Part 1

Project: Explainable AI-Driven Adverse Drug Reactions Prediction Toward Pediatric Drug Discovery & Development

All source:https://drive.google.com/drive/u/0/folders/1tgMaF3fOFp_wqFd7rTWBiovmlc-8i08_

Objective

To develop a deep learning model that classifies compounds based on their anticancer activity using integrated chemical and biological data. The project involves building **binary classifiers** using real-world datasets of drug compounds, gene expression, and cell line responses, with SMILES strings and molecular fingerprints representing chemical structures, and expression profiles capturing cellular context. The aim is to explore multi-modal deep learning approaches for predicting drug sensitivity, contributing to precision oncology and pediatric drug discovery.

Part 1 :Data Integration & Preparation

Data Loading and Cleaning:

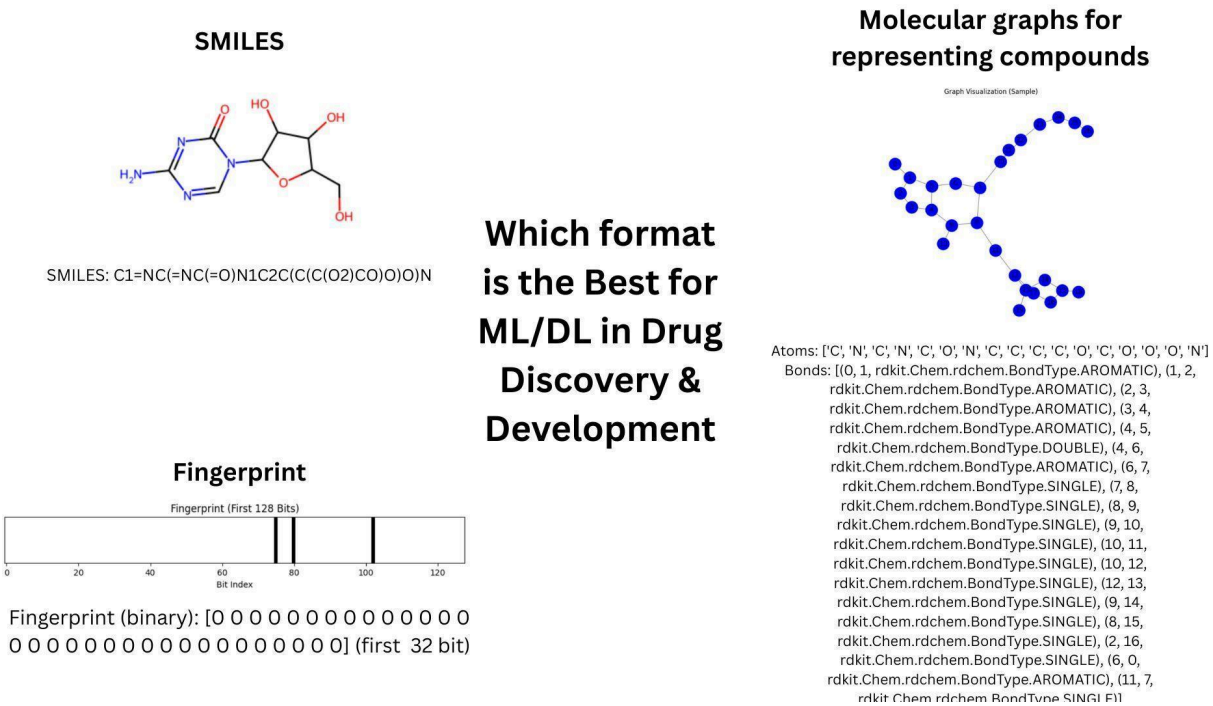
After I finish running **01_data_prep.ipynb** and **02_gen_dataset.ipynb** locally and I upload the dataset on my google drive for mount google drive and using GPU google collab (faster for model training) then I cleaned the dataset on collab by **then I cleaned the dataset on Google Colab by performing comprehensive data preprocessing steps including:**

1. **Loading the datasets** – Imported gene expression features and drug sensitivity training data from **Google Drive.**
2. **Cleaning gene expression data** – Removed columns with excessive missing values (more than 30%), and **imputed remaining missing values using column-wise means.**
3. **Feature standardization** – **Scaled gene expression features** to standardized units (e.g., z-scores) to ensure fair contribution from all genes, while maintaining the integrity of cell line mappings (e.g., Gene A: 15,000 → 0.5, Gene B: 2.1 → 0.3).
4. **Filtering sensitivity data** – **Removed invalid IC50 values (missing or non-positive)** and created binary sensitivity labels based on a log(IC50) threshold of 1.5.
5. **Data integration** – **Merged gene expression and sensitivity** datasets using cell line identifiers to form a unified training dataset.
6. **Quality validation** – **Verified final dataset** dimensions, confirmed no remaining missing values, and checked distributions of cancer types and sensitivity classes.

