LIGAND-CONDITIONED DISCRETE DIFFUSION FOR PROTEIN SEQUENCE-STRUCTURE CO-DESIGN

Anonymous authors
Paper under double-blind review

ABSTRACT

Proteins perform their biological functions through three-dimensional structures encoded by amino acid sequences, and ligand-binding protein co-design requires models that generate sequence-structure compatible proteins under explicit ligand constraints. Although continuous diffusion and flow-based models have enabled ligand-aware protein design in coordinate or latent feature spaces, existing discrete diffusion protein language models mainly operate over sequence or structure tokens without direct small-molecule conditioning. We introduce **ProtLiD**^{2}, a **Prot**ein **L**igand-conditioned **D**iscrete **D**iffusion model for protein sequence-structure co-design. ProtLiD^2 jointly generates amino-acid sequence and discrete structure tokens while incorporating ligand chemical and geometric information through geometry-aware cross-attention. Trained on over one million ligand-protein complexes, ProtLiD^2 extends masked discrete diffusion from general sequence-structure generation to ligand-aware functional protein design. For decoding, we combine MDLM-based stochastic candidate selection with confidence-margin ranking under a geometric retention budget, committing argmax predictions at retained positions and deferring the remaining proposals. Experimentally, ProtLiD^2 improves global fold confidence over Complexa in ligand-conditioned whole-protein design, increasing TM-score from 0.672 to 0.802 and pLDDT from 64.55 to 73.00. In ligand-binding pocket co-design, ProtLiD^2 reduces active-site BB-RMSD from 3.46Å for FAIR and 3.40Å for PocketGen to 1.97Å, and improves ligand-aware combined pass rates over PocketGen from 14.86% to 59.73% and from 6.08% to 23.49% under increasingly stringent docking thresholds. These results demonstrate the potential of ligand-conditioned discrete diffusion as an effective token-space framework for functional protein co-design.

1 INTRODUCTION

Proteins are fundamental biomolecules that fold from linear amino acid sequences into three-dimensional structures, enabling diverse functions that drive nearly every biological process across all forms of life, from catalysis and signaling to molecular recognition and cellular regulation. Due to their superior ability to learn the underlying distributions of large-scale training data, recent data-driven generative models have transformed protein design, shifting the field from traditional physics-based and evolutionary-profile-guided design methods Kortemme & Baker (2004); Pearce et al. (2019) toward diffusion-based generative models Watson et al. (2023); Ahern et al. (2026); Butcher et al. (2025); Zhang et al. (2023; 2024); Geffner et al. (2025); Didi et al. (2026); Alamdari et al. (2024); Wang et al. (2024; 2025b); Hsieh et al. (2025). These generative approaches can be broadly categorized into continuous diffusion models Watson et al. (2023); Ahern et al. (2026); Butcher et al. (2025); Zhang et al. (2024; 2023); Geffner et al. (2025); Didi et al. (2026) and discrete diffusion models Alamdari et al. (2024); Wang et al. (2024; 2025b); Hsieh et al. (2025) according to their generation space: continuous diffusion models perform denoising over continuous coordinate or feature representations, whereas discrete diffusion models generate proteins through iterative refinement in tokenized sequence, structure, or joint sequence-structure spaces.

Continuous diffusion- and flow-based protein design models Watson et al. (2023); Ahern et al. (2026); Butcher et al. (2025); Zhang et al. (2024; 2023); Geffner et al. (2025); Didi et al. (2026) have emerged as a powerful class of generative approaches that progressively denoise protein representations to

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

design structures, ranging from backbone-level models to fully atomistic frameworks capable of ligand- or protein-conditioned sequence-structure co-design. In parallel with continuous diffusion models, discrete diffusion models Alamdari et al. (2024); Wang et al. (2024; 2025b); Hsieh et al. (2025) have recently emerged as a complementary protein design paradigm, operating directly in amino-acid or tokenized structure space to generate, inpaint, and co-design protein sequences and structures through iterative denoising. **Although discrete diffusion protein language models have enabled unconditional sequence generation, motif scaffolding, inverse folding, folding, and sequence–structure co-generation, they still lack the explicit ligand-conditioning capability of continuous diffusion models, which can directly generate protein sequences and structures in the context of ligand constraints**. To fill this gap, we present **ProtLiD**^{2}, a **Prot**ein **Li**gand-conditioned **D**iscrete **D**iffusion model for protein sequence–structure co-design that jointly designs protein sequence and structure in discrete token space under explicit ligand conditioning, extending discrete diffusion protein design toward functional ligand-aware generation.

Existing discrete diffusion protein language models, including EvoDiff Alamdari et al. (2024), DPLM Wang et al. (2024), DPLM-2 Wang et al. (2025b) and Geo-DPLM Hsieh et al. (2025), generate proteins through iterative denoising in discrete sequence or structure-token space, typically unmasking multiple tokens in parallel via order-agnostic or confidence-ranked mask decoding. Recent advances in masked discrete diffusion language models Wang et al. (2025a); Arriola et al. (2025); Nie et al. (2025); Kim et al. (2025) suggest that the sampling trajectory, especially the token unmasking order, plays a critical role in generation quality. Building on these decoding strategies, we combine MDLM-based stochastic candidate selection with confidence-margin ranking under a geometric retention budget and term this algorithm as MCM-ReMask. This MCM-ReMask policy ranks only newly proposed non-mask positions and commits their argmax predictions, while a retained-fraction parameter controls additional deferral.

In summary, we highlight our main contributions as follows:

(i) We propose ProtLiD^2, a ligand-conditioned masked discrete diffusion model for joint protein sequence-structure co-design. ProtLiD^2 represents proteins in a unified sequence-structure token space and incorporates ligand chemical and geometric information through geometry-aware cross-attention, extending discrete diffusion protein language modeling from general sequence-structure generation to ligand-aware functional protein design.

(ii) We curate a large-scale ligand-protein complex dataset for ligand-conditioned sequence-structure co-design. After source-specific filtering and leakage removal against the PLINDER benchmark, the final training set contains 1 million ligand-protein complexes.

(iii) We systematically evaluate ProtLiD^2 across sequence-structure, ligand-conditioned whole-protein, and pocket co-design benchmarks, showing improved whole-protein fold confidence over Complexa and substantially better pocket-design accuracy than FAIR and PocketGen, including lower active-site BB-RMSD and higher ligand-aware combined pass rates.

2 Related Work

2.1 Generative Protein Design Model

Continuous diffusion and flow-based models have advanced protein design from backbone generation followed by inverse folding (Dauparas et al., 2022) to atomistic, ligand-aware, and conditioned generation. RFdiffusion-family models (Watson et al., 2023; Ahern et al., 2026; Butcher et al., 2025) extend motif and binder scaffolding to atom-level functional conditioning, while pocket-design models such as FAIR (Zhang et al., 2023) and PocketGen (Zhang et al., 2024) jointly generate ligand-conditioned pocket sequences and atomic structures using refinement or atom/residue/ligand interaction modeling. Partially latent flow-matching models, including La-Proteina (Geffner et al., 2025) and Proteina-Complexa (Didi et al., 2026), further enable joint sequence-structure and target-conditioned binder design in continuous latent spaces.

Discrete diffusion protein language models provide a complementary paradigm to continuous diffusion by generating proteins in amino-acid or tokenized structure spaces. EvoDiff Alamdari et al. (2024) and DPLM Wang et al. (2024) primarily target sequence generation and controllable tasks such as inpainting, motif scaffolding, and inverse folding, while DPLM-2 Wang et al. (2025b) extends

discrete diffusion to sequence-structure co-design through learned backbone tokenization and decoding Jing et al. (2021); Yu et al. (2024); Jumper et al. (2021). Geo-DPLM Hsieh et al. (2025) further improves structure-token modeling with enhanced supervision, refinement, and geometry-aware modules. Despite these advances, explicit ligand-conditioned co-design remains largely unexplored in discrete sequence-structure token space.

To fill this gap, rather than optimizing the structure tokenization module itself, this work adopts an existing protein structure tokenization model, GCP-VQVAE (Pourmirzaei et al., 2025), and focuses on demonstrating that explicit ligand conditioning can be effectively integrated into discrete diffusion models for high-performing ligand-aware protein sequence-structure co-design, providing a complementary token-space alternative to continuous diffusion-based protein design.

Molecular representation learning incorporates chemical, geometric, and language information through different architectures (Zhou et al., 2023; Wang et al., 2026; Liu et al., 2023; 2025; Luo et al., 2025). We use Uni-Mol (Zhou et al., 2023), which encodes ligand atom types and 3D coordinates into contextual atom and pair representations, to condition protein denoising through cross-attention.

2.2 Masked Discrete Diffusion Language Models

Discrete diffusion protein language models typically use absorbing-state corruption, where clean tokens are progressively replaced by `[MASK]`. The DPLM series (Wang et al., 2024; 2025b; Hsieh et al., 2025) adopts a reparameterized masked diffusion view (Zheng et al., 2024), treating generation as a route-and-denoise process that separates clean-token prediction from token selection.

In this work, we follow the masked discrete diffusion language model (MDLM) formulation (Sahoo et al., 2024; Shi et al., 2024). Given clean tokens x_0, the forward process independently preserves each token with probability α_t and replaces it with the mask token m with probability $1-\alpha_t$. The joint corruption distribution is $q(x_t \mid x_0) = \prod_i \mathrm{Cat}\Big(x_t^{(i)}; \alpha_t e_{x_0^{(i)}} + (1-\alpha_t)e_m\Big)$, where e_v is the one-hot vector for token v and α_t decreases from clean to masked states. The denoising model $\mu_\theta(x_t, t)$ predicts the clean-token distribution at masked positions, and the training objective reduces to a weighted masked cross-entropy loss:

$$\mathcal{L} = \int_0^1 w(t)\, \mathbb{E}_{q(x_t|x_0)} \left[\sum_{i: x_t^{(i)}=m} -\log \mu_\theta^{(i)}(x_t, t)_{x_0^{(i)}} \right] dt, \tag{1}$$

where $w(t) = -\alpha_t'/(1-\alpha_t)$.

Generation reverses the masking process from an all-masked or partially masked sequence. At each reverse step from t to $s < t$, unmasked tokens are copied unchanged, while each masked position is sampled as

$$x_s^{(i)} \sim \mathrm{Cat}\left(\frac{\alpha_s - \alpha_t}{1-\alpha_t} \mu_\theta^{(i)}(x_t, t) + \frac{1-\alpha_s}{1-\alpha_t} e_m \right), \quad \text{if } x_t^{(i)} = m, \tag{2}$$

where e_m denotes the one-hot mask token. This transition either reveals a token according to the denoising distribution or keeps it masked for later refinement.

2.3 Unmasking Strategies

In masked discrete diffusion models, token reveal order is a key factor in generation quality. While vanilla MDLM sampling unmasks tokens in an order-agnostic reverse process (Sahoo et al., 2024; Shi et al., 2024), adaptive strategies improve decoding by prioritizing high-confidence positions, such as those with large top-K confidence or top-1/top-2 probability margins (Kim et al., 2025). LLaDA (Nie et al., 2025) selects proposals by confidence at currently masked positions while preserving previously committed tokens. ReMDM (Wang et al., 2025a) introduces a reverse process that can remask previously decoded tokens. Building on these approaches, MCM-ReMask combines MDLM-based stochastic candidate selection with confidence-margin ranking. It ranks only newly proposed non-mask positions and commits their argmax predictions, while a retained-fraction parameter controls additional deferral.

