Master in Artificial Intelligence (UPC-URV-UB)

Master of Science Thesis

Data analysis and navigation in high-dimensional chemical and biological spaces

Josep Cester Bofarull

Advisor: Robert Rallo Moya

September 2012

Acknowledgements

Thanks to Dr. Robert Rallo for his supervision of the thesis and the dedication during the whole project. I would like to thank as well the BioCenit Research Group for giving me the opportunity of building and testing this complex and useful tool.

This project has been funded by the Osiris Project (Integrated Project within the Sixth EU Research Framework Programme, Contract No. 037017), the POPmap project (MICINN, CTM2011-24303) and the R2B grant (2011R2B-01) given by Universitat Rovira i Virgili.

Table of contents

List of figures

List of tables

1. Introduction

Chemical Space (which encompasses all possible small organic molecules, including those present in biological systems) is vast, vast in terms of number of chemicals that may contain and in terms of the number of available descriptors for each of these chemicals (van Deursen et al. 2007).

Theoretically the chemical space is formed by all possible stable molecules. Even small parts of this chemical space contain large amounts of information. Actually the largest chemical databases contain up to 25 million different molecules which are far from the amount of information contained in the total chemical space. However, the analysis of these large scale chemical databases could reveal which regions of the chemical space have been extensively explored and which remain relatively uncharted (Yamashita et al. 2006; Larsson et al. 2007). These would have a direct application in drug discovery, where there is always the need of generate new compounds and detect small molecules whose properties enable them to interact with biological molecules without generating adverse effects (O'Driscoll 2004).

The development of advanced visualization and navigation techniques is required to analyze these databases. It is know that the human brain is capable of visually detecting non evident relationships or patterns from data representations. Techniques for the virtual screening of chemicals could benefit from the information that can be visually extracted from a graphical representation of the chemical space. The combination of virtual screening techniques with tools for advanced visualization of high dimensional data would result in a new generation of virtual visual screening tools that will facilitate the extraction of relationships between the molecular structure of a chemical and their physicochemical and biological properties.

Nowadays computers hardware performance is good enough to manage and process a huge part of available chemical and biological data. However, the visual inspection and analysis of high dimensional data spaces is a difficult task which

requires the use of appropriate projection techniques to reduce the dimensions of the original space up to a dimension suitable for its visualization (usually 2D or 3D).

Our project is based on providing exploratory data analysis as a first step towards modeling and interpretation of the chemical and biological activity. Direct interaction from the user in the process of data analysis facilitates the discovery of cause-effect relationships between the studied parameters (e.g., establishing structure-activity relationships).

1.1. REACH regulatory framework

Current regulations for the use of chemicals in the European Union require the complete characterization of the potential environmental and human health impact of chemicals. Central to this characterization is the assessment of persistence, bioaccumulation and toxicity (PBT) profiles for chemicals produced or imported in amounts above 100 Tm/year. As a consequence, chemical industries must invest a significant amount of economic resources to accomplish with regulatory requirements. In this context, the use of non-testing methods is emerging as an alternative to reduce the costs, both economic and in terms of animal use, associated with the implementation of REACH (Registration, Evaluation, Authorization and Restriction of Chemicals) regulatory framework.

Non-testing methods for chemicals are based on the use of existing data to infer the PBT properties of new chemicals. The experimental assessment of new chemicals can be waived when non-testing methods provide enough evidence to support the waiving decision. Data analysis and modeling via diverse machine learning algorithms are the most used techniques to implement non-testing strategies for chemicals.

1.2. Similarity measures

The concept of molecular similarity is widely used in medicinal chemistry and cheminformatics. The basic idea is that when molecules have similar chemical or

structural properties, they will probably behave in a similar fashion in biological assays (Dean 1995; Maldonado et al. 2006). Choosing the appropriate combination of molecular properties (descriptors) and analysis methods (metrics) for the estimation of similarity between molecules is vital for this process. However, this approach has grown to be a key factor in improving the efficiency of modern virtual screening programs. Experienced medicinal chemists often use visual inspection of structurally distinct molecules to look for similarity in structure that might translate to biological activity in novel series (Glen et al. 2006).

The similar property principle, which states that structurally similar molecules tend to have similar properties, constitutes the underlying idea of many applications in cheminformatics such as compound ranking (chemical similarity searching), ligand-based virtual screening, and diversity analysis for compound library design. These applications are utilized by employing chemical descriptors together with a measure of (dis)similarity defined on the descriptors (Rupp et al. 2008).

1.3. Visual Data Mining

Data Mining is commonly defined as the extraction of patterns or models from observed data, usually as part of a more general process of extracting high-level, potentially useful knowledge, from low-level data. Data visualization and visual data exploration play an important role in this process. If the data is presented textually, it is very difficult to deal with data sets containing millions of data items. Analysts need tools for creating hypotheses about complex data sets, a process that requires capabilities for exploring and understanding them (Oliveira et al. 2003).

For data mining to be effective, it is important to include the human in the data exploration process and combine the flexibility, creativity, and general knowledge of the human with the enormous storage capacity and the computational power of today's computers. Visual data exploration aims at integrating the human in the data exploration process, applying its perceptual abilities to the large data sets available in today's computer systems. The basic idea of visual data exploration is to present the data in some visual form, allowing the human to get insight into the data, draw

conclusions, and directly interact with the data. Visual data mining techniques have proven to be of high value in exploratory data analysis and they also have a high potential for exploring large databases (Keim 2002).

2. Aims and scope of the project

The goal of this master thesis is to develop and validate a visual data-mining approach suitable for the screening of chemicals in the context of REACH. The proposed approach will facilitate the development and validation of non-testing methods via the exploration of environmental endpoints and their relationship with the chemical structure and physicochemical properties of chemicals.

The use of an interactive chemical space data exploration tool using 3D visualization and navigation will enrich the information available with additional variables like size, texture and color of the objects of the scene (compounds). The features that distinguish this approach and make it unique are (i) the integration of multiple data sources allowing the recovery in real time of complementary information of the studied compounds, (ii) the integration of several algorithms for the data analysis (dimensional reduction, generation of composite variables and clustering) and (iii) direct user interaction with the data through the virtual navigation mechanism. All this is achieved without the need for specialized hardware or the use of specific devices and high-cost virtual reality and mixed reality.

The space visualization and data analysis using a large number of variables is a difficult task that requires the use of appropriate projection techniques to reduce the number of variables of the original space to a size suitable for its visualization. The project aims to build a tool that provides multiple options to allow this visualization. A dimensional reduction module allows selecting desired variables and dimensional reduction algorithm to locate elements in a three-dimensional space. After this process that assigns the coordinates values in the space for each represented compound, the 3D space creation module also allows the user to specify the variables that are to be assigned to the size, color and texture of the elements (compounds) located in space. Note that the three spatial coordinates are not distances but values of the three variables (relevant physicochemical properties and molecular descriptors) chosen for viewing/representing the elements studied. So, user can see 6 variables, 3 coordinates and 3 characteristics of compounds.

In an area like chemical and biological data, the number of chemical compounds is very large and the number of properties for each chemical compound also is. To analyze similarities between different compounds we can represent them as spheres and place them in space in a particular way to visually see the differences between them. These differences will be stated not only by the different situation in 3D space but also by differences in color, size (radius sphere) and texture (e.g., transparency), as shown in **Figure 1**. This will make the user move from having large volumes of data in Excel sheets or 3D bar charts to have a representative and orderly 3D chemical space where it will be easy to see at a glance the differences and similarities between the compounds.

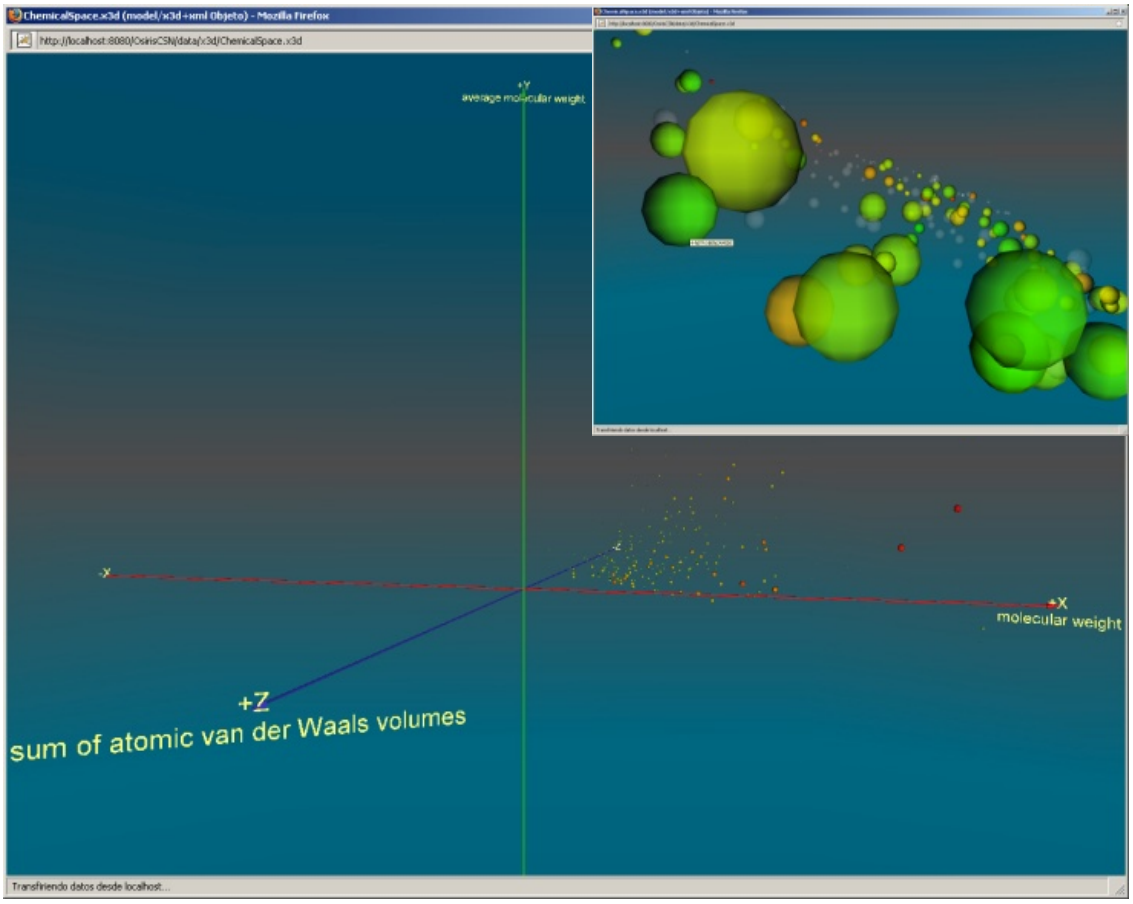

Figure 1. General appearance of the graphical interface of the tool

3. Related work and tools

Several projects and tools have tried to address the chemical space data analysis and exploration issue. From last decade, some of them have also tried to incorporate a 3D visualization of chemical space. Our project presents an innovative solution with respect to most of the existing works. At the conceptual level, our tool places the expert in the application domain within the data space to increase the effectiveness of the process of data analysis and decision-making mechanisms. In terms of hardware needed, the tool simplifies viewing and other mechanisms of interaction in the market that require the use of specialized and expensive devices. Another advantage of the tool is the support it has had from scientists from different disciplines (chemistry, bioinformatics and biostatistics).

3.1. State of art

(Dobson 2004) predicts that, using the types of computational methods pioneered by the flourishing bioinformatics community, the analysis of databases obtained from large-scale screening exercises of small molecules on biological systems should lead to major advances, both in our understanding of biological chemistry and in our ability to identify promising therapeutic compounds and therapeutic targets. Although progress is now being made in developing tools for mining chemical information, such progress is often limited by the difficulty in accessing much of the data of interest. He also proposes that to exploit fully chemical tools and new methodologies in molecular and structural biology, chemist must increasingly develop strong interactions with scientist from different disciplines.

A framework for compound classification and comparison is provided in (Larsson et al. 2007). The tool presented allows identifying volumes related to particular biological activities and track changes in chemical properties. It is also capable of charting biologically relevant chemical space and provides an efficient mapping

device for selection of high-probability hits and prediction of their properties and activities.

(Feher and Shcmidt 2003) show in their study how compounds can be separated in three different groups based on the value of its descriptors (number of chiral centers, the prevalence of aromatic rings, the introduction of complex ring systems, and the degree of the saturation of the molecule as well as the number and ratios of different heteroatoms). A PCA-based scheme is presented that differentiates the three classes of compounds.

Related to drug discovery, (van Deursen and Reymond 2007) opt to report a "spaceship" program in a known drugs region which travels from a starting molecule A to a target molecule B through a continuum of structural mutations, and thereby charts unexplored chemical space. The compounds encountered along the way may provide valuable starting points for virtual screening.

The use of neural networks based on self-organizing maps is proposed in (Matero et al. 2006). By using a tree-structured self-organizing map it is possible to construct a chemical space of compounds. Using neural networks based on Kohonen unsupervised learning, the neural networks train themselves without any external information. They learn the data and categorize it according to common features in the data. Thus, the user can visually inspect which of the original variables are responsible for the clustering results. The recognition of clusters, however, is more or less in the eyes of the observer and no formal clustering exists.

Another approach using a novel hierarchical data visualization technique (HeiankyoView) can be found in (Yamashita et al. 2006) study. With this technique it is possible to visualize large-scale multidimensional chemical information using 2D square images of subspaces, allowing the analysis of the structure-activity relationship of compounds. HeiankyoView represents hierarchically

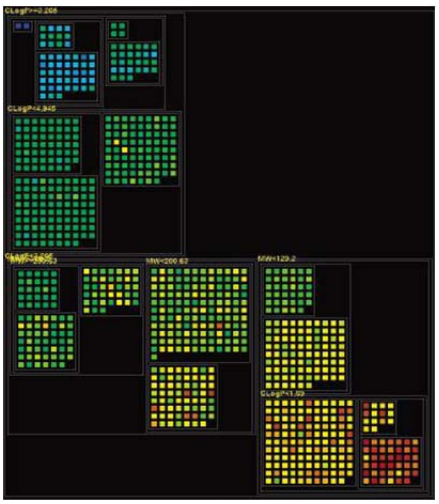

Figure 2. HeiankyoView hierarchical data visualization

organized data objects by mapping leaf nodes as colored square icons and non-leaf nodes as rectangular borders (**Figure 2**). In this way, data objects can be expressed as equishaped icons without overlapping one another in the two-dimensional display space. Thus, we can recognize trends in molecular physical properties relevant to a specific chemical class and optimize potential compounds.

(Petalson et al. 2007) explains the concept of activity landscapes, hypersurfaces in biologically relevant chemical space, where biological activity (compound potency) adds another dimension. In these landscapes smooth regions that are reminiscent of hills correspond to areas where gradual changes in chemical structure are accompanied by moderate changes in biological activity (Compounds mapping to such areas are related by so-called continuous Structure-Activity Relationships). By contrast, rugged regions in activity landscapes that are canyon-like correspond to areas where small chemical changes have dramatic effects on the biological response, and hence, compounds mapping to these areas form discontinuous Structure-Activity Relationships (**Figure 3**).

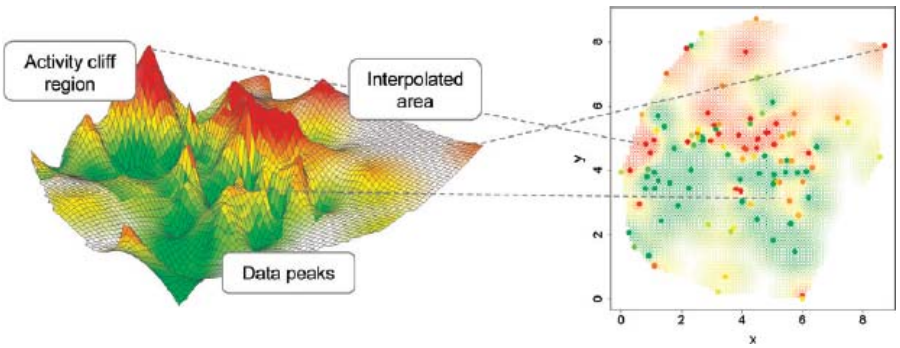

Figure 3. 3D and 2D activity landscape representations

3.2. Similar tools

By the finalization of our project, a very similar tool was developed by (Gütlein et al. 2012). Ches-Mapper is presented as an application to visualize and explore chemical datasets. In a preprocessing step, a chemical dataset is mapped into a virtual three-dimensional space. A key part of the preprocessing is the choice of features done by the user. The selected features are then used for clustering and 3D embedding. Thus,

compounds that have similar feature values are likely to be clustered together, and are closed to each other in 3D space (**Figure 4**).

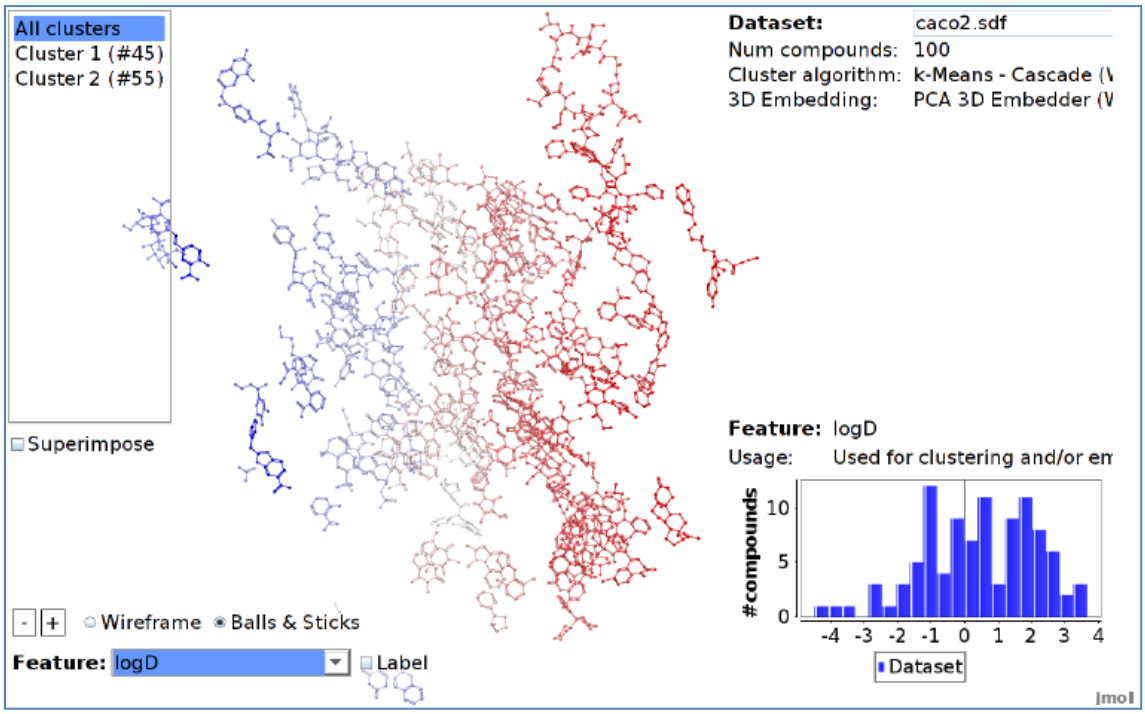

Figure 4. CheS-Mapper 3D chemical space visualization

The process of generating the 3D chemical space is very similar to the one followed in our tool:

- **Load dataset.** The first step is the dataset selection. User can select an existing dataset or import a new one.
- **Create 3D structures.** 3D structure can be calculated for the compounds in case it is not already present in the original dataset. CDK (Chemical Development Kit) and Open Babel libraries are used for this purpose.
- **Extract features.** User can select which features to employ in the subsequent steps (clustering and embedding). Three different types of features are available: included in dataset, CDK descriptors and structural fragments.
- **Cluster dataset.** Clustering divides the dataset into subgroups. Only features that have been selected in the previous step are used as input to the

clustering algorithm. Cluster algorithms from the statistics library *R* and the data-mining library *Weka* can be employed.

- **Align compounds.** User can chose the alignment method that will be used for the alignment of the compounds inside a cluster according to a common substructure.

After evaluating the tool, we realize that Ches-Mapper has many similarities with the work presented in this thesis, but there are some points that could make our tool preferred to this one:

- **Web application vs Java application.** Ches-Mapper is a Java application, which means that needs to be downloaded in each computer where we want to use it. Our tool is a web application which needs only one installation in a web server. User can access the tool from any computer having an internet connection, and the used data is always available from any computer. This also facilitates sharing data and visualization with different users.

- **Use of 6 variables.** In Ches-Mapper compounds are flying in the space and are colored depending on the cluster they belong to. In our tool, dimensional reduction and clustering are done in different steps, like this each compound have a fixed position in a 3 coordinates space (which provides the value of relevant physicochemical properties and molecular descriptors) and its size, color and texture, which enriches the visualization having up to 6 variables to differentiate the elements.

- **Communication with external components.** Once user is visualizing a compound in detail, our tool offers all compound available information from the external PubChem database.

- **Maximum number of compounds.** Ches-Mapper allows working with up to 6000 compounds, assuring a good response with up to 1500 compounds. Our tool offers an extra clustering option to group compounds inside a single element and showing them only when user is near their cluster. This way

navigation is smoother and users can work with large data sets (e.g., more than 10 thousand compounds).

- **Data mining module.** The tool provides a data mining module that offers the possibility to build 'visualization trees' in order to facilitate multiple chemical spaces visualization corresponding to a particular compound collection.

3.3. Commercial tools

There exist also in the market other tools for commercial use. These tools are based on the use of proprietary databases and focus to facilitate the use of various predictive schemes:

• LeadScope Inc.[1] is an American company leader in the field of predictive models for chemical compounds. In its portfolio of products and services offers access to proprietary databases with relevant information for the prediction of toxicity and / or chemical properties. The company allows the user to buy the software to develop their models, or alternatively the user can access the service "QSAR as You Go", which allows predictions for a single compound.

• Derek Nexus (Lhasa Technologies)[2] is a tool that allows the study of the toxicity of chemicals. The principle of operation is similar to the one by LeadScope mentioned before. The application provides access to a proprietary database that is used to predict certain end-points of toxicity (carcinogenicity, mutagenicity, skin irritation, etc.).

