CONTRASTIVE LEARNING OF IMAGE- AND STRUCTURE-BASED REPRESENTATIONS IN DRUG DISCOVERY

Ana Sanchez-Fernandez[*] **Elisabeth Rumetshofer**[*] **Sepp Hochreiter** [*†] **Günter Klambauer**[*]
[*]ELLIS Unit Linz and LIT AI Lab, Institute for Machine Learning,
Johannes Kepler University Linz, Austria
[†]Institute of Advanced Research in Artificial Intelligence (IARAI), Vienna, Austria

ABSTRACT

Contrastive learning for self-supervised representation learning has brought a strong improvement to many application areas, such as computer vision and natural language processing. With the availability of large collections of unlabeled data in vision and language, contrastive learning of language and image representations has brought impressive results. The contrastive learning methods CLIP and CLOOB have demonstrated that the learned representations are highly transferable to a large set of diverse tasks when trained on multi-modal data from two different domains. In drug discovery, similar large, multi-modal datasets comprising both cell-based microscopy images and chemical structures of molecules are available. However, contrastive learning has not been used for this type of multi-modal data in drug discovery, although transferable representations could be a remedy for the time-consuming and cost-expensive label acquisition in this domain. In this work, we present a contrastive learning method for image-based and structure-based representations of small molecules for drug discovery. Our method, Contrastive Leave One Out boost for Molecule Encoders (CLOOME), is based on CLOOB and comprises an encoder for microscopy data, an encoder for chemical structures and a contrastive learning objective. On the benchmark dataset ”Cell Painting”, we demonstrate the ability of our method to learn transferable representations by performing linear probing for activity prediction tasks. Additionally, we show that the representations could also be useful for bioisosteric replacement tasks.

1 INTRODUCTION

Contrastive learning has had a strong impact on computer vision and natural language processing. Over the last decade, supervised deep learning methods have achieved outstanding results in the field of computer vision (Krizhevsky et al., 2012; He et al., 2016). These supervised methods require large amounts of labeled data, which may be very costly or unfeasible to obtain, and they have limited generalization abilities (Sun et al., 2017; Marcus, 2018). This has led to the exploration of new methods that are able to learn robust representations of the data which can be transferred to different downstream tasks (Luo et al., 2017; Chen et al., 2020). With contrastive learning methods (Gutmann and Hyvärinen, 2010) and self-supervision these meaningful representations can be obtained without the need for large amounts of expensive manually-provided labels (He et al., 2020; Chen et al., 2020; Caron et al., 2020; Grill et al., 2020). While uni-modal methods typically use pre-text tasks (Chen et al., 2020), for multi-modal methods the self-supervision arises from the availability of two modalities of an instance, such as image and text (Radford et al., 2021; Devillers et al., 2021). Both uni-modal and multi-modal contrastive learning methods have recently had a substantial impact in computer vision and natural language processing (Jaiswal et al., 2020).

CLIP for multi-modal data yields spectacular performance at zero-shot transfer learning and has recently been improved by CLOOB. An outstanding multi-modal approach is Contrastive Language-Image Pre-training (CLIP) (Radford et al., 2021), which learns both image- and text-representations simultaneously. CLIP shows comparable performance to methods that are solely image-based and yields highly transferable representations, which is shown by its high performance at zero-shot transfer learning. However, CLIP has recently been shown to suffer from the ”explaining away" effect (Fürst et al., 2021; Pearl, 1988; Wellman and Henrion, 1993) (details in Section 2). Considering this caveat, the "Contrastive Leave One Out Boost" (CLOOB) method has been proposed (Fürst et al., 2021). CLOOB uses a different objective, the "InfoLOOB" (LOOB for "Leave One Out Bound") objective (Poole et al., 2019), which does not include the positive pair in the denominator to avoid saturation effects (Fürst et al., 2021). Moreover, continuous modern Hopfield networks (Ramsauer et al., 2021) are used to reinforce the covariance structure of the data. As a result, CLOOB has further improved zero-shot transfer learning. The ability to learn transferable representation from multi-modal data makes CLOOB the prime candidate for learning representations of molecules in drug discovery.

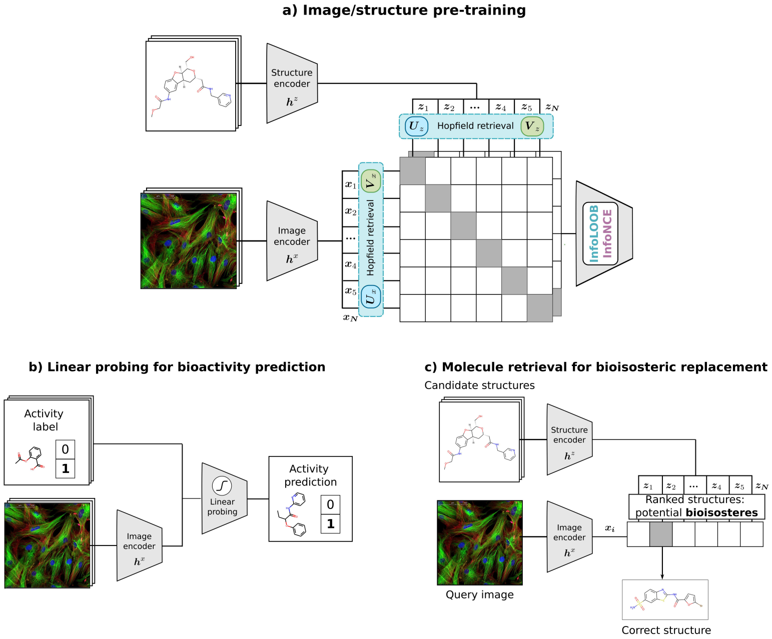
a) Image/structure pre-training
Structure encoder h^z
z_1 z_2 ... z_4 z_5 z_N
U_z Hopfield retrieval V_z
Image encoder h^x
x_1 x_2 ... x_4 x_5 x_N
V_x Hopfield retrieval U_x
InfoLOOB
InfoNCE
b) Linear probing for bioactivity prediction
Activity label
0
1
Image encoder h^x
Linear probing
Activity prediction
0
1
c) Molecule retrieval for bioisosteric replacement
Candidate structures
Structure encoder h^z
z_1 z_2 ... z_4 z_5 z_N
Ranked structures: potential bioisosteres
Image encoder h^x
x_i
Query image
Correct structure

Figure 1: Schematic representation of CLOOME. Contrastive pre-training of embeddings of the two modalities, microscopy image and chemical structure, of a molecule using the CLOOB (Fürst et al., 2021) approach. **b)** Using the CLOOME embeddings for activity prediction. A logistic regression model is trained for activity prediction tasks. **c)** The resulting embeddings can be used to rank molecules that could produce similar morphological changes, which can be considered a bioisosteric replacement task.

