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 the image. We present a novel retrieval system based on contrastive learning that is able to identify the chemical structure inducing the phenotype out of ∼2,000 candidates with a top-1 accuracy >70 times higher than a random baseline.

Brief Communication

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 methods, 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 approaches from sequence similarities to structural queries based on 3D shape. The chemical databases ChEMBL [6] and PubChem are huge corpora of chemical structures that contain billions of small molecules. The International Chemical Identifier (InChI) [7] was designed to facilitate querying 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 by 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 [8]. The contrastive learning methods CLIP and CLOOB embed natural language and images into the same representation space [9, 10]. Contrastive learning enforces that images and their matching 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 [9].These text-image embedding spaces enabled the generation of realistic images from short text prompts and led to the recent boom of "AI art" [11]. In this work, we use these powerful contrastive learning paradigms to enable retrieval or querying systems for microscopy images.

In order to characterize cell phenotypes, tissues, or cellular processes, microscopy imaging has been used as an informative and time- and cost-efficient biotechnology [12, 13]. Consequently, there have been efforts by the scientific community to use high-throughput microscopy imaging [14] as informative read-out and characterization of cellular systems and phenotypes under diverse perturbations [13, 15]. 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 [16]. The immense amount of microscopy imaging data are stored in large databases, many of which are publicly available. Their querying procedures, 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 content-based or associative 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 4C,F).