• MultiCase Inc.[3] is an American company that provides knowledge-based systems for predicting properties and biological activity of chemical compounds. Among the products provided we can find: MCASE/MC4PC, a system for building structure-activity relationships; CASETOX, a tool based on MCASE specializing in predicting toxicity.

---

[1] http://www.leadscope.com
[2] https://www.lhasalimited.org/derek_nexus/DX
[3] http://www.multicase.com

4. Technical development

The tool has been implemented using the Java language to ensure cross-platform compatibility and integration of new components. Open source libraries of machine learning algorithms have been included in the tool to provide the basic preprocessing, dimensional reduction and clustering capabilities. X3D language has been chosen to build the tri-dimensional chemical space scenes.

The project has been developed following a 3-tier programming model, thus separating the presentation (GUI), business logic (application functionality), and the logic of data (databases and other information sources). As can be seen in **Figure 5**, the client layer contains the browser (which will load HTML pages, JavaScript code and XML files). Using the browser the user will interact with the application, the requests will be sent to the application server (Apache Tomcat in our case), which will process and build new dynamic HTML pages with the results. In the application layer we can see the web server that receives the requests, the J2EE modules and external resources they can use (e.g., machine learning libraries). J2EE modules will interact with the database, updating or requesting information as needed.

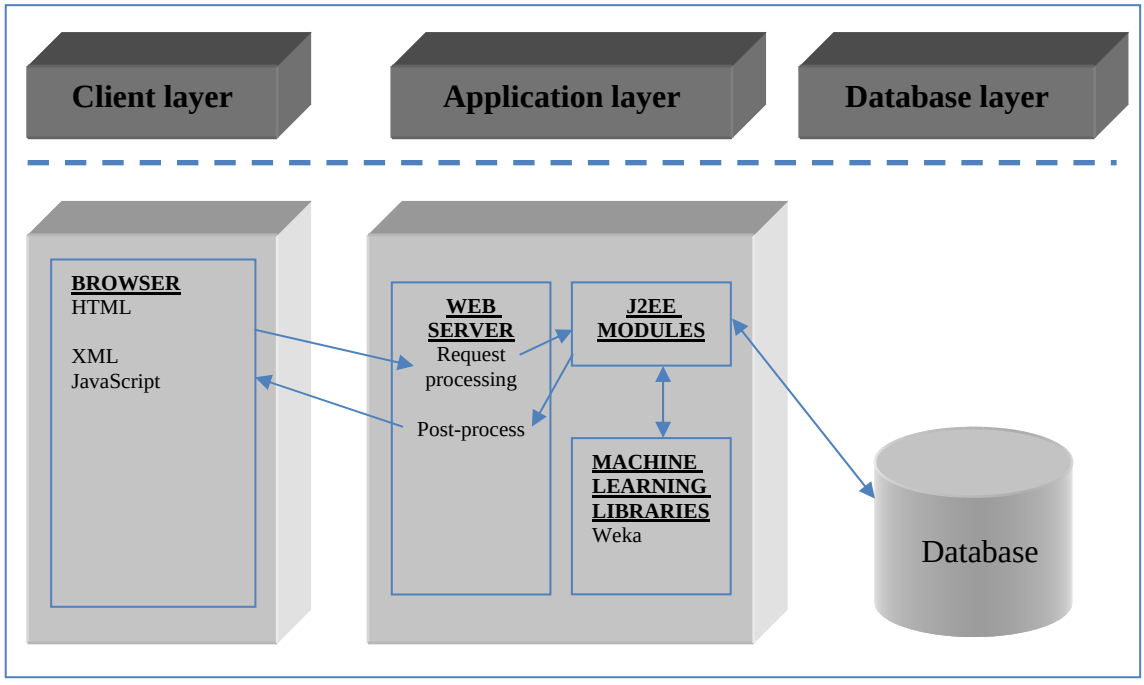

Figure 5. 3-tier programming model

4.1. Functional and non-functional requirements

The main functional requirements of the tool can be summarized as:

- **FR1. Data import and export capabilities**. The tool should be able to import and export its data to common formats like comma-separated text files.

- **FR2. Communication with external components**. The tool must be able to communicate through a well-defined interface with external components such as databases or web services.

- **FR3. Data selection and transformation**. The user has to be able to select and manipulate the data used by the tool to create the representation of the chemical space. This will include (i) the selection of the information to be used in the visualization from all the available information in the databases or external data sources, and (ii) the basic transformation of the data either as individual variables or as a group. The set of data transformations will include diverse normalization techniques and linear data transformations.

- **FR4. Projection techniques for dimension reduction**. To visualize these high dimensional data spaces the tool would require the implementation of a set of projection techniques capable of reducing the dimension of the input data and preserving the data relationships found in the original (untransformed) space. The result of this dimension reduction process will provide a set of coordinates defining the location of each compound in the chemical space representation.

- **FR5. Classification and labeling engine**. The user must be able to explore the clustering structure of the chemical space. To this end the tool must provide basic support and algorithms to classify and label the chemical compounds according to their structural characteristics or biological activity. The classification engine will also provide basic similarity metrics to generate similarity matrices for the components of the chemical space.

- **FR6. Three-dimensional navigation and interaction**. The user must be able to navigate interactively into a 3D representation of the chemical space

analyzed. The tool will incorporate basic 3D viewers that will permit the interaction of the user with the 3D scene. The user must be able to visualize physicochemical, molecular and biological information of the chemicals during the 3D navigation.

The additional non-functional requirements considered for the tool can be summarized as:

- **NFR1. Use of open source components**. The tool must be based in open source components to ensure the compatibility across systems and also to provide a solution independent of any specific software provider.
- **NFR2. Internationalization of the user interface**. The user interface of the tool will initially be provided in English. The tool will include the necessary mechanisms to ensure the easy internationalization of the user interface.
- **NFR3. Extension mechanism allowing the inclusion of user-defined components**. Several of the tool components will require the addition of new user-defined modules. To this end the tool will include a basic API definition for user added extensions.

4.2. Conceptual Model and building blocks

Figure 6 shows the conceptual model of the tool based on the functional requirements presented in the previous subsection. The main building blocks of the proposed tool are specified from the conceptual model.

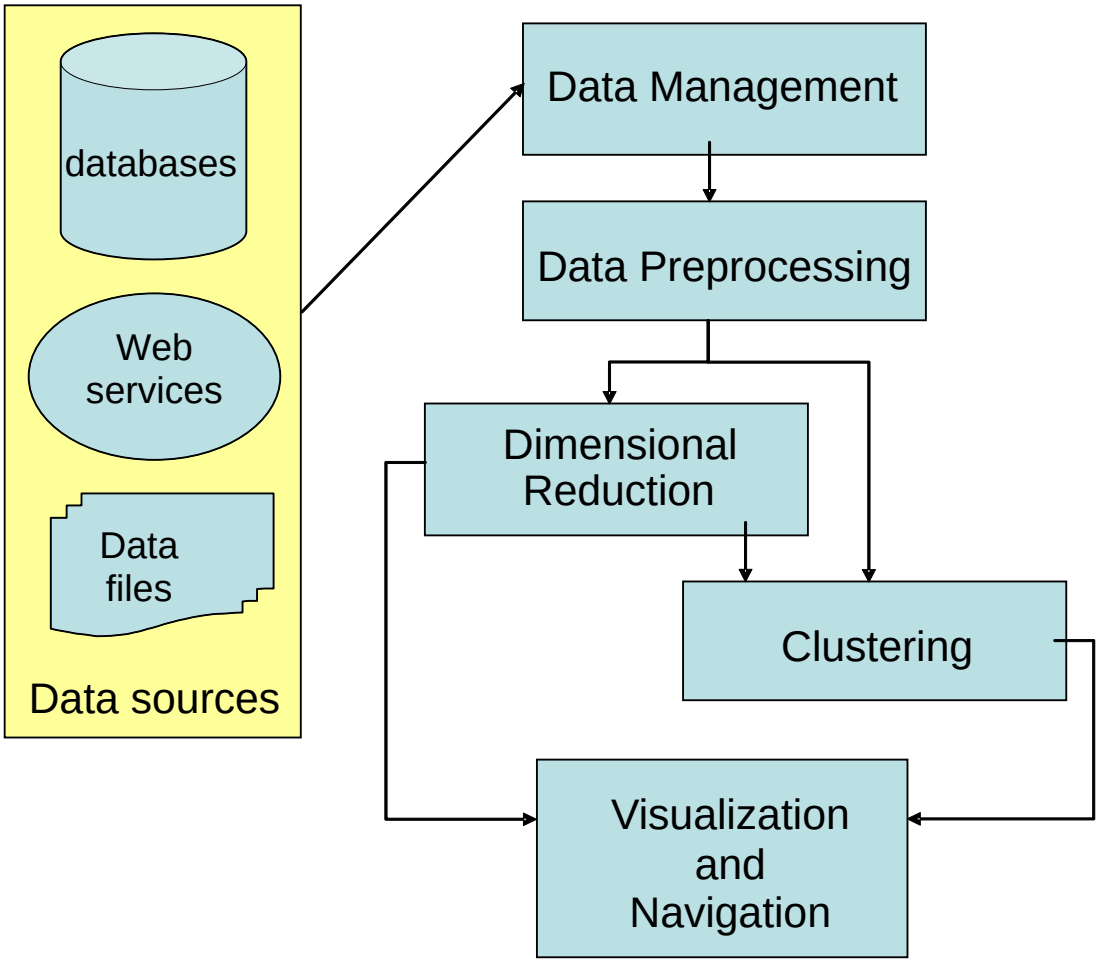

Figure 6. Conceptual model of the tool showing the data flows between components

Data Management Module. This will be the component responsible of the communication between the tool and their data providers (data sources). Three basic types of data sources are proposed for the tool:

- Data files. This data source will provide access and compatibility with data stored in most of the data management applications (for instance, CSV formatted files resulting from a data export from an EXCEL spreadsheet).
- Database connector. The tool should be able to use several databases as a native data source providing direct access to the chemical and biological information stored in the database. A mechanism to control and grant access levels should be included in order to preserve the confidentiality of the proprietary information stored in the database.

- Communication with web services. The tool will provide bidirectional data exchange between several chemical web services and the tool.

Preprocessing Module. According to the basic functional requirements of the tool, a basic set of preprocessing primitives will be needed. The proposed set of functionalities provided by this module is:

- Data filtering and selection. The preprocessing module should permit the user of the tool to select among all the available data the most appropriate information to be used for the generation of the map of the chemical or biological space. The tool must be able to distinguish between molecular information (i.e., the chemical space descriptors) and biological information (i.e., environmental endpoints).

- Basic data transformations including diverse normalization schemes (range, variance), linear transformation (scaling and shifting), as well as user defined non-linear transformations.

Projection and dimensional reduction Module. The set of chemical and biological space descriptors will define a high dimensional space in which the visualization and navigation tasks will be very difficult. The tool must incorporate a set of lower dimensional projection components aimed to reduce the dimensionality of the original input space. The dimensional reduction will usually transform the high dimensional original space into 1D, 2D, 3D or even 4D spaces suitable for its visualization and manipulation. The projection methods must allow preserving the relationships of data in the original high dimensional space. The dimensional reduction module provides the tool with the appropriate coordinate system to locate compounds in the chemical and biological space map.

Clustering Module. In addition to the dimensional reduction process, a set of automated procedures for data classification and labeling would be needed. The clustering module will complement the dimensional reduction component providing chemical similarity estimations.

Visualization and Navigation Module. The purpose of the navigation module is to offer a mechanism to interact in 3D with the visual representation of the chemical and biological space. This module will use the information provided by the dimensional reduction component to associate a point coordinate in the chemical space to each compound. A data mining 3D visualization option will offer the possibility to build 'visualization trees' in order to facilitate multiple chemical spaces visualization corresponding to a particular compound collection.

4.3. Database design

The tool needs to efficiently manage a huge amount of information to process all user selected data and show the results. Not only chemical information must be saved, but also the result data obtained applying dimensional reduction and clustering methods. Thus, we can use the processed data in future actions. **Figure 7** shows the Entity-Relationship diagram used at the beginning of the project to build the database. The database contains three different types of information: (i) chemical compounds data, (ii) dimensional reduction data, and (iii) clustering data.

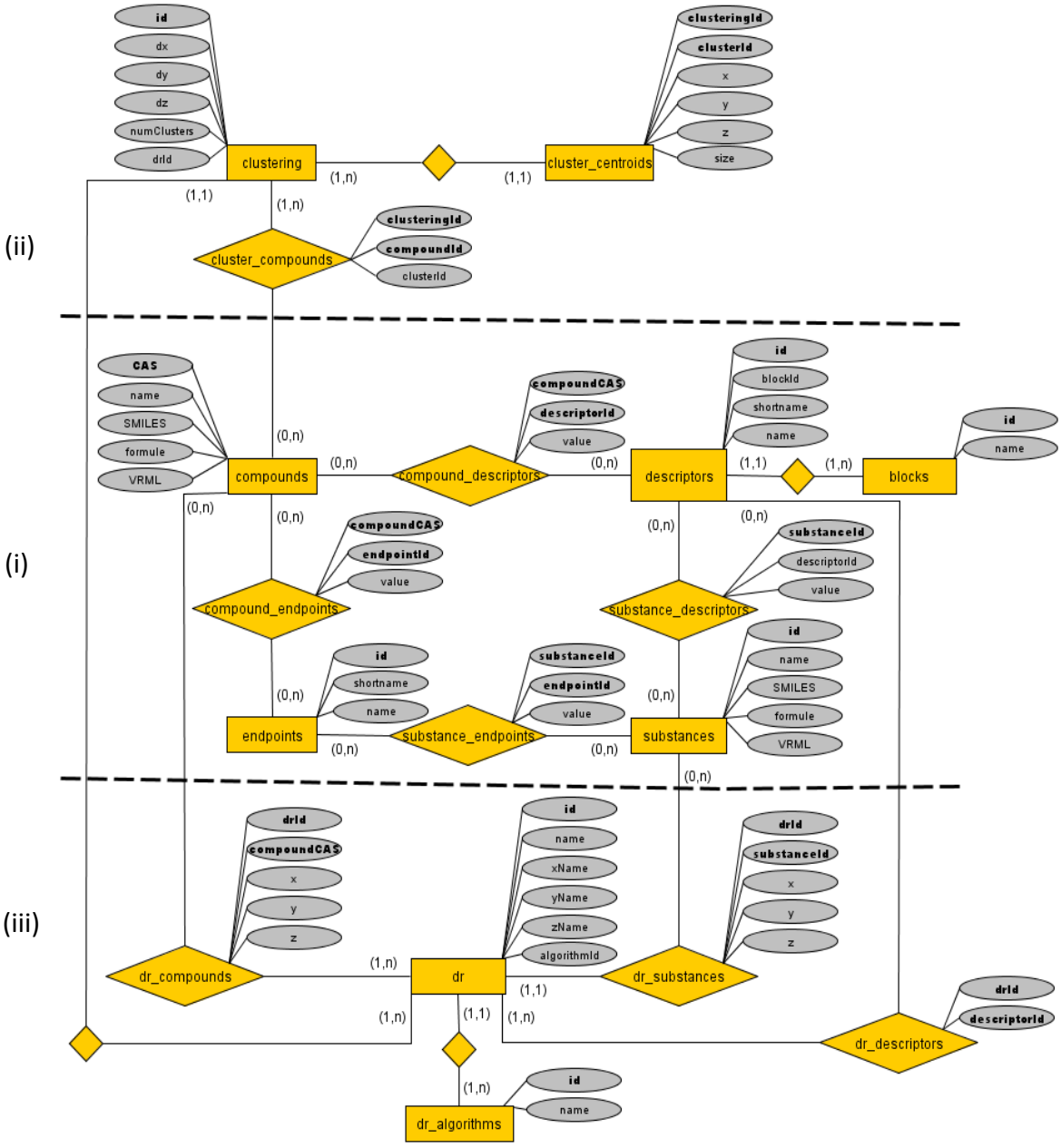

Figure 7. Entity-Relationship diagram

4.4. Implementation details

Following sections will describe how was addressed the implementation of the different tool conceptual blocks.

4.4.1. Data management

All Java database access classes implemented in the tool are fully independent and allow the use of different database management systems (through the Java Database Connectivity, JDBC). The database model has been designed to achieve a clear organization of data. The following code shows how database connection is initialized taking in account the specific connection parameters:

```
// get DB params
String useDatasource = servletContext.getInitParameter("useDatasource");
String dataSource = servletContext.getInitParameter("dataSource");
String jdbcDriver = servletContext.getInitParameter("jdbcDriver");
String connectStr = servletContext.getInitParameter("connectStr");
String dbUser = servletContext.getInitParameter("user");
String dbPassword = servletContext.getInitParameter("password");

// get DB connection
DatabaseConnectionFactory dbCF = new DatabaseConnectionFactory();
Connection conn = new Connection();

// set DB connection
conn.setConnection(dbCF.getDatabaseConnection(useDatasource, dataSource,
                            jdbcDriver, connectStr, dbUser, dbPassword));
servletContext.setAttribute("conn", conn);
```

The communication with other external sites that offer additional chemical or biological information is also implemented. In the 3D navigation window, user can access to the information provided by the *PubChem* data repository.

A data source management module has been developed allowing users to directly manipulate the application data. Sources can be created, edited and deleted. In source management window file headers are provided to build source data files related to compounds, endpoint values and descriptors values (**Figure 8**).

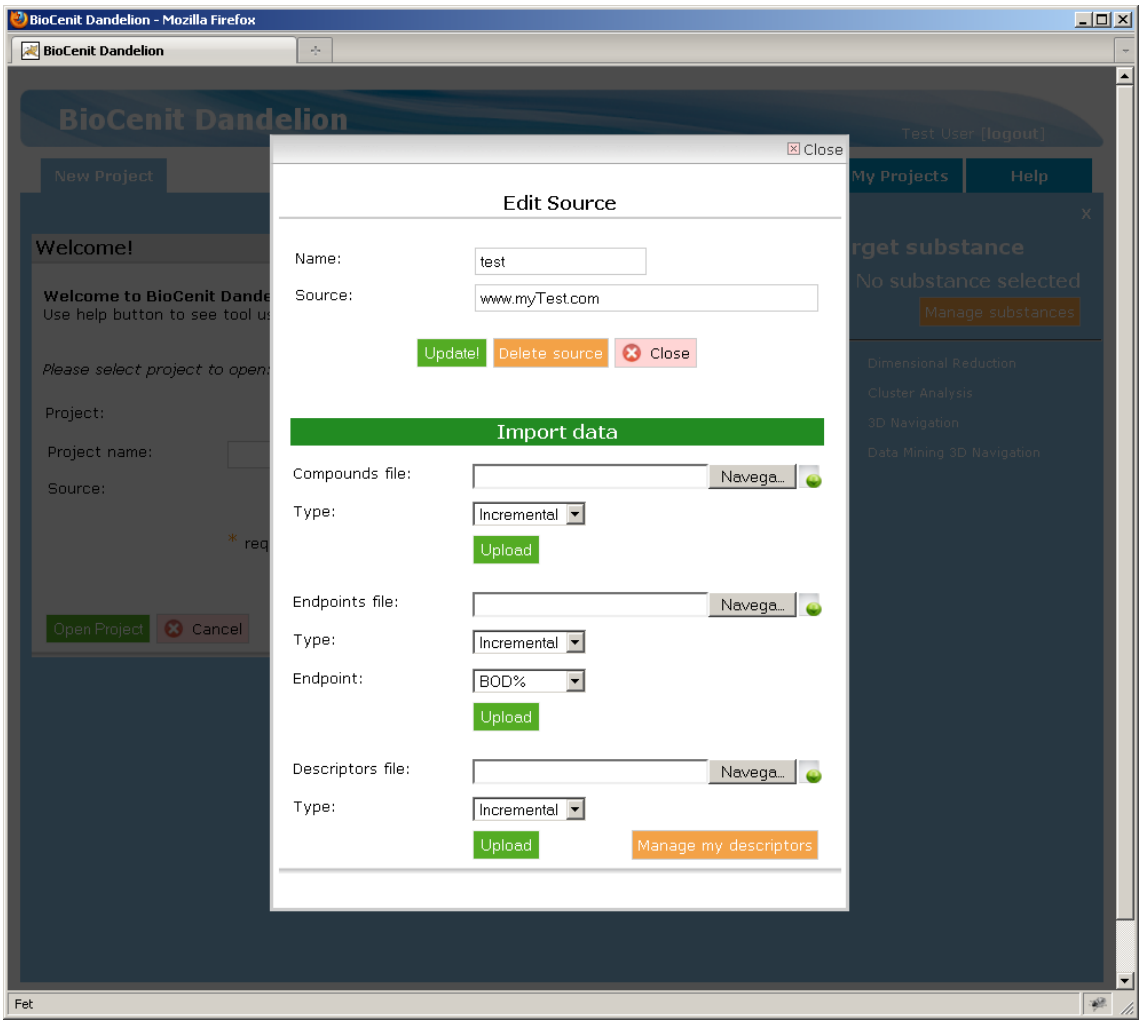

Figure 8. Data source management window

4.4.2. Data preprocessing

To simplify the process of data selection and increase the performance of the tool, the data filtering and selection is doing in the dimensional reduction module. When user creates or edits a dimensional reduction, he can use a filter to specify the set of rules to be satisfied by compounds he wants to use. These rules can affect to molecular descriptors or environmental endpoints. Compounds not satisfying selected rules will be added to a complementary chemical space. User can also choose if he wants compounds to satisfy all rules or only one of the specified rules (**Figure 9**). Issues related to data normalization are addressed in the 3D chemical space creation module.

