MULTI-MODAL CONTRASTIVE LEARNING FOR PROTEINS BY COMBINING DOMAIN-INFORMED VIEWS

Haotian Xu *
Applied Mathematics & Statistics
Stony Brook University
Stony Brook, NY 11794, USA
haotian.xu@stonybrook.edu

Yuning You & Yang Shen
Department of Electrical and Computer Engineering
Texas A&M University
College Station, TX 77843, USA
{yuning.you, yshen}@tamu.edu

ABSTRACT

Proteins, often represented as multi-modal data of 1D sequences and 2D/3D structures, provide a motivating example for the communities of machine learning and computational biology to advance multi-modal representation learning. Protein language models over sequences and geometric deep learning over structures learn excellent single-modality representations for downstream tasks. It is thus desirable to fuse the single-modality models for better representation learning. But it remains an open question on how to fuse them effectively into multi-modal representation learning, especially with a modest computational cost yet significant downstream performance gain. To answer the question, we propose to make use of separately pretrained single-modality models, integrate them in parallel connections, and continuously pretrain them end-to-end under the framework of multimodal contrastive learning. The technical challenge is to construct views for both intra- and inter-modality contrasts while addressing the heterogeneity of various modalities, particularly various levels of semantic robustness. We address the challenge by using domain knowledge of protein homology to inform the design of positive views, specifically protein classifications of families (based on similarities in sequences) and superfamilies (based on similarities in structures). We also assess the use of such views compared to, together with, and composed to other positive views such as identity and cropping. Extensive experiments on enzyme classification and protein function prediction benchmarks demonstrate the potential of domain-informed view construction and combination in multi-modal contrastive learning.

1 INTRODUCTION

Proteins are essential molecules in living organisms that not only serve as building blocks of cells but also play a pivotal role in various functional interactions across cells (Alberts, 2017). *In-silico* modeling of protein molecules is treated as one of the stepping stones toward the ultimate goal of creating “digital twins" for health care. The scope of the paper is on protein representation learning. The goal is to learn a mapping that transforms (or embeds) complicated protein molecules into machine-readable fixed-dimensional vectors, which can be utilized for various downstream applications.

Proteins are often described as data of a multi-modality nature such as 1D amino-acid sequences and 2D/3D residue- or atom-resolution structures. In nature, the roles and functions of proteins are believed to be associated with their diverse modality features. Single-modality models have been progressing to embed 1D sequences and 2D/3D structures. Recently multi-modal representation learning for proteins, demanded by the functional and evolutionary coupling across their individual modalities, is a nascent topic of growing interest. Related works are detailed in Sec. 2.

In this study, we leverage *multi-modal contrastive learning* for protein representation and address the following questions:

*This work started when the author was in an M.S. program at Texas A&M University.

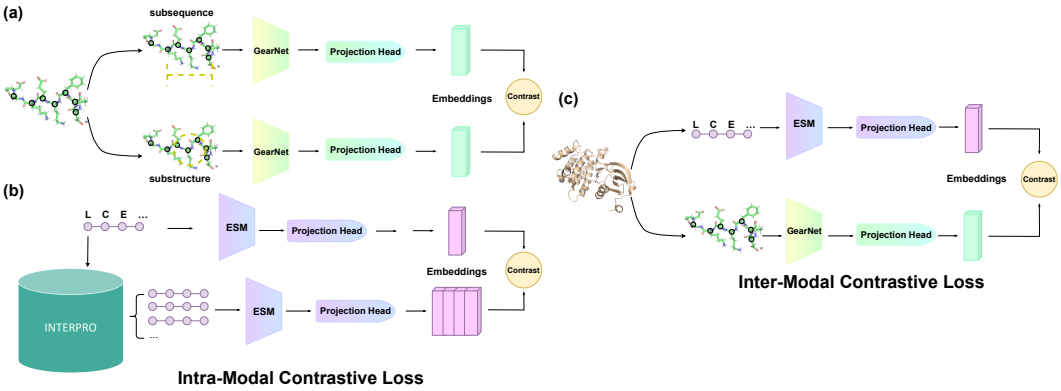

Figure 1: **(a)** uses multi-view contrast (cropping (Zhang et al., 2023b)); newly proposed **(b)** uses biological domain knowledge (protein homology) as defined in InterPro to inform the design of positive and negative view-pairs, and **(c)** adapts a CLIP cross-modal contrastive learning on top of **(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.

(i) How to achieve multi-modal representation learning, often with more complex models and larger data corpora than single-modal counterparts, while maintaining low computational cost? We utilize separately pre-trained uni-modal encoders, curate lean data with paired modalities, and continue training the integrated encoders (see **(ii)**) end to end through contrastive learning (see **(iii)**).

