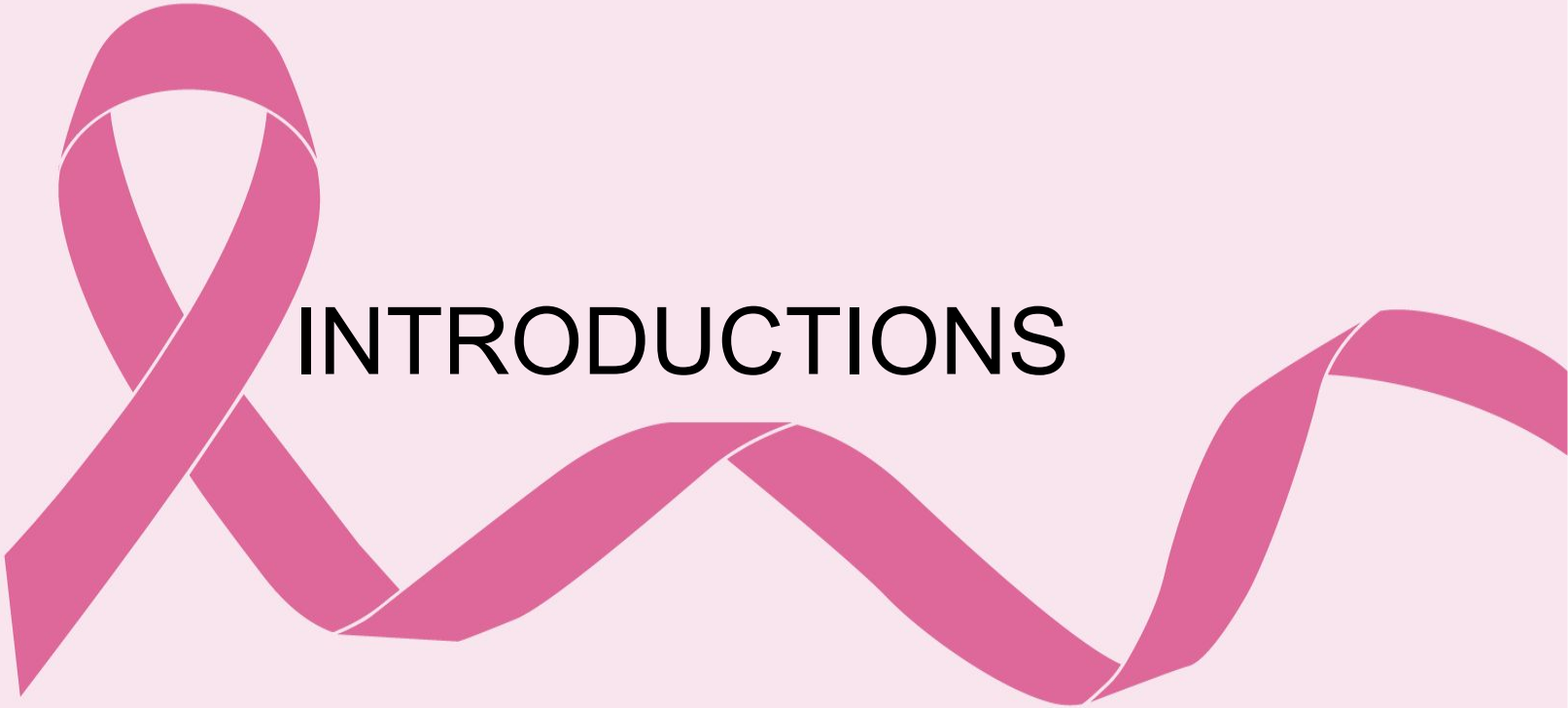
Precision Oncology

Group 1

Austin Ly, Joyce Yu, Sam McCarthy-Potter, Yanzhe Wang, Haobo Ling



INTRODUCTIONS

Intro + Background + Goal

According to the World Health Organization, breast cancer is the most commonly diagnosed cancer among women, with millions of new cases each year and hundreds of thousands of deaths.

Accurate diagnosis and therapeutic prescriptions are critical to improving patient outcomes and survival rates, yet challenges remain in ensuring timely and precise identification of the optimal medication.

Using a machine learning model we could quickly and accurately match a patient's breast cancer to the most effective known therapy potentially saving lives by reducing the time the disease could spread before the right treatments are applied.

DATA CLEANING, ACQUISITION, AND EXPLORATION

INITIAL 6 DATASETS

	Patient ID	Age at Diagnosis	Type of Breast Surgery	Cancer Type	Cancer Type Detailed	Cellularity	Chemotherapy	Pam50 + Claudin-low subtype	Cohort	ER status measured by IHC	...	Overall Survival Status	PR Status	Radio Therapy	Relapse Free Status (Months)	Relapse Free Status	Sex	3-Gene classifier subtype	Tumor Size	Tumor Stage	Patient's Vital Status
0	MB-0000	75.65	Mastectomy	Breast Cancer	Breast Invasive Ductal Carcinoma	NaN	No	claudin-low	1.0	Positive	...	Living	Negative	Yes	138.65	Not Recurred	Female	ER-/HER2-	22.0	2.0	Living
1	MB-0002	43.19	Breast Conserving	Breast Cancer	Breast Invasive Ductal Carcinoma	High	No	LumA	1.0	Positive	...	Living	Positive	Yes	83.52	Not Recurred	Female	ER+/HER2- High Prolif	10.0	1.0	Living
2	MB-0005	48.87	Mastectomy	Breast Cancer	Breast Invasive Ductal Carcinoma	High	Yes	LumB	1.0	Positive	...	Deceased	Positive	No	151.28	Recurred	Female	NaN	15.0	2.0	Died of Disease
3	MB-0006	47.68	Mastectomy	Breast Cancer	Breast Mixed Ductal and Lobular Carcinoma	Moderate	Yes	LumB	1.0	Positive	...	Living	Positive	Yes	162.76	Not Recurred	Female	NaN	25.0	2.0	Living
4	MB-0008	76.97	Mastectomy	Breast Cancer	Breast Mixed Ductal and Lobular Carcinoma	High	Yes	LumB	1.0	Positive	...	Deceased	Positive	Yes	18.55	Recurred	Female	ER+/HER2- High Prolif	40.0	2.0	Died of Disease
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
2504	MTS-T2428	70.05	NaN	Breast Cancer	Invasive Breast Carcinoma	NaN	NaN	NaN	1.0	Positive	...	NaN	NaN	NaN	4.93	Recurred	Female	NaN	27.0	1.0	NaN
2505	MTS-T2429	63.60	NaN	Breast Cancer	Invasive Breast Carcinoma	NaN	NaN	NaN	1.0	Positive	...	NaN	NaN	NaN	16.18	Recurred	Female	NaN	28.0	2.0	NaN
2506	MTS-T2430	NaN	NaN	Breast Cancer	Invasive Breast Carcinoma	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	Female	NaN	NaN	0.0	NaN
2507	MTS-T2431	NaN	NaN	Breast Cancer	Invasive Breast Carcinoma	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	Female	NaN	NaN	0.0	NaN
2508	MTS-T2432	NaN	NaN	Breast Cancer	Invasive Breast Carcinoma	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	Female	NaN	NaN	0.0	NaN

2509 rows × 34 columns

	HMS LINCS Batch ID	HMS LINCS ID	Name	Alternative Names	LINCS ID	PubChem CID	ChEBI ID	ChEMBL ID	Molecular Mass	
0	10006-101-1	10006	AZD7762	NaN	LSM-1006	67077825.0	NaN	NaN	362.12	InChI=1S/C17H19FN4O2S 11-4-1-3-10(7-11)
1	10018-101-1	10018	Neratinib	HKI-272	LSM-42778	53398697.0	NaN	180022.0	556.20	InChI=1S/C30H29ClN6O 4-39-28-16-25-23(1
2	10020-101-1	10020	Dasatinib	BMS-354825; Sprycel	LSM-1020	3062316.0	NaN	1421.0	487.16	InChI=1S/C22H26ClN7O2 14-4-3-5-16(23)2C

	Age	Race	Marital Status	Unnamed: 3	T Stage	N Stage	6th Stage	Grade	A Stage	Tumor Size	Estrogen Status	Progesterone Status	Regional Node Examined	Reginol Node Positive	Survival Months	Status
0	43	Other (American Indian/AK Native, Asian/Pacifi...	Married (including common law)	NaN	T2	N3	IIIC	Moderately differentiated; Grade II	Regional	40	Positive	Positive	19	11	1	Alive
1	47	Other (American Indian/AK Native, Asian/Pacifi...	Married (including common law)	NaN	T2	N2	IIIA	Moderately differentiated; Grade II	Regional	45	Positive	Positive	25	9	2	Alive
2	67	White	Married (including common law)	NaN	T2	N1	IIB	Poorly differentiated; Grade III	Regional	25	Positive	Positive	4	1	2	Dead
3	46	White	Divorced	NaN	T1	N1	IIA	Moderately differentiated; Grade II	Regional	19	Positive	Positive	26	1	2	Dead
4	63	White	Married (including common law)	NaN	T2	N2	IIIA	Moderately differentiated; Grade II	Regional	35	Positive	Positive	21	5	3	Dead
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
4019	52	White	Married (including common law)	NaN	T1	N1	IIA	Well differentiated; Grade I	Regional	10	Positive	Positive	19	1	107	Alive
4020	53	White	Married (including common law)	NaN	T1	N2	IIIA	Poorly differentiated; Grade III	Regional	9	Negative	Negative	13	5	107	Alive
4021	53	White	Divorced	NaN	T1	N1	IIA	Moderately differentiated; Grade II	Regional	9	Negative	Negative	4	2	107	Alive
4022	60	Other (American Indian/AK Native, Asian/Pacifi...	Married (including common law)	NaN	T1	N1	IIA	Moderately differentiated; Grade II	Regional	9	Positive	Positive	14	2	107	Alive
4023	62	White	Divorced	NaN	T1	N1	IIA	Moderately differentiated; Grade II	Regional	8	Positive	Positive	1	1	107	Alive

