Contrastive Learning Drug Response Models from Natural Language Supervision

Kun Li[1] , **Xiuwen Gong**[2] , **Jia Wu**[3] and **Wenbin Hu**[1*]
[1]School of Computer Science, Wuhan University, Wuhan, China
[2]UTS Faculty of Engineering and IT, University of Technology Sydney, Sydney, Australia
[3]Department of Computing, Macquarie University, Sydney, Australia
{li__kun, hwb}@whu.edu.com, gongxiuwen@gmail.com, jia.wu@mq.edu.au

Abstract

Deep learning-based drug response prediction (DRP) methods can accelerate the drug discovery process and reduce research and development costs. Despite their high accuracy, generating regression-aware representations remains challenging for mainstream approaches. For instance, the representations are often disordered, aggregated, and overlapping, and they fail to characterize distinct samples effectively. This results in poor representation during the DRP task, diminishing generalizability and potentially leading to substantial costs during the drug discovery. In this paper, we propose CLDR, a contrastive learning framework with natural language supervision for the DRP. The CLDR converts regression labels into text, which is merged with the drug response caption as a second sample modality instead of the traditional modes, i.e., graphs and sequences. Simultaneously, a common-sense numerical knowledge graph is introduced to improve the continuous text representation. Our framework is validated using the genomics of drug sensitivity in cancer dataset with average performance increases ranging from 7.8% to 31.4%. Furthermore, experiments demonstrate that the proposed CLDR effectively maps samples with distinct label values into a high-dimensional space. In this space, the sample representations are scattered, significantly alleviating feature overlap. The code is available at: https://github.com/DrugD/CLDR.

1 Introduction

Phenotypic drug discovery (PDD) [Chen *et al.*, 2023; Vincent *et al.*, 2022a] demonstrates its superiority over target-based approaches [Lu *et al.*, 2022; Li *et al.*, 2021] by identifying drugs, targets, and mechanisms of action (MoA) [Maillard and Pascoe, 2023]. When molecular insights into a disease are limited [Vincent *et al.*, 2022b; Eder *et al.*, 2014], PDD research provides a framework for exploring the uncharted "dark biological matter" territory, which includes biological molecules and cellular processes linked to diseases through the utilization of proteomics and other molecular methods. This contributes to a more efficient drug discovery process [Moffat *et al.*, 2017]. Moreover, the drug response prediction (DRP) task is an essential PDD step, which is gradually developing as a field and becoming a recognized discovery paradigm in academia and the pharmaceutical industry [Liu *et al.*, 2023]. This sustained interest stems from notable successes over the past decade, such as treating schizophrenia with SEP-363856 [Begni *et al.*, 2021], malaria with KAF156 [Ogutu *et al.*, 2023], and atopic dermatitis with Crisaborole [Kim *et al.*, 2023].

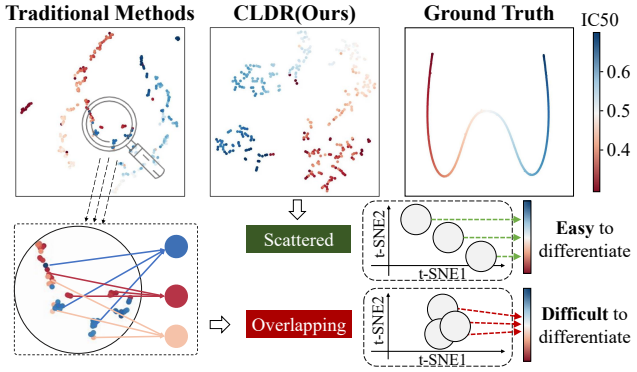

Figure 1: Visual comparison of various methods' learned representations on the GDSC2 dataset. Traditional DRP task regression methods poorly map samples with different regression labels, resulting in aggregated and overlapping features after visualization with the t-SNE method. In contrast, our method can efficiently represent different samples and provide scattered high-dimensional features for the regression task.

DRP primarily focuses on regression tasks. Historically, representation learning has received less attention for regression than classification tasks [Zha *et al.*, 2023]. This is because the DRP task inputs are one cell and drug type, and the output is the half maximal inhibitory concentration (IC_{50}) [Bébéar and Robertson, 1996] of this process. Notably, deep learning-based DRP methods (e.g., [Chu *et al.*, 2023; Liu *et al.*, 2020; Liu *et al.*, 2019b]) show excellent performance when both drug and cell line species are present in the training and test datasets (i.e., many/few-shot [Lee *et al.*, 2022]). However, in practical applications, the IC_{50} of

*Corresponding author

the tested compounds are unlabeled and not observed in the training datasets (i.e., zero-shot [Wang *et al.*, 2023]). Different drugs exhibit diverse structures and properties within the same cell line, resulting in a significant performance decline under the zero-shot learning condition. This factor presents a serious challenge to the development of PDD. For instance, in the work of [Nguyen *et al.*, 2022], the prediction performance of their models based on graph neural networks (GNNs) exceeded 90% in the Pearson correlation coefficient (PCC) under the many-shot condition. However, the PCC decreased to about 4–32% under the zero-shot condition.

Under the zero-shot condition, traditional methods represent sample features inadequately. This is mainly due to overlapping and disordered sample features. These challenges reduce the effectiveness of the DRP model. As illustrated in Figure 1, we compare our method's representations with that of traditional approaches using the genomics of drug sensitivity in cancer database (GDSC2) [Yang *et al.*, 2013], visualized with the t-distributed stochastic neighbor embedding (t-SNE) method. Applying the mean squared error (MSE) loss function as an example, traditional methods yield disordered, fragmented, aggregated, and overlapping mappings, leading to unsatisfactory results during various regression tasks. It is worth noting that when the model fails to capture the mapping pattern of numerical labels, it can only map unseen samples into the known space in a disordered manner, resulting in samples with different label values that cannot be distinguished. This has a detrimental impact on the DRP model's performance during practical applications. Therefore, there is an urgent need to enhance a model's ability to represent continuous values in an ordered and scattered manner to improve generalization performance under the zero-shot condition.

In recent years, contrastive language image pre-training (CLIP) [Zhang *et al.*, 2022] has provided a powerful multimodal representation learning framework, enabling computers to better understand and process the semantic relationships between images and text, like GLIDE [Nichol *et al.*, 2022]. Moreover, CLIP research indicates that state-of-the-art (SOTA) image representations can be achieved with a simple pre-training task using a dataset of 400 million image-text pairs. Furthermore, contrastive learning with labels has been theoretically proven to enhance the performance of learned representations [Ji *et al.*, 2023] in downstream tasks [Wang *et al.*, 2022]. Consequently, in the realm of drug discovery, it becomes possible to establish a connection between drug response data and annotated text, learning representations from the text [Fang *et al.*, 2022]. Subsequently, these representations may be used to enhance zero-shot learning performance during natural language supervision contrastive learning [Khosla *et al.*, 2020; Gunel *et al.*, 2020].

