Complex-based Ligand-Binding Proteins Redesign by Equivariant Diffusion-based Generative Models

Anonymous authors
Paper under double-blind review

Abstract

Proteins, serving as the fundamental architects of biological processes, interact with ligands to perform a myriad of functions essential for life. The design and optimization of ligand-binding proteins are pivotal for advancing drug development and enhancing therapeutic efficacy. In this study, we introduce ProteinReDiff, a novel computational framework designed to revolutionize the redesign of ligand-binding proteins. Distinguished by its utilization of Equivariant Diffusion-based Generative Models and advanced computational modules, 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 thorough evaluation across sequence diversity, structural preservation, and ligand binding affinity underscores ProteinReDiff's potential to significantly advance computational drug discovery and protein engineering. We will release our data and source code upon acceptance.

1 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 (Du et al., 2016). 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 (Listov et al., 2024). 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 (Yang & Lai, 2017), 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. 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 (Polizzi & DeGrado, 2020; Stärk et al., 2023; Dauparas et al., 2023), such as detailed protein configurations and specific binding pocket data, limiting their applicability, especially in urgent scenarios like the emergence of novel diseases. For instance, during the outbreak of a new disease like COVID-19 (Lv et al., 2020), the spike proteins of the virus may not have well-characterized binding sites, delaying the development of effective drugs. Furthermore, the complexity of binding mechanisms, including allosteric effects and cryptic pockets, adds another layer of difficulty. In addition, many proteins do not exhibit clear binding pockets until ligands are in close vicinity, necessitating extensive simulations to reveal potential binding interfaces (Meller et al., 2023). This complexity underscores the need for a drug design methodology that is agnostic to predefined binding pockets.

Our study addresses identified challenges by introducing ProteinReDiff, an innovative computational framework developed to enhance the process of redesigning ligand-binding proteins. Originating from the foundational concepts of the Equivariant Diffusion-Based Generative Model for Protein-Ligand Complexes (DPL) (Nakata et al., 2023), ProteinReDiff incorporates key improvements inspired by the unparalleled capabilities of advanced modules from the AlphaFold2 (AF2) model (Jumper et al., 2021). Specifically, we integrate the Outer Product Mean, Single Representation Attention (adapted from MSA row attention module of AF2), and Triangle Multiplicative Updates 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 seamlessly combines the generation of diverse protein sequences with advanced blind docking capabilities. Beginning with a selected protein-ligand pair, our approach strategically masks specific amino acids to enable targeted protein redesign. Central to our strategy is the diffusion model's proficiency in capturing the joint distribution of ligand and protein conformations, meticulously optimized to enhance the protein's affinity for its ligand. A 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, relying solely on initial protein sequences and ligand SMILES (Weininger, 1988) strings. This streamlined approach significantly reduces reliance on detailed structural data, facilitating the redesign process of ligand-binding proteins and expanding the scope for sequence-based exploration of protein-ligand interactions. The final outcome is a new protein sequence with an anticipated higher binding affinity for the ligand, underscoring the substantial potential of our model to significantly impact computational drug discovery and protein engineering.

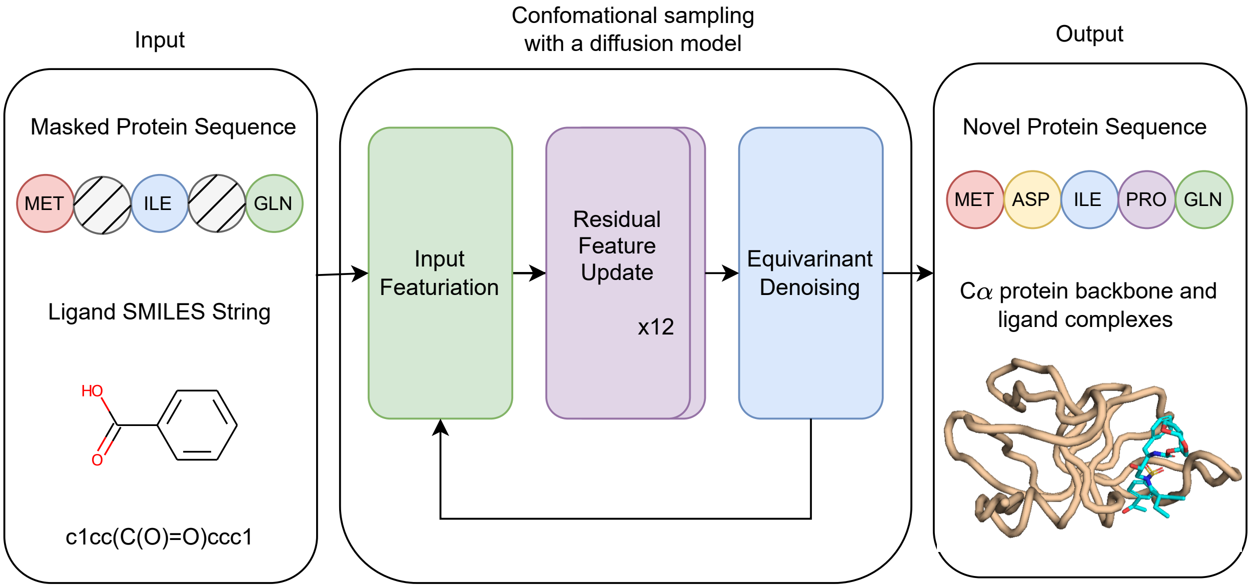


Figure 1: Overview of the proposed framework. The process begins with utilizing a protein amino acid sequence and a ligand SMILES string as inputs. The conformational sampling process includes iteratively applying input featurization, updating residual features, and denoising equivariantly, ultimately yielding novel protein sequences alongside their corresponding $C\alpha$ protein backbone and ligand complexes.

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

- We introduce ProteinReDiff, a novel computational framework for ligand-binding protein redesign, rooted in Equivariant Diffusion-based Generative Models. Our innovation lies in integrating advanced modules to enhance the framework's ability to capture intricate protein-ligand interactions.
- Our framework represents a significant advancement by enabling 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 dimensions, including sequence diversity, structure preservation, and ligand binding affinity, ensuring a holistic assessment of its effectiveness and applicability in various contexts.

2 Background

2.1 Protein language models (PLMs)

Protein Language Models (PLMs) leverage the principles of natural language processing (NLP) to decode the complex language inherent in protein sequences. By treating amino acid sequences as analogous to sentences in human language, PLMs can uncover profound insights into protein functions, interactions, and evolutionary histories. The foundational premise of PLMs is that sequences of amino acids can be conceptualized similarly to sentences composed of words, enabling the application of advanced text processing techniques to predict the structural, functional, and interactional properties of proteins based solely on their amino acid sequences. The field has seen the development of several influential PLMs (Brandes et al., 2022; Elnaggar et al., 2022; Rives et al., 2021; Lin et al., 2023a; Nguyen & Hy, 2023; Ngo & Hy, 2024), each contributing unique perspectives and capabilities to the understanding of protein sequences. The adoption of PLMs in protein design has spurred significant advancements, with numerous studies (Madani et al., 2023; Ruffolo & Madani, 2024; Min et al., 2024; Zheng et al., 2023; Tran & Hy, 2023; Ngo & Hy, 2024) leveraging these models to transform sequence data into detailed insights, thereby guiding the engineering of proteins with targeted functional properties.

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 predicted structural and functional properties:

$$X = F(S), \quad X \in R^d,$$

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 essential for determining the protein's three-dimensional structure and biological functionality. PLMs, through rigorous training on extensive databases of known protein sequences and structures, acquire the ability to discern the "grammar" that governs protein folding and function, facilitating accurate predictions about unseen proteins.

