4	6	2023	Average balanced accuracy	0.839
				Cohen's Kappa	0.856

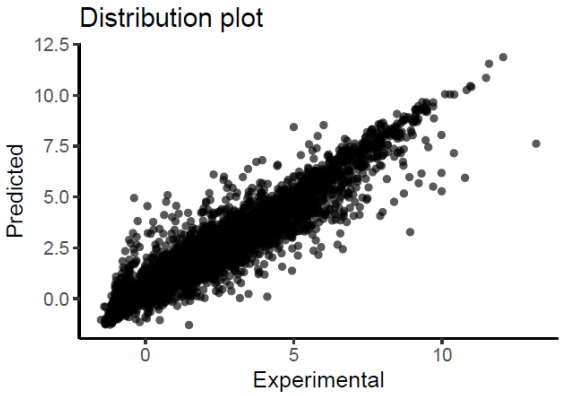
1	NA	NA

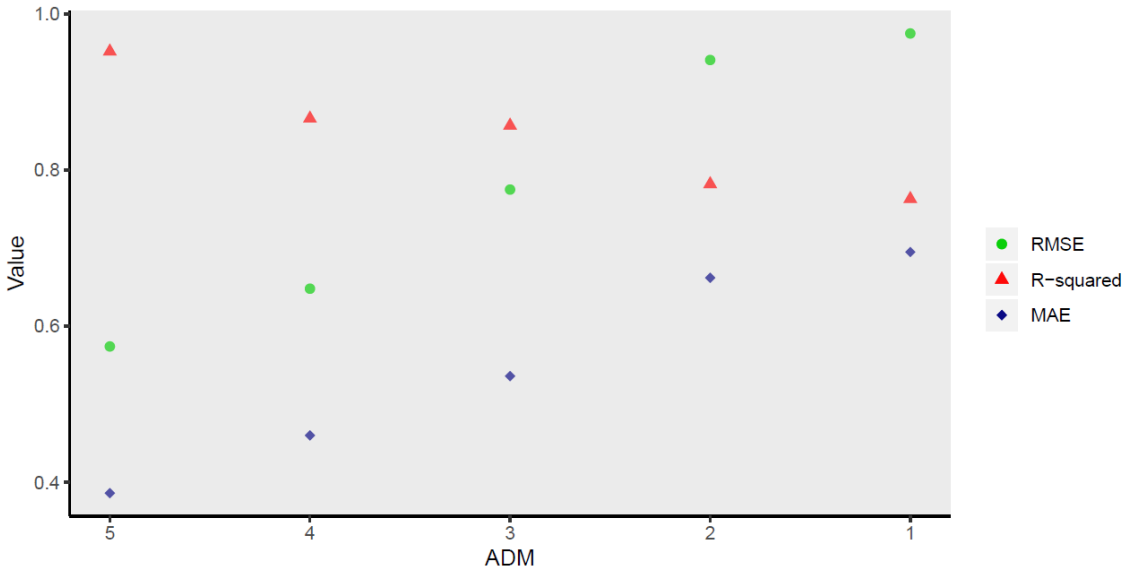
A.38 Vega - Water solubility

aiQSAR report
2020−07−24 18:32:17
Endpoint: Vega − Water solubility
Type: regression
Range: −1.494 to 13.172
Training set

Statistics

	N	RMSE	R-squared	MAE
Overall	4694	0.83	0.858	0.572
ADM = 1	818	0.975	0.763	0.695
ADM = 2	1560	0.941	0.782	0.662
ADM = 3	909	0.775	0.857	0.536
ADM = 4	718	0.648	0.866	0.46
ADM = 5	689	0.574	0.952	0.386





Appendix B

Differential Expression – Tested Chemicals and Genes

In this appendix, the genes and the chemicals related to the aiQSAR differential expression modelling experiment are provided (see Section 3.3.4). For the list of 210 tested chemicals, the columns of the table are the conventional name of the compound, chemical formula and the corresponding, distinct SMILES in canonicalized form. For the genes, all 478 modelled genes are listed with the corresponding number of active chemicals and statistical performance metrics.

B.1 Chemicals

Chemical Name	Formula	SMILES
(R)-Lansoprazole	$C_{16}H_{14}F_3N_3O_2S$	Cc1c(C[S@@](=O)c2[nH]c3ccccc3n2)nccc1OCC(F)(F)F
6-Thioguanine	$C_5H_5N_5S$	NC1=NC(=S)c2[nH]cnc2N1
Abacavir	$C_{14}H_{18}N_6O$	Nc1nc(NC2CC2)c3ncn([C@@H]4C[C@H](CO)C=C4)c3n1
Acetazolamide	$C_4H_6N_4O_3S_2$	CC(=O)Nc1sc(nn1)[S](N)(=O)=O
Acetohydroxamic acid	$C_2H_5NO_2$	CC(=O)NO
Acetylcysteine	$C_5H_9NO_3S$	CC(=O)NC(CS)C(O)=O
Acyclovir	$C_8H_{11}N_5O_3$	NC1=NC(=O)c2ncn(COCCO)c2N1
Adenosine	$C_{10}H_{13}N_5O_4$	Nc1ncnc2n(cnc12)C3OC(CO)C(O)C3O
Albendazole	$C_{12}H_{15}N_3O_2S$	CCCSc1ccc2nc(NC(=O)OC)[nH]c2c1
Allopurinol	$C_5H_4N_4O$	O=C1N=CN=C2NNC=C12
Alosetron (Hydrochloride)	$C_{17}H_{19}ClN_4O$	[H+].[Cl-].Cn1c2CCN(Cc3nc[nH]c3C)C(=O)c2c4ccccc14
Amiodarone (hydrochloride)	$C_{25}H_{30}ClI_2NO_3$	[H+].[Cl-].O=C(C1=CC(I)=C(OCCN(CC)CC)C(I)=C1)C=2C=3C=CC=CC3OC2CCCC
Amodiaquin (dihydrochloride dihydrate)	$C_{20}H_{28}Cl_3N_3O_3$	[H+].[H+].O.O.[Cl-].[Cl-].CCN(CC)Cc1cc(Nc2ccnc3cc(Cl)ccc23)ccc1O
Amoxapine	$C_{17}H_{16}ClN_3O$	Clc1ccc2Oc3ccccc3N=C(N4CCNCC4)c2c1
Anagrelide (hydrochloride)	$C_{10}H_8Cl_3N_3O$	[H+].[Cl-].Clc1ccc2NC3=NC(=O)CN3Cc2c1Cl
Anastrozole	$C_{17}H_{19}N_5$	CC(C)(C#N)c1cc(Cn2cncn2)cc(c1)C(C)(C)C#N
Antipyrine	$C_{11}H_{12}N_2O$	CN1N(C(=O)C=C1C)c2ccccc2
Aprepitant	$C_{23}H_{21}F_7N_4O_3$	C[C@@H](O[C@@H]1OCCN(CC2=NC(=O)NN2)[C@@H]1c3ccc(F)cc3)c4cc(cc(c4)C(F)(F)F)C(F)(F)F
Aspirin	$C_9H_8O_4$	CC(=O)Oc1ccccc1C(O)=O
Azasetron (hydrochloride)	$C_{17}H_{21}Cl_2N_3O_3$	[H+].[Cl-].O=C(NC1CN2CCC1CC2)C3=CC(Cl)=CC4=C3OCC(=O)N4C

Chemical Name	Formula	SMILES
Azathioprine	$C_9H_7N_7O_2S$	Cn1cnc(c1Sc2ncnc3nc[nH]c23)[N+]([O-])=O
Aztreonam	$C_{13}H_{17}N_5O_8S_2$	C[C@H]1[C@H](NC(=O)C(=N/OC(C)(C)C(O)=O)\c2csc([NH3+])n2)C(=O)N1[S]([O-])(=O)=O
Bendazol	$C_{14}H_{13}ClN_2$	[H+].[Cl-].N=1C=2C=CC=CC2NC1CC=3C=CC=CC3
Benzydamine (hydrochloride)	$C_{19}H_{24}ClN_3O$	[H+].[Cl-].CN(C)CCCOc1nn(Cc2ccccc2)c3ccccc13
Bethanechol (chloride)	$C_7H_{17}ClN_2O_2$	[Cl-].CC(C[N+](C)(C)C)OC(N)=O
Bicalutamide	$C_{18}H_{14}F_4N_2O_4S$	CC(O)(C[S](=O)(=O)c1ccc(F)cc1)C(=O)Nc2ccc(C#N)c(c2)C(F)(F)F
Bifonazole	$C_{22}H_{18}N_2$	c1ccc(cc1)C(n2ccnc2)c3ccc(cc3)c4ccccc4
Budesonide	$C_{25}H_{34}O_6$	CCC[C@@H]1O[C@@H]2C[C@H]3[C@@H]4CCC5=CC(=O)C=C[C@]5(C)[C@H]4[C@@H](O)C[C@]3(C)[C@@]2(O1)C(=O)CO
Buspirone (hydrochloride)	$C_{21}H_{32}ClN_5O_2$	[H+].[Cl-].O=C1CC2(CCCC2)CC(=O)N1CCCCN3CCN(CC3)c4ncccn4
Busulfan	$C_6H_{14}O_6S_2$	O=S(=O)(OCCCCOS(=O)(=O)C)C
Butoconazole (nitrate)	$C_{19}H_{18}Cl_3N_3O_3S$	O[N+]([O-])=O.Clc1ccc(CCC(Cn2ccnc2)Sc3c(Cl)cccc3Cl)cc1
Carbamazepine	$C_{15}H_{12}N_2O$	NC(=O)N1c2ccccc2C=Cc3ccccc13
Carbimazole	$C_7H_{10}N_2O_2S$	CCOC(=O)N1C=CN(C)C1=S
Cefdinir	$C_{14}H_{13}N_5O_5S_2$	Nc1scc(n1)\C(=N\O)C(=O)N[C@H]2[C@H]3SCC(=C(N3C2=O)C(O)=O)C=C
Cefepime (Dihydrochloride Monohydrate)	$C_{19}H_{28}Cl_2N_6O_6S_2$	O.[Cl-].[Cl-].CO\N=C(C(=O)N[C@H]1[C@H]2SCC(=C(N2C1=O)C(O)=O)C[N+]3(C)CCCC3)\c4csc(N)n4.[H+]
Cefixime	$C_{16}H_{15}N_5O_7S_2$	Nc1scc(n1)\C(=N\OCC(O)=O)C(=O)N[C@H]2[C@H]3SCC(=C(N3C2=O)C(O)=O)C=C
Cefoselis (sulfate)	$C_{19}H_{24}N_8O_{10}S_3$	CO\N=C(C(=O)N[C@H]1[C@H]2SCC(=C(N2C1=O)C(O)=O)CN3C=CC(=N)N3CCO)\c4csc(N)n4.O[S](O)(=O)=O
Cefozopran (hydrochloride)	$C_{19}H_{18}ClN_9O_5S_2$	[Cl-].CO\N=C(C(=O)N[C@H]1[C@H]2SCC(=C(N2C1=O)C(O)=O)Cn3cc[n+]4ncccc34)\c5nsc(N)n5
Cevimeline (hydrochloride)	$C_{10}H_{18}ClNOS$	[Cl-].C[C@@H]1O[C@@]2(CS1)CN3CCC2CC3.[H+]
Chlormezanone	$C_{11}H_{12}ClNO_3S$	CN1C(c2ccc(Cl)cc2)[S](=O)(=O)CCC1=O
Chlorzoxazone	$C_7H_4ClNO_2$	Clc1ccc2OC(=O)Nc2c1
Cinacalcet (hydrochloride)	$C_{22}H_{23}ClF_3N$	[Cl-].C[C@@H](NCCCc1cccc(c1)C(F)(F)F)c2cccc3ccccc23.[H+]
Ciprofibrate	$C_{13}H_{14}Cl_2O_3$	O=C(O)C(OC1=CC=C(C=C1)C2CC2(Cl)Cl)(C)C
Cladribine	$C_{10}H_{12}ClN_5O_3$	Nc1nc(Cl)nc2n(cnc12)[C@H]3C[C@H](O)[C@@H](CO)O3
Clemizole (hydrochloride)	$C_{19}H_{21}Cl_2N_3$	[H+].[Cl-].Clc1ccc(Cn2c(CN3CCCC3)nc4ccccc24)cc1
Clofarabine	$C_{10}H_{11}ClFN_5O_3$	Nc1nc(Cl)nc2n(cnc12)[C@@H]3O[C@H](CO)[C@@H](O)[C@@H]3F
Clofibrate	$C_{12}H_{15}ClO_3$	O=C(OCC)C(OC1=CC=C(Cl)C=C1)(C)C
Clonidine (hydrochloride)	$C_9H_{10}Cl_3N_3$	[H+].[Cl-].Clc1cccc(Cl)c1NC2=NCCN2
Colchicine	$C_{22}H_{25}NO_6$	COC1=CC=C2C(=CC1=O)[C@H](CCc3cc(OC)c(OC)c(OC)c23)NC(C)=O
Conivaptan (hydrochloride)	$C_{32}H_{27}ClN_4O_2$	[H+].[Cl-].Cc1[nH]c2CCN(C(=O)c3ccc(NC(=O)c4ccccc4c5ccccc5)cc3)c6ccccc6c2n1
Cyclosporin A	$C_{62}H_{111}N_{11}O_{12}$	O=C1NC(C(=O)N(C)C(C(=O)N(C)C(C(=O)N(C)C(C(=O)N(C)C(C(=O)NC(C(=O)N(C)C(C(=O)NC1C)CC(C)C)C(C)C)CC(C)C)CC)C(O)C(C)CC=CC)C(C)C)CC(C)C)CC(C)C)C
Danazol	$C_{22}H_{27}NO_2$	C[C@]12Cc3cnoc3C=C1CC[C@@H]4[C@@H]2CC[C@@]5(C)[C@H]4CC[C@@]5(O)C#C

