Molecule-Morphology Contrastive Pretraining for Transferable Molecular Representation

Cuong Q. Nguyen [1] **Dante Pertusi** [2] **Kim M. Branson** [1]

Abstract

Image-based profiling techniques have become increasingly popular over the past decade for their applications in target identification, mechanism-of-action inference, and assay development. These techniques have generated large datasets of cellular morphologies, which are typically used to investigate the effects of small molecule perturbagens. In this work, we extend the impact of such dataset to improving quantitative structure-activity relationship (QSAR) models by introducing Molecule-Morphology Contrastive Pretraining (MoCoP), a framework for learning multimodal representation of molecular graphs and cellular morphologies. We scale MoCoP to approximately 100K molecules and 600K morphological profiles using data from the JUMP-CP Consortium and show that MoCoP consistently improves performances of graph neural networks (GNNs) on molecular property prediction tasks in ChEMBL20 across all dataset sizes. The pretrained GNNs are also evaluated on internal GSK pharmacokinetic data and show an average improvement of 2.6% and 6.3% in AUPRC for full and low data regimes, respectively. Our findings suggest that integrating cellular morphologies with molecular graphs using MoCoP can significantly improve the performance of QSAR models, ultimately expanding the deep learning toolbox available for QSAR applications.

1. Introduction

Quantitative structure-activity relationship (QSAR) modeling is a critical step for virtual screening in drug discovery, helping researchers prioritize modifications to chemical structures that shift modeled properties in a favorable direction. Since the Merck Molecular Activity Challenge, applying deep learning techniques to QSAR modeling has gained significant attention due to their ability to extract complex nonlinear relationships between chemical structures and their associated activities. Typically, QSAR models are trained to predict the activity of a molecule based on its in silico representation, which can have varying levels of complexity ranging from computed chemical properties, 2- and 3-D descriptors (Rogers & Hahn, 2010; Sheridan et al., 1996; Carhart et al., 1985; Nilakantan et al., 1987; Schaller et al., 2020), and molecular graphs (Kearnes et al., 2016; Yang et al., 2019).

However, performance of QSAR models is limited by the amount of available data, especially when assays are low-throughput, expensive to run, or only commissioned at the later stages of the drug discovery process. To overcome this limitation, methods such as active learning (Reker & Schneider, 2015; Smith et al., 2018), large-scale multitask learning (Xu et al., 2017; Ramsundar et al., 2015; Kearnes et al., 2017) pretraining (Hu et al., 2020), and few-shot learning approaches (Altae-Tran et al., 2017; Nguyen et al., 2020) have been shown to improve model performance in low data regime.

Improving the in silico representation of molecules can also enhance performance of QSAR models. Recent trends in small-molecule drug discovery have shifted toward high-content screening approaches, with cellular imaging emerging as a relatively high-throughput (Kurita & Linington, 2015; Kraus et al., 2017; Chandrasekaran et al., 2021) method to profiling small molecules in relevant biological system. The Cell Painting assay (Bray et al., 2016) – an unbiased and scalable approach for capturing images of cells – have made large and reusable repositories of paired molecule and cell images possible (Bray et al., 2017; Fay et al., 2023; Chandrasekaran et al., 2023). These images contain cellular morphologies induced by small molecule perturbagens and can be used as an alternative in silico representation of these molecules (Kraus et al., 2017; Godinez et al., 2018; Hofmarcher et al., 2019; Stirling et al., 2021). Convolutional neural network-based approaches have been shown to improve the predictivity of QSAR models across a wide range of assays (Hofmarcher et al., 2019), leading to

[1]GSK, Artificial Intelligence and Machine Learning [2]GSK, Medicine Design. Correspondence to: Cuong Q. Nguyen <cuong.q.nguyen@gsk.com>.

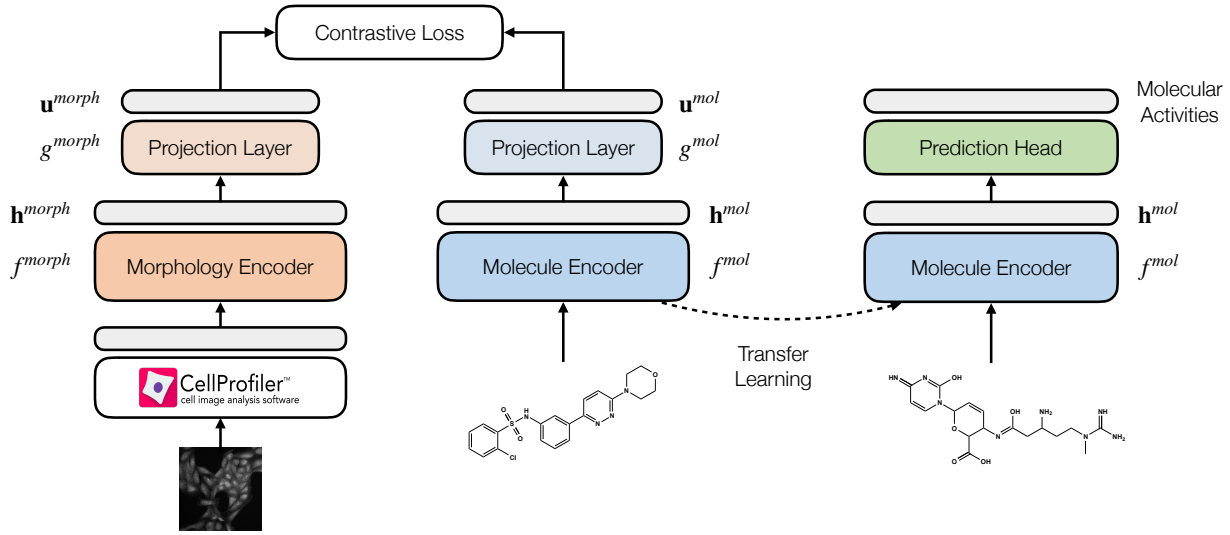

Figure 1. Molecule-morphology contrastive learning workflow. We first jointly learn a molecule encoder and morphology encoder using contrastive learning on paired (molecule, morphology) data in available in the JUMP-CP dataset (left). Transfer learning is then performed by fine-tuning the pretrained molecule encoder on specific downstream tasks (right).

increased hit rates and optimization of compounds to elicit a desired phenotype (Cuccarese et al., 2020). However, the use of such models is limited by two factors: (1) cellular images are commonly plagued by batch effects, requiring extensive engineering efforts to learn domain agnostic representation (Ando et al., 2017; Sypetkowski et al., 2023), and (2) only molecules that have paired cellular images can be used as input during inference, restricting the application of these models in virtual screening scenarios where such images are not available for the majority of molecules.

