Quantitative Structure Activity Relationships: An overview

Prachi Pradeep

Oak Ridge Institute for Science and Education Research Participant

National Center for Computational Toxicology

U.S. Environmental Protection Agency

Motivation: Current status and prospects of QSAR Modeling in Medical Devices Community

Toxicological and clinical computational analysis and the US FDA/CDER.

Review article

Benz RD. Expert Opin Drug Metab Toxicol. 2007.

Authors

Benz RD[1].

Author information

[1] US Food and Drug Administration, Office of Pharmaceutical Science, Center for Drug Evaluation and Research, 10903 New Hampshire Ave., Silver Spring, MD 20993, USA. R.Daniel.Benz@fda.hhs.gov

Use of QSAR Modeling to Predict the Carcinogenicity of Color Additives

Ronald Brown, Shannon White, Jennifer Goode, Prachi Pradeep and Stephen Merrill

[+] Author Affiliations

Paper No. FMD2013-16161, pp. V001T10A044; 2 pages
doi:10.1115/FMD2013-16161

From: ASME 2013 Conference on Frontiers in Medical Devices: Applications of Computer Modeling and Simulation

Altern Lab Anim. 2009 Nov;37(5):523-31.

Computational toxicology approaches at the US Food and Drug Administration.

Yang C[1], Valerio LG Jr, Arvidson KB.

Author information

1   Office of Food Additive Safety, Center for Food Safety and Applied Nutrition, US Food and Drug Administration, College Park, MD, USA. chihae.yang@fda.hhs.gov

Development of nanotoxicology: implications for drug delivery and medical devices

Sourav Bhattacharjee & David J Brayden

Published Online: 26 Jun 2015
Doi: https://doi.org/10.2217/nnm.15.69

QSAR: Definition

Structure-Activity Relationship (SAR) is an approach to find qualitative relationships between chemical structure and their biological activity

Quantitative Structure Activity Relationship (QSAR) models are theoretical models that relate a quantitative measure of chemical structure to a physical property, or a biological activity

Principle: Structurally similar chemicals are likely to have similar physicochemical and biological properties
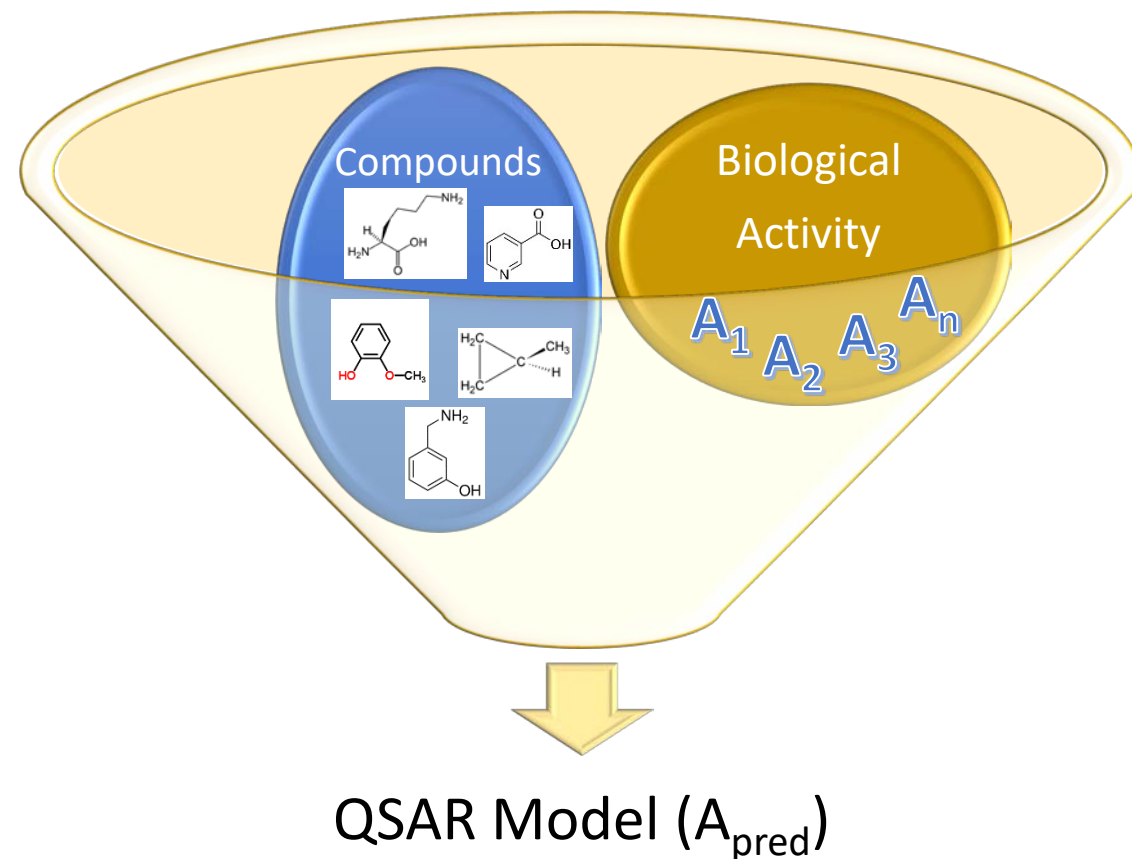
QSAR models are of the form:

$A_{pred} = f(D_1, D_2, \ldots D_n)$

where,

A_{pred}: biological activity (or toxicological endpoint)

$D_1, D_2, \ldots D_n$: chemical or structural properties (molecular descriptors)

$A_1, A_2, \ldots A_n$: biological activity of training chemicals

QSAR Model (A_{pred})

QSAR: Tools

A number of free and proprietary (Q)SAR tools are available that can predict the toxicity of a given chemical based on its chemical structure

