European Commission

JRC QSAR Model Database

EURL ECVAM DataBase service on ALternative Methods to animal experimentation

Guideline for Authors and Editors

QMRF

The European Commission's science and knowledge service
Joint Research Centre
Directorate F - Health, Consumers & Reference Materials
Chemical Safety & Alternative Methods Unit

Database Service on ALternative Methods to Animal Experimentation

Joint Research Centre

EUR 28714 EN

This publication is a template by the Joint Research Centre (JRC), the European Commission's science and knowledge service. It aims to provide user support. The scientific output expressed does not imply a policy position of the European Commission. Neither the European Commission nor any person acting on behalf of the Commission is responsible for the use that might be made of this publication.

Contact information
Email: JRC-COMPUTOX@ec.europa.eu

JRC Science Hub
https://ec.europa.eu/jrc

JRC107493
EUR 28714 EN

PDF ISBN 978-92-79-71408-5 ISSN 1831-9424 doi:10.2760/419511
Print ISBN 978-92-79-71407-8 ISSN 1018-5593 doi:10.2760/923423

Luxembourg: Publications Office of the European Union, 2017
Ispra: European Commission, 2017

How to cite this tutorial: Triebe J, Worth A, Janusch Roi A, Coe A. *JRC QSAR Model Database:EURL ECVAM DataBase service on ALternative Methods to animal experimentation Guideline for Authors and Editors 2017*, EUR 28714 EN, Publications Office of the European Union, Luxembourg, 2017, ISBN 978-92-79-71408-5, doi:10.2760/419511, JRC107493

DISCLAIMER: Addresses to Internet sites change constantly. Therefore, all links to third parties websites included in the document have been checked last June 2017. We will review it periodically.

JRC QSAR Model Database

Guideline for Authors and Editors

INTRODUCTION

In the regulatory assessment of chemicals, **Q**uantitative **S**tructure-**A**ctivity **R**elationship (**QSAR**) **models** are playing an increasingly important role in predicting properties needed for hazard and risk assessment. The JRC QSAR Model Database provides information on the validity of QSAR models that have been submitted to the JRC. The database is intended to help to identify valid QSARs, e.g. for the registration and authorisation purposes of chemical substances within the context of REACH Registration, Evaluation, Authorisation and Restriction of Chemicals (EC 1907/2006), a European Union regulation dated 18 December 2006.

The **QSAR Model Reporting Format (QMRF)** is a harmonised template for summarising and reporting key information on QSAR models, including the results of any validation studies. The information is structured according to the OECD principles for the validation of QSAR models.
PRINCIPLE 1: "A DEFINED ENDPOINT". ENDPOINT refers to any physicochemical, biological, or environmental effect that can be measured and therefore modelled. The intent of PRINCIPLE 1 (a (Q)SAR should be associated with a defined endpoint) is to ensure clarity in the endpoint being predicted by a given model, since a given endpoint could be determined by different experimental protocols and under different experimental conditions. It is therefore important to identify the experimental system that is being modelled by the Q)SAR.
PRINCIPLE 2: "AN UNAMBIGUOUS ALGORITHM". The (Q)SAR estimate of an endpoint is the result of applying an ALGORITHM to a set of structural parameters which describe the chemical structure. The intent of PRINCIPLE 2 (a (Q)SAR should be associated with a unambiguous algorithm) is to ensure transparency in the model algorithm that generates predictions of an endpoint from information on chemical structure and/or physicochemical properties. In this context, algorithm refers to any mathematical equation, decision rule or output approach.
PRINCIPLE 3: "A DEFINED DOMAIN OF APPLICABILITY". APPLICABILITY DOMAIN refers to the response and chemical structure space in which the model makes predictions with a given reliability. Ideally the applicability domain should express the structural, physicochemical and response space of the model.
The CHEMICAL STRUCTURE (x variable) space can be expressed by information on physicochemical properties and/or structural fragments. The RESPONSE (y –
variable) can be any physicochemical, biological or environmental effect that is being predicted. According to PRINCIPLE 3 a (Q)SAR should be associated with a defined domain of applicability. Section 5 can be repeated (e.g., 5.a, 5.b, 5.c, etc.) as many time as necessary if more than one method has been used to assess the applicability domain.
PRINCIPLE 4: "APPROPRIATE MEASURES OF GOODNESS-OF-FIT, ROBUSTENESS AND PREDICTIVITY". PRINCIPLE 4 expresses the need to perform validation to establish the performance of the model. GOODNESS-OF-FIT and ROBUSTNESS refer to the internal model performance.
PRINCIPLE 5: "A MECHANISTIC INTERPRETATION, IF POSSIBLE". According to PRINCIPLE 5, a (Q)SAR should be associated with a mechanistic interpretation, if possible.

