Searching Substances in Reaxys

Learning Objectives

- Understand that substances in Reaxys have different sources (e.g., Reaxys, PubChem) and can be found in Document, Reaction and Substance Records
- Recognize that substances may be identified in different fields in Reaxys and thus, identification information can be searched in different fields
- Become aware of different policies for the registration of substances in different databases
- Use the main approaches to search substances in Reaxys, including full/part structure, full/part name, and full/part formula
- Know that there are several more specific ways to search substances, including InChI Key, Linear Structure Formula, types of ligands and more

Background

The document Substance Searching Workflow in the Reaxys Support Hub discusses the 'mechanics' of drawing and searching substances by structure. It also describes ways to change search defaults and how to allow for variables in structure search queries.

However, there are other ways to find substances in Reaxys that can be quite effective. This resource summarizes the main options and discusses some important things to consider when planning a substance search.

Where do we find substances in Reaxys?

Substances can be found in different records in Reaxys:

1. In **Reaxys Substance Records**

   Indexed substances carefully selected from around 450 major chemistry and chemistry-related journal titles and key patent sources are catalogued in over 28 million Substance Records.

   The focus of this selection is on the most important substances, complemented with factual information about each substance excerpted from the original documents. This information is organized into >500 Property fields and subfields within the Substance Record.

2. In **PubChem Records**

   Reaxys includes >92 million substances from PubChem.[1] While these substances may be searched separately through PubChem, their integration into Reaxys makes them discoverable through Reaxys search tools and makes it easier for users to access information from two databases.

---

[1] Inclusion of Substance Records from PubChem in NEW Reaxys will be implemented in Q1, 2017.

3. In **eMolecules**

   Reaxys grants access to >8 million substances in eMolecules, which contains information on commercially available substances.

4. In **Reaxys Reaction Records**

   Currently, Reaxys has >43 million Reaction Records into which substances are entered. Many of these Reaction Records contain extensive information on all classes of substances including polymers, materials, and biologics.

5. In **Reaxys Document Records**

   Document Records come from >16,000 chemistry and chemistry-related literature and key patent sources.

   The names of substances appear in titles, abstracts and index headings; formulas for substances may also appear in some of these fields.

   Document Records contain information on many millions of substances including polymers, materials, and biologics.

The important thing to remember is that information about substances in Reaxys is not restricted to Substance Records. There is a wealth of information to be found in the other Reaxys records, as well as additional sources that feed into Reaxys. Thus, a substance search can involve multiple queries, each with a different approach.

Substance identification information in **Substance Records** in Reaxys

Substances are identified in Substance Records in many ways. We can find the list of substance identification fields in **Query builder** by expanding the menu under **Identification** (click the corresponding down arrow in the list of the right-hand panel). The list contains almost 50 fields, but the main ones are:

- **Structure**: Substances with "valence bond" structures — organics, organometallics, co-ordination substances and others — are entered and searched as structures.

- **Name**: Some names of substances are displayed in a Substance Record, but Reaxys also looks up additional substance names in dedicated dictionaries during searches.

- **Formula**: Most substances have formulas, but searching full formulas for organics and related substances may give hundreds of isomeric substances.

  Most simple inorganics have relatively unique formulas, so formula searches are likely to retrieve the required substance.

A note about CAS Registry Numbers

CAS Registry Numbers were developed by CAS to use as a single identifier for the systematic indexing of (precise) substances in CAS databases. Because many substances can be identified easily in this way, CAS Registry Numbers are quite commonly used worldwide.

However, nearly all database producers have alternative ways to (systematically) index substances. Many of these index terms are text-based and are often the most commonly used name for the substance. Some database producers also index substances such as salts and mixtures under the name of the parent substance, so, unlike with the CAS registry system, a single index name can be used to include many closely related substances.

Algorithm-based correlations between CAS Registry Numbers and the systematic substance terms used by other databases are quite difficult to establish. Commonly, intellectual correlations need to be undertaken. In practice, this means that CAS Registry Numbers should be used only as an *initial step* to find the corresponding Substance Records in Reaxys. They should not be considered as precise or comprehensive terms for use in Reaxys, as you may find the answer with an alternative search.

Searching for substances in **Substance Records** in Reaxys

As a first step, we need to consider how best to define the substance or group of substances that we want to find, and then search accordingly. When a substance and related substances of interest are clearly defined by valence bond structures, then (sub)structure searches may be the preferred option. However, keep in mind that Reaxys allows not only searches by full name and/or full formula. We can also search by part name and/or part formula. Thus, if the substances of interest have common part names (e.g., *fulleren…*) or formulas, then searches based on these may be the preferred option.

This applies specially to searches of the type:

 "find <information on> substances that contain <this/these element(s)>"

where Reaxys interprets the following symbols in relation to formulas:

> ? – indicates that the formula can have any number a given element

> * – indicates that the formula can have any number of any other element

Hence, searching *P?O?* in the **Molecular Formula** field retrieves substances consisting of phosphorus and oxygen only, but any number of each element, whereas searching *Au?** finds all substances containing gold.

