ProteinReDiff: Complex-based ligand-binding proteins redesign by equivariant diffusion-based generative models

Cite as: Struct. Dyn. **11**, 064102 (2024); doi: 10.1063/4.0000271
Submitted: 11 August 2024 · Accepted: 6 November 2024 ·
Published Online: 25 November 2024

Viet Thanh Duy Nguyen,[1] Nhan D. Nguyen,[2] and Truong Son Hy[3,a)]

AFFILIATIONS

[1]FPT Software AI Center, Ho Chi Minh City, Vietnam
[2]Pritzker School of Molecular Engineering, University of Chicago, Chicago, Illinois 60637, USA
[3]Department of Computer Science, University of Alabama at Birmingham, Birmingham, Alabama 35294, USA

Note: Paper published as part of the special topic on Artificial Intelligence and Structural Science.
[a)]**Author to whom correspondence should be addressed:** thy@uab.edu

ABSTRACT

Proteins, serving as the fundamental architects of biological processes, interact with ligands to perform a myriad of functions essential for life. Designing functional ligand-binding proteins is pivotal for advancing drug development and enhancing therapeutic efficacy. In this study, we introduce ProteinReDiff, an diffusion framework targeting the redesign of ligand-binding proteins. Using equivariant diffusion-based generative models, ProteinReDiff enables the creation of high-affinity ligand-binding proteins without the need for detailed structural information, leveraging instead the potential of initial protein sequences and ligand SMILES strings. Our evaluations across sequence diversity, structural preservation, and ligand binding affinity underscore ProteinReDiff's potential to advance computational drug discovery and protein engineering.

I. INTRODUCTION

Proteins, often referred to as the molecular architects of life, play a critical role in virtually all biological processes. A significant portion of these functions involves interactions between proteins and ligands, underpinning the complex network of cellular activities. These interactions are not only pivotal for basic physiological processes, such as signal transduction and enzymatic catalysis, but also have broad implications in the development of therapeutic agents, diagnostic tools, and various biotechnological applications.[1–3] Despite the paramount importance of protein–ligand interactions, the majority of existing studies have primarily focused on protein-centric designs to optimize specific protein properties, such as stability, expression levels, and specificity.[4–8] This prevalent approach, despite leading to numerous advancements, does not fully exploit the synergistic potential of optimizing both proteins and ligands for redesigning ligand-binding proteins. By embracing an integrated design approach, it becomes feasible to refine control over binding affinity and specificity, leading to applications such as tailored therapeutics with reduced side effects, highly sensitive diagnostic tools, efficient biocatalysis, targeted drug delivery systems, and sustainable bioremediation solutions,[9–11] thus illustrating the transformative impact of redesigning ligand-binding proteins across various fields.

Traditional methods for designing ligand-binding proteins have relied heavily on experimental techniques, characterized by systematic but often inefficient trial-and-error processes.[12–14] These methods, while foundational, are time-consuming, resource-intensive, and sometimes fall short in precision and efficiency. The emergence of computational design has marked a transformative shift, offering new pathways to accelerate the design process and gain deeper insights into the molecular basis of protein–ligand interactions. However, even with the advancements in computational approaches, significant challenges remain. Many existing models demand extensive structural information, such as protein crystal structures and specific binding pocket data, limiting their applicability, especially in urgent scenarios like the emergence of novel diseases.[15–17] For instance, during the outbreak of a new disease like COVID-19, the spike proteins of the virus may not have well-characterized binding sites, delaying the development of effective drugs.[18,19] Furthermore, the complexity of binding mechanisms, including allosteric effects and cryptic pockets, adds another layer of difficulty.[20,21] Specifically, many proteins do not exhibit clear binding pockets until ligands are in close vicinity, necessitating extensive simulations to reveal potential binding interfaces.[21,22] While molecular dynamics simulations offer detailed atomistic insights into binding mechanisms, they often prove inadequate for designing high-throughput sequences due to high computational cost.[9,23] This complexity underscores the need for a drug design methodology that is agnostic to predefined binding pockets.

Our study addresses those identified challenges by introducing ProteinReDiff, a **Protein** **Re**design framework based on **Diff**usion models. Originating from the foundational concepts of the Equivariant Diffusion-Based Generative Model for Protein–Ligand Complexes (DPL),[24] ProteinReDiff incorporates key improvements inspired by the representation learning modules from the AlphaFold2 (AF2) architecture.[25] Specifically, we integrate the Outer Product Update (adapted from outer product mean of AF2), single representation attention (SRA) [adapted from multiple sequence alignment (MSA) row attention module], and Triangle Multiplicative Update modules into our Residual Feature Update procedure. These modules collectively enhance the framework's ability to capture intricate protein–ligand interactions, improve the fidelity of binding affinity predictions, and enable more precise redesigns of ligand-binding proteins.

The framework integrates the generation of diverse protein sequences with blind docking capabilities. Starting with a selected protein–ligand pair, our approach stochastically masks amino acids and equivariantly denoises the diffusion model to capture the joint distribution of ligand and protein complex conformations (Fig. 1). Another key feature of our method is blind docking, which predicts how the redesigned protein interacts with its ligand without the need for predefined binding site information, while relying solely on initial protein sequences and ligand SMILES strings.[26] This streamlined approach significantly reduces reliance on detailed structural data, thus expanding the scope for sequence-based exploration of protein–ligand interactions.

In summary, the contributions of our paper are outlined as follows:

- We introduce ProteinReDiff, an efficient computational framework for ligand-binding protein redesign, rooted in equivariant diffusion-based generative models. Our innovation lies in integrating AF2's representational learning modules to enhance the framework's ability to capture intricate protein–ligand interactions.
- Our framework enables the design of high-affinity ligand-binding proteins without reliance on detailed structural information, relying solely on initial protein sequences and ligand SMILES strings.
- We comprehensively evaluate our model's outcomes across multiple design aspects, including sequence diversity, structure preservation, and ligand binding affinity, ensuring a holistic assessment of its effectiveness and applicability in various contexts.

II. RELATED WORK

A. Traditional approaches in protein design

Protein design has historically hinged on computational and experimental strategies that paved the way for modern advancements in the field. These foundational methodologies emphasized the balance between understanding protein structure and engineering novel functionalities, albeit with inherent limitations in scalability and precision. Key traditional approaches include the following:

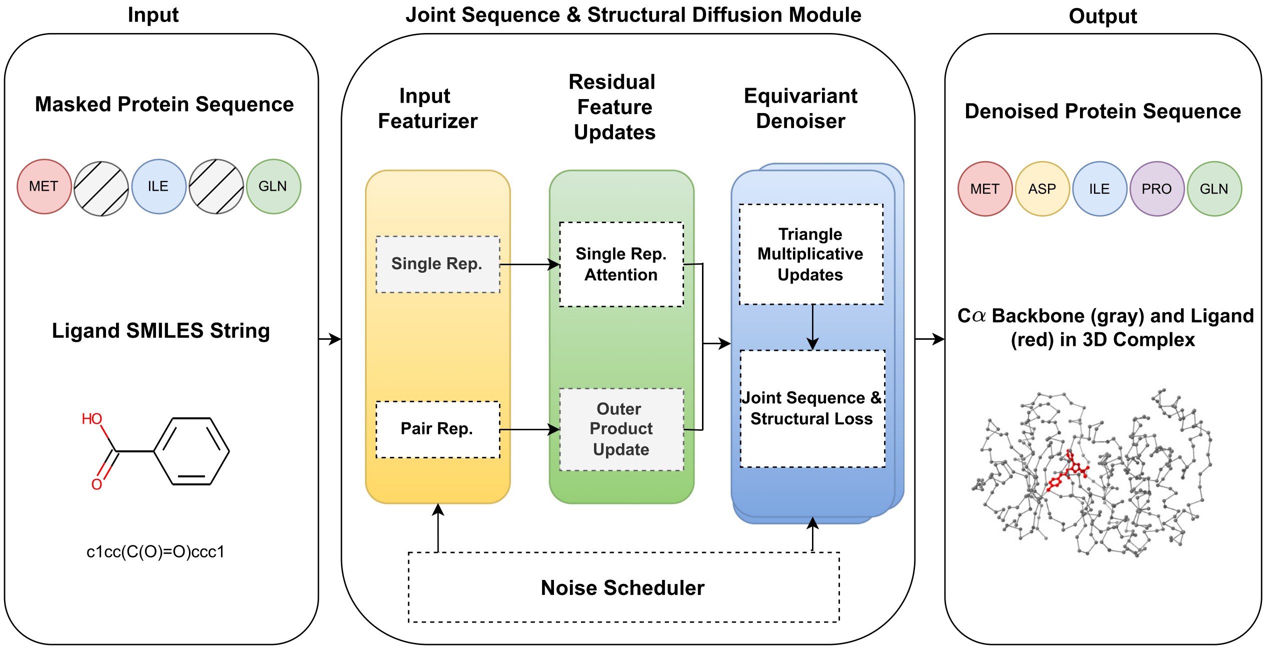


FIG. 1. Overview of the proposed framework. The process begins with utilizing a protein amino acid sequence and a ligand SMILES string as inputs. The joint sequence and structural diffusion process include input featurization, residual feature updates, and equivariant denoiser, ultimately yielding novel protein sequences alongside their corresponding $C\alpha$ protein backbone (gray) and ligand (red) in 3D complexes.

- **Rational Design**[27–29] focused on introducing specific mutations into proteins based on known structural and functional insights. This method required an in-depth understanding of the target protein structures and how changes might impact its function.
- **Directed Evolution**[30–33] mimicked natural selection in the laboratory, evolving proteins toward desired traits through iterative rounds of mutation and selection. Despite its effectiveness in discovering functional proteins, the process was often labor-intensive and time-consuming.

These traditional methods have been instrumental in advancing our understanding and capability in protein design. However, their limitations in terms of efficiency, specificity, and the broad applicability of findings highlighted the need for more versatile and scalable approaches. As the field progressed, the integration of computational power and biological understanding opened new avenues for innovation in protein design, leading to the exploration and adoption of more advanced methodologies.

B. Deep generative models in protein design

Since their inception, deep generative models have significantly advanced fields like computer vision (CV)[34] and natural language processing (NLP),[35] sparking interest in their application to protein design. This enthusiasm has led to numerous studies that harness these models for innovating within the protein design area. Among these, certain types of deep generative models have distinguished themselves through their effectiveness and the promising results they have achieved, including:

- **Variational Autoencoders (VAEs)** are utilized to explore diverse chemical spaces by learning rich latent representations of protein sequences, enabling the generation of novel sequences through latent space manipulation.[36–38]
- **Autoregressive models** predict the probability of each amino acid in a sequential manner, facilitating the generation of coherent and functionally plausible protein sequences.[39,40]
- **Generative adversarial networks (GANs)** employ two networks that work in tandem to produce protein sequences indistinguishable from real ones, enhancing the realism and diversity of generated designs.[41,42]
- **Diffusion models** represent a step forward by gradually transforming noise into structured data, simulating the complex process of folding sequences into functional proteins.[43–46]

However, the majority of these studies have focused on protein-centric designs, with a noticeable gap in research that integrates both proteins and ligands for the purpose of redesigning ligand-binding proteins. Such integration is crucial for a holistic understanding of the intricate dynamics between protein structures and their ligands, a domain that remains underexplored.

C. Current approaches in ligand-binding protein redesign

1. Heavy reliance on detailed structural information

Contemporary computational methodologies for designing proteins that target specific surfaces predominantly rely on structural insights from native complexes, underscoring the critical role of fine-tuning side chain interactions and optimizing backbone configurations for optimal binding affinity.[15–17,44,47,48] These strategies often initiate with the generation of protein backbones, employing inverse folding techniques to identify sequences capable of folding into these pre-designed structures.[6,7,48,49] This approach signifies a paradigm shift by prioritizing structural prediction ahead of sequence identification, aiming to produce proteins that not only fit the desired conformations for potential ligand interactions but also navigate around the challenge of undefined binding sites. Despite the advantages, including the potential of computational docking to create binders via manipulation of antibody scaffolds and varied loop geometries,[36,50,51] a notable challenge persists in validating these binding modes with high-resolution structural evidence. Additionally, the traditional focus on a limited array of hotspot residues for guiding protein scaffold placement often restricts the exploration of possible interaction modes, particularly in cases where target proteins lack clear pockets or clefts for ligand accommodation.[22,52]

2. Limited training data and lack of diversity

Existing approaches often rely on a limited set of training data, which can restrict the diversity and generalizability of the resulting models. For instance, datasets like PDBBind provide detailed ligand information, but their scope is limited.[53] This limitation is further compounded when protein datasets lack corresponding ligand data, reducing the effectiveness of the training process. Traditional methodologies also tend to focus on a narrow range of protein–ligand interactions, potentially overlooking the broader spectrum of possible interactions.

3. Single-domain denoising focus

Previous methodologies typically concentrate on denoising either in sequence space or structural space, but not both. Approaches like ProteinMPNN,[6] LigandMPNN,[17] and MIF[48] primarily operate in sequence space, while others like DPL function in structural space.[24] This single-domain focus can limit the ability to capture the full complexity of protein–ligand interactions, which inherently involve both sequence and structural dimensions. Consequently, these methodologies may fall short of accurately predicting the functional capabilities of redesigned proteins.

4. Challenges in generating diverse sequences with structural integrity

While some approaches prioritize sequence similarity to generate functional proteins, they often do so at the expense of structural integrity. For example, ProteinMPNN and CARP focus heavily on sequence similarity, which can result in a lack of diversity and flexibility in the generated sequences.[6,7] This limitation can hinder the ability to explore a wider range of functional conformations, reducing the effectiveness of the protein design process.

