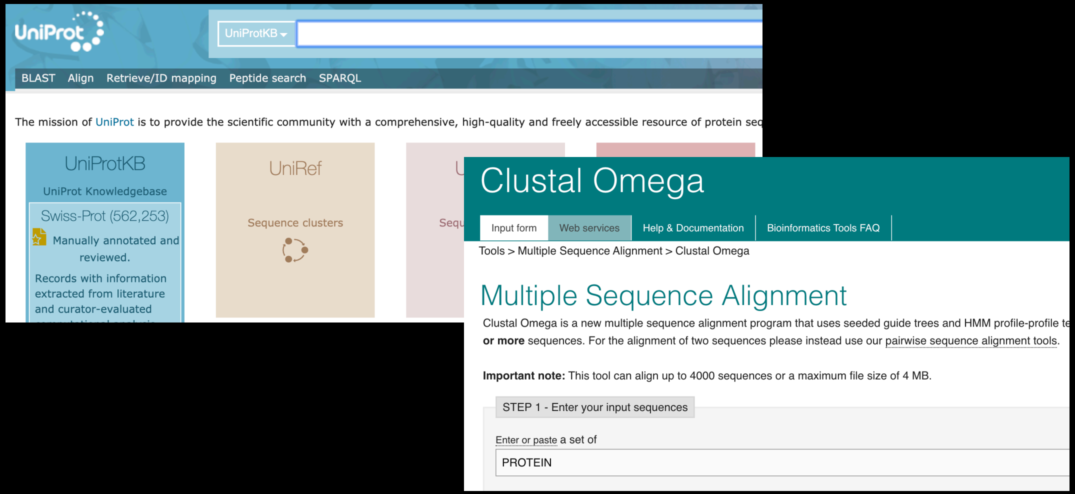
ChEMBL

EMBL-EBI



The mission of UniProt is to provide the scientific community with a comprehensive, high-quality and freely accessible resource of protein seq

UniProtKB

UniProt Knowledgebase

Swiss-Prot (562,253)

Manually annotated and reviewed.

Records with information extracted from literature and curator-evaluated

UniRef

Sequence clusters

Sequ

Clustal Omega

Tools > Multiple Sequence Alignment > Clustal Omega

Multiple Sequence Alignment

Clustal Omega is a new multiple sequence alignment program that uses seeded guide trees and HMM profile-profile te **or more** sequences. For the alignment of two sequences please instead use our pairwise sequence alignment tools.

Important note: This tool can align up to 4000 sequences or a maximum file size of 4 MB.

STEP 1 - Enter your input sequences

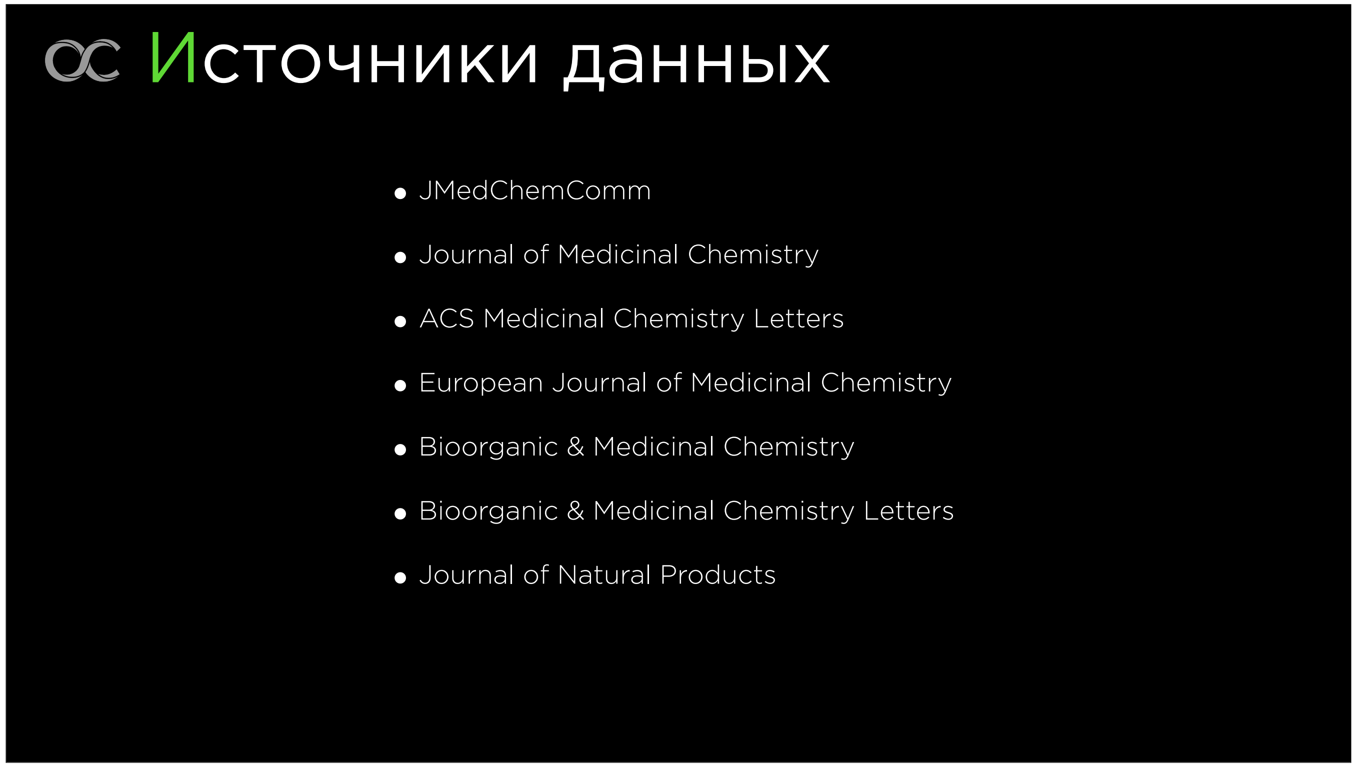
Enter or paste a set of

PROTEIN

Источники данных

ChEMBL is a large, open-access drug discovery database that aims to capture Medicinal Chemistry data and knowledge across the pharmaceutical research and development process

