(Q)SAR APPLICATION TOOLBOX

VERSION 1.1

Strategies for grouping chemicals for data gap filling for acute aquatic toxicity endpoints

APRIL 2010

Copyright © 2010 Organisation for Economic Cooperation and Development

TABLE OF CONTENTS

Introduction

It is the purpose of this document to provide guidance on the use of profilers and databases associated with the OECD (Q)SAR Application Toolbox Version 1.1 (here after noted as the Toolbox) with the express aim of providing non-prescriptive guidance to help the user building categories that are mechanistically and structurally robust to maximise success in filling data gaps for acute aquatic toxicity endpoints, in particular for a single target chemical.

This document is for users having some experience with the workflow of the Toolbox. OECD recommends that users first read the manual for getting started using the Toolbox and the "Guidance Document for Using the OECD (Q)SAR Application Toolbox to Develop Chemical Categories According to the OECD Guidance on Grouping of Chemicals". These and other documents and training aids are available at www.oecd.org/env/existingchemicals/qsar.

The Work Flow of the OECD (Q)SAR Application Toolbox

The Toolbox has six work modules which are used in a sequential work flow. These six modules and a summary of their functions are:

Step 1 Chemical Input: This module provides the user with several means of entering the chemical of interest or target chemical. Since all subsequent functions are based on chemical structure, the goal here is to make sure the molecular structure assigned is the correct one.

Step 2 Profiling: "Profiling" refers to the electronic process of identify structural and mechanistic properties of the target chemical, which are stored in the Toolbox and can subsequently be used in the module on category definition to group the target chemical with similar chemicals.

Step 3 Endpoints: "Endpoints" refer to the electronic process of retrieving physical chemical properties, the environmental fate and toxicity data which are stored in the Toolbox. This data gathering can be executed in a global fashion (i.e., collecting all data of all endpoints) or on a more narrowly defined basis (e.g., collecting data for a single or limited number of endpoints).

Step 4 Category Definition: This module provides the user with several means of grouping chemicals into a (toxicologically) meaningful category that includes the target molecule. This is the critical step in the workflow and several options are available in the Toolbox to assist the user in refining the category definition via subcategorization.

Step 5 Filling Data Gaps: This module provides the user with three options for making an endpoint-specific prediction for the untested chemical; in this case the target molecule. These options, in increasing order of complexity, are by read-across, by trend analysis, and through the use of QSAR models.

Step 6 Report: The final module provides the user with a downloadable written audit trail of what the Toolbox did to arrive at the prediction.

Step 4, defining the chemical category, is based on the outcome of the profiling (step 3). Since defining the chemical category for data gap filling is the critical step in the workflow of the Toolbox the selection and use of profiler is an important process. The Toolbox provides several options (i.e., profiler) to assist the user in defining and refining the category definition. However, the order, in which these profilers are used and the overall results when using selected profilers in sequence affect the structural boundaries of the category and the confidence that the category is correct. Thus, there is a need for guidance on using profilers associated with the Toolbox to define categories for acute aquatic toxicity endpoints. Since data gap filling for a single target chemical is different than working with an inventory of chemicals the profiling strategies will be different.

Aquatic Toxicity Endpoints

Acute toxicity to aquatic animals is an adverse outcomes typically measured as individual mortality or reduced population levels following short-term (i.e., 96 hours or less) exposure. Similar acute toxicity to exponentially growing single cell organisms such as algae, cyanobacteria or ciliates is an adverse outcome measured as an inhibition of the population growth. Acute aquatic toxic effects are the most data-rich endpoints in the Toolbox. Organism and endpoints (duration and exposure regimes) found in the Toolbox and typically integrated into such analyses including (1) fish 24-, 48-, 72-, or 96-hr mortality most often measured as the LC50 (concentration, which results in lethality of 50% of the tested individuals) value, (2) daphnid 24- or 48-hr EC_{50} (concentration, which results in immobility of 50% of the tested individuals or LC50 value, (3) algae 24- 48- 72- or 96-hr EC_{50} (concentration with results in a 50% reduction on the exponential growth rate) value[1], and (4) ciliate protozoa (*Tetrahymena pyriformis*) 2-day IGC_{50} (concentration with results in a 50% impairment in population growth as compared to controls) values.

The depth and breadth of the acute aquatic toxicity data plus more than three decades of testing and experience in modelling such acute effects makes it possible to use profilers a higher level of complexity for categorizing acute aquatic endpoints. This complexity is aided by several factors including (1) acute aquatic toxicity has a water solubility-related minimal toxicity, which, while it may be superseded by other modes of action toxicity, does form a baseline for potency, (2) the majority of the industrial organic chemicals, especially the most common ones, are baseline toxicants, which acting via the nonpolar narcosis mode of toxic action (either based on experimental data or because of the lack of information indicating otherwise), and (3) several profilers found in the Toolbox including **EcoSAR classification**, **OASIS acute toxicity mode of action** and **Verhaar classification** have been developed with the aid of these robust data. Each of these latter three profilers has their advantages and disadvantages and they are often used in parallel. An examination of structural alerts most often associated with excess acute aquatic toxicity reveals them to be remarkably similar to those found in the **Protein binding** profiler. Thus, the **Protein binding** profiler is included along with the three above noted profilers as an initial battery of profilers relevant to acute aquatic toxicity endpoints.

Initial Battery of Profilers Relevant to Acute Aquatic Toxicity Endpoints

The initial battery of Toolbox profilers relevant to acute aquatic toxicity endpoints includes the 1) **EcoSAR classification**, 2) **OASIS acute toxicity mode of action,** 3) **Protein binding**, and 4) **Verhaar classification** profilers.

The **EcoSAR classification** profiler is the most detailed profiler, which is applicable to acute aquatic endpoints. It is based on 40 years of experience in the Office of Pollution Prevention and Toxics of the United States Environmental Protection Agency. The current **EcoSAR classification** profiler has a large number of different chemical categories most of these categories are based on structural alerts, which have been linked to toxicity in excess of baseline. Many are related to specific modes of toxic action or specific

[1] Note: the preferred response variable for tests on exponentially growing organisms is inhibition of the exponential growth rate (= log final cell density, i.e. ErC-values) because this response variable is not, like the response variable final cell density (giving EbC –values), dependent of the duration of test and the absolute growth rate of the controls and therefore less prone to dependency of the slope of the response curve (cf. also OECD TG 201). It is recommended to use data reflecting the response as inhibition of the exponential growth rate or at least not to mix data representing the two different response variables. Use of data relating to reduction of final cell density may, however, be similar to those representing inhibition of the exponential growth rate constant in very standardized tests where exponential growth prevailed also in the controls during the whole testing period and where the absolute growth of the controls does not differ substantially between tests and for chemicals where the slope of the response curve is not particular steep or shallow. In these cases it may be considered to pool the two response variables for the endpoint in order to expand the size of the dataset.

molecular initiating events and thus have a sound mechanistic basis. Others, however, lack this mechanistic grounding. The major advantage to the **EcoSAR classification** profiler is its large number of categories, which are based on experimental evidence. The major disadvantage of the **EcoSAR classification** profiler is that it often lacks mechanistic transparency for the basis of the category. The neutral organic or basesurface narcotics are often arrived at because the target chemical does not fit in any other categories so there may be no experimental evidence for this category. Another disadvantage is that chemicals may be assigned to more than one of the EcoSAR categories. In such cases it is recommended to use the category, which provides the lowest predicted value.

The **OASIS acute toxicity mode of action** profiler was developed by the Laboratory of Mathematical Chemistry, Bourgas "Prof. As. Zlatarov" University, Bourgas, Bulgaria. It is based on a broader set of structural alerts gathered primarily from the fathead minnow toxicity testing and defined by Russom et al. (1997). It includes further rules elaborations based on simple metabolism. The major advantage to the **OASIS acute toxicity mode of action** profiler is it assigns categories based on modes of toxic action; thus there is usually a clear mechanistic foundation to the category, which improves transparency and aids acceptability. The major disadvantage of the **OASIS acute toxicity mode of action** profiler is that since it is based on fish toxicity data its applicability to other organisms may not be warranted and at least relevance of its use for other aquatic organisms should always critically be considered .

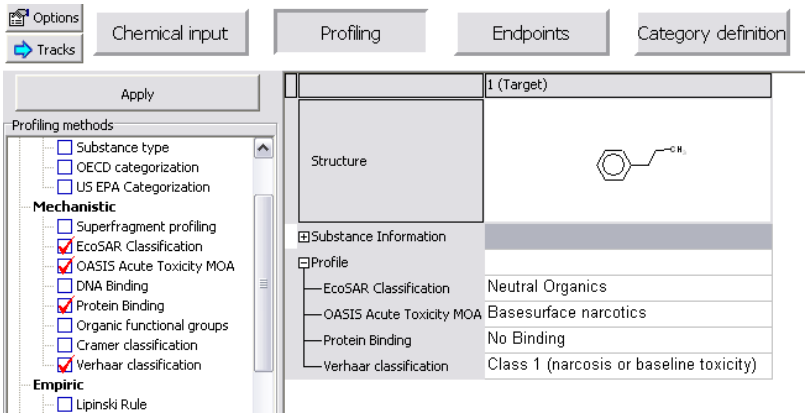
The **Protein binding** profiler includes many categories. Each category represents a chemical mechanism by which organic compounds can covalently react with protein moieties in particular thiol (SH) and amino (NH_2) groups. The associated mechanisms are in accordance with the existing knowledge on electro(nucleo)philic reactions of various electrophilic functionalities (Dimitrov et al., 2005). Briefly, covalent bonds between a substrate and a target molecule are formed by reactions between electron-rich nucleophiles and electron-poor electrophiles. Electron-rich groups usually contain heteroatoms (ones other than carbon or hydrogen), especially in nucleic acids and proteins. Although the other profilers mentioned here deal with a broad field of different mechanisms and biochemical pathways leading to different chemical categories, the **Protein binding** profiler comprises the covalent binding reactions between a small electrophile and thiol and amino-based endogenous nucleophile. The preference of a compound toward a specific molecular site of action can be explained by a classification of electrophiles and nucleophiles according to their polarisability, in other words, the chemical “hardness” and “softness” of the electrophilic or nucleophilic centre. Dissimilar hardness leads to a higher potential energy barrier of reactions between electrophiles and nucleophiles. Therefore, molecular interactions between electrophiles and nucleophiles are governed by properties, which follow hardness and softness patterns in general; hard Lewis acids bind strongly to hard Lewis bases and soft Lewis acids bind strongly to soft Lewis bases. In this context, Lewis acids are electron acceptors, which act as electrophiles, while Lewis bases are electron donors, acting as nucleophiles. Generally speaking soft electrophiles will react preferentially with thiol groups e.g. cysteine amino acids in proteins, while harder electrophiles will prefer to react with the amino groups of e.g. lysine amino acids in proteins. As can be inferred from this statement, firstly establishing whether a compound is electrophilic in nature and secondly the type of electrophile (and associate this with a reaction mechanism) can be of great benefit to predicting acute aquatic toxicity because electrophilic reactivity is often the molecular initiating event and potency-determining property for acute toxicity. The major advantage to the **Protein binding** profiler is it assigns categories based on well documented and well-understood chemical reactions; thus there is a clear mechanistic foundation to the category, which improves transparency and acceptability. The major disadvantage of the **Protein binding** profiler is that the vast majority of industrial organic compounds do not covalently bind to proteins so this profiler is of little value in placing these chemicals in a toxicologically meaningful category.

