Posetta: Language-Guided Protein Design

Stanford CS224N Custom Project

Haotian Du
Department of Chemistry
Stanford University
`dht14@stanford.edu`

Jingjia Liu
Department of Bioengineering
Stanford University
`ljjchem@stanford.edu`

Tianyu Lu
Department of Bioengineering
Stanford University
`tianyulu@stanford.edu`

Abstract

Protein research is essential across many fundamental disciplines. Generating scientific hypotheses in such research depends on comprehending complex protein structural features, a task traditionally reliant on the expertise of structural biologists, which creates a significant knowledge barrier. Recent progress in Large Language Models (LLM) offers a promising approach to using published texts for understanding protein structures. Here we present Posetta, a multimodal system that combines a protein encoder and an LLM to produce detailed descriptions of protein structures. By encoding protein `.pdb` files with state-of-the-art models and integrating them through a vision transformer, Posetta aims to democratize protein structure comprehension and accelerate advancements in protein engineering. A demo is available at `https://colab.research.google.com/drive/105FR-kbs-ax1WXpIlGico2EVR3ZDo0t-?usp=sharing`.

1 Key Information to include

- Mentor: Shijia Yang
- Project type: Custom
- Team Contributions:
  - H.D.: Conceptualization, Data preparation (keyword filtering of full-text), Code (evaluation metrics), Evaluation, Writing
  - J.L.: Conceptualization, Data preparation (extracting PDB titles), Code (protein embeddings), Evaluation, Writing
  - T.L.: Conceptualization, Data preparation (getting ProteinChat dataset and Pubmed full-text dataset), Code (dataloaders, model training and logging), Training, Evaluation, Writing
- External Collaborators (if you have any): None
- Sharing project: No

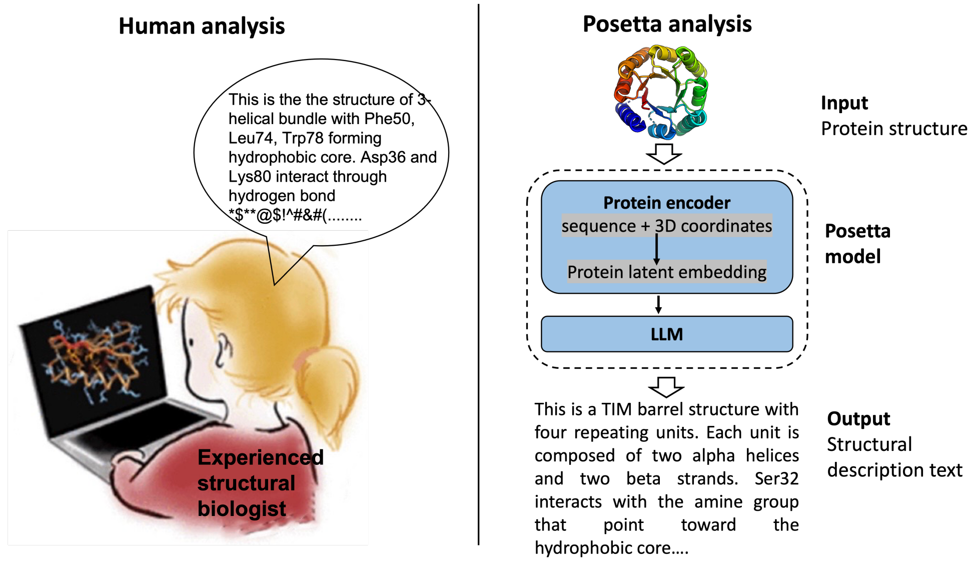
2 Introduction

Proteins play crucial roles in every perspective of biological processes. Comprehending complex protein structural information has always been essential in life science research. As protein designers, we recognize that the inspirations of protein engineering also heavily rely on the understanding of protein structures. However, translating these intricate structures into comprehensible text depends on the expertise of structural biologists. For example, our PI (trained for 20+ years) would examine a new structure first and explain to us (less experienced graduate students) the key structural features. The reliance on expert interpretation creates a significant knowledge barrier.

Recent advancements in Large Language Models (LLMs) have demonstrated significant progress in comprehending task-specific knowledge conditioned on data from multiple modalities. These advancements suggest the potential to leverage human language from published text for understanding protein structures and provide valuable insights for protein research.

Here we introduce Posetta, which aims at generating detailed protein structural descriptions after "looking at" the structure (Figure.1). This system integrates a protein encoder and an LLM decoder to achieve this goal. The input protein structure (`.pdb` file) is encoded by state-of-art structural encoder in the protein design field, ProteinMPNN (Dauparas et al., 2022) or LigandMPNN (Dauparas et al., 2023). The protein embeddings are then integrated with the LLM by patchifying and feeding the patches through a vision transformer (ViT) (Radford et al., 2021) following the GiT model (Wang et al., 2022).

Our ideal outcome is to build a model that has effectively "read" and integrated knowledge from all structural biology publications. By explaining important structural features, Posetta aims to serve as an "experienced consultant" that helps researchers understand protein structures and eventually facilitates scientific discovery in the field.


Human analysis
This is the the structure of 3-helical bundle with Phe50, Leu74, Trp78 forming hydrophobic core. Asp36 and Lys80 interact through hydrogen bond *$**@$!^#&#(........
Experienced structural biologist
Posetta analysis
Input
Protein structure
Protein encoder
sequence + 3D coordinates
Protein latent embedding
LLM
Posetta model
Output
Structural description text

This is a TIM barrel structure with four repeating units. Each unit is composed of two alpha helices and two beta strands. Ser32 interacts with the amine group that point toward the hydrophobic core....

