FLEXPROTEIN: JOINT SEQUENCE AND STRUCTURE PRETRAINING FOR PROTEIN MODELING

Anonymous authors
Paper under double-blind review

ABSTRACT

Protein foundation models have advanced rapidly, with most approaches falling into two dominant paradigms. Sequence-only language models (e.g., ESM-2) capture sequence semantics at scale but lack structural grounding. MSA-based predictors (e.g., AlphaFold 2/3) achieve accurate folding by exploiting evolutionary couplings, but their reliance on homologous sequences makes them less reliable in highly mutated or alignment-sparse regimes. We present FlexProtein, a pretrained protein model that jointly learns from amino acid sequences and three-dimensional structures. Our pretraining strategy combines masked language modeling with diffusion-based denoising, enabling bidirectional sequence-structure learning without requiring MSAs. Trained on both experimentally resolved structures and AlphaFold 2 predictions, FlexProtein captures global folds as well as flexible conformations critical for biological function. Evaluated across diverse tasks spanning interface design, intermolecular interaction prediction, and protein function prediction, FlexProtein establishes new state-of-the-art performance on 12 different tasks, with particularly strong gains in mutation-rich settings where MSA-based methods often struggle.

1 INTRODUCTION

Proteins are fundamental to nearly all biological processes, and modeling their sequences, structures, and functions underpins biomedical and biotechnological advances ranging from enzyme engineering to therapeutic antibody design. Recently, protein foundation models (PFMs) have emerged as a unifying framework that leverages large-scale data and deep learning to capture the principles of protein biology, offering new opportunities for both understanding and design. The development of PFMs has followed two main trajectories. One line of work builds on sequence-only language models (PLMs) such as ESM-2 (Lin et al., 2023) and ProtT5 (Pokharel et al., 2022), which leverage large corpora of protein sequences to learn universal embeddings. These models are broadly applicable and computationally efficient, but the lack of physical relevance, particularly information about three-dimensional geometry, limits their ability to capture the structural basis of protein function. Another line is represented by multiple sequence alignment (MSA) based structure predictors, exemplified by AlphaFold 2/3 (Jumper et al., 2021; Abramson et al., 2024), which exploit evolutionary couplings encoded in MSAs to achieve striking accuracy in structure prediction. Yet, this dependence on homologous sequences introduces sensitivity: when alignments are shallow, sparse, or disrupted by extensive mutation, the predictive signal degrades. As a result, critical scenarios such as antibody CDR loops, intrinsically disordered interfaces, and rapidly evolving pathogens remain inadequately addressed by either paradigm; in these settings, single-sequence models that bypass MSAs and directly model individual sequences provide a more faithful way to capture flexible and highly mutated regions where alignment signals are weak.

We introduce FlexProtein, a 3-billion-parameter pretrained protein model that learns directly from amino acid sequences and large-scale structural corpora, including experimentally resolved structures (Berman et al., 2000) and AlphaFold 2 predicted structures (Varadi et al., 2024). Unlike sequence-only models or predictors that impose a one-way sequence-to-structure mapping, FlexProtein integrates sequence and structure signals from the outset: each residue is represented by a single embedding that combines sequence identity with structural context. The training strategy couples masked language modeling on sequences with diffusion-based denoising on structures, enabling the model to capture bidirectional sequence-structure dependencies and support full-atom structure

generation. To address the variable confidence of predicted structures, we introduce an adaptive loss that selectively weights low-confidence regions, extracting useful signal while avoiding overfitting to unreliable geometry. Previous joint models were designed primarily for structure prediction, but the high memory cost of full-atom representations made the structural component difficult to scale, so most parameters ended up concentrated on the sequence side. FlexProtein overcomes this limitation with a hierarchical modeling strategy that allocates scalable capacity across both sequence and structure, allowing efficient large-scale structural learning alongside sequence semantics.

We systematically evaluate FlexProtein across three broad task families: (i) flexible interface prediction and design, such as antibody/nanobody CDR modeling and peptide binding; (ii) intermolecular interaction prediction, including protein-ligand docking prediction, ligand-induced conformational change, and protein-ligand affinity prediction; and (iii) protein function prediction, such as gene ontology and enzyme activity. Across these categories, FlexProtein achieves state-of-the-art performance, with especially strong improvements in mutation-rich settings where MSA-based methods often struggle. Beyond outperforming existing models, our results highlight the consistent advantages of joint sequence-structure pretraining. The key contributions are:

- Proposing a novel pretraining strategy for FlexProtein that unifies protein structure prediction and design by combining masked language modeling with diffusion-based denoising, thereby learning a bidirectional sequence-structure mapping rather than a one-way sequence-to-structure mapping.
- Introducing a hierarchical modeling strategy that balances scalable capacity across sequence and structure representations, overcoming the memory bottlenecks of full-atom models and enabling structural representations to scale effectively.
- Showing that FlexProtein enables co-design of protein sequence and structure, delivering substantial improvements on flexible and highly mutated regions such as antibody/nanobody CDR loops and peptide-binding interfaces, where MSA-based models struggle.
- Demonstrating consistent gains across 12 tasks spanning flexible interface modeling, intermolecular interactions, and protein function prediction, showing that sequence-structure pretraining transfers broadly beyond protein folding.

2 Related Works

Protein foundation models. PFMs learn transferable protein representations for diverse tasks. Early PFMs were sequence-only language models such as ESM-1b/ESM-2 (Rives et al., 2021; Lin et al., 2023) and ProtT5 (Pokharel et al., 2022), trained on large sequence corpora but limited by the absence of geometric priors. In contrast, structure-centric PFMs such as AlphaFold2/3 (Jumper et al., 2021; Abramson et al., 2024) leverage MSA and templates to achieve high-accuracy folding, yet degrade in highly mutated or low-homology regions. More recently, PFMs have moved toward multimodal, structure-aware pretraining. ESM-3 (Hayes et al., 2025) unifies sequence, structure, and function in a frontier generative model. DPLM-2 (Wang et al., 2025b) extends diffusion protein language models (PLMs) to jointly model both sequences and structures via structure tokenization.

Antibody design. Antibody design methods can be broadly categorized into sequence-based and structure-based approaches. On the sequence side, general-purpose PLMs such as ProtBert (Elnaggar et al., 2021) provide strong baselines for paratope prediction, mutation recovery, and antibody library generation. More specialized pretraining frameworks such as SFM-Protein (He et al., 2024) introduce masked language modeling with pairwise and span-level objectives, showing improved performance on CDR-H3 benchmarks. In addition, graph neural network methods like ABGNN (Gao et al., 2023) and RefineGNN (Jin et al., 2022) attempt to couple sequence embeddings with local structural context, while knowledge-driven frameworks such as RosettaAntibodyDesign (RAbD) (Adolf-Bryfogle et al., 2018) remain widely used in practice. On the structure side, diffusion-based models such as DiffAb (Luo et al., 2022) generate CDR loops conditioned on antigen structures, enabling co-design of sequence and structure, while dyMEAN (Kong et al., 2023b) and MEAN (Kong et al., 2023a) extend this direction with $E(3)$-equivariant architectures for full-atom design. More recently, IgGM (Wang et al., 2025a) expands design capabilities to antibodies and nanobodies by producing antigen-specific complexes. Together, these approaches demonstrate the promise of combining sequence information and structural priors for flexible and functional antibody design.

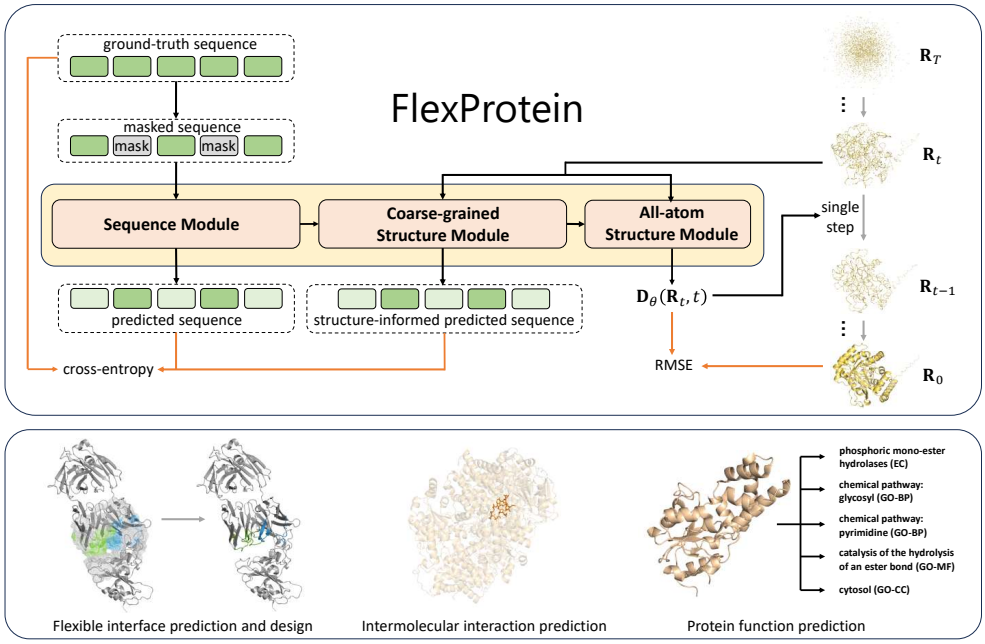

Figure 1: FlexProtein framework. The model architecture consists of three modules: (1) a sequence module that encodes masked protein sequences and ligand topologies, (2) a coarse-grained structure module that encodes residue-level structural information, and (3) an all-atom structure module that refines these representations into chemically consistent coordinates. The framework combines diffusion-based denoising with sequence recovery, enabling joint alignment of sequence, residue, and atomic representations for protein-ligand modeling. Various downstream tasks, including antibody/nanobody design, modeling of intermolecular interactions, and protein function prediction, are supported.