	QSAR TOOLS		
	Expert Systems/ Rule-based (SARs)	**Statistical model based (QSARs)**	**Hybrid**
Underlying Algorithm	• Structural Alerts (SA) • Expert Judgment	• Mathematical models • Data Mining • Machine Learning	• Rule-based • Statistical modeling
Application	• Toxic endpoints with known mechanism of action • Less training (chemical) data	• Toxic endpoints with little or no knowledge of mechanism of action • Significant training (chemical) data	Combines the best features of rule-based and statistical methods • Mechanistic interpretation • High accuracy
Example	**Freely available** • Toxtree **Commercial** • Derek Nexus	**Freely available** • EPA T.E.S.T • VEGA • LAZAR **Commercial** • MultiCASE	**Commercial** • TIMES • Catalogic

QSAR: Tools Review

Review of Software Tools for Toxicity Prediction

Mojca Fuart Gatnik and Andrew Worth

EUR 24489 EN · 2010

SOFTWARE (AND DEVELOPER)	AVAILABILITY	Acute oral toxicity	Repeat dose (chronic) oral toxicity	Genotoxicity (including mutagenicity)	Carcinogenicity	Reproductive (including developmental) toxicity	Endocrine activity / disruption	Hepatotoxicity	Nephrotoxicity (+ urinary tract toxicity)	Neurotoxicity	Cytotoxicity	Immunotoxicity (3)
ACD/Tox Suite (ToxBoxes)	Commercial	•		•			•					
ADMET Predictor (Simulations Plus Inc.)	Commercial		• (1)	•	•		•	•				
BioEpisteme	Commercial				•			•	•			
Caesar project models (Mario Negri Institute)	Freely available			•	•	•						
Derek (Lhasa Ltd)	Commercial			•	•	•	•	•		•		•
HazardExpert (CompuDrug)	Commercial			•	•					•		•
Lazar (In Silico Toxicology; Freiburg university)	Freely available		• (1)	•	•			•				
Leadscope (Leadscope)	Commercial			•	•	•		•	•	•		
MCASE/MC4PC (MultiCASE)	Commercial	•	•		•	•	•	•	•		•	
MDL QSAR (MDL)	Commercial	•	• (1)	•	•			•	•			
OASIS-TIMES (Laboratory of Mathematical Chemistry, Bourgas University)	Commercial			•			•					
OncoLogic (US EPA)	Freely available				•							
Pallas Suite including ToxAlert, Cytotoxicity (CompuDrug)	Commercial			•	•					•	•	
TerraQSAR (TerraBase)	Commercial	•					•					
TOPKAT (Accelrys)	Commercial	•	•	•	•	•						
Toxtree (JRC)	Freely available		• (2)	•	•							
Molcode Toolbox (Molcode Ltd)	Commercial		•	•	•		•				•	

http://publications.jrc.ec.europa.eu/repository/bitstream/JRC59685/reqno_jrc59685_software_tools_for_toxicity_prediction%5B1%5D.pdf

QSARs: Needs and Applications

Too many chemicals problem
- Many chemicals to evaluate for multiple toxicity endpoints
- More sensitive analytical chemistry methods for chemical identification
- Lack of sufficient and relevant in vivo data

Alternative to animal testing
- Broad applications as a faster and cheaper alternative to animal testing methods in academia, industry and government institutions

Regulatory uses
- Supplement experimental data
- Support prioritization in the absence of experimental data
- Substitute or replace experimental animal testing methods

Rational chemical design
- Design and development of new drugs, perfumes, dye etc. in an efficient manner

Promoting green chemistry
- Design of chemical products and processes that reduce or eliminate the use/generation of hazardous substances.

QSAR: Regulatory Applicability

Organization	Guidelines
Multi-National Consortium of 34 countries **OECD** - Organisation for Economic Co-operation and Development (Established 1961)	**OECD Principles for the Validation of (Q)SARs (2004)[1]** • A defined endpoint • An unambiguous algorithm • A defined domain of applicability • Appropriate measures of goodness-of-fit, robustness and predictivity • A mechanistic interpretation, if possible
European Union Driven by the requirements for safety assessment and characterization of existing and new chemicals, the European Chemicals Agency (ECHA) established the **REACH** (Registration, Evaluation, Authorization and Restriction of Chemicals) regulation (Came into force 2007) • Animal testing is only allowed as a last resort	**(Q)SARs in REACH (described in Annex XI of the REACH regulation)[2]** • Results are derived from a (Q)SAR model which is scientifically valid • The chemical of interest falls under the applicability domain of the (Q)SAR model • The predictions are adequate for the purpose of classification & labeling and/or risk assessment • Adequate and reliable documentation on the (Q)SAR model and its prediction is available (structured using the OECD principles)

Red: Statistical validation
Green: Scientific explanation

[1] http://www.oecd.org/env/ehs/risk-assessment/37849783.pdf
[2] https://echa.europa.eu/regulations/reach/legislation

QSAR: Workflow

1. Generation of molecular descriptors from chemical structure

2. Selection of most relevant molecular descriptors

3. Statistical mapping of the descriptors to a toxic endpoint

4. Model validation

5. Model application

6. Documentation

Molecular descriptors are a quantification of the various molecular properties of a chemical compound. There are different levels of chemical representation ranging from 1D to 4D[1]

Descriptor Types	Description
1D	They consider properties inferred only the chemical formula of a chemical
2D	They consider properties inferred about the structure of the chemical based on the 2 dimensional structural formula
3D	They consider properties inferred from the spatial shape of the chemical for one conformation
4D	They are similar to 3D descriptors extended to multiple conformations

[1] R Todeschini et al. Handbook of molecular descriptors

Tools to calculate molecular descriptors:

Descriptor Name	Descriptor Type	Availability
Chemistry Development Kit	Continuous	Free. https://cdk.github.io/
PADel	Continuous /Fingerprints	Free. http://www.yapcwsoft.com/dd/padeldescriptor
RDKit	Continuous /Fingerprints	Free. http://www.rdkit.org
MOE	Continuous	Free. https://www.chemcomp.com/journal/descr.htm
Dragon	Continuous	Commercial. http://www.talete.mi.it/products/dragon_description.htm
PubChem	Fingerprints	Free. ftp://ftp.ncbi.nlm.nih.gov/pubchem/specifications/pubchem_fingerprints.pdf
Chemotypes	Fingerprints	Free. https://toxprint.org

QSAR WORKFLOW: Molecular Descriptors

2D descriptors are the most commonly used molecular descriptors

2D Descriptor Types	Description	Examples
Constitutional Descriptors	They represent properties related to molecular structure	molecular weight, total number of atoms in the molecule, number of aromatic rings
Electrostatic	They represent properties related to the electronic nature of the compound	atomic net and partial charges
Topological Descriptors	They represent properties which can be inferred by treating the structure of the compound as a graph, with atoms as vertices and covalent bonds as edges	total number of bonds in shortest paths between all pairs of non-hydrogen atoms
Geometrical Descriptors	They represent properties related to spatial arrangement of atoms constituting the compound	Vander Waals Area
Fragment based Descriptors	They represent properties related to sub-structural motifs	MDL Keys and Molecular Fingerprints

QSAR: Workflow

QSAR WORKFLOW: Feature Selection

Improves Interpretation
- Less features, simpler models.
- Expert-driven feature selection enhances the mechanistic interpretation of the models.

Reduces Overfitting
- Less redundant data means lesser decisions based on noise.

Reduces Training Time
- Less data to learn from ensures quicker model development.

Univariate Feature Selection	Recursive Feature Elimination
Principal Component Analysis	Feature Importance
Correlated Feature Removal	Expert-driven Feature Selection
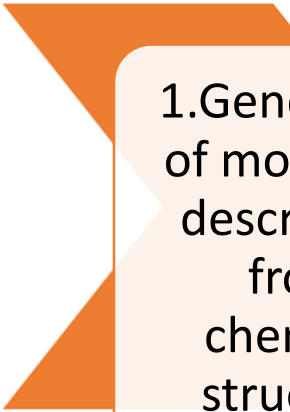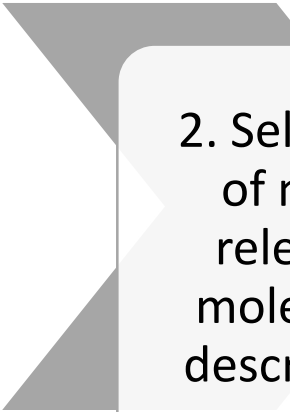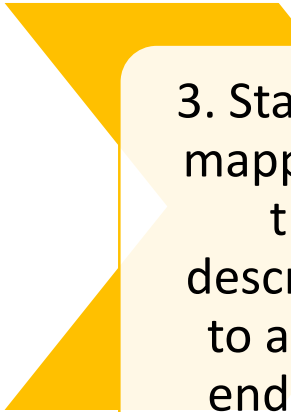
QSAR: Workflow
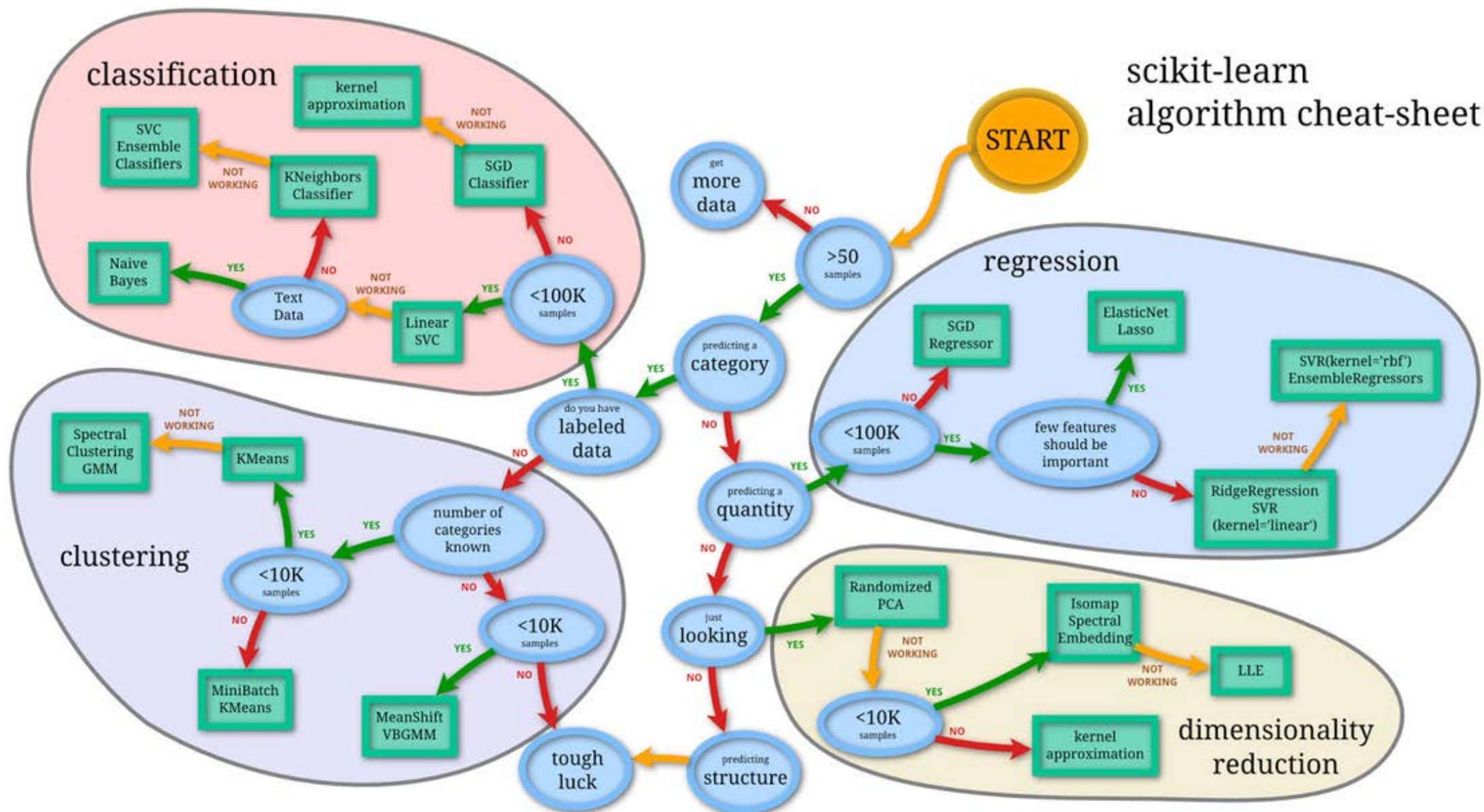
scikit-learn algorithm cheat-sheet

