PBL-I

Data Science for Chemistry Using Covid-19 Data Sets

2021/7/27

Schedule

- **7/27 Introduction < Now**
  - How to describe chemical compounds
- **Group work**
  - Determine objective(s)
  - Build machine learning models, if necessary.
  - Discuss the outcome of your results
  - Prepare the presentation materials
  - 8 groups / Log in to Slack workspace
- **9/28 Presentations & Discussion**
  - 9:20-12:30 L3 / Zoom
  - Presentation: 10 min (talk) + 3 min (discussion)
  - Presentation materials = English / talk = (English or Japanese)
- **Report**

Report

- Task:
  1. Summarize the works of your group.
     - Background and motivation
     - Materials and methods
     - Results
     - Discussion and evaluation
     - References
  2. Explain your contribution in the group.

     Describe “when” (date or period) explicitly (thus, regular meetings are recommended).

     Any types of contributions would be OK

     (i.g. implementation, gathering new data, data cleansing, active suggestion in discussion, evaluation of results etc…)
  - A4 2 pages (excluding figures and references)

A template doc file is available on the PBL web page

Report

- Deadline： 10/12 (Tue) 23:59

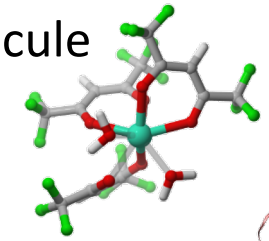
- Contents:
  A4 2 pages (PDF file)
  A template doc file is available at our web site.
  `http://www-dsc.naist.jp/dsc_jp/index.php/dsc-pbl/`

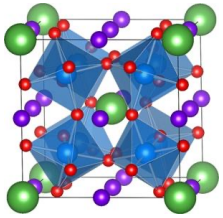
- How to submit
  Upload to the NAIST report submission service
  Report box name:
    データサイエンスPBL I / Data Science PBL I [ID: 5013]
  File name:
    `{Student ID}_{LastName}_{FirstName}.pdf`

Data Science in Chemistry

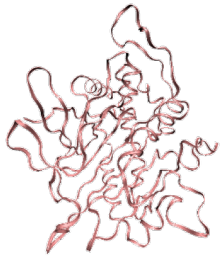
Compounds 化合物

Molecule





Solid



Polymer

AI

Learn the correlation

Functions 機能

Catalytic ability

Conductive property

Photo activity

Thermal durability

etc

INPUT

Unknown material

OUTPUT

Function (ON/OFF or intensity)

Point!

Compounds need to be represented as numerical numbers!

化合物を数値・ベクトル等で表現する必要あり

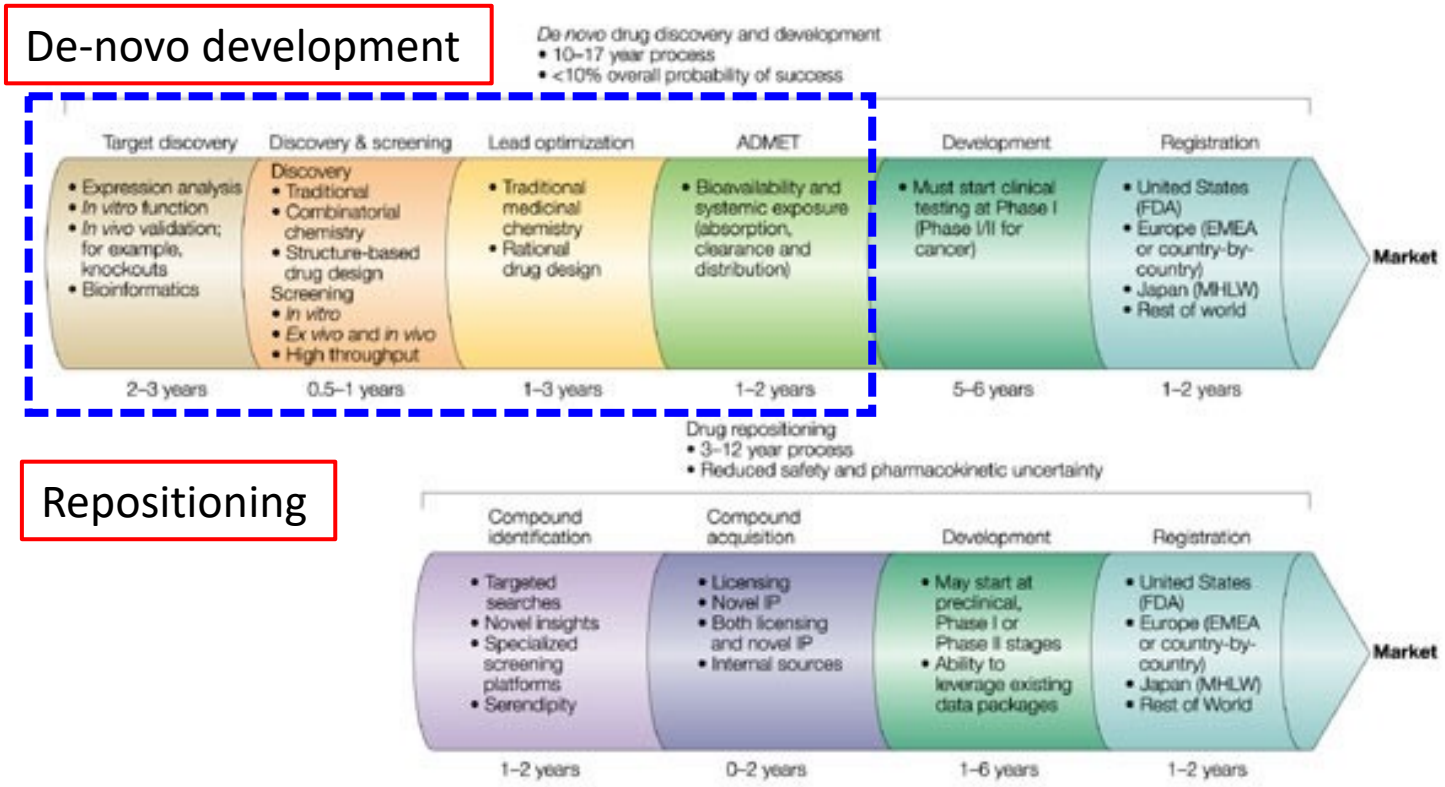
COVID-19 Outbreak

- The most impactful viral disease of the 21st century
  - Declared *pandemic* on 11th March 2020 by the WHO.
  - Affected more than 200 countries with millions of cases.
  - As of 23 July, around 200 million cases were reported and more than 4 million people died.
- Severe acute respiratory syndrome coronavirus-2(SARS-Cov-2) is the virus that causes COVID-19
- The virus has been massively studied
  - Infection mechanism (protein-protein interaction with ACE2)
  - SARS-Cov-2 main protease structure (PDB:6LU7)
  - **It would need some time to develop drugs based on mechanism…**

We need drugs against COVID-19 as soon as possible.

Drug Development Approach for COVID-19

- Drug repositioning (repurposing)
  - Reuse currently approved (clinically tested) drugs against COVID-19.
  - We can skip time-consuming clinical trials and other experiments.



