(Q)SAR Model Reporting Format (QMRF)

(The present QMRF v.2.1 is prepared in accordance with (Q)SAR Assessment Framework (QAF) document developed by OECD)
(https://one.oecd.org/document/ENV/CBC/MONO(2023)32/ANN1/en/pdf**)**

Welcome

Model version: *In vitro* Mouse Lymphoma v.07.07
Platform version: OASIS TIMES 2.35.1
Name: *In vitro* Mouse Lymphoma with S9 metabolic activation
Author: LMC, University "Prof. As. Zlatarov", Bourgas, Bulgaria
Date: 22 April 2026
E-mail: ovanes.mekenyan@oasis-lmc.org
www: http://www.oasis-lmc.org/

Section 1. QSAR identifier

1.1. QSAR identifier (title)

In vitro Mouse Lymphoma with S9 metabolic activation

1.2. Other related models

In vitro Ames mutagenicity with S9 metabolic activation; *In vitro* Chromosomal Aberrations S9 activated

1.3. Software coding of the model

Model version: *In vitro* Mouse Lymphoma v.07.07
Platform version: OASIS TIMES 2.35.1
Name: *In vitro* Mouse Lymphoma with S9 metabolic activation
Developer: LMC, University "Prof. As. Zlatarov", Bourgas, Bulgaria
Coding language: Delphi 10.2

Section 2. General information

2.0. Abstract

In vitro mouse lymphoma thymidine kinase assay (MLA) identifies agents that cause point mutations and chromosomal rearrangement [1, 2]. The *Tk* locus in mouse lymphoma

cells is the preferred target for mammalian cell genotoxicity assays because of the wealth of data which exists to support this locus and assay and because of the broad spectrum of damage detected at this locus.

2.1.Date of QMRF

22 April 2026

2.2. QMRF author(s) and contact details

Name: Laboratory of Mathematical Chemistry

Affiliation: Laboratory of Mathematical Chemistry, University "Prof. As. Zlatarov", "Yakimov" St. #1, 8010 Bourgas, BULGARIA

URL: http://www.oasis-lmc.org

E-mail: ovanes.mekenyan@oasis-lmc.org

2.3. Date of QMRF update(s)

17 Oct 2019; 19 Nov 2021; 1 March 2023; 31 March, 2024; 30 May 2025; 22 April 2026.

2.4. QMRF update(s)

Information which has been modified:

Sections 1.1 QSAR identifier (title)**; Sections 1.3** Software coding the model; **Section 2.** General information; **Sections 2.0** Abstract; **Sections 2.1** Date of QMRF; **Sections 2.3** Date of QMRF update(s); **Sections Section 4.6.** Software name and version for descriptor generation; **Section 5.4.** Limits of applicability;

2.5.Model developer(s) and contact details

Name: Ovanes Mekenyan, Petko Petkov, Stefan Kotov, Kalin Kirilov

Affiliation: Laboratory of Mathematical Chemistry, University "Prof. As. Zlatarov", "Yakimov" St. #1, 8010 Bourgas, BULGARIA

URL: http://www.oasis-lmc.org

E-mail: ovanes.mekenyan@oasis-lmc.org

2.6. Date of model development and/or publication

2018 May

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

1. Applegate, M. L., Moore, M. M., Broder, C. B., Burrell, A., Juhn, G., Kasweck, K. L., Lin, P. F., Wadhams, A. and Hozier, J. C. (1990) Molecular dissection of mutations at the heterozygous thymidine kinase locus in mouse lymphoma cells. *Proc. Natl Acad. Sci. USA*, 87, pp. 51–55.
2. Moore, M. M., Harrington-Brock, K., Doerr, C.L., Dearfield, K.L. (1989) Differential mutant quantitation at the mouse lymphoma tk and CHO hgprt loci. *Mutagenesis*, 4(5), pp. 394-403.
3. P.I. Petkov, T.W. Schultz, M. Honma, K. Kirilov, S. Kotov, O.G. Mekenyan. 2017. Predicting *in vitro* genotoxicity by mouse lymphoma L5178Y thymidine kinase mutation assay (MLA): Accounting for simulated metabolic activation of chemicals, *Computational Toxicology* 4: pp. 45–53.

2.8. Availability of information about the model

In vitro CA with S9 metabolic activation model is proprietary and its use is subject of licence agreement.

Information that cannot be disclosed:

- External validation sets,
- Proprietary chemicals,
- Source code.

For more details, please contact Professor Ovanes Mekenyan: ovanes.mekenyan@oasis-lmc.org.

