SurfDesign: Effective Protein Design on Molecular Surfaces

Fang Wu [1] **Shuting Jin** [2] **Xiangru Tang** [3] **Mark Gerstein** [3] **Xiangxiang Zeng** [4] **Jure Leskovec** [1] **Yejin Choi** [1] **Jinbo Xu** [5]

Abstract

Protein function is largely determined by molecular surface geometry and physicochemical complementarity, yet most protein design methods condition only on backbone structure. We introduce SurfDesign, a surface-conditioned protein design framework that models molecular surfaces as continuous geometric manifolds and integrates them with pretrained protein language models. SurfDesign employs surface-based equivariant message passing to capture surface normals, curvature, and directional geometry, together with a parameter-efficient fine-tuning strategy. Focusing on functional protein design, we show that SurfDesign consistently outperforms prior surface-conditioned and backbone-only methods on de novo binder and enzyme design benchmarks. We also report strong performance on inverse-folding benchmarks as a diagnostic of structural compatibility. Our results highlight manifold-aware surface representations as a principled foundation for functional protein and enzyme design. Code is available at https://github.com/smiles724/SurfDesign.

1. Introduction

Proteins are fundamental molecular machines that drive nearly all biological processes, including catalysis, molecular recognition, signaling, and regulation. Recent advances in deep learning (DL) (Huang et al., 2016; Song et al., 2020) have substantially accelerated progress in protein design, shifting the field from physics-based optimization toward data-driven generative modeling (Ingraham et al., 2019; Jing et al., 2020; Hsu et al., 2022; Zheng et al., 2023; Wang et al., 2024). A dominant paradigm in this space is structure-based inverse folding (Defresne et al., 2021), where a target backbone is specified, and a compatible amino-acid sequence is generated. A large body of work has demonstrated impressive improvements under this formulation, especially with graph neural networks (GNNs) and pretrained protein language models (PLMs) (Rives et al., 2021).

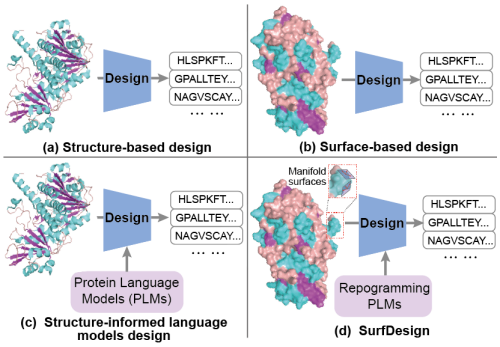

Figure 1. Protein design setups, conditioned on backbone structures or molecular surfaces.

However, the ultimate objective of protein design extends beyond folding correctness (Song et al., 2024; Tang et al., 2025). Many practical design tasks, such as enzyme engineering, receptor-ligand binding, and de novo interaction design, are governed not solely by backbone geometry, but by localized surface shape and physicochemical complementarity (Gainza et al., 2023; Wu & Li, 2024b; Wu et al., 2025c; 2026a; Li et al., 2025b). Proteins with nearly identical folds can exhibit drastically different functions if their surface charge distributions, curvature, or hydrophobic patterns differ. Consequently, backbone-only approaches may fail to adequately constrain functional interfaces, even when global folding is correct (Li et al., 2025a).

Molecular surfaces provide a natural and functionally grounded representation. Defined by smooth atomic boundaries, protein surfaces directly mediate physical interactions with substrates, ligands, and binding partners. Surface geometry encodes shape complementarity, while surface-level physicochemical patterns determine specificity and affinity. Prior work (Song et al., 2024; Li et al., 2025b) in protein surface modeling has demonstrated the effectiveness of surface representations for interaction prediction and binder design, highlighting their importance in functional protein discovery. Despite this, surface-conditioned generation re-

[1]Stanford University, Stanford, USA [2]Wuhan University of Science and Technology, Wuhan, China [3]Yale University, New Haven, USA [4]Hunan University, Changsha, China [5]Toyota Technological Institute at Chicago, Illinois, USA. Correspondence to: Fang Wu <fangwu97@stanford.edu>.

Accepted at the 2026 Workshop on Generative and Agentic AI for Biology (ICML 2026)

mains underexplored, and existing methods often rely on discretized point clouds or meshes that inadequately capture intrinsic continuity and geometry.

Two key challenges exist in surface-conditioned design. First, molecular surfaces are continuous and smooth manifolds (Lee et al., 2023; Sun et al., 2024), whereas existing methods (Song et al., 2024; Zhang et al., 2023) treat them as unordered collections of points or fixed meshes, ignoring local tangent structure, curvature, and directional consistency. This mismatch restricts the expressiveness of learned representations and hampers generalization to fine-grained functional regions such as binding pockets and catalytic sites. Second, crystal structures are limited in number relative to available sequences, making it difficult to train data-hungry models purely from surface-sequence pairs. Surface information alone may be insufficient in buried regions, where evolutionary and sequential priors play complementary roles.

To address them, we propose SurfDesign that explicitly models molecular surfaces as geometric manifolds and integrates them with PLMs (Zheng et al., 2023; Qiu et al., 2024; Wang et al., 2024; Mao et al., 2023) (see Fig. 2). We introduce a surface-conditioned equivariant message passing (SEMP) encoder that leverages surface normals, curvature, and directional relationships to capture local manifold structure while preserving roto-translation equivariance. To mitigate data scarcity and enhance sequence modeling, we further incorporate a hybrid parameter-efficient fine-tuning (PEFT) strategy that injects surface-derived structural information into PLMs without full retraining. Importantly, SurfDesign is designed primarily to generate functional proteins. In functional design tasks such as binding and enzyme recognition, there is no unique ground-truth sequence associated with a given structure or surface. Instead, multiple diverse sequences may satisfy the same functional constraints. Accordingly, we treat inverse folding benchmarks not as supervised label-recovery tasks, but as diagnostic evaluations that assess whether generated sequences are structurally compatible with the specified geometry. This distinction is crucial: **conditioning on molecular surfaces introduces richer physical constraints, not privileged access to native sequence identities.**

We evaluate SurfDesign on standard inverse folding benchmarks (Orengo et al., 1997) and achieve state-of-the-art amino-acid recovery (AAR) and perplexity, validating its ability to generate structurally consistent sequences. More importantly, we show that SurfDesign excels in functional protein design, achieving superior performance in de novo binder generation across six benchmark targets and in enzyme design for multiple enzyme-substrate systems. These findings indicate that expressive surface modeling enables protein design with greater functional fidelity than backbone-only approaches. Overall, this work advances protein design by elevating molecular surfaces from auxiliary features to first-class conditioning signals, demonstrating that explicitly modeling molecular surfaces as geometric manifolds is a powerful and complementary approach for moving beyond backbone-centric design.

2. Method

2.1. Preliminary and Background

Problem Statement. Neural structure-conditioned protein design aims to find the amino acid sequence $\mathcal{S} = \{s_i \in \text{Cat}(20) : 1 \leq i \leq n\}$ folding into a desired structure $\mathcal{X} = \{\boldsymbol{x}_i \in \mathbb{R}^{4\times 3} : 1 \leq i \leq n\}$, where s_i is one of the 20 residue types and $\mathcal{X}$ denotes the spatial coordinates for 4 backbone atoms (*i.e.*, C_α, C, N and O). It can be formulated as an end-to-end graph-to-sequence learning problem with a parameterized encoder-decoder neural network $\mathcal{F}_{\boldsymbol{\vartheta}}$: $\mathcal{X} \to \mathcal{S}$. Surface-conditioned design is analogous but yields functional proteins that fold into the expected surface $\mathcal{Q}$ with associated biochemical properties (Song et al., 2024). Our objective, therefore, transfers to learn a function $\mathcal{F}_{\boldsymbol{\vartheta}}(\cdot) : \mathcal{Q} \to \mathcal{S}$. Given sufficient surface-sequence paired data, the learning purpose is to maximize the conditional log-likelihood $p(\mathcal{S}|\mathcal{Q};\boldsymbol{\vartheta})$. This enables the design of sequences that either maximize likelihood or are generated via sampling algorithms to ensure diversity and novelty (Zheng et al., 2023). Remarkably, homologous proteins consistently share similar surfaces (Pearson & Sierk, 2005), so the surface-conditioned design is underdetermined.

Unlike supervised classification, protein design does not assume a unique ground-truth sequence $\mathcal{S}$ for a given structure (Gao et al., 2022a). A single backbone or surface geometry is compatible with many valid sequences, and the experimentally observed sequence is neither unique nor necessarily optimal. Accordingly, sequence recovery is used only as a diagnostic proxy for structural compatibility rather than a supervised learning target. Our objective is therefore to learn physically and functionally consistent sequence distributions conditioned on geometric constraints, not to recover native labels.

Surface Generation Surface geometry is crucial for interaction analysis. We employ PyMol (DeLano et al., 2002) to obtain the raw molecular surface, where a probe of a certain radius (~ 1 Å) is moved along the protein to calculate the Solvent Accessibility Surface (SAS) and Solvent Excluded Surface (SES). We treat the resulting surface vertices as oriented surface points, defined by an oriented point cloud $Q = \{q_i : 1 \leq i \leq m\}$ and $m >> n$. Each surface point q_i has a triplet of attributes $(\boldsymbol{x}_i, \boldsymbol{n}_i, \boldsymbol{h}_i)$, where $\boldsymbol{x}_i \in \mathbb{R}^3$ and $\boldsymbol{n}_i \in \mathbb{R}^3$ are the 3D coordinates and unit normal vector, and $\boldsymbol{h}_i \in \mathbb{R}^{\phi_h}$ indicates the physicochemical properties of q_i

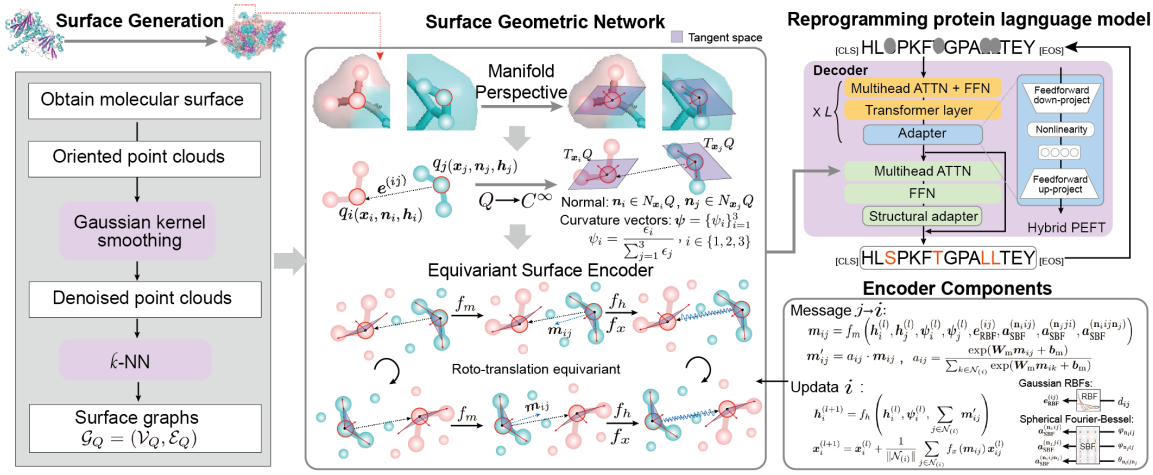

Figure 2. Illustration of SurfDesign. Smooth-surface graphs are obtained using PyMOL or MSMS and subsequently denoised. Then, an equivariant surface encoder is appended to extract manifold representations. These features are further incorporated into the structural adapter of protein language models to recover masked amino acids.

such as hydrophobicity, hbond, and charge. Then the surface graph is built via k-NN, resulting in $\mathcal{G}_Q = (\mathcal{V}_Q, \mathcal{E}_Q)$. We also investigate MSMS (Robinson et al., 2014) and BioPython (Cock et al., 2009) for surface generation and identify negligible differences in processing speeds across several toolkits. As raw point clouds generally carry noise that may limit the expressivity of molecular surfaces (Alexa et al., 2001), we apply the Gaussian kernel smoothing (Song et al., 2024) to raw cloud data:

$$\boldsymbol{x}_i \leftarrow \sum_{j \in \mathcal{N}_{(i)}} \frac{\mathcal{K}(\boldsymbol{x}_i, \boldsymbol{x}_j) \cdot \boldsymbol{x}_j}{\sum_{t \in \mathcal{N}_{(i)}} \mathcal{K}(\boldsymbol{x}_i, \boldsymbol{x}_t)}, \quad \mathcal{K}(\boldsymbol{x}, \boldsymbol{y}) = \exp^{-\frac{\|\boldsymbol{x}-\boldsymbol{y}\|^2}{\eta}}, \tag{1}$$

where $\mathcal{N}_{(i)}$ denotes the neighborhood of $\boldsymbol{x}_i$ and $\mathcal{K}(.,.)$ is the Gaussian kernel with η indicating distance scale in the point space. Here, η is set as $\max_{j \in \mathcal{N}_{(i)}, i \in [m]}(\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2)$. For clarity, $\boldsymbol{h}_i$ are computed from local atomic geometry and element types, not from residue identity labels. We do not use: (i) residue identities, (ii) multiple sequence alignments (MSA) or evolutionary profiles, or (iii) functional annotations or active-site labels, during the surface generation pipeline. While we compute surfaces from full-atom structures, these structures may be experimental, predicted, or generated (e.g., RFdiffusion + packing). The atom types specify geometric constraints but do not reveal the target sequence, as multiple valid sequences can fold into similar atomic arrangements.

