GRAF: a Geometry-Guided Adaptation Framework for Protein Representation in Isolated Low-Resource Tasks

1st Cheng Bi
College of Computer Science
Sichuan University
Chengdu, China
2024223040064@stu.scu.edu.cn

2nd Yi Feng
College of Computer Science
Sichuan University
Chengdu, China
2023223040046@stu.scu.edu.cn

3rd Mingyuan Yang
College of Computer Science
Sichuan University
Chengdu, China
yangly.scu@gmail.com

4th Han Wang
State Key Laboratory of Natural and Biomimetic Drugs
School of Pharmaceutical Sciences, Peking University
Beijing, China
wanghan0501@bjmu.edu.cn

5th Huawei Cai
West China Hospital
Sichuan University
Chengdu, China
huawei.cai@wchscu.edu.cn

6th Haixian Zhang*
College of Computer Science
Sichuan University
Chengdu, China
zhanghaixian@scu.edu.cn

***Abstract*—Effectively learning protein representations is crucial in isolated low-resource tasks, which are characterized by data scarcity and high sequence heterogeneity. In these scenarios, the prevailing paradigm of adapting broad representations of protein language models (PLMs) to specific biochemical contexts through fine-tuning or meta-learning is often constrained by limited sequence homology. To address these constraints, we propose GRAF, a geometry-guided representation adaptation framework that leverages intrinsic protein geometry to steer the adaptation process. We first establish the scale-aware geometric grounding module to restrict the search space within noisy representations. Subsequently, we devise the surface-guided geometric mapping module to decouple functional surface from global structural stability, allowing the extraction of conserved patterns without relying on external supervision. Extensive experiments demonstrate that GRAF achieves state-of-the-art performance on bacterial effector identification and antimicrobial peptide functional classification, validating its superiority in resolving the inherent challenges of biological isolation and data scarcity.**

***Index Terms*—Protein Representation Learning, Low-Resource, Geometric Deep Learning, Graph Neural Networks**

I. Introduction

Isolated low-resource tasks, characterized by data scarcity and high sequence heterogeneity, present a critical challenge in computational biology. Nevertheless, it is important to identify novel samples and enrich the known functional space within these limited existing data [1], [2]. Unlike housekeeping proteins that belong to dense and well-conserved evolutionary families [3], these specialized functional classes often arise through mechanisms such as horizontal gene transfer or convergent evolution [4], [5]. With the advancement of deep learning, the core of addressing this problem becomes how to effectively learn protein representations under such constrained conditions.

Exemplified by bacterial effector identification [1] and antimicrobial peptide (AMP) functional classification [2], these problems have traditionally been addressed by task-specific models. In the former, methodologies have progressed from system-specific architectures, such as T3SEpp [6], T4SEfinder [7] and T4SEpp [8], to unified frameworks like BastionX [9]. The latter has evolved from predictors based on pseudo-amino acid composition, represented by iAMP-CA2L [10], AMPDiscover [11] and iAMP-RAAC [12], to recent graph-based approaches including sAMPpred-GAT [13] and PGAT-ABPp [14]. However, these methods are tailored to specific biological contexts, largely relying on manually extracted physicochemical features or statistical sequence correlations.

The emergence of general-purpose protein language models (PLMs), such as ESM-2 [15], has prompted new approaches for isolated low-resource tasks, as demonstrated by recent works like DeepSecE [16] and CLEF [17]. While PLMs successfully encode evolutionary syntax through large-scale pre-training, the domain gap persists between general sequence representations and specific biological functions. To bridge this gap, the standard approach employs data-driven fine-tuning [18]. For instance, frameworks like GeSite [19] fine-tune ESM-2 to predict nucleic acid-binding sites, while BALM [20] applies efficient adaptation strategies to align pre-trained representations with drug-target binding affinity tasks. To handle cases of data insufficiency, other research investigates retrieval-augmented or meta-learning strategies. Strategies such as FSFP [21] employs meta-transfer learning to bridge the task gap via prior knowledge accumulation on Proteingym [22], whereas METALIC [23] leverages optimization-based meta-learning across diverse assays to extract transferable structural

*Corresponding author

patterns.

However, the success of these mainstream paradigms relies on sufficient data density and rich homologous information, resources that are scarce in isolated low-resource tasks. As noted by Wang et al. [24], traditional fine-tuning assumes sufficient label density to cover features. In our setting, where experimentally verified labels are limited, this data-driven approach tends to capture spurious sequence correlations rather than invariant biochemical mechanisms [25]. Conversely, retrieval-based and meta-learning strategies rely on dense homologues. However, for targets like bacterial effectors, significant sequence divergence and the scarcity of informative homologs restrict the utility of these methods [26]. Consequently, obtaining high-quality protein representations without relying on external resources becomes the core of our solution.

Therefore, we propose GRAF, a geometry-guided representation adaptation framework. Based on the fundamental sequence-structure-function principle, instead of forcing the model to find patterns in sparse labels, we exploit the intrinsic geometric properties of proteins to guide the adaptation process. We first employ the scale-aware geometric grounding (SAGG) module to capture fine-grained physicochemical micro-environments, establishing a stable structural foundation that restricts the search space to compensate for data scarcity. Subsequently, the surface-guided geometric mapping (SGGM) aligns these local cues with surface-exposed functional regions to decouple interaction semantics from global structural stability, effectively overcoming the limitation of low homology. By integrating these geometry-guided mechanisms, GRAF presents high-quality protein representations under isolated low-resource conditions and demonstrates effectiveness in downstream tasks.

