Introduction to contrastive self-supervised learning and its application in computational biology & drug discovery

Michael M. Pieler (MicPie)

Agenda

1. Introduction
2. Contrastive self-supervised learning
3. Contrastive supervised learning
4. Contrastive learning with different modalities
5. Contrastive learning in computational biology & drug discovery
6. Outlook

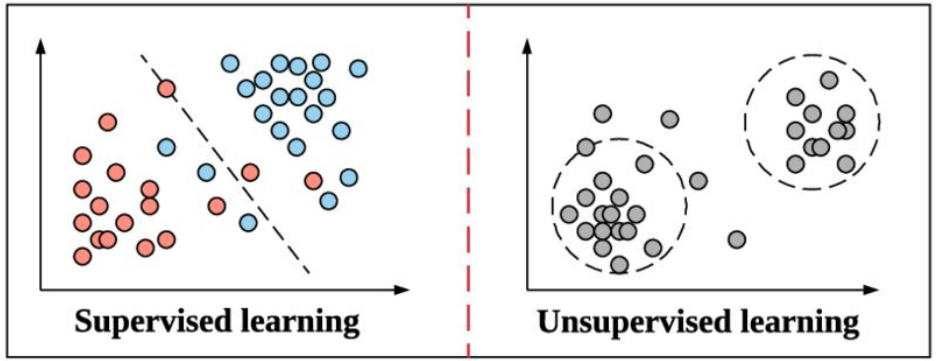
Supervised vs. self-supervised learning?

Supervised learning:

“Traditional setup”: we have input data and the associated labels.

Self-supervised learning:

We use unlabeled data with special learning objectives.



Self-Supervised Representation Learning

Image: Examples of Supervised Learning (Linear Regression) and Unsupervised... | Download Scientific Diagram

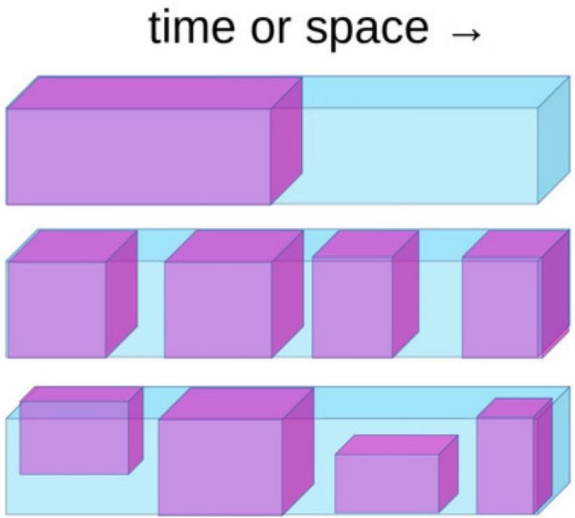
“How much information for learning” theory?

- **“Pure” Reinforcement Learning (cherry)**
  - The machine predicts a scalar reward given once in a while.
  - **A few bits for some samples**
- **Supervised Learning (icing)**
  - The machine predicts a category or a few numbers for each input
  - Predicting human-supplied data
  - **10→10,000 bits per sample**
- **Self-Supervised Learning (cake génoise)**
  - The machine predicts any part of its input for any observed part.
  - Predicts future frames in videos
  - **Millions of bits per sample**

Slide from Yann LeCun from https://www.gwern.net/newsletter/2020/05#baking-the-cake

Traditional self-supervised learning setup

- Predict any part of the input from any other part.
- Predict the future from the past.
- Predict the invisible from the visible.
- Predict any occluded, masked, or corrupted part from all available parts.



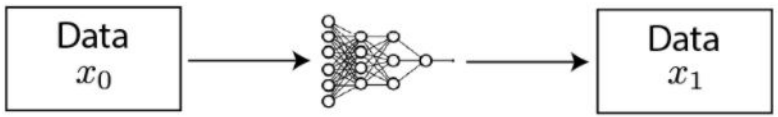
- Pretend there is a part of the input you don't know and predict that.
- Reconstruction = SSL when any part could be known or unknown

There are a lot of different approaches!

Slide from Y. LeCun ICLR 2020; detailed overview: Self-Supervised Representation Learning

Generative & contrastive self-supervised learning

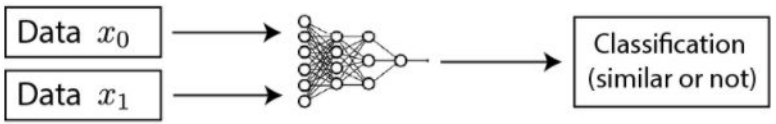
Generative / Predictive



Loss measured in the output space

Examples: Colorization, Auto-Encoders

Contrastive



Loss measured in the representation space

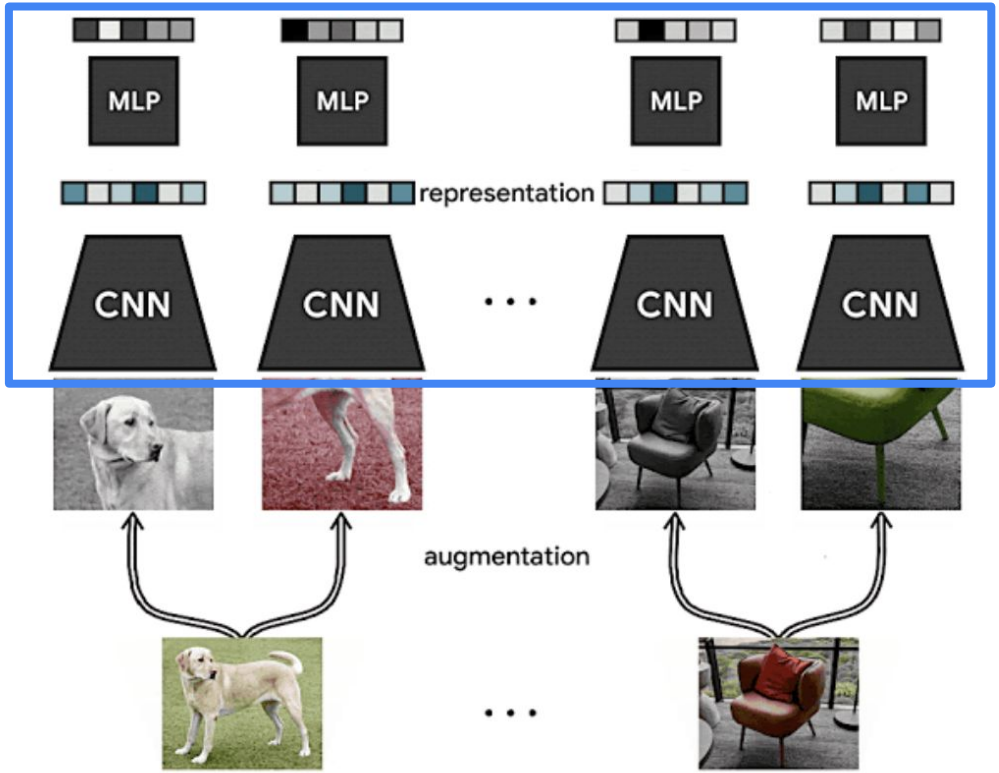
Examples: TCN, CPC, Deep-InfoMax

Contrastive Self-Supervised Learning

Agenda

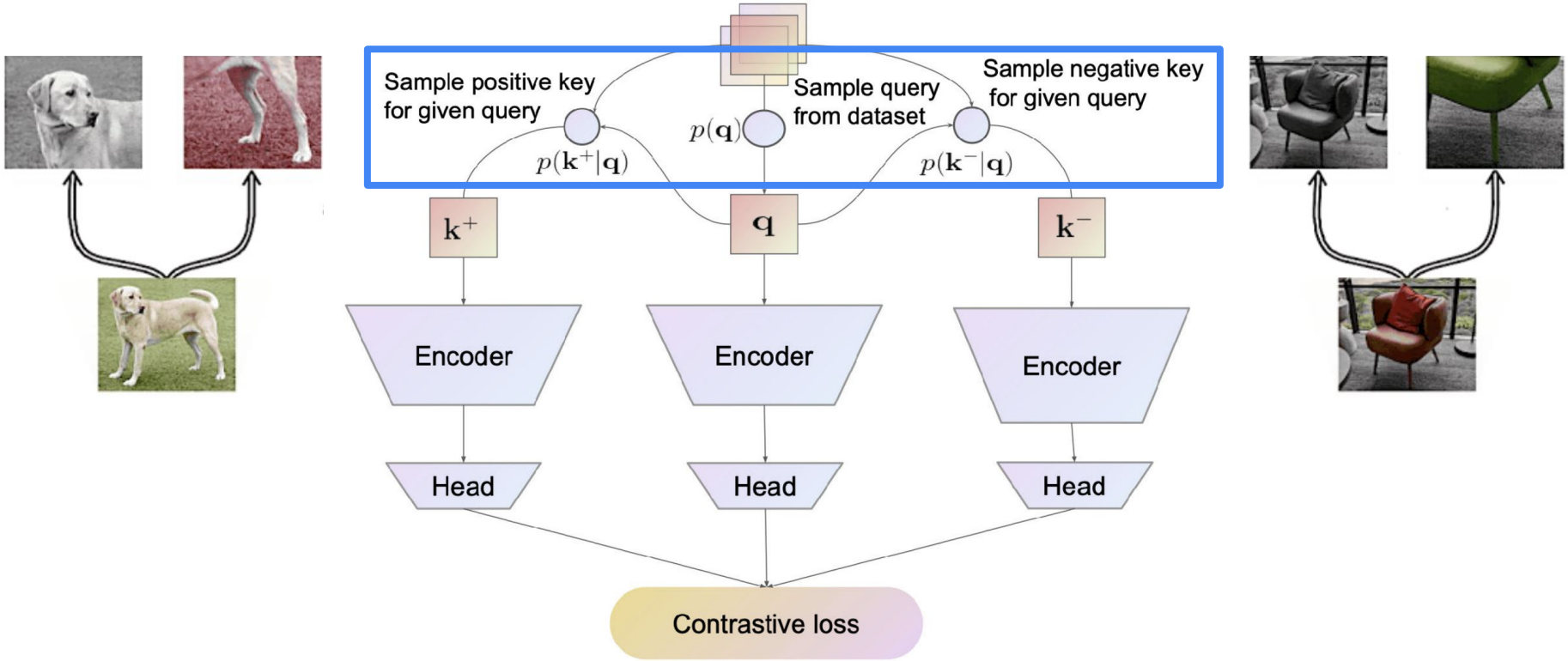
1. Introduction
2. Contrastive self-supervised learning
3. Contrastive supervised learning
4. Contrastive learning with different modalities
5. Contrastive learning in computational biology & drug discovery
6. Outlook

Contrastive self-supervised learning



Advancing Self-Supervised and Semi-Supervised Learning with SimCLR

Contrastive self-supervised learning



Contrastive Representation Learning: A Framework and Review & Advancing Self-Supervised and Semi-Supervised Learning with SimCLR

The InfoNCE loss

Information noise contrastive loss: cross-entropy loss for an N-way softmax classifier.

