Chemical Databases under *InstantJChem*

G. Marcou and A. Varnek

Data organisation

InstantJChem organizes data into a strict hierarchy:

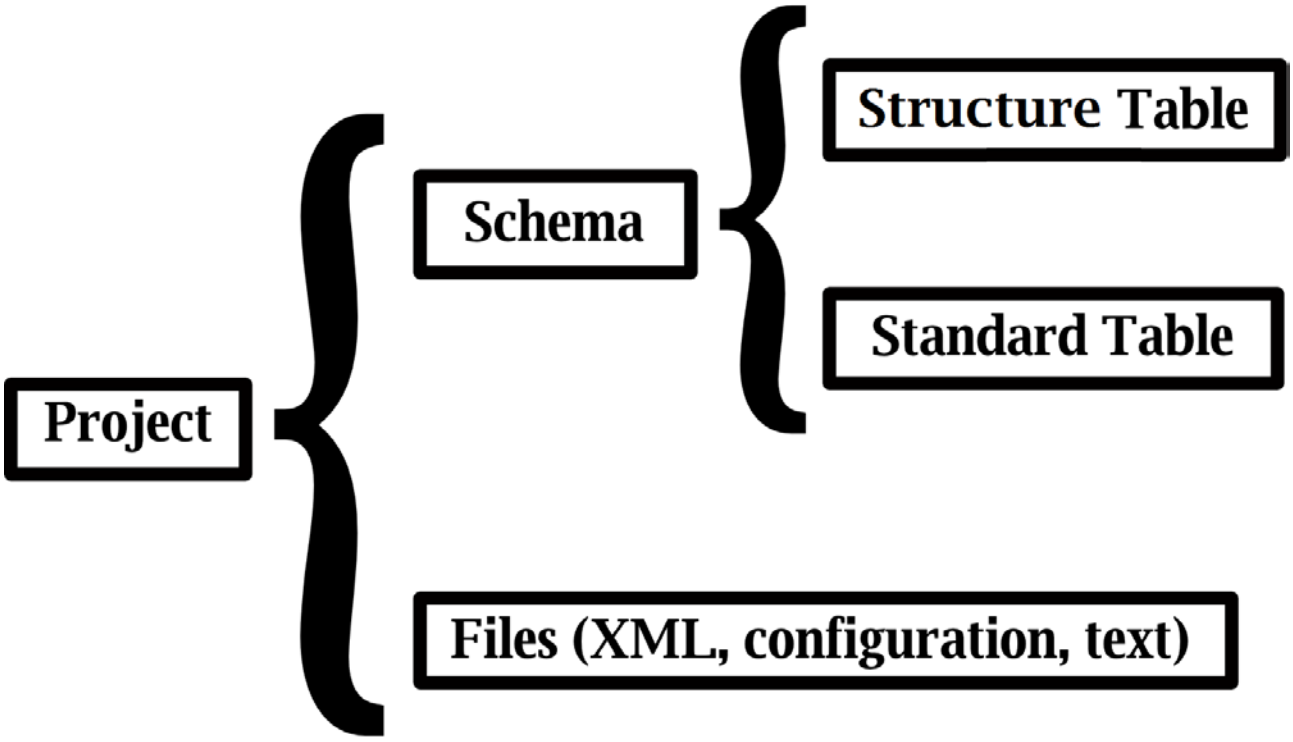


Figure 1: Different levels of data organisation in InstantJChem

Data are organized in **Projects**, which can contain both generic files and **Schema(s)**. All files placed to the project folder will appear in the graphical user interface as belonging to this project. These could be (*i*) configuration files for specific applications, (*ii*) molecular files used for the data exchange between different applications, and (*iii*) text files for comments. **Schema** could be considered as a layer between the graphical interface and a database management system (Derby, MySQL or Oracle). It represents an ensemble of related tables, each of which can be manipulated and displayed in *InstantJChem*. Two types of tables are used: *Standard* tables containing only standard data types and *Structure* tables containing molecular information as structures, calculated properties, etc. *Structure* tables are organised around a molecular graph representing a compound.

How to create a new empty database with InstantJchem

1. Click on **File**, **New Project**.
2. In the popup menu, select **IJC Project (empty)**, then click the **Next** button
3. In the form, fill the **Project Name**, **Project Location** fields, the **Project Folder** is generated according to the information in preceding fields.
4. Click on **Finish**.

The database is empty. Now it is needed to create at least one **Schema**.

How to create a local database with InstantJChem

1. Click on **File**, **New Project**.
2. In the popup menu, select **IJC Project (with local database)**, then click the **Next** button
3. In the next form, fill the **Project Name**, **Project Location** fields, the **Project Folder** is generated accroding to the information in preceding fields.
4. Click on **Finish**.

The database now contains a new schema.

