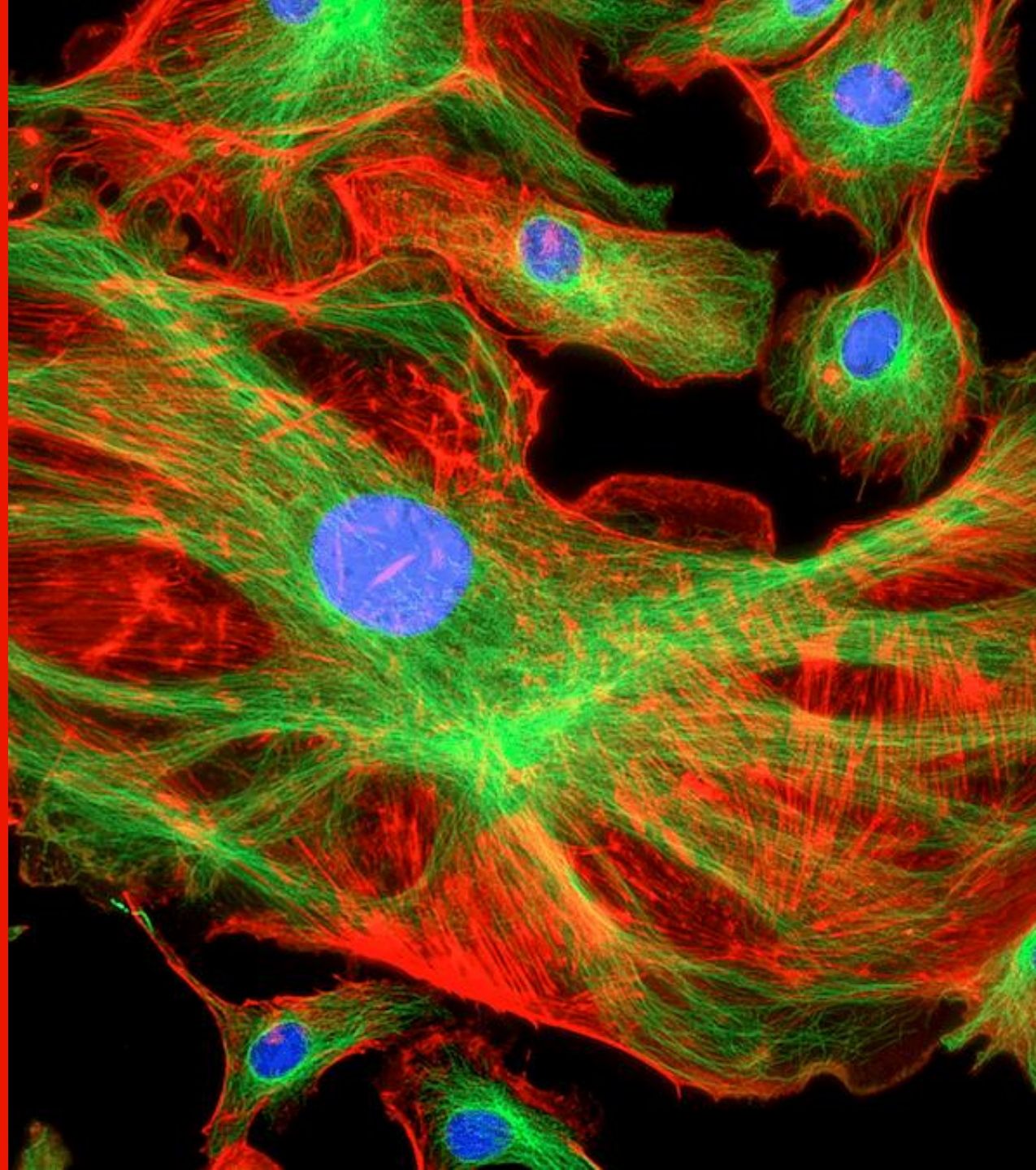
Contrastive Learning of Microscopy Image and Structure-Based Representations of Molecules

Ana Sanchez-Fernandez

AIDD 6th School

J&J

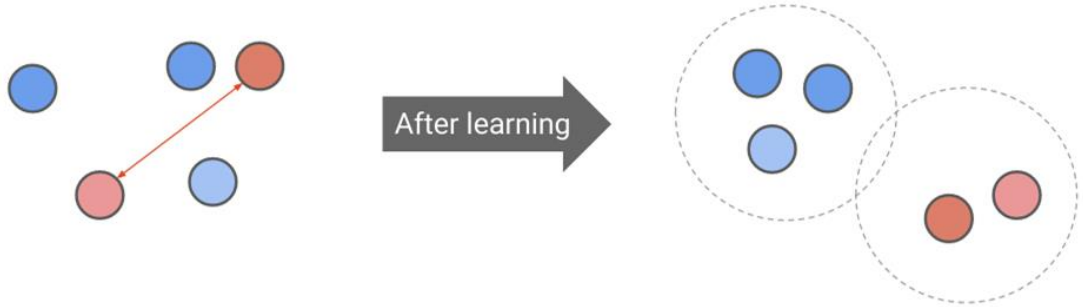


Contents

CLOOME recap

Self-supervised contrastive learning

- **Contrastive learning:** predict relationship between pairs of samples
- Learn an embedding space in which similar ("positive") sample pairs are close to each other and dissimilar ("negative") ones are far apart



