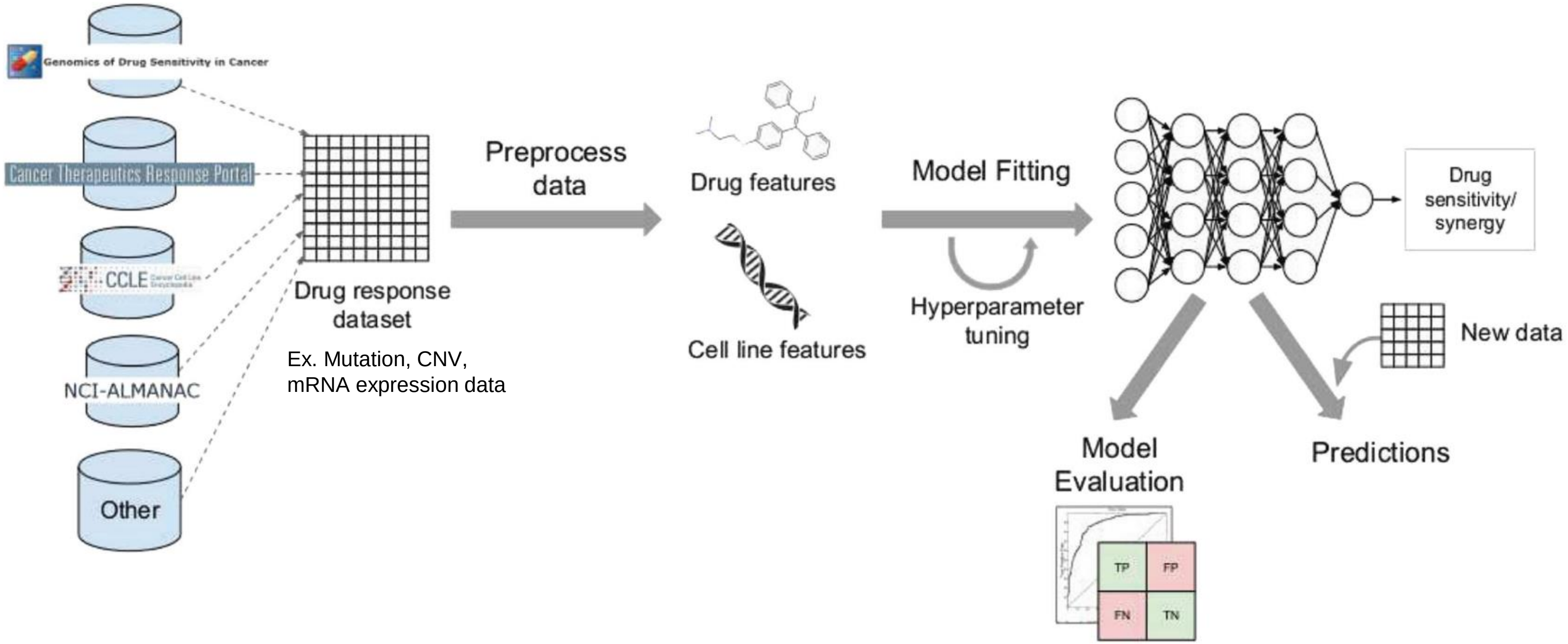
BIO-MEDICAL
INFORMATICS LAB

Multi-omics drug response prediction

박서연

Deep learning workflows for drug response prediction



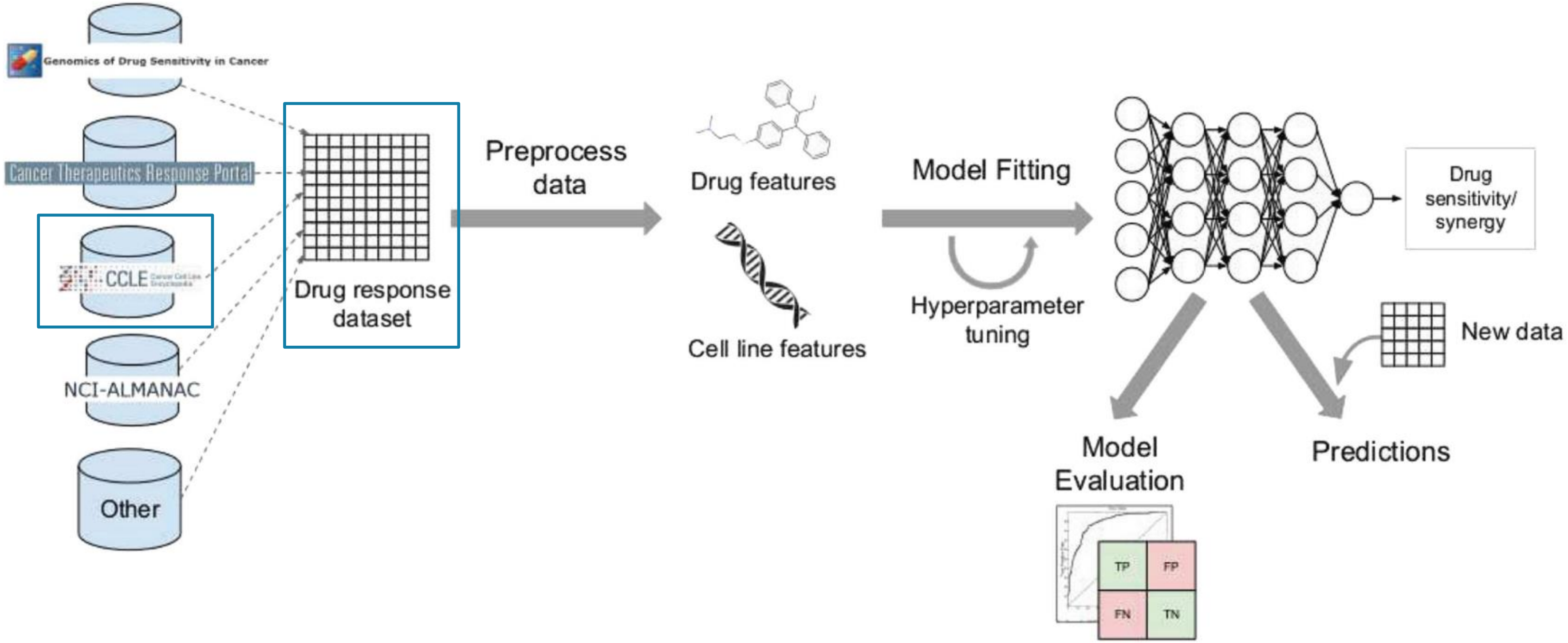
Ex. Mutation, CNV,
mRNA expression data

Delora Baptista, et.al. Deep learning for drug response prediction in cancer. Briefings in Bioinformatics, 22(1). (2021)

Deep learning workflows for drug response prediction

① Select Dataset
: CCLE

② Select Data types
: Gene expression
(Transcriptomic data)

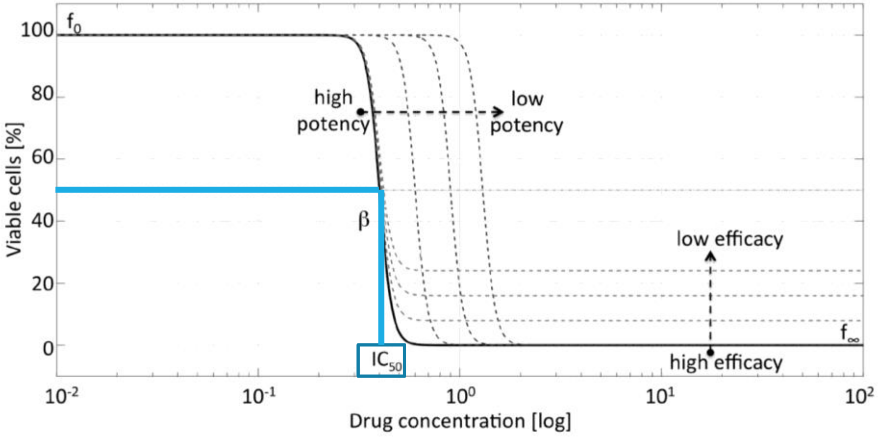


Delora Baptista, et.al. Deep learning for drug response prediction in cancer. Briefings in Bioinformatics, 22(1). (2021)

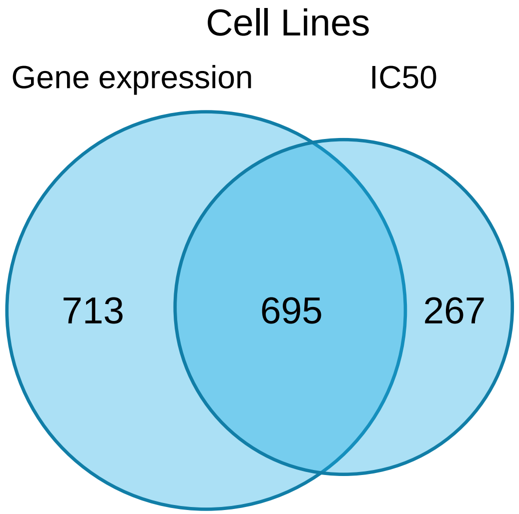
Dataset

- **CCLE database**
  - Gene expression data
    - CCLE Database Gene expression data 22Q4

