ChEBI

USER MANUAL

Contents

1. Introduction

Chemical **E**ntities of **B**iological **I**nterest (ChEBI) is a freely available dictionary of 'small molecular entities'. The term 'molecular entity' encompasses any constitutionally or isotopically distinct atom, molecule, ion, ion pair, radical, radical ion, complex, conformer, etc., identifiable as a separately distinguishable entity. The molecular entities in question are either products of nature or synthetic products used to intervene in the processes of living organisms (either deliberately, as for drugs, or unintentionally', as for chemicals in the environment). The qualifier 'small' implies the exclusion of entities directly encoded by the genome, and thus as a rule nucleic acids, proteins and peptides derived from proteins by cleavage are not included. Classes of molecular entities and part-molecular entities (in the form of substituent groups or atoms) are also included in ChEBI.

All data in the database is non-proprietary or is derived from a non-proprietary source. It is thus freely accessible and available to anyone. In addition, each data item is fully traceable and explicitly referenced to the original source.

This manual is set up as a guide for navigating the ChEBI database. For details on possible search options in the database see Chapter 8, Searching the ChEBI database. ChEBI employs nomenclature and terminology recommended by the following international bodies:

- International Union of Pure and Applied Chemistry (IUPAC)
- Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB)

In addition ChEBI incorporates an ontology, whereby the relationships between compounds, groups or classes of compounds and their parents, children and/or siblings are specified.

For each small molecular entity in the ChEBI database, information is displayed over several tabs. The Main tab is displayed by default and shows general information about the entity (ChEBI name, ID, structure, definition, etc. chapter 2). Other tabs displays more specialised information: ChEBI Ontology (chapter 3), Automatic X-references (chapter 4), Reactions (chapter 5), Pathways (chapter 6) and Models (chapter 7). The following chapters describe the different tabs and the information provided in more detail.

2. Main

2.1 ChEBI Name

The name for an entity recommended for use by the biological community. In general traditional names have been retained by ChEBI but these may have been modified to enhance clarity, avoid ambiguity and follow more closely current IUPAC recommendations on chemical nomenclature.

For more information see the Annotation Manual.

2.2 ChEBI ID

A unique and stable identifier for the entity, for example, CHEBI:16236. It has no chemical significance and may be cited by external users.

2.3 ChEBI ASCII Name

The ChEBI Name is also provided in ASCII format if the original includes special characters which require a Unicode presentation.

2.4 Definition

A short verbal definition is included in some entries (and for *all* new entries annotated after June 2009). For more information see the Annotation Manual.

Wikipedia: In addition to a definition, for those compounds or classes for which ChEBI provides a database accession link to Wikipedia, the first paragraph of the Wikipedia entry is reproduced, together with a link to the full article.

2.5 3-Star status

Entries which have been manually annotated by the ChEBI team are indicated by the presence of a '3-star' symbol. This is shown on the main display screen for an entity and on the search results page. An absence of a '3-star' symbol indicates that the entity has been manually annotated by a third party, or (occasionally) that it has been marked as deleted or obsolete.

2.6 Secondary ChEBI IDs

Here are listed the IDs of any entries which may have been subsumed into the parent via merging of duplicate entries. Searching for a secondary ID would always redirect to the most updated parent ID.

2.7 Submitter Information

If an entry is present by virtue of its having been submitted via the ChEBI Submission Tool, the name of the submitter is displayed here (unless the submitter has elected to remain anonymous).

2.8 Supplier Information 

The supplier information contains an eMolecules number and/or a zinc number if available. These numbers provide links which connect to free to use websites that provide information from a large range of suppliers. This information allows the user to directly access suppliers for the entity. If an eMolecules or zinc number is not available, the message "no supplier information found for this compound" will appear.

If an eMolecules link does not provide supplier information it is likely to be because the eMolecule has classified the compound as Prohibited/Forbidden, and a login account may be required to access the information.

2.9 Chemical Structure

When appropriate, a chemical structure of an entity will be displayed. Hovering the cursor over the structure will cause it to enlarge for extra clarity. The structure can be saved as a png file and as a molfile. If desired, the size of the structure can be altered by opening the copied image location and changing the dimensions in the address browser. In order to create a transparent background and remove the white background simply add "&transbg=true" to the end of text in the address browser. There are alternative structures shownincluding a JSmol which would give a 3D projection of the entity. By clicking on expand it is possible to rotate the entity in 3D.

Below the structure there are three options to search the ChEBI database based on the chemical structure. It is possible to search for compounds which contain the structure or which resemble the structure. A third option is to take the structure to the Advanced Search section.

2.10 Wikipedia

If available the Wikipedia introductory paragraph of an entity has been added including a link to read the full article on Wikipedia.

2.11 Formula

Where possible, molecular formulae are assigned for entities and groups. To facilitate searching and downloading of data from external sources, the use of subscripts to indicate multipliers is avoided.

The following conventions regarding ChEBI formulae are followed:

- Unless immediately following a dot '.' any numeral refers to the preceding element in the formula. Example: H2O means there are two hydrogen atoms and one oxygen atom.
- The dot '.' convention is used when dividing a formula into parts. Any numeral following a dot refers to all the elements within that part of the formula that follow it. Example: C2H3O2.Na.3H2O (CHEBI:32138) really means that after C2H3O2 there is one sodium (Na), six hydrogen and three oxygen atoms.
- Parentheses are used within ChEBI formulae to denote multiplication of elements.
- The 'n' convention is used to show an unknown quantity by which a formula is multiplied. For example: (C2H4)n from polyethylene, CHEBI:53227, really means that a C2H4 unit is multiplied by an unknown quantity.
- In some cases there can be a formula relating to a subclass of structures which have a general formula but with an R representing the variety of groups attached. For example: ceramides (CHEBI:17761) formula is C4H6NO2R3 which includes the general *N*-acyl-sphingoid base (C4H6NO2) and three different R groups.