2.2. Surface Geometric Network

A Manifold Perspective for Molecular Surfaces. Theoretically, molecular surfaces are continuous manifolds with infinite resolution (Lee et al., 2023), which cannot be fully expressed by existing mesh- (Gainza et al., 2020) or point-based (Sverrisson et al., 2021; Zhang et al., 2023; Song et al., 2024) mechanisms. The key distinguishing property of manifold surfaces relative to conventional point clouds or meshes is that every point on the manifold is locally Euclidean. Mathematically, for $\forall q_i \in Q$, there exists a neighborhood U_{q_i} and a homeomorphism $f_{\text{homo}}(\cdot)$ such that $f_{\text{homo}} : U_{q_i} \to V \subseteq \mathbb{R}^3$, where V is an open ball in $\mathbb{R}^3$. In order to describe the local geometry of a manifold point $q_i \in Q$, we need to know at least (1) the linear approximation of the manifold in its vicinity, which corresponds to *the tangent space*, and (2) how fast the surface bends or deviates from being a plane near this point, which can be measured by *curvature*.

Towards this goal, we assume that the surface Q is a C^∞ differentiable manifold and $T_{\boldsymbol{x}_i}Q$ denotes the tangent space of any point $\boldsymbol{x}_i \in Q$. Then we can acquire the unit normal vector $\boldsymbol{n}_i \in N_{\boldsymbol{x}_i}Q$ perpendicular to $T_{\boldsymbol{x}_i}Q$. If Q is implicitly described by a signed distance function (SDF) satisfying $f_{\text{SDF}}(\cdot) = 0$, then the normal at point $\boldsymbol{x}_i$ is equivalent to the gradient, *i.e.*, $\boldsymbol{n}_i = \nabla f_{\text{SDF}}(\boldsymbol{x}_i)$. Here, we draw the normal vector set $\{\boldsymbol{n}\}_{i=1}^m$ immediately from the software (*i.e.*, PyMol) and integrate this orientation knowledge into the geometric encoder to linearly approximate the manifold and achieve manifold-awareness. Prior studies (Zhang et al., 2023; Song et al., 2024) have seldom considered this specialty of molecular surfaces and merely handle naive clouds. One exception, dMaSIF (Strokach et al., 2020), notices this manifold uniqueness and computes the quasi-geodesic distance as $d_{ij} = \|\boldsymbol{x}_{ij}\|^2 \cdot \left(2 - \boldsymbol{n}_i^\top \cdot \boldsymbol{n}_j\right)$ to naively resemble the geodesic coordinates in the tangent space $T_{\boldsymbol{x}_i}Q$. However, its construction of tangent vectors destroys the equivariance.

Additionally, there are varying ways to define curvatures of 3D Riemannian manifolds intrinsically without reference to a larger space (Kobayashi & Nomizu, 1996), such as normal curvature k_n, geodesic curvature k_g, and geodesic

torsion τ_r. Those all relate the direction of curvatures to the unit normal vector $\boldsymbol{n}_i$. Given a non-singular curve $\gamma(q_i) \in Q$ parametrized by arc length, we can compute $\boldsymbol{T}_i = \gamma'(q_i) = \frac{d\gamma}{dq}$ and $\boldsymbol{t}_i = \boldsymbol{n}_i \times \boldsymbol{T}_i$ to form the Darboux frame. The triple $(\boldsymbol{T}_i, \boldsymbol{t}_i, \boldsymbol{n}_i)$ defines a positively oriented orthonormal basis attached to each point of the curve $\gamma(q_i)$. Then the above quantities are related by $\begin{pmatrix} \boldsymbol{T}' \\ \boldsymbol{t}' \\ \boldsymbol{u}' \end{pmatrix} = \begin{pmatrix} 0 & k_g & k_n \\ -k_g & 0 & \tau_r \\ -k_n & -\tau_r & 0 \end{pmatrix} \begin{pmatrix} \boldsymbol{T} \\ \boldsymbol{t} \\ \boldsymbol{u} \end{pmatrix}$. Inspired by progress in geometry processing (Tian et al., 2023; Wu & Li, 2024b; Zhang et al., 2008), we estimate these quantities in a closed form from local points $\mathcal{N}_{(i)}$. Specifically, we first compute a covariance matrix for q_i and its neighborhood $\mathcal{N}_{(i)}$:

$$\boldsymbol{\Sigma} = \frac{1}{\|\mathcal{N}_{(i)}\|} \sum_{\mathbf{x}_j \in \mathcal{N}_{(i)}} \mathbf{x}_j \mathbf{x}_j^\top - \bar{\mathbf{x}}\bar{\mathbf{x}}^\top, \quad \boldsymbol{\Sigma} \in \mathbb{R}^{3\times 3}. \tag{2}$$

where $\bar{\mathbf{x}}$ is the centroid of this point cluster. After the eigen-decomposition of $\boldsymbol{\Sigma}$ (*e.g.*, singular value decomposition or eigenvalue decomposition), eigenvalues can be obtained as ϵ_1, ϵ_2, and ϵ_3 ($\epsilon_1 \geq \epsilon_2 \geq \epsilon_3$). Those pseudo curvatures vectors $\boldsymbol{\psi} = \{\psi_i\}_{i=1}^3$ can be therefore computed as:

$$\psi_i = \frac{\epsilon_i}{\sum_{j=1}^3 \epsilon_j}, \quad i \in \{1, 2, 3\}. \tag{3}$$

We employ $\boldsymbol{\psi}$ as a rotation-invariant local shape descriptor that approximates curvature-related information, rather than exact differential curvatures. It can be proved that this curvature feature $\boldsymbol{\psi}$ is roto-translation invariant (see App. B).

Directionality in Surface Point Clouds. The manifold characteristic of molecular surfaces introduces additional directional information when considering pairwise or ternary interactions among connected particles. To be specific, for each neighboring point pair (i, j), two intersecting planes (see Fig. 3) are formulated with respective normals $(\boldsymbol{n}_i, \boldsymbol{n}_j)$. We denote the angles between normals and the connecting directed line of two points $(\boldsymbol{x}_{ij}, \boldsymbol{x}_{ji})$ by $\varphi_{\mathbf{n}_i ij} = \angle \boldsymbol{n}_i \boldsymbol{x}_{ij}$ and $\varphi_{\mathbf{n}_j ji} = \angle \boldsymbol{n}_j \boldsymbol{x}_{ji}$. We denote the dihedral angle between two half-phases as $\theta_{\mathbf{n}_i ij \mathbf{n}_j} = \angle \mathbf{n}_i \mathbf{n}_j \perp \boldsymbol{x}_{ij}$. In addition to the common distance $\|\boldsymbol{x}_{ij}\|^2$, these three angles provide a more comprehensive view of understanding the relative position of (q_i, q_j) lying in the surface manifold Q, which will also be incorporated into our surface modeling. For instance, for different values of $\left(\varphi_{\mathbf{n}_i ij}, \varphi_{\mathbf{n}_j ji}, \theta_{\mathbf{n}_i ij \mathbf{n}_j}\right)$, a triplet of $(\frac{\pi}{2}, \frac{\pi}{2}, 0)$ indicates a perfectly smooth region, while a triplet of (π, π, π) implies a severely sharp and steep curve.

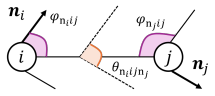

Figure 3. Angles hidden in the oriented surface point cloud, containing two intersection angles $\varphi_{\mathbf{n}_i ij} = \angle \boldsymbol{n}_i \boldsymbol{x}_{ij}$ and $\varphi_{\mathbf{n}_j ji} = \angle \boldsymbol{n}_j \boldsymbol{x}_{ji}$ as well as a dihedral angle $\theta_{\mathbf{n}_i ij \mathbf{n}_j}$.

Equivariant Surface Encoder. Finally, we draw inspiration from prevalent and modern equivariant algorithms (Gasteiger et al., 2020b;a; Satorras et al., 2021; Zhang et al., 2023) and propose a **s**urface-based **e**quivariant **m**essage **p**assing (SEMP) as the encoder of $\mathcal{F}_{\boldsymbol{\vartheta}}(\cdot)$. Our SEMP architecture is roto-translation equivariant and leverages both directional and curvature information. To begin with, by setting an interaction cutoff c_{int}, we calculate the 3D spherical Fourier-Bessel bases $\left(\boldsymbol{a}_{\text{SBF}}^{(\mathbf{n}_i ij)}, \boldsymbol{a}_{\text{SBF}}^{(\mathbf{n}_j ji)}\right) \in 2 \times \mathbb{R}^{N_{\text{CBF}} \times N_{\text{SBF}} \times N_{\text{RBF}}}$ for two angles $\varphi \in \left[\varphi_{\mathbf{n}_i ij}^{(l)}, \varphi_{\mathbf{n}_j ji}^{(l)}\right]$ to integrate orientation knowledge between each interactive particles in the surface:

$$a_{\text{SBF},ovt}^{(l)}\left(\boldsymbol{x}_{ij}^{(l)}, \varphi, \theta_{\mathbf{n}_i ij \mathbf{n}_j}^{(l)}\right) = \sqrt{\frac{2}{c_{\text{int}}^3 j_{o+1}^2(z_{ov})}} j_o\left(\frac{z_{ov}}{c_{\text{int}}} \left\|\boldsymbol{x}_{ij}^{(l)}\right\|^2\right) Y_o^t\left(\varphi_{\mathbf{n}_i ij}^{(l)}, \theta_{\mathbf{n}_i ij \mathbf{n}_j}^{(l)}\right), \tag{4}$$

where $o \in [N_{\text{CBF}}]$, $v \in [N_{\text{SBF}}]$, and $t \in [N_{\text{RBF}}]$ control the degree, root, and order of the radial basis functions, respectively. $\|\mathbf{x}_{ij}\|$ denotes the Euclidean distance between surface points i and j. Besides, $j_o(\cdot)$ is the o-th degree spherical Bessel functions and z_{ov} is its corresponding v-th root. $Y_o^t(\cdot)$ is the real o-th degree and t-th order spherical harmonics. Equ. 4 can be boiled down to a joint 2D basis if the order t is set to 0. By using $Y_o^0(\cdot)$, we obtain the 2D representation $\boldsymbol{a}_{\text{SBF}}^{(\mathbf{n}_i ij \mathbf{n}_j)} \in \mathbb{R}^{N_{\text{CBF}} \times N_{\text{SBF}}}$ based on $\theta_{\mathbf{n}_i ij \mathbf{n}_j}^{(l)}$.

Remarkably, those 2D/3D spherical Fourier-Bessel representations $\boldsymbol{a}_{\text{SBF}}^{(\mathbf{n}_i ij)}$, $\boldsymbol{a}_{\text{SBF}}^{(\mathbf{n}_j ji)}$, and $\boldsymbol{a}_{\text{SBF}}^{(\mathbf{n}_i ij \mathbf{n}_j)}$ enjoy the roto-translation invariant property due to their exploitation of the relative distance as well as the invariant angles. Then those directional vectors, along with pointwise curvature, are fed into SEMP to attain the initial messages $\boldsymbol{m}_{ij}$ as:

$$\boldsymbol{m}_{ij} = f_m\left(\boldsymbol{h}_i^{(l)}, \boldsymbol{h}_j^{(l)}, \boldsymbol{\psi}_i^{(l)}, \boldsymbol{\psi}_j^{(l)}, \boldsymbol{e}_{\text{RBF}}^{(ij)}, \boldsymbol{a}_{\text{SBF}}^{(\mathbf{n}_i ij)}, \boldsymbol{a}_{\text{SBF}}^{(\mathbf{n}_j ji)}, \boldsymbol{a}_{\text{SBF}}^{(\mathbf{n}_i ij \mathbf{n}_j)}\right), \tag{5}$$

where $f_m(\cdot)$ is a multi-layer perceptron (MLP) appended with an activation function like SiLU (Nwankpa et al., 2018). $\boldsymbol{e}_{\text{RBF}}^{(ij)}$ is the radial basis function representation of the interatomic distance $\|\boldsymbol{x}_{ij}\|^2$. Then a softmax is employed to reweight the messages:

$$\boldsymbol{m}'_{ij} = a_{ij} \cdot \boldsymbol{m}_{ij}, \quad a_{ij} = \frac{\exp(\boldsymbol{W}_{\text{m}} \boldsymbol{m}_{ij} + \boldsymbol{b}_{\text{m}})}{\sum_{k \in \mathcal{N}_{(i)}} \exp(\boldsymbol{W}_{\text{m}} \boldsymbol{m}_{ik} + \boldsymbol{b}_{\text{m}})} \tag{6}$$

where the weight matrix $\boldsymbol{W}_{\text{m}} \in \mathbb{R}^{\phi_m \times 1}$ and vector $\boldsymbol{b}_{\text{m}} \in \mathbb{R}$ are learnable. After that, messages are propagated from the vicinity of each point q_i to update its node feature as well

as coordinates:

$$\boldsymbol{h}_i^{(l+1)} = f_h\left(\boldsymbol{h}_i^{(l)}, \boldsymbol{\psi}_i^{(l)}, \sum_{j\in\mathcal{N}_{(i)}} \boldsymbol{m}'_{ij}\right), \tag{7}$$