The contributions of this study are summarized as follows:

- We propose GRAF, a generalizable geometry-guided framework designed to acquire high-quality protein representations specifically for isolated low-resource tasks characterized by data scarcity and high sequence heterogeneity.
- We introduce a geometry-guided strategy built upon general PLMs. By constructing hierarchical views to ground structural features and implementing surface-guided alignment to decouple functional semantics, GRAF effectively captures intrinsic biochemical patterns, yielding robust domain-adaptive representations.
- We evaluate the protein representations obtained by GRAF through extensive experiments in bacterial effector identification and antimicrobial peptide functional classification. With a specific focus on identifying 3 effectors and classifying 22 distinct AMP functional activities, GRAF achieves state-of-the-art performance, which validates its robustness.

The code is available at https://github.com/ChengBistu/GRAF.

II. RELATED WORK

Geometric Deep Learning (GDL). GDL is widely used in proteins to capture 3D spatial dependencies. For instance, Feng et al. modeled protein-ligand 3D geometric features via SE(3) equivariant geometric deep learning [27], and Wu et al. integrated pre-trained PLMs into GDL architectures [28]. Unlike these methods that treat geometry as static input, GRAF utilizes it as guidance to adapt the importance of residues with sequence representations.

III. METHODS

In this section, we introduce the details of GRAF. First, we present the overview and notations. Next, we describe the pre-training task design, including SAGG, SGGM and GCCL. The overall workflow of GRAF is described in Fig. 1.

A. Overview and Notations

Pre-training for Domain Alignment. Let $\mathcal{D}_{pre} = \{P_i\}_{i=1}^{M}$ denote the unlabeled dataset used for domain adaptation, where M is the number of proteins. For each protein P_i, we define three input modalities:

- **Sequence:** Let $S_i = \{r_1, r_2, \ldots, r_L\}$ be the amino acid sequence of length L, which serves as the input for the PLM.
- **Structure:** Let $X_i \in \mathbb{R}^{L\times a\times 3}$ represent the 3D coordinates of the residues and their constituent atoms (where a is the number of atoms per residue), serving as the structural basis for geometric guidance.
- **Biological Priors:** Let $\mathcal{B} = \{Attr_1, \ldots, Attr_m\}$ be a set of learnable tokens representing orthogonal biological attributes, such as functional motifs or physicochemical properties, extracted from the protein to guide the geometric mapping.

The pre-training objective is to learn a geometry-aware adapter space where the structure and biological priors are fused into $Z_{Geo} \in \mathbb{R}^d$ by SAGG and SGGM, aligning with H_{seq} by GCCL.

Downstream Tasks. For downstream tasks, we utilize a labeled dataset $\mathcal{D}_{task} = \{(S_j, y_j)\}_{j=1}^{N}$, where S_j serves as the input protein sequence for the PLM and y_j is the corresponding functional label. We freeze parameters of both the ESM-2 and the aligned 2-layer Transformer sequence encoder, training only a lightweight MLP head to predict $\hat{y}_j$.

B. Scale-Aware Geometric Grounding (SAGG)

By incorporating fine-grained atomic microenvironments into the residue-level topology, SAGG establishes a stable structural foundation to restrict the search space within noisy sequence representations.

Graph Construction. Based on the hierarchical nature of proteins, we construct a dual-level graph representation. The atom graph $\mathcal{G}_{atom} = (\mathcal{V}_{atom}, \mathcal{E}_{atom})$ captures local chemical micro-environments, where each node $a_i \in \mathcal{V}_{atom}$ represents a heavy atom initialized with one-hot element encodings and backbone flags. Nodes are connected to their K (e.g. $K = 24$) nearest neighbors, and a 3-layer GAT aggregates these features

