Instructions for the KNIME workflow "PrecisionTox Concentration response modelling Reporting Template BMD"

Author: Stefan Scholz, Helmholtz Centre for Environmental Research – UFZ, Dpt. Bioanaytical Ecotoxicology, Permoserstr. 15, 04318 Leipzig, Germany, stefan.scholz@ufz.de

Concentration response analysis for mortality, immobilisation and cytotoxicity endpoints in the EU project PrecisionTox.

Goals:

- Calculate effect concentrations (ECx) and BMD (benchmark dose) for a given BMR (benchmark response level). BMDs refer to a defined response level (e.g. 10 % more mortality in relation to controls).
- The KNIME workflow also provides a reporting file with raw data included plus the modelled effect concentration. The reporting file is structured to allow automatic compiling of effect data into a database.

Overview:

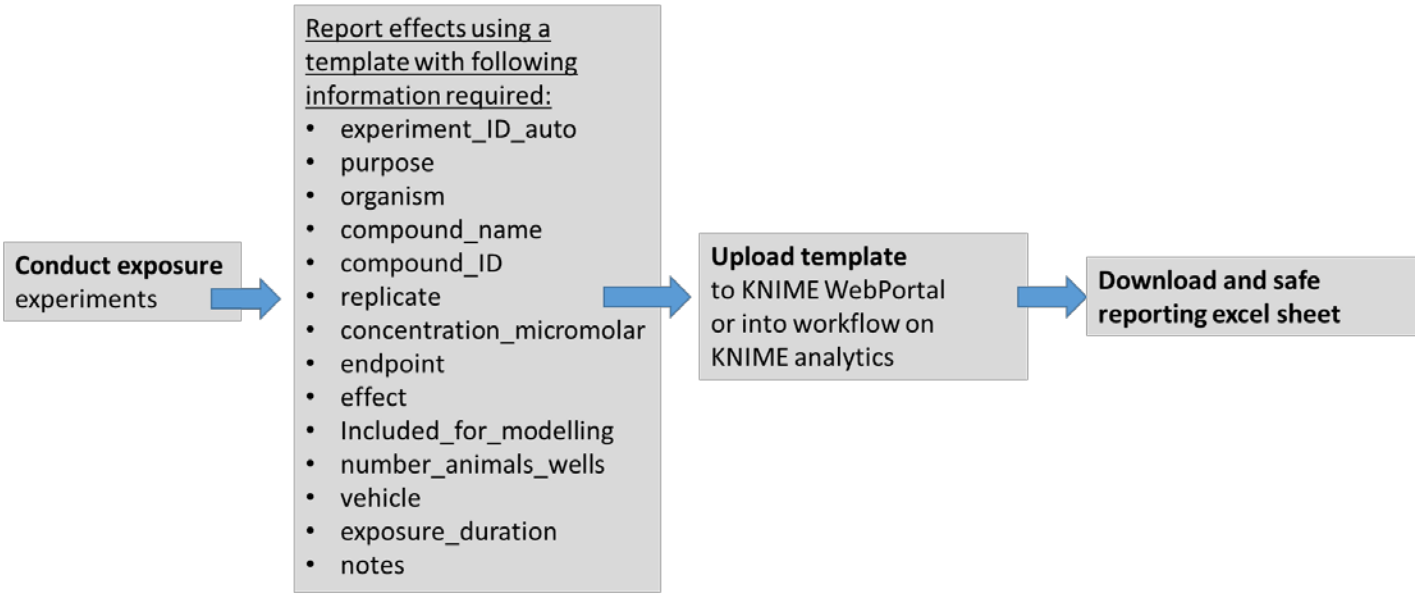


Fig. 1: Overview of the workflow needed to create the reporting file.

Procedure:

1. **Exposure experiments**
   Conduct exposure experiments and report data using the provided template.

2. **Estimate effect level for each chemical, replicate and effect concentration**

Report the effect level for each chemical, exposure concentration and replicate. The effect level can be a percent value (e.g. percent immobilized organisms), or a number (e.g. cell count). An important prerequisite is that effect data should be increasing, for instance in case of percent values and if an effect occurs, controls should have low percentage values and increasing concentrations should approach 100 %. In case your data are different (100 % for controls), please transform your data.

