	QMRF identifier (JRC Inventory): To be entered by JRC
	QMRF Title: P-Glycoprotein Activity Classification Model (NIC) version 1.0.1
	Printing Date: November 2022

1.QSAR identifier

1.1.QSAR identifier (title):

P-Glycoprotein Activity Classification Model (NIC) (version 1.0.1)

1.2.Other related models:

NA

1.3.Software coding the model:

VEGA (https://www.vegahub.eu/)

The VEGA software provides QSAR models to predict tox, ecotox, environ, phys-chem and toxicokinetic properties of chemical substances.

emilio.benfenati@marionegri.it

2.General information

2.1.Date of QMRF:

November 2022

2.2.QMRF author(s) and contact details:

[1] Liadys Mora Lagares National Institute of Chemistry NIC liadys.moralagares@ki.sihttps://www.researchgate.net/profile/Liadys-Mora-Lagares

[2] Marjana Novi National Institute of Chemistry NIC marjana.novic#ki.si

[3] Erika Colombo Istituto di Ricerche Farmacologiche Mario Negri - erika.colombo@marionegri.it

2.3.Date of QMRF update(s):

NA

2.4.QMRF update(s):

NA

2.5.Model developer(s) and contact details:

[1] Liadys Mora Lagares National Institute of Chemistry NIC liadys.moralagares@ki.sihttps://www.researchgate.net/profile/Liadys-Mora-Lagares

[2]Marjana Novi National Institute of Chemistry NIC marjana.novic@ki.si

[3]Nikola Minovski National Institute of Chemistry NIC nikola.minovski@ki.si

[4]Alberto Manganaro Kode srl info@kode-solutions.net

2.6.Date of model development and/or publication:

2019

2.7.Reference(s) to main scientific papers and/or software package:

[1] Multiclass Classifier for P-Glycoprotein Substrates, Inhibitors, and Non-Active Compounds https://www.mdpi.com/1420-3049/24/10/2006

[2] Homology Modeling of the Human P-glycoprotein (ABCB1) and Insights into Ligand Binding through Molecular Docking Studies https://www.mdpi.com/1422-0067/21/11/4058

[3] Benfenati E, Manganaro A, Gini G. VEGA-QSAR: AI inside a platform for predictive toxicology Proceedings of the workshop "Popularize Artificial Intelligence 2013", December 5th 2013, Turin, Italy Published on CEUR Workshop Proceedings Vol-1107

2.8.Availability of information about the model:

The model is non-proprietary and the training set is available.

2.9.Availability of another QMRF for exactly the same model:

Yes - please refer to Mansouri et al. 2016

3.Defining the endpoint - OECD Principle 1

3.1.Species:

Different species (Human, rodents)

3.2.Endpoint:

[1][2][3]QMRF 6. Other QMRF 6. 6. Other

3.3.Comment on endpoint:

The model provides a qualitative prediction of P-Glycoprotein inhibition/substrate activity

3.4.Endpoint units:

Inhibitor/Substrate/Non-active

3.5.Dependent variable:

NA

3.6.Experimental protocol:

NA

3.7.Endpoint data quality and variability:

The dataset was collected mainly from the admet SAR database (http://lmmd.ecust.edu.cn/admetsar2) and from the work of Li et al (https://doi.org/10.1021/mp400450m)

4.Defining the algorithm - OECD Principle 2

4.1.Type of model:

P-Glycoprotein Activity Classification Model (NIC) is Counter Propagation Artificial Neural Network (CP ANN) Multiclass classification model

4.2.Explicit algorithm:

Counter Propagation Artificial Neural Network (CP ANN) in combination with

The genetic algorithm (GA)Mora Lagares, L., Minovski, N., & Novic, M. (2019). Multiclass Classifier for P-Glycoprotein Substrates, Inhibitors, and Non-Active Compounds. Molecules, 24(10). doi:10.3390/molecules24102006

4.3.Descriptors in the model:

[1]H%: percentage of H atoms

[2]nR07: number of 7-membered rings

[3]D/Dtr11: distance/detour ring index of order 11

[4]MWC01: molecular walk count of order 1

[5]X2A: average connectivity index of order 2

[6]SIC3: Structural Information Content index (neighborhood symmetry of 3-order)

[7]VE1sign_B(s): coefficient sum of the last eigenvector from Burden matrix weighted by I-State