In parallel, contrastive learning has been shown to be effective for learning representations of multi-modal data. ConVIRT (Zhang et al., 2020) uses a modified InfoNCE objective (Oord et al., 2019) to learn a joint embedding space of medical images and human annotations. CLIP (Radford et al., 2021) scales up this approach to 400M (image, text) pairs, enabling zero-shot transfer to downstream image classification tasks. Recently, CLOOME (Sanchez-Fernandez et al., 2022) uses the InfoLOOB objective (Fürst et al., 2022) to jointly learn a molecule encoder and a morphology encoder for molecular retrieval task using the dataset introduced by Bray et al. (2017). Using the same dataset, Zheng et al. (2022) extends this approach by including masked-graph modeling objective for pretraining graph neural networks (GNNs), showing improved performances on downstream tasks in the Open Graph Benchmark (Hu et al., 2021).

In this work, we further demonstrate the scaling of GNN-based **M**olecule-m**o**rphology **Co**ntrastive **P**retraining – refered to as **MoCoP** – from 30K molecules and 120K images in Bray et al. (2017) to approximately 100K molecules and 600K images in JUMP-CP (Chandrasekaran et al., 2023). Using the modified InfoNCE objective (Zhang et al., 2020; Radford et al., 2021) and a gated graph neural network (GGNN) molecule encoder, we first show the effects of pretraining dataset sizes on morphology retrieval tasks. Transfer learning performances of GGNN molecule encoder pretrained with MoCoP is benchmarked on QSAR modeling task with varying training set sizes using the ChEMBL20 dataset (Gaulton et al., 2012). Finally, we demonstrate positive transfer of pretrained GGNNs on internal GSK pharmacokinetic data consisting of four different in vitro clearance assays.

2. Background

Learning multi-modal molecule and morphology representation with contrastive learning Contrastive learning is a member of the metric learning family which aims to learn an embedding space that pulls similar data together and pushes dissimilar data apart. Contrastive learning has experienced a resurgence in interest due to major advances in self-supervised learning. More recently, it has been increasingly employed to learn multi-modal data representation (Zhang et al., 2020; Desai & Johnson, 2021; Radford et al., 2021). For MoCoP, we employ a symmetric variant of InfoNCE loss for pretraining following prior works (Zhang et al., 2020; Radford et al., 2021).

Intuitively, we aim to simultaneously learn a molecular encoder f^{mol} and a morphology encoder f^{morph} by minimizing the modified InfoNCE loss. Specifically, the pretraining dataset consists of N molecule-morphology pairs, defined

as $\{(\mathbf{x}_i^{mol}, \mathbf{x}_i^{morph}) \mid i \in \{1, ..., N\}\}$. The i-th molecule-morphology pair $\mathbf{x}_i^{mol}$ and $\mathbf{x}_i^{morph}$ are first encoded by their corresponding encoders f^{mol} and f^{morph} to produce their respective representations

$$\mathbf{h}_i^{mol} = f^{mol}(\mathbf{x}_i^{mol})$$
$$\mathbf{h}_i^{morph} = f^{morph}(\mathbf{x}_i^{morph})$$

.

where $\mathbf{h}_i^{mol} \in \mathbb{R}^{d^{mol}}$ and $\mathbf{h}_i^{morph} \in \mathbb{R}^{d^{morph}}$ are the encoded representations of $\mathbf{x}_i^{mol}$ and $\mathbf{x}_i^{morph}$. Each encoder representation is transformed using projection functions g following

$$\mathbf{u}_i^{mol} = g^{mol}(\mathbf{h}_i^{mol})$$
$$\mathbf{u}_i^{morph} = g^{morph}(\mathbf{h}_i^{morph})$$

where $\mathbf{u}_i^{mol} \in \mathbb{R}^{proj}$ and $\mathbf{u}_i^{morph} \in \mathbb{R}^{proj}$ are vectors in the muli-modal embedding space. During training, f^{mol}, f^{morph}, g^{mol}, and g^{morph} are jointly optimized to minimize the loss function

$$\mathcal{L} = \alpha \cdot \mathcal{L}_{mol \to morph} + (1 - \alpha) \cdot \mathcal{L}_{morph \to mol}$$

where α is a weighting term and $\mathcal{L}_{mol \to morph}$ and $\mathcal{L}_{morph \to mol}$ are molecule- and morphology-specific InfoNCE losses, defined as

$$\mathcal{L}_{mol \to morph} = \frac{1}{N} \sum_{i=1}^{N} \log \frac{e^{\langle \mathbf{u}_i^{mol}, \mathbf{u}_i^{morph} \rangle / \tau}}{\sum_{k=1}^{N} e^{\langle \mathbf{u}_i^{mol}, \mathbf{u}_k^{morph} \rangle}}$$

$$\mathcal{L}_{morph \to mol} = \frac{1}{N} \sum_{i=1}^{N} \log \frac{e^{\langle \mathbf{u}_i^{mol}, \mathbf{u}_i^{morph} \rangle / \tau}}{\sum_{k=1}^{N} e^{\langle \mathbf{u}_k^{mol}, \mathbf{u}_i^{morph} \rangle}}$$

with $\langle \mathbf{u}, \mathbf{v} \rangle$ denoting the cosine similarity between vectors $\mathbf{u}$ and $\mathbf{v}$, and τ denotes a temperature scaling parameter.

Minimizing $\mathcal{L}$ produces encoders f^{mol} and f^{morph} that maximally preserve the mutual information between representations $\mathbf{h}_i^{mol}$ and $\mathbf{h}_i^{morph}$. The resulting f^{mol} is then fine-tuned on downstream tasks for transfer learning.

3. Methods

JUMP-CP dataset We use a subset of the dataset *cpg0016-jump*, available from the Cell Painting Gallery on the Registry of Open Data on AWS (*https://registry.opendata.aws/cellpainting-gallery/*) as part of the JUMP-CP Consortium (Chandrasekaran et al., 2023). This subset (as of February 2023) contains approximately 700K morphological profiles of 120K compounds in U2OS cells collected across 12 data generating centers.

Throughout our experiments, we use the precomputed well-level profiles provided with JUMP-CP. Each feature in a well-level profile is scaled independently using median and interquartile range statistics of the plate that the well belongs to. More concretely, the i-th feature of profile $x \in \mathbb{R}^d$ belonging to plate p – denoted as $x_{i,p}$ – is preprocessed as followed

$$x_{i,p}^{processed} = \frac{x_{i,p}^{raw} - med(X_{i,p})}{IQR(X_{i,p})}$$

Where $x_{i,p}^{raw}$ denotes the raw feature value, $X_{i,p}$ denotes the vector of all i-th features in plate p, and med and IQR denote the median and interquartile range.

