- **Data Science Slack ワークスペース**
  - 下記の招待リンクからSlackワークスペースに参加して下さい。
  - Invitation to Data Science PBL Slack workspace.

https://join.slack.com/t/naist-dsc-pbl/shared_invite/zt-2ngqmcqxf-9eEyTMjuacVXAPSIV_a7zQ

(注：このリンクは３週間で無効になります / This link will expire in three weeks)

今後、スケジュールについてのアナウンスなどもこちらのワークスペースで案内していきます。/ We will announce the schedule of PBL and other information in that workspace.

Note: Slackワークスペースへの参加方法 /
How to join a Slack workspace.
[https://slack.com/help/articles/212675257-Join-a-Slack-workspace]

Data Science PBL I

2024/07/29

- **Overview**
  - 7/29 Introduction < now

    About the task of this PBL

  - 7/29-9/26 Group work

    Build the model, evaluate the results

    Discuss the meaning of the analysis

    Prepare the presentation materials

    (Discuss at least once every two weeks...)

  - 9/27 (Fri.) 9:20-12:30 Group Presentations (IS:AI lecuter room)

    Presentation: 15 min (talk) + 3 min (discussion)

    Slides: in English

    Talk: either in English or Japanese

- **Task: Chemo-bioinformatics analysis**
  - Develop a model using machine learning to predict the biochemical properties of chemical molecules from their molecular structures. Based on the results, analyze the relationship between structure and function from chemical and biological perspectives.

  - Dataset1

    Toxicity prediction

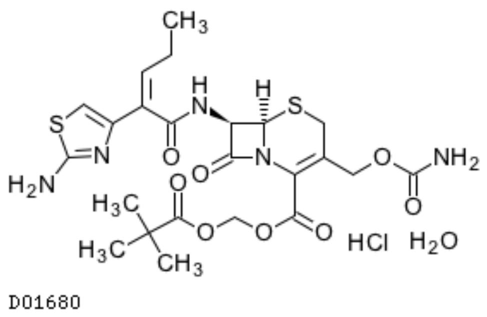
  - Dataset2

    Antibiotics screening

Dataset1:
Toxicity prediction

Quantitative structure-activity relationship (QSAR)

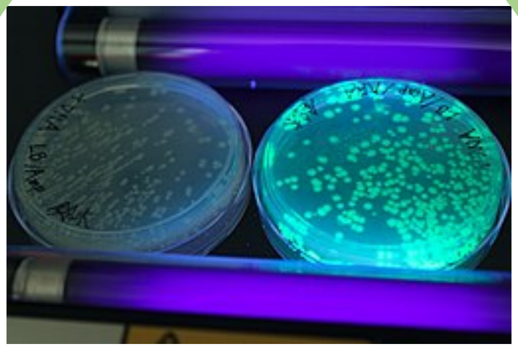
Material



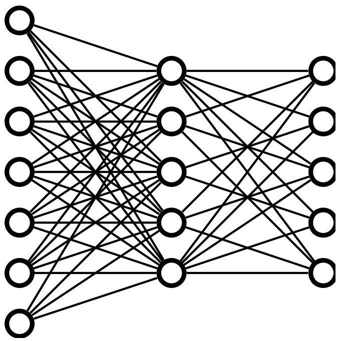
Chemical structures

Predict biological activity of chemical compounds using computational models

Biology



Biological activity



Machine learning

Data science

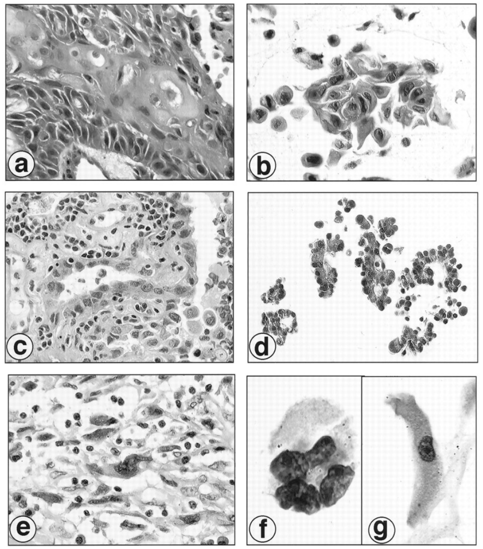
Cefcapene Pivoxil Hydrochloride, a.k.a.
「フロモックス」

Escherichia coli
a.k.a.
大腸菌

Cancer Cell Lines and Compounds Screening

Cell lines （細胞株）

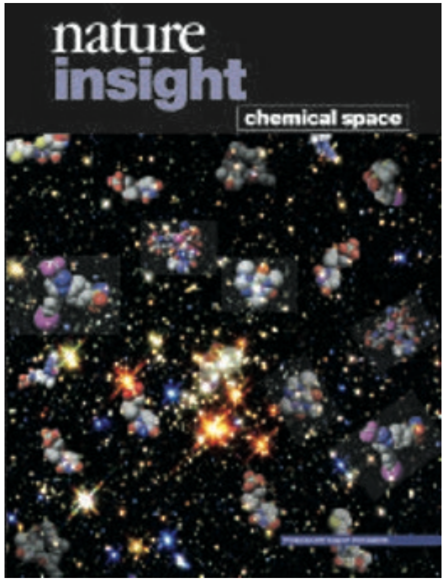
Thousands lines of immortalized cells (cancer tumor, stem cell etc.) have been isolated and cultivated continuously.



