PLI-CLIP: LEARNING TRANSFERABLE PROTEIN–LIGAND INTERACTION REPRESENTATIONS FROM NATURAL LANGUAGE SUPERVISION

Anonymous authors
Paper under double-blind review

ABSTRACT

Predicting the 3D structure of protein–ligand complexes is central to structure-based drug discovery, and recent approaches employ protein language models (PLMs) to address this challenge. While PLMs provide powerful representations of proteins, they are inherently ligand-agnostic and do not capture explicit protein–ligand interaction semantics. To address this shortcoming, we exploit rich, text-like biochemical annotations produced by tools such as PLIP Salentin et al. (2015) to learn ligand-aware protein representations. Specifically, we construct the Protein-Ligand Interaction Text Dataset (PLI-TD) by translating structured PLIP profiles into grounded, natural-language interaction descriptions. We then introduce PLI-CLIP, a CLIP-style contrastive framework that aligns protein features with PLI-TD to inject explicit interaction semantics. While these textual descriptions specify which interactions (e.g., hydrogen bonds, hydrophobic contacts) occur, they do not localize them. Therefore, PLI-CLIP is further trained to predict per-residue binding geometry, transforming geometry-blind representations into geometry-aware ones. Finally, these representations serve as a transferable module that integrates into existing structure prediction models without modifying their original architectures, eliminating the need for text annotations or interaction profiles at inference time. PLI-CLIP consistently improves docking accuracy across various recent downstream docking models, raising DiffDock's apo top-1 rate ($24.4 \rightarrow 30.3$), NeuralPLexer's < 2 Å fraction ($36.6 \rightarrow 40.2$), and FlowDock's docking success rate ($50.4 \rightarrow 54.1$). [1]

1 INTRODUCTION

Understanding the interactions between small molecules and proteins is a core problem in structural biology and underpins modern drug discovery pipelines, including virtual screening, lead optimization, and structure-based drug design Kitchen et al. (2004); Sadybekov & Katritch (2023). The ability to reliably predict protein–ligand complex structures enables researchers to determine binding poses and rank candidate molecules before committing to expensive experimental validation. Recently developed deep learning models for protein–ligand structure prediction and docking, including DiffDock Corso et al. (2023), NeuralPLexer Qiao et al. (2024), and FlowDock Morehead & Cheng (2025), have achieved remarkable progress by leveraging protein representations derived from large-scale protein language models (PLMs) such as ESM-2 Lin et al. (2023). These representations encode rich evolutionary and structural information directly from protein sequences, offering powerful priors for subsequent structural modeling tasks.

However, this dependence on PLMs reveals a crucial limitation: since conventional PLMs are trained only on amino acid sequences, they lack explicit information about how a specific protein interacts with a particular ligand. This is important because binding is intrinsically ligand-dependent: the very same residue can serve as a hydrogen-bond donor for one ligand yet form a hydrophobic interaction with another, a difference that a ligand-agnostic PLMs cannot capture. As a result, downstream docking models need to learn ligand-specific interaction patterns from task-specific structural supervision, rather than taking them as explicit pretrained interaction priors as inputs.

[1] Code and dataset will be made publicly available upon acceptance.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

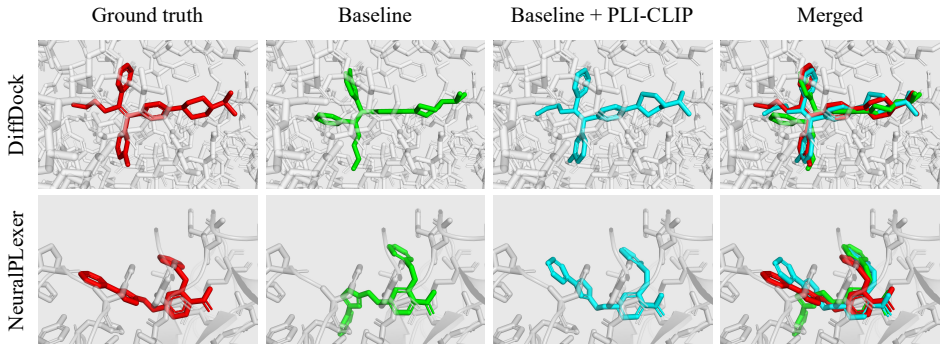


Figure 1: Visual comparison of predictions from protein-ligand docking models without (baseline) and with (baseline + PLI-CLIP) PLI-CLIP injection. With PLI-CLIP, the predicted ligand poses align more closely with the ground truth. Top: DiffDock is used as the baseline. Bottom: NeuralPLexer is used as the baseline.

Recently, this ligand-specific interaction information can be systematically extracted from experimentally resolved complexes. Structure-based tools such as the Protein–Ligand Interaction Profiler (PLIP) Salentin et al. (2015) apply rule-based geometric criteria to detect non-covalent interactions directly from 3D structures, providing structured, atom-level records of the interacting residues, their interaction types, and the associated geometric attributes. These profiles capture precisely the ligand-conditioned physicochemical knowledge that sequence-only PLMs cannot express; yet, they remain largely unexploited as direct supervision for representation learning in docking models. A straightforward approach to exploit them is to verbalize PLIP profiles and use the resulting text as a supervision signal, since protein-text pretraining has shown that language can augment sequence-only learning by providing protein encoders with explicit biological semantics Xu et al. (2023); Wang et al. (2024). However, existing protein-text datasets Xu et al. (2023); Bairoch & Apweiler (2000); Zhang et al. (2022) are protein-centric and ligand-agnostic, while protein-ligand resources provide structures and interaction records without accompanying text. To our knowledge, no public dataset bridges the two by pairing experimentally determined complexes with natural-language descriptions of residue level protein-ligand interactions. Beyond this missing pairing, another challenge is availability: a reliable PLIP profile requires a known binding pose, so profiles exist for experimentally resolved training complexes but not for unseen pairs before structure prediction. Collectively, these gaps give rise to the following research question:

> *Can structured protein–ligand interaction knowledge be transformed into text-based supervision and encoded into transferable representations that require no interaction annotations at inference?*