We follow Way et al. (2021) and remove features with low variance, features with extreme outlier values, and any blacklisted CellProfiler features that are known to be noisy unreliable (Way, 2019). This results in the final set of 3,475 features.

ChEMBL20 dataset We use the ChEMBL20 dataset processed by Mayr et al. (2018) to evaluate transfer learning. The dataset has been used extensively to evaluate and benchmark machine learning approaches for QSAR modeling (Wu et al., 2018; Yang et al., 2019; Nguyen et al., 2020). In short, the dataset consists of approximately 450K compounds, each with sparse annotations of 1,310 binary downstream tasks spanning ADME, toxicity, physicochemical, binding, and functional.

Internal GSK pharmacokinetic dataset Internal rodent in vitro metabolism data were collated from four different intrinsic clearance assay protocols: rat liver microsomes (CL_{int}^{RLM}), mouse liver microsomes (CL_{int}^{MLM}), rat hepatocytes (CL_{int}^{RH}), and mouse hepatocytes (CL_{int}^{MH}). We convert all readouts to intrinsic clearance based on percent hepatic blood flow (PHBF) and aggregate replicate experiments for the same compound and protocol by taking the median reported PHBF. This yielded a dataset of 105,172 unique compounds with available data across all four endpoints. Finally, the data is binarized based on the median PHBF value per endpoint.

Contrastive pretraining procedure Following notations from Section 2, f_{mol} and f_{morph} are a GGNN and a feed-forward neural network (FFNN), respectively, while both g_{mol} and g_{morph} are single feedforward layers. Following Zhang et al. (2020), g_{mol} and g_{morph} are non-linear transformations utilizing ReLU as the activation function.

The model is trained for 1,000 epochs – approximately 400,000 steps – with a batch size of 256 on approximately 100K of the 120K compounds and 600K of the 700K morphological profiles. We follow the protocol proposed by

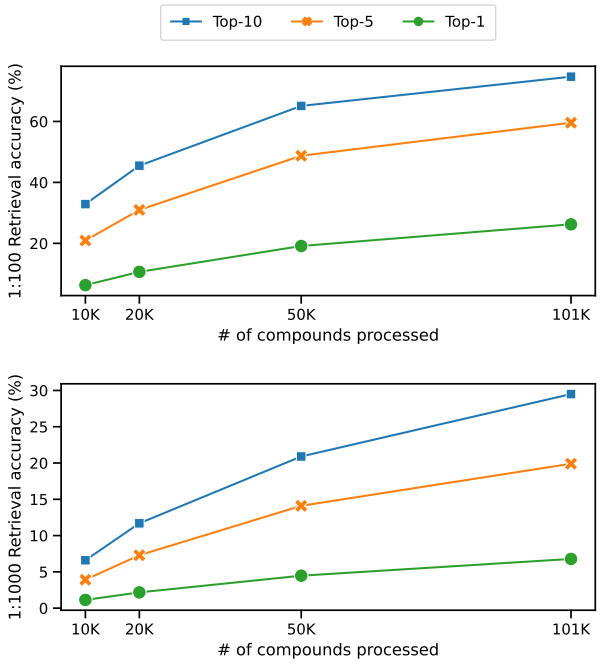

Figure 2. Molecule and morphology retrieval performance at positive-to-negative sampling ratio of 1:100 (top) and 1:1000 (bottom) using MoCoP trained with increasing number of compounds in JUMP-CP. Average top-k accuracy of retrieving molecule given morphology and vice versa is reported for $k \in \{1, 5, 10\}$ for each sampling ratio.

CLIP (Radford et al., 2021) and OpenCLIP (Cherti et al., 2022) to use the AdamW optimizer (Loshchilov & Hutter, 2019) with a learning rate of 10^{-3} and cosine annealing learning rate scheduler with 50 warm-up epochs. MoCoP hyperparameters are further detailed in Appendix B.1.

Transfer learning We explore two transfer learning strategies for MoCoP: linear probe and fine-tuning whole model, which we refer to as MoCoP-LP and MoCoP-FT respectively. We use the Adam optimizer (Kingma & Ba, 2017) with a learning rate of 5×10^{-5} and a batch size of 128 for both strategies.

Baselines We include two baselines: training from scratch and fine-tuning from GGNNs pretrained with multitask supervised learning, which we refer to as FS and Multitask-FT, respectively.

Hyperparameter optimization is performed to ensure FS baseline is competitive. Specifically we use ChEMBL 5% and down-sampled GSK phamacokinetic datasets to carry out a random search consisting of 50 parallel trials spanning the search space described in Appendix A to maximize validation performance. The down-sampling procedure is detailed in Section 4.

For Multitask-FT, we first pretrain GGNNs to directly predict morphological profiles in a multi-task regression setting. Pretraining hyperparameters are optimized using random search consisting of 20 trials while fine-tuning hyperparameters are hand-tuned for performances on validation set of ChEMBL 5%.

4. Experimental Results and Discussion

Scaling MoCoP to JUMP-CP We first evaluate if MoCoP is feasible with the JUMP-CP dataset following procedure detailed in Section 3. Similar approaches have been previously carried out on smaller datasets collected at a single site (Sanchez-Fernandez et al., 2022; Zheng et al., 2022), and the aim is to test its scalability on a larger and multi-site dataset. To evaluate the pretraining performance, the accuracy of molecule and morphology retrieval is measured. Specifically, the average top-k accuracy – where k can be 1, 5, or 10 – of retrieving molecule given morphology and vice versa is reported. The positive-to-negative sampling ratio is set to 1:100 and 1:1000.

Shown in Figure 2, the performance of pretraining improves as more compounds are included in the training process. The trend continues even beyond the maximum of 101K compounds, indicating pretraining can further benefit from obtaining more data. This observation highlights the importance of large public repositories of cellular imaging data. Additionally, we present training and validation curves in Appendix B.2, which demonstrates a stable and convergent training process.

Moreover, we have not extensively explored preprocessing pipelines for morphological profiles, and we anticipate that employing more advanced approaches to mitigate batch effects could improve performance.

Transfer learning performances on ChEMBL20 We aim to evaluate the quality of pretrained GGNN molecule encoder by using ChEMBL20 as the downstream task. Random splits based on compounds are carried out at an 80/10/10 ratio for training, validation, and test sets. For each split, we further subsample 1%, 5%, 10%, and 25%, and 50% of the training set to simulate an increasingly sparse data regime.