[Wistuba et al. Clinical Cancer Res. 1999]

Chemical space

‘Drug design’ is a painstaking search for candidates of new drug from billions of possible chemical compounds.

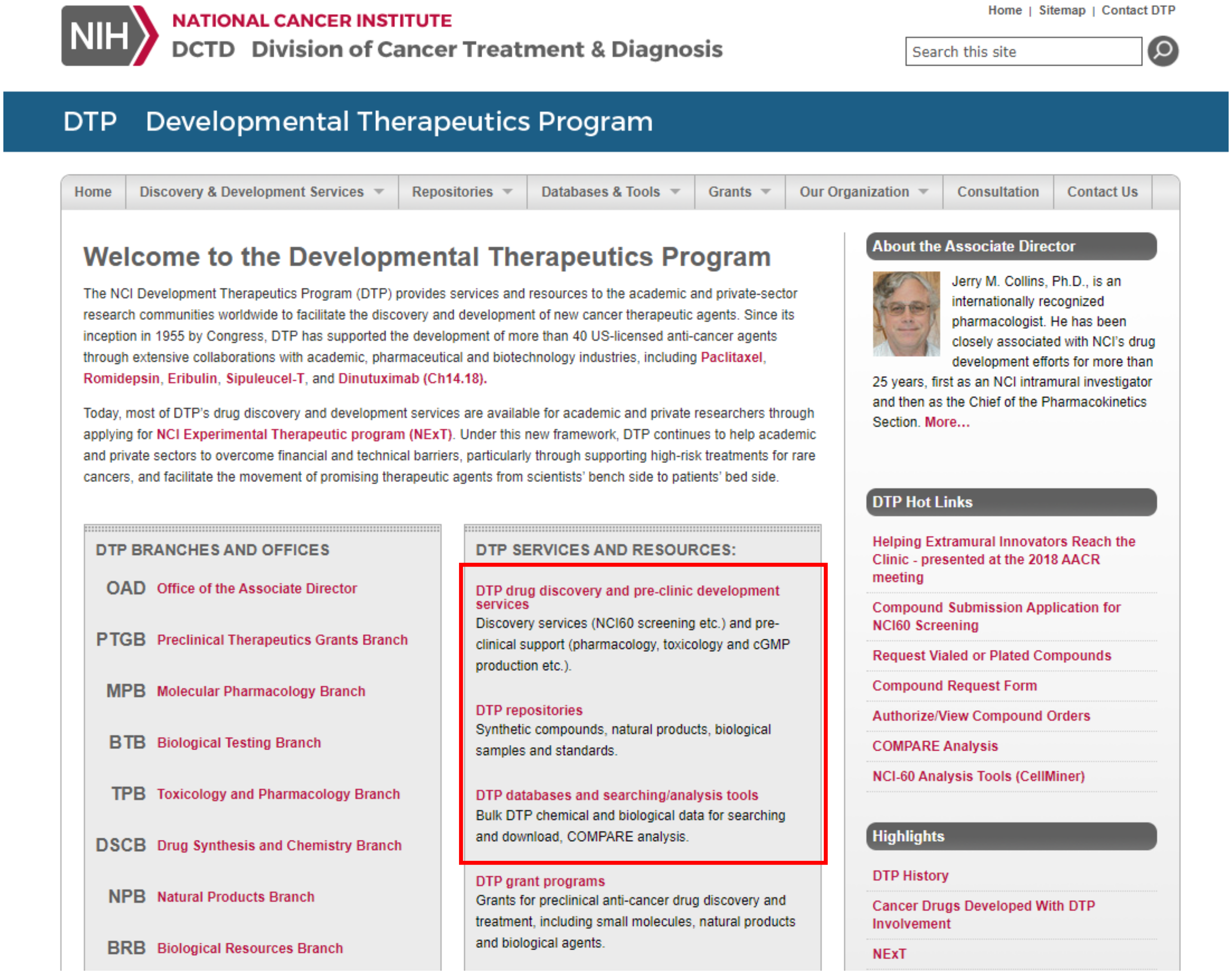


[Kirkpatrick & Ellis, Nature 2004]

Compounds Screening

A huge matrix of cell types and compound species to evaluate their biological effects have been accumulated through massive experimental assays.

Cell-line Screening Data Sets


NIH
NATIONAL CANCER INSTITUTE
DCTD Division of Cancer Treatment & Diagnosis
Home | Sitemap | Contact DTP
Search this site
DTP Developmental Therapeutics Program
Home
Discovery & Development Services
Repositories
Databases & Tools
Grants
Our Organization
Consultation
Contact Us

Welcome to the Developmental Therapeutics Program

The NCI Development Therapeutics Program (DTP) provides services and resources to the academic and private-sector research communities worldwide to facilitate the discovery and development of new cancer therapeutic agents. Since its inception in 1955 by Congress, DTP has supported the development of more than 40 US-licensed anti-cancer agents through extensive collaborations with academic, pharmaceutical and biotechnology industries, including Paclitaxel, Romidepsin, Eribulin, Sipuleucel-T, and Dinutuximab (Ch14.18).

Today, most of DTP's drug discovery and development services are available for academic and private researchers through applying for NCI Experimental Therapeutic program (NExT). Under this new framework, DTP continues to help academic and private sectors to overcome financial and technical barriers, particularly through supporting high-risk treatments for rare cancers, and facilitate the movement of promising therapeutic agents from scientists' bench side to patients' bed side.

