VEGA

Guide to Mutagenicity SarPy/IRFMN Model version 1.0.7

Table of Contents

1. Model explanation

1.1 Introduction

The model provides a qualitative prediction of mutagenicity on Salmonella typhimurium (Ames test). It is implemented inside the VEGA online platform, accessible at: http://www.vega-qsar.eu/

1.2 Model details

The model has been built as a set of rules, extracted with Sarpy software from the original training set from the Mutagenicity Caesar model. The model is based on T. Ferrari, D. Cattaneo, G. Gini, N. Golbamaki Bakhtyari, A. Manganaro, E. Benfenati, "Automatic knowledge extraction from chemical structures: the case of mutagenicity prediction", SAR and QSAR in Environmental Research (2013), vol. 24 issue 5, 365-83.

The original work has been extended, resulting in two sets of rules for mutagenicity (112 rules) and non-mutagenicity (93 rules). If at least one mutagenicity rule is matching with the given compound, a "mutagen" prediction is given; if only one or more non-mutagenicity rule is matching, a "non-mutagen" prediction is given; if no rules match with the given compound, a "possible non-mutagen" prediction is given.

1.3 Applicability Domain

The applicability domain of predictions is assessed using an Applicability Domain Index (ADI) that has values from 0 (worst case) to 1 (best case). The ADI is calculated by grouping several other indices, each one taking into account a particular issue of the applicability domain. Most of the indices are based on the calculation of the most similar compounds found in the training and test set of the model, calculated by a similarity index that consider molecule's fingerprint and structural aspects (count of atoms, rings and relevant fragments).

For each index, including the final ADI, three intervals for its values are defined, such that the first interval corresponds to a positive evaluation, the second one corresponds to a suspicious evaluation and the last one corresponds to a negative evaluation.

Following, all applicability domain components are reported along with their explanation and the intervals used. Note that for purpose of evaluating accuracy and concordance indices, prediction of "suspect mutagen" are considered as "mutagen".

- **Similar molecules with known experimental value**. This index takes into account how similar are the first three most similar compounds found. Values near 1 mean that the predicted compound is well represented in the dataset used to build the model, otherwise the prediction could be an extrapolation.

Defined intervals are:

1 >= index > 0.8	strongly similar compounds with known experimental value in the training set have been found
0.8 >= index > 0.6	only moderately similar compounds with known experimental value in the training set have been found
index <= 0.6	no similar compounds with known experimental value in the training set have been found

- **Accuracy of prediction for similar molecules**. This index takes into account the classification accuracy in prediction for the three most similar compounds found. Values near 1 mean that the predicted compounds falls in an area of the model's space where the model gives reliable predictions (no misclassifications), otherwise the lower is the value, the worse the model behaves. Defined intervals are:

1 >= index > 0.9	accuracy of prediction for similar molecules found in the training set is good
0.9 >= index > 0.5	accuracy of prediction for similar molecules found in the training set is not optimal
index <= 0.5	accuracy of prediction for similar molecules found in the training set is not adequate

- **Concordance for similar molecules** . This index takes into account the difference between the predicted value and the experimental values of the three most similar compounds. Values near 0 mean that the prediction made disagrees with the values found in the model's space, thus the prediction could be unreliable. Defined intervals are:

1 >= index > 0.9	similar molecules found in the training set have experimental values that agree with the predicted value
0.9 >= index > 0.5	some similar molecules found in the training set have experimental values that disagree with the predicted value
index <= 0.5	similar molecules found in the training set have experimental values that disagree with the predicted value

- **Atom Centered Fragments similarity check**. This index takes into account the presence of one or more fragments that aren't found in the training set, or that are rare fragments. First order atom centered fragments from all molecules in the training set are calculated, then compared with the first order atom centered fragments from the predicted compound; then the index is calculated as following: a first index RARE takes into account rare fragments (those who occur less than three times in the training set), having value of 1 if no such fragments are found, 0.85 if up to 2 fragments are found, 0.7 if more than 2 fragments are found; a second index NOTFOUND takes into account not found fragments, having value of 1 if no such fragments are found, 0.6 if a fragments is found, 0.4 if more than 1 fragment is found. Then, the final index is given as the product RARE * NOTFOUND. Defined intervals are:

index = 1	all atom centered fragment of the compound have been found in the compounds of the training set

