An adversarial scheme for integrating multi-modal data on protein function - MIRAGE

Rami Nasser[1], Leah V Schaffer[2], Trey Ideker[2], Roded Sharan[1]

Tel Aviv University[1]
University of California, San Diego[2]

Agenda

- Problem Formulation
- Background and Previous Work
- MIRAGE method
- Results

Data - HEK293T cells

Images

Immunofluorescence images from the Human Protein Atlas. $1,125$ immunofluorescence images.

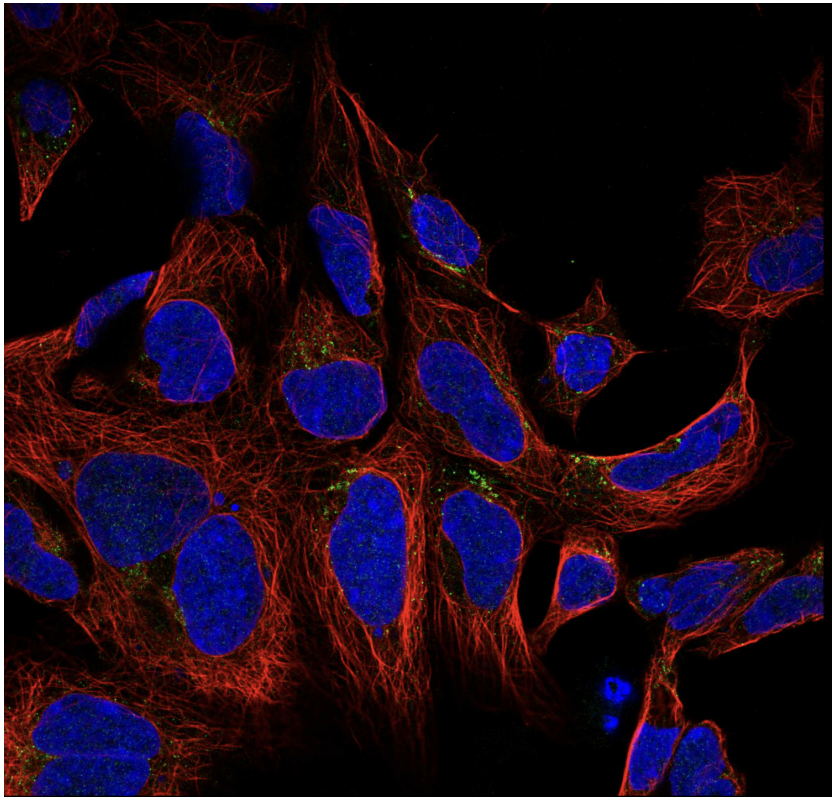
Network

Protein-protein interactions from Bioplex3.0
$14,032$ proteins and $127,732$ interactions.

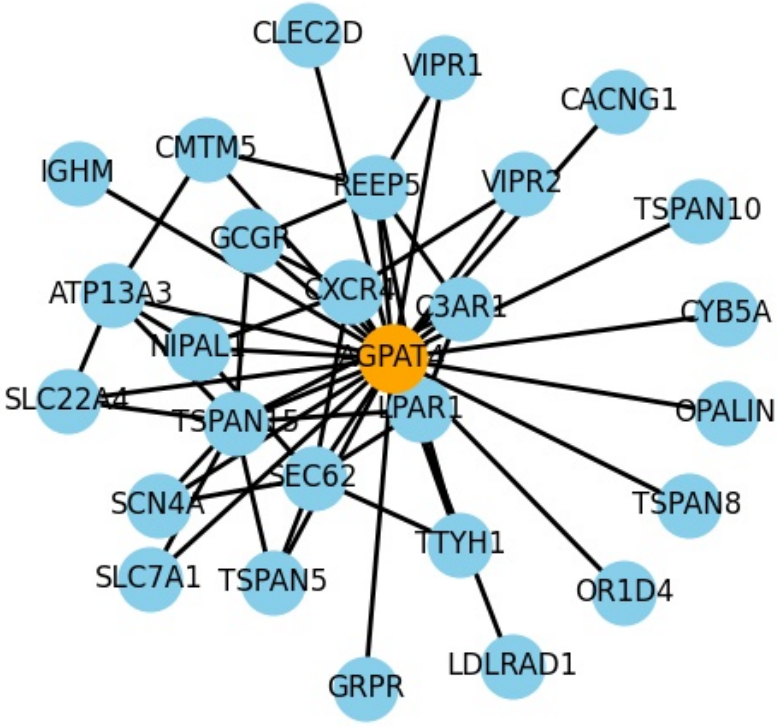
Sequence

UniProt for $20,218$ human proteins

Data - Example



(a) AGPAT4 - Golgi apparatus



(b) AGPAT4 - Neighborhood

AGPAT4-Sequence

MADTQCCPPPCEFISSAGTDLALGMGWDATLCLLPFTGFGKCAGIWNHMDEEP
DNGDDRGSRRTTGQGRKWAAHGTMAAPRVHTDYHPGGGSACSSVKVRSHVG
HTGVFFFVDQDPLAVSLTSQSLIPPLIKPGLLKAWGFLLLCAQPSANGHSLCCLLY
TDLVSHELPFRALCLGPSDAPSACASCNCLASTYYL

