DIFFGRAPHTRANS: A DIFFERENTIAL ATTENTION-BASED APPROACH FOR EXTRACTING MEANINGFUL FEATURES OF DRUG COMBINATIONS

Bingzheng Wu
Yantai Institute
China Agricultural University
Yantai 264670, Peoples R China
bzwoo@cau.edu.cn

Qi Wang*
Collage of Science
China Agricultural University
Beijing 100083, Peoples R China
wangqi_math@cau.edu.cn

1 INTRODUCTION

Drug combination therapy offers advantages like synergy and reduced resistance (Tyers & Wright, 2019), but traditional experimental methods are costly and slow. Computational approaches improve efficiency but lack explainability. In recent years, computational methods have boosted the efficiency of prediction (Madani Tonekaboni et al. (2018); Hu et al. (2023); Rafiei et al. (2023); Wang et al. (2025)), but their explainability still faces severe challenges (Azuaje (2017); Tsigelny (2019); Chen & Huang (2024)).

Most modern computational methods use the molecular structure of a drug within a drug combination to build a graph, whose each node represents an atom while each edge represents a bond. Models based on graph neural networks, such as DeepDDS (Wang et al., 2022) and ACDGNN(Yu et al., 2023), capture features of the molecular structure. Meanwhile, Transformer-based methods use the multi-head attention mechanism to structure the relations between atoms. However, due to molecular graph heterogeneity, existing transformers amplify high-frequency noise and suppress low-frequency signals critical to functional groups (Zhao et al. (2021); Shui & Karypis (2020)).

In response to these problems, we proposed the Differential Graph Transformer (DiffGraphTrans), a framework that combines the differential filter method and the multi-head attention mechanism. This model achieves the effective extraction and representation learning of molecular features by eliminating noise in the calculation process of attention scores based on the parameter λ, realizes the conservation of computing resources as much as possible while ensuring performance, and provides an embedding with a biochemical basis. We believe that this is a key step towards interpretable artificial intelligence-driven drug discovery.

2 METHODS

We used lung cancer-related effective drug combinations from DCDB 2.0 (Liu et al., 2014), C-DCDB (Shtar et al., 2022) and DrugMAP 2.0 (Li et al., 2025) as the train and test data, which are publicly available. We retrieved molecular structures in SMILES (Simplified molecular input line entry system) format (Weininger, 1988). We used RDKit (Landrum, 2013) to obtain the number of non-hydrogen atoms of the molecule and their spatial relationships to construct the graph data, where the atoms are represented as the nodes and the edges represents bonds, with node features (atom type, hydrogen count, etc.) encoded via one-hot vectors. The features N of all nodes in the graph were stored in a tensor x and served as the input of the attention module.

The differential multi-head attention computes two sets of queries Q, keys K, and values V for the features of each node on the graph, and calculates the difference between the two sets of Q, K, and V through a learnable parameter λ (Ye et al., 2024). Among them, the lambdas of these three are denoted as λ_q, λ_k and λ_v respectively, where the λ is used to ensure non-linear scaling, allowing our model to amplify meaningful attention differences while suppressing noise, like noise-canceling headphones and differential amplifiers. More details about the differential multi-head attention are

*Corresponding author.

introduced in more detail in the appendix. The tensor x whose shape is determined by the number of nodes in the graph and the dimension of the node feature vector, stored features of nodes, and is input into the multi-head attention module. Subsequently, the tensors calculated from the features of two drugs in a combination are added, as the embedding of this combination, passed through a three-layer fully connected layer, and the result is output. The structure of the entire model is shown in Figure 1, and the code is now available at `https://github.com/Endurernura/Differential-Graph-Transformer`.