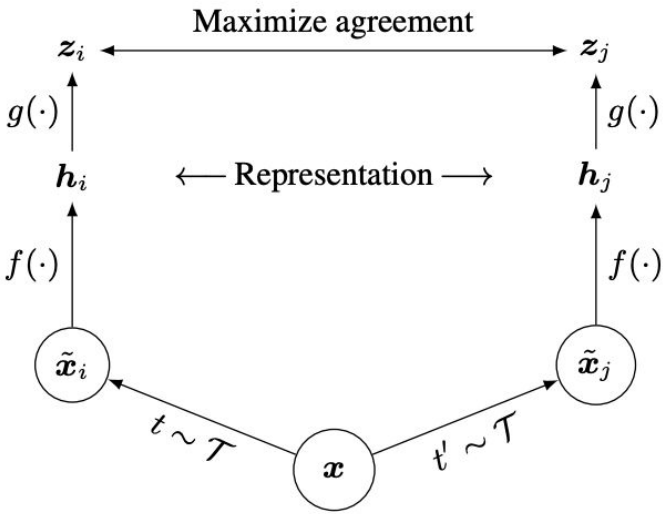
$$\text{score}(f(x), f(x^+)) >> \text{score}(f(x), f(x^-))$$

$$\mathcal{L}_N = -\mathbb{E}_X \left[\log \frac{\exp\left(f(x)^T f\left(x^+\right)\right)}{\exp\left(f(x)^T f\left(x^+\right)\right) + \sum_{j=1}^{N-1} \exp\left(f(x)^T f\left(x_j\right)\right)} \right]$$

Scoring function is the dot product.
Maximizes the mutual information between f(x) and f(x+).

Contrastive Self-Supervised Learning

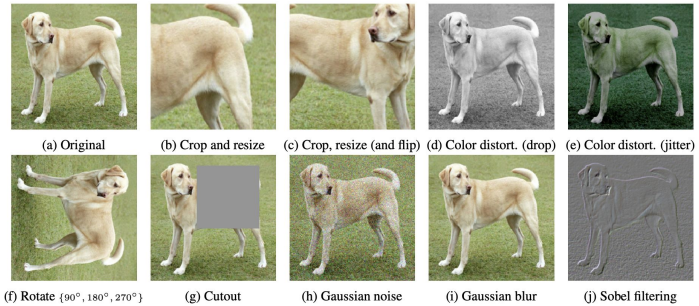
SimCLR - overview

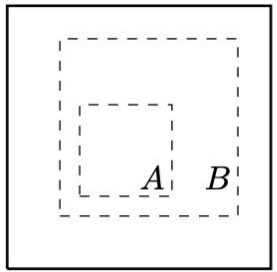


SimCLR v1 https://arxiv.org/abs/2002.05709,
SimCLR v2 https://arxiv.org/abs/2006.10029,
Advancing Self-Supervised and Semi-Supervised Learning with SimCLR

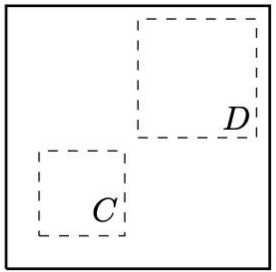
SimCLR - key findings

- Composition of multiple data augmentation operations is crucial and benefits from stronger data augmentation than supervised learning.
- Introducing a learnable nonlinear transformation between the representation and the contrastive loss substantially improves the quality of the learned representations.
- Benefits from normalized embeddings and an appropriately adjusted temperature parameter.
  (→ Features are embedded on a unit hypersphere.)
- Benefits from larger batch sizes and longer training compared to its supervised counterpart.





(a) Global and local views.



(b) Adjacent views.

SimCLR v1 https://arxiv.org/abs/2002.05709,
SimCLR v2 https://arxiv.org/abs/2006.10029,
Advancing Self-Supervised and Semi-Supervised Learning with SimCLR

Features on the unit hypersphere?

Features are (usually) L2 normalized to embed them on a unit hypersphere:

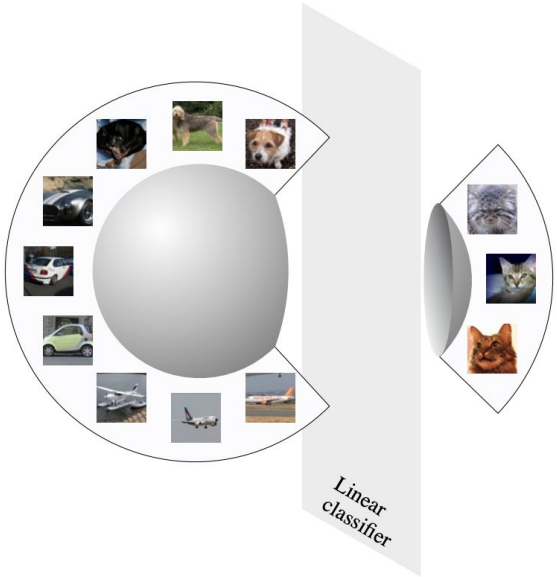
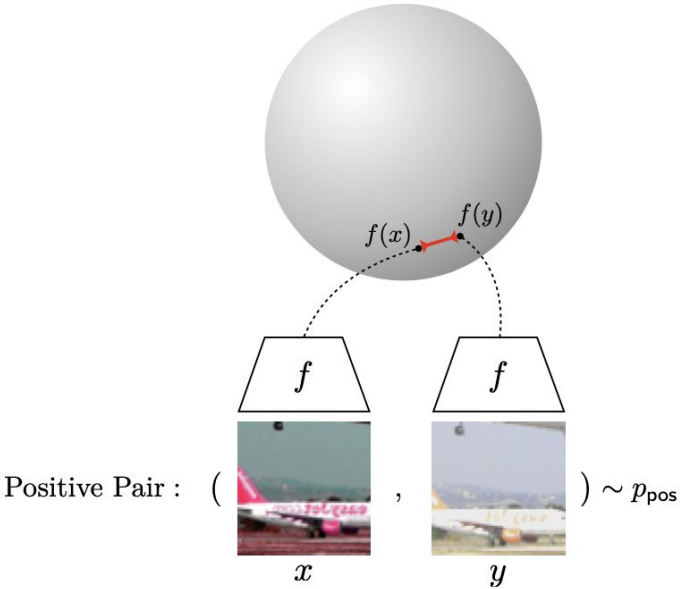


Figure 2: **Hypersphere:** When classes are well-clustered (forming spherical caps), they are linearly separable. The same does not hold for Euclidean spaces.

Understanding Contrastive Representation Learning through Alignment and Uniformity on the Hypersphere

Alignment & uniformity?

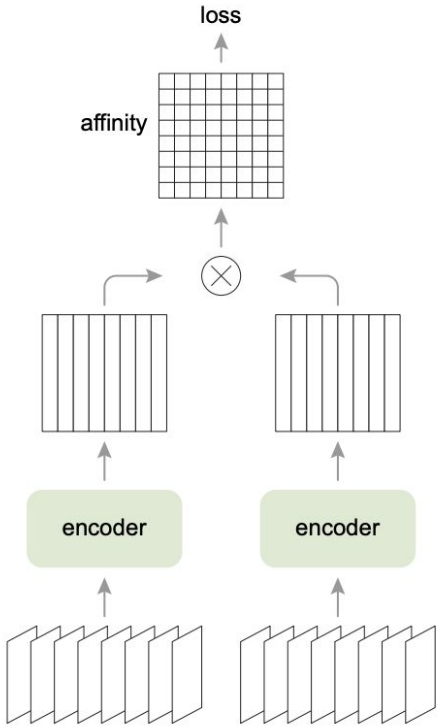


Alignment: Similar samples have similar features.
(Figure inspired by Tian et al. (2019).)

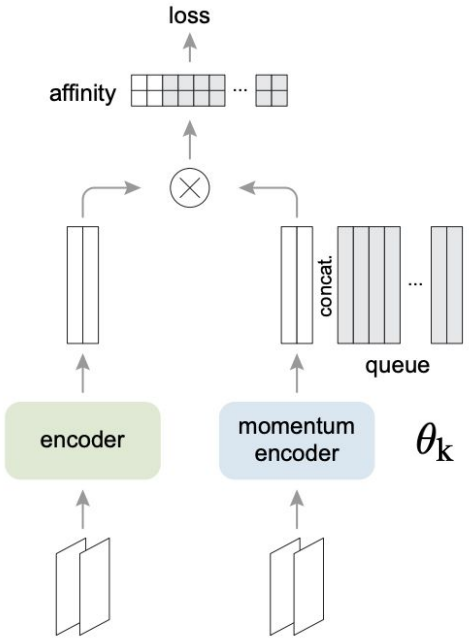
Uniformity: Preserve maximal information.