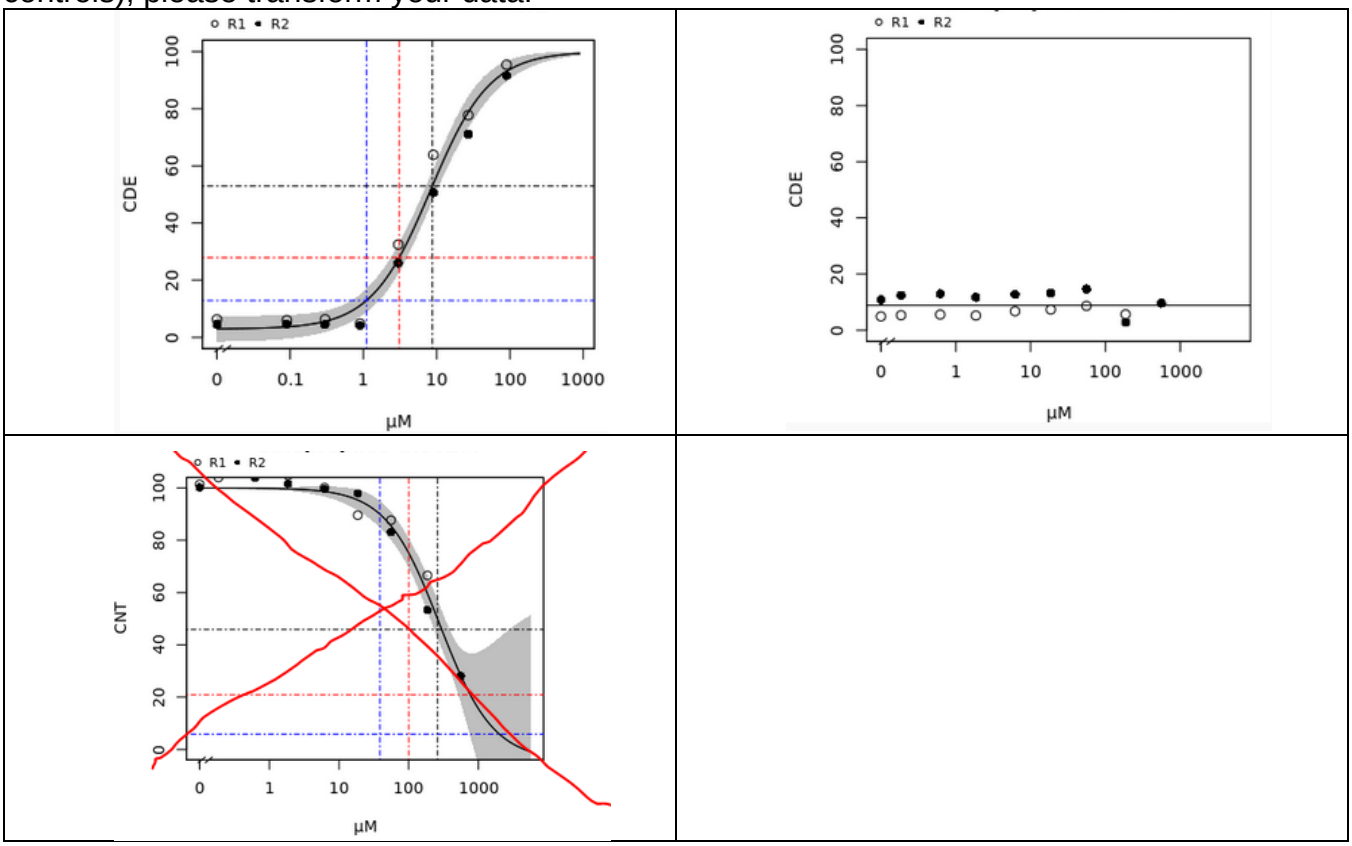


Fig. 2: Examples of concentration-response modelling using the provided KNIME workflow: The lower figure indicates a case with decreasing percentage values as effect estimate. This requires transforming of the data since otherwise BMDs are not calculated correctly. The lines refer to the BMDs with the different response levels (blue = BMR of 10 %, red = BMR of 25 %, black = BMR of 50 %)

3. Template

Use the provided excel template to upload data for concentration response modelling into the KNIME workflow

There are example data included in the template. Please remove these data prior to fill in your own data.

