Pisces: A cross-modal contrastive learning approach to synergistic drug combination prediction

Jiacheng Lin*[1], Hanwen Xu*[2], Addie Woicik[2], Jianzhu Ma[3], and Sheng Wang†[2]

[1]Department of Automation, Tsinghua University, Beijing, China
[2]Paul G. Allen School of Computer Science and Engineering, University of Washington, WA
[3]Institute for Artificial Intelligence, Peking University, Beijing, China.

Abstract

Drug combination therapy is a promising solution to many complicated diseases. Since experimental measurements cannot be scaled to millions of candidate combinations, many computational approaches have been developed to identify synergistic drug combinations. While most of the existing approaches either use SMILES-based features or molecular-graph-based features to represent drugs, we found that neither of these two feature modalities can comprehensively characterize a pair of drugs, necessitating the integration of these two types of features. Here, we propose Pisces, a cross-modal contrastive learning approach for synergistic drug combination prediction. The key idea of our approach is to model the combination of SMILES and molecular graphs as four views of a pair of drugs, and then apply contrastive learning to embed these four views closely to obtain high-quality drug pair embeddings. We evaluated Pisces on a recently released GDSC-Combo dataset, including 102,893 drug combinations and 125 cell lines. Pisces outperformed five existing drug combination prediction approaches under three settings, including vanilla cross validation, stratified cross validation for drug combinations, and stratified cross validation for cell lines. Our case study and ablation studies further confirmed the effectiveness of our novel contrastive learning framework and the importance of integrating the SMILES-based features and the molecular-graph-based features. Pisces has obtained the state-of-the-art results on drug synergy prediction and can be potentially used to model other pairs of drugs applications, such as drug-drug interaction. **Availability:** Implementation of Pisces and comparison approaches can be accessed at `https://github.com/linjc16/Pisces`.

1 Introduction

Drug combination therapy is a promising solution to many complex diseases, such as breast cancer [1–3], colorectal cancer [4, 5], Alzheimer's disease [6], and diabetes [7, 8]. However, previous studies have also pointed out the rareness of synergistic drug combinations [9–11]. Since experimentally testing millions of candidate combinations is not scalable, there is a pressing need to develop computational approaches to identify synergistic drug combinations [12–16].

We focus on cancer drug synergistic prediction and follow existing approaches [17] to form the synergistic drug combination prediction problem as a triplet classification problem. Each triplet consists of two drugs and a cancer cell line and will be classified into synergistic or not. Each cell line is represented using its genomics features, such as gene expression and somatic mutation.

*Contributed equally
†Corresponding authors. Emails: `swang@cs.washington.edu`

A key technical challenge is to derive an effective representation for a pair of drugs. Simplified molecular-input line-entry system (SMILES) sequences [18] and molecular graphs [19–21] are the two major modalities for representing a drug. Both of them have strengths and limitations: SMILES sequences are easier to embed by leveraging the recent progress in natural language processing, but are ambiguous on drugs with complex structures; molecular graphs precisely characterize the molecular information, but large graphs (i.e., large diameter) are often hard to embed [22]. This dilemma is even more severe when we want to embed a pair of drugs since one might be better represented using SMILES sequences and the other might be better represented using molecular graphs. As we showed in the experiments, a simple concatenation of these two kinds of features yields undesirable results.

To address this problem, we propose Pisces, a cross-modal contrastive learning approach for drug synergy prediction. Our intuition is that the SMILES sequence modality and the molecular graph modality complement each other, and thus should be integrated. To realize this intuition, we have developed a cross-modal contrastive learning framework. Contrastive learning has recently obtained great success in computer vision where an image is embedded closely to its augmentation (e.g., the rotated review) [23–30]. However, contrastive learning has never been applied to pairs of drugs since it is unclear what should be a proper augmentation. We propose to create four augmented views for each drug pair based on the combination of the SMILES sequence and the molecular graph modality. We hypothesize that these four combinations can offer a comprehensive view of a drug pair, thus enhancing the drug synergistic prediction.

We evaluated our method on a recently published large-scale cancer drug synergy dataset GDSC-Combo [9], which covers 102,893 drug combinations spanning over 63 drugs and 125 cell lines [2]. We first observed a substantial discrepancy between the prediction performance by using two different modalities. By contrasting these two modalities, Pisces substantially outperformed five existing drug combination prediction approaches under vanilla cross validation setting, stratified cross validation for drug combinations setting, and stratified cross validation for cell lines setting. Finally, we found that two drugs from the top performed drug pairs favored different modalities, again confirming the effectiveness of integrating SMILES and molecular graph modalities. Despite the large amount of triplet combinations that have been measured in GDSC-Combo [9], it still only covers 20.7% of all possible triplet combinations. Pisces offers an *in silico* solution to massively generalize these *in vitro* measurements. In addition to drug synergy prediction, Pisces can be broadly applied to other applications that require the modeling of drug pairs, such as drug-drug interaction prediction, as well as further integrating other drug modalities.

2 Methods

2.1 Problem setting

We model the synergistic drug combination prediction problem as a classification task. Given the drug set $\mathcal{D}$ and the cell line set $\mathcal{C}$, each drug combination in the dataset is denoted as a triplet $(d_A, d_B, c)_i$, where $d_A \in \mathcal{D}$ and $d_B \in \mathcal{D}$ represent possible pairs of two different drugs, and $c \in \mathcal{C}$ denotes the cell line. The prediction of drug synergy is modeled as:

$$\hat{y}_i = f_\theta((d_A, d_B, c)_i), \tag{1}$$

where $f_\theta : \mathcal{D} \times \mathcal{D} \times \mathcal{C} \rightarrow [0, 1]$ is a learned mapping function with θ as the learnable parameters. The output $\hat{y}_i \in [0, 1]$ is the probability of the synergistic drug combination prediction. For each drug combination, the binary synergy label y_i with a value of 1 indicates synergy, otherwise no synergy.

