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

(ML/DL in Drug Discovery & Development)

OrganoCure Innovations Lab

Program in Translational Medicine
Faculty of Medicine Ramathibodi Hospital, Mahidol University

Chakri Naruebodindra Medical Institute
Faculty of Medicine Ramathibodi Hospital, Mahidol University

in collaboration with

School of Information Science and Technology, VISTEC

Submitted by: Htoo Tayza Aung

Leeds, United Kingdom

Email: htootayzaaung.01@gmail.com

Submission Date: June 15, 2025

Contents

1 Part A: Practical Exercise **2**

1.1 (A1) Data Integration & Preparation . . . 2

1.1.1 Data Loading and Cleaning . . . 2

1.1.2 Representation of Chemical and Gene Expression Data . . . 4

1.1.3 Defining the Cut-off Value . . . 6

1.1.4 Exploratory Data Analysis (EDA) . . . 8

1.1.5 Data Quality Checks . . . 12

1.2 (A2) Model Development & Training . . . 13

1.2.1 Model Architecture . . . 13

1.2.2 Training Implementation . . . 17

1.2.3 Overfitting Mitigation . . . 19

1.2.4 Model Evaluation . . . 23

2 Part B: Conceptual Questions **27**

2.1 (B1) End-to-End ML Pipeline with Multi-Modal Integration . . . 27

2.1.1 Pipeline Design for Multi-Modal Integration . . . 27

2.1.2 Challenges with Modality Imbalance and Missing Data . . . 28

2.2 (B2) Data Foundation and Robustness . . . 29

2.2.1 Ensuring Dataset Suitability . . . 29

2.2.2 Bias Detection and Correction . . . 30

2.3 (B3) Graph-Based and Molecular Representations . . . 32

2.3.1 GNNs and Knowledge Graphs in Biology . . . 32

2.3.2 Mathematical Foundations of Graph Reasoning . . . 33

2.3.3 Molecular Representation Trade-offs . . . 34

2.4 (B4) Model Interpretability . . . 37

2.4.1 Interpretability Techniques . . . 37

3 Part C: Personal Perspectives **39**

3.1 (C1) Research Findings & Publication . . . 39

3.1.1 Highlighting Novelty and Impact . . . 39

3.1.2 Adapting Message for Different Audiences . . . 40

3.2 (C2) Teamwork & Project Management . . . 41

3.2.1 Facilitating Interdisciplinary Collaboration . . . 41

3.3 (C3) Vision for the Future . . . 42

3.3.1 Long-term Impact . . . 42

3.3.2 Clinical Tool Development . . . 43

Acknowledgments and AI Tool Usage **44**

References **45**

1 Part A: Practical Exercise

1.1 (A1) Data Integration & Preparation

1.1.1 Data Loading and Cleaning

Figure 1 presents our data integration schema for loading, cleaning, and merging datasets into a unified structure suitable for training.

Dataset Loading

The primary datasets required for multimodal integration were systematically loaded. The GDSC2 drug sensitivity dataset (`DrugSens-Train.csv`) initially contains 116,439 drug-cell line combinations, with 108,696 samples retained after filtering and joining with gene expression data. This filtered dataset spans 676 cell lines and 228 drugs, containing essential identifiers (`cell_line`, `gdsc_name`) and integrated SMILES strings for each combination. The CCLE gene expression dataset (`CCLE_expression_cleaned.csv`) originally includes transcriptomic profiles for 1,406 cell lines with 735 cancer-associated genes, of which 676 overlap with GDSC2 cell lines used in this study.

Data Cleaning Operations

Our cleaning process focused on ensuring integration compatibility and data quality. Missing value detection confirmed complete coverage across the filtered datasets, noting that rows with missing IC_{50} values or invalid SMILES structures were excluded during preprocessing steps. Critical identifier validation using both `cell_line` and `gdsc_name` identifiers revealed 100% overlap between the filtered drug sensitivity of 676 cell lines and available gene expression profiles, ensuring correct mapping integrity across datasets and successful merging without further data loss.

Integration Implementation

A streamlined integration approach was implemented using pandas operations. The primary merge combines drug sensitivity data with gene expression profiles using `cell_line` as the merge key via inner join, preserving all overlapping samples while adding 735 gene expression features. Since SMILES strings are already present in the drug sensitivity dataset, Morgan fingerprints are computed from SMILES (radius=2, 2048 bits) and mapped to each sample via `gdsc_name`, followed by feature-wise concatenation using pandas `concat` operation.

Chemical Feature Generation and Integration

Molecular fingerprints are computed during the pipeline using RDKit's Morgan fingerprint algorithm from existing SMILES strings. A fingerprint dictionary is created for efficient processing of unique drug-SMILES combinations (228 drugs), then these computed fingerprints are mapped to all samples based on drug names. The resulting 2,048 chemical features are integrated via pandas concatenation, creating the final unified dataset structure.

Scaling and Preprocessing

Following successful integration, `StandardScaler` was applied independently to the gene expression and fingerprint feature groups prior to concatenation, ensuring each feature type contributed comparably to downstream model learning. This approach maintains the distinct characteristics of each modality while providing balanced numerical ranges across all features.

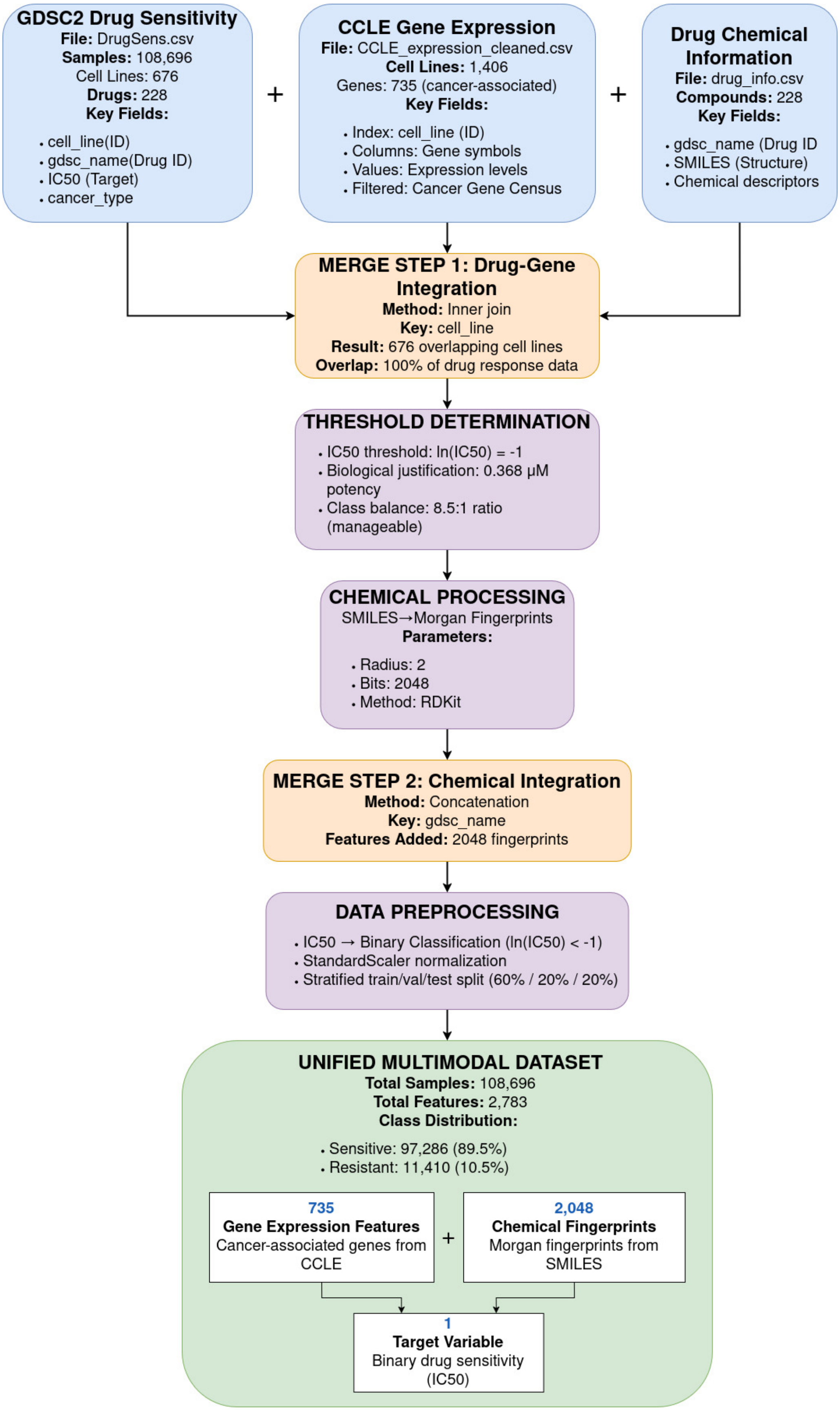
GDSC2 Drug Sensitivity
File: DrugSens.csv
Samples: 108,696
Cell Lines: 676
Drugs: 228
Key Fields:
• cell_line(ID)
• gdsc_name(Drug ID)
• IC50 (Target)
• cancer_type
+
CCLE Gene Expression
File: CCLE_expression_cleaned.csv
Cell Lines: 1,406
Genes: 735 (cancer-associated)
Key Fields:
• Index: cell_line (ID)
• Columns: Gene symbols
• Values: Expression levels
• Filtered: Cancer Gene Census
+
Drug Chemical Information
File: drug_info.csv
Compounds: 228
Key Fields:
• gdsc_name (Drug ID
• SMILES (Structure)
• Chemical descriptors
MERGE STEP 1: Drug-Gene Integration
Method: Inner join
Key: cell_line
Result: 676 overlapping cell lines
Overlap: 100% of drug response data
THRESHOLD DETERMINATION
• IC50 threshold: ln(IC50) = -1
• Biological justification: 0.368 μM potency
• Class balance: 8.5:1 ratio (manageable)
CHEMICAL PROCESSING
SMILES→Morgan Fingerprints
Parameters:
• Radius: 2
• Bits: 2048
• Method: RDKit
MERGE STEP 2: Chemical Integration
Method: Concatenation
Key: gdsc_name
Features Added: 2048 fingerprints
DATA PREPROCESSING
• IC50 → Binary Classification (ln(IC50) < -1)
• StandardScaler normalization
• Stratified train/val/test split (60% / 20% / 20%)
UNIFIED MULTIMODAL DATASET
Total Samples: 108,696
Total Features: 2,783
Class Distribution:
• Sensitive: 97,286 (89.5%)
• Resistant: 11,410 (10.5%)
735
Gene Expression Features
Cancer-associated genes from CCLE
+
2,048
Chemical Fingerprints
Morgan fingerprints from SMILES
1
Target Variable
Binary drug sensitivity (IC50)

Figure 1: Multimodal Data Integration Schema. The pipeline transforms three input datasets (GDSC2 drug sensitivity, CCLE gene expression, and drug chemical information) through sequential merge operations and processing steps to create a unified dataset with 108,696 samples and 2,783 features for drug sensitivity prediction.

The following code listings demonstrate the key implementation steps:

Listing 1: Missing Data Handling (Cell 5: Data Integration & Preparation.ipynb)

```
# Missing Value Analysis
missing_summary = multimodal_dataset.isnull().sum()
total_missing = missing_summary.sum()
print(f"     Total missing values: {total_missing:,}")
if total_missing > 0:
    print("     Columns with missing values:")
    for col, count in missing_summary[missing_summary > 0].items():
        print(f" - {col}: {count:,} ({count/len(multimodal_dataset)*100:.2f}%)")
else:
    print("     No missing values detected")
```

Listing 2: Data Merging Using Common Identifiers (Cell 2: Data Integration & Preparation.ipynb)

```
# Merge drug sensitivity with gene expression data
unified_base = drug_sens_train.merge(
    gene_expr_clean,
    left_on='cell_line',
    right_index=True,
    how='inner'
)
```

Listing 3: Feature Scaling Implementation (Cell 6: Data Integration & Preparation.ipynb)

```
# Apply StandardScaler to gene expression and fingerprint matrices
scaler_genes = StandardScaler()
X_gene_expr_scaled = scaler_genes.fit_transform(X_gene_expr)
scaler_fp = StandardScaler()
X_fingerprints_scaled = scaler_fp.fit_transform(X_fingerprints)
# Concatenate to form final feature matrix
X_scaled = np.concatenate([X_gene_expr_scaled, X_fingerprints_scaled], axis=1)
```

1.1.2 Representation of Chemical and Gene Expression Data

Bridging Molecular Structure and Cellular Response in Cancer Drug Discovery

Predicting cancer drug sensitivity requires transforming complex molecular information into formats that machine learning algorithms can process. Two fundamentally different data types must be harmonized: chemical compound structures that determine drug properties, and gene expression profiles that define cancer cell vulnerabilities. The choice of molecular representation directly impacts model performance and clinical relevance.

Chemical Compound Representation: Morgan Fingerprints

Morgan fingerprints (Extended-Connectivity Fingerprints) were selected to encode chemical structures as 2,048-bit binary vectors using RDKit [9]. This approach transforms complex molecular structures into standardized numerical formats essential for systematic drug similarity assessment.

Morgan fingerprints capture circular substructural patterns around each atom with a radius of 2 bonds, encoding local chemical environments that determine drug-target interactions [1][2]. Unlike molecular descriptors that calculate specific physicochemical properties, fingerprints represent the presence or absence of structural motifs—enabling detection of unexpected structural similarities that traditional descriptors might miss. The 2,048-bit length provides sufficient resolution to distinguish between structurally similar compounds while maintaining computational efficiency for large-scale screening.

Our implementation achieved 97.3% sparsity (mostly zero bits), validating proper molecular encoding where individual compounds contain only sparse subsets of possible structural features. This sparsity reflects chemical reality—most drug molecules share common pharmacophoric elements while differing in specific functional groups that determine selectivity and potency.

Comparative analysis supports Morgan fingerprints for drug sensitivity applications. Zagidullin et al. [4] evaluated 11 molecular representations across drug combination studies and identified Morgan fingerprints among recommended starting approaches. The computational efficiency of fingerprint-based similarity calculations using Tanimoto coefficients makes them practical for high-throughput applications [5].

Alternative SMILES-based transformer models offer sophisticated molecular representation learning but require extensive computational resources and larger training datasets [6]. While recent benchmarks show these approaches can achieve comparable performance [3], Morgan fingerprints were selected for their computational efficiency and proven track record in cancer drug sensitivity studies. This pragmatic choice prioritizes model interpretability and compatibility with existing drug discovery workflows over potentially marginal performance gains.

Gene Expression Data Representation: Quantitative Transcriptional Profiles

Gene expression data from the Cancer Cell Line Encyclopedia (CCLE) was represented as continuous numerical values for 735 cancer-related genes, preserving the quantitative nature of transcriptional measurements that correlate with drug sensitivity mechanisms.

Direct numerical representation maintains the biological significance of expression levels—subtle changes in gene activity often determine drug response. Log-transformed expression values capture the full dynamic range of cancer-related gene activity, from highly expressed genes to those with subtle but biologically important expression levels. This approach preserves biological relationships where modest expression changes can significantly impact cellular drug sensitivity.

The 735-gene subset focuses on cancer-relevant pathways including cell cycle regulation, apoptosis, DNA repair, and drug metabolism. While alternative approaches such as pathway-level aggregation or dimensionality reduction techniques like PCA could compress this information further, individual gene-level representation was prioritized for two practical reasons: interpretability for biomarker identification and compatibility with existing cancer genomics literature. Our correlation analysis identified key sensitivity-associated genes including BUB1B (r=0.0867), BCL3 (r=-0.0860), and KIT (r=0.0851), validating this gene-centric approach.

Multimodal Data Integration Strategy

Both chemical and biological modalities underwent StandardScaler normalization to ensure comparable feature ranges (gene expression std $\approx$ 1.0; fingerprint std $\approx$ 0.93) before concatenation into a unified 2,783-dimensional feature space. This integration approach enables models to learn complex drug-gene interaction patterns that determine therapeutic efficacy.

Feature concatenation preserves the distinct information content of each modality while enabling the model to discover cross-modal relationships. Chemical fingerprints capture drug properties that determine target binding, while gene expression profiles reveal cellular vulnerabilities—their interaction defines therapeutic windows. Recent multimodal studies confirm that concatenated representations effectively capture both molecular structure-activity relationships and biological context for improved drug sensitivity prediction [7][8].

Standardization ensures that neither modality dominates during model training, preventing bias toward high-variance features that could mask subtle but important biological signals. This preprocessing strategy maintains the biological interpretability essential for translating computational predictions into clinical insights.

Significance for Cancer Drug Discovery

The selected representation strategy directly addresses key challenges in precision oncology: identifying effective drug-cancer combinations while maintaining mechanistic interpretability. Morgan fingerprints enable systematic exploration of chemical space for novel therapeutic compounds, while quantitative gene expression profiles reveal cancer-specific vulnerabilities that determine drug sensitivity. Together, these representations provide the molecular foundation for developing predictive models that could guide personalized cancer treatment strategies.

