Volume

2

IDEACONSULT

Toxmatch User Manual

IDEACONSULT

Toxmatch User Manual

Nina Jeliazkova
Ideaconsult Ltd.
4 Angel Kanchev St.
1000 Sofia, Bulgaria
Phone +359 886802011
Email nina@acad.bg

Grace Patlewicz, Ana Gallegos Saliner
European Chemicals Bureau, TP 582
Institute for Health & Consumer Protection
Joint Research Centre, European Commission
21020 Ispra (VA), Italy
Phone +39 332789616, +39 332789291
Email grace.patlewicz@jrc.it
ana.gallegos@jrc.it

Table of contents

List of figures

Introduction

Toxmatch is a fully-featured and flexible user-friendly open source application, which encodes a variety of chemical similarity indices including several structural and descriptor based chemical similarity indices. A number of file formats are supported.

The Toxmatch application is suitable for use on a standalone PC. It has been designed with flexible capabilities for future extensions (e.g. other similarity schemes that might be developed in the future).

Background

Chemical similarity is a widely used concept in toxicology based on the hypothesis that similar compounds have similar biological activities. This hypothesis has a number of supporting examples as well as many examples where the relationship does not hold. In the latter case, a small change in chemical structure can lead to a significant change in biological activity (the so-called "similarity paradox"). Philosophy teaches us that similarity is not an absolute concept, but a relative one. This has important consequences on the precise definition of chemical similarity. Hence two objects cannot be similar in absolute terms, but only with respect to some property of the object. In the same way, two chemicals can only be similar with respect to some measurable key feature. A computerised analysis of similarity is based on numerical representation of the compound and a measure (similarity index) between these representations. There are a plethora of numerical representations (e.g. descriptors, structural or 3D representation) available for chemicals in addition to a range of approaches to measure the similarity. Toxmatch is able to determine the representation from structures supplied and to extract descriptors imported from a file. Since there is a vast number of published representations (more than 3000 descriptors available), Toxmatch provides implementations for only a subset. Two structural similarity approaches are encoded into Toxmatch. These are based on fingerprints and atom environments.

The objective of assessing chemical similarity in toxicology is to enable a more systematic identification of chemicals with similar biological activities. Given the "similarity paradox", it is unlikely that a single similarity measure exists which will be universally appropriate, instead similarity is best related to a given endpoint. Toxmatch provides several endpoint specific similarity measures, where descriptors are selected using a training set in combination with data mining methods.

Main features at a glance

- Ability to load datasets from *.mol*, *.sdf* files, or from a list of SMILES codes arranged in *txt*, *.csv* and *.xls* files;
- Structural information can be complemented with descriptors stored elsewhere i.e. *.sdf*, *.txt*, *.csv*, and *.xls* files. In this case, a primary key linking the structures and other information needs to be defined;
- Descriptors can be calculated on the fly or imported from separate files;
- Pair wise and composite (average) similarity measures can be calculated;
- Results are displayed in a table which can be subsequently exported;
- Chemicals may be ranked with respect to a chosen similarity index;
- Ready export of calculated similarity indices and similarity matrix in *sdf*, *.txt*, *.csv* and/or *.xls* formats;
- Various graphical displays including scatter plots, pair wise/composite similarity histograms and similarity matrices which can be exported as *.jpg* and *.png* formats;
- Implementation of a range of similarity indices, including:
    - Distance-Like similarity indices:
        - General definition:
          $D_{AB}(k,x) = \left[k(Z_{AA} + Z_{BB})/2 - xZ_{AB}\right]^{1/2}, D_{AB} = [0,\infty)$
        - Euclidean Distance Index (k=x=2):
          $D_{AB}(k,x) = \left[Z_{AA} + Z_{BB} - 2Z_{AB}\right]^{1/2}$
    - Correlation-like similarity indices:
        - General definition: $V_{AB}(k,x) = (k-x)Z_{AB}D_{AB}^{-2}(k,x), V_{AB} = [0,1]$
        - Hodgkin-Richards index: $H_{AB} = 2Z_{AB}[Z_{AA} + Z_{BB}]^{-1}$
        - Tanimoto index: $T_{AB} = 2Z_{AB}[Z_{AA} + Z_{BB} - Z_{AB}]^{-1}$

- Cosine-like similarity index or Carbó index: $C_{AB} = Z_{AB}[Z_{AA}Z_{BB}]^{-1/2}$

Development tools

The Toxmatch application is implemented in Java™. The basic cheminformatic functionality relies on the open source LGPL licensed Java ™ library The Chemistry Development Kit (CDK). The Integrated Development Environment (IDE) Eclipse, in conjunction with Apache Ant is the main development tool. Some of Toxmatch' capabilities are provided by the following open source libraries:

- JChemPaint – a structure diagram editor, which recently became part of CDK;
- org.xmlcml – CML support;
- gnujaxp – XML support;
- jgrapht – graph algorithms library;
- apache log4j – application logging;
- javax.vecmath – vector and matrix classes;
- junit – test suite framework;
- JFreeChart – scatter plots and histograms;
- Weka – classification and clustering algorithms;
- Toxtree – decision tree approach to estimating toxic hazard.

Launching Toxmatch

On a Windows ™ platform, Toxmatch may be launched either by using the "Start" menu, or by double clicking on the Toxmatch.jar file (the full path name is "C:\Ideaconsult\Toxmatch-vX.YZ\bin\Toxmatch.jar").

On all platforms (having Java™ 2 Runtime, Standard Edition 1.5 or newer installed), Toxmatch may be launched by executing the following command (after decompressing the ZIP archive distribution of Toxmatch):

java –jar Toxmatch.jar

Please, note that in the above mentioned command "java" and "Toxmatch.jar" should be eventually prefixed with the full path to java and Toxmatch on the destination platform.

Main screen layout

The main Toxmatch application window that appears upon launch is shown in Figure 1 below. The areas can be designated into the following: training/test dataset areas, training/test set groups and a visualisation area. The training dataset is akin to the terminology used in the (Q)SAR field. Here investigations on a starting dataset of chemicals and/or data are referred to as "training dataset". A dataset that is used to investigate the utility of the results derived from the training dataset is known as the test dataset.

In grouping approaches, the training dataset may represent a series of existing categories whereas the test set would represent the new chemical(s) under evaluation. Note the training dataset and test dataset can contain only one compound each as a minimum.

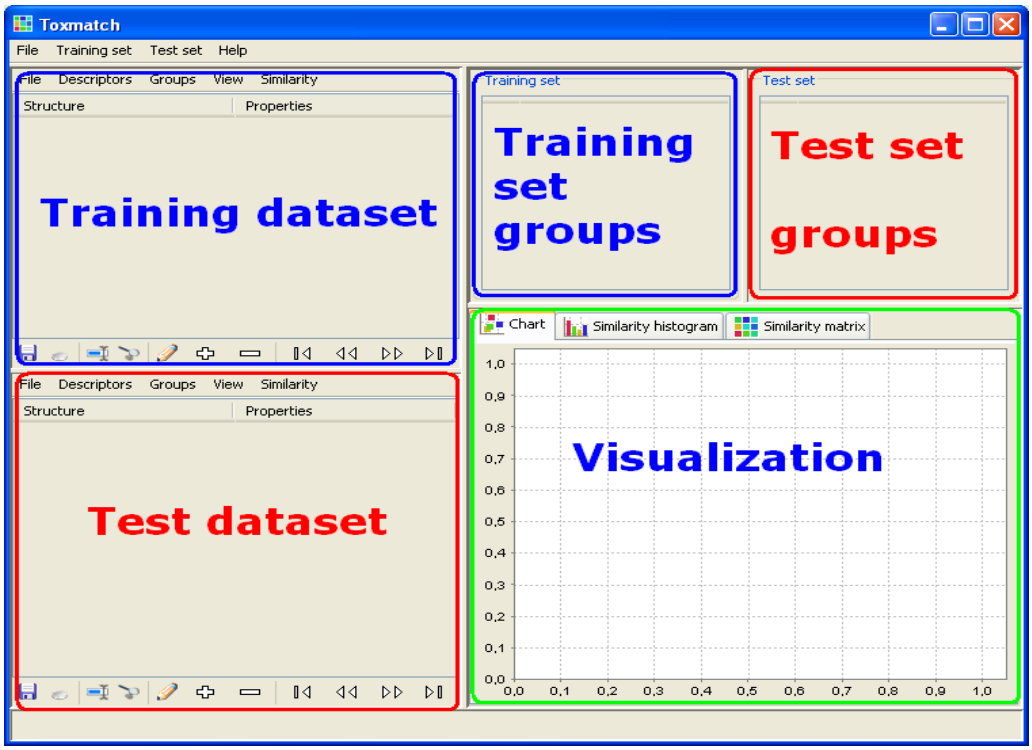


Figure 1: Main application screen upon launch

Training set

Toxmatch is able to open the following file types: CML, CSV, HIN, ICHI, INCHI, MDL MOL, MDL SDF, MOL2, PDB, SMI, TXT and XYZ file types.

> *Please, note that whilst CSV files can be read/written by MS Excel™ attention should be paid to the cell format (namely 'text'). Input of CSV and TXT requires a column with a "SMILES" heading in order for the structure to be read correctly. All other fields are optional and will be read and displayed as molecule properties.*

Toxmatch includes data sets for six toxicity endpoints – *Aquatic toxicity*, *Bioconcentration factor*, *Skin sensitization*, *Skin irritation*, *Carcinogenicity and Mutagenicity*

The aquatic toxicity dataset is a copy of DSSTox EPA Fathead Minnow Acute Toxicity dataset (April 2006 update) [1]. It contains 617 chemicals with information pertaining to organic chemical class assignments (ChemClass_FHM), acute toxicity in fathead minnow (LC50_mg), dose-response assessments (LC50_Ratio, ExcessToxicityIndex), behavioural assessments (FishBehaviourTest), joint toxicity

MOA evaluations of mixtures (MOA_MixtureTest), and additional MOA evaluation of fish acute toxicity syndrome (FishAcuteToxSyndrome) in rainbow trout [2].

The bioconcentration factor dataset comprises LogBCF values for 610 non-ionic chemicals taken from [3]. The dataset was used to develop the model in EPISuite BCFwin. CAS numbers, chemical names, LogP and LogBCF values are provided.

The skin sensitisation dataset contains local lymph node assay (LLNA) data for 210 chemicals, as published in [4]. In addition to the numeric LLNA EC3% values, a qualitative classification of potency category and reaction domain is reported for each compound, where potency category is one of non-sensitiser, weak sensitiser, moderate sensitiser, strong sensitiser, extreme sensitiser (as defined by Kimber et al., 2003 [5]) and reaction domain is one of Michael acceptors, S_NAr electrophiles, S_N2 electrophiles, Schiff base formers, Acylating agents, Nonreactive and special cases (as defined by Roberts et al., 2007 [6])

The skin irritation dataset consists of 72 chemicals, labelled according to EU/GHS classification for skin irritation potential (NI - not irritating, MI - mild irritation, R38 - irritating). The data arises from several sources including the US EPA TSCA (Toxic Substances Control Act) inventory and the ECETOC databank [7].

A second dataset with skin irritation data has been added since ToxMatch-v1.07, namely Skin irritation and corrosion: reference chemical data bank as per ECETOC Technical report No.66. Skin irritation potential is summarized as the Primary Irritation Index (PII), calculated from erythema and oedema grades. The maximum possible PII is 8.

The carcinogenicity dataset is a copy of ISSCAN - Istituto Superiore di Sanità, "CHEMICAL CARCINOGENS: STRUCTURES AND EXPERIMENTAL DATA", available at http://www.epa.gov/comptox/dsstox/sdf_isscan_external.html It contains 1153 chemicals with information pertaining to carcinogenicity studies of rat and mouse (male and female), as well as summary carcinogenicity data (field "Canc" with values 3 (carcinogen), 2 (equivocal) and 1 (noncarcinogen)). The ISSCAN dataset contains also data for mutagenicity in Salmonella typhimurium (Ames test), in the data field SAL, with possible values: 3 (mutagen), 2 (equivocal) and 1 (nonmutagen).

These may be opened by using the "Training set → File → Predefined sets" menu, and selecting the corresponding endpoint as shown in Figure 2. Clicking the "New" option will create and open a new empty file as a training set in which users are able to manually add training set compounds. At least one compound should be present in the training set in order to perform further similarity computations. Other training sets can be imported by clicking on the "Training set → File → Open" menu.

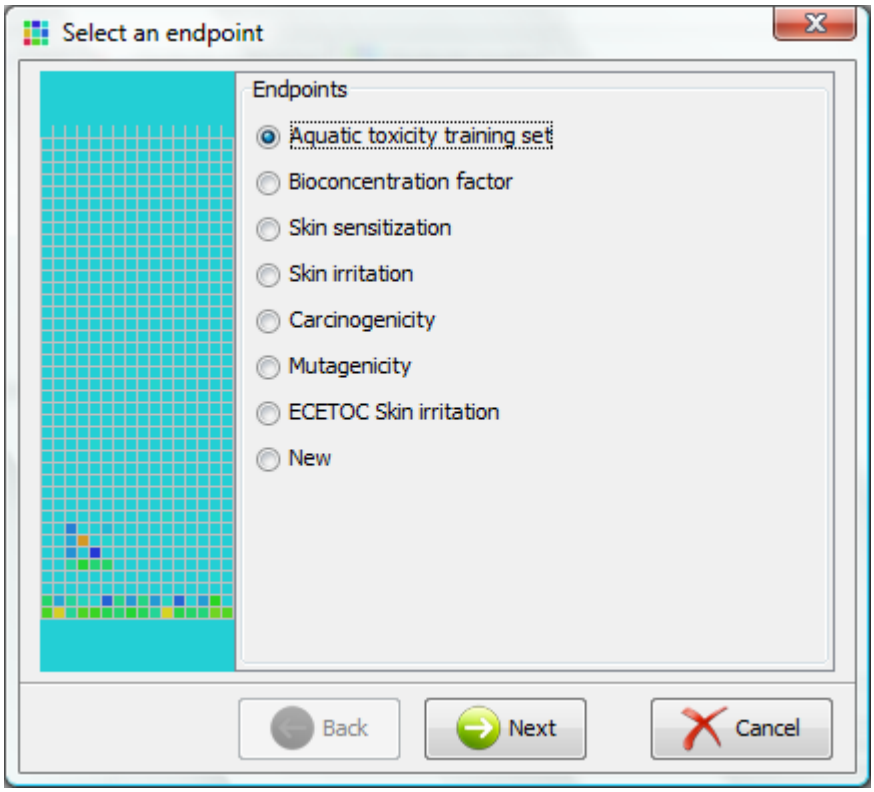


Figure 2: Selecting and opening a training set

Each of these training sets has been annotated and categorised in some way – either on the basis of mechanistic class, mode of action or potency class, etc. The user can select between these available categorisations (groups) or define new groups. The user is prompted to select one of the groups as shown in Figure 3 but note these can be modified at any later stage without reopening the file from the "Training set → Groups → Select groups" menu.

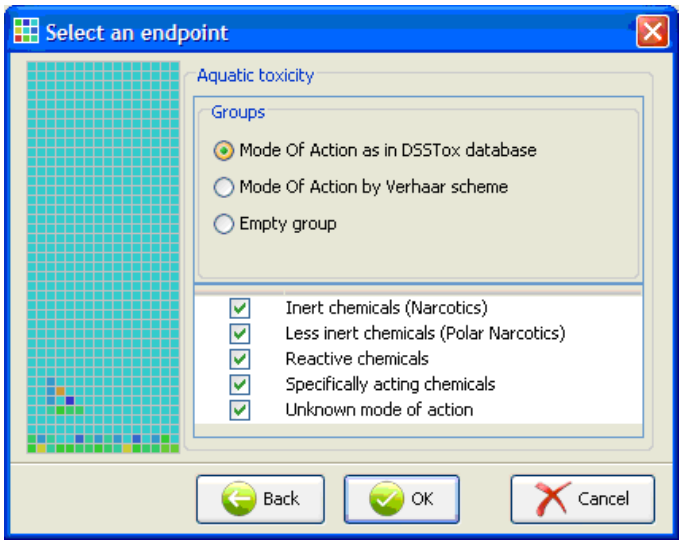


Figure 3: Group selection

The selected dataset and the groups are visible in the training set areas, as shown in Figure 4.

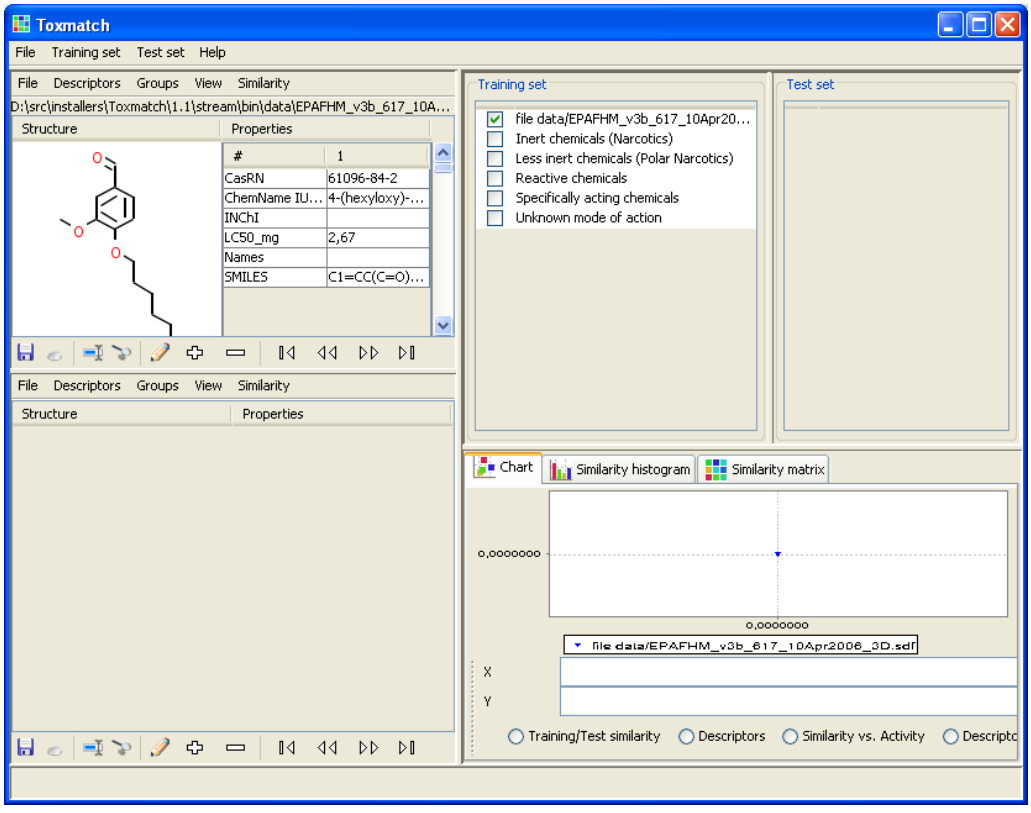


Figure 4: Main application window with training set loaded

The dataset, chemicals together with their properties and calculated similarity indices may be exported in a variety of formats e.g. *.sdf*, *.csv*, *.xls* or *.txt*, by clicking on the "Training set → File → Save" menu. The application will prompt the user for a file name as well as for any other fields to be saved. These can be selected using the dialog box as shown in Figure 5.

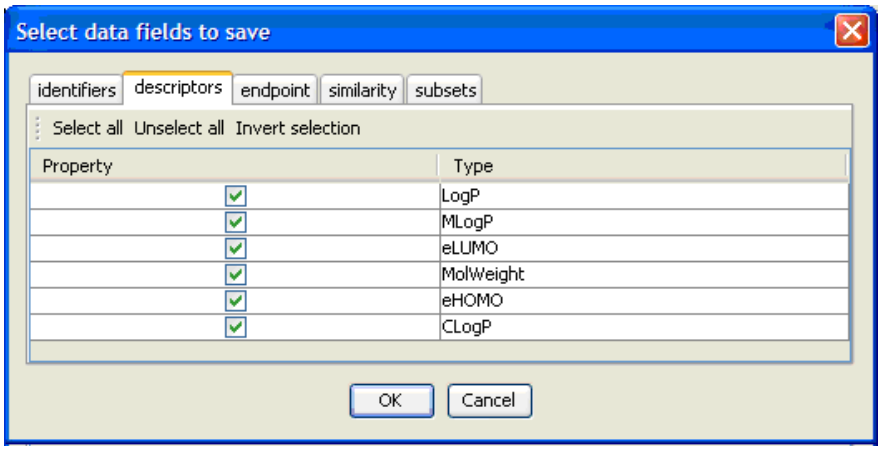


Figure 5: Selecting data fields to be saved

Descriptors

A range of descriptors can be calculated by Toxmatch. Those currently implemented are listed in Figure 6.

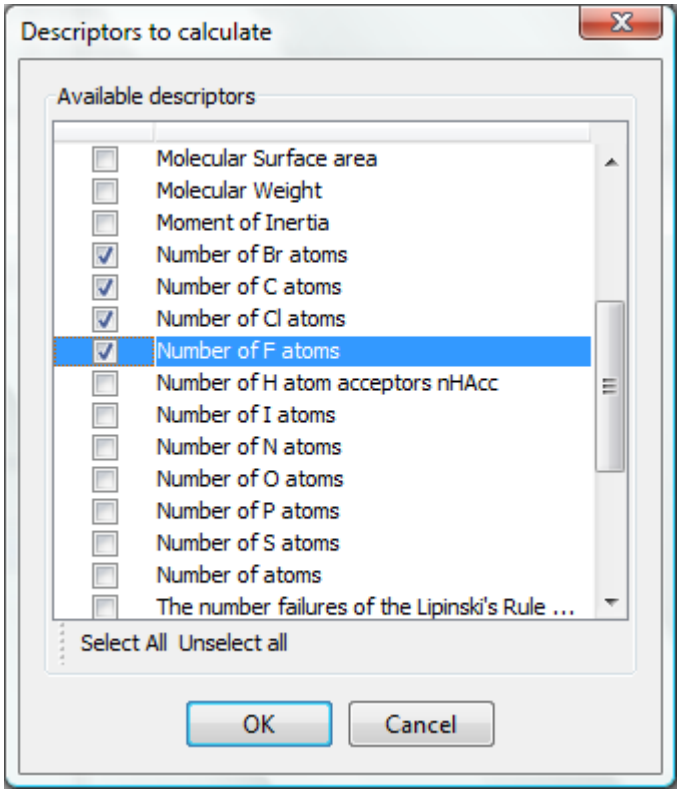


Figure 6: Descriptor calculation

A dataset with pre-calculated descriptors can be imported from a *.csv*, *.sdf*, *.txt* or *.xls* file. In this case, a dialog to select relevant descriptors is shown (Figure 7). If the descriptors have been calculated independently from the structures, the file of descriptor information can be subsequently imported so long as a primary key linking the two files is included. The primary key will ensure that the right descriptor information is attributed to the correct structure. In the example shown in Figure 7 the correspondence is established by CAS RN, which is specified in the Lookup field.