In this work, we address this question on three fronts. First, we introduce **PLI-TD**, a new protein–ligand interaction text dataset that converts heterogeneous PLIP profiles into multiple semantically equivalent but syntactically diverse natural-language descriptions, pairing each complex with interaction-grounded text. Second, we propose **PLI-CLIP**, a CLIP-style contrastive alignment framework that aligns ligand-aware residue representations with PLI-TD interaction texts by enabling residues to cross-attend to ligand atoms, distilling interaction knowledge into per-residue embeddings. Third, we make these representations **transferable**: a lightweight residual adapter adds these representations at the PLM-feature interface of standard docking backbones, enabling interaction-aware refinement without altering backbone architectures. Note that since interaction texts are used exclusively during the pretraining of PLI-CLIP, downstream inference does not require any PLIP profiles or textual annotations. Our experiments across DiffDock, NeuralPLexer, and FlowDock demonstrate consistent improvements in blind docking accuracy, with especially strong gains on difficult complexes as shown in Fig. 1.

2 Related Works

Protein–Ligand Docking and Structure Prediction. Blind molecular docking predicts the bound pose of a ligand without prior knowledge of the binding pocket. Early one-shot regression models

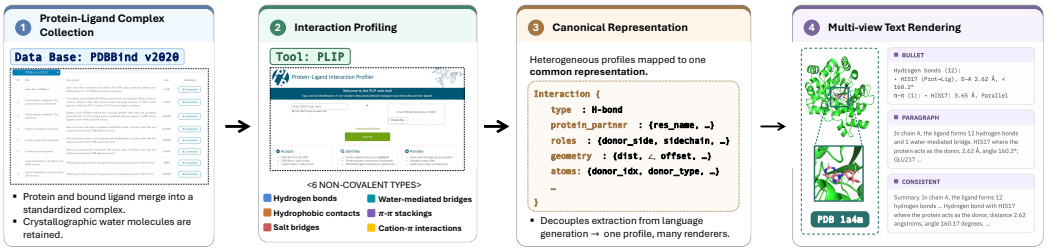
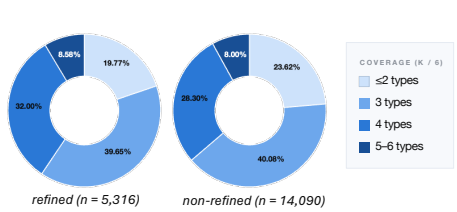


Figure 2: Proposed pipeline for constructing our protein-ligand interaction text dataset (PLI-TD).



(a) Interaction-Type Coverage

Interaction type	Complexes	Total
Hydrogen bond	18,688 (96.3%)	123,623
Hydrophobic contact	17,018 (87.7%)	88,630
Water bridge	12,560 (64.7%)	45,404
Salt bridge	7,106 (36.6%)	14,536
π-stacking	5,199 (26.8%)	7,944
π-cation	1,889 (9.7%)	2,521
Total	median 13 per complex	282,658

(b) Interaction-Type Statistics

Figure 3: PLI-TD statistics: (a) per-complex coverage of interaction types for the refined and non-refined subsets, and (b) the number of complexes containing each interaction type and the corresponding total instance count.

such as EquiBind Stärk et al. (2022) and TankBind Lu et al. (2022) replaced slow search-based programs with fast graph neural networks, and diffusion- and flow-based generative models have since largely closed the accuracy gap: DiffDock Corso et al. (2023; 2024) casts docking as SE(3) diffusion over the ligand pose, while DynamicBind Lu et al. (2024) and FlowDock Morehead & Cheng (2025) additionally model conformational flexibility and affinity. All-atom complex predictors such as NeuralPLexer Qiao et al. (2024; 2025) and AlphaFold3 Abramson et al. (2024) jointly generate protein and ligand atoms at atomistic resolution. Crucially, the sequence-based models in this family draw their protein features from pretrained PLMs (e.g., ESM-2 Lin et al. (2023)) and must infer protein–ligand interaction patterns from geometric supervision alone, with no explicit interaction knowledge available.

Protein Language Models. Protein language models (PLMs) learn residue-level representations through self-supervised training on large sequence corpora. ESM-2 Lin et al. (2023) and ProtT5 Elnaggar et al. (2022) capture evolutionary and structural signals that transfer broadly, and have become the default protein encoders for docking and structure prediction. More recent variants incorporate geometric context, either through a structure-aware vocabulary Su et al. (2024) or by extending the model to atom-level resolution Zheng et al. (2024). Despite their strength, these models are trained purely on sequences (or sequences and structures) and therefore encode no knowledge of how a particular protein interacts with a particular ligand. Our approach builds on a frozen PLM and supplement this missing interaction knowledge through contrastive pretraining.

Protein–Text Contrastive Learning. Contrastive vision–language learning, pioneered by CLIP Radford et al. (2021), aligns paired modalities in a shared embedding space, and has been extended to molecules through text-guided molecular models Liu et al. (2023) and pocket–molecule retrieval Gao et al. (2023). For proteins, ProtST Xu et al. (2023), ProtCLIP Zhou et al. (2025), ProteinCLIP Wu et al. (2024), and ProTrek Su et al. (2025) align sequences with natural-language descriptions, while OntoProtein Zhang et al. (2022) injects ontology knowledge. In addition, PLM-based models predict drug–target interaction or binding affinity directly from sequences and ligand SMILES Singh et al. (2023); Rose et al. (2024). These previous approaches align proteins with functional or property descriptions, or target interaction and affinity labels. To our knowledge, PLI-CLIP is the first to use natural-language descriptions of residue-level interaction records as the contrastive target for protein–ligand representation learning. Prior protein–text models align sequences