4024 rows × 16 columns

	Patient Information	MRI Technical Information	Demographics	Tumor Characteristics	MRI Findings	SURGERY	Radiation Therapy	Tumor Response	Recurrence	Unnamed: 9	...	Neoadjuvant Anti-Her2 Neu Therapy	Adjuvant Anti-Her2 Neu Therapy	Received Neoadjuvant Therapy or Not	Pathologic response to Neoadjuvant therapy: Pathologic stage (T) following neoadjuvant therapy	Pathologic response to Neoadjuvant therapy: Pathologic stage (N) following neoadjuvant therapy	Pathologic response to Neoadjuvant therapy: Pathologic stage (M) following neoadjuvant therapy	Overall Near-complete Response: Stricter Definition	Overall Near-complete Response: Looser Definition	Near-complete Response (Graded Measure)	Unnamed: 133
0	Breast_MRI_001	6	2	0	5	1	0	-191.8003 X -176.1259 X 86.6065	1.0	15.0	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1	Breast_MRI_002	12	0	4	1	3	0	154.724 X 176.048 X 94.5771	1.0	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2	Breast_MRI_003	10	0	3	2	3	0	174.658 X 228.317 X 88.4878	1.0	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
3	Breast_MRI_004	18	0	4	1	1	0	188.148 X 194.282 X 94.1832	1.0	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
4	Breast_MRI_005	12	2	0	5	1	1	-173.063 X -150.7869 X 59.161	1.0	5.0	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
917	Breast_MRI_918	6	0	4	1	1	0	179.537 X 165.877 X 100	4.0	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
918	Breast_MRI_919	24	0	4	1	1	0	172.595 X 192.108 X 130.246	4.0	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
919	Breast_MRI_920	21	2	0	5	1	1	-173.6078 X -147.4121 X 78.5315	3.0	5.0	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
920	Breast_MRI_921	21	0	1	1	1	0	206.292 X 221.499 X 118.632	3.0	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
921	Breast_MRI_922	19	0	1	1	1	0	187.894 X 204.024 X 118.165	3.0	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

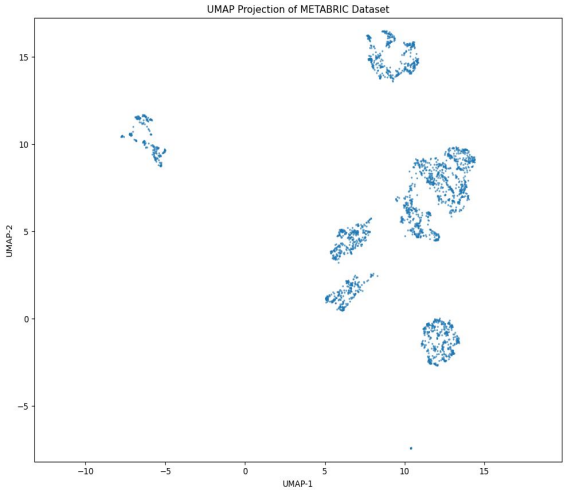
922 rows × 134 columns

	Cell HMS LINCS ID	Cell Name	Small Molecule HMS LINCS ID	Small Molecule Name	Small Mol Concentration (uM)	Primary Target	Pathway	Mean Normalized Growth Rate Inhibition Value	Increase Fracti De
0	50211-2	HCC1806	10390-103-1	Abemaciclib	0.001000	CDK4/6	Cell cycle	0.9779	0.003
1	50211-2	HCC1806	10390-103-1	Abemaciclib	0.003162	CDK4/6	Cell cycle	0.9667	-0.002
2	50211-2	HCC1806	10390-103-1	Abemaciclib	0.010000	CDK4/6	Cell cycle	0.9168	0.00434
3	50211-2	HCC1806	10390-103-1	Abemaciclib	0.031623	CDK4/6	Cell cycle	0.7658	0.01650
4	50211-2	HCC1806	10390-103-1	Abemaciclib	0.100000	CDK4/6	Cell cycle	0.7132	0.01025
...	...	...	...	...	...	...	...	...	...
10705	51083-2	SUM159PT	10194-106-1	Cabozantinib	0.100000	VEGFR2/MET	RTK	1.0115	-0.00152
10706	51083-2	SUM159PT	10194-106-1	Cabozantinib	0.316230	VEGFR2/MET	RTK	0.9965	0.00568
10707	51083-2	SUM159PT	10194-106-1	Cabozantinib	1.000000	VEGFR2/MET	RTK	0.9307	0.03130
10708	51083-2	SUM159PT	10194-106-1	Cabozantinib	3.162300	VEGFR2/MET	RTK	0.7480	0.08743
10709	51083-2	SUM159PT	10194-106-1	Cabozantinib	10.000000	VEGFR2/MET	RTK	0.1744	0.19675

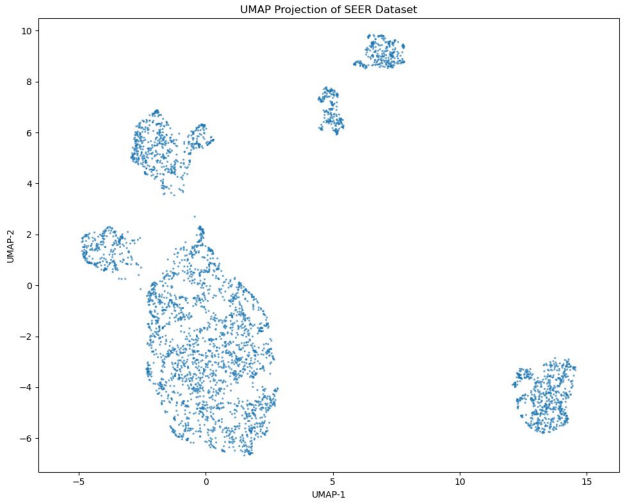
10710 rows × 9 columns

	HMS LINCS Batch ID	HMS LINCS ID	Name	Alternative Names	LINCS ID	Alternative ID	Reference Sc
0	50008-2	50008	CAL-51	NaN	LCL-1472	<a href = "http://purl.obolibrary.org/obo/CLO_...	<a h "https://www.dsmz.de/catalogues/d
1	50029-2	50029	MCF7	NaN	LCL-1460	<a href = "http://purl.obolibrary.org/obo/CLO_...	<a h "http://www.atcc.org/Products/All/

