CLOOME: contrastive learning unlocks bioimaging databases for queries with chemical structures

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

Currently, bioimaging databases cannot be queried by chemical structures that induce the phenotypic effects captured by an image. Through the advent of the contrastive learning paradigm, images and text could be embedded into the same space. We build on this contrastive learning paradigm, to present a novel retrieval system that is able to identify the correct bioimage given a chemical structure out of a database of ∼2,000 candidate images with a top-1 accuracy >70 times higher than a random baseline. Additionally, the learned embeddings of our method are highly transferable to various relevant downstream tasks in drug discovery, including activity prediction, microscopy image classification and mechanism of action identification.

1 Introduction

Biological and chemical databases and their querying mechanisms are at the heart of research in molecular biology. Sequence databases, such as RefSeq [1] or UniProt [2], contain DNA or protein sequences, and are often queried with a given sequence using BLAST [3] or its variants. Genome databases [4] usually allow for multiple types of querying mechanisms, such as genetic location, gene names, or accession numbers. Protein structure databases, for example, the Protein Data Bank (PDB) [5], offer a range of querying mechanisms from sequence similarities to structural queries based on 3D shape. The chemical databases ChEMBL [6] and PubChem [7] are huge corpora of chemical structures that contain billions of small molecules. The International Chemical Identifier (InChI) [8] was designed to facilitate searching for chemical structures in such databases, which is difficult because of the graph matching problem. While BLAST, the structural search in PDB, and the InChI-based queries can be considered as associative or content-based querying, bioimaging databases still rely on manual annotation and text-based search. However, querying large bioimaging databases with a chemical structure that induces the phenotypic effect captured by the image could considerably empower biomedical research. Additionally, unlocking chemical databases for queries with a microscopy image capturing the phenotypic effects of a chemical structure could be equally important. (see Figure 1A,B).

Recently, contrastive learning has emerged as a powerful paradigm to learn rich representations from data [9]. The contrastive learning methods CLIP and CLOOB embed natural language and images into the same representation space [10, 11]. Contrastive learning enforces that images and their matched text captions are close to one another in this embedding space, while un-matched images and captions are separated. Therefore, text prompts can query an image database by extracting nearby images in the embedding space and vice versa [10].These text-image embedding spaces enabled the generation of realistic images from short text prompts and led to the recent emergence of "AI art" [12]. In this work, we use these powerful contrastive learning paradigms to enable querying or retrieval systems for bioimaging databases.

Microscopy imaging has been used as an informative and time- and cost-efficient biotechnology to characterize cell phenotypes, tissues, or cellular processes [13, 14]. Consequently, there have been substantial efforts by the scientific community to use high-throughput microscopy imaging [15] as informative read-out and characterization of cellular systems and cell phenotypes under diverse perturbations [14, 16]. For instance, features computed from microscopy images [17] have already been proved useful to find sets of compounds with higher diversity performance in high throughput screening assays than those from chemical structures [18]. In addition to the wealth of information that is comprehensible and informative for human experts, these microscopy images also contain large amounts of biological information inaccessible to humans, but which can be successfully extracted by computational methods, such as Deep Learning [19]. The immense amount of microscopy images are stored in large bioimaging databases, many of which are publicly available. Their querying mechanisms, however, are still limited to queries by textual annotations. A common embedding space of (a) microscopy images capturing phenotypic effects of perturbations, and (b) chemical structures inducing those effects would allow for associative

or content-based querying of both imaging and chemical databases. Such an embedding space would represent cellular processes both in terms of the chemical structures that induce them and in terms of images that capture the cell phenotypes caused by these processes. New applications such as the detection of novel cell phenotypes are possible through such embedding spaces (see Figure 5C,F).

CLIP and CLOOB models have been constructed via contrastive learning on large image-text datasets [10]. Analogously to these image-text datasets, the Cell Painting dataset [16] contains image-structure pairs (see Figure 1D). Therefore, we were able to use contrastive learning to jointly train a microscopy image encoder and a chemical structure encoder to construct a common embedding space of microscopy images capturing cell phenotypes and chemical structures representing the perturbations. We propose a contrastive learning framework for image-structure pairs that we call CLOOME (see Figure 1D, E, F). Within the CLOOME framework, a microscopy image encoder and a chemical structure encoder are learned by contrasting representations of matched image-structure pairs against un-matched examples from other pairs. Because our framework extends the contrastive learning methods CLIP [10] and CLOOB [11] to image-structure pairs, we call it Contrastive Learning and leave-One-Out-boost for Molecule Encoders (CLOOME).

We introduce CLOOME as a novel tool that enables querying large chemical databases with microscopy images, and vice versa. Moreover, the experiments conducted in this study show that our method is useful beyond this cross-modal querying, because the encoders provide rich and transferable embeddings of bioimages. In many cases these embeddings are more informative than high content imaging platforms [17], for example, for bioactivity prediction and mechanism of action prediction (see Section 3). In summary, we illustrate four use case scenarios that we envision for CLOOME.

1. A query molecule is known to cause an interesting biological effect, and the goal is to find other chemical structures that show the same biological activity. If a microscopy image of this query molecule is available, CLOOME could be used to retrieve chemical structures from large chemical databases. These chemical structures are candidates to exhibit the same biological activity and to show an equivalent phenotype as the one shown in the query molecules. Note that this type of retrieval does not need any text or annotation, which is completely novel (see structure retrieval results in Section 3.1).
Moreover, this application could also be of interest for biological effects, for which no molecules are currently known. Assume that the biological effect concerns the inhibition of a certain protein. A gene knock-out procedure could be carried out for this protein (e.g., using CRISPR-Cas9) and the resulting modified cells, captured with microscopy. Using this microscopy image as a query, CLOOME could be used to search for chemical structures that induce the inhibition of this specific target. In contrast to target-based virtual screening, this procedure involves the full complexity of a biological system.
2. For a new chemical structure it is unknown which cellular process it induces. The molecule's chemical structure is used to query a database of microscopy images of cells treated with different chemical compounds that have well-known biological effects. CLOOME could be used to retrieve images showing the potential phenotype of the query molecule, which could give an indication about the induced cellular process. This type of querying mechanism is also novel (see image retrieval results in Section 3.1).
3. CLOOME could also be used to predict the activity of a compound in a given assay. Even though this task has been previously addressed by classical supervised learning methods [19, 20], the training process was compute-intensive and slow. CLOOME's pre-training reduces this task to fitting a logistic regression model, which is compute-efficient and fast. After making predictions with CLOOME for a set of molecules, the results could be used to guide the initial screening campaigns in the drug discovery process (see Section 3.2).
4. As in the second case scenario, we assume that the goal is to find the cellular process that is induced by a certain molecule. However, we use image-image relations instead of structure-image relations. We use the image treated with the molecule of interest to query an image database whose images capture well-studied mechanism of actions (MoA). Then, CLOOME retrieves the most similar images from that bioimaging database, and those are labelled by MoAs, such that the MoA of the molecule of interest can be inferred (see Section 3.4).

In the following sections, we:
a) present a new contrastive learning approach for learning rich and expressive representations of microscopy images and chemical structures without the need for human annotation or bioactivity data;
b) we demonstrate that our framework yields a retrieval system which allows to query bioimaging databases by chemical structures linked by the underlying captured or induced cellular process, and vice versa; and
c) show that the learned representations are highly transferable to several relevant downstream tasks in drug discovery, such as activity prediction, microscopy image classification, and mechanism of action identification.