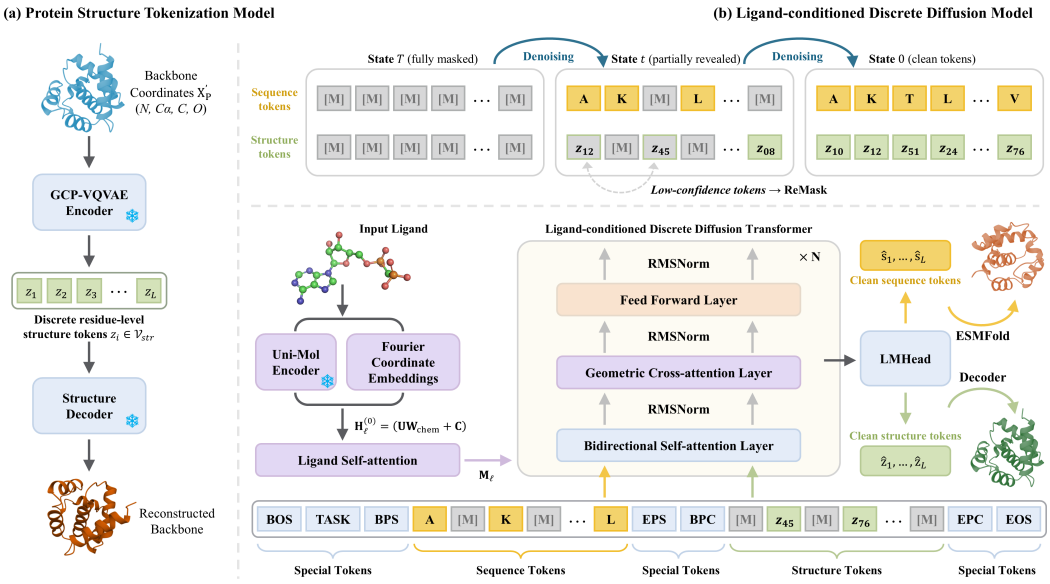


Figure 1: Overview of the proposed ProtLiD^2 model. (a) A frozen GCP-VQVAE tokenizer converts protein backbone coordinates into residue-level structure tokens. (b) ProtLiD^2 jointly denoises sequence and structure tokens with a ligand-conditioned masked discrete diffusion Transformer. Ligand chemical and geometric features are injected through geometry-aware cross-attention, and MCM-ReMask retains confident predictions while remasking uncertain tokens for refinement.

3 METHOD

An overview of the proposed architecture is shown in Fig. 1. ProtLiD^2 represents proteins as paired amino-acid sequence tokens and discrete structure tokens obtained from a frozen GCP-VQVAE tokenizer. Ligand atom features from Uni-Mol are fused with Fourier coordinate embeddings, refined by ligand self-attention, and supplied to a masked discrete diffusion Transformer through geometry-aware cross-attention. During inference, the model iteratively denoises sequence and structure tokens, while MCM-ReMask retains updates at high-confidence positions and remasks uncertain proposals for subsequent refinement. The generated structure tokens are decoded into a protein backbone, and ESMFold independently refolds the generated sequence to evaluate sequence-structure consistency.

3.1 DATASET CONSTRUCTION

To train ProtLiD^2, we constructed a large-scale ligand-conditioned protein sequence–structure dataset by integrating protein–ligand complexes from Protenix (Zhang et al., 2026), PLINDER (Durairaj et al., 2024), CrossDock (Francoeur et al., 2020), HiQBind (Wang et al., 2025c), and AlphaFill-derived complexes (Hekkelman et al., 2023). Complexes from non-Protenix sources were processed with the AlphaFold3 data-processing pipeline (Abramson et al., 2024) to obtain a unified representation of protein chains, ligand identities, ligand coordinates, and protein backbone geometry.

For each complex, we extracted one protein chain and its associated ligand, retaining examples with valid protein coordinates, ligand SMILES, and at least one ligand-contacting residue within a 6.0 Å cutoff. We further removed long proteins, large ligands, severe protein-ligand clashes, and low-confidence AlphaFill-derived complexes. After filtering and source-specific sampling, the merged dataset contained 1,125,038 ligand-protein complexes. To reduce redundancy, samples were indexed by protein sequence so that each unique sequence could be associated with one or more complexes. Detailed filtering criteria and source-specific statistics are provided in Appendix A.

For leakage-aware evaluation, every candidate training protein is compared against the complete PLINDER test set, before selecting the evaluation subset. We removed training examples with $\geq 30\%$ sequence identity to any PLINDER test protein using MMseqs2 (Steinegger & Söding, 2017),

yielding 1,026,766 training complexes. Due to the cost of full test set evaluation, we sampled 200 protein-ligand complexes with an approximately uniform sequence-length distribution as the final benchmark set.

3.2 LIGAND-CONDITIONED SEQUENCE-STRUCTURE CO-DESIGN MODEL

ProtLiD^2 formulates ligand-conditioned protein design as masked discrete diffusion over a unified sequence-structure token space. Given a protein-ligand complex, the protein is represented by paired amino-acid sequence and discrete structure tokens, while the ligand is encoded from atom-level chemical features and 3D coordinates. The model learns to recover clean protein tokens from `[MASK]`-corrupted inputs under ligand conditioning.

Protein and ligand preprocessing. For each protein-ligand complex, the protein is represented by its amino-acid sequence $\mathbf{a} = (a_1, \ldots, a_L)$ and backbone coordinates $\mathbf{X}_{\mathrm{p}} \in \mathbb{R}^{L\times 4\times 3}$, while the ligand is represented by heavy-atom coordinates $\mathbf{X}_\ell \in \mathbb{R}^{M\times 3}$ and atom-level features. We center each complex at the protein Cα centroid and apply a shared ligand-guided PCA rotation, preserving relative protein-ligand geometry. Together with geometric augmentation, this canonicalization aims to reduce sensitivity to the global coordinate frame, addressing a central objective of SE(3)-equivariant modeling. The rotation-sensitivity experiments in Section 4.5 indicating empirical rotational robustness without explicitly enforcing SE(3) equivariance. Preprocessing details appear in Appendix B. **Protein sequence and structure tokenization.** The amino-acid sequence is tokenized as $\mathbf{s} = (s_1, \ldots, s_L)$, where $s_i \in \mathcal{V}_{\mathrm{seq}}$. As illustrated in Fig. 1(a), a frozen GCP-VQVAE tokenizer maps the canonicalized backbone coordinates to residue-level discrete structure tokens $\mathbf{z} = (z_1, \ldots, z_L)$, where $z_i \in \mathcal{V}_{\mathrm{str}}$. Sequence and structure tokens are concatenated into a unified multimodal token sequence, $\mathbf{y}_0 = [\texttt{BOS}, \texttt{TASK}, \texttt{BPS}, s_1, \ldots, s_L, \texttt{EPS}, \texttt{BPC}, z_1, \ldots, z_L, \texttt{EPC}, \texttt{EOS}]$, where special tokens mark the sequence and structure spans. Structure-token indices are shifted so that both modalities are represented in a shared vocabulary while retaining modality-specific validity constraints. **Ligand representation.** As shown in Fig. 1(b), the ligand condition is encoded from both chemical and geometric information. We use Uni-Mol Zhou et al. (2023) to extract atom-level and pairwise atom–atom features, and encode ligand coordinates using Fourier coordinate embeddings. The projected chemical features and coordinate embeddings are fused and refined by stacked pairwise-aware ligand self-attention layers, producing the final ligand memory $\mathbf{M}_\ell$. Details of the ligand embedding and pair-bias attention are given in Appendix B. **Geometry-aware ligand cross-attention.** To inject ligand information into the protein denoising backbone, we insert a geometry-aware ligand cross-attention layer after each protein self-attention block. Protein hidden states are used as queries, while the ligand memory $\mathbf{M}_\ell$ provides keys and values. In addition to standard attention logits, we add a learned geometric bias derived from pairwise distances between protein-token proxy coordinates and ligand atom coordinates. This allows sequence and structure tokens to attend to chemically encoded ligand atoms while emphasizing spatially relevant protein–ligand interactions. The full formulation is provided in Appendix B.

Ligand-conditioned discrete diffusion transformer. The denoising backbone follows the architecture shown in Fig. 1(b). At diffusion time t, the clean sequence-structure token sequence $\mathbf{y}_0$ is corrupted into $\mathbf{y}_t$ by an absorbing-state masking process. The initial hidden states are obtained from the noisy tokens as $\mathbf{H}^{(0)} = \mathrm{Embed}(\mathbf{y}_t)$. These hidden states are processed by stacked Transformer blocks with bidirectional self-attention, ligand-conditioned geometric cross-attention, and feed-forward layers. After N layers, the final hidden states are passed through two fully connected layers, denoted as LMHead, to produce logits over the joint sequence-structure vocabulary: $\boldsymbol{\ell}_\theta(\mathbf{y}_t, \mathbf{M}_\ell) = \mathrm{LMHead}(\mathbf{H}^{(N)})$. The corresponding denoising distribution is $\mu_\theta(\mathbf{y}_t, \mathbf{M}_\ell) = \mathrm{Softmax}(\boldsymbol{\ell}_\theta(\mathbf{y}_t, \mathbf{M}_\ell))$, which predicts the clean sequence and structure tokens from the corrupted input.

Following the MDLM objective in Eq. (1), the model is trained with weighted masked cross-entropy over the joint sequence-structure token sequence. During training, only masked non-special tokens contribute to the loss, and modality-specific vocabulary constraints are used so that sequence positions are predicted over $\mathcal{V}_{\mathrm{seq}}$ and structure-token positions are predicted over $\mathcal{V}_{\mathrm{str}}$. During generation, tokens are progressively sampled from the reverse transition defined in Eq. (2), which either reveals a token according to the model predicted denoising distribution or keeps the position masked for later refinement.

3.3 Maximum Confidence-Margin Guided ReMask Decoding Strategy

We implement the MCM-ReMask decoding strategy by applying confidence-margin filtering to MDLM token proposals. At each reverse step, the strategy retains high-margin non-mask proposals and remasks the remaining proposals, with a retained-fraction parameter controlling additional deferral, and the detail of the procedure is shown in Algorithm 1 in Appendix B.4.

At reverse step j, let T_j denote the currently masked positions whose sampled MDLM proposals are non-mask tokens. The number of retained proposals is

$$k_j = \lfloor r|T_j| \rfloor, \qquad 0 < r \leq 1. \tag{3}$$

For each eligible position, the confidence margin is the difference between the two largest probabilities obtained from modality-constrained logits. The strategy retains the sampled token values at the k_j highest-margin positions in T_j and returns the remaining proposals to `[MASK]`. Previously visible tokens remain unchanged. Smaller r defers more proposals.

4 Experiment

In this section, we evaluate decoding strategies, whole-protein and pocket co-design, and component ablations and sensitivity analyses.