EDITORS

Developers and users of QSAR models can submit information on these models by using the QMRF. The JRC will then perform a quality control (i.e. **comprehensibility, consistency and completeness of the documentation**) of the QMRFs submitted, so that **properly documented** summaries of **QSARs** can be included in the JRC QSAR Model Database.

AUTHORS

Please, try to fill in the fields of the QMRF for the model of interest. If the field is not pertinent with the model you are describing, or if you cannot provide the requested information, please answer "no information available". **The set of information that you provide will be used to facilitate regulatory considerations of (Q)SARs.** For this purpose, the structure of the QMRF is devised to reflect as much as possible the OECD principles for the validation, for regulatory purposes, of (Q)SAR models. You are invited to consult the OECD "Guidance Document on the Validation of (Quantitative) Structure-Activity Relationship Models" that can aid you in filling in a number of fields of the QMRF.

CHAPTERS QMRF editor

1. QSAR identifier	1.QSAR identifier (title) 2.Other related models 3.Software coding the model
2. General information	1.Date of QMRF 2.QMRF author(s) and contact details 3.Date of QMRF update(s) 4.QMRF update(s) 5.Model developer(s) and contact details 6.Date of model development and/or publication7.Reference(s) to main scientific papers and/or software package 8.Availability of information about the model 9.Availability of another QMRF for exactly the same model
3. Defining the endpoint - OECD Principle 1	1.Species 2.Endpoint 3.Comment on endpoint 4.Endpoint units 5.Dependent variable 6.Experimental protocol 7.Endpoint data quality and variability
4. Defining the algorithm - OECD Principle 2	1.Type of model 2.Explicit algorithm 3.Descriptors in the model 4.Descriptor selection 5.Algorithm and descriptor generation 6.Software name and version for descriptor generation 7.Chemicals/Descriptors ratio
5. Defining the applicability domain - OECD Principle 3	1.Description of the applicability domain of the model 2.Method used to assess the applicability domain 3.Software name and version for applicability domain assessment 4.Limits of applicability
6. Internal validation - OECD Principle 4	1.Availability of the training set 2.Available information for the training set 3.Data for each descriptor variable for the training set 4.Data for the dependent variable for the training set 5.Other information about the training set 6.Pre-processing of data before modelling 7.Statistics for goodness-of-fit 8.Robustness - Statistics obtained by leave-one-out cross-validation 9.Robustness - Statistics obtained by leave-many-out cross-validation 10.Robustness - Statistics obtained by Y-scrambling 11.Robustness - Statistics obtained by bootstrap 12.Robustness - Stats obtained by other methods
7. External validation - OECD Principle 4	1.Availability of the external validation set 2.Available information for the external validation set 3.Data for each descriptor variable for the external validation set 4.Data for the dependent variable for the external validation set 5.Other information about the external validation set 6.Experimental design of test set 7.Predictivity - Statistics obtained by external validation 8.Predictivity - Assessment of the external validation set 9.Comments on the external validation of the model
8. Providing a mechanistic interpretation - OECD Principle 5	1.Mechanistic basis of the model 2.A priori or a posteriori mechanistic interpretation 3.Other information about the mechanistic interpretation
9. Miscellaneous information	1.Comments 2.Bibliography 3.Supporting information
10. Summary for the JRC QSAR Model Database	1.QMRF number 2.Publication date 3.Keywords 4.Comments
Supporting Datasets	Guidance is provided for submitting information about the training and test sets. Storage of searchable information about the training and the test sets in the database of the JRC Database will be possible if the submitter uses specific file formats (Excel file or preferably SDF files) with predefined fields.