(ii) How to integrate single-modality encoders for effective multi-modal learning? Unlike previously found powerful serial integration, we adopt parallel integration that are modality order-invariant and robustly competitive.

(iii) How to incorporate intra- and inter-modality contrast while addressing the heterogeneity in multi-modalities? Modalities not only differ in data forms, thus impacting encoders. They also differ in semantic robustness in the latent space (specifically here, how much unit perturbations of sequence or structure embeddings change protein functions), thus impacting views to contrast, which is under-explored for protein representation learning. We leverage protein homology classes (families of similar sequences and superfamilies of similar structures) to design novel augmentations and combine them in various ways with previously adopted intra- and inter-modality views (Figure 1).

2 Related Work

2.1 Protein Representation Learning

Previous studies on protein representation learning isolate or fuse different modalities. In uni-modal settings, Protein Language Models (PLMs) (Brandes et al., 2022; Lu et al., 2020; Rives et al., 2019; Lin et al., 2022; Zhang et al., 2022; Xu et al., 2023; Talukder et al., 2022) apply self-supervised learning objectives such as autoregressive modeling (Notin et al.; Madani et al., 2023; Hesslow et al.; Elnaggar et al., 2021; 2023), mask modeling (Brandes et al., 2022; Rives et al., 2019; Lin et al., 2022; He et al., 2021), and contrastive learning (Lu et al., 2020) on large-scale protein sequence corpora to learn informative and effective protein representations. On the other hand, structure-based encoders, especially geometric deep learning, can capture different granularities of protein structures, including residue-level structures (Gligorijević et al., 2021; Zhang et al., 2023b; Xu et al., 2022), atom-level structures (Jing et al., 2009; Hermosilla et al., 2020), and protein surfaces (Gainza et al., 2020; Sverrisson et al., 2021). Taking advantage of uni-modal encoders, multi-modal encoders can even boost protein function understanding and protein design through better awareness across modalities (Zhang et al., 2023a; Chen et al., 2023; Sun & Shen, 2023; Wang et al., 2022; You & Shen, 2022).

2.2 Contrastive Learning with Data Augmentations

Contrastive learning seeks to maximize the mutual information between anchors and their positive pairs, which are the outputs from various augmentation functions. In visual learning Chen et al. (2020) proposes SimCLR algorithm that contrasts augmented images and He et al. (2020) chooses to build

a dynamic dictionary and further facilitates contrastive unsupervised learning; in natural language processing, Gao et al. (2021) takes dropout randomness as a unique augmentation technique and shows contrastive objective effectively uniforms pre-trained embeddings' anisotropic space; Xu et al. (2021) utilizes retrieval augmentation to select harder negatives and achieves better performance even compared to supervised setting for video–text understanding; and in graph learning You et al. (2020; 2021; 2022; 2023); Wei et al. (2022) investigate the impact of parameterized graph augmentations' extents and patterns. Cross-modal contrastive learning has also been explored in CLIP (Radford et al., 2021) for text-image data. In the domain of protein sequence–structure data, Zhang et al. (2023b) designs subsequence and substructure (cropping) views for multi-view contrast to pre-train structure encoders; and Chen et al. (2023) uses identical proteins in sequence and structure views as positive pairs to pre-train sequence and structure encoders.

3 METHODOLOGY

Setup. 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.

Intra- and Inter-modality augmentations.

Multi-view contrast (Zhang et al., 2023b) uses augmentations $\mathcal{H}_{\text{seq}}$ and $\mathcal{H}_{\text{struc}}$, i.e., subsequence cropping and substructure (subspace) cropping. leading to the following loss:

$$\begin{aligned}\mathcal{L}_{\text{Multi-view}} = &-\frac{1}{n}\sum_i \log\left(\text{sim}((\mathcal{F}\circ g)(\mathcal{H}_{\text{seq}}(x^i)), (\mathcal{F}\circ g)(\mathcal{H}_{\text{struc}}(x^i)))\right) \\ &+\frac{\tau}{n}\sum_i \log\left(\sum_{j\neq i}\text{sim}((\mathcal{F}\circ g)(\mathcal{H}_{\text{seq}}(x^i)), (\mathcal{F}\circ g)(\mathcal{H}_{\text{struc}}(x^j)))\right),\end{aligned} \tag{1}$$

where $\mathcal{F}$ and g denote encoder and projection head; and $\text{Sim}(\cdot,\cdot)$ is the exponential cosine similarity function.