1 > index >= 0.7	some atom centered fragment of the compound have not been found in the compounds of the training set or are rare fragments
index < 0.7	a prominent number of atom centered fragments of the compound have not been found in the compounds of the training set or are rare fragments

- **Global AD Index**. The final global index takes into account all the previous indices, in order to give a general global assessment on the applicability domain for the predicted compound. Defined intervals are:

1 >= index >= 0.9	predicted substance is into the Applicability Domain of the model
0.9 > index >= 0.65	predicted substance could be out of the Applicability Domain of the model
index < 0.65	predicted substance is out of the the Applicability Domain of the model

1.4 Structural Alerts for mutagen compounds

Following, the list of the 112 rules for mutagenicity, expressed as SMARTS strings:

SM 1: O=[N+]([O-])c1ccc2ccccc2c1
SM 2: O=NN(C)C
SM 3: n1ccc(N)c2ccc(cc12)
SM 4: c1oc(cc1)[N+](=O)[O-]
SM 5: O=[N+]([O-])c1ccc(c2ccccc2)c(c1)
SM 6: Nc4ccc(N)cc4N
SM 7: C1C=Cc2ccccc2C1
SM 8: N1CC1
SM 9: c1cc([N+](=O)[O-])sc1
SM 10: n1ccnc2c1ccc(N)c2
SM 11: c1ccc2c(c1)cc3ccc(cc3c2)C
SM 12: Nc1c(ncn1)
SM 13: n1cc(nc2ccc3c(ncn3C)c12)
SM 14: O(c1ccccc1)CC2OC2
SM 15: O=C1c2ccccc2C(=O)c3c(O)ccc(O)c13
SM 16: N(O)c1ccc(C=C)cc1
SM 17: c1ccc2ccc3ccc(cc3c2c1)N
SM 18: O(Cc1cccc2ccccc12)C
SM 19: O=C(c1ccc(cc1)NO)
SM 20: O=C(c1ccccc1)Cl
SM 21: C1=Cc2cccc3cccc1c23
SM 22: O1CC1CCc2ccc(cc2)
SM 23: C(O)C=CCl
SM 24: OCC(CBr)
SM 25: [N-]=[N+]
SM 26: n1c2ccc(cc2c(cc1))C
SM 27: O=CC1(OC1)C

SM 28: n1cccc2c1ccc3c2ncn3
SM 29: c2nc3C(=O)C=CC(=O)c3cc2
SM 30: O=Nc1ccc(OC)cc1
SM 31: SC(=CCl)Cl
SM 32: O=C1c2cccc(N)c2C(=O)c3ccccc13
SM 33: Nc1ccc2c(c1)c3ccccc3n2
SM 34: Oc1ccc2ccc3ccc(cc3c2c1)
SM 35: C(c1ccccc1)COC=C
SM 36: c1cc2cccc3c4cc(ccc4c(c1)c23)
SM 37: N(O)c1ccc(Oc2ccccc2)cc1
SM 38: c1ccc2c(c1)c3ccccc3n2
SM 39: O(Cc1cccc2ccccc12)
SM 40: c1ccc2c3ccccc3CCc2c1
SM 41: n1cc2ccccc2s1
SM 42: P(=O)(N)N(C)CC
SM 43: C(N)Cl
SM 44: c1ccc(C=Cc2ccc(N)cc2)cc1
SM 45: c1cc(ccc1NCCCl)
SM 46: N(c1ccc(N=Nc2ccccc2)cc1)C
SM 47: O=C(NCc1ccccc1)C
SM 48: c1cc2ccc3cccc4ccc(c1)c2c34
SM 49: Nc1cccc(c1)c2ccccc2
SM 50: O=[N+]([O-])c1cccc2cccc(c12)
SM 51: O=C(Nc1ccc(cc1)c2ccccc2)
SM 52: O=Cc1cccc(c1)[N+]
SM 53: O=[N+]([O-])c1cc(N)c(c(N)c1)
SM 54: c1ccc(Oc2ccc(N)cc2)cc1
SM 55: COC=CC=CC
SM 56: N(=N)NC
SM 57: ONc1ccc(cc1)S
SM 58: O1CC1Cc2ccc(cc2)
SM 59: O=C(c1ccccc1O)c2ccccc2
SM 60: Nc1ccc(cc1)c2ccccc2
SM 61: c1ccc2c(c1)ccc3c2cc4ccccc4c3
SM 62: c1ccc2c(c1)cc3ccc(cc3c2C)
SM 63: c1ccc2c(ccc3c4ccccc4ccc23)c1
SM 64: c1c2ccccc2nc3ccccc13
SM 65: O=CC=C(C(=O)c1ccccc1)
SM 66: n1cc(cc2c1ncn2)
SM 67: Nc1nccn1C
SM 68: C1C(C=C(C))C1(C)C
SM 69: Nc1ccc(cc1)[N+](=O)[O-]
SM 70: Nc1ccc(cc1N)
SM 71: N=CC=C
SM 72: O=[N+]([O-])c1ccc(cc1)CO
SM 73: CCNCCCl
SM 74: O=S(=O)(OCC)

