Pisces: A multi-modal data augmentation approach for drug combination synergy prediction

Hanwen Xu[1,*], Jiacheng Lin[2,*], Addie Woicik[1], Zixuan Liu[1], Jianzhu Ma[3], Sheng Zhang[4], Hoifung Poon[4], Liewei Wang[5], Sheng Wang[#]

[1]School of Computer Science and Engineering, University of Washington, Seattle, WA

[2]Department of computer Science, University of Illinois Urbana-Champaign, Champaign, IL

[3]Department of Electronic Engineering, Tsinghua University, Beijing, China

[4]Microsoft Research, Redmond, WA

[5]Mayo Clinic, Rochester, MN

*Equal contribution

[#]Email: swang@cs.washington.edu

Abstract

Drug combination therapy is promising for cancer treatment through simultaneously reducing resistance and improving efficacy. Machine learning approaches to drug combination response prediction can prioritize experiments and discover new combinations, but require lots of training data in order to fit the nonlinearity of synergistic effect. Here, we propose Pisces, a novel machine learning approach for drug combination synergy prediction. The key idea of Pisces is to augment the sparse drug combination dataset by creating multiple views for each drug combination based on its different modalities. We combined eight different modalities of a single drug to create 64 augmented views for a pair of drugs, effectively expanding the size of the original data 64 times. Pisces obtained state-of-the-art results on cell-line-based drug synergy prediction, xenograft-based drug synergy prediction, and drug-drug interaction prediction. By interpreting Pisces's predictions using a genetic interaction network, we further identified a breast cancer drug-sensitive pathway from BRCA cell lines in GDSC. We validated this pathway on an independent TCGA-BRCA tumor dataset and found that patients with this pathway activated had substantially longer survival time. Collectively, Pisces effectively predicts drug synergy and drug-drug interactions through augmenting the original dataset 64 times, and can be broadly applied to various biological applications that involve a pair of drugs.

Introduction

Drug combination therapy, which exploits two synergistic drugs whose combined effect is greater than the sum of each drug's individual activity, is a promising treatment for cancer.[1–9] A few successful drug combinations,[6,9–13] such as lenvatinib and gefitinib for hepatocellular carcinoma (HCC),[10] have delivered encouraging results in clinics. As a result, large-scale cancer pharmacology projects, including the Genomics of Drug Sensitivity in Cancer (GDSC) projects,[14] Beat Acute Myeloid Leukemia project,[15] national cancer institute ALMANAC project,[16] and AstraZeneca's DREAM challenge,[17] have generated a large volume of genomics and pharmacological data for studying drug synergy, offering an unprecedented opportunity to discover new drug combinations and achieve precision medicine.[18] Since it is expensive to experimentally profile all combinations, there is a pressing need to develop machine learning approaches for predicting drug combination response.[19–24]

Developing machine learning approaches for predicting drug combination response is inherently challenging. On the one hand, we need an expressive model, which requires lots of training data, to capture the nonlinearity of synergistic effect.[19–23] On the other hand, we can only obtain limited training data due to the cubic growth of the drug combination space (i.e., two drugs and a cell line), which leads to the overfitting of an expressive model. Recently, data augmentation has transformed computer vision and natural language processing,[25–30] and became an essential component in generative artificial intelligence.[31–36] Data augmentation alleviates the overfitting of expressive machine learning models by creating multiple views of the same data instance and encouraging them to be similar to each other.[25]

Here we proposed Pisces, a novel machine learning model for drug combination synergy prediction. The key idea of Pisces is to augment the sparse drug combination dataset by creating multiple views for each drug combination based on different modalities. Our intuition is that different drug modalities are complementary to each other, and jointly considering them can enhance the prediction performance. Specifically, we consider eight drug modalities (**Fig. 1a**), including Simplified Molecular Input Line Entry System (SMILES),[37] molecular graphs,[37] three-dimensional molecular structures,[38] drug targets,[39] textual descriptions,[40,41] side effects,[42,43] drug response,[44] and drug ontology.[45,46] We combined these eight modalities of a single drug to create 64 augmented modality pairs for a pair of drugs, effectively expanding the number of training data instances of the original data 64 times (**Fig. 1b**). Moreover, another advantage of our framework is that we don't require to see every modality for every drug. Traditional approaches that can consider different modalities need to either collect or impute every modality for every drug. In contrast, we circumvent the missing modality challenge by treating a modality pair instead of a drug pair as a data instance, thus allowing us to train the model just using available modality pairs.

We systematically evaluated Pisces on three drug combination tasks: cell-line-based drug synergy prediction, xenograft-based drug synergy prediction, and drug-drug interaction prediction. On GDSC-combo,[47] which is a recently-released cell-line-based drug synergy dataset, Pisces achieved at least 21.4%, 23.8%, 10.2% improvements on F_1 score on three different data split settings compared to five comparison approaches.[19–23] We further demonstrated how Pisces obtained 0.8525 AUROC on three-drug combination synergy prediction when only trained on two-drug data. By interpreting Pisces's prediction using a genetic interaction network,[48] we identified a breast cancer (BRCA) drug-sensitive pathway from BRCA cell lines in GDSC-combo. We validated this pathway on an independent BRCA tumor dataset[49] and found that patients with this pathway have substantially longer survival time (log rank p-value < 0.02). Our second validation is a large-scale xenograft dataset.[50] We found that Pisces again outperforms other approaches on predicting minimum tumor volume changes. Furthermore, we showed that Pisces can accurately predict the tumor volume at an unmeasured time point with a Spearman of 0.47, while existing approaches cannot be applied to such temporal analyses. We visualized that the embeddings learnt by Pisces are ordered in chronological order and matched well with tumor volume, supporting clinical decision making and drug resistance detection. Finally, we applied Pisces to drug-drug interaction prediction, where the goal is to predict whether two drugs will have a certain interaction or adverse drug reaction. On DrugBank[39] and TwoSIDES,[51] Pisces outperformed existing approaches when one or both drugs in each test combination were never-before-seen. By examining drug-drug interactions that significantly occurred between drug classes, we constructed a novel DDI network that can be used to infer interactions for new drugs. Collectively, Pisces effectively models drug pairs through augmenting each drug pair to 64 modality pairs, and can be broadly applied to various biological applications that involve a pair of drugs.

Results

Overview of Pisces

Pisces takes a triplet of two drugs and a cell line as input and outputs the predicted synergy for this triplet (**Fig. 1b**). Pisces can handle drugs with different available modalities without imputing missing modalities. Pisces is divided into three components: projector, augmentor and aggregator. The projector will enable different modalities, which have unmatched features, to be comparable. Specifically, the projector learns eight neural networks to embed eight different modalities into a shared embedding space (**Supplementary Figure 1**). Each point in the shared embedding space is one augmented view of a particular drug based on a particular modality. For example, drug A with 5 available modalities will be mapped to five different points, each representing one of its modalities. As a result, the projector allows us to expand the original

dataset at most eight times. The augmentor combines the augmented views of single drugs into augmented views of drug combinations. In particular, drug A with 5 modalities and drug B with 4 modalities will be combined into 20 augmented views through a pairwise combination. These 20 augmented views will be combined with cell line features to become 20 new data points. Therefore, the augmentor allows us to expand the original dataset at most 64 times.

The aggregator aggregates the predictions from different augmented views into a final prediction (**Fig. 1c**). Pisces leverages the idea of noisy label learning[52] by only considering top *k* predictions among predictions from all augmented views for a triplet. Only these top predictions will be used by a ResNet[53] architecture to calculate the final prediction, allowing Pisces to exclude noise from low-quality augmented views. Pisces can be broadly applied to cell line, xenograft and patient data, and other biomedical applications involving a pair of drugs.

We first investigated the agreement among using different drug modalities for drug synergy prediction. We trained eight drug synergy prediction models (**Fig. 2a**), each only using one of the modalities. On one hand, we observed that two modalities often have a Spearman correlation greater than 0.6 (**Fig. 2b, Supplementary Figure 2-4**), indicating a good agreement among them. On the other hand, none of these Spearman correlations is greater than 0.8, even for the one between SMILES and graphs (**Fig. 2c**), which can be converted to each other programmatically.[54] This demonstrates the complementary information among these modalities, both from different information captured by each modality and from the different neural network architectures used to convert them into machine-readable embeddings. For example, SMILES is embedded using a Transformer, while the molecular graph is embedded using a graph convolutional network (**Supplementary Figure 1**). The agreement and complement of these modalities raises our confidence about the merit of Pisces and motivates us to evaluate it on large-scale drug synergy datasets.

Pisces improve cell-line-based drug synergy prediction

We first sought to evaluate Pisces on cell-line-based drug synergy prediction. We exploited a recently published drug combination dataset from GDSC,[47] and examined three data split settings: vanilla cross validation, split by drug combination, and split by cell line. On vanilla cross validation, Pisces substantially outperformed five existing approaches on all four metrics (**Fig. 3a-d**). This prominent performance of Pisces indicates the possibility of learning high-quality drug pair embeddings by integrating different modalities. We next examined the split by drug combination setting, where all test drug combinations have never been seen in the training set. Since experimental measurements can only cover a small proportion of all possible drug combinations, this setting can simulate whether our method can predict the responses for new drug combinations. We found that the performance of all methods dropped, indicating that

this is a more challenging setting compared to the vanilla cross validation. Nevertheless, Pisces still surpassed the next-best-performing approach by 24% and 26% in terms of the F_1 and KAPPA scores, demonstrating that the embeddings learnt by Pisces on a small collection of training drug pairs can be extended to never-before-seen drug combinations. Finally, we evaluated the split by cell line setting, where all test cell lines have never been seen in the training set. This setting examines whether we can recommend drug combinations to a new patient who has not been tested on any combinations. We found that this setting is the most challenging one among all the three settings according to the overall prediction performance. However, Pisces still obtained 10% improvements in terms of F_1 score against the next-best-performing approach. Collectively, the prominent performance of Pisces on three different settings demonstrates its broad applicability to new drug combinations and new cell lines, confirming the effectiveness of augmenting existing dataset using different drug modalities.

Existing drug synergy prediction approaches are mainly developed for combinations of two drugs. In practice, combinations of three drugs have been shown to alleviate side effects [55] and drug resistance.[56,57] Pisces can be extended to predict the synergy of three drugs by only training on the combinations of two drugs (see Methods). On the three-drug combination data in GDSC, Pisces obtained an AUROC of 0.82 and an AUPRC of 0.43 on ranking cell lines for a combination (**Fig. 3e**), and 0.88 AUROC and 0.71 AUPRC on ranking combinations for a cell line (**Supplementary Figure 5-6**). Afatinib and Trametinib have been reported to be an effective combination by co-inhibiting of *ErbB* family and *MEK/ERK* pathways.[58] Pisces predicted that Linsitinib is the most effective third drug when used together with Afatinib and Trametinib on 33 colon adenocarcinoma cell lines (average synergy probability 19.2%). Previous studies demonstrated that Linsitinib could overcome acquired Afatinib resistance by inhibiting *IGF1R* phosphorylation, making the combination of Linsitinib and Afatinib an effective combination therapeutics.[59] The promising performance of using Pisces to predict the synergy of three-drug combinations indicates that Pisces can be applied to applications that involve more than two drugs.

Pisces identifies a breast cancer drug-sensitive pathway

To further illustrate the clinical relevance of Pisces, we integrated the predictions of Pisces on 51 BRCA cell lines with a genetic interaction network[48] to identify a BRCA drug-sensitive pathway (**Fig. 3f**). In particular, since Pisces can be applied to never-before-seen combinations, we first used Pisces to predict the response of 176,344 triplets of cell line and a pair of drugs that have not been experimentally measured in GDSC. Pisces predicted 5.1% of these triplets to be synergistic, demonstrating the rareness of effective drug combinations. We then created 97,479

gene pairs from these synergistic triplets using the following criteria: One gene is over-expressed in the cell line and the other gene is the target of the drug combination. We next filtered these gene pairs using synthetic dosage lethality (SDL) interactions from a genetic interaction network.[48] The resulting gene pairs are connected as a pathway with 19 genes and 32 edges, including 8 drug targets and 12 SDL interactions. Many genes in this pathway are from *PI3K* (phosphoinositide 3-kinase) and *EGFR*/*HE*R families. *PI3K* plays a crucial role in *ER*+ (estrogen receptor-positive) tumors that have developed resistance to hormonal therapy.[60,61] At least 14 drugs in GDSC can inhibit one of these 8 targets (**Supplementary Figure 7**). We hypothesize that this is a breast cancer drug-sensitive pathway since cancer drugs applied to this pathway will inhibit their targets and exploit SDL to effectively kill the cancer cells.

