Comparative Analysis of Drug Compound Representation and Perceived Reactivity: Chemception Embedding vs.Chemical Fingerprinting

CogSys students:
Swapnil Jha
Milena Voskanyan
Module: Project Module

Goal of the Project

Problem Statement:
We aim to determine which method is more accurate for predicting cancer cell sensitivity to a given chemical and why.

Methods:

1. **Fingerprint Approach:**
    - Use chemical fingerprints (Morgan & MACCS)
    - Apply a numerical input-output model with machine learning techniques.
2. **Chemception Approach:**
    - Represent chemicals as images using Chemception embeddings.
    - Train a CNN to predict sensitivity.

Dataset

Dataset Composition:

- 981 genetically characterized human cancer cell lines.
- Screened with a wide range of anti-cancer drugs.

Objective:

- To correlate drug sensitivity patterns of cancer cell lines with extensive genomic and expression data.

Impact:

- Captures the genomic heterogeneity of human cancers, helping explain why patients have variable responses to the same treatment.

Availability:

- Freely accessible data for the academic and medical communities through the GDSC website.

Dataset

GDSC fitted dose response description

Possible columns in GDSC fitted data results file. Not every listed column is present in every file.

Column	Description	Notes
DATASET_VERSION	Each dataset is processed (curve fitted and ANOVA analysis) as a whole.	
IC50_RESULTS_ID	Identifier for the fitted dose response	
COSMIC_ID	Cell identifier from the COSMIC database	
CELL_LINE_NAME	Primary name for the cell line	
DRUG_ID	Unique identifier for a drug. Used for internal lab tracking	
DRUG_NAME	Primary name for the drug	
PUTATIVE_TARGET	Putative drug target	

Version 1.0.0 - 21 September 2017
https://docs.google.com/a/sanger.ac.uk/document/d/1YKK1BE5FNITaMVYfhEQ_QD6juzBZU_ZybKXGuN8epKM

MAX_CONC_MICROMOLAR	Maximum micromolar screening concentration of the drug	
MIN_CONC_MICROMOLAR	Minimum micromolar screening concentration of the drug	
LN_IC50	Natural log of the fitted IC50	To convert to micromolar take the exponent of this value, i.e. exp(IC50_nat_log)
AUC	Area Under the Curve for the fitted model. Presented as a fraction of the total area between the highest and lowest screening concentration.	
RMSE	Root Mean Squared Error, a measurement of how well the modelled curve fits the data points.	Curves with RMSE > 0.3 are excluded prior to release as part of quality control.
Z_SCORE	Z score of the LN_IC50 (x) comparing it to the mean (μ) and standard deviation (σ^2) of the LN_IC50 values for the drug in question over all cell lines treated.	$Z = \frac{x - \mu}{\sigma^2}$

Dataset

```
for col in data.columns[:7]:
    print(col, data[col].nunique())
    print(col, data[col].unique()[:10])
    print()
```

```
DATASET 1
DATASET ['GDSC2']

NLME_RESULT_ID 1
NLME_RESULT_ID [343]

NLME_CURVE_ID 242036
NLME_CURVE_ID [15946310 15946548 15946830 15947087 15947369 15947651 15947932 15948212
 15948491 15948772]

COSMIC_ID 969
COSMIC_ID [683667 684052 684057 684059 684062 684072 687448 687452 687455 687457]

CELL_LINE_NAME 969
CELL_LINE_NAME ['PFSK-1' 'A673' 'ES5' 'ES7' 'EW-11' 'SK-ES-1' 'COLO-829' '5637' 'RT4'
 'SW780']

SANGER_MODEL_ID 969
SANGER_MODEL_ID ['SIDM01132' 'SIDM00848' 'SIDM00263' 'SIDM00269' 'SIDM00203' 'SIDM01111'
 'SIDM00909' 'SIDM00807' 'SIDM01085' 'SIDM01160']

TCGA_DESC 32
TCGA_DESC ['MB' 'UNCLASSIFIED' 'SKCM' 'BLCA' 'CESC' 'GBM' 'LUAD' 'LUSC' 'SCLC'
 'MESO']
```