      : OmicsExpressionProteinCodingGenesTPMLogp1.csv
    - Genes: 16,383
    - Cell Lines: 1,408
    - Primary Diseases: 74
    - Lineages: 29
  - IC50 data
    - Drug sensitivity IC50 (Sanger GDSC1)
      - Cell Lines: 962
      - Drugs: 310



Dataset



	Cell Line	Gene expression	Drug Name	IC50
1	ACH-000242	[6.729417 0. ... 0.17632277]	cabozantinib	1.353158
2	ACH-000242	[6.729417 0. ... 0.17632277]	Torin 2	3.097715
3	ACH-000242	[6.729417 0. ... 0.17632277]	ZG-10	2.934035
	…	…	…	…
41,460	ACH-000052	[4.2494454 0. ... 0.13750352]	vinblastine	-6.64979
Total	**695 cell lines**	**16,384 genes**	**310 drugs**	**41,460 IC50 values (cell line-drug pairs)**

Data Preprocessing
- Gene expression

Dataset

- **CCLE Database Gene expression data 22Q4**

OmicsExpressionProteinCodingGenesTPMLogp1.csv

- Gene expression TPM values of the protein coding genes for DepMap cell lines
- Values are inferred from RNA-seq data using the RSEM tool and are reported after log2 transformation, using a pseudo-count of 1

$$log_2(TPM + 1)$$

Genes

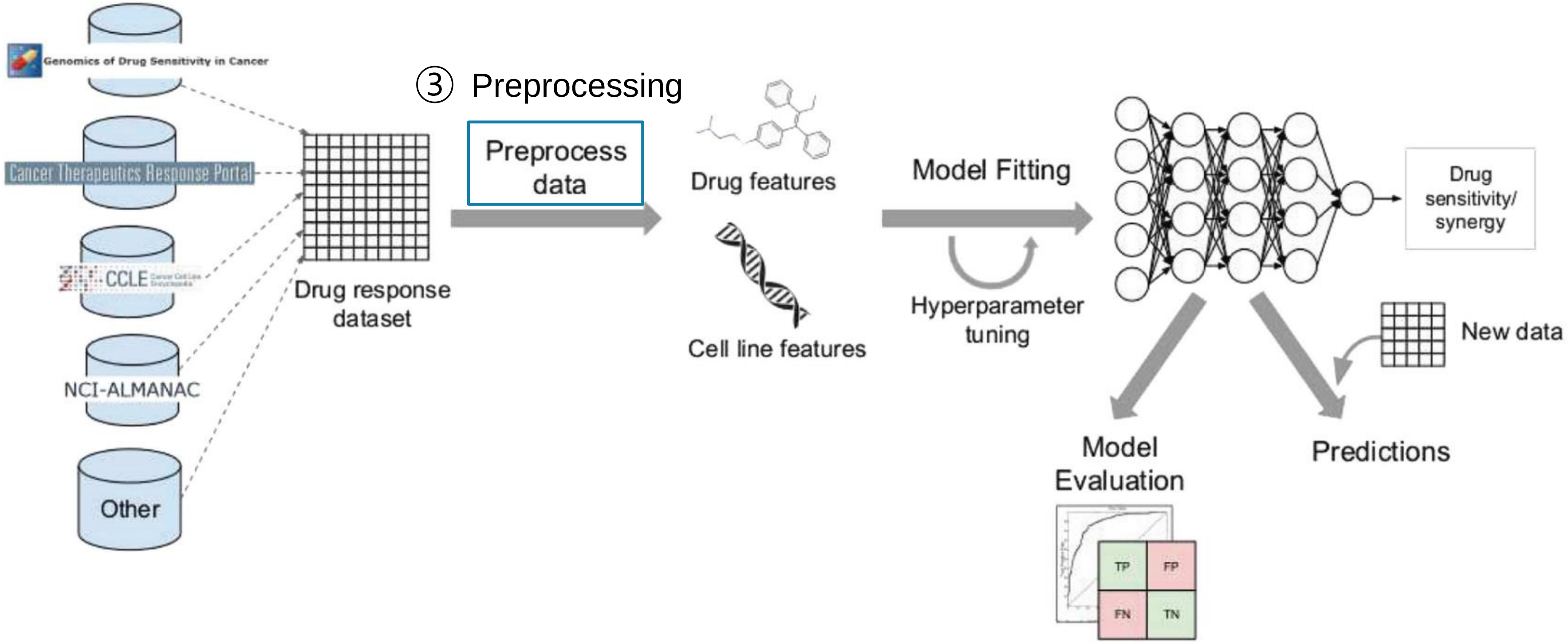
Cell lines

	TSPAN6	TNMD	DPM1	…	MAGEA6
ACH-001113	4.331992	0	7.36466	…	0.028569
…	…	…	…	…	…
ACH-000052	4.249445	0	6.175724	…	0.137504

Yu-Chiao Chiu, et.al. Predicting drug response of tumors from integrated genomic profiles by deep neural networks. BMC Medical Genomics 12, 18. (2019)
Qiao Liu, et.al. DeepCDR: a hybrid graph convolutional network for predicting cancer drug response. Bioinformatics 36. (2020)
Hossein Sharifi-Noghabi, et.al. MOLI:multi-omics late integration with deep neural networks for drug response prediction. Bioinformatics 15, 14. (2019)

Deep learning workflows for drug response prediction

③ Preprocessing



Delora Baptista, et.al. Deep learning for drug response prediction in cancer. Briefings in Bioinformatics, 22(1). (2021)