Chemical Name	Formula	SMILES
Dapagliflozin ((2S)-1,2-propanediol, hydrate)	$C_{24}H_{35}ClO_9$	ClC1=CC=C(C=C1CC2=CC=C(OCC)C=C2)C3OC(CO)C(O)C(O)C3O.O.OCC(O)C
Darifenacin (hydrobromide)	$C_{28}H_{31}BrN_2O_2$	[Br-].NC(=O)C([C@@H]1CCN(CCc2ccc3OCCc3c2)C1)(c4ccccc4)c5ccccc5.[H+]
Deracoxib	$C_{17}H_{14}F_3N_3O_3S$	COc1ccc(cc1F)c2cc(nn2c3ccc(cc3)[S](N)(=O)=O)C(F)F
Dexchlorpheniramine (maleate)	$C_{20}H_{23}ClN_2O_4$	CN(C)CC[C@@H](c1ccc(Cl)cc1)c2ccccn2.OC(=O)\C=C/C(O)=O
Diatrizoic acid	$C_{11}H_9I_3N_2O_4$	CC(=O)Nc1c(I)c(NC(C)=O)c(I)c(C(O)=O)c1I
Diclofenac	$C_{14}H_{11}Cl_2NO_2$	O=C(O)CC=1C=CC=CC1NC=2C(Cl)=CC=CC2Cl
Dihydroergotamine (mesylate)	$C_{34}H_{41}N_5O_8S$	CN1C[C@@H](C[C@H]2[C@H]1Cc3c[nH]c4cccc2c34)C(=O)N[C@]5(C)O[C@@]6(O)[C@@H]7CCCN7C(=O)[C@H](Cc8ccccc8)N6C5=O.C[S](O)(=O)=O
Dimesna	$C_4H_8Na_2O_6S_4$	[Na+].[Na+].O=S(=O)([O-])CCSSCCS(=O)(=O)[O-]
Dimethyl fumarate	$C_6H_8O_4$	O=C(OC)C=CC(=O)OC
Dimetridazole	$C_5H_7N_3O_2$	Cn1c(C)ncc1[N+]([O-])=O
Diphylline	$C_{10}H_{14}N_4O_4$	CN1C(=O)N(C)c2ncn(CC(O)CO)c2C1=O
Disulfiram	$C_{10}H_{20}N_2S_4$	CCN(CC)C(=S)SSC(=S)N(CC)CC
Domperidone	$C_{22}H_{24}ClN_5O_2$	Clc1ccc2N(C3CCN(CCCN4C(=O)Nc5ccccc45)CC3)C(=O)Nc2c1
Doxofylline	$C_{11}H_{14}N_4O_4$	CN1C(=O)N(C)c2ncn(CC3OCCO3)c2C1=O
Doxorubicin (hydrochloride)	$C_{27}H_{30}ClNO_{11}$	[H+].[Cl-].O=C1C=2C=CC=C(OC)C2C(=O)C=3C(O)=C4C(=C(O)C13)CC(O)(C(=O)CO)CC4OC5OC(C)C(O)C(N)C5
Droperidol	$C_{22}H_{22}FN_3O_2$	Fc1ccc(cc1)C(=O)CCCN2CCC(=CC2)N3C(=O)Nc4ccccc34
Dutasteride	$C_{27}H_{30}F_6N_2O_2$	C[C@]12CCC3[C@@H](CC[C@H]4NC(=O)C=C[C@]34C)[C@@H]1CC[C@@H]2C(=O)Nc5cc(ccc5C(F)(F)F)C(F)(F)F
Econazole (nitrate)	$C_{18}H_{16}Cl_3N_3O_4$	O[N+]([O-])=O.Clc1ccc(COC(Cn2ccnc2)c3ccc(Cl)cc3Cl)cc1
Efavirenz	$C_{14}H_9ClF_3NO_2$	FC(F)(F)[C@]1(OC(=O)Nc2ccc(Cl)cc12)C#CC3CC3
Eprosartan (mesylate)	$C_{24}H_{28}N_2O_7S_2$	CCCCc1ncc(\C=C(/Cc2sccc2)C(O)=O)n1Cc3ccc(cc3)C(O)=O.C[S](O)(=O)=O
Etofenamate	$C_{18}H_{18}F_3NO_4$	OCCOCCOC(=O)c1ccccc1Nc2cccc(c2)C(F)(F)F
Etofylline	$C_9H_{12}N_4O_3$	CN1C(=O)N(C)c2ncn(CCO)c2C1=O
Etomidate	$C_{14}H_{16}N_2O_2$	CCOC(=O)c1cncn1C(C)c2ccccc2
Ezetimibe	$C_{24}H_{21}F_2NO_3$	O=C1N(C2=CC=C(F)C=C2)C(C3=CC=C(O)C=C3)C1CCC(O)C4=CC=C(F)C=C4
Famciclovir	$C_{14}H_{19}N_5O_4$	CC(=O)OCC(CCn1cnc2cnc(N)nc12)COC(C)=O
Febuxostat	$C_{16}H_{16}N_2O_3S$	CC(C)COc1ccc(cc1C#N)c2sc(C(O)=O)c(C)n2
Flecainide (acetate)	$C_{19}H_{24}F_6N_2O_5$	CC(O)=O.FC(F)(F)COc1ccc(OCC(F)(F)F)c(c1)C(=O)NCC2CCCCN2
Flibanserin	$C_{20}H_{21}F_3N_4O$	FC(F)(F)c1ccc(cc1)N2CCN(CC2)CCN3C(=O)Nc4ccccc34
Fluconazole	$C_{13}H_{12}F_2N_6O$	OC(Cn1cncn1)(Cn2cncn2)c3ccc(F)cc3F
Fludarabine	$C_{10}H_{12}FN_5O_4$	Nc1nc(F)nc2n(cnc12)[C@@H]3O[C@H](CO)[C@@H](O)[C@@H]3O
Flufenamic acid	$C_{14}H_{10}F_3NO_2$	OC(=O)c1ccccc1Nc2cccc(c2)C(F)(F)F
Flumazenil	$C_{15}H_{14}FN_3O_3$	CCOC(=O)c1ncn2c1CN(C)C(=O)c3cc(F)ccc23
Fluoxetine (hydrochloride)	$C_{17}H_{19}ClF_3NO$	[H+].[Cl-].CNCCC(Oc1ccc(cc1)C(F)(F)F)c2ccccc2
Fluphenazine (dihydrochloride)	$C_{22}H_{28}Cl_2F_3N_3OS$	[H+].[H+].[Cl-].[Cl-].OCCN1CCN(CCCN2c3ccccc3Sc4ccc(cc24)C(F)(F)F)CC1
Flutamide	$C_{11}H_{11}F_3N_2O_3$	CC(C)C(=O)Nc1ccc(c(c1)C(F)(F)F)[N+]([O-])=O

Chemical Name	Formula	SMILES
Fluticasone (propionate)	$C_{25}H_{31}F_3O_5S$	CCC(=O)O[C@@]1([C@H](C)C[C@H]2C3C[C@H](F)C4=CC(=O)C=C[C@]4(C)[C@@]3(F)[C@@H](O)C[C@]12C)C(=O)SCF
Fluvoxamine (maleate)	$C_{19}H_{25}F_3N_2O_6$	COCCCC\C(=N/OCCN)c1ccc(cc1)C(F)(F)F.OC(=O)\C=C/C(O)=O
Fulvestrant	$C_{32}H_{47}F_5O_3S$	C[C@]12CC[C@H]3[C@@H]([C@H](CCCCCCCCC[S](=O)CCCC(F)(F)C(F)(F)F)Cc4cc(O)ccc34)[C@@H]1CC[C@@H]2O
Furosemide	$C_{12}H_{11}ClN_2O_5S$	N[S](=O)(=O)c1cc(C(O)=O)c(NCc2occc2)cc1Cl
Ganciclovir	$C_9H_{13}N_5O_4$	NC1=NC(=O)c2ncn(COC(CO)CO)c2N1
Gemfibrozil	$C_{15}H_{22}O_3$	O=C(O)C(C)(C)CCCOC1=CC(=CC=C1C)C
Goserelin (acetate)	$C_{61}H_{88}N_{18}O_{16}$	CC(C)CC(NC(=O)C(COC(C)(C)C)NC(=O)C(Cc1ccc(O)cc1)NC(=O)C(CO)NC(=O)C(Cc2c[nH]c3ccccc23)NC(=O)C(Cc4[nH]cnc4)NC(=O)C5CCC(=O)N5)C(=O)NC(CCCN=C(N)N)C(=O)N6CCCC6C(=O)NNC(N)=O.CC(O)=O
Granisetron (Hydrochloride)	$C_{18}H_{25}ClN_4O$	[H+].[Cl-].CN1C2CCCC1CC(C2)NC(=O)c3nn(C)c4ccccc34
Hydroxyzine (dihydrochloride)	$C_{21}H_{29}Cl_3N_2O_2$	[H+].[H+].[Cl-].[Cl-].OCCOCCN1CCN(CC1)C(c2ccccc2)c3ccc(Cl)cc3
Ibudilast	$C_{14}H_{18}N_2O$	CC(C)C(=O)c1c2ccccn2nc1C(C)C
Ibuprofen	$C_{13}H_{18}O_2$	CC(C)Cc1ccc(cc1)C(C)C(O)=O
Imatinib	$C_{29}H_{31}N_7O$	O=C(NC1=CC=C(C(=C1)NC=2N=CC=C(N2)C3=CN=CC=C3)C)C4=CC=C(C=C4)CN5CCN(C)CC5
Imiquimod	$C_{14}H_{16}N_4$	CC(C)Cn1cnc2c(N)nc3ccccc3c12
Indapamide	$C_{16}H_{16}ClN_3O_3S$	O=C(NN1C=2C=CC=CC2CC1C)C3=CC=C(Cl)C(=C3)S(=O)(=O)N
Iohexol	$C_{19}H_{26}I_3N_3O_9$	CC(=O)N(CC(O)CO)c1c(I)c(C(=O)NCC(O)CO)c(I)c(C(=O)NCC(O)CO)c1I
Iopamidol	$C_{17}H_{22}I_3N_3O_8$	C[C@H](O)C(=O)Nc1c(I)c(C(=O)NC(CO)CO)c(I)c(C(=O)NC(CO)CO)c1I
Iopanoic acid	$C_{11}H_{12}I_3NO_2$	CCC(Cc1c(I)cc(I)c(N)c1I)C(O)=O
Iopromide	$C_{18}H_{24}I_3N_3O_8$	COCC(=O)Nc1c(I)c(C(=O)NCC(O)CO)c(I)c(c1I)C(=O)N(C)CC(O)CO
Ioversol	$C_{18}H_{24}I_3N_3O_9$	OCCN(C(=O)CO)c1c(I)c(C(=O)NCC(O)CO)c(I)c(C(=O)NCC(O)CO)c1I
Iproniazid (phosphate)	$C_9H_{16}N_3O_5P$	CC(C)NNC(=O)c1ccncc1.O[P](O)(O)=O
Isoniazid	$C_6H_7N_3O$	NNC(=O)c1ccncc1
Isosorbide mononitrate	$C_6H_9NO_6$	O[C@H]1CO[C@@H]2[C@@H](CO[C@H]12)O[N+]([O-])=O
Isradipine	$C_{19}H_{21}N_3O_5$	COC(=O)C1=C(C)NC(=C(C1c2cccc3nonc23)C(=O)OC(C)C)C
Itraconazole	$C_{35}H_{38}Cl_2N_8O_4$	CCC(C)N1N=CN(C1=O)c2ccc(cc2)N3CCN(CC3)c4ccc(OC[C@H]5CO[C@@](Cn6cncn6)(O5)c7ccc(Cl)cc7Cl)cc4
Kanamycin (sulfate)	$C_{18}H_{38}N_4O_{15}S$	NC[C@H]1O[C@H](O[C@@H]2[C@@H](N)C[C@@H](N)[C@H](O[C@H]3O[C@H](CO)[C@@H](O)[C@H](N)[C@H]3O)[C@H]2O)[C@H](O)[C@@H](O)[C@@H]1O.O[S](O)(=O)=O
Ketoconazole	$C_{26}H_{28}Cl_2N_4O_4$	CC(=O)N1CCN(CC1)c2ccc(OC[C@@H]3CO[C@](Cn4ccnc4)(O3)c5ccc(Cl)cc5Cl)cc2
L-Thyroxine	$C_{15}H_{11}I_4NO_4$	NC(Cc1cc(I)c(Oc2cc(I)c(O)c(I)c2)c(I)c1)C(O)=O
Leflunomide	$C_{12}H_9F_3N_2O_2$	Cc1oncc1C(=O)Nc2ccc(cc2)C(F)(F)F
Letrozole	$C_{17}H_{11}N_5$	N#Cc1ccc(cc1)C(n2cncn2)c3ccc(cc3)C#N
Leuprolide Acetate	$C_{61}H_{88}N_{16}O_{14}$	CCNC(=O)C1CCCN1C(=O)C(CCCN=C(N)N)NC(=O)C(CC(C)C)NC(=O)C(CC(C)C)NC(=O)C(Cc2ccc(O)cc2)NC(=O)C(CO)NC(=O)C(Cc3c[nH]c4ccccc34)NC(=O)C(Cc5[nH]cnc5)NC(=O)C6CCC(=O)N6.CC(O)=O
Luliconazole	$C_{14}H_9Cl_2N_3S_2$	Clc1ccc([C@@H]2CSC(/S2)=C(/C#N)n3ccnc3)c(Cl)c1

