ChEMBL

Accessing big molecular data via the web interface and API

James Blackshaw & **Melissa Adasme**

ChEMBL Scientific Data Engineer

jblackshaw@ebi.ac.uk

Data Mining Scientist

adasme@ebi.ac.uk

https://www.ebi.ac.uk/chembl/

chembl-help@ebi.ac.uk

EMBL-EBI

Hybrid workshop on in silico Drug design, Jan 30, 2022

Training structure

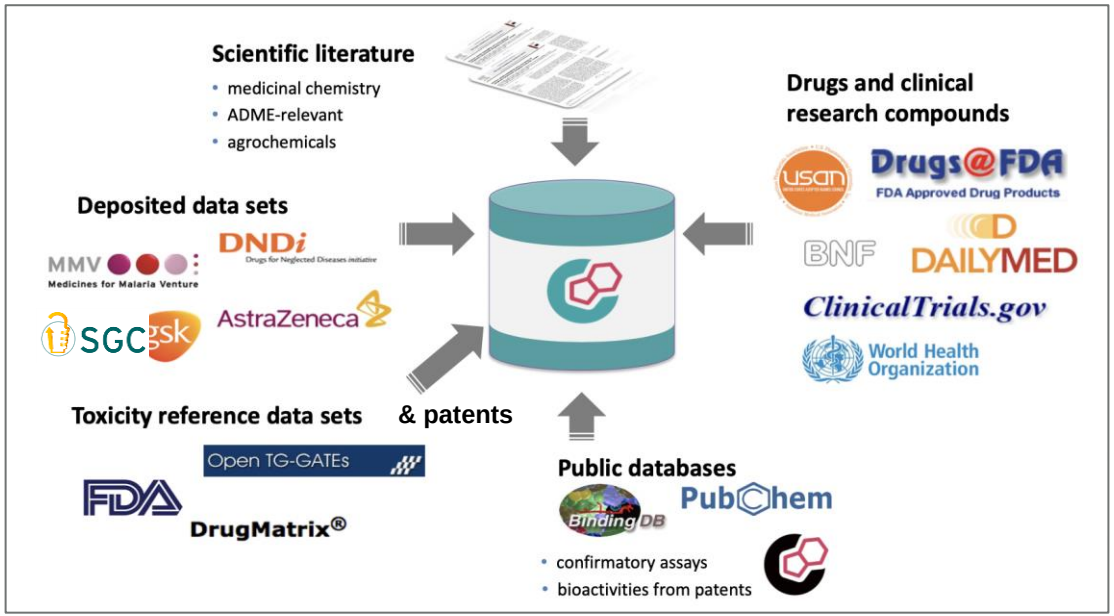
General introduction into ChEMBL (James Blackshaw) – ***25 min***

- What is ChEMBL and how is it structured?
- What data does ChEMBL contain?
- How are data extracted from scientific articles?
- How are the data in ChEMBL curated?
- How are drug data curated and annotated in ChEMBL?
- How can you use the ChEMBL GUI?

ChEMBL live tutorial (Melissa Adasme) – ***25 min***

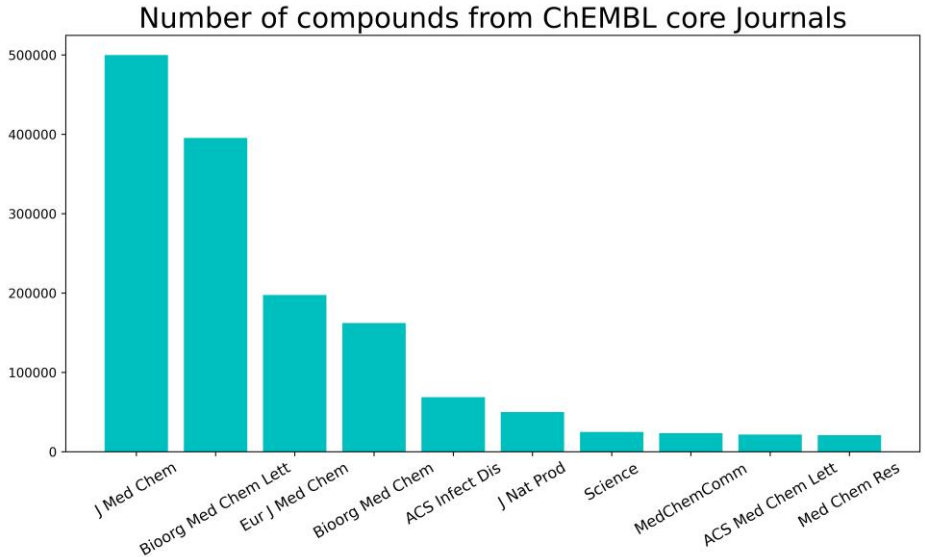
Q&A – 10 min

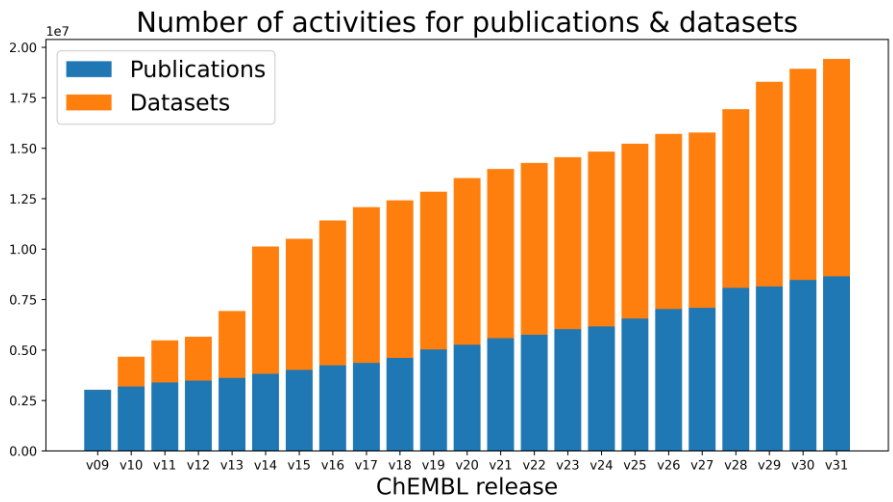
What is included in ChEMBL?



- "Open access data for drug discovery".
- Manually curated, high quality, FAIR.
- The core of the database is the primary MedChem literature & donated data sets.
- ChEMBL covers >40 years of research (starting 1980).

Distribution of sources





EMBL-EBI

ChEMBL 32 Summary (Releasing in mid Feb)

	**Datasets**	213
	**Publications**	86,364
	**Targets**	15,139
	**Compounds**	2,354,965
	**Activities**	20,038,828

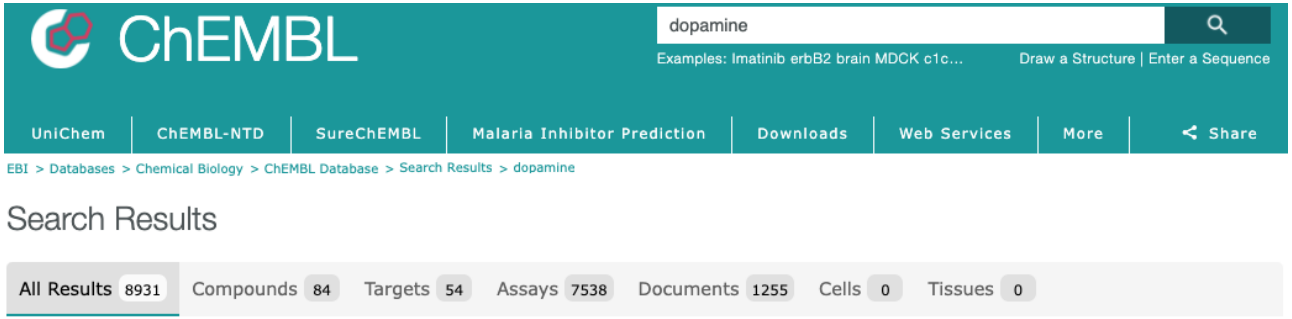
EMBL-EBI

Data statistics

Molecular properties	E.g. two-dimensional structures, calculated molecular properties (logP, molecular weight, Lipinski ‘Rule of Five’ parameters etc.)
Protein-target interactions	Information about how small molecules interact with their protein targets.
Cell and organism level data	Records on how these compounds affect cells and whole organisms. Information on absorption, distribution, metabolism, excretion and toxicity (ADMET) etc.
Bioactivity data	Binding constants and other activity results from deposited assays.
Curated drug and candidate set	A highly curated set of drugs and clinical candidates. These compounds are annotated with their known therapeutic targets and therapeutic indications.