Data Preprocessing
- Gene expression

Data preprocessing for CCLE 22Q4 dataset

- **Gene Filtering**

  : exclude 20% of genes with the lowest variance assuming them not informative

  ※ Calculate variance only with Train data

  ⇒ Do not calculate variance for the whole dataset

Genes

Cell lines

	TSPAN6	TNMD	DPM1	…	MAGEA6
ACH-001113	4.331992	0	7.36466	…	0.028569
…	…	…	…	…	…
ACH-000052	4.249445	0	6.175724	…	0.137504

Genes

Cell lines

	TSPAN6	DPM1	…	MAGEA6
ACH-001113	4.331992	7.36466	…	0.028569
…	…	…	…	…
ACH-000052	4.249445	6.175724	…	0.137504

Total: 16,384 → 13,106 genes

1,408 cell lines × 13,106 genes

Data Preprocessing
- Gene expression

Data preprocessing for CCLE 22Q4 dataset

- **Gene expression Normalization**

  : Standard Normalization

  → normalize gene expression to $N(0, 1)$

Genes

Cell lines

	TSPAN6	DPM1	…	MAGEA6
ACH-001113	4.331992	7.36466	…	0.028569
…	…	…	…	…
ACH-000052	4.249445	6.175724	…	0.137504

→

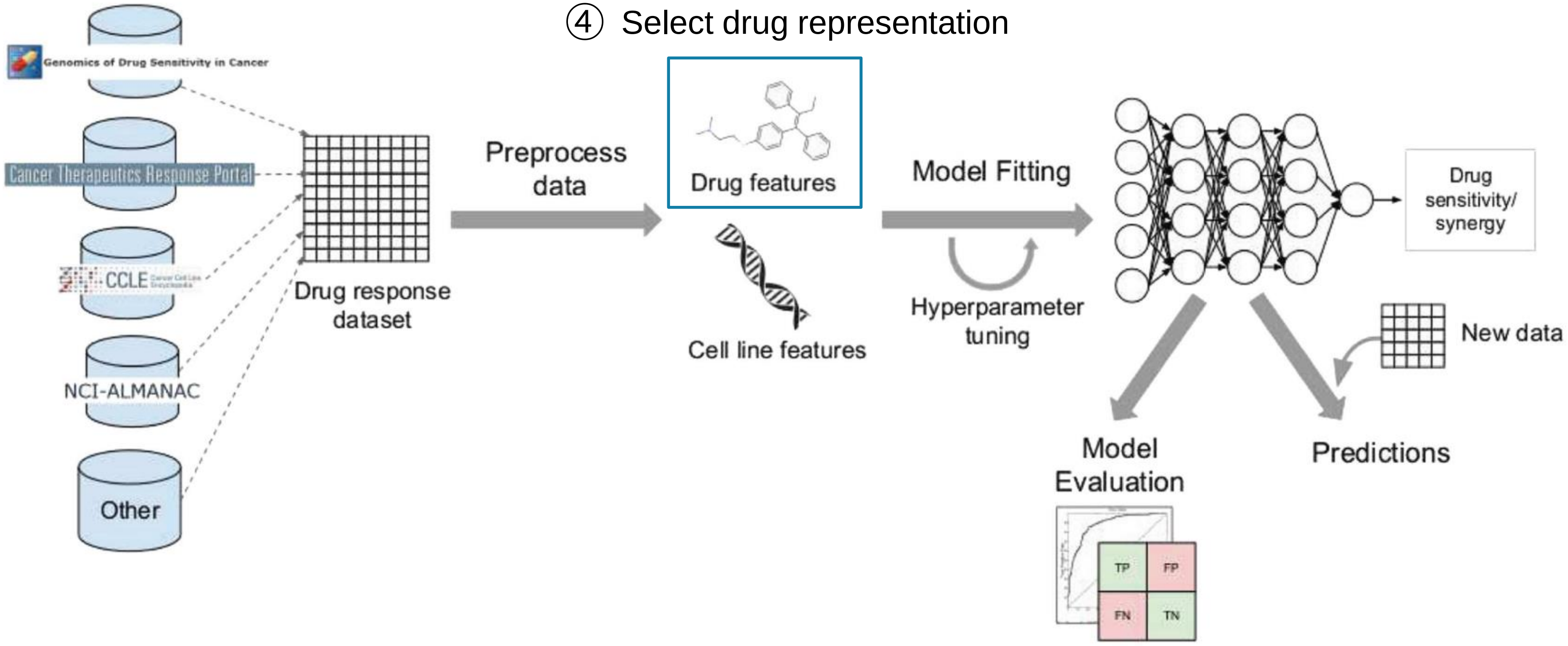
Genes

Cell lines

	TSPAN6	DPM1	…	MAGEA6
ACH-001113	0.584224	1.337422	…	-0.736650
…	…	…	…	…
ACH-000052	0.534013	-0.481543	…	-0.696717

Deep learning workflows for drug response prediction

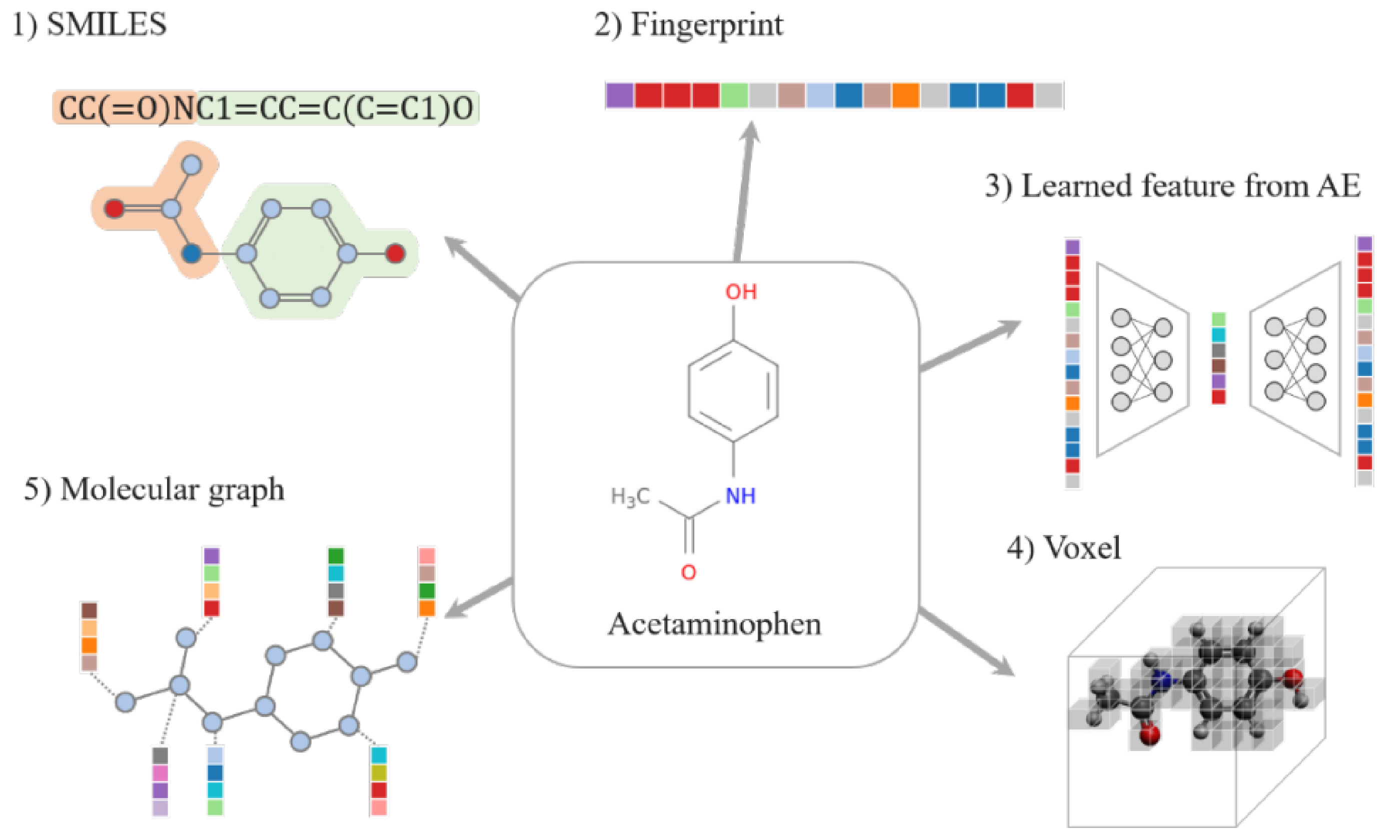
④ Select drug representation



Delora Baptista, et.al. Deep learning for drug response prediction in cancer. Briefings in Bioinformatics, 22(1). (2021)

Data Preprocessing
- Drug

- Drug Representation



Jintae Kim, et.al. Comprehensive Survey of Recent Drug Discovery Using Deep Learning. International Journal of Molecular Sciences. (2021)

Data Preprocessing
- Drug

-Fingerprint

- **Morgan fingerprints/Extended Connectivity Fingerprints (ECFP)**
- Daylight fingerprints

⇒ Previous models: DeepDSC, CDRScan

-SMILES Molecular Graph

- Molecular Structure → Graph
  - ConvMolFeaturizer
  - **MolGraphConvFeaturizer**
  - Manual

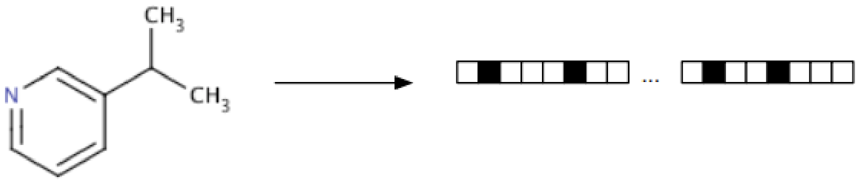
⇒ Previous models: PaccMann, DeepCDR

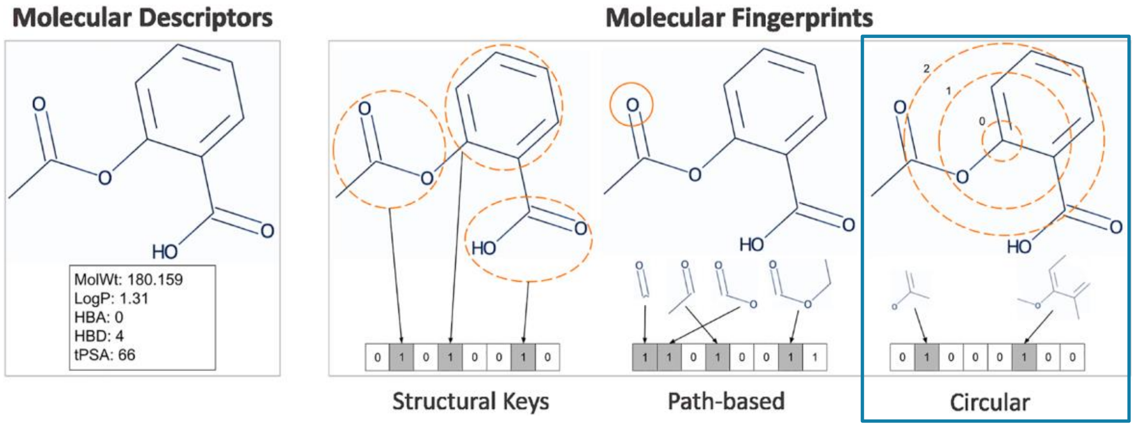
Data Preprocessing
- Drug: Fingerprint

Fingerprint

Morgan Fingerprints

- Numbering invariant atom information into an initial atom identifier
- Identifiers are generated independently of previous identifiers and intermediate results are discarded
- The iteration process is continued until every atom identifier is unique





Delora Baptista, et.al. Evaluating molecular representations in machine learning models for drug response prediction and interpretability. Journal of Integrative Bioinformatics. (2022)

Data Preprocessing
- Drug: Fingerprint

Fingerprint

Morgan Fingerprints

	Cell Line	Gene expression	Drug Name	Fingerprint	IC50
	ACH-000242	[2.042524 … 0.6824865]	cabozantinib	[00000000000000000000100 … 000000]	1.353158
	…	…	…	…	…
	ACH-000052	[0.5340133 … -0.6967168]	vinblastine	[00000000100000000000000 … 000000]	-6.64979
Total	**695 cell lines**	**13,106 genes**	**310 drugs**		**41,460** **IC50 values** **(cell line-drug pairs)**

Min Li, et.al. DeepDSC: A Deep Learning Method to Predict Drug Sensitivity of Cancer Cell Lines. IEEE/ACM Trans Comput Biol Bioinform. (2019)

