Drug Synergy Prediction Using Chemical Language Learning and Cross-Modal Attention Mechanism

Jiawei Chen[1], Chao Liu[2], Kwun Sy Lee[1], Debby D. Wang[1*]

[1]School of Science and Technology, Hong Kong Metropolitan University, Ho Man Tin, Hong Kong
[2]College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, Guangdong Province, China
[*]Correspondence author: dwang@hkmu.edu.hk

Abstract:

Combination therapy is a crucial strategy in cancer treatment to improve efficacy and overcome drug resistance. Although many deep learning models have been developed to predict drug synergy, they fail to capture complex biological interactions and leave the models as uninterpretable black boxes. To meet these challenges, we propose an interpretable, cross-modal attention framework for drug synergy prediction. It extracts the basic features of drugs through pre-trained chemical language models and characterizes cancer cell lines using their gene expression profiles. The features of drugs and cell lines are then fused through a hierarchical cross-modal attention mechanism. Experiments show that our approach outperforms previous baselines, particularly in discovering rare synergistic combinations on severely imbalanced datasets. Moreover, the model automatically emphasizes important chemical atoms, thereby improving the reliability of its predictions for further validation.

Keywords:

Drug synergy prediction; Cross-modal attention; Chemical language models; Explainable deep learning

1. Introduction

Combination therapy plays an important role in precision oncology and the treatment of complex diseases. By concurrently targeting multiple biological pathways, synergistic drug combinations can significantly enhance therapeutic efficacy, reduce required dosages, and mitigate the onset of drug resistance [1]. However, the vast combinatorial space of potential drugs renders exhaustive *in vitro* and *in vivo* screening intractable. In recent decades, *artificial intelligence-driven drug discovery* (AIDD) has emerged as a transformative paradigm, leveraging deep learning to predict synergistic interactions and prioritize candidates for experimental validation [2].

Despite their advances, many deep learning models for drug synergy face two critical limitations. *First*, existing learning architectures typically aggregate drug and cell-line representations in a static, context-agnostic process. They mostly rely on rigid mathematical operations, such as vector concatenation or element-wise multiplication [3]. These approaches failed to capture the interaction between two chemical entities, which is highly dynamic and profoundly modulated by the specific transcriptomic or mutational landscape of the host cell. *Second*, the pharmacological domain demands not only high accuracy but also greater interpretability of the predictions. The "black-box" nature of conventional deep neural networks (DNNs) prevents researchers from understanding the mechanistic rationale behind a positive synergy prediction. Without knowing *why* a combination works or *which* drug drives the synergistic effect in a specific cellular context, computational predictions struggle to gain trust in clinical settings or guide downstream biomarker discovery[4].

To bridge these gaps, we propose a novel, interpretable deep learning framework based on a **Cross-Modal Attention Mechanism**. Instead of relying on static feature concatenation or *post-hoc* explanation methods, our architecture models the drug-drug and drug-cell interactions as a dynamic querying process. We utilize a pre-trained chemical language model (e.g. `ChemBERTa`) to extract robust latent embeddings for each drug [5]. Crucially, we design a cross-attention module where the cellular microenvironment acts as an active query, adaptively allocating attention weights to the interactive features of two drugs. This mechanism natively generates a dynamic, instance-specific mask, quantifying the relative contribution of each drug to the synergistic outcome.

2 Related Work

Molecular Representation Learning in AIDD. Accurate and expressive molecular representation modeling is quite important in AIDD tasks. Traditional methods heavily relied on selected physicochemical descriptors or fixed-length structural fingerprints (e.g., Extended-Connectivity Fingerprints - ECFPs) [6]. In recent years, *Graph Neural Networks* (GNNs) that process molecular topological graphs [7, 8] and sequence-based *Transformers* that decode chemical strings [9] have become two dominant paradigms in this field. In particular, pre-trained chemical language models (CLMs), such as `ChemBERTa` [5] and `MolFormer` [10], have shown strong ability to capture deep semantic and structural representations through large-scale self-supervised learning, thereby reducing the need for manual feature engineering in downstream predictive tasks.

Computational Drug Synergy Prediction. Early computational approaches for drug synergy prediction often adopted traditional machine learning algorithms (e.g., Random Forests and Support Vector Machines), which suffered from limited capacity to capture non-linear, high-dimensional interactions [11]. Recent advances in deep learning have led to many new models, such as `DeepSynergy` [3] and `MatchMaker` [12], which typically integrate multimodal features of drugs and cell lines using multi-layer perceptrons (MLPs). However, the majority of these frameworks combine the drug and cell-line features through simple vector concatenation or element-wise product, failing to capture the comprehensive interplay between drug substructures in specific cell-line contexts.