3 METHODS

3.1 DIFFUSION PRETRAINING

We employ diffusion modeling as a generative pretraining objective for protein structures. A structure $\mathbf{R} \in \mathbb{R}^{3N}$ is represented by the 3D coordinates of all heavy atoms. Following Karras et al. (2022) (also adopted by AlphaFold 3), we connect the data distribution $p(\mathbf{R})$ with Gaussian noise p_{src} through a variance-exploding process (Song et al., 2021):

$$\mathbf{R}_t = \mathbf{R}_0 + \sigma_t \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}),$$

where σ_t increases with time t. Sampling amounts to reversing this process, which requires learning the score function $\nabla \log p_t$. We approximate it with a neural network $\mathbf{s}_\theta(\mathbf{R}, t)$, equivalently parameterized as:

$$\mathbf{s}_\theta(\mathbf{R}, t) := \frac{\mathbf{D}_\theta(\mathbf{R}, t) - \mathbf{R}}{\sigma_t^2}, \tag{1}$$

where $\mathbf{D}_\theta(\mathbf{R}, t)$ denotes model output. The effective learning objective amounts to denoising loss:

$$\min_\theta \mathbb{E}_{p(t)} w_t \, \mathbb{E}_{p(\mathbf{R}_0)} \mathbb{E}_{p(\mathbf{R}_t|\mathbf{R}_0)} \left\| \mathbf{D}_\theta(\mathbf{R}_t, t) - \mathbf{R}_0 \right\|^2, \tag{2}$$

with a time-step sampler $p(t)$ and weight w_t. A key challenge is ensuring invariance to rigid-body transformations. We remove translational freedom by centering structures on the center of mass, and enforce rotational invariance by augmenting data with random $\mathrm{SO}(3)$ rotations, instead of relying on heavy $\mathrm{SO}(3)$-equivariant architectures (Köhler et al., 2020) that may also introduce undesired reflection symmetry. We also found that alignment-based objectives (Xu et al., 2022; Abramson et al., 2024) did not improve training stability in our settings but added the risk of improper sampling (Wohlwend et al., 2025). Further details are provided in Appendix C.6.

3.2 ARCHITECTURE

Our architecture is organized in three stages: a sequence module, a coarse-grained structure module, and an all-atom structure module (Fig. 1 and 6). This design balances efficiency and expressivity: coarse-grained modeling captures global protein-ligand organization, while the all-atom stage ensures fine-grained structural accuracy. Table 4 shows the architectural hyperparameters.

Sequence Module. The sequence module jointly embeds protein residues and small-molecule atoms into a unified representation space. For protein residues, we apply a standard Transformer encoder (Appendix B) with rotary position embeddings (Su et al., 2023), focusing purely on sequence-derived semantics. For small molecules, we incorporate 2D topology with a learnable attention bias derived from atom types and bond types to capture chemical identity and connectivity. The combined representations define a residue-atom graph, which is further refined by a pair-feature update module that models residue-residue and residue-atom interactions.

Coarse-grained Structure Module. The coarse-grained structure module employs a Diffusion Transformer (DiT) (Peebles & Xie, 2023) to denoise coordinates at residue level for proteins and atom level for small molecules. Each residue is represented as a coarse structural anchor, while each ligand atom is represented by a position embedding derived from its noised coordinates. The module conditions on embeddings from the sequence module to guide denoising.

All-atom Structure Module. The all-atom structure module employs a DiT where each atom of proteins is represented explicitly. Noised 3D coordinates of all atoms are encoded into position features that serve as token inputs. The coarse-grained outputs are broadcast to all atoms of each residue, providing residue-level guidance as conditional input. To preserve chemical validity, learnable attention biases are added to atom pairs connected by covalent bonds, combining atom-type and bond-type embeddings as additive bias terms in the attention map. This refinement stage allows the model to reconcile global residue-level context with detailed atomic-level interactions, yielding chemically consistent and high-resolution structures.

3.3 STRUCTURE-INFORMED MASKED LANGUAGE MODEL (SIMLM)

Masked language modeling (MLM) (Kenton & Toutanova, 2019; Lin et al., 2023) has proven effective for predicting masked amino acids in protein sequences. In the spirit of unifying sequence and structure, the masked positions should be inferred from correlations within the surrounding sequence and reflect the structural context that these residues possess. To realize this principle, we extend MLM beyond sequence-only inputs by integrating diffusion-based noise into structural representations, yielding a structure-informed masked language model (SIMLM). This formulation couples sequence recovery with structural denoising, thereby reinforcing the mapping between amino acid identity and three-dimensional conformation.

Concretely, we integrate MLM and diffusion through three complementary training modes. **Mode 1 (Sequence-to-Structure):** standard diffusion-based structure reconstruction, where clean sequences condition the generation of noisy structures. **Mode 2 (Coupled Perturbation):** for 15% of residues selected at random, mask the amino acid type and add diffusion noise to their local structures, while leaving all other tokens and structures unperturbed. **Mode 3 (Sequence-Masked Global Perturbation):** randomly select 15% of residues for type masking, while applying diffusion noise to the structures of all residues.

Through these modes, the model alternates between one-way mapping, localized joint perturbation, and global perturbation, which together encourage robust learning of the bidirectional relationship between protein sequences and structures. These allow the model to capture not only sequence-level regularities but also the structural constraints and variability that underlie protein evolution and function. More details are provided in Appendix C.3.

3.4 TRAINING AND SAMPLING

Loss function. Our training objective integrates four complementary components to balance coordinate accuracy, sequence recovery, and structural plausibility. The overall loss is defined as

$$\mathcal{L} = \mathcal{L}_{\text{MSE}} + \mathcal{L}_{\text{MLM}} + \mathcal{L}_{\text{Dist}} + \mathcal{L}_{\text{smooth-lDDT}}.$$

Here, $\mathcal{L}_{\text{MSE}}$ denotes the diffusion pretraining loss function given in Eq. (2), $\mathcal{L}_{\text{MLM}}$ improves sequence-level representation through masked residue prediction, $\mathcal{L}_{\text{Dist}}$ regularizes predicted inter-residue distances to maintain realistic tertiary structure geometry, and $\mathcal{L}_{\text{smooth-lDDT}}$ aligns training with widely used structure quality metrics by emphasizing local geometric accuracy (Appendix C).

Training. Pretraining is organized into two progressive stages. Stage A optimizes all components except $\mathcal{L}_{\text{MLM}}$, training on proteins with up to 384 residues. Deferring MLM at this stage avoids the instability that arises when it is introduced too early, while the residue cap improves efficiency and helps the model prioritize learning core structural regularities. Stage B expands the input length to 768 residues and incorporates $\mathcal{L}_{\text{MLM}}$, enabling stable joint optimization of sequence and structure on larger scales. In both stages, we include a confidence-weighted diffusion loss that scales residue-level contributions by pLDDT-derived sigmoid weights, reducing noise from low-confidence regions while emphasizing reliable structural signals (Appendix C.5).

Sampling. The sampling procedure is the simulation of the reverse process. By leveraging the relation of the denoising model to the score model in Eq. (1), we have:

$$\begin{aligned} \bar{\mathbf{R}}_0 &\sim p_{\text{src}} = \mathcal{N}(\mathbf{0}, \sigma_T^2 \mathbf{I}), \\ \bar{\mathbf{R}}_{\bar{t}+h} &= \bar{\mathbf{R}}_{\bar{t}} + \frac{\mathbf{D}_\theta(\bar{\mathbf{R}}_{\bar{t}}, t) - \bar{\mathbf{R}}_{\bar{t}}}{\sigma_{T-\bar{t}}} (\sigma_{T-\bar{t}} - \sigma_{T-\bar{t}-h}). \end{aligned} \tag{3}$$

We follow similar modifications as used in AlphaFold 3 (Abramson et al., 2024; Karras et al., 2022), but forgo applying the random rotation at each sampling step as orientation alignment is not used in the loss function. Hence, the model learns the correct output orientation relative to the input. The detailed sampling algorithm is presented in Appendix C.7.

4 Experiments

We use entries from the AlphaFold Protein Structure Database (AFDB, CC-BY 4.0 License) (Varadi et al., 2024) and the Protein Data Bank (PDB, CC0 1.0 License) (Berman et al., 2000) released on or before 2021-09-30 for pretraining (Appendix A). For downstream evaluation, we consider three major task families: (i) **Flexible interface prediction and design**, spanning five tasks involving antigen–antibody, antigen–nanobody, and protein–peptide complexes; (ii) **Intermolecular interaction prediction**, including three tasks centered on protein–ligand binding; and (iii) **Protein function prediction**, comprising four tasks focused on functional annotation.

4.1 Flexible interface prediction and design

Biomolecules with flexible binding interfaces are difficult to model and design (Wu et al., 2025). Antibodies, nanobodies, and peptides are key examples, as their functions depend on flexible binding (Wu et al., 2023). This flexibility allows them to target diverse molecules, but also makes structure prediction challenging. To study this problem, we introduce tasks on flexible interface modeling, including antigen-antibody, antigen-nanobody, and protein-peptide complexes. Each task

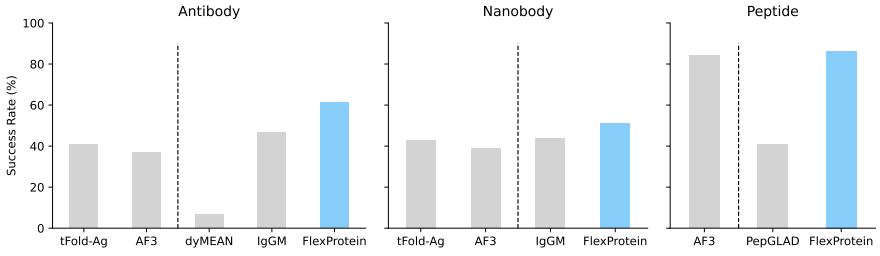

Figure 2: Success rates of structure prediction for antibodies, nanobodies, and peptides. tFold-Ag and AlphaFold 3 take MSA information as input. IgGM, dyMEAN, and FlexProtein leverage antigen structural information. All methods, except AlphaFold 3, additionally incorporate epitope information. Results except FlexProtein are taken from Wang et al. (2025a).

Table 1: Metrics for antibody and nanobody design. Both IgGM and FlexProtein leverage antigen structure and epitope information. All baseline results are taken from Wang et al. (2025a).