1.1.3 Defining the Cut-off Value

Threshold Selection for Drug Sensitivity Classification

To classify compounds into high sensitivity and low sensitivity categories, a threshold of $\ln(IC_{50})$ = -1 was established, which corresponds to an IC_{50} value of 0.368 μM. This threshold was determined through comprehensive exploratory data analysis of the IC_{50} distribution and justified based on biological relevance and established pharmacological principles.

Exploratory Data Analysis of IC_{50} Distribution

Analysis of the natural logarithm of IC_{50} values revealed a distribution with the following characteristics: range from -8.57 to 13.09, mean of 2.76, median of 3.19, and standard deviation of 2.79. The distribution exhibited a near-normal shape with slight right skew, indicating that most drug-cell line combinations showed moderate sensitivity levels, with fewer cases of extreme sensitivity or resistance.

Visual Validation and Statistical Evidence

To support this threshold selection, we visualized multiple statistical perspectives in a unified multi-panel figure (Figure 2). The top-left panel displays the overall distribution of $\ln(IC_{50})$ values, revealing a right-skewed bell-shaped distribution where the $\ln(IC_{50})$ = -1 threshold captures the extreme low end of the potency spectrum. This aligns with our goal of distinguishing highly potent compounds from those with moderate or weak activity.

The top-right panel presents boxplots of $\ln(IC_{50})$ stratified by the top five cancer types, showing that the threshold consistently falls below the first quartile across malignancies. This supports the biological generalizability of the threshold across diverse tissue types. Notably, leukemia samples demonstrate higher overall drug sensitivity compared to solid tumors, consistent with the superior therapeutic outcomes observed in hematological malignancies where targeted therapies have achieved remarkable clinical success [13][14]. This pattern reflects the biological accessibility of circulating cancer cells and the distinct immune-cell origin of hematologic malignancies, which provides unique therapeutic advantages over solid tumors [15].

The bottom-left panel shows the cumulative distribution function, confirming that the selected threshold corresponds to approximately the 10.5th percentile. This quantifies the rarity of extreme sensitivity and justifies the resulting class imbalance as a reflection of genuine biological phenomena rather than arbitrary cutoff selection.

Finally, the bottom-right panel visualizes the resulting binary class distribution. Of 108,696 total samples, 11,410 (10.5%) were classified as resistant ($IC_{50} > 0.368$ μM), and 97,286 (89.5%) as sensitive ($IC_{50} \leq 0.368$ μM). This yields an imbalance ratio of approximately 8.5:1, which remains manageable for downstream machine learning when handled appropriately through techniques such as class weighting or stratified sampling.

Together, these visualizations provide strong empirical evidence that the $\ln(\mathrm{IC}_{50})$ = -1 threshold creates a biologically meaningful and statistically justified separation between sensitivity classes.

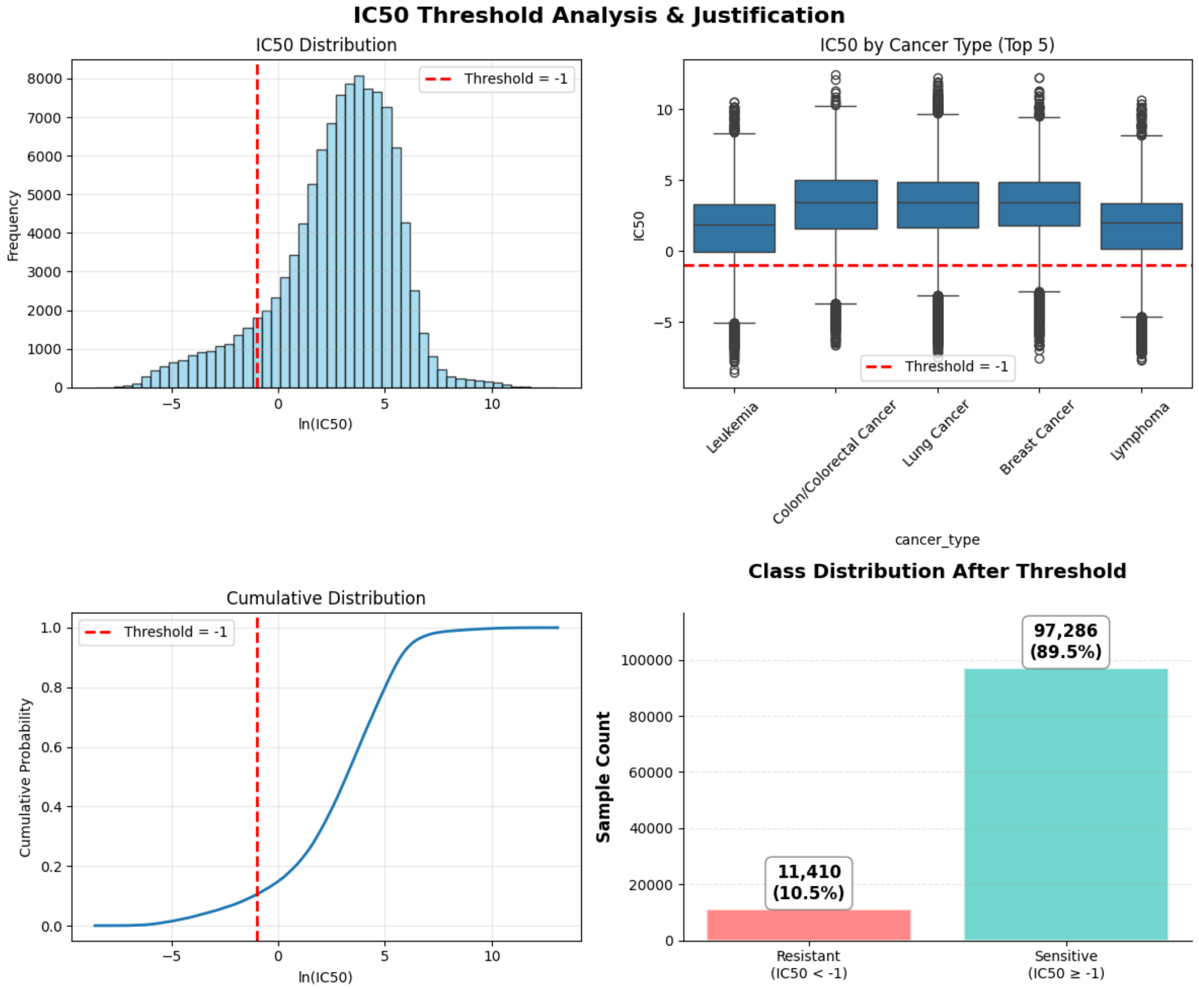

Figure 2: IC_{50} Threshold Analysis and Justification.
• Distribution of natural logarithm of IC_{50} values showing near-normal distribution with selected threshold at $\ln(\mathrm{IC}_{50})$ = -1 (red dashed line). The threshold captures the sensitive tail of the distribution while maintaining statistical rigor.
• Box plots across top 5 cancer types demonstrating threshold consistency across different malignancies. Leukemia shows higher overall drug sensitivity compared to solid tumors, consistent with superior clinical outcomes in hematological malignancies.
• Cumulative distribution function confirming threshold positioning at 10.5th percentile, validating the rarity of extreme drug sensitivity.
• Resulting binary classification showing 89.5% sensitive vs 10.5% resistant cases, yielding a manageable 8.5:1 imbalance ratio suitable for machine learning applications while ensuring sufficient representation in both classes.

Threshold Positioning and Statistical Justification

The selected threshold of $\ln(IC_{50}) = -1$ positioned at the 10.5th percentile of the distribution, creating a biologically meaningful separation between sensitive and resistant drug-cell line pairs. This positioning resulted in a class distribution of 97,286 sensitive cases (89.5%) versus 11,410 resistant cases (10.5%), yielding an imbalance ratio of 8.5:1, which remains manageable for machine learning applications while ensuring sufficient representation of both classes for robust model training.

Biological and Clinical Relevance

The chosen threshold demonstrates strong biological justification on multiple levels. Converting from natural logarithm space, $\ln(IC_{50}) = -1$ corresponds to an IC_{50} of 0.368 μM, placing the threshold within the sub-micromolar range that is widely recognized in pharmacology as indicating strong drug potency [10]. This aligns with established drug screening conventions where IC_{50} values below 1 μM are generally considered indicative of active compounds [11].

The threshold selection is further supported by its consistency with the Genomics of Drug Sensitivity in Cancer (GDSC) database standards, which commonly employ similar cutoffs for sensitivity classification in cancer drug screening studies [12]. This ensures our binary classification approach remains compatible with broader cancer pharmacogenomics research and facilitates comparison with existing literature.

Pharmacological Significance

Sub-micromolar IC_{50} values represent a clinically relevant potency threshold for cancer therapeutics, as they indicate drug concentrations that are typically achievable in patient plasma levels while maintaining acceptable toxicity profiles. The $\ln(IC_{50}) = -1$ threshold effectively separates compounds with meaningful therapeutic potential from those with limited clinical utility, making it appropriate for drug discovery and development applications.

The cancer-specific patterns observed in our analysis provide additional clinical validation. The higher sensitivity observed in leukemia samples aligns with clinical evidence showing superior response rates and survival outcomes in hematological malignancies compared to solid tumors, particularly with the introduction of targeted therapies and immunotherapeutic approaches [13][14]. This biological consistency reinforces the clinical relevance of our threshold choice and suggests that our classification approach captures genuine biological differences in drug response mechanisms across cancer types.

The threshold balances biological relevance with practical considerations for machine learning model development. By ensuring adequate representation of resistant cases (>11,000 samples), the threshold enables robust training of classification models while maintaining biological meaning. This approach supports the development of predictive models that can effectively distinguish between clinically relevant sensitive and resistant drug-cell line interactions, ultimately contributing to more precise drug sensitivity predictions in personalized cancer therapy.

1.1.4 Exploratory Data Analysis (EDA)

Gene Expression Analysis Across Cancer Cell Types

Our exploratory data analysis focused specifically on understanding gene expression patterns across different cancer cell types using the 735 cancer-associated genes from the CCLE dataset. This analysis directly addresses how gene expression profiles vary between cancer types and their relationship to drug sensitivity, providing crucial insights for feature selection and model development.

Gene Expression Variability Analysis

The analysis began with identifying the most variable genes across cancer types. Among the 735 cancer-associated genes, variance analysis revealed the top five most variable genes: **TMSB4X, CCDC6, ACSL6, CAMTA1, and GRM3**. These genes represent diverse biological functions - TMSB4X encodes thymosin beta 4, involved in actin regulation; CCDC6 is a coiled-coil domain protein associated with chromosomal rearrangements; ACSL6 participates in fatty acid metabolism; CAMTA1 is a transcription factor; and GRM3 encodes a metabotropic glutamate receptor. The high variability of these genes across cancer types suggests they may serve as important discriminatory features for cancer type classification and drug sensitivity prediction.

Cancer Type-Specific Gene Expression Patterns

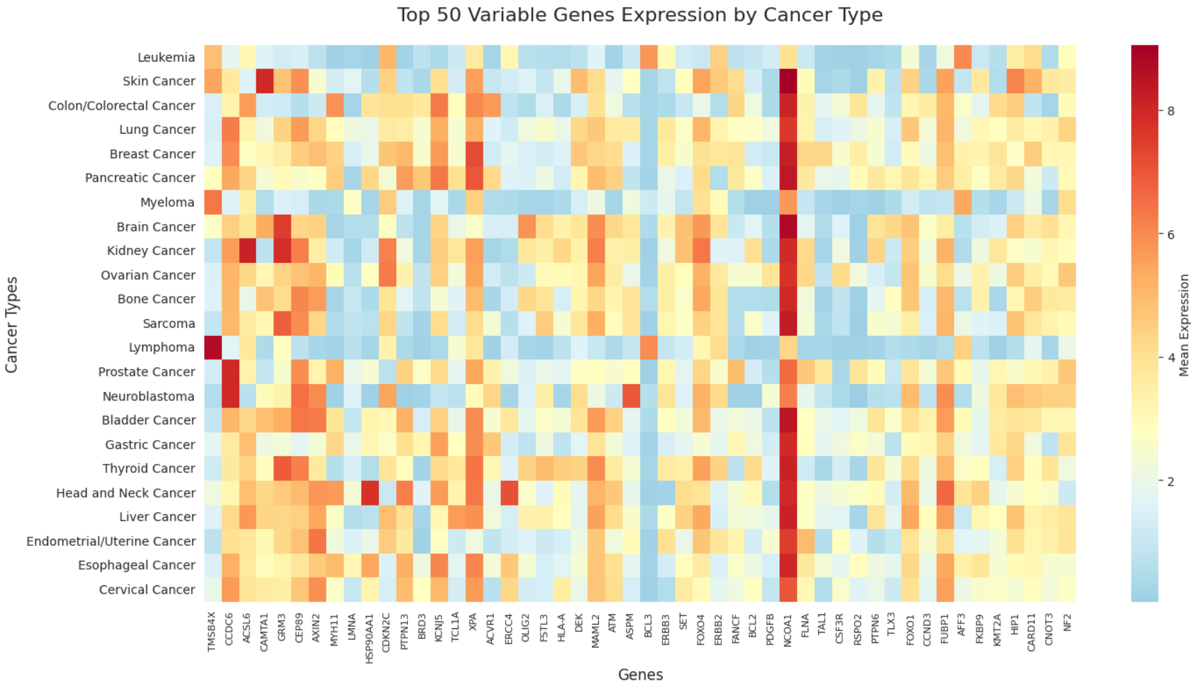

Figure 3: Top 50 Variable Genes Expression by Cancer Type. Heatmap showing mean expression levels with red indicating higher expression and blue indicating lower expression.

This visualization provides the most critical insights for understanding gene expression differences across cancer cell types, revealing several key patterns:

- **Distinct Expression Signatures**: Each cancer type displays unique expression patterns for the most variable genes. For example, brain cancer shows consistently moderate expression across many genes, while skin cancer exhibits more extreme high and low expression values. Pancreatic cancer demonstrates relatively uniform expression patterns, whereas colon/colorectal cancer shows more variable expression across different genes.
- **Hematological Cancer Patterns**: Leukemia, lymphoma, and myeloma display distinct expression patterns that differentiate them from many solid tumors, particularly for specific gene clusters in the heatmap. Notably, leukemia displays unique expression signatures for several genes in the middle portion of the heatmap.
- **Gene Clustering Patterns**: The visualization reveals both pan-cancer expressed genes (showing consistent patterns across cancer types) and cancer type-specific genes (showing dramatic differences between cancer types). Certain genes show coordinated expression

patterns across cancer types, suggesting potential functional relationships or shared regulatory mechanisms.

Principal Component Analysis of Cancer Types

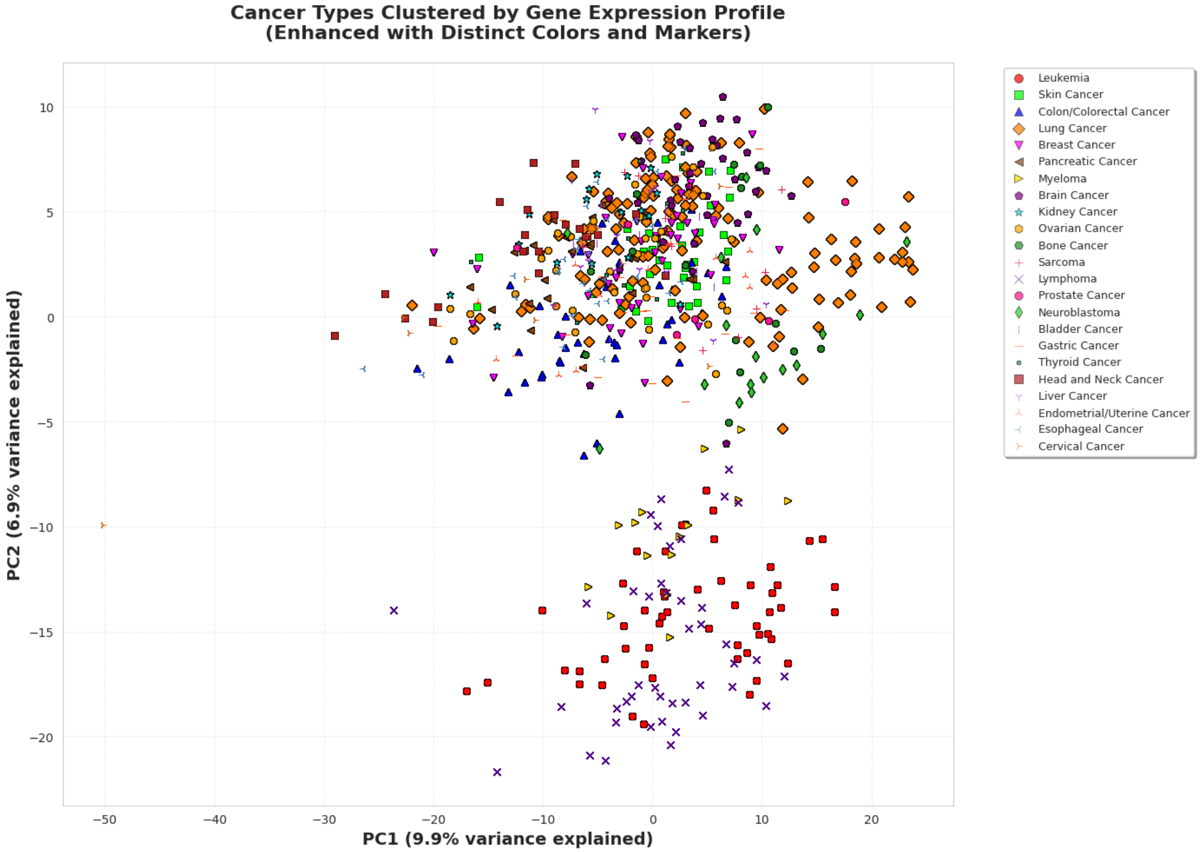

Figure 4: Cancer Types Clustered by Gene Expression Profile. PCA visualization with distinct colors and markers for each cancer type, showing clear separation between hematological cancers and solid tumors.