The **Verhaar classification** (Verhaar et al., 1992) was developed utilizing acute toxicity data collection for guppies and fathead minnows. This scheme based on structural alerts delineated chemicals into one of five classes. The Verhaar classes include 1) Class 1 or “inert” chemicals, which are nonpolar

narcosis or baseline toxicity, 2) Class 2 or “less inert” chemicals, which are the polar narcotics, 3) Class 3 or “reactivity” chemicals, which are typically non-selectively, covalently reactive with protein moieties, 4) Class 4 or “specifically-acting” chemicals, which specific reactivity with receptors, or 5) Class 5 or “unclassified” chemicals. Based on molecular structure there are rules for discrimination of the first three classes. An advantage of the **Verhaar classification** profiler that is it typically assigns the Class 1 status based on meeting structural alerts. The major disadvantage of the **Verhaar classification** profiler is that a large number of industrial organic compounds get relegated to Class 5 and furthermore that this classification approach only operates with five different classes where especially class 4 and 5 cover widely different modes of actions and toxicity ranges.

It is recommended as shown in Figure 1 to select all four of the above noted profilers in the **Profiling** phase of the work flow.

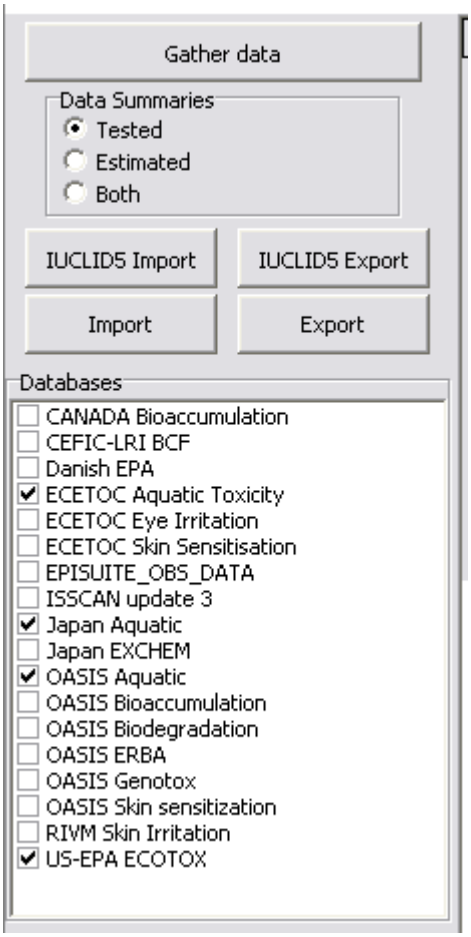
Figure 1: The Initial Battery of Profilers for Categorizing Acute Aquatic Endpoints.



Databases for Aquatic Toxicity Endpoints

The Toolbox is well populated with experimental data for acute aquatic effects. This reflects two facts, first that there is a long history of acute aquatic toxicity testing in fish, crustaceans, algae, and protozoa. And two, several groups have collated these data into searchable databases. Within the Toolbox, databases, which include acute aquatic effects potency data, include **ECETOC Aquatic Toxicity**, **Japan Aquatic**, **OASIS Aquatic**, and **US-EPA ECOTOX**. In order to collect the largest number of compounds possible within a chemical category, it is recommended as shown in Figure 2 to always use all four of the above noted databases.

Figure 2. The Toolbox Databases Relevant to Acute Aquatic Toxicity Endpoints.



Profiling Results: What They Tell About a Grouping Strategy

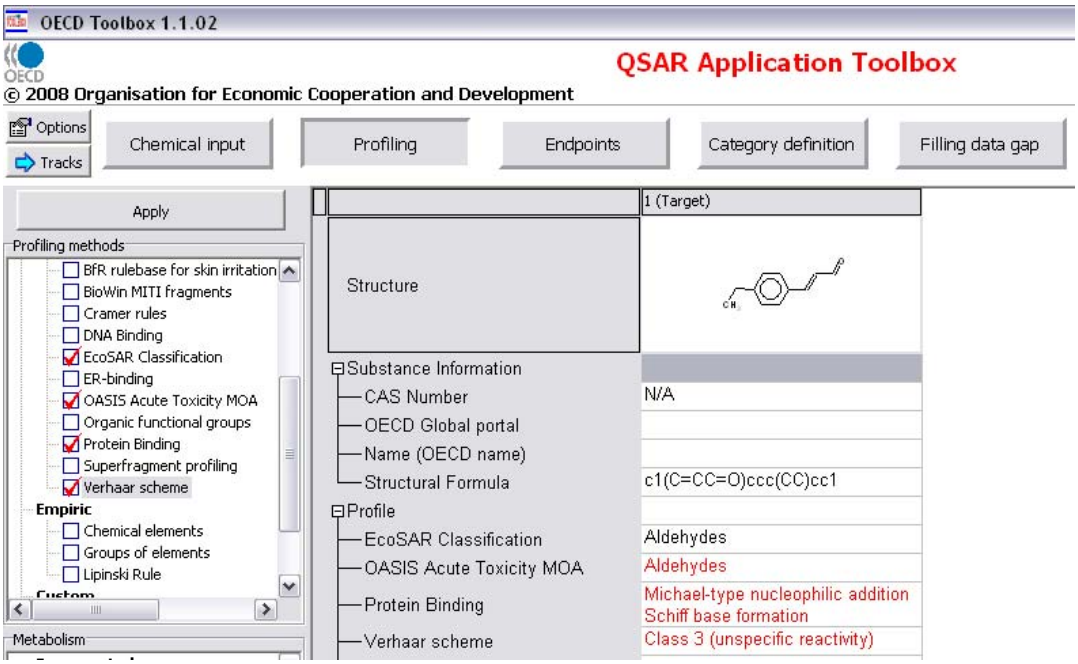
As noted earlier defining the chemical category is the critical step in the workflow of the Toolbox. The results from using the initial battery of Toolbox profilers relevant to acute aquatic toxicity endpoints (**EcoSAR classification**, **OASIS acute toxicity mode of action, Protein binding**, and **Verhaar classification** profilers) has the potential of aid one in determining if the grouping strategy and resulting chemical category is a good one. One can ask do we want to have a chemical category where all members have the same profiling result in each of the profiles? Experience has revealed the one word answer is yes. However, reality has shown this is not always a simple task. In the extreme each chemical is a category of

itself, but this extreme in not useful. Similarly using very general category definitions such as traditional chemical classes (e.g., alcohols, aldehydes or amines etc.) is also not useful.

Based on current knowledge in the Toolbox experience has shown that the consistency in the profiler groupings is an indicator of good category selection. For example, re-examining Figure 1 reveals that in this case there is a consistency among the profiling results so one is confident based on the current knowledge in the Toolbox that the target chemical is indeed a baseline, nonpolar narcotic compound because there is consistence in the **EcoSAR classification**, **OASIS acute toxicity mode of action,** and **Verhaar classification** and that **Protein binding** is **"No Binding"**. This consistency can be taken to reflect the grouping strategy and resulting chemical category is good.

Similar but different results are observed after initial profiling of an alpha-beta-unsaturated aldehyde (Figure 3).

Figure 3. Profiling of an Alpha-Beta-Unsaturated Aldehyde.

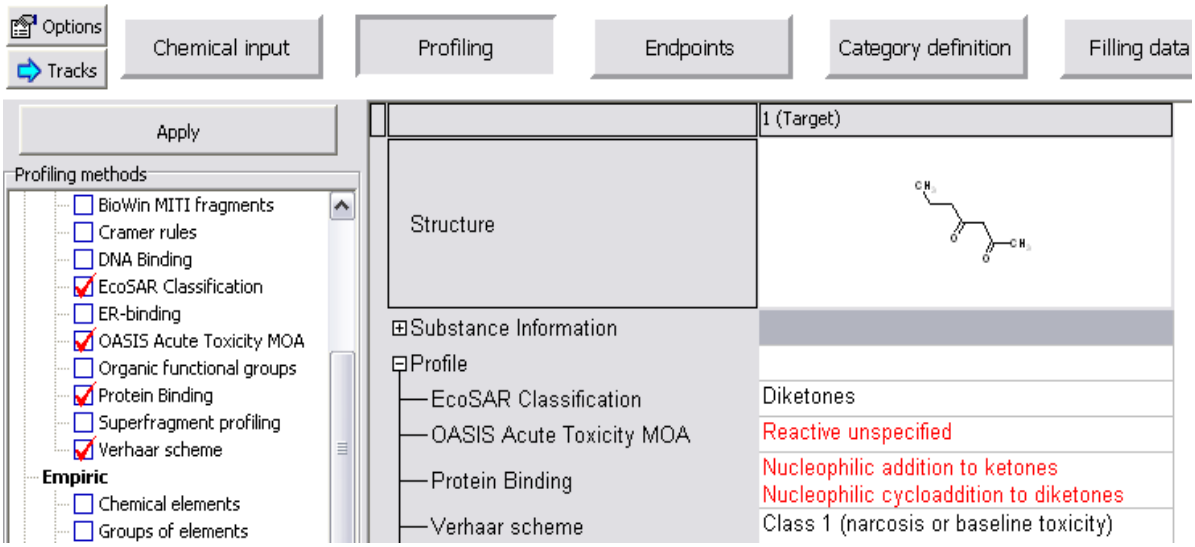


In this case the **Protein binding** profiling is reported as **"Michael-type nucleophilic addition"** and **"Schiff base formation"**. Both of these mechanisms of protein binding are associated with **"aldehydes"** and both mechanisms are types of **"Class 3 (unspecific reactivity)"**. Taken together, the consistence between the results for the initial profilers shows the grouping strategy is a good one. However, as noted above because there are two categories listed under **Protein binding** forming the final chemical category may require further subcategorization. The importance of further subcategorization is related to the facts that **Michael-type nucleophilic addition** is a more potent toxic reaction than is **Schiff base formation**. Moreover, **Schiff-bases formation** has a stronger dependency on Log Kow than does **Michael-type nucleophilic addition**. These facts are demonstrated by comparing Toolbox trend analyses for acrylates, which act as Michael-type acceptors and saturated aliphatic aldehydes, which act as Schiff base formers.