- Columns ‘DATASET’ and ‘NLME_RESULT_ID’ have only one value ‘GDSC2’ and ‘343’ respectively
- No information found on ‘NLME_CURVE_ID’ column. Assumption: R package (https://cran.r-project.org/web/packages/nlme/nlme.pdf)

- COSMIC_ID: Cell identifier from the COSMIC database.
- We could not find any information on how to incorporate the column in model training process by exploring official COSMIC educational videos and materials (*https://youtu.be/bvY7wt9djG4*)
- We failed to find the given COSMIC IDs on the official COSMIC dataset website

```
data[['COSMIC_ID', 'CELL_LINE_NAME', 'SANGER_MODEL_ID']].drop_duplicates()
```

	COSMIC_ID	CELL_LINE_NAME	SANGER_MODEL_ID
0	683667	PFSK-1	SIDM01132
1	684052	A673	SIDM00848
2	684057	ES5	SIDM00263
3	684059	ES7	SIDM00269
4	684062	EW-11	SIDM00203
...	...	...	...
964	1660035	SNU-61	SIDM00194
965	1660036	SNU-81	SIDM00193
966	1674021	SNU-C5	SIDM00498
967	1789883	DiFi	SIDM00049
2360	1290906	HCC202	SIDM00870

969 rows × 3 columns

```
for col in data.columns[:7]:
    print(col, data[col].nunique())
    print(col, data[col].unique()[:10])
    print()
```

```
COSMIC_ID 969
COSMIC_ID [683667 684052 684057 684059 684062 684072 687448 687452 687455 687457]

CELL_LINE_NAME 969
CELL_LINE_NAME ['PFSK-1' 'A673' 'ES5' 'ES7' 'EW-11' 'SK-ES-1' 'COLO-829' '5637' 'RT4'
 'SW780']

SANGER_MODEL_ID 969
SANGER_MODEL_ID ['SIDM01132' 'SIDM00848' 'SIDM00263' 'SIDM00269' 'SIDM00203' 'SIDM01111'
 'SIDM00909' 'SIDM00807' 'SIDM01085' 'SIDM01160']
```

Columns with Minimal Information Contribution

Removing the columns *'TCGA_DESC'*, *'PUTATIVE_TARGET'*, and *'PATHWAY_NAME'* results in less than 1% information loss

```
print(data.shape)
data2 = data[[
    'CELL_LINE_NAME',
    # 'TCGA_DESC',
    'DRUG_NAME',
    # 'PUTATIVE_TARGET',
    # 'PATHWAY_NAME',
    'MIN_CONC',
    'MAX_CONC',
    'LN_IC50']].drop_duplicates()
print(data2.shape)
data3 = data2[[
    'CELL_LINE_NAME',
    # 'TCGA_DESC',
    'DRUG_NAME',
    # 'PUTATIVE_TARGET',
    # 'PATHWAY_NAME',
    'MIN_CONC',
    'MAX_CONC'
    ]].drop_duplicates()
print(data3.shape)
print("Data Loss :", round((data2.shape[0]-data3.shape[0])*100/data2.shape[0], 2), "%")
```

```
(242036, 19)
(242036, 5)
(239995, 4)
Data Loss : 0.84 %
```

Columns with Minimal Information Contribution

	PUTATIVE_TARGET	count
0	NaN	27155
1	PARP1, PARP2	4714
2	MEK1, MEK2	4547
3	TOP1	4325
4	EGFR	3836
...	...	...
181	Induces reactive oxygen species	225
182	RSK, AURKB, PIM1, PIM3	225
183	EGLN1	225
184	TBK1, PDK1 (PDPK1), IKK, AURKB, AURKC	225
185	AR	225