In our research, we employ the ESM-2 model (Lin et al., 2023a), a state-of-the-art protein language model with 650 million parameters, pre-trained on nearly 65 million unique protein sequences from the UniRef (Suzek et al., 2014) database. Distinguished by its comprehensive training regimen encompassing a wide variety of protein sequences, ESM-2 adeptly identifies structural and phylogenetic patterns across a broad spectrum of proteins. Its capacity to infer accurate protein structures from their amino acid sequences marks a pivotal advancement in protein science, providing a robust tool for exploring protein functionality and evolutionary dynamics without resorting to traditional, labor-intensive structural determination methods. Employing ESM-2 enabled us to derive structural and phylogenetic information from the input sequences, significantly augmenting our understanding of the underlying protein mechanisms. This enhanced understanding is instrumental in the design and optimization of proteins for specific functions, including ligand-binding activities.

2.2 Equivariant diffusion-based generative models

In our research, we utilize a generative model driven by equivariant diffusion principles, drawing from the foundations laid by Variational Diffusion Models (Kingma et al., 2023) and E(3) Equivariant Diffusion Models (Hoogeboom et al., 2022).

2.2.1 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

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

where α_t and σ_t^2 represent strictly positive scalar functions of t, dictating the extent of signal preservation versus 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

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

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:

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

with

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

2.2.2 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:

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

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:

$$p(z_1) = N_x(0, I).$$

Furthermore, under the variance-conserving framework and considering $\alpha_0 \approx 1$, the distribution $q(z_0|x)$ is modeled as narrowly concentrated, allowing for the approximation of $p_{data}(x)$ as uniform across this concentration. This yields:

$$q(\mathbf{z}_0|\mathbf{x}) = \frac{q(\mathbf{z}_0|\mathbf{x}) p_{\text{data}}(\mathbf{x})}{\int_{\tilde{x}} q(\mathbf{z}_0|\tilde{\mathbf{x}}) p_{\text{data}}(\tilde{\mathbf{x}})} \approx \frac{q(\mathbf{z}_0|\mathbf{x})}{\int_{\tilde{x}} q(\mathbf{z}_0|\tilde{\mathbf{x}})} = \mathcal{N}(\mathbf{z}_0|\mu_0, \sigma_0^2/\alpha^2).$$

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

$$p(x|z_0) = N_x(x|z_0/\alpha_0, \sigma_0^2/\alpha_0^2 I).$$

The generative model's conditional distributions are then formulated as:

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

which mirrors $q(z_s|z_t, x)$ 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 $\hat{\epsilon}_\theta(z_t;t)$, is articulated as:

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

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

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

3 Method

In this section, we detail the methodology employed in our noise prediction model, which is depicted in Figure 1 and consists of three main procedure: (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.

3.1 Input featurization

We develop both single and pair representations from protein sequences and ligand SMILES string (Figure 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 (Jumper et al., 2021). 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 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.

3.2 Residual feature update procedure

Our methodology marks a notable departure from the residual feature update procedure utilized in the original DPL model (Nakata et al., 2023). While the DPL model relied on Folding blocks from ESMFold (Lin et al., 2023b) for updating single and pair representations, wherein the two representations mutually influence each other, our aim is to enhance the efficiency of this procedure. Specifically, we integrate enhancements such as the Outer Product Mean, Single Representation Attention, and Triangle Multiplicative updates, drawing inspiration from the AlphaFold2 (Jumper et al., 2021) model. Notably, we adapt and customize these modules to align with our model architecture, ensuring their optimal performance in representing the complex interplay between proteins and ligands.

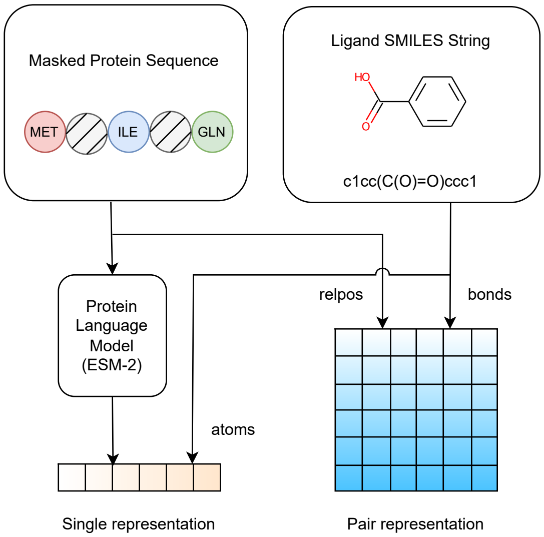


Figure 2: Overview of the input featurization procedure of the model.

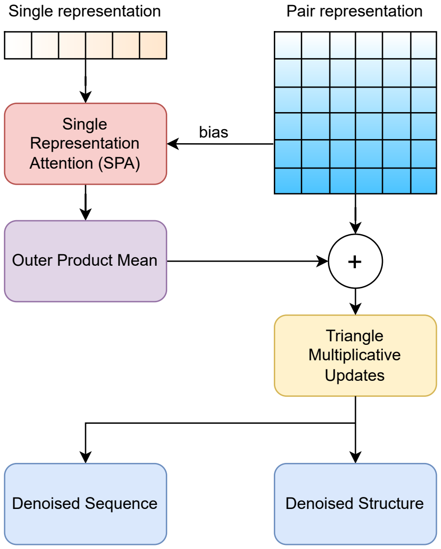


Figure 3: Overview of the residual feature update procedure of the model.

3.2.1 Single representation attention module

Our Single Representation Attention (SPA) module, derived from the Alphafold2 model's MSA row attention with pair bias, brings a significant enhancement to our framework. In the original Alphafold2, the Multiple Sequence Alignment (MSA) row attention mechanism is designed to process input from a single sequence, while the SPA module is tailored to incorporate representations from multiple protein-ligand complexes concurrently. Specifically, the pair bias component of the SPA attention module is strategically employed to capture the nuanced interactions between proteins and ligands. Through simultaneous consideration of both the single representation vector (which encodes the protein/ligand sequential representation) and the pair representation vector (which encodes protein-ligand interactions), this cross-attention mechanism adeptly preserves internal motifs, as evidenced by contact overlap metrics (Rao et al., 2021). Moreover, its efficacy extends to binding affinity prediction, as demonstrated in section 4.2.6. For a detailed description of the computational steps implemented in this module, refer to Algorithm 1.

Algorithm 1 Single Representation Attention pseudocode

Input: Single representation vector m_{si}, pair representation vector $z_{sij} of s sequences, C = 65, N_{head} = 4$

Output: Updated single representation vector $\tilde{m}_{si}$

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

3.2.2 Outer product mean

The outer product layer merges insights from SPA and channels them into pair representations, enabling further refinement. Within this layer, evolutionary cues generate plausible structural hypotheses, which are then transferred to pair representations using Algorithm 2. Analogous to Tensor Product Representations (TPR) in NLP, the role of the outer product is akin to consolidating congruent information from residue pairs (Huang et al., 2018; Smolensky, 1990; Huang et al., 2019). Although not directly translatable from NLP, this process integrates correlated information from sequence s for residues i and j, resulting in intermediate tensors. These tensors amalgamate all available data, culminating in coherent representations. Subsequently, mean computation aggregates these representations, followed by an affine transformation to derive hypotheses concerning the relative positions of residues i and j. This information is then conveyed to pair representations for a comprehensive assessment of plausibility, considering other data and physical constraints. For a detailed description of the computational steps implemented in this module, refer to Algorithm 2. Note that in this implementation, we have adapted outer product without the mean to maintain the pair representations of mutiple protein-ligand pairs.

Algorithm 2 Outer product mean pseudocode

Input: Single representation vector $m_{si} of s sequences, C = 32$

Output: Pair representation vector z_{sij}

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

3.2.3 Triangle multiplicative updates

After refining the pair representations, our model interprets the primary protein-ligand structure using principles from graph theory, treating each residue as a distinct entity interconnected through the pair representation. These connections are then refined through triangular multiplicative updates, which account for physical and geometric constraints, such as the triangular inequality on distance. While attention mechanisms help identify residues with significant influence, the utilization of triangular multiplicative updates becomes crucial in preventing excessive focus on specific residue subsequences while equivariantly transforming the denoised coordinates, evidenced in implementation of (Lin & AlQuraishi, 2023). By aggregating information from neighboring residues and considering the third edge of each triangle, 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.