5. Key improvements of ProteinReDiff

We address the weaknesses of available methodologies by integrating diverse datasets, employing a dual-domain denoising strategy, and ensuring the generation of diverse sequences while maintaining structural integrity. Our approach utilizes only protein sequences and ligand SMILES strings, eliminating the need for detailed structural information. By combining PDBBind[53] and CATH[54] datasets, we effectively double our training data, enhancing protein representations. Our equivariant and KL-divergence loss functions enable denoising across both sequence and structural dimensions, capturing the full complexity of protein–ligand interactions. This approach maintains structural fidelity and promotes sequence diversity, overcoming the limitations of methodologies prioritizing sequence similarity at the expense of diversity.

III. BACKGROUND

A. Protein language models (PLMs)

Protein language models (PLMs) harness the power of natural language processing (NLP) to unravel the intricate latency embedded within protein sequences. By analogizing amino acid sequences to human language sentences, PLMs unlock profound insights into protein functions, interactions, and evolutionary trajectories.[55] These models leverage advanced text processing techniques to predict structural, functional, and interactional properties of proteins based solely on their amino acid sequences.[56–59] Their adoption in protein design has catalyzed significant progress, with studies leveraging PLMs to translate protein sequence data[47,60–62] into actionable insights, thus guiding the precise engineering of proteins with targeted functional attributes.

Mathematically, a PLM can be represented as a function F that maps a sequence of amino acids $S = [s_1, s_2, \ldots, s_n]$, where s_i denotes the i-th amino acid in the sequence, to a high-dimensional feature space that encapsulates the protein's structural and functional properties

$$X = F(S), \quad X \in R^d, \tag{1}$$

where X represents the continuous representation or embedding derived from the sequence S and d represents the dimensionality of the embedding space, determined by the PLM's architecture. This embedding captures the complex dependencies and patterns underlying the protein's structural information and biological functionality. Through training on known sequences and structures, PLMs discern the "grammar" governing protein folding and function, facilitating accurate predictions.

We employ the ESM-2 model,[59] a state-of-the-art protein language model with 650×10^6 parameters, pre-trained on nearly 65×10^6 unique protein sequences from the UniRef[63] database, to feature initial masked protein sequences. ESM-2 enriches the latent representation of protein sequences, bypassing the need for conventional multiple sequence alignment (MSA) methods. By incorporating structural and evolutionary information from input sequences, ESM-2 enables us to unravel interaction patterns across protein families for effective ligand targeting. This understanding is crucial for designing and optimizing ligand-binding proteins.

B. Equivariant diffusion-based generative models

We utilize a generative model driven by equivariant diffusion principles, drawing from the foundations laid by variational diffusion models[64] and E(3) equivariant diffusion models.[65]

1. The diffusion procedure

First, we employ a diffusion procedure that is equivariant with respect to the coordinates of atoms x, alongside a series of progressively more perturbed versions of x, known as latent variables z_t, with t varying from 0 to 1. To maintain translational invariance within the distributions, we opt for distributions on a linear subspace that anchors the centroid of the molecular structure at the origin, and designate N_x as a Gaussian distribution within this specific subspace. The conditional distribution of the latent variable z_t given x, for any given t in the interval [0, 1], is defined as

$$q(z_t|x) = N_x(\alpha_t x, \sigma_t^2 I), \tag{2}$$

where α_t and σ_t^2 represent strictly positive scalar functions of t, dictating the extent of signal preservation vs noise introduction, respectively. We implement a variance-conserving mechanism where $\alpha_t = 1 - \sigma_t^2$ and posit that α_t smoothly and monotonically decreases with t, ensuring $\alpha_0 \approx 1$ and $\alpha_1 \approx 0$. Given the Markov property of this diffusion process, it can be described via transition distributions as

$$q(z_t|z_s) = N_x(\alpha_{t|s} z_s, \sigma_{t|s}^2 I), \tag{3}$$

for any $t > s$, where $\alpha_{t|s} = \alpha_t/\alpha_s$ and $\sigma_{t|s}^2 = \sigma_t^2 - \alpha_t^2 \sigma_s^2$. The Gaussian posterior of these transitions, conditional on x, can be derived using Bayes' theorem

$$q(z_s|z_t, x) = N_x(\mu_{t\to s}(z_t, x), \sigma_{t\to s}^2 I), \tag{4}$$

with

$$\mu_{t\to s} = \frac{\alpha_s \sigma_{t|s}^2}{\alpha_{t|s} \sigma_s^2} z_t + \frac{\sigma_s^2 \sigma_t^2}{\sigma_{t|s}^2} x, \quad \sigma_{t\to s}^2 = \frac{\sigma_t^2 \sigma_s^2}{\sigma_{t|s}^2}. \tag{5}$$

2. The generative denoising process

The construction of the generative model inversely mirrors the diffusion process, generating a reverse temporal sequence of latent variables z_t from $t = 1$ back to $t = 0$. By dividing time into T equal intervals, the generative framework can be described as

$$p_\theta(x) = \int_z p(z_1) p(x|z_0) \prod_{i=1}^{T} p_\theta(z_{t_i}|z_{t_{i-1}}), \tag{6}$$

with $s(i) = (i-1)/T$ and $t(i) = i/T$. Leveraging the variance-conserving nature and the premise that $\alpha_1 \approx 0$, we posit $q(z_1) = N_x(0, I)$, hence treating the initial distribution of z_1 as a standard Gaussian

$$p(z_1) = N_x(0, I). \tag{7}$$

Furthermore, under the variance-preserving framework and assuming $\alpha_0 \approx 1$, the distribution $q(z_0|x)$ is modeled as highly peaked.[64,66] This allows us to approximate $p_{\text{data}}(x)$ as nearly constant within this narrow peak region. This yields

$$q(x|z_0) = \frac{q(z_0|x) p_{\text{data}}(x)}{\int_{\tilde{x}} q(z_0|\tilde{x}) p_{\text{data}}(\tilde{x})} \approx \frac{q(z_0|x)}{\int_{\tilde{x}} q(z_0|\tilde{x})} = \mathcal{N}_x(x|z_0/\alpha_0, \sigma_0^2/\alpha_0^2 I). \tag{8}$$

Accordingly, we approximate $q(x|z_0)$ through

$$p(x|z_0) = \mathcal{N}_x(x|z_0/\alpha_0, \sigma_0^2/\alpha_0^2 I). \tag{9}$$

The generative model's conditional distributions are then formulated as

$$p_\theta(z_s|z_t) = q(z_s|z_t, x = \hat{x}_\theta(z_t; t)), \tag{10}$$

which mirrors $q(z_s|Fsz_t, x)$ but substitutes the actual coordinates x with the estimates from a temporal denoising model $\hat{x}_\theta(z_t; t)$, which employs a neural network parameterized by θ to predict x from its noisier version z_t. This denoising model's framework, predicated on noise prediction $\hat{\varepsilon}_\theta(z_t; t)$, is articulated as

$$\hat{x}_\theta(z_t; t) = \frac{(z_t - \sigma_t \hat{\varepsilon}_\theta(z_t; t))}{\alpha_t}. \tag{11}$$

Consequently, the transition mean $\mu_{t\to s}(z_t, \hat{x}_\theta(z_t; t))$ is determined by

$$\mu_{t\to s}(z_t, \hat{x}_\theta(z_t; t)) = \frac{\alpha_s \sigma_{t|s}^2}{\alpha_{t|s}\sigma_s^2} z_t + \frac{\alpha_s \sigma_t^2}{\sigma_{t|s}^2} x = \frac{1}{\alpha_{t|s}} z_t - \frac{\sigma_{t|s}^2}{\alpha_{t|s}\sigma_t} \hat{\varepsilon}_\theta(z_t; t). \tag{12}$$

IV. METHOD

In this section, we detail the methodology employed in our noise prediction model, which is depicted in Fig. 1 and consists of three main procedures: (1) input featurization, (2) residual feature update, and (3) equivariant denoising. Through these steps, we transform raw protein and ligand data into structured representations, iteratively refine their features, and leverage denoising techniques inherent in the diffusion model to improve sampling quality.

A. Input featurization

We develop both single and pair representations from protein sequences and ligand SMILES string (Fig. 2). For proteins, we initially applied stochastic masking to segments of the amino acid sequences. The protein representation is attained through the normalization and linear mapping of the output from the final layer of the ESM-2 model, which is subsequently combined with the amino acid and masked token embeddings. Additionally, for pair representations of proteins, we leveraged pairwise relative positional encoding techniques, drawing from established methodologies.[25] For ligand representations, we employed a comprehensive feature embedding approach, capturing atomic and bond properties such as atomic number, chirality, connectivity, formal charge, hydrogen attachment count, radical electron count, hybridization status, aromaticity, and ring presence for atoms and bond type, stereochemistry, and conjugation status for bonds. These representations are subsequently merged, incorporating radial basis function (RBF) embeddings of atomic distances and sinusoidal embeddings of diffusion times. Together, these steps culminate in the formation of preliminary complex representations, laying the foundation for our computational analyses.

B. Residual feature update procedure

Our Residual Feature Update Procedure, as illustrated in Fig. 3, deviates significantly from the approach employed in the original DPL model.[24] While the DPL model relied on Alphafold2's Triangular Multiplicative Update for updating single and pair representations, where these representations mutually influence each other, our objective is to optimize this procedure for greater efficiency. Specifically, we incorporate enhancements such as the Outer Product Update and single representation attention to formulate sequence representational hypotheses of protein structures and to model suitable motifs for binding target ligands specifically. These modules, integral to Evoformer, the sequence-based module of AF2, play a crucial role in extracting essential connections among internal motifs that serve structural functions (i.e., ligand binding) when structural information is not explicitly

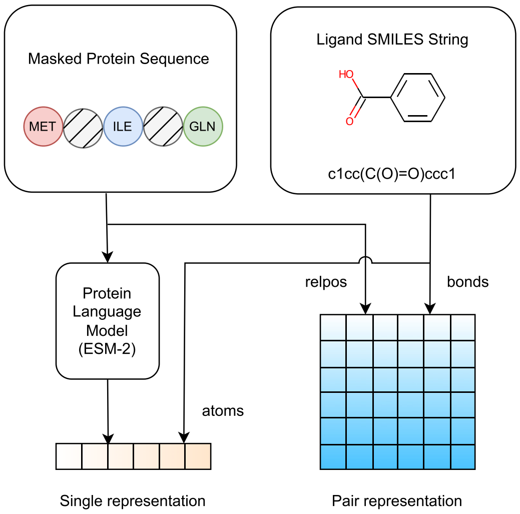


FIG. 2. Overview of the input featurization procedure of the model. Adapted from Ref. 24 to illustrate the specific adaptations made in our model.

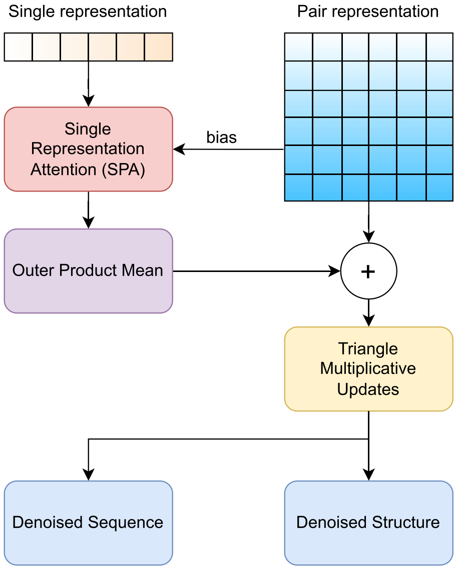


FIG. 3. Overview of the residual feature update procedure of the model. Adapted from Ref. 24 to illustrate the specific adaptations made in our model.

provided during training. Importantly, we adapt and tailor these modules to fit within our model architecture, ensuring their effectiveness in capturing the intricate interplay between proteins and ligands.

1. *Single representation attention module*

The single representation attention (SRA) module, derived from the Alphafold2 model's MSA row attention with pair bias, accounts for long-range interactions among residues and ligand atoms within a single protein–ligand embedding vector. In essence, the attention mechanism assigns importance to those involved in complex-based folding to denoise the equivariant loss (Sec. IV C) in a self-supervised manner. While the original Alphafold2 MSA row attention mechanism processes input for a single sequence, the SRA module is designed to incorporate representations from multiple protein–ligand complexes concurrently. Specifically, the pair bias component of the SRA attention module captures dependencies between proteins and ligands, which was shown to fit the attention score better than the regular self-attention model without bias terms.[67] By considering both the single representation vector (which encodes the protein/ligand sequential representation) and the pairwise representation vector (which encodes protein-protein and protein–ligand interactions), this cross-attention mechanism exchanges information between pairwise and single representation to effectively preserves internal motifs, as evidenced by contact overlap metrics.[55,68] As transformer architecture is widely used for predicting protein functions,[69] we observed similar efficacy to our binding affinity prediction in Results V B 5 and Appendix B and C. For a detailed description of the computational steps implemented in this module, refer to Algorithm 1.

2. *Outer product update*

Since the SRA encodings have a shape (s, r, c_m) and the pair representation has a shape (s, r, r, c_z), the outer product (OPU) layer merges insights by reshaping SRA encodings into pair representations. This module leverages evolutionary cues from ESM to generate plausible structural hypotheses for pair representations.[70] It first calculates the outer product of the SRA embeddings of protein–ligand pairs, then aggregates the outer products to yield a measure of co-evolution between every residue pair.[55] Analogous to tensor product representations (TPR) in NLP, the outer product is akin to the filler-and-role binding relationship, where each entity (i.e., amino acid residue) on a sequence is attached to a rich functional embedding based on its relationship to one another.[71–73]

This process integrates correlated information of residues i and j of a sequence s, resulting in the intermediate Kronecker product tensors (i.e., role embeddings in NLP).[67,74,75] Subsequently, an affine transformation projects those representations to hypotheses concerning the relative positions of residues i and j under biophysical constraints. Our implementation adapts the outer product without computing the mean to maintain the pair representations of multiple protein–ligand complexes. For a detailed description of the computational steps implemented in this module, refer to Algorithm 2.