For more information see the Annotation Manual.

2.12 Net Charge

The charge is the sum of all the positive and negative charges shown in the structure. For ions the magnitude of the charge is given in arabic numerals preceded by the sign of the charge. For neutral molecules the charge is indicated as a numerical zero. For instance, the charge of 5,10,15,20-tetrakis(1-methylpyridinium-4-yl)porphyrin (CHEBI:37447) is +4; the charge of borate (CHEBI:22908) is -3.

2.13 Average Mass

Relative molecular, atomic and ionic masses are shown for molecular, atomic and ionic entities respectively. The relative masses are calculated from tables of relative atomic masses (atomic weights) published by IUPAC.

For entities that have R groups included in their formula the average mass will be stated as: Average mass (excl. R groups).

2.14 Monoisotopic Mass

The monoisotopic mass relates to the sum of the masses of all atoms present in a molecule using the unbound, ground-state, rest mass of the principal (most abundant) isotope for each element.

For entities that have R groups included in their formula the monoisotopic mass will be stated as: Monoisotopic mass (excl. R groups).

2.15 IUPAC International Chemical Identifier (InChI)

The InChI is a non-proprietary identifier for chemical substances that can be used in printed and electronic data sources thus enabling easier linking of diverse data compilations. It expresses chemical structures in terms of atomic connectivity, tautomeric state, isotopes, stereochemistry and electronic charge in order to produce a string of machine-readable characters unique to the respective molecule. Further information regarding InChI is available at http://www.iupac.org/inchi/.sequence

A very useful InChI FAQ is also accessible at the inchi-trust.

2.16 InChIKey

The InChIKey is a 25-character hashed version of the full InChI, designed to allow for easy web searches of chemical compounds. InChIKeys consist of 14 characters resulting from a hash of the connectivity information from the full InChI string, followed by a hyphen, followed by 8 characters resulting from a hash of the remaining layers of the InChI, followed by a single character indicating the version of InChI used, followed by single checksum character. There is a very remote probability of finding two structures with the same InChIKey. However the probability for duplication of only the first block of 14 characters has been estimated as one duplication in 75 databases each containing one billion unique structures; such duplication therefore appears highly unlikely.

Further information regarding the InChIKey is available at http://old.iupac.org/inchi/release102.html.

2.17 SMILES

SMILES (Simplified Molecular Input Line Entry System) is a simple but comprehensive chemical line notation, created in 1986 by David Weininger and further extended by Daylight Chemical Information Systems, Inc. SMILES specifically represents a valence model of a molecule and is widely used as a data exchange format.

Further information on SMILES is available at http://www.daylight.com/smiles/.

2.18 Metabolite of Species

For entities which have been detected in or isolated from living organisms, the species name, link to a taxonomy and location (stem, root, blood, urine etc.) where the entity was found is provided, together with an appropriate citation.

2.19 Roles Classification

The Roles Classification is based on the ChEBI Ontology which is a structured classification of the entities contained within ChEBI. For more details about the ontology see Section 5 below. The roles classification has been divided into three sub-categories: 'chemical role' which classifies entities on the basis of their role within a chemical context, e.g. as ligand, inhibitor, surfactant; 'biological role' which classifies entities on the basis of their role within a biological context, e.g. antibiotic, antiviral agent, coenzyme, hormone; and 'application' which classifies on the basis of their intended use by humans, e.g. pesticide, antirheumatic drug, fuel.

2.20 ChEBI Ontology

This includes the outgoing and incoming relationships between a ChEBI entry and its immediate related entities. Further information on ontology is described in chapter 3.

2.21 IUPAC name

A name provided for an entity based on current recommendations of IUPAC. It need not be fully systematic as it makes use of 'retained names'.

Example: The IUPAC Name for abietic acid (CHEBI:28987) is abieta-7,13-dien-18-oic acid, based on the retained name 'abietane', rather than the fully systematic name (1*R*,4a*R*,10a*R*)-1,4a-dimethyl-7-(propan-2-yl)-1,2,3,4,4a,5,6,10,10a-decahydrophenant hrene-1-carboxylic acid (which is cited in ChEBI within the list of synonyms for this compound).

The IUPAC name provided is not necessarily the Preferred IUPAC Name (PIN). For further information on Preferred IUPAC names see Chapter 5 of Nomenclature of Organic Chemistry: IUPAC Recommendations and Preferred Names 2013, available via the Royal Society of Chemistry here.

2.22 INN

In cases where an entity is a pharmaceutical substance, an International Nonproprietary Name (INN) may be shown. The INN is the official non-proprietary or generic name given to a pharmaceutical substance, as designated by the World Health Organisation

(WHO). INNs may appear in ChEBI as English, Latin, Spanish and French language versions.

2.23 Synonyms

Alternative names for an entity which either have been used in EBI or external sources or have been devised by the curators based on recommendations of IUPAC, NC-IUBMB or their associated bodies. The source of each synonym is clearly identified (see References Manual). Systematic names may also be included in this section. In addition to English-language synonyms, versions may be shown in French  , German  , Spanish  and Latin  , the language being indicated by a flag.

For more information see the Annotation Manual.

2.23.1 Adapted Synonyms

Synonyms are normally reproduced in the exact form in which they appear in the source. However, where changes have been made, e.g. to correct syntax or to convert from an index style of presentation, then this is indicated by  .