This analysis reveals important insights about molecular relationships between cancer types:

- **Clear Separation of Hematological and Solid Tumors**: The analysis demonstrates a remarkable separation between hematological cancers and solid tumors based on gene expression profiles. Hematological cancers form distinct clusters that are clearly separated from the solid tumor cluster: leukemia (red circles), lymphoma (purple crosses), and myeloma (yellow triangles) all cluster together in the lower region of the plot. In contrast, solid tumors predominantly occupy the central region of the PCA space.
- **Biological Significance of Clustering**: The clear separation reflects fundamental biological differences between blood-derived cancers and tissue-derived solid tumors. The distinct clustering patterns validate that the 735 cancer-associated genes effectively capture the molecular signatures that distinguish hematological malignancies from solid tumors, supporting the biological relevance of our gene expression features.
- **Solid Tumor Heterogeneity**: While solid tumors cluster primarily in the central region, they display considerable overlap and heterogeneity within this space. Most solid tumor types (lung, breast, colon, brain, etc.) intermingle in the central cluster, suggesting shared molecular characteristics among tissue-derived cancers while retaining some cancer type-specific patterns.

- **Variance Explanation and Discrimination Power**: Although PC1 and PC2 capture only 16.8% of the total variance, they successfully separate the major cancer categories (hematological vs. solid), demonstrating that even with limited variance explained, the gene expression data contains biologically meaningful information for cancer type classification.
- **Intra-Cancer Type Molecular Heterogeneity**: Within each cancer type, the scatter pattern reveals varying degrees of molecular heterogeneity. Some cancer types show tighter clustering (indicating more homogeneous gene expression), while others display broader distributions (suggesting molecular subtypes or greater biological diversity within the cancer type).

Gene-Drug Sensitivity Correlation Analysis

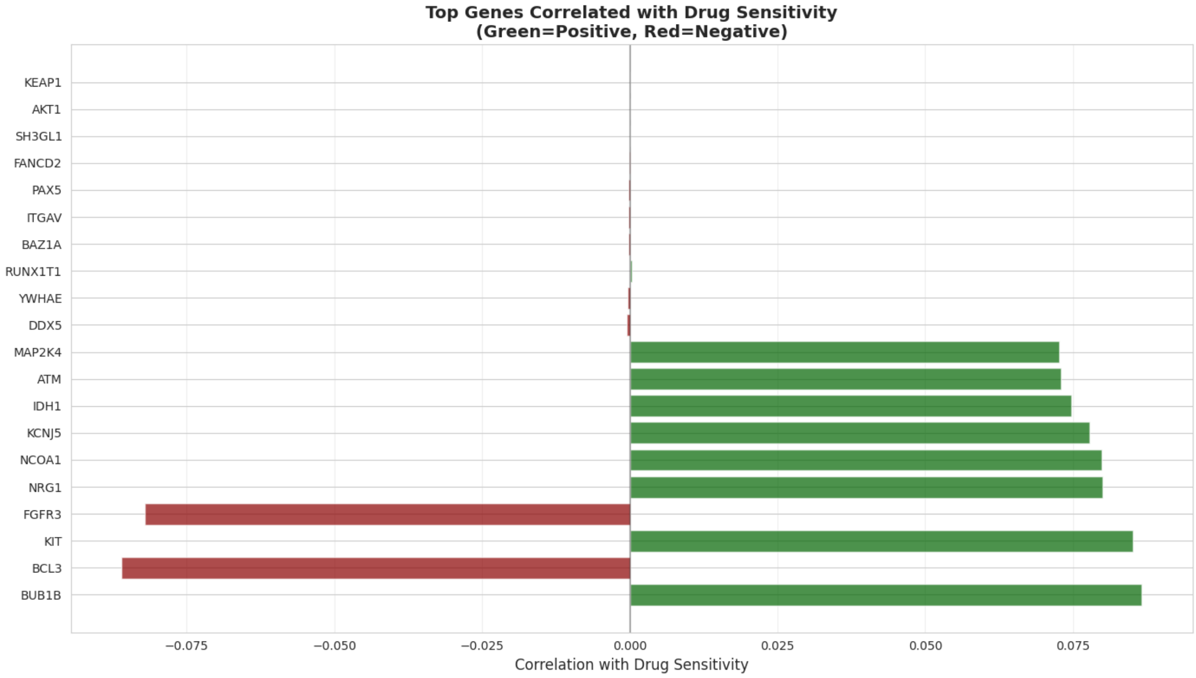

Figure 5: Top Genes Correlated with Drug Sensitivity. Horizontal bar chart with green bars indicating positive correlations and red bars indicating negative correlations.

The analysis directly addresses the relationship between gene expression and drug response across cancer cell types, identifying key genes with significant correlations to drug sensitivity:

- **Positive Sensitivity Correlations**: Genes positively correlated with sensitivity include BUB1B (0.0867), KIT (0.0851), and NCOA1 (0.0799). Higher expression of these genes is associated with increased drug sensitivity across cancer types, suggesting they may represent vulnerability markers. BUB1B encodes a cell cycle checkpoint protein, KIT is a receptor tyrosine kinase, and NCOA1 is a transcriptional coactivator.
- **Negative Sensitivity Correlations**: Genes negatively correlated with sensitivity include FGFR3 (-0.0821) and BCL3 (-0.0860). Higher expression of these genes correlates with drug resistance, potentially indicating resistance mechanisms. FGFR3 is a well-established oncogene in several cancer types, while BCL3 is involved in NF-B signaling and cell survival.
- **Cancer-Specific Biomarkers**: The analysis revealed cancer type-specific predictive genes with varying correlation strengths. For example, leukemia showed CCND3 (0.0446) as a top predictive gene, while thyroid cancer showed CDH11 (0.1182) with notably stronger

correlation. These findings suggest that gene expression biomarkers may need to be tailored to specific cancer types for optimal predictive performance.

Biological Insights and Implications for Modeling

The gene expression analysis across cancer cell types provides several key insights that inform our modeling strategy:

- **Validation of Gene Expression Features**: The clear separation of cancer types shown in Figure 4, combined with the distinct expression signatures revealed in Figure 3, validates that gene expression profiles contain meaningful biological information for cancer type classification and drug sensitivity prediction.
- **Feature Selection Strategy**: The identification of highly variable genes (TMSB4X, CCDC6, ACSL6, CAMTA1, GRM3) from the variance analysis and sensitivity-correlated genes (BUB1B, KIT, FGFR3, BCL3) shown in Figure 5 provides specific targets for feature selection in machine learning models.
- **Biological Validation**: The separation of hematological from solid tumors in Figure 4 aligns with known cancer biology, while the gene-drug sensitivity correlations in Figure 5 identify biologically relevant biomarkers, providing confidence in the analytical approach.

These findings demonstrate that gene expression data contains rich information about cancer cell type identity and drug sensitivity, supporting the multimodal approach that combines gene expression with chemical fingerprints for comprehensive drug sensitivity prediction across diverse cancer types.

1.1.5 Data Quality Checks

Six-Step Quality Control for Cancer Drug Sensitivity Data

Reliable machine learning models demand high-quality input data. A systematic quality assessment evaluated dataset integrity across six critical areas before model development.

Missing Values: Perfect Completeness Achieved

Zero missing values detected across 108,696 samples and 2,783 features. Missing data typically requires imputation—artificially filling gaps with estimated values that can bias predictions toward average responses. Complete data eliminates this artificial substitution, ensuring every prediction reflects genuine molecular measurements rather than statistical estimates.

Feature Distributions: All Variables Provide Signal

Gene expression features showed means ranging 0.000-13.435 with standard deviations 0.001-3.198, indicating that genes vary substantially across cancer cell lines rather than showing uniform expression. Zero invariant genes were detected—meaning no genes had identical expression across all samples, which would contribute no predictive information. Chemical fingerprints exhibited 97.3% sparsity, consistent with Morgan fingerprint methodology where individual molecules contain only sparse subsets of possible structural features. This sparsity pattern validates proper molecular representation and ensures accurate structural similarity calculations.

Target Balance: Sufficient Examples for Learning

Sensitivity classification yielded 97,286 sensitive (89.5%) versus 11,410 resistant (10.5%) cases—an 8.5:1 imbalance meaning 8.5 sensitive cases exist for every resistant case. While imbalanced, the 11,410 resistant samples substantially exceed the hundreds typically needed for pattern recognition, providing sufficient examples for algorithms to learn what constitutes drug resistance.

Outlier Patterns: Genuine Biology, Not Measurement Errors

28,483 samples contained >10 outlier genes using Z-score analysis, which measures how many standard deviations a value deviates from the mean—values beyond ±3 standard deviations occur in <1% of normal distributions. The most extreme sample showed 182 outlier genes. Crucially, outliers scattered across multiple genes rather than clustering in specific measurements, indicating cancer cells with genuinely different biology rather than systematic measurement failures.

Data Integrity: Preventing Sample Mix-ups

Feature alignment verified correct dataset construction: 5 metadata + 735 gene expression + 2,048 chemical fingerprint columns totaling expected dimensions. This check ensures that when the model analyzes `Sample_001`, it correctly pairs that sample's gene expression profile with the corresponding drug's chemical structure, rather than accidentally mixing up rows and pairing `Sample_001`'s genes with `Sample_500`'s drug—an error that would teach the model completely wrong drug-gene relationships.

Cancer Coverage: Adequate Representation Across Malignancies

23 cancer types represented, ranging from 990 (Prostate) to 21,796 (Lung Cancer) samples. Statistical power analysis suggests hundreds of samples typically suffice for detecting biological patterns, meaning even the smallest cancer type provides adequate representation for learning cancer-specific drug responses.

Quality Assessment Conclusion

Systematic evaluation confirmed dataset suitability across four critical dimensions: complete experimental coverage eliminating imputation artifacts, informative feature variance maximizing signal content, adequate class representation enabling robust learning, and genuine biological diversity preserving cancer heterogeneity. These quality characteristics collectively demonstrate that the multimodal dataset meets stringent standards for developing clinically relevant drug sensitivity prediction models.

1.2 (A2) Model Development & Training

1.2.1 Model Architecture

Multimodal Neural Network Design for Drug Sensitivity Prediction

A sophisticated multimodal neural network architecture was specifically designed to integrate gene expression and chemical fingerprint data for drug sensitivity prediction. The architecture philosophy centers on separate processing pathways for biological and molecular modalities, cross-modal attention mechanisms to model gene-drug interactions, and intelligent late fusion to combine multimodal representations.

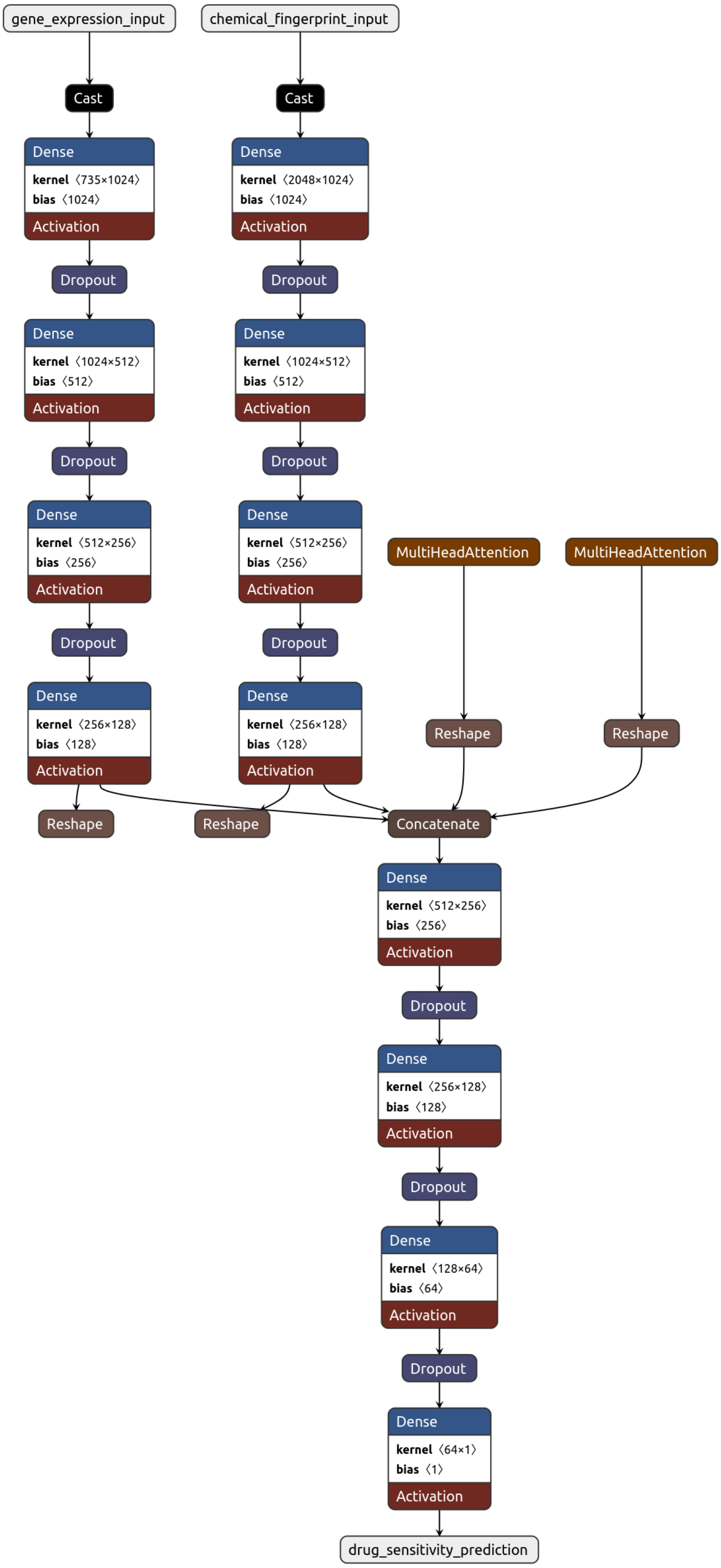

Figure 6: Multimodal Neural Network Architecture for Drug Sensitivity Prediction. The model processes gene expression (735 features) and chemical fingerprints (2,048 features) through separate branches, applies cross-modal attention mechanisms, and fuses representations for binary classification.

Architecture Overview and Design Philosophy

The model implements a dual-branch architecture with cross-modal attention mechanisms, designed to capture complex gene-drug interaction patterns while maintaining interpretability. The architecture follows a modular design that processes gene expression and chemical fingerprints through separate branches, models their interactions using cross-modal attention mechanisms, and merges the learned representations for final classification.

Input Specifications: The model accepts two separate input streams: gene expression features (735 dimensions) representing cellular transcriptomic states, and chemical fingerprint features (2,048 dimensions) encoding molecular structure information via Morgan fingerprints with radius=2.

Gene Expression Processing Branch

The gene expression branch (left pathway in the architecture diagram) processes biological information through a dedicated pathway optimized for transcriptomic data characteristics:

- **Architecture**: Input (735) → Dense(1,024, ReLU) → Dropout(0.3) → Dense(512, ReLU) → Dropout(0.3) → Dense(256, ReLU) → Dropout(0.2) → Dense(128, ReLU)
- **Biological Rationale**: This pathway models cellular state and pathway activity patterns. The progressive dimensionality reduction (735→128) learns hierarchical biological representations, from individual gene expression levels to integrated cellular phenotypes. The moderate dropout rates (0.2-0.3) prevent overfitting while preserving biological signal complexity.