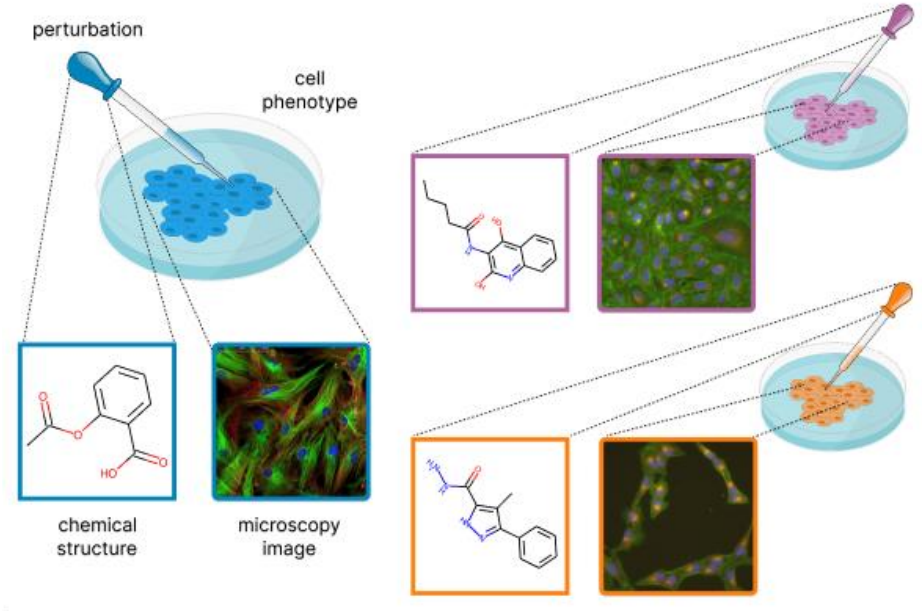
- **InfoNCE objective**

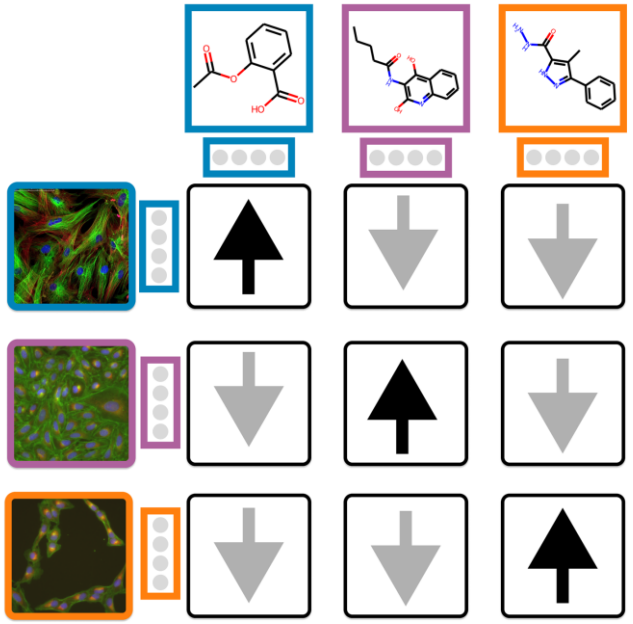
$$\mathrm{L}_{\mathrm{InfoNCE}} = -\ln \frac{\exp(\tau^{-1}\boldsymbol{x}_1^T\boldsymbol{y})}{\exp(\tau^{-1}\boldsymbol{x}_1^T\boldsymbol{y}) + \sum_{j=2}^{N}\exp(\tau^{-1}\boldsymbol{x}_j^T\boldsymbol{y})}$$

Weng, L. and Jong Wook Kim (2021). **Self-supervised learning. Self-prediction and contrastive learning.** [Tutorial presentation]. In *NeurIPS*, 2021.
Liu, X., Zhang, F., Hou, Z., Mian, L., Wang, Z., Zhang, J., Tang, J., and Fellow, I. (2021) **Self-supervised Learning: Generative or Contrastive.** *ArXiv, 1911.05722.*

CLOOME
Contrastive Learning and Leave-One-Out Boost for Molecule Encoders

Learn molecular **representations** with **contrastive** learning using **microscopy** images and molecular **structures**

