

Computational Notebooks for Cheminformatics

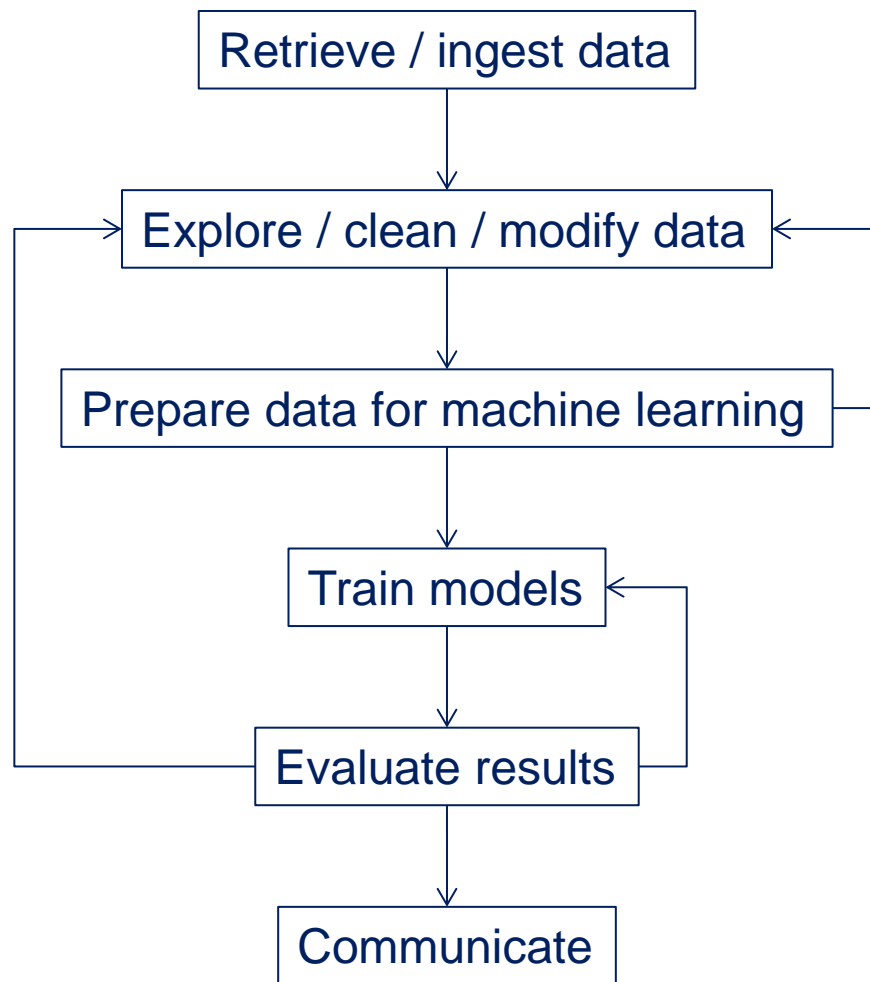
ACS Fall 2019 CHED 285

Paul J Kowalczyk
Senior Data Scientist

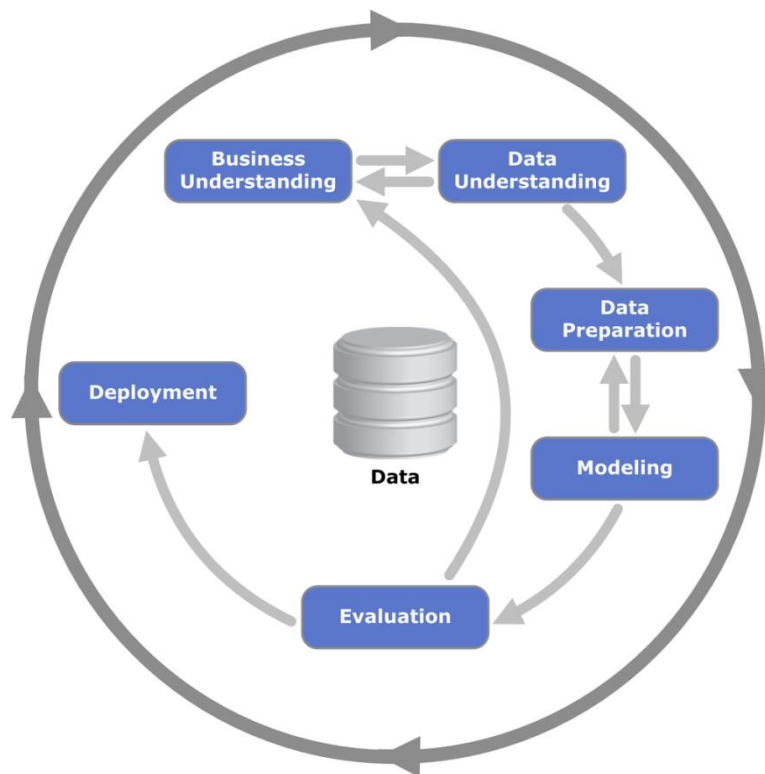
paul.kowalczyk@solvay.com

www.linkedin.com/in/PaulJKowalczyk

Machine Learning Workflow



Cross-Industry Standard Process for Data Mining

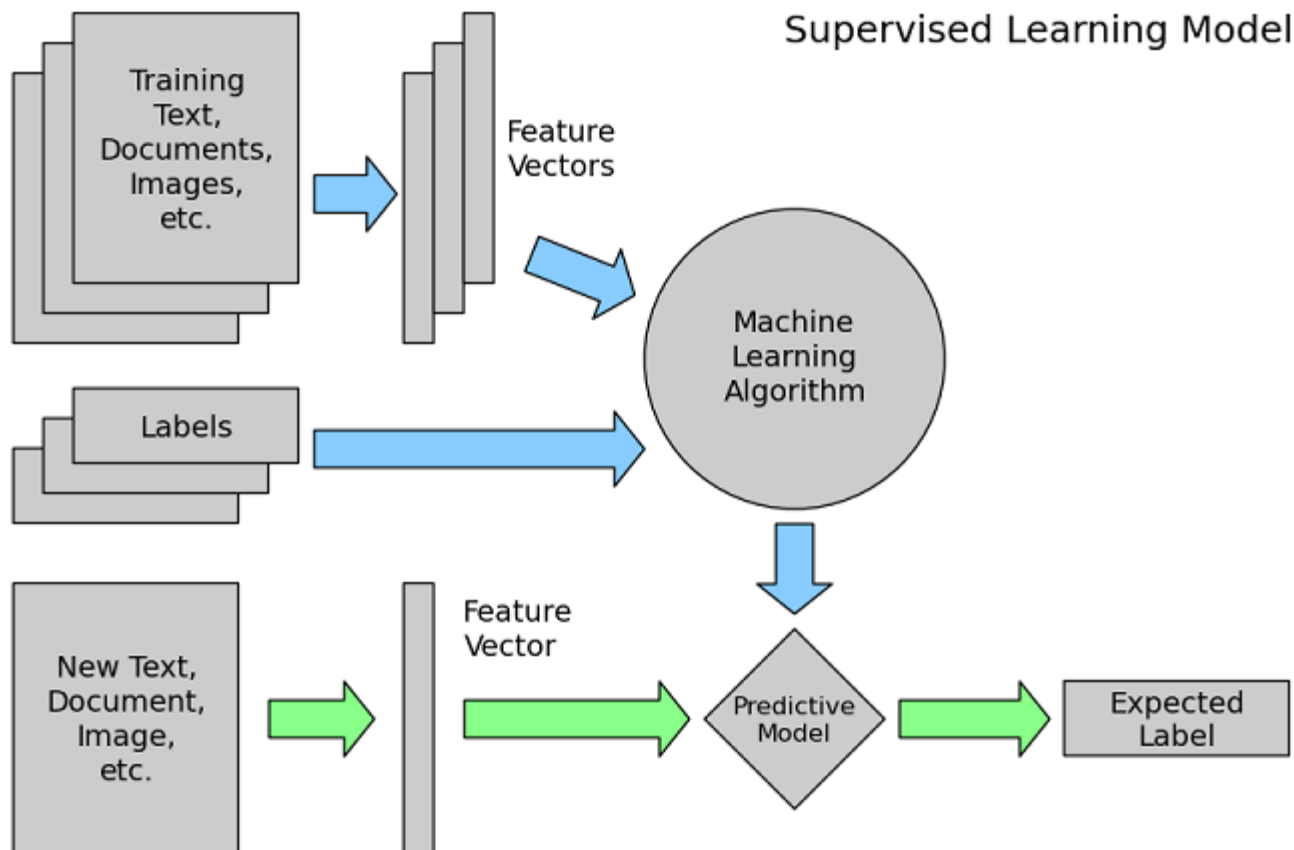


Cross-industry standard process for data mining, known as CRISP-DM, is an open standard process model that describes common approaches used by data mining experts.

The dataframe

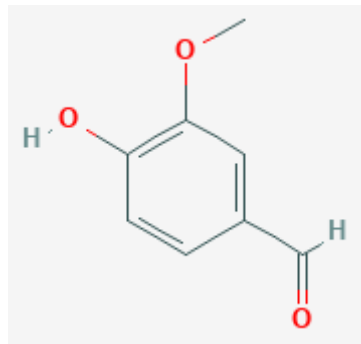
Molecule	Activity	Fingerprints and/or Descriptors				
	Y		←	X	→	

Supervised Learning Model



Representing molecules: vanillin

Vanillin



Canonocal SMILES:

COC1=C(C=CC(=C1)C=O)O

InChI=1S/C8H8O3/c1-11-8-4-6(5-9)2-3-7(8)10/h2-5,10H,1H3

InChI Key:

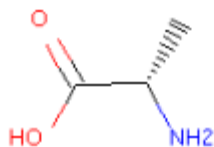
MWOOGGOJBHIARFG-UHFFFAOYSA-N

Representing molecules: alanine.mol

```
In [37]: alanine = Chem.MolFromInchi('InChI=1S/C3H7NO2/c1-2(4)3(5)6/h2H,4H2,1H3,(H,5,6)/t2-/m0/s1')
```

```
In [39]: Draw.MolToImage(alanine)
```

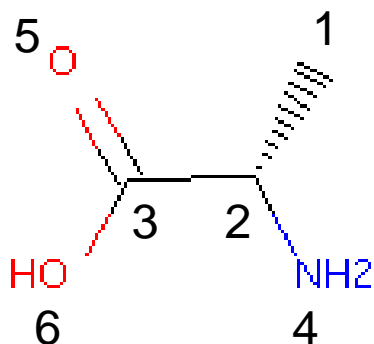
Out[39]:



```
In [40]: Chem.MolToMolFile(m1, 'alanine.mol')
```

```
RDKit          2D
6  5  0  0  0  0  0  0  0  0999 v2000
0.0000  0.0000  0.0000 C  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
1.2990  0.7500  0.0000 C  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
2.5981 -0.0000  0.0000 C  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
1.2990  2.2500  0.0000 N  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
2.5981 -1.5000  0.0000 O  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
3.8971  0.7500  0.0000 O  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
2  1  1  6
2  3  1  0
2  4  1  0
3  5  2  0
3  6  1  0
M  END
```

Representing molecules: alanine.mol



```

RDKit          2D
6   5   0   0   0   0   0   0   0   0   0999 v2000
    0.0000      0.0000      0.0000 C    0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
    1.2990      0.7500      0.0000 C    0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
    2.5981     -0.0000      0.0000 C    0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
    1.2990      2.2500      0.0000 N    0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
    2.5981     -1.5000      0.0000 O    0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
    3.8971      0.7500      0.0000 O    0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
2   1   1   6
2   3   1   0
2   4   1   0
3   5   2   0
3   6   1   0
M   END