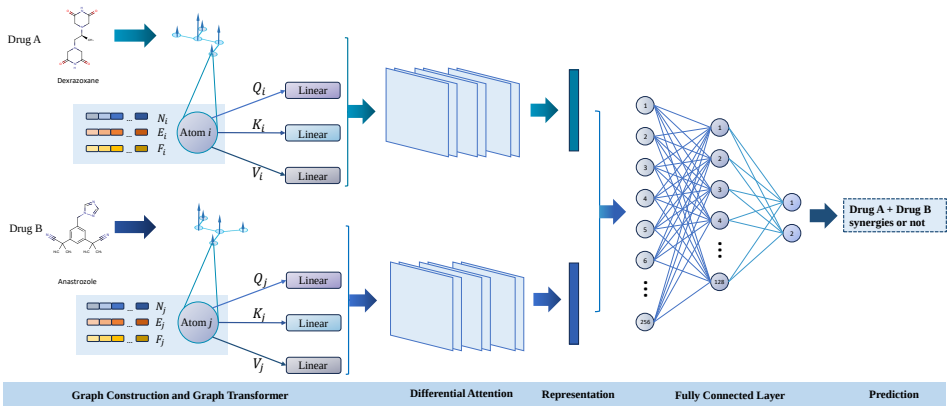

Figure 1: The structure of DiffGraphTrans.

The differential attention mechanism was first proposed in October 2024. Its earliest and only application was in natural language processing tasks. It was used to reduce the "hallucination phenomenon" of large language models and improve the accuracy of Transformers in time series prediction tasks. We noticed that the heterogeneity of molecular graphs makes it equally prone to cause the multi-head attention mechanism to amplify invalid information in molecules, increase the weight of meaningless nodes, and thereby lead to a decline in its performance. In the main drug combination databases, the number of drug combinations for a certain disease is often relatively small. When the model pays insufficient attention to key functional groups , its performance in the prediction task will deteriorate. We processed the original graph data, artificially changed the positions of some bonds and functional groups, and conducted a new round of training to simulate the situation when the model is used to develop new drugs or verify whether drugs or combinations that do not exist in the data set are synergistic. We compared our model with other methods, such as Transformer, Support Vector Machine(SVM), Random Forest(RF), and XGBoost. The experimental results are shown in Table 1.

3 Discussion and Further Work

The results of the experiment prove that the multi-head differential attention module is crucial to enhance the model's performance. Generally speaking, graph-based neural networks and drug molecular structures only serve as a part in predicting drug combinations or drug-drug interaction events. DiffGraphTrans addresses the high - noise problem in molecular representation under few - shot scenarios through differential attention, offering new ideas for the generalization of biological data. Future work will integrate multi-omics data (e.g., genomic profiles) and extend to multi-drug predictions via hypergraphs.

Meaningfulness Statement

A "meaningful representation of life" in drug discovery requires embeddings that reflect biochemical mechanisms and functional group interactions. Our work contributes by integrating differential attention into graph Transformers, which dynamically suppresses molecular noise and amplifies key functional groups (e.g., hydroxyl bonds). This enhances both prediction accuracy and interpretability, bridging AI-driven models with actionable biological insights. By linking attention weights to

biochemical relevance, DiffGraphTrans advances the design of transparent AI systems for precision medicine.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the College of Science at China Agricultural University for their support of this work.

REFERENCES

Francisco Azuaje. Computational models for predicting drug responses in cancer research. *Briefings in bioinformatics*, 18(5):820–829, 2017.

Xing Chen and Li Huang. Computational model for drug research. *Briefings in Bioinformatics*, 25 (3):bbae158, 2024.

Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in Neural Information Processing Systems*, 35:16344–16359, 2022.

Jie Hu, Xiaozhi Zhang, Desi Shang, Lijun Ouyang, Yue Li, and Dongping Xiong. Egtsyn: edge-based graph transformer for anti-cancer drug combination synergy prediction. *arXiv preprint arXiv: 230310312*, 2023.

Greg Landrum. Rdkit documentation. *Release*, 1(1-79):4, 2013.

Fengcheng Li, Minjie Mou, Xiaoyi Li, Weize Xu, Jiayi Yin, Yang Zhang, and Feng Zhu. Drugmap 2.0: molecular atlas and pharma-information of all drugs. *Nucleic Acids Research*, 53(D1): D1372–D1382, 2025.