ChEMBL entities

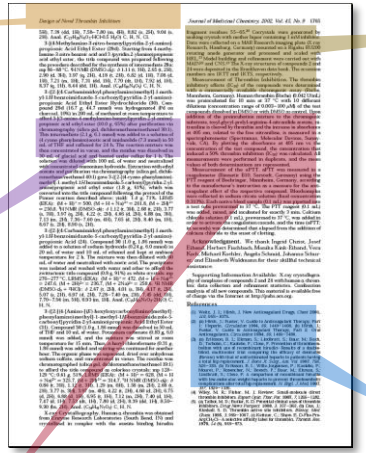
www.ebi.ac.uk/chembl



Compounds	Chemical structure and properties on which experimental data have been measured
Targets	Biological system in which bioactivity data is measured (e.g. protein, cell, organism)
Assays	The description of the specific experiment
Documents	Where the data are reported (e.g. scientific paper)
Cells	If applicable, the cell in which the bioactivity was measured
Tissues	If applicable, the tissue in which the bioactivity was measured

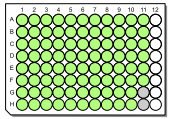
Literature data extraction

Reference

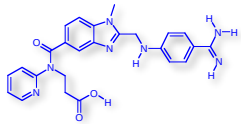


Organism

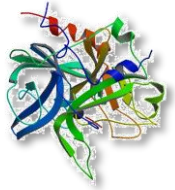
Assay



Compound



Target

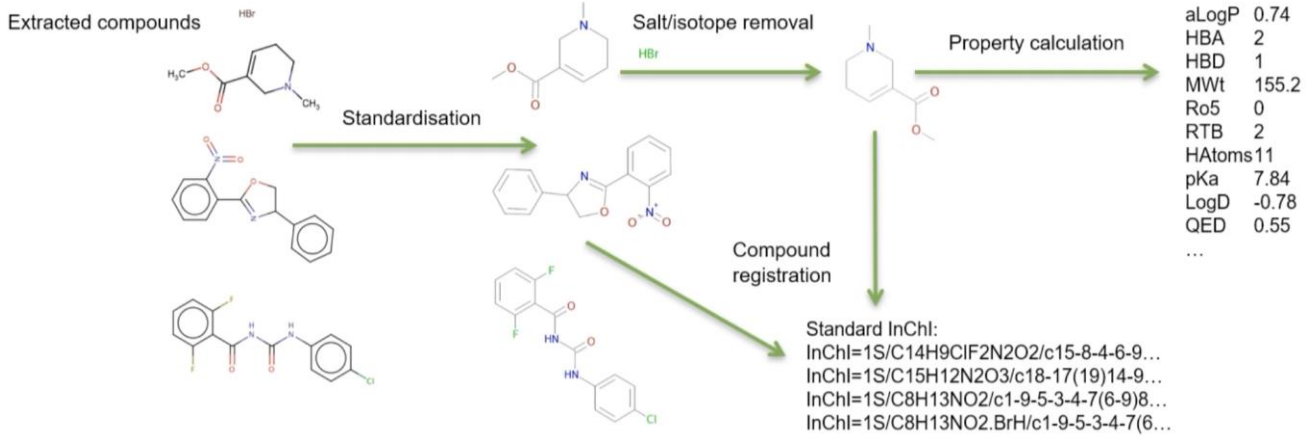


EMBL-EBI

Literature data extraction

Manual extraction from the literature	Data are extracted from full text articles
Comprehensive	All quantitative bioactivity measurements are extracted
Multiple data types	These include: Target binding data Functional assays In-vivo efficacy data In-vitro and in-vivo ADME/pharmacokinetics Toxicology measurements

Chemistry curation



We standardise the structures.	Automated error checking, standardization, and salt stripping.
We calculate additional data.	Automated calculation step for compound properties.
ChEMBL curators correct any remaining issues.	Manual curation of invalid structures, name-structure mismatches, missing stereochemistry etc.

Bento, A.P., Hersey, A., Félix, E. *et al.* An open source chemical structure curation pipeline using RDKit. *J Cheminform* **12**, 51 (2020)

Activity curation

	ACTIVITY_ID	PCHEMBL_VALUE	STANDARD_RELATION	STANDARD_VALUE	STANDARD_UNITS	RELATION	VALUE	UNITS	DATA_VALIDITY_COMMENT
1	33501	7.57	=	27	nM	=	27	nM	(null)
2	36800	(null)	=	19.8	ug.mL-1	=	19.8	ug ml-1	(null)
3	859055	4.8	=	16000	nM	=	16	uM	(null)
4	1273293	(null)	<=	100000	nM	<=	0.1	mM	(null)
5	1699014	4.07	=	86000	nM	=	0.086	mM	(null)

Standardised data	Depositor input

Mostly against protein targets	A large proportion of data is **dose-response measurements against protein targets** E.g. IC50 – half maximal inhibitory concentration
pChEMBL values allow approximate comparisons	pChEMBL is calculated when: • Relation is ‘=‘ • Units standardised to ‘nM’