2 Methods

We propose contrastive learning of representations from pairs of microscopy images and chemical structures to obtain a common embedding space of these two modalities, a retrieval system and highly transferable encoders (see Figure 2). In contrast to previous approaches, in which chemical structure encoders learned representations

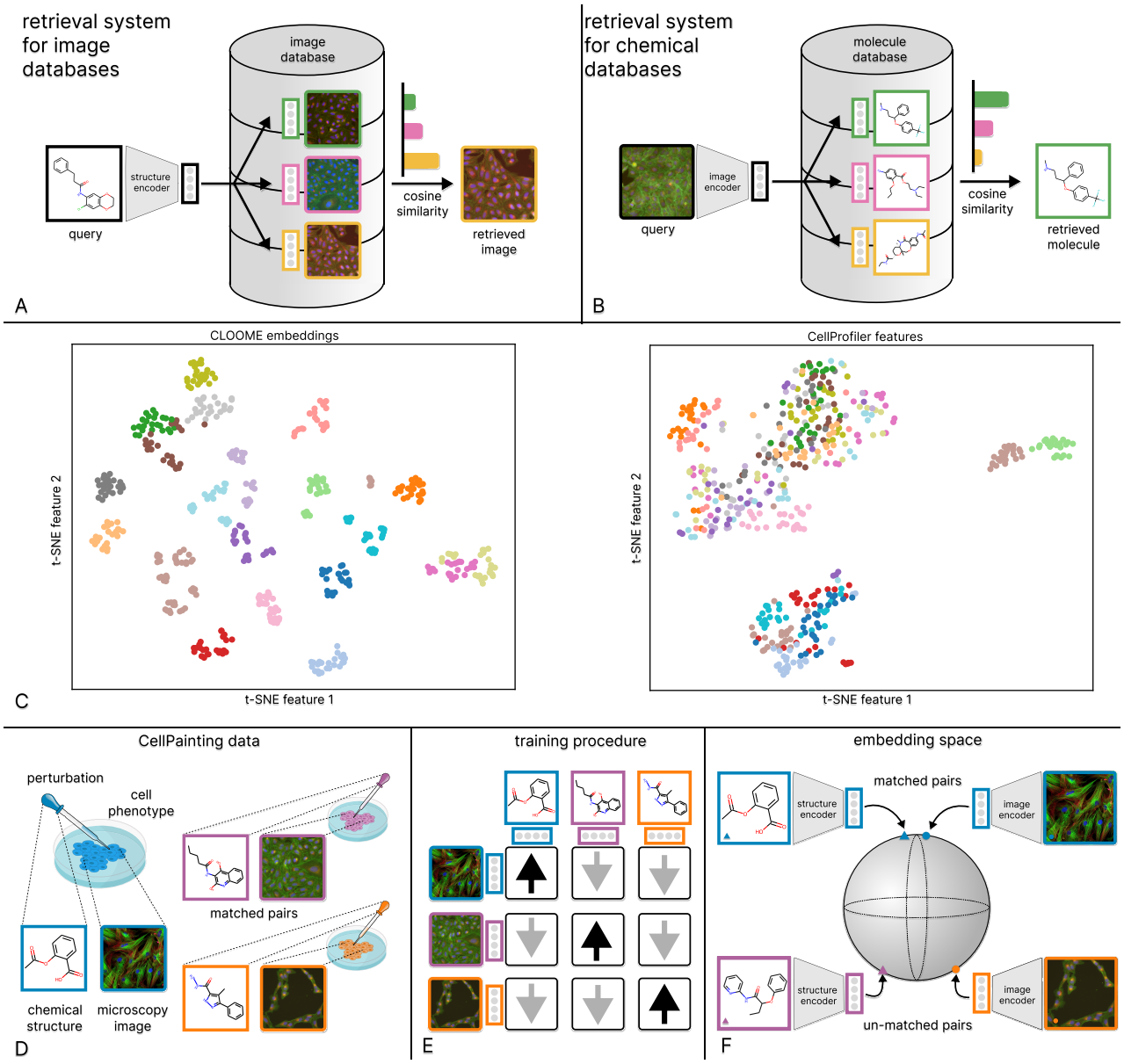

Figure 1: **A, B** The CLOOME encoders can be used to query a bioimaging database (**A**) by a chemical structure, and vice versa, query a chemical database by a microscopy image (**B**). **C** Visualization of the embedding space in terms of a t-SNE projection of image embeddings of new cell phenotypes. Each point represents a microscopy image from a hold-out set. The color indicates the cell phenotype, which was also withheld from training. The CLOOME embeddings (left) are indicative of the cell phenotype (clustered colors). CellProfiler features are less indicative of cell phenotypes (only a few colors cluster together). **D** A multi-modal setting for imaging cell phenotypes. Small molecules are administered to cells which are then imaged to capture potential phenotypic effects. In this way, matched image-structure pairs are obtained. **E** Schematic depiction of the training procedure of CLOOME. During training, the similarity of matched image-structure pairs is increased (black arrows), while the similarity of un-matched image-structure pairs is decreased (grey arrows). **F** The encoders of CLOOME map chemical structures and microscopy images to the same embedding space using a structure and a microscopy image encoder. Both encoders are deep neural networks. Matched pairs of chemical structures and microscopy images are mapped to embeddings that are close together, whereas un-matched pairs are mapped to embeddings that are separated.

using activity data [19, 20] or microscopy image encoders used hand-crafted representations [17, 21] or learned on manual annotations, CLOOME optimizes representations without activity data or human expertise.

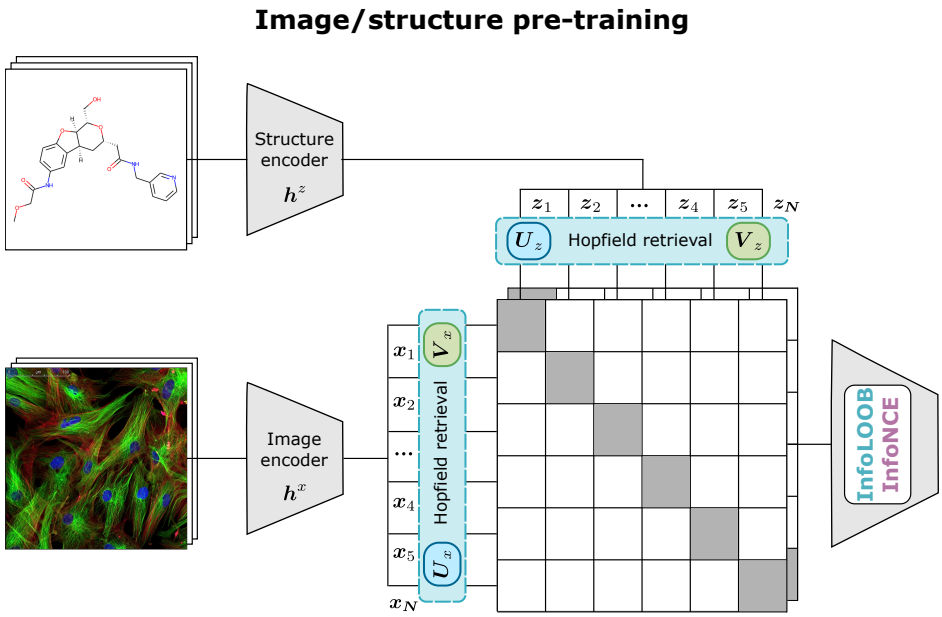

Figure 2: Schematic representation of CLOOME. Contrastive pre-training of embeddings of the two modalities, microscopy image and chemical structure, of a molecule using the CLOOB [11] approach.

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 microscopy 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 such that $\|\boldsymbol{x}_n\| = \|\boldsymbol{z}_n\| = 1 \; \forall n$ as in other contrastive learning approaches [22]. For notation, see also Supplementary Table A1.

In a contrastive learning setting, methods aim at increasing the similarity of matched pairs and decrease the similarity of un-matched pairs. This task has often been approached by maximizing the mutual information of the embeddings using the InfoNCE loss [9, 10, 22], which is also used in the CLIP approach [10]. 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 scales the cosine similarity of embedding pairs. This parameter can be regarded as a penalization of un-matched samples that are highly similar to the matched one. The lower this parameter is, the more the un-matched, but similar samples influence the loss. Conversely, as this value increases, all pairs will tend to contribute equally to the loss [23].

The contrastive learning method CLIP has the problem of "explaining away" [11, 24, 25]. 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 [11, 26, 27]. To ameliorate these drawbacks, CLOOB [11] has introduced the InfoLOOB objective together with continuous modern Hopfield networks [28] as a promising method for contrastive learning. Our contrastive learning framework CLOOME comprises both methods CLIP [10] and CLOOB [11].

For our extension of the CLOOB method, 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:

$$\boldsymbol{U}_{\boldsymbol{x}_i} = \boldsymbol{U}\ \mathrm{softmax}(\beta\ \boldsymbol{U}^T \boldsymbol{x}_i)\ , \quad (2) \qquad \boldsymbol{V}_{\boldsymbol{x}_i} = \boldsymbol{V}\ \mathrm{softmax}(\beta\ \boldsymbol{V}^T \boldsymbol{x}_i)\ , \quad (4)$$

$$\boldsymbol{U}_{\boldsymbol{z}_i} = \boldsymbol{U}\ \mathrm{softmax}(\beta\ \boldsymbol{U}^T \boldsymbol{z}_i)\ , \quad (3) \qquad \boldsymbol{V}_{\boldsymbol{z}_i} = \boldsymbol{V}\ \mathrm{softmax}(\beta\ \boldsymbol{V}^T \boldsymbol{z}_i)\ , \quad (5)$$

where β is a scaling parameter of the Hopfield network which is considered a hyperparameter. A β value of 0 corresponds to retrieving the average of the stored embeddings. Conversely, when β is set to a large value, the stored embeddings that are most similar to the query embeddings $\boldsymbol{x}_i$ and $\boldsymbol{z}_i$ are retrieved. The retrieved embeddings $\boldsymbol{U}_{\boldsymbol{x}_i}, \boldsymbol{U}_{\boldsymbol{z}_i}, \boldsymbol{V}_{\boldsymbol{x}_i}, \boldsymbol{V}_{\boldsymbol{z}_i}$ are normalized to unit norm. By default, we store the current minibatch in the continuous 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 supplementary Table A1. Then, the InfoLOOB objective [11, 29] for the retrieved embeddings is used as objective function:

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

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 [15, 20]. Although standard image encoders, such as Residual Networks [30] could be in principle used with minor adjustments, alternative approaches, such as multiple instance learning, could be required for very high resolution datasets [31].