Chemical Name	Formula	SMILES
Mebendazole	$C_{16}H_{13}N_3O_3$	COC(=O)Nc1[nH]c2cc(ccc2n1)C(=O)c3ccccc3
Mefenamic acid	$C_{15}H_{15}NO_2$	Cc1cccc(Nc2ccccc2C(O)=O)c1C
Mefloquine (hydrochloride)	$C_{17}H_{17}ClF_6N_2O$	[H+].[Cl-].OC(C1CCCCN1)c2cc(nc3c2cccc3C(F)(F)F)C(F)(F)F
Meloxicam	$C_{14}H_{13}N_3O_4S_2$	CN1C(=C(/O)Nc2sc(C)cn2)/C(=O)c3ccccc3[S]1(=O)=O
Mesna	$C_2H_5NaO_3S_2$	[Na+].O=S(=O)([O-])CCS
Methazolamide	$C_5H_8N_4O_3S_2$	CN1N=C(SC1=NC(C)=O)[S](N)(=O)=O
Methimazole	$C_4H_6N_2S$	CN1C=CNC1=S
Metronidazole	$C_6H_9N_3O_3$	Cc1ncc(n1CCO)[N+]([O-])=O
Metyrapone	$C_{14}H_{14}N_2O$	CC(C)(c1cccnc1)C(=O)c2cccnc2
Misoprostol	$C_{22}H_{38}O_5$	CCCCC(C)(O)C/C=C/[C@H]1[C@H](O)CC(=O)[C@@H]1CCCCCCC(=O)OC
Molsidomine	$C_9H_{14}N_4O_4$	CCOC(=O)[N-]c1cn(n[o+]1)N2CCOCC2
Naproxen (sodium)	$C_{14}H_{13}NaO_3$	[Na+].COc1ccc2cc(ccc2c1)[C@H](C)C([O-])=O
Nefazodone (hydrochloride)	$C_{25}H_{33}Cl_2N_5O_2$	[H+].[Cl-].CCC1=NN(CCCN2CCN(CC2)c3cccc(Cl)c3)C(=O)N1CCOc4ccccc4
Neomycin (sulfate)	$C_{23}H_{48}N_6O_{17}S$	O=S(=O)(O)O.OCC1OC(OC2C(O)C(N)CC(N)C2OC3OC(CN)C(O)C(O)C3N)C(O)C1OC4OC(CN)C(O)C(O)C4N
Netupitant	$C_{30}H_{32}F_6N_4O$	CN1CCN(CC1)c2cc(c(cn2)N(C)C(=O)C(C)(C)c3cc(cc(c3)C(F)(F)F)C(F)(F)F)c4ccccc4C
Niflumic acid	$C_{13}H_9F_3N_2O_2$	OC(=O)c1cccnc1Nc2cccc(c2)C(F)(F)F
Nilotinib	$C_{28}H_{22}F_3N_7O$	Cc1cn(cn1)c2cc(NC(=O)c3ccc(C)c(Nc4nccc(n4)c5cccnc5)c3)cc(c2)C(F)(F)F
Nimesulide	$C_{13}H_{12}N_2O_5S$	C[S](=O)(=O)Nc1ccc(cc1Oc2ccccc2)[N+]([O-])=O
Nitazoxanide	$C_{12}H_9N_3O_5S$	CC(=O)Oc1ccccc1C(=O)Nc2sc(cn2)[N+]([O-])=O
Nitisinone	$C_{14}H_{10}F_3NO_5$	[O-][N+](=O)c1cc(ccc1C(=O)C2C(=O)CCCC2=O)C(F)(F)F
Nitrofurantoin	$C_8H_6N_4O_5$	[O-][N+](=O)c1oc(cc1)/C=N/N2CC(=O)NC2=O
Olopatadine (hydrochloride)	$C_{21}H_{24}ClNO_3$	[H+].[Cl-].CN(C)CC\C=C/1c2ccccc2COc3ccc(CC(O)=O)cc13
Omeprazole	$C_{17}H_{19}N_3O_3S$	O=S(C1=NC=2C=CC(OC)=CC2N1)CC3=NC=C(C(OC)=C3C)C
Ornidazole	$C_7H_{10}ClN_3O_3$	Cc1ncc(n1CC(O)CCl)[N+]([O-])=O
Oxaprozin	$C_{18}H_{15}NO_3$	OC(=O)CCc1oc(c2ccccc2)c(n1)c3ccccc3
Oxiconazole nitrate	$C_{18}H_{14}Cl_4N_4O_4$	O[N+]([O-])=O.Clc1ccc(CO\N=C(/Cn2ccnc2)c3ccc(Cl)cc3Cl)c(Cl)c1
Paliperidone	$C_{23}H_{27}FN_4O_3$	CC1=C(CCN2CCC(CC2)c3noc4cc(F)ccc34)C(=O)N5CCCC(O)C5=N1
Pamidronic acid	$C_3H_{11}NO_7P_2$	O=P(O)(O)C(O)(CCN)P(=O)(O)O
Pancuronium (dibromide)	$C_{35}H_{60}Br_2N_2O_4$	[Br-].[Br-].CC(=O)O[C@H]1C[C@@H]2CC[C@@H]3[C@H](CC[C@@]4(C)[C@H]3C[C@@H]([C@@H]4OC(C)=O)[N+]5(C)CCCCC5)[C@@]2(C)C[C@@H]1[N+]6(C)CCCCC6
Parecoxib (Sodium)	$C_{19}H_{17}N_2NaO_4S$	[Na+].CCC(=O)[N-][S](=O)(=O)c1ccc(cc1)c2c(C)onc2c3ccccc3
Paromomycin (sulfate)	$C_{23}H_{47}N_5O_{18}S$	O=S(=O)(O)O.OCC1OC(OC2C(OC3OC(CO)C(OC4OC(CN)C(O)C(O)C4N)C3O)C(O)C(N)CC2N)C(N)C(O)C1O
Penciclovir	$C_{10}H_{15}N_5O_3$	NC1=NC(=O)c2ncn(CCC(CO)CO)c2N1
Penfluridol	$C_{28}H_{27}ClF_5NO$	OC1(CCN(CCCC(c2ccc(F)cc2)c3ccc(F)cc3)CC1)c4ccc(Cl)c(c4)C(F)(F)F
Phenytoin	$C_{15}H_{12}N_2O_2$	O=C1NC(=O)C(N1)(c2ccccc2)c3ccccc3
Pimecrolimus	$C_{43}H_{68}ClNO_{11}$	O=C1C(=O)C2(O)OC(C(OC)CC(C)CC(=CC(C(=O)CC(O)C(C)C(OC(=O)C3N1CCCC3)C(=CC4CCC(Cl)C(OC)C4)C)CC)C)C(OC)CC2C
Pimozide	$C_{28}H_{29}F_2N_3O$	Fc1ccc(cc1)C(CCCN2CCC(CC2)N3C(=O)Nc4ccccc34)c5ccc(F)cc5
Pioglitazone	$C_{19}H_{20}N_2O_3S$	O=C1SC(C(=O)N1)CC2=CC=C(OCCC3=NC=C(C=C3)CC)C=C2

Chemical Name	Formula	SMILES
Propranolol (hydrochloride)	$C_{16}H_{22}ClNO_2$	[H+].[Cl-].CC(C)NCC(O)COc1cccc2ccccc12
Propylthiouracil	$C_7H_{10}N_2OS$	CCCC1=CC(=O)NC(=S)N1
Propyphenazone	$C_{14}H_{18}N_2O$	CC(C)C1=C(C)N(C)N(c2ccccc2)C1=O
Proxyphylline	$C_{10}H_{14}N_4O_3$	CC(O)Cn1cnc2N(C)C(=O)N(C)C(=O)c12
Ramosetron (Hydrochloride)	$C_{17}H_{18}ClN_3O$	[Cl-].Cn1cc(C(=O)[C@@H]2CCc3nc[nH]c3C2)c4ccccc14.[H+]
Retinoic acid	$C_{20}H_{28}O_2$	O=C(O)C=C(C=CC=C(C=CC1=C(C)CCCC1(C)C)C)C
Ribavirin	$C_8H_{12}N_4O_5$	NC(=O)c1ncn(n1)[C@@H]2O[C@H](CO)[C@@H](O)[C@H]2O
Rifampicin	$C_{43}H_{58}N_4O_{12}$	O=C(OC1C(C)C(OC)C=COC2(OC3=C(C2=O)C4=C(O)C(C=NN5CCN(C)CC5)=C(NC(=O)C(=CC=CC(C)C(O)C(C)C(O)C1C)C)C(O)=C4C(O)=C3C)C)C
Rifaximin	$C_{43}H_{51}N_3O_{11}$	CO[C@H]1/C=C/O[C@@]2(C)Oc3c(C)c(O)c4c(O)c(NC(=O)C(=C\C=C\[C@H](C)[C@H](O)[C@@H](C)[C@@H](O)[C@@H](C)[C@H](OC(C)=O)[C@@H]1C)/C)c5n6ccc(C)cc6nc5c4c3C2=O
Riluzole	$C_8H_5F_3N_2OS$	Nc1sc2cc(OC(F)(F)F)ccc2n1
Ritonavir	$C_{37}H_{48}N_6O_5S_2$	CC(C)[C@H](NC(=O)N(C)Cc1csc(n1)C(C)C)C(=O)N[C@H](C[C@H](O)[C@H](Cc2ccccc2)NC(=O)OCc3scnc3)Cc4ccccc4
Ronidazole	$C_6H_8N_4O_4$	Cn1c(COC(N)=O)ncc1[N+]([O-])=O
Rosiglitazone	$C_{18}H_{19}N_3O_3S$	O=C1SC(C(=O)N1)CC2=CC=C(OCCN(C3=NC=CC=C3)C)C=C2
Salbutamol (hemisulfate)	$C_{26}H_{44}N_2O_{10}S$	CC(C)(C)NCC(O)c1ccc(O)c(CO)c1.CC(C)(C)NCC(O)c2ccc(O)c(CO)c2.O[S](O)(=O)=O
Secnidazole	$C_7H_{11}N_3O_3$	CC(O)Cn1c(C)ncc1[N+]([O-])=O
Sertaconazole (nitrate)	$C_{20}H_{16}Cl_3N_3O_4S$	O[N+]([O-])=O.Clc1ccc(C(Cn2ccnc2)OCc3csc4c(Cl)cccc34)c(Cl)c1
Sildenafil (citrate)	$C_{28}H_{38}N_6O_{11}S$	CCCc1nn(C)c2C(=O)N=C(Nc12)c3cc(ccc3OCC)[S](=O)(=O)N4CCN(C)CC4.OC(=O)CC(O)(CC(O)=O)C(O)=O
Silodosin	$C_{25}H_{32}F_3N_3O_4$	C[C@H](Cc1cc2CCN(CCCO)c2c(c1)C(N)=O)NCCOc3ccccc3OCC(F)(F)F
Simvastatin	$C_{25}H_{38}O_5$	CCC(C)(C)C(=O)O[C@H]1C[C@@H](C)C=C2C=C[C@H](C)[C@H](CC[C@@H]3C[C@@H](O)CC(=O)O3)[C@@H]12
Sitagliptin (phosphate monohydrate)	$C_{16}H_{20}F_6N_5O_6P$	O.N[C@@H](CC(=O)N1CCn2c(C1)nnc2C(F)(F)F)Cc3cc(F)c(F)cc3F.O[P](O)(O)=O
Sorafenib (Tosylate)	$C_{28}H_{24}ClF_3N_4O_6S$	CNC(=O)c1cc(Oc2ccc(NC(=O)Nc3ccc(Cl)c(c3)C(F)(F)F)cc2)ccn1.Cc4ccc(cc4)[S](O)(=O)=O
Succinylsulfathiazole	$C_{13}H_{13}N_3O_5S_2$	OC(=O)CCC(=O)Nc1ccc(cc1)[S](=O)(=O)Nc2sccn2
Sulfamethizole	$C_9H_{10}N_4O_2S_2$	Cc1sc(N[S](=O)(=O)c2ccc(N)cc2)nn1
Sulfamethoxazole	$C_{10}H_{11}N_3O_3S$	Cc1onc(N[S](=O)(=O)c2ccc(N)cc2)c1
Sulfaphenazole	$C_{15}H_{14}N_4O_2S$	Nc1ccc(cc1)[S](=O)(=O)Nc2ccnn2c3ccccc3
Sulfasalazine	$C_{18}H_{14}N_4O_5S$	OC(=O)C1=C\C(C=CC1=O)=N/Nc2ccc(cc2)[S](=O)(=O)Nc3ccccn3
Sulfisoxazole	$C_{11}H_{13}N_3O_3S$	Cc1noc(N[S](=O)(=O)c2ccc(N)cc2)c1C
Sulindac	$C_{20}H_{17}FO_3S$	CC\1=C(CC(O)=O)c2cc(F)ccc2C1=C\c3ccc(cc3)[S](C)=O
Tamoxifen	$C_{26}H_{29}NO$	CC/C(c1ccccc1)=C(c2ccccc2)/c3ccc(OCCN(C)C)cc3
Tazobactam	$C_{10}H_{12}N_4O_5S$	C[C@]1(Cn2ccnn2)[C@@H](N3[C@@H](CC3=O)[S]1(=O)=O)C(O)=O
Telmisartan	$C_{33}H_{30}N_4O_2$	CCCc1nc2c(C)cc(cc2n1Cc3ccc(cc3)c4ccccc4C(O)=O)c5nc6ccccc6n5C
Temozolomide	$C_6H_6N_6O_2$	CN1N=Nc2n(cnc2C(N)=O)C1=O