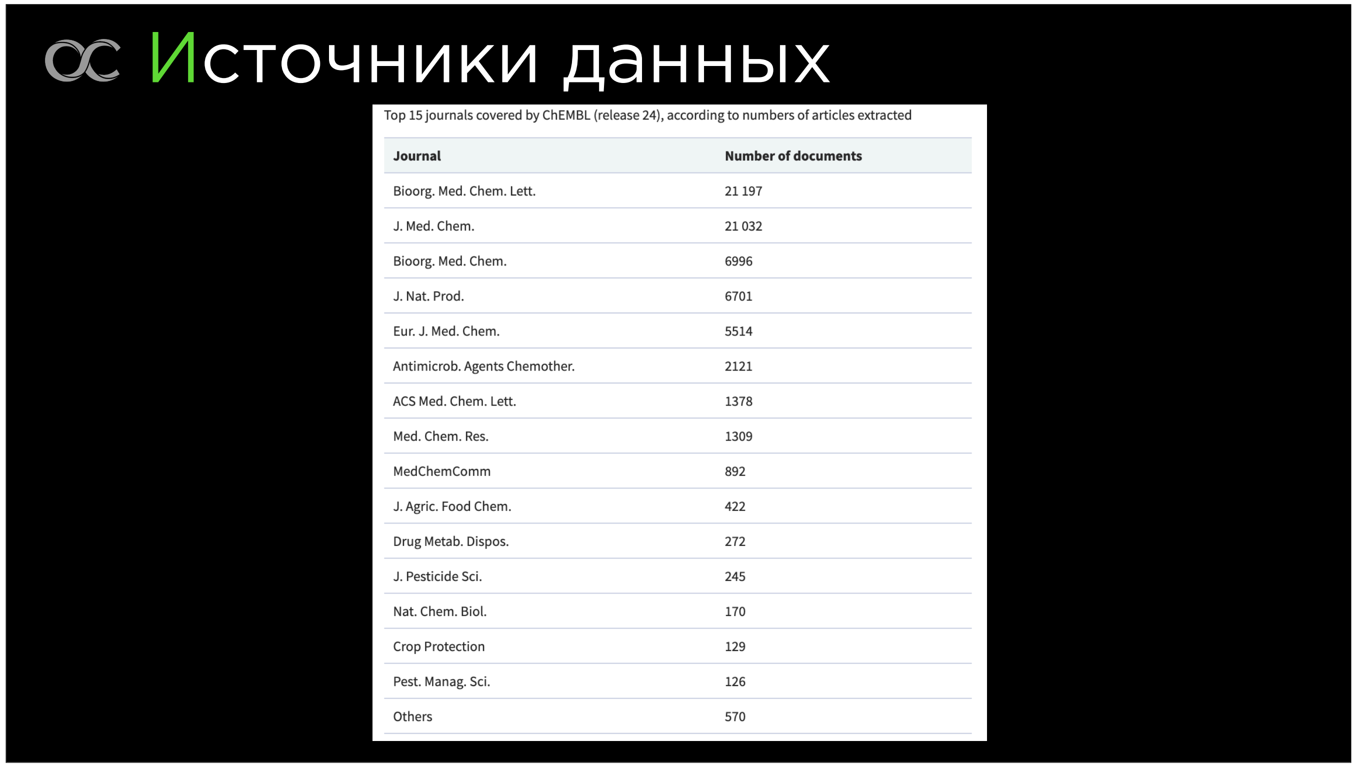
https://academic.oup.com/nar/article/47/D1/D930/5162468



Источники данных

- JMedChemComm
- Journal of Medicinal Chemistry
- ACS Medicinal Chemistry Letters
- European Journal of Medicinal Chemistry
- Bioorganic & Medicinal Chemistry
- Bioorganic & Medicinal Chemistry Letters
- Journal of Natural Products

For these journals, every article in each new issue is screened for the presence of quantitative small molecule (or peptide) bioactivity data.

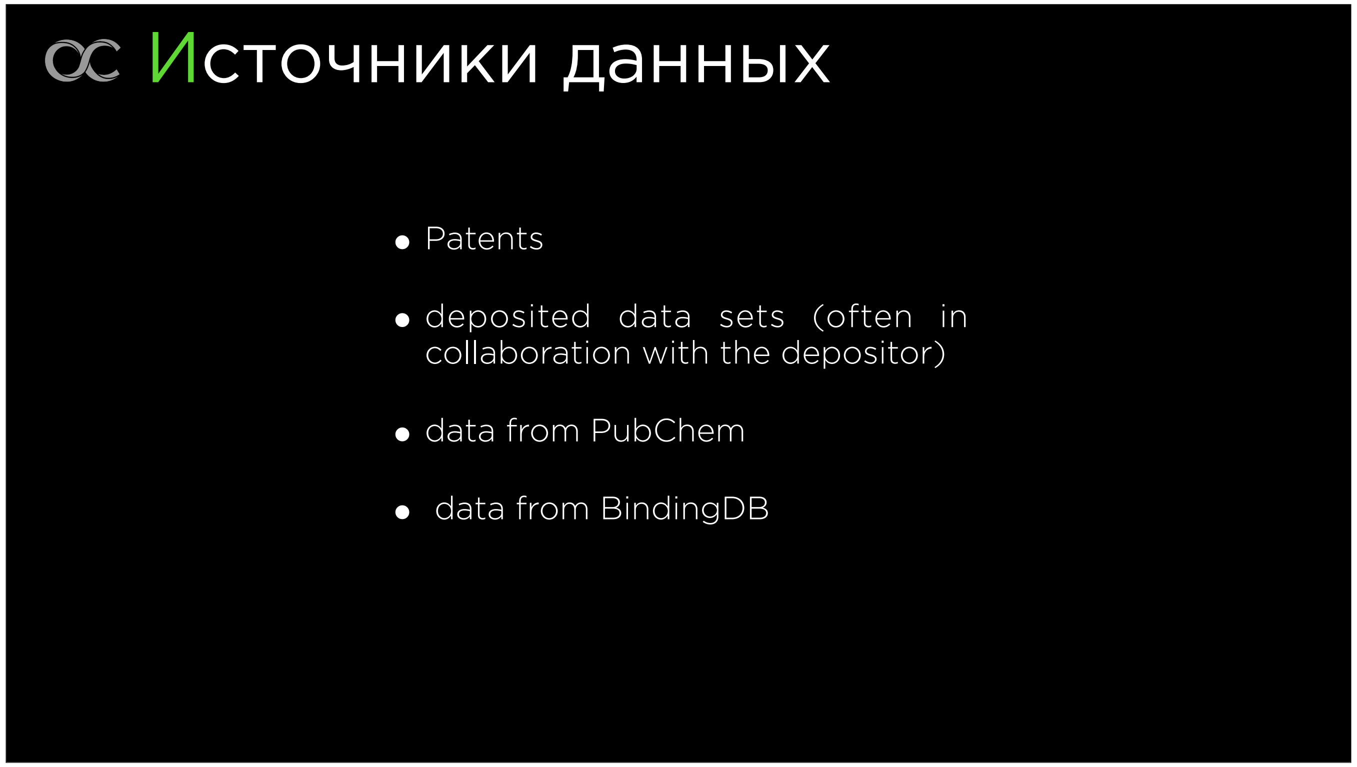


Источники данных

Top 15 journals covered by ChEMBL (release 24), according to numbers of articles extracted

Journal	Number of documents
Bioorg. Med. Chem. Lett.	21 197
J. Med. Chem.	21 032
Bioorg. Med. Chem.	6996
J. Nat. Prod.	6701
Eur. J. Med. Chem.	5514
Antimicrob. Agents Chemother.	2121
ACS Med. Chem. Lett.	1378
Med. Chem. Res.	1309
MedChemComm	892
J. Agric. Food Chem.	422
Drug Metab. Dispos.	272
J. Pesticide Sci.	245
Nat. Chem. Biol.	170
Crop Protection	129
Pest. Manag. Sci.	126
Others	570