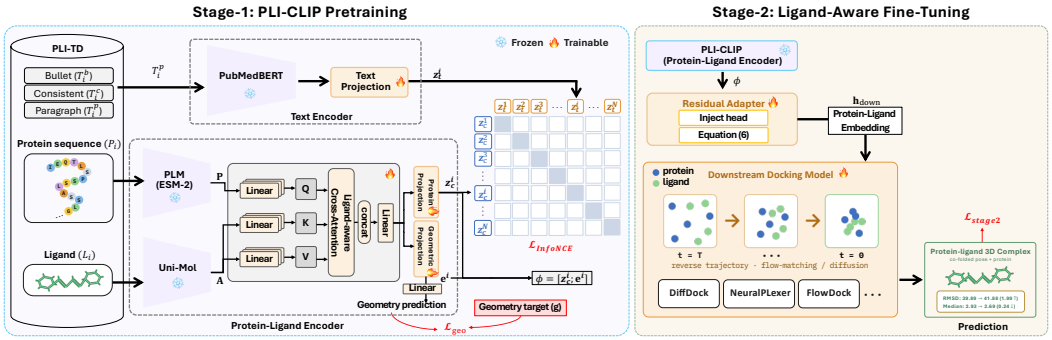


Figure 4: Overall architecture of PLI-CLIP. Stage 1 pretrains the ligand-aware protein encoder with contrastive alignment and geometry prediction; Stage 2 transfers the learned representation into protein-ligand docking models through a residual adapter.

with functional or property descriptions, while prior interaction-aware pretraining (e.g., Feng et al. (2024)) supervises with geometric targets rather than language. PLI-CLIP combines the two, distilling PLIP-derived interaction semantics into representations explicitly tailored for downstream docking.

3 PLI-TD

In this section, we describe the construction of the proposed protein–ligand interaction text dataset (PLI-TD). To create this dataset, we design an interaction-to-text pipeline that generates textual descriptions of interactions from experimentally resolved protein–ligand complexes. As illustrated in Fig. 2, the pipeline comprises four sequential steps: (1) Protein-Ligand Complex Collection, (2) Interaction Profiling, (3) Canonical Representation, and (4) Multi-view Text Rendering.

Step 1: Protein-Ligand Complex Collection. We construct PLI-TD from the PDBbind v2020 general set Wang et al. (2005), comprising 19,443 experimentally determined protein-ligand complexes, each pairing a protein structure with its bound ligand. For each complex, the ligand and protein are merged into a standardized structure, with crystallographic water retained to enable the detection of water-mediated interactions. See Appendix A.1 for details.

Step 2: Interaction Profiling. To extract text-based protein-ligand interaction information, we employ the PLIP Salentin et al. (2015) and retain six types of non-covalent interactions: hydrogen bonds, hydrophobic contacts, salt bridges, water-mediated bridges, π-π stacking, and cation-π interactions, as illustrated in Fig. 3 (b). For each interaction, PLIP records the participating protein residue together with interaction-specific attributes such as molecular roles, distances, and angles. Complexes for which interaction profiling or profile parsing fails are excluded from subsequent text generation.

Step 3: Canonical Representation. PLIP profiles are machine-readable, but each interaction type is associated with its own distinct attributes, so there is no single template that can render all interaction types. To address this, we convert every profile into a canonical interaction schema (Appendix A.3), which encodes each interaction using a unified set of fields: the interaction type, the participating residue, the molecular role, and the corresponding type-specific attributes. This canonical schema enables us to generate multiple textual representations from the same profile without repeatedly parsing each interaction type.

Step 4: Multi-view Text Rendering. For each protein-ligand complex, we generate from the canonical representation three complementary textual views that describe the same interaction records, but differ in linguistic organization: *bullet*, *paragraph*, and *consistent* (Fig. 2) views. The bullet view presents interactions as compact entries grouped by type, using symbols and abbreviations to make each interaction and its attributes machine-readable. The paragraph view expresses

the same content as fluent prose, yielding a human-readable form closer to the natural-language distribution seen by pretrained text encoders. The consistent view combines both aspects.: it verbalizes interactions as complete sentences yet follows a fixed category order and sentence template, thereby retaining the regularity of the bullet view while remaining in natural-language form. Together, the three views expose the encoder to the same interaction semantics under different linguistic surfaces. Since the optimal view remains unexplored, we generate diverse views to analyze their distinct characteristics in our experiments.

PLI-TD Composition. Each entry in PLI-TD is denoted as $\{P_i, L_i, T_i^b, T_i^c, T_i^p\}$ for complex i, where P_i is the protein sequence, L_i is the ligand 3D structure, and T_i^b, T_i^c, and T_i^p are the bullet, consistent, and paragraph interaction texts, respectively. Fig. 3 (a) shows the per-complex coverage of interaction types, which is nearly identical across the refined and non-refined subsets, and (b) summarizes the per-type statistics over all 19,406 complexes, where hydrogen bonds and hydrophobic contacts dominate, and each complex contains a median of 13 interactions.

4 PLI-CLIP

Overview. As shown in Fig. 4, PLI-CLIP first learns protein–ligand representations and then adapts existing molecular docking models to exploit this information in a two-stage process. In Stage-1, we pretrain PLI-CLIP with a CLIP-style two-tower contrastive objective. One tower encodes a protein-ligand pair: the Protein-Ligand Encoder updates each residue embedding using the input ligand. The other tower encodes the interaction text: the Text Encoder turns the text description into another embedding. During pretraining, these embeddings for correct pairs are pulled closer together, while those for incorrect pairs are pushed further apart. In Stage-2, the ligand-aware protein representation from Stage-1 is injected into a molecular docking backbone through a lightweight residual adapter, and the backbone is finetuned while PLI-CLIP remains frozen. Since the Protein-Ligand Encoder of PLI-CLIP produces this protein-ligand representation from only the protein sequence and the unbound ligand, it functions as a lightweight transferable module at inference.

Architecture. PLI-CLIP is built on three frozen pretrained backbones: ESM-2 (PLM), which produces residue-level protein representations; Uni-Mol Zhou et al. (2023), which generates atom-level ligand representations; and PubMedBERT Gu et al. (2021), which embeds the interaction description. On top of these, we introduce four trainable components: Text Projection, Ligand-aware Cross-Attention (LCA), Protein Projection, and Geometric Projection.

First, within the Text Encoder, for a given interaction text T_i from PLI-TD, the frozen PubMedBERT model generates a text embedding. This embedding is then fed into the Text Projection module, a two-layer MLP, which maps it into a shared d-dimensional space, yielding the final text embedding $\mathbf{z}_t^i$.