k-nearest **Neighbor** is a non-parametric method used in classification and regression problems.

Principle: The property of an instance (chemical) is similar to instances close to them, where closeness is defined by the appropriate *distance function* using the feature space (molecular descriptors).

Highlights

- Different distance functions available: Euclidean, Manhattan, Minkowski
- Simple to implement
- Easy to interpret (conceptually similar to read-across)
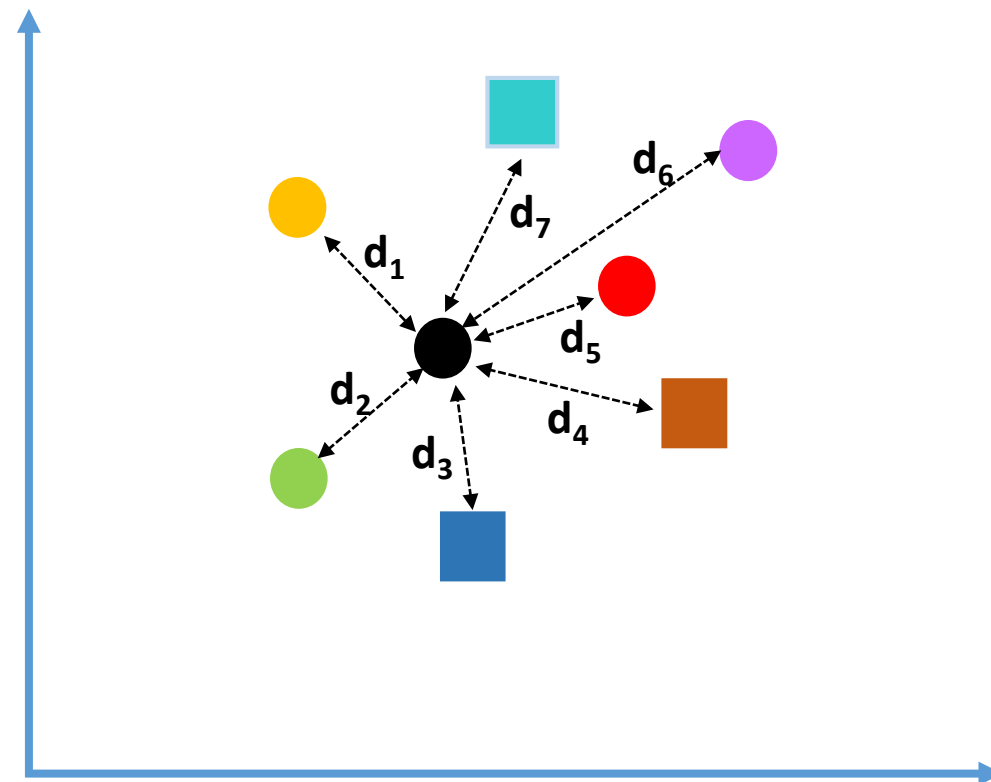
Support vector machine is a linear binary classifier which calculates an optimal hyper-plane for categorizing data.

The hyper-plane separates all data points of one class from those of the other class and is used to classify any new data points