DTP BRANCHES AND OFFICES

OAD	Office of the Associate Director
PTGB	Preclinical Therapeutics Grants Branch
MPB	Molecular Pharmacology Branch
BTB	Biological Testing Branch
TPB	Toxicology and Pharmacology Branch
DSCB	Drug Synthesis and Chemistry Branch
NPB	Natural Products Branch
BRB	Biological Resources Branch

DTP SERVICES AND RESOURCES:

DTP drug discovery and pre-clinic development services
Discovery services (NCI60 screening etc.) and pre-clinical support (pharmacology, toxicology and cGMP production etc.).

DTP repositories
Synthetic compounds, natural products, biological samples and standards.

DTP databases and searching/analysis tools
Bulk DTP chemical and biological data for searching and download, COMPARE analysis.

DTP grant programs
Grants for preclinical anti-cancer drug discovery and treatment, including small molecules, natural products and biological agents.

About the Associate Director

Jerry M. Collins, Ph.D., is an internationally recognized pharmacologist. He has been closely associated with NCI's drug development efforts for more than 25 years, first as an NCI intramural investigator and then as the Chief of the Pharmacokinetics Section. More...

DTP Hot Links

- Helping Extramural Innovators Reach the Clinic - presented at the 2018 AACR meeting
- Compound Submission Application for NCI60 Screening
- Request Vialed or Plated Compounds
- Compound Request Form
- Authorize/View Compound Orders
- COMPARE Analysis
- NCI-60 Analysis Tools (CellMiner)

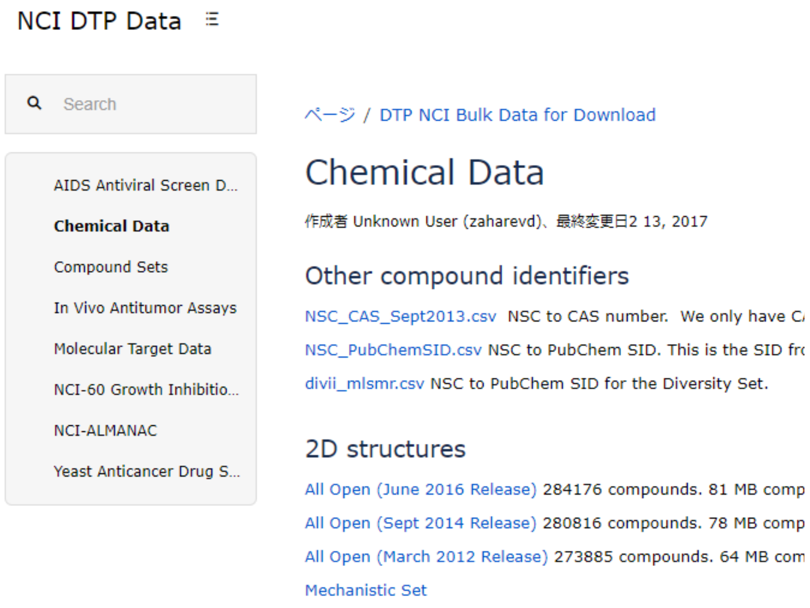
Highlights

- DTP History
- Cancer Drugs Developed With DTP Involvement
- NExT

We can get the data from the NIH-DTP web site

Data Preparation

Chemical structural data



NCI DTP Data

Search

- AIDS Antiviral Screen D...
- **Chemical Data**
- Compound Sets
- In Vivo Antitumor Assays
- Molecular Target Data
- NCI-60 Growth Inhibitio...
- NCI-ALMANAC
- Yeast Anticancer Drug S...

ページ / DTP NCI Bulk Data for Download

Chemical Data

作成者 Unknown User (zaharevd)、最終変更日2 13, 2017

Other compound identifiers

NSC_CAS_Sept2013.csv NSC to CAS number. We only have C

NSC_PubChemSID.csv NSC to PubChem SID. This is the SID fr

divii_mlsmr.csv NSC to PubChem SID for the Diversity Set.

