DTU Food
National Food
Institute

User Manual for the Danish (Q)SAR Database

Copyright notice, terms and conditions of use

Permission is granted to use information from the database as is. The database is an expert tool where the final assessment of properties is not dictated by the (Q)SAR estimates, but by the user's own scientific judgment. Aside from the fact that models are never perfect, the (Q)SAR field is under rapid development and models are regularly updated and improved. It is also impossible to provide the detailed information accompanying each individual prediction that is available to those who do not own licences to the software platforms. The structural information in the database stems from many sources and in some cases it may be wrong. The structures are also in some cases abbreviated in that possible anions and cations have been removed. This can have important toxicological significance (e.g. for Heavy Metal salts).

All access to the database should happen through the provided client-side software and without any use of automated workflow or scripting.

Reproduction of information from the database is permitted provided the source is acknowledged as follows: "Danish (Q)SAR Database, Division of Diet, Disease Prevention and Toxicology, National Food Institute, Technical University of Denmark, http://qsar.food.dtu.dk."

The Technical University of Denmark (DTU) is not responsible for any errors or inaccuracies the database may contain and is not liable for any use that may be made of the information contained therein. DTU do not warrant, and hereby disclaim any warranties, with respect to the accuracy, adequacy or completeness of any information obtained from this database. Nor do we warrant that the site will operate in an uninterrupted or error-free manner or that the site and its components are free of viruses or other harmful components. Use of information obtained from or through this site is at your own risk. As a user of this database, you agree to indemnify and hold DTU harmless from any claims, losses or damages, including legal fees, resulting from your use of this database, and to fully cooperate in DTU's defense against any such claims.

The user requests are processed by the server hosting the database which in the process stores information. Only authorized employees have authorized access to the server and reasonable measures are in place to protect the server from unauthorized access. DTU uses the stored user request information solely for error tracking and to collect anonymized statistics (number of users, number of searches, number of report downloads etc.), and we do not release any information at the level of individual searches. However, as the online user access to the database does not happen through a secure connection and as any server/PC/network that the requests pass through may be compromised by unauthorized access, we cannot guarantee that the information submitted by users does not fall into the hands of third parties.

These terms are governed by Danish Law, with the exception of international private law and conflict of law rules, to the extent that such rules would result in the application of another country's law. Any dispute arising between the parties in connection with the use of this database, including the interpretation of the above terms, which cannot be settled amicably by negotiation between the parties, shall be settled by the Court of Lyngby, Denmark, as the court of first instance.

Contents

Background

The Danish QSAR database has been freely available on the internet since 2004. It is a tool that allows industry, research, authorities and others to search for hazard information on chemical substances, especially those with little or no testing data. The information provided may be useful to identify chemical substances of potential concern.

With the EU chemicals legislations, e.g. the REACH regulation, there is increased focus on the use of alternatives to animal testing. The QSAR database is used for a wide variety of tasks such as screening for potentially harmful substances and for assessment of specific substances e.g. in relation to dossier evaluation under REACH. The results in the database have also been used to generate the Danish Advisory Self-classification List and to screen for potential PBTs.

Besides direct replacement of experimental tests in some cases, QSAR predictions can help prioritize further *in vitro* and *in vivo* testing of chemicals. In cases where animal testing is still needed, QSAR predictions of mechanistic properties for the chemical can contribute in optimizing the experimental design. In this way, QSARs can reduce the need for later animal testing. It is anticipated that the use of QSAR predictions, and hence the need for good tools will grow in the future.

The new version of the QSAR database has been rebuilt from scratch, and is an updated, extended and improved version of the previous 2004 version of the online QSAR predictions database. It was published in November 2015 and has since then been expanded and updated a number of times. It contains an improved, user-friendly interface, new functionalities and updated predictions for a considerably larger substance structure set than the previous database. The new database is a dynamic system, which will be updated continuously in terms of functionalities and content.

Introduction

The new Danish QSAR database is a repository of model estimates for more than 600,000 substances. The QSAR models include endpoints for physico-chemical properties, environmental fate, bioaccumulation, eco-toxicity, absorption, metabolism and toxicity. As far as possible all organic single constituent substances that were pre-registered or registered under REACH (around 80,000) are included in the structure set. In addition, chemical structures from other relevant databases are included leading to the new structure set of more than 600,000 unique chemical structures.

When possible, the endpoints have been modelled in the three software systems Leadscope, CASE Ultra and SciQSAR. All DTU in-house models and a number of commercial models have with the kind permission from MultiCASE® been modelled in two or three systems. The structure set has been predicted in the different systems and an overall battery prediction is made. With the battery approach it is in many cases possible to reduce “noise” from the individual model estimates and thereby improve accuracy and/or broaden the applicability domain.

All applied DTU QSAR models are documented in QMRFs (QSAR Model Reporting Format). Permissions to publish predictions for more than 600,000 substances were kindly provided by MultiCASE Inc., Leadscope Inc., SciMatics, ACD/Labs, and US EPA. The published predictions are abbreviated predictions (simple yes/no) and do not include detailed information about specific alerts identified. Applicability domain calls are however available.

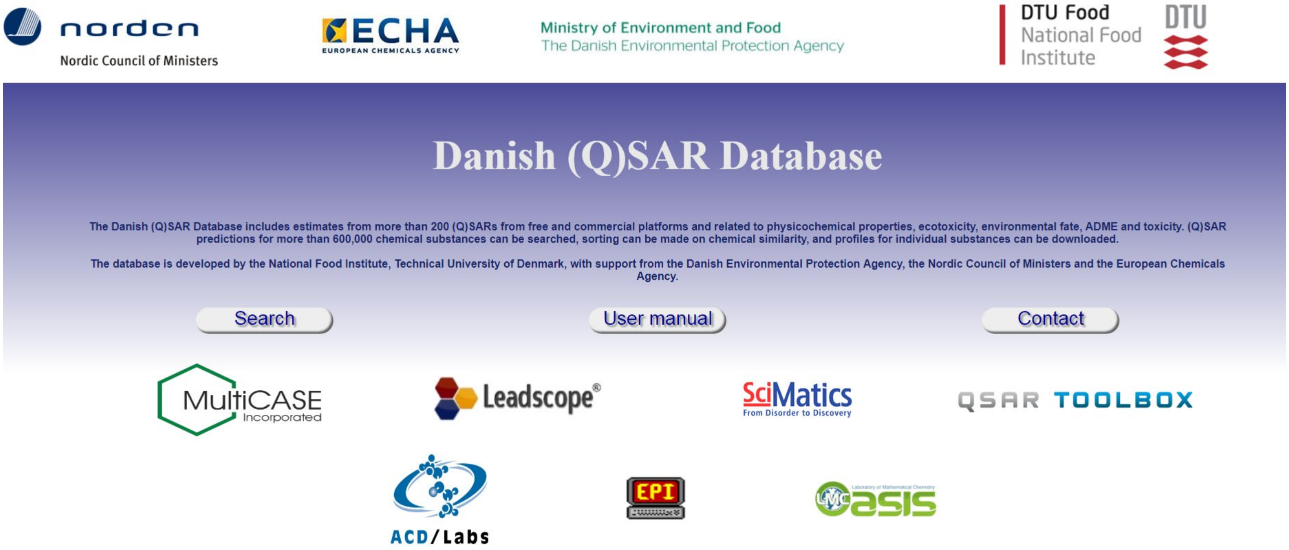
Predictions from a number of OECD QSAR Toolbox profilers have also been included as supporting information to the QSAR predictions. Besides predictions for the parent compounds, predictions for Toolbox simulated transformation products have been included for some profilers. If alerts were predicted, they are included in the database, and if none were found or the compound could not be predicted by the Toolbox this is likewise included. Reference is made to the QSAR Toolbox documentation for the individual profilers via direct link to the documents.

Main features at a glance

- Estimates for more than 600,000 chemicals in over 200 (Q)SAR models.
- Contains experimental training set data for DTU models, for which data are public.
- Search on substance ID and affiliation.
- Structure search on 2D structures as substructure or exact match.
- Search on all contained QSAR predictions and training set data.
- Combination of search results to make complex AND, OR and NOT algorithms.
- Download of QSAR predictions in an RTF format document compatible with Microsoft Word and OpenOffice.
- Sorting on chemical similarity to facilitate read-across groupings.

Launching the Danish QSAR Database

Type in the following link in the address bar of the web browser: http://qsar.food.dtu.dk



Danish (Q)SAR Database

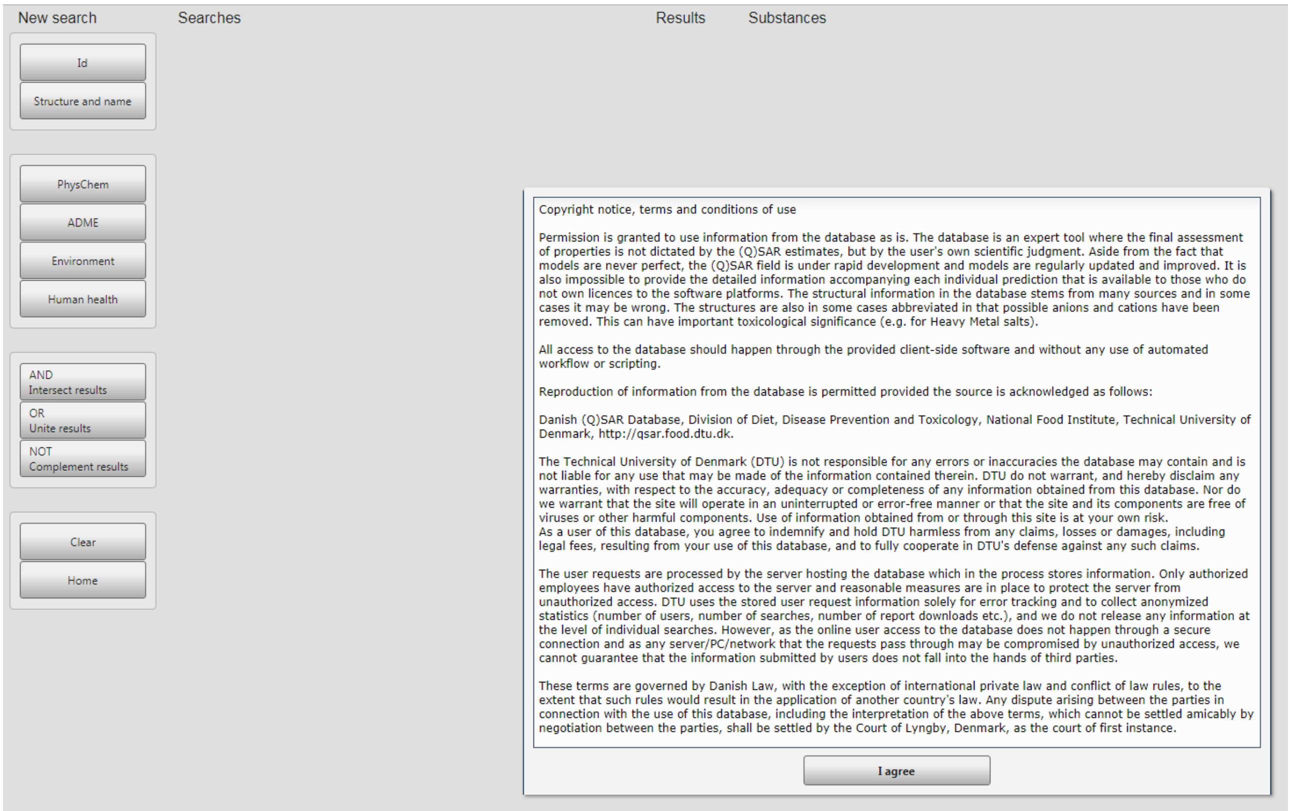
The Danish (Q)SAR Database includes estimates from more than 200 (Q)SARs from free and commercial platforms and related to physicochemical properties, ecotoxicity, environmental fate, ADME and toxicity. (Q)SAR predictions for more than 600,000 chemical substances can be searched, sorting can be made on chemical similarity, and profiles for individual substances can be downloaded.

The database is developed by the National Food Institute, Technical University of Denmark, with support from the Danish Environmental Protection Agency, the Nordic Council of Ministers and the European Chemicals Agency.

norden
Nordic Council of Ministers
ECHA
EUROPEAN CHEMICALS AGENCY
Ministry of Environment and Food
The Danish Environmental Protection Agency
DTU Food
National Food Institute
DTU
Search
User manual
Contact
MultiCASE Incorporated
Leadscope
SciMatics
From Disorder to Discovery
QSAR TOOLBOX
ACD/Labs
EPI
OASIS

Figure 1: Opening screen for the Danish QSAR Database.

To begin searching, click the ***Search*** button, and the screen shown in Figure 2 should appear. Click the button marked ***I agree*** to enter the database.



Copyright notice, terms and conditions of use

Permission is granted to use information from the database as is. The database is an expert tool where the final assessment of properties is not dictated by the (Q)SAR estimates, but by the user's own scientific judgment. Aside from the fact that models are never perfect, the (Q)SAR field is under rapid development and models are regularly updated and improved. It is also impossible to provide the detailed information accompanying each individual prediction that is available to those who do not own licences to the software platforms. The structural information in the database stems from many sources and in some cases it may be wrong. The structures are also in some cases abbreviated in that possible anions and cations have been removed. This can have important toxicological significance (e.g. for Heavy Metal salts).

All access to the database should happen through the provided client-side software and without any use of automated workflow or scripting.