Method	Antibody			Nanobody		
	H3-AAR	DockQ	SR	H3-AAR	DockQ	SR
dyMean	0.294	0.079	0.049	-	-	-
diffAb (AF3)	0.226	0.208	0.368	0.156	0.211	0.346
IgGM	0.360	0.246	0.433	0.183	**0.267**	0.415
FlexProtein	**0.414**	**0.273**	**0.460**	**0.218**	0.244	**0.437**

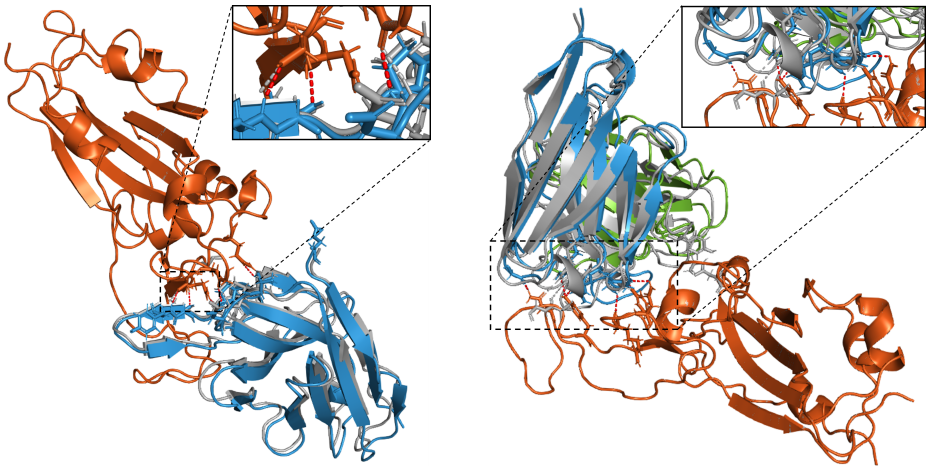

Figure 3: Predicted structures for (left) SARS-CoV-2 RBD with the Re30H02 nanobody (or our design) and (right) SARS-CoV-2 Omicron BA.4 RBD along with its antibody (or our design). Native structures are shown in grey, while those for sequences generated by FlexProtein are in colour. The insets show close-up views of the interaction region, with dashed lines indicating presumed hydrogen bonds. For SARS-CoV-2 RBD (left), the native and designed sequences are structurally similar, with our design manifesting one additional hydrogen bond, whereas for SARS-CoV-2 Omicron BA.4 RBD (right), the predicted structures differ significantly in the interaction region, with our design yielding a greater number of hydrogen bonds. These showcases demonstrate the ability of FlexProtein to generate sequences that are structurally sound.

is defined as: given the sequence or structure of the components, predict the structure of the complex. To avoid overlap, protein chains in the test sets share at most 40% sequence identity with the training data. We focus on two tasks: interface structure prediction and interface design.

Antibody and nanobody interface prediction is a cornerstone of structural immunology, as accurate modeling underpins antibody discovery and therapeutic engineering. Nanobodies can be regarded as single-domain antibodies derived from VHH fragments (Harmsen & De Haard, 2007), allowing both classes to be modeled within a shared framework. For this task, we follow the evaluation procedure of IgGM (Wang et al., 2025a), measuring performance by the success rate (SR) based on the DockQ (Mirabello & Wallner, 2024) score, with a threshold of DockQ ≥ 0.23. Experiments are conducted on the SAb23H2 test set from IgGM, where we compare FlexProtein against the structure prediction models AlphaFold 3 and tFold-Ag (Wu et al., 2024a), as well as the antibody design methods dyMEAN (Kong et al., 2023b) and IgGM. Following the IgGM protocol, we predict antigen-antibody (-nanobody) complex structures given the antigen sequence and antibody (nanobody) sequence. As shown in Fig. 2, FlexProtein achieves success rates of 61.3% for antigen-antibody and 51.1% for antigen-nanobody complexes, yielding absolute improvements of 14.6% and 7.1% over IgGM, respectively. These results demonstrate that FlexProtein effectively models antigen-antibody and antigen-nanobody interactions (Tables 6 and 7).

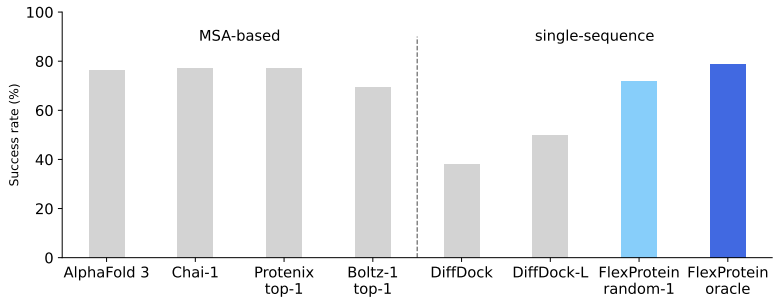

Figure 4: Evaluation of protein-ligand docking on the PoseBusters V1 benchmark. The methods are separated into (left) MSA-based and (right) single-sequence groups. The success rate is defined as the percentage of predictions with pocket-aligned ligand RMSD <2 Å. Apart from DiffDock and DiffDock-L, which predict the ligand pose with the protein structure given, all other methods jointly generate the structure of the protein-ligand complex. Results for AlphaFold 3 are taken from Jumper et al. (2021), Chai-1 from Chai Discovery (2024), DiffDock from Corso et al. (2023), and DiffDock-L from Corso et al. (2024). Results for Protenix, Boltz-1 and FlexProtein were generated locally using a single seed with five generated samples. Other methods report top-1 ranked predictions, while our model does not use a confidence head; thus, we report random single-sample performance, with "oracle" denoting the best prediction among five samples selected against the ground truth structure.

Protein-peptide interface prediction is another important scenario, as peptides often act as recognition motifs or regulators for diverse cellular processes. We use FoldBench (Xu et al., 2025) as the benchmark. We follow the FoldBench evaluation protocol, also reporting the success rate based on DockQ. FlexProtein is compared with the structure prediction model AlphaFold 3 and the peptide design method PepGLAD (Kong et al., 2025). As shown in Fig. 2, FlexProtein achieves an SR of 91.4%, exceeding AlphaFold 3 and PepGLAD by 7.0% and 10.2%, respectively. These findings suggest that FlexProtein generalizes well to flexible peptide-protein binding scenarios (Table 8).

Antibody and nanobody design is a key challenge in developing novel binders for therapeutic and diagnostic applications. In this task, the input is the antigen sequence and structure, and the objective is to design sequences for all antibody/nanobody CDR regions while jointly generating the full complex structure. We evaluate FlexProtein on the SAb23H2 test set from IgGM, following the same protocol. Performance is measured by amino acid recovery (AAR), DockQ, and success rate (SR, defined as the proportion of samples which have DockQ ≥ 0.23) relative to wild-type complexes. Baselines include dyMEAN, diffAb (Luo et al., 2022), and IgGM. As shown in Table 1, FlexProtein achieves 41.4% AAR and 46.0% SR for antibody design, surpassing IgGM and setting a new state-of-the-art. For nanobody design, FlexProtein also slightly outperforms IgGM, with higher AAR (21.8%) and SR (43.7%). In addition, FlexProtein supports user-specified CDR lengths, enabling flexible design. Figure 3 illustrates some representative designs: our framework generates up to six CDRs simultaneously, with AAR reported specifically for the highly flexible CDR-H3 region, which is also the most challenging to design. Detailed results are provided in Appendix D. Together, these results show that FlexProtein can generate realistic CDR sequences while maintaining structural fidelity to wild-type complexes.

4.2 Intermolecular interaction prediction

Protein-ligand interactions are fundamental in understanding protein conformational changes, binding affinities, and diverse biological functions. The accurate prediction of these interactions is thus crucial for elucidating molecular mechanisms and accelerating drug discovery. We evaluate FlexProtein on three downstream tasks: protein-ligand docking prediction, ligand-induced conformational changes and protein-ligand binding affinity, which sharing a common foundation in modeling protein-ligand complexes, while emphasizing different aspects of interaction.

Protein-ligand docking prediction is a key task for modeling intermolecular interactions with broad implications for life sciences and drug discovery. We follow the AlphaFold 3 protocol on the

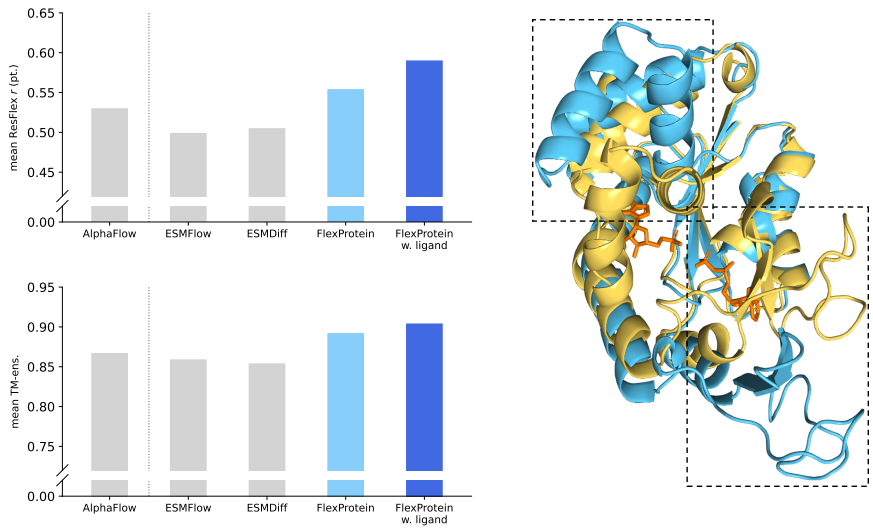

Figure 5: Evaluation of ligand-induced conformational change prediction. The left panels show the per-target mean correlations (top) and mean ensemble TM-scores (bottom). In the right panel, an overlay of the predicted apo (blue, PDB 4AKE) and holo (yellow, PDB 2ECK) structures of adenylate kinase is presented, illustrating the conformational changes (highlighted by the dashed boxed regions) induced by the presence of AMP and ADP molecules (orange). FlexProtein is able to accurately predict both states, with TM-scores of 0.985 and 0.984, respectively.

PoseBusters V1 benchmark (Buttenschoen et al., 2024) of 428 protein-ligand complexes, comparing against MSA-based models (AlphaFold3, Chai-1 (Chai Discovery, 2024), Protenix (ByteDance AML AI4Science Team et al., 2025), Boltz-1 (Wohlwend et al., 2025)) and single-sequence-based models (DiffDock (Corso et al., 2023), DiffDock-L (Corso et al., 2024)). Unlike previous single-sequence-based methods, which assume a fixed protein structure and only generate ligand poses, FlexProtein jointly predicts protein-ligand structures directly from sequence like AlphaFold 3. As shown in Fig. 4, FlexProtein achieves 71.82% in the random-1 regime and 78.70% under oracle selection, surpassing all single-sequence baselines by substantial margin and reaching parity with MSA-based approaches (Table 11).