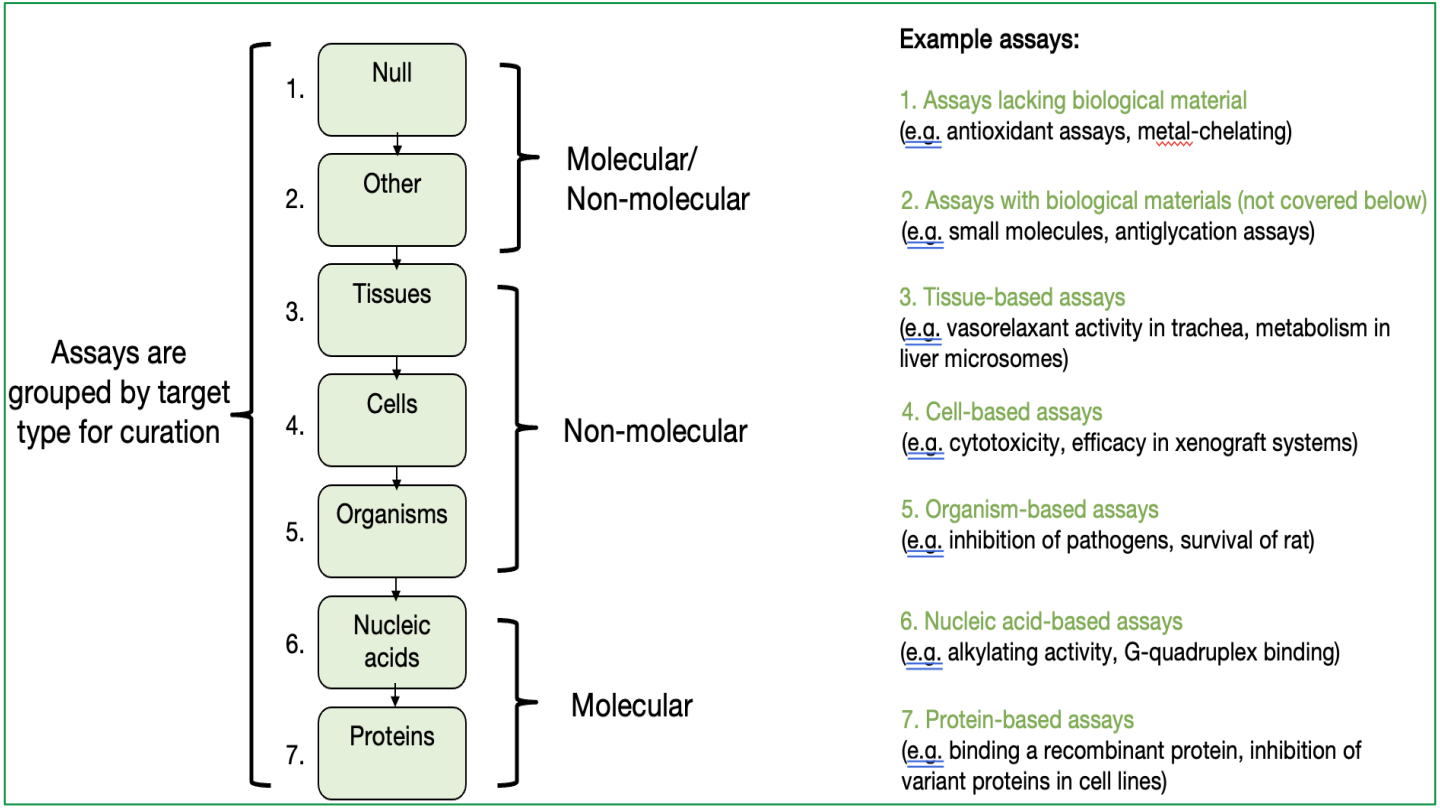
Assay Types

Assay type	Assay description	Example
F	Functional	Biological effect of a compound e.g., % cell death in a cell line, rat weight
A	ADME	ADME data e.g., t1/2, oral bioavailability
T	Toxicity	Toxicity of a compound e.g., cytotoxicity
B	Binding	Binding of a compound to a molecular target e.g., Ki, IC50, Kd
P	Physicochemical	Physiochemical properties of the compounds in the absence of biological material e.g., chemical stability, solubility
U	Unassigned	Cannot be classified into one of the above categories e.g., ratio of binding vs. efficacy

Bioassay curation

• Standardise activity types	e.g., IC-50, mean IC50, IC(50) -> **IC50**
• Antilog values	e.g., pKi -> **Ki**, -log IC50 -> **IC50**
• Standardise values/units	e.g., uM, 10'-6 mol/L, pmol/mL -> **nM**
• Flag potentially incorrect data	e.g., out of range, non-standard units
• Calculate pChEMBL values	$-\log_{10}$ (molar IC50, XC50, EC50, AC50, Ki, Kd or Potency)
• Flag potential duplicates	e.g., same compound, target, type & value
• Manual curation to fix erroneous data	

Bioassay curation



Bioassay curation

Curation Method

- **Target types** and **assay types** are both extracted from assay description
- Correct and standardize names for tissues, cell lines etc by using controlled vocabularies (e.g. UBERON, Cellosaurus)
- Assay formats are mapped to BAO ontology

Common curation tasks for the Assay type:

1. Curation of assay type ‘U’
2. Re-mapping of toxicity assays from ‘A’ to ‘T’
3. Re-mapping of GPCR assays measuring secondary messengers from ‘B’ to ‘F’

ChEMBL target types

Non-molecular

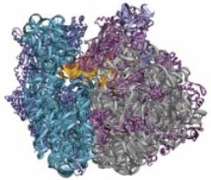
Molecular

Nucleic acid

DNA

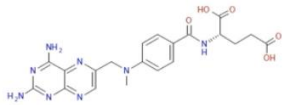
Protein

Protein Nucleic-Acid Complex



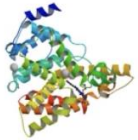
Ribosome

Small Molecule



Methotrexate

Single Protein



PDE5

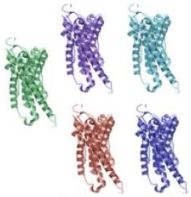
Protein Complex

Integrin alpha-4/beta-7

Protein Complex Group

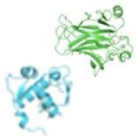
GABA-A receptors

Protein Family



Muscarinic receptors

Protein-Protein Interaction



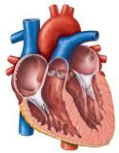
p53-MDM2

ChEMBL 31: ~8M bioactivities on protein targets; ~12M bioactivities on non-protein targets

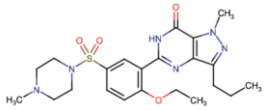
EMBL-EBI

Drugs and clinical candidates

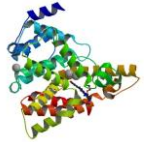
- Clinical Trials.gov
- DailyMed
- USAN applications
- FDA Orangebook
- ATC classification
- BNF
- INN applications
- FDA new drug approvals
- Withdrawn drugs
- Pharma pipelines



Disease annotation



Name & structure mapping



Mechanism annotation

Drug property annotation

ChEMBL

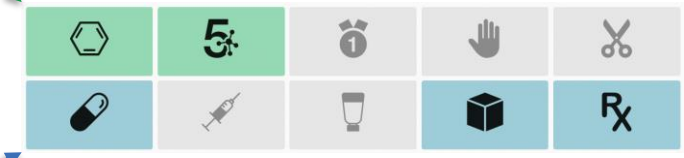
Open Targets

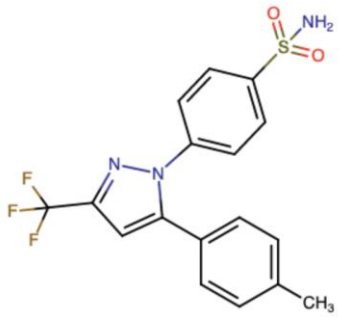
EMBL-EBI

Drug properties

Molecule-based **properties are assigned to all clinical-stage compounds (green icons)**	• Molecule type (small molecule, antibody, protein, oligonucleotide, oligosaccharide etc. • Rule-of-five compliant • First in class • Chiral/racemic/achiral • Prodrug
Product-based **properties are assigned to approved drugs (blue icons)**	• Oral/parenteral/topical administration • Black box warning/withdrawn • Availability type (prescription only, over the counter, discontinued, withdrawn)