Interpretability in Predictive Pharmacology. Despite the high predictive accuracy of deep learning models, their "black-box" nature remains a key barrier to clinical translation. Current attempts to interpret drug synergy prediction models mostly rely on *post-hoc* attribution methods, such as *SHapley Additive exPlanations* (SHAP) or *Integrated Gradients* [13, 14]. But these methods are computationally expensive and fail to provide the actual reasoning process. More recently, self-attention mechanisms have been utilized to highlight important molecular substructures [15, 16]. Inspired by this, we aim to develop an *intrinsic, cross-modal attention mechanism* to interpret the interplay between drugs within specific cell-line contexts. By quantifying the relative synergistic contributions of interacting drugs, our model is designed to achieve competitive predictive performance while providing instance-specific, transparent mechanistic insights.

3 Methodology

In this section, we describe the synergy prediction task and our token-level cross-attention framework. The overall architecture is illustrated in Figure 1.

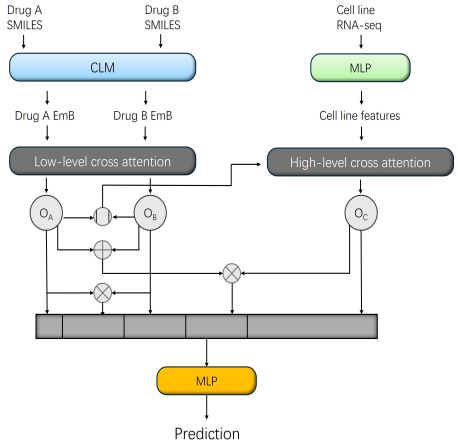

FIGURE 1. The overall architecture of the proposed learning framework. Pre-trained CLMs and MLPs extract drug and cell-line features. A hierarchical cross-attention module captures drug-drug and drug-cell line interactions.

3.1 Problem Formulation

We frame the drug synergy prediction problem as a straightforward binary classification task. The inputs of our model include a pair of drugs represented by their SMILES strings (S_A and S_B) and a target cancer cell line represented by its gene expression profile ($\mathbf{x}_c$). The model outputs a probability for the binary label $y \in \{0, 1\}$, where $y = 1$ indicates that the drugs are synergistic and $y = 0$ otherwise. Our objective is to predict whether the combination of the two drugs produces a synergistic effect on the given cell line $(S_A, S_B, \mathbf{x}_c) \rightarrow \hat{y}$.

3.2 Token-Level and Transcriptomic Feature Extraction

CLM-based Drug Feature Extraction. Unlike traditional fixed-length molecular fingerprints (e.g., ECFPs) or global GNN embeddings, the sequence representations of drugs (S_A and S_B) were extracted from pre-trained CLMs (e.g., `ChemBERTa` and `MoLformer` [5, 10]) in our work. Specifically, the SMILES sequences are tokenized and passed through the frozen pre-trained CLM encoder, yielding dense, token-

level hidden states $\mathbf{H}_A, \mathbf{H}_B \in \mathbb{R}^{L \times D_h}$, where L is the maximum sequence length and D_h represents the hidden dimension:

$$\mathbf{H}_i = \text{CLMencoder}(S_i), \quad i \in \{A, B\} \tag{1}$$

To ensure computational validity and prevent the model from attending to meaningless padded tokens in subsequent cross-attention layers, binary **attention masks** $\mathbf{M}_A, \mathbf{M}_B \in \{0, 1\}^L$ are inherently generated by the tokenizer alongside the hidden states. Here, an element value of 1 indicates a valid chemical token (allowing attention), while 0 denotes a padded token (blocking attention).

Cell Line Transcriptomic Processing and Feature Processing. To represent the cellular environment, we utilize RNA-seq expression profiles from the Cancer Dependency Map (DepMap) database[17][18]. For each cell line, its expression profile is extracted as a raw vector $\mathbf{u}_c \in \mathbb{R}^{1000}$, including the top 1000 genes with the highest variance. To ensure statistical stability in different cell lines, we apply z-score normalization to obtain the processed transcriptomic vector $\hat{\mathbf{x}}_c$ as

$$\hat{\mathbf{x}}_c = \frac{\mathbf{x}_c - \mu(\mathbf{x}_c)}{\sigma(\mathbf{x}_c) + \epsilon} \tag{2}$$

where $\mu(\cdot)$ and $\sigma(\cdot)$ denote the mean and standard deviation of the vector elements, and $\epsilon = 10^{-8}$ is a small constant used to avoid division by zero.