2.24 Brand names

Where an entity is an active ingredient of a proprietary pharmaceutical preparation, the brand name of the preparation may be shown. For instance Panadol and Tylenol are brand names for paracetamol (CHEBI:46195).

2.25 Database links

Direct links to the entries for an entity in the databases cited, as detailed in the References Manual.

2.26 Registry Number(s)

The Chemical Abstracts Service (CAS) Registry Number is a unique numeric identifier assigned to a substance when it enters the CAS REGISTRY database. Registry Numbers have no chemical significance and are assigned in sequential order to unique, new substances identified by CAS scientists for inclusion in the database.

Two principles of ChEBI are that (1) nothing held in the database must be proprietary or derived from a proprietary source that would limit its free distribution and/or availability and (2) every data item in the database should be fully traceable and explicitly referenced to the original source. As such, it is impossible for ChEBI to cite CAS as a source for Registry Numbers as this organization's products are not freely accessible.

ChEBI therefore cites other reliable and freely accessible sources for CAS Registry Numbers which are always fully referenced.

Other registry numbers which may be displayed are Beilstein and Gmelin Registry Numbers which are now part of the Reaxys database. For more information see the Annotation Manual.

2.27 Citation(s)

Publications which cite the entity are listed here, along with hyperlinks to the PubMed entry via Europe PMC. This is a web application managed and developed by EMBL-EBI that provides free access to citation and abstract or full text details of biomedical and life sciences literature. Clicking on the 'Show Abstract' link displays the abstract as contained within Europe PMC.

2.28 Last Modified

Date when the last modification of the entity took place.

2.29 Comment(s)

A free-text comment may be added to some terms especially in cases where confusing terminology has been historically used. A comment may relate to a single term or to the entry as a whole.

Further details regarding the data resources used for 2.23 - 2.27 are provided in the References Manual.

3. ChEBI Ontology

The ChEBI Ontology is a structured classification of the entities contained within ChEBI. Originally developed as 'Chemical Ontology' by Michael Ashburner and Pankaj Jaiswal, the initial alpha release was subsumed into ChEBI. Its structure is essentially that of a directed acyclic graph (DAG), which differs from a simple taxonomy in that a child term can have many parent terms. Additionally, a number of relationships are incorporated which are cyclic in nature.

The ChEBI Ontology section of the Main display page (see section 2.17) shows some of the information contained within the ontology; further details are provided on the Ontology display page (separate tab).

Information included on the main page includes the roles classification (chemical roles, biological roles and applications as mentioned below in 3.1) as well as the outgoing and incoming ontology relationships (i.e. a description in words of the relationships between a ChEBI entry and its immediate related entities).

3.1 The Ontologies

The ChEBI Ontology is subdivided into three separate sub-ontologies:

- Molecular structure, in which molecular entities or parts thereof are classified according to composition and structure, e.g. hydrocarbons, carboxylic acids, tertiary amines;
- Role, divided into three sub-categories: 'chemical role' which classifies entities on the basis of their role within a chemical context, e.g. as ligand, inhibitor, surfactant; biological role which classifies entities on the basis of their role within a biological context, e.g. antibiotic, antiviral agent, coenzyme, hormone; and 'application' which classifies on the basis of their intended use by humans, e.g. pesticide, antirheumatic drug, fuel;

- Subatomic Particle, which classifies particles which are smaller than atoms, e.g. electron, photon, nucleon.