To address this challenge, we present the **CLDR**, a **c**ontrastive **l**earning framework with natural language supervision for **d**rug **r**esponse prediction. The CLDR framework transforms numerical labels used to represent drug response into text using customized prompts. First, drugs and cell lines from the same samples are encoded using natural language and traditional fusion methods, respectively. Then, a contrastive learning strategy is employed to map the drug and cell line feature space and the text with labels using natural language into the shared representation space. This strategy maximizes the similarity between related samples while minimizing those between unrelated ones. In this study, we construct a common-sense numerical knowledge graph (CN-KG), drawing inspiration from the ordinal number definition [Agustito *et al.*, 2023]. The CN-KG constrains the text representation order, improving the fusion encoder representation for drugs and cell lines. Furthermore, through contrastive learning pre-training, ordered and scattered natural language representations are aligned and mapped to a unified high-dimensional space with those of the fusion encoder for drugs and cells.

For the practical evaluation, we validate the method using a dataset comprising over 150,000 samples from the GDSC2 dataset. Consequently, all the methods display increases of 7.88%, 19.49%, 17.83%, 14.29%, 13.04%, and 31.46% after adopting our framework. The proposed CLDR method effectively establishes connections between drug response data and the corresponding labeled text, enhancing generalizability and improving the PDD success rate. Section 3.4 provides detailed theoretical proof of the CLDR method's validity. Our contributions are as follows:

- We propose the CLDR method, a novel contrastive learning framework with natural language supervision for the DRP task. The CLDR method effectively constructs links between drug response data and labeled text.
- We construct a CN-KG to capture the continuous nature of sample order to improve the fusion encoder representation for drugs and cell lines.
- The extensive experiments using the GDSC2 dataset demonstrate that employing the CLDR method leads to a notable improvement in the DRP methods, enhancing results by at least 7.8% and up to 31.4%.

2 Related Work

DRP exploration has become feasible due to drug response studies on a large number of cell lines, exemplified by GDSC2 [Yang *et al.*, 2013] and cancer cell line encyclopedia (CCLE) [Barretina *et al.*, 2012]. Various DRP techniques have been introduced, and categorized into 1DCNN, graph, and transformer methods. Generally, non-graph-based methods employ convolutional neural networks (CNNs) and multilayer perceptrons (MLPs) for information retrieval. These methods encode drug molecules in the simplified molecular input line entry specification (SMILES) string format and directly extract gene sequence features using 1DCNN, overlooking the pharmacological and structural attributes of the molecule. Conversely, graph-based methods [**?**; Liu *et al.*, 2020] concentrate on converting drug molecules into graph structures, utilizing GNNs for representation instead of directly extracting features from the SMILES strings. Additionally, in contrast to GNNs, the transformer-based methods [Chu *et al.*, 2023; Jiang *et al.*, 2022] avoid introducing any structural inductive bias at intermediate layers [Chu *et al.*, 2023], thereby mitigating the GNNs' expressivity limitations [Park *et al.*, 2020].

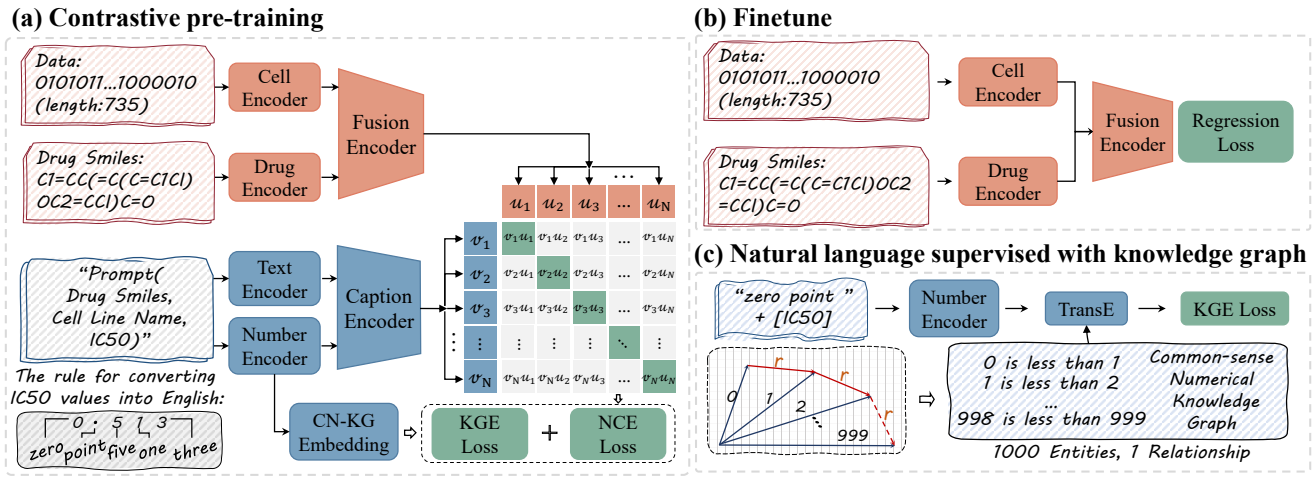

Figure 2: Summary of the CLDR framework. (a)During the pre-training stage, the CLDR jointly trains the cell and drug fusion and the caption encoders to predict the correct sample pairings. (b)When fine-tuning, the standard DRP model trains the drug-cell feature fusion encoder and a regressor to predict IC_{50}. (c)When the caption encoder describes the response process, the number encoder is restricted to learning the continuous nature of sample orders with the CN-KG.

In summary, the conventional DRP method employs a unimodal end-to-end approach for IC_{50} training and prediction. However, its effective applicability is limited under the zero-shot condition, affecting the PDD's efficiency.

3 Methods

During the pre-training stage, the CLDR method principally comprised the fusion encoder for the drug and cell line features extraction and fusion, the caption encoder for the drug response process text, and the natural language supervision module incorporating the CN-KG.

Subsequently, during the fine-tuning phase, standard regression loss functions are utilized. The primary goal of the pre-training phase is to map the feature spaces of the drug, cell line, and the labeled text, along with the labeled text, into a shared representation space.

3.1 The Fusion Encoder

The DRP task's inputs were the drugs and cell lines, and the specific method for the fusion encoder is not discussed in detail in this section.