Details of the model is provided in the sections bellow as well as in the following link: http://oasis-lmc.org/products/models/human-health-endpoints.aspx

2.9. Availability of another QMRF for exactly the same model

Not available.

Section 3. Defining the endpoint – OECD Principle 1

3.1. Species

Chemicals included in the training set of the TIMES MLA model are collected according to the recommendation in the OECD technical guideline 490 addressing L5178Y mouse lymphoma cells.

For more details, see: OECD GUIDELINE FOR TESTING OF CHEMICALS

http://www.oecd-ilibrary.org/environment/test-no-490-in-vitro-mammalian-cell-gene-mutation-tests-using-the-thymidine-kinase-gene_9789264242241-en

3.2. Endpoint

In vitro Mammalian Cell Gene Mutation

According to JRC pre-classification list of endpoints:
No. 207 QMRF Human Health Effects, QMRF 4.10 Mutagenicity.

3.3. Comment on endpoint

Direct binding of chemicals to DNA is one of the underlying mechanisms that is responsible for MLA mutagenicity. Disturbance of protein synthesis due to inhibition of topoisomerases and interaction of chemicals with nuclear proteins associated with DNA (e.g., histone proteins) are identified as additional mechanisms also leading to positive MLA effect in the *Tk* locus. Based on that, the toxicodynamic component of the MLA models, i.e. alerts (functionalities responsible for eliciting positive MLA effect) are derived to account for interactions with DNA and/or proteins.

3.4. Endpoint units

Qualitative – positive/ negative

3.5. Dependent variable

Observed MLA with S9

3.6. Experimental protocol

Standard test for *In vitro* Mammalian Cell Gene Mutation

3.7. Endpoint data quality and variability

National Toxicology Program, Toxicology and Carcinogenesis Studies; different scientific papers. CCRIS Toxnet database; ECHA CHEM database.

References associated with each documented mutagenicity data (except for proprietary data) included in the training set of the model are provided in ***Appendix 1***.

Section 4. Defining the algorithm – OECD Principle 2

4.1. Type of model

(Q)SAR

4.2. Explicit algorithm

Prediction of *in vitro* Mammalian Cell Gene Mutations.

Reactivity component of the MLA model describing interactions of chemicals with DNA and/or proteins is based on an alerting group approach. Only those toxicophores having clear interpretation for the molecular mechanism causing the ultimate effect were included in the model.

In the MLA model (+S9), the reactivity component is combined with a metabolic simulator, which is trained to reproduce documented maps for mammalian rat liver metabolism for 441 chemicals. Parent chemicals and each of the generated metabolites are submitted to a battery of models to screen for a general effect and mutagenicity mechanisms. Thus, chemicals are predicted to be mutagenic as parents only, parents and metabolites, and metabolites only.

4.3. Descriptors in the model

Descriptors in the model are structural boundaries associated with alerting groups related to interactions with DNA/proteins. Alerts in the TIMES MLA (+S9) constitute expertly-derived sets of structural fragments incorporating knowledge for the interactions of chemicals (parents and metabolites) with DNA/proteins. Application of the alerts on the training set of the model forms fractions of representative chemicals for the alerts, i.e. so-called 'local' training sets. All chemicals captured by the alerts are considered as validation sets of the introduced expert knowledge addressing reactivity of chemicals with DNA/proteins. The procedure for obtaining local training sets includes applying the structural boundaries of the alert searching among all chemicals from the training set of the model after application of S9 metabolic simulator. According to this, local training sets contain parent chemicals in which general fragments are:

- found in their structures;
- not found in the parent structures but found in their metabolite(s).

Description of these alerts is provided in the next sections.

4.4. Descriptor section

The main characteristics of each DNA and protein alerts in TIMES MLA (+S9) are:

- Alert name (corresponding to the name of the chemical class which is addressed);

- Performance of alert (correct/incorrect predictions) which is estimated based on proportion of observed positive chemicals from all chemicals captured by the alert. Performance of each alert is provided with its confidence range. As smaller is the size of local training sets as wider are the confidence ranges and vice versa.
- P-values addressing the reliability of alert performance estimation and taking into account possible bias of positive/negative chemicals in the training set of the model. Low p-values could be obtained only if both are satisfied:
    - The number of chemicals in local training set is high enough;
    - The alert performance is significantly higher than the proportion of positive/negative chemicals in the model training set, i.e., so-called naïve alert.

Analogically, high p-values could be obtained in case of:

- Small number of local training set chemicals (1-2 chemicals); or
- Performance comparable to the performance of the naïve alert.

High performance associated with low *p-values* indicate for High Reliability of alerts.