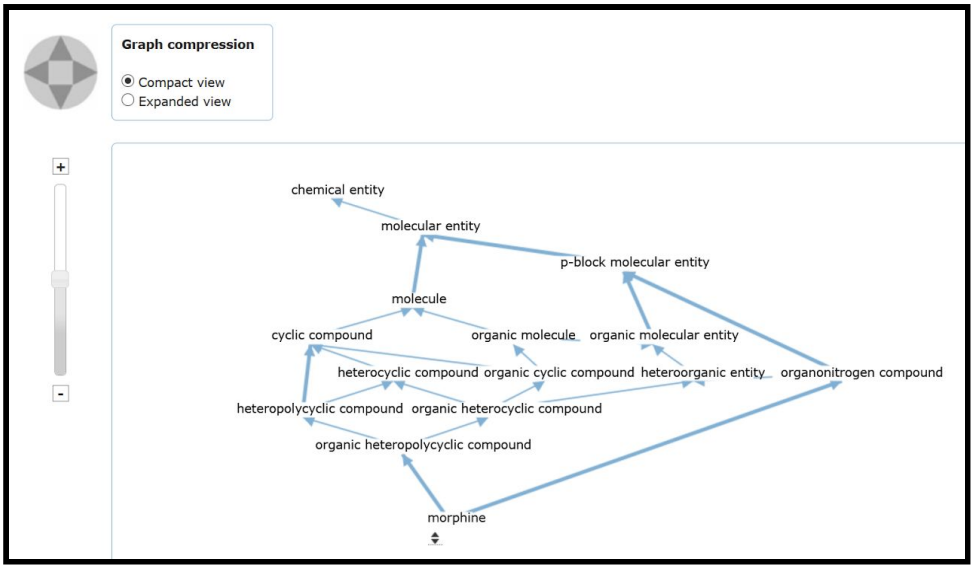
3.2 The Views

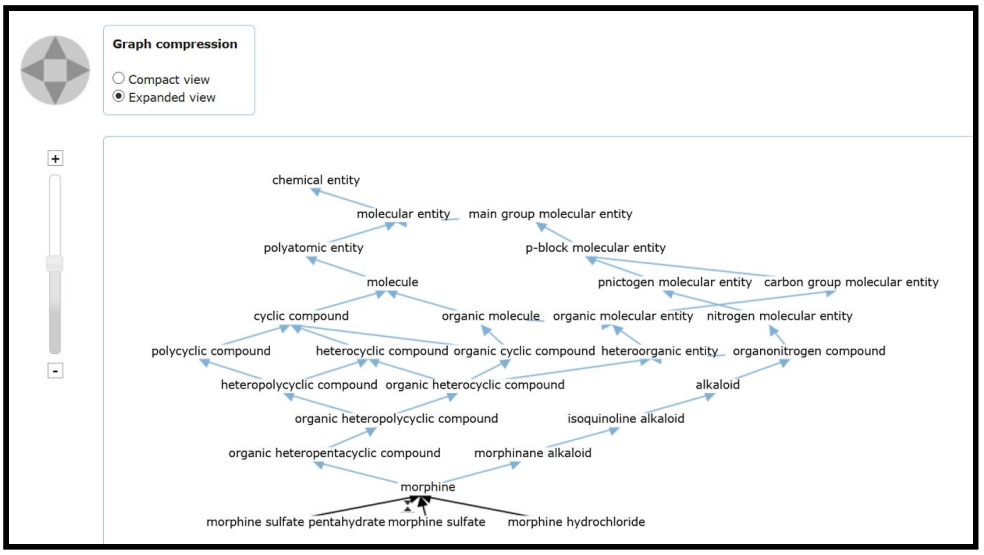
The roles and applications shown on the Main display page are repeated on the Ontology page. Additionally, related structures (for instance functional parent or structural derivative) and two graph representations of the ontology relationships (tree view and graph view).

In addition to the incoming and outgoing view on the main page, the ontology tab provides two options for visualising the ontology relationships of an entry in ChEBI, namely a graph view and a tree view.

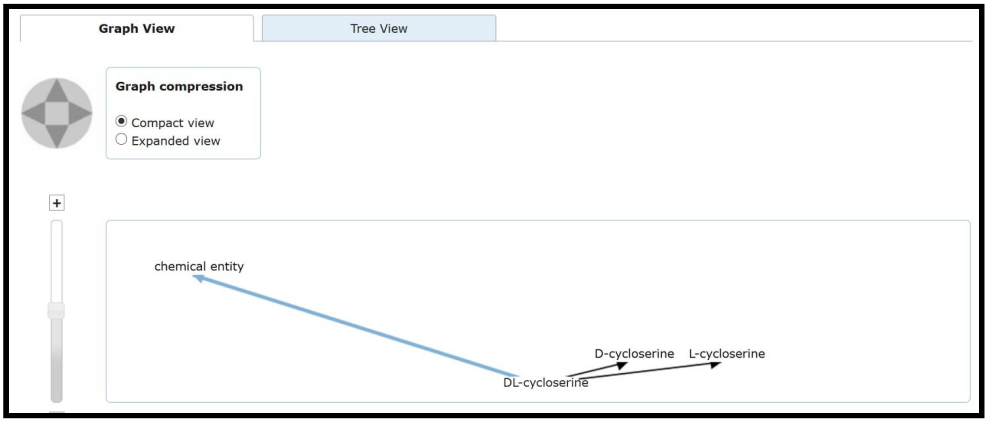
3.2.1 Graph View

Graph View represents a visualization of relationships "is a" and "has part of" between the ontologies of a certain entity. With the graph view there are two options, the compact view or expanded view. In the compact view only the outgoing ontologies are displayed while the other ontology paths are hidden and represented by thicker drawn lines. These hidden paths can subsequently be displayed by clicking on these specific connection lines. Both graphs have the ability to zoom in and out if required As an example, the compact and expanded ontology graph views of morphine (CHEBI:17303) are shown below.



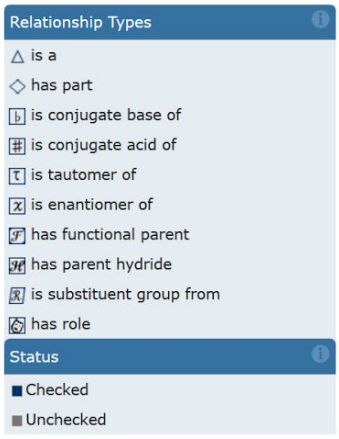


Clicking on ontology terms within the graph view will give you a definition for that specific term. One extra option is available for substructures in graph view for entities that contains an ontology relationship of "has part". For example DL-cycloserine (CHEBI:27792) which contains part D-cycloserine and L-cycloserine which can be seen by the black links in the graph view.



3.2.2 Tree View

Another option of viewing the ontology relationships of an entity is by a tree view where all parents within the hierarchy are shown, as well as the immediate children. Adjacent is a key identifying the relationships used within the tree structure.



All different relationship types between the ontologies of an entity are described with a specific symbol as shown. Details of each relationship is described below.

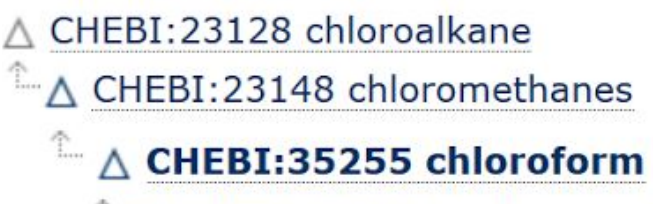
Entries and relationships which have been checked by a curator are shown in blue while those added externally (unchecked) are in grey. Clicking on a node within the tree will take the user to the ChEBI entry for that node. Unchecked ChEBI entries accessed by this route will display the heading 'Preliminary ChEBI Entry'.

3.3 The Relationships

For each relationship a formal definition is included beneath the description.