Figure 1: Schematic overview of the Posetta multimodal system

3 Related Work

Comprehending protein structures and capturing the structure-function relationship has been an active research topic in bioinformatics, traditionally people use rules or statistical techniques for function prediction. In recent years, machine learning methods have started to emerge. Earlier methods leverage ML models to embed structure features, for example, Gligorijević et al. (2021) used a graph convolutional network for structure embedding to predict Gene Ontology terms. However, using protein function classification (Gene Ontology terms, enzyme commission (EC) numbers, etc.) as function labels, these models are limited to predicting these low-resolution function features, omitting the uniqueness of each protein. Free-form text allows more sophisticated descriptions of protein structure and function. An earlier work of Xu et al. (2023) pioneered learning from biomedical text descriptions, the ProtST model they developed was evaluated on protein function classification and functional protein retrieval (text to protein).

The recent advance of large language models (e.g., ChatGPT) has laid the foundations for specialized language generation tasks. Research on Vision-Language-Pretraining (VLP) has demonstrated the possibility on learning multimodal foundation models and performed well on vision-and-language tasks (Radford et al., 2021). Meanwhile, work in protein neural-network has produced diverse protein encoding solutions. Inspired by these advances, LLM based multimodal text generation models for protein have been explored recently. In 2023, Guo et al. (2023) developed the ProteinChat model. They constructed the PDB Description Dataset containing PDB structures and the abstract of their corresponding publication. Their model includes a fixed pre-trained protein structure encoder ESM-IF1 and the LLM model Vicuna-13b. They only trained the projection layer between the protein encoder and the LLM. The model can perform ChatGPT-like function on simple prompts, however, the paper only provided one example output, making it hard to compare with. In 2024, Wang et al. (2024) reported the ProtChatGPT model, which uses the same PDB Description Dataset to achieve ChatGPT like function. In addition to the pre-trained ESM-IF1 structure encoder, they used an additional pre-trained protein sequence encoder ESM-1b. They also used Vicuna-13b as their LLM decoder. Their main advantage is introducing a transformer-based model instead of a simple projection layer to connect the protein encoder and the LLM decoder. Their training also only occurs on this transformer (i.e. freezing protein encoder and LLM decoder). The model can generate text related to the function of the input protein, and the model was evaluated on a spectrum of metrics.

Since the goal of the Posetta model is to generate structural description (e.g. This protein is a tyrosine kinase which contains 3 helices and 2 beta-sheets) instead of only general functional description (e.g. This protein is an enzyme), we recognize several limitations in ProtChatGPT and ProteinChat. First of all, the protein encoder used by both models are from ESMfold (Hsu et al., 2022), which is a model trained on the protein structure prediction task (predicting a protein structure from a sequence). A drawback of the ESMfold model is its tendency to memorize protein folds by aligning to native sequences rather than learning the protein's biophysical features. This limitation means that the protein embedding might lack important structural information, such as amino-acid interaction. Recent models for *de novo* protein design (Dauparas et al., 2022; Dauparas et al., 2023) could potentially provide better protein encoding solutions for our purpose, as their ability to capture protein structural feature and sequence-structure correlation have been verified in wet labs with proteins unseen by nature. In Posetta, we used the protein encoders from protein design models. Secondly, ProtChatGPT and ProteinChat models are only trained on paper abstract text, which might not include the desired information about a specific protein structure. In Posetta, we included sentences from the publication full-text that are related to structural features. Finally, the PDB Description Dataset only contains the first chain (chain A) of each pdb file, which excludes important information on protein-protein and protein-small molecule interactions between chains. For Posetta, considering the multi-chain protein complex with its ligands is essential to generate meaningful structural feature descriptions.

4 Approach

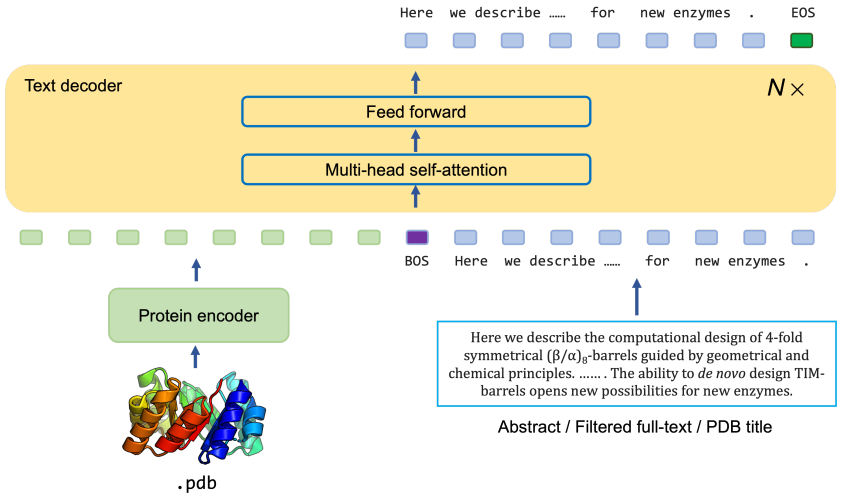
The Posetta model is composed of a protein encoder and a LLM. The protein encoder generates protein latent embeddings from .pdb files with protein sequence and structural information (Figure 2). The embeddings are reshaped and normalized to be compatible with the input shape of the LLM.

We used the pre-trained protein encoder from ProteinMPNN (Dauparas et al., 2022) and LigandMPNN (Dauparas et al., 2023) in out model. Both of the models were developed for designing protein sequences from backbone structures. Both models used a sparse graph to represent protein structure, a message-passing neural network (MPNN) to encode the protein structure, and an auto-regressive decoder to predict protein sequences. We modified the code to extract the node feature from the encoder embedding to input to the LLM decoder.

We experimented with both ProteinMPNN and LigandMPNN to investigate the effect of encoded information on the resolution of the generated text descriptions. ProteinMPNN only takes in backbone atoms as context and thus does not have access to sidechain coordinate information. This makes it suitable for descriptions involving e.g. secondary structure or topology. In contrast, LigandMPNN (side chain only) takes in backbone and sidechain atoms as context, making it suitable for descriptions involving amino acid mutation or atomic interactions such as hydrogen bonds and pi-stacking. LigandMPNN (side chain and heteroatoms) additionally encodes the coordinate of ligands. (e.g. ATP molecules).