In contrast are the results from using the initial battery of Toolbox profilers relevant to acute aquatic toxicity endpoints for the target chemical in Figure 4. In this case there are inconsistencies in the profiling

results. While the **EcoSAR classification**, giving the category **"Diketone"** the **OASIS acute toxicity mode of action** profiler gives **"Reactive unspecified"**. The **Protein binding** again gives two mechanism **"Nucleophilic addition to ketones"** and **"Nucleophilic cycloaddition to diketones"**. While these results may be resolved by following subcategorization, the results of the **Verhaar classification** profiler, **"Class 1 (narcosis or baseline toxicity)"** is inconsistent with the other primary profilers. Taken together, these results suggest that a more detailed look at the grouping strategy is required before settling on a category definition and moving on to filling data gaps.

Figure 4. Profiling of an Aliphatic Alpha-Gamma-Diketone.



The latter example raises the question whether some profiling results are more important than other? Experience has revealed that the answer is yes. The top tier profilers for acute aquatic toxicity are **EcoSAR classification** and **Protein binding** the second tier profiler is the **OASIS acute toxicity mode of action** and the lowest tier is the **Verhaar classification** profiler. For most part, these tiers reflect the age and complexity of the profilers. Older and simpler profilers are of less value in determining group strategies and settling on chemical categories, especially for more structurally complex compounds.

Secondary Profilers Relevant to Acute Aquatic Toxicity Endpoints

Toolbox profilers, which can be used as secondary profilers in a subcategorization scheme relevant to acute aquatic toxicity endpoints includes the 1) **Organic functional groups**, 2) **Metabolism**, and 3) **Superfragment** profilers. These are not used in parallel, but rather one at a time. Since initial profiling is designed to selects broad chemical categories, it is often necessary to subcategorize to more narrowly define the chemical category. However, since the secondary profilers are based on imperfect structural similarity it is important to review the list of structures provide with each subcategorization routine to assure one is not eliminating analogues for unknown reasons.

The **Organic functional groups** profiler simply identifies the classic organic functional groups present in the molecule. It is recommended to be used iteratively for conducting secondary groupings to identify more structurally limited compounds with which to conduct read-across. Since there is no generally preferred way of identifying structural similarity, it is recommended to always use **Organic functional groups** profiler as a first option.

In selected cases of acute aquatic toxicity, biotransformation of a non-protein-binding parent compound can lead to a protein-binding metabolite thus, metabolism profilers are often included either as a primary or secondary profiler for acute aquatic effects (see section 3.7 of "Guidance Document for Using the OECD (Q)SAR Application Toolbox to Develop Chemical Categories According to the OECD Guidance on Grouping of Chemicals" [www.oecd.org/env/existingchemicals/qsar]. For aquatic toxicity endpoints the **Liver metabolism simulator** is recommended. When the **Liver metabolism simulator** is used the Toolbox inserts the results into a data matrix. The **Liver metabolism simulator** identifies reactive metabolites that are captured by primary profilers especially the **EcoSAR classification**, **OASIS acute toxicity mode of action**, and **Protein Binding** profilers; thus, the **Liver metabolism simulator** is typically used in conjunction with these other profilers. Since the use of the **Liver metabolism simulator** dramatically increases the computing time it is recommended to be used as a secondary profiler.

The **Superfragment** profiler is based on Clog P calculations of coefficient of partitioning of chemicals in the n-octanol/water system by a group contribution method that includes structural correction factors and interaction factors. A superfragment consists of a combination of simple polar fragments that are in such close proximity that their solvation behavior as evidenced by the Clog P calculation is markedly affected. Presently, the **Superfragment** profiler is restricting to extended fragments, which are separated by one or two isolating carbons, designating as Y-C-Y and Y-C-C-Y, - respectively. Superfragments are not correlated with covalent reactivity they appear to be useful as a secondary categorization tool related to chemical bioavailability and/or non-covalent protein interactions. The **Superfragement** profiler works well in identifying chemical categories where category members all have specific combinations of multiple polar groups. (e.g. diketones).

While not encouraged, if the **Structural Similarity** profiler is used it should be done with extreme caution as results have been shown to be inconsistent when using this profiler.

Subcategorization with Initial Profilers Relevant to Acute Aquatic Toxicity Endpoints

Sometimes it is necessary to more narrowly define the chemical category by subcategorizing using one or more of the initial profilers relevant to acute aquatic toxicity endpoints. This is more often the case with the **EcoSAR classification** and **Protein binding** profilers. The aliphatic-CH=O is a structural alert for Schiff base forming protein binders, while the aliphatic-C=CCH=O is a structural alert for Michael addition protein binders. However, since the [–CH=O] fragment is common to both reactions the initial profiling with the **Protein binding** profiler will capture chemical with substructure alerts for both of these protein binding reactions (Figure 3). Therefore, subcategorization with the **Protein binding** profiler will be necessary to separate the two chemical spaces and narrow the structural boundaries of the applicability domain to either Schiff base formers or Michael acceptors. When the results from a profiling appear in red (e.g., figure 3) it is an indication that some sort of subcategorization will be necessary to ensure a good chemical category.

Endpoint Comparisons that are Relevant to Acute Aquatic Toxicity Endpoints

Experience has shown that there may be a too small amount of experimental data in some categories to derive robust estimates for certain endpoints. In these cases it may be considered to use inter-endpoint comparisons, but only if the same grouping strategy leads to similar models for the same category. For example one can use results from *Tetrahymena* acute toxicity data to support predictions for algae, daphnids, and fish. For example, in filling the data gap for the aliphatic alpha,beta-unsaturated aldehyde 2-octen-1-al [CCCCCC=CC=O] we find there are seven *Pimephales promelas* 96hr LC_{50} data and 23 *Tetrahymena pyriformis* 48hr IGC_{50} data. Trend analyses of both these data sets reveal a similar log Kow-

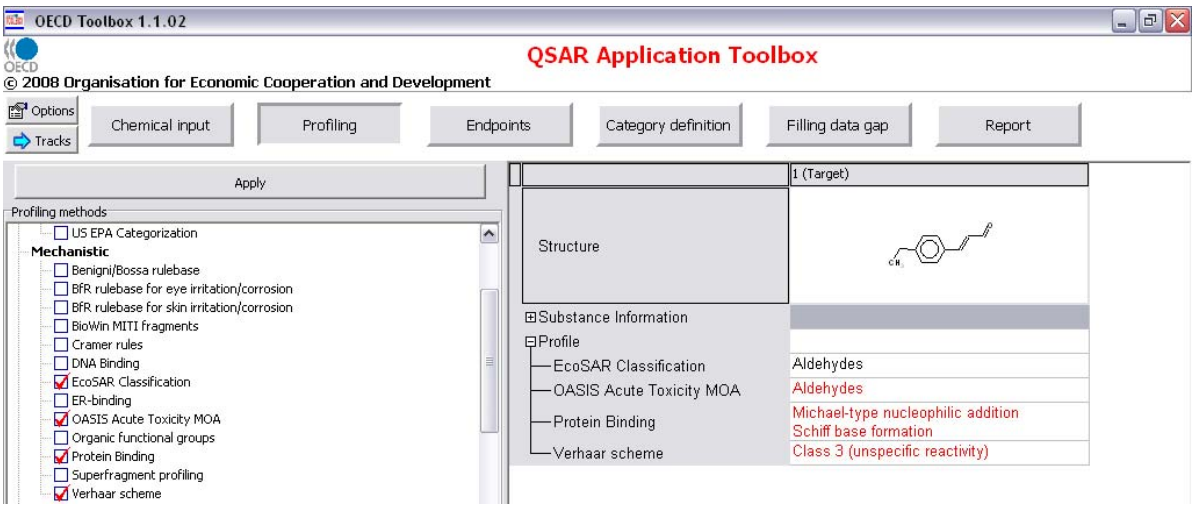
dependent relationship (i.e., the slope for the fish data regression is 0.391, while the slope of the protozoan data regression is 0.354). Thus, the 23-point protozoan model supports the data gap filling based on the 7-point fish model. Other types of endpoint comparisons or rather pooling of response variables concerning a toxicity endpoint are e.g. pooling of data from different species in the same higher taxonomic group (e.g. fish or crustaceans) or pooling data from different test periods for the same species (e.g. 24 and 48 h EC_{50} for *Daphnia magna)*. Caution is advised in using such pooled data. It should only be done when one has a good working knowledge of the acute aquatic toxicity endpoints and a rational for making such pooling. It is suggested when such data pooling is done to record the explicit justification for doing so.

Data pooling can also be used for *Daphnia magna* and other crustaceans for the acute endpoint 48 h LC_{50} and EC_{50}. These two endpoints are a measure of the same biological effect (50 % effect concentration for immobilization) and can therefore be pooled to increase the size of the dataset. Care should be taken, however, to make sure that the same values are not represented two times in the pooled dataset.

Example: Data Gap Filling for 4-Ethylcinnamicaldehyde

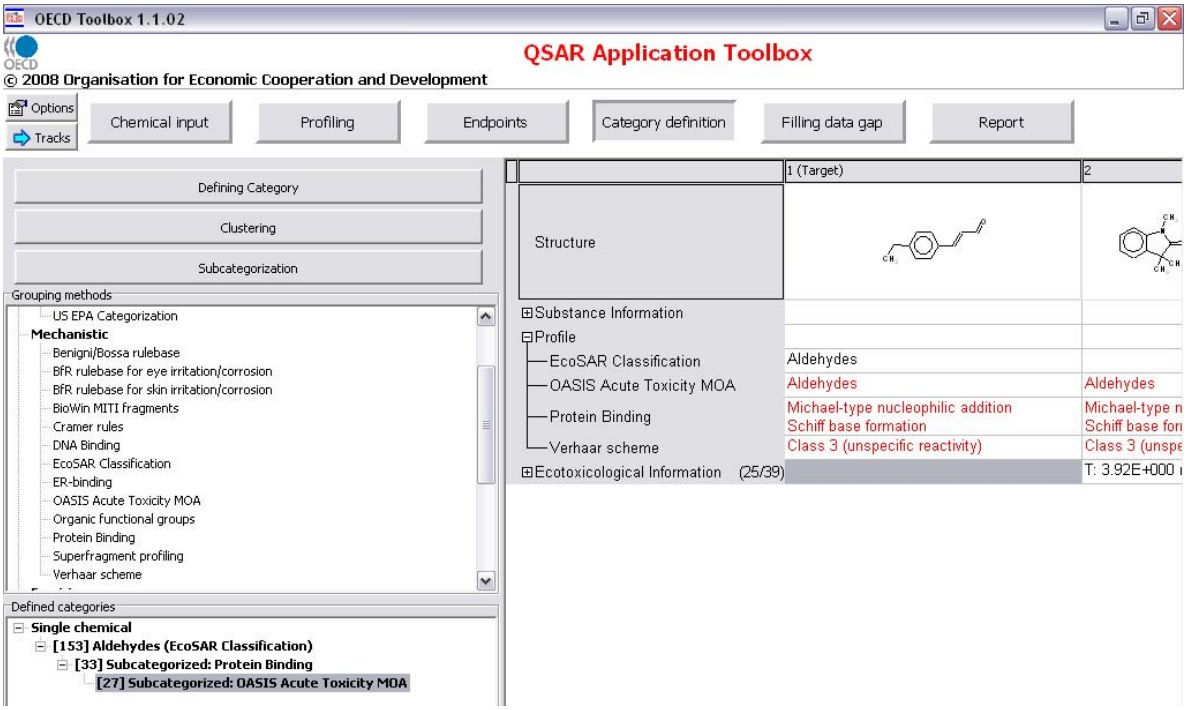
The profiling for 4-Ethylcinnamicaldehyde [c1cc(CC)ccc1C=C=O] is shown in Figure 5.

