PBL-I

Data Science for Chemistry

2022/7/28 (Thu.)

Schedule

- 7/28 **Introduction** < now
  - About the task of this PBL

- 7/28-9/27 **Group work**
  - Build the model, evaluate the results
  - Discuss the meaning of the analysis
  - Prepare the presentation materials
  - (Have a discussion at least once every two weeks)

- 9/28 (Tue.) 9:20-12:30 **Group Presentations** (L3)
  - **Presentation**: 10 min (talk) + 3 min (discussion)
  - **Slides**: in English
  - **Talk**: either in English or Japanese

Final Report

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

     Describe “when” (date or period) and "what" you've done explicitly. Any types of contributions would be OK

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

     A4 2 pages (excluding figures and references)

     A template doc file is available on the PBL web page

Final Report

- Deadline : 10/12 (Wed.) 23:59 (JST)

Upload a PDF file to the NAIST Databox

データサイエンスPBL I / Data Science PBL I [ID: 5013]

URL: https://nrss.naist.jp/group/158

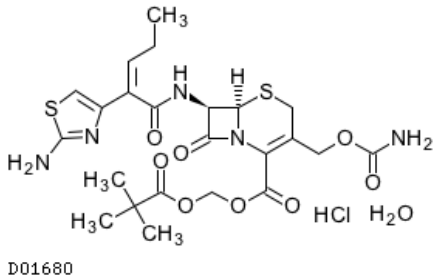
`DSPBL1_{Student ID}_{LastName}_{FirstName}.pdf`

(For example, `DSPBL1_01382812_ONO_Naoaki.pdf`)

Task

Quantitative structure-activity relationship (QSAR)

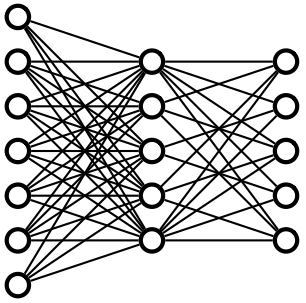
Material



Chemical structures

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

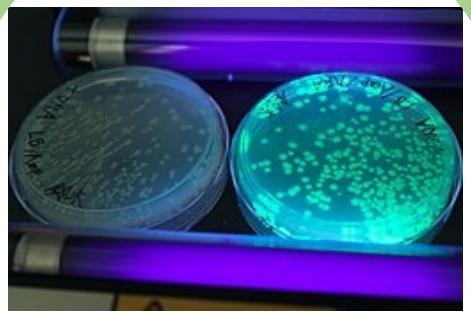
Predict biological activity of chemical compounds using computational models



Machine learning

Data science

Biology



Biological activity

Escherichia coli
a.k.a.
大腸菌

Dataset

Basic: Cancer Cell Line

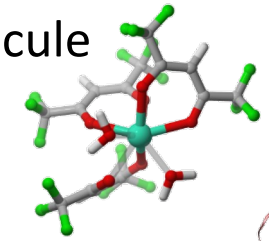
Advanced: COVID-19 protein inhibitor
(See introduction in 2021 for detail)

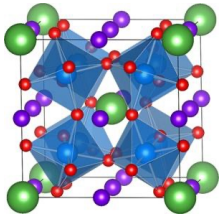
Or you may look for other public databases...

Data Science in Material Design

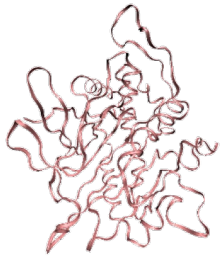
Compounds 化合物

Molecule

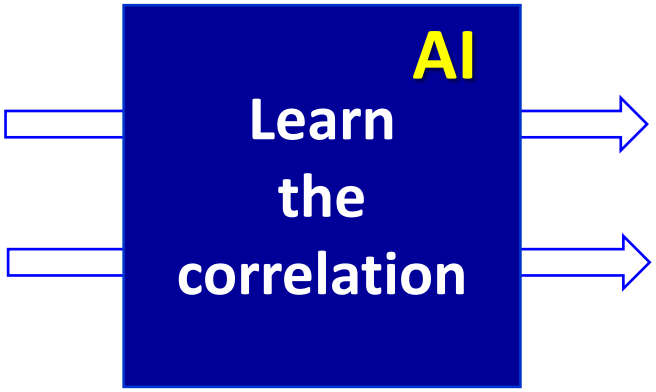




Solid



Polymer



Functions 機能

Catalytic ability

Conductive property

Photo activity

Thermal durability

etc

INPUT

Unknown material

OUTPUT

Function (ON/OFF or intensity)