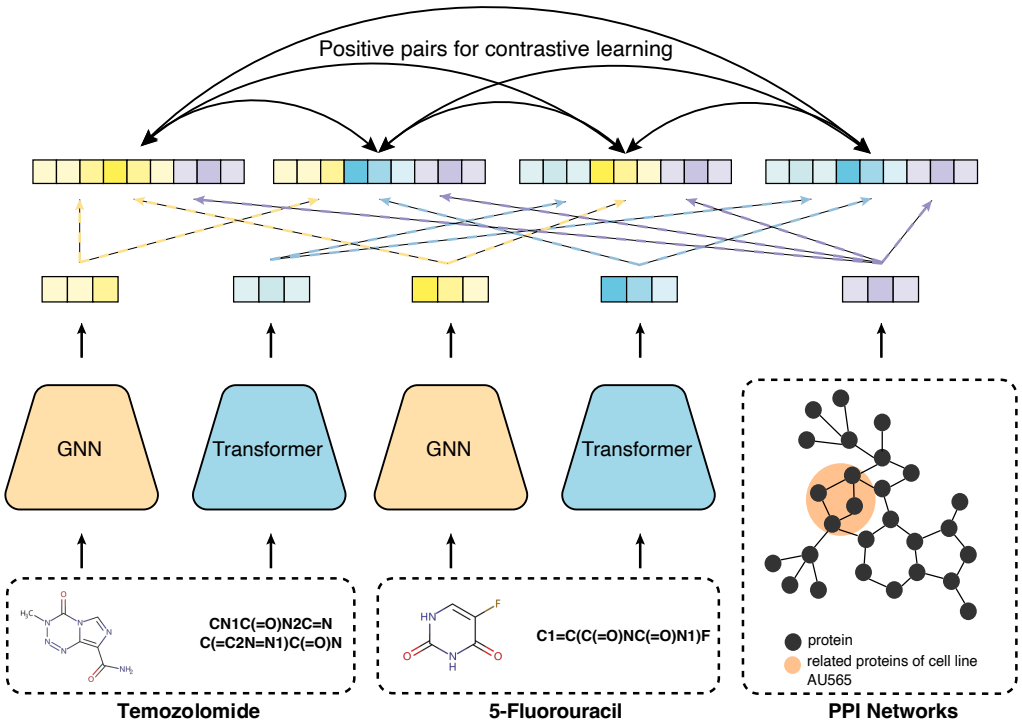

Figure 1: **Flowchart of Pisces.** Pisces considers both the SMILES sequence and the molecular graph of each drug. It first uses Transformer to embed SMILES sequences and graph neural networks to embed molecular graphs. Pisces embeds each cell line by aggregating neighbors of over-expressed genes in the protein-protein interaction network. It then concatenates the SMILES embedding and the graph embedding between two drugs, which creates four different views for each drug combination. These four different views are treated as augmentations in contrastive learning.

We aim to find the best f_θ to enable the predictions on test set $\{\hat{y_i}\}_{test}$ to approximate the labels $\{y_i\}_{test}$. Moreover, following the previous works [12, 14–16], we view the prediction probability greater than 0.5 as synergy, otherwise no synergy.

Furthermore, for drug representations, we define the SMILES features and molecular graph features of drugs as $\mathbf{s}$ and $\mathbf{g}$. Each SMILES string can be represented as $\mathbf{s} = (s_1, \ldots, s_l)$ where s denotes the tokens and l denotes the string length. Molecular graph features can be represented as $\mathbf{g} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = (g_1, \ldots, g_n)$ are the n atoms in the molecular graph and $\mathcal{E}$ are the bonds between atoms. The cell line features are provided as a fixed-size vector $\mathbf{c} \in \mathbb{R}^M$ for M genes, with each dimension representing transcripts per million (TPM) for one gene. Since Pisces uses both kinds of features, the input would be $(\mathbf{s_A}, \mathbf{g_A}, \mathbf{s_B}, \mathbf{g_B}, \mathbf{c})_i$.

2.2 Overview of Pisces

We propose a cross-modal contrastive learning approach Pisces, as shown in (**Figure 1**). Pisces makes full use of the complementary information in the structural and SMILES-based inputs. A graph neural network [31] and Transformer [32] model are applied to encode each drug's graph-based features and SMILES-based features respectively. Pisces also encodes cell line features by assembling over-expressed genes and their associated neighbors with a Multi-layer Perceptron

(MLP) [33]. We then concatenate the drug feature vectors and cell feature vectors to produce the drug combination representations. We introduce a contrastive learning loss term to integrate the feature representations. Finally, Pisces uses a binary classifier to predict the drug synergy label.

2.3 Embedding drugs using SMILES-based features

Transformer has obtained great success in processing string data, such as natural languages [34, 35], combinatorial optimization [36–38], protein sequences [39, 40] and molecular features [22, 41]. We first encode the SMILES string using a Transformer [32] model. Specifically, we used the encoder architecture with multi-head self-attention modules. Each encoding layer includes one self-attention module and a feed forward network (FFN). Skip-connections are added to each module to build up the residual blocks [42]. The special token `[CLS]` is added to the beginning of each SMILES string $\mathbf{s}$. We use the contextualized embedding corresponding to `[CLS]` from the output as the SMILES representation and denote the SMILES embeddings of drug A and drug B as $\mathbf{z}_s^A$ and $\mathbf{z}_s^B$.

2.4 Embedding drugs using molecular graph-based features

The molecular graph-based feature is encoded with a DeeperGCN encoder [31], which is stacked by identical layers. It takes a set of node vectors and an adjacency matrix as input and propagates information between nodes based on the graph structure [43]. In each layer, the output from the previous layer is fed and processed sequentially by a layer normalization module [44], a nonlinear ReLU module [45], and a residual graph convolutional block where each node will aggregate both its neighbor edge and neighbor nodes with an aggregation module. Moreover, the aggregation module comprises the concatenation of maximize, minimize and average pooling. We denoted the output node vectors as $(\widetilde{g}_1, \ldots, \widetilde{g}_n)$. Then we calculated molecular graph-based embeddings using mean pooling and max pooling.

$$\mathbf{z}_g^A = \text{Concat}(\text{MeanPool}((\widetilde{g}_1^A, \ldots, \widetilde{g}_n^A)), \text{MaxPool}((\widetilde{g}_1^A, \ldots, \widetilde{g}_n^A))) \tag{2}$$

$$\mathbf{z}_g^B = \text{Concat}(\text{MeanPool}((\widetilde{g}_1^B, \ldots, \widetilde{g}_n^B)), \text{MaxPool}((\widetilde{g}_1^B, \ldots, \widetilde{g}_n^B))) \tag{3}$$

Finally, for drug A and drug B, we produce the SMILES-based representations $\mathbf{z}_s^A$ and $\mathbf{z}_s^B$ and the molecular graph-based representations $\mathbf{z}_g^A$ and $\mathbf{z}_g^B$.