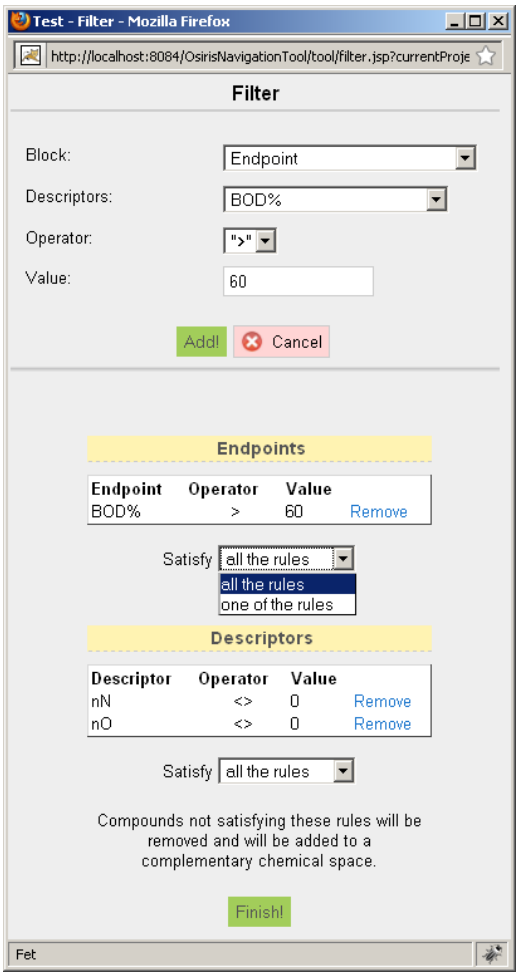

Figure 9. Data filtering and selection popup

4.4.3. Dimensional reduction

To offer different dimensional reduction algorithms we have implemented an interface to allow the use of several machine learning libraries. In the actual version of the tool, *Weka* library[1] is used and these 2 dimensional reduction algorithms are available:

- **Principal Component Analysis (PCA)**: performs an orthogonal transformation to convert a set of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. The number of principal components is less than or equal to the number of original variables. The first principal component has the largest

[1] http://www.cs.waikato.ac.nz/ml/weka

possible variance (that is, accounts for as much of the variability in the data as possible), and each succeeding component in turn has the highest variance possible under the constraint that it must be uncorrelated with the preceding components. For the chemical space visualization we will use the first three principal components (one for each axis coordinate).

- **Random Projection**: Reduces the dimensionality of the data by projecting it onto a lower dimensional subspace using a random matrix with columns of unit length (i.e., normalized).

The following code shows how *Weka* library is used for the PCA execution. PrincipalComponents and Ranker objects are assigned to an AttributeSelection object in order to achieve the dimensional reduction. Specific algorithm parameters (like variance covered, original or principal components variables selection, standardize or center values) are provided by the user in the dimensional reduction module:

```
_filter = new AttributeSelection();
_pca = new PrincipalComponents();
_ranker = new Ranker();
_pca.setVarianceCovered(_varianceParameter);
_pca.setTransformBackToOriginal(_originalSpaceParameter);
_pca.setCenterData(_centerParameter);
_filter.setEvaluator(_pca);
_filter.setSearch(_ranker);
_filter.SelectAttributes(_originalInstances);
newInstances =
          _filter.reduceDimensionality(_originalInstances).enumerateInstances();
```

4.4.4. Clustering

Different clustering algorithms are also implemented using machine learning libraries. In the actual version of the tool, also *Weka* library is used and these 3 clustering algorithms are available:

- **K-Means**: performs a partition of N observations into K clusters in which each observation belongs to the cluster with the nearest mean.

- **Hierarchical Clustering**: is a general approach to cluster analysis in which the object is to group together objects building a hierarchy of clusters.

- **Expectation Maximization**: assigns a probability distribution to each instance which indicates the probability of it belonging to each of the clusters.

Endpoints and descriptors can be selected in order to use their values in clustering algorithm. User can also specify the number of clusters to be obtained.

4.4.5. 3D visualization and navigation

Before the implementation of the tool, various 3D engines have been evaluated in order to choose the one who best fits our requirements. The standard X3D language[1] has been chosen to build the tri-dimensional chemical space scenes (**Figure 10**).

```
<Scene>
 <Shape>
  <Sphere DEF='S'/>
   <Appearance>
    <ImageTexture
     url=' "earth-topo.png" "earth-topo-small.gif" "http://www.web3d.org/x3d/content/examples/Basic/earth-topo.png"
              "http://www.web3d.org/x3d/content/examples/Basic/earth-topo-small.gif" '/>
   </Appearance>
  </Shape>
  <Transform rotation='0 1 0 1.57' translation='0 -2 1.25'>
   <Shape>
    <Text string='"Hello" "world!"' solid='false'/>
    <Appearance>
     <Material diffuseColor='0.1 0.5 1'/>
    </Appearance>
   </Shape>
  </Transform>
</Scene>
```

Figure 10. Example of X3D code

X3D is a standard mark-up file format (*XML, eXtensible Markup Language*) designed to represent 3D computer graphics. The specification was developed by the Web3D Consortium and approved by the International Standards Organization (ISO). This

[1] http://www.web3d.org

language is the successor of the Virtual Reality Modelling Language (VRML). The main advantages of X3D are:

- Use of XML: structured data, strict grammar rules, modular, platform independent and well supported.
- Various available browsers (e.g., *Xj3D*: open source browser, easy to use with java applications, implements several interfaces which facilitate the interaction within user and application).
- Various available XML managers, which allow building automatically, X3D files from MS-Excel sheets or database records.
- Excellent guide available and small learning curve.

The main drawbacks of the X3D approach are mainly related to graphics performance issues and information and support:

- Open source browsers such as *Xj3D* are not as fast as other commercial solutions (e.g., *BS Contact*).
- The use of specific hardware acceleration optimizers such as *OpenGL* engines strongly depends of X3D browser.
- The community of developers supporting X3D is not very large.

Additional 3D engines that could be suitable to be used as a replacement for X3D have also been evaluated. The main advantage of these engines resides in the use of native image libraries that fully support the hardware acceleration capabilities of current graphic cards. The graphic engines assessed have been:

- **OGRE (Object-Oriented Graphics Rendering Engine).** *OGRE* is an open source graphics engine written in C++ and designed to facilitate the implementation of applications using hardware-accelerated 3D graphics. The engine can be integrated in java applications by using the native libraries *Ogre4j*. The main advantage of this approach resides in its improved performance with respect to the X3D engines. The main drawbacks of OGRE

are that (i) require a C++ compiler and libraries to work, (ii) it is difficult to include as a simple library into a Java project, and (iii) requires advanced OpenGL programming skills.

- **JMonkey Engine (High performance scene graph-based graphics API).** *JMonkey* is a full featured graphics engine written in Java. Its main features reside in the organization of the graphical data into a tree structure, where a parent node can contain any number of children nodes. The use of a tree structure results in improved performance in the manipulation of 3D scenarios. The main drawback of this approach resides in that it requires important OpenGL programming skills resulting in a slow learning curve.

There exist many open source VRML/X3D auxiliary applications that have been assessed as potential components of the tool:

- **Xj3D** is a Java-based toolkit and X3D browser for creating X3D-compliant products. Xj3D is often used to develop new extensions and features for X3D. It is highly componentized and can be used as the basis to develop lightweight X3D applications.

- **FreeWRL** is a VRML/X3D browser for Mac OS X and Linux with support for JavaScript interfacing, the External Authoring Interface (EAI), and the X3D Scene Authoring Interface (SAI).

- **OpenVRML** includes a cross-platform VRML/X3D runtime library written in C++ and available for use under the LGPL as well as a Mozilla browser plug-in for platforms using the X Window system.

- **BS Contact** can be used as standalone viewer or embedded in web browsers based on DirectX or OpenGL. Is a cross-platform tool and can be tailored by functionality.

After testing the different alternatives, the *BS Contact* plugin was the one that fitted better our tool.

4.5. Case Study: Analysis of the chemical space for aerobic biodegradation using the MITI-1 assay

In this chapter the different parts of the tool are tested showing the full application capabilities and performance. Taking a chemical dataset as example, we will show how a user can extract valuable information from the chemical space visualization. For a detailed description of tool functional features see the User Guide (**Annex I**).

For the evaluation, an *Apache Tomcat* webserver and a *MySQL* database have been used. The *BS Contact* X3D web plugin[1] has been used for the 3D navigation.

4.5.1. Background

Persistent organic pollutants (POPs) and Persistent, Bioaccumulative and Toxic (PBT) substances are carbon based chemicals that resist degradation in the environment and accumulate in tissues of living organism, where they can produce undesirable effects on human health or the environment at certain exposure levels.

Persistent substances resist physical, biological and chemical degradation. The molecular structure of these compounds resists degradation processes that break down other pollutants in the atmosphere, water, and biota. A bioaccumulative substance concentrates in fatty tissue and tends to build-up higher concentrations in humans and other living organisms. These substances are also more likely to transfer and accumulate in the upper levels of the food chain. Usually, bioaccumulation is measured and modeled in terms of the bioconcentration factor (BCF). Some of these persistent or bioaccumulative chemicals are toxic since they cause or are suspected to cause adverse effects to humans and wildlife in ways that range from minor skin irritation to cancer.

POP and PBT chemicals are of particular concern if their release rates are higher than their rate of disappearance because in this case will accumulate in the environment.

[1] http://www.bitmanagement.com/products/interactive-3d-clients/bs-contact

The concern is that their accumulation may result in effects that are difficult to detect in early stages and that once detected are difficult to reverse.

The duration and level of exposure of living organisms to a toxic substance increases when it is persistent and bioaccumulative, in which cases it leads to higher risk of harm. Potential chronic effects resulting from long-term exposure to low levels of a toxin are relatively difficult to predict from current laboratory tests. This results in a high uncertainty in the corresponding evaluation of risk.

POP and PBT substances are at present the subject of growing attention and interest, with risk management procedures and regulations being implemented all over the world. The United Nations Environment Program has two POP initiatives: the UN-ECE Protocol (*Aarhus Protocol)*, and the UNEP POPs Convention (*Stockholm Convention)*. The convention includes a set of procedures for identifying new POPs put under global control and surveillance. Modeling has been introduced as one of the new criteria for persistence and long-range transport (LRT) evaluation (Pavan et al. 2006).

The new REACH (Registration, Evaluation and Authorization of Chemicals) legislation in the EU requires companies to assess PBT and vPvB (very Persistent and very Bioaccumulative) characteristics of chemicals being manufactured or imported into EU (European Commission, 2003) over certain annual amounts. Other countries such as Canada and Japan have already started the screening process for chemicals in their national inventory lists, and implementing restrictions whenever necessary.

4.5.2. Overview of Biodegradation modeling

The persistence of manufactured chemicals in the environment is governed by the rate of their biochemical and chemical transformations in the environment. Biodegradation is often the most important transformation process occurring in water, soil and sediments. However, the generation of reliable experimental data is very difficult (Aroson et al. 2006). Generally speaking, biodegradability can be defined as the molecular degradation of a substance resulting from the complex action of microorganisms. According to the Organization for Economic Co-operation and Development (OECD) Guidelines for QSAR validation (OECD, 2007),

biodegradation is an endpoint where special care is needed for the development of models as well as the interpretation of their predictions.

Many methods have been developed to estimate the biodegradation potential of chemicals with the purpose of predicting their ultimate fate. These methods have initially relied on correlations of degradation rates with physicochemical properties and subsequently evolved towards the use of molecular information. Examples of modeling techniques based on molecular descriptors include group contribution approaches (Alikhannidi and Takahashi 2004; Aronson et al. 2006) and quantitative and qualitative-structure-biodegradability relations (QSBRs/SBR); (Baker et al. 2004; Jaworska et al. 2003; Dzeroski et al. 1999). Additional modeling efforts have focused in the metabolic transformations that occur during the degradation process. The most probable biodegradation pathway is used in the CATABOL approach (Dimitrov et al. 2007; Sakuratani et al. 2005) which is a probabilistic approach with the assumption of first order kinetics for catabolic transformations. QSBR/SBR models rely on the complete characterization of the chemical structure to understand the mechanisms of biodegradability as well as to reliably predict biodegradation rates for new chemicals (Baker et al. 2004; Jaworska et al. 2003; Aronson et al. 2006; Hongwei et al. 2006). Several reviews have been published recently (Raymond et al. 2001; Jaworska et al. 2002; Cronin et al. 2003).

Several initiatives have recently emerged to increase acceptance of QSARs for regulatory purposes. The OECD principles for validity, applicability and acceptance of QSARs are becoming a standard in Europe. These principles can be summarized as follows: defined endpoint; unambiguous algorithm; defined domain of applicability; appropriate measures of goodness of fit, robustness and predictability; and mechanistic interpretation, if possible.

4.5.3. Defined endpoint

The data set consist of experimental biodegradation rates. Data fulfill the first principle, since it is referred to OECD guideline 301-C. The biodegradation rates were

obtained from the National Institute of Technology and Evaluation web site[1] (MITI-I Data peer-reviewed by the Chemical Products Council of the Ministry of Economy, Trade and Industry, Japan). A total of 1456 compounds were selected with their biodegradability values. 178 compounds were not processed due to specific conditions in compounds structure which did not allow descriptor calculations. The remaining set of 1278 chemicals characterizes the current endpoint for biodegradation that will be used for model training and testing.

The MITI-I test is a screening test for "ready" biodegradability in an aerobic aqueous medium, as described by OECD and EU test guidelines (OECD 301-C; EU C.4-F). The MITI-I test was developed in Japan and it constitutes one of the six standardized "ready" biodegradability tests described by EU and OECD regulations. For the MITI-I test, 100 mg/l of test substance are inoculated and incubated with 30 mg/l of sludge. Biological oxygen demand (BOD) is measured continuously during a 28-day test period. The pass level for ready biodegradability is reached if the BOD amounts to ≥60% of theoretical oxygen demand (ThOD). The reported data for each chemical consists of biodegradation yields measured indirectly, through biological oxygen demand (% BOD), the test period (usually 4 weeks) and direct biodegradation measures using total organic carbon (TOC) determined by chromatographic techniques (high performance liquid chromatography and gas chromatography).

Other techniques related to data uncertainty reduction may be applied to further refine the quality of these experimental data. This would be of specific interest in those cases in which contradictory information is present.

[1] http://www.safe.nite.go.jp

4.5.4. Algorithms selection

Calculation of molecular descriptors.

Molecular descriptors of the 1278 compounds were calculated by Talete Dragon 6 software[1]. **Table 1** shows the block of constitutional descriptors that we will use as input for the Weka algorithms and for the navigation tool.

Name	Symbol	Min value	Max value	Avg. value	Std. Dev.
molecular weight	MW	26.04	1177.8	202.010	128.320
average molecular weight	AMW	4.01	50.544	8.484	4.598
sum of atomic van der Waals volumes (scaled on Carbon atom)	Sv	2.527	110.03	16.119	10.627
sum of atomic Sanderson electronegativities (scaled on Carbon atom)	Se	3.884	196.71	26.850	18.949
sum of atomic polarizabilities (scaled on Carbon atom)	Sp	2.67	122.43	17.335	11.826
sum of first ionization potentials (scaled on Carbon atom)	Si	4.415	228.1	30.292	22.062
mean atomic van der Waals volume (scaled on Carbon atom)	Mv	0.455	1.144	0.623	0.103
mean atomic Sanderson electronegativity (scaled on Carbon atom)	Me	0.95	1.316	1.014	0.049
mean atomic polarizability (scaled on Carbon atom)	Mp	0.496	1.367	0.665	0.107
mean first ionization potential (scaled on Carbon atom)	Mi	1.022	1.381	1.132	0.033
number of atoms	nAT	4	201	26.755	19.422
number of non-H atoms	nSK	2	85	13.141	8.192
number of bonds	nBT	3	200	26.739	19.668
number of non-H bonds	nBO	1	88	13.125	8.711
number of multiple bonds	nBM	0	48	5.459	5.448
sum of conventional bond orders (H-depleted)	SCBO	1	104	16.420	10.705
number of rotatable bonds	RBN	0	68	3.750	5.946
rotatable bond fraction	RBF	0	0.389	0.104	0.094
number of double bonds	nDB	0	9	1.006	1.248
number of triple bonds	nTB	0	2	0.041	0.245
number of aromatic bonds	nAB	0	48	4.411	5.307
number of Hydrogen atoms	nH	0	124	13.614	12.440
number of Carbon atoms	nC	1	73	9.890	7.174
number of Nitrogen atoms	nN	0	8	0.649	1.063
number of Oxygen atoms	nO	0	12	1.714	1.715
number of Phosphorous atoms	nP	0	2	0.037	0.192
number of Sulfur atoms	nS	0	6	0.113	0.440

[1] http://www.talete.mi.it/products/dragon_description.htm

number of Fluorine atoms	nF	0	27	0.201	1.794
number of Chlorine atoms	nCL	0	12	0.412	1.146
number of Bromine atoms	nBR	0	10	0.110	0.712
number of Iodine atoms	nI	0	1	0.002	0.048
number of Boron atoms	nB	0	0	0.000	0.000
number of heavy atoms	nHM	0	12	0.687	1.389
number of heteroatoms	nHet	0	29	3.251	2.826
number of halogen atoms	nX	0	27	0.726	2.200
percentage of H atoms	H%	0	72.7	47.773	14.257
percentage of C atoms	C%	9.1	60.7	36.741	9.356
percentage of N atoms	N%	0	40	3.047	5.300
percentage of O atoms	O%	0	50	7.271	7.446
percentage of halogen atoms	X%	0	80	4.499	11.933
number of sp3 hybridized Carbon atoms	nCsp3	0	61	4.646	6.233
number of sp2 hybridized Carbon atoms	nCsp2	0	32	5.189	5.058
number of sp hybridized Carbon atoms	nCsp	0	2	0.055	0.291

Table 1. Constitutional descriptors

Algorithms for feature selection.

The key to success in a classification task is the selection of the attributes used as the input to the algorithm. Finding the most suitable set of descriptors is a task that occurs in many contexts and involves techniques such as machine learning, pattern recognition and data mining. Feature selection methods are grouped in two categories: filter methods, which evaluate the parameters on heuristic-based general characteristics of the data (for example, correlations), and wrapper methods, which use the modeling algorithm as the feature evaluation function (Hall 1998).

The correlation based feature selection (CFS) filter is an effective way for parameter selection (Hall 1998). It selects a parameter if it correlates with the decision outcome but not with any other parameter that has already been selected. In this study genetic algorithms provided the global search framework for the CFS filter, which in turn used its built-in functionality to optimize the parameters selected. CFS uses a best-first-search heuristic. This heuristic algorithm takes into account the usefulness of individual features for predicting the class along with the level of inter-correlation

among features. The method is based in the idea that "good feature subsets contain features highly correlated with the class, yet uncorrelated with each other". The CFS first calculates a matrix of feature-class and feature-feature correlations from the training data. Then a score of a subset of features is assigned by the heuristic defined as:

$$merit_s = \frac{k\bar{r}_{cf}}{\sqrt{k + k(k+1)\bar{r}_{ff}}} \quad [1]$$

where $merit_s$ is the merit of a feature subset S containing k features, r_{cf} is the mean feature-class correlation, and r_{ff} is the average feature-feature correlation. The numerator of equation (1) can be considered as an indicator of how predictive of the class group the selected features are and the denominator as an indicator of how much redundancy there is among features.

A measure based on conditional entropy is used to measure correlations between features and class, and between features. Continuous features are transformed to categorical features using discretization methods. If X and Y are discrete random variables, equations (2) and (3) give the entropy of Y before and after observing X,

$$H(Y) = -\sum_{y \in Y} p(y) \log_2 p(y) \quad [2]$$

$$H(Y|X) = -\sum_{x \in X} p(x) \sum_{y \in Y} p(y|x) \log_2 p(y|x) \quad [3]$$

Equation (3) is known as the information gain and accounts for the amount of information gained about Y after observing X, which is equal to the amount of information gained about X after observing Y (Quinlan 1993).

Algorithms for model development.

SAR/QSAR models will focus on predicting the target endpoint for a specific compound from a vector representation of its molecular structure. A widely accepted family of machine learning methods is the decision tree, also known as recursive partitioning (Breiman et al. 1984; Quinlan 1993). Decision trees represent a

supervised approach to classification. A decision tree is a simple structure where non-terminal nodes represent tests on one or more attributes and terminal nodes reflect decision outcomes (Bauer and Kohavi 1999). The algorithms examined in this study are the tree J48 (C4.5 derivative), the instance-base learners IBk and Kstar, Random Tree, the ensemble of trees Random Forest, and logistic model trees (LMT). The Random Forest algorithm is the one that yields consistently better results because it is an ensemble technique based in random trees (Breiman 2001). The Weka software package provides implementations of the above classification algorithms.

The J48 classifier is a simple C4.5 decision tree for classification which induces a binary tree structure in the data. A decision tree algorithm involves the following actions:

(i) Choose an attribute that best differentiates the output attribute values

(ii) Create a separate tree branch for each value of the chosen attribute.

(iii) Divide the instances into subgroups so as to reflect the attribute values of the chosen node.

(iv) Terminate the attribute selection process for each subgroup if

- all members of a subgroup have the same value for the output attribute. In this case label the branch on the current path with the specified value;
- the subgroup contains a single node or no further distinguishable attributes can be determined. Also in this case label the branch with the output value seen by the majority of remaining instances.

The instance-based learner IBk (Aha and Kibler 1991) is similar to Instance Based classification (IB1) that is equivalent to the well-known K-nearest neighbor classifier (KNN). The main difference with KNN is that IBk processes the training sets incrementally and ignores missing values. In IBk it is possible to define the desired number of nearest neighbors. The advantage of this is to widen the numbers of

instances considered. However, this is very memory intensive, increasing memory requirements with the number of additional nearest neighbors considered.

The Instance based learned K-star is another instance-based learning algorithm that uses entropy as a distance measure in the K-Nearest Neighbor transformation (Clearly and Trigg 1995). As a consequence, it shows good results in the management of missing values, real valued attributes and symbolic data.

The logistic model tree (LMT) constructs a tree-structured classifier with logistic regression functions at the leaves. The classic logistic regression approach models log(p/(1−p) as a linear function of the features where p represents the probability of a feature vector x belonging to class i. It can be written as,

$$\log\left(p/_{1-p}\right) = \beta_0 + \beta^T x \qquad [5]$$

where the β vector and the scalar βo are parameters to be determined and x denotes the feature vector for each molecule. The LMT algorithm follows the "divide and conquer" principle in which a complex set of data is divided into smaller subsets in a way that a simple linear logistic regression model can adequately fit the data in each subset.

Random Forest (RF), which was developed by (Breiman 2001), is an ensemble method that combines several individual classification trees. Prediction is made by aggregating (majority vote for classification) the predictions of the ensemble. Two types of randomness, bootstrap sampling and random selection of input variables, are used in the algorithm to ensure that all the classification trees are dissimilar and uncorrelated. Growing a forest of trees and using randomness in building each classification tree in the forest leads to better predictions compared to a single classification tree and helps to make the algorithm robust to noise an uncertainty in the data set. Similar to most classifiers, RF can suffer from the curse of learning from an extremely imbalanced training data set. As it is constructed to minimize the overall error rate, it will tend to focus more on the prediction accuracy of the majority class, which often results in poor accuracy for the minority class.