3.3 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 3D space for each atom, as shown in the equation (Nakata et al., 2023):

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

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

$$\begin{aligned}
\hat{e}_i(\mathbf{Rz} + \mathbf{t}) &= \sum_j \frac{W_{ij}(\mathbf{Rz} + \mathbf{t})}{\|(\mathbf{Rz}_i + \mathbf{t}) - (\mathbf{Rz}_j + \mathbf{t})\|} \cdot ((\mathbf{Rz}_i + \mathbf{t}) - (\mathbf{Rz}_j + \mathbf{t})) \\
&= \mathbf{R} \sum_j \frac{W_{ij}(\mathbf{Rz} + \mathbf{t})}{\|\mathbf{z}_i - \mathbf{z}_j\|} \cdot (\mathbf{z}_i - \mathbf{z}_j) \\
&= \mathbf{R} \sum_j \frac{W_{ij}(\mathbf{z})}{\|\mathbf{z}_i - \mathbf{z}_j\|} \cdot (\mathbf{z}_i - \mathbf{z}_j) \\
&= \mathbf{R}\hat{e}_i(\mathbf{z})
\end{aligned}$$

for any rotation $\mathbf{R}$ and translation $\mathbf{t}$. 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.

4 Experiments

4.1 Training process

4.1.1 Materials

Our training strategy harnesses a meticulously curated dataset encompassing a broad range of protein structures, including both ligand-bound (holo) and ligand-free (apo) forms, sourced from two key repositories: PDBBind v2020 (Wang et al., 2004) and CATH 4.2 (Sillitoe et al., 2018). PDBBind v2020 offers a diverse collection of protein-ligand complexes, while CATH 4.2 provides a substantial repository of protein structures. Each dataset was selected for its unique contributions to our understanding of protein-ligand interactions and structural diversity. This strategic selection of datasets ensures our model is exposed to a wide and varied spectrum of protein-ligand interactions and structural configurations, enabling a comprehensive evaluation against diverse inverse folding benchmarks. By training on both holo and apo structures, our approach not only aims to imbue the model with a robust understanding of protein-ligand dynamics but also equips it to adeptly navigate the complexities of unseen protein-ligand interaction scenarios. To ensure robust model training and evaluation, we employ careful data partitioning techniques. The dataset is divided into distinct subsets, including training, validation, and test sets. Table 1 provides an overview of the partitioning details, facilitating a clear understanding of the distribution of samples across different subsets of the dataset.

- **PDBBind v2020**: For consistency and comparability with previous studies, we adhered closely to the test/training/validation split settings outlined in established literature, specifically following the configurations defined in the respective sources for the PDBBind v2020 datasets (Koh et al., 2023).
- **CATH 4.2**: In our approach, we deliberately focused on proteins with fewer than 400 amino acids 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 training and validation phases, SMILES strings of CATH 4.2 proteins were represented as asterisks (masked tokens) to denote missing ligands. Notably, CATH 4.2 was excluded

from the test set due to the absence of corresponding ligands required for evaluating protein-ligand interactions.

Table 1: Data Partitioning Overview (Unit: number of samples)

Dataset	**Training**	**Validation**	**Test**
PDBBind v2020	9430	552	207
CATH 4.2	15261	939	-

4.1.2 Loss functions

The optimization of our model for ligand-binding protein predesign is governed by a composite loss function L, structured to facilitate the intricate balance required for predicting and enhancing protein-ligand interactions. The loss function is formulated as follows:

$$L = L_{\text{WS}} + L_{\text{KL}} + L_{\text{CE}},$$

The optimization of our model for ligand-binding protein pre-design is governed by a composite loss function L, structured to facilitate the intricate balance required for predicting and enhancing protein-ligand interactions. The loss function is formulated as follows:

$$L = L_{\text{WS}} + L_{\text{KL}} + L_{\text{CE}}.$$

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 denoising step, enabling accurate noise prediction for atoms in the protein-ligand complex. Defined as:

$$L_{\text{WS}} = \sum_j \frac{W_{ij}(\mathbf{z})(\mathbf{z}_i - \mathbf{z}_j)}{\|\mathbf{z}_i - \mathbf{z}_j\|},$$

where $W_{ij}(\mathbf{z})$ are the weights for the differences between atom i and atom j, and $(\mathbf{z}_i - \mathbf{z}_j)$ is the vector difference, normalized by its magnitude.

Kullback-Leibler divergence (L_{KL}) (Joyce, 2011) This component quantifies the divergence between the model's predictions and actual sequence data at timestep $t-1$, playing a pivotal role in the denoising process. Defined as $KL(x_{\text{pred\_t-1}}, seq_{\text{t-1}})$, it contrasts the predicted distribution, $x_{\text{pred\_t-1}}$, against the true sequence distribution, $seq_{\text{t-1}}$, leveraging the diffusion process's γ parameter for temporal adjustment. This approach ensures the model's predictions progressively align with actual data distributions, significantly enhancing the accuracy of sequence and structure generation by minimizing the expected divergence.

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 assigns each amino acid 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. The result is a marked improvement in the precision of sequence predictions.

4.1.3 Training performance

Throughout the training phase, we meticulously observed the model's performance, paying close attention to the dynamics between training and validation losses, as demonstrated in Figure 4. While the training loss consistently diminished, indicating effective learning, the validation loss presented a more erratic behavior—fluctuating around the training loss yet trending downwards overall. Such fluctuations in validation loss, despite its general decline, suggest the model's adaptive optimization in the face of complex data patterns. The overall downward trend in both metrics, with validation loss closely mirroring the training loss albeit with fluctuations, highlights the model's capacity for generalization to unseen data without significant overfitting.

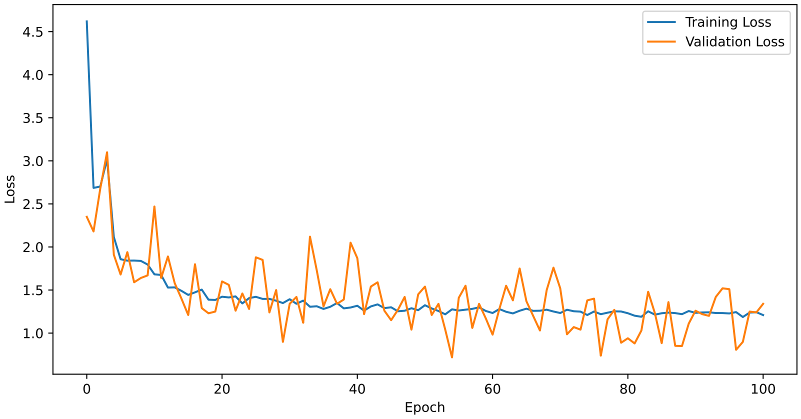


Figure 4: Training history chart of ProteinReDiff, showcasing the evolution of training and validation losses over epochs.

4.2 Evaluation process

4.2.1 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. Within this context, our model, ProteinReDiff, is evaluated based on its ability to tailor 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.

4.2.2 Sequence diversity

In computational protein design, sequence diversity is crucial for fostering innovation. 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) (Miller et al., 2009) 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:

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

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.

4.2.3 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: the Template Modeling Score (TM Score) (Zhang & Skolnick, 2004), the Root Mean Square Deviation (RMSD) (Laskowski & de Beer, 2014), and the Contact Overlap (CO) (Bastolla et al., 2023). These two metrics collectively provide a comprehensive assessment of structural integrity and similarity, essential for evaluating the success of our protein redesign efforts.

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:

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

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 root mean square of distances between corresponding atoms, after optimal superposition, which involves finding the best-fit rotation R and translation t that aligns the two sets of points. A lower RMSD value indicates a higher degree of structural similarity, making it a direct measure of the extent to which structural deviation has been minimized. Achieving a low RMSD is desirable, as it signifies that the redesign process has successfully preserved the core structural features of the original protein.

