SurfPro: Functional Protein Design Based on Continuous Surface

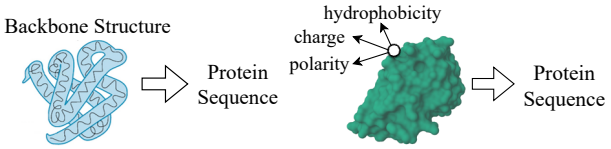
Zhenqiao Song [1] **Tinglin Huang** [2] **Lei Li** [1] **Wengong Jin** [3]

Abstract

How can we design proteins with desired functions? We are motivated by a chemical intuition that both geometric structure and biochemical properties are critical to a protein's function. In this paper, we propose SurfPro, a new method to generate functional proteins given a desired surface and its associated biochemical properties. SurfPro comprises a hierarchical encoder that progressively models the geometric shape and biochemical features of a protein surface, and an autoregressive decoder to produce an amino acid sequence. We evaluate SurfPro on a standard inverse folding benchmark CATH 4.2 and two functional protein design tasks: protein binder design and enzyme design. Our SurfPro consistently surpasses previous state-of-the-art inverse folding methods, achieving a recovery rate of 57.78% on CATH 4.2 and higher success rates in terms of protein-protein binding and enzyme-substrate interaction scores.

1. Introduction

Proteins serve diverse functions crucial to cellular processes in our biological system. In recent years, the remarkable achievements of generative AI have transformed the field of protein design (Huang et al., 2016; Rives et al., 2021; Watson et al., 2023). One prevalent approach involves first choosing or designing a target backbone structure and then identifies a sequence that folds into this backbone (Dauparas et al., 2022; Anishchenko et al., 2021; Wang et al., 2022; Yeh et al., 2023). The first step specifies the geometry of the desired protein (without amino acid types) and the second step (also known as inverse folding, as illustrated in Figure 1 (a)), determines the amino acid composition corresponding to the given shape.



(a) Inverse folding (b) Surface based design

Figure 1. Problem setups of protein design. (a) Inverse folding: protein design conditioned on geometric constraints only. (b) Surface based design: protein design conditioned on both geometric shape and biochemical properties.

However, the goal of protein design goes beyond predicting a sequence that folds into a target backbone (Defresne et al., 2021). The ultimate goal is to design proteins with desired functions, such as enzymes binding to specific substrates or proteins inhibiting given targets. The limitation of inverse folding is that it only specifies geometric constraints through the given backbone structure. To dictate the desired functions, we need to impose not only geometric constraints but also biochemical property constraints. For example, two proteins with complementary shapes may still not bind due to poorly placed charges, polarity, or hydrophobicity at their binding interface (Gainza et al., 2023).

To address this issue, we propose SurfPro, a method to design functional proteins given a biochemical property augmented point cloud (also called surface based design, Figure 1 (b)). Each point on the surface is labeled with a three-dimensional (3D) coordinate and a set of biochemical properties. SurfPro generates an amino acid sequence based on the surface's geometric shape and biochemical properties. SurfPro comprises a hierarchical encoder and an autoregressive decoder. The encoder progressively models the geometric and biochemical features of the surface through a series of local graph convolutions, followed by global self-attention layers that focus on modeling long-range interactions. The decoder generates a protein sequence based on the learned geometric and biochemical representations of the surface, with the goal that the generated sequence folds into the given surface.

Our contributions are listed as follows:

- We propose SurfPro to design functional proteins based on continuous surfaces augmented with biochemical properties.

[1]Language Technologies Institute, Carnegie Mellon University, Pittsburgh, the United States. [2]Yale University, New Haven, United States. [3]Broad Institute of MIT and Harvard, Boston United States. Correspondence to: Zhenqiao Song <zhenqiaosong@cmu.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

- We evaluate SurfPro on a standard inverse folding benchmark CATH 4.2. SurfPro achieves 57.78% sequence recovery rate and 3.13 perplexity, significantly outperforming previous inverse folding methods including ProteinMPNN (Dauparas et al., 2022), PiFold (Gao et al., 2022), and LM-DESIGN (Zheng et al., 2023).
- We setup a binder design task, and we use AlphaFold2 (Jumper et al., 2021) pAE_interaction (Bennett et al., 2023; Watson et al., 2023) to evaluate the binding of designed proteins. SurfPro exhibits superior capability of designing binders with stronger interaction with target proteins than experimentally confirmed positive binders, with an average success rate of 26.22% across six targets, outperforming the best prior method by 6.9%.
- We setup an enzyme design task, and we use ESP score (Kroll et al., 2023a;b) to measure the binding between designed enzymes and their substrates. SurfPro is able to design enzymes with higher enzyme-substrate interaction scores than natural enzymes, achieving an average success rate of 43.46% across five enzyme datasets, outperforming the best prior method by 2.98%.

2. Related Work

Methods for Protein Sequence Design. Protein sequence design has been studied with a wide variety of methods. Most studies in this field have adopted one of the three main paradigms: (i) guided by fitness landscape (function scores), (ii) conditioning on a fixed backbone structure, and (iii) finetuned from a pretrained model on large-scale data.

Protein sequence design guided by fitness landscape includes traditional directed evolution (Arnold, 1998; Dalby, 2011; Packer & Liu, 2015; Arnold, 2018) and machine learning methods. The mainly used machine learning algorithms include reinforcement learning (Angermueller et al., 2019; Jain et al., 2022), Bayesian optimization (Moss et al., 2020; Terayama et al., 2021) and search using deep generative models (Brookes & Listgarten, 2018; Brookes et al., 2019; Kumar & Levine, 2020; Ren et al., 2022; Song & Li, 2023). Protein design based on a fixed backbone structure is also called inverse folding (Fleishman et al., 2011; Ingraham et al., 2019; Hsu et al., 2022; Gao et al., 2022; Zheng et al., 2023), which ensures the preservation of a stable structure while also allows for the sampling of diverse sequences. Due to the available massive sequence data, recent studies have successfully employed machine learning models pretrained on such data, including ESM (Rives et al., 2021), ProtGPT2 (Ferruz et al., 2022) and ProGen (Madani et al., 2023). This approach has led to notable advancements in addressing various downstream tasks within the field of protein design, such as protein mutation (Meier et al., 2021) and structure-informed protein design (Zheng et al., 2023).

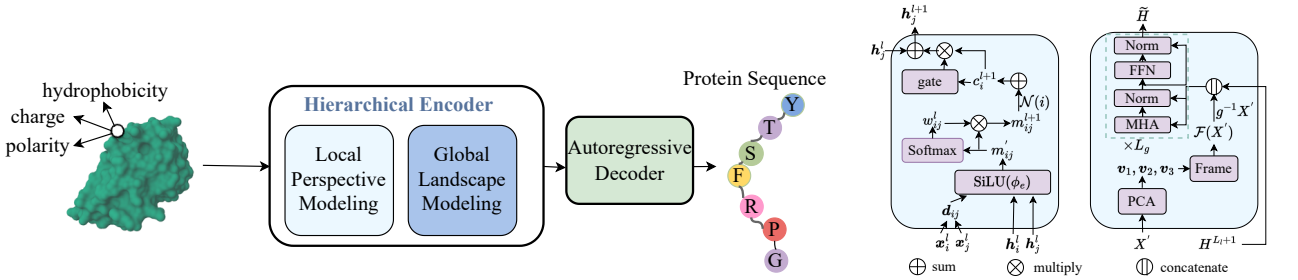
Protein Surface Modeling. Protein design based on its surface is an under-explored area, and most existing studies have not taken into account the significant role of protein molecular surface plays in various biological processes. Traditionally, molecular surfaces are defined using Connelly surfaces (Connolly, 1983; Sanner et al., 1996) based on van der Waals radii, often represented as mesh-based structures derived from signed distance functions. Seminal work for modeling protein molecular surfaces is MaSIF (Gainza et al., 2020), which fingerprints molecular surfaces expressed as molecular meshes using pre-defined and pre-calculated physical and geometrical features. To remove the high pre-computation costs of featurization, Sverrisson et al. (2021) propose dMaSIF, showing that modeling molecular surface as a point cloud with atom categories per point is competitive to mesh-based methods. However, both works target at protein understanding tasks, such as protein-protein interactions, instead of protein design. Gainza et al. (2023) then expands MaSIF to enable de novo binder design. Initially, they utilize the generated surface fingerprints to predict target binding sites. Following this, they search for binders containing complementary structural motifs, which are subsequently transplanted to protein scaffolds. In our work, we develop a generative model to directly generate functional proteins from their surfaces, eliminating the need for handcrafted feature calculation.

