A CIL-750K DETAILS

The original CIL dataset includes 919,265 five-channel fields of view containing 30,616 test compounds. It also includes metadata files which record morphological features for each cell in each image, both at the single cell level and at the population average level (i.e. per well); a workflow for image analysis to generate morphological features is also provided. Quality control indicators are provided as metadata, indicating fields of view that are out of focus or contain highly fluorescent material or debris. Chemical annotations are also provided for the application of compound processing. Figure 1 shows the molecular data distribution and the number of view per molecule in CIL dataset.

In CIL, each molecular intervention is imaged from multiple views in an experimental well and the experiment was repeated several times, resulting in an average of 30 views for each molecule. In order to keep the data balanced, we restricted each molecule to a maximum of 30 images, resulting in a cross-modal graph-image benchmark containing 750K views. Each view has a resolution of 692×520 pixels and 5 channels. These images were imaged with the ImageXpress Micro XLS automated microscope at 20×magnification. We resize the images to 128×128 without any cropping to fit the CNN models' input format. Figure 2 shows examples of molecules and corresponding images from the CIL dataset and Figure 3 shows the multiple views of a random selected molecule.

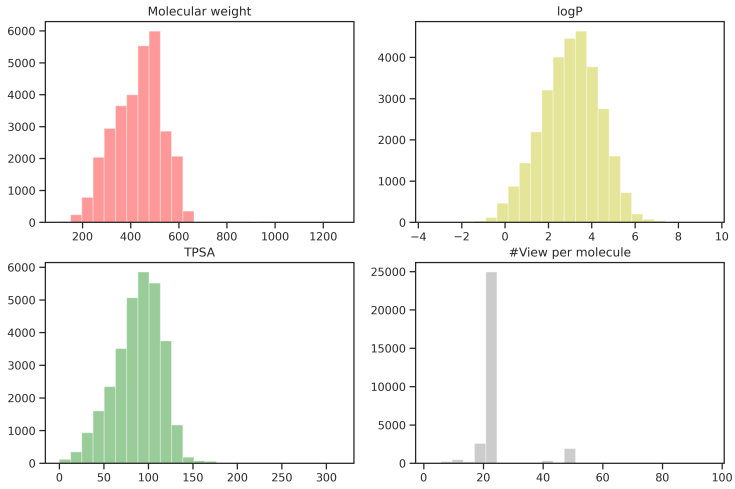

Figure 1: Data Distribution of CIL.

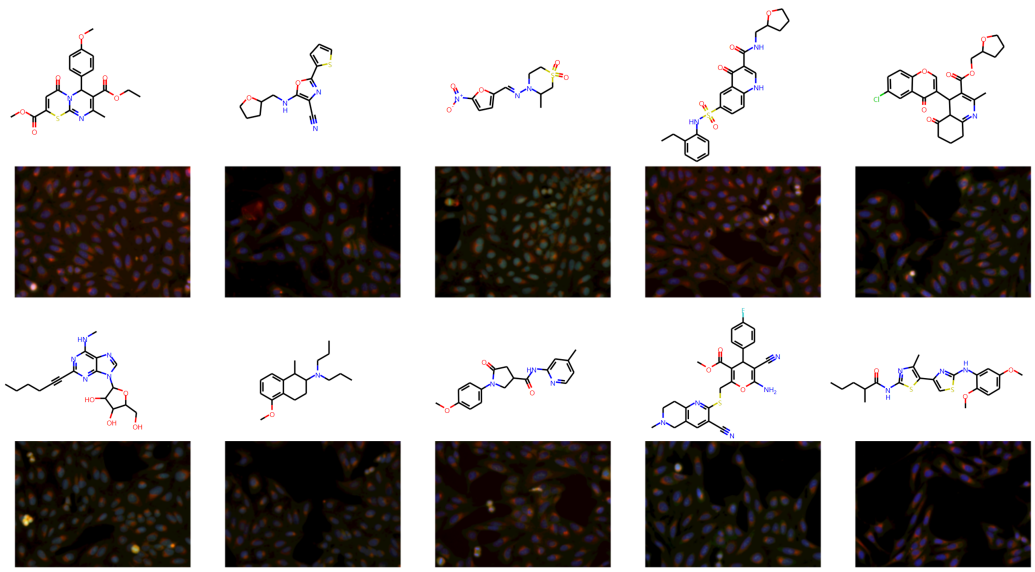

Figure 2: A random selection of 10 molecules and corresponding cellular images (1 view).

