ChEMBL: Quick tour

Anne Hersey [1]

- Chemical biology
- Beginner
- 0.5 hour

This quick tour provides a brief introduction to ChEMBL, the EMBL-EBI's chemogenomics resource. For a more detailed walkthrough of ChEMBL, have a look at our ChEMBL: Exploring bioactive drug-like molecules [2] tutorial.

Learning objectives:

- Describe ChEMBL is and how it can help you to understand the interactions between drugs or drug-like molecules and their targets
- Recall where to find out more about ChEMBL

What is ChEMBL?

The human genome sequence provided a complete molecular ‘parts list’ for researchers interested in improving human health. A key task now is to catalogue how the gene products interact with drugs and drug-like molecules.

ChEMBL [3] is a ‘chemogenomic’ database that brings together chemical, bioactivity and genomic data to aid the translation of genomic information into effective new drugs. The resource is therefore of interest to drug discovery researchers in large pharmaceutical companies, as well as small companies and academic institutions.

What data is included in ChEMBL?

ChEMBL [4] is a manually curated, freely available database of bioactive molecules with drug-like properties. It brings together chemical, bioactivity and genomic data to aid the translation of genomic information into effective new drugs.

The database (Figure 1) is unique because of its focus on all aspects of drug discovery and its size, containing information on more than 1.8 million compounds and over 15 million records of their effects on biological systems.

ChEMBL includes information about how small molecules [5] interact with their protein targets, how these compounds affect cells and whole organisms, and information on absorption, distribution, metabolism [6], excretion and toxicity (ADMET [7]). ChEMBL holds two-dimensional structures, calculated molecular properties (e.g. logP [8], molecular weight, Lipinski ‘Rule of Five’ parameters) and bioactivity data (such as binding constants and pharmacology). The bioactivity data is tagged to show links between molecular targets and published assays.
The data is primarily manually extracted and curated structure-activity relationship (SAR) data from the primary medicinal chemistry and pharmacology literature. Additionally, the ChEMBL database contains data deposited by researchers and data extracted from other data sources.

**Figure 1** Data included in ChEMBL.

Lastly data on the clinical progress of compounds has being integrated into ChEMBL (Figure 2). Specifically we have a highly curated set of "drug" molecules that includes marketed compounds and compounds that are or have previously been in clinical development. These compounds are annotated with their known therapeutic targets and their therapeutic indications.
Figure 2 Information on drug molecules included in ChEMBL.

How is ChEMBL data curated?

We identify scientific facts in a journal article and then extract [9] the information and add it to the ChEMBL [3] database in a structured format (Figure 3).

For example an article might have structure activity relationship (SAR) data for a series of related compounds binding to a specific protein in the form of a number of IC50 [10] values. In this case we will draw the chemical structures as molfiles or smiles [11]. We will identify the Uniprot [12] accession [13] number for the protein and the organism it is from and extract a short description of the experiment (or assay [14] as we call it) that the IC50 values were measured in. This information is then added to the ChEMBL database. You can then search it for specific target by name or sequence or do substructure or similarity searches to find data on specific chemotypes.

If an article has data on measurements made in cell based assays, in vivo pharmacology assays or pharamcokinetic studies for compounds we will also extract that and add it to the ChEMBL database.
Figure 3 Inclusion of information into the ChEMBL database.

We then perform additional curation [15] and annotation [16] of the data so that we can provide a high quality database for you to use in your research. This curation process is shown in the figure below.
Figure 4 Curation of data included in ChEMBL.

Accessing ChEMBL data

You can get access to the data in the ChEMBL [3] database in a number of different ways (Figure 5):

1. Using our website [4]
2. Via our Web Services [17]
3. Downloading the whole database [18] in a number of different formats
4. Via the EBI RDF Platform [19]

The following pages provide more information on these methods of access.
**Searching using the web interface**

The [ChEMBL](https://www.ebi.ac.uk/chembl) web interface provides a flexible and easy way to access ChEMBL’s core bioactivity data. Searching is via an encrypted and secure protocol.

You can perform searches of the ChEMBL data in many ways:

- Explore ChEMBL by viewing complete entity sets
- Explore bioactivity data on protein target families
- A flexible text matching search for ChEMBL entities such as compounds, targets, assays, documents, cells, tissues
- Compound substructure and similarity searches
- Search by protein sequence similarity ([BLAST](https://blast.ncbi.nlm.nih.gov/blast.cgi))

Explore ChEMBL's web interface by clicking on the [web interface](https://www.ebi.ac.uk/chembl) below:
Viewing results from the web interface

The results can be viewed as:

- Tables of entity results which can be filtered
- Entity report cards (compound, target etc)
- Tables of bioactivity data which link numeric experimental results to compounds, assays and targets

Data can also be downloaded in either CSV for TSV [21] format using the download results tool on the search page.

Click the below to learn how to navigate the two different search result pages.

ChEMBL target search results

ChEMBL compound search results

Programmatic access via web services

The CheMBL [3] web services [22] provide a flexible way of exploring the ChEMBL data without you having to download the whole database yourself.

If you want an introduction to ChEMBL webservices, please watch the video below. For detailed information on using the webservices, please refer to the documentation [17].

Your feedback

Please tell us what you thought about this course. Your feedback is invaluable and helps us to improve our courses and thus enhance your learning experience.

Get help and support on ChEMBL

Support and find out more

- For any queries regarding any of the CheMBl [3] databases, please contact our chembl-help [at] ebi.ac.uk (support team)
- For news from the ChEMBL team, visit the CheMBl blog [23]
- To keep up to date with ChEMBL news and data releases subscribe to the Chembl-announce mailing list [24]
References


Funding

The ChEMBL resources are made available due to funding from the organisations [30] listed here [30].

Contributors

Anne Hersey [1]
EMBL-EBI
Group Coordinator - Leach team: ChEMBL

Anne has been Acting Group Leader of the ChEMBL chemogenomics data resource, which brings together chemical and bioactivity data that can be used by scientists working in drug discovery. Before joining EMBL-EBI, Anne worked for many years in the pharmaceutical industry. She initially worked on physicochemical property measurement, then led a team developing models for the prediction of ADMET properties for use by drug discovery teams. Anne obtained a degree in chemistry and PhD from the University of Kent.

Anna Gaulton [31]
EMBL-EBI