	QMRF identifier (JRC Inventory): **To be entered by JRC**
	QMRF Title: **ADMET Predictor - Bacterial mutagenicity model (MUT_98)**
	Printing Date: **Jul 22, 2021**

1.QSAR identifier

1.1.QSAR identifier (title):

ADMET Predictor - Bacterial mutagenicity model (MUT_98)

1.2.Other related models:

The bacterial mutagenicity panel within ADMET Predictor Toxicity Module features a series of 10 MUT_*** models that predict Ames Mutagenicity in 5 individual strains of Salmonella (and/or E.coli), with or without metabolic activation, i.e.: MUT_97+1537; MUT_m97+1537; MUT_98; MUT_m98; MUT_100; MUT_m100; MUT_102+wp2; MUT_m102+wp2; MUT_1535; MUT_m1535. The ten TOX_MUT* Artificial Neural Network Ensembles (ANNE) are qualitative models, predicting the mutagenicity of new compounds as “Positive” (i.e., mutagenic) or “Negative”. Two additional mutagenicity models are included: i) ADMET Risk™ rule file, called "MUT_Risk", which predicts overall mutagenicity by counting instances of “Positive"; ii) "MUT_NIHS", classification model based on the proprietary Ames database provided by the Division of Genetics and Mutagenesis, National Institute of Health Sciences of Japan (DGM/NIHS).

1.3.Software coding the model:

ADMET Predictor 10.0
ADMET property prediction and QSAR model-building application
Simulations Plus, Inc., 42505 10th Street West, Lancaster, 93534-7059, CA, USA.
https://www.simulations-plus.com/software/admetpredictor/

2.General information

2.1.Date of QMRF:

2 February 2021

2.2.QMRF author(s) and contact details:

Simona Kovarich TOXIT of S-IN Soluzioni S.r.l. Via Ferrari 14, I-36100 Vicenza (Italy)
simona.kovarich@s-in.it https://www.toxit.it/en/; http://www.s-in.it

2.3.Date of QMRF update(s):

n/a

2.4.QMRF update(s):

n/a

2.5.Model developer(s) and contact details:

Pankaj Daga Simulation Plus, Inc. 42505 10th Street West, Lancaster, 93534- 7059, CA, USA
pankaj@simulations-plus.com https://www.simulations-plus.com/

2.6.Date of model development and/or publication:

The model was developed in June of 2017 and was first released in ADMET Predictor 8.5. The model is currently implemented in ADMET Predictor 10.0 (2020).

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

[1]ADMET Predictor 10.0 https://www.simulations-plus.com/software/admetpredictor/
[2]Ghosh J, Lawless MS, Waldman M, Gombar V, Fraczkiewicz R (2016) Modeling ADMET. Methods Mol Biol. 1425:63-83. https://pubmed.ncbi.nlm.nih.gov/27311462/
[3]Clark R.D., Daga P.R. (2019) Building a Quantitative Structure-Property Relationship (QSPR) Model. In: Larson R., Oprea T. (eds) Bioinformatics and Drug Discovery. Methods in Molecular Biology, vol 1939. Humana Press, New York, NY.

2.8.Availability of information about the model:

The model is proprietary and implemented in the commercial software ADMET Predictor (by Simulation Plus). Training and test set are not publicly available.

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

No

3.Defining the endpoint - OECD Principle 1

3.1.Species:

Salmonella typhimurium (strain TA98)

3.2.Endpoint:

TOX 7.6.1. Genetic toxicity in vitro

3.3.Comment on endpoint:

Mutagenicity assessment based on bacterial reverse mutation test using Salmonella typhimurium strain TA98, without metabolic activation.

3.4.Endpoint units:

Unitless

3.5.Dependent variable:

Qualitative variable: the model classifies compounds as "Positive" (mutagenic) or "Negative" (non-mutagenic).

3.6.Experimental protocol:

The source of the dataset is Bacha et al. (2002) [4], according to which bacterial mutagenicity data were primarily obtained from the Chemical Carcinogenesis Research Information System (CCRIS). This toxicology data file is maintained by the National Cancer Institute and made public through the National Library of Medicine’s Toxicology Data Network (TOXNET). It is a scientifically evaluated and fully referenced database with mutagenicity results for individual bacterial indicator strains (S. typhimurium TA97, TA1537, TA98, TA100, TA1535, and TA102 and E. coli WP2 uVrA) with and without addition of rat liver microsomal preparation to measure metabolic activation. These data were supplemented both with information from the Genetic Activity Profile database maintained by the Environmental Protection Agency in association with the International Agency for Research on Cancer and with data from a series of literature references. Only data referred to S. typhimurium strain TA98, without metabolic activation, were considered for model development.
Curation of chemical structures was performed automatically and/or manually within the ADMET Modeler/Predictor, and included the following:
i) extraction of the active moiety from salts and other multicomponent

compounds; ii) standardization of substructural representations (e.g., nitro groups); iii) standardization of tautomers (rule-based system that strikes a balance between consistency and accuracy; the microstate analysis tool was used to check cases where automatic tautomer assignments were questionable).