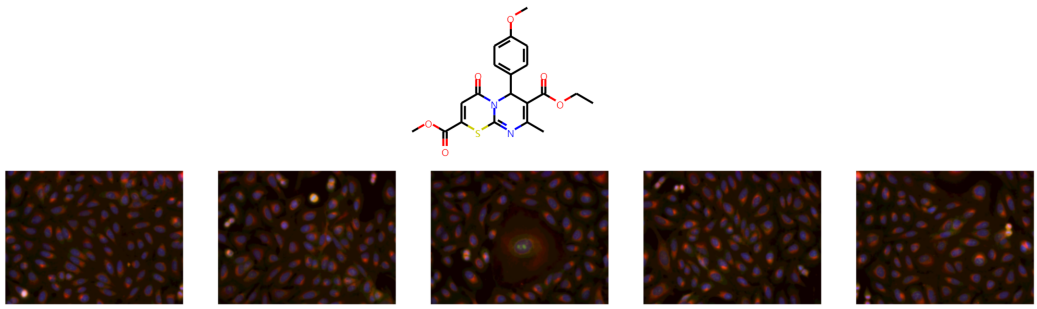

Figure 3: Five different views on the same molecule.

B IMPLEMENTATION DETAILS AND HYPERPARAMETERS.

Here we describe the implementation details for pre-training and fine-tuning stages.

Generative Graph-image matching. We employ Variational Auto-Encoders (VAE) as generative agents, which are asked to recover the representation of one modality given the parallel representation from the other modality. For example, when generating the cellular image from their corresponding molecular graph, we need to model the conditional likelihood $p(z_I|z_G)$. The reparameterized variable could be defined as $z_G = \mu_G + \sigma_G \cdot \zeta$ with mean μ_G, covariance $sigma_G$, and $\zeta \sim \mathcal{N}(0,1)$. Therefore, we have the following lower bound:

$$logp(z_I|z_G) \geq \mathbb{E}_{q(z_G|z_I)}[\log p(z_I|z_G)] - \mathcal{D}_{KL}(q(z_G|z_I)||p(z_G)) \tag{1}$$

Similarly, when generating the molecular graph from their corresponding cellular image, we have:

$$logp(z_G|z_I) \geq \mathbb{E}_{q(z_I|z_G)}[\log p(z_G|z_I)] - \mathcal{D}_{KL}(q(z_I|z_G)||p(z_I)) \tag{2}$$

Both the above objectives are composed of a conditional log-likelihood and a KL-divergence.

Following the variation representation reconstruction (VRR) of (1), we use the mean-squared error (MSE) for reconstruction on the representation space:

$$\mathbb{E}_{q(z_G|z_I)}[\log p(z_I|z_G)] = \mathbb{E}_{q(z_G|z_I)}[\left\| z_G - q(z_G^{'}|z_I) \right\|_2^2] + C \tag{3}$$

$$\mathbb{E}_{q(z_I|z_G)}[\log p(z_G|z_I)] = \mathbb{E}_{q(z_I|z_G)}[\left\| z_I - q(z_I^{'}|z_G) \right\|_2^2] + C \tag{4}$$

Thus, combining both two regularizers mentioned above, the final GM loss function can be formulated as:

$$\begin{aligned} \mathcal{L}_{GM} = &-\frac{\lambda_{kl}}{2}(\mathcal{D}_{KL}(q_\phi(z_I|z_G)||p(z_I)) + \mathcal{D}_{KL}(q_\phi(z_G|z_I)||p(z_G))) \\ &+\frac{1}{2}(\mathbb{E}_{q(z_I|z_G)}[\left\| z_I - q(z_I^{'}|z_G) \right\|_2^2] + \mathbb{E}_{q(z_G|z_I)}[\left\| z_G - q(z_G^{'}|z_I) \right\|_2^2]) \end{aligned} \tag{5}$$

Pre-training Our pre-training model consists of a Graph Isomorphism Network (GIN) from (2) with 5 layers and 300 hidden dimensions and a residual convolutional neural network (ResNet-34) (3) with 63.5M parameters. We pre-train the model for 100 epochs using a batch size of 1024 on 8 NVIDIA 3090TI GPUs. We use the Adam optimizer with an initial learning rate of 3e-4 and weight decay of 0.02. We take image with resolution of 128×128. The margin γ is set to 4. Pre-training on 750k graph-image pairs for MIGA takes 8 hours, far less than 26 hours for ContextPred and 48 hours for GraphCL on 280k molecules of GEOM-Drugs.