As in ProteinChat and ProChatGPT, we decided to freeze the pre-trained encoders. ProteinMPNN and LigandMPNN are trained on all .pdb files in the protein data bank, and have been extensively verified in wet-lab for their ability to capture protein structure features correctly. We believe that they generate ideal proteins structural embeddings and do not want to perturb/bias it with text descriptions.

We use the GiT model architecture described by Wang et al. (2022) and Guo et al. (2023) as the LLM component. The original GiT model was trained to caption (224×224) RGB images. We repurpose the model to take in $(128 \times N)$ protein embeddings, where 128 is the per-amino acid embedding dimension from ProteinMPNN and N is the protein length. The protein embeddings are zero-padded to $(128 \times M)$ where M is the next greatest integer divisible by 3, reshaped to $(128 \times M/3 \times 3)$ and then zero-padded to $(224 \times 224 \times 3)$. Since LigandMPNN embeddings include information from multiple chains, the total protein length can become very large. To fit information from longer proteins, we average pool adjacent embedding vectors of LigandMPNN embeddings such that each position in the final embedding corresponds to a pair of adjacent residues.


Here we describe ...... for new enzymes . EOS
Text decoder
$N\times$
Feed forward
Multi-head self-attention
BOS Here we describe ...... for new enzymes .
Protein encoder
.pdb
Abstract / Filtered full-text / PDB title

Here we describe the computational design of 4-fold symmetrical $(\beta/\alpha)_8$-barrels guided by geometrical and chemical principles. .... The ability to *de novo* design TIM-barrels opens new possibilities for new enzymes.

Figure 2: Model Architecture

5 Experiments

5.1 Data

We used the PDBs in the PDB Description dataset. The dataset was randomly split into 90% train and 10% test, giving 73,508 training and 8303 test examples for the backbone-only ProteinMPNN encodings and 89,846 training and 10052 test examples for the atomic-context LigandMPNN encodings. Only chain A of the PDB file was used for ProteinMPNN encoding for compatibility. The full-length PDBs containing all chains were used for LigandMPNN encoding. Approximately 76% of the dataset include articles not in the open access full text dataset of PubMed so for such examples we fall back to using their abstracts. For structures that has open access full text available, we filter the text based on structural description keywords for extracting the structural description paragraphs.

5.2 Evaluation method

For quantitative evaluation, we employ six commonly used metrics described in ProtChatGPT (Wang et al., 2024), including BLEU (Papineni et al., 2002), ROUGE-L (Lin and Hovy, 2002), METEOR (Banerjee and Lavie, 2005), CIDEr (Vedantam et al., 2015), and BertScore (Zhang et al., 2019), as a systematic comparison between the generated text and the reference text. BLEU score quantifies the similarity between the candidate and reference text through n-gram matching. ROUGE-L calculates the longest common subsequence between the candidate and reference text. METEOR evaluates the text by also considering synonyms and stemming. CIDEr is designed to evaluate generated image captions. BertScore measures the similarity between the candidate text and the reference text using contextual embeddings from BERT models. To better access the quality of generated protein-related descriptions, PubMedBERT (Gu et al., 2021) was applied, which is a BERT model pre-trained on the biomedical corpus.

5.3 Experimental details

We train five models, each with the same architecture but with differences which are summarized below. Model 1 was trained for 57 epochs for 3 days and 22 hours on an A100 before a force restart on our node. We then continued training from the epoch 57 checkpoint for an additional 26 epochs. Model 2 was trained for 41 epochs but its validation loss remained much higher than the fully trainable model (Figure.3). Model 3 was trained for 13 epochs before another force restart on our node. We then continued training from the epoch 13 checkpoint for an additional 5 epochs.

The model architecture is identical for all models and taken from Wang et al. (2022), consisting of 12 layers of ViT (self-attention on patchified structure embeddings) followed by 6 layers of self and cross attention for generating text tokens. The vocabulary size is 30522, the embedding and hidden dimensions are 768 throughout, except in intermediate MLP blocks where the dimensionality expands 4× to 3072 then back down. A dropout probability of 0.1 is applied to the input word embeddings and each output of the text attention blocks. The model was trained with a batch size of 32, learning rate of 1e-5 using the AdamW optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$ (Loshchilov and Hutter, 2017). We experimented with Low-Rank Adaptation (Hu et al., 2021) by only training adaptors to the query, key, value, and output projection linear layers of both the ViT attention blocks and the text attention blocks.

We initially planned to train the Posetta Model using PubMed full text and the ProteinMPNN embedding (Model 1 and 2). However, we observed issues with repetitive text and inaccurate scientific descriptions in the initial results. To address these issues, we tried several different strategies. Model 3 was trained on short PDB titles, which are usually a general description of the structure identity, instead of long PubMed texts. This is because the GiT model was pre-trained on image captioning texts, which are typically short. Models 4 and 5 were trained using LigandMPNN, incorporating side chain and all-atom information, respectively. This was done to: first, test whether more complex protein encoders would provide more comprehensive encoding of structural features; second, enable protein encoding on protein-protein complex and protein-ligand complex.

Model	Dataset	Structure Encoding	Sidechain	Heteroatom	LoRA
1	PubMed full-text	ProteinMPNN	No	No	No
2	PubMed full-text	ProteinMPNN	No	No	Yes
3	PDB Title	ProteinMPNN	No	No	No
4	PubMed full-text	LigandMPNN	Yes	No	No
5	PubMed full-text	LigandMPNN	Yes	Yes	No

5.4 Results

We observe a steady decrease in both train and test losses for models using ProteinMPNN structure encodings, with training loss still decreasing at epoch 57 but test loss reaching a plateau (Figure.3). This trend holds for the LigandMPNN structure encodings as well (Figure.4). Interestingly, the LigandMPNN training losses are higher which could be explained by the average pooling reducing the resolution of the embeddings compared to ProteinMPNN.

We compute metrics on 100 randomly sampled holdout structures. The results are summarized in the table below. In all cases, higher score is better. We do not evaluate on the LoRA trained model as the validation loss was much higher than the other models. For multinomial sampling, we set temperature = 0.1, `top_k = 50`, and `top_p = 0.95` with a fixed seed of 1.