SM 75: c1ccc2c3ccccc3Cc2c1
SM 76: c1c2ccccc2n(c1)C
SM 77: C(CBr)Br
SM 78: Nc1ccccc1F
SM 79: c1ccc(N)c(c1N)C
SM 80: c1ccc2c(c1)ccc3cc(ccc23)
SM 81: c1ccc2c(c1)cc3ccccc3c2
SM 82: c1ccc2ccccc2c1C
SM 83: Nc1ccc(cc1)Cc2ccccc2
SM 84: Oc1ccc2Cc3ccccc3Oc2c1
SM 85: C(Cl)(Cl)Cl
SM 86: O(c1ccccc1N)C
SM 87: NN(c1ccccc1)
SM 88: n1c(N)n(c2ccccc12)
SM 89: O=C(N(O))C
SM 90: n1ccnc2c1cccc2
SM 91: c1cc(c(N)cc1N)C
SM 92: OCC1OC1
SM 93: C(C)Br
SM 94: C(OCC)N
SM 95: Nc1cccc(N)c1
SM 96: c1c(nn(c1))
SM 97: C1OC1
SM 98: C(O)N
SM 99: c1ccc2cccnc2c1
SM 100: N=NC
SM 101: O=CC(=C)Cl
SM 102: n1cnc2c(ncn2)c1N
SM 103: NNCC
SM 104: Cc2ccc(N)cc2
SM 105: Nc1ccc(N)cc1
SM 106: CCCl
SM 107: C=NN1N=Nc2c([nH]c3ccccc23)C1=O
SM 108: NC([N+])
SM 109: n1c2ccc(cc2[s+]c3cc(N)c(cc13))N
SM 110: O=Nn1cc(c2ccccc12)CC
SM 111: O=CC(=CC)C=CC
SM 112: O=C1OCC1

1.5 Structural Alerts for non mutagen compounds

Following, the list of the 93 rules for non mutagenicity, expressed as SMARTS strings:

SM 113: C(O)CCCCCCCC=CCC
SM 114: CCOc1ccc(Cl)cc1
SM 115: C(NC(C(=O))C)C(NC)
SM 116: c1(c(ccc(c1)CCCCCC))O
SM 117: c1c(c(C(C)(C)C)cc(c1)C)
SM 118: c1(c(C(=O)O)cccc1)C(=O)
SM 119: S(=O)(=O)(N)c1ccc(N)cc1
SM 120: n1c(nc(nc1))
SM 121: C(=O)(C(CCC(=O)O))O
SM 122: CCOCCOCCOCCO
SM 123: CC(=O)OCC(CC)CCCC
SM 124: P(O)OCCCCC
SM 125: N(c1ccccc1)CCCNC
SM 126: c1(C(=O)OC)c(N)cccc1
SM 127: C(C)(Oc1ccccc1)(C)C
SM 128: N(CCO)(CCCC)C
SM 129: C(=C)CCCl
SM 130: C(=O)(C(=C)C)OCCCCC
SM 131: Oc1ccc2C=C(COc2c1)c3ccccc3
SM 132: c1(nc2c(o1)cccc2)c3ccccc3
SM 133: c1(c(c(cc(c1)))O)c2c(c(cc(c2)))O
SM 134: n1c(cc(c2c1cccc2))CO
SM 135: c1c(c(ccc1C(O)CNC)O)
SM 136: O(C(=O))C(=O)
SM 137: C(=O)(CCC(=O)O)
SM 138: S(=O)(=O)(c1ccccc1)N
SM 139: c1c(c(cc(c1Cl)Cl)Cl)
SM 140: C(C=C)(CCC=C)
SM 141: c1(CCCCCC)ccccc1
SM 142: N(C)(CCCC)CCCC
SM 143: C(=O)(O)CCCCCCCC
SM 144: C(=O)(N(c1ccccc1))N
SM 145: C1(C(CCC(C1)C)C)O
SM 146: c1(C(=O)OCC)ccccc1
SM 147: c1c(C(F)(F)F)cccc1
SM 148: C(=O)CS
SM 149: O(c1ccccc1)CC(CNC(C))O
SM 150: C(F)(F)C
SM 151: c1cc(c(c(c1)Br)O)
SM 152: c1(Br)ccc(cc1)C
SM 153: SCCCC
SM 154: C(=O)CC(=O)C

SM 155: N(CCO)(CCO)C
SM 156: C(=O)N(C)CCC
SM 157: CCCCCCCCCCCC
SM 158: c1(c(ccc(c1)C=CC)O)
SM 159: c1cc(c(cc1)Cl)Cl
SM 160: C=CC=C(CCC)C
SM 161: c1ccc(NC(C)C)cc1
SM 162: C1CC(CC(C1))(C)C
SM 163: CCCCCCC
SM 164: C(C(OC)(C)C)
SM 165: N#CCC
SM 166: C(=C(CC)O)
SM 167: CN1CN=CC=C1
SM 168: S(=O)(=O)c1ccc(N)cc1
SM 169: CC(CC)CCC
SM 170: C(=O)(c1ccccc1)OC
SM 171: c1(F)ccc(cc1)C
SM 172: C(=Cc1ccccc1)C(=O)c2ccccc2
SM 173: [nH]1cc(c2c1cccc2)CC
SM 174: P(=S)(OCC)OC
SM 175: [N+](C)(C)C
SM 176: OCCN(C)C
SM 177: C(=O)CCCCC
SM 178: C(=O)(CC(C))OCC
SM 179: CC#N
SM 180: C(=C(CC)C)C=O
SM 181: O=C1NC=NC=C1
SM 182: CCCC(C)C
SM 183: c1(cc(ccc1)Cl)Cl
SM 184: c12c(cc(cc2)O)ccc(S)c1
SM 185: PC
SM 186: OCC(CO)(C)C
SM 187: OC(=O)C(=C)C
SM 188: C(=O)(C)OCCCC
SM 189: c1(C(=O)O)c(ccc(c1))O
SM 190: c1(c(Cl)cccc1)C
SM 191: Cc1cc(c(cc1)OC)OC
SM 192: C#N
SM 193: Cc1ccc(Cl)cc1
SM 194: Nc1ccc2C=CCOc2c1
SM 195: C(=O)(O)Cc1ccccc1
SM 196: c1(cc(ccc1CCC)O)O
SM 197: N(c1ccccc1)c2ccccc2
SM 198: C=NO
SM 199: C(=S)(N)
SM 200: c1(CC)cc(OC)ccc1
SM 201: c1(nc2c(s1)cccc2)

SM 202: OCCOCCO
SM 203: N(C)(CCN)c1ccccc1
SM 204: [N+]([O-])CC
SM 205: S(=O)(=O)(c1cc(c(cc1))N)O

1.6 Model statistics

Following, statistics obtained applying the model to its original dataset:

- Training set: n = 3367; Accuracy = 0.82; Specificity = 0.77; Sensitivity = 0.86
- Test set: n = 837; Accuracy = 0.81; Specificity = 0.76; Sensitivity = 0.86

2. Model usage

2.1 Input

The model accepts as input two molecule formats: SDF (multiple MOL file) and SMILES. All molecules found as input are preprocessed before the calculation of molecular descriptors, in order to obtain a standardized representation of compound. For this reason, some cautions should be taken.