Yanbin Liu, Qiang Wei, Guisheng Yu, Wanxia Gai, Yongquan Li, and Xin Chen. Dcdb 2.0: a major update of the drug combination database. *Database*, 2014, 2014.

Seyed Ali Madani Tonekaboni, Laleh Soltan Ghoraie, Venkata Satya Kumar Manem, and Benjamin Haibe-Kains. Predictive approaches for drug combination discovery in cancer. *Briefings in bioinformatics*, 19(2):263–276, 2018.

Tony S Mok, Yi-Long Wu, Myung-Ju Ahn, Marina C Garassino, Hye R Kim, Suresh S Ramalingam, Frances A Shepherd, Yong He, Hiroaki Akamatsu, Willemijn SME Theelen, et al. Osimertinib or platinum–pemetrexed in egfr t790m–positive lung cancer. *New England Journal of Medicine*, 376(7):629–640, 2017.

Fatemeh Rafiei, Hojjat Zeraati, Karim Abbasi, Jahan B Ghasemi, Mahboubeh Parsaeian, and Ali Masoudi-Nejad. Deeptrasynergy: drug combinations using multimodal deep learning with transformers. *Bioinformatics*, 39(8):btad438, 2023.

Guy Shtar, Louise Azulay, Omer Nizri, Lior Rokach, and Bracha Shapira. Cdcdb: A large and continuously updated drug combination database. *Scientific data*, 9(1):263, 2022.

Zeren Shui and George Karypis. Heterogeneous molecular graph neural networks for predicting molecule properties. In *2020 IEEE International Conference on Data Mining (ICDM)*, pp. 492–500. IEEE, 2020.

Jean-Charles Soria, Yuichiro Ohe, Johan Vansteenkiste, Thanyanan Reungwetwattana, Busyamas Chewaskulyong, Ki Hyeong Lee, Arunee Dechaphunkul, Fumio Imamura, Naoyuki Nogami, Takayasu Kurata, et al. Osimertinib in untreated egfr-mutated advanced non–small-cell lung cancer. *New England journal of medicine*, 378(2):113–125, 2018.

Martin Steins, Michael Thomas, and Michael Geißler. Erlotinib. In *Recent Results in Cancer Research*, Recent results in cancer research. Fortschritte der Krebsforschung. Progres dans les recherches sur le cancer, pp. 1–17. Springer International Publishing, Cham, 2018.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Igor F Tsigelny. Artificial intelligence in drug combination therapy. *Briefings in bioinformatics*, 20 (4):1434–1448, 2019.

Mike Tyers and Gerard D Wright. Drug combinations: a strategy to extend the life of antibiotics in the 21st century. *Nature Reviews Microbiology*, 17(3):141–155, 2019.

Jinxian Wang, Xuejun Liu, Siyuan Shen, Lei Deng, and Hui Liu. Deepdds: deep graph neural network with attention mechanism to predict synergistic drug combinations. *Briefings in Bioinformatics*, 23(1):bbab390, 2022.

Qi Wang, Xiya Liu, and Guiying Yan. Predicting effective drug combinations for cancer treatment using a graph-based approach. *Synthetic and Systems Biotechnology*, 10(1):148–155, 2025.

Beth A Weaver. How taxol/paclitaxel kills cancer cells. *Molecular biology of the cell*, 25(18): 2677–2681, 2014.

David Weininger. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of chemical information and computer sciences*, 28(1):31–36, 1988.

Tianzhu Ye, Li Dong, Yuqing Xia, Yutao Sun, Yi Zhu, Gao Huang, and Furu Wei. Differential transformer. *arXiv preprint arXiv:2410.05258*, 2024.

Hui Yu, KangKang Li, WenMin Dong, ShuangHong Song, Chen Gao, and JianYu Shi. Attention-based cross domain graph neural network for prediction of drug–drug interactions. *Briefings in Bioinformatics*, 24(4):bbad155, 2023.

Biao Zhang and Rico Sennrich. Root mean square layer normalization. *arXiv preprint arXiv:1910.07467*, 2024.