3. Proposed Method: SurfPro

A molecular surface defines the shape of a protein in 3D Euclidean space and the biochemical properties, such as hydrophobicity and charge. The surface shape and the associated biochemical properties co-determine the underlying protein functions. Given a desired surface with geometric and biochemical constraints, how can we generate protein sequences fitting the surface? In this section, we introduce SurfPro, a new functional protein design method based on protein surfaces. Our method works on successive geometric representations of a protein. SurfPro consists of a hierarchical encoder that progressively model the 3D geometric shape and the biochemical features from a local perspective to a global landscape, and an autoregressive decoder that generates a protein sequence based on the geometric and biochemical constraints of the corresponding surface. Figure 2 (a) gives an overview of SurfPro.

3.1. Surface Generation

Our method works on successive point clouds of protein surfaces. A high-quality surface should satisfy the following two properties: (1) Smooth: The surface defined by the point cloud should exhibit sufficient smoothness; (2) Compact:



(a) model overview

(b) left: local modeling, right: global modeling

Figure 2. (a) The overview of our proposed SurfPro. (b) Left: local perspective modeling, right: global landscape modeling.

The point cloud should remove redundant information by down-sampling to improve efficiency.

Raw Surface Construction. We use MSMS (Ewing & Hermisson, 2010) to compute the raw molecular surface of a protein, which is provided as a point cloud with N vertices $\{\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N\} \in \mathbb{R}^{N\times 3}$. Suppose the protein is a L-residue sequence $y = \{y_1, y_2, ..., y_L\} \in \mathcal{A}^L$ where $\mathcal{A}$ is the set of 20 common amino acids and $N \gg L$. We associate the biochemical features of each vertex to its nearest atom belonging to one of the L residues. Specifically, we utilize two biochemical features for each vertex $\boldsymbol{x_i}$, which are its hydrophobicity t_i and charge c_i. Then we sort all vertices based on the residue index of their nearest atoms. Appendix Figure 6 (a) depicts an example of raw surface.

Surface Smoothing. As mentioned in previous methods (Alexa et al., 2001; Lv et al., 2021), raw point clouds generally carry noise, which may limit the expressivity of the molecular surface. Therefore, point cloud denoising and smoothing are necessary. We apply Gaussian kernel smoothing on raw point cloud data:

$$\boldsymbol{x}_i^{'} = \sum_{\boldsymbol{x}_j \in \mathcal{N}(\boldsymbol{x}_i)} \frac{\mathcal{K}(\boldsymbol{x}_i, \boldsymbol{x}_j)\boldsymbol{x}_j}{\sum_{\boldsymbol{x}_t \in \mathcal{N}(\boldsymbol{x}_i)} \mathcal{K}(\boldsymbol{x}_i, \boldsymbol{x}_t)}, \quad \mathcal{K}(x, y) = e^{-\frac{(x-y)^2}{\eta}} \tag{1}$$

where $i \in \{1, 2, ..., N\}$. $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ denote the coordinates of i-th and j-th vertices on the raw point cloud, respectively. $\mathcal{N}(\boldsymbol{x}_i)$ are K-nearest neighbors of $\boldsymbol{x}_i$. $\mathcal{K}(\cdot, \cdot)$ is the Gaussian kernel with η indicating distance scale in the point space. In our paper, we set $\eta = \max(\text{dist}(\boldsymbol{x}_i, \mathcal{N}(\boldsymbol{x}_i)))$ where $i \in \{1, ..., N\}$. The number of nearest neighbors K is set to 8. An in-depth analysis of kernel smoothing supports that the surface is infinitely smooth, i.e., $\theta \in C^\infty$ (Levin, 1998; 2004). Appendix Figure 6 (b) depicts an example of smoothed surface.

Surface Compression. To reduce surface points and improve sample efficiency, we use an octree-based compression method to down-sample a protein surface (Schnabel & Klein, 2006). We use an octree to convert the surface into small cubes and estimate local densities of each cube. Every octree node is recursively divided into eight equal octants. After each division, the number of points in each node is examined to determine whether or not to continue dividing the current node. The cubes with fewer points than the specific threshold N_{min} are taken as leaf nodes and not divided further. After all the nodes are processed, a point cloud is converted into a number of unequal-volume cubes based on the point distribution. Lower density regions result in larger cubes. The desired number of points for each cube is $N_s = V_s * r$, where V_s is the number of points in the s-th cube and r is the desired down-sampling ratio. Figure 6 (c) depicts an example of compressed surface.

3.2. Hierarchical Surface Encoder

We design a hierarchical encoder to model the geometric shape and biochemical properties of a protein surface.

Local Perspective Modeling. Residues nearing each other exhibit strong interactions. To model such interactions among nearest vertices on a surface, we design a variant of equivariant graph convolutional layer (EGCL) proposed by Satorras et al. (2021) to capture local geometric and biochemical features (Figure 2 (b) left module). Specifically, after surface compression, the surface has $N^{'}$ vertices ($N^{'} \leq N$), each of which has a 3D coordinate $\boldsymbol{x}_i^{'} \in \mathbb{R}^3$ and two biochemical features $\boldsymbol{h}_i = [t_i, c_i]^T$ where t_i denotes its hydrophobicity, c_i denotes its charge and $i \in \{1, ..., N^{'}\}$. We calculate local messages as:

$$\begin{aligned} \boldsymbol{m}_{ij}^{'} &= \text{SiLU}(\phi_e([(\boldsymbol{h}_i^l; \boldsymbol{h}_j^l; ||\boldsymbol{x}_i^{'} - \boldsymbol{x}_j^{'}||_2)])) \\ w_{ij}^l &= \frac{\exp(W_s^l \boldsymbol{m}_{ij}^{'} + b_s^l)}{\sum_{k \in \mathcal{N}(\boldsymbol{x}_i)} \exp(W_s^l \boldsymbol{m}_{ik}^{'} + b_s^l)} \\ \boldsymbol{m}_{ij}^{l+1} &= w_{ij}^l * \boldsymbol{m}_{ij}^{'} \end{aligned} \tag{2}$$

where vertex $j \in \mathcal{N}(\boldsymbol{x}_i)$ belongs to the K-nearest neighbors of vertex i. We set $K = 30$ here. $\boldsymbol{h}_i^1 = W_m \boldsymbol{h}_i$ where $W_m \in \mathbb{R}^{256\times 2}$ is a mapping matrix and $l = 1$ to L_l is the layer number for the local perspective modeling module. $W_s^l \in \mathbb{R}^{1\times 256}$ and $b_s^l \in \mathbb{R}$ are learnable parameters. [;] denotes concatenation operation. ϕ_e denotes multi-

layer perceptron (MLP). SiLU denotes SiLU activation function (Elfwing et al., 2018). For each vertex, we propagate messages from its neighbors to update the node feature:

$$\begin{aligned} \boldsymbol{c}_i^{l+1} &= \sum\nolimits_{j \in \mathcal{N}(\boldsymbol{x}_i)} \boldsymbol{m}_{ij}^{l+1} \\ \boldsymbol{h}_i^{l+1} &= \boldsymbol{h}_i^{l} + \mathrm{gate}(\boldsymbol{c}_i^{l+1}) \odot \boldsymbol{c}_i^{l+1} \end{aligned} \tag{3}$$

where gate is a gating mechanism achieved by a MLP followed by a sigmoid function, which is used to control the information flow over the local geometric shapes.