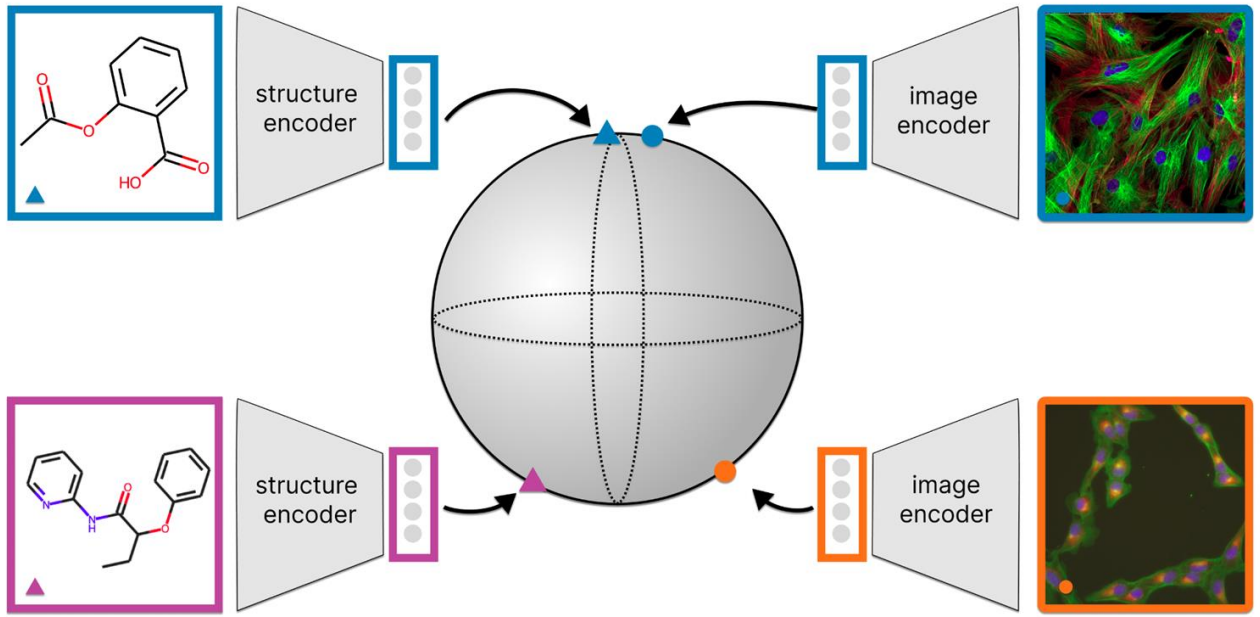




Sanchez-Fernandez, A., Rumetshofer, E., Hochreiter, S. and Klambauer G. **CLOOME: contrastive learning unlocks bioimaging databases for queries with chemical structures.** *Nat Commun* **14**, 7339 (2023).

CLOOME
Contrastive Learning and Leave-One-Out Boost for Molecule Encoders

Learn molecular **representations** with **contrastive** learning using **microscopy** images and molecular **structures**



Sanchez-Fernandez, A., Rumetshofer, E., Hochreiter, S. and Klambauer G. **CLOOME: contrastive learning unlocks bioimaging databases for queries with chemical structures.** *Nat Commun* **14**, 7339 (2023).

Transferability benchmark

—

Ardigen project

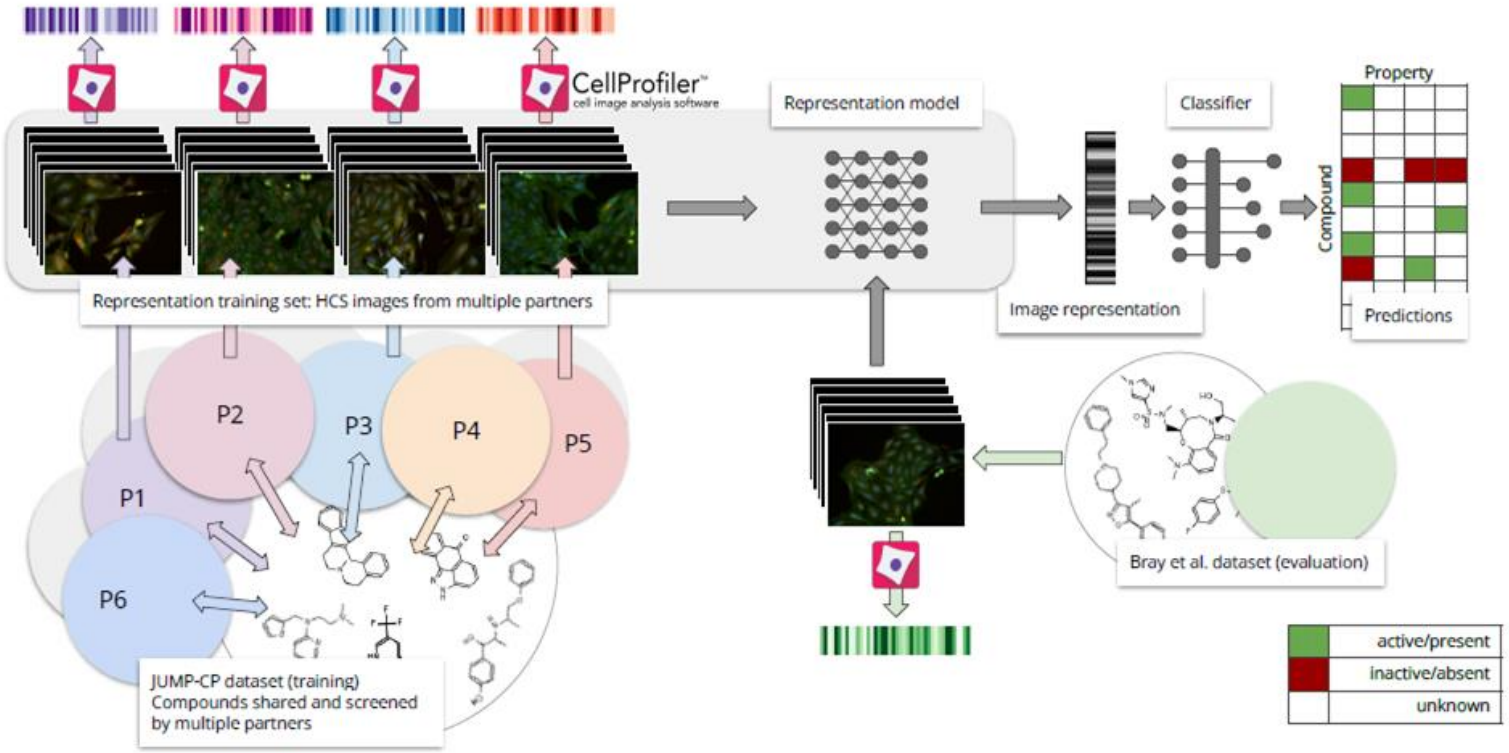
JUMP-CP dataset

- **Large microscopy image dataset**, released by a consortium of 10 pharma and 2 academic partners
- Three different perturbation modalities:
    - Chemical compounds (small molecules)
    - Overexpression of genes
    - Gene knockout by CRISPR
- **120,000 compounds**
    - Public compound structures or could be released by the company
    - High purity (> 90 %)
- Chemically perturbed samples: **3,127,224 images**