ALGORITHM 1. Single Representation Attention pseudocode.

Input: Single representation vector m_{si}, pair representation vector z_{sij} of the i-th sequence in the set of sequences s, C = 65, $N_{head} = 4$.
Output: Updated single representation vector $\tilde{m}_{si}$ with the dimension of C_m.

```
$m_{si} \leftarrow \mathrm{LayerNorm}(m_{si})$
$q^h_{si}, k^h_{si}, v^h_{si} \leftarrow \mathrm{LinearNoBias}(m_{si}) \quad q^h_{si}, k^h_{si}, v^h_{si} \in \mathbb{R}^C, h \in \{1, \ldots, N_{\mathrm{head}}\}$
$b^h_{sij} \leftarrow \mathrm{LinearNoBias}(\mathrm{LayerNorm}(z_{sij}))$
$g^h_{si} \leftarrow \mathrm{sigmoid}(\mathrm{Linear}(m_{si})) \quad g^h_{si} \in \mathbb{R}^C$
$a^h_{sij} \leftarrow \mathrm{softmax}_j\left(\frac{1}{\sqrt{C}} q^h_{si} k^{h\,T}_{sj} + b^h_{sij}\right)$
$o^h_{si} \leftarrow g^h_{si} \odot \sum_j a^h_{sij} v^h_{sj}$
$\tilde{m}_{si} \leftarrow \mathrm{Linear}(\mathrm{concat}_h(o^h_{si})) \quad \tilde{m}_{si} \in \mathbb{R}^{C_m}$
return$\{\tilde{m}_{si}\}$
```

ALGORITHM 2. Outer product update pseudocode.

Input: Single representation vector m_{si} of the i-th sequence in the set of sequences s, $C = 32$.
Output: Pair representation vector z_{sij} with the dimension of $s \times C_z$.

```
$m_{si} \leftarrow \mathrm{LayerNorm}(m_{si})$
$a_{si}, b_{si} \leftarrow \mathrm{Linear}(m_{si}) \quad a_{si}, b_{si} \in \mathbb{R}^C$
$o_{sij} \leftarrow \mathrm{flatten}(a_{si} \otimes b_{si}) \quad o_{sij} \in \mathbb{R}^{C \times C}$
$z_{sij} \leftarrow \mathrm{Linear}(o_{sij}) \quad z_{sij} \in \mathbb{R}^{s \times C_z}$
return$\{z_{sij}\}$
```

3. *Triangle multiplicative updates*

After refining the pair representation, our model interprets the primary protein–ligand structure using principles from graph theory, treating each residue as a distinct entity interconnected through the pairwise matrix. These connections are then refined through triangular multiplicative updates to account for physical and geometric constraints, such as triangular inequality. While the SRA weights the importance of residues, the triangular multiplicative update acts as another stack of transformer-based layers where any two edges affect the third one to enforce triangle equivariance.[55,76] The starting and ending nodes propagate information in and out of neighbors in similar fashion as the message-passing framework.[67] These mechanisms enable the model to generate more accurate representations of protein–ligand complexes, leading to improved predictive performance in predicting binding affinities and structural characteristics.

C. Equivariant denoising

During the equivariant denoising process, the final pair representation undergoes symmetrization and is then transformed using a multi-layer perceptron (MLP) into a weight matrix W. This matrix is utilized to compute the weighted sum of all relative differences in three-dimensional (3D) space for each atom, as shown in the equation[24]

$$\hat{\varepsilon}_i(z) = \sum_j W_{ij}(z) \cdot \frac{(z_i - z_j)}{||z_i - z_j||}. \tag{13}$$

Afterward, the centroid is subtracted from this computation, resulting in the output of our noise prediction model $\hat{\varepsilon}$. Additionally, it is important to note that the described model maintains SE(3) equivariance, meaning that

$$\hat{\varepsilon}_i(\mathbf{R}z + \mathbf{t}) = \sum_j \frac{W_{ij}(\mathbf{R}z + \mathbf{t})}{||(\mathbf{R}z_i + \mathbf{t}) - (\mathbf{R}z_j + \mathbf{t})||} \cdot ((\mathbf{R}z_i + \mathbf{t}) - (\mathbf{R}z_j + \mathbf{t})), \tag{14}$$

$$= \mathbf{R} \sum_j \frac{W_{ij}(\mathbf{R}z + \mathbf{t})}{||z_i - z_j||} \cdot (z_i - z_j), \tag{15}$$

$$= \mathbf{R} \sum_j \frac{W_{ij}(z)}{||z_i - z_j||} \cdot (z_i - z_j), \tag{16}$$

$$= \mathbf{R}\hat{\varepsilon}_i(z), \tag{17}$$

for any rotation **R** and translation **t**. This property is derived from the fact that the final representation, and hence the weight matrix W, depends solely on atom distances that are invariant to rotation and translation.

V. EXPERIMENTS

A. Training process

1. Data curation

We curated a broad range of protein structures, including both ligand-bound (holo) and ligand-free (apo) forms, sourced from two key repositories: PDBBind v2020[53] and CATH 4.2.[54] PDBBind v2020 offers a diverse collection of protein–ligand complexes, while CATH 4.2 provides a substantial repository of protein structures. This strategic selection of datasets ensures our model is exposed to a wide and varied spectrum of protein–ligand interactions and structural configurations, enabling comprehensive evaluation against diverse inverse folding benchmarks. By training on both holo and apo structures, our approach imbues the model with a robust understanding of protein–ligand dynamics to navigate the complexities of unseen protein–ligand interactions.

To ensure robust model training and evaluation, we partitioned the datasets by MMseqs2.[77] The protein sets were clustered for training, validation, and testing to maintain sequence similarities between 40% and 50% and ensure unbiased training and predictions. Similar protocols were implemented in other protein models.[25,48] For ligands, we cluster based on the Tanimoto similarity of Morgan fingerprints[78] on ligand structures. Incorporating CATH 4.2 data into PDBBind not only preserves the objectivity of the train/test/validation partitions but also substantially decreases the similarities within ligand sets, as shown in Table I.

Table II provides an overview of the partitioning details, facilitating a clear understanding of the distribution of samples across different subsets of the dataset.

TABLE I. Similarity between train/validation/test sets of proteins and ligands. The values represent similarity percentages for the original PDBBind dataset vs combined PDBBind with CATH datasets in parentheses.

Protein	Validation	Test
Train	36.0% (36.2%)	38.0% (42.2%)
Validation	...	39.08% (43.5%)
Ligand	Validation	Test
Train	72.2% (36.1%)	9.41% (3.11%)
Validation	...	9.37% (3.17%)

- **PDBBind v2020**: For consistency and comparability with previous studies, we first adhered to the test/training/validation split settings outlined in the established literature,[79] specifically following the configurations defined in the respective sources for the PDBBind v2020 datasets.[80] Then, we filtered out those highly similar sequences (above 95%) to keep the average similarities between 40% and 50%.
- **CATH 4.2**: In our approach, we deliberately focused on proteins with fewer than 400 amino acids and less similar (below 90%) sequences from the CATH 4.2 database. This selective criterion was chosen to prioritize smaller proteins, which often represent more druggable targets of interest in drug discovery and development endeavors. During both the training and validation phases, SMILES strings of CATH 4.2 proteins were represented as asterisks (masked tokens) to denote unspecified ligands. Notably, CATH 4.2 was excluded from the test set due to the absence of corresponding ligands required for evaluating protein–ligand interactions.

2. Loss functions

Previous models typically denoise in only one domain, such as ProteinMPNN,[6] LigandMPNN,[17] and MIF[48] in sequence space, and DPL[24] in structural space. These limitations restrict their ability to fully capture the intricate interactions between proteins and ligands. To address this, we have introduced significant modifications to the loss function to better suit the task of ligand-binding protein redesign. By tailoring the loss function to both sequence and structural spaces, our approach addresses the unique challenges of protein–ligand interactions. Specifically, the optimization of our model for ligand-binding protein redesign is governed by a composite loss function L, formulated as follows:

$$L = L_{\mathrm{WS}} + L_{\mathrm{KL}} + L_{\mathrm{CE}}. \tag{18}$$

a. Weighted sum of relative differences (L_{WS}). This component ensures the model's sensitivity to the directional influence between atoms, supporting the accurate prediction of the denoised structure while maintaining physical symmetries. It is crucial for the equivariant

TABLE II. Data partitioning overview (unit: number of samples).

Dataset	Train	Validation	Test
PDBBind v2020	9430	552	207
CATH 4.2	15261	939	...

denoising step, enabling accurate noise prediction for atoms in the protein–ligand complex. The loss is defined as

$$L_{\mathrm{WS}} = \sum_{t=1}^{T} ||\varepsilon - \hat{\varepsilon}_\theta(z;t)||, \qquad (19)$$

where T is the total number of time steps in the diffusion process, ε is the Gaussian noise vector $\mathcal{N}(\mathbf{0}, \mathbf{I})$, and $\hat{\varepsilon}_\theta(z;t)$ is the loss prediction at time step t parameterized by a weight MLP in Sec. IV C.

b. Kullback–Leibler divergence (L_{KL}). This component quantifies the divergence between the model's predictions and actual sequence data at time step $t-1$. Defined as $KL(x_{\mathrm{pred\_t-1}}, seq_{\mathrm{t-1}})$, it contrasts the predicted distribution, $x_{\mathrm{pred\_t-1}}$, against the true sequence distribution, $seq_{\mathrm{t-1}}$, leveraging the diffusion process's β parameter for temporal adjustment. This loss is also applied in the Protein Generator[5] model to ensure the model's predictions progressively align with actual data distributions, enhancing the accuracy of sequence and structure generation by minimizing the expected divergence.[81]

c. Cross-entropy loss (L_{CE}). This loss function is crucial for the accurate prediction of protein sequences, aligning them with the ground truth through effective classification. It denoises each amino acid from masked latent embedding to a specific class, leveraging categorical cross-entropy to rigorously penalize discrepancies between the model's predicted probability distributions and the actual distributions for each amino acid type.

3. Training performance

Throughout the training phase, we observed the model's performance between training and validation losses, as demonstrated in Fig. 4. While the training loss consistently diminished, indicating effective learning, the validation loss exhibited more variability. Despite these fluctuations, the validation loss showed an overall downward trend, suggesting that the model is improving its generalization capabilities over time. The general alignment between the downward trends of training and validation losses indicates that the model is learning effectively without significant overfitting.

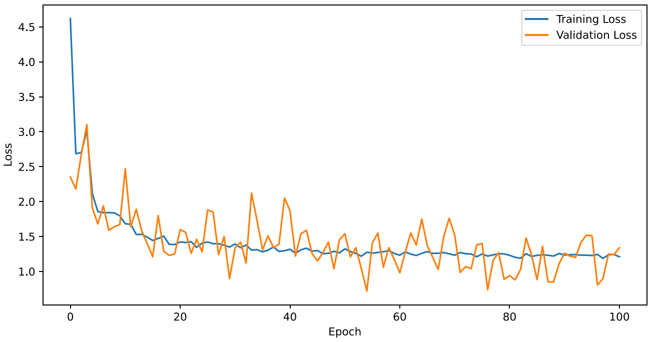


FIG. 4. Training history chart of ProteinReDiff, showcasing the evolution of training and validation losses over epochs.

B. Evaluation process

1. Ligand binding affinity (LBA)

Ligand binding affinity is a fundamental measure that quantifies the strength of the interaction between a protein and a ligand. This metric is crucial as it directly influences the effectiveness and specificity of potential therapeutic agents; higher affinity often translates to increased drug efficacy and lower chances of side effects.[82] Within this context, ProteinReDiff is evaluated on its ability to generate protein sequences for significantly improved binding affinity with specific ligands. We utilize a docking score-based approach for this assessment, where the docking score serves as a quantitative indicator of affinity. Expressed in kcal/mol, these scores inversely relate to binding strength—lower scores denote stronger, more desirable binding interactions.

2. Sequence diversity

Sequence diversity is crucial for exploring protein's functional space.[83] It reflects the capacity of our model, ProteinReDiff, to traverse the vast landscape of protein sequences and generate a wide array of variations. To quantitatively assess this diversity, we utilize the average edit distance (Levenshtein distance)[84] between all pairs of sequences generated by the model. This metric offers a nuanced measure of variability, surpassing traditional metrics that may overlook subtle yet significant differences. The diversity score is calculated using the formula

$$\text{Diversity Score} = \frac{1}{\binom{n}{2}} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} d(S_i, S_j), \qquad (20)$$

where $d(S_i, S_j)$ represents the edit distance between any two sequences S_i and S_j. This calculation provides an empirical gauge of ProteinReDiff's ability to enrich the protein sequence space with novel and diverse sequences, underlining the practical variance introduced by our model.