Table 1. Main characteristics of the DNA and protein alerts in the TIMES_CA model (+S9).

No.	Alert name	Correct	Incorrect	Performance	p-value
1	N-Substituted Aromatic Amines	21	0	0.957 (0.873 ÷ 1.000)	0.0021
2	Haloalkane Derivatives with Labile Halogen	17	0	0.947 (0.847 ÷ 1.000)	0.0066
3	alpha,beta-Unsaturated Carboxylic Acids and Esters	16	0	0.944 (0.838 ÷ 1.000)	0.0088
4	Haloalkanes Containing Heteroatom	14	0	0.938 (0.819 ÷ 1.000)	0.015
5	Fused-Ring Primary Aromatic Amines	13	0	0.933 (0.807 ÷ 1.000)	0.021
6	Heterocyclic Aromatic Amines	11	0	0.923 (0.779 ÷ 1.000)	0.037
7	Polycyclic Aromatic Hydrocarbon and Naphthalenediimide Derivatives	11	0	0.923 (0.779 ÷ 1.000)	0.037
8	Vicinal Dihaloalkanes	11	0	0.923 (0.779 ÷ 1.000)	0.037
9	Single-ring Substituted Primary Aromatic Amines	22	1	0.920 (0.816 ÷ 0.998)	0.010

10	C-Nitroso Compounds	32	2	0.917 (0.827 ÷ 0.990)	0.0038
11	Nitroaniline Derivatives	10	0	0.917 (0.762 ÷ 1.000)	0.050
12	Substituted Anilines	52	4	0.914 (0.841 ÷ 0.977)	0.0006
13	Hydroxylated phenols	51	4	0.912 (0.839 ÷ 0.977)	0.0007
14	Quinones and Trihydroxybenzenes	19	1	0.909 (0.792 ÷ 0.997)	0.021
15	alpha-Activated Haloalkanes	9	0	0.909 (0.741 ÷ 1.000)	0.068
16	N-Nitroso Compounds	9	0	0.909 (0.741 ÷ 1.000)	0.068
17	Piperidindiones	9	0	0.909 (0.741 ÷ 1.000)	0.068
18	Quinoneimines protein binding	9	0	0.909 (0.741 ÷ 1.000)	0.068
19	Epoxides and Aziridines	35	3	0.900 (0.808 ÷ 0.980)	0.0064
20	Aldoxime derivatives	7	0	0.889 (0.688 ÷ 1.000)	0.12
21	Quinoneimines	7	0	0.889 (0.688 ÷ 1.000)	0.12
22	Thioureas and thiocarbamates	7	0	0.889 (0.688 ÷ 1.000)	0.12
23	N-HydroxylAmines	50	6	0.879 (0.795 ÷ 0.956)	0.0062
24	Polarised Haloalkenes Derivatives	6	0	0.875 (0.652 ÷ 1.000)	0.16
25	Dicarbonyl Compounds	19	2	0.870 (0.735 ÷ 0.983)	0.065
26	alpha,beta-Unsaturated Carbonyls and Related Compounds	25	3	0.867 (0.746 ÷ 0.972)	0.048
27	Benzoquinolines and Acridines derivatives	5	0	0.857 (0.607 ÷ 1.000)	0.22
28	Haloepoxides and Halooxetanes	5	0	0.857 (0.607 ÷ 1.000)	0.22
29	Heterocyclic phenylureas	5	0	0.857 (0.607 ÷ 1.000)	0.22
30	Nitrogen Mustards	5	0	0.857 (0.607 ÷ 1.000)	0.22
31	Quinoline Derivatives	5	0	0.857 (0.607 ÷ 1.000)	0.22
32	Substituted Phenols	32	5	0.846 (0.733 ÷ 0.949)	0.061