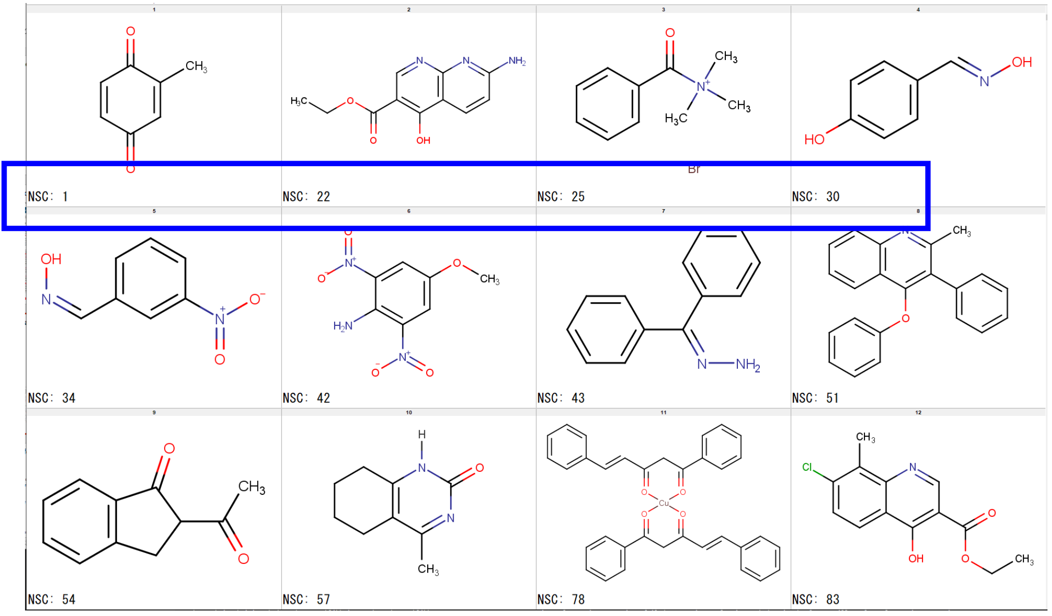
2D structures

All Open (June 2016 Release) 284176 compounds. 81 MB comp

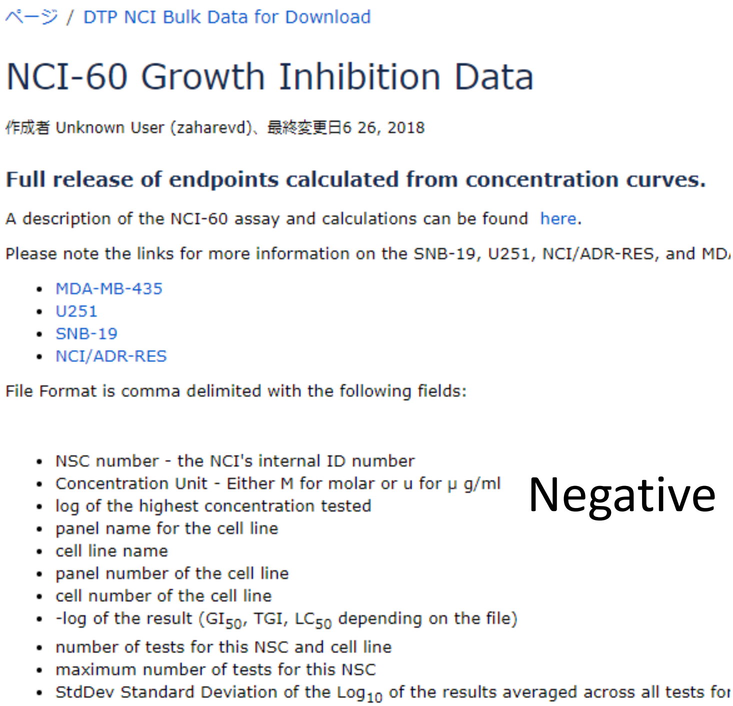
All Open (Sept 2014 Release) 280816 compounds. 78 MB comp

All Open (March 2012 Release) 273885 compounds. 64 MB com

Mechanistic Set



Assay data



ページ / DTP NCI Bulk Data for Download

NCI-60 Growth Inhibition Data

作成者 Unknown User (zaharevd)、最終変更日6 26, 2018

Full release of endpoints calculated from concentration curves.

A description of the NCI-60 assay and calculations can be found here.

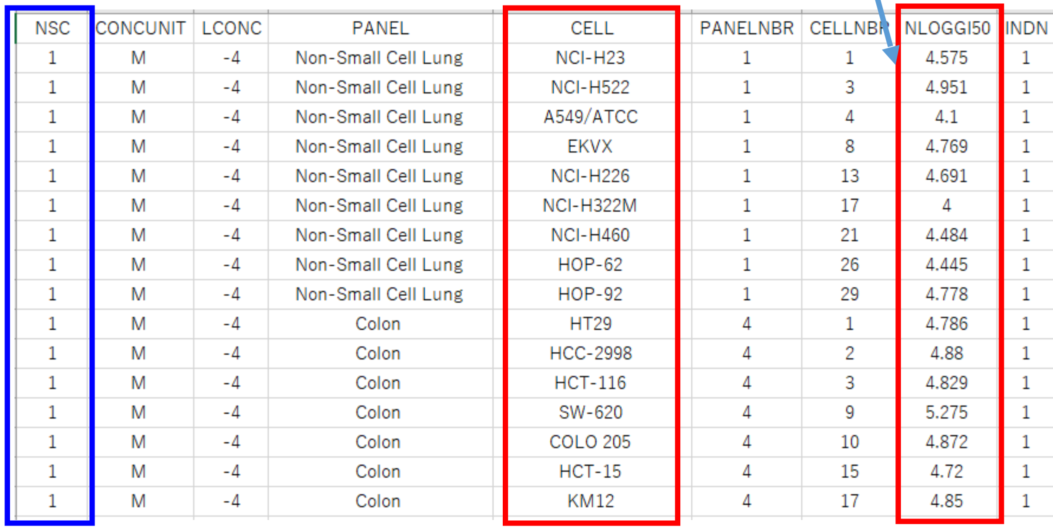
Please note the links for more information on the SNB-19, U251, NCI/ADR-RES, and MD

- MDA-MB-435
- U251
- SNB-19
- NCI/ADR-RES

File Format is comma delimited with the following fields:

- NSC number - the NCI's internal ID number
- Concentration Unit - Either M for molar or u for μ g/ml
- log of the highest concentration tested
- panel name for the cell line
- cell line name
- panel number of the cell line
- cell number of the cell line
- -log of the result (GI_{50}, TGI, LC_{50} depending on the file)
- number of tests for this NSC and cell line
- maximum number of tests for this NSC
- StdDev Standard Deviation of the Log_{10} of the results averaged across all tests for

Negative log(GI50)