186 rows × 2 columns

	PATHWAY_NAME	count
0	Unclassified	24979
1	PI3K/MTOR signaling	22724
2	Other	21402
3	DNA replication	17650
4	Other, kinases	17277
5	ERK MAPK signaling	13350
6	Genome integrity	12221
7	Cell cycle	11620
8	Apoptosis regulation	10828
9	Chromatin histone methylation	10612

	DRUG_ID	DRUG_NAME
0	1803	Acetalax
1	1804	Acetalax
2	1811	Dactinomycin
3	1911	Dactinomycin
4	1007	Docetaxel
5	1819	Docetaxel
6	1200	Fulvestrant
7	1816	Fulvestrant
8	1627	GSK343
9	2037	GSK343

Chosen Columns from the Dataset

	CELL_LINE_NAME	DRUG_NAME	MIN_CONC	MAX_CONC	LN_IC50
0	PFSK-1	Camptothecin	0.000100	0.1	-1.463887
1	A673	Camptothecin	0.000100	0.1	-4.869455
2	ES5	Camptothecin	0.000100	0.1	-3.360586
3	ES7	Camptothecin	0.000100	0.1	-5.044940
4	EW-11	Camptothecin	0.000100	0.1	-3.741991
...	...	...	...	...	...
242031	SNU-175	N-acetyl cysteine	2.001054	2000.0	10.127082
242032	SNU-407	N-acetyl cysteine	2.001054	2000.0	8.576377
242033	SNU-61	N-acetyl cysteine	2.001054	2000.0	10.519636
242034	SNU-C5	N-acetyl cysteine	2.001054	2000.0	10.694579
242035	DiFi	N-acetyl cysteine	2.001054	2000.0	10.034825

242036 rows × 5 columns

Data Preprocessing Chemception

XAV939 ⇒ C1CSCC2=C1N=C(NC2=O)C3=CC=C(C=C3)C(F)(F)F ⇒ <rdkit.Chem.rdchem.Mol object at 0x7d610a1e8dd0> ⇒ [[[0.0, 0.0, 0.0, 0.0], [0.0, 0.0, 0.0, 0.0], …]]

molimage

Input: Takes a molecular object (`mol`).

Grid Creation: It creates a 2D grid based on the molecule's dimensions, which helps in organizing data.

Coordinates: Calculates the positions of atoms and bonds in the molecule.

Bonds: For each bond, it records:

- **Bond Order**: Indicates the strength/type of the bond.

Atoms: For each atom, it captures:

- **Atomic Number**: Identifies the type of element.
- **Gasteiger Charges**: Represents the charge on the atom.
- **Hybridization**: Describes the bonding characteristics of the atom.