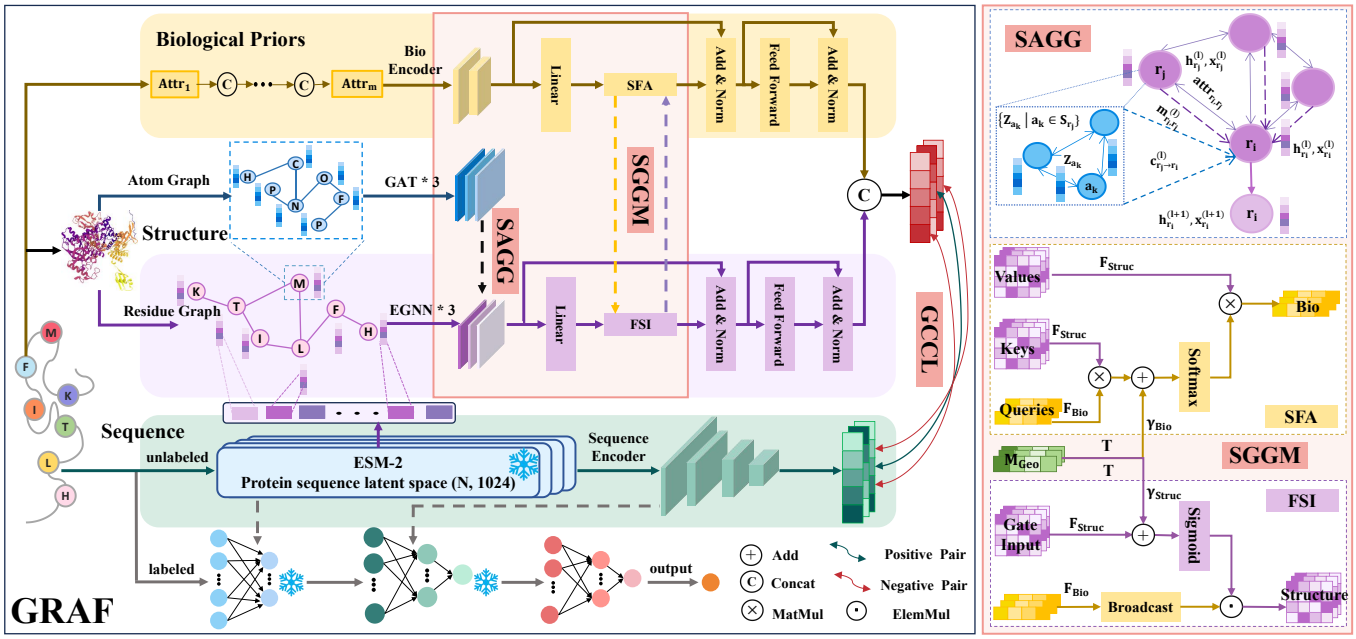

Fig. 1. The overall architecture of GRAF. (A) Sequence features are extracted via ESM-2, while structural representations are encoded by the SAGG module using coupled atom-residue graphs. (B) The SGGM module fuses biological priors with structural contexts to generate unified geometry-guided representations. (C) The GCCL module aligns the sequence space with the geometry-guided representations via circular contrastive learning, enabling efficient downstream tasks with a lightweight head.

to produce static atomic representations $Z_{atom} = \{h_{a_i}^{(3)}\}$. Simultaneously, we define a residue graph $\mathcal{G}_{res} = (\mathcal{V}_{res}, \mathcal{E}_{res})$, where each node $r_i \in \mathcal{V}_{res}$ is initialized with ESM-2 embeddings to preserve evolutionary context. The residue topology $\mathcal{E}_{res}$ is established using a 14 Å Euclidean distance threshold [19] between C_α atoms, with these distances assigned as edge weights.

Scale-Aware Message Passing. We employ an E(n) Equivariant Graph Neural Network (EGNN) [29] consisting of 3 stacked Equivariant Graph Convolutional Layers (EGCLs) to learn structural representations while maintaining geometric symmetry. To compensate for the inevitable information loss associated with coarse-grained residue representations, we introduce a cross-scale injection strategy. Specifically, we utilize the static atomic representations Z_{atom} processed by the GAT as invariant fine-grained priors. In each EGCL, these static priors are re-injected to calibrate the dynamic residue updates, ensuring that the learning of global topology remains enriched with local physicochemical details. As formulated in (1) and (2), taking a center residue r_i and its neighbor r_j at the l-th layer as an example, let S_{r_j} denote the set of atoms belonging to r_j and Z_{a_k} be the static feature of atom $a_k \in S_{r_j}$. We aggregate the fine-grained atomic context from neighbor r_j to center r_i via a cross-scale attention mechanism, where the sequence-aware residue feature $h_{r_i}^{(l)}$ queries the atomic features $Z_{S_{r_j}}$ that serve as both keys and values:

$$c_{r_j \to r_i}^{(l)} = \sum_{a_k \in S_{r_j}} \text{softmax}\left(\frac{(h_{r_i}^{(l)} W_Q)(Z_{a_k} W_K)^T}{\sqrt{d_k}}\right)(Z_{a_k} W_V), \tag{1}$$

where $c_{r_j \to r_i}^{(l)}$ represents the aggregated atomic context vector flowing from neighbor j to i at layer l. This vector is then injected into the edge operation ϕ_e of the EGCL [29]:

$$m_{r_i,r_j}^{(l)} = \phi_e\left(h_{r_i}^{(l)}, h_{r_j}^{(l)}, \|x_{r_i}^{(l)} - x_{r_j}^{(l)}\|^2, \text{attr}_{r_i,r_j}, c_{r_j \to r_i}^{(l)}\right). \tag{2}$$

Finally, the coordinates and features of the center residue are updated based on the messages $m_{r_i,r_j}^{(l)}$:

$$x_{r_i}^{(l+1)} = x_{r_i}^{(l)} + C \sum_{r_j \in \mathcal{N}(r_i)} (x_{r_i}^{(l)} - x_{r_j}^{(l)}) \phi_x(m_{r_i,r_j}^{(l)}), \tag{3}$$

$$h_{r_i}^{(l+1)} = \phi_h\left(h_{r_i}^{(l)}, \sum_{r_j \in \mathcal{N}(r_i)} m_{r_i,r_j}^{(l)}\right), \tag{4}$$

where ϕ_x and ϕ_h are Multi-Layer Perceptrons (MLPs), and $C = 1/(N-1)$ is a normalization factor. After 3 layers of iteration, the final residue representation $H_{struc} = \{h_{r_i}^{(3)}\}$ encodes both the global topological structure and the precise local atomic constraints.

C. Surface-Guided Geometric Mapping (SGGM)

As shown in the right panel of Fig. 1, this phase aligns structural features with biological priors through a surface-biased cross-modal attention mechanism. We define a geometric exposure vector $M_{Geo} \in \mathbb{R}^{1\times L}$ by calculating the normalized Euclidean distance from each residue to the protein centroid. Biological attributes are encoded into $H_{bio} \in \mathbb{R}^{1\times d}$ via a 2-layer MLP before interacting with the structural representations H_{struc}. The SGGM module leverages M_{Geo} as a spatial inductive bias to prioritize solvent-accessible regions, thereby

decoupling interaction-prone functional semantics from the stable protein core.