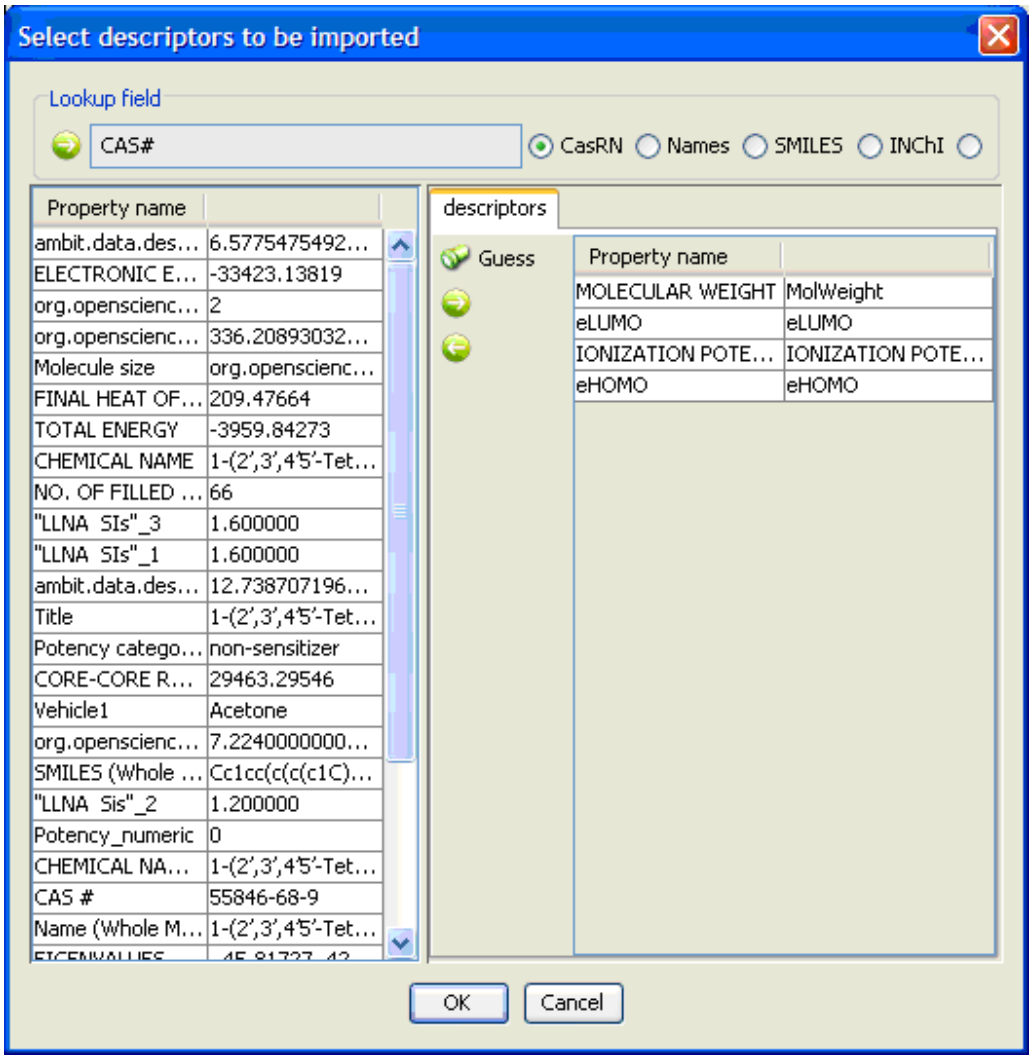


Figure 7: Descriptors import dialog

Groups

Toxmatch allows a user to categorise datasets into groups (e.g. potency categorisation for skin sensitisation, a set of different ranges for a BCF endpoint value, a mechanism of action, etc. These can subsequently be used to classify new chemicals into groups of potency, endpoint value ranges or mechanism of action, based on similarity values (see Similarity section). The classification procedure is dependent on the group definition; if no groups are defined for the particular dataset then it cannot be performed.

For the available example datasets, it is easy to switch between groups as shown in Figure 8. The current groups are simply replaced by the newly selected ones and displayed in the "*Training groups area*".

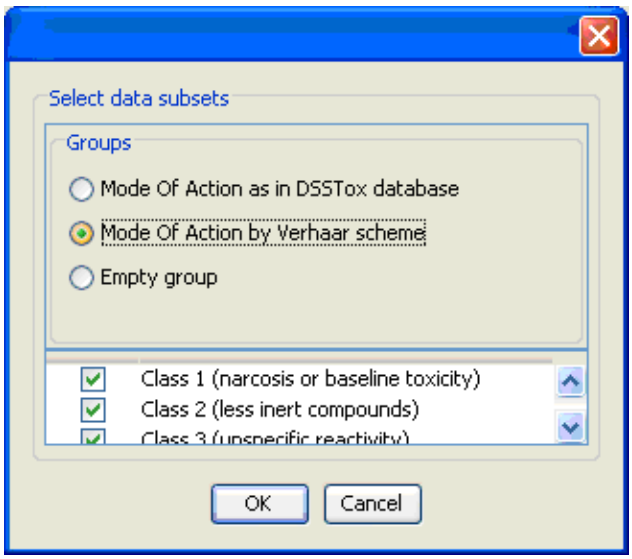


Figure 8: Groups selection

Alternatively new groups may be defined by following the dialog shown in Figure 9. New groups can be either empty, parameter value or range based.. This allows the user to define their own categorisation of the dataset thus the potential to apply classification based on similarity values for user defined groups. For example, the user could define groups based on ranges of LLNA EC3% values and subsequently predict ranges of LLNA EC3% values for new query chemicals. Alternatively, the user could define groups based on a parameter having a fixed value (e.g. Potency = "non sensitiser" would define a group of non sensitisers, whilst Potency = "moderate" would define a group of moderate sensitisers, etc.). The classification procedure, based on similarity values, can then be used to assign potency labels for new chemicals.

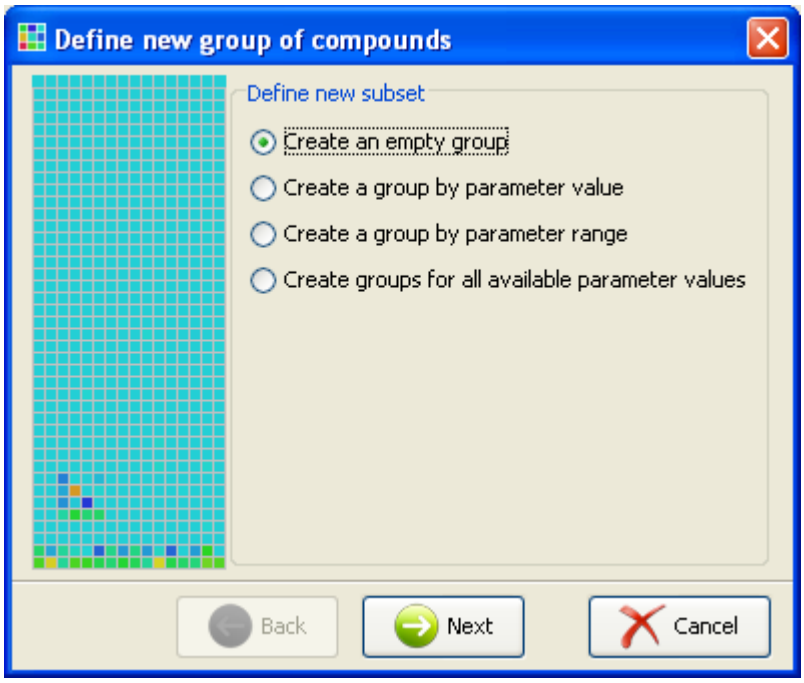


Figure 9: Defining a new group by different criteria

On defining a new group, the first step is to enter the parameter name and value, as shown in Figure 10.

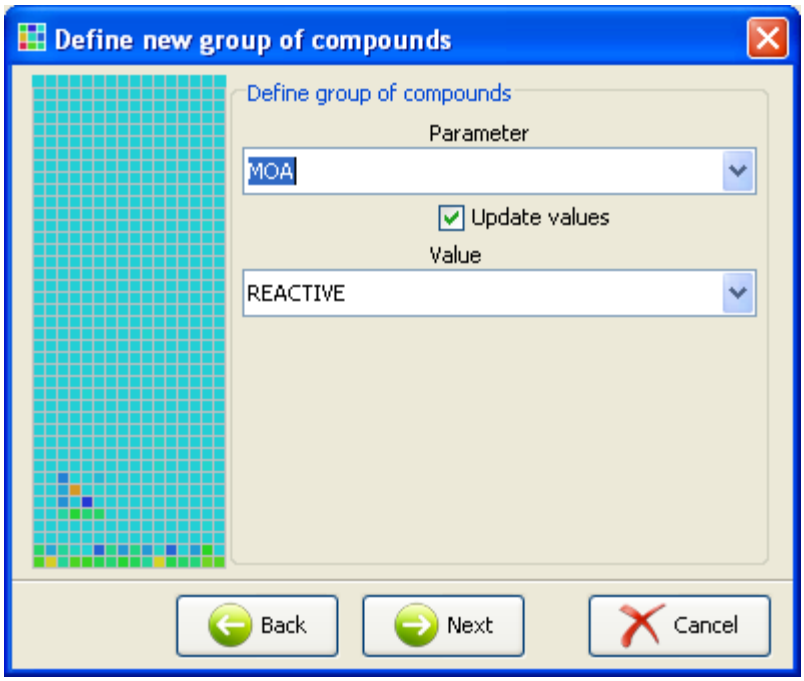


Figure 10: Defining a group consisting of chemicals with reactive mode of action (MOA=REACTIVE)

The next step is to specify other attributes of the group, as shown in Figure 11.

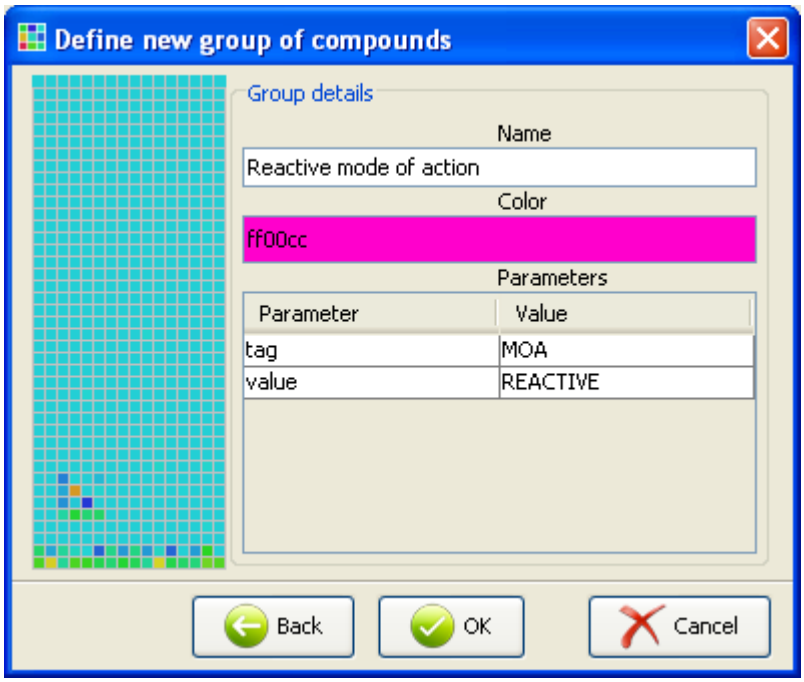


Figure 11: Group attributes

To define a group by parameter range, select the corresponding option (see Figure 9), then define parameter ranges, as shown in Figure 12.

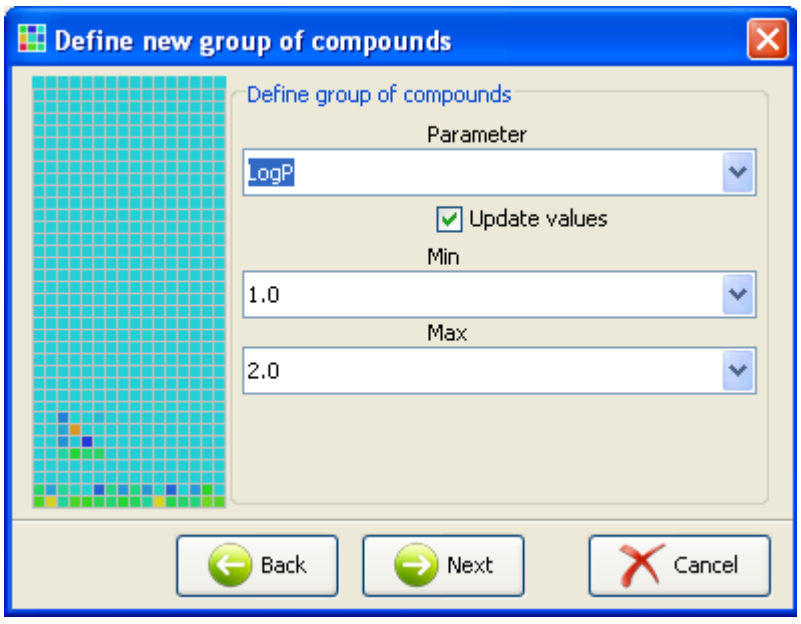


Figure 12: Defining new group by parameter range (1.0 < LogP < 2.0)

The selected compounds may be displayed by clicking on the new group line in the *"Training groups data area"*. The compounds will appear both on the scatter plot and in the training data area as shown in Figure 13. The compounds in the group can also be exported and saved as a new file by clicking the leftmost button on the dataset toolbar (highlighted in red in Figure 13).

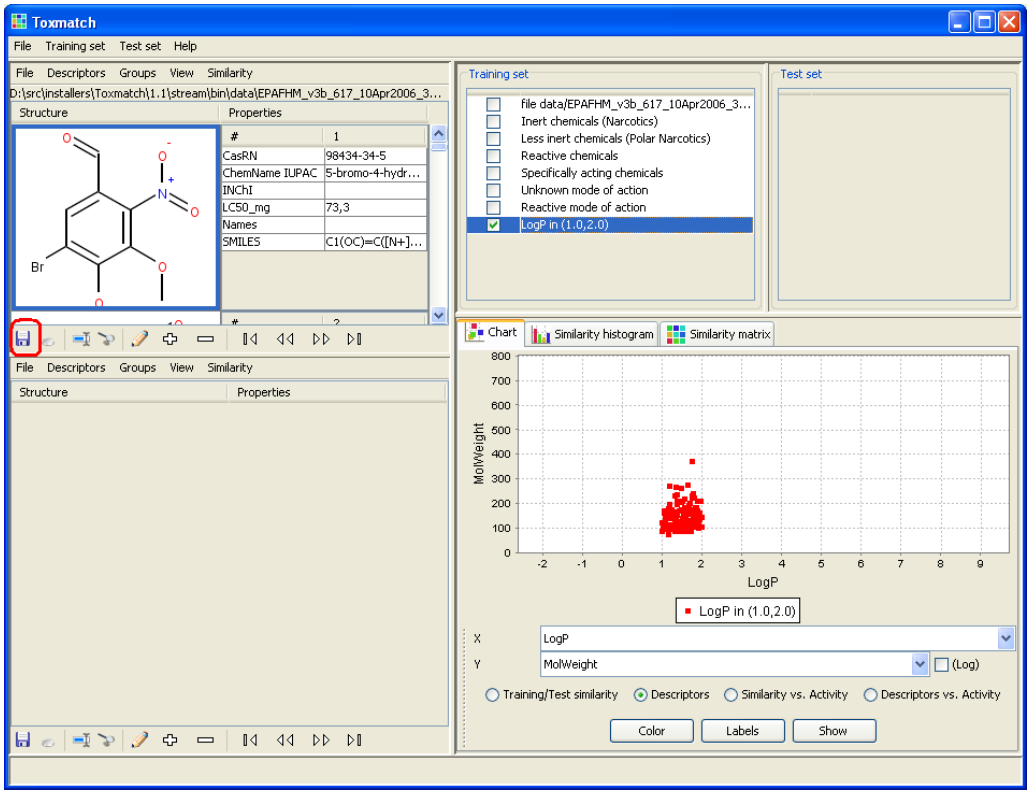


Figure 13: Group visualization

View modes and fields

The dataset can either be viewed as a table (text) or on a per chemical basis with the corresponding structure diagram. A dialog to select additional data fields for display is available (see Figure 14).

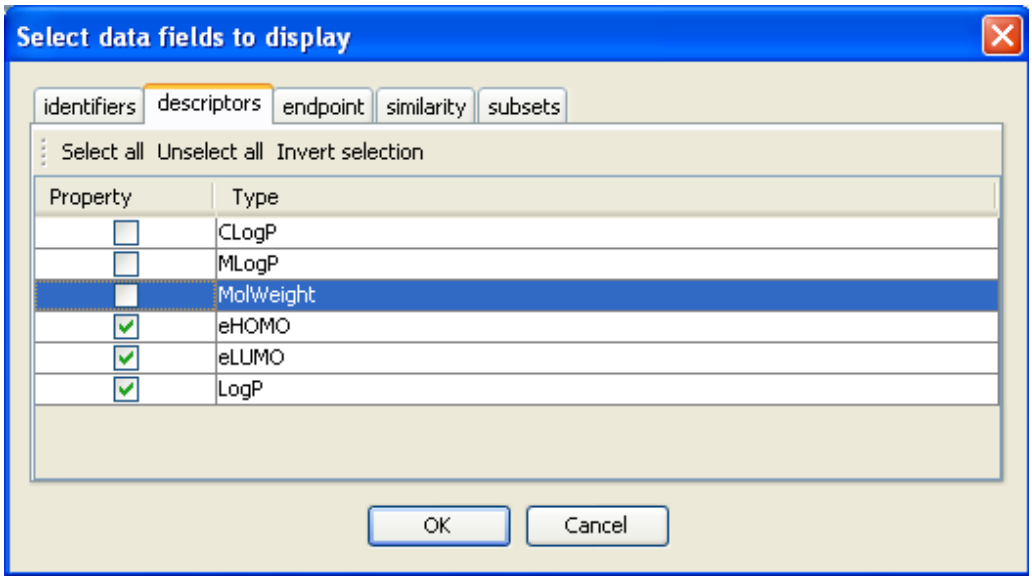


Figure 14: Data fields selection

Similarity

Background

Toxmatch provides the following pair wise similarity indices:

1. Similarity in descriptor space:
   - Euclidean Distance (k=x=2): $D_{AB}(k,x) = \left[Z_{AA} + Z_{BB} - 2Z_{AB}\right]^{1/2}$
   - Hodgkin-Richards index: $H_{AB} = 2Z_{AB}[Z_{AA} + Z_{BB}]^{-1}$
   - Tanimoto index: $T_{AB} = 2Z_{AB}[Z_{AA} + Z_{BB} - Z_{AB}]^{-1}$
   - Cosine-like similarity index or Carbó index: $C_{AB} = Z_{AB}[Z_{AA}Z_{BB}]^{-1/2}$
2. Structure similarity:
   - Tanimoto index: $T_{AB} = 2Z_{AB}[Z_{AA} + Z_{BB} - Z_{AB}]^{-1}$ between 1024 bits length hashed fingerprints. The fingerprint generation is based on the fingerprint implementation of the open source cheminformatics library The Chemistry Development Kit (CDK) [7] and follows the ideas of

Daylight's fingerprint theory [8] that states: 1) for a given molecule all possible paths for a predefined length (default is 7) are generated, 2) the path is submitted to a hash function which uses it as a seed to a pseudo-random generator, 3) the hash function outputs a set of bits, and 4) the set of bits thus produced is added (with a logical OR) to the fingerprint. We use 1024 bit fingerprints in this study.

- Hellinger distance between atom environments $H_{a,b} = \sum_{i=1}^{N} \left(\sqrt{a_i} + \sqrt{b_i} \right)^2$.

  Atom Environments (AE) [9, 10] can be regarded as fragments, surrounding each atom in a molecule, up to a predefined level. The calculation procedure is explained in [11]. First, atom types to be included in the generation of AEs are selected. We use 34 atom types, as listed in [11]. Next, a vector of length (34 * L+1) is constructed for each atom, where L is the maximum level for generating atom environments and L=3 by default. Third, for each atom, neighbours at level 1, 2, 3 are identified and corresponding counts stored in the vector. For example, if there are several C.sp2 atoms with the same neighbours up to 3rd level in the molecule, they will have the same string representation. We refer to this representation as a "fragment". In the equation above, N is the number of all fragments found in structures a and b, a_i refers to the probability (normalized frequency) of i^{th} fragment in chemical a and b_i refers to the probability of i^{th} fragment to be found in chemical b

- Maximum Common Substructure similarity (MCSS)

  $SI_{s,t} = \frac{(A+B)_{MCS}}{(A+B)_s} \frac{(A+B)_{MCS}}{(A+B)_t}$, where (A+B) is the sum of atoms and bonds in the maximum common substructure (MCS), in the s^{th} compound and in the t^{th} compound, respectively [12]. The MCS of two compounds is the largest possible substructure that is present in both structures.

Given a similarity index between two compounds, a similarity index between a chemical and a set of chemicals can be defined. Toxmatch provides the following options for such composite similarity assessment:

1. Similarity index between a representative (e.g. centre) point (compound) of the set and the query compound. This is available through the following options, appearing in the *Similarity* menu:

   - The **"Tanimoto distance (Fingerprints, kNN)"** method calculates Tanimoto similarity between fingerprint of a chemical and a consensus fingerprint. The consensus fingerprint is 1024 bit fingerprint where each

bit is set on, if at least one compound from the set has the corresponding bit on.

- The "**Hellinger distance (atom environments, summary atom environment)**" method calculates Hellinger distance between atom environments of a chemical and a atom environments of the set. Atom environments of the set is simply the set of all "fragments" (as defined above) found in the set.

2. Average similarity between a query chemical and the nearest *k* chemicals.

$$S_Q = \frac{1}{k}\sum_{i=1}^{k} S_{i,Q} \ ,$$

where S_{iq} is the pair wise similarity and *k* is either specified by the user (default is 10), or selected by a cross validation procedure. Note that the actual list of *k* most similar chemicals will depend on the similarity measure used (e.g. the 10 most similar chemicals selected by Euclidean distance may not coincide with the 10 most similar chemicals, selected by fingerprints similarity). This average similarity is available through the following options, appearing in the *Similarity* menu:

- The "**Euclidean distance (descriptors, kNN)**" method calculates average Euclidean distance between selected descriptors for the query chemical and *k* most similar chemicals from the set. The most similar chemicals are those with the smallest Euclidean distance.

- The "**Cosine similarity (descriptors, kNN)**" method calculates average Cosine index between selected descriptors for the query chemical and *k* most similar chemicals from the set. The most similar chemicals are those with the highest Cosine index values.