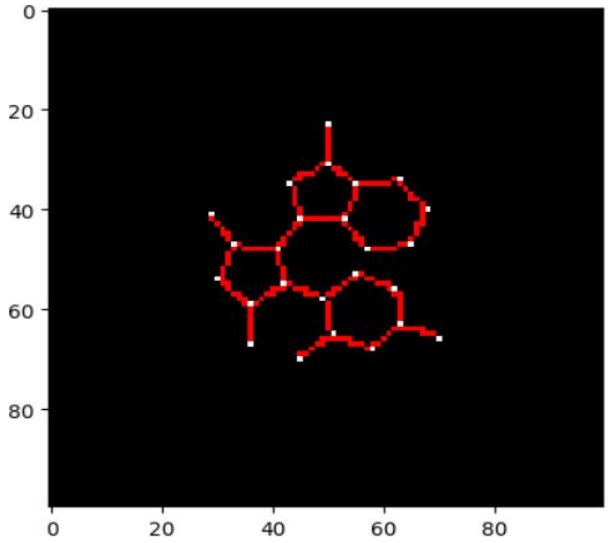


Image Augmentation

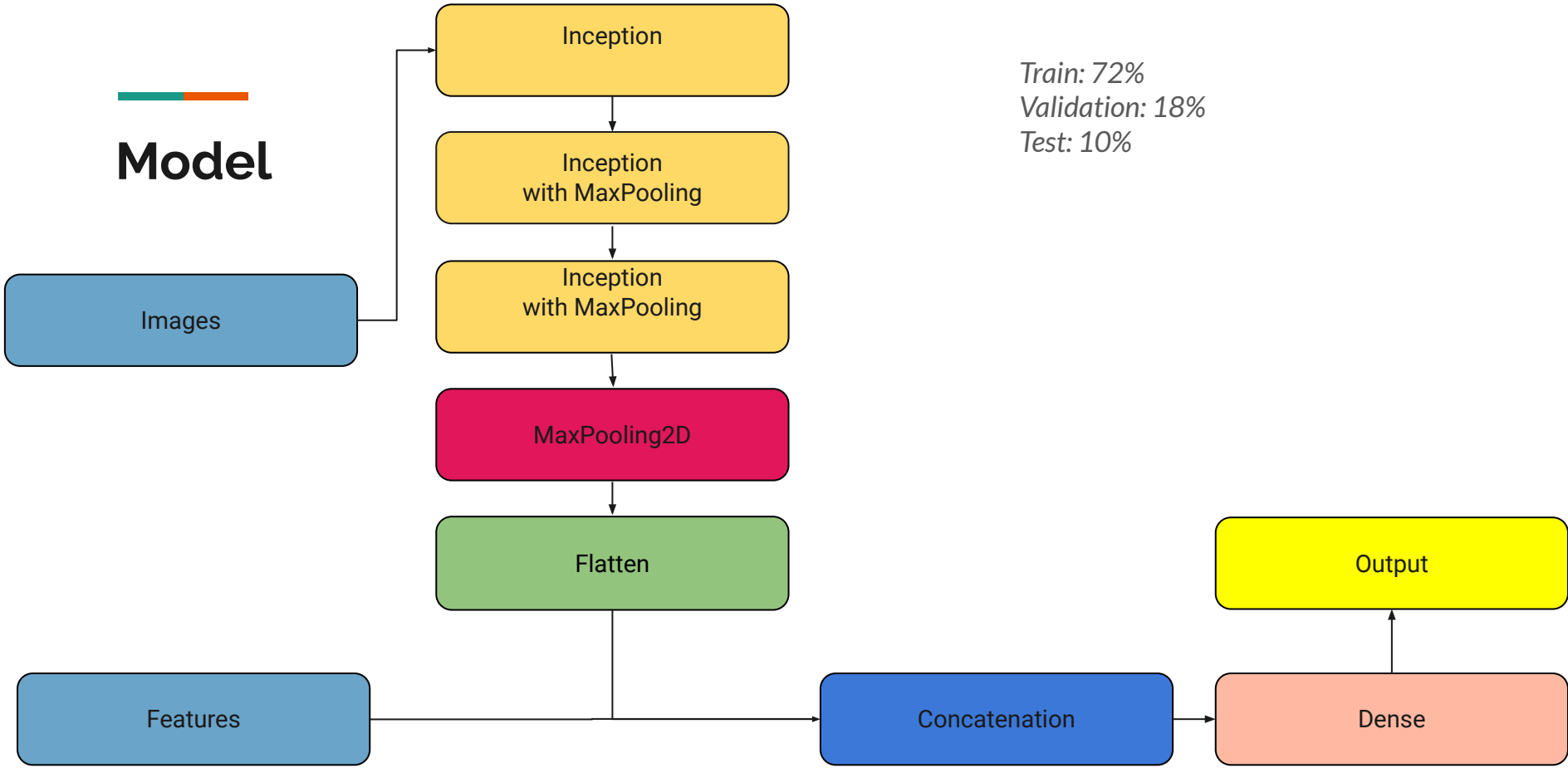
Rotation: Randomly rotates images up to 180 degrees.