Molecule structure encoder. Since the advent of Deep Learning, a large number of architectures to encode molecules have been suggested [32–36]. In contrast to computer vision and natural language processing, in which only few prominent architectures have emerged, there is yet no standard choice for chemical structure encoders. Because of their computational efficiency and good predictive performance, CLOOME uses a descriptor-based fully-connected network [37, 38] with 4 hidden layers of 1024 units with ReLU activations and batch normalization (for further details see supplementary Section B). However, also any graph [35, 39–41], message-passing [42], or sequence-based [43] neural network with an appropriate pooling operation can be used as structure encoder.

Dataset and preprocessing. *Cell Painting.* We use matched pairs of microscopy images and molecules from the Cell Painting [16, 21] dataset. This dataset is a collection of high-throughput fluorescence microscopy images of U2OS cells treated with different small molecules [16]. 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 disregarding erratic images (out of focus or containing high fluorescence material) as well as images of untreated cells that were used as controls, 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.* [20], which consisted of converting the original TIF images from 16-bit to 8-bit, simultaneously removing the 0.0028% of pixels with highest values. Moreover, the images were normalized using the mean and standard deviation calculated for the training split. Concerning molecules, two types of preprocessing yielded the best results. For the bioactivity prediction and zero-shot tasks, the SMILES strings were transformed to 1024-bit Morgan fingerprints with a radius of 3, taking chirality into account [44, 45]. For the retrieval task, a max-pooling combination of Morgan and RDKit count-based fingerprints, with a final length of 8192 bits, provided better results.

Data splits. We split our dataset into training, validation, and test set, using the splits of Hofmarcher *et al.* [20]. 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 [46] and CLOOB [11] wherever applicable, and tuned a few critical hyperparameters, such as learning rate and the β parameter of the Hopfield layer on the validation set. The architecture of the structure encoder was inspired by previous successful models [38] and was not subject to substantial hyperparameter optimization. We used the Adam optimizer [47] with decoupled weight decay regularization [48]. 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 [49]. 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 the retrieval and zero-shot image classification tasks, a 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 images were cropped to a pixel resolution of 520x520, based on performance in the validation set.

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 data augmentation and to allow large batch sizes, for the bioactivity prediction task, the images were cropped and re-scaled from the original 520x696 pixel resolution to 320x320 during training, whereby the original aspect ratio was mostly maintained.

Hence, different pre-training settings have been found to yield best results for bioactivity prediction and for both the retrieval and zero-shot image classification 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 supplementary Section B.

CellProfiler features calculation and preprocessing. The CellProfiler software [17] can also be considered as a microsopy image encoder that supplies image embeddings. These features were calculated using the latest CellProfiler pipelines provided in the CellPainting gallery [50]. The resulting embeddings consist of 1,240 features aggregated in one vector per image. Features were aggregated to allow comparability across all methods' results. After filtering out features with a standard deviation of zero across all samples as well as trivial features (e.g., file paths), these embeddings consist of 1,081 features. Before training, these features were standardized using the mean and standard deviation calculated for the embeddings corresponding to the training set.

3 Experiments and Results

First, we demonstrate the abilities of CLOOME as a retrieval system for bioimaging and chemical databases (see Figure 3A). Then, we use CLOOME's image embeddings to predict assay bioactivities by fitting merely a logistic regression model (see Figure 3B). Lastly, we evaluate the performance of CLOOME image embeddings in molecule and mechanism of action classification tasks. Figure 3 C and D depict the experimental setting for these tasks.

3.1 A retrieval system for imaging and chemical databases

In this experiment, we assessed the ability of CLOOME to correctly retrieve the matched chemical structure given a microscopy image of cells treated with this molecule. Notably, this is an extremely challenging task for human experts, even considered close-to-impossible: given a microscopy image of cells, the task is to select the chemical structure with which they have been treated from a set of thousands of candidate structures. Since cells often do not exhibit any or only subtle phenotypic changes, this task is highly ambitious.

This image-based retrieval task can also be understood as a bioisosteric replacement task [51]: bioisosteres are molecules with roughly the same biological properties or activities. Bioisoterism is highly relevant in drug discovery when a molecule should be replaced with another, but at the same time its biological activity should be kept. An extreme case of bioisosteric replacement is scaffold hopping, because molecules with a highly different chemical structures but still the same bioactivity should be found. With this experiment, we evaluate the ability of CLOOME to correctly rank the matched molecular structure given the corresponding image. Other high-ranked structures could be potential bioisosteres, which makes this experiment a proxy for the bioisteric replacement problems (see Figure 2 **b)**).

As a baseline, we included a bilinear model defined by $y = \boldsymbol{x}\boldsymbol{W}\boldsymbol{z}$, where $\boldsymbol{x}$ and $\boldsymbol{z}$ are one image and molecule feature vectors, respectively, and $\boldsymbol{W}$ is a trainable weight matrix. Analogously to the CLIP method, the InfoNCE objective was symmetrically computed between modalities. The explored hyperparameters for this model can be found in Table A6. Additionally, a random baseline is included to demonstrate the performance improvement of the different methods in comparison to selecting the correct image or molecule by chance.