4. Load raw data (with the reporting template)

Use the **sheet** (“Raw_data”) of the reporting template to report the effects for each test concentration, chemical and replicate. Multiple test systems or organism, respectively, can be reported in the same sheet (one file per partner).

The following information is required per row:

Parameter	Description
experiment_ID_auto	User specific experiment ID.
purpose	Purpose of experiment e.g. dose range finding for omics; to be selected from drop-down list, at present only one option.
organism	species Latin name, to be selected from pop-up menu
compound_name	Free text that can be added by the user. Will be displayed in headers of graphs.
compound_ID	Chemical ID, user-defined
replicate	Indicate the experiment replicate number (this is not the number of animals per replicate.
concentration_micromolar	The test concentration in µM
endpoint	Endpoint abbreviation can be selected from a drop-down list. See excel sheet in template and report "Endpoint_explanation" for description of endpoints.
effect	Depending on the endpoint this can be a percent value, or a count.
Included_for_modelling	Indicate by yes/no (from drop-down menu), whether the result should be included in the subsequent concentration-response modelling.
number_animals_wells	The number of animals (or wells for cell cultures) tested per each concentration and replicate.
vehicle	indicates whether test chemical was initially solubilised in water, the exposure medium or DMSO (select from a drop down list)
exposure_duration	Exposure duration (hours of exposure) (select from a drop-down list)
notes	Any comment you would like to provide. E.g. give reasons when you want to exclude an effect value or have observation to be reported for a particular effect concentration.

Details for each of the columns in the reporting table can also be found in sheet 2 (“Parameter_description”).
To avoid syntax errors and facilitate fusion of tables and automated assessment some columns are filled automatically or information can only be provided from a drop-down list. **For this purpose, it is very important that you extend the table by copy/paste of rows to preserve the field codes and that you do not edit fields and codes**.

5. Calculation of effect concentration

Effect concentrations, BMDs and BMDLs are automatically calculated for all compounds and endpoints by uploading the template to the KNIME workflow. At present the workflow is only able to deal with concentration response modelling of percentage values and for log-logistic modelling of increasing effect levels (Hill). The modelling of effect concentrations is based on the R packages drc and bmd. If minimum and maximum of the Hill model are 0 and 100, respectively the ECx is equal to the BMD of the same effect level. For more explanation of BMD estimation consider e.g. the following web page:
https://www.chemsafetypro.com/Topics/CRA/What_Is_Benchmark_Dose_%28BMD%29_and_How_to_Calculate_BMDL.html

When the workflow is executed on a KNIME server, you will be guided through the workflow (upload data, specify asymptotic levels, selecting and downloading results). If the workflow is executed within KNIME analytics, make sure that you execute the components of the workflow subsequently and add the required information at each component.

In the workflow, flag any results where the model does not describe the data appropriately. You can identify these data sets by visual inspection and checking various statistical parameters (p-values from trend analysis, variance_explained, AIC_delta). Your flag will be transmitted to the final reporting file. You only need to select the first row of the appropriate data set (with the raw data graph; see Fig. 3).