The Posetta Models overall have worse BLEU, ROUGE, METEOR, and CIDEr scores compared to the ProChatGPT model, indicating that the generated language is less similar to their reference texts. However, they achieve better PubMed BERT scores, suggesting that the generated content has better scientific context. **Model 1 behaved the best for overall scoring metrics, and should be the default choice when analyzing monomeric proteins. Model 4 should be considered when analyzing protein-protein complex structures. Model 5 should be considered when analyzing protein-ligand complex structures. Model 1 can not be applied for these examples because the ProteinMPNN encode excluded these features. However, it is important to note that the increased complexity of the LigandMPNN encoder makes training more challenging, potentially affecting performance.**

Model	Sampling	BLEU-1	ROUGE-L	METEOR	CIDEr	PubMed BERTScore
1	Greedy	0.20	0.22	0.23	0.15	0.66
1	Multinomial	0.17	0.19	0.22	0.05	0.65
3	Greedy	0.07	0.31	0.28	0.00	0.70
3	Multinomial	0.07	0.26	0.23	0.00	0.68
4	Greedy	0.04	0.16	0.15	0.10	0.61
4	Multinomial	0.20	0.17	0.19	0.10	0.62
5	Greedy	0.27	0.15	0.14	0.05	0.60
5	Multinomial	0.14	0.15	0.18	0.01	0.62
ProtChatGPT	Multinomial	0.61	0.49	0.29	0.64	0.46

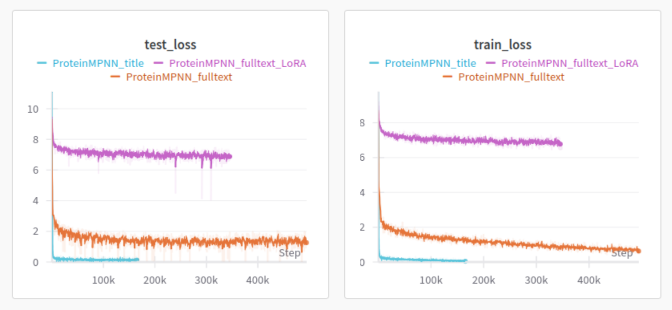


Figure 3: Test and train loss curves for models using ProteinMPNN structure encodings (models 1, 2, and 3). Curves are shown with a smoothing factor of 0.5.

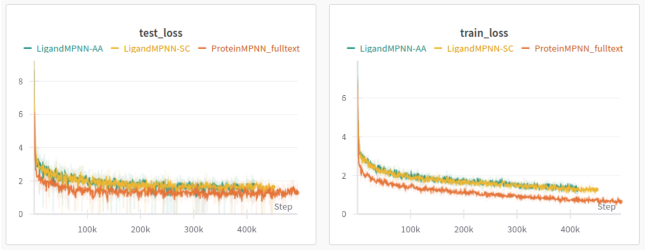


Figure 4: Test and train loss curves for models using LigandMPNN structure encodings (models 4 and 5) with the ProteinMPNN fulltext model shown for comparison (model 1). Curves are shown with a smoothing factor of 0.5.

Model 3 which was trained on PDB titles, appears to have the lowest train and test losses. When looking at the generated examples, they can generate grammatically correct titles that sound like PDB titles (reflected by the highest BERT Score), without repetition issues, which echos our hypothesis that the GiT model is better at handling short content. However, the contents are wrong, as reflected by the low BLEU score and CIDEr score.

We explored the multinomial sampling strategy as an alternative of the initial greedy sampling strategy, as we observed that greedy sampling tend to give highly repetitive phrases and sentences. **Multinomial sampling with the `top_k` and `top_p` settings noted above tend to produce much more coherent text with much fewer repetitions (see Appendix 9.1).** We note that multinomial sampling does not have a large effect on the metrics except BLEU score. The observed repetitions in the greedy sampling strategies are not reflected in this set of metrics as repetition can artifically inflate BLEU score, making it better than it actually is.

We also evaluate the decoder's sensitivity to the protein structure embeddings by perturbing the embeddings with Gaussian noise and observing differences in the decoded text. We observe that the model is very sensitive to the input embeddings. Adding Gaussian noise with standard deviation of as little as 0.001 changes the decoded text (see Appendix 9.1).

6 Analysis

Evaluating the protein structure-to-text model requires a comprehensive approach that addressed both the linguistic quality of the generated description and their scientific accuracy and utility. From a linguistic standpoint, the evaluation focuses on how well the model generates clear, coherent, and understandable language. From a scientific perspective, the evaluation focuses on whether the model can identify the function class the of a unseen protein (e.g., enzyme, antibody, etc) and whether it can capture correct structural features (secondary, ternary structure, amino acid interaction, etc).

Here we focus on analyzing two examples, which are newly deposited structures in the RCSB that were **not included in the dataset** (Our dataset included all structures published before June 2023, but both examples are new structures released in 2024.). We also aim to compare our generated text with that produced by the reference ProtChatGPT model. However, since the ProtChatGPT model has not been open-sourced, we currently cannot perform this comparison.

Example 1 PDB ID: 9BKP, Crystal structure of rubredoxin from Pyrococcus furiosus reconstituted with FeCl3 solved by Fe/S-SAD
This is a small monomeric protein, therefore ProteinMPNN encoding with multinomial sampling was applied. Detailed settings see Appendix 9.2.