- The "**Hodgkin-Richards (descriptors, kNN)**" method calculates average Hodgkin-Richards index between selected descriptors for the query chemical and *k* most similar chemicals from the set. The most similar chemicals are those with the highest Hodgkin-Richards index values.

- The "**Tanimoto distance (descriptors, kNN)**" method calculates average Tanimoto index between selected descriptors for the query chemical and *k* most similar chemicals from the set. The most similar chemicals are those with the highest Tanimoto index values.

- The "**Tanimoto distance (fingerprints, kNN)**" method calculates average Tanimoto index between 1024 bit fingerprints (as above) for the query chemical and fingerprints for the *k* most similar chemicals from the

set. The most similar chemicals are those with the highest Tanimoto index values.

- The "**Hellinger distance (atom environments, kNN)**" method calculates average Hellinger distance between atom environments (as above) for the query chemical and atom environments for the *k* most similar chemicals from the set. The most similar chemicals are those with the highest Tanimoto index values.

Toxmatch goes further than merely calculating similarity values. The K-nearest neighbour classifier and learner algorithms [13] allow users to exploit the similarity information in two different ways:

- to predict the activity of a chemical of interest by considering the activities of similar compounds (nearest neighbours), or

- by "clustering"/"grouping" similar chemicals together on the basis of similarity values.

The implementation for the descriptor-based similarity is based on the open source Weka data mining software [14], while for the rest of the available similarity measures, the same algorithm for K-nearest neighbour classification and prediction is implemented specifically for Toxmatch.

The first approach results in a prediction of activity based on the weighted average of the activity values of the *k* nearest neighbours i.e. the activity of the most similar (closest) chemicals are averaged proportionately and used to estimate the activity of a chemical of interest.

$$P_Q = \frac{1}{\sum_{i=1}^{k} S_{i,Q}} \sum_{i=1}^{k} S_{i,Q} A_i \ ,$$

where k is the number of most similar chemicals, A_i are the activity values of these chemicals and S_{iQ} are the pair wise similarity values and P_Q is the predicted activity of the query chemical. Figure 15 is an illustration of the approach.

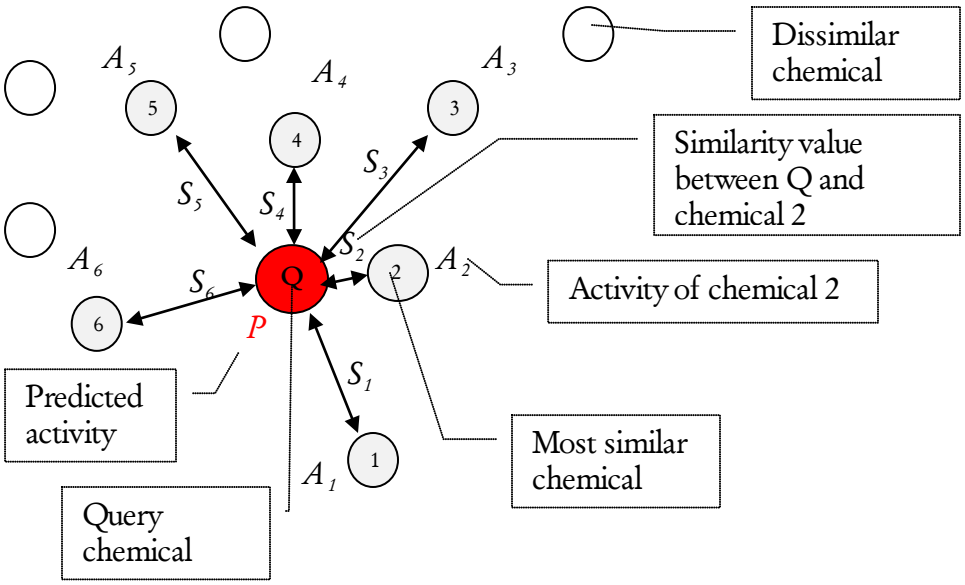


Figure 15: Prediction based on similarity and k-nearest neighbours (k most similar chemicals)

In this case the most similar chemicals could be referenced as source chemicals and the chemical of interest, the target chemical. The process of relating these source chemicals to a target chemical could be termed as a many to one read-across. The actual set of *k* most similar compounds will depend on the similarity measure. The weights are proportional to the pair wise similarities (e.g. the activity value of most similar compound has largest weight and vice versa). In order to predict dependent variable (activity), the measured activity values should be available for the training set. Two values are reported per each compound– averaged similarity to the *k* nearest neighbours and predicted activity value.

The second approach performs a classification into groups of activity, based on similarity values. The read-across is slightly different whereby the source chemicals are binned into one group and the similarity measure provides the means to define the likelihood that the target chemical falls into one or other bin. The procedure also relies on *k* nearest neighbours and classifies the target compound into the group where most of the *k* most similar compounds belong. For this purpose activity groups should be available for the training set (e.g. potency classes or other grouping). The values reported are: Probability to belong to a group (*m/k , where m is the number of compounds in the group)* and the predicted group. Figure 16 illustrates the classification of the query chemical Q as an "Active" chemical, based on the majority of the most similar chemicals (4 out of 6) being active, while only two out of six are inactive.

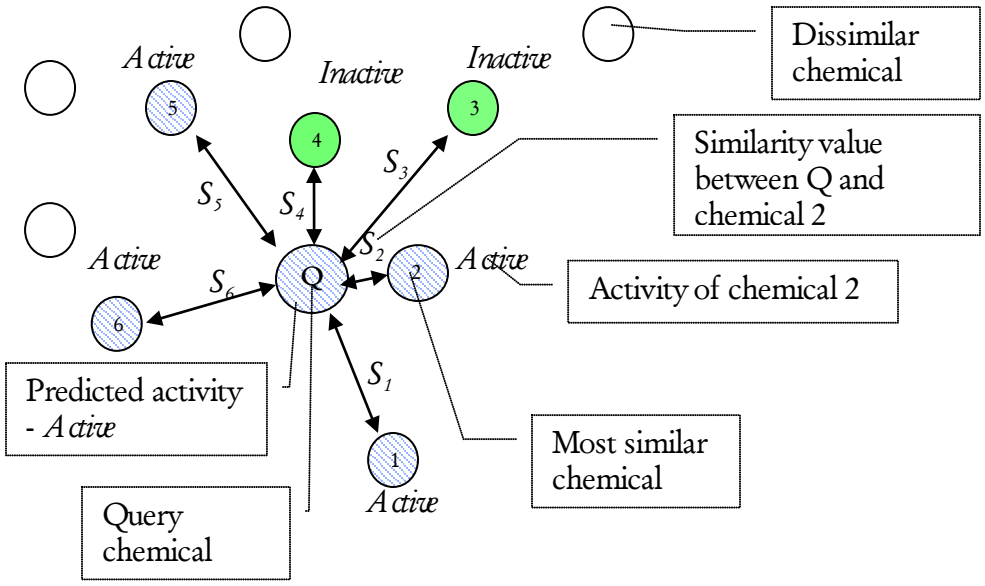


Figure 16: Classification based on similarity and k-nearest neighbours (k most similar chemicals)

Toxmatch provides means to calculate similarity of a query chemical to the training and similarity to the test set. The basic algorithm is as explained above, but in case of similarity to the training set, the *k* most similar chemicals are selected from within the training set, while when calculating similarity to the test set, the *k* most similar chemicals are selected from within the test set.

The most common usage is to calculate similarity to the training set, since the training set is expected to contain endpoint data, either numeric (used for prediction) or nominal (used for classification). Calculating similarity to the test set makes sense in cases when a comparison of the two sets is necessary (as in Figure 26).

How to

The first step is to select a similarity method. The choice of method is context dependent on the chemicals of interest and on the knowledge of the activity measure or endpoint. A variety of approaches may be performed and each should be carefully evaluated in the context of the problem of interest. Toxmatch has the following measures included: Euclidean distance, Cosine similarity, Hodgkin-Richards Index and Tanimoto distance, Fingerprints and Atom environments. The first four of these are descriptor based methods; the latter two are structure-based. Toxmatch also has the Verhaar scheme encoded. This is a set of structural rules that facilitates the grouping of chemicals into modes of action [15]. The Cluster option performs a clustering in the descriptor space into a predefined number of clusters.

The dialog for selecting a similarity method is shown on Figure 17. After the similarity method is selected, further information for the selected method, as shown on Figure 18

is provided. If it is a descriptor based method, the *Add* button allows more descriptors to be added from the list of available descriptors.

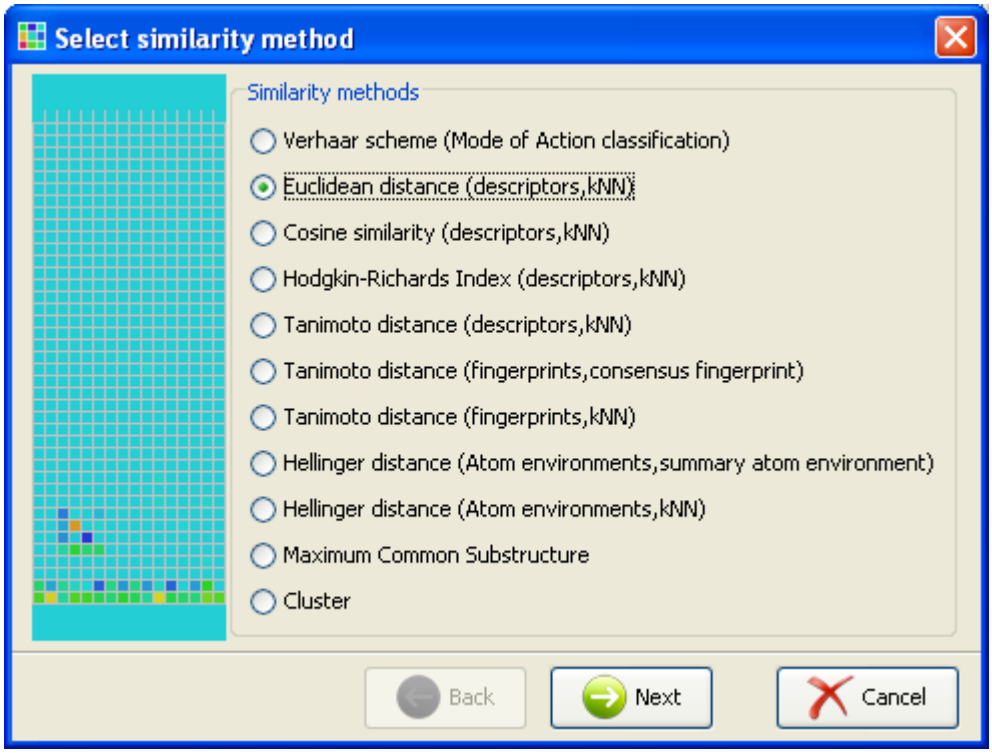


Figure 17: Similarity method selection

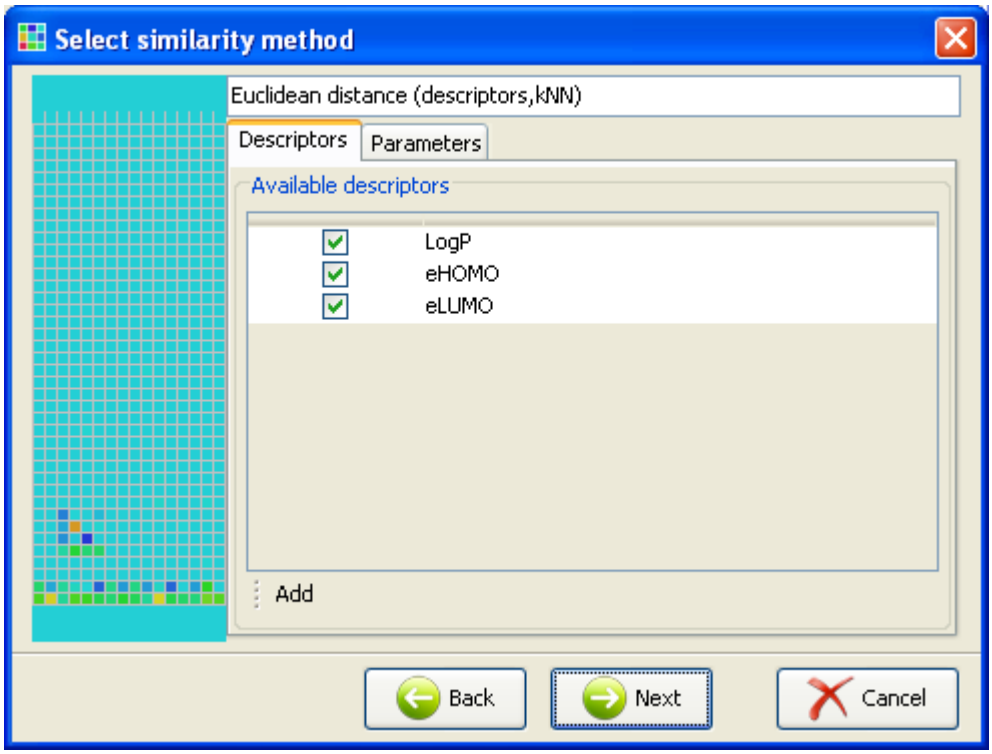


Figure 18: Similarity method configuration

Aside from calculating the similarity measure, the first option (*"Calculate similarity and predict activity"*) allows a prediction of activity to be made based on the nearest neighbours. The second option (*"Calculate similarity and dassify into groups")* enables a classification between groups (see Figure 19) to be made. Note that the estimation may take several minutes depending on the size and complexity of the dataset. When the processing has finished, a number of visualisation options will become available.

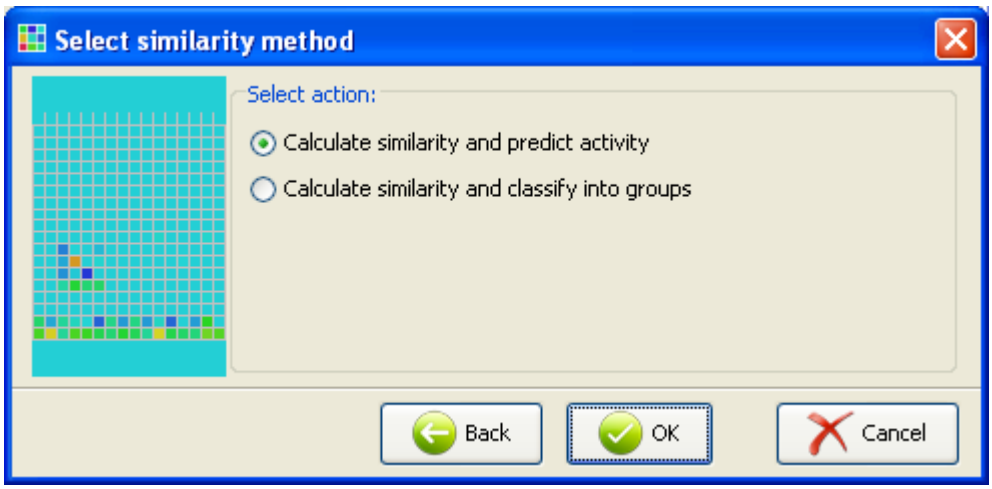


Figure 19: Further similarity options

The *"Similarity vs. activity"* radio button on the bottom of the scatter plot fills in existing similarity measures into the X drop down box (see Figure 20).

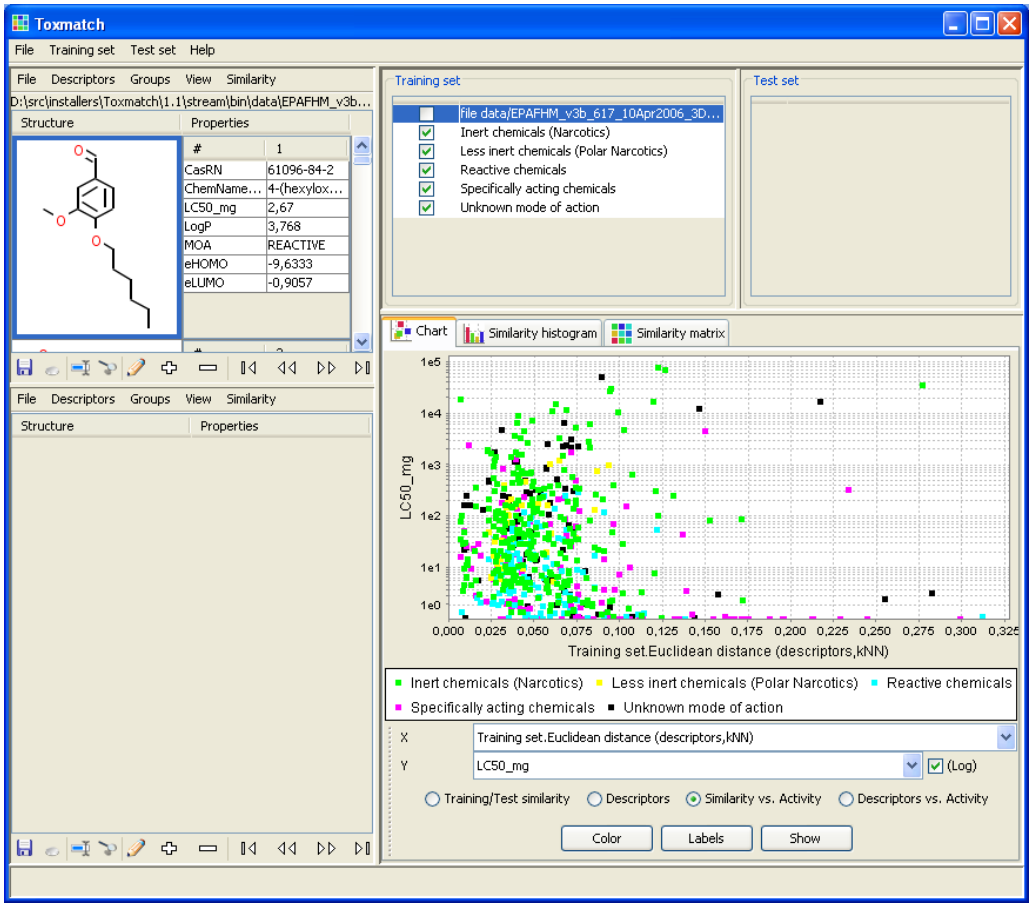


Figure 20: Similarity vs. activity plot

For descriptor-based similarity, the calculated distance is the averaged distance between the point and its *k* nearest neighbours. This is usually available under ***"Training set.Distance"*** (see Figure 20).

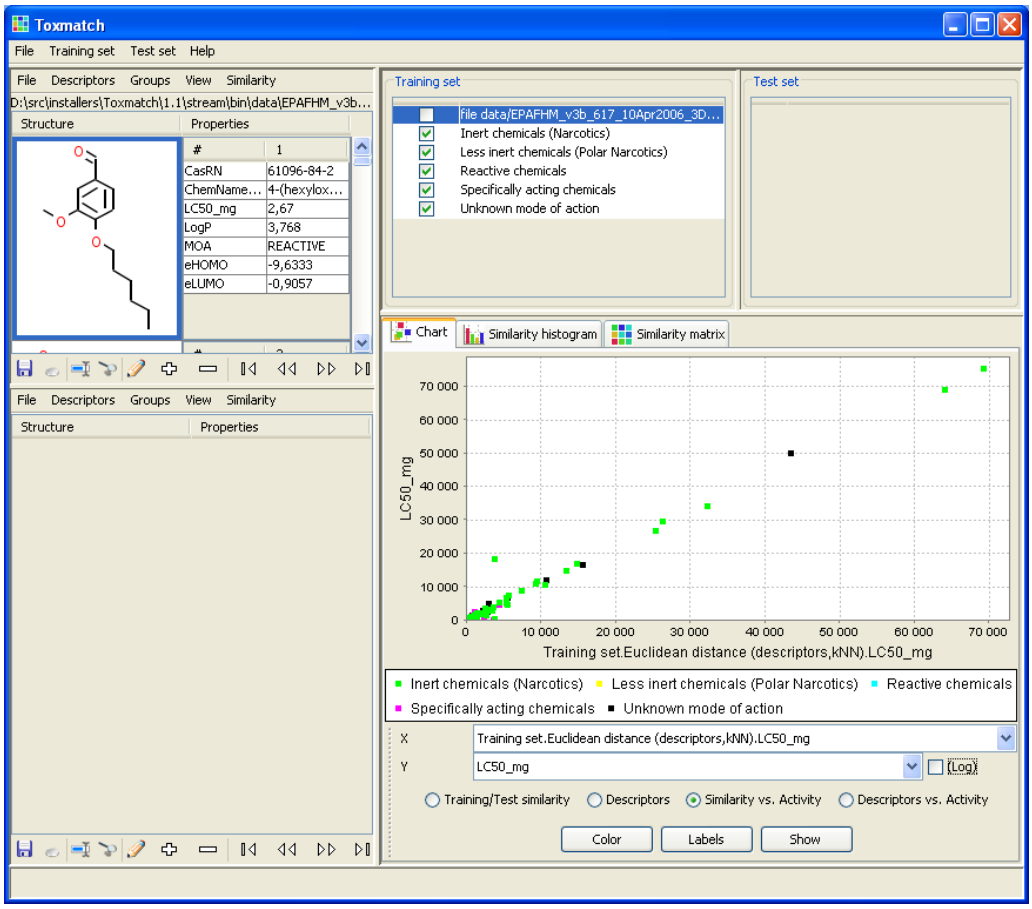


Figure 21: Predicted activity by kNN vs. measured activity plot

When the "Calculate similarity and predict activity" option is selected, the nearest neighbours are also used to estimate the activity. The result is found as a data field labelled as ***"Training set.Distance.Activity"*** (e.g. Training set.Euclidean distance.LC50_mg in Figure 21).

When the "Calculate similarity and classify into groups" option is selected, the similarity measure is used to perform classification between the groups i.e. each compound is assigned a probability of belonging to each group. These probabilities are located as ***"Training set.Distance.Group name"*** fields (e.g. "*Training set.Euclidean distance.Inert chemicals*") and can be displayed in both the chart and data areas. A scatter plot with these fields is displayed when "*Training/Test similarity*" radio button on the bottom of the chart is selected (see Figure 22).

The group with the highest probability is considered to be the one assigned and this is available in the *Training set.Distance.Group* field.