To make the cell-line features compatible with that of the drug features, $\mathbf{x}_c$ is projected into the semantic space with D_h dimensions. This projection consists of a two-layer MLP with Batch Normalization and ReLU non-linearity:

$$\mathbf{h}_c = \mathbf{W}_2 \cdot \sigma_{ReLU}\left(\text{BN}(\mathbf{W}_1 \hat{\mathbf{x}}_c + \mathbf{b}_1)\right) + \mathbf{b}_2 \tag{3}$$

where $\mathbf{W}_1$, $\mathbf{W}_2$, $\mathbf{b}_1, \mathbf{b}_2$ are learnable weights or biases, and σ_{ReLU} denotes the ReLU activation function. The resulting vector $\mathbf{h}_c$ serves as the cellular context features for the subsequent layers.

3.3 Hierarchical Multi-Modal Cross-Attention

Instead of simply concatenating drug and gene expression features, our model uses a hierarchical multi-modal cross-attention architecture to fuse them. This architecture captures the complex dependencies between drug pairs and cancer cell lines.

Low-level: drug and drug interaction. In our model, we apply the multi-head cross-attention layers to capture the interaction inside a drug pair (drug A and drug B). The feature embedding of each drug serves as the `Query` ($\mathbf{Q}$), attending to the `Key` ($\mathbf{K}$) and `Value` ($\mathbf{V}$) embeddings of the other drug.

$$\mathbf{O}_A, \boldsymbol{\alpha}_{A \to B} = \text{LayerNorm}\big(\mathbf{H}_A + \text{MHA}(\mathbf{Q} = \mathbf{H}_A, \mathbf{K} = \mathbf{H}_B, \mathbf{V} = \mathbf{H}_B)\big), \tag{4}$$

$$\mathbf{O}_B, \boldsymbol{\alpha}_{B \to A} = \text{LayerNorm}\big(\mathbf{H}_B + \text{MHA}(\mathbf{Q} = \mathbf{H}_B, \mathbf{K} = \mathbf{H}_A, \mathbf{V} = \mathbf{H}_A)\big). \tag{5}$$

This stage contextualizes each drug's sub-structures based on the presence of the partner drug.

High-Level: Cell and Drug Contextualization. We concatenate the updated drug embeddings to form a joint drug context $\mathbf{H}_{ctx} = [\mathbf{O}_A \parallel \mathbf{O}_B] \in \mathbb{R}^{2L \times D_h}$. To project the cellular microenvironment into the attention computation, the dense cell line vector $\mathbf{h}_c \in \mathbb{R}^{D_h}$ is explicitly expanded to $\mathbb{R}^{1 \times D_h}$, acting as a global `Query` to dynamically contextualize and highlight relevant pharmacological patterns from the concatenated `Key` and `Value` embeddings of the two drugs:

$$\mathbf{O}_C, \boldsymbol{\alpha}_{C \to drugs} = \text{MHA}(\mathbf{Q} = \mathbf{h}_c, \mathbf{K} = \mathbf{H}_{ctx}, \mathbf{V} = \mathbf{H}_{ctx}). \tag{6}$$

The resulting output $\mathbf{O}_C \in \mathbb{R}^{1 \times D_h}$ is subsequently squeezed back to $\mathbb{R}^{D_h}$, capturing the cell-line-specific synergistic representation.

Explicit Feature Fusion. Based on the representative drug features $O_A, O_B \in \mathbb{R}^{D_h}$, we calculate two interaction terms to help the model decipher synergistic mechanisms: an inter-drug feature $z_{syn} = O_A \odot O_B$ and a drug-cell interaction feature $z_{int} = (O_A + O_B) \odot O_C$, where $\odot$ denotes the element-wise product.

Unlike the standard concatenation approach, we employ an additive fusion strategy to integrate all latent representations into a single holistic representation:

$$z_{fused} = O_A + O_B + O_C + z_{syn} + z_{int} \tag{7}$$

This vector $z_{fused} \in \mathbb{R}^{D_h}$ is finally processed by an MLP with Batch Normalization and Dropout to generate the synergy prediction probability $\hat{y}$.