Ligand-induced conformational change prediction is key to understanding how proteins adapt upon ligand binding. We follow the ESMDiff protocol (Lu et al., 2025) and use the Apo/Holo dataset (Saldaño et al., 2022) to evaluate FlexProtein with ensemble TM-score (TM-ens) and residue flexibility correlations (ResFlex r) at both global and per-target levels. Baselines include AlphaFlow (MSA-based), ESMFlow (Jing et al., 2024) (sequence-based), and ESMDiff (structure–language). Unlike these methods, FlexProtein can model protein-ligand complexes in two modes: (i) protein-only (5 samples) and (ii) mixed (3 apo + 2 holo samples) with ligand guidance. Using a zero-shot pretrained checkpoint, FlexProtein achieves a TM-ens score of 0.889 (improving upon ESMDiff by 0.038) and stronger flexibility correlations. Adding ligands further improves TM-ens by 0.012, showing the importance of ligand context. Figure 5 and Table 12 summarize the results.

Protein-ligand binding affinity prediction is a cornerstone of drug discovery, enabling efficient prioritization of candidate compounds for therapeutic targets. Traditional high-throughput screening is costly and limited in scope, motivating computational approaches that estimate binding affinities directly from protein-ligand structures. We evaluate FlexProtein on the CASF-2016 benchmark (Su et al., 2018), using the standard metrics of root mean square error (RMSE) and Pearson's correlation coefficient (R). Comparisons are made against state-of-the-art baselines SIGN (Li et al., 2021), GLANT (Li et al., 2023), and SPIN (Choi et al., 2024). As shown in Table 2, FlexProtein achieves the best performance on both criteria. These results highlight the value of pretrained embeddings derived from joint protein-ligand structures as a strong basis for accurate affinity prediction.

Table 2: Evaluation of protein-ligand binding affinity prediction on CASF-2016.

Method	RMSE (↓)	R (↑)
SIGN	1.316	0.797
GIANT	1.269	0.814
SPIN	1.258	0.826
FlexProtein	**1.150**	**0.848**

4.3 PROTEIN FUNCTION PREDICTION

Protein function prediction is central to characterizing novel proteins, understanding disease, and guiding therapeutic discovery. We evaluate this by finetuning FlexProtein on Gene Ontology (GO) (Ashburner et al., 2000) and Enzyme Commission (EC) (Bairoch, 2000) annotation tasks using the DeepFRI (Gligorijević et al., 2021) setup. Baselines include sequence-only models (ESM-2-3B, SFM-Protein-3B (He et al., 2024)) and a sequence–structure hybrid (ESM-GearNet (Zhang et al., 2023)). Unlike these methods, FlexProtein jointly predicts structures and embeddings without external structural input. Performance is measured by maximum F_1 score.

Table 3: F_1 for the Enzyme Commission (EC) and Gene Ontology (GO) tasks. The GO task is comprised of three independent sub-tasks, namely biological process (BP), molecular function (MF), and cellular component (CC).

Method	EC	GO-BP	GO-MF	GO-CC
ESM-2-3B	0.863	0.476	0.659	0.497
SFM-Protein-3B	0.869	0.495	0.673	0.510
ESM-GearNet	0.890	0.488	0.681	0.464
FlexProtein	**0.891**	**0.539**	**0.694**	**0.560**

EC number prediction provides a controlled benchmark for catalytic function annotation, formulated as a binary classification task. As shown in Table 3, FlexProtein achieves an F_1 score of 0.891, slightly exceeding ESM-GearNet and clearly outperforming sequence-only baselines (ESM-2 and SFM-Protein). These results highlight the importance of structural information in accurately capturing enzymatic function.

GO term prediction. GO term prediction evaluates protein function across biological processes (BP), molecular functions (MF), and cellular components (CC), each framed as an independent multi-label classification task, consistent with the EC setup. As shown in Table 3, FlexProtein achieves F1 scores of 0.539 (BP), 0.694 (MF), and 0.560 (CC), outperforming ESM-GearNet by 0.051, 0.013, and 0.096, respectively. These gains indicate that FlexProtein provides more informative representations for finetuning, enabling more accurate functional annotation across ontologies.

5 CONCLUSION

We introduce FlexProtein, a pretrained protein foundation model that integrates both sequence and structural information into a unified framework. Unlike prior approaches that impose a one-way sequence-to-structure mapping, FlexProtein implements joint training via masked language modeling and diffusion-based denoising, enabling bidirectional sequence-structure representations that support both prediction and design. Extensive evaluations across a diverse set of downstream tasks spanning antibody/nanobody and peptide interface modeling, ligand-induced conformational change, protein-ligand binding affinity, and functional annotation demonstrate strong and consistent gains, with especially notable improvements on flexible and mutation-rich interfaces where existing methods struggle. These results highlight the effectiveness of joint sequence-structure pretraining and show that its benefits extend broadly beyond protein folding, establishing FlexProtein as a general-purpose foundation model for protein science.

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*, 630(8016):493–500, 2024.

Jared Adolf-Bryfogle, Oleg Kalyuzhniy, Maureen Kubitz, Brian D. Weitzner, Xiaozhen Hu, Yasunori Adachi, William R. Schief, and Roland L. Dunbrack Jr. RosettaAntibodyDesign (RAbD): A general framework for computational antibody design. *PLOS Computational Biology*, 14(4): e1006112, 2018.

Brian D.O. Anderson. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3):313–326, 1982.

Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, et al. Gene Ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29, 2000.

Amos Bairoch. The ENZYME database in 2000. *Nucleic acids research*, 28(1):304–305, 2000.

Helen M Berman, John Westbrook, Zukang Feng, Gary Gilliland, Talapady N Bhat, Helge Weissig, Ilya N Shindyalov, and Philip E Bourne. The protein data bank. *Nucleic acids research*, 28(1): 235–242, 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*, 15(9):3130–3139, 2024.

ByteDance AML AI4Science Team, Xinshi Chen, Yuxuan Zhang, Chan Lu, Wenzhi Ma, Jiaqi Guan, Chengyue Gong, Jincai Yang, Hanyu Zhang, Ke Zhang, et al. Protenix-advancing structure prediction through a comprehensive AlphaFold3 reproduction. *BioRxiv*, pp. 2025–01, 2025.

Chai Discovery. Chai-1: Decoding the molecular interactions of life. *bioRxiv*, 2024.

Seungyeon Choi, Sangmin Seo, and Sanghyun Park. SPIN: SE(3)-invariant physics informed network for binding affinity prediction. In *27th European Conference on Artificial Intelligence (ECAI 2024)*, 2024.

Gabriele Corso, Hannes Stärk, Bowen Jing, Regina Barzilay, and Tommi Jaakkola. DiffDock: Diffusion steps, twists, and turns for molecular docking. In *International Conference on Learning Representations (ICLR)*, 2023.

Gabriele Corso, Arthur Deng, Nicholas Polizzi, Regina Barzilay, and Tommi S. Jaakkola. Deep confident steps to new pockets: Strategies for docking generalization. In *The Twelfth International Conference on Learning Representations*, 2024.

James Dunbar, Konrad Krawczyk, Jinwoo Leem, Terry Baker, Angelika Fuchs, Guy Georges, Jiye Shi, and Charlotte M. Deane. SAbDab: the structural antibody database. *Nucleic Acids Research*, 42(D1):D1140–D1146, 11 2013.

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.

Kaiyuan Gao, Lijun Wu, Jinhua Zhu, Tianbo Peng, Yingce Xia, Liang He, Shufang Xie, Tao Qin, Haiguang Liu, Kun He, and Tie-Yan Liu. Pre-training antibody language models for antigen-specific computational antibody design. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 506–517. ACM, 2023.

Vladimir Gligorijević, P. Douglas Renfrew, Tomasz Kosciolek, Julia Koehler Leman, Daniel Berenberg, Tommi Vatanen, Chris Chandler, Bryn C. Taylor, Ian M. Fisk, Hera Vlamakis, Ramnik J. Xavier, Rob Knight, Kyunghyun Cho, and Richard Bonneau. Structure-based protein function prediction using graph convolutional networks. *Nature Communications*, 12(1):3168, May 2021.

Michiel M Harmsen and Hans J De Haard. Properties, production, and applications of camelid single-domain antibody fragments. *Applied microbiology and biotechnology*, 77(1):13–22, 2007.

Thomas Hayes, Roshan Rao, Halil Akin, Nicholas J Sofroniew, Deniz Oktay, Zeming Lin, Robert Verkuil, Vincent Q Tran, Jonathan Deaton, Marius Wiggert, et al. Simulating 500 million years of evolution with a language model. *Science*, 387(6736):850–858, 2025.

Liang He, Peiran Jin, Yaosen Min, Shufang Xie, Lijun Wu, Tao Qin, Xiaozhuan Liang, Kaiyuan Gao, Yuliang Jiang, and Tie-Yan Liu. SFM-Protein: Integrative co-evolutionary pre-training for advanced protein sequence representation. *arXiv preprint arXiv:2410.24022*, 2024.

Wengong Jin, Jeremy Wohlwend, Regina Barzilay, and Tommi S. Jaakkola. Iterative refinement graph neural network for antibody sequence-structure co-design. In *International Conference on Learning Representations*, 2022.

Bowen Jing, Bonnie Berger, and Tommi Jaakkola. AlphaFold meets flow matching for generating protein ensembles. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 22277–22303. PMLR, 21–27 Jul 2024.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with AlphaFold. *nature*, 596(7873):583–589, 2021.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35:26565–26577, 2022.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, pp. 2. Minneapolis, Minnesota, 2019.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations (ICLR2015)*, 2015.

Jonas Köhler, Leon Klein, and Frank Noé. Equivariant flows: exact likelihood generative learning for symmetric densities. In *International conference on machine learning*, pp. 5361–5370. PMLR, 2020.

Xiangzhe Kong, Wenbing Huang, and Yang Liu. Conditional antibody design as 3d equivariant graph translation. In *The Eleventh International Conference on Learning Representations*, 2023a.

Xiangzhe Kong, Wenbing Huang, and Yang Liu. End-to-end full-atom antibody design. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 17409–17429. PMLR, 23–29 Jul 2023b.

Xiangzhe Kong, Yinjun Jia, Wenbing Huang, and Yang Liu. Full-atom peptide design with geometric latent diffusion. *Advances in Neural Information Processing Systems*, 37:74808–74839, 2025.

Shuangli Li, Jingbo Zhou, Tong Xu, Liang Huang, Fan Wang, Haoyi Xiong, Weili Huang, Dejing Dou, and Hui Xiong. Structure-aware interactive graph neural networks for the prediction of protein-ligand binding affinity. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, pp. 975–985, 2021.

Shuangli Li, Jingbo Zhou, Tong Xu, Liang Huang, Fan Wang, Haoyi Xiong, Weili Huang, Dejing Dou, and Hui Xiong. GIaNt: Protein-ligand binding affinity prediction via geometry-aware interactive graph neural network. *IEEE Transactions on Knowledge and Data Engineering*, 36(5): 1991–2008, 2023.