$$\boldsymbol{x}_i^{(l+1)} = \boldsymbol{x}_i^{(l)} + \frac{1}{\|\mathcal{N}_{(i)}\|}\sum_{j\in\mathcal{N}_{(i)}} f_x\left(\boldsymbol{m}_{ij}\right)\boldsymbol{x}_{ij}^{(l)}, \tag{8}$$

where $f_h(\cdot)$ is another MLP and $f_x : \mathbb{R}^{\phi_m} \to \mathbb{R}$ transforms $\boldsymbol{m}_{ij}$ into a scalar score to control the impact of directional vector $\boldsymbol{x}_{ij}^{(l)}$. Notably, as the position of each point $\boldsymbol{x}_i^{(l)}$ is moving as the layer $l \in [L]$ goes deeper with $\boldsymbol{x}_i^{(0)} = \boldsymbol{x}_i$, it is optional but recommended to adjust and recalculate the curvature $\boldsymbol{\psi}_i$ and relevant angles $\left(\varphi_{\mathbf{n}_i ij}, \varphi_{\mathbf{n}_j ji}, \theta_{\mathbf{n}_i ij\mathbf{n}_j}\right)$ simultaneously. As angles $\left(\varphi_{\mathbf{n}_i ij}, \varphi_{\mathbf{n}_j ji}, \theta_{\mathbf{n}_i ij\mathbf{n}_j}\right)$ depend on each normal vector pair ($\boldsymbol{n}_i$ and $\boldsymbol{n}_j$), we adopt the local least fitting method (Mitra & Nguyen, 2003) to estimate and renew $\{\boldsymbol{n}_i\}_{i=1}^m$. In specific, for q_i's updated coordinates $\boldsymbol{x}_i^{(l)}$ at the l-th layer, we compute the covariance $\boldsymbol{\Sigma}^{(l)}$ according to Equ. 2 and decompose it to obtain three sorted eigenvalues as well as their corresponding eigenvectors $(\boldsymbol{\nu}_1, \boldsymbol{\nu}_2, \boldsymbol{\nu}_3)$. Then $\boldsymbol{\nu}_3$ with the least eigenvalue is selected as the normal vector $\boldsymbol{n}^{(l)}$ at the l-th layer.

All geometric quantities used as inputs to the message function, including radial distances r_{ij}, curvature descriptors $\boldsymbol{\psi}_i$, and angular features $(\varphi_{\mathbf{n}_i ij}, \varphi_{\mathbf{n}_j ji}, \theta_{\mathbf{n}_i ij\mathbf{n}_j})$, are invariant under global rigid transformations. Equivariance is preserved by propagating directional information exclusively through relative displacement vectors $\mathbf{x}_{ij}$ in the coordinate update, ensuring SE(3)-equivariant behavior by construction.

2.3. Reprogramming Protein Language Models

PEFT for SurfDesign. Recent works have explored the possibility of transforming PLMs (Rives et al., 2021; Lin et al., 2022; Hu et al., 2022) into protein design models, and massive evidence demonstrates that the emergent evolutionary knowledge hidden in those PLMs can vastly facilitate the structure-conditioned protein design. Concretely, LM-Design (Zheng et al., 2023), InstructPLM (Qiu et al., 2024), KW-Design (Gao et al., 2023), and VFN-IF-ESM (Mao et al., 2023) report improvements in CATH 4.2 of 10.8% (AAR $50.22\% \to 55.65\%$), 73.9% (perplexity $10.28 \to 2.68$), 14.4% (AAR $54.74\% \to 62.67\%$), and 17.6% (AAR $51.66\% \to 60.77\%$), respectively.

Motivated by this progress, we also use PLMs as the decoder for $\mathcal{F}_{\boldsymbol{\vartheta}}(\cdot)$ and stack several parameter-efficient fine-tuning (PEFT) techniques to fully exploit the potential of PLMs and significantly reduce the memory footprint. Specifically, we utilize a hybrid PEFT method combined with a structural adapter (Zheng et al., 2023) and LoRA (Hu et al., 2021) with a rank of $r = 4$ and a scaling constant of $\alpha = 8$. It is worth noting that there is still no consensus on which type of PEFT strategy is most suitable for PLMs (Sledzieski et al., 2024), and we have found our hybrid mechanism to be more effective than a single strategy for surface-conditioned protein design.

Training. We employ the conditional masked language modeling (CMLM) (Zheng et al., 2023) to better accommodate PLMs that are tasked with MLM (Devlin et al., 2018) as the training objective. Given the surface $\mathcal{Q}$, CMLM decomposes the sequence into masked and observed ones as $\mathcal{S} = \mathcal{S}_{\text{masked}} \cup \mathcal{S}_{\text{obs}}$ and assumes a conditional independence over identities of target residues $s_i \in \mathcal{S}_{\text{masked}}$. Then it requires the model to predict a set of target amino acids $\mathcal{S}_{\text{masked}}$ from the remaining observed residues $\mathcal{S}_{\text{obs}}$:

$$p(\mathcal{S}_{\text{masked}}|\mathcal{S}_{\text{obs}}, \mathcal{Q}; \theta) = \Pi_{s_i\in\mathcal{S}_{\text{masked}}} p(s_i|\mathcal{S}_{\text{obs}}, \mathcal{Q}; \theta) \tag{9}$$

where $\mathcal{S}_{\text{masked}}$ is randomly masked. Moreover, Zheng et al. (2023) presents a coarse-to-fine manner to reconstruct a protein's native sequence from its corrupted version. We also explore this inference scheme with iterative refinement (Savinov et al., 2021) but discover no benefit.

3. Experiments

SurfDesign unifies surface-conditioned inverse folding and functional protein design within a single geometric framework; inverse folding serves as a stress test of geometric compatibility, while binding and catalysis assess functional utility in the absence of ground-truth sequence data. Unless otherwise specified, *Average* denotes a sample-weighted average over the entire test set rather than a simple mean over targets. More experimental details, dataset statistics, and additional results are elaborated in App. C.

3.1. Protein Binder Design

Dataset and Setups. We focus on optimizing proteins for high-affinity binding under functional and stability constraints and evaluate our method on the benchmark introduced by Song et al. (2024), in which the task is to design protein binders that strongly interact with given target receptors. The dataset comprises experimentally validated positive binder-target pairs from six categories, curated in Bennett et al. (2023). Following prior work, we use AlphaFold2 predicted aligned error (AF2 pAE_interaction) as the evaluation metric, as it quantitatively reflects interface confidence and effectively separates positive binders from negative ones under a fixed binding geometry, without conflating docking-search effects inherent to end-to-end multimer prediction. A designed binder is considered successful if its AF2 pAE_interaction is lower than that of the corresponding native positive binder. As a control, we include randomly sampled non-binding sequences of the same length as negative binders. All models, including ours, are fine-tuned on the same binder design dataset derived from

Table 1. AF2 pAE_interaction (↓) for all models in the binder design task. The AF2 pAE_interaction for randomly sampled negative binders of the same length is also provided, along with the positive ones.

Models	InsulinR	PDGFR	TGFb	H3	IL7Ra	TrkA	Average
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 Baselines	19.9880	21.2690	21.4971	24.4997	24.1541	23.1147	22.2020
ProteinMPNN	18.3393	25.2919	25.8559	24.5968	25.5278	27.0980	23.4462
PiFold	12.9809	21.8230	24.4737	23.3924	26.6738	19.7172	20.5785
LM-DESIGN	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
SurfDesign	**8.9827**	**16.3462**	**17.4338**	**21.0642**	**20.8207**	**10.4288**	**15.8460**

Table 2. Success rate (%, ↑) of different models on the binder design task. SurfDesign achieves the highest overall success rate.

Model	Seen Class			Zero-Shot			Average
	InsulinR	PDGFR	TGFb	H3	IL7Ra	TrkA	
ProteinMPNN	3.22	5.71	20.71	18.68	24.10	7.50	11.96
PiFold	20.64	3.57	19.19	29.21	22.85	20.00	19.32
LM-DESIGN	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
SurfDesign	34.87	24.28	29.46	32.89	34.28	27.50	**30.14**

CATH 4.2 pretraining to ensure a fair comparison.

Results. As shown in Tab. 1, SurfDesign achieves an average AF2 pAE_interaction of 15.85, the lowest among all models. This represents a clear improvement over the prior surface-conditioned method, SurfPro (16.95), and significantly outperforms non-surface models such as LM-DESIGN (20.77), PiFold (20.58), and ProteinMPNN (23.45). SurfDesign consistently yields top performance on five of six targets: PDGFR (16.35), TGFb (17.43), H3 (21.06), IL7Ra (20.82), and TrkA (10.43). Lower AF2 pAE_interaction directly reflects increased structural certainty at the designed interface, though it is not a direct measure of binding affinity, indicating that SurfDesign produces binders whose surfaces more consistently support stable receptor engagement. While the improvement over SurfPro is moderate, it reflects the benefits of enhanced surface modeling and improved geometric representation of interaction sites. In contrast, negative and random binders yield much higher AF2 pAE_interaction scores (21.13 and 22.20, respectively), further validating the discriminative power of this metric. These findings support the hypothesis that surface-informed representations can enhance the design of functional proteins with superior binding fidelity. SurfDesign thus represents a promising direction for high-accuracy functional protein generation grounded in surface geometry.

3.2. Enzyme Design

Enzyme design poses a stricter test than binder design, as successful generation requires precise pocket geometry and localized physicochemical patterns rather than global shape complementarity.

Dataset and Setups. We evaluate SurfDesign on the *enzyme design* task, where the objective is to generate enzyme sequences that bind specific small-molecule substrates. It tests whether surface-conditioned generation can capture fine-grained geometric and physicochemical patterns required for enzyme–substrate interactions. We adopt the benchmark compiled by Kroll et al. (2023) comprising five enzyme categories, each associated with a distinct substrate. To prevent data leakage, we explicitly exclude all enzymes that appear in the CATH 4.2 dataset. For enzyme categories containing more than 100 samples, we perform clustering-based splitting and randomly partition the data into training, validation, and test sets using an 8:1:1 ratio. For categories with fewer samples, all enzymes are used exclusively for testing in a zero-shot setting.

To assess enzyme-substrate binding affinity, we adopt the Enzyme-Substrate Potential (ESP) score proposed by Kroll et al. (2023). The ESP model predicts enzyme-substrate interactions with approximately 91% accuracy across multiple benchmarks, with higher values indicating greater predicted enzyme-substrate compatibility. We use the official implementation released by Kroll et al. (2023) to compute ESP scores for all designed enzymes. Following the protocol used in protein binder design, we report (i) the average ESP score obtained via greedy decoding, and (ii) the average success rate computed from sequences generated by sampling with temperature 0.1. For each enzyme-substrate pair, the success rate is defined as the fraction of generated sequences whose ESP score is better than that of the corresponding native enzyme.

Consistent with the binder design setup, we fine-tune all baseline models on the enzyme design dataset, starting from models pretrained on the inverse folding task. We additionally report results for a random baseline and SurfPro-Pretrain under the same settings as the binder design. Importantly, the pretraining corpus for SurfPro-Pretrain explicitly excludes all enzymes used in this evaluation to avoid any potential data leakage.

Results. Quantitatively, SurfDesign improves over prior surface and non-surface baselines on both metrics. For ESP score (Tab. 3), SurfDesign achieves the best overall average of **0.9058**, compared to SurfPro (**0.8931**) and ProteinMPNN (**0.8676**), and remains competitive with LM-DESIGN (**0.9037**) despite LM-DESIGN benefiting from large-scale PLM pretraining. For success rate (Tab. 4), SurfDesign attains the highest overall success rate among the compared methods at **47.30%**, exceeding SurfPro (**42.23%**) and SurfPro-Pretrain (**43.63%**), and outperforming PiFold (**40.65%**) and LM-DESIGN (**37.58%**). Notably, gains persist in the zero-shot setting (C00001), where

Table 3. ESP score (↑) of different models under greedy decoding on the enzyme design task. Substrates are denoted by their KEGG IDs.

Model	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	0.9711	0.7375	0.2614	0.6699	0.9763	0.8676
PiFold	0.9142	0.8816	0.4296	0.8212	0.9616	0.8865
LM-DESIGN	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
SurfDesign	0.9487	0.9012	0.4523	0.8189	0.9801	**0.9058**

Table 4. Success rate (%, ↑) of different models on the enzyme design task. Substrates are denoted by their KEGG database IDs. SurfDesign achieves the highest average success rate.

Model	Seen Class				Zero-Shot	Average
	C00002	C00677	C00019	C00003	C00001	
ProteinMPNN	47.54	31.63	58.82	44.72	27.65	39.23
PiFold	48.54	41.72	58.29	37.54	24.97	40.65
LM-DESIGN	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
SurfDesign	52.73	45.45	60.18	51.82	34.36	**47.30**

SurfDesign achieves **34.36%** success, compared with SurfPro's **33.55%**. These suggest that more expressive manifold-aware surface representations can better capture localized pocket geometry and physicochemical patterns critical for enzyme-substrate interactions.

3.3. Inverse Folding as Compatibility Analysis

We include inverse folding results to assess whether surface-conditioned generation preserves global fold compatibility, rather than treating it as a primary design objective. Various benchmarks are used to design fixed-backbone protein sequences, including single-chain monomers and multi-chain protein complexes.