Understanding Contrastive Representation Learning through Alignment and Uniformity on the Hypersphere

Momentum contrast (MoCo) - overview



(a) end-to-end

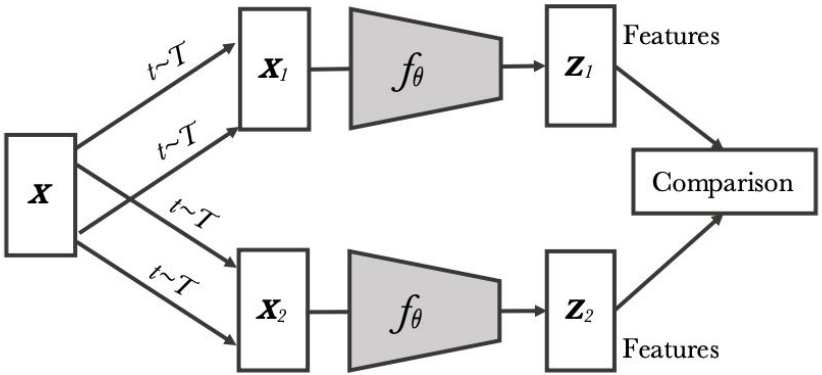


(b) Momentum Contrast

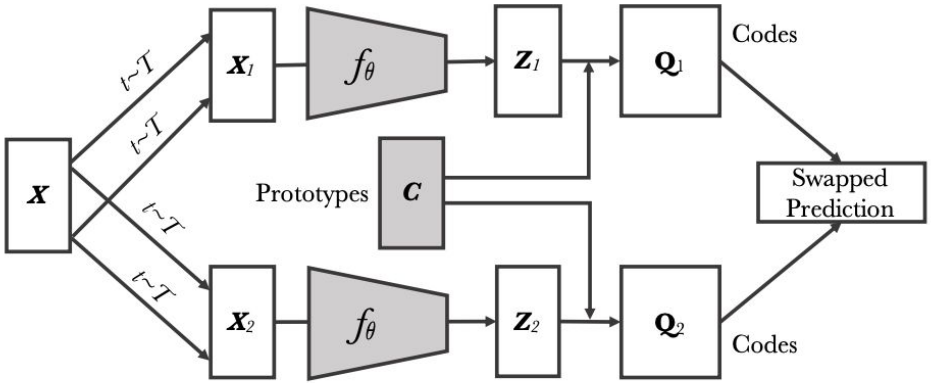
$$\theta_{\mathrm{k}} \leftarrow m\theta_{\mathrm{k}} + (1-m)\theta_{\mathrm{q}}.$$

MoCo v1 https://arxiv.org/abs/1911.05722, MoCo v2 https://arxiv.org/abs/2003.04297

Swapping Assignments between Views (SwAV)



Contrastive instance learning



Swapping Assignments between Views (Ours)

[Unsupervised Learning of Visual Features by Contrasting Cluster Assignments](#) &
https://github.com/facebookresearch/swav

Swapping Assignments between Views (SwAV)

Unsupervised Learning of Visual Features by Contrasting Cluster Assignments &
https://github.com/facebookresearch/swav

SwAV: Multi-crop

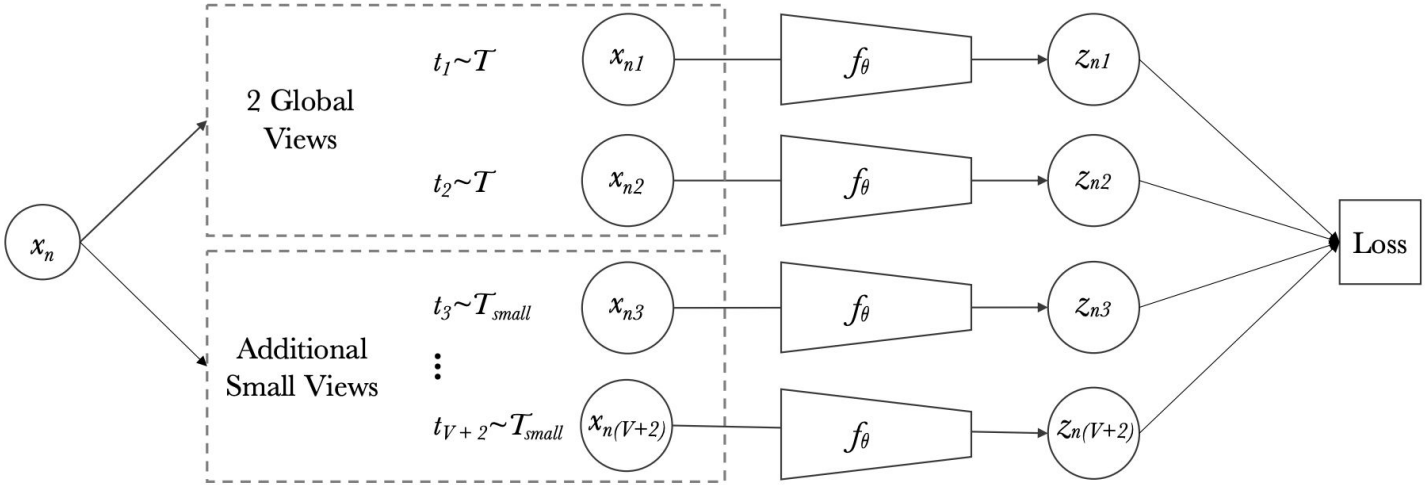


Figure 5: **Multi-crop**: the image x_n is transformed into $V + 2$ views: two global views and V small resolution zoomed views.

[Unsupervised Learning of Visual Features by Contrasting Cluster Assignments](#) &
https://github.com/facebookresearch/swav

SwAV with a 10B CNN model

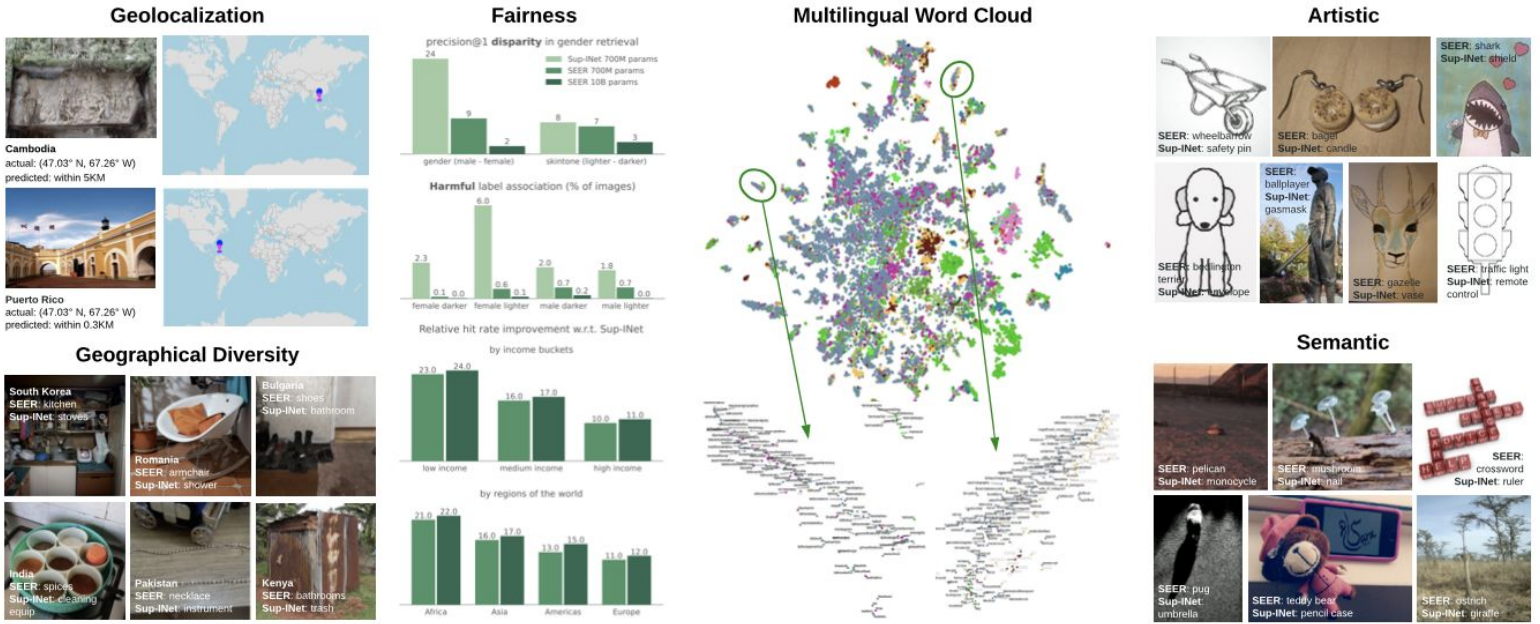
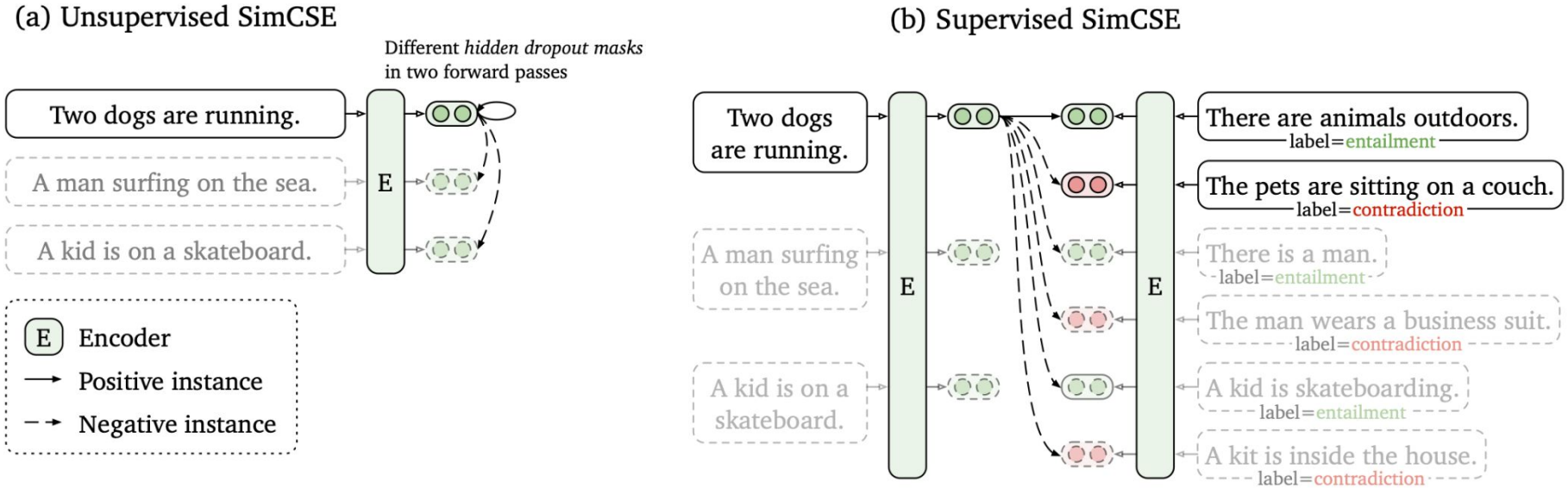


Figure 1. Self-supervised training on diverse, real, and unfiltered internet data leads to interesting properties emerging like geolocalization, fairness, multilingual hashtag embeddings, artistic and better semantic information. See supplemental material for license information.