4.1 Experimental Setup

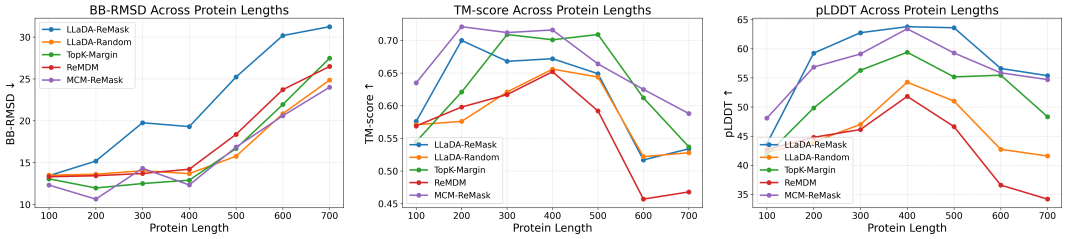
ProtLiD^2 is implemented as a Transformer-based model (Vaswani et al., 2017) with approximately 370M parameters. It contains 16 Transformer layers with hidden dimension 1280, configured feed-forward dimension 5120, and 10 query heads with 2 key/value heads. The model was trained with PyTorch on 8 NVIDIA RTX Pro 6000 GPUs, each with 96GB memory, for 100,000 optimization steps over approximately 11 days. Training used bfloat16 precision, gradient accumulation over 8 steps, with a maximum of 45,000 tokens per batch and maximum protein length of 1024 residues. We optimized the model with AdamW using learning rate 6×10^{-4}, $\beta = (0.9, 0.95)$, weight decay 0.1, gradient clipping at 1.0, and a cosine learning-rate schedule. Geometric data augmentation was applied using random rotations with probability 0.3 and coordinate noise with scale 0.07.

We compare MCM-ReMask with ReMDM (Wang et al., 2025a), LLaDA-ReMask (Nie et al., 2025), and TopK-Margin (Kim et al., 2025) using the sampled ligands, and 500 reverse steps. Each policy generates 100 sequence-structure pairs at each length from 100 to 700 residues. Comparing GCP-VQVAE-decoded backbones with ESMFold (Lin et al., 2023) refoldings measures self-consistency through TM-score (Zhang & Skolnick, 2004) and BB/CA-RMSD (Maiorov & Crippen, 1994); pLDDT (Jumper et al., 2021) measures prediction confidence. Appendix C.1 gives detailed results.

For ligand-conditioned whole-protein and pocket co-design evaluations, each method generates 10 candidates per target under the same ligand condition. Whole-protein co-design additionally conditions on the target protein length, whereas pocket co-design fixes the non-pocket context and redesigns ligand-contacting residues within a 6.0 Å protein-ligand heavy-atom cutoff. We evaluate sequence-structure self-consistency by folding generated sequences with ESMFold and comparing them with model-generated structures using BB/CA-RMSD, TM-score, and pLDDT. For ligand-aware evaluation, we select the highest-pLDDT candidate per target, predict the protein-ligand complex with AlphaFold3, and compute an AF3-Vina score using AutoDock Vina (Eberhardt et al., 2021) with a ligand-centered docking box. We also report combined pass-rate criteria integrating fold confidence, structural accuracy, and ligand-aware docking quality, with full definitions in Appendix C.2 and C.3.

4.2 Controlled Decoding-Strategy Ablation

Under the shared-checkpoint protocol, MCM-ReMask has the highest mean TM-score at six of seven lengths and the lowest mean BB/CA-RMSD at five (Fig. 2; Table 7). We use it for subsequent comparisons because of its overall self-consistency performance. We set the retention fraction to 1.0, and present a sensitivity analysis of this hyperparameter in Section 4.5.



(a) BB-RMSD (b) TM-score (c) pLDDT

Figure 2: Comparison of different unmasking strategies across protein lengths.

Table 1: Comparison between ProtLiD^2 and Complexa. Structure metrics are evaluated on 200 targets, while AF3-Vina scores are evaluated on 191 valid targets.

Method	Global RMSD ↓		TM-score ↑ ($n = 200$)	pLDDT ↑ ($n = 200$)	AF3-Vina ↓ ($n = 191$)
	BB ($n = 200$)	CA ($n = 200$)			
Complexa	$\mathbf{10.35 \pm 11.24}$	$\mathbf{10.40 \pm 11.27}$	0.672 ± 0.333	64.55 ± 18.10	$\mathbf{-7.11 \pm 2.26}$
ProtLiD^2	12.07 ± 13.16	12.13 ± 13.15	$\mathbf{0.802 \pm 0.175}$	$\mathbf{73.00 \pm 12.85}$	-6.82 ± 1.75

4.3 Ligand-Conditioned Whole Protein Co-design

We next evaluate ProtLiD^2 on ligand-conditioned whole-protein co-design. Existing discrete diffusion protein language models such as DPLM do not directly support small-molecule ligand conditioning, and are therefore not directly comparable in this setting. We instead compare ProtLiD^2 with Complexa (Didi et al., 2026), using its ligand-conditioned small-molecule setting.

The comparison is conducted on the 200-target benchmark described in Section 3.1. Following the protocol in Section 4.1, both methods are conditioned on the same ligand and target protein length for each target. Structure metrics are evaluated on all 200 targets, while AF3-Vina scores are reported on 191 targets because 9 ligands could not be converted into valid AutoDock Vina-compatible representations. Full evaluation details are provided in Appendix C.2.

As shown in Table 1, Complexa achieves lower global RMSD and slightly better AF3-Vina scores, whereas ProtLiD^2 yields higher TM-score and pLDDT, indicating stronger fold consistency and predicted structural confidence. In the combined pass-rate comparison for in silico design (Table 2), ProtLiD^2 improves FC from 51.00% to 58.50%, BC-5 from 47.12% to 52.88%, and BC-7 from 29.32% to 30.37%, while also achieving a higher SWPS pass rate. These results suggest that ProtLiD^2 is competitive for ligand-conditioned whole-protein design when global fold consistency, sequence–structure compatibility, and ligand-aware evaluation are considered jointly.

4.4 Pocket Co-design

We further evaluate ProtLiD^2 on ligand-binding pocket co-design, following the pocket definition and evaluation protocol in Section 4.1. Starting from the 200-target benchmark described in Section 3.1, we remove 50 multichain complexes to focus on single-chain pocket design, resulting in 150 valid targets. Vina-related metrics and pass rates are computed on 149 targets because one target could not be processed by AutoDock Vina. Full evaluation details are provided in Appendix C.3.

Table 2: Computational composite pass-rate comparison between methods. For each criterion, we report the number of passed designs and pass rate. Full criterion definitions are provided in Appendix C.2.

Method	FC		HCF		BC-5		BC-7		SWPS	
	Count	Rate	Count	Rate	Count	Rate	Count	Rate	Count	Rate
Complexa	102/200	51.00	13/200	6.50	90/191	47.12	56/191	29.32	8/191	4.19
ProtLiD^2	**117/200**	**58.50**	13/200	6.50	**101/191**	**52.88**	**58/191**	**30.37**	**11/191**	**5.76**

Table 3: Overall comparison of pocket-design methods. Values are reported as mean ± standard deviation. Lower is better for RMSD and Vina score; higher is better for TM-score and pLDDT.

Method	Active-site RMSD ↓		Global RMSD ↓		TM-score ↑	pLDDT ↑	Vina ↓
	BB	CA	BB	CA			
FAIR	3.46 ± 2.62	3.37 ± 2.69	3.78 ± 4.75	3.81 ± 4.77	0.866 ± 0.194	79.83 ± 11.31	-6.94 ± 1.74
PocketGen	3.40 ± 2.54	3.50 ± 2.55	3.70 ± 4.67	3.74 ± 4.68	0.869 ± 0.192	$\mathbf{80.83 \pm 11.59}$	$\mathbf{-8.84 \pm 3.80}$
ProtLiD^2	$\mathbf{1.97 \pm 1.69}$	$\mathbf{2.06 \pm 1.72}$	$\mathbf{3.63 \pm 4.45}$	$\mathbf{3.69 \pm 4.46}$	$\mathbf{0.915 \pm 0.127}$	79.17 ± 11.49	-6.93 ± 1.48

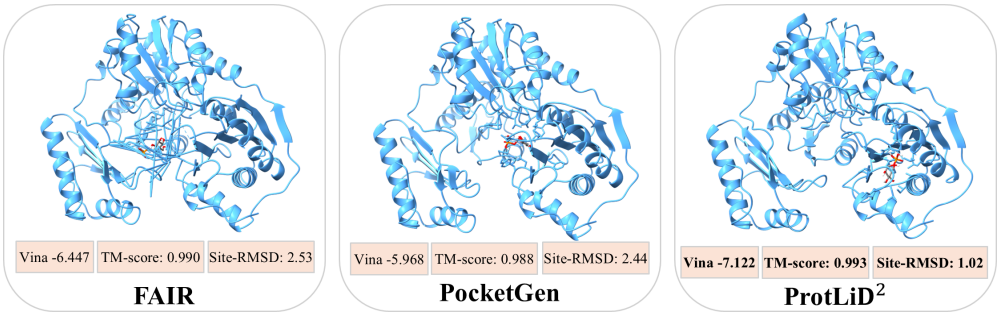


Figure 3: Qualitative pocket co-design case study on 3BKQ. ProtLiD^2 has lower Site-RMSD in this computational comparison; the illustration does not establish ligand binding.

Table 4: Computational composite pass-rate comparison among FAIR, PocketGen, and ProtLiD^2. For each criterion, we report the number of passed designs and pass rate. Full criterion definitions are provided in Appendix C.3.

Method	FC		HCF		PGC		BC-5		BC-7		SDS	
	Count	Rate	Count	Rate	Count	Rate	Count	Rate	Count	Rate	Count	Rate
FAIR	**127/149**	**85.23**	68/149	45.64	15/149	10.07	10/149	6.71	2/149	1.34	0/149	0.00
PocketGen	126/149	84.56	**70/149**	**46.98**	30/149	20.13	22/148	14.86	9/148	6.08	0/148	0.00
ProtLiD^2	**127/149**	**85.23**	63/149	42.28	**96/149**	**64.43**	**89/149**	**59.73**	**35/149**	**23.49**	**6/149**	**4.03**

Table 5: Component-removal ablations. Means over 600 designs per setting; RMSD in Å.

Setting	CA-RMSD ↓	BB-RMSD ↓	TM ↑	pLDDT ↑
Full model	**17.61**	**17.52**	**0.536**	**40.61**
No geometric bias	18.20	18.10	0.520	39.29
No Uni-Mol pair bias	17.54	17.44	0.533	40.29
No ligand conditioning	23.69	23.59	0.385	26.03

As shown in Table 3, ProtLiD^2 achieves the lowest active-site BB/CA-RMSD (1.97/2.06 Å), compared with FAIR (3.46/3.37 Å) and PocketGen (3.40/3.50 Å), alongside the lowest global RMSD and highest TM-score. Although PocketGen obtains the best average Vina score and slightly higher pLDDT, ProtLiD^2 provides more accurate pocket geometry and stronger global fold consistency. The representative case in Fig. 3 illustrates these improvements, with additional examples provided in Appendix Fig. 6. In the combined pass-rate comparison for in silico design (Table 4), ProtLiD^2 matches the best FC pass rate and improves PGC from 20.13% to 64.43%, BC-5 from 14.86% to 59.73%, and BC-7 from 6.08% to 23.49% relative to PocketGen. It is also the only evaluated method to achieve a nonzero SDS pass rate. These results support its effectiveness for pocket co-design under joint structural and ligand-aware criteria.