Point!

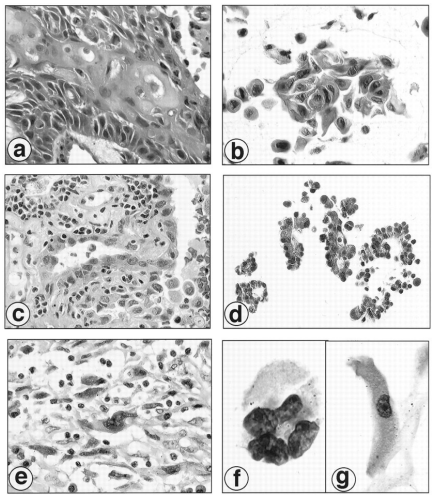
The compounds need to be represented as numbers !

化合物を数値・ベクトル等で表現する必要あり

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

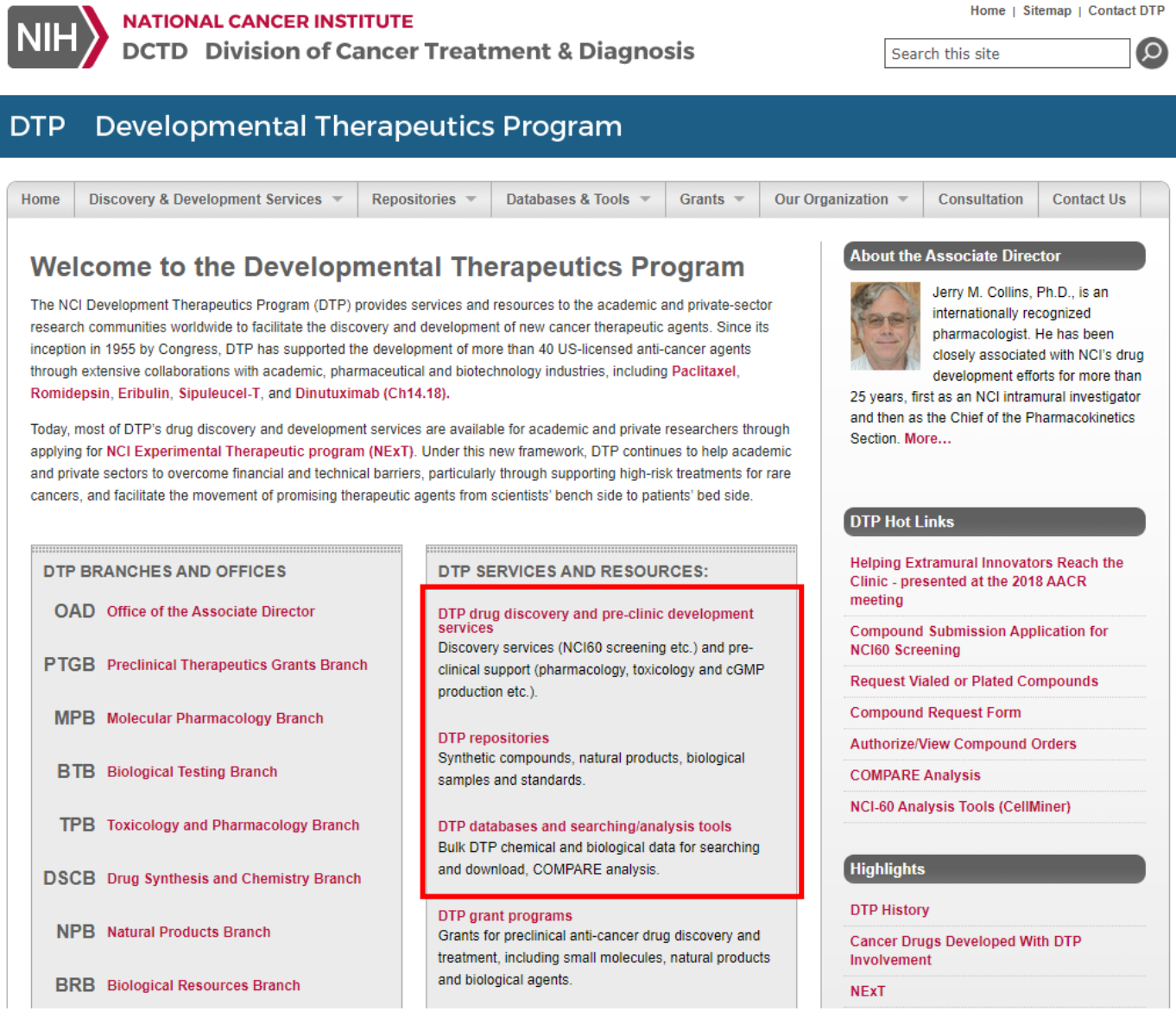
‘Drug design’ is a painstaking search for candidates of new drug from billions of possible chemical compounds.