```


Prepare data for machine learning: Descriptors (features)

List of Available Descriptors

Descriptor/Descriptor Family	Notes
Gasteiger/Marsili Partial Charges	<i>Tetrahedron</i> 36 :3219–28 (1980)
BalabanJ	<i>Chem. Phys. Lett.</i> 89 :399–404 (1982)
BertzCT	<i>J. Am. Chem. Soc.</i> 103 :3599–601 (1981)
Ipc	<i>J. Chem. Phys.</i> 67 :4517–33 (1977)
HallKierAlpha	<i>Rev. Comput. Chem.</i> 2 :367–422 (1991)
Kappa1 – Kappa3	<i>Rev. Comput. Chem.</i> 2 :367–422 (1991)
Chi0, Chi1	<i>Rev. Comput. Chem.</i> 2 :367–422 (1991)
Chi0n – Chi4n	<i>Rev. Comput. Chem.</i> 2 :367–422 (1991)
Chi0v – Chi4v	<i>Rev. Comput. Chem.</i> 2 :367–422 (1991)
MolLogP	Wildman and Crippen <i>JCICS</i> 39 :868–73 (1999)
MolMR	Wildman and Crippen <i>JCICS</i> 39 :868–73 (1999)
MolWt	
ExactMolWt	
HeavyAtomCount	
HeavyAtomMolWt	
NHOHCount	
NOCOUNT	
NumHAcceptors	
NumHDonors	
NumHeteroatoms	
NumRotatableBonds	
NumValenceElectrons	
NumAmideBonds	
Num{Aromatic,Saturated,Aliphatic}Rings	
Num{Aromatic,Saturated,Aliphatic}{Hetero,Carbo}cycles	
RingCount	
FractionCSP3	

NumSpiroAtoms	Number of spiro atoms (atoms shared between rings that share exactly one atom)
NumBridgeheadAtoms	Number of bridgehead atoms (atoms shared between rings that share at least two bonds)
TPSA	<i>J. Med. Chem.</i> 43 :3714–7, (2000) See the section in the RDKit book describing differences to the original publication.
LabuteASA	<i>J. Mol. Graph. Mod.</i> 18 :464–77 (2000)
PEOE_VSA1 – PEOE_VSA14	MOE-type descriptors using partial charges and surface area contributions http://www.chemcomp.com/journal/vsadesc.htm
SMR_VSA1 – SMR_VSA10	MOE-type descriptors using MR contributions and surface area contributions http://www.chemcomp.com/journal/vsadesc.htm
SlogP_VSA1 – SlogP_VSA12	MOE-type descriptors using LogP contributions and surface area contributions http://www.chemcomp.com/journal/vsadesc.htm
ESate_VSA1 – ESate_VSA11	MOE-type descriptors using ESate indices and surface area contributions (developed at RD, not described in the CCG paper)
VSA_ESate1 – VSA_ESate10	MOE-type descriptors using ESate indices and surface area contributions (developed at RD, not described in the CCG paper)
MQNs	Nguyen et al. <i>ChemMedChem</i> 4 :1803–5 (2009)
Topliss fragments	implemented using a set of SMARTS definitions in \$(RDBASE)/Data/FragmentDescriptors.csv
Autocorr2D	New in 2017.09 release. Todeschini and Consoni “Descriptors from Molecular Geometry” Handbook of Chemoinformatics http://dx.doi.org/10.1002/9783527618279.ch37

Prepare data for machine learning: Fingerprints

List of Available Fingerprints

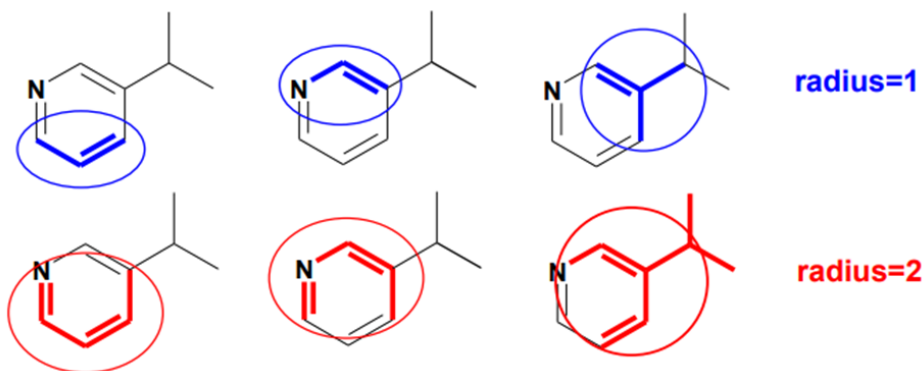
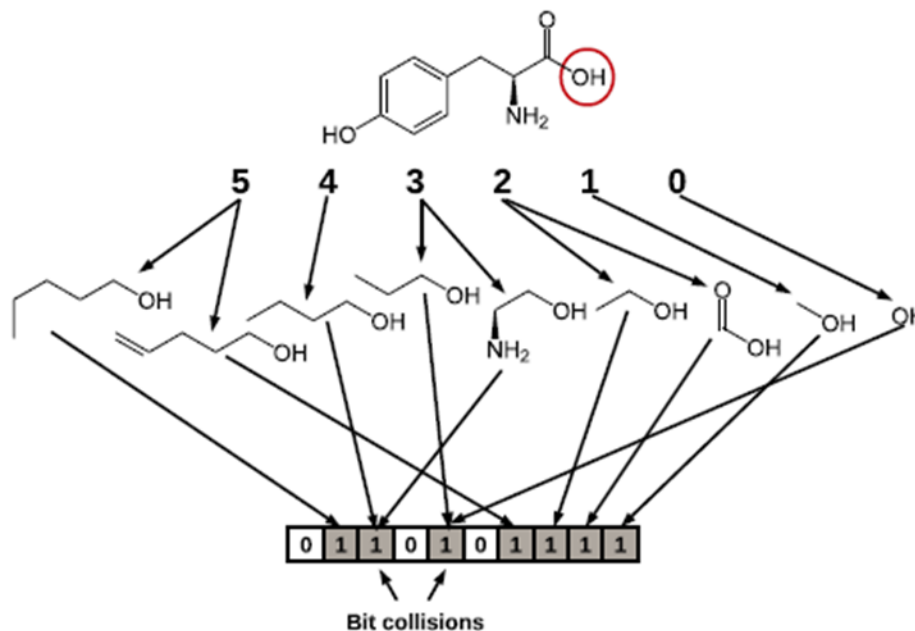
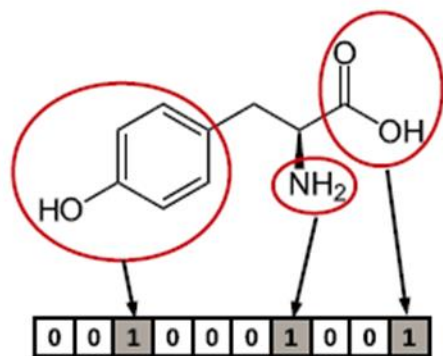
Fingerprint Type	Notes
RDKit	a Daylight-like fingerprint based on hashing molecular subgraphs
Atom Pairs	<i>JCICS</i> 25:64-73 (1985)
Topological Torsions	<i>JCICS</i> 27:82-5 (1987)
MACCS keys	Using the 166 public keys implemented as SMARTS
Morgan/Circular	Fingerprints based on the Morgan algorithm, similar to the ECFP/FCFP fingerprints <i>JCIM</i> 50:742-54 (2010).
2D Pharmacophore	Uses topological distances between pharmacophoric points.
Pattern	a topological fingerprint optimized for substructure screening
Extended	Derived from the ErG fingerprint published by Stiefl et al. in <i>JCIM</i> 46:208-20 (2006).
Reduced Graphs	NOTE: these functions return an array of floats, not the usual fingerprint types