Chemical Name	Formula	SMILES
Tenofovir (Disoproxil Fumarate)	$C_{23}H_{34}N_5O_{14}P$	CC(C)OC(=O)OCO[P](=O)(CO[C@H](C)Cn1cnc2c(N)ncnc12)OCOC(=O)OC(C)C.OC(=O)\C=C\C(O)=O
Tetrabenazine	$C_{19}H_{27}NO_3$	COc1cc2CCN3CC(CC(C)C)C(=O)CC3c2cc1OC
Theophylline	$C_7H_8N_4O_2$	CN1C(=O)N(C)c2nc[nH]c2C1=O
Ticlopidine (hydrochloride)	$C_{14}H_{15}Cl_2NS$	[H+].[Cl-].Clc1ccccc1CN2CCc3sccc3C2
Tinidazole	$C_8H_{13}N_3O_4S$	CC[S](=O)(=O)CCn1c(C)ncc1[N+]([O-])=O
Tioconazole	$C_{16}H_{13}Cl_3N_2OS$	Clc1ccc(C(Cn2ccnc2)OCc3ccsc3Cl)c(Cl)c1
Tiratricol	$C_{14}H_8I_3NaO_4$	[Na+].Oc1ccc(Oc2c(I)cc(CC([O-])=O)cc2I)cc1I
Tobramycin	$C_{18}H_{37}N_5O_9$	OCC1OC(OC2C(O)C(OC3OC(CN)C(O)CC3N)C(N)CC2N)C(O)C(N)C1O
Trapidil	$C_{10}H_{15}N_5$	CCN(CC)c1cc(C)nc2ncnn12
Trazodone (hydrochloride)	$C_{19}H_{23}Cl_2N_5O$	[H+].[Cl-].Clc1cccc(c1)N2CCN(CCCN3N=C4C=CC=CN4C3=O)CC2
Trifluoperazine (dihydrochloride)	$C_{21}H_{26}Cl_2F_3N_3S$	[H+].[H+].[Cl-].[Cl-].CN1CCN(CCCN2c3ccccc3Sc4ccc(cc24)C(F)(F)F)CC1
Triflupromazine (hydrochloride)	$C_{18}H_{20}ClF_3N_2S$	[H+].[Cl-].CN(C)CCCN1c2ccccc2Sc3ccc(cc13)C(F)(F)F
Trifluridine	$C_{10}H_{11}F_3N_2O_5$	OCC1OC(CC1O)N2C=C(C(=O)NC2=O)C(F)(F)F
Triflusal	$C_{10}H_7F_3O_4$	CC(=O)Oc1cc(ccc1C(O)=O)C(F)(F)F
Troglitazone	$C_{24}H_{27}NO_5S$	O=C1SC(C(=O)N1)CC2=CC=C(OCC3(OC=4C(=C(C(O)=C(C4CC3)C)C)C)C)C=C2
Valdecoxib	$C_{16}H_{14}N_2O_3S$	Cc1onc(c2ccccc2)c1c3ccc(cc3)[S](N)(=O)=O
Valganciclovir (hydrochloride)	$C_{14}H_{23}ClN_6O_5$	[Cl-].CC(C)[C@H](N)C(=O)OCC(CO)OCn1cnc2C(=O)N=C(N)Nc12.[H+]
Valproic acid (sodium salt)	$C_8H_{15}NaO_2$	[Na+].CCCC(CCC)C([O-])=O
Valrubicin	$C_{34}H_{36}F_3NO_{13}$	CCCCC(=O)OCC(=O)[C@@]1(O)C[C@H](OC2CC(NC(=O)C(F)(F)F)C(O)C(C)O2)c3c(O)c4C(=O)c5c(OC)cccc5C(=O)c4c(O)c3C1
Venlafaxine (hydrochloride)	$C_{17}H_{28}ClNO_2$	[H+].[Cl-].COc1ccc(cc1)C(CN(C)C)C2(O)CCCCC2
Voriconazole	$C_{16}H_{14}F_3N_5O$	C[C@@H](c1ncncc1F)[C@](O)(Cn2cncn2)c3ccc(F)cc3F
Vorinostat	$C_{14}H_{20}N_2O_3$	O=C(NO)CCCCCCC(=O)NC=1C=CC=CC1
Zonisamide	$C_8H_8N_2O_3S$	N[S](=O)(=O)Cc1noc2ccccc12

B.2 Genes

Gene	Active Chem.	Accuracy	Balanced Acc.	Kappa	MCC
ABCA1	23	0.895	0.579	0.227	0.275
ABCB10	20	0.89	0.537	0.104	0.12
ABCC1	43	0.781	0.577	0.187	0.2
ABCF1	17	0.9	0.516	0.047	0.054
ACAA1	14	0.91	0.487	-0.036	-0.042
ACAT2	15	0.919	0.526	0.078	0.097
ACBD3	17	0.905	0.519	0.056	0.068
ACP1	30	0.852	0.553	0.149	0.182
ACSL1	11	0.929	0.49	-0.029	-0.033
ACTA2	16	0.919	0.583	0.225	0.247
ACTG2	17	0.914	0.497	-0.009	-0.021

Gene	Active Chem.	Accuracy	Balanced Acc.	Kappa	MCC
AHSA1	24	0.862	0.523	0.063	0.072
AKAP8	12	0.924	0.49	-0.029	-0.034
AKR7A2	20	0.9	0.587	0.23	0.252
AKT1	15	0.929	0.562	0.186	0.232
AMDHD2	17	0.9	0.516	0.047	0.054
AMY2B	18	0.895	0.515	0.042	0.05
ANLN	12	0.948	0.62	0.33	0.365
ANO10	22	0.876	0.529	0.082	0.094
ANXA1	14	0.952	0.643	0.427	0.521
ANXA5	11	0.924	0.487	-0.034	-0.037
AP1M2	13	0.924	0.492	-0.024	-0.031
APOE	13	0.924	0.492	-0.024	-0.031
AREG	11	0.948	0.586	0.246	0.28
ARIH1	18	0.886	0.484	-0.045	-0.053
ARL1	12	0.943	0.617	0.307	0.327
ARSG	11	0.938	0.538	0.108	0.124
ASS1	29	0.848	0.55	0.134	0.154
ATAD2	12	0.938	0.537	0.113	0.143
ATP5C1	23	0.886	0.593	0.24	0.26
ATP6V0B	27	0.862	0.495	-0.018	-0.038
ATP6V0E1	17	0.895	0.514	0.038	0.042
ATP6V1F	20	0.881	0.487	-0.04	-0.051
ATRAID	21	0.876	0.508	0.023	0.027
BAX	13	0.938	0.572	0.212	0.253
BCAS2	11	0.948	0.543	0.14	0.197
BCL2L1	14	0.91	0.487	-0.036	-0.042
BMP4	11	0.952	0.588	0.269	0.332
BRD3	14	0.914	0.49	-0.031	-0.037
BRMS1L	14	0.933	0.5	0	0
BTG3	13	0.924	0.528	0.079	0.089
C1D	51	0.738	0.574	0.172	0.181
C1QBP	22	0.9	0.583	0.237	0.283
CA2	16	0.924	0.586	0.241	0.274
CAPN1	14	0.91	0.487	-0.036	-0.042
CARMIL1	15	0.905	0.487	-0.037	-0.043
CAST	20	0.905	0.567	0.195	0.236
CAV1	14	0.924	0.495	-0.017	-0.026
CAV2	24	0.881	0.57	0.192	0.22
CB	14	0.933	0.5	0	0
CBLB	30	0.881	0.597	0.286	0.379
CBX4	23	0.89	0.557	0.169	0.214
CCNA2	17	0.919	0.607	0.283	0.306
CCNB1	28	0.881	0.629	0.334	0.364
CCNB2	21	0.89	0.601	0.248	0.26
CCNC	21	0.876	0.487	-0.04	-0.052
CCNE2	12	0.938	0.576	0.209	0.231
CCNG1	13	0.929	0.495	-0.017	-0.025
CCNG2	11	0.929	0.49	-0.029	-0.033
CCNH	12	0.919	0.487	-0.035	-0.038
CD24	18	0.91	0.573	0.202	0.227
CD52	11	0.938	0.495	-0.016	-0.023
CD9	21	0.881	0.489	-0.033	-0.046
CDC16	18	0.895	0.49	-0.032	-0.043
CDC20	21	0.89	0.537	0.109	0.133
CDC45	19	0.905	0.545	0.134	0.169

Gene	Active Chem.	Accuracy	Balanced Acc.	Kappa	MCC
CDCA3	12	0.948	0.659	0.396	0.411
CDCA5	29	0.848	0.521	0.061	0.079
CDCA8	16	0.91	0.521	0.061	0.073
CDK1	14	0.914	0.49	-0.031	-0.037
CDKN3	31	0.829	0.566	0.162	0.172
CEBPB	17	0.919	0.58	0.228	0.263
CENPH	24	0.881	0.57	0.192	0.22
CENPK	21	0.89	0.537	0.109	0.133
CENPU	11	0.943	0.54	0.123	0.152
CEP55	13	0.933	0.497	-0.009	-0.018
CHN1	11	0.933	0.492	-0.023	-0.028
CHST11	17	0.905	0.492	-0.025	-0.036
CIRBP	19	0.886	0.487	-0.039	-0.049
CIT	14	0.924	0.495	-0.017	-0.026
CKS2	20	0.895	0.629	0.298	0.303
CLIC4	24	0.862	0.505	0.014	0.017
CLTB	20	0.876	0.484	-0.046	-0.056
CLTC	25	0.857	0.504	0.011	0.014
CNN1	15	0.905	0.518	0.048	0.052
COG4	31	0.833	0.529	0.08	0.096
COL12A1	15	0.919	0.526	0.078	0.097
COL1A1	15	0.924	0.528	0.089	0.122
COL1A2	18	0.9	0.543	0.118	0.133
COL2A1	48	0.771	0.581	0.199	0.221
COL5A1	14	0.919	0.492	-0.024	-0.032
COPZ1	27	0.852	0.489	-0.034	-0.054
CRIM1	16	0.924	0.557	0.175	0.223
CRISP3	12	0.943	0.539	0.129	0.187
CRY1	14	0.929	0.564	0.182	0.209
CSNK1D	16	0.914	0.581	0.21	0.224
CST6	16	0.914	0.495	-0.017	-0.028
CTH	14	0.933	0.566	0.198	0.242
CXCL1	12	0.938	0.537	0.113	0.143
CXCL14	19	0.9	0.495	-0.018	-0.031
CXCL2	11	0.948	0.586	0.246	0.28
CYBRD1	17	0.91	0.575	0.198	0.215
CYP24A1	68	0.571	0.461	-0.087	-0.092
CYP26A1	11	0.952	0.588	0.269	0.332
CYP26B1	29	0.876	0.566	0.203	0.3
CYP51A1	15	0.924	0.559	0.17	0.199
DAG1	15	0.919	0.526	0.078	0.097
DCK	17	0.905	0.546	0.125	0.139
DDB2	12	0.933	0.495	-0.017	-0.024
DEK	11	0.952	0.631	0.354	0.384
DEPDC1	11	0.948	0.586	0.246	0.28
DEPDC1B	15	0.933	0.595	0.274	0.321
DLG4	23	0.881	0.552	0.145	0.169
DNAJB9	14	0.919	0.526	0.073	0.083
DPP10	12	0.957	0.625	0.386	0.489
DTL	21	0.91	0.59	0.264	0.324
DUSP1	14	0.924	0.528	0.084	0.102
DUSP22	14	0.91	0.52	0.053	0.057
ECH1	11	0.948	0.543	0.14	0.197
ECI1	18	0.89	0.537	0.097	0.103
EDEM1	15	0.948	0.664	0.454	0.507