T. Ashburn and KB. Thor, Nature Reviews Drug Discovery **3**, 673–683 (2004)

Currently Investigated Drugs (Repositioning)

Pharmacological class	Drug	Proposed mechanism in the treatment of SARS-CoV-2 infection
Kinase inhibitors	Baricitinib	It could exert anti-viral effects by its affinity for AP2-associated protein AAK1, reducing SARS-CoV-2 endocytosis
	Imatinib	It accumulates in lysosomes resulting in some antiviral activities by lysosomal alkalization required for virus/cell fusion
Antibacterials	Doxycycline	It could reduce pro-inflammatory cytokines levels and chelate matrix metalloproteinases used for cell fusion and viral replication
Antidiabetic drugs	Dapagliflozin	During virus infection, serum lactate dehydrogenase level excessively rises. Dapagliflozin has been reported to reduce lactate levels by various mechanisms. It also reduces oxygen consumption in tissues and causes the use of glucose in the erobic pathway
	Linagliptin, sitagliptin	Since SARS-CoV-2 could use DPP4 receptor to invade cells, the inhibition of DPP4 could be useful in mild COVID-19 patients
Antimalarials	Artemisinin/artesunate	Anti-inflammatory activity, NF-κB-coronavirus effect and chloroquine-like endocytosis inhibition mechanism
	Atovaquone	It could inhibit SARS-CoV-2 through targeting of the viral RdRp or 3C-like protease
	Chloroquine, hydroxychloroquine	They showed interference with the glycosylation of ACE-2 receptors; they increase the pH of acidic cellular organelles, counteracting virus replication
	Mefloquine	It inhibited SARS-CoV-2 replication *in vitro* experimental models
Antitumorals	Plitidepsin	It could inhibit the multiplication and propagation of SARS-CoV-2
	Selinexor	It could inhibit the replication of SARS-CoV-2 and mediate anti-inflammatory and anti-viral effects
Antivirals	Atazanavir, danoprevir, darunavir	Potential SARS-CoV-2 protease inhibition
	Clevudine	It acts as a potent inhibitor of RdRp protein, preventing RNA replication
	Daclatasvir	It could target different proteins of the SARS-CoV-2 life cycle, affecting both viral RNA replication and virion assembly
	Emtricitabine	RNA synthesis nucleos(t)ide analogue inhibitors could have an effect against SARS-CoV-2 infection
	Favipiravir, galidesivir	They inhibit RdRp of RNA viruses, blocking SARS-CoV-2 replication
	Lopinavir/ritonavir	They could inhibit SARS-CoV-2 replication by blocking $3CL^{pro}$ and $PL2^{pro}$ proteases
	Nelfinavir	It may bind to the S trimer structure inhibiting the membrane fusion process
	Nitazoxanide	It exerts antiviral effects through the phosphorylation of protein kinase activated by double-stranded RNA, which leads to an increase in phosphorylated factor 2-alpha, an intracellular protein with antiviral effects
	Oseltamivir	It could inhibit virus replication and virion release
	Remdesivir	It could inhibit the RNA synthesis of SARS-CoV-2
	Ribavirin	It could inhibit SARS-CoV-2 replication
	Sofosbuvir	It is a chain terminator for SARS-CoV-2 RNA polymerase. In human brain organoids, it protected from SARS-CoV-2-induced cell death
	Tenofovir alafenamide	It could inhibit SARS-CoV-2 RdRp
	Umifenovir	It could block trimerization of the spike glycoprotein, essential for host cell adhesion
Immunosuppressants	Cyclosporine	It can block viral replication and thus transcription of pro-inflammatory cytokines
	Leflunomide	In *vitro* studies have shown antiviral effects of leflunomide against SARS-CoV-2
	Sirolimus	It could block viral protein expression and virion release
	Tacrolimus	
Interferons	Alpha and beta interferons	
	Peginterferon lambda-1A	natural immune system into action

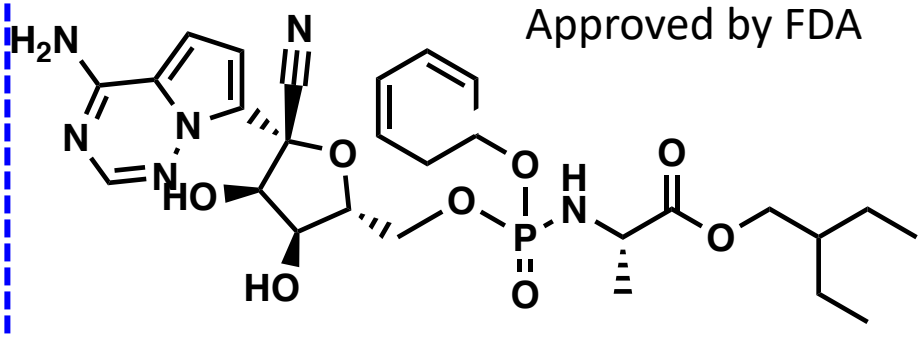
and more…

As of 2020 July.

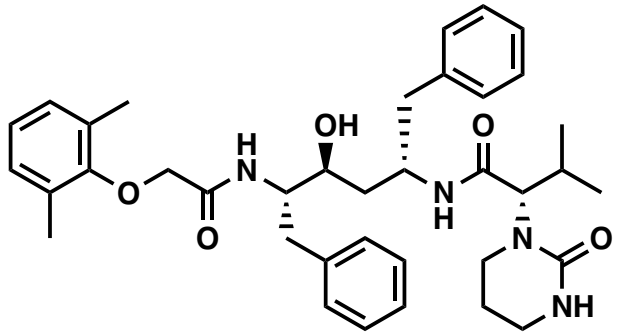
From Sultana et al., Front. Pharmacol., 06, 2020 | https://doi.org/10.3389/fphar.2020.588654

Drug Candidates for COVID-19

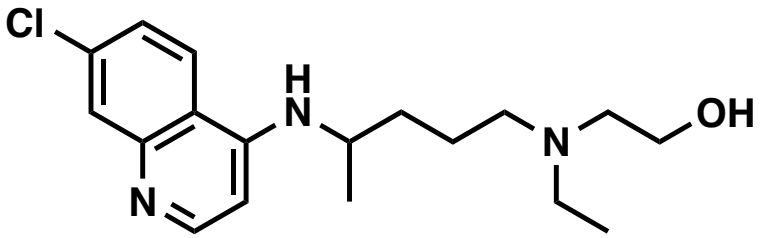
Approved by FDA



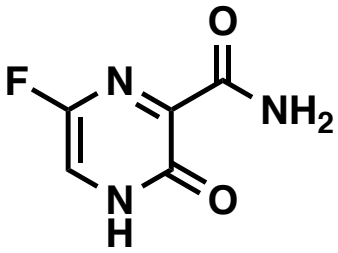
Remdesivir (hepatitis C treatment)



Lopinavir (HIV treatment)



Hydroxychloroquine (Malaria treatment)