Table 1 shows transfer learning performance on ChEMBL20. We report performance averaged across all tasks following existing works utilizing this dataset (Mayr et al., 2018; Wu et al., 2018; Yang et al., 2019). Our results indicate that fine-tuning GGNNs pretrained with MoCoP (MoCoP-FT) consistently outperformed training-from-scatch (FS) baseline across all data regimes. This improvement is also observed by simply applying a linear probe on the frozen molecule encoder (MoCoP-LP). We also observe that MoCoP-LP out-

Table 1. Performance on held-out test sets of different subsets of ChEMBL20 averaged across all tasks. FS baseline: GGNNs trained from scratch; Multitask-FT baseline: Fine-tuning GGNNs pretrained using multitask supervised learning and fine-tuned; MoCoP-LP: Linear probe on GGNNs pretrained with MoCoP; MoCoP-FT: Fine-tuning GGNNs pretrained with MoCoP. Mean and standard deviation are obtained from 9 repeats from 3 splits and 3 seeds (see Section 3 for details). The best and second best values are in bold and regular text, respectively.

METRIC	DATASET	FS	MULTITASK-FT	MOCOP-LP	MOCOP-FT
AUROC	CHEMBL20 - 1%	0.511 ± 0.008	0.508 ± 0.007	$\mathbf{0.545 \pm 0.017}$	0.542 ± 0.010
	CHEMBL20 - 5%	0.571 ± 0.010	0.574 ± 0.004	$\mathbf{0.624 \pm 0.018}$	0.621 ± 0.022
	CHEMBL20 - 10%	0.597 ± 0.014	0.588 ± 0.009	0.638 ± 0.017	$\mathbf{0.646 \pm 0.021}$
	CHEMBL20 - 25%	0.648 ± 0.017	0.643 ± 0.020	0.678 ± 0.015	$\mathbf{0.689 \pm 0.018}$
	CHEMBL20 - 50%	0.669 ± 0.016	—	—	$\mathbf{0.693 \pm 0.030}$
	CHEMBL20 - 100%	0.706 ± 0.022	—	—	$\mathbf{0.721 \pm 0.020}$
AUPRC	CHEMBL20 - 1%	0.487 ± 0.013	0.482 ± 0.015	$\mathbf{0.511 \pm 0.024}$	0.510 ± 0.016
	CHEMBL20 - 5%	0.528 ± 0.010	0.525 ± 0.013	$\mathbf{0.576 \pm 0.026}$	0.569 ± 0.023
	CHEMBL20 - 10%	0.550 ± 0.022	0.539 ± 0.023	0.588 ± 0.032	$\mathbf{0.597 \pm 0.036}$
	CHEMBL20 - 25%	0.600 ± 0.028	0.595 ± 0.026	0.623 ± 0.027	$\mathbf{0.640 \pm 0.031}$
	CHEMBL20 - 50%	0.623 ± 0.026	—	—	$\mathbf{0.654 \pm 0.037}$
	CHEMBL20 - 100%	0.662 ± 0.033	—	—	$\mathbf{0.681 \pm 0.033}$

Table 2. Performance on held-out test sets of GSK internal pharmacokinetic data. Mean and standard deviation are obtained from 9 repeats from 3 splits and 3 seeds (see Section 3 for details). The best values are in bold text.

METRIC	DATASET	FS	MOCOP-FT
AUROC	CL_{int}^{RH}	0.762 ± 0.008	$\mathbf{0.788 \pm 0.014}$
	CL_{int}^{MH}	0.763 ± 0.031	$\mathbf{0.791 \pm 0.026}$
	CL_{int}^{RLM}	0.845 ± 0.011	$\mathbf{0.864 \pm 0.013}$
	CL_{int}^{MLM}	0.839 ± 0.018	$\mathbf{0.852 \pm 0.024}$
	AVERAGE	0.802 ± 0.013	$\mathbf{0.824 \pm 0.014}$
AUPRC	CL_{int}^{RH}	0.760 ± 0.023	$\mathbf{0.790 \pm 0.030}$
	CL_{int}^{MH}	0.775 ± 0.030	$\mathbf{0.795 \pm 0.031}$
	CL_{int}^{RLM}	0.851 ± 0.006	$\mathbf{0.870 \pm 0.004}$
	CL_{int}^{MLM}	0.831 ± 0.009	$\mathbf{0.845 \pm 0.014}$
	AVERAGE	0.804 ± 0.011	$\mathbf{0.825 \pm 0.014}$

Table 3. Performance on held-out test sets of GSK internal pharmacokinetic data with down-sampled training data. Mean and standard deviation are obtained from 9 repeats from 3 splits and 3 seeds (see Section 3 for details). The best values are in bold text.

METRIC	DATASET	FS	MOCOP-FT
AUROC	CL_{int}^{RH}	0.716 ± 0.046	$\mathbf{0.763 \pm 0.057}$
	CL_{int}^{MH}	0.716 ± 0.056	$\mathbf{0.805 \pm 0.049}$
	CL_{int}^{RLM}	0.800 ± 0.011	$\mathbf{0.824 \pm 0.018}$
	CL_{int}^{MLM}	0.779 ± 0.015	$\mathbf{0.805 \pm 0.023}$
	AVERAGE	0.752 ± 0.028	$\mathbf{0.799 \pm 0.033}$
AUPRC	CL_{int}^{RH}	0.715 ± 0.053	$\mathbf{0.768 \pm 0.049}$
	CL_{int}^{MH}	0.710 ± 0.044	$\mathbf{0.799 \pm 0.046}$
	CL_{int}^{RLM}	0.820 ± 0.011	$\mathbf{0.842 \pm 0.018}$
	CL_{int}^{MLM}	0.818 ± 0.019	$\mathbf{0.846 \pm 0.027}$
	AVERAGE	0.766 ± 0.025	$\mathbf{0.814 \pm 0.031}$

performs MoCoP-FT in lower data regime. Notably, we encounter challenges with Multitask-FT, in which GGNNs are first trained to directly predict morphological features in a multi-task regression setting. This approach fails to produce any improvements over FS baseline. Our finding is consistent with previous research that highlights the superior learning efficiency of contrastive objectives over predictive objectives.(Chen et al., 2020; Tian et al., 2020; Radford et al., 2021).

