

QMRF DOSSIER

**ProtoQSAR model for *in vivo*
mammalian somatic cell genotoxicity
study (micronucleus assay)**



www.protoqsar.com



Parque Tecnológico de Valencia
Carrer de Nicolau Copèrnic 6
46980 Paterna (Valencia, Spain)



protopred@protoqsar.com



+34 962 021 811



QMRF: ProtoQSAR model for *in vivo* mammalian somatic cell genotoxicity study (micronucleus assay) (v1.0)

1. QSAR identifier

1.1. QSAR identifier (title):

ProtoQSAR model for *in vivo* mammalian somatic cell genotoxicity study (micronucleus assay) (v1.0)

1.2. Other related models:

None

1.3. Software coding the model:

ProtoPRED[®] (ProtoQSAR proprietary software) v1.0
<https://protoqsar.com>

2. General information

2.1. Date of QMRF:

7th July 2022

2.2. QMRF author(s) and contact details:

- [1] Moncho, Salvador
- [2] Goya, Addel
- [3] Serrano-Candelas, Eva
- [4] Gozalbes, Rafael

ProtoQSAR S.L.

+34 96 202 18 11

protopred@protoqsar.com

2.3. Date of QMRF update(s):

15th October 2025

2.4. QMRF update(s):

Some texts have been improved to clarify the information provided and enhance the relationship with regulatory applications. Additionally, the model's name has been revised to align with those used by regulatory bodies.

2.5. Model developer(s) and contact details:

- [1] Vallés-Pardo, José Luis
- [2] Gómez-Ganau, Sergi
- [3] Roca-Martínez, Joel
- [4] Barigye, Stephen Jones

[5] Serrano-Candelas, Eva

[6] Gozalbes, Rafael

ProtoQSAR S.L.

Contact: Parque Tecnológico de Valencia. Carrer de Nicolau Copèrnic 6, Local 8. 46980 Paterna (Valencia)

+34 96 202 18 11

protopred@protoqsar.com

2.6. Date of model development and/or publication:

January 2022

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

Not published.

2.8. Availability of information about the model:

The model and the algorithm are proprietary, but the dataset is non-proprietary and is available upon request.

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

None.

3. Defining the endpoint - OECD Principle 1

3.1. Species:

Rodents

3.2. Endpoint:

Human health effects: Mutagenicity/Genotoxicity. Mammalian erythrocyte micronucleus test.

3.3. Comment on endpoint:

Mutagenicity refers to the induction of permanent transmissible changes in the amount or structure of the genetic material of cells or organisms. Genotoxicity is a broader term and refers to processes which alter the structure, information content or segregation of DNA and are not necessarily associated with mutagenicity. *In vivo* mammalian erythrocyte micronucleus test is a measure of *in vivo* chromosomal mutagenicity. The test identifies substances that cause micronuclei in erythroblasts sampled from bone marrow and/or peripheral blood cells of animals, usually rodents. These micronuclei originate from acentric fragments or whole chromosomes, and the test thus has the potential to detect both clastogenic and aneugenic substances (see reference for "Guidance on information requirements and chemical safety assessment, Chapter R.7a" in section 9.2).

3.4. Endpoint units:

N/A

3.5. Dependent variable:

The dependent variable for modelling purposes is a binary classification in two categories. Original data was retrieved as a binary classification: positive (cytotoxic) / negative (non-cytotoxic).

3.6. Experimental protocol:

Endpoint following the OECD: Test No. 474: Mammalian erythrocyte micronucleus test

Animals are exposed to the test chemical by an appropriate route. If bone marrow is used, the animals are humanely euthanised at an appropriate time(s) after treatment, the bone marrow is extracted. When peripheral blood is used, the blood is collected at an appropriate time(s) after treatment. Preparations are made and stained in both cases. When treatment is administered acutely, it is important to select bone marrow or blood harvest times at which the treatment-related induction of micronucleated immature erythrocytes can be detected. In the case of peripheral blood sampling, enough time must also have elapsed for these events to appear in circulating blood. Preparations are analysed for the presence of micronuclei, either by visualisation using a microscope, image analysis, flow cytometry, or laser scanning cytometry.

3.7. Endpoint data quality and variability:

The data for developing the model was extracted from ISSMIC public database retrieved from QSAR Toolbox. After curation and preprocessing the database is formed by 272 experimental results, with a 56.6% of positive values (154) and a 43.4% of negative values (118).