The cleaning process resulted in:

- **Final dataset**: 92,563 samples with 740 features (735 gene expression features + 5 metadata columns)
- **Balanced representation** across 23 cancer types, with Lung Cancer (18,957), Breast Cancer (6,367), and Leukemia (6,240) being most prevalent
- **Zero missing values** after preprocessing

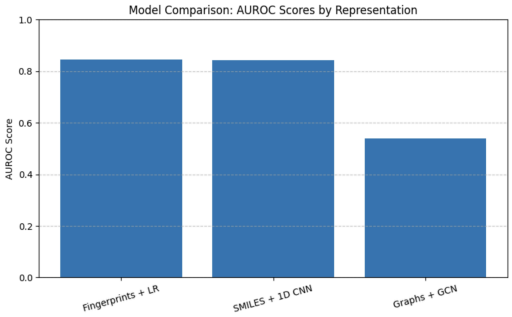
Representation of Chemical and Gene Expression Data:



For gene expression data, I represent each cell line as a high-dimensional vector of standardized expression values across protein-coding genes. This enables the model to learn biologically relevant transcriptional patterns associated with drug response.

For chemical compounds, I use two complementary representations: **molecular fingerprints** and **SMILES strings**. Fingerprints (e.g., ECFP) provide fixed-length binary vectors capturing chemical substructures and are highly efficient for training models like logistic regression or MLPs. In contrast, SMILES strings retain sequential structure and are tokenized for use with deep models such as 1D CNNs or transformers, which can learn more nuanced patterns from chemical syntax. [links]

As shown in the benchmark results (more Detail is on Part B1):

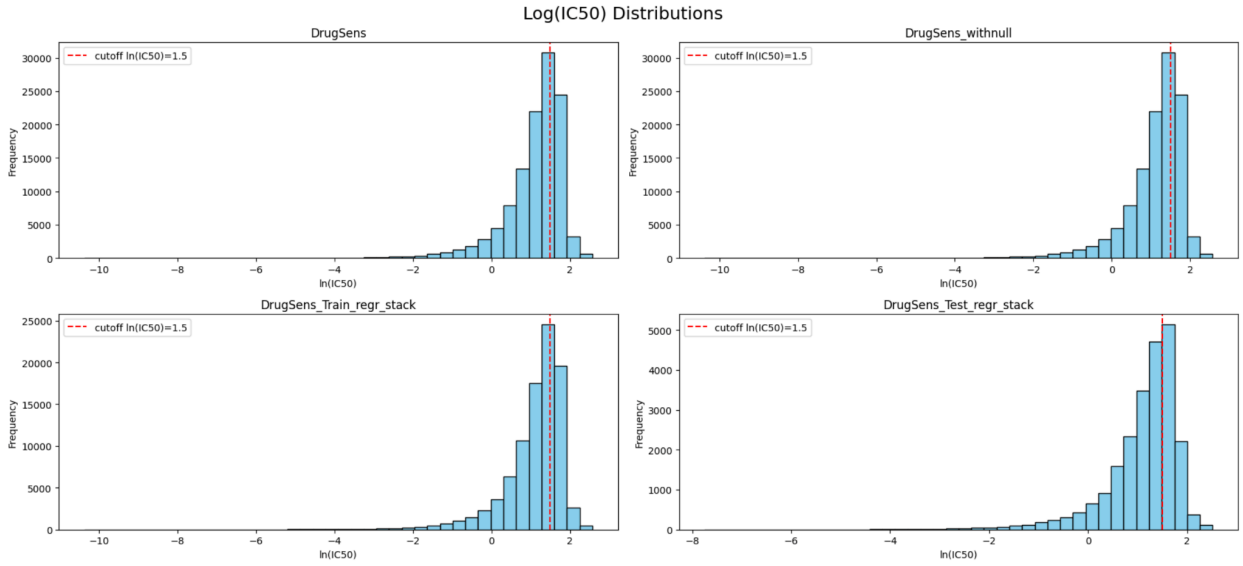


Model	AUROC	Times(s)	Remarks
Fingerprints + Logistic Regression()	0.846	9.49	Fastest, strong baseline
SMILES + 1D CNN (Keras) ()	0.843	90.47	Better generalization
Graphs + GCN (PyTorch Geometric)()	0.540	128.82	Poor score, slowest