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:

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

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 successful structural preservation.

Contact Overlap (CO) provides a complementary perspective to RMSD and TM Score by focusing on the preservation of specific structural interactions rather than overall geometric similarity. This focus makes CO an essential metric in evaluating the efficacy of protein redesign strategies aimed at maintaining or enhancing protein function while modifying other properties. CO quantitatively measures the conservation of inter-atomic contacts between the original and redesigned protein structures, which are crucial for the protein's structural integrity and functional capabilities. The metric is defined as:

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

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 that many of the original contacts are preserved in the redesigned structure, suggesting that the redesign maintains much of the original protein's structural network, crucial for its stability and function.

4.2.4 Ablation study

Efficiency of our innovations To further validate the effectiveness of the enhancements introduced in ProteinReDiff, we conducted experiments comparing two variants of our model: one incorporating our proposed enhancements and the original DPL model, which was initially designed to generate ensembles of complex structures rather than for targeted protein redesign. To adapt DPL for our purposes, significant adjustments were made to align it with the specific requirements of ligand-binding protein redesign. These modifications allowed us to evaluate how well our enhancements improve upon the base model in terms of ligand binding affinity. This comparative assessment not only highlighted the added value of our improvements but also showcased the adaptability and potential of the original DPL framework when reconfigured for protein redesign tasks.

Impact of masking ratios We examined the impact of varying the percentage of masked amino acids on ProteinReDiff's efficacy. This investigation is key to understanding the optimal level of sequence modification needed to enhance ligand binding affinity and ensure sequence diversity as well as structural preservation. We adjusted the percentage of masked amino acids in the original sequences, observing the effects on performance metrics such as ligand binding affinity and structural integrity. This approach allowed us to determine how changes in the sequence masking strategy influence the model's redesign capabilities. The initial setup involved a minimal percentage of sequence masking, essential for the redesign process, gradually increasing to explore the model's flexibility and its ability to generate sequences with improved characteristics. The study aims to identify a balance that maximizes protein function enhancement while maintaining crucial structural and functional motifs. Findings will inform strategies for sequence modification, highlighting its significance in computational protein design.

4.2.5 Experimental setup

In our experimental setup, we began by employing Omegafold (Wu et al., 2022) to predict the three-dimensional structures of all designed protein sequences. This step was essential since AutoDock Vina (Trott & Olson, 2010), the molecular docking software utilized in our study, necessitates 3D structures to conduct docking simulations and evaluate the binding affinity between the proteins and their respective ligands. To ensure fair comparisons and mitigate potential biases introduced by pre-docked structures, we aligned our redesigned proteins with reference structures. 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, particularly in light of the limited studies addressing similar problems, we compared the docking scores of our redesigned proteins not only with those of the original proteins but also with proteins generated by advanced inverse folding models. Contrasting the docking scores of proteins from both approaches allows us to elucidate the impact of sequence context on ligand binding. Proteins generated by inverse folding models may exhibit different sequence characteristics compared to those explicitly designed for ligand binding affinity. Understanding how these differences influence docking outcomes provides valuable insights into the interplay between protein sequence and structure in determining ligand interactions, enriching the interpretation of our findings and advancing the understanding of protein-ligand interactions. For a detailed comparison of the inputs and outputs of each model, refer to Table 2.

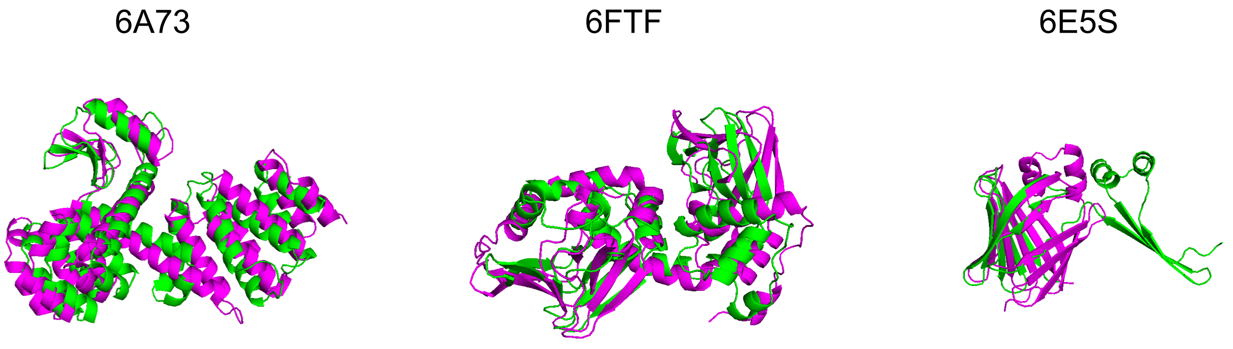
Table 2: Comparison of protein design models based on input and output characteristics

Model	Input			Output		
	Protein Sequence	Protein Structure	Ligand Structure	Protein Sequence	Protein Structure	Ligand Structure
CARP (Yang et al., 2023)	✓	×	×	✓	×	×
ESMIF (Hsu et al., 2022)	×	✓	×	✓	×	×
MIF (Yang et al., 2022)	✓	✓	×	✓	×	×
MIF-ST (Yang et al., 2022)	✓	✓	×	✓	×	×
ProteinMPNN (Dauparas et al., 2022)	×	✓	×	✓	×	×
DPL (Nakata et al., 2023)	✓	×	✓	×	✓	✓
ProteinReDiff (Ours)	✓	×	✓	✓	✓	✓

4.2.6 Results and discussion

Our comprehensive evaluation of ProteinReDiff, as detailed in Table 3 and visually represented in Figure 6, across the metrics of ligand binding affinity, sequence diversity, and structure preservation, has yielded insightful findings. 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 Figure 5, we presented several instances where the contact overlap appeared to be maintained, yet the RMSD remained 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 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.



CO	RMSD
0.931	6.017

C0	RMSD
0.928	7.720

CO	RMSD
0.910	15.381

Figure 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, illustrating the precise structural changes and enhancements achieved through the redesign.

A pivotal observation from our study is ProteinReDiff's unparalleled ability to enhance ligand binding affinity, particularly pronounced at a 15% masking ratio. This configuration not only eclipses the performance of Inverse Folding (IF) models and the original DPL framework, but also crucially exceeds the binding efficiencies of the original protein designs. By incorporating advanced computational modules from AlphaFold2, ProteinReDiff has successfully represented the complex interplay between proteins and ligands, demonstrating its effectiveness in enhancing ligand-binding affinity beyond the capabilities of the original DPL model. While other masking ratios within ProteinReDiff exhibit a range of effectiveness, lower ratios, though better than baseline models, fall short of the peak performance observed at 15%. Conversely, higher ratios miss the mark on achieving the necessary balance between introducing beneficial modifications and maintaining functional precision, highlighting the critical nature of optimizing the masking ratio.

In analyzing sequence diversity and structure preservation metrics, we found an intricate balance that underscores the complexity of protein redesign. The 15% masking ratio, our model's sweet spot for enhancing ligand binding affinity, also achieves outcomes equivalent to baseline methods in both sequence diversity and structure preservation. This equilibrium suggests that while ProteinReDiff excels in optimizing ligand interactions, a paramount objective in protein redesign, it does so without compromising on the diversity of sequences explored or the structural integrity of the original proteins. Notably, 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. This nuanced understanding of the interplay between these critical metrics emphasizes the sophistication of ProteinReDiff's approach to protein redesign, promising a method that respects the delicate balance necessary for successful therapeutic development.

Moreover, the capability of ProteinReDiff to operate efficiently using merely protein sequences and ligand SMILES as inputs, without the requirement for detailed 3D structural data, marks a significant leap in computational protein design. This feature gains particular importance considering that the model not only achieves structure preservation at levels competitive with Inverse Folding (IF) models, which rely heavily on structural information for sequence optimization but also excels in enhancing ligand binding affinity. This approach insightfully captures and applies the core principles of protein-ligand interactions and structural fidelity directly from sequence data.