3. Structure preservation

Structural preservation is paramount in the redesign of proteins, ensuring that essential functional and structural characteristics are maintained post-modification. To effectively measure structural preservation between the original and redesigned proteins, three key metrics are the template modeling score (TM Score),[85] the root mean square deviation (RMSD),[86] and the contact overlap (CO).[87] These three metrics collectively provide a comprehensive assessment of structural integrity and similarity.

a. The root mean square deviation (RMSD). The root mean square deviation (RMSD) is a measure used to quantify the distance between two sets of points. In the context of protein structures, these points are the positions of the atoms in the protein. The RMSD is given by the formula

$$\mathrm{RMSD}(\mathbf{p}, \mathbf{p}') = \min_{(R,t)\in \mathrm{SO}(3)\times \mathbb{R}_3} \left[\frac{1}{N}\sum_{i=1}^{N} ||p_i - (Rp'_i + t)||_2^2\right]^{1/2}, \qquad (21)$$

where $\mathbf{p} = (x_i, y_i, z_i)_{i=1}^{N}$ and $\mathbf{p}' = (x'_i, y'_i, z'_i)_{i=1}^{N}$ denote two sequences of N 3D coordinates representing the atomic positions in the original and redesigned proteins, respectively. This formula calculates the minimum RMSD between corresponding atoms after optimal alignment, using the best-fit rotation R and translation t. A lower RMSD indicates higher structural similarity, reflecting successful preservation of the protein's core structure.

b. TM score. TM score provides a normalized measure of structural similarity between protein configurations, which is less sensitive to local variations and more reflective of the overall topology. The TM score is defined as follows:

$$\text{TM Score}(\mathbf{p}, \mathbf{p}') = \max_{(R,t)\in \mathrm{SO}(3)\times\mathbb{R}_3} \left[\frac{1}{1 + \frac{1}{N}\sum_{i=1}^{N} \frac{\|p_i - (Rp'_i + t)\|_2^2}{d_0^2}} \right], \tag{22}$$

where d_0 is a scale parameter typically chosen based on the size of the proteins. The closer the TM Score is to 1, the more similar the structures are, indicating global structural alignment.

c. Contact overlap (CO). Contact overlap (CO) provides a complementary perspective to RMSD and TM score by focusing on the preservation of local structural motifs rather than overall geometric similarity. Several studies show that having high CO indicates protein's residue pairs having co-evolutionary signals[87,88] and performing related functions.[89] CO quantifies the conservation of inter-atomic contacts between original and redesigned protein structures, essential for structural integrity and function. The metric is defined as

$$\mathrm{CO}(\mathbf{p}, \mathbf{p}') = \frac{|C \cap C'|}{|C \cup C'|}, \tag{23}$$

where $C = \{(i,j) : \|p_i - p_j\| < r_c, i \neq j\}$ and $C' = \{(i,j) : \|p'_i - p'_j\| < r_c, i \neq j\}$ represent the sets of contacts in the original and redesigned proteins, respectively. Here, p_i and p'_i are the positions of atoms in the original and redesigned proteins, and r_c is a predefined cutoff distance that determines when two atoms are considered to be in contact. A high CO score indicates strong preservation of the original contacts in the redesigned structure.

4. Experimental setup

To evaluate ProteinReDiff, we employed Omegafold[90] to predict the three-dimensional structures of all designed protein sequences. The choice of Omegafold over AF2 was favorable because Omegafold can more accurately fold proteins with low similarity to existing proteomes, making it suitable for proteins lacking available ligand-binding conformations. Next, we utilized AutoDock Vina[91] to conduct docking simulations and evaluate the binding affinity between the redesigned proteins and their respective ligands based on the predicted 3D structures. To ensure fair comparisons and mitigate potential biases introduced by pre-docked structures, we aligned our redesigned protein structures with reference structures before docking. This approach is crucial, particularly because the use of pre-docked structures may favor certain conformations, leading to inaccurate evaluations. Additionally, to provide context for our results, we compared the binding scores of our redesigned proteins not only with those of the original proteins but also with proteins generated by other protein design models. While these models may differ in sequence characteristics from those optimized for ligand binding, comparing their scores provides insights into the relationship between protein sequence, structure, and ligand interactions, deepening our understanding of protein–ligand dynamics.

a. Benchmark model selection. In selecting benchmark models for performance comparison, we focused on state-of-the-art approaches, particularly those relevant to protein design tasks. Traditionally, protein design has been primarily based on inverse folding, utilizing protein structure information. Our choices encompass a range of methodologies:

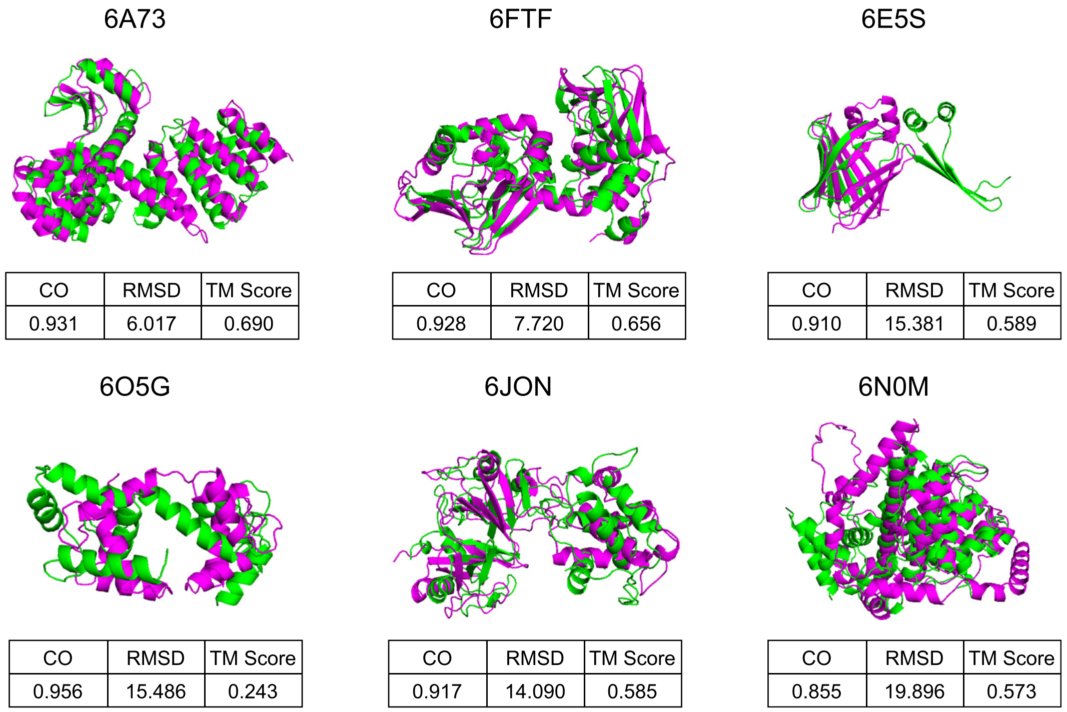
- MIF,[48] MIF-ST,[48] and ProteinMPNN[6] are notable for generating sequences with high identity and experimental significance, utilizing protein structure information.
- The Protein Generator,[5] a representative of RosettaFold models,[44] employs diffusion-based methods, making it an intriguing comparative candidate. The model also shares a similar loss function, L_{KL}, in sequence space with our model but diverges in modules and training procedures (i.e., stochastic masking).
- ESMIF,[49] belonging to the ESM model family,[59] stands as another competitive benchmark, emphasizing the generation of high-quality sequences.

TABLE III. Comparison of protein design models based on input and output characteristics.

	Input				Output		
Model	Protein Sequence	Protein Structure	Ligand SMILES	Binding Pocket	Protein Sequence	Protein Structure	Ligand Structure
CARP[7]	✓	×	×	×	✓	×	×
ESMIF[49]	×	✓	×	×	✓	×	×
MIF[48]	✓	✓	×	×	✓	×	×
MIF-ST[48]	✓	✓	×	×	✓	×	×
ProteinMPNN[6]	×	✓	×	×	✓	×	×
LigandMPNN[17]	×	✓	✓	✓	✓	×	×
Protein generator[5]	×	✓	×	×	✓	×	×
DPL[24]	✓	×	✓	×	×	✓	✓
ProteinReDiff (ours)	✓	×	✓	×	✓	✓	✓

TABLE IV. Comparison of method performance across multiple metrics: ligand binding affinity (LBA), sequence diversity, and structure preservation. Ligand binding affinity (LBA), TM score, and RMSD are reported as mean values with their respective margins of error.

				Structure preservation		
Category	Method	LBA (kcal/mol) ↓	Sequence diversity ↑	TM Score ↑	RMSD (Å) ↓	CO ↑
Baseline	CARP[7]	−5.658 ± 0.301	185.532	0.850 ± 0.023	3.768 ± 0.553	0.922 ± 0.003
	MIF[48]	−5.518 ± 0.381	185.600	**0.877** ± 0.020	**2.986** ± 0.468	0.938 ± 0.002
	MIF-ST[48]	−5.596 ± 0.330	185.584	0.872 ± 0.021	3.026 ± 0.451	0.937 ± 0.003
	ESMIF[49]	−5.555 ± 0.326	187.512	0.837 ± 0.021	4.000 ± 0.501	0.915 ± 0.003
	ProteinMPNN[6]	−5.423 ± 0.225	188.792	0.714 ± 0.026	6.806 ± 0.616	0.859 ± 0.004
	LigandMPNN[17]	−5.717 ± 0.287	**191.384**	0.782 ± 0.024	4.512 ± 0.668	0.915 ± 0.008
	Protein generator[5]	−5.674 ± 0.266	186.962	0.806 ± 0.022	4.431 ± 0.523	0.899 ± 0.003
	DPL[24]	−5.551 ± 0.459	188.139	0.788 ± 0.024	5.094 ± 0.537	0.896 ± 0.009
	Reference cases	−5.847 ± 0.263	...	...	...	...
ProteinReDiff (Ours)	5% masking	−5.805 ± 0.252	185.935	0.864 ± 0.022	3.197 ± 0.470	**0.942** ± 0.007
	15% masking	−**6.803** ± 0.329	186.627	0.845 ± 0.023	3.690 ± 0.508	0.935 ± 0.007
	30% masking	−5.769 ± 0.244	187.877	0.803 ± 0.024	4.467 ± 0.544	0.916 ± 0.008
	40% masking	−5.617 ± 0.366	188.600	0.756 ± 0.026	5.639 ± 0.625	0.896 ± 0.008
	60% masking	−5.467 ± 0.318	190.425	0.305 ± 0.024	18.056 ± 0.773	0.735 ± 0.010
	70% masking	−5.470 ± 0.199	187.291	0.147 ± 0.004	23.197 ± 0.497	0.689 ± 0.007



6A73

CO	RMSD	TM Score
0.931	6.017	0.690

6FTF

CO	RMSD	TM Score
0.928	7.720	0.656

6E5S

CO	RMSD	TM Score
0.910	15.381	0.589

6O5G

CO	RMSD	TM Score
0.956	15.486	0.243

6JON

CO	RMSD	TM Score
0.917	14.090	0.585

6N0M

CO	RMSD	TM Score
0.855	19.896	0.573

FIG. 5. Comparative visualizations of protein structures, each annotated with its corresponding PDB ID. The figure includes a succinct table detailing contact overlap (CO) and root mean square deviation (RMSD) metrics. Original protein structures are highlighted in green, and the redesigned versions by ProteinReDiff are depicted in pink.

- CARP, while lacking ligand information, shares similar protein input and output characteristics with our models, warranting inclusion for comparison.
- DPL,[24] originally geared toward protein–ligand complex generation, was adapted for our purposes by modifying loss functions and incorporating a sequence prediction module, given its alignment with our model architecture.
- LigandMPNN,[17] resembling the most to our task in designing ligand-binding proteins, necessitates binding pocket information, unlike our model, which emphasizes a simplified yet effective approach for ligand-binding protein tasks.

Our model's design prioritizes simplicity in input while achieving effectiveness in output for ligand-binding protein tasks. For a comprehensive comparison of input–output dynamics across each model, please consult Table III.

5. Results and discussion

We conducted comprehensive evaluation of ProteinReDiff, as detailed in Table IV and visually represented in Fig. 6, across the metrics of ligand binding affinity, sequence diversity, and structure preservation. These evaluations provide a clear depiction of the model's performance relative to established baselines and within its variations.