4.5 Ablation Studies

Using the 370M ProtLiD^2 model, we examine ligand conditioning, decoding retention, coordinate-frame sensitivity, and screening thresholds. For experiments in this subsection, metrics are averaged over all generated proteins within each setting, without selecting the best-performing candidates.

Table 6: Retained-fraction sensitivity. Means over 600 designs per setting; RMSD in Å.

r	CA-RMSD ↓	BB-RMSD ↓	TM-score ↑	pLDDT ↑
0.5	17.20	17.12	0.656	**57.52**
0.7	16.18	16.10	0.653	55.92
0.9	16.24	16.16	0.651	56.04
1.0	**15.15**	**15.06**	**0.666**	55.90

Ligand conditioning. We ablate ligand information, geometric attention bias, and Uni-Mol pair bias to assess their contributions to generation quality. Removing ligand conditioning lowers TM-score from 0.536 to 0.385 and pLDDT from 40.61 to 26.03 (Table 5; Appendix D.1). Removing the explicit geometric attention bias produces smaller decreases, to 0.520 and 39.29, while removing Uni-Mol pair bias has little effect. These results indicate a strong reliance on ligand conditioning for sequence–structure consistency and prediction confidence, with a smaller contribution from explicit geometric bias.

Retained fraction and decoding cost. We vary r to assess how proposal deferral affects structural self-consistency and prediction confidence. Across $r \in \{0.5, 0.7, 0.9, 1.0\}$, TM-score spans 0.015 and pLDDT spans 1.62, whereas CA/BB-RMSD span 2.05/2.06 Å (Table 6; Appendix D.2). The $r = 1$ setting achieves the lowest RMSDs and highest TM-score, while $r = 0.5$ yields the highest pLDDT. These results suggest a trade-off between structural self-consistency and prediction confidence, favoring $r = 1$ for backbone agreement.

Coordinate-frame sensitivity. To test rotation robustness under PCA canonicalization and geometric augmentation, we generate six predictions per orientation for 100 stratified targets, using the original orientation and three global rotations and recomputing PCA each time. The mean changes in BB-RMSD and TM-score are $+0.26$ Å (paired 95% CI $[-0.14, 0.66]$) and $+0.001$ ($[-0.009, 0.012]$), respectively. Pocket-design evaluation likewise shows no average deterioration, with TM-score and pLDDT changes of $+0.016$ and $+1.22$. These results support stable average generation quality under the tested rotations without explicit architectural equivariance. Appendix D.3 details the results, including the six low-eigengap targets.

Screening thresholds. To test the sensitivity of combined pass-rate comparisons to docking thresholds, we vary the Vina cutoff from -9 to -4. ProtLiD^2 achieves higher pocket-design pass rates than FAIR and PocketGen at every tested cutoff. For whole-protein design, ProtLiD^2 and Complexa achieve broadly similar pass rates, with their relative ranking changing at stricter cutoffs. These results support the robustness of the pocket-design comparison to threshold choice (Appendix D.4).

5 Discussion

ProtLiD^2 demonstrates that ligand conditioning can be integrated into masked discrete diffusion over unified sequence-structure tokens. The proposed MCM-ReMask decoding first improves sequence-structure self-consistency through lightweight inference-time self-correction. Building on this decoding strategy, ProtLiD^2 improves whole-protein TM-score and pLDDT over Complexa, and substantially reduces active-site RMSD while increasing ligand-aware pass rates over FAIR and PocketGen in pocket co-design. These results suggest that ProtLiD^2 combines robust token-space generation with geometry-aware ligand conditioning for functional protein co-design.

Several limitations and broader-impact considerations remain. ProtLiD^2 relies on a frozen backbone tokenizer, lacks explicit full-atom side-chain and ligand-flexibility modeling, and is evaluated mainly with computational proxies such as ESMFold, AlphaFold3, and AutoDock Vina; thus, generated proteins require experimental validation. While ProtLiD^2 may accelerate ligand-aware protein and enzyme design, generative protein design also carries dual-use risks and should be accompanied by expert review, biosafety screening, and safeguards for future releases. To support reproducibility, we will release the inference and evaluation code after refactoring, together with the training and validation datasets and reproduction instructions. Future work will explore improved tokenization, full-atom refinement, stronger ligand-aware objectives, and experimental validation.

AI USE STATEMENT

In this work, we used generative AI tools to assist with manuscript writing and editing and to support code debugging. We have reviewed all AI-assisted text and code and made corrections where necessary. We take responsibility for the final content of this work, including its text, claims, and code produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Dataset filtering and protein-sequence de-overlap are described in Section 3.1 and Appendix A; architecture and decoding are specified in Section 3 and Appendix B. Section 4.1 and the experimental appendices define the evaluation procedures. Section 4.5 summarizes component removals, retained-fraction sensitivity, rotation tests, and screening-threshold sensitivity; Appendix D gives detailed protocols and additional results.

REFERENCES

Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J. Ballard, Joshua Bambrick, Sebastian W. Bodenstein, David A. Evans, Chia-Chun Hung, Michael O'Neill, David Reiman, Kathryn Tunyasuvunakool, Zachary Wu, Akvilė Žemgulytė, Eirini Arvaniti, Charles Beattie, Ottavia Bertolli, Alex Bridgland, Alexey Cherepanov, Miles Congreve, Alexander I. Cowen-Rivers, Andrew Cowie, Michael Figurnov, Fabian B. Fuchs, Hannah Gladman, Rishub Jain, Yousuf A. Khan, Caroline M. R. Low, Kuba Perlin, Anna Potapenko, Pascal Savy, Sukhdeep Singh, Adrian Stecula, Ashok Thillaisundaram, Catherine Tong, Sergei Yakneen, Ellen D. Zhong, Michal Zielinski, Augustin Žídek, Victor Bapst, Pushmeet Kohli, Max Jaderberg, Demis Hassabis, and John M. Jumper. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, 630(8016):493–500, June 2024. ISSN 1476-4687. doi: 10.1038/s41586-024-07487-w. URL https://doi.org/10.1038/s41586-024-07487-w.

Woody Ahern, Jason Yim, Doug Tischer, Saman Salike, Seth M. Woodbury, Donghyo Kim, Indrek Kalvet, Yakov Kipnis, Brian Coventry, Han Raut Altae-Tran, Magnus S. Bauer, Regina Barzilay, Tommi S. Jaakkola, Rohith Krishna, and David Baker. Atom-level enzyme active site scaffolding using RFdiffusion2. *Nature Methods*, 23(1):96–105, January 2026. ISSN 1548-7091, 1548-7105. doi: 10.1038/s41592-025-02975-x.

Sarah Alamdari, Nitya Thakkar, Rianne van den Berg, Neil Tenenholtz, Robert Strome, Alan M. Moses, Alex X. Lu, Nicolò Fusi, Ava P. Amini, and Kevin K. Yang. Protein generation with evolutionary diffusion: sequence is all you need. *bioRxiv*, 2024. doi: 10.1101/2023.09.11.556673. URL https://www.biorxiv.org/content/early/2024/11/04/2023.09.11.556673.

Marianne Arriola, Subham Sekhar Sahoo, Aaron Gokaslan, Zhihan Yang, Zhixuan Qi, Jiaqi Han, Justin T Chiu, and Volodymyr Kuleshov. Block diffusion: Interpolating between autoregressive and diffusion language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=tyEyYT267x.

Jasper Butcher, Rohith Krishna, Raktim Mitra, Rafael I. Brent, Yanjing Li, Nathaniel Corley, Paul Kim, Jonathan Funk, Simon Mathis, Saman Salike, Aiko Muraishi, Helen Eisenach, Tuscan Rock Thompson, Jie Chen, Yuliya Politanska, Enisha Sehgal, Brian Coventry, Odin Zhang, Bo Qiang, Kieran Didi, Max Kazman, Frank DiMaio, and David Baker. De novo design of all-atom biomolecular interactions with rfdiffusion3. *bioRxiv*, 2025. doi: 10.1101/2025.09.18.676967. URL https://www.biorxiv.org/content/early/2025/09/18/2025.09.18.676967.

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.

Kieran Didi, Zuobai Zhang, Guoqing Zhou, Danny Reidenbach, Zhonglin Cao, Sooyoung Cha, Tomas Geffner, Christian Dallago, Jian Tang, Michael M. Bronstein, Martin Steinegger, Emine Kucukbenli, Arash Vahdat, and Karsten Kreis. Scaling atomistic protein binder design with generative pretraining and test-time compute. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=qmCpJtFZra.

Janani Durairaj, Yusuf Adeshina, Zhonglin Cao, Xuejin Zhang, Vladas Oleinikovas, Thomas Duignan, Zachary McClure, Xavier Robin, Gabriel Studer, Daniel Kovtun, Emanuele Rossi, Guoqing Zhou, Srimukh Veccham, Clemens Isert, Yuxing Peng, Prabindh Sundareson, Mehmet Akdel, Gabriele Corso, Hannes Stärk, Gerardo Tauriello, Zachary Carpenter, Michael Bronstein, Emine Kucukbenli, Torsten Schwede, and Luca Naef. PLINDER: The protein-ligand interactions dataset and evaluation resource. *bioRxiv*, 2024. doi: 10.1101/2024.07.17.603955. URL https://www.biorxiv.org/content/10.1101/2024.07.17.603955v3.

Jerome Eberhardt, Diogo Santos-Martins, Andreas F. Tillack, and Stefano Forli. Autodock vina 1.2.0: New docking methods, expanded force field, and python bindings. *Journal of Chemical Information and Modeling*, 61(8):3891–3898, 2021. doi: 10.1021/acs.jcim.1c00203. URL https://doi.org/10.1021/acs.jcim.1c00203. PMID: 34278794.

Paul G. Francoeur, Tomohide Masuda, Jocelyn Sunseri, Andrew Jia, Richard B. Iovanisci, Ian Snyder, and David R. Koes. Three-dimensional convolutional neural networks and a cross-docked data set for structure-based drug design. *Journal of Chemical Information and Modeling*, 60(9): 4200–4215, 2020. doi: 10.1021/acs.jcim.0c00411. URL https://doi.org/10.1021/acs.jcim.0c00411. PMID: 32865404.

Tomas Geffner, Kieran Didi, Zhonglin Cao, Danny Reidenbach, Zuobai Zhang, Christian Dallago, Emine Kucukbenli, Karsten Kreis, and Arash Vahdat. La-Proteina: Atomistic Protein Generation via Partially Latent Flow Matching, July 2025.

Maarten L. Hekkelman, Ida de Vries, Robbie P. Joosten, and Anastassis Perrakis. AlphaFill: enriching alphafold models with ligands and cofactors. *Nature Methods*, 20(2):205–213, February 2023. ISSN 1548-7105. doi: 10.1038/s41592-022-01685-y. URL https://doi.org/10.1038/s41592-022-01685-y.

Cheng-Yen Hsieh, Xinyou Wang, Daiheng Zhang, Dongyu Xue, Fei Ye, Shujian Huang, Zaixiang Zheng, and Quanquan Gu. Elucidating the design space of multimodal protein language models. In *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, Proceedings of Machine Learning Research. PMLR / OpenReview.net, 2025. URL https://proceedings.mlr.press/v267/hsieh25a.html.

Bowen Jing, Stephan Eismann, Patricia Suriana, Raphael John Lamarre Townshend, and Ron O. Dror. Learning from protein structure with geometric vector perceptrons. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=1YLJDvSx6J4.

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, August 2021. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-021-03819-2.