3.3.1 △ is a

Implies that 'Entity A' is a subtype of 'Entity B'. E.g.



or, in words, chloroform ($CHCl_3$, CHEBI:35255) is a subtype (member) of the class of chloromethanes (CHEBI:23148), which in turn is a subtype of the class 'chloroalkane' (CHEBI:23128); hence all instances of chloromethanes are also instances of chloroalkanes.

Definition: "C is_a C' if and only if: given any c that instantiates C at a time t, c instantiates C' at t."

3.3.2 ◇ has part

Used to indicate the relationship between part and whole. E.g.

or, in words, potassium tetracyanonickelate(2−) (CHEBI:30071) has part tetracyanonickelate(2−) (CHEBI:49928).

Definition: "C has_part C' if and only if: given any c that instantiates C at a time t, there is some c' such that c' instantiates C' at time t, and c has c' as a part at t."

3.3.3 ♭ is conjugate base and ♯ is conjugate acid of

Cyclic relationships used to connect acids with their conjugate bases. E.g.

and

Thus, the neutral pyruvic acid (CHEBI:32816) is the conjugate acid of the pyruvate anion (CHEBI:15361), while as a corollary pyruvate is the conjugate base of the acid.

Definition: "A is_conjugate_acid_of B if and only if, given any a, a instantiates A and has the disposition to be a Bronsted Acid, then there is some b, such that b instantiates B and has the disposition to be a Bronsted Base, such that b derives from a through the removal of a proton as the result of a chemical transformation process."

3.3.4 ⁊ is tautomer of

A cyclic relationship used to show the interrelationship between two tautomers, where the differences between the structures are significant enough to warrant their separate inclusion in ChEBI. E.g.

and

Thus, isocyanuric acid (CHEBI:17696) and cyanuric acid (CHEBI:38028) are tautomers.

Definition: "A is_tautomer_of B if and only if, given any a which instantiates A and has composition ca and is described by a molecular graph ag, there is some b that instantiates B, has composition cb and is described by a molecular graph bg, such that ca equals cb, ag is different from bg and a derives from b as the result of an

intramolecular chemical transformation process (i.e. a chemical transformation process which has only one participant), in which only bonds to hydrogen are broken or formed."

3.3.5 is enantiomer of

A cyclic relationship used in cases when two entities are mirror images of and non-superposable upon each other. E.g.

and

Each relationship shows that (*S*)-camphor (CHEBI:15397) is an enantiomer of (*R*)-camphor (CHEBI:15396) and vice versa.

Definition: "A is_enantiomer_of B if and only if, given any a that instantiates A, has molecular graph ag, there is some b such that b instantiates B, is described by molecular graph bg, such that ca is equal to cb and ag is transformed into bg through a C2 symmetric transform."

3.3.6 has functional parent

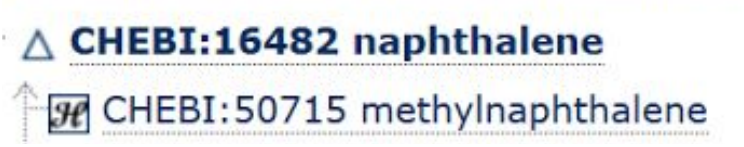
Used to denote the relationship between two molecular entities (or classes of entities), one of which possesses one or more characteristic groups from which the other can be derived by functional modification. E.g.

Or, in words, 11-deoxycorticosterone (CHEBI:16973) can be derived by functional modification of progesterone (CHEBI:17026).

Definition: "A has_functional_parent B if and only if given any a, a instantiates A , has molecular graph ag and a obo: has_part some functional group fg, then there is some b such that b instantiates B, has molecular graph bg and has functional group fg' such that bg is the result of a graph transformation process on ag resulting in the conversion of fg into fg'."

3.3.7 has parent hydride

Denotes the relationship between an entity and its parent hydride (defined by IUPAC as "an unbranched acyclic or cyclic structure or an acyclic/cyclic structure having a semisystematic or trivial name to which only hydrogen atoms are attached"). E.g.



Thus methylnaphthalene (CHEBI:50715) has as its parent hydride the cyclic hydrocarbon naphthalene (CHEBI:16482).

Definition: "A has_parent_hydride B if and only if given any a, a instantiates A , has molecular graph ag and a obo: has_part some functional group fg, then there is some b such that b instantiates B, has molecular graph bg such that bg is the result of a graph transformation process on ag resulting in the removal of fg and its replacement by a hydrogen atom."

3.3.8 is substituent group from

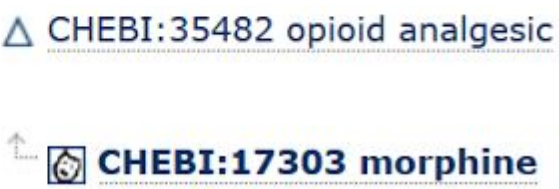
Indicates the relationship between a substituent group (or atom) and its parent molecular entity, from which it is formed by loss of one or more protons or simple groups such as hydroxy groups. E.g.

The L-valyl group (CHEBI:32853) is derived by hydroxide loss of L-valine (CHEBI:16414).

Definition: "A is_substituent_group_from B if and only if A is a group and B is a molecular entity; given any a that instantiates A, a has molecular graph ag and specified attachment point agap, and there is some b that instantiates B and has molecular graph bg, then it is the case that bg is the result of a graph transformation process on ag resulting in the replacement of agap by some group bgg (which may be a hydrogen atom or a more complex group)."

3.3.9 has role

Indicates the particular behaviour which an entity may exhibit, either naturally or by human application. E.g.



Thus morphine (CHEBI:17303) has a role opioid analgesic (CHEBI:35482).