4.5.5. Applicability domain

The applicability domain of a SAR model is the chemical and response joint space in which the model makes predictions with a given reliability (Netzeva et al. 2005). Thus, it is the information space on which the training of the model has been carried out and for which it is applicable to make predictions for new compounds.

The characterization of the chemical space involves several actions: (i) data cleaning and conditioning; (ii) selection of the most relevant information to develop the model; and (iii) design of proper training and test sets. The most demanding validation procedure is to use an external set of compounds (validation set) that were not used at any stage of model development. These compounds should be structurally representative of the studied chemical domain (Jaworkska et al. 2007; Tropsha et al. 2006). The proper establishment of the application domain for a predictive model defines its validity limits. Predictions corresponding to compounds defined within the domain can be interpreted as interpolations. Accordingly, the response of compounds outside the domain are extrapolations and thus unreliable.

4.5.6. Characterization of the Chemical Space for Biodegradation via the Navigation Tool

The most relevant molecular descriptors were selected using the Weka implementation of the CFS algorithm. The merit of the best subset found is of 0.16 and includes the following ten descriptors: {MW, Mp, Mi, RBF, nN, nHM, nHet, O%, X%, nCsp2}.

Figure 11 depicts the distribution of BOD and non-BOD values for each of these descriptors.

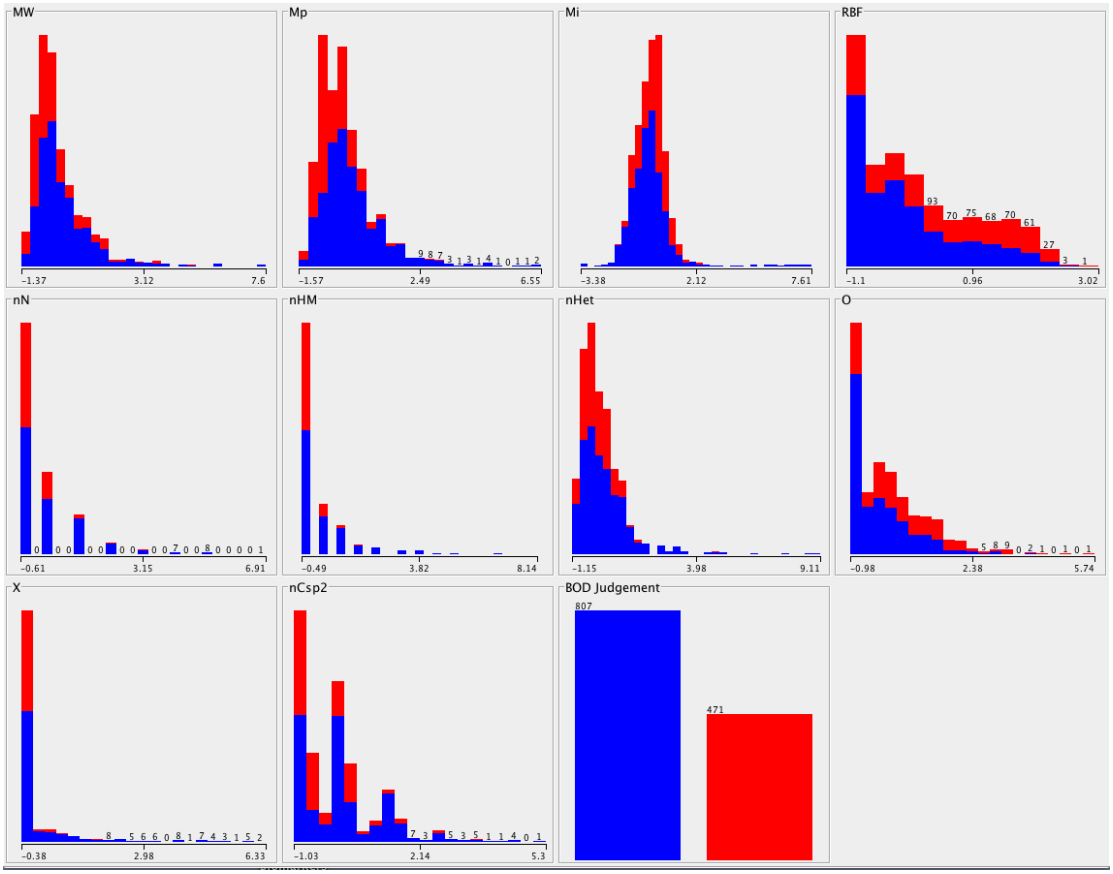

Figure 11. Frequency histograms of the two biodegradation classes for each descriptor (nonBOD in blue and BOD in red)

It is clear from the inspection of **Figure 11** that there is no single descriptor which is capable of distinguishing between BOD and non-BOD chemicals. If we use the navigation tool to visualize the chemical space as a function of the three most relevant variables (MW, Mp, Mi) we obtain the following visualization:

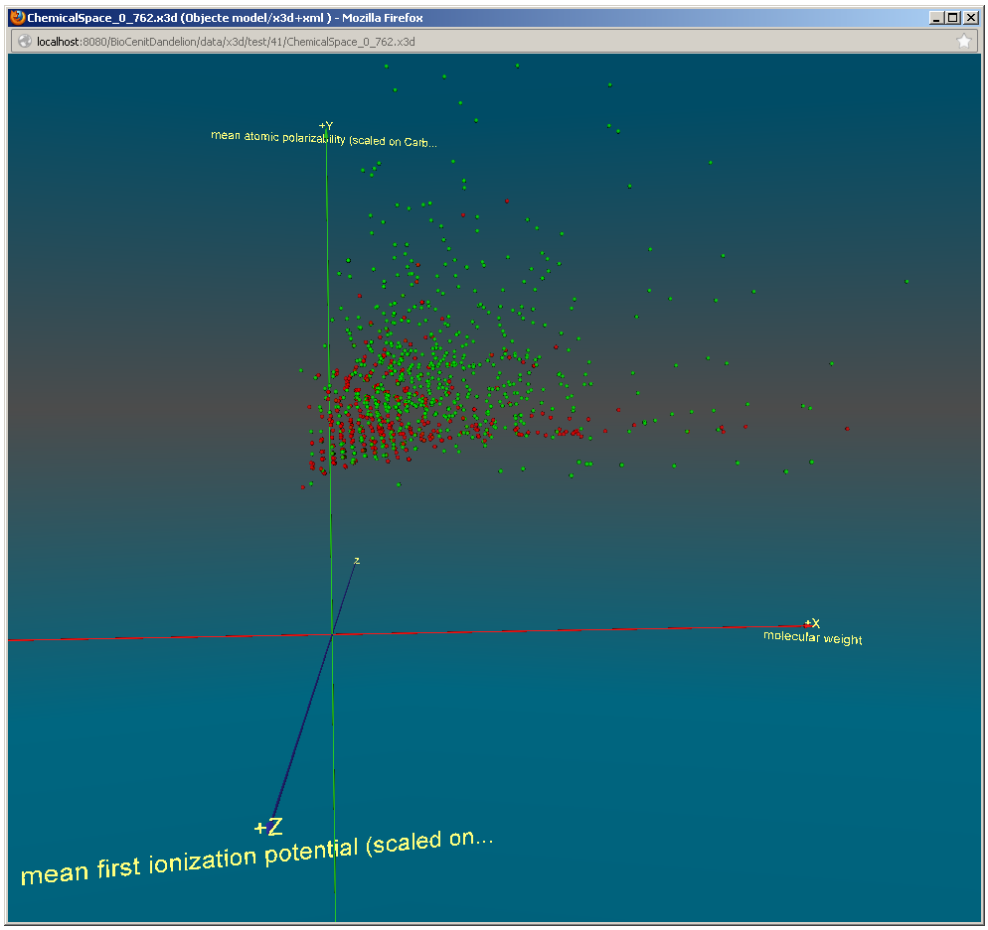

Figure 12. Visualization of the three most relevant variables (green color indicates non BOD chemicals whereas red spheres correspond to BOD chemicals)

In **Figure 12** it can be observed that chemicals become persistent (non-biodegradable) when the atomic polarizability increases. Also the proportion of persistent chemicals increases with molecular weight (MW).

The tool can also be used to embed the complete ten dimensional space into a three dimensional representation via dimension reduction. **Figure 13** shows the application of PCA to project the complete chemical space over the three main principal components.

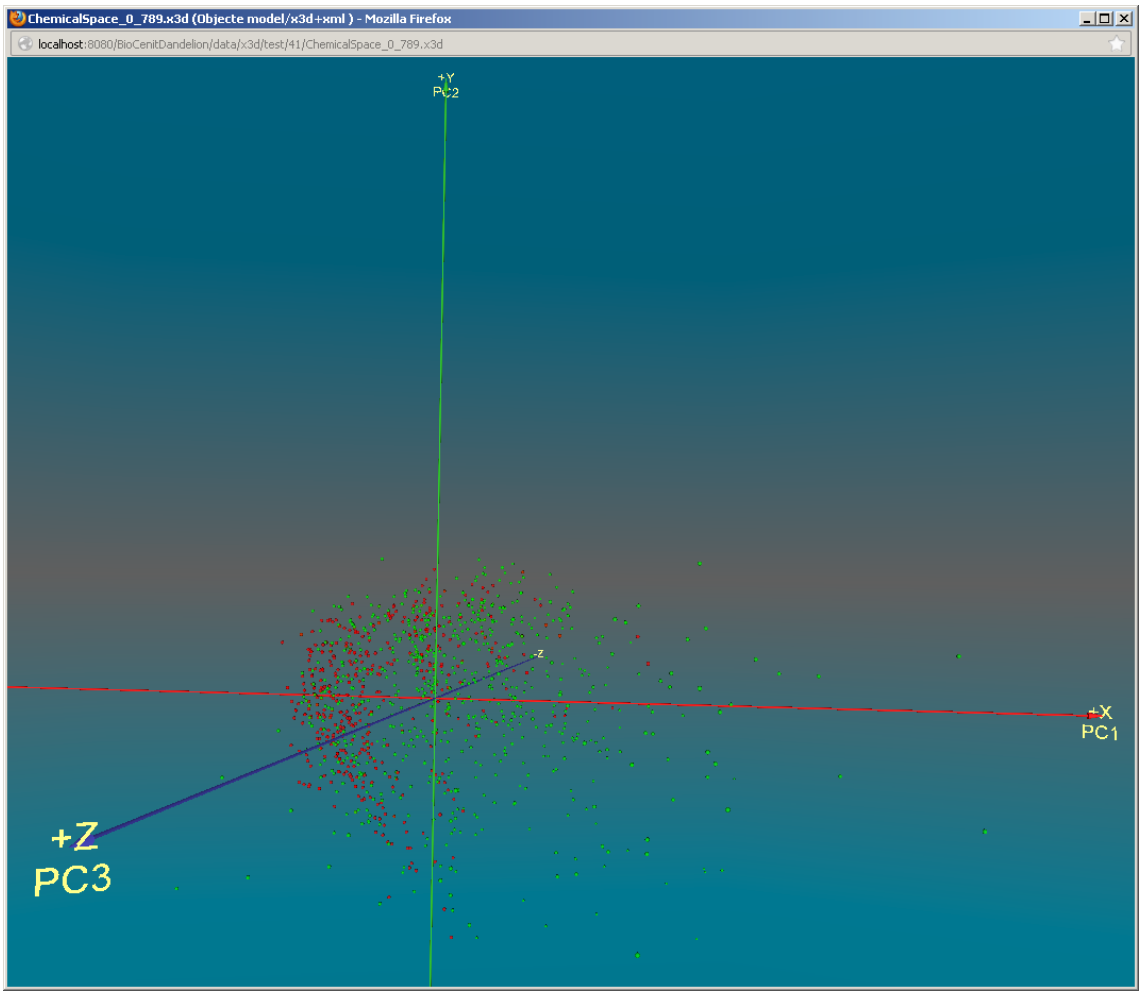

Figure 13. PCA-based projection of the chemical space (green color indicates nonBOD chemicals whereas red spheres correspond to BOD chemicals)

From the analysis of **Figure 13** it can be inferred that PC1 and PC3 are the most influential for grouping the biodegradable chemicals (red dots). The expressions corresponding to the first three PCs are:

PC1= 0.529Mp+0.499nHM+0.385X+0.319MW+0.246nHet+0.234nCsp2-0.23Mi-0.177RBF-0.142O+0.02nN

PC2=0.567nHet+0.45 Mi+0.422RBF+0.35 MW+0.242O-0.19Mp+0.186nN+0.162X-0.129nCsp2+0.052nHM

PC3=0.584nCsp2-0.487X-0.368Mi+0.337MW+0.303nN+0.192RBF-0.166nHM+0.082O+0.072nHet-0.07Mp

It can be seen that the mean atomic polarizability and the number of heteroatoms are the most influential variables for PC1 and PC2. These findings are consistent with previous results reported in the literature regarding the role of polarizability and the

presence of atoms different than C and H in the biodegradation potential of chemicals.

The similarity in the profiles of the components of the chemical space can be analyzed with the Navigation Tool using the clustering feature. **Figure 14** shows the K-means clustering obtained from the set of ten molecular descriptors.

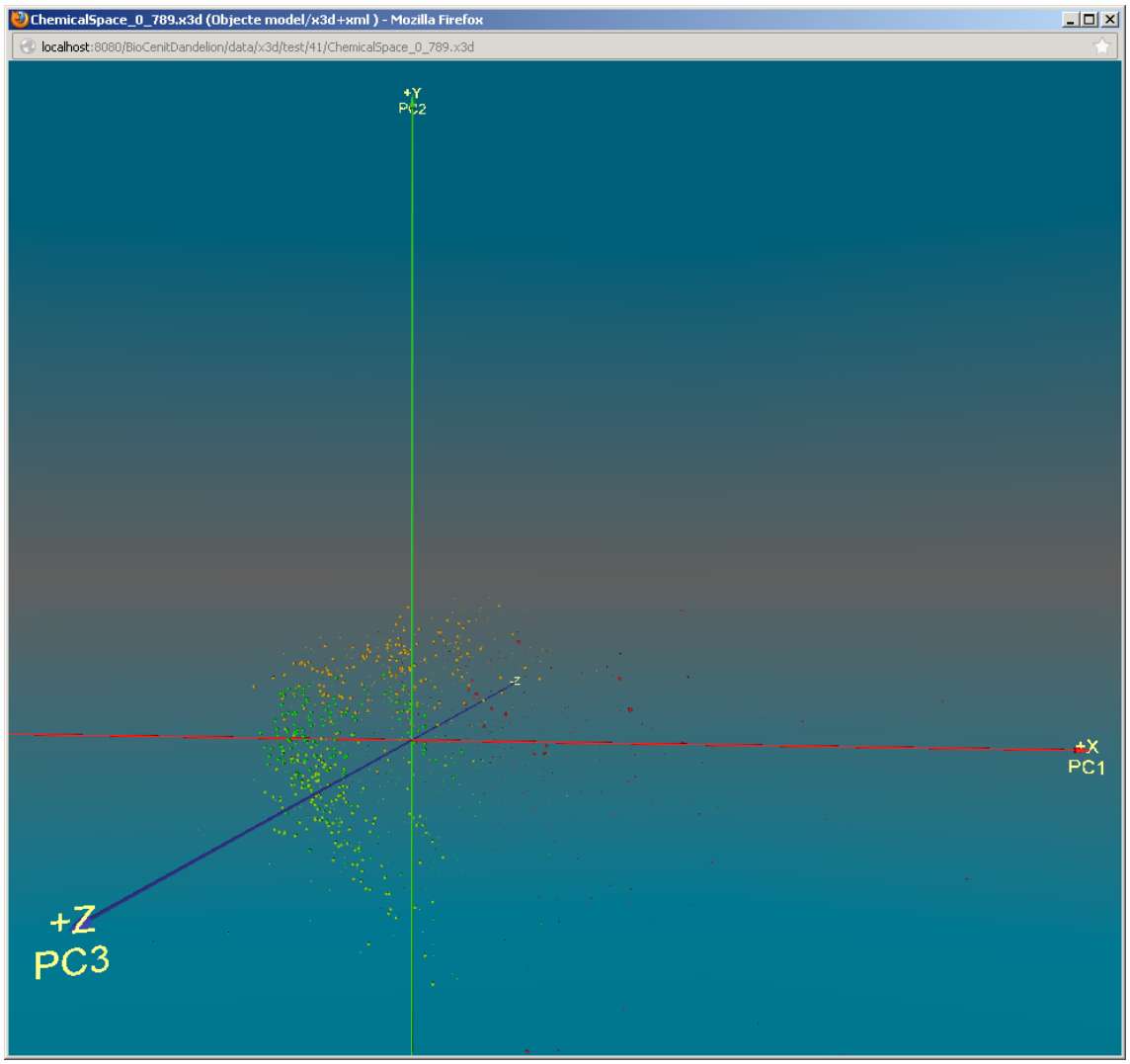

Figure 14. PCA projection of the clustered chemical space represented by the set of ten molecular descriptors. Colors indicate cluster assignment

The analysis of figure 14 reveals similarities in different regions of the chemical space. In the above representation the size of the spheres is proportional to the biodegradation potential of each chemical (small: non BOD; large: BOD).

Outliers can be identified and inspected via the "near compounds view". **Figure 15** and **Figure 16** show the neighborhood of 2-tert-Butoxyethanol (white non BOD outlier located within a group of red BOD chemicals; grey spheres represent compound collisions).

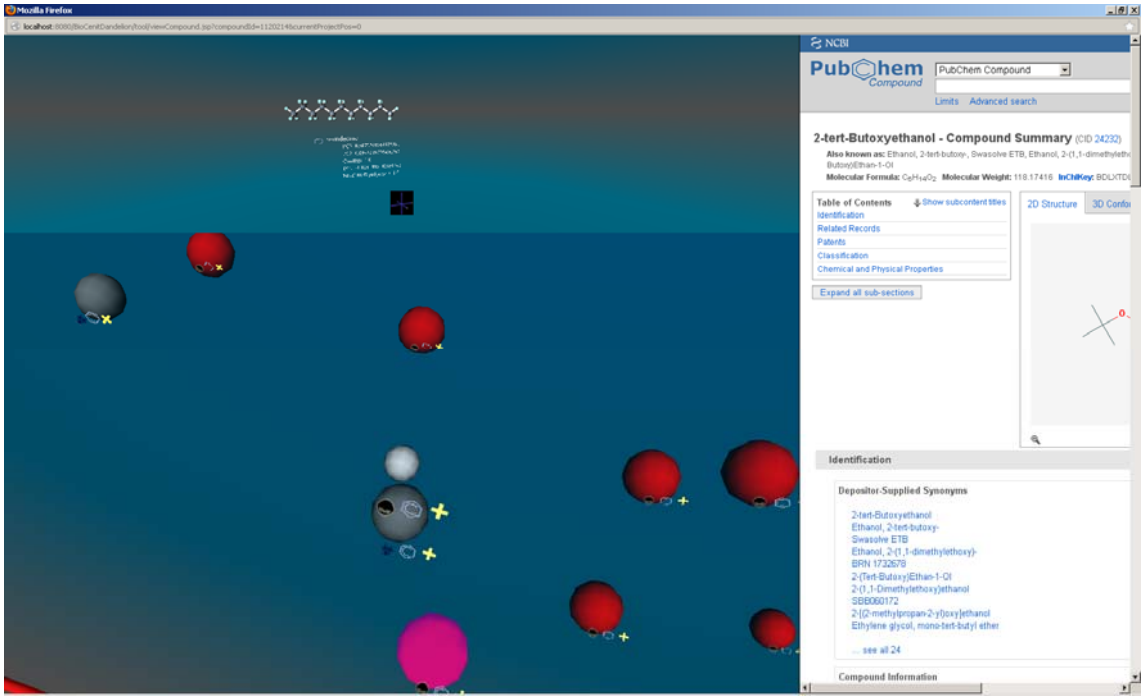

Figure 15. Neighborhood of 2-tert-Butoxyethanol (white sphere)

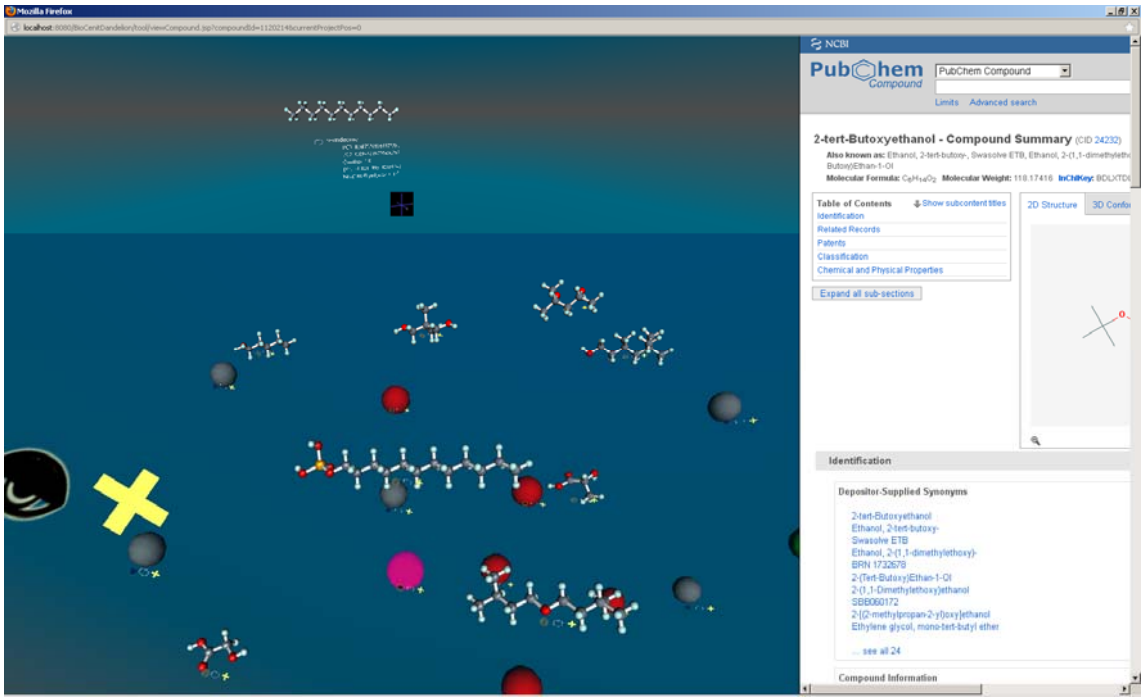

Figure 16. Neighborhood of 2-tert-Butoxyethanol showing some molecular structures

Although there are similarities in the molecular descriptor profiles for this group of chemicals the analysis of the chemical structures in **Figure 16** reveals significant structural differences.