[Vision Models Are More Robust And Fair When Pretrained On Uncurated Images Without Supervision]

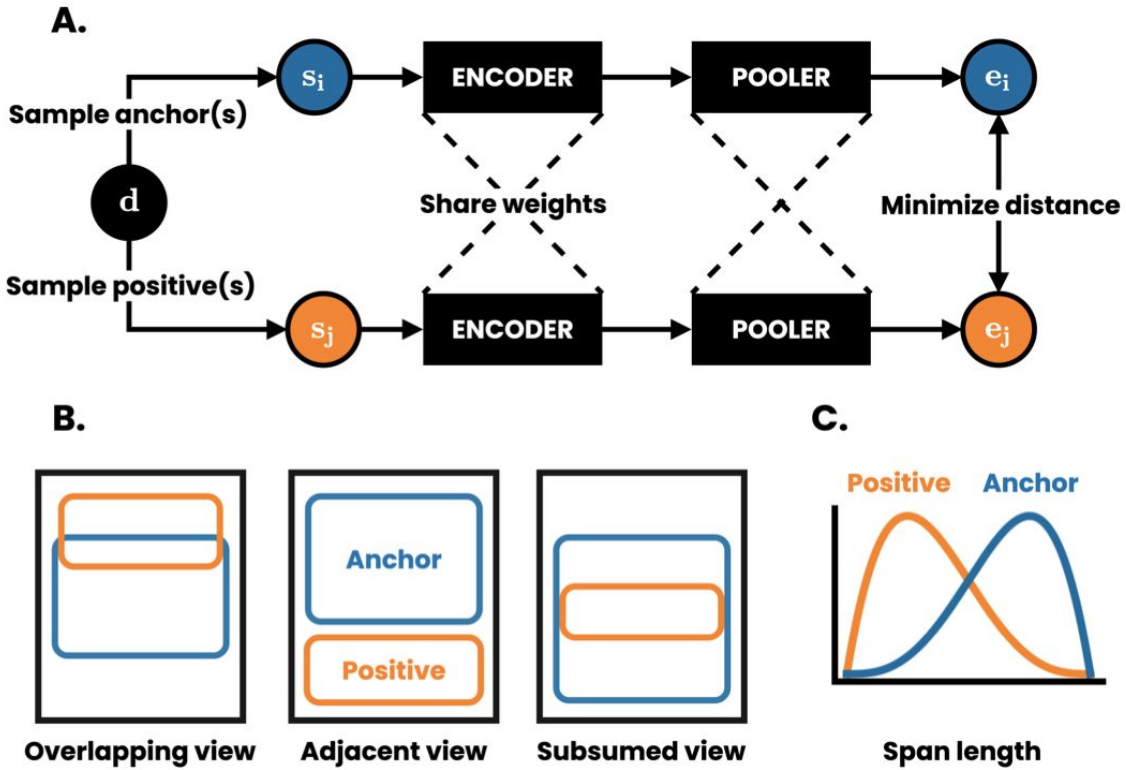
Simple Contrastive Learning of Sentence Embeddings



(more on supervised CL later)

SimCSE: Simple Contrastive Learning of Sentence Embeddings

DeCLUTR



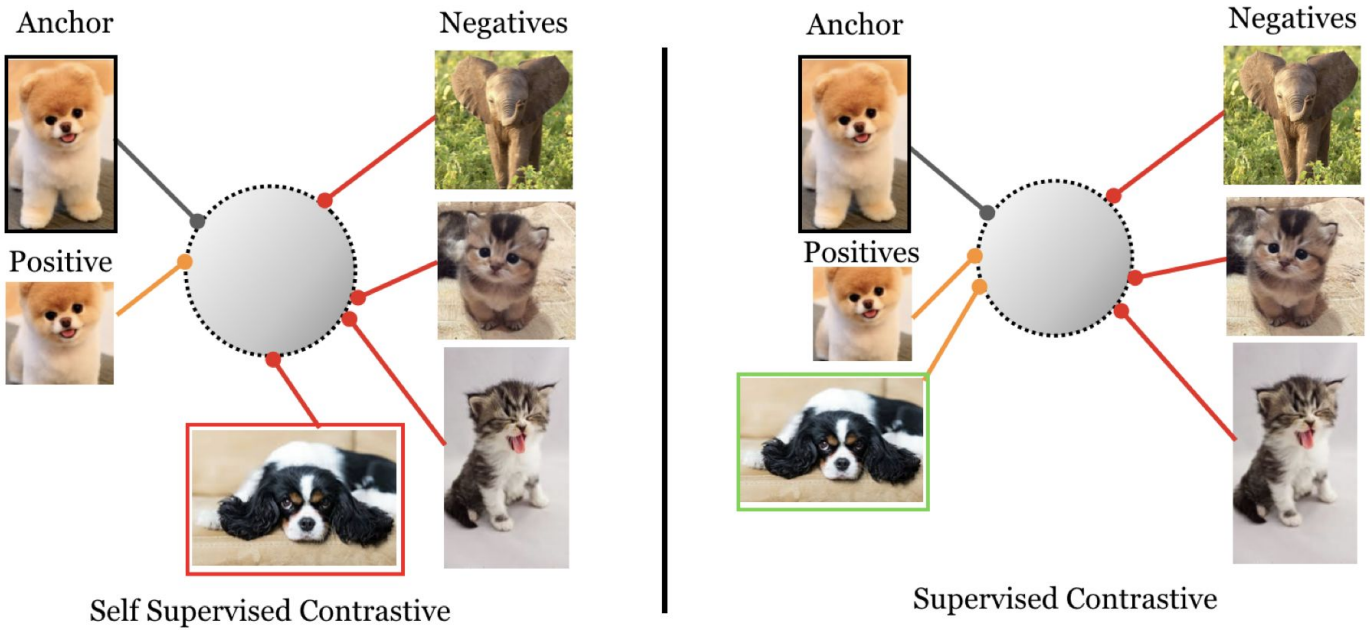
→ Similar setups (should) work for biological sequence and graph data!

DeCLUTR: Deep Contrastive Learning for Unsupervised Textual Representations, https://github.com/JohnGiorgi/DeCLUTR

Agenda

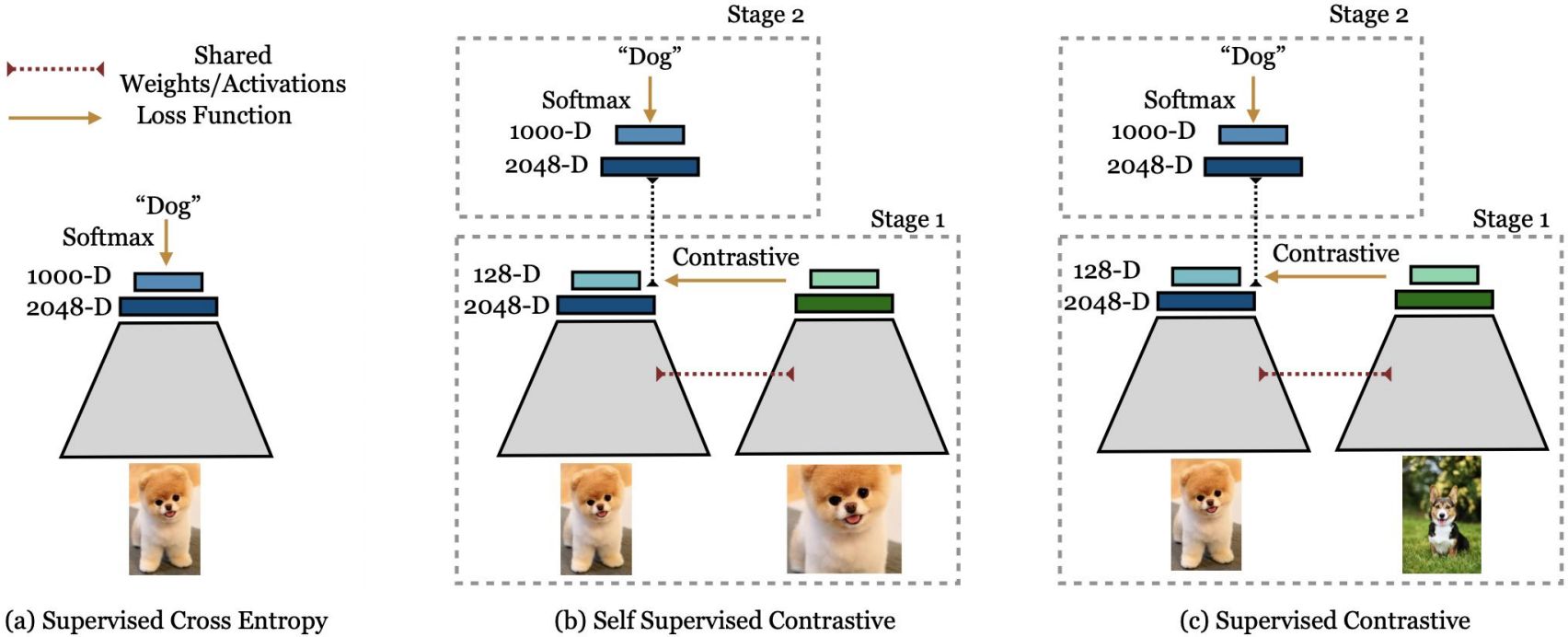
1. Introduction
2. Contrastive self-supervised learning
3. Contrastive supervised learning
4. Contrastive learning with different modalities
5. Contrastive learning in computational biology & drug discovery
6. Outlook

Supervised, self-supervised contrastive, and supervised contrastive learning



Supervised Contrastive Learning, https://github.com/HobbitLong/SupContrast

Supervised, self-supervised contrastive, and supervised contrastive learning



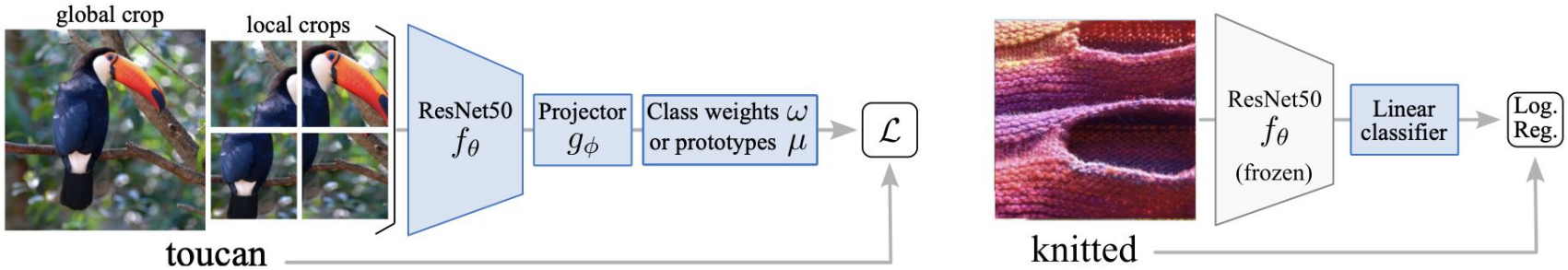
(a) Supervised Cross Entropy

(b) Self Supervised Contrastive

(c) Supervised Contrastive

Supervised Contrastive Learning, https://github.com/HobbitLong/SupContrast

Contrastive approaches can help supervised setups



(a) Supervised learning using multi-crop and a projector.