Width/Height Shift: Shifts images horizontally and vertically by 10% of the total size.

Fill Mode: Fills any empty pixels (due to transformations) with a constant value (0, black).

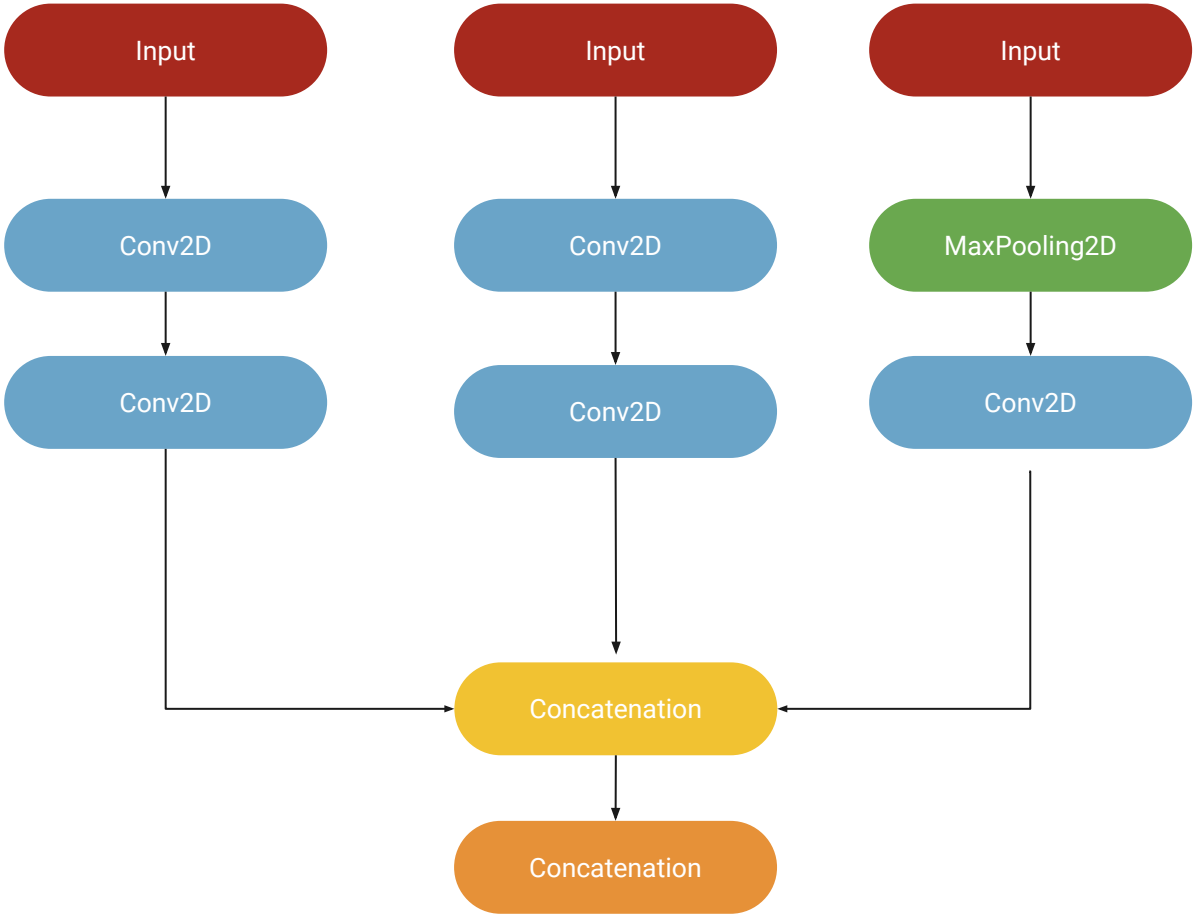
Flipping: Randomly flips images horizontally and vertically.