These results demonstrate that **fingerprints offer a fast and reliable baseline**, while **SMILES paired with CNNs provide better generalization** at a higher computational cost. In contrast, **graph-based models** underperformed significantly in both accuracy and efficiency. This empirical comparison supports the dual use of fingerprints and SMILES, offering flexibility depending on architectural needs and compute constraints.

Defining the Cut-off Value:

- IC50 values -> how much drug is needed to inhibit growth or Half maximal inhibitory concentration (IC50) is a measure of the potency of a substance in inhibiting a specific biological or biochemical function.[links]



This shows a clear left-skewed distribution. Most ln (IC50) values are between 0 and 2, with a peak near 1.0–1.5. A red dashed vertical line at ln (IC50) = 1.5

- Classification models need binary labels:
  - High sensitivity (IC50 >1.5, label = 1)
  - Low sensitivity (IC50 <1.5,label = 0)

Dataset Splitting (Imbalance?):

I split the dataset into training (70%), validation (15%), and test sets (15%) while maintaining class balance through stratification and then verify the class distribution in each set by calculating the relative frequency of the sensitivity labels using value_counts(normalize=True).

distribution:	High sensitivity (1)	Low sensitivity (0)
Train label	0.666389	0.333611
Validation label	0.666379	0.333621
Test label	0.666403	0.333597

61,683 sensitive samples (67%) vs 30,880 resistant samples (33%) 0.66:0.33 is **still balanced across all splits**

Exploratory Data Analysis (EDA) ,focusing on cancer cell types**:**

- Definition of technical words (In chemistry):
    - **A gene expression profile** [links]**:** may be used to help diagnose a disease or condition, such as cancer. It may also be used to help plan treatment, find out how well treatment is working, make a prognosis, or predict whether cancer will come back or spread to other parts of the body.
    - **Frequency Mean in Gene Expression** [links]**:** gene frequency or allele frequency is a measurement to show the genetic diversity of a population species or in other terms the abundance of its gene pool. Mathematically, it is expressed in terms of percentage or proportion.
    - **Cancer can be described as a disease of altered gene expression** [links]: There are many proteins that **are turned on or off (gene activation or gene silencing**) that dramatically alter the overall activity of the cell. A gene that is not normally expressed in that cell can be switched on and expressed at high levels.
        - **Gene** = blueprint
        - **Expression level** = how often that blueprint is being used to make parts (mRNA -> proteins)
            - **Higher expression** = more activity / importance in the cell
            - **Lower expression** = gene is off or less active
        - **Gene expression** is the level of activity of a gene in a cell.
            - It's typically measured as mRNA abundance how much a gene is **"turned on"** or **"off."**
            - In datasets like CCLE or GDSC, each row is a cell line, and each column is a gene.
            - The values are often log-transformed and standardized before use in machine learning
    - **High expression ≠ always cancer :**
        - **Gene expression profiles in cancers and their therapeutic implications** [links], It shows that cancer isn't just one disease - different patients can have different types even within the same kind.
            - Importantly, it reminds us that **just because a gene is highly active doesn't mean it causes cancer**. Sometimes high activity is a normal part of how the body responds or repairs itself. What matters is *which* genes are active, *where*, and *why*.
                - **Identify molecular subtypes** (e.g., HER2+ or basal-like breast cancer), where *some* high-expression markers correlate with therapy response, but not all.
                - **Predict prognosis**, where *increased* expression of certain genes may correlate with either better or worse survival, depending on the gene's role.
                - **Predict drug sensitivity**, using cell line models, but expression-response correlations vary by drug and genetic background.

