Deliverable D1.2

GUI Design

Grant Agreement	Health-F5-2008-200787
Acronym	OpenTox
Name	An Open Source Predictive Toxicology Framework
Coordinator	Douglas Connect

Contract No.	Health-F5-2008-200787	
Document Type:	Deliverable Report Master	
WP/Task:	WP1 / D1.2	
Name	GUI Design	
Document ID:	OpenTox Deliverable Report WP1-D1.2	
Date:	Aug 31, 2009	
Status:	Final Version	
Organisation:	David Gallagher (DG)	
Contributors	David Gallagher (author) Barry Hardy, Nicki Douglas (review)	DG DC

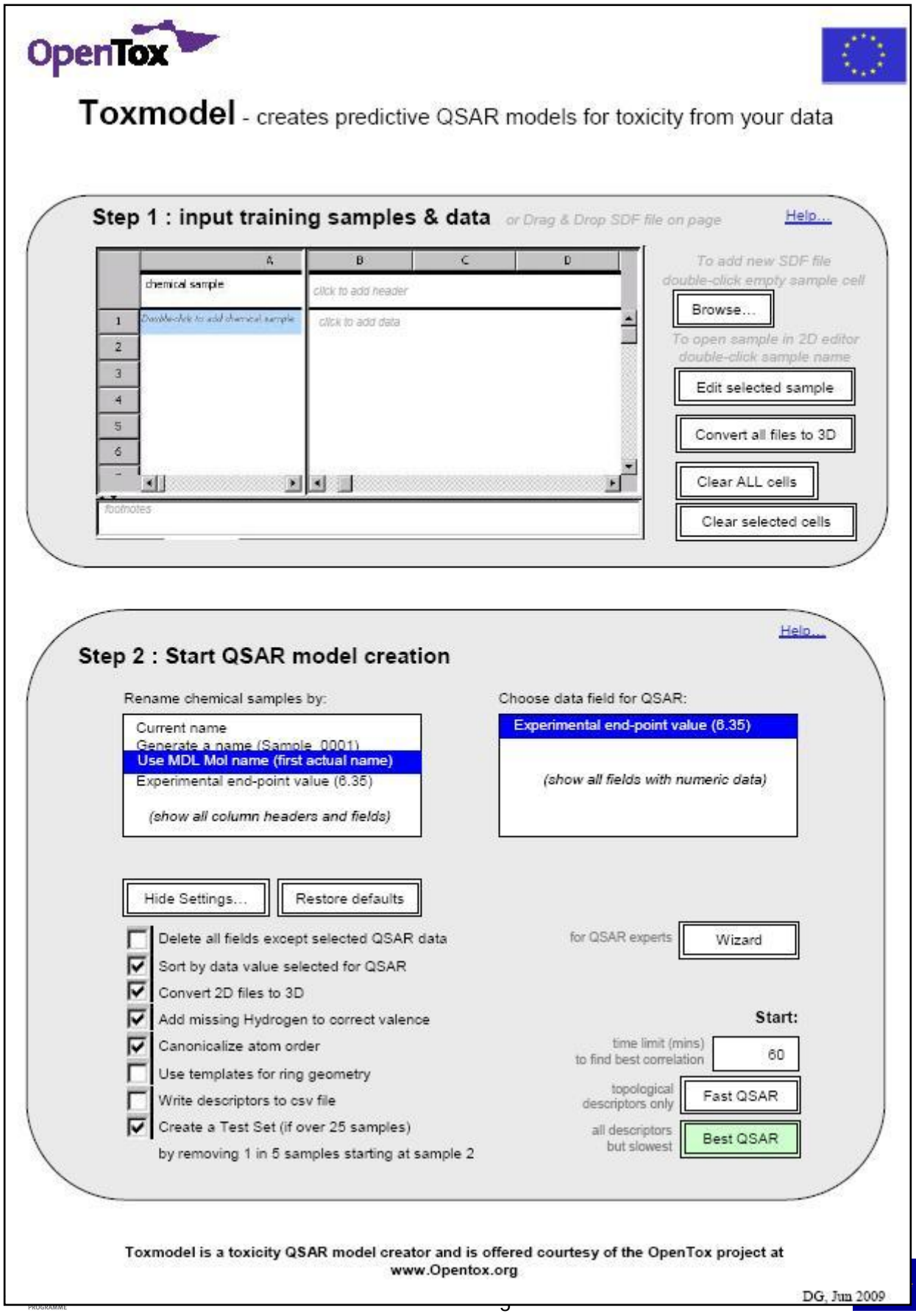
Distribution:	Public

Purpose of Document:	To document results for this deliverable

Document History:	1 – Initial Draft by David Gallagher, August 7, 2009
	2 – Updated report v1, August 14, 2009
	3 – Completed report, August 31 2009
	4 – Updated version with updated links, 24 February 2010

Table of Contents

Figure 1, Draft layout for GUI interface



Summary

This document describes a proposed graphical user interface for the OpenTox suite of tools to enable non-QSAR-experts to easily use OpenTox to develop new QSAR models to predict toxicity.

1. Graphical User Interface (GUI) for creating new toxicity models

QSAR methods require a significant amount of experience and understanding of the underlying methodology, as well as skilful interpretation of the results in order to be able to get reliable and meaningful results. Hence, this has restricted the practical use of QSAR methods to a limited number of experts. The complexity and potential pitfalls of QSAR presents a significant barrier to the use of these methods for scientists such as Toxicologists who could most benefit from the increased use of the tools.

In order to extend the usefulness of QSAR to scientists without QSAR experience, OpenTox plans to create a GUI that incorporates 'default' settings and artificial intelligence based on the experience and skills of the partners in the OpenTox project and publications from other experts in the field.

1.1 Description

This document describes a graphical user interface (GUI) "Toxmodel" designed primarily for the 'novice' with little or no knowledge of QSAR or computational chemistry, to automate most of the process of building predictive toxicity models using the OpenTox toxicity prediction software package. Once an acceptable predictive model has been created, it can be saved and used via the "Fastox" GUI to make predictions.

The software package can be installed as a complete stand-alone application for Windows and Linux, or a web-based version can be accessed without any local installation using a browser interface via the internet (for example, hosted by http://www.opentox.org/toxicity-prediction or linked server). Toxmodel automatically checks the input data, calculates a wide range of descriptors and applies a wide range of regression techniques to find the best correlation within a set time period. Then a detailed report on the performance and limitations of the best few models is created. The model can be saved for use by Fastox to make predictions.

1.2 Assumptions

The user needs to make toxicity predictions for a novel class of compounds that are not well represented in available databases, and he has a minimum of 6 representative compounds with experimental data with which to create a predictive model.

The user has an SDF file containing all the structures and associated experimental data or, the user has a number of chemical structure files in a common format (e.g. SDF, PDF, MOL, etc.) and can manually enter the associated experimental end-point data.

The user is able to use a browser such as Internet Explorer and has access to the web page for "ToxModel", or the stand-alone version of ToxModel.

Procedure for using "Toxmodel" GUI

See the graphical layout design (Figure 1)

1.3 Section: "Step 1", entering the chemical sample files and data.

1. User navigates to the "Toxmodel" page (e.g. www.Toxmodel.com) or opens the Toxmodel application (if installed locally). It may be appropriate to access the web page via a login so that preferences and results can be saved.