Structure to Function Aggregation (SFA). To integrate structural details into biological priors, we employ a geometry-biased attention mechanism. Unlike standard attention, we enforce a spatial inductive bias where the contribution of each residue is strictly weighted by its surface accessibility. As formulated in (5), the attention score α_{agg} is adjusted by the geometric constraint M_{Geo}:

$$\alpha_{agg} = \text{Softmax}(\frac{(H_{bio}W_Q)(H_{struc}W_K)^T}{\sqrt{d}} + \gamma_1 \cdot M_{Geo}), \quad (5)$$

where γ_1 is a learnable scalar controlling the geometric penalty. Consequently, the aggregated representation H'_{bio} is updated via (6):

$$H'_{bio} = H_{bio} + \alpha_{agg}(H_{struc}W_V), \quad (6)$$

This formulation forces the model to attend primarily to interaction-competent regions, filtering out noise from the buried core.

Function to Structure Injection (FSI). Conversely, to inject biological priors into specific structural sites, we design a gating scheme conditioned on both semantic compatibility and geometric feasibility. We compute a semantic gate G_{sem} derived from structural features and a geometric gate G_{geo} derived from physical location as shown in (7):

$$G_{sem} = H_{struc}W_{gate}, \quad G_{geo} = \gamma_2 \cdot M_{Geo}^T, \quad (7)$$

where W_{gate} is a learnable projection. In (8), the injection is modulated by a Sigmoid-activated summation of these gates:

$$H'_{struc} = H_{struc} + \sigma(G_{sem} + G_{geo}) \odot \text{Broadcast}(H_{bio}). \quad (8)$$

Here, the use of sigmoid (σ) enables independent feature selection, allowing the simultaneous activation of multiple surface-exposed functional residues without the signal dilution inherent in softmax normalization.

D. Global Feature Alignment via Circular Contrastive Learning (GCCL)

GCCL projects the geometry-guided representations from SGGM into the sequence space to calibrate the feature space. This module resolves the twin challenges of data imbalance and memory overflow by utilizing a circular mechanism to maximize the utility of limited positive samples. Specifically, sequence features H_{seq} and fused geometric representations $Z_{Geo} = \text{Concat}(H'_{bio}, H'_{struc})$ from the same protein form positive pairs, while those from different proteins serve as negatives.

Decoupled Sigmoid Loss. Instead of the standard softmax-based InfoNCE loss which induces gradient coupling, we adopt the pairwise sigmoid loss inspired by SigLIP [30]. As formulated in (9), this objective treats alignment as independent binary classification problems:

$$\mathcal{L} = -\frac{1}{B}\sum_{i=1}^{B}\sum_{j=1}^{B}\log\sigma\left(\eta_{ij}\cdot\text{sim}(H_{seq}^{(i)}, Z_{Geo}^{(j)}) - \beta\right), \quad (9)$$

where i and j index the sequence and fused geometric representations respectively. $\eta_{ij} \in \{1, -1\}$ serves as the label indicator for positive or negative pairs, and β is a learnable threshold bias introduced to stabilize training initialization dynamics. By decoupling the gradients of positive pairs from the dominant negative samples, this formulation effectively suppresses noise in imbalanced biological datasets.

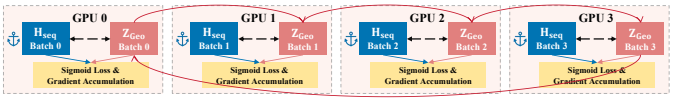

Fig. 2. Illustration of the Circular Contrastive Learning mechanism, where geometric representations are rotated across the GPU ring to expand negative sample coverage without increasing local memory overhead.

Memory-Efficient Ring Negative Expansion. Standard contrastive learning requires large batch sizes to cover sufficient negative samples, which causes memory overflow with heavy geometric graph encoders. To resolve this trade-off, we implement a gradient-free circular rotation mechanism. As illustrated in Fig. 2, while sequence features H_{seq} remain on local GPUs, the fused geometric features Z_{Geo} are passed across the GPU ring $N-1$ times. At each step, the model computes the loss between the local sequence batch and the visiting geometric batch, accumulating the gradients before passing the features to the next neighbor. This strategy expands the effective negative sample size to $N \times B$ without increasing the per-GPU memory complexity $O(B)$.

IV. EXPERIMENTS

A. Datasets

To evaluate GRAF in isolated low-resource tasks, we conducted experiments on two biological tasks, using predicted structures generated by ESMFold [15].

Bacterial Effector Identification. Following Peng et al. [17], the pre-training dataset comprises 10,831 unlabeled protein sequences, gathering effectors from existing predictors and non-effector proteins from UniProt and VFDB. The downstream phase employs the labeled dataset curated by Zhang et al. [16], consisting of 2,918 training proteins (1,341 secreted, 1,577 non-secreted) and an independent test set of 260 samples. For biological priors, we utilized DPC-PSSM for local evolutionary context and annotation texts for functional semantics.

Antimicrobial Peptide (AMP) Functional Classification. We utilized the datasets curated by Xu et al. [31], aggregating 40,056 sequences (6,405 AMPs and 33,651 non-AMPs) for pre-training and employing the functional activity dataset, approximately 10,000 for training and 5,000 for test, for downstream classifications. The biological priors encoded Amino Acid Composition (AAC), hydrophobic moment, and global physicochemical properties.

B. Implementation Details

We utilized the pre-trained ESM-2-650M model with an initial embedding dimension of 1,280 as the sequence backbone. For the pre-training phase, we employed the AdamW

optimizer with a learning rate of 2×10^{-5} and a weight decay of 1×10^{-2}. To stabilize the contrastive alignment, the model was trained for 50 epochs with a batch size of 48 per GPU, utilizing gradient accumulation to maximize the effective negative sample coverage.