How to add a Schema to a Project

1. Into the **Projects** window, right click onto the name of your project
2. In the popup menu, select **New Schema**.
3. In the following menu, select **Embedded Derby**, then click the **Next** button -Note: if you want to connect to an existing Oracle or MySQL database, it is needed to select them here.
4. In the following form, give a name in to the **New Schema name** text area, and then click **Finish**.

How to import an SDF file into Schema

1. Into the **Projects** window, develop the content of you project
2. Into the content of your project, right click onto a **Schema**.
3. In the popup menu select **Import File**.
4. In the next pop up menu, fill the **File to import** field.
5. Edit the other fields. Pay a particular attention to the **Table details** since it is the interface for many details of the management of chemical structures (mainly absolute stereo, duplicates filtering, empty structures, fingerprints, standardization). Fields of the table are automatically detected.
6. Click **Next**.
7. Edit the list of fields of the table to be created and loaded from the SDF file, then click **Next**.
8. The SDF is loaded. Examine the monitoring of the import and click **Finish**.

How to delete a project or an element of the project

1. Right click on any element appearing into the Project window.
2. Select Delete.
3. In the popup window, it is recommended to require deletion of every related files and schema, otherwise *InstantJChem* or the database management system will continue to record part of it that might come across the way in subsequent operations.

How to view a database in a grid

1. Right click on a table.
2. Click on **New View**.
3. In the popup window, select **Default Grid View**.
4. Click on **Finish**.

How to browse a database

1. Right click on a table.
2. Click on **New View**.
3. In the popup window, select **Empty Form View**.
4. Click on **Finish**. A **Design** mode is activated.
5. In the **Design** mode, add a molecule panel area and as many text area as required to view database fields.
6. Switch to **Browse** mode.

Many elements to the form can be added. It includes also labels, checkboxes for boolean properties, date pane for date fields, list for list fields, multiline text area and a table.

How to perform substructure search query

1. Right click on the main blank area in the **Query** window.
2. Select a field of type molecule on which to perform a query. It generates an interface for molecular structure search.
3. Double click onto the blank area of this interface to open a sketcher and draw your query.
4. When finished click **Set Query.**
5. Select the **Substructure** key word into the rolling menu of the interface.
6. Click **Run Query**.

How to perform similarity search query

1. Right click on the main blank area in the **Query** window.
2. Select a field of type molecule on which to perform a query. It generates an interface for molecular structure search.
3. Double click onto the blank area of this interface to open a sketcher and draw your query.
4. When finished click **Set Query.**
5. Select the **Similarity** key word into the rolling menu of the interface.
6. Click on the **Options** button to access the similarity threshold.
7. Click **Run Query**.

The similarity search is based on the hashed fingerprint computed by InstantJChem during the loading of the structures. Settings of the fingerprints are a property of the shema. It can be edited into the **Edit Schema** item into the popup menu activated by a right click on a schema name.

Note: the similarity search can't be performed on Markush structures.

How to combine queries

Queries are organised in a tree. Each node is a logical word AND/OR. Each leaf is a query. If there

is only one query, it belongs to a default AND node.

1. Right Click on a logical node, on the main area in the **Query** window, to change it to OR or to add an OR/AND node to the tree.
2. Right click on an existing query element and in the popup menu click on **Add Field**.
3. Select a field on which to perform a query.
4. Customize the specific generated interface.
5. When finished click **Run Query.**

Note: any query element can be edited with right clicking

How to save a query results

1. When the query result is displayed in the main **Grid View**, click on **List**.
2. In the menu, select the **Save as List** option.
3. In the popup menu, select **Permanent** as **New list type** and **All rows** in the **Include in list** area.
4. Click **OK**.

How to combine query results

The **Lists and queries** tab gives access to the history of all queries prepared during a session. Items labelled **Temporary** are destroyed at the end of the session. To keep them for the future work the labels should be changed to **Permanent**.

1. Each query result must be saved in a List.
2. Click on the **Lists and queries** tab near the **Projects** tab.
3. Select some lists.
4. Click on the **Lists** item in the tool bar.
5. Develop the **List Operations** item and select an appropriate operation.
6. Configure the popup menu (check list orders, operation and resulting list saving options) and click **OK**.

How to perform a substructure search using lists of atoms

1. Set up a substructure search.
2. Edit the substructure.
3. In the sketcher, click on the **Periodic System** icon.
4. In the popup menu, click on the **Atom list** button.
5. While this button is activated, activate any combination of element button.
6. When setup, click the **Close** button.
7. Click on all atoms of the structure that should be replaced by the atom list.
8. When finished, click on the **Set Query**.
9. End the setup of the query and click **Run Query**.