Table 3: Comparison of method performance across multiple metrics: Ligand binding affinity, sequence diversity, and structure preservation. Ligand binding affinity (LBA) and structure preservation metrics are reported as mean values derived from the dataset's samples.

Category	Method	LBA (kcal/mol) ↓	Sequence diversity ↑	Structure preservation		
				TM Score ↑	RMSD (Å) ↓	CO ↑
Baseline	CARP	-5.657	185.532	0.849	3.767	0.922
	MIF	-5.488	185.600	**0.876**	**2.986**	0.938
	MIF-ST	-5.596	185.584	0.871	3.026	0.936
	ESMIF	-5.555	187.512	0.837	4.000	0.915
	ProteinMPNN	-5.422	188.792	0.714	6.805	0.859
	DPL	-5.300	183.231	0.780	5.200	0.870
	Reference cases	-5.847	-	-	-	-
ProteinReDiff (Ours)	5% Masking	-5.804	185.935	0.863	3.196	**0.942**
	15% Masking	**-6.803**	186.627	0.844	3.689	0.934
	30% Masking	-5.769	187.877	0.803	4.467	0.916
	40% Masking	-5.617	188.600	0.756	5.639	0.896
	60% Masking	-5.467	**190.425**	0.305	18.056	0.734
	70% Masking	-5.469	187.291	0.147	23.196	0.688

In the visualizations provided in Figure 7, the nuanced distinctions between the original and redesigned proteins underscore the targeted precision of ProteinReDiff. This precision manifests in strategic adjustments at the molecular level, specifically engineered to amplify ligand binding efficiency. Despite the profound impact on binding affinity, these adjustments are executed with such finesse that structural deviations from the original framework are minimal. This approach not only enhances the protein's interaction with ligands but also preserves its structural and functional integrity. The modifications, though seemingly minor, are the result of a complex optimization process, balancing the need for improved functionality with the imperative to maintain the protein's overall architecture.

5 Conclusions

In conclusion, this study presents ProteinReDiff, a groundbreaking computational framework designed for the advanced redesign of ligand-binding proteins. Leveraging state-of-the-art techniques inspired by Equivariant Diffusion-Based Generative Models and the transformative insights of AlphaFold2, ProteinReDiff showcases a profound ability to decipher and enhance complex protein-ligand interactions. Uniquely, our model excels in optimizing ligand binding affinity based solely on the initial protein sequences and ligand SMILES strings, circumventing the traditional reliance on comprehensive structural data. The experimental validations of ProteinReDiff illuminate its remarkable capacity not only to elevate ligand binding affinity but also to

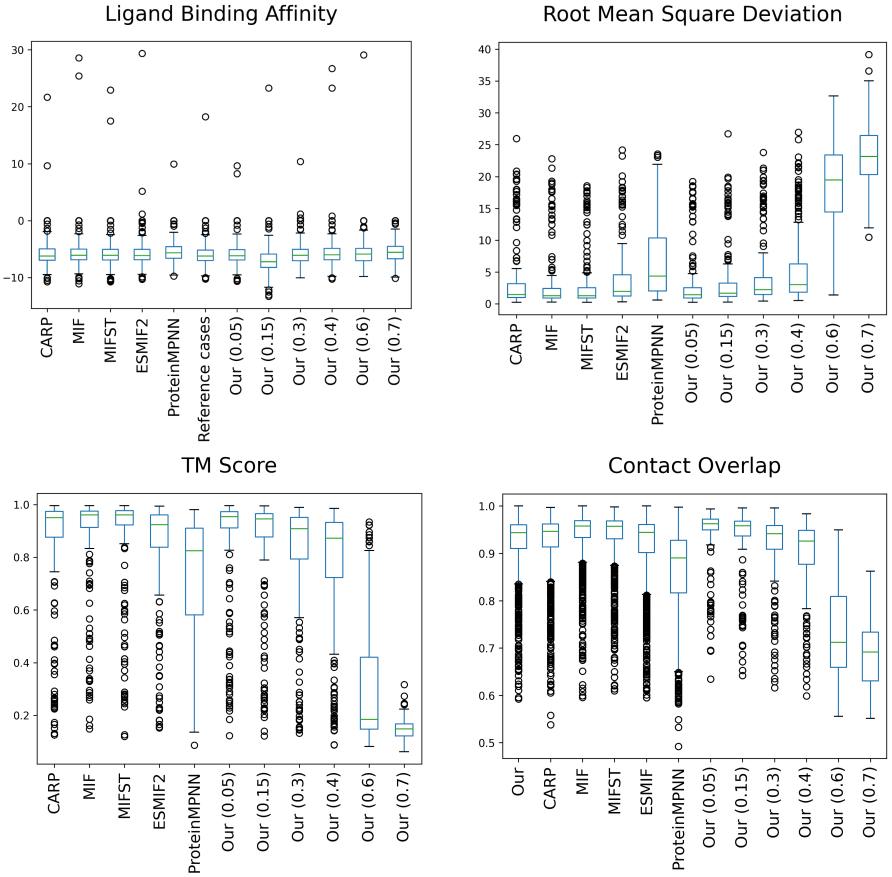


Figure 6: Boxplot illustrating the distribution of ligand binding affinities, and structure preservation metrics (TM Score and RMSD) across all methods evaluated, including baseline models and variations of ProteinReDiff. Each boxplot showcases the median, quartiles, and outliers within the data, providing insight into the variability and central tendency of each metric across the dataset's samples.

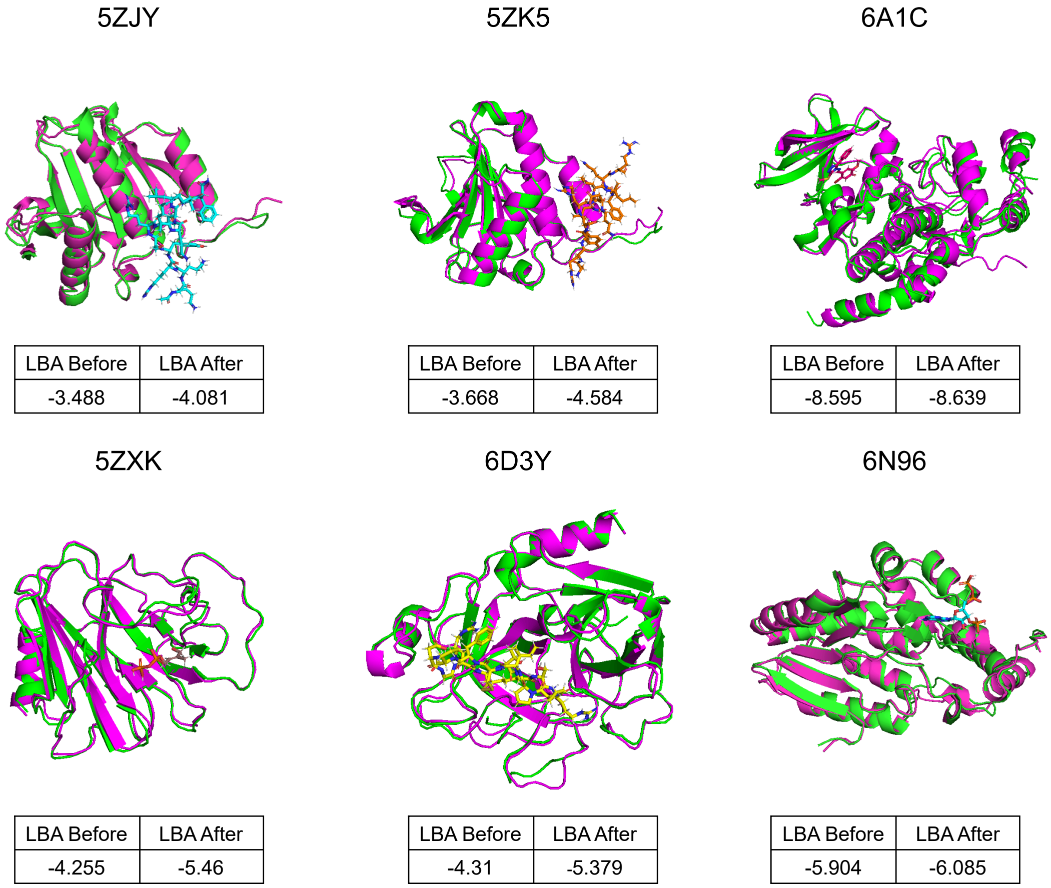
maintain essential sequence diversity and structural integrity. These findings herald a significant leap forward in protein-ligand complex modeling, suggesting a bright future for ProteinReDiff in various biotechnological and pharmaceutical domains. The success of ProteinReDiff paves the way for its further development and broad application, promising to revolutionize approaches to drug design and protein engineering.