In multi-modal learning, we adapt CLIP (Radford et al., 2021) training objective and regard modalities as views (augmentations). Positives are sequence and structure representations from identical proteins:

$$\begin{aligned}\mathcal{L}_{\text{Multi-modal}} = &-\frac{1}{n}\sum_i \log\left(\text{sim}((\mathcal{F}\circ g)_{\text{seq}}(x^i_{\text{seq}}), (\mathcal{F}\circ g)_{\text{struc}}(x^i_{\text{struc}}))\right) \\ &+\frac{\tau}{n}\sum_i \log\left(\sum_{j\neq i}\text{sim}((\mathcal{F}\circ g)_{\text{seq}}(x^i_{\text{seq}}), (\mathcal{F}\circ g)_{\text{struc}}(x^j_{\text{struc}}))\right).\end{aligned} \tag{2}$$

Further, we propose homology-informed augmentations as in Eq. 3.

We further extend the idea of data augmentation by using bio-informed protein classifications based on homology, such as family (similar in sequences) and super-family (similar in structures). Let $\mathcal{C}$ denote a mapping that returns a corresponding set of family members or super-family members for input proteins. Thus, we have a new contrastive objective:

$$\begin{aligned}\mathcal{L}_{\text{Bio-informed}} = &-\frac{1}{n}\sum_i \log\left(\sum_{x^j\in\mathcal{C}(x^i)}\text{sim}((\mathcal{F}\circ g)(x^i), (\mathcal{F}\circ g)(x^j))\right) \\ &+\frac{\tau}{n}\sum_i \log\left(\sum_{x^j\notin\mathcal{C}(x^i)}\text{sim}((\mathcal{F}\circ g)(x^i), (\mathcal{F}\circ g)(x^j))\right),\end{aligned} \tag{3}$$

where $\mathcal{F}$ and g denote encoder and projection head; and $\text{Sim}(\cdot,\cdot)$ is the exponential cosine similarity function. $\mathcal{L}_{\text{Bio-informed}}$ can be either $\mathcal{L}_{\text{family}}$ or $\mathcal{L}_{\text{superfamily}}$ by choosing the corresponding classification $\mathcal{C}$. One can even combine $\mathcal{L}_{\text{family}}$ and $\mathcal{L}_{\text{superfamily}}$ to learn the hierarchy between them.

A visualization of the three data augmentations can be reached in Figure 1. The 3 data augmentation approaches yield different loss terms, and one can trivially show that those augmentations are orthogonal to each other in terms of methodology. According to Chen et al. (2021), the first term in contrastive objective called alignment loss is to reduce the uncertainty of the other views given one view and the latter distribution loss can be considered as a proxy to entropy $H(\mathcal{X})$ for maximizing the entropy in the representation of the example. The 3 augmentations above all encourage representations of augmented views to be consistent and to match a prior evolutionary origin distribution.

Composed augmentations. The three positive pairs, identity in Eq. 2, cropping in Eq. 1 (multi-view), and homology in Eq. 3 (bio-informed) can be composed, leading to the following loss:

$$\begin{aligned}
&\mathcal{L}_{\text{Multi-modal}\circ\text{ Bio-informed }\circ\text{ Multi-view}} \qquad (4)\\
&= -\frac{1}{n}\sum_i \log\left(\sum_{(x^m,x^n)\in\mathcal{C}_f(x^i)\times\mathcal{C}_s(x^i)} \text{sim}((\mathcal{F}\circ g)_{\text{seq}}(\mathcal{H}_{\text{seq}}(x^m_{\text{seq}})),(\mathcal{F}\circ g)_{\text{struc}}(\mathcal{H}_{\text{struc}}(x^n_{\text{struc}})))\right)\\
&\quad + \frac{\tau}{n}\sum_i \log\left(\sum_{(x^m,x^n)\notin\mathcal{C}_f(x^i)\times\mathcal{C}_s(x^i)} \text{sim}((\mathcal{F}\circ g)_{\text{seq}}(\mathcal{H}_{\text{seq}}(x^m_{\text{seq}})),(\mathcal{F}\circ g)_{\text{struc}}(\mathcal{H}_{\text{struc}}(x^n_{\text{struc}})))\right),
\end{aligned}$$

where $\mathcal{C}_f$ and $\mathcal{C}_s$ denote the sets of families and superfamilies, respectively, for any given protein. The losses for composing multi-modal contrasts with multi-view or bio-informed augmentations can be found in Appendix A.

4 Experiments and Results

4.1 Experiment Setup

Pre-training Dataset We apply foldseek clustering (Barrio-Hernandez et al., 2023), a structural-alignment based clustering, on AlphaFold database v1 and v2 of predicted protein structures (Varadi et al., 2022; Zhang et al., 2023b). In this way, we select about 55k cluster representatives out of 1M instances. We also used InterPro to classify those 55K proteins into families and superfamilies, when applicable. More details can be found in Appendix B.