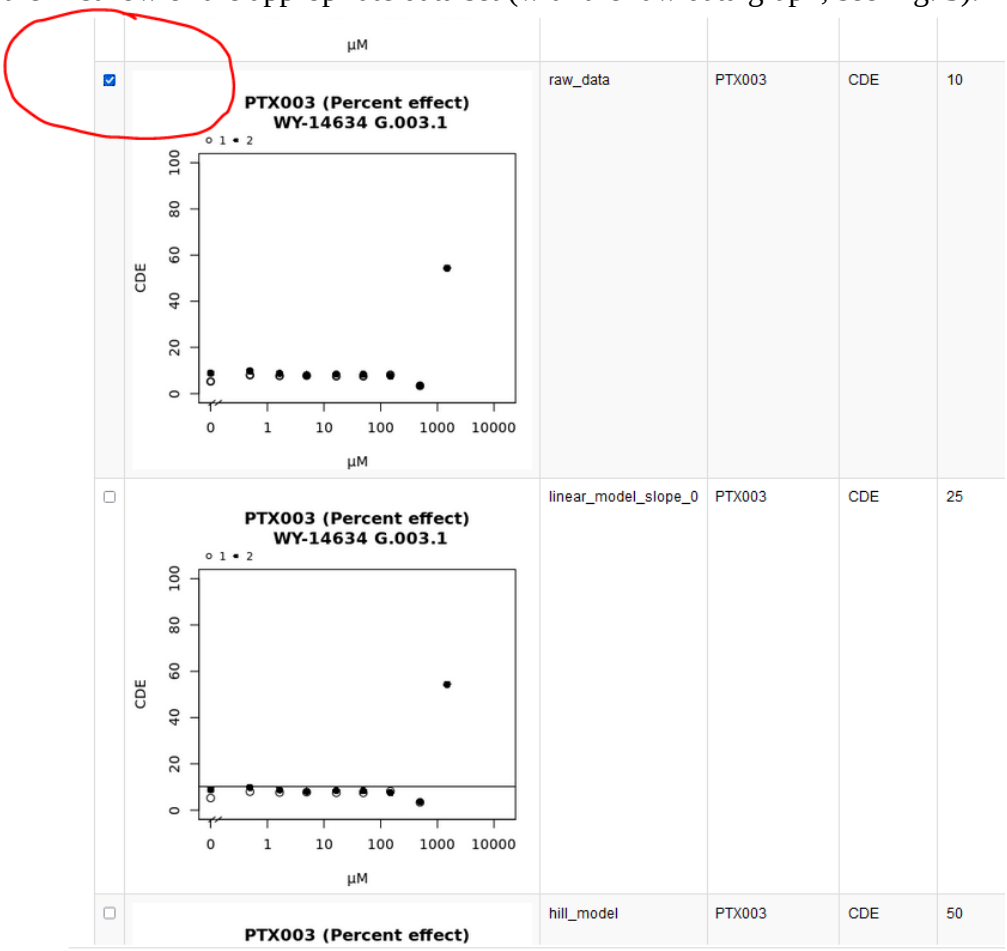


Fig. 3: Close-up from the KNIME workflow. Select datasets that should be flagged by checking the first row for the appropriate dataset

For each dataset three rows are provided referring to the BMDs 10, 25 and 50. Furthermore, the three rows host raw data plots, a linear plot for a slope of zero and a plot for the Hill model.

The primary aim of the workflow is BMD calculation and reporting the calculated values. It is not intended to provide publication-ready graphs.

After checking the tick box for flagging the user can review the selection and assignment of the flag in the next window and go back if a revision is needed.

Results of BMD calculation and curve fittings are provided also as a separate sheet in the excel report)