C. Comparison with State-of-the-Art Methods

As shown in TABLE I, GRAF achieves superior performance compared to previous SOTA methods designed for bacterial effector identification. For instance, GRAF outperforms the runner-up models with gains of 0.39%, 2.00%, and 1.82% in ACC, MCC, and F1 in T4SE identification. Regarding AMP functional classification, TABLE II reports results for broad-spectrum antimicrobial activities. Taking the antifungal task as an example, GRAF achieves improvements of 2.80%, 7.89%, and 4.05% in ACC, MCC, and F1 over the SOTA method iAMPCN. TABLE III focuses on specialized therapeutic and toxicity activities. In the anticancer task, our method surpasses iAMPCN with gains of 2.71%, 8.77%, and 2.27% in ACC, MCC, and F1, respectively.

D. Case Study

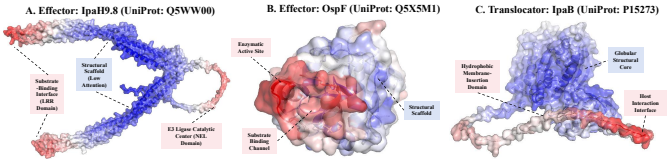

Fig. 3. Visualization of attention weights learned by the SGGM module on three representative bacterial effectors: (A) IpaH9.8, (B) OspF, (C) IpaB. The heatmaps visualize the interpretability of GRAF, where red denotes high attention concentrations on surface-exposed functional motifs such as binding interfaces, enzymatic grooves, and membrane-insertion arms, while blue indicates low attention on conserved structural backbones.

To validate the physical interpretability of GRAF and confirm that it captures genuine biochemical semantics rather than spurious correlations, we conducted a qualitative analysis by visualizing the attention weights from the SGGM module. Specifically, we extracted the attention maps during the inference phase for three distinct effectors and projected them onto their 3D structures. As illustrated in Fig. 3, the model exhibits a strong topological bias towards surface-exposed regions. The high-attention areas, highlighted in red, precisely align with the ground-truth functional sites, including the substrate-binding LRR tips of IpaH9.8, the docking groove of the phosphothreonine lyase OspF, and the hydrophobic membrane-insertion domain of the translocator IpaB. Conversely, the internal α-helical scaffolds responsible for thermodynamic stability are consistently down-weighted, as shown in blue. This demonstrates that the model adaptively decouples biochemical functions from structural stability features via geometric guidance.

E. Analysis of Limitations in Mainstream Paradigms

We select bacterial effector identification to investigate the specific constraints of gradient-based fine-tuning and retrieval-based meta-learning in isolated low-resource tasks.

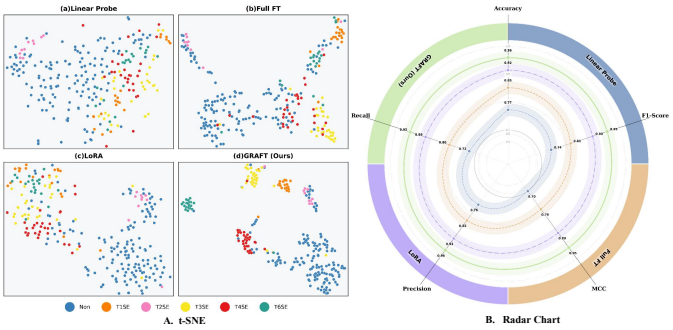

Fig. 4. A: t-SNE visualization of feature embeddings. B: Radar chart

Analysis of Fine-Tuning Overfitting. As illustrated in Fig. 4, we compared GRAF against three baselines: Linear Probe (frozen ESM-2), Full FT (full fine-tuning), and LoRA, analyzing their feature distributions via t-SNE and quantitative performance via a radar chart. The Linear Probe yields the lowest performance, with an ACC of 0.7725 and MCC of 0.6983, indicating that general PLM representations are lacking in the specific functional semantics required for this task. Notably, Full FT underperforms LoRA across all metrics. This inversion confirms our hypothesis that in data-scarce settings with biological heterogeneity, updating all parameters leads to overfitting on spurious sequence correlations rather than learning invariant mechanisms. While LoRA mitigates this by restricting the trainable parameter space, it remains a data-driven optimization process limited by the available labels. In contrast, GRAF achieves the best performance, recording an ACC of 0.9596 and MCC of 0.9504, significantly surpassing both Full FT and LoRA. The t-SNE visualizations further corroborate this: while baseline methods exhibit blurred class boundaries and fragmented clusters due to the sparse feature distribution, GRAF produces distinct, compact clusters for different effector classes. This demonstrates that by injecting internal geometry-guided constraints rather than relying solely on gradient-based updates, our method successfully acquires superior protein representations under isolated low-resource conditions.

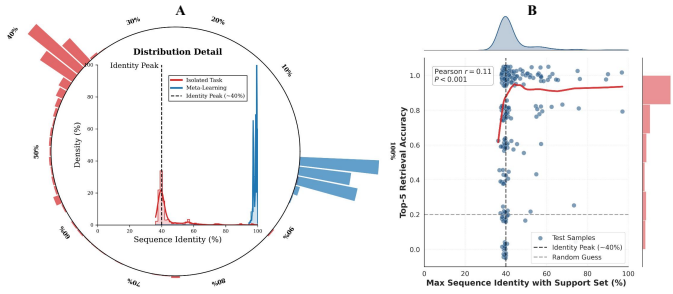

Fig. 5. Analysis of data distribution and retrieval reliability. (A) Comparison of sequence homology distributions between Proteingym and effector dataset. (B) Correlation analysis between retrieval accuracy and sequence identity.