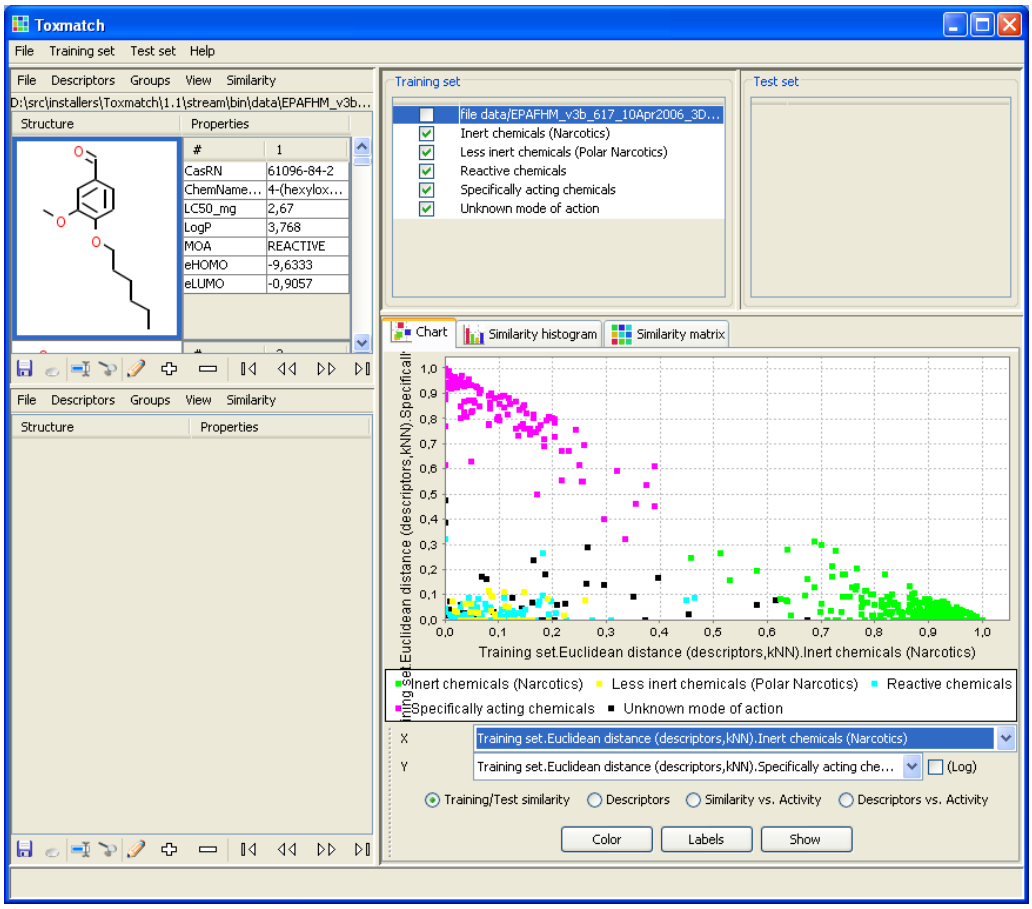


Figure 22: Training/Test similarity

The similarity to a test set option is only available if a test set has been already loaded and a similarity to the test set estimated by running the "Test set → Similarity → Similarity to the test set" operation.

Test set

A supported file type with test set data can be opened by using the "Test set → File → Open" menu. A multi-tab dialog for choosing identifiers and descriptors is presented, as shown in Figure 23 and Figure 24.

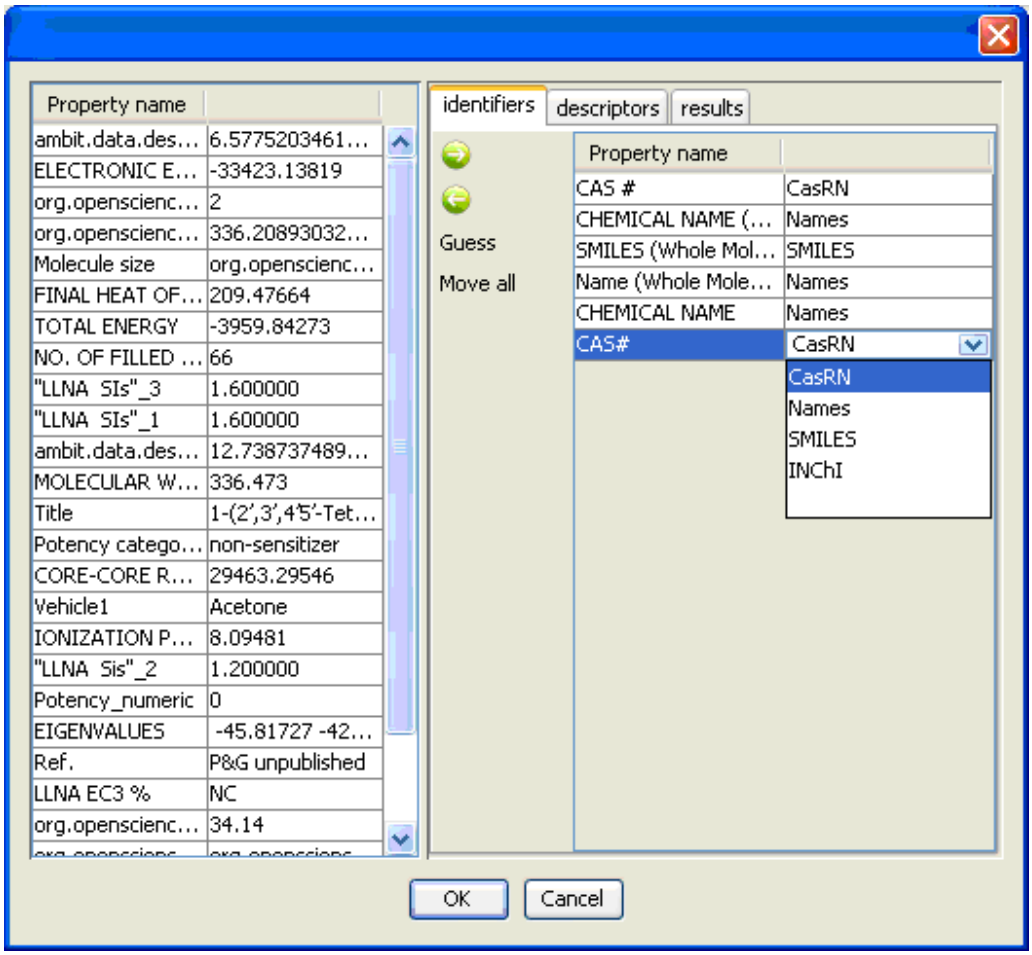


Figure 23: Identifiers selection tab

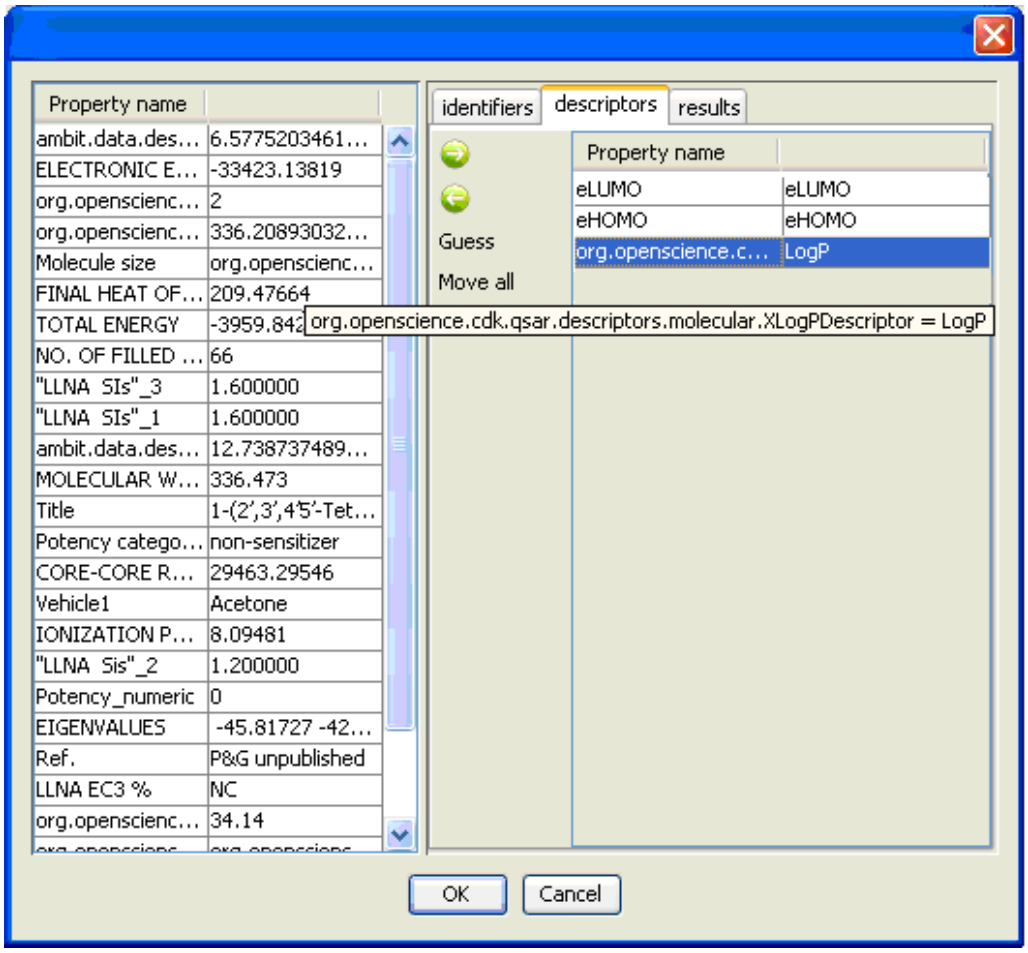


Figure 24: Descriptors selection tab

The dataset, chemicals together with their properties and calculated similarity indices may be exported in a variety of formats e.g. *.sdf*, *.csv*, *.xls* or *.txt*, by using the "Test set → File → Save" menu.

Test set descriptors, groups and views menus/actions are the same as those used for training sets (as already described in previous sections of the manual).

The *Similarity to the training set* option is only available if both training and test sets have been loaded and a similarity to the training set estimated by selecting the "Training set → Similarity → Similarity to the training set" menu. A choice between available similarity measures is offered, as shown in Figure 25. Each compound from the test set will be assigned values for the corresponding similarity measure.

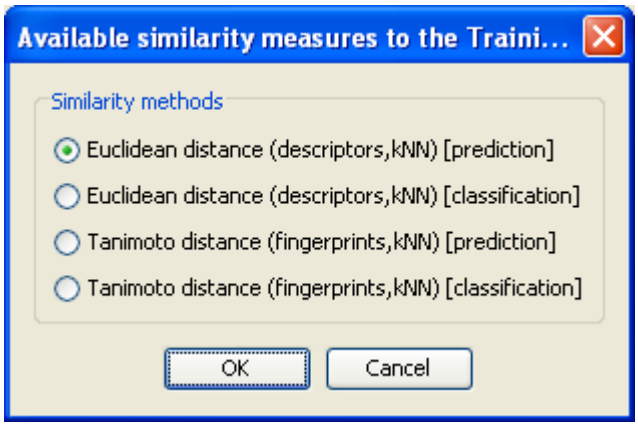


Figure 25: Available similarity measures to the training set

Similarity to the test set is essentially the same as *Similarity to the training set*, but here the similarity is calculated with respect to the test set. Note that since the data field, representing the measured activity is usually missing in test sets, predicting activity value or performing classification with respect to the test set groups may not be available. The recommended use of this option is to calculate similarity to the entire test data set and then to visualise *Test set.Similarity* vs. *Training set.Similarity*. This allows a user to make a comparison between the training and test sets as shown in Figure 26.

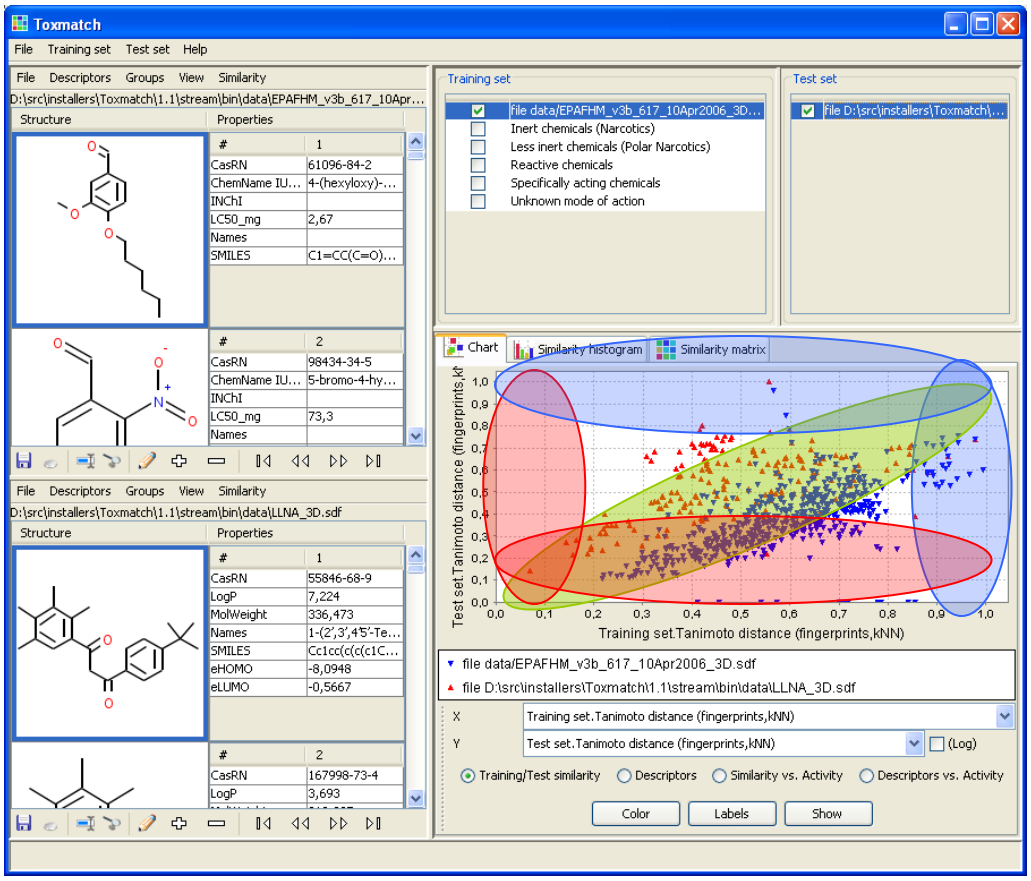


Figure 26: Similarity to the training set vs. similarity to the test set. (Euclidean distance)

One possible interpretation of the training set vs. test set comparison plot is illustrated in Figure 27.

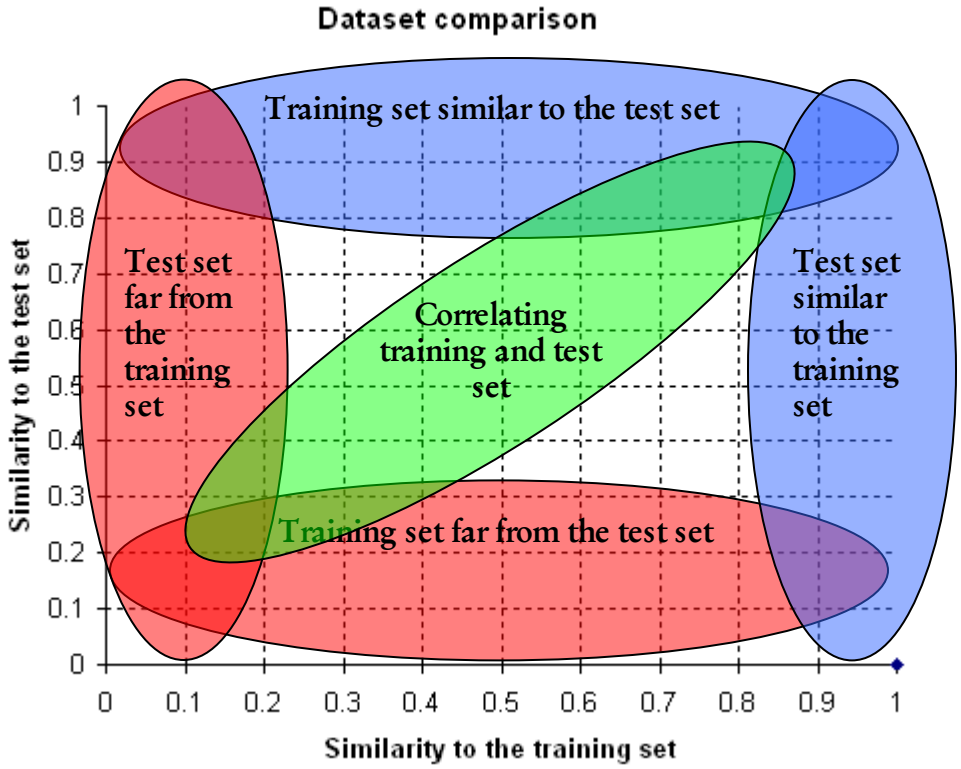


Figure 27: The interpretation of training / test comparison plot

Data area

The data area is where compounds are displayed. The "Training set → View → View" menu allows the user to switch between two views, as shown in Figure 28 and Figure 29 respectively.

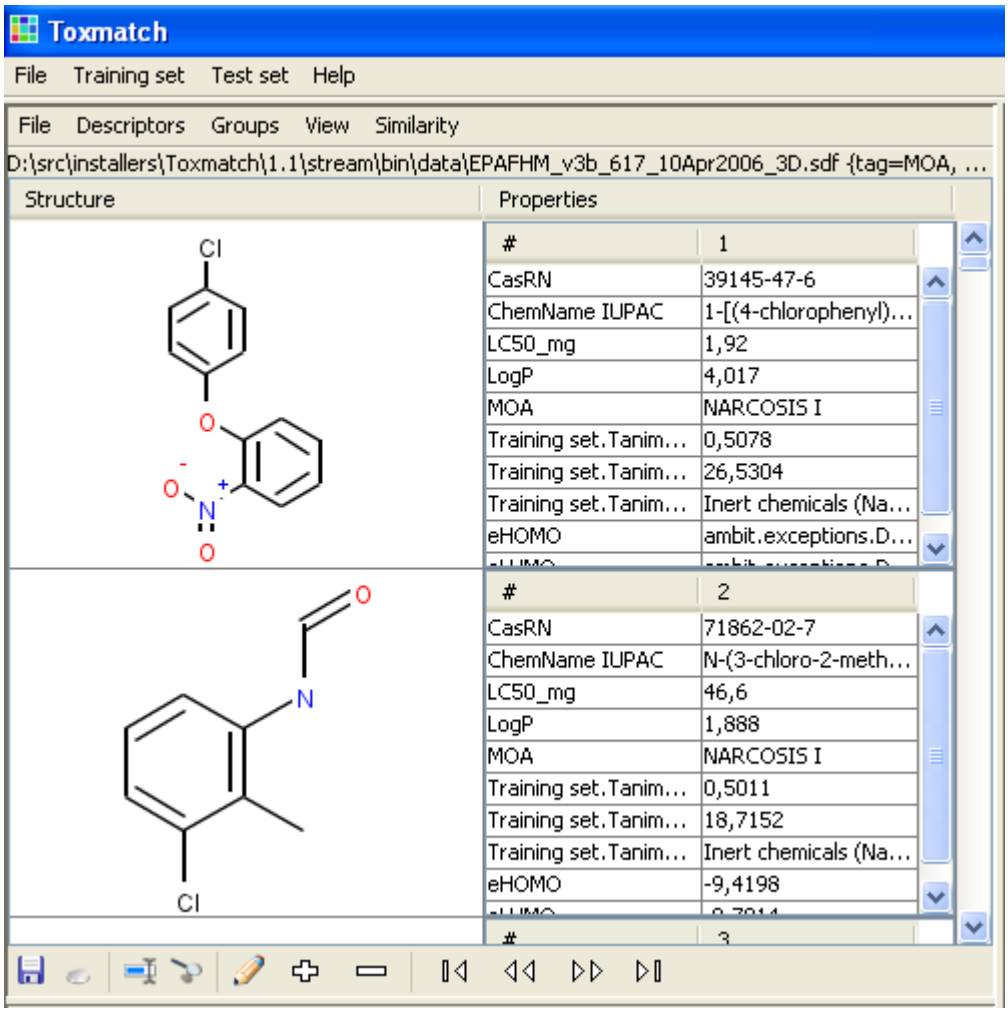
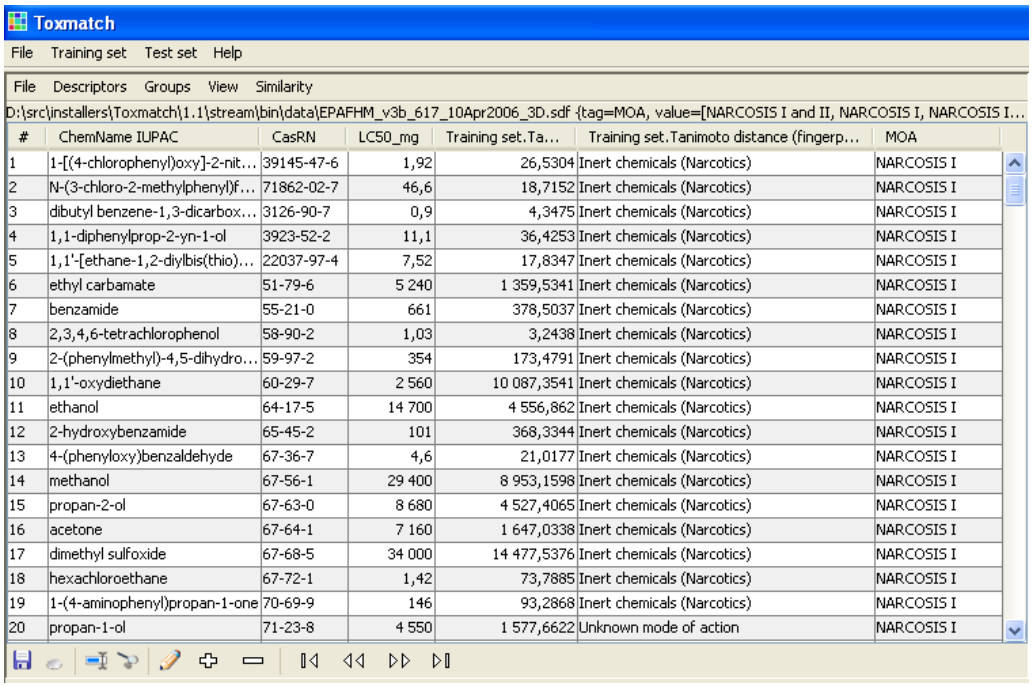


Figure 28: Structure view



#	ChemName IUPAC	CasRN	LC50_mg	Training set.Ta...	Training set.Tanimoto distance (fingerp...	MOA
1	1-[(4-chlorophenyl)oxy]-2-nit...	39145-47-6	1,92	26,5304	Inert chemicals (Narcotics)	NARCOSIS I
2	N-(3-chloro-2-methylphenyl)f...	71862-02-7	46,6	18,7152	Inert chemicals (Narcotics)	NARCOSIS I
3	dibutyl benzene-1,3-dicarbox...	3126-90-7	0,9	4,3475	Inert chemicals (Narcotics)	NARCOSIS I
4	1,1-diphenylprop-2-yn-1-ol	3923-52-2	11,1	36,4253	Inert chemicals (Narcotics)	NARCOSIS I
5	1,1'-[ethane-1,2-diylbis(thio)...	22037-97-4	7,52	17,8347	Inert chemicals (Narcotics)	NARCOSIS I
6	ethyl carbamate	51-79-6	5 240	1 359,5341	Inert chemicals (Narcotics)	NARCOSIS I
7	benzamide	55-21-0	661	378,5037	Inert chemicals (Narcotics)	NARCOSIS I
8	2,3,4,6-tetrachlorophenol	58-90-2	1,03	3,2438	Inert chemicals (Narcotics)	NARCOSIS I
9	2-(phenylmethyl)-4,5-dihydro...	59-97-2	354	173,4791	Inert chemicals (Narcotics)	NARCOSIS I
10	1,1'-oxydiethane	60-29-7	2 560	10 087,3541	Inert chemicals (Narcotics)	NARCOSIS I
11	ethanol	64-17-5	14 700	4 556,862	Inert chemicals (Narcotics)	NARCOSIS I
12	2-hydroxybenzamide	65-45-2	101	368,3344	Inert chemicals (Narcotics)	NARCOSIS I
13	4-(phenyloxy)benzaldehyde	67-36-7	4,6	21,0177	Inert chemicals (Narcotics)	NARCOSIS I
14	methanol	67-56-1	29 400	8 953,1598	Inert chemicals (Narcotics)	NARCOSIS I
15	propan-2-ol	67-63-0	8 680	4 527,4065	Inert chemicals (Narcotics)	NARCOSIS I
16	acetone	67-64-1	7 160	1 647,0338	Inert chemicals (Narcotics)	NARCOSIS I
17	dimethyl sulfoxide	67-68-5	34 000	14 477,5376	Inert chemicals (Narcotics)	NARCOSIS I
18	hexachloroethane	67-72-1	1,42	73,7885	Inert chemicals (Narcotics)	NARCOSIS I
19	1-(4-aminophenyl)propan-1-one	70-69-9	146	93,2868	Inert chemicals (Narcotics)	NARCOSIS I
20	propan-1-ol	71-23-8	4 550	1 577,6622	Unknown mode of action	NARCOSIS I

Figure 29: Table view

The fields displayed in table view could be selected by "Training set → View → Fields" menu.