Jianan Zhao, Xiao Wang, Chuan Shi, Binbin Hu, Guojie Song, and Yanfang Ye. Heterogeneous graph structure learning for graph neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 4697–4705, 2021.

A Appendix

A.1 Differential Transformer

The differential attention is defined as

$$DiffAttn(X) = (softmax(\frac{Q_1 K_1^T}{\sqrt{d_n}}) - \lambda \cdot softmax(\frac{Q_2 K_2^T}{\sqrt{d_n}}))V$$

where λ is parameterized by

$$\lambda = exp(\lambda_{q1} \cdot \lambda_{k1}) - exp(\lambda_{q2} \cdot \lambda_{k2}) + \lambda_{init}$$

and the multi-head differential attention is defined as

$$head_i = DiffAttn(X; W_i^q, W_i^K, W_i^V, \lambda)$$

$$\overline{\mathrm{head_i}} = (1 - \lambda_{init}) \cdot LN(head_i)$$

$$Multihead(X) = Concat(\overline{\mathrm{head_1}}, \ldots, \overline{\mathrm{head_n}}) \cdot W^O$$

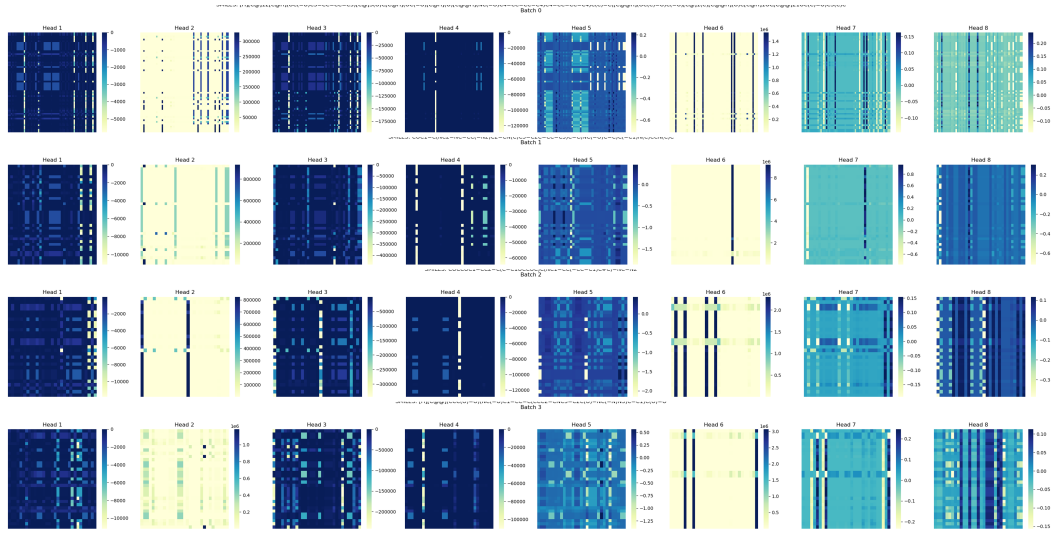

Figure 2: The differential attention heatmaps of drugs.