Highlights

- Different kernel methods available for linear and non-linear data separation
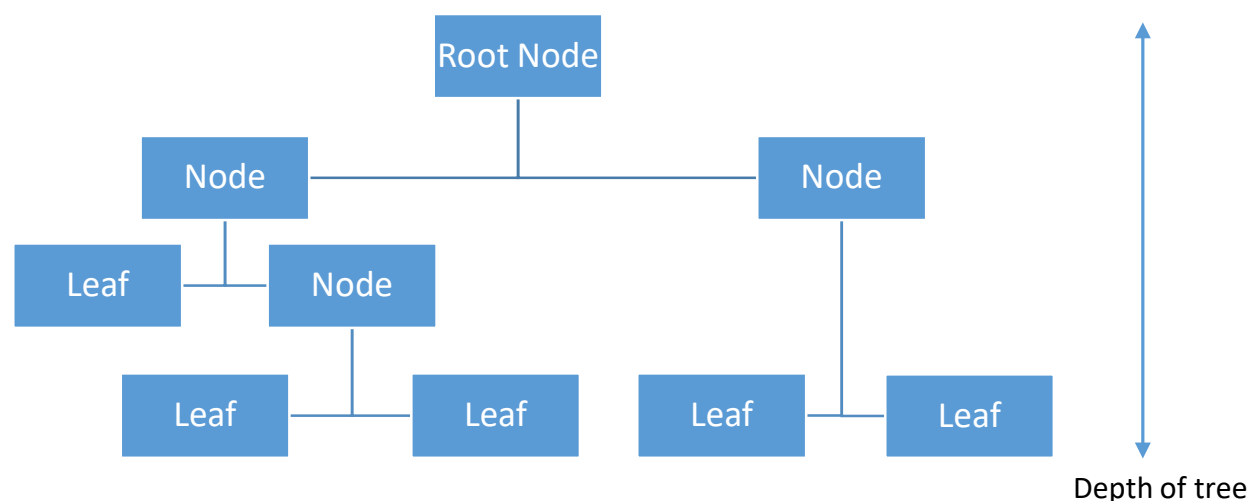- Especially suited for problems with small sized training data and binary classifiers
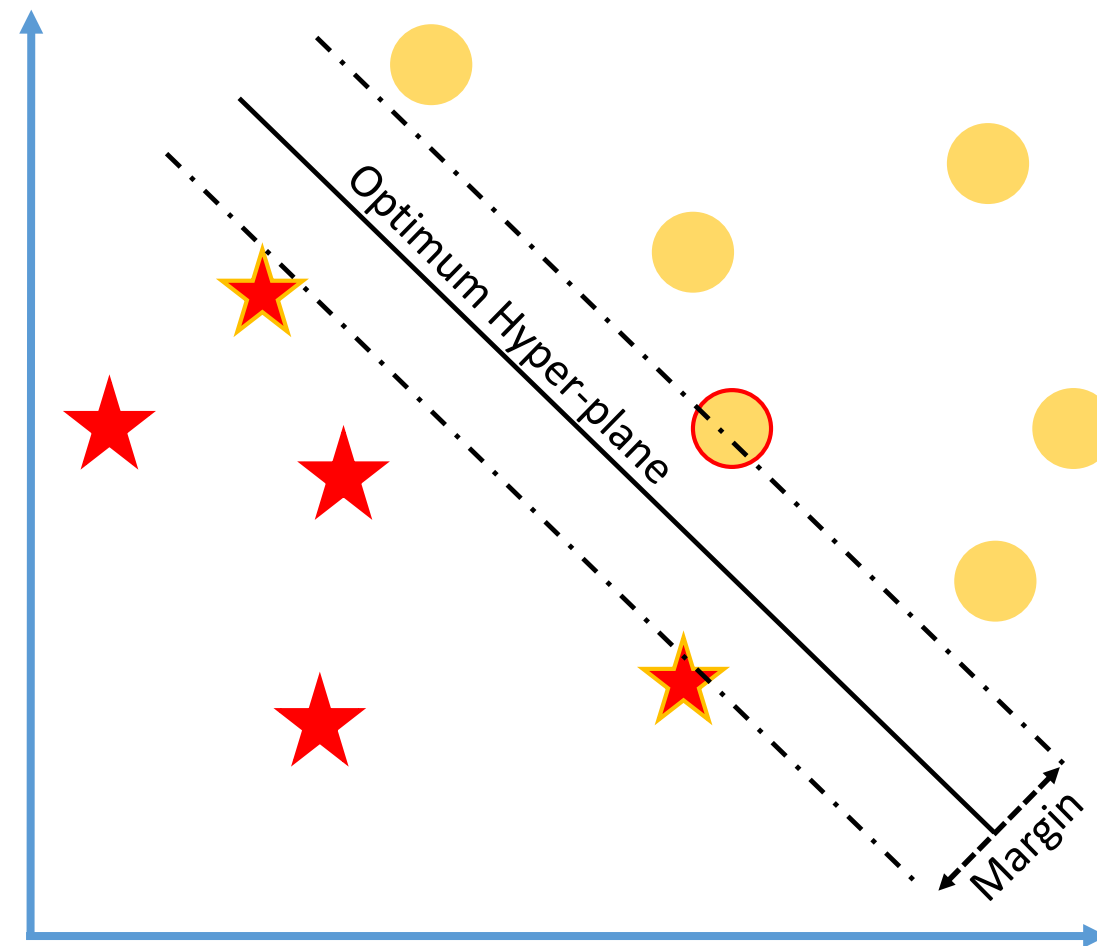
Decision tree is a non-parametric supervised learning method used for classification and regression. It is a divide and conquer algorithm that works by partitioning the data into subsets that contain data with similar values

Decision Tree Components
- **Root node** is the starting point of the tree
- **Node** is the decision point from where data is partitioned into subsets
- **Branches** are the decision outcome path that lead to a node/leaf
- **Leaf node** is the last stage of the decision path when an outcome is reached

Image: http://grannysuesnews.blogspot.com/2011/05/tree-of-hearts.html

Root Node

Node

Node

Leaf

Node

Leaf

Leaf

Leaf

Leaf

Depth of tree

Decision Tree Hyper-parameters
- Depth of tree
- Minimum number of samples to split at a node
- Maximum number of features to consider at each split

Decision Tree Limitations:
- Overfitting
- Underfitting
- High variance

Random forest constructs an ensemble of random decision trees. The new data is classified based on the majority prediction of all the trees in the ensemble.

Principle
High variance can be mitigated by averaging predictions from multiple decision trees.