3.4 Permutation Invariance and Optimization

Permutation Invariance. Considering the biological symmetry of synergistic effects, the order of drugs A and B should not affect the prediction score [3]. Thus, for each drug pair (S_A, S_B) in the training set, we add its swapped counterpart (S_B, S_A) with the same label.

Optimization with Entropy Regularization. We used a weighted Binary Cross-Entropy (BCE) loss to mitigate the class imbalance problem in drug synergy datasets. In addition, to improve interpretability and avoid attention collapse, we apply a Joint Information Entropy Penalty to the attention distributions ($\boldsymbol{\alpha}_{A\to B}$, $\boldsymbol{\alpha}_{B\to A}$, and $\boldsymbol{\alpha}_{C\to drugs}$). This penalty ensures that the model focuses on specific and important molecular structures, rather than distributing its attention equally across all parts. The total loss function is calculated as follows:

$$\begin{aligned}\mathcal{L}_{total} = \mathcal{L}_{BCE}(\hat{y}, y, w_{pos}) + \lambda_{cell}\mathcal{H}(\boldsymbol{\alpha}_{C\to drugs}) \\ + \lambda_{drug}\big(\mathcal{H}(\boldsymbol{\alpha}_{A\to B}) + \mathcal{H}(\boldsymbol{\alpha}_{B\to A})\big),\end{aligned} \tag{8}$$

where w_{pos} is the positive class weight, $\mathcal{H}(\cdot)$ denotes the calculation of information entropy over the attention probabilities, and λ_{cell} and λ_{drug} are hyperparameters controlling the sparsity of biological targeting and chemical interactions, respectively.

4 Experiment

To rigorously evaluate the proposed framework, we benchmarked our model on standardized repositories including **DrugCombDB** [20]and the **Sanger Institute combination dataset**[19]. All experiments are conducted within a stratified 5-fold cross-validation protocol.

4.1 Experiment Setup

Datasets. We carried out experiments on two large-scale cancer benchmarks: (1) **DrugCombDB** [20], from which we used a high-quality subset of 2,353 unique drug combinations across 73 human cancer cell lines; and (2) the **Sanger Institute dataset** [19], which includes 2,025 experimentally proven combinations across 125 molecularly pancreatic cancer cell lines.

Evaluation Metrics. The predictive performance of the model is comprehensively evaluated using the Area Under the Receiver Operating Characteristic curve (ROC-AUC), Precision-Recall AUC (PR-AUC / AUPR), Accuracy (ACC), and F1-score. Given the severe class imbalance inherent in experimental drug synergy screening, we highlight **PR-AUC** as the primary evaluation metric, as it provides a more sensitive and reliable assessment for the minority positive class.

Implementation Details. We used the AdamW optimizer[21] with a learning rate of 10^{-4} and a weight decay of 10^{-2}. The batch size was set to 256. To ensure stable training with the entropy penalty, we used a linear warm-up strategy for the first 10 epochs, gradually increasing the penalty weights λ_{cell} and λ_{drug} to their maximum values (0.001 and 0.0005, respectively). We employed a stratified 5-fold cross-validation to evaluate the model performance. To prevent overfitting, early stopping is applied if the ROC-AUC on the validation set does not improve for 10 consecutive epochs.

4.2 Result

The performance of our proposed framework was benchmarked against established architectures (DeepSynergy[3], DeepDDS [22], and SDCInterpreter[23]). The comparative results are detailed in Table 1.

TABLE 1. Performance comparison of different models

DrugCombDB				
Model	**ROC AUC**	**PR AUC**	**F1**	**ACC**
DeepSynergy	0.6109	0.5676	0.4817	0.5926
SDCInterpreter	0.8309	**0.8536**	**0.7984**	0.7477
DeepDDS	0.7256	0.5677	0.3828	0.7502
Ours[1]	0.8390	0.8210	0.7330	0.7590
Ours[2]	**0.8448**	0.8220	0.7425	**0.7637**
Sanger Institute Dt.				
Model	**ROC AUC**	**PR AUC**	**F1**	**ACC**
DeepSynergy	0.7512	0.1856	0.1834	**0.8845**
SDCInterpreter	0.8023	0.3156	0.2658	0.8623
DeepDDS	0.7845	0.2634	0.2412	0.8512
Ours[1]	0.8115	0.3340	0.2915	0.8522
Ours[2]	**0.8345**	**0.3872**	**0.3215**	0.8436

[1] ChemBERTa is used as the backbone architecture for our model.
[2] MolFormer is used as the backbone architecture for our model.