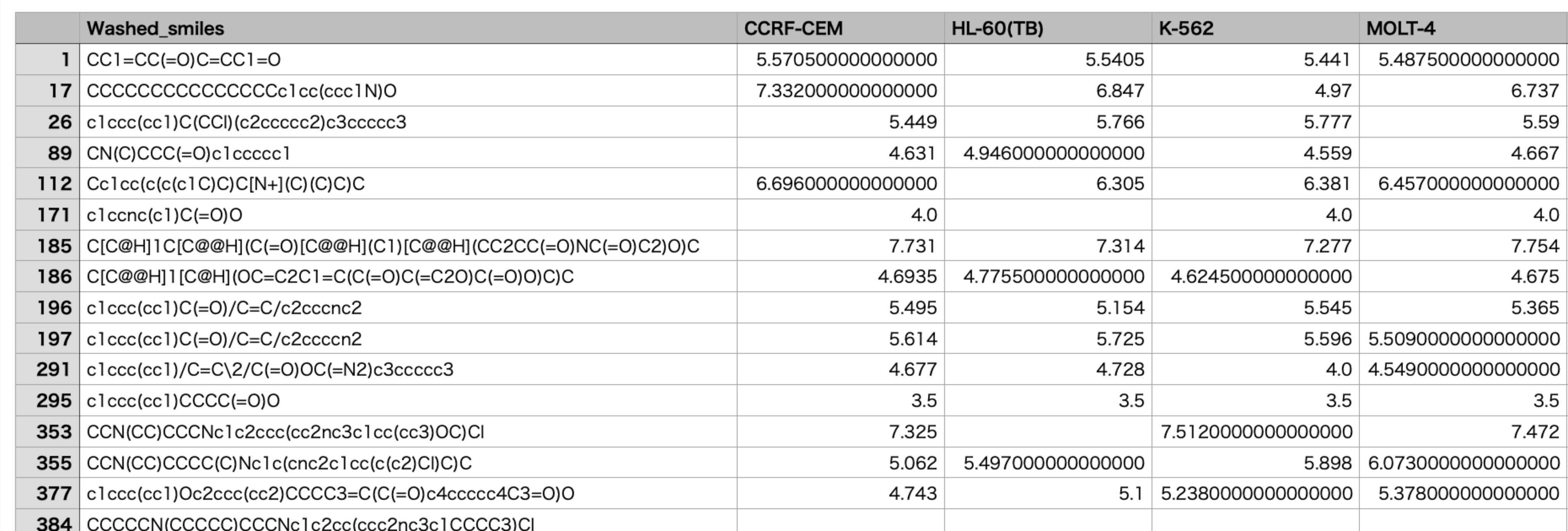
NSC	CONCUNIT	LCONC	PANEL	CELL	PANELNBR	CELLNBR	NLOGGI50	INDN
1	M	-4	Non-Small Cell Lung	NCI-H23	1	1	4.575	1
1	M	-4	Non-Small Cell Lung	NCI-H522	1	3	4.951	1
1	M	-4	Non-Small Cell Lung	A549/ATCC	1	4	4.1	1
1	M	-4	Non-Small Cell Lung	EKVX	1	8	4.769	1
1	M	-4	Non-Small Cell Lung	NCI-H226	1	13	4.691	1
1	M	-4	Non-Small Cell Lung	NCI-H322M	1	17	4	1
1	M	-4	Non-Small Cell Lung	NCI-H460	1	21	4.484	1
1	M	-4	Non-Small Cell Lung	HOP-62	1	26	4.445	1
1	M	-4	Non-Small Cell Lung	HOP-92	1	29	4.778	1
1	M	-4	Colon	HT29	4	1	4.786	1
1	M	-4	Colon	HCC-2998	4	2	4.88	1
1	M	-4	Colon	HCT-116	4	3	4.829	1
1	M	-4	Colon	SW-620	4	9	5.275	1
1	M	-4	Colon	COLO 205	4	10	4.872	1
1	M	-4	Colon	HCT-15	4	15	4.72	1
1	M	-4	Colon	KM12	4	17	4.85	1

GI50: 50 % Growth Inhibition, 細胞の増殖を50% 阻害する濃度

- **Dataset1**
  - A tab-separated text file named "pGI50_mols.tsv"
  - 51178 samples, 61 cell lines
  - With values of pGI50 (negative log GI50)

https://royalsocietypublishing.org/doi/pdf/10.1098/rstb.2018.0226

(Roughly speaking, higher pGI50 implies strong toxicity)



	Washed_smiles	CCRF-CEM	HL-60(TB)	K-562	MOLT-4
1	CC1=CC(=O)C=CC1=O	5.570500000000000	5.5405	5.441	5.487500000000000
17	CCCCCCCCCCCCCCc1cc(ccc1N)O	7.332000000000000	6.847	4.97	6.737
26	c1ccc(cc1)C(CCl)(c2ccccc2)c3ccccc3	5.449	5.766	5.777	5.59
89	CN(C)CCC(=O)c1ccccc1	4.631	4.946000000000000	4.559	4.667
112	Cc1cc(c(c(c1C)C)C[N+](C)(C)C)C	6.696000000000000	6.305	6.381	6.457000000000000
171	c1ccnc(c1)C(=O)O	4.0		4.0	4.0
185	C[C@H]1C[C@@H](C(=O)[C@@H](C1)[C@@H](CC2CC(=O)NC(=O)C2)O)C	7.731	7.314	7.277	7.754
186	C[C@@H]1[C@H](OC=C2C1=C(C(=O)C(=C2O)C(=O)O)C)C	4.6935	4.775500000000000	4.624500000000000	4.675
196	c1ccc(cc1)C(=O)/C=C/c2cccnc2	5.495	5.154	5.545	5.365
197	c1ccc(cc1)C(=O)/C=C/c2ccccn2	5.614	5.725	5.596	5.509000000000000
291	c1ccc(cc1)/C=C\2/C(=O)OC(=N2)c3ccccc3	4.677	4.728	4.0	4.549000000000000
295	c1ccc(cc1)CCCC(=O)O	3.5	3.5	3.5	3.5
353	CCN(CC)CCCNc1c2ccc(cc2nc3c1cc(cc3)OC)Cl	7.325		7.512000000000000	7.472
355	CCN(CC)CCCC(C)Nc1c(cnc2c1cc(c(c2)Cl)C)C	5.062	5.497000000000000	5.898	6.073000000000000
377	c1ccc(cc1)Oc2ccc(cc2)CCCC3=C(C(=O)c4ccccc4C3=O)O	4.743	5.1	5.238000000000000	5.378000000000000
384	CCCCCN(CCCCC)CCCNc1c2cc(ccc2nc3c1CCCC3)Cl				