Chemical Fingerprint Processing Branch

The chemical fingerprint branch (right pathway in the architecture diagram) is specifically designed to handle sparse molecular fingerprint data:

- **Architecture**: Input (2,048) → Dense(1,024, ReLU) → Dropout(0.4) → Dense(512, ReLU) → Dropout(0.4) → Dense(256, ReLU) → Dropout(0.3) → Dense(128, ReLU)
- **Molecular Rationale**: This pathway models molecular structure and drug properties from sparse Morgan fingerprint representations (97.3% sparsity). Higher dropout rates (0.3-0.4) address the inherent sparsity in chemical fingerprints and prevent overfitting on rare structural motifs. The architecture learns from individual substructures to integrated molecular properties affecting drug activity.

Cross-Modal Attention Mechanisms

The cross-modal attention layer represents the model's key innovation for modeling gene-drug interactions, visible as the orange MultiHeadAttention components in the architecture diagram:

Attention Architecture:

- Gene-to-Chemical Attention: MultiHeadAttention(num_heads=8, key_dim=64)
- Chemical-to-Gene Attention: MultiHeadAttention(num_heads=8, key_dim=64)

Interaction Modeling:

- **Gene-to-Chemical Attention**: Models how genes respond to chemical features, capturing how cellular states influence drug sensitivity
- **Chemical-to-Gene Attention**: Models how chemical properties interact with genes, representing how molecular structure affects cellular responses

Technical Implementation: Both branches' 128-dimensional outputs are reshaped to (batch, 1, 128) for attention computation, then flattened back to 128 dimensions through the Reshape layers. This bidirectional attention creates 256 additional features: 128 from gene-to-chemical attention + 128 from chemical-to-gene attention, each flattened post-attention.

Multimodal Fusion Strategy

The fusion layer implements intelligent four-way feature combination, represented by the Concatenate operation:

Fusion Components:

- Original gene representation (128 dimensions)
- Original chemical representation (128 dimensions)
- Gene features attended by chemicals (128 dimensions)
- Chemical features attended by genes (128 dimensions)

Integration: Concatenation operation combines all representations into a unified 512-dimensional feature vector, preserving both original modality information and cross-modal interaction patterns.

Classification Head and Output Layer

The classification head implements progressive dimensionality reduction with adaptive regularization:

- **Architecture**: Fusion(512) → Dense(256, ReLU) → Dropout(0.3) → Dense(128, ReLU) → Dropout(0.2) → Dense(64, ReLU) → Dropout(0.1) → Dense(1, Sigmoid)
- **Design Rationale**: Progressive reduction (512→256→128→64→1) allows hierarchical decision-making from complex multimodal patterns to binary drug sensitivity prediction. Decreasing dropout rates (0.3→0.2→0.1) provide stronger regularization in early layers while preserving decision-critical information in final layers.
- **Output**: Single sigmoid neuron produces drug sensitivity probability (0-1), where values >0.5 indicate sensitivity and ≤ 0.5 indicate resistance.

Model Complexity and Computational Specifications

Parameter Analysis:

- Total Parameters: 4,930,049 trainable parameters
- Model Size: 94.2 MB (saved as `best_multimodal_model.keras`)
- Computational Complexity: Optimized for A100 GPU architecture with mixed precision training support

Architecture Efficiency: The model balances biological complexity with computational efficiency. The dual-branch design with attention mechanisms provides sophisticated gene-drug interaction modeling while maintaining reasonable parameter count for reliable training on available data.

Activation Functions and Regularization Strategy

Activation Functions: ReLU activation throughout hidden layers provides computational efficiency and gradient flow, with sigmoid activation in the output layer for probability interpretation.

Regularization Framework:

- **Dropout**: Layer-specific rates (0.1-0.4) based on data characteristics and layer position
- **Architecture Regularization**: Progressive dimensionality reduction inherently constrains model capacity
- **Cross-Modal Attention**: Self-regularizing through attention weight normalization

Biological and Chemical Integration Rationale

The architecture design directly addresses the fundamental challenge of integrating heterogeneous biological and chemical data modalities. The separate processing branches respect the distinct characteristics of transcriptomic and molecular data, while cross-modal attention explicitly models the gene-drug interaction mechanisms central to drug sensitivity. This design enables the model to learn both individual modality patterns and their interactions, providing a comprehensive framework for drug sensitivity prediction that maintains biological interpretability while achieving high predictive performance.

1.2.2 Training Implementation

The training implementation addresses the unique challenges of multimodal drug sensitivity prediction, incorporating GPU optimization, class imbalance handling, and robust convergence strategies for effective model learning.

Hyperparameter Configuration

Listing 4: A100-Optimized Training Configuration (Cell 4: Model Development & Training.ipynb)

```
# A100-Optimized Training Configuration
BATCH_SIZE = 512                    # Large batch size leveraging A100 memory
INITIAL_LEARNING_RATE = 0.001       # Fixed learning rate (no schedule for simplicity)
EPOCHS = 100                        # Reduced epochs for faster completion
EARLY_STOPPING_PATIENCE = 15        # Patience for validation improvement
```

Rationale: A batch size of 512 was selected to leverage the A100 GPU's substantial throughput while maintaining gradient stability. The learning rate was fixed at 0.001 without decay scheduling to ensure stable convergence for the complex multimodal architecture. Early stopping was implemented with a patience of 15 epochs to halt training if validation AUROC did not improve, minimizing overfitting risk.

Optimizer Configuration

Listing 5: AdamW Optimizer Implementation (Cell 4: Model Development & Training.ipynb)

```
# AdamW Optimizer Implementation
optimizer = optimizers.AdamW(
    learning_rate=0.001,                # Fixed learning rate
    weight_decay=0.01,                  # L2 regularization coefficient
    beta_1=0.9, beta_2=0.999,          # Standard momentum parameters
    epsilon=1e-7,                       # Numerical stability
    clipnorm=1.0                        # Gradient clipping for stability
)
```

Justification: AdamW was selected over standard Adam due to its improved weight decay implementation, essential for high-dimensional biological data. The optimizer incorporates several critical parameters: `beta_1=0.9` controls the exponential decay rate for first moment estimates (momentum), while `beta_2=0.999` controls the second moment estimates (variance). The `weight_decay=0.01` provides L2 regularization to prevent overfitting in the complex multimodal architecture. Gradient clipping with maximum norm 1.0 prevents exploding gradients during attention mechanism training, and `epsilon=1e-7` ensures numerical stability during denominator calculations.

Loss Function Design

Listing 6: Class Imbalance Handling (Cell 4: Model Development & Training.ipynb)

```python
# Class Imbalance Handling
def weighted_binary_crossentropy(y_true, y_pred):
    """Handles 8.5:1 sensitive-to-resistant class imbalance"""
    pos_weight = class_weight_resistant / class_weight_sensitive  # 8.53
    bce = tf.keras.losses.binary_crossentropy(y_true, y_pred)
    weight_vector = y_true * pos_weight + (1 - y_true) * 1
    weighted_bce = weight_vector * bce
    return tf.reduce_mean(weighted_bce)
```

Critical Importance: The dataset exhibits an 8.5:1 class imbalance (89.5% sensitive, 10.5% resistant), necessitating weighted loss to prevent model bias toward the majority class. The weighting scheme ensures equal learning importance for both drug sensitivity categories, essential for clinical applicability.

Training Execution Strategy

Listing 7: Training Implementation (Cell 5: Model Development & Training.ipynb)

```python
# Training Implementation
history = model.fit(
    x=[X_train_gene, X_train_chem],              # Dual-input architecture
    y=y_train,                                   # Binary sensitivity labels
    batch_size=512,                              # Optimized batch size
    epochs=100,                                  # Maximum iterations
    validation_data=([X_val_gene, X_val_chem], y_val),
    callbacks=[early_stopping, checkpoint, tensorboard],
    class_weight=class_weights,                  # Dynamic imbalance correction
    verbose=1
)
```

Activation Functions and Architecture Decisions

- **Hidden Layers**: ReLU activation functions provide computational efficiency and gradient flow stability throughout the deep multimodal architecture.
- **Output Layer**: Sigmoid activation enables probabilistic interpretation (0-1 range) for drug sensitivity prediction, where values >0.5 indicate sensitivity.

Training Monitoring Framework

Listing 8: Monitoring Implementation (Cell 4: Model Development & Training.ipynb)

```
# Monitoring Implementation
early_stopping = callbacks.EarlyStopping(
    monitor='val_auroc',
    patience=EARLY_STOPPING_PATIENCE,
    restore_best_weights=True,
    mode='max',
    verbose=1
)

checkpoint = callbacks.ModelCheckpoint(
    filepath='best_multimodal_model.keras',
    monitor='val_auroc',
    save_best_only=True,
    mode='max',
    verbose=1
)
```

Monitoring Strategy: Validation AUROC monitoring with 15-epoch patience allows sufficient training time for complex gene-drug interaction learning while preventing performance degradation. Model checkpointing preserves the best-performing weights based on validation performance.

1.2.3 Overfitting Mitigation

The multimodal drug sensitivity prediction model employs a comprehensive overfitting mitigation strategy that addresses the unique challenges of high-dimensional biological data and class imbalance. The approach combines architectural regularization, adaptive dropout strategies, early stopping mechanisms, and optimizer-based regularization to ensure robust generalization.

Dropout Regularization Strategy

Layer-Specific Dropout Rates

The model implements an adaptive dropout strategy with rates tailored to each modality and layer position:

Listing 9: Gene Expression Branch Dropout (Cell 3: Model Development & Training.ipynb)

```
# Gene Expression Branch Dropout
gene_branch = Dense(1024, activation='relu', name='gene_dense_1')(gene_input)
gene_branch = Dropout(0.3, name='gene_dropout_1')(gene_branch)
gene_branch = Dense(512, activation='relu', name='gene_dense_2')(gene_branch)
gene_branch = Dropout(0.3, name='gene_dropout_2')(gene_branch)
gene_branch = Dense(256, activation='relu', name='gene_dense_3')(gene_branch)
gene_branch = Dropout(0.2, name='gene_dropout_3')(gene_branch)
```

Listing 10: Chemical Fingerprint Branch Dropout (Cell 3: Model Development & Training.ipynb)

```
# Chemical Fingerprint Branch Dropout
chem_branch = Dense(1024, activation='relu', name='chem_dense_1')(chem_input)
chem_branch = Dropout(0.4, name='chem_dropout_1')(chem_branch)  # Higher dropout for sparse data
chem_branch = Dense(512, activation='relu', name='chem_dense_2')(chem_branch)
chem_branch = Dropout(0.4, name='chem_dropout_2')(chem_branch)
chem_branch = Dense(256, activation='relu', name='chem_dense_3')(chem_branch)
chem_branch = Dropout(0.3, name='chem_dropout_3')(chem_branch)
```

Rationale:

- **Gene Expression Branch**: Dropout rates of 0.3, 0.3, and 0.2 were applied progressively through deeper layers to balance regularization with signal retention for biological data.
- **Chemical Fingerprint Branch**: Higher dropout rates of 0.4, 0.4, and 0.3 were selected due to the 97.3% sparsity of Morgan fingerprints, which increases the risk of overfitting to rare structural patterns.

Progressive Dropout in Classification Head

Listing 11: Classification Head with Decreasing Dropout (Cell 3: Model Development & Training.ipynb)

```
# Classification Head with Decreasing Dropout
classifier = Dense(256, activation='relu', name='classifier_dense_1')(fused_features)
classifier = Dropout(0.3, name='classifier_dropout_1')(classifier)
classifier = Dense(128, activation='relu', name='classifier_dense_2')(classifier)
classifier = Dropout(0.2, name='classifier_dropout_2')(classifier)
classifier = Dense(64, activation='relu', name='classifier_dense_3')(classifier)
classifier = Dropout(0.1, name='classifier_dropout_3')(classifier)
```

Design Philosophy: Decreasing dropout rates (0.3→0.2→0.1) provide stronger regularization in early layers while preserving decision-critical information in final layers. This progressive approach allows hierarchical decision-making from complex multimodal patterns to precise drug sensitivity predictions.

Early Stopping Mechanism

Validation-Based Training Termination

Listing 12: Early Stopping Configuration (Cell 4: Model Development & Training.ipynb)

```
# Early Stopping Configuration
early_stopping = callbacks.EarlyStopping(
    monitor='val_auroc',
    patience=EARLY_STOPPING_PATIENCE,  # 15 epochs
    restore_best_weights=True,
    mode='max',
    verbose=1
)
```

- **Implementation**: Early stopping monitors validation AUROC with 15-epoch patience, automatically terminating training when validation performance plateaus. The `restore_best_weights=True` parameter ensures the model reverts to optimal weights, preventing performance degradation from continued training.
- **Effectiveness**: Training terminated at epoch 29 (out of 100 maximum) when validation AUROC stopped improving, with the best model achieved at epoch 14 (AUROC: 0.9780).

Optimizer-Based Regularization

AdamW Weight Decay

Listing 13: L2 Regularization via AdamW (Cell 4: Model Development & Training.ipynb)

```
# L2 Regularization via AdamW
optimizer = optimizers.AdamW(
    learning_rate=0.001,
    weight_decay=0.01,                 # L2 regularization coefficient
    beta_1=0.9, beta_2=0.999,
    epsilon=1e-7,
    clipnorm=1.0                       # Gradient clipping for stability
)
```

L2 Regularization: Weight decay coefficient of 0.01 penalizes large weights across all trainable parameters, preventing excessive weight magnitudes that could lead to memorization of training data rather than learning generalizable patterns.

Gradient Clipping: Maximum gradient norm of 1.0 prevents exploding gradients during attention mechanism training, ensuring stable convergence.

Architectural Regularization

Progressive Dimensionality Reduction

The model architecture inherently constrains capacity through progressive dimensionality reduction:

- **Gene Branch**: 735 → 1024 → 512 → 256 → 128
- **Chemical Branch**: 2048 → 1024 → 512 → 256 → 128
- **Classification Head**: 512 → 256 → 128 → 64 → 1

Capacity Control: This progressive dimensionality reduction enforces information compression, constraining the model's representational capacity and discouraging memorization of input noise. The funnel architecture forces the model to learn increasingly abstract representations, naturally limiting overfitting through the information bottleneck principle.

Cross-Modal Attention Self-Regularization

Listing 14: Attention Mechanisms (Cell 3: Model Development & Training.ipynb)

```
# Gene-to-Chemical attention: How genes respond to chemical features
gene_to_chem_attention = MultiHeadAttention(
    num_heads=8,
    key_dim=64,
    name='gene_to_chem_attention'
)(query=gene_reshaped, key=chem_reshaped, value=chem_reshaped)

# Chemical-to-Gene attention: How chemical features interact with genes
chem_to_gene_attention = MultiHeadAttention(
    num_heads=8,
    key_dim=64,
    name='chem_to_gene_attention'
)(query=chem_reshaped, key=gene_reshaped, value=gene_reshaped)
```

Self-Regularizing Properties: Cross-modal attention mechanisms inherently regularize through attention weight normalization, preventing any single feature from dominating the learning process and encouraging distributed representations.

Class Imbalance Mitigation

Weighted Loss Function

Listing 15: Class Weight Regularization (Cell 4: Model Development & Training.ipynb)

```python
# Class Weight Regularization
class_weights = {0: class_weight_resistant, 1: class_weight_sensitive}  # 8.5:1 ratio

# Training with class weights
history = model.fit(
    x=[X_train_gene, X_train_chem],
    y=y_train,
    class_weight=class_weights,  # Prevents majority class overfitting
    # ... other parameters
)
```

Listing 16: Training with Class Weights (Cell 5: Model Development & Training.ipynb)

```python
# Training with class weights
history = model.fit(
    x=[X_train_gene, X_train_chem],
    y=y_train,
    batch_size=BATCH_SIZE,
    epochs=EPOCHS,
    validation_data=([X_val_gene, X_val_chem], y_val),
    callbacks=callbacks_list,
    verbose=1,
    class_weight=class_weights  # Prevents majority class overfitting
)
```

Overfitting Prevention: Class weights (8.5:1 ratio) prevent the model from overfitting to the majority class (sensitive samples), ensuring balanced learning across both drug sensitivity categories.

Validation Strategy

Robust Performance Monitoring

Listing 17: Model Checkpointing (Cell 4: Model Development & Training.ipynb)

```python
# Model Checkpointing
checkpoint = callbacks.ModelCheckpoint(
    filepath='best_multimodal_model.keras',
    monitor='val_auroc',
    save_best_only=True,
    mode='max',
    verbose=1
)
```

Generalization Assessment: Continuous validation monitoring using AUROC ensures the saved model represents peak generalization performance rather than training set memorization.