2. User then enters structure as one or more SD files or other supported file formats:

i) **Browse**: the user double-clicks in an empty sample cell in the spreadsheet to open a browser dialogue box to select one or more SDF files (or SD, Mol, SMILES, Cartesian & XYZ files) on the local computer or network. After clicking to select one or more files, the "Open" button opens the file which is then loaded into the spreadsheet, one sample per sample cell. All associated data is placed in the same row, under the appropriate header which is created automatically.

ii) **Drag and Drop** an SDF file anywhere on the page (SD, Mol, SMILES, Cartesian & XYZ files). The file is then loaded into the spreadsheet as above.

(this may be too complex for first implementation, so may be deferred)

iii) If the structure files do not include numerical data for the QSAR, the user can enter this manually after the structure files have been loaded. The user can click in an empty header cell to position the cursor to type a header name. Data values can be entered into the cells below, by placing the cursor in each cell in order to type the value. The "Return" key completes the data entry in that cell and moves the cursor into the next cell below.

1.3.1 File types

ToxModel will automatically recognize the following file types: SDF, SD, Mol, PDB, SMILES. Files containing a batch of structures must be in SDF format.

1.3.2 Clicking in the spreadsheet cells

Clicking in a chemical sample cell (under column A) highlights that cell.
Clicking in any other cell (not under column A) places a text cursor, allowing the user to type data or copy and paste into that cell. Press and drag allows selection of a range of cells

1.3.3 Browsing to chemical sample files

Double clicking on an empty chemical sample cell opens the "browse to…" dialog window so the user can navigate to and select a file to load into that cell. If more than one file is selected, the additional files are added alphabetically and sequentially in the sample cells below the first selected one.

Alternatively, the "Browse to…" dialog window can be opened by selecting the empty sample cell, then clicking the "Browse…" button

1.3.4 Editing chemical samples

Double-clicking on a sample cell with a file name in it, opens that structure in a 2D editor which can be toggled between 2D and 3D views with a simultaneous 1D (text) editor. Both windows are interactive (edits in one window automatically update the other immediately) and editing is allowed in both (see FasTox). When the 3D view is selected, an option button appears allowing the user to automatically "Convert to 3D".

Alternatively, the editor can be opened by selecting the sample name, then clicking "Edit selected sample"

If an empty chemical sample cell is selected, clicking "Edit selected sample"
opens an empty editor window allowing the user to draw, name, and save a new structure into that cell. If an existing file is edited, the original file cannot be overwritten, but is saved with an incremented suffix number.

1.3.5 Deleting chemical samples

Chemical samples can be deleted by clicking the name to select the cell, then clicking the "Clear selected cells" button. If there are associated fields of data, a dialog box asks the user if the additional data should be deleted or retained. "Clear all cells" clears and resets the whole spreadsheet. The "Back" arrow on the browser allows the user to undo the "Clear all cells" action.

1.3.6 Checking the input file for errors

While loading, after the file is checked for errors, various alerts appear in a dialog box:

"Multiple molecules in sample #, non-relevant molecules should be deleted…

"incorrect valence on atoms: etc.

"unrecognized atom type:

"Structure has a non-zero net charge of ? (ion)

"Structure for CAS number ??? not found

"File could not be read due to unrecognized format

"Etc.

Critical errors invoke an error dialog box describing the error and sample number (if appropriate). Non-critical errors do not prevent the sample from being loaded, but are listed as footnotes at the bottom of the spreadsheet and an associated superscript added to the sample name.

Examples of non-critical 3D* errors:

Bond distance for atoms X and Y is outside normal range

Atoms X and Y are too close

Incorrect valence on atoms: C26, C28, etc.

Total charge is non-zero charge

more than one molecule in file

unrecognized element present

incorrect valence on one or more atoms

etc.

The error report dialog box has a button to "Edit…", "Cancel" and for some errors a "Ignore and continue" button will be active. The "Edit…" button open a 2D/3D and simultaneous 1D (text) editor. Both windows are interactive (edits in one window automatically update the other immediately) and editing is allowed in both (see Fastox).

Non-critical errors are listed as footnotes at the bottom of the spreadsheet and an associated superscript added to the sample name. Critical errors invoke an error dialog box describing the error and sample number (if appropriate).

"Step 2", Start QSAR model creation

This section explains the various fields in "Step 2" of Figure 1

1.4 "Step 2", Start QSAR model creation

1.4.1 Rename chemical samples by:

This field show a list of all fields in the SDF file. The selected field will be used as the sample name in the final results spreadsheet. The default (selected) setting is "Use MDL MOL name".

1.4.2 Choose data field for QSAR:

A list of the data fields that contain numeric data is show in this window. The selected field will be used to correlate with the descriptors generated.

1.4.3 Hide Settings / Show settings

On the left side a list of check-box options can be displayed or hidden by clicking to toggle the "Hide settings" or "Show settings" button. By default all check boxes are hidden. "Reset defaults" sets all check boxes to the default settings (as shown).

***3D structures**

Assuming that the 3D conversion is fast and that OpenTox would eventually include some 3D QSARs based on QM methods or 3D-substructure searching, then this option should be checked by default. The extra conversion time would be minor compared to the descriptor generation and regression analysis time.

1.4.4 Restore Defaults

Sets all check boxes back to default values

1.4.5 Wizard

The Wizard takes the user through a series of steps to select various options such as descriptors and regression criteria etc., before running the QSAR model builder (not yet designed).

1.4.6 Time Limit

Time limit sets the maximum time that the system will continue to look for the best correlation. If the time limit is reached before the calculations are finished, then calculations will be stopped and the result will be presented showing the best correlation(s) found in the time allowed.

1.4.7 Fast QSAR

To save time, only 'linear' topological descriptors are computed and used in the regression

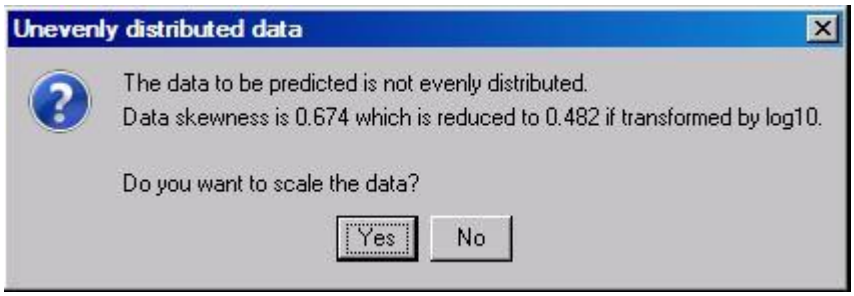
1.4.8 Best QSAR

All descriptors, including their non-linear functions, are computed and used for the regression analysis.

"Step 3" automatically computing the new QSAR Model

1.4.9 Data analysis:

On starting the QSAR analysis, the training data values are analyzed for evenness of spread. If the "skew" value does not fall below a preset threshold, then various functions are tested (e.g. reciprocal, square, square root, logarithm, etc.). The new skew values are computed and the user is presented with a warning "The data to be predicted is not evenly spread and could lead to a misleadingly high r^2", and a list of functions that improve the skew value (best first) and has the option to select one or continue with the original linear data.



1.4.10 Descriptors:

For "Best QSAR", all descriptors are calculated by default including the following:

1. Topological descriptors including atom and group counts and Kier and Hall indices
2. Electrotopological descriptors
3. Quantum descriptors (MOPAC PM6 recommended for speed)
4. "Density descriptors" of most meaningful descriptors (above), e.g. Number of hydroxyl groups divided by molecular weight (or /volume or /surface area).
5. Non-linear functions of all above descriptors, e.g. reciprocal, square, square root, logarithm

"Quick QSAR" uses only linear topological, Electrotopological descriptors and "Density descriptors".