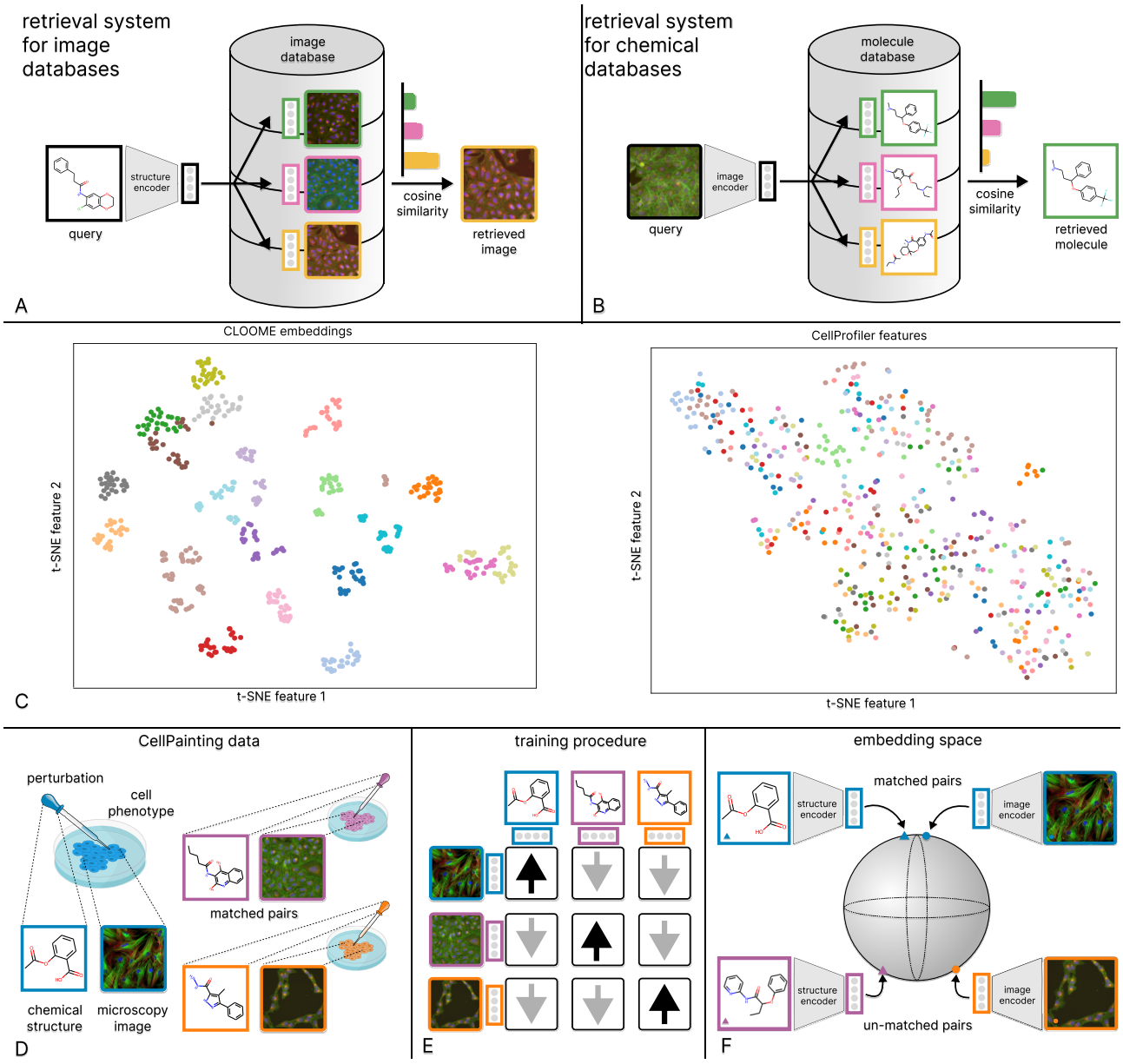

Figure 1: **A, B** The CLOOME encoders can be used to query a microscopy imaging database (**A**) by a chemical structure, and vice versa, query a chemical database by a microscopy image (**B**). **C** Visualization of the embedding space in terms of a t-SNE projection of image embeddings of new cell 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 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 changes. 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.

CLIP and CLOOB models have been constructed via contrastive learning on large image-text datasets [9]. Analogously to image-text pair datasets, the Cell Painting dataset [15] 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 (see Figure 1E, F). We propose a contrastive learning framework for image-structure pairs that we call CLOOME (see Online Methods). The training process of CLOOME would yield a) an image encoder that can map microscopy images to an informative embedding space, b) a structure encoder that can map chemical structures to the same embedding space (see Figure 1F). Both encoders are deep neural networks that build the basis of a search engine, which we also call CLOOME, for microscopy images and chemical structures (Figure 1F). If CLOOME achieves similar results for microscopy images to CLIP or CLOOB for natural images [9, 10], the image encoder should produce features, or equivalently cell profiles [17], that are highly transferable and robust to distribution shifts.

We trained the encoders of CLOOME on 674,357 pairs of microscopy images and chemical structures of the Cell Painting dataset, setting aside 28,632 for validation and 56,793 pairs for testing, ensuring that no data leakage occurred between these sets (see Online Methods). We used a Residual Network [18] to encode microscopy images [19], and a fully-connected neural network to encode chemical structures [20]. For each training step, 256 image-structure pairs are randomly drawn from the training set and the encoders are updated to increase the cosine similarities of the matched pairs and decrease the cosine similarity of un-matched pairs (Figure 1E). We trained the CLOOME encoders for the retrieval system for 51 epochs, based on validation performance (details in Online Methods). After the training process, we investigated CLOOME models for a) the use as a retrieval systems for microscopy images and chemical structures, b) the quality of the image embeddings to predict bioactivities c) the expressiveness of the image embeddings to distinguish between unseen cell phenotypes.

a) The CLOOME encoders as retrieval system for microscopy images and chemical structures. On a hold-out set of new 2,115 molecules and images, we tested whether CLOOME is able to correctly identify the chemical structure with which the cells have been treated, and vice versa. This, to our knowledge, is a task that is considered almost impossible for human experts. The trained CLOOME system is able to identify the matched microscopy image given the chemical structure, and vice versa, with an top-10 accuracy of 8.4% (95%-CI 7.3-9.7%) an 7.9% (95%-CI 6.8-9.1%), respectively. The task is extremely difficult because there are many chemical structures that induce similar cell phenotypes or no phenotypic changes at all and because one correct image or chemical structure has to be selected from a large set of ∼2,000 candidates. Therefore, the performance of human experts would likely be close to random, over which CLOOME exhibits a 70-fold improvement for image retrieval and 64-fold improvement for structure retrieval over this random baseline (see Table 1). This means that by just investigating the cell phenotype displayed on the microscopy image, CLOOME is able to identify the matched chemical structure from a large database, and vice versa (see Table 1). Therefore, the CLOOME encoders can be used as a content-based or associative retrieval system for microscopy images and chemical structures [1].