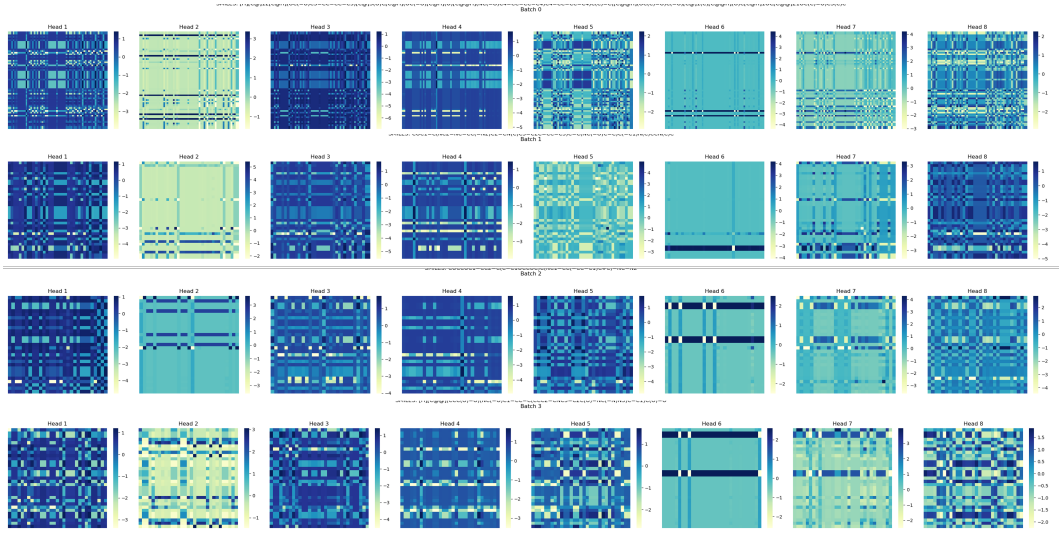

Figure 3: The differential attention heatmaps of drugs, when attention scores are standardized.

where W^O is a learnable projection matrix, and the $LN(\cdot)$ uses RMSNorm (Zhang & Sennrich, 2024) for each head to normalize.

In the model presented in the original "Differential Transformer", rotational position encoding (RoPE,Su et al. (2024)) and optimization methods including Apex and flash attention(Dao et al., 2022) were used. Given that there is no problem of very long sequences in our prediction task, in order to reduce the complexity of the model, compared to the traditional Transformer structure, we only modified the attention part and calculated the attention score according to the differential multi-head attention method proposed in the article.

To present the performance of the DiffGraphTrans, We selected four drugs, namely Paclitaxel, Olmutinib, Erlotinib, and Pemetrexed, respectively, as demonstrations to showcase the model's performance in feature extraction. The attention heatmaps are shown in Figure 2. Also, we conducted a z-score standardization on the attention scores of the drugs. The heatmaps of the standardized scores are presented as Figure 3. From the top to the bottom of the figures 2 and 3, they represent Paclitaxel, Olmutinib, Erlotinib, and Pemetrexed respectively.

Palitaxel is a complex drug, which can only be derived from nature before, and can now be fully synthesized artificially. Its structure contains a special part called Taxane Ring, and a hydroxyl group ad a carbamate group, which are on the side chain and can form hydrogen bonds with proteins. Therefore, the paclitaxel can bind to β - tubulin and form hydrogen bonds, inhibiting its function, resulting in failed chromosome segregation and apoptosis (Weaver, 2014).

Olmutinib is a third - generation epidermal growth factor receptor - tyrosine kinase inhibitor (EGFR - TKI), belonging to the same type of drugs as afatinib. Its chemical structure contains a Pyrimidoindole Core and an Acrylamide Group, enabling it to interact with EGFR, inhibit the kinase activity of EGFR mutants, and block the signaling pathway (Soria et al., 2018) .

Erlotinib is a first - generation epidermal growth factor receptor - tyrosine kinase inhibitor (EGFR - TKI). Its chemical structure contains a Quinazoline Core, allowing it to competitively bind to the ATP - binding site of the EGFR kinase, blocking the signaling pathway and inhibiting the proliferation of tumor cells. It is worth noting that there are certain differences between its structure and that of olmutinib (Steins et al., 2018) .

Pemetrexed is a multi-target, anti-metabolic and anti-tumor drug. Its structure contains the Pyrrolo[2,3 - d]pyrimidine and Glutamate moiety. It can interact with enzymes and targets such as thymidylate synthase (TS), dihydrofolate reductase (DHFR), and glycinamide ribonucleotide formyltransferase (GARFT), and exhibits broad spectrum anti - tumor activity (Mok et al., 2017).

A.2 DATABASE

In this article, we chose Drug Combination Database 2.0 (DCDB 2.0), Continuous-Drug Combination Database (C-DCDB), and DrugMAP 2.0 as the training and validation databases. Since all the provided drug combinations in these databases are synergistic drug combinations, we manually selected the drug combination methods that did not exist in all the databases as artificial negative samples for model training. A total of approximately 350 artificial negative samples were added, accounting for approximately 20% of the total number of samples.