We validated this hypothesis on BRCA tumors in TCGA.[49] We clustered 977 BRCA tumors into four clusters based on whether this pathway was activated or not and Estrogen receptor (ER) status. We found that *ER*+ tumors that activated this pathway had substantially longer overall survival time (log rank p-value < 0.011) and disease-free survival time (log rank p-value < 0.017) (**Fig. 3g-h**). To further confirm that the survival prediction ability is not simply from clustering patients to *ER*+ and *ER*-, we clustered 977 BRCA tumors into two groups based on the activation of this pathway. We found that the activation of this pathway still enabled a significantly longer overall survival time (log rank p-value < 0.003) and disease-free survival time (log rank p-value < 0.008) (**Supplementary Figure 8**), reflecting the possibility to use this pathway as clinical biomarkers for breast cancer prognostic.

Pisces improves xenograft-based drug synergy prediction

After observing the promising performance of Pisces on predicting drug synergy for cell lines, we next validated it on patient-derived tumor xenografts. We exploited a xenograft dataset that spans 1,238 drug combinations across 277 xenograft models. We first used Pisces to predict the BestResponse, which is defined as the minimum tumor volume change after 10 days, for a pair of drugs on a xenograft model in the cross validation setting. Pisces outperformed other drug synergy prediction approaches using both Spearman correlation and Pearson correlation (**Fig. 4a, Supplementary Figure 9-10**), demonstrating the promising performance of using Pisces to predict tumor growth.

Moreover, predicting tumor growth at a future time point is important for cancer prognosis and survival prediction. Therefore, we extended Pisces to perform such temporal prediction for tumor growth. In particular, we incorporated a time embedding into our framework so that Pisces can now take the time point as an additional input. Pisces can then predict the tumor growth of a drug combination on a xenograft model at any unobserved time point. We found that Pisces obtained a 0.47 Pearson correlation on predicting tumor growth at holdout time

points (p-value < $1 * 10^{-100}$) (**Fig. 4b**). Moreover, we examined an extrapolation setting, where only the last time point of each xenograft model is held out, and observed that Pisces also accurately predicted the tumor growth (Pearson r = 0.47, p-value < $1.6 * 10^{-25}$) (**Fig. 4c**), demonstrating its applicability to perform temporal prediction on tumor growth.

After observing Pisces's prominent performance on temporal tumor growth prediction, we are motivated to examine the tumor growing trajectories using embeddings generated by Pisces. To this end, we visualize the embedding of test triplets (i.e., a drug combination on a xenograft model) in the two-dimensional space. We first found that these triplets are clustered by xenograft models and branches in each cluster represent different drug combinations (**Fig. 4d, Supplementary Figure 11**). By coloring each triplet using time points, we find that each branch grows from an earlier time point to a later time point, reassuring Pisces's ability to capture the temporal tumor growth. By coloring each triplet using tumor volume change and adding contours to characterize time points, this embedding space can illustrate both tumor growth and time point, supporting the identification of drug resistance (**Fig. 4e, Supplementary Figure 12**). For example, INC424 and binimetinib showed diverse drug responses among different xenograft models, where the tumor grows most aggressively on X-3846 and X-2564, suggesting drug resistance. Collectively, the promising performance of Pisces on xenograft indicates its applicability to the *in vivo* setting, as well as unique tumor temporal analysis that can hardly be achieved by existing drug combination prediction methods.

Pisces accurately predicts drug-drug interactions

The promising results of Pisces on drug synergy prediction further motivates us to apply it to other tasks that involve a pair of drugs. To this end, we investigated whether Pisces can be used to predict drug-drug interactions (DDI), where the goal is to classify a pair of drugs into hundreds of predefined interaction types. We evaluated our method on two large-scale DDI datasets: DrugBank, which contains 191,402 drug pairs over 86 interactions, and TwoSIDES, which contains 63,472 drug pairs over 963 interactions. On DrugBank, we found that Pisces substantially outperformed six existing approaches under three different settings, including vanilla cross validation, one new drug in each test pair and two new drugs in each test pair (**Fig. 5a-c**). The improvement against competing methods is larger under the more challenging settings of one new drug in each test pair and two new drugs in each test pair, indicating the strong generalizability of Pisces through integrating different modalities. On a larger dataset TwoSIDES, Pisces outperformed five comparison approaches and achieved a comparable performance with R^2-DDI (**Fig. 5d-f**), again highlighting its broad applicability to drug-drug interaction prediction.

Finally, we examined DDIs that are not collected in DrugBank, but are predicted with high confidence by Pisces. We first group drugs based on their classes in the drug ontology. We then used Fisher's exact test to find interaction types that significantly occurred between two drug classes (**Supplementary Table 1**). We form these significant interactions as a DDI network, where nodes are drugs and edges are colored by the most likely interaction between two drugs (**Fig. 5g**). This novel DDI network offers a global view of DDI and can extend our analysis to never-before-seen drugs according to its class in the drug ontology.

Discussion

We have proposed Pisces, a novel machine learning approach for drug synergy prediction. Pisces can augment existing sparse drug combination dataset at most 64 times using 8 different modalities of each drug. We have shown the prominent performance of Pisces on cell-line-based drug synergy prediction, xenograft-based drug synergy prediction, and drug-drug interaction prediction tasks. We also demonstrated the interpretability and clinical relevance of Pisces by identifying a breast cancer drug-sensitive pathway, and verified using survival data from TCGA. Collectively, Pisces will facilitate the future discovery of drug combinations and drug-drug interactions, and can be applied to a wide range of biological applications that involve multiple drugs.

`

Pisces has at least two limitations that we would like to address in future works. First, the large number of modality pairs increase the predictive performance, but also makes it hard to understand what modalities are most important for a certain prediction. We plan to exploit local interpretation methods to interpret the prediction results.[62,63] Second, while we have considered different modalities as features, we didn't consider different types of labels jointly. Labels for drug pairs include synergy effects over cell lines and xenografts, side effects, and interactions. Since our framework can handle missing modality, it can also be extended to handle missing labels. We plan to incorporate multi-task learning[64] into our framework so that Pisces can be simultaneously optimized for different biological tasks that involve a pair of drugs.

There are a few existing cross-modal drug representation learning approaches,[38,65–68] such as MolCLR[66] and the geometry enhanced GEM[65]. There are at least two key differences contrasting Pisces with these works. First, these existing cross-modal approaches mainly study two drug modalities, such as SMILES and molecular graphs. In contrast, Pisces integrates eight different modalities, including structure features, mode of action features, phenotypes, and pharmacodynamic features, allowing Pisces to learn much more comprehensive embeddings for drugs. Second, these approaches only model a single drug, while Pisces is designed for a pair of drugs. Our pairwise combination of single drug augmentation is a novel approach to augment a pair of drugs, effectively augmenting existing sparse datasets up to 64 times. Compared to

existing drug combination prediction approaches,[19–23] Pisces also has at least three key differences. First, Pisces exploits novel machine learning techniques, including data augmentation, contrastive learning and noisy label learning, which alleviate the sparsity and noise in drug combination dataset. Second, existing drug synergy approaches were designed specifically for synergy prediction, while Pisces is a general framework for tasks that study a pair of drugs. Third, Pisces can be applied to combinations with more than two drugs, while existing approaches can only be applied to two-drug combinations, demonstrating the broad applicability of Pisces.

Figure legend

Fig. 1 Pisces model overview. a, Pisces considers eight different modalities for a single drug, including Simplified Molecular Input Line Entry System (SMILES), molecular graphs, three-dimensional molecular structures, drug targets, textual descriptions, side effects, drug response and drug ontology. Pisces can handle missing modalities by only training on available modality pairs, circumventing the need to impute missing modalities. **b,** Pisces takes a pair of drugs and a cell line as input. Each drug is represented by its available modalities, which could vary between two drugs. Pisces has three components: projector, augmentor and aggregator. The projector trains eight different encoders for eight modalities to project all modalities of all drugs into a shared embedding space. Each point in this space is one modality of one drug, allowing us to expand the original single drug data at most eight times. The augmentor then creates embeddings of a drug pair through the pairwise combination of each drug's embedding, allowing us to expand the original drug pair data approximately 64 times. **c,** Pisces predicts one score for each of the 64 views, producing 64 predicted scores for a drug pair on a cell line. The aggregator first finds the top *k* largest predictions and zero-out the remaining ones, allowing us to exclude the noise from low-quality augmentations. It then uses a ResNet structure to integrate these *k* largest predictions as the final prediction.

Fig. 2 The agreement among predictions of different modalities. a, Heatmap showing the agreement between predicted synergy probabilities on GDSC cell lines using the modality in the x-axis and y-axis. Agreement is measured using the Spearman correlation. The same single modality is used for both drugs. Predictions are evaluated using the AUPRC. Ratio of missingness means the percent of drugs that do not have this modality. **b,** Scatter plot comparing the prediction scores of using drug-ontology-based modality and text-based modality. **c,** Scatter plot comparing the prediction scores of using graph-based modality and SMILES-based modality.

Fig. 3 Drug synergy prediction on cell lines. a-d, Bar plots comparing the drug synergy prediction on GDSC under three data split settings (x-axis) using BACC (**a**), AUPRC (**b**), F1 (**c**), Cohen's Kappa (**d**). Split by combination means all test combinations have never been seen in the training data. Split by cell line means all test cell lines have never been seen in the training data. The * indicates that Pisces outperforms the next-best-performing model in the metric, with significance levels of t-test p-value < 5e-2 for *, t-test p-value < 1e-2 for **, and t-test p-value < 1e-3 for ***. All t-tests are one-sided. **e**, Bar plot showing Pisces's prediction performance on three-drug combination prediction when trained only on two-drug and single-drug data. AUROC and AUPRC are only calculated for three-drug combinations with at least one measured cell line. Afatinib and Trametinib are two of the three drugs and the third drug is shown in x-axis. Each colored point is Pisces's predicted synergy (y-axis) of this three-drug combination on a cell line. **f,** A BRCA drug-sensitive pathway identified by interpreting Pisces's drug synergy predictions using a genetic interaction network. **g,h,** Survival plots showing the significantly different overall survival (**g**) and disease-free survival (**h**) among four groups of TCGA-BRCA patients. These four groups are classified using the gene expression in the BRCA drug-sensitive pathway and *ER* status.

Fig. 4 Drug synergy prediction on xenografts. a, Box plot comparing the drug synergy prediction on xenografts using Spearman correlation. The * indicates that Pisces outperforms the next-best-performing model in the metric, with significance levels of t-test p-value < 5e-2 for *, t-test p-value < 1e-2 for **, and t-test p-value < 1e-3 for ***. All t-tests are one-sided. **b**, Scatter plots comparing the predicted tumor change and the observed tumor change at holdout time points. Bar plots show the frequency of points in the scatter plot. **c**, Scatter plots comparing the predicted tumor change and the observed tumor change when only the last time point is held out. Bar plots show the frequency of points in the scatter plot. **d**, Each point in the UMAP plot is the embedding of a particular pair of drugs on a particular cell line at a particular time point. Points are colored and marked according to the drug pair. The contours connect tumors that have the same time point. **e**, Each point in the UMAP plot is the embedding of a particular pair of drugs on a particular cell line at a particular time point. Points are colored and marked according to the drug pair. The contours connect tumors that have the same time point. The arrows are from early time points to later time points.

Fig. 5 Drug-drug interaction predictions on DrugBank and TwoSIDES. **a-c**, Bar plots comparing the prediction performance of drug-drug interactions on DrugBank under three cross validation settings using Accuracy (**a**), AUROC (**b**), and AUPRC (**c**). One new drug in each test pair means that one and only one of the two drugs in a test drug pair has never been seen in the training data. Two new drugs in each test pair means that both of the two drugs in a test drug pair have never been seen in the training data. **d-f**, Bar plots comparing the prediction

performance of drug-drug interactions on TwoSIDES using Accuracy (**d**), AUROC (**e**), and AUPRC (**f**). **g**, Visualization of significantly occurred interaction types between two drug classes determined by Fisher's exact test (p-value < 0.05). Nodes are colored by drug classes in the drug ontology. Edges are colored by interaction types.

Methods

Problem setting

Let $\mathcal{D}$ be the set of drugs and C be the additional features. The input drug combination dataset was defined as multiple triplets $\left\{(d_a, d_b, c)_i\right\}_{i=1}^{N}$, where $d_a \in \mathcal{D}$ and $d_b \in \mathcal{D}$ is a pair of drugs. We used $c \in C$ to represent the additional features of this drug pair. C can be the cell lines and the xenograft models in drug synergy prediction tasks or the indication of interaction types in DDI prediction. Both tasks can be modeled as the binary prediction task:

$$\hat{y}_i = f_\theta((d_a, d_b, c)_i),$$

where $f_\theta: \mathcal{D} \times \mathcal{D} \times C$ is a learned mapping function with θ as parameters. The output $\hat{y}_i$ denotes the predicted property of the input triplet $(d_a, d_b, c)_i$ being positive, which can be the probability of being synergistic in cell-line-based drug synergy prediction, the tumor volume change in xenograft-based drug synergy prediction and the probability of the drug pair exhibiting a specific interaction type in drug-drug interaction prediction.