Method	Top-k accuracy (%)					
	Top-1 (rel.)	95%-CI	Top-5 (rel.)	95%-CI	Top-10 (rel.)	95%-CI
CLOOME (structure retr.)	**3.03** (64.4x)	[2.34, 3.85]	**6.62** (25.4x)	[5.60, 7.76]	**8.42** (17.8x)	[7.27, 9.68]
CLOOME (image retr.)	**3.31** (70.4x)	[2.59, 4.16]	**6.24** (24.0x)	[5.24, 7.36]	**7.90** (16.7x)	[6.78, 9.13]
Random	0.0473	[0.0012, 0.263]	0.236	[0.0768, 0.551]	0.473	[0.227, 0.868]

Table 1: Results for the database retrieval (retr.) task. Given a molecule-perturbed microscopy image, the matched molecule, i.e. chemical structure, must be selected from a set of ∼2,000 candidate molecules (first row). Vice versa, given a chemical structure, the matched microscopy image, capturing the phenotype induced by the chemical perturbation, has to be selected from ∼2,000 candidate images (second row). 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.

b) Bio-activity prediction. Next, we investigated whether the embeddings are transferable to other tasks. To this end, we used bioactivity prediction tasks as these have been approached before with cell profiling [16] and convolutional neural networks (CNNs) [19]. We found that without the need for re-training or fine-tuning any neural network, the CLOOME image embeddings could predict 209 activity prediction tasks with an AUC of 0.714 ± 0.20, which is on par with the best method CNNs trained in an end-to-end fashion. For CLOOME, we just trained a logistic regression model, and, thus, we can conclude that the embeddings are highly transferable and predictive for a diverse set of activity prediction tasks.

c) Zero-shot image classification. Lastly, we investigated how well the CLOOME image embeddings characterize new cell phenotypes and whether new phenotypes could potentially be detected. We again used a

[1] The search engine is available here `https://huggingface.co/spaces/anasanchezf/cloome`

hold-out set of microscopy images. We took the simplified assumption that each chemical structure induces a separate new cell phenotype. If different structures in reality induce the same phenotype, this would even make the prediction task simpler. Since multiple different samples, i.e. images, were treated with the same chemical, they should produce similar embeddings or cell profiles. Therefore, a t-SNE plot, in which each point represents a microscopy image embedding colored by phenotype, should show clusters of datapoints of the same color (see Figure 1C left). For comparison, we also produced the same plot using the cell profiles computed with the CellProfiler [17] software (see Figure 1C right). Besides the visual confirmation, the classification accuracy can also be quantified (see Online Methods). Indeed, the CLOOME embeddings are similar for images capturing the same, but previously unknown cell phenotype, and are thus highly expressive features of cells.

To conclude, we have demonstrated that self-supervised contrastive learning methods can be readily used for multi-modal data arising from informative biotechnologies, such as microscopy images. Our contrastive learning framework CLOOME was used to construct a common embedding space for microscopy images capturing cell phenotypes and chemical structures inducing cellular processes. This enabled us to build a content-based retrieval system for microscopy images and chemical structures. Furthermore, the learned microscopy image encoder has been shown to produce highly transferable and expressive embeddings, or cell profiles, that can be efficiently used to predict bioactivities or detect new phenotypes. We envision that our work paves the way for retrieval systems for other pairs of modalities, for example querying microscopy imaging databases with transcriptomics signatures or vice versa. We provide the CLOOME search engine as a free web application, the CLOOME framework, the encoders and all code on github `https://github.com/ml-jku/cloome`.

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

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. Heller, S., McNaught, A., Stein, S., Tchekhovskoi, D. & Pletnev, I. InChI - the worldwide chemical structure identifier standard. *Journal of Cheminformatics* **5,** 1–9 (2013).
8. van den Oord, A., Li, Y. & Vinyals, O. Representation Learning with Contrastive Predictive Coding. arXiv: `1807.03748` (2018).
9. 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).

10. 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).

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

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

13. 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).

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

15. 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).

16. 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).

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

19. 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).

20. 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).