Global Landscape Modeling. To facilitate message passing over the whole desired surface, we design a global landscape encoder called FAMHA (Figure 2 (b) right module). Its key idea is to incorporate the frame averaging technique (FA, (Puny et al., 2021)) into a multi-head attention layer. The resulting operation not only enables the global biochemical features to spread out but also guarantees its SE(3) equivariance. Specifically, from the compressed point cloud $\boldsymbol{X}^{'} \in \mathbb{R}^{N^{'} \times 3}$, we calculate three principle components $\boldsymbol{v}_1$, $\boldsymbol{v}_2$, $\boldsymbol{v}_3 \in \mathbb{R}^3$ through principle component analysis (PCA). With these three base coordinates, we define a frame $\mathcal{F}(\boldsymbol{X}^{'})$ as a function:

$$\mathcal{F}(\boldsymbol{X}^{'}) = \{([\alpha_1 \boldsymbol{v}_1, \alpha_2 \boldsymbol{v}_2, \alpha_3 \boldsymbol{v}_3], \boldsymbol{t}) | \alpha_i \in \{-1, +1\}\} \tag{4}$$

where $\boldsymbol{t}$ is the centroid of $\boldsymbol{X}^{'}$. The frame function forms an algebraic group of eight transformations. We calculate the global message passing as follows:

$$\widetilde{\boldsymbol{H}} = \frac{1}{|\mathcal{F}(\boldsymbol{X}^{'})|} \sum_{g \in \mathcal{F}(\boldsymbol{X}^{'})} \mathrm{FAMHA}(\boldsymbol{H}^{L_l+1}; g^{-1}\boldsymbol{X}^{'}) \tag{5}$$

where $\boldsymbol{H}^{L_l+1} = [\boldsymbol{h}_1^{L_l+1}, ..., \boldsymbol{h}_{N^{'}}^{L_l+1}]^T \in \mathbb{R}^{N^{'} \times 256}$ is the output vertex features from the local perspective modeling. $g^{-1}\boldsymbol{X}^{'}$ denotes translating $\boldsymbol{X}^{'}$ with $\boldsymbol{t}$ and rotating $\boldsymbol{X}^{'}$ with rotation matrix $[\alpha_1 \boldsymbol{v}_1, \alpha_2 \boldsymbol{v}_2, \alpha_3 \boldsymbol{v}_3]$ for $g \in \mathcal{F}(\boldsymbol{X}^{'})$. FAMHA is composed of L_g stacked multi-head attention (MHA) sub-layers and fully connected feed-forward networks (FFN). A residue connection and a layer normalization are performed after each of the two sub-layers. Accordingly, the FAMHA can be formulated as follows:

$$\begin{aligned} \boldsymbol{h}_i^{l+1} &= \mathrm{LayerNorm}\left(\mathrm{FFN}(\tilde{\boldsymbol{h}}_i^{l}) + \tilde{\boldsymbol{h}}_i^{l}\right), \\ \tilde{\boldsymbol{h}}_i^{l} &= \mathrm{LayerNorm}\left(\mathrm{MHA}(\boldsymbol{h}_i^{l}, \boldsymbol{H}_g^{l}) + \boldsymbol{h}_i^{l}\right) \end{aligned} \tag{6}$$

where $l \in \{1, ..., L_g\}$ and $i \in \{1, ..., N^{'}\}$. $\boldsymbol{h}_i^1 = [\boldsymbol{h}_i^{L_l+1}; g^{-1}\boldsymbol{X}^{'}]$ for $g \in \mathcal{F}(\boldsymbol{X}^{'})$. $\boldsymbol{H}_g^l = [\boldsymbol{h}_1^l, ..., \boldsymbol{h}_{N^{'}}^l]^T$.

3.3. Autoregressive Protein Decoder

Given the hidden representations encoding both geometric shapes and biochemical features, we use an autoregressive Transformer decoder (i.e., GPT, (Vaswani et al., 2017)) to generate the protein sequence for a given surface:

$$p(y_t) = \mathrm{TransDec}(y_{<t}, \widetilde{\boldsymbol{H}}; \theta_{dec}) \tag{7}$$

where $p(y_t)$ is the probability of t-th residue in the protein sequence and θ_{dec} denotes the learnable parameters.

We minimize the negative log likelihood to train the overall model:

$$\mathcal{L} = \sum_{t=1}^{L} -\log p(y_t; \theta) \tag{8}$$

where $\theta = \{\theta_{enc}, \theta_{dec}\}$ denotes the parameter set of our hierarchical encoder and autoregressive protein decoder.

4. Experiments

In this section, we first describe our implementation details in Section 4.1. Then we conduct extensive experiments and evaluate our proposed SurfPro on one general protein design task, **Inverse Folding** (Section 4.2) and two functional protein design tasks, **Binder Design** (Section 4.3) and **Enzyme Design** (Section 4.4). The specific experimental settings are introduced in each task section.

4.1. Implementation Details

We set a maximum limit of $5,000$ vertices for each surface. Surfaces with fewer than $5,000$ vertices remain unchanged, while those exceeding this limit are compressed with a down-sampling ratio r set to $5,000/N$, where N denotes the original vertex count. The minimum vertex number in a cube N_{min} in surface compression is set to 32. Local perspective modeling utilizes three layers, and global landscape modeling employs a two-layer FAMHA. The two biochemical features are mapped to a hidden space with a dimensionality of 256. The autoregressive decoder is built with 3-layer Transformer decoder. The mini-batch size and learning rate are set to $4,096$ tokens and 5e-4, respectively. The model, trained with one NVIDIA RTX A6000 GPU card, utilizes the Adam optimizer (Kingma & Ba, 2014). The detailed values for biochemical features are provided in Appendix Table 9.

4.2. Inverse Folding

This task is to design protein sequences that fold into given backbone structures. In our method, we design a protein sequence from a coarser structure – a protein surface, instead of a rigid backbone structure.

Datasets. Following previous work (Dauparas et al., 2022; Gao et al., 2022), we use the CATH 4.2 dataset curated by Ingraham et al. (2019) and follow the same data splits of Jing et al. (2020). Due to the occasional failures in raw surface construction by the MSMS tool, we filter out these

Methods	Perplexity (↓)	Recovery Rate (%, ↑)
ProteinMPNN	5.19	44.95
PiFold	4.88	52.61
LM-DESIGN	4.47	54.16
SurfPro	**3.13**	**57.78**

Table 1. Perplexity and Recovery Rate of different approaches on CATH 4.2 dataset. (↑): the higher the better. (↓): the lower the better. Among all the baselines, SurfPro achieves the highest recovery rate and the lowest perplexity.

instances as well as proteins longer than $1,024$ residues. As a consequence, the training, validation, and test splits consist of 14525, 468, and 887 samples, respectively. We use the same data splits for all models rigorously for a fair comparison. The vertex count statistics for the curated CATH 4.2 dataset is provided in Appendix Table 10.

Baseline Models. We compare with the following baseline models: (1) **ProteinMPNN** is a representative inverse folding model. (2) **PiFold** and (3) **LM-DESIGN** are state-of-the-art-methods for inverse folding task. The used architecture for LM-DESIGN is LM-DESIGN (pretrained ProteinMPNN-CMLM: fine-tune). We use all their released codes on GitHub and the experimental settings in their official implementations to ensure a fair comparison.