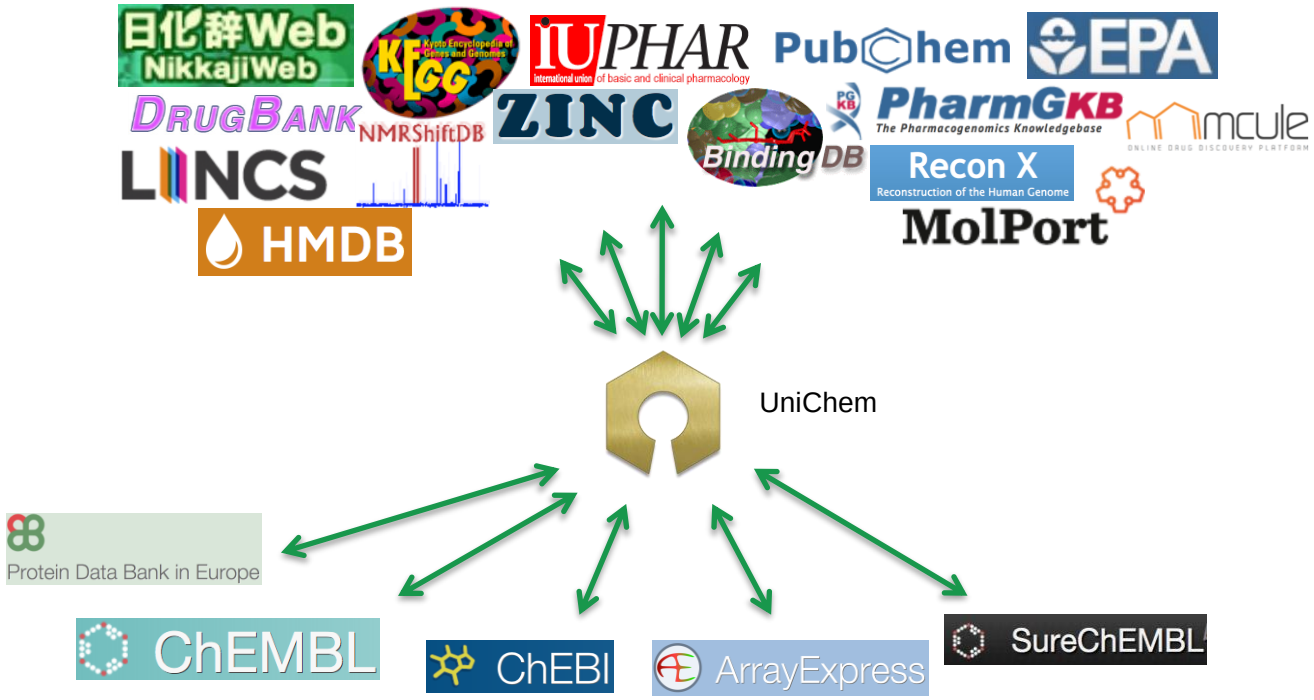
Celecoxib





EMBL-EBI

UniChem generates ChEMBL linkouts



- UniChem maps database identifiers and structures for 170 million compounds from 40 different source databases

Cross-references from UniChem

Results rendered in ChEMBL interface


```
---
-
  aux_for_url: 0
  base_id_url: https://www.ebi.ac.uk/chembldb/compound/inspect/
  base_id_url_available: 1
  description: A database of bioactive drug-like small molecules and bioactivities abstracted fro
  name: chembl
  name_label: ChEMBL
  name_long: ChEMBL
  src_compound_id:
    - CHEMBL490
  src_id: 1
  src_url: https://www.ebi.ac.uk/chembl/
-
  aux_for_url: 0
  base_id_url: http://www.drugbank.ca/drugs/
  base_id_url_available: 1
  description: A database that combines drug (i.e. chemical, pharmacological and pharmaceutical)
  name: drugbank
  name_label: DrugBank
  name_long: DrugBank
  src_compound_id:
    - DB00715
  src_id: 2
  src_url: http://drugbank.ca/
-
  aux_for_url: 0
  base_id_url: http://www.ebi.ac.uk/pdbe-srv/pdbechem/chemicalCompound/show/
  base_id_url_available: 1
  description: The European resource for the collection, organisation and dissemination of data o
  name: pdb
  name_label: PDBe
  name_long: PDBe (Protein Data Bank Europe)
  src_compound_id:
    - 8PR
  src_id: 3
  src_url: http://www.ebi.ac.uk/pdbe/
-
  aux_for_url: 0
  base_id_url: http://www.guidetopharmacology.org/GRAC/LigandDisplayForward?ligandId=
  base_id_url_available: 1
  description: The IUPHAR (International Union of Basic and Clinical Pharmacology)/BPS (British P
  name: gtopdb
  name_label: Guide to Pharmacology
  name_long: Guide to Pharmacology
  src_compound_id:
    - 4790
  src_id: 4
  src_url: http://www.guidetopharmacology.org
-
```


Programmatic output e.g., JSON, XML

UniChem Cross References

New View the UniChem Connectivity matches for CHEMBL490

ACToR	110429-35-1 61869-08-7
Atlas	paroxetine
BindingDB	50331515 22416
ChEBI	7936
DrugBank	DB00715
DrugCentral	2068
eMolecules	1935652
EPA CompTox Dashboard	DTXSID3023425
FDA SRS	41VRH5220H
Guide to Pharmacology	4790
Human Metabolome Database	HMDB14853
IBM Patent System	BA9CB78987006F7DD38940BBB3C63F01
KEGG Ligand	C07415
LINCS	LSM-2843
Nikkaji	J259.859K
PDBe	8PR
PharmGKB	PA450801
PubChem	43815
PubChem: Drugs of the Future	12013591
PubChem: Thomson Pharma	14924270 14850948
SureChEMBL	SCHEMBL27799
ZINC	ZINC000000527386

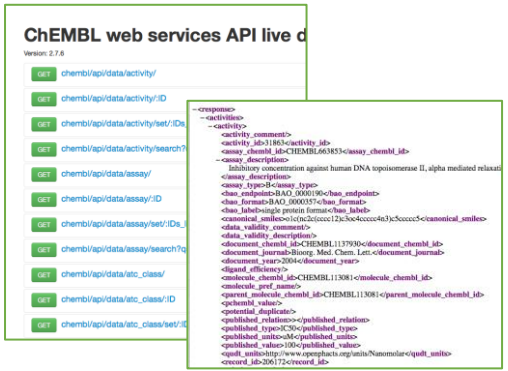
Link from ChEMBL to SureChEMBL

EMBL-EBI

Access to ChEMBL data

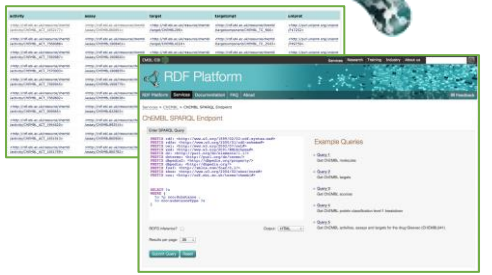
Website

Web Services



Downloads

Semantic Web



www.ebi.ac.uk/chembl/

EMBL-EBI

How should I access ChEMBL?

Use API if:	Need a repeatable query, or embed in a script? Don't want to download full database Want easy access to data via url weblink?
Use Web Interface if:	Only need to access small amounts of data, e.g. for one compound or one target For initial or exploratory investigation
Download Database if:	Need to run multiple queries with much input data? Have complex, bespoke queries across multiple tables? (Construct an sql query) Need data from a previous release? (API is for the current release only)

Web Interface Structure

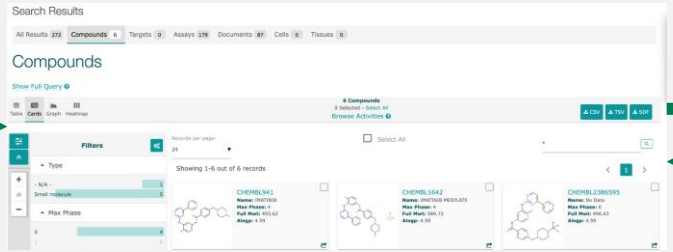
Landing Page

Search by:

- Free Text.
- Chemical Structure.
- Identifiers.
- BLAST.

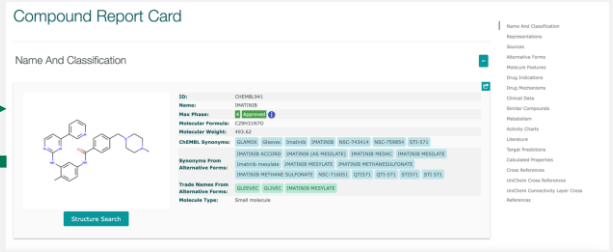
Explore all items in compounds, assays, etc.

“Browse Items” Pages



- Explore datasets.
- See data distribution.
- Apply filters.
- See relationships among entities.