Data Preprocessing

- Drug: SMILES

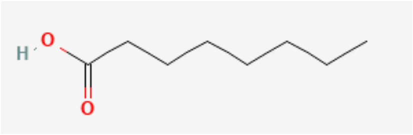
SMILES

MolGraphConvFeaturizer

- **Node Features: Atom properties**
  - Feature length: 30
    - Atom type: A one-hot vector of this atom, “C”, “N”, “O”, “F”, “P”, “S”, “Cl”, “Br”, “I”, “other atoms”.
    - Formal charge: Integer electronic charge.
    - Hybridization: A one-hot vector of “sp”, “sp2”, “sp3”.
    - Hydrogen bonding: A one-hot vector of whether this atom is a hydrogen bond donor or acceptor.
    - Aromatic: A one-hot vector of whether the atom belongs to an aromatic ring.
    - Degree: A one-hot vector of the degree (0-5) of this atom.
    - Number of Hydrogens: A one-hot vector of the number of hydrogens (0-4) that this atom connected.
    - Chirality: A one-hot vector of the chirality, “R” or “S”. (Optional)
    - Partial charge: Calculated partial charge. (Optional)

- **Edge Index**
  - [src, dest]

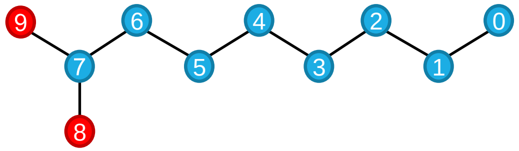
Octanoic acid



SMILES

CCCCCCCC(=O)O

Molecular Graph



Atom features: (# of nodes, 30)

1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0
1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0
1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0
1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0
1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0
1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0
1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0
1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0
0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	1	0	0	0	0	0	1	0	0	0	0	0
0	0	1	0	0	0	0	0	0	0	0	0	1	0	1	1	0	0	0	1	0	0	0	0	0	1	0	0	0	0

10 × 30

Adjacency list: (2, 2 × # of edges)

0	1	1	2	2	3	3	4	4	5	5	6	6	7	7	8	7	9
1	0	2	1	3	2	4	3	5	4	6	5	7	6	8	7	9	7

2 × 18

- GraphData(node_features=atom_features, edge_index=np.asarray([src, dest], dtype=int), edge_features=bond_features, node_pos_features=node_pos_features)

Qiao Liu, et.al. DeepCDR: a hybrid graph convolutional network for predicting cancer drug response. Bioinformatics, 36. (2020)

Data Preprocessing
- Drug: SMILES

SMILES

MolGraphConvFeaturizer

- Node Features: Atom properties
- Edge Index
- **GraphData(node_features=atom_features, edge_index=np.asarray([src, dest], dtype=int), edge_features=bond_features, node_pos_features=node_pos_features**)

	Cell Line	Gene expression	Drug Name	SMILES	Graph (V, E)	IC50
	ACH-000242	[2.042524 … 0.6824865]	cabozantinib	COC1=CC2=C(C=CN=C2C=C1OC)OC3=CC=C(C=C3)NC(=O)C4(CC4)C(=O)NC5=CC=C(C=C5)F	[GraphData(node_features=[37, 30], edge_index=[2, 82], edge_features=None, pos=[0])]	1.353158
	…	…	…	…	…	…
	ACH-000052	[0.5340133 … -0.6967168]	vinblastine	CCC1(CC2CC(C3=C(CCN(C2)C1)C4=CC=CC=C4N3)(C5=C(C=C6C(=C5)C78CCN9C7C(C=CC9)(C(C(C8N6C)(C(=O)OC)O)OC(=O)C)CC)OC)C(=O)OC)O	[GraphData(node_features=[59, 30], edge_index=[2, 134], edge_features=None, pos=[0])]	-6.64979
Total	**695 cell lines**	**13,106 genes**		**310 drugs**		**41,460 IC50 values (cell line-drug pairs)**

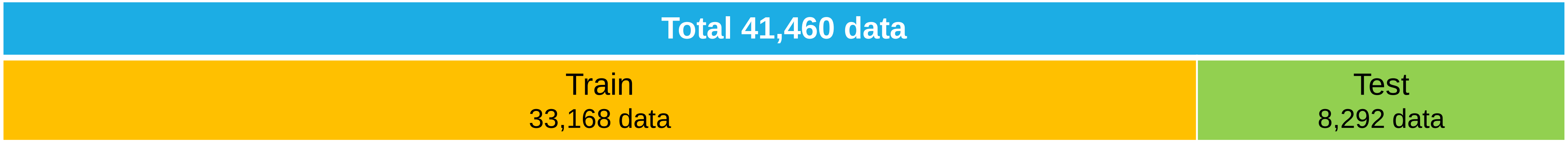
Qiao Liu, et.al. DeepCDR: a hybrid graph convolutional network for predicting cancer drug response. Bioinformatics, 36. (2020)

Data

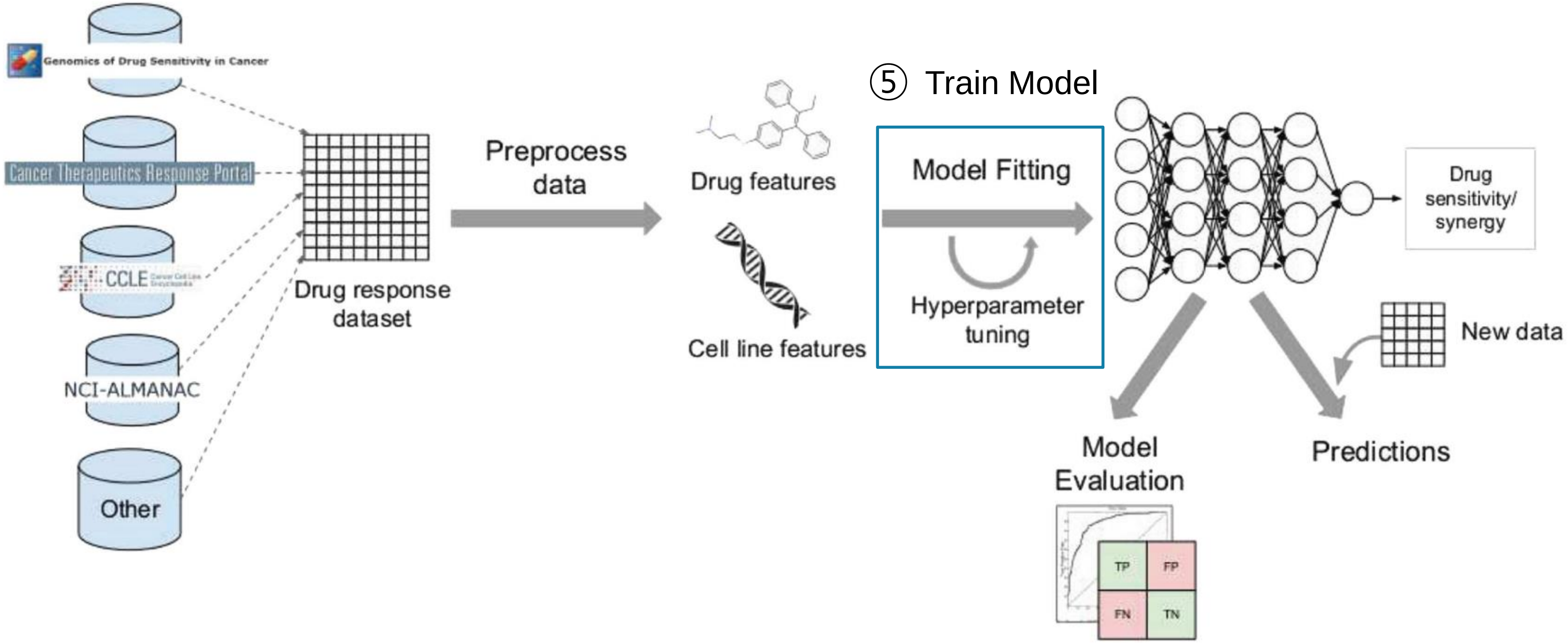
- Total data

	Cell Line	Gene expression	Drug Name	SMILES	Graph (V, E)	IC50
	ACH-000242	[2.042524 … 0.6824865]	cabozantinib	COC1=CC2=C(C=CN=C2C=C1OC)OC3=CC=C(C=C3)NC(=O)C4(CC4)C(=O)NC5=CC=C(C=C5)F	[GraphData(node_features=[37, 30], edge_index=[2, 82], edge_features=None, pos=[0])]	1.353158
	…	…	…	…	…	…
	ACH-000052	[0.5340133 … -0.6967168]	vinblastine	CCC1(CC2CC(C3=C(CCN(C2)C1)C4=CC=CC=C4N3)(C5=C(C=C6C(=C5)C78CCN9C7C(=CC9)(C(C(C8N6C)(C(=O)OC)O)OC(=O)C)CC)OC)C(=O)OC)O	[GraphData(node_features=[59, 30], edge_index=[2, 134], edge_features=None, pos=[0])]	-6.64979
Total	**695 cell lines**	**13,106 genes**		**310 drugs**		**41,460 IC50 values (cell line-drug pairs)**

