A Foundational Multi-Modal Knowledge Graph for Pancreatic Cancer Drug Effects Prediction

Jingwen Hui
Shu Chien-Gene Lay Department of Bioengineering
University of California San Diego
La Jolla, CA 92093
jwhui@ucsd.edu

Shengchao Liu*
Independent
shengchao1224@gmail.com

Xiaohua Huang*
Shu Chien-Gene Lay Department of Bioengineering
University of California San Diego
La Jolla, CA 92093
x2huang@ucsd.edu

Anima Anandkumar*
Computing + Mathematical Sciences
California Institute of Technology
Pasadena, CA 91125
anima@caltech.edu

Abstract

AI-assisted drug discovery has revolutionized healthcare by accelerating virtual screening methods as compared to traditional processes. Many advanced AI models have been developed to predict and generate drug candidates, with potential applications across various diseases. However, challenges still remain in applying AI models in clinical settings. These include the lack of heterogeneity and insufficient consideration of patient-specific treatment plans. To mitigate these challenges, we propose PanRX, a cell-line-specific pancreatic cancer drug effect model using multi-modal knowledge graphs. It aims at achieving a personalized drug discovery framework by incorporating rich genetic and chemical information. We first construct a multi-modal knowledge graph dataset PanCan-DrugsGenes. It extracts textual genetic information from NCBI, mutation status from the Genomics of Drug Sensitivity in Cancer (GDSC) dataset, textual descriptions of drugs from PubChem, and chemical geometry from the PM6 dataset. Then, PanRX utilizes a geometric model to learn chemical conformation, a language model to learn textual description, and a graph neural network to fuse all information and predict the target drug effects. We verify the effectiveness of PanRX by achieving the generalization performance with very low MSE (< 0.0000) and MAE (0.0009). This work emphasizes the potential of merging knowledge graphs and deep learning in the fields of genomics and medicine, enriching the intersection of human biological expertise and AI in drug discovery and design tasks.

1 Introduction

Personalized medicine is becoming the pillar of modern medicine because it considers varying phenotypes of diseases as a result of differences in genetic information. Genetic information, including gene and copy number alterations (CNA) mutation status, gene regulatory networks, gene function, and gene location are critical information that collectively determine the wellbeing of individuals [1, 2]. Considering these factors is essential to understanding cancer mechanisms and therapeutic requirements in addition to the chemical/medicinal properties of various drugs. Databases

*Equal advising author

Foundation Models for Science Workshop,38th Conference on Neural Information Processing Systems (NeurIPS 2024).

like GDSC [3] have collected drug effect experiments and genomic association tests on over 1,000 cancer cell lines. However, predicting drug effects given genetic information is still limited by data insufficiency and a lack of insights into the genetic circuit. Nevertheless, deep learning (DL) tools can integrate multi-modal information from all human knowledge and are expected to generalize to unseen tasks. They can be adapted to uncover hidden relationships of the genetic circuit to generalize partial data and offer robust predictions efficiently compared to human methods.

Recent breakthroughs in deep learning have accelerated the drug discovery process due to its ability for pattern recognition. Among these, conventional (transductive) knowledge graphs (KG) gained wide attraction in computational biology due to their network-oriented structure [4, 5, 6]. Transductive KG works with a fixed set of nodes and edges. In the biological context, transductive KG is constructed only with known entities (e.g. genes, proteins, diseases) and their interactions (e.g. genetic pathways, gene-protein interactions, gene-disease associations). Existing DL pipelines for drug discovery predominantly use SMILES, 2D molecular graphs, and conformation data of chemical compounds [7]. However, the complexity of the interactions within the living systems due to biophysical conditions necessitates a more advanced method of drug effect predictions. Recent advancements in geometric learning further popularized drug effect prediction, catalyzing state-of-the-art models to capture these complex biological interactions. For instance, B. Kuenzi et al predicted the drug response of human cancer cells using chemical structure data, genomics data, drug sensitivity data, and protein activity data [8]. However, the choice of 2D structure instead of 3D topology and a lack of comprehensive gene descriptions/networks limit the model's ability to fully capture complex interactions of drugs and intricate biological pathways. In addition, drug response is often heavily influenced by one's genetic makeup. The gene regulatory network poses a great challenge for researchers in this field because of the hidden interactions and a lack of biological understanding of this subject. To address these challenges, this research implements multi-modal data frameworks that are descriptive (geometric/textual information) and relationship-focused.