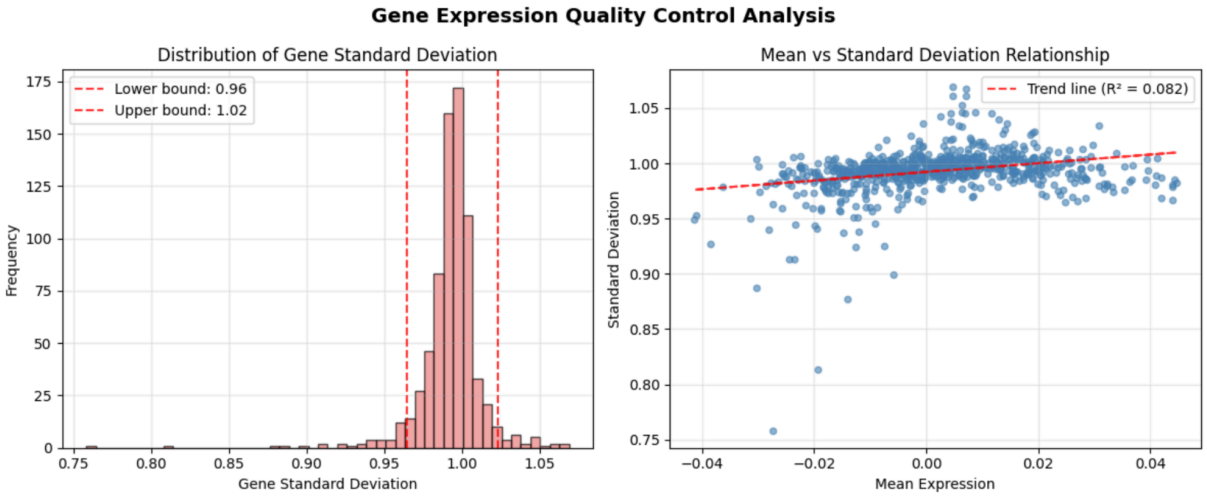
- **Drug sensitivity or drug intolerance** [links]: refers to an inability to tolerate the adverse effects of a medication, generally at therapeutic or subtherapeutic doses, **the proportion of cell lines (or patients) that respond to a drug** typically classified as **“sensitive”** based on a predefined threshold (e.g., ln(IC50) ≤ 1.5).

Insightful analysis:

Outlier analysis:

Threshold starts with the 95th percentile (gene_stats['mean'].quantile(0.95)) as a threshold to flag highly expressed genes in a data-adaptive way. Genes with extremely high coefficients of variation (CV), often due to low mean expression, were identified as potentially unstable. These genes may be excluded or treated with caution in downstream modeling. Overall, **the dataset appears high quality, with minimal technical artifacts and successful standardization**, as shown by the consistent spread of standard deviations and weak correlation between mean and variance.

```
=== PROBLEMATIC GENES IDENTIFIED ===

Top 5 highest expressing genes:
           mean       std
KCNJ5  0.044737  0.982071
IDH1   0.044384  0.985083
KIT    0.043978  0.980503
NCOA1  0.043905  0.966385
BUB1B  0.043851  0.978719

Top 5 most variable outlier genes:
           mean       std
GNA11  0.004762  1.069160
YAP1   0.007008  1.067659
FOXA1  0.007011  1.061431
MYO5A  0.004766  1.061355
SSX4   0.006345  1.052942

Top 5 highest CV genes (potentially noisy):
             mean       std             cv
JAZF1    0.000002  0.995305  426225.579376
USP9X    0.000267  1.008760    3772.429638
KDSR     0.000322  0.996009    3088.513909
BCL7A    0.000333  1.002865    3008.123784
FIP1L1   0.000383  0.992961    2591.128305
```