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*, 379(6637):1123–1130, 2023.

Zhihai Liu, Minyi Su, Li Han, Jie Liu, Qifan Yang, Yan Li, and Renxiao Wang. Forging the basis for developing protein–ligand interaction scoring functions. *Accounts of chemical research*, 50 (2):302–309, 2017.

Jiarui Lu, Xiaoyin Chen, Stephen Zhewen Lu, Chence Shi, Hongyu Guo, Yoshua Bengio, and Jian Tang. Structure language models for protein conformation generation. In *The Thirteenth International Conference on Learning Representations*, 2025.

Shitong Luo, Yufeng Su, Xingang Peng, Sheng Wang, Jian Peng, and Jianzhu Ma. Antigen-specific antibody design and optimization with diffusion-based generative models for protein structures. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 9754–9767. Curran Associates, Inc., 2022.

Claudio Mirabello and Björn Wallner. DockQ v2: improved automatic quality measure for protein multimers, nucleic acids, and small molecules. *Bioinformatics*, 40(10):btae586, 2024.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4195–4205, 2023.

Suresh Pokharel, Pawel Pratyush, Michael Heinzinger, Robert H Newman, and Dukka B Kc. Improving protein succinylation sites prediction using embeddings from protein language model. *Scientific reports*, 12(1):16933, 2022.

Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C Lawrence Zitnick, Jerry Ma, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15):e2016239118, 2021.

Tadeo Saldaño, Nahuel Escobedo, Julia Marchetti, Diego Javier Zea, Juan Mac Donagh, Ana Julia Velez Rueda, Eduardo Gonik, Agustina García Melani, Julieta Novomisky Nechcoff, Martín N Salas, et al. Impact of protein conformational diversity on alphafold predictions. *Bioinformatics*, 38(10):2742–2748, 2022.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding, 2023.

Minyi Su, Qifan Yang, Yu Du, Guoqin Feng, Zhihai Liu, Yan Li, and Renxiao Wang. Comparative assessment of scoring functions: the CASF-2016 update. *Journal of chemical information and modeling*, 59(2):895–913, 2018.

Mihaly Varadi, Damian Bertoni, Paulyna Magana, Urmila Paramval, Ivanna Pidruchna, Malarvizhi Radhakrishnan, Maxim Tsenkov, Sreenath Nair, Milot Mirdita, Jingi Yeo, et al. AlphaFold Protein Structure Database in 2024: providing structure coverage for over 214 million protein sequences. *Nucleic acids research*, 52(D1):D368–D375, 2024.

Rubo Wang, Fandi Wu, Xingyu Gao, Jiaxiang Wu, Peilin Zhao, and Jianhua Yao. IgGM: A generative model for functional antibody and nanobody design. In *The Thirteenth International Conference on Learning Representations*, 2025a.

Xinyou Wang, Zaixiang Zheng, Fei YE, Dongyu Xue, Shujian Huang, and Quanquan Gu. DPLM-2: A multimodal diffusion protein language model. In *The Thirteenth International Conference on Learning Representations*, 2025b.

Jeremy Wohlwend, Gabriele Corso, Saro Passaro, Noah Getz, Mateo Reveiz, Ken Leidal, Wojtek Swiderski, Liam Atkinson, Tally Portnoi, Itamar Chinn, et al. Boltz-1 democratizing biomolecular interaction modeling. *BioRxiv*, pp. 2024–11, 2025.

Fandi Wu, Yu Zhao, Jiaxiang Wu, Biaobin Jiang, Bing He, Longkai Huang, Chenchen Qin, Fan Yang, Ningqiao Huang, Yang Xiao, et al. Fast and accurate modeling and design of antibody-antigen complex using tFold. *bioRxiv*, pp. 2024–02, 2024a.

Kejia Wu, Hua Bai, Ya-Ting Chang, Rachel Redler, Kerrie E McNally, William Sheffler, TJ Brunette, Derrick R Hicks, Tomos E Morgan, Tim J Stevens, et al. De novo design of modular peptide-binding proteins by superhelical matching. *Nature*, 616(7957):581–589, 2023.

Kejia Wu, Hanlun Jiang, Derrick R Hicks, Caixuan Liu, Edin Muratspahić, Theresa A Ramelot, Yuexuan Liu, Kerrie McNally, Sebastian Kenny, Andrei Mihut, et al. Design of intrinsically disordered region binding proteins. *Science*, 389(6757):eadr8063, 2025.

Kevin E Wu, Kevin K Yang, Rianne van den Berg, Sarah Alamdari, James Y Zou, Alex X Lu, and Ava P Amini. Protein structure generation via folding diffusion. *Nature communications*, 15(1): 1059, 2024b.

Minkai Xu, Lantao Yu, Yang Song, Chence Shi, Stefano Ermon, and Jian Tang. GeoDiff: A geometric diffusion model for molecular conformation generation. In *International Conference on Learning Representations*, 2022.

Sheng Xu, Qiantai Feng, Lifeng Qiao, Hao Wu, Tao Shen, Yu Cheng, Shuangjia Zheng, and Siqi Sun. FoldBench: An all-atom benchmark for biomolecular structure prediction. *bioRxiv*, pp. 2025–05, 2025.

Zuobai Zhang, Chuanrui Wang, Minghao Xu, Vijil Chenthamarakshan, Aurélie Lozano, Payel Das, and Jian Tang. A systematic study of joint representation learning on protein sequences and structures. *arXiv preprint arXiv:2303.06275*, 2023.

A DATA

The pretraining dataset is constructed from two primary sources: **AFDB** and **PDB**.

The AlphaFold Protein Structure Database (AFDB), released by Google DeepMind and EMBL-EBI, contains over 200 million predicted structures spanning nearly the entire **UniProt_2021_04** release. To reduce redundancy, we cluster sequences at 90% identity and retain only entries with a global pLDDT score greater than 50, discarding low-confidence structures. This yields approximately 78 million AFDB samples.

For experimentally resolved structures in Protein Data Bank (PDB), we use **PDB_20210930**, adopting the same cutoff date as AlphaFold 3. Following their filtering protocol, we exclude structures with more than 300 chains, resolution worse than 9 Å,or fewer than 4 residues. After filtering, we obtain roughly 181 thousand PDB samples.

In total, our pretraining corpus comprises more than 78 million protein structures.

The datasets used for each downstream task are detailed in Appendix D.

B ARCHITECTURE

Parameters of architecture is shown in Table 4 and details of model architecture is shown in Figure 6.

Table 4: Key architectural hyperparameters of FlexProtein.

Component	Hyperparameter	Value
Sequence Module	Number of layers	32
	Hidden size	2048
	FFN dimension	8192
	Attention heads	32
Coarse-grained structure module	Number of layers	16
	Hidden dimension	2048
	FFN dimension	8192
	Attention heads	32
All-atom structure module	Number of layers	8
	Embedding dimension	256
	FFN dimension	256
	Attention heads	4

C PRE-TRAINING

C.1 CONFIDENCE-WEIGHTED DIFFUSION LOSS.

To incorporate structural reliability, we scale the diffusion MSE loss using residue-level pLDDT scores with a sigmoid weighting function. Specifically, residues with very low confidence (pLDDT $\lesssim 60$) are down-weighted toward zero, while those with very high confidence (pLDDT $\gtrsim 80$) receive weights close to one. The transition between these regimes is smoothed using a sigmoid:

$$w = \sigma\left(\beta \cdot \frac{\text{pLDDT} - 70}{10}\right),$$

where $\sigma(\cdot)$ is the logistic sigmoid and β controls the steepness of the curve. In practice, we set $\beta = 5$ such that weights are near zero at pLDDT $= 60$ and near one at pLDDT $= 80$. This formulation avoids hard thresholds while ensuring that uncertain structural regions contribute less to the optimization, and high-confidence regions dominate the learning signal.

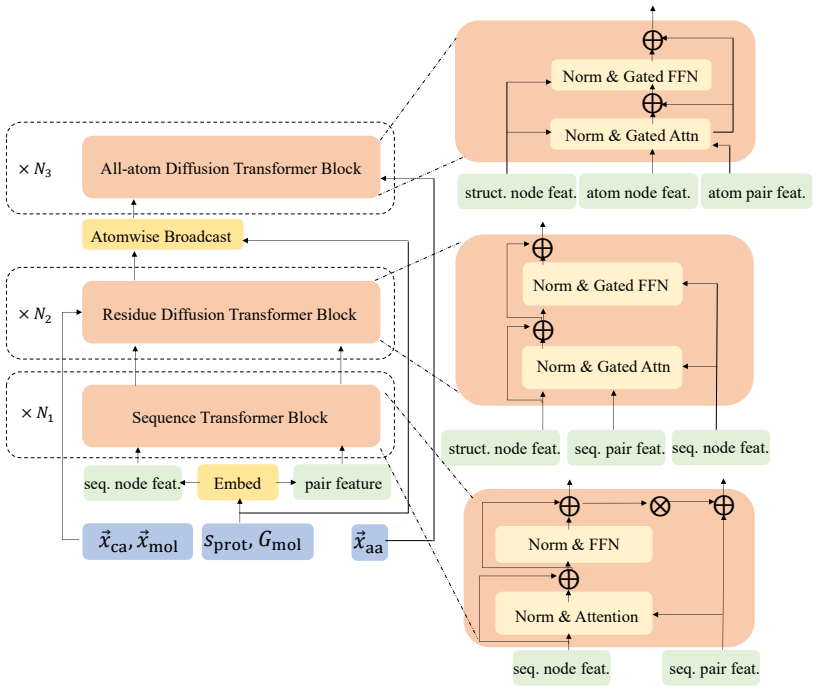

Figure 6: Detailed Model Architecture of FlexProtein.

C.2 Inter-residue distance loss.

We apply $\mathcal{L}_{\text{Dist}}$ on top of the sequence module to regularize predicted inter-residue distances. Specifically, the sequence encoder outputs residue-level embeddings, which are combined through an outer product to form pairwise features. A lightweight MLP head then predicts the C_α-C_α distance for each residue pair. The loss penalizes deviations between predicted and ground-truth distances, encouraging the encoder to capture geometric constraints directly at the sequence-pair feature level. This design provides the model with explicit supervision on tertiary structure geometry while avoiding direct reliance on coordinate-level regression.

C.3 Structure-informed Masked Language Modeling (SIMLM) loss.