Reproduction of information from the database is permitted provided the source is acknowledged as follows:

Danish (Q)SAR Database, Division of Diet, Disease Prevention and Toxicology, National Food Institute, Technical University of Denmark, http://qsar.food.dtu.dk.

The Technical University of Denmark (DTU) is not responsible for any errors or inaccuracies the database may contain and is not liable for any use that may be made of the information contained therein. DTU do not warrant, and hereby disclaim any warranties, with respect to the accuracy, adequacy or completeness of any information obtained from this database. Nor do we warrant that the site will operate in an uninterrupted or error-free manner or that the site and its components are free of viruses or other harmful components. Use of information obtained from or through this site is at your own risk.
As a user of this database, you agree to indemnify and hold DTU harmless from any claims, losses or damages, including legal fees, resulting from your use of this database, and to fully cooperate in DTU's defense against any such claims.

The user requests are processed by the server hosting the database which in the process stores information. Only authorized employees have authorized access to the server and reasonable measures are in place to protect the server from unauthorized access. DTU uses the stored user request information solely for error tracking and to collect anonymized statistics (number of users, number of searches, number of report downloads etc.), and we do not release any information at the level of individual searches. However, as the online user access to the database does not happen through a secure connection and as any server/PC/network that the requests pass through may be compromised by unauthorized access, we cannot guarantee that the information submitted by users does not fall into the hands of third parties.

These terms are governed by Danish Law, with the exception of international private law and conflict of law rules, to the extent that such rules would result in the application of another country's law. Any dispute arising between the parties in connection with the use of this database, including the interpretation of the above terms, which cannot be settled amicably by negotiation between the parties, shall be settled by the Court of Lyngby, Denmark, as the court of first instance.

Figure 2. Main search screen with disclaimer box.

Main search screen

In the left part of this screen a number of buttons and the headline "New Search" is shown. There are three basic search options in the interface window: **Id**, **Structure** and **Model endpoint** (divided into PhysChem, ADME, Environment and Human health). These are explained in more detail below.

Each search can be combined with others in order to form more complex search queries. The combined searches are performed using the three buttons ***AND***, ***OR*** and ***NOT*** and are described in more detail below in the section: **Combining searches**.

The **Clear** button is used to clear the previous searches from the screen.

Searching by identification data

The ID search button is designed for queries by **Single ID**, **ID List** or **Affiliation** (see Figure 3).

When choosing Single ID, a number of options are possible: **Registry Number**, **EC Number**, **PubChem CID**. To start a Single ID search, type in the query in the white box and click the ***Search*** button. The Registry number can be entered with or without hyphens. The structures matching your search will be listed in a result window similar to the window shown in Figure 8 and give the possibility to download a report containing the prediction results of the resulting substance. The search section and the result window are further described below in the sections **Searches and Results** and **Results window with substances.**

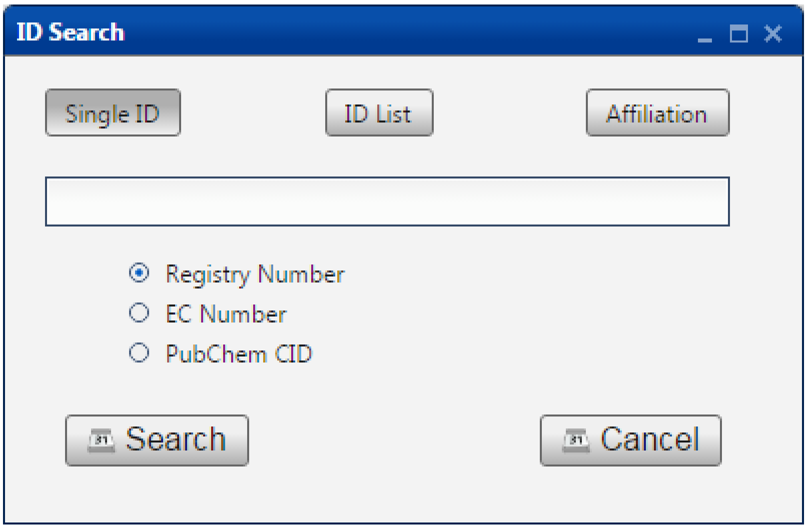


Figure 3. The ID search box.

When choosing ID list (in the ID Search box), two options are possible: **Registry numbers** and **PubChem CIDs**. To start an ID list search, type or paste in the query in the white box and click the ***Search*** button.

Choosing Affiliation gives two options for retrieving database structures: **REACH Pre-registration list** and **PubChem**. To retrieve the structures, choose the database of interest and click the ***Search*** button.

A search example is given in Appendix 1.

Structure, name and similarity search

The **Structure** section offers a 2D molecular editor, where it is possible to input structure fragments to search for. Click the ***Structure*** button to the left on the main search screen to open the molecular editor (see Figure 4). Structures can be drawn or entered from either SMILES or MOL/SDF records. Substructure search, exact structure matching and similarity search are available. Chemical name search is also performed using the molecular editor window.

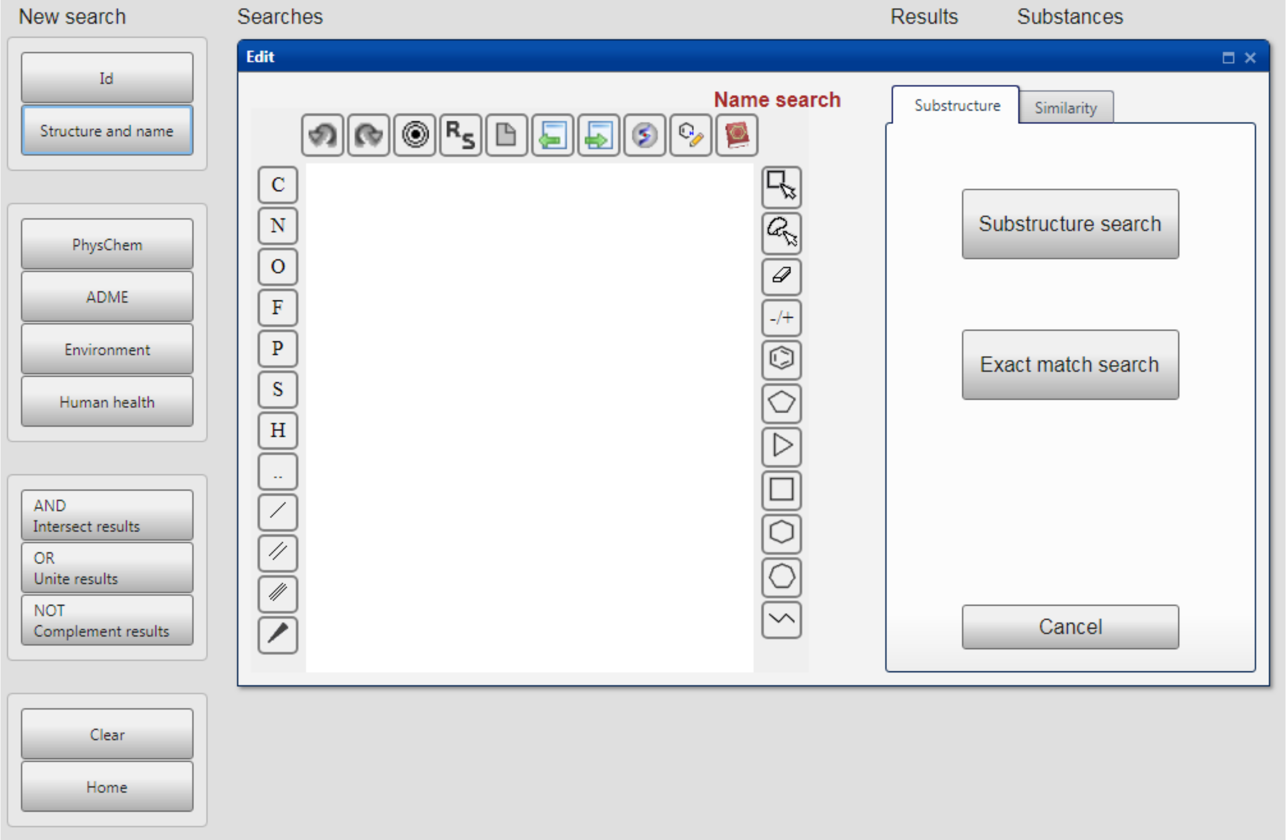


Figure 4. The Structure and name search interface.

A chemical structure can be either built using the molecular editor or entered from a SMILES string or a MOL/SDF record. When building a fragment, to add an item, click on the corresponding button and then click on the blank canvas. Add atoms/fragments/bonds one by one. Click the .. button if you need other atoms than shown in the molecular editor, and then type the element symbol in the blank field, which is case-sensitive; confirm with ‘Ok’ and click the desired atom position in the structure.

To start a structure search, select either **Substructure** or **Similarity** to the right on the screen. When choosing **Substructure**, two options are possible: Select ***Substructure search*** to search for the built fragment as a subfragment within the database, or ***Exact match search*** to search for exactly the same structure. When choosing **Similarity**, the entire database will be ordered in the order of similarity to the query structure. Alternatively, select a number of closest analogs that will be sorted in the order of similarity and displayed. The search will generate a result window similar to the window shown in Figure 8 and include the similarity coefficient for each analog. The result window is further described below in the section: **Results window with substances.**

Editor operations:

Undo: Undoes the last operation.

Redo: Repeats the last undone operation.

Center: Moves the fragment to the center of the canvas.

Toggle R/S labels: Marks R/S isomerism.

Clear: Clears the Edit window.

Import: Imports MOL or SMILES structure decsription.

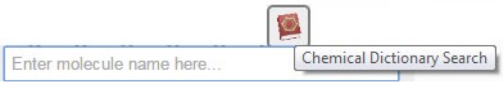
Export: Displays MOL or SMILES description for the current structure.

About: Displays version number etc. of the fragment editor.

2D cleanup/depiction: Corrects bond angles etc.

Chemical name search:

- type a chemical name or a part of it. The structure will be looked up in the US NIH chemical dictionary and displayed in the molecular editor.



- proceed with Exact match to search for the structure in the Danish (Q)SAR Database.

Search examples are given in Appendix 2.

Searching by model endpoint

The ***PhysChem***, ***Environment***, ***ADME*** and ***Human health*** buttons to the left on the main screen can be used to search for specific model endpoints. Each of the four categories covers a number of different endpoints. To start a search by model endpoint, click the category button of interest, e.g. ***Human health***. This will generate a drop-down menu with a list of subcategories as shown in Figure 5. Figure 5 shows an example query to search for prediction results in the model for Bacterial Reverse Mutation Test (Ames test in S. typhimurium (in vitro)), which is found in the genotoxicity subcategory.