Downstream Tasks We adopt two function prediction tasks proposed in Gligorijević et al. (2021) for downstream evaluation. Gene Ontology (GO), a standardized system for annotating genes and gene products across different species, aims to describe the molecular functions (MF), biological processes (BP), and cellular components (CC) of genes and gene products in a structured and hierarchical manner. Enzyme Commission (EC) classification is for categorizing enzymes based on their catalytic functions.

Baselines We choose separately pretrained ESM and GearNet and use them in parallel (dubbed ESM-GearNet). ESM-GearNet without continue-training and with solely inter-modality contrastive learning (Multi-modal) serve as baselines. Additional baselines include ESM and GearNet in series (denoted ESM→GearNet, which was found empirically powerful (Zhang et al., 2023b)).

Training We take the pre-trained checkpoints for ESM-1b (Rives et al., 2019) and GearNet-Edge (Zhang et al., 2023b) as model initializations and continue training the parallelly connected ESM-GearNet end-to-end for 5 epochs, using various contrastive learning on our tinyAlphaFoldDB dataset. The resulting models will then fine-tune on EC and GO tasks for 50 epochs. We follow the default hyperparameter configuration in Zhang et al. (2023a) for continued training and fine-tuning. For Multiview Contrast, we use the cropping length of subsequence as 50, and the mask rate of random edge masking as 0.15. The temperature in InfoNCE loss is set as 0.07. The model is continue-trained with batch size 8 and learning rate 2e-4 on 4 Tesla A100 GPUs. We use batch size 8 and learning rate 1e-4 on 1 Tesla A100 GPU for downstream evaluation. All these models are implemented with TorchDrug library (Zhu et al., 2022).

4.2 ANALYSIS

Sequential v.s. Parallel Integration

Figures 2 and 3 (Appendix C) show the order sensitivity of sequential integration. Though the number of parameters is the same between the two sequential orders, the Sequence-to-Structure order (ESM→GearNet as adopted in Zhang et al. (2023a)) was better performing. To determine the optimal order in sequential integration, $n!$ sequentia 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.

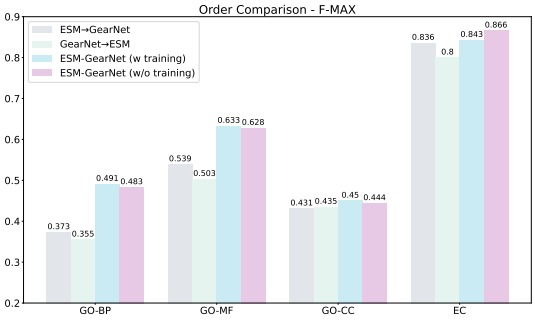

Figure 2: Comparing Fmax over the EC and GO benchmark sets between *Sequential Integration* with various orders (Zhang et al., 2023a) and *Parallel Integration.*

Intra- and Inter-Modality Contrast: As shown in Table 1, models trained under multi-modal setting perform better than those under uni-modal setting (only sequence or structure inputs). Moreover, using Bio-informed loss to guide intra-contrastive learning, which addresses the heterogeneity across modalities, achieved robust performances when used together with multi-modal contrastive learning. Performances in molecular function (GO-MF) and cellular component (GO-CC) are much improved compared to intra-modality Multi-view loss. Overall, our models outperformed the baselines in $F_{\max}$ for 2 of the 4 downstream tasks and were less successful in improving AUPR.

We also note that intra-modality contrastive learning with Multi-view loss fails to outperform baselines in some of the downstream tasks. The reason might be that 1) GearNet was pre-trained with Multi-view contrast on 1M AlphFoldDB dataset, as our foldseek dataset is a small subset of the original dataset, continue-training GearNet with Mutli-view contrast makes the structure encoder overfitting to a small set of proteins, which could degrade model generalizability and result in deficient performances; and 2) ESM uses mask modeling (MK) in pre-training, and there can be a conflict between MK loss and Contrastive Learning (CL) loss in multi-task setting (Jiang et al., 2023).

Table 1: Combining Intra- and Inter-Modality Contrast: ① are reproduced uni-modal results; ② is directly fine-tuned two encoders together (thus multimodal) on downstream tasks, without any continuing training (GearNet was pre-trained on 1M AlphaFoldDB protein structures); whereas ③–⑤ are our multimodal models that continue training parallel ESM-GearNet on 55k tinyAlphaFoldDB proteins (paired sequences and structures) using intra- and inter-modality contrastive learning. Our models are named ESM-GearNet(*intra-contrast type*, *inter-contrast type*) where the contrast types include Mm, Mv, and Bio that stand for Multi-modal, Multi-view, and Bio-informed, respectively. Boldfaced in each column of task–metric combination is the best performance.