How to perform a Chemical Term search

1. Right click on the main blank area in the **Query** window.
2. Select a field of type molecule (field Structure) on which to perform a query. It generates an interface for molecular structure search.
3. Click on the **Chemical Term** button.
4. In the popup window select and edit the required **Chemical Term** (for instance in the **Favourites** list, the **Bioavailability**).
5. Click **Run Query**.

How to create an empty Structure table

1. Double click on the name of an existing Schema to open the connection.
2. Right click on the Schema and select in the menu **New Structure entity table**.
3. Right click on the Schema and select **Edit Schema**.
4. Click on the **Add Standard Field** or the **Add Chemical Terms Field** icons.
5. In the popup form that opens, select the field type.
6. Configure the fields. Key properties are: the name, whether it is required or not, a default value.

How to create an empty Standard table

1. Double click on the name of an existing Schema to open the connection.
2. Right click on the Schema and select in the menu **New standard entity table**.
3. Right click on the Schema and select **Edit Schema**.
4. Click on the **Add Standard Field** or the **Add Chemical Terms Field** icons
5. In the popup form that opens, select the field type
6. Configure the fields. Key properties are: the name, whether it is required or not, a default value.

How to add an item to existing table

1. Select a view of a database table. For instance select the **Grid View**.
2. Click on the **Add row** icon.
3. In the dialog popup menu, fill all required informations: draw structure, give values for numerical and text fields.
4. Click **Add**.
5. If needed edit a new item and click **Add** again when it is setup
6. When finished, click **Close**.

How to create a relationship between tables (Many to One)

1. Open the **Edit Schema** interface and go in the **Data trees** window. Both tables to be linked should be already present at this point.

2. The table being on the “Many” side, should be added a **Required** integer field to contain the table keys of the “One” side table.
3. Go in the **Entities** window and click on the **New Relationship** icon.
4. In the pop menu select, **New Simple Relationship**.
5. In the Basic tab, give **Name** to the new relationship, then select **Many to one** as the type of the relationship.
6. Tick the **Create DB constraints**.
7. Select the **From** table, that should be the “Many” side, and the relevant **Field** created before.
8. Select the **To** table, that should be the “One” side, and the relevant **Field** that should be the table's key.
9. In the **DB Constraints** tab, select the relevant rules in case of changes or deletion of entries in one or the other table. For instance select **Restrict** as the **On Delete** rule. Note: once a rule has been set, it can be hardly changed without deleting and recreating the relationship.
10. Click **Finish.**
11. Click on the main table and then on the **New Edge** icon. The relationship should be automatically found.
12. Click **Finish**.

Note that any relationship between tables in a database must be designed before loading the database. Otherwise, many complications are to be expected.

TUTORIALS with InstantJChem

G. Marcou and A. Varnek

Datasets

- **All files are to found in the IJC directory.**
- **SC100.sdf: A database of 99 diverse compounds from Chemaxon**
- **ISICCRsm.mrv/ISICCRsm.RDF: A database of 239 reactions in both RDF and MRV formats.**

Exercise 1.

Create a new Project named ***IJCExercises*** and import the file SC100.sdf in it. Customize a browser for it. A new database table should be created named ***SC100***.

Exercise 2.

In the ***SC100*** database, perform a search on fluorobenzene and pyridine molecules using *Substructure* or *Similarity* options. Compare results of these two types of search.

Exercise 3.

Combine the compound 89 and 25 into one query editor and create a bond between them. Search the database with the resulting query with the option *superstructure*. Comment the result compared to the previous exercise.

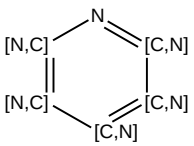
Exercise 4.

Use the entry 46 as a query. Edit the query and remove the Bromide. Perform a search setting the query as *Full* first, then as *Full fragment*.

Exercise 5. *Combined Searching*
Perform a search on benzene as substructure and 'pyrimidin' containing *Product name* field and *Aq Sol* being Good.

Exercise 6.

Perform a substructural search on cyclic aromatic fragments of 6 atoms containing at least one nitrogen atom.



Exercise 7.

Perform a search on molecules for which MolWeight > 200 and which don't contain the benzene ring.

Exercise 8.

Same question, but perform the search on two steps (*i*) search for compounds for which MolWeight > 200 then (*ii*) search for compounds containing benzene ring. Cross the two result lists.

Exercise 9.

Use the Chemical Term field to search compounds possessing more than 4 microspecies at pH=4.0. Export your hit list in an SDF file called **HitList.sdf**.

Exercise 10.

Import into your project, the file ISICCRsm.RDF. Customize a browser for it. A new database table should be created named ***ISICCRsm***.

Exercise 11.

In the ***ISICCRsm*** table, perform a search of imydazole as a substructure of the reactant then as a substructure of the product.

Exercise 12.