Moreover, the input types for the drug extractor were the SMILES sequence or molecular graph. We uniformly describe the feature representations process for the drug (denotes as d_i) and the cell line (denotes as c_i) as follows:

$$\mathbb{P}_i = \Phi_{\text{drug}}(d_i), \mathbb{Q}_i = \Phi_{\text{cell}}(c_i), \tag{1}$$

where, N represents the total number of samples $i \in [0, N]$, Φ_{drug} and Φ_{cell} denotes the drug and cell line encoders of the initial DRP methods, while $\mathbb{P}_i$ and $\mathbb{Q}_i$ are the representations of the drug d_i and the cell line c_i. The fusion encoder can be expressed as follows:

$$F = \Phi_{\text{f}}([\mathbb{P}_i, \mathbb{Q}_i]), \tag{2}$$

where $F \in \mathbb{R}^{N \times m}$ is the feature after encoding using the fusion encoder, Φ_{f} denotes the fusion encoder, which can be replaced in different methods.

3.2 The Caption Encoder

In the caption encoder, the text-based description was the second coding modality for the drug response process. The inputs $\mathbf{a}$, $\mathbf{b}$ and the output $\mathbf{c}$ in the DRP task are described uniformly with the prompt $\mathcal{F}_{\text{prompt}}(\mathbf{a}, \mathbf{b})$ as follows:

The drug response value between [a] and [b] is [c]

where $\mathbf{a}$ denotes the drug's SMILES string representation, $\mathbf{b}$ represents the name of the cell line, and $\mathbf{c}$ is the quantitative output value of the reaction level between the drug and the cell lines (i.e., the IC_{50}). To represent IC_{50} efficiently, we defined a transformation method $\mathcal{F}_{\text{num2str}}(\mathbf{c})$ that converts it into English words in character order, where the decimal point is encoded as "point". For example, as shown in Figure 2(a), "0.513" can be converted to "zero point five one three."

To facilitate continuity constraints on the values, the caption encoder was designed to consist of two encoders, including Φ_{text}, Φ_{number}, which encodes text $T_{\text{text}} = \mathcal{F}_{\text{prompt}}(\mathbf{a}, \mathbf{b})$ and value-based descriptions $T_{\text{number}} = \mathcal{F}_{\text{num2str}}(\mathbf{c})$, respectively.

Then, the two feature vectors were merged, fusing the descriptive information of the drug reaction process using the multi-layer transformer Φ_{cap} as follows:

$$T = \Phi_{\text{cap}}\left([\Phi_{\text{text}}\left(T_{\text{text}}\right), \Phi_{\text{number}}\left(T_{\text{number}}\right)]\right), \tag{3}$$

where $T \in \mathbb{R}^{N \times m}$ is the description text feature with the caption encoder encoding.

3.3 Supervising Natural Language Using a Knowledge Graph

To guarantee perceiving the number continuity when encoding numerical values in natural language, we constructed a

CN-KG according to [Liu *et al.*, 2017; Duan *et al.*, 2021], which uses linear structures to construct graphs that accurately and intuitively represent numbers and their relationships.

As shown in Figure 2(c), the CN-KG's entity sets are sequences of ordered numbers, denoted as E. These CN-KG entities are integers, the significant number of **c** multiplied by the specified precision, where the minimum and maximum numbers are customized for the specific task. The E are linked by a single relationship type called "is less than," denoted as L, which ensures the numbers' transfer properties are captured.

When dealing with numerical information, we must consider how to incorporate numerical features into the framework to represent their relationships more accurately. To enhance the number encoder's efficiency in the numerical relationships using the CN-KG, we proposed a margin-based loss function $\mathcal{L}_{\mathrm{KGE}}$ for the CN-KG embedding. We aim to minimize the differences in embedding vectors between the entities' set E and the single relationship L (i.e. is less than).

$$\mathcal{L}_{\mathrm{KGE}} = \sum_{(h,l,t)\in S} \left[\gamma + d(\mathbf{h}+\mathbf{l},\mathbf{t}) - d(\mathbf{t}+\mathbf{l},\mathbf{h})\right]_+, \quad (4)$$

where, $[x]_+$ denotes the positive part of x, $\gamma > 0$, and is a margin hyperparameter. The set S is composed of the triplets (h, l, t), with $h, t \in E$, $l \in L$. The embeddings $\mathbf{h}, \mathbf{l}, \mathbf{t}$ obtain values in $\mathbb{R}^k$ (k is a hyperparameter) and are denoted with the same letters in boldface characters. In addition, the L_1 or L_2-norm can be used for the similarity measure d.

Considering that the CN-KG can restrict number relationships, the set of entities E are encoded by the number encoder $\Phi_{\mathrm{number}}\left(T_{\mathrm{number}}(E)\right)$, and the relationship l is represented by a learnable embedding.

Pre-training For the i-th representations (d_i, c_i) generated by the fusion encoder and the j-th captions (d_j, c_j, y_j) produced by the caption encoder in a batch $\mathcal{B}$, we normalized the feature vectors in a hyper-sphere using $u_i := \frac{\Phi_{\mathrm{f}}(d_i,c_i)}{\|\Phi_{\mathrm{f}}(d_i,c_i)\|}$ and $v_j := \frac{\Phi_{\mathrm{cap}}(d_j,c_j,y_j)}{\|\Phi_{\mathrm{cap}}(d_j,c_j,y_j)\|}$. The similarity between u_i and v_j was calculated as $u_i^{\mathrm{T}} v_j$. Finally, a supervised contrastive loss function was used to train the model:

$$\mathcal{L}_{\mathrm{NCE}} = -\frac{1}{N}\left(\sum_i^N \log\frac{\exp(u_i^{\mathrm{T}} v_i/\sigma)}{\sum_{j=1}^N \exp(u_i^{\mathrm{T}} v_j/\sigma)} + \sum_i^N \log\frac{\exp(v_i^{\mathrm{T}} u_i/\sigma)}{\sum_{j=1}^N \exp(v_i^{\mathrm{T}} u_j/\sigma)}\right), \quad (5)$$

where, N is the size of the batch $\mathcal{B}$, and σ is the temperature for scaling the logits.

CLDR's goal during the pre-training phase is to jointly optimize the following contrast and CN-KG embedding loss functions:

$$\mathcal{L}_{\mathrm{All}} = \alpha\mathcal{L}_{\mathrm{NCE}} + (1-\alpha)\mathcal{L}_{\mathrm{KGE}}, \quad (6)$$

where α represents the joint optimization weight adjustment factor for the two loss functions.