Chandrasekaran, S. N., Ackerman, J., Alix, E., Ando, D. M., ... Carpenter, A. E. (2023). **JUMP Cell Painting dataset: morphological impact of 136,000 chemical and genetic perturbations.** *BioRxiv*, 2023.03.23.534023. https://doi.org/10.1101/2023.03.23.534023

Transferability benchmark

Check **transferability** of supervised and self-supervised methods from one dataset (JUMP-CP. Chandrasekaran et al., 2023) to another (CellPainting. Bray et al., 2016)



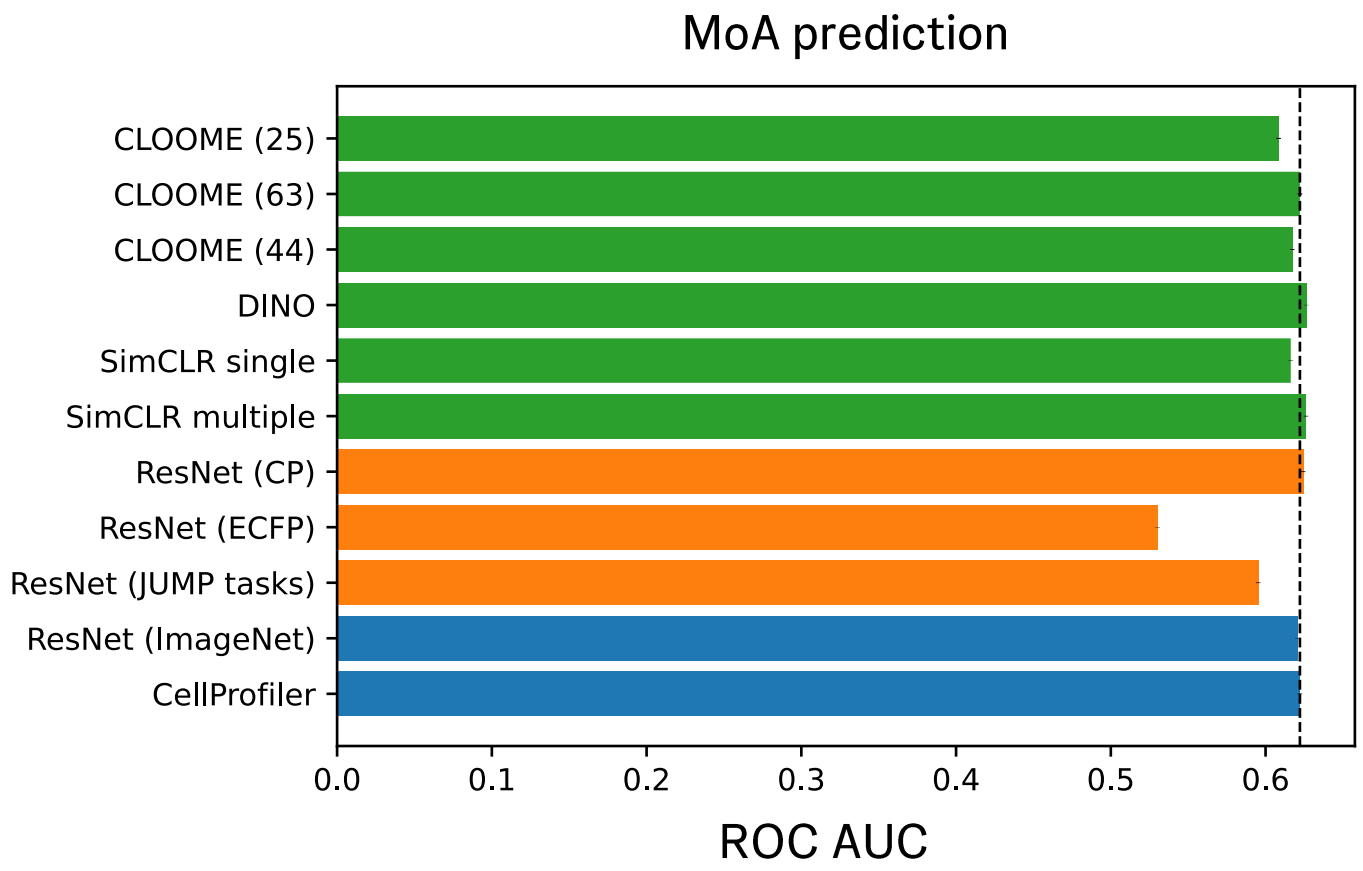
Results - MoA classification

- **Baselines**
  - CellProfiler features
  - ResNet50 pretrained with ImageNet
- **Supervised**
  - ResNet