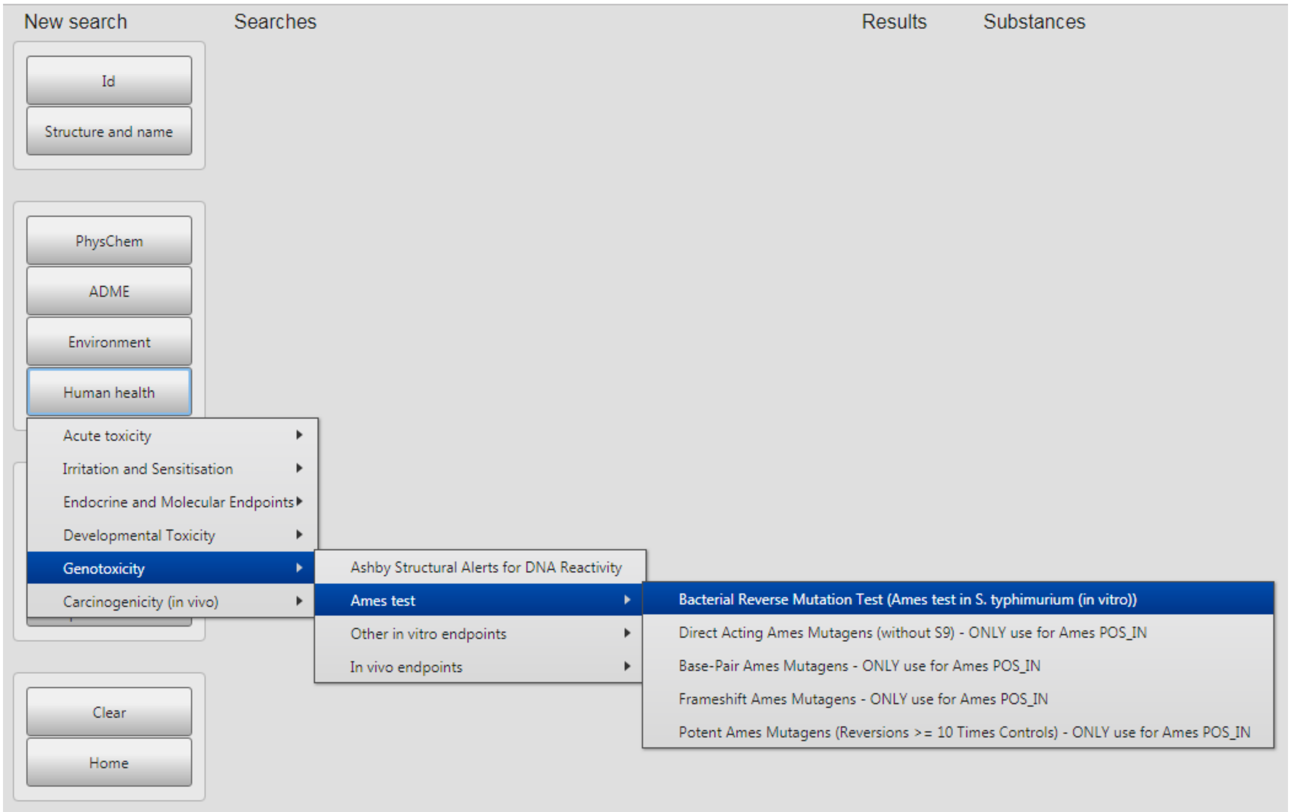
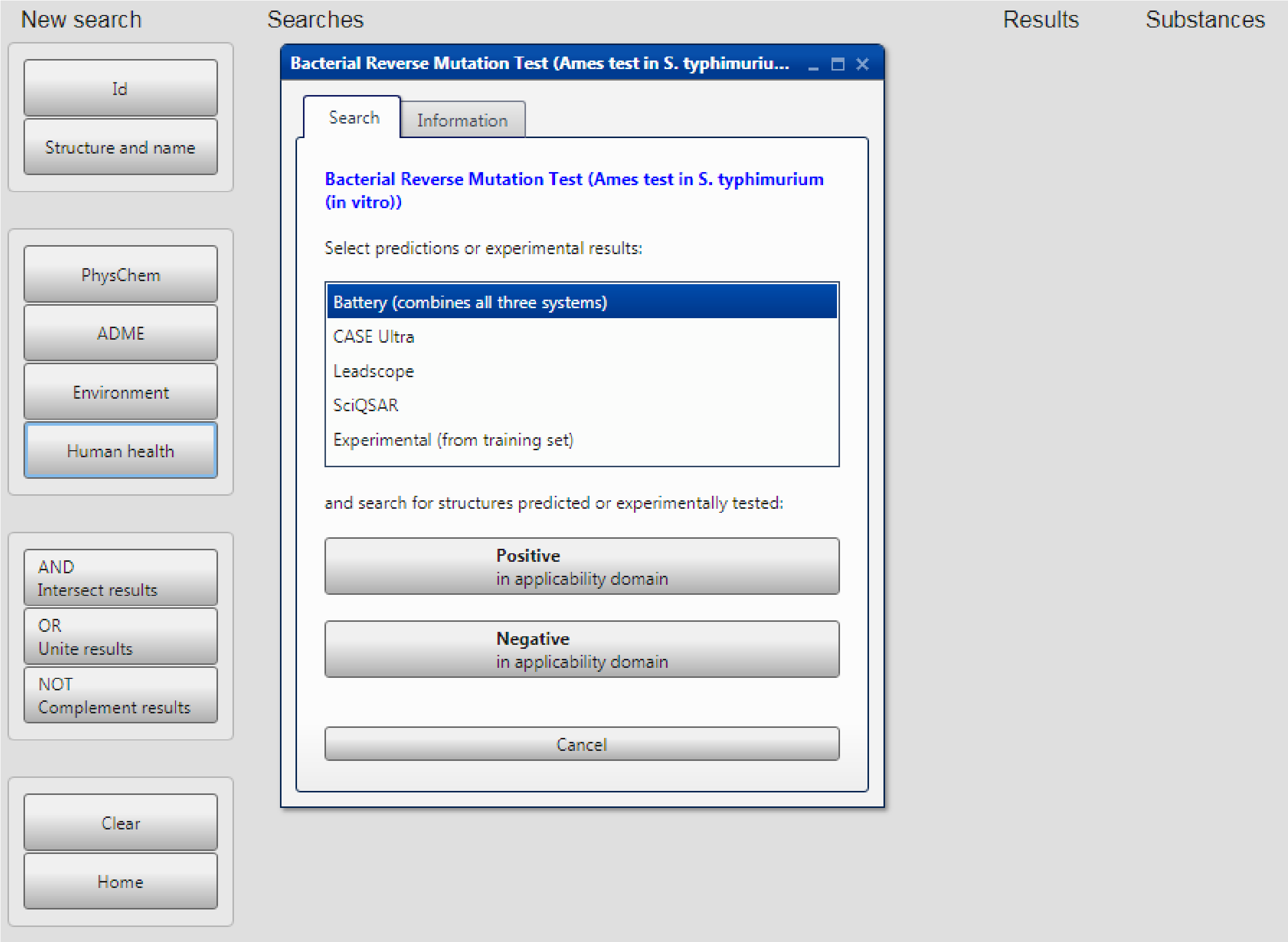

New search
Searches
Results
Substances
Id
Structure and name
PhysChem
ADME
Environment
Human health
Acute toxicity
Irritation and Sensitisation
Endocrine and Molecular Endpoints
Developmental Toxicity
Genotoxicity
Carcinogenicity (in vivo)
Ashby Structural Alerts for DNA Reactivity
Ames test
Other in vitro endpoints
In vivo endpoints
Bacterial Reverse Mutation Test (Ames test in S. typhimurium (in vitro))
Direct Acting Ames Mutagens (without S9) - ONLY use for Ames POS_IN
Base-Pair Ames Mutagens - ONLY use for Ames POS_IN
Frameshift Ames Mutagens - ONLY use for Ames POS_IN
Potent Ames Mutagens (Reversions >= 10 Times Controls) - ONLY use for Ames POS_IN
Clear
Home

Figure 5. An example query to search for prediction results in Ames test. As shown, a number of submodels are available.

When the model of interest is chosen, a dialog box appears (Figure 6). Select the heading **Search** at the top of the dialog box to start a search. The menu in the dialog box depends on whether the selected model is made in one or more software systems. The selected model in Figure 6 is made in three systems, CASE Ultra, Leadscope and SciQSAR. Based on predictions from the three systems, a fourth and overall battery prediction is made. These four predictions (three predictions from the individual systems and the battery prediction) can be selected individually. The battery prediction approach is further described below in the section: **Battery algorithm**. It is also possible to select and search for experimental results from the training set.



Bacterial Reverse Mutation Test (Ames test in S. typhimurium (in vitro))

Select predictions or experimental results:

- Battery (combines all three systems)
- CASE Ultra
- Leadscope
- SciQSAR
- Experimental (from training set)

and search for structures predicted or experimentally tested:

Figure 6. Dialog box from query shown in Figure 5.

Select the relevant results type (predictions/experimental) in the dialog box and then click either the ***positive*** or ***negative*** button to start the search. Only the predictions within applicability domain will be searched and displayed.

The search will generate the browser window shown in Figure 8 and give the possibility to download a report containing the prediction results in the rtf.file format. The results window is further described below in the section: **Results window with substances.**

Information about the selected model can be found by selecting the **information** tab at the top of the dialog box. A list of options will appear enabling you to download QMRFs of the relevant model versions.

Search examples are given in Appendix 3.

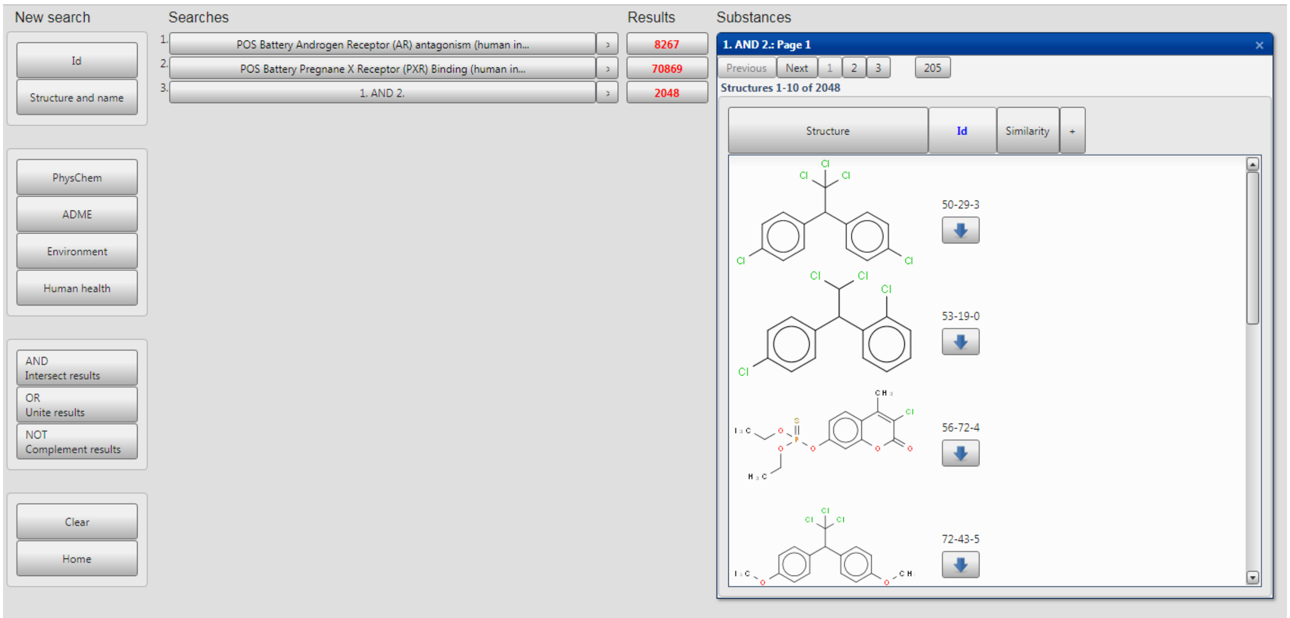
Combining searches

Combinations of searches are also possible. These are performed using the two buttons to the left on the main screen, ***AND*** and ***OR.*** Using the ***OR*** button will display all substances from two or more searches, whereas the ***AND*** button will only display the intersection of the individual searches. The individual queries are made as described in the previous text so that they appear under “Searches” on the main search screen.

To combine searches, click the search definition buttons for the searches of interest that appear under the field “Searches”. This will highlight the text in the selected buttons, which change the state to ‘selected’ and the foreground color to green. Then select either the ***AND*** or ***OR*** button to start the search. The results of the combination search are displayed to the right under “Searches”, “Results” and “Substances”. The example in figure 7 shows the result of a combination of searches for AR antagonism and PXR binding using the ***AND*** button. The result window is further described below in the section: **Results window with substances.**

The ***NOT*** button to the left on the main search screen is for inverting a search. Click the search definition button of interest (only one) under the field “Searches” and select ***NOT***. Inverted searches, as well as results of AND and OR searches, can in turn be combined with other individual or combined ones to form more complex combined searches.

Search example is given in Appendix 4.



Searches	Results
1. POS Battery Androgen Receptor (AR) antagonism (human in...	8267
2. POS Battery Pregnane X Receptor (PXR) Binding (human in...	70869
3. 1. AND 2.	2048

1. AND 2.: Page 1

Structures 1-10 of 2048

Structure	Id
	50-29-3
	53-19-0
	56-72-4
	72-43-5

Figure 7. An example where searches for AR antagonism and PXR binding are combined by using the ***AND*** button.

Searches and Results sections

Every time you perform a search, several new screen elements will appear. A search definition button will be added to the Searches section (Figure 7). It can be used for combining searches, which is described in the section **“Combining searches”**.

Another button in the Results section will display the number of structures resulting from the search. The actual structures will be listed in a browser window similar to the window shown in Figure 8 (described below in the section **Results window with substances**).

The Searches and Results sections will keep track of all searches you have performed. You can clear them by clicking the Clear button, or delete individual searches using the small '>' button next to the search definition. You can revisit previous search results at any time by clicking the button displaying the number of structures in the Results section. The relevant searches are not executed again but instead retrieved quickly from a repository of searches.

Results window with substances

The searches described in the previous sections will generate a results window and give the possibility to download a report containing database results for selected substances. One report per substance will be generated.

The example in Figure 8 shows the result of a Model endpoint search of the Bacterial Reverse Mutation Test. The results of the search are displayed to the right under "Results" and "Substances".

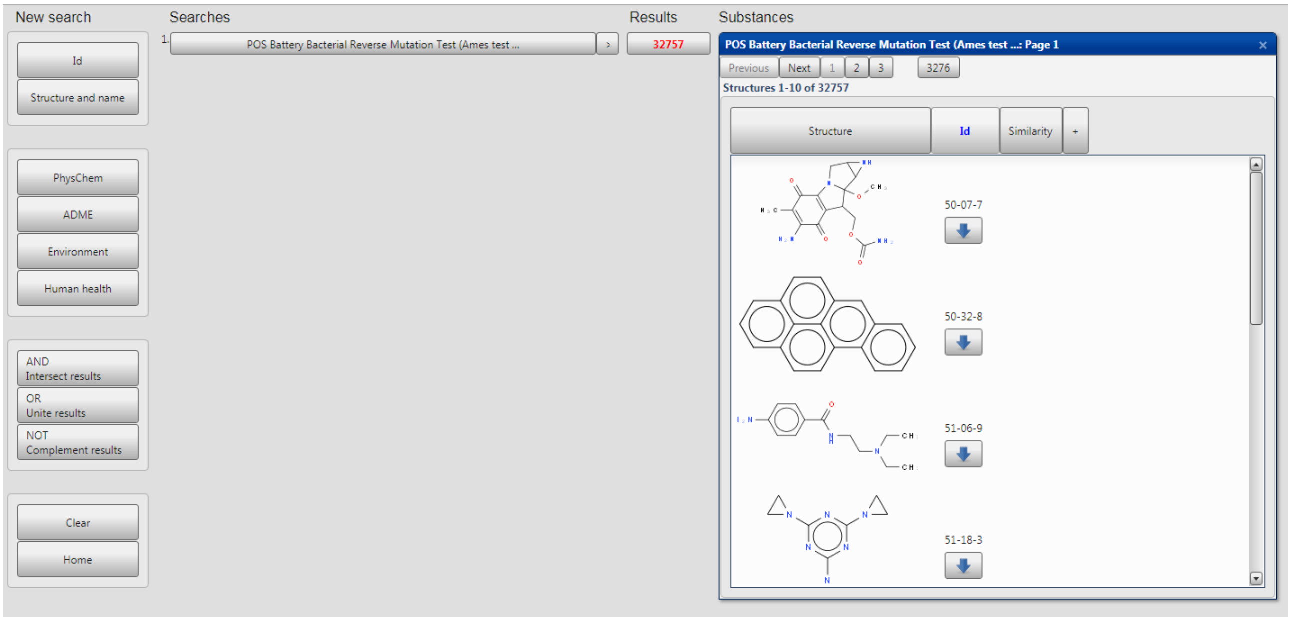


Figure 8. An example of a results window from a Model endpoint search of the Bacterial Reverse Mutation Test.

