ProtoICH

QPRF DOSSIER

ProtoQSAR model for *in vitro* gene mutation study in bacteria (Ames test)

Prediction for CCCCC

www.protoqsar.com

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

protopred@protoqsar.com

+34 962 021 811

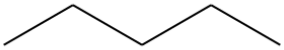
ProtoQSAR
Computational toxicology:
fast, economical and ethical

Summary

Model: ProtoQSAR model for *in vitro* gene mutation study in bacteria (Ames test) (v1.0)

Human health effects: Mutagenicity. OECD 471: Bacterial reverse mutation test. Mutagenicity refers to the induction of permanent transmissible changes in the amount or structure of the genetic material of cells or organisms. The Bacterial reverse mutation test evaluates gene mutations. The test uses amino-acid requiring strains of bacteria to detect (reverse) gene mutations (point mutations and frameshifts).

Molecule: CCCCC



Prediction: Non-mutagenic

A compound is classified Ames positive if it significantly induces revertant colony growth in at least one of out of five strains, following OECD Test No. 471 guidance.

Applicability domain: Inside

The compound falls inside the applicability domain by 4 different methods: Tanimoto-Jaccard similarity, leverage, Euclidean distance and descriptors range. This has been automatically assessed by ProtoPRED® using the criteria detailed in 6.1.b.

Reliability: Highly reliable

Highly reliable. The prediction has been considered highly reliable, with a score of 94% based on the reliability criteria described in section 7.6

QPRF: ProtoQSAR model for *in vitro* gene mutation study in bacteria (Ames test)

1. General information

1.1. Date of QPRF:

23-Jan-2026

1.2. QPRF author and contact details:

a. Authorship: ProtoQSAR

b. Address: Carrer de Nicolau Copèrnic 6

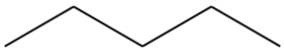
c. Phone:

d. Email: protopred@protoqsar.com

2. Substance

2.1. CAS number:

-



2.2. EC number:

-

2.3. Other regulatory numerical identifiers:

-

2.4. Chemical name:

-

2.5. Structural formula:

-

2.6. Structural and composition information:

a. SMILES: CCCCC

b. InChI: InChI=1S/C5H12/c1-3-5-4-2/h3-5H2,1-2H3

c. Other structural representation: A graphical representation above (not used in the prediction).

d. Stereochemical features: There is no stereochemical information codified in the SMILES because the substance has no identified stereochemical centres.

e. Composition information: The prediction corresponds to a single molecule, fully defined by its SMILES. It is unknown if this is the only component of the substance.

3. Model and software

3.1. Model:

a. Model or submodel name: ProtoQSAR model for *in vitro* gene mutation study in bacteria (Ames test)

b. Model version: Version v1.0, calculated with ProtoPRED® (ProtoQSAR proprietary software) (v1.0)

c. Reference to QMRF: The corresponding QMRF, named *ProtoQSAR model for in vitro gene mutation study in bacteria (Ames test)* can be downloaded from https://protopred.protoqsar.com, more details can be requested to ProtoQSAR S.L., the owner of the model, by email to protopred@protoqsar.com.

3.2. Software:

a. Software name: ProtoPRED® (ProtoQSAR proprietary software)

b. Software version: v1.0

c. Software reference: The software is proprietary and there is no reference to a publication. Software can be referenced as: ProtoQSAR S.L. (2021). *ProtoPRED® v1.0* [Web service] https://protopred.protoqsar.com.

d. Software availability: The software to develop the models is proprietary and not publicly available. However, evaluators can contact protopred@protoqsar.com to get further details or to try the code.

4. Prediction

4.1. Property:

a. Predicted property: *In vitro* gene mutation study in bacteria (Ames test)

b. Test guideline covered: Endpoint following the OECD: Test No. 471: Bacterial reverse mutation test.

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

4.2. Value:

a. Predicted value: Non-mutagenic

b. Predicted value (comments): A compound is classified Ames positive if it significantly induces revertant colony growth in at least one of out of five strains, following OECD Test No. 471 guidance.

c. Unit: N/A

5. Input

5.1. Structure:

a. Input structure: The prediction uses as input the SMILES of the molecule as shown in point 2.6a

b. Stereochemical features: There is no stereochemical information codified in the SMILES, so the substance is a non-stereochemical molecule or a racemic mixture.

c. Tautomerism: There is not automatic treatment of tautomers. The exact structure depicted in the SMILES is used.