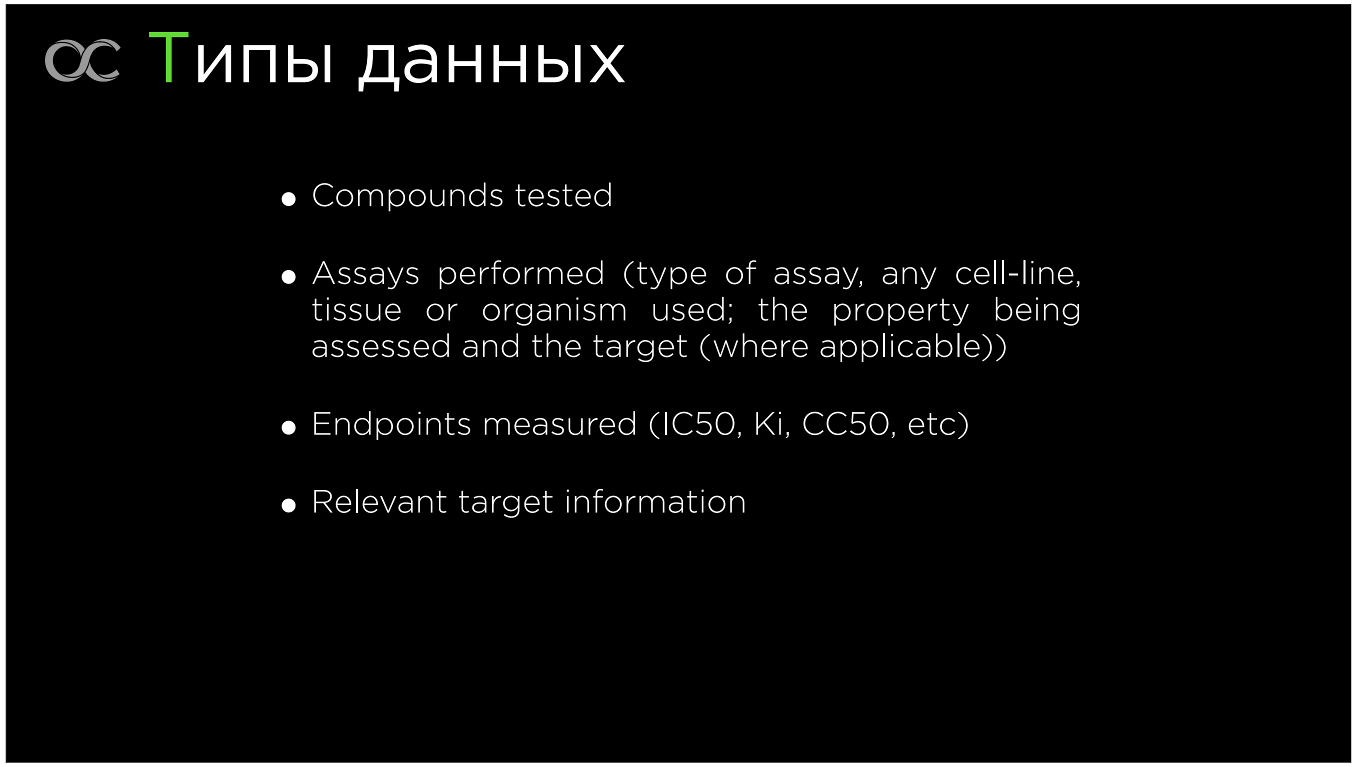
Data are also sometimes extracted from other journals and articles, when relevant to an area of particular interest



Источники данных

- Patents
- deposited data sets (often in collaboration with the depositor)
- data from PubChem
- data from BindingDB

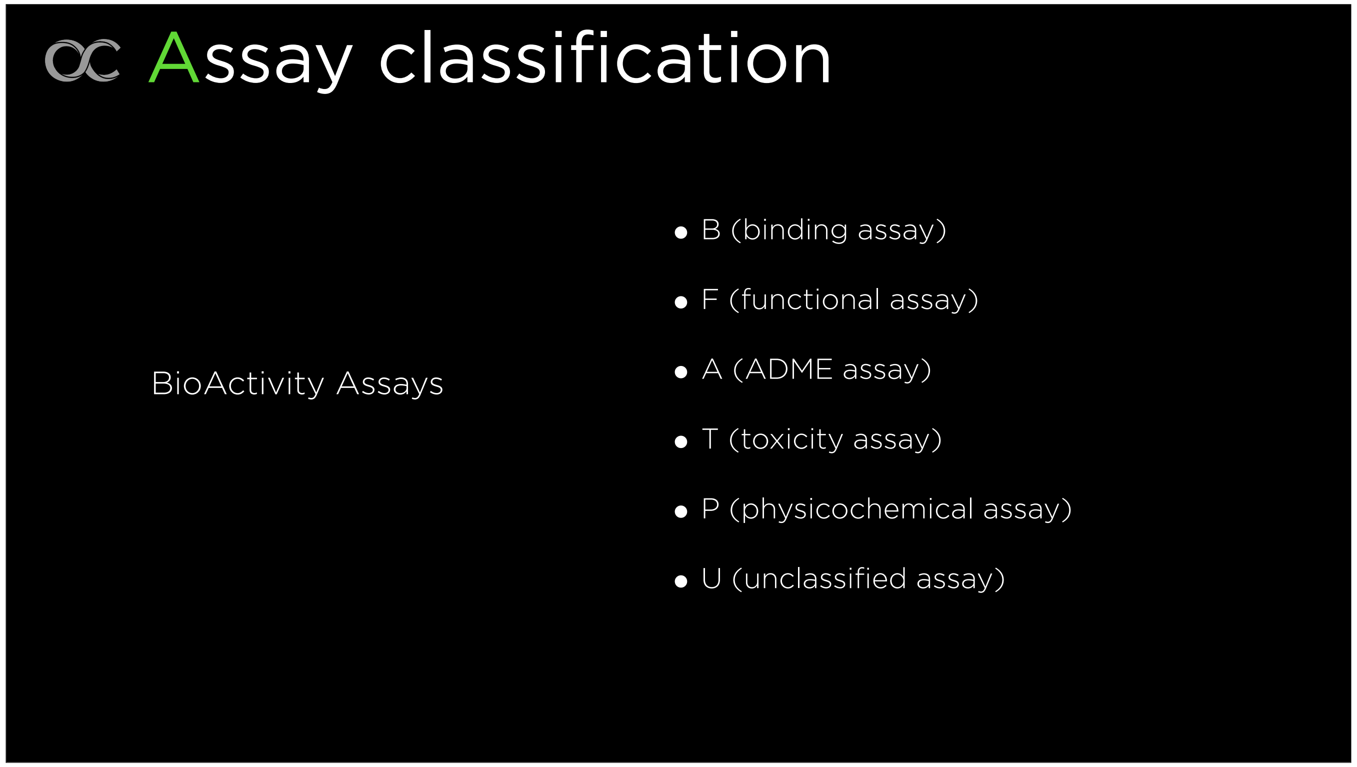
In addition to data from the peer-reviewed literature, ChEMBL now incorporates some data extracted from patent documents. This data extraction effort is focused on patents containing compounds and bioactivity data for targets that are not currently well represented in ChEMBL (and in particular, those of interest to the NIH Illuminating the Druggable Genome project: https://commonfund.nih.gov/IDG/understudiedproteins (7)). To date, 74 050 activity measurements have been extracted from 241 patents, and these are available in ChEMBL as source 38. Additional, focused sets may also be included in future releases, where these add value to the existing ChEMBL data. A larger set of bioactivity data extracted from granted US patents by BindingDB (described previously) is also available in ChEMBL as source 37 (3).



Типы данных

- Compounds tested
- Assays performed (type of assay, any cell-line, tissue or organism used; the property being assessed and the target (where applicable))
- Endpoints measured (IC50, Ki, CC50, etc)
- Relevant target information

From each article, details of compounds tested, assays performed, endpoints measured and relevant target information, are extracted. Compound structures are drawn in full (including any salt present) and saved in V2000 Mol file format. A brief description of the assay is abstracted; this typically includes the type of assay being performed; any cell-line, tissue or organism used; the property being assessed and the target (where applicable). Biological activity data is recorded; this includes (but is not limited to) binding measurements, efficacy in functional assays, pharmacokinetic data and toxicity endpoints.