Feature Definitions Used in the Morgan Fingerprints

These are adapted from the definitions in Gobbi, A. & Poppinger, D. "Genetic optimization of combinatorial libraries." *Biotechnology and Bioengineering* 61, 47-54 (1998).

Feature	SMARTS
Donor	[$\$([N; !H0; v3, v4 \& +1]), \$([O, S; H1; +0]), n \& H1 \& +0]$]
Acceptor	[$\$([O, S; H1; v2; !\$ (* - * = [O, N, P, S])]), \$([O, S; H0; v2]), \$([O, S; -]), \$([N; v3; !\$ (N - * = [O, N, P, S])]), n \& H0 \& +0, \$([O, s; +0; !\$ ([O, s] : n); !\$ ([O, s] : c : n)])]$]
Aromatic	[a]
Halogen	[F, Cl, Br, I]
Basic	[$\#7; +, \$([N; H2 \& +0] [\$ ([C, a]); !\$ ([C, a] (=0))]), \$([N; H1 \& +0] ([\$ ([C, a]); !\$ ([C, a] (=0))]) [\$ ([C, a]); !\$ ([C, a] (=0))]), \$([N; H0 \& +0] ([C; !\$ (C (=0))]) ([C; !\$ (C (=0))]) [C; !\$ (C (=0))])]$]
Acidic	[$\$([C, S] (= [O, S, P]) - [O; H1, -1])]$]

Prepare data for machine learning: Fingerprints



Evaluate results: Cohen's kappa

kappa takes into account the possibility of the agreement occurring by chance

Predicted	Reference	
	Event	No Event
Event	A	B
No Event	C	D

A: true positive

B: false positive (Type I error)

C: false negative (Type II error)

D: true negative

$$\kappa = \frac{\text{Pr}(a) - \text{Pr}(e)}{1 - \text{Pr}(e)}$$

$$\text{Pr}(a) = \frac{A+D}{A+B+C+D}$$

$$P_{\text{YES}} = \frac{A+B}{A+B+C+D} \times \frac{A+C}{A+B+C+D}$$

$$P_{\text{NO}} = \frac{C+D}{A+B+C+D} \times \frac{B+D}{A+B+C+D}$$

$$\text{Pr}(e) = P_{\text{YES}} + P_{\text{NO}}$$

Evaluate results: Cohen's kappa

Kappa	Agreement
< 0	Less than chance agreement
0.01 – 0.20	Slight agreement
0.21 – 0.40	Fair agreement
0.41 – 0.60	Moderate agreement
0.61 – 0.80	Substantial agreement
0.81 – 0.99	Almost perfect agreement

Retrieve / ingest data

Table 1 Endpoint datasets in the PHYSPROP database

Property abbreviation	Property
AOH	Atmospheric hydroxylation rate
BCF	Bioconcentration factor
BioHL	Biodegradability half-life
BP	Boiling point
HL	Henry's Law constant
KM	Fish biotransformation half-life
KOA	Octanol–air partition coefficient
KOC	Soil adsorption coefficient
logP	Octanol–water partition coefficient
MP	Melting point
RB	Readily biodegradable
VP	Vapor pressure
WS	Water solubility

Mansouri, K., Grulke, C. M., Judson, R. S., & Williams, A. J. (2018). OPERA models for predicting physicochemical properties and environmental fate endpoints. *Journal of cheminformatics*, 10(1), 10.

OPEn structure-activity/property Relationship App

Ready Biodegradability

OECD*i*Library

Search all content by title or author



EN ▾



My Favorites



Login

Browse by Theme ▾

Browse by Country ▾

Browse by Theme and Country ▾

Catalogue ▾

Statistics

Home > Books > OECD Guidelines for the Testing of Chemicals, Section 3 > Test No. 301: Ready Biodegradability

OECD Guidelines for the Testing of Chemicals, Section 3



Subscribe to the RSS feed

Environmental fate and behaviour

The OECD Guidelines for the Testing of Chemicals is a collection of about 150 of the most relevant internationally agreed testing methods used by government, industry and independent laboratories to identify and characterise potential hazards of chemicals. They are a set of tools for professionals, used primarily in regulatory safety testing and subseque...