Second, within the Protein-Ligand Encoder (protein-ligand side), we enhance the ligand-agnostic PLM features by conditioning each residue on the candidate ligand. To be specific, given a protein sequence P_i and a ligand L_i from PLI-TD, we extract frozen per-residue PLM features $\mathbf{P}$ and frozen atom-level Uni-Mol features $\mathbf{A}$. We then project both modalities into a common d-dimensional space and treat the residues as queries ($\mathbf{Q}$), the ligand atoms as keys ($\mathbf{K}$), and values ($\mathbf{V}$) as follows:

$$\mathbf{Q} = \mathbf{P} \times W_q, \qquad \mathbf{K} = \mathbf{V} = \mathbf{A} \times W_{kv}, \tag{1}$$

where W_q and W_{kv} are learnable query and key/value projections. The residues are then conditioned on the ligand through LCA, each with a residual connection and layer normalization as:

$$\mathbf{Q} \leftarrow \mathrm{LayerNorm}\big(\mathbf{Q} + \mathrm{LCA}\big(\mathbf{Q}, \mathbf{K}, \mathbf{V}\big)\big), \tag{2}$$

where LCA denotes two-layer multi-head cross-attention Vaswani et al. (2017) from the residue queries ($\mathbf{Q}$) to the ligand keys ($\mathbf{K}$) and values ($\mathbf{V}$). Because the ligand atoms serve as keys and values, the post-attention embedding $\mathbf{Q} \in \mathbb{R}^{n \times d}$, where n denotes the number of residues and d is the shared embedding dimension, is ligand-aware: the representation of each residue is conditioned on the specific candidate ligand. Finally, given $\mathbf{Q}$, two two-layer MLP heads generate distinct outputs: the Protein Projection produces the complex-level embedding $\mathbf{z}_c^i$ employed for contrastive alignment directly, and the Geometric Projection produces the per-residue geometric embedding $\mathbf{e}^i$. We provide these features to downstream models in the form of a single concatenated vector, denoted as $\boldsymbol{\phi} = [\mathbf{z}_c^i; \mathbf{e}^i]$.

PLI-CLIP Pretraining (Stage-1). We pretrain PLI-CLIP by jointly optimizing two loss functions: an Interaction-Aware Contrastive Alignment loss ($\mathcal{L}_{\text{InfoNCE}}$) and a Geometric Prediction loss ($\mathcal{L}_{\text{geo}}$). The former guides the protein embeddings to capture the interaction semantics expressed in the text, while the latter endows these representations with geometric information describing the spatial relationship between the protein and the ligand. By integrating these two objectives, PLI-CLIP is pretrained in the first stage using

$$\mathcal{L}_{\text{stage1}} = \mathcal{L}_{\text{InfoNCE}} + \lambda\, \mathcal{L}_{\text{geo}}, \tag{3}$$

where λ determines the trade-off between the two loss terms.

• Interaction-Aware Contrastive Alignment Loss: The interaction text T_i specifies which protein residues are in contact with the ligand, a type of information that a sequence-only PLM is unable to represent. Inspired by the effectiveness of contrastive alignment in cross-modal learning Radford et al. (2021); Zhou et al. (2025); Xu et al. (2023), we adopt this approach to project both the ligand-aware protein representation and the interaction text into a common d-dimensional space. Given a batch of N complex-text pairs $\{(\mathbf{z}_c^i, \mathbf{z}_t^i)\}_{i=1}^N$, we optimize the symmetric InfoNCE van den Oord et al. (2018) objective as:

$$\mathcal{L}_{\text{InfoNCE}} = -\frac{1}{2N}\sum_{i=1}^{N}\left[\log\frac{\exp(\text{sim}(\mathbf{z}_c^i, \mathbf{z}_t^i)/\tau)}{\sum_{j=1}^{N}\exp(\text{sim}(\mathbf{z}_c^i, \mathbf{z}_t^j)/\tau)} + \log\frac{\exp(\text{sim}(\mathbf{z}_t^i, \mathbf{z}_c^i)/\tau)}{\sum_{j=1}^{N}\exp(\text{sim}(\mathbf{z}_t^i, \mathbf{z}_c^j)/\tau)}\right], \tag{4}$$

where $\text{sim}(\cdot,\cdot)$ is the cosine similarity and τ is a learnable temperature. This pulls matched complex-text pairs together while pushing unmatched pairs apart, injecting interaction semantics into the encoder.

Table 1: Performance of existing docking backbones before and after PLI-CLIP injection. (↑ higher is better; ↓ lower is better).

Dataset: PDBBind2020 benchmark	Holo Crystal Proteins				Apo ESMFold Proteins			
	Top-1 RMSD		Top-5 RMSD		Top-1 RMSD		Top-5 RMSD	
	%<2Å↑	Median ↓	%<2Å↑	Median ↓	%<2Å↑	Median ↓	%<2Å↑	Median ↓
DiffDock	39.89	2.93	47.01	2.16	24.36	6.01	35.41	3.66
DiffDock + PLI-CLIP	**41.88**	**2.69**	**49.57**	**2.02**	**30.31**	**4.72**	**38.81**	**3.22**

Dataset: PDBBind2020 benchmark	Top-1 Ligand RMSD				Top-5 Ligand RMSD			
	Percentiles ↓		Below threshold ↑		Percentiles ↓		Below threshold ↑	
	25th	50th	RMSD<5Å	RMSD<2Å	25th	50th	RMSD<5Å	RMSD<2Å
NeuralPLexer	1.65	3.21	65.1	36.6	1.55	3.09	75.1	46.0
NeuralPLexer + PLI-CLIP	**1.30**	**2.62**	**69.5**	**40.2**	**1.05**	**2.14**	**77.3**	**47.9**

Dataset: PoseBusters	No Post-Processing		With Relaxation	
	RMSD<2Å↑	RMSD<2Å& PB-Valid ↑	RMSD<2Å↑	RMSD<2Å& PB-Valid ↑
FlowDock	71.4	34.0	71.0	50.4
FlowDock + PLI-CLIP	**72.2**	**38.4**	**72.2**	**54.1**

• Geometry Prediction Loss: Our contrastive alignment framework identifies which interactions are present but does not encode the spatial relationship between the protein and ligand; because the encoder input is pose-free, the resulting representation would otherwise lack information about the geometry of the binding pocket. Since downstream molecular docking is inherently a geometric problem, we introduce an auxiliary objective that anchors the representation to the underlying binding geometry.