	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

We performed additional experiments to show the robustness of our proposed methods across the different random settings, as reflected in the low standard deviation values of model performances in Table 2. Due to the resource limits, we only fine-tuned the last model in Table 1, ESM-GearNet($\mathcal{L}_{\text{Bio}\circ\text{Mv}}$, $\mathcal{L}_{\text{Mm}}$), with 4 different random seeds, each for 30 epochs on the GO benchmarks and for 20 epochs on the EC benchmarks. Clearly, by composing bio-informed and multi-view intra-modality contrasts

and adding the inter-modality contrasts, our model improved their performances significantly and consistently in all tasks, compared to the starting point of no continual training.

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). The better performance in each column task is boldfaced.

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

Composing Augmentations: We further compose augmentations for inter-modality contrast. As shown in Table 3, fusing different augmentations can match or even surpass the naive cross-modal model in some of the benchmarks. Similar to earlier observations, when composed together, bio-informed views enhance the identity views in multi-modal CLIP-type learning and the cropped views in multi-view contrast, especially in GO-MF where individual proteins themselves are the determinants of molecular functions. Bio-informed views are based on protein homology classes (family and superfamily notions for sequence homology and structure similarity from a shared common evolutionary origin), which extends the semantic robustness compared to that of identity views used in CLIP. Meanwhile, composing such homology views with previously proposed multi-view cropping could further mitigate the latent feature suppression (Chen et al., 2021) issue between sequence and structure. We note that GO-BP and GO-CC can largely depend on the protein interaction partners as well, which may not manifest the improvement of representation learning for individual proteins.

Table 3: Downstream fine-tuned performances (Fmax (AUPR)) using composed inter-modal augmentations. The best performance in each task is boldfaced.

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)

Combining results from Tables 1, 2, and 3, we conclude that adding our newly designed bio-informed augmentations though view augmentation can enhance model performance in different training settings by aligning the inductive biases in models to domain knowledge.

5 Conclusion

In this work, we aim at multimodal representation learning for proteins (sequences and structures) while overcoming resource limitations and modality heterogeneity. To mitigate the computational cost, we start with separately pre-trained signal-modality encoders [including protein language models for sequences (ESM-1b) and geometric deep learning models for structures (GearNet)], integrate them in parallel, curate lean data of paired sequences and structures, and leverage multi-modal contrastive learning to continue-training the parallelly integrated single-modality encoders. To address a type of under-explored heterogeneity for protein modalities, namely the different levels of semantic robustness (unit perturbations of sequence and structure embeddings in the latent space lead to different levels of functional changes), we leverage domain-informed protein homology classes (families for similar sequences and superfamilies for similar structures) to design novel data views that can be combined together with previously proposed views for both 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.

ACKNOWLEDGMENTS

This project was in part supported by the National Science Foundation under grant CCF-1943008. Portions of this research were conducted with the advanced computing resources provided by Texas A&M High Performance Research Computing.

REFERENCES

Bruce Alberts. *Molecular biology of the cell*. Garland science, 2017.

Inigo Barrio-Hernandez, Jingi Yeo, Jürgen Jänes, Tanita Wein, Mihaly Varadi, Sameer Velankar, Pedro Beltrao, and Martin Steinegger. Clustering predicted structures at the scale of the known protein universe. *Nature*, 2023. doi: https://doi.org/10.1038/s41586-023-06510-w.

Nadav Brandes, Dan Ofer, Yam Peleg, Nadav Rappoport, and Michal Linial. Proteinbert: a universal deep-learning model of protein sequence and function. *Bioinformatics*, 38(8):2102–2110, 2022.

Tianlong Chen, Chengyue Gong, Daniel Jesus Diaz, Xuxi Chen, Jordan Tyler Wells, qiang liu, Zhangyang Wang, Andrew Ellington, Alex Dimakis, and Adam Klivans. HotProtein: A novel framework for protein thermostability prediction and editing. In *The Eleventh International Conference on Learning Representations*, 2023.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020.

Ting Chen, Calvin Luo, and Lala Li. Intriguing properties of contrastive losses. *Advances in Neural Information Processing Systems*, 34:11834–11845, 2021.

Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, et al. Prottrans: Toward understanding the language of life through self-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 44(10):7112–7127, 2021.

Ahmed Elnaggar, Hazem Essam, Wafaa Salah-Eldin, Walid Moustafa, Mohamed Elkerdawy, Charlotte Rochereau, and Burkhard Rost. Ankh: Optimized protein language model unlocks general-purpose modelling. biorxiv, 2023.