For ProteinReDiff, we aimed to capture the diverse conformations of ligand-binding proteins, recognizing that they can adopt multiple structural states. To assess these conformations, we employed alignment metrics such as TM score, RMSD, and contact overlap (CO). In Fig. 5, we presented several instances where the contact overlap appeared to be maintained, yet the RMSD is large and TM score is low. This discrepancy suggests that while global alignment metrics like TM score and RMSD may not adequately capture the domain shift within these complex ensembles, the preservation of local motifs, as

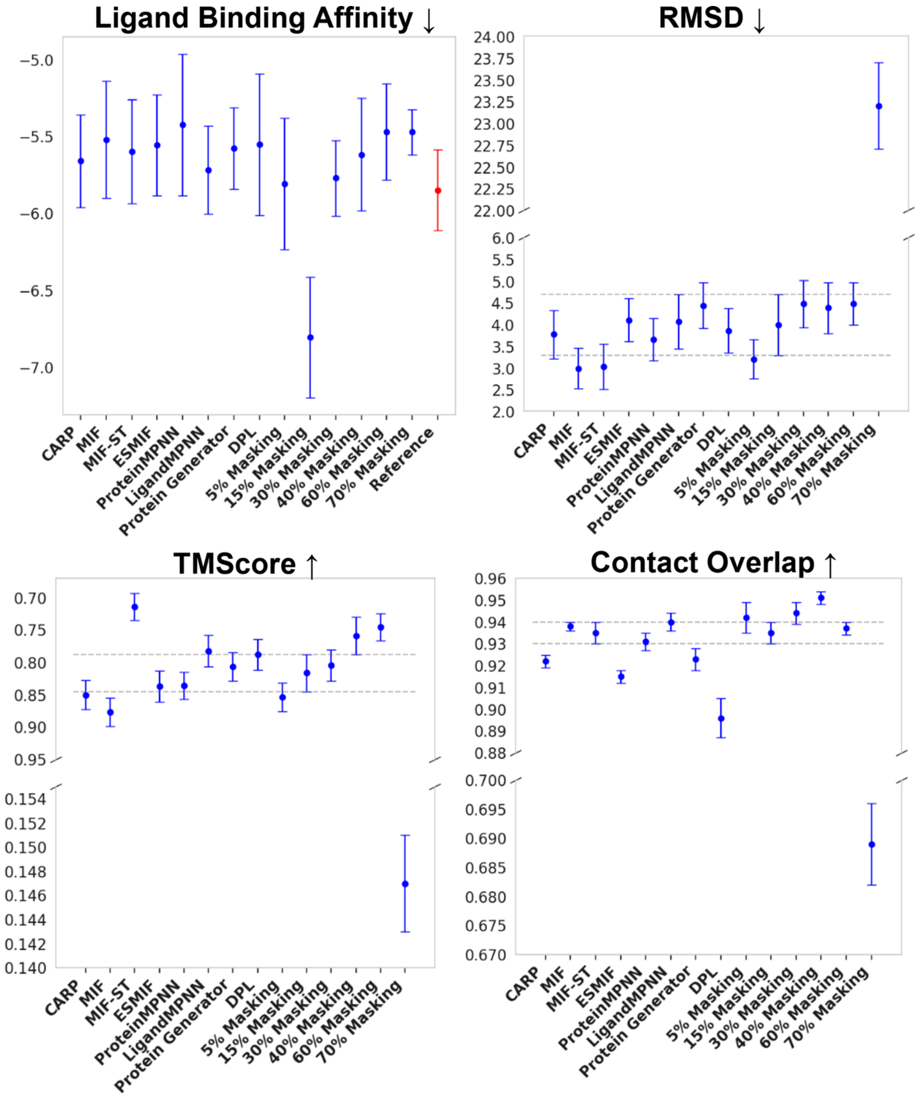
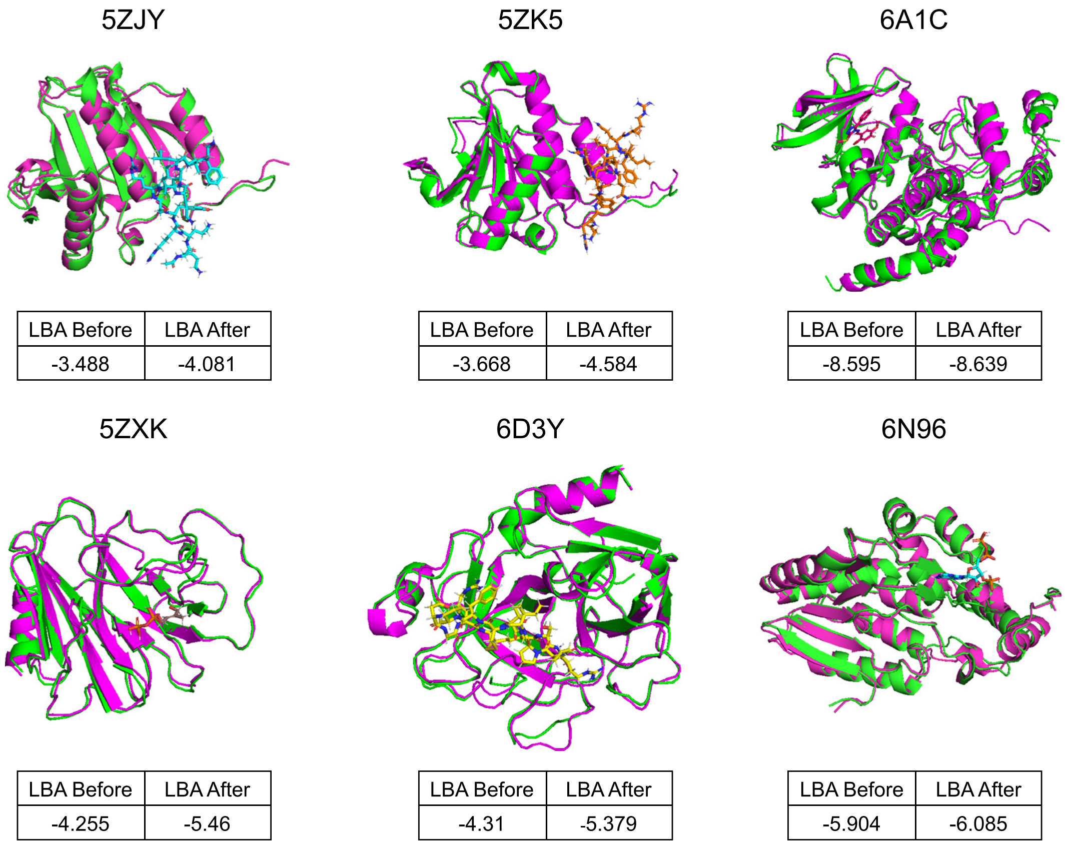


FIG. 6. Visualization of method performance across metrics. The metrics are plotted with mean values and margins of error. For LBA, the red bar (top right) shows the docking score of reference complexes. The horizontal dash lines indicate the regions of 15% masking model which is our standard for comparison. Detailed descriptions of each baseline are provided in Sec. V B 4.



5ZJY

LBA Before	LBA After
-3.488	-4.081

5ZK5

LBA Before	LBA After
-3.668	-4.584

6A1C

LBA Before	LBA After
-8.595	-8.639

5ZXK

LBA Before	LBA After
-4.255	-5.46

6D3Y

LBA Before	LBA After
-4.31	-5.379

6N96

LBA Before	LBA After
-5.904	-6.085

FIG. 7. Comparative visualizations of protein–ligand complexes, each labeled with corresponding PDB IDs and accompanied by a small table showing ligand binding affinity (LBA) before and after the redesign. Original structures are highlighted in green, while redesigned versions by ProteinReDiff appear in pink. Ligands are depicted in various colors to emphasize specific binding sites and molecular interaction enhancements post-redesign.

indicated by contact overlap, remains crucial in our framework. This underscores the importance of capturing both global and local structural features for a comprehensive understanding of protein–ligand interactions.

A pivotal observation from our study is ProteinReDiff's unparalleled ability to enhance ligand binding affinity, particularly at a 15% masking ratio in Fig. 6. This configuration not only surpasses the performance of inverse folding (IF) models and the original DPL framework but also exceeds the binding efficiencies of the original protein designs. By incorporating attention modules from AlphaFold2, ProteinReDiff effectively captures the complex interplay between proteins and ligands, demonstrating its superiority over the original DPL model. While other masking ratios within ProteinReDiff show varying degrees of effectiveness, lower ratios, though at the same par as reference, do not achieve the peak LBA performance observed at 15%. For instance, the 5% masked model emphasizes structural consistency with a high TM score and low RMSD, but does not exhibit the same level of binding capability as the 15% masking. These findings are also consistent with ablation studies shown in Appendix C. Conversely, higher masking ratios fail to strike the necessary balance between introducing beneficial modifications and maintaining functional precision, underscoring the importance of optimizing the masking ratio.

Our analysis of sequence diversity and structure preservation metrics reveals a delicate balance essential in protein redesign. The 15% masking ratio, identified as optimal for enhancing ligand binding affinity in our model, also aligns closely with benchmark methods in both sequence diversity and structure preservation. For instance,

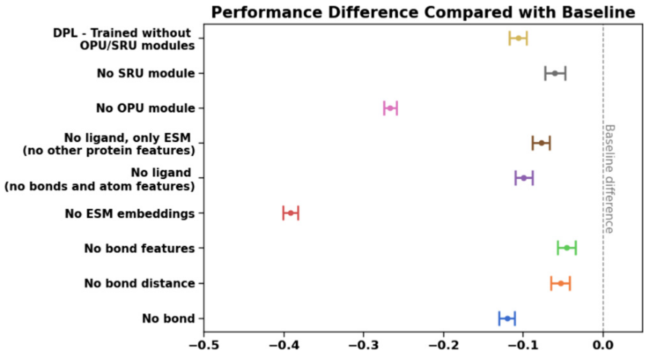


FIG. 8. Ablation studies on ProteinReDiff's model architecture and featurization. The dash line indicates the baseline's average score obtained from ProteinReDiff without ablations.

LigandMPNN excels in sequence diversity but faces challenges in obtaining binding pocket inputs for various design tasks, unlike our approach. Moreover, our models (at 30% and 40% maskings) significantly outperform others in contact overlap, crucial for diversifying structures while preserving functional motifs in protein redesign tasks. This equilibrium underscores ProteinReDiff's ability to optimize ligand interactions without compromising the exploration of sequence diversity or the integrity of original protein structures alone.

In contrast, extreme values in either sequence diversity or structure preservation, which could be seen in other masking ratios, do not lead to optimal ligand binding affinities. This finding highlights an inverse relationship between pushing the limits of diversity and preservation and achieving the primary goal of binding enhancement. Thus, the 15% masking ratio not only stands out for its ability to significantly improve ligand binding affinity but also for maintaining a balanced approach, ensuring that enhancements in functionality do not detract from the protein's structural and functional viability.

In Fig. 7, we compare the ligand-binding affinity (LBA) of original and redesigned proteins by ProteinReDiff. The redesigned proteins maintain their original folds while significantly enhancing LBA. In ablation studies (Sec. V B 6), we can apply various masking strategies to adjust both sequence diversity and structural integrity. This approach has potential applications in different settings to control the affinity of ligand binders.

6. Ablation studies

Here, we conducted thorough ablation studies on ProteinReDiff's model architecture, featurization, and masking ratios. For complete ablation setup, please refer to Table VII (Appendix C).

a. Interpreting model architecture. We trained ablated versions of ProteinReDiff without the SRA or OPU modules and compared them to the original DPL model. Initially designed for generating ensembles of complex structures, DPL was adapted for targeted protein redesign by adding sequence-based loss functions to generate new target sequences.

In Fig. 8, we computed the performance score by averaging the sum of five evaluation metrics introduced in Secs. V B 1–V B 3. Since the sequence diversity is not within the [0,1] range, we applied Min-Max normalization. For LBA and RMSD, we used inverse normalization to ensure that a score closer to 1.0 indicates better model performance. The average score is then compared with the score of baseline ProteinReDiff, which was trained without any ablations.

We observed that our model outperformed DPL by a large margin. Incorporating just the OPU module (without the SRA module) yields better performance than DPL, indicating OPU's ability to exchange insights between single and pair representations. First, the equivariant loss function is parameterized on the structural space, making the pairwise representations from the OPU critical to that loss. Second, without OPU, the model performs poorly on TM score (the bottom brown line in Fig. 11, Appendix C), which measures global structural preservation. Additionally, introducing SRA only without OPU hurts our model performance, suggesting the model would have been over-parameterized as the SRA updates primarily on the sequence representation. Therefore, combining both the OPU and SRA modules provides an effective approach for enhancing the representational learning of ProteinReDiff. A complete comparative assessment is presented in Table IV and Appendix C.

b. Ablations on input featurization methods. We conducted ablation studies to evaluate different input featurization methods, including

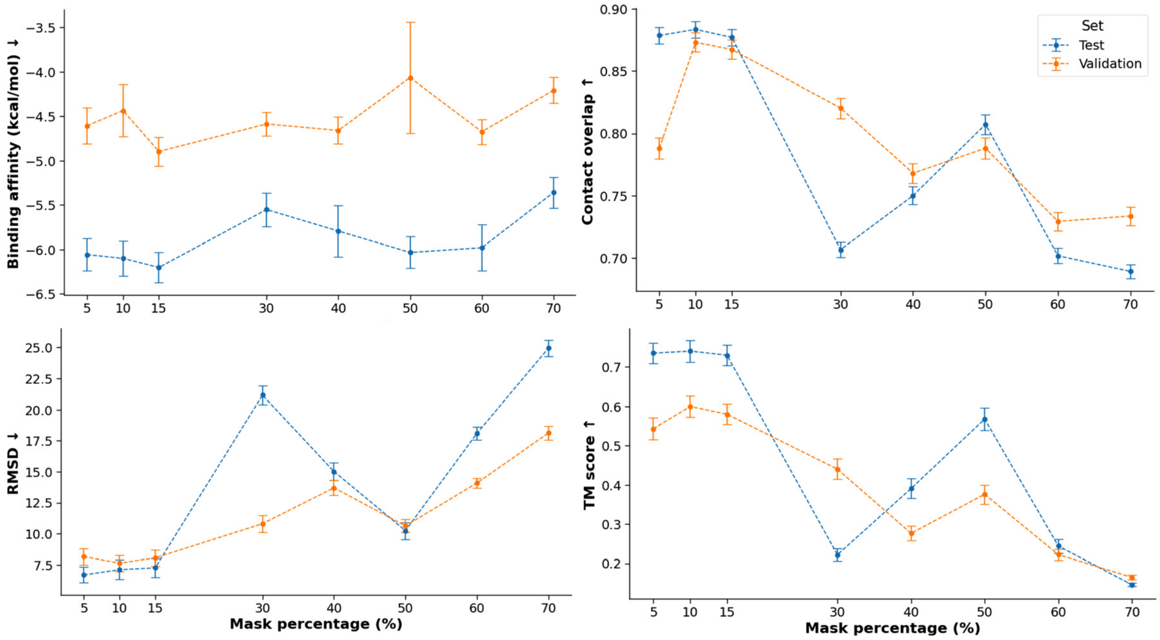


FIG. 9. Mask ablation studies on both validation and test sets. Each of the mask ratios (5, 10, 15, 30, 40, 50, 60, and 70%) is a hyperparameter and represented by a model. The performances of the masked models are evaluated for all metrics. The arrows on y-axes show directions of better performance.

manual feature engineering for ligands and the use of ESM-2 as a pre-trained LLM (Large Language Model) for protein featurization.

