BEYOND THE RIGID POCKET: PROTEIN FLEXIBILITY-INFORMED LIGAND GENERATION FOR STRUCTURE-BASED DRUG DESIGN

Anonymous authors
Paper under double-blind review

ABSTRACT

Structure-based ligand generation commonly assumes a fixed target protein structure, overlooking conformational heterogeneity and the structural adjustments that accompany ligand binding. Modeling protein flexibility remains challenging: matched apo-holo structure pairs are scarce, and pocket-focused methods often omit structural context beyond the binding site. Here, we present FLEXIDD (**Flex**ibility-**I**nformed **D**rug **D**esign), a generative framework for *de novo* drug design that explicitly models full-protein conformational adaptation at atomic resolution during ligand generation. To address the scarcity of paired apo–holo data, FLEXIDD introduces a transferable coordinate denoising strategy that uses coordinate perturbations to bring apo and holo conformations into overlapping noisy distributions. This enables training solely on holo complexes, while supporting both apo and holo inputs at inference. By modeling the full protein rather than isolated binding pockets, FLEXIDD incorporates global structural context and constraints into local binding-site adaptation. Experiments on a leakage-audited benchmark demonstrate strong ligand generation performance, with favorable predicted binding affinities and higher PoseBusters validity rates across both holo and apo inputs. Under apo conditioning, FLEXIDD achieves 91.6% PoseBusters validity, compared with less than 46.1% for the evaluated baselines that also account for target flexibility. These results demonstrate the potential of holo-only training and full-protein conformational modeling for ligand generation from apo structures, supporting FLEXIDD as a practical approach to flexibility-informed structure-based drug design.

1 INTRODUCTION

In structure-based drug design (SBDD) (Zhang et al., 2025c), deep generative models have emerged as a powerful paradigm to generate ligands targeting protein binding pockets (Ragoza et al., 2022; Peng et al., 2022; Guan et al., 2023; Schneuing et al., 2024; Qu et al., 2024; Cremer et al., 2026a). By capturing geometric and chemical dependencies between pockets and ligands, these methods can efficiently sample target-compatible molecules across vast chemical space.

Existing SBDD approaches typically treat the protein as a rigid template and generate only the ligand (Guan et al., 2023; 2024; Schneuing et al., 2024; Cremer et al., 2026a; Peng et al., 2026), which often suffices when a suitable ligand-bound structure is available. However, ligand binding can stabilize pre-existing protein conformations or induce structural adjustments, leading to well-documented differences between the unbound (apo) and bound (holo) states (Teague, 2003; Boehr et al., 2009). Rigid-protein methods discard this conformational coupling. When conditioning on apo conformations or computationally predicted structures (*e.g.*, AlphaFold (Jumper et al., 2021)), they constrain ligand generation to a pocket geometry that may be incompatible with the bound state, limiting practical utility when no suitable holo structure is available (Zheng et al., 2026).

These limitations have motivated recent flexible *de novo* SBDD methods, which have primarily explored two directions. The first explicitly learns protein conformational transitions from paired apo-holo states (Zhang et al., 2024; Zhou et al., 2025; Zheng et al., 2026). This provides direct supervision for protein adaptation but requires either experimental apo counterparts or an addi-

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.

tional structure-construction procedure. The second direction learns protein flexibility directly from protein-ligand complexes. However, existing methods typically confine protein adaptation to limited degrees of freedom, such as side-chain torsions with fixed backbone anchors, or to the local binding pocket (Schneuing et al., 2025; Wang et al., 2026), overlooking the non-pocket environment that constrains pocket geometry and makes globally consistent adaptation difficult.

To address these limitations, we propose **FLEXIDD** (**Flex**ibility-**I**nformed **D**rug **D**esign), a generative framework for *de novo* drug design that explicitly models full-protein conformational adaptation at atomic resolution during ligand generation. First, to alleviate the data bottleneck of paired apo-holo structures, FLEXIDD sidesteps direct apo-to-holo mapping through transferable protein coordinate denoising. Moderate coordinate perturbations bring apo and holo conformations of the same protein into a shared noisy space while preserving structural identity. The learned denoiser therefore generalizes across conformations at inference (Fig. 1). This design enables FLEXIDD to leverage abundant holo complexes directly, without requiring curated apo counterparts. Second, to move beyond pocket-restricted modeling, FLEXIDD denoises the full protein structure while jointly generating ligand 3D geometry and chemistry. Full-protein adaptation and ligand generation thus proceed in tandem, each informing the other throughout sampling.