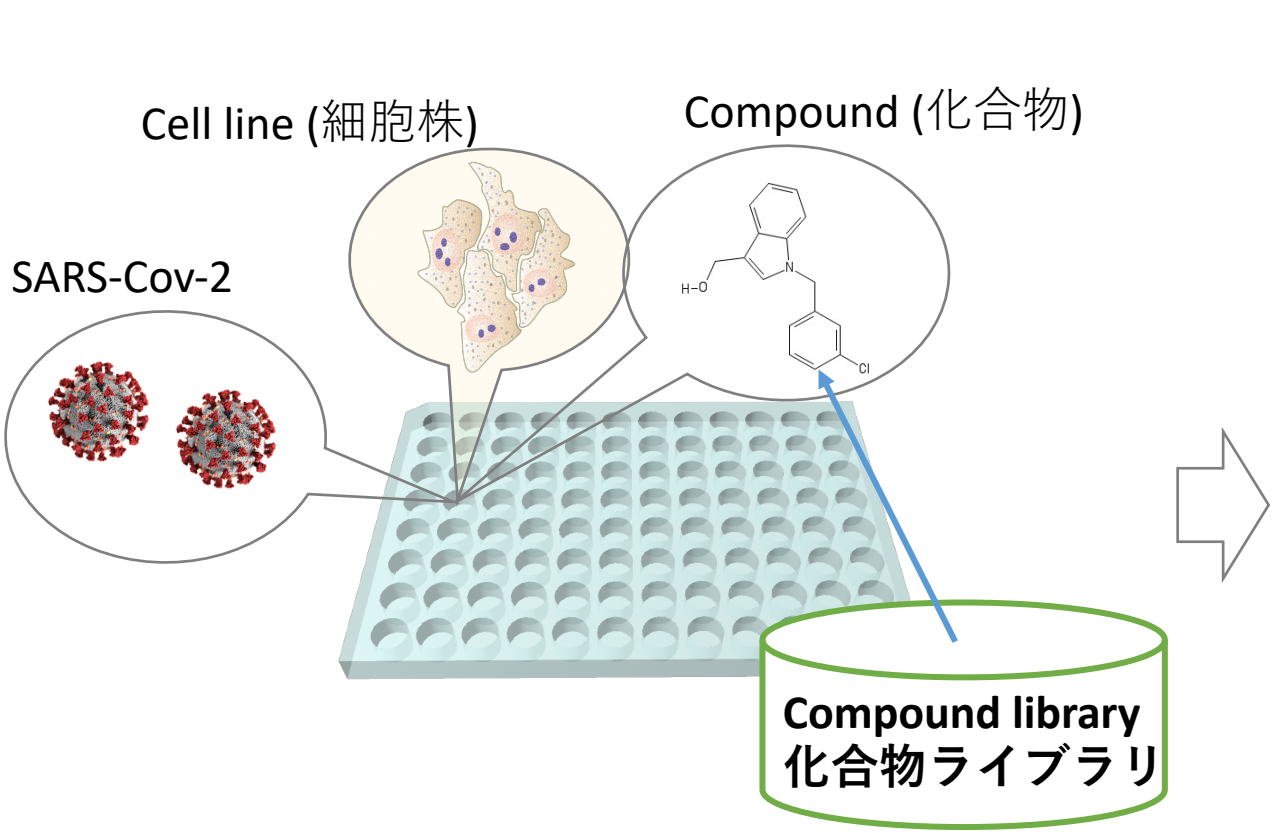


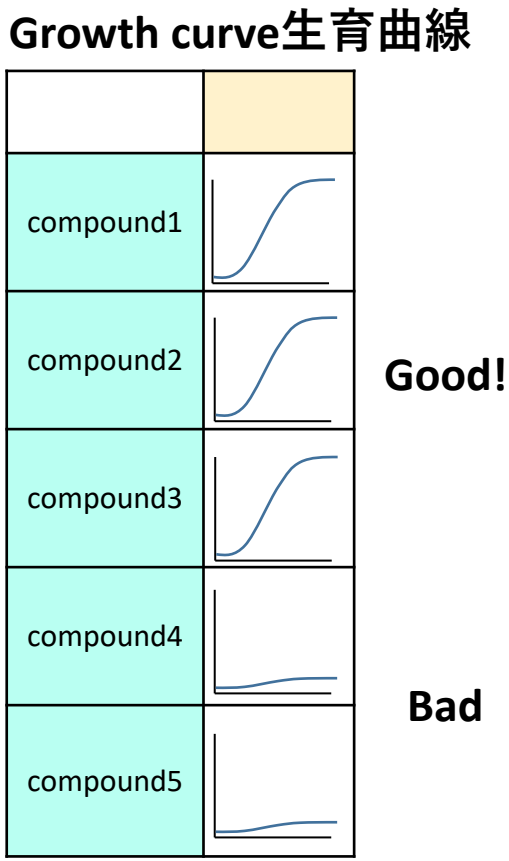
Favipiravir: Avigan (Influenza treatment)

In vitro Testing for Drug Repositioning

The cytopathic effect (CPE) reduction assay

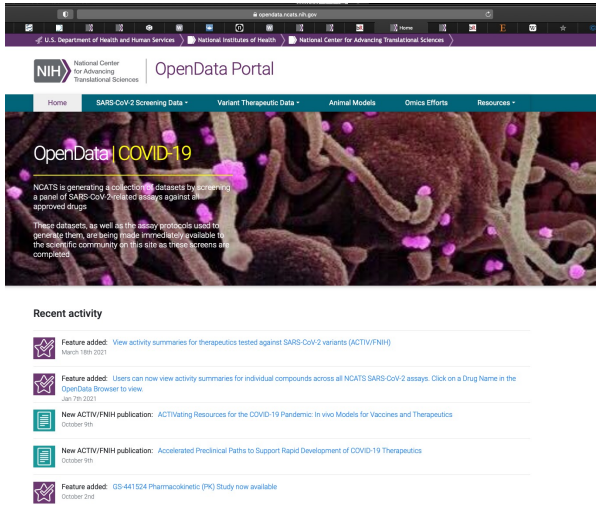
- Widely used assay format to screen for antiviral agents
- Easy to measure the effects of compounds
- The effectiveness of the compounds can be evaluated by the host cell viability.
- Indirectly monitoring the compound ability to inhibit viral infection and replication.

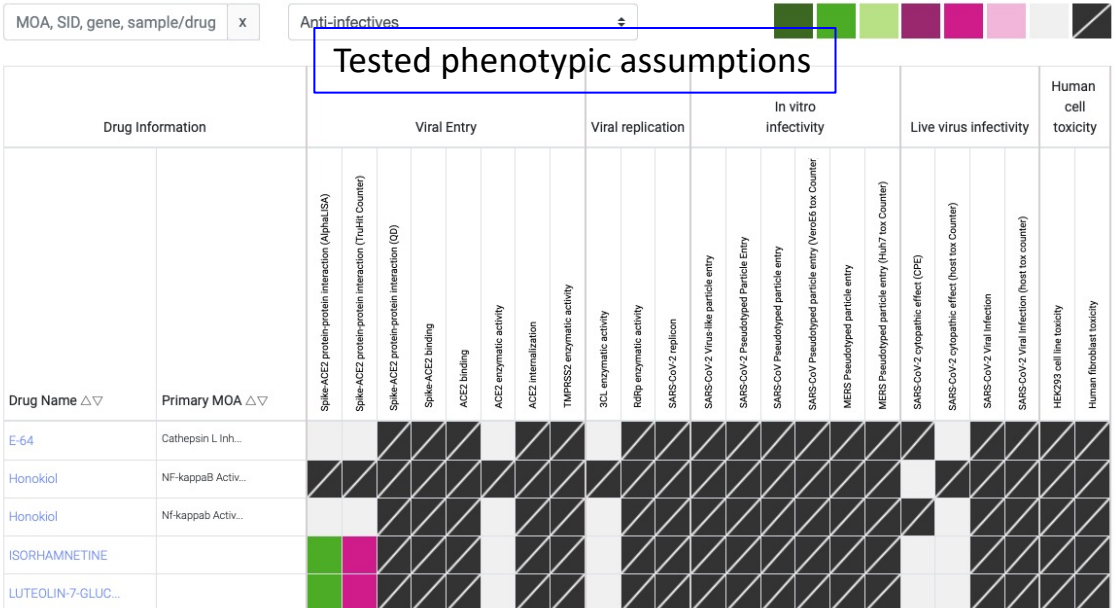




Data Sets for COVID-19

- NIH (National Center for Advancing Translational Sciences)
  - Lunch portal site for COVID-19 related databases (OpenData Portal) https://opendata.ncats.nih.gov/covid19
  - Not only screening data but other data set types, such as multi-omics data in COVID-19 are collected.