We considered 8 different feature modalities for each drug: the SMILES, molecular graphs, drug target genes, drug three-dimensional structures, textual descriptions, drug ontology views, drug side effects and drug sensitivity, which we used

$$\mathcal{M} = \left\{m_{SMILES}, m_{g2D}, m_{target}, m_{g3D}, m_{text}, m_{dron}, m_{sides}, m_{sens}\right\}$$

to denote them respectively. We denote the input sample with a pair of modalities as $x[m_a, m_b] = (d_a[m_a], d_a[m_a], c)$, where $m_a, m_b \in \mathcal{M}$. Considering $|\mathcal{M}|$ different modalities, each drug combination produces a set of $|\mathcal{M}| \times |\mathcal{M}|$ modality pairs $\left\{x[m_a, m_b] | \forall m_a, m_b \in \mathcal{M}\right\}$. By treating each modality pair as an augmented drug combination view, we now expand the original data by $|\mathcal{M}| \times |\mathcal{M}|$ times.

Projector with modality-specific drug encoders

The eight modalities have different data structures. In order to integrate them together, we designed a projector module that transformed these modalities into a shared space, e.g. dense embeddings with the same length. The diverse data structures of different modalities motivate

us to learn a set of embedding functions $\left\{f_{\theta[m]} | m \in \mathcal{M}\right\}$ separately with parameters $\theta[m]$ for $|\mathcal{M}|$ modalities.

1. **SMILES modality:** Each drug SMILES string can be represented as $m_{SMILES} = \langle s_1, ..., s_l \rangle$ where s_i denotes the i-th token of the l length string. Since Transformer has emerged as an efficient model to embed a wide range of sequence features, such as natural languages,[31,32] protein sequences[69,70] and molecular features,[71] we adopted the encoder part of Transformer as $f_{\theta[m_{SMILES}]}$. We appended the special token $[CLS]$ to each SMILES string: $\widehat{m}_{SMILES} = \langle s_1, ..., s_l, [CLS] \rangle$ and get the contextualized embedding at the position corresponding to the special token as the embeddings $\widetilde{\mathbf{z}}_{SMILES}$ for this SMILES string. We finally used one transformation MLP layer to get the final embedding: $\mathbf{z}_{SMILES} = MLP\left(\widetilde{\mathbf{z}}_{SMILES}\right)$, where $\mathbf{z}_{SMILES} \in \mathbb{R}^{d_{hid}}$.
2. **2D graphs:** Each molecule can be represented as a 2D molecular graph $m_{g2D} = (V, E)$, where $V = \left\{v_i\right\}_{i=1}^{n}$ are n atoms in this molecule and E are the set of bonds between atoms. We used DeeperGCN[72] as the graph encoder $f_{\theta[m_{g2D}]}$ to calculate embeddings for each atom: $\left\{\widehat{v}_i\right\}_{i=1}^{n} = f_{\theta[m_{g2D}]}(\left\{v_i\right\}_{i=1}^{n}, E)$. Finally, we used mean pooling and max pooling to get the graph embedding: $\mathbf{z}_{g2D} = MLP(Concat(MeanPool(\left\{\widehat{v}_i\right\}_{i=1}^{n}), MaxPool(\left\{\widehat{v}_i\right\}_{i=1}^{n})))$, where $\mathbf{z}_{g2D} \in \mathbb{R}^{d_{hid}}$.
3. **Drug target:** We denote the target gene set of each drug as $m_{target} = \left\{\mathfrak{g}_i\right\}_{i=1}^{n_{target}}$. To utilize the physical and functional interactions between genes, we find the $k_{\mathfrak{g}}$ nearest neighbors for each $\mathfrak{g} \in m_{target}$ in the protein-protein interaction networks (PPI), defined as $\mathcal{N}_{k_{\mathfrak{g}}}(\mathfrak{g})$ and finally compute the expanded gene set $\widehat{m}_{target} = \bigcup_{\mathfrak{g} \in m_{target}} \{\mathfrak{g}\} \cup \mathcal{N}_{k_{\mathfrak{g}}}(\mathfrak{g})$. For each gene $\mathfrak{g} \in \widehat{m}_{target}$, we used one learnable embedding as its representation: $\mathcal{E}_{\mathfrak{g}} \in \mathbb{R}^{d_{\mathfrak{g}}}$. We finally compute the drug target embedding using an MLP layer $f_{\theta[m_{target}]}: \mathbb{R}^{|\widehat{m}_{target}| \times d_{\mathfrak{g}}} \to \mathbb{R}^{d_{hid}} = MLP_{\theta[m_{target}]}$ as:

$$\mathbf{z}_{target} = f_{\theta[m_{target}]}(Concat(\{\mathcal{E}_{\mathfrak{g}} | \mathfrak{g} \in \widehat{m}_{target}\})).$$

4. **Drug 3D geometric views:** We also sought to learn the drug representations from 3D geometric views. By projecting the drug 2D graph modalities towards the 3D geometric

views, we are able to enhance each drug with 3D information. Let $\phi_{2D}(V,E)$ be the 2D graph encoder and $\phi_{3D}(V,O)$ be the 3D geometric encoder, where O denotes the 3D coordinates of each atom. The 2D graph encoder is aligned with the 3D encoder using a contrastive learning loss:

$$\mathcal{L}_{infoNCE} = -\frac{1}{2}\mathbb{E}_{p(V,E,O)}\left[\frac{\exp\left(s_\theta\left(\phi_{2D}(V,E),\phi_{3D}(V,O)\right)\right)}{\exp\left(s_\theta\left(\phi_{2D}(V,E),\phi_{3D}(V,O)\right)\right)+\sum_j \exp\left(s_\theta\left(\phi_{2D}(V^j,E^j),\phi_{3D}(V,O)\right)\right)}\right]$$
$$-\frac{1}{2}\mathbb{E}_{p(V,E,O)}\left[\frac{\exp\left(s_\theta\left(\phi_{2D}(V,E),\phi_{3D}(V,O)\right)\right)}{\exp\left(s_\theta\left(\phi_{2D}(V,E),\phi_{3D}(V,O)\right)\right)+\sum_j \exp\left(s_\theta\left(\phi_{2D}(V,E),\phi_{3D}(V^j,O^j)\right)\right)}\right],$$

where s_θ is the scoring function for computing similarities between 2D and 3D views and j corresponds to the randomly sampled complementary views. Using the fixed 2D graph encoder aligned with 3D view, we could get the drug embedding augmented by both views. We use our encoder $f_{\theta[m_{g3D}]}$ to calculate embeddings:

$$\mathbf{z}_{g3D} = f_{\theta[m_{g3D}]}\left(\phi_{2D}(V,E)\right) = MLP_{\theta[m_{g3D}]}\left(\phi_{2D}(V,E)\right),$$

where $\mathbf{z}_{g3D} \in \mathbb{R}^{d_{hid}}$ represents the output embedding with 3D geometric views. In our implementation, we utilized the pretrained GraphMVP[38] as the $\phi_{2D}(V,E)$ encoder.

5. **Drug textual descriptions:** For each drug, we used their textual descriptions from the ChEBI dataset. For drugs that are not collected in ChEBI,[41] we used MolT5,[40] a translation framework from molecules to natural languages, to generate their textual descriptions. We denote the each textual description as a token sequence $m_{text} = \langle x_1^{(text)},\ldots,x_l^{(text)}\rangle$. To learn a dense representations from textual descriptions, we used a parameter-fixed biomedical domain specific pretrained language model PubMedBERT[73] together with an MLP mapping function as the encoder to embed textual descriptions for each drug $\mathbf{z}_{text} = f_{\theta[m_{text}]}\left(m_{text}\right) = MLP_{\theta[m_{text}]}\left(\phi_{PubMedBERT}\left(\langle x_1^{(text)},\ldots,x_l^{(text)}\rangle\right)\right)$, where $\mathbf{z}_{text} \in \mathbb{R}^{d_{hid}}$ represents the output text modality embeddings.
6. **Drug ontology structures:** Drugs can be queried as a node in the drug ontology, which is a structured controlled vocabulary to classify drugs. We followed Mashup[74] to get node embeddings. We constructed the transition matrix $\boldsymbol{\mathcal{T}}$ from the ontology structure using the Random Walk with Restart (RWR):$\boldsymbol{t}_j^{(t+1)} = (1-\alpha)\boldsymbol{\mathcal{T}}^T\boldsymbol{t}_j^{(t)} + \alpha\boldsymbol{\varepsilon}_j$ where $\boldsymbol{t}_j^{(t)}$ denotes the probability vector of transiting from node j to all nodes. α is the restart probability and $\boldsymbol{\varepsilon}_j$ is the basis vector. Then we got node embeddings by singular value

decomposition: $\tilde{\mathbf{z}}_{dron} = \sqrt{\sigma_i}\vec{v_i}, \mathbf{U\Sigma V} = SVD\left(\left[\boldsymbol{t}_1^{(\infty)};...;\boldsymbol{t}_{n^{node}}^{(\infty)}\right]\right)$. Let's use $\mathbf{z}_{dron} \in \mathbb{R}^{d_{hid}}$ to denote the ontology embeddings. Finally, we got $\mathbf{z}_{dron}$ using our drug ontology encoder: $\mathbf{z}_{dron} = f_{\theta[m_{dron}]}(m_{dron}) = MLP_{\theta[m_{dron}]}(\tilde{\mathbf{z}}_{dron})$.

7. **Drug side effects:** We collected the drug side effects from the SIDER 4.1 dataset[42,43] which includes 1,428 drugs and 27 side effect types. For each drug, we used the one hot embedding $\tilde{\mathbf{z}}_{sides} \in \mathbb{R}^{27}$ to represent their side effects, where $\tilde{\mathbf{z}}_{sides}[i] = 1$ means this drug has the i-th side effect. We transformed the side effects into dense embeddings $\mathbf{z}_{sides} \in \mathbb{R}^{d_{hid}}$ using: $\mathbf{z}_{sides} = MLP_{\theta[m_{sides}]}(\tilde{\mathbf{z}}_{sides})$.
8. **Drug sensitivity:** The drug sensitivity data was the drug response measured across NCI 60 cell lines.[44] Each drug response can be represented as $\tilde{\mathbf{z}}_{sens} \in \mathbb{R}^{60}$ where each element represents the response of a single cell line. We transformed $\tilde{\mathbf{z}}_{sens}$ to get the drug sensitivity representations $\mathbf{z}_{sens} \in \mathbb{R}^{d_{hid}}$ using one MLP layer. The transformation function can be denoted as $f_{\theta[m_{sens}]}: \mathbb{R}^{60} \to \mathbb{R}^{d_{hid}} = MLP_{\theta[m_{sens}]}$.

This projector module enables us to embed each drug with $|\mathcal{M}|$ modalities into $|\mathcal{M}|$ dense embeddings. Next, we sought to learn representations for a pair of drugs.

Augmentor for creating augmented drug combinations

The $|\mathcal{M}|$ modalities enable us to obtain $|\mathcal{M}| \times |\mathcal{M}|$ views for a pair of drugs, resulting in augmented drug combinations. For the drug combination triplet $(d_a, d_b, c)_i$, if we use the embedding $\mathbf{z}_m$ from a single modality $m \in \mathcal{M}$ to represent one drug, the embedding of additional feature $\boldsymbol{c} \in \mathbb{R}^{d_{hid}}$ to represent the additional features c, we were able to combine them into $|\mathcal{M}| \times |\mathcal{M}|$ triplets: $\left\{\left(\mathbf{z}_{m_a}, \mathbf{z}_{m_b}, \boldsymbol{c}\right) | \forall\, m_a, m_b \in \mathcal{M}\right\}$. The augmentor module applied one transformation function to generate augmented drug combination views:

$$\forall m_a, m_b \in \mathcal{M},\ \hat{\mathbf{z}}_i[m_a, m_b, \boldsymbol{c}] = \Psi_{aug}\left(\mathbf{z}_{m_a}, \mathbf{z}_{m_b}, \boldsymbol{c}\right)$$

Note that with these new data instances, our training dataset is $|\mathcal{M}| \times |\mathcal{M}|$ times larger. Passing all augmented drug combination views through a shared classifier, we were able to collect $|\mathcal{M}| \times |\mathcal{M}|$ predictions for each triplet. Let's denote the classifier as $\psi_{output}\left(\hat{\mathbf{z}}_i[m_a, m_b, \boldsymbol{c}]\right)$, and the collected predictions for triplet $(d_a, d_b, c)_i$ is $\left\{\psi_{output}\left(\hat{\mathbf{z}}_i[m_a, m_b, \boldsymbol{c}]\right) | \forall\, m_a, m_b \in \mathcal{M}\right\}$.