Figure 5. Profiles for Ethylcinnamicaldehyde.



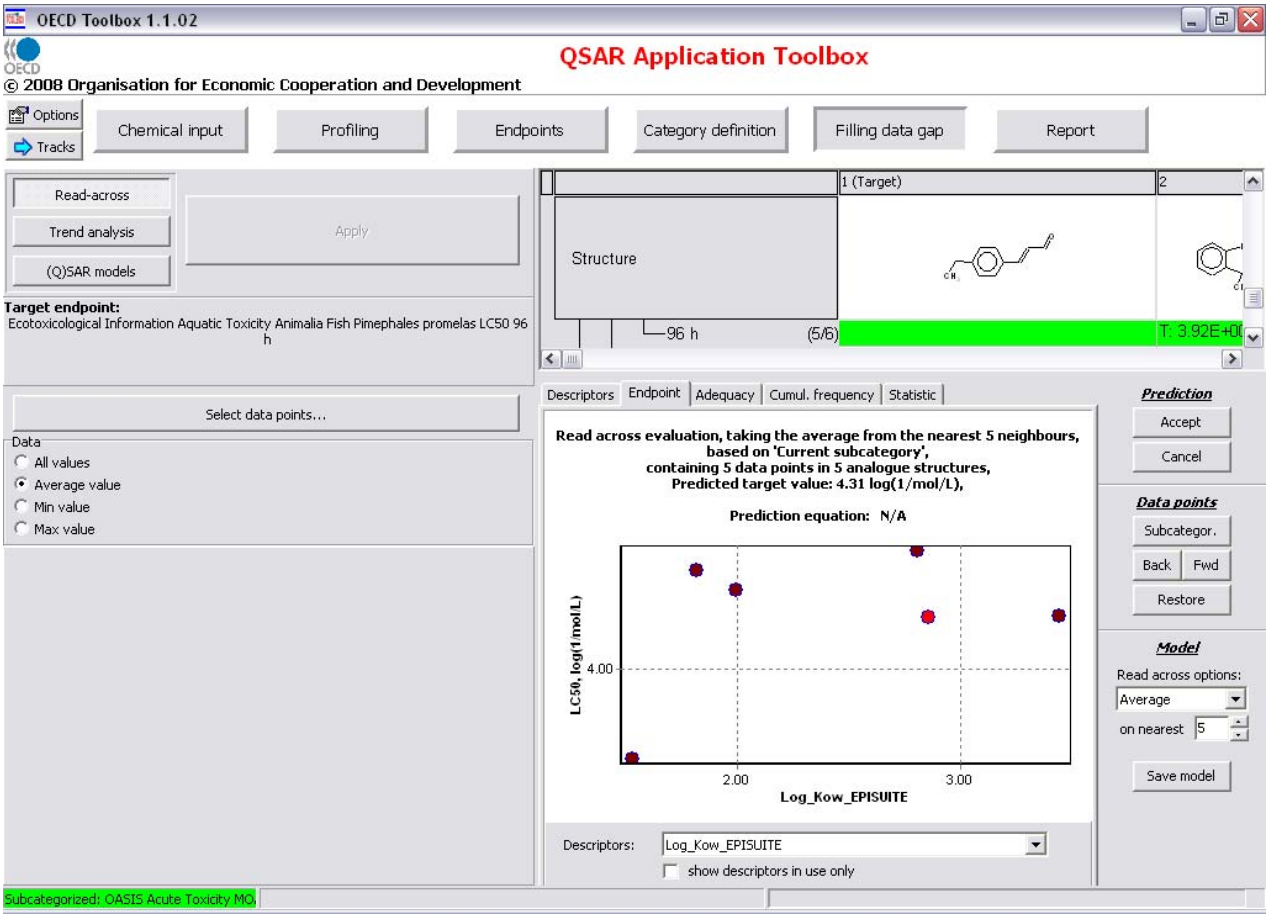
If one does a step-wise categorization / subcategorization using the four initial profilers in order of first tier **EcoSAR classification** and **Protein binding,** second tier **OASIS acute toxicity mode of action** and third tier **Verhaar** classification the result is a subcategory of 25 compounds with ecotoxicity data in the Toolbox (Figure 6).

Figure 6. Step-Wise Subcategorization for 4-Ethylcinnamicaldehyde.



Data gap filling for the target compound and the *Pimephales promelas* 96hr LC_{50} endpoint by read across is shown in the screen shot in Figure 7. It is based on only those chemicals which are **"Aldehydes"** that react by both **"Michael-type nucleophilic addition"** and **"Schiff base formation"**.

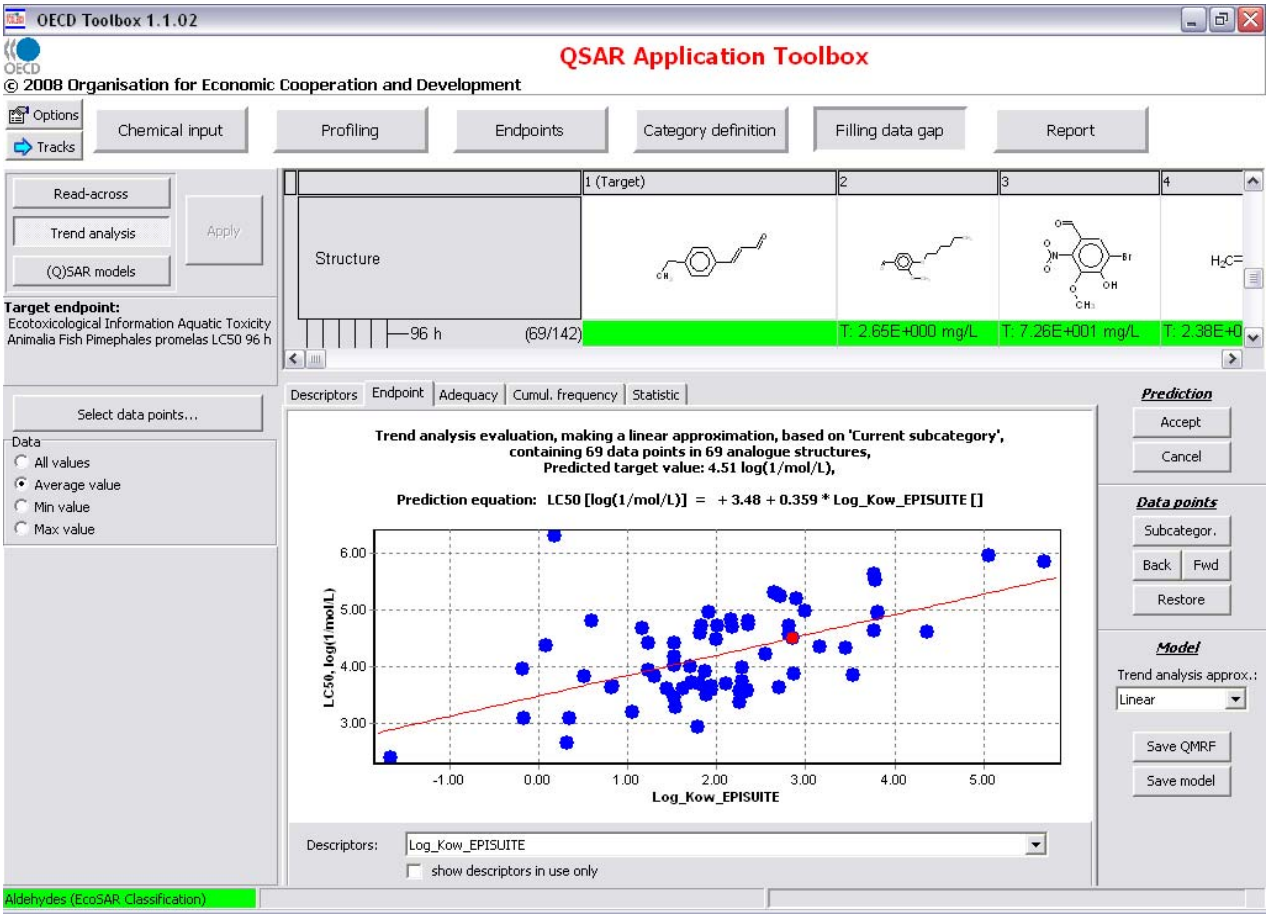
Figure 7. Filling Data Gap for 4-Ethylcinnamicaldehyde by Read-Across.



The conclusions from this example are the experimental data is limited and there is no trend among these data; at best a read-across from a close analogue(s) can be performed. Since the OECD is not responsible for the quality of the endpoint data in the Toolbox, it is recommended that the data used in any data gap filling exercise be checked. There are, however, other ways of using the profilers in the Toolbox to form a category for filling the data gap for 4-Ethylcinnamicaldehyde.