Descriptors with variance below a preset threshold are removed from the set.

1.4.11 Regression analysis, systematic

Regression analysis uses all computed descriptors and starts by systematically looking for and saving the best ten single-descriptor-models (based on highest r^2). Then it looks for the best two-descriptor-models by comparing all permutations of two descriptors, and saves the best 10 models. Then it continues by looking for the best three-descriptor-models by comparing all permutations of three-descriptor models. Finally it looks for the best four-descriptor model.

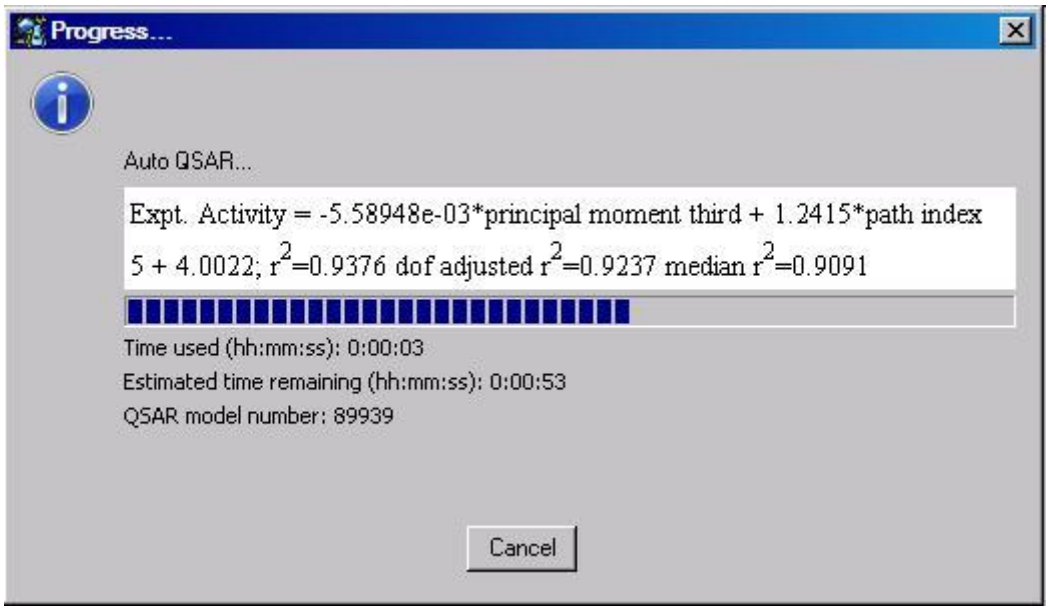
However, to avoid over-fitting, the number of descriptors in the model is not allowed to exceed the number of training samples divided by 5.

1.4.12 Other analyses:

After applying the systematic regression analysis, other methods such as genetic algorithms, neural nets are tried next, until the preset time limit is reached.

1.4.13 Stopping the calculation

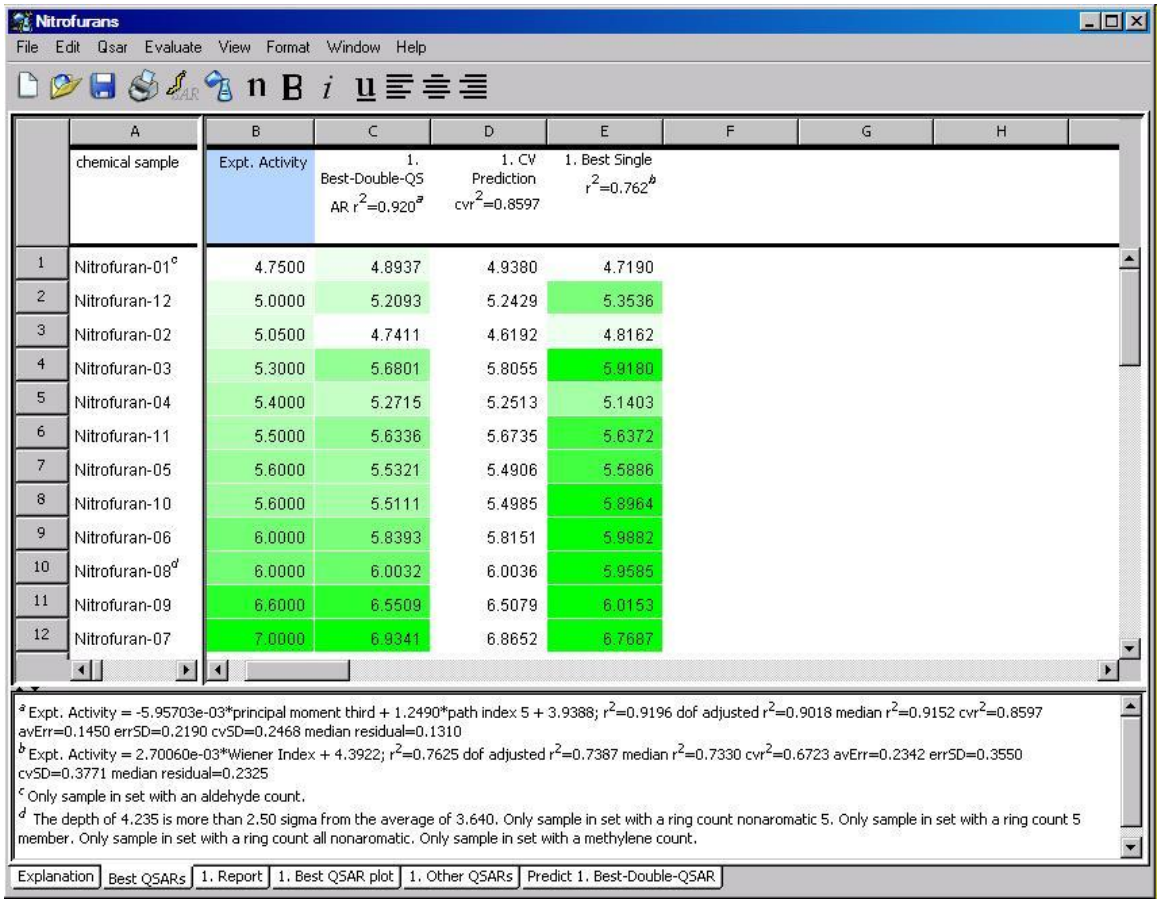
The best QSAR (highest r^2 adjusted for number of degrees of freedom) found to date is always displayed on the screen during the computation, along with a progress bar and an estimated time to complete. The user has the option of stopping further calculation (using the “Cancel” or “Stop and Save” button) and accepting the model shown, at any time. Models that include two descriptors that correlate more than 95% are discarded.



Final QSAR Report

1.4.14 Spreadsheet format report

The data and results in ToxModel should be presented as an interactive spreadsheet format. Shading the data cells proportional to the data value is desirable (note green shading):



1.4.15 PDF Format Report

At the end of the calculation a report is created automatically on the "Report" page, and as a downloadable pdf file. The following is an example of the style and content of the QSAR report in PDF format:

QSAR report for model 1 in the Nitrofurans project

User: David
Date: Aug 6, 2009
SW version: ToxModel build 2590

CAUTION: This report contains warnings.