Baselines and Datasets. A wide variety of approaches are established for comparison, most of which are open source. Among them, StructGNN (Ingraham et al., 2019), GraphTrans (Ingraham et al., 2019), GVP (Jing et al., 2020), ProteinMPNN (Dauparas et al., 2022), AlphaDesign (Gao et al., 2022b), PiFold (Gao et al., 2022a), UniIF (Gao et al., 2024), etc. are GNN-based algorithms, while DenseCPD (Qi & Zhang, 2020) is CNN-based. DPLM (Wang et al., 2024), InstructPLM (Qiu et al., 2024), LM-Design (Zheng et al., 2023), KW-Design (Gao et al., 2023) and VFN-IF-ESM (Mao et al., 2023) leverage and integrate PLMs' knowledge. GRADE-IF (Yi et al., 2023) and DMRA (Wang et al., 2024) rely on diffusion. SurfPro (Song et al., 2024) is a surface-conditioned framework. Using the same splitting strategy as the compared systems (Jing et al., 2020; Dauparas et al., 2022; Gao et al., 2022a), proteins in CATH 4.2 were partitioned into 18,024/608/1,120 samples for training, validation, and testing, respectively. To compare with ESM-IF (Hsu et al., 2022), structures in CATH 4.3 were split into 16,153/1,457/1,797 samples for training, validation, and testing, respectively. To provide a head-to-head comparison with ESM-IF, no additional data, such as AF2DB (Varadi et al., 2022), was used to train SurfDesign. To evaluate generative quality thoroughly, we report perplexity and the median AAR rate in the short-chain, single-chain, and all-chain settings, as usual. The multi-chain protein design employs the dataset curated by Dauparas et al. (2022), which was pre-processed by clustering sequences at 30% sequence identity, yielding 25,361 clusters. Following ProteinMPNN's setup, the clusters were randomly divided into 23,358/1,464/1,539 samples for training, validation, and testing, respectively. This strategy ensures that none of the target chain's chains or biounits were present in the other two sets.

Table 5. Sequence design performance and ablation studies on CATH 4.2 held-out test split. The **best performance** is shown in bold, while the best baseline is indicated with an underline. ESM-IF is tested on CATH 4.2, although it was originally trained and evaluated on CATH 4.3.

Models	Trainable/Total Params.	Perplexity (↓)			Median AAR (↑)		
		Short	Single-chain	All	Short	Single-chain	All
StructGNN (Ingraham et al., 2019)	1.4M / 1.4M	8.29	8.74	6.40	29.44	28.26	35.91
GraphTrans (Ingraham et al., 2019)	1.56M / 1.56M	8.39	8.83	6.63	28.14	28.46	35.82
GCA (Tan et al., 2023)	2.1M / 2.1M	7.09	7.49	6.05	32.62	31.10	37.64
GVP (Jing et al., 2020)	1.0M / 1.0M	7.23	7.84	5.36	30.60	28.95	39.47
AlphaDesign (Gao et al., 2022b)	3.6M / 3.6M	7.32	7.63	6.30	34.16	32.66	41.31
ProteinMPNN (Dauparas et al., 2022)	1.9M / 1.9M	6.21	6.68	4.61	36.35	34.43	45.96
ESM-IF (Hsu et al., 2022)	142M / 142M	6.93	6.65	3.96	35.28	33.78	48.95
PiFold (Gao et al., 2022a)	6.6M / 6.6M	6.04	6.31	4.55	39.84	38.53	51.66
LM-Design-MPNN (Zheng et al., 2023)	5.0M / 659M	7.01	6.58	4.41	35.19	40.00	54.41
LM-Design-PiFold (Zheng et al., 2023)	11.9M / 664M	6.77	6.46	4.52	37.88	42.47	55.65
DPLM (Wang et al., 2024)	5.0M / 659M	–	–	–	–	–	54.54
InstructPLM (Qiu et al., 2024)	89.1M / 6.6B	3.22	3.17	2.68	61.59	59.29	57.51
KW-Design (Gao et al., 2023)	6.4M / 798M	5.48	5.16	3.46	44.66	45.45	60.77
VFN-IF (Mao et al., 2023)	5.4M / 5.4M	5.70	5.86	4.17	41.34	40.98	54.74
VFN-IF-ESM (Mao et al., 2023)	5.4M / 15B	4.92	4.22	3.36	50.00	52.13	62.67
SurfPro (Song et al., 2024)	5.8M / 5.8M	–	–	3.13	–	–	57.78
PRISM (Mahbub et al., 2025)	– / –	3.74	2.68	2.71	60.89	40.98	60.43
GRADE-IF (Yi et al., 2023)	– / –	5.49	6.21	4.35	45.27	42.77	52.21
DMRA (Wang et al., 2024)	– / –	4.06	4.76	2.93	53.57	48.95	64.07
MapDiff (Bai et al., 2025)	14.7M / 14.7M	3.96	4.41	3.43	54.04	49.34	60.93
SurfDesign-backbone	5.3M / 656M	3.28	3.11	3.16	63.45	64.32	65.12
SurfDesign (w/o PLMs)	5.3M / 5.3M	3.21	3.10	3.08	62.70	64.88	65.35
SurfDesign (w/o SEMP)	4.8M / 655M	3.08	2.93	2.76	65.43	67.06	66.27
SurfDesign	5.3M / 656M	**2.43**	**2.44**	**2.41**	**73.74**	**75.17**	**74.13**

Single-chain Protein Design. Tab. 5 and 9 document the results on the CATH (Orengo et al., 1997) benchmark, where SurfDesign consistently achieves state-of-the-art performance in distinct settings. Under a controlled CATH-only training setting, SurfDesign is the first surface-conditioned model to exceed 70% AAR on CATH 4.2 and CATH 4.3, demonstrating its superior ability to restore effective protein sequences. On the full CATH 4.2, SurfDesign achieves a perplexity of 2.41 and an AAR of 74.13%, outperforming the previous state-of-the-art VFN-IF-ESM (Mao et al., 2023) by 28.27% and 18.28%, respectively. It also induces AAR improvements of 19.72% and 26.78% on the short and single-chain subsets, respectively. Furthermore, SurfDesign surpasses SurfPro, another surface-conditioned algorithm, by 23.00% and 28.29% in the overall metrics, respectively. The outstanding phenomenon also exists for the CATH 4.3 benchmark, where SurfDesign outperforms the strongest

Table 6. Structure recovery based on the self-consistent protocol (Yim et al., 2023). ‡: results are quoted from Mao et al. (2023).

Metrics	PiFold‡	LM-Design‡	VFN-IF-ESM‡	SurfDesign
scTM > 0.5	90.98%	89.42%	93.29 %	**96.17%**
scRMSD < 2.0	60.35 %	58.41%	64.16%	**72.83%**

competitor, KW-Design (Gao et al., 2023), by 10.60% in perplexity and 19.49% in AAR. To summarize, SurfDesign enhances surface-conditioned sequence generation with greater efficiency, thanks to the significant advancements and open-source contributions from the entire community, building on the foundation laid by previous pioneers. To address data leakage concerns and demonstrate that SurfDesign's ¿70% recovery results more from manifold-aware geometric reasoning than from privileged sequence information, our backbone-only variant achieves 65.12% AAR and still outperforms backbone-only baselines, while testing on noisy ESMFold-predicted structures maintains 68.5% AAR, demonstrating robustness independent of native side-chain accuracy.

Multi-chain Protein Complex Design. A protein functions only when it docks, associates, and interacts with other macromolecules, forming multi-chain protein complexes. Thus, studying protein sequence design for multi-chain assembled structures is crucial, motivating us to assess whether SurfDesign can design a protein complex. From Appendix Tab. 10, we conclude that the AAR is generally higher for longer proteins, and all models achieve higher AAR rates on PDB than CATH datasets. More importantly, SurfDesign achieves the best performance, with an AAR exceeding 80%. This phenomenon indicates that SurfDesign can design both single-chain proteins and multi-chain complexes. This makes SurfDesign more versatile in the categories and scenarios in which it can be deployed, creating opportunities to use it to design specific protein complexes.

Zero-shot Generalization to New Protein Families. TS50 and TS500 are commonly used independent test sets to assess model generalization on unseen proteins, as introduced by Li et al. (2014). Towards this goal, we evaluate SurfDesign trained on CATH 4.2 and 4.3, respectively, and report the results in Tab. 11. We find that SurfDesign outperforms prior studies by a large margin across all benchmarks. Specifically, it achieves a perplexity of 2.05 and an AAR rate of 82.16 on TS50, outperforming the previous state-of-the-art algorithm, VFN-IF-ESM, by 18.65% and 12.08%, respectively. Meanwhile, on the TS500 dataset, SurfDesign obtains a perplexity of 1.98 and an AAR rate of 84.70. These numbers are better than VFN-IF-ESM by 22.04% and 16.80%, respectively. In addition, for those trained in CATH 4.3, SurfDesign consistently achieves the best. In a nutshell, SurfDesign is the first to transcend 82% and 84% AAR on the TS50 and TS500.

3.4. Discussion and Analysis

By elevating molecular surfaces to first-class conditioning signals, SurfDesign shifts protein design from a fold-centric to an interaction-centric paradigm. This perspective aligns naturally with emerging pipelines that generate or refine structures prior to sequence design, and suggests a modular future in which surface geometry, sequence priors, and functional objectives can be composed rather than entangled.

Ablation Studies. We systematically investigate the contributions of SurfDesign's components, shown in Tab. 5. It can be observed that the knowledge of PLMs provides a large improvement of 13.43% in AAR (65.35% $\rightarrow$ 74.13%) and a decrease of 24.29% in perplexity (3.21 $\rightarrow$ 2.43). Moreover, the incorporation of directionality and curvatures also contributes to the superiority of SurfDesign, with improvements of 11.86% in AAR and 12.68% in perplexity.

Table 7. Evaluation on the surface recovery on CATH 4.2.

Models	IoU (↑)	CD (↓)	NC (↑)
LM-Design	0.90	5.972	0.4236
VFN-IF-ESM	0.92	4.688	0.4859
SurfDesign	**0.98**	**2.873**	**0.6241**

Surface Recovery. The ultimate goal of our surface-conditioned design is to generate proteins with higher surface similarity of key regions, such as the binding or interaction site (Lai et al., 2024). To measure the similarity between two 3D molecular shapes, we use three evaluation metrics (Sun et al., 2024) commonly used in 3D modeling from three aspects: volume, distance, and normal vectors. They are Volumetric Intersection over Union (IoU), Chamfer distance (CD), and Normal Consistency (NC) (computational details are in App. C.4). As shown in Tab. 7, SurfDesign reconstructs molecular surfaces well, aligning with the motivation for our surface-conditioned design. Visualization of the generated and ground-truth surfaces is provided in App. D.4.

Structure Recovery. We compare SurfDesign with strong baselines in terms of protein structure recovery on CATH 4.2, reported in Tab. 6. Following standard evaluation procedures (Yim et al., 2023; Mao et al., 2023), ESMFold was used to predict structures of designed sequences. A case study of visualization comparison using Alphafold-3 is displayed in App. D.3. Two self-consistent metrics, scTM (↑) and scRMSD (↓) are leveraged to assess the similarity between desired and designed protein structures. SurfDesign is more likely to generate protein sequences with the expected structures. More analysis of refoldability is elaborated in App. D.2.

Scalability of PLMs. The scaling law w.r.t model sizes of PLMs has recently been studied (Zheng et al., 2023; Qiu et al., 2024). To understand the influence of PLM

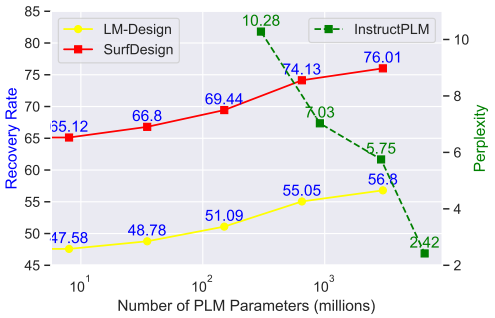

Figure 4. Performance of different PLM scales.

model size on SurfDesign's capacity, we increase the ESM-2 parameter count from 8M to 3B. A similar phenomenon has been discovered in Fig. 4, where the performance of SurfDesign improves as PLMs scale. When integrating knowledge from the largest PLM (3B), SurfDesign achieves a recovery rate of 76.01% on CATH 4.2. This coincidence highlights the significant potential of integrating surface-conditioned design with state-of-the-art PLMs (Kaplan et al., 2020).

Structural Contexts. We dissect the action mechanism of SurfDesign according to different structural contexts in Fig. 5. Structure-based LM-Design shows high AAR on structurally constrained residues in the folding core, while low AAR in structurally less constrained residues on surface areas and loops. SurfDesign significantly enhances the recovery of structurally constrained and less-constrained residues, particularly those on the surface regions.

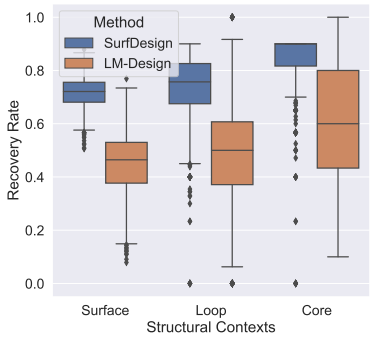

Figure 5. Sequence recovery w.r.t. structural contexts regarding SASA and interaction interface, on CATH 4.2 single-chain proteins.

