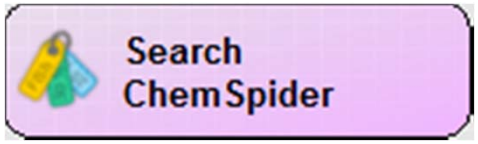
Compound Discoverer 2.1 ChemSpider Search Node



Parameters of 'Search ChemSpider'

Show Advanced Parameters

1. Search Settings

Database(s)	
Mass Tolerance	5 ppm
Max. # of results per compound	100
Max. # of Predicted Compositions	3

- The Search ChemSpider node can be given either
  - Compound molecular weight (connected directly to Group Unknown Compounds)
  - Elemental Compositions (connected to Predict Compositions)
  - Both (connect to both nodes).
- It will provide a list of candidates based on the information searched.
- A key parameter is what Database(s) (called “Data Sources” on ChemSpider) are searched.
- The following are general suggestions on good Databases to use.

Data sources to Avoid

Typically, do NOT include datasources that have many millions of compounds as ChemSpider will be slow to process and return results for a dataset.

- Aurora Feinchemie
  - >25 million, compounding library (limited applicability)
- PubChem
  - >10 million, too diverse/large
  - Could consider using it for a “MaxID” workflow
- MOlPort
  - >5 million
  - Compounding library source (novel structures)

Useful “General” Datasources

These datasources contain compounds that are widely relevant to a number of sample types and applications.

- ChEBI (84000) – Typically endogenous or compounds used as research tools. Relevant for environmental and metabolomics.
- FDA UNII – NLM (62000) – approved and retired food and drug ingredients and additives. Good when analyzing any human or environmental sample.

Data sources for Metabolomics

- General
  - Biocyc (7000) – focused on known compounds from multiple species
  - Cayman Chemical (8000) – mixed endogenous and drug of abuse. Covers prostaglandins.
  - LipidMAPS (7000) – lipid data source.
  - MCISB (15000) – biological pathways (mixed), Manchester Center for Integrative Systems Biology
  - SMPDB (645) – general small molecule biological pathway compounds.
  - KEGG (23000) – general “pathways of all life” compounds.
- Mamalian
  - Human Metabolome Database (38000) – common metabolites as well as drugs, food, and cosmetic ingredients and environmental exposure contaminants.
- Plant
  - AnalytiCon Discovery (38k) – Good, but has significant amount of synthetics mixed in with endogenous metabolites. Still, one of the most sizable databases for plants.
  - AraCyc (1900) – small but plant specific (Arabidopsis).
  - Baoji Herbest Bio-Tech (500) – very small, overlaps with some of the larger options.
  - Extrasynthese (1000) – chemical supplier of isolated plant standards.
  - Golm Metabolome Database (1300) – primarily plant, will overlap with some of the larger data sources.
  - Indofine (11000) – company selling isolated natural product standards. Datasource also has other synthetic compounds included.
  - PlantCyc (4000) – general plant data source.
  - Sequoia Research Products (2300) – general plant standard provider.
- Specific
  - E Coli Metabolome Database (700) – very small, useful to give additional coverage.
  - Yeast metabolome database (1000)

Data sources for Forensics

- FDA UNII – NLM (62000) – approved and retired food and drug ingredients and additives.
- Drugbank (7000) – therapeutic drugs.
- Cayman Chemical (8000) – mixed endogenous and drug of abuse.
- LGC Standards (10000) – mixed but contains several designer drugs
- Toxin Toxin-Target Database (1800) – small but specific data source on toxins, natural and synthetic.

Data sources for Environmental

- ACToR – EPA database of environmental chemicals of interest
- Drugbank (7000) – therapeutic drugs, common contaminant in wastewater.
- EAWAG Biocatalysis/Biodegradation Database (1000) – predicted breakdown products / mineralization endpoints for several compounds.
- EPA DSSTox (670000) – very large database of compounds of environmental interest.
- FDA UNII – NLM (62000) – approved and retired food and drug ingredients and additives.
- MICAD (136) – very small but has medical imaging compounds which are common contaminants.
- Toxin Toxin-Target Database (1800) – small but specific data source on toxins, natural and synthetic.

Data sources for Food

- FDA UNII – NLM (62000) – approved and retired food and drug ingredients and additives.
- Food and Agriculture Organization of the UN (1500) – Approved food ingredients and additives in the EU (subset only)
- FooDB (245) – known common food ingredients/endogenous substances.
- Toxin Toxin-Target Database (1800) – small but specific data source on toxins, natural and synthetic.