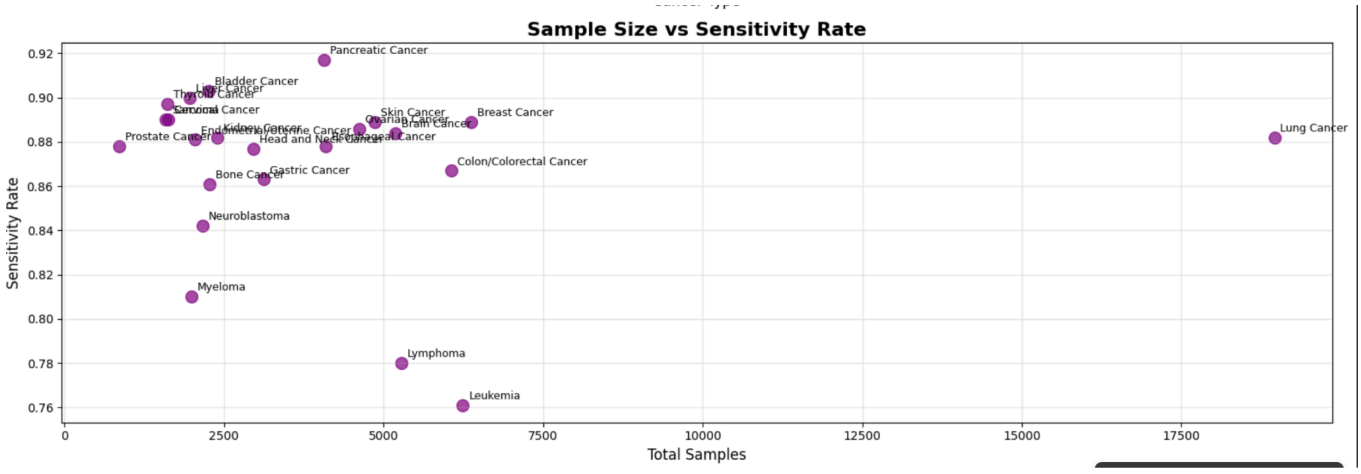


The gene expression quality assessment shows our dataset is properly prepared and reliable for analysis. Most genes have **standard deviations clustered tightly** around 1.0 (ranging from 0.96 to 1.02), demonstrating consistent behavior across the dataset with very few problematic outliers.

The weak relationship between gene expression levels and their variability ($R^2 = 0.082$) is excellent news - it means high-expressing genes won't unfairly dominate the analysis simply because of their magnitude. This prevents bias where louder genes overshadow quieter but equally important ones.

The results confirm successful data standardization with minimal technical errors, creating a stable foundation for machine learning models. **The dataset is clean, balanced, and ready for robust statistical analysis** without concerns about data quality issues affecting model performance.

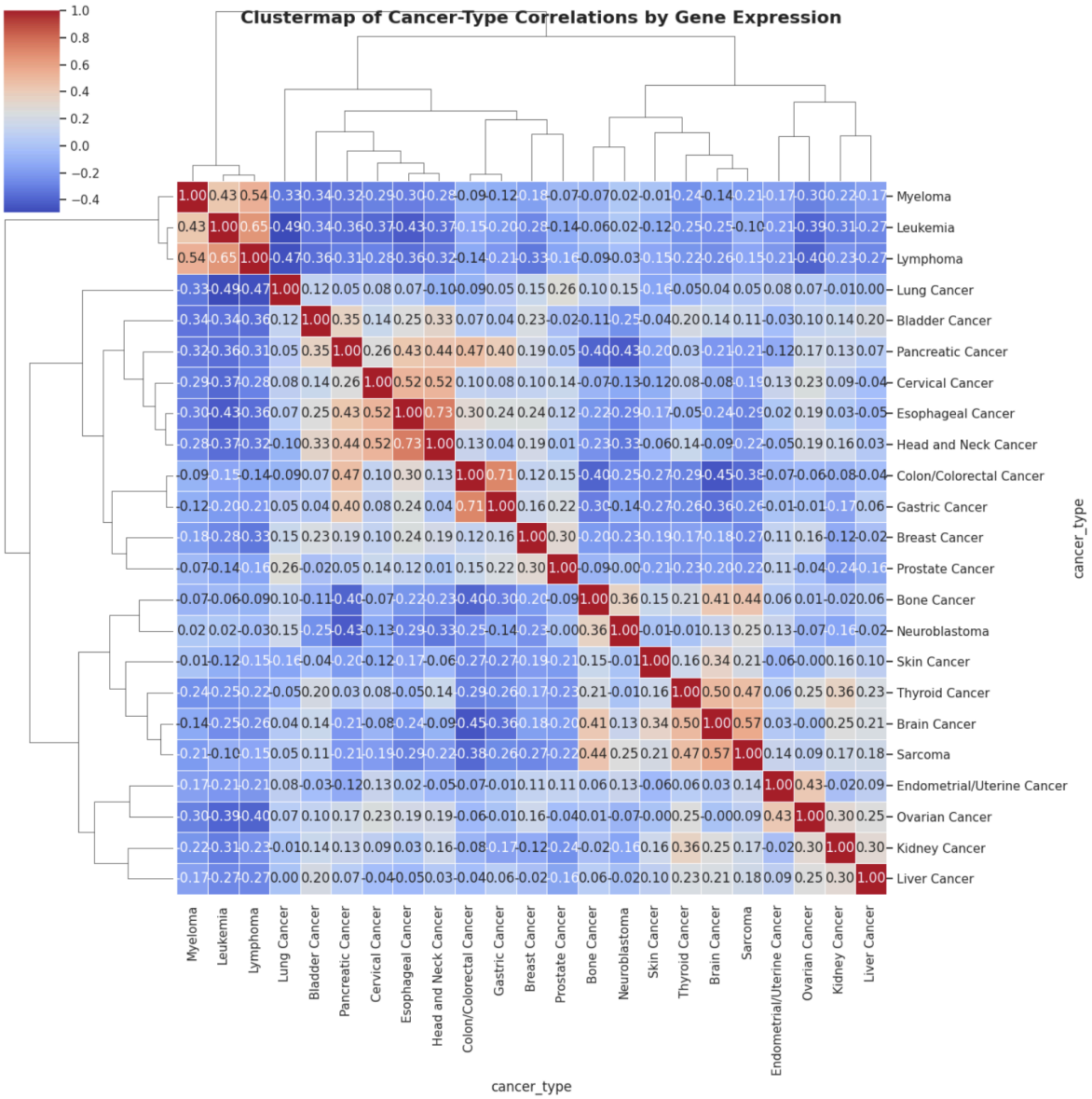
Sensitivity Rate analysis:



The scatter plot illustrates the relationship between sample size and drug sensitivity rates across cancer types, based on a binary classification derived from IC50 values. Specifically,

samples with **IC50 ≤ 1.5** were labeled as **sensitive**, and those with **IC50 > 1.5** as resistant. The **drug sensitivity rate** for each cancer type was computed as the **proportion of sensitive samples**. The analysis reveals that **Pancreatic Cancer** (มะเร็งเม็ดเลือดขาว) and **Bladder Cancer** exhibit **exceptionally high sensitivity rates** of approximately **92% and 91%**, respectively. Most cancer types fall within the **85–92% sensitivity range**, indicating generally strong drug responses. However, **Lymphoma** (มะเร็งต่อมน้ำเหลือง) and **Leukemia** (มะเร็งไขกระดูก) show **notably lower sensitivity rates** around **76–78%**, suggesting higher resistance. Additionally, **Lung Cancer** appears as a distinct **outlier**, exhibiting a lower-than-expected sensitivity rate given its large sample size.

Cancer Similarity analysis:



Method: Calculated correlation matrix from mean gene expression profiles across cancer types and used hierarchical clustering to identify cancers with similar transcriptional patterns.

- Most Similar Cancer Pairs:
    - **Esophageal**(หลอดอาหาร) **& Head/Neck Cancer** (73% similar) - Both affect upper digestive tract
    - **Colon** (มะเร็งลำไส้ใหญ่) **& Gastric (Stomach) Cancer** (71% similar) - Both are digestive system cancers
    - **Leukemia** (มะเร็งเม็ดเลือดขาว) **& Lymphoma** (มะเร็งต่อมน้ำเหลือง) (65% similar) - Both are blood cancers
    - **Brain & Sarcoma**(เนื้อเยื่อเกี่ยวพัน) (57% similar) - Both affect connective tissues
    - **Lymphoma** (ต่อมน้ำเหลือง) **& Myeloma** (ไขกระดูก) (53.5% similar) - Both are blood system cancers

Cancers from the **same body system** (digestive, blood, etc.) **share genetic patterns.**