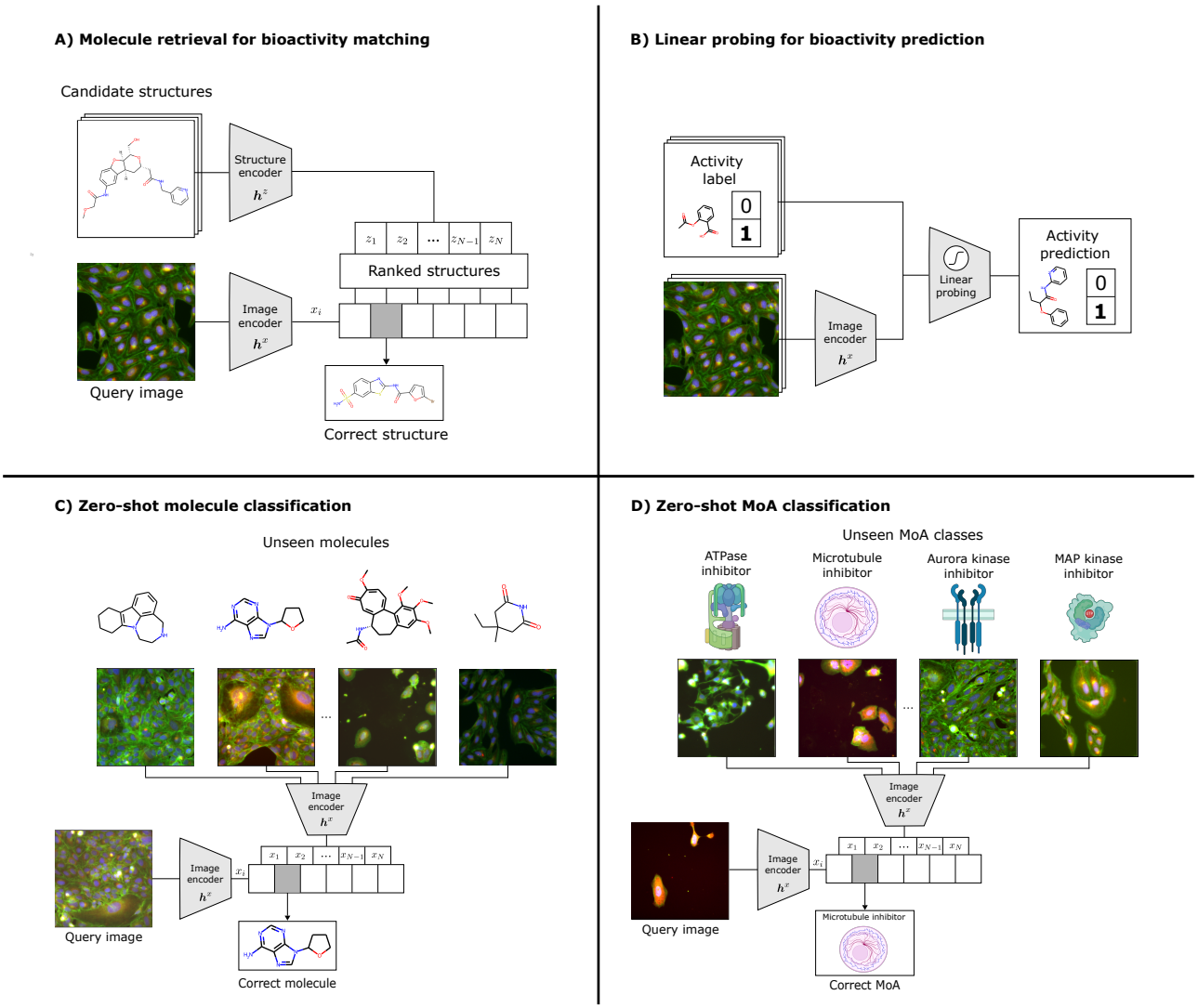

Figure 3: **A** Multi-modal retrieval task using CLOOME image and molecule embeddings. The resulting embeddings can be used to rank chemical structures that induce similar phenotypic effects, and vice versa. **B** Using the CLOOME embeddings for activity prediction. A logistic regression model is trained for activity prediction tasks. **C** Zero-shot classification task using CLOOME image embeddings for molecule prediction. A set of representative images, one for each molecule, are used to infer which compound was applied in a query image. **D** Zero-shot classification task using CLOOME image embeddings for mechanism of action prediction. In this case, the set of representative images depict each mechanism of action. A query image is classified into the most likely MoA category based on similarity with its corresponding representative image.

Split	Method	Top-k accuracy (%)					
		Top-1	95%-CI	Top-5	95%-CI	Top-10	95%-CI
Random	CLOOME (structure retr.)	**3.78**	[3.01, 4.69]	**7.94**	[6.83, 9.18]	**9.46**	[8.24, 10.8]
	CLOOME (image retr.)	**3.22**	[2.51, 4.06]	**8.42**	[7.27, 9.68]	**9.88**	[8.64, 11.2]
	Bilinear model (structure retr.)	0.473	[0.227, 0.868]	2.27	[1.68, 2.10]	3.22	[3.14, 4.84]
	Bilinear model (image retr.)	0.804	[0.469, 1.28]	3.64	[2.51, 4.06]	5.53	[3.95, 5.82]
	Random	0.0473	[0.0012, 0.263]	0.236	[0.0768, 0.551]	0.473	[0.227, 0.868]
Scaffold	CLOOME (structure retr.)	**2.79**	[1.99, 3.79]	**6.29**	[5.08, 7.70]	**7.58**	[6.25, 9.10]
	CLOOME (image retr.)	**2.50**	[1.75, 3.46]	**6.58**	[5.34, 8.01]	**8.08**	[6.71, 9.64]
	Bilinear model (structure retr.)	0.787	[0.393, 1.40]	2.72	[1.93, 3.71]	4.36	[3.35, 5.57]
	Bilinear model (image retr.)	0.930	[0.496, 1.58]	2.65	[1.87, 3.63]	4.51	[3.48, 5.73]
	Random	0.0715	[0.00181, 0.398]	0.357	[0.116, 0.833]	0.715	[0.344, 1.31]

Table 1: Results for the retrieval task **among 2,115** and **1,398** candidates, in the random and scaffold split, respectively. Given a molecule-perturbed microscopy image, the matched molecule must be selected from a set of candidates, and vice versa. 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.

Split	Method	Top-k accuracy (%)					
		Top-1	95%-CI	Top-5	95%-CI	Top-10	95%-CI
Random	CLOOME (structure retr.)	**11.3**	[9.94, 12.7]	**21.7**	[19.9, 23.5]	**31.7**	[29.7, 33.7]
	CLOOME (image retr.)	**10.7**	[9.45, 12.1]	**22.0**	[20.3, 23.9]	**31.0**	[29.0, 33.0]
	Bilinear model (structure retr.)	5.48	[4.55, 6.54]	15.4	[13.9, 17.0]	22.1	[20.4, 24.0]
	Bilinear model (image retr.)	6.48	[5.47, 7.61]	16.0	[14.5, 17.7]	22.6	[20.9, 24.5]
	Random	0.992	[0.616, 1.51]	5.01	[4.12, 6.03]	10.0	[8.78, 11.4]
Scaffold	CLOOME (structure retr.)	**8.08**	[6.71, 9.64]	**18.6**	[16.6, 20.7]	**26.8**	[24.5, 29.2]
	CLOOME (image retr.)	**8.37**	[7.00, 9.95]	**17.7**	[15.8, 19.8]	**25.9**	[23.6, 28.3]
	Bilinear model (structure retr.)	4.36	[3.36 5.57]	11.9	[10.3, 13.8]	18.3	[16.3, 20.4]
	Bilinear model (image retr.)	4.58	[3.54, 5.81]	13.2	[11.4, 15.0]	19.8	[17.8, 22.0]
	Random	1.00	[0.549, 0.167]	5.01	[3.92, 6.28]	10.0	[8.49, 11.7]

Table 2: Results for the retrieval task **among 100** candidates. Given a molecule-perturbed microscopy image, the matched molecule must be selected from a set of candidates, and vice versa. 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.

On hold-out data of 2,115 image and molecule pairs, CLOOME ranked the matched molecule in the first place for 3% of the cases. A random method would achieve a value of $1/2,115 \approx 0.047\%$, which indicates a $\sim$ 70-fold improvement of CLOOME. For this task, different hyperparameters and model were selected based on the appropriate validation metric (see supplementary Section B). The top-1, top-5, top-10 accuracy are given in Table 1 for retrieving from a database of 2,115 instances. Additionally, we report the same metrics for a sampling rate of 1%, or equivalently, 1 matched example together with 99 un-matched ones – a setting often used to evaluate retrieval systems, see Table 2. Further, some examples are displayed in Figure 4. This is, to our knowledge, the first system of cell-image-based retrieval of molecular structures.

3.2 Bio-activity prediction as downstream task

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 [22, 52] on downstream tasks is often performed for contrastive learning approaches to check the transferability of learned features. In such experiments, the representations of the pre-trained 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 [22], 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.* [20]. 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.