Aggregator module for effective aggregations of all predictions

For each drug triplet, with the projector and augmentor modules, we now collected $|\mathcal{M}| \times |\mathcal{M}|$ predictions from augmented drug combination views. Finally, we aggregated the output predictions together. Inspired by the noisy label learning, we sought to consider the top predictions as highly confident predictions and treat others as predictions from low-quality augmented views. We set k data instances with larger prediction values to be kept while other data instances to be ignored. One $NLL()$ layer was added to perform this operation. This layer basically keeps the top k neurons with higher prediction values while setting the output of other neurons to be zero. Then we applied a linear affine matrix $\boldsymbol{\omega}_{NLL}$ together with a shortcut to aggregate the top predictions. Taken together, the output of this aggregator module can be calculated using the following equation:

$$\hat{y}_i^{\bigstar} = \psi_{NLL}\left(\left\{\psi_{output}\left(\hat{\mathbf{z}}_i[m_a, m_b, \boldsymbol{c}]\right)\right\}\right) = \boldsymbol{\omega}_{NLL}{}^{T} \cdot NLL\left(\left\{\psi_{output}\left(\hat{\mathbf{z}}_i[m_a, m_b, \boldsymbol{c}]\right)\right\}\right) + \frac{\Sigma\left(\left\{\psi_{output}\left(\hat{\mathbf{z}}_i[m_a, m_b, \boldsymbol{c}]\right)\right\}\right)}{k}.$$

Training objective of Pisces

The training objective of Pisces consisted of two parts: the supervised training objective and the consistency objective. The supervised training objective aims to minimize the loss between the ground truth label and the predictions either from each augmented data instance or from the aggregated predictions. We denote the supervised training objective as:

$$\mathcal{L}_{SL} = -\sum_{i=1}^{n_{batch}}\left(y_i \log\left(\hat{y}_i^{\bigstar}\right) + \left(1 - y_i\right)\log\left(1 - \hat{y}_i^{\bigstar}\right)\right)$$
$$- \lambda_{aux} \sum_{(m_a, m_b) \in \mathcal{M}^*} \sum_{i=1}^{n_{batch}}\left(y_i \log\left(\hat{y}_i^{(m_a, m_b)}\right) + \left(1 - y_i\right)\log\left(1 - \hat{y}_i^{(m_a, m_b)}\right)\right),$$

where y_i denotes the ground truth label of the i-th sample and λ_{aux} is a hyperparameter. We also used $\hat{y}_i^{\bigstar}$ to denote the aggregated output and used $\hat{y}_i^{(m_a, m_b)}$ to denote the output when drug a and drug b use m_a and m_b modalities respectively. In the second term, we used a sampling strategy to reduce the computational costs. We used $\mathcal{M}^* = \left\{(m_a^{(1)}, m_b^{(1)}), \ldots, (m_a^{(r)}, m_b^{(r)})\right\}$ to denote a set of modality pairs uniformly sampled from all combination set $\mathcal{M} \times \mathcal{M}$. In our experiments, we set $r = 2$.

Since the augmented data instances from the same drug triplet are not statistically independent, we introduced a consistency objective to encourage these instances to correlate with each other. We considered these correlations in terms of both the embedding space and the output space. A contrastive learning loss was introduced to improve the consistency among augmented views in

terms of the embedding space. We adopted the InfoNCE loss,[75] which has been broadly applied in diverse machine learning areas,[76–79] to improve correlations among augmented views within one batch:

$$\mathcal{L}_{CSS}^{(InfoNCE)} = -\sum_{i=1}^{n_{batch}} \log\left(\frac{\exp\left(\hat{\mathbf{z}}_i[m_a^{(1)}, m_b^{(1)}, \boldsymbol{c}]^T \cdot \hat{\mathbf{z}}_i[m_a^{(2)}, m_b^{(2)}, \boldsymbol{c}] / \left(||\hat{\mathbf{z}}_i[m_a^{(1)}, m_b^{(1)}, \boldsymbol{c}]||_2 \cdot ||\hat{\mathbf{z}}_i[m_a^{(2)}, m_b^{(2)}, \boldsymbol{c}]||_2\right)\right)}{\sum_{j=1}^{n_{batch}} \exp\left(\hat{\mathbf{z}}_i[m_a^{(1)}, m_b^{(1)}, \boldsymbol{c}]^T \cdot \hat{\mathbf{z}}_j[m_a^{(2)}, m_b^{(2)}, \boldsymbol{c}] / \left(||\hat{\mathbf{z}}_i[m_a^{(1)}, m_b^{(1)}, \boldsymbol{c}]||_2 \cdot ||\hat{\mathbf{z}}_j[m_a^{(2)}, m_b^{(2)}, \boldsymbol{c}]||_2\right)\right)}\right).$$

In addition to the contrastive learning loss at the embedding space level, we added another loss term to improve consistency in the output. We applied the mean squared error (MSE) loss to let the aggregated prediction $\hat{y}_i^{\bigstar}$ to be correlated with prediction from each augmented data instance:

$$\mathcal{L}_{CSS}^{(MSE)} = -\sum_{i=1}^{n_{batch}} \sum_{(m_a, m_b) \in \mathcal{M}^*} (\hat{y}_i^{\bigstar} - \hat{y}_i^{(m_a, m_b)}).$$

Inspired by noisy label learning,[52,80,81] the key intuition of this consistency loss is to exploit high-quality refined prediction $\hat{y}_i^{\bigstar}$ as a soft label to correct the rest of noisy predictions using each modality pair $\hat{y}_i^{(m_a, m_b)}$. If $\hat{y}_i^{(m_a, m_b)}$, the noisy predictions do not agree with the refined soft label, we encourage the model to adjust them to be close to the refined soft labels. This consistency loss benefits the model to have predictions with high fidelity. Finally, we combined the supervised training loss and consistency loss together as the training objective:

$$\mathcal{L}^{(Pisces)} = \mathcal{L}_{SL} + \lambda_{CSS}^{(InfoNCE)} \mathcal{L}_{CSS}^{(InfoNCE)} + \lambda_{CSS}^{(MSE)} \mathcal{L}_{CSS}^{(MSE)}.$$

Collect multi-modalities of drugs

We assume that the drug combination dataset provides the SMILES modalities for all drugs. Graph modalities can be converted from the SMILES modalities using the RDKit python package.[54] We collected the drug target modalities from both the Genomics of Drug Sensitivity in Cancer (GDSC)[14] and the DrugBank dataset.[39] The GDSC dataset contains 621 drugs with target genes and the DrugBank dataset has 14,398 drugs. We aligned them with existing drugs using their names and synonyms to obtain drug targets. To collect the 3D geometric views, we used a 2D graph encoder from GraphMVP, pre-trained with 3D chemical structures using contrastive learning loss, to project our molecules into the representation space augmented with 3D information. The textual descriptions are collected with the MolT5 code base, which both curated textual descriptions for molecules in ChEBI dataset and provides the machine translation tool to generate texts for molecules outside the ChEBI dataset. The drug ontology dataset, with 8,248 drugs in total, was obtained from the National Center for Biomedical

Ontology.[82] We used the version released by October 22, 2022. The drug side effects were collected from SIDER v4.1 dataset[42,43] and we finally obtained 1,428 drugs and 27 side effect types. The drug sensitivity data, which measure 20,861 drugs across NCI 60 cell lines[44] in terms of average z score was obtained from the CellMiner website (https://discover.nci.nih.gov/cellminer/loadDownload.do).

GDSC-combo dataset processing

The GDSC-combo dataset was a recently published Genomics of Drug Sensitivity in Cancer dataset.[47] The original dataset presents the samples formulated by (Drug A, Drug B, cell line, synergy or not) tuples, and we view each triplet (drug A, drug B, cell line) as synergy if there exists one, otherwise no synergy. We consider the drug combination to be synergistic as long as it is indicated as synergistic under one circumstance. We chose samples that have a combination of 2 drugs and finally obtained 102,893 samples, including 63 drugs and 125 cell lines. We evaluated three settings: vanilla cross validation, split by drug combination and split by cell line settings. In the vanilla cross validation setting, we randomly split the dataset by 60% for training, 20% for validation and 20% for testing. In the split by drug combination cross validation setting, we split the dataset according to drug pairs. We used the splitting ratio of 60%, 20% and 20% for training, validation and testing to split the drug combinations, and distributed samples to different sets based on the drug pairs. We used the same way to split the 125 cell lines features in the split by cell line cross validation setting and collected samples into train, validation and test set.

Comparison approaches for drug synergy prediction on cell lines

We compared Pisces to five existing drug synergistic prediction approaches. **PRODeepSyn**[23] 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.[83] Then drug descriptor is a 200-dimensional real vector used representing molecules' physical or chemical properties of interest, such as lipophilicity or molecular refractivity.[84] Both the fingerprints and the descriptors are obtained from RDKit.[54] Cell lines are embedded using Graph Convolutional Networks[85] by integrating the protein-protein interaction (PPI) network with the gene expression vector. Finally, PRODeepSyn then used an MLP to predict drug synergy. **AuDNNsynergy**[22] utilizes the same features as those in PRODeepSyn. Different from PRODeepSyn, it trains three autoencoders to predict drug synergy. **DeepSynergy**[20] 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**[21] 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**[19] takes molecule graphs

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 rely 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 preprocessing to increase the usability for fair comparison. We have made our implementations of all five comparison approaches available for future studies.

Implementation details for cell-line based drug synergy prediction task

In the drug synergy prediction, the input of our model is a triplet, including a pair of drugs and the cell line feature, and the output is the probability of the drug pairs being synergistic on this cell line. Following the previous works,[19–23] we used a threshold of 0.5 to determine the class label for the input triplet $x = (d_a, d_b, c)_i$. The additional features are gene expression values of one cell line in terms of TPM (transcripts per million), which is a fixed-size vector $c^{\bigstar} \in \mathbb{R}^{n_G}$ representing gene expression levels across n_G genes. To model the cell line features, we first determined an overexpressed gene set from $c^{\bigstar}$. We denote this gene set as $\widetilde{G} = \left\{\mathfrak{g}_i | 1 \leq i \leq n_G, c_i^{\bigstar} \geq T_H\right\}$ where T_H is a threshold to determine whether a gene is overexpressed in this cell line. In practice, we set $T_H = 400$. Similar to the processing of the drug target modality, we also added one-hop information on the PPI network: $\widehat{G} = \bigcup_{\mathfrak{g} \in \widetilde{G}} \{\mathfrak{g}\} \cup \mathcal{N}_{k_{\mathfrak{g}}}(\mathfrak{g})$. Then the cell line features, which is the context features embeddings, were calculated by aggregating embeddings of all genes in $\widehat{G}$ together: $\boldsymbol{c} = f_{cell\ line}(Concat(\{\mathcal{E}_{\mathfrak{g}} | \mathfrak{g} \in \widehat{G}\}))$, where $\mathcal{E}_{\mathfrak{g}}$ is learnable embedding for each gene and shared with the drug target modality processing. In the augmentor, we first concatenated the drug embeddings and context features together: $\left\{\mathbf{z}_i[m_a, m_b, \boldsymbol{c}] \triangleq \left[\mathbf{z}_{m_a} | \mathbf{z}_{m_b} | \boldsymbol{c}\right]_i | m_a, m_b \in \mathcal{M}\right\}$. We used one layer MLP as the augmentor transformation function :$\Psi_{aug}\left(\mathbf{z}_{m_a}, \mathbf{z}_{m_b}, \boldsymbol{c}\right) := \psi_{\omega}(\mathbf{z}_i[m_a, m_b, \boldsymbol{c}]) = Dropout(ReLU(\boldsymbol{\omega}^T \cdot \mathbf{z}_i[m_a, m_b, \boldsymbol{c}]))$, where $\boldsymbol{\omega} \in \mathbb{R}^{(3*d_{hid}, 2*d_{hid})}$. Here $\widehat{\mathbf{z}}_i[m_a, m_b, \boldsymbol{c}]$ refers to the embeddings for each augmented drug combination. Finally, we applied a classifier to get the predicted probability for each data instance. We denote this classifier as:

$\psi_{output}\left(\widehat{\mathbf{z}}_i[m_a, m_b, \boldsymbol{c}]\right) = \boldsymbol{\omega}_2^{\ T} \cdot Dropout(ReLU(\boldsymbol{\omega}_1^{\ T} \cdot \mathbf{z}_i[m_a, m_b, \boldsymbol{c}]))$,where $\boldsymbol{\omega}_1 \in \mathbb{R}^{(2*d_{hid}, d_{hid})}$ and $\boldsymbol{\omega}_2 \in \mathbb{R}^{d_{hid}}$.