The buttons on the bottom of the data area have the following functionality:

- Save currently displayed dataset.
- Dataset information – displays information about the currently displayed dataset (see Figure 30);
- Go to record number (entered by user) – asks for a record number and makes it the current one;
- Search by CAS RN – asks for CAS RN and makes the record with this CAS number the current one (if such record exists);
- Edit current structure – launches structure diagram editor for editing the current structure (see Figure 31).
- Append empty record – appends an empty record;
- Delete current record – deletes the current record;
- First record – selects the first record;

- Previous record
- Next record
- Last record – selects the last record;

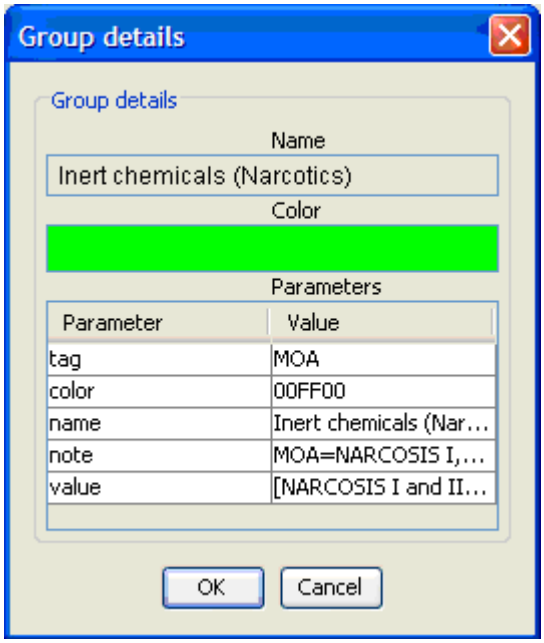


Figure 30: Dataset information

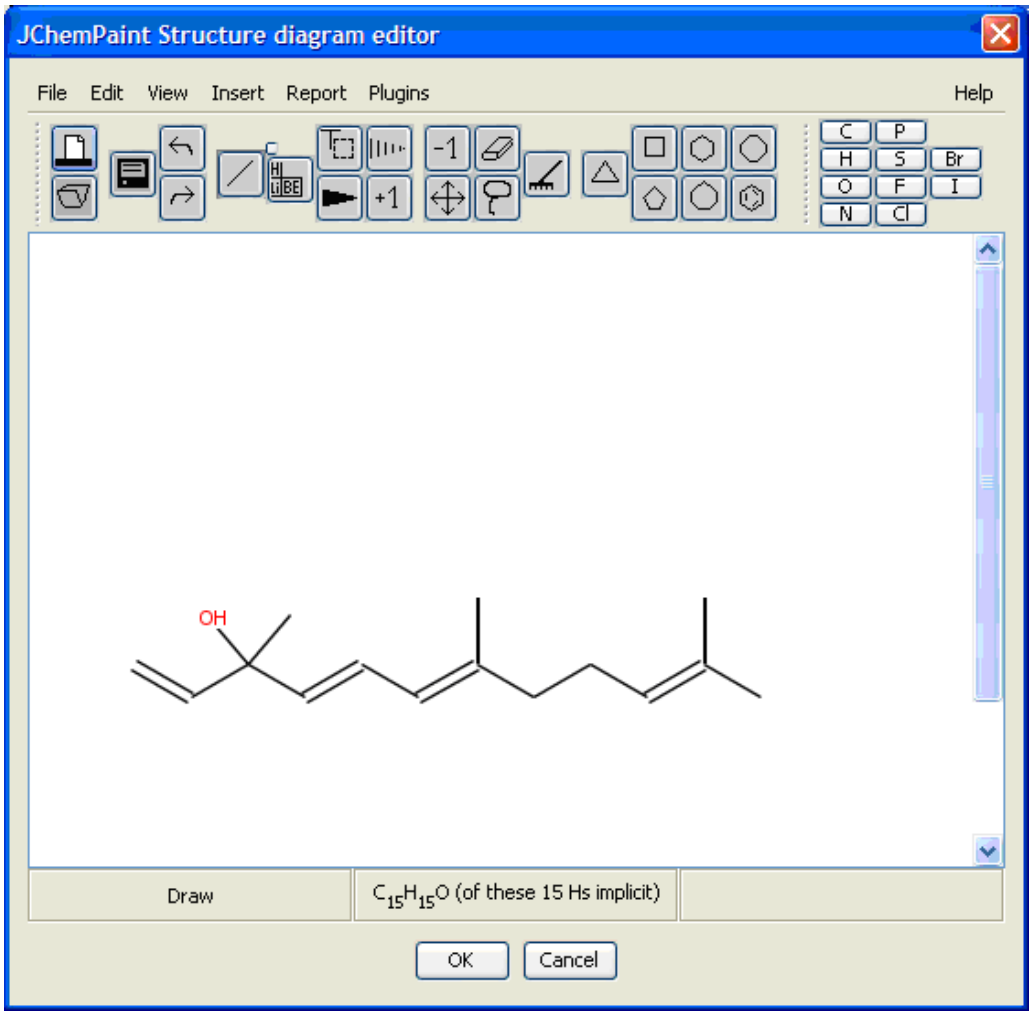


Figure 31: Structure diagram editor

Chart

The chart can display data points right after a file is loaded. By default all points from the loaded dataset are displayed. The descriptors on *X* and *Y* axis can be selected by the corresponding drop down boxes. The *Y* axis can be toggled between logarithmic and normal mode. The chart has several display modes, controlled by radio buttons at the bottom (see Figure 32):

- Training /Test similarity – clicking on this button fills the X and Y drop down boxes with available similarity indices names. Note: similarity names only become available after launching certain similarity calculations;

- Descriptors – clicking on this button fills the X and Y drop down boxes with available descriptor names;
- Similarity vs. Activity – clicking on this button fills the X drop down box with available similarity indices names and the Y box – with the endpoint name;
- Descriptors vs. Activity - clicking on this button fills the X drop down box with available descriptor names and the Y box - with the endpoint name;
- Colour button – clicking this button displays drop down box with available similarity indices names plus default "*groups*" entry. The later is used to colour the points with the colour defined for each group;
- Labels button – clicking this button brings up a list of possible labels that can be displayed for each point;
- Show button - allows displaying all visible points in the training or test dataset area, correspondingly. This functionality is useful when the chart has been zoomed and allows displaying a subset of points.

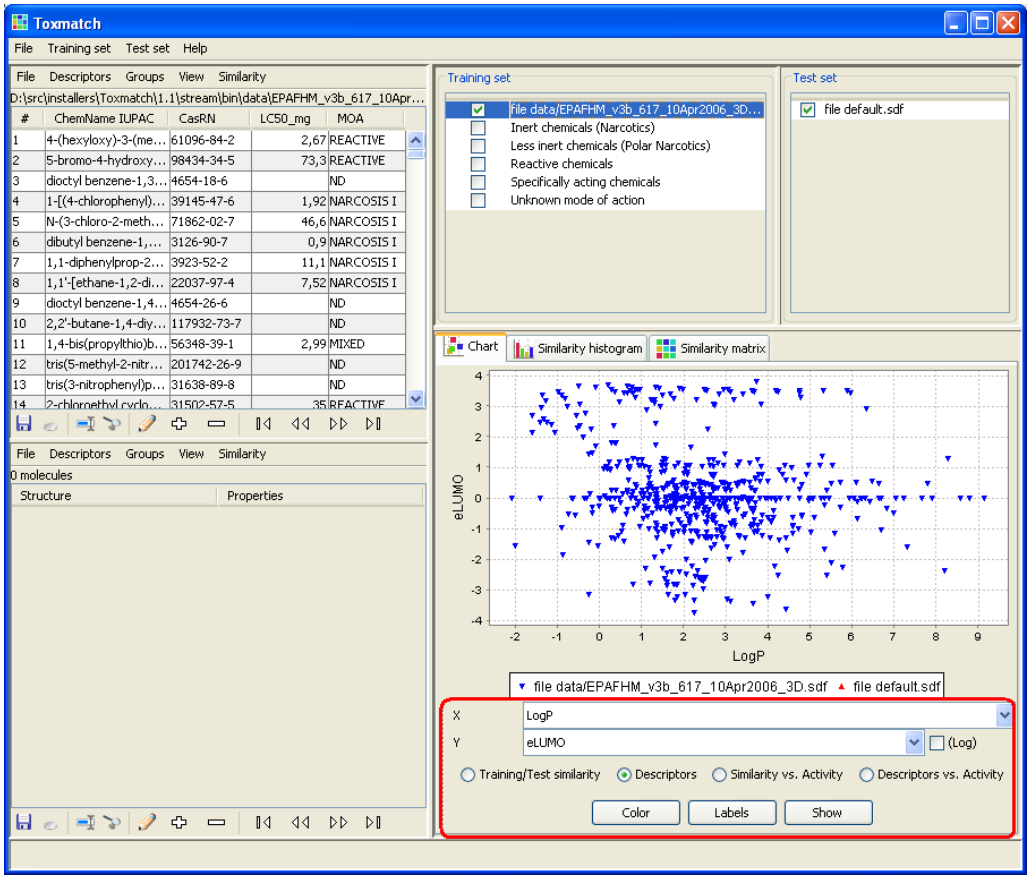


Figure 32: Scatter plot of the training data set for Aquatic toxicity, X=LogP, Y=eLUMO

The *"Training set groups"* area allows toggling on or off the visibility of the corresponding groups on the chart. Additionally, clicking on a group will display all compounds from this group into data set area on the left.

Clicking on a point displays all the compounds of the corresponding group in the data set area on the left, whilst the selected compound corresponds to the clicked point.

Similarity matrix

The similarity matrix displays pair wise similarity of the last calculated similarity measure between selected datasets (or groups). The colour range is from blue (pair wise similarity value =0) to red (pair wise similarity value =1). Note that the same structures will have value of 0 when using a distance-based similarity (e.g. Euclidean distance), but a value of 1 when using a similarity index such as Tanimoto coefficient. A left button mouse click on a cell will display the compound of the selected row. Aright button mouse click on a cell will display the compound of the selected column. The content of the similarity matrix is configured by the *rows* and *columns* drop down boxes, found above the matrix:

- *Rows: Training set* – the rows of the similarity matrix correspond to the selected group/dataset of the training set. The selected group can be changed by clicking a row on the *training data groups area* above the similarity matrix;
- *Columns: Training set* – the columns of the similarity matrix correspond to the selected group/dataset of the training set. The selected group can be changed by clicking a row on the *training data groups area* above the similarity matrix;
- *Rows: Test set* – the rows of the similarity matrix correspond to the selected group/dataset of the test set. The selected group can be changed by clicking a row on the *test data groups area* above the similarity matrix;
- *Columns: Training set* – the columns of the similarity matrix correspond to the selected group/dataset of the test set. The selected group can be changed by clicking a row on the *test data groups area* above the similarity matrix.

If rows and columns correspond both to the training set or both to the test set, then the diagonal matrix is displayed (see Figure 33).

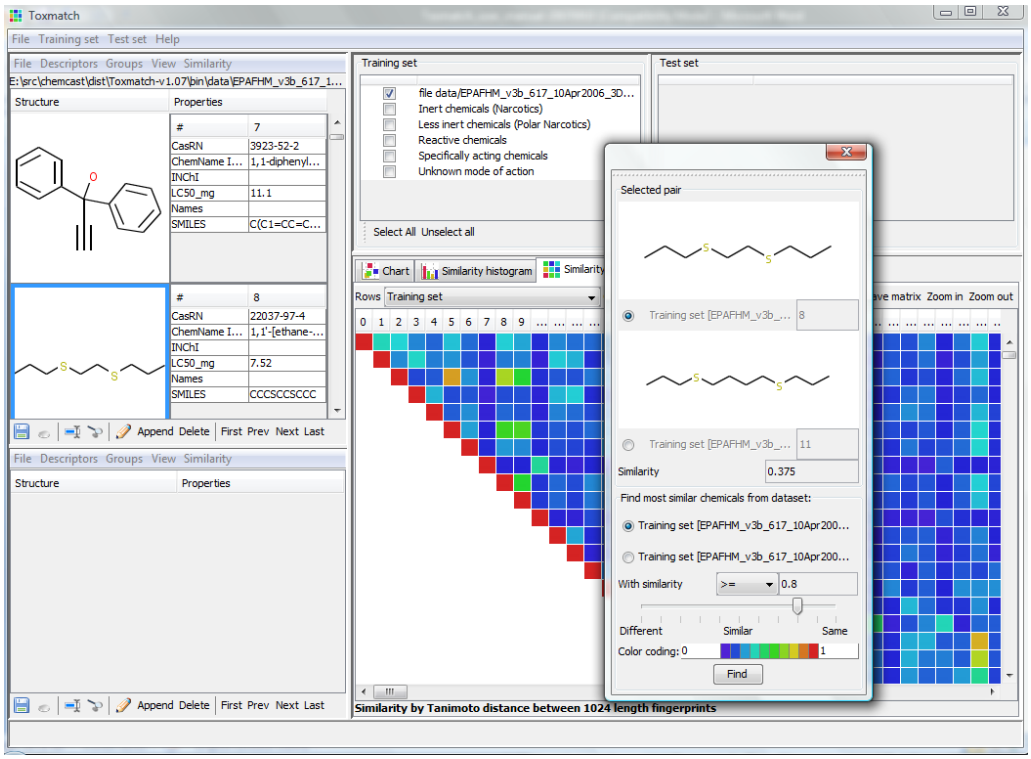


Figure 33: Pair wise similarity between compounds from group "Inert chemicals (narcotics)"

When rows and columns correspond to different sets (training set and test set), the full matrix is displayed.

The *Save matrix* button allows the matrix to be saved as a picture (*.png* file) or as a spreadsheet (*.csv* file).

File processing

The "File → File processing" menu option performs batch calculation of descriptors. Compounds are read from a file and the results are stored in another file. The results file can then be later loaded as training or test sets.

Help

The "Help → About" menu option displays information about application's version as shown on Figure 34.

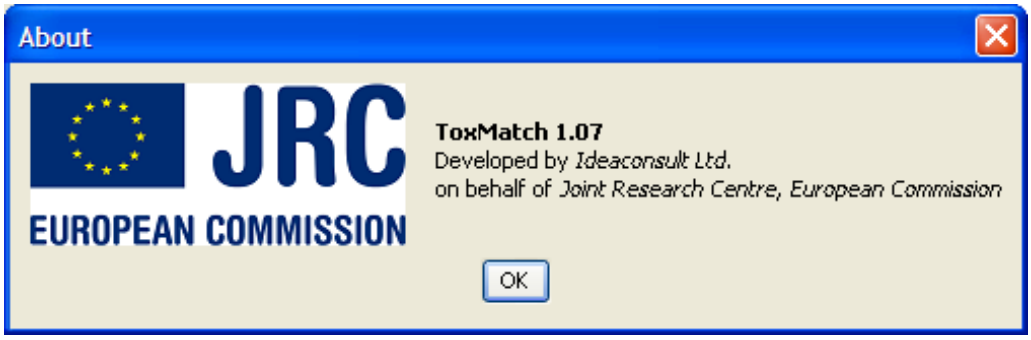


Figure 34: Application version

Configuration

The "Help → Configuration" menu option provides graphical user interface for editing Toxmatch configuration file. Toxmatch configuration is an XML file, containing information about the datasets, endpoints, similarity measures and groups within datasets.

The configuration is organized as a tree. There are four top level entries, namely [property], [similarity], [descriptors] and [endpoints]. Endpoints configuration is of main interest to end users. An endpoint entry consists of configuration for the training set ("[trainingset]"), where the relevant file is specified, as well as the field, containing the assay data "[result]". Multiple groups, splitting the dataset by different criteria can be defined.

The recommended way for editing is via copy/paste. To create a new endpoint entry, click on existing one and use menu *Edit/Copy*. Then click on the top-level [endpoints] entry and use *Edit/Paste*. This will create new endpoint tree, which can be edited to modify name, training dataset and groups.

Modifications are not saved, unless *File/Save* menu option is used. The configuration changes will be only visible after Toxmatch is restarted.

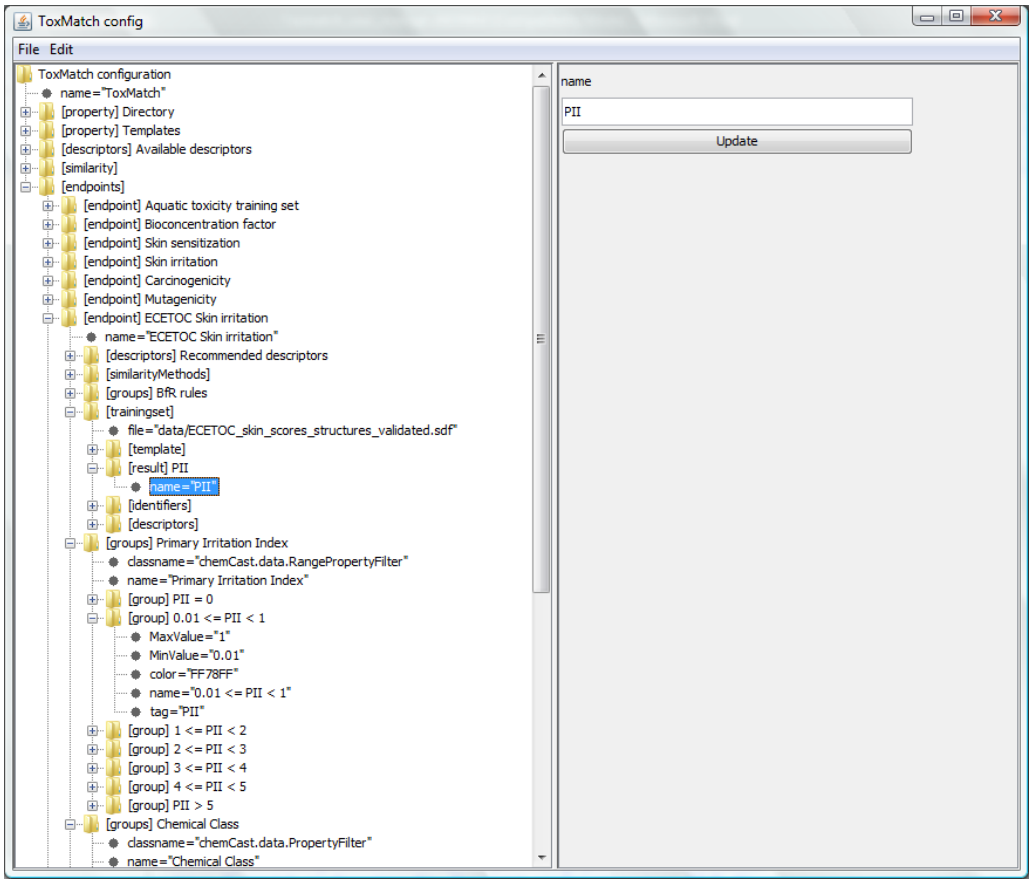


Figure 35: Toxmatch configuration

Exit

The “File Exit” menu option quits the application.

Case study – BCF read across

This case study describes a read across for a chemical with respect to the bioconcentration factor endpoint. The approach will be performed in 2 stages. Firstly a set of source data (the BCF training set within Toxmatch) will be characterised using one or other similarity method. In this case, a weighted similarity of nearest neighbours will be carried out using 3 methods: Euclidean distance in descriptor space, Tanimoto distance on 1024 bit length fingerprints and Hellinger distance on atom environments (circular fingerprints).The second stage will focus on the target chemical (the test set) and compare it with the most similar training set chemicals to arrive at an estimate for BCF score.

The approach will be outlined as a series of procedural steps.

Load training set

First select BCF training set from *File/Predefined training sets* menu.