Problem Formulation

Given

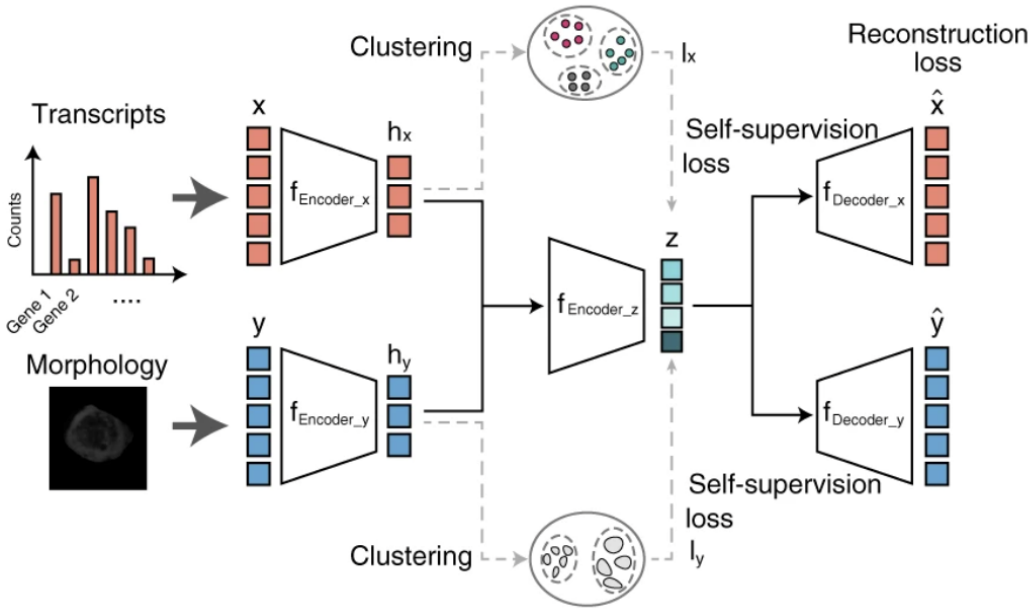
$X = \{X_1, X_2, \ldots, X_M\}$ represent the set of M different modalities for protein representation (e.g., sequence (SEQ), interaction (PPI) and localization (IMG)), where $X_i \in \mathbb{R}^{d_i}$.

Goal

Embed $X_i \underset{E_i}{\longrightarrow} Z$ into a shared latent space $Z \in \mathbb{R}^l$, and to generate $Z \underset{G_i}{\longrightarrow} X_i$ modality vector from any latent modality j, where E_i and G_i are parameterized mapping functions.

Previous Work

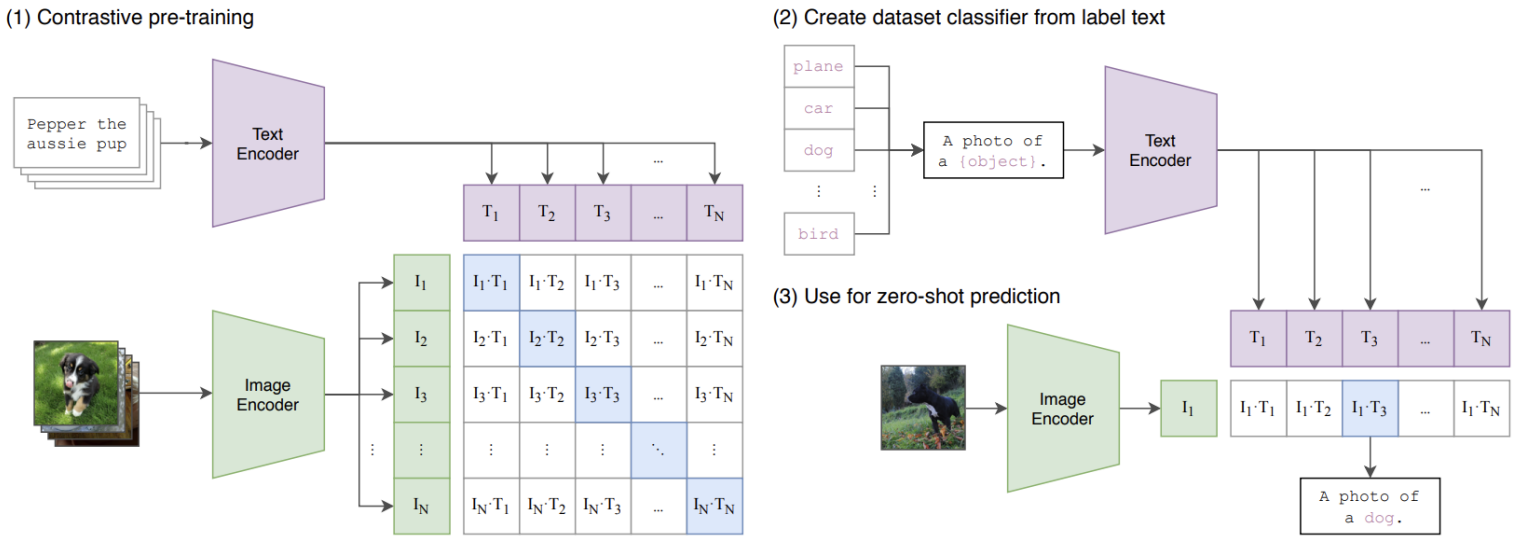
MUSE



Bao, F., Deng, Y., Wan, S. et al. Integrative spatial analysis of cell morphologies and transcriptional states with MUSE.