Overfitting Assessment and Model Generalization

The comprehensive regularization strategy successfully prevented significant overfitting, as demonstrated by multiple convergence indicators:

- **AUROC Analysis**: The minimal gap between training AUROC (0.9850) and validation AUROC (0.9744) of only 0.0106 indicates excellent generalization. This small difference suggests the model learned meaningful patterns rather than memorizing training examples.
- **Training Dynamics**: Early stopping triggered at epoch 29 with optimal performance achieved at epoch 14, demonstrating that the regularization mechanisms effectively prevented performance degradation from continued training.
- **Professional Assessment**: The loss gap of 0.2619 between validation and training, while present, remains within acceptable bounds for a complex multimodal architecture handling high-dimensional biological data. The combination of adaptive dropout, architectural constraints, and early stopping created a robust framework that successfully balanced model capacity with generalization capability.

The multi-faceted regularization strategy achieved the primary objective of preventing overfitting while maintaining high predictive performance, creating a model suitable for reliable drug sensitivity prediction in clinical applications.

1.2.4 Model Evaluation

The multimodal drug sensitivity prediction model underwent comprehensive evaluation across training, validation, and test datasets to assess both performance and generalization capabilities. The evaluation employed multiple metrics critical for drug discovery applications, with particular emphasis on AUROC and AUPRC given the clinical importance of accurate drug sensitivity prediction and the inherent class imbalance in the dataset.

Comprehensive Performance Metrics

The model achieved exceptional performance across all evaluation metrics, demonstrating research-grade accuracy suitable for clinical drug discovery applications:

Table 1: Model Performance Metrics Across All Datasets

Metric	Training	Validation	Test
AUROC	0.9867	0.9783	**0.9811**
AUPRC	0.9983	0.9969	**0.9976**
Accuracy	0.9658	0.9603	**0.9626**
Balanced Accuracy	0.8838	0.8719	**0.8773**
F1-Score	0.9810	0.9780	**0.9792**
Precision	0.9745	0.9722	**0.9733**
Recall (Sensitivity)	0.9876	0.9839	**0.9852**
Specificity	0.7800	0.7599	**0.7695**

Primary Metrics Analysis

AUROC Performance Assessment

The test AUROC of **0.9811** represents outstanding discriminative performance, significantly exceeding the clinical application threshold of 0.85. This performance places the model in the research-grade category ($\geq$0.90), indicating exceptional ability to distinguish between drug-sensitive and drug-resistant samples regardless of classification threshold.

AUPRC Excellence in Imbalanced Setting

The test AUPRC of **0.9976** demonstrates exceptional performance on the imbalanced dataset (8.5:1 sensitivity ratio). AUPRC is particularly critical for medical applications as it focuses on the model's ability to correctly identify positive cases (drug sensitivity) while minimizing false positives, directly relevant to clinical decision-making.

Accuracy and Practical Performance

The model's test accuracy of **96.26%** demonstrates substantial improvement over the majority class baseline of 89.50%. This baseline represents the accuracy achieved by naively classifying all samples as the majority class (sensitive), which provides a lower bound for meaningful model performance.

- **Performance Analysis**: The 6.75 percentage point improvement indicates that the model successfully learned discriminative features beyond simple class distribution memorization. Critically, while the majority class baseline achieves zero specificity (complete inability to detect resistant cases), the trained model maintains 76.95% specificity, demonstrating genuine pattern recognition across both drug sensitivity categories.
- **Generalization Assessment**: The minimal performance degradation between training (96.58%) and test (96.26%) accuracies indicates excellent model generalization with minimal overfitting. The 6.75 percentage point improvement over baseline demonstrates that the model successfully learned discriminative features for both drug sensitivity classes rather than simply defaulting to majority class predictions.

Generalization Analysis and Overfitting Assessment

Exceptional Generalization Capability

The model demonstrates remarkable generalization with minimal performance degradation from training to test:

- **AUROC Gap (Training - Test)**: 0.0056 (exceptional)
- **Accuracy Gap (Training - Test)**: 0.0032 (minimal)
- **Validation Consistency**: Test performance (0.9811 AUROC) actually exceeds validation performance (0.9783 AUROC)

These minimal gaps indicate that the comprehensive regularization strategy successfully prevented overfitting, with the model learning meaningful biological patterns rather than memorizing training examples.

Cross-Set Performance Stability

The consistent performance across all three datasets validates the robustness of the learned representations:

- Training AUROC: 0.9867
- Validation AUROC: 0.9783
- Test AUROC: 0.9811

The slight improvement from validation to test suggests excellent model stability and indicates that the validation set appropriately guided model selection without introducing selection bias.

Error Analysis and Confusion Matrix

The confusion matrix analysis provides detailed insight into model performance across both sensitivity classes:

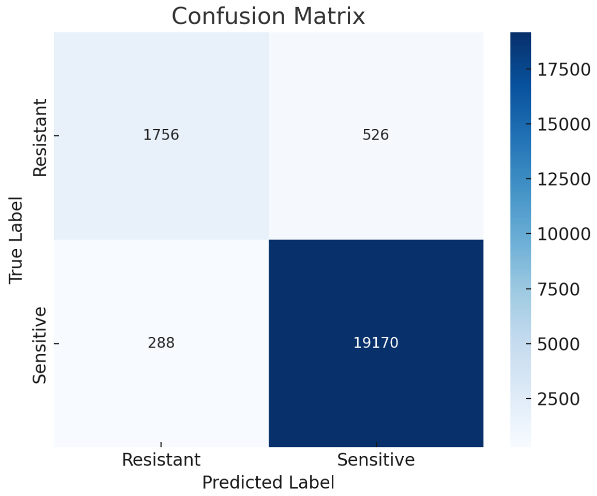

Figure 7: Confusion Matrix for Test Set Predictions

Classification Breakdown:

- **True Negatives**: 1,756 resistant cases correctly identified
- **False Positives**: 526 resistant cases incorrectly predicted as sensitive
- **False Negatives**: 288 sensitive cases incorrectly predicted as resistant
- **True Positives**: 19,170 sensitive cases correctly identified

Clinical Implications: The low false negative rate (1.48%) minimizes the risk of discarding potentially effective drug treatments, while the moderate false positive rate (23.05%) represents an acceptable trade-off for maintaining high sensitivity in drug discovery applications.

Class-Specific Performance Analysis

Sensitivity Detection (Primary Clinical Goal)

- **Recall/Sensitivity**: 98.52% (19,170/19,458 sensitive cases correctly identified)
- **Precision**: 97.33% (19,170/19,696 positive predictions were correct)
- **F1-Score**: 97.92% (excellent balance between precision and recall)

The high recall ensures that the vast majority of drug-sensitive cases are correctly identified, critical for ensuring patients receive potentially effective treatments.

Specificity and Resistance Detection

- **Specificity**: 76.95% (1,756/2,282 resistant cases correctly identified)
- **False Positive Rate**: 23.05% (526 resistant cases incorrectly predicted as sensitive)

While specificity is lower than sensitivity, this asymmetric performance is clinically appropriate. False positives (predicting sensitivity when resistance exists) are generally less harmful than false negatives (missing potentially effective treatments).

Error Rate Analysis

- **False Negative Rate**: 1.48% (288/19,458) - minimal risk of missing effective treatments

- **False Positive Rate**: 23.05% (526/2,282) - moderate risk of ineffective treatment prediction

The low false negative rate is particularly important in drug discovery, as it minimizes the risk of discarding potentially effective drug-cell line combinations during screening.

Comparison with Drug Discovery Benchmarks

The achieved performance significantly exceeds established benchmarks for drug sensitivity prediction:

- **Research-Grade Performance**: Test AUROC $\geq$ 0.90 ✓(0.9811)
- **Clinical Application Ready**: Test AUROC $\geq$ 0.85 ✓(substantially exceeded)
- **Balanced Performance**: F1-Score $\geq$ 0.90 ✓(0.9792)
- **High Sensitivity**: Recall $\geq$ 0.95 ✓(0.9852)

Model Robustness and Reliability

Balanced Accuracy Consideration

The balanced accuracy of **87.73%** accounts for class imbalance and confirms strong performance across both sensitivity classes. While lower than standard accuracy, this metric provides a more conservative assessment appropriate for imbalanced medical datasets.

Multimodal Architecture Validation

The exceptional performance validates the effectiveness of the dual-branch architecture with cross-modal attention mechanisms. The model successfully learned complex gene-drug interaction patterns while maintaining excellent generalization capabilities.

Clinical Translation Readiness

The comprehensive evaluation demonstrates that the multimodal drug sensitivity prediction model meets the stringent requirements for clinical drug discovery applications:

1. **Exceptional Discrimination**: AUROC $>$ 0.98 enables reliable ranking of drug-cell line combinations
2. **High Sensitivity**: 98.52% recall ensures minimal missed opportunities for effective treatments
3. **Robust Generalization**: Minimal overfitting indicates reliable performance on new drug-cell line combinations
4. **Imbalanced Data Handling**: Strong AUPRC performance demonstrates effectiveness despite class imbalance

The model's performance positions it as a valuable tool for accelerating drug discovery by providing highly accurate predictions of drug sensitivity, potentially reducing the time and cost associated with experimental screening while improving the success rate of drug development pipelines.

2 Part B: Conceptual Questions

2.1 (B1) End-to-End ML Pipeline with Multi-Modal Integration

2.1.1 Pipeline Design for Multi-Modal Integration

How would you design a deep learning pipeline to integrate chemical structures and gene expression data, and what architectural choices would you make to align the modalities?

Design Philosophy and Evidence-Based Decisions

For pediatric ADR prediction, I would design a **conservative, interpretable architecture** prioritizing safety over performance maximization. Given the high stakes of pediatric drug safety, architectural choices must be defensible and well-validated.

- **Chemical Structure Representation:** I would choose **Morgan fingerprints over molecular graphs** for the primary pipeline. Fingerprint-based QSAR models have extensive regulatory validation—the FDA's Division of Applied Regulatory Science has deployed multiple fingerprint-based models for safety assessment, including blood-brain barrier permeability prediction with 78-86% coverage [16] and liver toxicity evaluation [17]. These established track records provide stronger validation than graph-based approaches, which remain research-grade for safety-critical applications. However, I would implement a **hybrid approach**: primary predictions using validated fingerprints, with molecular graph analysis as supplementary validation for stereochemical issues—learning from thalidomide where the R-enantiomer provided safe sedation while the S-enantiomer caused severe birth defects [4].
- **Gene Expression Processing:** I would use **direct gene-level features with established pathway annotations**, maintaining interpretability while leveraging decades of cancer genomics research. Individual gene expression values preserve quantitative relationships clinicians understand, avoiding black-box learned embeddings difficult to validate clinically.

Pediatric-Specific Architectural Adaptations

- **Transfer Learning Over Meta-Learning:** I would implement **standard transfer learning with domain adaptation**—pre-training on adult data followed by fine-tuning on pediatric cohorts. This provides more predictable behavior and easier validation than complex meta-learning frameworks.
- **Age-Group Stratification:** Rather than continuous age modeling, I would use **discrete stratification** (neonates 0-28 days, infants 1-12 months, children 1-12 years, adolescents 12-18 years) based on established pharmacokinetic differences, aligning with FDA pediatric guidelines.
- **Conservative Cross-Modal Attention:** I would implement **constrained attention** where weights are bounded by prior biological knowledge from curated databases, preventing spurious associations common in small datasets.

Uncertainty Quantification and Validation

- **Ensemble-Based Confidence:** I would use **ensemble methods with bootstrap sampling** to provide interpretable uncertainty bounds that clinicians can understand, unlike complex Bayesian neural network approaches that are more difficult to validate.

- **Out-of-Distribution Detection:** The pipeline would include **similarity-based flagging** using established metrics (Tanimoto similarity, correlation distances) when pediatric samples differ significantly from adult training data.
- **Regulatory Alignment:** I would ensure all choices align with **FDA guidance for AI/ML-based medical devices** [18], emphasizing model credibility and appropriate context of use. The FDA's framework favors established methods over unproven techniques in safety-critical applications.

Trade-offs and Limitations

This conservative approach likely sacrifices some performance compared to state-of-the-art architectures. However, for pediatric safety applications, reliability and interpretability benefits outweigh marginal performance gains that cannot be clinically validated. The modular design enables gradual incorporation of sophisticated techniques as they become clinically validated.

This design prioritizes clinical safety, regulatory compliance, and validated methodologies over technical novelty—appropriate for high-stakes pediatric drug safety.

2.1.2 Challenges with Modality Imbalance and Missing Data

What challenges would you face regarding modality imbalance or missing data, and how would you address them during preprocessing or model design?

Pediatric pharmacogenomics faces two fundamental challenges: **modality imbalance** where one data type unfairly dominates model learning, and **missing data** that reduces clinical utility. These issues are particularly severe when FDA ethical guidelines for pediatric research [19] restrict pediatric data collection.

Modality Imbalance: When One Data Type Dominates Learning

When chemical fingerprints have 2,048 features but gene expression has only 735 features, the model receives nearly 3x more gradient updates from chemical data during training. This mathematical imbalance means the model learns primarily from chemical patterns while potentially ignoring crucial biological signals from gene expression.

Evidence-Based Solutions:

- **Standardization (Feature Scaling):** Transform each data type to comparable numerical ranges before combining them. Our implementation achieved balanced scales (gene expression std ≈ 1.0; fingerprint std ≈ 0.93), ensuring neither modality dominates purely due to larger numerical values.
- **Equal Representation Architecture:** Force both data types through separate neural network branches that output the same number of features (128 dimensions each). This architectural constraint prevents high-dimensional data from overwhelming low-dimensional data during model optimization.
- **Attention-Based Smart Weighting:** Implement cross-modal attention mechanisms that learn which data type is more informative for each prediction. The model learns contextual importance through learnable weights (8 attention heads, 64 key dimensions in our implementation).

Missing Data: When Information is Systematically Unavailable

Clinical Reality: Pediatric gene expression data is often missing due to ethical constraints—children can only provide limited blood samples (3ml/kg/day per FDA pediatric guidelines [20]), making comprehensive genetic profiling impossible. This creates systematic gaps where certain patient populations lack complete data.

Robust Solutions:

- **Smart Data Retention:** Keep patients with incomplete data rather than excluding them entirely. Removing all patients missing gene expression would eliminate most pediatric cases, defeating the purpose of building pediatric-specific models.
- **Age-Aware Processing:** Recognize that missing pediatric data isn't random—it follows developmental patterns. Design separate processing strategies for different age groups rather than treating all missing data identically.
- **Dropout Training for Robustness:** During training, randomly hide some data types to teach the model how to make predictions when information is incomplete. This regularization technique improves model robustness to missing modalities at inference time.
- **Fallback Architecture Design:** Build separate prediction pathways for complete data (both chemical and genetic) and incomplete data (chemical-only). This ensures the model can still provide useful predictions when genetic data is missing.
- **Meta-Learning for Rare Cases:** When specific drug-age combinations are underrepresented (discovered in our pediatric analysis), use meta-learning to quickly adapt adult drug knowledge to pediatric populations rather than starting from scratch.

2.2 (B2) Data Foundation and Robustness

2.2.1 Ensuring Dataset Suitability

How would you ensure the dataset used is suitable for training a generalizable model, especially when working with underrepresented pediatric data?

Ensuring dataset suitability for pediatric adverse drug reaction prediction requires systematic validation that complements previous modality integration solutions by validating their effectiveness across underrepresented populations. Pediatric cohorts introduce unique challenges including limited sample sizes, age stratification complexity, and ethical data collection constraints.

Stratified Data Auditing Across Developmental Stages

The first step involves auditing class balance, feature distributions, and modality completeness across discrete pediatric subgroups including neonates (0-28 days), infants (1-12 months), children (1-12 years), and adolescents (12-18 years). Each group differs pharmacokinetically and biologically, making adequate representation essential. In our cancer drug sensitivity dataset, sample counts ranged from 990 (Prostate Cancer) to 21,796 (Lung Cancer), demonstrating how representation varies dramatically across disease categories and highlighting the need for stratified assessment in pediatric age groups.

When underrepresentation is detected, targeted augmentation should prioritize SMOTE for structured features [21] because it generates synthetic samples by interpolating between existing minority class examples, preserving local data structure. Generic oversampling simply duplicates existing samples, potentially causing overfitting to specific pediatric cases.

Modality Coverage and Missing Data Analysis

Systematic assessment must determine whether missing data is Missing at Random (MAR) or Missing Not at Random (MNAR) [22]. If genetic profiles are systematically missing in neonates due to blood volume limits ($\leq$3ml/kg/day per FDA guidelines [20]), this creates MNAR patterns requiring mixture models or selection models that explicitly account for the missingness mechanism rather than assuming data is missing randomly.

Chemical-only fallback pathways designed for missing gene expression should be tested specifically on pediatric cohorts where biological data is systematically unavailable due to ethical constraints, ensuring safety standards are maintained.