For the **DrugCombDB** dataset, our MolFormer-based model showed competitive performance, with a ROC-AUC of 0.8448 and an Accuracy of 0.7637, outperforming most of the baseline models. Although SDCInterpreter achieved slightly higher

PR-AUC, our model maintained a more balanced F1-score (0.7425).

The **Sanger Institute dataset** is more challenging because of the much higher imbalance rate (rare synergistic pairs). Accordingly, the PR-AUC and F1 scores of all models dropped significantly. For example, DeepSynergy had a high accuracy (0.8845) but a poor PR-AUC (0.1856). In comparison, our MolFormer-based model dealt with this high imbalance quite effectively. It earned the highest PR-AUC (0.3872) and F1-score (0.3215). This shows that our cross-attention strategy can successfully detect rare synergistic signals that other models miss.

4.3 Interpretability

Our cross-modal framework offers natural interpretability. The model, driven by entropy regularization, produces highly sparse distributions while strongly reducing widespread organic noise to emphasize essential synergistic factors.

Identifying a Flexible Synergistic Partner. Figure 2 shows that `[As+3]` and `[O-2]` have a significant impact on token-level attention. Chemically, these tokens correspond to arsenic trioxide (As_2O_3), which is commonly regarded as a highly versatile "partner drug" in the field of clinical. It has been shown to generate synergistic combinations with a large array of varied treatment agents—ranging from classic chemotherapies (e.g., cisplatin, 5-FU) to current targeted inhibitors—across various cancer types [24, 25]. The attention paid to this substructure demonstrates that our data-driven algorithm may ignore weak interactions in order to identify strong synergistic partners.

5. Conclusions

In this study, we presented an interpretable, cross-modal attention framework for drug synergy prediction. By using RNA-seq expression profiles as an active query to explain the token-level chemical language representations of interacting drugs, our approach mitigates the "black-box" limitations of most deep learning models. Preferable prediction performance of this framework was validated on two well-acknowledged benchmarks, **DrugCombDB** and **Sanger Institute Datasets**. Importantly, using information entropy regularization, our model filters background noise to discover the synergistic drugs. In future work, we aim for a greater level of token-to-gene interpretability, specifically mapping which molecular tokens within a medication pair are most sensitive to the expression levels of individual genes.

Acknowledgements

This work is supported by Hong Kong Research Grants Council (Project UGC/FDS16/E11/25).

References

[1] Mokhtari, R. B. et al. "Combination therapy in combating cancer," *Oncotarget*, vol. 8, no. 23, pp. 38022–38043, 2017.

[2] Paul, D. et al. "Artificial intelligence in drug discovery and development," *Drug Discovery Today*, vol. 26, no. 1, pp. 80–93, 2021.

[3] Preuer, K., Lewis, R. P., Hochreiter, S., Bender, A., Bulusu, K. C., and Klambauer, G. "DeepSynergy: predicting anti-cancer drug synergy with Deep Learning," *Bioinformatics*, vol. 34, no. 9, pp. 1538–1546, 2018.

[4] Jiménez-Luna, J., Grisoni, F., and Schneider, G. "Drug discovery with explainable artificial intelligence," *Nature Machine Intelligence*, vol. 2, no. 10, pp. 573–584, 2020.

[5] Chithrananda, S., Grand, G., and Ramsundar, B. "ChemBERTa: large-scale self-supervised pretraining for molecular property prediction," *arXiv preprint arXiv:2010.09885*, 2020.

[6] Rogers, D., and Hahn, M. "Extended-connectivity fingerprints," *Journal of Chemical Information and Modeling*, vol. 50, no. 5, pp. 742–754, 2010.

[7] Duvenaud, D. K. et al. "Convolutional networks on graphs for learning molecular fingerprints," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 28, 2015.

[8] Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., and Dahl, G. E. "Neural message passing for quantum chemistry," in *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pp. 1263–1272, 2017.

[9] Vaswani, A. et al. "Attention is all you need," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017.

[10] Ross, J. et al. "Large-scale chemical language representations capture molecular structure and properties," *Nature Machine Intelligence*, vol. 4, no. 12, pp. 1256–1264, 2022.