[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



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.

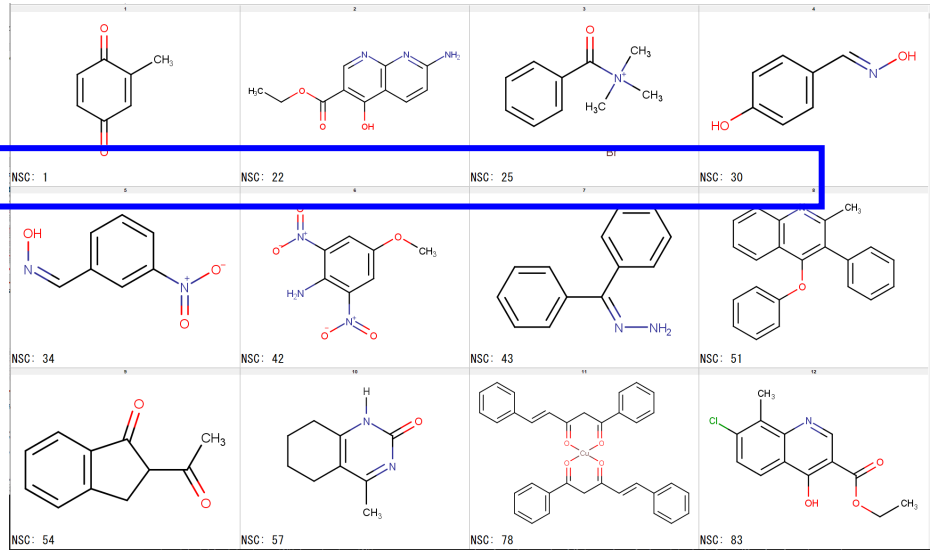
2D structures

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

About the data

“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

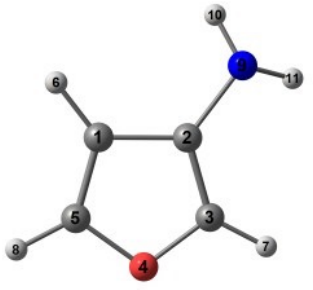
✓ SDF file

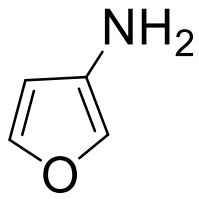
- Series of MOL format
- Properties can be also included