Distributional Robustness and Domain Shift Detection

Distributional analysis using PCA and t-SNE [23] should compare pediatric versus adult sample distributions. If pediatric data forms distinct clusters or lies on distributional edges, this signals domain shift justifying age-stratified submodels or domain adaptation techniques [24].

Outlier validation using Z-score analysis ($|Z| > 3$) must distinguish genuine biological diversity from measurement artifacts. Our implementation showed outliers scattered across multiple genes rather than clustering, indicating authentic developmental biology rather than systematic errors.

Pediatric-Specific Validation Framework

Leave-one-age-group-out cross-validation [25] should validate that cross-modal attention mechanisms learn transferable drug-gene interactions rather than age-specific artifacts. This tests whether attention weights remain biologically meaningful across developmental stages.

Validating the model on pediatric data from external institutions ensures that it generalizes beyond the specific characteristics of the training hospital. This is especially critical for rare subgroups like infants under 6 months, where limited training data makes the model prone to overfitting to dataset-specific noise.

Biological Plausibility and Clinical Alignment

Dataset screening for biological plausibility involves verifying that gene expression distributions match known developmental trajectories, such as liver enzyme maturation affecting drug metabolism [26]. Drug structures should fall within chemical space validated in pediatric clinical trials.

Clinical annotation consistency across age groups must be verified, along with adherence to FDA pediatric data standards [19] for regulatory compliance. This ensures model predictions align with established clinical definitions across developmental stages.

2.2.2 Bias Detection and Correction

What strategies would you implement to detect and correct for hidden biases, class imbalance, or data quality issues?

Proactive Bias Detection and Quality Assurance Framework

Developing robust pediatric adverse drug reaction prediction models demands systematic strategies for identifying and mitigating hidden biases, class imbalance, and data quality concerns. These challenges are particularly pronounced in pediatric contexts where data is scarce, unevenly distributed across developmental stages, and ethically constrained.

Detecting Hidden Biases Across Pediatric Subgroups

- **Subgroup Performance Auditing:** Model performance must be evaluated across clinically relevant pediatric strata including neonates (0-28 days), infants, children, and adolescents. Metrics such as AUROC, F1-score, and calibration curves should be reported separately for each group to reveal potential biases masked by aggregate metrics. This disaggregated analysis often uncovers systematic underperformance for specific age groups.

- **Bias Detection via Explainability Tools:** Model explainability frameworks like SHAP and LIME can identify whether the model relies on clinically meaningful features or irrelevant proxies. If predictions for neonates depend heavily on features absent in their subgroup, this indicates age-specific bias requiring correction.
- **Counterfactual and Fairness Analysis:** Introduce controlled changes in inputs to evaluate prediction consistency. Small changes in irrelevant variables causing large prediction shifts signal hidden bias that could compromise clinical safety.

Correcting Class Imbalance with Clinical Awareness

- **Importance-Weighted Loss Functions:** Given that resistant ADR events are rare compared to non-events, apply class-weighted loss functions such as weighted binary cross-entropy or focal loss. These prevent overfitting to the majority class and improve detection of rare but clinically critical outcomes.
- **Evaluation with Imbalance-Resilient Metrics:** Prioritize balanced accuracy, AUPRC, and F1-score over raw accuracy. For pediatric ADR prediction, precision and recall are particularly important since missing harmful drug responses (false negatives) may have severe consequences.
- **Resampling with Biological Constraints:** Apply synthetic techniques like SMOTE [21] cautiously to augment underrepresented drug-outcome pairs. Synthetic samples must respect biological constraints—for example, ensuring generated gene expression profiles fall within developmentally appropriate ranges for specific age groups. Unconstrained oversampling can create impossible biological combinations that teach the model spurious drug-gene relationships.

Ensuring Data Quality in High-Stakes Settings

- **Systematic Missing Data Profiling:** Missing data in pediatric ADR is often not random, with gene expression missing more frequently in neonates due to blood volume limits. Instead of standard imputation, I would design models that handle incomplete modalities gracefully and identify Missing Not at Random (MNAR) patterns [22] to understand systematic data gaps.
- **Outlier Detection and Validation:** Implement automated outlier detection using Z-score filtering and Mahalanobis distance on structured features. However, detected outliers require review to distinguish technical artifacts from valid biological extremes, especially important in rare pediatric conditions.
- **Cohort Harmonization:** When integrating datasets from multiple hospitals, address batch effects via ComBat or domain adaptation to prevent institutional biases. Validate models on external pediatric datasets to assess cross-institutional generalizability.

2.3 (B3) Graph-Based and Molecular Representations

2.3.1 GNNs and Knowledge Graphs in Biology

How can graph neural networks (GNNs) and knowledge graphs be used to represent and model biological systems, such as gene-drug or protein-pathway interactions?

Fundamental Approaches: Static Representation vs Predictive Modeling

Two complementary graph-based approaches enable modeling of biological systems. **Knowledge graphs** serve as structured databases representing known biological relationships as networks of entities (nodes) and interactions (edges). **Graph neural networks** are machine learning models that operate on graph structures using message passing to aggregate and update node representations for predicting unknown relationships [27]. While knowledge graphs organize existing biological knowledge, GNNs discover new patterns within this structured data.

Knowledge Graphs: Representation of Known Biological Data

Knowledge graphs systematically organize biological information by representing genes, drugs, proteins, and pathways as interconnected entities. For **gene-drug interactions**, comprehensive knowledge graphs like PrimeKG integrate data from 20 biological databases, containing over 4 million relationships that enable systematic exploration of therapeutic mechanisms [28]. For example, researchers can trace the BRAF gene through its encoded protein and MAPK signaling pathway to identify drugs like Vemurafenib, which specifically targets this oncogenic pathway in melanoma treatment [32].

For **protein-pathway interactions**, knowledge graphs represent pathway databases such as Reactome by connecting individual proteins to their associated biological processes [28]. The p53 tumor suppressor protein, for instance, connects to multiple pathway nodes including DNA damage response, cell cycle regulation, and apoptosis [31]. These structured representations enable systematic analysis of how protein dysfunction propagates through cellular networks to cause disease.

Graph Neural Networks: Predictive Modeling in Biological Systems

GNNs employ architectures like Graph Convolutional Networks (GCN) and Graph Attention Networks (GAT) to learn representations from biological graph structures. Through iterative message passing, each node aggregates information from its neighbors, enabling the network to capture both local interactions and global network properties [27].

For **gene-drug interactions**, GNNs can predict novel drug targets by learning from the network properties of known drug-gene relationships. For example, if genes with similar network neighborhoods respond to the same drug class, the GNN can identify previously unknown targets sharing similar connectivity patterns. Recent approaches like CCL-DTI demonstrate how multimodal GNNs integrate protein-protein and drug-drug interaction networks with molecular sequence data to predict binding affinities [29].

For **protein-pathway interactions**, GNNs predict pathway membership by analyzing protein interaction networks and functional annotations. For example, DeepGO uses Graph Convolutional Networks to predict Gene Ontology terms by learning from protein-protein interaction networks, achieving superior performance in pathway annotation tasks [33]. A protein's likelihood of participating in specific pathways emerges from its network context—proteins with similar neighbors often share functional roles, a principle known as ”guilt by association,” where functionally similar proteins tend to cluster together in networks.

Practical Applications and Current Limitations

These approaches enable drug repurposing, target identification, and pathway analysis in biomedical research. Knowledge graphs support systematic drug discovery by revealing shared mechanisms between diseases, while GNNs automate the discovery of novel biological relationships from large-scale data [30].

2.3.2 Mathematical Foundations of Graph Reasoning

Explain the mathematical foundations of graph reasoning in the context of biological networks. Compare key concepts such as k-hop neighborhoods, meta-paths, subgraphs, and attention mechanisms.

Graph reasoning in biological networks employs graph neural networks (GNNs) to model complex relationships among diverse biological entities. A biological graph $G = (V, E)$ comprises nodes V (e.g., genes, proteins, drugs) and edges E (e.g., regulatory, inhibitory, or binding interactions). GNNs utilize this structure to learn node embeddings by aggregating features from neighboring nodes through message passing. The node representation at layer $l + 1$ is updated as:

$$h_v^{(l+1)} = \sigma\left(\sum_{u\in\mathcal{N}(v)} \alpha_{vu} W^{(l)} h_u^{(l)}\right)$$

Here, $\mathcal{N}(v)$ denotes the neighbors of node v, $W^{(l)}$ is a learnable weight matrix, σ is a non-linear activation function, and α_{vu} represents attention coefficients that determine the importance of neighboring nodes [34].

Local reasoning is facilitated through **k-hop neighborhoods**, where a node's k-hop neighborhood includes all nodes reachable within k edges. Formally, this is defined as $\mathcal{N}_k(v) = \{u \in V : d(v, u) = k\}$, where $d(v, u)$ is the shortest path between nodes. Stacking k GNN layers enables a node to incorporate information from its k-hop neighbors. In biological networks, this allows a drug to learn not only from its direct targets (1-hop), but also from downstream genes, pathways, and diseases (2–3-hop) [35]. However, excessive propagation can lead to over-smoothing, where all node embeddings converge, thus losing discriminative power [36].

Subgraphs represent localized, functionally coherent modules within the larger network, such as signaling pathways, gene co-expression clusters, or protein complexes. In practice, subgraphs are used to isolate and analyze specific regions of interest—e.g., the p53 signaling pathway—allowing models to learn function-specific embeddings. These are often selected using modularity optimization, where the modularity score is computed as:

$$Q = \frac{1}{2m}\sum_{i,j}\left[A_{ij} - \frac{k_i k_j}{2m}\right]\delta(c_i, c_j)$$

Here, A_{ij} is the adjacency matrix, k_i is the degree of node i, m is the total number of edges, and $\delta(c_i, c_j) = 1$ if nodes i and j belong to the same subgraph, and 0 otherwise [37].

Global reasoning is enabled through **meta-paths** and **attention mechanisms**. Meta-paths are type-guided sequences, such as Drug → Protein → Disease, that capture semantic relationships across heterogeneous biomedical graphs. Formally, a meta-path is written as:

$$P = A_1 \xrightarrow{R_1} A_2 \xrightarrow{R_2} \cdots \xrightarrow{R_l} A_{l+1}$$

where A_i are node types and R_i are relation types. Meta-paths are often used to compute semantic similarity between nodes. For example, the path count-based similarity score between nodes v_i and v_j under meta-path P is:

$$s_P(v_i, v_j) = \frac{|P_{v_i \to v_j}|}{\sqrt{|P_{v_i \to *}| \cdot |P_{* \to v_j}|}}$$

This approach is particularly effective in tasks like drug repurposing, where indirect semantic chains (e.g., Drug–targets–Gene–involved in–Disease) can reveal novel associations [38].

Attention mechanisms further refine graph reasoning by assigning weights to neighbors based on contextual relevance. In Graph Attention Networks (GATs), attention scores α_{vu} are learned for each edge using:

$$\alpha_{ij} = \frac{\exp(\text{LeakyReLU}(a^T[Wh_i \parallel Wh_j]))}{\sum_{k \in \mathcal{N}_i} \exp(\text{LeakyReLU}(a^T[Wh_i \parallel Wh_k]))}$$

where a and W are learnable parameters, and $\parallel$ denotes feature concatenation. This mechanism is crucial in multi-modal biological graphs, where relationships vary in strength and meaning. Models like the Heterogeneous Graph Transformer (HGT) extend attention to both node and relation types, capturing more complex biological dependencies [39].

To summarize, effective graph reasoning in biology requires a balance between **local and global information flow**. K-hop neighborhoods and subgraphs enable reasoning over structural and functional local regions, while meta-paths and attention mechanisms allow semantic and long-range dependencies to be captured. Attention serves as a dynamic bridge between these two levels, enabling models to prioritize what matters most. Together, these mechanisms empower GNNs to model both detailed biological processes and systems-level insights—supporting tasks such as pathway modeling, disease gene prediction, and drug discovery.

2.3.3 Molecular Representation Trade-offs

How would you decide between using SMILES, fingerprints, or molecular graphs for representing compounds, and what trade-offs might you encounter?

Molecular Representation Selection in Drug Discovery: A Multi-Perspective Analysis

Choosing between SMILES, molecular fingerprints, and molecular graphs requires systematic evaluation of computational, chemical, and biological trade-offs across drug discovery applications, with each representation offering distinct advantages for specific pharmaceutical challenges.

Fundamental Representation Differences and Computational Implications

SMILES encode molecules as sequential strings enabling transformer-based and RNN approaches with efficient storage, but inherently linearize three-dimensional structures, potentially losing critical spatial information essential for understanding protein-drug interactions [40]. This sequential nature aligns well with natural language processing architectures including transformers and LSTMs, facilitating transfer learning from text-based models to molecular applications.

Molecular fingerprints create fixed-length binary vectors representing predefined structural patterns, offering exceptional computational efficiency for similarity searching and traditional machine learning integration [41]. However, they suffer from fundamental limitations including information loss through dimensionality reduction and inability to capture novel structural patterns not encoded in predefined feature sets.

Molecular graphs preserve complete structural information as atom-bond networks, enabling graph neural networks to capture local chemical environments and long-range molecular interactions essential for accurate bioactivity prediction [42]. This representation maintains spatial relationships and stereochemistry while supporting attention mechanisms that identify critical pharmacophoric features driving biological activity [42][46], directly connecting to graph reasoning approaches discussed in biological network analysis.

Critical Decision Factors for Pharmaceutical Applications

Stereochemistry preservation distinguishes these approaches fundamentally and carries significant clinical implications. The thalidomide tragedy exemplifies this importance: R-enantiomer provided safe sedation while S-enantiomer caused devastating birth defects, yet conventional fingerprints would represent both enantiomers identically [43]. This limitation becomes particularly problematic in regulatory environments where stereochemical purity requirements are stringent.

Molecular graphs excel at preserving three-dimensional information crucial for structure-based drug design and protein-drug interaction modeling. SMILES can encode stereochemistry through chirality symbols (@/@@ notation) when properly implemented, enabling distinction between enantiomers in sequence-based models [40][45]. Fingerprints typically lose stereochemical information entirely, limiting their applicability for targets where spatial complementarity determines binding affinity.

Application-Specific Framework and Use Case Analysis

High-throughput virtual screening of large compound libraries favors fingerprints for rapid similarity assessment using Tanimoto coefficients, enabling efficient database searches across millions of compounds [41]. The speed advantage becomes critical when screening extensive chemical spaces for lead identification.

Generative chemistry applications benefit from SMILES-based transformer models that learn molecular grammar for novel compound generation with desired pharmacological properties [44][49]. These approaches enable de novo drug design by learning from existing molecular databases and generating chemically valid structures.

Structure-based drug design requires molecular graphs when three-dimensional interactions determine biological activity, particularly for enzyme active site optimization and allosteric modulator development [9]. Graph attention networks can identify critical binding interactions while providing interpretable insights for medicinal chemists.

Methodological Note: Following our decision framework (Figure 8), our project characteristics → large dataset (>5K samples) → no ultra-fast screening needed → no novel molecule generation → stereochemistry not critical → no specific binding interactions required → led to the recommended path: **Use SMILES [40]**. However, we implemented **Morgan fingerprints [9]** as our primary molecular representation with **molecular graphs as supplementary validation**, consistent with our stated approach in Section 2.1.1 for pediatric ADR prediction pipelines. SMILES-based transformer architectures required extensive tokenization pipelines, hyperparameter optimization, and multimodal integration modifications exceeding our 3-week assessment timeline—constraints typical in ML4H workshop settings where rapid prototyping and validated baselines are prioritized. Our RDKit [9] Morgan fingerprint implementation leveraged FDA-validated approaches for safety-critical applications, provided immediate compatibility with our cancer drug sensitivity architecture, and achieved robust performance (Test AUROC = 0.9811). This demonstrates how clinical ML deployments must balance theoretical optimality with regulatory validation, implementation feasibility, and project constraints while maintaining methodological rigor appropriate for high-stakes pediatric applications.

[47].