CHAPTER 1

1. QSAR identifier

Requirements		Guideline	Examples
1.1. QSAR identifier (title)	Please provide a clear and concise title that allows the end user to decide whether the model is relevant for their needs? Please provide keywords which specify the endpoint modelled and the name of the expert system where appropriate.	Capitalise first letter only No full stop Delete “model”, since this is redundant Choose informative title	**Good examples** would be “QSAR model for polar narcosis to fathead minnow”, “BCF model based on theoretical descriptors developed by Grammatica et al”, "Quantitative Mechanistic Model for the skin sensitisation potential of Schiff Base formers”, “Derek for Windows – skin sensitisation” or "QSAR for rat chronic LOAEL" **Bad examples** would be “Epiwin” without specifying the endpoint or “Aquatic toxicity QSAR model”.
1.2. Other related models	Some models in particular those encoded into expert systems might invoke the use of a sub-model or several sub models. This heading is to flag such instances.		**E.g.** a first model would identify the presence of hazard and a second model would quantify the strength of the effect. An example might be the TOPKAT model for skin irritation; one set of models discriminate between mild/non-irritant and moderate/strong irritant and a second set discriminate between mild and non-irritant as well as between moderate and strong.
1.3. Software coding the model	Please provide the version number of the software model! Failure to provide this information might invalidate the remainder of the QMRF as the version number determines the status of development at the given point in time. Expert systems are typically updated on a periodic basis,	Please include: • Name • Description • Contact: • www	**e.g.** Derek for Windows Version 9, or TOPKAT Version 6.2. **e.g.** Derek for Windows is a particular example that released each year with a new version number.

CHAPTER 2

2. General information

Requirements		Guideline	Examples
2.1. Date of QMRF	Please provide a timeline. A timeline is needed to start the audit trail of the documentation of the model..	dd/mm/yyyy or dd month yyyy	20 August 2008
2.2. QMRF author(s) and contact details	Please provide a contact person/organisation. This is particularly useful if the QMRF author is not the same as the model developer and to provide a point of reference for further information.	Please include: • Name • Affiliation • Contact • e-mail • www (if available)	Christoph Helma; in silico toxicology gmbh; Rastatterstr. 41, CH-4057 Basel; helma@in-silico.ch; www.in-silico.ch
2.3. Date of QMRF update(s)	This should be left blank only if the model is the first to be described. In all other instances, it provides the audit trail to track additions/modifications that have been made to an existing QSAR Model. The QMRF can be updated for a number of reasons such as additions of new information (e.g. addition of new validation studies in section 7) and corrections of information.	If this is an updated version, please mention here the date of the previous versions.	

Requirements		Guideline	Examples
2.4. QMRF update(s)	Please clearly specify any updates. Any specific changes should be noted under this field. Indicate the name and the contact details of the author(s) of the updates QMRF (see field 2.3) and list which sections and fields have been modified	Please give the title of the previous version (s)	
2.5. Model developer(s) and contact details	Please complete this field. This is particularly relevant if the QMRF author and model developer are different. It also provides another point of reference for obtaining further information. Indicate the name of developer(s)/author(s), and the corresponding contact details; possibly report the contact details of the corresponding author.	Please include: • Name • Affiliation • Contact • e-mail • www (if available)	Stefano Cassani Insubria University, Department of Theoretical and Applied Sciences (DiSTA), via J.H. Dunant 3, 21100 Varese (Italy) +390332421439 stefano.cassani@uninsubria.it http://www.qsar.it/
2.6. Date of model develop-ment and/or publication	Please provide a date . This provides some indication of whether the model is leading edge science at the time of development or not. It is important information for an end user to help them determine what “value” to place on the model in a risk assessment scenario. A reference citation for the model development should also be provided in the case of models published in the peer review literature as a source of background information.		Developed in 2013, Published in 2014 [ref 4; sect 9.2]

Requirements		Guideline	Examples
2.7. Reference(s) to main scientific papers and/or software package	Please provide key published references that describe the model development. List the main bibliographic references (if any) to original paper(s) explaining the model development and/or software implementation. Any other reference such as references to original experimental data and related models can be reported in field 9.2 “Bibliography”.	Please follow reference convention.	Karelson M, Dobchev D, Tamm T, Tulp I, Jänes J, Tämm K, Lomaka A, Savchenko D & Karelson G (2008). Correlation of blood-brain penetration and human serum albumin binding with theoretical descriptors. ARKIVOC 16, 38-60.
2.8. Availability of information about the model	Please specify: Does the information provided give an appreciation of the extent of information available about the model? Is the algorithm proprietary? Is the training data set available? Indicate whether the model is proprietary or non-proprietary and specify (if possible) what kind of information about the model cannot be disclosed or are not available (e.g., training and external validation sets, source code, and algorithm).	Please avoid vague sentences like “All information in full detail is available” Specify exactly what is available	Training, selection and test sets are available. Model algorithm is available (snn file). e.g. Derek for Windows has the alerts available for inspection within the software, representative examples are provided to illustrate a given alert if available but the training set underpinning a given alert is proprietary and not available. A model published in the peer-reviewed literature might have its full training dataset made available in addition to the QSAR algorithm.
2.9. Availability of another QMRF for exactly the same model	Please identify existing QMRF(s) for the same model, but produced by a different author. Indicate if you are aware or suspect that another QMRF is available for the current model you are describing. If possible, identify this other QMRF.	If there is none, please note: None to date.	