Generated text: a novel, highly efficient rubredoxin from thermus thermophilum. the three - dimensional structure of the reduced rubredoxin from pyrococcus furiosus was determined with an ncs - type distance of 18 + / - 1. 2 a. rubredoxin was solved by single - wavelength anomalous dispersion using selenomethionine and single - quantum - mechanical (sad) techniques. the structure is very similar to other rubredoxins. however, it exhibits a more compact fold that has a more extensive network of relatively weak hydrogen bonds. two distinct clusters with interdigitated residues per turn form a network of protein -protein and water - and disulfide bonds. rubredoxin r1 consists of three different, two and four other auxiliary clusters, which, consisting of one small and two adjacent aspartates, are located with no redox potentials. the structure of the rubredoxin r1 refined to 2. 0 a reveals a network of interactions between residues 29 and 50 of an extended loop with residues 30 - 37. this interface includes multiple weak but non - stable, concerted, concerted rearrangements along aromatic and non - catalytic residues including residues 32 - 32 of highly reduced rubredoxin r1. nevertheless, this high number of clusters in this study has not been accurately characterized. this finding provides a new paradigm for multi - enzyme interactions, particularly for rubredoxins, which act as protein and rubredoxins.

Language analysis: The paragraph provides a substantial amount of detail about the protein structure, which can help in conveying a comprehensive picture of the research findings. It follows a logical structure, beginning with an introduction to the protein, followed by details about its structural characteristics, and ending with broader implications. However, it contains repetition of phrases (e.g., "concerted, concerted") and awkward constructions that make the text hard to read. There is redundancy in phrases like "protein and rubredoxins" and ambiguity in statements like "three different, two and four other auxiliary clusters." The paragraph also lacks coherence, with abrupt shifts between topics and ideas.

Scientific analysis: The description successfully captured that this protein is a rubredoxin. The paragraph is detail-oriented and uses technical vocabulary effectively for an expert audience. The functional structural description is also highly-related to the rubredoxin, for example, clusters with redox potential, hydrogen bond networks, etc. It also successfully captured secondary structure features, for example, the extended loop between residue 30 - 37. However, some scientific context are vague and even nonsensical, such as "a network of protein -protein and water - and disulfide bonds", and "three different, two and four other auxiliary clusters".

Example 2 PDB ID: 8U9G, Human Class I MHC HLA-A2 bound to sorting nexin 24 (127-135) neoantigen KLSHQLVLL

This is protein-protein complex, therefore LigandMPNN-SC encoding with multinomial sampling was applied. Detailed settings see Appendix 9.2.

Generated text: t cell recognition between viral and host cell receptors is considered a method for immunotherapy. the structural basis for the highly optimized t cell receptor (tcr) cross - reactivity, however, remains elusive. we found that the non - immunogenic peptide, p2 - nc5b (residues 34 - 70), is a potent tcr, in which binding of different and functionally unrelated peptides was replaced by the t cell receptor cd8 + t cells. p2 - nc5b is a prevalent public tcr in the hiv - 1 hla - a2 strain, and it has been noted that the unmutated p2 - nc5b structure can adopt all of the same cross - reactive conformations. the same tcr cross - reactivity of the mart - 1 - a tcr with three different mhc ligands is the structural basis for the high tcr cross - reactivity of immunotherapy. peptides cross - reactivity was observed in the peptides, but their conformations were different for the two presented peptides. we conclude that p2 - nc5b is specifically recognized by class ii mhc molecules, which can be used as all peptides to recognize the same major groove on the peptide. the structural and biochemical data presented here provide a clear explanation for the broad tcr cross -reactivity of class i mhc molecules.

Language analysis: This paragraph follows the correct logic of introducing a protein, starting from the relevance in immunotherapy into detailed descriptions. It uses the correct scientific terms related to this topic. However, it has grammarly incorrect sentence constructions such as "the binding of different and functionally unrelated peptides was replaced by the T cell receptor CD8+ T cells" make the text difficult to follow. There is also ambiguity in phrases like "peptides cross-reactivity was observed in the peptides," making it unclear what is being discussed.

Scientific analysis: The description generated descriptions related to TCR-MHC interactions and peptide cross-reactivity, which is related to the input structure from a scientific topic perspective. However, the contents about MHC peptide presentation and peptide conformation are not correctly described. And it's also confusing regarding whether this protein is a MHC or a TCR.

7 Conclusion

7.1 Summary of contribution

In this project, we introduce Posetta, a multimodal system designed to generate detailed descriptions of protein structures by integrating a protein encoder with a large language model (LLM). A key contribution is our successful fine-tuning of the GiT model, originally designed for image captioning, for scientific contexts on protein structure description. Another key contribution is enabling the inclusion of protein-protein complex and protein-ligand complex information in the structural descriptions, which was not explored by earlier models. We developed five models using protein encoders derived from state-of-the-art protein design models, experimenting with different training data and settings to enhance text generation accuracy and context. This project demonstrates the potential of combining advanced protein encoders and LLMs to facilitate protein structure comprehension and accelerate advancements in protein engineering.

7.2 Limitations and future work

Although the model is able to generate structure-related text for proteins and capture correct structural and functional features, we note that Posetta is prone to hallucinations where generated text contain truthful statements that are obfuscated by falsehoods. The generated text has issues with sentence constructions and grammar. The connection between sentences overall lack of logic coherency.

The first limitation of this work is dataset availability. While we aimed to use the full text from PubMed for training, only a small subset is open access. Alternatively, we could rely on human expertise to generate the dataset, but this approach would be highly labor-intensive.

Improvements to the model could be made by modifying the model architecture via a hyperparameter sweep. Currently, no dropout is applied in the attention blocks processing the input protein structure embedding. Adding dropout could make the model less sensitive to possibly spurious features in the structure embedding. Noising the input protein structure embedding is another alternative, either by noising the embeddings themselves or noising the input structure used to compute the embedding. The goal is to make the decoder more robust to insignificant perturbations in the input structure and embedding. However, this must be balanced preserving information content in the structure embedding as the downstream decoder should be able to detect subtle changes in the input structure via the structure embeddings. Finally, to combat hallucinations, we anticipate that contrastive learning of protein structure-description pairs can help by pushing incorrect but plausible generations to have embeddings more distant than correct structure-description pairs. Fine-tuning with human guidance can then be feasible once a non-trivial fraction of generated samples contain mostly truths which can form the “preferred” subset of finetuning data.