We design a structure-informed masked language modeling loss to align sequence and structure representations. Only protein sequences (FASTA) are masked, following the BERT-style policy: 15% of residues are selected for corruption, with 80% replaced by `[MASK]`, 10% replaced by a random amino acid, and 10% left unchanged. For each masked residue i with ground-truth identity y_i, we compute hidden features from both the sequence encoder f_i^{seq} and the coarse-grained structure encoder f_i^{struct} (e.g., based on C_α geometry). Two independent prediction heads are applied: one maps f_i^{seq} to a distribution $p_\theta^{(\text{seq})}(y_i)$ and the other maps f_i^{struct} to $p_\theta^{(\text{struct})}(y_i)$. The loss averages the negative log-likelihoods from both heads:

$$\mathcal{L}_{\text{S-MLM}} = -\frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} \Big[\log p_\theta^{(\text{seq})}(y_i \mid f_i^{\text{seq}}) + \log p_\theta^{(\text{struct})}(y_i \mid f_i^{\text{struct}}) \Big],$$

where $\mathcal{M}$ is the set of masked positions. This formulation encourages both the sequence and structure pathways to retain predictive signal for residue identity, thereby improving cross-modal consistency. In practice, we interleave the three perturbation modes (Mode 1, Mode 2, and Mode 3) during training with a ratio of $6:2:2$, balancing standard sequence-to-structure generation with increasingly challenging coupled and global perturbations.

C.4 SMOOTH-LDDT LOSS.

Following AlphaFold2 (Jumper et al., 2021), we compute the smooth local distance difference test (lDDT) loss to assess local structural accuracy. The smooth lDDT metric measures the agreement of predicted inter-residue distances with the ground truth in a differentiable manner. Specifically, for each residue i, we evaluate all neighboring residues j within a cutoff radius (typically 15 Å). For each pair (i, j), the absolute deviation of the predicted C_α-C_α distance from the reference is mapped to a soft score using a piecewise linear function with thresholds at 0.5, 1, 2, and 4 Å. The residue-wise scores are averaged across neighbors and then across residues to produce the overall smooth lDDT. In training, we only use C_α atoms to compute this loss, consistent with AlphaFold2. The resulting value serves both as a differentiable accuracy proxy and as a regularizer encouraging the model to capture local geometric consistency.

C.5 TRAINING RECIPE.

We adopt a two-stage pretraining strategy (see Table 5). Stage A focuses on diffusion-based denoising with proteins of up to 384 residues. Training uses the Adam optimizer (Kingma & Ba, 2015) in bfloat16 mixed precision with a batch size of 4,096 on 128 A100 GPUs for 200k steps and a learning rate of 1×10^{-4}. This stage builds the core ability to reconstruct clean structures from noisy inputs while incorporating structural regularization via distance and smooth-lDDT losses. Stage B extends the maximum protein length to 768 residues and adds the masked language modeling (MLM) objective. Training uses a batch size of 2,048 on the same hardware for 100k steps with a learning rate of 6×10^{-5}. This stage enables the model to handle larger proteins and integrate sequence-level supervision, while continuing to optimize diffusion, distance, and smooth-lDDT objectives.

Table 5: Two-stage pretraining configuration.

	Stage A	**Stage B**
Max residues	384	768
Losses	$\mathcal{L}_{\text{MSE}} + \mathcal{L}_{\text{Dist}} + \mathcal{L}_{\text{smooth-lDDT}}$	$\mathcal{L}_{\text{MSE}} + \mathcal{L}_{\text{Dist}} + \mathcal{L}_{\text{smooth-lDDT}} + \mathcal{L}_{\text{MLM}}$
Batch size	4,096	2,048
Steps	200k	100k
Learning rate	1×10^{-4}	6×10^{-5}

C.6 DIFFUSION TRAINING DETAILS

We provide here the complete derivations and formulation details omitted from the main text.

Forward process. Diffusion-based generative modeling aims to approximate a target distribution $p(\mathbf{R})$ by connecting it to a tractable source distribution p_{src}. We represent a protein structure $\mathbf{R} \in \mathbb{R}^{3N}$ by the 3D coordinates of all heavy atoms. The forward noising process is defined as

$$\mathbf{R}_t = \mathbf{R}_0 + \sigma_t \boldsymbol{\epsilon}_t, \quad \boldsymbol{\epsilon}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \quad t \in [0, T],$$

with σ_t monotone increasing in t. For sufficiently large σ_T, $\mathbf{R}_T$ approximates a Gaussian distribution $p_{\text{src}} = \mathcal{N}(\mathbf{0}, \sigma_T^2 \mathbf{I})$.

This corresponds to the SDE

$$\mathrm{d}\mathbf{R}_t = \sqrt{(\sigma_t^2)'}\, \mathrm{d}\mathbf{w}_t,$$

where $\mathbf{w}_t$ is a Wiener process on $\mathbb{R}^{3N}$.

Reverse process. By stochastic process theory (Anderson, 1982), one can recover $p(\mathbf{R})$ by simulating the reverse diffusion process. A deterministic equivalent is given by the probability-flow ODE (Song et al., 2021):

$$\mathrm{d}\bar{\mathbf{R}}_{\bar{t}} = \tfrac{1}{2}(\sigma_t^2)'|_{t=T-\bar{t}} \nabla \log p_{T-\bar{t}}(\bar{\mathbf{R}}_{\bar{t}})\, \mathrm{d}\bar{t},$$

where $\bar{t}$ denotes reversed time and $\bar{\mathbf{R}}_{\bar{t}}$ denotes the reversed sample trajectory.

Score estimation. The only unknown term is $\nabla \log p_t$, which we approximate with $\mathbf{s}_\theta(\mathbf{R}, t)$. Minimizing the score-matching objective

$$\mathbb{E}_{p_t(\mathbf{R}_t)} \left\| \mathbf{s}_\theta(\mathbf{R}_t, t) - \nabla \log p_t(\mathbf{R}_t) \right\|^2$$

is equivalent to a denoising objective with conditional distribution $p(\mathbf{R}_t \mid \mathbf{R}_0) = \mathcal{N}(\mathbf{R}_t \mid \mathbf{R}_0, \sigma_t^2 \mathbf{I})$:

$$\min_\theta \, \mathbb{E}_{p(t)} w_t \, \mathbb{E}_{p(\mathbf{R}_0)} \mathbb{E}_{p(\mathbf{R}_t \mid \mathbf{R}_0)} \left\| \mathbf{D}_\theta(\mathbf{R}_t, t) - \mathbf{R}_0 \right\|^2,$$

where we use the parameterization

$$\mathbf{s}_\theta(\mathbf{R}, t) := \frac{\mathbf{D}_\theta(\mathbf{R}, t) - \mathbf{R}}{\sigma_t^2}.$$

Intuitively, the network predicts the clean structure $\mathbf{R}_0$ from its noisy version $\mathbf{R}_t$, hence the name "denoising model."[1]

Rigid-body invariances. Protein structures are equivalent up to rigid-body transformations. Translations are removed by centering at the center of mass. Rotational invariance is harder: while $\mathrm{SO}(3)$-equivariant networks (Köhler et al., 2020) can guarantee invariance, they often require heavy operations and introduce reflection symmetry. We instead use a standard architecture and provide rotational invariance information via random $\mathrm{SO}(3)$ data augmentation. Some works apply explicit rotational alignment in the loss (Xu et al., 2022; Abramson et al., 2024), but such alignment lacks a consistent orientation correspondence and complicates sampling (Wohlwend et al., 2025). In our experiments, the plain denoising objective already yielded stable and effective training, so we removed the alignment operation in the loss.

Algorithm 1 Sampling procedure.

Require: A trained diffusion model in denoising form $\mathbf{D}_\theta(\mathbf{R}, t)$ under the noise schedule choice $\sigma_t = t$, sampling time schedule $0 = \bar{t}_0 < \bar{t}_1 < \cdots < \bar{t}_N = T$, recursion ratio γ_{recur}, recursion threshold γ_{min}, noise scale λ.

1: Initialize $\bar{\mathbf{R}}_0 \sim p_{\text{src}} := \mathcal{N}(\mathbf{0}, T^2 \mathbf{I})$;
2: **for** $i = 0$ to $N - 1$ **do**
3: Center $\bar{\mathbf{R}}_i$ to its center of mass;
4: $\gamma = \gamma_{\text{recur}}$ if $T - \bar{t}_i > \gamma_{\text{min}}$ else 0;
5: $\hat{\bar{t}}_i = (1 + \gamma)\bar{t}_i - \gamma T$;
6: $\bar{\mathbf{R}}_{\hat{\bar{t}}_i} = \bar{\mathbf{R}}_{\bar{t}_i} + \lambda \sqrt{(T - \hat{\bar{t}}_i)^2 - (T - \bar{t}_i)^2}\, \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$;
7: $\bar{\mathbf{R}}_{\bar{t}_{i+1}} = \bar{\mathbf{R}}_{\hat{\bar{t}}_i} + \frac{\mathbf{D}_\theta(\bar{\mathbf{R}}_{\hat{\bar{t}}_i}, T - \hat{\bar{t}}_i) - \bar{\mathbf{R}}_{\hat{\bar{t}}_i}}{T - \hat{\bar{t}}_i} (\hat{\bar{t}}_i - \bar{t}_{i+1})$;
8: **end for**
9: **return** $\bar{\mathbf{R}}_{\bar{t}_N}$.

C.7 Sampling procedure

From the reverse sampling formulation in Eq. (3), what essentially controls the progression of the diffusion process is the discretization of σ_t. A convenient choice is thus to let $\sigma_t = t$ (Karras et al., 2022). The sampling process is then specified by a discretization of the reverse time $0 = \bar{t}_0 < \bar{t}_1 < \cdots < \bar{t}_N = T$, where N is the number of discretization steps. Following (Karras et al., 2022) (which is also adopted in Alphafold 3 (Abramson et al., 2024)), in each step, the update starts not directly from the current time step $\bar{t}_i$. Instead, the clock is first recurred back to $\hat{\bar{t}}_i := (1+\gamma)\bar{t}_i - \gamma T$ (which comes from increasing the forward time by $(1 + \gamma)$, *i.e.*, $T - \hat{\bar{t}}_i = (1 + \gamma)(T - \bar{t}_i)$) with a more noisy state, which can be implemented by simulating the forward process from $T - \bar{t}_i$ to $T - \hat{\bar{t}}_i$ as $\bar{\mathbf{R}}_{\hat{\bar{t}}_i} = \bar{\mathbf{R}}_{\bar{t}_i} + \sqrt{(T - \hat{\bar{t}}_i)^2 - (T - \bar{t}_i)^2}\, \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. The simulation then proceeds by an update from $\hat{\bar{t}}_i$ to $\bar{t}_{i+1}$ following Eq. (3). In contrast to the sampling process by Alphafold 3, we do not need a random rotation in each step as we do not use rotational alignment in training. The complete procedure is presented in Alg. C.6.