Concretely, for each residue r of a given complex, we derive a target $\mathbf{g}_r \in \mathbb{R}^6$ from the crystal structure that summarizes how closely and how densely the residue contacts the ligand across multiple distance scales (Appendix B.1). Subsequently, a lightweight Geometric Projection transforms each residue feature $\mathbf{Q}_r \in \mathbb{R}^d$ from $\mathbf{Q} \in \mathbb{R}^{n\times d}$ into a compact geometry embedding $\mathbf{e}_r^i \in \mathbb{R}^{64}$, yielding $\mathbf{e}^i \in \mathbb{R}^{n\times 64}$ for the full sequence. A linear head $\text{Linear}(\cdot) : \mathbb{R}^{64} \to \mathbb{R}^6$ then regresses the target from this embedding, trained with a mean-squared error,

$$\mathcal{L}_{\text{geo}} = \frac{1}{n}\sum_{r=1}^{n}\left\|\text{Linear}(\mathbf{e}_r^i) - \mathbf{g}_r\right\|^2, \tag{5}$$

where n is the number of residues. The geometry objective not only provides a more expressive residue representation but also generates a reusable feature: we append $\mathbf{e}^i$ to $\mathbf{z}_c^i$ to obtain the spatially interaction-aware per-residue representation $\boldsymbol{\phi}$, ensuring that the predicted binding geometry is propagated to later stages.

Table 2: PLI-CLIP with different ESM-2 sizes (150M/650M/3B), injected into DiffDock and NeuralPLexer on Holo (crystal) proteins of the PDBBind2020 benchmark.

DiffDock	Top-1 RMSD		Top-5 RMSD		NeuralPLexer	Top-1 RMSD		Top-5 RMSD	
	%<2Å ↑	%<5Å ↑	%<2Å ↑	%<5Å ↑		%<2Å ↑	%<5Å ↑	%<2Å ↑	%<5Å ↑
Baseline	39.89	60.97	47.01	74.36	Baseline	36.6	65.1	46.0	75.1
+ PLI-CLIP (150M)	39.89	62.96	**50.14**	76.64	+ PLI-CLIP (150M)	39.6	67.6	46.5	77.0
+ PLI-CLIP (650M)	**41.88**	**66.38**	49.57	**77.49**	+ PLI-CLIP (650M)	40.2	**69.5**	**47.9**	77.3
+ PLI-CLIP (3B)	40.74	64.96	47.58	76.92	+ PLI-CLIP (3B)	**41.0**	69.0	47.4	**77.8**

Table 3: Ablation on the text view (bullet, consistent, paragraph) used to render each interaction profile. Each backbone is evaluated with its own metrics.

Text view	**DiffDock** on PDBBind				**NeuralPLexer** on PDBBind				**FlowDock** on PoseBusters			
	Holo Crystal		Apo ESMFold		Top-1		Top-5		No Post-proc.		With Relaxation	
	Top-1 ↑	Top-5 ↑	Top-1 ↑	Top-5 ↑	<2 Å ↑	<5 Å ↑	<2 Å ↑	<5 Å ↑	<2Å ↑	& PB-valid ↑	<2Å ↑	& PB-valid ↑
Bullet	40.17	48.15	24.08	34.84	36.0	64.8	45.2	77.0	72.2	37.5	71.9	49.8
Consistent	39.60	49.29	30.03	**39.66**	37.1	63.2	46.5	**79.8**	**72.7**	35.1	**73.4**	52.1
Paragraph	**41.88**	**49.57**	**30.31**	38.81	**40.2**	**69.5**	**47.9**	77.3	72.2	**38.4**	72.2	**54.1**

Transferring to Downstream Models (Stage-2). Existing downstream docking models for protein–ligand structure prediction typically rely on protein representations derived from a pretrained PLM. In contrast, our PLI-CLIP generates a ligand-aware, interaction-centric embedding for each residue from the same inputs. In Stage–2 of the proposed approach, we inject this learned knowledge precisely at the PLM-feature interface, enriching the downstream model's per-residue representation via a lightweight residual adapter as follows:

$$\mathbf{h}_{\text{down}} = \mathbf{h}_{\text{orig}} + \exp(s)\ \tanh\left(W_{\text{inj}}\boldsymbol{\phi} + b_{\text{inj}}\right), \tag{6}$$

where $\mathbf{h}_{\text{orig}}$ and $\mathbf{h}_{\text{down}}$ are the backbone's per-residue representations before and after injection, and $\boldsymbol{\phi} = [\mathbf{z}_c^i; \mathbf{e}^i]$ is the transferred representation from the Protein-Ligand Encoder of PLI-CLIP. Trainable parameters W_{inj} and b_{inj} are initialized to zero, s is a learnable scalar, and $\tanh$ is used to nonlinearly transform the representation, while $\exp(s)$ adjusts its magnitude. Note that, since the projection is initialized to all zeros, the adapter initially behaves like an identity function and exactly matches the pretrained model, ensuring that fine-tuning only deviates from the original feature distribution to the extent warranted by the data.

In Stage-2, we introduce no additional objectives: the Stage-2 loss $\mathcal{L}_{\text{stage2}}$ is simply the structure-prediction loss employed by the downstream models, and PLI-CLIP is kept fixed (frozen) throughout this stage.

5 Experiments

Implementation Details. In Stage-1, the learnable parameters of PLI-CLIP shown in Fig. 4 are trained using the AdamW optimizer with a learning rate of 3×10^{-4} and a weight decay of 0.01, a batch size of 64, and gradient clipping of 1.0 for 25 epochs, with ESM-2 (650M), Uni-Mol, and PubMedBERT as frozen encoders. In Stage-2, the learnable parameters, the Inject head of residual adapter and the docking backbone, are finetuned on each backbone's original training data for 10 epochs using a learning rate of 1×10^{-4}, while PLI-CLIP remains frozen. For evaluation, PLI-CLIP consumes the same unbound ligand input as the host backbone, with no bound-state conformational information available at inference. We additionally use an exponential moving average (EMA) of the finetuned weights for more stable predictions. Further training details for each downstream model are provided in Appendix B.