Transfer learning performances on internal GSK pharmacokinetic data The quality of pretrained GGNNs is further evaluated using a subset of GSK internal pharmacokinetic data as downstream tasks. This dataset consists of 4 tasks as detailed in Section 3. Unlike the previous experiment with ChEMBL20, here we employ scaffold splitting, which has been shown to provide better estimates of model performances in QSAR tasks (Kearnes et al., 2017; Wu et al., 2018). The compounds are first clustered using the Butina algorithm implemented in RDKit with a Euclidean distance function and a distance cutoff of 0.6. The clusters are ordered by size, and for every of six clusters, four are assigned to the training set, one to the validation set, and one to the test set. The procedure is repeated with random cluster ordering to create two additional splits. For each split, a down-sampled version is created randomly selecting a single compound from each cluster to uniformly sample the chemical space in our dataset.

Using results from the previous experiment, we benchmark the most performant approach MoCoP-FT, where each model is repeated 9 times with 3 splits and 3 seeds. We

again observe that MoCoP-FT consistently outperforms FS baseline across both full and down-sampled datasets, shown in Table 2 and 3, respectively. On the full dataset, pretrained GGNNs show an average improvement of 2.6% in AUPRC across the 4 individual tasks. This effect is increased to 6.3% in AUPRC when less data is available for training. We expect performance can be further improved by considering using related endpoints as descriptors, as demonstrated by Broccatelli et al. (2022).

This result offers a glimpse at the potential of using datasets not directly related to the learning task at hand in improving QSAR models. While the results in this study are limited to a single publicly available high-content imaging dataset, other high-dimensional readouts such as transcriptomics and proteomics can be used to augment QSAR modeling in similar manners. Further investigation of routine re-use of high-dimensional data in standard QSAR workflows is warranted in future works.

5. Conclusion

In this study, we explore MoCoP as a means to improve the performance of QSAR models. We scale MoCoP to approximately 100K molecules and 600K morphological profiles, and evaluate pretrained GGNNs molecule encoder on both public and internal downstream tasks.

Our results demonstrate that MoCoP consistently improves the performance of GGNNs in QSAR tasks, especially in low-data regimes when compared to training-from-scratch and multitask supervised pretraining baselines. We observe this trend in both the ChEMBL20 dataset and GSK internal pharmacokinetic data, indicating that the approach is applicable across a range of datasets and tasks.

Our work also suggests that data from unbiased high-dimensional assays, beyond cellular imaging, can improve QSAR models via contrastive pretraining. Future works will further explore this approach with other data sources such as transcriptomics and proteomics. Overall, we believe our work can be combined with existing methods to improve model performances and expands the deep learning toolbox available for QSAR applications.

References

Altae-Tran, H., Ramsundar, B., Pappu, A. S., and Pande, V. Low Data Drug Discovery with One-Shot Learning. *ACS Central Science*, 3(4):283–293, April 2017. ISSN 2374-7943. doi: 10.1021/acscentsci.6b00367. Publisher: American Chemical Society.

Ando, D. M., McLean, C. Y., and Berndl, M. Improving Phenotypic Measurements in High-Content Imaging Screens, July 2017.

Bray, M.-A., Singh, S., Han, H., Davis, C. T., Borgeson, B., Hartland, C., Kost-Alimova, M., Gustafsdottir, S. M., Gibson, C. C., and Carpenter, A. E. Cell Painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes. *Nature Protocols*, 11 (9):1757–1774, September 2016. ISSN 1750-2799. doi: 10.1038/nprot.2016.105.

Bray, M.-A., Gustafsdottir, S. M., Rohban, M. H., Singh, S., Ljosa, V., Sokolnicki, K. L., Bittker, J. A., Bodycombe, N. E., Dančík, V., Hasaka, T. P., Hon, C. S., Kemp, M. M., Li, K., Walpita, D., Wawer, M. J., Golub, T. R., Schreiber, S. L., Clemons, P. A., Shamji, A. F., and Carpenter, A. E. A dataset of images and morphological profiles of 30 000 small-molecule treatments using the Cell Painting assay. *GigaScience*, 6(12):giw014, December 2017. ISSN 2047-217X. doi: 10.1093/gigascience/giw014.

Broccatelli, F., Trager, R., Reutlinger, M., Karypis, G., and Li, M. Benchmarking Accuracy and Generalizability of Four Graph Neural Networks Using Large In Vitro ADME Datasets from Different Chemical Spaces. *Molecular Informatics*, 41(8):2100321, 2022. ISSN 1868-1751. doi: 10.1002/minf.202100321.

Carhart, R. E., Smith, D. H., and Venkataraghavan, R. Atom pairs as molecular features in structure-activity studies: definition and applications. *Journal of Chemical Information and Computer Sciences*, 25(2):64–73, May 1985. ISSN 0095-2338. doi: 10.1021/ci00046a002. Publisher: American Chemical Society.

Chandrasekaran, S. N., Ceulemans, H., Boyd, J. D., and Carpenter, A. E. Image-based profiling for drug discovery: due for a machine-learning upgrade? *Nature Reviews Drug Discovery*, 20(2):145–159, February 2021. ISSN 1474-1784. doi: 10.1038/s41573-020-00117-w. Number: 2 Publisher: Nature Publishing Group.

Chandrasekaran, S. N., Ackerman, J., Alix, E., Ando, D. M., Arevalo, J., Bennion, M., Boisseau, N., Borowa, A., Boyd, J. D., Brino, L., Byrne, P. J., Ceulemans, H., Ch'ng, C., Cimini, B. A., Clevert, D.-A., Deflaux, N., Doench, J. G., Dorval, T., Doyonnas, R., Dragone, V., Engkvist, O., Faloon, P. W., Fritchman, B., Fuchs, F., Garg, S., Gilbert, T. J., Glazer, D., Gnutt, D., Goodale, A., Grignard, J., Guenther, J., Han, Y., Hanifehlou, Z., Hariharan, S., Hernandez, D., Horman, S. R., Hormel, G., Huntley, M., Icke, I., Iida, M., Jacob, C. B., Jaensch, S., Khetan, J., Kost-Alimova, M., Krawiec, T., Kuhn, D., Lardeau, C.-H., Lembke, A., Lin, F., Little, K. D., Lofstrom, K. R., Lotfi, S., Logan, D. J., Luo, Y., Madoux, F., Zapata, P. A. M., Marion, B. A., Martin, G., McCarthy, N. J., Mervin, L., Miller, L., Mohamed, H., Monteverde, T., Mouchet, E., Nicke, B., Ogier, A., Ong, A.-L., Osterland, M., Otrocka, M., Peeters, P. J., Pilling, J., Prechtl, S., Qian, C., Rataj,

K., Root, D. E., Sakata, S. K., Scrace, S., Shimizu, H., Simon, D., Sommer, P., Spruiell, C., Sumia, I., Swalley, S. E., Terauchi, H., Thibaudeau, A., Unruh, A., Waeter, J. V. d., Dyck, M. V., Staden, C. v., Warchoł, M., Weisbart, E., Weiss, A., Wiest-Daessle, N., Williams, G., Yu, S., Zapiec, B., Żyła, M., Singh, S., and Carpenter, A. E. JUMP Cell Painting dataset: morphological impact of 136,000 chemical and genetic perturbations, March 2023.