Chapter 3

3. Defining the endpoint – OECD Principle 1

Requirements		Guideline	Examples
3.1. Species	Please provide the name of the species modelled.	If not applicable leave blank.	Pimephales promelas (Fathead minnow)
3.2. Endpoint	Please select the endpoint from the pre-defined drop down list. Choose the endpoint (physicochemical, biological, or environmental effect) from the pre-defined classification. If the pre-defined classification does not include the endpoint of interest, select "Other" and report the endpoint in the subsequent field 3.3.	If the endpoint is not listed in the dropdown list please specify and give additional information in 3.3.	3.Ecotoxic effects 3.3.Acute toxicity to fish (lethality)
3.3. Comment on endpoint	Please provide information of the underlying experimental data that has been used as the basis of developing a model. Include in this field any other information to define the endpoint being modelled. Specify the endpoint further if relevant, e.g. according to test organism such as species, strain, sex, age or life stage; according to test duration and protocol; according to the detailed nature of endpoint etc. You can also define here the endpoint of interest in case this is not listed in the pre-defined classification (see field 3.2).		Good examples include: log TA98 model; LC50 – the 50% lethality in a given strain after a 96 hour exposure period.

Requirements		Guideline	Examples
3.4. Endpoint units	Please clearly specify units of measurement.		The median lethal concentrations are reported as the logarithm of the inverse molar concentration: log(1/LC50) M
3.5. Dependent variable	Please describe clearly whether any processing was carried out to the experimental raw data to transform the endpoint to a different form for deriving a model.		Examples include transforming scores into their logarithmic values - Log EC50 rather than EC50; binning of values to give rise to bands of potency i.e. strong/moderate/mild irritancy; conversion of percentage concentrations into molar values - Log(1/EC3molar) instead of EC3%. In the case of “bands” of potency, it is important that these bands are clearly defined so that the user can evaluate the output as results such as “mild” can be ambiguous.
3.6. Experimental protocol	Please list a test procedure or protocol that provides some background information about the raw data being used. Please provide any important experimental conditions that affect the measurement and therefore the prediction.	The description of the Quality Assurance (QA) procedure for choosing the results from the literature is relevant here.	Useful information would be an OECD Test guideline, or a detailed explanation of the procedure in the primary literature reference.

Requirements		Guideline	Examples
3.7. Endpoint data quality and variability	Please clearly specify units of measurement. Provide available information about the test data selection and evaluation and include a description of the data quality used to develop the model. This includes provision of information about the variability of the test data, i.e. repeatability (variability over time) and reproducibility (variability between laboratories) and sources of error (confounding factors which may influence testing results).		Experimentally determined LC50 values for 468 industrial organic chemicals were collected from Russom et al. (1997) [ref 5; sect 9.2] (original source: U.S.-E.P.A. Duluth Fathead minnow Database)

CHAPTER 4

4. Defining the algorithm – OECD Principle 2

Requirements		Guideline	Examples
4.1. Type of model	Explain what approach has been used to derive the model.	Not too wordy	○ QSAR ○ Neural network A clear example would be "QSAR model derived from linear regression analysis", "QSAR model derived by two-group Linear Discriminant Analysis", "SAR model for mutagenicity"
4.2. Explicit algorithm	Please provide an explicit definition of the algorithm including definitions of all descriptors (including substructures where relevant).	Complete "definition" (same as 4.1), capitalise first letter only The "description" field can be more wordy	QSAR; Neural network; Multilinear regression QSAR Nonlinear QSAR: Backpropagation Neural Network (Multilayer Perceptron) regression Nonlinear QSAR: artificial neural network for classification of in vitro polyploidy in mammalian cells **Good examples** include: "LogLC50 = -0.723LogKow -2.16", "LogKp = 2.26LogP -0.5MW + 3", a substructural fragment with exclusion/inclusion conditions, e.g. RCHO for sensitisation but where R = aromatic.
4.3. Descriptors in the model	Please identify the number and the name or identifier of the descriptors included in the model. In this context, descriptors refers to e.g. physicochemical parameters, structural fragments etc.	Explain the approach used to select the descriptors. Define the descriptors/substructures adequately. This should include: ○ the approach used to select the initial set of descriptors ○ the initial number of descriptors considered ○ the approach used to select a smaller, final set of descriptors from a larger, initial set ○ the final number of descriptors included in the model ○ the type of descriptors	e.g. physicochemical parameters, structural fragments, log(1/LC50) 96h , VP-1 , MLFER_BH , nAtomLAC , HybRatio , naasC , nN, etc. .