33	Alfa,Beta-Unsaturated Aldehydes	4	0	0.833 (0.549 ÷ 1.000)	0.30
34	Aminoacridine DNA Intercalators	4	0	0.833 (0.549 ÷ 1.000)	0.30
35	Carbamates	4	0	0.833 (0.549 ÷ 1.000)	0.30
36	Conjugated Nitroalkenes and Five-Membered Aromatic Nitroheterocyclics	4	0	0.833 (0.549 ÷ 1.000)	0.30
37	Nitrophenols, Nitrophenyl Ethers and Nitrobenzoic Acids	8	1	0.818 (0.602 ÷ 0.993)	0.27
38	Acridone, Thioxanthone, Xanthone and Phenazine Derivatives	3	0	0.800 (0.473 ÷ 1.000)	0.40
39	Alkylated nitrosoureas and nitrosoguanidines	3	0	0.800 (0.473 ÷ 1.000)	0.40
40	Arenesulphonamides	3	0	0.800 (0.473 ÷ 1.000)	0.40
41	DNA Intercalators with Carboxamide and Aminoalkylamine Side Chain	3	0	0.800 (0.473 ÷ 1.000)	0.40
42	Isocyanates and Diisocyanates	3	0	0.800 (0.473 ÷ 1.000)	0.40
43	Nitrogen and Sulfur Mustards	3	0	0.800 (0.473 ÷ 1.000)	0.40
44	Pyrimidines and Purines	3	0	0.800 (0.473 ÷ 1.000)	0.40
45	Sulfonates and Sulfates	3	0	0.800 (0.473 ÷ 1.000)	0.40
46	Geminal Polyhaloalkane Derivatives	14	3	0.789 (0.610 ÷ 0.951)	0.32
47	Halogenated Vicinal Hydrocarbons	10	2	0.786 (0.579 ÷ 0.968)	0.36
48	Alkylphosphates, Alkylthiophosphates and Alkylphosphonates	6	1	0.778 (0.524 ÷ 0.989)	0.42
49	Alkylnitrites	5	1	0.750 (0.473 ÷ 0.987)	0.51
50	Alpha-Beta Conjugated Alkene Derivatives with Geminal Electron-Withdrawing Groups	2	0	0.750 (0.368 ÷ 1.000)	0.54
51	Carboxylic Acid Amides	2	0	0.750 (0.368 ÷ 1.000)	0.54
52	Dialkyl Alkylphosphonates	2	0	0.750 (0.368 ÷ 1.000)	0.54
53	Dialkyldithiocarbamate Derivatives	2	0	0.750 (0.368 ÷ 1.000)	0.54

54	Diazenes and Azoxyalkanes	2	0	0.750 (0.368 ÷ 1.000)	0.54
55	Ethenyl Pyridines	2	0	0.750 (0.368 ÷ 1.000)	0.54
56	Haloalkenes with Electron-Withdrawing Groups	2	0	0.750 (0.368 ÷ 1.000)	0.54
57	Nitro Azoarenes and p-Substituted Azobenzenes	2	0	0.750 (0.368 ÷ 1.000)	0.54
58	p-Substituted Mononitrobenzenes	2	0	0.750 (0.368 ÷ 1.000)	0.54
59	Phosphoramides	2	0	0.750 (0.368 ÷ 1.000)	0.54
60	Quinone Methides	2	0	0.750 (0.368 ÷ 1.000)	0.54
61	Specific Acetate Esters	2	0	0.750 (0.368 ÷ 1.000)	0.54
62	Specific Imine and Thione Derivatives	2	0	0.750 (0.368 ÷ 1.000)	0.54
63	Amino Anthraquinones	3	1	0.667 (0.330 ÷ 0.974)	0.72
64	Cyanohydrins	3	1	0.667 (0.330 ÷ 0.974)	0.72
65	Arendiazonium Salts	1	0	0.667 (0.224 ÷ 1.000)	0.73
66	Arenecarboxylic Acid Esters	1	0	0.667 (0.224 ÷ 1.000)	0.73
67	Bipyridilium Herbicides	1	0	0.667 (0.224 ÷ 1.000)	0.73
68	Flavonoids	1	0	0.667 (0.224 ÷ 1.000)	0.73
69	Four-and Five-Membered Lactones	1	0	0.667 (0.224 ÷ 1.000)	0.73
70	Fused-Ring Nitroaromatics	1	0	0.667 (0.224 ÷ 1.000)	0.73
71	Gallic Acid Esters	1	0	0.667 (0.224 ÷ 1.000)	0.73
72	Monohaloalkanes	1	0	0.667 (0.224 ÷ 1.000)	0.73
73	N-Acetoxyamines	1	0	0.667 (0.224 ÷ 1.000)	0.73
74	N-Aryl-N-Acetoxy(Benzoyloxy) Acetamides	1	0	0.667 (0.224 ÷ 1.000)	0.73
75	Nitroalkanes	1	0	0.667 (0.224 ÷ 1.000)	0.73
76	Nitroarenes with Other Active Groups	1	0	0.667 (0.224 ÷ 1.000)	0.73