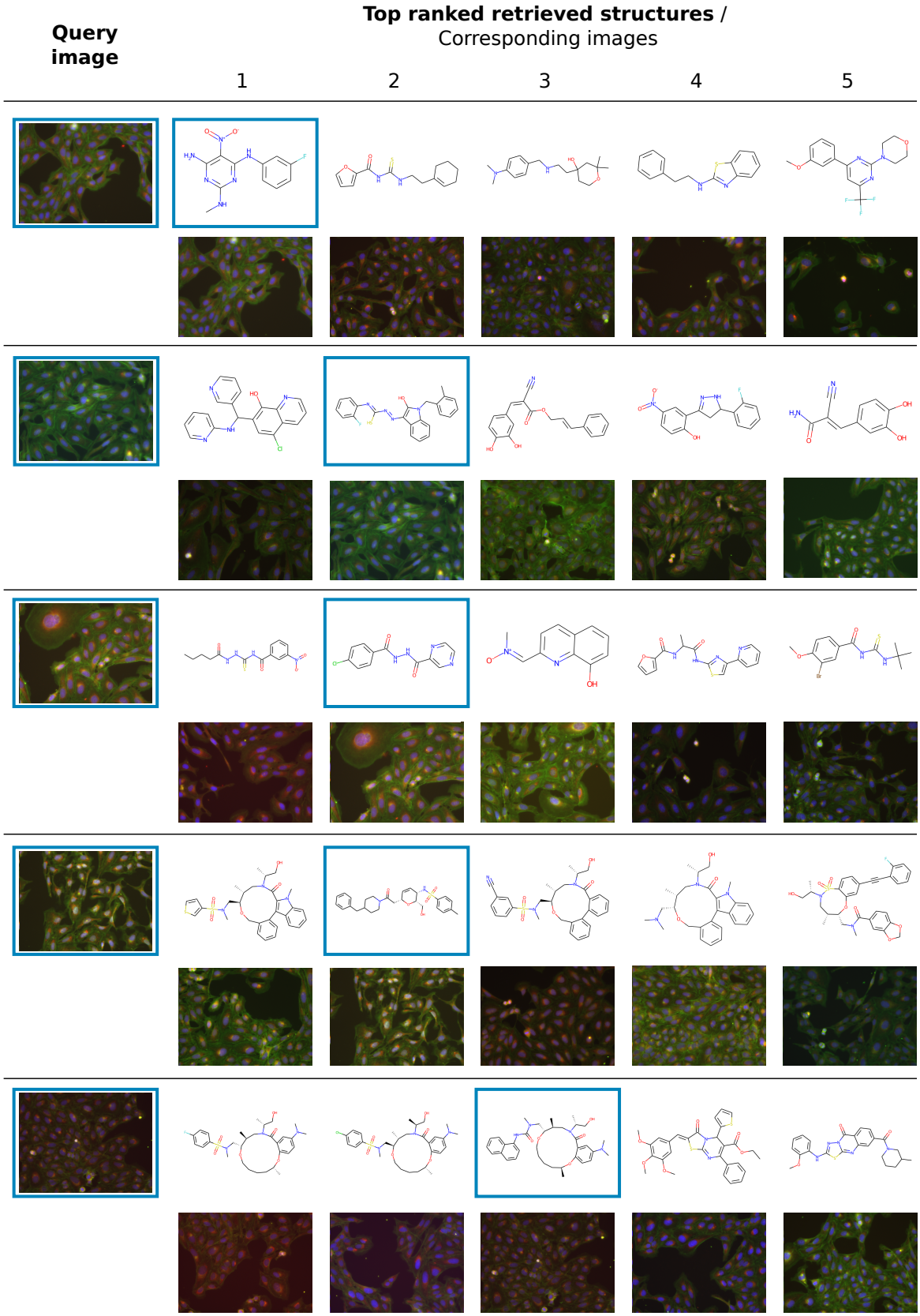

Figure 4: Example results for the retrieval task. On a hold-out test set, the five molecules for which representations are the most similar to the query image are shown along with their corresponding images. Blue boxes mark the query image and its matched molecular structure, i.e. the matched pair. CLOOME can be used to retrieve molecules that could produce similar biological effects on treated cells, i.e. bioisosteres.

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 [19, 20], 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 [19, 20] because models within those categories lead to certain levels of enrichment of hit rates in drug discovery projects.

Baselines. As baselines we consider methods reported in Hofmarcher *et al.* [20]. They are the best performing methods for bioactivity prediction using microscopy images to date and consist of different convolutional neural network architectures, used in a fully supervised setting, and a method ("FNN") that uses expert-designed cell features [17, 19, 21]. The compared methods were trained in a multi-task setting to predict activity labels for 209 tasks, extracted from ChEMBL.

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 3: Comparison of the linear probing evaluation of the learned representations against fully supervised methods [20]. 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 [22]. The best method in each category is marked bold.

Results. The predictive performance on the downstream activity prediction tasks is reported in Table 3. CLOOME reached an average AUC of 0.714 ± 0.20 across prediction tasks, which indicates that the learned representations are indeed transferable since no activity data had been used to train the CLOOME encoders. CLOOME even outperformed fully supervised methods, such as M-CNN [53] and SC-CNN [20], with respect to AUC.

3.3 Zero-shot molecule classification

The goal of this analysis is to evaluate the potential of image embeddings of CLOOME to distinguish the specific applied molecule. Note that, instead of using both molecule and image embeddings, as in the retrieval task, only image embeddings were used in this case. However, this task is not intended to be considered a real use-case scenario, but more of a proxy task for technical validation. This evaluation can also be regarded as a quantitative evaluation of how closely clustered are image embeddings of cells treated with the same molecule.

A zero-shot classification setting is considered, which means that the test set contains new, i.e. "unseen", image classes that had not been included in the training set. Concretely, one image for each of the molecules in a hold- out test set was randomly selected, which means that each unseen molecule class is represented by a single image. Then, samples from this set as well as samples corresponding to both the same molecule and plate were removed from the full test set, in order to ensure that the classification was not due to plate effects. We will refer to the remaining samples as the "test set". Finally, the set of unseen classes consisted of 2,115 images in the random split and 1,398 in the scaffold split. Overall, the test set comprised 44,102 for the random split, and 28,248 for the scaffold split.

To obtain the embeddings, all samples were pushed through the CLOOME image encoder and normalized. Then, cosine similarity was computed between embeddings from the former and the latter, and the softmax function was applied. The output from the softmax function was then used to calculate the metrics shown in Table. 4

We compared the image embeddings of CLOOME to embeddings of a microscopy image encoder trained in a supervised fashion and to CellProfiler features computed as detailed in Section 2. Regarding the image

embeddings of GapNet, the images were encoded using the model weights provided by Hofmarcher *et al.* [20], removing the last layer of its classifier, which resulted in a 1024-dimension embedding space.

As shown in Table 4, CLOOME exhibits higher performance in top-1, top-5 and top-10 accuracy metrics in comparison to both GapNet and CellProfiler-extracted features.

Split	Method	Accuracy[%]					
		Top-1	95% CI	Top-5	95% CI	Top-10	95% CI
Random	CLOOME	**18.4**	[18.1, 18.8]	**41.5**	[41.1, 42.0]	**56.2**	[55.8, 56.7]
	GapNet	0.36	[0.31, 0.42]	1.07	[0.973, 1.17]	1.78	[1.66, 1.91]
	CellProfiler	2.68	[2.53, 2.84]	8.24	[7.99, 8.51]	12.6	[12.3, 12.9]
Scaffold	CLOOME	**22.3**	[21.9 22.9]	**50.8**	[50.2, 51.4]	**67.1**	[66.6, 67.7]
	GapNet	0.517	[0.437, 0.608]	1.62	[1.47, 1.77]	2.56	[2.38, 2.75]
	CellProfiler	3.64	[3.42, 3.86]	10.4	[10.0, 10.7]	15.6	[15.1, 16.0]

Table 4: Results for the zero-shot molecule classification. Given a molecule-perturbed microscopy image, the image corresponding to the matched 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 resulting proportion.

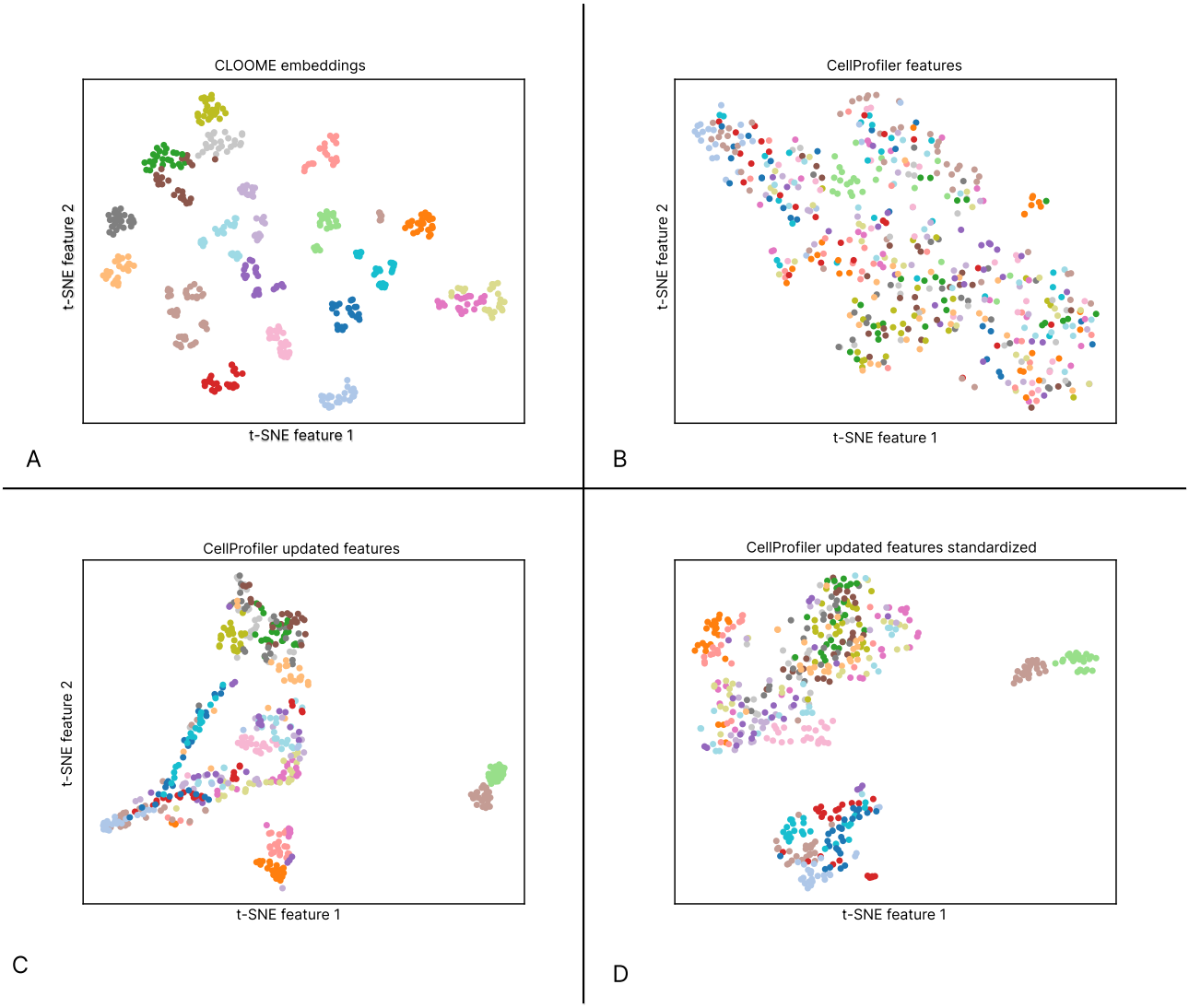

Figure 5: t-SNE downprojection of the CLOOME learned embeddings (**A**), the CellProfiler extracted features provided in [21] (**B**), and the CellProfiler extracted features using an updated pipeline, non standardized and standardized (**C** and **D**, respectively) of all microscopy images corresponding to 20 randomly selected molecules from the test set. The colors represent different molecules.