2.5 Embedding cell line features using gene expression level and PPI topological relationships

Based on the TPM, we set a threshold T and determine an over-expressed gene set $\tilde{G} = \{g = 1, \ldots, M_H : \mathbf{c}_g \geq T\}$. Then, we find the k nearest neighbors for each gene $g \in \tilde{G}$ in the protein-protein interaction (PPI) [46] network, defined as $\mathcal{N}_k(g)$, and finally compute a candidate set $\hat{G} = \bigcup_{g \in \tilde{G}} \{g\} \cup \mathcal{N}_k(g)$. We then produce the embeddings of cell line features as

$$\mathbf{z}_c = \text{MLP}(\text{Concat}(\text{MeanPool}(\{\mathbf{c}_g | g \in \hat{G}\}), \text{MaxPool}(\{\mathbf{c}_g | g \in \hat{G}\})), \tag{4}$$

where $\mathbf{c}_g$ is a learnable embedding for each protein.

2.6 Improved synergistic prediction using contrastive learning

Contrastive learning has been broadly used in both supervised and unsupervised learning [23, 25, 47]. The key idea behind contrastive learning is to push positive sample pairs to be close within the embedding space while pushing negative pairs far away from each other. In this formulation, positive pairs are instances defined to be similar in some important way, while negative pairs keep

the space from collapsing to a single point. By defining positive pairs as two triplets from the same drug combinations while negative pairs as triplets from different drug combinations, we can apply contrastive learning to integrate different features of the drug combination learning task. In particular, we randomly select one type of feature embedding (SMILES or molucule graphs) for each drug, i.e.,

$$\mathbf{z}_i(u,v) = (\mathbf{z}_u^A, \mathbf{z}_v^B, \mathbf{c})_i \quad u,v \in \{s,g\} \tag{5}$$

where subscript s and g represent modalities of SMILES and molecule graphs. We use $\mathbf{z}_i(u,v)$ and $\mathbf{z}_i(u',v')$ to denote two different random choice of the i-th drug combination in one training batch. We choose infoNCE [48] as the contrastive learning loss in our study, defined as

$$\mathcal{L}^{(infoNCE)} = -\sum_{i=1}^{n_{batch}} \log\left(\frac{\exp((\mathbf{z}_i(u,v))^{\mathrm{T}} \cdot \mathbf{z}_i(u',v'))}{\sum_{j=1}^{n_{batch}} \exp((\mathbf{z}_i(u,v))^{\mathrm{T}} \cdot \mathbf{z}_j(u',v'))}\right), \tag{6}$$

where n_{batch} denotes the number of samples in one training batch. Finally, Pisces uses one linear layer and MLP layer P_ψ to generate the output $\hat{y}$, trained with binary cross entropy loss for this classification task as

$$\hat{y}_i = \sigma(P_\psi([\mathrm{Linear}(\mathbf{z}_s^A, \mathbf{z}_g^A), \mathrm{Linear}(\mathbf{z}_s^B, \mathbf{z}_g^B), \mathbf{c}])) \tag{7}$$

$$\mathcal{L}^{(CE)} = -\sum_{i=1}^{n_{batch}} (y_i \log \hat{y}_i + (1-y_i)\log(1-\hat{y}_i)). \tag{8}$$

where σ represents the Sigmoid [49] activation function. To better enhance the consistency between different types of features, we introduce an additional auxiliary loss. First, we define the auxiliary score as the following:

$$\widetilde{y}_i = \sigma(\mathrm{Mean}[\mathrm{P}_\psi(\mathbf{z}_i(s,s)), \mathrm{P}_\psi(\mathbf{z}_i(g,g))]), \tag{9}$$

where σ denotes the Sigmoid activation function. Therefore the auxiliary loss is defined to minimize between the auxiliary scores and classification outputs:

$$\mathcal{L}^{(Aux)} = \sum_{i=1}^{n_{batch}} (\widetilde{y}_i - \hat{y}_i)^2, \tag{10}$$

Finally, we combine infoNCE loss in (6), classification loss in (8) and auxiliary loss in (10) as our final cross-modal contrastive learning loss with λ_1, λ_2 as hyperparameters:

$$\mathcal{L} = \mathcal{L}^{(CE)} + \lambda_1 \mathcal{L}^{(infoNCE)} + \lambda_2 \mathcal{L}^{(Aux)}. \tag{11}$$

3 Experimental settings

Dataset. We obtained a recently released drug combination dataset from Genomics of Drug Sensitivity in Cancer (GDSC-combo) [9]. Since there are two replicates for each drug pair cell line triplet, We first obtained the samples formulated by (Drug A, Drug B, cell line, synergy or not) tuples by following the rule: for each triple (Drug A, Drug B, cell line) in the original dataset, we view it as synergy if there exists one, otherwise no synergy. We excluded the samples that have a combination of three or more drugs. Finally, we obtained 102,893 samples, including 63 drugs and 125 cell lines. We observed that this is a highly imbalanced dataset where only 5,362 samples are synergistic. We evaluated three settings, including the vanilla cross validation (CV) setting, the

stratified cross validation for drug combinations setting and the stratified cross validation for cell lines setting. Specifically, in the stratified CV for drug combinations setting, drug combinations in the test set have never been seen in the training set. Likewise, in the stratified CV for cell lines setting, cell lines in the test set have never been seen in the training set. Moreover, PRODeepSyn, GraphSynergy and Pisces share the PPI network which is obtained from the STRING database [50] as in [14].