Edit the Schema of your project and add a new empty Structure table called ***AlkanBoilingPoint***. The new table must contain a field names BoilingPoint receiving floating point values.

Exercise 13.

Add to the ***AlkanBoilingPoint*** table the following compounds along with their boiling point temperature.

Alkan	**Boiling Point (in K)**
Pentane	231.05
Butane	272.65
Cyclopentane	322.4
1,1-dimethyl cyclopropane	293.75

Tableau 1: Boiling point for sample alkans

Exercise 14.

Add a new field named Date, for the date of the entry and fill it.

Exercise 15.

Add a Chemical Term field to add an automatically estimated value of the logP to each entry.

InstantJChem

Full structure search, substructure search (extracted from ChemAxon Manual)

Chemists are most often interested in *substructure search*, that is, whether one molecular structure contains the other one as a substructure or not. By definition, the examined molecule is called a target, the structure we are looking for is called a query, and a target molecule matching the query structure is called a *hit* (Table 1).

If special molecular features are present on the query (eg. stereochemistry, charge, etc.), only those targets match which also contain the feature. However, if a feature is missing from the query, it is not checked by default.

A *full structure search* finds molecules that are equal (in size) to the query structure. (No additional fragments or heavy atoms are allowed.) Molecular features (by default) are evaluated the same way as described above for substructure search.

Other search types

Besides the above, InstantJChem supports *similarity*, *duplicate*, *superstructure* and *full fragment* type searches.

Similarity is only used in database searches, and its similarity concept is based on hashed binary chemical fingerprints with Tanimoto metrics. (For a more detailed description, see the Developers Guide.) For a more sophisticated approach of similarity, we provide the Screen package.

Duplicate search is mainly used before database inserts to check whether the given molecule is already contained in the database or not. All molecular features need to be equal here, eg. non-stereo query will only match non-stereo target, etc.

Superstructure search is the opposite of substructure search: It searches for those target molecules which can be found in the given superstructure query. (In this case the roles of the query and target molecules are simply exchanged, so query properties should be specified on the target!)

Full fragment search is between substructure and full search: the query must fully match to a fragment of the target. Other fragments may be present in the target, they are ignored. This search type is useful to perform a "Full search" that ignores salts or solvents beside the main structure in the target.

Table 1. Full structure search, substructure search

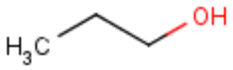
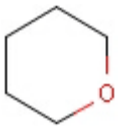
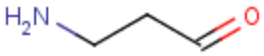
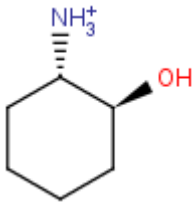
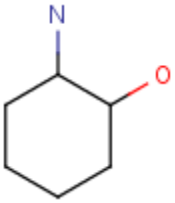
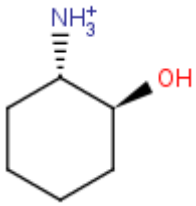
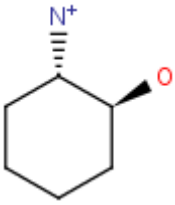
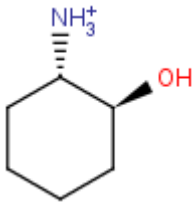
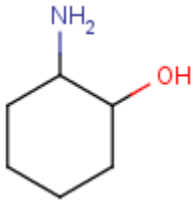
query	target	hit	
		full structure search	*substructure search*
		✓	✓
		✗	✓
		✗	✗
		✗	✓
		✓	✓
		✓	✓
		✗	✗

Table 2. details the main differences amongst these search types.

Table 2. Search type differences

Search type	Search feature

	Similarity	Tests if target contains query	Tests if query contains target	Full fragment coverage	Exact topology matching	Exact stereo matching	Exact atom features matching	Exact bond matching
SUBSTRUCTURE	n/a							
SUPERSTRUCTURE	n/a							
FULL_FRAGMENT	n/a							
FULL	n/a							
DUPLICATE	n/a							
SIMILARITY		n/a	n/a	n/a	n/a	n/a	n/a	n/a