Gene	Active Chem.	Accuracy	Balanced Acc.	Kappa	MCC
EGFL6	11	0.943	0.497	-0.009	-0.016
EIF1	14	0.943	0.638	0.375	0.413
EIF3A	46	0.762	0.535	0.091	0.108
EIF3C	16	0.924	0.557	0.175	0.223
EIF4B	15	0.933	0.564	0.203	0.278
EIF4G1	12	0.933	0.534	0.099	0.116
ELMOD1	12	0.929	0.492	-0.023	-0.03
EMX2	21	0.876	0.508	0.023	0.027
EPHA3	18	0.914	0.576	0.216	0.254
ERO1A	24	0.881	0.606	0.265	0.28
ETFA	16	0.914	0.552	0.146	0.166
ETS1	18	0.929	0.609	0.323	0.398
EXOSC5	42	0.771	0.545	0.111	0.122
FAH	22	0.881	0.492	-0.026	-0.041
FAM174A	14	0.914	0.49	-0.031	-0.037
FAM57A	24	0.857	0.484	-0.048	-0.062
FAM72A	17	0.919	0.554	0.165	0.214
FAM96A	14	0.919	0.526	0.073	0.083
FAM96B	18	0.895	0.49	-0.032	-0.043
FASN	24	0.871	0.546	0.127	0.146
FASTKD5	18	0.9	0.492	-0.025	-0.037
FBXO11	21	0.886	0.492	-0.026	-0.04
FDFT1	11	0.943	0.497	-0.009	-0.016
FDPS	17	0.9	0.516	0.047	0.054
FGF2	34	0.867	0.612	0.313	0.387
FKBP5	30	0.833	0.486	-0.043	-0.064
FLRT3	11	0.948	0.586	0.246	0.28
FOS	11	0.943	0.54	0.123	0.152
FOXA1	20	0.886	0.512	0.034	0.042
FOXO4	14	0.919	0.492	-0.024	-0.032
FTH1	17	0.924	0.61	0.3	0.334
FTL	19	0.895	0.492	-0.025	-0.038
FUT1	14	0.933	0.566	0.198	0.242
FZD2	17	0.924	0.61	0.3	0.334
G3BP2	11	0.933	0.492	-0.023	-0.028
GADD45A	26	0.876	0.583	0.224	0.255
GADD45B	17	0.919	0.554	0.165	0.214
GALNT3	24	0.876	0.531	0.092	0.118
GAPDH	19	0.886	0.534	0.09	0.097
GATA2	11	0.943	0.54	0.123	0.152
GBF1	19	0.91	0.595	0.256	0.284
GBP1	11	0.938	0.495	-0.016	-0.023
GCHFR	27	0.852	0.536	0.1	0.115
GDF15	27	0.89	0.653	0.385	0.411
GDNF	21	0.895	0.497	-0.009	-0.023
GINS2	11	0.948	0.543	0.14	0.197
GLRX	43	0.762	0.505	0.013	0.017
GMNN	11	0.943	0.54	0.123	0.152
GMPR2	17	0.905	0.519	0.056	0.068
GNAI2	15	0.929	0.592	0.255	0.285
GNB1	14	0.924	0.528	0.084	0.102
GPI	22	0.881	0.492	-0.026	-0.041
GPR87	15	0.919	0.495	-0.017	-0.027
GPR89A	18	0.914	0.525	0.084	0.145
GPR89B	15	0.924	0.559	0.17	0.199

Gene	Active Chem.	Accuracy	Balanced Acc.	Kappa	MCC
GPX4	23	0.886	0.555	0.157	0.19
GRB10	31	0.833	0.489	-0.035	-0.058
GSK3A	13	0.929	0.495	-0.017	-0.025
GTPBP8	13	0.933	0.605	0.268	0.283
H2AFX	11	0.948	0.5	0	0
HACD3	14	0.924	0.495	-0.017	-0.026
HADHA	16	0.905	0.547	0.122	0.13
HBD	13	0.924	0.528	0.079	0.089
HDAC6	27	0.838	0.497	-0.009	-0.011
HERC5	29	0.838	0.544	0.115	0.126
HERPUD1	20	0.895	0.539	0.115	0.139
HIKESHI	12	0.933	0.573	0.191	0.204
HIST1H1C	18	0.91	0.548	0.142	0.175
HIST1H1E	29	0.857	0.584	0.219	0.241
HIST1H2AC	25	0.867	0.578	0.198	0.211
HIST1H2AE	17	0.914	0.578	0.213	0.237
HIST1H2AG	33	0.824	0.55	0.131	0.147
HIST1H2AK	33	0.824	0.55	0.131	0.147
HIST1H2BB	50	0.738	0.553	0.128	0.14
HIST1H2BC	38	0.81	0.566	0.17	0.191
HIST1H2BD	32	0.843	0.561	0.168	0.198
HIST1H2BE	20	0.895	0.517	0.053	0.073
HIST1H2BF	24	0.876	0.567	0.18	0.201
HIST1H2BG	52	0.752	0.59	0.214	0.229
HIST1H2BH	23	0.9	0.601	0.282	0.328
HIST1H2BI	34	0.843	0.598	0.252	0.28
HIST1H2BL	25	0.871	0.564	0.172	0.194
HIST1H2BM	28	0.862	0.558	0.162	0.194
HIST1H2BN	28	0.857	0.555	0.151	0.175
HIST1H3B	17	0.924	0.583	0.245	0.297
HIST1H3C	18	0.905	0.57	0.188	0.206
HIST1H3D	21	0.905	0.587	0.248	0.292
HIST1H3H	48	0.733	0.527	0.066	0.073
HIST2H2AA4	26	0.876	0.583	0.224	0.255
HLA-B	16	0.91	0.492	-0.025	-0.035
HLA-E	13	0.929	0.495	-0.017	-0.025
HMGCR	12	0.943	0.578	0.228	0.266
HMMR	16	0.924	0.586	0.241	0.274
HMOX1	16	0.919	0.555	0.16	0.191
HOXA1	11	0.943	0.54	0.123	0.152
HOXA2	20	0.895	0.562	0.169	0.19
HSP90AA1	13	0.91	0.485	-0.041	-0.044
HSP90AB1	18	0.886	0.484	-0.045	-0.053
HSPA1B	18	0.905	0.545	0.129	0.152
HSPA2	14	0.919	0.492	-0.024	-0.032
HSPA4	17	0.9	0.543	0.114	0.123
HSPA5	14	0.929	0.564	0.182	0.209
HSPA9	14	0.924	0.495	-0.017	-0.026
HSPB8	14	0.933	0.566	0.198	0.242
ID1	17	0.895	0.541	0.104	0.11
ID3	22	0.886	0.515	0.046	0.066
IFIT1	11	0.933	0.492	-0.023	-0.028
IFITM1	14	0.919	0.492	-0.024	-0.032
IFRD1	18	0.924	0.581	0.249	0.331
IGFBP2	25	0.867	0.509	0.028	0.039

Gene	Active Chem.	Accuracy	Balanced Acc.	Kappa	MCC
IGFBP3	11	0.933	0.492	-0.023	-0.028
IL23A	16	0.905	0.49	-0.031	-0.04
INA	38	0.805	0.543	0.114	0.136
INHBE	16	0.919	0.555	0.16	0.191
INPP4B	12	0.943	0.578	0.228	0.266
INSIG1	17	0.914	0.524	0.078	0.111
ITGA4	71	0.614	0.523	0.05	0.053
ITGA5	12	0.938	0.576	0.209	0.231
ITPKC	13	0.952	0.687	0.477	0.503
ITPR2	12	0.943	0.578	0.228	0.266
JMJD6	14	0.919	0.492	-0.024	-0.032
KDM3A	21	0.89	0.579	0.207	0.224
KDM5A	14	0.919	0.526	0.073	0.083
KIF11	33	0.829	0.578	0.194	0.208
KIF14	19	0.919	0.624	0.335	0.371
KIF1BP	19	0.895	0.516	0.047	0.06
KIF20A	21	0.886	0.492	-0.026	-0.04
KIF23	15	0.924	0.59	0.238	0.258
KITLG	12	0.938	0.576	0.209	0.231
KLHL24	14	0.924	0.561	0.167	0.183
KRT6A	19	0.876	0.482	-0.051	-0.059
L1CAM	11	0.943	0.497	-0.009	-0.016
LAGE3	14	0.905	0.485	-0.042	-0.046
LAMP1	13	0.938	0.572	0.212	0.253
LAMP3	18	0.938	0.689	0.489	0.523
LDHA	16	0.929	0.589	0.259	0.308
LDLR	17	0.905	0.546	0.125	0.139
LMNB1	17	0.91	0.575	0.198	0.215
LOC100506621	19	0.89	0.49	-0.032	-0.044
LRP1	38	0.81	0.576	0.192	0.21
LRP10	14	0.933	0.566	0.198	0.242
LRRC41	12	0.938	0.497	-0.009	-0.017
LSM2	24	0.876	0.549	0.138	0.163
LSM5	42	0.79	0.539	0.106	0.133
LSS	16	0.919	0.555	0.16	0.191
MAD2L1	17	0.924	0.637	0.348	0.368
MAFG	16	0.929	0.617	0.316	0.347
MAL2	25	0.867	0.544	0.121	0.14
MALAT1	19	0.881	0.532	0.081	0.085
MANF	24	0.905	0.638	0.369	0.412
MAP1B	23	0.871	0.508	0.025	0.031
MAP7	23	0.89	0.538	0.12	0.174
MARCKS	30	0.819	0.506	0.015	0.017
MBNL1	18	0.89	0.512	0.034	0.038
MBTPS1	25	0.848	0.481	-0.055	-0.068
MCL1	11	0.943	0.54	0.123	0.152
MCM3	11	0.952	0.631	0.354	0.384
MCM4	16	0.929	0.617	0.316	0.347
MDH1	13	0.929	0.531	0.091	0.109
MDK	23	0.862	0.503	0.008	0.01
MDM2	21	0.89	0.579	0.207	0.224
ME1	11	0.952	0.588	0.269	0.332
MED12	28	0.829	0.478	-0.063	-0.078
MED21	28	0.871	0.593	0.251	0.285
MKI67	11	0.943	0.54	0.123	0.152