Evaluation Metrics. Following previous work (Jing et al., 2020; Gao et al., 2022), we use **perplexity** and **recovery rate** to evaluate the quality of designed protein sequences. Since the surface does not include residues buried beneath it, we report the recovery rate after pairwise alignment for all autoregressive models to ensure a fair comparison:

$$\text{recovery rate} = \frac{\text{number of recovered residue}}{\text{aligned sequence length}} \tag{9}$$

We provide the recovery rates after pairwise alignment for all models in Appendix Table 11, wherein the non-autoregressive models consistently exhibit lower recovery rates compared to the ones before alignment.

Main Results. Table 1 shows that SurfPro achieves the highest recovery rate and the lowest perplexity among all the compared baselines. These findings demonstrate that incorporating both geometric and biochemical constraints of protein surfaces is beneficial for general protein design, leading SurfPro to achieve the highest recovery rate across diverse protein folds in CATH 4.2 dataset.

4.3. Protein Binder Design

In this section, we aim to use SurfPro to design proteins that bind to a target protein with high affinity.

Function Evaluator. Following previous work (Bennett et al., 2023; Watson et al., 2023), we use the **AlphaFold2** (AF2) **pAE_interaction** to evaluate the binding affinities between the designed binders and target proteins. Bennett et al. (2023) discover that AF2 pAE_interaction is very effective in distinguishing experimentally validated binders from non-binders, with a success rate ranging from 1.5% to 7% on target protein IL7Ra, TrkA, InsulinR and PDGFR. We use the official code of Bennett et al. (2023) to calculate the AF2 pAE_interaction. The lower the AF2 pAE_interaction, the better the designed binder.

Evaluation Metrics. We calculate the average **AF2 pAE_interaction** for the entire test set using greedy decoding, and the average **success rate**. For each positive binder, the success rate is defined as the proportion of designed binders with a lower pAE_interaction than it. For each <positive binder, target protein> pair, we generate 10 new binders using sampling with a temperature=0.1. To calculate the pAE_interaction, we first use ESMFold (Lin et al., 2023) to predict the structure of the designed binder sequence, and then we superimpose this structure to the real complex. Finally we calculate the AF2 pAE_interaction for the new complex. As the AF2 pAE_interaction model will automatically revise the input complex structures, there is little difference between AlphaFold2 and ESMFold predicted binder structures. A comparison is provided in Appendix Table 14.

Datasets. We collect experimentally confirmed positive complexes of <binder, target protein> pairs across six categories from Bennett et al. (2023). Among the 10 categories available, 4 exhibit indistinguishable AF2 pAE_interaction between negative and positive binders. Therefore, we choose to evaluate on the remaining 6 categories for a reliable evaluation. For categories with over 50 complexes, we employ an $8:1:1$ random split for training, validation, and test sets; otherwise, all complexes are included in the test set, establishing a zero-shot scenario. Detailed data statistics is provided in Appendix Table 12.

Baseline Models. Using the binder design dataset, we finetune all baseline models (ProteinMPNN, PiFold, LM-DESIGN) and our SurfPro that are respectively pretrained on the CATH 4.2 dataset as detailed in Section 4.2. In addition, we also provide a random baseline by randomly mutating 20% residues of a binder. Furthermore, to leverage all available crystal structures and fully explore the design capability of our SurfPro, we pretrain the model using all surfaces generated from the entire Protein Data Bank (PDB). This pretraining dataset, collected up to March 10, 2023, comprises $179,278$ <surface, sequence> pairs. Detailed data preprocessing steps and pretraining details are provided in the Appendix E. We then finetune this model on the binder design dataset, and we refer to the resulting model as SurfPro-Pretrain.

Main Results. The AF2 pAE_interaction and success rate

Models	Seen Class			Zero-Shot			Average
	InsulinR	PDGFR	TGFb	H3	IL7Ra	TrkA	
Positive Binder	**5.9996**	**14.1366**	**15.4884**	**21.2631**	**20.9102**	**10.2791**	**14.7061**
Negative Binder	19.7167	18.0937	23.2664	22.4556	26.0540	24.7567	21.1335
Random Baseline	19.9880	21.2690	21.4971	24.4997	24.1541	23.1147	22.2020
ProteinMPNN (Dauparas et al., 2022)	18.3393	25.2919	25.8559	24.5968	25.5278	27.0980	23.4462
PiFold (Gao et al., 2022)	12.9809	21.8230	24.4737	23.3924	26.6738	19.7172	20.5785
LM-DESIGN (Zheng et al., 2023)	13.6440	22.0749	23.3725	23.8332	24.3937	22.3987	20.7728
SurfPro	**10.2608**	**17.9862**	**17.7364**	**21.2916**	**20.8594**	**10.6535**	**16.9485**
SurfPro-Pretrain	11.2530	18.4141	**15.4011**	22.2704	**20.5700**	21.3515	17.6699

Table 2. AF2 pAE_interaction (↓) for all models in the binder design task. "Average" denotes the average AF2 pAE_interaction across the entire test set instead of the direct average on different target proteins. We also provide the AF2 pAE_interaction for randomly sampled negative binders of the same length as the positive ones. Our SurfPro outperforms all previous methods on AF2 pAE_interaction.

Models	Seen Class			Zero-Shot			Average
	InsulinR	PDGFR	TGFb	H3	IL7Ra	TrkA	
ProteinMPNN (Dauparas et al., 2022)	3.22	5.71	**20.71**	18.68	24.10	7.50	11.96
PiFold (Gao et al., 2022)	20.64	3.57	19.19	**29.21**	22.85	20.00	19.32
LM-DESIGN (Zheng et al., 2023)	7.74	15.00	15.71	22.29	**24.28**	**25.00**	16.37
SurfPro	**31.57**	**19.99**	11.61	23.21	19.28	**25.00**	**22.29**
SurfPro-Pretrain	5.48	**27.14**	**33.57**	**37.63**	**38.57**	**25.00**	**26.22**

Table 3. Success rate (%, ↑) for all models in the binder design task. "Average" denotes the average success rate across the entire test set instead of the direct average on different target proteins. Our SurfPro-Pretrain outperforms all previous methods in success rate by a big margin.

results of different models are reported in Table 2 and Table 3, respectively. **The results show that our SurfPro achieves the lowest average AF2 pAE_interaction and the highest average success rate across six target proteins.** In particular, our SurfPro achieves the best pAE_interaction on all six categories and the highest success rate in three out of six categories. The pAE_interaction of our model is even slightly lower than the experimentally confirmed positive binders on IL7Ra. These findings demonstrate that leveraging protein surface properties is effective for functional binder design. Furthermore, SurfPro achieves the highest success rates in two zero-shot testing categories, affirming its capability to directly capture valuable protein properties from surfaces. Consequently, even without specific training on binder proteins binding to particular targets, SurfPro can generate binders in those categories with lower pAE_interaction than the positive ones. After pretraining on the entire PDB, the functions of binders generated by greedy decoding show minor difference. However, the success rate is improved significantly, from 22.29% to 26.22%. It demonstrates that pretraining on larger dataset helps to improve the design capability of our SurfPro, ensuring that more binders with better pAE_interaction can be designed.

4.4. Enzyme Design

In our work, we target at designing enzymes which bind to specific substrates.

Function Evaluator. To evaluate the binding affinity between enzyme and substrate, we use the **ESP score** developed by Kroll et al. (2023a). Their model predicts enzyme-substrate interaction with 91% accuracy across multiple benchmarks. We use their official code to calculate ESP score.

Evaluation Metrics. Similar to binder design, we report the average **ESP score** using greedy decoding and average **success rate** using sampling with temperature=0.1.

Datasets. We collect five categories of enzymes from Kroll et al. (2023a), each of which binds to a specific substrate. We exclude enzymes in CATH 4.2 to prevent data leakage issue. For enzyme categories containing more than 100 samples, we randomly split the data into training, validation, and test sets using an $8:1:1$ ratio after clustering; otherwise, all data are taken as the test set. The detailed data statistics is provided in Appendix Table 13.

Baseline Models. Similar to binder design, we finetune all baseline models (ProteinMPNN, PiFold, LM-DESIGN) and our SurfPro, using the enzyme design dataset, on the pretrained models from inverse folding task, respectively.