4.5.7. Development of BOD models

After screening the performance of several classifiers, the Random Forest ensemble of trees was selected as the most suitable algorithm to build the current biodegradation model with the selected set of ten molecular descriptors. The model consisted of an ensemble of ten random trees and was evaluated via stratified 10-fold cross validation. The summary report of the model performance is:

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        1014               79.3427 %
Incorrectly Classified Instances       264               20.6573 %
Kappa statistic                          0.5514
Mean absolute error                      0.2535
Root mean squared error                  0.3901
Relative absolute error                 54.4539 %
Root relative squared error             80.8698 %
Total Number of Instances             1278

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.851     0.306      0.827     0.851     0.839      0.851    0
                 0.694     0.149      0.732     0.694     0.712      0.851    1
Weighted Avg.    0.793     0.248      0.792     0.793     0.792      0.851

=== Confusion Matrix ===

   a   b   <-- classified as
 687 120 |   a = 0
 144 327 |   b = 1
```

It is interesting to note that the outlier identified in **Figure 15** and **Figure 16** is also one of the misclassified chemicals in the above model. The navigation tool can be used as a diagnostic utility to inspect model predictions. Thus, by inspecting the structure of the chemical space near misclassified chemicals the user is able to improve model accuracy. The tool can also be used to explain the output of the models by identifying potential mechanistic interpretations via the inspection of the chemical space.

5. Conclusions and future work

After the evaluation of the tool, we can conclude that the main objectives of the project have been achieved. The application provides an advanced virtual platform for visual screening of the chemical and biological space compounds that facilitates the analysis of the general structure of the chemical and biological space in terms of the environmental endpoints.

As further work, the implementation of a user interface based in augmented reality is proposed. Augmented reality (AR) (Azuma 1997; Haller et al., 2006) is an environment that includes both virtual reality and real-world elements. For instance, an AR user might wear translucent goggles; through these, he could see the real world, as well as computer-generated images projected on top of that world. The fundamental characteristics of an augmented reality system are that (i) it combines real and virtual scenarios, (ii) the system is interactive in real-time, and (iii) it is presented in three dimensions. The implementation of a user interface based in augmented reality techniques in the navigation tool would increase the ways in which users may interact with the representation of the chemical space. These interfaces will permit the direct manipulation of the virtual models and will enable collaborative exploration of the chemical space. The prototype implementation of the AR interface will use open-source AR toolkits and low cost visualization devices.

6. References

Alikhannidi S., Takahashi Y. (2004) **Pesticide Persistence in the Environment. Collected Data and Structure-Based Analysis**. J. Comput. Jpn, 3(2):59-70.

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

Aronson D., Boethling R., Howard P., Stiteler W. (2006) **Estimating biodegradation half-lives for use in chemical screening**. *Chemosphere*, 63:1953-1960.

Azuma, RT. (1997) **A Survey of Augmented Reality**. Presence: Teleoperators and Virtual Environments 6, 4:355-385.

Baker JR., Gamberger D., Mihelcic JR., Sabjić A. (2004) **Evaluation of artificial intelligence based models for chemical biodegradability prediction**. Molecules, 9:989-1004.

Bauer E., Kohavi R. (1999) **An Emprical comparison of voting classification algorithms: Bagging, Boosting and variants**. Mach. Learn., 36:105-139.

Breiman L., Friedman J., Stone C., Olshen R. (1984) **Classification and Regression Trees**. Chapman & Hall: Boca Raton, FL.

Breiman L. (2001) **Random Forests**. Mach. Learn., 45:5-32.

Clearly JG., Trigg L. (1995) **K*. An instance-based learner using an entropic distance measure**. Proc. Int. Conference on Machine learning, Morgan Kaufmann.

Cronin MTD., Walker JD., Jaworska JS., Comber MHI., Watts CD., Worth AP. (2003) **Use of QSARs in International Decision Making Frameworks to Predict Ecologic Effects and Environmental Fate of Chemical Substances**. Environmental Health Perspectives, 11 (10):1376-1390.

Dean PM. (1995) **Molecular Similarity in Drug Design**, Chapman & Hall, 291-331.

Dimitrov S., Pavlov T., Nedelcheva D., Reuschenbach P., Silvani M., Bias R., Combers M., Low L., Lee C., Pakerton T., Mekeyan O. (2007) **A kinetic model for predicting biodegradation**. SAR and QSAR in Environmental Res., 18(5-6): 443-457.

Dzeroski S., Blockeel H., Kramer S., Kompare B., Pfahringer B., and Van Laer W. (1999) **Experiments in predicting biodegradability**. Proceedings of the Ninth International Workshop on Inductive Logic Programming (S. Dzeroski and P. Flach, eds.), LNAI, vol. 1634, Springer, 80-91.

Feher M, Schmidt JM. (2003) **Property distributions: differences between drugs, natural products, and molecules from combinatorial chemistry**. J Chem Inf Comput Sci. 2003 Jan-Feb, 43(1):218-27.

Glen RC., Adams SE. (2006) **Similarity Metrics and Descriptor Spaces – Which Combinations to Choose?**. QSAR & Combinatorial Science, 25:1133-1142