Gene	Active Chem.	Accuracy	Balanced Acc.	Kappa	MCC
MRPL12	14	0.919	0.492	-0.024	-0.032
MRPL14	32	0.829	0.527	0.076	0.092
MRPS6	13	0.933	0.569	0.195	0.219
MSI2	12	0.943	0.578	0.228	0.266
MSL2	11	0.933	0.492	-0.023	-0.028
MT2A	11	0.952	0.631	0.354	0.384
MXRA5	28	0.843	0.532	0.085	0.096
MYBL2	11	0.943	0.497	-0.009	-0.016
MYC	11	0.957	0.591	0.296	0.417
MYL9	12	0.938	0.497	-0.009	-0.017
NAA20	49	0.762	0.532	0.089	0.123
NCBP1	13	0.929	0.531	0.091	0.109
NCR3LG1	29	0.862	0.601	0.259	0.28
NEAT1	15	0.943	0.692	0.472	0.489
NEIL3	13	0.929	0.531	0.091	0.109
NFKB2	12	0.952	0.583	0.274	0.398
NFKBIA	11	0.952	0.588	0.269	0.332
NME2	22	0.886	0.535	0.103	0.128
NOTCH2	12	0.943	0.539	0.129	0.187
NPC1	20	0.89	0.514	0.044	0.056
NPRL2	13	0.938	0.644	0.35	0.362
NQO1	22	0.895	0.58	0.223	0.257
NRG1	14	0.952	0.676	0.479	0.527
NUDCD3	17	0.91	0.522	0.066	0.086
NUF2	16	0.924	0.615	0.298	0.318
NUSAP1	21	0.914	0.635	0.362	0.4
OLFML2A	23	0.886	0.555	0.157	0.19
ORC6	19	0.905	0.545	0.134	0.169
OSTC	14	0.919	0.526	0.073	0.083
PAFAH1B3	16	0.914	0.552	0.146	0.166
PANK3	17	0.9	0.49	-0.032	-0.041
PDGFB	13	0.924	0.492	-0.024	-0.031
PDHA1	23	0.881	0.495	-0.018	-0.034
PDIA6	11	0.933	0.492	-0.023	-0.028
PEBP1	14	0.924	0.561	0.167	0.183
PEG10	18	0.91	0.548	0.142	0.175
PFKFB4	17	0.919	0.58	0.228	0.263
PFN1	20	0.876	0.484	-0.046	-0.056
PGAM1	22	0.895	0.6	0.264	0.289
PGRMC1	12	0.924	0.49	-0.029	-0.034
PHF5A	12	0.929	0.492	-0.023	-0.03
PHGDH	13	0.938	0.536	0.119	0.178
PHLDB2	33	0.81	0.48	-0.058	-0.08
PIGP	28	0.838	0.529	0.076	0.084
PLAU	18	0.933	0.661	0.432	0.472
PLIN2	25	0.862	0.524	0.067	0.08
PLK1	15	0.924	0.559	0.17	0.199
PLXDC2	12	0.933	0.495	-0.017	-0.024
POLE3	13	0.924	0.528	0.079	0.089
POLR2I	11	0.943	0.54	0.123	0.152
POLR2K	38	0.829	0.608	0.273	0.298
PPCS	24	0.871	0.583	0.207	0.218
PPP1R13B	18	0.895	0.49	-0.032	-0.043
PPP1R15A	21	0.89	0.495	-0.018	-0.033
PPP2R1A	11	0.948	0.5	0	0

Gene	Active Chem.	Accuracy	Balanced Acc.	Kappa	MCC
PRAC1	19	0.905	0.545	0.134	0.169
PRC1	14	0.929	0.564	0.182	0.209
PRDX2	18	0.905	0.495	-0.017	-0.03
PRKG1	13	0.919	0.49	-0.03	-0.036
PRR7	19	0.91	0.571	0.205	0.245
PSAP	48	0.757	0.564	0.157	0.172
PSEN2	12	0.938	0.537	0.113	0.143
PSMA1	18	0.91	0.523	0.073	0.106
PSMB3	20	0.914	0.595	0.276	0.334
PSMD8	12	0.933	0.495	-0.017	-0.024
PSRC1	11	0.948	0.629	0.328	0.345
PTPN1	17	0.905	0.492	-0.025	-0.036
RAB25	11	0.943	0.583	0.225	0.244
RAB6C	36	0.819	0.505	0.016	0.029
RAD9A	13	0.933	0.497	-0.009	-0.018
RAN	11	0.938	0.495	-0.016	-0.023
RFC3	17	0.905	0.519	0.056	0.068
RGS2	31	0.838	0.545	0.123	0.143
RNF145	17	0.914	0.524	0.078	0.111
RNF19B	11	0.948	0.629	0.328	0.345
RPA3	12	0.957	0.664	0.452	0.5
RPL13A	54	0.757	0.588	0.217	0.247
RPL13AP5	19	0.886	0.487	-0.039	-0.049
RPL15	32	0.829	0.553	0.137	0.152
RPL17	33	0.819	0.511	0.03	0.038
RPL19	17	0.91	0.548	0.137	0.159
RPL22	34	0.8	0.501	0.003	0.003
RPL26	16	0.924	0.557	0.175	0.223
RPL27	28	0.824	0.49	-0.026	-0.029
RPL28	18	0.895	0.54	0.107	0.117
RPL29	20	0.886	0.534	0.094	0.105
RPL32	12	0.919	0.487	-0.035	-0.038
RPL35	25	0.857	0.504	0.011	0.014
RPL6	30	0.829	0.511	0.031	0.037
RPLP0	18	0.919	0.628	0.332	0.355
RPS18	21	0.881	0.532	0.088	0.1
RPSA	19	0.886	0.487	-0.039	-0.049
RRAGA	11	0.938	0.538	0.108	0.124
RRM2	15	0.924	0.497	-0.009	-0.019
RSL24D1	26	0.881	0.618	0.301	0.323
RUFY3	13	0.924	0.492	-0.024	-0.031
S100A10	14	0.924	0.528	0.084	0.102
S100A9	17	0.91	0.495	-0.017	-0.029
SAT1	19	0.914	0.597	0.272	0.311
SCOC	17	0.9	0.49	-0.032	-0.041
SDK2	48	0.79	0.622	0.292	0.312
SEMA3A	11	0.952	0.588	0.269	0.332
SEPT10	11	0.929	0.49	-0.029	-0.033
SERPINE2	16	0.919	0.583	0.225	0.247
SESN2	28	0.852	0.567	0.174	0.19
SGK1	11	0.943	0.54	0.123	0.152
SGO2	14	0.924	0.528	0.084	0.102
SGPP1	33	0.795	0.509	0.022	0.024
SH3BGRL3	13	0.929	0.495	-0.017	-0.025
SKP1	42	0.762	0.503	0.008	0.01

Gene	Active Chem.	Accuracy	Balanced Acc.	Kappa	MCC
SLC25A5	72	0.662	0.577	0.17	0.18
SLC27A2	11	0.952	0.588	0.269	0.332
SLC2A1	14	0.924	0.594	0.236	0.246
SLC2A6	12	0.929	0.532	0.087	0.096
SLC35B1	41	0.8	0.562	0.162	0.19
SLC3A2	13	0.938	0.572	0.212	0.253
SLC7A5	16	0.924	0.586	0.241	0.274
SLIT2	12	0.933	0.534	0.099	0.116
SNAI2	17	0.905	0.492	-0.025	-0.036
SNCA	12	0.933	0.534	0.099	0.116
SNRPG	16	0.933	0.649	0.385	0.412
SNX6	23	0.89	0.538	0.12	0.174
SPAG5	13	0.933	0.533	0.104	0.136
SPEN	18	0.91	0.598	0.254	0.271
SPP1	15	0.91	0.521	0.057	0.063
SPRR1B	20	0.881	0.509	0.026	0.03
SPTAN1	19	0.905	0.568	0.191	0.219
SRSF10	18	0.895	0.49	-0.032	-0.043
SSB	27	0.862	0.558	0.158	0.181
ST3GAL1	20	0.91	0.592	0.26	0.301
STC1	17	0.9	0.49	-0.032	-0.041
STMN1	25	0.867	0.544	0.121	0.14
STXBP1	51	0.686	0.459	-0.105	-0.127
SUMO1	13	0.933	0.569	0.195	0.219
SYNE2	13	0.919	0.49	-0.03	-0.036
TAGLN	12	0.943	0.5	0	0
TCEAL4	16	0.91	0.492	-0.025	-0.035
TGFB2	15	0.914	0.523	0.067	0.078
TIMM23B	14	0.914	0.49	-0.031	-0.037
TIMP3	13	0.943	0.61	0.31	0.349
TINAGL1	31	0.81	0.488	-0.032	-0.038
TJP1	52	0.719	0.504	0.01	0.012
TK1	13	0.929	0.495	-0.017	-0.025
TMED10	13	0.929	0.567	0.178	0.193
TMEM126A	11	0.943	0.583	0.225	0.244
TMEM128	17	0.871	0.474	-0.064	-0.066
TNC	27	0.867	0.545	0.13	0.166
TNFRSF10B	14	0.952	0.676	0.479	0.527
TNFRSF12A	11	0.943	0.497	-0.009	-0.016
TNFRSF21	14	0.924	0.495	-0.017	-0.026
TOP2A	17	0.924	0.61	0.3	0.334
TP53BP1	17	0.9	0.516	0.047	0.054
TPBG	17	0.914	0.524	0.078	0.111
TPI1	14	0.919	0.492	-0.024	-0.032
TPPP3	28	0.848	0.489	-0.034	-0.055
TPX2	39	0.8	0.521	0.06	0.08
TRIM22	15	0.914	0.523	0.067	0.078
TSC22D3	15	0.924	0.528	0.089	0.122
TSKU	17	0.89	0.484	-0.044	-0.051
TST	12	0.919	0.487	-0.035	-0.038
TTK	11	0.943	0.497	-0.009	-0.016
TUBA1B	11	0.938	0.495	-0.016	-0.023
TUBB2B	13	0.933	0.569	0.195	0.219
TUBB6	14	0.938	0.602	0.291	0.334
TUBG1	13	0.938	0.536	0.119	0.178

Gene	Active Chem.	Accuracy	Balanced Acc.	Kappa	MCC
TWIST1	15	0.933	0.564	0.203	0.278
TXNDC5	12	0.948	0.659	0.396	0.411
TYMS	16	0.929	0.589	0.259	0.308
UBA3	21	0.886	0.534	0.098	0.115
UBA52	11	0.938	0.538	0.108	0.124
UBE2L6	16	0.905	0.49	-0.031	-0.04
UBE2T	19	0.919	0.624	0.335	0.371
UBE2V2	28	0.876	0.596	0.264	0.307
UGDH	13	0.929	0.531	0.091	0.109
UQCRC1	14	0.914	0.49	-0.031	-0.037
UQCRQ	26	0.843	0.481	-0.055	-0.07
USMG5	61	0.69	0.559	0.136	0.146
USO1	51	0.719	0.521	0.053	0.06
USP7	13	0.91	0.485	-0.041	-0.044
VDAC2	25	0.852	0.518	0.05	0.056
VEGFA	11	0.957	0.634	0.383	0.436
VGLL1	12	0.943	0.539	0.129	0.187
VPS72	12	0.948	0.62	0.33	0.365
WARS	14	0.938	0.569	0.217	0.29
WASF3	18	0.89	0.487	-0.039	-0.048
WDR61	19	0.89	0.49	-0.032	-0.044
WFDC2	14	0.924	0.495	-0.017	-0.026
WNT4	26	0.857	0.522	0.063	0.076
ZBTB41	16	0.919	0.583	0.225	0.247
ZNF274	16	0.9	0.487	-0.038	-0.045
ZNF395	12	0.962	0.706	0.538	0.573
ZWINT	11	0.948	0.5	0	0

References

[1] C. Hansch and T. Fujita, "p-σ-π analysis: A method for the correlation of biological activity and chemical structure," *J. Am. Chem. Soc,* vol. 86, p. 1616–1626, 1964.

[2] A. K. Ghose and W. J. Crippen, "Atomic physicochemical parameters for three-dimensional structure-directed quantitative structure-activity relationships," *J. Comput. Chem,* vol. 7, no. 6, p. 565–577, 1986.

[3] G. Gini, "QSAR methods," in *Methods in Molecular Biology*, vol. 1425, 2016, p. 1–20.

[4] A. Cherkasov et al., "QSAR modeling: Where have you been? Where are you going to?," *J. Med. Chem,* vol. 57, no. 12, p. 4977–5010, 2014.

[5] C. Nantasenamat et al., "A practical overview of quantitative structure-activity relationship," *EXCLI J,* vol. 8, p. 74–88, 2009.

[6] K. Todeschini and V. Consonni, Molecular Descriptors for Chemoinformatics, Weinheim, Germany: Wiley-VCH, 2009.

[7] Kode, *DRAGON 7.0.8 Software,* 2017.

[8] D. Gadaleta et al., "Integrating computational methods to predict mutagenicity of aromatic azo compounds," *J. Environ. Sci. Health C,* vol. 35, no. 1, p. 139–162, 2017.

[9] K. Roy, P. Ambure, S. Kar and O. P. K, "Is it possible to improve the quality of predictions from an 'intelligent' use of multiple QSAR/QSPR/QSTR models?," *J. Chemom.,* vol. 32, no. 12, 2018.

[10] C. Zhao and E. Benfenati et al., "A new hybrid system of QSAR models for predicting bioconcentration factors (BCF)," *Chemosphere,* vol. 73, no. 11, p. 1701–1707, 2008.

[11] M. Hewitt et al., "Consensus QSAR models. Do the benefits outweigh the complexity?," *J. Chem. Inf. Model,* vol. 47, no. 4, p. 1460–1468, 2007.

[12] N. C. Kleinstreuer et al., "Predictive models for acute oral systemic toxicity: A workshop to bridge the gap from research to regulation," *Comput. Toxicol,* vol. 8, p. 21–24, 2018.