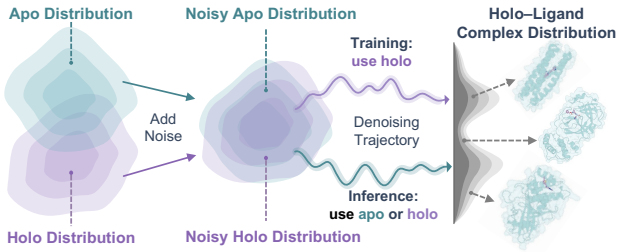

Figure 1: Bridging the apo-holo conformational gap via perturbation. Noise creates an overlapping representation that allows models trained on holo structures to generalize to both inputs at inference.

To equip the model with sufficient structural priors, we adopt a two-stage training curriculum: In the first stage, the protein module is independently trained on a protein-only corpus, comprising 302k PDB structures (Burley et al., 2017) and 133k synthetic data from OpenFold (Ahdritz et al., 2024), to acquire structural priors for protein modeling. In the second stage, we jointly train protein and ligand modules on 332k protein-ligand complexes from GatorAffinity-DB (Wei et al., 2025), PLINDER (Durairaj et al., 2024), and HiQBind (Wang et al., 2025), to learn the joint distribution over protein conformation, ligand geometry, and chemical composition. We evaluate FLEXIDD on a leakage-audited apo-holo benchmark. Under apo inputs, FLEXIDD achieves 91.6% PoseBusters validity, nearly doubling that of the best-performing flexible baseline among the compared methods (46.1%), while maintaining competitive performance across apo and holo proteins.

To sum up, our main contributions are as follows:

- **FLEXIDD** is, to our knowledge, the first generative framework for flexibility-informed *de novo* drug design that jointly models full-protein conformational adaptation and ligand generation at all-atom resolution. Full-protein modeling provides global structural context and geometric constraints for consistent local conformational adaptation.
- We introduce a transferable protein denoising strategy that bridges apo and holo conformations through a shared noisy coordinate space. This enables the protein denoiser, trained on holo structures without paired supervision, to generalize to both apo and holo conformations at inference, while allowing FLEXIDD to leverage large-scale protein-ligand complexes for training.
- We conduct extensive evaluation on leakage-audited benchmarks. FLEXIDD supports both apo and holo target protein inputs, achieves consistently strong ligand validity while retaining competitive predicted binding affinity scores against rigid-protein baselines, and substantially outperforms existing flexible SBDD methods under apo inputs.

2 RELATED WORK

2.1 STRUCTURE-BASED DRUG DESIGN

Structure-based drug design (SBDD) generates ligands conditioned on the geometry and chemistry of a target binding site. Early methods evolved from protein-conditioned atomic density grid models to autoregressive generation of ligand atoms, coordinates, and, in some cases, covalent bonds (Ragoza

et al., 2022; Luo et al., 2021; Liu et al., 2022; Peng et al., 2022). More recent approaches jointly model ligand geometry and chemistry using diffusion, Bayesian flow networks, or flow matching, with extensions incorporating spatial decomposition, explicit bond modeling, protein-ligand interactions, and property guidance (Guan et al., 2023; 2024; Schneuing et al., 2024; Huang et al., 2024b; Gu et al., 2024; Cremer et al., 2024; Qu et al., 2024; Qiu et al., 2025; Cremer et al., 2026a). Generalist models extend beyond single-task generation: PocketXMol (Peng et al., 2026) supports docking and molecular design through atomic prompts, while FLOWR.ROOT (Cremer et al., 2026b) combines ligand generation with affinity and confidence prediction. Despite these advances, most SBDD methods treat the protein as a rigid template, leaving ligand-induced conformational adaptation underexplored.