- **Dataset1**
  - Analyze the dataset to understand the distribution of chemical features

  - Train machine learning models and predict toxicity from molecular features

    # You may choose some of the cell lines, or, may use all targets.

    ## Note that not all combinations have been evaluated

  - Try various types of molecular descriptors, and various models of machine learning.

  - Find the relationship between molecular features and biological activities

Example programs

You can access the notebook of the sample program from

https://colab.research.google.com/drive/1ryR7DpuXAO_mMjWzCjfT-r_mf6Gb1bco?usp=sharing

Dataset2:
Antibiotics Screening

- **Dataset2**
  - Part1: mechanism of action

    From KEGG DRUG database (https://www.genome.jp/kegg/drug/)

    List of antibiotic compounds classified by mechanism of actions

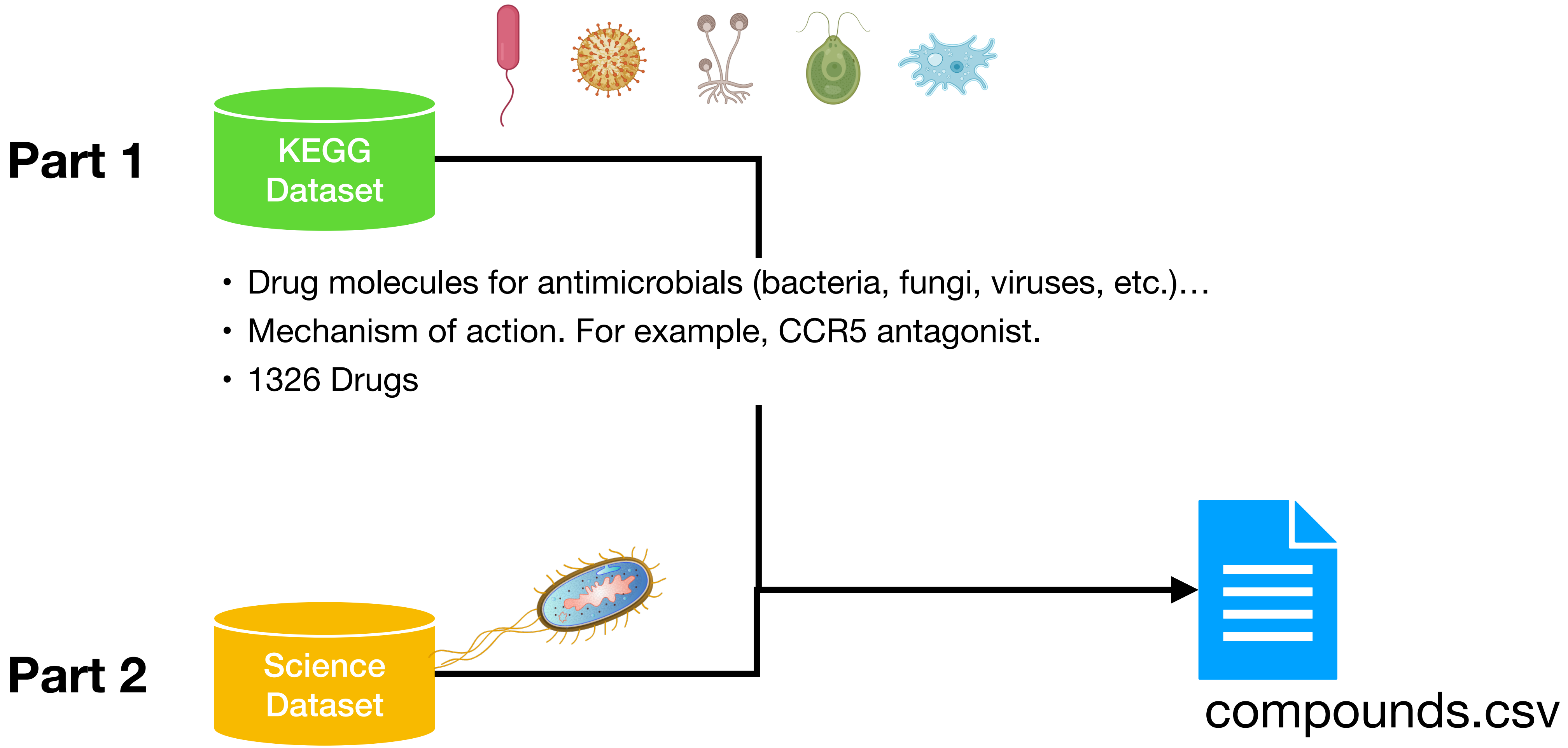
  - Part2: inhibition ratio

    From "A Deep Learning Approach to Antibiotic Discovery"