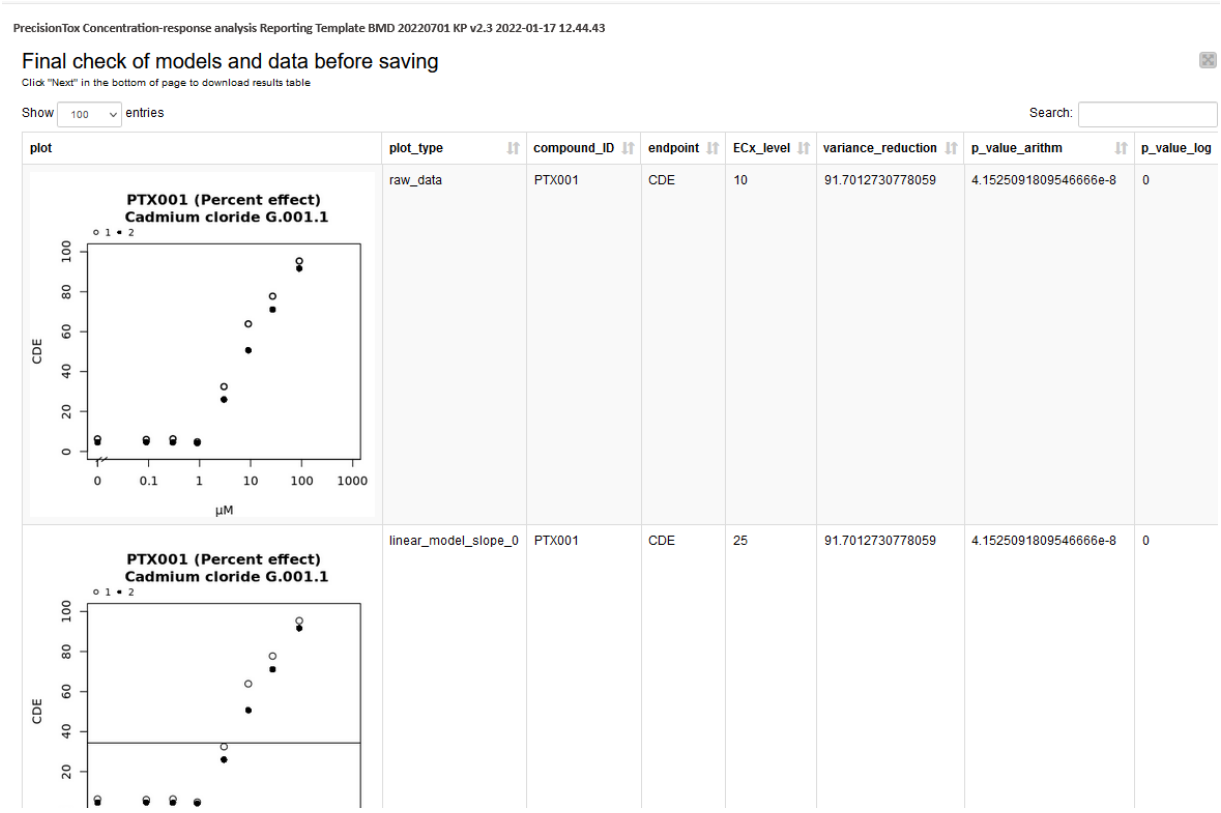


Fig. 6: Close-up of the Excel sheet that shows curve fitting data, statistics and modelled effect concentrations

5. Parameter descriptions

The excel reporting file contains a sheet in which all parameters of all sheets are described.

6. Reporting

The excel reporting file contains a sheet called “Reporting” that combines raw data, experimental parameters and modelling results. This represents the final data that will be aggregated into a database.
The reporting sheet contains the headers of the raw data sheet along modelling data:

Parameter	Description
experiment_ID_auto	Experiment ID
purpose	Purpose of experiment e.g. dose range finding
organism	species Latin name, to be selected from pop-up menu
compound_name	Free text that was added by the user. Will be displayed in headers of graphs.
compound_ID	The PrecisionTox Chemical number, was selected from a drop-down list.
replicate	Indicate the experiment replicate number (this is not the number of animals per replicate.
concentration_micromolar	The test concentration in µM

endpoint	Endpoint abbreviation which was selected from a drop-down list. See excel sheet in template and report "Endpoint_explanation" for description of endpoints.
effect	Depending on the endpoint this can be a percent value, or a count.
Included_for_modelling	User indicated by yes/no (from drop-down menu), whether the result should be included in the subsequent concentration-response modelling.
number_animals_wells	The number of animals (or wells for cell cultures) tested per each concentration and replicate.
vehicle	indicates whether test chemical was initially solubilised in water, the exposure medium or DMSO (select from a drop down list)
exposure_duration	Exposure duration (hours of exposure) (select from a drop-down list)
notes	Any comment provided by the user. E.g. reasons whey excluding an effect value or have observation to be reported for a particular effect concentration.
variance_reduction	Percent variance explained by the Hill (log-logistic) modelling
p_value_arithm	A p-value for indicating whether data show a trend for increase
p_value_log	A p-value for indicating whether data show a trend for increase
AIC _delta	Difference of the AICs for linear and Hill model. A value >>0 indicates that the Hill model is better describing the data. Values close to 0 indicate that the data are variable, with a weak or no trend to be described by a Hill model
model_status	Indicates whether the modelling could be computed (is not in indicator for validity)
user_flag	Indicates whether the user has considered the model as invalid (based on AIC ratio, p-values and/or VarianceRed)
BMD10	Modelled BMD for the ECx level (=BMR) of 10
BMD25	Modelled BMD for the ECx level (=BMR) of 25
slope	Slope of the Hill curve
df	The degrees of freedom (number of observations minus 2)
min	The minimum of the modelled Hill curve
max	The maximum of the modelled Hill curve

The reporting table is automatically calculated from the user input and the modelling.