    → Chemical activity prediction

    → ECFP features prediction

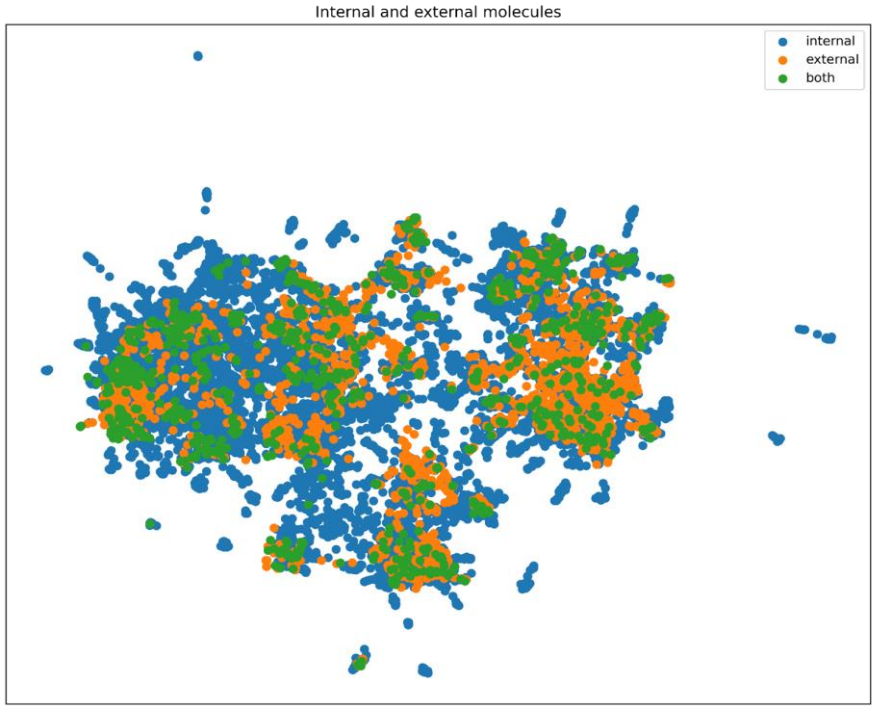
    → CP features prediction
- **Self-supervised**
  - SimCLR
  - DINO
  - CLOOME



Library design

Library design. Introduction

- **Goal:** extend internal deck of compounds
- Ideally, find compounds with new and diverse biological effects
- Usually, this search is guided by chemical structure similarity
- Including phenotypic information could enrich the search



Library design. Phenotype clustering

Experiment

Goal: Assess clustering ability of Acapella features wrt. phenotypic effect

Experiment:

1. Select Acapella features that correspond to certain assay
2. Calculate pairwise distances
3. Calculate mean average precision

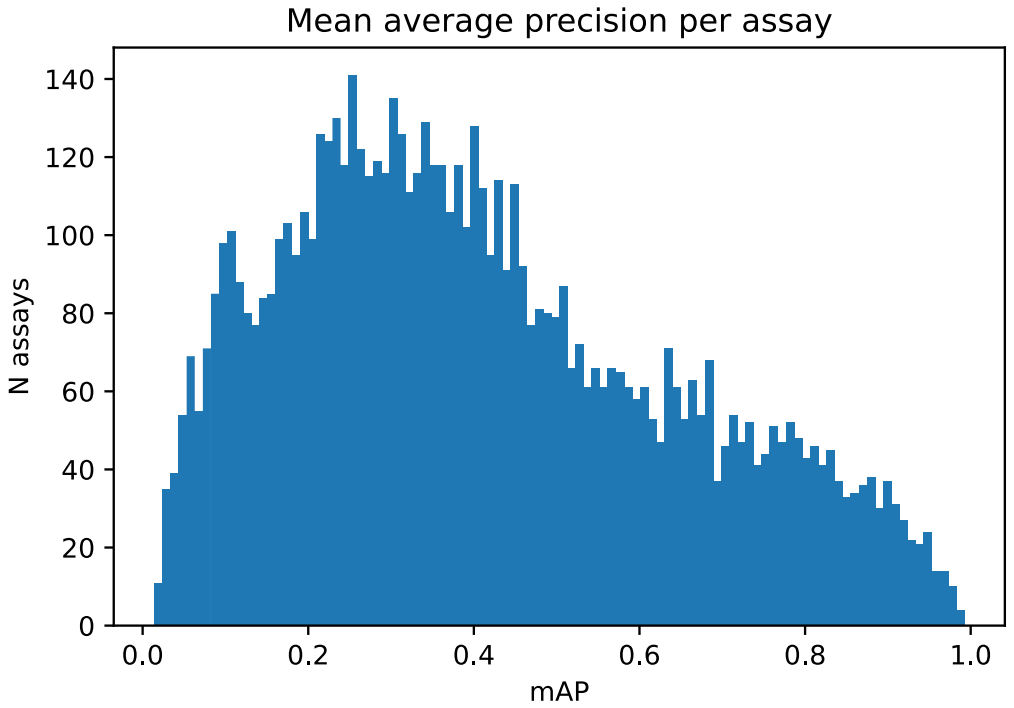
$$AP@n = \frac{1}{\#\ positives} \sum_{k=1}^{n} Precision@k\ \times\ rel(k) \qquad rel(k)\ = \begin{cases} 1, & k^{th}\ \text{element is positive} \\ 0, & k^{th}\ \text{element not positive} \end{cases}$$