Requirements		Guideline	Examples
4.4. Descriptor selection	Please provide a justification detailing how descriptors were selected. Indicate the number and the type (name) of descriptors initially screened, and explain the method used to select the descriptors and develop the model from them.		
4.5. Algorithm and descriptor generation	Please provide sufficient information to enable the model to be re-derived. Explain the approach used to derive the algorithm and the method (approach) used to generate each descriptor.		
4.6. Software name and version for descriptor generation	If numerical descriptors are included in the model, please provide sufficient information that enables an end user to regenerate the descriptors for a new compound. Specify the name and the version of the software used to generate the descriptors. If relevant, report the specific settings chosen in the software to generate a descriptor.	The “name” field has to be filled in. Put in software and version number only, not additional details such as the address. If no software mentioned, put “N/A”	e.g. chemical structures energy minimised using which procedure, LUMO calculations performed using VAMP, MOPAC etc.

Requirements		Guideline	Examples
4.7. Chemicals/ Descriptors ratio	Are there sufficient compounds per descriptor used in the model? This is important to judge whether the model may have been overfitted. A rule of thumb might be "5 data points per descriptor" included in the model, e.g. a linear regression model with 2 descriptors should be based on at least 10 data points (chemicals). Models with the same ratio of compounds to descriptors are questionable, due to possible overfitting. Report the following ratio: number of chemicals (chemicals from the training set) to number of descriptors , if applicable (if not, explain why).	Make sure the ratio is chemicals: descriptors	15.2 (76 chemicals / 5 descriptors)

CHAPTER 5

5. Defining the applicability domain – OECD Principle 3

Requirements		Guideline	Examples
5.1. Description of the applicability domain of the model	Please provide information which characterises the scope of the model such that the end user can determine whether the model is applicable for a specific chemical of interest or not.	Tell whether fixed or probabilistic boundaries have been used to define the applicability domain. Describe if inclusion and/or exclusion rules have been defined for the descriptor variable values. Have inclusion and/or exclusion rules regarding the chemical classes to which a substructure is applicable been included, e.g. Have inclusion and/or exclusion rules been defined for the response variable? e.g. Have rules been described for the modularity effects of the substructure's molecular environment	
5.2. Method used to assess the applicability domain	Describe the method used to assess the applicability domain of the model.	This is only relevant for models with numerical descriptors where a statistical approach has been used to define the domain of a training set. Models such as structural alerts or where a mechanistic domain has been defined can be captured in 5.1.	
5.3. Software name and version for applicability domain assessment	Examples of software might include AMBIT or an in-house algorithm. This can be left blank if no specific software was used to characterise the domain.	The "name" field has to be filled in. Put in software and version number only, not additional details such as the address. If no software mentioned, put "N/A"	QSARModel 4.0.4

Requirements		Guideline	Examples
5.4. Limits of applicability	Describe for example the inclusion and/or exclusion rules (fixed or probabilistic boundaries, structural features, descriptor space, response space) that defines the applicability domain. This will depend on what information has been provided in 5.1.		

CHAPTER 6

6. Internal validation – OECD Principle 4

Requirements		Guideline	Examples
6.1. Availability of the training set	Indicate whether the training set is somehow available (e.g., published in a paper, embedded in the software implementing the model, stored in a database) and appended to the current QMRF as supporting information (field 9.3). If it is not available, explain why. This will allow the end user to inspect the underlying basis of the model?		Example: “It is available and attached” “It is available but not attached”; “It is not available because the data set is proprietary”; “The data set could not be retrieved”.
6.2. Available information for the training set	Indicate whether the following information for the training set is reported as supporting information (see field 9.3): a) Chemical names (common names and/or IUPAC names); b) CAS numbers; c) SMILES; d) InChI codes; e) MOL files; f) Structural formula; g) Any other structural information.		
6.3. Data for each descriptor variable for the training set	Indicate whether the descriptor values of the training set are available and are attached as supporting information (see field 9.3).		○ All ○ Some ○ No ○ unknown