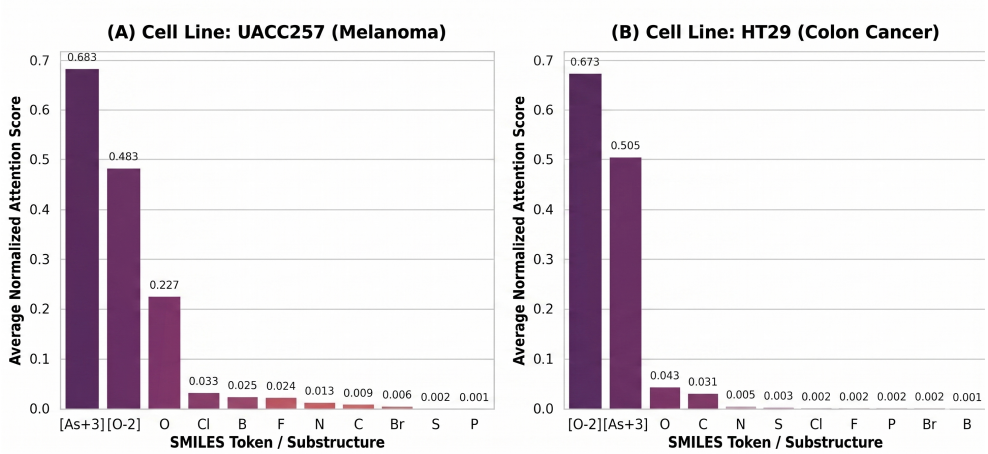

FIGURE 2. Cell-line-specific adaptation of token-level attention weights. The bar charts present the most attended SMILES tokens for **(A) UACC257 (Melanoma)** and **(B) HT29 (Colon Cancer)**. The attention is highly concentrated on `[As+3]` and `[O-2]`, separately identifying the potent Arsenic Trioxide (As_2O_3) synergistic drugs.

[11] Wildenhain, J. et al. "Prediction of synergism from chemical-genetic interactions by machine learning," *Cell Systems*, vol. 1, no. 6, pp. 383–395, 2015.

[12] Kuru, H. I., Tastan, O., and Cicek, A. E. "MatchMaker: A Deep Learning Framework for Drug Synergy Prediction," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 19, no. 4, pp. 2334–2344, 2022.

[13] Lundberg, S. M., and Lee, S. I. "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017.

[14] Sundararajan, M., Taly, A., and Yan, Q. "Axiomatic attribution for deep networks," in *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pp. 3319–3328, 2017.

[15] Maziarka, Ł. et al. "Molecule attention transformer," *arXiv preprint arXiv:2002.08264*, 2020.

[16] Lee, K. S., Chen, J., Chung, F. S. F., Zhao, T., Chen, Z., and Wang, D. D. "Task-Specific Sparse Feature Masks for Molecular Toxicity Prediction with Chemical Language Models," *arXiv preprint arXiv:2512.11412*, 2025.

[17] E. Gonçalves, R. C. Poulos, Z. Cai, et al., "Pan-cancer proteomic map of 949 human cell lines", *Cancer Cell*, vol. 40, no. 8, pp. 835-849.e8, 2022.

[18] Tsherniak, A. et al. "Defining a cancer dependency map," *Cell*, vol. 170, no. 3, pp. 564–576, 2017.

[19] Jaaks, P. et al. "Effective drug combinations in breast, colon and pancreatic cancer cells," *Nature*, vol. 603, no. 7899, pp. 166–173, 2022.

[20] Liu, H. et al. "DrugCombDB: a comprehensive database of drug combinations toward the discovery of combinatorial therapy," *Nucleic Acids Research*, vol. 48, no. D1, pp. D871–D881, 2020.

[21] Loshchilov, I., and Hutter, F. "Decoupled weight decay regularization," in *Proceedings of the 7th International Conference on Learning Representations (ICLR)*, 2019.

[22] Wang, J. et al. "DeepDDS: deep graph neural network with attention mechanism to predict synergistic drug combinations," *Briefings in Bioinformatics*, vol. 23, no. 1, p. bbab390, 2022.

[23] Wang, S. et al. "Path-Based Graph Neural Network for Drug Synergy Prediction and Interpretation," *Journal of Chemical Information and Modeling*, vol. 66, no. 1, pp. 337–348, 2026.

[24] Khademi, F. et al. "Therapeutic strategy of arsenic trioxide in the fight against cancers and other diseases," *Metallomics*, vol. 12, no. 5, pp. 682–693, 2020.

[25] Chow, S. K. et al. "Arsenic trioxide as a broad-spectrum sensitizer for solid tumors," *International Journal of Molecular Sciences*, vol. 20, no. 21, p. 5262, 2019.