Contrastive learning for molecule representations in drug discovery. In drug discovery, the effect of the limited availability of data on molecules is even more severe, since the acquisition of a single bioactivity data point can cost several thousand dollars and take several weeks or months (MacArron et al., 2011; Knight et al., 2006). Therefore, methods that can learn transferable representations from unlabelled data are highly demanded. Thus, a handful of contrastive learning approaches have been recently developed for different tasks in drug discovery. MolCLR (Wang et al., 2022) uses contrastive molecule-to-molecule training by augmenting molecular graphs. Stärk et al. (2021) contrastively learn 3D and 2D molecule representations to inform the learned molecule encoder with 3D information. Lee et al. (2021) and Seidl et al. (2022) use contrastive learning for molecules and chemical reactions, and Vall et al. (2021) utilizes text representations of wet-lab procedures to enable zero-shot predictions. However, none of these methods have exploited the wealth of information contained in microscopy images of molecule-perturbed cells (Bray et al., 2016) and demonstrated strong transferability of the learned molecule encoders.

Image-based profiling of small molecules has strongly improved the drug discovery process. Characterizing a small molecule by the morphological changes it induces to a cell, is considered promising for accelerating drug discovery (Bray et al., 2016; Caicedo et al., 2017; Chandrasekaran et al., 2021; Simm et al., 2018). The advantages of this biotechnology are that it is time- and cost-effective as compared to standard activity measurements. Measuring the effects of a molecule on a biological system early in the drug discovery process might be useful to improve clinical success rates (Bender and Cortés-Ciriano, 2021). Particularly, microscopy image-based profiles of small molecules have been suggested to be effective together with deep learning methods (Chandrasekaran et al., 2021). However, the current efforts are still in standard supervised learning settings based on extracted features (Simm et al., 2018) or deep architectures (Hofmarcher et al., 2019). The amount of labeled images is in the range of few tens of thousands, although international efforts are currently building datasets which are magnitudes larger (JUMP-CP Consortium, 2022). Instead of the currently used activity measurements as labels (Simm et al., 2018; Hofmarcher et al., 2019), we propose a self-supervised contrastive learning strategy of image- and structure-based molecule encoders: Contrastive Leave One Out boost for Molecule Encoders (CLOOME). CLOOME extends recent successful contrastive learning methods to the fields of biological imaging and drug discovery. Our approach intends to overcome the limited transferability of current molecule encoders (Cai et al., 2020; Stanley et al., 2021).

Contributions

- We introduce a new contrastive learning approach for image- and structure-based representations of molecules, which is based on the contrastive learning method CLOOB.
- We show that the learned representations are highly transferable to relevant downstream tasks in drug discovery by linear probing on activity prediction tasks.
- We demonstrate that our approach learns meaningful representations of molecules which allow to retrieve potential bioisosteres.

2 CLOOME: Contrastive Leave One Out Boost for Molecule Encoders

We propose contrastive learning of representations of molecules from pairs of microscopy images and chemical structures to obtain highly transferable molecule encoders (see Figure 1). In contrast to previous approaches, in which molecule encoders learned representations using activity data (Hofmarcher et al., 2019; Simm et al., 2018) or used hand-crafted representations (Carpenter et al., 2006; Bray et al., 2017), CLOOME optimizes representations without activity data or human expertise.

The training dataset consists of N pairs of microscopy images of molecule-perturbed cells and chemical structures of molecules $\{(x_1, z_1), \ldots, (x_N, z_N)\}$. We assume that an adaptive image-encoder $\boldsymbol{h}^x(.)$ and an adaptive structure-encoder $\boldsymbol{h}^z(.)$ are available that map the images and chemical structures to their embeddings $\boldsymbol{x}_n = \boldsymbol{h}^x(x_n)$ and $\boldsymbol{z}_n = \boldsymbol{h}^z(z_n)$, respectively. Note that the original image is denoted as x_n, which is mapped to an image embedding $\boldsymbol{x}_n$ by a neural network $\boldsymbol{h}^x(.)$, e.g. a ResNet. The stacked microscopy image embeddings are denoted as $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)$ and the stacked structure embeddings as $\boldsymbol{Z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_N)$. The embeddings are normalized to such that $\|\boldsymbol{x}_n\| = \|\boldsymbol{z}_n\| = 1 \ \forall n$. For notation, see also Table A1.

In a contrastive learning setting, methods aim to increase the similarity of matched pairs and decrease the similarity of unmatched pairs. This task has often been approached by maximizing the mutual information of the embeddings using the InfoNCE loss (Radford et al., 2021; van den Oord et al., 2018; Chen et al., 2020), which is also used in the CLIP approach (Radford et al., 2021).

The InfoNCE objective function has the following form:

$$\mathrm{L}_{\mathrm{InfoNCE}} \ = \ - \ \frac{1}{N} \sum_{i=1}^{N} \ln \frac{\exp(\tau^{-1} \ \boldsymbol{x}_i^T \boldsymbol{z}_i)}{\sum_{j=1}^{N} \exp(\tau^{-1} \ \boldsymbol{x}_i^T \boldsymbol{z}_j)} \ - \ \frac{1}{N} \sum_{i=1}^{N} \ln \frac{\exp(\tau^{-1} \ \boldsymbol{x}_i^T \boldsymbol{z}_i)}{\sum_{j=1}^{N} \exp(\tau^{-1} \ \boldsymbol{x}_j^T \boldsymbol{z}_i)} \ , \tag{1}$$

where τ^{-1} is the inverse temperature parameter, which is a hyperparameter of the method.

CLIP has the problem of ”explaining away” (Pearl, 1988; Wellman and Henrion, 1993; Fürst et al., 2021). Explaining away describes the effect in which few features are over-represented while others are neglected. This effect can be present a) when learning focuses only on few features and/or b) when the covariance structure in the data is insufficiently extracted. Explaining away can be caused by saturation of the InfoNCE objective (Yeh et al., 2021; Zhang et al., 2022; Fürst et al., 2021). To ameliorate these drawbacks, CLOOB (Fürst et al., 2021) has introduced the InfoLOOB objective together with Hopfield networks as a promising method for contrastive learning. Therefore, we base our method on CLOOB.