1.1 Preliminaries: Drug Discovery Process

Drug discovery involves five main phases; 1) the pre-discovery stage where disease mechanisms are explored; 2) the discovery stage where scientists search for appropriate small-molecule therapeutics that interfere with the disease mechanism; 3) the preclinical stage where drug candidates are tested for their efficacy on various in-vitro or in-vivo models; 4) the clinical stage for human testing; 5) the post-market reviewing and approval of this drug [9]. Our ML pipeline aims to optimize the pre-clinical phase by predicting in-vitro drug effects on various cell lines.

Traditional Preclinical Development

Extensive in vitro tests are performed during preclinical development. These experiments test potential drug efficacy before proceeding to in-vivo studies and clinical trials. IC50, AUC, and Z-score are some of the most informative metrics that shed light on different aspects of the drug's behavior. The IC50 value measures the drug's potency. It shows the concentration needed to inhibit a biological or biochemical function by 50% [10]. The AUC is derived from a dose-response curve that represents the effect of various drug concentrations on each cell line. This value summarizes the drug efficacy across a range of concentrations [11]. The Z-score compares an IC50 value with those from other cell lines, showcasing the effectiveness of a drug in comparison to the average response, and highlighting whether the drug is more or less potent in a specific cell line relative to others.

Traditionally, screening these values costs hundreds of millions of dollars and requires years of clinical testing [12]. On average, out of 10,000 molecules screened, only one may eventually lead to the market. Due to the immense efforts of validating one drug molecule, there is limited flexibility in the traditional pipeline. Such inefficiencies pose an economic barrier to drug development for serious diseases like cancer. Consequently, companies place a lower priority on these endeavors and often shift their focus toward more cost-effective avenues, potentially delaying the development of critical treatments for diseases like cancer.

AI-Assisted Preclinical Development

Artificial Intelligence (AI) is revolutionizing the drug discovery process. The power of AI can be hugely manifested in medicine due to its ability to recognize patterns in vast amounts of data with varying modalities, personalize treatment plans, and predict patient outcomes [13]. For instance, supervised learning is heavily used for the prediction of molecular properties, pharmacokinetics,

chemical synthesis, etc [14, 15]. However, the biggest challenge lies in the complexity of data. This arises from the diversity of data types, the complexity of chemical interactions in biological systems, and the difficulty of accurately encoding this information into an ML pipeline.

1.2 Our contributions.

We introduce PanRX, a deep-learning model designed to predict drug effects by integrating multi-modal drug and genetic information through a transductive KG. We leverage the extensive training capabilities of language models and geometric models with the relational structure of KGs to capture complex interactions between drug molecules and cell lines. On the drug side, we incorporate 3D geometric information such as atomic numbers, bonding details, and 3D coordinates of atomic positions as well as SMILES string representation of molecular topology. Also, textual information (including drug summaries, pharmacodynamics, indications, and mechanism of action) and numeric data (including charge, enthalpy, and free energy) further enrich the model. On the genetic side, PanRX integrates gene-related textual data including gene summaries, expression patterns, cellular locations, interactions, and binary mutation status for genes and copy number alterations. This multi-modal fusion of 3D geometry and comprehensive genomic descriptions within a KG framework emphasizes extensive biological relationships, providing a nuanced understanding of drug-cell line interactions. Therefore, PanRX bridges the gap between current models and real-world biological complexities, offering improved predictive accuracy and possibility in drug discovery.

To verify the effectiveness of PanRX, we created PanCan-DrugsGenes, a dataset specifically designed for pancreatic cancer research. PanCan-DrugsGenes is a Lightning Memory-Mapped Database (LMDB) with 204 pancreatic cancer drugs and 142 genes that are highly correlated with the disease. Each drug or gene has multi-modal data that conveys rich information. Each KG is constructed from the outputs of this dataset. For evaluation of PanRX, we utilize IC50, AUC, and Z-Score values to measure the confidence of predicted drug effects.