Fine-tuning During the fine-tuning stage, we employed the MSE loss function for supervised regression on the fusion encoder. A regression output layer Φ_{mlp} based on the MLP was designed after the fusion encoder as follows:

$$\mathcal{L}_{\mathrm{REG}} = \frac{1}{|N|}\sum_{i=0}^{|N|}(\Phi_{\mathrm{mlp}}(\Phi_{\mathrm{fusion}}(d_i, c_i)) - y_i)^2. \quad (7)$$

where y_i is the normalized value of IC_{50} corresponding to $\{d_i, c_i\}$.

3.4 Theoretical Analysis

In this section, we theoretically prove that jointly optimizing $\mathcal{L}_{\mathrm{NCE}}$ and $\mathcal{L}_{\mathrm{KGE}}$ enables the fusion encoder Φ_{f} to obtain continuous regression-aware representation.

Notations Let $\{x_i, y_i\}$ be the inputs and outputs respectively, where y_i is a sorted label with the ordering $y_1 \le y_2 \le \cdots \le y_n$. The $\delta \in (0, 1)$ denotes the minimum interval of normalized labels $y_i \in [0, 1]$.

First, based on the loss function $\mathcal{L}_{\mathrm{KGE}}$ and its expectation that **t** should be a nearest neighbor of $\mathbf{h} + \mathbf{l}$ [Bordes *et al.*, 2013] (see Section 3.3), we formulated:

$$d(\Phi_{\mathrm{number}}(y_i), \Phi_{\mathrm{number}}(y_{i+1})) := \mathbf{l} + \epsilon, \quad (8)$$

where $\mathbf{l}$ is the only relationship embedding type in the CN-KG and $\epsilon \in \mathbb{R}$ is a model perturbation. If we denote Φ_{number} as $\mathcal{N}$, then the following lemma can be inferred:

Lemma 1 (Equal interval representation of $\mathcal{N}$)**.** *For any $0 < \delta < 1$, two perturbations ϵ_0, ϵ_1 exist to make:*

$$\mathcal{N}(y_i) - \mathcal{N}(y_{i+1}) = \epsilon_0 \cdot \mathbf{l} + \epsilon_1.$$

Lemma 1 implies that Φ_{number} can learn continuous representations that capture the intrinsic sample order of the regression target. Next, based on the $\mathcal{L}_{\mathrm{NCE}}$ constraints on positive x_i, y_i and negative samples x_i^-, y_i^-, we expect the following:

$$\begin{aligned} d(\mathcal{C}(x_i, y_i), \mathcal{F}(x_i)) &\ll d(\mathcal{C}(x_i^-, y_i^-), \mathcal{F}(x_i)) \\ d(\mathcal{C}(x_i, y_i), \mathcal{F}(x_i)) &\ll d(\mathcal{C}(x_i, y_i), \mathcal{F}(x_i^-)), \end{aligned} \quad (9)$$

where $\mathcal{C}$ and $\mathcal{F}$ denote Φ_{cap} and Φ_{f}, respectively. Also, $\mathcal{M} := \mathcal{C}(x_i, y_i) - \mathcal{F}(x_i)$ denotes the degree of alignment between two modalities. The subsequent lemma can be derived from Equation (9):

Lemma 2 (Upper bound of $\mathcal{L}_{\mathrm{NCE}}$)**.** *For any $i \in j$, the formula $\epsilon_2 > 0$ such that $\|\mathcal{M}\| \le \epsilon_2$ exists.*

where ϵ_2 is a value related to ϵ_0 and ϵ_1. Lemma 2 states that the upper bound on $\mathcal{L}_{\mathrm{NCE}}$ is equal to the upper bound on the distance between the distributions of features $\mathcal{C}$ and $\mathcal{F}$ in a uniform representation space.

Thus, under the $\mathcal{L}_{\mathrm{NCE}}$ constraint, the property of the equal interval representation of $\mathcal{N}$ will conditionally transfer to $\mathcal{F}$:

Theorem 1 (Main theorem)**.** *For any $i \in j$, the formula $\theta \in (0, 1)$ such that if δ is close to 0, then $\|\mathcal{F}(x_i) - \mathcal{F}(x_{i+1})\| \le 2\epsilon_2$ exists.*

Methods		Drug					Total				
		RMSE ↓	MSE ↓	PCC ↑	SPC ↑	Rank ↓	RMSE ↓	MSE ↓	PCC ↑	SPC ↑	Rank ↓
tCNNs [Liu *et al.*, 2019b]	Original	0.0548	0.0036	0.4710	0.4682	0.0257	0.0596	0.0036	0.5342	0.4632	0.0267
	+CLDR	**0.0539**	**0.0034**	**0.5272**	**0.5321**	**0.0218**	**0.0580**	**0.0034**	**0.5451**	**0.4894**	**0.0227**
	(Improv.)	1.64%	5.56%	11.93%	13.65%	15.18%	2.71%	5.34%	2.05%	5.64%	15.13%
DeepTTC [Jiang *et al.*, 2022]	Original	0.0620	0.0057	0.4405	0.4409	0.0291	0.0729	0.0054	0.3231	0.1986	0.0302
	+CLDR	**0.0590**	**0.0044**	**0.4778**	**0.4797**	**0.0200**	**0.0659**	**0.0043**	**0.3873**	**0.2756**	**0.0208**
	(Improv.)	4.84%	22.81%	8.47%	8.80%	31.27%	9.55%	19.16%	19.89%	38.75%	31.32%
DeepCDR [Liu *et al.*, 2020]	Original	0.0598	0.0050	0.4599	0.4546	0.0330	0.0726	0.0053	0.3419	0.2282	0.0342
	+CLDR	**0.0560**	**0.0040**	**0.5360**	**0.5402**	**0.0264**	**0.0640**	**0.0041**	**0.4262**	**0.2696**	**0.0275**
	(Improv.)	6.35%	20.00%	16.55%	18.83%	20.00%	11.82%	22.25%	24.68%	18.13%	19.66%
GraphDRP [Nguyen *et al.*, 2022]	Original	0.0637	0.0056	0.4402	0.4447	0.0337	0.0729	0.0053	0.3096	0.2037	0.0312
	+CLDR	**0.0565**	**0.0047**	**0.5285**	**0.5348**	**0.0287**	**0.0699**	**0.0049**	**0.3410**	**0.2712**	**0.0297**
	(Improv.)	11.38%	16.23%	20.05%	20.25%	14.91%	4.09%	7.77%	10.14%	33.14%	4.91%
GratransDRP [Chu *et al.*, 2023]	Original	0.0593	0.0047	0.4738	0.4770	0.0283	0.0666	0.0044	0.4080	0.3255	0.0292
	+CLDR	**0.0523**	**0.0039**	**0.5288**	**0.5333**	**0.0264**	**0.0633**	**0.0040**	**0.4665**	**0.4424**	**0.0274**
	(Improv.)	11.80%	17.02%	11.61%	11.80%	6.71%	5.01%	9.77%	14.33%	35.92%	6.39%
TransEDRP [Li and Hu, 2022]	Original	0.0624	0.0052	0.5060	0.5040	0.0295	0.0694	0.0048	0.3040	0.1635	0.0331
	+CLDR	**0.0547**	**0.0038**	**0.5149**	**0.5294**	**0.0229**	**0.0612**	**0.0037**	**0.4768**	**0.3721**	**0.0239**
	(Improv.)	12.38%	26.23%	1.77%	5.05%	22.42%	11.93%	22.43%	56.86%	127.62%	27.92%