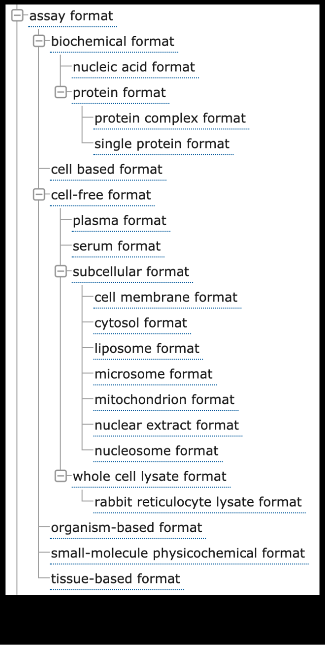
Assay classification

BioActivity Assays

- B (binding assay)
- F (functional assay)
- A (ADME assay)
- T (toxicity assay)
- P (physicochemical assay)
- U (unclassified assay)

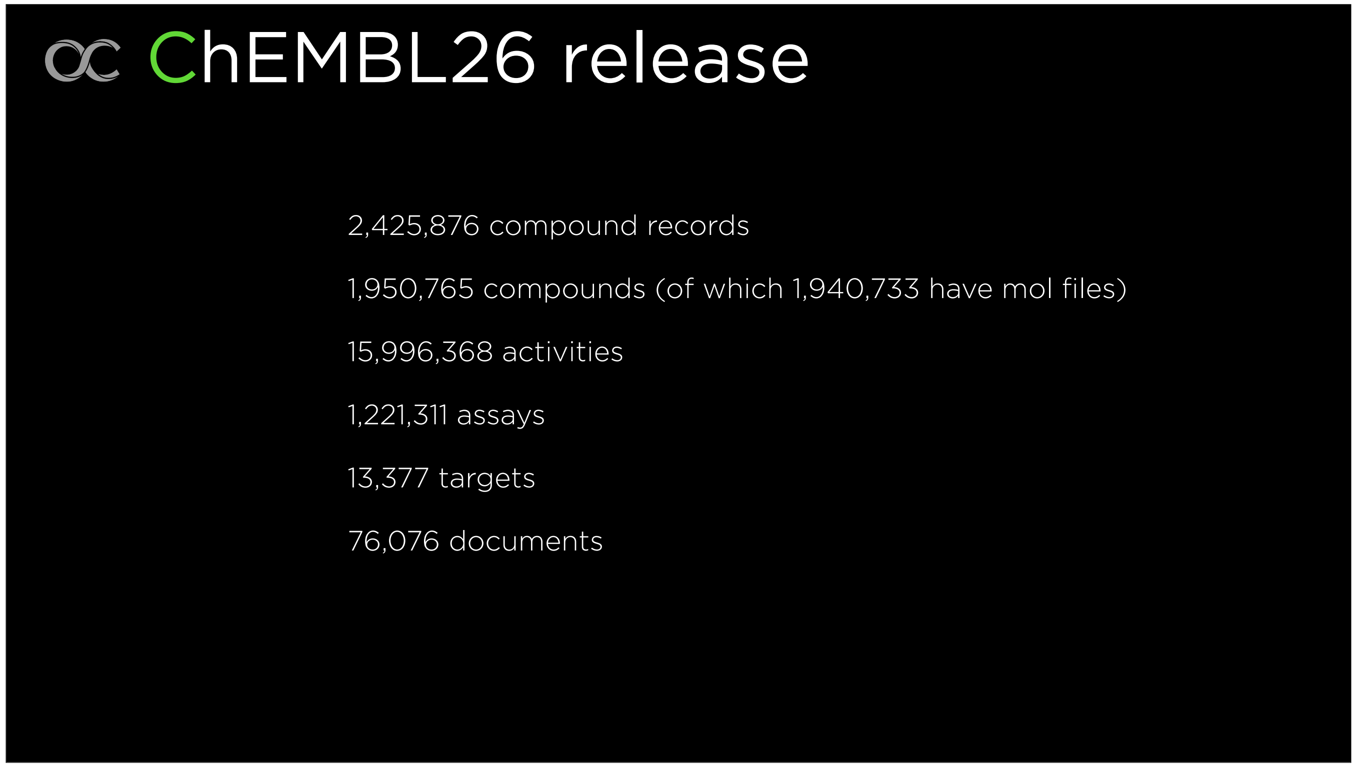
Each assay is classified into one of the following categories: B (binding assay), F (functional assay), A (ADME assay), T (toxicity assay), P (physicochemical assay) or U (unclassified assay). Where the assay describes the interaction with a molecular target, this target is also recorded in the form of a UniProt (15) accession, or list of accessions. Quantitative and qualitative activity measurements are extracted in the form reported in the publication, with their respective activity types, units and qualifiers.

BAO format



- assay format
  - biochemical format
    - nucleic acid format
    - protein format
      - protein complex format
      - single protein format
  - cell based format
  - cell-free format
    - plasma format
    - serum format
    - subcellular format
      - cell membrane format
      - cytosol format
      - liposome format
      - microsome format
      - mitochondrion format
      - nuclear extract format
      - nucleosome format
    - whole cell lysate format
      - rabbit reticulocyte lysate format
  - organism-based format
  - small-molecule physicochemical format
  - tissue-based format

https://www.ebi.ac.uk/ols/ontologies/bao



ChEMBL26 release

2,425,876 compound records

1,950,765 compounds (of which 1,940,733 have mol files)

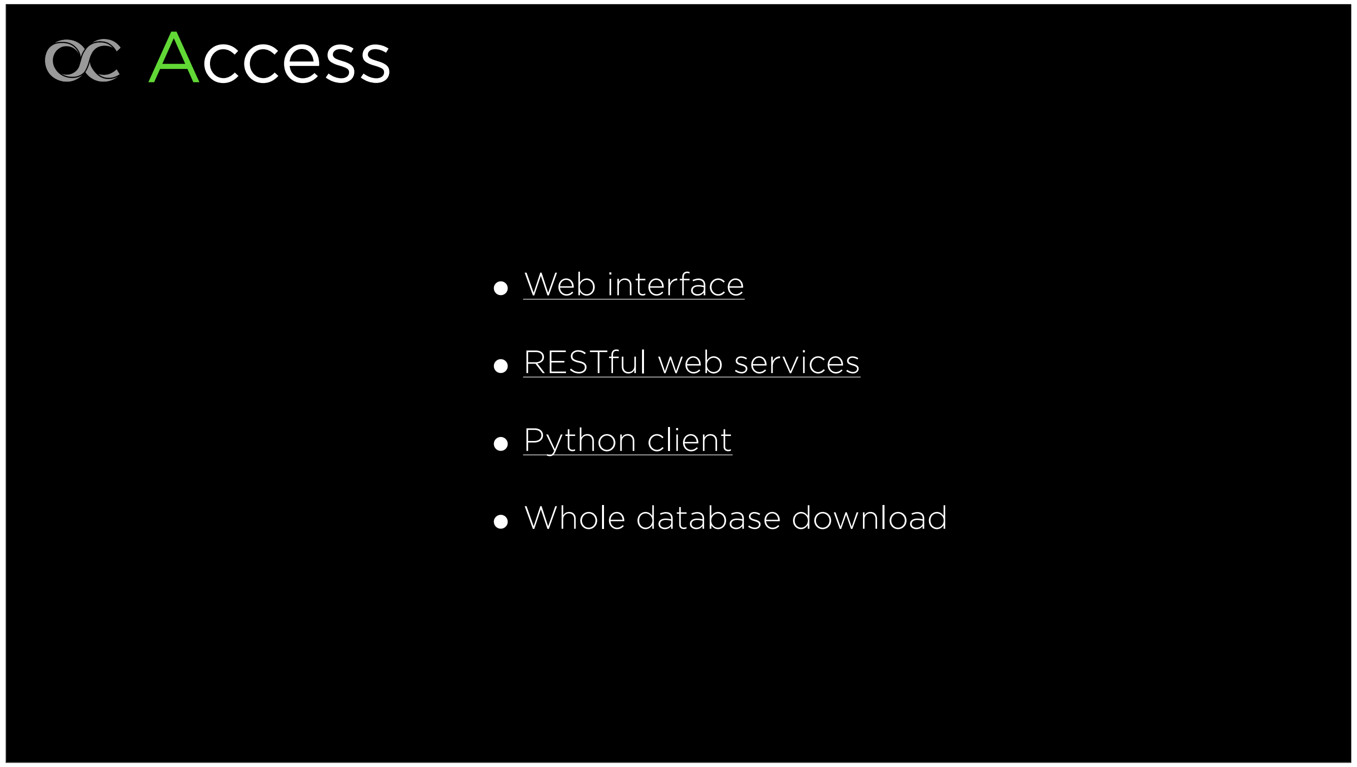
15,996,368 activities

1,221,311 assays

13,377 targets

76,076 documents

Following initial data extraction, an extensive data curation and standardization process is applied before incorporating the information into ChEMBL. In brief, compound structures are standardized, salt-stripped and assigned identifiers based on Standard InChI (17); assay descriptions are mapped to controlled vocabularies such as the Cell Line Ontology (18), Uberon (19) and BioAssay Ontology (20); activity measurements and units are converted to a standard form; and multi-protein targets are created where required. This same curation process is also applied to other data sources within ChEMBL such as data extracted from patents, deposited data sets (often in collaboration with the depositor) and data from PubChem BioAssay and BindingDB.