Searching for substances in **Document Records** in Reaxys

The main two reasons to search substances in Document Records are:

1. The search leverages information in over 54 million records (remember, that Substance Records come from approximately 450 major chemistry and

chemistry-related journals and key patent sources that are manually processed whereas Document Records come from >16,000 sources)

2. The search will include all classes of substances (including polymers, materials, biologics and so forth).

The indexing of substances in Document Records appears in the **Index Terms**. The list of Index Terms in Document Records is an important tool for our searches. Browsing through the index terms in answers helps ensure that our search retrieved the relevant indexing for our substance of interest and can help improve our queries.

Some examples to try

The best way to become familiar with substance searches in Reaxys is to just try different strategies. The exercises below work through many of the different ways to find substances in Reaxys, illustrating some basic principles. Look carefully through results to understand how a search strategy leads to the resulting set of hits.

Structure searches

1. Compare results from a text search for a substance with results from a structure search (As drawn) for the same substance.

   a. In **Quick search**, enter *fluticasone* into the **Search Reaxys** entry box.

      Look at the resulting hit set of substances and note the number of records retrieved.

      Look at the first two substances displayed in the list (the list ranking should be the default **No of References**). How are the two substances different? What we wish to illustrate here is that substances are entered into Reaxys exactly as they were described in the original publication.

      Access the resulting hit set of documents, noting the number of records retrieved.

      What are some of the EMTREE drug terms listed in the Index Terms? Make sure to look through a couple records.

   b. Now use the **Create Structure or Reaction Drawing** entry box of Quick search to perform a structure search for *fluticasone*

      Hint: In the Marvin JS structure editor, click **Create structure template from name** and enter *fluticasone.* Reaxys creates a structure from the name and inserts it into the editor workspace. Make sure the option **As drawn** is checked on the right-hand side and then click **Transfer to query**. Then click **Search**.

      In the list of hit sets that result from the search, note the number of substances. Why is that number different from the search in 1.a.?

      Note also that now we also obtain some **Reaction Records.**

Main learning point: We all want to have a search system in which we simply enter a query and we get the answers we want. To achieve this, however, a search engine must make assumptions about the intent of a query and then implement algorithms based on these assumptions. In the text-based search we did (1.a.), Reaxys implemented algorithms that retrieved a set of substances that did <u>not</u> include salts or mixtures. These algorithms were built to best answer the most common queries made by users. There are times, however, when users want 'something different'. For example, pharmaceutical and agrochemical researchers may want information on salts and mixtures. Thus, it helps to understand how the various searches are performed.

2. A colleague comes to you and says: "When I try a substructure search in a simple query like *2,3-dimethylnaphthalene*, I get substances with multi-ring systems and with further substituents on the two methyl groups. But I just want naphthalenes and I don't want substituents on the two methyl groups. How do I do that?" Create and search a query that meets your colleague's requirements.

   Some steps to help you out:

   Build the *2,3-dimethylnaphthalene* query and check the box **As substructure (On all atoms)**

   Make sure to uncheck the box **Additional ring closures**

   Then in the editor workspace, right-click on a methyl group and select **Atom properties**. A menu with 2 tabs appears; select the **Advanced** tab. Then set the **Substitutions** dropdown menu to **as drawn**. Repeat these steps for the other methyl group.

   Click **Transfer to query** and then **Search**.

   Main learning point: There are many features in substructure searches that have default settings. Among others, these include:

   a) rings in query structures may be isolated or part of larger ring systems
   b) chain bonds in query structures may be in chains or in rings
   c) all positions allow for substitution.

   However, these defaults can be changed to meet specific search criteria.

Name searches

3. Find Substance Records for *fullerenes*.

   Hint: Building structures for all the possible types of fullerenes can be a challenge. Try using the **Chemical Name** field in **Query builder** (do you remember how to find and use search fields in Query builder?) and enter **fulleren** as your search term. Click **Search** and then **Substances**.

   Main learning point: There are cases when building a structure query is challenging or when groups of substances can be described by parts of chemical names. Reaxys allows searches by part name.

4. Find Substance Records of *fullerenes* that contain information on luminescence spectra.

   Start with the answer set of Substance Records from example 3 (above). In the **Filters and Analysis** panel on the left, click **Available Data**. This will allow you to filter the answer set according to your choice of data. Click **More** at the bottom of the **Available Data** menu to get an expanded list of options. Check the box for **Luminescence Spectroscopy** and click **Limit to**. Browse through the results.

   Main learning point: At times, the **Filters and Analysis** options can quickly narrow down hit sets to specific types of information. An alternative approach is to search the two different fields simultaneously. So, in the **Query builder** drag both the **Chemical Name** field and the **Luminescence Spectroscopy** field into the workspace. Enter **fulleren** into the **Chemical Name** field and check **Find any** in the **Luminescence Spectroscopy** field.