[8]ATSC7m: Centred Broto-Moreau autocorrelation of lag 7 weighted by mass

[9]MATS6v: Moran autocorrelation of lag 6 weighted by van der Waals volume

[10]GATS4s: Geary autocorrelation of lag 4 weighted by I-state

[11]P_VSA_LogP_3: P_VSA-like on LogP, bin 3

[12]P_VSA_ppp_D: P_VSA-like on potential pharmacophore points, D - hydrogen-bond donor[13]nRCOOR: number of esters (aliphatic)

[14]nArCONHR: number of secondary amides (aromatic)

[15]nArCO: number of ketones (aromatic)

[16]H-048: H attached to C2(sp3)/C1(sp2)/C0(sp)

[17]SdsCH: Sum of dsCH E-states

[18]CATS2D_01_DN: CATS2D Donor-Negative at lag 01

[19]CATS2D_05_PP: CATS2D Positive-Positive at lag 05

[20]CATS2D_02_PL: CATS2D Positive-Lipophilic at lag 02

[21]B07[O-F]: Presence/absence of O - F at topological distance 7

[22]F01[C-C]: Frequency of C - C at topological distance 1

[23]F02[C-O]: Frequency of C - O at topological distance 2

[24]F04[C-P]: Frequency of C - P at topological distance 4

[25]F04[C-Br]: Frequency of C - Br at topological distance 4

[26]F07[O-F]: Frequency of O - F at topological distance 7

4.4.Descriptor selection:

Initially, a total of 1,229 2D molecular descriptors were calculated, and their values were normalized. With the intention to eliminate the uninformative descriptors (noise) as well as to prevent over-fitting of the model, a variable reduction was performed on the initial set of descriptors before the modelling. For this reason, the descriptors with constant values as well as those with a standard deviation of less than 0.0001 were removed, as they provide little information for the construction of the model. In addition, descriptors that are orthogonal to each other were identified by pair-wise correlations using the Pearson correlation coefficient; if two descriptors have an absolute correlation coefficient above the desired threshold, only one of them is retained, i.e., redundancy is avoided. Descriptors with an absolute pair correlation coefficient value greater than or equal to 0.95 were removed. To further reduce the likelihood of correlations between descriptors, a Kohonen top-map was used (Drganet al., 2017). In this way, the remaining descriptors were mapped onto a network with a 7 by 7 architecture of neurons using the transpose of the descriptor matrix; two descriptors were selected from each neuron, those with the largest and the shortest Euclidean distance to the central neuron, yielding a final set of 96 molecular descriptors for further use. To further reduce the likelihood of correlations between descriptors, a Kohonen top-map was used (Drganet al., 2017). In this way, the remaining descriptors were mapped onto a network with a 7 by 7 architecture of neurons using the transpose of the descriptor matrix; two descriptors were selected from each neuron, those with the largest and the shortest Euclidean distance to the central neuron, yielding a final set of 96 molecular descriptors for further use

4.5.Algorithm and descriptor generation:

NA

4.6.Software name and version for descriptor generation:

Dragon (Version 7.0.8) software for molecular descriptor calculationhttps://chm.kode-solutions.net

4.7.Chemicals/Descriptors ratio:

1785 chemicals (training set) / 96 descriptors = 18.6

5.Defining the applicability domain - OECD Principle 3

5.1.Description of the applicability domain of the model:

The Applicability Domain (AD) is assessed using the original algorithm implemented within VEGA. An overall AD index is calculated, based on a number of parameters, which relate to the results obtained on similar substances within the training.

ADI is defined in this way for this QSAR model´s predictions:

If 1 ≥ AD index > 0.80, the predicted substance is regarded in the Applicability Domain of the model. It corresponds to “good reliability” of prediction.

If 0.80 ≥ AD index > 0.60, the predicted substance could be out of the Applicability Domain of the model. It corresponds to “moderate reliability” of prediction.

If AD index ≤ 0.60, the predicted substance is regarded out of the Applicability Domain of the model. It corresponds to “low reliability” of prediction.

5.2.Method used to assess the applicability domain:

The Applicability Domain and the chemical similarity is measured with the algorithm developed for VEGA. Full details are in the VEGA website (www.vegahub.eu), including the open access paper describing it [3]. The VEGA AD also evaluates the correctness of the prediction on similar compounds (accuracy), the consistency between the predicted value for the target compound and the experimental values of the similar compounds, the range of the descriptors, and the presence of unusual fragments, using atom centred fragments.