[More](#)

English | Also available in: [French](#)

ISSN: 2074577X (online) | <https://doi.org/10.1787/2074577x>



Test No. 301: Ready Biodegradability

This Test Guideline describes six methods that permit the screening of chemicals for ready biodegradability in an aerobic aqueous medium. The methods are: the DOC Die-Away, the CO₂ Evolution (Modified Sturm Test), the MITI (I) (Ministry of International Trade and Industry, Japan), the Closed Bottle, the Modified OECD Screening and the Manometric Respirometry. A solution, or suspension, of the test substance, well determined/described, in a mineral medium is inoculated and incubated under aerobic conditions in the dark or in diffuse light. The running parallel blanks with inoculum but without test substance permits to determined the endogenous activity of the inoculum. A reference compound (aniline, sodium acetate or sodium benzoate) is run in parallel to check the operation of the procedures. Normally, the test lasts for 28 days. At least two flasks or vessels containing the test substance plus inoculum, and at least two flasks or vessels containing inoculum only should be used; single vessels are sufficient for the reference compound. In general, degradation is followed by the determination of parameters such as DOC, CO₂ production and oxygen uptake. The pass levels for ready biodegradability are 70% removal of DOC and 60% of ThOD or ThCO₂ production for respirometric methods. These pass values have to be reached in a 10-d window within the 28-d period of the test.

● You have access to all formats

CITE THIS PUBLICATION

EMAIL THIS PAGE

Authors
OECD

17 Jul 1992

62 pages

ISBN:
9789264070349 (PDF)

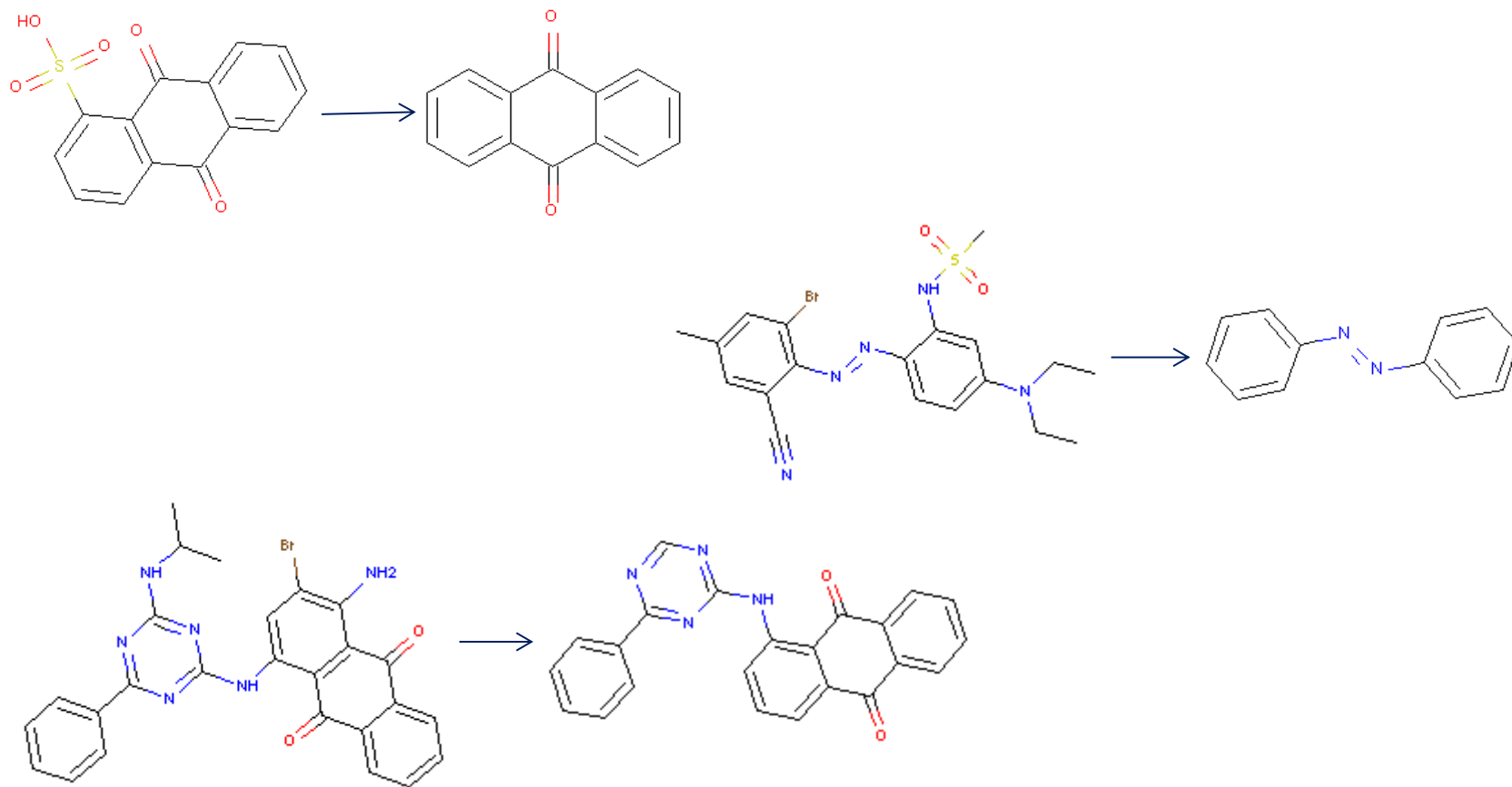
<https://doi.org/10.1787/9789264070349>
en

Ready Biodegradable (RB): 681
Not Ready Biodegradable (NRB): 1304

Explore / clean / modify data: structure curation

- Ingest the 3 biodegradability datasets
 - Cheng (JChemInfModel_52_655)
 - Mansouri (JCIM_53_867)
 - OPERA (OPERA)
- Sanitize molecules
- Identify replicates in the datasets
- Compare / contrast the datasets, *e.g.*, compare molecular weight distributions, TPSA, logP, ...