4. Related Work

Structure-based Protein Design. Advances in AI-driven structure prediction, notably AlphaFold (Jumper et al., 2021), have reinvigorated the complementary task of *inverse folding*. Early inverse folding methods formulated the problem as per-residue classification using multi-layer perceptrons (MLPs) with handcrafted structural features. SPIN (Li & Koehl, 2014) combined torsion angles, sequence profiles, and energy descriptors to achieve 30% AAR on TS50, while SPIN2 (O'Connell et al., 2018) incorporated backbone angles, contact numbers, and inter-residue distances, improving AAR to 34%. Wang et al. (2018) leveraged backbone dihedrals, solvent-accessible surface area, secondary-structure annotations, and unit direction vectors, attaining 33% AAR. Subsequent work replaced MLPs with convolutional architectures to better capture spatial context. SPROF (Chen et al., 2019) applied 2D CNNs to Cα–Cα distance maps, achieving 40.25% AAR on TS500. In three dimensions, ProDCoNN (Zhang et al., 2022a) employed a multi-scale 3D CNN to reach 42.2%, while DenseCPD (Qi & Zhang, 2020) further improved performance to 55.53%.

Recognizing proteins as inherently graph-structured objects, recent methods adopt GNNs to better respect geometric constraints. GraphTrans (Ingraham et al., 2019) introduced a graph-attention encoder with an autoregressive decoder, while GVP (Jing et al., 2020) incorporated geometric vector perceptrons to jointly process scalar and vector features. Building on this foundation, GCA (Tan et al., 2023) integrates global attention across residue graphs; AlphaDesign (Gao et al., 2022b) proposes a streamlined GVP-based encoder with a constraint-aware decoder; ProteinMPNN (Dauparas et al., 2022) combines autoregressive decoding with iterative message passing; and PiFold (Gao et al., 2022a) introduces virtual atoms and explicit backbone dihedral modeling. VFN (Mao et al., 2023) employs learnable vector operations over frame-anchored virtual atoms, pushing AAR to 62.67%.

Despite steady improvements, limited structural data constrain sequence diversity. To mitigate this, ESM-IF (Hsu et al., 2022) leverages large-scale AlphaFold2 predictions to pretrain a GVP-based model. LM-Design (Zheng et al., 2023) fine-tunes ESM-2 conditioning on pretrained structural embeddings, and InstructPLM (Qiu et al., 2024) incorporates explicit *structure prompts* via cross-modal alignment in ProGen2 (Nijkamp et al., 2023). KW-Design (Gao et al., 2023) enhances low-confidence residues using knowledge from ESM and GearNet (Zhang et al., 2022b), while recent work (Wang et al., 2024) shows that self-supervised discrete diffusion models can serve as general protein learners for structure-conditioned sequence generation.

5. Conclusion

In this work, we presented SurfDesign, a surface-conditioned protein design framework that elevates molecular surfaces from auxiliary inputs to first-class conditioning signals for generative protein modeling.

References

Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A., Ronneberger, O., Willmore, L., Ballard, A. J., Bambrick, J., et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, pp.

1–3, 2024.

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.

Bai, P., Miljković, F., Liu, X., De Maria, L., Croasdale-Wood, R., Rackham, O., and Lu, H. Mask-prior-guided denoising diffusion improves inverse protein folding. *Nature Machine Intelligence*, pp. 1–13, 2025.

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

Berman, H. M., Battistuz, T., Bhat, T. N., Bluhm, W. F., Bourne, P. E., Burkhardt, K., Feng, Z., Gilliland, G. L., Iype, L., Jain, S., et al. The protein data bank. *Acta Crystallographica Section D: Biological Crystallography*, 58(6):899–907, 2002.

Chen, S., Sun, Z., Lin, L., Liu, Z., Liu, X., Chong, Y., Lu, Y., Zhao, H., and Yang, Y. To improve protein sequence profile prediction through image captioning on pairwise residue distance map. *Journal of chemical information and modeling*, 60(1):391–399, 2019.

Cock, P. J., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., et al. Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11):1422, 2009.

Connolly, M. L. Analytical molecular surface calculation. *Journal of applied crystallography*, 16(5):548–558, 1983.

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.

DeLano, W. L. et al. Pymol: An open-source molecular graphics tool. *CCP4 Newsl. Protein Crystallogr*, 40(1): 82–92, 2002.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Ektefaie, Y., Viessmann, O., Narayanan, S., Dresser, D., Kim, J. M., and Mkrtchyan, A. Reinforcement learning on structure-conditioned categorical diffusion for protein inverse folding. *arXiv preprint arXiv:2410.17173*, 2024.

Elnaggar, A., Heinzinger, M., Dallago, C., Rihawi, G., Wang, Y., Jones, L., Gibbs, T., Feher, T., Angerer, C., Steinegger, M., et al. Prottrans: towards cracking the language of life's code through self-supervised deep learning and high performance computing. *arXiv preprint arXiv:2007.06225*, 2020.

Gainza, P., Sverrisson, F., Monti, F., Rodola, E., Boscaini, D., Bronstein, M. M., and Correia, B. E. 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*, 617 (7959):176–184, 2023.

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

Gao, Z., Tan, C., and Li, S. Z. Alphadesign: A graph protein design method and benchmark on alphafolddb. *arXiv preprint arXiv:2202.01079*, 2022b.

Gao, Z., Tan, C., Chen, X., Zhang, Y., Xia, J., Li, S., and Li, S. Z. Kw-design: Pushing the limit of protein design via knowledge refinement. In *The Twelfth International Conference on Learning Representations*, 2023.

Gao, Z., Wang, J., Tan, C., Wu, L., Huang, Y., Li, S., Ye, Z., and Li, S. Z. Uniif: Unified molecule inverse folding. *arXiv preprint arXiv:2405.18968*, 2024.

Gasteiger, J., Giri, S., Margraf, J. T., and Gunnemann, S. Fast and uncertainty-aware directional message passing for non-equilibrium molecules. *arXiv preprint arXiv:2011.14115*, 2020a.

Gasteiger, J., Gross, J., and Gunnemann, S. Directional message passing for molecular graphs. *arXiv preprint arXiv:2003.03123*, 2020b.

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. *bioRxiv*, 2022.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Hu, M., Yuan, F., Yang, K., Ju, F., Su, J., Wang, H., Yang, F., and Ding, Q. Exploring evolution-aware &-free protein language models as protein function predictors. *Advances in Neural Information Processing Systems*, 35:38873–38884, 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.

Jing, B., Eismann, S., Suriana, P., Townshend, R. J., and Dror, R. Learning from protein structure with geometric vector perceptrons. *arXiv preprint arXiv:2009.01411*, 2020.

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

Kabsch, W. and Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers: Original Research on Biomolecules*, 22(12):2577–2637, 1983.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Kobayashi, S. and Nomizu, K. *Foundations of differential geometry, volume 2*, volume 61. John Wiley & Sons, 1996.

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, 2023.

Lai, H., Wang, L., Qian, R., Ye, G., Huang, J., Wu, F., Wu, F., Zeng, X., and Liu, W. Interformer: An interaction-aware model for protein-ligand docking and affinity prediction. 2024.

Lee, Y., Yu, H., Lee, J., and Kim, J. Pre-training sequence, structure, and surface features for comprehensive protein representation learning. In *The Twelfth International Conference on Learning Representations*, 2023.

Lester, B., Al-Rfou, R., and Constant, N. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*, 2021.

Li, D., Shen, L., Song, M., Li, D., Liu, J., and Liu, X. Surffold: A unified model for protein inverse folding by integrating surface and structural information. *Bioinformatics*, pp. btaf666, 2025a.

Li, G., Zhao, X., Wu, F., and Laue, S. Joint design of protein surface and backbone using a diffusion bridge model. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025b.

Li, J. and Koehl, P. 3d representations of amino acids—applications to protein sequence comparison and classification. *Computational and structural biotechnology journal*, 11(18):47–58, 2014.

Li, Z., Yang, Y., Faraggi, E., Zhan, J., and Zhou, Y. Direct prediction of profiles of sequences compatible with a protein structure by neural networks with fragment-based local and energy-based nonlocal profiles. *Proteins: Structure, Function, and Bioinformatics*, 82(10):2565–2573, 2014.

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., dos Santos Costa, A., Fazel-Zarandi, M., Sercu, T., Candido, S., et al. Language models of protein sequences at the scale of evolution enable accurate structure prediction. *bioRxiv*, 2022.

Mahbub, S., Kundu, S., and Xing, E. P. Prism: Enhancing protein inverse folding through fine-grained retrieval on structure-sequence multimodal representations. *arXiv preprint arXiv:2510.11750*, 2025.

Mao, W., Zhu, M., Chen, H., and Shen, C. Modeling protein structure using geometric vector field networks. *bioRxiv*, pp. 2023–05, 2023.

Mitra, N. J. and Nguyen, A. Estimating surface normals in noisy point cloud data. In *Proceedings of the nineteenth annual symposium on Computational geometry*, pp. 322–328, 2003.

Nijkamp, E., Ruffolo, J. A., Weinstein, E. N., Naik, N., and Madani, A. Progen2: exploring the boundaries of protein language models. *Cell systems*, 14(11):968–978, 2023.

Nwankpa, C., Ijomah, W., Gachagan, A., and Marshall, S. Activation functions: Comparison of trends in practice and research for deep learning. *arXiv preprint arXiv:1811.03378*, 2018.

O'Connell, J., Li, Z., Hanson, J., Heffernan, R., Lyons, J., Paliwal, K., Dehzangi, A., Yang, Y., and Zhou, Y. Spin2: Predicting sequence profiles from protein structures using deep neural networks. *Proteins: Structure, Function, and Bioinformatics*, 86(6):629–633, 2018.

Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B., and Thornton, J. M. Cath–a hierarchic classification of protein domain structures. *Structure*, 5(8):1093–1109, 1997.

Park, J. J., Florence, P., Straub, J., Newcombe, R., and Lovegrove, S. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 165–174, 2019.

Pearson, W. R. and Sierk, M. L. The limits of protein sequence comparison? *Current opinion in structural biology*, 15(3):254–260, 2005.

Qi, Y. and Zhang, J. Z. Densecpd: improving the accuracy of neural-network-based computational protein sequence design with densenet. *Journal of chemical information and modeling*, 60(3):1245–1252, 2020.

Qiu, J., Xu, J., Hu, J., Cao, H., Hou, L., Gao, Z., Zhou, X., Li, A., Li, X., Cui, B., et al. Instructplm: Aligning protein language models to follow protein structure instructions. *bioRxiv*, pp. 2024–04, 2024.

Rao, R., Bhattacharya, N., Thomas, N., Duan, Y., Chen, P., Canny, J., Abbeel, P., and Song, Y. Evaluating protein transfer learning with tape. *Advances in neural information processing systems*, 32, 2019.

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.

Robinson, E. C., Jbabdi, S., Glasser, M. F., Andersson, J., Burgess, G. C., Harms, M. P., Smith, S. M., Van Essen, D. C., and Jenkinson, M. Msm: a new flexible framework for multimodal surface matching. *Neuroimage*, 100:414–426, 2014.

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.

Savinov, N., Chung, J., Binkowski, M., Elsen, E., and Oord, A. v. d. Step-unrolled denoising autoencoders for text generation. *arXiv preprint arXiv:2112.06749*, 2021.

Sledzieski, S., Kshirsagar, M., Baek, M., Dodhia, R., Lavista Ferres, J., and Berger, B. Democratizing protein language models with parameter-efficient fine-tuning. *Proceedings of the National Academy of Sciences*, 121(26): e2405840121, 2024.

Somnath, V. R., Bunne, C., and Krause, A. Multi-scale representation learning on proteins. *Advances in Neural Information Processing Systems*, 34:25244–25255, 2021.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

Song, Z., Huang, T., Li, L., and Jin, W. Surfpro: Functional protein design based on continuous surface. *arXiv preprint arXiv:2405.06693*, 2024.

Strokach, A., Becerra, D., Corbi-Verge, C., Perez-Riba, A., and Kim, P. M. Fast and flexible protein design using deep graph neural networks. *Cell systems*, 11(4):402–411, 2020.

Su, J., Han, C., Zhou, Y., Shan, J., Zhou, X., and Yuan, F. Saprot: Protein language modeling with structure-aware vocabulary. *bioRxiv*, pp. 2023–10, 2023.

Sun, D., Huang, H., Li, Y., Gong, X., and Ye, Q. Dsr: dynamical surface representation as implicit neural networks for protein. *Advances in Neural Information Processing Systems*, 36, 2024.

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.

Tan, C., Gao, Z., Xia, J., Hu, B., and Li, S. Z. Global-context aware generative protein design. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Tan, Y., Li, M., Zhou, B., Zhong, B., Zheng, L., Tan, P., Zhou, Z., Yu, H., Fan, G., and Hong, L. Simple, efficient and scalable structure-aware adapter boosts protein language models. *arXiv preprint arXiv:2404.14850*, 2024.

Tang, X., Ye, X., Wu, F., Shao, D., Xu, D., and Gerstein, M. Bc-design: A biochemistry-aware framework for highly accurate inverse protein folding. In *ICML 2025 Generative AI and Biology (GenBio) Workshop*, 2025.

Tian, X., Ran, H., Wang, Y., and Zhao, H. Geomae: Masked geometric target prediction for self-supervised point cloud pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13570–13580, 2023.