Access

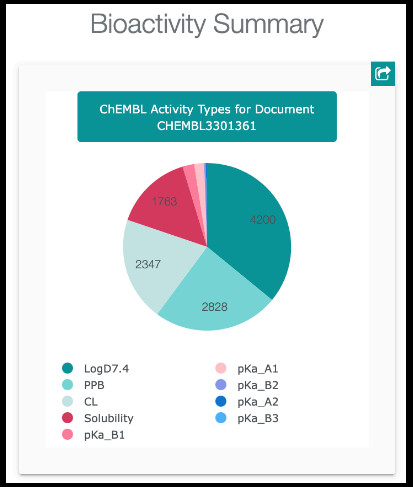
- Web interface
- RESTful web services
- Python client
- Whole database download

ftp://ftp.ebi.ac.uk/pub/databases/chembl/ChEMBLdb/latest/

In vitro DMPK and physchem data

AstraZeneka

DOI: 10.6019/CHEMBL3301361



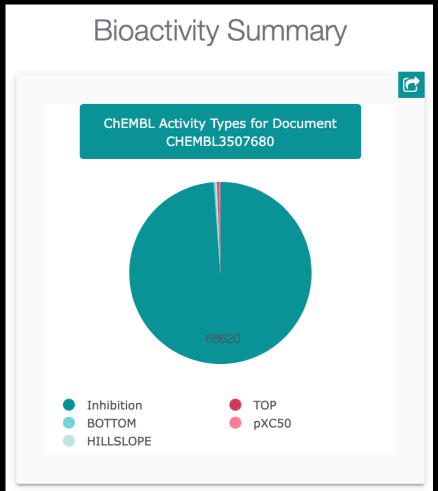
Experimental data determined at AstraZeneca on a set of compounds in the following assays: pKa, lipophilicity (LogD7.4), aqueous solubility, plasma protein binding (human, rat, dog , mouse and guinea pig), intrinsic clearance (human liver microsomes, human and rat hepatocytes). The references provided for the assays exemplify the experimental procedures used in generating the data.

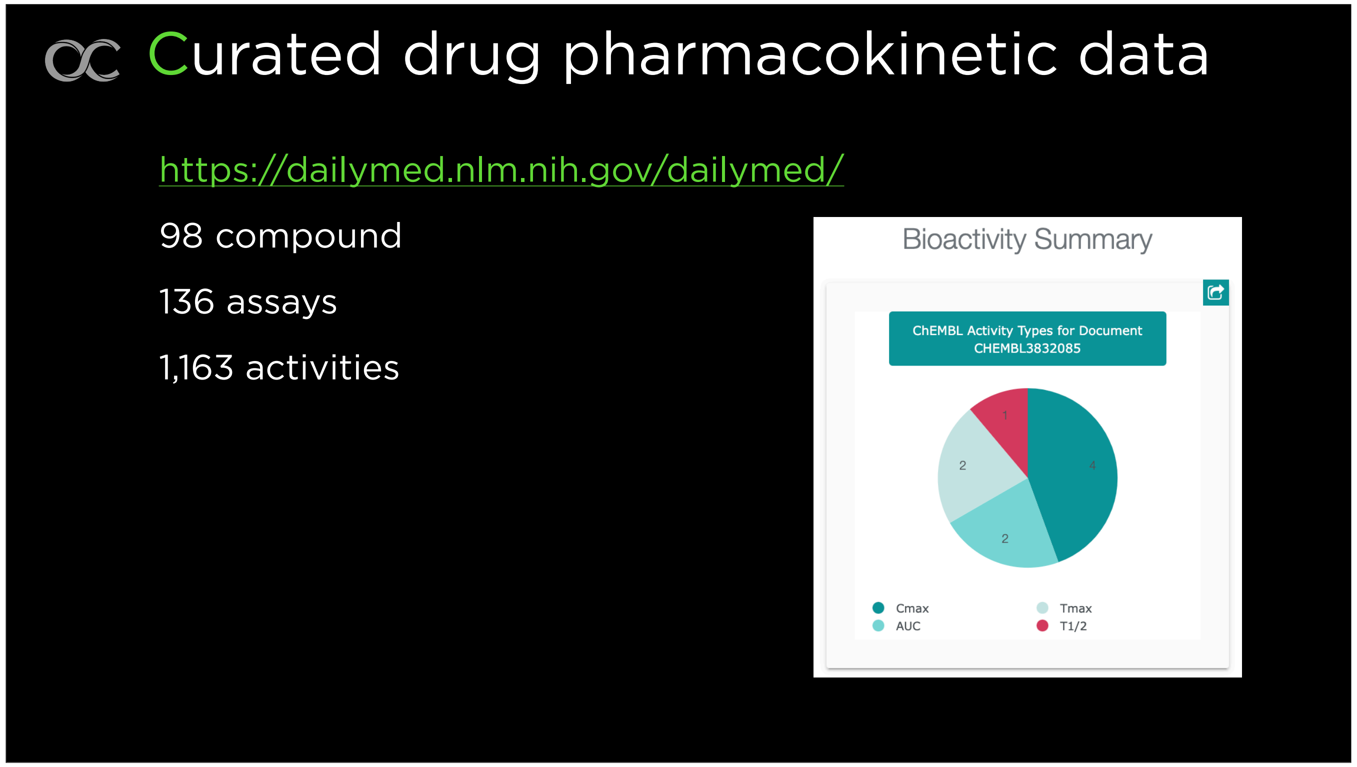
Toxicity

Gates Library compound collection

DOI:10.6019/CHEMBL3507680

Human HepG2 cell viability assay in 384 well format





Curated drug pharmacokinetic data

https://dailymed.nlm.nih.gov/dailymed/

98 compound

136 assays

1,163 activities

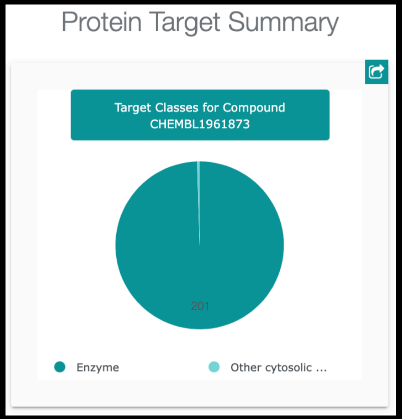
Bioactivity Summary

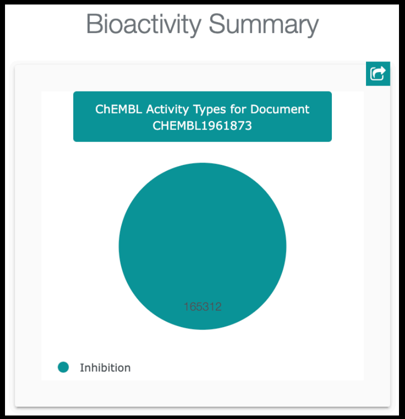
ChEMBL Activity Types for Document CHEMBL3832085

Knowledge of the pharmacokinetic properties of drugs is critical in understanding their safety and efficacy profiles, yet such data are often not readily available in a structured form. In order to begin to address this, pharmacokinetic measurements for 85 drugs have been extracted from reference books (21) and drug prescribing information (https://dailymed.nlm.nih.gov/dailymed/) and incorporated into ChEMBL as source 39