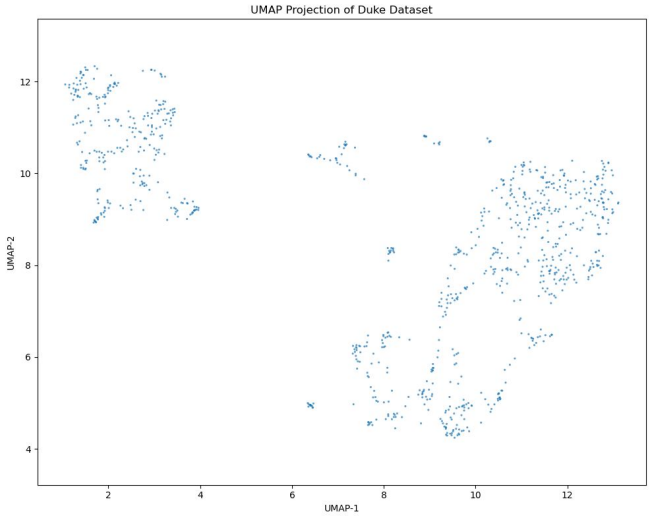
Visualizations of 3 raw datasets



METABRIC



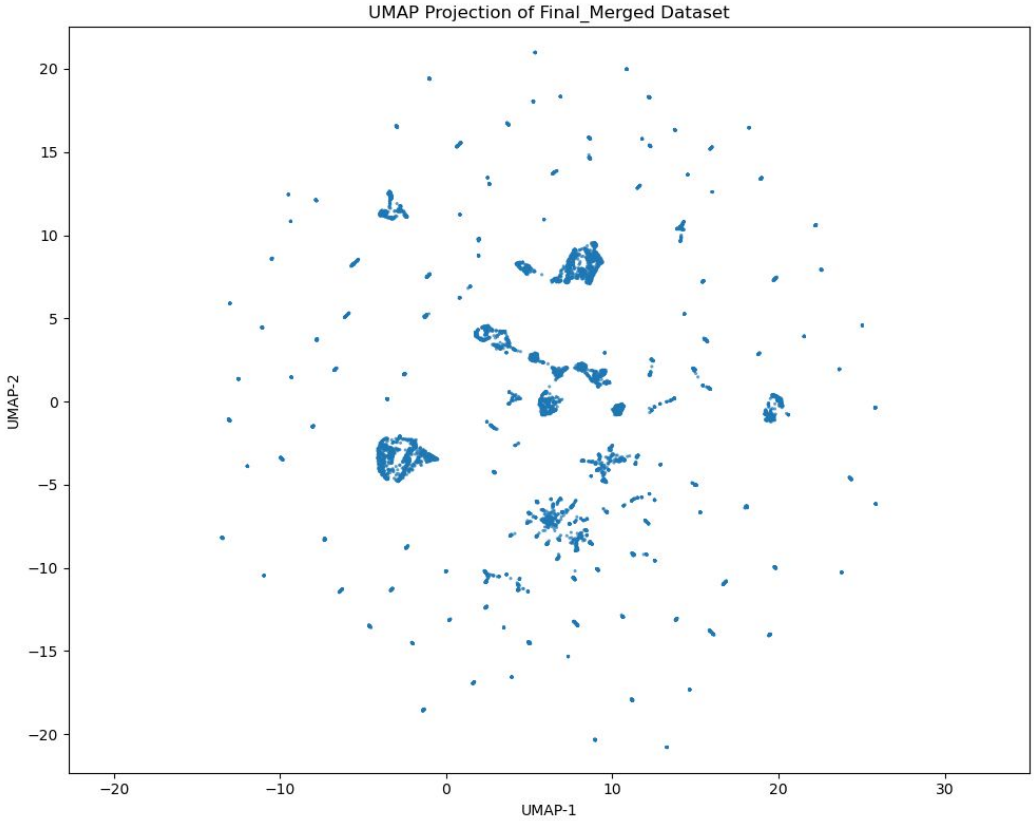
SEER



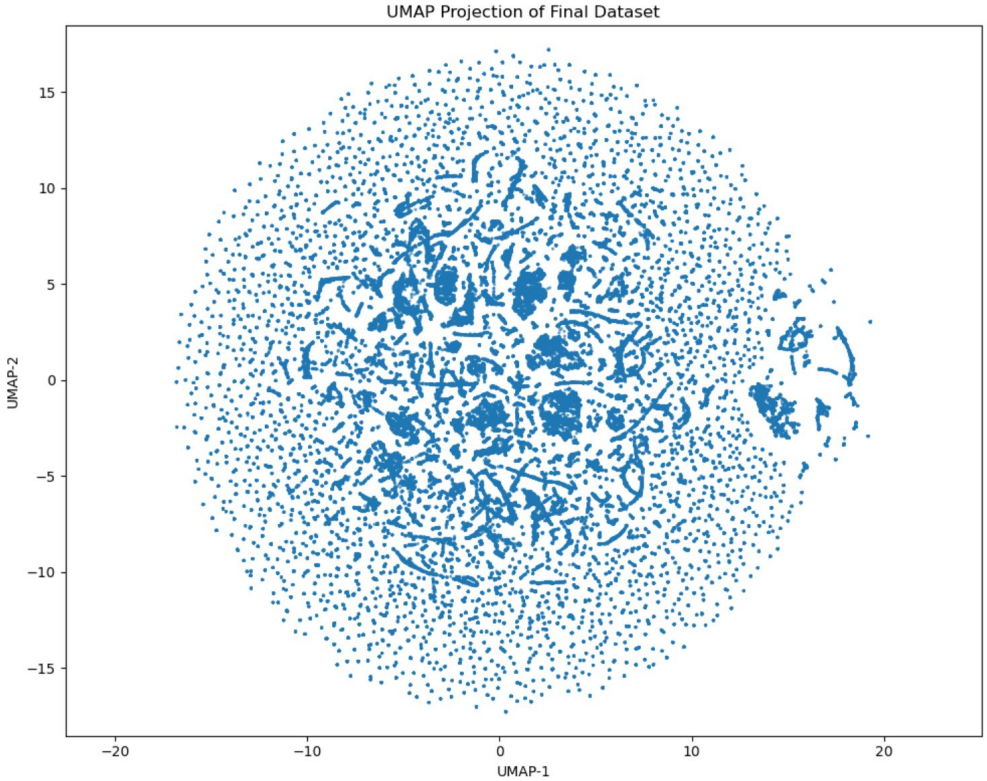
DUKE

Why these UMAP results all contain highly concentrated clusters?

Visualizations of merged and imputed patient demographics



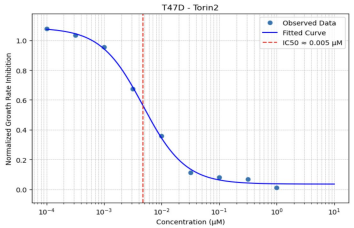
Visualizations of patient demographic x molecular information

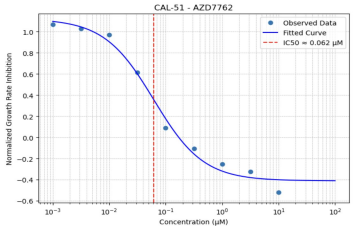


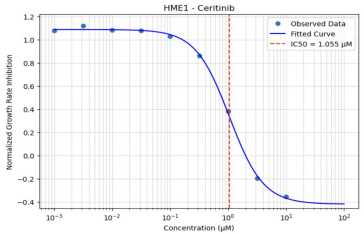
After adding chemical information, the UMAP result changes a lot …

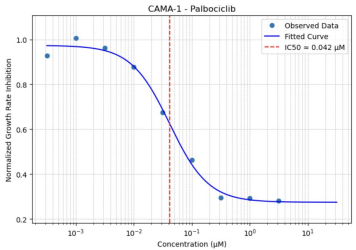
Drug Characterization - IC50 Determinations

Examples of **Valid** IC50 Curves

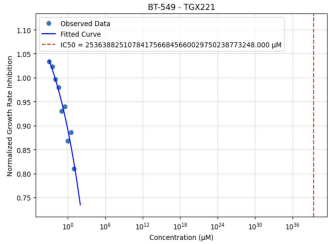


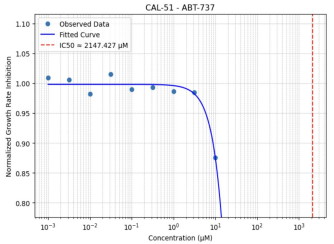


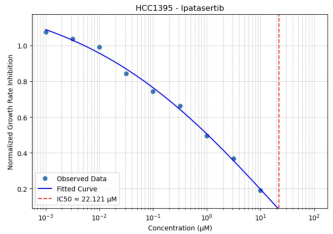


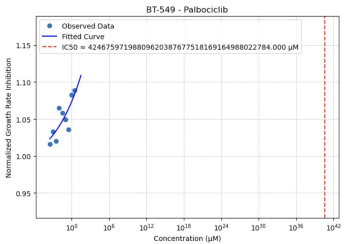


Examples of **Invalid** IC50 Curves









Equation for Mean Normalized Growth Rate (GR) Inhibition = 2^[log2(x(c)/x0)/log2(xctrl/x0)]-1

Equations for IC50 Determination:

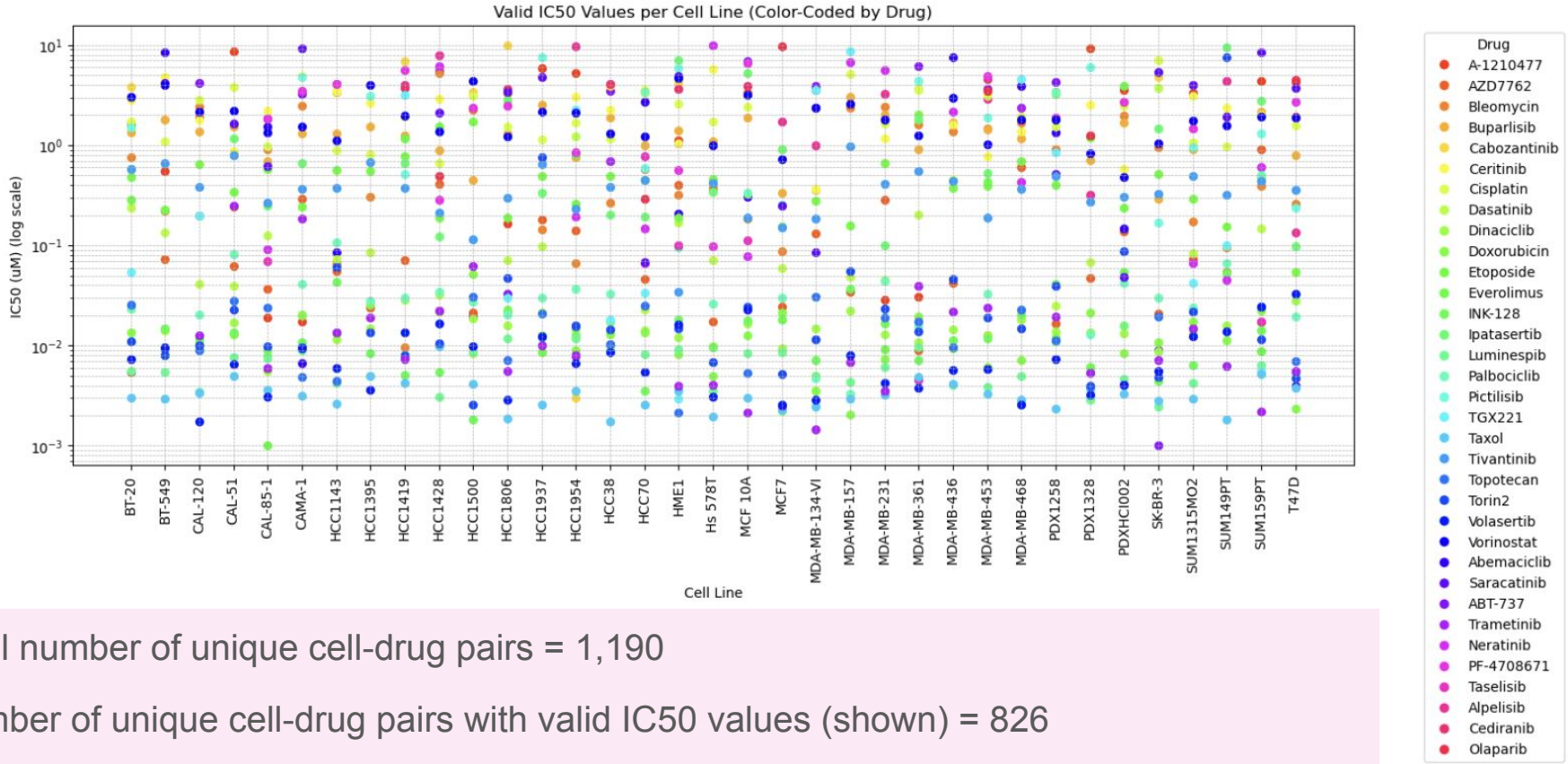
- Y = Bottom + (Top - Bottom) / (1 + 10^((LogEC50 - X) * HillSlope) → Isolate EC50 term
- 50% Inhibition activity =(Top+Baseline)/2

X-axis: Concentration of drug in units of log(uM)

- IC50 (or EC50) is the most widely used measure for drug efficacy (SOURCE: 1 , 2).
- IC50 is the concentration that provokes an inhibitory response half way between the maximal (Top) response and the maximally inhibited (Bottom) response (SOURCE 1).
- The lower the IC50, the more potent / effective the drug is.