5.2. Descriptors:

These are the calculated values of the descriptors of the model for the predicted molecule. In case that the descriptor cannot be calculated, the imputed and scaled value is indicated in red italics.

Descriptor	Value
GATS5c	0.83
GATS1d	1.62
BELd-1	0.99
D/Dr3	0.00
SIC1	0.28
N-071	0.00
MAXDN	0.16
MATS1s	-0.37
BELc-1	-0.18
BELi0	11.44
MIC1	10.02
AATS1i	146.54
AATSC1i	-0.17

a. Units: ProtoPRED® uses standardised values (unitless) for the molecular descriptors, both in the prediction and model development. Original values of the descriptors can have different units, but in most cases are also unitless (such as ratios and counters).

b. Experimental/calculated: All molecular descriptors in this model are fully calculated from the SMILES. Any experimental parameter included in the calculation of a descriptor is read from the database implemented in the descriptor.

c. Source: The molecular descriptors in this model were selected statistically from a panel of more than 4000 descriptors calculated by ProtoQSAR S.L. proprietary software. This software uses different packages such as RDKit and Mordred and also has its own implementation of descriptors based on literature.

[1] Todeschini, R. & Consonni, V. (2009). *Molecular Descriptors for Chemoinformatics*, Wiley-VCH

[2] Ghose, A.K. Viswanadhan, V.N. & Wendoloski, J. (1998). *J. Phys. Chem. A.*, 102:3762-3772

d. Differences with model development and validation: All molecular descriptors are calculated in the prediction using the same methodology and software than in the model development. Nevertheless, to avoid any unexpected deviation, all the descriptors have been recalculated in the same server to produce the final model and obtain the performance indicators provided here.

5.3. Model and/or software settings:

The model has been applied as described here and in the QMRF. ProtoPRED® does not have any customizable setting.

6. Applicability domain (AD) and limitations

6.1. Applicability domain (AD) and limitations:

a. AD assessment: The compound falls inside the applicability domain by 4 different methods: Tanimoto-Jaccard similarity, leverage, Euclidean distance and descriptors range. This has been automatically assessed by ProtoPRED® using the criteria detailed in 6.1.b.

b. AD assessment justification: The applicability domain (AD) of the prediction was calculated by using the Tanimoto-Jaccard coefficient assessment, the Euclidean distance calculation, the leverage analysis and the descriptors range analysis (more details in the QMRF report). A prediction is considered to fall outside the AD if it does not match any of these criteria.

- ***Tanimoto:*** The Tanimoto-Jaccard coefficient for the structurally more similar compound in the training set is 1.000. It is considered inside the AD if the value is higher than 0.528 (corresponding to a 10% of compounds in an external set).
- ***Euclidean distance:*** The Euclidean distance of this compound with the furthest one in the training set is 18.09 and with the nearest one is 0.00. It is considered inside the AD if it is lower than the largest compound-compound distance in the training set (22.01). As a reference, the average pairwise distance between in the training set is 4.74.
- ***Leverage:*** The leverage of the compound is 0.00. It is considered inside the AD if it is lower than 0.01 (the threshold value is calculated as $3p/n$ considering the number of descriptors, p, and training data, n, following the standard criterium).
- ***Descriptors range:*** There are 0 descriptors with a standardised value outside the range in the training set. It is considered inside the AD if none of the descriptors is outside the range.

c. Any other limitation: 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.

7. Reliability assessment

7.1. Reproducibility:

The algorithm is unambiguous and perfectly defined by the equations encoded in the model, as well as the calculation of the descriptors (including adequate and consistent seeds if there is any randomness). Hence, the prediction can be reproduced using https://protopred.protoqsar.com (please, contact protopred@protoqsar.com if you need access to assess a prediction).

7.2. Overall performance of the model:

The model included in this study has been validated (both internal and external validation) as described in the QMRF. A summary of the main performance metrics is shown below. Note that this data corresponds to the internal model units as discussed in the section 4.2.b and in the QMRF, which can differ from the reported numerical prediction.