Graph-Image Retrieval We randomly split the CIL-750K dataset into a training set of 27.6K molecules corresponding to 680K images, and hold out the remaining of the data for testing. The held-out data consists of 3K molecules and the corresponding 50K images. We formulate the retrieval task as a ranking problem. In the inference phase, given a query molecular graph in the held-out set, we take images in the held-out set as a candidate pool and rank candidate images according to the L2 distance between the image embeddings and the molecular embeddings, and vice versa. The negative sampling rate is set to positive: negative = 1:100. We use GraphCL (4) and cross-modal pre-learning methods, CLIP (5)and ALIGN (6) as well as baselines. The encoder part of these methods has been changed to the same setting as MIGA, but the decoder, training part and technical tricks have not been changed. We also re-implemented CLOOME (7) following the setting shown in the paper, except that we did not use the trick in warm-up iterations as no improvements were observed. After pre-training, We use the pre-trained model to output embeddings of molecular graphs and celluar images, then rank the candidate pool based on their L2 similarity. Experiments are performed 5 times with different seeds. The average of MRR, AUC, Hit@1, Hit@5 and Hit@10 are reported.

Zero-shot Graph Retrieval We use MIGA to prioritize the functional molecules from a compound pool. This task mimics the real-world virtual screening scenario using morphological features observed when overexpressing a specific gene by cDNA interventions. We collected cellular images that were overexpressed with cDNA open reading frames for 6 genes by (8), including BRCA1, HIF1A, JUN, STAT3, TP53 and HSPA5. We used ExCAPEDB database (9) to retrieve gene-specific agonists and inactive molecules that have not been observed in training set. For each gene, we constructed a candidate pool with 20 agonists and 100 negative molecules, denoted as $P(I, G_A, G_N)$. For each gene, given a random selected input image I with resolution of 128×128, we ask the model to rank the G_A in front of the G_N. We use Hit@10 as the metric to evaluate our model in such a zero-shot graph retrieval task, where the random Hit@10 is 0.17 (20/120).

Clinical Outcome Prediction DATASET To standardize the clinical-trial-outcome predictions, we use the Trial Outcome Prediction (TOP) benchmark constructed by HINT, which incorporate rich data components including drug molecule information, disease information, trial eligibility criteria and trial outcome information. Herein, we consider phase-level evaluation on the trial outcome, where we predict the outcome of a single-phase study. Since each phase has different goals (e.g., phase I is for safety, whereas phases II and III are for efficacy), we evaluate phases I, II, and III separately. We follow the data splitting proposed by HINT and data statistics are shown in Table 1

Phase-level	Molecule	Successes	Failures
Phase I	944	564	380
Phase II	2865	1396	1469
Phase III	1752	1203	549

Table 1: Statistics of Clinical Outcome Datasets.

BASELINES We first include three machine learning-based methods (RF, LR, XGBoost) and a knowledge-aware GNN model HINT as our baseline. **Random Forest (RF)** is a bagging algorithm for classification or regression problems, which obtains the prediction by voting or averaging of each base learner (decision tree). **Logistic regression (LR)** is a simple, parallelizable classification method that uses maximum likelihood estimation for parameter estimation. **XGBoost**, also called an extreme gradient boosting tree, uses CART regression tree or linear classifier as a base learner to ensemble model predictions. These machine learning baselines utilize 1024-dimensional Morgan fingerprint features for trial outcome prediction. **HINT** is a hierarchical interaction network designed for clinical-trial-outcome predictions. It uses (1) 1024-dimensional Morgan fingerprint features, (2) a pre-trained BERT model to encode eligibility criteria into sentence embedding and (3) a graph-based attention model GRAM to encode disease information. Furthermore, we also include the self-supervised learning methods to constitute our baselines, including **ContextPred**, **GraphLoG**, **GROVER**, **GraphCL** and **JOAO**. For this downstream task, we use the molecule encoders over input molecule graphs for the fine-tuning of clinical outcome prediction.

FUNE-TUNING HYPERPARAMETER For fine-tuning, an extra linear classifier is appended to the pre-trained GNN. We fine-tune the model for 100 epochs using a batch size of 32 with a dropout rate

of 50%. We use the Adam optimizer with an initial learning rate of 1e-3. Experiments are performed for 5 times, with mean and standard deviation of ROC-AUC and PR-AUC are reported.