Explore / clean / modify data: Murcko frameworks

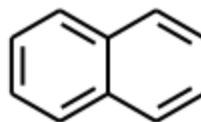


Bemis, Guy W., and Mark A. Murcko. "The properties of known drugs. 1. Molecular frameworks." *Journal of medicinal chemistry* 39.15 (1996): 2887-2893.

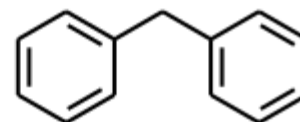
Explore / clean / modify data: Murcko frameworks



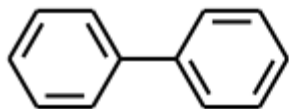
512



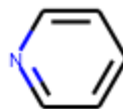
43



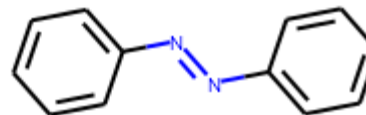
36



34

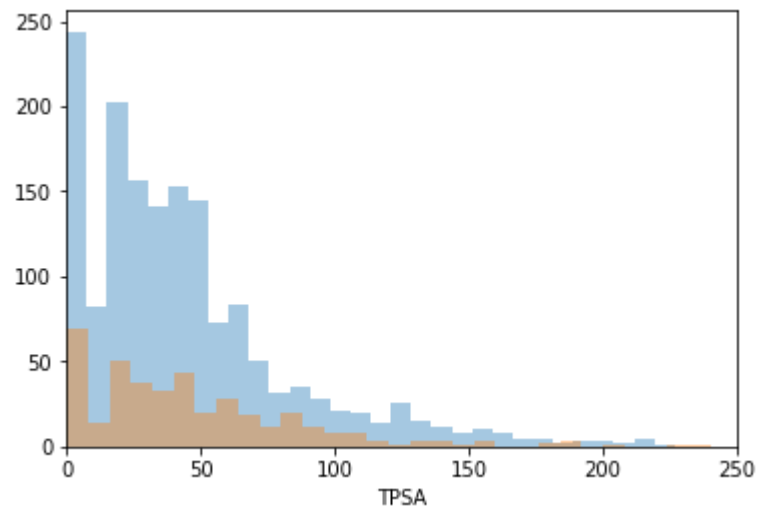
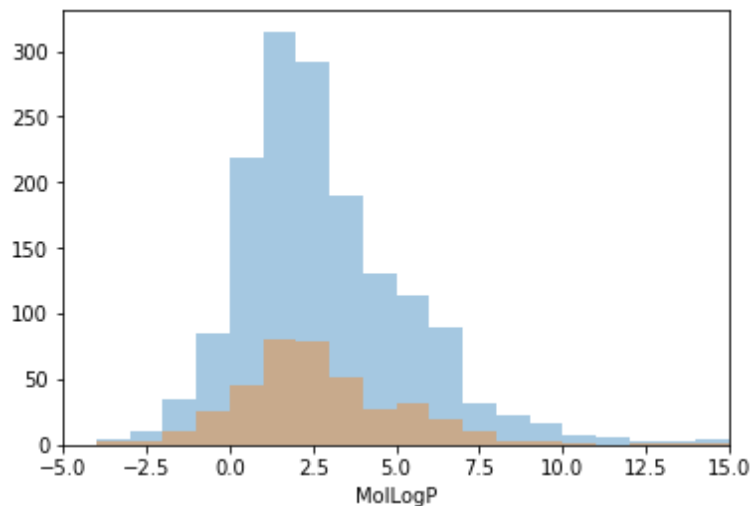
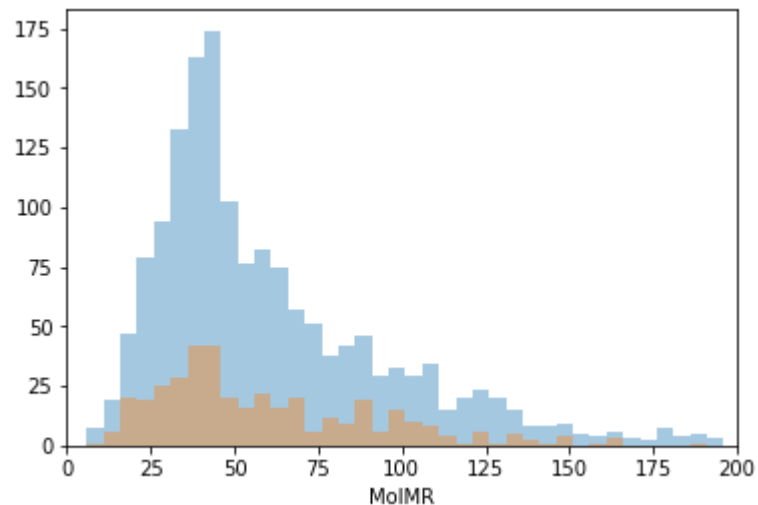
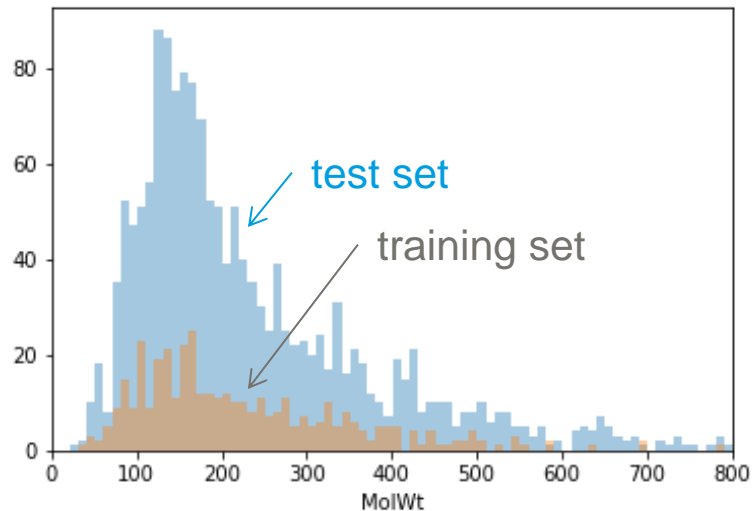