4. Defining the algorithm - OECD Principle 2

4.1. Type of model:

QSAR

4.2. Explicit algorithm:

Support Vector Machine Classifier (SVC). Support Vector Machines (SVM) are a set of supervised learning methods used for classification and regression problems. The SVM classify the data by finding the hyperplane that maximizes the margin between the two classes. The vectors (cases) that define the hyperplane are the support vectors.

4.3. Descriptors in the model:

- **C-016**: =CHR.
- **ATSC5i**: Centred Broto-Moreau autocorrelation of lag 5 (log function) weighted by ionization potential.
- **AATSC0se**: Averaged centred Broto-Moreau autocorrelation of lag 0 (log function) weighted by Sanderson electronegativity.
- **AATSC5i**: Averaged centred Broto-Moreau autocorrelation of lag 5 (log function) weighted by ionization potential.
- **MATS4p**: Moran autocorrelation of lag 4 (log function) weighted by polarizability.
- **GATS4i**: Geary autocorrelation of lag 4 (log function) weighted by ionization potential.
- **B02(C-C)**: Presence/absence of C-C at topological distance 2.
- **B02(C-O)**: Presence/absence of C-O at topological distance 2.

- **B04(C-O)**: Presence/absence of C-O at topological distance 4.
- **B04(O-O)**: Presence/absence of O-O at topological distance 4.
- **B08(C-N)**: Presence/absence of C-N at topological distance 8.
- **B09(O-O)**: Presence/absence of O-O at topological distance 9.
- **EState_VSA6**: EState VSA descriptor 6.

4.4. Descriptor selection:

The descriptor selection is performed by eliminating non-variant descriptors, as well as filtering collinear descriptors ($R^2 > 0.9$). Afterwards, by Recursive Feature Elimination (RFE) based on Support Vector Machine (SVM) using linear kernel ($C = 1$), the number of descriptors was reduced based on their correlation with the values of the independent variable.

4.5. Algorithm and descriptor generation:

Descriptors are calculated by an in-house software module in which these are implemented as described in Todeschini & Consonni, 2009 and Consonni & Todeschini, 2010 (full references in section 9.2).

4.6. Software name and version for descriptor generation:

ProtoPRED[®] (ProtoQSAR proprietary software) v1.0

4.7. Chemicals/Descriptors ratio:

Ratio: 203/13 = 15.62

5. Defining the applicability domain - OECD Principle 3

5.1. Description of the applicability domain of the model:

The applicability domain is defined by the training set based on several criteria:

- **by chemical similarity:**

Evaluated by Tanimoto-Jaccard similarity index based on molecular fingerprints (≥ 0.528)

- **by molecular descriptors by three different criteria:**

Evaluated by the Leverage of model descriptors (≤ 0.19).

Evaluated by the Euclidean distance of model descriptors (see 4.3).

Evaluated by the range of values for each descriptor.

5.2. Method used to assess the applicability domain:

- **Tanimoto**: The Tanimoto-Jaccard coefficient allows to compare the structural similarity of two chemical structures by computing a set of MACCS fingerprints for each chemical compound. A value from 0 to 1 is obtained, where 1 corresponds to identical structures and is closer to zero if they are very different.

- **Euclidean distance**: The Euclidean distance is a measure of the separation between two points in Euclidean space. We compute the distance of the descriptor values of the molecule to the descriptor values of the molecules present in the training set and determine if it is inside the applicability domain or not.

- **Leverage:** The leverage of a compound measures the distance of this compound to the structural centroid of the training set and is a measure of its influence on the model.
- **Descriptors range:** The range of standardised values for each descriptor in the train is independently evaluated by determining its maximum and minimum values. External values are standardised using the same rules and compared with the maximum and minimum value.

5.3. Software name and version for applicability domain assessment:

ProtoPRED[®] (ProtoQSAR proprietary software) v1.0

5.4. Limits of applicability:

The model was built only for discrete organic chemicals. A prediction is considered to fall outside the AD if it does not match any of the criteria specified in QMRF section 5.2.

6. Internal validation - OECD Principle 4

6.1. Availability of the training set:

The curated training and validation sets are not included as supporting information, but they can be provided upon request for regulatory assessment.