Molecular property Prediction DATASET **BBBP**: The Blood-brain barrier penetration dataset includes binary labels for 2035 compounds on their permeability properties.**Tox21**: The Tox21 dataset was created in the Tox21 data challenge, which contains qualitative toxicity measurements for 7821 compounds on 12 different targets, including nuclear receptors and stress response pathways. **HIV**: 41K compounds with binary labels for HIV virus replication inhibition. **ToxCast** includes 8576 drug compounds with binary labels of toxicity experiment outcomes with 617 targets. **ESOL**: The ESOL is a small dataset consisting of water solubility data for 1128 compounds.**Lipophilicity**: Experimental data for the octanol/water distribution coefficient of 4200 molecules.

Dataset	Tasks	Type	Molecule	Metric
BBBP	1	GC	2,035	ROC-AUC
Tox21	12	GC	7,821	ROC-AUC
HIV	1	GC	41K	ROC-AUC
ToxCast	617	GC	8,576	ROC-AUC
ESOL	1	GR	1,128	RMSE
Lipophilicity	1	GR	4,198	RMSE

Table 2: Statistics of datasets. GC for Graph Classification, GR for Graph Regression.

Features	Size	Description
Atom type	101	type of atom (e.g C,N,O)
Hybridization	6	sp, sp2, sp3, sp3d, sp3d2 or unknown
Number of H	1	number of bond hydrogen atoms
Degrees	1	number of neighbor atoms
Formal Charges	1	number of formal charge
Valences	1	number of valences

Table 3: Atom features

Features	Size	Description
Bond type	4	single, double, triple, aromatic
Stereo	6	none, any, E/Z or cis/trans
In ring	1	whether the bond is part of a ring
conjugated	1	whether the bond is conjugated

Table 4: Bond features

FEATURIZATION EXTRACTION The feature extraction contains three parts: 1) Node feature extraction. 2) Bond feature extraction. 3) Topology connection matrix. We use RDKit to extract all features as the input of GNN. Table 3 and Table 4 show the atom and bond features we used in MIGA.

FUNE-TUNING HYPERPARAMETER For fine-tuning, we followed the GraphCL's (4) settings. An extra linear layer is appended to the pre-trained GNN to perform classification and regression, repectively. We fine-tune the model for 100 epochs using a batch size of 32 with a dropout rate of 50%. We use the Adam optimizer with an initial learning rate of 1e-3. Experiments are performed for 5 times, with mean and standard deviation of AUC and RMSE are reported.

C CASE STUDY FOR IMAGE RETRIEVAL.

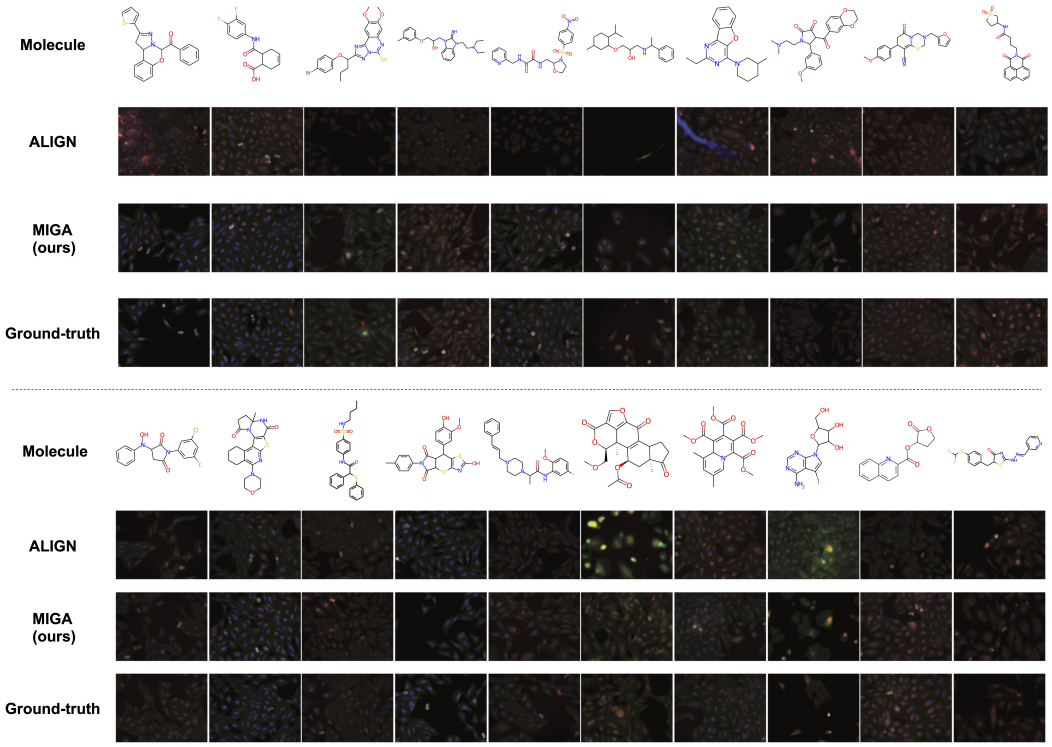

Figure 4: Case study for image retrieval task. The images retrieved by our method and baseline (ALIGN) are shown.