7. Validating the curve fits

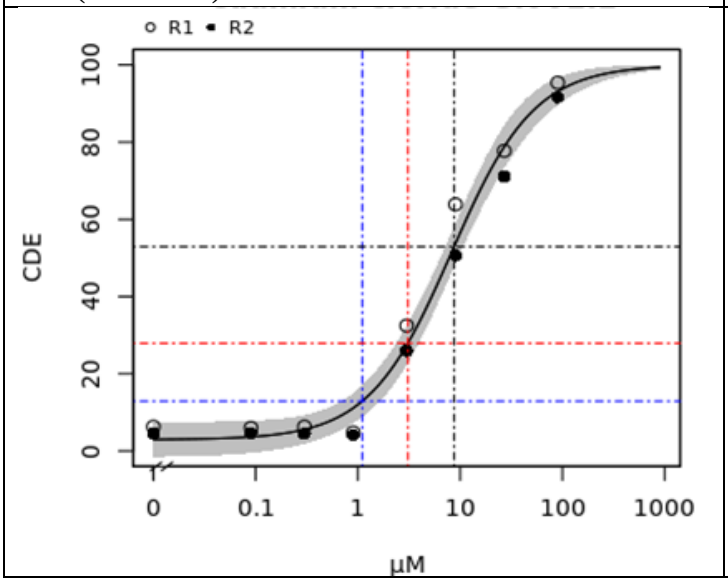
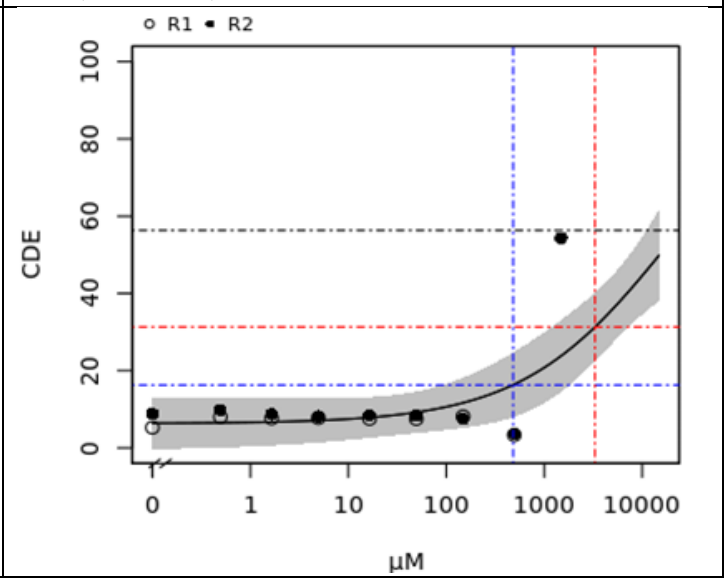
Please note that the KNIME workflow is an automated workflow. Hence, it will try to fit a curve to data, even in case that it may make no sense (e.g. no effect, very variable data). There are, however, a few metrics you can use to decide, whether the calculated BMDs should be considered or whether it may rather be concluded that there is no effect. **Hence, for reporting of the BMDs, this may require your manual editing of the table to delete BMDs in the reporting sheet when they do not match certain criteria (see below).**

Aikaike information criterion (AIC)

The AIC is a parameter for a goodness of fit that allows to compare different models (developed by the Japanese statistician Aikaike). The AIC includes a kind of penalty for increasing numbers of parameters (if you have many parameters e.g. an exponential curve of xxx order, you may always be able to fit a curve that hits every data point). The AIC **is not** useful to compare different datasets since absolute values impacts on the magnitude of the value. However, you can use it to compare the fitting of different models to the same data set. In the PrecisionTox workflow this is implemented by comparison of the Hill model with a linear model of slope

0. The model with the lower AIC has a better fit. For instance, if the linear model with slope 0 has a lower AIC this indicates that there rather is no trend or the variability is so large that it is difficult to model the data, or the Hill model is not appropriate. In Fig. 3 two examples are shown. In A the AIC is very different for the Hill model and the linear curve (slope 0) model. In example B – not surprising – both AIC are very close indicating that the Hill model fit is not very good. Here a different model may be considered, but more data for the higher effect range would probably needed as well and with more data the Hill model fit may become better.