Follow the wizard as below (Figure 36):

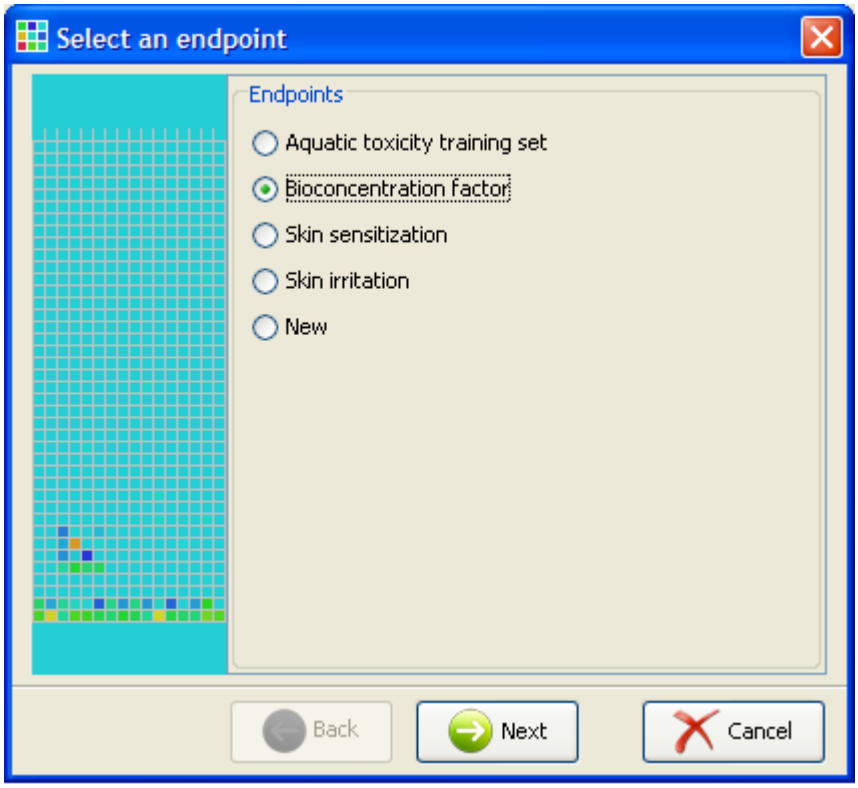


Figure 36: Endpoints selection

Whilst the training set can be divided into different groups, here a classification procedure will be carried out to assign query chemicals into specific LogBCF ranges.

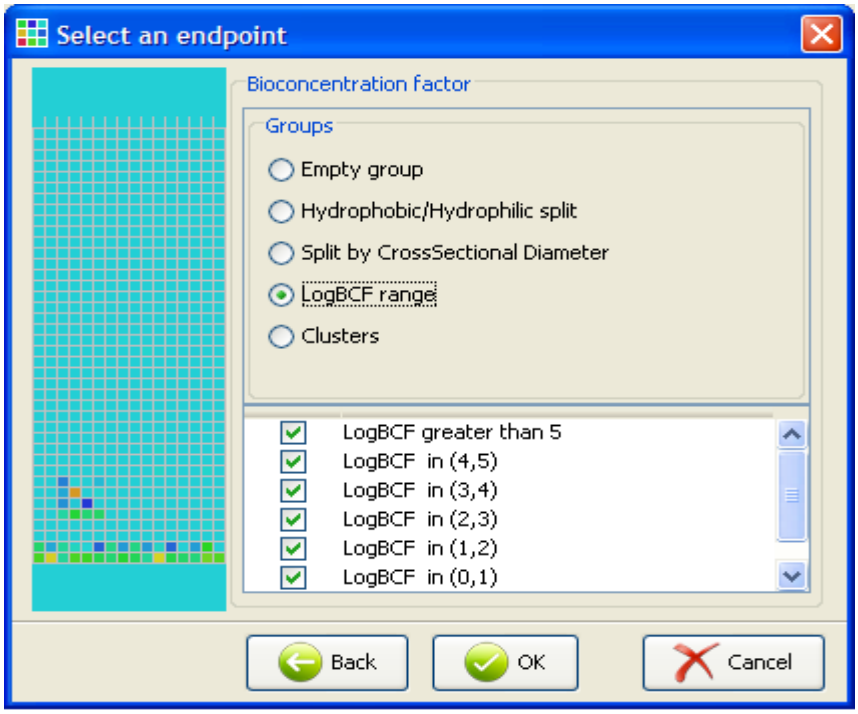


Figure 37: Groups selection

After the dataset is loaded, the Toxmatch screen should look as in Figure 38.

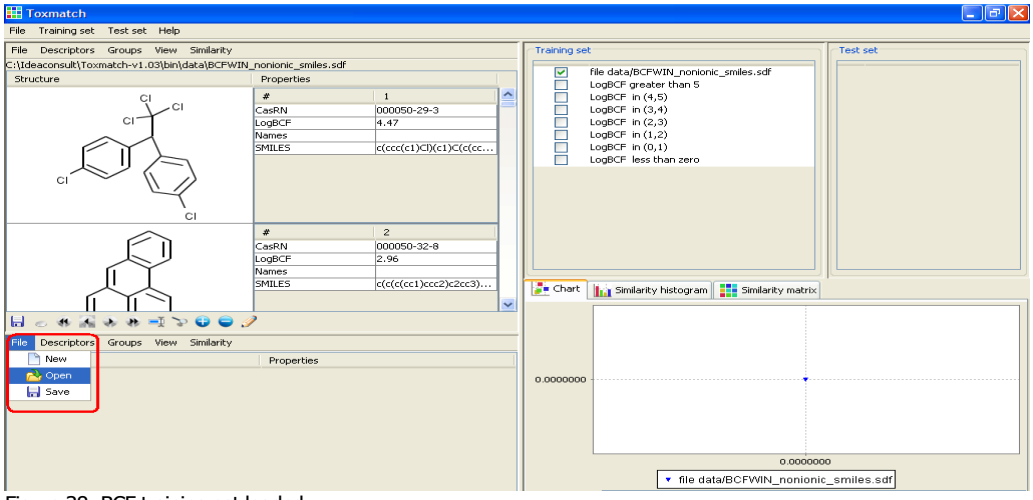


Figure 38: BCF training set loaded

Load test set

The next step is to load the test set. In this case, a single query or target chemical will be used. The details of which are given below. A query chemical can be introduced

into Toxmatch in a variety of ways as previously described. Here the text from Figure 39 should be copied into a text editor or into MS Excel™ and saved as 66-25-1.csv file.

```
CasRN,SMILES,NSC,XLogPDescriptor,CrossSectionalDiameterDescriptor
[Angstrom],MaximumDiameterDescriptor [Angstrom],WeightDescriptor

66-25-1,O=CCCCCC,2596,1.753,2.4897,8.1759,100.0888
```

Figure 39: Prepare *.csv* file with text editor

Note if using MS Excel™, insert the content as in Figure 40 and use File/Save As to save as CSV (Comma delimited) (*.csv) file.

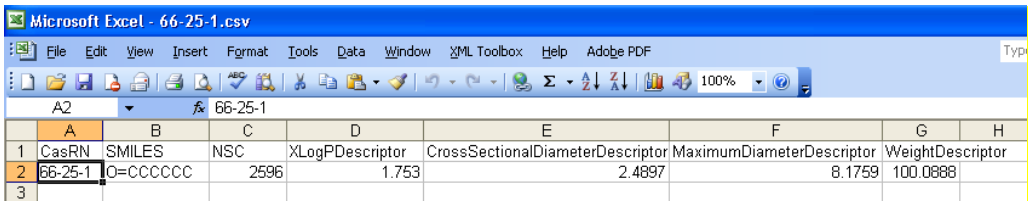


Figure 40: Prepare *.csv* file with MS Excel™

Use the *File/Open* menu from the bottom panel to load the test set.

A window will appear (Figure 41), requesting field names to be specified. If the descriptors in the file have different names compared with the descriptors in the original training set then a correspondence needs to be done. For example: "*Eff Diam*" is a new name for "*CrossSectionalDiameter*"

Once field names have been assigned click *OK* to load the query chemical into the test set panel.

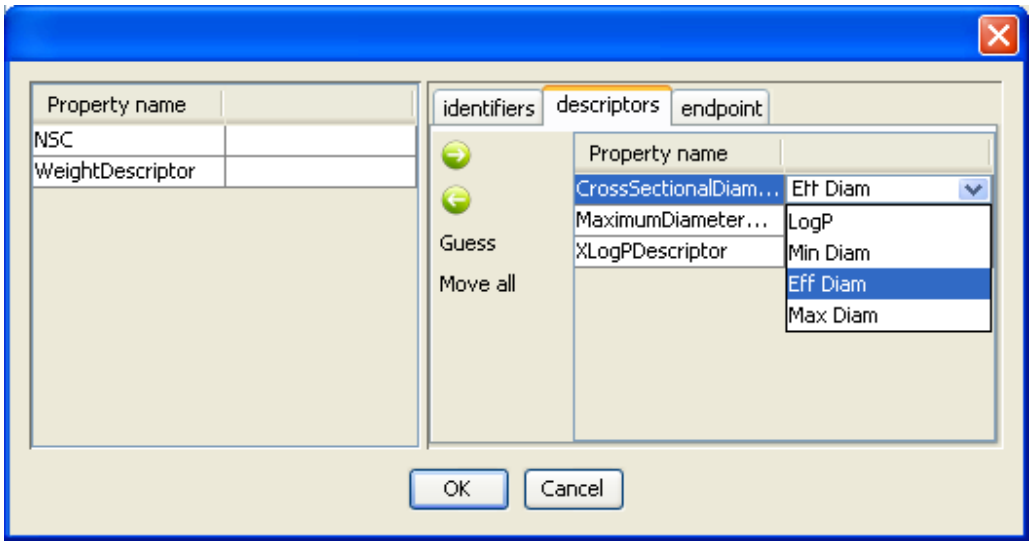


Figure 41: Descriptors selection

Exploring the descriptor space

The descriptor space can be explored using the chart functionality. Click on *Descriptor* radio button to select *X* and *Y* descriptors from the available list (Figure 42).

In Figure 42, *Eff Diam* and *LogP* have been selected as the X and Y descriptors respectively. The plot shows the scatter distribution of these two descriptors for both the training set (blue) and test set (red) chemicals.

In this descriptor space the query chemical appears in a region that is not densely populated.

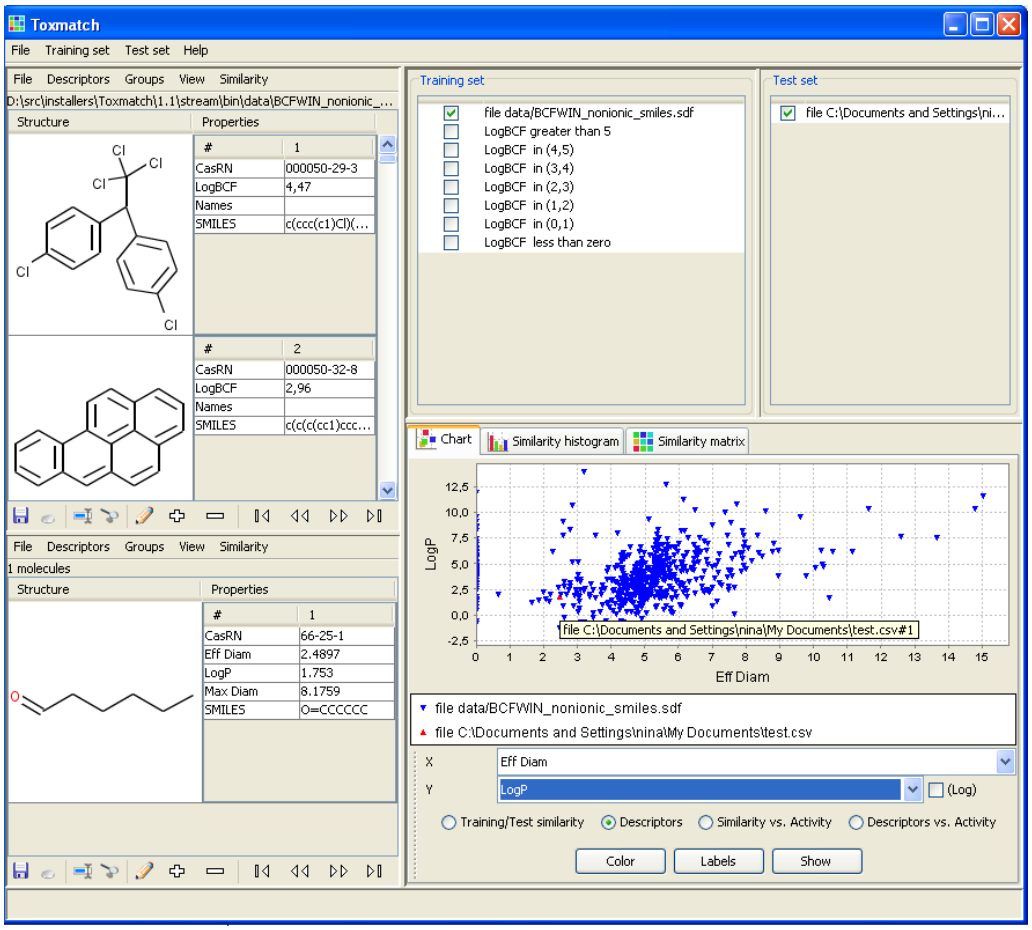


Figure 42: Exploring descriptor space with scatter plot

Regions of the chart may be enlarged by using the zoom in/out functionality. The mouse can be used in a click and drag mode to select a region of descriptor space to be viewed in greater detail as shown in Figure 43.

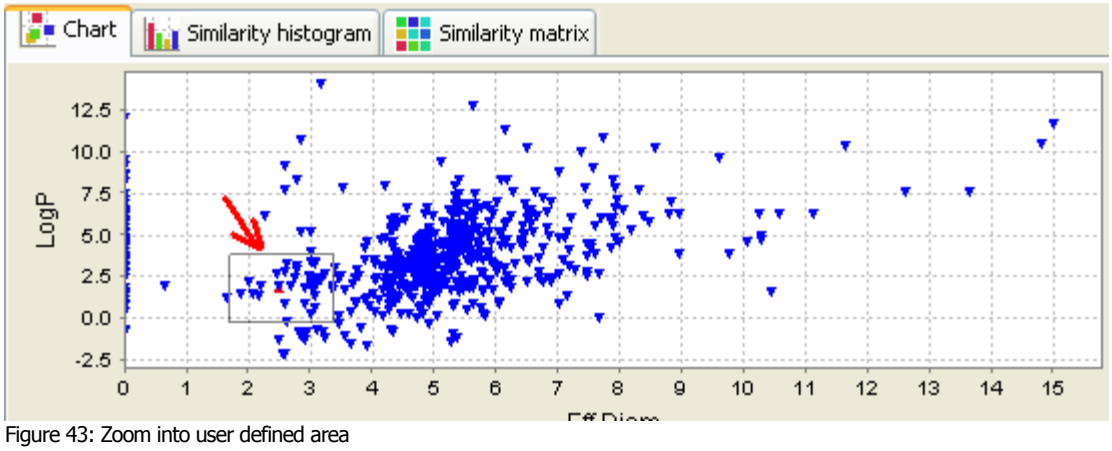


Figure 43: Zoom into user defined area

The points in the plot can also be labelled if desired by using the *Labels* button (at the right bottom of the chart). Available options are descriptors, endpoint, CAS RN, Names (Figure 44).

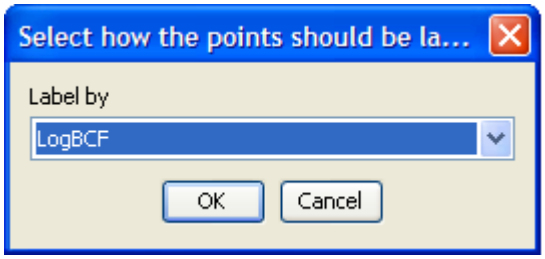


Figure 44: Labels

The action shown in Figure 44 results in each point being labelled by the LogBCF value (Figure 45). Clicking on a point will display the corresponding compound. Clicking on the *Show* button will display all the visible points in the left dataset area.

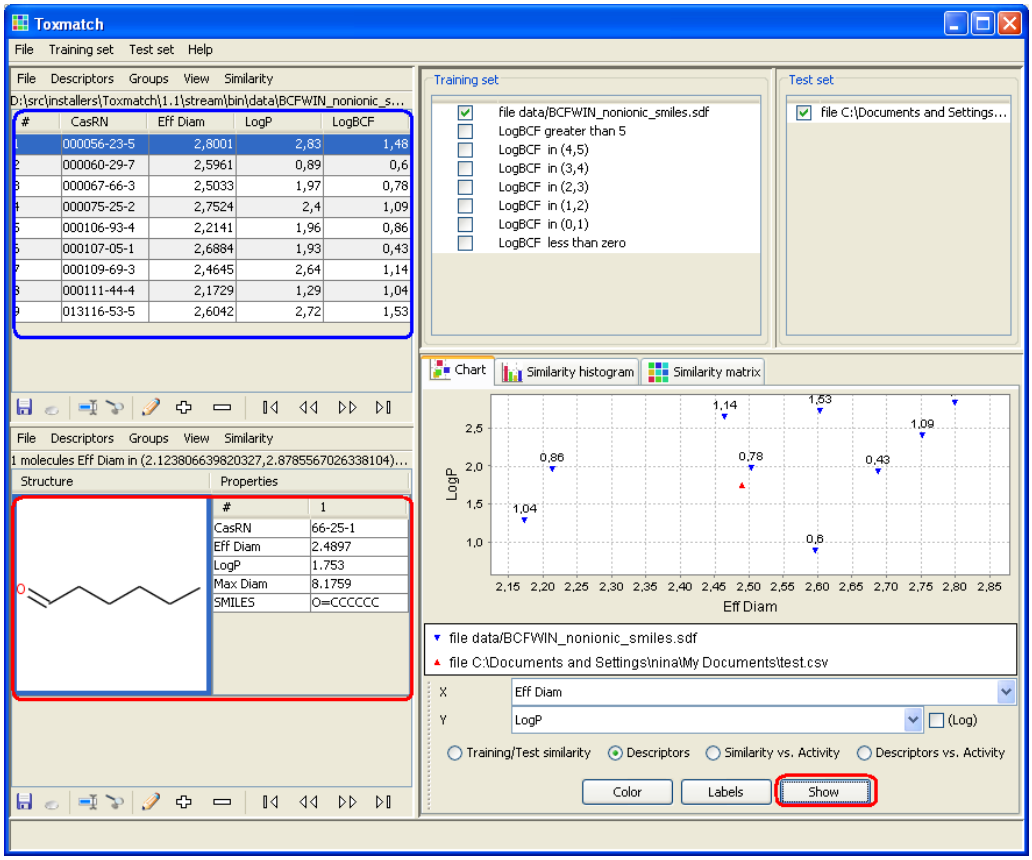


Figure 45: Zoomed area with LogBCF as labels

Similarity in descriptor space – Euclidean distance

The next step is to calculate similarity to the training set. Use the training (top) panel menu *Similarity/Similarity to the training set* (Figure 46) and follow the wizard.

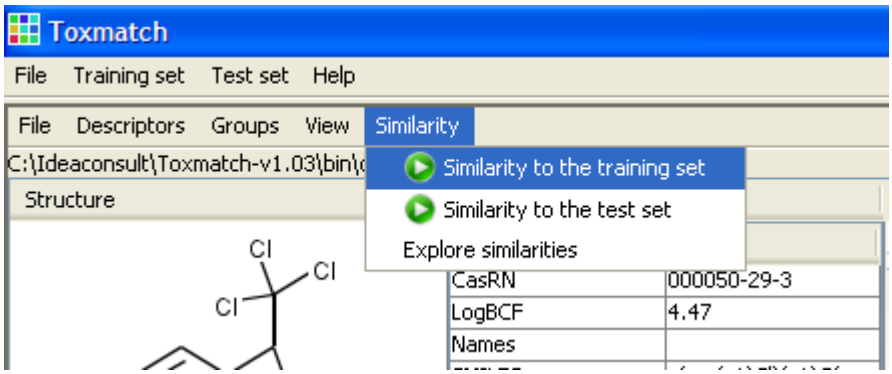


Figure 46: The Similarity menu for the training set (top panel)

Select *Euclidean distance* (Figure 47)

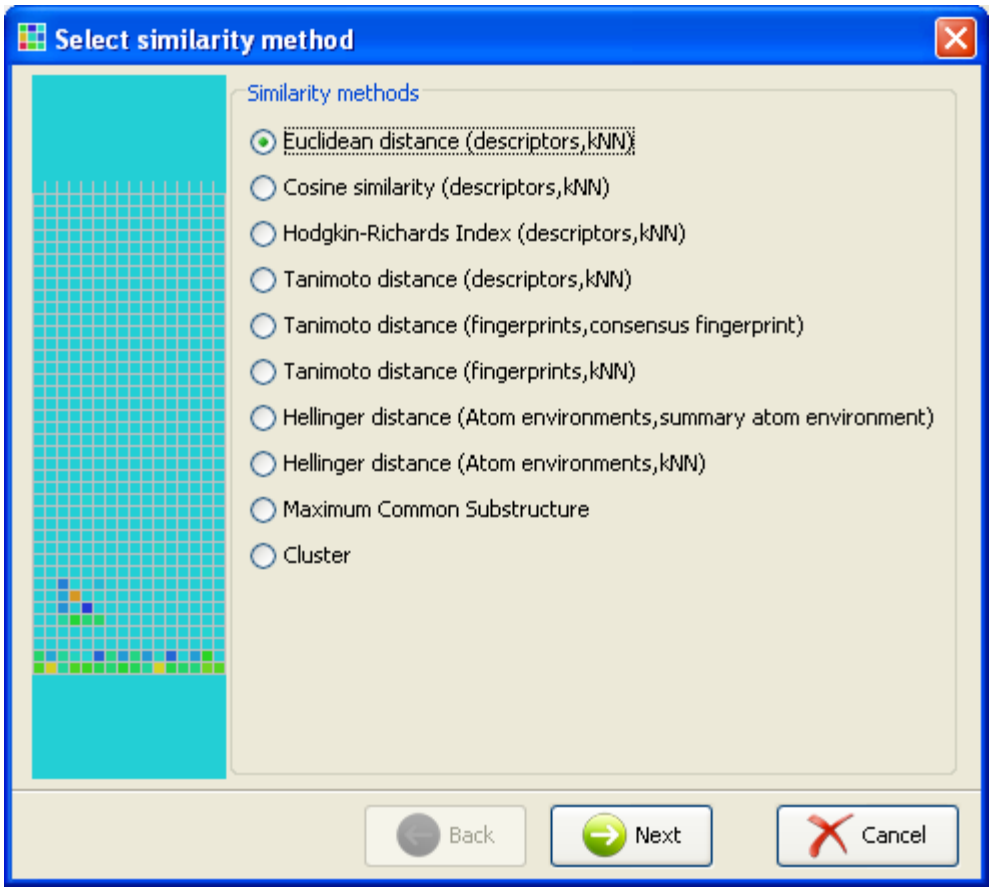


Figure 47: Similarity methods

Click *Next* on the following window (Figure 48)

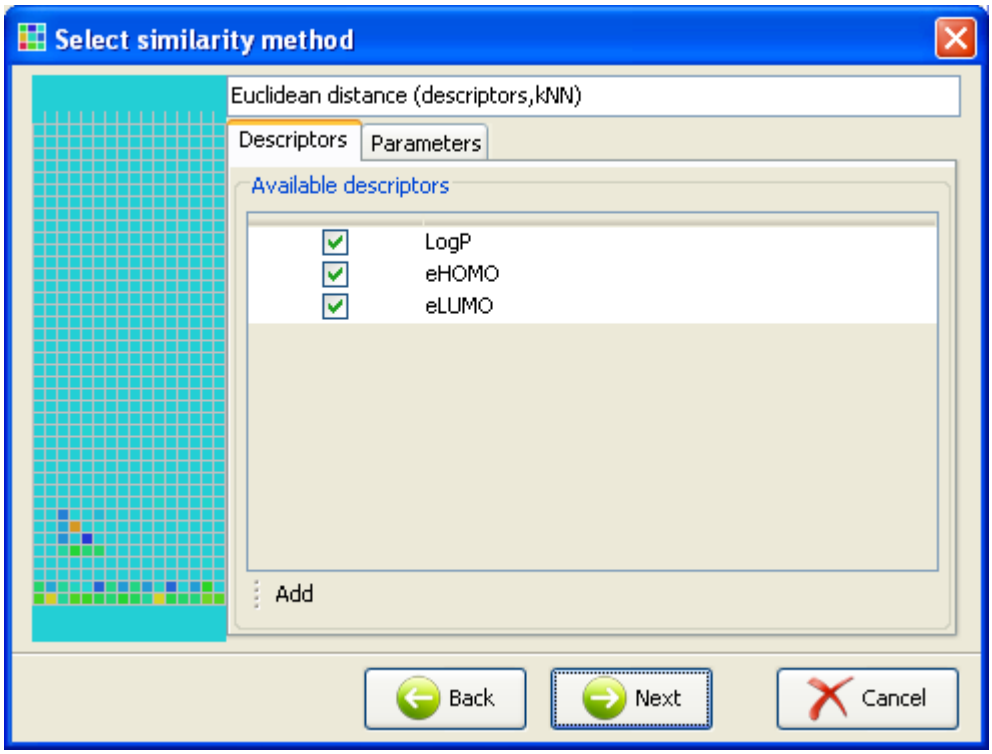


Figure 48: Similarity method options

Select *Calculate similarity and predict activity* on the following window (Figure 49) and click OK.