Curation of mutagenicity data was performed automatically and/or manually within the ADMET Predictor, and included the following: i) removal of duplicate entries (based on shared name or structure or based on tautomeric equivalence), eliminating all but one example that represents a consensus of the replicates; ii) handling of structures with conflicting results (positive and negative) from different data sources: data are further verified for correctness analysing the original data source(s) (e.g., journal articles); if the conflict can't be resolved, then the records are removed.

3.7.Endpoint data quality and variability:

Endpoint quality was dependent on the original literature. Experimental variability was not taken directly into account, but is known historically to be about 85% between labs.

4.Defining the algorithm - OECD Principle 2

4.1.Type of model:

QSAR

4.2.Explicit algorithm:

ANNE (Artificial Neural Network Ensembles)

Ensemble of 33 artificial neural network sub-models sharing the same "architecture" (i.e. same inputs and number of neurons). A total of 165 models are initially created and the best 33 are selected for the ensemble. More detail on the ANNE approach can be found in the ADMET Predictor Manual [1].

4.3.Descriptors in the model:

The list of modeling descriptors and their description is provided in Supporting Information

4.4.Descriptor selection:

The ADMET Modeler module (which is part of the ADMET Predictor) automates the key steps necessary for model building, including the reduction and selection of descriptors. In a first stage, the number of descriptors is reduced based on their own properties and how they relate to other descriptors ("unsupervised" process). The filtering of molecular descriptors is performed to eliminate those that are underrepresented, those with very small variance, and those that are highly correlated with other descriptors. The second, "supervised" stage takes the relationship of the descriptors to the dependent variable - their "sensitivity" - into account in prioritizing them for incorporation during model building. For the building of MUT_98 model the Genetic algorithm (once per cell) method was selected for variable selection. More detail on this method can be found in the ADMET Predictor Manual [1].

4.5.Algorithm and descriptor generation:

Molecular descriptors were calculated within ADMET Predictor software. ADMET Predictor generates 341 molecular descriptors from 2D structures, including Textual Description descriptors and indicators (not used for modeling), Simple Constitutional descriptors, Topological Indices, Atom-type Electrotopological State indices, Charge-based descriptors, Hydrogen bonding descriptors, Molecular Ionization descriptors and Functional groups (a description of the available descriptors is provided within the ADMET Predictor Manual [1]).

4.6.Software name and version for descriptor generation:

ADMET Predictor
ADMET property prediction and QSAR model-building application
Simulations Plus, Inc., 42505 10th Street West, Lancaster, 93534-7059, CA, USA.
https://www.simulations-plus.com/software/admetpredictor/

4.7.Chemicals/Descriptors ratio:

2892 compounds / 32 descriptors

5.Defining the applicability domain - OECD Principle 3

5.1.Description of the applicability domain of the model:

DESCRIPTOR DOMAIN: applicability domain defined by the descriptor space of training set compounds (hypercubes in the model's standardized space). Predictions computed for compounds lying outside the applicability domain of the model should be assessed as low reliable.
RESPONSE DOMAIN: positive/negative

5.2.Method used to assess the applicability domain:

The applicability domain is defined by hypercubes in the model's standardized space: the range of training set values for each descriptor used in the model is mapped to the interval [0,1]. A compound for which any of those descriptors is below -0.1 or above 1.1 is flagged as "out of scope" - i.e., as lying outside the applicability domain of the model. The prediction for such a compound may be correct but it would be unwise to put much faith in it (i.e., low reliable prediction).

5.3.Software name and version for applicability domain assessment:

ADMET Predictor 10.0
ADMET property prediction and QSAR model-building application
Simulations Plus, Inc., 42505 10th Street West, Lancaster, 93534-7059, CA, USA.
https://www.simulations-plus.com/software/admetpredictor/

5.4.Limits of applicability:

The limits of the model's applicability domain are defined by the DESCRIPTOR SPACE of training set compounds (see section 5.2). For new compounds, the standardised modelling descriptors' values should fall within the interval [0,1]; if any of those descriptors is below -0.1 or above 1.1, the compound is flagged as "out of scope" (i.e. outside the applicability domain of the model).

6.Internal validation - OECD Principle 4

6.1.Availability of the training set:

No

6.2.Available information for the training set:

CAS RN: No

Chemical Name: No

Smiles: No

Formula: No

INChI: No

MOL file: No

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

No

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

No

6.5.Other information about the training set:

The training set consists of 2892 compounds, including 686 positive compounds (ca. 24%) and 2206 negative compounds.

For model development, the training pool was further randomly splitted into a training set (66%) and a verification set (33%), with each individual model in the ensemble "seeing" a different split ("fold").

The verification set is involved in building individual models, albeit only indirectly: for early stopping and parameter setting in ANNE classification model. More critically, predictive performance on the verification set is used to determine which models get included in the final ensemble. Hence verification statistics are a better indicator of what to expect for compounds from outside the data set than training set statistics.

6.6.Pre-processing of data before modelling:

Pre-processing of mutagenicity data was performed automatically and/or manually within the ADMET Predictor, and included the following: i) removal of duplicate entries (based on shared name or structure or based on tautomeric equivalence), eliminating all but one example that represents a consensus of the replicates; ii) handling of structures with conflicting results (positive and negative) from different data sources: data are further verified for correctness analysing the original data source(s) (e.g., journal articles); if the conflict can't be resolved, then the records are removed.