The window under "Substances" shows the resulting structures. When there are more than 10 structures, they are shown in pages with 10 structures per page. Use the top button row (Previous, Next, First, Current, Last etc.) to navigation through the result pages. All result pages are directly accessible the moment the search is executed, so you can e.g. view any page directly without having to go first through the preceding ones.

To download a single substance report, click the ⬇ button in the id column next to the substance of interest. This will provide an .RTF file containing all predictions as well as training set data when available. The .RTF document format is supported by Microsoft Word, OpenOffice and other viewers/editors.

Clicking the ***Similarity*** button will open the 2D fragment editor, where it is possible to search for substances similar to a query substance within the current result set. The current result set will be ordered by decreasing similarity to the query substance.

To revert back to the Id order, click the Id button above the structure list.

Clicking the + button next to ***Similarity*** opens a dialog box, where you can select any database property (experimental or predicted in any model and predictive system) and display its values in the result window. You can select up to eight properties to display. The extra information will be displayed in new columns and refresh as you navigate through result pages.

The Substances window can be resized and moved and scroll bars will automatically appear if necessary.

Technical requirements and notes

All operations with the Danish (Q)SAR database are performed in a web browser. There is no need to download or install any software. Likewise, there is no need to install any browser plugin or add-on (the previous version of the web site used Java).

The system is can be accessed from both personal computers and mobile devices. The minimum screen resolution for using the system is 640x480 pixels. For convenience, higher resolution display settings can be recommended (preferably 1280 or more pixels on the horizontal axis).

The client-side software is implemented entirely in JavaScript and is compatible with all major browsers and operating systems without the need for third-party software. Depending on the security settings of your browser, you may need to enable JavaScript in order to use the website.

The system has been tested with the following browser versions: Google Chrome 46.0, Microsoft Internet Explorer 11, Opera 33.0, Mozilla Firefox 37.0.2.

Battery algorithm

Some of the models are made in two or three of the following independent systems: CASE Ultra (CU), Leadscope Predictive Data Miner (LS) and SciQSAR (SQ). The systems are described in Appendix 5. Based on predictions from each of the applied systems, a battery prediction is made using a so-called battery algorithm. The battery approach can give more reliable predictions and can also expand the applicability domain, which was shown in a previous pilot project including 32 different models and the three systems mentioned above (not published).

For a given effect, QSAR predictions are made in each of the independent QSAR model systems and combined into a battery prediction by using the criteria shown in Table 1. The first column shows the total number of predictions (positive/negative) in domain. The next two columns show the number of positive and negative predictions, respectively. The final battery prediction based on the individual predictions is shown in the fourth column.

Table 1. Battery algorithm.

Total POS/NEG in domain	POS in domain	NEG in domain	Battery prediction[a]	Remarks
3	3	0	POS_IN	
3	0	3	NEG_IN	
3	2	1	POS_IN	
3	1	2	INC_OUT or (see remark) NEG_IN	EXCEPT when CU and LS are both NEG_IN, in this case the battery call is NEG_IN
2	2	0	POS_IN	
2	1	1	INC_OUT	
2	0	2	NEG_IN	
1	1	0	POS_OUT	
1	0	1	NEG_OUT	
0	0	0	INC_OUT	If minimum one prediction (out of domain)
0	0	0	-	None predicted

[a] POS, positive; NEG, negative; INC, inconclusive; IN, inside applicability domain; OUT, outside applicability domain. [b] Less weight is put on an SQ POS compared to LS or CU POS in cases where LS and CU agree on a NEG in AD prediction, because SQ in many cases has lower specificity than LS and CU.

Table 2. Training set numbers and performance results. See QMRFs for more information.

Endpoint	N in training set	Software	Performance result (%)[a]
Not ready biodegradability (POS=Not Ready)	735	CASE Ultra	Sens=68.9, Spec=87.8, Conc=77.2
		Leadscope	Sens=87.3, Spec=85.2, Conc=86.4
		SciQSAR	Sens=63.0, Spec=92.7, Conc=77.8
Fathead minnow 96h LC50 (mg/L)	565	Leadscope	R^2=0.75, Q^2=0.73
		SciQSAR	R^2=0.74, Q^2=0.72
Daphnia magna 48h EC50 (mg/L)	626	Leadscope	R^2=0.67, Q^2=0.64
		SciQSAR	R^2=0.65, Q^2=0.63
Pseudokirchneriella s. 72h EC50 (mg/L)	531	Leadscope	R^2=0.74, Q^2=0.71
		SciQSAR	R^2=0.64, Q^2=0.60
Cytochrome P450 2D6 (CYP2D6) substrates (human clinical data)	746	CASE Ultra	Sens=43.9, Spec=87.0, Conc=74.1
		Leadscope	Sens=60.0, Spec=89.4, Conc=80.1
		SciQSAR	Sens=59.5, Spec=79.8, Conc=73.1
Cytochrome P450 2C9 (CYP2C9) substrates (human clinical data)	736	CASE Ultra	Sens=30.6, Spec=83.6, Conc=68.8
		Leadscope	Sens=30.0, Spec=89.6, Conc=75.4
		SciQSAR	Sens=26.3, Spec=91.5, Conc=74.7
Rat oral	6,464	ACDLabs	Ext. validation, RI>0.5, Q^2=0.64
Rat intraperitoneal	3,751	ACDLabs	Ext. validation, RI>0.5, Q^2=0.56
Mouse oral	14,678	ACDLabs	Ext. validation, RI>0.5, Q^2=0.55
Mouse intraperitoneal	27,004	ACDLabs	Ext. validation, RI>0.5, Q^2=0.61
Mouse intravenous	14,972	ACDLabs	Ext. validation, RI>0.5, Q^2=0.66
Mouse subcutaneous	6,432	ACDLabs	Ext. validation, RI>0.5, Q^2=0.57
Maximum recommended daily dose (MRDD) in humans ≤ 2.69 mg/kg-2bw/d	1,222	CASE Ultra	Sens=69.4, Spec=92.5, Conc=82.5
		Leadscope	Sens=78.6, Spec=82.5, Conc=80.7
		SciQSAR	Sens=73.1, Spec=77.3, Conc=75.3
Severe skin irritation in rabbit	836	CASE Ultra	Sens=63.4, Spec=86.7, Conc=75.8
		Leadscope	Sens=79.5, Spec=81.7, Conc=80.6
		SciQSAR	Sens=77.3, Spec=71.3, Conc=74.3
Allergic contact dermatitis in guinea pig and human	1,032	CASE Ultra	Sens=76.7, Spec=93.9, Conc=89.3
		Leadscope	Sens=75.0, Spec=96.3, Conc=90.8
		SciQSAR	Sens=61.6, Spec=96.8, Conc=85.8

Endpoint	N in training set	Software	Performance result (%)[a]
Respiratory sensitisation in humans	80	CASE Ultra	Sens=68.2, Spec=96.3, Conc=86.4
		Leadscope	Sens=91.7, Spec=95.5, Conc=93.9
		SciQSAR	Sens=80.0, Spec=87.5, Conc=83.8
Profiler[b]: Protein binding by OASIS 1.4	N/A (101 alerts)	OECD QSAR Toolbox	N/A
Profiler[b]: Protein binding by OECD	N/A (104 alerts)	OECD QSAR Toolbox	N/A
Profiler[b]: Protein binding potency Cys (DPRA 13%)	229 (77 alerts)	OECD QSAR Toolbox	N/A
Profiler[b]: Protein binding potency Lys (DPRA 13%)	228 (73 alerts)	OECD QSAR Toolbox	N/A
Profiler[b]: Keratinocyte gene expression	~300 (21 alerts)	OECD QSAR Toolbox	N/A
Estrogen Receptor α binding (human in vitro) ALL	802	CASE Ultra	Sens=60.9, Spec=95.2, Conc=85.7
		Leadscope	Sens=75.2, Spec=90.1, Conc=84.7
		SciQSAR	Sens=67.3, Spec=89.0, Conc=81.3
Estrogen Receptor α binding (human in vitro) Balanced	595	CASE Ultra	Sens=81.7, Spec=89.2, Conc=85.4
		Leadscope	Sens=83.7, Spec=89.0, Conc=86.3
		SciQSAR	Sens=76.1, Spec=83.3 Conc=79.8
Estrogen Receptor α activation (human in vitro)	481	CASE Ultra	Sens=73.7, Spec=86.6, Conc=80.9
		Leadscope	Sens=73.1, Spec=86.6, Conc=80.7
		SciQSAR	Sens=77.9, Spec=80.8, Conc=79.6
Profiler[b]: Estrogen Receptor Binding	N/A	OECD QSAR Toolbox	N/A
Profiler[b]: rtER Expert System - USEPA	N/A	OECD QSAR Toolbox	N/A
Androgen Receptor antagonism (human in vitro)	874	CASE Ultra	Sens=57.4, Spec=87.2, Conc=78.3
		Leadscope	Sens=51.7, Spec=91.2, Conc=80.4
		SciQSAR	Sens=56.3 Spec=91.1, Conc=81.9
Arylhydrocarbon (AhR) Activation – Rational final model	4,625	Leadscope	TBA (manuscript under review)
Arylhydrocarbon (AhR) Activation – Random final model	4,625	Leadscope	TBA (manuscript under review)
Thyroid receptor α binding – log(IC_{50} in µM) (human in vitro)	118	CASE Ultra	Q^2=0.59
		Leadscope	R^2=0.83, Q^2=0.68
		SciQSAR	R^2=0.64, Q^2=0.57
	130	CASE Ultra	Q^2=0.61

Endpoint	N in training set	Software	Performance result (%)[a]
Thyroid receptor β binding – log(IC_{50} in µM) (human in vitro)		Leadscope	R^2=0.83, Q^2=0.64
		SciQSAR	R^2=0.65, Q^2=0.58
Thyroperoxidase (TPO) inhibition QSAR1 (rat in vitro)	877	Leadscope	Sens=72.4, Spec=89.0, BA=80.6 Ext. validation: Sens=79.7, Spec=90.8, BA=85.3
Thyroperoxidase (TPO) inhibition QSAR2 (rat in vitro)	1,519	Leadscope	Sens=75.6, Spec=89.8, BA=82.7
Pregnane X receptor binding (human in vitro)	631	CASE Ultra	Sens=72.3, Spec=89.0, Conc=78.5
		Leadscope	Sens=80.4, Spec=80.4, Conc=80.4
		SciQSAR	Sens=79.9, Spec=82.7, Conc=81.4
Pregnane X Receptor (PXR) Binding (Human *in vitro*) NEW	1,504	Leadscope	Sens=86.6, Spec=98.5, Conc=92.6 (Leadscope 2*5-fold cross-validation)
Pregnane X Receptor (PXR) Activation (Human *in vitro*)	2,176	Leadscope	Sens=89.1, Spec=98.6, BA=93.9 (Leadscope 2*5-fold cross-validation)
Pregnane X Receptor (PXR) Activation (Rat *in vitro*)	2,330	Leadscope	Sens=86.5, Spec=97.4, BA=92.0 (Leadscope 2*5-fold cross-validation)
CYP3A4 Induction (Human *in vitro*)	2,271	Leadscope	Sens=86.7, Spec=98.2, BA=92.5 (Leadscope 2*5-fold cross-validation)
Teratogenic potential in Humans	323	CASE Ultra	Sens=65.0, Spec=85.1, Conc=76.4
		Leadscope	Sens=72.0, Spec=85.5, Conc=80.1
		SciQSAR	Sens=64.6, Spec=92.7, Conc=81.4
Ashby structural alerts	782	CASE Ultra	Sens=89.7, Spec=95.1, Conc=91.9
		Leadscope	Sens=87.5, Spec=90.7, Conc=88.5
		SciQSAR	Sens=81.7, Spec=80.6, Conc=81.1
Bacterial reverse mutation test (Ames test in S. typhimurium in vitro)	4,102	CASE Ultra	Sens=83.9, Spec=89.1, Conc=86.4
		Leadscope	Sens=84.3, Spec=85.7, Conc=84.9
		SciQSAR	Sens=79.3, Spec=79.1, Conc=79.2
Direct acting Ames mutagens (without S9) – ONLY use for Ames POS_IN	388	CASE Ultra	Sens=63.5, Spec=90.4, Conc=79.5
		Leadscope	Sens=66.9, Spec=78.9, Conc=74.0
		SciQSAR	Sens=56.5, Spec=72.9, Conc=68.6
Base pair Ames mutagens - ONLY use for Ames POS_IN	204	CASE Ultra	Sens=52.8, Spec=88.4, Conc=71.9
		Leadscope	Sens=70.2, Spec=66.4, Conc=68.4
		SciQSAR	Sens=68.6, Spec=67.7, Conc=68.1
	309	CASE Ultra	Sens=73.5, Spec=84.1, Conc=78.9