Table 1: Overall method experiment summary. Each method shows the results of the original and those based on our framework (+CLDR), where Improv. denotes the enhancement percentage. Positive improvements are highlighted in red.

More generally, if $\mathcal{Q} := \mathcal{F}(x_i) - \mathcal{F}(x_{i+1})$, then we can derive the following expression:

$$\|\mathcal{Q}\| \leq \left\| \frac{\partial \mathcal{C}}{\partial y} \left(x_i, \theta y_i + (1-\theta)\, y_{i+1}\right) \right\| (y_{i+1} - y_i) + 2\epsilon_2. \tag{10}$$

Theorem 1 suggests an upper bound on the learned continuous representation of $\mathcal{F}$. The upper bound is jointly determined by that of $\mathcal{M}$ and δ, corresponding to $\mathcal{L}_{\mathrm{NCE}}$ and $\mathcal{L}_{\mathrm{KGE}}$, respectively. In addition, we observed that reducing δ and $\mathcal{L}_{\mathrm{NCE}}$ enhances the representation of Φ_{f}.

3.5 Algorithm Complexity

When a model encounters many unlabeled compounds, prediction time and accuracy become the most critical factors, while the training time and calculated cost may be deemed negligible. Our method exhibits a time complexity of $\mathcal{O}(2n^2 + 3n)$ during pre-training. This complexity shows moderate variations when integrated with specific DRP models. During the fine-tuning and inference phases, the algorithm's complexity is solely determined by the specific model, and our method does not play a role in these processes.

4 Experiments

4.1 Experiment Settings

Validation strategy In the zero-shot learning context, simulating the PDD practical application scenario required clustering the response data by drug type. Then, the data was randomly divided into training, validation, and testing sets by a ratio of 8:1:1 and the drug type as the division standard. As a result, the DRP model encounters unknown compounds during the testing phase. This zero-shot learning condition presents a greater level of complexity compared to the random splitting of the entire dataset commonly employed in supervised learning [Bai *et al.*, 2023; Wang *et al.*, 2023]. During the ablation study, all the fusion encoder experiment branches adopted the TransEDRP structure [Li and Hu, 2022].

Metrics To comprehensively evaluate the CLDR's impact, we employed several evaluation metrics: root mean square error (RMSE) to gauge deviation, PCC [Cohen *et al.*, 2009] to assess linear correlation, Spearman's rank correlation coefficient (SRC) [Sedgwick, 2014] to measure monotonicity, and margin ranking loss (Rank) [Liu *et al.*, 2019a] to evaluate ranking performance.

4.2 Overall Experiment

This paper proposes the CLDR framework which leverages the caption encoder and CN-KG's powerful representation capabilities to align the fusion encoder for drug response features, improving the DRP model's generalizability for zero-shot learning.

Therefore, we broadly selected representative DRP methods based on deep learning models such as CNNs, GNNs, and transformers, to verify the CLDR's effectiveness and generalizability. In the overall experiment, we selected and tested six methods on the most widely employed GDSC2 dataset.

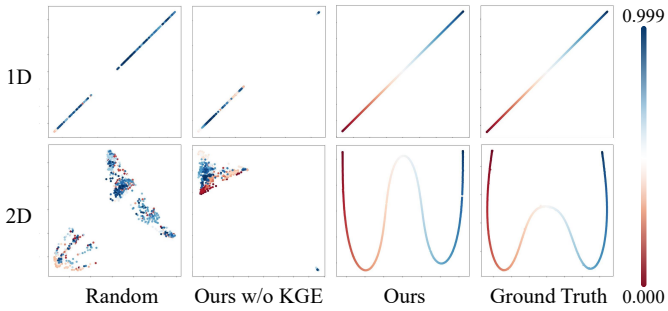

Figure 3: Illustrations of learned representations of 0.000 to 0.999 with different strategies, where the dimensions of the text representations are reduced in one- and two-dimensions by t-SNE to compare the effect of the CN-KG.

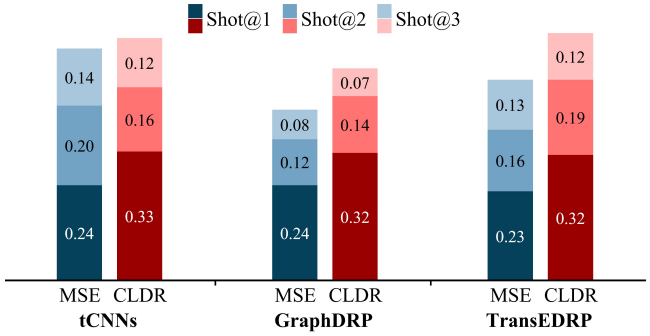

Figure 4: DRP model's drug screening capabilities.

As shown in Table 1, the original models' prediction results are denoted by **Original**, while the combinedCLDR framework's outcomes are represented by **+CLDR** .

All the methods demonstrated increases of 7.88%, 19.49%, 17.83%, 14.29%, 13.04%, and 31.46% after our framework was adopted. The results show that the CLDR framework can be generally applied to various DRP methods to enhance performance under the zero-shot condition.

4.3 Ablation Study

CN-KG To verify that the CN-KG enhances the continuous numerical text representation, we designed the experiment without using the knowledge graph embedding. Specifically, we did not constrain the number encoder with $\mathcal{L}_{\text{KGE}}$. Also, we maintained pre-training and the fine-tuning approach remained unchanged. Based on the ablation experiments in Table 2, the model's performance increased by 9.6% with $\mathcal{L}_{\text{KGE}}$.

