Guide to Hepatotoxicity IRFMN Model version 1.0.0

Table of Contents

1. Model explanation

1.1 Introduction

The model provides a qualitative prediction of hepatotoxicity (drug-induced liver injury). 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, both extracted with Sarpy software and manually created by expert from a data set of 950 compounds using data from the literature and further tested on an external validation set of 101 compounds. These rules are expressed SMARTS representing molecular fragments, and each rule is related to toxicity or non-toxicity. The development of the model is reported in:

F. Pizzo, A. Lombardo, A. Manganaro, E. Benfenati, "A New Structure-Activity Relationship (SAR) Model for Predicting Drug-Induced Liver Injury, Based on Statistical and Expert-Based Structural Alerts", *Frontiers in Pharmacology* (2016), 7.

If at least one rule is matching with the given compound, a prediction is given, otherwise the model reports a "unknown" activity prediction. When one or more rules are matching, if the number of rules related to toxicity is higher or equal than the number of rules for non-toxicity, a "toxic" prediction is provided; otherwise, a "non-toxic" prediction is provided.

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.

- **Similar molecules with known experimental value**. This index takes into account how similar are the first two 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 two 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 > 0.8	accuracy of prediction for similar molecules found in the training set is good
0.8 >= index > 0.6	accuracy of prediction for similar molecules found in the training set is not optimal
index <= 0.6	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 two 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 > 0.8	similar molecules found in the training set have experimental values that agree with the predicted value
0.8 >= index > 0.6	some similar molecules found in the training set have experimental values that disagree with the predicted value
index <= 0.6	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 >= 0.8	predicted substance is into the Applicability Domain of the model
0.8 > index >= 0.6	predicted substance could be out of the Applicability Domain of the model
index < 0.6	predicted substance is out of the the Applicability Domain of the model

1.4 Structural Alerts for toxic compounds

Following, the list of the 51 rules for hepatotoxicity, expressed as SMARTS strings:

Hepatotoxicity toxic alert no. 1: C(=CC(C)C)CCCCCCC
Hepatotoxicity toxic alert no. 2: O=CC(NC)CO
Hepatotoxicity toxic alert no. 3: O(CCCC)CC(NC)CC
Hepatotoxicity toxic alert no. 4: O=C(NCCCC)c1ccccc1
Hepatotoxicity toxic alert no. 5: OCC(Oc1ccccc1)(C)C
Hepatotoxicity toxic alert no. 6: c1ccc(cc1)SCC
Hepatotoxicity toxic alert no. 7: c1ccc(cc1)CCCc2ccccc2
Hepatotoxicity toxic alert no. 8: O=COC(CC(O)C)C(C)C
Hepatotoxicity toxic alert no. 9: O=CCc1ccccc1(N)
Hepatotoxicity toxic alert no. 10: N(CC)CCCCNCC
Hepatotoxicity toxic alert no. 11: O=C(CO)C(O)CCO
Hepatotoxicity toxic alert no. 12: n2cnc1c(ncn1CCOC)c2N
Hepatotoxicity toxic alert no. 13: O(c1cccc(c1)CNCCCC)C
Hepatotoxicity toxic alert no. 14: O=C(OCCc1ccccc1)C
Hepatotoxicity toxic alert no. 15: OCC(O)C(O)CCCNC
Hepatotoxicity toxic alert no. 16: O=[N+]([O-])c1cccc(O)c1
Hepatotoxicity toxic alert no. 17: c1c(cc(cc1C)CC)C
Hepatotoxicity toxic alert no. 18: O=CNC(CC=O)CCC
Hepatotoxicity toxic alert no. 19: O=CNC(C(=O)O)CCC
Hepatotoxicity toxic alert no. 20: O=SCCNCCC
Hepatotoxicity toxic alert no. 21: O=C(N)N(N)CC
Hepatotoxicity toxic alert no. 22: FC(F)CCl
Hepatotoxicity toxic alert no. 23: O=C(NC)CCc1ccccc1
Hepatotoxicity toxic alert no. 24: O=COC(C)COC
Hepatotoxicity toxic alert no. 25: Oc1ccc(c(OC)c1)C
Hepatotoxicity toxic alert no. 26: N(C)C(C)CNCCCC

Hepatotoxicity toxic alert no. 27: NCCCCCNC
Hepatotoxicity toxic alert no. 28: O=CC(N)Cc1ccccc1
Hepatotoxicity toxic alert no. 29: c1ccc(c(c1)C=CC)C
Hepatotoxicity toxic alert no. 30: n1cc[nH]c1
Hepatotoxicity toxic alert no. 31: NCNc1ccccc1
Hepatotoxicity toxic alert no. 32: O=CC(c1ccccc1)CN
Hepatotoxicity toxic alert no. 33: FC(F)(F)c1ccccc1
Hepatotoxicity toxic alert no. 34: n1ccccc1Cc2ccccc2
Hepatotoxicity toxic alert no. 35: N(CC)CCCl
Hepatotoxicity toxic alert no. 36: c1cc(C)sc1
Hepatotoxicity toxic alert no. 37: C(=C(Cl))
Hepatotoxicity toxic alert no. 38: O=S(=O)(N)c1ccccc1
Hepatotoxicity toxic alert no. 39: Nc1ccc(cc1)S(=O)=O
Hepatotoxicity toxic alert no. 40: [n,o]1n[c,n][c,n,s,nH][c,n]1
Hepatotoxicity toxic alert no. 41: O=C1CCCCCCCCCCCCO1
Hepatotoxicity toxic alert no. 42: [n,c]1ccn[n,c]c1
Hepatotoxicity toxic alert no. 43: CNC(=O)N(CCCl)N=O
Hepatotoxicity toxic alert no. 44: OC(=O)C1[C,S][S,O,C]C2CC(=O)N12
Hepatotoxicity toxic alert no. 45: Nc1[n,c]cc2C(=O)C(=CNc2[c,n]1)C(O)=O
Hepatotoxicity toxic alert no. 46: O=C1N~CC=C[N,C]1C2C~[S,C]CO2
Hepatotoxicity toxic alert no. 47: NS(=O)(=O)c1ccccc1
Hepatotoxicity toxic alert no. 48: C1[S,C,N,O]c2ccccc2[N,C,S,O]c3ccccc13
Hepatotoxicity toxic alert no. 49: *N(*)CCC(c1cccc[n,c]1)c2cccc[n,c]2
Hepatotoxicity toxic alert no. 50: CC=C(C)C=CC=C(C)C=C[R,a]
Hepatotoxicity toxic alert no. 51: Nc1[n,c]cnc2[n,c]cccc12