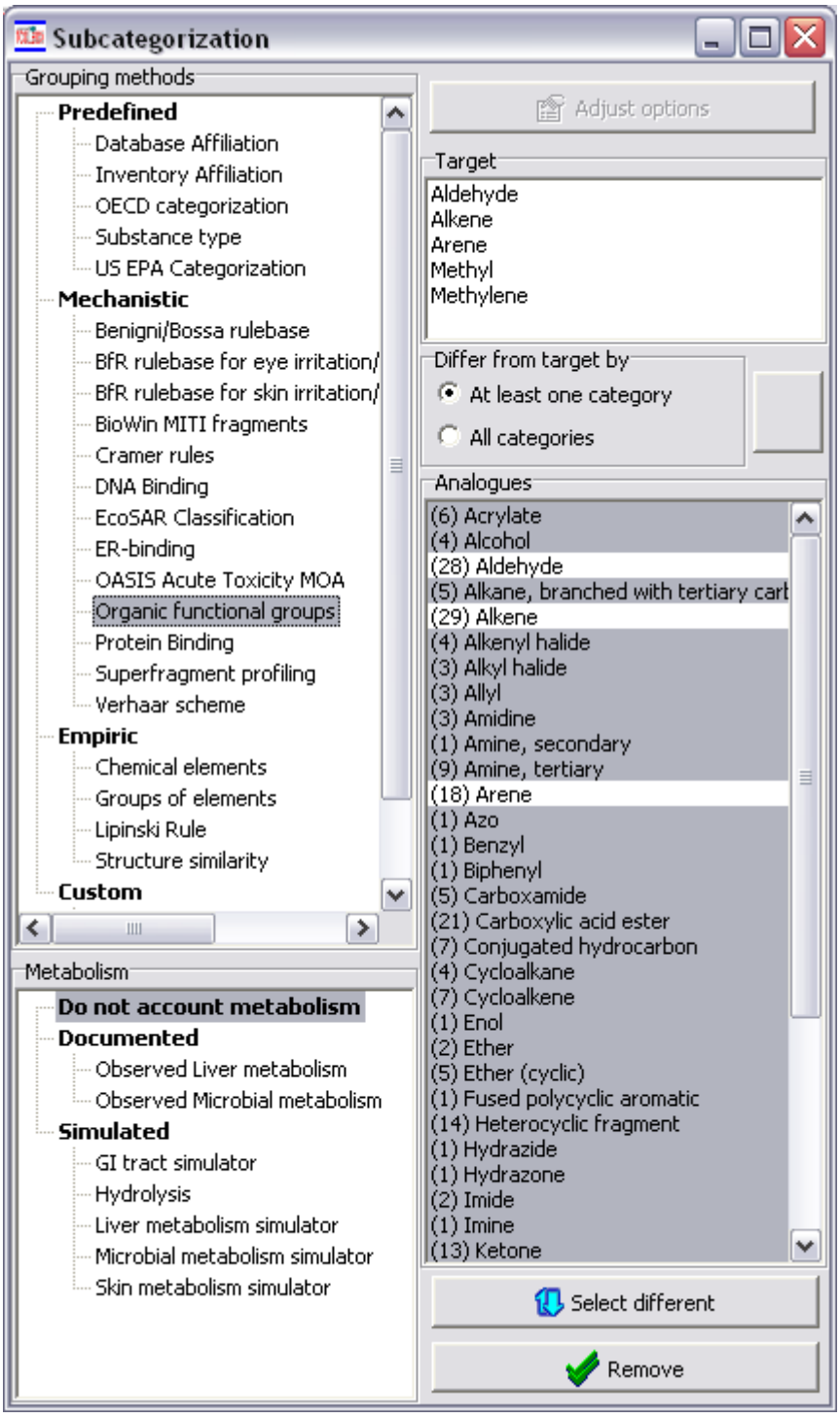
Using the **EcoSAR Classification** profiler gives the category **"Aldehyde".** This option leads to a trend analysis based on 96 experimental data for the *Pimephales promelas* 96hr LC_{50} endpoint (Figure 8). However, since this category includes aliphatic and aromatic aldehydes as well as saturated and unsaturated aldehydes, it is not sufficiently defined to accurately fill the data gap for this endpoint.

Figure 8. Trend Analysis for 4-Ethylcinnamicaldehyde Based on the Aldehyde Category.



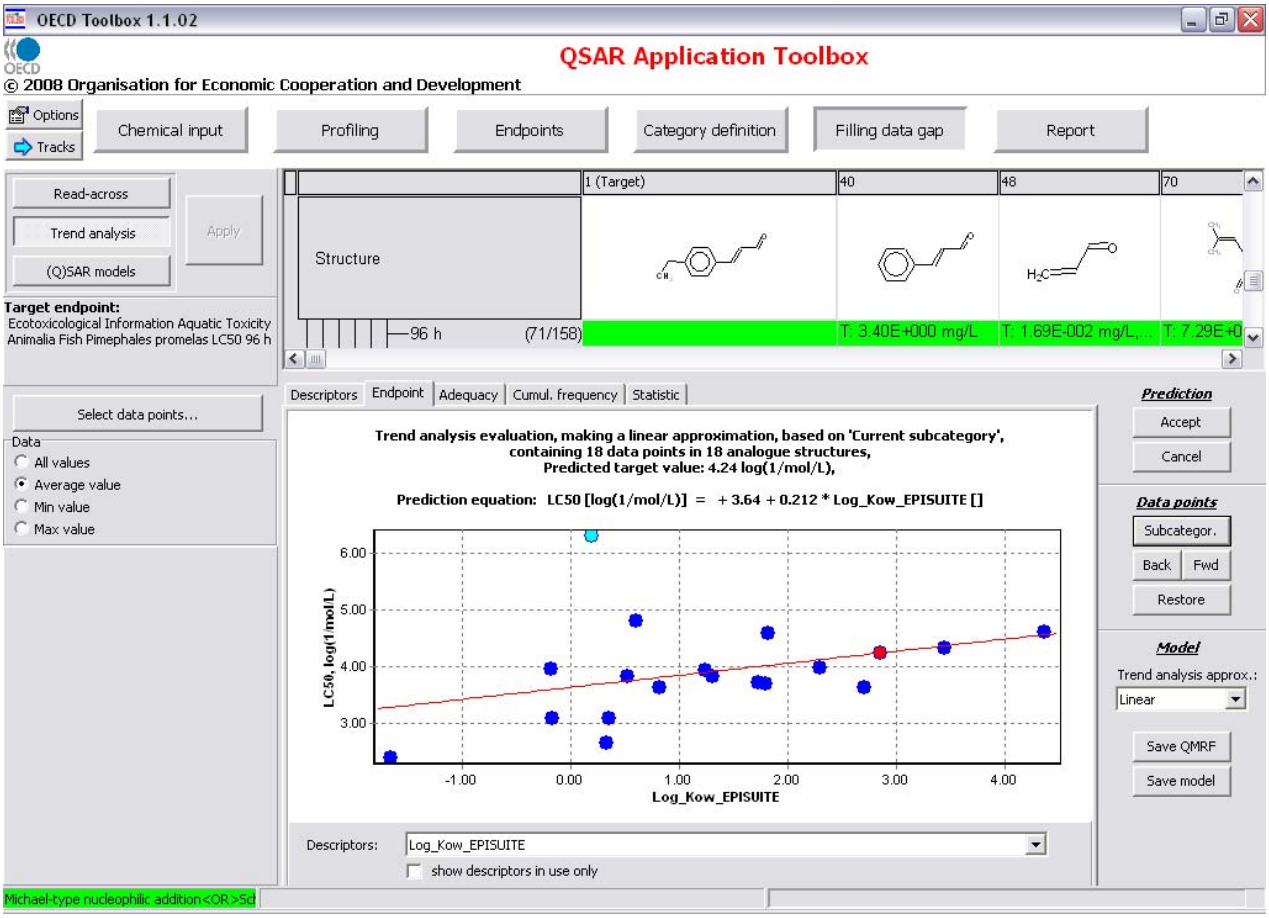
One way of achieving a better category is subcategorization the **"Aldehydes"** in figure 8 using the **Organic functional groups** profiler (Figure 9).

Figure 9. Subcategorization Using the Organic Functional Group Profiler.



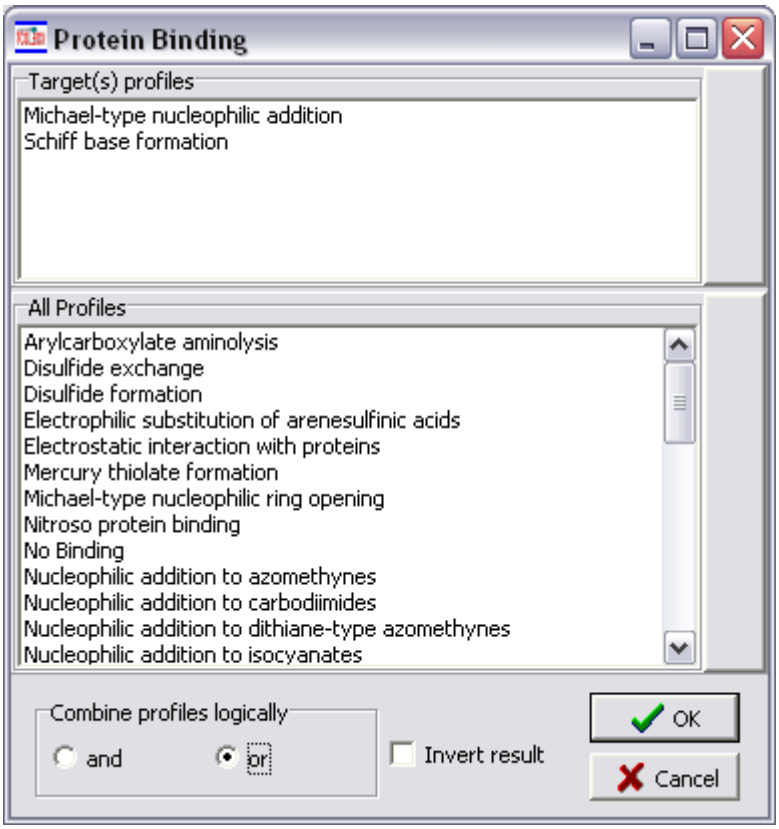
This subcategorization results in a trend analysis based on the 18 analogous that are alpha,beta-unsaturated aldehydes (Figure 10). The single outlier, which is more toxic and predicted by the trend analysis is acrolein [C=CC=O]. Acrolein is unique in that it is the only compound that has both a terminal carboxyl group [C=O] and a terminal vinyl group [C=C].