For our approach, first image- and structure-embeddings are retrieved from stored image embeddings $\boldsymbol{U}$ and structure embeddings $\boldsymbol{V}$. $\boldsymbol{U}_{\boldsymbol{x}_i}$ denotes an image-retrieved image embedding, $\boldsymbol{U}_{\boldsymbol{z}_i}$ a structure-retrieved image embedding, $\boldsymbol{V}_{\boldsymbol{x}_i}$ an image-retrieved structure embedding and $\boldsymbol{V}_{\boldsymbol{z}_i}$ a structure-retrieved structure embedding. In analogy to CLOOB, these retrievals from continuous modern Hopfield networks are computed as follows:

$$\boldsymbol{U}_{\boldsymbol{x}_i} \ = \ \boldsymbol{U} \, \mathrm{softmax}(\beta \, \boldsymbol{U}^T \boldsymbol{x}_i) \ , \qquad (2)$$

$$\boldsymbol{U}_{\boldsymbol{z}_i} \ = \ \boldsymbol{U} \, \mathrm{softmax}(\beta \, \boldsymbol{U}^T \boldsymbol{z}_i) \ , \qquad (3)$$

$$\boldsymbol{V}_{\boldsymbol{x}_i} \ = \ \boldsymbol{V} \, \mathrm{softmax}(\beta \, \boldsymbol{V}^T \boldsymbol{x}_i) \ , \qquad (4)$$

$$\boldsymbol{V}_{\boldsymbol{z}_i} \ = \ \boldsymbol{V} \, \mathrm{softmax}(\beta \, \boldsymbol{V}^T \boldsymbol{z}_i) \ , \qquad (5)$$

where β is a scaling parameter of the Hopfield network which is considered a hyperparameter. These retrieved embeddings $\boldsymbol{U}_{\boldsymbol{x}_i}, \boldsymbol{U}_{\boldsymbol{z}_i}, \boldsymbol{V}_{\boldsymbol{x}_i}, \boldsymbol{V}_{\boldsymbol{z}_i}$ are also normalized to unit norm. By default, we store the current minibatch in the modern Hopfield networks, that is, $\boldsymbol{U} = \boldsymbol{X}$ and $\boldsymbol{V} = \boldsymbol{Z}$. Note that $\boldsymbol{X}$ contains the image embeddings ($\boldsymbol{Z}$ the structure embeddings) and we use N ambiguously both as dataset size, but also as mini-batch size to keep the notation uncluttered. The choice that $\boldsymbol{U} = \boldsymbol{X}$ and $\boldsymbol{V} = \boldsymbol{Z}$ is mostly taken because of computational constraints, while $\boldsymbol{U}$ and $\boldsymbol{V}$ could hold the whole dataset or, alternatively, exemplars. For further details on notation, see Table A1.

Then, the InfoLOOB objective (Fürst et al., 2021; Poole et al., 2019) for the retrieved embeddings is used as objective function:

$$\mathrm{L}_{\mathrm{InfoLOOB}} \ = \ - \ \frac{1}{N} \sum_{i=1}^{N} \ln \frac{\exp(\tau^{-1} \, \boldsymbol{U}_{\boldsymbol{x}_i}^T \boldsymbol{U}_{\boldsymbol{z}_i})}{\sum_{j \neq i}^{N} \exp(\tau^{-1} \, \boldsymbol{U}_{\boldsymbol{x}_i}^T \boldsymbol{U}_{\boldsymbol{z}_j})} \ - \ \frac{1}{N} \sum_{i=1}^{N} \ln \frac{\exp(\tau^{-1} \, \boldsymbol{V}_{\boldsymbol{x}_i}^T \boldsymbol{V}_{\boldsymbol{z}_i})}{\sum_{j \neq i}^{N} \exp(\tau^{-1} \, \boldsymbol{V}_{\boldsymbol{x}_j}^T \boldsymbol{V}_{\boldsymbol{z}_i})} \ . \qquad (6)$$

Microscopy image encoder. Microscopy images differ from natural images in several aspects, for example the variable number of channels that depends on the staining procedure (Pepperkok and Ellenberg, 2006; Hofmarcher et al., 2019). Although standard image encoders, such as Residual Networks (He et al., 2016) could be in principle used with minor adjustments, alternative architectures, such as multiple instance learning approaches, could be required for very high resolution datasets (Ilse et al., 2020). In all our experiments, we use a ResNet-50 encoder with five input channels and downsized the microscopy images to 320x320.

Molecule structure encoder. Since the advent of Deep Learning, a large number of architectures to encode molecules have been suggested (Lusci et al., 2013; Dahl et al., 2014; Unterthiner et al., 2014; Kearnes et al., 2016; Jiang et al., 2021). In contrast to computer vision and natural language processing, in which only few prominent architectures have emerged, there is yet no standard choice for molecule structure encoders. Due to their computational efficiency and good predictive performance, CLOOME uses descriptor-based fully-connected networks (Mayr et al., 2016; 2018) with 4 hidden layers of 1024 units with ReLU activations and batch normalization (for further details see Sec. 3 and Sec. A.2). However, also any graph (Merkwirth and Lengauer, 2005; Scarselli et al., 2008; Kearnes et al., 2016; Xu et al., 2018), message-passing (Gilmer et al., 2017), or sequence-based (Alperstein et al., 2019) neural network with an appropriate pooling operation can be used as structure encoder.