$$mAP = \frac{1}{N} \sum_{i=1}^{N} AP_i$$

4. Filter out assays with less than 25 actives, 25 inactives or 100 total samples

Library design. Phenotype clustering

Results



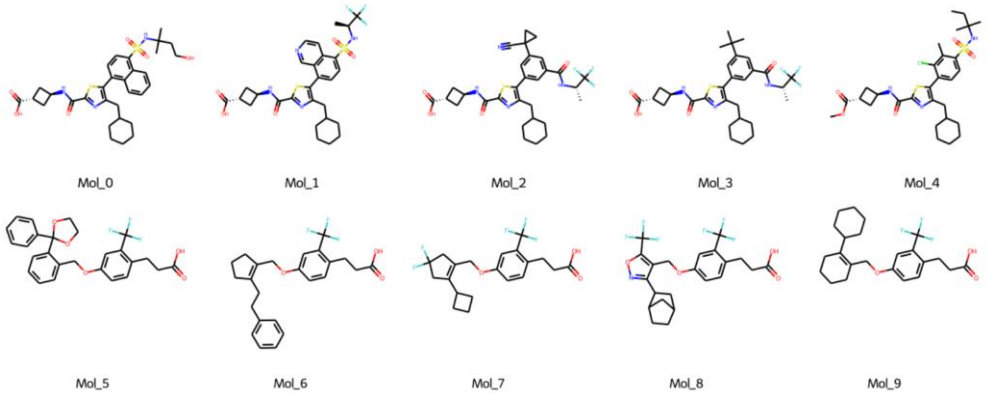
- 6,645 assays
- Among the top 1% ranked assays:

Protein family / Assay type	Number of assays
GPCR Receptor	17
Transferase (Kinase)	15
Ion Channel	10
Proliferation assay	3

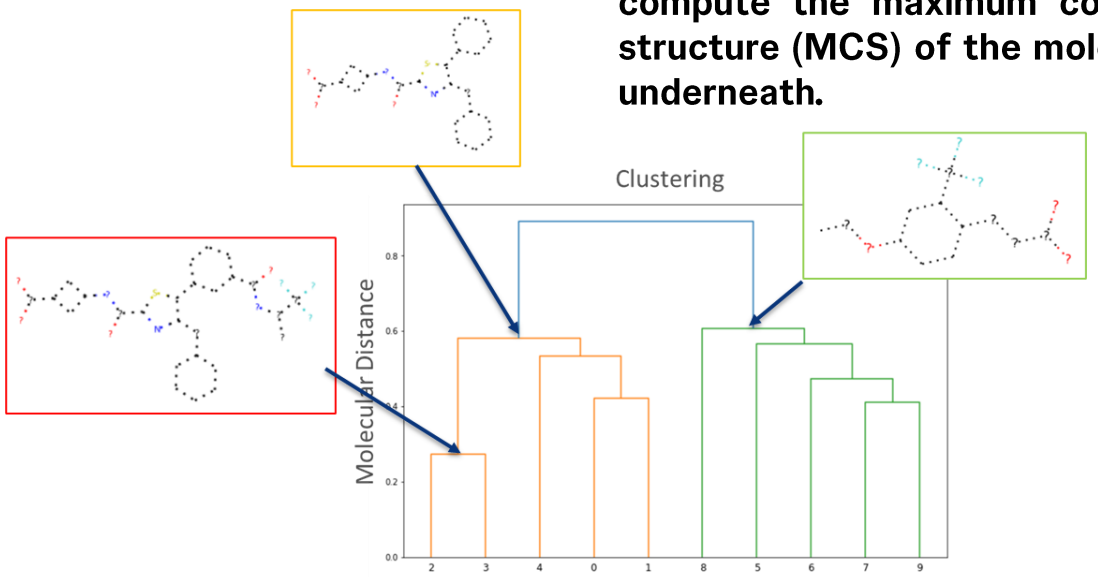
Library design. Chemotype clustering

MCS clustering. Credit: Xinhao Li

- **Maximum common structure (MCS)-based clustering** provides a fully automated approach for chemical series identification which closely mimics human chemical series conception. The cluster is defined by a single scaffold. Molecules are assigned by substructure matching and can be assigned to multiple clusters.



From top to down, for each node, compute the maximum common structure (MCS) of the molecules underneath.



Library design. Chemotype clustering

Experiment

Goal: Assess clustering ability of Acapella features wrt. chemical series

Experiment:

1. Select Acapella features of molecules that belong to certain MCS cluster
2. Calculate pairwise distances
3. Calculate mean average precision

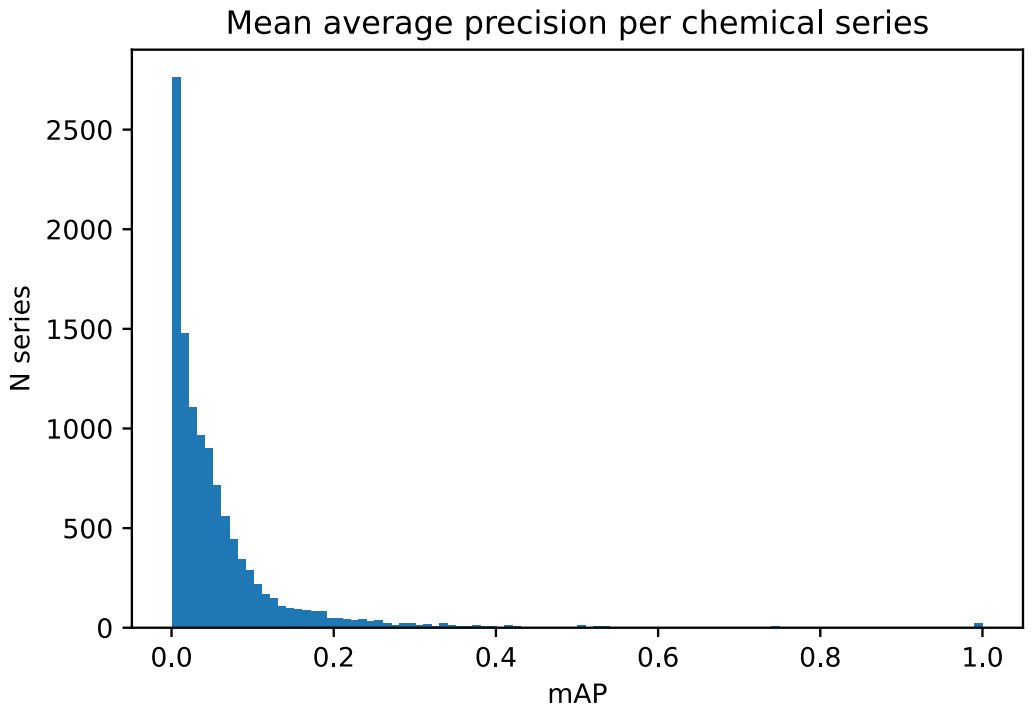
$$AP@n = \frac{1}{\#\ positives} \sum_{k=1}^{n} Precision@k\ \times\ rel(k)$$