These indices are defined in this way for this QSAR model:

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:

If 1 ≥ index > 0.8, strongly similar compounds with known experimental value in the training set have been found

If 0.8 ≥ index > 0.6, only moderately similar compounds with known experimental value in the training set have been found

If index ≤ 0.6, no similar compounds with known experimental value in the training set have been found

Accuracy (average error) 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 fall 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:

If index < 0.5, accuracy of prediction for similar molecules found in the training set is good

If 0.9 > index ≥ 0.5, accuracy of prediction for similar molecules found in the training set is not optimal

If index ≥ 0.9, 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:

If index < 0.5, molecules found in the training set have experimental values that agree with the target compound predicted value

If 0.9 > index ≥ 0.5, similar molecules found in the training set have experimental values that slightly disagree with the target compound predicted value

If index ≥ 0.9, similar molecules found in the training set have experimental values that completely disagree with the target compound predicted value

Model descriptors range check.

This index checks if the descriptors calculated for the predicted compound are inside the range of descriptors of the training and test set. The index has value 1 if all descriptors are inside the range, 0 if at least one descriptor is out of the range. Defined intervals are:

index = True, descriptors for this compound have values inside the descriptor range of the compounds of the training set

index = False, descriptors for this compound have values outside the descriptor range of the compounds of the training set

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:

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

If 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

If 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

5.3.Software name and version for applicability domain assessment:

VEGA

Included in the VEGA software and automatically displayed when running the model

emilio.benfenati@marionegri.it

https://www.vegahub.eu/

5.4.Limits of applicability:

The model is not applicable to inorganic chemicals and substances containing unusual elements (i.e., different from C, O, N, S, P, Cl, Br, F, I). Salts can be predicted only if converted to the neutralized form.

6.Internal validation - OECD Principle 4

6.1. Availability of the training set:

Yes

6.2. Available information for the training set:

CAS RN: No

Chemical Name: No

Smiles: Yes

Formula: No

INChI: No

MOL file: No

6.3. Data for each descriptor variable for the training set:

All

6.4. Data for the dependent variable for the training set:

All

6.5. Other information about the training set:

The training set contains 1785 compounds.

6.6. Pre-processing of data before modelling:

To increase the size of the dataset and to extend the chemical space, we collected and included compounds derived from different references which use different types of experimental assays to assess the P-gp class.

Therefore, before constructing the model, pre-processing of the data was required to detect duplicate compounds and compounds with both experimental classes (or overlapping classified compounds).

The P-gp non-inhibitor and non-substrate compounds were merged into the non-active class. The overlap of negative compounds in both sets was desirable, so they were included in the non-active class, while all other overlaps that might introduce uncertainty into the model (S/I = 42S/NI = 29 I/NS = 10; S: substrate I: inhibitor NI: non-inhibitor NS: non-substrate) were removed. The final dataset includes 2,512structurally diverse compounds, e.g., acridone derivatives, flavonoids, azoles, antidepressants of the selective serotonin reuptake inhibitor (SSRI) class, persistent organic pollutants (POPs), B-lactam antibiotics, and benzodiazepines, among others, which can be divided into three main classes, i.e., 1,178 P-gp inhibitors, 477 substrates, and 857 non-active compounds. The data curation was mainly performed utilizing the software Pipeline Pilot 9.2(Accelrys, 2014). To facilitate the data curation, it was necessary to convert the original SMILES notations into a uniform representation, running a Pipeline Pilot protocol. The protocol used includes the Canonical SMILES component, which adds canonical smiles as a new property to the dataset. All newly generated SMILES were then combined into a single SDF file format along with their Pgp class notation. Duplicate compounds were identified and removed from further analysis byrunning a pipeline pilot protocol that includes the Remove Duplicate Molecules component. In this component, the canonical SMILES was set as a filter to find duplicates. Additionally, compounds classified as belonging to more than one class, defined as overlapping compounds, were also discarded from the analysis. After running the pipeline pilot protocol, some duplicate compounds were still present in the dataset. Removal of the remaining duplicates and overlapping compounds was performed manually based on the descriptor values for the molecules

6.7. Statistics for goodness-of-fit:

Following, statistics obtained applying the model to its original dataset, where the test set and validation set from the original work have been merged into a single test set. Considering only the “Inhibitor” class vs the remaining (“Substrate” and “Non active”):

Training set: n = 1777 (+ 8 non predicted compounds)