Part 1: Summary

a. Property predicted

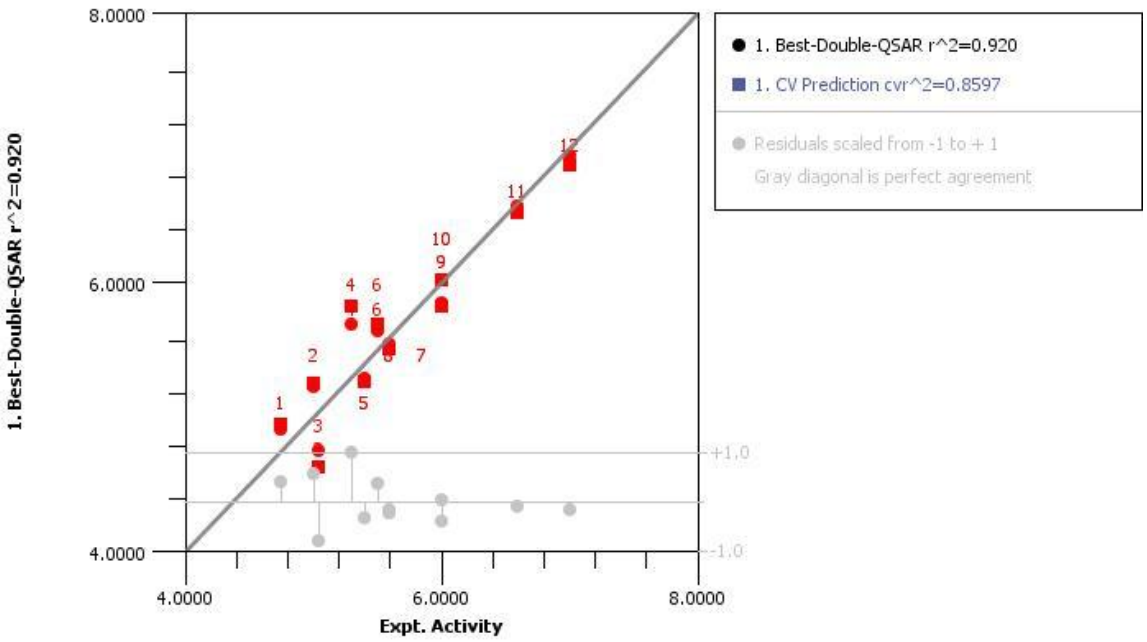
The property used to develop the QSAR is named *Expt. Activity* and its values were obtained from a training set of 12 chemical samples.

b. Best QSAR equation

Using the *Fast QSAR* option, the following regression equation, which is the best from 5,356 possible double combinations of 104 descriptors, gave the highest *r^2*=0.9196.

Expt. Activity = -5.95703e-03**principal moment third* + 1.2490**path index 5* + 3.9388

1. Best QSAR plot



chemical sample	Expt. Activity	1. Best-Double-QSAR r^2=0.920 (a)	1. CV Prediction cvr^2=0.8597	1. Best Single r^2=0.762 (b)
Nitrofuran-01 (c)	4.75	4.8937	4.938	4.719
Nitrofuran-12	5	5.2093	5.2429	5.3536
Nitrofuran-02	5.05	4.7411	4.6192	4.8162
Nitrofuran-03	5.3	5.6801	5.8055	5.918
Nitrofuran-04	5.4	5.2715	5.2513	5.1403
Nitrofuran-11	5.5	5.6336	5.6735	5.6372
Nitrofuran-05	5.6	5.5321	5.4906	5.5886
Nitrofuran-10	5.6	5.5111	5.4985	5.8964
Nitrofuran-06	6	5.8393	5.8151	5.9882
Nitrofuran-08 (d)	6	6.0032	6.0036	5.9585
Nitrofuran-09	6.6	6.5509	6.5079	6.0153
Nitrofuran-07	7	6.9341	6.8652	6.7687

Footnotes

(a) Expt. Activity = -5.95703e-03*principal moment third + 1.2490*path index 5 + 3.9388; r2=0.9196 dof adjusted r2=0.9018 median r2=0.9152 cvr2=0.8597 avErr=0.1450 errSD=0.2190 cvSD=0.2468 median residual=0.1310

(b) Expt. Activity = 2.70060e-03*Wiener Index + 4.3922; r2=0.7625 dof adjusted r2=0.7387 median r2=0.7330 cvr2=0.6723 avErr=0.2342 errSD=0.3550 cvSD=0.3771 median residual=0.2325

(c) Only sample in set with an aldehyde count.

(d) The depth of 4.235 is more than 2.50 sigma from the average of 3.640. Only sample in set with a ring count nonaromatic 5. Only sample in set with a ring count 5 member. Only sample in set with a ring count all nonaromatic. Only sample in set with a methylene count.

c. Quality of the best QSAR equation

The cross-validated correlation coefficient (cvr^2= 0.8597) suggests that the stability of the equation on addition of similar training data is likely to be reasonable as it is above 0.70. A more detailed analysis is provided in Part 2.

The average error for the training set is 0.1450 and the standard deviation is 0.2190.

The F-ratio is 30.5123. The probability that a greater F-ratio can be obtained by chance alone is 0.0000. Since the probability is less than 0.05 (1 in 20), there is at least one significant descriptor in the model, i.e. this is a valid and stable equation. A probability above 0.05 indicates that the equation might be a chance correlation and not stable.

Based on the partial-F value of each descriptor, there is a greater than 99% probability that all descriptors are significant.

The training data and QSAR predictions were checked for the following and warnings were noted and are discussed in more detail in Part 2:

1. There are enough observed data values per descriptor.
2. The data is distributed evenly enough.
3. The training set had these notes:

   Nitrofuran-01:

   Only sample in set with an aldehyde count.

   Nitrofuran-08:

   The depth of 4.235 is more than 2.50 sigma from the average of 3.640. Only sample in set with a ring count nonaromatic 5. Only sample in set with a ring count 5 member. Only sample in set with a ring count all nonaromatic. Only sample in set with a methylene count.

4. No outliers or problems with the predicted values were found.

An independent set of chemical samples should be used to test this equation.

d. Applicable prediction range and chemical space

This equation should be used to estimate only data values that fall within the training range from 4.75 to 7. Predictions that fall outside this range should be treated with caution as there is no way to know if the correlation holds outside the training set range.

The QSAR should be used to predict values only for chemical samples that have properties in these ranges:

1. *principal moment third* from 108.0989 to 451.8945.
2. *path index 5* from 1.2801 to 4.5534.

Predictions for chemical samples with properties that fall outside the training set property range should be treated with caution as there is no way to know if the correlation holds outside the training set range.

The QSAR should be used to predict values only for chemical samples that are chemically similar to the training set or share a common mode of action. The 12 samples in the training set included the following elements (min:max:# samples): H(3:10:12) C(5:15:12) N(1:6:12) O(3:5:12). The 12 samples in the training set included the following functional groups: *ring nonaromatic*, *aldehyde*, *H-bond donor*, *rotatable bond nonterminal*, *H-bond acceptor*, *donatable hydrogen*, *ring aromatic*, *molecule*, *rotatable bond*, *methylene*, *alkene*, *nitro*, *ring size largest*, *ring*, *ring nonaromatic 5*, *ring 6 member*, *ring size smallest*, *ring aromatic 6*, *ring aromatic 5*, *ring 5 member*, *guanidine*, *amine*, *amide*.

e. Mechanistic interpretation

The descriptors and their relative importance are listed below:

Descriptor	Relative importance
principal moment third	-0.5625
path index 5	1.0000

Part 2: Detailed Analysis

Data distribution

There are 6.000 data values per descriptor in the QSAR model.
The QSAR equation was derived using a training set of 12 chemical samples with a three-sigma range for Expt. Activity from 3.7576 to 7.5424. The average was 5.65 and the standard deviation was 0.6308 with a minimum data value of 4.75 and a maximum of 7. The data skewness measure is 0.6744. The data skewness is between -2.0 and 2.0 which indicates that the data is not skewed. Partitioning the data into equal thirds from lowest to highest data values gives three bins with these counts: 6:4:2.