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:

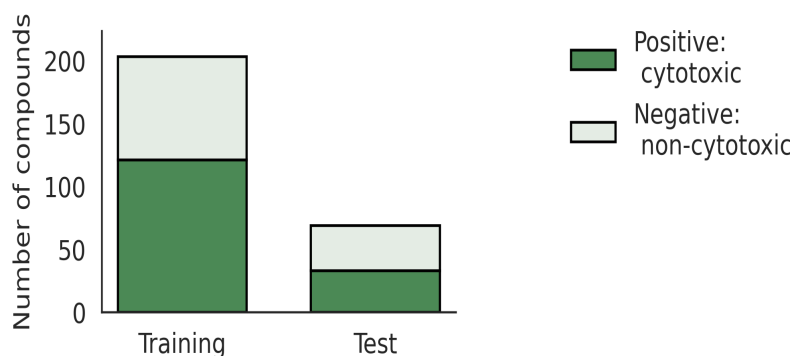
The descriptor values for training set are not included as supporting information.

6.4. Data for the dependent variable (response) for the training set:

The dependent variable for training set is not included as supporting information, but it is available upon request.

6.5. Other information about the training set:

The training set is comprised of 203 (75%) compounds from a curated dataset of 272 compounds. Compounds were selected for the training set using the Kmeans algorithm and a random split of resulting clusters, ensuring a balanced distribution on positive and negative values.



6.6. Pre-processing of data before modelling:

The experimental data of this dataset was curated following a standard procedure in order to guarantee its quality. Compounds with unclearly defined chemical structures were deleted, as well as inorganics compounds, metal complexes, salts containing organic polyatomic counterions, mixtures and substances of unknown or variable composition (UVCB). Also, duplicates and tautomers were checked.

6.7. Statistics for goodness-of-fit:

Experimental values	QSAR predictions		
	non-cytotoxic	cytotoxic	
non-cytotoxic	60	22	73.2% (TNR)
cytotoxic	29	92	76.0% (TPR)
	67.4 % (NPV)	80.7% (PPV)	74.9% (ACC)

Parameters	Training
Accuracy (ACC)	0.749
Sensitivity, recall or true positive rate (TPR)	0.760
Specificity or true negative rate (TNR)	0.732
Precision or positive predictive value (PPV)	0.807
Area under the ROC (AUC)	0.746
Negative predictive value (NPV)	0.674
F-score	0.783
Critical Success Index (CSI)	0.643
Matthews Correlation Coefficient (MCC)	0.487

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

Not available.

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

The dataset used to train the model was divided with a (stratified) K-fold algorithm in five parts to check the robustness of the model. The average metrics of the 5 folds and their standard deviation (80% train - 20% validation) are presented here.

Parameters	Training (CV)	Validation (CV)
Accuracy (ACC)	0.77 ± 0.02	0.68 ± 0.03
Sensitivity, recall or true positive rate (TPR)	0.80 ± 0.02	0.74 ± 0.04
Specificity or true negative rate (TNR)	0.74 ± 0.05	0.58 ± 0.07
Precision or positive predictive value (PPV)	0.82 ± 0.02	0.73 ± 0.03
Area under the ROC (AUC)	0.77 ± 0.02	0.66 ± 0.03
Negative predictive value (NPV)	0.71 ± 0.02	0.61 ± 0.05
F-score	0.81 ± 0.01	0.73 ± 0.03
Critical Success Index (CSI)	0.68 ± 0.02	0.58 ± 0.03
Matthews Correlation Coefficient (MCC)	0.54 ± 0.04	0.33 ± 0.07

6.10. Robustness – Statistics obtained by Y-scrambling:

Not reported.

6.11. Robustness – Statistics obtained by bootstrap:

Not available.

6.12. Robustness – Statistics obtained by other methods:

Not available.

7. External validation - OECD Principle 4

7.1. Availability of the external validation set:

The curated training and validation sets are not included as supporting information, but they can be provided upon request for regulatory assessment.

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:

The descriptor values for validation set are not included as supporting information.

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

The dependent variable for validation set is not included as supporting information, but it is available upon request.

7.5. Other information about the external validation set:

The external validation set is comprised of 69 compounds extracted from the dataset compiled to develop the model.

7.6. Experimental design of test set:

The external validation set is comprised of 69 (25%) compounds extracted from the curated dataset of 272 compounds used to train the model. Compounds were selected for the external validation set using the Kmeans algorithm and a random split of resulting clusters, ensuring a balanced distribution on positive and negative values.