77	Nitrobiphenyls and Bridged Nitrobiphenyls	1	0	0.667 (0.224 ÷ 1.000)	0.73
78	Non-Cyclic Alkyl Phosphoramides and Thionophosphoramides	1	0	0.667 (0.224 ÷ 1.000)	0.73
79	Polynitroarenes	1	0	0.667 (0.224 ÷ 1.000)	0.73
80	Pyrazolone and Pyrazolidine Derivatives	1	0	0.667 (0.224 ÷ 1.000)	0.73
81	Pyrrolizidine derivatives	1	0	0.667 (0.224 ÷ 1.000)	0.73
82	Thiols	1	0	0.667 (0.224 ÷ 1.000)	0.73
83	p-Aminobiphenyl Analogs	4	2	0.625 (0.315 ÷ 0.919)	0.81
84	Hydrazine Derivatives	2	1	0.600 (0.228 ÷ 0.956)	0.83
85	Hydroxamic acid	1	1	0.500 (0.094 ÷ 0.906)	0.93
86	Acyclic Triazenes	0	0	N/A	N/A
87	alpha,omega-Dihaloalkanes	0	0	N/A	N/A
88	Alpha-Haloethers	0	0	N/A	N/A
89	Amidoxime Esters and Amidoximes	0	0	N/A	N/A
90	Anthrones	0	0	N/A	N/A
91	Bleomycin and Structurally Related Chemicals	0	0	N/A	N/A
92	Carboxylic acid Anhydrides	0	0	N/A	N/A
93	Coumarins	0	0	N/A	N/A
94	Diazoalkanes	0	0	N/A	N/A
95	Haloalkene Cysteine S-Conjugates	0	0	N/A	N/A
96	Halofuranones	0	0	N/A	N/A
97	Haloisothiazolinones	0	0	N/A	N/A
98	Heterocyclic N-Hydroxylamines	0	0	N/A	N/A
99	Heterocyclic nitro compounds	0	0	N/A	N/A
100	Heterocyclic Nitroso compounds	0	0	N/A	N/A
101	Hexahydrotriazine Derivatives	0	0	N/A	N/A
102	Isothiocyanates	0	0	N/A	N/A
103	N-Acyloxy(Alkoxy) Arenamides	0	0	N/A	N/A
104	N-Alkyl-N-nitrosocarbamates	0	0	N/A	N/A
105	N-Hydroxyethyl Lactams	0	0	N/A	N/A
106	N-methylol derivatives	0	0	N/A	N/A
107	N-Nitrosoamine Derivatives	0	0	N/A	N/A
108	N-Trihalomethyl Imides	0	0	N/A	N/A
109	Non-Aromatic Hydroxylamine	0	0	N/A	N/A

	Derivatives				
110	Organic Azides	0	0	N/A	N/A
111	Organic Diselenides and Ditellurides	0	0	N/A	N/A
112	Organic Peroxy Compounds	0	0	N/A	N/A
113	Perfluorinated Hypofluorites	0	0	N/A	N/A
114	Peroxyacyl Nitrates	0	0	N/A	N/A
115	Propargyl Derivatives	0	0	N/A	N/A
116	Propyne Derivatives	0	0	N/A	N/A
117	Quinolone Derivatives	0	0	N/A	N/A
118	Quinoxaline-Type 1,4-Dioxides	0	0	N/A	N/A
119	Short-Chain Alkyltin and Alkylgermanium Halides	0	0	N/A	N/A
120	Specific 5-Substituted Uracil Derivatives	0	0	N/A	N/A
121	Sterically Hindered Piperidine Derivatives	0	0	N/A	N/A
122	Sulfonyl Azides	0	0	N/A	N/A
123	Sulfonyl Halides	0	0	N/A	N/A
124	Sultones	0	0	N/A	N/A
125	Tertiary Heterocyclic Amines	0	0	N/A	N/A
126	Thiadiazole-dioxide derivatives	0	0	N/A	N/A
127	Thiazolidinediones	0	0	N/A	N/A
128	Tri-Methylindole derivatives	0	0	N/A	N/A
129	Triarylimidazole and Structurally Related DNA Intercalators	0	0	N/A	N/A

4.5. Algorithm and descriptor generation

The structural boundaries of the alerts are derived from the chemicals included in the local training sets (see Section 4.3). For derivation of each alert mechanistically justifiable structural fragments for interaction with DNA and/or proteins are identified from the chemicals having positive data in the local training set. Additional structural fragments from the other parts of the molecules which could affect (enhance or reduce) the endpoint effect are also introduced to complete definition of most alerts.

4.6. Software name and version for descriptor generation

TIMES 2.35.1, *In vitro* Mouse Lymphoma v.07.07

4.7. Chemicals/Descriptors ratio

Provided in Section 4.4.