Datasets. We pretrain PLI-CLIP on PLI-TD complexes for which per-residue crystal-contact geometry targets are available, excluding cases of sequence-structure mismatch (e.g., residues unresolved in the crystal) or ligand parsing failure (Appendix B.1), which yields $12{,}690$ complexes. To prevent any information leakage into the downstream evaluation, we then remove from this pool the union of all evaluation complexes used in our experiments, namely the test splits of DiffDock, NeuralPLexer, and FlowDock together with the PoseBusters set, rather than holding out a single benchmark split. This exclusion leaves $9{,}789$ training and $1{,}185$ validation complexes, and guarantees that no complex seen during Stage-1 pretraining appears in any benchmark on which PLI-CLIP is subsequently

Table 4: Ablation on the two Stage-1 objectives $\mathcal{L}_{\text{InfoNCE}}$ and $\mathcal{L}_{\text{geo}}$ on NeuralPLexer. Removing an objective also drops the component it supervises from $\phi = [\mathbf{z}_c^i; \mathbf{e}^i]$.

NeuralPLexer on PDBBind									
Stage-1 objective		Top-1 Ligand RMSD				Top-5 Ligand RMSD			
		Percentiles ↓		Below threshold ↑		Percentiles ↓		Below threshold ↑	
Text ($\mathcal{L}_{\text{InfoNCE}} \to \mathbf{z}_c^i$)	Geometry ($\mathcal{L}_{\text{geo}} \to \mathbf{e}^i$)	25th	50th	<5 Å	<2 Å	25th	50th	<5 Å	<2 Å
✗	✗	1.65	3.21	65.1	36.6	1.55	3.09	75.1	46.0
✗	✓	1.45	3.01	66.1	36.5	1.31	2.72	75.8	46.2
✓	✗	1.46	3.19	63.4	36.8	1.07	2.38	76.5	45.9
✓	✓	**1.30**	**2.62**	**69.5**	**40.2**	**1.05**	**2.14**	**77.3**	**47.9**

evaluated. All design choices, including the text rendering view and the ESM-2 scale, are selected on this validation split and never on the evaluation benchmarks. For evaluation, we use two standard benchmarks, PDBBind Wang et al. (2005) and PoseBusters Buttenschoen et al. (2024), both of which consist of experimentally determined protein–ligand complexes held out from training. On PDBBind, we sample multiple candidate poses per complex and rank them by the model's ligand confidence, reporting the fraction of complexes whose top-ranked pose falls within 2 Å and 5 Å ligand RMSD of the crystal structure. On PoseBusters, we report the fraction of predictions that are both within 2 Å RMSD and pass the PoseBusters physical-validity checks (PB-valid), which assess steric and geometric plausibility.

Downstream Models. To evaluate how PLI-CLIP can be effectively integrated, we inject the learned representations into three protein–ligand structure prediction backbones, DiffDock Corso et al. (2023), NeuralPLexer Qiao et al. (2024), and FlowDock Morehead & Cheng (2025), and then finetune each backbone in Stage–2 using the injected features, optimizing only with their respective original loss functions. For each backbone, we adopt the evaluation protocol and metrics of its original paper and compare against baselines that we reproduce with the official released code, ensuring that the comparison with and without PLI-CLIP is made under identical conditions.

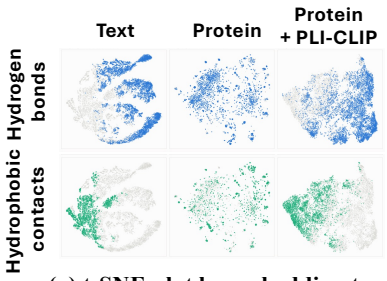
5.1 PLI-CLIP on Downstream Models

Table 1 reports each docking backbone with and without PLI-CLIP injection, where PLI-CLIP is pretrained on the paragraph view in PLI-TD. Across all three backbones, injecting the learned representation consistently improves blind docking accuracy, confirming that the PLI-CLIP embedding is readily transferable. For DiffDock, the top-1 success rate ($\% < 2$ Å) rises from 39.89 to 41.88 on holo proteins and, more markedly, from 24.36 to 30.31 on the harder apo proteins, with the median RMSD dropping in every setting. NeuralPLexer with PLI-CLIP improves at both the 25th and 50th RMSD percentiles and raises the fraction of predictions below 5 Å (e.g., $65.1 \to 69.5$ on holo). FlowDock shows its largest gains on physical validity, where the fraction of poses that are both accurate and PB-valid rises from 34.0 to 38.4 without post-processing and from 50.4 to 54.1 with relaxation. Notably, the gains are largest on apo structures and on the less strict quality metrics (median, < 5 Å, top-5), indicating that PLI-CLIP helps most precisely where the backbone's own geometric priors are weakest.

In Table 2, to verify that these gains are not tied to a particular PLM in PLI-CLIP, we pretrain PLI-CLIP with a smaller PLM (ESM2-150M) and a larger PLM (ESM2-3B) instead of the default ESM2-650M, and inject PLI-CLIP into DiffDock and NeuralPLexer. PLI-CLIP improves both backbones at nearly every scale, confirming that the transferred representation is not tied to a specific PLM. Performance, however, saturates beyond 650M: scaling to 3B yields no consistent gain over 650M on either backbone. We therefore adopt ESM2-650M as the default encoder throughout experiments, since the larger model offers a marginal accuracy benefit at a substantially higher cost.

Ablation Studies. We conduct ablation studies to analyze the impact of textual views and geometry prediction, and further examine how the learned embeddings change with and without PLI-CLIP via t-SNE visualization and supervised classifiers.