Models	Seen Class				Zero-Shot	Average
	C00002	C00677	C00019	C00003	C00001	
Real Enzyme	0.9573	0.8642	0.4497	0.8076	0.9892	0.9091
Random Baseline	0.5523	0.2475	0.1673	0.4705	0.7891	0.5292
ProteinMPNN (Dauparas et al., 2022)	**0.9711**	0.7375	0.2614	0.6699	0.9763	0.8676
PiFold (Gao et al., 2022)	0.9142	0.8816	0.4296	**0.8212**	0.9616	0.8865
LM-DESIGN (Zheng et al., 2023)	0.9498	0.8836	**0.4585**	0.8078	0.9650	**0.9037**
SurfPro	0.9264	**0.8921**	0.3892	0.7631	**0.9772**	0.8931
SurfPro-Pretrain	**0.9376**	0.8631	**0.3949**	**0.7668**	0.9691	0.8900

Table 4. ESP score (↑) for all models by greedy decoding in enzyme design task. The substrate is denoted as its KEGG database ID. "Average" here denotes the average ESP score across the entire test set instead of the direct average on different enzyme categories.

Models	Seen Class				Zero-Shot	Average
	C00002	C00677	C00019	C00003	C00001	
ProteinMPNN (Dauparas et al., 2022)	47.54	31.63	58.82	44.72	27.65	39.23
PiFold (Gao et al., 2022)	**48.54**	41.72	58.29	37.54	24.97	40.65
LM-DESIGN (Zheng et al., 2023)	45.00	42.54	43.76	**53.63**	20.13	37.58
SurfPro	43.36	**46.00**	**59.41**	45.45	**33.55**	**42.23**
SurfPro-Pretrain	**50.90**	41.81	52.94	36.36	**34.21**	**43.63**

Table 5. Success rate (**%**, ↑) for all models in the enzyme design task. "Average" denotes the average success rate across the entire test set instead of the direct average on different enzyme categories. The substrate is denoted as its KEGG database ID. Notice that our SurfPro achieves the highest average success rate.

Methods	Perplexity (↓)	Recovery Rate (%, ↑)
SurfPro	**3.13**	**57.78**
– w five	3.15	56.78
– w/o global	15.97	15.51
– w/o local	7.40	34.70
– w hydrophobicity	4.12	51.60
– w charge	11.23	21.34
– w/o feature	17.36	6.45
– w unsorted	18.98	14.95

Table 6. Ablation study on CATH 4.2 dataset. (↑): the higher the better. (↓): the lower the better.

Likewise, we also provide the results of a random baseline and SurfPro-Pretrain with the same settings as binder design. Note that the pretraining dataset excludes all enzymes here to prevent data leakage issue.

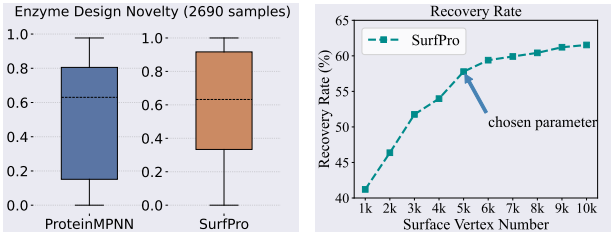
Main Results. Table 4 and Table 5 show that **our SurfPro achieves the highest average success rate and comparable average ESP score to LM-DESIGN across five categories.** It is important to note that LM-DESIGN is finetuned from the 650M ESM-1b (Rives et al., 2021), which is pretrained on the extensive UniRef50 dataset. Consequently, there is a potential for data leakage, enabling it to achieve the best performance on average ESP score. However, our SurfPro outperforms LM-Design with a significantly higher success rate of 42.23% compared to LM-Design's 37.58%. This performance is further improved to 43.63% after pretraining on the entire PDB surfaces. These findings indicate that our SurfPro is able to design enzymes with stronger enzyme-substrate interaction functions than real enzymes, validating that surface properties are helpful for functional protein design again. Furthermore, our SurfPro demonstrates zero-shot design capability, displaying a success rate of 33.55% for designing enzymes binding to substrate C00001.

5. Analysis: Diving Deep into SurfPro

5.1. Ablation Study: How Does Each Component Work?

Both geometric and biochemical constraints facilitate protein design. To better analyze the influence of different components in our model, we conduct ablation tests on inverse folding task. The models to be compared are listed as follows: (1) SurfPro-w-five uses five biochemical features, which are hydrophobicity, charge, polarity, acceptor and donor; (2) SurfPro-w/o-global removes the global landscape modeling; (3) SurfPro-w/o-local removes the local perspective modeling; (4) SurfPro-w-hydrophobicity only uses hydrophobicity feature; (5) SurfPro-w-charge only uses charge feature; (6) SurfPro-w/o-feature does not use any biochemical features; (7) SurfPro-w-unsorted does not sort the vertices on the raw surface.

The results in Table 6 indicate that incorporating five chemi-



(a) enzyme novelty (b) surface vertex number effect

Figure 3. (a) Novelty (1-recovery rate) of designed enzymes with temperature=0.1. (b) Recovery rates on CATH 4.2 dataset for models with varying numbers of sampled surface vertices.

Methods	Perplexity (↓)	Recovery Rate (%, ↑)
SurfPro	**3.13**	**57.78**
– w NAD	12.51	9.61

Table 7. Ablation study on CATH 4.2 dataset. (↑): the higher the better. (↓): the lower the better. NAD denotes non-autoregressive decoder.

cal features does not yield additional benefits. Removal of either global landscape modeling or local perspective modeling results in significant performance degradation. Utilizing only the hydrophobicity feature slightly decreases the performance, while relying solely on the charge feature seriously damages performance. The absence of both biochemical features further decreases performance. These observations validate the crucial roles played by both geometric shapes and biochemical features in surface representation learning, emphasizing the necessity of incorporating both into the protein design process. It is worth noting that unsorting the vertices on the raw surface significantly decreases performance. Our interpretation is the local shape in different areas might be similar, and the model is hard to align each local shape to a specific protein fragment without sorting the vertices, especially for extremely long sequences.

5.2. Ablation Study: Can Autoregressive Decoder Be Replaced?

The autoregressive decoder demonstrates strong performance within our proposed SurfPro architecture. To assess its significance, we compare SurfPro with an alternative employing a non-autoregressive decoder, incorporating SoftCopy and glancing learning strategies inspired by non-autoregressive machine translation in natural language processing (Gu & Kong, 2021; Qian et al., 2021). Table 7 presents the comparative results. It unequivocally demonstrates that SurfPro outperforms the non-autoregressive decoder variant, affirming the efficacy of the autoregressive decoder within our proposed framework.

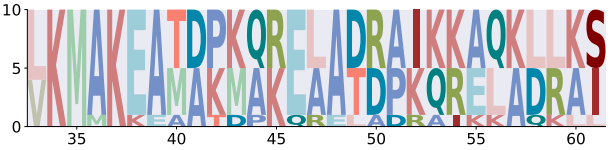


Figure 4. Amino acid distribution at different positions in designed binders for the target protein InsulinR.

5.3. Can SurfPro Design Novel and Diverse Proteins?

SurfPro is able to generate novel and diverse proteins. The novelty distribution among the designed enzymes, sampled with a temperature of 0.1, is visualized in Figure 3 (a). Novelty here is calculated as 1 - recovery rate. The figure shows our SurfPro demonstrates a superior average novelty of 58.51% in comparison to ProteinMPNN, which achieves an average novelty of 49.46%. Additionally, we analyze the amino acid distribution in designed binders across different positions. An illustrative example is shown for target protein InsulinR in Figure 4, demonstrating a diversified residue distribution. Notably, all these designed binders (provided in Appendix F.2) achieve a pAE_interaction lower than 10, a threshold identified by Bennett et al. (2023) to significantly enhance success rate. These findings affirm that SurfPro is able to design diverse proteins with desired functions.