Pablo Gainza, Freyr Sverrisson, Frederico Monti, Emanuele Rodola, D Boscaini, MM Bronstein, and BE Correia. Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. *Nature Methods*, 17(2):184–192, 2020.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*, 2021.

Vladimir Gligorijević, P Douglas Renfrew, Tomasz Kosciolek, Julia Koehler Leman, Daniel Berenberg, Tommi Vatanen, Chris Chandler, Bryn C Taylor, Ian M Fisk, Hera Vlamakis, et al. Structure-based protein function prediction using graph convolutional networks. *Nature communications*, 12 (1):3168, 2021.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738, 2020.

Liang He, Shizhuo Zhang, Lijun Wu, Huanhuan Xia, Fusong Ju, He Zhang, Siyuan Liu, Yingce Xia, Jianwei Zhu, Pan Deng, et al. Pre-training co-evolutionary protein representation via a pairwise masked language model. *arXiv preprint arXiv:2110.15527*, 2021.

Pedro Hermosilla, Marco Schäfer, Matěj Lang, Gloria Fackelmann, Pere Pau Vázquez, Barbora Kozlíková, Michael Krone, Tobias Ritschel, and Timo Ropinski. Intrinsic-extrinsic convolution and pooling for learning on 3d protein structures. *arXiv preprint arXiv:2007.06252*, 2020.

Daniel Hesslow, Niccoló Zanichelli, Pascal Notin, Iacopo Poli, and Debora Marks. Rita: a study on scaling up generative protein sequence models, 2022. *URL https://arxiv. org/abs/2205.05789*.

Ziyu Jiang, Yinpeng Chen, Mengchen Liu, Dongdong Chen, Xiyang Dai, Lu Yuan, Zicheng Liu, and Zhangyang Wang. Layer grafted pre-training: Bridging contrastive learning and masked image modeling for label-efficient representations. In *The Eleventh International Conference on Learning Representations*, 2023.

Bowen Jing, Stephan Eismann, Patricia Suriana, Raphael JL Townshend, and Ron Dror. Learning from protein structure with geometric vector perceptrons. 2020, 2009.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Sal Candido, et al. Language models of protein sequences at the scale of evolution enable accurate structure prediction. *BioRxiv*, 2022:500902, 2022.

Amy X Lu, Haoran Zhang, Marzyeh Ghassemi, and Alan Moses. Self-supervised contrastive learning of protein representations by mutual information maximization. *BioRxiv*, pp. 2020–09, 2020.

Ali Madani, Ben Krause, Eric R Greene, Subu Subramanian, Benjamin P Mohr, James M Holton, Jose Luis Olmos Jr, Caiming Xiong, Zachary Z Sun, Richard Socher, et al. Large language models generate functional protein sequences across diverse families. *Nature Biotechnology*, pp. 1–8, 2023.

Pascal Notin, Mafalda Dias, Jonathan Frazer, Javier Marchena-Hurtado, Aidan Gomez, Debora S Marks, and Yarin Gal. Tranception: protein fitness prediction with autoregressive transformers and inference-time retrieval, 2022. *URL https://arxiv. org/abs/2205.13760*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Alexander Rives, Siddharth Goyal, Joshua Meier, Demi Guo, Myle Ott, C Lawrence Zitnick, Jerry Ma, and Rob Fergus. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. biorxiv, 2019.

Yuanfei Sun and Yang Shen. Structure-informed protein language models are robust predictors for variant effects. 2023.

Freyr Sverrisson, Jean Feydy, Bruno E Correia, and Michael M Bronstein. Fast end-to-end learning on protein surfaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15272–15281, 2021.

Arghamitra Talukder, Rujie Yin, Yuanfei Sun, Yang Shen, and Yuning You. Does inter-protein contact prediction benefit from multi-modal data and auxiliary tasks? *bioRxiv*, pp. 2022–11, 2022.

Mihaly Varadi, Stephen Anyango, Mandar Deshpande, Sreenath Nair, Cindy Natassia, Galabina Yordanova, David Yuan, Oana Stroe, Gemma Wood, Agata Laydon, et al. Alphafold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic acids research*, 50(D1):D439–D444, 2022.

Zichen Wang, Steven A Combs, Ryan Brand, Miguel Romero Calvo, Panpan Xu, George Price, Nataliya Golovach, Emmanuel O Salawu, Colby J Wise, Sri Priya Ponnapalli, et al. Lm-gvp: an extensible sequence and structure informed deep learning framework for protein property prediction. *Scientific reports*, 12(1):6832, 2022.

Tianxin Wei, Yuning You, Tianlong Chen, Yang Shen, Jingrui He, and Zhangyang Wang. Augmentations in hypergraph contrastive learning: Fabricated and generative. *Advances in neural information processing systems*, 35:1909–1922, 2022.

Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. Videoclip: Contrastive pre-training for zero-shot video-text understanding. *arXiv preprint arXiv:2109.14084*, 2021.

Minghao Xu, Yuanfan Guo, Yi Xu, Jian Tang, Xinlei Chen, and Yuandong Tian. Eurnet: Efficient multi-range relational modeling of spatial multi-relational data. *arXiv preprint arXiv:2211.12941*, 2022.

Minghao Xu, Xinyu Yuan, Santiago Miret, and Jian Tang. Protst: Multi-modality learning of protein sequences and biomedical texts. *arXiv preprint arXiv:2301.12040*, 2023.

Yuning You and Yang Shen. Cross-modality and self-supervised protein embedding for compound–protein affinity and contact prediction. *Bioinformatics*, 38(Supplement_2):ii68–ii74, 09 2022. ISSN 1367-4803. doi: 10.1093/bioinformatics/btac470.

Yuning You, Tianlong Chen, Yongduo Sui, Ting Chen, Zhangyang Wang, and Yang Shen. Graph contrastive learning with augmentations. *Advances in neural information processing systems*, 33: 5812–5823, 2020.

Yuning You, Tianlong Chen, Yang Shen, and Zhangyang Wang. Graph contrastive learning automated. In *International Conference on Machine Learning*, pp. 12121–12132. PMLR, 2021.

Yuning You, Tianlong Chen, Zhangyang Wang, and Yang Shen. Bringing your own view: Graph contrastive learning without prefabricated data augmentations. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, pp. 1300–1309, 2022.

Yuning You, Ruida Zhou, Jiwoong Park, Haotian Xu, Chao Tian, Zhangyang Wang, and Yang Shen. Latent 3d graph diffusion. In *The Twelfth International Conference on Learning Representations*, 2023.

Ningyu Zhang, Zhen Bi, Xiaozhuan Liang, Siyuan Cheng, Haosen Hong, Shumin Deng, Jiazhang Lian, Qiang Zhang, and Huajun Chen. Ontoprotein: Protein pretraining with gene ontology embedding. *arXiv preprint arXiv:2201.11147*, 2022.

Zuobai Zhang, Minghao Xu, Vijil Chenthamarakshan, Aurelie Lozano, Payel Das, and Jian Tang. Enhancing protein language models with structure-based encoder and pre-training. In *International Conference on Learning Representations Machine Learning for Drug Discovery Workshop*, 2023a.

Zuobai Zhang, Minghao Xu, Arian Jamasb, Vijil Chenthamarakshan, Aurelie Lozano, Payel Das, and Jian Tang. Protein representation learning by geometric structure pretraining. In *International Conference on Learning Representations*, 2023b.

Zhaocheng Zhu, Chence Shi, Zuobai Zhang, Shengchao Liu, Minghao Xu, Xinyu Yuan, Yangtian Zhang, Junkun Chen, Huiyu Cai, Jiarui Lu, et al. Torchdrug: A powerful and flexible machine learning platform for drug discovery. *arXiv preprint arXiv:2202.08320*, 2022.

A CONTRASTIVE LOSSES FOR COMPOSED MULTI-MODAL VIEWS

We have the following inter-modality losses by composing the identity in Eq. 2 with the cropping in Eq. 1 (multi-view), or the homology in Eq. 3 (bio-informed), or both:

$$\mathcal{L}_{\text{Multi-modal} \circ \text{Multi-view}} \tag{5}$$
$$= -\frac{1}{n}\sum_{i=j}\log\Big(\text{sim}((\mathcal{F}\circ g)_{\text{seq}}(\mathcal{H}_{\text{seq}}(x^i_{\text{seq}})), (\mathcal{F}\circ g)_{\text{struc}}(\mathcal{H}_{\text{struc}}(x^j_{\text{struc}}))\Big)$$
$$+\frac{\tau}{n}\sum_{i}\log\left(\sum_{j\neq i}\text{sim}((\mathcal{F}\circ g)_{\text{seq}}(\mathcal{H}_{\text{seq}}(x^i_{\text{seq}})), (\mathcal{F}\circ g)_{\text{struc}}(\mathcal{H}_{\text{struc}}(x^j_{\text{struc}}))\right)$$

$$\mathcal{L}_{\text{Multi-modal} \circ \text{Bio-informed}} \tag{6}$$
$$= -\frac{1}{n}\sum_{i}\log\left(\sum_{(x^m,x^n)\in\mathcal{C}_f(x^i)\times\mathcal{C}_s(x^i)}\text{sim}((\mathcal{F}\circ g)_{\text{seq}}(x^m_{\text{seq}}), (\mathcal{F}\circ g)_{\text{struc}}(x^n_{\text{struc}}))\right)$$
$$+\frac{\tau}{n}\sum_{i}\log\left(\sum_{(x^m,x^n)\notin\mathcal{C}_f(x^i)\times\mathcal{C}_s(x^i)}\text{sim}((\mathcal{F}\circ g)_{\text{seq}}(\mathcal{H}_{\text{seq}}x^m_{\text{seq}}), (\mathcal{F}\circ g)_{\text{struc}}(x^n_{\text{struc}}))\right),$$