Report Card Pages

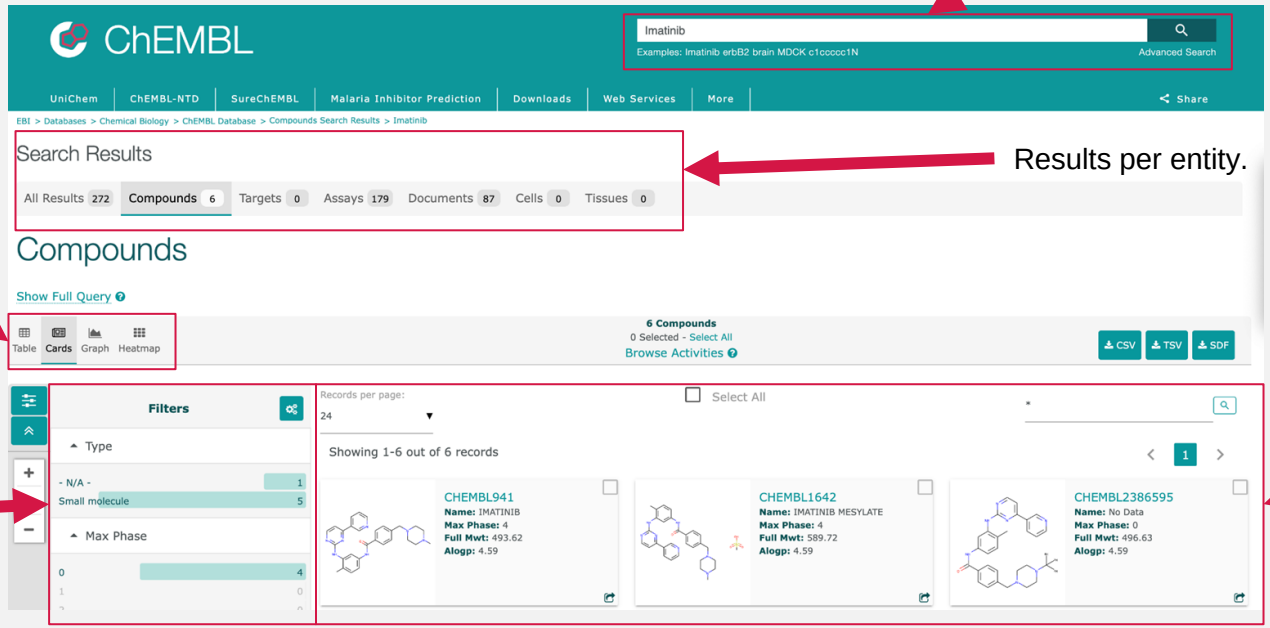


- See the details of one item.
- Start a structure search from the item (compounds).
- See the relationship with other entities.

EMBL-EBI

Free Text Search

Enter a text term on the search bar



Results per entity.

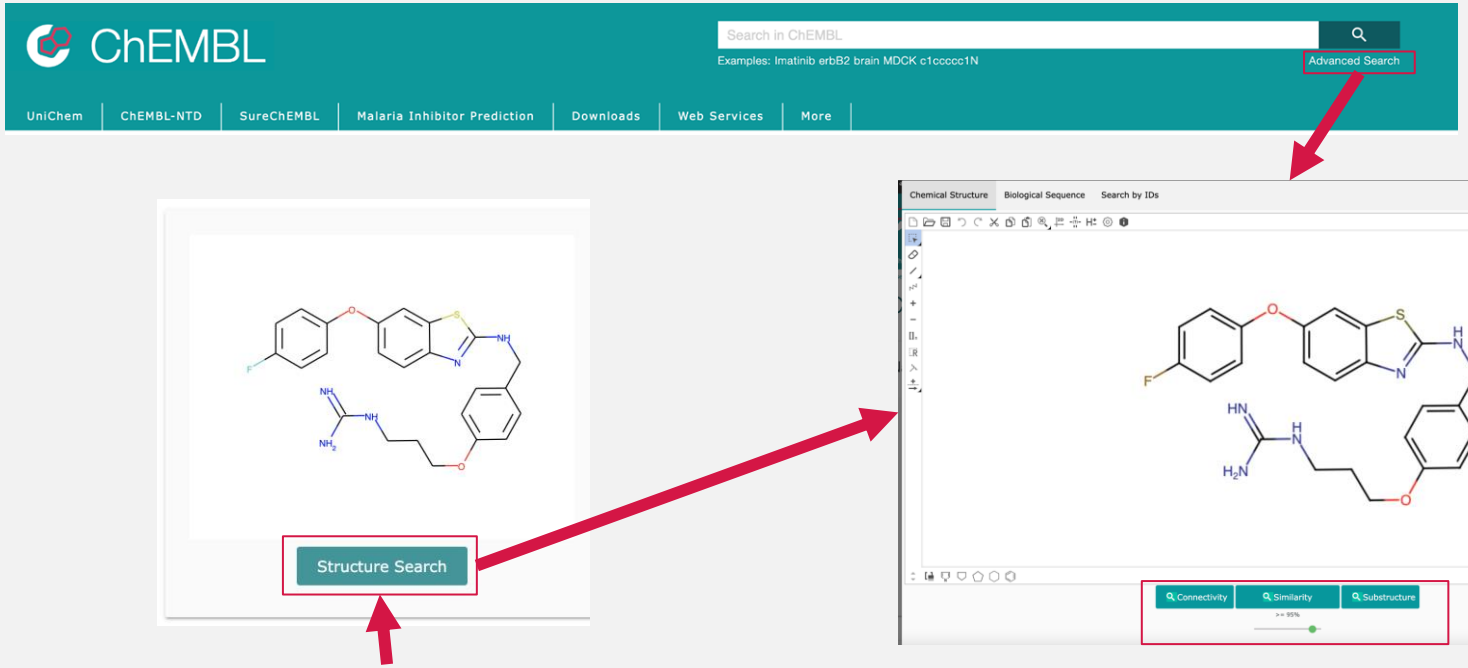
View data in different presentations.

See distribution of data and apply filters.

Specific Items

EMBL-EBI

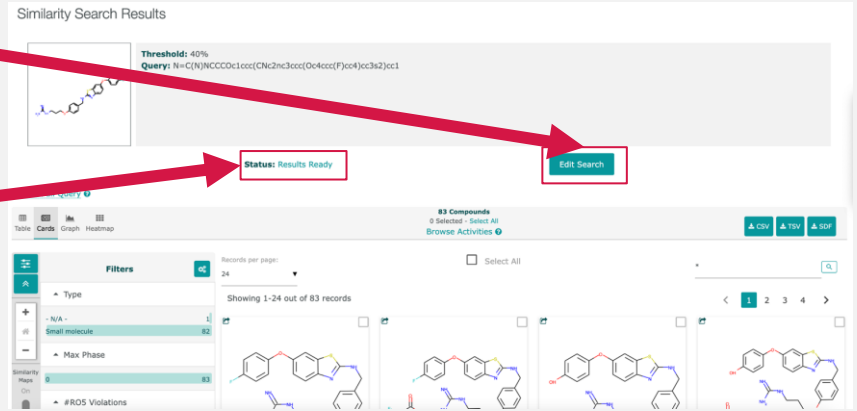
Structure Search



https://www.ebi.ac.uk/chembl/compound_report_card/CHEMBL4301448/

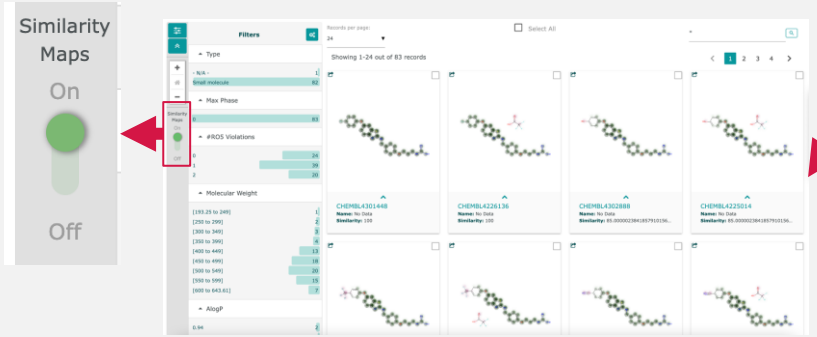
Structure Search - Result

Edit search and start a new one.