・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

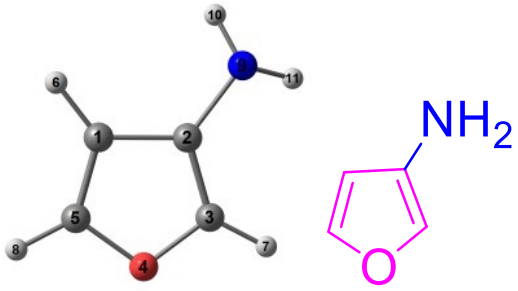
“SMILES”: Simplified Molecular Input Line Entry Specification

✓ 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

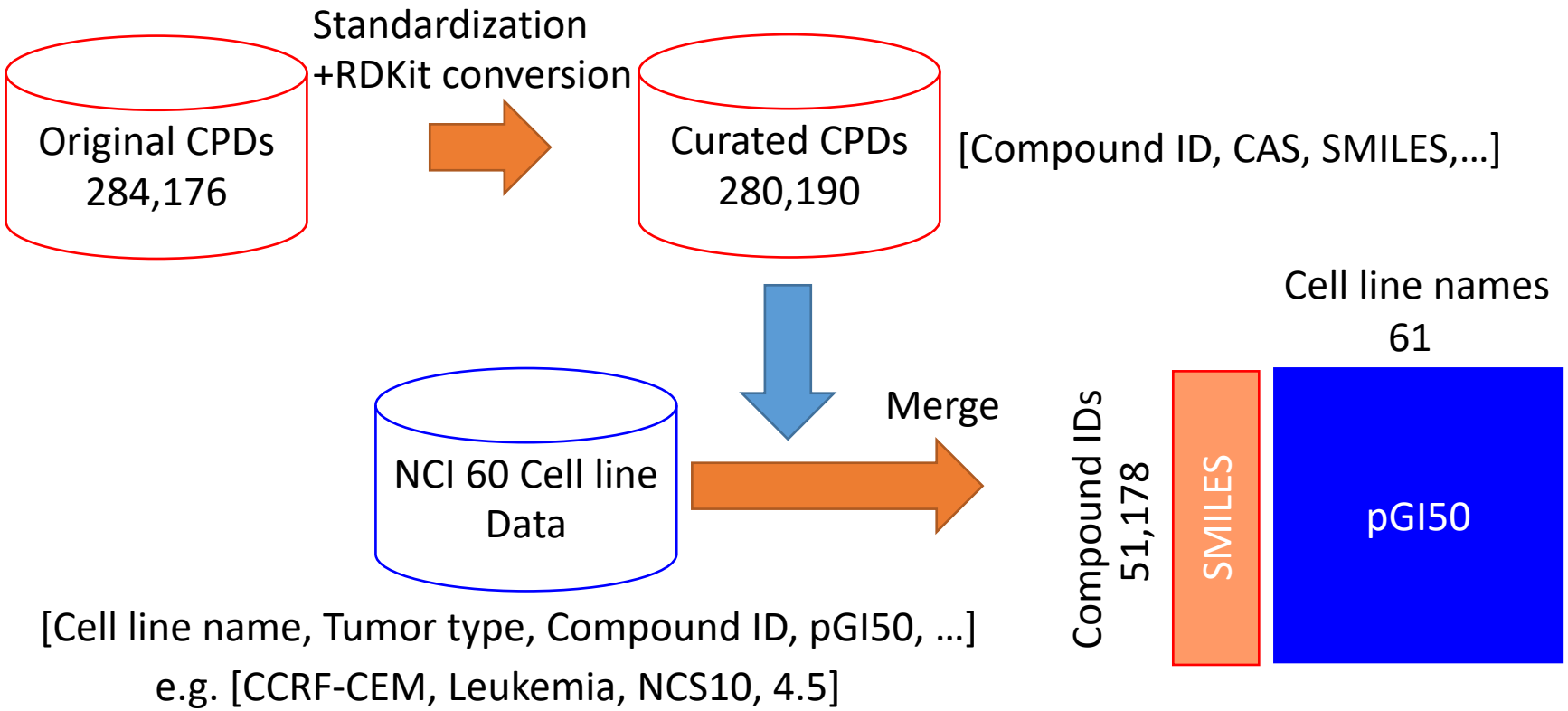
- Atoms
  - element name: C, N, Cl
  - aromatic/aliphatic: c/C
- Bonds
  - single, double, triple: -, =, #
- Branching and rings
  - substituents are put in round brackets ( )
  - rings are indicated by **digits** following ring atoms