• **Effect of Textual Views:** Table. 3 compares the three textual views used to render each interaction profile: bullet, consistent, and paragraph. The paragraph view achieves the highest downstream accuracy across all backbones, while the bullet view is consistently worst, especially on apo proteins



(a) t-SNE plot by embedding type

Embedding	Logistic regression	MLP
Protein	0.812	0.816
Protein + Ligand	0.901	0.909
Protein + PLI-CLIP	0.964	0.967

(b) Classification performance

Figure 5: Interaction awareness of the learned embeddings by embedding type. (a) t-SNE plots. (b) Classification AUROC.

(e.g., DiffDock apo $\% < 2$ Å of 24.08 versus 30.31 for paragraph). We attribute the advantage of the paragraph view to the Text Encoder: as natural prose, it better matches the biomedical text on which PubMedBERT was pretrained, providing richer, more semantically structured supervision for residue representations than the terser bullet and consistent views. We therefore use the paragraph view for all main experiments.

• **Effect of Text and Geometry Supervision:** Table 4 isolates the contribution of each supervision signal by ablating the two Stage-1 objectives, where removing an objective also drops the component of $\phi = [\mathbf{z}_c^i; \mathbf{e}^i]$ it supervises. Each objective is individually beneficial: geometry prediction alone lowers the top-1 median RMSD from 3.21 to 3.01 Å and raises the <5 Å rate to 66.1, while contrastive alignment alone improves the top-5 25th-percentile RMSD from 1.55 to 1.07 Å. Their individual gains are nevertheless modest, and only their combination delivers a substantial improvement, raising the top-1 rates to 40.2 and 69.5 and attaining the best value on every top-5 metric. This super-additive behavior suggests that the two objectives play distinct and mutually dependent roles. The geometric objective supplies dense spatial targets but no account of what the predicted contacts mean, leaving the encoder without a basis for deciding which residues warrant precise modeling. Natural-language supervision resolves this ambiguity by specifying which interaction types are formed and with which residues, yet it does not localize them in space. The two are therefore effective only together: interaction semantics determine where geometric precision matters, and geometry grounds those semantics into pose-level accuracy.

• **Interaction Awareness of the Learned Embeddings:** Beyond downstream accuracy, Fig. 5 analyzes how the PLI-CLIP representation clusters by grouping complexes according to their dominant interaction type and comparing embeddings before and after pretraining. As shown in Fig. 5 (a), hydrogen-bond-dominant and hydrophobic-dominant complexes are entangled in the original PLM features but form separate clusters after PLI-CLIP. Fig. 5 (b) quantifies this separation using logistic regression and 3-layer MLP classifiers trained on the frozen embeddings and evaluated on unseen proteins. AUROC rises from 0.812 with protein features alone to 0.901 when the ligand is included without training, and further to 0.964 with PLI-CLIP, showing that contrastive pretraining adds discriminative interaction information beyond ligand conditioning itself.

6 Conclusion

We presented PLI-CLIP, a framework that distills protein-ligand interaction knowledge into transferable residue-level representations for structure prediction. It aligns ligand-aware residue representations with PLIP-derived interaction text through contrastive pretraining, and, since text captures which interactions occur but not where, additionally predicts and fuses the binding geometry to make the representation pose-aware. The learned features are transferable through a lightweight residual adapter, requiring no interaction profiles, text, or bound pose at inference, and consistently improve blind docking accuracy across DiffDock, NeuralPLexer, and FlowDock, with the largest gains on challenging apo structures. Promising directions for future work include extending PLI-CLIP to all-atom cofolding models such as AlphaFold3 Abramson et al. (2024) and NeuralPLexer3 Qiao et al. (2025), and developing a self-contained injection module that improves docking accuracy without updating the backbone.

AI USE STATEMENT

In this work, we used generative AI tools for writing assistance. We have not used generative AI tools for research ideation, experimental design, or the generation of experimental results. Retrieval and summarization of related work through generative AI tools is not applicable to this work, as the literature survey was conducted manually. Additionally, we used generative AI tools to assist with aggregating and cross-checking numerical results reported in our tables. We have reviewed all AI-assisted work: every number appearing in the text was verified against the corresponding table by the authors, and all AI-drafted text was rewritten or edited by the authors before inclusion. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work does not involve human subjects, personally identifiable information, or sensitive data. PLI-TD is derived from PDBbind, a publicly available database of experimentally determined protein–ligand complexes, and we release only the textual interaction descriptions generated by our pipeline, in accordance with the terms under which the source data are distributed. Improvements in protein–ligand structure prediction primarily benefit therapeutic discovery, but the same capability could in principle be applied to the design of harmful compounds. We consider this risk low, as our method improves the accuracy of existing public docking models rather than enabling any new capability, and it provides no information that is not already obtainable from those models. We declare no conflicts of interest.

REPRODUCIBILITY STATEMENT

The construction of PLI-TD is described in the PLI-TD section of the main text and detailed in Appendix A, which specifies the source database, the tool versions used for ligand parsing and interaction profiling, the complex standardization rules, the canonical schema fields retained for each interaction type, and the deterministic rendering rules that fix ordering and numeric precision for the three textual views. The construction of the per-residue geometry target, together with the alignment requirement that determines which complexes admit a target, is given in Appendix B.1. Architecture and training settings for both stages, including all hyperparameters and the per-backbone injection settings, are listed in Appendix B. Evaluation follows the protocol and metrics of each backbone's original paper, and all baselines are reproduced with the officially released code so that the comparison with and without PLI-CLIP is made under identical conditions. **Code and the PLI-TD dataset will be released upon acceptance.**

REFERENCES

Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J. Ballard, Joshua Bambrick, et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, 2024.

Amos Bairoch and Rolf Apweiler. The swiss-prot protein sequence database and its supplement trembl in 2000. *Nucleic acids research*, 2000.

Martin Buttenschoen, Garrett M. Morris, and Charlotte M. Deane. Posebusters: Ai-based docking methods fail to generate physically valid poses or generalise to novel sequences. *Chemical Science*, 2024.

Gabriele Corso, Hannes Stärk, Bowen Jing, Regina Barzilay, and Tommi S. Jaakkola. Diffdock: Diffusion steps, twists, and turns for molecular docking. In *ICLR*, 2023.