- Train : test = 80 : 20

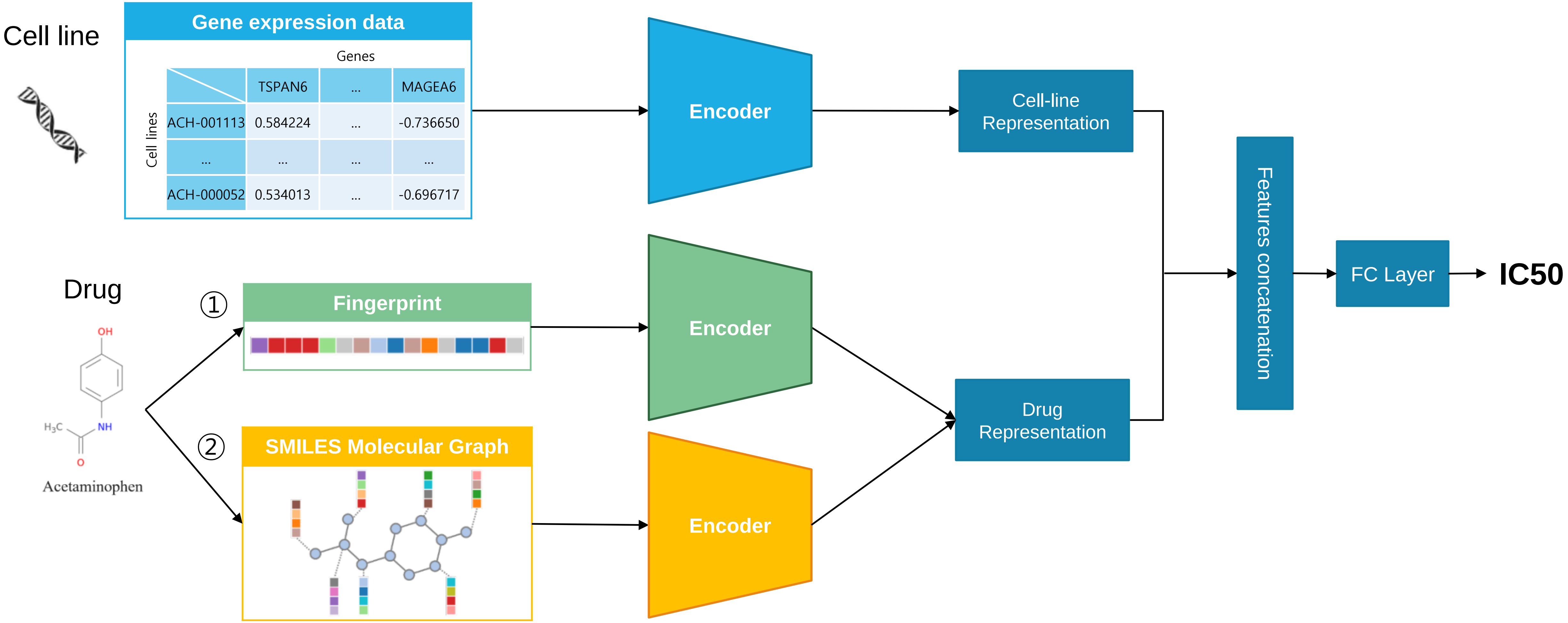


Deep learning workflows for drug response prediction



Delora Baptista, et.al. Deep learning for drug response prediction in cancer. Briefings in Bioinformatics, 22(1). (2021)

Model Architecture



Cell line

Gene expression data

Genes

Cell lines

	TSPAN6	...	MAGEA6
ACH-001113	0.584224	...	-0.736650
...	...	...	...
ACH-000052	0.534013	...	-0.696717

Drug

Acetaminophen

① Fingerprint

② SMILES Molecular Graph

Encoder

Encoder

Encoder

Cell-line Representation

Drug Representation

Features concatenation

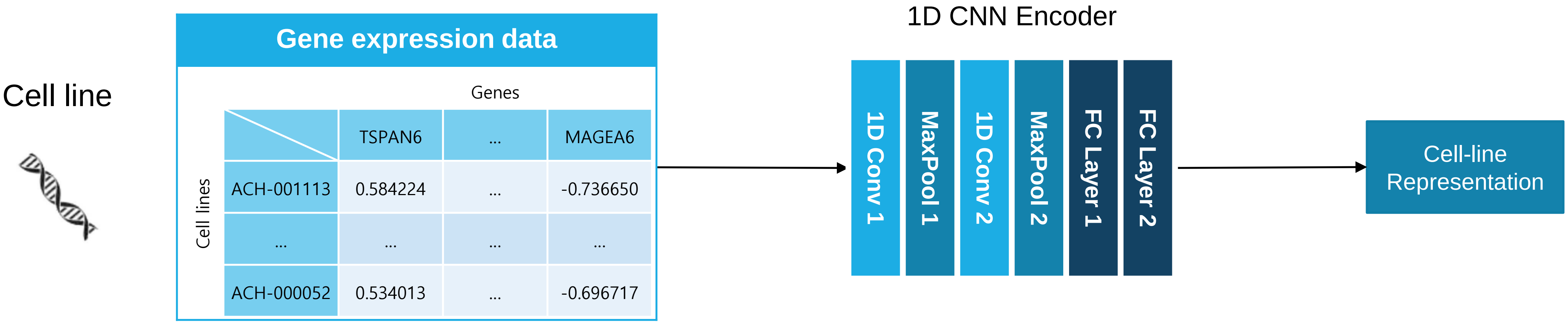
FC Layer

IC50

Train

- Gene expression

1) 1D CNN Encoder



Cell line

Gene expression data

Genes

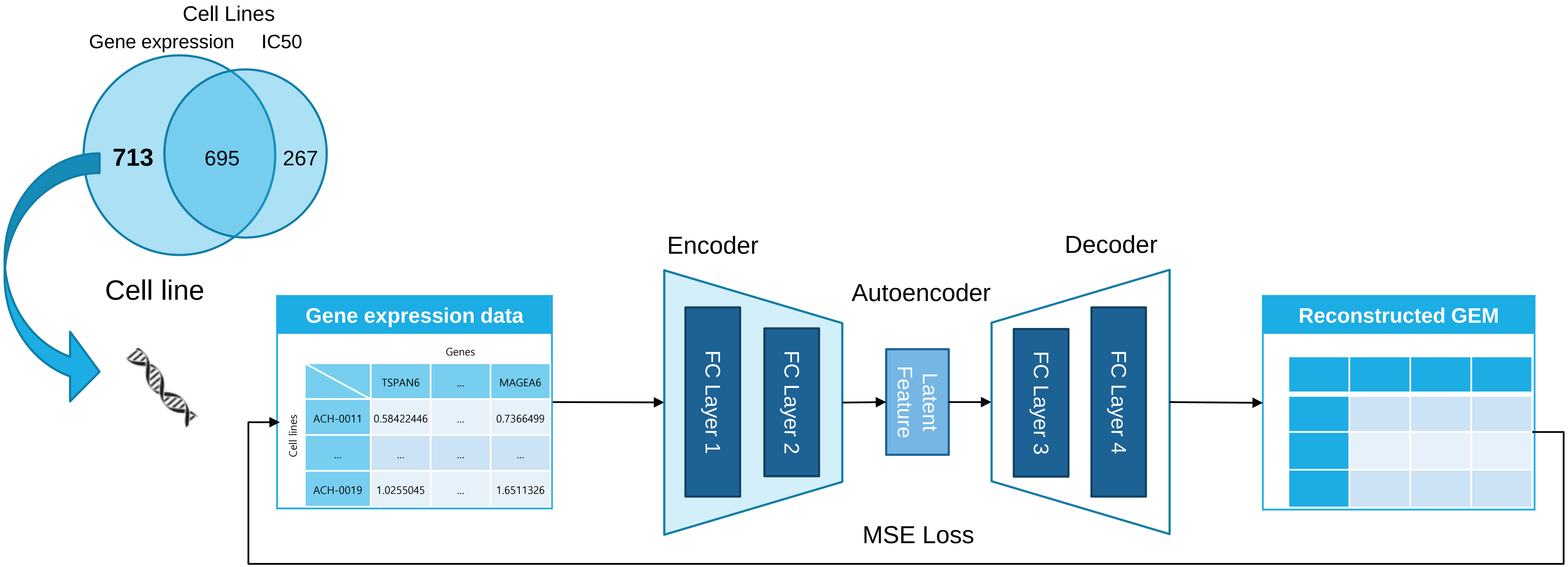
Cell lines

	TSPAN6	…	MAGEA6
ACH-001113	0.584224	…	-0.736650
…	…	…	…
ACH-000052	0.534013	…	-0.696717

1D CNN Encoder

Train
- Gene expression

2) Pretrained Encoder from Autoencoder



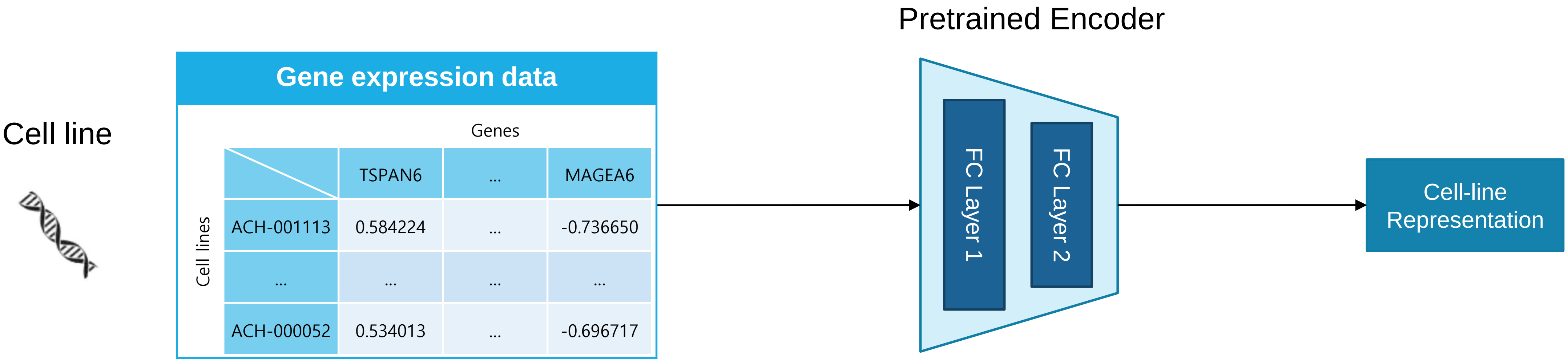
Genes

Cell lines	TSPAN6	...	MAGEA6
ACH-0011	0.58422446	...	0.7366499
...	...	...	...
ACH-0019	1.0255045	...	1.6511326