Comparison approaches. We compare Pisces to five existing drug synergistic prediction approaches. **PRODeepSyn** [14] takes molecular fingerprints and descriptors for drugs as the inputs. The molecular fingerprint is a 256-dimensional binary vector for each drug, representing the existence of a set of predefined substructures [51]. Then drug descriptor is a 200-dimensional real vector used in [52], representing molecules' physical or chemical properties of interest, such as lipophilicity or molecular refractivity. Both the fingerprints and the descriptors are obtained from RDKit [53]. Cell lines are embedded using Graph Convolutional Networks [54] by integrating the protein-protein interaction (PPI) network with the gene expression vector. Finally, PRODeepSyn then used an MLP to predict drug synergy. **AuDNNsynergy** [12] utilizes the same features as those in PRODeepSyn. Different from PRODeepSyn, it trains three autoencoders to predict drug synergy. **DeepSynergy** [13] takes molecular fingerprints, descriptors and drug-target interactions for drugs and Transcripts Per Million (TPM) for cell lines as the inputs. All these features are then fed into an MLP to predict drug synergy. **GraphSynergy** [15] utilizes Graph Convolutional Networks to extract drug and cell line features from the PPI network. These features are then fed into an MLP to predict drug synergy. **DeepDDS** [16] takes molecule graph for each drug and TPM for cell lines as the inputs. It then uses an MLP to predict drug synergy. Notably, these comparison approaches often relies on very different features that might not be available in any dataset. The original implementations of these methods often use hard-coded or pre-processed features that cannot be generalized to GDSC-combo. The details of pro-processing are also not comprehensively revealed and make it hard to reproduce their results. Therefore, we have re-implemented many of them and use the same feature pre-processing to increase the usability for fair comparison. We have made our implementations of all five comparison approaches available for future studies.

Model architecture. We utilized DeeperGCN and Transformer to embed drugs. Specifically, DeeperGCN is stacked by 6 layers of the graph convolutional blocks, where the dimension of the hidden size was set to 384. The hidden size, the FFN size, and the number of Transformer encoder layers was set to 512, 1024 and 6. The number of the attention heads in Transformer was set to 4. The dropout rates of DeeperGCN and Transformer were both 0.1. When determining over-expressed genes, the threshold T was set to 400. For the comparison approaches, we followed the original papers for the model architectures and the training details.

Training details. We trained our models using the Adam [55] optimizer with $\beta_1 = 0.9, \beta_2 = 0.98, \epsilon = 10^{-6}$, a weight decay of 0.01 and a batch of 128 for 100,000 training steps. We used a linear decay scheduler with 4,000 warm-up steps. We ran a grid search within $[1e-5, 5e-5, 8e-5, 1e-4, 5e-4]$ for the learning rate. λ_1, λ_2 were both set as 0.01.

Metrics. Since our dataset is highly imbalanced, we consider the following four metrics for evaluation: balanced accuracy (BACC) which is the average of sensitivity (true positive rate) and specificity (true negative rate), area under the precision-recall curve (AUPRC), F1 score and Cohen's Kappa statistic. All metrics are higher the better.

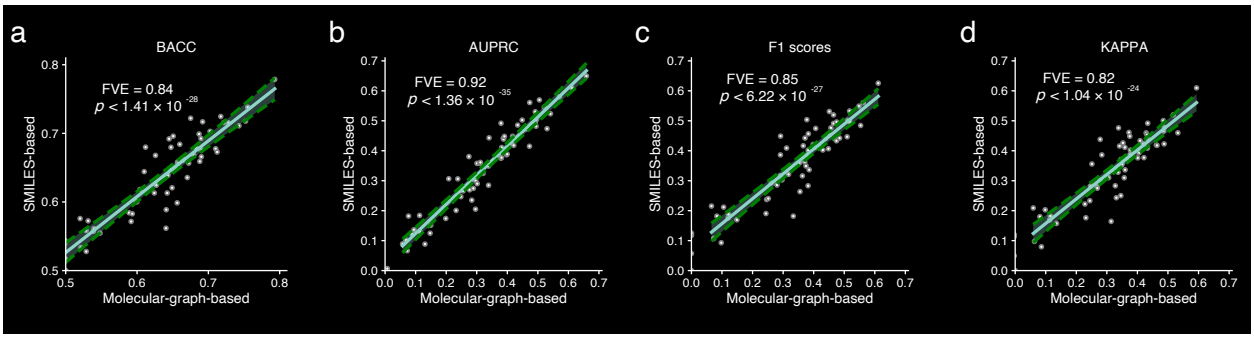

Figure 2: **Comparison between SMILES-based features and molecular-graph-based features. a-d**, Scatter plots comparing the performance of using SMILES-based feature and molecular graph-based feature in terms of BACC (**a**), AUPRC (**b**), F_1 (**c**), KAPPA (**d**). Each point is a drug. Shaded areas represent 95 % confidence intervals of a linear regression line. FVE stands for fraction of inter-species variance explained. p-values are obtained using F-test.

4 Results

4.1 SMILES-based and molecular-graph-based features are complementary for drug synergy prediction

Pisces is developed based on the hypothesis that SMILES-based features and molecular-graph-based features complement each other, thus an integration of them might enhance the drug synergy prediction. Therefore, we first sought to validate this hypothesis by examining the performance of using either SMILES-based or molecular-graph-based features. Despite the consistent performance between these two types of features in (**Figure 2**), we also observed that they showed substantial performance discrepancy on many drug combinations. For example, drug Sapitinib and drug Entinostat obtained a KAPPA score of 0.7897 when using SMILES-based features, while the KAPPA score drops to 0.3691 when using molecular-graph-based features. The number of drug combinations that can be better predicted by the SMILES-based features and the molecular-based features is comparable, necessitating the importance of considering both features.

4.2 Pisces achieves substantial improvement on drug synergy prediction

After validating our hypothesis that molecular graphs and SMILES provide complementary information, we next sought to compare Pisces to other drug synergy prediction approaches under three cross validation settings. On vanilla cross validation, we found that Pisces substantially outperformed all comparison approaches on all four metrics (**Figure 3**). For example, Pisces obtained a 0.4474 KAPPA score, which is 21.05% higher than the best comparison approach. Since Pisces is the only approach considering both types of drug features, the prominent performance of Pisces confirms the effectiveness of contrasting molecular-graph-based features and SMILES-based features.