Y-Axis: Activity quantified as Normalized Growth Rate Inhibition (SOURCE: 3)

Drug Characterization - IC50 Determinations

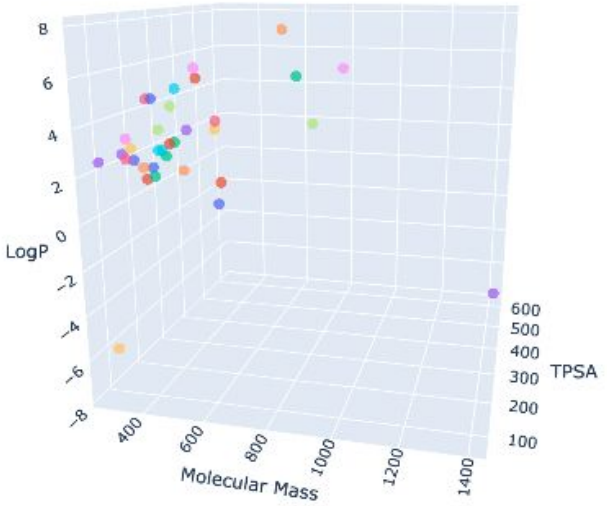


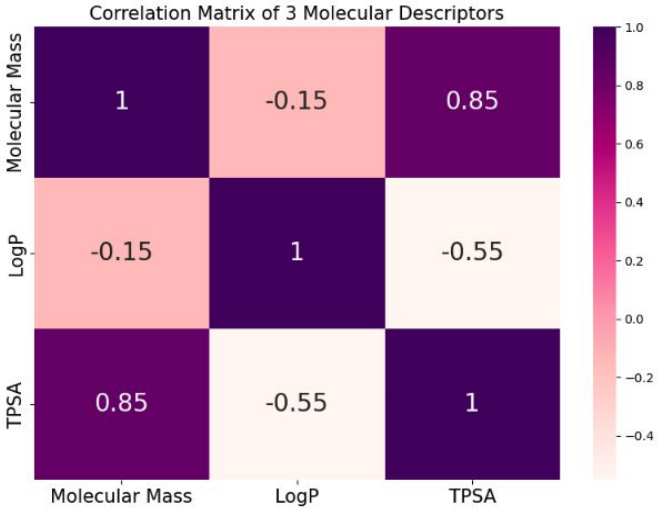
Total number of unique cell-drug pairs = 1,190

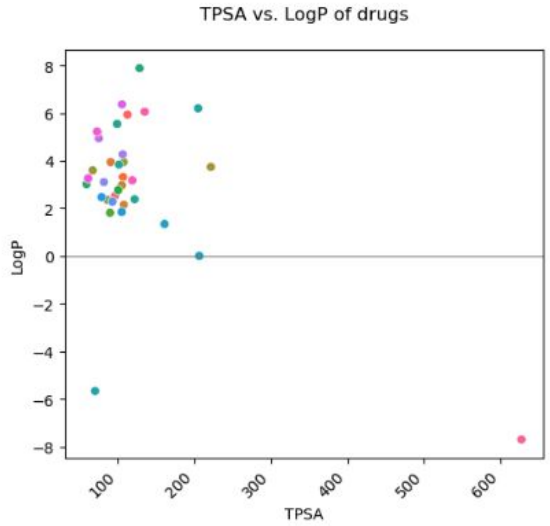
Number of unique cell-drug pairs with valid IC50 values (shown) = 826

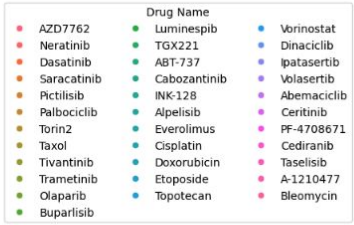
Number of extremely potent drugs (IC50 < 0.001nM) = 50

Drug Characterization - LogP and TPSA





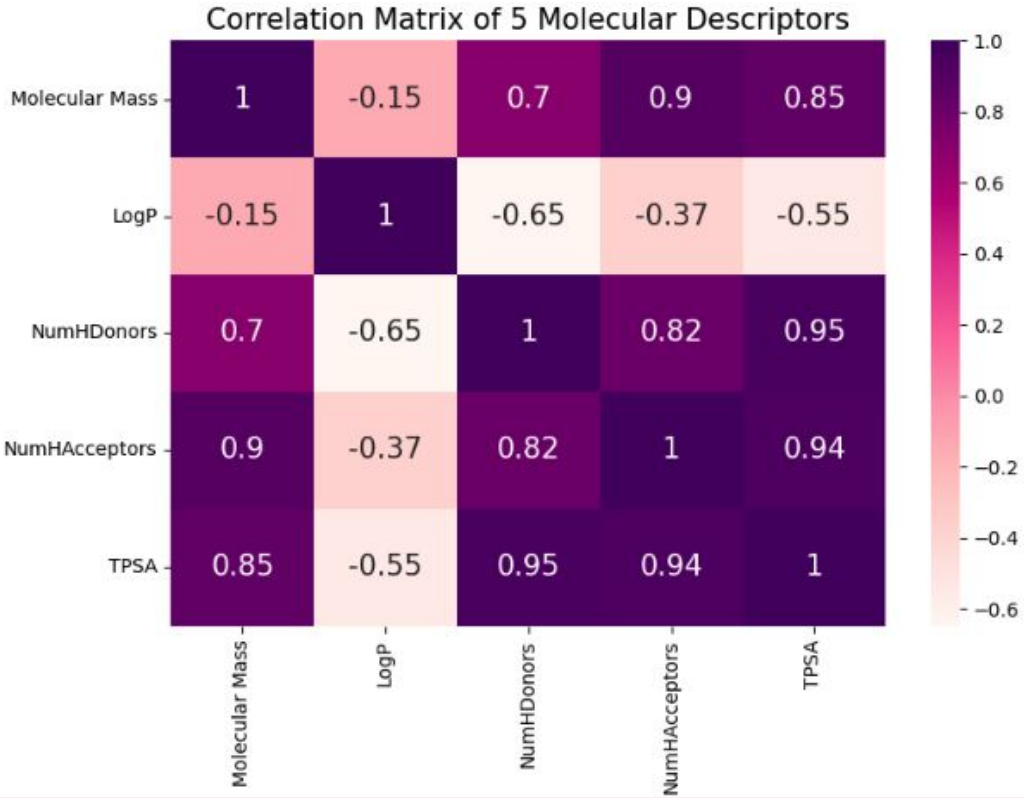




Good measures of In-Vivo Drug Response: **LogP and TPSA**

- LogP → Good measure of drug-likeness, permeability, and solubility [SOURCE: 4]
- TPSA → Good measure for pharmacodynamics/kinetics (i.e. drug metabolism and clearance in the body)
    - Drugs with lower TPSA values tend to be more extensively metabolised since they are more lipid soluble and thus are more likely to be reabsorbed extensively in the kidneys. Once TPSA is known for a drug, the value can be used to predict the extent to which an oral dose of the drug will be absorbed from the GI tract into the portal circulation, or the extent to which the drug will partition into the brain from the plasma. [SOURCE: 5]

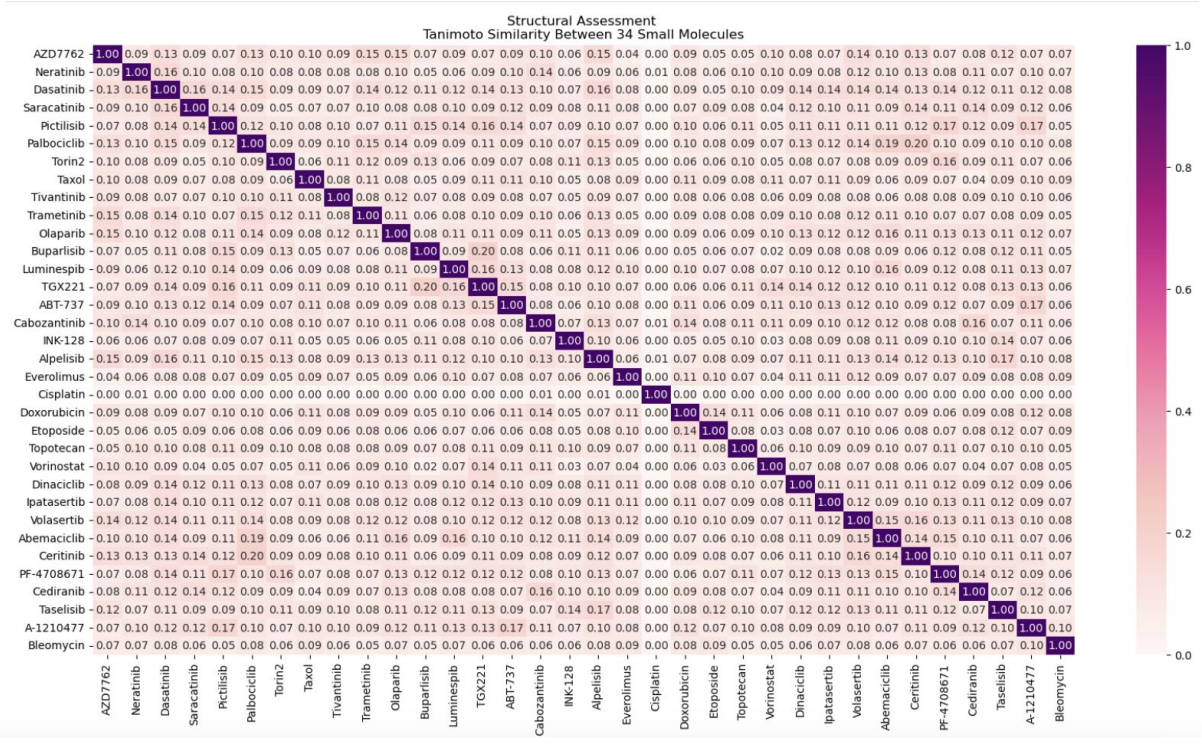
Drug Characterization - Other Descriptors