Model



Train: 72%
Validation: 18%
Test: 10%

Inception Layer



Evaluation Metrics

Root Mean Squared Error

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

Mean Absolute Error

$$MAE - \frac{1}{n}\sum_{i-1}^{n}|y_i - \hat{y}_i|$$

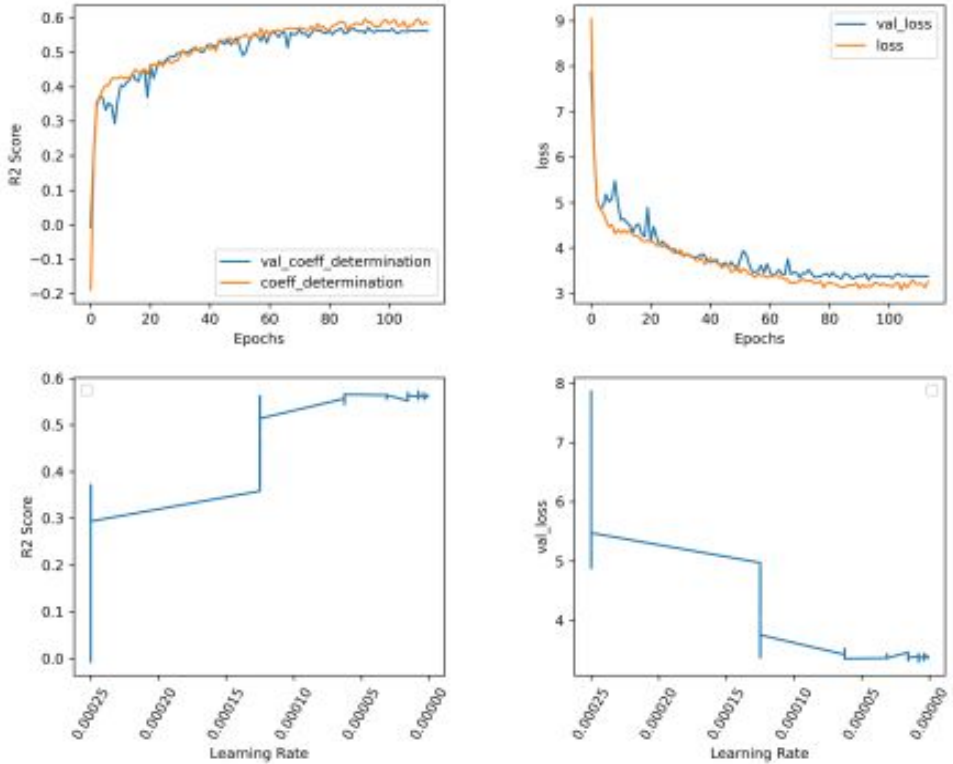
Mean absolute percentage error

$$MAPE - \frac{1}{n}\sum_{i-1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100$$

R^2

$$R^2 - 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$$

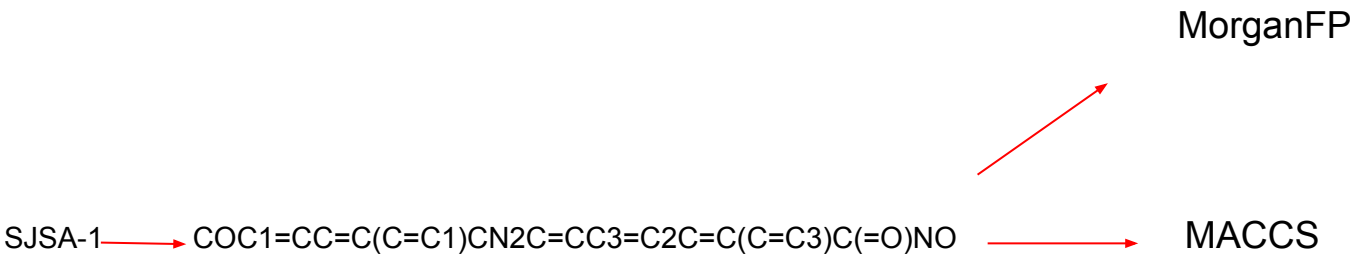
Results


R2 Score
Epochs
val_coeff_determination
coeff_determination
loss
Epochs
val_loss
loss
R2 Score
Learning Rate
val_loss
Learning Rate