$$rel(k)\ = \begin{cases} 1, & k^{th} \text{ element is positive} \\ 0, & k^{th} \text{ element not positive} \end{cases}$$

$$mAP\ = \frac{1}{N} \sum_{i=1}^{N} AP_i$$

Library design. Chemotype clustering

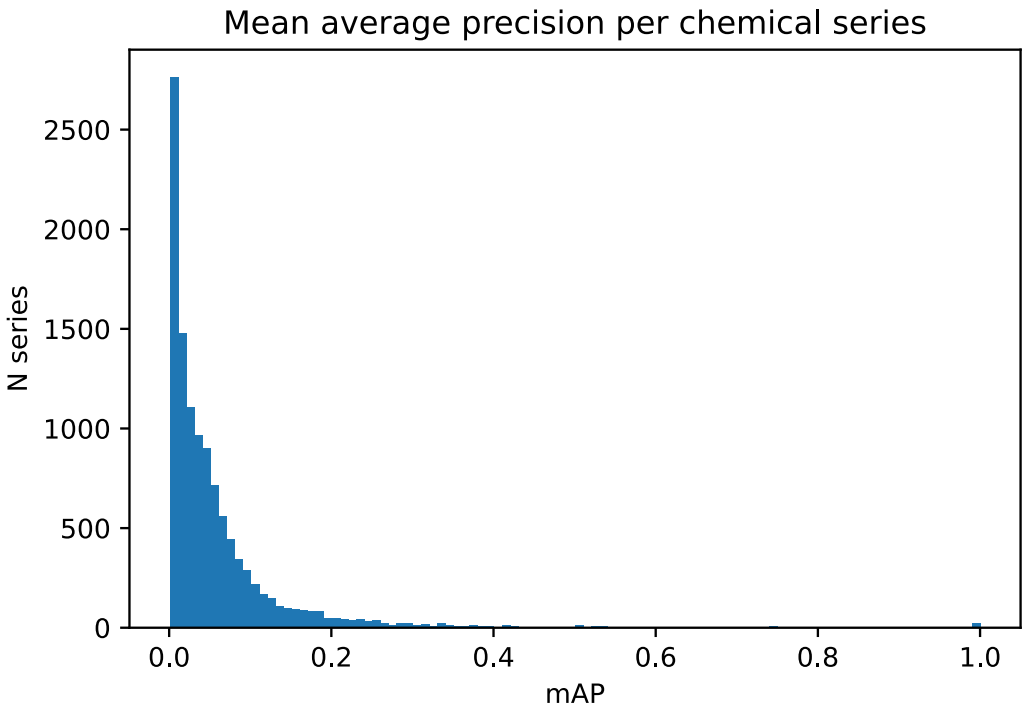
Results



- 11,262 chemical series
- Chemical series with high MAP ⟶ series that can be well clustered with Acapella features ⟶ potential series with unique biology