To the best of our knowledge, PanRX is the first to combine 3D drug chemical structures with genetic information to enhance personalized drug design. In addition, PanRX emphasizes complex interactions within biological networks, labeling each connection to provide detailed information about the nature of these relationships. This structure prepares for future downstream applications such as entity/relationship predictions. Extensive experiments on clinical datasets were conducted. The results show powerful drug effect predictive capabilities for pancreatic cancer.

2 Related Work

Multi-Modal Modeling on Small Molecules. In existing ML for the drug discovery community, there are multiple modalities describing molecules, and they can be roughly divided into two venues: internal chemical structure and external functional description [16]. For internal chemical structures, GraphMVP [17] initiates the molecule pretraining by utilizing the 2D topology and 3D geometry. Follow-up works like MoleculeSDE [18] extend this line by proposing a more advanced geometric pretraining algorithm, and MoleculeJAE [19] utilizes the molecular dynamics for pretrianing. On the other hand, MolT5 [20] and MoleculeSTM [21] are the first two works to align both the internal chemical structures and the external functional description for molecule design and editing.

Multi-Modal Modeling on Genomes. Genomic information is organized in various ways depending on the research objective. FASTA formats are used to represent sequences (DNA, RNA, protein) and are predominantly used for gene function prediction or drug-target interactions [22]. FASTA provides efficient storing of biological sequence strings that allows for straightforward retrieval and analysis. Feature tables gained popularity due to their ability to map genomic regions to biological functions, assisting the understanding of gene regulation and expression under various conditions. Sparse matrices are used to handle high-dimensional gene expression data used during drug response prediction tasks [23]. Sparse matrices are also computationally efficient due to the majority of elements being zero.

This research team utilizes textual descriptions of the genome (genes and CNAs) because text descriptions can encapsulate information that is represented by any other types of genomic data structures. Textual data can effectively represent nuanced information such as genomic interactions, regulatory mechanisms, and expression conditions. By integrating these descriptions that are often lost

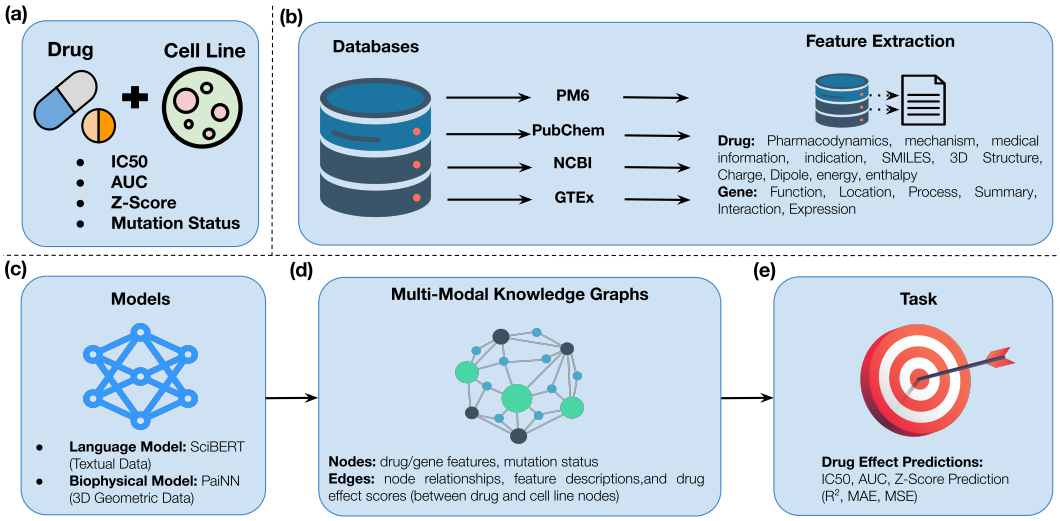

Figure 1: An illustration of the pancreatic cancer drug prediction (PanRX) pipeline. (a) The bioassays (IC50, AUC, Z-Score) of each drug on diseased cell lines is extracted from GDSC. In addition, the mutation status of key genes and CNAs are recorded. (b) Drug and genetic features were extracted from open-access databases like PM6, PubChem, NCBI, and PTEx. (c) Textual and geometric data are encoded by SciBERT and PaiNN respectively. (d) Multi-modal knowledge graphs are constructed for each pair of drug-cell line interactions. (e) The bioassays (IC50, AUC, Z-Score) are predicted.

in purely numeric data modalities, we enhance the application of genomic information in personalized medicine.