Jaeyeon Kim, Kulin Shah, Vasilis Kontonis, Sham M. Kakade, and Sitan Chen. Train for the worst, plan for the best: Understanding token ordering in masked diffusions. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, Proceedings of Machine Learning Research. PMLR / OpenReview.net, 2025. URL https://proceedings.mlr.press/v267/kim25ah.html.

Tanja Kortemme and David Baker. Computational design of protein–protein interactions. *Current Opinion in Chemical Biology*, 8(1):91–97, 2004. ISSN 1367-5931. doi: https://doi.org/10.1016/j.cbpa.2003.12.008.

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, 2023. doi: 10.1126/science.ade2574. URL https://www.science.org/doi/abs/10.1126/science.ade2574.

Zhiyuan Liu, Sihang Li, Yanchen Luo, Hao Fei, Yixin Cao, Kenji Kawaguchi, Xiang Wang, and Tat-Seng Chua. Molca: Molecular graph-language modeling with cross-modal projector and uni-modal adapter. In *EMNLP*, 2023. URL https://openreview.net/forum?id=14WRhMNq7H.

Zhiyuan Liu, Yanchen Luo, Han Huang, Enzhi Zhang, Sihang Li, Junfeng Fang, Yaorui Shi, Xiang Wang, Kenji Kawaguchi, and Tat-Seng Chua. NEXT-MOL: 3d diffusion meets 1d language modeling for 3d molecule generation. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=p66a00KLWN.

Yanchen Luo, Zhiyuan Liu, Yi Zhao, Sihang Li, Hengxing Cai, Kenji Kawaguchi, Tat-Seng Chua, Yang Zhang, and Xiang Wang. Towards unified and lossless latent space for 3d molecular latent diffusion modeling. *arXiv preprint arXiv:2503.15567*, 2025.

Vladimir N. Maiorov and Gordon M. Crippen. Significance of root-mean-square deviation in comparing three-dimensional structures of globular proteins. *Journal of Molecular Biology*, 235 (2):625–634, 1994. ISSN 0022-2836. doi: https://doi.org/10.1006/jmbi.1994.1017. URL https://www.sciencedirect.com/science/article/pii/S0022283684710175.

Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. Large language diffusion models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=KnqiC0znVF.

Robin Pearce, Xiaoqiang Huang, Dani Setiawan, and Yang Zhang. EvoDesign: Designing Protein–Protein Binding Interactions Using Evolutionary Interface Profiles in Conjunction with an Optimized Physical Energy Function. *Journal of Molecular Biology*, 431(13):2467–2476, June 2019. ISSN 00222836. doi: 10.1016/j.jmb.2019.02.028.

Mahdi Pourmirzaei, Alex Morehead, Farzaneh Esmaili, Jarett Ren, Mohammadreza Pourmirzaei, and Dong Xu. GCP-VQVAE: A geometry-complete language for protein 3d structure. *bioRxiv*, 2025. doi: 10.1101/2025.10.01.679833. URL https://www.biorxiv.org/content/early/2025/10/03/2025.10.01.679833.

Subham S. Sahoo, Marianne Arriola, Yair Schiff, Aaron Gokaslan, Edgar Marroquin, Justin T. Chiu, Alexander Rush, and Volodymyr Kuleshov. Simple and effective masked diffusion language models. In Amir Globerson, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/eb0b13cc515724ab8015bc978fdde0ad-Abstract-Conference.html.

Jiaxin Shi, Kehang Han, Zhe Wang, Arnaud Doucet, and Michalis K. Titsias. Simplified and generalized masked diffusion for discrete data. In Amir Globerson, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/bad233b9849f019aead5e5cc60cef70f-Abstract-Conference.html.

Martin Steinegger and Johannes Söding. Mmseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology*, 35(11):1026–1028, November 2017. ISSN 1546-1696. doi: 10.1038/nbt.3988. URL https://doi.org/10.1038/nbt.3988.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 5998–6008, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Guanghan Wang, Yair Schiff, Subham Sekhar Sahoo, and Volodymyr Kuleshov. Remasking discrete diffusion models with inference-time scaling. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025a. URL https://openreview.net/forum?id=IJryQAOy0p.

Lin Wang, Yifan Wu, Hao Luo, Minglong Liang, Yihang Zhou, Cheng Chen, Chris Liu, Jun Zhang, and Yang Zhang. Learned conformational space and pharmacophore into molecular foundational model. *Advanced Science*, 13(17):e13556, 2026. doi: https://doi.org/10.1002/advs.202513556. URL https://advanced.onlinelibrary.wiley.com/doi/abs/10.1002/advs.202513556.

Xinyou Wang, Zaixiang Zheng, Fei Ye, Dongyu Xue, Shujian Huang, and Quanquan Gu. Diffusion language models are versatile protein learners. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, Proceedings of Machine Learning Research, pp. 52309–52333. PMLR / OpenReview.net, 2024. URL https://proceedings.mlr.press/v235/wang24ct.html.

Xinyou Wang, Zaixiang Zheng, Fei Ye, Dongyu Xue, Shujian Huang, and Quanquan Gu. DPLM-2: A multimodal diffusion protein language model. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025b. URL https://openreview.net/forum?id=5z9GjHgerY.

Yingze Wang, Kunyang Sun, Jie Li, Xingyi Guan, Oufan Zhang, Dorian Bagni, Yang Zhang, Heather A. Carlson, and Teresa Head-Gordon. A workflow to create a high-quality protein–ligand binding dataset for training, validation, and prediction tasks. *Digital Discovery*, 4:1209–1220, 2025c. doi: 10.1039/D4DD00357H. URL http://dx.doi.org/10.1039/D4DD00357H.

Joseph L. Watson, David Juergens, Nathaniel R. Bennett, Brian L. Trippe, Jason Yim, Helen E. Eisenach, Woody Ahern, Andrew J. Borst, Robert J. Ragotte, Lukas F. Milles, Basile I. M. Wicky, Nikita Hanikel, Samuel J. Pellock, Alexis Courbet, William Sheffler, Jue Wang, Preetham Venkatesh, Isaac Sappington, Susana Vázquez Torres, Anna Lauko, Valentin De Bortoli, Emile Mathieu, Sergey Ovchinnikov, Regina Barzilay, Tommi S. Jaakkola, Frank DiMaio, Minkyung Baek, and David Baker. De novo design of protein structure and function with RFdiffusion. *Nature*, 620(7976):1089–1100, August 2023. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-023-06415-8.

Lijun Yu, José Lezama, Nitesh Bharadwaj Gundavarapu, Luca Versari, Kihyuk Sohn, David Minnen, Yong Cheng, Agrim Gupta, Xiuye Gu, Alexander G. Hauptmann, Boqing Gong, Ming-Hsuan Yang, Irfan Essa, David A. Ross, and Lu Jiang. Language model beats diffusion - tokenizer is key to visual generation. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=gzqrANCF4g.

Yang Zhang and Jeffrey Skolnick. Scoring function for automated assessment of protein structure template quality. *Proteins: Structure, Function, and Bioinformatics*, 57(4):702–710, 2004. doi: https://doi.org/10.1002/prot.20264. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/prot.20264.

Yuxuan Zhang, Chengyue Gong, Hanyu Zhang, Wenzhi Ma, Zhenyu Liu, Xinshi Chen, Jiaqi Guan, Lan Wang, Yanping Yang, Yu Xia, and Wenzhi Xiao. Protenix-v1: Toward high-accuracy open-source biomolecular structure prediction. *bioRxiv*, 2026. doi: 10.64898/2026.02.05.703733. URL https://www.biorxiv.org/content/early/2026/02/22/2026.02.05.703733.1.

Zaixi Zhang, Zepu Lu, Zhongkai Hao, Marinka Zitnik, and Qi Liu. Full-atom protein pocket design via iterative refinement. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/365a6f71486ecdfa7eb8d61cbe168782-Abstract-Conference.html.

Zaixi Zhang, Wan Xiang Shen, Qi Liu, and Marinka Zitnik. Efficient generation of protein pockets with PocketGen. *Nature Machine Intelligence*, 6(11):1382–1395, November 2024. ISSN 2522-5839. doi: 10.1038/s42256-024-00920-9.

Lin Zheng, Jianbo Yuan, Lei Yu, and Lingpeng Kong. A reparameterized discrete diffusion model for text generation. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=PEQFHRUFca.

Gengmo Zhou, Zhifeng Gao, Qiankun Ding, Hang Zheng, Hongteng Xu, Zhewei Wei, Linfeng Zhang, and Guolin Ke. Uni-mol: A universal 3d molecular representation learning framework. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=6K2RM6wVqKu.

A Dataset Processing Details

We integrated ligand-protein complexes from Protenix (Zhang et al., 2026), PLINDER (Durairaj et al., 2024), CrossDock (Francoeur et al., 2020), HiQBind (Wang et al., 2025c), and AlphaFill-derived complexes (Hekkelman et al., 2023). Ligand-protein complexes were first extracted from the Protenix training set. Complexes from the remaining sources were processed using the AlphaFold3 data-processing pipeline (Abramson et al., 2024), which provides unified parsing and cleanup of biomolecular complexes, including resolving alternative atom locations, removing waters, normalizing residue names, and expanding biological assemblies.

For each candidate complex, we extracted a single protein chain and its associated ligand. The protein component was represented by the amino-acid sequence and residue-wise coordinates of the four main-chain atoms N, C_α, C, and O. The ligand component was represented by atom coordinates, atom types, ligand identifiers, and SMILES strings. Complexes were retained only when both protein and ligand could be parsed successfully, the protein sequence was consistent with the coordinate-derived sequence, a valid ligand SMILES was available, and at least one ligand-contacting residue was identified within a 6.0 Å protein-ligand distance cutoff.

We removed proteins longer than 1000 residues and ligands containing more than 100 atoms. Complexes with severe protein-ligand steric clashes were discarded. A clash was defined by either an absolute interatomic distance below 0.8 Å or a van der Waals overlap criterion with a source-dependent tolerance of 0.6 Å. For AlphaFill-derived complexes, we retained only complexes with mean predicted confidence greater than 80.

After filtering and source-specific selection, the dataset contained 395,142 complexes from Protenix, 116,134 from CrossDock, 27,150 from HiQBind, and 115,610 from PLINDER. For AlphaFill-derived data, AlphaFold Database protein models are enriched with small molecules, cofactors, and ions transplanted from homologous experimentally determined structures. After filtering, this source yielded 5,281,501 processed complexes, from which we randomly sampled 587,136 complexes to balance the training data across sources. After removing duplicate protein entries across data sources, the final merged dataset comprised 1,125,038 ligand-protein complexes.

To prevent training-test leakage, we compared training protein sequences against PLINDER test proteins using MMseqs2 (Steinegger & Söding, 2017) and removed training examples with sequence identity $\geq 30\%$ to any benchmark protein. This de-overlap procedure produced the final training set of 1,026,766 ligand-protein complexes.