A Online Methods

"CLOOME" is a contrastive learning framework for multi-modal microscopy imaging data. 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 [1] and CLOOB [2] to image-structure pairs, we call it Contrastive Learning and leave-One-Out-boost for Molecule Encoders (CLOOME). In the following we provide details on the method, data, training, assessment and evaluation. Concretely,

- we introduce a new contrastive learning approach for image- and structure-based representations of molecules,
- 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 rich representations of molecules which allow to retrieve potential bioisosteres from image or chemical databases.

A.1 CLOOME: Contrastive Learning and Leave-One-Out Boost for Molecule Encoders

We propose contrastive learning of representations from pairs of microscopy images and chemical structures to obtain a common embedding space of these two modalities, and to obtain highly transferable encoders (see Figure 2). In contrast to previous approaches, in which chemical structure encoders learned representations using activity data [3, 4] or microscopy image encoders used hand-crafted representations [5, 6], CLOOME optimizes representations without activity data or human expertise.

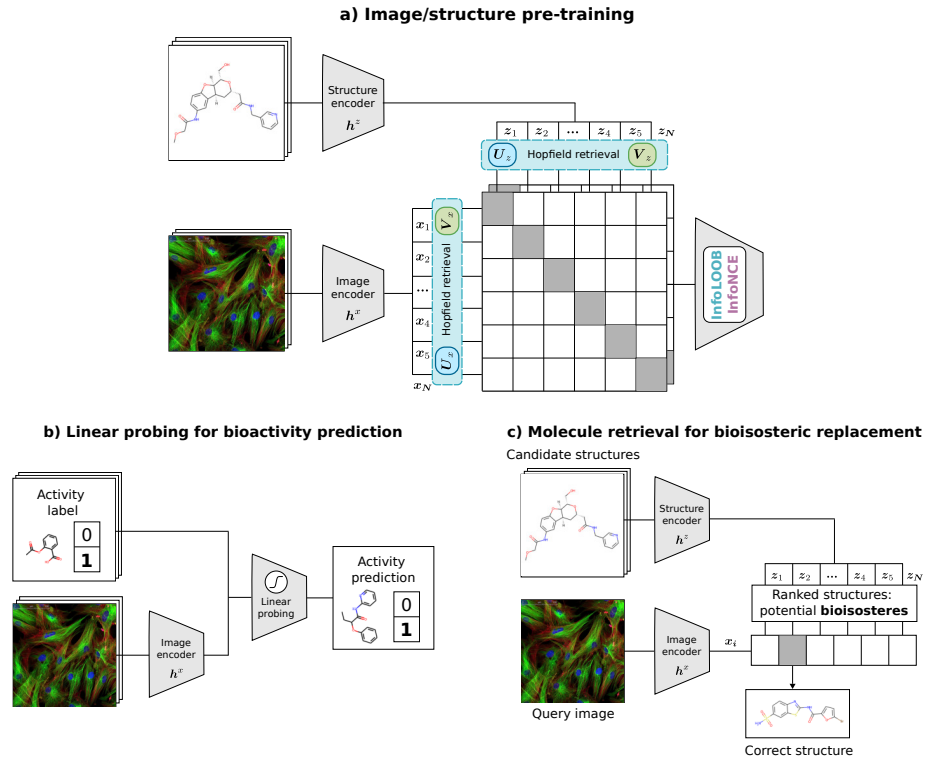

Figure 2: Schematic representation of CLOOME. Contrastive pre-training of embeddings of the two modalities, microscopy image and chemical structure, of a molecule using the CLOOB [2] approach. **b)** Using the CLOOME embeddings for activity prediction. A logistic regression model is trained for activity prediction tasks. **c)** The resulting embeddings can be used to rank chemical structures that induce similar phenotypic effects, which can be considered a bioisosteric replacement task.

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 $\|\boldsymbol{x}_n\| = \|\boldsymbol{z}_n\| = 1 \; \forall n$. For notation, see also Table 6.

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 [1, 7, 8], which is also used in the CLIP approach [1]. 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.

The contrastive learning method CLIP has the problem of "explaining away" [2, 9, 10]. 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 [2, 11, 12]. To ameliorate these drawbacks, CLOOB [2] has introduced the InfoLOOB objective together with Hopfield networks as a promising method for contrastive learning. Our contrastive learning framework CLOOME comprises both methods CLIP [1] and CLOOB [2].