Index of each atom

Bond order

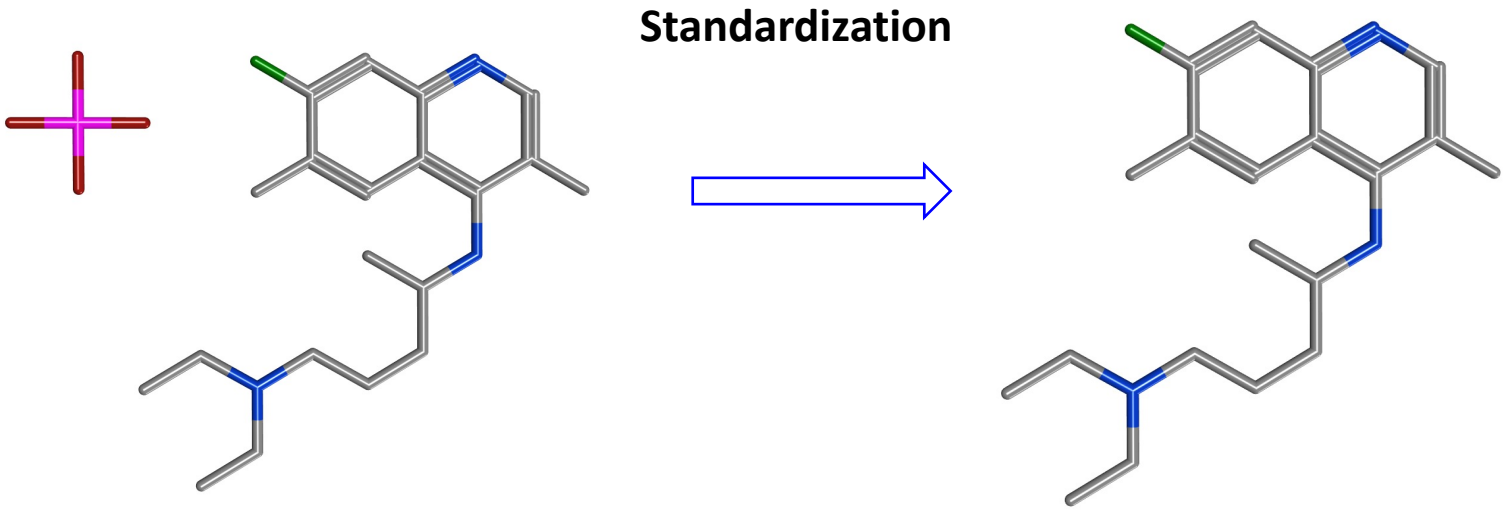
Data Curation Procedure

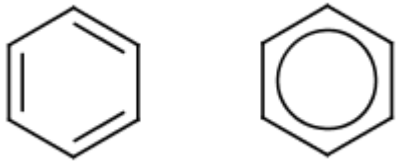


pGI50 = negative logarithm of growth inhibition of 50%
Bigger is better: * basically pGI50 > 6 can be considered as active.

https://pubchem.ncbi.nlm.nih.gov/bioassay/79#section=Description

Data Curation (Chemical Structure Standardization)





- Aromatic format or Kekule?
- Deprive salts
- Hydrogen forms (protonation or deprotonation)? at PH7.4
- Tautomers (keto or enol)
- Erroneous corrections (e.g. carbon atom with 5 bonds)

Prepared Data Sets

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	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

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!)