Requirements		Guideline	Examples
6.4. Data for the dependent variable for the training set	Indicate whether dependent variable values of the training set are available and attached as supporting information (see field 9.3).		○ All ○ Some ○ No ○ unknown
6.5. Other information about the training set	Indicate any other relevant information about the training set Give any extra information that characterises the training set in more detail?	respect data set convention	18 data points: 1 negative values; 17 positive values e.g. number and type of compounds in the training set (e.g. for models predicting positive and negative results the number of positives and the number of negatives in the training set). e.g. balanced/unbalanced dataset, type of chemicals – pharmaceuticals vs. pesticides vs. fragrances etc. Has the number of substances been specified?
6.6. Pre-processing of data before modelling	Indicate whether raw data have been processed before modelling (e.g. averaging of replicate values); if yes, report whether both raw data and processed data are given. Make it clear whether some processing of the data has been carried out.		e.g. replicate values have been averaged, data points have been omitted from the modelling for a specific reason. Ex:: Transformation of LC50 into Log1/LC50 (mol/L)

Requirements		Guideline	Examples
6.7. Statistics for goodness-of-fit	Report here goodness-of-fit statistics (r2, r2 adjusted, standard error, sensitivity, specificity, false negatives, false positives, predictive values etc).	Make proper use of superscripts and subscripts	R^2 = 0.725 (Correlation coefficient) R^2_{CVMO} = 0.744
6.8. Robustness - Statistics obtained by leave-one-out cross-validation	Has a cross validation been carried out, if so what procedure was used (LOO, LMO etc.)? Is the information sufficient to allow a judgement of the extent of model robustness to be made?	Make proper use of superscripts and subscripts	R^2 = 0.725 (Correlation coefficient) R^2_{CVMO} = 0.744
6.9. Robustness - Statistics obtained by leave-many-out cross-validation	In case cross-validation was used, is the cross-validation method clearly described or referenced? For example, there are different ways of performing leave-many-out validation.	Make proper use of superscripts and subscripts	R^2 = 0.725 (Correlation coefficient) R^2_{CVMO} = 0.744

Requirements		Guideline	Examples
6.10. Robustness - Statistics obtained by Y-scrambling	Report here the corresponding statistics and the number of iterations. In case Y-sampling was applied, please add the resulting statistics.	Make proper use of superscripts and subscripts	R^2 = 0.725 (Correlation coefficient) R^2_{CVMO} = 0.744
6.11. Robustness - Statistics obtained by bootstrap	Report here the corresponding statistics and the number of iterations. In case bootstrapping was applied, please add the methodological details and resulting statistics.	Make proper use of superscripts and subscripts	R^2 = 0.725 (Correlation coefficient) R^2_{CVMO} = 0.744
6.12. Robustness - Statistics obtained by other methods	Report here the corresponding statistics in case another cross-validation methods was applied, please describe this clearly and provide the resulting statistics.	Make proper use of superscripts and subscripts	R^2 = 0.725 (Correlation coefficient) R^2_{CVMO} = 0.744

CHAPTER 7

7. External validation – OECD Principle 4

Requirements		Guideline	Examples
7.1. Availability of the external validation set	Has an external validation been carried out? If not, has an explanation been provided as to why an external validation was not carried out? Is the test set available? Is information provided that allows the end-user to determine whether the representativeness of the dataset was taken into account when selecting the chemicals in the test set? Is information available about the experimental data for the test set of chemicals?	Indicate whether an external validation set is available and appended to the current QMRF as supporting information (field 9.3). If it is not available, explain why.	○ YES ○ NO
7.2. Available information for the external validation set	Please provide the test (validation) set of chemicals and identifiers (e.g. Name, SMILES, CAS#, InChI, MOL file, Formula).	Chemname: Yes/No SMILES: Yes/No CAS RN: Yes/No InChI: Yes/No MOL file: Yes/No Formula: Yes/No	
7.3. Data for each descriptor variable for the external validation set	Please provide the descriptor values.		○ All ○ Some ○ No ○ unknown

Requirements		Guideline	Examples
7.4. Data for the dependent variable for the external validation set	Indicate whether dependent variable values of the external validation set are somehow available and attached as supporting information (see field 9.3).	Chemname: Yes/No SMILES: Yes/No CAS RN: Yes/No InChI: Yes/No MOL file: Yes/No Formula: Yes/No	
7.5. Other information about the external validation set	Has the approach for selecting test set chemicals been described?	respect data set convention	Ex.1: 18 data points: 1 negative values; 17 positive values Ex.2: The external prediction set consists of 200 heterogeneous organic compounds with a range of log(1/LC50) values from 0.84 to 6.72. Training and prediction sets are structurally balanced, since the splitting was based on the structural similarity analysis (SOM, see Section 6.5) Ex.3: “External validation set with 56 compounds appended”.
7.6 . Experimental design of test set	Indicate any experimental design for getting the test set (In case that experimental testing was based on prior chemicals selection, make sure that the method for selecting chemicals is described clearly.		e.g. by randomly setting aside chemicals before modelling, by literature search after modelling, by prospective experimental testing after modelling, etc.).