Variance reduced (VarienceRed)
There is another intuitive value included in the KNIME workflow. This is based on the calculation of residuals, i.e. the distance of the data points to the fitted curve. If the curve fit describes the data very well, the residuals become very small after curve fitting. If you compare them to the original non fitted data (relative to the mean) you can calculate a value that describes to which extend the variance (regarding residuals) was reduced. Values such as 80 or 90 percent indicate a very good fit. The formula will be added to the ANNEX.
Below are two examples with a rather different VarianceRed values (Fig. 3). As you can see from the curve in B, the data points are still quite distant to the model in contrast to the graph in A. You may consider using a different model but there is also a range of data where more data points might be needed to conduct a proper curve fitting.

A: VarianceRed = 92 % AIC (linear model slope 0) =162 AIC (Hill model) = 101	B: VarianceRed = 33 % AIC (linear model slope 0) = 134 AIC (Hill model) = 130
	

8. References

For more details on the R packages used for concentration-response modelling please refer to the following publications:

Jensen, S.M., Kluxen, F.M., Streibig, J.C., Cedergreen, N., Ritz, C., 2020. bmd: an R package for benchmark dose estimation. PeerJ. 8, e10557.
Ritz, C., Baty, F., Streibig, J.C., Gerhard, D., 2015. Dose-response analysis using R. PLoS ONE. 10, e0146021.
Ritz, C., Streibig, J., Bioassay Analysis Using R. Journal of Statistical Software. 12, 1-22.
Ritz, C., Streibig, J.C., 2005. Bioassay analysis using R. J Stat Softw. 12, 1-22.

ANNEX

The annex describes main elements of calculations
conducted in the KNIME workflow. Please note that details of coding and calculation can also be obtained via the R-scripts embedded in the KNIME workflow.

R-code for BMD calculation

Required R-packages: drc **and bmd**

Step	Command	Explanation
Hill (log-logistic) curve modelling:	LL_4P<- drm(probe~conc , fct = LL.4(fixed = c(NA, Hill_min, Hill_max, NA)))	Probe-conc refers to the data table with concentration and effect data. Hill_min and Hill_max specify the asymptotes of the Hill model, can be user defined or modelled.
Calculation of BMR level in modelled curve (to account for offsets, I.e. background effects in controls)	BMR10<-10/((upasymp-lowasymp)/100) BMR25<-25/((upasymp-lowasymp)/100) BMR50<-50/((upasymp-lowasymp)/100)	Upasymp and lowasym refer to the modelled or specified minimum and maximum of the Hill model). A factor of 100 is required to convert from percent to frequency.
Calculation of BMDs	bmd_values10<- bmd(LL_4P, BMR10/100, def="relative") bmd_values25<- bmd(LL_4P, BMR25/100, def="relative") bmd_values50<- bmd(LL_4P, BMR50/100, def="relative")	The previous defined BMR (including the offset) is used to calculate the corresponding BMD

R-code for calculation of variance reduction

Step	Command	Explanation
Get predicted values (based on Hill model) for each concentration)	Predicted<-predict(LL_4P)	LL_4P is the name of the previously calculated Hill model
Calculate standard deviation	SDdata<-sd(probe)	Standard deviation of effect data prior to modelling
Residuals calculation	Residuals<-abs(Predicted-probe)	Residuals represent the distance of effect data to the modelled curve
Standard deviation of residuals	SDResiduals<-sd(Residuals)	
Estimate the variance reduction obtained by modelling	VarianceRed<-100-SDResiduals/SDdata*100	If the model describes the observation very well, the residuals become small and hence, also the resulting standard deviation of all residuals. The value “VarianceRed” is reported as the reduction of variance obtained by the model.