Multi-Modal Modeling on Knowledge Graph. In this work, we are merging the multi-modal information of small molecules and genomes using a knowledge graph. Existing knowledge graph papers use either single-modal information on small molecules [24, 25, 26] or single-modal information on genomes [27, 28]. More recent works have started to merge the information of different entities, such as small molecules and proteins [29]. However, as illustrated in recent benchmark works [30], the geometric information has been more informative for molecule representation.

3 Methods

In drug development, performing precise drug effect predictions on specific cancer cell lines is crucial yet challenging. Genetic variations such as gene and copy number alteration (CNA) mutation status pose extreme obstacles to anticipating drug performance in different biological contexts. Traditional approaches rely on time-consuming procedures, significantly delaying the drug delivery pipeline.

Predicting drug effects in pancreatic cancer is a multi-dimensional barrier involving integrating various levels of knowledge such as molecular properties, genetic information, and relevant bioassays. Unique genetic profiles like the variation of copy number alternations (CNA) and mutation statuses remain the root cause of this challenge. Traditional methods rely on repetitive in vitro tests and linear models which lack insights against complex biological interactions that define pancreatic cancer drug responses.

This paper aims to build a predictive pipeline that captures the interaction between drugs and pancreatic cancer cell lines. We focus on genetic variations across cell lines by leveraging cell-line genetic characteristics and chemical properties. Specifically, we utilize numerical data to represent drug characteristics, including charge, dipole moment, energy, and enthalpy, alongside bioassay measurements such as IC50, AUC, and Z-score Additionally, we incorporate binary data to capture the mutation status of genes and copy number alterations (CNA) within each cell line. By constructing a multi-modal knowledge graph with other chemical and genetic features, the problem is now framed as an edge prediction task a multi-modal knowledge graph where the aim is to predict drug-cell line interaction outcomes (IC50, AUC, Z-Score).

Our approach involves the construction of knowledge graphs where nodes represent drugs, genes, and their inherent features to emphasize variability. This task requires multi-modal data including textual descriptions, chemical structures, and numerical data.

3.1 Problem Formulation

Our model can be described by $p(x, y, z \mid KG)$, where x, y, and z represent the IC50, AUC, and Z-Score values, respectively, and KG refers to the multi-modal knowledge graphs we constructed. We use graph neural networks (GNN) to fuse the multi-modal information discussed above. Leveraging graph structure, we encode drug molecules (geometric data) and cell line genetic information (textual data) as nodes and relationships between these entities as edges. This multi-modal approach enables the model to learn from diverse data types of the biological network that are critical to predicting drug responses. In this project, we consider two architects of the GNN to perform such fusion: graph convolutional networks (GCNs) and graph attention networks (GATs).

Geometric Modeling on Conformation. In our study, we employ a geometric model to represent 3D drug conformations to further investigate their interactions with biological targets. The chemical geometry is described using the PaiNN model $f_g = \text{PaiNN}(a, r)$, where a is atom types and r is atom coordinates in the molecule [31]. By leveraging a geometrically pre-trained PaiNN [18], our method assures all drug spatial configurations are precisely represented. This approach facilitates improved accuracy interaction and effect prediction.

Language Model on Functional Description. On the other hand, we leverage the BERT architecture by utilizing SciBERT to effectively capture the existing single-modal data. Specifically, drug textual descriptions include drug summaries, pharmacodynamics, drug mechanisms, and indications. Genetic textual descriptions include Gene functions, locations, processes, summary, interactions, and expression. The model has a representation of $f_t = \text{SciBERT}(w_t)$, where w_t is the textual description for each chemical/gene feature node. These embeddings capture intricate details contained within textual descriptions, improving the overall model with an extra layer of biological context.