Chen, Y.-C., Li, L., Yu, L., Kholy, A. E., Ahmed, F., Gan, Z., Cheng, Y., and Liu, J. UNITER: UNiversal Image-TExt Representation Learning, July 2020. arXiv:1909.11740 [cs].

Cherti, M., Beaumont, R., Wightman, R., Wortsman, M., Ilharco, G., Gordon, C., Schuhmann, C., Schmidt, L., and Jitsev, J. Reproducible scaling laws for contrastive language-image learning, December 2022. arXiv:2212.07143 [cs].

Cuccarese, M. F., Earnshaw, B. A., Heiser, K., Fogelson, B., Davis, C. T., McLean, P. F., Gordon, H. B., Skelly, K.-R., Weathersby, F. L., Rodic, V., Quigley, I. K., Pastuzyn, E. D., Mendivil, B. M., Lazar, N. H., Brooks, C. A., Carpenter, J., Probst, B. L., Jacobson, P., Glazier, S. W., Ford, J., Jensen, J. D., Campbell, N. D., Statnick, M. A., Low, A. S., Thomas, K. R., Carpenter, A. E., Hegde, S. S., Alfa, R. W., Victors, M. L., Haque, I. S., Chong, Y. T., and Gibson, C. C. Functional immune mapping with deep-learning enabled phenomics applied to immunomodulatory and COVID-19 drug discovery. Technical report, bioRxiv, August 2020. Section: New Results Type: article.

Desai, K. and Johnson, J. VirTex: Learning Visual Representations from Textual Annotations, September 2021. arXiv:2006.06666 [cs].

Fay, M. M., Kraus, O., Victors, M., Arumugam, L., Vuggumudi, K., Urbanik, J., Hansen, K., Celik, S., Cernek, N., Jagannathan, G., Christensen, J., Earnshaw, B. A., Haque, I. S., and Mabey, B. RxRx3: Phenomics Map of Biology, February 2023. Pages: 2023.02.07.527350 Section: New Results.

Fürst, A., Rumetshofer, E., Lehner, J., Tran, V., Tang, F., Ramsauer, H., Kreil, D., Kopp, M., Klambauer, G., Bitto-Nemling, A., and Hochreiter, S. CLOOB: Modern Hopfield Networks with InfoLOOB Outperform CLIP, November 2022. arXiv:2110.11316 [cs].

Gaulton, A., Bellis, L. J., Bento, A. P., Chambers, J., Davies, M., Hersey, A., Light, Y., McGlinchey, S., Michalovich, D., Al-Lazikani, B., and Overington, J. P. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Research*, 40(D1):D1100–D1107, January 2012. ISSN 0305-1048. doi: 10.1093/nar/gkr777.

Godinez, W. J., Hossain, I., and Zhang, X. Unsupervised phenotypic analysis of cellular images with multi-scale convolutional neural networks, July 2018.

Hofmarcher, M., Rumetshofer, E., Clevert, D.-A., Hochreiter, S., and Klambauer, G. Accurate Prediction of Biological Assays with High-Throughput Microscopy Images and Convolutional Networks. *Journal of Chemical Information and Modeling*, 59(3):1163–1171, March 2019. ISSN 1549-9596. doi: 10.1021/acs.jcim.8b00670. Publisher: American Chemical Society.

Hu, W., Liu, B., Gomes, J., Zitnik, M., Liang, P., Pande, V., and Leskovec, J. Strategies for Pre-training Graph Neural Networks, February 2020. arXiv:1905.12265 [cs, stat].

Hu, W., Fey, M., Zitnik, M., Dong, Y., Ren, H., Liu, B., Catasta, M., and Leskovec, J. Open Graph Benchmark: Datasets for Machine Learning on Graphs, February 2021. arXiv:2005.00687 [cs, stat].

Kearnes, S., McCloskey, K., Berndl, M., Pande, V., and Riley, P. Molecular graph convolutions: moving beyond fingerprints. *Journal of Computer-Aided Molecular Design*, 30(8):595–608, August 2016. ISSN 1573-4951. doi: 10.1007/s10822-016-9938-8.

Kearnes, S., Goldman, B., and Pande, V. Modeling Industrial ADMET Data with Multitask Networks, January 2017. arXiv:1606.08793 [stat].

Kingma, D. P. and Ba, J. Adam: A Method for Stochastic Optimization, January 2017. arXiv:1412.6980 [cs].

Kraus, O. Z., Grys, B. T., Ba, J., Chong, Y., Frey, B. J., Boone, C., and Andrews, B. J. Automated analysis of high-content microscopy data with deep learning. *Molecular Systems Biology*, 13(4):924, April 2017. ISSN 1744-4292. doi: 10.15252/msb.20177551.

Kurita, K. L. and Linington, R. G. Connecting Phenotype and Chemotype: High-Content Discovery Strategies for Natural Products Research. *Journal of Natural Products*, 78(3):587–596, March 2015. ISSN 0163-3864. doi: 10.1021/acs.jnatprod.5b00017. Publisher: American Chemical Society.

Loshchilov, I. and Hutter, F. Decoupled Weight Decay Regularization, January 2019. arXiv:1711.05101 [cs, math].

Mayr, A., Klambauer, G., Unterthiner, T., Steijaert, M., Wegner, J. K., Ceulemans, H., Clevert, D.-A., and Hochreiter, S. Large-scale comparison of machine learning methods for drug target prediction on ChEMBL. *Chemical Science*, 9(24):5441–5451, June 2018. ISSN 2041-6539. doi: 10.1039/C8SC00148K.

Nguyen, C. Q., Kreatsoulas, C., and Branson, K. M. Meta-Learning GNN Initializations for Low-Resource Molecular Property Prediction, July 2020. arXiv:2003.05996 [physics, stat].