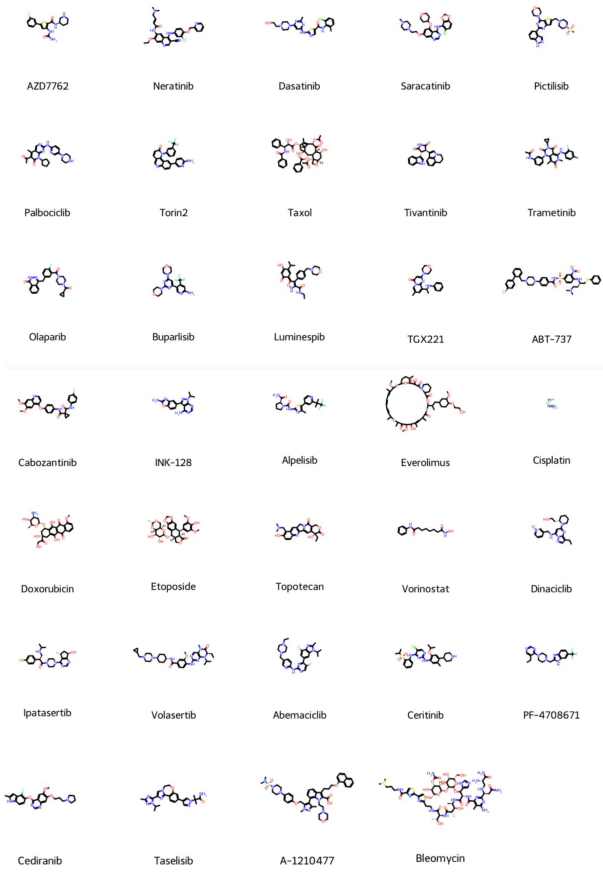


	Name	Molecular Mass	LogP	NumHDonors	NumHAcceptors	TPSA
0	AZD7762	362.12	2.52660	4	4	96.25
1	Neratinib	556.20	5.93248	2	8	112.40
2	Dasatinib	487.16	3.31354	3	9	106.51
3	Saracatinib	541.21	3.93950	1	10	90.44
4	Pictilisib	513.16	2.14840	1	9	107.55
5	Palbociclib	447.24	2.96582	2	9	105.04
6	Torin2	432.12	5.20190	1	5	73.80
7	Taxol	853.33	3.73570	4	14	221.29
8	Tivantinib	369.15	3.59260	2	3	66.89
9	Trametinib	615.08	3.94012	2	8	107.13
10	Olaparib	434.18	2.34740	1	4	86.37
11	Buparlisib	410.17	1.81280	1	8	89.63
12	Luminespib	465.23	2.76190	3	7	100.13
13	TGX221	364.19	3.01262	1	6	58.87
14	ABT-737	812.26	7.88060	2	10	128.13
15	Cabozantinib	501.17	5.54080	2	6	98.78
16	INK-128	309.13	2.37980	2	8	121.67
17	Alpelisib	441.14	3.83502	2	5	101.21
18	Everolimus	957.58	6.19720	3	14	204.66
19	Cisplatin	298.96	-5.67050	2	2	70.00
20	Doxorubicin	543.17	0.00130	6	12	206.07
21	Etoposide	588.18	1.33860	3	13	160.83
22	Topotecan	421.16	1.84680	2	8	104.89
23	Vorinostat	264.15	2.47110	3	3	78.43
24	Dinaciclib	396.23	2.27850	2	7	92.63
25	Ipatasertib	457.22	3.10100	2	6	81.59
26	Volasertib	618.40	4.26720	2	9	106.17
27	Abemaciclib	506.27	4.93692	1	8	75.00
28	Ceritinib	557.22	6.36192	3	8	105.24
29	PF-4708671	390.18	3.25630	1	5	60.94
30	Cediranib	450.21	5.22422	1	6	72.50
31	Taselisib	460.23	3.17422	1	9	118.67
32	A-1210477	849.39	6.05282	1	11	134.84
33	Bleomycin	1414.52	-7.70358	20	31	627.07

* Drug features used for training. Note that TPSA, NumHDonors, and NumHAcceptors are highly correlated.

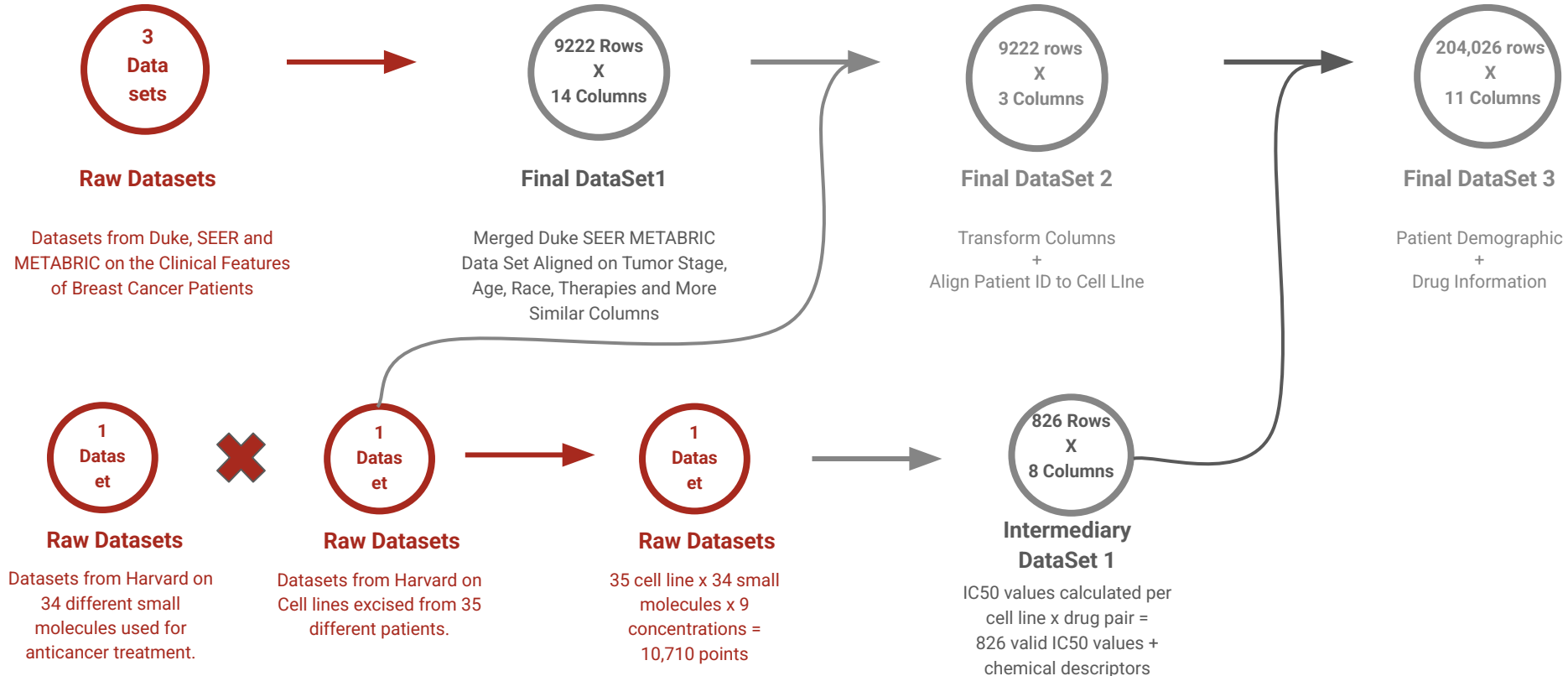
Drug Characterization - Structural diversity