[1] Recovering the exact $\mathbf{R}_0$ is impossible due to information loss; the model in fact predicts $\mathbb{E}[\mathbf{R}_0 \mid \mathbf{R}_t]$.

D EXPERIMENTS DETAILS

Downstream tasks are application-specific benchmarks designed to evaluate how effectively a pretrained foundation model can be adapted to solve targeted scientific problems. While pretraining provides the model with general sequence-structure representations, downstream tasks assess its transferability to practical domains such as protein design, intermolecular interaction prediction, and functional annotation. These tasks typically involve fine-tuning the model on smaller, curated datasets and comparing its performance against established baselines. By systematically evaluating across diverse downstream tasks, we demonstrate not only the generality of the pretrained model, but also its ability to capture biologically meaningful features that enable real-world scientific discovery. Details of the fine-tuning procedure, encompassing dataset partitioning, optimization strategies, and evaluation metrics, are provided for each of the eight downstream tasks.

D.1 ANTIBODY AND NANOBODY INTERFACE PREDICTION

Datasets. High-quality datasets are essential for evaluating antibody and nanobody interface modeling. We use SAbDab (Dunbar et al., 2013) as the training and validation dataset and adopt the same training, validation, and test splits as in IgGM (Wang et al., 2025a) to ensure fair comparison. Moreover, we removed anti-ligand pattern from the dataset. In total, we constructed 10028 samples from 5146 unique PDB ids, in which 2023 samples are nanobody and make up 1108 unique PDB ids. for training and validation and evaluate performance on 60 antigen-antibody docking structures (SAb-23H2-Ab) and 27 antigen-nanobody docking structures (SAb-23H2-Nano).

Finetuning and inference. Accurate antibody modeling requires leveraging both sequence and structural information. We incorporate epitope annotations, which have been shown to be critical for reliable antibody prediction (Wang et al., 2025a), by labeling residues with at least one heavy atom within 10 Å of an antibody or nanobody chain. During fine-tuning, we adopt four complementary training modes to balance complex structure prediction and antibody design: (i) with 30% probability, the model receives full sequences and predicts the antibody-antigen complex structure; (ii) with 40% probability, the model is provided with the antibody backbone sequence, antigen sequence, and antigen structure, and is tasked with designing antibody CDR sequences and structures; (iii) with 15% probability, the model receives antibody and antigen sequences along with the antibody structure and predicts the antigen structure; and (iv) with 15% probability, the model receives antibody and antigen sequences along with the antigen structure and predicts the antibody structure. During inference, we follow the IgGM protocol for fair comparison, generating 5 samples per test instance. The input consists of the antigen sequence, antigen structure, epitope annotations, and antibody sequence, and the model predicts the final antigen-antibody or antigen-nanobody complex structure.

Evaluation metrics. Model performance is evaluated using DockQ, interface RMSD (iRMS), ligand RMSD (LRMS), and success rate (SR, defined as the proportion of samples which have DockQ ≥ 0.23), which are widely used in the antibody modeling community (Mirabello & Wallner, 2024; Wu et al., 2024b; Wang et al., 2025a). DockQ, iRMS, and LRMS are averaged across all generated samples, while the success rate is computed as the fraction of all generated samples with DockQ ≥ 0.23. As shown in Table 6 and Table 7, FlexProtein consistently achieves the best performance across all metrics, substantially outperforming existing baselines.

D.2 PROTEIN-PEPTIDE INTERFACE PREDICTION

Datasets. High-quality, non-redundant datasets are essential for training accurate protein-peptide interface models. We constructed the training dataset from PepGLAD (Kong et al., 2025) and applied a temporal filter to exclude entries released after September 30, 2021, ensuring that the pretrained model had no prior exposure to test-like data. After filtering, the dataset contains 5,202 non-redundant protein-peptide complexes, reduced from the original 6,105 entries. For evaluation, we use FoldBench (Xu et al., 2025), which comprises 51 protein-peptide pairs, all sharing less than 40% sequence identity with the training and validation sets.

Finetuning and inference. Effective fine-tuning is crucial for adapting a foundation model to specific downstream tasks. We adopt the same hyperparameters as in the antibody and nanobody interface prediction task. During inference, we follow the AlphaFold 3 procedure, generating 5 samples

Table 6: Metrics for prediction of antigen-antibody docking structure. tFold-Ag and AlphaFold 3 use MSA information as input. dyMEAN, IgGM, and FlexProtein use antigen structure information. All methods except AlphaFold 3 use epitope information. AlphaFold 3, dyMEAN, tFold-Ag, and IgGM results are taken from Wang et al. (2025a). Methods marked with † use MSA as input.

Method	DockQ↑	iRMS↓	LRMS↓	SuccessRate↑
tFold-Ag→HDock†	0.022	16.652	48.157	0.0000
tFold-Ag†	0.252	6.796	21.035	0.4068
AlphaFold 3†	0.295	10.965	32.408	0.3684
dyMEAN	0.101	8.923	27.423	0.0667
IgGM	0.299	6.220	19.489	0.4667
FlexProtein	**0.384**	**5.704**	**14.533**	**0.6133**

Table 7: Metrics for structure prediction for nanobody. tFold-Ag and AlphaFold 3 use MSA information as input. IgGM, and FlexProtein use antigen structure information. All methods except AlphaFold 3 use epitope information. AlphaFold 3, tFold-Ag, and IgGM results are taken from Wang et al. (2025a). Methods marked with † use MSA as input.

Method	DockQ↑	iRMS↓	LRMS↓	SuccessRate↑
tFold-Ag†	0.288	6.349	15.081	0.4296
AlphaFold 3†	0.287	11.219	32.676	0.3885
IgGM	0.291	7.988	22.017	0.4400
FlexProtein	**0.336**	**5.380**	**11.632**	**0.5111**

per test instance. The input consists of the protein sequence, protein structure, epitope annotations, and the peptide sequence, and the model predicts the corresponding protein-peptide complex structure. Additional details are provided in Section D.1.

Evaluation metrics. Standardized metrics are important for consistent assessment of interface prediction performance. We evaluate model performance using DockQ, interface RMSD (iRMS), ligand RMSD (LRMS), and success rate (SR, defined as the proportion of samples which have DockQ ≥ 0.23). DockQ, iRMS, and LRMS are averaged across all generated samples, while the success rate is computed as the fraction of samples with DockQ ≥ 0.23. When computing DockQ, the heavy and light chains of the antibody are merged into a single chain, with the antigen treated as a separate chain. As shown in Table 8, FlexProtein outperforms all other methods across these metrics, demonstrating more accurate modeling of protein-peptide interfaces.

Table 8: Metrics for peptide structure prediction. AlphaFold 3 and Boltz-1 use MSA information as input, while PepGLAD and FlexProtein leverage protein structure and epitope information. Results for Boltz-1 and AlphaFold 3 are taken from (Xu et al., 2025), where DockQ scores are not reported. PepGLAD results are obtained by running the method on this benchmark. Methods marked with † use MSA as input.

Method	DockQ↑	iRMS↓	LRMS↓	SuccessRate↑
Boltz-1†	-	2.88	8.82	0.8039
AlphaFold 3†	-	2.81	6.56	0.8431
PepGLAD	0.413	3.18	6.72	0.8118
FlexProtein	**0.558**	**1.93**	**5.17**	**0.9137**

D.3 ANTIBODY AND NANOBODY DESIGN

Datasets. High-quality and consistent datasets are essential for evaluating CDR design performance. We use the same training, validation, and test datasets as described in Section D.1 to ensure comparability and reproducibility.

Training and inference. Effective CDR design requires careful integration of sequence and structural information. During training, we follow the procedure in Section D.1. At inference, the CDR regions are masked to enable the model to design new sequences based on the antigen structure and epitope annotations. Users can also specify different CDR lengths to generate diverse designs. Following the IgGM evaluation protocol, we generate 5 samples per test case using the same CDR lengths as the wild-type sequences, jointly designing all six CDR regions for antibodies and all three CDR regions for nanobodies.

Evaluation metrics. Quantitative metrics are necessary to assess both sequence and structural fidelity in CDR design. We use amino acid recovery (AAR) (Wang et al., 2025a) to measure sequence similarity to the wild-type, with higher values indicating closer resemblance. For antibodies, AAR is computed separately for each of the six CDR regions (three from the heavy chain and three from the light chain) and averaged across samples. Structural evaluation is conducted using DockQ, interface RMSD (iRMS), ligand RMSD (LRMS), and success rate (SR, defined as the proportion of samples which have DockQ ≥ 0.23), which compare the designed structures against wild-type complexes. When computing DockQ, the heavy and light chains of the antibody are merged into a single chain, with the antigen treated as a separate chain. As shown in Tables 9 and 10, FlexProtein achieves performance comparable to IgGM across all metrics, demonstrating that it produces realistic CDR sequences while maintaining structural integrity.

Table 9: Comparison of antibody modeling methods for antibody design, reporting CDR loop accuracy (AAR, RMSD) and docking performance. Higher values of AAR, DockQ, and SR indicate better performance, while lower values of RMSD, iRMS, and LRMS are preferable. DockQ scores are computed by comparing the designed structures against the corresponding wild-type complexes.

Model	DiffAb (IgFold)	dyMEAN	IgGM	FlexProtein
AAR ↑				
L1	0.597	0.633	**0.750**	0.727
L2	0.598	0.634	0.743	**0.773**
L3	0.421	0.570	0.635	**0.653**
H1	0.642	0.742	0.740	**0.745**
H2	0.363	0.627	0.644	**0.683**
H3	0.214	0.294	0.360	**0.414**
Docking with wild-type				
DockQ ↑	0.022	0.079	0.246	**0.273**
iRMS ↓	17.034	9.698	**6.579**	6.961
LRMS ↓	48.163	28.764	19.678	**19.599**
Success Rate ↑	0.000	0.049	0.433	**0.460**

D.4 PROTEIN-LIGAND DOCKING PREDICTION

Datasets. Benchmarking zero-shot performance is a key way to assess a model's generalization ability without task-specific fine-tuning. For this task, we directly evaluate our model in the zero-shot setting. The test set is PoseBusters V1 (Buttenschoen et al., 2024), which contains 428 protein-ligand complexes deposited in the PDB between January 1, 2021 and May 30, 2023. For pretraining, we follow the same protocol as Boltz-1 and Protenix, using all PDB structures released before 2021-09-30. Since these three methods share the same data cutoff time, comparisons on the test set remain fair.