Chemical samples

Chemical samples were located in directory:

1. C:\Users\David\Desktop\QSAR Exercise files\2-Nitrofurans\nitrofuran_3

Chemical samples in the training set had molecular weights from 141.0816 to 266.2515 and these elements and counts:

Element	Lowest	Highest	Count
Hydrogen	3	10	12
Carbon	5	15	12
Nitrogen	1	6	12
Oxygen	3	5	12

All chemical samples used to develop the model were uncharged, neutral species.
The chemical samples had these groups and counts:

Group	Lowest	Highest	Count
ring nonaromatic	0	1	1
aldehyde	0	1	1
H-bond donor	0	2	10
rotatable bond nonterminal	2	5	12
H-bond acceptor	4	8	12
donatable hydrogen	0	4	10

ring aromatic	1	3	12
molecule	1	1	12
rotatable bond	2	6	12
methylene	0	1	1
alkene	0	2	6
nitro	1	1	12
ring size largest	5	6	12
ring	1	3	12
ring nonaromatic 5	0	1	1
ring 6 member	0	2	5
ring size smallest	5	5	12
ring aromatic 6	0	2	5
ring aromatic 5	1	1	12
ring 5 member	1	2	12
guanidine	0	2	3
amine	0	2	7
amide	0	2	2

Chemical samples were used as-is without preconditioning the geometry.
The preconditioning geometry model was MOPAC:PM5.
The training set samples had these notes:

Nitrofuran-01:

Only sample in set with an aldehyde count.

Nitrofuran-08:

The depth of 4.235 is more than 2.50 sigma from the average of 3.640. Only sample in set with a ring count nonaromatic 5. Only sample in set with a ring count 5 member. Only sample in set with a ring count all nonaromatic. Only sample in set with a methylene count.

Analysis of QSAR equation

The following equation predicts Expt. Activity:

Expt. Activity = -5.95703e-03**principal moment third* + 1.2490**path index 5* + 3.9388

The average error for the training set is 0.1450. The standard deviation of the error is 0.2190.
r^2 is 0.9196; the degrees-of-freedom adjusted r^2 is 0.9018;; the leave-one-out cross-validated r^2 is 0.8597. and the median r^2 is 0.9152. The standard deviation in the error predicted by leave-one-out cross-validation is 0.2468. The F-ratio is 30.5123. The probability that a greater F-ratio can be obtained by chance alone is 0.0000. Since the probability is less than 0.05, there is at least one significant descriptor in the model.
Use the normalized coefficients in the following analysis section to interpret the relative importance of each descriptor.
92% of the variability in Expt. Activity is explained by this equation.

The relative weight of each normalized contribution is:

Statistics for model 1 in the Nitrofurans project *Aug 6, 2009*

	Descriptor Coefficient	Normalized coefficient	Descriptor standard deviation	Partial-F	Probability of greater F-ratio	Variance Inflation Factor
principal moment third	-0.0060	-0.5625	106.9021	18.2960	0.0021	6.24
path index 5	1.2490	1.0000	0.9064	57.8228	0.0000	6.24
Constant	3.9388					

r^2 =	*dof* r^2 =	*SD* =	*cvr*^{2} =	*median* r^2 =	*cv SD* =	*F* =	*P* =
0.9196;	0.9018;	0.2190;	0.8597;	0.9152;	0.2468;	30.5123;	0.0000;

Correlation

	principal moment third	*path index 5*
principal moment third	1.0000	
path index 5	0.9163	1.0000

The equation is the best from 5,356 possible double combinations of 104 descriptors. The correlation between any pair of descriptors that appear in the equation is less than 0.9500. All descriptors have fewer than 90.00% identical values. All descriptors also have a fractional standard deviation of at least 0.0001.
A plot of predicted values against the original data may be found in the 1. Best QSAR plot. Other equations with *r^2* from 0.9196 to 0.8521 may be found in the 1. Other QSARs table.
This equation should be used only to predict values of similar chemicals that have molecular weights, elements, groups and charges within the lowest to highest ranges of the training set.

List of 104 descriptors evaluated

Descriptors

1. molecular weight (mass_au)
2. log P
3. length
4. width
5. length/width
6. depth
7. width/depth
8. log P/(width/depth)
9. box volume (angstrom^3)
10. log P/box volume
11. box area (angstrom^2)
12. log P/box area
13. box cross section
14. log P/box cross section
15. principal moment first
16. principal moment second
17. principal moment third
18. principal moment ratio 1st/2nd
19. principal moment ratio 1st/3d

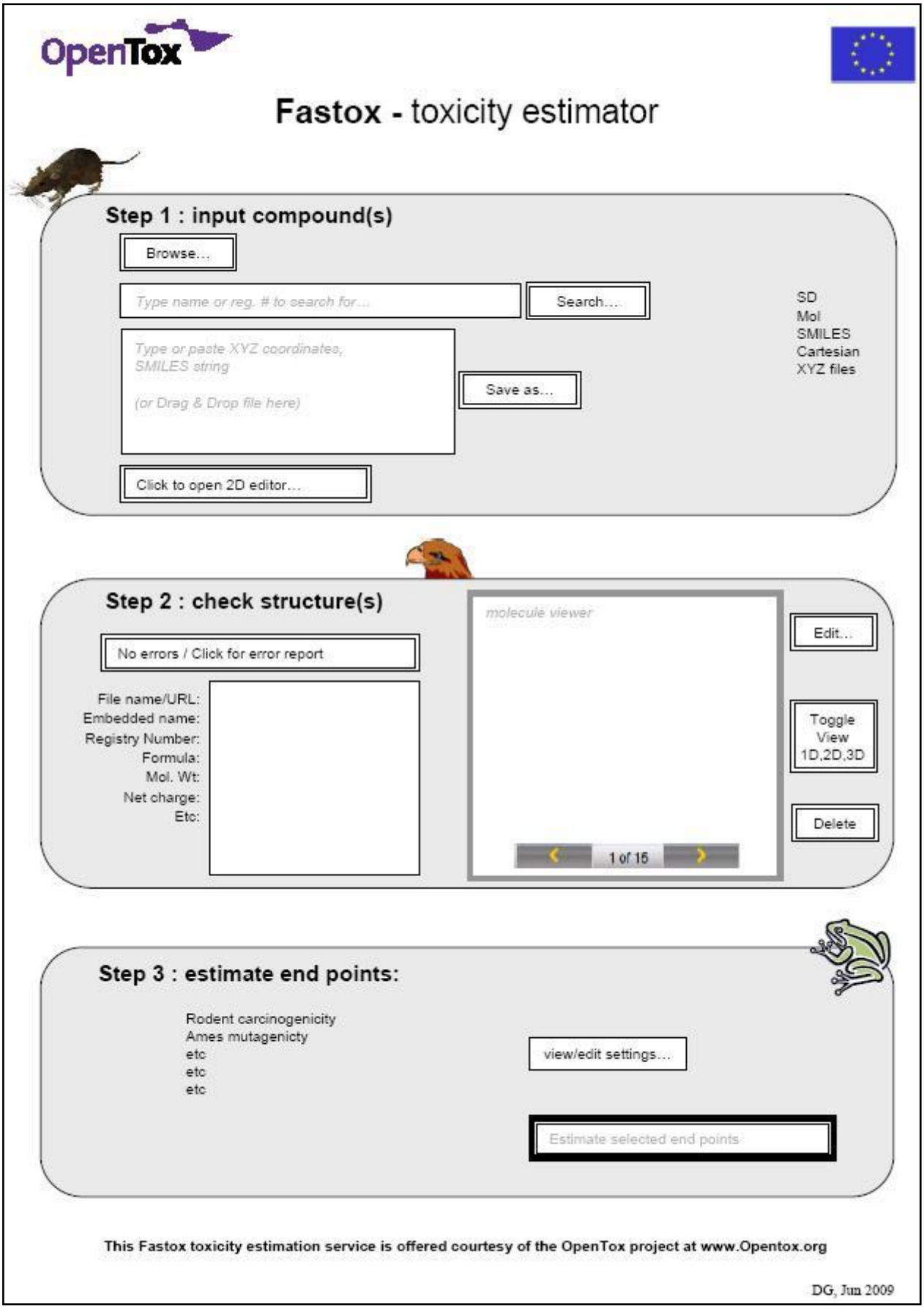
20. principal moment ratio 2nd/3rd
21. radius of gyration
22. donatable hydrogen count
23. H-bond donor count
24. H-bond acceptor count
25. rotatable bond count
26. rotatable bond count nonterminal
27. all atom count
28. Hydrogen count
29. Carbon count
30. Nitrogen count
31. Oxygen count
32. all bond count
33. single bond count
34. double bond count
35. bonded gravitational index
36. cube root(bonded gravitational index)
37. path index 0
38. path index 1
39. path index 2
40. path index 3
41. path index 4
42. path index 5
43. valence path index 0
44. valence path index 1
45. valence path index 2
46. valence path index 3
47. valence path index 4
48. valence path index 5
49. cluster index 3
50. cluster index 5
51. valence cluster index 3
52. valence cluster index 5
53. path-cluster index 4
54. path-cluster index 5
55. valence path-cluster index 4
56. valence path-cluster index 5
57. chain index 6
58. valence chain index 6
59. shape index order 1
60. shape index order 2
61. flexibility index
62. graph diameter
63. shortest path count
64. shortest path average
65. shortest path standard deviation
66. Csp^2 bonded to 2 C
67. alkene count
68. amide count
69. amine count