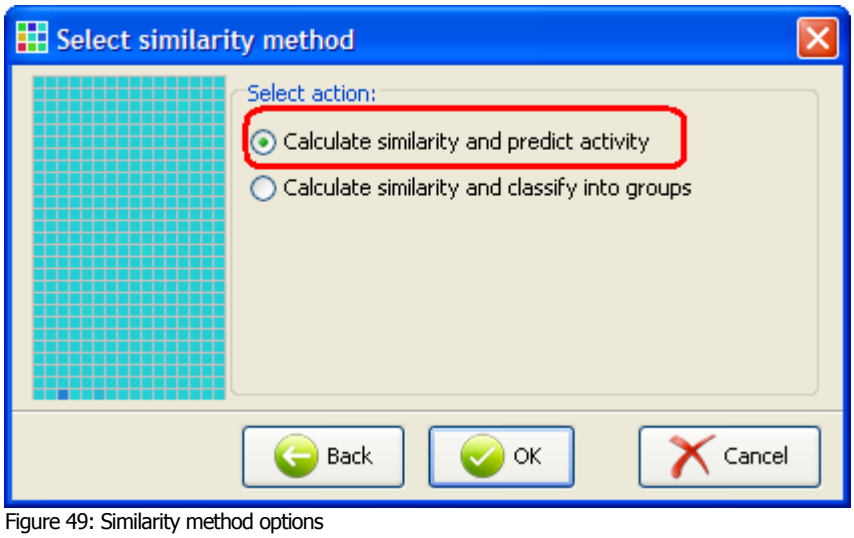


Figure 49: Similarity method options

This will run a calculation of Euclidean distance for the training set. The Euclidean distances will be available in the *Training set.Euclidean distance* field, whilst the predicted LogBCF values will be available in the *Training set.Euclidean distance.LogBCF* field.

Use menu *View/View Fields* from the top panel to select which fields will be visualised. Switch to the similarity tab and select *Training set.Euclidean distance* and *Training set.Euclidean distance.LogBCF* (Figure 50). Switch to descriptors panel and select all descriptors. Click OK.

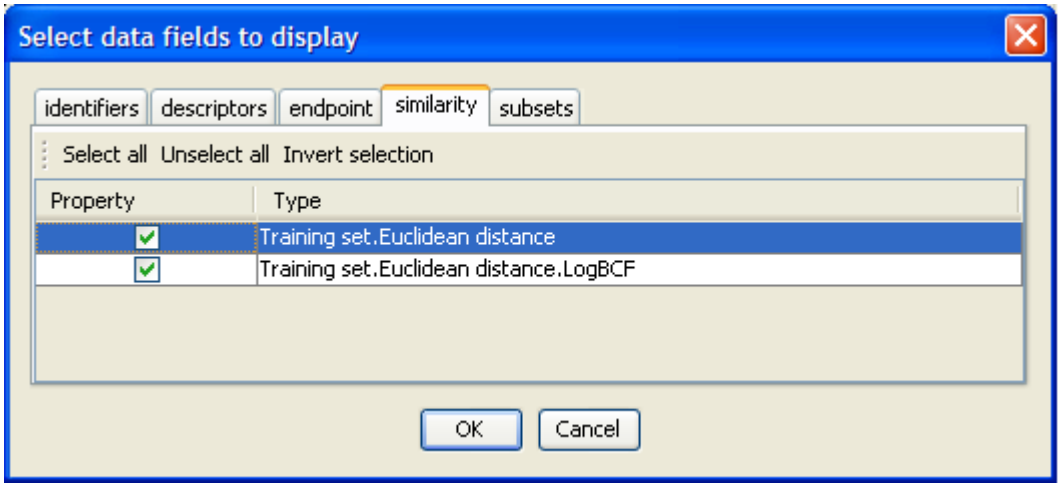


Figure 50: Selecting fields to display

The selected fields and their values will appear in the training panel.

Similarity assessment of the test compound

Now run the similarity calculation for the test compound. Use *Similarity/Similarity to the training set* from the bottom (test set) panel and follow the wizard.

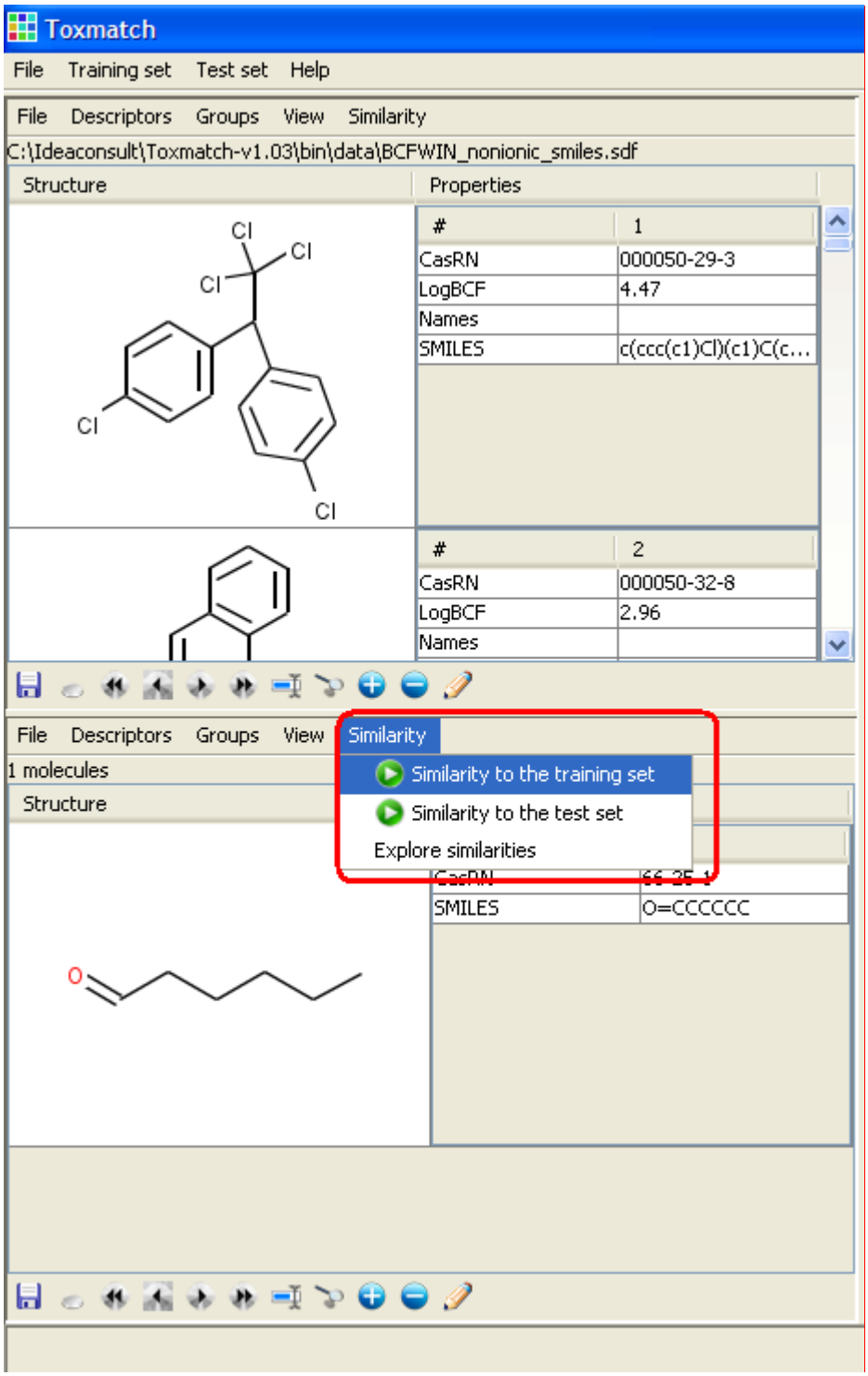


Figure 51: Similarity calculation

Exploring similarity assessment results

When ready, use menu *View/View Fields* from the bottom panel and select the same fields as for the training set. The selected fields and their values will appear in the test panel. The chart options can also be changed to label compounds by predicted LogBCF or calculated distance.

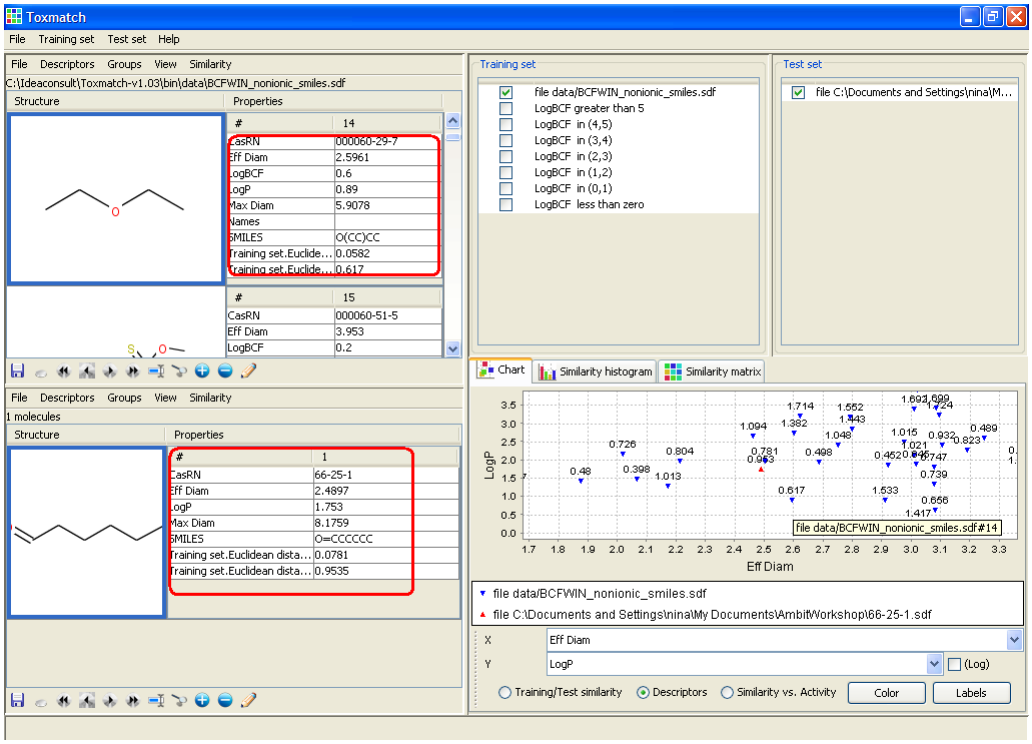


Figure 52: Displaying results

The chart can also be configured to display observed LogBCF vs. predicted LogBCF or observed LogBCF vs. Euclidean distance.

The procedure outlined has characterised the training set data using the Euclidean distance as a similarity measure. The test set query chemical has been loaded and a Euclidean distance measure calculated. A comparison between the two sets has enabled the test set chemical to be visualised in a range of ways in order to compare the most similar chemicals from the training set and make an estimate of its own likely BCF range. This could be termed as a many to one read-across.

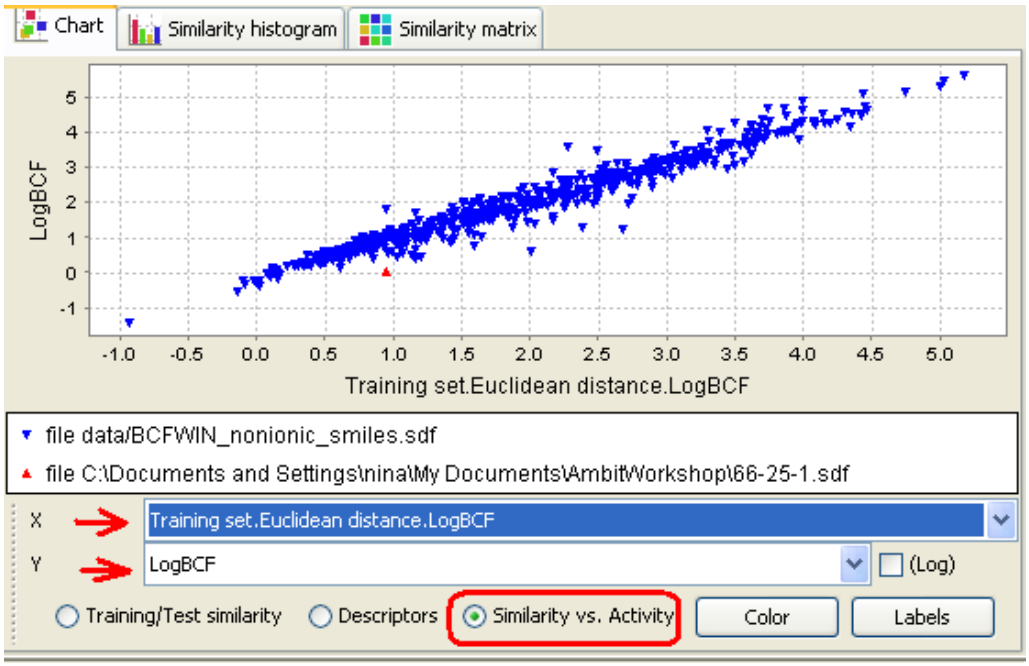


Figure 53: Predicted vs. observed LogBCF

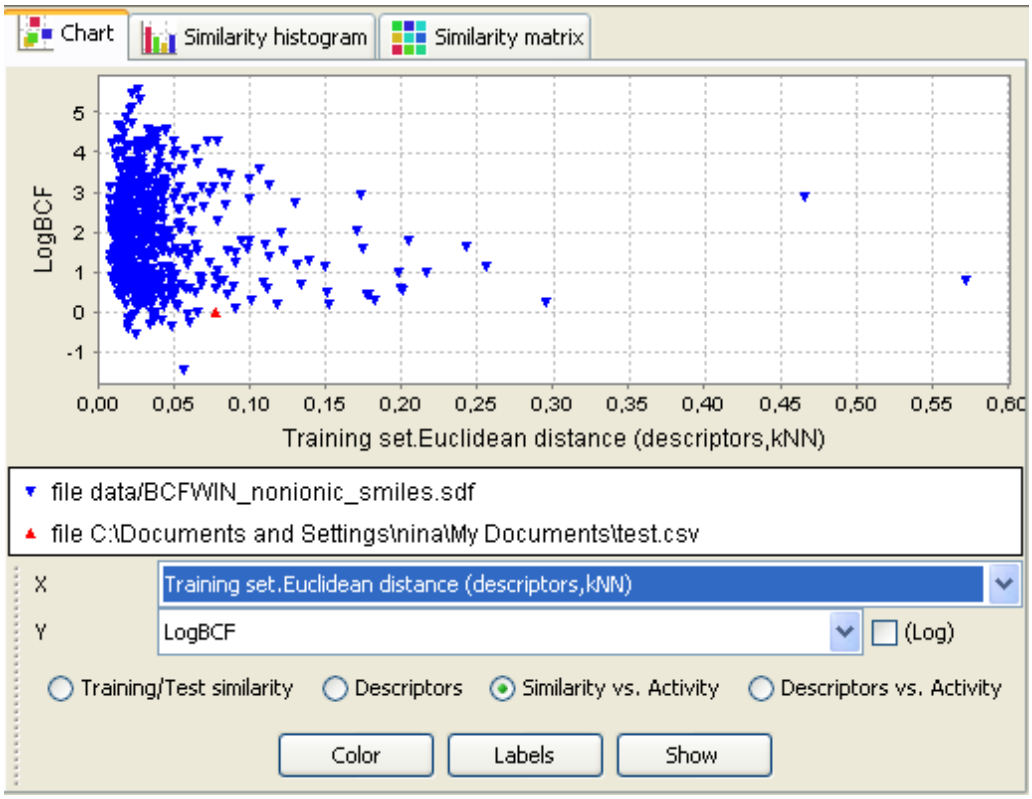


Figure 54: Predicted vs. observed LogBCF

Structural similarity - Fingerprints

Here the same menu as in Figure 46, and the same procedure will be used but instead, *Fingerprints (nearest neighbours)* will be selected instead of Euclidean distance.

Training set

Use *Similarity/Similarity to training set* from the top panel and follow the wizard (Figure 55)

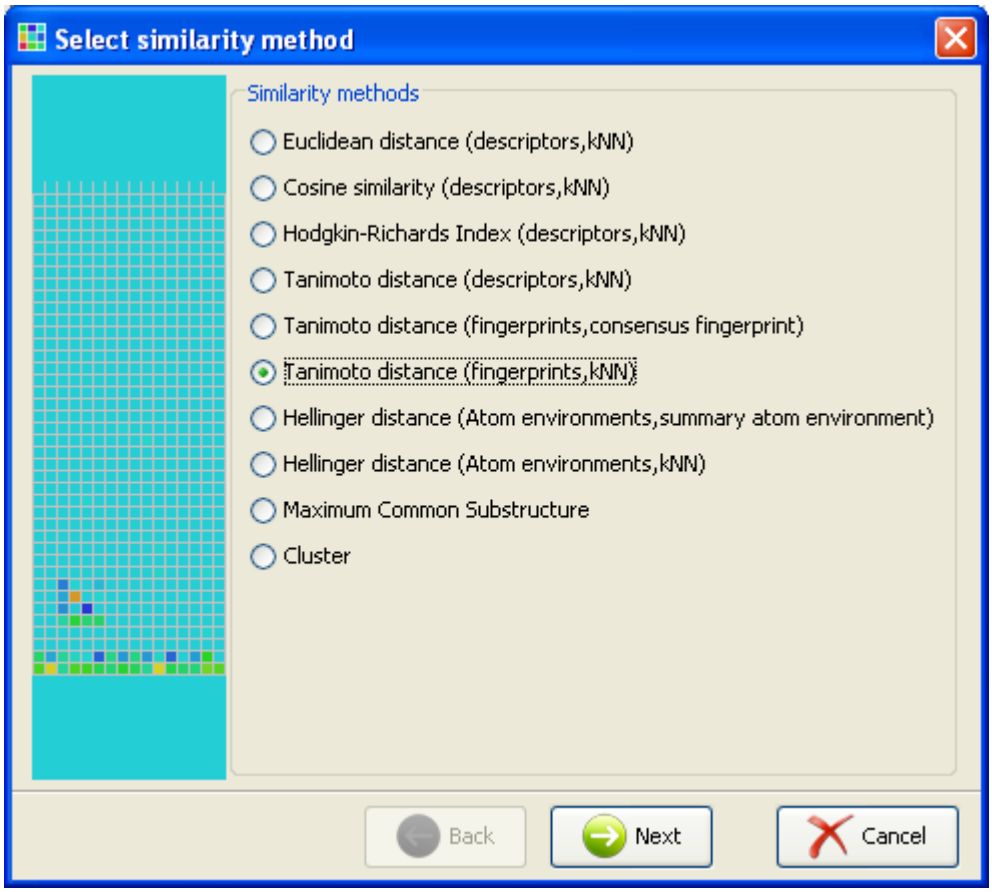


Figure 55: Selecting fingerprints for similarity assessment

Test set

Use *Similarity/Similarity to training set* from the bottom panel and follow the wizard.

Follow the same steps explained when calculating *Euclidean distance* for the test set (Similarity assessment of the test compound - p.49), but this time select *Tanimoto distance (fingerprints, kNN)* from the following menu (Figure 56).

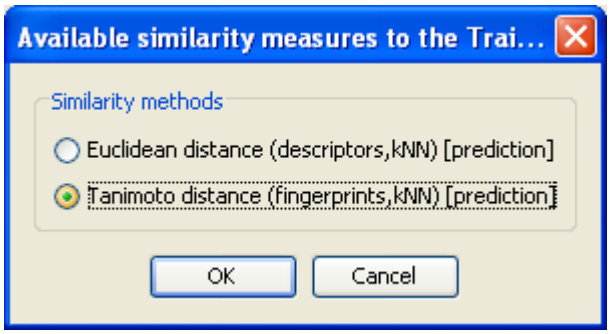


Figure 56: Similarity assessment of the test set

Exploring similarity assessment results – (scatter plots)

The field *Training set.Fingerprints (nearest neighbours)* now contains the calculated Tanimoto distance, where *Training set.Fingerprints (nearest neighbours).LogBCF* contains predicted LogBCF value, based on fingerprints similarity. We could use the same menus to visualise these fields in the training set panel and in the chart.