Table 10: Comparison of nanobody modeling methods for nanobody design, reporting CDR accuracy (AAR), RMSD, and docking performance. Higher values of AAR, DockQ, and SR indicate better performance, while lower RMSD, iRMS, and LRMS are preferable. DockQ scores are calculated by comparing the designed structures to the corresponding wild-type complexes.

Method	CDR1 ↑	CDR2 ↑	CDR3 ↑	DockQ ↑	iRMS ↓	LRMS ↓	Success Rate ↑
DiffAb (AF3)	0.533	0.291	0.156	0.211	13.265	35.805	0.346
IgGM	**0.565**	0.330	0.183	**0.267**	6.927	14.966	0.415
FlexProtein	0.500	**0.441**	**0.218**	0.244	**5.571**	**13.506**	**0.437**

Training and inference. Evaluating zero-shot inference provides insight into a model's ability to directly predict complex structures from minimal inputs. Given protein sequences and ligand SMILES strings, we generate full protein-ligand complex structures in a manner similar to AlphaFold 3. For FlexProtein, Protenix, and Boltz-1, we generate five samples per complex using a single random seed.

Evaluation metrics. Standardized metrics are critical to ensure reliable comparison across methods. Following the AlphaFold 3 protocol, we report the success rate, defined as the percentage of predictions with a pocket-aligned RMSD <2 Å. The pocket-alignment procedure is consistent with AlphaFold 3: the pocket is defined as all heavy atoms within 10 Å of any ligand heavy atom, restricted to the primary polymer chain or modified residue of the ligand, and further limited to protein backbone atoms. Baselines include MSA-based methods (AlphaFold 3, Chai-1, Protenix, Boltz-1) and single-sequence methods (DiffDock, DiffDock-L). For Protenix and Boltz-1, results are reported using the top-ranked sample out of five diffusion-generated predictions. For FlexProtein, which does not include a confidence head, we report both the *random-1* score (performance of a randomly chosen sample) and the *oracle* score (the best of five samples selected against the ground truth). Note that Boltz-1 failed on two targets (7M31_TDR and 7SUC_COM) due to residue number restrictions.

Table 11: Success Rate (SR) comparison of different methods.

Method	SR (%) ↑
MSA-based	
AlphaFold3 2019	76.40
Chai-1	77.00
Protenix top1	77.10
Protenix oracle	81.31
Boltz-1 top1	69.62
Boltz-1 oracle	74.06
Single-sequence-based	
DiffDock	37.90
DiffDock-L	50.00
FlexProtein random1	71.82
FlexProtein oracle	78.70

D.5 Ligand-induced conformational change prediction

Datasets. Benchmarking zero-shot performance provides insights into a model's ability to generalize without task-specific adaptation. For this task, we do not perform finetuning and directly evaluate the zero-shot capability of our model. The test set, originally derived from Saldaño et al. (2022), consists of 90 apo-holo protein pairs.

Training and inference. Zero-shot inference allows us to assess the model's structural prediction ability under different input conditions. Given protein sequences and ligand SMILES strings, we generate structural predictions without any fine-tuning. Specifically, we produce five predictions for each case without ligands (apo) and five predictions with ligands obtained from the original holo complexes in the PDB. For fair comparison, we report two evaluation settings: (1) all five apo samples, and (2) a mixed set of three apo samples and two holo samples.

Evaluation metrics. Rigorous evaluation metrics are essential to capture both accuracy and diversity in structural predictions. Following the protocol in AlphaFlow (Jing et al., 2024), we use two types of metrics. The first is the Pearson correlation (r) between sampled diversity and ground-truth

Table 12: Evaluation of ligand-induced conformation changes: (1) Pearson correlation (r) between sampled and ground-truth diversity as measured by the residue flexibility (ResFlex, absolute deviation after alignment), and (2) the ensemble TM-score (TM-ens). For residue flexibility, both global (gl.) correlations and mean/median per-target (pt.) correlations are reported; for TM-ens, mean/median correlations are reported. Higher values indicate better performance. Methods marked with † use MSA as input.

Method	ResFlex r (gl.)	ResFlex r (pt.)	TM-ens
Benchmark results compared against baselines			
AlphaFlow†	0.455	0.527/0.527	0.864/0.893
ESMFlow	0.416	0.496/0.522	0.856/0.893
ESMDiff	0.424	0.502/0.517	0.851/0.883
FlexProtein	0.503	0.551/0.542	0.889/0.920
FlexProtein with ligand	**0.519**	**0.587/0.615**	**0.901/0.931**
Ablation study			
FlexProtein 5apo	0.503	0.551/0.542	0.889/0.920
FlexProtein 5holo	0.454	0.524/0.540	0.888/0.918
FlexProtein 3apo+2holo	0.519	0.587/0.615	0.901/0.931
FlexProtein 5apo+5holo	0.535	0.620/0.638	0.907/0.936

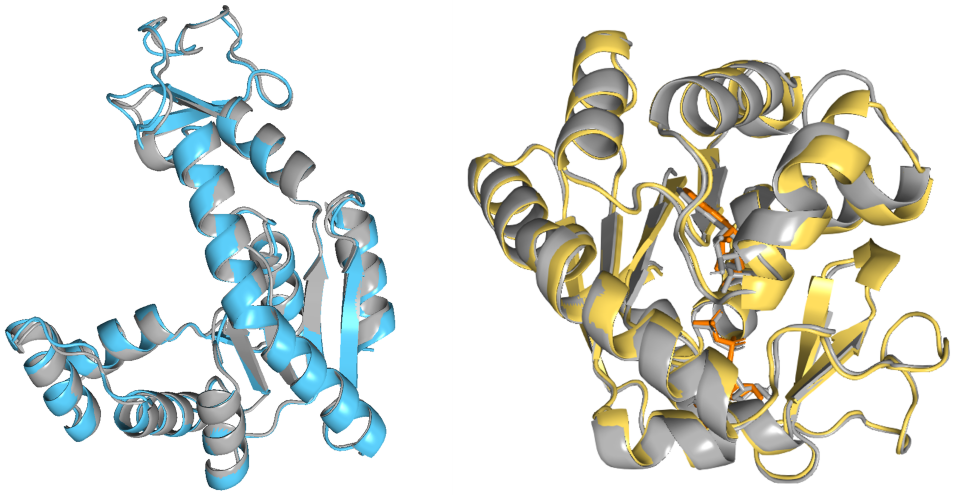

Figure 7: Overlay of the predicted structures of the (left) apo and (right) holo states of adenylate kinase, overlaid with the native structures (grey) from the PDB. The holo state includes an AMP and an ADP molecule (orange), whose presence induces the kinase to fold inwards to hold the molecules in place.

diversity, measured by residue flexibility (ResFlex, absolute deviation after alignment), reported as global (gl.) mean and per-target (pt.) mean/median correlations. The second is the ensemble TM-score (TM-ens), reported as mean and median. Results are presented in Section 4.2. In addition, we conduct an ablation study with different ligand conditions, showing that FlexProtein can generate structures in multiple conformational states, highlighting its potential to address the protein multi-state problem.

D.6 Protein-ligand binding affinity prediction

Datasets. Reliable benchmarking requires consistent training and evaluation protocols. Following the strategy of SPIN (Choi et al., 2024), we use the same training and test sets. The training data is drawn from PDBbind v2020 (Liu et al., 2017), comprising 19,443 protein-ligand complexes. For evaluation, we adopt the CASF-2016 (Su et al., 2018) benchmark, which includes 285 samples. To prevent data leakage, any overlapping entries between CASF-2016 and the training set were removed.

Training and inference. Model training is formulated as a regression task to predict binding affinity values. The input is the three-dimensional structure of protein-ligand complexes, and the target output is a continuous affinity score. During inference, the model predicts one affinity score per sample, which is directly compared against the ground-truth value.

Evaluation metrics. Standard regression metrics are used to assess predictive accuracy and correlation with experimental data. Specifically, we report the Root Mean Square Error (RMSE) and Pearson's correlation coefficient (R). Detailed results are presented in Section 4.2.

D.7 EC number prediction

Datasets. Accurate enzyme function prediction requires high-quality annotation datasets. We follow the dataset setup used in DeepFRI (Gligorijević et al., 2021), where enzyme annotations are derived from UniProtKB with experimentally validated Enzyme Commission (EC) numbers. The training, validation, and test set contains 15551, 1729, 1919 protein samples respectively.

Training and inference. The task is formulated as a multi-label classification problem, where each protein sequence may be associated with one or more EC numbers. During inference, the model outputs probability scores over all possible EC labels, which is then used to compute the precision-recall curve. For both training and inference, the protein sequences are passed through our base model once for structure prediction, after which both sequence and structural information are used for model finetuning and evaluation.

Evaluation metrics. Model performance is evaluated using the maximum F-score (F_1), which balances precision and recall. Specifically, F_1 is defined as the maximum F-score across all probability thresholds:

$$F_1 = \max_{t \in [0,1]} \frac{2 \cdot \text{Precision}(t) \cdot \text{Recall}(t)}{\text{Precision}(t) + \text{Recall}(t)},$$

where t is the threshold applied to predicted probabilities.

D.8 GO term prediction

Datasets. Gene Ontology (GO) provides a comprehensive representation of protein function, covering three sub-ontologies: Molecular Function (MF), Biological Process (BP), and Cellular Component (CC). Following DeepFRI (Gligorijević et al., 2021), we construct the training, validation, and test set as shown in Table 13.

Training and inference. GO term prediction is also framed as a multi-label classification problem. For each protein, the model outputs probability scores over GO terms independently for MF, BP, and CC. The training and inference procedures are identical to those used for EC number prediction, including structure prediction.

Evaluation metrics. Performance is measured by F_1, defined as the maximum F1-score across thresholds. The metric captures the balance between precision and recall in predicting GO terms and is widely adopted in functional annotation benchmarks.

Table 13: Size of data samples used for the Gene Ontology (GO) task.

Sub-ontology	Training samples	Validation samples	Test samples
BP	23514	2624	3415
MF	24952	2747	3415
CC	11298	1299	3415

E USAGE OF LLM

We employed GPT-5 to assist in refining the writing of this manuscript. Specifically, GPT-5 was used to polish grammar, improve readability, and streamline phrasing, while all scientific content, experimental design, and data analysis were developed and verified by the authors. The use of GPT-5 was limited to language refinement and did not influence the technical contributions or conclusions of this work.