D CASE STUDY FOR ZERO-SHOT GRAPH RETRIEVAL.

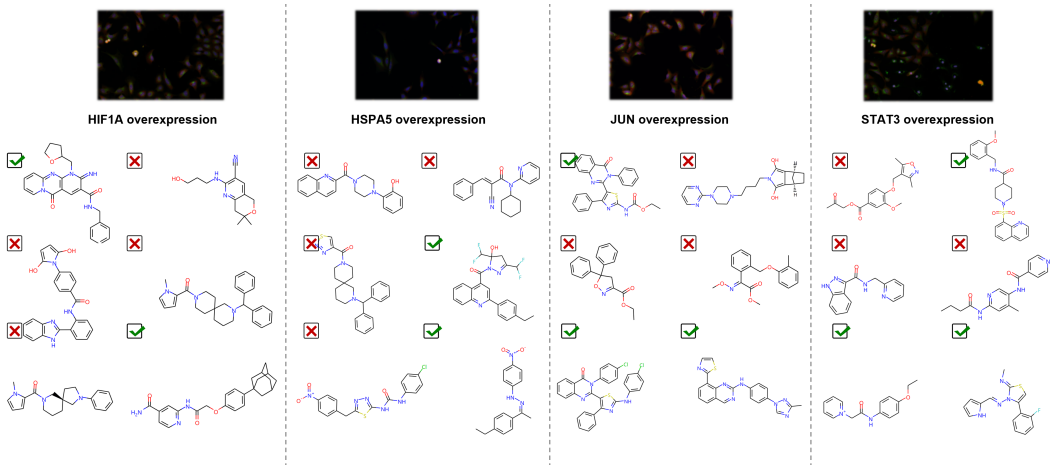

Figure 5: Case study for zero-shot graph retrieval task. The figure shows the cells induced by the cDNA interventions for specific genes (HIF1A, HSPA5, JUN, STAT3) and our model can find diverse molecules that have similar functions to these cDNA interventions (ticked).

E PERFORMANCE COMPARISON OF MIGA AND BASELINE APPROACHES FOR PHASE-LEVEL-OUTCOME PREDICTIONS (WITH STANDARD DEVIATIONS)

Task	Phase I		Phase II		Phase III	
Metrics	PR-AUC	AUC	PR-AUC	AUC	PR-AUC	AUC
LR	0.634 (0.007)	0.487 (0.006)	0.509 (0.014)	0.534 (0.017)	0.675 (0.010)	0.528 (0.003)
RF	0.651 (0.013)	0.488 (0.009)	0.488 (0.005)	0.523 (0.009)	0.722 (0.011)	0.588 (0.013)
XGBoost	0.646 (0.003)	0.508 (0.006)	0.481 (0.004)	0.516 (0.007)	0.712 (0.009)	0.597 (0.015)
HINT	0.683 (0.015)	0.516 (0.005)	0.537 (0.004)	0.584 (0.003)	0.689 (0.003)	0.621 (0.006)
ContextPred	0.693 (0.006)	0.541 (0.019)	0.544 (0.019)	0.586 (0.003)	0.710 (0.023)	0.554 (0.036)
GraphLoG	0.681 (0.016)	0.539 (0.016)	0.550 (0.043)	**0.593 (0.043)**	0.719 (0.024)	0.554 (0.024)
GROVER	0.711 (0.015)	0.559 (0.024)	0.521 (0.005)	0.574 (0.011)	0.713 (0.013)	0.575 (0.028)
GraphCL	0.721 (0.020)	0.578 (0.018)	0.543 (0.008)	0.588 (0.004)	**0.733** (0.011)	**0.601 (0.008)**
JOAO	**0.736 (0.019)**	**0.586 (0.018)**	**0.546 (0.018)**	0.587 (0.000)	0.720 (0.000)	0.563 (0.006)
MIGA	**0.758** (0.010)	**0.601** (0.031)	**0.562** (0.010)	**0.605** (0.022)	**0.729** (0.008)	**0.654** (0.016)