Definition: "Chemical entity C has_role role R if and only if: given any c that instantiates C at t, there exists some r that instantiates R at t, and c is the bearer of r at t."

3.4 Open and closed classes

There are different types of classes in the ChEBI ontology. The majority of classes have an infinite number of possible members and are referred to as open classes. For instance, toluenes (CHEBI:27024) includes toluene itself and all substituted toluenes. Large open classes can be divided into sub-classes, e.g. hydroxytoluenes (CHEBI:24751) is a subclass of toluenes.

In some instances, classes have restricted numbers of members. These are referred to as closed classes. Some examples of closed classes are described below.

- alanine (CHEBI:16449) can contain two and only two stereoisomers: L-alanine (CHEBI:16977) and D-alanine (CHEBI:15570).
- cresol (CHEBI:25399) can have three structural isomers; *ortho*-cresol (CHEBI:28054), *meta*-cresol (CHEBI:17231) and *para*-cresol (CHEBI:17847).
- pyrrole (CHEBI:35556) can have three different tautomers; 1*H*-pyrrole (CHEBI:19203), 2*H*-pyrrole (CHEBI:35558) and 3*H*-pyrrole (CHEBI:35557). The corresponding open class is named pyrroles (CHEBI:26455) and encompasses pyrrole itself and all substituted pyrroles.

4. Automatic Xrefs

The Automatic Xrefs are generated by matching names and/or structures against those present in various freely-available databases, described in detail in the References Manual. The references are divided into the seven categories shown below including the data sources used for the automatic generated x-references:

- **Protein Sequences**
  - UniProtKB which is the central access point for extensive curated protein information, including function, classification, and cross-reference
- **Reactions and Pathways**
  - BioModels which is a database of containing mathematical models of biological interest
  - Rhea is a freely available, manually annotated database of biochemical reactions.
  - Reactome contains a curated knowledge base of biological pathways
  - SABIO-RK is a system for the analysis of biochemical pathways focussing on the kinetics of reactions
- **Small molecules**
  - NMRShiftDB which contains a NMR database of organic structures and their nuclear magnetic resonance (nmr) spectra
  - PubChem also provides information on the biological activities
  - Golm contains a database with the mass spectra from metabolites quantified using gas chromatography coupled to mass spectroscopy (GC-MS)
  - MassBank which contains high quality mass spectral databases
- **Enzymes**
  - BRENDA which is a comprehensive enzyme information system
  - IntEnz contains an integrated relational enzyme database
  - Enzyme Portal integrates publicly available information about enzymes
- **Gene expression**
  - ArrayExpress (Repository of Microarray data) is a public repository for transcriptomics and related data

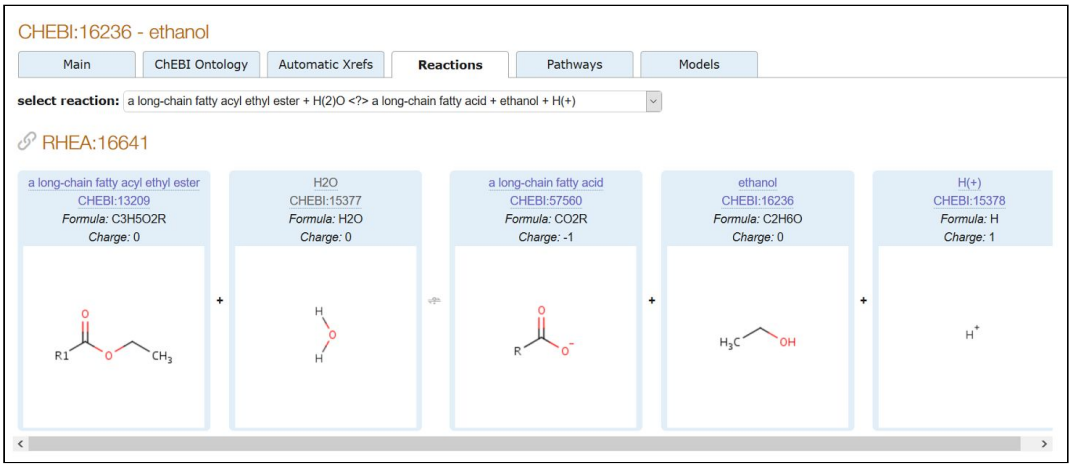
- ArrayExpress (Gene Expression Atlas) Database servicing amongst others, queries for condition-specific gene expression patterns
- NURSA the nuclear receptor signaling atlas is an online information resource for the nuclear receptor signaling community

- **Molecular interactions**
  - IntAct Interactions is a database of evidence for molecular interactions
  - IEDB a database of epitopes involved in infectious disease, allergy, autoimmunity and transplant
  - ChEMBL is a database of bioactive drug-like small molecules and associated bioactivities abstracted from the scientific literature
  - CompTox provides a high quality public chemistry resource for supporting improved predictive toxicology
- **Literature**
  - Patents a collection of biology-related abstracts of patent applications.

5. Reactions

The reactions page is based on information provided by the Rhea database which is a manually annotated database of chemical reactions in which all reaction participants (reactants and products) are linked to ChEBI. While its main focus is enzymatic reactions, other biochemical reactions are included.

For instance ethanol (CHEBI:16236) is included in twelve different enzyme-mediated reactions. All the entities (reactants and products) are presented when a specific reaction is selected. Each reaction pathway shows the link to all the ChEBI entities taking part in the reaction and includes the ChEBI ID, structure, formula and charge. The example below focusses on the reaction of a long-chain fatty acyl ethyl ester and water in equilibrium with a long-chain fatty acid and ethanol in an acidic environment.