3 Experiments

Dataset and preprocessing. *CellPainting.* We use pairs of microscopy images and molecules from the CellPainting (Bray et al., 2016; 2017) dataset. This dataset is a collection of high-throughput fluorescence microscopy images of U2OS cells treated with different small molecules (Bray et al., 2016). The dataset consists of 919,265 five-channel images corresponding to 30,616 different molecules. The experiment to obtain the microscopy images was conducted using 406 multi-well plates, and each one of the before mentioned individual images are views from a sample spanning the space in the corresponding well, so that six adjacent views belong to one single sample. After

Type	Method	AUC	F1	AUC >0.9	AUC >0.8	AUC >0.7
Linear probing on self-supervised	CLOOME	**0.714**±0.20	0.395±0.32	57	84	109
Supervised	ResNet	**0.731**±0.19	0.508±0.30	68	94	119
	DenseNet	0.730±0.19	0.530±0.30	61	98	121
	GapNet	0.725±0.19	0.510±0.29	63	94	117
	MIL-Net	0.711±0.18	0.445±0.32	61	81	105
	M-CNN	0.705±0.19	0.482±0.31	57	78	105
	SC-CNN	0.705±0.20	0.362±0.29	61	83	109
	FNN	0.675±0.20	0.361±0.31	55	71	90

Table 1: Comparison of the linear evaluation of the learned representations against fully supervised methods (Hofmarcher et al., 2019). Note that the CLOOME encoders do not have access to any activity data. The features produced by the CLOOME encoder are still predictive for activity data as shown by fitting a logistic regression model, considered as linear probing. CLOOME reaches the performance of the several supervised methods, which indicates transferability of the learned representations (Chen et al., 2020). The best method in each category is marked bold.

disregarding images of untreated cells used for control as well as erratic images (out of focus or containing high fluorescence material), our final dataset comprises 759,782 microscopy images treated with 30,404 different molecules.

Pre-processing. We followed the pre-processing protocol of Hofmarcher et al. (2019), which consisted of converting the original TIF images from 16-bit to 8-bit, simultaneously removing the 0.0028% of pixels with highest values. For data augmentation and to allow large batch sizes, the images were cropped and re-scaled from the original 520x696 pixel resolution to 320x320 during training. Moreover, the images were normalized using the mean and standard deviation calculated for the training split. Concerning molecules, their corresponding SMILES strings were transformed to 1024-bit Morgan fingerprints with a radius of 3, taking chirality into account (Morgan, 1964; Rogers and Hahn, 2010).

Data splits. We split our dataset into training, validation, and test set, using the splits of Hofmarcher et al. (2019). Samples which have not been used in the previous study due to missing activity data, are assigned to the training split. Note that all images belonging to the same molecular structure are moved into the same set. Finally, training, validation and test set consist of 674,357, 28,632 and 56,793 image and molecule pairs, respectively.

Pre-training, architecture and hyperparameters. We use the suggested hyperparameters of OpenCLIP (Wortsman et al., 2021) and CLOOB (Fürst et al., 2021) wherever applicable, and tuned a few critical hyperparameters, such as learning rate and the β parameter of the Hopfield layer on a validation set. The architecture of the structure encoder was inspired by previous successful models (Mayr et al., 2016) and was not subject to substantial hyperparameter optimization. Due to computational constraints, the search was limited to the hyperparameters shown in Table A2. We used the Adam optimizer (Kingma et al., 2014) with decoupled weight decay regularization (Loshchilov and Hutter, 2019). The value for weight decay was 0.1. For the learning rate scheduler, we used cosine annealing with a warm-up of 20,000 steps and hard restarts every 7 epochs (Loshchilov and Hutter, 2017). We set the dimension of the embedding space to $d = 512$, which determines the size of the output of both encoders. We use a batch size of 256 as default due to computational constraints. For activity prediction as downstream task, the inverse temperature parameter $\tau^{-1} = 30$ was used. For the Hopfield layers, the scaling hyperparameter $\beta = 22$ was selected, and the model was trained for 63 epochs based on linear probing results in the corresponding validation set. For the bioisosteric replacement task, higher validation performance was achieved by a CLIP-like architecture directly using the embeddings returned from the image and structure encoders and the InfoNCE loss. In this case, the inverse temperature parameter τ^{-1} was set to 14.3, and the model was trained for 51 epochs based on the top-1 accuracy in validation. Hence, different pre-training settings have been found to yield best results for bioactivity prediction and for the bioisosteric replacement task, respectively. However, the large majority of hyperparameters were shared in both strategies. Because of the limited exploration of the vast hyperparameter space, we expect potential improvements from further investigations. For further details on the hyperparameter selection, see Sec. A.2.

Activity prediction as downstream tasks. In this experiment, we tested whether the representations learned by CLOOME are transferable by linear probing on 209 downstream activity prediction tasks. The *linear probing* test (Alain and Bengio, 2016; Chen et al., 2020) on downstream tasks is often performed for contrastive learning approaches to check the transferability of learned features. In such experiments, the representations of the pretrained encoders are used and only a single-layer network, such as logistic regression, is fit to the given labels for the supervised task. If the linear probing test yields good predictive quality, usually below a fully supervised approach (Chen et al., 2020), the representations are considered transferable.

Linear probing evaluation. The prediction tasks that we employed for linear probing evaluation is the same as used in Hofmarcher et al. (2019). It is a subset of the Cell Painting dataset, consisting of 284,035 images for which the activity labels of the compound treatments were retrieved from ChEMBL. The retrieved labels correspond to 10,574 compounds across 209 activity prediction tasks, which are binary classification problems. However, activity data points are not available for all compounds in all of the tasks, which results in a sparse label matrix. The data was split into 70% training, 10% validation, and 20% test sets. This split had been carried out by grouping views from samples treated with the same molecule.

We use image features taken from the penultimate layer of the image encoder, omitting the classification layer. We train a logistic regression classifier, and report the corresponding metric for each task. The L2 regularization strength λ was tuned individually for each one of the tasks, considering the values $\{10^{-6}, 10^{-5}, \ldots, 10^{6}\}$.

In order to evaluate model performance for this downstream task, we use the area under the ROC curve (AUC), which is one of the most prevalent metrics for drug discovery Simm et al. (2018); Hofmarcher et al. (2019), as it considers the order of the molecules regarding their activity. We also show the number of tasks for which this metric is higher than the thresholds 0.9, 0.8 and 0.7, respectively. These thresholds have been used in previous studies (Simm et al., 2018; Hofmarcher et al., 2019) because models within those categories lead to certain levels of enrichment of hit rates in drug discovery projects.

Baselines. As baselines we propose methods with the best performance in bioactivity prediction using microscopy images to date, reported in Hofmarcher et al. (2019). These consist of different convolutional neural network architectures, used in a fully supervised setting, and a method (”FNN”) that uses expert-designed cell features (Simm et al., 2018; Carpenter et al., 2006; Bray et al., 2017). The compared methods were trained in a multi-task setting to predict activity labels for 209 tasks, extracted from ChEMBL.