About the task

Group work

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

- 特定のセルラインに対しての高精度活性予測モデルの構築
- 特定の腫瘍細胞に対しての活性（毒性）予測（マルチタスク）
- 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
- Other analyzes will be welcomed

(2) Conduct the analysis / 解析

- Train your machine learning model and evaluate its outputs/ 予測モデルを構築し、学習の精度を評価する

(3) Discuss the outcome and get insight about compounds /
解析結果を考察し目的と照らし合わせよう

- Evaluate the most important molecular feature / 活性に重要となる要素を特定できるか評価してみよう
- Discuss biological or chemical explanation of the important descriptors / 重要な記述子の生物学的/化学的意味を調べてみよう

Exemplary Case (1)

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

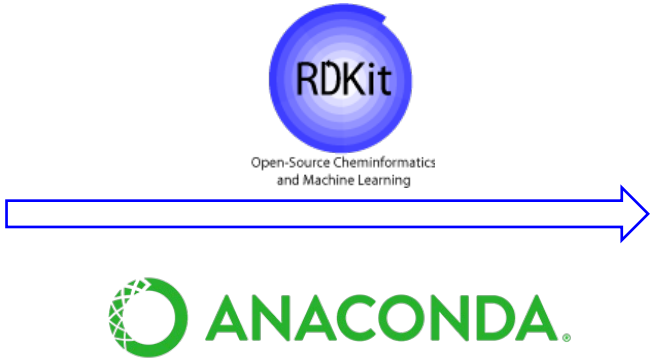
I want to make a model distinguishing active and inactive compounds for the CCRF-CEM cell line (Leukemia).