70. guanidine count
71. ring count all
72. ring count all aromatic
73. ring count 6 member
74. ring count aromatic 6
75. ring size largest
76. Balaban J Index
77. Wiener Index
78. total accessible surface area (angstrom^2)
79. total accessible surface volume
80. highest partial charge on H
81. highest partial charge on donatable H
82. lowest partial charge on free H acceptor
83. lowest partial charge on O
84. lowest partial charge on N
85. highest partial charge on C
86. lowest partial charge on C
87. second highest partial charge on H
88. third highest partial charge on H
89. highest partial charge on N
90. dipole moment from partial charges (debye)
91. bond surface area
92. hydrogen donor partial surface area
93. (hydrogen donor partial surface area/total accessible surface area)
94. (hydrogen donor partial surface area/total accessible surface volume)
95. hydrogen acceptor partial surface area
96. (hydrogen acceptor partial surface area/total accessible surface area)
97. (hydrogen acceptor partial surface area/total accessible surface volume)
98. min(hydrogen acceptor partial surface area, hydrogen donor partial surface area)
99. hydrophobic dipole
100. hydrophobicity weighted area
101. hydrophobicity weighted positive area
102. hydrophobicity weighted negative area
103. (hydrophobicity weighted positive area/total accessible surface area)
104. (hydrophobicity weighted negative area/total accessible surface area)

Descriptors with no variance have been omitted from the list above.

End of QSAR report for model 1 in the Nitrofurans project

Figure 2. Draft layout of "Fastox" Predictor



2. Graphical User Interface (GUI) for predicting toxicity

This section describes a proposed graphical user interface for the OpenTox suite of tools to enable non-QSAR-experts to easily use OpenTox's QSAR methods to predict toxicity.

2.1 Description

This chapter describes a graphical user interface (GUI) "Fastox" designed primarily for the 'novice' with little or no knowledge of QSAR or computational chemistry, to make predictions using the OpenTox toxicity prediction software package. This Fastox interface does not allow for the building of new QSAR models, as this is a separate function with a different interface (Toxmodel).

The software package can be installed as a complete stand-alone application for Windows and Linux. It can also be accessed without any local installation via a browser interface with remote computing via the internet (for example, hosted by http://www.opentox.org/toxicity-prediction or linked server). This particular interface uses in-built QSARs and methods, and does not include any facility for building new models.

2.2 Assumptions

The user needs a detailed report on the potential toxicity of a specified compound for REACH submission (for example).

The user knows the structure of the compound, or the CAS registry number, or has an electronic file (e.g. MOL). OpenTox provides one web page (or application) called "Fastox" for predicting toxicity end-points, and a separate page (or application) for building QSAR models.

2.3 Procedure for "Step 1"

1. User navigates to the "Fastox" page or opens the Fastox application (if installed locally). It may be appropriate to access the web page via a login so that preferences and results can be saved.

2. User then enters structure via one of several different methods:

i) **Browse...** opens a browser to local and networked files and to OpenTox databases, for various file types (SD, Mol, SMILES, Cartesian & XYZ files), The "Open" button opens the file which then appears in the window in "Step 2".

ii) **Search...** A name is typed in the box, then "Search…" opens a dialog to choose various search options such as: local computer, local network, internet, CAS registry, database xyz, all options, etc. The "Open" button opens the file which then appears in the window in "Step 2".

iii) "**Type or paste...**" window allows input of SMILES string or XYZ coordinates, etc.? The "Save as…" button creates a file which then appears in the window in "Step 2"

iv) **Click to open 2D editor...** opens a 2D drawing editor (can we find a plug-in, such as at http://qsardb.jrc.it/qmrf/search_substances.jsp?). If possible, this window can accept copy and paste from ChemDraw, etc. Once the structure is complete, the user needs to click "Save as..." to create the file which then appears in the "Step 2" window.

v) **Drag and Drop** the file anywhere on page (SD, Mol, SMILES, Cartesian & XYZ files). The file then appears in the "Step 2" window.

(this may be too complex for first implementation, so may be deferred)

OpenTox will automatically recognise the following file types: SD, Mol, PDB, MOPAC, ChemDraw, other common file types? etc. Files containing a batch of structures have to be in SD format.

2.4 "Step 2"

The user needs a detailed report on the potential toxicity of a specified compound for REACH submission (for example).

2.4.1 Error Checking

After the file is loaded or saved, the structure is automatically checked for various errors. Any warning or error reports will be automatically presented at this point as a separate pop-up error text window. If there are no errors, the first box of Step 2 shows "No errors". If there are errors, a pop-up error report window opens automatically, and the box shows "Click for error report". Clicking this box opens up the same separate pop-up error text window or brings it to the foreground if already open.

Possible errors & warnings include:

"{Structure number(s) with error (if multiple structures in file)

"Multiple molecules in window, non-relevant molecules should be deleted

"incorrect valence on atoms: etc.

"unrecognized atom type:

"Structure has a non-zero net charge of ? (ion)

"Structure for CAS number ??? not found

"File could not be read due to unrecognized format

"Etc.

Examples of 3D* errors:

Bond distance for atoms X and Y are outside normal range

Atoms X and Y are too close

Incorrect valence on atoms: C26, C28, etc.