5.4. How Does Vertex Number Affect Protein Design?

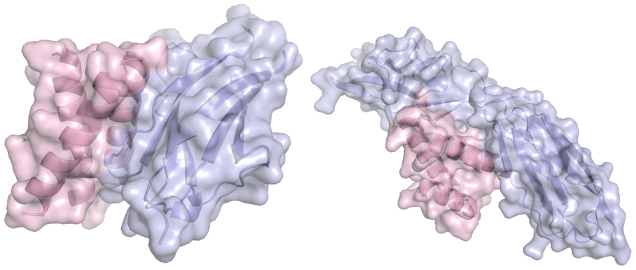
The more vertices sampled, the more effectively SurfPro performs. To thoroughly explore the impact of surface vertex sampling size on model performance, we train models with maximum vertex numbers ranging from 1k to 10k on CATH 4.2 dataset. Figure 3 (b) demonstrates an improved recovery rate as the number of sampled vertices increases. However, after reaching a sampling size exceeding 5k, the improvement rate slightly decreases. Additionally, with more vertices sampled, the inference speed will decrease. To ensure both design quality and inference efficiency, we set the maximum vertex number to 5k.

5.5. Comparing with MaSIF

In this section, we conduct a comparative analysis between our SurfPro and the well-established surface-based binder design model MaSIF (Gainza et al., 2020). Initially, we employ ProteinMPNN to generate 100 binder sequences for each positive binder. Subsequently, ESMFold is utilized to predict the structures of these designed binder sequences. Following this, MaSIF is employed to rank these candidate binders, with the top-performing one selected as the final designed binder. The resulting pAE_interaction scores across six datasets are summarized in Table 8. The comparison reveals that MaSIF outperforms our model in 2 out of 6 target proteins, while our SurfPro outperforms in 4 out of

Models	Seen Class			Zero-Shot			Average	Inference Time (s/sample)
	InsulinR	PDGFR	TGFb	H3	IL7Ra	TrkA		
MaSIF	**9.0737**	**10.8734**	18.8502	21.9093	25.9689	25.9112	**16.1280**	210.42
SurfPro	10.2608	17.9862	**17.7364**	**21.2916**	**20.8594**	**10.6535**	16.9485	**0.45**

Table 8. Comparison with MaSIF on binder design task. "Average" denotes the average AF2 pAE_interaction across the entire test set instead of the direct average on different target proteins.



(a) pAE score with TrkA: 4.75 (b) pAE score with PDGFR: 5.58

Figure 5. Case study of complexes involving our SurfPro designed binders (in red) and target proteins (in purple): (a) target protein TrkA with pAE_interaction=4.75, (b) target protein PDGFR with pAE_interaction=5.58.

6 target proteins. Despite MaSIF demonstrating a slightly superior average pAE_interaction score, it's noteworthy that the average time taken for MaSIF to identify a promising binder is 210.42 seconds, significantly longer than the mere 0.45 seconds required by SurfPro. This underscores SurfPro's efficacy as a standout generative model, capable of swiftly and directly generating functional protein sequences in an end-to-end manner.

5.6. Case Study

To get an insight on the designed functional proteins by our SurfPro, we visualize two complexes of our model designed binders and target proteins belonging to TrkA (Figure 5 (a)) and PDGFR (Figure 5 (b)). Both the complexes have AF2 pAE_interaction lower than 6, indicating a strong protein-protein binding. As Bennett et al. (2023) state in their work that success rates for designs will be sharply increased with AF2 pAE_interaction < 10. It intuitively shows that our SurfPro is able to design functional binders with high protein-protein binding affinities.

6. Discussion

Our SurfPro exhibits superior performance in rapidly and directly generating functional protein sequences based on provided protein surfaces. However, despite its impressive capabilities, certain limitations persist, which we will discuss in this section. While our approach excels in protein optimization, it leans more towards refinement rather than de novo protein design. This distinction is significant and meaningful. Particularly in binder design, achieving a high-affinity binder from scratch is seldom feasible in practical scenarios. Thus, starting from existing positive binders can accelerate the design process. Nonetheless, this approach also imposes constraints on the practical utility of our method. Firstly, locating a positive initial point isn't always feasible. Secondly, commencing from a favorable starting point may result in limited improvements, as our SurfPro necessitates consideration of both geometric and biochemical constraints. Our future work could explore methods for de novo designing protein surfaces. For example, integrating diffusion models to generate point clouds could substantially enhance the versatility and applicability of our existing framework.

7. Conclusion

In this work, we propose SurfPro, a new generative model to design functional proteins based on desired surface. SurfPro incorporates a hierarchical encoder that progressively captures geometric and biochemical features, transitioning from a local perspective to a global landscape. Additionally, an autoregressive decoder is employed to generate a protein sequence based on the learned geometric and biochemical representations of the surface. Our approach consistently outperforms prior strong inverse folding methods on a general protein design benchmark CATH 4.2, with a sequence recovery rate of 57.78% , and two functional protein design tasks, with higher success rates in terms of protein-protein binding and enzyme-substrate interaction scores.

Acknowledgement

This research is partially supported by the UC Santa Barbara Faculty Research Grant (to L.L.). The authors thank the anonymous reviewers and Wenxian Shi, Siqi Ouyang, Danqing Wang, and Yujia Gao for their excellent suggestions.

Impact Statement

This paper contributes to the design of functional proteins from their surfaces, thereby advancing the field of genera-

tive molecule design. Our work may have various societal implications, but we do not think they need to be specifically highlighted here.

References

Alexa, M., Behr, J., Cohen-Or, D., Fleishman, S., Levin, D., and Silva, C. T. Point set surfaces. In *Proceedings Visualization, 2001. VIS'01.*, pp. 21–29. IEEE, 2001.

Angermueller, C., Dohan, D., Belanger, D., Deshpande, R., Murphy, K., and Colwell, L. Model-based reinforcement learning for biological sequence design. In *International conference on learning representations*, 2019.

Anishchenko, I., Pellock, S. J., Chidyausiku, T. M., Ramelot, T. A., Ovchinnikov, S., Hao, J., Bafna, K., Norn, C., Kang, A., Bera, A. K., et al. De novo protein design by deep network hallucination. *Nature*, 600(7889):547–552, 2021.

Arnold, F. H. Design by directed evolution. *Accounts of chemical research*, 31(3):125–131, 1998.

Arnold, F. H. Directed evolution: bringing new chemistry to life. *Angewandte Chemie International Edition*, 57(16): 4143–4148, 2018.

Bennett, N. R., Coventry, B., Goreshnik, I., Huang, B., Allen, A., Vafeados, D., Peng, Y. P., Dauparas, J., Baek, M., Stewart, L., DiMaio, F., Steven, M., Savvides, S., and Baker, D. Improving de novo protein binder design with deep learning. *Nature Communications*, 14(1):2625, 2023.

Brookes, D., Park, H., and Listgarten, J. Conditioning by adaptive sampling for robust design. In *International conference on machine learning*, pp. 773–782. PMLR, 2019.

Brookes, D. H. and Listgarten, J. Design by adaptive sampling. *arXiv preprint arXiv:1810.03714*, 2018.

Connolly, M. L. Solvent-accessible surfaces of proteins and nucleic acids. *Science*, 221(4612):709–713, 1983.

Dalby, P. A. Strategy and success for the directed evolution of enzymes. *Current opinion in structural biology*, 21(4): 473–480, 2011.

Dauparas, J., Anishchenko, I., Bennett, N., Bai, H., Ragotte, R. J., Milles, L. F., Wicky, B. I., Courbet, A., de Haas, R. J., Bethel, N., et al. Robust deep learning–based protein sequence design using proteinmpnn. *Science*, 378 (6615):49–56, 2022.

