Multi-Modal Contrastive Learning for Proteins by Combining Domain-Informed Views

Haotian Xu [1] Yuning You [2] Yang Shen [2]

[1]Stony Brook University, Department of Applied Mathematics & Statistics
[2]Texas A&M University, Department of Electrical and Computer Engineering

Introduction & Motivation

In this project, we focus on learning a multi-modal representation of proteins, specifically how to integrate different pretrained single-modality encoders and how to explore heterogeneity for protein modalities efficiently. We compare sequential and parallel integration paradigms and propose a novel data augmentation that leverages domain-informed protein homology classes (families of similar sequences and superfamilies of similar structures) for intra- and inter-modality contrast. Numerical results indicate that the novel views and the novel ways to compose views can facilitate multi-modal synergy toward better downstream performances.

Training Diagram & Training Objective

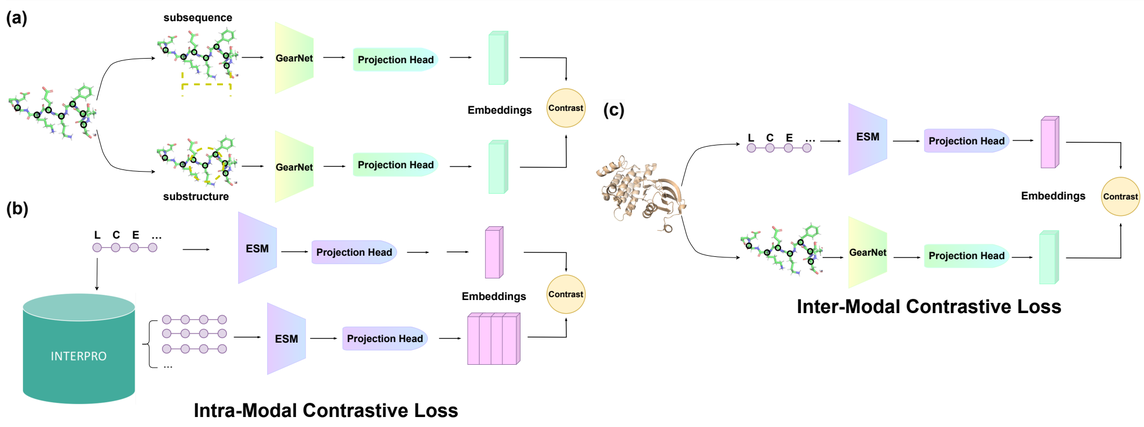

Figure 1. **(a)** uses multi-view contrast; newly proposed **(b)** uses biological domain knowledge (protein homology) as defined in InterPro to inform the design of positive and negative view-pairs, and **(c)** adapts a CLIP cross-modal contrastive learning on top of **(a)** and **(b)** that can be done for either intra-modality contrastive learning (sequence or structure). We further propose to compose views in **(a)** and **(b)** with **(c)** for (additional) inter-modality contrastive learning.

$$\mathcal{L}_{\text{Multi-modal} \circ \text{Bio-informed} \circ \text{Multi-view}} \tag{1}$$
$$= -\frac{1}{n}\sum_i \log\left(\sum_{(x^m,x^n)\in\mathcal{C}_f(x^i)\times\mathcal{C}_s(x^i)} \text{sim}((\mathcal{F}\circ g)_{\text{seq}}(\mathcal{H}_{\text{seq}}(x^m_{\text{seq}})), (\mathcal{F}\circ g)_{\text{struc}}(\mathcal{H}_{\text{struc}}(x^n_{\text{struc}})))\right)$$
$$+\frac{\tau}{n}\sum_i \log\left(\sum_{(x^m,x^n)\notin\mathcal{C}_f(x^i)\times\mathcal{C}_s(x^i)} \text{sim}((\mathcal{F}\circ g)_{\text{seq}}(\mathcal{H}_{\text{seq}}(x^m_{\text{seq}})), (\mathcal{F}\circ g)_{\text{struc}}(\mathcal{H}_{\text{struc}}(x^n_{\text{struc}})))\right)$$

Let $\mathcal{X} = \{x^1, x^2, \ldots, x^N\}$ represent the set of proteins in our dataset, and x^i_{seq} be the sequence form of x^i and x^i_{struc} be the structural description. Two encoders $F_{\text{seq}} : \mathcal{X}_{\text{seq}} \to \mathcal{Z}_{\text{seq}}$ and $F_{\text{struc}} : \mathcal{X}_{\text{struc}} \to \mathcal{Z}_{\text{struc}}$, where $\mathcal{Z}_{\text{seq}}$ and $\mathcal{Z}_{\text{struc}}$ are the hidden representation space. Two projection heads $g_{\text{seq}} : \mathcal{Z}_{\text{seq}} \to \mathcal{Z}$ and $g_{\text{struc}} : \mathcal{Z}_{\text{struc}} \to \mathcal{Z}$, where $\mathcal{Z}$ is the shared hidden space to perform inter-modality contrastive learning. $\mathcal{C}_f$ and $\mathcal{C}_s$ denote the sets of families and superfamilies.