6.7.Statistics for goodness-of-fit:

Concordance= 88.1%; Sensitivity = 82.7%; Specificity = 89.8%

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

n/a

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

n/a

6.10.Robustness - Statistics obtained by Y-scrambling:

n/a

6.11.Robustness - Statistics obtained by bootstrap:

n/a

6.12.Robustness - Statistics obtained by other methods:

The model had an acceptable uncertainty profile (Clark et al. J Cheminfo 2014, 6(1), 1-19.).

7.External validation - OECD Principle 4

7.1.Availability of the external validation set:

No

7.2.Available information for the external validation set:

CAS RN: No

Chemical Name: No

Smiles: No

Formula: No

INChI: No

MOL file: No

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

No

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

No

7.5.Other information about the external validation set:

The test set consists of 723 compounds, including 155 positive compounds (ca. 21.4%) and 568 negative compounds.

7.6.Experimental design of test set:

The dataset was partitioned into a training pool (i.e., subset of compounds that are used to train the model) and a test set (i.e., group of compounds that are set aside before training begins, and is not involved in the training in any way). The splitting was performed by random selection of test set chemicals from the dataset.

7.7.Predictivity - Statistics obtained by external validation:

Concordance= 85.9%; Sensitivity = 80%; Specificity = 87.5%

7.8.Predictivity - Assessment of the external validation set:

The partitioning of the dataset was aimed at maximizing the size and diversity of training and test set. The external validation set is considered as sufficiently large (since represents nearly 20% of the data set) and representative of the applicability domain, especially considering the response representation (training and test set exhibit a similar balance between positive and negative compounds).

7.9.Comments on the external validation of the model:

n/a

8.Providing a mechanistic interpretation - OECD Principle 5

8.1.Mechanistic basis of the model:

The high number of modeling descriptors involved in ANNE models doesn't allow an easy mechanistic interpretation of the model. However, ADMET Predictor allows the user to perform descriptor sensitivity analysis for

this model [1].

The DESCRIPTOR SENSITIVITY ANALYSIS (DSA) allows the user to explore the relationship between one specific descriptor and model output in detail, for one data record at a time. For classification models, the sensitivities can be visualized with the Gradient Bar graph. Within this plot, the bars show the average sensitivities of descriptors used by the selected model for the selected molecule. The direction of the bars shows the sign of the sensitivity while the size of the bars shows the magnitude of the sensitivity. The longer the bar is, the greater the impact the respective descriptor has on the prediction of the selected molecule. By default, descriptors are automatically sorted by the magnitude of their sensitivity.
Calculation of sensitivity for binary classification models: first, the program calculates the minimal change d, either positive or negative or both, of a given descriptor “d” to flip the current prediction for the selected molecule. The smaller the change d is, the higher the descriptor d sensitivity. Therefore, the d’s sensitivity is defined as the reciprocal of the minimal flipping change: S = 1/(d+1). Thus, if the DSA for a particular descriptor is positive, then increasing the value of the descriptor may cause the prediction to flip. On the other hand, if the sensitivity bar is negative then decreasing the magnitude of the descriptor may cause the prediction to be flipped. However, since a descriptor can go both ways to flip a prediction, some descriptors will have both negative and positive bars.

8.2.A priori or a posteriori mechanistic interpretation:

A posteriori mechanistic interpretation, with the support of DESCRIPTOR SENSITIVITY ANALYSIS window available within the ADMET Predictor (see section 8.1)

8.3.Other information about the mechanistic interpretation:

n/a

9.Miscellaneous information

9.1.Comments:

n/a

9.2.Bibliography:

[1]Simulation Plus. ADMET Predictor Manual. September 2020 ver. 10.0
[2]Ghosh J, Lawless MS, Waldman M, Gombar V, Fraczkiewicz R (2016) Modeling ADMET. Methods Mol Biol. 1425:63-83. https://pubmed.ncbi.nlm.nih.gov/27311462/
[3]Clark R.D., Daga P.R. (2019) Building a Quantitative Structure-Property Relationship (QSPR) Model. In: Larson R., Oprea T. (eds) Bioinformatics and Drug Discovery. Methods in Molecular Biology, vol 1939. Humana Press, New York, NY.
[4]Bacha PA, Gruver HS, Den Hartog BK, Tamura SY, Nuttet RF (2002) Rule Extraction from a Mutagenicity Data Set Using Adaptively Grown Phylogenetic-like Trees. J. Chem. Inf. Comput. Sci. 42, 1104-1111.
[5]Lawless M and Clark RD (2016) QSAR-based Prediction of Ames Mutagenicity for ICH M7

Submissions. Poster presented at Society of Toxicology (SOT) 55th Annual Meeting. New Orleans, Louisiana. March 13-17 2016.

9.3.Supporting information:

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

SI_MUT_98-descriptors.xlsx	file:///C:\Users\utente\Documents\S+_QMRF_SI\SI_MUT_98-descriptors.xlsx

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