Library design. Chemotype clustering

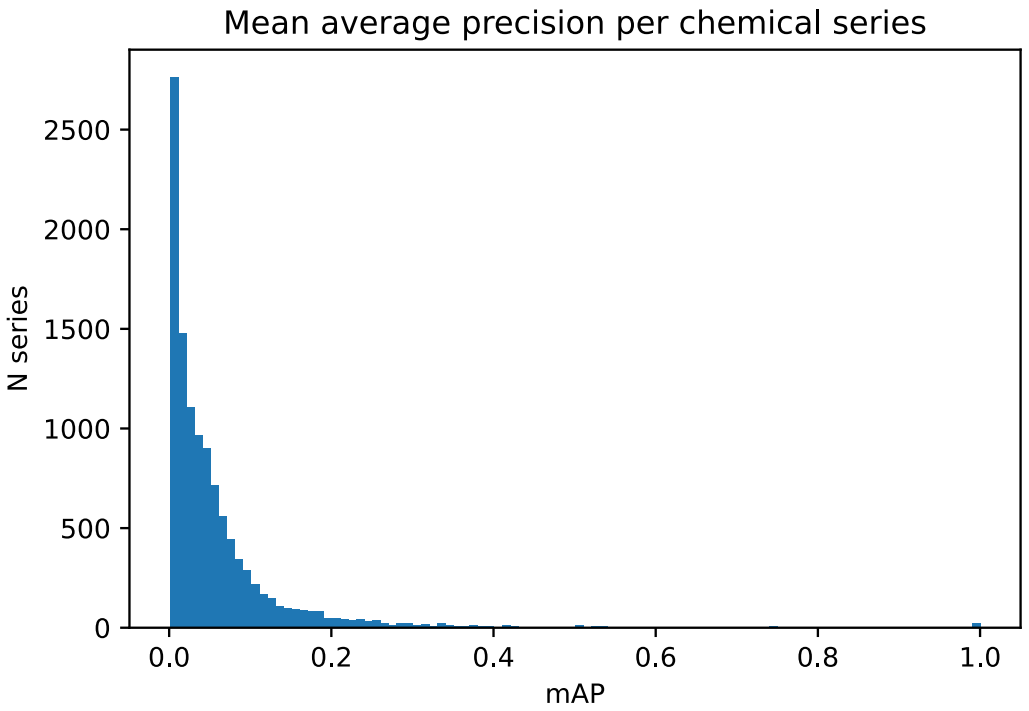
Results. Overlap with bioactivity



- For closely clustered chemical series in Acapella features (> 0.2 MAP), there are **1,050 tasks**, for which more than 60% of compounds in said cluster are actives ⟶ potential compounds with unique biology

Library design. Chemotype clustering

Results. Overlap with bioactivity



Protein family / Assay type	Number of assays
Transferase (Kinase)	190
GPCR Transmembrane Receptor	105
Hydrolase (other)	57
Proliferation assay	39

Library design. CLOOME embeddings

Results

- **Ongoing:** compute phenotype and chemotype clustering with CLOOME embeddings
- **Hypothesis:** a higher number of chemotypes are closely clustered in comparison to Acapella features
- Analyze assays that with highest MAP difference between CLOOME embeddings and Acapella features

Conclusions

Conclusions

- CLOOME pretraining achieves comparable performance to other self-supervised methods and fully supervised baselines
- Chemotypes for which acapella features are closely clustered are series with potentially unique biological effect

References

Niranj Chandrasekaran, S., Ceulemans, H., Boyd, J. D., & Carpenter, A. E. (2021). **Image-based profiling for drug discovery: due for a machine-learning upgrade?** *Nature Reviews Drug Discovery*. https://doi.org/10.1038/s41573-020-00117-w

Simm, J., Klambauer, G., Arany, A., Steijaert, M., Wegner, J. K., Gustin, E., Chupakhin, V., Chong, Y. T., Vialard, J., Buijnsters, P., Velterm Ingrid, Vapirev, A., Singh, S., Carpenter, A. E., Wuyts, R., Hochreiter, S., Moreau, Y., & Ceulemans, H. (2018). **Repurposing high-throughput image assays enables biological activity prediction for drug discovery.** *Cell Chem Biol*, *25*(5), 611–618. https://doi.org/10.1016/j.chembiol.2018.01.015

Hofmarcher, M., Rumetshofer, E., Clevert, D., Hochreiter, S., & Klambauer, G. (2019). **Accurate Prediction of Biological Assays with High-Throughput Microscopy Images and Convolutional Networks.** *Journal Of Chemical Information And Modeling*, *59*(3), 1163-1171. doi: 10.1021/acs.jcim.8b00670

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

Hochreiter, S. (2022). **Toward a broad AI.** *Communications of the ACM*, *65*(4), 56-57.

Chollet, F. (2019). **On the measure of intelligence.** *arXiv preprint arXiv:1911.01547*.

Liu, X., Zhang, F., Hou, Z., Mian, L., Wang, Z., Zhang, J., Tang, J., & Fellow, I. (2021) **Self-supervised Learning: Generative or Contrastive.** ArXiv, 1911.05722.

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

K. He, H. Fan, Y. Wu, S. Xie & R. Girshick. (2020). **Momentum Contrast for Unsupervised Visual Representation Learning.** 2020 IEEE/CVF *Conference on Computer Vision and Pattern Recognition* (CVPR), 2020, pp. 9726-9735, doi: 10.1109/CVPR42600.2020.00975.

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. (2021). **Learning transferable visual models from natural language supervision.** In *Proceedings of the 38th International Conference on Machine Learning* (ICML)

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. (2022). **Hierarchical Text-Conditional Image Generation with CLIP Latents.** ArXiv.2204.06125

Fürst, A., Rumetshofer, E., Tran, V., Ramsauer, H., Tang, F., Lehner, J., David Kreil, Michael Kopp, Günter Klambauer, Angela Bitto-Nemling & Hochreiter, S. (2021). **CLOOB: Modern Hopfield Networks with InfoLoob outperform CLIP**. *ArXiv*. 2110.11316

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. (2016). **Cell Painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes.** *Nature Protocols 2016 11:9*, *11*(9), 1757–1774. https://doi.org/10.1038/nprot.2016.105

Chandrasekaran, S. N., Ackerman, J., Alix, E., Ando, D. M., … Carpenter, A. E. (2023). **JUMP Cell Painting dataset: morphological impact of 136,000 chemical and genetic perturbations.** *BioRxiv*, 2023.03.23.534023. https://doi.org/10.1101/2023.03.23.534023

Sanchez-Fernandez, A., Rumetshofer, E., Hochreiter, S. & Klambauer, G. (2022, March). **Contrastive Learning of Image-and Structure-Based Representations in Drug Discovery.** In ICLR 2022 Machine Learning for Drug Discovery.

Thank you

Johnson&Johnson