3.2 Multi-modal Fusion

Graph convolution network (GCN). Our proposed GCN model consists of six convolutional layers (GCNConv), each followed by an Exponential Linear Unit (ELU) activation to capture non-linear relationships. To prevent overfitting, a dropout rate of 0.3 is applied after each layer. These layers aggregate node features on a global level based on node connectivity, allowing the convolutional layers to propagate node feature embeddings across the graph, with each layer adjusting the embeddings by considering adjacent nodes. A layer normalization is applied after the final convolutional layer to ensure stable learning dynamics. The model is trained using the Adam optimizer with a learning rate of 0.00001 and a weight decay of 0.0001. Performance is evaluated using Mean Squared Error (MSE), Mean Absolute Error (MAE), and the coefficient of determination (R^2).

Graph attention network (GAT). The GAT setup builds upon the GCN architecture by integrating attention mechanisms to dynamically assess the significance of neighboring nodes during the aggregation phase. The model comprises six GAT layers, each employing multi-head attention to explore various facets of a node's vicinity. The initial five layers utilize four attention heads each, while the final layer uses a single head to achieve dimensionality reduction of the embedding. This approach permits selective integration of neighboring information, fostering a context-sensitive aggregation compared to the GCN. The GAT model utilizes the ELU activation function and a dropout rate of 0.3. The optimization is performed using the Adam optimizer with a learning rate of 0.00001 and a weight decay of 0.0001. The evaluation metrics are consistent with those used in the GCN model, including MSE, MAE, and R^2.

4 Experiments

4.1 Data Acquisition and Feature Extraction

Our primary data source was the GDSC dataset [3]. To evaluate the effectiveness of multi-modal knowledge graphs, we compiled a set of 5014 pancreatic cancer drug-cell line pairs from GDSC. Descriptive features of drugs and cell lines are pulled from public databases. For drugs, we extracted

3D geometry, compound charge, dipole moment, energy, and enthalpy from PM6 [32], and pharmacodynamics, drug mechanisms, medical information, and drug indication from PubChem. For each cell line, GDSC provides the mutation status of key genes and copy number alterations (CNA). We extract descriptive information about each gene from NCBI, Biopython, and GTEx such as function, location, process, summary, interactions, and expression level.

GDSC dataset provides comprehensive drug response information on various cell lines. We build a knowledge graph for each drug-cell pair and extract supporting descriptive features of the drug and related genes from public databases like PubChem, PM6, and NCBI. By feeding them into a deep learning pipeline, we assess the accuracy of drug effect predictions and graph relationship predictions. We selected graph attention networks (GAT) and graph convolution networks (GCN) as our deep learning architects because they can process graph-structured data and capture local and global dependencies. Both architects are designed for edge attribute predictions within the graph data, emphasizing the edge attributes that connect nodes 0 and 1 (IC50, AUC, Z-score).

4.2 Data Pre-Processing and Integration

We built an automated pipeline that streamlines the extraction process of relevant drug and gene information systematically. Firstly, this pipeline extracts data from our primary sources and stores them in the appropriate format in a Lightning Memory-Mapped Database (LMDB) to ensure efficient access to large-scale genomic and pharmacological data of various modalities. Then, we implemented a custom heterogeneous data loader in which each output instance contains a specific drug and the complete set of genes required for constructing a corresponding knowledge graph representing the interaction between a drug and a cell line. Data of various modalities are encoded using the models described in section 4.3 before they are assembled into nodes and organized into a knowledge graph structure. Appropriate edges are included to represent relationships that enhance the logical connectivity of each graph. It is critical to note that the connection between node 0 (drug) and node 1 (cell line) is characterized by three edges each representing a different bioassay measurement - IC50, AUC, Z-Score.

4.3 Experimental Setup

The objective of this experiment is to predict edge attributes between the drug node (node 0) and the cell line node (node 1). These edge attributes represent bioassays, namely IC50, AUC, and Z-score values. Our models were trained at 500 epochs with a batch size of 400 to ensure sufficient stability. To evaluate model performances, we split the dataset into training, validation, and testing sets of 80/10/10 and 50/25/25 ratios. Training, validation, and testing results (MSE, MAE, and R^2) are reported. On average, training out models (GAT and GCN) with the current parameters requires approximately 10 hours.

Hardware All experiments were run on the Nvidia 4090 GPU with 24GB of GDDR6X RAM and 16,384 CUDA cores. We utilized the Windows operating system with Python version 3.12.1 for all computations.