In the drug synergy prediction task, we set the number of layers of both the DeeperGCN and Transformer to 6 and the hidden dimension size to 384 and 512 respectively. We set the transformed hidden dimension d_{hid} for all the modalities to 512. The architecture of transformation layers for all modalities $\left\{MLP_{\theta[m]} | m \in \mathcal{M}\right\}$ is one linear layer mapping function with different input dimension sizes. We trained the model using the Adam optimizer with a batch size of 128 for 100,000 steps. We adopted a polynomial decay scheduler with 4,000 warm-up steps. We performed a grid search for the learning rate within $\left[1 \times 10^{-5}, 5 \times 10^{-5}, 8 \times 10^{-5}, 1 \times 10^{-4}, 5 \times 10^{-4}\right]$. λ_{aux}, $\lambda_{CSS}^{(InfoNCE)}$ and $\lambda_{CSS}^{(MSE)}$ were set to 0.01. We chose top $k = 8$ predictions in the aggregator. To evaluate all comparison approaches and our method on this highly imbalanced dataset, we used four metrics: balanced accuracy (BACC), that is the average of sensitivity (true positive rate) and specificity (true negative rate), area under the precision-recall curve (AUPRC), F_1 score and Cohen's Kappa statistic. All metrics are higher the better.

In **Fig. 2**, we investigated the agreement between different modalities on the GDSC-Combo dataset. Here we trained eight models based on Pisces's architecture and each of them only uses one type of modality. For missing modalities, we assigned learnable embeddings and optimized them in the training procedure. In **Fig. 3e**, we also extended Pisces that is trained on two-drug combination to three-drug combination prediction. We first separated each three-drug combination into three two-drug combinations, and got a prediction for each of them. Finally, we used the average predictions to be the synergy probability of these three drug combinations. In **Supplementary Fig. 13**, we investigated how the performance varied if we chose different top k in the aggregator. We evaluated the vanilla cross validation setting. First, we observed that balanced accuracy (BACC) was sensitive to the value of k. We observed that BACC did not change too much when k is smaller than 32, indicating our approach was robust to the selection of k. Notably, the performance deteriorated a lot in terms of BACC when we aggregated all 64 predictions to calculate the output, demonstrating that our strategy could exclude noisy predictions.

Identification of a breast cancer drug-sensitive pathway

The GDSC-Combo dataset has 102,893 triplets (a pair of drugs and a cell line) which encompasses 63 drugs and 125 cell lines. These 125 cell lines include breast, colorectal and

pancreas cancer cell lines. Each triplet was represented as $(d_a, d_b, c)_i$. Here d_a is the anchor compound and d_b is the library compound. Notably, $(d_a, d_b, c)_i$ and $(d_b, d_a, c)_i$ represent two triplets with the same drug pair and cell line, but the order of anchor and library compounds was swapped. Considering all possible combinations of these 63 drugs and 125 cell lines, there were 488,250 possible triplets. For each triplet $(d_a, d_b, c)_i$, if neither $(d_a, d_b, c)_i$ or $(d_b, d_a, c)_i$ appeared in the collected GDSC-Combo dataset, we collected it as a novel triplet. When collecting novel triplets, if $(d_b, d_a, c)_i$ had already been collected, $(d_a, d_b, c)_i$ would not be collected. We finally collected 176,344 novel triplets. We applied Pisces to these novel triplets and filtered out triplets with predicted probabilities less than 0.9. For each triplet, we collected the drug target genes $\left\{\mathrm{g}_i^{(d)}\right\}_{i=1}^{n_{target}}$ and overexpressed genes $\widetilde{G} = \left\{\mathrm{g}_j^{(c)} | 1 \le j \le n_G, c_i^{\star} \ge T_H\right\}$ in the cancer cell lines. Therefore, we constructed multiple gene pairs, each pair being represented as $\left(\mathrm{g}_i^{(d)}, \mathrm{g}_j^{(c)}\right)$ and obtained 97,479 gene pairs in total. For each pair, we treated $\mathrm{g}_j^{(c)}$ as the overactive genes and the drug target $\mathrm{g}_i^{(d)}$ as the essential genes to the cancer cells. This logic matched the genetic interactions in synthetic dosage lethality (SDL) networks.[48] Therefore, we used a published SDL networks database to filter out gene pairs not in the SDL networks. To focus on studying breast cancer, we selected gene pairs collected from triplets on 51 breast cancer cell lines. We finally collected 12 SDL gene pairs with 19 genes as a breast cancer drug-sensitive pathway.

We verified this breast cancer drug-sensitive pathway on breast cancer patients from the TCGA dataset. We collected the gene mutation data and gene expression data. If the expression level of one gene from a patient was higher than 70% of all patients, we determined this gene as a highly-expressed gene. If one gene was mutated or its expression level was lower than 70% patients, we determined it as inhibited genes. We determined the pathway to be activated as long as there existed one gene pair of the 12 SDL pairs, from which $\mathrm{g}_i^{(d)}$ was inhibited and $\mathrm{g}_j^{(c)}$ was highly-expressed. We also noticed that the breast cancer patients can be classified into ER+ breast cancer patients and ER- breast cancer patients. Therefore, based on the status of ER and the activation of our identified pathway, we now clustered these breast cancer patients into four groups: pathway activated (ER+), pathway not activated (ER+), pathway activated (ER-), and pathway not activated (ER-) patients. We therefore perform survival analysis using the lifelines python package.[86]

Xenograft dataset processing

The xenograft dataset refers to the drug response data for the patient tumor derived xenograft models. Hui Gao et al. released the xenograft data for both single drugs and drug combinations that provided tumor volume changes at different time points after transplantation.[50] We denote the tumor changes at time t as $\Delta Vol_t = \frac{Vol_t - Vol_0}{Vol_0}$, where Vol_t represents the tumor sizes at time t and Vol_0 represents the tumor size when $t = 0$. We collected the response of 26 drugs across 191 mice with gene expression features, and 13 drugs of them can find target genes in our collected DrugBank and GDSC drug target dataset. In the raw dataset, the gene expression features were normalized into the FPKM unit. We then converted the FPKM unit into TPM unit:

$$c_{i,j}^{TPM} = \frac{c_{i,j}^{FPKM}}{\sum_k c_{k,j}^{FPKM}} \times 10^6$$

We evaluated 3 settings on the xenograft dataset: (1) We used our model to predict the BestResponse for one drug combination across all time points. The BestResponse means the minimum value for all tumor changes when $t \geq 10d$: $\Delta Vol_{BestReponse} = \min_{t \geq 10d} \Delta Vol_t$. In BestResponse prediction, we finally collected BesResponse of 1,769 single drugs and 476 drug combinations. We used all single drug samples and 60% of the drug combination samples as the training set, and used 20% drug combination samples as validation and test set respectively. For each single drug sample, we replicated them into a drug pair as the input. (2) We evaluated our model's ability to predict the volume changes at each time point. We also used all the single drug samples as part of the training set and split the drug combination samples into the training, validation and test set. Notably, the drug combinations in all the test samples are never seen in the training set. (3) We applied our model to the extrapolation setting where we only predict the last time point. We evaluated all the three settings using 10 fold cross validation. We also collected 22,156 single drug responses and 9,035 drug combination responses across all time points.

Implementation details for drug response prediction on xenograft models

The drug response prediction task aims to predict the volume changes for a xenograft model when exposed to a specific drug combination. Since the tumor volume changes represent continuous values, we opted to modify our supervised training objective by replacing the cross entropy loss with the mean squared error loss:

$$\mathcal{L}_{SL} = -\sum_{i=1}^{n_{batch}} \left(y_i - \hat{y}_i^{\bigstar} \right)^2 - \lambda_{aux} \sum_{(m_a, m_b) \in \mathcal{M}^*} \sum_{i=1}^{n_{batch}} \left(y_i - \hat{y}_i^{(m_a, m_b)} \right)^2$$

In BestResponse prediction, the input of our model is a triplet (drug A, drug B, model c), where we used the gene expression features $\boldsymbol{c}^{exp}$ to represent the model c. We employed the same

approach utilized in the drug synergy prediction task to calculate the gene expression embeddings. This gene expression feature embeddings $\boldsymbol{c}^{exp}$ were used as the model embeddings.

When predicting drug response at time point t, the input of our model can be represented as (d_a, d_b, c, t). Here the time t is an integer representing the number of days post to the transplantation. Inspired by the position embedding in Transformer,[87] we used the following equation to calculate the time embedding:

$$\boldsymbol{c}_t^{time}[2i] = \sin\left(\frac{t}{10000^{2i/d_{hid}}}\right), \boldsymbol{c}_t^{time}[2i+1] = \cos\left(\frac{t}{10000^{2i/d_{hid}}}\right),$$

where $\boldsymbol{c}_t^{time}$ is a fixed length vector with the dimension size of $d_{hid} = 512$. In both the drug response prediction at time point t and the extrapolation tasks, we concatenated the gene expression embeddings and the time embeddings together as the additional feature embeddings:$\boldsymbol{c} = \boldsymbol{c}^{exp} + \boldsymbol{c}_t^{time}$.

We used the same hyperparameters of model architectures as the model used in cell-line-based drug synergy prediction. We compared our model to five baselines used in the drug synergy prediction task: DeepDDS,[19] AuDNNsynergy,[22] PRODeepSyn[23] and DeepSynergy.[20] In this particular setting, we did not evaluate GraphSynergy due to its requirement for all of the drugs to have known drug target genes. However, we could only collect drug targets for 13 drugs using existing DrugBank and GDSC dataset. We trained our model for 10,000 iterations with a batch size of 64. We set the warm-up steps of the polynomial decay learning scheduler to 2,000. We evaluated the correlations between the predicted volume changes and the observed volume changes using the Spearman correlations coefficients and Pearson correlation coefficients.

DrugBank and TwoSIDES dataset processing

The drug-drug interaction dataset was collected from the DrugBank and TwoSIDES dataset. We obtained 1,706 drugs with 86 interaction types from the DrugBank dataset and 645 drugs with 963 interaction types. In the raw data, each sample can also be represented as a triplet (drug A, drug B, interaction c), which means the drug A and drug B have an interaction type c. All the triplets appearing in the raw dataset were treated as positive samples. We additionally generated samples that were not included in the original dataset as negative samples. These negative samples were created to match the quantity of the positive samples. We evaluated the vanilla cross validation, one new drug in each test pair and split by combination setting for the DrugBank dataset. The one new drug in each test pair setting means that each combination in the validation set has one drug that is never seen in the training set. The split by combination setting means that all validation combinations never appeared in the training set. We followed the partition procedure in R^2-DDI[88] and GMPNN-CS[89] to split the dataset.

Comparison approaches for drug-drug interaction predictions

We compared Pisces to six different approaches on drug-drug interaction prediction tasks. **R^2-DDI** is a recently proposed DDI prediction approach by refining interaction type features to improve the prediction accuracy. R^2-DDI simultaneously learned embeddings for different interaction types and drug pairs. The probability of an interaction type between a pair of drugs is calculated using dot product between their embeddings. **GMPNN-CS** utilized a gated message passing mechanism to capture chemical substructure information at different sizes. The final prediction was calculated using the pairs between the learned substructures. **SSI-DDI** utilized graph attention layers to extract substructure interactions from molecular graphs. The final predictions were calculated by aggregating the interaction scores between substructures. **MR-GNN** utilized the multi-resolution based neural networks to capture the node features of molecular graphs and used these features for drug-drug interaction prediction. **MHCADDI** predicts side effects between drugs based on the message passing and co-attention mechanisms. **GAT-DDI** directly used graph attentions for predicting drug-drug interactions. We adopted experimental results of these approaches reported in R^2-DDI and GMPNN-CS. These results come from validation set accuracy on the DrugBank and TwoSIDES dataset. MHCADDI did not provide their accuracy and AUPRC on the TwoSIDES dataset. The reported AUROC of MHCADDI on TwoSIDES dataset was 0.882, much lower than our method (an AUROC of 0.914), therefore we did not make comparisons in **Fig. 5d-f**.

Implementation details for drug-drug interaction prediction task

In the DDI prediction task, we aim to predict which type of interaction is associated with one pair of drugs. Here we formulated it as k_c binary classification tasks, where k_c refers to the number of DDI types. We queried the model with an input of triplet $(d_a, d_b, c)_i$. The output $\hat{y}_i$ denotes the probability of the input drug pair (d_a, d_b) has this interaction c. To model the interaction type features in each triplet, we assigned a learnable embedding $\varepsilon_{DDI}^{(i)}$ for each interaction type c as the additional feature embeddings $\boldsymbol{c}$. In the augmentor for DDI prediction, the input of the augmentor module only contains a pair of drugs. We first concatenated the drug pair embeddings: $\left\{\mathbf{z}_i[m_a, m_b] \triangleq \left[\mathbf{z}_{m_a} | \mathbf{z}_{m_b}\right]_i | m_a, m_b \in \mathcal{M}\right\}$. The augmentor transformation layer was used to embed the drug pairs: $\hat{\mathbf{z}}_i[m_a, m_b] = \Psi_{aug}\left(\mathbf{z}_{m_a}, \mathbf{z}_{m_b}\right) = Dropout(Tanh(\boldsymbol{\omega}^T \cdot \mathbf{z}_i[m_a, m_b]))$, where $\boldsymbol{\omega} \in \mathbb{R}^{(2*d_{hid}, 2*d_{hid})}$. We modified the classification head inspired by R^2-DDI.[88] Remember that here we transformed DDI predictions into k_c binary classification tasks and denoted the

learnable embeddings of possible interaction types as $\boldsymbol{c}$. We first embedded the interaction type features $\hat{\boldsymbol{c}} = Dropout(Tanh(\boldsymbol{c} + \boldsymbol{\omega}_c^T[\hat{\mathbf{z}}_i[m_a, m_b]|\boldsymbol{c}])), \boldsymbol{\omega}_c \in \mathbb{R}^{(3*d_{hid}, d_{hid})}$. Then we output the predicted probability using a bilinear model: $\psi_{output}\left(\hat{\mathbf{z}}_i[m_a, m_b], \boldsymbol{c}\right) = Dropout(Tanh(\boldsymbol{\omega}_z^T \cdot \hat{\mathbf{z}}_i[m_a, m_b]))^T \cdot \hat{\boldsymbol{c}}, \boldsymbol{\omega}_z \in \mathbb{R}^{(2*d_{hid}, d_{hid})}$.