Requirements		Guideline	Examples
7.7. Predictivity - Statistics obtained by external validation	Report here the corresponding statistics. In the case of classification models, include false positive and negative rates. Report statistics based on external validation.	Make proper use of superscripts and subscripts	R^2 = 0.725 (Correlation coefficient) R^2_{CVMO} = 0.744
7.8. Predictivity – Assessment of the external validation set	Discuss whether the external validation set is sufficiently large and representative of the applicability domain. Describe for example the descriptor and response range or space for the validation test set as compared with that for the training set.	Make proper use of superscripts and subscripts Here the descriptor values of the chemicals predicted by the model (training set) should be compared with the descriptor value range of the test set. In addition the distribution of the response values of the chemicals in the training set should be compared to the distribution of the response values of the test set.	R^2 = 0.725 (Correlation coefficient) R^2_{CVMO} = 0.744
7.9. Comments on the external validation of the model	Add any other useful comments about the external validation procedure.	Make proper use of superscripts and subscripts. Do not use subjective words, like “good”	R^2 = 0.725 (Correlation coefficient) R^2_{CVMO} = 0.744

CHAPTER 8

8. Providing a mechanistic interpretation – OECD Principle 5

Requirements		Guideline	Examples
8.1. Mechanistic basis of the model	Provide information on the mechanistic basis of the model (if possible). In the case of SAR, you may want to describe (if possible) the molecular features that underlie the properties of the molecules containing the substructure (e.g. a description of how sub-structural features could act as nucleophiles or electrophiles, or form part or all of a receptor-binding region). In the case of QSAR, you may give (if possible) a physicochemical interpretation of the descriptors used (consistent with a known mechanism of biological action). If it is not possible to provide a mechanistic interpretation, try to explain why.		e.g. In the case of a SAR-based model, have the underlying molecular features (substructures) been provided? e.g. In the case of a QSAR has a plausible interpretation of the descriptors/substructures that is consistent with known mechanism of (biological) action been provided?
8.2. *A priori* or *a posteriori* mechanistic inter-pretation	Indicate whether the mechanistic basis of the model was determined a priori (i.e. before modelling, by ensuring that the initial set of training structures and/or descriptors were selected to fit pre-defined mechanism of action) or a posteriori (i.e. after modelling, by interpretation of the final set of training structures and or descriptors).		Ex.: A posteriori mechanistic interpretation, even if in this case is particularly difficult: The equation of the PaDEL-Descriptor model included in QSARINS 2.2 is : log(1/LC50) 96h =………………….
8.3. Other information about the mechanistic inter-pretation	Report any other useful information about the (purported) mechanistic interpretation described in the previous fields (8.1 and 8.2) such as any reference supporting the mechanistic basis. Give literature references that support the (purported) mechanistic basis.		Ex.: No other information available.

CHAPTER 9

9. Miscellaneous information

Requirements		Guideline	Examples
9.1. Comments	Please add any additional comments to help build up an appreciation of the level of use of the model or particular scenarios where it has been successfully applied. Equally it would be useful to highlight scenarios where the model was not successfully applied so as to gain an appreciation of the limitations of the model.		e.g. if a model was primarily useful in screening; the functionality of the similarity assessment in TOPKAT was of value in “validating” a prediction for a chemical of interest that could be helpful as part of an integrated testing strategy; a model was of sufficient robustness and accuracy that it could be used as a standalone models.
9.2. Bibliography	Please add references that might provide further background information or context of use of the model.	Please respect reference convention. Please provide a URL to the paper.	Ex.1: García-Domenech R, de Julián-Ortiz JV & Besalú E (2006). True prediction of lowest observed adverse effect levels. Molecular Diversity 10,159-168. Ex.2: Chirico N & Gramatica P (2012). Real External Predictivity of QSAR Models. Part 2. New Intercomparable Thresholds for Different Validation Criteria and the Need for Scatter Plot Inspection, Journal of Chemical Information &.Modeling 52 (8), 2044–2058. DOI: 10.1021/ci300084j
9.3. Supporting information	Please add any other supporting information (e.g. external documents) to the QMRF.	Please make sure that the sdf files are readable, and that the structures render properly Please make sure that the numbers of chemical structures in the sdf files correspond with the stated sizes of the training set (Section 6.5) and test set (Section 7.5)	