Figure 10. Trend Analysis for 4-Ethylcinnamicaldehyde Based on Aldehydes Having the Same Organic Functional Groups (i.e., Alpha,Beta-Unsaturated Aldehydes)



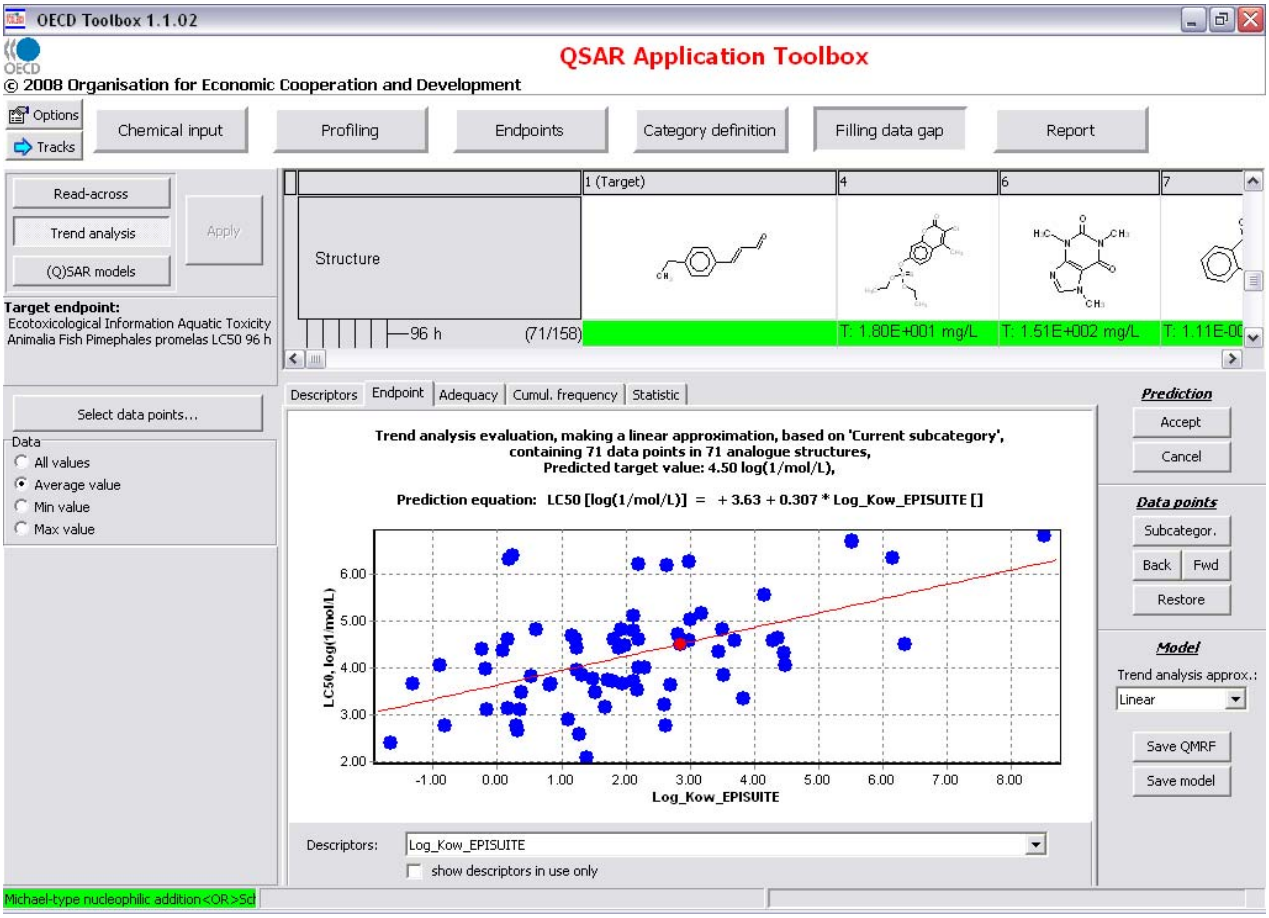
One can also profiles 4-Ethylcinnamicaldehyde starting with the **Protein binding** profiler. If one starts by selecting the **"or"** option (Figure 11).

Figure 11. Selecting the "or" Option for Protein Binding Profiling of 4-Ethylcinnamicaldehyde.



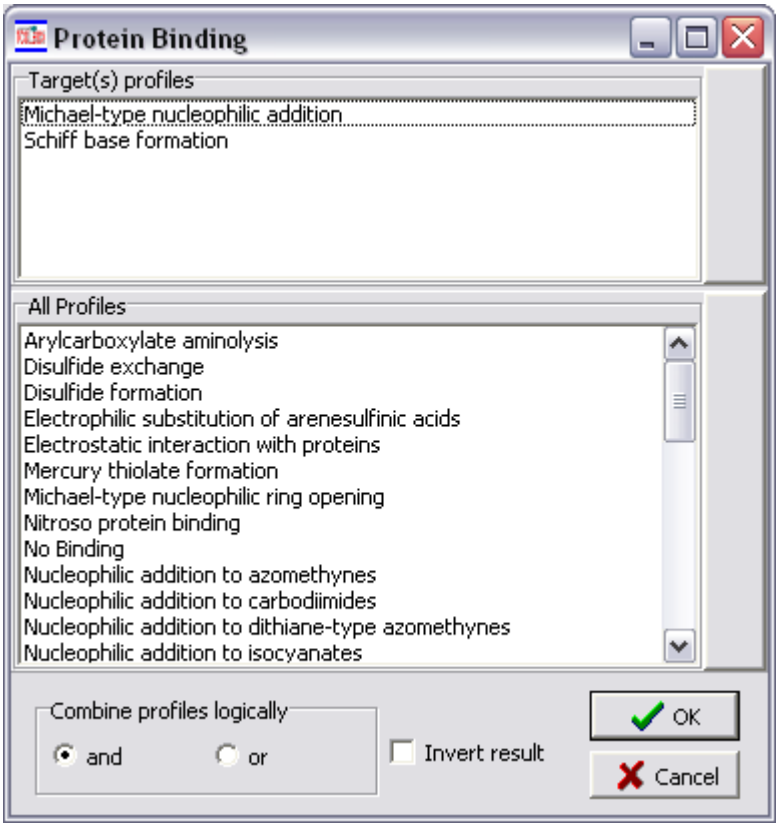
This option leads to a trend analysis based on 71 experimental data for the *Pimephales promelas* 96hr LC_{50} endpoint (Figure 12). Again, this profiling does not sufficiently defined the category to accurately fill the data gap as this profiling option combined aldehydes, which act either by **Michael-type nucleophilic addition** or **Schiff base formation**.

Figure 12. Trend Analysis for 4-Ethylcinnamicaldehyde Based on Either Michael-Type Nucleophilic Addition or Schiff Base Formation.



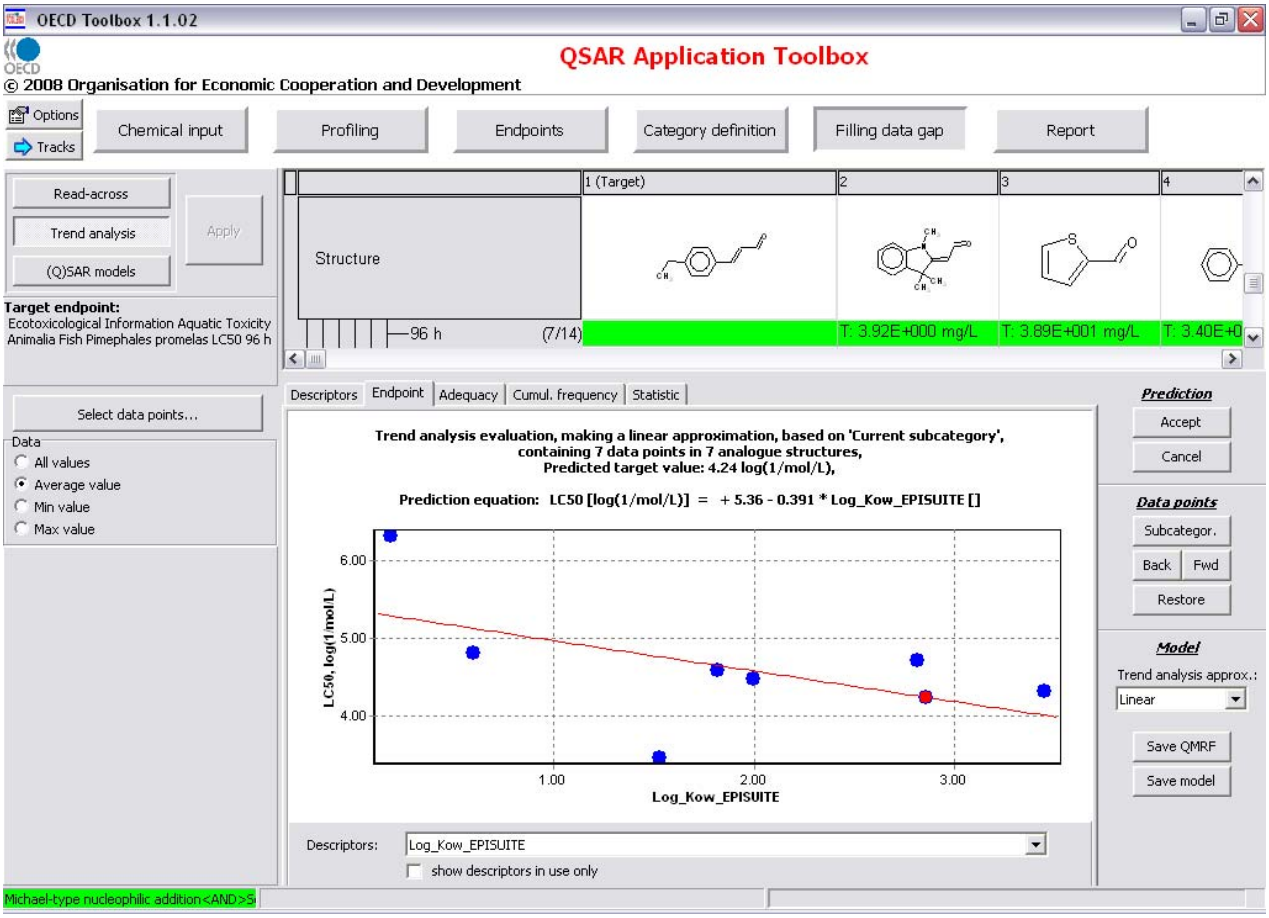
There are two options, which give rise to a better category. The first alternative is to select the **"and"** option so only chemical exhibiting structural alerts for both **Michael-type nucleophilic addition** and **Schiff base formation** are selected for data gap filling (Figure 13).

Figure 13. Selecting the "and" Option for Protein Binding Profiling of 4-Ethylcinnamicaldehyde.