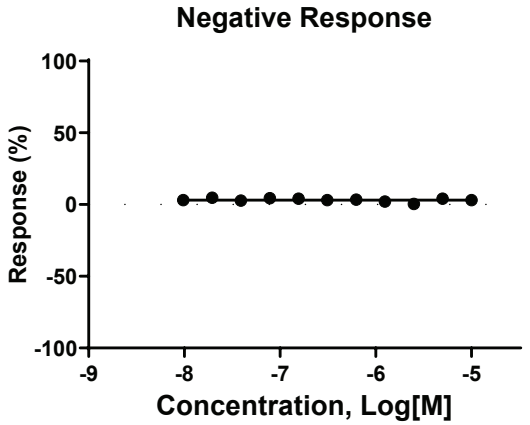


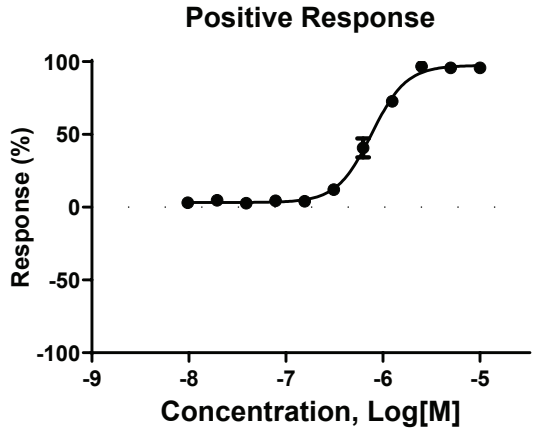
Data Sets of Cytopathic Effect Assay

- High throughput experimental data sets (HTE) for searching candidate drugs against COVID-19.

Assay description

Target Category	Live virus infectivity
Target	Viral infection and replication
Assay Type	Cell viability
Cell Line	Vero E6
Detection Type	Luminescence
Date Screened	2020-04-29
Throughput	384-well
Positive control	Cells without SARS-Cov-2
Negative control	DMSO





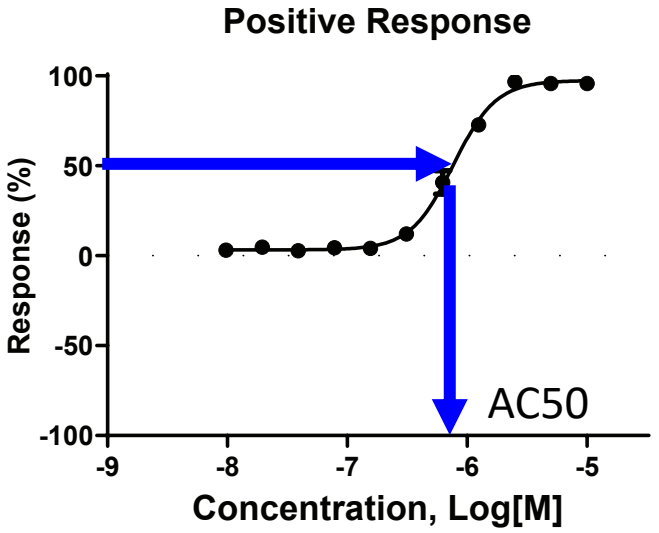
https://opendata.ncats.nih.gov/covid19/assay?aid=14

Data Sheet (cytopathic effect data)

	A	B	G	I	J	K
1	SAMPLE_ID	SAMPLE_NAME	CAS	AC50(uM)	LOG_AC50	AUC
2	NCGC00166226-02	4-Propylbenzoic acid	2438/05/03			7.87015726
3	NCGC00166271-02	Pregnenolone acetat	1778-02-5			-0.5120982
4	NCGC00387409-02	Ertapenem	153832-46-3			-3.6504244
5	NCGC00378604-01	ABT-333	1132935-63-	12.5892541	-4.9	39.4890464
6	NCGC00181030-02	Uracil	66-22-8			-2.0335411
7	NCGC00181159-03	Trabectedin	114899-77-3			15.9259229
8	NCGC00159339-10	Streptomycin sulfate	3810-74-0			-5.3577235
10	NCGC00521082-01	Mebezonium Iodide	7681-78-9			-4.6265005
11	NCGC00188443-01	Dapiprazole hydrochl	72822-13-0			-0.7015852
12	NCGC00250402-09	Dapagliflozin	461432-26-8			1.27343879

Blank cells
Inactive compounds

Active compound

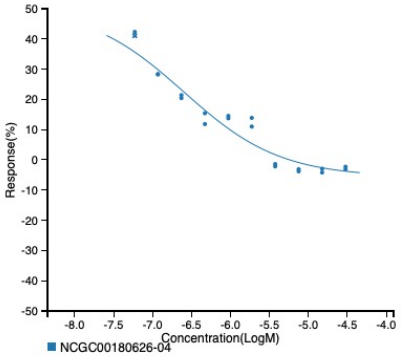


Example:
AC50 = 5.0 X 10^{-7} [M]
pAC50 = $-\log_{10}(5.0 \times 10^{-7})$
= 6.3

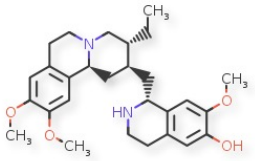
The higher a pAC50 value is, the stronger the cytopathic reduction ability of the compound is.

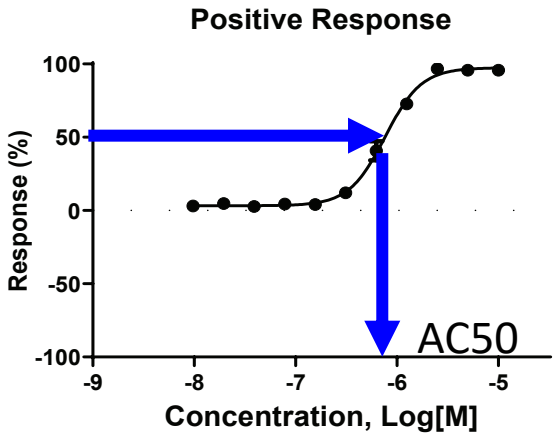
Data Curation for Assay Results

Sample Name:	N/A
Area under the curve (AUC) :	77.606
AC50 (μM) :	0.251
Efficacy (%) :	-56.310
Curve Class:	-1.2
Assay Name :	SARS-CoV-2 cytopathic effect (CPE)
Library :	Annotated/Bioactive Compound Collection
Primary Mechanism of Action (MOA) :	N/A
Curve:	



AC50 is assigned, but this compound reduced cell activity (killing host cells)





We have to curate the data so as to

- identify “truly” active compounds
- remove inconsistent assay results if multiple assays were conducted
  - E.g. inconsistent AC50 values

“MDL”: a standard format for chemical structure

✓ MDL format

- A standard format
- Cartesian coordinates of atoms
- Connectivity between atoms