8 Ethics Statement

The development and deployment of a protein structure-to-text model present significant ethical considerations that must be carefully addressed. These considerations primarily revolve around the following two perspectives:

1. **Lowering the barrier for non experts**
   By translating complex protein structures into easily interpretable language, our model could significantly lower the knowledge barrier in structural biology and protein design. While this can drive scientific progress and innovation, it can also introduce substantial risks. The simplifications and accessibility of protein data might enable the design and production of novel proteins without the necessary expertise to understand their full implications. This could lead to the creation of proteins with harmful biological activities, which could pose significant dangers to human health. Furthermore, the combination of our model with advanced protein design generative models could facilitate the exploration of vast protein spaces, which could increase the likelihood of inadvertently generating hazardous proteins such as viral components. Implementing access controls to ensure that only qualified individuals can use the model and establishing robust monitoring systems to track the usage could help with identifying potentially harmful applications.
2. **Potential for hallucinating misleading information**
   The model's ability to generate textual descriptions of protein structures, while powerful, also carries the risk of producing misleading or inaccurate information. Such inaccuracies could misguide scientists, leading to flawed experimental designs, and potentially detrimental conclusions. Therefore, ensuring the fidelity and reliability of the model's output is crucial to maintaining the integrity of scientific research. Continuous validation, peer review, and cross-referencing with established databases are essential to mitigate this risk.

By addressing these ethical concerns proactively, we aim to leverage the transformative potential of our protein structure-to-text model while safeguarding against its misuse and ensuring that it serves as a valuable tool for the advancement of science and human health.

9 Appendix

9.1 Greedy vs. multinomial sampling

All results are from giving PDB 7UEK as input which we confirmed is not in the training data. The random seed is fixed to 1 in all generations.

ProteinMPNN Title Greedy

crystal structure of protein at 1. 8 a resolution from archaeoglobus fulgidus

ProteinMPNN Title Greedy 0.001 Gaussian Noise

crystal structure of apo protein from escherichia coli

ProteinMPNN Title Greedy 0.01 Gaussian Noise

crystal structure of psest3 from thermus thermophilus hb8

ProteinMPNN Title Greedy 0.1 Gaussian Noise

structural studies of the non - ribose 5 - phosphate hydrolase from streptococcus pyogenes

ProteinMPNN Title Multinomial

crystal structure of soluble biotin protein from escherichia coli

ProteinMPNN Full Text Greedy

the crystal structure of recombinant aspartate kinase from salmonella typhimurium has been determined at 2. 0 a resolution. in this paper we present the structure of the aspartate chaperone bound to the active site of the enzyme, a noncleavable atp analogue of the natural substrate, mg2 +, and a noncleavable atp analogue. the overall structure is similar to that of the previously published crystal structure of the wild - type protein, except for two loops, that are disordered in the crystal. in addition, a hydrophobic trough located at the center of the dimer is observed in the crystal. the hydrophobic trough is surrounded by helices that is surrounded by helices. the hydrophobic trough is surrounded by helices that are similar to those of other known structures. the hydrophobic trough is surrounded by helices 3, 5, 6, and 5. the hydrophobic trough is surrounded by helices 3, 5, 5, and 5. the hydrophobic trough is surrounded by helices 4, 5, 5, 6, and 5. the hydrophobic trough is surrounded by helices 4, 5, 5, 6, and 8. the hydrophobic trough is surrounded by helices 4, 5, 5, 6, and 8. the hydrophobic trough is surrounded by helices 4, 5, 5, 6, and 8. the hydrophobic trough is formed by residues 5, 5, 6, 8, and 8. the hydrophobic trough is surrounded by helices 4, 5, 5, 8, and 8. the hydrophilic binding is facilitated by a network of conserved residues, phe (9), phe (9), and tyr (9).

ProteinMPNN Full Text Multinomial

in the quest for new potent inhibitors to treat protein tyrosine phosphorylation, the crystal structures of wild - type and two double mutant forms of escherichia coli typhospholipase d (2s, z3 +) were determined and refined to 1. 92 and 2. 75 a, respectively, for the first time both structures have been refined against data of the same crystal form (1. 65 a resolution) and a crystallographic r - factor of 18. 9 % for both models. however, only two of these conformations are observed, confirming that this is a structural feature of the molecule. in the second structure, two differences are seen in the monomeric form. the conformation of the active site is different from those of the wild - type enzyme, in our structure, indicating that the effect of the mutations is not conserved. this is a key step in the catalysis of atp synthesis and subsequent transfer of atp to the catalytic site. furthermore, the absence of important residues suggests that residues in the second monomer are essential for correct positioning of the phosphate group. the conformational differences between the two structures may be related to its functional importance.

ProteinMPNN Full Text Multinomial 0.001 Gaussian Noise

in the quest to identify new potent inhibitors of protein tyrosine phosphatases, we previously reported two crystal structures of wild - type and mutated forms of the enzyme involving two mutations. the crystal structure of the c166s mutant of protein yscum in complex with the physiological substrates tyr and trp was solved at 1. 9 - a resolution. however, the precise mode of the enzyme - protein interactions are unknown. however, the precise interactions between yscum and wild - type were not conserved between both wild - type and wt. consequently, our data set is becoming nearly equivalent to the whole protein. in the wt protein, only three mg2 + ions have been shown to be required for efficient catalysis. while the ability of yscum is essential, the precise nature of the effect of this mutation is unknown. the precise nature of the contacts between yscum and wild - type enzyme has not been elucidated, and the structural basis for this and functional investigation has not been fully elucidated. these results provide a foundation for further mechanistic studies aimed at the future design of inhibitors as potential antipodial agents.

ProteinMPNN Full Text Multinomial 0.01 Gaussian Noise

in the present paper we report the 2. 0 a crystal structure of the universal sequence - selective tyrosine phosphorylase (cdp - plcp) from escherichia coli which crystallizes in a 1 : 1 stoichiometry. in the presence of atp and mg (2 +), the crystal structure reveals a prenylated molecule in the asymmetric unit. in the initial model, two copies of the protein are closely associated in the crystal, in a similar arrangement to the monoclinic crystal form. the exact conformation of the chalcone ring and the mode of atp binding are similar to those of the wild - type enzyme. in addition, in the crystal structure at 2. 4 - a resolution, all of the molecules make extensive contacts to the active site, indicating a novel mechanism for the operation of the physiological enzyme. the results provide a new foundation for the future design of new cdp - plcps inhibitors.