Analysis of Retrieval Reliability. To validate the limita-

TABLE I
COMPARISON RESULTS OF EFFECTOR IDENTIFICATION ON THE INDEPENDENT TEST SET

Metrics	T3SE				T4SE					T6SE			
	GRAF (Ours)	CLEF [17]	DeepSecE [16]	T3SEpp [6]	**GRAF (Ours)**	CLEF [17]	T4SEpp [8]	DeepSecE [16]	T4SEfinder [7]	**GRAF (Ours)**	CLEF [17]	DeepSecE [16]	BastionX [9]
ACC ↑	**0.9923**	0.9885	0.9654	0.9538	**0.9962**	0.9923	0.9654	0.9770	0.8460	0.9962	**1.0000**	0.9810	0.6769
F1 ↑	**0.9667**	0.9508	0.8615	0.7778	**0.9831**	0.9655	0.8696	0.9063	0.5833	0.9744	**1.0000**	0.8718	0.3115
MCC ↑	**0.9623**	0.9444	0.8452	0.7582	**0.9811**	0.9619	0.8597	0.8954	0.5639	0.9727	**1.0000**	0.8617	0.3297

TABLE II
COMPARISON RESULTS ON BROAD-SPECTRUM ANTIMICROBIAL ACTIVITIES

Metrics	Antibacterial				Antifungal				Antiviral			
	GRAF (Ours)	iAMPCN [31]	iAMP-CA2L [10]	AMPDiscover-RNN [11]	**GRAF (Ours)**	iAMPCN [31]	iAMP-RAAC [12]	iAMP-CA2L [10]	**GRAF (Ours)**	iAMPCN [31]	iAMP-RAAC [12]	iAMP-CA2L [10]
ACC ↑	**0.7125**	0.7016	0.4987	0.6611	**0.6903**	0.6715	0.5982	0.5000	**0.7964**	0.7878	0.6249	0.4980
F1 ↑	**0.7308**	0.7262	0.5766	0.7203	**0.6455**	0.6204	0.6273	0.2998	**0.7823**	0.7750	0.6395	0.0461
MCC ↑	**0.4286**	0.4099	-0.0028	0.3557	**0.3842**	0.3561	0.1989	0.0000	**0.5951**	0.5794	0.2506	-0.0126

TABLE III
COMPARISON RESULTS ON SPECIALIZED THERAPEUTIC AND TOXICITY ACTIVITIES

Metrics	Anticancer				Anti-HIV			Anti-mammalian cells			
	GRAF (Ours)	iAMPCN [31]	iAMP-RAAC [12]	dbAMP [32]	**GRAF (Ours)**	iAMPCN [31]	iAMP-CA2L [10]	**GRAF (Ours)**	iAMPCN [31]	iAMP-RAAC [12]	dbAMP [32]
ACC ↑	**0.7245**	0.7054	0.4914	0.4774	**0.8115**	0.8050	0.4969	**0.7205**	0.7058	0.5379	0.5451
F1 ↑	**0.7388**	0.7224	0.1873	0.5483	0.7905	**0.7919**	0.0123	**0.7355**	0.7297	0.2768	0.5878
MCC ↑	**0.4502**	0.4139	-0.0259	-0.0476	**0.6233**	0.6149	-0.0325	**0.4420**	0.4181	0.1096	0.0923

tions of retrieval-based strategies such as meta-learning in isolated tasks, we compared the sequence homology distribution of the bacterial effector dataset against Proteingym [22], a widely utilized benchmark for protein fitness prediction. As shown in the density plot in Fig.5(A), Proteingym samples represented in blue predominantly share over 90% sequence identity with their support sets, a condition that facilitates reliable reference retrieval. In contrast, the effector dataset shown in red exhibits a severe distribution shift towards the isolated zone, centering around a low homology threshold of 40%. We further evaluated the impact of this scarcity on retrieval reliability by calculating the Top-5 label consistency of neighbors retrieved via the ESM-2 baseline as illustrated in Fig.5(B). In the low-identity region below 40%, the retrieval performance exhibits high instability, with a dense cluster of samples falling near or below the random guess baseline. This empirical evidence confirms that in isolated low-resource tasks, the insufficiency of close evolutionary neighbors renders distance-based retrieval unreliable, as the model tends to capture spurious sequence similarities rather than conserved functional signals.

TABLE IV
ABLATION STUDIES ON EFFECTOR IDENTIFICATION TEST SET

Pre-training Model	SAGG	SGGM	GCCL	ACC ↑	F1 ↑	MCC ↑
Linear Probe	×	×	×	0.7725	0.7418	0.6983
w/o Circular Loss	✓	✓	×	0.8146	0.7932	0.7655
ESM-2 + Adapter	×	×	×	0.8540	0.8328	0.8153
w/o Geometric Gate	✓	×	✓	0.9168	0.9056	0.8972
Residue only	×	✓	✓	0.9482	0.9415	0.9380
GRAF (Ours)	✓	✓	✓	**0.9596**	**0.9531**	**0.9504**

F. Ablation Study

As shown in TABLE IV, we conduct ablation studies on the bacterial effector identification task to evaluate the contribution of each component within the GRAF. Specifically, the removal of the SAGG module leads to a decline in MCC from 0.9504 to 0.9380, indicating that fine-grained atomic information provides more detailed micro-environmental constraints than coarse residue-level graphs. Furthermore, replacing the SGGM module with standard cross-attention results in a performance gap of 4.67% in ACC compared to the full model. This confirms that the hard geometric constraint of residue-to-surface proximity is superior to generic attention, effectively decoupling functional surface semantics from buried core stability features. Regarding optimization, the GCCL module via circular loss outperforms standard contrastive learning; without it, memory overhead from multi-modal features constrains local batch sizes, degrading performance below even the standard adapter. Finally, when removing all geometric components and directly inputting the protein representations derived from the PLM into a standard adapter or a simple MLP, the performance degrades, with ACC dropping to 0.8540 and 0.7725, respectively. Consequently, the naive utilization of representations without specific geometry-guided adaptation proves insufficient for isolated low-resource tasks.