Method: Each tree is developed by
i.   Selecting a bootstrap sample from the training data with replacement,
ii. Randomly selecting the best descriptor variables at each node and growing the tree, and then
iii. Estimating the classification error by testing the tree on the remaining data.
The new data is classified based on the majority prediction of all the trees in the ensemble

Highlights
- Intrinsic feature selection
- Cross-validation not necessary
- 2 key hyper-parameters need tuning
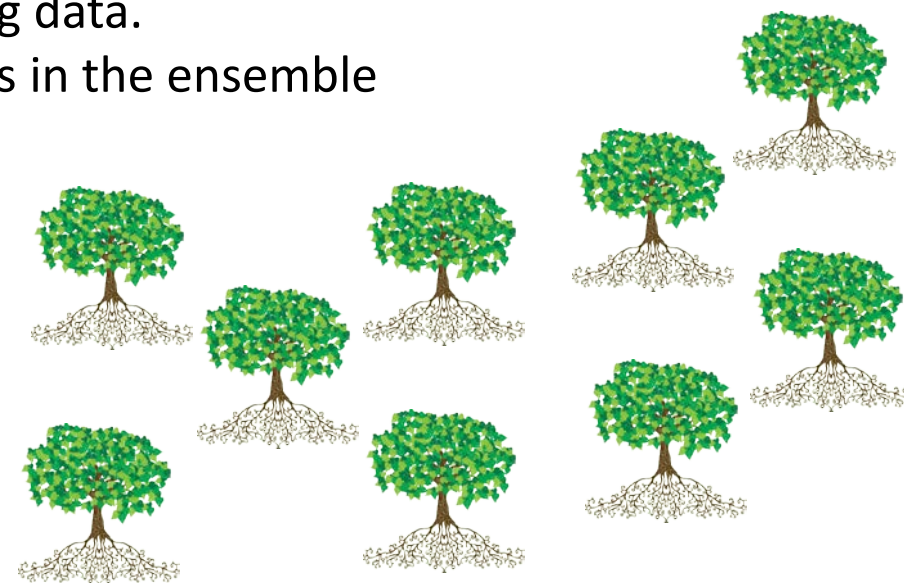
QSAR: Workflow

1. Internal validation [x%]
   - K-fold cross validation: The dataset is split into K parts. K models are developed using (K-1) sets and the Kth set is used as the test set.
   - Leave one out cross-validation: N models are developed each with (N − 1) chemicals as training set and 1 chemical as the test set.
2. External test set validation [(100- x)%]

Classification Model Metrics	Regression Model Metrics
• Accuracy • Sensitivity • Specificity • Balance Accuracy • Positive Predictivity • Negative Predictivity • Receiver operating curves	• Root-mean-squared-error • Mean Average Error • Coefficient of Determination

QSAR: Workflow

QSAR: Applicability Domain

The applicability domain (AD) of a QSAR model is defined as the "the response and chemical structure space in which the model makes predictions with a given reliability".[1]

AD evaluation enables the assessment whether the model will be useful and applicable to new chemicals.

A Stepwise Approach for Defining the Applicability Domain of SAR and QSAR Models

Sabcho Dimitrov,[†]Gergana Dimitrova,[†]Todor Pavlov,[†]Nadezhda Dimitrova,[†]Grace Patlewicz,[‡]Jay Niemela,[§] andOvanes Mekenyan*[†]
Laboratory of Mathematical Chemistry, University "Prof. As. Zlatarov", 8010 Bourgas, Bulgaria, Safety and Environmental Assurance Centre (SEAC), Unilever Colworth, Sharnbrook, Bedford MK44 1LQUK, United Kingdom, and Danish Institute for Food and Veterinary Research, 19 Mørkhøj Bygade, DK-2860 Søborg, Denmark

J. Chem. Inf. Model., **2005**, *45* (4), pp 839–849
DOI: 10.1021/ci0500381
Publication Date (Web): June 22, 2005
Copyright © 2005 American Chemical Society

Molecules **2012**, *17*(5), 4791-4810; doi:10.3390/molecules17054791

Open Access   Article

Comparison of Different Approaches to Define the Applicability Domain of QSAR Models

Faizan Sahigara ✉, Kamel Mansouri ✉, Davide Ballabio ✉, Andrea Mauri ✉,
Viviana Consonni ✉ⓘ and Roberto Todeschini *✉ⓘ

Milano Chemometrics and QSAR Research Group, Department of Environmental Sciences, University of Milano-Bicocca, P.za della Scienza 1-20126 Milano, Italy

*   Author to whom correspondence should be addressed.

[1] Current status of methods for defining the applicability domain of (quantitative) structure-activity relationships. The report and recommendations of ECVAM Workshop 52.

QSAR: Workflow

1. Generation of molecular descriptors from chemical structure

2. Selection of most relevant molecular descriptors

3. Statistical mapping of the descriptors to a toxic endpoint

4. Model validation

5. Model application

6. Documentation

QSAR Model Reporting Format (QMRF)
"The QSAR Model Reporting Format (QMRF) was developed by the JRC and EU Member State authorities as a harmonised template for summarising and reporting key information on QSAR models, including the results of any validation studies. The information is structured according to the OECD validation principles."

QSAR Prediction Reporting Format (QPRF)
"The QSAR Prediction Reporting Format (QPRF) is a harmonised template for summarizing and reporting substance-specific predictions generated by (Q)SAR models."

Details available at:
https://eurl-ecvam.jrc.ec.europa.eu/databases/jrc-qsar-model-database

- QMRF Editor is a Java application, which provides user friendly interface for editing QMRF files.
- JRC QSAR Model Database is a web application, designed to accommodate users to search QMRF files.