Table 5: Performance comparison of MIGA and several baseline approaches for phase-level-outcome predictions on TOP dataset. We report the mean (and standard deviation) PR-AUC and ROC-AUC of five times for clinical trial outcome prediction. The best and second best results are marked **bold** and **bold**, respectively

F PERFORMANCE COMPARISON OF MIGA AND BASELINE APPROACHES FOR MOLECULAR PROPERTY PREDICTIONS (WITH STANDARD DEVIATIONS)

Dataset	Classification (AUC)					Regression (RMSE)		
	HIV	Tox21	ToxCast	BBBP	Avg.	ESOL	Lipo	Avg.
Non-pretrain	70.30 (0.51)	68.90 (0.80)	58.60 (1.20)	65.40(2.4)	65.80	1.278 (0.24)	0.744 (0.14)	1.011
ContextPred	74.17 (1.33)	71.44 (0.11)	60.05 (0.15)	69.87 (0.99)	68.88	1.141 (0.03)	0.724 (0.02)	**0.933**
AttrMask	75.55 (1.00)	74.58 (0.66)	59.51 (0.36)	68.88 (2.65)	69.63	1.194 (0.04)	0.736 (0.02)	0.965
EdgePred	72.53 (1.20)	68.86 (0.38)	57.39 (0.80)	63.45 (1.39)	65.56	1.146 (0.05)	0.751 (0.02)	0.949
InfoGraph	**76.22 (0.24)**	69.22 (0.78)	59.87 (0.35)	63.75 (1.52)	67.27	1.242 (0.01)	0.725 (0.01)	0.984
GraphLoG	73.29 (2.64)	69.80 (0.41)	59.22 (1.05)	68.43 (2.86)	67.68	1.194 (0.02)	0.766 (0.01)	0.980
GraphCL	74.85 (1.71)	74.19 (0.43)	61.37 (0.10)	66.13 (1.68)	69.14	1.151 (0.04)	0.745 (0.02)	0.948
GROVER	74.35 (0.92)	74.02 (0.79)	61.30 (0.13)	**69.88 (0.58)**	**69.89**	1.199 (0.02)	**0.721 (0.01)**	0.960
JOAO	74.91 (0.66)	74.60 (0.49)	**61.62 (0.37)**	68.33 (0.58)	69.87	**1.117 (0.05)**	0.753 (0.02)	0.935
MIGA	**76.38 (0.55)**	**75.23 (0.71)**	**62.34 (0.23)**	**71.52 (0.43)**	**71.37**	**1.123 (0.01)**	**0.717 (0.00)**	**0.919**

Table 6: Comparison of SSL baselines against MIGA on six OGB datasets. Mean ROC-AUC and Root Mean Squared Error (RMSE) (with the SD) of 5 times independent test are reported. The best and second best results are marked **bold** and **bold**, respectively.

REFERENCES

[1] Shengchao Liu, Hanchen Wang, Weiyang Liu, Joan Lasenby, Hongyu Guo, and Jian Tang. Pre-training molecular graph representation with 3D geometry. In *International Conference on Learning Representations*, 2022.

[2] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In *7th International Conference on Learning Representations, ICLR*. OpenReview.net, 2019.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[4] Yuning You, Tianlong Chen, Yongduo Sui, Ting Chen, Zhangyang Wang, and Yang Shen. Graph contrastive learning with augmentations. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS*, 2020.

[5] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[6] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021.

[7] Ana Sanchez-Fernandez, Elisabeth Rumetshofer, Sepp Hochreiter, and Günter Klambauer. Contrastive learning of image-and structure-based representations in drug discovery. In *ICLR2022 Machine Learning for Drug Discovery*, 2022.

[8] Mohammad Hossein Rohban, Shantanu Singh, Xiaoyun Wu, Julia B Berthet, Mark-Anthony Bray, Yashaswi Shrestha, Xaralabos Varelas, Jesse S Boehm, and Anne E Carpenter. Systematic morphological profiling of human gene and allele function via cell painting. *Elife*, 6:e24060, 2017.

[9] Jiangming Sun, Nina Jeliazkova, Vladimir Chupakhin, Jose-Felipe Golib-Dzib, Ola Engkvist, Lars Carlsson, Jörg Wegner, Hugo Ceulemans, Ivan Georgiev, Vedrin Jeliazkov, et al. Excape-db: an integrated large scale dataset facilitating big data analysis in chemogenomics. *Journal of cheminformatics*, 9(1):1–9, 2017.