Formula searches

5. Find information on the density of lanthanoids.

   Start in Query builder and click the **Molecular Formula** icon at the top of the workspace. The Molecular Formula field is automatically added to the workspace. Now click the <u>same</u> icon on the right-hand side of the field.
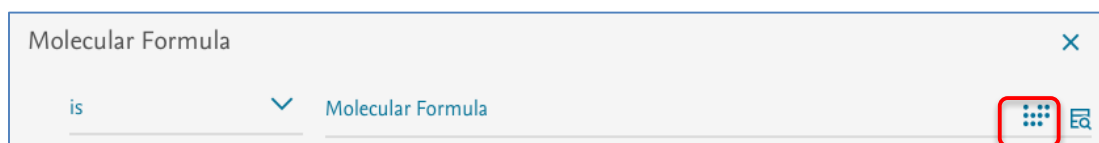
   A new dialog box appears with a periodic table and several chemical groups. This is a graphical user interface (GUI) to build formulas.

   Click **Lanthanoids** at the bottom of the GUI. A short-cut name for that group **(lan)** will appear at the top of the GUI. Click **Use this formula**.

   Click **Search** (**Substances**) and note the number of substances found. Browse some of the Substance Records to understand the types of answers retrieved.

   Use **Filters and Analysis** to find the substances in this hit set that have **Available Data** (**Density**). You should get 26 substances (though that number may change as new data are added to Reaxys). Hint: you can arrange the list of Available Data alphabetically to find **Density** faster.

   You can create a report of your results by clicking **Export** at the top of the hit set. Select **Hit data only** in the resulting menu.

   Once the export is ready, it appears at the bottom of the screen. Click **Download** and browse the report. We can do many things from here; for example, prepare a lecture on the density of lanthanoids.

6. Let's learn more about opportunities for searching Molecular Formulas in Reaxys.

   Go back to the **Query builder**. You should still see the following screen:

Molecular Formula	✕
is ﹀	Molecular Formula (lan) ⋮⋮ 🗒
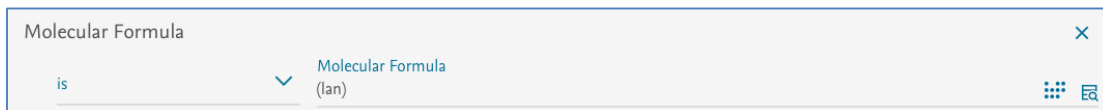
   Compare the answers obtained when you change the search criterion as follows (enter the changes manually):

(lan)	Note: this is the search you just did
(lan)?	
(lan)?*	
(lan)2O[2-5]	Note: allows 2, 3, 4 or 5 oxygens
(lan)S2	
any other formula-based search you would like!	

   Main learning point: Molecular Formulas or part Molecular Formulas are very easy to search in Reaxys, and are particularly useful for searching inorganic substances.

Other search forms

7. Find substances with the general carotene structure that have been studied by time-of-flight mass spectrometry (TOFMS).

   TOFMS is a technique in mass spectrometry, and we know that there are hundreds of techniques specifically indexed in **Substance Records**. Carotenes are a class of C40 terpenes with specific structures, and perhaps the structure editor in Reaxys recognizes the carotene structure. Thus, we could combine a property search with a structure search.

   The search steps would be:

   In **Quick search**, open the structure editor by clicking **Create Structure or Reaction Drawing.** Then use the **Create structure template from name** option to enter *carotene* and transfer the structure to the editor.

   Since double bonds can be almost anywhere in the substances required, we need to specify all bonds as **any**. To do this:

   a. Select the entire structure by dragging the **Selection tool** (top icon to the left of the editor workspace) across the structure.
   b. Right-click any bond and select **Bond properties** from the menu.
   c. Set the **Type** menu to **any**.

Finally, click **As substructure** in the options to the right of the structure editor and click **Transfer to query.**

Back in **Quick search**, enter *tofms* into the **Search Reaxys** entry box.

Click **Search**. Reaxys produces 4 hit sets, the first of which consists of Substance Records (290). Browse the **Hit Data** contained in the records.

Often we want to start a search with a property and then find substances that have that property. This type of search is discussed in the resource "Finding Property Information in Reaxys" posted on the Explore Reaxys module as Reaxys_PropertiesSearch_200117.

For the most part, search opportunities using the other substance identification fields available in **Query builder** are straightforward, except perhaps the **One-Center Ligands** and **Multi-Center Ligands** fields. These fields contain information on a metal and atoms in ligand(s) attached to the metal (in the form of codes) that may be present in a compound.
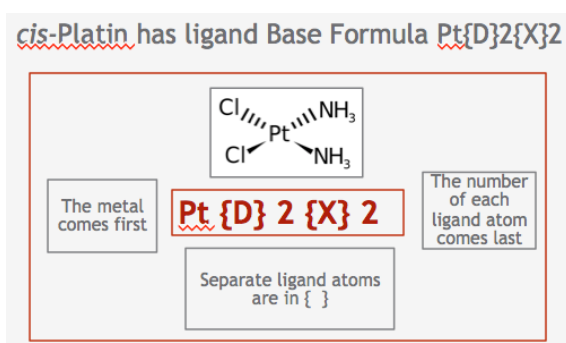
Example of a Ligand Base Formula

Field codes for co-ordinating atoms

Searches in these fields are beyond the scope of this resource but are useful for finding co-ordination compounds. Those working in the field are encouraged to contact Help in Reaxys for further information.