Results

Model	R-Squared	RMSE	MAE	MAPE
Chemception (128)	0.58	1.81	1.43	2.27%
Chemception (64)	0.56	1.96	1.42	2.45%

Data Preprocessing
Fingerprints



Morgan vs. MACCS Fingerprints

Morgan Fingerprints

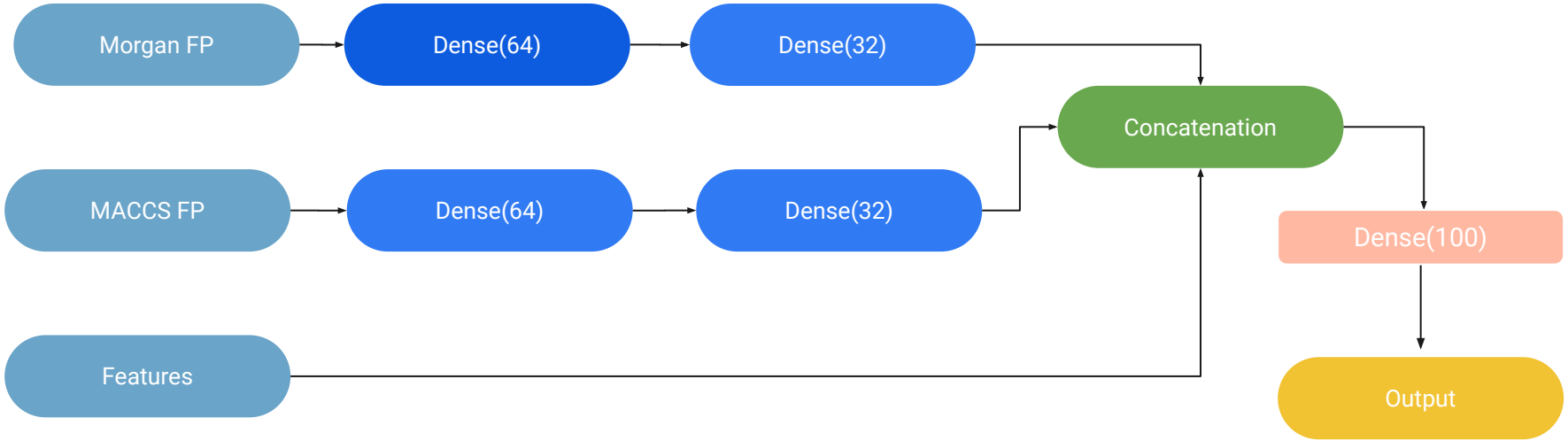
- **Type**: Circular, data-driven (e.g., ECFP).
- **Generated from**: Local atom environments (up to a defined radius).
- **Length**: Varies, commonly 1024 or 2048 bits.
- **Key Features**:
  - Captures detailed atom connectivity and environment.
  - Customizable (adjust radius and bit size).
  - Commonly used in virtual screening and machine learning.
- **Advantages**: Highly flexible and scalable for large datasets.