Furthermore, as shown in Figure 3, the numerical text representation vectors employing the CN-KG are compared. This is because contrastive learning enables the fusion encoder to learn the number encoder's continuous numerical representation capability. Although the fusion encoder without $\mathcal{L}_{\text{KGE}}$ can still represent the continuous sample information in a regular manner (as shown in Figure 1), it fails to capture the continuous nature of sample orders, resulting in feature overlap and making distinguishing between different samples difficult.

Precision During pre-training, the constraint of continuous values with different precision has various effects on the performance of models based on Equation (10). Thus, we designed ablation experiments to investigate specific effects. With all other conditions consistent, the precision grids 0.1, 0.01, and 0.001 were created to pre-train the model. The results are shown in Table 2, where the performance of the model at 0.01 significantly exceeds that at 0.1 and 0.001. This is mainly because the model needs to balance the precision and perturbation. As shown in Table 1, the MSE loss of various DRP methods is around 0.005.

As a result, the model cannot accurately predict the thousandth part. Additionally, there is a possibility that the value of the thousandth part will be a perturbation factor during contrastive text learning.

Caption During the pre-training phase, the caption encoder's prompts contain the drug response inputs $\mathbf{a}$, $\mathbf{b}$ and the IC_{50} results $\mathbf{c}$. Since the inputs contain drugs and cell lines, to verify whether the inputs containing the drug reaction captions improve the fusion encoder's ability to discretely map various drug reaction pairs, we removed the input texts to compare the results. In the caption encoder, we used $\mathbf{a}$ and $\mathbf{b}$ as **Text**, and $\mathbf{c}$ as **Number**. As shown in Table 2, the model improves on all metrics with an average increase of 9.1% after adding the inputs to the text description. This is because the description of the drug response, a real biological phenomenon, requires at least three kinds of information, drug, cell type, and reaction result. Otherwise, the drug response process will collapse into a one-dimensional space, affecting the transformation of the feature space from the caption to the fusion encoder.

4.4 Computational Analysis

We performed a computational analysis to compare the time cost and parameter increases to the profits made during the pre-training phase.

The experiment utilized an Intel Xeon E5-2690 v3 processor with 12 cores (i.e., 24 threads) and a clock frequency of 2.60GHz. Additionally, the RTX 4090 GPU was utilized. During the training time experiments, each model underwent testing 10 times in the GPU environment.

As shown in Table 3, **AVG w. CLDR** represents an average metric enhancement facilitated by the CLDR method. Incorporating the CLDR framework significantly extends the training time and parameter count by an additional 4.21 seconds and 5,811M, respectively. Notably, the CLDR method demands fewer training epochs in the pre-training and fine-tuning phases, resulting in a shorter training duration than the original models.

In summary, integrating the CLDR framework while increasing the parameter count substantially enhances generalization performance. Our work represents a breakthrough when compared to traditional DRP methods.

4.5 Case Study Analysis

During the overall experiments, the CLDR method demonstrated high accuracy and generalizability, presenting an ad-

Finetune	Caption	Precision	Loss			Total					Drug				
			MSE	CNE	KGE	RMSE ↓	MSE ↓	PCC ↑	SPC ↑	Rank ↓	RMSE ↓	MSE ↓	PCC ↑	SPC ↑	Rank ↓
×	-	-	√			0.069	0.005	0.304	0.163	0.033	0.062	0.005	0.506	0.504	0.030
×	Number	0.001		√		0.089	0.008	0.299	0.248	0.038	0.077	0.008	0.434	0.466	0.037
√	Number	0.001	√			0.063	0.004	0.414	0.361	0.025	0.056	0.004	0.516	0.525	0.024
×	Number	0.001		√	√	0.084	0.007	0.262	0.150	0.041	0.073	0.007	0.446	0.467	0.040
√	Number	0.001	√			0.065	0.004	0.403	0.301	0.027	0.055	0.004	0.514	0.520	0.026
×	Text+Number	0.1		√	√	0.451	0.203	0.006	0.051	0.156	0.132	0.054	0.008	0.006	0.048
×	Text+Number	0.01		√	√	0.428	0.183	-0.085	-0.175	0.143	0.197	0.084	0.008	0.004	0.040
×	Text+Number	0.001		√	√	0.332	0.110	0.167	0.178	0.078	0.238	0.081	-0.011	-0.009	0.076
×	Text+Number	0.001		√		0.311	0.097	-0.279	-0.211	0.112	0.243	0.085	-0.008	-0.005	0.096
√	Text+Number	0.1	√	√	√	0.065	0.004	0.411	0.287	0.023	0.056	0.004	0.519	0.527	**0.022**
√	Text+Number	0.01	√	√	√	**0.061**	**0.004**	0.484	**0.362**	0.024	**0.054**	**0.004**	**0.531**	**0.540**	0.023
√	Text+Number	0.001	√	√	√	0.061	0.004	**0.502**	0.354	**0.024**	0.056	0.004	0.517	0.523	0.023
√	Text+Number	0.001	√	√		0.066	0.004	0.371	0.321	0.027	0.057	0.004	0.521	0.528	0.026

Table 2: Our method's ablation experiments. Our method's design rationality was tested using the GDSC2 dataset and employing TransEDRP as a drug-response fusion encoder. All numerical values in the table are rounded to three significant figures. The best performer is highlighted in bold.

Model	Training time(s)	Params(M)	Gain w. CLDR(%)
tCNNs	2.35	233	7.88
DeepCDR	1.69	179	17.83
DeepTTC	1.28	1573	19.49
GraphDRP	1.58	529	14.29
GratransDRP	1.71	2361	13.04
TransEDRP	0.98	2244	31.46
AVG w. CLDR	+4.21	+5811	+17.3

Table 3: Computational analysis and performance comparison of different models under the CLDR framework. **Training time** signifies the duration required for processing one batch, while **Gain w. CLDR** denotes the level of performance improvement within the CLDR.

vanced DRP framework in drug discovery research. To validate that our framework improves numerical values and increases drug screening hit rates, we conducted a case study analysis as illustrated in Figure 4. We used unknown drugs as the screen molecules in our test dataset during preclinical drug screening.

Then, we compared the screening results of different methods with and without incorporating our framework. According to $HR@\mathrm{n}$ in recommendation systems [Lee *et al.*, 2010], we defined a novel evaluation measure specifically designed for drug screening, denoted as $Shot@\mathrm{x}$:

$$Shot@\mathrm{x} = \frac{\sum_{c=1}^{N_{cell}} Hit(\mathrm{T}_c^{top_1}, \mathrm{P}_c^{top_\mathrm{x}})}{N_{cell}}. \tag{11}$$

where T and P represent the true labels and predicted results, respectively. top_x represents top number of drugs with the smallest drug response values, N_{cell} represents the number of c cell lines in the test dataset, and $Hit(\cdot,\cdot)$ is a counting function that is 1 when the two inputs are equal and 0 otherwise.