- Least Similar Cancer Pairs:
    - **Leukemia**(เม็ดเลือดขาว) & **Lung Cancer** (- 49% similar) - Blood vs. organ cancer
    - **Lung** & **Lymphoma** (ต่อมน้ำเหลือง) (- 47% similar) - Organ vs. blood cancer
    - **Brain** & **Colon Cancer** (มะเร็งลำไส้ใหญ่) (- 45% similar) - Neural vs. digestive cancer
    - **Esophageal** (หลอดอาหาร) & **Leukemia** (มะเร็งเม็ดเลือดขาว) (- 43% similar) - Digestive vs. blood cancer
    - **Neuroblastoma** (เซลล์ประสาทในเด็ก) & **Pancreatic** (ตับอ่อน) (- 43% similar) - Childhood brain vs. adult organ cancer

Blood cancers are completely different from solid organ cancers - they're almost genetic opposites

	Cancer_Type	Total_Correlation_Sum
0	Ovarian Cancer	1.162118
1	Bladder Cancer	0.929126
2	Thyroid Cancer	0.905991
3	Cervical Cancer	0.893943
4	Head and Neck Cancer	0.866622
5	Liver Cancer	0.745675
6	Esophageal Cancer	0.736545
7	Kidney Cancer	0.508976
8	Pancreatic Cancer	0.481716
9	Endometrial/Uterine Cancer	0.437945
10	Sarcoma	0.166107
11	Brain Cancer	0.158718
12	Breast Cancer	-0.226489
13	Gastric Cancer	-0.314584
14	Prostate Cancer	-0.421143
15	Lung Cancer	-0.423842
16	Bone Cancer	-0.479138
17	Colon/Colorectal Cancer	-0.697139
18	Skin Cancer	-0.911528
19	Neuroblastoma	-1.298209
20	Myeloma	-2.876995
21	Lymphoma	-3.803219
22	Leukemia	-3.991795

Most Similar Cancers:

Method: Computed total correlation scores by summing pairwise gene expression correlations for each cancer type (excluding self-correlations) to measure overall transcriptional similarity to other cancer types

- **Ovarian**(รังไข่)**, Bladder** (กระเพาะปัสสาวะ)**, Thyroid**(ต่อมไทรอยด์)**, Cervical** (ปากมดลูก) - these cancers have gene patterns that look like many other cancer types

Most Different Cancers:

- **Blood cancers** (Leukemia (เลือดขาว), Lymphoma (ต่อมน้ำเหลือง) , Myeloma(ไขกระดูก)) - completely different from solid tumors
- **Neuroblastoma** (เซลล์ประสาทในเด็ก), **Skin, Colon, Bone** - each has very specialized gene patterns

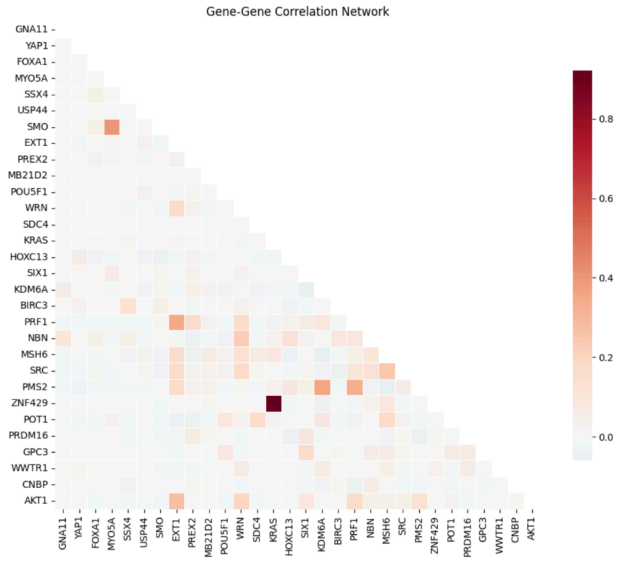
```
Sorted Cancer Types by Mean Gene Expression:

cancer_type
Brain Cancer                    0.130693
Bone Cancer                     0.125633
Lung Cancer                     0.122086
Prostate Cancer                 0.104060
Sarcoma                         0.093442
Neuroblastoma                   0.070846
Skin Cancer                     0.039039
Bladder Cancer                  0.029668
Breast Cancer                   0.029093
Liver Cancer                    0.006661
Gastric Cancer                 -0.002198
Leukemia                       -0.032790
Thyroid Cancer                 -0.035825
Endometrial/Uterine Cancer     -0.036619
Ovarian Cancer                 -0.058144
Myeloma                        -0.063620
Pancreatic Cancer              -0.080849
Colon/Colorectal Cancer        -0.083294
Kidney Cancer                  -0.086183
Esophageal Cancer              -0.105742
Lymphoma                       -0.125348
Head and Neck Cancer           -0.224498
Cervical Cancer                -0.277937
```