Train

- Gene expression

2) Pretrained Encoder from Autoencoder

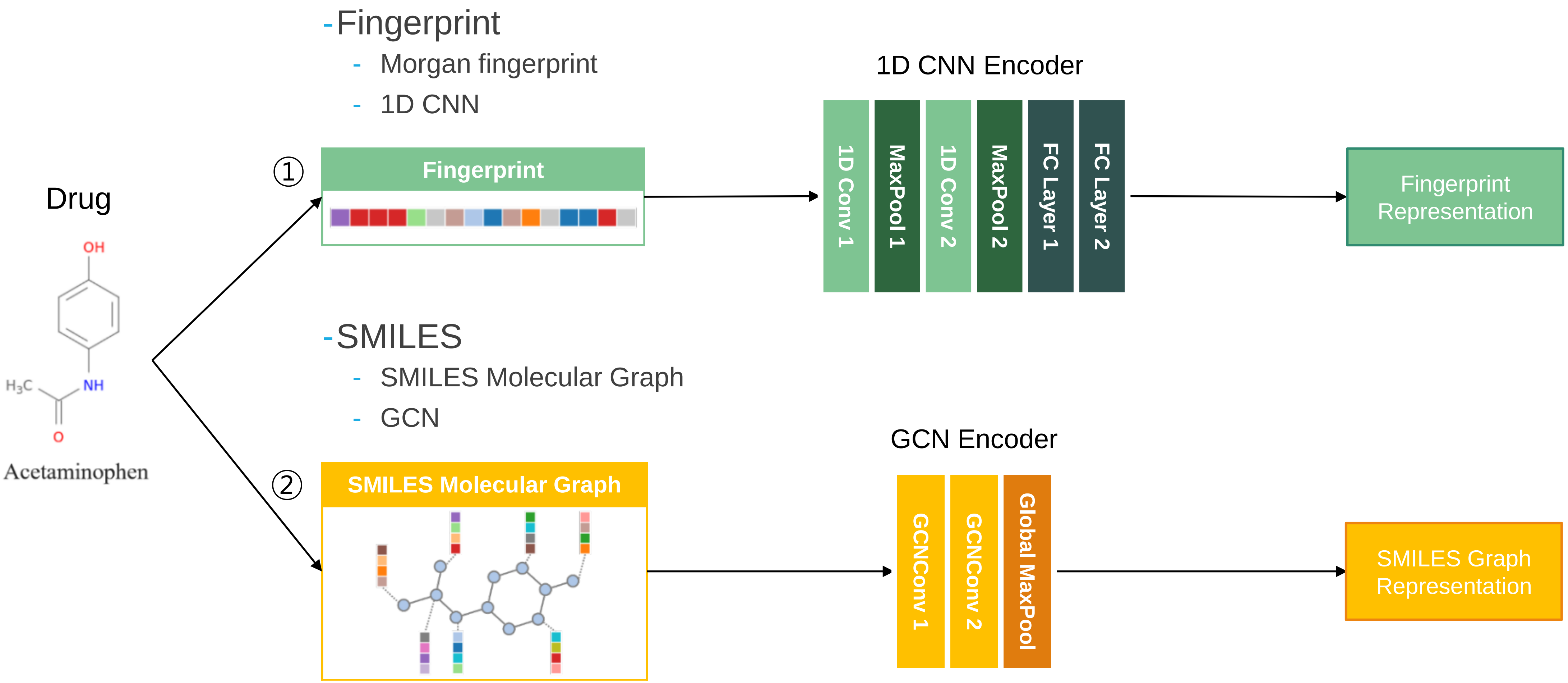


Gene expression data

Genes

Cell lines		TSPAN6	...	MAGEA6
	ACH-001113	0.584224	...	-0.736650
	...	...	...	...
	ACH-000052	0.534013	...	-0.696717

Train
- Drug



- Fingerprint
  - Morgan fingerprint
  - 1D CNN

- SMILES
  - SMILES Molecular Graph
  - GCN

Drug

Acetaminophen

Integration

- Early integration

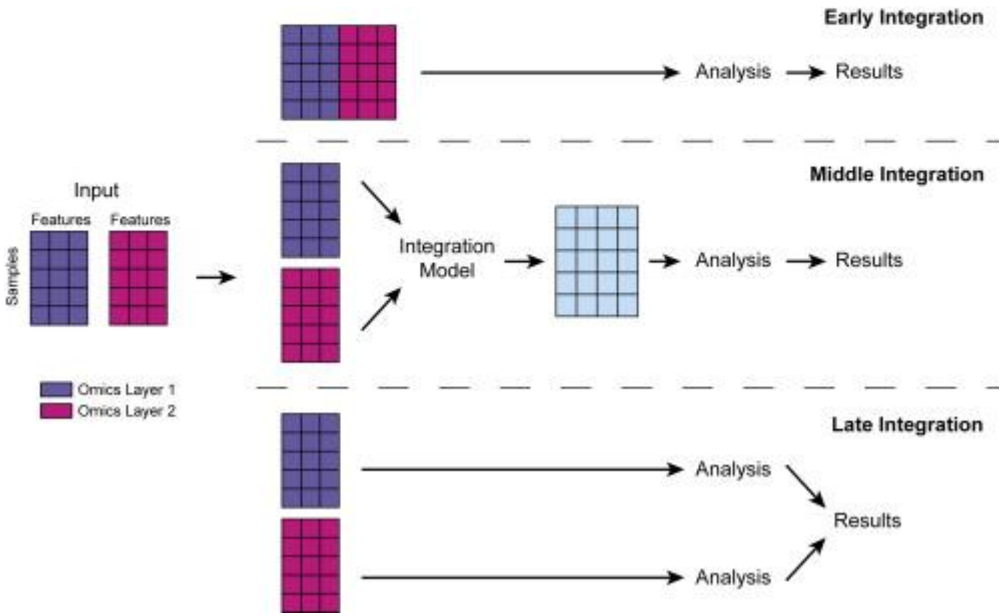
  : features from different data matrices are concatenated

- Middle integration

  : uses machine learning models to consolidate data without concatenating features or merging results

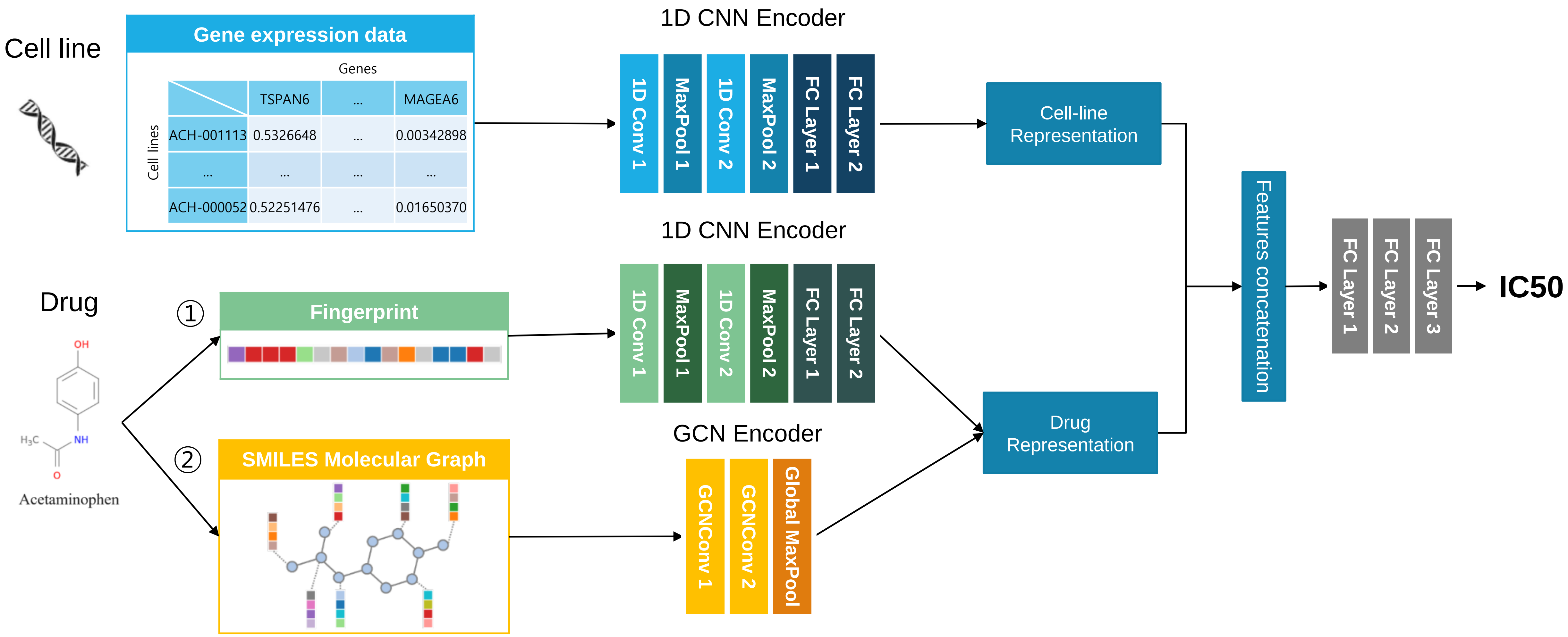
- Late integration

  : each omics layer is analyzed independently, and results are combined at the end