In the vanilla cross validation setting on both DrugBank dataset and TwoSIDES dataset, we used a 12 layers DeeperGCN and Transformer model. We set the hidden dimension size to 384 and 768 for DeeperGCN and Transformer respectively. The output dimension of the transformation layer d_{hid} was 512. In the one new drug in each test pair and two new drugs in each test pair setting on DrugBank, we used a 6 layers DeeperGCN with a hidden dimension of 384 and a 6 layers Transformer with the hidden dimension of 512. We set d_{hid} to 512. $\left\{MLP_{\theta[m]} | m \in \mathcal{M}\right\}$ are one-depth linear layers. We trained the model with a learning rate of 1×10^{-4} for 50,000 iterations. We chose the polynomial decay scheduler for the learning rate with 4000 warm-up steps. We ran the grid-search for the dropout rate within $[0.1, 0.2, 0.3, 0.4, 0.5]$ and find the optimal dropout rate of 0.1 for vanilla cross validation setting, 0.3 for one new drug in each test pair setting and 0.5 for split by combination setting. The training batch size was set to 128.

We also used predictions from Pisces to construct a DDI network. We first selected ten representative biomedical related drug classes from the drug ontology. Then we obtained 834 drugs, with each of them belonging to at least one of the selected drug classes. We narrowed down our scope on predicting 50 most frequent drug-drug interaction types in DrugBank and collected 868,400 predictions. Next, by filtering out the drug-drug interactions with a predicted probability less than 0.9 and choosing 100 most confident interactions for each drug, we collected 27,510 interactions. We finally applied Fisher's exact test to identify interaction types that were significantly associated with a pair of classes. Based on these identified interaction types and corresponding drug classes, we were able to generate a DDI network.

Declaration of Interests

The authors declare that there is no competing interest or any potential conflicts of interests associated with this work.

Data Availability

The drug synergy prediction on cell line dataset, drug response prediction on xenografts dataset and the drug-drug interaction dataset are available at https://figshare.com/articles/dataset/Pisces_dataset/23272049.

Code Availability

Pisces code is available at https://github.com/HanwenXuTHU/Pisces.

Reference

1. Al-Lazikani, B., Banerji, U. & Workman, P. Combinatorial drug therapy for cancer in the post-genomic era. *Nat. Biotechnol.* **30**, 679–692 (2012).
2. Jia, J. *et al.* Mechanisms of drug combinations: interaction and network perspectives. *Nat. Rev. Drug Discov.* **8**, 111–128 (2009).
3. Lehár, J. *et al.* Synergistic drug combinations tend to improve therapeutically relevant selectivity. *Nat. Biotechnol.* **27**, 659–666 (2009).
4. Dancey, J. E. & Chen, H. X. Strategies for optimizing combinations of molecularly targeted anticancer agents. *Nat. Rev. Drug Discov.* **5**, 649–659 (2006).
5. Hao, J. *et al.* Combination treatment with FAAH inhibitors/URB597 and ferroptosis inducers significantly decreases the growth and metastasis of renal cell carcinoma cells via the PI3K-AKT signaling pathway. *Cell Death Dis.* **14**, 247 (2023).
6. Sicklick, J. K. *et al.* Molecular profiling of cancer patients enables personalized combination therapy: the I-PREDICT study. *Nat. Med.* **25**, 744–750 (2019).
7. Schafer, J. M. *et al.* Targeting MYCN-expressing triple-negative breast cancer with BET and MEK inhibitors. *Sci. Transl. Med.* **12**, (2020).
8. Manchado, E. *et al.* A combinatorial strategy for treating KRAS-mutant lung cancer. *Nature* **534**, 647–651 (2016).
9. Hashimoto, M. *et al.* PD-1 combination therapy with IL-2 modifies CD8+ T cell exhaustion program. *Nature* **610**, 173–181 (2022).

10. Jin, H. *et al.* EGFR activation limits the response of liver cancer to lenvatinib. *Nature* **595**, 730–734 (2021).

11. To, C. *et al.* An allosteric inhibitor against the therapy-resistant mutant forms of EGFR in non-small cell lung cancer. *Nat Cancer* **3**, 402–417 (2022).

12. Gavert, N. *et al.* Ex vivo organotypic cultures for synergistic therapy prioritization identify patient-specific responses to combined MEK and Src inhibition in colorectal cancer. *Nat Cancer* **3**, 219–231 (2022).

13. Lee, J. *et al.* Effective breast cancer combination therapy targeting BACH1 and mitochondrial metabolism. *Nature* **568**, 254–258 (2019).

14. Yang, W. *et al.* Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Res.* **41**, D955–61 (2013).

15. Burd, A. *et al.* Precision medicine treatment in acute myeloid leukemia using prospective genomic profiling: feasibility and preliminary efficacy of the Beat AML Master Trial. *Nat. Med.* **26**, 1852–1858 (2020).

16. Holbeck, S. L. *et al.* The national cancer institute ALMANAC: A comprehensive screening resource for the detection of anticancer drug pairs with enhanced therapeutic activity. *Cancer Res.* **77**, 3564–3576 (2017).

17. Menden, M. P. *et al.* Community assessment to advance computational prediction of cancer drug combinations in a pharmacogenomic screen. *Nat. Commun.* **10**, 2674 (2019).

18. Collins, F. S. & Varmus, H. A New Initiative on Precision Medicine. *N. Engl. J. Med.* **372**, 793–795 (2015).

19. Wang, J., Liu, X., Shen, S., Deng, L. & Liu, H. DeepDDS: deep graph neural network with

attention mechanism to predict synergistic drug combinations. *Brief. Bioinform.* **23**, (2022).

20. Preuer, K. *et al.* DeepSynergy: predicting anti-cancer drug synergy with Deep Learning. *Bioinformatics* **34**, 1538–1546 (2017).
21. 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. Med. Inform. Assoc.* **28**, 2336–2345 (2021).
22. Zhang, T., Zhang, L., Payne, P. R. O. & Li, F. Synergistic Drug Combination Prediction by Integrating Multiomics Data in Deep Learning Models. *Methods Mol. Biol.* **2194**, 223–238 (2021).
23. Wang, X. *et al.* PRODeepSyn: predicting anticancer synergistic drug combinations by embedding cell lines with protein–protein interaction network. *Brief. Bioinform.* **23**, bbab587 (2022).
24. Janizek, J. D. *et al.* Uncovering expression signatures of synergistic drug responses via ensembles of explainable machine-learning models. *Nat Biomed Eng* (2023) doi:10.1038/s41551-023-01034-0.
25. Chen, T., Kornblith, S., Norouzi, M. & Hinton, G. A Simple Framework for Contrastive Learning of Visual Representations. in *Proceedings of the 37th International Conference on Machine Learning* (eds. Iii, H. D. & Singh, A.) vol. 119 1597–1607 (PMLR, 13--18 Jul 2020).
26. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet classification with deep convolutional neural networks. *Commun. ACM* **60**, 84–90 (2017).
27. Caron, M. *et al.* Emerging properties in self-supervised vision transformers. in *Proceedings of the IEEE/CVF international conference on computer vision* 9650–9660 (2021).

28. Wei, J. & Zou, K. EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks. *arXiv [cs.CL]* (2019).

29. Feng, S. Y. *et al.* A Survey of Data Augmentation Approaches for NLP. *arXiv [cs.CL]* (2021).

30. Kumar, A., Ahuja, K., Vadapalli, R. & Talukdar, P. Syntax-guided controlled generation of paraphrases. *Transactions of the Association for Computational Linguistics* **8**, 330–345 (2020).

31. Bubeck, S. *et al.* Sparks of Artificial General Intelligence: Early experiments with GPT-4. *arXiv [cs.CL]* (2023).

32. Brown, T. *et al.* Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* **33**, 1877–1901 (2020).

33. Radford, A. *et al.* Language Models are Unsupervised Multitask Learners. https://life-extension.github.io/2020/05/27/GPT%E6%8A%80%E6%9C%AF%E5%88%9D%E6%8E%A2/language-models.pdf (2019).

34. Radford, A., Narasimhan, K., Salimans, T. & Sutskever, I. Improving language understanding by generative pre-training. https://www.cs.ubc.ca/~amuham01/LING530/papers/radford2018improving.pdf (2018).

35. Goodfellow, I. *et al.* Generative adversarial networks. *Commun. ACM* **63**, 139–144 (2020).

36. Karras, T., Laine, S. & Aila, T. A style-based generator architecture for generative adversarial networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **43**, 4217–4228 (2021).

37. Kim, S. *et al.* PubChem 2023 update. *Nucleic Acids Res.* **51**, D1373–D1380 (2023).

38. Liu, S. *et al.* Pre-training Molecular Graph Representation with 3D Geometry. *arXiv [cs.LG]* (2021).

39. Wishart, D. S. *et al.* DrugBank 5.0: a major update to the DrugBank database for 2018.

Nucleic Acids Res. **46**, D1074–D1082 (2018).

40. Edwards, C. *et al.* Translation between Molecules and Natural Language. in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* 375–413 (Association for Computational Linguistics, 2022).
41. Hastings, J. *et al.* ChEBI in 2016: Improved services and an expanding collection of metabolites. *Nucleic Acids Res.* **44**, D1214–9 (2016).
42. Kuhn, M., Letunic, I., Jensen, L. J. & Bork, P. The SIDER database of drugs and side effects. *Nucleic Acids Res.* **44**, D1075–9 (2016).
43. Kuhn, M., Campillos, M., Letunic, I., Jensen, L. J. & Bork, P. A side effect resource to capture phenotypic effects of drugs. *Mol. Syst. Biol.* **6**, 343 (2010).
44. Shoemaker, R. H. The NCI60 human tumour cell line anticancer drug screen. *Nat. Rev. Cancer* **6**, 813–823 (2006).
45. Whetzel, P. L. *et al.* BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. *Nucleic Acids Res.* **39**, W541–5 (2011).
46. Martínez-Romero, M. *et al.* NCBO Ontology Recommender 2.0: an enhanced approach for biomedical ontology recommendation. *J. Biomed. Semantics* **8**, 21 (2017).
47. Jaaks, P. *et al.* Effective drug combinations in breast, colon and pancreatic cancer cells. *Nature* **603**, 166–173 (2022).
48. Jerby-Arnon, L. *et al.* Predicting cancer-specific vulnerability via data-driven detection of synthetic lethality. *Cell* **158**, 1199–1209 (2014).
49. Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast

tumours. *Nature* **490**, 61–70 (2012).

50. Gao, H. *et al.* High-throughput screening using patient-derived tumor xenografts to predict clinical trial drug response. *Nat. Med.* **21**, 1318–1325 (2015).
51. Tatonetti, N. P., Ye, P. P., Daneshjou, R. & Altman, R. B. Data-driven prediction of drug effects and interactions. *Sci. Transl. Med.* **4**, 125ra31 (2012).
52. Li, J., Socher, R. & Hoi, S. C. H. DivideMix: Learning with Noisy Labels as Semi-supervised Learning. *arXiv [cs.CV]* (2020).
53. 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)* 770–778 (IEEE, 2016).
54. Landrum, G. RDKit: Open-source cheminformatics. 2006. *Google Scholar* (2006).
55. Durie, B. G. M. *et al.* Bortezomib with lenalidomide and dexamethasone versus lenalidomide and dexamethasone alone in patients with newly diagnosed myeloma without intent for immediate autologous stem-cell transplant (SWOG S0777): a randomised, open-label, phase 3 trial. *Lancet* **389**, 519–527 (2017).
56. Haley, C. A. *et al.* Novel 6-Month Treatment for Drug-Resistant Tuberculosis, United States. *Emerg. Infect. Dis.* **27**, 332–334 (2021).
57. Goldman, A. *et al.* Targeting tumor phenotypic plasticity and metabolic remodeling in adaptive cross-drug tolerance. *Sci. Signal.* **12**, (2019).
58. Lin, X., Liao, J., Geng, X., Dan, H. & Chen, L. Concurrent inhibition of ErbB family and MEK/ERK kinases to suppress non-small cell lung cancer proliferation. *Am. J. Transl. Res.* **12**, 847–856 (2020).
59. Lee, Y., Wang, Y., James, M., Jeong, J. H. & You, M. Inhibition of IGF1R signaling abrogates

resistance to afatinib (BIBW2992) in EGFR T790M mutant lung cancer cells. *Mol. Carcinog.* **55**, 991–1001 (2016).