4.4 Results

Table 1: Regression Analysis of Predicted Bioassay Values (80/10/10 split)

GAT (500 epochs)			
Test Set	R^2	MSE	MAE
Training	0.3327	0.0482	0.0184
Validation	0.3330	0.0002	0.0090
Testing	0.3331	0.0002	0.0089

GCN (500 epochs)			
Test Set	R^2	MSE	MAE
Training	0.3253	0.0348	0.0152
Validation	0.3321	0.0009	0.0171
Testing	0.3325	0.0009	0.0169

As depicted in Table 1 and Table 2, both the GAT and GCN architectures performed consistently across the 80/10/10 and 50/25/25 splits, with R^2 at approximately 0.333 across the training, validation, and testing sets, highlighting its stable predictive performance. Also, the MSE was notably low across both architects and splits. For instance, in the 80/10/10 split, the GAT model achieved an MSE of 0.0002

Table 2: Regression Analysis of Predicted Bioassay Values (50/25/25 split)

GAT (500 epochs)			
Test Set	R^2	MSE	MAE
Training	0.3332	0.0049	0.0070
Validation	0.3333	0.0000	0.0010
Testing	0.3333	0.0000	0.0009

GCN (500 epochs)			
Test Set	R^2	MSE	MAE
Training	0.3332	0.0052	0.0073
Validation	0.3333	0.0000	0.0028
Testing	0.3333	0.0000	0.0025

during validation and testing, while the GCN exhibited a slightly higher MSE of 0.0009. Additionally, in the 50/25/25 split, the MSE of both the GAT and GCN showed 0.0000 across these phases.

On the other hand, the GAT consistently achieved lower values relative to GCN in terms of MAE. In the 80/10/10 split, the GAT had an MAE of 0.0090 during validation and 0.0089 during testing compared to the GCN's higher values of 0.0171 and 0.0169. Similarly, in the 50/25/25 split, the GAT maintained MAE values of 0.0010 and 0.0009 for validation and testing, while the GCN resulted in slightly higher values of 0.0028 and 0.0025.

We conducted experiments using multiple training batch configurations, and the results demonstrated an exceptional degree of similarity across the different settings. These results demonstrate the outstanding performance of both models, with GAT exhibiting a slightly enhanced performance in error minimization.

5 Conclusion

In this paper, we proposed PanRX, a novel multi-modal deep-learning pipeline for prediction of pancreatic cancer drug effects. Unlike previous work that used 2D molecular representations, we combined 3D molecular topology with textual information to comprehensively capture each pair of drugs, cell line information, and relationships. Specifically, embeddings were created via the PaiNN model for geometric data, the SciBERT model for textual data, and direct encoding for numerical data. We then constructed a multi-modal knowledge graph for 5014 drug-cell line pairs and employed a GAT and GCN model to predict IC50, AUC, and Z-scores. Preliminary experiments show that PanRX achieves high accuracies on both GNN architects, showcasing the importance of heterogeneity of modalities with the addition of 3D molecular conformation.

However, our model requires further consideration before advancing to the clinical level. Firstly, GDSC performed in vitro experiments which drastically differ from in vivo physiological conditions within an organism. In addition, mutation status is not necessarily binary due to the continuous spectrum of functional consequences. Lastly, PanRX is currently only trained and experimented on pancreatic cancer data which cannot be generalized to other diseases. To overcome these challenges, we plan to further our research with the following points: 1) Train and experiment our model with other cancer types found in the GDSC dataset to validate the effectiveness of multi-model knowledge graphs. 2) Transition to a large language model (LLM) due to its improved efficiency for handling large texts, as it can represent the same information conveyed by knowledge graphs in pure text form. 3) Verify the model with clinicians on the patient level.

Acknowledgement

The authors would like to thank Divin Yan from Fudan University and Haorui Li from Huazhong University of Science and Technology (HUST) for their interesting discussions, suggestions, and data extraction help during this research.

References

[1] S. L. Carey-Smith, R. S. Kotecha, L. C. Cheung, and S. Malinge. Insights into the clinical, biological and therapeutic impact of copy number alteration in cancer. *International Journal of Molecular Sciences*, 25(13):6815, 2024. doi: 10.3390/ijms25136815.