B Model Details

B.1 Coordinate Canonicalization

Given protein backbone coordinates $\mathbf{X}_{\mathrm{p}} \in \mathbb{R}^{L\times 4\times 3}$ and ligand heavy-atom coordinates $\mathbf{X}_\ell \in \mathbb{R}^{M\times 3}$, we canonicalize each protein-ligand complex to reduce global translational and rotational variation. We first translate the complex by the protein Cα centroid,

$$\mathbf{c}_{\mathrm{p}} = \frac{1}{L}\sum_{i=1}^{L} \mathbf{X}_{\mathrm{p}}^{(i,\mathrm{C}\alpha)}.$$

This yields centered coordinates $\widetilde{\mathbf{X}}_{\mathrm{p}} = \mathbf{X}_{\mathrm{p}} - \mathbf{c}_{\mathrm{p}}$ and $\widetilde{\mathbf{X}}_\ell = \mathbf{X}_\ell - \mathbf{c}_{\mathrm{p}}$. We then compute a deterministic ligand-guided PCA rotation matrix $\mathbf{R} \in \mathbb{R}^{3\times 3}$ from the centered ligand coordinates $\widetilde{\mathbf{X}}_\ell$ and apply the same rigid rotation to both the protein and ligand:

$$\mathbf{X}'_{\mathrm{p}} = \widetilde{\mathbf{X}}_{\mathrm{p}}\mathbf{R}, \qquad \mathbf{X}'_\ell = \widetilde{\mathbf{X}}_\ell\mathbf{R}.$$

This shared transformation preserves relative protein-ligand geometry. Selected degenerate geometries use an identity fallback; eigenvector signs are resolved deterministically and the frame is made right-handed ($\det R = +1$). Training includes PCA-axis sign changes, small rigid rotations, and coordinate perturbations. A protein-centered translation is defined for an observed protein-ligand complex. Appendix D.3 reports empirical sensitivity.

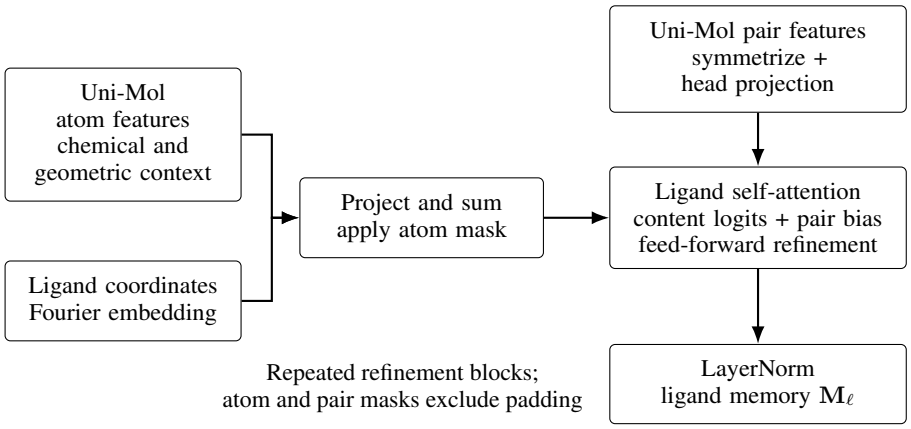


Figure 4: Ligand embedding module. Uni-Mol atom features and Fourier coordinate embeddings are fused to obtain initial ligand atom representations. Uni-Mol pair features and pair masks are then used by stacked pairwise refinement blocks to inject relative atom-atom geometry into ligand embeddings. The final normalized ligand embedding is used as the ligand memory for cross-attention.

B.2 Ligand Embedding Module

The ligand embedding module is shown in Fig. 4. For each ligand with M atoms, the input consists of Uni-Mol atom features, ligand coordinate features, atom masks, Uni-Mol pair features, and pair masks. Let $\mathbf{U} \in \mathbb{R}^{B\times M\times d_{\text{atom}}}$ denote Uni-Mol atom features and $\mathbf{P} \in \mathbb{R}^{B\times M\times M\times d_{\text{pair}}}$ denote Uni-Mol pair features. In our implementation, the Uni-Mol atom features have dimension $d_{\text{atom}} = 512$, while the pair features have dimension $d_{\text{pair}} = 64$.

The Uni-Mol atom features are first projected into the model hidden dimension by a chemical adapter,

$$\mathbf{F}_{\text{chem}} = \mathbf{U}\mathbf{W}_{\text{chem}}.$$

In parallel, the ligand coordinates are encoded by Fourier coordinate embeddings, producing coordinate features $\mathbf{F}_{\text{coord}} \in \mathbb{R}^{B\times M\times d}$. Invalid atoms are removed by the atom mask, and the initial ligand representation is obtained by feature fusion:

$$\mathbf{H}_\ell^{(0)} = (\mathbf{F}_{\text{chem}} + \mathbf{F}_{\text{coord}}) \odot \mathbf{m}_\ell,$$

where $\mathbf{m}_\ell \in \{0,1\}^{B\times M}$ is the ligand atom mask, broadcast along the hidden dimension.

To further refine ligand atom embeddings, we use a stack of pairwise-aware refinement blocks. These blocks inject Uni-Mol pairwise atom-atom information into atom representations through pair-biased self-attention. Specifically, the pair representation is symmetrized and projected into a head-wise attention bias:

$$\mathbf{B}_{\text{pair}} = \text{Proj}_{\text{pair}}\left(\frac{\mathbf{P} + \mathbf{P}^\top}{2}\right).$$

The ligand self-attention is then computed as

$$\text{Attn}_\ell = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_h}} + \alpha_{\text{pair}}\mathbf{B}_{\text{pair}} + \mathbf{B}_{\text{mask}}\right)\mathbf{V},$$

where α_{pair} is a learnable pair-bias scale, and $\mathbf{B}_{\text{mask}}$ is the attention mask derived from the ligand atom and pair masks. After K pairwise refinement layers, the final ligand memory is obtained by layer normalization:

$$\mathbf{M}_\ell = \text{LayerNorm}\left(\mathbf{H}_\ell^{(K)}\right), \qquad \mathbf{M}_\ell \in \mathbb{R}^{B\times M\times d}.$$

This ligand memory contains both semantic atom-level chemical information and coordinate-aware geometric information, and is used as the conditioning memory in the protein denoising Transformer.

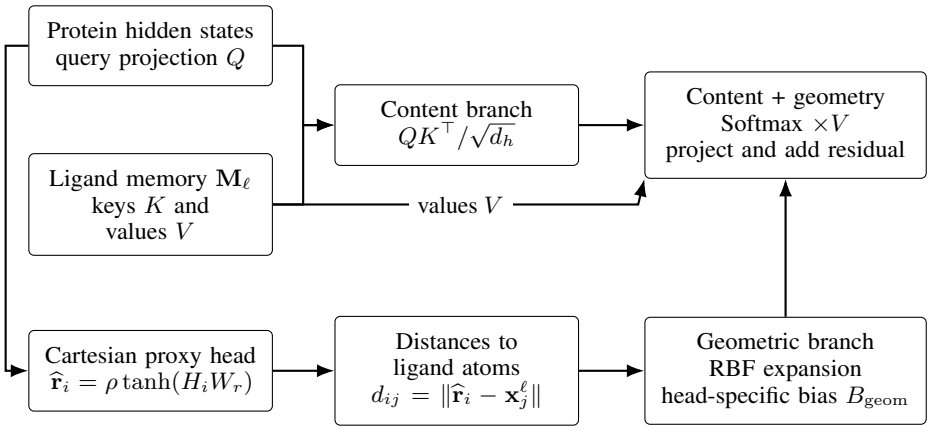


Figure 5: Geometry-aware ligand cross-attention module. Protein hidden states provide queries, while ligand embeddings provide keys and values. A separate geometric branch predicts protein-token proxy coordinates, computes protein-ligand pairwise distances against ligand atom coordinates, and converts them into a head-wise geometric bias. The final cross-attention logits combine content-based attention and geometric bias.

B.3 Geometry-Aware Ligand Cross-Attention

The geometry-aware ligand cross-attention module is illustrated in Fig. 5. Given protein hidden states $\mathbf{H} \in \mathbb{R}^{B\times T\times d}$ and ligand memory $\mathbf{M}_\ell \in \mathbb{R}^{B\times M\times d}$, the protein hidden states are used as queries, while ligand embeddings provide keys and values:

$$\mathbf{Q} = \mathrm{RMSNorm}(\mathbf{H}\mathbf{W}_Q), \qquad \mathbf{K}_\ell = \mathrm{RMSNorm}(\mathbf{M}_\ell \mathbf{W}_K), \qquad \mathbf{V}_\ell = \mathbf{M}_\ell \mathbf{W}_V.$$

The standard content-based cross-attention logits are

$$\mathbf{A}_{\text{content}} = \frac{\mathbf{Q}\mathbf{K}_\ell^\top}{\sqrt{d_h}}.$$

To make ligand conditioning sensitive to protein-ligand geometry, we add a geometric bias branch. Each protein token hidden state is first mapped to a learned proxy coordinate:

$$\widehat{\mathbf{R}} = \rho \cdot \tanh(\mathbf{H}\mathbf{W}_r), \qquad \widehat{\mathbf{R}} \in \mathbb{R}^{B\times T\times 3},$$

where $\rho = 20$ constrains the proxy coordinates to a bounded spatial range. We then compute pairwise distances between the predicted protein-token proxy coordinates and the ligand atom coordinates:

$$d_{ij} = \left\|\widehat{\mathbf{r}}_i - \mathbf{x}_j^{\ell\prime}\right\|_2, \qquad \mathbf{D} \in \mathbb{R}^{B\times T\times M}.$$

The distance matrix is expanded with radial basis functions and projected into a head-wise geometric bias:

$$\mathbf{B}_{\text{geom}} = \mathrm{Proj}_{\text{rbf}}\left(\mathrm{RBF}(\mathbf{D})\right), \qquad \mathbf{B}_{\text{geom}} \in \mathbb{R}^{B\times H\times T\times M}.$$

The final attention logits combine content-based similarity and geometric bias:

$$\mathbf{A} = \frac{\mathbf{Q}\mathbf{K}_\ell^\top}{\sqrt{d_h}} + \mathbf{B}_{\text{geom}}.$$

The ligand-conditioned cross-attention output is then

$$\mathrm{CrossAttn}_\ell = \mathrm{Softmax}(\mathbf{A})\mathbf{V}_\ell.$$

The output is projected back to the model dimension and added to the protein hidden states through a residual connection.

B.4 MCM-ReMask Decoding

Algorithm 1 specifies the procedure of the MCM-ReMask policy.