Screenshot of QMRF Editor 2.0.0

Source:
https://sourceforge.net/p/qmrf/wiki/JRC%20QSAR%20Model%20Database/

QSAR: Limitations and Challenges

1. Lack of proper chemical coverage in the training datasets which affects the applicability domain of the models and subsequently their suitability across different chemical classes

2. Low predictivity for mechanistically complex endpoints

3. Effect of quality and quantity of underlying training data
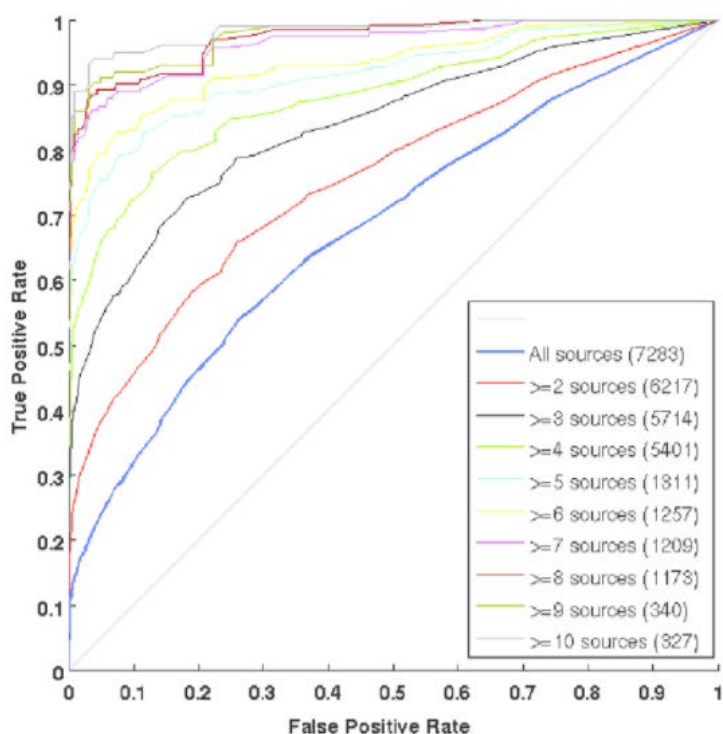

Image: Mansouri et al. "CERAPP: Collaborative Estrogen Receptor Activity Prediction Project"
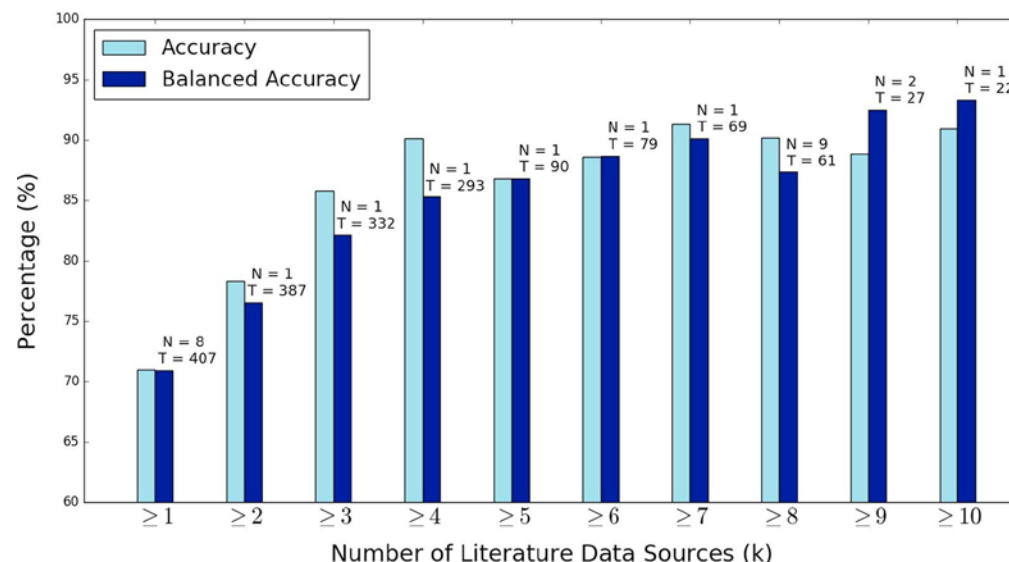

Image: Pradeep et al. "A systematic evaluation of analogs and automated read-across prediction of estrogenicity: A case study using hindered phenols"

4. Conflicting predictions by different QSAR models[1]

Chemical	Toxtree	Lazar	OECD Toolbox	Danish QSAR
Biphenyl (Carcinogen)	✗	✗	✗	✗
1,3-Butadiene (Carcinogen)	✗	✓	✗	✓
Crotonaldehyde (Carcinogen)	✗	✗	✓	✓
Chlorodifluoromethane (Non-carcinogen)	✓	✗	✓	✓
1-Phenyl-2-thiourea (Non-carcinogen)	✓	✗	✓	✗

Table 2.3: Misleading carcinogenicity predictions by QSAR tools. The ✓ represents carcinogenic and ✗ represent non-carcinogenic predictions, respectively.

5. Predictive performance of QSAR tools varies with the chemical set under study[2]

Table 2
Performance metrics for air toxins dataset

Model	Accuracy (%)	SN (%)	SP (%)	BA (%)
Toxtree	75.56	68.18	79.51	73.85
Lazar	75.24	74.55	75.61	75.08
Danish QSAR	74.29	80.91	70.73	75.82
OECD toolbox	76.19	69.09	80.00	74.55

Table 3
Performance metrics for the CPDB dataset

Model	Accuracy (%)	SN (%)	SP (%)	BA (%)
Toxtree	66.04	84.50	44.59	64.55
Lazar	80.63	86.05	74.32	80.19
Danish QSAR	65.00	91.09	34.68	62.89
OECD toolbox	64.79	84.50	41.89	63.20

[1] P. Pradeep. Hybrid Computational Toxicology Models for Regulatory Risk Assessment
[2] Pradeep et al. An ensemble model of QSAR tools for regulatory risk assessment. J. Cheminform., 8 (2016), p. 48.

QSAR: Limitations and Challenges (Contd.)