[2] P. Unger Avila, T. Padvitski, A. C. Leote, H. Chen, J. Saez-Rodriguez, M. Kann, and A. Beyer. Gene regulatory networks in disease and ageing. *Nature Reviews Nephrology*, 20:616–633, 2024. doi: 10.1038/s41581-024-00849-7.

[3] W. Yang, J. Soares, P. Greninger, E. J. Edelman, H. Lightfoot, S. Forbes, N. Bindal, D. Beare, J. A. Smith, I. R. Thompson, et al. Genomics of drug sensitivity in cancer (gdsc): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Research*, 41(D1):D955–D961, 2013. doi: 10.1093/nar/gks1111.

[4] H. Askr, E. Elgeldawi, H. Aboul Ella, et al. Deep learning in drug discovery: an integrative review and future challenges. *Artificial Intelligence Review*, 56:5975–6037, 2023. doi: 10.1007/s10462-022-10306-1.

[5] M. Sun, S. Zhao, C. Gilvary, O. Elemento, J. Zhou, and F. Wang. Graph convolutional networks for computational drug development and discovery. *Briefings in Bioinformatics*, 21(3):919–935, 2020. doi: 10.1093/bib/bbz042.

[6] B. Tang, Z. Pan, K. Yin, and A. Khateeb. Recent advances of deep learning in bioinformatics and computational biology. *Frontiers in Genetics*, 10, 2019. doi: 10.3389/fgene.2019.00214.

[7] J. Kim, S. Park, D. Min, and W. Kim. Comprehensive survey of recent drug discovery using deep learning. *Int. J. Mol. Sci.*, 22(18):9983, 2021. doi: 10.3390/ijms22189983.

[8] B. M. Kuenzi, J. Park, S. H. Fong, K. S. Sanchez, J. Lee, J. F. Kreisberg, and T. Ideker. Predicting drug response and synergy using a deep learning model of human cancer cells. *Cancer Cell*, 38 (5):672–684.e6, 2020. doi: 10.1016/j.ccell.2020.09.014.

[9] N. Singh, P. Vayer, K. Tsaioun, B. O. Voultireix, S. Tanwar, and J.-L. Poyet. Drug discovery and development: introduction to the general public and patient groups. *Front. Drug Discov.*, 3: Article 101419, 2023. doi: 10.3389/fdds.2023.1021419. URL `https://doi.org/10.3389/fdds.2023.1021419`.

[10] S. Aykul and E. Martinez-Hackert. Determination of half-maximal inhibitory concentration using biosensor-based protein interaction analysis. *Analytical Biochemistry*, 508:97–103, 2016. doi: 10.1016/j.ab.2016.06.025.

[11] J. D. Scheff, R. R. Almon, D. C. DuBois, W. J. Jusko, and I. P. Androulakis. Assessment of pharmacologic area under the curve when baselines are variable. *Pharmaceutical Research*, 28 (5):1081–1089, 2011. doi: 10.1007/s11095-010-0363-8.

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

[13] G. Obaido, I. D. Mienye, I. D. Emmanuel, A. Ogunleye, O. F. Egbelowo, P. Miene, and K. Aruleba. Supervised machine learning in drug discovery and development: Algorithms, applications, challenges, and prospects. *Machine Learning with Applications*, 17:100576, 2024. doi: 10.1016/j.mlwa.2024.100576.

[14] Z. A. Rollins, A. C. Cheng, and E. Metwally. Molprop: Molecular property prediction with multimodal language and graph fusion. *J Cheminform*, 16:56, 2024. doi: 10.1186/s13321-024-00846-9. URL `https://doi.org/10.1186/s13321-024-00846-9`.

[15] M. Galushka, C. Swain, F. Browne, et al. Prediction of chemical compounds properties using a deep learning model. *Neural Comput Appl*, 33:13345–13366, 2021. doi: 10.1007/s00521-021-05961-4. URL `https://doi.org/10.1007/s00521-021-05961-4`.

[16] S. Liu. Ai for molecule discovery with multi-modal knowledge. 2023.

[17] S. Liu, H. Wang, W. Liu, J. Lasenby, H. Guo, and J. Tang. Pre-training molecular graph representation with 3d geometry. 2022. URL `https://openreview.net/forum?id=xQUe1pOKPam`.