(b) Transfer learning with a frozen model.

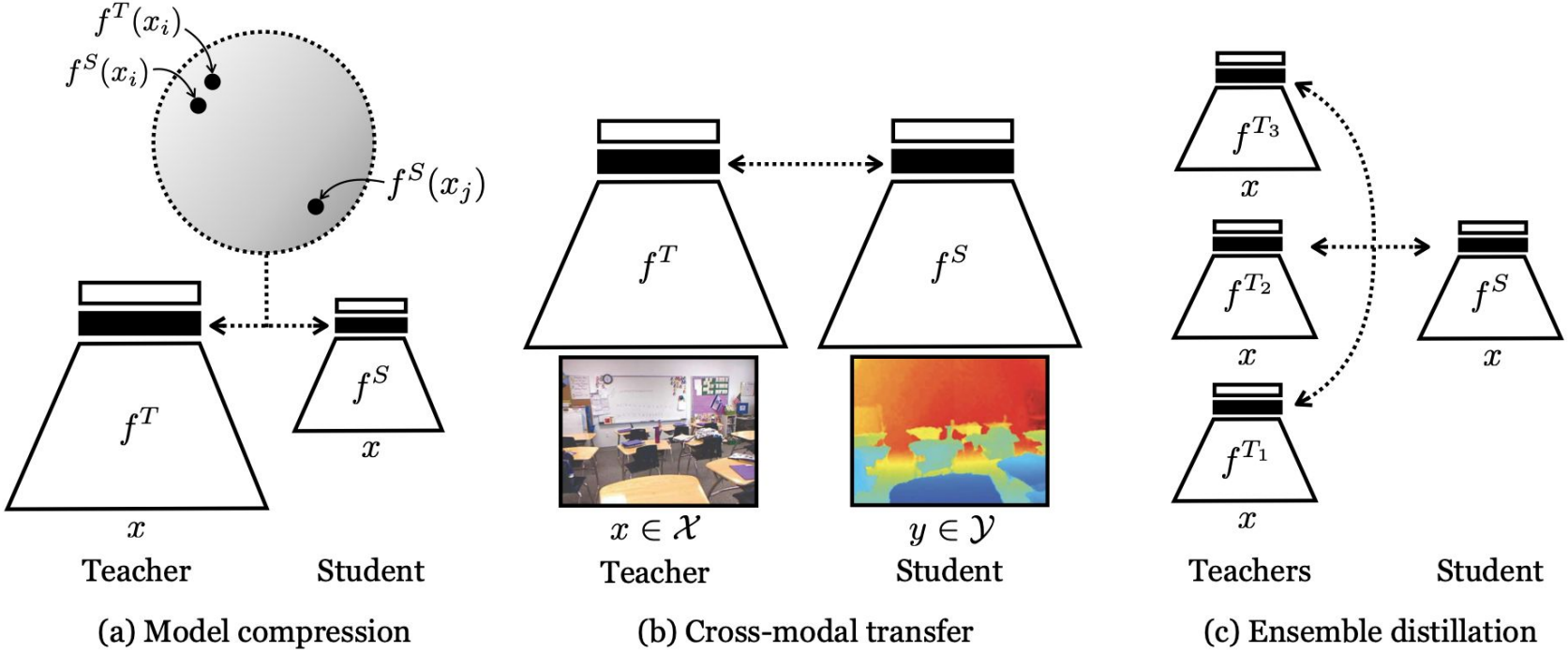
Figure 2: **Our proposed supervised learning setup** borrows multi-crop [6] and projectors [8] from SSL to train on IN1K (*left*). The projector g is discarded after training, and the ResNet backbone f is used as a feature extractor in combination with a linear classifier trained for each task, e.g., for texture classification on DTD [11] (*right*).

Improving the Generalization of Supervised Models

Agenda

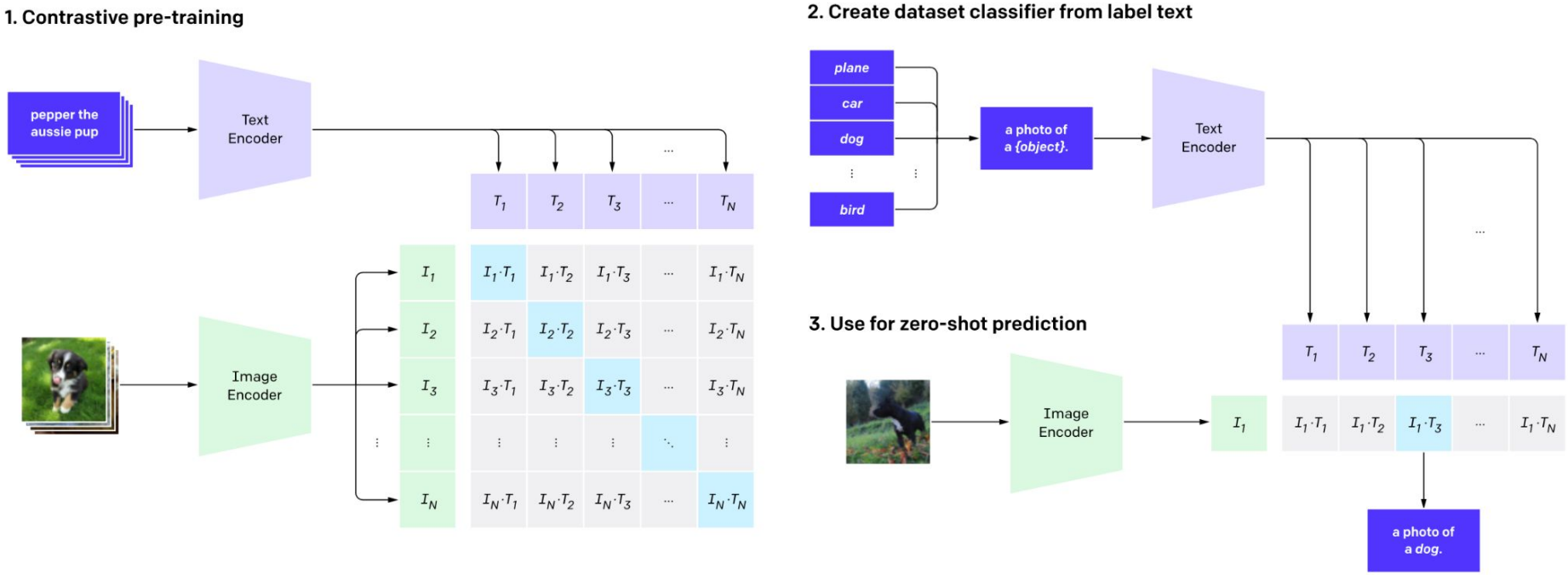
1. Introduction
2. Contrastive self-supervised learning
3. Contrastive supervised learning
4. Contrastive learning with different modalities
5. Contrastive learning in computational biology & drug discovery
6. Outlook

Contrastive learning with different modalities and setups



[Contrastive Representation Distillation]

Contrastive Language–Image Pre-training (CLIP)



Learning Transferable Visual Models From Natural Language Supervision,CLIP: Connecting Text and Images

CLIP loss

$$\mathcal{L}_{\text{CLIP}} =$$

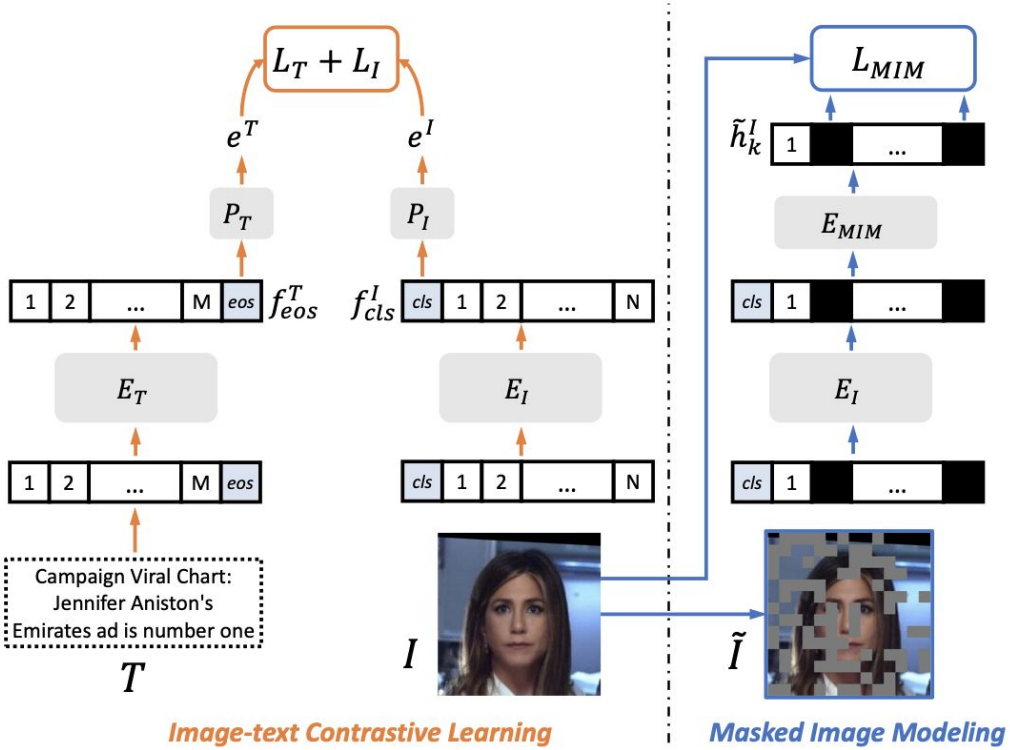
$$-\frac{1}{2N}\sum_{j=1}^{N}\log\underbrace{\left[\frac{\exp\left(\left\langle I_j^e, T_j^e\right\rangle/\tau\right)}{\sum_{k=1}^{N}\exp\left(\left\langle I_j^e, T_k^e\right\rangle/\tau\right)}\right]}_{\text{Contrasting images with texts}}$$