・各原子のXYZ座標と
原子間の結合の様子で分子を表現

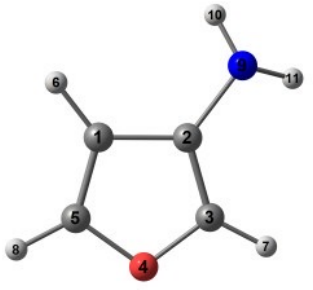
✓ SDF file

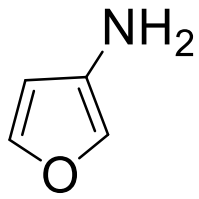
- Series of MOL format
- Properties can be also included

・複数の化合物のMOLを集めたもの
・化合物の性質も書き込める

Title-1 | Cartesian Coordinate | Atom Name

```
11 11  0  0  0  0  0  0  0  0999 V2000
   -0.3985    1.2800   -5.6905 C   0  0  0  0  0  0  0  0  0  0  0  0
   -0.3985    0.2517   -4.6890 C   0  0  0  0  0  0  0  0  0  0  0  0
   -0.3985    0.8935   -3.4894 C   0  0  0  0  0  0  0  0  0  0  0  0
   -0.3985    2.2456   -3.6705 O   0  0  0  0  0  0  0  0  0  0  0  0
   -0.3985    2.4623   -5.0173 C   0  0  0  0  0  0  0  0  0  0  0  0
   -0.3985    1.1510   -6.7631 H   0  0  0  0  0  0  0  0  0  0  0  0
   -0.3985    0.5573   -2.4644 H   0  0  0  0  0  0  0  0  0  0  0  0
   -0.3985    3.4958   -5.3264 H   0  0  0  0  0  0  0  0  0  0  0  0
   -0.3985   -1.2026   -4.9030 N   0  0  0  0  0  0  0  0  0  0  0  0
   -0.7770   -1.5012   -5.7170 H   0  0  0  0  0  0  0  0  0  0  0  0
   -0.7770   -1.7446   -4.1381 H   0  0  0  0  0  0  0  0  0  0  0  0
  1  5  2  0  0  0  0
  1  2  1  0  0  0  0
  2  3  2  0  0  0  0
  3  7  1  0  0  0  0
  4  3  1  0  0  0  0
  5  4  1  0  0  0  0
  6  1  1  0  0  0  0
  8  5  1  0  0  0  0
  9  2  1  0  0  0  0
  9 11  1  0  0  0  0
 10  9  1  0  0  0  0
M  END
$$$$
```





Index of each atom

Bond order

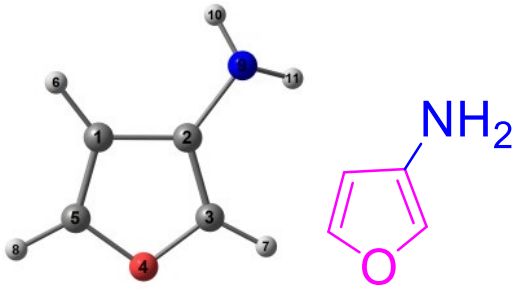
“SMILES”: Simplified Molecular Input Line Entry Specification

✓ SMILES format

・A linear notation
・Coordinates are not stored
・Compact than connectivity table
・各原子のつながりを線形で表記
・分子構造のコンパクトな表現

NC1=COC=C1

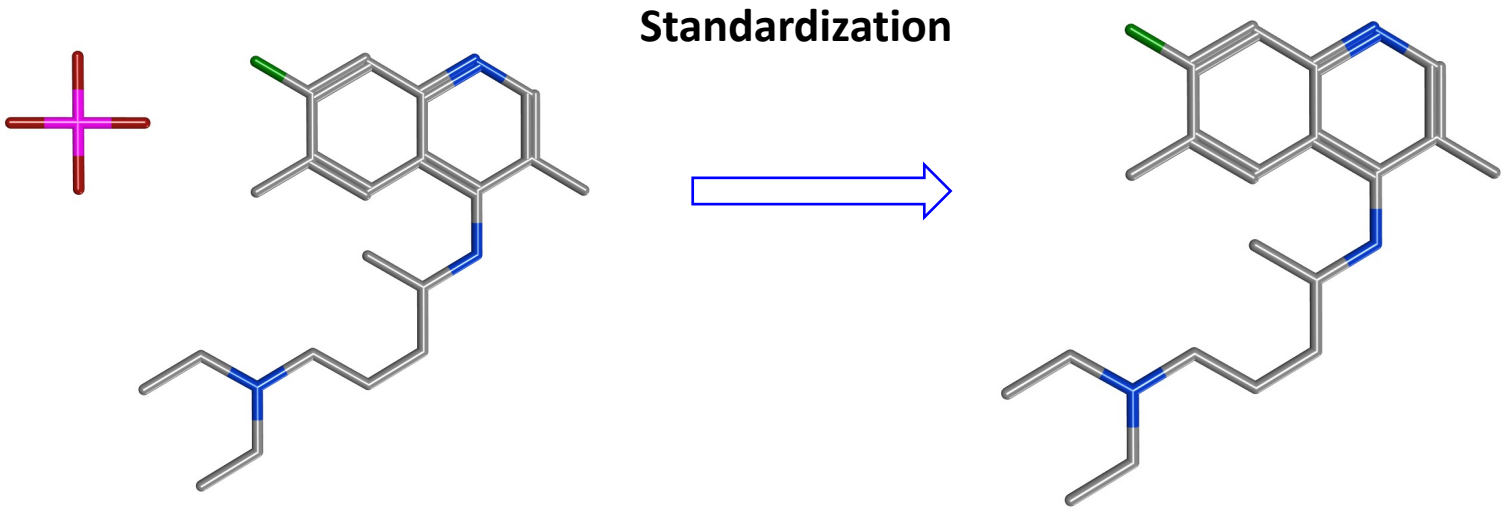
- Atoms
  - element name: C, N, Cl
  - aromatic/aliphatic: c/C
- Bonds
  - single, double, triple: -, =, #
- Branching and rings
  - substituents are put in round brackets ( )
  - rings are indicated by **digits** following ring atoms

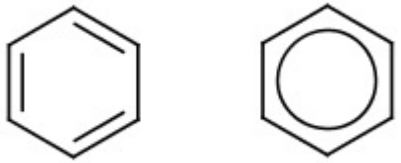


Index of each atom

Bond order

Data Curation (Chemical Structure Standardization)





- Aromatic format or Kekule?
- Deprive salts
- Hydrogen forms (protonation or deprotonation)? neutralization
- Tautomers (keto or enol)
- Erroneous corrections (e.g. carbon atom with 5 bonds)

Data Curation Procedure (Cytopathic effect (CPE) data as an example)

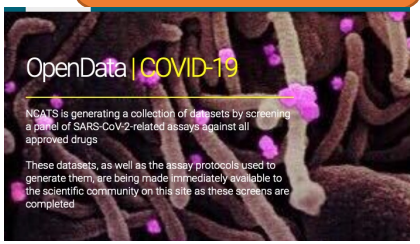
Measuring toxicity against
host cells without SARS-Cov-2 virus