We present the capabilities of three representative methods, tCNNs, GraphDRP, and TransEDRP, under the original strategy (i.e., MSE) and the CLDR framework. As shown in Figure 4, We assessed the predicted IC_{50} rankings of multiple drugs using the model with the actual values in order. Then, the probability of scoring a hit with the first optimal drug when recommending x compounds.

Due to the application of the CLDR framework, the probability of the model scoring a hit with the first optimal drug was consistently above 32%. In other words, for a set of screened compounds, our model achieves a hit rate of about 32% on a single chance for about 320 cell lines, and it has a cumulative hit probability of between 46% and 51% on the second attempt. Notably, as the number of x attempts increases, the results between methods become more similar; however, the number of attempts is limited in real drug studies. Thus, the DRP task requires the model to efficiently recommend the most effective drug in as few attempts as possible. In this context, our method improves the success rate by about 10% compared to traditional methods when only one attempt is available. This is an astounding development.

5 Conclusion

In this paper, we propose the CLDR framework, a contrastive learning framework with natural language supervision for DRP. The CLDR framework converts regression labels into text, which is merged with the drug response captions as a second modality for the samples in contrast to traditional encoding modalities. To enhance the continuous representation capability of the numerical text, the CN-KG was proposed to constrain the caption encoder's ability to perceive continuous values. In addition, we provided detailed theoretical evidence of the CLDR method's validity and conducted validation experiments on the GDSC2 dataset. This demonstrated the CLDR framework's ability to establish a link between drug-response data and valuable labeled text, improving the PDD's generalizability and success rate.

Acknowledgments

The work of Wenbin Hu was supported by the National Key Research and Development Program of China (2023YFC2705700). This work was supported in part by the Natural Science Foundation of China (No. 82174230), Artificial Intelligence Innovation Project of Wuhan Science and Technology Bureau (No. 2022010702040070), Natural Science Foundation of Shenzhen City (No. JCYJ20230807090211021).

References

[Agustito *et al.*, 2023] Denik Agustito, Krida Singgih Kuncoro, Istiqomah Istiqomah, and Agus Hendriyanto. Construction of ordinal numbers and arithmetic of ordinal numbers. *JTAM (Jurnal Teori dan Aplikasi Matematika)*, 7(3):781–792, 2023.

[Bai *et al.*, 2023] Peizhen Bai, Filip Miljković, Bino John, and Haiping Lu. Interpretable bilinear attention network with domain adaptation improves drug–target prediction. *Nature Machine Intelligence*, 5(2):126–136, 2023.

[Barretina *et al.*, 2012] Jordi Barretina, Giordano Caponigro, Nicolas Stransky, Kavitha Venkatesan, Adam A Margolin, Sungjoon Kim, Christopher J Wilson, Joseph Lehár, Gregory V Kryukov, Dmitriy Sonkin, et al. The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, 483(7391):603–607, 2012.

[Bébéar and Robertson, 1996] Christiane Bébéar and JA Robertson. Determination of minimal inhibitory concentration. *Molecular and diagnostic procedures in mycoplasmology*, 2:189–197, 1996.

[Begni *et al.*, 2021] Veronica Begni, Alice Sanson, Alessia Luoni, Federica Sensini, Ben Grayson, Syeda Munni, Joanna C Neill, and Marco A Riva. Towards novel treatments for schizophrenia: molecular and behavioural signatures of the psychotropic agent sep-363856. *International Journal of Molecular Sciences*, 22(8):4119, 2021.

[Bordes *et al.*, 2013] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.

[Chen *et al.*, 2023] Lili Chen, Suling Huang, Yangliang Ye, Yu Shen, Tifei Xu, Li Qin, Lili Du, Ying Leng, and Jianhua Shen. Phenotypic screening-based drug discovery of furan-2-carboxylic acid derivatives for the amelioration of type 2 diabetes mellitus (t2dm). *European Journal of Medicinal Chemistry*, 246:114994, 2023.

[Chu *et al.*, 2023] Thang Chu, Thuy Trang Nguyen, Bui Duong Hai, Quang Huy Nguyen, and Tuan Nguyen. Graph transformer for drug response prediction. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 20(2):1065–1072, 2023.

[Cohen *et al.*, 2009] Israel Cohen, Yiteng Huang, Jingdong Chen, Jacob Benesty, Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. Pearson correlation coefficient. *Noise reduction in speech processing*, pages 1–4, 2009.

[Duan *et al.*, 2021] Hanyu Duan, Yi Yang, and Kar Yan Tam. Learning numeracy: a simple yet effective number embedding approach using knowledge graph. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2597–2602, 2021.

[Eder *et al.*, 2014] Jörg Eder, Richard Sedrani, and Christian Wiesmann. The discovery of first-in-class drugs: origins and evolution. *Nature Reviews Drug Discovery*, 13(8):577–587, 2014.

[Fang *et al.*, 2022] Yin Fang, Qiang Zhang, Haihong Yang, Xiang Zhuang, Shumin Deng, Wen Zhang, Ming Qin, Zhuo Chen, Xiaohui Fan, and Huajun Chen. Molecular contrastive learning with chemical element knowledge graph. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 3968–3976, 2022.

[Gunel *et al.*, 2020] Beliz Gunel, Jingfei Du, Alexis Conneau, and Veselin Stoyanov. Supervised contrastive learning for pre-trained language model fine-tuning. In *International Conference on Learning Representations*, 2020.

[Ji *et al.*, 2023] Wenlong Ji, Zhun Deng, Ryumei Nakada, James Zou, and Linjun Zhang. The power of contrast for feature learning: A theoretical analysis. *Journal of Machine Learning Research*, 24(330):1–78, 2023.

[Jiang *et al.*, 2022] Likun Jiang, Changzhi Jiang, Xinyu Yu, Rao Fu, Shuting Jin, and Xiangrong Liu. DeepTTA: a transformer-based model for predicting cancer drug response. *Briefings in Bioinformatics*, 23(3), 03 2022. bbac100.

[Khosla *et al.*, 2020] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020.

[Kim *et al.*, 2023] Madeline Kim, Ester Del Duca, Julia Cheng, Britta Carroll, Paola Facheris, Yeriel Estrada, Amy Cha, John Werth, Robert Bissonnette, Karl Nocka, et al. Crisaborole reverses dysregulation of the mild to moderate atopic dermatitis proteome toward nonlesional and normal skin. *Journal of the American Academy of Dermatology*, 89(2):283–292, 2023.