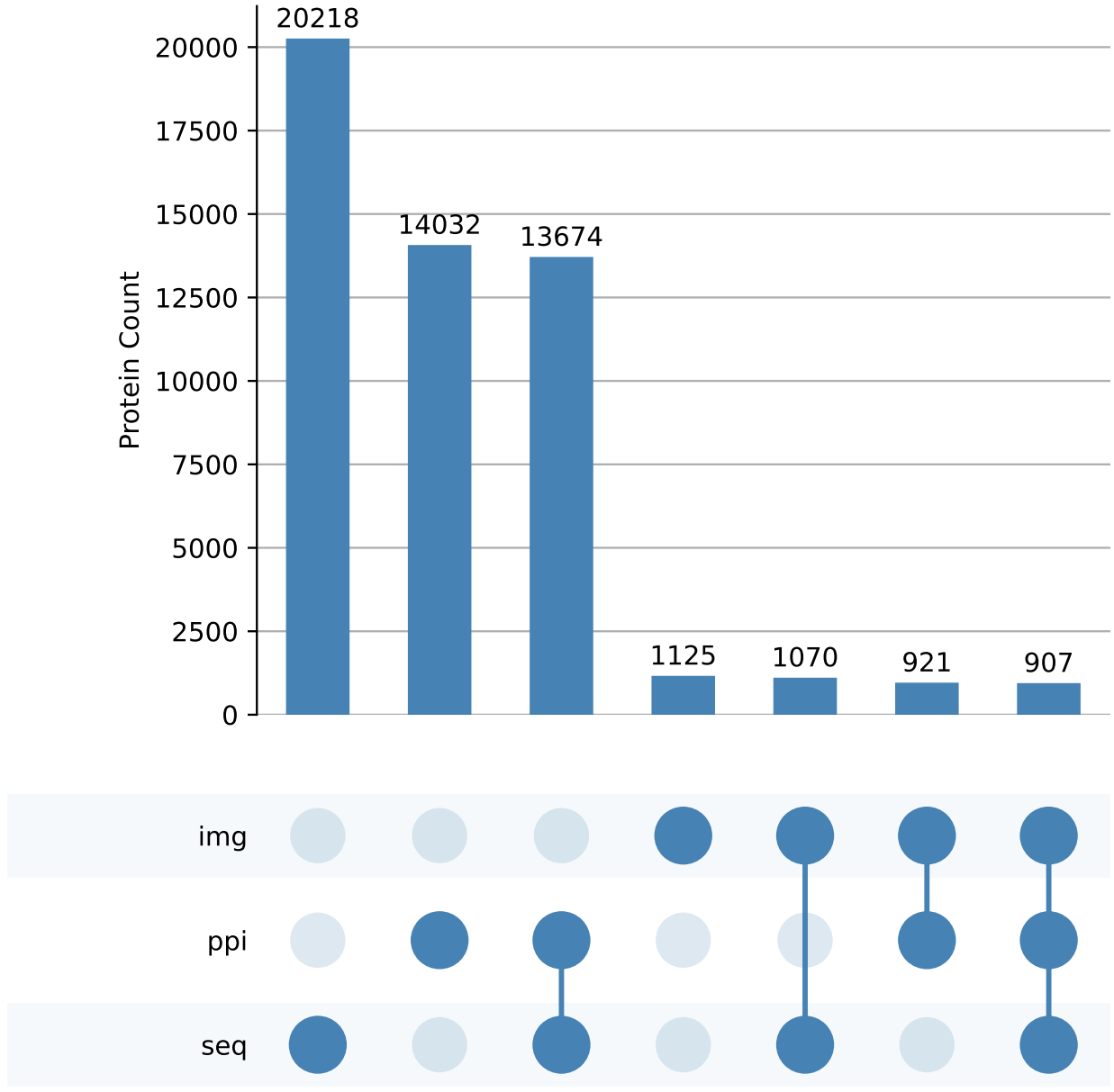
Recent Work

CLIP

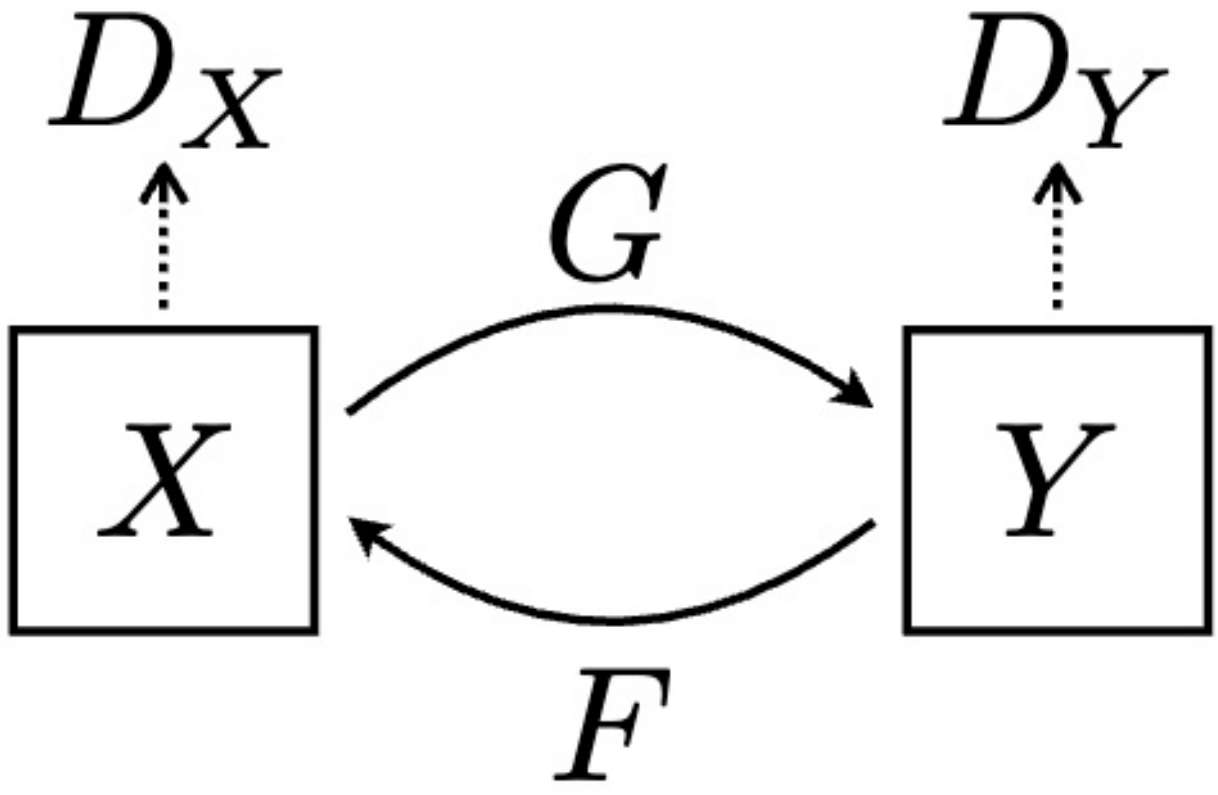


Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021, July). Learning transferable visual models from natural language supervision. In International conference on machine learning (pp. 8748-8763). PMLR.