Section 5. Defining the applicability domain of the model – OECD Principle 3

5.1. Description of the applicability domain of the model

The domain consists of the following sub-domain layers:

1. General parametric requirements.

The variations of molecular parameters that may affect the quality of the measured endpoint significantly are included here (such as molecular weight, etc.). The domain of general parametric includes the range of variation of hydrophobicity (log K_{OW}) and Molecular weight (MW) of chemicals in training set.

2. Structural domain.

The structural component of the model is based on the structural similarity between chemicals in the training set which were correctly predicted by the model. The structural neighborhood of atom-centered fragments (accounting for the first neighbours) extracted from correctly and incorrectly predicted parent structures from the training set is used to determine this similarity.

The target chemical could contain the following types of ACF:

- Fragments present in correctly predicted training chemicals only (i.e., correct fragments),
- Fragments found both in correctly and non-correctly predicted training chemicals (i.e., fuzzy fragments). These fragments are treated as correct fragments,
- Fragments present in non-correctly predicted training chemicals only (i.e., incorrect fragments),
- Fragments not present in the training chemicals (i.e., unknown fragments).

A chemical belongs to the structural domain of the model if it could be partitioned only on correct fragments. The user is able to analyse how important are unknown and incorrect fragments (if present in the target) and to make a decision about their effect on the quality of prediction. The distribution of structural characteristics of the target chemical and accepted thresholds is used as a criterion to determine how well the target is represented

in the structural space of correctly predicted chemicals. The accepted domain thresholds for Mutagenicity are as follows:

- Correct = 100%
- Incorrect = 0%

A chemical is considered In Domain if it is classified to belong to all sub-domain levels. The information implemented in the applicability domain is extracted from the correctly predicted training chemicals used to build the model and in this respect the applicability domain determines practically the interpolation space of the model.

5.2. Method used to assess the applicability domain

The approach used to determine and assess the domain is described elsewhere:

Dimitrov S, Dimitrova G., Pavlov T., Dimitrova N., Patlewicz G., Niemela J., Mekenyan O., A stepwise approach for defining the applicability domain of SAR and QSAR models, *J. Chem. Inf. Model.*, 45, 839-849 (2005).

5.3. Software name and version for the applicability domain assessment

The LMC software OASIS Domain Manager v.1.13 (which is embedded in OASIS platform) is used to determine the applicability domain.

http://oasis-lmc.org/products/software/domain-manager.aspx

5.4. Limits of applicability

In order to belong to the model domain a target structure must meet the requirements of all layers of the domain.

- General properties requirements:

Property	Domain	Target chemical
log K_{OW}	[-10.09; 12.453]	-0.176
MW, Da	[46.01; 1256.62]	178.199

* K_{OW} is calculated by EPI Suite

- Structural domain extracted from 432 training chemicals contains:

- 1479 correct fragments,
- 175 fuzzy fragments (treated as correct fragments),
- 234 incorrect fragments.

Section 6. Defining goodness-of-fit and robustness (internal validation) – OECD Principle 4

6.1. Availability of the training set

The training set of MLA model includes 432 organic chemicals from different chemical classes and is embedded in the software implementation of the model.

6.2. Available information for the training set

Chemical names, CAS numbers, SMILES, documented data, literature sources are available.

6.3. Data for each descriptor variable for the training set

Descriptors in the models are structural alerts. The main characteristics of each alert are provided in Table 1 (Section 4.4).

6.4. Data for the dependent variable for the training set

The training set of 432 chemicals include:

- 320 chemicals having positive observed MLA data
- 112 chemicals having negative observed MLA data.

Distribution of positive/negative chemicals in the training set of model is used for estimating performance and confidence range of the so-called naïve alert which is 0.740 (0.698 ÷ 0.781)[1)].

1) Confidence range is calculated at 95% confidence level

6.5. Other information about the training set

The training set is compiled according to the recommendations described in the OECD TG490.

6.6. Pre-processing of data before modelling

Not available

6.7. Statistics for goodness-of-fit

During the internal validation the original training set is separated many times randomly into two parts – one becomes a training set and the other becomes a test set. The model is re-derived many times using each new training set. Then, performance is estimated for the training sets and test sets. The averaged value of all training set performances is compared to the averaged value of all test set performances in order to assess the amount of optimism in the goodness-of-fit (GOF optimism) in the original model. GOF optimism is calculated as average performance over training sets minus average performance over test sets. Results are provided in Table 2.

Table 2. Performance of the original model over its training set (goodness-of-fit, GOF) vs. expected performance over set different from the training set (GOF – GOF optimism).