[Lee *et al.*, 2010] Dongjoo Lee, Sung Eun Park, Minsuk Kahng, Sangkeun Lee, and Sang-goo Lee. *Exploiting Contextual Information from Event Logs for Personalized Recommendation*, pages 121–139. Springer Berlin Heidelberg, 2010.

[Lee *et al.*, 2022] Eunjoo Lee, Jiho Yoo, Huisun Lee, and Seunghoon Hong. Metadta: Meta-learning-based drug-target binding affinity prediction. In *ICLR2022 Machine Learning for Drug Discovery*, 2022.

[Li and Hu, 2022] Kun Li and Wenbin Hu. Transedrp: Dual transformer model with edge emdedded for drug respond prediction, 2022.

[Li *et al.*, 2021] Tianjiao Li, Xing-Ming Zhao, and Limin Li. Co-vae: Drug-target binding affinity prediction by co-regularized variational autoencoders. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):8861–8873, 2021.

[Liu *et al.*, 2017] Weiwei Liu, Ivor W Tsang, Müller Klaus-Robert, et al. An easy-to-hard learning paradigm for multiple classes and multiple labels. *Journal of Machine Learning Research*, 18(94):1–38, 2017.

[Liu *et al.*, 2019a] Lihao Liu, Qi Dou, Hao Chen, Jing Qin, and Pheng-Ann Heng. Multi-task deep model with margin ranking loss for lung nodule analysis. *IEEE transactions on medical imaging*, 39(3):718–728, 2019.

[Liu *et al.*, 2019b] Pengfei Liu, Hongjian Li, Shuai Li, and Kwong Sak Leung. Improving prediction of phenotypic drug response on cancer cell lines using deep convolutional network. *BMC Bioinformatics*, 20(1):1–14, 2019.

[Liu *et al.*, 2020] Qiao Liu, Zhiqiang Hu, Rui Jiang, and Mu Zhou. DeepCDR: a hybrid graph convolutional network for predicting cancer drug response. *Bioinformatics*, 36(26):i911–i918, 12 2020.

[Liu *et al.*, 2023] Changtong Liu, Yingchao Wang, Yixin Zeng, Zirong Kang, Hong Zhao, Kun Qi, Hongzhi Wu, Lu Zhao, and Yi Wang. Use of deep-learning assisted assessment of cardiac parameters in zebrafish to discover cyanidin chloride as a novel keap1 inhibitor against doxorubicin-induced cardiotoxicity. *Advanced Science*, page 2301136, 2023.

[Lu *et al.*, 2022] Wei Lu, Qifeng Wu, Jixian Zhang, Jiahua Rao, Chengtao Li, and Shuangjia Zheng. Tankbind: Trigonometry-aware neural networks for drug-protein binding structure prediction. *Advances in neural information processing systems*, 35:7236–7249, 2022.

[Maillard and Pascoe, 2023] Jean-Yves Maillard and Michael Pascoe. Disinfectants and antiseptics: Mechanisms of action and resistance. *Nature Reviews Microbiology*, pages 1–14, 2023.

[Moffat *et al.*, 2017] John G Moffat, Fabien Vincent, Jonathan A Lee, Jörg Eder, and Marco Prunotto. Opportunities and challenges in phenotypic drug discovery: an industry perspective. *Nature reviews Drug discovery*, 16(8):531–543, 2017.

[Nguyen *et al.*, 2022] Tuan Nguyen, Giang T. T. Nguyen, Thin Nguyen, and Duc-Hau Le. Graph convolutional networks for drug response prediction. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 19(1):146–154, 2022.

[Nichol *et al.*, 2022] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob Mcgrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In *International Conference on Machine Learning*, pages 16784–16804. PMLR, 2022.

[Ogutu *et al.*, 2023] Bernhards Ogutu, Adoke Yeka, Sylvia Kusemererwa, Ricardo Thompson, Halidou Tinto, Andre Offianan Toure, Chirapong Uthaisin, Amar Verma, Afizi Kibuuka, Moussa Lingani, et al. Ganaplacide (kaf156) plus lumefantrine solid dispersion formulation combination for uncomplicated plasmodium falciparum malaria: an open-label, multicentre, parallel-group, randomised, controlled, phase 2 trial. *The Lancet Infectious Diseases*, 23(9):1051–1061, 2023.

[Park *et al.*, 2020] Chanhee Park, Jinuk Park, and Sanghyun Park. Agcn:attention-based graph convolutional networks for drug-drug interaction extraction. *Expert Systems with Applications*, 159:113538, 2020.

[Sedgwick, 2014] Philip Sedgwick. Spearman's rank correlation coefficient. *BMJ: British Medical Journal (Online)*, 349, 2014.

[Vincent *et al.*, 2022a] Fabien Vincent, Arsenio Nueda, Jonathan Lee, Monica Schenone, Marco Prunotto, and Mark Mercola. Phenotypic drug discovery: recent successes, lessons learned and new directions. *Nature Reviews Drug Discovery*, 21(12):899–914, 2022.

[Vincent *et al.*, 2022b] Fabien Vincent, Arsenio Nueda, Jonathan Lee, Monica Schenone, Marco Prunotto, and Mark Mercola. Publisher correction: Phenotypic drug discovery: recent successes, lessons learned and new directions. 21:541–541, 2022.

[Wang *et al.*, 2022] Yuyang Wang, Jianren Wang, Zhonglin Cao, and Amir Barati Farimani. Molecular contrastive learning of representations via graph neural networks. *Nature Machine Intelligence*, 4(3):279–287, 2022.

[Wang *et al.*, 2023] Yuxuan Wang, Ying Xia, Junchi Yan, Ye Yuan, Hong-Bin Shen, and Xiaoyong Pan. Zerobind: a protein-specific zero-shot predictor with subgraph matching for drug-target interactions. *Nature Communications*, 14(1):7861, 2023.

[Yang *et al.*, 2013] Wanjuan Yang, Jorge Soares, Patricia Greninger, and et al. Genomics of drug sensitivity in cancer (gdsc): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic acids research*, 41:D955—61, January 2013.

[Zha *et al.*, 2023] Kaiwen Zha, Peng Cao, Jeany Son, Yuzhe Yang, and Dina Katabi. Rank-n-contrast: Learning continuous representations for regression. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 17882–17903. Curran Associates, Inc., 2023.

[Zhang *et al.*, 2022] Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. Contrastive learning of medical visual representations from paired images and text. In *Machine Learning for Healthcare Conference*, pages 2–25. PMLR, 2022.