MACCS Fingerprints

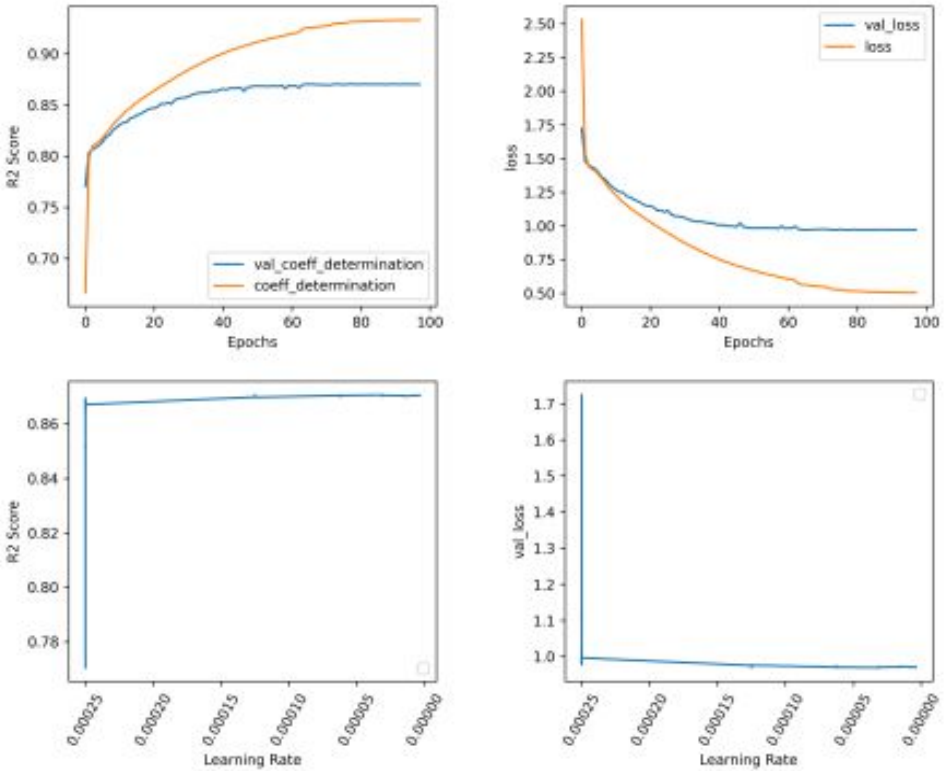
- **Type**: Predefined substructure-based.
- **Generated from**: 166 predefined chemical substructures.
- **Length**: Fixed at 166 bits.
- **Key Features**:
  - Simple, easy-to-interpret (each bit maps to a known substructure).
  - Good for substructure searching.
- **Advantages**: Fast, standardized, and intuitive.

Morgan and MACCS capture two distinct kinds of information, which reduces redundancy.

Fingerprint



Results


R2 Score
Epochs
val_coeff_determination
coeff_determination
loss
val_loss
loss
Epochs
R2 Score
Learning Rate
val_loss
Learning Rate

Results

Model	R-Squared	RMSE	MAE	MAPE
Fingerprint (128)	0.88	0.99	0.74	0.80%
Fingerprint (64)	0.87	1.01	0.78	0.81%

Significance Testing

- **T-statistics** determines whether to reject the null hypothesis, which typically states that there is no effect or no difference between groups.
- **P-value** indicates the likelihood that the observed difference between the true and predicted values could have occurred by random chance under the null hypothesis

Model	T-statistic	P-value
Chemception(128)	-18.66	0.00
Fingerprint(128)	**1.60**	**0.11**

Conclusion

1. **Superior Performance**: The Morgan and MACCS fingerprints-based model outperforms the Chemception model across all validation metrics.
2. **Variance Capture**: The fingerprint model explains a significant portion of the variance in drug reactivity, whereas the Chemception model struggles in this regard.
3. **Predictive Accuracy**: The fingerprint model demonstrates better predictive accuracy and lower error rates compared to the Chemception approach.
4. **Effectiveness of Traditional Methods**: Traditional chemical fingerprinting methods are more effective for modeling drug sensitivity in cancer cells than the more complex Chemception method.
5. **Chemception Limitations**: Chemception has difficulty capturing detailed molecular relationships, especially for molecules with polar covalent bonds.

Github for further reading-
https://github.com/swapniljha001/PrecisionMedicine/