3.4 Zero-shot mechanism of action (MoA) classification

We now apply CLOOME to a challenging zero-shot MoA-classification task. This task assesses how well distinguished these embeddings are according to their mechanism of action. In order to evaluate this, labels were taken from the Drug Repurposing Hub [54] of molecules that are present in our validation and test sets.

The same procedure as in the zero-shot molecule classification task (see Subsection above) was followed, with the difference that the set of unseen classes consists of one image per mechanism of action. Again, images corresponding to the same molecule and plate as those from the representative class set were removed in order to ensure that the classification was not influenced by plate effects.

For the random split, the representative set consisted of 126 MoAs and 202 molecules. For the scaffold, this set contained 68 MoAs and 93 molecules. Regarding the test set, it has 8,826 samples for the random split and 4,056 for the scaffold. We employed the same baselines as in the zero-shot molecule classification task.

Results displayed in Table 5 show that CLOOME presents better performance than GapNet and CellProfiler features in mechanism of action prediction.

Split	Method	Accuracy[%]					
		Top-1	95% CI	Top-5	95% CI	Top-10	95% CI
Random	CLOOME	**13.0**	[12.3, 13.7]	**35.3**	[34.3, 36.3]	**48.0**	[46.9, 49.0]
	GapNet	1.09	[0.882, 1.33]	4.68	[4.25, 5.14]	8.63	[8.06, 9.24]
	CellProfiler	2.87	[2.53, 3.24]	9.46	[8.86, 10.1]	14.9	[14.2 15.7]
Scaffold	CLOOME	**24.3**	[23.1 25.6]	**51.9**	[50.4, 53.4]	**64.3**	[62.8, 65.7]
	GapNet	1.92	[1.52, 2.39]	8.00	[8.14, 9.92]	17.2	[16.1, 18.4]
	CellProfiler	4.54	[3.92, 5.22]	15.3	[14.2, 16.4]	24.5	[23.2, 25.8]

Table 5: Results for the zero-shot MoA classification. Given a molecule-perturbed microscopy image, the image corresponding to the matched 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 resulting proportion.

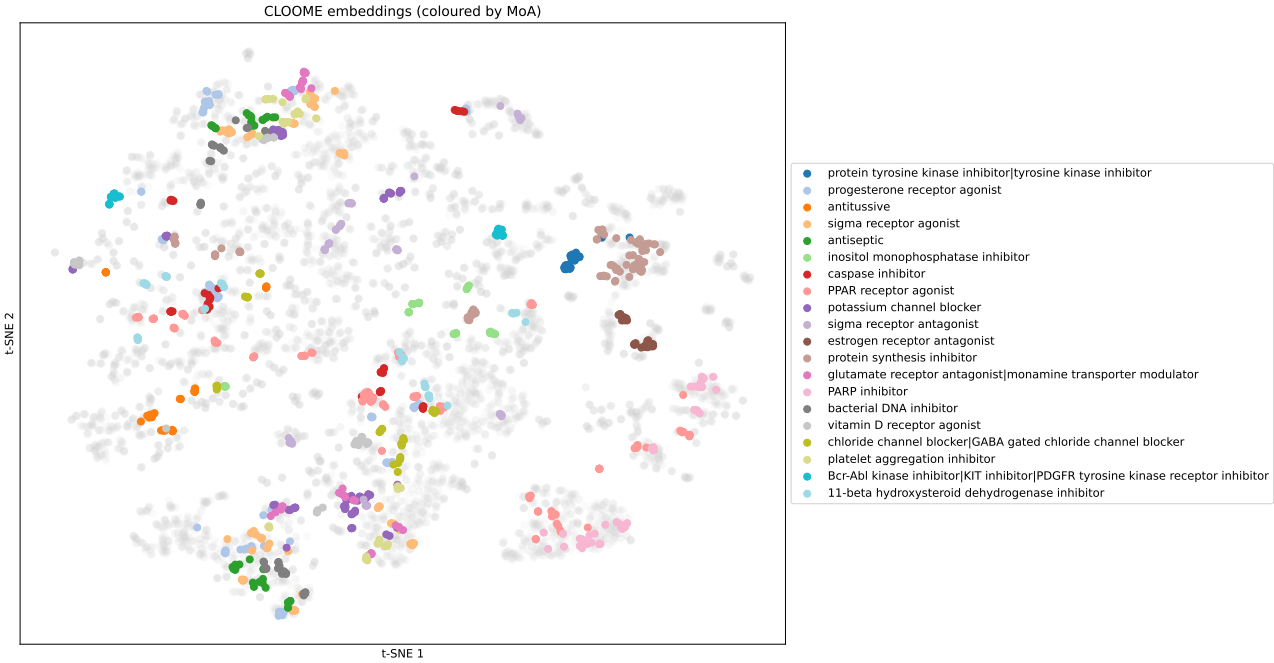

Figure 6: t-SNE downprojection of the learned CLOOME image embeddings. Samples corresponding to 20 randomly selected MoAs are shown in different colors, while the remaining samples are shown in gray. Embeddings of images capturing the same MoAs the are close together in the t-SNE plot, which indicates that the CLOOME embeddings are highly useful for identifying new MoAs.

3.5 Statistics and reproducibility

All confidence intervals reported in this study were calculated using the Clopper-Pearson interval. To compute

these intervals, the number of test samples in each task were used and have been reported in their specific subsections in Section 3.

4 Related work

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 [30, 55]. These supervised methods require large amounts of labeled data, which may be very costly or unfeasible to obtain, and they have limited generalization abilities [56, 57]. 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 [22, 58]. With contrastive learning methods [59] and self-supervision these meaningful representations can be obtained without the need for large amounts of expensive manually-provided labels [22, 60–62]. While uni-modal methods typically use pre-text tasks [22], for multi-modal methods the self-supervision arises from the availability of two modalities of an instance, such as image and text [10, 63]. Both uni-modal and multi-modal contrastive learning methods have recently had a substantial impact in computer vision and natural language processing [64].

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) [10], 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 [11, 24, 25] (details in Section 2). Considering this caveat, the "Contrastive Leave One Out Boost" (CLOOB) method has been proposed [11]. CLOOB uses a different objective, the "InfoLOOB" (LOOB for "Leave One Out Bound") objective [29], which does not include the positive pair in the denominator to avoid saturation effects [11]. Moreover, continuous modern Hopfield networks [65] 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.

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 [66, 67]. Therefore, methods that can learn transferable representations from unlabelled data are highly demanded. Thus, several contrastive learning approaches have been recently developed for different tasks in drug discovery. MolCLR [68] uses contrastive molecule-to-molecule training by augmenting molecular graphs. Stärk *et al.* [69] contrastively learn 3D and 2D molecule representations to inform the learned molecule encoder with 3D information. Lee *et al.* [70] and Seidl *et al.* [71] use contrastive learning for molecules and chemical reactions, and Vall *et al.* [72] 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 [16] 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 phenotypic changes it induces to a cell, is considered promising for accelerating drug discovery [16, 19, 73, 74]. 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 [75]. Particularly, microscopy image-based profiles of small molecules have been suggested to be effective together with deep learning methods [74]. However, the current efforts are still in standard supervised learning settings based on extracted features [19] or deep architectures [20]. 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 [76]. Instead of the currently used activity measurements as labels [19, 20], 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 [77, 78].

5 Discussion and conclusion

We have introduced a contrastive learning method for learning representations of microscopy images and chemical structures. On the largest available dataset of this type, we demonstrate that the encoders of CLOOME can be used as a powerful cross-modal retrieval system between chemical structures and bioimages. Additionally, we demonstrated that the CLOOME embeddings are both rich and transferable representations. This opens the possibility to re-use the learned representations for activity or property prediction and for other tasks, such as retrieval tasks from microscopy image or chemical databases, which we demonstrate in a series of challenging downstream tasks.