Method	Top-k accuracy (%)					
	Top-1	95% CI	Top-5	95% CI	Top-10	95% CI
CLOOME	**3.215**	[2.505, 4.058]	**6.998**	[5.947, 8.170]	**8.936**	[7.754, 10.233]
Random	0.047	[0.001, 0.263]	0.236	[0.077, 0.551]	0.473	[0.227, 0.868]

Table 2: Results for the bioisosteric replacement task. Given a molecule-perturbed microscopy image, the correct molecule must be selected from a set of candidates. Top-1, top-5 and top-10 accuracy in percentage are shown for a hold-out test set, along with the upper and lower limits for a 95% confidence interval on the proportion.

Results. The predictive performance on the downstream activity prediction tasks is reported in Table 1. CLOOME reached an average AUC of 0.714, which indicates that the learned representations are indeed transferable. CLOOME even outperformed fully supervised methods, such as M-CNN (Godinez et al., 2017) and SC-CNN (Hofmarcher et al., 2019), with respect to AUC.

Retrieval for bioisoteric replacement and scaffold hopping. In this experiment, we assessed the ability of CLOOME to correctly retrieve the correct molecular structure given a microscopy image of cells treated with this molecule. Notably, this is an extremely challenging task for human experts: given a microscopy image of cells, select the molecular structure with which they have been

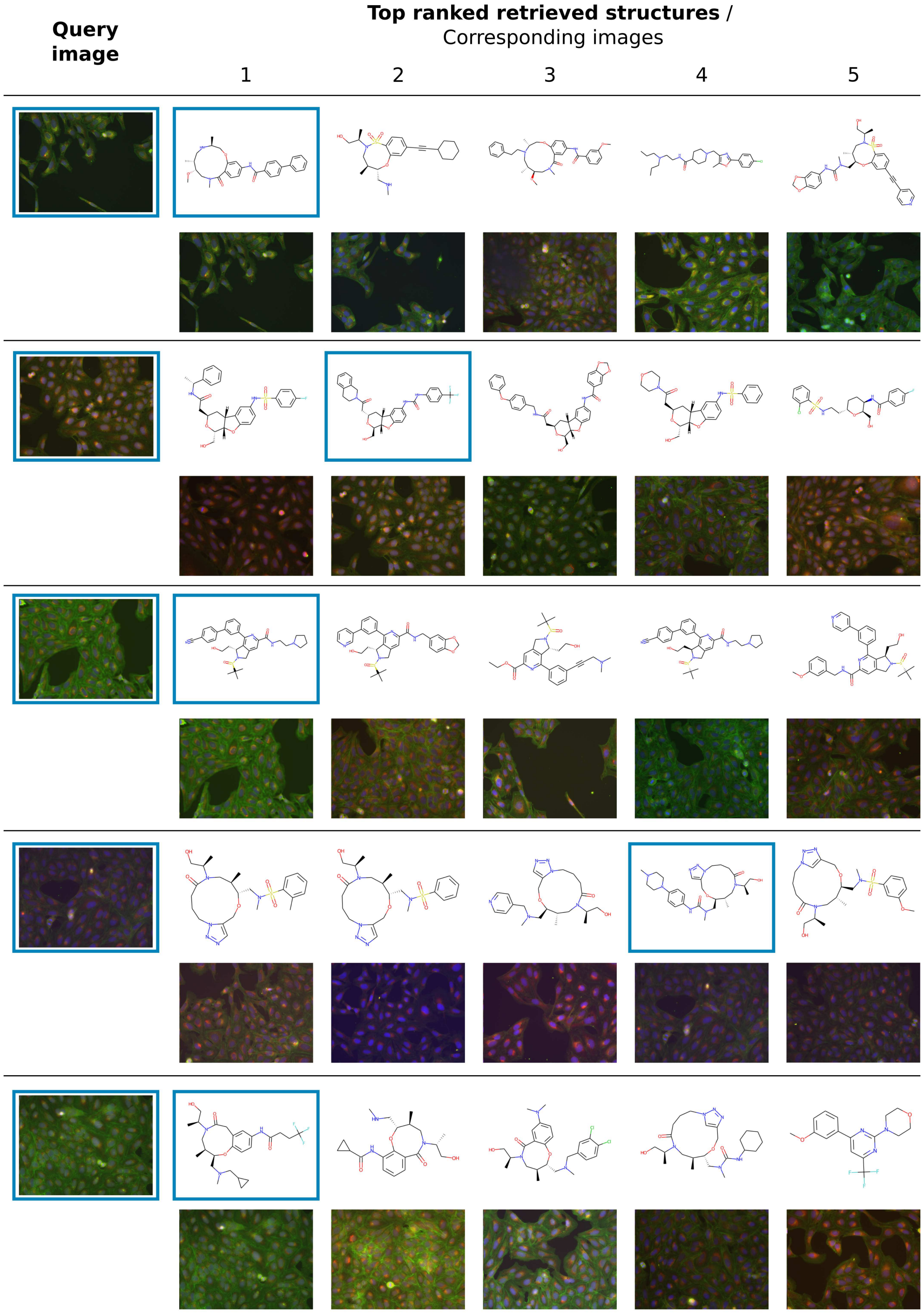

Figure 2: Example results for the retrieval task. On a hold-out test set, the five molecules for which representations are the most similar to the query image are shown along with their corresponding images. Blue boxes mark the query image and its matching molecular structure , i.e., the correct pair. CLOOME can be used to retrieve molecules that could produce similar biological effects on treated cells, i.e. bioisosteres.

treated from a set of thousand candidate structures. Since cells often do not exhibit any or only subtle morphological changes, this task is highly ambitious.

This image-based retrieval task can also be understood a bioisosteric replacement task (Lipinski, 1986): Bioisosteres are molecules with roughly the same biological properties or activities, which is highly relevant in drug discovery when a chemical scaffold should be replaced with another, but at the same time its biological activity should be kept. With this experiment, we evaluate the ability of CLOOME to correctly rank the matched molecular structure given the image. Other high-ranked structures could be potential bioisosteres, which makes this experiment a proxy for the bioisteric replacement problems (see Figure 1 **b)**).