[18] S. Liu, W. Du, Z. M. Ma, H. Guo, and J. Tang. A group symmetric stochastic differential equation model for molecule multi-modal pretraining. In *International Conference on Machine Learning*, pages 21497–21526. PMLR, 2023.

[19] W. Du, J. Chen, X. Zhang, Z. Ma, and S. Liu. Molecule joint auto-encoding: trajectory pretraining with 2d and 3d diffusion. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 55077–55096, 2023.

[20] C. Edwards, T. Lai, K. Ros, G. Honke, K. Cho, and H. Ji. Translation between molecules and natural language. *arXiv preprint arXiv:2204.11817*, 2022.

[21] S. Liu, W. Nie, C. Wang, J. Lu, Z. Qiao, L. Liu, J. Tang, C. Xiao, and A. Anandkumar. Multi-modal molecule structure–text model for text-based retrieval and editing. *Nature Machine Intelligence*, 5(12):1447–1457, 2023.

[22] W. R. Pearson. *Using the FASTA Program to Search Protein and DNA Sequence Databases*, volume 24 of *Methods in Molecular Biology*, page 307. Humana Press, Totowa, NJ, 1994. ISBN 978-0-89603-246-0. doi: 10.1385/0-89603-246-9:307.

[23] A. Serra, P. Coretto, M. Frattello, and R. Tagliaferri. Robust and sparse correlation matrix estimation for the analysis of high-dimensional genomics data. *Bioinformatics*, 34(4):625–634, 2018. doi: 10.1093/bioinformatics/btx642.

[24] A. Kroll, S. Ranjan, M. K. M. Engqvist, et al. A general model to predict small molecule substrates of enzymes based on machine and deep learning. *Nature Communications*, 14:2787, 2023. doi: 10.1038/s41467-023-38347-2. URL `https://doi.org/10.1038/s41467-023-38347-2`.

[25] A. Rossi, D. Barbosa, D. Firmani, A. Matinata, and P. Merialdo. Knowledge graph embedding for link prediction: a comparative analysis. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 15(2):14, 2020. doi: 10.1145/3424672. URL `https://doi.org/10.1145/3424672`.

[26] M. Wang, L. Qiu, and X. Wang. A survey on knowledge graph embeddings for link prediction. *Symmetry*, 13(3):485, 2021. doi: 10.3390/sym13030485. URL `https://doi.org/10.3390/sym13030485`.

[27] J. Vilela, M. Asif, A. R. Marques, J. X. Santos, C. Rasga, A. Vicente, and H. Martiniano. Biomedical knowledge graph embeddings for personalized medicine: Predicting disease-gene associations. *Expert Systems*, 2022. doi: 10.1111/exsy.13181.

[28] Z. Gao, P. Ding, and R. Xu. Kg-predict: A knowledge graph computational framework for drug repurposing. *Journal of Biomedical Informatics*, 132:104133, 2022. doi: 10.1016/j.jbi.2022.104133.

[29] S. Liu, M. Qu, Z. Zhang, H. Cai, and J. Tang. Structured multi-task learning for molecular property prediction. In *International conference on artificial intelligence and statistics*, pages 8906–8920. PMLR, 2022.

[30] S. Liu, Y. Li, Z. Li, Z. Zheng, C. Duan, Z. M. Ma, O. Yaghi, A. Anandkumar, C. Borgs, J. Chayes, et al. Symmetry-informed geometric representation for molecules, proteins, and crystalline materials. *Advances in neural information processing systems*, 36, 2024.

[31] K. Schütt, O. Unke, and M. Gastegger. Equivariant message passing for the prediction of tensorial properties and molecular spectra. In *International Conference on Machine Learning*, pages 9377–9388. PMLR, 2021.

[32] M. Nakata, T. Shimazaki, M. Hashimoto, and T. Maeda. Pubchemqc pm6: Data sets of 221 million molecules with optimized molecular geometries and electronic properties. *J. Chem. Inf. Model.*, 60:5891–5899, 2020. doi: 10.1021/acs.jcim.0c00740. URL `https://doi.org/10.1021/acs.jcim.0c00740`.