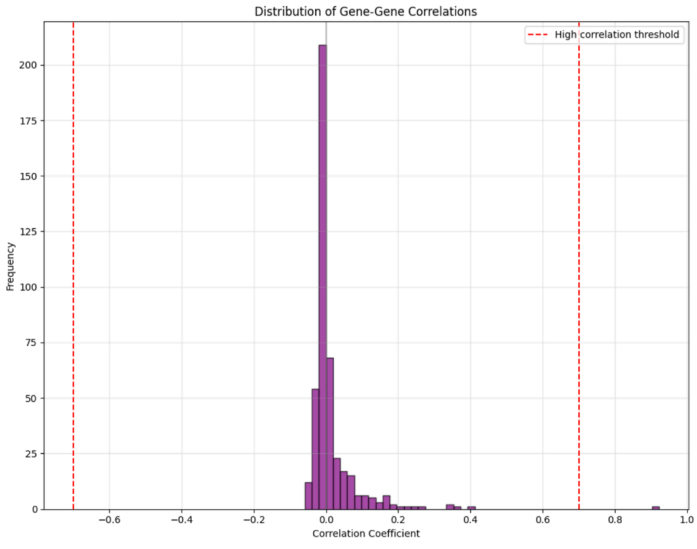
Gene Expression analysis:

Method: Calculated mean gene expression for each cancer type by averaging all gene expression values within each type. This approach quantifies overall transcriptional activity per cancer type, enabling comparison of gene expression patterns and identification of outliers across tumor types.

Gene expression levels don't predict cancer severity - Brain and Bone cancers show high gene activity, while aggressive cancers like **Leukemia** (เม็ดเลือดขาว) and **Pancreatic** (ตับอ่อน) cancer show low activity.

Gene activity patterns help understand cancer biology and guide treatment decisions, but **higher gene expression doesn't mean more dangerous cancer.**





Method: selected the 30 most variable genes and calculated how strongly each pair is correlated using **Pearson correlation** gene pairs with a correlation above 0.7 were considered

highly related, but the mean correlation is approximately -0.1. I visualized the results with a heatmap and a histogram to show overall patterns and detect strongly co-expressed gene pairs.

Among 435 gene pair combinations, only one pair (**KRAS** and **ZNF429**) showed high correlation (r = 0.92), while remaining pairs demonstrated largely independent expression patterns.This analysis **reveals minimal redundancy in the gene set**, with each gene contributing unique biological information valuable for predictive modeling, though the the strong correlation between KRAS and ZNF429 represents biologically meaningful coordination rather than redundancy. KRAS is a key cancer-driving gene [link], while ZNF429 regulates cell death pathways and tumor suppressor genes [link]. Their coordinated expression suggests they function as part of the same cancer regulatory network, making this correlation likely reflects **biological coordination** rather than mere statistical overlap.

Summarize all insight:

- **Drug Sensitivity Patterns**
  - **Highest sensitivity:** Pancreatic (92%) and Bladder (91%) cancers respond best to treatments
  - **Lowest sensitivity:** Blood cancers (Leukemia ~76%, Lymphoma ~78%) show more drug resistance
  - **Outlier:** Lung cancer has unexpectedly low sensitivity despite large sample size
- **Cancer Biology Insights**
  - **Blood vs. Solid tumors**: Blood cancers are genetically opposite to solid organ cancers (-45% to -49% similarity)
  - **System-based similarities:** Cancers from **same body systems** cluster together (digestive cancers 70-73% similar, blood cancers 53-65% similar)
  - **Expression ≠ Severity:** High gene activity doesn't mean more dangerous cancer (Brain/Bone high activity, but Leukemia/Pancreatic low activity yet aggressive)
- **Gene Expression Characteristics**
  - **Minimal redundancy: Only 1 out of 435 gene pairs highly correlated** (KRAS-ZNF429), confirming independent biological information
  - **Specialized patterns:** Some cancers (Neuroblastoma, Skin, Colon, Bone) have unique gene signatures
  - **Universal patterns:** Other cancers (Ovarian, Bladder, Thyroid, Cervical) share similarities with multiple cancer types