Zhaoxiang Cai, et.al. Machine learning for multi-omics data integration in cancer. iScience, 25(2). (2022)

Model Architecture



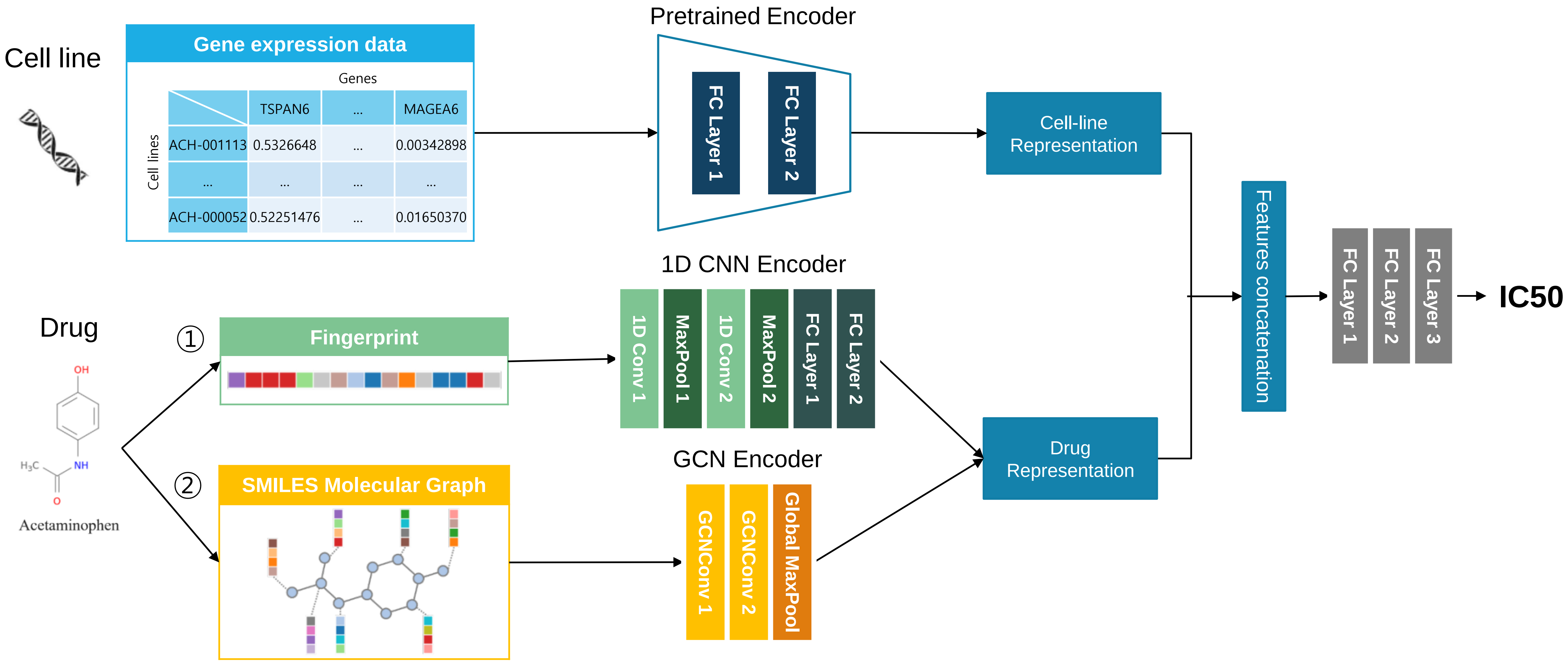
Gene expression data

Genes

Cell lines

	TSPAN6	...	MAGEA6
ACH-001113	0.5326648	...	0.00342898
...	...	...	...
ACH-000052	0.52251476	...	0.01650370

Model Architecture



Pretrained Encoder

Cell line

Gene expression data

Genes

Cell lines

	TSPAN6	...	MAGEA6
ACH-001113	0.5326648	...	0.00342898
...	...	...	...
ACH-000052	0.52251476	...	0.01650370

FC Layer 1

FC Layer 2

Cell-line Representation

Features concatenation

FC Layer 1

FC Layer 2

FC Layer 3

IC50

Drug

Acetaminophen

① Fingerprint

1D CNN Encoder

1D Conv 1

MaxPool 1

1D Conv 2

MaxPool 2

FC Layer 1

FC Layer 2

② SMILES Molecular Graph

GCN Encoder

GCNConv 1

GCNConv 2

Global MaxPool

Drug Representation

Hyperparameters

- Epoch: 100
- Batch size: 64
- Learning rate: 0.0001
- Dropout ratio: 0.1
- Optimization: Adam Optimizer

Evaluation

- MSE loss

$$MSE = \frac{1}{n}\sum_{i}^{n}(y_i - \hat{y}_i)^2$$

- Pearson Correlation

$$CC_P = \frac{\sum_i^n (y_i - \hat{y}_i)^2}{\sigma_O \sigma_Y}$$

Result

- Performance comparison in terms of CC_p and $RMSE$ on CCLE dataset

		CC_p	$RMSE$
Fingerprint	Original	0.9417	1.535
	Pretrained GE Autoencoder	0.9417	1.535
SMILES Graph	Original	0.9348	1.719
	Pretrained GE Autoencoder	0.9361	1.681

Train
- Gene expression

- Gene expression
    - 1D CNN

Name	Output Size	Layer
Initial	(64, 13106)	
Conv 1	(64, 8, 13099)	Conv1d(1, 8, kernel_size=(8,), stride=(1,))
Pool 1	(64, 8, 4366)	MaxPool1d(kernel_size=3, stride=3, padding=0, dilation=1, ceil_mode=False)
Conv 2	(64, 16, 4359)	Conv1d(8, 16, kernel_size=(8,), stride=(1,))
Pool 2	(64, 16, 1453)	MaxPool1d(kernel_size=3, stride=3, padding=0, dilation=1, ceil_mode=False)
Fully Connected (FC) 1	(64, 1024)	Linear(in_features=23248, out_features=1024, bias=True)
Fully Connected (FC) 2	(64, 128)	Linear(in_features=1024, out_features=128, bias=True)

Train
- Drug Fingerprint

- Drug
  - Fingerprint

Name	Output Size	Layer
Initial	(64, 2048)	
Conv 1	(64, 8, 2041)	Conv1d(1, 8, kernel_size=(8,), stride=(1,))
Pool 1	(64, 8, 680)	MaxPool1d(kernel_size=3, stride=3, padding=0, dilation=1, ceil_mode=False)
Conv 2	(64, 16, 673)	Conv1d(8, 16, kernel_size=(8,), stride=(1,))
Pool 2	(64, 16, 224)	MaxPool1d(kernel_size=3, stride=3, padding=0, dilation=1, ceil_mode=False)
Fully Connected (FC) 1	(64, 1024)	Linear(in_features=3584, out_features=1024, bias=True)
Fully Connected (FC) 2	(64, 128)	Linear(in_features=1024, out_features=128, bias=True)

Train
- Drug SMILES Graph

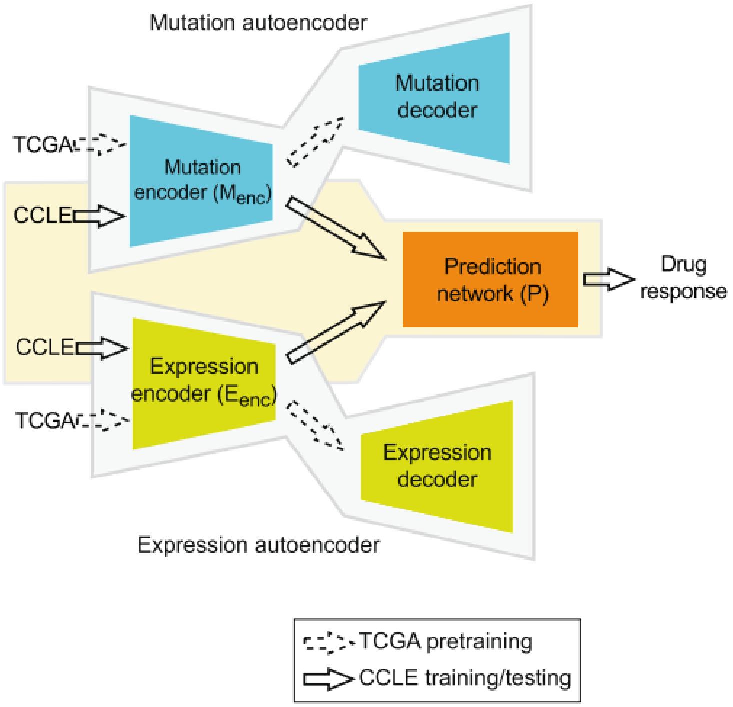
- Drug
  - SMILES Graph

Name	Output Size	Layer
Initial	(64, - ,30)	
GCNConv 1	(- , 64)	GCNConv(30, 64)
GCNConv 2	(- , 128)	GCNConv(64, 128)
GlobalMaxpool	(64, 128)	