	Performance $_{est.}$[1)] model	*p-value*[1)]	Performance[2) 3)] different set
All predictions (accuracy)	0.848 (0.814 ÷ 0.881)	< 1.0E-10	0.819
Positive chemicals (sensitivity)	0.879 (0.843 ÷ 0.914)	6.8E-7	0.837
Negative chemicals (specificity)	0.754 (0.675 ÷ 0.831)	< 1.0E-10	0.730

[1]) Confidence ranges and *p-value* are calculated at 95% confidence level

[2]) Estimated performance for training set minus GOF optimism calculated from internal validation

[3]) Estimation of expected performance over external sets (different from training sets)

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

Not performed

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

Method 1. k-fold cross-validation

In ***k-fold cross-validation*** the original training set is partitioned into k equally sized subsets. Each time a single subset is used as a test set and the remaining k-1 subsets are used as training set. In this manner the process is repeated k times and each data from the original training set is used once as a test data and k-1 times as a training data. The advantage of this method is that any data is used for both training and validation and each data is used exactly once as a test data. Commonly the 10-fold cross-validation is used

(90% training data, 10% test data). In addition, 4-fold cross validation (75% training data, 25% test data) is also performed and the results from both procedures are provided in Table 3.

Table 3. Results from *k-fold (10-fold and 4-fold)* cross-validation.

Results from k-fold cross-validation	**10-fold**		**4-fold**	
	Training sets	**Test sets**	**Training sets**	**Test sets**
Unique chemicals, %	90.0 (89.8 ÷ 90.2)	10.0 (9.8 ÷ 10.2)	75.0 (75.0 ÷ 75.0)	25.0 (25.0 ÷ 25.0)
Performance $_{est.}$, all predictions (accuracy)	0.848 (0.837 ÷ 0.858)	0.817 (0.704 ÷ 0.929)	0.847 (0.828 ÷ 0.867)	0.814 (0.740 ÷ 0.887)
p-value, accuracy	< 1.0E-10	0.0007	< 1.0E-10	2.8E-7
Performance $_{est.}$, positive chemicals (sensitivity)	0.879 (0.869 ÷ 0.889)	0.832 (0.703 ÷ 0.962)	0.878 (0.855 ÷ 0.901)	0.825 (0.729 ÷ 0.920)
p-value, sensitivity	2.2E-6	0.091	1.3E-5	0.070
Performance $_{est.}$, negative chemicals (specificity)	0.754 (0.727 ÷ 0.780)	0.671 (0.364 ÷ 0.978)	0.752 (0.710 ÷ 0.793)	0.732 (0.568 ÷ 0.896)
p-value, specificity	< 1.0E-10	5.6E-8	< 1.0E-10	< 1.0E-10

[1]) Confidence ranges and *p-value* are calculated at 95% confidence level

Method 2. *Monte Carlo* cross-validation

In ***Monte Carlo cross-validation*** the original training set is split randomly into training and test set. The advantage of this method (compared to *k-fold* cross validation) is that the proportion between training and test sets does not depend on the number of repetitions in the internal validation procedure. The ***Monte Carlo cross-validation*** (similarly to the *bootstrapping*) suppose creating a large number of new training/test sets (1000 – 10000). Results from application of this statistical method are provided in Table 5.

Table 4. Results from Monte Carlo cross-validation (1000 repetitions).

Results from Monte-Carlo cross-validation (1000 samples)	**75% training set**		**63% training set**	
	Training sets	**Test sets**	**Training sets**	**Test sets**
Unique chemicals, %	75.0 (75.0 ÷ 75.0)	25.0 (25.0 ÷ 25.0)	63.0 (63.0 ÷ 63.0)	37.0 (37.0 ÷ 37.0)
Performance $_{est.}$, all predictions (accuracy)	0.848 (0.828 ÷ 0.867)	0.821 (0.759 ÷ 0.883)	0.847 (0.821 ÷ 0.873)	0.820 (0.771 ÷ 0.869)

p-value, accuracy	< 1.0E-10	1.4E-7	< 1.0E-10	3.0E-10
Performance $_{est.}$, positive chemicals (sensitivity)	0.878 (0.858 ÷ 0.898)	0.843 (0.776 ÷ 0.910)	0.878 (0.851 ÷ 0.904)	0.841 (0.788 ÷ 0.893)
p-value, sensitivity	1.4E-5	0.026	0.0001	0.013
Performance $_{est.}$, negative chemicals (specificity)	0.754 (0.707 ÷ 0.801)	0.738 (0.603 ÷ 0.873)	0.752 (0.691 ÷ 0.813)	0.748 (0.648 ÷ 0.849)
p-value, specificity	< 1.0E-10	< 1.0E-10	< 1.0E-10	< 1.0E-10

[1]) Confidence ranges and *p-value* are calculated at 95% confidence level

6.10. Robustness - Statistics obtained by Y-scrambling

Not performed

6.11. Robustness - Statistics obtained by bootstrap

In bootstrapping a newly derived training sets is populated from the original training set of the model by random sampling with replacement until the size of the new training set reaches the size of the original training set. The data not selected for the new training set becomes the new test set. On average, about 63% of original training set data goes into the new training set (some data appear more than once) and 37% remains in the new test set. One of the advantages of this method is that the new training sets and the original training set are equally sized. The process is repeated many times and the average results are provided in Table 5.

