

Database of experimental data

Features overview

In brief, the distinguishing features of the OCHEM database are as follows.

- The wiki principle: most of the data can be accessed, introduced and modified by users
- Different access levels: guests, registered users, verified users, administrators
- Tracking of all the changes
- Obligatory indications of the source of the data
- Possibility to indicate conditions of the experiment, which can be later used for QSAR modeling
- Search by substructure, molecule names, by publication where the measurements were referenced, by conditions of experiments, etc.
- Control of duplicated records
- Batch upload and batch modification of large amounts of data
- Different units of measurements and utilities to interconvert between units
- Organizing the records in re-usable sets ("baskets")

Structure overview

The database contains experimentally measured biological and physicochemical properties of small molecules together with the conditions under which the experiments have been conducted and references to the sources where the data were published.

The structure of the database is shown schematically in the following figure:

The *experimental measurements* are the central entities of the database. They combine all the information related to the experiment, in particular the result of the measurement, which can be either numeric or qualitative depending on the measured property. The central system component, where the experimental measurements can be introduced, searched and manipulated, is the [compound property browser](#).

An *experimental measurement* includes information about the *property* that was measured and the associated *chemical compound*. The compounds and the properties can be marked with particular keywords, also known as *tags*, that allow convenient filtering and grouping of the data.

For every measurement stored in the OCHEM, it is obligatory to specify the *source of the data*. The source is usually a publication in a scientific journal or a book. The strict policy of OCHEM is to accept only those experimental records, that have their source of information specified. This improves the quality of the data and allows it to be verified by checking the original publication. Although a user can also introduce an unpublished article and link the data to it, records from such sources should be treated with caution.

Every numeric property has a corresponding category of *units*, for example, the category of units for *Inhibition Concentration 50%* (IC50) is "Concentration". By default the OCHEM database keeps experimental endpoints in the original format (i.e., in *units* as reported in the publication). For this purpose all units are grouped into strictly defined unit categories, for example Kelvin, Celsius and Fahrenheit degrees belong to the "Temperature" category. For the purpose of compatibility and for modeling of the combined sets from different publications, the system provides on the fly conversion between different units.

An important feature of our database, which is also unique among other chemical databases, is the possibility to store the *conditions of experiments*. This information is crucial for modeling: in many cases, the result of an experimental measurement is senseless without knowing the conditions under which the experiment has been

conducted. For example, it does not make sense to specify the boiling point for a compound without specifying the air pressure. Such conditions should be introduced as *obligatory* conditions, i.e., a new record will be rejected by the system if there is no information about these conditions provided. Conditional values stored in the database can be numerical (with units of measurement), qualitative or descriptive (textual). Moreover, in the “conditions” section it is possible to note additional information related to the experiment, even if it is not a “condition” in the classical sense. Examples of such additional information are assay descriptions, a target of the ligand (the receptor) or species on which the biological activity has been tested. For simplicity, we further universally refer to all this information as “conditions”.