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 [16]. This protocol has been published and currently there are community efforts [76] 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 [79]. 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 could 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 [10, 11], if the latter are very high dimensional, this method presents the risk of overfitting [52]. Having addressed these limitations, we nevertheless believe that both the retrieval system and representations obtained with CLOOME could be highly useful for both the community using bioimaging as well as for drug discovery.

Data availability

The model weights are publicly available in the HuggingFace repository: `https://huggingface.co/anasanchezf/cloome/tree/main`. The dataset for training is available at: `http://gigadb.org/dataset/100351`. A script with the functions used to preprocess data as in this study can be found at: `https://github.com/ml-jku/cloome/blob/main/src/preprocess/preprocess_image.py`. Code is available at: `https://github.com/ml-jku/cloome`.

References

1. O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research* **44,** D733–D745 (2016).
2. UniProt Consortium. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Research* **47,** D506–D515 (2019).
3. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *Journal of Molecular Biology* **215,** 403–410 (1990).
4. Karolchik, D., Baertsch, R., Diekhans, M., Furey, T. S., Hinrichs, A., Lu, Y., Roskin, K. M., Schwartz, M., Sugnet, C. W., Thomas, D. J., *et al.* The UCSC genome browser database. *Nucleic Acids Research* **31,** 51–54 (2003).
5. Burley, S. K., Berman, H. M., Kleywegt, G. J., Markley, J. L., Nakamura, H. & Velankar, S. Protein Data Bank (PDB): the single global macromolecular structure archive. *Protein Crystallography,* 627–641 (2017).
6. Gaulton, A., Bellis, L. J., Bento, A. P., Chambers, J., Davies, M., Hersey, A., Light, Y., McGlinchey, S., Michalovich, D., Al-Lazikani, B. & Overington, J. P. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Research* **40,** D1100–D1107 (2011).
7. Wang, Y., Xiao, J., Suzek, T. O., Zhang, J., Wang, J., Zhou, Z., Han, L., Karapetyan, K., Dracheva, S., Shoemaker, B. A., Bolton, E., Gindulyte, A. & Bryant, S. H. PubChem's BioAssay Database. *Nucleic Acids Research* **40,** D400–D412 (2012).
8. Heller, S., McNaught, A., Stein, S., Tchekhovskoi, D. & Pletnev, I. InChI - the worldwide chemical structure identifier standard. *Journal of Cheminformatics* **5,** 1–9 (2013).
9. van den Oord, A., Li, Y. & Vinyals, O. Representation Learning with Contrastive Predictive Coding. arXiv: `1807.03748` (2018).

10. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G. & Sutskever, I. Learning Transferable Visual Models From Natural Language Supervision. *International Conference on Machine Learning (ICML)* (2021).

11. Fürst, A., Rumetshofer, E., Lehner, J., Tran, V., Tang, F., Ramsauer, H., Kreil, D. P., Kopp, M., Klambauer, G., Bitto-Nemling, A. & Hochreiter, S. CLOOB: Modern Hopfield Networks with InfoLOOB Outperform CLIP. *Advances in Neural Infomation Processing Systems (NeurIPS)* (2022).

12. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C. & Chen, M. Hierarchical text-conditional image generation with CLIP latents. arXiv: `2204.06125` (2022).

13. Zanella, F., Lorens, J. B. & Link, W. High content screening: seeing is believing. *Trends in biotechnology* **28,** 237–245 (2010).

14. Thul, P. J., Åkesson, L., Wiking, M., Mahdessian, D., Geladaki, A., Ait Blal, H., Alm, T., Asplund, A., Björk, L., Breckels, L. M., *et al.* A subcellular map of the human proteome. *Science* **356,** eaal3321 (2017).

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

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

17. 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.* CellProfiler: image analysis software for identifying and quantifying cell phenotypes. *Genome Biology* **7,** 1–11 (2006).

18. Wawer, M. J., Li, K., Gustafsdottir, S. M., Ljosa, V., Bodycombe, N. E., Marton, M. A., Sokolnicki, K. L., Bray, M. A., Kemp, M. M., Winchester, E., Taylor, B., Grant, G. B., Hon, C. S. Y., Duvall, J. R., Wilson, J. A., Bittker, J. A., Dančík, V., Narayan, R., Subramanian, A., Winckler, W., Golub, T. R., Carpenter, A. E., Shamji, A. F., Schreiber, S. L. & Clemons, P. A. Toward performance-diverse small-molecule libraries for cell-based phenotypic screening using multiplexed high-dimensional profiling. *Proceedings of the National Academy of Sciences of the United States of America* **111,** 10911–10916. ISSN: 10916490 (30 July 2014).

19. Simm, J., Klambauer, G., Arany, A., Steijaert, M., Wegner, J. K., Gustin, E., Chupakhin, V., Chong, Y. T., Vialard, J., Buijnsters, P., *et al.* Repurposing high-throughput image assays enables biological activity prediction for drug discovery. *Cell Chemical Biology* **25,** 611–618 (2018).

20. Hofmarcher, M., Rumetshofer, E., Clevert, D.-A., Hochreiter, S. & Klambauer, G. Accurate Prediction of Biological Assays with High-Throughput Microscopy Images and Convolutional Networks. *Journal of Chemical Information and Modeling* **59,** 1163-1171 (2019).

21. Bray, M.-A., Gustafsdottir, S. M., Rohban, M. H., Singh, S., Ljosa, V., Sokolnicki, K. L., Bittker, J. A., Bodycombe, N. E., Dancik, V., Hasaka, T. P., *et al.* A dataset of images and morphological profiles of 30 000 small-molecule treatments using the Cell Painting assay. *Gigascience* **6,** 1–5 (2017).

22. Chen, T., Kornblith, S., Norouzi, M. & Hinton, G. A Simple Framework for Contrastive Learning of Visual Representations. *International Conference on Machine Learning (ICML). Proceedings of Machine Learning Research (PMLR)* 1597–1607 (2020).

23. Wang, F. & Liu, H. *Understanding the Behaviour of Contrastive Loss* in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021), 2495–2504.

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

25. Wellman, M. P. & Henrion, M. Explaining 'Explaining Away'. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **15,** 287–292 (1993).

26. Yeh, C.-H., Hong, C.-Y., Hsu, Y.-C., Liu, T.-L., Chen, Y. & LeCun, Y. Decoupled Contrastive Learning. arXiv: `2110.06848` (2021).

27. Zhang, C., Zhang, K., Pham, T. X., Niu, A., Qiao, Z., Yoo, C. D. & Kweon, I. S. Dual Temperature Helps Contrastive Learning Without Many Negative Samples: Towards Understanding and Simplifying MoCo. arXiv: `2203.17248` (2022).

28. Ramsauer, H., Schäfl, B., Lehner, J., Seidl, P., Widrich, M., Gruber, L., Holzleitner, M., Adler, T., Kreil, D., Kopp, M. K., Klambauer, G., Brandstetter, J. & Hochreiter, S. Hopfield networks is all you need. *International Conference on Learning Representations* (2021).

29. Poole, B., Ozair, S., van den Oord, A., Alemi, A. A. & Tucker, G. On Variational Bounds of Mutual Information. *Proceedings of Machine Learning Research (PMLR)* **97,** 5171–5180 (2019).

30. He, K., Zhang, X., Ren, S. & Sun, J. Deep Residual Learning for Image Recognition. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016).

31. Ilse, M., Tomczak, J. M. & Welling, M. in *Handbook of Medical Image Computing and Computer Assisted Intervention* 521–546 (Elsevier, 2020).

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

33. Dahl, G. E., Jaitly, N. & Salakhutdinov, R. Multi-task Neural Networks for QSAR Predictions. *ArXiv.* eprint: `1406.1231` (2014).

34. Unterthiner, T., Mayr, A., Klambauer, G., Steijaert, M., Ceulemans, H., Wegner, J. K. & Hochreiter, S. Deep Learning as an Opportunity in Virtual Screening. *Advances in Neural Information Processing Systems (NeurIPS), Workshop on Deep Learning and Representation Learning* (2014).

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

36. Jiang, D., Wu, Z., Hsieh, C.-Y., Chen, G., Liao, B., Wang, Z., Shen, C., Cao, D., Wu, J. & Hou, T. 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,** 1–23 (2021).

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

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

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

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

41. Xu, K., Hu, W., Leskovec, J. & Jegelka, S. How Powerful are Graph Neural Networks? *International Conference on Learning Representations (ICLR)* (2018).

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

43. Alperstein, Z., Cherkasov, A. & Rolfe, J. T. All SMILES variational autoencoder. arXiv: `1905.13343` (2019).

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

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

46. Wortsman, M., Ilharco, G., Li, M., Kim, J. W., Hajishirzi, H., Farhadi, A., Namkoong, H. & Schmidt, L. Robust fine-tuning of zero-shot models. arXiv: `2109.01903` (2021).