Accuracy = 0.95 Specificity = 0.95 Sensitivity = 0.95

Considering only the “Substrate” class vs the remaining (“Inhibitor” and “Non active”): Training set: n = 1777 (+ 8 non predicted compounds)

Accuracy = 0.97 Specificity = 0.98 Sensitivity = 0.92

Considering three classes: Accuracy 0.93

	Reference		
Prediction	Inhibitor	Substrate	Non Active
Inhibitor	799	13	37
Substrate	14	306	19
Non Active	24	14	551

6.8.Robustness - Statistics obtained by leave-one-out cross-validation:

NA

6.9.Robustness - Statistics obtained by leave-many-out cross-validation:

NA

6.10.Robustness - Statistics obtained by Y-scrambling:

NA

6.11.Robustness - Statistics obtained by bootstrap:

NA

6.12.Robustness - Statistics obtained by other methods:

NA

7.External validation - OECD Principle 4

7.1.Availability of the external validation set:

Yes

7.2.Available information for the external validation set:

CAS RN: No

Chemical Name: No

Smiles: Yes

Formula: No

INChI: No

MOL file: No

7.3.Data for each descriptor variable for the external validation set:

All

7.4.Data for the dependent variable for the external validation set:

All

7.5.Other information about the external validation set:

The external validation set counts 726 compounds

7.6.Experimental design of test set:

The global dataset was splitted utilizing the Kohonen mapping as implemented in the CPANN at NIC software (Drganet al., 2017). The dataset was mapped onto the network and the validation set, containing 385 compounds was selected and not used during the model construction and optimization procedures. The training and test sets were selected from the remaining compounds; 1,785 compounds were chosen for training set and 341 for test set. The network parameters used for mapping the dataset were: 20 × 20 neurons, 100 learning epochs, maximum learning rate 0.5 and minimum learning rate 0.01

7.7.Predictivity - Statistics obtained by external validation:

Following, statistics obtained applying the model to its original dataset, where the test set and validation set from the original work have been merged into a single test set. Considering only the “Inhibitor” class vs the remaining (“Substrate” and “Non active”): Test set: n = 717 (+ 9 non predicted compounds) Accuracy = 0.85 Specificity = 0.86 Sensitivity = 0.84

Considering only the “Substrate” class vs the remaining (“Inhibitor” and “Non active”): Test set: n = 717 (+ 9 non predicted compounds) Accuracy = 0.88 Specificity = 0.93 Sensitivity = 0.69

Considering all three classes:

Not predicted 18, Accuracy 0.80

	Reference		
Prediction	Inhibitor	Substrate	Non Active
Inhibitor	278	22	31
Substrate	19	92	19
Non Active	36	21	192

7.8.Predictivity - Assessment of the external validation set:

NA

7.9.Comments on the external validation of the model:

NA

8.Providing a mechanistic interpretation - OECD Principle 5

8.1.Mechanistic basis of the model:

The classification model is based on a structural similarity approach represented by the model's descriptors

8.2.A priori or a posteriori mechanistic interpretation:

NA

8.3.Other information about the mechanistic interpretation:

NA

9.Miscellaneous information

9.1.Comments:

The training set is available on the Vega documentation

9.2.Bibliography:

[1] Mora Lagares, L., Minovski, N., & Novic, M. (2019). Multiclass Classifier for P-Glycoprotein Substrates, Inhibitors, and Non-Active Compounds. Molecules, 24(10). doi:10.3390/molecules24102006

[2]Mora Lagares, L., Minovski, N., Caballero Alfonso, A. Y., Benfenati, E., Wellens, S., Culot, M., . . .Novi, M. (2020). Homology modeling of the human P-glycoprotein (ABCB1) and insights into ligandbinding through molecular docking studies. International journal of molecular sciences, 21(11), 4058. doi: 10.3390/ijms21114058

[3] Floris, Matteo, Alberto Manganaro, Orazio Nicolotti, Ricardo Medda, Giuseppe Felice Mangiatordi, e Emilio Benfenati. «A generalizable definition of chemical similarity for read-across». Journal of Cheminformatics 6, n. 1 (18 october 2014): 39. https://doi.org/10.1186/s13321-014-0039-1

9.3.Supporting information:

Training set(s)Test set(s)Supporting information:

All available dataset are present in the model inside the VEGA software.

10.Summary (JRC QSAR Model Database)

10.1.QMRF number:

To be entered by JRC

10.2.Publication date:

To be entered by JRC

10.3.Keywords:

To be entered by JRC

10.4.Comments:

To be entered by JRC