Problem - Curse of Alignment



CycleGAN



MIRAGE-Data Preprocessing

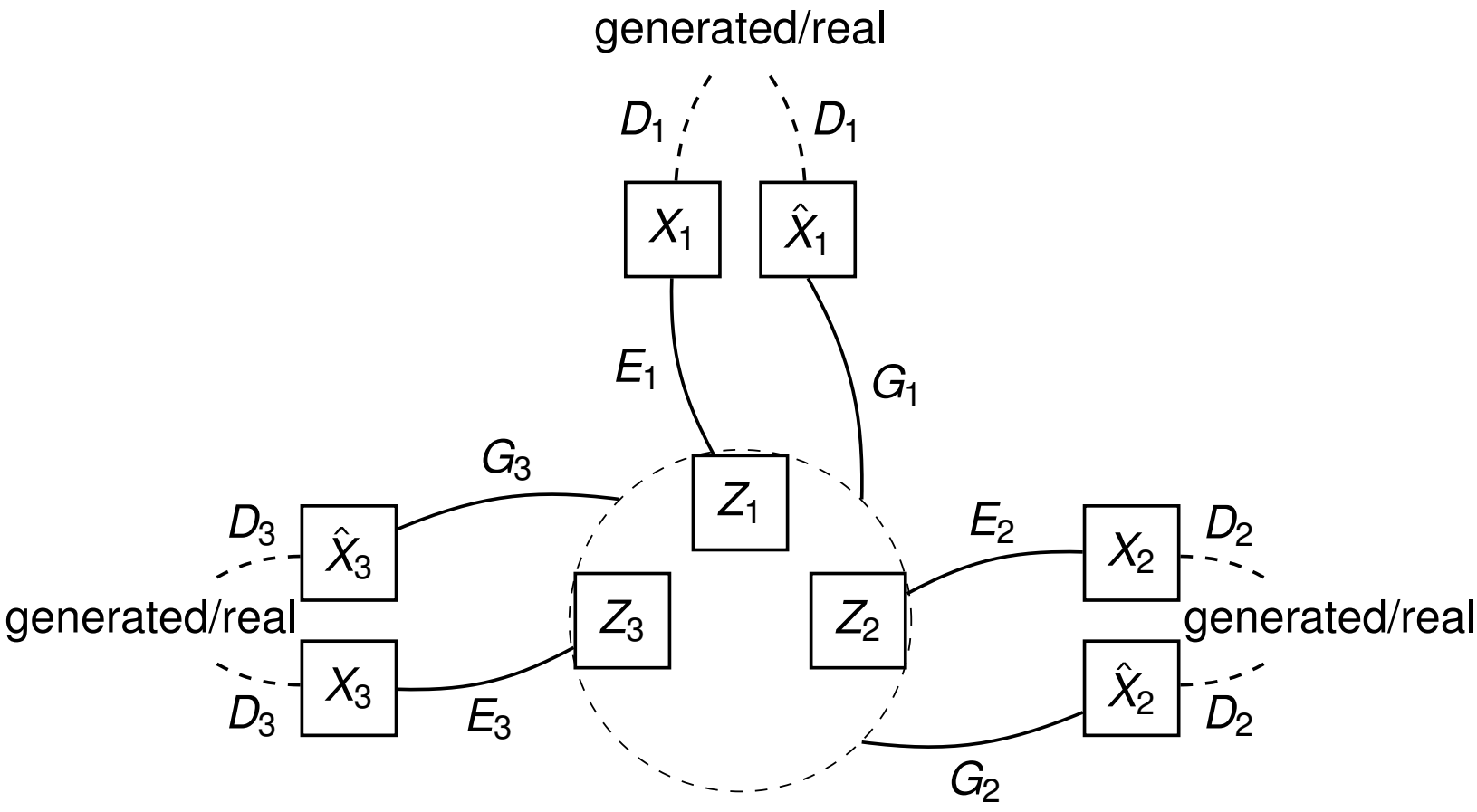
Approach

Breaking large tasks into smaller, using intermediate representations.

3 Modalities

- sequence encoding, we utilized the ESM-2.
- network encoding, we employed node2vec.
- image encoding, we use DenseNet.

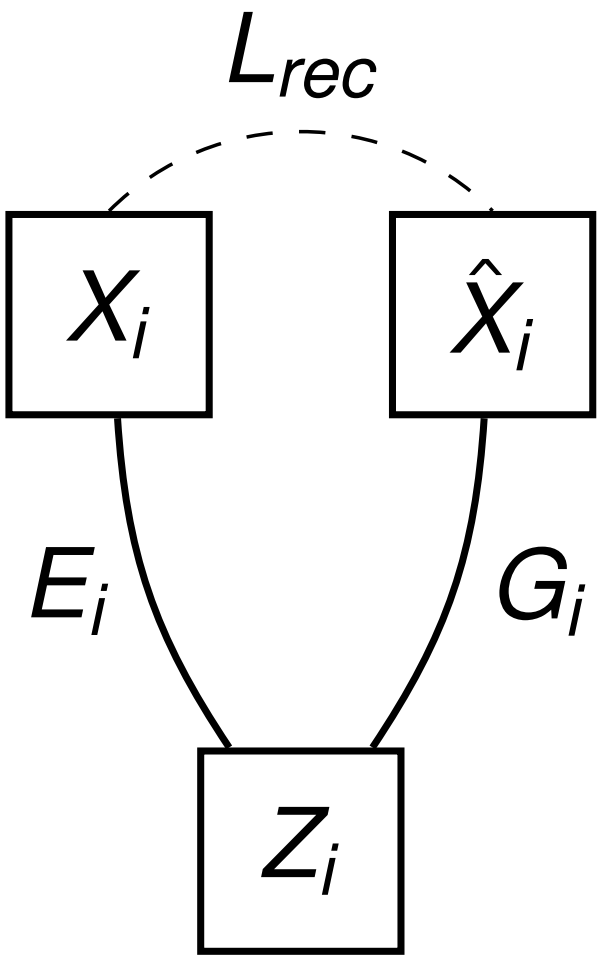
MIRAGE-Schema

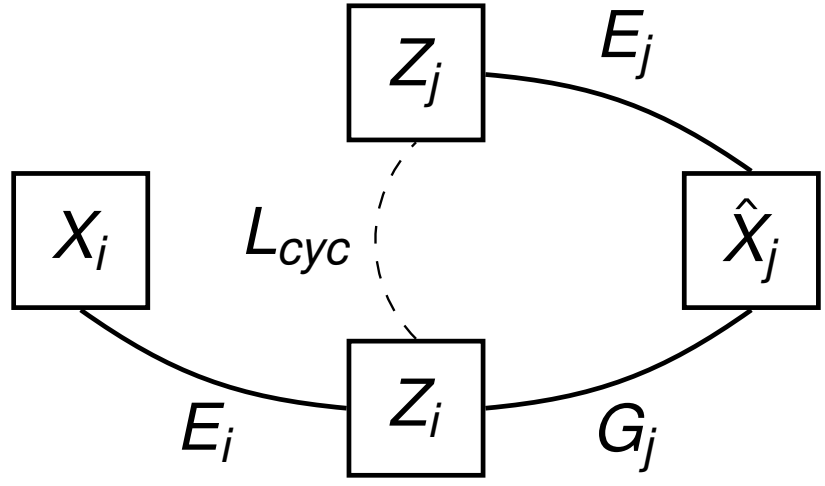


Loss

$$\mathcal{L}_{total} = \lambda_{gan}\mathcal{L}_{gan} + \lambda_{cyc}\mathcal{L}_{\text{cyc}} + \lambda_{rec}\mathcal{L}_{\text{rec}}$$