Endpoint	N in training set	Software	Performance result (%)[a]
Frame shift Ames mutagens - ONLY use for Ames POS_IN		Leadscope	Sens=74.4, Spec=78.6, Conc=76.6
		SciQSAR	Sens=68.3, Spec=78.2, Conc=73.8
Potent Ames mutagens, reversions ≥ 10 times controls - ONLY use for Ames POS_IN	187	CASE Ultra	Sens=73.7, Spec=87.7, Conc=81.2
		Leadscope	Sens=68.9, Spec=70.0, Conc=69.8
		SciQSAR	Sens=75.0, Spec=74.7, Conc=74.9
Profiler[b]: DNA alerts for AMES by OASIS	N/A (85 alerts)	OECD QSAR Toolbox	See OECD QSAR Toolbox documentation on internal predictivity for individual alerts
Profiler[b]: In vitro mutagenicity (Ames test) alerts by ISS	N/A (85 alerts)	OECD QSAR Toolbox	See OECD QSAR Toolbox documentation on internal predictivity for individual alerts
Chromosome aberrations in CHO cells (*in vitro*)	233	CASE Ultra	Sens=40.4, Spec=94.5, Conc=74.4
		Leadscope	Sens=54.1, Spec=79.3, Conc=68.8
		SciQSAR	Sens=50.5, Spec=84.3, Conc=70.3
Chromosome aberrations in CHL cells (*in vitro*)	600	CASE Ultra	Sens=63.3, Spec=86.7, Conc=76.4
		Leadscope	Sens=74.6, Spec=75.2, Conc=74.9
		SciQSAR	Sens=73.0, Spec=72.8, Conc=72.9
Mutations in thymidine kinase locus in mouse lymphoma cells (*in vitro*)	555	CASE Ultra	Sens=76.5, Spec=86.3, Conc=81.2
		Leadscope	Sens=85.1, Spec=83.8, Conc=84.4
		SciQSAR	Sens=79.1, Spec=80.5, Conc=79.8
Mutations in HGPRT locus in CHO cells (*in vitro*)	239	CASE Ultra	Sens=75.4, Spec=84.5, Conc=78.9
		Leadscope	Sens=81.7, Spec=78.4, Conc=80.5
		SciQSAR	Sens=80.0, Spec=73.0, Conc=76.5
Unscheduled DNA synthesis (UDS) in rat hepatocytes (*in vitro)*	415	CASE Ultra	Sens=60.6, Spec=87.0, Conc=74.1
		Leadscope	Sens=74.1, Spec=70.1, Conc=72.4
		SciQSAR	Sens=69.6, Spec=72.5, Conc=71.1
Syrian hamster embryo (SHE) cell transformation (*in vitro)*	363	CASE Ultra	Sens=50.8, Spec=86.9, Conc=74.0
		Leadscope	Sens=71.6, Spec=76.5, Conc=74.5
		SciQSAR	Sens=76.1, Spec=66.5, Conc=71.3
Profiler[b]: DNA alerts for CA and MNT by OASIS	N/A (85 alerts)	OECD QSAR Toolbox	See OECD QSAR Toolbox documentation on internal predictivity for individual alerts
Profiler[b]: Protein binding alerts for Chromosomal aberration by OASIS	N/A (35 alerts)	OECD QSAR Toolbox	See OECD QSAR Toolbox documentation on internal predictivity for individual alerts
Sex-linked recessive lethal (SLRL) test in Drosophila m. (*in vivo*)	367	CASE Ultra	Sens=75.4, Spec=92.0, Conc=83.6
		Leadscope	Sens=79.1, Spec=80.3, Conc=79.6

Endpoint	N in training set	Software	Performance result (%)[a]
		SciQSAR	Sens=74.2, Spec=78.3, Conc=76.2
Micronucleus test in mouse erythrocytes (*in vivo*)	357	CASE Ultra	Sens=31.2, Spec=95.2, Conc=75.7
		Leadscope	Sens=64.1, Spec=77.6, Conc=72.3
		SciQSAR	Sens=52.1, Spec=83.3, Conc=69.7
Dominant lethal mutations in rodents (*in vivo*)	191	CASE Ultra	Sens=42.4, Spec=92.7, Conc=73.7
		Leadscope	Sens=61.5, Spec=80.4, Conc=71.8
		SciQSAR	Sens=57.7, Spec=81.4, Conc=71.7
Sister chromatid exchange in mouse bone marrow cells (*in vivo*)	265	CASE Ultra	Sens=91.8, Spec=94.8, Conc=93.9
		Leadscope	Sens=88.6, Spec=95.9, Conc=94.0
		SciQSAR	Sens=76.7, Spec=93.2, Conc=86.8
Comet assay in mouse (*in vivo*)	286	CASE Ultra	Sens=60.1, Spec=93.1, Conc=82.9
		Leadscope	Sens=86.6, Spec=80.8, Conc=83.1
		SciQSAR	Sens=82.4, Spec=82.0, Conc=82.2
Profiler[b]: In vivo mutagenicity (Micronucleus) alerts by ISS	N/A (35 alerts)	OECD QSAR Toolbox	See OECD QSAR Toolbox documentation on internal predictivity for individual alerts
FDA RCA cancer male rat (*in vivo*)	1,324	CASE Ultra	Sens=34.2, Spec=95.0, Conc=63.9
		Leadscope	Sens=62.6, Spec=74.7, Conc=69.2
FDA RCA cancer female rat (*in vivo*)	1,321	CASE Ultra	Sens=44.4, Spec=93.3, Conc=71.6
		Leadscope	Sens=57.7, Spec=83.6, Conc=72.7
FDA RCA cancer rat (*in vivo*)	1,379	CASE Ultra	Sens=41.7, Spec=94.0, Conc=66.9
		Leadscope	Sens=57.1, Spec=82.3, Conc=71.2
FDA RCA cancer male mouse (*in vivo*)	1,197	CASE Ultra	Sens=38.4, Spec=86.1, Conc=66.1
		Leadscope	Sens=58.6, Spec=81.4, Conc=71.9
FDA RCA cancer female mouse (*in vivo*)	1,208	CASE Ultra	Sens=41.5, Spec=85.9, Conc=65.6
		Leadscope	Sens=59.2, Spec=80.6, Conc=71.3
FDA RCA cancer mouse (*in vivo*)	1,221	CASE Ultra	Sens=43.1, Spec=86.9, Conc=66.9
		Leadscope	Sens=56.5, Spec=83.9, Conc=72.7
FDA RCA cancer rodent (*in vivo*)	1,530	CASE Ultra	Sens=51.4, Spec=88.3, Conc=68.2
		Leadscope	Sens=65.9, Spec=76.2, Conc=71.3
Profiler[b]: Carcinogenicity (genotox and nongenotox) alerts by ISS	N/A (55 alerts)	OECD QSAR Toolbox	See OECD QSAR Toolbox documentation on internal predictivity for individual alerts
Profiler[b]: Oncologic Primary Classification	N/A (48 alerts)	OECD QSAR Toolbox	N/A

Endpoint	N in training set	Software	Performance result (%)[a]
Liver specific cancer in rat or mouse (*in vivo*)	320	CASE Ultra	Sens=31.1, Spec=92.0, Conc=70.9
		Leadscope	Sens=35.6, Spec=88.6, Conc=69.3
		SciQSAR	Sens=38.5, Spec=84.8, Conc=69.1

[a] Where nothing else is mentioned, the results stem from DTU 5 * 2-fold cross-validation. Sens: sensitivity; Spec: specificity; Conc: concordance; BA: balanced accuracy; Ext. validation: external validation; RI: reliability index.

[b] OECD Toolbox profilers predictions: Supporting information to be used together with relevant QSAR predictions

Table 3. Software names and versions for physical-chemical and environmental models

Predicted property	Software
Melting Point (deg C), Boiling Point (deg C), Melting Point Experimental (deg C), Boiling Point Experimental (deg C), Vapour Pressure (mm Hg), Vapour Pressure (Pa), Vapour Pressure Experimental (mm Hg), Vapour pressure Subcooled Liquid (Pa)	EPI MPBPWIN v1.43
HLC Bond Method (atm-m3/mole), HLC Group Method (atm-m3/mole), HLC Via VP/WSol (atm-m3/mole), HLC Via VP/WSol (Pa-m3/mole), Henrys Law Const. Exp db (Pa-m3/mole), Henrys Law Const. Exp (atm-m3/mole)	EPI HENRYWIN v3.20
Water solubility from Kow (mg/L), Water solubility Exp (mg/L), Water solubility Exp Ref, Log Kow, Log Kow Exp, Log Kow Exp Ref,	EPI WSKOW v1.42
Water solubility from Fragments (mg/L)	EPI WATERNT v1.01
Hydrolysis half-life at pH 7 and 8	EPI HYDROWIN v2.00
pKa Acid, pKa Base	ACD/ ToxSuite 2.95.1 Ionization\ACD/Labs pKa module (GALAS)
LogD	ACD/ ToxSuite 2.95.1 Ionization\ACD/Labs LogD module (GALAS)
Log Koa, Log Kaw	EPI KOAWIN v1.10
Kp (m3/ug) Mackay-based, Kp (m3/ug) Koa-based, Phi Junge-Pankow-based, Phi Mackay-based, Phi Koa-based	EPI AEROWIN v1.00
Koc from MCI (L/kg), Log Koc from MCI, Koc from Kow (L/kg), Log Koc from Kow	EPI KOCWIN v2.00
Mass Amount (%), Half-Life (hr), Emissions (kg/hr)	EPI Level III Fugacity Model (EPI Suite v4.11)
Sewage Treatment Plant (STP) overall chemical mass balance using 10,000 hr (%)	EPI STPWIN (EPI Suite v4.11)
Half-Life (d), Half-Life (hr), Overall Rate Const. (OH: E-12 cm3/molecule-sec and OZ: E-17 cm3/molecule-sec)	EPI AOPWIN v1.92
Biowin1 (linear model) Probability of Rapid Biodegradation, Biowin2 (non-linear model) Probability of Rapid Biodegradation, Biowin3 Expert Survey Ultimate Biodegradation, Biowin3 Expert Survey Ultimate Timeframe, Biowin4 Expert Survey Primary Biodegradation, Biowin4 Exp. Survey Primary Timeframe, Biowin5 (MITI linear model) Biodegradation Probability, Biowin6 (MITI non-linear model) Biodegradation Probability, Biowin7 (Anaerobic Linear) Biodegradation Probability, Petroleum Hydrocarbon Biodegradation Half-Life (days)	EPI BIOWIN v4.10
BCF (L/kg wet-wt), Log BCF (L/kg wet-wt), Whole Body Primary Biotransformation Fish Half-Life (days), BCF Arnot-Gobas (upper trophic) Including Biotransformation (L/kg wet-wt), BCF Arnot-Gobas (upper trophic) Zero Biotransformation (L/kg wet-wt), BAF Arnot-Gobas (upper trophic) Including Biotransformation (L/kg wet-wt), BAF Arnot-Gobas (upper trophic) Zero Biotransformation (L/kg wet-wt)	EPI BCFBAF v3.01
LC50 (Fish) or EC50 (Daphnid and Algae) for Most Toxic Class (mg/L), Max. Log Kow for Most Toxic Class, Most Toxic Class	EPI ECOSAR v1.11
Lipinski's Rule-of-five (bioavailability), Absorption from gastrointestinal tract for 1 mg dose (%), Absorption from gastrointestinal tract for 1000 mg dose (%), Log brain/blood partition coefficient	Equations from literature
Dermal absorption (mg/cm2/event)	EPI DERMWIN v2.02

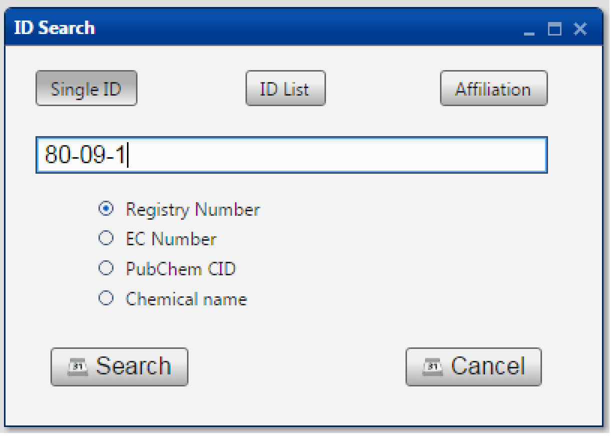
Predicted property	Software
Acute toxicity in rodents: Rat Oral, Rat Intraperitoneal, Mouse Oral, Mouse Intraperitoneal, Mouse Intravenous, Mouse Subcutaneous	ACD/ ToxSuite 2.95.1

Appendix 1: Searching by identification number

Example 1: Single ID query.

Search for registry number 80-09-1.