Sequential v.s. Parallel Integration

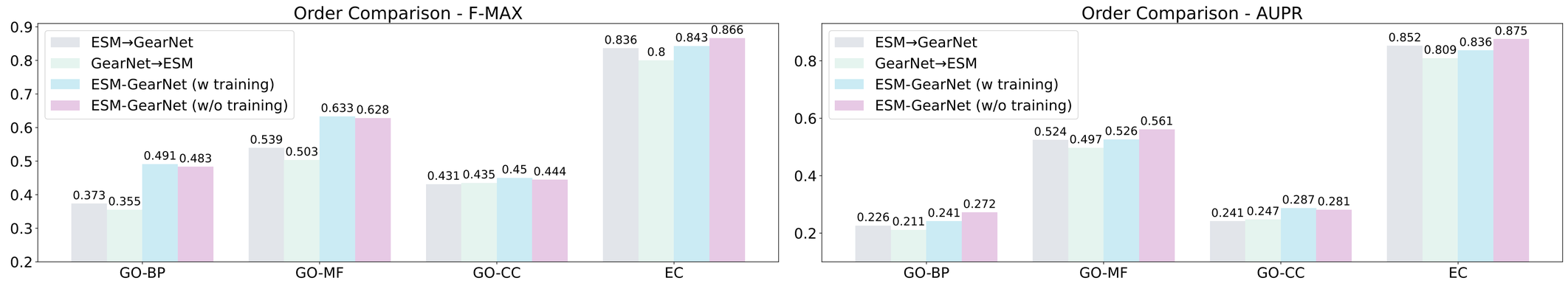

Figure 2. shows the order sensitivity of sequential integration. Though the number of parameters is the same between the two sequential orders, the Sequence-to-Structure order was better performing. To determine the optimal order in sequential integration, $n!$ sequential models need to be trained over all permutations ($n = 2$ for sequences–structure and $n = 3$ for videos of images, texts, and audios). In contrast, parallel integration is order-invariant, only needs to be trained once, and can outperform the best serial integration.

Intra- and Inter-Modality Contrast

Table 1. Combining Intra- and Inter-Modality Contrast: ① and ② are baselines, whereas ③–⑤ are our multimodal models that continue training parallel ESM-GearNet on 55k tinyAlphaFoldDB proteins.

	Method	GO-BP		GO-MF		GO-CC		EC	
		$F_{\max}$	AUPR	$F_{\max}$	AUPR	$F_{\max}$	AUPR	$F_{\max}$	AUPR
①	GearNet-Edge(Reproduce)	0.493	0.234	**0.635**	0.531	0.447	0.242	0.845	0.834
	ESM-1b(Reproduce)	0.394	**0.277**	0.519	0.521	0.403	**0.326**	0.809	0.824
②	ESM-GearNet (no continue training)	0.483	0.272	0.628	0.561	0.444	0.281	**0.866**	**0.875**
③	ESM-GearNet(N/A, $\mathcal{L}_{\text{Mm}}$)	0.500	0.226	0.632	0.540	**0.450**	0.253	0.850	0.850
④	ESM-GearNet($\mathcal{L}_{\text{Mv}}$, $\mathcal{L}_{\text{Mm}}$)	**0.503**	0.229	0.627	0.547	0.440	0.255	0.846	0.827
	ESM-GearNet($\mathcal{L}_{\text{Bio}}$, $\mathcal{L}_{\text{Mm}}$)	0.491	0.241	0.633	0.526	**0.450**	0.287	0.843	0.836
⑤	ESM-GearNet($\mathcal{L}_{\text{Bio}\circ\text{Mv}}$, $\mathcal{L}_{\text{Mm}}$)	0.492	0.224	0.621	**0.565**	0.447	0.247	0.840	0.822

Table 2. Mean values and standard deviations of $F_{\max}$ compared between the starting point (no continual training) and one of our models (⑤ in Table 1).

Method	GO-BP $F_{\max}$	GO-MF $F_{\max}$	GO-CC $F_{\max}$	EC $F_{\max}$
ESM-GearNet(no train)	48.38%(±0.17%)	**64.48%**(±0.27%)	45.49%(±0.24%)	86.37%(±0.06%)
ESM-GearNet($\mathcal{L}_{\text{Bio}\circ\text{Mv}}$, $\mathcal{L}_{\text{Mm}}$)	**48.59%**(±0.12%)	64.45%(±0.24%)	**45.92%**(±0.25%)	**86.92%**(±0.23%)

Table 3. Downstream fine-tuned performances.

Augmentations \ Tasks	GO-BP	GO-MF	GO-CC	EC
Mm (Identity)	0.500 (0.226)	0.632 (0.540)	**0.450** (0.253)	**0.850** (**0.850**)
Mm ∘ Mv (Identity cropping)	0.498 (**0.229**)	0.638 (0.514)	0.448 (**0.263**)	0.843 (0.831)
Mm ∘ Bio (Homology)	**0.503** (0.222)	0.633 (**0.556**)	0.444 (**0.263**)	0.845 (0.830)
Mm ∘ Bio ∘ Mv (Homology cropping)	0.501 (0.226)	**0.643** (0.530)	0.448 (0.250)	0.841 (0.827)

Dataset

Table 4. Statistics of TinyAlphaFoldDB data with homology classifications

Size \ Data	Foldseek	Family	Superfamily	Prot Has Family	Prot Has Superfamily
# samples	55,189	9,361	2,363	26,883	30,068

We apply foldseek clustering, a structural-alignment-based clustering, on AlphaFold database v1 and v2 of predicted protein structures. In this way, we select about 55k cluster representatives out of 1M instances. We used InterPro to classify those 55K proteins into families and superfamilies.

Conclusion & Discussion

In this work, we aim at multimodal representation learning for proteins while overcoming resource limitations and modality heterogeneity. We first investigate and show the effects of different model integration strategies, in which parallel integration stands out. Next, we leverage domain-informed protein homology classes to design novel data views, which further addresses a type of under-explored heterogeneity for protein modalities. Numerical results indicate that the novel views and the novel ways to compose views can facilitate multi-modal synergy toward better downstream performances.