$$\mathcal{L}_{\text{Multi-modal} \circ \text{Bio-informed} \circ \text{Multi-view}} \tag{7}$$
$$= -\frac{1}{n}\sum_{i}\log\left(\sum_{(x^m,x^n)\in\mathcal{C}_f(x^i)\times\mathcal{C}_s(x^i)}\text{sim}((\mathcal{F}\circ g)_{\text{seq}}(\mathcal{H}_{\text{seq}}(x^m_{\text{seq}})), (\mathcal{F}\circ g)_{\text{struc}}(\mathcal{H}_{\text{struc}}(x^n_{\text{struc}})))\right)$$
$$+\frac{\tau}{n}\sum_{i}\log\left(\sum_{(x^m,x^n)\notin\mathcal{C}_f(x^i)\times\mathcal{C}_s(x^i)}\text{sim}((\mathcal{F}\circ g)_{\text{seq}}(\mathcal{H}_{\text{seq}}(x^m_{\text{seq}})), (\mathcal{F}\circ g)_{\text{struc}}(\mathcal{H}_{\text{struc}}(x^n_{\text{struc}})))\right),$$

B DATA CURATION

Foldseek classifies the AlphaFold dataset into 2M clusters, each cluster is assigned with a representative protein and a set of member proteins. We take an intersection between the original 1M AlphaFoldDB dataset with foldseek clusters: if we have the representative protein, select the representation protein into our new dataset; otherwise, we check if we encounter members, and randomly choose one from our matched member proteins into our dataset. Note that we only select one protein per foldseek cluster. The procedure leads to a new dataset of 55,189 representative proteins, called tinyAlphaFoldDB. Table 4 provides the statistics of InterPro in our tinyAlphaFoldDB dataset: we are able to query 9,361 and 2,363 unique Family and SuperFamily classifications in our 55,189 instances; among those 55k proteins, 26,883 proteins have Family classifications and 30,068 proteins are assigned with SuperFamily classifications. When proteins do not associate with any Family or SuperFamily classification, proteins still enter the loss with identity positives.

Table 4: Statistics of TinyAlphaFoldDB data with homology classifications

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

C ORDERING IN SEQUENTIAL INTEGRATION

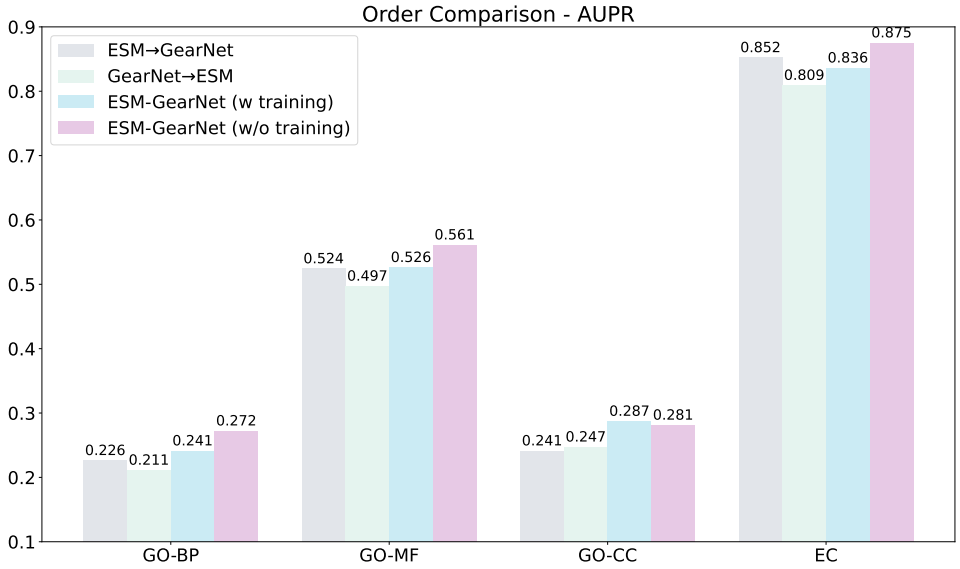

Figure 3: Comparing AUPR over the EC and GO benchmark sets between *Sequential Integration* with various orders (Zhang et al., 2023a) and *Parallel Integration.*