On hold-out data of 2,115 image and molecule pairs, CLOOME ranked the correct molecule in the first place for 3.215% of the cases. A random method would achieve a value of $1/2,115 \approx 0.047\%$, which indicates a ~ 70-fold improvement of CLOOB. For this task, different hyperparameters and model were selected based on the appropriate validation metric (see Sec. A.2). The top-1, top-5, top-10 accuracy are given in Table 2. Further, some examples are displayed in Figure 2. This is, to our knowledge, the first system of cell-image-based retrieval of molecular structures.

4 Discussion and conclusion

We have introduced a contrastive learning method for learning representations of molecules based on microscopy images and chemical structures. On the largest available dataset of this type, we demonstrate that CLOOME is able to learn transferable representations of molecules. This opens the possibility to re-use the learned representations for activity or property prediction and for other tasks, such as finding bioisosteric replacements of molecules.

Limitations. Our method currently has several limitations. Our trained networks are restricted to a particular type of microscopy images, which are acquired with the Cell Painting protocol (Bray et al., 2016). This protocol has been published and currently there are community efforts (JUMP-CP Consortium, 2022) to increase the amount of available data. Large and more diverse datasets of molecule-perturbed cells or internal pharmaceutical company datasets will likely improve the learned representations, both image and structure encoder (Sturm et al., 2020). Due to the computational complexity, the hyperparameter and architecture space is currently under-explored such that we expect our method to further improve with better hyperparameters or encoder architectures. Furthermore, it has not escaped our notice that the learned structure encoder can also be used for transfer learning on molecular activities and properties. Also, it is worth noting that, although linear probing has been extensively used for the purpose of evaluating the quality of representations (Radford et al., 2021; Fürst et al., 2021), if the latter are very high dimensional, this method presents the risk of overfitting (Alain and Bengio, 2016). Having addressed these limitations, we nevertheless believe that the representations obtained with CLOOME could be highly useful for the community with respect to different drug discovery efforts.

Future work. For the reasons mentioned above, we intend to further analyze and test the representations produced by our method for different datasets and in different settings in order to better assess its capabilities and pitfalls.

Acknowledgements

ASF's research position was funded by the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie Innovative Training Network - European Industrial Doctorate grant agreement No. 956832, "Advanced machine learning for Innovative Drug Discovery". The ELLIS Unit Linz, the LIT AI Lab, the Institute for Machine Learning, are supported by the Federal State Upper Austria. IARAI is supported by Here Technologies. We thank the projects AI-MOTION (LIT-2018-6-YOU-212), AI-SNN (LIT-2018-6-YOU-214), DeepFlood (LIT-2019-8-YOU-213), Medical Cognitive Computing Center (MC3), INCONTROL-RL (FFG-881064), PRIMAL (FFG-873979), S3AI (FFG-872172), DL for GranularFlow (FFG-871302), AIRI FG 9-N (FWF-36284, FWF-36235), ELISE (H2020-ICT-2019-3 ID: 951847). We thank Audi.JKU Deep Learning Center, TGW LOGISTICS GROUP GMBH, Silicon Austria Labs (SAL), FILL Gesellschaft mbH, Anyline GmbH, Google, ZF Friedrichshafen AG, Robert Bosch GmbH, UCB Biopharma SRL, Merck

Healthcare KGaA, Verbund AG, Software Competence Center Hagenberg GmbH, TÜV Austria, Frauscher Sensonic and the NVIDIA Corporation.

REFERENCES

Alain, G. and Bengio, Y. (2016). Understanding intermediate layers using linear classifier probes. *ArXiv*, 1610.01644.

Alperstein, Z., Cherkasov, A., and Rolfe, J. T. (2019). All SMILES variational autoencoder. *ArXiv*, 1905.13343.

Bender, A. and Cortés-Ciriano, I. (2021). Artificial intelligence in drug discovery: what is realistic, what are illusions? part 1: ways to make an impact, and why we are not there yet. *Drug discovery today*, 26(2):511–524.

Bray, M.-A., Gustafsdottir, S. M., Rohban, M. H., Singh, S., Ljosa, V., Sokolnicki, K. L., Bittker, J. A., Bodycombe, N. E., Dančík, V., Hasaka, T. P., et al. (2017). A dataset of images and morphological profiles of 30 000 small-molecule treatments using the cell painting assay. *Gigascience*, 6(12):giw014.

Bray, M.-A., Singh, S., Han, H., Davis, C. T., Borgeson, B., Hartland, C., Kost-Alimova, M., Gustafsdottir, S. M., Gibson, C. C., and Carpenter, A. E. (2016). Cell Painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes. *Nature Protocols*, 11(9):1757–1774.

Cai, C., Wang, S., Xu, Y., Zhang, W., Tang, K., Ouyang, Q., Lai, L., and Pei, J. (2020). Transfer learning for drug discovery. *Journal of Medicinal Chemistry*, 63(16):8683–8694.

Caicedo, J. C., Cooper, S., Heigwer, F., Warchal, S., Qiu, P., Molnar, C., Vasilevich, A. S., Barry, J. D., Bansal, H. S., Kraus, O., et al. (2017). Data-analysis strategies for image-based cell profiling. *Nature Methods*, 14(9):849–863.

Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., and Joulin, A. (2020). Unsupervised learning of visual features by contrasting cluster assignments. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H., editors, *Advances in Neural Information Processing Systems 33*, pages 9912–9924. Curran Associates, Inc.

Carpenter, A. E., Jones, T. R., Lamprecht, M. R., Clarke, C., Kang, I. H., Friman, O., Guertin, D. A., Chang, J. H., Lindquist, R. A., Moffat, J., et al. (2006). Cellprofiler: image analysis software for identifying and quantifying cell phenotypes. *Genome biology*, 7(10):1–11.

Chandrasekaran, S. N., Ceulemans, H., Boyd, J. D., and Carpenter, A. E. (2021). Image-based profiling for drug discovery: due for a machine-learning upgrade? *Nature Reviews Drug Discovery*, 20(2):145–159.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In Daumé, H. and Singh, A., editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research (PMLR)*, pages 1597–1607.

Dahl, G. E., Jaitly, N., and Salakhutdinov, R. (2014). Multi-task neural networks for QSAR predictions. *ArXiv*, 1406.1231.

Devillers, B., Bielawski, R., Choski, B., and VanRullen, R. (2021). Does language help generalization in vision models? *ArXiv*, 2104.08313.

Fürst, A., Rumetshofer, E., Lehner, J., Tran, V., Tang, F., Ramsauer, H., Kreil, D. P., Kopp, M., Klambauer, G., Bitto-Nemling, A., and Hochreiter, S. (2021). CLOOB: modern Hopfield networks with InfoLOOB outperform CLIP. *ArXiv*, 2110.11316.

Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., and Dahl, G. E. (2017). Neural message passing for quantum chemistry. In *International Conference on Machine Learning*, pages 1263–1272. PMLR.

Godinez, W. J., Hossain, I., Lazic, S. E., Davies, J. W., and Zhang, X. (2017). A multi-scale convolutional neural network for phenotyping high-content cellular images. *Bioinformatics*, 33(13):2010–2019.

Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P. H., Buchatskaya, E., Doersch, C., Pires, B. Á., Guo, Z. D., Azar, M. G., Piot, B., Kavukcuoglu, K., Munos, R., and Valko, M. (2020). Bootstrap your own latent - a new approach to self-supervised learning. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 21271–21284. Curran Associates, Inc.

Gutmann, M. and Hyvärinen, A. (2010). Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In Teh, Y. W. and Titterington, M., editors, *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 297–304. JMLR Workshop and Conference Proceedings.

He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. B. (2020). Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Hofmarcher, M., Rumetshofer, E., Clevert, D.-A., Hochreiter, S., and Klambauer, G. (2019). Accurate prediction of biological assays with high-throughput microscopy images and convolutional networks. *Journal of Chemical Information and Modeling*, 59(3):1163–1171.

Ilse, M., Tomczak, J. M., and Welling, M. (2020). Deep multiple instance learning for digital histopathology. In *Handbook of Medical Image Computing and Computer Assisted Intervention*, pages 521–546. Elsevier.

Jaiswal, A., Babu, A. R., Zadeh, M. Z., Banerjee, D., and Makedon, F. (2020). A survey on contrastive self-supervised learning. *Technologies*, 9(1):2.

Jiang, D., Wu, Z., Hsieh, C.-Y., Chen, G., Liao, B., Wang, Z., Shen, C., Cao, D., Wu, J., and Hou, T. (2021). Could graph neural networks learn better molecular representation for drug discovery? a comparison study of descriptor-based and graph-based models. *Journal of Cheminformatics*, 13(1):1–23.

JUMP-CP Consortium (2022). Joint undertaking in morphological profiling `https://jump-cellpainting.broadinstitute.org/`. Accessed March 9,2022.

Kearnes, S., McCloskey, K., Berndl, M., Pande, V., and Riley, P. (2016). Molecular graph convolutions: moving beyond fingerprints. *Journal of Computer-Aided Molecular Design*, 30(8):595–608.

Kingma, D. P., Mohamed, S., Rezende, D. J., and Welling, M. (2014). Semi-supervised learning with deep generative models. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 27*, pages 3581–3589. Curran Associates, Inc.

Knight, A., Bailey, J., and Balcombe, J. (2006). Animal carcinogenicity studies: 3. alternatives to the bioassay. *Alternatives to Laboratory Animals*, 34(1):39–48.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In Pereira, F., Burges, C. J. C., Bottou, L., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc.

Lee, H., Ahn, S., Seo, S.-W., Song, Y. Y., Yang, E., Hwang, S.-J., and Shin, J. (2021). RetCL: A selection-based approach for retrosynthesis via contrastive learning. *ArXiv*, 2105.00795.

Lipinski, C. A. (1986). Bioisosterism in drug design. *Annual Reports in Medicinal Chemistry*, 21:283–291.

Loshchilov, I. and Hutter, F. (2017). SGDR: stochastic gradient descent with warm restarts. In *International Conference on Learning Representations (ICLR)*.

Loshchilov, I. and Hutter, F. (2019). Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*.

Luo, Z., Zou, Y., Hoffman, J., and Fei-Fei, L. F. (2017). Label efficient learning of transferable representations acrosss domains and tasks. In *Advances in Neural Information Processing Systems 30*.

Lusci, A., Pollastri, G., and Baldi, P. (2013). Deep architectures and deep learning in chemoinformatics: the prediction of aqueous solubility for drug-like molecules. *Journal of Chemical Information and Modeling*, 53(7):1563–1575.

MacArron, R., Banks, M. N., Bojanic, D., Burns, D. J., Cirovic, D. A., Garyantes, T., Green, D. V., Hertzberg, R. P., Janzen, W. P., Paslay, J. W., Schopfer, U., and Sittampalam, G. S. (2011). Impact of high-throughput screening in biomedical research. *Nature Reviews Drug Discovery*, 10(3):188–195.

Marcus, G. (2018). Deep learning: A critical appraisal. *ArXiv*, 1801.00631.

Mayr, A., Klambauer, G., Unterthiner, T., and Hochreiter, S. (2016). DeepTox: toxicity prediction using deep learning. *Frontiers in Environmental Science*, 3:80.

Mayr, A., Klambauer, G., Unterthiner, T., Steijaert, M., Wegner, J. K., Ceulemans, H., Clevert, D.-A., and Hochreiter, S. (2018). Large-scale comparison of machine learning methods for drug target prediction on ChEMBL. *Chemical Science*, 9(24):5441–5451.

Merkwirth, C. and Lengauer, T. (2005). Automatic generation of complementary descriptors with molecular graph networks. *Journal of Chemical Information and Modeling*, 45(5):1159–1168.

Morgan, H. L. (1964). The Generation of a Unique Machine Description for Chemical Structures - A Technique Developed at Chemical Abstracts Service. *Journal of Chemical Documentation*, 5(2):107–113.

Pearl, J. (1988). Embracing causality in default reasoning. *Artificial Intelligence*, 35(2):259–271.

Pepperkok, R. and Ellenberg, J. (2006). High-throughput fluorescence microscopy for systems biology. *Nature Reviews Molecular Cell Biology*, 7(9):690–696.

Poole, B., Ozair, S., van den Oord, A., Alemi, A. A., and Tucker, G. (2019). On variational bounds of mutual information. In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research (PMLR)*, pages 5171–5180.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*.

Ramsauer, H., Schäfl, B., Lehner, J., Seidl, P., Widrich, M., Gruber, L., Holzleitner, M., Pavlović, M., Sandve, G. K., Greiff, V., Kreil, D., Kopp, M., Klambauer, G., Brandstetter, J., and Hochreiter, S. (2021). Hopfield networks is all you need. In *International Conference on Learning Representations (ICLR)*.