Gabriele Corso, Arthur Deng, Benjamin Fry, Nicholas Polizzi, Regina Barzilay, and Tommi Jaakkola. Deep confident steps to new pockets: Strategies for docking generalization. In *ICLR*, 2024.

Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, Debsindhu Bhowmik, and

Burkhard Rost. Prottrans: Toward understanding the language of life through self-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

Shikun Feng, Minghao Li, Yinjun Jia, Wei-Ying Ma, and Yanyan Lan. Protein-ligand binding representation learning from fine-grained interactions. In *ICLR*, 2024.

Bowen Gao, Bo Qiang, Haichuan Tan, Yinjun Jia, Minsi Ren, Minsi Lu, Jingjing Liu, Wei-Ying Ma, and Yanyan Lan. Drugclip: Contrastive protein-molecule representation learning for virtual screening. In *NeurIPS*, 2023.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 2021.

Douglas B. Kitchen, Hélène Decornez, John R. Furr, and Jürgen Bajorath. Docking and scoring in virtual screening for drug discovery: methods and applications. *Nature Reviews Drug Discovery*, 2004.

Greg Landrum et al. RDKit: Open-source cheminformatics software. https://www.rdkit.org, 2025. Version 2025.03.6.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 2023.

Shengchao Liu, Weili Nie, Chengpeng Wang, Jiarui Lu, Zhuoran Qiao, Ling Liu, Jian Tang, Chaowei Xiao, and Anima Anandkumar. Multi-modal molecule structure–text model for text-based retrieval and editing. *Nature Machine Intelligence*, 2023.

Wei Lu, Qifeng Wu, Jixian Zhang, Jiahua Rao, Chengtao Li, and Shuangjia Zheng. Tankbind: Trigonometry-aware neural networks for drug-protein binding structure prediction. In *NeurIPS*, 2022.

Wei Lu, Jixian Zhang, Weifeng Huang, Ziqiao Zhang, Xiangyu Jia, Zhenyu Wang, Leilei Shi, Chengtao Li, Peter G. Wolynes, and Shuangjia Zheng. Dynamicbind: Predicting ligand-specific protein-ligand complex structure with a deep equivariant generative model. *Nature Communications*, 15(1):1071, 2024.

Alex Morehead and Jianlin Cheng. Flowdock: Geometric flow matching for generative protein-ligand docking and affinity prediction. *Bioinformatics*, 41(Supplement_1):i198–i206, 2025.

Zhuoran Qiao, Weili Nie, Arash Vahdat, Thomas F. Miller III, and Animashree Anandkumar. State-specific protein–ligand complex structure prediction with a multiscale deep generative model. *Nature Machine Intelligence*, 6(2):195–208, 2024.

Zhuoran Qiao, Feizhi Ding, Thomas Dresselhaus, Mia A. Rosenfeld, Xiaotian Han, Owen Howell, Aniketh Iyengar, Stephen Opalenski, Anders S. Christensen, Sai Krishna Sirumalla, Frederick R. Manby, Thomas F. Miller III, and Matthew Welborn. Neuralplexer3: Accurate biomolecular complex structure prediction with flow models. In *NeurIPS*, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

Joshua David Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. Contrastive learning with hard negative samples. In *ICLR*, 2021.

Tyler Rose, Nicolò Monti, Navvye Anand, and Tianyu Shen. Plapt: Protein-ligand binding affinity prediction using pretrained transformers. *bioRxiv*, 2024.

Anastasiia V. Sadybekov and Vsevolod Katritch. Computational approaches streamlining drug discovery. *Nature*, 2023.

Sebastian Salentin, Sven Schreiber, V. Joachim Haupt, Melissa F. Adasme, and Michael Schroeder. Plip: Fully automated protein-ligand interaction profiler. *Nucleic Acids Research*, 2015.

Rohit Singh, Samuel Sledzieski, Bryan Bryson, Lenore Cowen, and Bonnie Berger. Contrastive learning in protein language space predicts interactions between drugs and protein targets. *PNAS*, 2023.

Hannes Stärk, Octavian Ganea, Lagnajit Pattanaik, Regina Barzilay, and Tommi Jaakkola. Equibind: Geometric deep learning for drug binding structure prediction. In *ICML*, 2022.

Jin Su, Chenchen Han, Yuyang Zhou, Junjie Shan, Xibin Zhou, and Fajie Yuan. Saprot: Protein language modeling with structure-aware vocabulary. In *ICLR*, 2024.

Jin Su, Xibin Zhou, Xuting Zhang, Fajie Yuan, et al. A trimodal protein language model enables advanced protein searches. *Nature Biotechnology*, 2025.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017.

Renxiao Wang, Xueliang Fang, Yipin Lu, Chao-Yie Yang, and Shaomeng Wang. The pdbbind database: methodologies and updates. *Journal of medicinal chemistry*, 2005.

Zeyuan Wang, Qiang Zhang, Keyan Ding, Ming Qin, Xiang Zhuang, Xiaotong Li, and Huajun Chen. InstructProtein: Aligning human and protein language via knowledge instruction. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1114–1136, Bangkok, Thailand, August 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.acl-long.62.

Kevin E. Wu, Howard Chang, and James Zou. Proteinclip: Enhancing protein language models with natural language. *bioRxiv*, 2024.

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

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

Kangjie Zheng, Siyu Long, Tianyu Lu, Junwei Yang, Xinyu Dai, Ming Zhang, Zaiqing Nie, Wei-Ying Ma, and Hao Zhou. Esm all-atom: Multi-scale protein language model for unified molecular modeling. In *ICML*, 2024.

Gengmo Zhou, Zhifeng Gao, Qiankun Ding, Hang Zheng, Hongteng Xu, Zhewei Wei, Linfeng Zhang, and Guolin Ke. Uni-mol: A universal 3d molecular representation learning framework. In *ICLR*, 2023.

Hanjing Zhou, Mingze Yin, Wei Wu, Mingyang Li, Kun Fu, Jintai Chen, Jian Wu, and Zheng Wang. Protclip: Function-informed protein multi-modal learning. In *AAAI*, 2025.