Gütlein M., Karwath A., Kramer S. (2012) **CheS-Mapper - Chemical Space Mapping and Visualization in 3D**. Journal of Cheminformatics 2012, 4:7.

Hall MA. (1998) **Correlation-based Feature Selection for Machine Learning**. Ph.D. diss. Dept. of Computer Science, Waikato Univ.

Hongwei Y., Jiang Z., Shi S. (2006) **Aromatic compounds biodegradation under anaerobic conditions and their QSBR models**. Science of the Total Environment, 358(1-3):265-276.

Jaworska J., Dimitrov S., Nikolova N., Mekenyan O. (2002) **Probabilistic assessment of biodegradability based on metabolic pathways: CATABOL system**. SAR QSAR Environ. Res., 13:307-323.

Jaworksa J., Boethling R., Howerds P. (2003) **Recent developments in broadly applicable structure-biodegradability relationships**. Environmental Toxicology and Chemistry, 22(8):1710-1723.

Jaworska J., Nikolova-Jeliazkova N. (2007) **How can structural similarity analysis help in category formation**. SAR & QSAR in Environmental Res, 18(3-4):195-197.

Keim DA. (2002) **Information Visualization and Visual Data Mining**. IEEE Transactions on Visualization and Computer Graphics, 8(1):1-8

Haller, M., Billinghurst, M., Thomas, B. (2006) **Emerging Technologies of Augmented Reality: Interfaces and Design**. Idea Group Publishing 2006.

Larsson, J., Gottfries, J., Muresan, S., Blacklund, A. (2007) **ChemGPS-NP: Tuned for navigation in Biologically relevant Chemical space**. J. Nat. Prod., 70:789-794.

Maldonado AG., Doucet JP., Petitjean M., Fan BT. (2006) **Molecular similarity and diversity in cheminformatics: from theory to applications**, Mol. Divers., 10:39-79.

Matero S., Lahtela-Kakkonen M., Korhonen O., Ketolainen J., Lappalainen R., Poso A. (2006) **Chemical space of orally active compounds**. Chemometrics and Intelligent Laboratory Systems, 84: 134-141.

Netzeva TI., Aptula A., Benfenati E., Cronin M., Gini G., Lessigiarska I., Maran U., Vracko M., Schüürmann G. (2005) **Description of the electronic structure of organic chemicals using semi empirical and ab initio methods for development of toxicological QSAR**. Journal of Chemical Information and Modeling, 45:106-114.

O'Driscoll C. (2004) **A Virtual Space Oyssey**. Horizon Symposia, Charting Chemical Space.

Oliveira MCFD., Levkowitz H. (2003) **From Visual Data Exploration to Visual Data Mining: A Survey**. IEEE Trans. Vis. Comput. Graph, 378-394.

Pavan M., Wroth AP., Netzeva TI. (2006) **Review of QSAR for bioconcentration**. Joint Research Centre.

Quinlan, JR. (1993) **C4.5. Programs for Machine Learning**. Morgan Kaufmann.

Raymond JW., Rogers TN., Shonnard DR., Kline AA. (2001) **A review of structure-based biodegradation estimation methods**. Journal of hazard Materials, 84:189-215.

Rupp M., Schneider P., Schneider G. (2008) **Distance Phenomena in High-Dimensional Chemical Descriptor Spaces: Consequences for Similarity-Based Approaches**. J Comput Chem. 2009 Nov 15, 30(14):2285-96.

Sakuratani Y., Yamada J., Kasai K., Noguchi Y., Nishihara T. (2005) **External validation of the biodegradability prediction model CATABOL using data sets of existing and new chemicals under the Japanese Chemical Substances Control Law**. SAR and QSAR in Environmental Res., 16(5):403-431.

Tropsha A., Gramatica P., Gombar VK. (2003) **The importance of being Earnest: validation is the absolute essential for successful application and interpretation of QSPR models**. QSAR & Combinatorial Science, 22:69-76.

Van Deursen, R., Reymond, J-L. (2007) **Chemical Space Travel**. ChemMedChem, 2:636-640.

Yamashita, F., Itoh, T., Hara, H., Hashida, M. (2006) **Visualization of Large-Scale Aqueous Solubility data using a novel hierarchical data visualization technique**. J. Chem. Inf. Model., 46:1054-1059.

Annex I – User Guide

To get a 3D chemical space, user may follow next steps:

- *User login*: user provides username and password to access application.
- *Open project*: user selects an existing project or creates a new one selecting the source he wants to use.
- *Dimensional reduction*: selection of desired molecular descriptors and application of dimensional reduction algorithm. Rules can be added to filter compounds in the exceptions filter pop-up.
- *Cluster analysis (optional)*: selection of desired endpoints and molecular descriptors and application of clustering algorithm. Once it is created, it can be assigned to radius or colour in the 3D navigation form.
- *3D chemical space creation*: selection of dimensional reduction and endpoints (or clustering) to use for 3D coordinates, radius, and colour. User can also choose a normalization algorithm for each attribute and apply clustering to optimize graphics renderization in 3D navigation. When selecting colour endpoint, user can choose a "proximity clustering" in order to differ compounds groups.

1. Entry point

The entry point of the tool is the login page shown in Figure 1, where the user will introduce his credentials in the form of a username and password to access application.

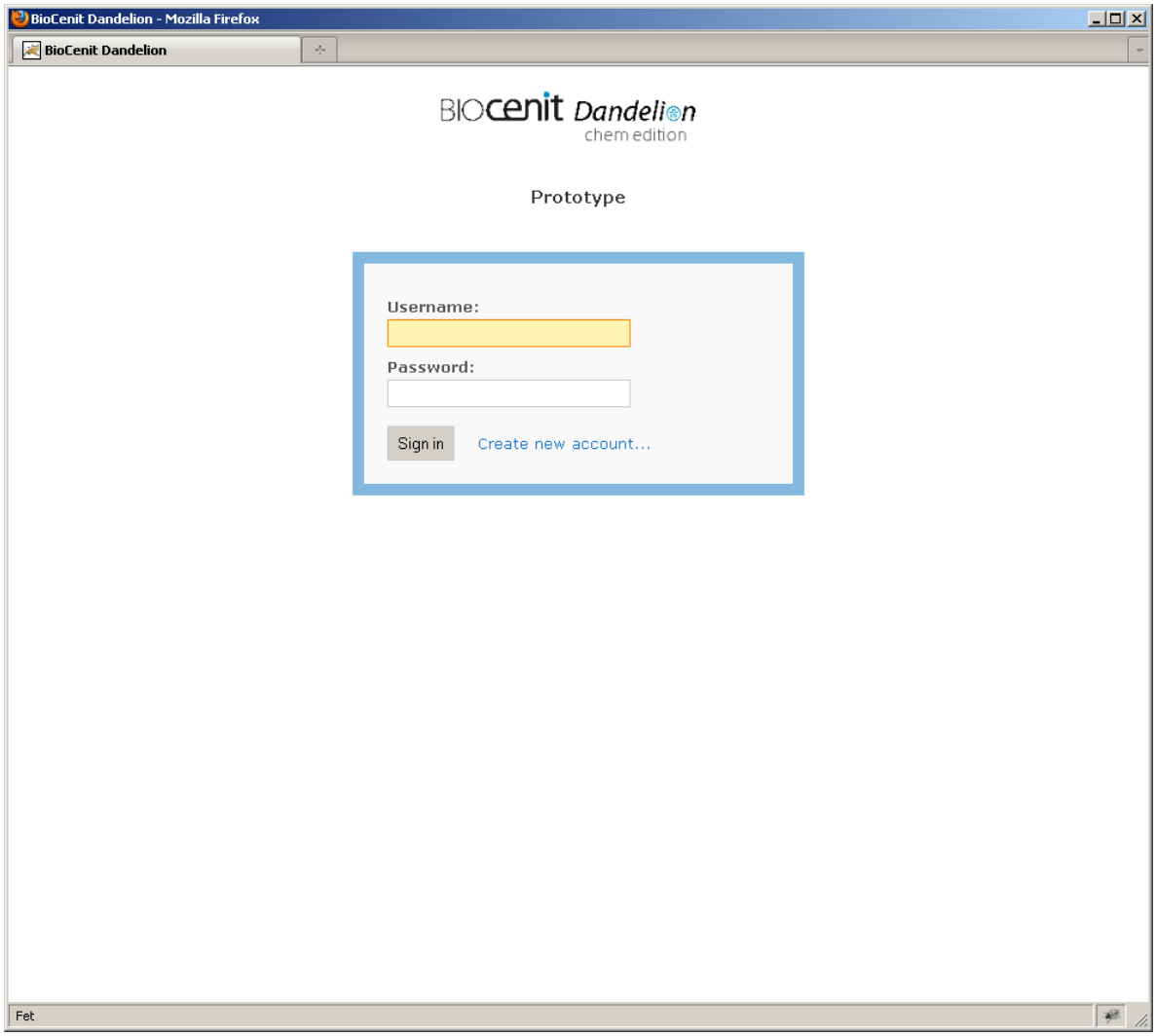

Figure 1. Initial user validation window to access the application

Once user has been authenticated, the welcome page is showed, asking the user to select a project to open (Figure 2). In this page the user also has the option to manage data sources and target substances.

From every page the user can access the help page, manage projects, create new projects, and change current project tab. The user can have a maximum of 8 projects opened at the same time. A logout link is also provided. The lateral menu is the central navigation element of the tool once a project has been selected, and provides access to the main functionalities: dimensional reduction, cluster analysis, 3D navigation and data mining 3D navigation.