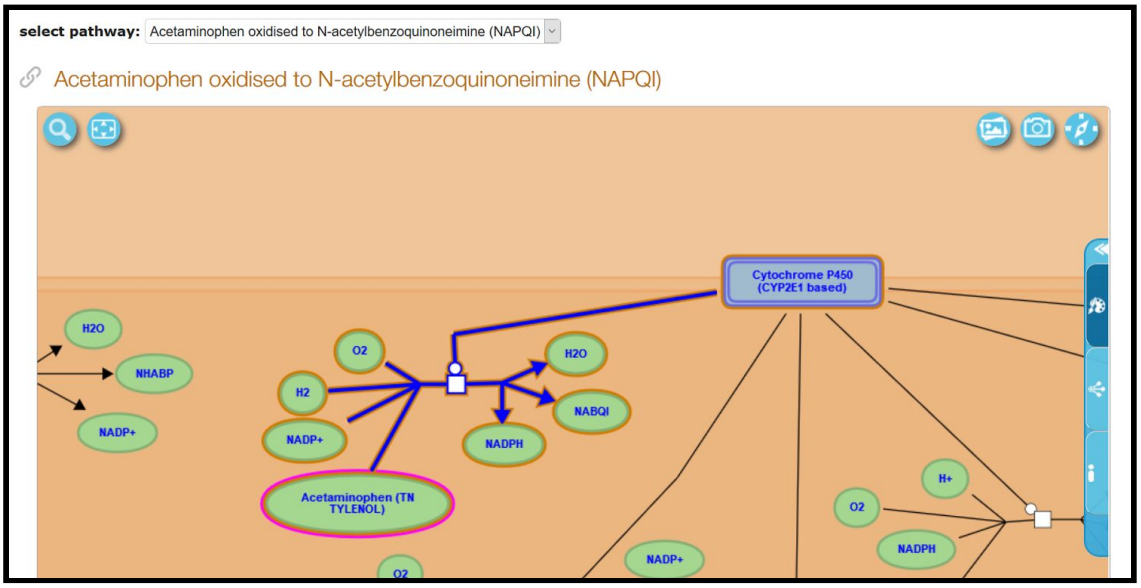


The link to the specific Rhea website in presented at the top of the page which provides more data including citations of reaction involved.

6. Pathways

The pathways represented in the ChEBI database are sourced from the Reactome database. Reactome is a curated database of pathways and reactions (pathway steps) focusing on human biology. The Reactome definition of a 'reaction' includes many events in biology that are changes in state, such as binding, activation, translocation and degradation, in addition to classical biochemical reactions.

For example paracetamol (CHEBI:46195) contains three separate pathway of which the oxidation of acetaminophen to N-acetylbenzoquinoneimine (NAPQI) pathway is shown below in bold dark blue lines:

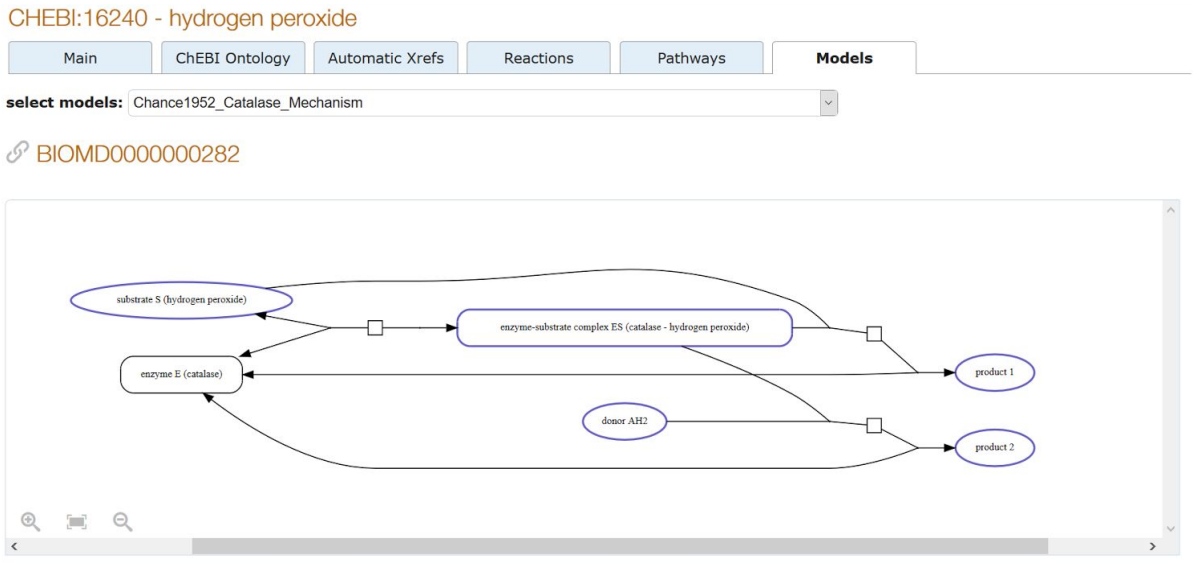


The specific pathway includes a link to the Reactome web page providing more detailed information. It includes the type of pathway, in what organism the pathway takes place, which specific compartment of the organism, literature references and links for the participants of the specific pathway. There are two entities flagged within this pathway. The entity itself indicated by the heavy pink border and where the reaction takes place.

7. Models

The BioModels Database is a repository of computational models of biological processes. Models described from the literature are manually curated and enriched with cross-references. The Models tab shows the biological processes that the entity participates in and for each process displays the entity itself and also a ChEBI link to other entities that are involved in the same process. When hovering over the entity it produces the structure, name, ID, definition, formula, average mass and net charge.