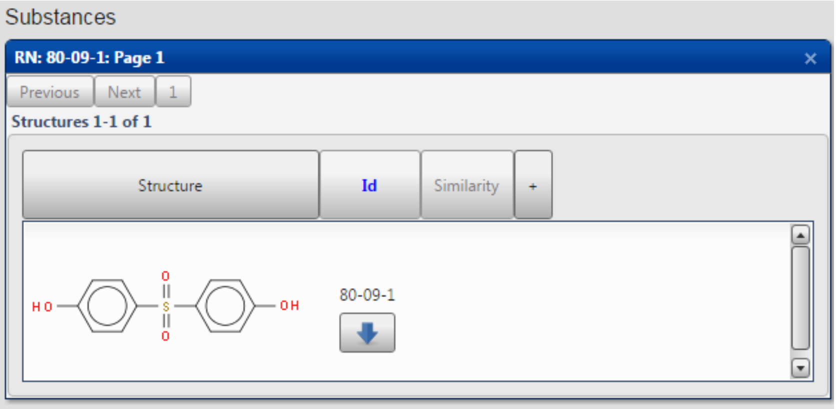
Start by clicking the ***Id*** button to the right on the main search screen. The ID Search box will appear on the screen. Select Single ID and Registry number from the list in the ID Search box. Once you've done this, type in the registry number with or without hyphens in the input box:



80-09-1

● Registry Number
○ EC Number
○ PubChem CID
○ Chemical name

To start the search, click the ***Search*** button. You will now see the result to the right on the screen under "Substances":



Structures 1-1 of 1

Structure	Id	Similarity	+
	80-09-1		

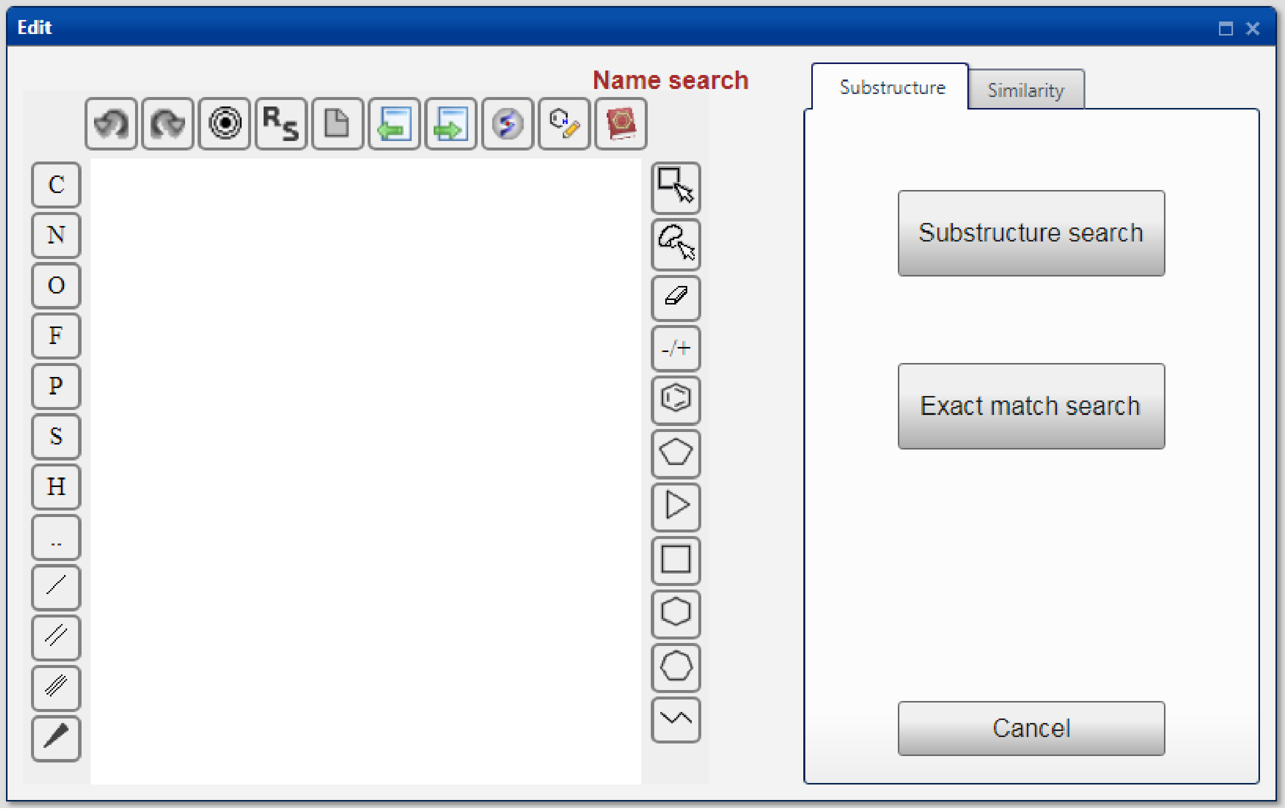
To download a single substance report, click the ⬇ button in the id column next to the substance of interest.

Appendix 2: Searching by structure and similarity

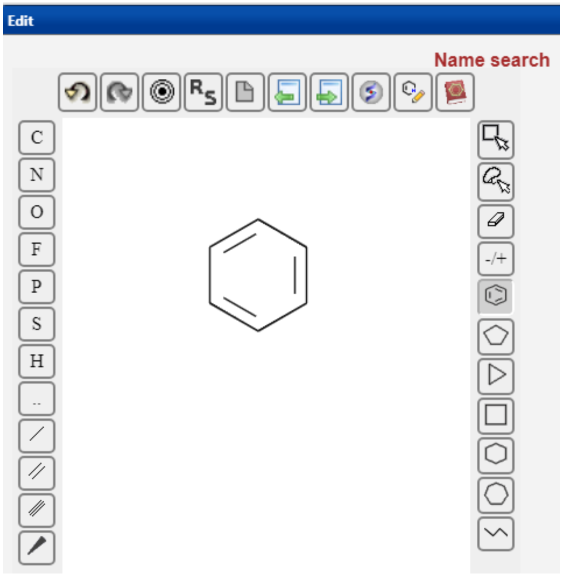
Example 1: Substructure search.

Search for molecules containing a fluorobenzene fragment.

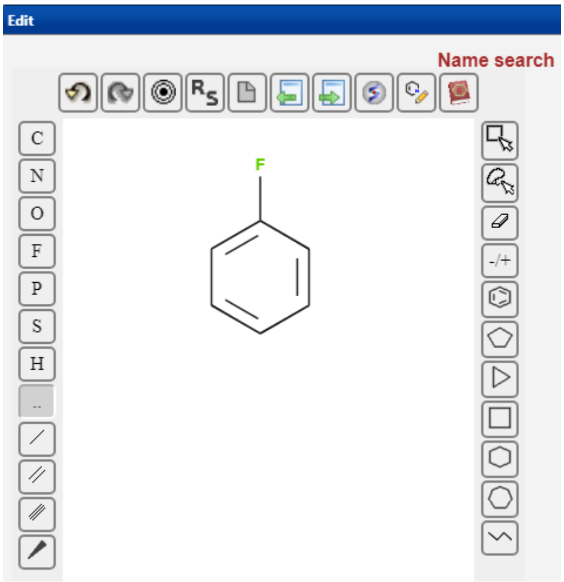
It is possible to search for molecules which contain specific molecular fragments. Start by clicking the ***Structure*** button to the left on the main search screen. This will open the molecular editor shown below:



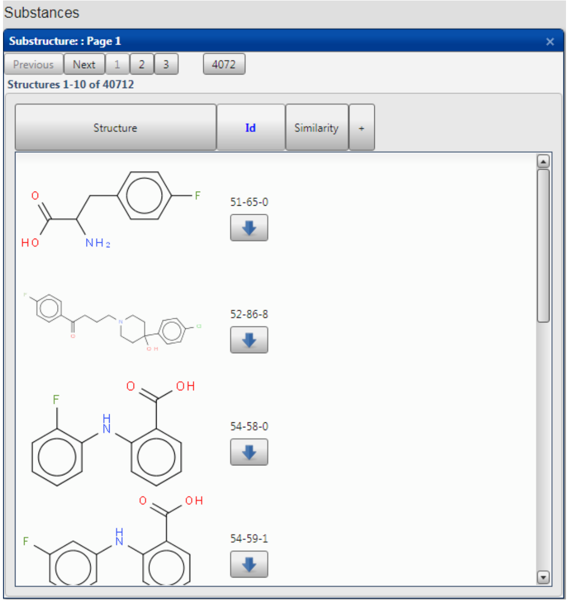
In this example we will use the drawing tools in the molecular editor to create the fluorobenzene fragment. Alternatively, the fragment can be imported as a SMILES/MOL file or from the chemical dictionary. When drawing a fragment, the different atoms, bonds etc. are added one by one. Start by adding the benzene part of the molecule: left-click on the ***benzene*** button and then click on the blank canvas. The canvas will look like this:



Now click on the single bond button and then click on one of the atoms in the benzene ring on the canvas. Then add the fluorine (F) atom. The fluorobenzene fragment is now completed:



To start the search, click the ***Substructure search*** button under the **Substructure** heading to the right in the molecular editor. The molecules that contain the fluorobenzene fragment will be shown to the right on the screen:



The total number of molecules containing the fluorobenzene fragment are shown under the heading “Results”. The functionalities of the results window are described in detail in the section “Results window with substances” in the manual.

Example 2: Exact match search.

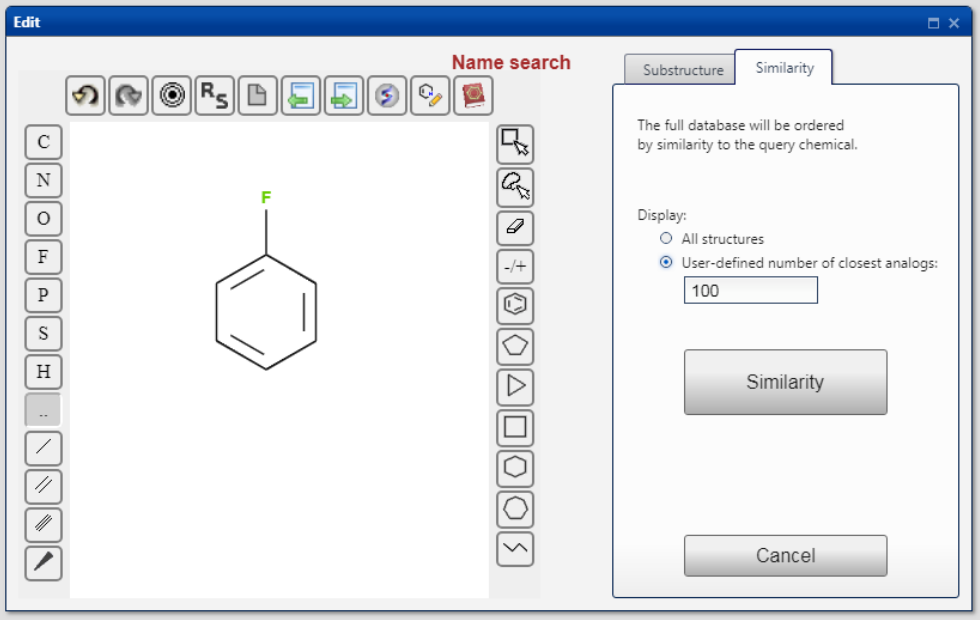
Find molecules exactly matching the fluorobenzene structure.

Start by drawing the fluorobenzene structure as described in example 1. To start the search, click the ***Exact match search*** button under the **Substructure** heading to the right in the molecular editor. You will now see the search result in the window under “Substances” to the right on the screen.

Example 3: Similarity search.

Find the chemicals that are most similar to fluorobenzene.

Start by drawing the fluorobenzene structure as described in example 1. Then click the heading **Similarity** to the right in the molecular editor. Before you start the search you need to select if all structures or a user-defined number (e.g. 100) should be displayed. In this example the 100 closest analogs is selected:



The full database will be ordered by similarity to the query chemical.

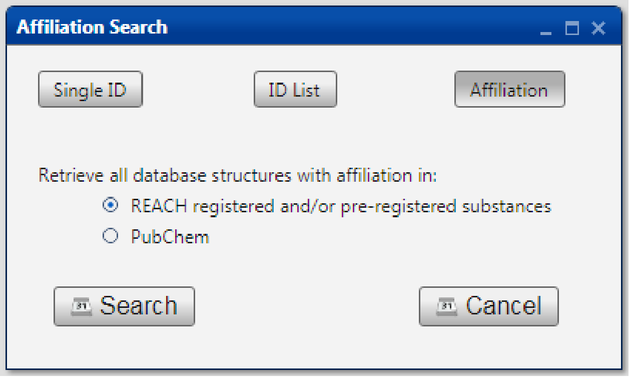
Display:
○ All structures
● User-defined number of closest analogs: 100

To start the search, click the ***Similarity*** button. The resulting substances will be ordered by similarity to the query chemical.

Example 4: Similarity search.