    (JM Stokes et al, Cell, 2020, https://doi.org/10.1016/j.cell.2020.01.021)

    Experimental dataset for screening of new antibiotics

Dataset 2 (Antibiotic Candidates)



Part 1

KEGG Dataset

- Drug molecules for antimicrobials (bacteria, fungi, viruses, etc.)…
- Mechanism of action. For example, CCR5 antagonist.
- 1326 Drugs

Part 2

Science Dataset

compounds.csv

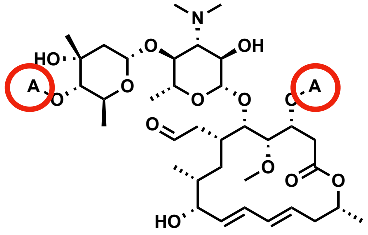
- Growth inhibition ratio against E.coli
- 2,335 diverse molecules were tested

Workflow of Curation 2024. 07. 11

compounds.csv
3,427 entries

Drop polymers and molecules with undetermined atoms
Peginterferon alfa-2b : *C(=O)OCCOC (combination of PEG and Interferon)
Kitasamicyn :[1*]O[C@@H]1CC(=O)O[C@H](C)C/C=C/C=C/[C@H](O)[C@H](C)C[C@H](CC=O)[C@H](O[C@@H]2O[C@H](C)[C@@H](O[C@H]3C[C@@](C)(O)[C@@H](O[2*])[C@H](
[C@H]1OC

3,407 entries



Arbitrary atoms (or representing further connection)

Standardization
1. Remove Salts
2. Neutralize molecules

Merge Inhibition & Class (if several standardized smiles exist) and Drop duplicates.
If several inhibition values are assigned to a single SMILES, the averaged one is used

3,407 entries

standardized_compounds.tsv
2941 Unique SMILES

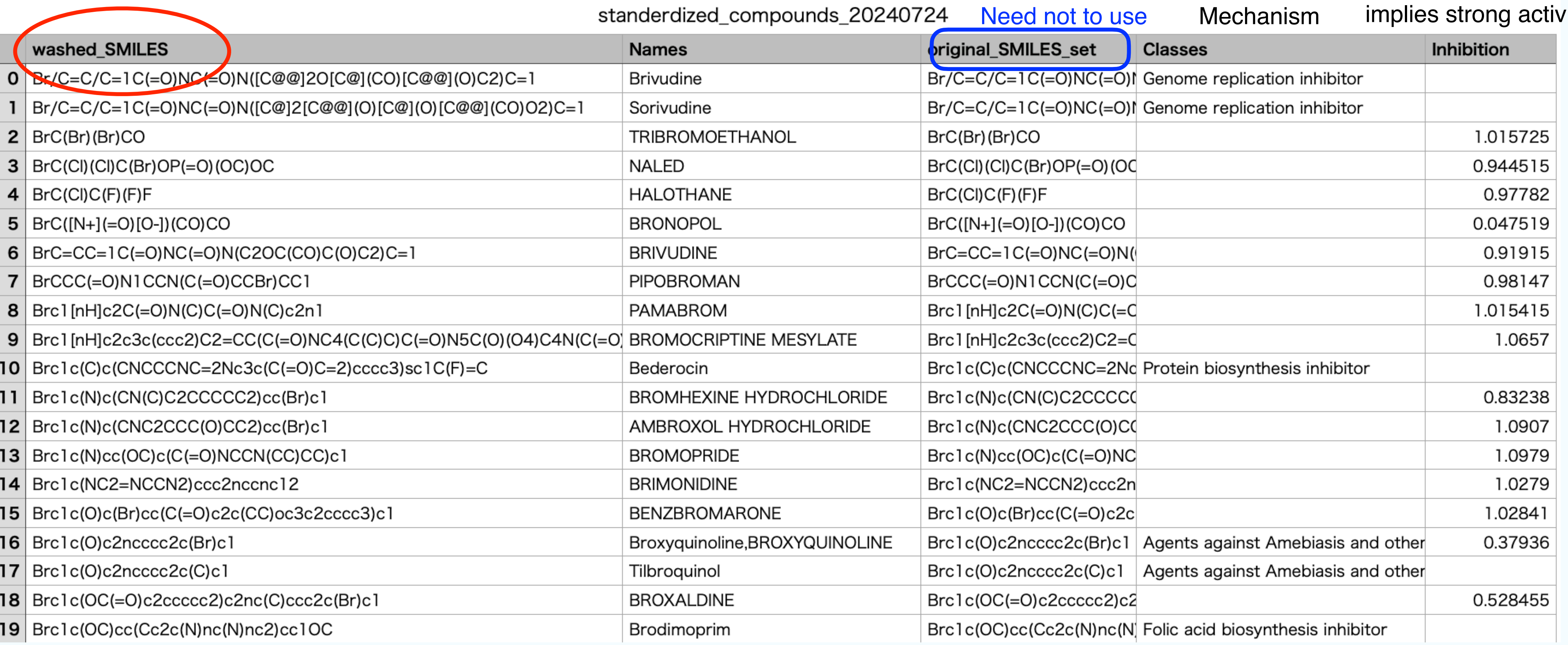
- **Dataset2:**
  - 748 samples with 13 classes
  - 2289 samples with inhibition (relative growth) values