34



28

Explore / clean / modify data



Model: recursive partitioning / random forests

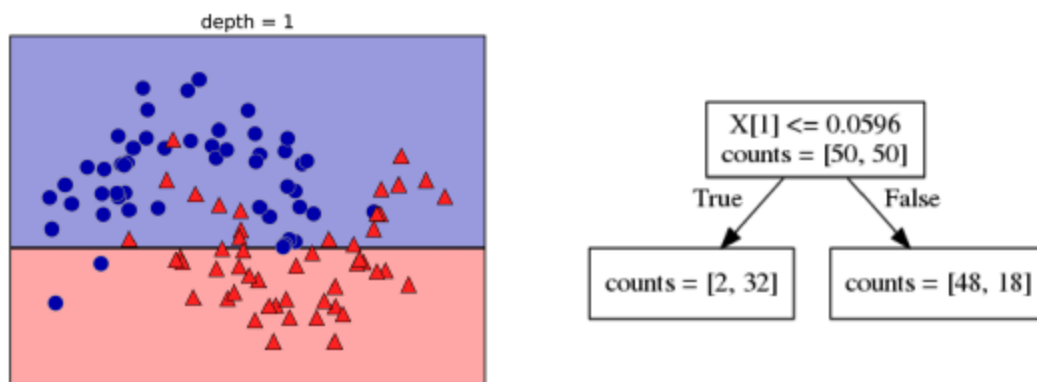


Figure 2-24. Decision boundary of tree with depth 1 (left) and corresponding tree (right)

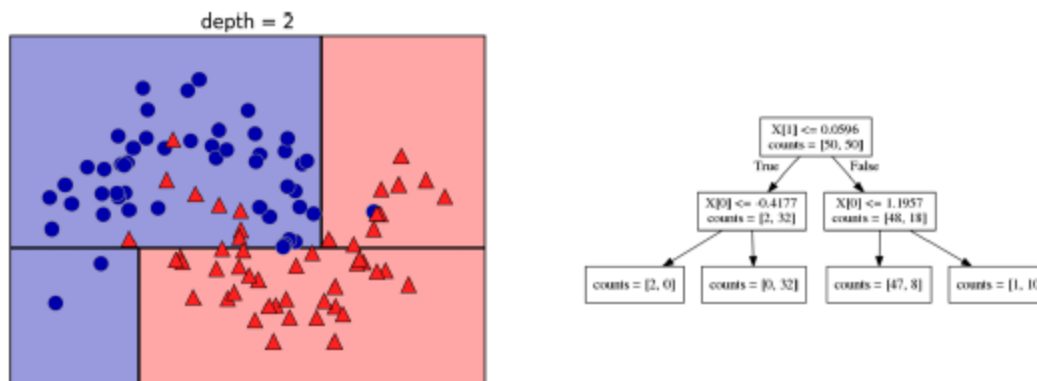
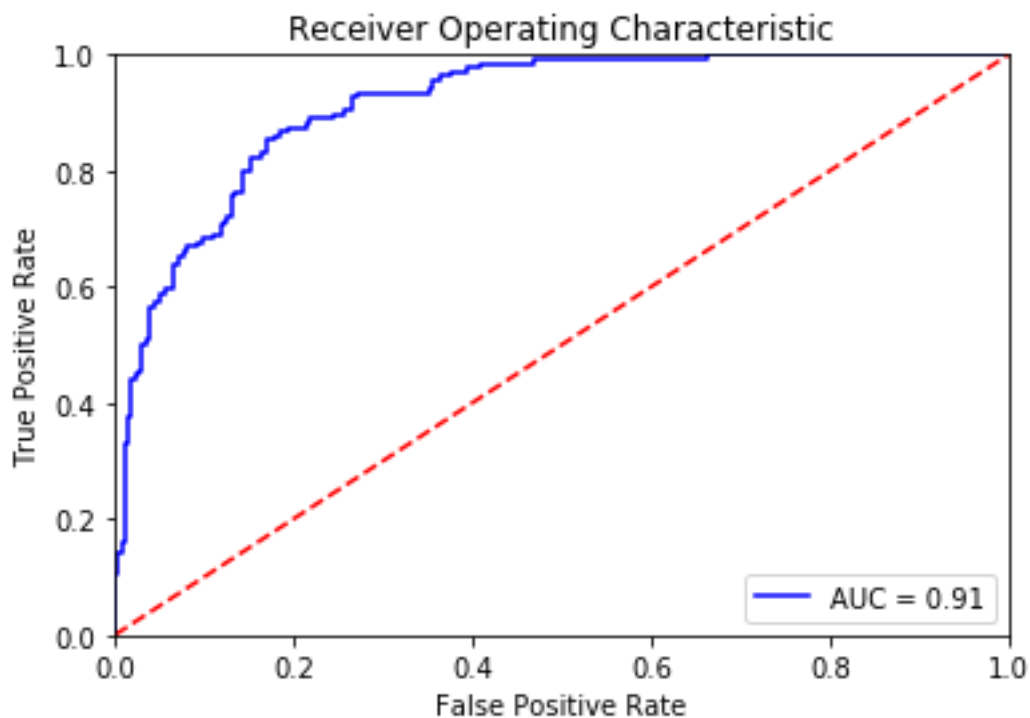


