ChEMBL: Quick tour

Louisa Bellis [1]

- Chemical biology
- Beginner
- 0.5 hour

This quick tour provides a brief introduction to ChEMBL, the 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:

- Basic understanding of ChEMBL and how it can help you to understand the interactions between drugs or drug-like molecules and their targets
- Know where to find out more about ChEMBL

What is ChEMBL?

Why do we need ChEMBL [3]?

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 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.

About ChEMBL

ChEMBL [4] is a publicly available database of drugs, drug-like small molecules [5] and their targets. The database is unique because of its focus on all aspects of drug discovery and its size, containing information on more than 1.4 million compounds and over 12 million records of their effects on biological systems.

The database includes information about how small molecules bind to their 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, with a set of varying confidence levels. Additional data on the clinical progress of compounds is being integrated into ChEMBL. The database holds manually extracted and curated structure–activity relationship (SAR) data from the primary medicinal chemistry and pharmacology literature.

ChEMBL data is also a core component of the SARfaris – integrated chemogenomics workbenches for drug discovery that focus on protein kinases (Kinase SARfari [9]) and G protein-coupled receptors
(GPCR SARfari [9]). They are central resources for target class knowledge, combining both biological and chemical information.

Protein kinases are often regulators of cell signalling and are therefore key candidates for drug discovery research. Kinase SARfari contains a reference alignment of each protein kinase family, three-dimensional structures, bound ligand [10] conformations and binding sites. GPCRs [11] are transmembrane receptors implicated in many diseases, including inflammation and neurotransmission. Both SARfaris contain SAR data and clinical candidate data extracted from ChEMBL.

**ChEMBL data**

**What data does ChEMBL contain?**

The ChEMBL [3] data is a combination of extracted data from literature and donated data sets from companies such as GSK and PubChem. The curated data is made up of targets, organisms, compounds and their associated bioactivities (Figure 3).

![ChEMBL resources](image)

Figure 3 ChEMBL resources.

**ChEMBLdb**
Figure 2 Different ways to search the data in ChEMBLdb:

[A] Browse target views

[B] Target search using protein sequence

[C] Search results; displays bioactivities, chemistry-related information, etc.

[D] Compound search view.

Neglected Tropical Disease archive
Neglected Tropical Disease (NTD) archive

A repository for open access primary screening and medicinal chemistry data directed at neglected diseases – endemic tropical diseases of the developing world.

<table>
<thead>
<tr>
<th>Assay</th>
<th>Min Value</th>
<th>Adjust Range</th>
<th>Max Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>TCMDC: % inhibition of P. falciparum strain 3D7 (at 2uM)</td>
<td>80</td>
<td></td>
<td>100</td>
</tr>
<tr>
<td>TCMDC: % inhibition of P. falciparum Dd2 (at 2uM)</td>
<td>0</td>
<td></td>
<td>100</td>
</tr>
<tr>
<td>TCMDC: pXC50 determination of P. falciparum 3D7 growth</td>
<td>5</td>
<td></td>
<td>9</td>
</tr>
<tr>
<td>TCMDC: % inhibition of human HepG2 cell line (at 10uM)</td>
<td>0</td>
<td></td>
<td>100</td>
</tr>
<tr>
<td>TCMDC: Inhibition Frequency Index (IFI)</td>
<td>0</td>
<td></td>
<td>44</td>
</tr>
</tbody>
</table>

Figure 3 Searching for compound activity in the Neglected Tropical Disease archive

Kinase SARfari
Figure 4 Resource for searching and analysing protein kinase structure-activity data

GPCR SARfari
Figure 5 Resource for searching and analysing G protein-coupled receptor (GPCR) data.

**Searching and getting data from ChEMBL**

**Web interface**

The [ChEMBL web page](http://www.ebi.ac.uk/training/online) provides flexible and easy access to ChEMBL’s core bioactivity data. You can perform searches using compound names, compound structures, target names, target sequences, bioactivity and target classes. Searching is via an encrypted and secure protocol.

**Target search**

You can browse all data relating to a particular protein target using the [protein family](http://www.ebi.ac.uk/training/online) classification tree, searching by keyword, or using the sequence similarity search option. You can filter the resulting data (for example by [IC50](http://www.ebi.ac.uk/training/online) or [Ki](http://www.ebi.ac.uk/training/online)), download bioactivity data and identify interesting molecules on the basis of substructures, physicochemical properties, potency, and [ligand](http://www.ebi.ac.uk/training/online) efficiency.

**Compound search**

You can search for data on compounds of interest by drawing a molecular structure or substructure, or by inputting keywords or compound IDs. The results will list all molecules containing that substructure or molecules similar to the input structure. You can then identify molecules with interesting properties (such as their target interactions, selectivity, [ADMET](http://www.ebi.ac.uk/training/online)) and the data can be
downloaded for further analysis.

**Analysis of key gene families**

The SARfaris (Kinase SARfari [9], GPCR SARfari [15]) enable in-depth searching and analysis of target class data. You can browse families, search for specific protein or chemical (sub)structure and view binding sites, make comparisons with other related sequences and retrieve 3D structural data.

**Neglected Tropical Disease (NTD) archive**

You can search data on neglected tropical diseases using the ChEMBL-NTD [16] interface. For instance, ChEMBL-NTD contains a data set of 13,500 compounds that are known to inhibit the growth of *Plasmodium falciparum* strain 3D7, which is one of the most deadly parasites that cause malaria in humans. This data set was donated by GlaxoSmithKline [17] (GSK).

**Data download**

You can download ChEMBL data for free in a number of standard formats from our FTP site [18].

**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 support team [19]
- For news from the ChEMBL team, visit the ChEMBL blog [20]
- To keep up to date with ChEMBL news and data releases subscribe to the Chembl-announce mailing list. [21]

**References**


**Funding**

The data, maintenance and support of ChEMBL are funded through a £4.7 million Strategic Award from the Wellcome Trust [26].
Contributors

John Overington [27]

Stratified Medical
Director of Bioinformatics

John Overington is the Director of Bioinformatics at Stratified Medical. John previously worked as Team Leader for the computational Chemical Biology team at the European Bioinformatics Institute in Cambridge, UK. He studied for a PhD in Crystallography at Birkbeck College, University of London in 1991, working of protein modelling and sequence template bioinformatic methods development. Since this time John has held a number of positions in large Pharma and Biotech sectors prior to joining the EBI in 2008.

Louisa Bellis [1]

EMBL-EBI
Chemical Content Curator

Louisa Bellis is the Chemical Content Curator for the ChEMBL group at the European Bioinformatics Institute. Louisa is also responsible for coordinating the outreach within the ChEMBL group. After completing a degree in chemistry with pharmaceutical and forensic Science, Louisa went on to study for a PhD in pharmacology at the University of Bradford. Her PhD was focused on the development of potential new drugs for the treatment of bladder cancer. Following this and prior to joining the EBI in 2009, Louisa worked for a not-for-profit company, training users on their drug toxicity and metabolism prediction software.

Source URL: http://www.ebi.ac.uk/training/online/course/chembl-quick-tour

Links