$$-\frac{1}{2N}\sum_{k=1}^{N}\log\underbrace{\left[\frac{\exp\left(\left\langle I_k^e, T_k^e\right\rangle/\tau\right)}{\sum_{j=1}^{N}\exp\left(\left\langle I_j^e, T_k^e\right\rangle/\tau\right)}\right]}_{\text{Contrasting texts with images}}$$

```
# image_encoder - ResNet or Vision Transformer
# text_encoder  - CBOW or Text Transformer
# I[n, h, w, c] - minibatch of aligned images
# T[n, l]       - minibatch of aligned texts
# W_i[d_i, d_e] - learned proj of image to embed
# W_t[d_t, d_e] - learned proj of text to embed
# t             - learned temperature parameter

# extract feature representations of each modality
I_f = image_encoder(I) #[n, d_i]
T_f = text_encoder(T)  #[n, d_t]

# joint multimodal embedding [n, d_e]
I_e = l2_normalize(np.dot(I_f, W_i), axis=1)
T_e = l2_normalize(np.dot(T_f, W_t), axis=1)

# scaled pairwise cosine similarities [n, n]
logits = np.dot(I_e, T_e.T) * np.exp(t)

# symmetric loss function
labels = np.arange(n)
loss_i = cross_entropy_loss(logits, labels, axis=0)
loss_t = cross_entropy_loss(logits, labels, axis=1)
loss   = (loss_i + loss_t)/2
```

Learning Transferable Visual Models From Natural Language Supervision,CLIP: Connecting Text and Images
https://github.com/lucidrains/x-clip, https://github.com/mlfoundations/open_clip

Masked Auto Encoders (MAE) & contrastive learning

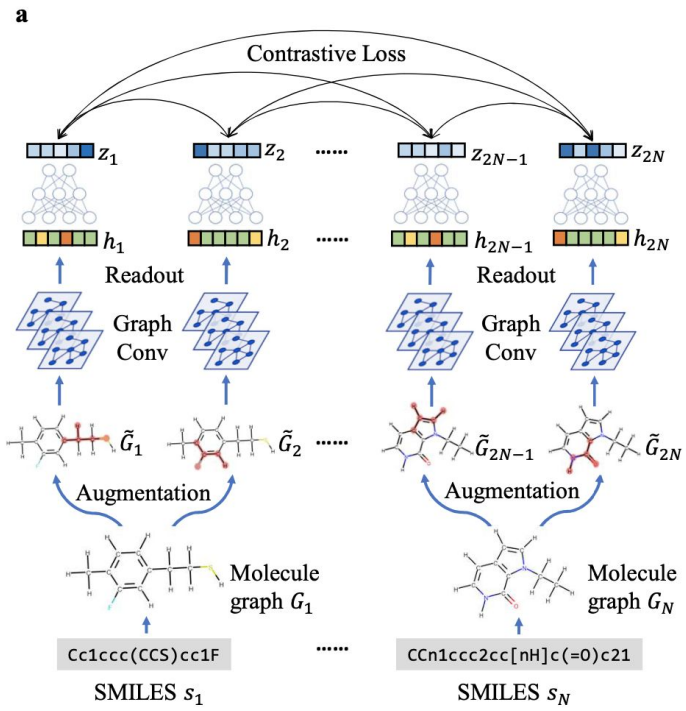


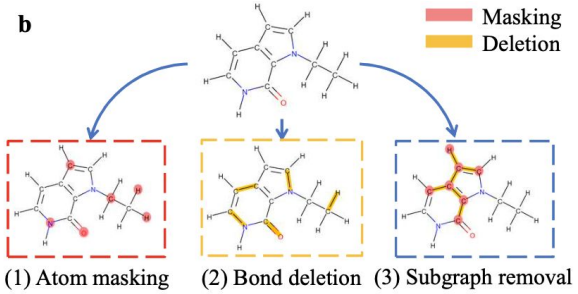
General Facial Representation Learning in a Visual-Linguistic Manner

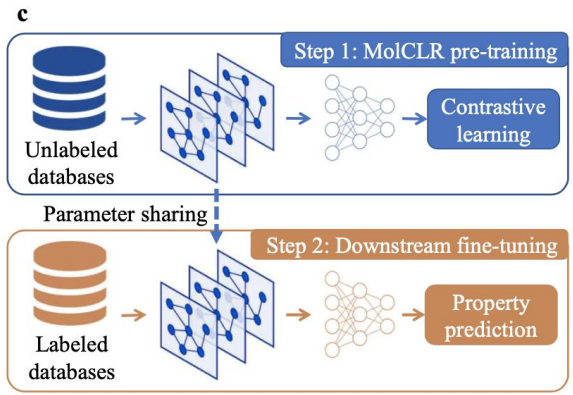
Agenda

1. Introduction
2. Contrastive self-supervised learning
3. Contrastive supervised learning
4. Contrastive learning with different modalities
5. Contrastive learning in computational biology & drug discovery
6. Outlook

MolCLR via GNN

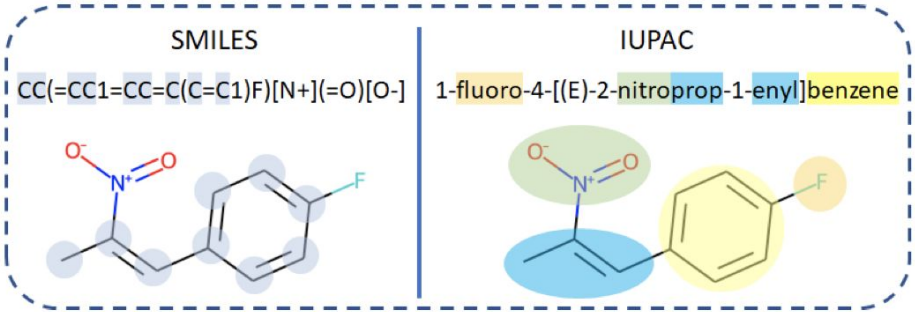


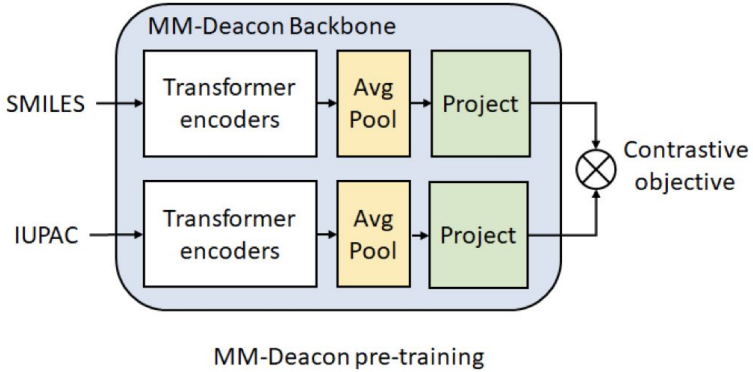




[Molecular Contrastive Learning of Representations via Graph Neural Networks]

Multilingual molecular domain embedding analysis via contrastive learning (MM-Deacon)





MM-Deacon pre-training

Multilingual Molecular Representation Learning via Contrastive Pre-training

3D Infomax improves GNNs for Molecular Property Preds.

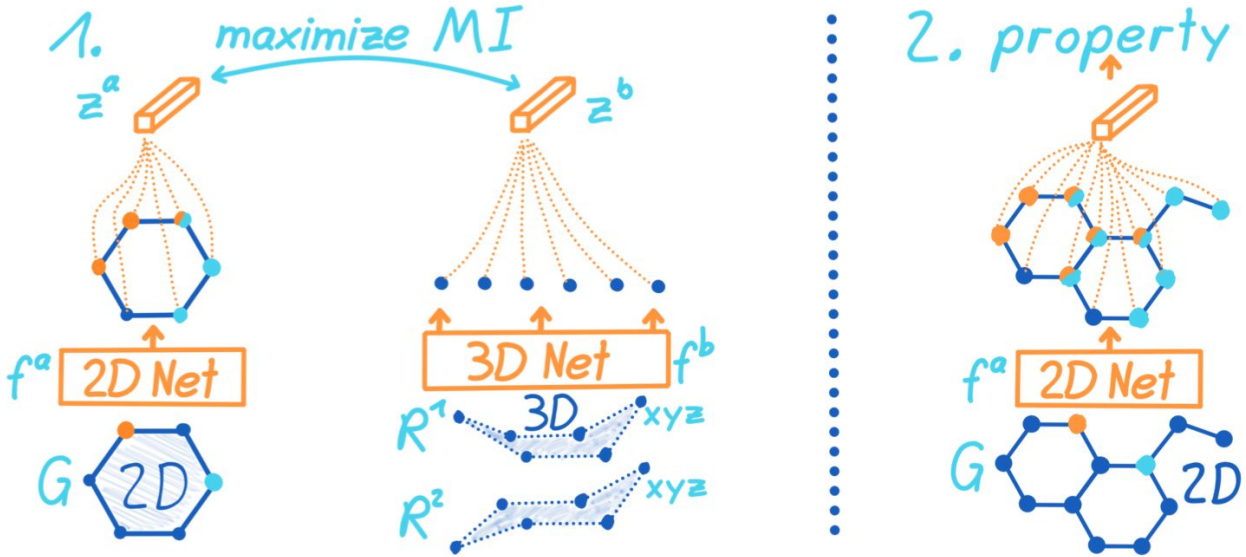
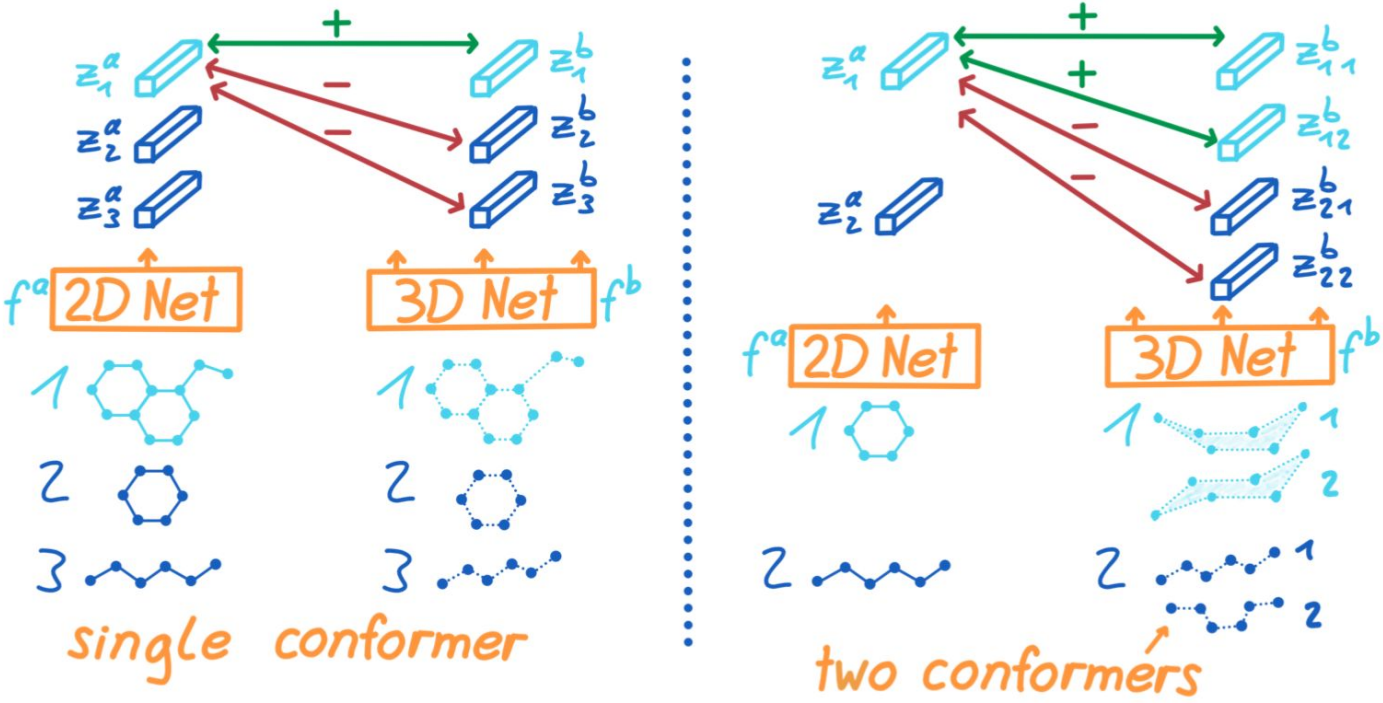