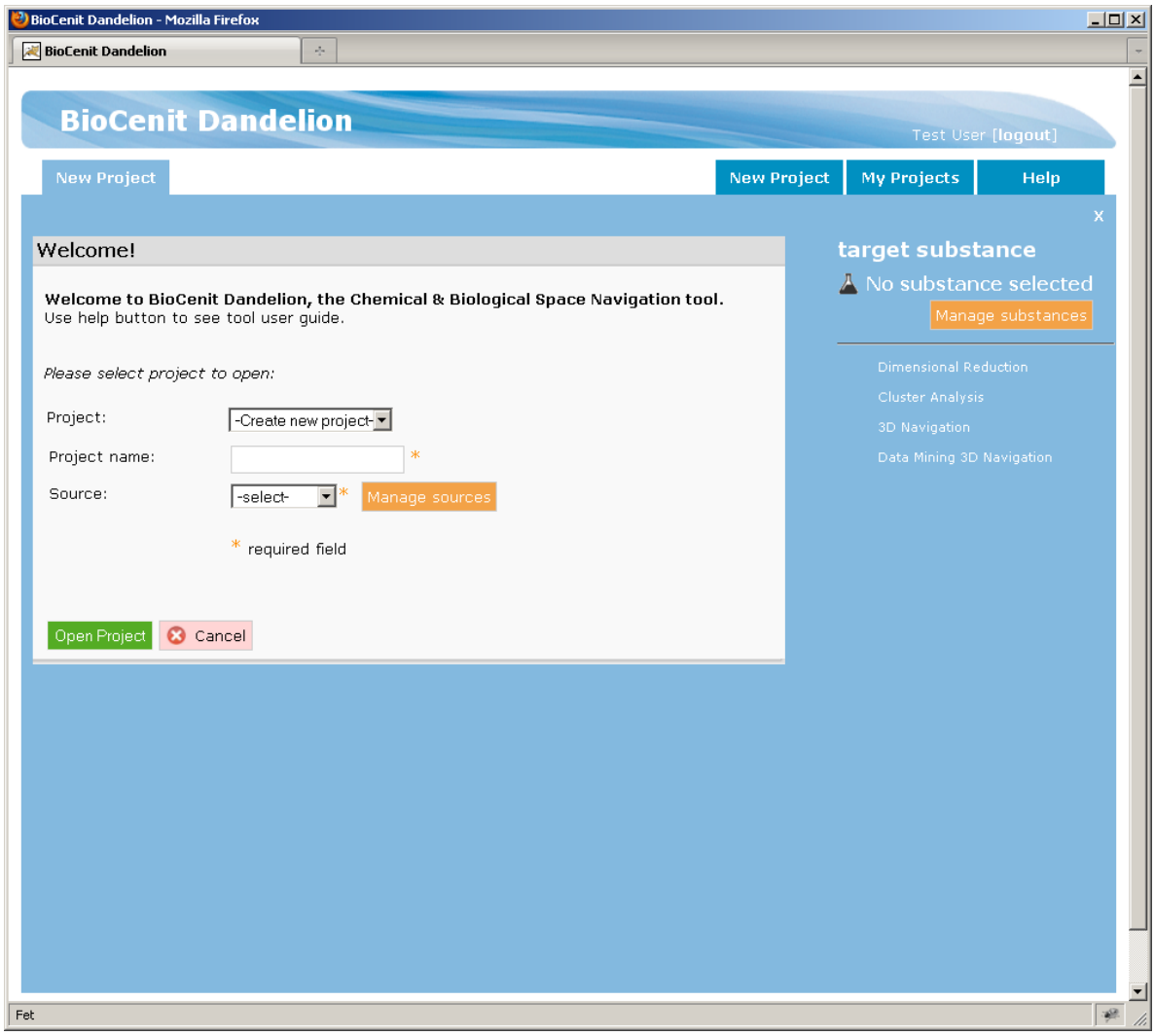

Figure 2. Welcome page of the tool showing the greeting message and asking the project to open

2. ***Source management***

The default dataset used in the tool is composed by the 382 compounds of CEFIC database. User can also add his own sources or edit the existing ones clicking the "Manage sources" button in the "New Project" window, as showed in figure 3.

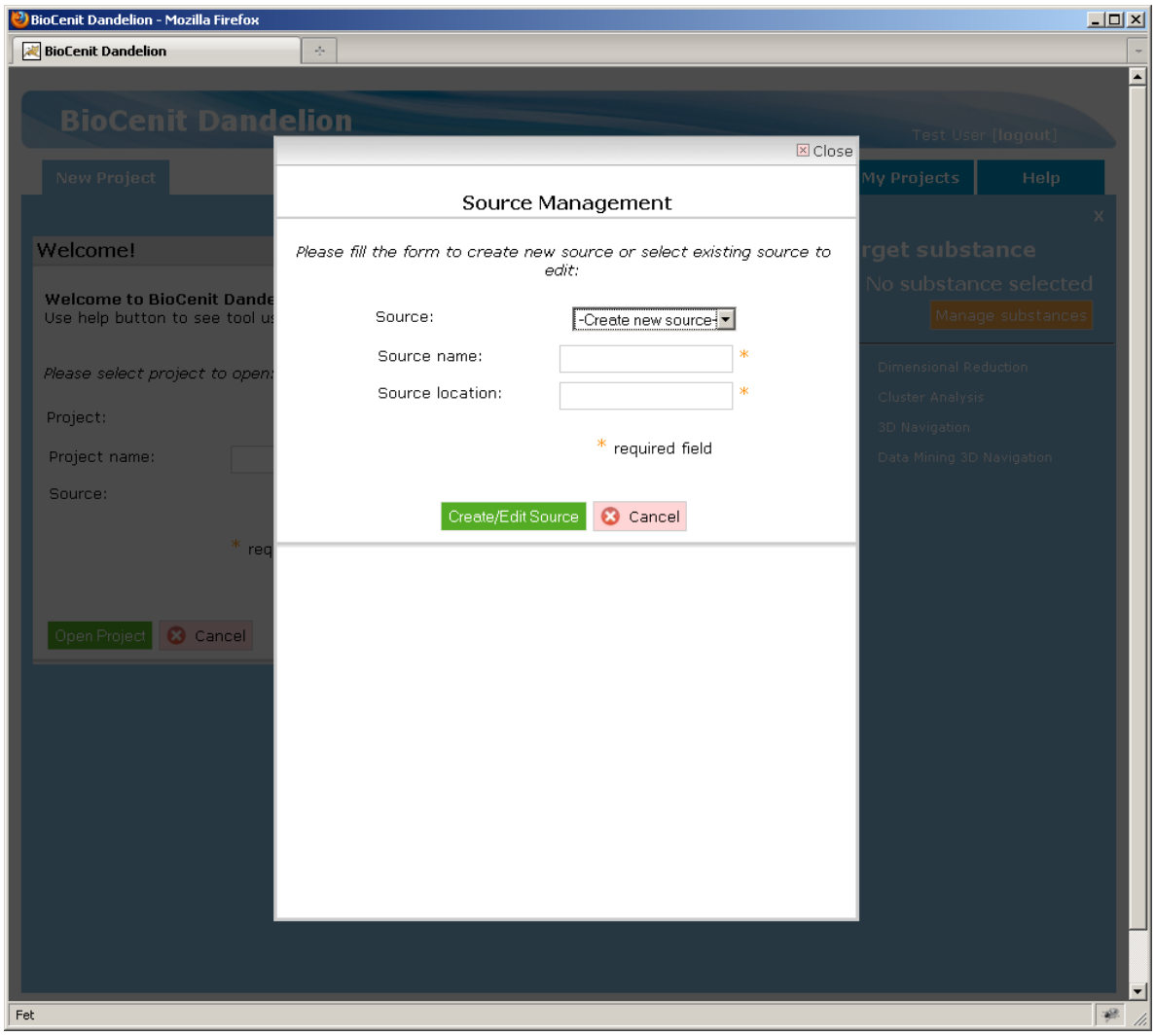

Figure 3. Source management popup

Once user has provided a source name or has selected an existing source, he will be able to upload his own files with data related to compounds, descriptors and endpoints (figure 4). For every type of data user will specify:

- Location: file path in user's computer. User can also download the file header with the corresponding icon ( ).
- Type: incremental (for modifications) or total (remove old data and put the new one).

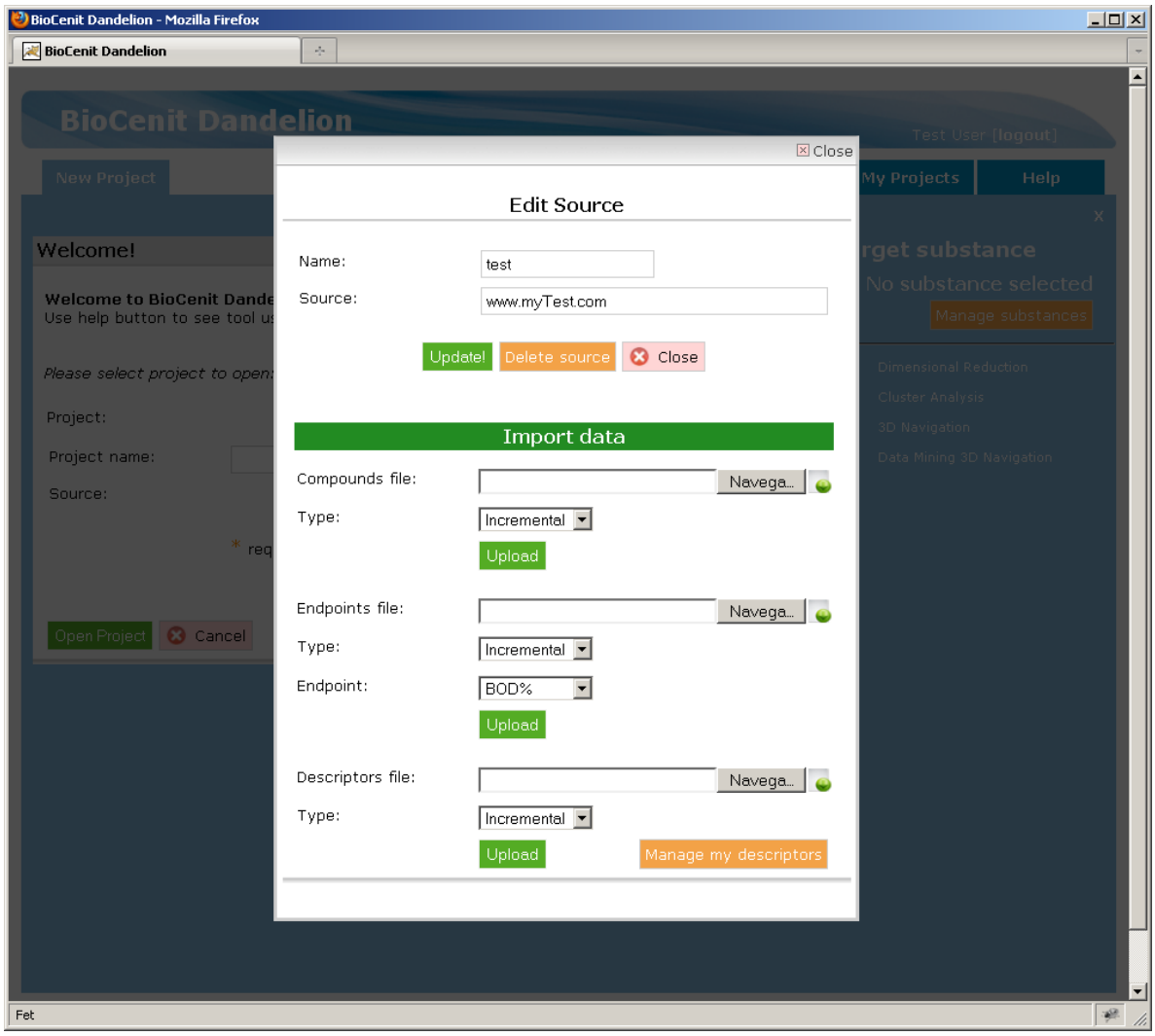

Figure 4. Source edit window allowing data upload

User can also manage his descriptors clicking the corresponding button in the source management window. This action will show a new window for user descriptors management (figure 5). Once descriptor is created, the application will provide a descriptor id which may be used to add/update descriptor values (figure 6).

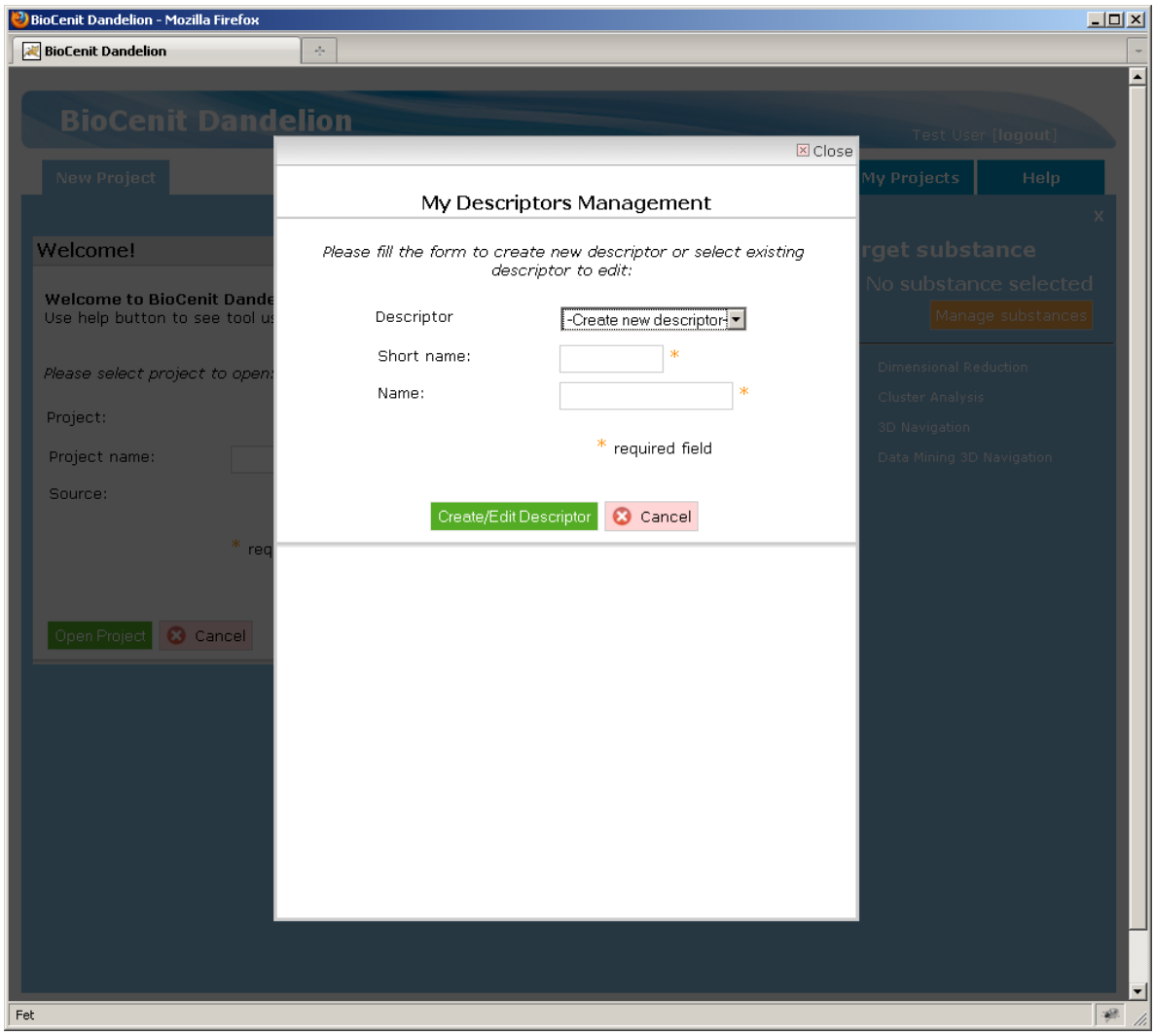

Figure 5. Descriptors management popup

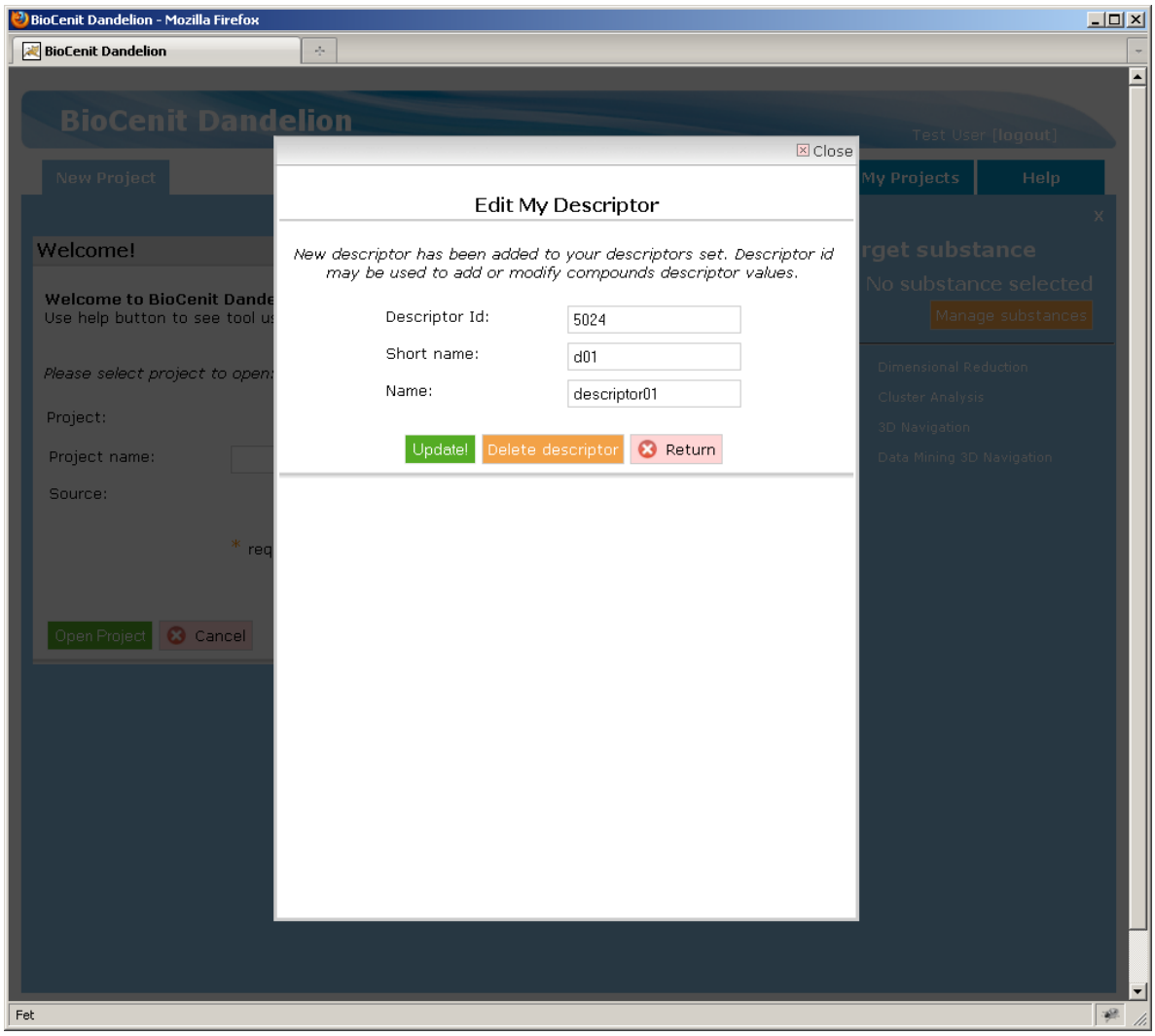

Figure 6. Descriptor edit window providing descriptor id

3. ***Dimensional reduction***

Figure 7 shows the dimensional reduction option which presents a web form that permits the selection of the set of molecular descriptors that will be used in the dimension reduction calculations. Descriptors are grouped by blocks following the classification scheme used in *Dragon 6*[1]. There is also a special block for user descriptors called "My Descriptors". Descriptors can be selected from different blocks and included into the active descriptors list for its processing. A button for descriptor removal from the active list is also provided.

[1] http://michem.disat.unimib.it/chm/Help/edragon/index.html

The next step is the selection of the dimension reduction algorithm (in the current version Principal Component Analysis and Random Projection algorithms are implemented). Finally, a name for this specific dimension reduction process is required since the application stores the results in an internal database for its posterior use in the generation of 3D scenes. The names for each projection coordinate are also required. If no dimensional reduction algorithm is selected, descriptor names are directly given to coordinates, assigning X, Y and Z to 1st, 2nd and 3rd descriptors of the list respectively.

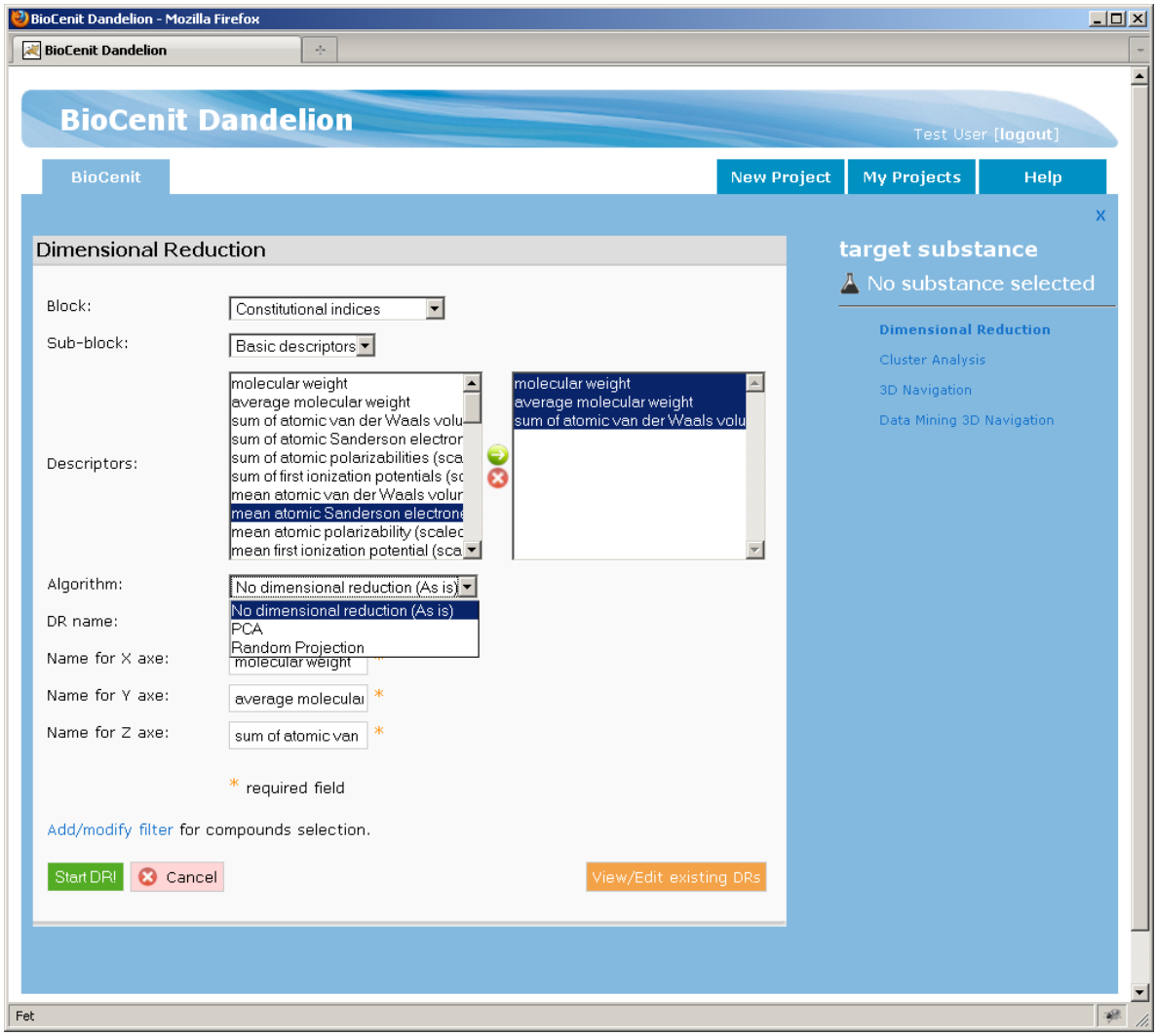

Figure 7. Dimensional Reduction form showing the list of molecular descriptor, the selected descriptors list and the reduction algorithm

Before submitting the form, the user has the possibility of adding or modifying rules for compounds selection by clicking the link "Add/modify filter". Compounds not

complying selected rules in filter will be added to a complementary chemical space (figure 8). Once the form is submitted, a new popup window appears showing the algorithm parameters and asking for confirmation (figure 9). These parameters are specific for each dimension reduction technique.

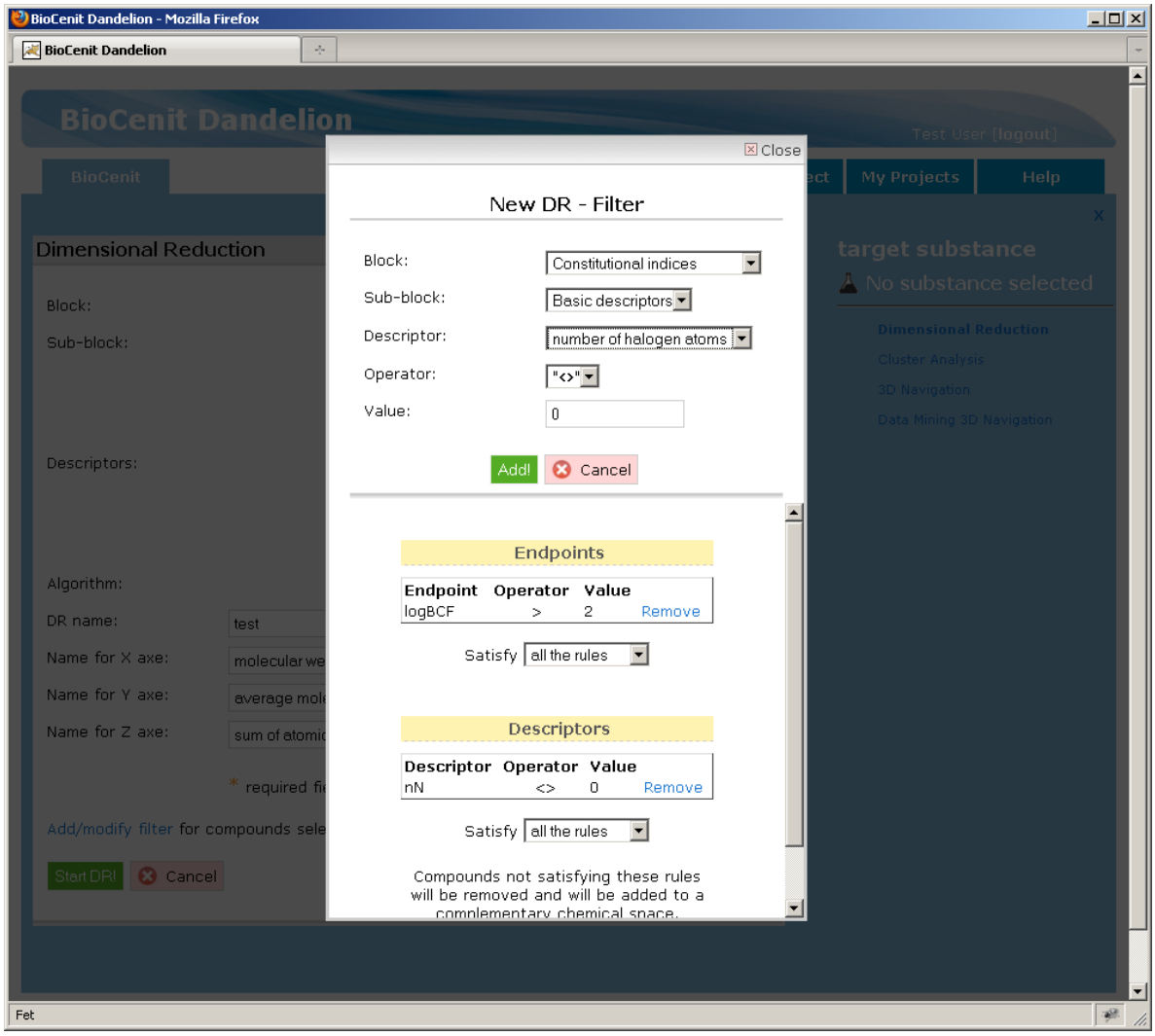

Figure 8. Popup window with filter

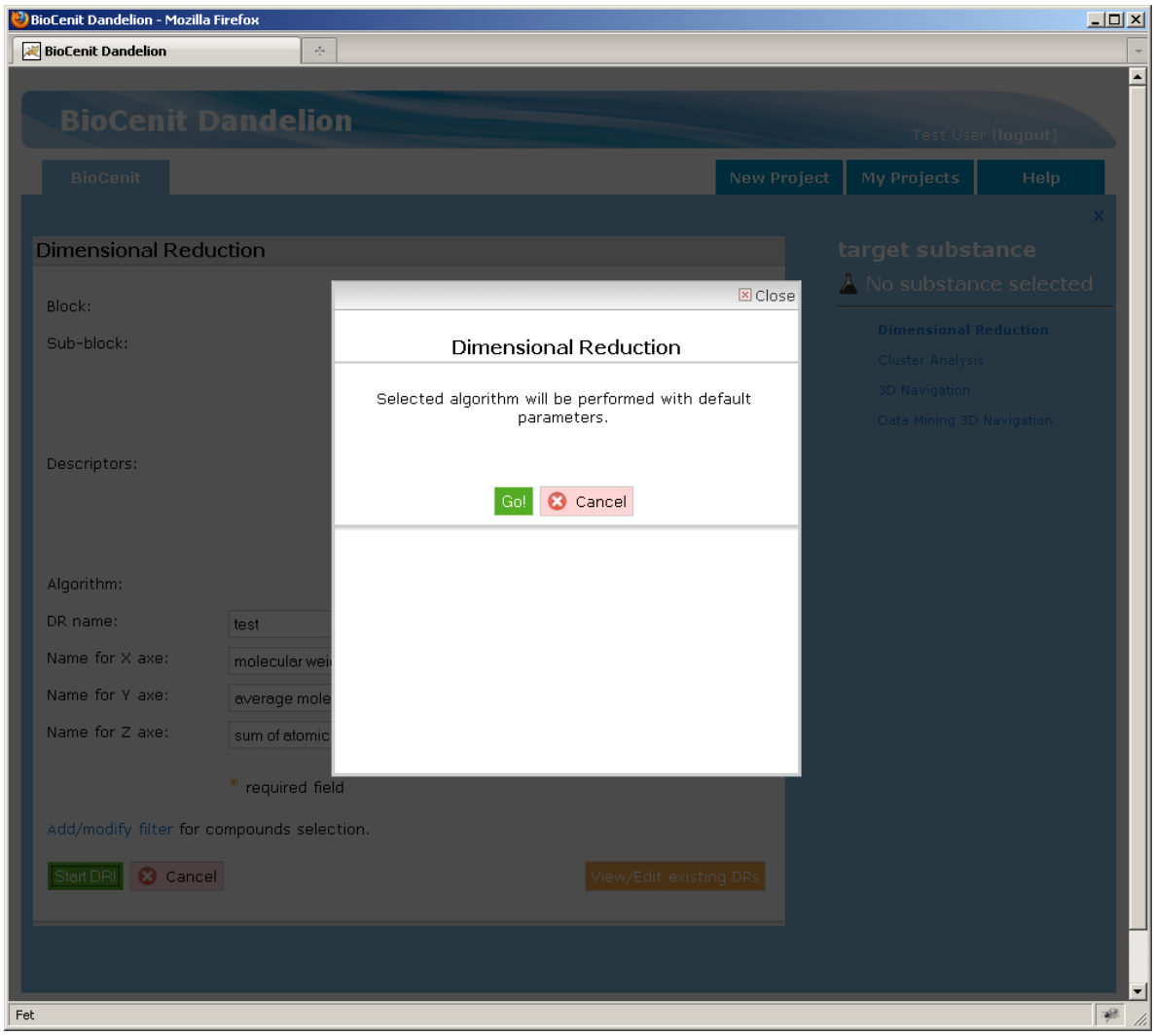

Figure 9. Popup window with algorithm parameters (if exist)

Finally a new popup window appears showing information about compounds selection for the Dimensional Reduction (figure 10). From this windows user can go to the 3D navigation module or close the windows and continue in the dimensional reduction module.

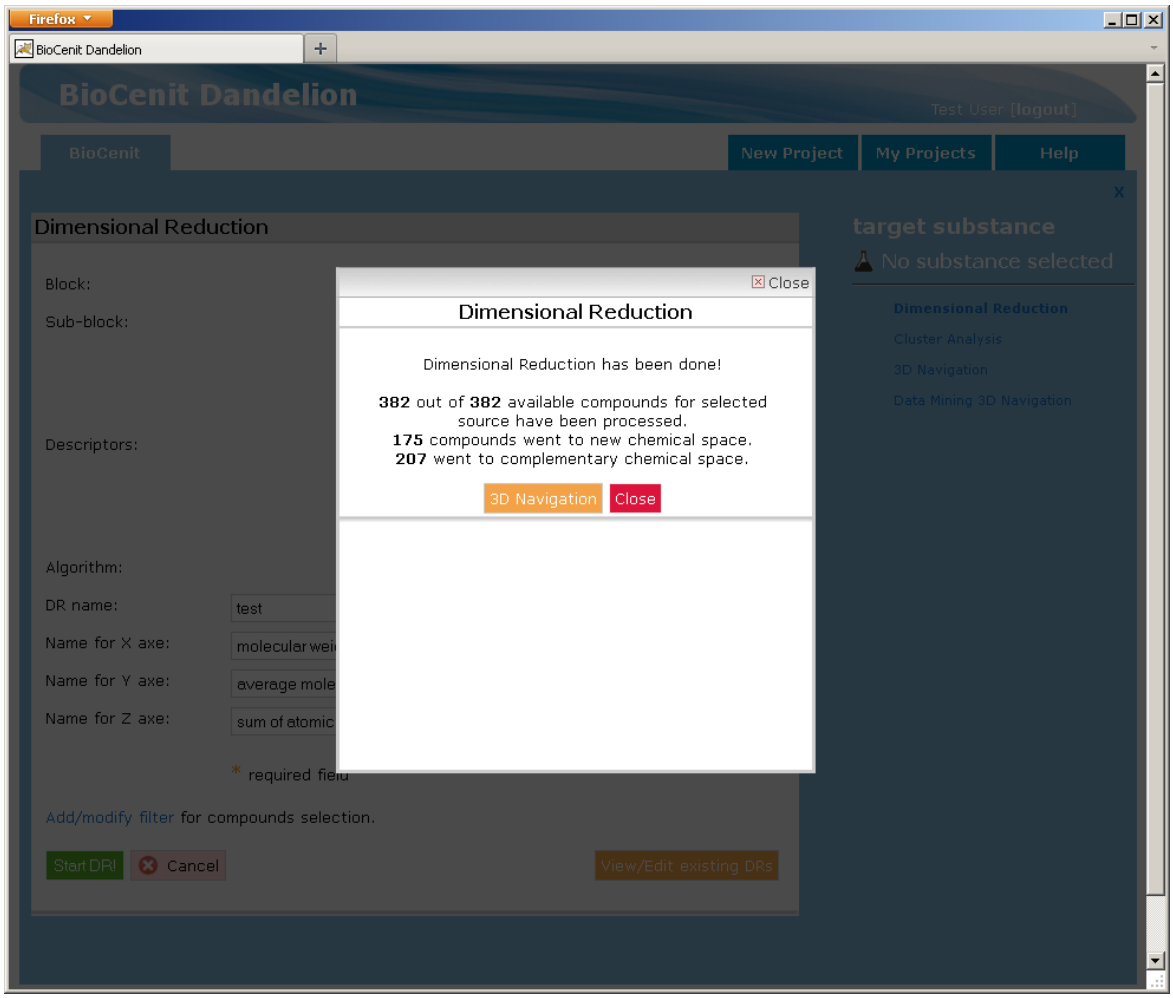

Figure 10. Popup window with compounds selection information

4. ***Cluster analysis***

Another optional previous step to chemical space visualization is the cluster analysis. This module, showed in figure 11, presents a web form that permits the selection of the set of endpoints and molecular descriptors that will be used in the cluster analysis calculations. Descriptors appear in the same way as they did in dimensional reduction module (grouped by blocks following the classification scheme used in *Dragon 6* plus special block with user descriptors). Endpoints and descriptors can be selected and included into the active endpoints and descriptors lists for its processing. A button for endpoint and descriptor removal from the active list is also provided.

The next step is the selection of the cluster analysis algorithm (in the current version K-Means, Hierarchical Clustering and Expectation Maximization algorithms are

implemented). Finally, a name for this specific cluster analysis process is required since the application stores the results in an internal database for its posterior use in the generation of 3D scenes. The number of desired clusters is also required. When the user submits the form, a new popup window appears showing the algorithm parameters and asking for confirmation (figure 12). These parameters are specific for each cluster analysis technique.

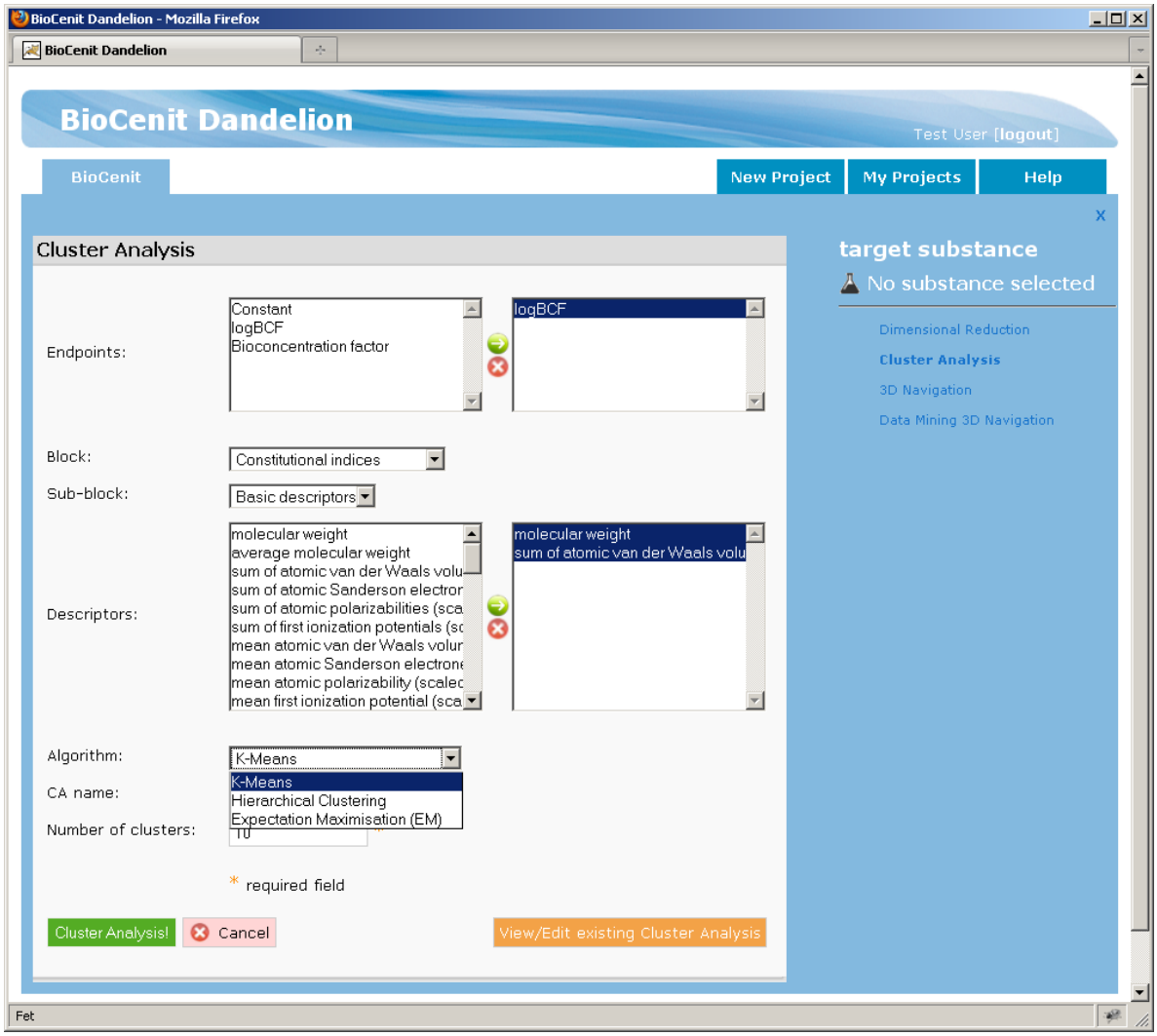

Figure 11. Cluster Analysis form showing endpoints, molecular descriptors, selected endpoints, selected descriptors and the clustering algorithm

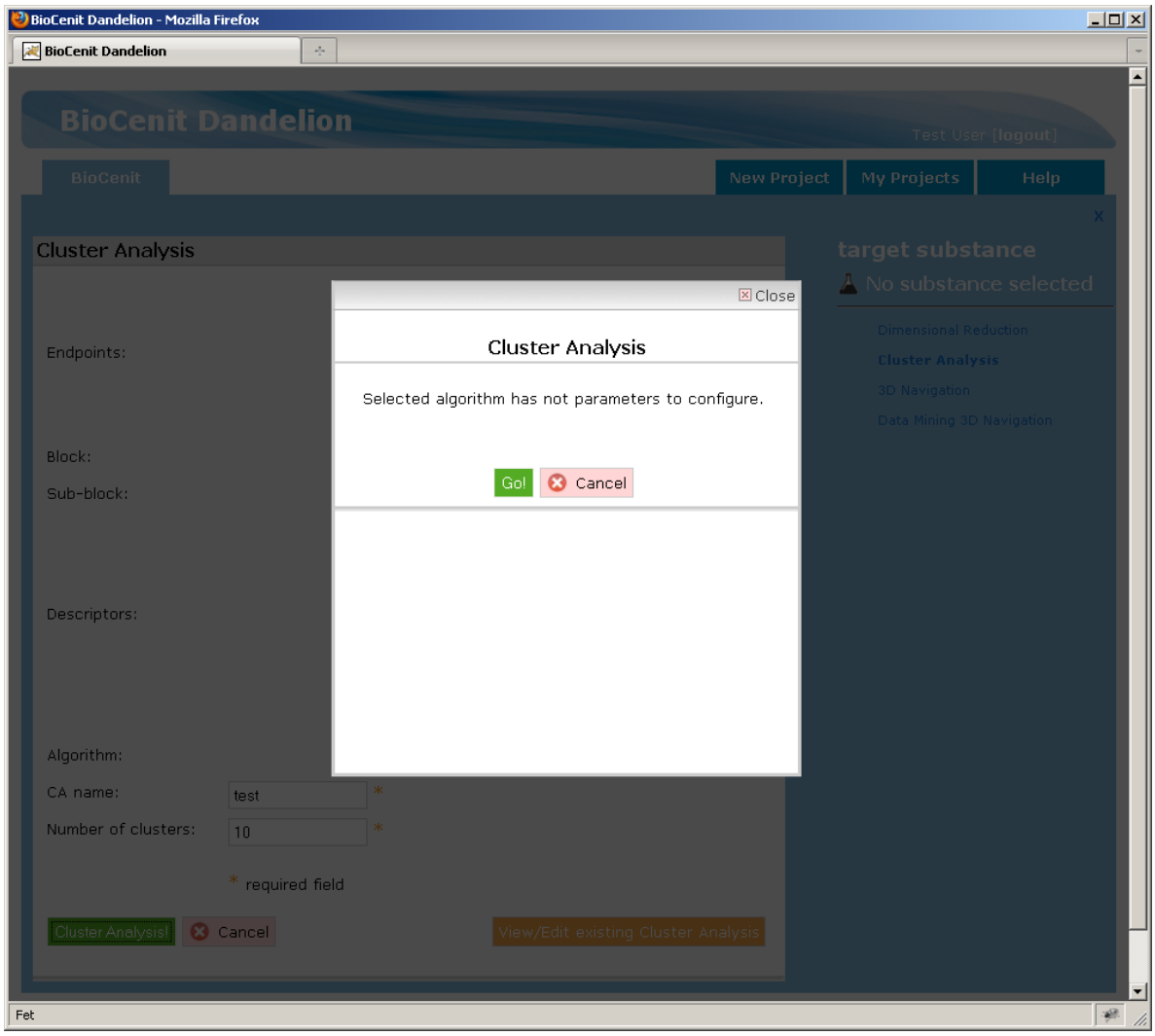

Figure 12. Popup window with algorithm parameters (if exist)

5. ***Chemical space navigation***