Nilakantan, R., Bauman, N., Dixon, J. S., and Venkataraghavan, R. Topological torsion: a new molecular descriptor for SAR applications. Comparison with other descriptors. *Journal of Chemical Information and Computer Sciences*, 27(2):82–85, May 1987. ISSN 0095-2338. doi: 10.1021/ci00054a008. Publisher: American Chemical Society.

Oord, A. v. d., Li, Y., and Vinyals, O. Representation Learning with Contrastive Predictive Coding, January 2019. arXiv:1807.03748 [cs, stat].

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning Transferable Visual Models From Natural Language Supervision, February 2021. arXiv:2103.00020 [cs].

Ramsundar, B., Kearnes, S., Riley, P., Webster, D., Konerding, D., and Pande, V. Massively Multitask Networks for Drug Discovery, February 2015. arXiv:1502.02072 [cs, stat].

Reker, D. and Schneider, G. Active-learning strategies in computer-assisted drug discovery. *Drug Discovery Today*, 20(4):458–465, April 2015. ISSN 1359-6446. doi: 10.1016/j.drudis.2014.12.004.

Rogers, D. and Hahn, M. Extended-Connectivity Fingerprints. *Journal of Chemical Information and Modeling*, 50(5):742–754, May 2010. ISSN 1549-9596. doi: 10.1021/ci100050t. Publisher: American Chemical Society.

Sanchez-Fernandez, A., Rumetshofer, E., Hochreiter, S., and Klambauer, G. Contrastive learning of image- and structure-based representations in drug discovery. May 2022.

Schaller, D., Šribar, D., Noonan, T., Deng, L., Nguyen, T. N., Pach, S., Machalz, D., Bermudez, M., and Wolber, G. Next generation 3D pharmacophore modeling. *WIREs Computational Molecular Science*, 10(4):e1468, 2020. ISSN 1759-0884. doi: 10.1002/wcms.1468. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/wcms.1468.

Sheridan, R. P., Miller, M. D., Underwood, D. J., and Kearsley, S. K. Chemical Similarity Using Geometric Atom Pair Descriptors. *Journal of Chemical Information and Computer Sciences*, 36(1):128–136, January 1996. ISSN 0095-2338. doi: 10.1021/ci950275b. Publisher: American Chemical Society.

Smith, J. S., Nebgen, B., Lubbers, N., Isayev, O., and Roitberg, A. E. Less is more: sampling chemical space with active learning. *The Journal of Chemical Physics*, 148 (24):241733, June 2018. ISSN 0021-9606, 1089-7690. doi: 10.1063/1.5023802. arXiv:1801.09319 [physics, stat].

Stirling, D. R., Swain-Bowden, M. J., Lucas, A. M., Carpenter, A. E., Cimini, B. A., and Goodman, A. CellProfiler 4: improvements in speed, utility and usability. *BMC Bioinformatics*, 22(1):433, September 2021. ISSN 1471-2105. doi: 10.1186/s12859-021-04344-9.

Sypetkowski, M., Rezanejad, M., Saberian, S., Kraus, O., Urbanik, J., Taylor, J., Mabey, B., Victors, M., Yosinski, J., Sereshkeh, A. R., Haque, I., and Earnshaw, B. RxRx1: A Dataset for Evaluating Experimental Batch Correction Methods, January 2023. arXiv:2301.05768 [cs].

Tian, Y., Krishnan, D., and Isola, P. Contrastive Multiview Coding, December 2020. arXiv:1906.05849 [cs].

Way, G. P. Blocklist Features - Cell Profiler. November 2019. doi: 10.6084/m9.figshare.10255811.v3. Type: dataset.

Way, G. P., Kost-Alimova, M., Shibue, T., Harrington, W. F., Gill, S., Piccioni, F., Becker, T., Shafqat-Abbasi, H., Hahn, W. C., Carpenter, A. E., Vazquez, F., and Singh, S. Predicting cell health phenotypes using image-based morphology profiling. *Molecular Biology of the Cell*, 32(9):995–1005, April 2021. ISSN 1059-1524. doi: 10.1091/mbc.E20-12-0784.

Wu, Z., Ramsundar, B., Feinberg, E. N., Gomes, J., Geniesse, C., Pappu, A. S., Leswing, K., and Pande, V. MoleculeNet: a benchmark for molecular machine learning. *Chemical Science*, 9(2):513–530, January 2018. ISSN 2041-6539. doi: 10.1039/C7SC02664A.

Xu, Y., Ma, J., Liaw, A., Sheridan, R. P., and Svetnik, V. Demystifying Multitask Deep Neural Networks for Quantitative Structure–Activity Relationships. *Journal of Chemical Information and Modeling*, 57(10):2490–2504, October 2017. ISSN 1549-9596. doi: 10.1021/acs.jcim.7b00087. Publisher: American Chemical Society.

Yang, K., Swanson, K., Jin, W., Coley, C., Eiden, P., Gao, H., Guzman-Perez, A., Hopper, T., Kelley, B., Mathea, M., Palmer, A., Settels, V., Jaakkola, T., Jensen, K., and Barzilay, R. Analyzing Learned Molecular Representations for Property Prediction. *Journal of Chemical Information and Modeling*, 59(8):3370–3388, August 2019. ISSN 1549-9596. doi: 10.1021/acs.jcim.9b00237.

Zhang, Y., Jiang, H., Miura, Y., Manning, C. D., and Langlotz, C. P. Contrastive Learning of Medical Visual Representations from Paired Images and Text, October 2020. arXiv:2010.00747 [cs] version: 1.

Zheng, S., Rao, J., Zhang, J., Cohen, E., Li, C., and Yang, Y. Cross-modal Graph Contrastive Learning with Cellular Images, September 2022.

A. FS Baseline Hyperparameter Tuning

Hyperparameter optimization is done on the search space below using a random search consisting of 50 parallel trials. Bold and underscored values denote the selected hyperparameters for ChEMBL20 and pharmacokinetic data, respectively.

HYPERPARAMETER	SEARCH SPACE
LEARNING RATE	$\{10^{-5}, 5\times10^{-5}, 10^{-4}, 5\times10^{-4}, \underline{\mathbf{10^{-3}}}\}$
# OF GGNN LAYERS	4, 5, 6, <u>**7**</u>, 8
BATCH SIZE	64, **128**, <u>256</u>, 512

B. Training MoCoP

B.1. Hyperparameters

MoCoP hyperparameters used in this work are provided in table B.1 below.