Table 5. Results from bootstrapping (1000 repetitions).

Results from cross-validation with bootstrapping (1000 samples)	**Training sets**	**Test sets**
Unique chemicals, %	63.3 (60.4 ÷ 66.2)	36.7 (33.8 ÷ 39.6)
Performance $_{est.}$, all predictions (accuracy)	0.848 (0.813 ÷ 0.882)	0.821 (0.772 ÷ 0.869)
p-value, accuracy	< 1.0E-10	3.7E-10
Performance $_{est.}$, positive chemicals (sensitivity)	0.880 (0.844 ÷ 0.915)	0.840 (0.786 ÷ 0.894)
p-value, sensitivity	3.6E-7	0.015
Performance $_{est.}$, negative chemicals (specificity)	0.752 (0.673 ÷ 0.831)	0.753 (0.654 ÷ 0.852)
p-value, specificity	< 1.0E-10	< 1.0E-10

[1]) Confidence ranges and *p-value* are calculated at 95% confidence level

6.12. Robustness - Statistics obtained by other methods

Not performed

6.13. Comment on the internal validation of the model

The model shows good predictive performance for positive chemicals (Sensitivity) – 88% for training set, around 84% expected Sensitivity for chemicals different from the training set. On the contrary, prediction performance for negative chemicals (Specificity) is not that high (~76%) indicating for possible over-prediction. Lower Specificity could be due to the fact that the model training set is relatively small and not balanced – 75% of the training chemicals are observed positive which favours training of positive chemicals rather than negative ones.

Section 7. Defining predictivity (external validation) – OECD Principle 4

7.1. Availability of the external validation set

Not available

7.2. Available information for the external validation set

Not available

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

Not available

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

Not available

7.5. Other information about the external validation set

Not available

7.6. Experimental design of test set

Not available

7.7. Predictivity – Statistics obtained by external validation

Not available

7.8. Predictivity – Assessment of the external validation set

Not available

7.9. Comment on the external validation of the model

Not available

Section 8. Providing a mechanistic interpretation – OECD Principle 5

8.1. Mechanistic basis of the model

In vitro mouse lymphoma thymidine kinase assay (MLA) identifies agents that cause point mutations and chromosomal rearrangement. The *Tk* locus in mouse lymphoma cells is the preferred target for mammalian cell genotoxicity assays because of the broadest spectrum of damages detected at this locus. Moreover, wealth of data exists to support MLA and particularly its *Tk* locus.

Direct binding of chemicals to DNA is one of the underlying mechanisms that is responsible for MLA mutagenicity. Another type of mechanisms is associated with disturbance of protein synthesis due to inhibition of topoisomerases and interaction of chemicals with nuclear proteins associated with DNA (e.g., histone proteins). Based on that, the toxicodynamic component of the MLA models i.e. alerts (functionalities responsible for eliciting positive MLA effect) are derived to account for interactions with DNA and/or proteins. Alerts are developed by chemicals which are positive as parents only. Alerts associated with the MLA models consist of boundaries, which provide the minimum structural requirements for interacting with DNA and/or proteins. Usually, additional structural and parametric boundaries are required for completing definition of alerts.

8.2. *A priori* or *a posteriori* mechanistic interpretation

The model building followed the traditional approach:

a. Building a hypothesis for the modelled event,
b. Defining the alerting groups based on parent structures,
c. Fitting of model variable to the observed data,
d. Verification of model quality,
e. Depending on the results found in step *d* model building could continue with step *a*, *b* or *f*,
f. Determination of the applicability domain and practical application of the model.

8.3. Other information about the mechanistic interpretation

Additional information about the mechanistic interpretation could be fund in Section 2 (2.7).

Section 9. Miscellaneous information

9.1. Comments

Model predictions are fully transparent. The user is able to analyse the whole prediction process and to verify whether it concise with his/her knowledge or purposes.

For other related models, see Section 1 (1.2).

9.2. Bibliography

Additional references are not provided.

9.3. Supporting information

Additional supporting information is not provided.