Figure 2-25. Decision boundary of tree with depth 2 (left) and corresponding decision tree (right)

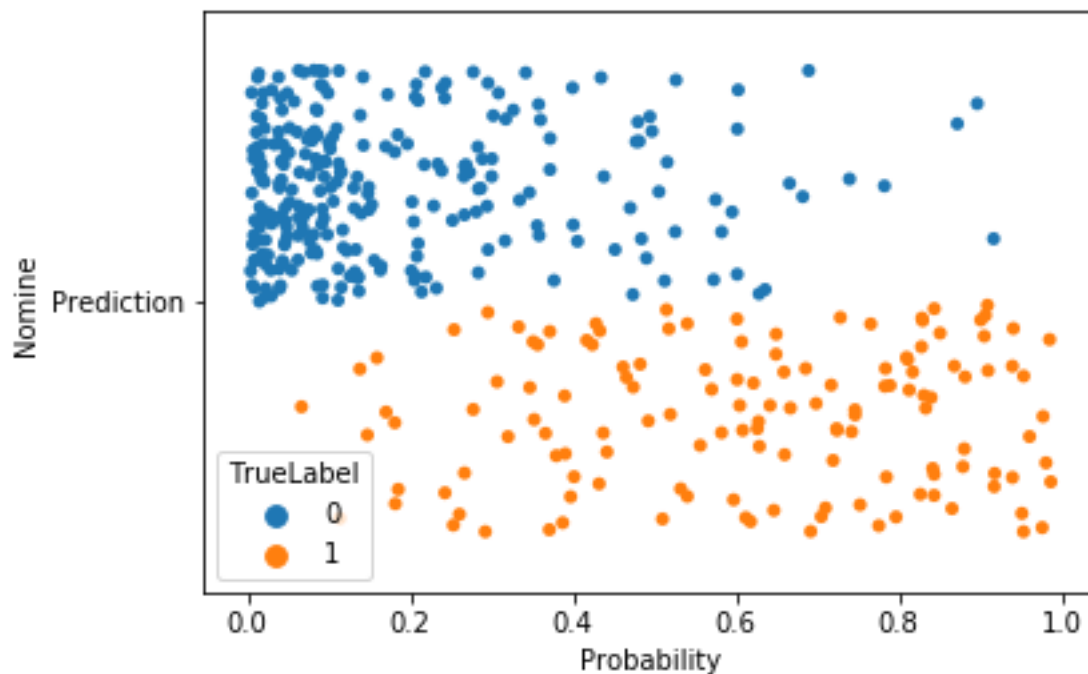
Model Performance

ROC & AUC



Model Performance: Confusion Matrix

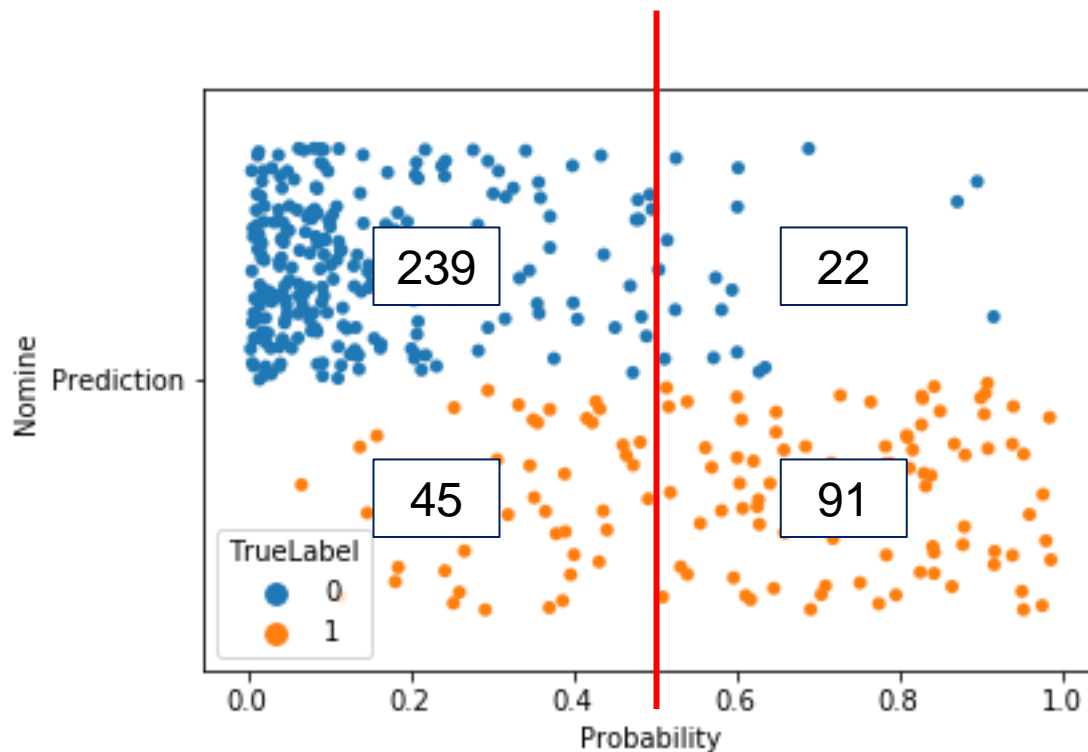
kappa = 0.61

$$\begin{bmatrix} 239 & 22 \\ 45 & 91 \end{bmatrix}$$


Model Performance: Confusion Matrix

kappa = 0.61

$\begin{bmatrix} 239 & 22 \\ 45 & 91 \end{bmatrix}$



notebook

Merci!

Thank You!

Dziękuję

paul.kowalczyk@solvay.com
www.linkedin.com/in/PaulJKowalczyk

www.solvay.com



SOLVAY

asking more from chemistry®