The promising performance of Pisces on vanilla CV motivates us to further evaluate it in two more challenging settings. We first examined the stratified CV for drug combinations setting where all test drug pairs have never been seen in the training set (**Figure 3**). We found that the performances of all approaches dropped, confirming that this is a more challenging setting compared to vanilla CV. Nevertheless, our method still outperformed all comparison approaches and the improvement is even larger on this challenging setting than that on the vanilla CV setting. We attribute this improvement to Pisces' consideration of both types of features, which offers us a

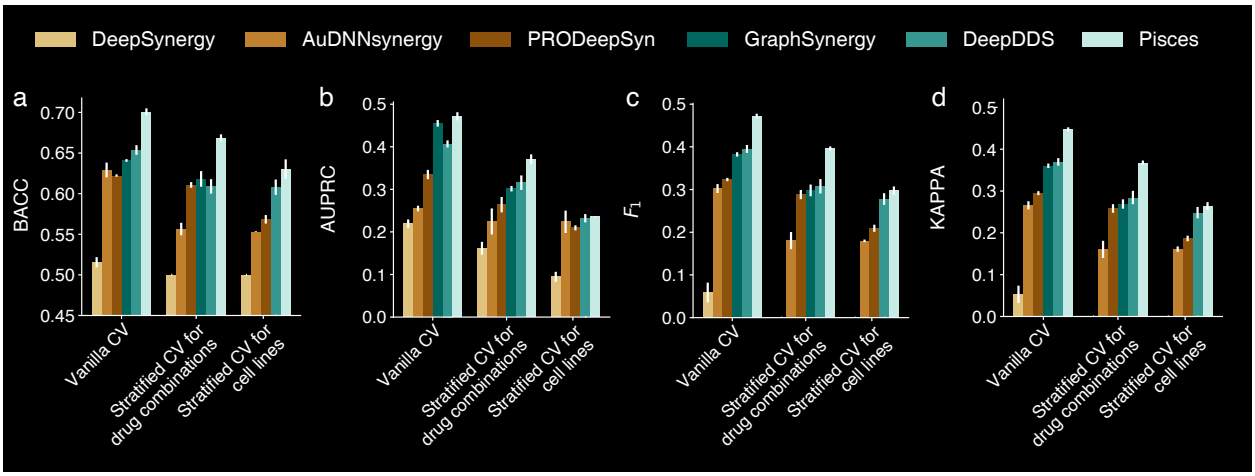

Figure 3: **Comparison on drug synergy prediction. a-d**, Bar plots comparing Pisces to five existing approaches under vanilla cross validation, stratified cross validation for drug combinations and stratified cross validation for cell lines, in terms of BACC (**a**), AUPRC (**b**), F_1 (**c**), KAPPA (**d**). GraphSynergy cannot be applied to the stratified cross validation for cell lines setting. Since DeepSynergy predicts all drug combinations as no synergy in stratified cross validation for drug combinations and cell lines setting, the F_1 and KAPPA values are zero there.

more robust drug combination representation that can be generalized to unseen drug combinations.

Finally, we evaluated the second challenging setting of stratified CV for cell lines, where all test cell lines have never been seen in the training set (**Figure 3**). This setting is much closer to real-world clinical applications since it can perform predictions for a new patient who has not been treated by any drug combinations. We again found that the performance of all methods dropped substantially. Nevertheless, Pisces still achieved the overall best performance, suggesting its applicability in real-world applications. Collectively, the prominent performance of Pisces on three different settings demonstrates the importance of integrating molecular-graph-based and SMILES-based features and the effectiveness of our cross-modal contrastive learning framework.

4.3 Pisces obtains larger improvement on drug combinations that favor different modalities

After observing the promising performance of Pisces, we then sought to understand what kind of drug combinations can obtain larger improvement using Pisces. We first visualized the embedding space of all test triplets by Pisces and two best-performed comparison approaches DeepDDS and GraphSynergy (**Figure 4**). We found that Pisces obtained a more visible pattern contrasting synergistic and non-synergistic triplets, further validating the promising performance of our method. We also observed a clear pattern on cell line OCUB-M, where our method achieved a large improvement compared to other approaches, suggesting the effectiveness of contrasting two types of modalities.

Next, we studied whether Pisces can obtain large improvement on a drug pair where two drugs favored different modalities. For each single drug, we determine the modality it favors using the same analysis as in (**Figure 2**). If both drugs in a combination favor the same modality (i.e., SMILES or molecular graph), we denote it as same in (**Figure 5a**). We found that drug pairs that favor different modalities have substantially larger improvement by Pisces than those favor the same modality, supporting our hypothesis that these drug pairs are relatively more challenging to model

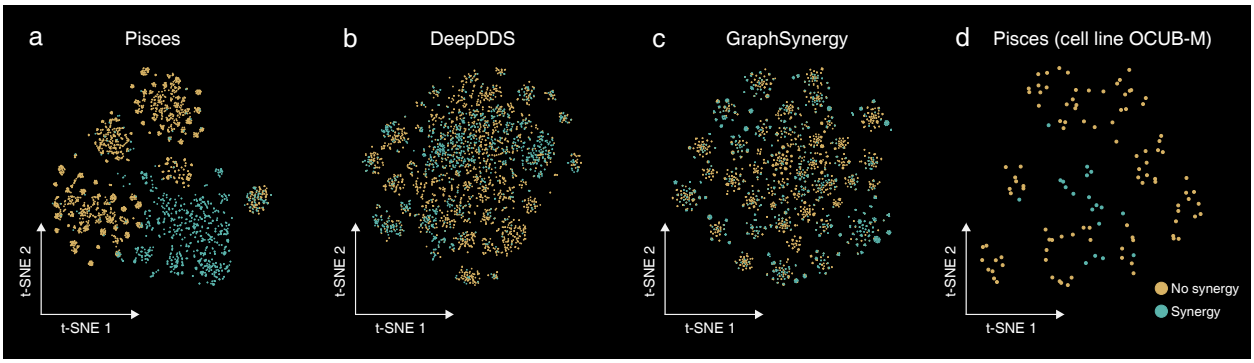

Figure 4: t-SNE visualization of the embedding space of Pisces on all cell lines (**a**), DeepDDS on all cell lines (**b**), GraphSynergy on all cell lines (**c**), and Pisces on cell line OCUB-M (**d**). Each point is a triplet of drug combination and cell line.