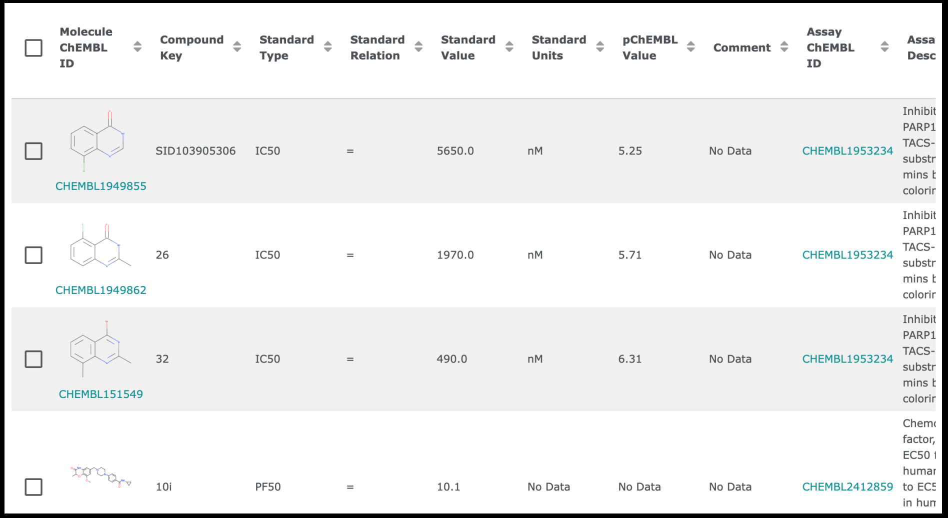
GSK Published Kinase Inhibitor Set

DOI:10.6019/CHEMBL2007661
DOI:10.6019/CHEMBL1961873



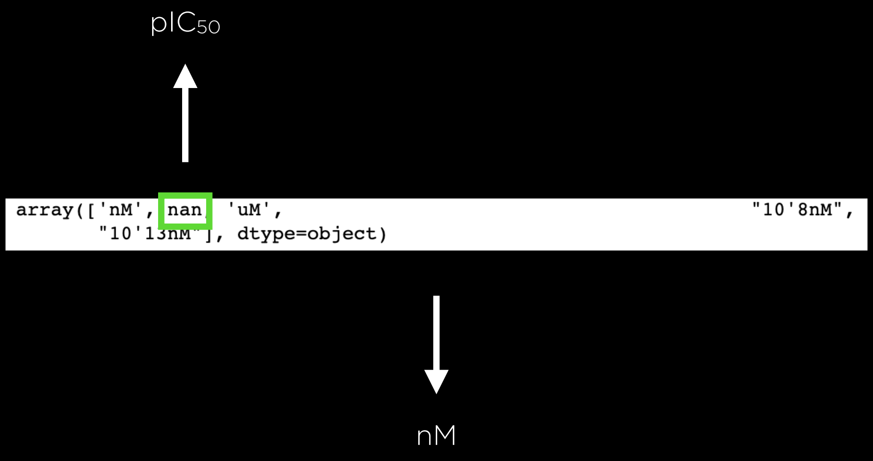


PARP1 case



□	Molecule ChEMBL ID	Compound Key	Standard Type	Standard Relation	Standard Value	Standard Units	pChEMBL Value	Comment	Assay ChEMBL ID	Assa Desc
□	CHEMBL1949855	SID103905306	IC50	=	5650.0	nM	5.25	No Data	CHEMBL1953234	Inhibit PARP1 TACS- substr mins t colorir
□	CHEMBL1949862	26	IC50	=	1970.0	nM	5.71	No Data	CHEMBL1953234	Inhibit PARP1 TACS- substr mins t colorir
□	CHEMBL151549	32	IC50	=	490.0	nM	6.31	No Data	CHEMBL1953234	Inhibit PARP1 TACS- substr mins t colorir
□		10i	PF50	=	10.1	No Data	No Data	No Data	CHEMBL2412859	Chem factor, EC50 t humar to EC5 in hun

Единицы измерения

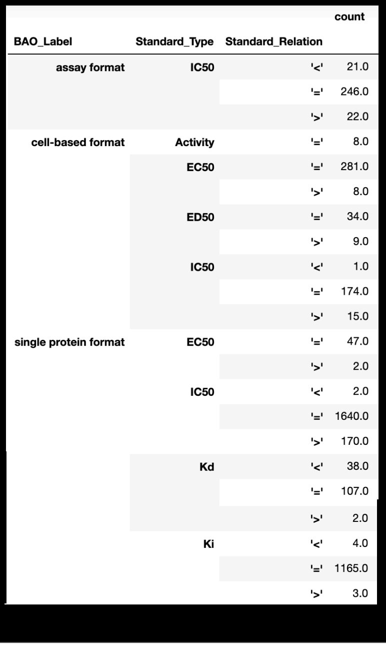


pIC$_{50}$

```
array(['nM', nan, 'uM',                                                  "10'8nM",
       "10'13nM"], dtype=object)
```

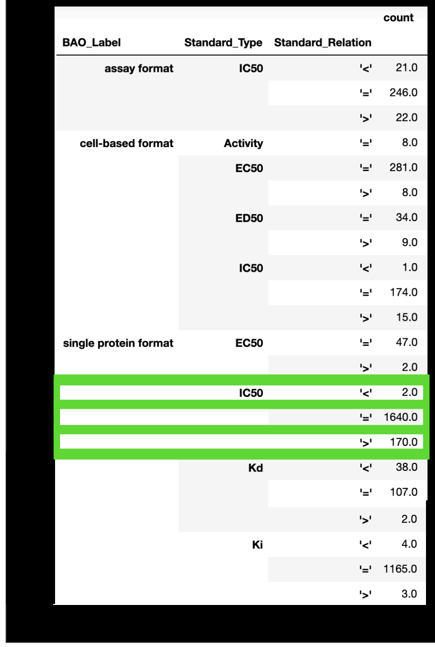
nM

BAO Label & Standard Type



			count
BAO_Label	Standard_Type	Standard_Relation	
assay format	IC50	'<'	21.0
		'='	246.0
		'>'	22.0
cell-based format	Activity	'='	8.0
	EC50	'='	281.0
		'>'	8.0
	ED50	'='	34.0
		'>'	9.0
	IC50	'<'	1.0
		'='	174.0
		'>'	15.0
single protein format	EC50	'='	47.0
		'>'	2.0
	IC50	'<'	2.0
		'='	1640.0
		'>'	170.0
	Kd	'<'	38.0
		'='	107.0
		'>'	2.0
	Ki	'<'	4.0
		'='	1165.0
		'>'	3.0

BAO Label & Standard Type



			count
BAO_Label	Standard_Type	Standard_Relation	
assay format	IC50	'<'	21.0
		'='	246.0
		'>'	22.0
cell-based format	Activity	'='	8.0
	EC50	'='	281.0
		'>'	8.0
	ED50	'='	34.0
		'>'	9.0
	IC50	'<'	1.0
		'='	174.0
		'>'	15.0
single protein format	EC50	'='	47.0
		'>'	2.0
	IC50	'<'	2.0
		'='	1640.0
		'>'	170.0
	Kd	'<'	38.0
		'='	107.0
		'>'	2.0
	Ki	'<'	4.0
		'='	1165.0
		'>'	3.0

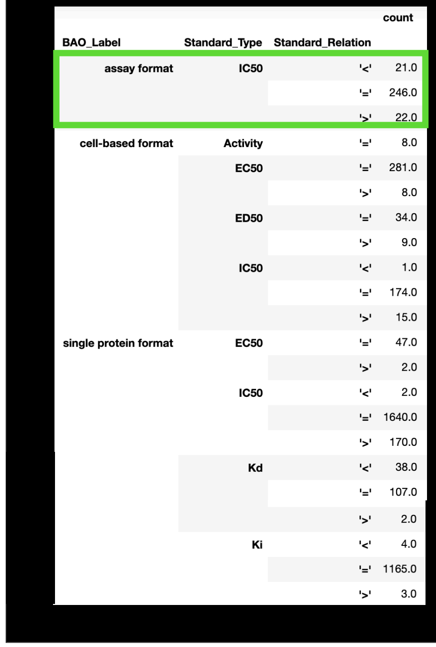
- Биохимическая активность (на белке)