Algorithm 1 Max Confidence-Margin Guided ReMask Decoding

Require: Current sequence x_t, MDLM reverse transition $q(x_s \mid x_t, \hat{x}_0)$ defined by Eq. (2), model logits ℓ, mask token [MASK], retained fraction $r \in (0, 1]$
Ensure: Updated sequence x_s

$M_{\text{active}} \leftarrow (x_t = [\texttt{MASK}]), \quad x_s \leftarrow x_t$
for each i with $M^{(i)}_{\text{active}} = \text{True}$ **do**
    Sample $x_s^{(i)} \sim q(x_s^{(i)} \mid x_t, \hat{x}_0)$ according to Eq. (2)
end for
$k \leftarrow \left\lfloor r \sum_i \mathbf{1}\left[x_t^{(i)} = [\texttt{MASK}] \wedge x_s^{(i)} \neq [\texttt{MASK}]\right] \right\rfloor$
if $k \leq 0$ **then** set $x_s^{(i)} \leftarrow [\texttt{MASK}]$ for all i with $M^{(i)}_{\text{active}}$ and **return** x_s
Construct constrained logits $\tilde{\ell}$ by suppressing [MASK] and invalid token types
for each position i **do**

$$c^{(i)} \leftarrow \begin{cases} \left|p_1^{(i)} - p_2^{(i)}\right|, & M^{(i)}_{\text{active}} = \text{True}, \\ -\infty, & \text{otherwise}, \end{cases}$$

    where $(p_1^{(i)}, p_2^{(i)}) = \text{TopK}(\text{Softmax}(\tilde{\ell}^{(i)}), 2)$
end for
$k \leftarrow \min(k, |\{i : c^{(i)} \neq -\infty\}|)$
$S \leftarrow \text{TopK}(c, k)$ ▷ positions whose sampled candidates are retained
while $\exists i \in S$ such that $x_s^{(i)} = [\texttt{MASK}]$ **do**
    Set $c^{(i)} \leftarrow -\infty$ for all $i \in S$ with $x_s^{(i)} = [\texttt{MASK}]$
    if no valid candidates remain **then**
        return x_s
    end if
    $k \leftarrow \min(k, |\{i : c^{(i)} \neq -\infty\}|)$
    $S \leftarrow \text{TopK}(c, k)$
end while
for each i with $M^{(i)}_{\text{active}} = \text{True}$ **do**

$$x_s^{(i)} \leftarrow \begin{cases} x_s^{(i)}, & i \in S \quad \text{retain sampled candidate}, \\ [\texttt{MASK}], & i \notin S \quad \text{remask uncertain position}. \end{cases}$$

end for
return x_s

C Experiment

C.1 Unmasking Strategy Evaluation for Protein Co-design

We provide the complete unmasking strategy comparison across protein lengths from 100 to 700 residues. For each decoding strategy and target length, 100 protein sequence-structure pairs were generated and evaluated by comparing the GCP-VQVAE-decoded backbone structure with the ESMFold-predicted structure from the generated sequence. The results are reported as mean $\pm$ standard deviation for CA-RMSD, BB-RMSD, TM-score, and pLDDT, where lower RMSD and higher TM-score indicate closer agreement between the decoded and independently refolded structures; pLDDT measures ESMFold prediction confidence.

C.2 Ligand-Conditioned Whole Protein Co-design

For each benchmark target, we use the native ligand and the length of the corresponding ligand-binding protein from the PDB complex as the input condition. Each method generates 10 candidate protein designs for the same ligand and target length. We evaluate sequence-structure self-consistency by folding each generated amino-acid sequence with ESMFold and comparing the ESMFold-predicted structure with the model-generated protein structure. We report backbone RMSD, Cα RMSD, TM-score, and pLDDT. For each target, the candidate with the highest ESMFold pLDDT is selected as the representative design for downstream ligand-aware evaluation.

Table 7: Comparison of different decoding methods across protein lengths.

Method	Length	CA-RMSD ↓	BB-RMSD ↓	TM-score ↑	pLDDT ↑
LLaDA-ReMask	100	13.55 ± 5.85	13.45 ± 5.86	0.576 ± 0.164	43.69 ± 13.40
LLaDA-Random	100	13.62 ± 4.49	13.51 ± 4.49	0.571 ± 0.149	42.24 ± 10.44
TopK-Margin	100	13.18 ± 4.36	13.07 ± 4.37	0.545 ± 0.152	41.80 ± 10.41
ReMDM	100	13.43 ± 3.40	13.32 ± 3.41	0.569 ± 0.135	42.70 ± 10.17
MCM-ReMask	100	$\mathbf{12.43 \pm 5.24}$	$\mathbf{12.32 \pm 5.25}$	$\mathbf{0.635 \pm 0.144}$	$\mathbf{48.09 \pm 12.54}$
LLaDA-ReMask	200	15.28 ± 14.12	15.19 ± 14.14	0.700 ± 0.214	$\mathbf{59.25 \pm 18.53}$
LLaDA-Random	200	13.72 ± 6.05	13.62 ± 6.04	0.576 ± 0.183	43.99 ± 17.18
TopK-Margin	200	12.08 ± 6.77	11.98 ± 6.75	0.621 ± 0.217	49.83 ± 19.85
ReMDM	200	13.54 ± 6.26	13.44 ± 6.25	0.598 ± 0.199	44.79 ± 18.50
MCM-ReMask	200	$\mathbf{10.74 \pm 6.31}$	$\mathbf{10.65 \pm 6.31}$	$\mathbf{0.721 \pm 0.169}$	56.85 ± 17.16
LLaDA-ReMask	300	19.84 ± 22.05	19.77 ± 22.06	0.668 ± 0.266	$\mathbf{62.76 \pm 17.46}$
LLaDA-Random	300	14.13 ± 8.00	14.04 ± 8.00	0.621 ± 0.199	47.01 ± 16.69
TopK-Margin	300	$\mathbf{12.60 \pm 10.85}$	$\mathbf{12.51 \pm 10.85}$	0.709 ± 0.223	56.30 ± 18.61
ReMDM	300	13.80 ± 8.06	13.71 ± 8.06	0.617 ± 0.216	46.13 ± 17.96
MCM-ReMask	300	14.41 ± 14.32	14.33 ± 14.33	$\mathbf{0.712 \pm 0.183}$	59.11 ± 14.69
LLaDA-ReMask	400	19.37 ± 17.71	19.31 ± 17.71	0.672 ± 0.254	$\mathbf{63.79 \pm 16.79}$
LLaDA-Random	400	13.77 ± 8.17	13.69 ± 8.17	0.656 ± 0.223	54.26 ± 18.82
TopK-Margin	400	12.98 ± 9.89	12.91 ± 9.90	0.701 ± 0.235	59.40 ± 16.92
ReMDM	400	14.30 ± 8.09	14.22 ± 8.08	0.652 ± 0.217	51.82 ± 19.24
MCM-ReMask	400	$\mathbf{12.42 \pm 10.54}$	$\mathbf{12.34 \pm 10.55}$	$\mathbf{0.716 \pm 0.222}$	63.43 ± 16.19
LLaDA-ReMask	500	25.30 ± 20.11	25.23 ± 20.13	0.649 ± 0.277	$\mathbf{63.61 \pm 15.28}$
LLaDA-Random	500	$\mathbf{15.85 \pm 8.81}$	$\mathbf{15.77 \pm 8.82}$	0.644 ± 0.208	51.01 ± 20.75
TopK-Margin	500	16.78 ± 12.60	16.70 ± 12.61	$\mathbf{0.709 \pm 0.236}$	55.17 ± 20.00
ReMDM	500	18.46 ± 9.15	18.38 ± 9.15	0.592 ± 0.221	46.64 ± 20.12
MCM-ReMask	500	16.93 ± 11.45	16.86 ± 11.46	0.664 ± 0.240	59.28 ± 16.39
LLaDA-ReMask	600	30.25 ± 18.73	30.18 ± 18.74	0.517 ± 0.273	$\mathbf{56.62 \pm 18.57}$
LLaDA-Random	600	20.93 ± 7.71	20.85 ± 7.71	0.522 ± 0.195	42.73 ± 17.23
TopK-Margin	600	22.02 ± 12.49	21.95 ± 12.49	0.612 ± 0.253	55.46 ± 19.11
ReMDM	600	23.79 ± 7.01	23.72 ± 7.01	0.457 ± 0.169	36.59 ± 15.64
MCM-ReMask	600	$\mathbf{20.66 \pm 11.73}$	$\mathbf{20.60 \pm 11.73}$	$\mathbf{0.625 \pm 0.247}$	55.87 ± 16.75
LLaDA-ReMask	700	31.31 ± 12.81	31.24 ± 12.81	0.534 ± 0.247	$\mathbf{55.38 \pm 15.75}$
LLaDA-Random	700	24.93 ± 5.81	24.86 ± 5.81	0.528 ± 0.177	41.59 ± 13.84
TopK-Margin	700	27.54 ± 10.64	27.48 ± 10.64	0.537 ± 0.230	48.33 ± 16.03
ReMDM	700	26.56 ± 4.55	26.49 ± 4.55	0.468 ± 0.161	34.18 ± 11.92
MCM-ReMask	700	$\mathbf{24.07 \pm 11.56}$	$\mathbf{24.01 \pm 11.56}$	$\mathbf{0.588 \pm 0.259}$	54.71 ± 17.87

To assess ligand compatibility, we use AlphaFold3 to predict the complex structure between the selected designed protein and the input ligand. Based on the AF3-predicted protein-ligand complex, we center the docking box on the ligand and compute the AF3 Vina score using AutoDock Vina.

Table 8 defines the combined pass-rate criteria used for ligand-conditioned whole-protein co-design evaluation. These criteria combine different thresholds for global fold similarity, model confidence, backbone-level structural agreement, and docking scores. FC and HCF assess whether a generated protein forms a confident and globally consistent fold, BC-5 and BC-7 further require favorable AF3-Vina scores under two docking thresholds, and SWPS represents the strictest criterion by requiring high fold confidence, low backbone RMSD, and a favorable docking score simultaneously, without implying experimental binding.

Among the 200 benchmark targets, AF3-Vina scores were obtained for 191 targets. The remaining 9 targets were excluded from Vina-score analysis because their ligands could not be converted into valid AutoDock Vina-compatible representations. These failures were mainly caused by RDKit sanitization errors from invalid valence assignments after ligand format conversion, or unsupported AutoDock atom types, such as Au or B, in the generated PDBQT files. Since these errors occurred during ligand preparation and PDBQT parsing, Vina scoring could not be performed for these cases. Therefore, Vina-based metrics are reported on the 191 successfully processed targets, while structure-based metrics are reported on the full target set when available.

Table 8: Definitions of the combined pass-rate criteria used for ligand-conditioned whole-protein co-design.

Shortcut	Criterion
FC	$\mathrm{TM}_{BB} > 0.7 \ \wedge \ \mathrm{pLDDT} > 70$
HCF	$\mathrm{TM}_{BB} > 0.85 \ \wedge \ \mathrm{pLDDT} > 85 \ \wedge \ \mathrm{BB\text{-}RMSD} < 2.0$
BC-5	$\mathrm{TM}_{BB} > 0.7 \ \wedge \ \mathrm{pLDDT} > 70 \ \wedge \ \mathrm{AF3\text{-}Vina} \leq -5.0$
BC-7	$\mathrm{TM}_{BB} > 0.7 \ \wedge \ \mathrm{pLDDT} > 70 \ \wedge \ \mathrm{AF3\text{-}Vina} \leq -7.0$
SWPS	$\mathrm{TM}_{BB} > 0.85 \ \wedge \ \mathrm{pLDDT} > 85 \ \wedge \ \mathrm{BB\text{-}RMSD} < 2.0 \ \wedge \ \mathrm{AF3\text{-}Vina} \leq -7.0$

Table 9: Definitions of the combined pass-rate criteria used for pocket co-design.