Trippe, B. L., Yim, J., Tischer, D., Baker, D., Broderick, T., Barzilay, R., and Jaakkola, T. Diffusion probabilistic modeling of protein backbones in 3d for the motif-scaffolding problem. *arXiv preprint arXiv:2206.04119*, 2022.

Tunyasuvunakool, K., Adler, J., Wu, Z., Green, T., Zielinski, M., Žídek, A., Bridgland, A., Cowie, A., Meyer, C., Laydon, A., et al. Highly accurate protein structure prediction for the human proteome. *Nature*, 596(7873):590–596, 2021.

Van Kempen, M., Kim, S. S., Tumescheit, C., Mirdita, M., Lee, J., Gilchrist, C. L., Soding, J., and Steinegger, M. Fast and accurate protein structure search with foldseek. *Nature biotechnology*, 42(2):243–246, 2024.

Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G., Yuan, D., Stroe, O., Wood, G., Laydon, A., et al. Alphafold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic acids research*, 50(D1):D439–D444, 2022.

Wang, C., Zhong, B., Zhang, Z., Chaudhary, N., Misra, S., and Tang, J. Pdb-struct: A comprehensive benchmark for structure-based protein design. *arXiv preprint arXiv:2312.00080*, 2023.

Wang, J., Cao, H., Zhang, J. Z., and Qi, Y. Computational protein design with deep learning neural networks. *Scientific reports*, 8(1):1–9, 2018.

Wang, X., Zheng, Z., Ye, F., Xue, D., Huang, S., and Gu, Q. Diffusion language models are versatile protein learners. *arXiv preprint arXiv:2402.18567*, 2024.

Wu, F. A semi-supervised molecular learning framework for activity cliff estimation. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pp. 6080–6088, 2024.

Wu, F. Diffantiseq: A controllable diffusion model for efficient antibody library design. In *LLM for Scientific Discovery: Reasoning, Assistance, and Collaboration*, 2025.

Wu, F. A semi-supervised molecular learning framework for activity cliff estimation. *arXiv preprint arXiv:2601.04507*, 2026.

Wu, F. and Li, S. Z. Diffmd: A geometric diffusion model for molecular dynamics simulations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 5321–5329, 2023.

Wu, F. and Li, S. Z. A hierarchical training paradigm for antibody structure-sequence co-design. *Advances in Neural Information Processing Systems*, 36, 2024a.

Wu, F. and Li, S. Z. Surface-vqmae: Vector-quantized masked auto-encoders on molecular surfaces. In *International Conference on Machine Learning*, pp. 53619–53634. PMLR, 2024b.

Wu, F. and Li, S. Z. Dynamics-inspired structure hallucination for protein-protein interaction modeling. *arXiv preprint arXiv:2601.06214*, 2026.

Wu, F., Jin, S., Jiang, Y., Jin, X., Tang, B., Niu, Z., Liu, X., Zhang, Q., Zeng, X., and Li, S. Z. Pre-training of equivariant graph matching networks with conformation flexibility for drug binding. *Advanced Science*, 9(33): 2203796, 2022a.

Wu, F., Li, S., Wu, L., Li, S. Z., Radev, D., and Zhang, Q. Discovering the representation bottleneck of graph neural networks from multi-order interactions. *arXiv preprint arXiv:2205.07266*, 2022b.

Wu, F., Courty, N., Jin, S., and Li, S. Z. Improving molecular representation learning with metric learning-enhanced optimal transport. *Patterns*, 4(4), 2023a.

Wu, F., Qin, H., Gao, W., Li, S., Coley, C. W., Li, S. Z., Zhan, X., and Xu, J. Instructbio: A large-scale semi-supervised learning paradigm for biochemical problems. *arXiv preprint arXiv:2304.03906*, 2023b.

Wu, F., Radev, D., and Li, S. Z. Molformer: Motif-based transformer on 3d heterogeneous molecular graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 5312–5320, 2023c.

Wu, F., Wu, L., Radev, D., Xu, J., and Li, S. Z. Integration of pre-trained protein language models into geometric deep learning networks. *Communications Biology*, 6(1): 876, 2023d.

Wu, F., Hu, B., and Li, S. Z. Generalized implicit neural representations for dynamic molecular surface modeling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 877–885, 2025a.

Wu, F., Xuan, W., Qi, H., Lu, X., Tu, A., Li, L. E., and Choi, Y. Deepsearch: Overcome the bottleneck of reinforcement learning with verifiable rewards via monte carlo tree search. *arXiv preprint arXiv:2509.25454*, 2025b.

Wu, F., Zhou, Z., Jin, S., Zeng, X., Leskovec, J., and Xu, J. Surface-based molecular design with multi-modal flow matching. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 3192–3203, 2025c.

Wu, F., Jin, S., Tang, X., Gerstein, M., Zeng, X., Choi, Y., Leskovec, J., and Xu, J. Surfdesign: Effective protein design on molecular surfaces, 2026a. URL https://arxiv.org/abs/2606.07567.

Wu, F., Jin, S., Tang, X., Xu, J., Gerstein, M., Li, L. E., and Zou, J. D-flow: Multi-modality flow matching for d-peptide design. *IEEE Journal of Biomedical and Health Informatics*, 2026b.

Wu, F., Xuan, W., Qi, H., Cao, H., Chang, H.-J., Zhou, Z., Zhao, H., Jian, M., Ma, C., Cheng, Y.-C., et al. Proteo-r1: Reasoning foundation models for de novo protein design. *arXiv preprint arXiv:2605.02937*, 2026c.

Wu, R., Ding, F., Wang, R., Shen, R., Zhang, X., Luo, S., Su, C., Wu, Z., Xie, Q., Berger, B., et al. High-resolution de novo structure prediction from primary sequence. *BioRxiv*, pp. 2022–07, 2022c.

Xu, J., Gao, Z., Zhou, X., Hu, J., Cheng, X., Song, L., Chen, G., Heng, P.-A., and Qiu, J. Protein inverse folding from structure feedback. *arXiv preprint arXiv:2506.03028*, 2025.

Yi, K., Zhou, B., Shen, Y., Liò, P., and Wang, Y. Graph denoising diffusion for inverse protein folding. *Advances in Neural Information Processing Systems*, 36:10238–10257, 2023.

Yim, J., Trippe, B. L., De Bortoli, V., Mathieu, E., Doucet, A., Barzilay, R., and Jaakkola, T. Se (3) diffusion model with application to protein backbone generation. *arXiv preprint arXiv:2302.02277*, 2023.

Zhang, N., Bi, Z., Liang, X., Cheng, S., Hong, H., Deng, S., Lian, J., Zhang, Q., and Chen, H. Ontoprotein: Protein pretraining with gene ontology embedding. *arXiv preprint arXiv:2201.11147*, 2022a.

Zhang, X., Li, H., Cheng, Z., et al. Curvature estimation of 3d point cloud surfaces through the fitting of normal section curvatures. *Proceedings of ASIAGRAPH*, 2008 (23-26):2, 2008.

Zhang, Y. and Skolnick, J. Scoring function for automated assessment of protein structure template quality. *Proteins: Structure, Function, and Bioinformatics*, 57(4):702–710, 2004.

Zhang, Y., Huang, W., Wei, Z., Yuan, Y., and Ding, Z. Equipocket: an e (3)-equivariant geometric graph neural network for ligand binding site prediction. *arXiv preprint arXiv:2302.12177*, 2023.

Zhang, Z., Xu, M., Jamasb, A., Chenthamarakshan, V., Lozano, A., Das, P., and Tang, J. Protein representation learning by geometric structure pretraining. *arXiv preprint arXiv:2203.06125*, 2022b.

Zheng, Z., Deng, Y., Xue, D., Zhou, Y., Ye, F., and Gu, Q. Structure-informed language models are protein designers. In *International conference on machine learning*, pp. 42317–42338. PMLR, 2023.

A. More Related Work

Protein Surface Modeling. The characteristics of the molecular surface dictate the types and strengths of interactions a protein can have with other molecules (Wu & Li, 2024a; Wu, 2024; 2026; 2025; Wu et al., 2022a; 2026c; 2023c; 2022b; Wu & Li, 2023; Wu et al., 2023b;a;d; 2026b; 2025a; Wu & Li, 2026). It is defined by van der Waals (vdW) radii (Connolly, 1983) and is commonly represented as meshes derived from signed distance functions. MaSIF (Gainza et al., 2020) pioneered the use of mesh-based geometric DL to abstract the internal parts of the protein fold and explore protein interactions. A subsequent study (Sverrisson et al., 2021) reduced pre-computation costs by modeling molecular surfaces as point clouds, assigning atom categories to each point. Other seminal works have linked protein surfaces with structural information in a multimodal manner (Somnath et al., 2021), incorporating comprehensive pretraining strategies (Wu & Li, 2024b) using implicit neural representations (INRs) (Park et al., 2019) for self-supervised learning (Lee et al., 2023) and dynamic structure modeling (Sun et al., 2024). Despite these efforts, surface-conditioned design remains underexplored. Recent advancements, such as the work by Gainza et al. (2023) on expanding MaSIF for *de novo* binder design and SurfPro (Song et al., 2024), which eliminates the need for handcrafted feature calculations, have begun to address this gap by directly generating functional proteins from surface data.

Parameter-efficient Fine-tuning. Training and storing full copies of large PLMs(Lin et al., 2022; Rao et al., 2019; Elnaggar et al., 2020; Wu et al., 2025b) for various downstream tasks is increasingly impractical. PEFT techniques (Sledzieski et al., 2024), such as LoRA (Hu et al., 2021) and prompt tuning (Lester et al., 2021), achieve competitive or superior performance compared to full fine-tuning, significantly reducing memory for tasks like interaction prediction and homo-oligomer symmetry prediction. Recent work integrates structural information into PLMs via PEFT. LM-Design (Zheng et al., 2023) introduces a lightweight adapter to realize structural awareness, referred to as *structural surgery* on PLMs. SES-Adapter (Tan et al., 2024) integrates structural data by converting it into sequential vectors using tools such as FoldSeek (Van Kempen et al., 2024) and DSSP (Kabsch & Sander, 1983), thereby enabling cross-modal attention calculations. It outperforms structure-aware PLMs such as SaProt (Su et al., 2023) on standard datasets, including thermostability, metal-ion binding, gene ontology annotations, and subcellular localization prediction.

B. Mathematical Analysis

Here we demonstrate that the curvature feature ψ is roto-translation invariant. Firstly, suppose we translate the entire neighborhood $\mathcal{N}_{(i)}$ by a vector $\mathbf{t} \in \mathbb{R}^3$, so each point $\mathbf{x}_j \in \mathcal{N}_{(i)}$ is transformed to $\mathbf{x}'_j = \mathbf{x}_j + \mathbf{t}$. - When computing the covariance matrix $\mathbf{\Sigma}$, the centroid $\overline{\mathbf{x}}$ is subtracted from each point in $\mathcal{N}_{(i)}$. The centroid after translation becomes $\overline{\mathbf{x}}' = \overline{\mathbf{x}} + \mathbf{t}$, so the translated covariance matrix becomes:

$$\mathbf{\Sigma}' = \frac{1}{\left\|\mathcal{N}_{(i)}\right\|} \sum_{\mathbf{x}_j \in \mathcal{N}_{(i)}} (\mathbf{x}_j + \mathbf{t}) (\mathbf{x}_j + \mathbf{t})^\top - \overline{\mathbf{x}}' \overline{\mathbf{x}}'^\top. \tag{10}$$

Expanding this, we get:

$$\mathbf{\Sigma}' = \frac{1}{\left\|\mathcal{N}_{(i)}\right\|} \sum_{\mathbf{x}_j \in \mathcal{N}_{(i)}} \mathbf{x}_j \mathbf{x}_j^\top + \mathbf{t}\mathbf{t}^\top + 2\mathbf{t} \cdot \sum_{\mathbf{x}_j \in \mathcal{N}_{(i)}} \mathbf{x}_j^\top / \left\|\mathcal{N}_{(i)}\right\| - (\overline{\mathbf{x}} + \mathbf{t})(\overline{\mathbf{x}} + \mathbf{t})^\top. \tag{11}$$

This simplifies back to the original $\mathbf{\Sigma}$ since $\mathbf{t}$ terms cancel out in the computation of $\mathbf{\Sigma}$ after translating by $\mathbf{t}$. Therefore, the covariance matrix $\mathbf{\Sigma}$ is invariant under translations.