MIRAGE-Losses





Evaluation Criteria - BIONIC

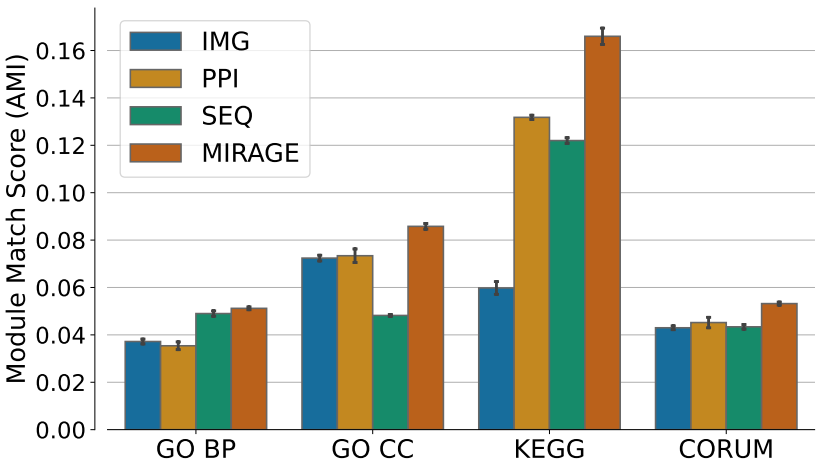
Tasks

- Module detection: Louvain clustered.
- Gene function prediction.