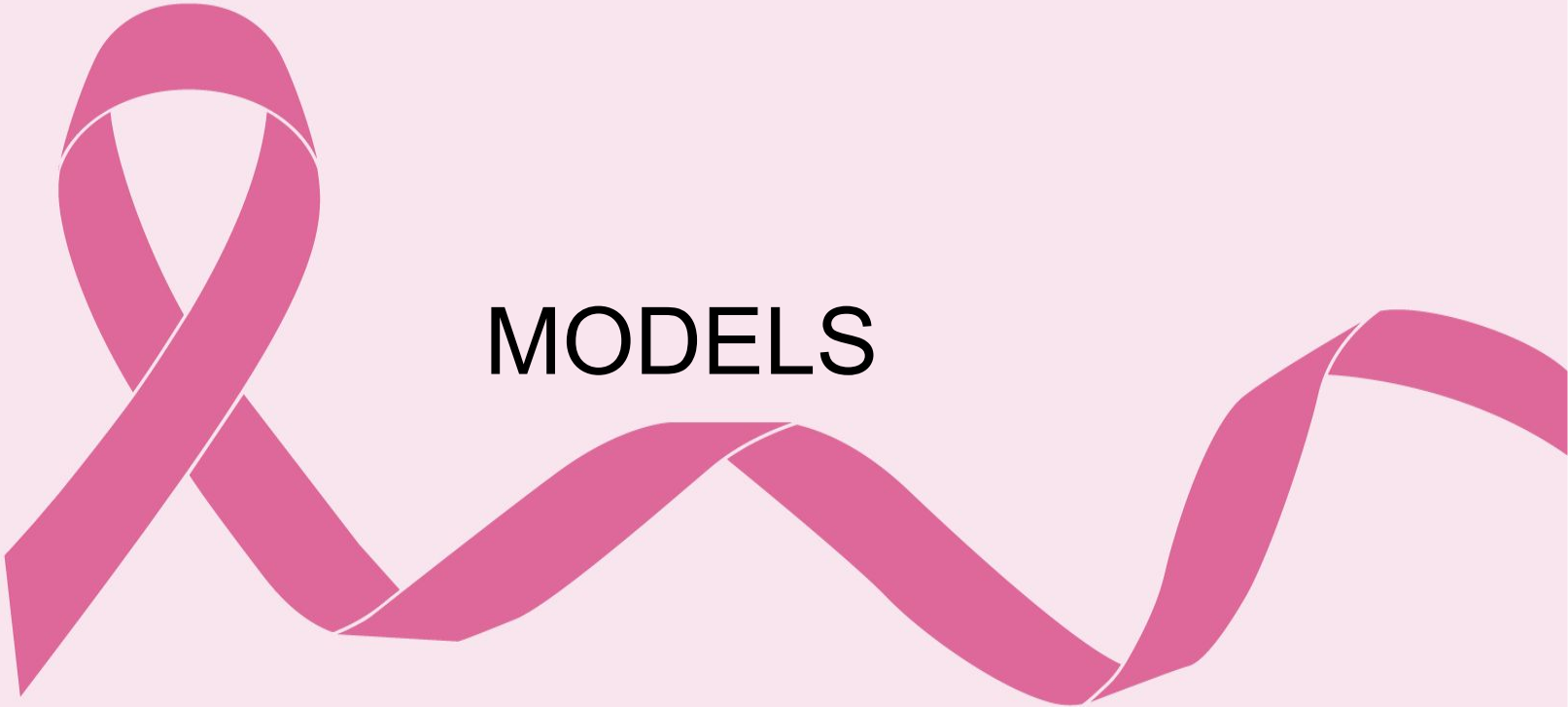


*Given **IC50 values, descriptors, and chemical structures**; we have a relatively good set of **effective, safe, and diverse** chemotherapies to recommend to patients with breast cancer.

Final datasets used for feature extraction & model training



MODELS



ML/DL Models

1. Racial Imputations/Predictions
    a. Gaussian Naive Bayes
    b. Summaries/Findings: Indications of correlated features, over representation of racial feature.
2. Translational Imputations/Predictions
    a. Transform Numerical and Categorical Features
    b. NearestNeighbors Model for Patient ID and Cell Line Alignment
3. Patient Drug response (IC50) predictions
    a. Simple feed forward neural network
    b. Summaries/Findings: Quantify the predictive power of translational research
4. Treatment to survival prediction
    a. Binary classification neural net
    b. Uses common features of merged dataset to predict Five Year survival rate

Model 1 - Predicting the Race Feature

Dataset Used:

- Combined full dataset
- Shape: (9222 rows, 140 features)

Why Predict Race?

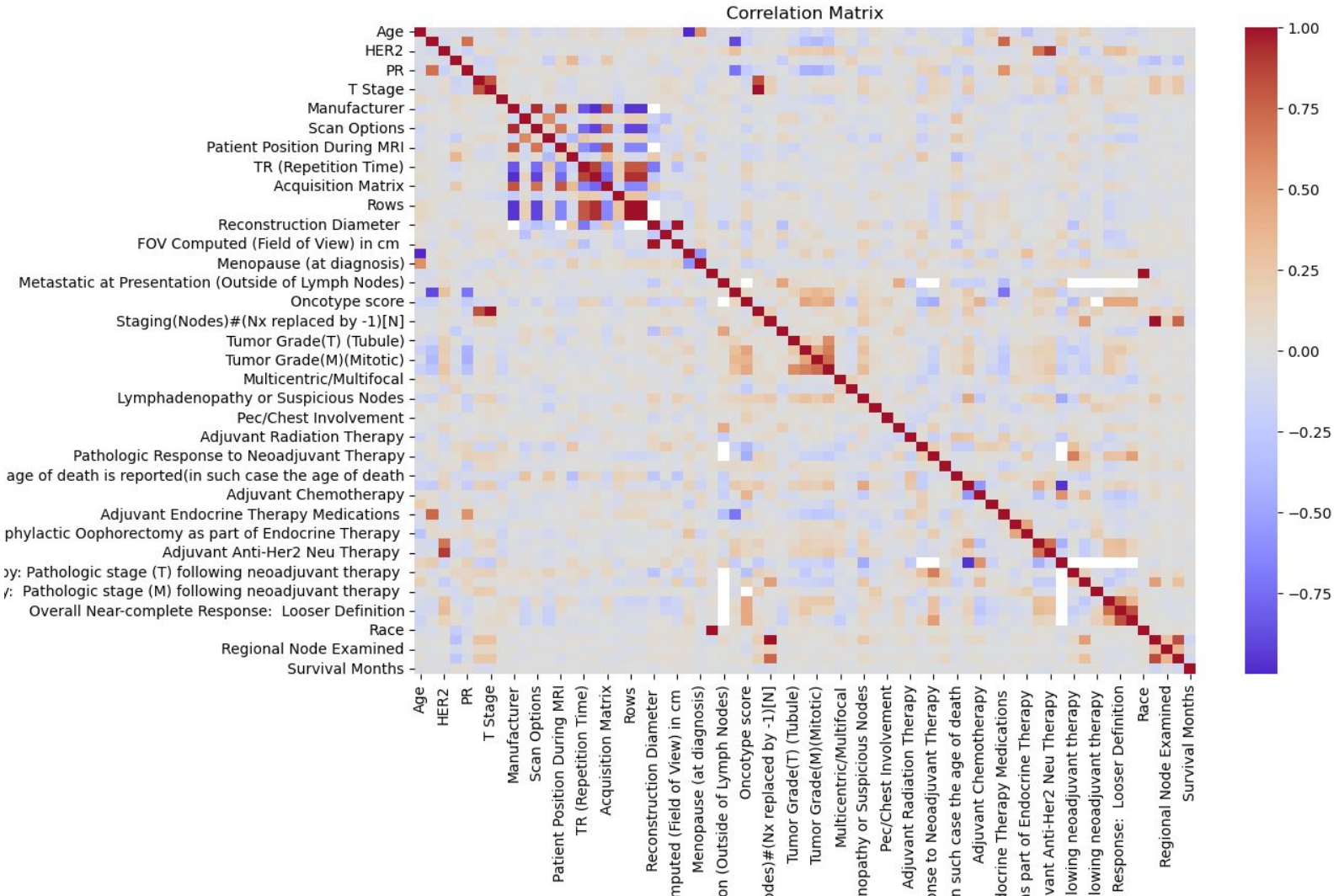
- Identifying race-correlated patterns highlights systemic bias and disparities in healthcare access, treatment, and outcomes.
- Helps understand which clinical, genomic, and imaging features may be unevenly distributed across racial groups.

Societal Impact

- Detecting bias: Predicting race can help surface inequalities in medical systems.
- Improving fairness: Understand features driving unintended differences in diagnosis, treatment, and survival.
- Caution: Race prediction must be handled ethically, recognizing race as a social and biological intersection, not a purely biological label.

Model 1 - Exploring the Data

- Data Preprocessing
  - Drop Rows
  - Drop Features where >80% of values are missing.
- Normalize Features and PCA
- Class Imbalance
  - Heavily skewed toward one majority group (Caucasian)
- Early Observations suggest differences driven by a mix of biological, clinical and societal factors.


Correlation Matrix
Age
HER2
PR
T Stage
Manufacturer
Scan Options
Patient Position During MRI
TR (Repetition Time)
Acquisition Matrix
Rows
Reconstruction Diameter
FOV Computed (Field of View) in cm
Menopause (at diagnosis)
Metastatic at Presentation (Outside of Lymph Nodes)
Oncotype score
Staging(Nodes)#(Nx replaced by -1)[N]
Tumor Grade(T) (Tubule)
Tumor Grade(M)(Mitotic)
Multicentric/Multifocal
Lymphadenopathy or Suspicious Nodes
Pec/Chest Involvement
Adjuvant Radiation Therapy
Pathologic Response to Neoadjuvant Therapy
age of death is reported(in such case the age of death
Adjuvant Chemotherapy
Adjuvant Endocrine Therapy Medications
phylactic Oophorectomy as part of Endocrine Therapy
Adjuvant Anti-Her2 Neu Therapy
oy: Pathologic stage (T) following neoadjuvant therapy
y: Pathologic stage (M) following neoadjuvant therapy
Overall Near-complete Response: Looser Definition
Race
Regional Node Examined
Survival Months
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75