CPE counter assay data set

Toxicity against the cell (pAC50 > 5) = False positive

Download form OpenDataPortal

CPE data set 9,479



Active 747

Remove false positives

Active 601

Drop duplicated molecule according to curated SMILES

Active 567

Inactive 8,752

Drop duplicated molecules according to curated SMILES

Inactive 7,939

Drop actives with Inconsistent active class

Active 515

Prepared Data Sets

Standardized molecule (SMILES)

primary mode of action
= original drug mechanism

	A	B	C	D	E	F	G	H
1	SAMPLE_ID	washed_SMILES	mean_pAC50	Class	PRIMARY_MOA	range_pAC50	count_IDs	SAMPLE_NAME
2	NCGC00343768-0!	Brc1c(C)cc(OC[C@H]2OC	4.9000001	Active	Chk1 Inhibitor	0	1	Rabusertib
3	NCGC00386831-0:	Brc1c(Cl)cc(NC(=O)c2ccc(	4.9000001	Active		0	1	NCGC00386831-01
4	NCGC00263232-0:	Brc1c(N(C)C)ccc(C2Nc3c(	4.9499998	Active	Glutaminase Inhibitor	0	1	GLS-968
5	NCGC00246387-0(	Brc1c(O)c(CN(C)C)c2c(C(=	4.96249985	Active	antiinfluenza drug	0.0250001	2	Arbidol
6	NCGC00263100-0:	Brc1cc(F)c(COc2c(C(=O)N	4.9000001	Active	VEGFR-2 Inhibitor	0	1	OSI-632

Active / Inactive

Including “drug” name

For detecting inconsistency

The number of duplicated chemical structures

Objective of this project

Get “useful” insight or tools for future COVID-19 drug design (Any analysis would be acceptable!)

Other Analysis Strategies

Analyzing a different HTS data set

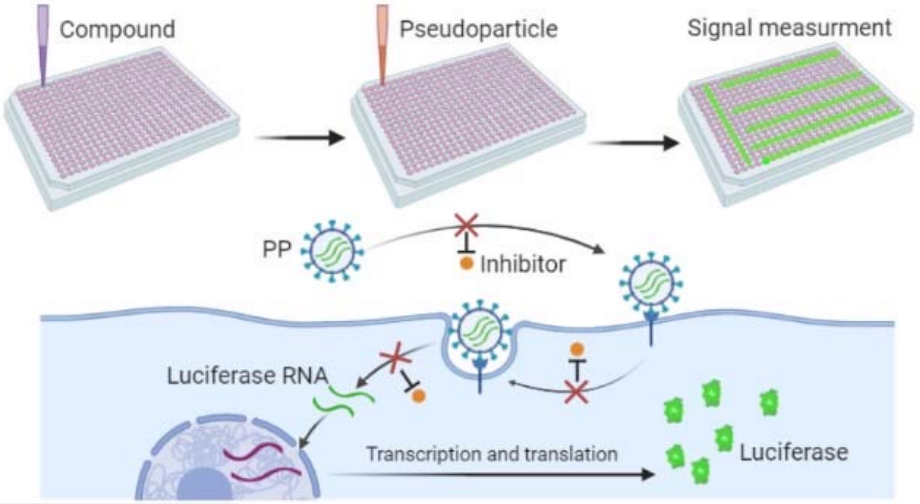
- Inhibitors of spike (SARS-CoV-2) mediated cell entry.
  Mechanism of the protection of virus infection is taken into account.

NCATS Data Portal

SARS-CoV-2 Assays

The assays below have been developed to cover a wide spectrum of the SARS-CoV-2 life cycle, including both viral and human (host) targets. This list will be updated continuously as more assays are developed and screened, and all protocols and screening datasets will be made freely available below.

Assay Name	Assay Type	Target Category	Detection Type	Cell Line	Status	Data
Spike-ACE2 protein-protein interaction (AlphaLISA)	Proximity	Viral Entry	AlphaLISA			
Spike-ACE2 protein-protein interaction (TruHit Counterscreen)	Proximity	Counterscreen	AlphaLISA			
Spike-ACE2 binding	Biophysical	Viral Entry	Bio-Layer Interferometry			
ACE2 binding	Biophysical	Viral Entry	Microscale thermophoresis			
ACE2 enzymatic activity	Biochemical	Viral Entry	Fluorescence			
TMPRSS2 enzymatic activity	Biochemical	Viral Entry	Fluorescence	Not appli		
3CL enzymatic activity	Biochemical	Viral replication	Fluorescence			
SARS-CoV-2 cytopathic effect (CPE)	Cell viability	Live virus infectivity	Luminescence	Vero E6		
SARS-CoV-2 cytopathic effect (host tox counterscreen)	Cell viability	Counterscreen	Luminescence	Vero E6		
SARS-CoV Pseudotyped particle entry	Cell-based	In vitro infectivity	Luminescence	Vero E6	Screening	



CZ. Chen et al., ACS Pharmacol. Transl. Sci. 2020, 3, 6, 1165–1175

Other Analysis Strategies

Mode of action identification (the main protease: M^{pro} in SARS-Cov-2)

The main protease (M^{pro}: aka papain-like protease 3CL) in SARS-Cov-2 is responsible for cleaving poly proteins into 16 non-strucual proteins.

- Improtant for viral replication
- Dissimilar to human proteins
- Promising candidate for SARS-Cov 2 drugs

SARS-CoV-2 Assays

The assays below have been developed to cover a wide spectrum of the SARS-CoV-2 life cycle, including both viral and human (host) targets. This list will be updated continuously as more assays are developed and screened, and all protocols and screening datasets will be made freely available below.

Assay Name	Assay Type	Target Category	Detection Type	Cell Line	Status	Data
Spike-ACE2 protein-protein interaction (AlphaLISA)	Proximity	Viral Entry	AlphaLISA		Screening	
Spike-ACE2 protein-protein interaction (TruHit Counterscreen)	Proximity	Counterscreen	AlphaLISA		Screening	
Spike-ACE2 binding	Biophysical	Viral Entry	Bio-Layer Interferometry		Screening	
ACE2 binding	Biophysical	Viral Entry	Microscale thermophoresis		Screening	
ACE2 enzymatic activity	Biochemical	Viral Entry	Fluorescence		Screening	
TMPRSS2 enzymatic activity	Biochemical	Viral Entry	Fluorescence	Not applicable	Screening	
3CL enzymatic activity	Biochemical	Viral replication	Fluorescence		Screening	
SARS-CoV-2 cytopathic effect (CPE)	Cell viability	Live virus infectivity	Luminescence	Vero E6	Screening	

Prepared Data Sets

	# Active	# Inactive	Total
Cytopathic effect	515	7939	8454
Pseudotyped particle entry	1872	2317	4189
Main protease (3CL)	311	9154	9465

Although you can use the data sets above, using other data sets is acceptable!

Group work

(1) Determine the objective ／ 解析の目的を決めよう

Examples (例)

- コロナウイルスの増殖(侵入)を防ぐ化合物を予測するモデルの構築
- データマイニングによる、活性化合物に共通する特徴量抽出
- コロナウイルスの増殖を防ぐメカニズム（e.g.M^{pro}の阻害）を推定
- Other analyzes will be welcomed !! (Original approaches are highly evaluated)

(2) Conduct the analysis / 解析

- Using machine learning

(3) Discuss the outcome and get insight about compounds or models you get.