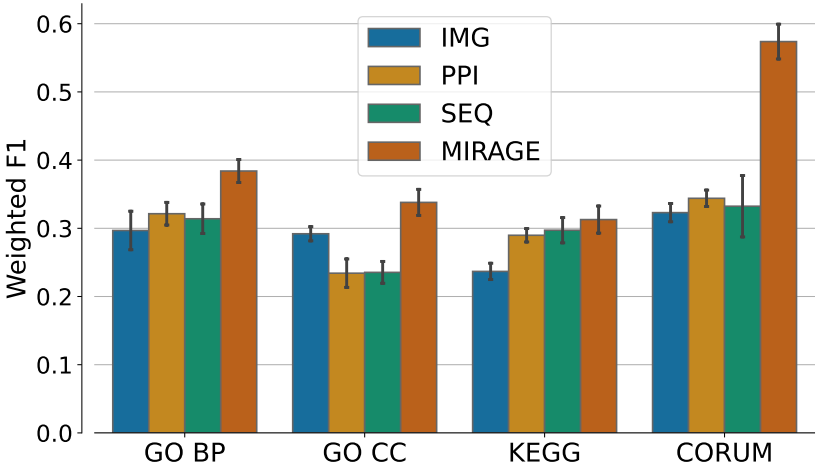
Annotated Collections of Proteins

- CORUM
- KEGG
- GO BP
- GO CC

Results-Power of Integration

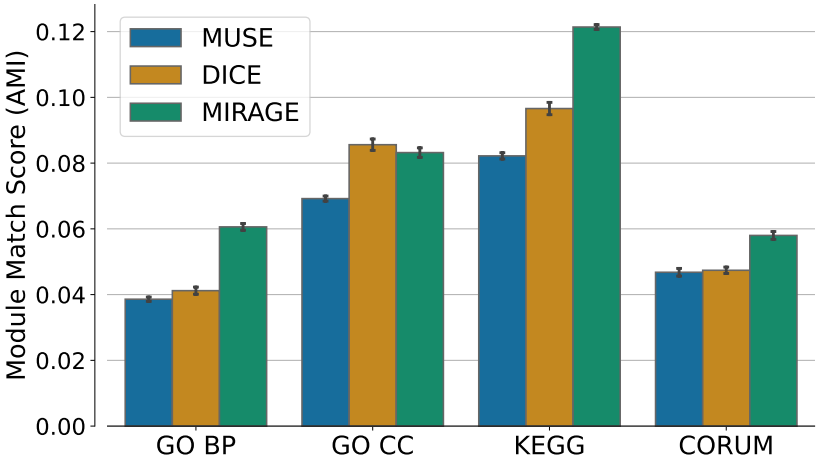


(a) Module detection

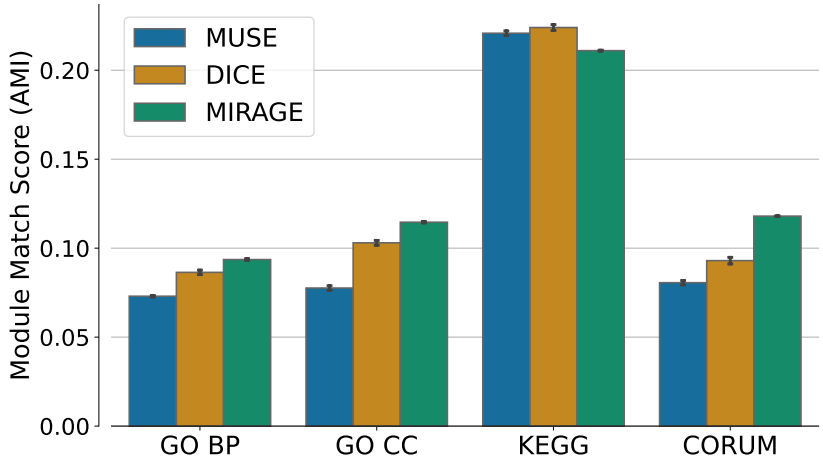


(b) Function Prediction

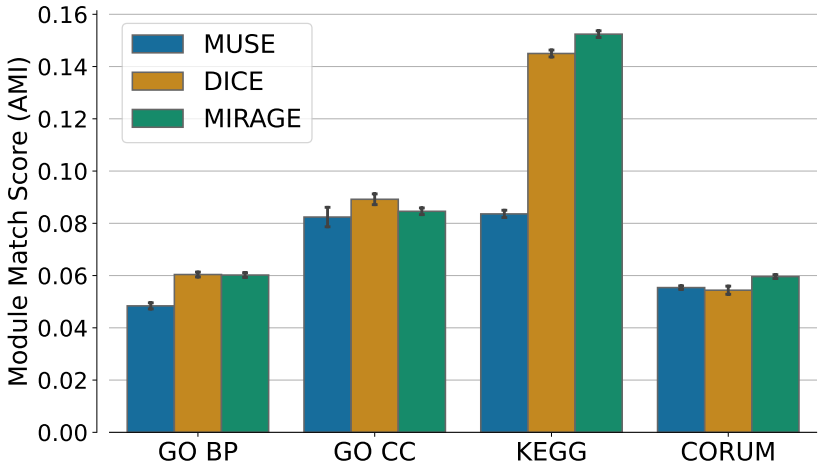