We gradually reduced ligand features, starting with ligand distance and bond information (e.g., types and ring), and even omitted the entire bond and ligand. In Fig. 8, omitting bond features and distance caused less reduction in model performance than omitting the entire ligand. Ligand bond information is crucial for the model to learn the relative positions of ligand atoms and adhere to geometric constraints within the triangular update module (Sec. IV B 3).

We observed a significant decrease in model performance when ESM embeddings were excluded (the red bar in Fig. 8). The ESM features alone (the brown bar) significantly boosted performance when training without ligand data, as these embeddings are enriched with protein evolutionary and biophysical information needed for both single and pair representations. Other protein features, such as position encodings and amino acid types, provided slight improvements, though they were minimal. However, excluding ligand information led to a reduction in model performance compared to the baseline, as the model relies on learning the overall structure of the complexes.

Therefore, using pre-trained featurization methods, such as ESM and other protein BERT-like models, in combination with ligand input, significantly enhances model training and performance.

c. Impact of masking ratios. We examined ProteinReDiff's performance with various percentages of masked amino acids, adjusting the masking ratio as a hyperparameter and retraining our model. In Fig. 9, we observed consistent top performance across the metrics with masking ratios between 5% and 15%. This range is crucial for the protein redesign strategy, enhancing binding affinity while preserving the structural and functional motifs of the target protein. The 15% masking ratio achieved the best ligand binding affinity, the most important metric for capturing protein function.

Interestingly, we noticed performance spikes for 50% masking in contact overlap and TM-score. This is because applying stochastic masks allows the model to learn representations with varied masking from 0 up to the set ratio. Although the 50% masking does not surpass the 15% masking's performance, the improvement in the high masking regime demonstrates the robustness of our training scheme.

Overall, this investigation highlights the optimal level of sequence masking needed to enhance ligand binding affinity, sequence diversity,

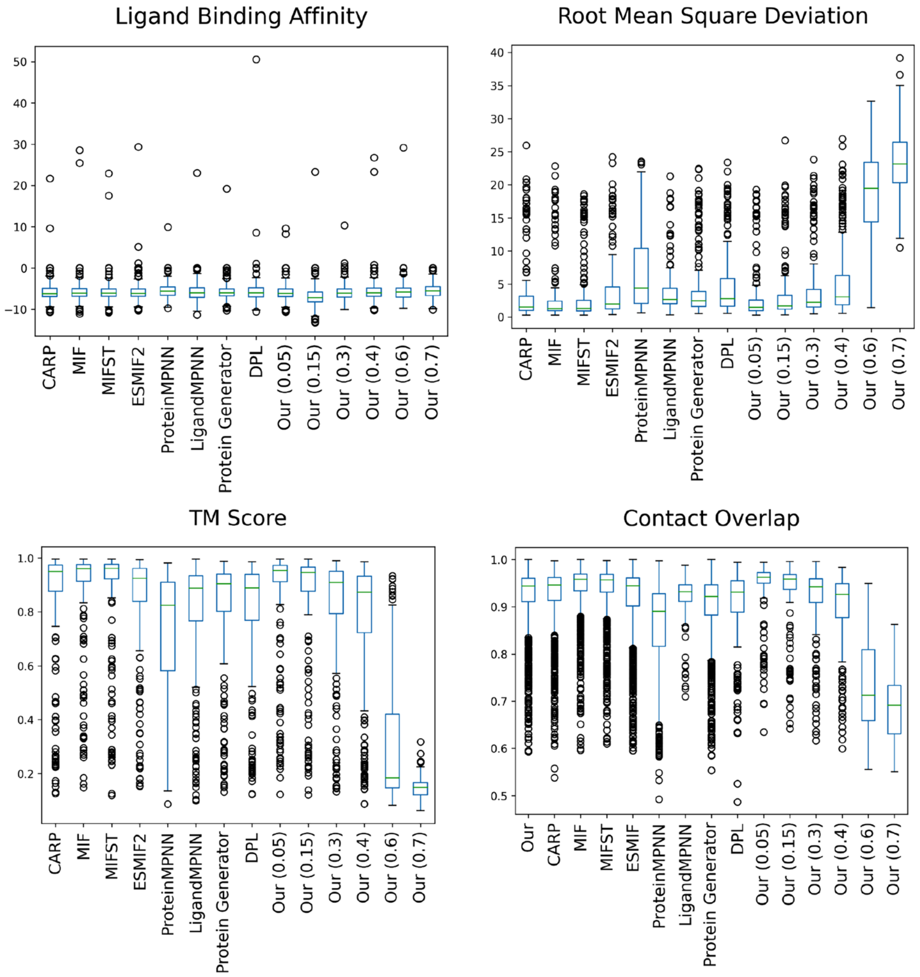


FIG. 10. Boxplot illustrating the distribution of ligand binding affinities, and structure preservation metrics (TM Score and RMSD) across all methods evaluated, including baseline models and variations of ProteinReDiff. Each boxplot showcases the median, quartiles, and outliers within the data, providing insight into the variability and central tendency of each metric across the dataset's samples. Detailed descriptions of each baseline are provided in Sec. V B 4.

and structural preservation. It also reinforces training strategies for protein redesign as shown on the Discussion Sec. V B 5.

VI. CONCLUSIONS

This study introduces ProteinReDiff, a computational framework developed to redesign ligand-binding proteins. By utilizing advanced techniques inspired by Equivariant Diffusion-Based Generative Models and the attention mechanism from AlphaFold2, ProteinReDiff demonstrates its ability to enhance complex protein–ligand interactions. Our model excels in optimizing ligand binding affinity based solely on initial protein sequences and ligand SMILES strings, bypassing the need for detailed structural data. Experimental validations highlight ProteinReDiff's capability to improve ligand binding affinity while preserving essential sequence diversity and structural integrity. These findings open new possibilities for protein–ligand complex modeling, indicating significant potential for ProteinReDiff in various biotechnological and pharmaceutical applications.

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to disclose.

Author Contributions

Viet Thanh Duy Nguyen: Data curation (equal); Formal analysis (equal); Methodology (equal); Software (equal); Validation (equal); Visualization (equal); Writing – original draft (equal); Writing – review & editing (equal). **Nhan D. Nguyen:** Conceptualization (equal); Data curation (equal); Formal analysis (equal); Methodology (equal); Software (equal); Validation (equal); Visualization (equal); Writing – original draft (equal); Writing – review & editing (equal). **Truong Son Hy:** Conceptualization (lead); Funding acquisition (lead); Investigation (lead); Methodology (lead); Project administration (lead); Resources (lead); Software (equal); Supervision (lead); Writing – original draft (equal); Writing – review & editing (equal).

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author upon reasonable request. The data that support the findings of this study are openly available in Ref. 106.

APPENDIX A: BENCHMARKING PROTEINREDIFF AGAINST RELATED MODELS

These plots, shown in Fig. 10, demonstrate the comparative performance of ProteinReDiff against other relevant models. The results indicate that our model consistently ranks among the high performers.

APPENDIX B: EVALUATING PROTEIN–LIGAND COMPLEX REPRESENTATION

In the continuation of our study's exploration of protein–ligand complex representations, we extended the use of the PDBBind v2020 dataset,[53] previously detailed in our training process, to specifically evaluate the effectiveness of the Input Featurizer from ProteinReDiff. By using embeddings generated by the Input Featurizer as input features, we trained a Gaussian Process (GP) model to predict ligand binding affinity. The choice of a GP model, recognized for its probabilistic nature and adaptability to the nuanced, uncertain dynamics of biological interactions, was pivotal in assessing how well the embeddings capture predictive information about protein–ligand interactions. The GP model employed a Gaussian likelihood, suitable for regression tasks, along with a radial basis function (RBF) kernel, chosen for its effectiveness in modeling smooth, continuous variations characteristic of binding affinities. The GP model's parameters were optimized to ensure a robust fit to the training data.

TABLE V. Experimental results for the ligand binding affinity prediction task on the PDBBind v2020 dataset. Results for comparative reference models are sourced from Ref. 80.

Approach	RMSE ↓ ($-\log K_d/K_i$)	MAE ↓ ($-\log K_d/K_i$)	Pearson ↑	Spearman ↑
Pafnucy[92]	1.435	1.144	0.635	0.587
OnionNet[93]	1.403	1.103	0.648	0.602
IGN[94]	1.404	1.116	0.662	0.638
SIGN[95]	1.373	1.086	0.685	0.656
SMINA[96]	1.466	1.161	0.665	0.663
GNINA[97]	1.740	1.413	0.495	0.494
dMaSIF[98]	1.450	1.136	0.629	0.588
TankBind[99]	1.345	1.060	0.718	**0.689**
GraphDTA[100]	1.564	1.223	0.612	0.570
TransCPI[101]	1.493	1.201	0.604	0.551
MolTrans[102]	1.599	1.271	0.539	0.474
DrugBAN[103]	1.480	1.159	0.657	0.612
DGraphDTA[104]	1.493	1.201	0.604	0.551
WGNN-DTA[105]	1.501	1.196	0.605	0.562
STAMP-DPI[104]	1.503	1.176	0.653	0.601
PSICHIC[80]	**1.314**	**1.015**	0.710	0.686
ProteinReDiff (our)	1.443	1.168	**0.721**	0.639

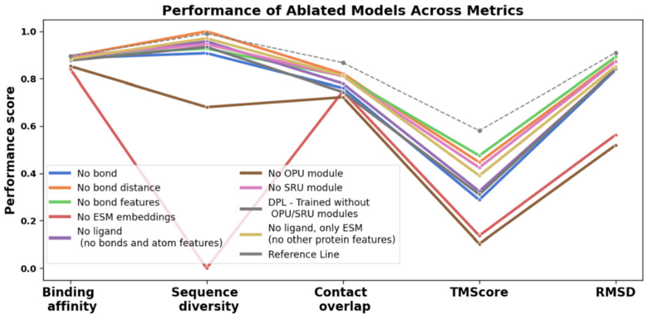


FIG. 11. Breakdown of metrics for ablation models based on different featurization methods and architectural adjustments. The dashed line indicates the baseline ProteinReDiff model trained without any ablations.

TABLE VI. Ablation study results on mask ratios. The table shows the impact of different mask ratios on validation and test set performance metrics.

	Valid					Test				
Mask ratio	LBA ↓	Sequence diversity ↑	TM-score ↑	RMSD ↓	CO ↑	LBA ↓	Sequence diversity ↑	TM-score ↑	RMSD ↓	CO ↑
5%	−4.602 ± 0.377	87.252	0.555 ± 0.023	8.225 ± 0.510	0.788 ± 0.008	−6.058 ± 0.182	180.800	0.734 ± 0.025	**6.685** ± 0.629	0.879 ± 0.010
10%	−4.410 ± 0.541	89.472	**0.598** ± 0.022	**7.808** ± 0.544	**0.873** ± 0.008	−6.101 ± 0.194	184.564	**0.739** ± 0.027	7.108 ± 0.784	**0.883** ± 0.010
15%	−**4.890** ± 0.303	89.601	0.581 ± 0.022	8.252 ± 0.537	0.867 ± 0.008	−**6.202** ± 0.167	184.925	0.729 ± 0.025	7.257 ± 0.768	0.877 ± 0.010
30%	−4.596 ± 0.257	**90.643**	0.453 ± 0.022	10.707 ± 0.604	0.820 ± 0.008	−5.553 ± 0.188	181.978	0.221 ± 0.015	21.166 ± 0.740	0.707 ± 0.009
40%	−4.668 ± 0.281	89.091	0.297 ± 0.016	14.309 ± 0.497	0.768 ± 0.008	−5.794 ± 0.286	185.136	0.390 ± 0.024	15.014 ± 0.717	0.750 ± 0.011
50%	−4.052 ± 1.162	90.445	0.390 ± 0.020	10.886 ± 0.424	0.788 ± 0.009	−6.034 ± 0.177	**188.163**	0.567 ± 0.029	10.239 ± 0.688	0.807 ± 0.012
60%	−4.678 ± 0.262	88.643	0.226 ± 0.011	14.142 ± 0.337	0.729 ± 0.007	−5.981 ± 0.258	184.356	0.243 ± 0.017	18.092 ± 0.525	0.702 ± 0.009
70%	−4.214 ± 0.264	81.333	0.165 ± 0.004	18.226 ± 0.456	0.733 ± 0.007	−5.360 ± 0.175	162.841	0.145 ± 0.004	24.944 ± 0.646	0.689 ± 0.008

TABLE VII. Ablation setup of featurization and model architecture.

		Ablation studies								
		No bond distance	No bond feats	No bond	No ligand	No ligand, only ESM	No ESM	No SRA	No OPU	DPL (No SRA/OPU)
	Bond distance		✓				✓	✓	✓	✓
Ligand	Bond feats (type, ring, etc.)	✓					✓	✓	✓	✓
	Ligand atom feats (chirality, charge, degree, etc.)	✓	✓	✓			✓	✓	✓	✓
Protein	ESM embeddings	✓	✓	✓	✓	✓		✓	✓	✓
	Residue feats (pos. encodings, res. type)	✓	✓	✓	✓		✓	✓	✓	✓
Model architecture	Single representation attention (SRA)	✓	✓	✓	✓	✓	✓		✓	
	Outer product update (OPU)	✓	✓	✓	✓	✓	✓	✓		

TABLE VIII. Ablation study results on input featurization methods. The table presents the impact of various feature removals on performance metrics.