Shortcut	Criterion
FC	$\mathrm{TM}_{BB} > 0.7 \ \wedge \ \mathrm{pLDDT} > 70$
HCF	$\mathrm{TM}_{BB} > 0.8 \ \wedge \ \mathrm{pLDDT} > 80 \ \wedge \ \mathrm{BB\text{-}RMSD} < 2.0$
PGC	$\mathrm{AS\text{-}local\ BB\text{-}RMSD} < 2.0 \ \wedge \ \mathrm{TM}_{BB} > 0.7 \ \wedge \ \mathrm{pLDDT} > 70$
BC-5	$\mathrm{AS\text{-}local\ BB\text{-}RMSD} < 2.0 \ \wedge \ \mathrm{TM}_{BB} > 0.7 \ \wedge \ \mathrm{pLDDT} > 70 \ \wedge \ \mathrm{Vina} \leq -5.0$
BC-7	$\mathrm{AS\text{-}local\ BB\text{-}RMSD} < 2.0 \ \wedge \ \mathrm{TM}_{BB} > 0.7 \ \wedge \ \mathrm{pLDDT} > 70 \ \wedge \ \mathrm{Vina} \leq -7.0$
SDS	$\mathrm{AS\text{-}local\ BB\text{-}RMSD} < 1.0 \ \wedge \ \mathrm{TM}_{BB} > 0.8 \ \wedge \ \mathrm{pLDDT} > 80 \ \wedge \ \mathrm{Vina} \leq -7.0$

C.3 Pocket Co-design

The pocket co-design benchmark is constructed from the 200-target benchmark dataset described in Section 3.1. Since this evaluation focuses on single-chain ligand-binding pocket design, we exclude 50 multichain complexes, resulting in 150 valid pocket-design targets. For each target, ligand-contacting active-site residues are defined as protein residues with any heavy atom within 6.0 Å of any ligand heavy atom in the native protein-ligand complex. Each method is then tasked with redesigning these pocket residues while keeping the remaining protein context fixed.

For each target, each method generates 10 candidate pocket designs under the same ligand and structural context. We fold each generated amino-acid sequence using ESMFold and compare the ESMFold-predicted structure with the model-generated structure. We report both global and active-site RMSD using backbone atoms and Cα atoms, as well as TM-score and pLDDT. Active-site RMSD is computed over ligand-contacting residues, while global RMSD and TM-score are computed over the full protein chain. For each target, the candidate with the highest ESMFold pLDDT is selected as the representative design for ligand-aware evaluation.

To evaluate ligand-binding plausibility, we compute AutoDock Vina scores using a ligand-centered docking box. Among the 150 valid pocket-design targets, Vina scoring was successfully performed for 149 targets. One target was excluded from Vina-based evaluation because its ligand could not be converted into a valid AutoDock Vina-compatible representation. Therefore, structure-based metrics are reported on 150 targets, while Vina-related pass-rate criteria are evaluated on the targets with valid Vina scores.

Figure 6 shows additional qualitative pocket co-design examples on 6U5Y and 7AC8. In both cases, ProtLiD^2 maintains high global fold similarity while achieving the lowest Site-RMSD among the compared methods, indicating more accurate ligand-binding pocket geometry. These visual examples are consistent with the aggregate results, where ProtLiD^2 improves pocket accuracy and ligand-aware pass rates over FAIR and PocketGen.

Table 9 defines the combined pass-rate criteria used for pocket co-design evaluation. These criteria progressively combine global fold confidence, active-site geometric accuracy, and ligand-aware docking quality. FC and HCF measure global fold confidence under standard and stricter thresholds, PGC additionally requires accurate active-site geometry, BC-5 and BC-7 further impose ligand-aware docking-score thresholds, and SDS denotes the strictest success criterion by requiring accurate active-site geometry, high fold confidence, and favorable predicted ligand binding simultaneously.

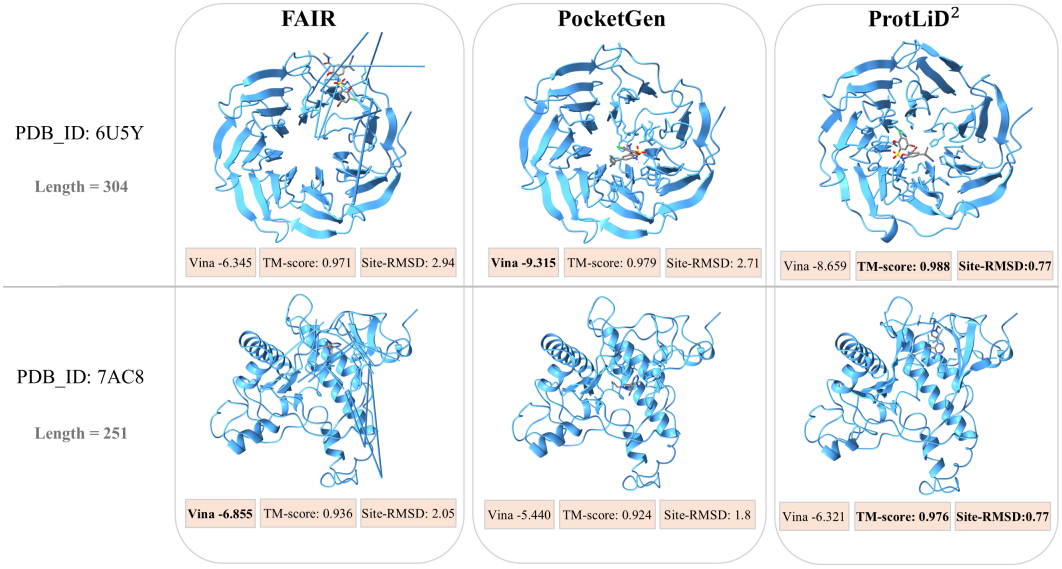


Figure 6: Additional qualitative pocket co-design case studies on 6U5Y and 7AC8. ProtLiD^2 has lower Site-RMSD and high TM-score in these selected examples, indicating closer agreement with independently refolded structures under the reported selection rules.

D Additional Ablation and Sensitivity Details

Section 4.5 presents the main results for the 370M-parameter ProtLiD^2 model. This appendix details component removals, decoder policies and retention, rotation sensitivity, and screening thresholds.

D.1 Same-Checkpoint Component Removals

We evaluate four settings with the checkpoint, tokenizer, conditioning ligands, sampling procedure, and evaluation protocol otherwise fixed. Each setting includes 100 generations at each length $100, 200, \ldots, 600$. We independently refold the generated sequences with ESMFold and compare them with the decoded backbones. Table 5 in Section 4.5 reports aggregate means without uncertainty estimates.

D.2 Decoder Policies, Retention, and Cost

We compare LLaDA-ReMask, LLaDA-Random, TopK-Margin, ReMDM, and MCM-ReMask using the same protein-specific, modality-constrained logits and 500 reverse steps. Table 7 reports the length-wise results. MCM-ReMask ranks newly proposed non-mask positions by confidence margin and retains $k_j = \lfloor r|T_j| \rfloor$ proposals.

For retained-fraction sensitivity, we evaluate $r \in \{0.5, 0.7, 0.9, 1.0\}$ using the 370M model. Each setting generates 100 designs at each length $100, 200, \ldots, 600$, yielding 600 designs per setting and 2,400 in total. Table 6 reports the mean evaluation metrics.

D.3 Rotation and Recanonicalization Sensitivity

For descending ligand covariance eigenvalues $\lambda_1 \geq \lambda_2 \geq \lambda_3$, define

$$g_{\min} = \min\left(\frac{\lambda_1 - \lambda_2}{\lambda_1 + \epsilon}, \frac{\lambda_2 - \lambda_3}{\lambda_1 + \epsilon}\right), \qquad \epsilon = 10^{-12}.$$

We stratify 100 targets from the 200-target benchmark by eigengap and observed PCA instability under 0.05 Å ligand-coordinate perturbations, retaining at most one target per PDB and representing every nonempty stability stratum. There are 81 well-conditioned ($g_{\min} \geq 0.10$), 13 mildly ambiguous

($0.05 \leq g_{\min} < 0.10$), five nearly symmetric ($0.01 \leq g_{\min} < 0.05$), and one strongly degenerate ($g_{\min} < 0.01$) ligand.

For each target we generate six predictions at the original orientation and six under each of three independently sampled global rotations, recomputing ligand-guided PCA each time. The whole-protein analysis thus uses 2,400 predictions. We average the six predictions within each target and condition, then average the three rotated conditions within each target. Paired target-level bootstrap intervals quantify the rotated-minus-original change (Table 10).

Table 10: Whole-protein rotation-and-recanonicalization sensitivity on 100 targets. Means ± SD are across target-level averages. Paired changes are reproduced as reported; displayed means are rounded. RMSD is in Å.

Metric	Original PCA	Rotated input + PCA	Change [95% CI]
CA-RMSD ↓	19.59 ± 7.67	19.85 ± 7.67	$+0.26\ [-0.14, +0.66]$
BB-RMSD ↓	19.50 ± 7.67	19.76 ± 7.68	$+0.26\ [-0.14, +0.66]$
TM-score ↑	0.536 ± 0.181	0.538 ± 0.179	$+0.001\ [-0.009, +0.012]$
pLDDT ↑	39.42 ± 16.62	39.68 ± 16.56	$+0.26\ [-0.62, +1.18]$

Across the 100-target analysis, the paired confidence intervals for all metric changes include zero. For the six low-eigengap targets, the mean changes are $+0.65$ Å for both CA- and BB-RMSD, $+0.010$ for TM-score, and -0.49 for pLDDT, with paired intervals also spanning zero. These results show no consistent average deterioration under the tested rotations.

Under pocket-design settings, we evaluate the same 100 targets using six predictions per original or rotated condition (Table 11). Mean TM-score and pLDDT increase by 0.016 and 1.22, respectively, indicating no average deterioration in these metrics.

Table 11: Additional rotation study under pocket-design settings on the same 100-target subset. RMSD is in Å; Δ is the mean over rotations minus the original.

Metric	Original	Rotation 1	Rotation 2	Rotation 3	Δ
CA-RMSD ↓	17.05 ± 10.16	16.30 ± 9.37	16.19 ± 9.42	16.21 ± 10.05	-0.82
BB-RMSD ↓	16.98 ± 10.16	16.23 ± 9.37	16.11 ± 9.42	16.13 ± 10.05	-0.82
TM-score ↑	0.658 ± 0.211	0.673 ± 0.192	0.671 ± 0.200	0.678 ± 0.191	$+0.016$
pLDDT ↑	57.32 ± 15.79	58.12 ± 15.15	58.22 ± 15.35	59.27 ± 15.33	$+1.22$

D.4 Screening-Threshold Sensitivity

To assess whether the combined pass-rate comparisons depend on the chosen docking threshold, we vary the Vina cutoff from -9 to -4. Table 12 presents results at -5 and -7. For pocket design, ProtLiD^2 maintains higher pass rates than PocketGen and FAIR across all tested cutoffs. At -5, the respective pass rates are 59.46%, 14.86%, and 6.76%; at -7, they are 23.65%, 6.08%, and 1.35%. For whole-protein design, the advantage of ProtLiD^2 over Complexa narrows from 5.26 percentage points at -5 to 0.53 points at -7. These results indicate that the pocket-design advantage persists across the tested screening thresholds, whereas the whole-protein advantage is modest and more sensitive to cutoff strictness.

Table 12: Reported secondary threshold-sensitivity summary, in percent. Rows compare ProtLiD^2 with a baseline; these rates are distinct from the original available-case tables. Target-level counts and cohort manifests are not provided with the summary.

Comparison	Vina ≤ -5	Vina ≤ -7
Whole protein: ProtLiD^2 / Complexa	52.63 / 47.37	30.00 / 29.47
Pocket: ProtLiD^2 / PocketGen	59.46 / 14.86	23.65 / 6.08
Pocket: ProtLiD^2 / FAIR	59.46 / 6.76	23.65 / 1.35

The reported sweep favors ProtLiD^2 for pocket screening; whole-protein ordering is threshold-dependent.