The error report window has a button to "Edit", "Delete" and for some errors a "Continue anyway" button will be active. The "Edit..." button open a 2D and simultaneous 1D (text) editor. Both windows

are interactive (edits in one window automatically update the other immediately) and editing is allowed in both..

2.4.2 3D structures

Assuming that the 3D conversion is fast and that OpenTox would eventually include some 3D QSARs based on QM methods or 3D-substructure searching, then it may be best to automatically create the 3D structure at Step 2. This would enable it to be viewed and verified as a 3D structure before proceeding with the calculations.

If it is certain that no 3D structure would be needed by any OpenTox methods and/or if the conversion is slow (> 1 second), then this would not be necessary unless the 3D view is selected.

2.4.3 Molecule viewer window

If a single structure file is opened, the picture of the structure automatically appears in the "molecule viewer" window in Step 2. This window shows the 2D structure by default

1D/2D/3D View

This button toggles the view between 2D, 1D (text: elements, XYZ, flags, connectivity, etc.) and 3D. The first time the 3D view is selected, a 3D structure is computed for viewing, if the file is not already 3D.

2.4.4 "Edit…" button

The "Edit…" button open a 2D and simultaneous 1D (text) editor. . If an existing file is edited, the original file cannot be overwritten, but is saved with an incremented suffix number.

2.4.5 "Delete" button

deletes everything in "Step 2" and returns the user to "Step 1".

2.4.6 Multiple structures

If an SD file with multiple structures is entered, only the first structure is shown in the windows, but navigation bar appears and allow the user to step through and view each structure, sequence numbering is also displayed at the base of the window (e.g. 4/150).

2.4.7 Information window

In the information window on the left are displayed the file name and path (or URL), embedded name, registry number, empirical formula, Mol. Wt. and net charge, etc.

"Step 3"

If there are no errors, the "Estimate selected end-points" button becomes active.

A list of the end-points selected for computation is shown (the default would be ALL available end-points)

2.5 "View/Edit settings..."

This button allows the users to view the detailed settings dialogue box and select or de-select endpoints and methods. Any changes to the selected end-points will be updated on the main page.

2.5.1 Default settings

End-points: All available end-points would be selected. The selected list is shown on the main page in section 3.

Estimation methods: All available methods e.g., Toxtree, Lazar, other-QSAR, read-across, etc. for each end-point would be used and reported. (We need to consider if it is practical to create a consensus report and how.)

Database information: OpenTox (and other?) databases would be searched for any existing data and this would be reported also.

2.5.2 "Estimate selected end-points"

This button starts the calculation.

A message "estimating end-points for compound 'XXXX', please wait" and a progress bar comes up with an elapsed time and a time-to-go (if possible).

Finally, the results page is presented as a PDF file (see draft report example below) that can be downloaded or saved.

FastTox Toxicity Estimation Report

Upon completion of all calculations, an HTML file is created and opened. A button is provided that allows the user to "Save this report as a PDF file…".

An example report format follows:

Note: red text is alternative text to (black) statement immediately above, depending on results

data for blue text is mined automatically from results of calculations

2.6 Example of Toxicity Report

Fastox toxicity report for: new_compound.sdf

Compound file name:	new compound.sdf
Compound name:	stuff
Empirical formula:	$C_{12}H_{18}O_4N_2$
Molecular weight:	226
Endpoint predicted:	pKa
Method used:	Toxtree 1.1
Technology:	QSAR
Prediction Model I.D.	AM.1256
Fastox build No:	build 2590 (web service)
User:	David
Date & time:	09.30, 23 June 2009

CAUTION: This report contains warnings.

Part 1: Summary

A. Database search results:

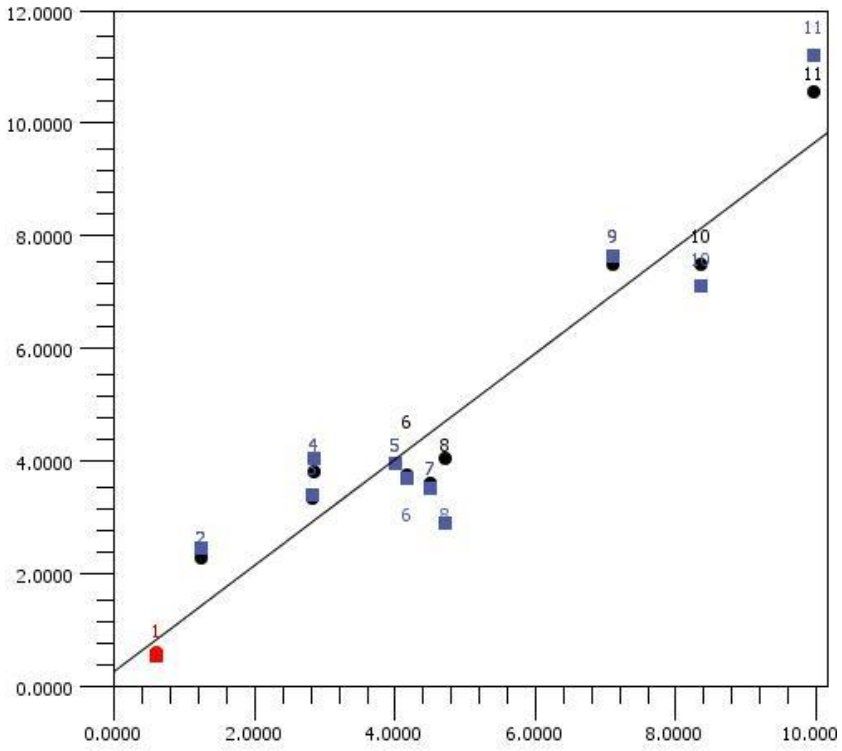
The compound of interest was found in the XXXX database with the following information:

xxxxxxxxxxxxxxxxxxxx
xxxxxxxxxxxxxxxxxxxxxxx
xxxxxxxxxxxxxxxx

The compound of interest was not found in any database:

B. Prediction results:

pKa	= 5.3
average unsigned error (test set)	= 0.1965
standard deviation (test set)	= 0.2989



Predicted value shown in red, calibration points shown in black

C. QSAR equation used:

Experimental pKa Value = 1.93732e-05*(principal moment third)^2 - 482.5944*(highest partial charge on H)^2 + 50.3412

Correlation coefficient, r^2 = 0.9952

Number of data points used in training set = 11

average error (training set) = 0.1465

standard deviation (training set) = 0.2389

D. Applicability domain for above prediction

Prediction range:

The predicted value for the compound of interest (5.30) falls within the range of predicted (or experimental?) values for the training set (1.25 – 10.50), i.e. it is interpolated.

WARNING: The predicted value for the compound of interest (5.30) falls nn% outside the range of experimental values of the training set (1.25 – 10.50), i.e. it is extrapolated and so the prediction should be treated with caution.