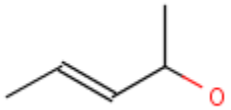
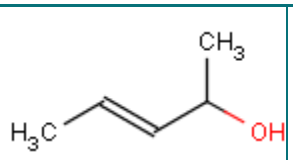
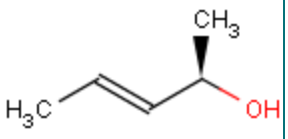
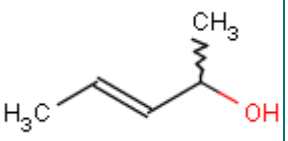
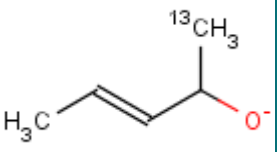
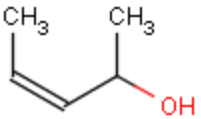
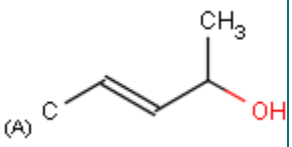
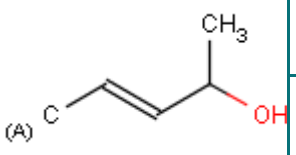
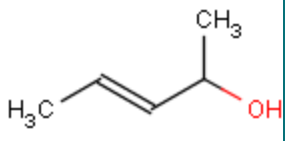
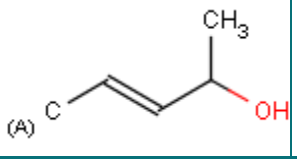
The definition of the search features are:

- Similarity: similarity search using chemical hashed binary fingerprint and Tanimoto metrics.
- Full fragment coverage: the query must cover a whole fragment of the target, but the target may contain other fragments. (Implicit and explicit hydrogens are treated equal.)
- Exact topology matching: the whole molecular graph must match (Implicit and explicit hydrogens are treated equal.)
- Exact stereo matching: equality is needed in stereochemistry, eg. non-stereo query only matches non-stereo target.
- Exact atom features matching: whether matching of certain atom properties should be switched to "exact". This requires equality of the properties (eg. uncharged query only matches uncharged target), and includes the following switches:
    - chargeMatching : exact
    - isotopeMatching : exact
    - exactQueryAtomMatching : true
    - radicalMatching : exact
    - valenceMatching : exact
- Exact bond matching: generic bonds are not evaluated, equality is needed.

Table 3. illustrates the most important differences between FULL and DUPLICATE searches.

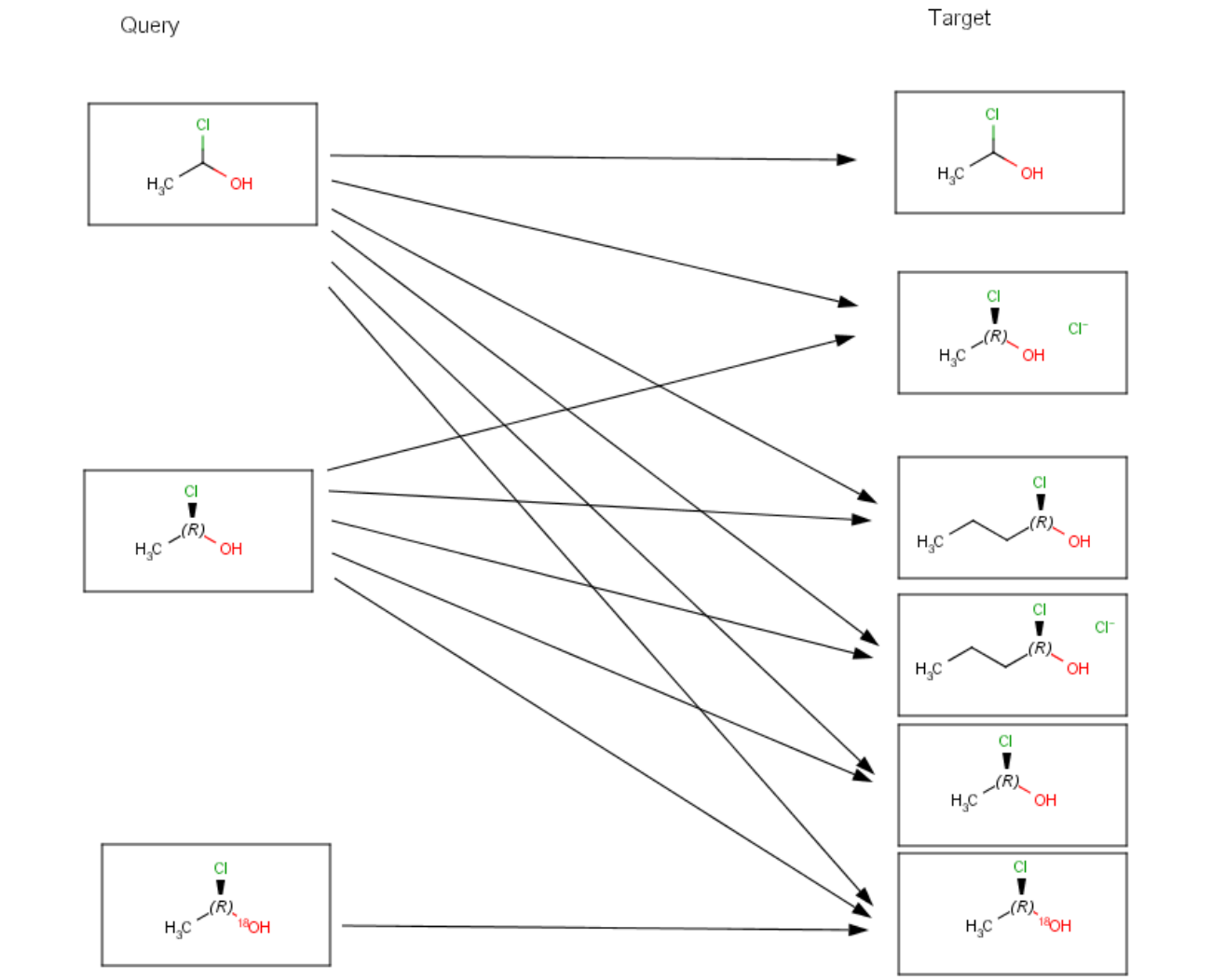
Table 3. FULL and DUPLICATE search differences