Once the dimensional reduction process or cluster analysis complete the user is redirected to the 3D navigation options. This form can also be accessed directly from the main menu. Figure 13 shows the input information template corresponding to the configuration of the chemical space visualization. The parameters governing the visualization process include the stored dimensional reduction to use, the desired endpoints or cluster analysis to be mapped to colour and radius, the normalization mode for each attribute, and the number of clusters for visualization (0 indicates no clustering). In the case where a cluster analysis is selected, a new select box will appear to choose the desired cluster analysis. There is also a special option for colour:

the proximity clustering. In this case another text box will appear to indicate the number of proximity clusters.

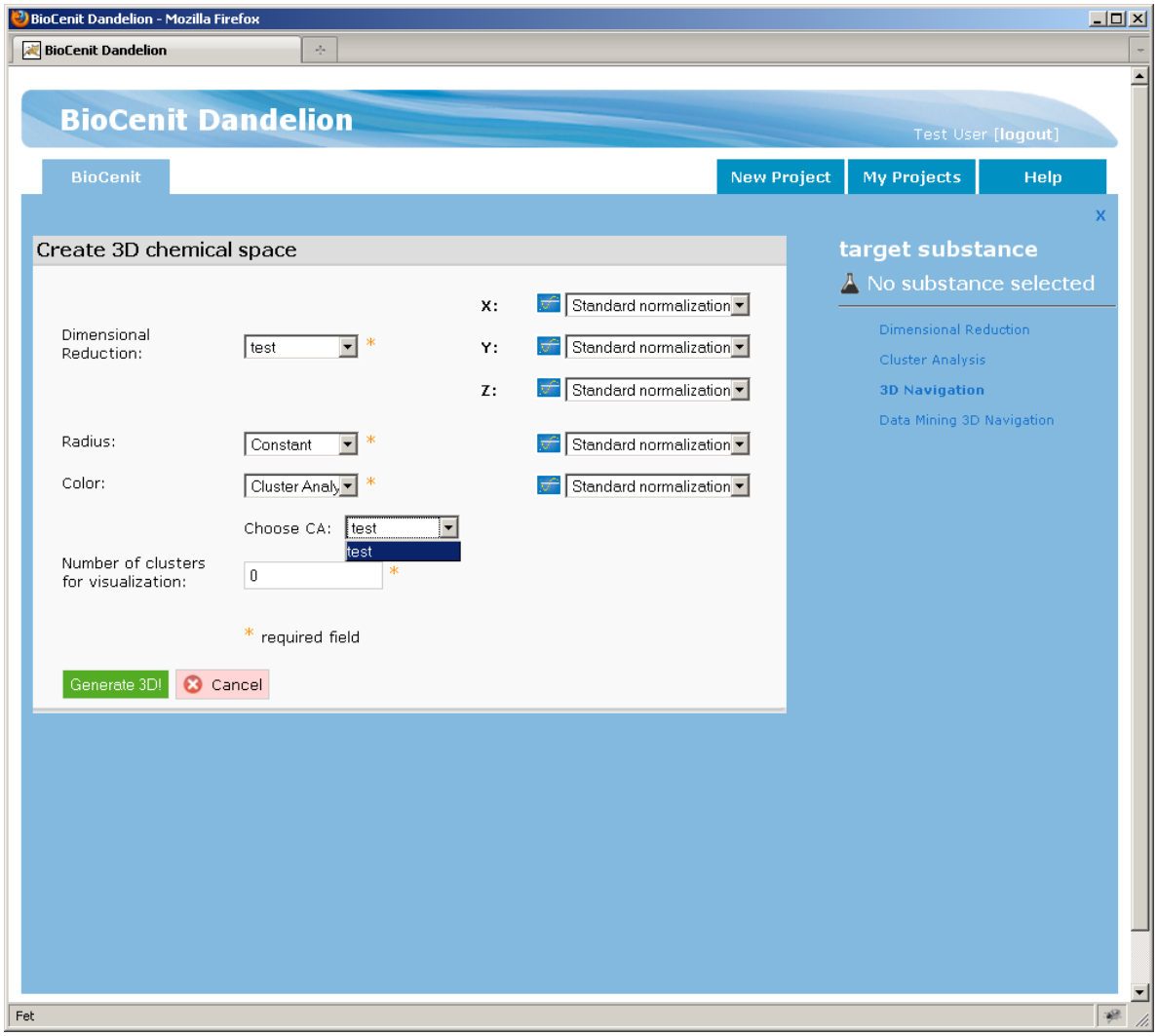

Figure 13. Parameters to generate the 3D representation of the chemical space

When user clicks in the *Generate 3D* button the application searches in its database for stored clustering partitions (if number of clusters is greater than 0) and a new window is opened that allows the use of a pre-calculated clustering or to build a new one. Figure 14 shows the chemical space representation embedded in the web browser through the *X3D* plug-in.

The modes available to navigate through the chemical space will depend on the installed browser plug-in. Generally a user can use both, keyboard and mouse to

move inside the scene. Right-clicking in the navigation window presents a menu of settings including:

- *Viewpoints*: user can selected between four pre-defined viewpoints: the global view, the molecule structure, view focused in the substance, view from the substance studied.
- *Movement*: *fly* is the default selected option for movement inside the scene. The user can select different modes such as *walk*, *slide*, *examine*, *panoramic*, *game-like* and *jump*.
- *Speed*: allows the control of the navigation speed.

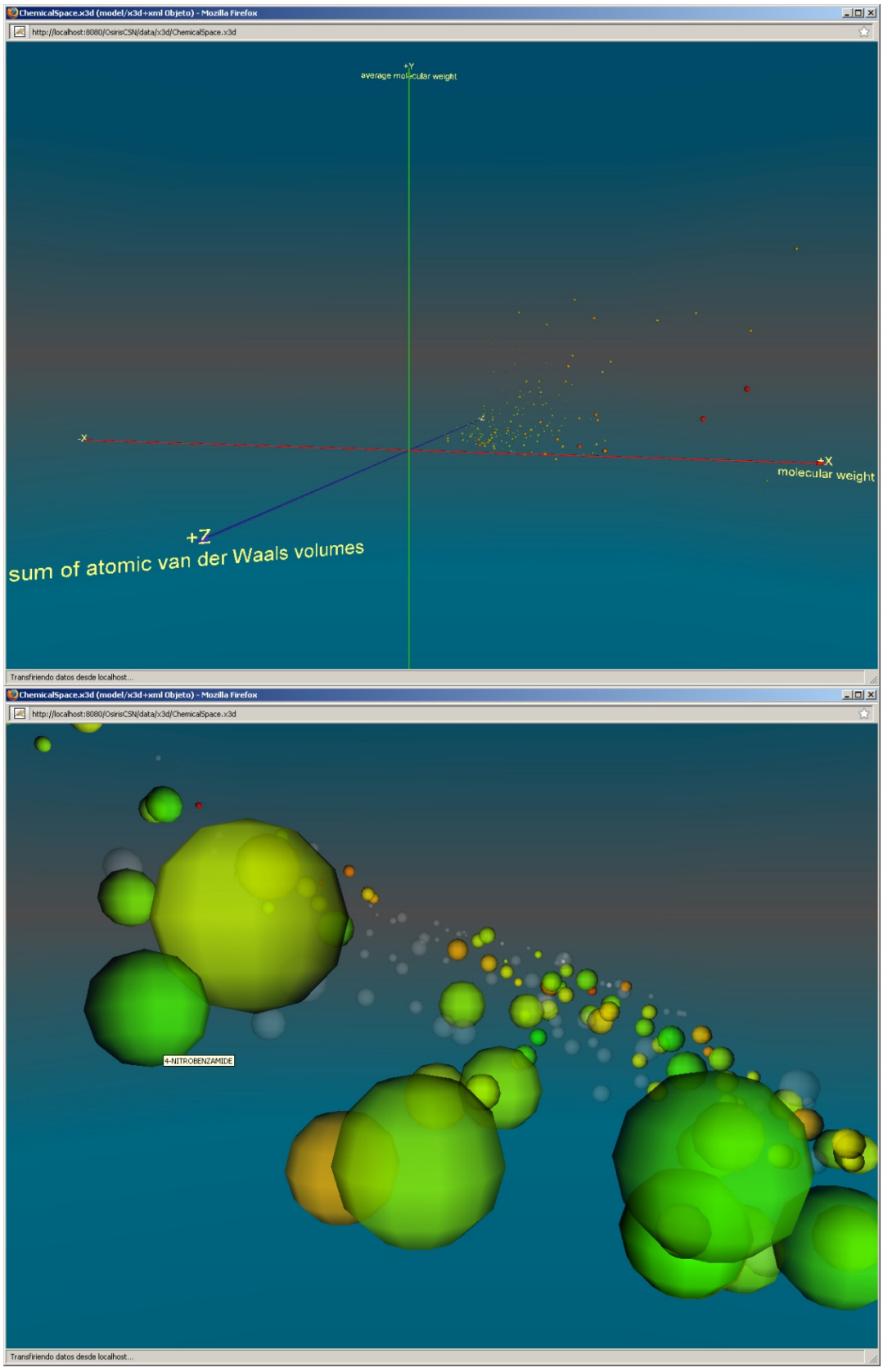

Figure 14. Initial 3D Chemical Space visualization showing the global view and a zoomed perspective

When the mouse pointer is over a compound a popup window shows its name and *clicking* in the selected chemical gives access to access a new 3D scene focused in the selected compound and its closer topological neighbours.

6. ***Near compounds navigation***

In the *near compounds mode*, the user is able to visualize/examine the chemical structures and the projection properties in order to detect chemical similarity patterns.

As shown in Figure 15, the window has three frames. The upper frame shows the selected compound, the lower frame shows the scene with the compounds surrounding the selected chemical, and finally the right frame shows the results of a query to *PubChem* database for the target compound. When different compounds belong to the same position in the space, user can click collision button to watch collision compounds in the upper frame (Figure 16).

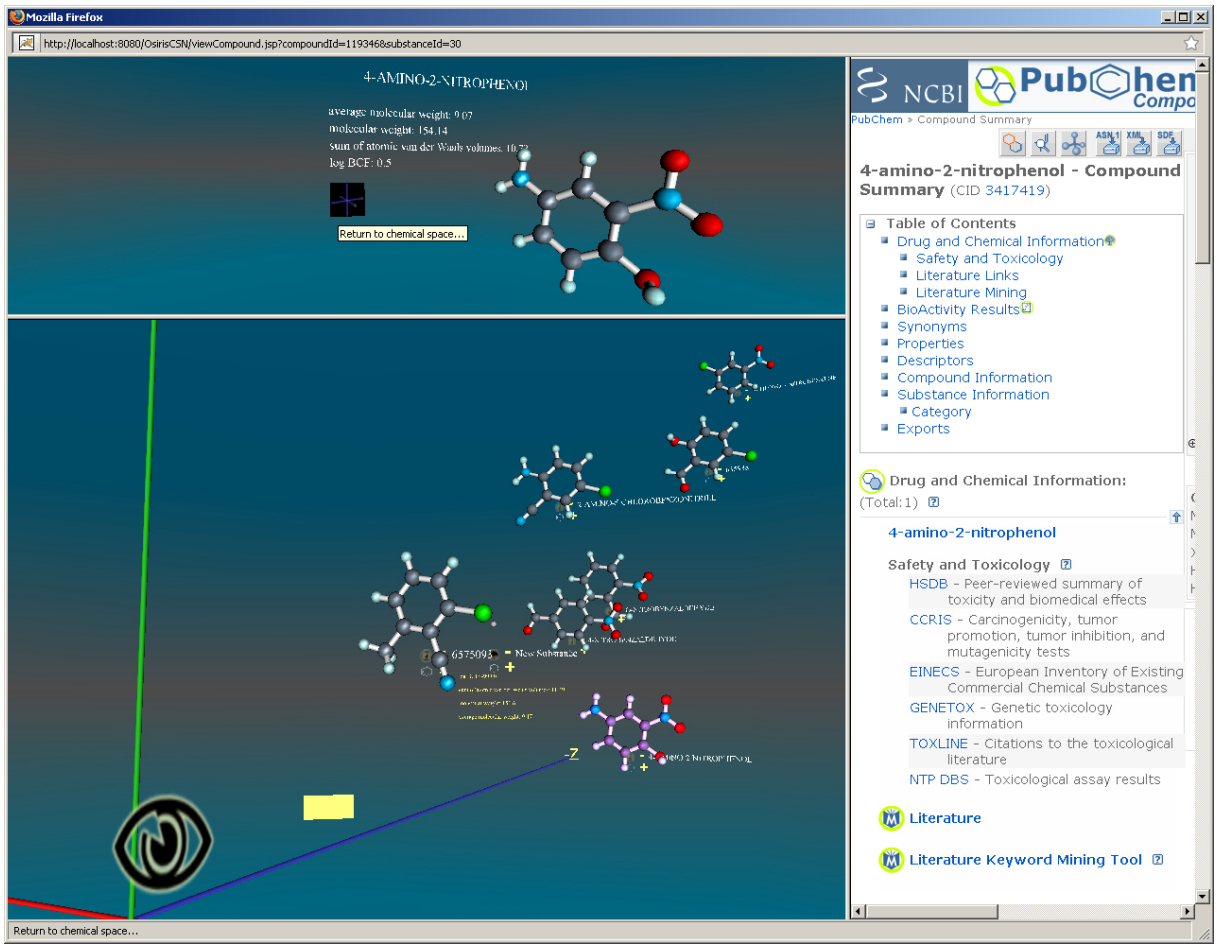

Figure 15. Structural similarity screening in the neighbourhood of the selected chemical

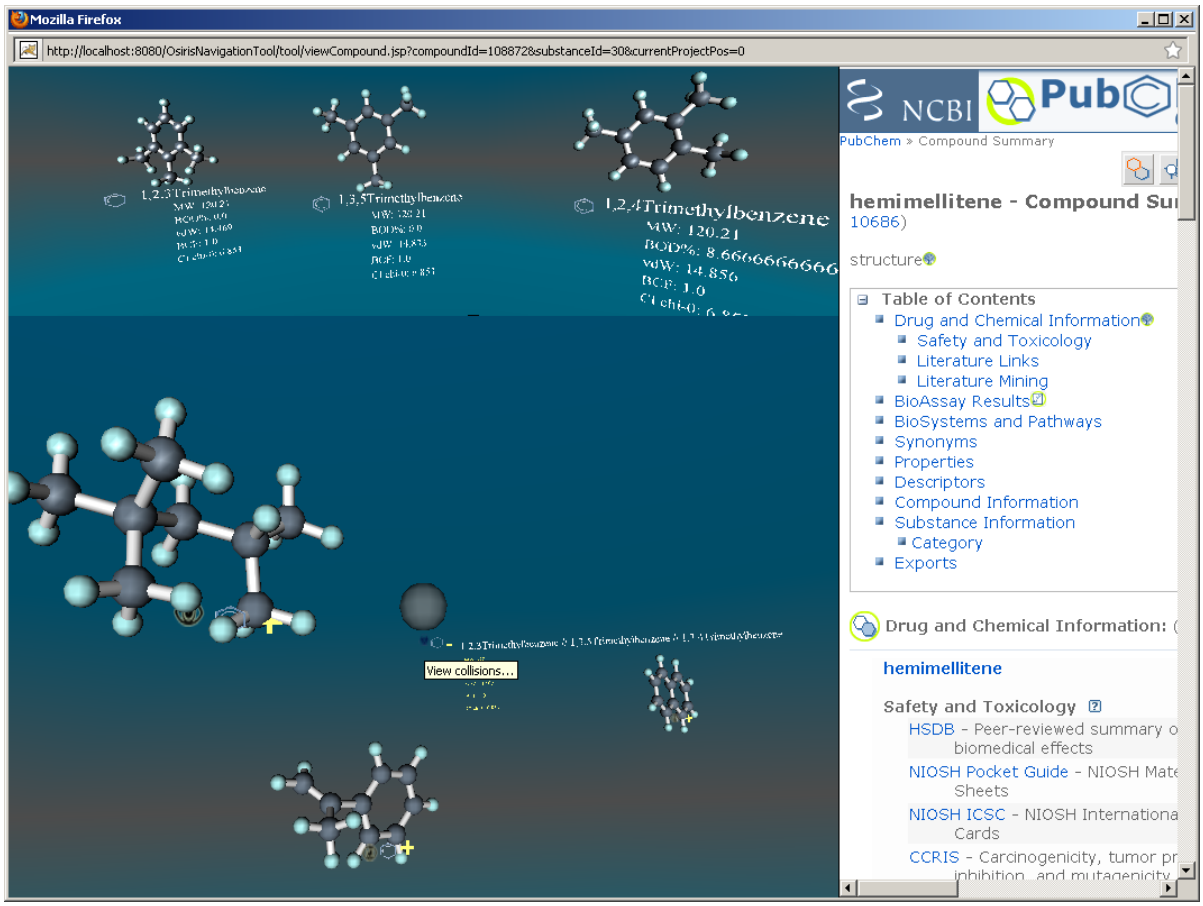

Figure 16. Collision treatment in the neighbourhood of the selected chemical

The 'Viewpoints' menu permits the selection of multiple viewpoints linked to the chemicals in the current scene. Specific molecule viewpoint can be directly accessed by clicking on the molecule.

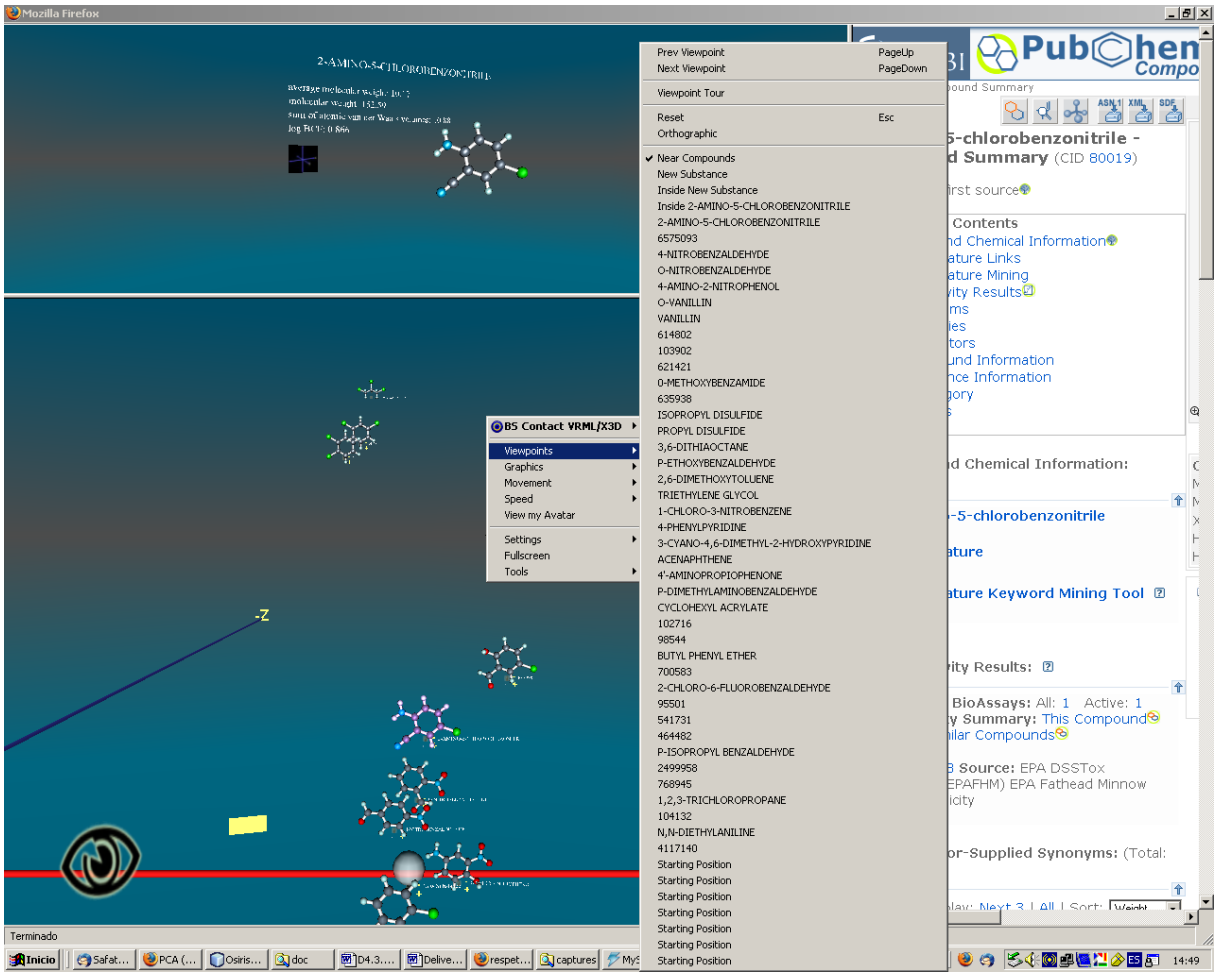

Figure 17. Selection of the pre-defined set of viewpoints in the neighbourhood of the chemical studied

By default all molecules are rendered showing its 3D structure. The complexity of the scene can be reduced by clicking the *'eye'* icon that converts the 3D molecule structure back to a sphere increasing the visualization performance. There is also the 'x' icon that hides all molecules icons if user wants to see only molecules structures. The specific attributes of each molecule can we shown or hidden by using the '+' and **'-'** icons. Each molecule also has a '*PubChem*' icon, which will refresh the right frame showing specific *PubChem* data for the selected chemical.

7. ***Data mining with the chemical space navigation***

The tool offers the possibility to build 'visualization trees' in order to facilitate multiple chemical spaces visualization corresponding to a particular compound collection. Figure 18 shows the input information template corresponding to data mining and chemical space visualization. In this window user can choose between build a new tree or choose an existing one. The parameters governing the visualization process are the same as in the chemical space navigation window.

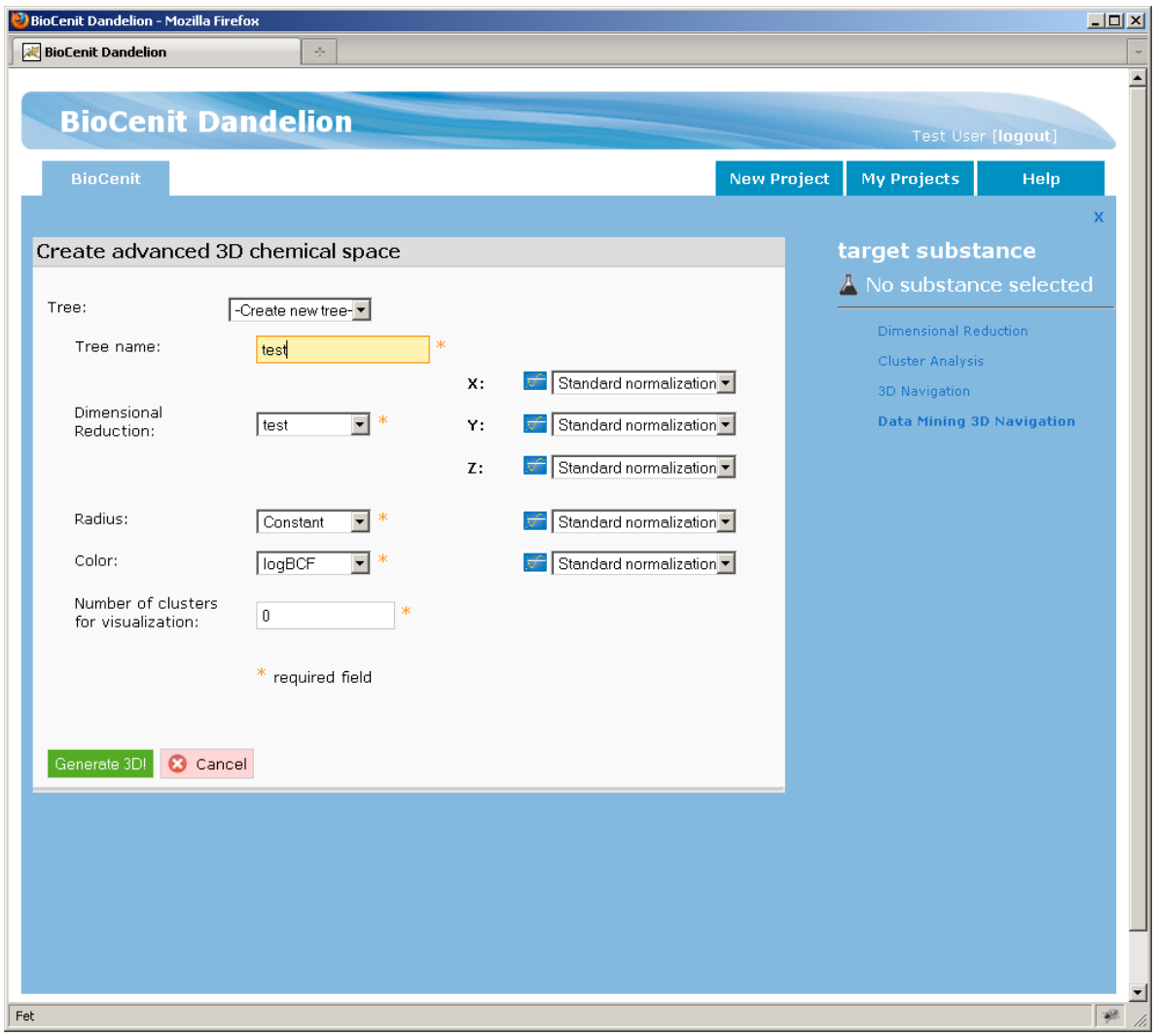

Figure 18. Parameters to generate the data mining 3D representation

As in standard 3D navigation window, when user clicks in the *Generate 3D* button the application searches in its database for selected tree or builds a new one depending on user selection. Figure 19 shows the window which contains the visualization of

different tree nodes. The upper-left chemical space corresponds to selected node by user, while complementary node is placed upper-right and parent node is in the bottom. The right frame offers the possibility to change node selection ( ) and edit ( ) or delete ( ) tree nodes.

User can also create new tree nodes providing node name and filter rules in the "New tree node" section. By clicking Split! button, new node and its complementary will be created and right frame will be refreshed.

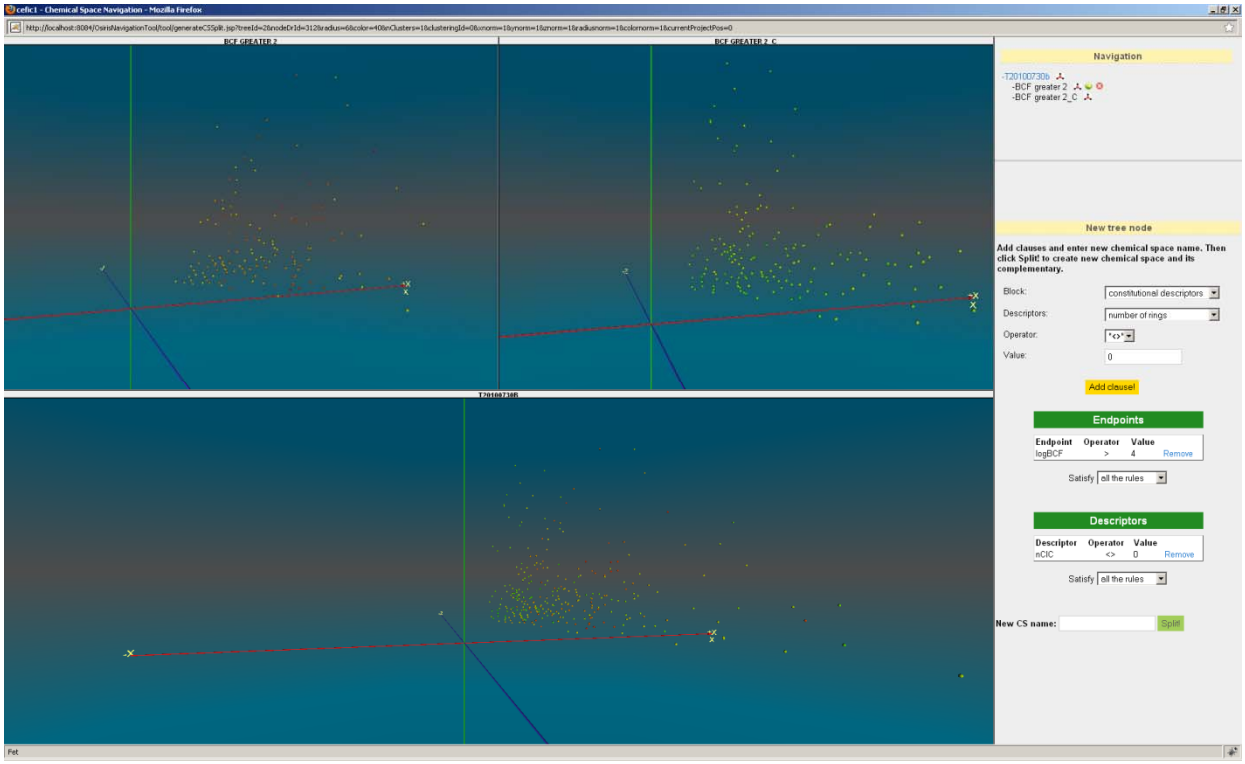

Figure 19. Chemical Space visualization showing the different tree nodes