Use these SMILES

Need not to use

Mechanism

(Roughly speaking, lower "Inhibition" implies strong activit



standerdized_compounds_20240724

	washed_SMILES	Names	original_SMILES_set	Classes	Inhibition
0	Br/C=C/C=1C(=O)NC(=O)N([C@@]2O[C@](CO)[C@@](O)C2)C=1	Brivudine	Br/C=C/C=1C(=O)NC(=O)N	Genome replication inhibitor	
1	Br/C=C/C=1C(=O)NC(=O)N([C@]2[C@@](O)[C@](O)[C@@](CO)O2)C=1	Sorivudine	Br/C=C/C=1C(=O)NC(=O)N	Genome replication inhibitor	
2	BrC(Br)(Br)CO	TRIBROMOETHANOL	BrC(Br)(Br)CO		1.015725
3	BrC(Cl)(Cl)C(Br)OP(=O)(OC)OC	NALED	BrC(Cl)(Cl)C(Br)OP(=O)(OC		0.944515
4	BrC(Cl)C(F)(F)F	HALOTHANE	BrC(Cl)C(F)(F)F		0.97782
5	BrC([N+](=O)[O-])(CO)CO	BRONOPOL	BrC([N+](=O)[O-])(CO)CO		0.047519
6	BrC=CC=1C(=O)NC(=O)N(C2OC(CO)C(O)C2)C=1	BRIVUDINE	BrC=CC=1C(=O)NC(=O)N(		0.91915
7	BrCCC(=O)N1CCN(C(=O)CCBr)CC1	PIPOBROMAN	BrCCC(=O)N1CCN(C(=O)C		0.98147
8	Brc1[nH]c2C(=O)N(C)C(=O)N(C)c2n1	PAMABROM	Brc1[nH]c2C(=O)N(C)C(=O		1.015415
9	Brc1[nH]c2c3c(ccc2)C2=CC(C(=O)NC4(C(C)C)C(=O)N5C(O)(O4)C4N(C(=O)	BROMOCRIPTINE MESYLATE	Brc1[nH]c2c3c(ccc2)C2=C		1.0657
10	Brc1c(C)c(CNCCCNC=2Nc3c(C(=O)C=2)cccc3)sc1C(F)=C	Bederocin	Brc1c(C)c(CNCCCNC=2Nc	Protein biosynthesis inhibitor	
11	Brc1c(N)c(CN(C)C2CCCCC2)cc(Br)c1	BROMHEXINE HYDROCHLORIDE	Brc1c(N)c(CN(C)C2CCCCC		0.83238
12	Brc1c(N)c(CNC2CCC(O)CC2)cc(Br)c1	AMBROXOL HYDROCHLORIDE	Brc1c(N)c(CNC2CCC(O)CC		1.0907
13	Brc1c(N)cc(OC)c(C(=O)NCCN(CC)CC)c1	BROMOPRIDE	Brc1c(N)cc(OC)c(C(=O)NC		1.0979
14	Brc1c(NC2=NCCN2)ccc2nccnc12	BRIMONIDINE	Brc1c(NC2=NCCN2)ccc2n		1.0279
15	Brc1c(O)c(Br)cc(C(=O)c2c(CC)oc3c2cccc3)c1	BENZBROMARONE	Brc1c(O)c(Br)cc(C(=O)c2c		1.02841
16	Brc1c(O)c2ncccc2c(Br)c1	Broxyquinoline,BROXYQUINOLINE	Brc1c(O)c2ncccc2c(Br)c1	Agents against Amebiasis and other	0.37936
17	Brc1c(O)c2ncccc2c(C)c1	Tilbroquinol	Brc1c(O)c2ncccc2c(C)c1	Agents against Amebiasis and other	
18	Brc1c(OC(=O)c2ccccc2)c2nc(C)ccc2c(Br)c1	BROXALDINE	Brc1c(OC(=O)c2ccccc2)c2		0.528455
19	Brc1c(OC)cc(Cc2c(N)nc(N)nc2)cc1OC	Brodimoprim	Brc1c(OC)cc(Cc2c(N)nc(N)	Folic acid biosynthesis inhibitor	

- **Dataset2**
  - Analyze the dataset to understand the distribution of chemical features.
  - Train machine learning models and predict new antibiotics.
  - You may apply clustering using **unsupervised learning** to group the molecules.
  - Alternatively, you may build

    a **classification** model using the Part 1 dataset to classify the molecules.

    a **regression** model with the Part 2 data to predict the antibiotic activity.
  - You may also predict using some new compound datasets using your trained model.
  - After conducting your analysis, add insights from both biological and chemical perspectives.

- **Group Presentations**
  - 9/27 (Fri.) 9:20-12:30 (IS：AI lecutre room)
  - Presentation: 15 min (talk) + 3 min (discussion)
  - Slides: in English
  - Talk: either in English or Japanese

Order	1	2	3	4	5	6
Group						

- **Report Task:**
  - Summarize the works of your group.

    Background and motivation

    Materials and methods

    Results and discussion

    References
  - Explain your contribution to the group.

    Describe “when” (date or period) and "what" you've done explicitly.

    Any type of contributions would be OK

    (i.g. implementation, gathering new data, data cleansing, active suggestion in discussion, evaluation of results, etc…)
  - Submission

    up to A4 2 pages (excluding figures and references)

    Submit by 10/11(Fri.) via Educational Affairs Portal (UNIPA)