HYPERPARAMETER	
# OF GGNN LAYERS IN f^{mol}	6
FF LAYERS DIMENSIONS IN f^{morph}	[512, 256, 128]
d^{mol}	1024
d^{morph}	128
d^{proj}	128
LEARNING RATE	10^{-3}
LEARNING RATE SCHEDULER	COSINE ANNEALING WITH LINEAR WARM-UP
# OF WARM-UP EPOCHS	50
# OF EPOCHS	1,000
BATCH SIZE	256

B.2. Training

We develop a simple sampling procedure to accommodate the one-to-many nature of molecule-to-morphology mapping. Specifically, for each batch of size N, we first randomly select N unique compounds, and for each compound randomly select a single corresponding morphology. We detail the procedure in Algorithm 1.

Algorithm 1 MoCoP Batch Sampling

Input:
Batch size N
Number of unique molecules K
All unique molecules $\mathbf{X}^{mol} = \{\mathbf{x}_i^{mol} \mid i \in \{1, ..., K\}\}$
Mapping of unique molecules to corresponding morphologies $M = \{(\mathbf{x}_i^{mol}, \mathbf{X}_i^{morph}) \mid i \in \{1, ..., K\}\}$

$batch \leftarrow \{\}$
for $i = 1$ **to** N **do**
Sample $\mathbf{x}_i^{mol}$ from $\mathbf{X}^{mol}$
Collect corresponding $\mathbf{X}_i^{morph}$ from mapping M
Sample $\mathbf{x}_i^{morph}$ from $\mathbf{X}_i^{morph}$
$\mathbf{X}^{mol} \leftarrow \mathbf{X}^{mol} \setminus \{\mathbf{x}^{mol}\}$
$batch \leftarrow batch \cup \{(\mathbf{x}_i^{mol}, \mathbf{x}_i^{morph})\}$
end for
Return $batch$

The sampling procedure above produces stable training behaviors for MoCoP, demonstrated in the training and validation

curves in Figure 1. Training is carried out on a single NVIDIA V100 GPU over 6 days.

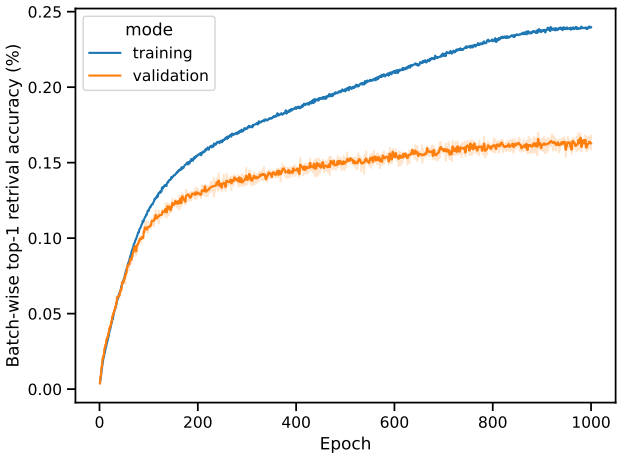

Figure 1. Training and validation curves of MoCoP across 3 different JUMP-CP splits and random initializations. The reported metric is calculated as the average top-1 accuracy for retrieving molecule and morphology in a batch.

C. Effects of Batch Size and d^{proj} on Transfer Learning

A small hyperparameters study was conducted to investigate the effects of batch size and d^{proj} on transfer learning performance using the down-sampled GSK pharmacokinetic dataset as the downstream tasks. We observe that smaller batch size produces more transferable molecule encoder while d^{proj} does not significantly affect overall performances.

METRIC	DATASET	BATCH SIZE = 1024 d^{proj} = 128	BATCH SIZE = 512 d^{proj} = 128	BATCH SIZE = 256 d^{proj} = 128
AUROC	CL_{int}^{RH}	0.747 ± 0.077	0.756 ± 0.053	$\mathbf{0.780 \pm 0.080}$
	CL_{int}^{MH}	0.772 ± 0.045	0.801 ± 0.042	$\mathbf{0.831 \pm 0.066}$
	CL_{int}^{RLM}	0.817 ± 0.008	0.825 ± 0.008	$\mathbf{0.836 \pm 0.030}$
	CL_{int}^{MLM}	0.791 ± 0.013	0.796 ± 0.003	$\mathbf{0.816 \pm 0.039}$
	AVERAGE	0.782 ± 0.032	0.795 ± 0.024	$\mathbf{0.815 \pm 0.053}$
AUPRC	CL_{int}^{RH}	0.741 ± 0.059	0.764 ± 0.048	$\mathbf{0.779 \pm 0.071}$
	CL_{int}^{MH}	0.761 ± 0.013	0.794 ± 0.010	$\mathbf{0.841 \pm 0.062}$
	CL_{int}^{RLM}	0.831 ± 0.002	0.840 ± 0.013	$\mathbf{0.856 \pm 0.027}$
	CL_{int}^{MLM}	0.829 ± 0.024	0.834 ± 0.017	$\mathbf{0.857 \pm 0.034}$
	AVERAGE	0.800 ± 0.023	0.808 ± 0.016	$\mathbf{0.833 \pm 0.046}$

METRIC	DATASET	d^{proj} = 128 BATCH SIZE = 1024	d^{proj} = 256 BATCH SIZE = 1024	d^{proj} = 512 BATCH SIZE = 1024
AUROC	CL_{int}^{RH}	0.747 ± 0.077	$\mathbf{0.745 \pm 0.068}$	$\mathbf{0.737 \pm 0.057}$
	CL_{int}^{MH}	0.772 ± 0.045	$\mathbf{0.819 \pm 0.049}$	$\mathbf{0.819 \pm 0.004}$
	CL_{int}^{RLM}	0.817 ± 0.008	0.817 ± 0.002	$\mathbf{0.819 \pm 0.003}$
	CL_{int}^{MLM}	0.791 ± 0.013	$\mathbf{0.806 \pm 0.009}$	0.803 ± 0.017
	AVERAGE	0.782 ± 0.032	0.794 ± 0.031	$\mathbf{0.815 \pm 0.053}$
AUPRC	CL_{int}^{RH}	0.741 ± 0.059	$\mathbf{0.758 \pm 0.058}$	0.735 ± 0.066
	CL_{int}^{MH}	0.761 ± 0.013	0.774 ± 0.029	$\mathbf{0.803 \pm 0.043}$
	CL_{int}^{RLM}	0.831 ± 0.002	0.833 ± 0.004	$\mathbf{0.837 \pm 0.010}$
	CL_{int}^{MLM}	0.829 ± 0.024	$\mathbf{0.843 \pm 0.015}$	0.840 ± 0.011
	AVERAGE	0.800 ± 0.023	0.802 ± 0.025	$\mathbf{0.803 \pm 0.010}$