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:

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

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

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

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

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 6. Then, the InfoLOOB objective [2, 13] for the retrieved embeddings is used as objective function:

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

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 [3, 14]. Although standard image encoders, such as Residual Networks [15] could be in principle used with minor adjustments, alternative architectures, such as multiple instance learning approaches, could be required for very high resolution datasets [16]. In all our experiments, we use a ResNet-50 encoder with five input channels and downsized the microscopy images to 320x320 pixels.

Molecule structure encoder. Since the advent of Deep Learning, a large number of architectures to encode molecules have been suggested [17–21]. 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 [22, 23] with 4 hidden layers of 1024 units with ReLU activations and batch normalization (for further details see Sec. A.2 and Sec. A.6). However, also any graph [20, 24–26], message-passing [27], or sequence-based [28] neural network with an appropriate pooling operation can be used as structure encoder.

A.2 Experiments

Dataset and preprocessing. *Cell Painting.* We use pairs of microscopy images and molecules from the Cell Painting [6, 29] dataset. This dataset is a collection of high-throughput fluorescence microscopy images of U2OS cells treated with different small molecules [29]. 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.* [3], 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, their corresponding SMILES strings were transformed to 1024-bit Morgan fingerprints with a radius of 3, taking chirality into account [30, 31].

Data splits. We split our dataset into training, validation, and test set, using the splits of Hofmarcher *et al.* [3]. 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 [32] and CLOOB [2] 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 [23] and was not subject to substantial hyperparameter optimization. Due to computational constraints, the search was limited to the hyperparameters shown in Table 7. We used the Adam optimizer [33] with decoupled weight decay regularization [34]. 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 [35]. 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 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. 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 the model was trained for 51 epochs based on the top-1 accuracy in validation. In this case, images were cropped and re-scaled to a pixel resolution of 520x520, based on performance in the validation set. 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 Sec. A.6.

a) A retrieval system for imaging and chemical databases to enable bioisoteric replacement and scaffold hopping. 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: 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 [36]: 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 2 **b)**).

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 Sec. A.6). The top-1, top-5, top-10 accuracy are given in Table 2 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 3. Further, some examples are displayed in Figure 3. This is, to our knowledge, the first system of cell-image-based retrieval of molecular structures.

Method	Top-k accuracy (%)					
	Top-1 (rel.)	95%-CI	Top-5 (rel.)	95%-CI	Top-10 (rel.)	95%-CI
CLOOME (structure retr.)	**3.03** (64.4x)	[2.34, 3.85]	**6.62** (25.4x)	[5.60, 7.76]	**8.42** (17.8x)	[7.27, 9.68]
CLOOME (image retr.)	**3.31** (70.4x)	[2.59, 4.16]	**6.24** (24.0x)	[5.24, 7.36]	**7.90** (16.7x)	[6.78, 9.13]
Random	0.0473	[0.0012, 0.263]	0.236	[0.0768, 0.551]	0.473	[0.227, 0.868]

Table 2: Results for the retrieval task **among 2,115** 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.

Method	Top-k accuracy (%)					
	Top-1	95% CI	Top-5	95% CI	Top-10	95% CI
CLOOME (structure retrieval)	**10.4**	[9.10, 11.7]	**21.3**	[19.6, 23.1]	**30.6**	[28.7, 32.7]
CLOOME (image retrieval)	**9.64**	[8.42, 11.0]	**20.7**	[19.0, 22.4]	**29.0**	[27.1, 31.0]
Random	0.992	[0.616, 1.51]	5.01	[4.12, 6.03]	10.0	[8.78, 11.4]

Table 3: 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.

b) Bio-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 [8, 37] 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 [8], 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.* [3]. 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 [3, 4], 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 [3, 4] 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.* [3]. 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 [4–6]. The compared methods were trained in a multi-task setting to predict activity labels for 209 tasks, extracted from ChEMBL.

Results. The predictive performance on the downstream activity prediction tasks is reported in Table 4. 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 [38] and SC-CNN [3], with respect to AUC.

c) Zero-shot microscopy image classification. The goal of this analysis is to evaluate the potential of the CLOOME image embeddings to distinguish between cell phenotypes. Classifying the phenotype captured by the microscopy image is a highly relevant biological question. Especially for drug discovery, where phenotypes are induced by chemical perturbations, embeddings that can identify novel phenotypes would provide some understanding about its possible mode of action and therefore its potential adverse effects.