Rogers, D. and Hahn, M. (2010). Extended-Connectivity Fingerprints. *Journal of Chemical Information and Modeling*, 50(5):742–754.

Scarselli, F., Gori, M., Tsoi, A. C., Hagenbuchner, M., and Monfardini, G. (2008). The graph neural network model. *IEEE Transactions on Neural Networks*, 20(1):61–80.

Seidl, P., Renz, P., Dyubankova, N., Neves, P., Verhoeven, J., Wegner, J. K., Segler, M., Hochreiter, S., and Klambauer, G. (2022). Improving few-and zero-shot reaction template prediction using modern hopfield networks. *Journal of Chemical Information and Modeling*.

Simm, J., Klambauer, G., Arany, A., Steijaert, M., Wegner, J. K., Gustin, E., Chupakhin, V., Chong, Y. T., Vialard, J., Buijnsters, P., et al. (2018). Repurposed high-throughput image assays enables biological activity prediction for drug discovery. *Cell Chemical Biology*, page 108399.

Stanley, M., Bronskill, J. F., Maziarz, K., Misztela, H., Lanini, J., Segler, M., Schneider, N., and Brockschmidt, M. (2021). FS-Mol: A few-shot learning dataset of molecules. In *35th Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Stärk, H., Beaini, D., Corso, G., Tossou, P., Dallago, C., Günnemann, S., and Liò, P. (2021). 3d infomax improves gnns for molecular property prediction. *ArXiv*, 2110.04126.

Sturm, N., Mayr, A., Le Van, T., Chupakhin, V., Ceulemans, H., Wegner, J., Golib-Dzib, J.-F., Jeliazkova, N., Vandriessche, Y., Böhm, S., et al. (2020). Industry-scale application and evaluation of deep learning for drug target prediction. *Journal of Cheminformatics*, 12(1):1–13.

Sun, C., Shrivastava, A., Singh, S., and Gupta, A. (2017). Revisiting Unreasonable Effectiveness of Data in Deep Learning Era. *ArXiv*, 1707.02968.

Unterthiner, T., Mayr, A., Klambauer, G., Steijaert, M., Wegner, J. K., Ceulemans, H., and Hochreiter, S. (2014). Deep learning as an opportunity in virtual screening. In *Workshop on Deep Learning and Representation Learning at Conference of the Neural Information Processing Systems Foundation (NIPS 2014)*.

Vall, A., Hochreiter, S., and Klambauer, G. (2021). BioassayCLR: Prediction of biological activity for novel bioassays based on rich textual descriptions. *ELLIS ML4Molecules workshop*.

van den Oord, A., Li, Y., and Vinyals, O. (2018). Representation learning with contrastive predictive coding. *ArXiv*, 1807.03748.

Wang, Y., Wang, J., Cao, Z., and Barati Farimani, A. (2022). Molecular contrastive learning of representations via graph neural networks. *Nature Machine Intelligence*, pages 1–9.

Wellman, M. P. and Henrion, M. (1993). Explaining 'explaining away'. *IEEE Trans. Pattern Anal. Mach. Intell.*, 15(3):287–292.

Wortsman, M., Ilharco, G., Li, M., Kim, J. W., Hajishirzi, H., Farhadi, A., Namkoong, H., and Schmidt, L. (2021). Robust fine-tuning of zero-shot models. *ArXiv*, 2109.01903.

Xu, K., Hu, W., Leskovec, J., and Jegelka, S. (2018). How powerful are graph neural networks? In *International Conference on Learning Representations (ICLR)*.

Yeh, C.-H., Hong, C.-Y., Hsu, Y.-C., Liu, T.-L., Chen, Y., and LeCun, Y. (2021). Decoupled contrastive learning. *ArXiv*, 2110.06848.

Zhang, C., Zhang, K., Pham, T. X., Niu, A., Qiao, Z., Yoo, C. D., and Kweon, I. S. (2022). Dual temperature helps contrastive learning without many negative samples: Towards understanding and simplifying moco. *arXiv preprint arXiv:2203.17248*.

A Appendix

A.1 Notation overview

Definition	Symbol/Notation	Dimension
molecule-perturbed microscopy image	x	image dimension, e.g. $320 \times 320 \times 5$
chemical structure of molecule	z	symbolic, e.g. graph
image embedding	$\boldsymbol{x}$	d
structure embedding	$\boldsymbol{z}$	d
stacked image embeddings	$\boldsymbol{X}$	$d \times N$
stacked structure embeddings	$\boldsymbol{Z}$	$d \times N$
stored image embeddings	$\boldsymbol{U}$	$d \times N$
stored structure embeddings	$\boldsymbol{V}$	$d \times N$
image-retrieved image embedding	$\boldsymbol{U}_{\boldsymbol{x}_i}$	d
structure-retrieved image embedding	$\boldsymbol{U}_{\boldsymbol{z}_i}$	d
image-retrieved structure embedding	$\boldsymbol{V}_{\boldsymbol{x}_i}$	d
structure-retrieved structure embedding	$\boldsymbol{V}_{\boldsymbol{z}_i}$	d
microscopy image encoder	$\boldsymbol{h}^x(.)$	$\mathbb{R}^{320\times320\times5} \rightarrow d$
molecule structure encoder	$\boldsymbol{h}^z(.)$	$\mathcal{M} \rightarrow d$
temperature parameter of the loss functions	τ	
scaling parameter of Hopfield net	β	
embedding dimension	d	
batch or dataset size	N	
chemical space	$\mathcal{M}$	
indices	i, j, n	

Table A1: Symbols and notations used in this paper.

A.2 Hyperparameter search

	Hyperparameter	Explored space
Learning	Optimizer	{AdamW}
	Learning rate	{0.0005, **0.001**, 0.005}
	Scheduler	{Cosine annealing with restarts}
	Weight decay	{0.1}
	Batch size	{**256**, 512}
	Warm-up iterations	{10000, **20000**}
	Inverse temperature	{**30**}
Image encoder	Image resolution	{**320**, 520}
	Model	{ResNet50}
Structure encoder	Number of layers	{4}
	Layer dimension	{1024}
	Activation	{ReLU}
	Batch normalization	{False, **True**}
Hopfield layers	β	{8, 14.3, **22**}
Embedding space	Number of dimensions	{**512**}

Table A2: Considered hyperparameter space of CLOOME models. The selected configurations for downstream activity prediction based on manual search on validation set shown in bold.