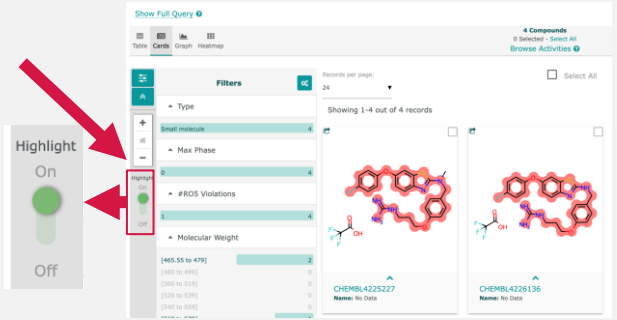


Status of the search.

- Similarity and connectivity are immediate.
- Substructure is run by a job in LSF.

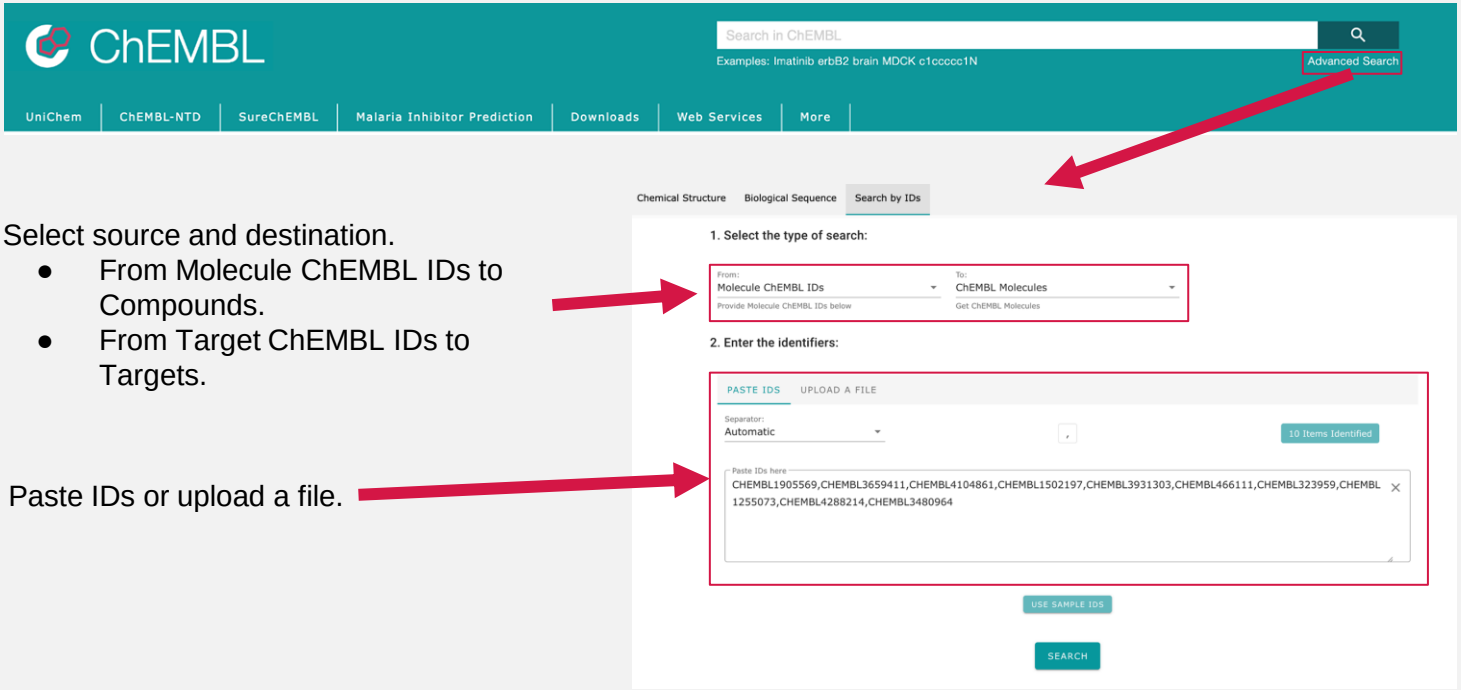


Structure highlights and similarity maps.



EMBL-EBI

Search by IDs



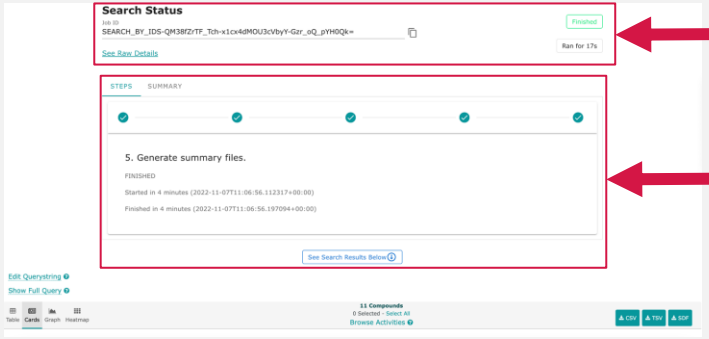
Select source and destination.

- From Molecule ChEMBL IDs to Compounds.
- From Target ChEMBL IDs to Targets.

Paste IDs or upload a file.

EMBL-EBI

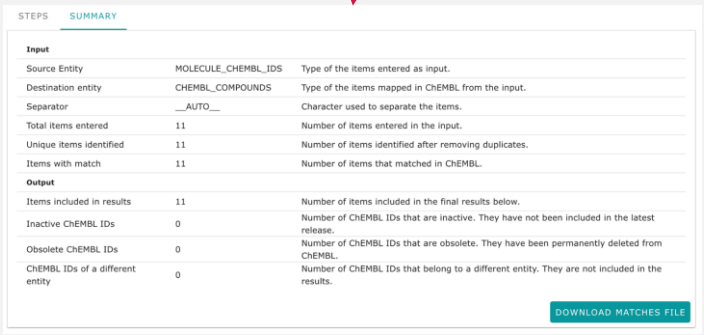
Search by IDs

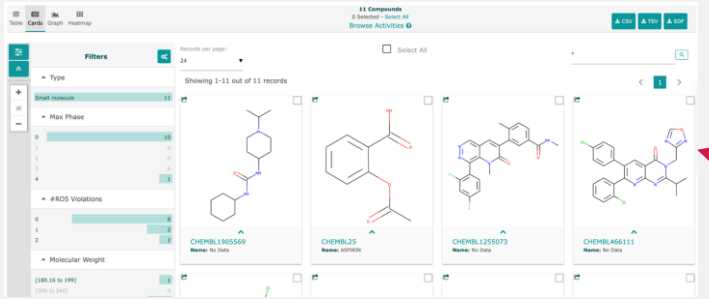


See the status of the search. It is run by a job in LSF.

Details of the process.

Details of the results.