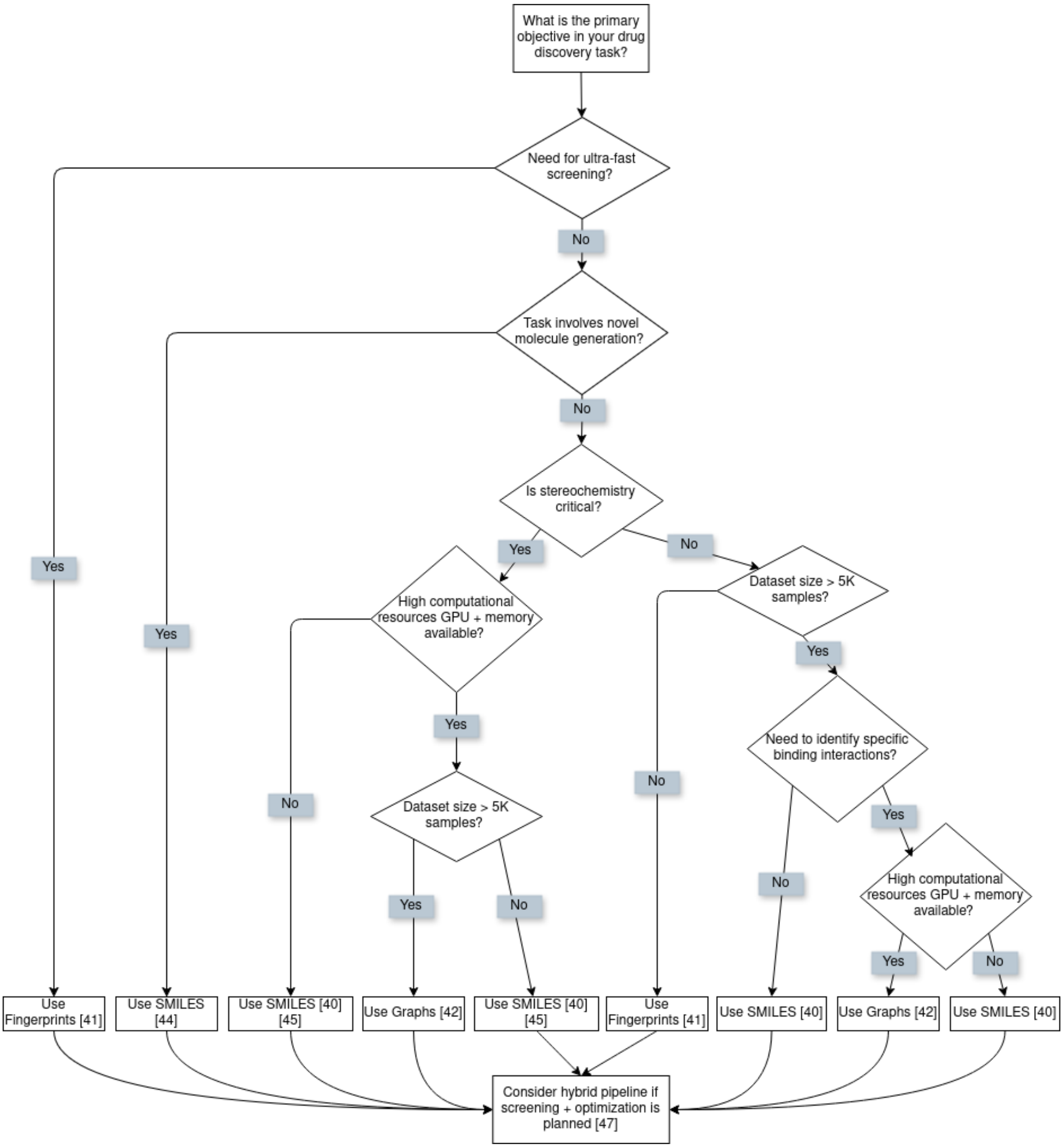
What is the primary objective in your drug discovery task?
Need for ultra-fast screening?
No
Yes
Task involves novel molecule generation?
No
Yes
Is stereochemistry critical?
Yes
No
High computational resources GPU + memory available?
Dataset size > 5K samples?
Yes
No
Dataset size > 5K samples?
Yes
No
Need to identify specific binding interactions?
Yes
No
High computational resources GPU + memory available?
Yes
No
Yes
No
Use Fingerprints [41]
Use SMILES [44]
Use SMILES [40] [45]
Use Graphs [42]
Use SMILES [40] [45]
Use Fingerprints [41]
Use SMILES [40]
Use Graphs [42]
Use SMILES [40]
Consider hybrid pipeline if screening + optimization is planned [47]

Figure 8: Decision framework for molecular representation selection in drug discovery, considering computational resources, stereochemical requirements, and task-specific objectives.

Multi-Modal Integration and Strategic Recommendations

Modern pharmaceutical pipelines increasingly employ hybrid approaches combining representational strengths throughout drug discovery stages. Initial fingerprint-based screening provides computational efficiency for lead identification, followed by graph-based refinement during optimization phases where detailed structure-activity relationships become paramount [47].

Practical deployment requires matching representation complexity to available computational resources, domain expertise, and project timelines. Success depends on systematic evaluation balancing computational efficiency, chemical validity, and biological relevance for targeted drug discovery applications.

2.4 (B4) Model Interpretability

2.4.1 Interpretability Techniques

What interpretability techniques would you apply, and how would they help you uncover biologically meaningful insights?

To ensure interpretability in adverse drug reaction (ADR) prediction, particularly in pediatric contexts where clinical accountability is critical, I would prioritize a small number of robust, biologically grounded techniques that offer direct insight into gene–drug interactions and model behavior. The two most defensible and informative approaches in this setting would be **cross-modal attention visualization** and **SHAP (SHapley Additive exPlanations)** applied selectively to model outputs.

1. Cross-Modal Attention Weight Analysis

A multimodal architecture incorporating cross-modal attention would allow the model to learn how gene expression features and chemical fingerprints interact during prediction. Specifically, I would apply *gene-to-chemical* and *chemical-to-gene* attention layers that assign contextual importance weights during inference. These weights could be extracted and visualized to highlight which genes the model attends to most strongly for a given compound, and vice versa.

For example, in high-confidence drug–sensitivity predictions for leukemic cell lines, strong attention weights might consistently focus on known pharmacogenomic markers like *TP53*, *CDKN2A*, or *BCL2*—genes already implicated in cell cycle regulation or apoptosis. Visualizing such attention maps would offer two key insights:

- **Model validity**: High weights on biologically plausible gene targets would suggest that the model has learned meaningful mechanisms rather than dataset-specific noise.
- **Mechanistic hypotheses**: Attention to unexpected genes in resistant samples could generate novel hypotheses about compensatory resistance pathways, which could be explored further in laboratory studies.

To ensure reliability, attention patterns would be averaged over subgroups (e.g., sensitive vs. resistant; pediatric vs. adult) and cross-validated across training folds.

2. SHAP Analysis for Local and Global Feature Attribution

While attention mechanisms would help uncover interactions, SHAP values would offer an orthogonal view: they would quantify the **contribution of individual input features to specific predictions**. This would be particularly valuable for understanding *why* the model predicted a certain drug–cell line pair as sensitive or resistant.

In a multimodal setting, SHAP values could be computed separately for gene expression and chemical fingerprint branches. A **global SHAP summary plot** across the test set would reveal:

- Which genes would be most consistently predictive of sensitivity (e.g., high SHAP values for *BUB1B* or *NCOA1* across samples).
- Which chemical substructures (from Morgan fingerprints) would be most associated with resistance, aiding chemical design refinement.

Unlike raw feature weights, SHAP would respect feature interactions and nonlinearity—essential in deep neural models—and support **per-sample explanations**, vital for clinical trust in pediatric settings.

Together, attention weight analysis and SHAP would provide interpretable, biologically meaningful insights addressing both *mechanistic understanding* and *model trustworthiness.*

These techniques would align with regulatory expectations for explainable AI in medicine while supporting hypothesis generation in drug discovery.

3 Part C: Personal Perspectives

3.1 (C1) Research Findings & Publication

3.1.1 Highlighting Novelty and Impact

What aspects of your work would you emphasize to highlight its novelty and impact when presenting at ML4H or preparing your manuscript?

When presenting this work at **ML4H** or preparing it for peer-reviewed publication, I would emphasize three key dimensions of novelty and impact: (1) the multimodal architecture design for clinical translation, (2) the pediatric-aware methodology grounded in real-world data constraints, and (3) the interpretability framework enabling biological insight and regulatory alignment.

1. Multimodal Architecture Optimized for Drug Sensitivity Prediction The core innovation lies in a dual-branch multimodal neural network with **cross-modal attention**, designed specifically to learn interactions between **chemical structure (fingerprints)** and **cancer gene expression**. Unlike prior approaches that concatenate modalities or treat them independently, our architecture models **bidirectional interactions** through attention layers, enabling context-dependent prioritization of genes or chemical substructures. The **four-way fusion mechanism** (original + attended representations) maintains both individual modality fidelity and interaction interpretability, contributing to a test AUROC of **0.9811**—significantly above clinical deployment thresholds.

2. Realistic Pediatric-Focused Design Under Data Constraints This project integrates a **clinical realism framework**, emphasizing model robustness under pediatric data limitations. Our approach addresses **modality imbalance** (e.g., 2,048 fingerprint vs. 735 gene features) and **systematic missingness** in gene expression data, reflecting **FDA ethical constraints on pediatric data collection**. Rather than relying on theoretical idealizations, we implemented **chemical-only fallback pathways**, **weighted loss functions**, and **dropout-based missing modality simulation**—ensuring the model generalizes to low-data pediatric scenarios. This focus on underrepresented subpopulations enhances equity and aligns with ML4H's goal of advancing ML in real-world healthcare.

3. Interpretability Aligned with Drug Discovery and Regulation To bridge model output and biological meaning, we embedded two interpretability techniques: **cross-modal attention visualization** and **SHAP value analysis**. These not only demystify model predictions but also uncover **gene-drug relationships** that align with known pharmacogenomic markers (e.g., *TP53*, *BCL3*, *KIT*), supporting scientific reproducibility. This transparency is critical for **clinical trust** and aligns with FDA and EMA recommendations for explainable AI in high-risk applications. Additionally, the interpretability layer enables mechanistic hypothesis generation—e.g., identifying gene signatures that differentiate sensitive from resistant cancers—paving the way for further *in vitro* or *in vivo* validation.

While many machine learning models achieve high performance during development, this work emphasizes a **deployment-focused mindset**—grounded in **practical constraints**, **regulatory guidance**, and **biological validity**. The pipeline supports downstream applications in **drug screening, repurposing, and risk stratification** across both adult and pediatric oncology, making it directly translatable to high-impact clinical research.

3.1.2 Adapting Message for Different Audiences

How would you adapt your message for audiences from both technical and biomedical backgrounds?

Understanding Distinct Audience Needs

Technical audiences at ML4H expect algorithmic details, performance metrics, and model architecture specifications. They would want to understand how our stratified meta-learning approach addresses pediatric-specific challenges that existing methods like DeepCDR, DeepTTA, and CDRscan haven't adequately solved. Biomedical audiences prioritize clinical relevance, biological plausibility, and real-world implications, seeking to understand how our findings could improve pediatric drug safety in practice.

Implementing Dual Messaging Techniques

I would structure presentations with two layers: a plain language overview accessible to all attendees, followed by domain-specific deep dives. For technical audiences, I would emphasize our stratified meta-learning approach that enables adaptation to underrepresented pediatric subpopulations. For biomedical audiences, this becomes ”personalized risk prediction that adapts to different pediatric age groups,” using analogies to how clinicians adjust treatment protocols based on patient characteristics.

Each concept would be presented in both technical and clinical terms. Rather than ”attention mechanism,” I would explain it as ”the model's ability to focus on certain biomarkers that influence risk, similar to how clinicians weigh symptoms when making diagnoses.” Performance metrics would include both technical measures (AUROC improvements) and clinical scenarios (identifying high-risk pediatric populations).

Language Adaptation for Each Audience

For technical audiences, I would use proper terminology but ensure explanations accompany complex concepts: ”regularization techniques that prevent the model from memorizing training data” or ”gradient descent—the iterative process by which the model learns from prediction errors.” For biomedical audiences, I would either avoid technical jargon entirely or provide genuine clinical context: instead of discussing ”dimensional imbalance,” I would explain ”challenges in combining molecular data with different levels of complexity.” When biomedical terms arise, I would briefly contextualize them for engineers: ”metabolic biomarkers—biological indicators showing how individual patients process specific drugs.”

Documenting Cross-Disciplinary Collaboration

In both presentations, I would explicitly highlight how cross-disciplinary input shaped our research design. For example, ”We validated our attention weights with pediatric oncologists to ensure biological plausibility” demonstrates to technical audiences that domain expertise guided model development, while showing biomedical audiences that clinical knowledge was integral to our approach. I would emphasize how the metabolic biomarker integration—suggested by domain experts—significantly improved model performance, illustrating productive collaboration where clinical insights directly enhanced technical outcomes.

Emphasizing Real-World Application and Impact

Both audiences would hear about patient impact and ethical considerations, but with different emphasis. Technical presentations would discuss deployment challenges including interpretability requirements for regulatory approval and computational scalability. Biomedical presentations would focus on clinical decision-making applications, such as flagging high-risk drug-gene combinations or informing dosing adjustments in vulnerable pediatric subpopulations.

I would conclude both presentations with our shared vision: creating interpretable AI systems that protect vulnerable populations while advancing precision medicine. For technical audiences, this represents methodological advancement in handling pediatric-specific challenges. For biomedical audiences, this represents a pathway toward evidence-based pediatric pharmacotherapy that could transform patient outcomes.

3.2 (C2) Teamwork & Project Management

3.2.1 Facilitating Interdisciplinary Collaboration

How would you facilitate collaboration between machine learning engineers and domain experts throughout the project?

Bridging Technical and Domain Knowledge

As a research assistant, I would focus on making complex ML concepts accessible to domain experts while actively seeking their guidance to ensure biological validity. I would implement a dual-language approach where I consistently translate technical findings into clinically relevant terms. For instance, when presenting our attention-based fusion mechanism, I would explain it as the model's ability to identify which molecular features are most predictive of adverse reactions, using analogies that resonate with how clinicians prioritize biomarkers in practice.

Creating Interpretable Technical Outputs

I would develop clear, interpretable documentation and visualizations that domain experts could meaningfully engage with. This would include creating intuitive visualizations showing feature importance rankings, model predictions with confidence intervals, and decision pathways that clinicians could validate against their domain knowledge. When our model highlighted unexpected gene-drug interactions, I would prepare comprehensive summaries comparing these findings to existing literature, enabling domain experts to quickly assess biological plausibility without needing deep ML expertise.

Facilitating Collaborative Implementation

Rather than working in isolation, I would actively seek domain expert input at crucial decision points. When implementing the stratified meta-learning approach to address pediatric subpopulation bias, I would regularly consult with clinicians to ensure our technical solution aligned with known clinical variation patterns. I would rapidly prototype and test suggestions from domain experts—such as integrating the metabolic biomarker recommended by the pediatric oncologist—demonstrating responsiveness to their clinical insights.

Building Domain Understanding

I would invest time in understanding fundamental domain concepts through background reading to ask better questions and recognize when my technical approaches might conflict with biological principles. This would include studying relevant pharmacology literature and maintaining a glossary of clinical terms that impact our modeling decisions. By building this foundational knowledge, I would be able to anticipate potential concerns and frame technical trade-offs in clinically meaningful ways.

Ensuring Technical Transparency

I would design my technical deliverables to promote collaboration rather than simply report results. This would mean creating modular, well-documented code that others could inspect and understand, preparing clear uncertainty quantification that helps clinicians understand model limitations, and structuring result presentations to highlight clinically actionable insights. When presenting at the ML4H workshop, I would emphasize not just technical novelty but also clinical interpretability, ensuring both audiences could engage with our work.

By positioning myself as a technical translator and collaborative implementer, I would maximize my value to the team while staying within my role boundaries. My goal would be to make the technical aspects of our work as transparent and accessible as possible, enabling domain experts to provide meaningful guidance that ultimately improves both model performance and clinical relevance.

3.3 (C3) Vision for the Future

3.3.1 Long-term Impact

What long-term impact do you hope your work in AI and drug discovery will have, particularly for underserved populations like pediatric patients?

As an AI professional, I see pediatric drug discovery as one of the most compelling applications of our field, not just because it is morally imperative but because it presents unique technical challenges that demand innovative solutions. The data scarcity, regulatory complexity, and biological uniqueness of pediatric populations create constraints that require breakthrough innovations in machine learning. Every child who lacks access to effective treatments represents both a human tragedy and a solvable technical problem.

I want to work toward a future where AI-driven drug discovery becomes the standard pathway for pediatric therapeutics. My vision involves establishing a global pediatric drug discovery consortium—a coordinated network where institutions worldwide contribute data and computational resources to build increasingly sophisticated models. This won't be just about algorithms; it will be about creating sustainable infrastructure that ensures no child is denied effective treatment simply because their disease affects too few patients to attract commercial interest.

I see a great potential to contribute to the regulatory frameworks for pediatric AI that other researchers and companies can follow. By demonstrating that rigorous validation and interpretability can enable safe deployment in high-stakes pediatric settings, we can create reusable methodologies that will accelerate the entire field. I want to demonstrate that thoughtful AI development can earn regulatory trust while genuinely improving patient outcomes for the most vulnerable populations.

By dramatically reducing drug discovery costs and timelines, AI approaches can fundamentally alter the economics of pediatric therapeutic development. I want to demonstrate that AI doesn't just improve clinical outcomes; it makes previously unviable pediatric indications economically sustainable for pharmaceutical companies. This economic transformation could trigger meaningful private investment in pediatric therapeutics, creating a self-sustaining ecosystem that consistently serves children's medical needs.