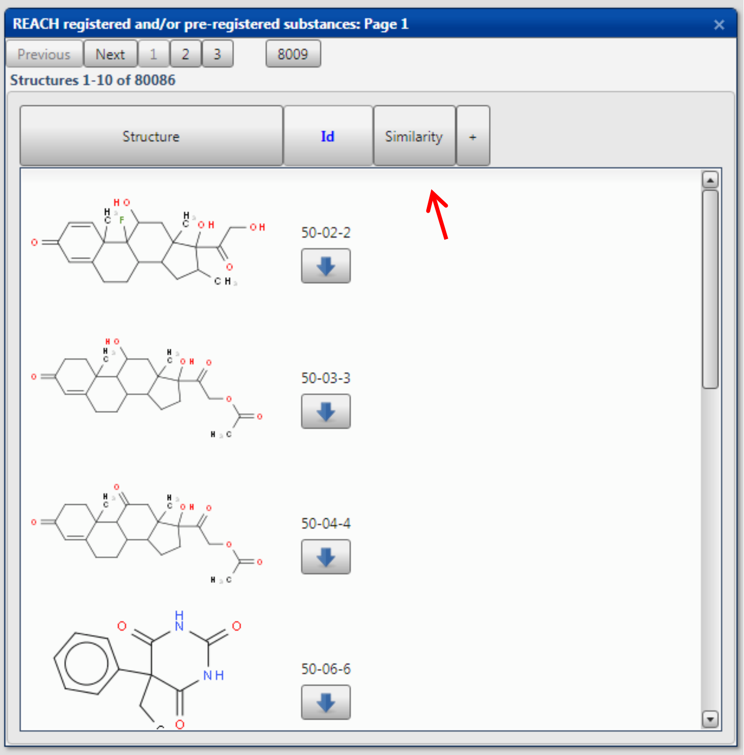
Find the REACH chemicals that are most similar to fluorobenzene.

Start by searching the REACH chemicals in the database: Click the ***Id*** button to the left on the main search screen and then click the heading **Affiliation** in the ID search box. The box should now look like this:



Retrieve all database structures with affiliation in:
● REACH registered and/or pre-registered substances
○ PubChem

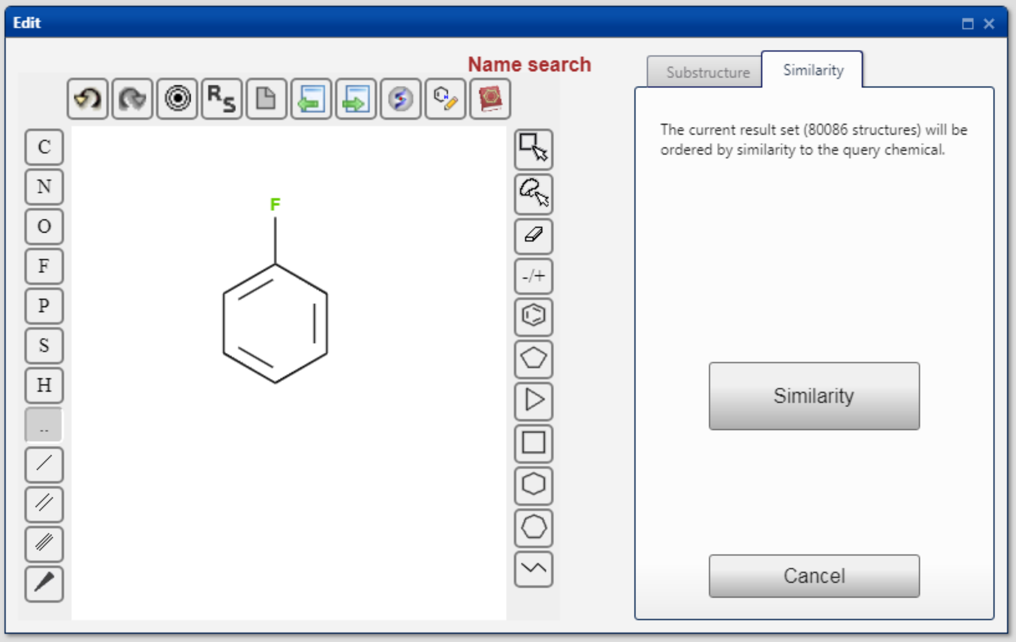
Select “REACH pre-registration list” and then click ***Search***. This will generate a results window showing all REACH pre-registered chemicals:



REACH registered and/or pre-registered substances: Page 1

Structures 1-10 of 80086

Now you can search for similarity within the REACH pre-registration list. Click the ***Similarity*** button marked with a red arrow in the results window above. At this point, a new box will open:



The current result set (80086 structures) will be ordered by similarity to the query chemical.

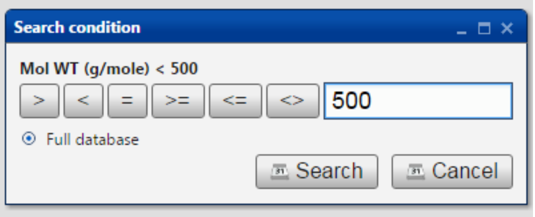
Draw the fluorobenzene structure (described in example 1) on the blank canvas and click the ***Similarity*** button to start the search. The REACH substances will now be ordered by similarity to the query chemical.

Appendix 3: Searching by model endpoint

Example 1: Simple model endpoint search.

Search for molecules with molecular weight < 500 g/mole.

Click the ***PhysChem*** button to the left on the main search screen and then select Mol WT (g/mole) from the drop-down menu:



Search condition

Mol WT (g/mole) < 500

500

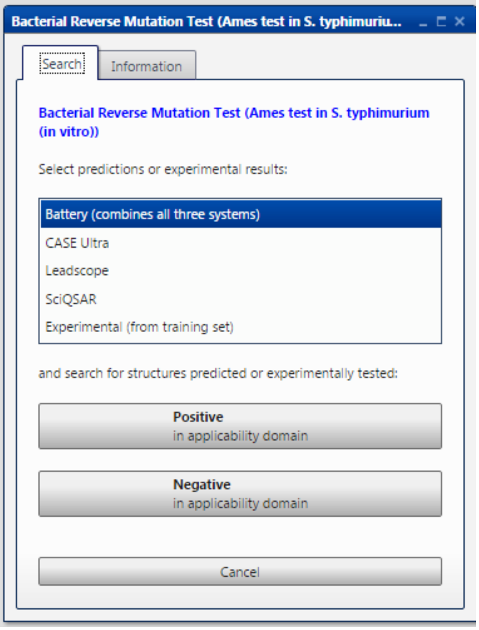
● Full database

Type in 500 in the blank field in the box and click the ***Search*** button. The substances of interest will appear in the results window.

Example 2: Simple model endpoint search.

Search for molecules that are positive for Ames test.

Click the ***Human health*** button to the left on the main search screen and then select "Genotoxicity - Ames test - Bacterial Reverse Mutation Test" from the drop-down menu. The following dialogue box should now appear:



Bacterial Reverse Mutation Test (Ames test in S. typhimuriu...

Bacterial Reverse Mutation Test (Ames test in S. typhimurium (in vitro))

Select predictions or experimental results:

- Battery (combines all three systems)
- CASE Ultra
- Leadscope
- SciQSAR
- Experimental (from training set)

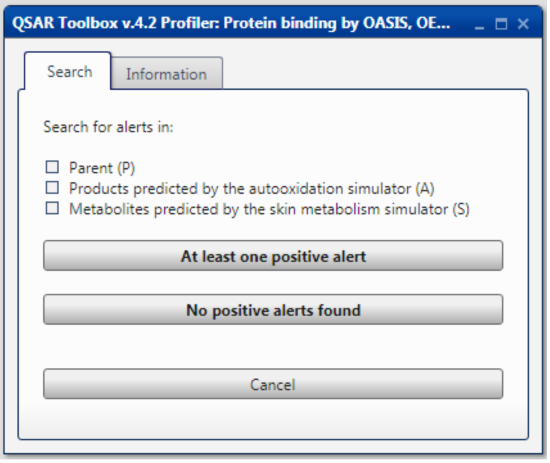
and search for structures predicted or experimentally tested:

The Ames test model is made in each of the three systems CASE Ultra, Leadscope and SCIQSAR and in this example we will select the overall battery prediction. Once the 'Battery' is selected, click the ***Positive*** button to start searching for molecules that are positive for Ames test. The results will appear on the screen.

Example 3: Simple model endpoint search.

Search for molecules where the QSAR Toolbox predicts alerts for protein binding by the profiler “Protein binding by OASIS”.

Click the ***Human health*** button to the left on the main search screen and then select “Irritation and Sensitisation – Profilers – Protein binding by OASIS, OECD QSAR Toolbox v.4.2” from the drop-down menu. The following dialogue box should now appear:


QSAR Toolbox v.4.2 Profiler: Protein binding by OASIS, OE...
Search
Information
At least one positive alert
No positive alerts found
Cancel

Search for alerts in:

□ Parent (P)
□ Products predicted by the autooxidation simulator (A)
□ Metabolites predicted by the skin metabolism simulator (S)

The database can be searched to find molecules where at least one alert or where no alerts were predicted in the parent structure or in transformation products predicted by Toolbox simulators. Tick of the small boxes depending of your choice: “Parent”, “Products predicted by the autooxidation simulator” and/or “Metabolites predicted by the skin metabolism simulator”. Here we will search for molecules for which positive alerts were predicted in transformation products predicted by the included two simulators but not necessarily in the parent, so we tick of the two boxes for autooxidation products and skin metabolites and leave the box for parent unchecked. Click the ***At least one positive alert*** button to start searching for molecules that are predicted to contain positive alerts for protein binding. The results will appear on the screen.

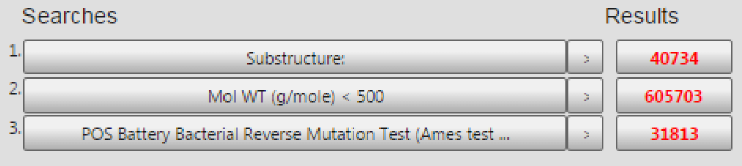
Appendix 4: Combining searches

Example 1: Complex search containing AND.

Search for molecules that have:

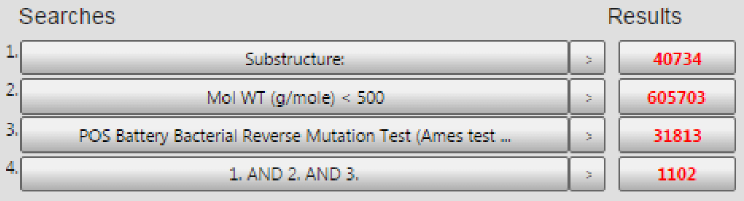
- a fluorobenzene fragment
- molecular weight < 500 g/mole
- a positive Ames test

We have already made searches for each of the three criteria in the previous examples (Appendix 2, example 1; Appendix 3, example 1 and 2). When the three individual searches are made one by one, the screen should look like this:



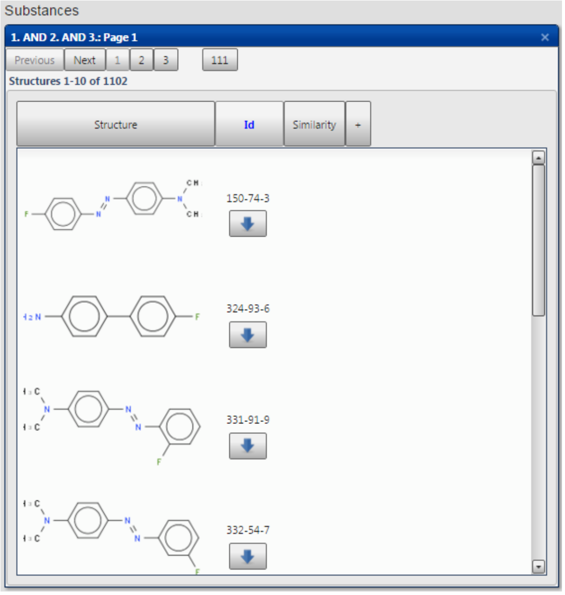
	Searches		Results
1.	Substructure:	>	40734
2.	Mol WT (g/mole) < 500	>	605703
3.	POS Battery Bacterial Reverse Mutation Test (Ames test ...	>	31813

To search for molecules that meet all three criteria, left-click on the three search definition buttons. This will change their state to "selected" and the foreground color to green. Now click on the ***AND*** button to the right on the main search screen to start the search. At this point, you should see this on the screen:



	Searches		Results
1.	Substructure:	>	40734
2.	Mol WT (g/mole) < 500	>	605703
3.	POS Battery Bacterial Reverse Mutation Test (Ames test ...	>	31813
4.	1. AND 2. AND 3.	>	1102

The molecules that meet all three criteria will be shown in the window under "Substances" to the right on the screen:



Substances

1. AND 2. AND 3.: Page 1

Structures 1-10 of 1102

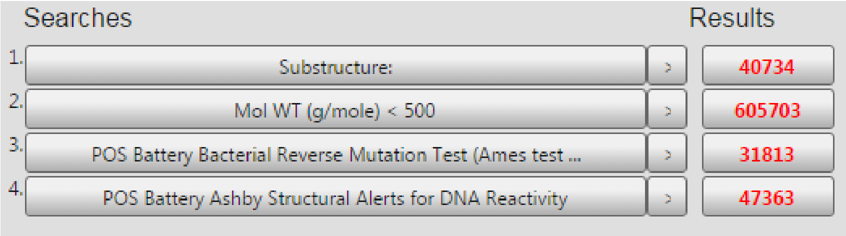
Structure	Id	Similarity
	150-74-3	
	324-93-6	
	331-91-9	
	332-54-7	

Example 2: Complex search containing AND, OR and NOT.