Suppose we apply a rotation $\mathbf{R} \in \mathrm{SO}(3)$ to all points in $\mathcal{N}_{(i)}$, where $\mathbf{R}$ is an orthogonal matrix with determinant 1. Then each point $\mathbf{x}_j \in \mathcal{N}_{(i)}$ is transformed to $\mathbf{x}'_j = \mathbf{R}\mathbf{x}_j$. The centroid $\overline{\mathbf{x}}$ also transforms under the rotation, so the new centroid is $\overline{\mathbf{x}}' = \mathbf{R}\overline{\mathbf{x}}$. The covariance matrix $\mathbf{\Sigma}'$ after rotation becomes:

$$\mathbf{\Sigma}' = \frac{1}{\left\|\mathcal{N}_{(i)}\right\|} \sum_{\mathbf{x}_j \in \mathcal{N}_{(i)}} \mathbf{R}\mathbf{x}_j (\mathbf{R}\mathbf{x}_j)^\top - \overline{\mathbf{x}}' \overline{\mathbf{x}}'^\top. \tag{12}$$

Expanding the terms, we obtain:

$$\mathbf{\Sigma}' = \mathbf{R} \left(\frac{1}{\left\|\mathcal{N}_{(i)}\right\|} \sum_{\mathbf{x}_j \in \mathcal{N}_{(i)}} \mathbf{x}_j \mathbf{x}_j^\top - \overline{\mathbf{x}}\overline{\mathbf{x}}^\top \right) \mathbf{R}^\top = \mathbf{R}\mathbf{\Sigma}\mathbf{R}^\top. \tag{13}$$

Since a rotation is a similarity transformation, the eigenvalues of $\mathbf{\Sigma}'$ are the same as those of $\mathbf{\Sigma}$. Therefore, the eigenvalues ϵ_1, ϵ_2, and ϵ_3, which are used to compute ψ, remain unchanged under rotations.

C. Experimental Details

C.1. Training and metrics

The models were trained for 50 epochs by default, using the Adam optimizer on 4 A100 GPUs. We used the same training settings as ProteinMPNN (Dauparas et al., 2022) and LM-Design (Zheng et al., 2023), with a batch size of approximately 6000 residues and the Adam optimizer configured with a NOAM learning rate scheduler. Following previous works, perplexity and *median* AAR scores are reported. In Tab. 5 and 9, two subsets of the entire test set are also reported. Particularly, the SHORT set contains proteins up to length 100, and the SINGLE CHAIN set contains proteins recorded as a single chain in PDB (Berman et al., 2002).

Protein Binder Design. We evaluate SurfDesign on a curated benchmark of experimentally validated protein-protein binding complexes spanning multiple target categories. For target categories with sufficient data, complexes are split into training, validation, and test sets using an 8:1:1 ratio; categories with limited data are evaluated in a zero-shot setting and used only for testing. All models are fine-tuned on the same training split to ensure fair comparison.

Given a target protein structure, SurfDesign generates binder sequences conditioned on the target surface geometry. We report two complementary evaluation settings. (i) *Greedy decoding*, where a single binder sequence is generated per target and evaluated directly. (ii) *Stochastic sampling*, where we sample $K = 10$ binder sequences per target using a softmax temperature of 0.1.

Functional binding quality is assessed using the AlphaFold2-predicted aligned error between the binder and target chains ($\text{pAE}_{\text{interaction}}$), which has been shown to correlate with binding likelihood. Lower $\text{pAE}_{\text{interaction}}$ indicates stronger predicted binding. For each designed binder, we first predict its monomer structure using ESMFold, superimpose it onto the binder chain position of the native complex, and then compute $\text{pAE}_{\text{interaction}}$ using AlphaFold2. A designed binder is considered *successful* if its $\text{pAE}_{\text{interaction}}$ is lower than that of the corresponding native positive binder. We additionally report randomly sampled sequences of matched length as a negative control.

Enzyme Design. We evaluate enzyme design using a benchmark comprising multiple enzyme classes, each paired with a specific substrate. To avoid potential data leakage from inverse folding pretraining corpora, all enzymes that overlap with the pretraining datasets are removed prior to data splitting and evaluation.

For enzyme classes with more than 100 samples, we perform clustering-based splitting followed by an 8:1:1 train/validation/test partition. Smaller enzyme classes are evaluated exclusively in a zero-shot setting. All methods are fine-tuned on the same training split when available.

Enzyme functionality is evaluated using the Enzyme–Substrate Potential (ESP) score, which quantifies the compatibility between a designed enzyme structure and its substrate; lower ESP indicates more favorable interactions. As in binder design, we report results for both greedy decoding and stochastic sampling ($K = 10$, temperature=0.1). A generated enzyme sequence is considered *successful* if its ESP score improves upon that of the native enzyme for the same substrate.

C.2. Implementation for Surface Generation

PyMol is used to generate surfaces in our implementation. We have tried the fast sampling algorithm introduced by dMaSIF (Sverrisson et al., 2021) and used by later studies (Wu & Li, 2024b), which approximates the protein surface as the level set of a smooth distance function. However, this sampling mechanism exhibits unacceptable randomness and is therefore abandoned in favor of SurfDesign. As for the biochemical feature computation, we follow MaSIF (Gainza et al., 2020) and calculate three key invariant point inputs, including the Poisson Boltzmann electrostatics using APBS [1], the hydrophobicity [2], and the free electrons/protons [3]. After a further ablation study, we discover that the hydrophobicity and

[1] https://github.com/LPDI-EPFL/masif/blob/master/source/triangulation/computeAPBS.py

[2] https://github.com/LPDI-EPFL/masif/blob/master/source/triangulation/computeHydrophobicity.py

[3] https://github.com/LPDI-EPFL/masif/blob/master/source/triangulation/compuSurfDesignteCharges.py

the charge are pivot to the performance improvement while the electrostatics is not necessary.

For Fig. 5, we employ RSA to determine the surface and core. To be specific, residues with an RSA greater than 0.25 are considered on the surface, while residues with an RSA less than 0.1 are regarded as core residues. We use the DSSP algorithm to decide the loop regions.

C.3. Dataset Information

Tab. 8 documents the vertex count statistics for the CATH datasets. We observe an equal distribution of vertices across different splits. Besides, comparing our surface with SurfPro (Song et al., 2024), it can be found that our surface is more sparse, with nearly half of the average number of vertices per residue. This difference is due to the different computational techniques employed by various software for surface generation (*e.g.*, PyMOL and MSMS).

Table 8. Vertex counts statistics for surfaces from the CATH 4.2 and CATH 4.3 datasets.

Vertex Count	CATH 4.2			CATH 4.3		
	Train	Validation	Test	Train	Validation	Test
Average Vertex Count Per Residue	53.47	53.56	53.31	53.36	55.27	53.11
Maximum Vertex Count	27,817	25,614	25,433	27,110	27,817	25,968
Minimum Vertex Count	1,923	2,315	2,022	1,923	2,011	2,000
Preprocess Time Per Protein		0.38s			0.36s	

C.4. Surface Comparison

C.4.1. Evaluation Metrics

Motivated by DSR (Sun et al., 2024), we employ IoU, CD, and NC to assess the similarity between the molecular surfaces of designed proteins and target proteins. For simplicity, these three metrics are normalized to the range $0-1$. They provide a comprehensive evaluation of the model's performance from different perspectives and are defined as follows.

IoU. IoU compares the reconstructed volume with the ground truth shape (higher is better). For two arbitrary shapes $A, B \subseteq \mathbb{S} \in \mathbb{R}^n$ is attained by $\text{IoU} = \frac{|A \cap B|}{|A \cup B|}$.

CD. CD is a standard metric to evaluate the distance between two point sets $\mathcal{X}_1, \mathcal{X}_2 \subset \mathbb{R}^n$ (lower is better) as $\mathrm{d_C}\left(\mathcal{X}_1, \mathcal{X}_2\right) =$
$\frac{1}{2}\left(\mathrm{d}_{\overrightarrow{\mathrm{C}}}\left(\mathcal{X}_1, \mathcal{X}_2\right) + \mathrm{d_C}\left(\mathcal{X}_2, \mathcal{X}_1\right)\right)$, where $\mathrm{d}_{\overrightarrow{\mathrm{C}}}\left(\mathcal{X}_1, \mathcal{X}_2\right) = \frac{1}{|\mathcal{X}_1|}\sum_{\boldsymbol{x}_1 \in \mathcal{X}_1} \min_{\boldsymbol{x}_2 \in \mathcal{X}_2} \|\boldsymbol{x}_1 - \boldsymbol{x}_2\|$.

NC. NC evaluates estimated surface normals (higher is better). Normal consistency between two normalized unit vectors n_i and n_j is defined as the dot product between the two vectors. For evaluating the surface normals, given the object surface points and normal vectors: $X_{\text{pred}} = \{(\boldsymbol{x}_i, \overrightarrow{n_i})\}$, and the ground truth surface points and normal vectors: $X_{gt} = \left\{\left(\boldsymbol{y}_j, \overrightarrow{m_j}\right)\right\}$, the surface normal consistency between X_{pred} and X_{gt}, denoted as Γ, is defined as: $\Gamma\left(X_{gt}, X_{\text{pred}}\right) = \frac{1}{|X_{gt}|}\sum_{j \in |X_{gt}|}\left|\overrightarrow{n_j} \cdot \vec{m}_{\theta\left(\boldsymbol{y}_j, X_{\text{pred}}\right)}\right|$, where $\theta\left(\boldsymbol{y}_j, X_{\text{pred}} := \{(\boldsymbol{x}_i, \overrightarrow{n_i})\}\right) = \underset{i \in |X_{\text{pred}}|}{\arg\min}\left\|\boldsymbol{y}_j - \boldsymbol{x}_i\right\|_2^2$.

D. Additional Results and Visualization

D.1. More Results of Non-functional Inverse Design

While the primary goal of SurfDesign is functional protein design—including protein-protein binding and enzyme-substrate interactions —we also report extended inverse folding results as a *diagnostic analysis* of structural compatibility. Importantly, these experiments are **not** intended to evaluate functional optimality, nor do they assume a unique ground-truth sequence for a given structure.

Purpose and interpretation. Inverse folding benchmarks assess whether a model can generate amino-acid sequences that are globally consistent with a specified backbone or molecular surface geometry. As discussed in the main text, a single protein structure or surface can accommodate many valid sequences, and the experimentally observed sequence is neither

unique nor necessarily optimal. Accordingly, metrics such as AAR and perplexity should be interpreted as *proxies for geometric and structural consistency*, rather than as supervised label-recovery objectives.

Extended results. Tab. 9, 10 and 11 report additional inverse folding results on CATH 4.3, PDB, and TS50/TS500 benchmarks under settings consistent with prior work. Across all benchmarks, SurfDesign maintains strong performance relative to both backbone-only and prior surface-conditioned baselines, achieving consistently low perplexity and high median AAR. These results indicate that incorporating manifold-aware surface representations does not compromise global fold compatibility, and in fact improves geometric conditioning compared to backbone-only models.

No functional supervision or privileged information. We emphasize that no residue identities, evolutionary information (e.g., MSAs), functional annotations, or active-site labels are used as inputs during surface generation or model training. Surface features are computed solely from atomic geometry and physicochemical attributes derived from structure. Furthermore, inverse-folding datasets are disjoint from the functional design benchmarks used for binder and enzyme evaluation, thereby preventing cross-task leakage.

Relation to functional design. The strong inverse folding performance of SurfDesign should be viewed as evidence that surface-conditioned generation preserves global structural validity, rather than as the model's primary objective. In functional design settings—where no unique target sequence exists—success is instead measured by downstream functional proxies (e.g., AF2 $\text{pAE}_{\text{interaction}}$ and ESP score). Together, these results suggest that SurfDesign achieves a favorable balance: it maintains structural compatibility while enabling more precise control over surface geometry, which is critical for functional protein design.

Overall, the extended inverse folding results support the central claim of this work: explicitly modeling molecular surfaces as continuous geometric manifolds provides a strong and reliable conditioning signal that generalizes beyond functional tasks, without introducing shortcut learning or label memorization.

Table 9. Sequence design on CATH 4.3. †: SINGLE-CHAIN in Hsu et al. (2022) is defined differently.

Models	**Perplexity** (↓)			**AAR** (↑)		
	Short	Single-chain	All	Short	Single-chain	All
GVP (Hsu et al., 2022)	7.68	†6.12	6.17	32.60	39.40	39.20
ProteinMPNN (Dauparas et al., 2022)	6.31	6.32	4.85	40.30	39.02	48.25
ESM-IF (Hsu et al., 2022)	8.18	†6.33	6.44	31.30	38.50	38.30
+ 1.2M AF2 Data	6.05	†4.00	4.01	38.10	51.50	51.60
PiFold (Gao et al., 2022a)	5.88	5.55	4.47	42.86	43.69	50.68
VFN-IF (Mao et al., 2023)	–	–	–	45.34	53.70	52.18
UniIF (Gao et al., 2024)	–	–	–	45.41	54.46	53.05
LM-Design-MPNN (Zheng et al., 2023)	5.88	5.66	4.19	45.71	46.15	56.38
LM-Design-PiFold (Zheng et al., 2023)	5.66	5.52	4.01	46.84	48.63	56.63
KW-Design (Gao et al., 2023)	5.47	5.23	3.49	43.86	45.95	60.38
MapDiff (Bai et al., 2025)	–	–	–	55.56	54.99	60.68
SurfDesign	**5.08**	**4.97**	**3.12**	**66.74**	**71.30**	**72.14**

Table 10. Performance on multi-chain protein complex dataset (*i.e.*, PDB).

Models	**AAR** (↑)			
length	$L < 100$	$100 \leq L < 500$	$500 \leq L < 1000$	Full
StructGNN (Ingraham et al., 2019)	0.41	0.41	0.42	0.41
GraphTrans (Ingraham et al., 2019)	0.40	0.39	0.40	0.40
GCA (Tan et al., 2023)	0.41	0.41	0.42	0.41
GVP (Jing et al., 2020)	0.44	0.42	0.45	0.43
AlphaDesign (Gao et al., 2022b)	0.48	0.49	0.50	0.49
ProteinMPNN (Dauparas et al., 2022)	0.52	0.53	0.55	0.53
PiFold (Gao et al., 2022a)	0.54	0.58	0.60	0.58
LM-Design-MPNN (Zheng et al., 2023)	–	–	–	0.61
LM-Design-GVP (Zheng et al., 2023)	–	–	–	0.62
KWDesign (Gao et al., 2023)	0.59	0.66	0.67	0.66
SurfDesign	**0.74**	**0.79**	**0.82**	**0.81**

Table 11. Performance comparison on TS50 and TS500. Following prior work, we primarily report results from models trained on CATH 4.2. Numbers in the brackets are results from models trained on CATH 4.3.