Results-2 Modalities



(a) PPI-IMG

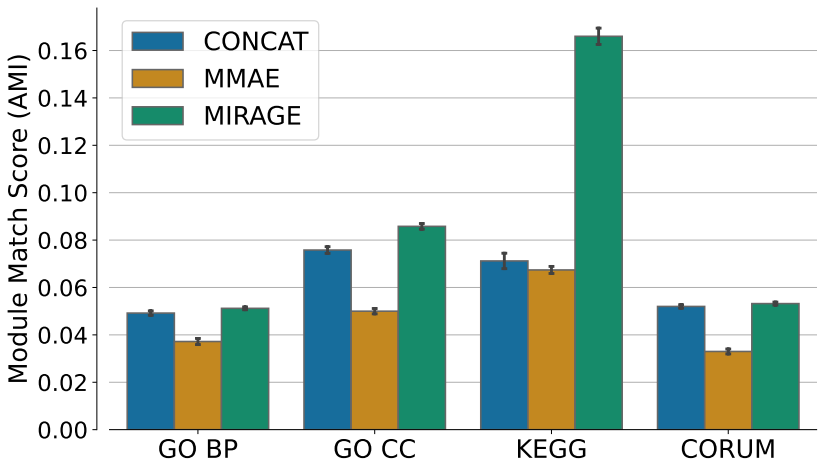


(b) PPI-IMG

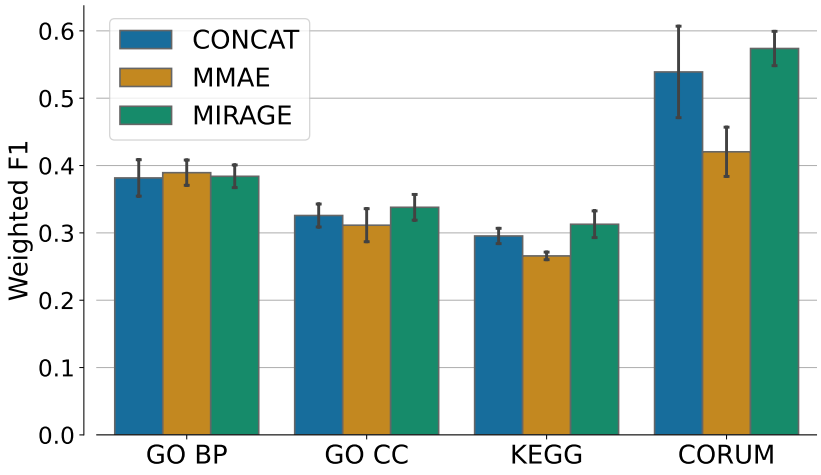


(c) IMG-SEQ

Results-Multi Modals

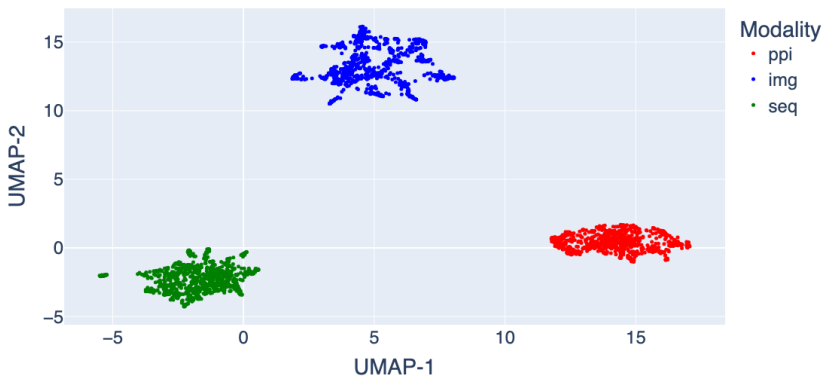


(a) Module detection

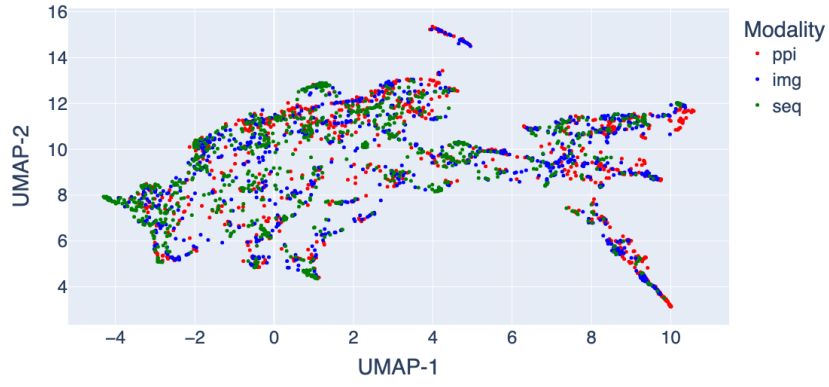