Search for molecules that:

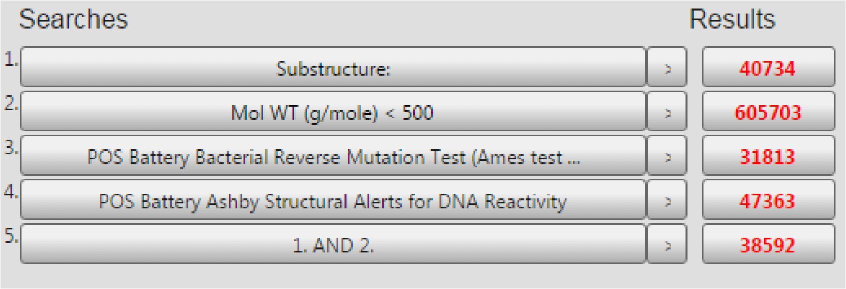
- have a fluorobenzene fragment AND molecular weight < 500 g/mole
- are NOT positive for Ames test OR Ashby alerts

You can use the individual searches from example 1, but you will have to add a few extra searches to find the molecules that fulfill all the criteria above. First, you may add the chemicals that are positive for Ashby alerts. Click the ***Human health*** button to the left on the main search screen and then select “Genotoxicity – Ashby Structural Alerts for DNA Reactivity” from the drop-down menu. Select ‘Battery’ and click the ***Positive*** button to start the search. At this point, the “Searches” on the screen should contain the following search definition buttons:



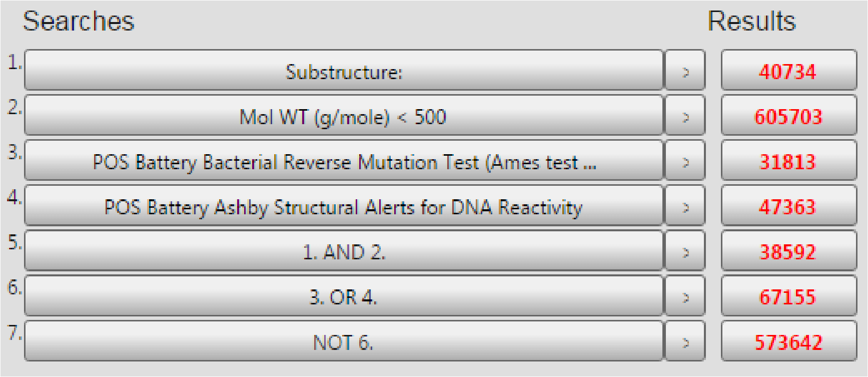
	Searches		Results
1.	Substructure:	>	40734
2.	Mol WT (g/mole) < 500	>	605703
3.	POS Battery Bacterial Reverse Mutation Test (Ames test ...	>	31813
4.	POS Battery Ashby Structural Alerts for DNA Reactivity	>	47363

Now you need to find the molecules that have both a fluorobenzene fragment and MW < 500 g/mole. To do this, click and highlight search definition buttons 1 and 2, then click the AND button. You will now see the resulting molecules on the screen and a new search definition button will appear:



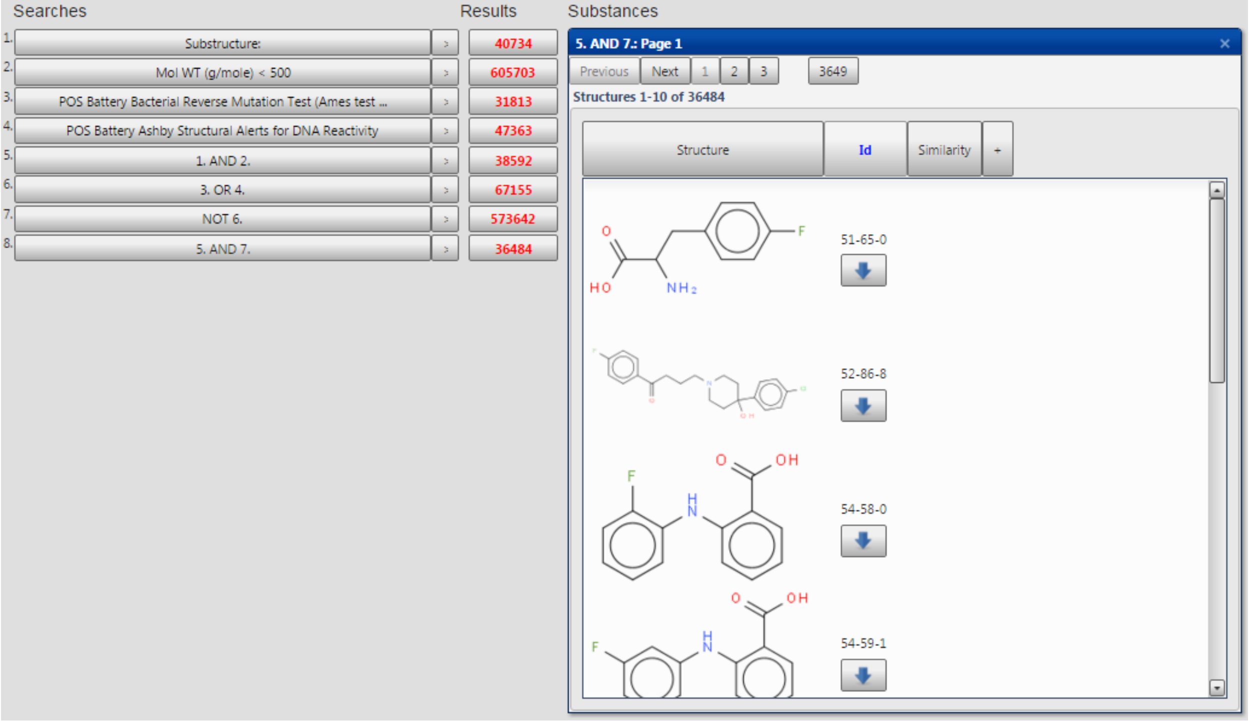
Searches		Results
1. Substructure:	>	40734
2. Mol WT (g/mole) < 500	>	605703
3. POS Battery Bacterial Reverse Mutation Test (Ames test ...	>	31813
4. POS Battery Ashby Structural Alerts for DNA Reactivity	>	47363
5. 1. AND 2.	>	38592

Next, you need to find the molecules that are not positive for either Ames test or Ashby alerts. This procedure is performed in two steps. Click on search definition buttons 3 and 4, then click the OR button and a new search definition button appears. Click this search definition button and then click the NOT button. You should now see the following search definition buttons on the screen:



Searches		Results
1. Substructure:	>	40734
2. Mol WT (g/mole) < 500	>	605703
3. POS Battery Bacterial Reverse Mutation Test (Ames test ...	>	31813
4. POS Battery Ashby Structural Alerts for DNA Reactivity	>	47363
5. 1. AND 2.	>	38592
6. 3. OR 4.	>	67155
7. NOT 6.	>	573642

You have now added all the searches required to find the molecules that fulfill all criteria listed in this example. The last step is to click on search definition buttons 5 and 7, then click the AND button. Once this is done, the molecules of interest are shown in the result window:



Searches		Results
1. Substructure:	>	40734
2. Mol WT (g/mole) < 500	>	605703
3. POS Battery Bacterial Reverse Mutation Test (Ames test ...	>	31813
4. POS Battery Ashby Structural Alerts for DNA Reactivity	>	47363
5. 1. AND 2.	>	38592
6. 3. OR 4.	>	67155
7. NOT 6.	>	573642
8. 5. AND 7.	>	36484

Substances

5. AND 7.: Page 1

Structures 1-10 of 36484

Appendix 5: Software systems used for modeling

More detailed information about the software systems described below can be found in the QMRFs.

Case Ultra

CASE Ultra is a fragment-based statistical model system. The methodology involves breaking down the structures of the training set into all possible fragments from 2 to 10 heavy (non-hydrogen) atoms in length. The fragment generation procedure produces simple linear chains of varying lengths and branched fragments as well as complex substructures generated by combining the simple fragments. A structural fragment is considered as a positive alert if it has a statistical significant association with chemicals in the active category. It is considered a deactivating alert if it has a statistically significant relation with the inactive category. Once final lists of positive and deactivating alerts are identified, CASE Ultra attempts to build local (Q)SARs for each alert in order to explain the variation in activity within the training set chemicals covered by that alert. The program calculates multiple molecular descriptors from the chemical structure such as molecular orbital energies and two-dimensional distance descriptors. A stepwise regression method is used to build the local (Q)SARs based on these molecular descriptors. For each step a new descriptor (modulator) is added if the addition is statistically significant and increases the cross-validated R2 (the internal performance) of the model. The number of descriptors in each local model is never allowed to exceed one fifth of the number of training set chemicals covered by that alert. If the final regression model for the alert does not satisfy certain criteria (R2 ≥ 0.6 and Q2 ≥ 0.5) it is rejected. Therefore, not all alerts will necessarily have a local (Q)SAR. The collection of positive and deactivating alerts with or without a local (Q)SAR constitutes a global (Q)SAR model for a particular endpoint and can be used for predicting the activity of a test chemical.

Leadscope

Leadscope Predictive Data Miner is a software program for systematic sub-structural analysis of a chemical using predefined structural features stored in a template library, training set-dependent generated structural features (scaffolds) and calculated molecular descriptors. The feature library contains approximately 27,000 pre-defined structural features such as functional groups, heterocycles and pharmacophores. The training set-dependent structural features (scaffold generation) can be added to the pre-defined structural features from the library and be included in the descriptor selection process. The program also calculates a number of physico-chemical descriptors such as logP, molecular weight and the number of hydrogen bond acceptors and donors. Leadscope has a default automatic descriptor selection procedure. This procedure selects the top 30% of the descriptors (structural features and molecular descriptors) according to X2-test for a binary variable or the top and bottom 15% descriptors according to t-test for a continuous variable. After selection of descriptors the program performs partial least squares (PLS) regression for a continuous response variable, or partial logistic regression (PLR) for a binary response variable, to build a predictive model.

SciQSAR

The SciQSAR software provides over 400 built-in molecular descriptors such as connectivity indices, electrotopological (atom E and HE-state) indices, and other descriptors. Furthermore, the program provides a variety of statistical tools that can be used to build predictive models for binary and continuous data. SciQSAR uses discriminant analysis for binary data and includes the capability to perform parametric and nonparametric discriminant analyses. For continuous data, regression analysis is used to build the predictive model, and a number of different regression methods are available such as regression on principal components (PCR) and partial least squares regression (PLS).

Appendix 6: A short introduction to QSARs

Structure-activity relationships (SARs) and quantitative structure-activity relationships (QSARs), collectively referred to as QSARs, are mathematical models that can be used to predict the physicochemical, biological (e.g. toxicological) and environmental fate properties of molecules based on their chemical structure. A QSAR is a mathematical model (often a statistical correlation) relating one or more quantitative parameters derived from chemical structure to a quantitative measure of a property or activity. QSARs are quantitative models that yield either a continuous or categorical (yes/no) result.

A QSAR model thus links information on the chemical structure of compounds with a specific property, and is subsequently used for predicting the same property for unknown compounds. Reliable predictions can be obtained for compounds that are within the domain of the developed QSAR model, i.e. for compounds that are sufficiently structurally similar to the compounds used to train the model. QSAR models are powerful tools for predicting chemically induced adverse effects and thus for filling data gaps. The reliability of QSAR predictions depends on numerous parameters relating to the mathematical methods used, the number and precision of the underlying data used for developing the model and how suitable the model is for the particular substance. In general the uncertainty of QSARs is caused predominantly by two different reasons: a) the inherent variability of the input data used to establish the model (training set); and b) the uncertainty resulting from the fact that a model can only be a partial representation of reality. However, as a model averages the uncertainty over all substances, it is possible for an individual model estimate to be more accurate than an individual measurement.

Validation is a trial of the model performance for a set of substances independent of the training set, but within the domain of the model. The model predictions for these substances are compared with measured endpoints for the substances in order to establish the predictive performance of the model. Ideally all models should be assessed by checking how well they predict the activity of substances, which were not used to make them. This is, however, not always simple. In part valuable information may be left out by setting aside substances to be used in such an evaluation, and in part it can be extremely difficult to assess how "external" substances relate to the model's domain; that is, if they represent a random distribution within this applicability domain and thereby giving a fair picture of the predictive performance of the model.

This problem is often addressed by using one or another form of cross-validation, where a number of partial models are "externally validated" by dividing the training set into a reduced training set and a testing set. The reduced training set is used to develop a partial model, while the remaining data are used as a test set to evaluate the model predictivity. This is repeated a number of times and the results are used to calculate the predictivity measures for the models; for quantitative (continuous) models in the form of Q2 and SDEP (standard deviation error of prediction), and for qualitative (categorical yes/no) models in the form of sensitivity (ability to correctly positives), specificity (ability to correctly predict negatives) and concordance (overall accuracy). In the majority of validations carried out on the models applied in this database the stable leave-many-out (LMO) cross-validation approach was used. The training set was split by random (however keeping the positive / negative balance in the subsets) into two portions of 50% of the substances, models on each of the reduced sets were made, and the one model was run to predict the training set of the other model and the other way around, repeating this 5 times. Leaving out 50% of the substances in the partial validation models is a large perturbation of the training set, which generally leads to realistic, and often pessimistic, measures of the predictivity of the model.

Concordance will vary depending on both the method used, and the endpoint in question. In general, contemporary QSAR systems can often correctly predict the activity of about 70 – 85% of the substances examined, provided that the query structures are within the domains of the models.

When applying QSAR's it is important to assure that an obtained prediction falls within the applicability domain (AD) of the models i.e., that there is sufficient similarity (in relevant descriptors) between the query substance and substances in the training set of the model. There is no single and absolute applicability domain for a given model. Generally, the broader the applicability domain is defined the lower predictivity can be expected. The applicability domain should be clearly defined and the validation results should correspond to this defined domain, which is again used when the model is applied for predictions.