10. Summary for the JRC QSAR Model Database

The summary section is specific for the JRC QSAR Model Database. If the model is submitted to JRC for inclusion in the JRC Database of QSAR models, then this summary is compiled by JRC after QMRF submission. **The QMRF author does not have to fill in any of the fields of the summary section.**

10.1 QMRF number: A unique number (numeric identifier) is assigned to any QMRF that is published in the JRC QSAR Model Database. The number encodes the following information: Q YEAR-ENDPOINT-No ***Example: Q11-417-002 refers to a QMRF published in 2011, for the endpoint 4.17. It is the second QMRF published in 2011.*** The number is unique for any QMRF uploaded and stored in the JRC QSAR Model Database.
10.2 Publication date: The date (day/month/year) of publication in the JRC Database is reported here.
10.3 Keywords: Any relevant keywords associated with the present QMRF are reported here.
10.4 Comments: Any comments that are relevant for the publication of the QMRF in the JRC Database (e.g., comments about updates and about supporting information) are reported here.

Supporting Datasets

Guidance is provided here for submitting information about the training and test sets. Storage of searchable information about the training and the test sets in the database will be possible if the submitter adheres to the following predefined fields.

Mandatory:

- Chemical Name (IUPAC)
- Chemical Name (Not IUPAC)
- CAS Number
- SMILES
- InChI
- MOL *(file name is reported for Excel files; if it is an SDF file, coordinate can be simply included in it).*
- Structural Formula
- Dependent Variable
- Descriptor1 Value *(the name of the descriptors should be specified by the user)*
- Descriptor 2 Value
- Descriptor 3 Value
- Descriptor 4 Value
- Descriptor X Value

Data should ideally be provided in SDF format, but failing that in EXCEL format. Any other supporting information (e.g. background documents) should be provided in PDF format.

Additional Advice

FILENAMING

- Use meaningful names for the QMRF and attachments
- QMRF should name developer and model
- The training and test sets should include the number of compounds
- Example report: Molcode_polyploidy
- Example training set/ test set: Soil_sorption_108_training ; Soil_sorption_54_test

SENDING FILES BY E-MAIL

- E-mail servers sometimes remove xml files as potentially harmful attachments, therefore convert them to zip files
- Example: Molcode_polyploidy.zip

GENERAL

- When there are multiple tabs in a field (e.g. 3.2 Endpoint), delete empty fields
- If a section is empty, leave empty. Delete terms such as “N/A” or “n/a” (this is implicit)

ATTACHMENTS

- Check that sdf files are readable, and that the structures render properly
- Check that the numbers of chemical structures in the sdf files correspond with the fstated sizes of the training set (Section 6.5) and test set (Section 7.5)
- To check how it will look when published, use “save as pdf file”, then view the pdf. Note that the pdf format introduces spaces which are not seen in the html version.

GETTING IN TOUCH WITH THE EU

In person

All over the European Union there are hundreds of Europe Direct information centres. You can find the address of the centre nearest you at: http://europea.eu/contact

On the phone or by email

Europe Direct is a service that answers your questions about the European Union. You can contact this service:

- by freephone: 00 800 6 7 8 9 10 11 (certain operators may charge for these calls),

- at the following standard number: +32 22999696, or

- by electronic mail via: http://europa.eu/contact

FINDING INFORMATION ABOUT THE EU

Online

Information about the European Union in all the official languages of the EU is available on the Europa website at: http://europa.eu

EU publications

You can download or order free and priced EU publications from EU Bookshop at: http://bookshop.europa.eu. Multiple copies of free publications may be obtained by contacting Europe Direct or your local information centre (see http://europa.eu/contact).

KJ-NA-28714-EN-N

JRC Mission

As the science and knowledge service of the European Commission, the Joint Research Centre's mission is to support EU policies with independent evidence throughout the whole policy cycle.

Publications Office

EU Science Hub
ec.europa.eu/jrc

@EU_ScienceHub

EU Science Hub - Joint Research Centre

Joint Research Centre

EU Science Hub

doi 10.2760/419511
ISBN 978-92-79-71408-5