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

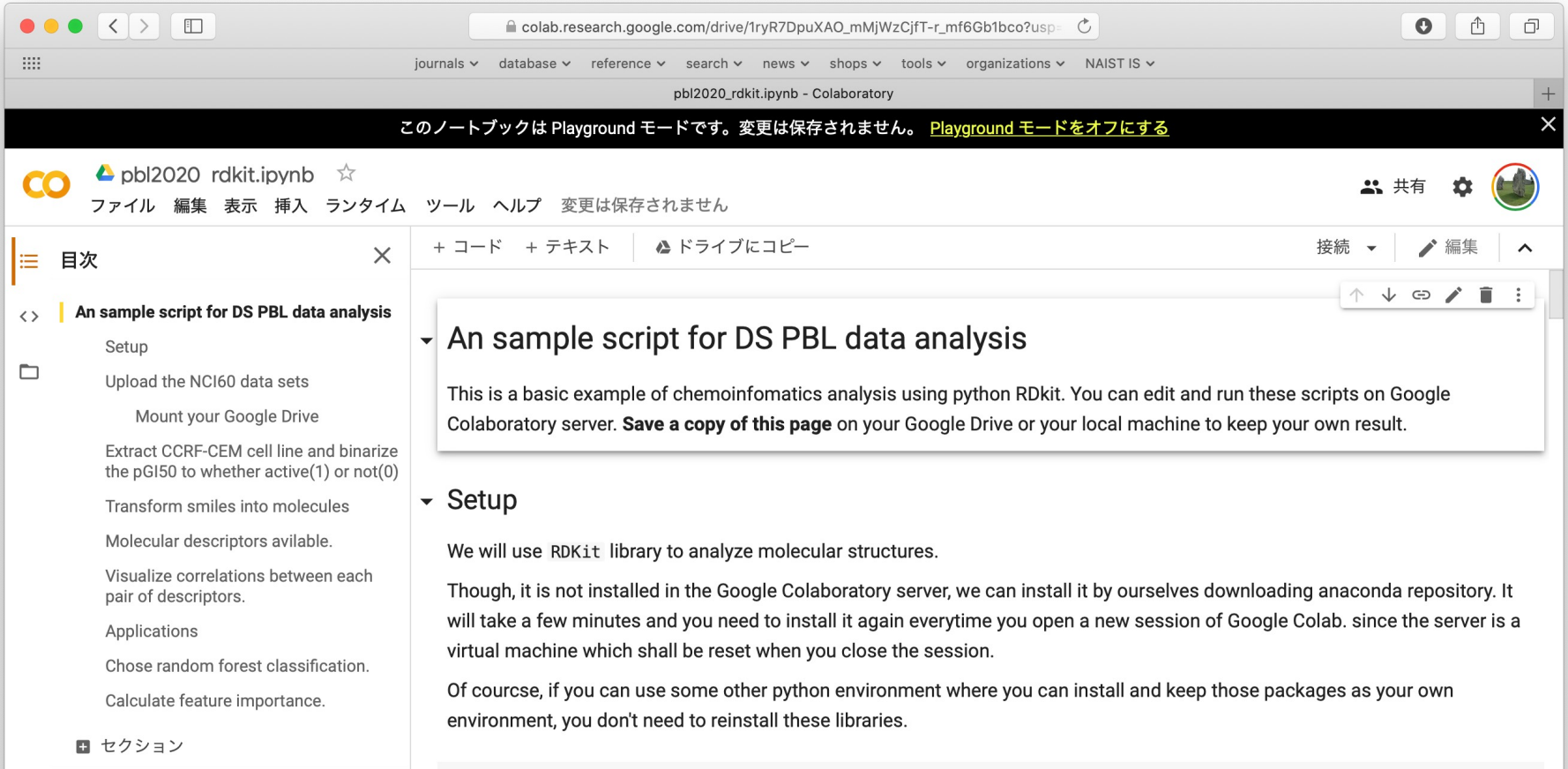
Example programs

You can access the notebook of the sample program from

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

Example programs

You can find a `notebook` of sample scripts on the PBL web page. The notebook will be opened on Google Colaboratory server, so that you can edit and try your analysis without any preparation (of course, you may use any environment you prefer).



Setup

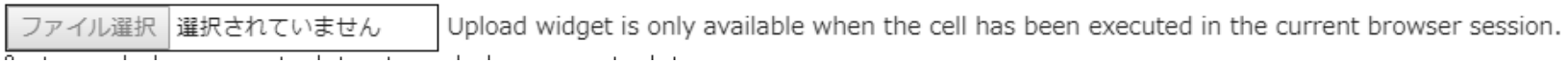
Import packages and modules.

```
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
```

Read Mol data

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

`Mol` object

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

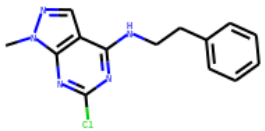
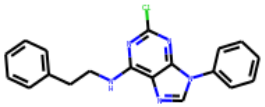
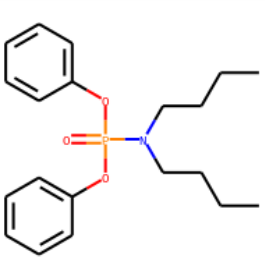
Convert SMILES to Molecule (ROMol)