Descriptor range:

All descriptors values for the compound of interest fall within the descriptor ranges of the training set, i.e. they are interpolated

WARNING: The following descriptors for the compound of interest fall outside the descriptor ranges of the training set, i.e. they are extrapolated and so the prediction should be treated with caution.

highest partial charge on H

compound of interest = 0.35 Training set range = 0.04 – 0.33

Structural fragments:
All structural fragments of the compound of interest are represented in at least 5% of the training set compounds

WARNING: The following structural fragments of the compound of interest are present in less than 5% of the training set and so the prediction should be treated with caution:

CF3 is present in nn training compounds
N-NO2 is present in nn training compounds

Elemental composition:
All elements in the compound of interest are present in at least 5% of the training set compounds

WARNING: The following element(s) in the compound of interest are present in less than 5% of the training set and so the prediction should be treated with caution:

F is present in nn training compounds
Si is present in nn training compounds

Other property comparisons to training set:

Molecular weight:
The molecular weight for the compound of interest (5.30) falls within the range of Molecular weights for the training set (1.25 – 10.50),

WARNING: The molecular weight for the compound of interest (5.30) falls outside the range of molecular weights for the training set (1.25 – 10.50)

Water-octanol partition coefficient (LogP, Kow):
The LogP (estimated by the atom-additivity Ghose and Crippen method) for the compound of interest (5.30) falls within the range of LogP for the training set (1.25 – 10.50),

WARNING: The LogP (estimated by the atom-additivity Ghose and Crippen method) for the compound of interest (5.30) falls outside the range of LogP for the training set (1.25 – 10.50),

Water solubility (LogS, Sw):
The LogS (estimated by the atom-additivity Ghose and Crippen method) for the compound of interest (5.30) falls within the range of LogS for the training set (1.25 – 10.50),

WARNING: The LogS (estimated by the atom-additivity Ghose and Crippen method) for the compound of interest (5.30) falls outside the range of LogS for the training set (1.25 – 10.50),

E. Quality of the QSAR equation used

Experimental pKa Value = 1.93732e-05*(principal moment third)^2 - 482.5944*(highest partial charge on H)^2 + 50.3412

Correlation coefficient, r^2	= 0.9952.
Cross-validated correlation coefficient cvr^2	= 0.9911
Number of data points used in training set	= 11

average error (training set) = 0.1465
standard deviation (training set) = 0.2389
average unsigned error (test set) = 0.1965
standard deviation (test set) = 0.2989

The cross-validated correlation coefficient (cvr^2= 0.9911) suggests that the stability of the equation on addition of similar training data is likely to be good as it is above 0.90. A more detailed analysis is provided in Part 2.
The average error for the training set is 0.1465 and the standard deviation is 0.2389.
The F-ratio is 485.0260. The probability that a greater F-ratio can be obtained by chance alone is 0.0000. Since the probability is less than 0.05 (1 in 20), there is at least one significant descriptor in the model, i.e. this is a valid and stable equation. A probability above 0.05 indicates that the equation might be a chance correlation and not stable.
Based on the partial-F value of each descriptor, there is a greater than 99% probability that all descriptors are significant.
The training data and QSAR predictions were checked for the following and warnings were noted and are discussed in more detail in Part 2:

1. There are enough observed data values per descriptor.
2. The data is distributed evenly enough.
3. The training set had these notes:

   trichloroacetic:

   Only sample in the training set with a trichloro count.

   Only sample in set with a trihalogen count.
4. No outliers or problems with the predicted values were found.
5. An independent test set of chemical samples was used to test this equation.

F. Test Set Predictions

Comparison of predicted and experimental results for compounds used in the test set.

Compound	Experimental	Predicted	Error
cmpd1	27	29	2
cmpd2	35	31	-4
Averages			

G. Mechanistic interpretation

The descriptors and their relative importance are listed below:

Descriptor	Relative importance
(principal moment third)^2	0.1817
(highest partial charge on H)^2	-1.0000

Part 2: Detailed Analysis of QSAR Equation used

Data distribution

There are 5.500 data values per descriptor in the QSAR model.

The QSAR equation was derived using a training set of 11 chemical samples with a three-sigma range for Experimental pKa Value from -3.6542 to 12.8724. The average was 4.6091 and the standard deviation was 2.7544 with a minimum data value of 0.63 and a maximum of 10. The data skewness measure is 0.4799.

The data skewness is between -2.0 and 2.0 which indicates that the data is not skewed. Partitioning the data into equal thirds from lowest to highest data values gives three bins with these counts: 4:4:3.

Chemical samples

Chemical samples were located in directory:

1. C:\Users\David\Desktop\C-SAR Exercise files\1-pKa\pka_2

Chemical samples in the training set had molecular weights from 60.052 to 163.3871 and these elements and counts:

Element	Lowest	Highest	Count
Hydrogen	1	7	11
Carbon	2	7	11
Nitrogen	0	1	2
Oxygen	1	3	11
Chlorine	0	3	6

All chemical samples used to develop the model were uncharged, neutral species.
The chemical samples had these groups and counts:

Group	Lowest	Highest	Count
trichloro	0	1	1
carboxyl	0	1	8
methyl	0	1	3
phenol	0	1	3
H-bond donor	1	1	11
rotatable bond nonterminal	0	3	8
H-bond acceptor	1	3	11
donatable hydrogen	1	1	11
trihalogen	0	1	1
ring aromatic	0	1	4
molecule	1	1	11
rotatable bond	1	4	11
methylene	0	3	4
nitro	0	1	2
ring size largest	0	6	4
ring	0	1	4
ring 6 member	0	1	4
ring size smallest	0	6	4
ring aromatic 6	0	1	4

Chemical samples were preconditioned by Hf at PM6 geometry (QSAR_PM5_Geo_Tab_Polar).
The preconditioning geometry model was MOPAC:PM6.
The training set samples had these notes:

trichloroacetic:

Only sample in set with a trichloro count. Only sample in set with a trihalogen count.

Analysis of QSAR equation

The following equation predicts Experimental pKa Value:

Experimental pKa Value = 1.93732e-05*(*principal moment third*)^2 - 482.5944*(*highest partial charge on H*)^2 + 50.3412

The average error for the training set is 0.1465. The standard deviation of the error is 0.2389. r^2 is 0.9952; the degrees-of-freedom adjusted r^2 is 0.9940; the leave-one-out cross-validated r^2 is 0.9911. and the median r^2 is 0.9920. The standard deviation in the error predicted by leave-one-out cross-validation is 0.2721. The F-ratio is 485.0260. The probability that a greater F-ratio can be obtained by chance alone is 0.0000. Since the probability is less than 0.05, there is at least one significant descriptor in the model.
Use the normalized coefficients in the following analysis section to interpret the relative importance of each descriptor.
99.5% of the variability in Experimental pKa Value is explained by this equation.
The relative weight of each normalized contribution is:

Statistics for model 1 in the pKa project *Jun 23, 2009*

Descriptor	Coefficient	Normalized coefficient	Descriptor standard deviation	Partial-F	Probability of greater F-ratio	Variance Inflation Factor
(principal moment third)^2	0.0000	0.1817	28567.3913	42.9435	0.0002	1.57
(highest partial charge on H)^2	-482.5944	-1.0000	0.0063	1300.5994	0.0000	1.57
Constant	50.3412					

r^2 = 0.9952; *dof* r^2 = 0.9940; *SD* = 0.2389; *cvr*^{2} = 0.9911; *median* r^2 = 0.9920; *cv SD* = 0.2721; *F* = 485.0260; *P* = 0.0000;

Correlation

	(principal moment third)^2	*(highest partial charge on H)*^{2}
(principal moment third)^2	1.0000	
(highest partial charge on H)^2	0.6028	1.0000

The equation is the best from 537,166 possible double combinations of 1037 descriptors. The correlation between any pair of descriptors that appear in the equation is less than 0.9500. All descriptors have fewer than 90.00% identical values. All descriptors also have a fractional standard deviation of at least 0.0001.

End of Fastox report

3. Conclusion

This report offers a first design draft of GUI for non-expert users to predict toxicity using existing QSAR models. As it is difficult for "QSAR-Experts" to adequately anticipate the needs of non-experts, the design will likely evolve after use and feedback by 'real' non-expert users.