解析結果の考察：目的と照らし合わせよう。

- Most important points in this PBL
- ここがこのPBLで一番大事なポイント
- For example
  - 活性に重要となる要素を特定する
  - Importance of each descriptor
  - 各記述子の意味・重要性を調べてみよう

Materials

- Curated Data Sets

- Original Data:
  - NCATS OpenData Portal

  https://opendata.ncats.nih.gov/covid19/index.html
- Related information
  - Machine learning approaches using the same data sets (cell entry): NCATS

    H. Sun et al., Bioorg. Med. Chem., 2021, 38, 116119

    https://doi.org/10.1016/j.bmc.2021.116119
  - Review article on computational approaches for drug developments for SARS-CoV-2

    Alves V.M. et al, Mol. Inf. **2021**, *40,* 2000113
    https://doi.org/10.1002/minf.202000113
  - There are so many publications by googling with a key word “SARS-CoV-2 drug repositioning”

Data Science PBL1 in 2020

Exemplary Analysis from 2020 PBL

Cancer cell line data sets

cancer cell + compounds → cell response (GI: growth inhibition)

From 2020 PBL

Cell line name

NSC	Washed_smiles	CCRF-CEM	HL-60(TB)	K-562	MOLT-4	RPMI-8226	SR
1	CC1=CC(=O)C=CC1=O	5.6	5.5	5.4	5.5	5.4	5.4
17	CCCCCCCCCCCCCCCc1cc(ccc1N)O	7.3	6.8	5.0	6.7	6.7	5.7
26	c1ccc(cc1)C(CCl)(c2ccccc2)c3ccccc3	5.4		5.8	5.6	5.4	5.7
89	CN(C)CCC(=O)c1ccccc1	4.6	4.9	4.6	4.7	5.0	4.5
112	Cc1cc(c(c(c1C)C)C[N+](C)(C)C)C	6.7	6.3	6.4	6.5	6.2	6.1
171	c1ccnc(c1)C(=O)O	4.0		4.0	4.0	4.0	4.0

pGI50

No data

Objective of this project

Get "useful" insight or tools for future anti-cancer drug design or tumor killing compounds.
(Any analysis would be acceptable!)

Exemplary Case of Chemoinformatics Analysis

NSC	Washed_smiles	CCRF-CEM	HL-60(TB)	K-562	MOLT-4	RPMI-8226	SR
1	CC1=CC(=O)C=CC1=O	5.6	5.5	5.4	5.5	5.4	5.4
17	CCCCCCCCCCCCCCCCc1cc(ccc1N)O	7.3	6.8	5.0	6.7	6.7	5.7
26	c1ccc(cc1)C(CCl)(c2ccccc2)c3ccccc3	5.4		5.8	5.6	5.4	5.7
89	CN(C)CCC(=O)c1ccccc1	4.6	4.9	4.6	4.7	5.0	4.5
112	Cc1cc(c(c(c1C)C)C[N+](C)(C)C)C	6.7	6.3	6.4	6.5	6.2	6.1
171	c1ccnc(c1)C(=O)O	4.0		4.0	4.0	4.0	4.0

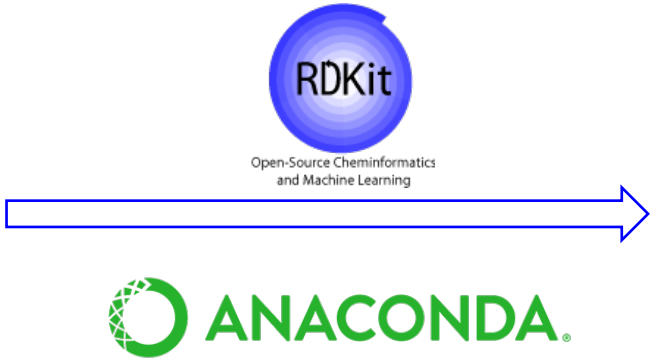
I want to make a model distinguishing active and inactive compounds for the CCRF-CEM cell line (Leukemia).

pGI50 > 6: active
pGI50 < 6: inactive

NSC	Washed_smiles	CCRF-CEM	Active
1	CC1=CC(=O)C=CC1=O	5.6	0
17	CCCCCCCCCCCCCCCCc1cc(ccc1N)O	7.3	1
26	c1ccc(cc1)C(CCl)(c2ccccc2)c3ccccc3	5.4	0
89	CN(C)CCC(=O)c1ccccc1	4.6	0
112	Cc1cc(c(c(c1C)C)C[N+](C)(C)C)C	6.7	1
171	c1ccnc(c1)C(=O)O	4.0	0

Convert the “SMILES” into descriptors

SMILES



Descriptors
記述子

Descriptors
記述子

Active or Inactive
活性の有無

Examples of program (python)

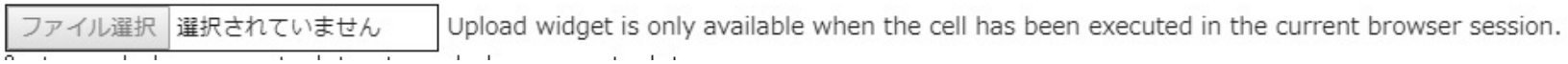
Import packages and modules.

```
import pandas as pd
import numpy as np
from numpy import vectorize as vec
import scipy as sp
import sklearn
from sklearn.model_selection import train_test_split
import matplotlib
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns

from rdkit import Chem
from rdkit.Chem.Draw import IPythonConsole
from rdkit.Chem import Descriptors,PandasTools
from rdkit.ML.Descriptors import MoleculeDescriptors
```

An example of the program (python)

Upload the NCI60 data sets to google colab.

```
from google.colab import files
files.upload() # upload the SMILES file (washed-nr-er-neutral.tsv)
```



Load the SMILES file

```
mols = pd.read_csv('pGI50_mols.tsv', sep='¥t', index_col=0)
mols.head(3)
```

NSC	Washed_smiles	CCRF-CEM	HL-60(TB)	K-562	MOLT-4	RPMI-8226	SR
1	CC1=CC(=O)C=CC1=O	5.5705	5.5405	5.441	5.4875	5.3855	5.4095
17	CCCCCCCCCCCCCCCc1cc(ccc1N)O	7.3320	6.8470	4.970	6.7370	6.7450	5.6610
26	c1ccc(cc1)C(CCl)(c2ccccc2)c3ccccc3	5.4490	5.7660	5.777	5.5900	5.4290	5.7150

An example of the program (python)

Extract CCRF-CEM cell line and convert pGI50 to active(1) or not (0)

```
ccrf = mols[['Washed_smiles', 'CCRF-CEM']]
ccrf['Active'] = ccrf['CCRF-CEM'] > 6
ccrf.head(3)
```

NSC	Washed_smiles	CCRF-CEM	Active
1	CC1=CC(=O)C=CC1=O	5.5705	False
17	CCCCCCCCCCCCCCCc1cc(ccc1N)O	7.3320	True
26	c1ccc(cc1)C(CCl)(c2ccccc2)c3ccccc3	5.4490	False

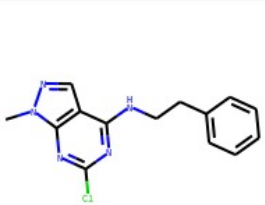
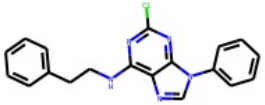
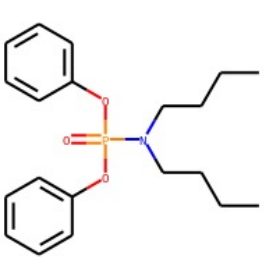
Convert SMILES to Molecule (ROMol)

```
Chem.PandasTools.AddMoleculeColumnToFrame(ccrf, smilesCol='Washed_smiles',
molCol='ROMol')
```

An example of the program (python/RDKit)

Visualize molecules (random selection)

```
ccrf.iloc[1000:1003,:] # random visualization
```

	Washed_smiles	CCRF-CEM	Active	ROMol
19139	Cn1c2c(cn1)c(nc(n2)Cl)NCCc3ccccc3	NaN	False	
19147	c1ccc(cc1)CCNc2c3c(nc(n2)Cl)n(cn3)c4ccccc4	NaN	False	
19192	CCCCN(CCCC)P(=O)(Oc1ccccc1)Oc2ccccc2	5.078	False	