V. CONCLUSIONS

In this paper, we present GRAF, a geometry-guided representation adaptation framework effectively addressing the inherent challenges of isolated low-resource tasks. By implementing the SAGG and the SGGM module, our method compensates for the limited homologous information in sparse datasets. Extensive experiments on bacterial effector identification and antimicrobial peptide functional classification confirm the superiority of GRAF over state-of-the-art baselines. Notably, comparative experiments against standard fine-tuning, combined with the analysis of retrieval reliability, verify GRAF's specific suitability for tasks characterized by

data scarcity and high sequence heterogeneity. Furthermore, interpretability study reveals that our model successfully identifies genuine surface-exposed interaction motifs rather than spurious sequence correlations, ensuring robust generalization. In the future, we plan to extend GRAF to more complex interaction tasks, such as predicting protein-protein interaction interfaces or guiding the de novo design of functional peptides, particularly in data-scarce environments.

ACKNOWLEDGMENT

We thank for the support of National Natural Science Foundation of China under Grant 62476183 and Natural Science Foundation from the Science & Technology Department of Sichuan Province under Grant 2024NSFTD0051.

REFERENCES

[1] Z. Zhao, Y. Hu, Y. Hu, A. P. White, and Y. Wang, "Features and algorithms: facilitating investigation of secreted effectors in Gram-negative bacteria," *Trends in Microbiology*, vol. 31, no. 11, pp. 1162–1178, 2023.

[2] N. G. Oliveira Júnior, C. M. Souza, D. F. Buccini, M. H. Cardoso, and O. L. Franco, "Antimicrobial peptides: structure, functions and translational applications," *Nature Reviews Microbiology*, vol. 23, no. 11, pp. 687–700, 2025.

[3] R. E. Ferguson, H. P. Carroll, A. Harris, E. R. Maher, P. J. Selby, and R. E. Banks, "Housekeeping proteins: a preliminary study illustrating some limitations as useful references in protein expression studies," *Proteomics*, vol. 5, no. 2, pp. 566–571, 2005.

[4] B. Segerman, "The genetic integrity of bacterial species: the core genome and the accessory genome, two different stories," *Frontiers in cellular and infection microbiology*, vol. 2, p. 116, 2012.

[5] J. Thomas, S. S. Watve, W. C. Ratcliff, and B. K. Hammer, "Horizontal gene transfer of functional type VI killing genes by natural transformation," *MBio*, vol. 8, no. 4, pp. 10–1128, 2017.

[6] X. Hui, Z. Chen, M. Lin, J. Zhang, Y. Hu, Y. Zeng, X. Cheng, L. Ou-Yang, M.-a. Sun, A. P. White, et al., "T3SEpp: an integrated prediction pipeline for bacterial type III secreted effectors," *Msystems*, vol. 5, no. 4, pp. 10–1128, 2020.

[7] Y. Zhang, Y. Zhang, Y. Xiong, H. Wang, Z. Deng, J. Song, and H.-Y. Ou, "T4SEfinder: a bioinformatics tool for genome-scale prediction of bacterial type IV secreted effectors using pre-trained protein language model," *Briefings in Bioinformatics*, vol. 23, no. 1, 2022.

[8] Y. Hu, Y. Wang, X. Hu, H. Chao, S. Li, Q. Ni, Y. Zhu, Y. Hu, Z. Zhao, and M. Chen, "T4SEpp: A pipeline integrating protein language models to predict bacterial type IV secreted effectors," *Computational and Structural Biotechnology Journal*, vol. 23, pp. 801–812, 2024.

[9] J. Wang, J. Li, Y. Hou, W. Dai, R. Xie, T. T. Marquez-Lago, A. Leier, T. Zhou, V. Torres, I. Hay, et al., "BastionHub: a universal platform for integrating and analyzing substrates secreted by Gram-negative bacteria," *Nucleic Acids Research*, vol. 49, no. D1, pp. D651–D659, 2021.

[10] X. Xiao, Y.-T. Shao, X. Cheng, and B. Stamatovic, "iAMP-CA2L: a new CNN-BiLSTM-SVM classifier based on cellular automata image for identifying antimicrobial peptides and their functional types," *Briefings in Bioinformatics*, vol. 22, no. 6, p. bbab209, 2021.

[11] S. A. Pinacho-Castellanos, C. R. García-Jacas, M. K. Gilson, and C. A. Brizuela, "Alignment-free antimicrobial peptide predictors: improving performance by a thorough analysis of the largest available data set," *Journal of Chemical Information and Modeling*, vol. 61, no. 6, pp. 3141–3157, 2021.

[12] G.-F. Dong, L. Zheng, S.-H. Huang, J. Gao, and Y.-C. Zuo, "Amino acid reduction can help to improve the identification of antimicrobial peptides and their functional activities," *Frontiers in Genetics*, vol. 12, p. 669328, 2021.