Figure 2. We first pre-train a 2D network f^a by maximizing the mutual information (MI) between its representation z^a of a molecular graph G and a 3D representation z^b produced from the molecules' conformers R^j. In step 2, the weights of f^a are transferred and fine-tuned to predict properties.

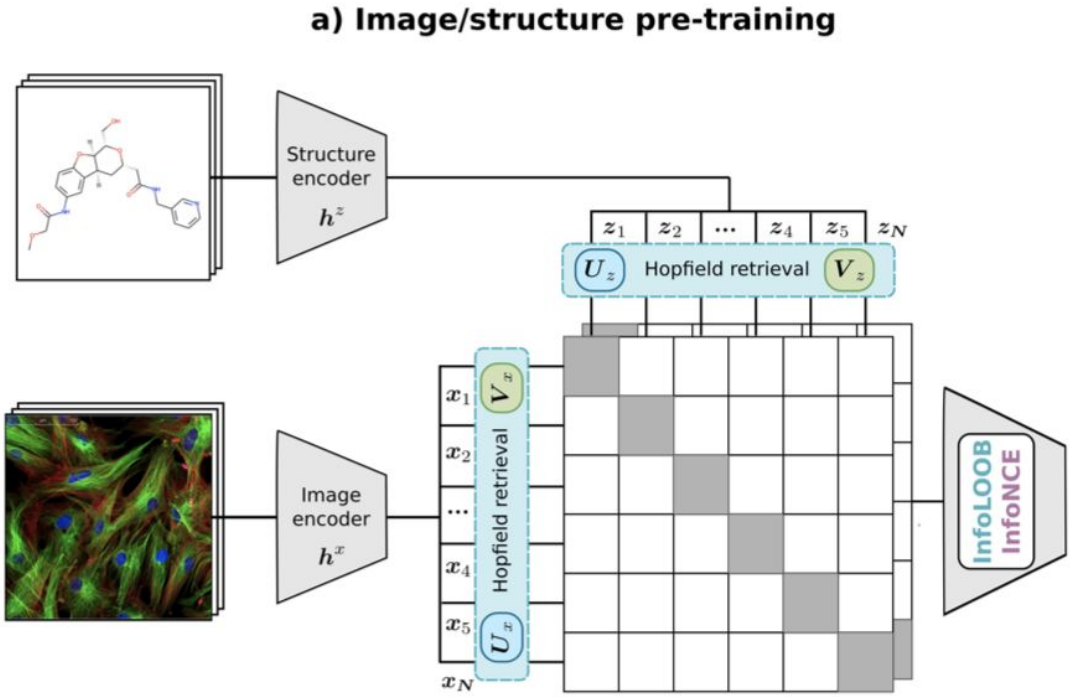
[3D Infomax improves GNNs for Molecular Property Prediction]

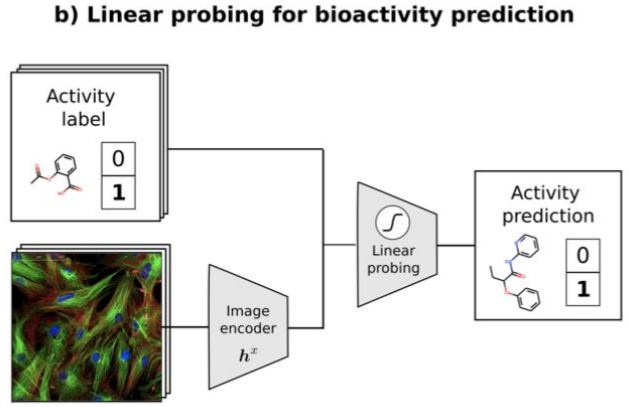
3D Infomax improves GNNs for Molecular Property Preds.