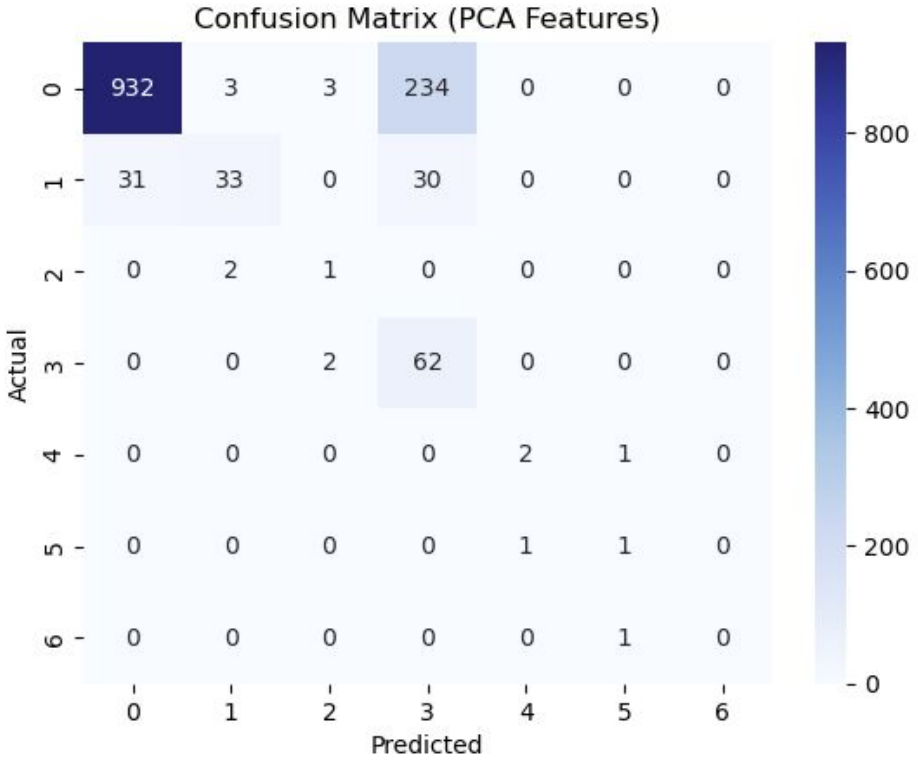
Model 1 - Gaussian NB & Results

- Why Gaussian NB?
  - Simple and interpretable baseline model
  - Assumes features are independent and Gaussian-distributed — aligns well after scaling
- PCA improved recall for minority groups
  - Some loss in precision compared to full feature set
- Results
  - .77 accuracy
  - Models predicted race with high accuracy for majority class
- Key takeaways
  - Race-correlated patterns exist in the combined dataset
  - Gaussian Naive Bayes served as a lightweight, explainable baseline
  - Societal implications and outside features influences this model (Low representation of minority groups)

Classification Report:

	precision	recall	f1-score	support
1.0	0.97	0.80	0.87	1172
2.0	0.87	0.35	0.50	94
3.0	0.17	0.33	0.22	3
4.0	0.19	0.97	0.32	64
5.0	0.67	0.67	0.67	3
6.0	0.33	0.50	0.40	2
8.0	0.00	0.00	0.00	1
accuracy			0.77	1339
macro avg	0.46	0.52	0.43	1339
weighted avg	0.92	0.77	0.82	1339

Model 1 - Confusion Matrix



Model 2 - Align Patient ID to Cell Line

- Dataset Used:
  - Combined full dataset (Shape: 9222 rows, 140 features)
  - Cell line dataset (Shape: 34 rows, 39 features)
- Steps:
  - Transform numerical and categorical features
  - Align Patient ID to Cell Line with NearestNeighbors
- Why
  - Given the existing dataset of different small molecules used for anticancer treatment and the aim of providing individual treatment, the alignment between patient ID and cell line contribute to out model 3 of IC50 prediction.

Model 2 - Align Patient ID to Cell Line

- Transform Column
  - For numerical features - StandardScaler
  - For categorical features - OneHotEncoder
  - Preprocessor - ColumnTransformer
- NearestNeighbors Model for Mapping
  - Process cell lines and patients ID with preprocessor
  - Fit NearestNerighbors to the cell lines
  - Calculate distances and indices
  - Flatten the outcome for the nearest
  - Similarity distance equals to nearest distance

Figure of Outcome Values

	Patient ID	Matched Cell Line ID	Similarity Distance
1	MB-0156	50056-1	0.6251790387932401
2	MB-0472	50056-1	0.5358677475370629
3	MTS-T0243	50029-2	1.4170308771321272
4	MB-0539	50056-1	0.9824242038179487
5	MB-0159	50056-1	0.1786225825123543
6	MB-0299	50056-1	1.3396693688426573
7	MB-0230	50583-6	0.4465564562808857
8	MB-0592	50105-2	1.4142135623730951
9	MB-0573	50105-2	1.4254494122849055
10	MB-0110	50056-1	0.7144903300494172

Model 3 - Simple feed forward Neural Network

Use Case:

- Predict patient response (IC50) to a particular (set) of potential anti-cancer drugs.

Assumptions:

- Cell lines are a proxy for the patient (translational medicine approach)

DataSet: ~204,600 rows X 11 columns

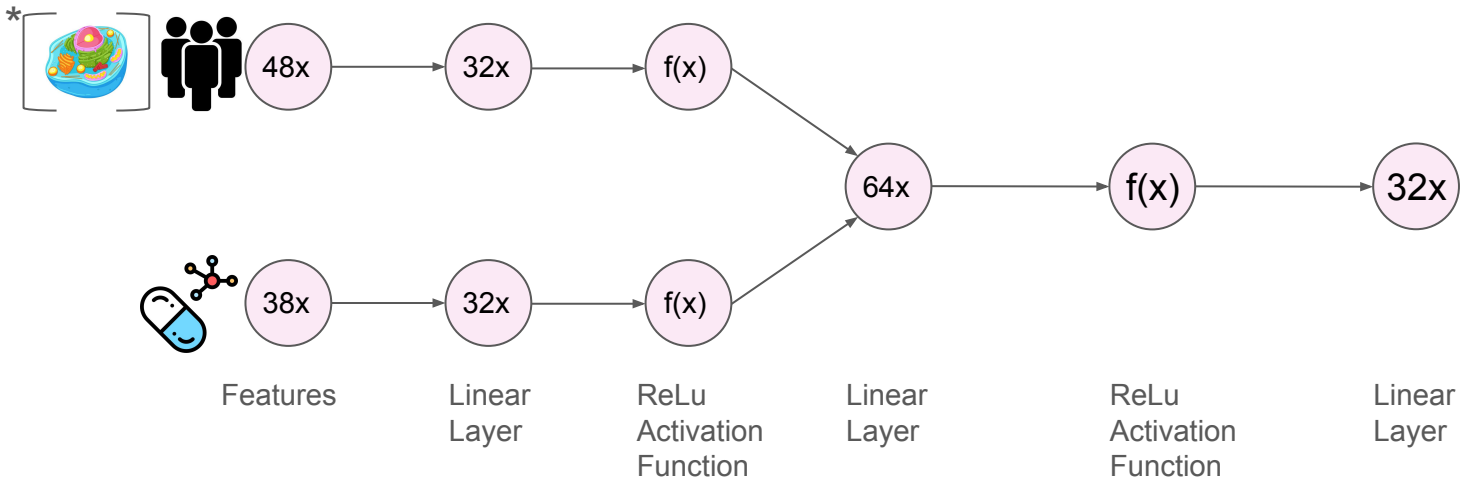
Features: Age, Race, T-stage, *Cell Line, Small Molecule, Molecular Mass, LogP, TPSA, NumHDonors, NumHAcceptors,

Target feature: IC50

* Note: Model was trained with and without ‘Cell Line’ feature for comparison.

Model 3 - Simple feed forward Neural Network

Architecture:



Epochs: 300

Activation Functions: ReLu

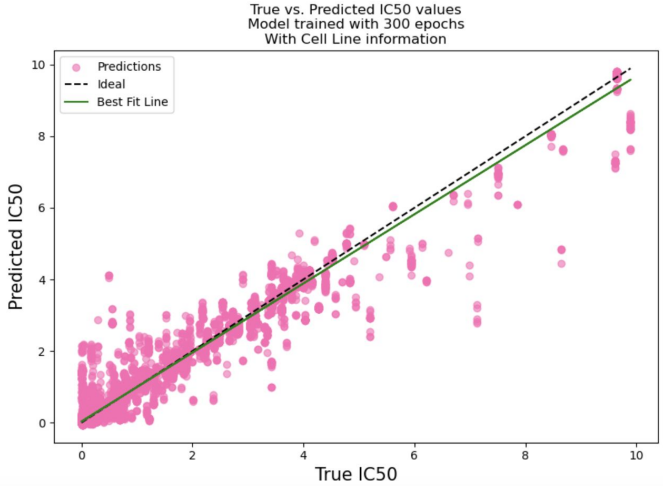
Optimizer: Adam Optimizer + learning rate of 0.001 + MSE

Final Evaluation/Loss Metric: MSE, MAE, and R^2

* Note: Model was trained with and without ‘Cell Line’ feature for comparison.