(b) Function Prediction

Results-UMAP

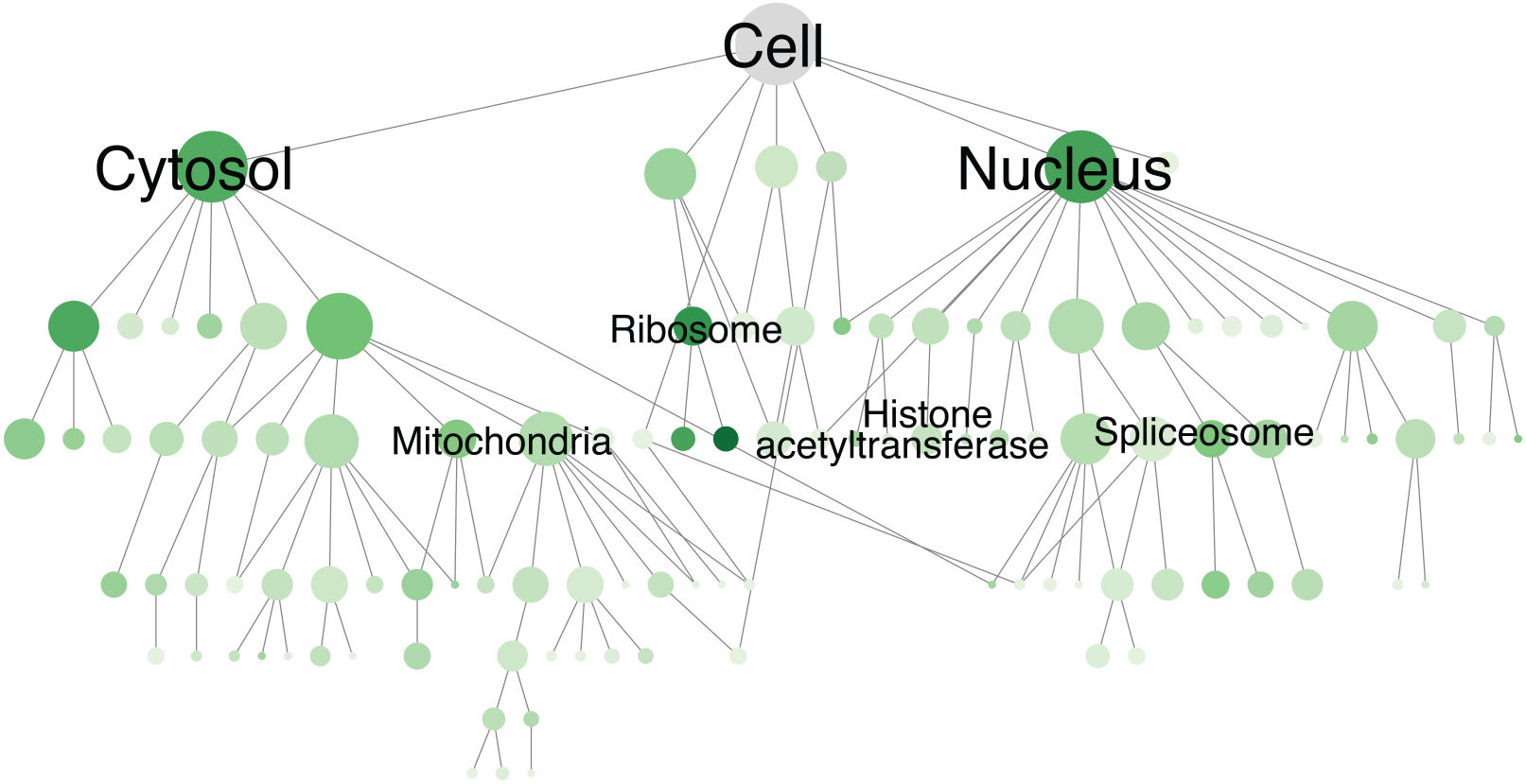


(a) Raw



(b) MIRAGE

Results-Hierarchy



Biological Findings

- 111 protein assemblies.
- 62 had significant overlap (GO, CORUM, HPA).
- recovered assemblies across scales.

Results-Fréchet Inception Distance (FID)