[13] N. A. Fonseca, J. Rung, A. Brazma and J. C. Marioni, "Tools for mapping high-throughput sequencing data," *Bioinformatics,* vol. 28, no. 24, p. 3169–3177, 2012.

[14] D. Gadaleta, A. Lombardo, C. Toma and E. Benfenati, "A new semi-automated workflow for chemical data retrieval and quality checking for modeling applications," *J. Cheminform,* vol. 10, 2018.

[15] D. Ballabio, F. Grisoni and R. Todeschini, "Multivariate comparison of classification performance measures," *Chemom. Intell. Lab. Syst.,* vol. 174, pp. 33-44, 2018.

[16] J. Dong et al., "ChemSAR: an online pipelining platform for molecular SAR modeling," *J. Cheminform.,* vol. 9, 2017.

[17] O. Soufan et al., "DPubChem: a web tool for QSAR modeling and high-throughput virtual screening," *Sci. Rep.,* vol. 8, 2018.

[18] N. M. O'Boyle et al., "Open Babel: an open chemical toolbox," *J. Cheminform.,* vol. 3, 2011.

[19] N. W. Chawla, K. W. Bowyer, L. O. Hall and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *J. Artif. Intell. Res.,* vol. 16, no. 1, p. 321–357, 2002.

[20] M. T. D. Cronin et al., "Making in silico predictive models for toxicology FAIR," *Regulatory Toxicology and Pharmacology,* vol. 140, 2023.

[21] A. Mayr, G. Klambauer, T. Unterthiner and S. Hochreiter, "DeepTox: toxicity prediction using deep learning," *Front. Environ. Sci.,* vol. 3, 2016.

[22] X. Menghang and R. Huang, "Editorial: Tox21 challenge to build predictive models of nuclear receptor and stress response pathways as mediated by exposure to environmental toxicants and drugs," *Front. Environ. Sci.,* vol. 5, 2017.

[23] D. Gunning, "Explainable artificial intelligence (XAI)," U.S. Defense Advanced Research Projects Agency, [Online]. Available: https://www.darpa.mil/program/explainable-artificial-intelligence. [Accessed 11 01 2026].

[24] W. M. S. Russell and R. L. Burch, The Principles of Humane Experimental Technique, London: Methuen, 1959.

[25] E. Commission, "Regulation (EC) No 1907/2006 of the European Parliament and of the Council of 18 December 2006 concerning the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH)," *Official Journal of the European Union,* vol. L396, p. 1–849, December 2006.

[26] I. C. Harmonisation, "Assessment and Control of DNA Reactive Mutagenic Impurities in Pharmaceuticals to Limit Potential Carcinogenic Risk," 2017. [Online]. Available: https://database.ich.org/sites/default/files/M7R1Guideline.pdf. [Accessed 11 01 2026].

[27] OECD, Guidance Document on the Validation of (Quantitative) Structure-Activity Relationship [(Q)SAR] Models, Paris: OECD Publishing, 2007.

[28] ECA, "Chapter R.6: QSARs and grouping of chemicals," in *Guidance on information requirements and chemical safety assessment.*, Helsinki, European Chemicals Agency, 2008.

[29] EFSA, "Guidance on the use of the weight of evidence approach in scientific assessments," *EFSA J.,* vol. 15, no. 8, 2017.

[30] A. Amberg et al., "Principles and procedures for handling out-of-domain and indeterminate results as part of ICH M7 recommended (Q)SAR analyses," *Regulatory Toxicology and Pharmacology,* vol. 102, p. 53–64, 2019.

[31] R. S. Foster et al., "The importance of expert review to clarify ambiguous situations for (Q)SAR predictions under ICH M7," *Genes Environ.,* 2020.

[32] L. C. Hasselgren et al., "Management of pharmaceutical ICH M7 (Q)SAR predictions – The impact of model updates," *Regulatory Toxicology and Pharmacology,* vol. 118, 2020.

[33] T. Hanser et al., "Applicability Domain: Towards a More Formal Definition," *SAR and QSAR in Environmental Research,* vol. 27, no. 11, pp. 893-909, 2016.

[34] E. Benfenati, A. Manganaro and G. Gini, "VEGA-QSAR: AI inside a platform for predictive toxicology," *CEUR Workshop Proc,* vol. 1107, p. 21–28, 2013.

[35] T. Hastie, R. Tibshirani and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer, 2009.

[36] P. Domingos, "A few useful things to know about machine learning," *Communications of the ACM,* vol. 55, no. 10, p. 78–87, 2012.

[37] R. Tibshirani, "Regression shrinkage and selection via the Lasso.," *Journal of the Royal Statistical Society. Series B,* vol. 58, no. 1, p. 267–288, 1996.

[38] J. Friedman, T. Hastie and R. Tibshirani, "Regularization paths for generalized linear models via coordinate descent," *Journal of Statistical Software,* vol. 33, no. 1, p. 1–22, 2010.

[39] R. E. Schapire and Y. Freund, Boosting: Foundations and Algorithms, Cambridge, MA: MIT Press, 2012.

[40] B. Hofner, A. Mayr, N. Robinzonov and M. Schmid, "Model-based boosting in R: A hands-on tutorial using the R package mboost," *Comput. Stat,* vol. 29, pp. 3–35,, 2014.

[41] J. Friedman, T. Hastie and R. Tibshirani, "Additive logistic regression: A statistical view of boosting," *Ann. Stat,* vol. 28, pp. 337–407,, 2000.

[42] L. Breiman, J. H. Friedman, R. A. Olshen and C. J. Stone, Classification and Regression Trees, Chapman and Hall/CRC, 1984.

[43] L. Breiman, "Random forests.," *Mach. Learn.,* vol. 45, no. 1, p. 5–32, 2001.

[44] P. Geurts, D. Ernst and L. Wehenkel, "Extremely randomized trees," *Mach. Learn,* vol. 63, pp. 3–42,, 2006.

[45] V. N. Vapnik, The Nature of Statistical Learning Theory, Springer, 2000.

[46] J. S. Marron, M. J. Todd and J. Ahn, "Distance-weighted discrimination," *Journal of the American Statistical Association,* vol. 102, pp. 1267–1271,, 2012.

[47] B. Wang and H. Zou, "Sparse Distance Weighted Discrimination," *Journal of Computational and Graphical Statistics,* vol. 25, no. 3, p. 826–838, 2016.

[48] N. Landwehr, M. Hall and E. Frank, "Logistic model trees," *Machine Learning,* vol. 59, pp. 161–205,, 2005.

[49] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Networks,* vol. 61, p. 85–117, 2015.

[50] Y. LeCun, Y. Bengio and G. Hinton, "Deep learning," *Nature,* vol. 521, p. 436–444, 2015.

[51] R. Hecht-Nielsen, "Counterpropagation networks," *Applied Optics,* vol. 26, no. 23, p. 4979–4983, 1987.

[52] J. Zupan, M. Novič and I. Ruisanchez, "Kohonen and counterpropagation artificial neural networks in analytical chemistry," *Chemometrics and Intelligent Laboratory Systems,* vol. 38, no. 1, p. 1–23, 1997.

[53] V. Drgan, K. Venko, J. Sluga and M. Novič, "Merging counter-propagation and back-propagation algorithms: Overcoming the limitations of counter-propagation neural network models," *Int. J. Mol. Sci.,* vol. 25, no. 8, 2024.

[54] Enamine, "Enamine Screening Collection," Enamine, [Online]. Available: https://enamine.net/compound-collections/screening-collection. [Accessed 11 01 2026].

[55] D. Weininger, A. Weininger and J. L. Weininger, "SMILES. 2. Algorithm for generation of unique SMILES notation," *Journal of Chemical Information and Computer Sciences,* vol. 29, no. 2, pp. 97-101, 1989.

[56] D. C. I. Systems, Daylight Theory Manual, 2008.

[57] P. Willett, "Similarity-based virtual screening using 2D fingerprints," *Drug Discovery Today,* vol. 11, pp. 1046-1053, 2006.

[58] PubChem, "PubChem Substructure Fingerprint," [Online]. Available: https://ftp.ncbi.nlm.nih.gov/pubchem/specifications/pubchem_fingerprints.txt. [Accessed 11 01 2026].

[59] NIH, "PubChem," [Online]. Available: https://pubchem.ncbi.nlm.nih.gov/. [Accessed 11 01 2026].

[60] J.-L. Faulon, D. P. Visco and R. S. Pophale, "The signature molecular descriptor. 1. Using extended valence sequences in QSAR and QSPR studies," *Journal of Chemical Information and Computer Sciences,* vol. 43, pp. 707–720,, 2003.

[61] Chemaxon, "Extended Connectivity Fingerprint (ECFP)," [Online]. Available: https://docs.chemaxon.com/display/docs/fingerprints_extended-connectivity-fingerprint-ecfp.md. [Accessed 11 01 2026].

[62] J. L. Durant et al., "Reoptimization of MDL Keys for Use in Drug Discovery," *Journal of Chemical Information and Computer Sciences,* vol. 42, no. 6, 2002.

[63] R. Nilakantan et al., "Topological torsion: a new molecular descriptor for SAR applications. Comparison with other descriptors," *Journal of Chemical Information and Computer Sciences,* vol. 27, no. 2, 1987.

[64] RDKit, "RDKit: Open-Source Cheminformatics Software," [Online]. Available: https://www.rdkit.org. [Accessed 11 01 2026].

[65] J. Klekota and F. P. Roth, "Chemical substructures that enrich for biological activity," *Bioinformatics,* vol. 24, no. 21, pp. 2518-2525, 2008.

[66] D. Probst and J.-L. Reymond, "A probabilistic molecular fingerprint for big data settings," *Journal of Cheminformatics ,* vol. 10, 2018.

[67] A. Capecchi, D. Probst and J.-L. Reymond, "One molecular fingerprint to rule them all: Drugs, biomolecules, and the metabolome," *Journal of Cheminformatics,* vol. 12, 2020.

[68] A. Manganaro et al., "A generalizable definition of chemical similarity for read-across," *Journal of Cheminformatics,* vol. 6, 2014.

[69] M. Bajusz, A. Rácz and K. Héberger, "Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations?," *Journal of Cheminformatics,* vol. 7, p. 1–13, 2015.

[70] Chemaxon, "Maximum Common Substructure (MCS) search," [Online]. Available: https://docs.chemaxon.com/display/docs/jchem-base_maximum-common-substructure-mcs-search.md. [Accessed 11 01 2026].

[71] C. M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006.

[72] O. Rainio, J. Teuho and R. Klén, "Evaluation metrics and statistical tests for machine learning.," *Scientific Reports,* vol. 14, 2024.

[73] Wikipedia, "Confusion matrix," [Online]. Available: https://en.wikipedia.org/wiki/Confusion_matrix. [Accessed 11 01 2026].

[74] K. H. Brodersen, C. S. Ong, K. E. Stephan and J. M. Buhmann, "The balanced accuracy and its posterior distribution," in *20th International Conference on Pattern Recognition*, 2010.

[75] B. W. Matthews, "Comparison of predicted and observed secondary structure of T4 phage lysozyme," *Biochimica et Biophysica Acta,* vol. 405, no. 2, p. 442–451, 1975.

[76] D. M. Powers, "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation," *Journal of Machine Learning Technologies,* vol. 2, no. 1, p. 37–63, 2008.

[77] P. Stoica and P. Babu, "Pearson–Matthews correlation coefficients for binary and multinary classification," *Signal Processing,* vol. 222, 2024.

[78] J. Gorodkin, "Comparing two K-category assignments by a K-category correlation coefficient," *Computational Biology and Chemistry,* vol. 28, no. 5, pp. 367-374, 2004.

[79] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and Psychological Measurement,* vol. 20, no. 1, p. 37–46, 1960.

[80] C. J. Willmott and K. Matsuura, "Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance," *Climate Research,* vol. 30, no. 1, pp. 79-82, 2005.

[81] R. L. Wasserstein and N. A. Lazar, "The American Statistician statement on statistical significance and p-values," *The American Statistician,* vol. 70, no. 2, pp. 129-133, 2016.

[82] H. B. Mann and D. R. Whitney, "On a test of whether one of two random variables is stochastically larger than the other," *Annals of Mathematical Statistics,* vol. 18, no. 1, pp. 50-60, 1947.

[83] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics Bulletin,* vol. 1, no. 6, p. 80–83, 1945.

[84] E. Brunner and U. Munzel, "The nonparametric Behrens-Fisher problem: asymptotic theory and a small-sample approximation," *Biometrical Journal,* vol. 42, no. 1, p. 17–25, 2000.

[85] S. Greenland et al., "Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations," *European Journal of Epidemiology,* vol. 31, p. 337–350, 2016.

[86] R. A. Fisher, The Design of Experiments, London: Oliver and Boyd, 1935.

[87] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society Series B,* vol. 57, no. 1, p. 289–300, 1995.