For example hydrogen peroxide (CHEBI:16240) takes part in 17 different pathways. One is shown below: Chance1952_Catalase_Mechanism.



5 CHEBI entities are flagged in blue lines.

Also included is a link directly to the specific data from the BioModels database which includes information about the mathematical expressions, physical entities and global parameters.

8. Searching the ChEBI database

The search interface for the ChEBI database comprises two parts: a quick text search and an Advanced Search. Text searching in both the quick and Advanced searches employs Lucene, a full-featured text search engine library written entirely in Java, while the structure search facility of the Advanced Search uses the new chemical structure search algorithm OrChem, an Oracle chemistry plug-in using the Chemistry Development Kit.

8.1 Quick text search

A text search box is provided on the home page. This enables users to enter either a precise search term or one employing wild cards. **The wild-card character is the asterisk (*)**. The search engine will then search for that term through all of the text fields within the ChEBI entries, and list the results using a scoring mechanism based on similarity, the compound with the highest score being listed first. In the table for each result is shown the structure (if one exists within the database), the ChEBI ID and Name, the Text Search Score and the status (2- or 3-star) of the entry. Clicking on the ChEBI ID takes the user directly to that entry, while hovering the cursor over the structure enlarges the structure. With quick search there is also the option to just search within the 3-star entities.

[Note that the quick text search does not search within the ontology or the citation fields.]

8.2 Advanced Search

The chemical structure search algorithm OrChem allows identical, substructure and similarity searching to be performed on an Oracle 11g database. It allows the user to search on groups and residues, as well as on complete molecular entities. OrChem works in combination with the Ketcher applet, a chemical structure editor for 2D chemical structures, and converts chemical structures into fingerprints, each fingerprint representing the occurrence of a particular structural feature. It is important to remember that fingerprints have limitations: while they can indicate that a particular structure feature is definitely absent, they do not necessarily indicate that a particular structural feature is definitely present.

Entity search

Draw the chemical structure using the Ketcher applet and select 'find this entity' using the radio button on the right-hand side. If present in the database it will bring you straight to the specific ChEBI page for the entity.

Substructure Search

Fingerprints are used to eliminate candidates for further examination in substructure searching. For molecule A to be a substructure of molecule B then all parts set in the fingerprint of molecule A should be present in molecule B. Once this initial screening is performed, the potential substructure candidates are subjected to a more rigorous inspection to determine whether molecule A is a substructure of molecule B.

To perform a substructure search in ChEBI draw your chemical structure using the Ketcher applet. Then select the 'find compounds that contain the substructure' option using the radio button on the right-hand side and click 'Search'. It is possible to search the entire ChEBI database or to select only the three stars entries (using the radio buttons above the search button). If the substructure is found within the database the results will be displayed in a table showing the structure, the ChEBI ID and Name and the status (2- or 3-star) of each entry.

Similarity Search

Similarity searching is performed by calculating the Tanimoto coefficient for each structure within the database against the query structure. The Tanimoto coefficient calculates how many structural features two chemical structures have in common based on the fingerprint described above. A Tanimoto score of 1.0 indicates that the two structures are very similar. However, as the fingerprints are calculated on a chemical structure path depth of eight it means that many structures will have similar fingerprints and very high similarity scores even though they might not bear any obvious resemblance to the original structure.

To obtain a similarity search select the 'find compounds which resemble this structure' again with the option to search the entire ChEBI database or just the three stars entries.

Advanced text search

The text search facility of the Advanced Search allows users to search all the data or to filter a search by category (see below). Mass and charge can be searched within ranges: for example, one can search for all entities with a mass of between 150 and 300 atomic mass units. Furthermore, searches can be filtered by database: for example, one can search for entities used in the NMRShiftDB or PubChem databases.

As in the Quick search, the asterisk (*) is provided as the wildcard character. You can place wildcards in any of the search options and in any of the search combinations, making this character very valuable in terms of searching.

All the above searches can be combined by using the Boolean logic operators AND, OR and BUT NOT, and there are options on the Results page for exporting the search results in either MDL SD file, tab delimited or XML format.

As mentioned above, users can also search by category. This option allows searches to be narrowed down by selecting from the categories provided. For instance All Names would include a search for ChEBI Names, IUPAC Names and Synonyms while searching Registry Numbers will search for database (e.g. CAS, Reaxys, HMDB) numbers.

8.3 OntoQuery

In addition to the Quick and Advanced Search options, it is also possible to conduct a search based on ontology classes. The specific search tool, OntoQuery, is located by clicking the 'Tools' tab on the ChEBI website.

The OntoQuery is an online OWL query tool designed to improve the experience of retrieving data from ChEBI while enhancing the power of the existing advanced tool search. It should be noted that this is a strict ontology search and thus any information outside the ontology will not be searchable using this tool (e.g. the structures). For more information visit the OntoQuery help page on the ChEBI website.

9. Keeping up to date with ChEBI

Please follow us on Twitter to keep up to date with all the new entities and general news from the ChEBI database.

9.1 Twitter

At the end of each month a tweet is sent out announcing the monthly release of the ChEBI database and contains the up-to-date number of fully annotated entries together with a link to the Entity of the Month. The twitter account also includes any news updates related to the ChEBI database. To follow ChEBI on twitter: https://twitter.com/chebit

For any queries please contact us via: https://github.com/ebi-chebi/ChEBI/issues or for urgent matter contact the ChEBI team by e-mail to: chebi-help@ebi.ac.uk.

9.2. Train online

An online quick tour (0.5 h) and a more detailed tour (1 h) covering much of the material in this manual can be found on the EBI Train Online site here.