60. Araki, K. & Miyoshi, Y. Mechanism of resistance to endocrine therapy in breast cancer: the important role of PI3K/Akt/mTOR in estrogen receptor-positive, HER2-negative breast cancer. *Breast Cancer* **25**, 392–401 (2018).
61. O'Brien, N. A. *et al.* Targeting activated PI3K/mTOR signaling overcomes acquired resistance to CDK4/6-based therapies in preclinical models of hormone receptor-positive breast cancer. *Breast Cancer Res.* **22**, 89 (2020).
62. Lundberg, S. M. & Lee, S.-I. A Unified Approach to Interpreting Model Predictions. in *Advances in Neural Information Processing Systems* (eds. Guyon, I. et al.) vol. 30 (Curran Associates, Inc., 2017).
63. Lundberg, S. M. *et al.* From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence* **2**, 56–67 (2020).
64. Luong, M.-T., Le, Q. V., Sutskever, I., Vinyals, O. & Kaiser, L. Multi-task Sequence to Sequence Learning. *arXiv [cs.LG]* (2015).
65. Fang, X. *et al.* Geometry-enhanced molecular representation learning for property prediction. *Nature Machine Intelligence* **4**, 127–134 (2022).
66. Wang, Y., Wang, J., Cao, Z. & Barati Farimani, A. Molecular contrastive learning of representations via graph neural networks. *Nature Machine Intelligence* **4**, 279–287 (2022).
67. Liu, S. *et al.* Multi-modal Molecule Structure-text Model for Text-based Retrieval and Editing. *arXiv [cs.LG]* (2022).
68. Zhu, J. *et al.* Dual-view Molecule Pre-training. *arXiv [q-bio.QM]* (2021).

69. Lin, Z. *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).

70. Rao, R. M. *et al.* MSA Transformer. in *Proceedings of the 38th International Conference on Machine Learning* (eds. Meila, M. & Zhang, T.) vol. 139 8844–8856 (PMLR, 18--24 Jul 2021).

71. Ross, J. *et al.* Large-scale chemical language representations capture molecular structure and properties. *Nature Machine Intelligence* **4**, 1256–1264 (2022).

72. Li, G., Xiong, C., Qian, G., Thabet, A. & Ghanem, B. DeeperGCN: All You Need to Train Deeper GCNs. *https://openreview.net › forumhttps://openreview.net › forumhttps://openreview.net › pdfhttps://openreview.net › pdf* (2023).

73. Gu, Y. *et al.* Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. *ACM Trans. Comput. Healthcare* **3**, 1–23 (2021).

74. Cho, H., Berger, B. & Peng, J. Compact Integration of Multi-Network Topology for Functional Analysis of Genes. *Cell Syst* **3**, 540–548.e5 (2016).

75. van den Oord, A., Li, Y. & Vinyals, O. Representation Learning with Contrastive Predictive Coding. *arXiv [cs.LG]* (2018).

76. Belghazi, M. I. *et al.* Mutual Information Neural Estimation. in *Proceedings of the 35th International Conference on Machine Learning* (eds. Dy, J. & Krause, A.) vol. 80 531–540 (PMLR, 10--15 Jul 2018).

77. Kong, L. *et al.* A Mutual Information Maximization Perspective of Language Representation Learning. *arXiv [cs.CL]* (2019).

78. Chi, Z. *et al.* InfoXLM: An Information-Theoretic Framework for Cross-Lingual Language Model Pre-Training. *arXiv [cs.CL]* (2020).

79. 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 (openaccess.thecvf.com, 2020).

80. Natarajan, N., Dhillon, I. S., Ravikumar, P. K. & Tewari, A. Learning with noisy labels. *Adv. Neural Inf. Process. Syst.* **26**, (2013).

81. Zhang, Z. & Sabuncu, M. Generalized cross entropy loss for training deep neural networks with noisy labels. *Adv. Neural Inf. Process. Syst.* **31**, (2018).

82. Musen, M. A. *et al.* The National Center for Biomedical Ontology. *J. Am. Med. Inform. Assoc.* **19**, 190–195 (2012).

83. Rogers, D. & Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **50**, 742–754 (2010).

84. Fabian, B. *et al.* Molecular representation learning with language models and domain-relevant auxiliary tasks. *arXiv [cs.LG]* (2020).

85. Kipf, T. N. & Welling, M. Semi-Supervised Classification with Graph Convolutional Networks. *arXiv [cs.LG]* (2016).

86. Davidson-Pilon, C. lifelines: survival analysis in Python. *J. Open Source Softw.* **4**, 1317 (2019).

87. Vaswani, A. *et al.* Attention is all you need. *Adv. Neural Inf. Process. Syst.* **30**, (2017).

88. Lin, J. *et al.* R2-DDI: relation-aware feature refinement for drug–drug interaction prediction. *Briefings in* (2023).

89. Nyamabo, A. K., Yu, H., Liu, Z. & Shi, J.-Y. Drug-drug interaction prediction with learnable size-adaptive molecular substructures. *Brief. Bioinform.* **23**, (2022).

Figure 1

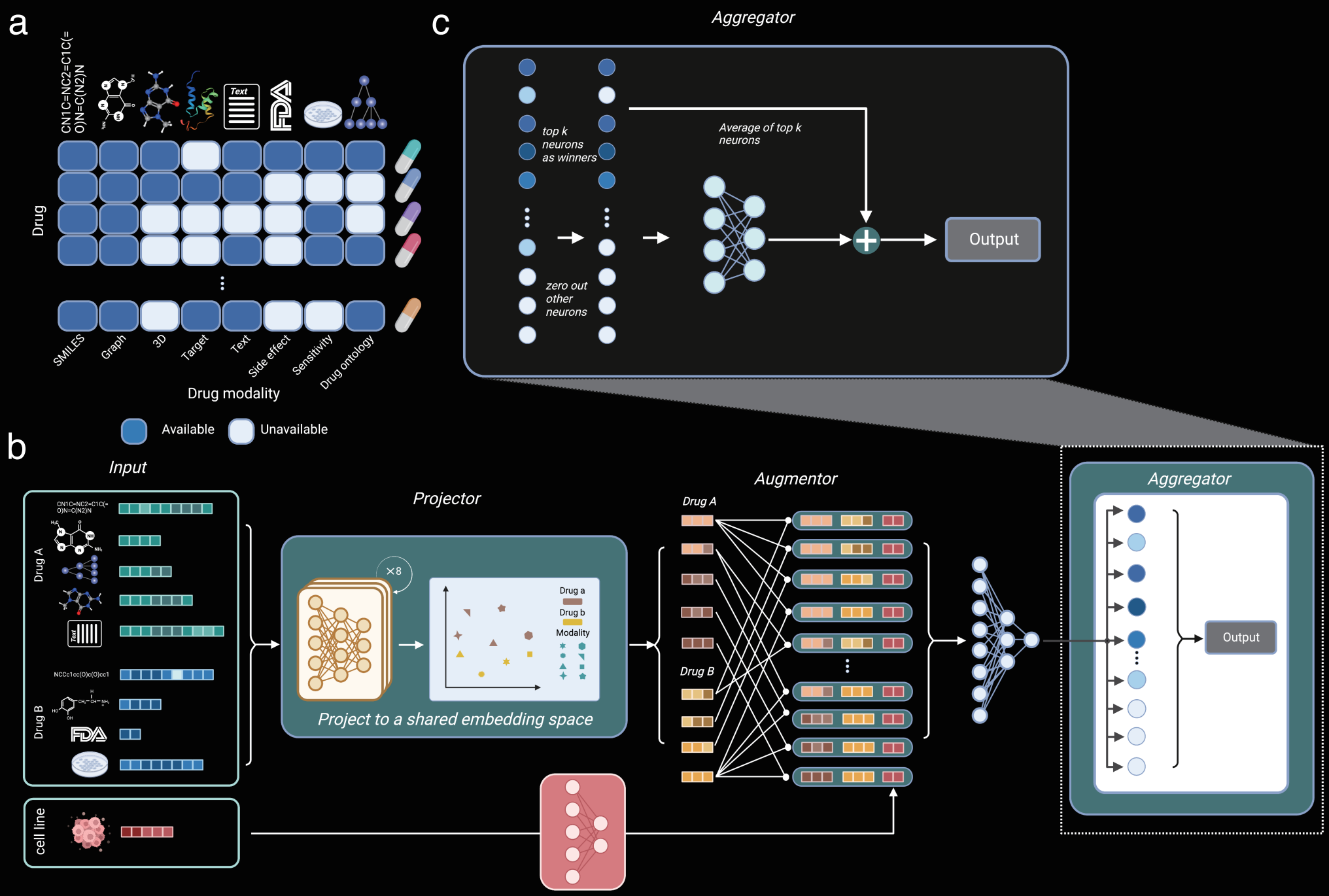
a
CN1C=NC2=C1C(=O)N=C(N2)N
Text
Drug
SMILES
Graph
3D
Target
Text
Side effect
Sensitivity
Drug ontology
Drug modality
Available
Unavailable
c
Aggregator
top k neurons as winners
zero out other neurons
Average of top k neurons
Output
b
Input
Drug A
Drug B
cell line
Projector
×8
Drug a
Drug b
Modality
Project to a shared embedding space
Augmentor
Drug A
Drug B
Aggregator
Output

Figure 2

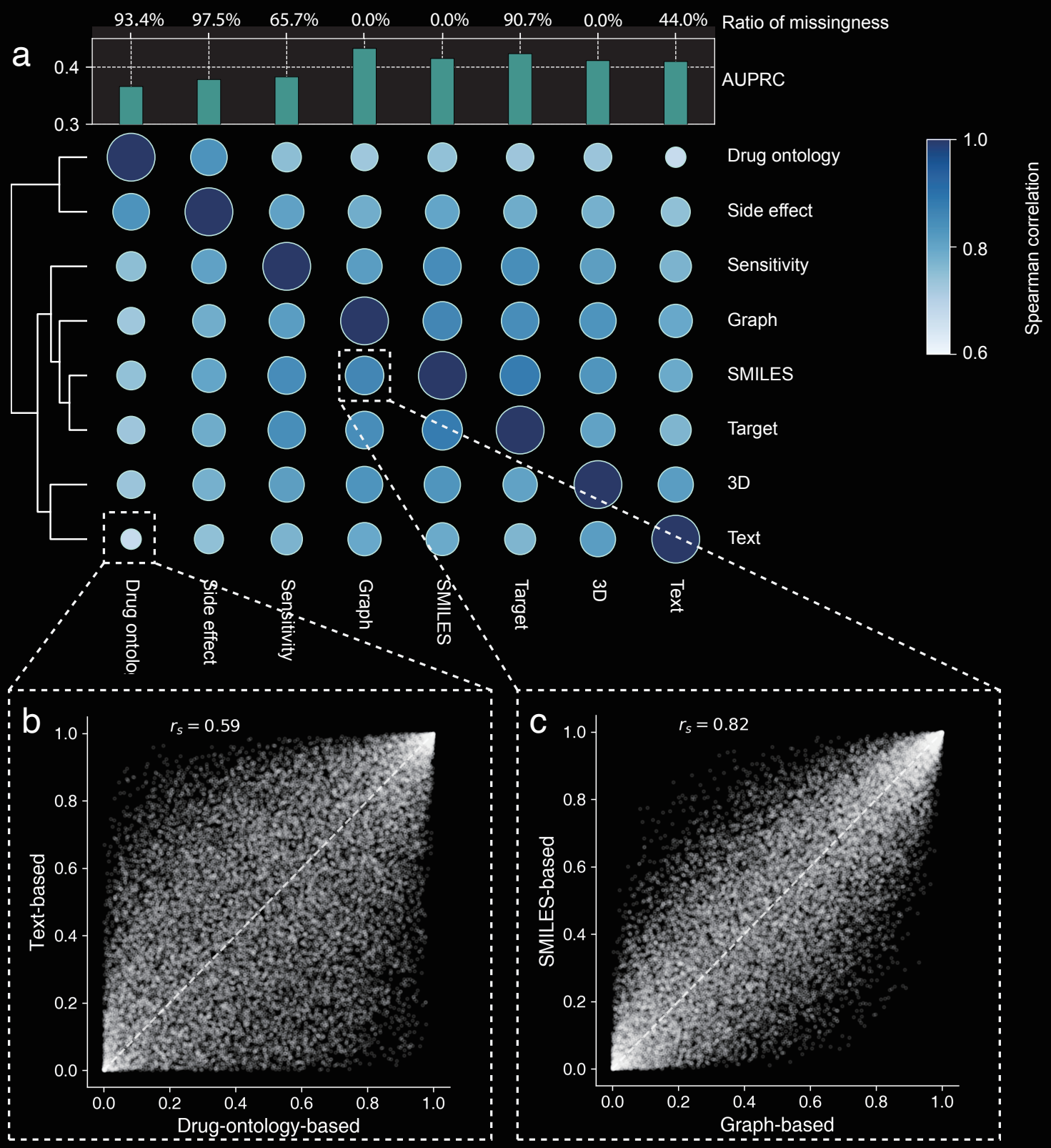
a
93.4%
97.5%
65.7%
0.0%
0.0%
90.7%
0.0%
44.0%
Ratio of missingness
AUPRC
0.4
0.3
Drug ontology
Side effect
Sensitivity
Graph
SMILES
Target
3D
Text
1.0
0.8
0.6
Spearman correlation
b
$r_s = 0.59$
Text-based
Drug-ontology-based
c
$r_s = 0.82$
SMILES-based
Graph-based