- **Hydrogen atoms**. In SDF files, hydrogen atoms should be explicit. As some times SDF file store only skeleton atoms, and hydrogen atoms are implicit, during the processing of the molecule the system tries to add implicit hydrogens on the basis of the known standard valence of each atom (for example, if a carbon atoms has three single bonds, an hydrogen atom will be added such to reach a valence of four). In SMILES molecules, the default notation uses implicit hydrogen. Anyway please note that in some cases it is necessary to explicitly report an hydrogen; this happens when the conformation is not unambiguous. For example, when a nitrogen atom is into an aromatic ring with a notation like "cnc" it is not clear whether it corresponds to C-N=C or to C-[NH]-C, thus if the situation is the latter, it should be explicitly reported as "c[nH]c".

- **Aromaticity**. The system calculates aromaticity using the basic Hueckel rule. Note that each software for drawing and storing of molecules can use different approaches to aromaticity (for instance, commonly the user can choose between the basic Hueckel rule and a loose approach that lead to considering aromatic a greater number of rings). As in the input files aromaticity can be set explicitly (for instance, in SMILES format by using lowercase letters), during the processing of the molecule the system removes aromaticity from rings that don't satisfy the Hueckel rule. Please note that when aromaticity is removed from a ring, it is not always possible to rebuild the original structure in Kekule form (i.e. with an alternation of single and double bonds, like in the SMILES for benzene, C=1C=CC=CC1), in this case all bonds are set to single. Furthermore, please note that aromaticity detection is a really relevant issue, some molecular descriptors can have significantly different values whether a ring is perceived as aromatic or not. For this reason it is strongly recommended:
- Always use explicit hydrogens in SDF file.
- Avoid explicit aromaticity notation in original files; in this way, the perception of aromaticity is left to the preprocessing step and there is no chance of mistakes due to the transformation of rings that were set to aromatic in the original format but not recognized as aromatic in VEGA.

Note that when some modification of the molecule are performed during the preprocessing (e.g. adding of lacking hydrogens, correction of aromaticity), a warning is given in the remark field of the results.

2.2 Output

Results given as text file consist of a plain-text tabbed file (easily importable and processable by any spreadsheet software) containing in each row all the information about the prediction of a molecule. Note that if some problems were encountered while processing the molecule structure, some warnings are reported in the last field (Remarks).

Results given as PDF file consists of a document containing all the information about the prediction. For each molecule, results are organized in sections with the following order:

1 – Prediction summary

Here is reported a depiction of the compound and the final assessment of the prediction (i.e. the prediction made together with the analysis of the applicability domain). Note that if some problems were encountered while processing the molecule structure, some warnings are reported in the last field (Remarks).

A graphical representation of the evaluation of the prediction and of its reliability is also provided, using the following elements:

🟢 Compound is classified as non-mutagen

🟡 Compound is classified as possible non-mutagen

🔴 Compound is classified as mutagen

★☆☆ Prediction has low reliability (compound out of the AD)

★★☆ Prediction has moderate reliability (compound could be out of the AD)

★★★ Prediction has high reliability (compound into the AD)

3.1 – Applicability Domain: Similar compounds, with predicted and experimental values

Here it is reported the list of the six most similar compounds found in the training and test set of the model, along with their depiction and relevant information (mainly experimental value and predicted value).

3.2 – Applicability Domain: Measured Applicability Domain scores

Here it is reported the list of all Applicability Domain scores, starting with the global Applicability Domain Index (ADI). Note that the final assessment on prediction reliability is given on the basis of the value of the ADI. For each index, it is reported its value and a brief explanation of the meaning of that value.

4.1 – Reasoning: Relevant chemical fragments and moieties

If some rare and/or missing Atom Centered Fragments are found, they are reported here with a depiction of each fragment.

If some relevant fragments are found (see section 1.4 and 1.5 of this guide), they are reported here (one for each page) with a brief explanation of their meaning and the list of the three most similar compounds that contain the same fragment. Note that if no relevant fragments are found, this section is not shown.