Features	LBA ↓	Sequence diversity ↑	TM-score ↑	RMSD ↓	CO ↑
Reference	**−4.890** ± 0.303	89.601	**0.581** ± 0.022	**8.252** ± 0.537	**0.877** ± 0.008
No bond	−4.549 ± 0.272	84.837	0.287 ± 0.016	14.325 ± 0.491	0.761 ± 0.009
No bond distance	−4.869 ± 0.277	**90.186**	0.447 ± 0.021	11.068 ± 0.579	0.821 ± 0.008
No bond feats	−4.985 ± 0.289	85.974	0.475 ± 0.022	9.748 ± 0.476	0.811 ± 0.010
No ESM	−2.723 ± 0.176	32.222	0.136 ± 0.007	37.322 ± 1.032	0.748 ± 0.008
No ligand	−4.478 ± 0.252	87.723	0.324 ± 0.018	14.125 ± 0.571	0.780 ± 0.008
No OPU	−3.197 ± 0.304	71.669	0.102 ± 0.006	40.969 ± 1.246	0.723 ± 0.004
No SRA	−4.878 ± 0.282	87.054	0.424 ± 0.023	11.391 ± 0.556	0.810 ± 0.009
DPL	−4.153 ± 0.631	86.379	0.311 ± 0.019	13.931 ± 0.527	0.744 ± 0.009
No ligand, only ESM	−4.429 ± 0.270	88.481	0.390 ± 0.020	13.108 ± 0.702	0.813 ± 0.007

The evaluation results in Table V demonstrate the performance of embeddings generated by the Input Featurizer on the PDBBind v2020 dataset compared to baseline methods. Notably, these embeddings achieved the highest Pearson correlation (0.721) for predicting ligand binding affinity, highlighting the Input Featurizer's effectiveness in capturing meaningful protein–ligand interactions. This strong performance is further supported by competitive RMSE, MAE, and Spearman correlation metrics.

APPENDIX C: ABLATION STUDIES

Here, we present additional results from the mask and feature ablation studies. Figure 11 illustrates the performance of ablated models across five key metrics. The impact of different mask ratios on validation and test set metrics is summarized in Table VI. For each model, Table VII specifies the features included or excluded, while Table VIII highlights the resulting effects of these feature ablations on performance.

REFERENCES

[1] X. Du, Y. Li, Y.-L. Xia, S.-M. Ai, J. Liang, P. Sang, X.-L. Ji, and S.-Q. Liu, “Insights into protein-ligand interactions: Mechanisms, models, and methods,” Int. J. Mol. Sci. **17**, 144 (2016).

[2] K. Wanat, “Biological barriers, and the influence of protein binding on the passage of drugs across them,” Mol. Biol. Rep. **47**, 3221–3231 (2020).

[3] J. Skolnick and H. Zhou, “Implications of the essential role of small molecule ligand binding pockets in protein–protein interactions,” J. Phys. Chem. B **126**, 6853–6867 (2022).

[4] D. Listov, C. A. Goverde, B. E. Correia, and S. J. Fleishman, “Opportunities and challenges in design and optimization of protein function,” Nat. Rev. Mol. Cell Biol. **25**, 639 (2024).

[5] S. L. Lisanza, J. M. Gershon, S. Tipps, L. Arnoldt, S. Hendel, J. N. Sims, X. Li, and D. Baker, “Joint generation of protein sequence and structure with rosettafold sequence space diffusion,” bioRxiv (2023).

[6] J. Dauparas, I. Anishchenko, N. Bennett, H. Bai, R. J. Ragotte, L. F. Milles, B. I. M. Wicky, A. Courbet, R. J. de Haas, N. Bethel, P. J. Y. Leung, T. F. Huddy, S. Pellock, D. Tischer, F. Chan, B. Koepnick, H. Nguyen, A. Kang, B. Sankaran, A. K. Bera, N. P. King, and D. Baker, “Robust deep learning–based protein sequence design using ProteinMPNN,” Science **378**, 49–56 (2022).

[7] K. K. Yang, N. Fusi, and A. X. Lu, “Convolutions are competitive with transformers for protein sequence pretraining,” bioRxiv (2023).

[8] S. Iqbal, F. Ge, F. Li, T. Akutsu, Y. Zheng, R. B. Gasser, D.-J. Yu, G. I. Webb, and J. Song, “Prost: AlphaFold2-aware sequence-based predictor to estimate protein stability changes upon missense mutations,” J. Chem. Inf. Model. **62**, 4270–4282 (2022).

[9] W. Yang and L. Lai, “Computational design of ligand-binding proteins,” Curr. Opin. Struct. Biol. **45**, 67–73 (2017).

[10] S. B. Ebrahimi and D. Samanta, “Engineering protein-based therapeutics through structural and chemical design,” Nat. Commun. **14**, 2411 (2023).

[11] A. Ruscito and M. C. DeRosa, “Small-molecule binding aptamers: Selection strategies, characterization, and applications,” Front. Chem. **4**, 14 (2016).

[12] R. Creutznacher, T. Maass, B. Veselkova, G. Ssebyatika, T. Krey, M. Empting, N. Tautz, M. Frank, K. Kölbel, C. Uetrecht, and T. Peters, “NMR experiments provide insights into ligand-binding to the SARS-COV-2 spike protein receptor-binding domain,” J. Am. Chem. Soc. **144**, 13060–13065 (2022).

[13] C. Munk, K. Harpsøe, A. S. Hauser, V. Isberg, and D. E. Gloriam, “Integrating structural and mutagenesis data to elucidate GPCR ligand binding,” Curr. Opin. Pharmacol. **30**, 51–58 (2016).

[14] D. Tavares and J. R. van der Meer, “Ribose-binding protein mutants with improved interaction towards the non-natural ligand 1,3-cyclohexanediol,” Front. Bioeng. Biotechnol. **9**, 705534 (2021).

[15] N. F. Polizzi and W. F. DeGrado, “A defined structural unit enables de novo design of small-molecule–binding proteins,” Science **369**, 1227–1233 (2020).

[16] H. Stärk, B. Jing, R. Barzilay, and T. Jaakkola, “Harmonic self-conditioned flow matching for multi-ligand docking and binding site design,” arXiv:2310.05764 (2023).

[17] J. Dauparas, G. R. Lee, R. Pecoraro, L. An, I. Anishchenko, C. Glasscock, and D. Baker, Atomic context-conditioned protein sequence design using LigandMPNN,” bioRxiv (2023).

[18] M. Lv, X. Luo, J. Estill, Y. Liu, M. Ren, J. Wang, Q. Wang, S. Zhao, X. Wang, S. Yang, X. Feng, W. Li, E. Liu, X. Zhang, L. Wang, Q. Zhou, W. Meng, X. Qi, Y. Xun, X. Yu, Y. Chen, and COVID-19 evidence and recommendations working Group, “Coronavirus disease (COVID-19): A scoping review,” Eurosurveillance **25**, 2000125 (2020).

[19] J. M. Schaub, C.-W. Chou, H.-C. Kuo, K. Javanmardi, C.-L. Hsieh, J. Goldsmith, A. M. DiVenere, K. C. Le, D. Wrapp, P. O. Byrne *et al.*, “Expression and characterization of SARS-CoV-2 spike proteins,” Nat. Protoc. **16**, 5339–5356 (2021).

[20] S. Agajanian, M. Alshahrani, F. Bai, P. Tao, and G. M. Verkhivker, “Exploring and learning the universe of protein allostery using artificial intelligence augmented biophysical and computational approaches,” J. Chem. Inf. Model. **63**, 1413–1428 (2023).

[21] V. Oleinikovas, G. Saladino, B. P. Cossins, and F. L. Gervasio, “Understanding cryptic pocket formation in protein targets by enhanced sampling simulations,” J. Am. Chem. Soc. **138**, 14257–14263 (2016).

[22] A. Meller, M. Ward, J. Borowsky, M. Kshirsagar, J. M. Lotthammer, F. Oviedo, J. L. Ferres, and G. R. Bowman, “Predicting locations of cryptic pockets from single protein structures using the PocketMiner graph neural network,” Nat. Commun. **14**, 1177 (2023).

[23] E. P. Barros, J. M. Schiffer, A. Vorobieva, J. Dou, D. Baker, and R. E. Amaro, "Improving the efficiency of ligand-binding protein design with molecular dynamics simulations," J. Chem. Theory Comput. **15**, 5703–5715 (2019).

[24] S. Nakata, Y. Mori, and S. Tanaka, "End-to-end protein–ligand complex structure generation with diffusion-based generative models," BMC Bioinf. **24**, 233 (2023).

[25] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, and D. Hassabis, "Highly accurate protein structure prediction with AlphaFold," Nature **596**, 583–589 (2021).

[26] D. Weininger, "Smiles, a chemical language and information system. 1. Introduction to methodology and encoding rules," J. Chem. Inf. Comput. Sci. **28**, 31–36 (1988).

[27] I. V. Korendovych, "Rational and semirational protein design," *Protein Engineering: Methods and Protocols* (Springer, 2018), pp. 15–23.

[28] Z. Song, Q. Zhang, W. Wu, Z. Pu, and H. Yu, "Rational design of enzyme activity and enantioselectivity," Front. Bioeng. Biotechnol. **11**, 1129149 (2023).

[29] E. C. Alley, G. Khimulya, S. Biswas, M. AlQuraishi, and G. M. Church, "Unified rational protein engineering with sequence-based deep representation learning," Nat. Methods **16**, 1315–1322 (2019).

[30] F. H. Arnold and A. A. Volkov, "Directed evolution of biocatalysts," Curr. Opin. Chem. Biol. **3**, 54–59 (1999).

[31] M. Wang and H. Zhao, "Combined and iterative use of computational design and directed evolution for protein–ligand binding design," Methods Mol. Biol. **1414**, 139–153 (2016).

[32] G. Guntas, T. J. Mansell, J. R. Kim, and M. Ostermeier, "Directed evolution of protein switches and their application to the creation of ligand-binding proteins," Proc. Natl. Acad. Sci. U. S. A. **102**, 11224–11229 (2005).

[33] Y. Waltenspühl, J. R. Jeliazkov, L. Kummer, and A. Plückthun, "Directed evolution for high functional production and stability of a challenging g protein-coupled receptor," Sci. Rep. **11**, 8630 (2021).

[34] G. Raut and A. Singh, "Generative AI in vision: A survey on models, metrics and applications," arXiv:2402.16369 (2024).

[35] T. Iqbal and S. Qureshi, "The survey: Text generation models in deep learning," J. King Saud Univ. - Comput. Inf. Sci. **34**, 2515–2528 (2022).

[36] S. Lyu, S. Sowlati-Hashjin, and M. Garton, "Proteinvae: Variational autoencoder for translational protein design," bioRxiv (2023).

[37] J. G. Greener, L. Moffat, and D. T. Jones, "Design of metalloproteins and novel protein folds using variational autoencoders," Sci. Rep. **8**, 16189 (2018).

[38] D. Brookes, H. Park, and J. Listgarten, "Conditioning by adaptive sampling for robust design," in *Proceedings of the 36th International Conference on Machine Learning* (PMLR, 2019), Vol. 97, pp. 773–782.

[39] J. Trinquier, G. Uguzzoni, A. Pagnani, F. Zamponi, and M. Weigt, "Efficient generative modeling of protein sequences using simple autoregressive models," Nat. Commun. **12**, 5800 (2021).

[40] C. Fannjiang, S. Bates, A. N. Angelopoulos, J. Listgarten, and M. I. Jordan, "Conformal prediction under feedback covariate shift for biomolecular design," Proc. Natl. Acad. Sci. U. S. A. **119**, e2204569119 (2022).

[41] T. Kucera, M. Togninalli, and L. Meng-Papaxanthos, "Conditional generative modeling for de novo protein design with hierarchical functions," Bioinformatics **38**, 3454–3461 (2022).

[42] N. Anand and P. Huang, "Generative modeling for protein structures," in *Advances in Neural Information Processing Systems*, edited by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Curran Associates, Inc., 2018), Vol. 31.

[43] N. Gruver, S. Stanton, N. C. Frey, T. G. J. Rudner, I. Hotzel, J. Lafrance-Vanasse, A. Rajpal, K. Cho, and A. G. Wilson, "Protein design with guided discrete diffusion," arXiv:2305.20009 (2023).

[44] J. L. Watson, D. Juergens, N. R. Bennett, B. L. Trippe, J. Yim, H. E. Eisenach, W. Ahern, A. J. Borst, R. J. Ragotte, L. F. Milles, B. I. M. Wicky, N. Hanikel, S. J. Pellock, A. Courbet, W. Sheffler, J. Wang, P. Venkatesh, I. Sappington, S. V. Torres, A. Lauko, V. De Bortoli, E. Mathieu, S. Ovchinnikov, R. Barzilay, T. S. Jaakkola, F. DiMaio, M. Baek, and D. Baker, "De novo design of protein structure and function with rfdiffusion," Nature **620**, 1089–1100 (2023).

[45] K. E. Wu, K. K. Yang, R. van den Berg, J. Y. Zou, A. X. Lu, and A. P. Amini, "Protein structure generation via folding diffusion," arXiv:2209.15611 (2022).

[46] C. Fu, K. Yan, L. Wang, W. Y. Au, M. McThrow, T. Komikado, K. Maruhashi, K. Uchino, X. Qian, and S. Ji, "A latent diffusion model for protein structure generation," arXiv:2305.04120 (2023).

[47] Z. Zheng, Y. Deng, D. Xue, Y. Zhou, F. Ye, and Q. Gu, "Structure-informed language models are protein designers," in *Proceedings of the 40th International Conference on Machine Learning, ICML'23* (JMLR.org, 2023).

