BROAD INSTITUTE - A leading biomedical and genomic research institute.

- RNAi (RNA interference) Dependency:
  - What it is: Data showing how essential each gene is for cell survival in different cell lines, measured by the impact of reducing gene expression using RNA interference.
  - Why it is used: Crucial for understanding gene function and identifying potential drug targets; connects gene essentiality to cell viability, relevant to drug response.
- RNA Expression Public:
  - What it is: Levels of gene expression (mRNA) in various cell lines.
  - Why it is used: Provides a baseline measure of gene activity, essential for comparing changes after drug treatment.
- CRISPR-Cas9 gene knockout dependency:
  - What it is: Data showing how essential each gene is for cell survival in different cell lines, measured by the impact of completely removing the gene using CRISPR-Cas9.
  - Why it is used: Similar to RNAi, but a more complete knockout; provides complementary information about gene function and essentiality, strengthens target identification.
- WES Log Copy Number:
  - What it is: The number of copies of each gene in the genome of different cell lines, obtained by Whole Exome Sequencing.
  - Why it is used: Gene copy number alterations can affect gene expression and drug sensitivity; important for understanding the context of drug response.
- Damaging Mutations:
  - What it is: Information about mutations in genes that are predicted to disrupt protein function.
  - Why it is used: Mutations can influence drug response by altering drug targets or downstream pathways; critical for personalized medicine.

CCLE (Cancer Cell Line Encyclopedia) - A collaboration between the Broad Institute and the Novartis Institutes for Biomedical Research that provides genomic data from a large panel of cancer cell lines.

- Methylation:
  - What it is: The level of DNA methylation at specific sites in the genome of different cell lines.
  - Why it is used: DNA methylation can regulate gene expression and is influenced by drug treatment; provides another layer of information about gene regulation.
- TPM RNA Seq:
  - What it is: Transcript Per Million, a measure of gene expression (mRNA) in various cell lines, obtained by RNA sequencing.
  - Why it is used: Similar to the Broad Institute's RNA Expression data, but from a different source and possibly processed differently; provides a more robust measure of gene expression.
- Protein Expression:
  - What it is: The amount of each protein present in different cell lines.
  - Why it is used: Protein levels are more directly related to cellular function than mRNA levels; provides crucial information about the downstream effects of drug treatment.

GO Consortium (Gene Ontology Consortium) - Provides a standardized, hierarchical classification of gene function.

- GO Profiles:
  - What it is: Annotations of genes with terms from the Gene Ontology, describing their molecular function, biological process, and cellular component.
  - Why it is used: GO profiles provide functional context for genes; helps understand the biological pathways and processes affected by drug treatment and can be used for pathway enrichment analysis.

Sanger Institute - A world leader in genomics research.

- CFGs (Cancer Functional Genes):
  - What it is: A curated list of genes whose mutations are known to drive cancer development.
  - Why it is used: CFGs are often drug targets or involved in drug response pathways; including this information helps prioritize genes and understand the impact of drugs on cancer-related processes.

EMBL-EBI (European Molecular Biology Laboratory - European Bioinformatics Institute) - Provides resources and services for biological data, including genomic information.

- GC Content:
  - What it is: The percentage of guanine (G) and cytosine (C) bases in a gene sequence.
  - Why it is used: GC content can influence gene expression and other genomic features; might be helpful in feature engineering for the model.
- Gene Length:
  - What it is: The number of base pairs in a gene.
  - Why it is used: Gene length can be correlated with other features, such as expression level; could be a useful feature for the model.
- Transcript Count:
  - What it is: The number of different mRNA transcripts produced by a gene.
  - Why it is used: Different transcripts can have different functions; this information adds complexity and detail about gene expression.
- Chromosome Position:
  - What it is: The location of a gene on a chromosome.
  - Why it is used: Chromosome position can be relevant to gene regulation and other genomic features; might be useful for understanding gene context
  -

ENCODE (Encyclopedia of DNA Elements) - An international consortium aiming to identify all functional elements in the human genome.

- Histone ChIP-Seq:
  - What it is: Chromatin Immunoprecipitation sequencing that identifies where specific histone modifications are located across the genome (e.g., H3K4me3, H3K4me1, H3K27ac). These modifications are often associated with active or repressed gene expression.
  - Why it is used: Histone modifications play a crucial role in regulating gene expression; this data helps understand the chromatin landscape and how it changes with drug treatment. Different histone marks have distinct roles (e.g., H3K4me3 at active promoters, H3K27me3 at repressed regions).
- TF ChIP-Seq:
  - What it is: Chromatin Immunoprecipitation sequencing that identifies where specific transcription factors (TFs) bind to DNA (e.g., POLR2A, CTCF).
  - Why it is used: TFs control gene expression by binding to specific DNA sequences; this data reveals which TFs are active and where they are binding, providing insights into gene regulatory mechanisms.
- ATAC-Seq (Assay for Transposase-Accessible Chromatin with high-throughput sequencing):
  - What it is: A technique to identify regions of open chromatin, where DNA is accessible to proteins.
  - Why it is used: Open chromatin is generally associated with active gene expression; this data helps map regulatory regions and understand how chromatin accessibility changes with drug treatment.
- DNase-Seq (DNase I hypersensitive sites sequencing):
  - What it is: A technique to identify regions of open chromatin, similar to ATAC-seq, but using DNase I enzyme to cut accessible DNA.
  - Why it is used: Provides another measure of chromatin accessibility; can be used in conjunction with ATAC-seq to get a more comprehensive view of the open chromatin landscape.