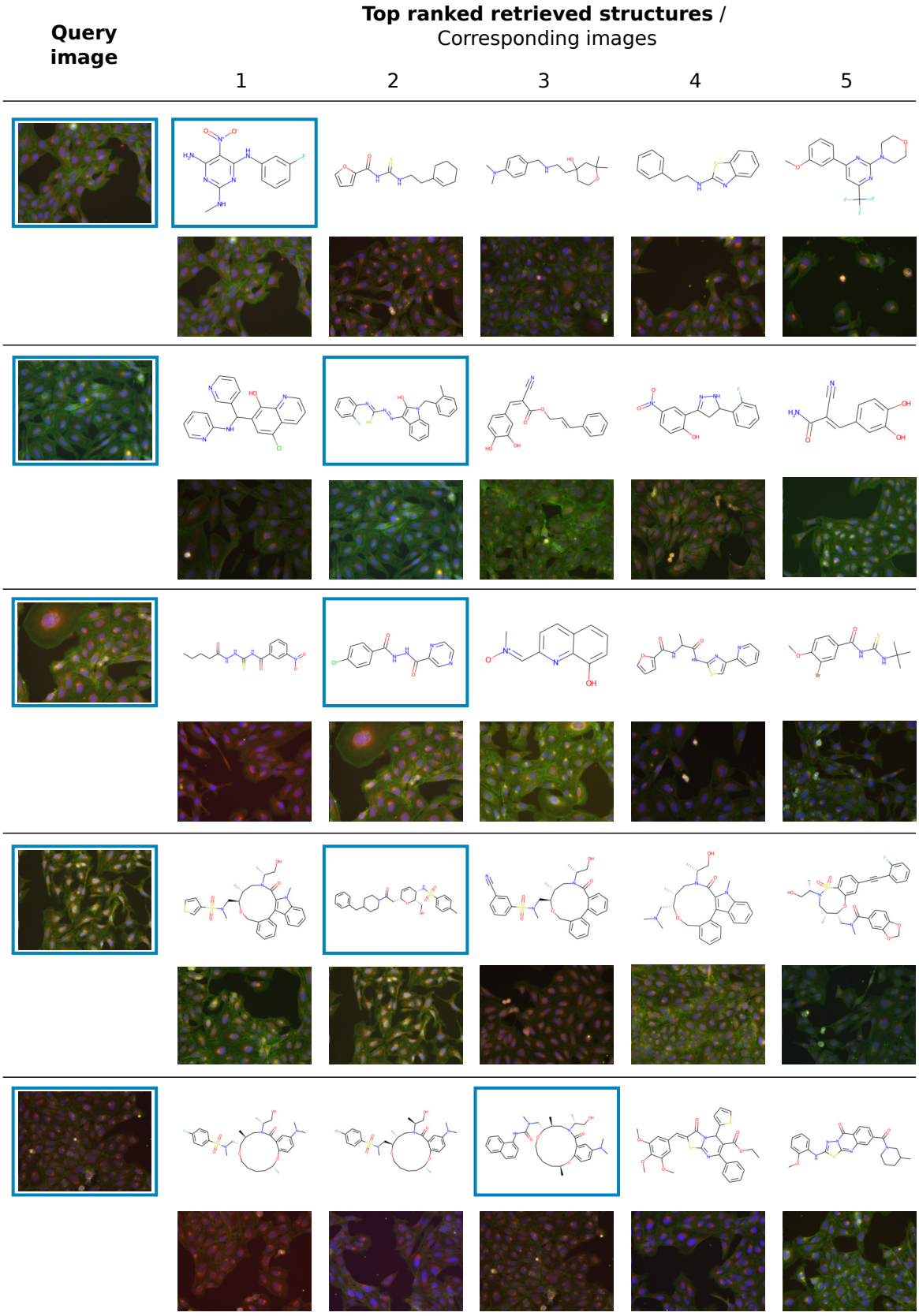

Figure 3: Example results for the retrieval task. On a hold-out test set, the five molecules for which representations are the most similar to the query image are shown along with their corresponding images. Blue boxes mark the query image and its matching molecular structure, i.e. the matching pair. CLOOME can be used to retrieve molecules that could produce similar biological effects on treated cells, i.e. bioisosteres.