1.5 Structural Alerts for non-toxic compounds

Following, the list of the 37 rules for non-hepatotoxicity, expressed as SMARTS strings:

Hepatotoxicity NON-toxic alert no. 1: OC(c1ccccc1)c2ccccc2
Hepatotoxicity NON-toxic alert no. 2: O=C(O)CCc1ccc(OC)cc1
Hepatotoxicity NON-toxic alert no. 3: OC1OC(CN)CCC1(N)
Hepatotoxicity NON-toxic alert no. 4: OCCNC(C)(C)C
Hepatotoxicity NON-toxic alert no. 5: O=C(OC2CCC3C4CCc1cc(O)ccc1C4(CCC23(C)))CC
Hepatotoxicity NON-toxic alert no. 6: O=C(N(c1ccccc1)CC)C
Hepatotoxicity NON-toxic alert no. 7: O=C(Nc1ccccc1C)CC
Hepatotoxicity NON-toxic alert no. 8: O=COCC(C)(C)COC=O
Hepatotoxicity NON-toxic alert no. 9: Oc4ccc1c2c4(OC3CCCC(O)(C(N(C)CC)C1)C23(C))
Hepatotoxicity NON-toxic alert no. 10: O=C(c1ccccc1)c2ccccc2(O)
Hepatotoxicity NON-toxic alert no. 11: c1ccc(cc1)Cc2ccccc2Cl
Hepatotoxicity NON-toxic alert no. 12: O=C(NC)CNC(=O)C(CC)CCC
Hepatotoxicity NON-toxic alert no. 13: OC(c1cccc(OC)c1)CNC
Hepatotoxicity NON-toxic alert no. 14: O(c1ccccc1)CCN(C)CC
Hepatotoxicity NON-toxic alert no. 15: Oc1ccc(cc1)NCCCC

Hepatotoxicity NON-toxic alert no. 16: OCCC1(C)(CCCCC1)
Hepatotoxicity NON-toxic alert no. 17: c1ccc(c(c1)CCC)Cl
Hepatotoxicity NON-toxic alert no. 18: C1C2CC3CC1CC(C2)C3
Hepatotoxicity NON-toxic alert no. 19: N1CN(C)CC(C)C1
Hepatotoxicity NON-toxic alert no. 20: OCCSCCC
Hepatotoxicity NON-toxic alert no. 21: O=S(=O)(NC)C
Hepatotoxicity NON-toxic alert no. 22: O=Cc1cccc(c1)NCC
Hepatotoxicity NON-toxic alert no. 23: OCCC(CO)CCO
Hepatotoxicity NON-toxic alert no. 24: O=C(N(c1ccccc1C)C)C
Hepatotoxicity NON-toxic alert no. 25: O=CC(C)(C)CN
Hepatotoxicity NON-toxic alert no. 26: CC[N+](C)(C)C
Hepatotoxicity NON-toxic alert no. 27: O=C1CCCCC1
Hepatotoxicity NON-toxic alert no. 28: c1ccc(cc1)I
Hepatotoxicity NON-toxic alert no. 29: O=C(N)c1cccc(c1)S(=O)=O
Hepatotoxicity NON-toxic alert no. 30: Nc1ccc2ccccc2(c1)
Hepatotoxicity NON-toxic alert no. 31: O(c1ccc(cc1)C)CCC
Hepatotoxicity NON-toxic alert no. 32: OC1OCC(O)C(O)C1(N)
Hepatotoxicity NON-toxic alert no. 33: c1cc(c(cc1Cl)Cl)C
Hepatotoxicity NON-toxic alert no. 34: NN=C
Hepatotoxicity NON-toxic alert no. 35: CC(=O)Nc1ccccc1C
Hepatotoxicity NON-toxic alert no. 36: CC(=O)NC1C2[S,O]CC=C(N2C1=O)C(O)=O
Hepatotoxicity NON-toxic alert no. 37: C1CC2CCC3C(CC[C,c]4[C,c][C,c][C,c][C,c][C,c]34)C2C1

1.6 Model statistics

Following, statistics obtained applying the model to its original dataset, where the 101 compounds from the external validation set have been included in the test set:

- Training set: n = 760; Predicted = 495; Accuracy = 0.82; Specificity = 0.67; Sensitivity = 0.94
- Test set: n = 291; Predicted = 157; Accuracy = 0.66; Specificity = 0.37; Sensitivity = 0.85

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-toxic

Compound is classified as toxic

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.