[48] K. K. Yang, N. Zanichelli, and H. Yeh, "Masked inverse folding with sequence transfer for protein representation learning," Protein Eng., Des. Sel. **36**, gzad015 (2023).

[49] C. Hsu, R. Verkuil, J. Liu, Z. Lin, B. Hie, T. Sercu, A. Lerer, and A. Rives, "Learning inverse folding from millions of predicted structures," in *International Conference on Machine Learning* (PMLR, 2022).

[50] N. R. Bennett, J. L. Watson, R. J. Ragotte, A. J. Borst, D. L. See, C. Weidle, R. Biswas, E. L. Shrock, P. J. Y. Leung, B. Huang, I. Goreshnik, R. Ault, K. D. Carr, B. Singer, C. Criswell, D. Vafeados, M. G. Sanchez, H. M. Kim, S. V. Torres, S. Chan, and D. Baker, "Atomically accurate de novo design of single-domain antibodies," bioRxiv (2024).

[51] M. F. Chungyoun and J. J. Gray, "AI models for protein design are driving antibody engineering," Curr. Opin. Biomed. Eng. **28**, 100473 (2023).

[52] L. Gagliardi and W. Rocchia, "SiteFerret: Beyond simple pocket identification in proteins," J. Chem. Theory Comput. **19**, 5242–5259 (2023).

[53] R. Wang, X. Fang, Y. Lu, and S. Wang, "The PDBbind database: Collection of binding affinities for protein-ligand complexes with known three-dimensional structures," J. Med. Chem. **47**, 2977–2980 (2004).

[54] I. Sillitoe, N. Dawson, T. E. Lewis, S. Das, J. G. Lees, P. Ashford, A. Tolulope, H. M. Scholes, I. Senatorov, A. Bujan, F. Ceballos Rodriguez-Conde, B. Dowling, J. Thornton, and C. A. Orengo, "CATH: Expanding the horizons of structure-based functional annotations for genome sequences," Nucl. Acids Res. **47**, D280–D284 (2019).

[55] Z. Yang, X. Zeng, Y. Zhao, and R. Chen, "AlphaFold2 and its applications in the fields of biology and medicine," Signal Transduction Targeted Ther. **8**, 115 (2023).

[56] N. Brandes, D. Ofer, Y. Peleg, N. Rappoport, and M. Linial, "ProteinBERT: A universal deep-learning model of protein sequence and function," Bioinformatics **38**, 2102–2110 (2022).

[57] A. Elnaggar, M. Heinzinger, C. Dallago, G. Rehawi, Y. Wang, L. Jones, T. Gibbs, T. Feher, C. Angerer, M. Steinegger, D. Bhowmik, and B. Rost, "Prottrans: Toward understanding the language of life through self-supervised learning," IEEE Trans. Pattern Anal. Mach. Intell. **44**, 7112–7127 (2022).

[58] A. Rives, J. Meier, T. Sercu, S. Goyal, Z. Lin, J. Liu, D. Guo, M. Ott, C. L. Zitnick, J. Ma, and R. Fergus, "Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences," Proc. Natl. Acad. Sci. U. S. A. **118**, e2016239118 (2021).

[59] Z. Lin, H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, N. Smetanin, R. Verkuil, O. Kabeli, Y. Shmueli, A. dos Santos Costa, M. Fazel-Zarandi, T. Sercu, S. Candido, and A. Rives, "Evolutionary-scale prediction of atomic-level protein structure with a language model," Science **379**, 1123–1130 (2023).

[60] A. Madani, B. Krause, E. R. Greene, S. Subramanian, B. P. Mohr, J. M. Holton, J. L. Olmos, C. Xiong, Z. Z. Sun, R. Socher, J. S. Fraser, and N. Naik, "Large language models generate functional protein sequences across diverse families," Nat. Biotechnol. **41**, 1099–1106 (2023).

[61] J. A. Ruffolo and A. Madani, "Designing proteins with language models," Nat. Biotechnol. **42**, 200–202 (2024).

[62] X. Min, C. Yang, J. Xie, Y. Huang, N. Liu, X. Jin, T. Wang, Z. Kong, X. Lu, S. Ge, J. Zhang, and N. Xia, "Tpgen: A language model for stable protein design with a specific topology structure," BMC Bioinf. **25**, 35 (2024).

[63] B. E. Suzek, Y. Wang, H. Huang, P. B. McGarvey, C. H. Wu, and the UniProt Consortium, "UniRef clusters: A comprehensive and scalable alternative for improving sequence similarity searches," Bioinformatics **31**, 926–932 (2015).

[64] D. P. Kingma, T. Salimans, B. Poole, and J. Ho, "Variational diffusion models," arXiv:2107.00630 (2023).

[65] E. Hoogeboom, V. G. Satorras, C. Vignac, and M. Welling, "Equivariant diffusion for molecule generation in 3d," arXiv:2203.17003 (2022).

[66] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," arXiv:2011.13456 (2020).
[67] T. Xu, Q. Xu, and J. Li, "Toward the appropriate interpretation of alphafold2," Front. Artif. Intell. **6**, 1149748 (2023).
[68] R. M. Rao, J. Liu, R. Verkuil, J. Meier, J. Canny, P. Abbeel, T. Sercu, and A. Rives, "MSA transformer," in *Proceedings of the 38th International Conference on Machine Learning* (PMLR, 2021), Vol. 139, pp. 8844–8856.
[69] N. Buton, F. Coste, and Y. Le Cunff, "Predicting enzymatic function of protein sequences with attention," Bioinformatics **39**, btad620 (2023).
[70] F. Ju, J. Zhu, B. Shao, L. Kong, T.-Y. Liu, W.-M. Zheng, and D. Bu, "CopulaNet: Learning residue co-evolution directly from multiple sequence alignment for protein structure prediction," Nat. Commun. **12**, 2535 (2021).
[71] Q. Huang, P. Smolensky, X. He, L. Deng, and D. Wu, "Tensor product generation networks for deep NLP modeling," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)* (Association for Computational Linguistics, New Orleans, Louisiana, 2018), pp. 1263–1273.
[72] P. Smolensky, "Tensor product variable binding and the representation of symbolic structures in connectionist systems," Artif. Intell. **46**, 159–216 (1990).
[73] Q. Huang, L. Deng, D. Wu, C. Liu, and X. He, "Attentive tensor product learning," AAAI **33**, 1344–1351 (2019).
[74] I. Schlag and J. Schmidhuber, "Learning to reason with third-order tensor products," arXiv:1811.12143 (2018).
[75] C. Chen, Q. Lu, A. Beukers, C. Baldassano, and K. A. Norman, "Learning to perform role-filler binding with schematic knowledge," PeerJ **9**, e11046 (2021).
[76] Y. Lin and M. AlQuraishi, "Generating novel, designable, and diverse protein structures by equivariantly diffusing oriented residue clouds," in *Proceedings of the 40th International Conference on Machine Learning, ICML'23* (JMLR.org, 2023).
[77] M. Steinegger and J. Söding, "MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets," Nat. Biotechnol. **35**, 1026–1028 (2017).
[78] H. L. Morgan, "The generation of a unique machine description for chemical structures-a technique developed at chemical abstracts service," J. Chem. Doc. **5**, 107–113 (1965).
[79] J. Ingraham, V. Garg, R. Barzilay, and T. Jaakkola, "Generative models for graph-based protein design," in *Advances in Neural Information Processing Systems*, edited by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Curran Associates, Inc., 2019), Vol. 32.
[80] H. Y. Koh, A. T. Nguyen, S. Pan, L. T. May, and G. I. Webb, "Psichic: Physicochemical graph neural network for learning protein-ligand interaction fingerprints from sequence data," bioRxiv (2023).
[81] J. M. Joyce, "Kullback-leibler divergence," in *International Encyclopedia of Statistical Science*, edited by M. Lovric (Springer Berlin Heidelberg, Berlin, Heidelberg, 2011), pp. 720–722.
[82] R. Sawada, Y. Sakajiri, T. Shibata, and Y. Yamanishi, "Predicting therapeutic and side effects from drug binding affinities to human proteome structures," iScience **27**, 110032 (2024).
[83] C. Ziegler, J. Martin, C. Sinner, and F. Morcos, "Latent generative landscapes as maps of functional diversity in protein sequence space," Nat. Commun. **14**, 2222 (2023).
[84] F. P. Miller, A. F. Vandome, and J. McBrewster, *Levenshtein Distance: Information Theory, Computer Science, String (Computer Science), String Metric, Damerau?Levenshtein Distance, Spell Checker, Hamming Distance* (Alpha Press, 2009).
[85] Y. Zhang and J. Skolnick, "Scoring function for automated assessment of protein structure template quality," Proteins **57**, 702–710 (2004).
[86] R. Laskowski and T. de Beer, "Root mean square deviation (RMSD)," in *Dictionary of Bioinformatics and Computational Biology* (John Wiley and Sons, Ltd, 2014).
[87] U. Bastolla, D. Abia, and O. Piette, "PC_ali: A tool for improved multiple alignments and evolutionary inference based on a hybrid protein sequence and structure similarity score," Bioinformatics **39**, btad630 (2023).
[88] L. Cheng, P. Liu, D. Wang, and K.-S. Leung, "Exploiting locational and topological overlap model to identify modules in protein interaction networks," BMC Bioinf. **20**, 23 (2019).
[89] M. Iyer, Z. Li, L. Jaroszewski, M. Sedova, and A. Godzik, "Difference contact maps: From what to why in the analysis of the conformational flexibility of proteins," PLoS One **15**, e0226702 (2020).
[90] R. Wu, F. Ding, R. Wang, R. Shen, X. Zhang, S. Luo, C. Su, Z. Wu, Q. Xie, B. Berger, J. Ma, and J. Peng, "High-resolution de novo structure prediction from primary sequence," bioRxiv (2022).
[91] O. Trott and A. J. Olson, "AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading," J. Comput. Chem. **31**, 455–461 (2010).
[92] M. M. Stepniewska-Dziubinska, P. Zielenkiewicz, and P. Siedlecki, "Development and evaluation of a deep learning model for protein–ligand binding affinity prediction," Bioinformatics **34**, 3666–3674 (2018).
[93] L. Zheng, J. Fan, and Y. Mu, "Onionnet: A multiple-layer intermolecular-contact-based convolutional neural network for protein–ligand binding affinity prediction," ACS Omega **4**, 15956–15965 (2019).
[94] D. Jiang, C.-Y. Hsieh, Z. Wu, Y. Kang, J. Wang, E. Wang, B. Liao, C. Shen, L. Xu, J. Wu, D. Cao, and T. Hou, "Interactiongraphnet: A novel and efficient deep graph representation learning framework for accurate protein–ligand interaction predictions," J. Med. Chem. **64**, 18209–18232 (2021).
[95] S. Li, J. Zhou, T. Xu, L. Huang, F. Wang, H. Xiong, W. Huang, D. Dou, and H. Xiong, "Structure-aware interactive graph neural networks for the prediction of protein-ligand binding affinity," in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, KDD '21* (Association for Computing Machinery, New York, NY, 2021), pp. 975–985.
[96] D. R. Koes, M. P. Baumgartner, and C. J. Camacho, "Lessons learned in empirical scoring with smina from the CSAR 2011 benchmarking exercise," J. Chem. Inf. Model. **53**, 1893–1904 (2013).
[97] A. T. McNutt, P. Francoeur, R. Aggarwal, T. Masuda, R. Meli, M. Ragoza, J. Sunseri, and D. R. Koes, "Gnina 1.0: Molecular docking with deep learning," J. Cheminf. **13**, 43 (2021).
[98] F. Sverrisson, J. Feydy, B. E. Correia, and M. M. Bronstein, "Fast end-to-end learning on protein surfaces," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, 2021), pp. 15267–15276.
[99] W. Lu, Q. Wu, J. Zhang, J. Rao, C. Li, and S. Zheng, "Tankbind: Trigonometry-aware neural networks for drug-protein binding structure prediction," bioRxiv (2022).
[100] T. Nguyen, H. Le, T. P. Quinn, T. Nguyen, T. D. Le, and S. Venkatesh, "GraphDTA: Predicting drug–target binding affinity with graph neural networks," Bioinformatics **37**, 1140–1147 (2021).
[101] L. Chen, X. Tan, D. Wang, F. Zhong, X. Liu, T. Yang, X. Luo, K. Chen, H. Jiang, and M. Zheng, "TransformerCPI: Improving compound–protein interaction prediction by sequence-based deep learning with self-attention mechanism and label reversal experiments," Bioinformatics **36**, 4406–4414 (2020).
[102] K. Huang, C. Xiao, L. M. Glass, and J. Sun, "MolTrans: Molecular interaction transformer for drug–target interaction prediction," Bioinformatics **37**, 830–836 (2021).
[103] P. Bai, F. Miljković, B. John, and H. Lu, "Interpretable bilinear attention network with domain adaptation improves drug–target prediction," Nat. Mach. Intell. **5**, 126–136 (2023).
[104] M. Jiang, Z. Li, S. Zhang, S. Wang, X. Wang, Q. Yuan, and Z. Wei, "Drug–target affinity prediction using graph neural network and contact maps," RSC Adv. **10**, 20701–20712 (2020).
[105] P. Wang, S. Zheng, Y. Jiang, C. Li, J. Liu, C. Wen, A. Patronov, D. Qian, H. Chen, and Y. Yang, "Structure-aware multimodal deep learning for drug-protein interaction prediction," J. Chem. Inf. Model. **62**, 1308–1317 (2022).
[106] V. T. D. Nguyen, N. D. Nguyen, and T. S. Hy (2024). "Complex-based ligand-binding proteins redesign by equivariant diffusion-based generative models," Cold Spring Harbor Laboratory. https://github.com/HySonLab/Protein_Redesign.