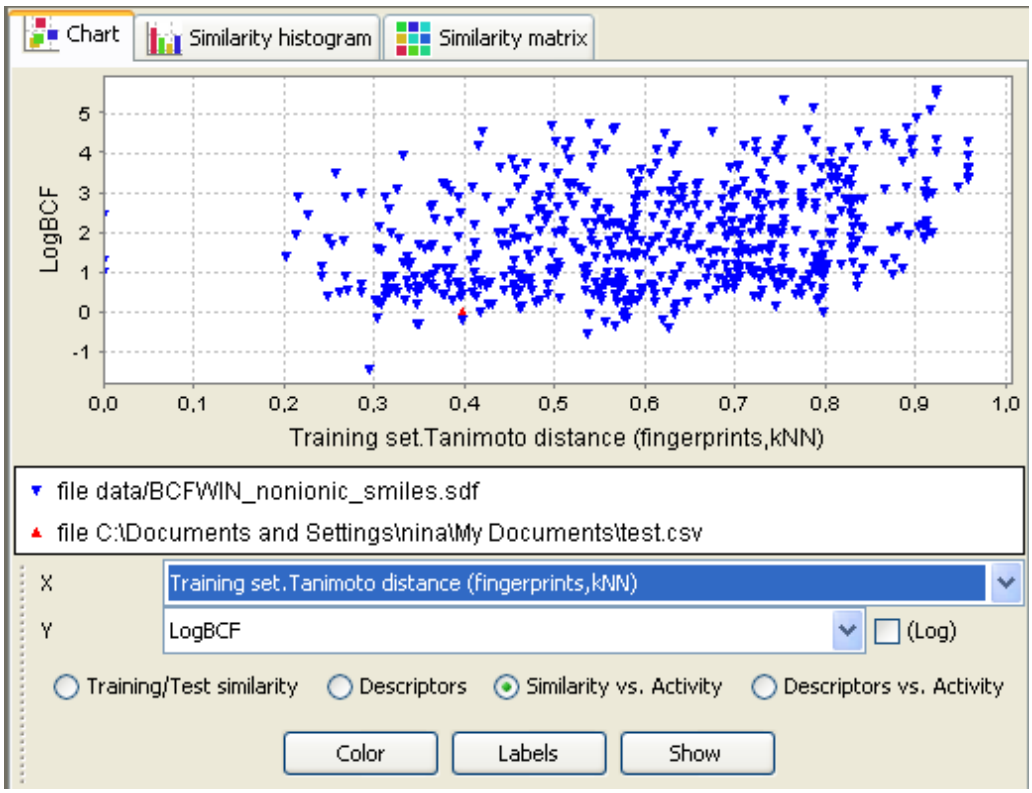


Figure 57: Tanimoto distance vs. observed LogBCF

Exploring similarity assessment results – (similarity matrix)

The most similar structures to the structure in the text set can also be explored using the similarity matrix. In the visualisation area (bottom right of the main screen), switch to the *Similarity matrix* tab.

The similarity matrix displays the pair wise similarities for the selected similarity measure. The toolbar just above the matrix configures what is displayed in rows and columns. The default setting is to compare the structures from the training set, therefore the toolbar is configured for *Training set* for both rows and columns Figure 58.

A click on a cell displays the compound on the selected row, while double click displays the compound on the selected column. The information about the clicked cell is available on the panel at the right of the matrix.

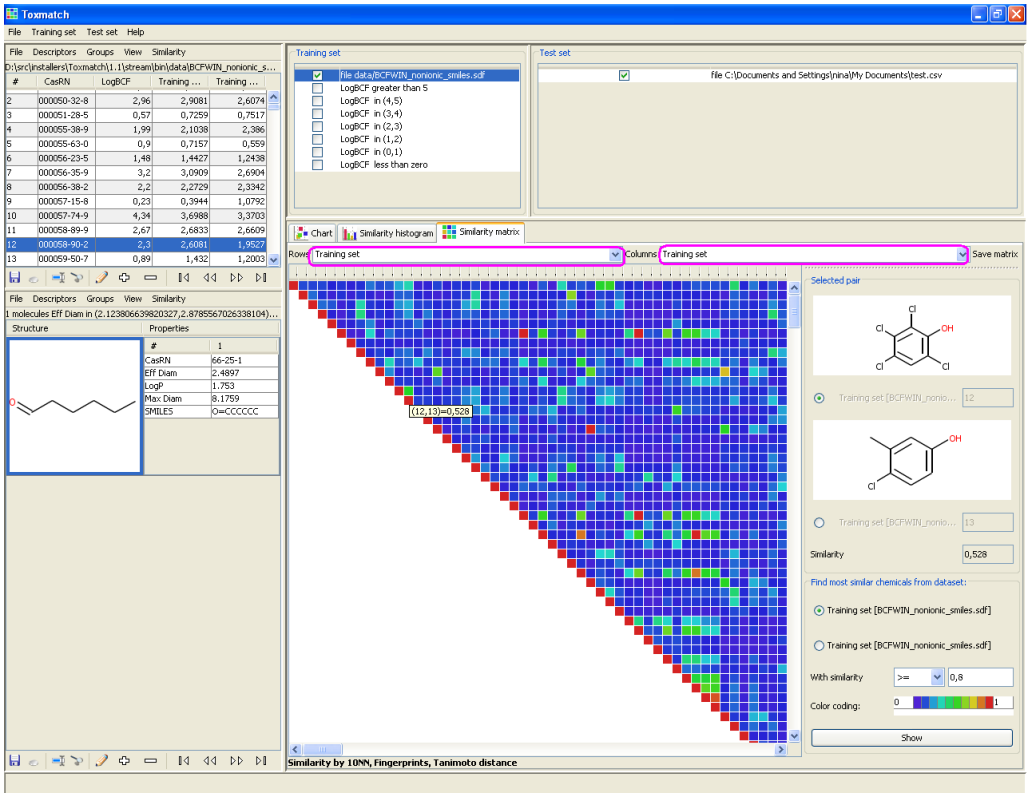


Figure 58: Similarity matrix for the training set

We will use this panel to select the most similar compounds to the test set. For this purpose:

Select *Test set* for columns

Click on the test set line in the top right corner (Figure 59)

The matrix will change to a single column, since there is only one compound in the test set. Click on each cell to see the corresponding structure from the training set and its pair wise similarity with the test structure.

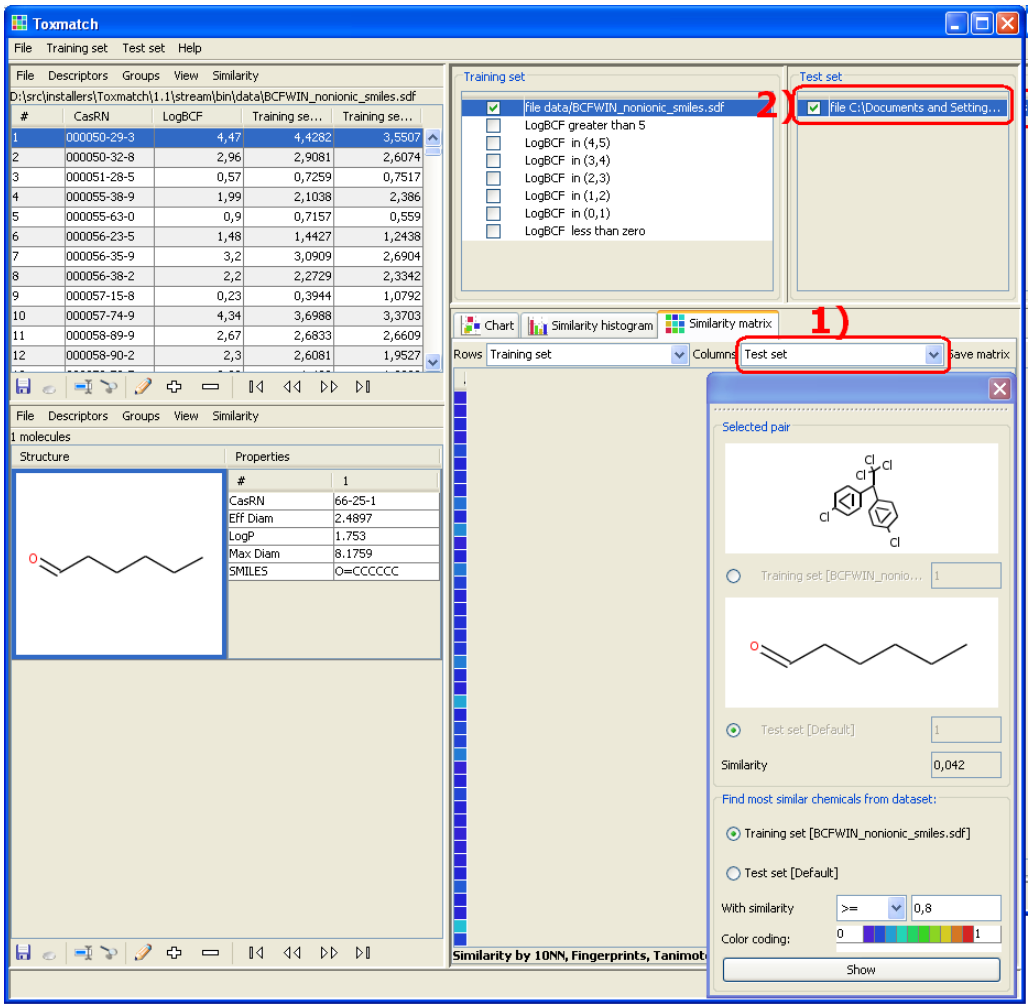


Figure 59: Similarity matrix for the test set

Identifying the most similar compounds could be achieved through performing the following steps:

Click on any cell.

Select the *Test set* radio box (lower structure in the matrix control panel).

> *The similarity matrix provides a useful interface to visualise and extract the most similar chemicals up to a given similarity threshold. This range goes from 0 to 1. Selecting a threshold of 0.31 will display the most similar structures with a similarity value less than 0.31.*

Type in 0.31 in the Threshold box.

Click the *Show* button.

This will filter the matrix to show only 10 structures with Tanimoto distance with test structure less than 0.31. The top left (training) area will also display only these 10 structures (Figure 60).

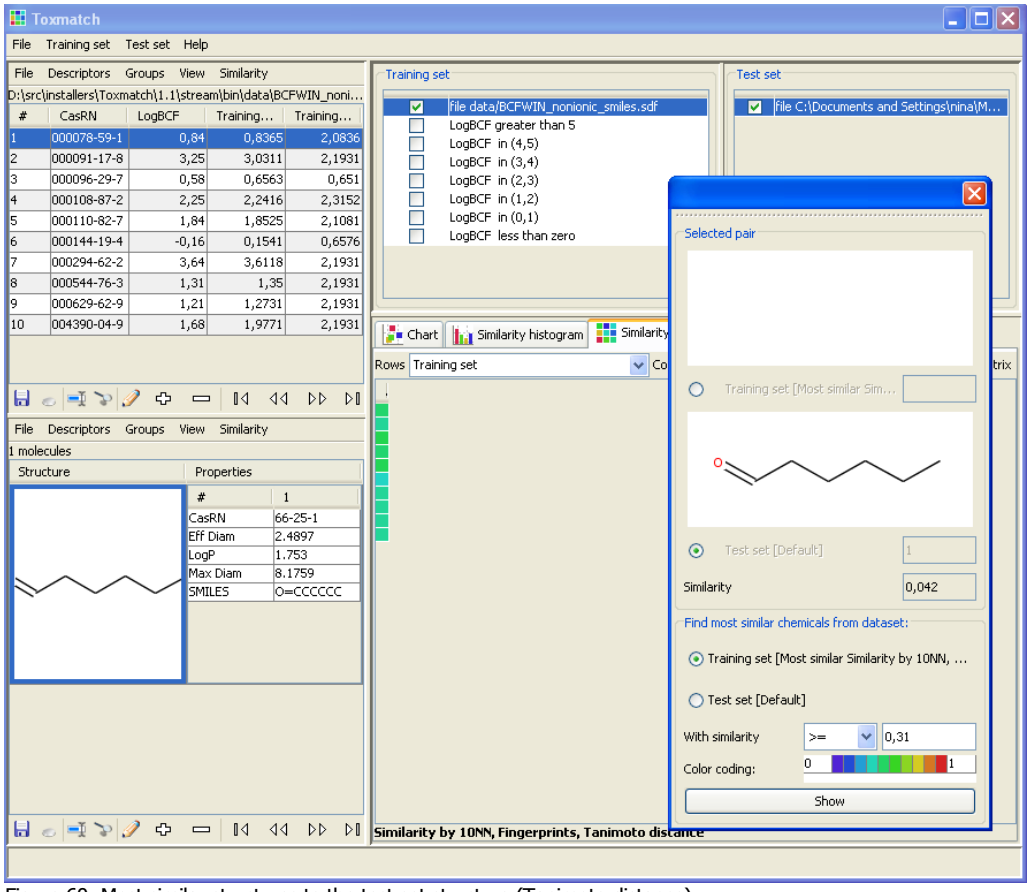


Figure 60: Most similar structures to the test set structure (Tanimoto distance)

The properties of the nearest neighbours can be additionally explored in a table form. Use menu *View/View Fields* from the top panel to select which fields will be visualised.

Switch to the *descriptors* tab and check the *Eff Diam*, *Max diam*, *LogP* descriptors (Figure 61).

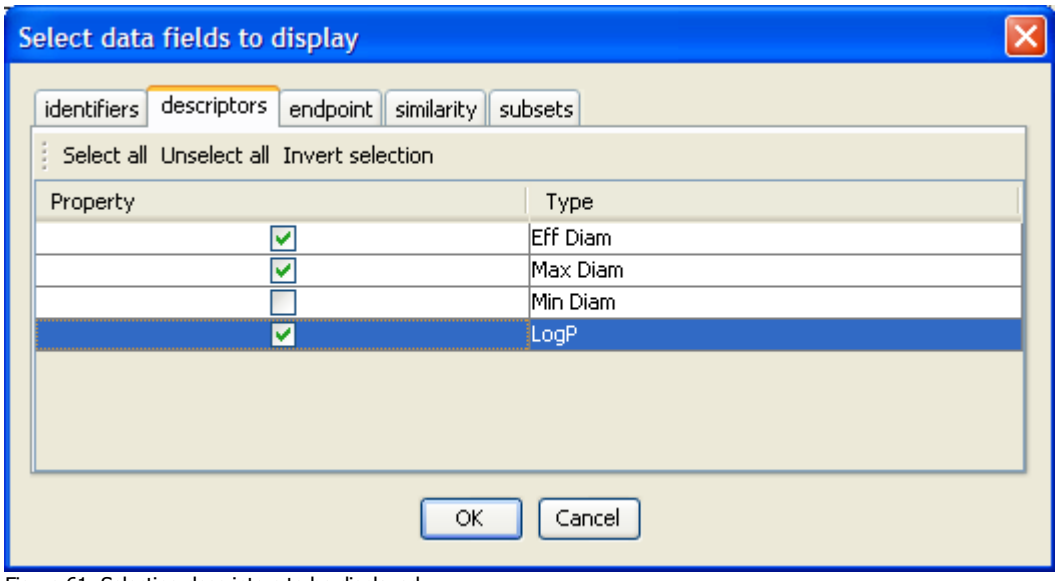


Figure 61: Selecting descriptors to be displayed

Switch to the similarity tab and select Training set.Euclidean distance.LogBCF and Training set.Fingerprints (nearest neighbours).LogBCF (Figure 62).

Click OK.

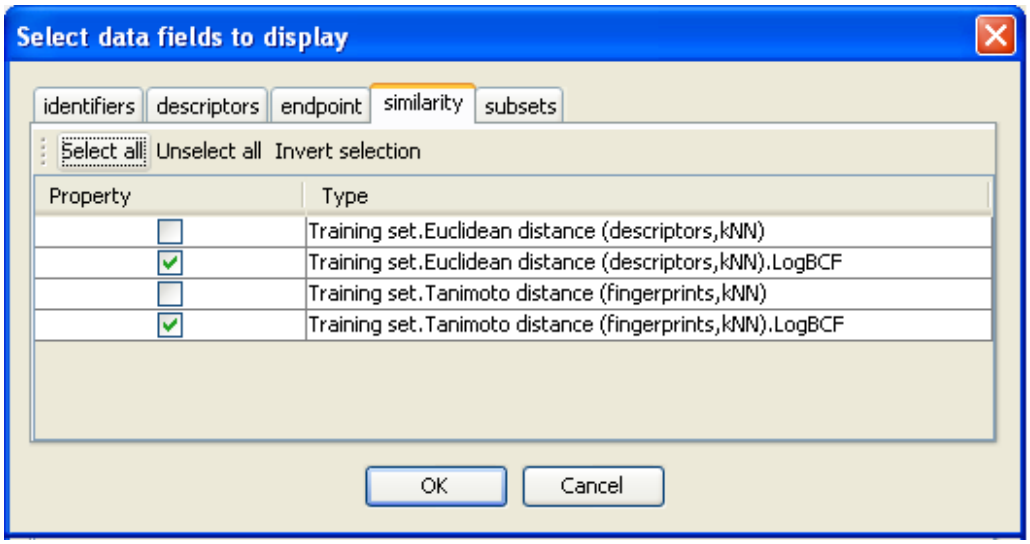
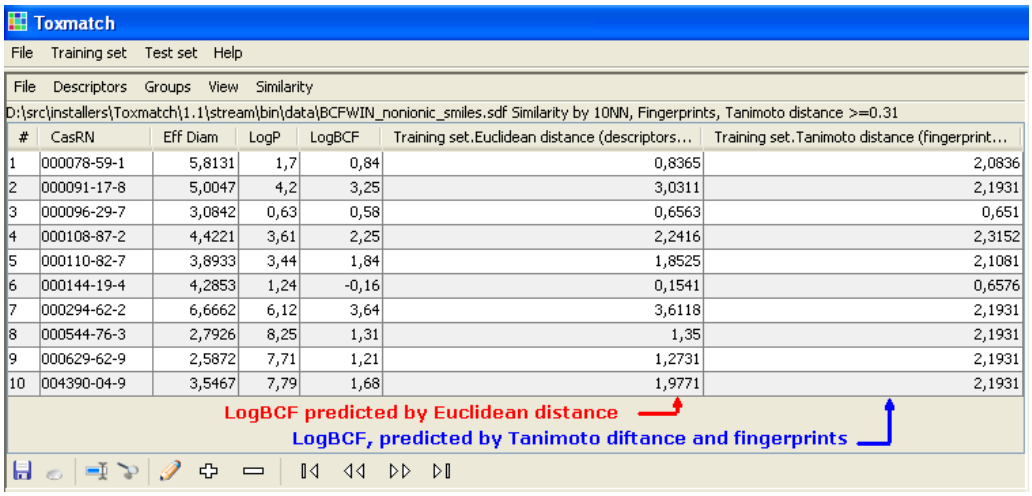


Figure 62: Selecting similarity indices and predicted LogBCF to be displayed

Use *View/View menu* to switch the display to table mode (Figure 63). Explore descriptor values and LogBCF values predicted by weighted average of nearest neighbours.



#	CasRN	Eff Diam	LogP	LogBCF	Training set.Euclidean distance (descriptors...	Training set.Tanimoto distance (fingerprint...
1	000078-59-1	5,8131	1,7	0,84	0,8365	2,0836
2	000091-17-8	5,0047	4,2	3,25	3,0311	2,1931
3	000096-29-7	3,0842	0,63	0,58	0,6563	0,651
4	000108-87-2	4,4221	3,61	2,25	2,2416	2,3152
5	000110-82-7	3,8933	3,44	1,84	1,8525	2,1081
6	000144-19-4	4,2853	1,24	-0,16	0,1541	0,6576
7	000294-62-2	6,6662	6,12	3,64	3,6118	2,1931
8	000544-76-3	2,7926	8,25	1,31	1,35	2,1931
9	000629-62-9	2,5872	7,71	1,21	1,2731	2,1931
10	004390-04-9	3,5467	7,79	1,68	1,9771	2,1931

Figure 63: Dataset display in table mode

This example has shown that it is possible to characterise the BCF dataset with respect to fingerprints and compute the weighted activity of nearest neighbours in order to then estimate the BCF for the query chemical. A similarity matrix is a convenient means of rapidly visualising the training and test set. Setting a similarity threshold can help to identify the most similar chemicals.

Figure 63 shows the compiled table of predicted LogBCF values using the two different similarity measures. The predicted values vary depending on the compounds tested, further investigation to identify which types of chemicals are most accurately predicted compared to the experimental values will help to identify which similarity approach is most optimal for a given type of chemical.

Structural similarity – atom environments

Here a similarity weighted similarity of nearest neighbours will be carried out using Atom environments (select *Hellinger distance (Atom Environments, kNN)* from the corresponding wizard menu.

Use the similarity matrix to extract most similar compounds, with a threshold of 0.75 – this will give result in 5 most similar compounds.

References

1. http://www.epa.gov/ncct/dsstox/sdf_epafhm.html, accessed 27 June 2007.

2. Russom CL, Bradbury SP, Broderius SJ, Hammermeister DE, Drummond R.A. (1997) Predicting modes of action from chemical structure: Acute toxicity in the fathead minnow (Pimephales promelas). *Environmental Toxicology and Chemistry* 16(5): 948-967.

3. Meylan WM, Howard PH, Boethling RS, Aronson D, Printup H, Gouchie S. (1999) Improved Method for Estimating Bioconcentration/Bioaccumulation Factor from Octanol/Water Partition Coefficient. *Environmental Toxicology and Chemistry* 18(4): 664–672.

4. Gerberick GF, Ryan CA, Kern PS, Schlatter H, Dearman RJ, Kimber I, Patlewicz G, Basketter DA. (2005) Compilation of historical local lymph node assay data for the evaluation of skin sensitization alternatives. *Dermatitis* 16(4): 157-202.

5. Kimber I, Basketter DA, Butler M, Gamer A, Garrigue JL, Gerberick GF, Newsome C, Steiling W, Vohr HW. (2003) Classification of contact allergens according to potency: proposals. *Food and Chemical Toxicology* 41(12): 1799-1809.

6. Roberts DW, Patlewicz G, Kern PS, Gerberick F, Kimber I, Dearman RJ, Ryan CA, Basketter DA, Aptula AO. (2007) Mechanistic Applicability Domain Classification of a Local Lymph Node Assay Dataset for Skin Sensitization. *Chemical Research in Toxicology* 20(7): 1019-1030.

7. Steinbeck C, Han Y, Kuhn S, Horlacher O, Luttmann E, Willighagen E. (2003) The Chemistry Development Kit (CDK): An Open-Source Java Library for Chemo- and Bioinformatics. *Journal of Chemical Information and Modeling* 43(2): 493-500.

8. http://www.daylight.com/dayhtml/doc/theory/theory.finger.html (accessed July 10, 2007).

9. Xing L, Glen RC. (2002) Novel Methods for the Prediction of logP, pKa, and logD. *Journal of Chemical Information and Computer Science* 42(4): 796-805.

10. Bender A, Mussa HY, Glen RC, Reiling S. (2004) Molecular Similarity Searching Using Atom Environments, Information-Based Feature Selection, and a Naïve Bayesian Classifier, *Journal of Chemical Information and Computer Science* 44(1): 170-178.

11. Jaworska J, Nikolova-Jeliazkova N. (2007) How can structural similarity analysis help in category formation. *SAR and QSAR in Environmental Research* 18(3-4): 195-207.

12. Todeschini R, Consonni V. (2000) Handbook of molecular descriptors. Wiley-VCH: Weinheim, Germany.

13. Aha D, Kibler D. (1991) Instance-based learning algorithms. *Machine Learning* 6(1): 37-66.

14. Witten IH, Frank E. (2005) Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, San Francisco, 2nd Edition.

15. Verhaar HJM, van Leeuwen CJ, Hermens JLM. (1992) Classifying environmental pollutants. 1. Structure-activity relationships for prediction of aquatic toxicity. *Chemosphere* 25: 471-491.