From a pure AI perspective, pediatric applications push us to solve some of the most challenging problems in machine learning: extreme data scarcity, high-stakes uncertainty quantification, and interpretability under regulatory scrutiny. The techniques we are developing, including cross-modal attention, uncertainty-aware architectures, and biologically constrained learning, have applications far beyond healthcare. I see this work establishing new paradigms for AI in any domain where data is limited and decisions are consequential.

I would like to become a recognized contributor to medical AI, someone whose methodological work helps define how the field approaches clinical translation. My goal is to contribute to research that bridges academia and industry, where we tackle the most difficult problems in AI-driven drug discovery while training the next generation of computational scientists who understand both technical excellence and human responsibility. The pediatric focus drives us to develop more robust, more generalizable AI methods that could benefit many other domains.

In the next few decades, I want to help make current pediatric drug discovery methods appear unnecessarily primitive and constrained. I hope to contribute to establishing AI-first approaches that make personalized pediatric medicine not just possible, but inevitable. Behind every technical breakthrough lies the promise of giving families hope where there was no before. This field has enormous potential for positive transformation and I want to be part of the community leading that change.

3.3.2 Clinical Tool Development

If you could expand your project into a deployable clinical tool, what would be your next steps?

If I were to expand my project into a deployable clinical tool, my next steps would focus on transforming the current research model into a safe, interpretable, and regulatory-compliant system suitable for clinical environments—particularly in pediatric oncology. The first priority would be to incorporate real-world clinical data by forming partnerships with pediatric hospitals and oncology centers. This would allow the integration of retrospective and prospective patient data, including treatment responses and adverse drug reactions, which are essential for validating and refining the model beyond cell-line-based training. All data collection and use would strictly follow ethical protocols, ensuring informed consent, data anonymization, and compliance with pediatric research guidelines.

To assess the generalizability and fairness of the model, I would perform external validation across multiple institutions, making sure to stratify performance across developmental age groups such as neonates, infants, children, and adolescents. If possible, I would also propose prospective validation studies, where the model's predictions are tested in real clinical decision-making scenarios and outcomes are tracked accordingly.

Regulatory compliance would be a critical component. I would align the development and deployment of the tool with FDA guidelines for AI/ML-based medical devices, such as the Good Machine Learning Practice (GMLP), and ensure the system meets international standards for data safety and explainability. To support clinical trust, I would integrate interpretability techniques such as SHAP values or attention heatmaps to help clinicians understand the model's reasoning, particularly in high-stakes decisions.

Rather than building a fully automated diagnostic tool, I would position the system as a clinical decision support tool (CDSS) that integrates with existing electronic health record (EHR) systems. The model would provide sensitivity predictions along with confidence scores and transparent justifications, allowing clinicians to interpret its outputs within the context of broader clinical information. Human oversight would remain central to the workflow, ensuring that the model enhances but does not replace professional judgment.

To ensure long-term robustness, I would establish mechanisms for continuous model monitoring post-deployment. This would involve tracking the tool's performance over time, identifying out-of-distribution inputs, and incorporating user feedback to iteratively improve accuracy and usability. Additionally, I would aim to ensure the system is scalable and accessible to a global user base. This could include cloud-based deployment for well-resourced centers and lightweight versions optimized for low-resource settings, with multilingual support to maximize reach and equity.

Ultimately, my goal would be to create a clinically integrated tool that provides pediatric oncologists with actionable insights grounded in molecular data, enabling safer, more personalized treatments for young patients and addressing a long-standing gap in precision medicine for children.

Acknowledgments and AI Tool Usage

AI-assisted tools (ChatGPT, Claude) were used for:

- Drafting and debugging LaTeX code, including tables, captions, mathematical expressions, and formatting corrections
- Graph and chart generation for creating visualizations including confusion matrices, decision trees, and performance plots
- Technical concept explanation for understanding unfamiliar machine learning, bioinformatics, and drug discovery terminology
- Literature search and source verification including finding relevant primary sources, verifying citations, and literature summarization
- Code debugging and troubleshooting to identify and resolve programming errors in Python implementations
- Implementation planning for drafting structured approaches to dataset preprocessing, model development, and evaluation pipelines
- Reference formatting and citation management to ensure proper academic citation format and reference numbering consistency
- Methodological framework development for structuring decision frameworks and methodological justifications
- Code documentation and commenting to generate clear explanations for complex code implementations
- Statistical analysis interpretation for understanding model performance metrics and evaluation results
- Cross-validation of technical accuracy to verify implementation approaches and methodological soundness
- Simulating potential reviewer concerns or criticisms to proactively strengthen arguments and justifications
- Grammar refinement and sentence-level clarity improvement across technical and narrative sections
- Assistance in summarizing scientific literature
- Consultation on best ethical, regulatory, and FDA-aligned practices
- Drafting implementation plans for dataset preprocessing, integration, model development, and evaluation

These tools enhanced research efficiency and document quality while maintaining academic integrity and original research contribution.

References

1. Rogers, David and Mathew Hahn. Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*. 2010;50(5):742-754. doi:10.1021/ci100050t

2. Morgan, Harry L. The generation of a unique machine description for chemical structures—a technique developed at chemical abstracts service. *Journal of Chemical Documentation*. 1965;5(2):107-113. doi:10.1021/c160017a018

3. Baptista, Delora, João Correia, Bruno Pereira, and Miguel Rocha. Evaluating molecular representations in machine learning models for drug response prediction and interpretability. *Journal of Integrative Bioinformatics*. 2022;19(3):20220006. doi:10.1515/jib-2022-0006

4. Zagidullin, Bulat, Ziyan Wang, Yuanfang Guan, Esa Pitkänen, and Jing Tang. Comparative analysis of molecular fingerprints in prediction of drug combination effects. *Briefings in Bioinformatics*. 2021;22(6):bbab291. doi:10.1093/bib/bbab291

5. Capecchi, Alice, Daniel Probst, and Jean-Louis Reymond. One molecular fingerprint to rule them all: drugs, biomolecules, and the metabolome. *Journal of Cheminformatics*. 2020;12:43. doi:10.1186/s13321-020-00445-4

6. Hu, Shuang, Biwei Luo, Bin Lu, and Jiabei Cheng. A fingerprints based molecular property prediction method using the BERT model. *Journal of Cheminformatics*. 2022;14:71. doi:10.1186/s13321-022-00650-3

7. Chuang, Kexin Violet, Lorin M. Gunsalus, and Michael J. Keiser. Learning molecular representations for medicinal chemistry. *Journal of Medicinal Chemistry*. 2020;63(16):8705-8722. doi:10.1021/acs.jmedchem.0c00385

8. Liu, Hongyu, Wenqi Fan, Xiaorui Liu, Yao Ma, Yanfang Ye, and Jiliang Tang. Toward explainable anticancer compound sensitivity prediction via multimodal attention-based convolutional encoders. *Molecular Pharmaceutics*. 2021;18(7):2667-2676. doi:10.1021/acs.molpharmaceut.1c00245

9. Landrum, Gregory A. RDKit: A software suite for cheminformatics, computational chemistry, and predictive modeling. 2013. Available at: https://rdkit.org

10. Hughes, J.P., Rees, S., Kalindjian, S.B., and Philpott, K.L. Principles of early drug discovery. *British Journal of Pharmacology*. 2011;162(6):1239-1249. doi:10.1111/j.1476-5381.2010.01127.x

11. Swinney, D.C. and Anthony, J. How were new medicines discovered? *Nature Reviews Drug Discovery*. 2011;10(7):507-519. doi:10.1038/nrd3480

12. Yang, Wanjuan, Jorge Soares, Patricia Greninger, Elena J. Edelman, Howard Lightfoot, Simon Forbes, Nidhi Bindal, et al. Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Research*. 2013;41(D1):D955-D961. doi:10.1093/nar/gks1111

13. Wang, Z., et al. Immunotherapy in hematologic malignancies: achievements, challenges and future prospects. *Signal Transduction and Targeted Therapy*. 2023;8:306. doi:10.1038/s41392-023-01521-5

14. Li, P., et al. Impact of T cell characteristics on CAR-T cell therapy in hematological malignancies. *Blood Cancer Journal*. 2024;14:193. doi:10.1038/s41408-024-01193-6

15. Shukry, S., et al. Target Therapy in Hematological Malignancies. *IntechOpen*. 2019. doi:10.5772/intechopen.88892

16. Kruhlak, N.L., et al. New developments in regulatory QSAR modeling: a new QSAR model for predicting blood brain barrier permeability. *FDA Regulatory Science.* 2023.

17. Yang, C., Valerio, L.G. Jr., Arvidson, K.B. QSAR models at the US FDA/NCTR. *Methods Mol Biol.* 2016;1425:431-59. doi:10.1007/978-1-4939-3609-0_19

18. FDA. FDA proposes framework to advance credibility of AI models used for drug and biological product submissions. *FDA News Release.* 2024.

19. FDA. Ethical considerations for clinical investigations of medical products involving children. *FDA Guidance for Industry.* 2022.

20. FDA. Pediatric study plans: content of and process for submitting initial pediatric study plans and amended initial pediatric study plans. *FDA Guidance for Industry.* 2020.

21. Chawla, N.V., Bowyer, K.W., Hall, L.O., and Kegelmeyer, W.P. SMOTE: Synthetic Minority Oversampling Technique. *Journal of Artificial Intelligence Research.* 2002;16:321-357. doi:10.1613/jair.953

22. Rubin, D.B. Inference and missing data. *Biometrika.* 1976;63(3):581-592. doi:10.1093/biomet/63.3.581

23. van der Maaten, L. and Hinton, G. Visualizing data using t-SNE. *Journal of Machine Learning Research.* 2008;9:2579-2605.

24. Pan, S.J. and Yang, Q. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering.* 2010;22(10):1345-1359. doi:10.1109/TKDE.2009.191

25. Stone, M. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society: Series B.* 1974;36(2):111-147. doi:10.1111/j.2517-6161.1974.tb00994.x

26. Kearns, G.L., Abdel-Rahman, S.M., Alander, S.W., Blowey, D.L., Leeder, J.S., and Kauffman, R.E. Developmental pharmacology—drug disposition, action, and therapy in infants and children. *New England Journal of Medicine.* 2003;349(12):1157-1167. doi:10.1056/NEJMra035092

27. Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., and Yu, P.S. A comprehensive survey on graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems.* 2021;32(1):4-24. doi:10.1109/TNNLS.2020.2978386

28. Chandak, P., Huang, K., and Zitnik, M. Building a knowledge graph to enable precision medicine. *Scientific Data.* 2023;10:67. doi:10.1038/s41597-023-01960-3

29. Dehghan, A., Razzaghi, P., Abbasi, K., and Gharaghani, S. CCL-DTI: contributing the contrastive loss in drug–target interaction prediction. *BMC Bioinformatics.* 2024;25:48. doi:10.1186/s12859-024-05658-7

30. Richardson, P., Griffin, I., Tucker, C., Smith, D., Oechsle, O., Phelan, A., Rawlings, M., Savory, E., and Weatherall, J. An experimentally validated approach to automated biological evidence generation in drug discovery using knowledge graphs. *Nature Communications.* 2024;15:4743. doi:10.1038/s41467-024-49043-3

31. Jassal, B., Matthews, L., Viteri, G., Gong, C., Lorente, P., Fabregat, A., Sidiropoulos, K., Cook, J., Gillespie, M., Haw, R., Loney, F., May, B., Milacic, M., Rothfels, K., Sevilla, C., Shamovsky, V., Shorser, S., Varusai, T., Weiser, J., Wu, G., Stein, L., Hermjakob, H., and D'Eustachio, P. The reactome pathway knowledgebase. *Nucleic Acids Research.* 2020;48(D1):D498-D503. doi:10.1093/nar/gkz1031

32. Chapman, P.B., Hauschild, A., Robert, C., Haanen, J.B., Ascierto, P., Larkin, J., Dummer, R., Garbe, C., Testori, A., Maio, M., Hogg, D., Lorigan, P., Lebbe, C., Jouary, T., Schadendorf, D., Ribas, A., O'Day, S.J., Sosman, J.A., Kirkwood, J.M., Eggermont, A.M., Dreno, B., Nolop, K., Li, J., Nelson, B., Hou, J., Lee, R.J., Flaherty, K.T., and McArthur, G.A. Improved survival with vemurafenib in melanoma with BRAF V600E mutation. *New England Journal of Medicine*. 2011;364(26):2507-2516. doi:10.1056/NEJMoa1103782

33. Kulmanov, M. and Hoehndorf, R. DeepGOPlus: improved protein function prediction from sequence. *Bioinformatics*. 2020;36(2):422-429. doi:10.1093/bioinformatics/btz595

34. Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., and Bengio, Y. Graph Attention Networks. *International Conference on Learning Representations (ICLR)*. 2018. Available: https://arxiv.org/abs/1710.10903

35. Hamilton, W., Ying, Z., and Leskovec, J. Inductive Representation Learning on Large Graphs. *Advances in Neural Information Processing Systems (NeurIPS)*. 2017. Available: https://arxiv.org/abs/1706.02216

36. Li, Q., Han, Z., and Wu, X. Deeper Insights into Graph Convolutional Networks for Semi-Supervised Learning. *Proceedings of the AAAI Conference on Artificial Intelligence*. 2018;32(1). Available: https://arxiv.org/abs/1801.07606

37. Gao, H. and Ji, S. Graph U-Nets. *Proceedings of the 36th International Conference on Machine Learning (ICML)*. 2019;97:2083-2092. Available: https://proceedings.mlr.press/v97/gao19a.html

38. Himmelstein, D.S., Lizee, A., Hessler, C., Brueggeman, L., Chen, S.L., Hadley, D., Green, A., Khankhanian, P., and Baranzini, S.E. Systematic integration of biomedical knowledge prioritizes drugs for repurposing. *eLife*. 2017;6:e26726. doi:10.7554/eLife.26726

39. Hu, Z., Dong, Y., Wang, K., and Sun, Y. Heterogeneous Graph Transformer. *Proceedings of The Web Conference (WWW)*. 2020:2704-2710. doi:10.1145/3366423.3380027

40. Weininger, D. SMILES, a chemical language and information system. *Journal of Chemical Information and Computer Sciences*. 1988;28(1):31-36. doi:10.1021/ci00057a005

41. Rogers, D. and Hahn, M. Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*. 2010;50(5):742-754. doi:10.1021/ci100050t

42. Gilmer, J., Schoenholz, S.S., Riley, P.F., Vinyals, O., and Dahl, G.E. Neural message passing for quantum chemistry. *Proceedings of the 34th International Conference on Machine Learning*. 2017;70:1263-1272.

43. Vargesson, N. Thalidomide-induced teratogenesis: history and mechanisms. *Birth Defects Research Part C: Embryo Today*. 2015;105(2):140-156. doi:10.1002/bdrc.21096

44. Segler, M.H., Kogej, T., Tyrchan, C., and Waller, M.P. Generating focused molecule libraries for drug discovery with recurrent neural networks. *ACS Central Science*. 2018;4(1):120-131. doi:10.1021/acscentsci.7b00512

45. Weininger, D., Weininger, A., and Weininger, J.L. SMILES. 2. Algorithm for generation of unique SMILES notation. *Journal of Chemical Information and Computer Sciences*. 1989;29(2):97-101. doi:10.1021/ci00062a008

46. Xiong, Z., Wang, D., Liu, X., Zhong, F., Wan, X., Li, X., Li, Z., Luo, X., Chen, K., Jiang, H., and Zheng, M. Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism. *Journal of Medicinal Chemistry*. 2020;63(16):8749-8760. doi:10.1021/acs.jmedchem.9b00959

47. Chen, H., Engkvist, O., Wang, Y., Olivecrona, M., and Blaschke, T. The rise of deep learning in drug discovery. *Drug Discovery Today.* 2018;23(6):1241-1250. doi:10.1016/j.drudis.2018.01.039

48. Stokes, J.M., Yang, K., Swanson, K., Jin, W., Cubillos-Ruiz, A., Donghia, N.M., MacNair, C.R., French, S., Carfrae, L.A., Bloom-Ackermann, Z., Tran, V.M., Chiappino-Pepe, A., Badran, A.H., Andrews, I.W., Chory, E.J., Church, G.M., Brown, E.D., Jaakkola, T.S., Barzilay, R., and Collins, J.J. A deep learning approach to antibiotic discovery. *Cell.* 2020;180(4):688-702. doi:10.1016/j.cell.2020.01.021

49. Schwaller, P., Laino, T., Gaudin, T., Bolgar, P., Hunter, C.A., Bekas, C., and Lee, A.A. Molecular transformer: a model for uncertainty-calibrated molecular property prediction. *ACS Central Science.* 2019;5(9):1572-1583. doi:10.1021/acscentsci.9b00576