Query	**Target**	**Hit**		**Remark**
		FULL	*DUPLICATE*	

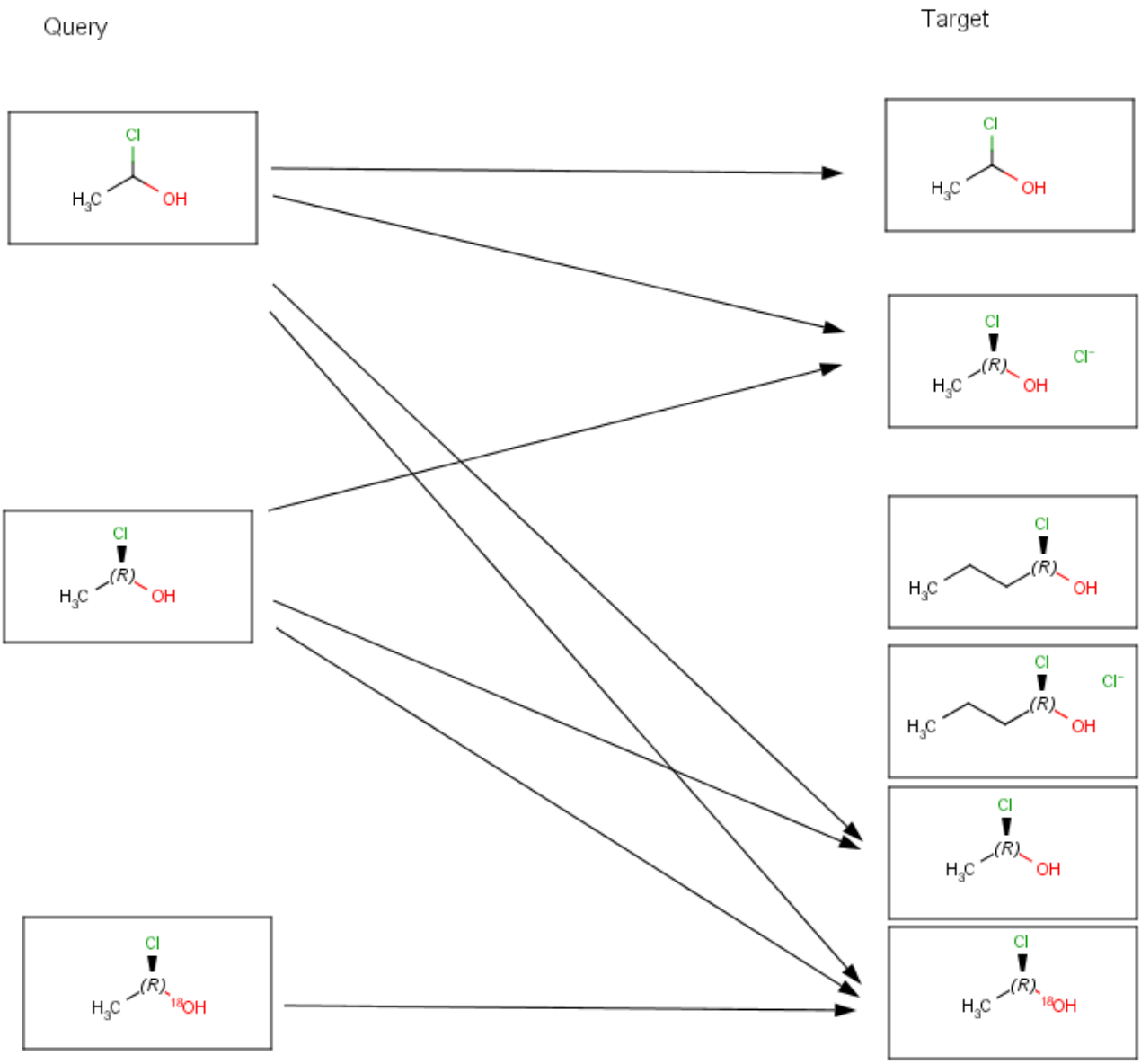
		✓	✓	
		✓	✗	
		✓	✗	
		✓	✗	
		✓	✗	with option DoubleBondStereoMatching set to DBS_MARKED (default)
		✓	✗	(A) denotes aliphatic query property
		✓	✗	
		✓	✓	

The diagrams below show further examples of substructure, full fragment, full and duplicate searches. The arrow between a query and target molecules denotes matching.