Training set confusion matrix:

Experimental values	QSAR predictions		
	non-mutagenic	**mutagenic**	
non-mutagenic	2048	203	91.0% (TNR)
mutagenic	154	2463	94.1% (TPR)
	93.0 % (NPV)	92.4% (PPV)	**92.7% (ACC)**

Validation set confusion matrix:

Experimental values	QSAR predictions		
	non-mutagenic	**mutagenic**	
non-mutagenic	553	196	73.8% (TNR)
mutagenic	190	685	78.3% (TPR)
	74.4 % (NPV)	77.8% (PPV)	**76.2% (ACC)**

Parameters	Training	Validation
Accuracy (ACC)	0.93	0.76
Sensitivity, recall or true positive rate (TPR)	0.94	0.78
Specificity or true negative rate (TNR)	0.91	0.74
Precision or positive predictive value (PPV)	0.92	0.78
Negative predictive value (NPV)	0.93	0.74
Miss rate or false negative rate (FNR)	0.06	0.22
Fall-out or false positive rate (FPR)	0.09	0.26
False discovery rate (FDR)	0.08	0.22
False omission rate (FOR)	0.07	0.26
F-score	0.93	0.78
Matthews Correlation Coefficient (MCC)	0.85	0.52

Critical Success Index (CSI)	0.87	0.64
Area under the ROC (AUC)	0.93	0.76

The machine learning algorithm used for the prediction provides, in addition of a binary response, a numerical score that can be related to the internal probability of being Non-mutagenic. In this case the value is 94.56 %.

7.3. Additional reliability aspects based on the training set:

a. Descriptor space: The value of all the descriptors of the target substance is inside the range of values for that descriptor on the training molecules.

Further details on the relationship of the target molecule with the training set in the descriptors space can be found in the section 6.1.b relative to the applicability domain.

b. Structural fragment space: There are no structural fragments among those considered in this analysis in the target substance. Hence the comparison with the training set has not been done.

c. Response space: The categorical model used for the prediction classifies the target substance as one of the available categories.

d. Mechanism considerations: Mechanistic considerations have not been automatically included in the model. The model is expected to be general for the predicted property as described above and in the QMRF. The user should consider if the target substance expected mechanism requires further investigation on the modelling database.

e. Metabolic considerations: Metabolic considerations have not been automatically included in the model. The model predicts the properties of the target substance exactly as reported above. Details on the metabolic stability of the substance and the effects of the potential metabolites should be considered separately.

7.4. Analogues:

Similar structural analogues to the target compound have been found inside the training and test sets. A similarity threshold of 0.575 is used to identify the more similar substances, which statistically correspond to the 5%. The 10 most similar substances are used as analogues and presented in the table of the Annex 1. However, there are 16 substances that reach the similarity threshold. The information reported includes:

a. Identifiers: Each analogue is identified by its SMILES and, if available, by its CAS number and chemical name.

b. Source of the analogue: All analogues are collected from the database used to develop the model (both training and test sets).

c. Experimental value for the property of interest: See individual values in Annex 1. 90.0% have the same observed value than the predicted molecule.

d. Reference for experimental value: The source of the training and test datasets is referenced in the QMRF of the model (section 9.2).

e. Predicted value of the property of interest: See individual values in Annex 1.

f. Accuracy of the prediction: See individual values in Annex 1.

g. Comments on the similarity: Analogues are calculated using the Tanimoto-Jaccard similarity index with MACCS fingerprints. The use of fingerprints as structural description ensures that the

similarity is not biased by the descriptors used in the model.

- Considerations on analogues: Regarding the target property, 90.0% of the presented analogues have the same observed value than the predicted molecule. The local performance of the model (calculated in the selected analogues) is described by an accuracy of 1.0.

7.5. Other reliable information on the property:

An experimental value has been found in the database used to develop the model: Non-mutagenic. Details on the source of the data can be found in the QMRF.

7.6. Conclusion on reliability:

Highly reliable. The prediction has been considered highly reliable, with a score of 94% based on the following reliability criteria.

A prediction is considered highly reliable if fulfils a 75% of the criteria, reliable if fulfils a 50% and lowly reliable if fulfils a 25%.

Criteria used for the reliability assessment

- Reproducibility: YES (1); the prediction is reproducible.

- Overall training performance: YES (0.93); the accuracy/r-square for the model on the training set is higher than 0.80 (see Annex 1 and section 7.4).

- Validation performance: YES (0.76); the accuracy/r-square for the model on the validation set is higher than 0.60 (see Annex 1 and section 7.4).