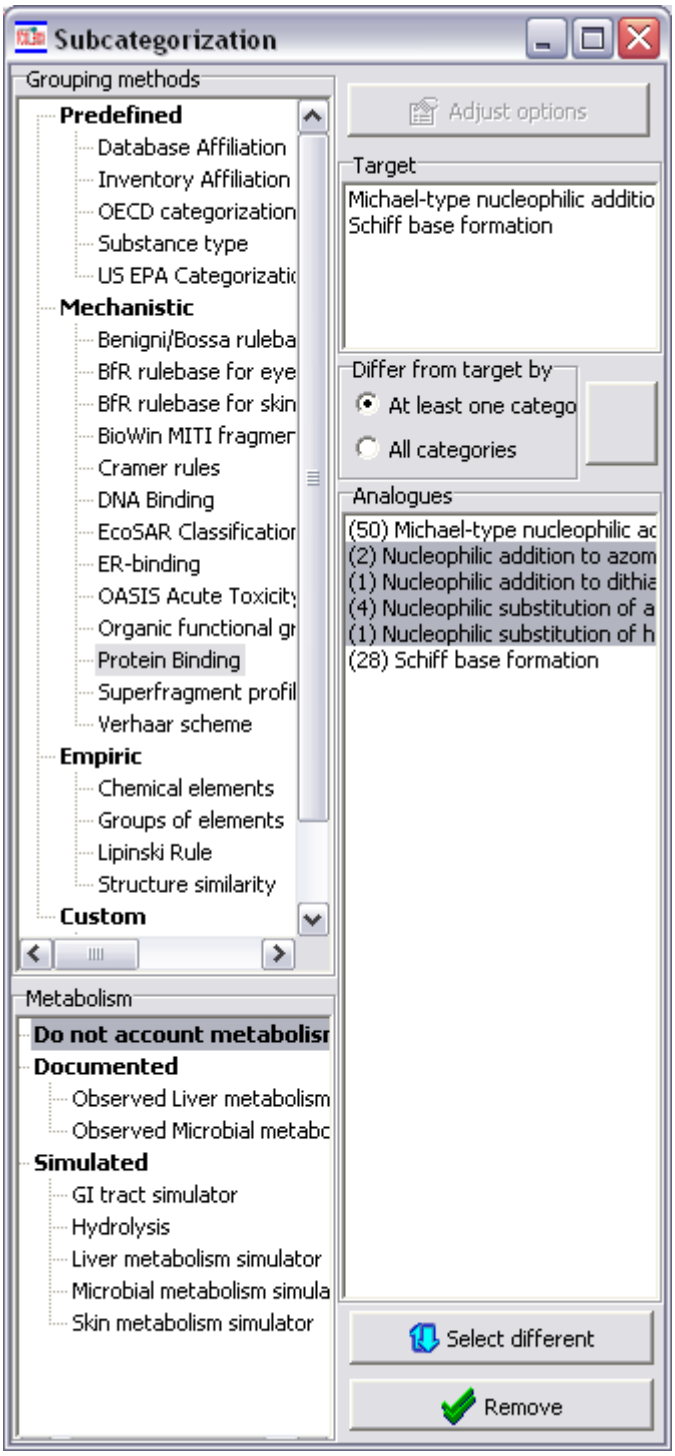
This option leads to a trend analysis based on only 7 experimental data for *Pimephales promelas* 96hr LC_{50} (Figure 14).

Figure 14. Trend Analysis for 4-Ethylcinnamicaldehyde Based on Both Michael-Type Nucleophilic Addition and Schiff Base Formation.



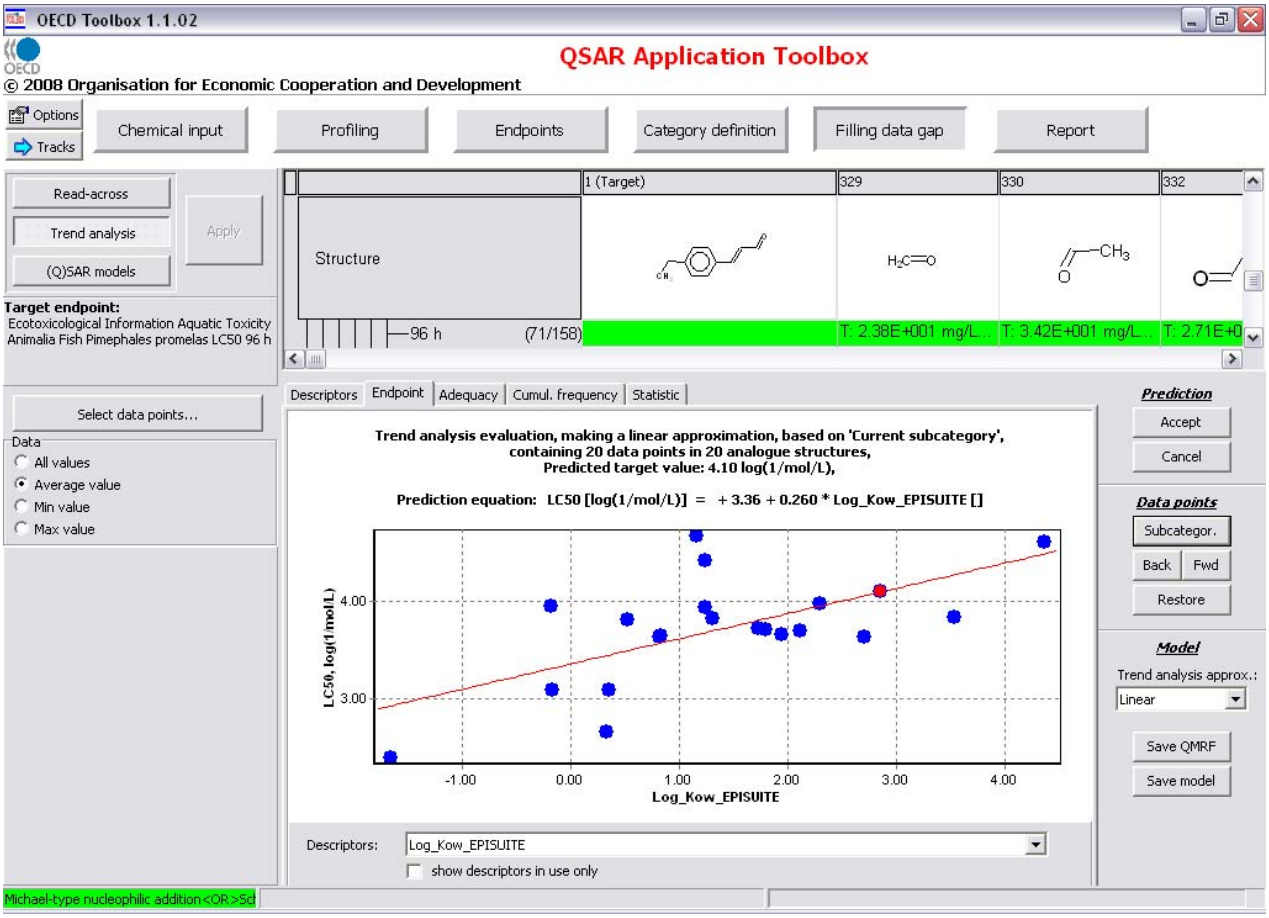
The second alternative is to subcategorize the 71 experimental data for the *Pimephales promelas* 96hr LC_{50} endpoint shown in Figure 12. This is done in a series of iteration. The first iteration is to eliminate all protein binding reactions that are neither **Michael-type nucleophilic addition** nor **Schiff base formation** (Figure 15).

Figure 15. Subcategorization and Elimination of Protein Binding Reactions that are Neither Michael-Type Nucleophilic Addition nor Schiff Base Formation.



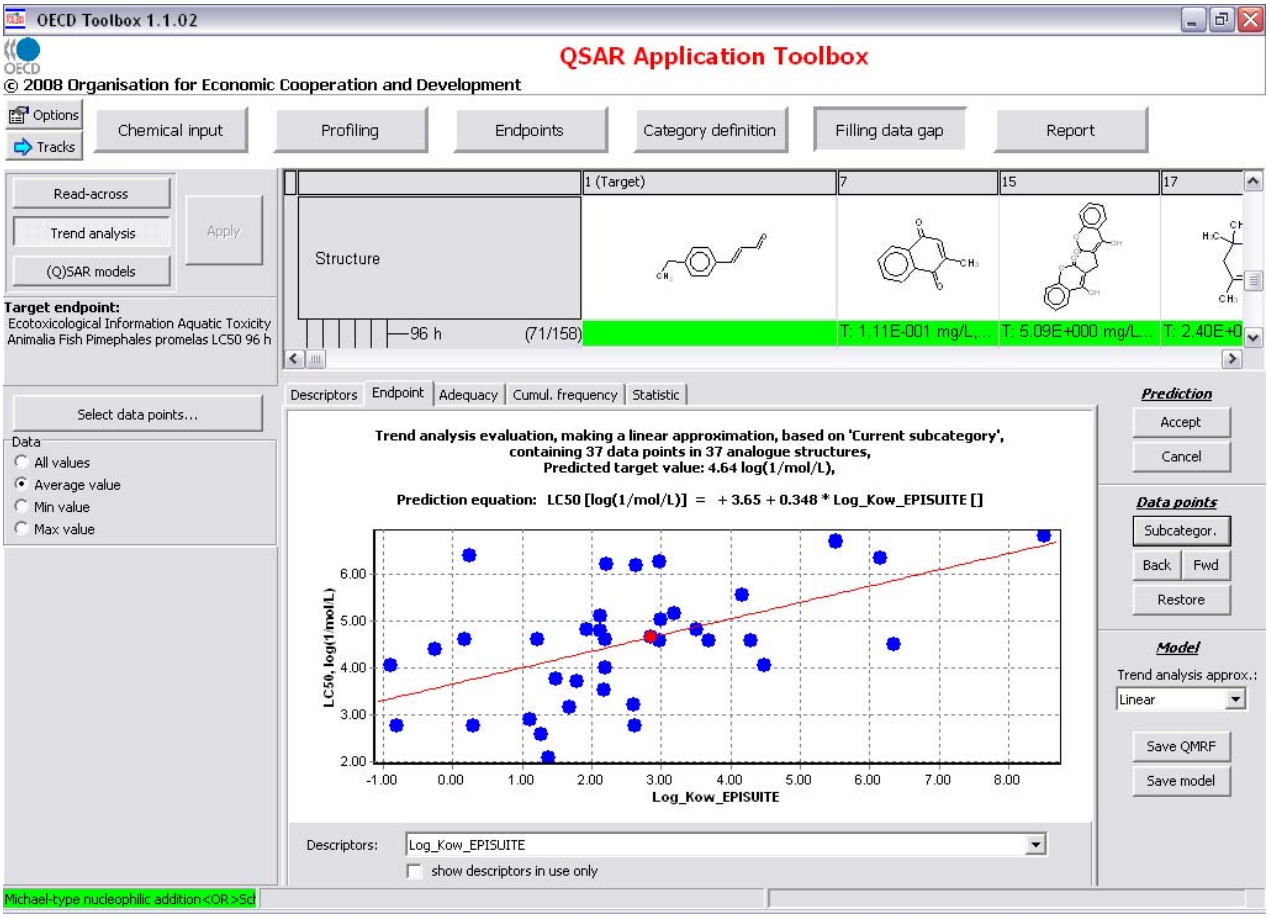
This subcategorization still leaves us with a mixture of two protein binding reactions and the question which reaction is the proper one? Conducting a trend analysis following selection for only **Schiff base formation** is shown in Figure 16.

Figure 16. Trend Analysis for 4-Ethylcinnamicaldehyde Based on Schiff Base Formation.