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 4: Comparison of the linear probing evaluation of the learned representations against fully supervised methods [3]. 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 [8]. The best method in each category is marked bold.

In the same hold-out test used for this zero-shot image classification task, each molecule is assumed to cause a different phenotype. While this assumption is not true, and distinct chemical structures can induce the same phenotype, this is the most difficult setting and the more realistic setting would make the classification task even easier. To provide details on this classification task, one image for each of the molecules was randomly selected, resulting in 2,115 classes. Then, samples corresponding to both the same molecule and plate as those from the class set were removed in order to ensure that the classification was not due to plate effects. This yielded a 44,102 image test set.

We compared the CLOOME embeddings, to embeddings of a microscopy image encoder trained in supervised fashion and to profiles from CellProfiler. Regarding GapNet embeddings, the images were encoded using the model weights provided by Hofmarcher *et al.* [3], removing the last layer of its classifier, which resulted in a 1024-dimension embedding space. CellProfiler embeddings consist in 148 features aggregated in one vector per image, as made availabe in Bray *et al.* [6].

Method	Accuracy[%]					
	Top-1	95% CI	Top-5	95% CI	Top-10	95% CI
CLOOME	**17.8**	[17.4, 18.2]	**40.6**	[40.2, 41.1]	**55.3**	[54.8, 55.8]
GapNet (CNN)	0.363	[0.309, 0.423]	1.07	[0.981, 1.18]	1.80	[1.67, 1.92]
CellProfiler	0.497	[0.433, 0.567]	1.75	[1.63, 1.88]	2.89	[2.74, 3.05]

Table 5: Results for the zero-shot microscopy image 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.

A.3 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 [15, 39]. These supervised methods require large amounts of labeled data, which may be very costly or unfeasible to obtain, and they have limited generalization abilities [40, 41]. 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 [8, 42]. With contrastive learning methods [43] and self-supervision these meaningful representations can be obtained without the need for large amounts of expensive manually-provided labels [8, 44–46]. While uni-modal methods typically use pre-text tasks [8], for multi-modal methods the self-supervision arises from the availability of two modalities of an instance, such as image and text [1, 47]. Both uni-modal and multi-modal contrastive learning methods have recently had a substantial impact in computer vision and natural language processing [48].

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-

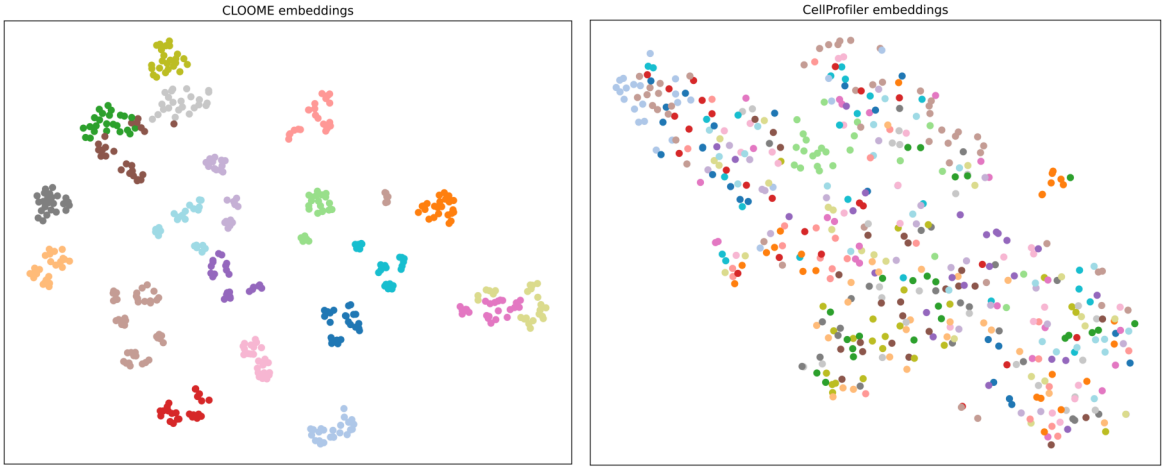

Figure 4: t-SNE downprojection of the CLOOME learned embeddings (left) and the CellProfiler extracted features (right) of all microscopy images corresponding to 20 molecules randomly selected from the test set. The colors represent different molecules.