2.2 Flexible Protein-Ligand Joint Generation

Flexible protein-ligand modeling can be broadly categorized by whether ligand chemistry is provided or generated. When the ligand is given, flexible docking methods jointly refine ligand poses and protein degrees of freedom (Zhu et al., 2024; Huang et al., 2024a; Lu et al., 2024; Corso et al., 2025; Zhang et al., 2025d; Feng et al., 2026; Morehead & Cheng, 2025; Zhang et al., 2025b). Cofolding models instead predict complex structures from protein sequences and specified molecular components (Krishna et al., 2024; Qiao et al., 2024; Abramson et al., 2024; Wohlwend et al., 2025). The more challenging setting, flexible SBDD, jointly generates ligand chemistry and adapts the protein. FlexSBDD (Zhang et al., 2024) refines pocket backbone frames and side-chain torsions from experimental or approximated apo structures, whereas FlexFlow (Schneuing et al., 2025) samples ligand variables together with pocket side-chain torsions under a fixed protein backbone. DynamicFlow (Zhou et al., 2025) and Apo2Mol (Zheng et al., 2026) learn from aligned apo-holo data, while YuelDesign (Wang et al., 2026) learns from bound complexes but retains fixed protein Cα anchors. Overall, existing flexible SBDD methods either rely on paired apo-holo supervision or restrict protein adaptation to local or predefined degrees of freedom. To the best of our knowledge, FLEXIDD is the first framework to jointly generate ligands and perform full-protein all-atom conformational adaptation without paired apo-holo supervision.

3 Method

FLEXIDD combines protein conformational adaptation with ligand generation, as illustrated in Fig. 2. We first define the task in Sec. 3.1, then describe the generative processes, network, and optimization in Sec. 3.2. Finally, we present the data and training stages in Sec. 3.3.

3.1 Problem Formulation

We consider flexible SBDD at a specified protein binding site, where ligand generation is coupled with protein conformational adaptation. A target protein with $n_{\mathcal{P}}$ atoms and n_R residues is represented as $\mathcal{P} = (F^{\mathcal{P}}, \mathbf{X}^{\mathcal{P}})$, where $\mathbf{X}^{\mathcal{P}} \in \mathbb{R}^{n_{\mathcal{P}} \times 3}$ contains all-atom coordinates and $F^{\mathcal{P}}$ contains invariant chemical and structural attributes, including residue identities, atom identities, covalent connectivity, *etc*. The target binding site is specified by a residue-level binary mask $\mathbf{m}^{\mathcal{P}} \in \{0,1\}^{n_R}$, where $m_r^{\mathcal{P}} = 1$ indicates that residue r belongs to the binding site. A ligand with $n_{\mathcal{M}}$ heavy atoms is represented as $\mathcal{M} = (\mathbf{X}^{\mathcal{M}}, \mathbf{a}, \mathbf{q}, \mathbf{b})$, where $\mathbf{X}^{\mathcal{M}} \in \mathbb{R}^{n_{\mathcal{M}} \times 3}$ contains the atomic coordinates, $\mathbf{a}^{(i)} \in \{0,1\}^{K_a}$ and $\mathbf{q}^{(i)} \in \{0,1\}^{K_q}$ are one-hot vectors encoding the atom type and formal charge of atom i, respectively, and $\mathbf{b}^{(ij)} \in \{0,1\}^{K_b}$ is a one-hot vector encoding the bond state between atoms i and j. Here, K_a, K_q, and K_b denote the number of categories for atom types, formal charges, and bond states, respectively.

Given an input protein conformation $\mathbf{X}_{\text{in}}^{\mathcal{P}}$ with a specified binding site and a prescribed ligand heavy-atom count $n_{\mathcal{M}}$, FLEXIDD models the joint conditional distribution:

$$p_\theta\big(\mathbf{X}_{\text{gen}}^{\mathcal{P}}, \mathcal{M}_{\text{gen}} \mid \mathbf{X}_{\text{in}}^{\mathcal{P}}, F^{\mathcal{P}}, \mathbf{m}^{\mathcal{P}}, n_{\mathcal{M}}\big)\,, \tag{1}$$

where $\mathcal{M}_{\text{gen}}$ denotes the generated ligand and $\mathbf{X}_{\text{gen}}^{\mathcal{P}}$ denotes the refined protein coordinates. During generation, the invariant protein attributes $F^{\mathcal{P}}$ and the binding-site specification $\mathbf{m}^{\mathcal{P}}$ remain fixed, while the protein coordinates are allowed to adapt.