ProteinMPNN Full Text Multinomial 0.1 Gaussian Noise

in eukaryotes, the synthesis and signal (or gtp) signals by a conserved transmembrane motor domain (tmrec) that delivers the key lipids to the cell, is a key player in this process. purified recombinant full - length human recombinant full - length recombinant full - length recombinant recombinant recombinant full - length recombinant recombinant recombinant full - length recombinant full - length recombinant full - length recombinant full - length recombinant full - length recombinant full - length recombinant recombinant full - length recombinant full - length recombinant full - length recombinant full - length recombinant full - length recombinant full - length recombinant full - length recombinant recombinant full - length recombinant recombinant full - length recombinant full - length recombinant recombinant full - length recombinant full - length recombinant full - length recombinant full - length recombinant recombinant full - length recombinant recombinant full - length recombinant full - length recombinant recombinant full - length recombinant full - length recombinant recombinant full - length recombinant full - length recombinant recombinant recombinant full - length recombinant full - length recombinant full - length recombinant full - length recombinant full - length recombinant recombinant full - length recombinant full - length recombinant full - length recombinant recombinant full - length recombinant recombinant recombinant recombinant recombinant full - length recombinant full - length recombinant full - length recombinant full - length recombinant recombinant full - length recombinant recombinant recombinant full - length recombinant full - length recombinant full - length recombinant full - length recombinant full - length recombinant recombinant recombinant full - length rec

LigandMPNN-SC Greedy

the crystal structure of the pep

LigandMPNN-SC Multinomial

peptostreptoma toxin, from the gram - negative bacterium bacillus subtilis, encodes the three peptidomimerases that cleave peptidic substrates into peptidomimetic membranes. crystal structures of apo - peptidomimetic and the catalytic mechanism of three peptidomimetic peptidomimetic peptidomimetic inhibitors were solved to 2. 0 and 1. 5 a resolution, respectively. the peptidomimetic inhibitors were used in specific docking into the peptidomimetic peptidomimetic enzyme and compared to the peptidomimetic peptidomimetic inhibitors. the peptidomimetic inhibitors bind with its carbon - carbon group pointing into the active site. the peptidomimetic inhibitors in two structures were found to bind into the active site. the peptidomimetic inhibitors bind to the active site in a fashion similar to those previously reported for peptidomimetic inhibitors.

9.2 Sampling strategy for example proteins

PDB ID 9BKP

Text generated from greedy sampling strategy The crystal structures of a new crystal form of a new crystal form of a new crystal form are reported at 1. 5 a resolution. this is the first report of a study of a new crystal form of a diheme protein. this is the smallest known solution structure, and the smallest form reported to date. a simple four - residue disulfide bridge protein, such as disulfide bonds, is formed by the four - residue clusters in the 3 - fold region of a3b, and the polypeptide chain of the small and the 3 - fold symmetry - related molecule is found in the crystallographic asymmetric unit. the protein is more ' 'compact and more compact than the typical 3 - fold symmetry. the disulfide ' 'bonds are formed in the crystallographic asymmetric unit, but there are some ' 'minor differences in the positions of the disulfide bonds. the positions of ' 'the four residues in the protein molecule

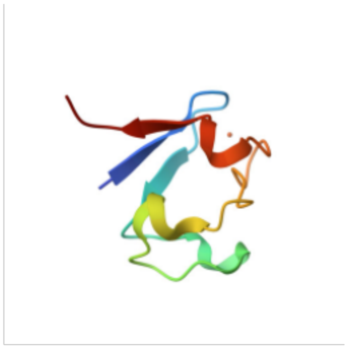
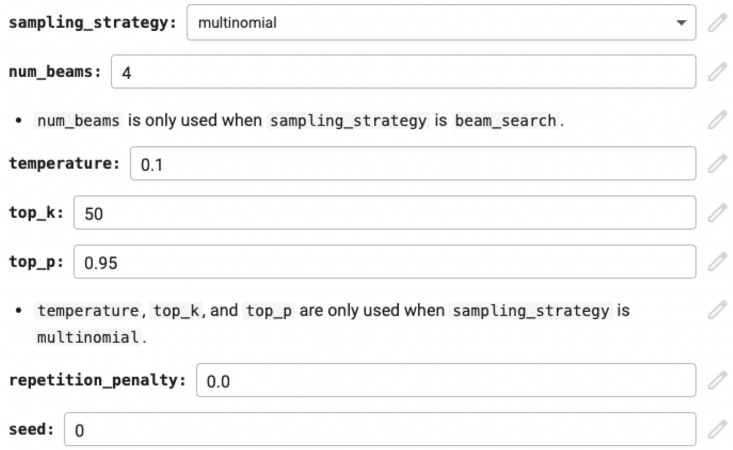


Figure 5: Protein structure for PDB ID 9BKP

are similar to those found in the ' 'crystal structures of other known crystal forms. however, there is no ' 'evidence of the pattern of conformational changes in the crystal as well as ' 'in solution. these changes have been found to be important in the crystal ' 'packing and to be caused by thermodynamic and kinetic studies of proteins.'