- AD: YES (1); the molecule is inside the applicability domain of the model by all methods considered (see details at section 6).

- Structural space: There are no structural fragments among those considered in this analysis in the target substance (see section 7.3b).

- Property range: The predicted response value cannot be considered as a reliability indicator for classification models (see section 7.3c).

- Descriptors space: YES (1); all the descriptors of the target have a value included in the range of values of the training molecules (see details at section 7.3a).

- Analogues: YES (0.9); more than a 66% of the structural analogues have the same experimental value than the prediction for the target (see Annex 1 and section 7.4).

- Local performance: YES (1.0); the accuracy/r-square for the model on the local space of the structural analogues is higher than 0.75 (see Annex 1 and section 7.4).

8. Purpose of use (for regulatory applications)

8.1. Regulatory purpose:

This prediction has been performed with the purpose of following the guidelines for ICH M7 regulation.

8.2. Approach for regulatory interpretation of the prediction:

According to the ICH M7 regulation, if not experimental data on mutagenicity is available, an *in silico* approach combining two types of models: SAR models (based on expert rules) and QSAR models (statistical) can be used.

8.3. Regulatory interpretation of the result:

For ICH-M7 regulatory purposes, the negative prediction (non-mutagenic) reported above should be confirmed by a matching result from the expert rules based model. In case of contradiction, the result is considered uncertain.

8.4. Uncertainty:

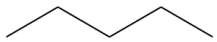
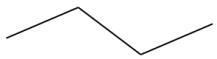
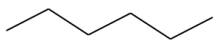
The uncertainty of a QSAR prediction could be estimated by reviewing the assessment elements outlined by the OECD in the prediction checklist. As a guide, find in Annex 2 below our proposed assessment combining information gathered in this QPRF and the related QMRF (This preliminary assessment is based on the information included in the platform only, further considerations by the user should affect the final assessment).

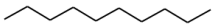
8.5. Conclusion:

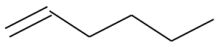
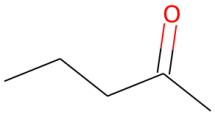
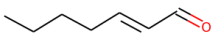
The ProtoQSAR model has been used for the prediction of *in vitro* gene mutation study in bacteria (Ames test) of the chemical substance with SMILES CCCCC. The compound has been predicted as *non-mutagenic*, and the prediction falls inside the applicability domain. An automatic, preliminary reliability score of 94% was obtained in basis to the criteria described in section 7.6, corresponding to a highly reliable prediction.

Annex 1: Structural analogs

These are the most similar substances to the target found in the model database, according to the Tanimoto-Jaccard similarity coefficient (Tc) based on the bit-based version of MACCS fingerprints. According to a benchmark on 25000 molecules, a 95% of molecules have a similarity below 0.575 (used as a threshold, more details in section 7.4). Otherwise, a threshold of 0.431 approximately includes a 70% and 0.655 a 99%.

analog 1 **b.** Train set	**a.** SMILES:	CCCCC	
			
	c. Observed value:	Negative	
	e. QSAR prediction:	Negative	**f.** Correct
	g. Similarity (Tanimoto):	1.000	
analog 2 **b.** Train set	**a.** SMILES:	CCCC	
			
	c. Observed value:	Negative	
	e. QSAR prediction:	Negative	**f.** Correct
	g. Similarity (Tanimoto):	0.880	
analog 3 **b.** Train set	**a.** SMILES:	CCCCCC	
			
	c. Observed value:	Negative	
	e. QSAR prediction:	Negative	**f.** Correct
	g. Similarity (Tanimoto):	0.800	

analog 4 **b.** Train set	**a.** SMILES:	CCCCCCCCCCC	
	c. Observed value:	Negative	
	e. QSAR prediction:	Negative	**f.** Correct
	g. Similarity (Tanimoto):	0.730	
analog 5 **b.** Train set	**a.** SMILES:	CCCCCCCCCCCCCCCCCCCC	
	c. Observed value:	Negative	
	e. QSAR prediction:	Negative	**f.** Correct
	g. Similarity (Tanimoto):	0.730	
analog 6 **b.** Train set	**a.** SMILES:	CCCCCCCCCC	
			
	c. Observed value:	Negative	
	e. QSAR prediction:	Negative	**f.** Correct
	g. Similarity (Tanimoto):	0.730	
analog 7 **b.** Train set	**a.** SMILES:	CCCCCCCCC	
			