47. Kingma, D. P., Mohamed, S., Rezende, D. . & Welling, M. Semi-supervised Learning with Deep Generative Models. *Advances in Neural Information Processing Systems (NeurIPS),* 3581–3589 (2014).

48. Loshchilov, I. & Hutter, F. Decoupled Weight Decay Regularization. *International Conference on Learning Representations (ICLR)* (2019).

49. Loshchilov, I. & Hutter, F. SGDR: Stochastic Gradient Descent with Warm Restarts. *International Conference on Learning Representations (ICLR)* (2017).

50. *CellPainting gallery.* `https://registry.opendata.aws/cellpainting-gallery`. Accessed 20/04/2023.

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

52. Alain, G. & Bengio, Y. Understanding intermediate layers using linear classifier probes. arXiv: `1610.01644` (2016).

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

54. Corsello, S. M., Bittker, J. A., Liu, Z., Gould, J., McCarren, P., Hirschman, J. E., Johnston, S. E., Vrcic, A., Wong, B., Khan, M., Asiedu, J., Narayan, R., Mader, C. C., Subramanian, A. & Golub, T. R. The Drug Repurposing Hub: a next-generation drug library and information resource. *Nature Medicine 2017 23:4* **23,** 405–408. ISSN: 1546-170X (4 Apr. 2017).

55. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet Classification with Deep Convolutional Neural Networks. *Advances in Neural Information Processing Systems (NeurIPS),* 1097–1105 (2012).

56. Sun, C., Shrivastava, A., Singh, S. & Gupta, A. Revisiting Unreasonable Effectiveness of Data in Deep Learning Era. arXiv: 1707.02968v2 (2017).

57. Marcus, G. Deep learning: A critical appraisal. arXiv: 1801.00631 (2018).

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

59. Gutmann, M. & Hyvärinen, A. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. *International Conference on Artificial Intelligence and Statistics. Proceedings of Machine Learning Research* **9,** 297–304 (2010).

60. He, K., Fan, H., Wu, Y., Xie, S. & Girshick, R. B. Momentum Contrast for Unsupervised Visual Representation Learning. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2020).

61. Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P. & Joulin, A. Unsupervised Learning of Visual Features by Contrasting Cluster Assignments. *Advances in Neural Information Processing Systems (NeurIPS),* 9912–9924 (2020).

62. 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. & Valko, M. Bootstrap Your Own Latent - A New Approach to Self-Supervised Learning. *Advances in Neural Information Processing Systems (NeurIPS)* **33,** 21271–21284 (2020).

63. Devillers, B., Bielawski, R., Choski, B. & VanRullen, R. Does language help generalization in vision models? arXiv: 2104.08313 (2021).

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

65. 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. & Hochreiter, S. *Hopfield networks is all you need* in *International Conference on Learning Representations (ICLR)* (2021).

66. 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. & Sittampalam, G. S. Impact of high-throughput screening in biomedical research. *Nature Reviews Drug Discovery* **10,** 188–195. ISSN: 14741776 (2011).

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

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

69. Stärk, H., Beaini, D., Corso, G., Tossou, P., Dallago, C., Günnemann, S. & Liò, P. 3D Infomax improves GNNs for Molecular Property Prediction. arXiv: 2110.04126 (2021).

70. Lee, H., Ahn, S., Seo, S.-W., Song, Y. Y., Yang, E., Hwang, S.-J. & Shin, J. RetCL: A Selection-based Approach for Retrosynthesis via Contrastive Learning. arXiv: 2105.00795 (2021).

71. Seidl, P., Renz, P., Dyubankova, N., Neves, P., Verhoeven, J., Wegner, J. K., Segler, M., Hochreiter, S. & Klambauer, G. Improving Few-and Zero-Shot Reaction Template Prediction Using Modern Hopfield Networks. *Journal of Chemical Information and Modeling* (2022).

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

73. 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.* Data-analysis strategies for image-based cell profiling. *Nature Methods* **14,** 849–863 (2017).

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

75. Bender, A. & Cortés-Ciriano, I. 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,** 511–524 (2021).

76. Chandrasekaran, S. N., Cimini, B. A., Goodale, A., Miller, L., Kost-Alimova, M., Jamali, N., Doench, J., Fritchman, B., Skepner, A., Melanson, M., Arevalo, J., Caicedo, J. C., Kuhn, D., Hernandez, D., Berstler, J., Shafqat-Abbasi, H., Root, D., Swalley, S., Singh, S. & Carpenter, A. E. Three million images and morphological profiles of cells treated with matched chemical and genetic perturbations. bioRxiv: 10.1101/2022.01.05.475090 (2022).

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

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

79. 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.* Industry-scale application and evaluation of deep learning for drug target prediction. *Journal of Cheminformatics* **12,** 1–13 (2020).

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), 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), EPILEPSIA (FFG-892171), AIRI FG 9-N (FWF-36284, FWF-36235), ELISE (H2020-ICT-2019-3 ID: 951847), Stars4Waters (HORIZON-CL6-2021-CLIMATE-01-01). We thank Audi.JKU Deep Learning Center, TGW LOGISTICS GROUP GMBH, Silicon Austria Labs (SAL), FILL GmbH, Anyline GmbH, Google, ZF Friedrichshafen AG, Robert Bosch GmbH, UCB Biopharma SRL, Merck Healthcare KGaA, Verbund AG, GLS (Univ. Waterloo) Software Competence Center Hagenberg GmbH, TÜV Austria, Frauscher Sensonic and the NVIDIA Corporation.

Author contributions

ASF and ER implemented the methods and algorithms. ASF, ER designed and performed the experiments. ASF, ER, SH, and GK analyzed and interpreted the results. ASF, ER, and GK wrote the manuscript. SH and GK conceived and designed the study.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to klambauer@ml.jku.at.

A 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.

B CLOOME hyperparameter search space

Below, we state the hyperparameter selection for the results reported in this study, based on performance in a validation set for each one of the downstream tasks. For the retrieval task, the model was trained for 70 epochs for the random split and for 60 epochs for the scaffold split, based on the top-1 accuracy in validation. For the zero-shot molecule classification, the selected models were trained for 63 and 57 epochs for the random and scaffold split, respectively. For the zero-shot mechanism of action classification, the selected models were trained for 70 and 69 epochs for the random and scaffold split, respectively.

B.1 Retrieval task hyperparameter search space

	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***}
	Initial inverse temperature	{**14.3**, 30}
	Learnable inverse temperature	{***True***, False}
Image encoder	Image resolution	{320, ***520***}
	Model	{ResNet50}
Structure encoder	Input size	{1024, ***8192***}
	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 on manual search on validation set shown in bold for the random split and in italics for the scaffold split.

B.2 Bioactivity prediction hyperparameter search space

	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**}
	Initial inverse temperature	{14.3, **30**}
	Learnable inverse temperature	{True, **False**}
Image encoder	Image resolution	{**320**, 520}
	Model	{ResNet50}
Structure encoder	Input size	{1024}
	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 A3: Considered hyperparameter space of CLOOME models. The selected configurations based on manual search on validation set shown in bold.

B.3 Zero-shot molecule prediction hyperparameter search space

	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	{***14.3***, 30}
Image encoder	Image resolution	{320, ***520***}
	Model	{ResNet50}
Structure encoder	Input size	{***1024***, 8192}
	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 A4: Considered hyperparameter space of CLOOME models. The selected configurations based on manual search on validation set shown in bold for the random split and in italics for the scaffold split.

B.4 Zero-shot MoA prediction hyperparameter search space

	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	{***14.3***, 30}
Image encoder	Image resolution	{320, ***520***}
	Model	{ResNet50}
Structure encoder	Input size	{***1024***, 8192}
	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 A5: Considered hyperparameter space of CLOOME models. The selected configurations based on manual search on validation set shown in bold for the random split and in italics for the scaffold split.

C Bilinear model hyperparameter search space

The hyperparameters of the bilinear models that yielded the results shown in this paper were also selected based on performance in a validation set, and are shown in Table C. The selected models were trained for 52 and 54 epochs for the random and scaffold split, respectively.

	Hyperparameter	Explored space
Learning	Optimizer	{SGD, ***AdamW***}
	Learning rate	{**0.0005**, *0.001*}
	Scheduler	{No scheduler, ***Cosine annealing with restarts***}
	Weight decay	{0.1}
	Batch size	{**256**, *512*, 1024}
	Warm-up iterations	{5000, 10000, ***20000***}
Weight matrix	Size	{(1081, 8192)}

Table A6: Considered hyperparameter space for the bilinear model. The selected configurations based on manual search on validation set shown in bold for the random split and in italics for the scaffold split.