Multinomial sampling strategy setting



sampling_strategy: multinomial

num_beams: 4

- num_beams is only used when sampling_strategy is beam_search.

temperature: 0.1

top_k: 50

top_p: 0.95

- temperature, top_k, and top_p are only used when sampling_strategy is multinomial.

repetition_penalty: 0.0

seed: 0

Figure 6: Sampling stragety for PDB ID 9BKP

PDB ID 8U9G

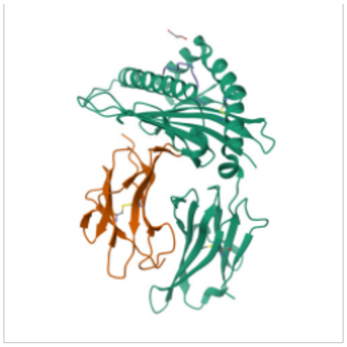
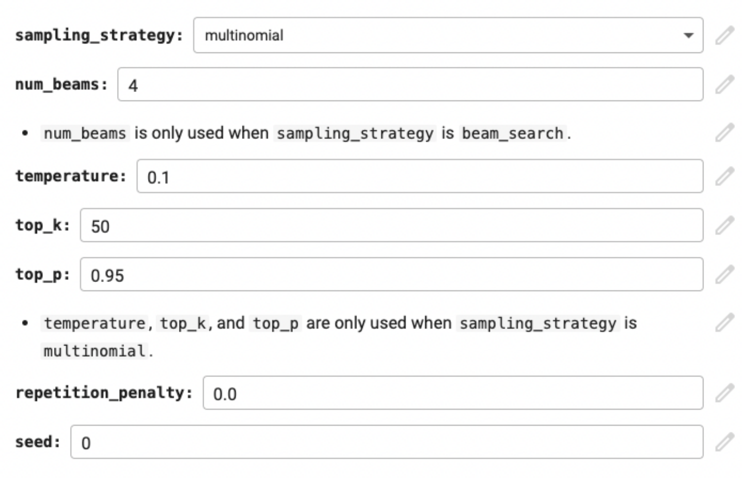


Figure 7: Protein structure for PDB ID 8U9G

Text generated from greedy sampling strategy ’the t cell receptor (tcr) of the major histocompatibility complex (mhc) ’ ’class i molecule hla - a2 is presented by the hla - a2 - restricted t cell ’ ’receptor (tcr). the tcr recognizes the hla - a2 - restricted peptide (hla ’ ’- a2) and the hla - a2 - restricted t cell receptor (tcr), which is ’ ’recognized by the hla - a2 - restricted tcr. the tcr recognizes the peptides ’ ’presented by hla - a2 and hla - a2. the tcr recognizes the peptide with a k ’ ’(d) of - fold greater affinity than the hla - a2 - peptide. the tcr ’ ’recognizes the peptide with a k (d) of - fold greater than that of hla - ’ ’a2. the tcr is accommodated by the peptide - mhc, which is accommodated in ’ ’the hla - a2 - peptide complex. the tcr recognizes the peptide - mhc, but ’ ’not the peptide - mhc, in the hla - a2 - peptide complex. the tcr recognizes ’ ’the peptide - mhc, but is not restricted in the hla - a2 - peptide complex. ’ ’the tcr recognizes the peptide - mhc, but is not restricted in the hla - a2 ’ ’- peptide - mhc complex. the tcr is accommodated in the hla - a2 - peptide - ’ ’mhc complex, where the mhc peptide - mhc complex binds with a k (d) of - ’ ’fold greater than that of hla - a2. the tcr - peptide - mhc complex is bound ’ ’in the hla - a2 - peptide - mhc complex, but the tcr - peptide - mhc complex ’ ’is bound in a similar orientation to the hla - a2 - peptide - mhc complex. ’ ’the tcr - peptide - mhc complex is bound in the hla - a2 - peptide - mhc ’ ’complex and is bound in a similar orientation to the hla - a2 - peptide - ’ ’mhc complex. the tcr - peptide - mhc complex is bound in a similar ’ ’orientation to the hla - a2 - peptide - mhc complex, but the tcr - peptide - ’ ’mhc complex is bound in a similar orientation to the hla - a2 - peptide - ’ ’mhc complex. the tcr - peptide - mhc complex is bound in a similar ’ ’orientation to the hla - a2 - peptide - mhc complex. the tcr - peptide - mhc ’ ’complex is bound in a’

Multinomial sampling strategy setting



sampling_strategy: multinomial

num_beams: 4

- num_beams is only used when sampling_strategy is beam_search.

temperature: 0.1

top_k: 50

top_p: 0.95

- temperature, top_k, and top_p are only used when sampling_strategy is multinomial.

repetition_penalty: 0.0

seed: 0

Figure 8: Sampling stragety for PDB ID 8U9G

References

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Justas Dauparas, Ivan Anishchenko, Nathaniel Bennett, Hua Bai, Robert J Ragotte, Lukas F Milles, Basile IM Wicky, Alexis Courbet, Rob J de Haas, Neville Bethel, et al. 2022. Robust deep learning–based protein sequence design using proteinmpnn. *Science*, 378(6615):49–56.

Justas Dauparas, Gyu Rie Lee, Robert Pecoraro, Linna An, Ivan Anishchenko, Cameron Glasscock, and D. Baker. 2023. Atomic context-conditioned protein sequence design using ligandmpnn. *bioRxiv*.

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

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.

Han Guo, Mingjia Huo, Ruiyi Zhang, and Pengtao Xie. 2023. Proteinchat: Towards achieving chatgpt-like functionalities on protein 3d structures. *Authorea Preprints*.

Chloe Hsu, Robert Verkuil, Jason Liu, Zeming Lin, Brian Hie, Tom Sercu, Adam Lerer, and Alexander Rives. 2022. Learning inverse folding from millions of predicted structures. In *International conference on machine learning*, pages 8946–8970. PMLR.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.

Chin-Yew Lin and Eduard Hovy. 2002. Manual and automatic evaluation of summaries. In *Proceedings of the ACL-02 workshop on automatic summarization*, pages 45–51.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

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

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

Chao Wang, Hehe Fan, Ruijie Quan, and Yi Yang. 2024. Protchatgpt: Towards understanding proteins with large language models. *arXiv preprint arXiv:2402.09649*.

Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. 2022. Git: A generative image-to-text transformer for vision and language. arXiv preprint arXiv:2205.14100.

Minghao Xu, Xinyu Yuan, Santiago Miret, and Jian Tang. 2023. Protst: Multi-modality learning of protein sequences and biomedical texts. In *International Conference on Machine Learning*, pages 38749–38767. PMLR.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.