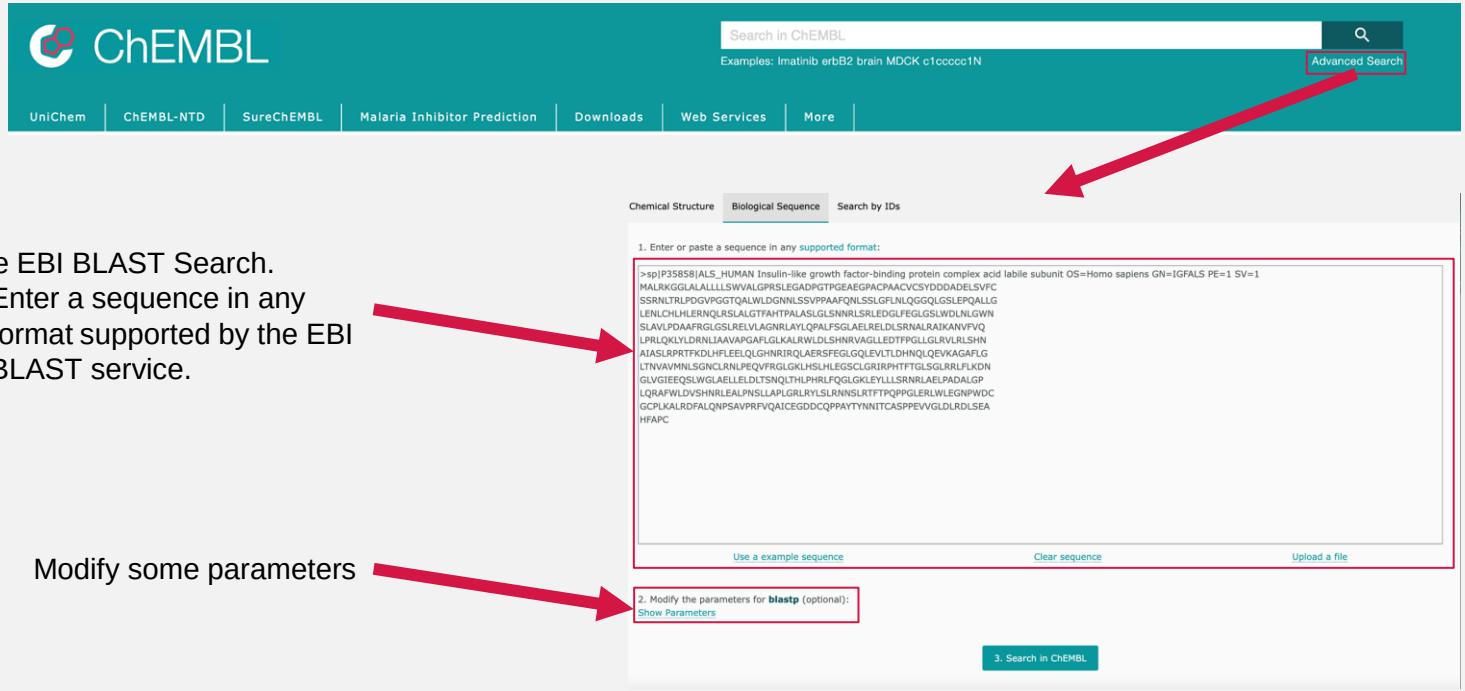




Explore all results as usual.

EMBL-EBI

BLAST Search



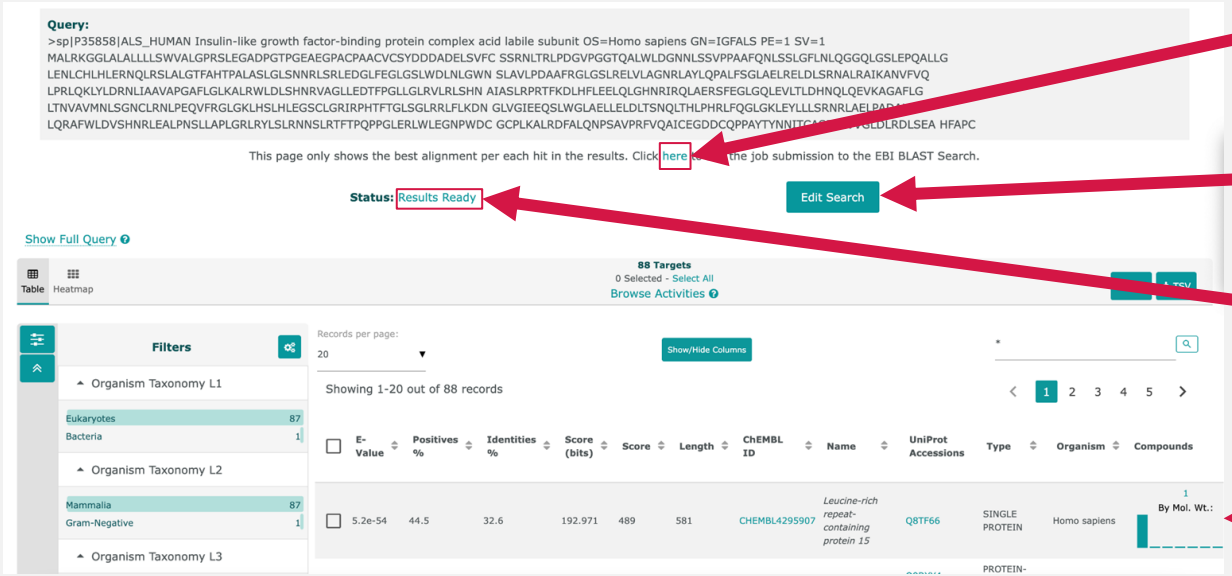
Uses the EBI BLAST Search.

- Enter a sequence in any format supported by the EBI BLAST service.

Modify some parameters

EMBL-EBI

BLAST Search



Go to EBI - BLAST search results

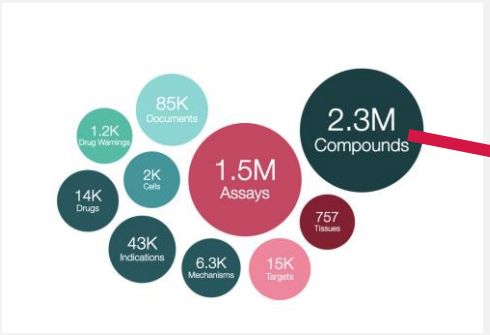
Edit Search

See progress of Search

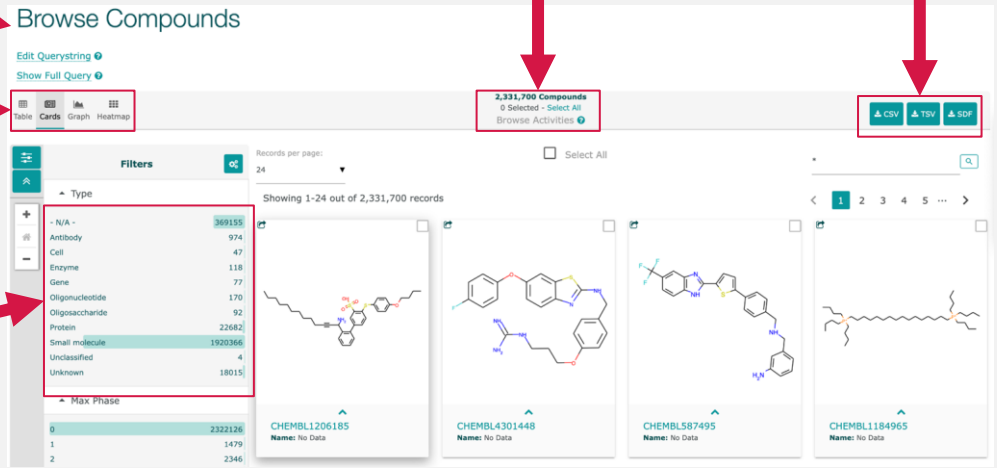
Results in Targets

EMBL-EBI

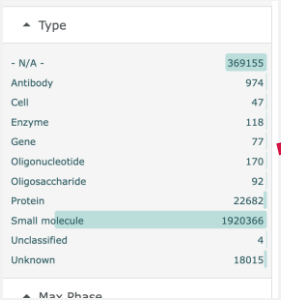
‘Browse’ Pages



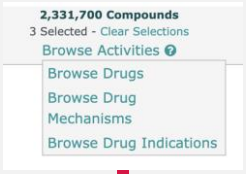
View data in different ways.



See data distribution and filter by several properties.



Go to related entities from a selection of items.