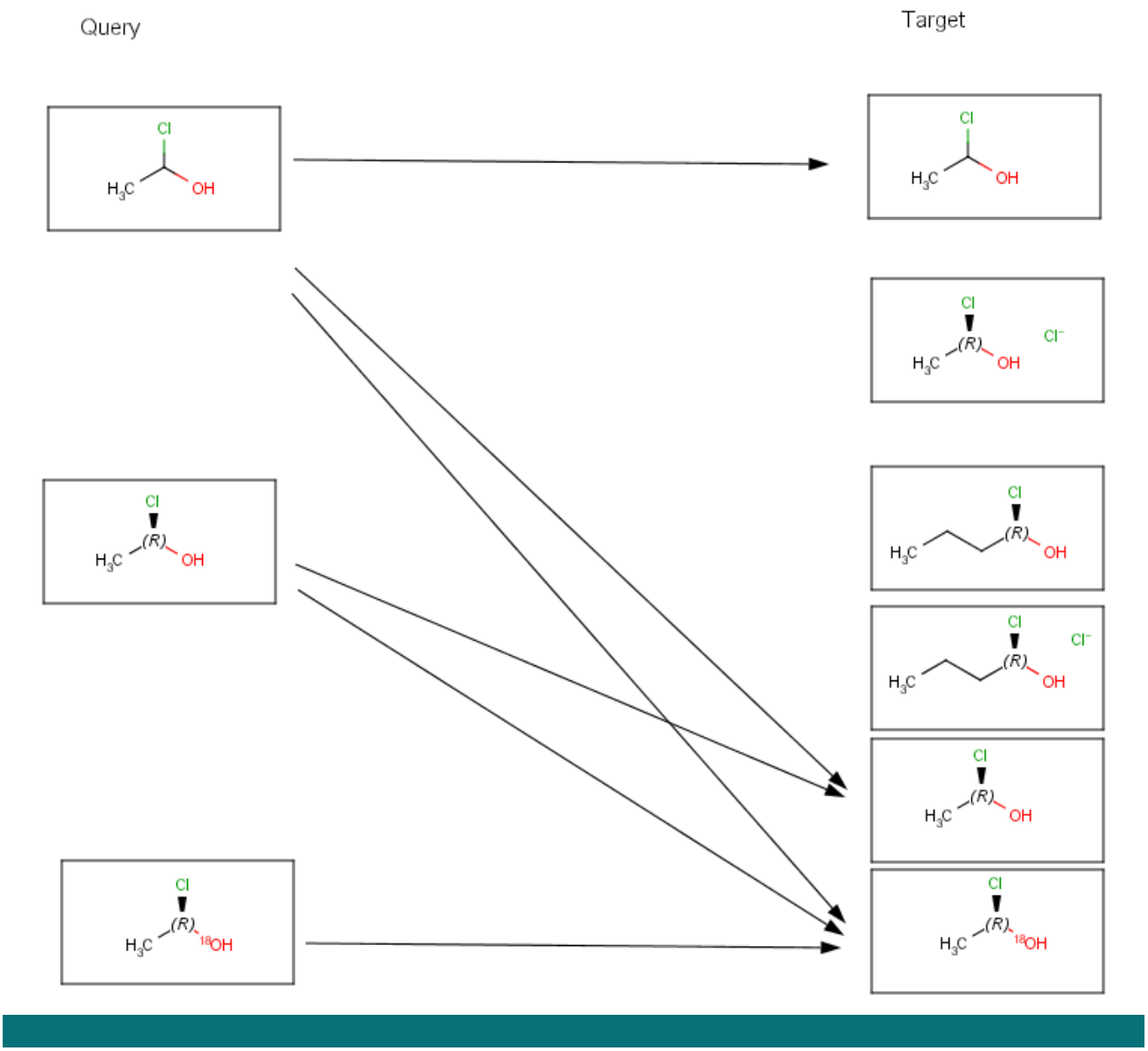
SUBSTRUCTURE search



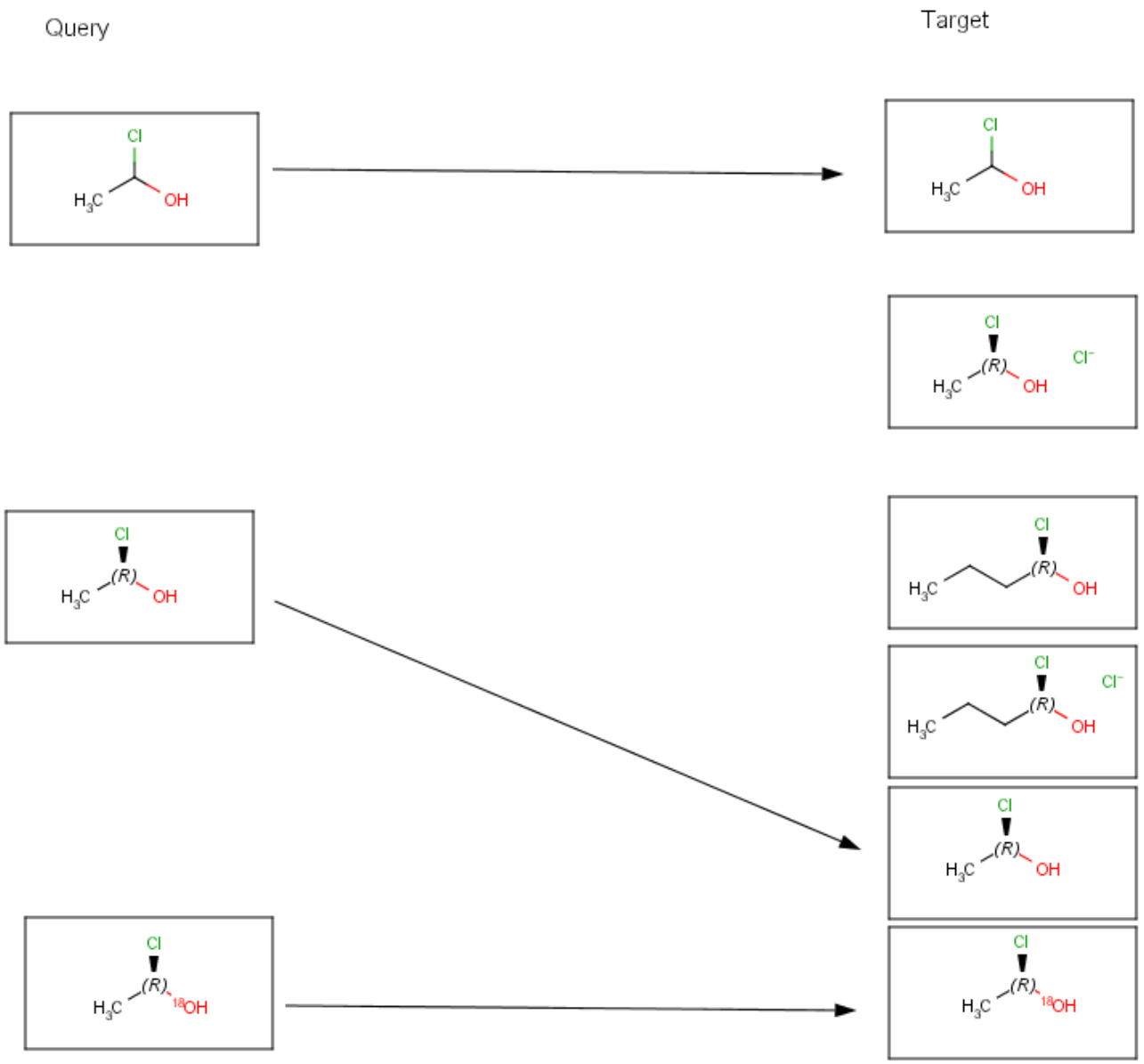
Full fragment search



Full structure search



Duplicate search



Searching in the database

Searching in the database contains a rapid prefiltering step, which screens out many of the targets not matching the query. This step uses chemical hashed fingerprints. To learn more about this step and how to fine-tune fingerprint generation to your needs, see the following document: Parameters for Generating Chemical Hashed Fingerprints

Comparison levels

Graph topology

Graphs consist of *nodes* and *edges*. When we compare structures represented as graphs, the graph patterns must match. Atoms correspond to nodes and bonds are edges.

Atom types

In the case of molecular structures, it is certainly not enough to simply compare the graph patterns, the type of atoms and bonds must be checked as well.

Stereo configuration

Even if both the topology and the type of the corresponding atoms and bonds are matching, we still have to examine the stereochemical configuration. Two molecules having the same kind of atoms connected by the same kind of bonds can be stereochemically different. The relative position of ligands connected to a chiral atom (*R/S* isomers), the enhanced stereo labels on chiral atoms and relative position of atoms located on rings or double bonds (*cis/trans* or *E/Z* isomers) determine the stereochemical configuration of the molecule.