Defresne, M., Barbe, S., and Schiex, T. Protein design with deep learning. *International Journal of Molecular Sciences*, 22(21):11741, 2021.

Elfwing, S., Uchibe, E., and Doya, K. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *Neural networks*, 107:3–11, 2018.

Ewing, G. and Hermisson, J. Msms: a coalescent simulation program including recombination, demographic structure and selection at a single locus. *Bioinformatics*, 26(16): 2064–2065, 2010.

Ferruz, N., Schmidt, S., and Höcker, B. Protgpt2 is a deep unsupervised language model for protein design. *Nature communications*, 13(1):4348, 2022.

Fleishman, S. J., Leaver-Fay, A., Corn, J. E., Strauch, E.-M., Khare, S. D., Koga, N., Ashworth, J., Murphy, P., Richter, F., Lemmon, G., et al. Rosettascripts: a scripting language interface to the rosetta macromolecular modeling suite. *PloS one*, 6(6):e20161, 2011.

Gainza, P., Sverrisson, F., Monti, F., Rodola, E., Boscaini, D., Bronstein, M., and Correia, B. Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. *Nature Methods*, 17(2):184–192, 2020.

Gainza, P., Wehrle, S., Van Hall-Beauvais, A., Marchand, A., Scheck, A., Harteveld, Z., Buckley, S., Ni, D., Tan, S., Sverrisson, F., et al. De novo design of protein interactions with learned surface fingerprints. *Nature*, pp. 1–9, 2023.

Gao, Z., Tan, C., Chacón, P., and Li, S. Z. Pifold: Toward effective and efficient protein inverse folding. *arXiv preprint arXiv:2209.12643*, 2022.

Gu, J. and Kong, X. Fully non-autoregressive neural machine translation: Tricks of the trade. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 120–133, 2021.

Hsu, C., Verkuil, R., Liu, J., Lin, Z., Hie, B., Sercu, T., Lerer, A., and Rives, A. Learning inverse folding from millions of predicted structures. In *International Conference on Machine Learning*, pp. 8946–8970. PMLR, 2022.

Huang, P.-S., Boyken, S. E., and Baker, D. The coming of age of de novo protein design. *Nature*, 537(7620): 320–327, 2016.

Ingraham, J., Garg, V., Barzilay, R., and Jaakkola, T. Generative models for graph-based protein design. *Advances in neural information processing systems*, 32, 2019.

Jain, M., Bengio, E., Hernandez-Garcia, A., Rector-Brooks, J., Dossou, B. F., Ekbote, C. A., Fu, J., Zhang, T., Kilgour, M., Zhang, D., et al. Biological sequence design with gflownets. In *International Conference on Machine Learning*, pp. 9786–9801. PMLR, 2022.

Jing, B., Eismann, S., Suriana, P., Townshend, R. J. L., and Dror, R. Learning from protein structure with geometric vector perceptrons. In *International Conference on Learning Representations*, 2020.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. 2014.

Kroll, A., Ranjan, S., Engqvist, M. K., and Lercher, M. J. A general model to predict small molecule substrates of enzymes based on machine and deep learning. *Nature Communications*, 14(1):2787, 2023a.

Kroll, A., Rousset, Y., Hu, X.-P., Liebrand, N. A., and Lercher, M. J. Turnover number predictions for kinetically uncharacterized enzymes using machine and deep learning. *Nature Communications*, 14(1):4139, 2023b.

Kumar, A. and Levine, S. Model inversion networks for model-based optimization. *Advances in Neural Information Processing Systems*, 33:5126–5137, 2020.

Levin, D. The approximation power of moving least-squares. *Mathematics of computation*, 67(224):1517–1531, 1998.

Levin, D. Mesh-independent surface interpolation. In *Geometric modeling for scientific visualization*, pp. 37–49. Springer, 2004.

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637): 1123–1130, 2023.

Lv, C., Lin, W., and Zhao, B. Voxel structure-based mesh reconstruction from a 3d point cloud. *IEEE Transactions on Multimedia*, 24:1815–1829, 2021.

Madani, A., Krause, B., Greene, E. R., Subramanian, S., Mohr, B. P., Holton, J. M., Olmos Jr, J. L., Xiong, C., Sun, Z. Z., Socher, R., et al. Large language models generate functional protein sequences across diverse families. *Nature Biotechnology*, pp. 1–8, 2023.

Meier, J., Rao, R., Verkuil, R., Liu, J., Sercu, T., and Rives, A. Language models enable zero-shot prediction of the effects of mutations on protein function. *Advances in Neural Information Processing Systems*, 34:29287–29303, 2021.

Moss, H., Leslie, D., Beck, D., Gonzalez, J., and Rayson, P. Boss: Bayesian optimization over string spaces. *Advances in neural information processing systems*, 33: 15476–15486, 2020.

Packer, M. S. and Liu, D. R. Methods for the directed evolution of proteins. *Nature Reviews Genetics*, 16(7): 379–394, 2015.

Puny, O., Atzmon, M., Smith, E. J., Misra, I., Grover, A., Ben-Hamu, H., and Lipman, Y. Frame averaging for invariant and equivariant network design. In *International Conference on Learning Representations*, 2021.

Qian, L., Zhou, H., Bao, Y., Wang, M., Qiu, L., Zhang, W., Yu, Y., and Li, L. Glancing transformer for non-autoregressive neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 1993–2003, 2021.

Ren, Z., Li, J., Ding, F., Zhou, Y., Ma, J., and Peng, J. Proximal exploration for model-guided protein sequence design. *bioRxiv*, 2022.

Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., Guo, D., Ott, M., Zitnick, C. L., Ma, J., 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.

Sanner, M. F., Olson, A. J., and Spehner, J.-C. Reduced surface: an efficient way to compute molecular surfaces. *Biopolymers*, 38(3):305–320, 1996.

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

Schnabel, R. and Klein, R. Octree-based point-cloud compression. *PBG@ SIGGRAPH*, 3, 2006.

Song, Z. and Li, L. Importance weighted expectation-maximization for protein sequence design. In *International Conference on Machine Learning*, pp. 32349–32364. PMLR, 2023.

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

Terayama, K., Sumita, M., Tamura, R., and Tsuda, K. Black-box optimization for automated discovery. *Accounts of Chemical Research*, 54(6):1334–1346, 2021.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Wang, J., Lisanza, S., Juergens, D., Tischer, D., Watson, J. L., Castro, K. M., Ragotte, R., Saragovi, A., Milles, L. F., Baek, M., et al. Scaffolding protein functional sites using deep learning. *Science*, 377(6604):387–394, 2022.

Watson, J. L., Juergens, D., Bennett, N. R., Trippe, B. L., Yim, J., Eisenach, H. E., Ahern, W., Borst, A. J., Ragotte, R. J., Milles, L. F., et al. De novo design of protein structure and function with rfdiffusion. *Nature*, 620(7976): 1089–1100, 2023.

Yeh, A. H.-W., Norn, C., Kipnis, Y., Tischer, D., Pellock, S. J., Evans, D., Ma, P., Lee, G. R., Zhang, J. Z., Anishchenko, I., et al. De novo design of luciferases using deep learning. *Nature*, 614(7949):774–780, 2023.

Zheng, Z., Deng, Y., Xue, D., Zhou, Y., Ye, F., and Gu, Q. Structure-informed language models are protein designers. *bioRxiv*, pp. 2023–02, 2023.

Appendix

A. Surface Biochemical Features

The specific values for biochemical features are sourced from the ImMunoGeneTics information system, which are provided in Table 9. In our paper, we utilize two biochemical features, which are hydrophobicity and charge. Additionally, we present the results of our model applying five biochemical features in Table 6. The extra three biochemical features include polarity, acceptor and donor, with their respective values provided in Table 9.