The DCDB is the first database dedicated to collecting and organizing information on drug combinations, containing 1,363 drug combinations involving 1,735 individual drugs. This study focuses on the subset of drug combinations related to lung cancers in the DCDB. We excluded the drug combinations with the classification of "Effective type: Need further study" or containing the classification of "Not approved" in the DCDB because the experimental results of these combinations are unclear.

The C-DCDB encompasses over 30,000 drug combinations and 4,000 related drugs, and this database is updated in real time. The data we used was downloaded on October 9, 2024. The data in the database mainly come from ClinicalTrials.gov, the FDA Orange Book, and patent literature. In our research, we selected the drug combinations which treats lung cancer and include 395 combinations and 119 different kinds of drugs.

The DrugMAP 2.0 (DrugMAP 2025) is a comprehensive platform designed to map drug combinations within biological networks, with a particular emphasis on drug combination therapies. This database contains more than 20,831 drug combinations covering various diseases and is updated regularly. The data we used was updated in January 2025. In this study, we selected 1,543 drug combinations related to the treatment of lung cancer, which included 111 different drugs. DrugMAP 2.0 offers extensive experimental data on drug synergy, molecular mechanisms, and the efficacy of various drug combinations, making it a valuable resource for both research and clinical applications.

A.3 DESCRIPTION OF TRAINING PARAMETERS

During the Transformer and DiffGraphTrans training process, we trained 40 epochs for each model, using 1800 drug combinations,and set the parameters as: head=8,learning rate=0.01, and the initial learning rate is 0.01, adjusted in real time. The batch size for mini batch training is 32, and the dimension of the hidden layer is 32. Specifically, we set the λ_{init} of DffGraphTrans as 0.8.

For the setting of the λ_{init} ,we tried several different parameters. It is certainly that when the λ_{init} is set as 0.8, the model has the best performance. It is worth noting that this is consistent with the performance of the differential transformer in the original article.

Table 1: Experimental outcomes with/without noise

Condition	Model	ROC AUC	PR AUC	ACC	BACC	PREC	TPR	F1
No Noise	Ours	**0.9051**	**0.9612**	**0.8796**	0.7463	0.8801	0.9666	0.9210
	Transformer	0.9011	0.9589	0.8793	0.7486	0.8797	**0.9803**	**0.9269**
	RF	0.7932	0.8910	0.8740	0.7932	0.8983	0.9436	0.9201
	SVM	0.8553	0.9188	0.8520	**0.7994**	0.9094	0.8968	0.9030
	XGBoost	0.8888	0.8907	0.7476	0.7270	**0.9155**	0.7598	0.8293
With Noise	Ours	**0.8932**	**0.9580**	**0.8787**	**0.7471**	**0.8779**	**0.9817**	**0.9265**
	Transformer	0.8806	0.9544	0.8645	0.7324	0.8729	0.9747	0.9206
	RF	0.7424	0.8005	0.8440	0.6762	0.8742	0.9315	0.9017
	SVM	0.6223	0.8512	0.7407	0.6722	0.8598	0.8124	0.8350
	XGBoost	0.6650	0.7601	0.7368	0.6550	0.7690	0.7063	0.7732

A.4 EXTENDED ANALYSES

In our experiments, the predictive performance of the DiffGraphTrans compared to the Transformer, Support Vector Machine(SVM), Random Forest(RF), and XGBoost in two environments, presented as Table 1, clearly demonstrate the differences in performance between scenarios with and without noise. In table 1, we represent the area under the receiver operating characteristic curve using ROC AUC, represent the area under the precision curve using PR AUC, represent the accuracy and balanced accuracy using ACC and BACC, represent precision using PREC, represent true positive rate or recall using TPR. F1 is the harmonic mean of precision and recall.

In the noise experiment, we randomly deleted 30% of the points and the connected edges in the graph. For some special chemical structures, such as benzene rings, we chose to skip them in the deletion method to ensure that the chemical formula after the operation conformed to the requirements of the kekule formula.