References

Peizhen Bai, Filip Miljković, Bino John, and Haiping Lu. Interpretable bilinear attention network with domain adaptation improves drug–target prediction. *Nature Machine Intelligence*, 5(2):126–136, Feb 2023a. ISSN 2522-5839. doi: 10.1038/s42256-022-00605-1. URL `https://doi.org/10.1038/s42256-022-00605-1`.

Peizhen Bai, Filip Miljković, Bino John, and Haiping Lu. Interpretable bilinear attention network with domain adaptation improves drug–target prediction. *Nature Machine Intelligence*, 5(2):126–136, Feb 2023b. ISSN 2522-5839. doi: 10.1038/s42256-022-00605-1. URL `https://doi.org/10.1038/s42256-022-00605-1`.

Ugo Bastolla, David Abia, and Oscar Piette. PC_ali: a tool for improved multiple alignments and evolutionary inference based on a hybrid protein sequence and structure similarity score. *Bioinformatics*, 39(11):btad630, 10 2023. ISSN 1367-4811. doi: 10.1093/bioinformatics/btad630. URL `https://doi.org/10.1093/bioinformatics/btad630`.



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

Figure 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 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.

Nadav Brandes, Dan Ofer, Yam Peleg, Nadav Rappoport, and Michal Linial. ProteinBERT: a universal deep-learning model of protein sequence and function. *Bioinformatics*, 38(8):2102–2110, 02 2022. ISSN 1367-4803. doi: 10.1093/bioinformatics/btac020. URL `https://doi.org/10.1093/bioinformatics/btac020`.

Lifan Chen, Xiaoqin Tan, Dingyan Wang, Feisheng Zhong, Xiaohong Liu, Tianbiao Yang, Xiaomin Luo, Kaixian Chen, Hualiang Jiang, and Mingyue Zheng. TransformerCPI: improving compound–protein interaction prediction by sequence-based deep learning with self-attention mechanism and label reversal experiments. *Bioinformatics*, 36(16):4406–4414, 05 2020. ISSN 1367-4803. doi: 10.1093/bioinformatics/btaa524. URL `https://doi.org/10.1093/bioinformatics/btaa524`.

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(6615):49–56, 2022. doi: 10.1126/science.add2187. URL `https://www.science.org/doi/abs/10.1126/science.add2187`.

Justas Dauparas, Gyu Rie Lee, Robert Pecoraro, Linna An, Ivan Anishchenko, Cameron Glasscock, and D. Baker. Atomic context-conditioned protein sequence design using ligandmpnn. *bioRxiv*, 2023. doi: 10.1101/2023.12.22.573103. URL `https://www.biorxiv.org/content/early/2023/12/23/2023.12.22.573103`.

Xing Du, Yi Li, Yuan-Ling Xia, Shi-Meng Ai, Jing Liang, Peng Sang, Xing-Lai Ji, and Shu-Qun Liu. Insights into protein-ligand interactions: Mechanisms, models, and methods. *Int. J. Mol. Sci.*, 17(2):144, January 2016.

Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, Debsindhu Bhowmik, and Burkhard Rost. Prottrans: Toward understanding the language of life through self-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):7112–7127, 2022. doi: 10.1109/TPAMI.2021.3095381.

Emiel Hoogeboom, Victor Garcia Satorras, Clément Vignac, and Max Welling. Equivariant diffusion for molecule generation in 3d, 2022.

Chloe Hsu, Robert Verkuil, Jason Liu, Zeming Lin, Brian Hie, Tom Sercu, Adam Lerer, and Alexander Rives. Learning inverse folding from millions of predicted structures. *ICML*, 2022. doi: 10.1101/2022.04.10.487779. URL `https://www.biorxiv.org/content/early/2022/04/10/2022.04.10.487779`.

Kexin Huang, Cao Xiao, Lucas M Glass, and Jimeng Sun. MolTrans: Molecular Interaction Transformer for drug–target interaction prediction. *Bioinformatics*, 37(6):830–836, 10 2020. ISSN 1367-4803. doi: 10.1093/bioinformatics/btaa880. URL `https://doi.org/10.1093/bioinformatics/btaa880`.

Qiuyuan Huang, Paul Smolensky, Xiaodong He, Li Deng, and Dapeng Wu. Tensor product generation networks for deep NLP modeling. In Marilyn Walker, Heng Ji, and Amanda Stent (eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 1263–1273, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1114. URL `https://aclanthology.org/N18-1114`.

Qiuyuan Huang, Li Deng, Dapeng Wu, Chang Liu, and Xiaodong He. Attentive tensor product learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):1344–1351, Jul. 2019. doi: 10.1609/aaai.v33i01.33011344. URL `https://ojs.aaai.org/index.php/AAAI/article/view/3934`.

Dejun Jiang, Chang-Yu Hsieh, Zhenxing Wu, Yu Kang, Jike Wang, Ercheng Wang, Ben Liao, Chao Shen, Lei Xu, Jian Wu, Dongsheng Cao, and Tingjun Hou. Interactiongraphnet: A novel and efficient deep graph representation learning framework for accurate protein–ligand interaction predictions. *Journal of Medicinal Chemistry*, 64(24):18209–18232, 2021. doi: 10.1021/acs.jmedchem.1c01830. URL `https://doi.org/10.1021/acs.jmedchem.1c01830`. PMID: 34878785.

Mingjian Jiang, Zhen Li, Shugang Zhang, Shuang Wang, Xiaofeng Wang, Qing Yuan, and Zhiqiang Wei. Drug–target affinity prediction using graph neural network and contact maps. *RSC Adv.*, 10:20701–20712, 2020. doi: 10.1039/D0RA02297G. URL `http://dx.doi.org/10.1039/D0RA02297G`.

James M. Joyce. *Kullback-Leibler Divergence*, pp. 720–722. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011. ISBN 978-3-642-04898-2. doi: 10.1007/978-3-642-04898-2_327. URL `https://doi.org/10.1007/978-3-642-04898-2_327`.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, Aug 2021. ISSN 1476-4687. doi: 10.1038/s41586-021-03819-2. URL `https://doi.org/10.1038/s41586-021-03819-2`.

Diederik P. Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models, 2023.

David Ryan Koes, Matthew P. Baumgartner, and Carlos J. Camacho. Lessons learned in empirical scoring with smina from the csar 2011 benchmarking exercise. *Journal of Chemical Information and Modeling*, 53(8):1893–1904, Aug 2013. ISSN 1549-9596. doi: 10.1021/ci300604z. URL `https://doi.org/10.1021/ci300604z`.

Huan Yee Koh, Anh TN Nguyen, Shirui Pan, Lauren T May, and Geoffrey I Webb. Psichic: physicochemical graph neural network for learning protein-ligand interaction fingerprints from sequence data. *bioRxiv*, pp. 2023–09, 2023.

Roman Laskowski and Tjaart de Beer. *Root Mean Square Deviation (RMSD)*. John Wiley and Sons, Ltd, 2014. ISBN 9780471650126. doi: https://doi.org/10.1002/9780471650126.dob0640.pub2. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/9780471650126.dob0640.pub2`.