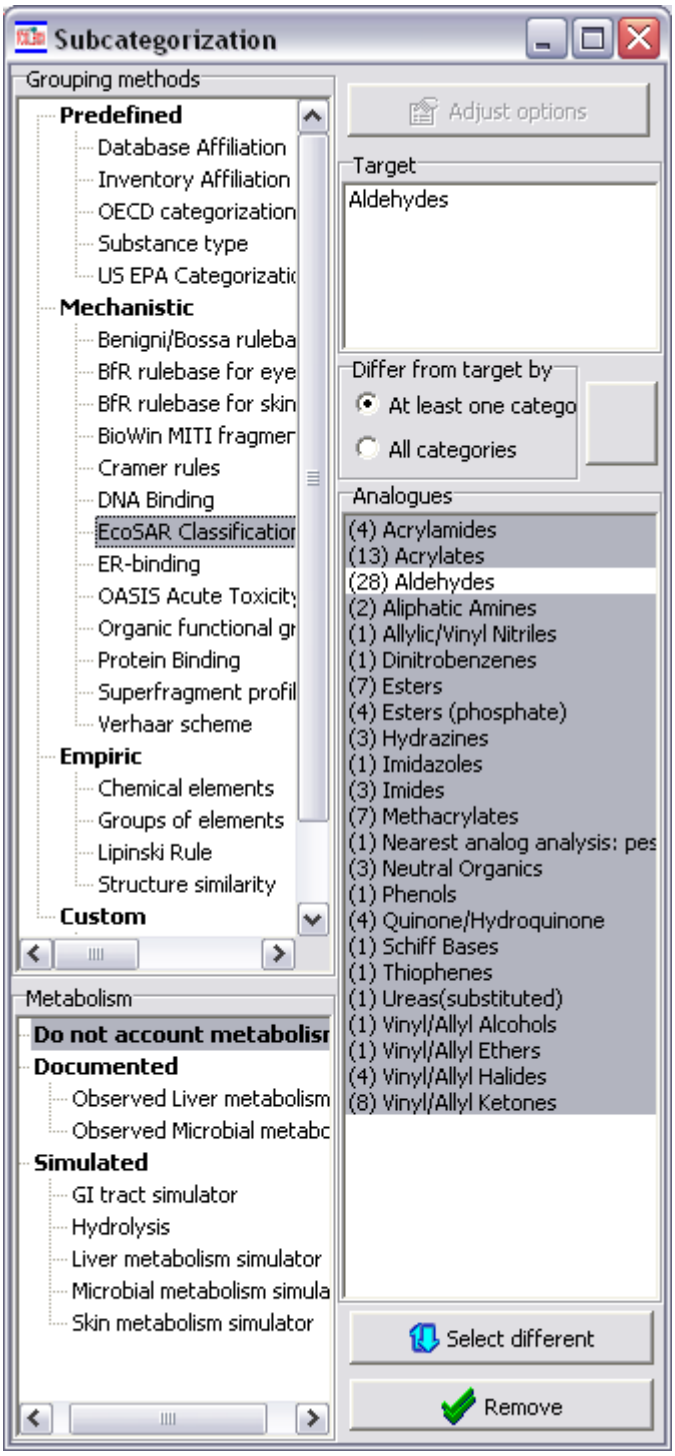
In contrast, conducting a trend analysis following the selection of only **Michael-type nucleophilic addition** is shown in Figure 17.

Figure 17. Trend Analysis for 4-Ethylcinnamicaldehyde Based on Michael-Type Nucleophilic Addition.



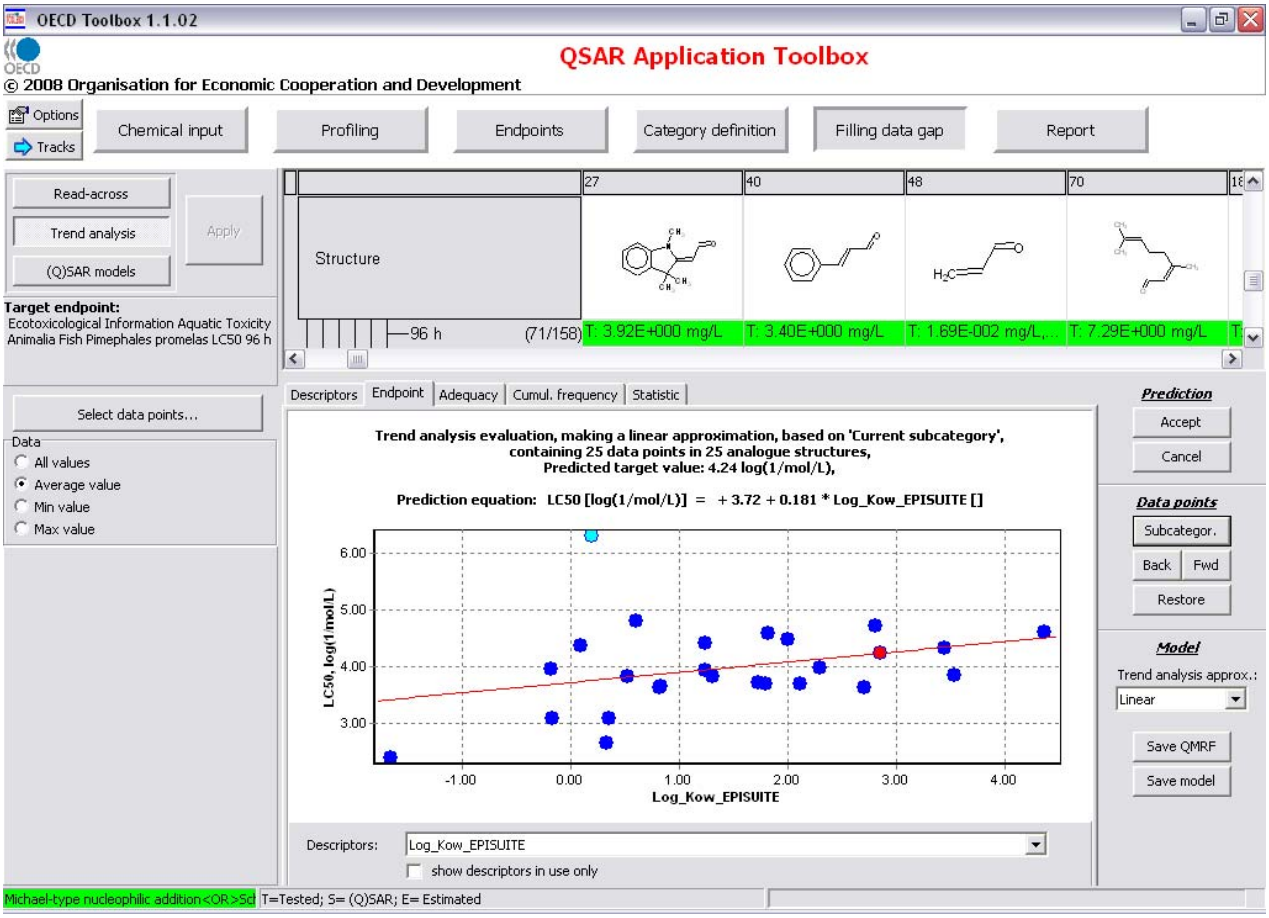
Neither the trend analysis in Figure 16 (r^2 = 0.38) nor the one in Figure 17 (r^2 = 0.30) are very promising. Therefore, a further subcategorization is undertaken. In this case one wants to eliminate all Michael-Type acceptors in figure 17, which are not aldehydes (i.e., the category should only include alpha,beta-unsaturated aldehydes). To achieve this better category one does a subcategorization with the **EcoSAR Classification** profiler and selects only **"Aldehydes"** (Figure 18).

Figure 18. Subcategorization and Elimination of all Chemicals that are not Aldehydes.



This subcategorization leads to a trend analysis based on the 25 analogous that are all alpha,beta-unsaturated aldehydes (Figure 19).

Figure 19. Trend Analysis for 4-Ethylcinnamicaldehyde Based on Alpha,Beta-Unsaturated Aldehydes Acting Via Michael-Type Nucleophilic Addition.



As with Figure 10 the outlier in Figure 19 is acrolein, the only chemical, which has both a terminal carboxyl group and a terminal vinyl group.

In summary simple, step-wise categorization / subcategorization using the 1) **EcoSAR classification**, 2) **Protein binding**, 3) **OASIS acute toxicity mode of action,** and 4) **Verhaar classification** profilers to fill the data gap for 4-Ethylcinnamicaldehyde [c1cc(CC)ccc1C=C=O] and the *Pimephales promelas* 96hr LC_{50} endpoint is not satisfactory. However, one may use **EcoSAR Classification**, **Protein binding**, and **Organic function group** profilers in different ways to achieve satisfactory results. The goal in any case is to narrow the final trend analysis to chemicals, which are alpha,beta-unsaturated aldehydes that act as Michael-type acceptors. Trend analyses-based data filling using all aldehydes or all Michael acceptors results in poorer (Q)SAR predictions.

Summary of Steps for Single Target Chemical Data Gap Filling for an Aquatic Toxicity Endpoint

1. Begin the profiling with a battery of mechanistic profilers including the **EcoSAR classification**, **OASIS acute toxicity mode of action, Protein binding**, and **Verhaar classification** profilers. Based on age and complexity of the profilers the **EcoSAR classification** and **Protein binding** are the top tier profilers for acute aquatic toxicity, while **OASIS acute toxicity mode of action** is the second tier profiler and the **Verhaar classification** is the lowest tier profiler.

2. Consistency in the profiler groupings results is an indication that a good chemical category has been achieved.

3. Inconsistency in the profiling grouping or when the profiling results appear in red is an indication that some sort of subcategorization will be necessary to achieve a good chemical category.

4. Since initial profiling (step 1) is designed to selects broad chemical categories, it is often necessary to subcategorize to more narrowly define the chemical category. Subcategorization can be done with either the primary profiles, especially the **EcoSAR classification** or **Protein binding** or done with the secondary profilers **Organic functional groups**, **Metabolism**, and **Superfragment**. In any case subcategorization is done as an iterative process. Since secondary profilers are based on imperfect structural similarity, it is important to review the list of structures provide with each subcategorization routine to assure analogues are not eliminating for unknown reasons.

5. Databases within the Toolbox, which include acute aquatic effects potency data are **ECETOC Aquatic Toxicity**, **Japan Aquatic**, **OASIS Aquatic**, and **US-EPA ECOTOX**. In order to collect the largest number of compounds possible within a chemical category, always use all four databases.

6. Since the OECD is not responsible for the quality of the endpoint data in the Toolbox, it is recommended that the data used in any data gap filling exercise be checked.

References

Russom, C.L. Bradbury, S.P. Broderius, S.J. Hammermeister, D.E. and Drummond, R.A. 1997. Predicting modes of toxic action from chemical structure: acute toxicity in the fathead minnow (*Pimephales promelas*), Environ. Toxicol. Chem. 16:948–967.

Dimitrov, S.D. Low, L.K. Patlewicz, G.Y. Kern, P.S. Dimitrova, G. D. Comber, M.H.I. Phillips, R.D. Niemela, J. Bailey, P.T. Mekenyan, O.G. 2005. Skin sensitization: Modeling based on skin metabolism simulation and formation of protein conjugates. Internat. J. Toxicol., 24:189-204.

Verhaar, H. J. M. van Leeuwen, C. J. Hermens, J. L. M. 1992. Classifying environmental pollutants. 1: structure-activity relationships for prediction of aquatic toxicity. Chemosphere 25:471-491.