7.7. Predictivity - Statistics obtained by external validation:

Experimental values	QSAR predictions		
	non-cytotoxic	cytotoxic	
non-cytotoxic	23	13	63.9% (TNR)
cytotoxic	10	23	69.7% (TPR)
	69.7 % (NPV)	63.9% (PPV)	66.7% (ACC)

Parameters	Validation
Accuracy (ACC)	0.667
Sensitivity, recall or true positive rate (TPR)	0.697
Specificity or true negative rate (TNR)	0.639
Precision or positive predictive value (PPV)	0.639
Area under the ROC (AUC)	0.668
Negative predictive value (NPV)	0.697
F-score	0.667
Critical Success Index (CSI)	0.500
Matthews Correlation Coefficient (MCC)	0.336

The full dataset of the model (including the external validation set) has been divided with a (Stratified) K-fold algorithm in five parts to check the robustness of the model. The average metrics of the 5 folds and their standard deviation (80% train - 20% validation) are presented here.

Parameters	Training (CV)	Validation (CV)
Accuracy (ACC)	0.73 ± 0.01	0.70 ± 0.04
Sensitivity, recall or true positive rate (TPR)	0.73 ± 0.03	0.71 ± 0.08
Specificity or true negative rate (TNR)	0.73 ± 0.01	0.69 ± 0.04
Precision or positive predictive value (PPV)	0.78 ± 0.01	0.75 ± 0.03
Area under the ROC (AUC)	0.73 ± 0.01	0.70 ± 0.04
Negative predictive value (NPV)	0.67 ± 0.02	0.65 ± 0.06
F-score	0.75 ± 0.02	0.73 ± 0.05
Critical Success Index (CSI)	0.60 ± 0.02	0.57 ± 0.06
Matthews Correlation Coefficient (MCC)	0.45 ± 0.03	0.40 ± 0.08

7.8. Predictivity – Assessment of the external validation set:

Among the validation set, 100.0% of the molecules are inside the applicability domain by at least one method. Particularly, 89.9% by the Tanimoto-Jaccard method, 92.8% by the leverage method and 98.6% by the Euclidean distance method.

7.9. Comments on the external validation of the model:

None

8. Providing a mechanistic interpretation - OECD Principle 5

8.1. Mechanistic basis of the model:

The presented model identifies chemical structural features and physicochemical properties, which during the construction of the model were found to be of relevance to *in vivo* mammalian somatic cell genotoxicity study (micronucleus assay).

8.2. A priori or a posteriori mechanistic interpretation:

A posteriori mechanistic interpretation: The identified chemical structural features and physicochemical properties may serve as starting point for a posteriori mechanistic interpretation.

8.3. Other information about the mechanistic interpretation:

None

9. Miscellaneous information

9.1. Comments:

All ProtoPRED models are developed in accordance with the OECD (2007) Guidance Document on the Validation of (Q)SAR Models and are assessed following the (Q)SAR Assessment Framework: Guidance for the Regulatory Assessment of (Q)SAR Models and Predictions, (2024), making them suitable for regulatory purposes, (full references in section 9.2).

A substance negative in this endpoint can be considered as non-genotoxic (*in vivo* chromosomal mutagenicity). There may however be specific cases where the *in vitro* mammalian cell test can still be justified even though *in vivo* cytogeneticity data exist. For example, in the *in vivo* micronucleus test, certain substances may not reach the bone marrow due to low bioavailability or specific tissue/organ distribution and would result negative (see reference for "Guidance on information requirements and chemical safety assessment, Chapter R.7a" in section 9.2).

9.2. Bibliography:

- [1] Guidance on information requirements and chemical safety assessment, Chapter R.7a: https://echa.europa.eu/documents/10162/17224/information_requirements_r7a_en.pdf
- [2] OECD guideline: OECD: Test No. 474: Mammalian erythrocyte micronucleus test. https://www.oecd-ilibrary.org/environment/test-no-474-mammalian-erythrocyte-micronucleus-test_9789264264762-en
- [3] Benigni R, Bossa C, Tcheremenskaia O, Battistelli CL, Crettaz P. The new ISSMIC database on *in vivo* micronucleus and its role in assessing genotoxicity testing strategies. *Mutagenesis*. 2012 Jan;27(1):87-92.
- [4] QSAR Toolbox: <https://qsartoolbox.org/>
- [5] Todeschini, R. & Consonni, V. (2009). *Molecular Descriptors for Chemoinformatics*, Wiley-VCH
- [6] Consonni, V., & Todeschini, R. (2010). Molecular descriptors. In Puzyn, T., Leszczynski, J. & Cronin, M. T. (Eds.) *Recent advances in QSAR studies* (pp. 29-102). Springer

9.3. Supporting information:

Files with the training and validation datasets and other additional data can be provided upon request to regulatory agencies and institutions for assessment (a non-disclosure agreement might be needed).