[13] K. Yan, H. Lv, Y. Guo, W. Peng, and B. Liu, "sAMPpred-GAT: prediction of antimicrobial peptide by graph attention network and predicted peptide structure," *Bioinformatics*, vol. 39, no. 1, p. btac715, 2023.

[14] Y. Hao, X. Liu, H. Fu, X. Shao, and W. Cai, "PGAT-ABPp: harnessing protein language models and graph attention networks for antibacterial peptide identification with remarkable accuracy," *Bioinformatics*, vol. 40, no. 8, p. btae497, 2024.

[15] Z. Lin, H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, N. Smetanin, R. Verkuil, O. Kabeli, Y. Shmueli, et al., "Evolutionary-scale prediction of atomic-level protein structure with a language model," *Science*, vol. 379, no. 6637, pp. 1123–1130, 2023.

[16] Y. Zhang, J. Guan, C. Li, Z. Wang, Z. Deng, R. B. Gasser, J. Song, and H.-Y. Ou, "DeepSecE: a deep-learning-based Framework for multiclass Prediction of secreted Proteins in gram-negative bacteria," *Research*, vol. 6, p. 0258, 2023.

[17] Y. Peng, J. Wu, Y. Sun, Y. Zhang, Y. Wang, and S. Shao, "Contrastive-learning of language embedding and biological features for cross modality encoding and effector prediction," *Nature Communications*, vol. 16, no. 1, p. 1299, 2025.

[18] H. Zheng, L. Shen, A. Tang, Y. Luo, H. Hu, B. Du, Y. Wen, and D. Tao, "Learning from models beyond fine-tuning," *Nature Machine Intelligence*, vol. 7, no. 1, pp. 6–17, 2025.

[19] W. Zeng, L. Pan, B. Ji, L. Xu, and S. Peng, "Accurate Nucleic Acid-Binding Residue Identification based Domain-Adaptive Protein Language Model and Explainable Geometric Deep Learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 1, 2025, pp. 1004–1012.

[20] R. Gorantla, A. P. Gema, I. X. Yang, Á. Serrano-Morrás, B. Suutari, J. Juárez-Jiménez, and A. S. J. S. Mey, "Learning binding affinities via fine-tuning of protein and ligand language models," *Journal of Chemical Information and Modeling*, vol. 65, no. 22, pp. 12279–12291, 2025.

[21] Z. Zhou, L. Zhang, Y. Yu, B. Wu, M. Li, L. Hong, and P. Tan, "Enhancing efficiency of protein language models with minimal wet-lab data through few-shot learning," *Nature Communications*, vol. 15, no. 1, p. 5566, 2024.

[22] P. Notin, A. Kollasch, D. Ritter, L. Van Niekerk, S. Paul, H. Spinner, N. Rollins, A. Shaw, R. Orenbuch, R. Weitzman, et al., "Proteingym: Large-scale benchmarks for protein fitness prediction and design," *Advances in Neural Information Processing Systems*, vol. 36, pp. 64331–64379, 2023.

[23] J. Beck, S. Surana, M. McAuliffe, O. Bent, T. D. Barrett, J. J. G. Luis, and P. Duckworth, "Metalic: Meta-learning in-context with protein language models," 2024, arXiv:2410.08355. [Online]. Available: https://arxiv.org/abs/2410.08355

[24] L. Wang, Q. Liu, S. Liu, X. Sun, and S. Wu, "Pin-tuning: Parameter-efficient in-context tuning for few-shot molecular property prediction," *Advances in Neural Information Processing Systems*, vol. 37, pp. 70765–70792, 2024.

[25] M. H. Çelik and X. Xie, "Efficient inference, training, and fine-tuning of protein language models," *Iscience*, vol. 28, no. 10, 2025.

[26] N. Wagner, O. Avram, D. Gold-Binshtok, B. Zerah, D. Teper, and T. Pupko, "Effectidor: an automated machine-learning-based web server for the prediction of type-III secretion system effectors," *Bioinformatics*, vol. 38, no. 8, pp. 2341–2343, 2022.

[27] Y. Feng, C. Bi, M. Tian, H. Wang, and H. Zhang, "An SE(3) Equivariant Unified Multi-Modal Framework for Protein-Ligand Interaction Prediction," in *2025 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2025, pp. 7243–7250..

[28] F. Wu, L. Wu, D. Radev, J. Xu, and S. Z. Li, "Integration of pre-trained protein language models into geometric deep learning networks," *Communications Biology*, vol. 6, no. 1, p. 876, 2023.

[29] V. G. Satorras, E. Hoogeboom, and M. Welling, "E(n) equivariant graph neural networks," in *International conference on machine learning*, 2021, pp. 9323–9332.

[30] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer, "Sigmoid loss for language image pre-training," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 11975–11986.

[31] J. Xu, F. Li, C. Li, X. Guo, C. Landersdorfer, H.-H. Shen, A. Y. Peleg, J. Li, S. Imoto, J. Yao, et al., "iAMPCN: a deep-learning approach for identifying antimicrobial peptides and their functional activities," *Briefings in Bioinformatics*, vol. 24, no. 4, p. bbad240, 2023.

[32] J.-H. Jhong, L. Yao, Y. Pang, Z. Li, C.-R. Chung, R. Wang, S. Li, W. Li, M. Luo, R. Ma, et al., "dbAMP 2.0: updated resource for antimicrobial peptides with an enhanced scanning method for genomic and proteomic data," *Nucleic acids research*, vol. 50, no. D1, pp. D460–D470, 2022.