Models	TS50			TS500		
	Perplexity (↓)	AAR (↑)	Worst (↑)	Perplexity (↓)	AAR (↑)	Worst (↑)
DenseCPD (Qi & Zhang, 2020)	–	50.71	–	–	55.53	–
StructGNN (Ingraham et al., 2019)	5.40	43.89	26.92	4.98	45.69	0.05
GraphTrans (Ingraham et al., 2019)	5.60	42.20	29.22	5.16	44.66	0.03
GVP (Jing et al., 2020)	4.71	44.14	33.73	4.20	49.14	0.09
GCA (Tan et al., 2023)	5.09	47.02	28.87	4.72	47.74	0.03
AlphaDesign (Gao et al., 2022b)	5.25	48.36	32.31	4.93	49.23	0.03
KW-Design (Gao et al., 2023)	3.10	62.79	39.31	2.86	69.19	0.02
VFN-IF (Mao et al., 2023)	3.58	59.54	–	3.19	63.65	–
VFN-IF-ESM (Mao et al., 2023)	2.52	73.30	–	2.54	72.49	–
InstructPLM (Qiu et al., 2024)	2.29	67.99	–	2.42	64.22	–
PRISM (Mahbub et al., 2025)	2.43	67.92	–	2.43	67.92	–
ProteinMPNN-DPO (Xu et al., 2025)	4.85	45.91	–	4.26	48.23	–
InstructPLM-DPO (Xu et al., 2025)	2.52	62.01	–	2.17	66.37	–
ProteinMPNN (Dauparas et al., 2022)	3.93 (3.62)	54.43 (54.22)	37.24 (41.18)	3.53 (3.27)	58.08 (57.23)	0.03 (0.04)
PiFold (Gao et al., 2022a)	3.86 (3.70)	58.72 (59.68)	37.93 (38.14)	3.44 (3.70)	60.42 (59.95)	0.03 (0.05)
LM-Design-MPNN (Zheng et al., 2023)	3.82 (3.60)	56.92 (58.13)	35.17 (39.14)	2.13 (2.15)	64.30 (63.76)	0.04 (0.04)
LM-Design-PiFold (Zheng et al., 2023)	3.50 (3.27)	57.89 (61.38)	39.74 (46.75)	3.19 (3.09)	67.78 (66.56)	0.02 (0.04)
SurfDesign	**2.05 (2.03)**	**82.16 (83.44)**	**41.30 (47.81)**	**1.98 (1.96)**	**84.70 (85.12)**	**0.10 (0.08)**

D.2. Refoldability Analysis

Following Wang et al. (2023), firstly, to assess whether the generated sequences can respect the structure condition, we evaluate the agreement of the ground truth structure with the predicted structures using the TM-score (Zhang & Skolnick, 2004). We refer to this metric as Ref-TM. Furthermore, to evaluate the folding stability of the generated sequences, we compute the mean per-residue confidence estimate, pLDDT, predicted by the structure prediction models, which we refer to as Ref-pLDDT. As pLDDT is a reliable predictor of disorder (Tunyasuvunakool et al., 2021), AlphaFold-2, OmegaFold (Wu et al., 2022c), and ESMFold (Lin et al., 2022) are leveraged as a structure prediction model, which helps minimize deviations due to the choice of model.

We evaluate SurfDesign on the same 82 test samples as in the CATH dataset, and the results are reported in Tab. 12. We observe that SurfDesign stands out as the leading design method across the refoldability metrics, competitive with ProteinMPNN. It achieves 0.89 Ref-TM and 89.42 Ref-pLDDT with AlphaFold-2 prediction. ProteinMPNN is slightly behind with a 0.87 Ref-TM and 87.89 Ref-pLDDT, followed by LM-Design.

Table 12. Refoldability metric and AAR metric on the CATH dataset. We employ **bold** and underline to highlight the best and suboptimal results on each metric. We use TM and pLDDT to represent Ref-TM and Ref-pLDDT.

Design method	ESMFold		OmegaFold		AlphaFold-2		AAR%
	TM	pLDDT	TM	pLDDT	TM	pLDDT	
Wildtype	0.80	74.91	0.75	78.39	0.90	89.87	100
Uniform	0.05	27.68	0.05	31.53	0.06	33.68	5.00
Natural frequencies	0.07	30.53	0.07	35.59	0.06	35.02	5.84
AF-Design	0.53	61.37	0.53	72.04	0.52	75.29	15.95
ESM-Design	0.38	59.65	0.38	62.66	0.37	60.02	17.33
StructTrans	0.72	68.85	0.64	70.35	0.79	80.66	35.89
GVP	0.73	69.67	0.67	74.33	0.83	84.29	39.46
ProteinMPNN	0.80	76.53	0.76	**80.75**	0.87	87.89	41.44
PiFold	0.71	67.55	0.64	70.21	0.82	82.54	44.86
LM-Design	0.73	72.12	0.70	77.58	0.85	87.26	51.23
SurfDesign	**0.81**	**79.35**	**0.76**	80.11	**0.89**	**89.42**	**70.19**

In addition to the AAR rate, we have incorporated Foldable Diversity and sc-TM, as recommended by Ektefaie et al. (2024), to further assess the diversity and self-consistency of the generated sequences. Foldable Diversity evaluates only those sequence pairs that are structurally consistent with the input protein backbone, providing a more targeted diversity metric

that avoids penalizing high-quality, diverse designs. Self-consistency TM score (sc-TM), following Trippe et al. (2022), gauges the consistency of structural predictions for generated sequences, leveraging a fixed threshold of $TM_{\min} = 0.7$ as implemented by Ektefaie et al. (2024). We refer to https://github.com/flagshippioneering/pi-rldif for computation, and the results are shown below. The analysis shows that SurfDesign maintains high structural consistency while exhibiting competitive diversity, outperforming other methods on foldable diversity metrics and providing substantial evidence of the model's ability to generate high-quality, diverse sequences that remain faithful to the structural constraints of input proteins.

Table 13. Foldable diversity on CATH-all.

Model	Foldable Diversity ↑	sc-TM ↑
ProteinMPNN (T=0, RD)	20%	0.80
ProteinMPNN (T=0.1)	23%	0.67
ProteinMPNN (T=0.2)	3%	0.30
ProteinMPNN (T=0.3)	0.1%	0.14
PiFold (T=0.1)	23%	0.72
PiFold (T=0.2)	8%	0.38
KWDesign (T=0.1)	18%	0.79
KWDesign (T=0.2)	23%	0.58
SurfDesign	**23%**	**0.84**

D.3. Structure Visualization

Here, we visualize several protein structure restoration results of SurfDesign, as shown in Fig. 6 and Fig. 7. The designed structures were obtained using the latest AlphaFold 3 (Abramson et al., 2024) [4]

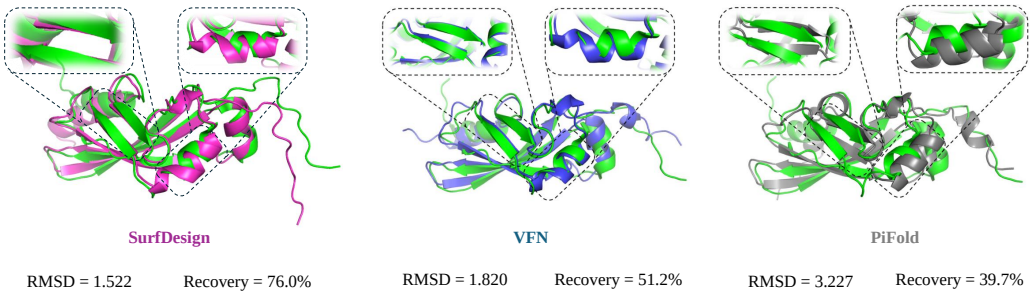

Figure 6. Visualization results of a challenging sample (PDB 2KRT). We use AlphaFold3 to recover the structure from the predicted sequence and compare it with the experimentally determined ground-truth structure.

D.4. Surface Visualization

In addition to restoring protein structure, we quantify surface similarity between the designed and ground-truth proteins. We envision the surfaces of the designed and target proteins in Fig. 8. A substantial overlap is observed between the point clouds of the designed protein surface and the ground-truth protein surface, with a relatively low CD and a significantly high IoU. All of these indicate that SurfDesign produces proteins with the expected surface shapes.

[4]We employed the Alphafold Server for inference at https://alphafoldserver.com/.

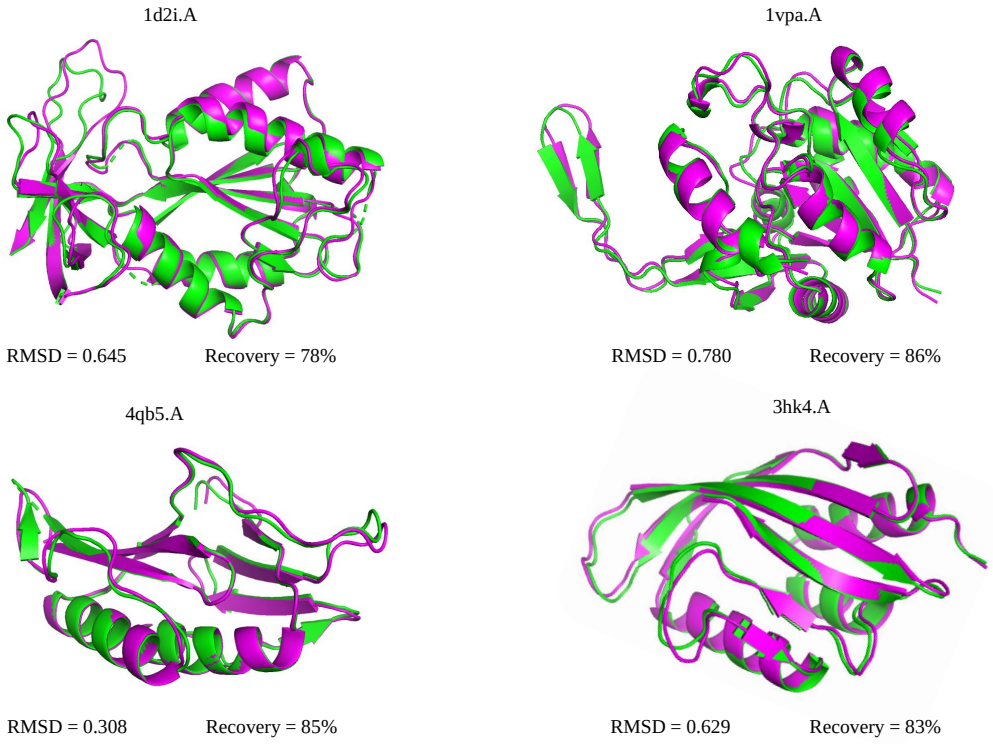

Figure 7. Visualization of SurfDesign, where the green and pink ones are ground truth and designed structures, respectively.

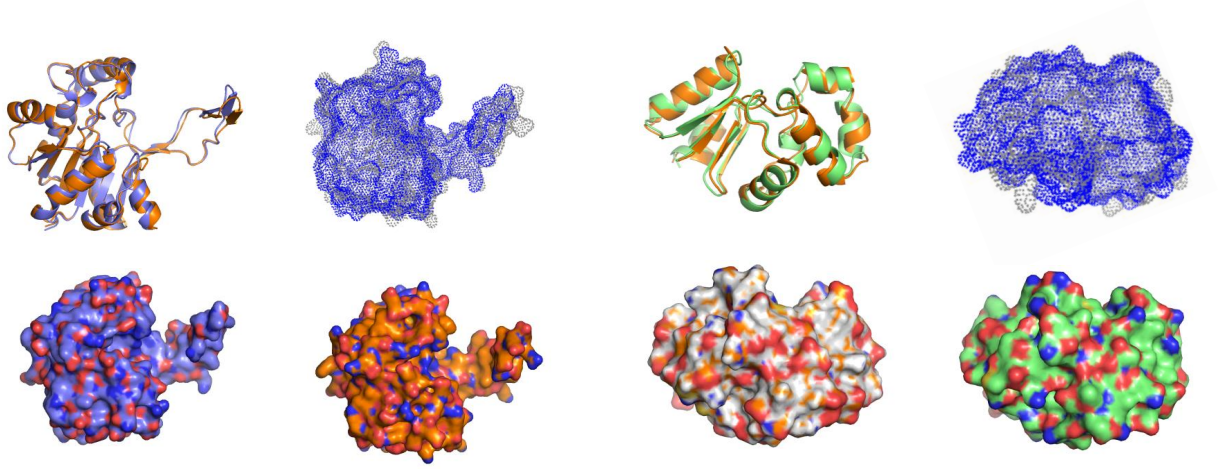

Figure 8. Comparison between original and designed surfaces, where molecular surfaces are visualized from two perspectives: the point cloud view and the manifold view.