Feature	Description	Value
hydrophobicity	the hydrophobicity level of a residue, the higher the hydrophobicity, the more hydrophobic the residue	I: 4.5, V: 4.2, L: 3.8, F: 2.8, C: 2.5, M: 1.9, A: 1.8 W: -0.9, G: -0.4, T: -0.7, S: -0.8, Y: -1.3, P: -1.6, H: -3.2 N: -3.5, D: -3.5, Q: -3.5, E: -3.5, K: -3.9, R: -4.5
charge	the charge value of a residue	R: 1, K: 1, D: -1, E: -1, H: 0.1, others: 0
polarity	polarity is a separation of electric charge leading to a molecule	R: 1, N: 1, D: 1, Q: 1, E: 1, H: 1, K: 1, S: 1, T: 1, Y: 1 others: 0
acceptor	accepting electrons from another compound	D: 1, E: 1, N: 1, Q: 1, H: 1, S: 1, T: 1, Y: 1, others: 0
donor	transferring electrons to another compound	R: 1, K: 1, W: 1, N: 1, Q: 1, H: 1, S: 1, T: 1, Y: 1, others: 0

Table 9. Detailed values for biochemical features.

B. Protein Surface Generation Example

We provide an example of protein surface generation process in Figure 6, including raw surface construction (Figure 6 (a)), surface smoothing (Figure 6 (b)) and surface compression (Figure 6 (c)).

(a) raw surface (b) smoothed surface (c) compressed surface

Figure 6. Protein surface generation: (a) raw surface, (b) smoothed surface, and (c) compressed surface.

C. Additional Information on Inverse Folding Task

C.1. Surface Vertex Count for CATH 4.2 Benchmark

Table 10 presents the vertex count statistics for the CATH 4.2 dataset.

Vertex Count	Training	Validation	Test
Average Vertex Count/Residue	103	114	119
Maximum Vertex Count	529, 271	393, 522	385, 008

Table 10. Vertex count statistics for surfaces from the CATH 4.2 dataset.

C.2. Recovery Rate After Pairwise Alignment

The amino acid recovery rates of different approaches after pairwise alignment are provided in Table 11. It shows our SurfPro achieves the highest recovery rate among all the compared baselines. In addition, for all non-autoregressive models, the recovery rate decreases after pairwise alignment.

Methods	Recovery Rate (%, ↑)
ProteinMPNN (Dauparas et al., 2022)	35.89
PiFold (Gao et al., 2022)	39.29
LM-DESIGN (Zheng et al., 2023)	38.21
SurfPro	**57.78**

Table 11. Recovery rate after pairwise alignment of different approaches on CATH 4.2 dataset. (↑): the higher the better. Among all the baselines, SurfPro achieves the highest recovery rate.

D. Data Statistics for Two Functional Protein Design Tasks

D.1. Binder Design Datasets

The detailed data statistics for protein binder design is provided in Table 12.

Target Protein	Positive Binder	Training	Validation	Test
InsulinR	238	184	23	31
PDGFR	262	208	26	28
TGFb	95	72	9	14
H3	38	0	0	38
IL7Ra	7	0	0	7
TrkA	4	0	0	4
Total	644	464	58	122

Table 12. Data statistics for the six categories of <positive binder, target protein> complexes in the binder design task.

D.2. Enzyme Design Datasets

The detailed data statistics for enzyme design is provided in Table 13.

Substrate	Enzyme	Training	Validation	Test
C00002	1101	881	110	110
C00677	555	445	55	55
C00019	175	141	17	17
C00003	112	90	11	11
C00001	76	0	0	76
Total	1979	1557	193	269

Table 13. Data statistics for the five enzyme categories, each of which binds to a specific substrate identified as its KEGG ID.

E. PDB Surface Pretraining

To fully explore the design capability of our model, we pretrain our proposed SurfPro on surfaces from the entire PDB. Specifically, we gather all proteins in the PDB until March 10, 2023, resulting in a total of 198,726 samples. Subsequently, we employ MSMS to compute the continuous surface for each protein and extract chain A as the target sequence. Following a similar data processing approach to the CATH 4.2 dataset, we filter out failed instances during the raw surface construction process by MSMS tool and sequences longer than 1024 residues. To prevent the potential data leakage issue, we exclude proteins included in the enzyme design datasets. Consequently, we obtain 179,278 <surface, sequence> pairs. Among them, We randomly split 50 cases for the validation set, leaving the rest for the training set. The model undergoes pretraining for 1,000,000 steps. The learning rate and batch size are set to 5e-4 and 4096 tokens, respectively. After pretraining process, we obtain SurfPro-Pretrain.

Then we separately finetune SurfPro-Pretrain on the binder design and enzyme design tasks, with the corresponding performance reported in Table 2, 3, 4 and 5. The outcomes illustrate a notable improvement in the design capabilities of our SurfPro, as evidenced by substantially increased success rates in both tasks. This enhancement implies that our model

can effectively design proteins with stronger binding functions than experimentally confirmed positive binders or natural enzymes when pretrained on a larger dataset.

F. Additional Experimental Results

F.1. Binder Design Success Rate Comparing AlphaFold2 and ESMFold Predicted Binder Structures

We present the success rate for the binder design task, comparing the binder structures predicted by AlphaFold2 and ESMFold in Table 14. The results reveal a minor difference of only 0.41%, indicating that there is little distinction in performance between AlphaFold2 and ESMFold for this task.

Models	Seen Class			Zero-Shot			Average
	InsulinR	PDGFR	TGFb	H3	IL7Ra	TrkA	
SurfPro-AlphaFold2	**33.15**	17.14	11.29	21.42	17.14	**25.00**	21.88
SurfPro-ESMFold	31.57	**19.99**	**11.61**	**23.21**	**19.28**	**25.00**	**22.29**

Table 14. Success rate (%) for SurfPro with AlphaFold2 predicted binder structures (SurfPro-AlphaFold2) and SurfPro with ESMFold predicted binder structures (SurfPro-ESMFold) on the binder design task. "Average" denotes the average success rate across the entire test set instead of the direct average on different target proteins.

F.2. Sequences for Case Analysis

We provide the designed binder sequences and the positive one in Table 15 for cases analyzed in Figure 4 .

Binders	Sequence
positive one	DEFTEIVKELVKLAEEAVKKNDEDSVKFIEAMLKMMKEAATDPKQRELADRAIKKVQKLLKS
binder1	DEFTEIVKELVKLAEEAVKKNDEESVKFIEAMVKMAKEAMAKMAKEAATDPKQRELADRAIK
binder2	DEFTEIVKELVKLAEEAVKKNDEESVKFIEAMLKMAKEATDPKQRELADRAIKKAQKLLKS
binder3	DEFTEIVKELVKLAEEAVKKNDEESVKFIEAMVKMAKEAMAKMAKEAATDPKQRELADRAIK
binder4	DEFTEIVKELVKLAEEAVKKNDEESVKFIEAMLKMAKEATDPKQRELADRAIKKAQKLLKS
binder5	DEFTEIVKELVKLAEEAVKKNDEESVKFIEAMVKMAKEAMAKMAKEAATDPKQRELADRAIK
binder6	DEFTEIVKELVKLAEEAVKKNDEESVKFIEAMLKMAKEATDPKQRELADRAIKKAQKLLKS
binder7	DEFTEIVKELVKLAEEAVKKNDEESVKFIEAMVKMAKEAMAKMAKEAATDPKQRELADRAIK
binder8	DEFTEIVKELVKLAEEAVKKNDEESVKFIEAMLKMAKEATDPKQRELADRAIKKAQKLLKS
binder9	DEFTEIVKELVKLAEEAVKKNDEESVKFIEAMLKMMKKEAATDPKQRELADRAIKKAQKLLK
binder10	DEFTEIVKELVKLAEEAVKKNDEESVKFIEAMLKMAKEATDPKQRELADRAIKKAQKLLKS

Table 15. Sequences for cases analyzed in Figure 4: designed binders and the experimentally confirmed positive one for the target protein InsulinR.