An example of the program (python/RDKit)

List up the descriptors in RDKit

```
names = [x[0] for x in Descriptors._descList]
print("Number of descriptors in the rdkit: ", len(names))
np.array(names)
```

```
Number of descriptors in the rdkit: 200

array(['MolWt', 'HeavyAtomMolWt', 'ExactMolWt',
'NumValenceElectrons', 'NumRadicalElectrons',
'MaxPartialCharge', 'MinPartialCharge',
'MaxAbsPartialCharge', 'MinAbsPartialCharge',
'MaxEStateIndex', 'MinEStateIndex', 'MaxAbsEStateIndex',
'MinAbsEStateIndex', 'BalabanJ', 'BertzCT', 'Chi0', 'Chi0n',
'Chi0v', 'Chi1', 'Chi1n', 'Chi1v', 'Chi2n', 'Chi2v', 'Chi3n',
**snip**
```

An example of the program (python/RDKit)

Select descriptors (in blue) and calculate them. You can choose all.

```
# Arbitrary selection
desc_for_now =
['TPSA','SlogP_VSA1','EState_VSA1','SMR_VSA1','MolLogP','MolMR','BalabanJ','HallKie
rAlpha','Kappa1','Kappa2','Kappa3','RingCount','NumHAcceptors','NumHDonors']

calculator = MoleculeDescriptors.MolecularDescriptorCalculator(desc_for_now)
from collections import OrderedDict
desc = OrderedDict()

for mol in ccrf.index:
    desc[mol] = calculator.CalcDescriptors(ccrf.loc[mol, 'ROMol'])

desc_mols = pd.DataFrame.from_dict(desc, orient='index', columns=desc_for_now)
```

Just in case. Save the descriptors calculated.

```
desc_mols.to_csv('descriptors.tsv', sep='¥t')
```

Check the descriptors

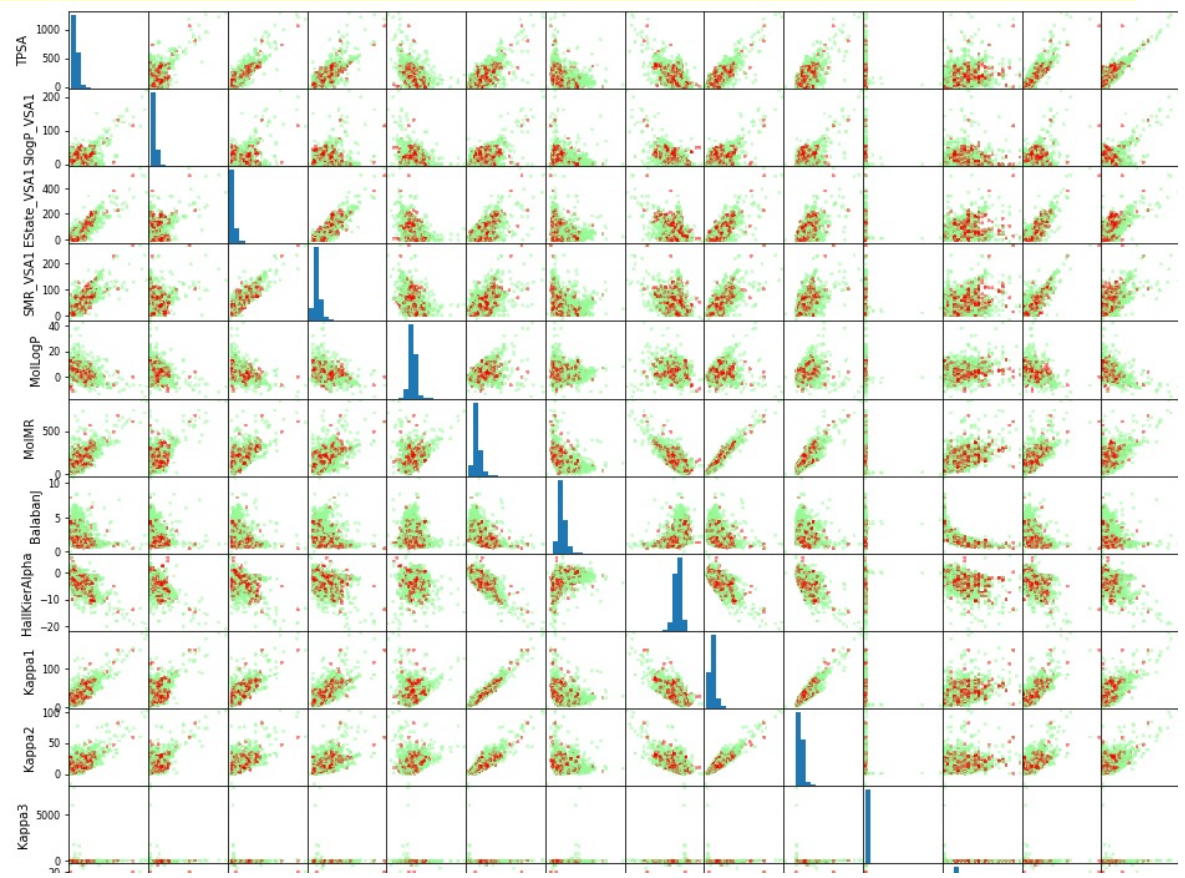
Visualize correlations between each pair of descriptors.

```
def set_color(L):
    O = []
    for l in L:
        if l == '1':
            O.append("red")
        else:
            O.append("palegreen")
    return O

pd.plotting.scatter_matrix(mols_desc,figsize=(16,16), hist_kwds={'bins':15},
                           marker='+', s=8, alpha=.5, c=set_color(ccrf.Active))
plt.show()
```

Check the descriptors

Visualize correlations between each pair of descriptors.

```
def set_color(L):
    O = []
    for l in L:
        if l == '1':
            O.append("red")
        else:
            O.append("palegreen"
    return O

pd.plotting.scatter_matrix(m
                           marker='+', s=
plt.show()
```



Separate into training and test data

```
X_train, X_test, y_train, y_test = train_test_split(desc_mols, ccrf.Active,
                                                    train_size=0.25, test_size=0.75, random_state=0)
```

Check the number of data

```
print("Training Data")
print("Number of active molecules: ", list(y_train).count(1))
print("Number of inactive molecules: ", list(y_train).count(0))
print("Test Data")
print("Number of active molecules: ", list(y_test).count(1))
print("Number of inactive molecules: ", list(y_test).count(0))
```

```
Training Data
Number of   active molecules:   4101
Number of inactive molecules:   34282
Test Data
Number of   active molecules:   1386
Number of inactive molecules:   11409
```

Build the model

例えば…　RandomForestを使ってみる

```
from sklearn.ensemble import RandomForestClassifier
model = RandomForestClassifier(random_state=0)
model.fit(X_train, y_train)
print("Accuracy on training set: {:.3f}".format(model.score(X_train, y_train)))
print("Accuracy on test set:       {:.3f}".format(model.score(X_test,  y_test)))
```

```
Accuracy on training set: 0.997
Accuracy on test set:     0.909
```

他にも色々試してみよう
例　K近傍法, Neural Network,　etc…

特徴量の重要度を測れないか検討してみよう
ヒント　feature_importances_

記述子の意味を調べる際は、Rdkitのマニュアルを参照すること
http://www.rdkit.org/RDKit_Docs.2012_12_1.pdf