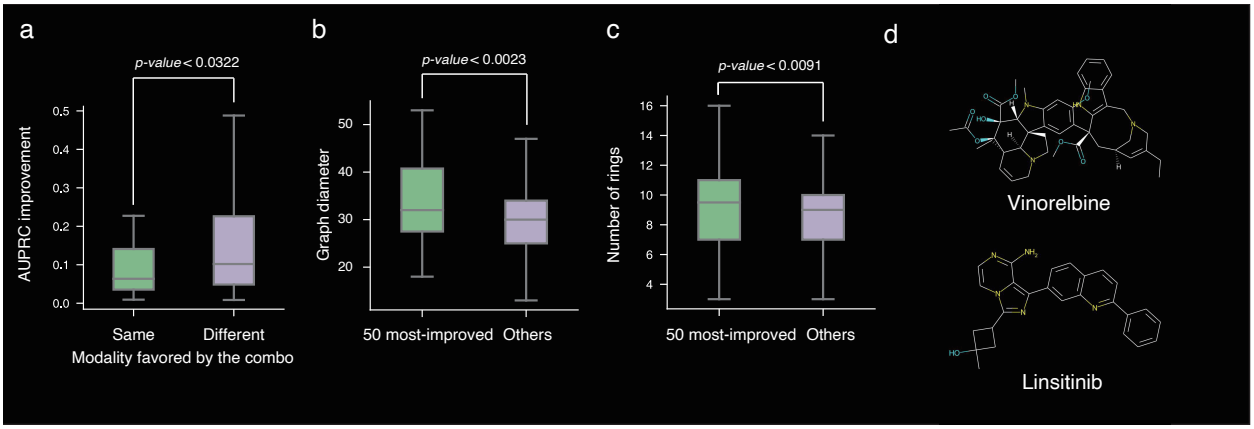

Figure 5: **Performance analysis of Pisces. a**, Box plot comparing the improvement of AUPRC by Pisces on drug pairs that favor the same modality to drug pairs that favor different modalities. **b,c,** Box plots comparing the graph diameter (**b**) and number of rings (**c**) between 50 most-improved drug combinations and other drug combinations. The improvement is calculated using the relative improvement between Pisces and the best comparison approach in terms of AUPRC. **d**, Molecular graphs of Vinorelbine and Linsitinib, on which Pisces obtained large improvements.

and our cross-modal contrastive learning approach can effectively embed them by integrating two modalities.

Finally, we investigated if molecular-graph properties are also related to the improvement of Pisces. We found that the 50 most-improved combinations by Pisces have significantly larger graph diameter (p-value < 0.0023) and number of rings in the graph (p-value < 0.0091) (**Figure 5b,c**). For instance, the drug pair Linsitinib and Vinorelbine, which obtained a 211.87% AUPRC improvements against the best comparison approach, have 31 graph diameter and 15 rings in total(**Figure 5d**). Large and complicated graphs are often difficult to embed using graph-based approaches. Pisces mitigated this issue by additionally considering SMILES-based features, thus resulting in a better performance on these combinations.

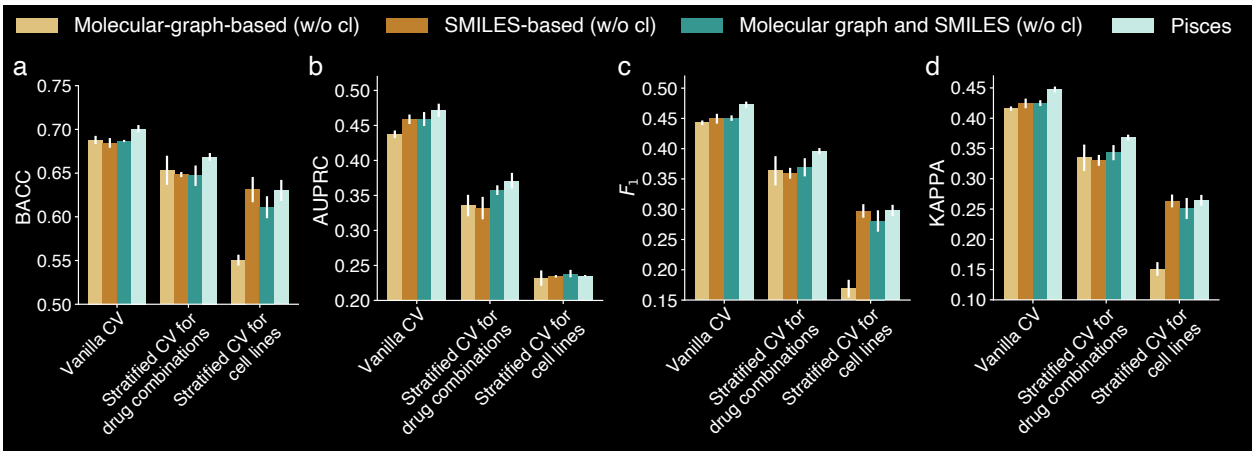

Figure 6: **Ablation studies**. **a-d,** Bar plots examining the importance of different types of features and contrastive learning in terms of BACC (**a**), AUPRC (**b**), F_1 (**c**), and KAPPA (**d**). w/cl denotes to using the proposed cross-modal contrastive learning. w/o cl denotes to not using the proposed cross-modal contrastive learning.