Image Pre-training (CLIP) [1], 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 [2, 9, 10] (details in Section A.1). Considering this caveat, the "Contrastive Leave One Out Boost" (CLOOB) method has been proposed [2]. CLOOB uses a different objective, the "InfoLOOB" (LOOB for "Leave One Out Bound") objective [13], which does not include the positive pair in the denominator to avoid saturation effects [2]. Moreover, continuous modern Hopfield networks [49] 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 [50, 51]. 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 [52] uses contrastive molecule-to-molecule training by augmenting molecular graphs. Stärk *et al.* [53] contrastively learn 3D and 2D molecule representations to inform the learned molecule encoder with 3D information. Lee *et al.* [54] and Seidl *et al.* [55] use contrastive learning for molecules and chemical reactions, and Vall *et al.* [56] 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 [29] 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 [4, 29, 57, 58]. 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 [59]. Particularly, microscopy image-based profiles of small molecules have been suggested to be effective together with deep learning methods [58]. However, the current efforts are still in standard supervised learning settings based on extracted features [4] or deep architectures [3]. 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 [60]. Instead of the currently used activity measurements as labels [3, 4], 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 [61, 62].

A.4 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 CLOOME is able to learn 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.

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 [29]. This protocol has been published and currently there are community efforts [60] 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 [63]. 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 [1, 2], if the latter are very high dimensional, this method presents the risk of overfitting [37]. Having addressed these limitations, we nevertheless believe that the representations obtained with CLOOME could be highly useful for both the community using bioimaging as well as for drug discovery.

A.5 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 6: Symbols and notations used in this paper.

A.6 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	{**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 7: Considered hyperparameter space of CLOOME models. The selected configurations for downstream activity prediction based on manual search on validation set shown in bold.

References

1. 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).
2. 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).
3. 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).
4. 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).
5. 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).
6. 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).
7. van den Oord, A., Li, Y. & Vinyals, O. Representation Learning with Contrastive Predictive Coding. arXiv: `1807.03748` (2018).
8. 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).
9. Pearl, J. Embracing causality in default reasoning. *Artificial Intelligence* **35,** 259–271 (1988).
10. Wellman, M. P. & Henrion, M. Explaining 'Explaining Away'. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **15,** 287–292 (1993).
11. Yeh, C.-H., Hong, C.-Y., Hsu, Y.-C., Liu, T.-L., Chen, Y. & LeCun, Y. Decoupled Contrastive Learning. arXiv: `2110.06848` (2021).
12. 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).
13. 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).
14. Pepperkok, R. & Ellenberg, J. High-throughput fluorescence microscopy for systems biology. *Nature Reviews Molecular Cell Biology* **7,** 690–696 (2006).
15. He, K., Zhang, X., Ren, S. & Sun, J. Deep Residual Learning for Image Recognition. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016).
16. Ilse, M., Tomczak, J. M. & Welling, M. in *Handbook of Medical Image Computing and Computer Assisted Intervention* 521–546 (Elsevier, 2020).
17. 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).
18. Dahl, G. E., Jaitly, N. & Salakhutdinov, R. Multi-task Neural Networks for QSAR Predictions. *ArXiv.* eprint: `1406.1231` (2014).
19. 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).
20. Kearnes, S., McCloskey, K., Berndl, M., Pande, V. & Riley, P. Molecular graph convolutions: moving beyond fingerprints. *Journal of Computer-Aided Molecular Design* **30,** 595–608 (2016).
21. 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).

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

23. 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).

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

25. 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).

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

27. 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).

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

29. 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).

30. 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).

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

32. 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).

33. 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).

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

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

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

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

38. 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).

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

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

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

42. 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).

43. 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).

44. 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).

45. 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).

46. 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).

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

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

49. 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).

50. 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).

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

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

53. 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).

54. 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).

55. 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).

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

57. 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).

58. 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).

59. 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).

60. 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).

61. 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).

62. 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).

63. 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).