Shuangli Li, Jingbo Zhou, Tong Xu, Liang Huang, Fan Wang, Haoyi Xiong, Weili Huang, Dejing Dou, and Hui Xiong. Structure-aware interactive graph neural networks for the prediction of protein-ligand binding affinity. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, KDD '21, pp. 975–985, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383325. doi: 10.1145/3447548.3467311. URL `https://doi.org/10.1145/3447548.3467311`.

Yeqing Lin and Mohammed 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.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023a. doi: 10.1126/science.ade2574. URL `https://www.science.org/doi/abs/10.1126/science.ade2574`.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023b. doi: 10.1126/science.ade2574. URL `https://www.science.org/doi/abs/10.1126/science.ade2574`.

Dina Listov, Casper A. Goverde, Bruno E. Correia, and Sarel Jacob Fleishman. Opportunities and challenges in design and optimization of protein function. *Nature Reviews Molecular Cell Biology*, Apr 2024. ISSN 1471-0080. doi: 10.1038/s41580-024-00718-y. URL `https://doi.org/10.1038/s41580-024-00718-y`.

Wei Lu, Qifeng Wu, Jixian Zhang, Jiahua Rao, Chengtao Li, and Shuangjia Zheng. Tankbind: Trigonometry-aware neural networks for drug-protein binding structure prediction. *bioRxiv*, 2022. doi: 10.1101/2022.06.06.495043. URL `https://www.biorxiv.org/content/early/2022/06/06/2022.06.06.495043`.

Meng Lv, Xufei Luo, Janne Estill, Yunlan Liu, Mengjuan Ren, Jianjian Wang, Qi Wang, Siya Zhao, Xiaohui Wang, Shu Yang, Xixi Feng, Weiguo Li, Enmei Liu, Xianzhuo Zhang, Ling Wang, Qi Zhou, Wenbo Meng, Xiaolong Qi, Yangqin Xun, Xuan Yu, Yaolong Chen, and COVID-19 evidence and recommendations working group. Coronavirus disease (COVID-19): a scoping review. *Euro Surveill.*, 25(15), April 2020.

Ali Madani, Ben Krause, Eric R. Greene, Subu Subramanian, Benjamin P. Mohr, James M. Holton, Jose Luis Olmos, Caiming Xiong, Zachary Z. Sun, Richard Socher, James S. Fraser, and Nikhil Naik. Large language models generate functional protein sequences across diverse families. *Nature Biotechnology*, 41(8):1099–1106, Aug 2023. ISSN 1546-1696. doi: 10.1038/s41587-022-01618-2. URL `https://doi.org/10.1038/s41587-022-01618-2`.

Andrew T. McNutt, Paul Francoeur, Rishal Aggarwal, Tomohide Masuda, Rocco Meli, Matthew Ragoza, Jocelyn Sunseri, and David Ryan Koes. Gnina 1.0: molecular docking with deep learning. *Journal of Cheminformatics*, 13(1):43, Jun 2021. ISSN 1758-2946. doi: 10.1186/s13321-021-00522-2. URL `https://doi.org/10.1186/s13321-021-00522-2`.

Artur Meller, Michael Ward, Jonathan Borowsky, Meghana Kshirsagar, Jeffrey M. Lotthammer, Felipe Oviedo, Juan Lavista Ferres, and Gregory R. Bowman. Predicting locations of cryptic pockets from single protein structures using the pocketminer graph neural network. *Nature Communications*, 14(1): 1177, Mar 2023. ISSN 2041-1723. doi: 10.1038/s41467-023-36699-3. URL `https://doi.org/10.1038/s41467-023-36699-3`.

Frederic P. Miller, Agnes F. Vandome, and John McBrewster. *Levenshtein Distance: Information theory, Computer science, String (computer science), String metric, Damerau?Levenshtein distance, Spell checker, Hamming distance*. Alpha Press, 2009. ISBN 6130216904.

Xiaoping Min, Chongzhou Yang, Jun Xie, Yang Huang, Nan Liu, Xiaocheng Jin, Tianshu Wang, Zhibo Kong, Xiaoli Lu, Shengxiang Ge, Jun Zhang, and Ningshao Xia. Tpgen: a language model for stable protein design with a specific topology structure. *BMC Bioinformatics*, 25(1):35, Jan 2024. ISSN 1471-2105. doi: 10.1186/s12859-024-05637-5. URL `https://doi.org/10.1186/s12859-024-05637-5`.

Shuya Nakata, Yoshiharu Mori, and Shigenori Tanaka. End-to-end protein–ligand complex structure generation with diffusion-based generative models. *BMC Bioinformatics*, 24(1):233, Jun 2023. ISSN 1471-2105. doi: 10.1186/s12859-023-05354-5. URL `https://doi.org/10.1186/s12859-023-05354-5`.

Nhat Khang Ngo and Truong Son Hy. Multimodal protein representation learning and target-aware variational auto-encoders for protein-binding ligand generation. *Machine Learning: Science and Technology*, 2024. URL `http://iopscience.iop.org/article/10.1088/2632-2153/ad3ee4`.

Thin Nguyen, Hang Le, Thomas P Quinn, Tri Nguyen, Thuc Duy Le, and Svetha Venkatesh. GraphDTA: predicting drug–target binding affinity with graph neural networks. *Bioinformatics*, 37(8):1140–1147, 10 2020. ISSN 1367-4803. doi: 10.1093/bioinformatics/btaa921. URL `https://doi.org/10.1093/bioinformatics/btaa921`.

Viet Thanh Duy Nguyen and Truong Son Hy. Multimodal pretraining for unsupervised protein representation learning. *bioRxiv*, 2023. doi: 10.1101/2023.11.29.569288. URL `https://www.biorxiv.org/content/early/2023/12/07/2023.11.29.569288`.

Nicholas F. Polizzi and William F. DeGrado. A defined structural unit enables de novo design of small-molecule–binding proteins. *Science*, 369(6508):1227–1233, 2020. doi: 10.1126/science.abb8330. URL `https://www.science.org/doi/abs/10.1126/science.abb8330`.

Roshan M Rao, Jason Liu, Robert Verkuil, Joshua Meier, John Canny, Pieter Abbeel, Tom Sercu, and Alexander Rives. Msa transformer. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8844–8856. PMLR, 18–24 Jul 2021. URL `https://proceedings.mlr.press/v139/rao21a.html`.

Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C. Lawrence Zitnick, Jerry Ma, and Rob Fergus. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15):e2016239118, 2021. doi: 10.1073/pnas.2016239118. URL `https://www.pnas.org/doi/abs/10.1073/pnas.2016239118`.

Jeffrey A. Ruffolo and Ali Madani. Designing proteins with language models. *Nature Biotechnology*, 42(2): 200–202, Feb 2024. ISSN 1546-1696. doi: 10.1038/s41587-024-02123-4. URL `https://doi.org/10.1038/s41587-024-02123-4`.

Ian Sillitoe, Natalie Dawson, Tony E Lewis, Sayoni Das, Jonathan G Lees, Paul Ashford, Adeyelu Tolulope, Harry M Scholes, Ilya Senatorov, Andra Bujan, Fatima Ceballos Rodriguez-Conde, Benjamin Dowling, Janet Thornton, and Christine A Orengo. CATH: expanding the horizons of structure-based functional annotations for genome sequences. *Nucleic Acids Research*, 47(D1):D280–D284, 11 2018. ISSN 0305-1048. doi: 10.1093/nar/gky1097. URL `https://doi.org/10.1093/nar/gky1097`.

Paul Smolensky. Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial Intelligence*, 46(1):159–216, 1990. ISSN 0004-3702. doi: https://doi.org/10.1016/0004-3702(90)90007-M. URL `https://www.sciencedirect.com/science/article/pii/000437029090007M`.