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

`Mol` object

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	

Descriptors in the 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**
```

Descriptors in the 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','HallKierAlpha','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

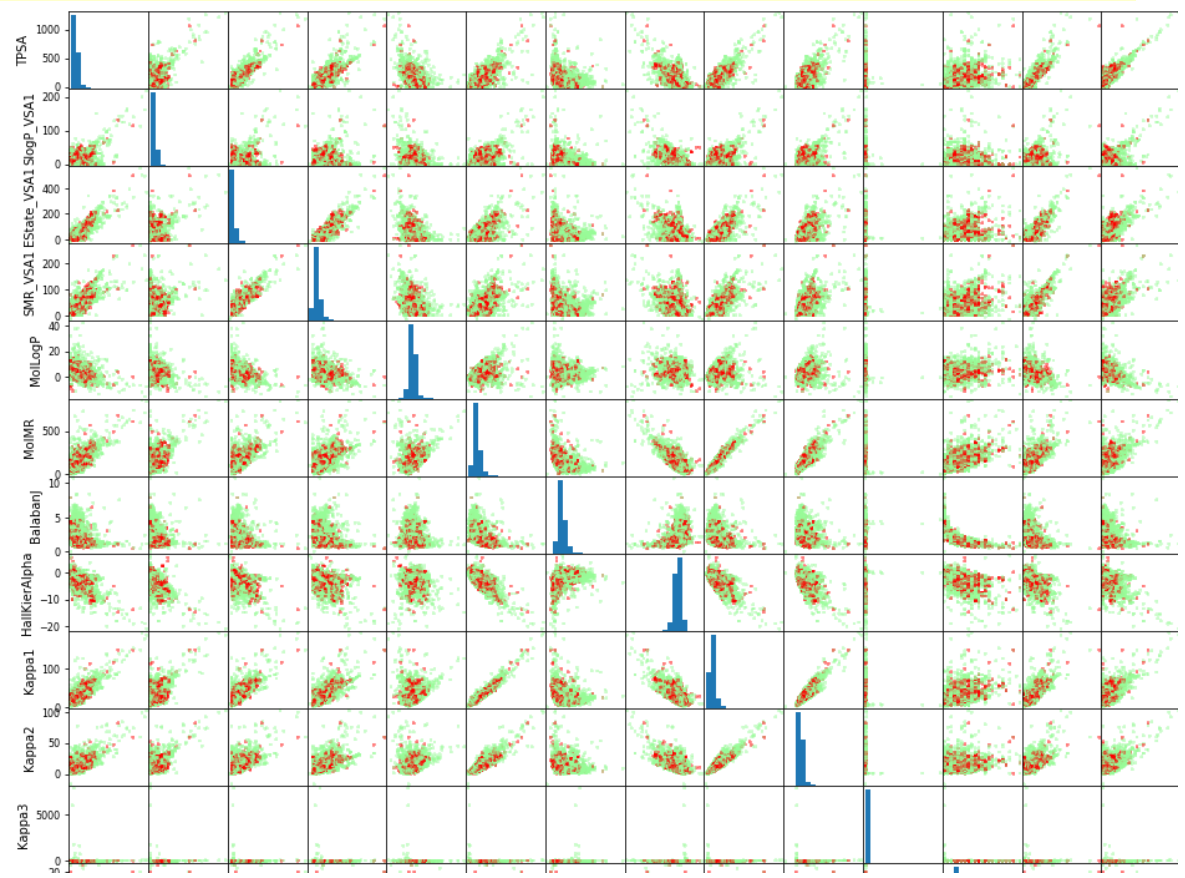
Visualize correlations between each pair of descriptors.

```
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 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
```

Apply some models

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

Materials

- Curated Data Sets < included in the example
  - pGI50 data sets:
  - NCI60 Cell line and Panel (type of cancer cell)
- Original Data:
  - Growth Inhibition Data:

    https://wiki.nci.nih.gov/display/NCIDTPdata/NCI-60+Growth+Inhibition+Data
  - Chemical Data:

    https://wiki.nci.nih.gov/display/NCIDTPdata/Chemical+Data
- **Data for Further Analysis**
  - In vivo antitumor assays:

    https://wiki.nci.nih.gov/display/NCIDTPdata/In+Vivo+Antitumor+Assays
  - Molecular target data:

    https://wiki.nci.nih.gov/display/NCIDTPdata/Molecular+Target+Data
  - Cell Miner CDB: Cell Line data (gene expression data)

    https://discover.nci.nih.gov/cellminercdb/
  - Related article:

    Nakano T. and J.B. Brown, Journal of Computer Aided Chemistry **2020**, 21, 1-10 https://doi.org/10.2751/jcac.21.1

Materials

- **http://www-dsc.naist.jp/dsc_naist/dsc-pbl/**
- **References**
  - Notebook with the sample python script (Google Colaboratory).
  - Curated dataset (21.4MB tsv file)
- **References (2021)**
  - Introduction (2021/07/27)
  - Curated dataset (1.3MB zipped CSV)
  - Presentation slides of 2021 groups
- **References (2020)**
  - Introduction (2020)