Future Work

- Gene expression
  - Feature Embedding
    - Autoencoder

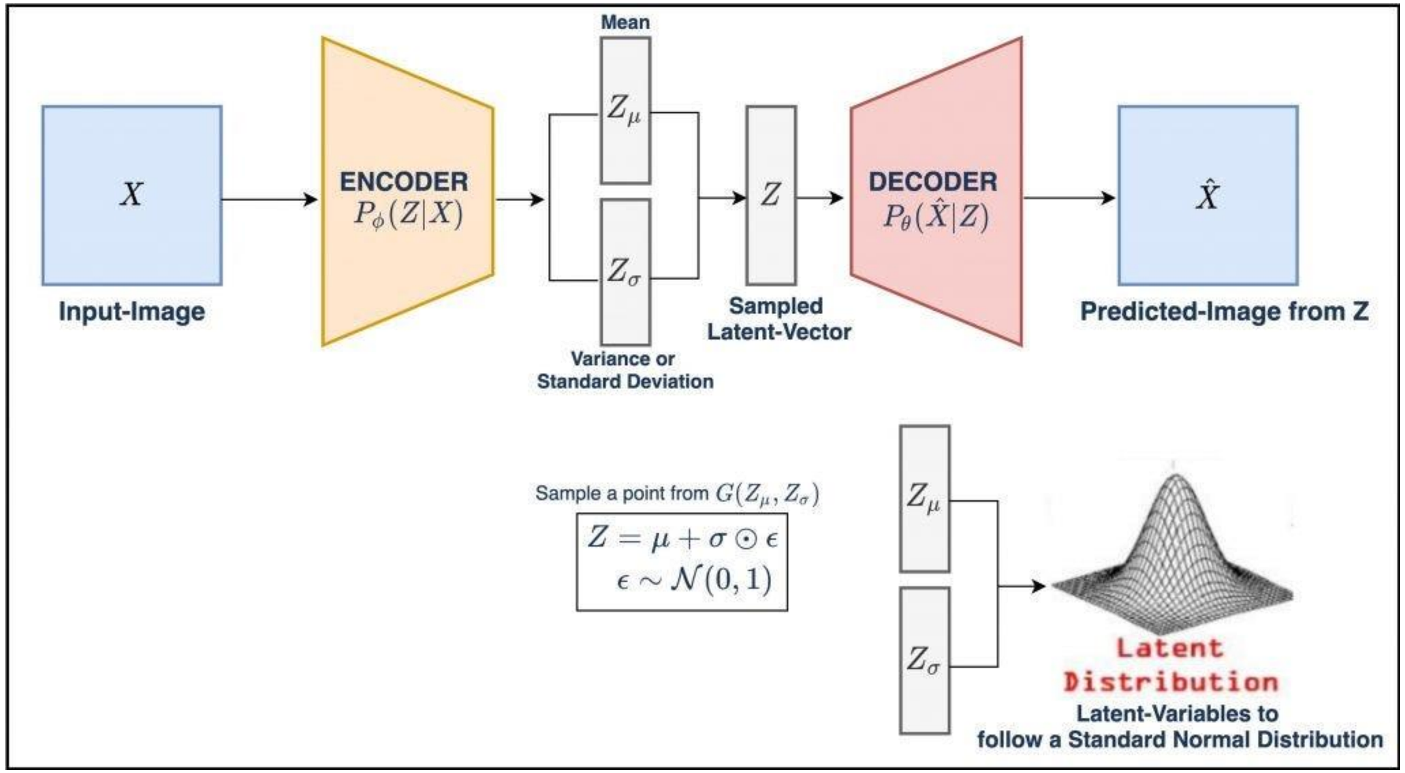
    : pretraining with TCGA database



Yu-Chiao Chiu, et.al. Predicting drug response of tumors from integrated genomic profiles by deep neural networks. BMC Medical Genomics. (2019)

Future Work

- Gene expression
  - Feature Embedding
    - Variational Autoencoder

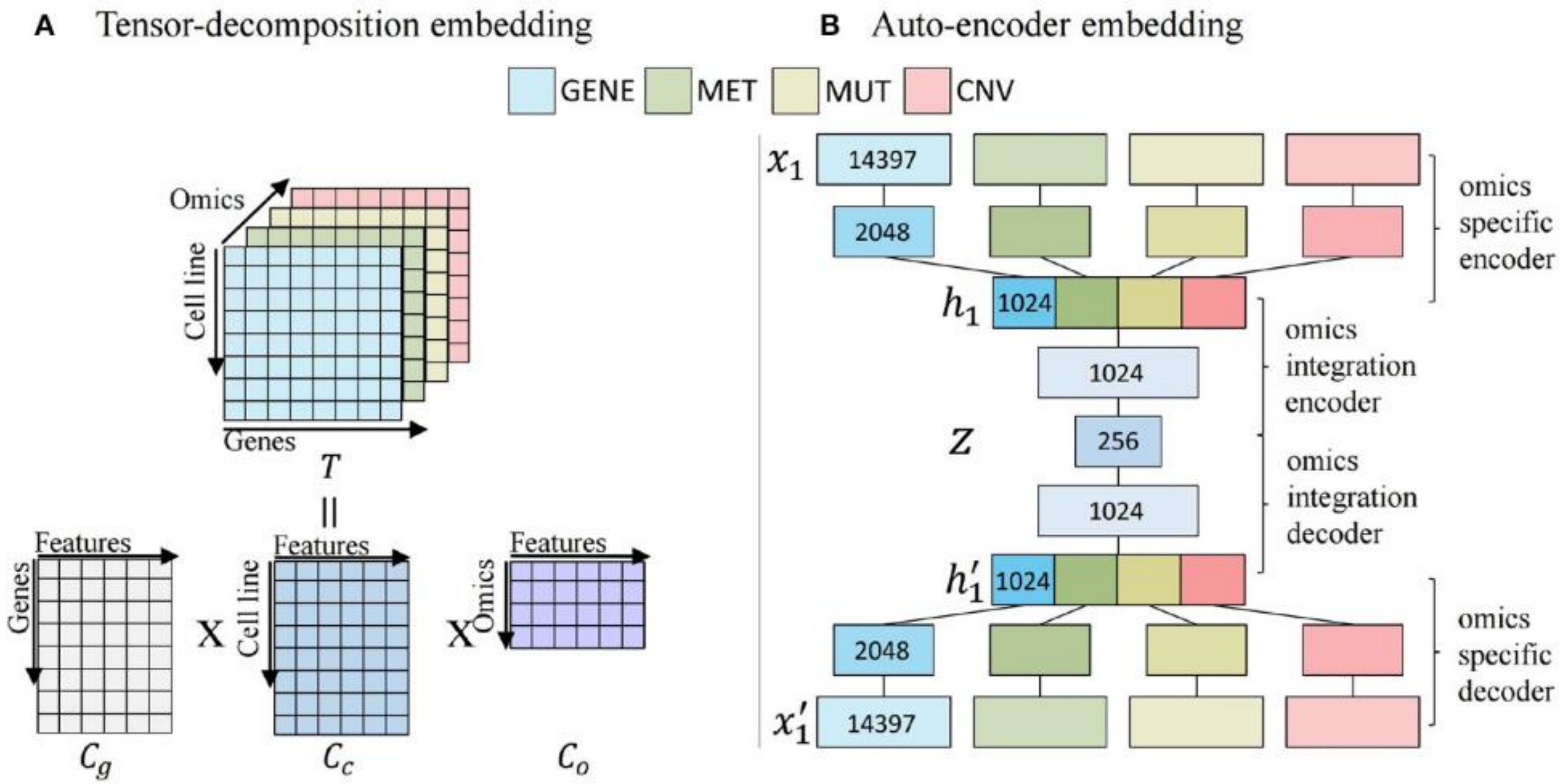


Mean
Z_μ
X
ENCODER
$P_\phi(Z|X)$
Z
DECODER
$P_\theta(\hat{X}|Z)$
$\hat{X}$
Z_σ
Input-Image
Sampled Latent-Vector
Predicted-Image from Z
Variance or Standard Deviation
Sample a point from $G(Z_\mu, Z_\sigma)$
$Z = \mu + \sigma \odot \epsilon$
$\epsilon \sim \mathcal{N}(0,1)$
Z_μ
Z_σ
Latent Distribution
Latent-Variables to follow a Standard Normal Distribution

From https://learnopencv.com/variational-autoencoder-in-tensorflow/

Future Work

- Multi-omics data



Minsik Oh, et.al. DRIM: A Web-Based System for Investigating Drug Response at the Molecular Level by Condition-Specific Multi-Omics Data Integration. Frontiers in Genetics. (2020)