4.4 Ablation studies

Finally, we performed ablation studies to examine the importance of each component in Pisces (**Figure 6**). We first examined the importance of cross-modal contrastive learning by comparing Pisces to a baseline approach that used a simple concatenation to integrate molecular-graph-based and SMILES-based features (Molecular graph and SMILES (w/o cl)). We found that our method outperformed this approach on 11 out of 12 comparisons, indicating the effectiveness of contrastive learning. Next, we investigated the importance of using both types of features by comparing Pisces to a model that only uses molecular-graph-based features and a model that only uses SMILES-based features. Cross-modal contrastive learning cannot be applied to these two models. We found that Pisces substantially outperformed both methods, again confirming the importance of using both types of features. Interestingly, we found that the simple concatenation-based approach is only slightly better than these two models that only use one type of the features, further confirming the effectiveness of our cross-modal contrastive learning framework.

5 Conclusion and future work

We have developed a synergistic drug combination prediction approach Pisces. Based on the intuition that different drug combinations might favor different types of features, Pisces exploits cross-modal contrastive learning to integrate SMILES-based features and molecular-graph-based features. We have evaluated our method on a recently-published large-scale dataset GDSC-Combo [9] and observed that Pisces substantially outperformed five existing approaches. In addition to drug synergy prediction, our framework of contrasting different modalities can also be applied to other drug combination tasks, such as drug-drug side effect prediction. In the future, we would like to further improve Pisces by incorporating more drug molecule modalities, such as drug three-dimensional structure, drug textual descriptions, and pharmacodynamics features.

References

[1] Liu, T. *et al.* Combinatorial effects of lapatinib and rapamycin in triple-negative breast cancer cellscombined treatment in triple-negative breast cells. *Molecular cancer therapeutics* **10**, 1460–1469 (2011).

[2] Lee, J. *et al.* Effective breast cancer combination therapy targeting bach1 and mitochondrial metabolism. *Nature* **568**, 254–258 (2019).

[3] Lee, J. V. *et al.* Combinatorial immunotherapies overcome myc-driven immune evasion in triple negative breast cancer. *Nature communications* **13**, 1–12 (2022).

[4] Kopetz, S. *et al.* Encorafenib, binimetinib, and cetuximab in braf v600e–mutated colorectal cancer. *New England Journal of Medicine* **381**, 1632–1643 (2019).

[5] Elez, E. *et al.* Rnf43 mutations predict response to anti-braf/egfr combinatory therapies in brafv600e metastatic colorectal cancer. *Nature Medicine* 1–9 (2022).

[6] Schmitt, B., Bernhardt, T., Moeller, H.-J., Heuser, I. & Frölich, L. Combination therapy in alzheimer's disease. *CNS drugs* **18**, 827–844 (2004).

[7] Home, P. D. *et al.* Rosiglitazone evaluated for cardiovascular outcomes in oral agent combination therapy for type 2 diabetes (record): a multicentre, randomised, open-label trial. *The Lancet* **373**, 2125–2135 (2009).

[8] Amoaku, W., Saker, S. & Stewart, E. A review of therapies for diabetic macular oedema and rationale for combination therapy. *Eye* **29**, 1115–1130 (2015).

[9] Jaaks, P. *et al.* Effective drug combinations in breast, colon and pancreatic cancer cells. *Nature* **603**, 166–173 (2022).

[10] Flobak, Å. *et al.* A high-throughput drug combination screen of targeted small molecule inhibitors in cancer cell lines. *Scientific data* **6**, 1–10 (2019).

[11] O'Neil, J. *et al.* An unbiased oncology compound screen to identify novel combination strategies. *Molecular cancer therapeutics* **15**, 1155–1162 (2016).

[12] Zhang, T., Zhang, L., Payne, P. R. O. & Li, F. Synergistic drug combination prediction by integrating multi-omics data in deep learning models. *CoRR* **abs/1811.07054** (2018). URL `http://arxiv.org/abs/1811.07054. 1811.07054`.

[13] Preuer, K. *et al.* Deepsynergy: predicting anti-cancer drug synergy with deep learning. *Bioinform.* **34**, 1538–1546 (2018). URL `https://doi.org/10.1093/bioinformatics/btx806`.

[14] Wang, X. *et al.* Prodeepsyn: predicting anticancer synergistic drug combinations by embedding cell lines with protein-protein interaction network. *Briefings Bioinform.* **23** (2022). URL `https://doi.org/10.1093/bib/bbab587`.

[15] Yang, J., Xu, Z., Wu, W. K. K., Chu, Q. & Zhang, Q. Graphsynergy: a network-inspired deep learning model for anticancer drug combination prediction. *J. Am. Medical Informatics Assoc.* **28**, 2336–2345 (2021). URL `https://doi.org/10.1093/jamia/ocab162`.

[16] Wang, J., Liu, X., Shen, S., Deng, L. & Liu, H. Deepdds: deep graph neural network with attention mechanism to predict synergistic drug combinations. *Briefings Bioinform.* **23** (2022). URL `https://doi.org/10.1093/bib/bbab390`.

[17] Menden, M. P. *et al.* Community assessment to advance computational prediction of cancer drug combinations in a pharmacogenomic screen. *Nature communications* **10**, 1–17 (2019).

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

[19] Wu, Z. *et al.* Moleculenet: a benchmark for molecular machine learning. *Chemical science* **9**, 513–530 (2018).

[20] Kearnes, S., McCloskey, K., Berndl, M., Pande, V. & Riley, P. Molecular graph convolutions: moving beyond fingerprints. *Journal of computer-aided molecular design* **30**, 595–608 (2016).

[21] Pires, D. E., Blundell, T. L. & Ascher, D. B. pkcsm: predicting small-molecule pharmacokinetic and toxicity properties using graph-based signatures. *Journal of medicinal chemistry* **58**, 4066–4072 (2015).

[22] Zhu, J. *et al.* Dual-view molecule pre-training. *CoRR* **abs/2106.10234** (2021). URL `https://arxiv.org/abs/2106.10234`. `2106.10234`.

[23] Chen, T., Kornblith, S., Norouzi, M. & Hinton, G. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, 1597–1607 (PMLR, 2020).

[24] Falcon, W. & Cho, K. A framework for contrastive self-supervised learning and designing a new approach. *arXiv preprint arXiv:2009.00104* (2020).

[25] He, K., Fan, H., Wu, Y., Xie, S. & Girshick, R. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9729–9738 (2020).