- Активно: IC$_{50}$ <= 50 нМ

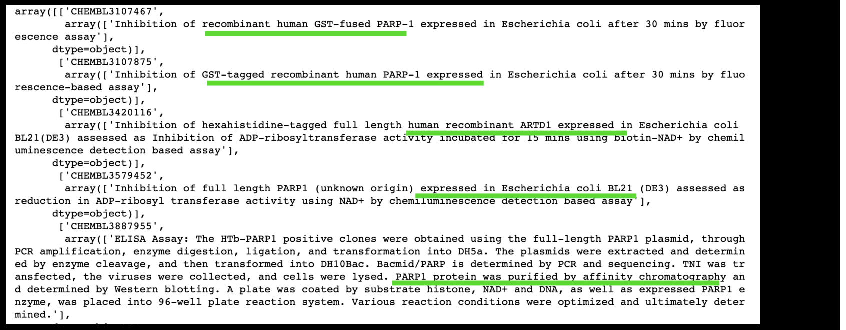
- Неактивно: IC$_{50}$ > 50 нМ

- Собрать как можно больше данных

BAO Label & Assay format



			count
BAO_Label	Standard_Type	Standard_Relation	
assay format	IC50	'<'	21.0
		'='	246.0
		'>'	22.0
cell-based format	Activity	'='	8.0
	EC50	'='	281.0
		'>'	8.0
	ED50	'='	34.0
		'>'	9.0
	IC50	'<'	1.0
		'='	174.0
		'>'	15.0
single protein format	EC50	'='	47.0
		'>'	2.0
	IC50	'<'	2.0
		'='	1640.0
		'>'	170.0
	Kd	'<'	38.0
		'='	107.0
		'>'	2.0
	Ki	'<'	4.0
		'='	1165.0
		'>'	3.0

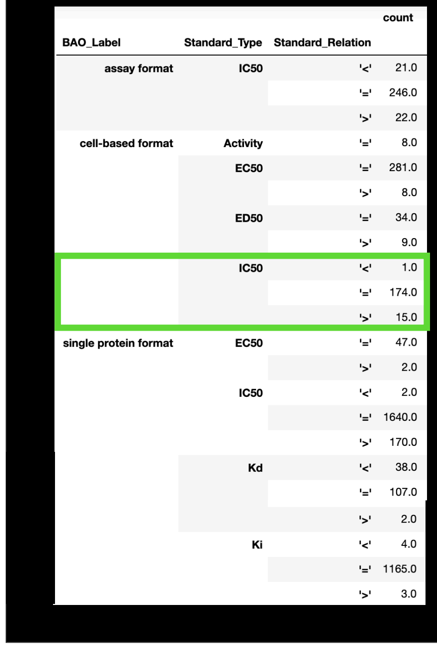


```
array([['CHEMBL3107467',
        array(['Inhibition of recombinant human GST-fused PARP-1 expressed in Escherichia coli after 30 mins by fluor
escence assay'],
      dtype=object)],
       ['CHEMBL3107875',
        array(['Inhibition of GST-tagged recombinant human PARP-1 expressed in Escherichia coli after 30 mins by fluo
rescence-based assay'],
      dtype=object)],
       ['CHEMBL3420116',
        array(['Inhibition of hexahistidine-tagged full length human recombinant ARTD1 expressed in Escherichia coli
BL21(DE3) assessed as Inhibition of ADP-ribosyltransferase activity incubated for 15 mins using biotin-NAD+ by chemil
uminescence detection based assay'],
      dtype=object)],
       ['CHEMBL3579452',
        array(['Inhibition of full length PARP1 (unknown origin) expressed in Escherichia coli BL21 (DE3) assessed as
reduction in ADP-ribosyl transferase activity using NAD+ by chemiluminescence detection based assay'],
      dtype=object)],
       ['CHEMBL3887955',
        array(['ELISA Assay: The HTb-PARP1 positive clones were obtained using the full-length PARP1 plasmid, through
PCR amplification, enzyme digestion, ligation, and transformation into DH5a. The plasmids were extracted and determin
ed by enzyme cleavage, and then transformed into DH10Bac. Bacmid/PARP is determined by PCR and sequencing. TNI was tr
ansfected, the viruses were collected, and cells were lysed. PARP1 protein was purified by affinity chromatography an
d determined by Western blotting. A plate was coated by substrate histone, NAD+ and DNA, as well as expressed PARP1 e
nzyme, was placed into 96-well plate reaction system. Various reaction conditions were optimized and ultimately deter
mined.'],
```

↓

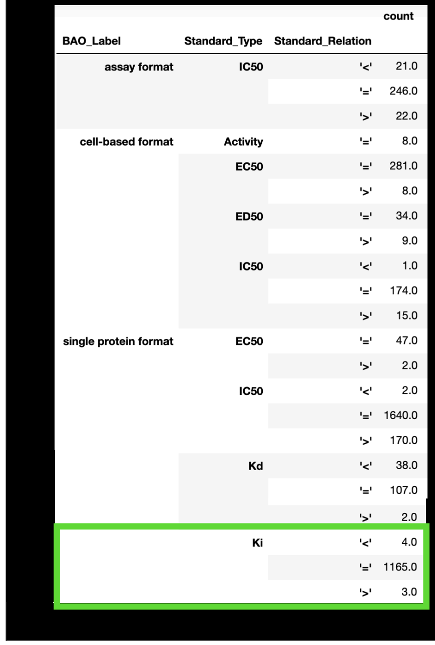
Single Protein Format

BAO Label & Assay format



			count
BAO_Label	Standard_Type	Standard_Relation	
assay format	IC50	'<'	21.0
		'='	246.0
		'>'	22.0
cell-based format	Activity	'='	8.0
	EC50	'='	281.0
		'>'	8.0
	ED50	'='	34.0
		'>'	9.0
	IC50	'<'	1.0
		'='	174.0
		'>'	15.0
single protein format	EC50	'='	47.0
		'>'	2.0
	IC50	'<'	2.0
		'='	1640.0
		'>'	170.0
	Kd	'<'	38.0
		'='	107.0
		'>'	2.0
	Ki	'<'	4.0
		'='	1165.0
		'>'	3.0

BAO Label & Standard Type

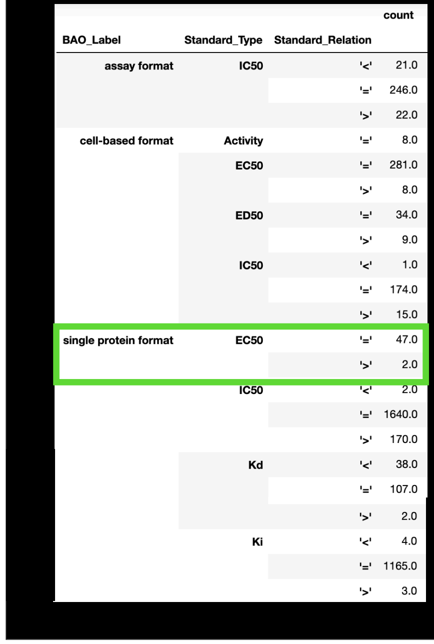


			count
BAO_Label	Standard_Type	Standard_Relation	
assay format	IC50	'<'	21.0
		'='	246.0
		'>'	22.0
cell-based format	Activity	'='	8.0
	EC50	'='	281.0
		'>'	8.0
	ED50	'='	34.0
		'>'	9.0
	IC50	'<'	1.0
		'='	174.0
		'>'	15.0
single protein format	EC50	'='	47.0
		'>'	2.0
	IC50	'<'	2.0
		'='	1640.0
		'>'	170.0
	Kd	'<'	38.0
		'='	107.0
		'>'	2.0
	Ki	'<'	4.0
		'='	1165.0
		'>'	3.0