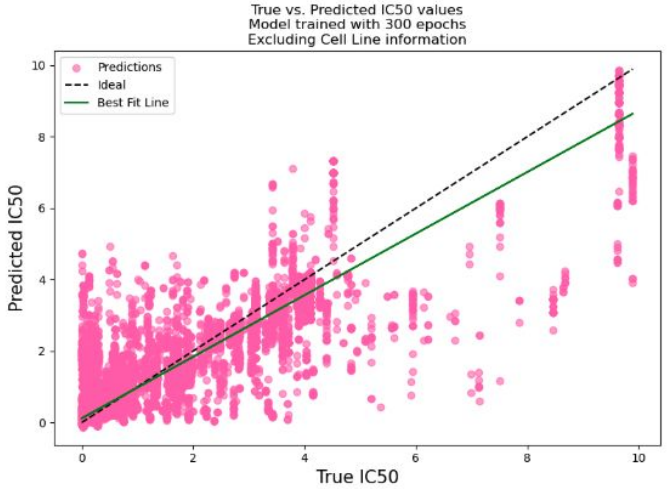
Model 3 - Simple feed forward Neural Network

With Cell Line Feature



MSE: 0.0519
MAE: 0.0830
R^2: 0.9765

Without Cell Line Feature



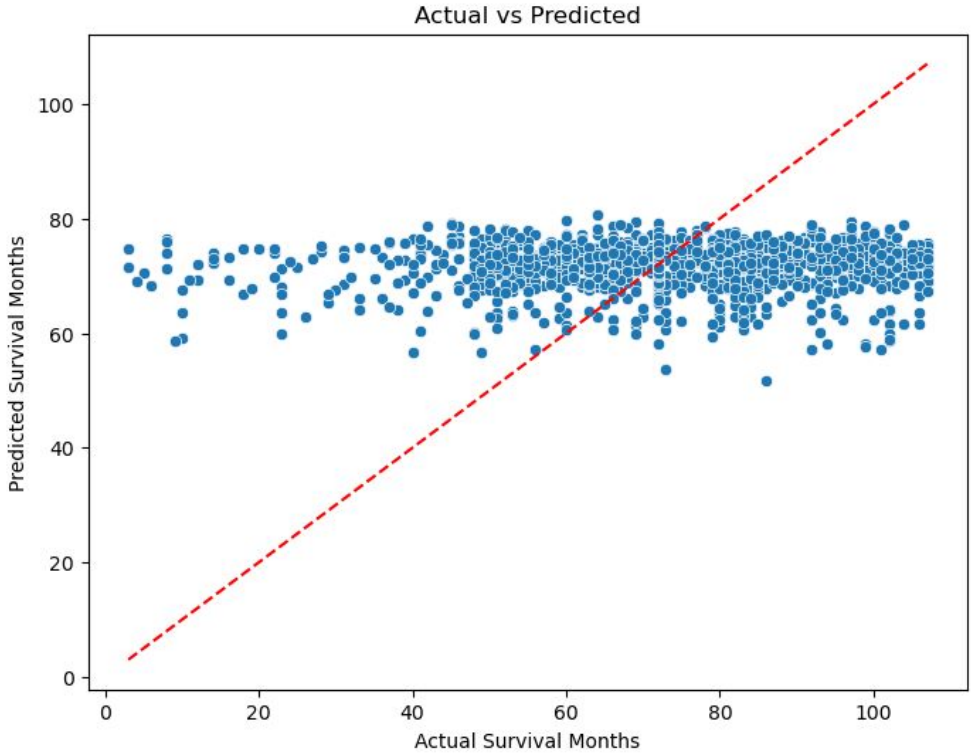
MSE: 0.2989
MAE: 0.2416
R^2: 0.8649

* Takeaway: Predictive power of translational research in precision oncology. Much of translational research applies scientific discoveries from bench to bedside by using in-vitro and (non-human) in-vivo models to proxy human patients. Although not perfect, this data shows helps quantify the predictive power of translational research when recommending new drugs to patients and can help narrow down drug candidates to speed delivery of cancer therapeutics onto market.

Data cleaning for Survival rate

Initially all 76 numeric columns were used to train a linear regression model.

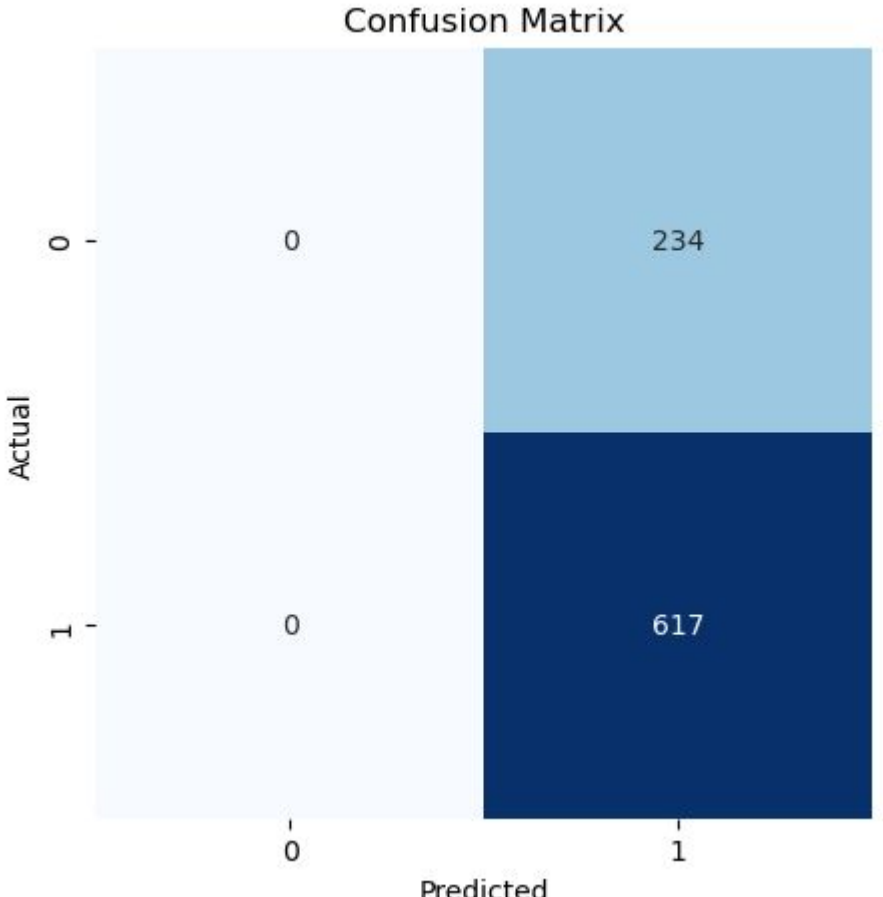
Similar columns from different data sets such as Radiotherapy and adjuvant Radiotherapy were reformation to be the same data type and merged.



First attempted at Binary classification

Due to the failure to get good accuracy with a linear regression model we instead tried for a binary classification model based on the 5 year survival rate of the patient.

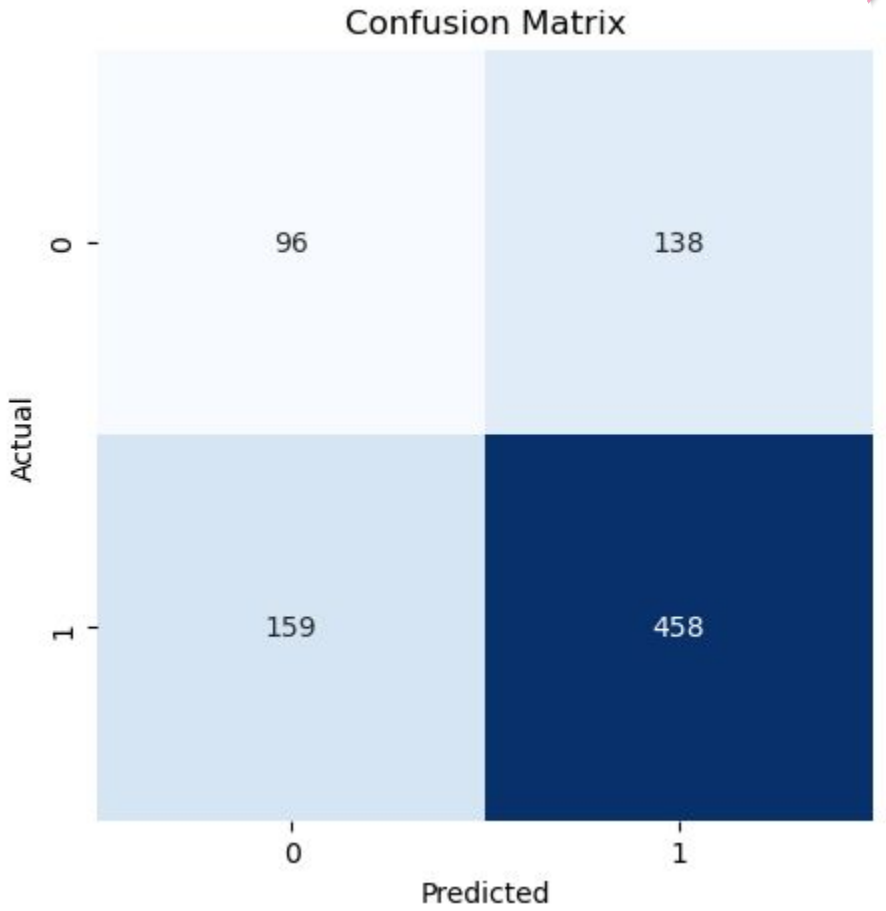
Due to the uneven distribution of classes the initial model defaulted to only predicting patients having a greater then 5 year life expectancy



Rebalancing Class weights

To fix the uneven distribution of classes the model was trained using the complete training dataset but the weight of each class was rescaled to be inversely proportional to how prevalent it was in the data set.

```
compute_class_weight(class_weight='balanced')
```



LIMITATIONS

Limitations and future improvements

- Data acquisition
  - Many authors are not willing to share their dataset.
  - Much of easily accessible data revolves around tumor size and image analysis.
- Messy datasets
  - Need to elaborate on imputations and merging
  - Skewness: Race and T-stage
- Lack of feature information on patients and donors (HIPAA)
  - -omics data of patients and cell lines
  - Family history
  - Culture medium vs. human diet
  - Other environmental factors
  - Other factors influencing drug safety
  - Most cell lines excised at T-stage 4
  - Most patients are at T-stage 1 to 3
- Small set of (35) cell lines and (34) drugs
  - Acquiring this data in a wet-lab is a lot of effort.
  - Not fully representative of the human population and existing drugs in practice.
- We are not doctors!