Figure 3

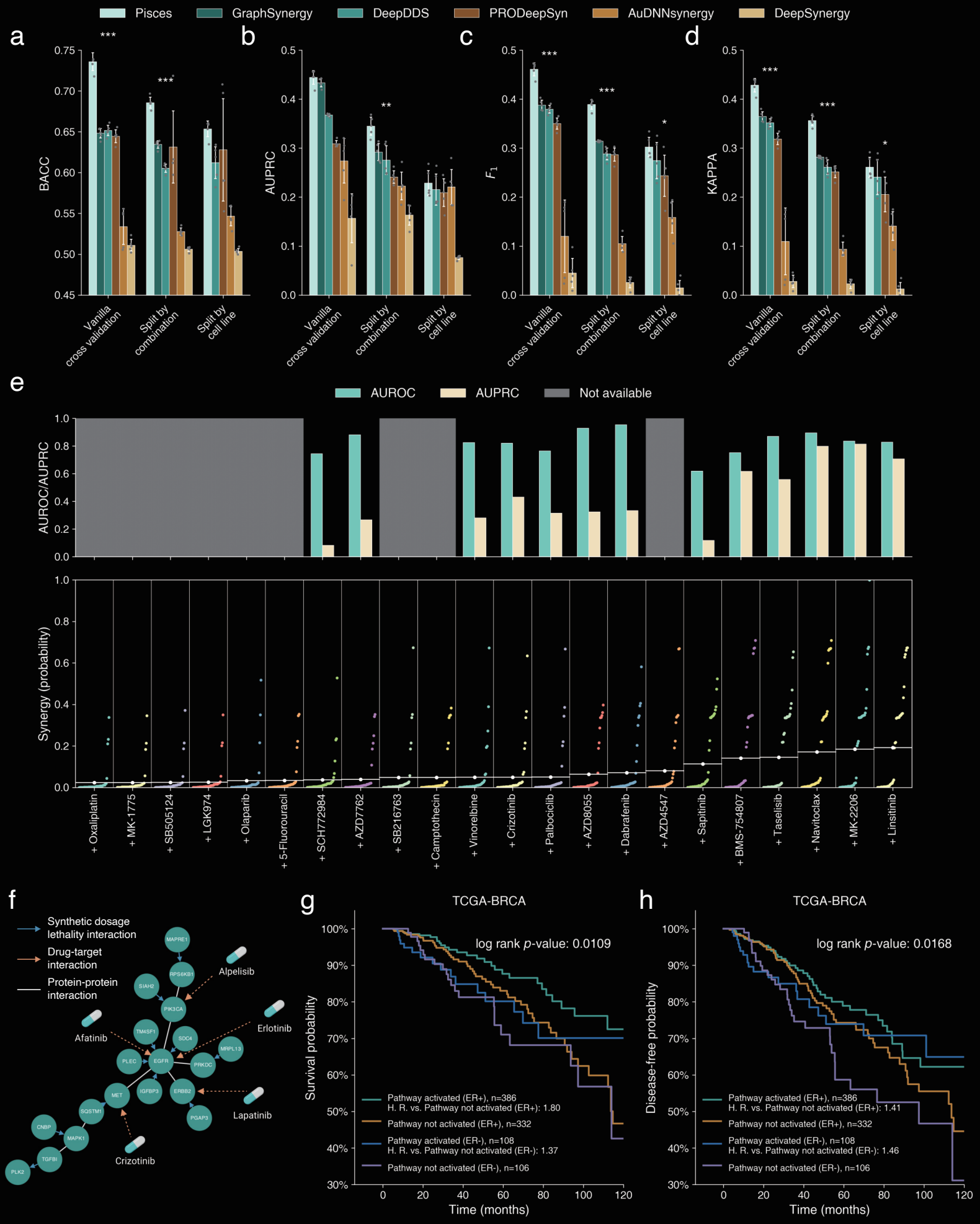
Pisces
GraphSynergy
DeepDDS
PRODeepSyn
AuDNNsynergy
DeepSynergy
a
b
c
d
BACC
AUPRC
F1
KAPPA
Vanilla cross validation
Split by combination
Split by cell line
e
AUROC
AUPRC
Not available
AUROC/AUPRC
Synergy (probability)
+ Oxaliplatin
+ MK-1775
+ SB505124
+ LGK974
+ Olaparib
+ 5-Fluorouracil
+ SCH772984
+ AZD7762
+ SB216763
+ Camptothecin
+ Vinorelbine
+ Crizotinib
+ Palbociclib
+ AZD8055
+ Dabrafenib
+ AZD4547
+ Sapitinib
+ BMS-754807
+ Taselisib
+ Navitoclax
+ MK-2206
+ Linsitinib
f
Synthetic dosage lethality interaction
Drug-target interaction
Protein-protein interaction
MAPRE1
RPS6KB1
SIAH2
PIK3CA
TM4SF1
SDC4
MRPL13
PLEC
EGFR
PRKDC
MET
IGFBP3
ERBB2
SQSTM1
PGAP3
CNBP
MAPK1
TGFBI
PLK2
Alpelisib
Afatinib
Erlotinib
Lapatinib
Crizotinib
g
TCGA-BRCA
log rank p-value: 0.0109
Survival probability
Pathway activated (ER+), n=386
H. R. vs. Pathway not activated (ER+): 1.80
Pathway not activated (ER+), n=332
Pathway activated (ER-), n=108
H. R. vs. Pathway not activated (ER-): 1.37
Pathway not activated (ER-), n=106
Time (months)
h
TCGA-BRCA
log rank p-value: 0.0168
Disease-free probability
Pathway activated (ER+), n=386
H. R. vs. Pathway not activated (ER+): 1.41
Pathway not activated (ER+), n=332
Pathway activated (ER-), n=108
H. R. vs. Pathway not activated (ER-): 1.46
Pathway not activated (ER-), n=106
Time (months)

Figure 4

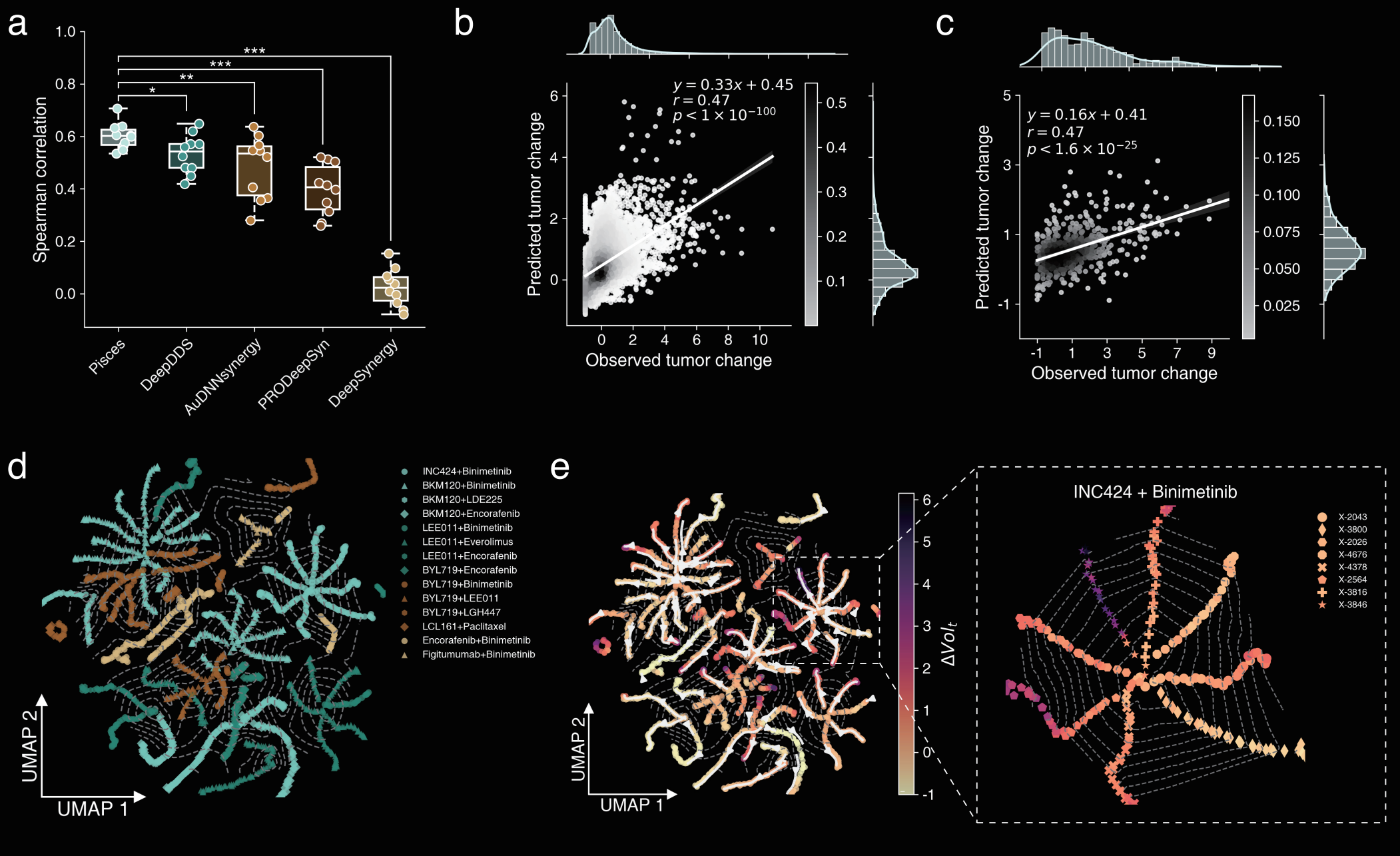
a
Spearman correlation
1.0
0.8
0.6
0.4
0.2
0.0
***
***
**
*
Pisces
DeepDDS
AuDNNsynergy
PRODeepSyn
DeepSynergy
b
$y = 0.33x + 0.45$
$r = 0.47$
$p < 1 \times 10^{-100}$
Predicted tumor change
Observed tumor change
c
$y = 0.16x + 0.41$
$r = 0.47$
$p < 1.6 \times 10^{-25}$
Predicted tumor change
Observed tumor change
d
INC424+Binimetinib
BKM120+Binimetinib
BKM120+LDE225
BKM120+Encorafenib
LEE011+Binimetinib
LEE011+Everolimus
LEE011+Encorafenib
BYL719+Encorafenib
BYL719+Binimetinib
BYL719+LEE011
BYL719+LGH447
LCL161+Paclitaxel
Encorafenib+Binimetinib
Figitumumab+Binimetinib
UMAP 2
UMAP 1
e
ΔVol_t
UMAP 2
UMAP 1
INC424 + Binimetinib
X-2043
X-3800
X-2026
X-4676
X-4378
X-2564
X-3816
X-3846

Figure 5

Drugbank
Pisces
R2-DDI
GMPNN-CS
SSI-DDI
MR-GNN
MHCADDI
GAT-DDI
a
b
c
Accuracy
AUPRC
AUROC
Vanilla cross validation
One new drug in each test pair
Two new drugs in each test pair
TwoSIDES
d
e
f
Olefinic compound
g
Organochlorine compound
Sulfonic acid derivative
Zwitterion
Hydroxy steroid
Organohalogen compound
Homopolycyclic compound
Organic fundamental parent
Aromatic amine
Phenols
Decrease the diuretic activities.
Increase the constipating activities.
Decrease the sedative activities.
Increase the thrombogenic activities.
Increase the anticholinergic activities.
Decrease the vasoconstricting activities.
Increase the risk or severity of QTc prolongation.