	c. Observed value:	Negative	
	e. QSAR prediction:	Negative	**f.** Correct
	g. Similarity (Tanimoto):	0.730	

analog 8 **b.** Test set	**a.** SMILES:	C=CCCCC	
			
	c. Observed value:	Negative	
	e. QSAR prediction:	Negative	**f.** Correct
	g. Similarity (Tanimoto):	0.700	
analog 9 **b.** Test set	**a.** SMILES:	CCCC(C)=O	
			
	c. Observed value:	Negative	
	e. QSAR prediction:	Negative	**f.** Correct
	g. Similarity (Tanimoto):	0.667	
analog 10 **b.** Train set	**a.** SMILES:	CCCC/C=C/C=O	
			
	c. Observed value:	Positive	
	e. QSAR prediction:	Positive	**f.** Correct
	g. Similarity (Tanimoto):	0.640	

Annex 2: Uncertainty assessment using the QSAR prediction checklist

The uncertainty of a QSAR prediction should be estimated by reviewing the assessment elements outlined by the OECD's prediction checklist.

A thorough review requires considering not only the internal data of the prediction, but the regulatory purpose, the characterization of the substance of interest and additional information.

However, as a guide, we propose an assessment below about the fulfilment of the assessment criteria on basis to the information gathered in this QPRF and the related QMRF.

Assessment Element	*Weight*	*Outcome*	*Uncertainty*	*Details*	*Comment*
AE.1. Correct input(s) to the model					
AE.1.1. Clear and complete description of the input and model setting	High	Fulfilled	Low	QPRF 5.	A complete description of the input is presented in the QPRF.
AE.1.2. Input representative of the substance under analysis	High	Fulfilled	Medium	QPRF 2. & 5.	The level of definition of the input data is properly described, but its adequacy to the substance of interest depends on the user.
AE.1.3. Reliable input (parameters)	Medium	Fulfilled	Low	QPRF 5.2.	Molecular structure is properly defined and all descriptors are fully calculated from it with reproducible and validated scripts.
AE.2. Substance within the applicability domain of a valid model					
AE.2.1. Substance within the applicability domain	High	Fulfilled	Low	QPRF 6.1.a-b	The substance is inside the applicability domain of the model by all methods considered.
AE.2.2. Any other limitation of the model is considered	High	Fulfilled	Low	QPRF 6.1.c	ProtoPRED® filters input that do not comply with the limitations of the model.
AE.3. Reliable prediction					
AE.3.1. Reproducibility	High	Fulfilled	Low	QPRF 7.1.	The prediction can be reproduced at ProtoPRED® (evaluators can contact protopred@protoqsar.com to obtain access)
AE.3.2. Overall performance of the model	High	Fulfilled	Low	QPRF 7.2.	The performance is considered adequate both in the validation and training set: the accuracy/r-square is 0.93 (training) and 0.76 (validation).

Assessment Element	*Weight*	*Outcome*	*Uncertainty*	*Details*	*Comment*
AE.3.3. Fit within the physicochemical, structural and response spaces of the training set of the model	Medium	Fulfilled	Medium	QPRF 7.3.a-c	The prediction fits within the training set by 2 of those 3 different spaces (physicochemical, structural and response).
AE.3.4. Performance of the model for similar substances	High	Fulfilled	Low	QPRF 7.4.	The accuracy/r-square for the model on the local space of the structural analogues (1.0) is higher than 0.75. All structural analogues are above the 95% similarity threshold.
AE.3.5. Mechanistic and/or metabolic considerations	High	Not documented	Medium	QPRF 7.3.d-e	The mechanistic and metabolic considerations are not explicitly included in the QPRF.
AE.3.6. Consistency of information	High	Not documented	Medium	QPRF 7.5.	ProtoPRED® does not include such analysis.
AE.4. Outcome is fit for the regulatory purpose					
AE.4.1. Compliance with additional requirements	High	Fulfilled	Low	QPRF 8.1.	The model has been developed to be appliable to the REACH regulatory framework.
AE.4.2. Correspondence between predicted property and property required by the regulation	High	Fulfilled	Low	QMRF 3.	The modelled property is described with sufficient level of detail.
AE.4.3. Decidability within the specific framework	High	Fulfilled	Low	QPRF 8.	The prediction output is described and related with the thresholds and conditions stablished in the regulation.