- Conflicting predictions raise interpretation, validation and adequacy concerns

- Optimization of false positives and false negatives is important. E.g.,
  - A chemical that is falsely predicted non-carcinogenic may pass regulatory approval but will cause exposure risk to cancer
  - A drug that is known to cure depression can be approved if it causes skin sensitization but not if it induces tumors

- Choice of an appropriate tool for evaluation of toxic effects in the absence of experimental data is difficult. E.g.
  - January 2014 Elk River 4-methylcyclohexanemethanol (MCHM) spill, West Virginia

QSAR ADVANCES: Nano-QSAR or QNAR
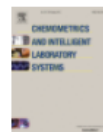
Software Description

"NanoBRIDGES" software: Open access tools to perform QSAR and nano-QSAR modeling

Pravin Ambure [a, 1], Rahul Balasaheb Aher [a, 1], Agnieszka Gajewicz [b], Tomasz Puzyn [b], Kunal Roy [a]

⊞ Show more

Exploring Quantitative Nanostructure-Activity Relationships (QNAR) Modeling as a Tool for Predicting Biological Effects of Manufactured Nanoparticles

Author(s): Denis Fourches, Dongqiuye Pu, Alexander Tropsha.

Quantitative Nanostructure–Activity Relationships: Methods, Case Studies, and Perspectives

Authors   Authors and affiliations

Denis Fourches , Ryan Lougee

Metal Oxide Nanomaterial QNAR Models: Available Structural Descriptors and Understanding of Toxicity Mechanisms

Jiali Ying [1,2,†], Ting Zhang [1,2,†,*] and Meng Tang [1,2,*]

Using nano-QSAR to predict the cytotoxicity of metal oxide nanoparticles

Tomasz Puzyn, Bakhtiyor Rasulev, Agnieszka Gajewicz, Xiaoke Hu, Thabitha P. Dasari, Andrea Michalkova, Huey-Min Hwang, Andrey Toropov, Danuta Leszczynska & Jerzy Leszczynski

Environmental, health and safety issues

Nanoparticles   Nanotoxicology

Nano-QSAR or QNAR: Challenges

The recent status and proof-of-concept studies demonstrate that QSAR modeling technique can be extended to successfully predict the biological effects of nanoparticles.

Challenges

- Lack of systematic studies for the determination of physicochemical properties of nanoparticles
- Limited strategies for the characterization (molecular descriptors) of nanomaterials unlike chemicals
- Lack of experimental data for training the models
- Limited understanding on the mechanisms of interactions between nanoparticles and biological systems

CHAPTER 10

Nano-QSAR: Advances and Challenges

B. RASULEV[a], A. GAJEWICZ[b], T. PUZYN[b], D. LESZCZYNSKA[c] AND J. LESZCZYNSKI*[a]

Review Article

Nano(Q)SAR: Challenges, pitfalls and perspectives

Ratna Tantra ✉, Ceyda Oksel, Tomasz Puzyn, Jian Wang, Kenneth N. Robinson, Xue Z. Wang,

Pages 636-642 | Received 06 May 2014, Accepted 05 Aug 2014, Published online: 11 Sep 2014

66 Download citation   ⤢ http://dx.doi.org/10.3109/17435390.2014.952698   Check for updates

QSAR: Useful Resources

QSAR Reviews

- OECD Quantitative Structure-Activity Relationships Project (http://www.oecd.org/chemicalsafety/risk-assessment/oecdquantitativestructure-activityrelationshipsprojectqsars.htm)
- The Use of Computational Methods for the Assessment of Chemicals in REACH (http://www.clbme.bas.bg/bioautomation/2009/vol_13.4/files/13.4_3.04.pdf)
- Joint research center and European Union backgroung on QSARs (https://eurl-ecvam.jrc.ec.europa.eu/laboratories-research/predictive_toxicology/background)
- Predicting Chemical Toxicity and Fate (ISBN: 9780415271806)
- Exploring QSAR: Fundamentals and Applications in Chemistry and Biology by Corwin Hansch et al (ISBN-13:9780841229877)
- QSAR: Hansch Analysis and Related Approaches by R Mannhold et al (ISBN: 978-3-527-61683-1)
- Practical guide How to use and report (Q)SARs (https://echa.europa.eu/documents/10162/13655/pg_report_qsars_en.pdf)
- Quantitative structure—activity relationships (QSAR) (DOI: 10.1016/0169-7439(89)80083-8)
- Best Practices for QSAR Model Development, Validation, and Exploitation (DOI: 10.1002/minf.201000061)
- Predictive QSAR Modeling Workflow, Model Applicability Domains, and Virtual Screening (DOI: 10.2174/138161207782794257)
- How not to develop a quantitative structure-activity or structure-property relationship (QSAR/QSPR). (DOI: 10.1080/10629360902949567)
- QSAR Modeling: Where Have You Been? Where Are You Going To? (DOI: 10.1021/jm4004285)
- How Qsars and read-across can help address REACH 2018 (https://chemicalwatch.com/22878/how-qsars-and-read-across-can-help-address-reach-2018)

QSAR: Useful Resources

QSAR Methods Reviews

- Descriptor Selection Methods in Quantitative Structure–Activity Relationship Studies: A Review Study (DOI: 10.1021/cr3004339)
- New approaches to QSAR: Neural networks and machine learning (DOI: 10.1007/BF02174529)
- Machine Learning: An Artificial Intelligence Approach (ISBN: 366212405X, 9783662124055)
- Scikit-learn: Machine Learning in Python (http://scikit-learn.org/stable/)
- Machine Learning in R for beginners (https://www.datacamp.com/community/tutorials/machine-learning-in-r)
- http://dataconomy.com/2017/03/beginners-guide-machine-learning/
- http://machinelearningmastery.com/start-here/#algorithms

ACKNOWLEDGEMENTS

All mentors and collaborators!

Thank you!