[88] F. H. C. Crick, "On protein synthesis," *Symposia of the Society for Experimental Biology,* vol. 12, p. 139–163, 1958.

[89] E. S. Lander et al., "Initial sequencing and analysis of the human genome," *Nature,* vol. 409, no. 6822, pp. 860-921, 2001.

[90] B. E. Bernstein, A. Meissner and E. S. Lander, "The mammalian epigenome," *Cell,* vol. 128, no. 4, p. 669–681, 2007.

[91] F. Sanger and A. R. Coulson, "A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase," *Journal of Molecular Biology,* vol. 94, no. 3, p. 441–448, 1975.

[92] J. Shendure et al., "DNA sequencing at 40: past, present, and future," *Nature,* vol. 550, p. 345–353, 2017.

[93] M. L. Metzker, "Sequencing technologies—the next generation," *Nature reviews genetics,* vol. 11, pp. 31-46, 2010.

[94] M. Ronaghi, M. Uhlén and P. Nyren, "A sequencing method based on real-time pyrophosphate," *Science,* vol. 281, 1998.

[95] J. H. Leamon et al., "A massively parallel PicoTiterPlate based platform for discrete picoliter-scale polymerase chain reactions," *Electrophoresis,* vol. 24, pp. 3769-3777, 2003.

[96] K. Vukovic, "Amino acid variation in human proteome," Faculty of Science, University of Zagreb, Master's thesis, Zagreb, 2016.

[97] D. R. Bentley et al., "Accurate whole human genome sequencing using reversible terminator chemistry," *Nature,* vol. 456, pp. 53-59, 2008.

[98] A. Rhoads and K. F. Au, "PacBio sequencing and its applications," *Genomics, Proteomics & Bioinformatics,* vol. 13, no. 5, p. 278–289, 2015.

[99] M. J. Levene et al., "Zero-mode waveguides for single-molecule analysis at high concentrations," *Science,* vol. 299, no. 5607, pp. 682-686, 2003.

[100] C.-S. Chin et al., "Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data," *Nature Methods,* vol. 10, p. 563–569, 2013.

[101] J. Eid et al., "Real-time DNA sequencing from single polymerase molecules," *Science,* vol. 323, no. 5910, pp. 133-138, 2009.

[102] M. Jain et al., "Nanopore sequencing and assembly of a human genome with ultra-long reads," *Nature Biotechnology,* vol. 36, no. 4, p. 338–345, 2018.

[103] N. Loman, "How a small backpack for fast genomic sequencing is helping combat Ebola," [Online]. Available: https://theconversation.com/how-a-small-backpack-for-fast-genomic-sequencing-is-helping-combat-ebola-41863. [Accessed 11 01 2026].

[104] A. Oshlack, M. D. Robinson and M. D. Young, "From RNA-seq reads to differential expression results," *Genome Biology,* vol. 11, no. 12, 2010.

[105] BioSpyder, "TempO-Seq White Paper: Sensitive Gene Expression Profiling with TempO-Seq: Advantages and Applications".

[106] C. J. Trejo et al., "Extraction-free whole transcriptome gene expression analysis of FFPE sections and histology-directed subareas of tissue," *PLoS One,* vol. 14, no. 2, 2019.

[107] J. M. Yeakley et al., "A trichostatin A expression signature identified by TempO-Seq targeted whole transcriptome profiling," *PLoS One,* vol. 12, no. 5, 2017.

[108] P. R. Bushel, R. S. Paules and S. S. Auerbach, "A Comparison of the TempO-Seq S1500+ Platform to RNA-Seq and Microarray Using Rat Liver Mode of Action Samples," *Frontiers in Genetics,* vol. 9, 2018.

[109] T. M. Martin, User's guide for T.E.S.T. (Toxicity Estimation Software Tool) version 5.1, Cincinnati, OH: U.S. Environmental Protection Agency, 2021..

[110] S. Zhang et al., "A Novel Automated Lazy Learning QSAR (ALL-QSAR) Approach: Method Development, Applications, and Virtual Screening of Chemical Databases Using Validated ALL-QSAR Models," *Journal of Chemical Information and Modeling,* vol. 46, no. 5, p. 1836–1847, 2006.

[111] A. Maunz et al., "Lazar: A modular predictive toxicology framework," *Frontiers in Pharmacology,* vol. 4, 2013.

[112] C. L. Russom, S. P. Bradbury, S. J. Broderius, D. E. Hammermeister and R. A. Drummond, "Predicting modes of toxic action from chemical structure: Acute toxicity in the fathead minnow," *Environmental Toxicology and Chemistry,* vol. 16, no. 5, p. 948–967, 1997.

[113] K. Vuković, D. Gadaleta and E. Benfenati, "Methodology of aiQSAR: a group-specific approach to QSAR modelling," *Journal of Cheminformatics,* vol. 11, no. 1, 2019.

[114] T. Cosimo, "VegaJavaR," [Online]. Available: https://github.com/marcabrus/VegaJavaR. [Accessed 11 01 2026].

[115] R. Guha, "Chemical informatics functionality in R," *Journal of Statistical Software,* vol. 18, no. 5, pp. 1-16, 2007.

[116] M. Kuhn, "Building predictive models in R using the caret package," *Journal of Statistical Software,* vol. 28, no. 5, pp. 1-26, 2008.

[117] B. Hofner, A. Mayr, N. Robinzonov and M. Schmid, "Model-based boosting in R: a hands-on tutorial using the R package mboost," *Computational Statistics,* vol. 29, no. 1, pp. 3-35, 2014.

[118] A. Karatzoglou, A. Smola, K. Hornik and A. Zeileis, "Kernlab—An S4 package for kernel methods in R," *Journal of Statistical Software,* vol. 11, no. 9, pp. 1-20, 2004.

[119] A. Liaw and M. Wiener, "Classification and Regression by Randomforest," *R News,* vol. 2, pp. 18-22, 2002.

[120] J. Simm, I. de Abril and M. Sugiyama, "Tree-based ensemble multi-task learning method for classification and regression," *IEICE Transactions on Information and Systems,* vol. 97, no. 6, p. 1677–1681, 2014.

[121] M. Dietze, "Package "caTools"," [Online]. Available: https://cran.r-project.org/web/packages/caTools/caTools.pdf. [Accessed 11 01 2026].

[122] D. Meyer et al., "Package "e1071"," [Online]. Available: https://cran.r-project.org/web/packages/e1071/e1071.pdf. [Accessed 11 01 2026].

[123] K. Hornik, C. Buchta and A. Zeileis, "Open-source machine learning: R meets Weka," *Computational Statistics,* vol. 24, no. 2, pp. 225-232, 2009.

[124] B. Wang, "Package "sdwd"," [Online]. Available: https://cran.r-project.org/web/packages/sdwd/sdwd.pdf. [Accessed 11 01 2026].

[125] NTP, "Predictive models for acute oral systemic toxicity," ICCVAM Acute Toxicity Workgroup, National Institute of Environmental Health Sciences, [Online]. Available: https://ntp.niehs.nih.gov/whatwestudy/niceatm/test-method-evaluations/acute-systemic-tox/models. [Accessed 11 01 2026].

[126] E. Walum, "Acute oral toxicity," *Environmental Health Perspectives,* vol. 106, pp. 497-503, 1998.

[127] K. Mansouri and K. Vukovic et al., "CATMoS: Collaborative Acute Toxicity Modeling Suite," *Environmental Health Perspectives,* vol. 129, no. 4, 2021.

[128] D. Gadaleta, K. Vuković, C. Toma, G. J. Lavado, A. L. Karmaus, K. Mansouri, N. C. Kleinstreuer, E. Benfenati and A. Roncaglioni, "SAR and QSAR modeling of a large collection of LD50 rat acute oral toxicity data," *Journal of Cheminformatics,* vol. 11, no. 1, 2019.

[129] U. N. A. a. Records, "Toxicity category," Office of the Federal Register, [Online]. Available: https://www.ecfr.gov/current/title-40/chapter-I/subchapter-E/part-156/subpart-D/section-156.62. [Accessed 11 01 2026].

[130] K. Vukovic, "aiQSAR," VEGA HUB, [Online]. Available: https://www.vegahub.eu/portfolio-item/aiqsar/. [Accessed 11 01 2026].

[131] EPA, "Tox21 Data Challenge," National Center for Advancing Translational Sciences, [Online]. Available: https://tripod.nih.gov/tox21/challenge/index.jsp. [Accessed 11 01 2026].

[132] L. Mora Lagares, N. Minovski and M. Novic, "Multiclass classifier for P-glycoprotein substrates, inhibitors, and non-active compounds," *Molecules,* vol. 24, no. 10, 2019.

[133] N.-N. Wang et al., "Predicting human intestinal absorption with modified random forest approach: a comprehensive evaluation of molecular representation, unbalanced data, and applicability domain issues," *RSC Advances,* vol. 7, 2017.

[134] I. Ezkurdia et al., "Multiple evidence strands suggest that there may be as few as 19,000 human protein-coding genes.," *Human Molecular Genetics,* vol. 23, no. 22, p. 5866–5878, 2014.

[135] M. Seo et al., "Computational modeling approaches for developing a synergistic effect prediction model of estrogen agonistic activity," *Chemosphere,* vol. 349, 2024.

[136] T. Uehara et al., "Prediction model of potential hepatocarcinogenicity of rat hepatocarcinogens using a large-scale toxicogenomics database," *Toxicology and Applied Pharmacology,* vol. 255, no. 3, pp. 297-306, 2011.

[137] GeneCards, "GeneCards: The Human Gene Database," [Online]. Available: https://www.genecards.org. [Accessed 11 01 2026].

[138] C. M. Clemson et al., "An architectural role for a nuclear noncoding RNA: NEAT1 RNA is essential for the structure of paraspeckles," *Molecular Cell,* vol. 33, no. 6, pp. 717-726, 2009.

[139] Z. Mazidi, M. Wieser, N. Spinu, A. Weidinger, A. V. Kozlov, K. Vukovic, S. Wellens, C. Murphy, P. Singh, L. Mora Lagares, M. R. Bobbili, L. Liendl, M. Schosserer and A. Diendorfer, "Cyclosporin A toxicity on endothelial cells differentiated from induced pluripotent stem cells: Assembling an adverse outcome pathway," *Toxicology in Vitro,* vol. 103, 2025.

Bibliography

Journal Articles

K. Vuković, D. Gadaleta and E. Benfenati, "Methodology of aiQSAR: a group-specific approach toQSAR modelling," *Journal of Cheminformatics,* vol. 11, no. 1, 2019.

K. Mansouri, K. Vukovic et al., "CATMoS: Collaborative Acute Toxicity Modeling Suite," *Environmental Health Perspectives*, vol. 129, no. 4, 2021.

D. Gadaleta, K. Vuković, C. Toma, G. J. Lavado, A. L. Karmaus, K. Mansouri, N. C. Kleinstreuer, E. Benfenati and A. Roncaglioni, "SAR and QSAR modeling of a large collection of LD50 rat acute oral toxicity data," *Journal of Cheminformatics,* vol. 11, no. 1, 2019.

Z. Mazidi, M. Wieser, N. Spinu, A. Weidinger, A. V. Kozlov, K. Vukovic, S. Wellens, C. Murphy, P. Singh, L. Mora Lagares, M. R. Bobbili, L. Liendl, M. Schosserer and A. Diendorfer, "Cyclosporin A toxicity on endothelial cells differentiated from induced pluripotent stem cells: Assembling an adverse outcome pathway," *Toxicology in Vitro,* vol. 103, 2025.

Software

K. Vukovic, "aiQSAR," VEGA HUB, [Online]. Available: https://www.vegahub.eu/portfolio-item/aiqsar/. [Accessed 11 01 2026].

Biography

Kristijan Vuković received his Bachelor's degree in Chemistry (2014) and Master's degree in Biochemistry and Physical Chemistry (2016, *summa cum laude*) from the Department of Chemistry, Faculty of Science, University of Zagreb, Croatia, where he was a recipient of the City of Zagreb scholarship (2011-2016). He completed a supplementary Computational Biology module (2015-2016) at the Department of Molecular Biology of the same institution. In 2016, he was awarded the Australian Government's Endeavour Research Fellowship for research at the Centre for Systems Genomics, University of Melbourne, under Prof. Michael Inouye.

From 2017 through 2020, he worked as a researcher at Mario Negri Institute for Pharmacological Research in Milan, Italy, group of Prof. Emilio Benfenati and at the National Institute of Chemistry in Ljubljana, Slovenia, group of Prof. Marjana Novič, supported by a Marie Skłodowska-Curie Actions ITN Fellowship. In 2018, he received the Best Oral Presentation Award at the 18th International Conference on QSAR in Environmental and Health Sciences in Bled, Slovenia. Since 2021, he has worked in research consulting and ML/AI development and served as Principal Investigator of a UK Biobank-associated project on predicting adverse outcomes from genomic data using machine learning models and localization.