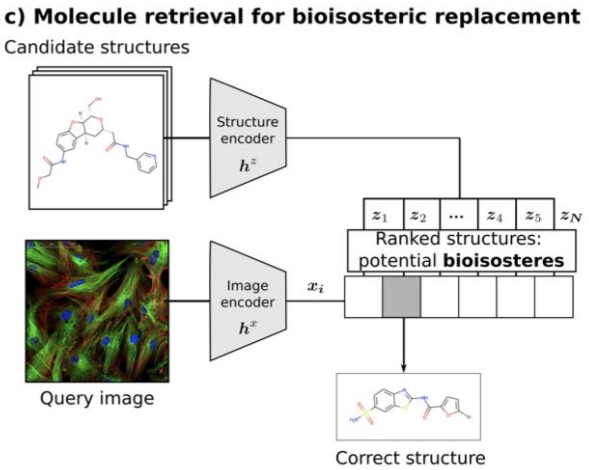


3D Infomax improves GNNs for Molecular Property Prediction

Contrastive learning of image- and structure-based representations

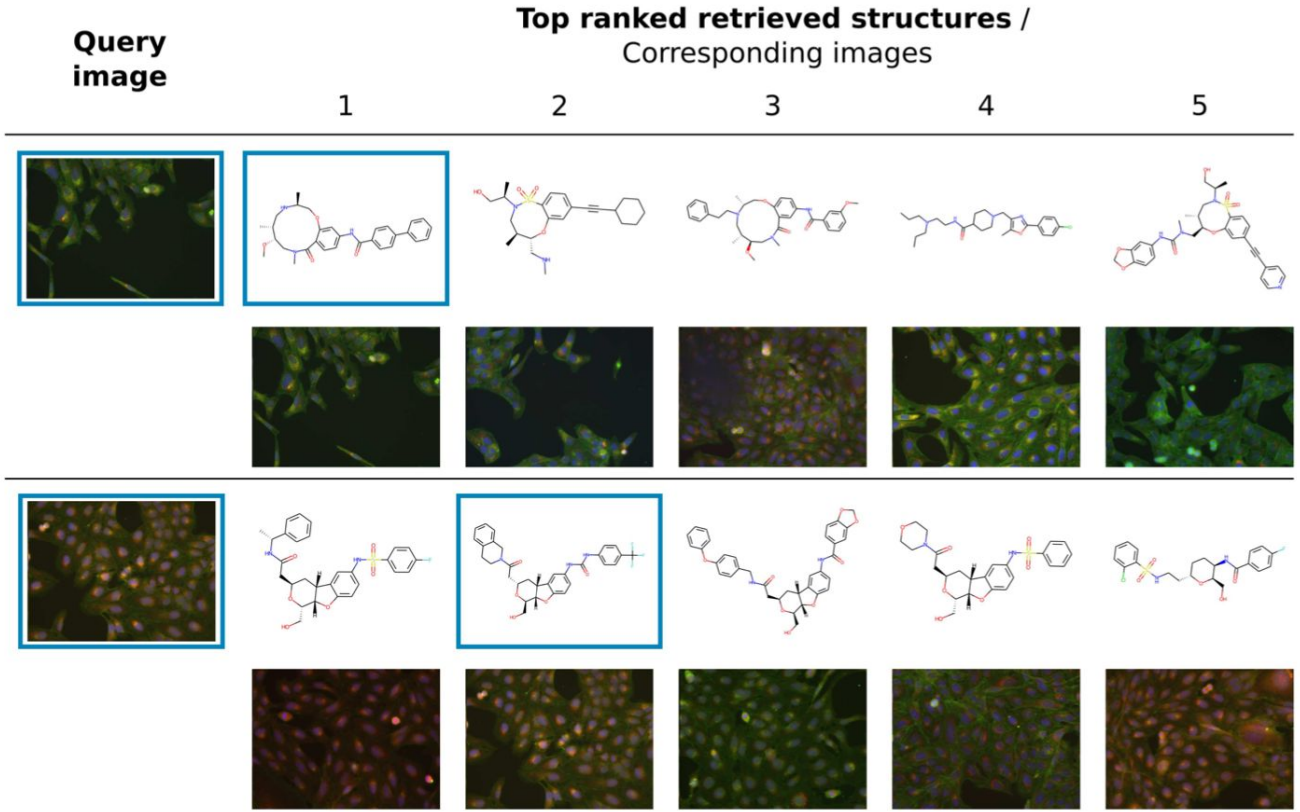






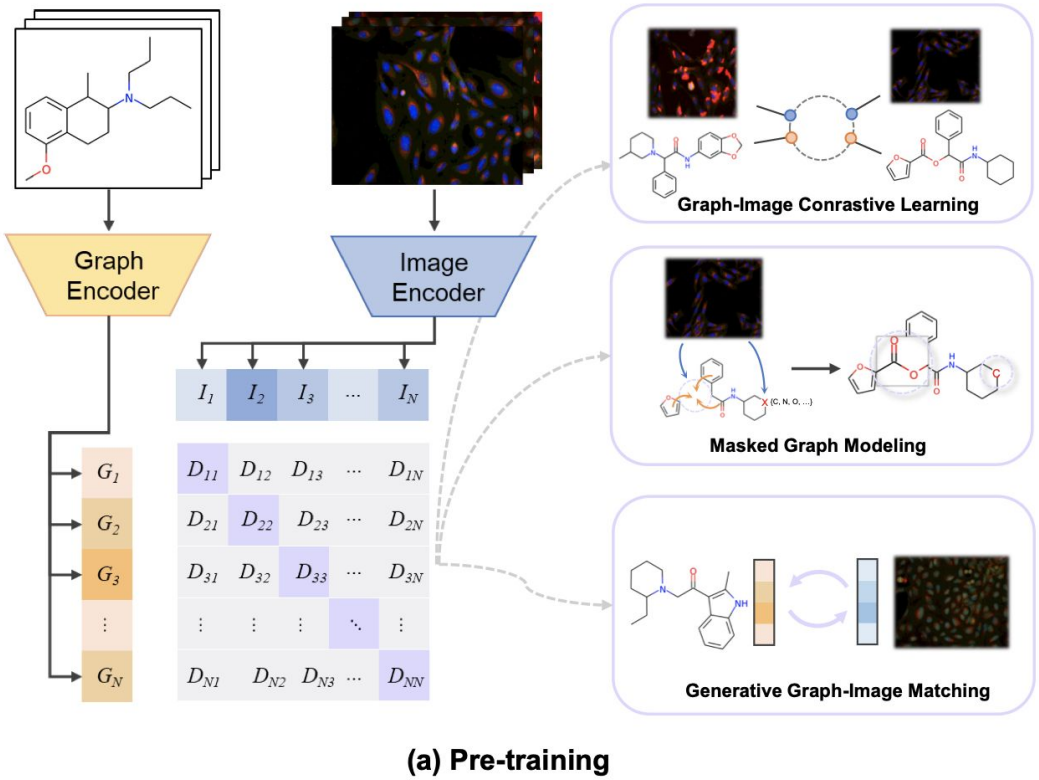
[Contrastive learning of image- and structure-based representations in drug discovery]

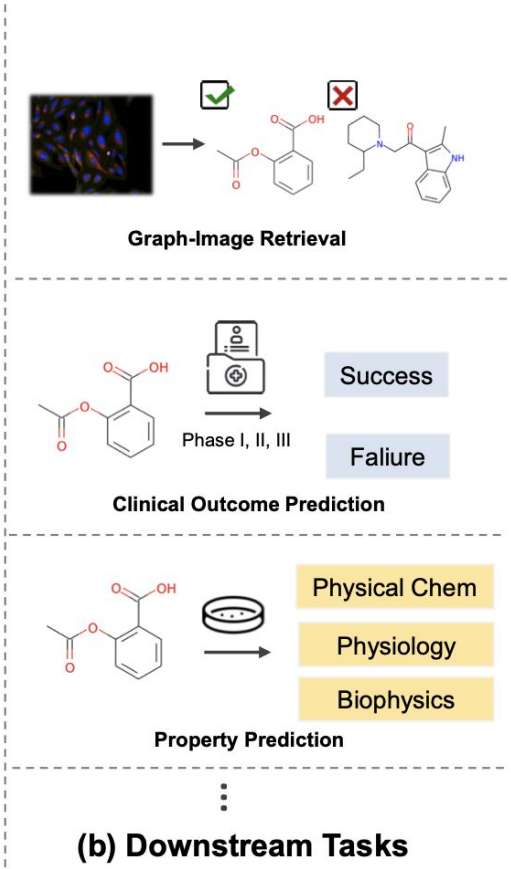
Contrastive learning of image- and structure-based representations



Contrastive learning of image- and structure-based representations in drug discovery

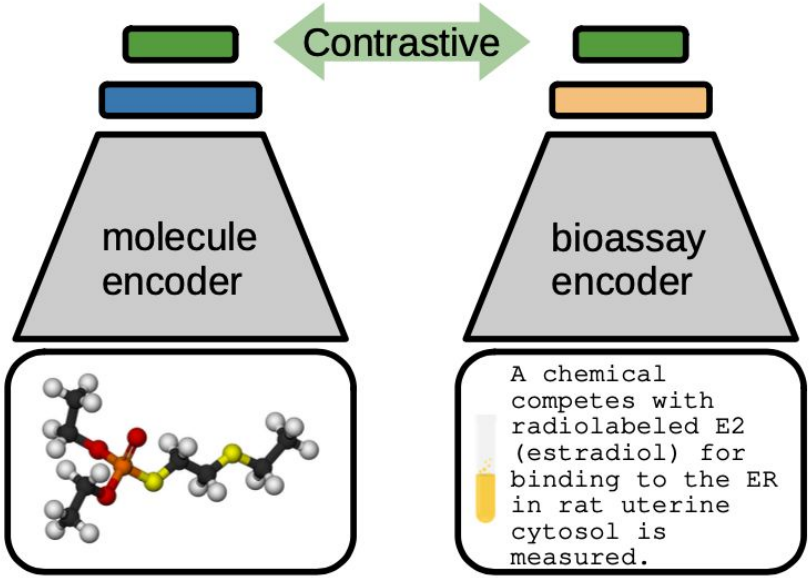
Cross-modal Graph Contrastive Learning with Cellular Images

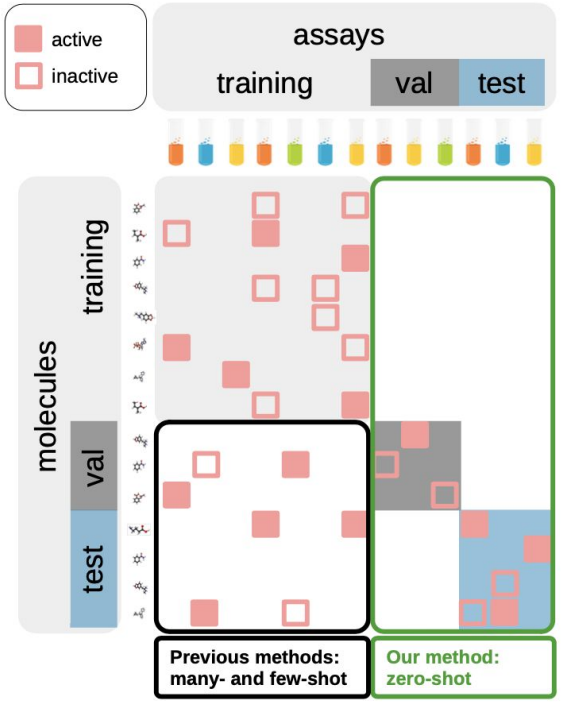




[Cross-modal Graph Contrastive Learning with Cellular Images | bioRxiv]

BioassayCLR: Prediction of biological activity for novel bioassays based on rich textual descriptions





http://www.bioinf.jku.at/people/klambauer/02.pdf

RetCL: A Selection-based Approach for Retrosynthesis via CL

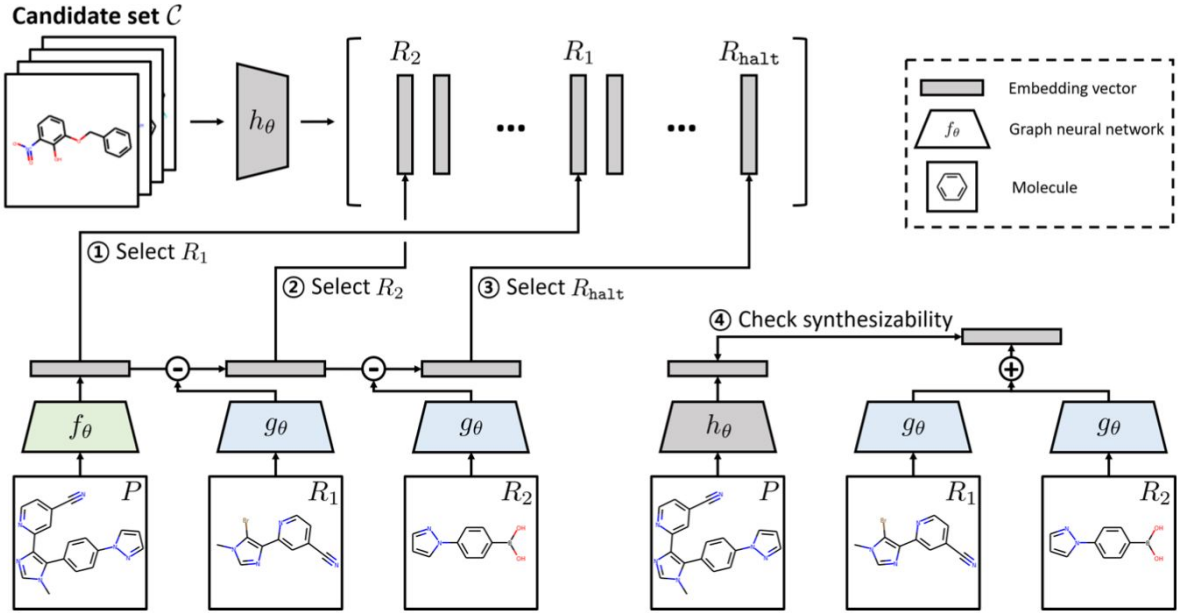


Figure 2: Illustration of the search procedure in RetCL. It first (1-3) selects reactants sequentially based on $\psi(R|P, \mathcal{R}_{\texttt{given}})$, and then (4) check the synthesizability of the selected reactant-set based on $\phi(P|\mathcal{R})$. The overall score is the average over all scores from (1) to (4).

RetCL: A Selection-based Approach for Retrosynthesis via Contrastive Learning

Outlook

- Use more than two modalities!
- How to obtain / derive multimodal datasets?
- Can we leverage the knowledge present in large LMs?
- New setups, e.g., contrastive learning & MAE?
- Special embedding spaces for special applications?

→ See also the #multimodal-drug-discovery project in the OpenBioML discord and join!

Thank you for your attention! Questions? Discussion?