[26] Chen, X., Fan, H., Girshick, R. B. & He, K. Improved baselines with momentum contrastive learning. *CoRR* **abs/2003.04297** (2020). URL `https://arxiv.org/abs/2003.04297`. `2003.04297`.

[27] Chen, T., Kornblith, S., Swersky, K., Norouzi, M. & Hinton, G. E. Big self-supervised models are strong semi-supervised learners. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. & Lin, H. (eds.) *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual* (2020). URL `https://proceedings.neurips.cc/paper/2020/hash/fcbc95ccdd551da181207c0c1400c655-Abstract.html`.

[28] Caron, M. *et al.* Unsupervised learning of visual features by contrasting cluster assignments. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. & Lin, H. (eds.) *Advances in Neural Information Processing Systems 33: Annual Conference*

on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual (2020). URL https://proceedings.neurips.cc/paper/2020/hash/70feb62b69f16e0238f741fab228fec2-Abstract.html.

[29] Grill, J. *et al.* Bootstrap your own latent - A new approach to self-supervised learning. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. & Lin, H. (eds.) *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual* (2020). URL https://proceedings.neurips.cc/paper/2020/hash/f3ada80d5c4ee70142b17b8192b2958e-Abstract.html.

[30] Chen, X. & He, K. Exploring simple siamese representation learning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, 15750–15758 (Computer Vision Foundation / IEEE, 2021). URL https://openaccess.thecvf.com/content/CVPR2021/html/Chen_Exploring_Simple_Siamese_Representation_Learning_CVPR_2021_paper.html.

[31] Li, G., Xiong, C., Thabet, A. K. & Ghanem, B. Deepergcn: All you need to train deeper gcns. *CoRR* **abs/2006.07739** (2020). URL https://arxiv.org/abs/2006.07739. 2006.07739.

[32] Vaswani, A. *et al.* Attention is all you need. In Guyon, I. *et al.* (eds.) *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, 5998–6008 (2017). URL https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

[33] Rosenblatt, F. Principles of neurodynamics. perceptrons and the theory of brain mechanisms. Tech. Rep., Cornell Aeronautical Lab Inc Buffalo NY (1961).

[34] Brown, T. *et al.* Language models are few-shot learners. *Advances in neural information processing systems* **33**, 1877–1901 (2020).

[35] Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[36] Lin, J., Zhu, J., Wang, H. & Zhang, T. Learning to branch with tree-aware branching transformers. *Knowl. Based Syst.* **252**, 109455 (2022). URL https://doi.org/10.1016/j.knosys.2022.109455.

[37] Kool, W., van Hoof, H. & Welling, M. Attention, learn to solve routing problems! In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019* (OpenReview.net, 2019). URL https://openreview.net/forum?id=ByxBFsRqYm.

[38] Wu, Y., Song, W., Cao, Z., Zhang, J. & Lim, A. Learning improvement heuristics for solving routing problems. *IEEE Trans. Neural Networks Learn. Syst.* **33**, 5057–5069 (2022). URL https://doi.org/10.1109/TNNLS.2021.3068828.

[39] Rao, R. M. *et al.* Msa transformer. In *International Conference on Machine Learning*, 8844–8856 (PMLR, 2021).

[40] Rao, R., Meier, J., Sercu, T., Ovchinnikov, S. & Rives, A. Transformer protein language models are unsupervised structure learners. *Biorxiv* (2020).

[41] Wang, S., Guo, Y., Wang, Y., Sun, H. & Huang, J. SMILES-BERT: large scale unsupervised pre-training for molecular property prediction. In Shi, X. M., Buck, M., Ma, J. & Veltri, P. (eds.) *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics, BCB 2019, Niagara Falls, NY, USA, September 7-10, 2019*, 429–436 (ACM, 2019). URL `https://doi.org/10.1145/3307339.3342186`.

[42] He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, 770–778 (IEEE Computer Society, 2016). URL `https://doi.org/10.1109/CVPR.2016.90`.

[43] Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O. & Dahl, G. E. Neural message passing for quantum chemistry. In *International conference on machine learning*, 1263–1272 (PMLR, 2017).

[44] Ba, L. J., Kiros, J. R. & Hinton, G. E. Layer normalization. *CoRR* **abs/1607.06450** (2016). URL `http://arxiv.org/abs/1607.06450. 1607.06450`.

[45] Agarap, A. F. Deep learning using rectified linear units (relu). *CoRR* **abs/1803.08375** (2018). URL `http://arxiv.org/abs/1803.08375. 1803.08375`.

[46] Titeca, K., Lemmens, I., Tavernier, J. & Eyckerman, S. Discovering cellular protein-protein interactions: Technological strategies and opportunities. *Mass spectrometry reviews* **38**, 79–111 (2019).

[47] Khosla, P. *et al.* Supervised contrastive learning. *Advances in Neural Information Processing Systems* **33**, 18661–18673 (2020).

[48] Oord, A. v. d., Li, Y. & Vinyals, O. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748* (2018).

[49] Han, J. & Moraga, C. The influence of the sigmoid function parameters on the speed of backpropagation learning. In *International workshop on artificial neural networks*, 195–201 (Springer, 1995).

[50] Szklarczyk, D. *et al.* String v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic acids research* **47**, D607–D613 (2019).

[51] Rogers, D. & Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **50**, 742–754 (2010). URL `https://doi.org/10.1021/ci100050t`.

[52] Fabian, B. *et al.* Molecular representation learning with language models and domain-relevant auxiliary tasks. *CoRR* **abs/2011.13230** (2020). URL `https://arxiv.org/abs/2011.13230. 2011.13230`.

[53] Landrum, G. Rdkit: Open-source cheminformatics software (2016). URL `https://github.com/rdkit/rdkit/releases/tag/Release_2016_09_4`.

[54] Kipf, T. N. & Welling, M. Semi-supervised classification with graph convolutional networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings* (OpenReview.net, 2017). URL `https://openreview.net/forum?id=SJU4ayYgl`.

[55] Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. In Bengio, Y. & LeCun, Y. (eds.) *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings* (2015). URL `http://arxiv.org/abs/1412.6980`.