Marta M Stepniewska-Dziubinska, Piotr Zielenkiewicz, and Pawel Siedlecki. Development and evaluation of a deep learning model for protein–ligand binding affinity prediction. *Bioinformatics*, 34(21):3666–3674, 05 2018. ISSN 1367-4803. doi: 10.1093/bioinformatics/bty374. URL `https://doi.org/10.1093/bioinformatics/bty374`.

Hannes Stärk, Bowen Jing, Regina Barzilay, and Tommi Jaakkola. Harmonic self-conditioned flow matching for multi-ligand docking and binding site design, 2023.

Baris E. Suzek, Yuqi Wang, Hongzhan Huang, Peter B. McGarvey, Cathy H. Wu, and the UniProt Consortium. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*, 31(6):926–932, 11 2014. ISSN 1367-4803. doi: 10.1093/bioinformatics/btu739. URL `https://doi.org/10.1093/bioinformatics/btu739`.

Freyr Sverrisson, Jean Feydy, Bruno E. Correia, and Michael M. Bronstein. Fast end-to-end learning on protein surfaces. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15267–15276, 2021. doi: 10.1109/CVPR46437.2021.01502.

Trong Thanh Tran and Truong Son Hy. Protein design by directed evolution guided by large language models. *bioRxiv*, 2023. doi: 10.1101/2023.11.28.568945. URL `https://www.biorxiv.org/content/early/2023/11/29/2023.11.28.568945`.

Oleg Trott and Arthur J. Olson. Autodock vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of Computational Chemistry*, 31(2):455–461, 2010. doi: https://doi.org/10.1002/jcc.21334. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/jcc.21334`.

Penglei Wang, Shuangjia Zheng, Yize Jiang, Chengtao Li, Junhong Liu, Chang Wen, Atanas Patronov, Dahong Qian, Hongming Chen, and Yuedong Yang. Structure-aware multimodal deep learning for drug–protein interaction prediction. *Journal of Chemical Information and Modeling*, 62(5):1308–1317, Mar 2022. ISSN 1549-9596. doi: 10.1021/acs.jcim.2c00060. URL `https://doi.org/10.1021/acs.jcim.2c00060`.

Renxiao Wang, Xueliang Fang, Yipin Lu, and Shaomeng Wang. The pdbbind database: Collection of binding affinities for protein-ligand complexes with known three-dimensional structures. *Journal of Medicinal Chemistry*, 47(12):2977–2980, 2004. doi: 10.1021/jm030580l. URL `https://doi.org/10.1021/jm030580l`. PMID: 15163179.

David Weininger. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences*, 28(1):31–36, 1988. doi: 10.1021/ci00057a005. URL `https://doi.org/10.1021/ci00057a005`.

Ruidong Wu, Fan Ding, Rui Wang, Rui Shen, Xiwen Zhang, Shitong Luo, Chenpeng Su, Zuofan Wu, Qi Xie, Bonnie Berger, Jianzhu Ma, and Jian Peng. High-resolution de novo structure prediction from primary sequence. *bioRxiv*, 2022. doi: 10.1101/2022.07.21.500999. URL `https://www.biorxiv.org/content/early/2022/07/22/2022.07.21.500999`.

Kevin K Yang, Niccolò Zanichelli, and Hugh Yeh. Masked inverse folding with sequence transfer for protein representation learning. *Protein Engineering, Design and Selection*, 36:gzad015, 10 2022. ISSN 1741-0126. doi: 10.1093/protein/gzad015. URL `https://doi.org/10.1093/protein/gzad015`.

Kevin K. Yang, Nicolo Fusi, and Alex X. Lu. Convolutions are competitive with transformers for protein sequence pretraining. *bioRxiv*, 2023. doi: 10.1101/2022.05.19.492714. URL `https://www.biorxiv.org/content/early/2023/02/23/2022.05.19.492714`.

Wei Yang and Luhua Lai. Computational design of ligand-binding proteins. *Current Opinion in Structural Biology*, 45:67–73, 2017. ISSN 0959-440X. doi: https://doi.org/10.1016/j.sbi.2016.11.021. URL `https://www.sciencedirect.com/science/article/pii/S0959440X16301464`. Engineering and design: New trends in designer proteins.

Yang Zhang and Jeffrey Skolnick. Scoring function for automated assessment of protein structure template quality. *Proteins*, 57(4):702–710, December 2004.

Liangzhen Zheng, Jingrong Fan, and Yuguang Mu. Onionnet: a multiple-layer intermolecular-contact-based convolutional neural network for protein–ligand binding affinity prediction. *ACS Omega*, 4(14):15956–15965, 2019. doi: 10.1021/acsomega.9b01997. URL `https://doi.org/10.1021/acsomega.9b01997`. PMID: 31592466.

Zaixiang Zheng, Yifan Deng, Dongyu Xue, Yi Zhou, Fei Ye, and Quanquan Gu. Structure-informed language models are protein designers. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

A Evaluating Protein-Ligand Complex Representation

Evaluation Methodology In the continuation of our study's exploration of protein-ligand complex representations, we extended the use of the PDBBind v2020 dataset, previously detailed in our training process, to evaluate the effectiveness of embeddings generated by ProteinReDiff. Employing these embeddings as input features, we trained a Gaussian Process (GP) model aimed at predicting 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 our embeddings encapsulate predictive information about protein-ligand interactions. The GP model used a Gaussian likelihood, which is appropriate for regression tasks, along with a Radial Basis Function (RBF) kernel. We chose the RBF kernel due to its effectiveness in modeling smooth, continuous variations, which is characteristic of protein-ligand binding affinities. The training of the GP model focused on optimizing the parameters to ensure a robust fit to the training data.

Table 4: Experimental results of ligand binding affinity prediction task on PDBBind v2020 dataset.

Approach	RMSE ↓ ($-\log K_d/K_i$)	MAE ↓ ($-\log K_d/K_i$)	Pearson ↑	Spearman ↑
Pafnucy (Stepniewska-Dziubinska et al., 2018)	1.435	1.144	0.635	0.587
OnionNet (Zheng et al., 2019)	1.403	1.103	0.648	0.602
IGN (Jiang et al., 2021)	1.404	1.116	0.662	0.638
SIGN (Li et al., 2021)	1.373	1.086	0.685	0.656
SMINA (Koes et al., 2013)	1.466	1.161	0.665	0.663
GNINA (McNutt et al., 2021)	1.740	1.413	0.495	0.494
dMaSIF (Sverrisson et al., 2021)	1.450	1.136	0.629	0.588
TankBind (Lu et al., 2022)	1.345	1.060	0.718	**0.689**
GraphDTA (Nguyen et al., 2020)	1.564	1.223	0.612	0.570
TransCPI (Chen et al., 2020)	1.493	1.201	0.604	0.551
MolTrans (Huang et al., 2020)	1.599	1.271	0.539	0.474
DrugBAN (Bai et al., 2023a)	1.480	1.159	0.657	0.612
DGraphDTA (Jiang et al., 2020)	1.493	1.201	0.604	0.551
WGNN-DTA (Bai et al., 2023b)	1.501	1.196	0.605	0.562
STAMP-DPI (Wang et al., 2022)	1.503	1.176	0.653	0.601
PSICHIC (Koh et al., 2023)	**1.314**	**1.015**	0.710	0.686
ProteinReDiff (Our)	1.443	1.168	**0.721**	0.639

Results and discussion The evaluation of our ProteinReDiff model on the PDBBind v2020 dataset demonstrates competitive results in predicting ligand binding affinity using protein-ligand complex representations, as evidenced in Table 4. It's important to note that this experiment aimed to verify the effectiveness of our protein-ligand complex representation rather than to fine-tune the model for this specific task. Consequently, while our results are promising and competitive with specialized studies focused solely on ligand binding affinity prediction, the primary goal was to validate the representation's capability within the protein redesign framework of ProteinReDiff. This underscores the model's utility in guiding the redesign process effectively, affirming the robustness and applicability of our protein-ligand complex representation strategy.