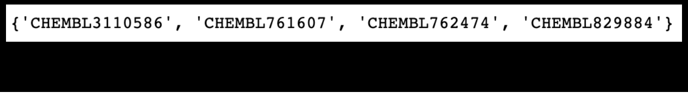
$$K_i = \frac{IC_{50}}{1 + [L]/K_D}$$

- Можно извлечь только неактивные

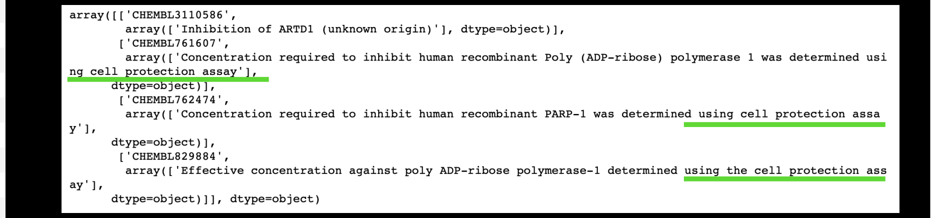
BAO Label & Standard Type



BAO_Label	Standard_Type	Standard_Relation	count
assay format	IC50	'<'	21.0
		'='	246.0
		'>'	22.0
cell-based format	Activity	'='	8.0
	EC50	'='	281.0
		'>'	8.0
	ED50	'='	34.0
		'>'	9.0
	IC50	'<'	1.0
		'='	174.0
		'>'	15.0
single protein format	EC50	'='	47.0
		'>'	2.0
	IC50	'<'	2.0
		'='	1640.0
		'>'	170.0
	Kd	'<'	38.0
		'='	107.0
		'>'	2.0
	Ki	'<'	4.0
		'='	1165.0
		'>'	3.0



```
{'CHEMBL3110586', 'CHEMBL761607', 'CHEMBL762474', 'CHEMBL829884'}
```

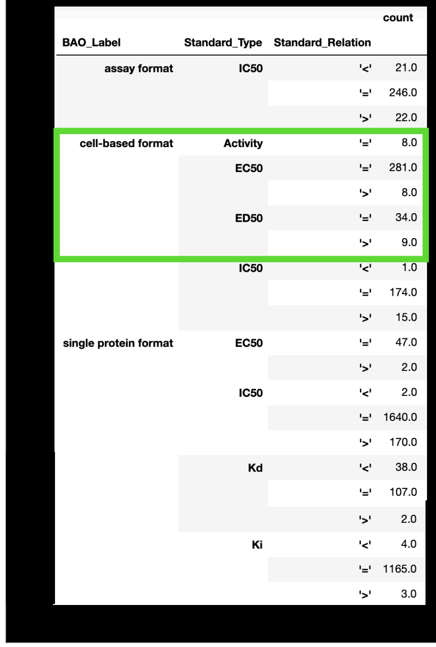


```
array([['CHEMBL3110586',
        array(['Inhibition of ARTD1 (unknown origin)'], dtype=object)],
       ['CHEMBL761607',
        array(['Concentration required to inhibit human recombinant Poly (ADP-ribose) polymerase 1 was determined usi
ng cell protection assay'],
      dtype=object)],
       ['CHEMBL762474',
        array(['Concentration required to inhibit human recombinant PARP-1 was determined using cell protection assa
y'],
      dtype=object)],
       ['CHEMBL829884',
        array(['Effective concentration against poly ADP-ribose polymerase-1 determined using the cell protection ass
ay'],
      dtype=object)]], dtype=object)
```

↓

Cell-based Format

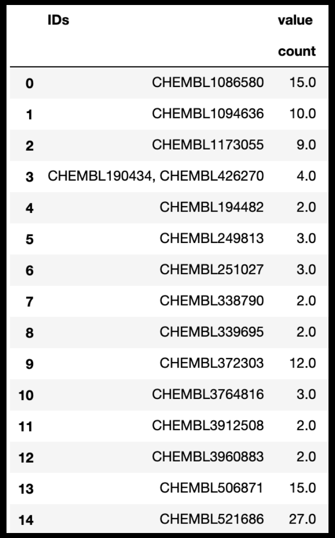
BAO Label & Standard Type



			count
BAO_Label	Standard_Type	Standard_Relation	
assay format	IC50	'<'	21.0
		'='	246.0
		'>'	22.0
cell-based format	Activity	'='	8.0
	EC50	'='	281.0
		'>'	8.0
	ED50	'='	34.0
		'>'	9.0
	IC50	'<'	1.0
		'='	174.0
		'>'	15.0
single protein format	EC50	'='	47.0
		'>'	2.0
	IC50	'<'	2.0
		'='	1640.0
		'>'	170.0
	Kd	'<'	38.0
		'='	107.0
		'>'	2.0
	Ki	'<'	4.0
		'='	1165.0
		'>'	3.0

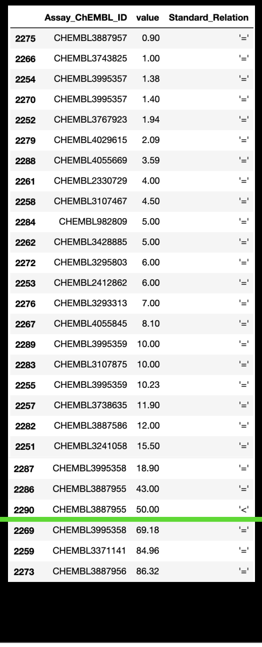
- Исключим cell death assays
- Исключим cell proliferation assay
- Учтем специфичные к PARP1 методики
- Учтем только активные соединения

Duplicates



	IDs	value
		count
0	CHEMBL1086580	15.0
1	CHEMBL1094636	10.0
2	CHEMBL1173055	9.0
3	CHEMBL190434, CHEMBL426270	4.0
4	CHEMBL194482	2.0
5	CHEMBL249813	3.0
6	CHEMBL251027	3.0
7	CHEMBL338790	2.0
8	CHEMBL339695	2.0
9	CHEMBL372303	12.0
10	CHEMBL3764816	3.0
11	CHEMBL3912508	2.0
12	CHEMBL3960883	2.0
13	CHEMBL506871	15.0
14	CHEMBL521686	27.0

'CHEMBL521686'



	Assay_ChEMBL_ID	value	Standard_Relation
2275	CHEMBL3887957	0.90	'='
2266	CHEMBL3743825	1.00	'='
2254	CHEMBL3995357	1.38	'='
2270	CHEMBL3995357	1.40	'='
2252	CHEMBL3767923	1.94	'='
2279	CHEMBL4029615	2.09	'='
2288	CHEMBL4055669	3.59	'='
2261	CHEMBL2330729	4.00	'='
2258	CHEMBL3107467	4.50	'='
2284	CHEMBL982809	5.00	'='
2262	CHEMBL3428885	5.00	'='
2272	CHEMBL3295803	6.00	'='
2253	CHEMBL2412862	6.00	'='
2276	CHEMBL3293313	7.00	'='
2267	CHEMBL4055845	8.10	'='
2289	CHEMBL3995359	10.00	'='
2283	CHEMBL3107875	10.00	'='
2255	CHEMBL3995359	10.23	'='
2257	CHEMBL3738635	11.90	'='
2282	CHEMBL3887586	12.00	'='
2251	CHEMBL3241058	15.50	'='
2287	CHEMBL3995358	18.90	'='
2286	CHEMBL3887955	43.00	'='
2290	CHEMBL3887955	50.00	'<'
2269	CHEMBL3995358	69.18	'='
2259	CHEMBL3371141	84.96	'='
2273	CHEMBL3887956	86.32	'='