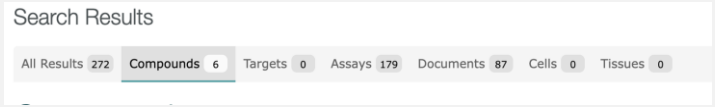


Download data

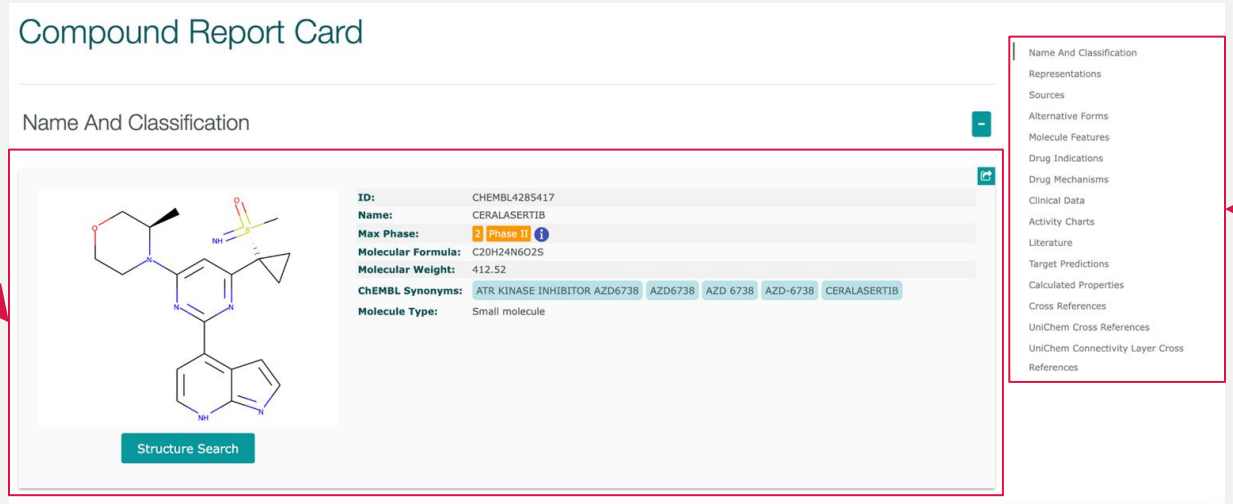
EMBL-EBI

Report Card Pages

Report Cards for Compounds, Targets, Assays, Documents, Cell Lines and Tissues.



Basic Information of the item.



Summary of the sections of the page.

EMBL-EBI

Formatting API calls

URL Pattern: stem resource field filter type value

https://www.ebi.ac.uk/chembl/api/data/molecule?pref_name__icontains=acetaminophen

single underscore!

double underscore!

Filters:

- contains (icontains)
- startswith (istartswith)
- regex (iregex)
- isnull …

(case insensitive)

Formatting:

Ordering -

insert '-' here for descending

https://www.ebi.ac.uk/chembl/api/data/target?order_by=-pref_name

Formats -

- XML(default), JSON, YAML, (SVG, SDF)

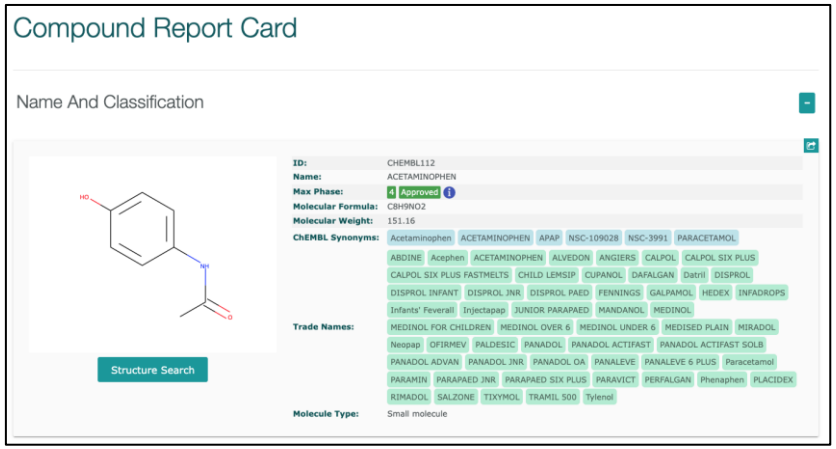
https://www.ebi.ac.uk/chembl/api/data/target.json

https://www.ebi.ac.uk/chembl/api/data/target?format=json

Full list at: https://chembl.gitbook.io/chembl-interface-documentation/web-services/chembl-data-web-services

Accessing data: website versus programmatically

ChEMBL web interface



- Search for acetaminophen
- View the interactive compound report card

ChEMBL web services

https://www.ebi.ac.uk/chembl/api/data/molecule/CHEMBL112

Stem | Resource | Entity

```
−<molecule>
  −<atc_classifications>
      <level5>N02BE51</level5>
      <level5>N02BE01</level5>
      <level5>N02BE71</level5>
   </atc_classifications>
   <availability_type>2</availability_type>
   <biotherapeutic/>
   <black_box_warning>1</black_box_warning>
   <chebi_par_id>46195</chebi_par_id>
   <chirality>2</chirality>
  −<cross_references>
    −<molecule>
        <xref_id>acetaminophen</xref_id>
        <xref_name>acetaminophen</xref_name>
        <xref_src>DailyMed</xref_src>
```

- Direct URL to access compound data
- Computer-readable output

EMBL-EBI

Best practice ideas

First steps:	Use the web interface for initial exploration, then progress onto the API
Get started with API:	Look at some example URLs: e.g. return all compounds: https://www.ebi.ac.uk/chembl/api/data/molecule
Check data output	Remember to spot check the API output e.g. For one compound: have I captured all relevant data? Are activity data for both parent and salt drug forms present?
Schema:	A schema description is available at e.g. https://www.ebi.ac.uk/chembl/api/data/activity/schema (although it can be easier to examine an actual example instead)
Try it out interactively:	Use our interactive API documentation https://www.ebi.ac.uk/chembl/api/data/docs

www.ebi.ac.uk/chembl

www.ebi.ac.uk/unichem/
www.ebi.ac.uk/unichem/beta/

- **Webinars:**
  - ChEMBL Webinar: https://www.ebi.ac.uk/training/events/guide-explore-drug-compounds-and-their-biological-targets-using-chembl/
  - ChEMBL quick tour: https://www.ebi.ac.uk/training-beta/online/courses/chembl-quick-tour/
- **Documentation:**
  - ChEMBL web services: https://chembl.gitbook.io/chembl-interface-documentation/web-services
  - UniChem documentation: https://chembl.gitbook.io/unichem/ , Beta: https://chembl.gitbook.io/unichem/unichem-2.0/
- **FTP sites:**
  - UniChem: ftp://ftp.ebi.ac.uk/pub/databases/chembl/UniChem/data/
  - ChEMBL: https://ftp.ebi.ac.uk/pub/databases/chembl/ChEMBLdb/releases/
- **Contact us:** chembl-help@ebi.ac.uk, unichem@ebi.ac.uk
- **Blog:** https://chembl.blogspot.com/
- **FAQs:** https://chembl.gitbook.io/chembl-interface-documentation/frequently-asked-questions
- **For news and data releases subscribe to:** http://listserver.ebi.ac.uk/mailman/listinfo/chembl-announce

EMBL-EBI

Acknowledgements

Chemical Biology Team

- Andrew R. Leach (Team Leader)
- **Barbara Zdrazil (Group Coordinator)***
- Eloy Felix (Technical Lead)
- **Fiona Hunter (Drug Data Leader)***
- Ricardo Arcila
- Melissa Adasme
- James Blackshaw
- Nicolas Bosc
- Sybilla Corbett
- Marleen De Veij
- Harris Ioannidis
- Tevfik Kiziloren
- Emma Manners
- M. Paula Magarinos
- **David Mendez***
- Juan F. Mosquera

* Contributed slides to these training materials