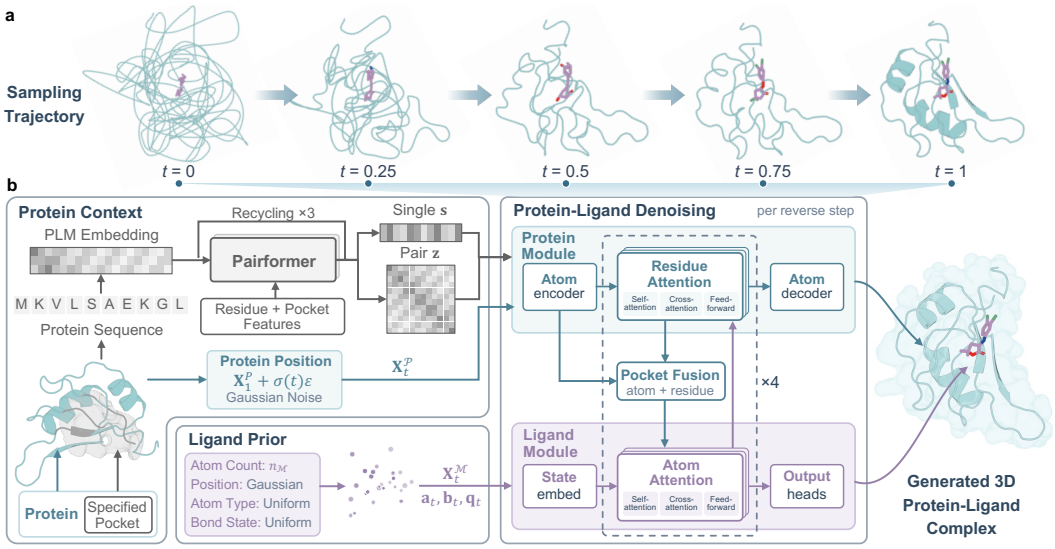

Figure 2: Overview of FLEXIDD. (**a**) FLEXIDD generates ligands while adapting the full protein conformation. (**b**) Given a protein with a specified pocket and a prescribed ligand heavy-atom count $n_{\mathcal{M}}$, the protein module pairs fixed protein context features with noisy coordinates $\mathbf{X}_t^{\mathcal{P}}$, while the ligand is initialized from priors over coordinates $\mathbf{X}_t^{\mathcal{M}}$, atom types $\mathbf{a}_t$, bond states $\mathbf{b}_t$, and formal charges $\mathbf{q}_t$. The two streams iteratively exchange geometric information, jointly updating the protein conformation and ligand structure to produce a complete 3D protein-ligand complex.

3.2 FLEXIDD DESIGN

FLEXIDD couples a protein coordinate denoiser with a ligand generation model, with the two modules exchanging information during coordinate and chemical prediction (Fig. 2).

3.2.1 PROTEIN MODELING

Directly learning protein adaptation from apo-to-holo transitions requires paired structures, which are scarce. Rather than learning this mapping explicitly, we exploit the conformational proximity between apo and holo states (Brylinski & Skolnick, 2008; Clark et al., 2019) and apply controllable coordinate perturbations to reduce their distributional discrepancy, inspired by Ambient Proteins (Daras et al., 2026).

For a given protein, let p^{apo} and p^{holo} denote the coordinate distributions of its aligned apo and holo conformations, respectively. After Gaussian corruption with noise scale σ, we define $p_\sigma^u = p^u * \mathcal{N}(0, \sigma^2\mathbf{I})$ for $u \in \{\text{apo}, \text{holo}\}$. The discrepancy between the perturbed distributions satisfies:

$$D_{\text{KL}}\left(p_\sigma^{\text{apo}} \,\|\, p_\sigma^{\text{holo}}\right) \leq \frac{W_2^2\left(p^{\text{apo}}, p^{\text{holo}}\right)}{2\sigma^2}, \tag{2}$$

where W_2 denotes the 2-Wasserstein distance. As the noise level increases, the upper bound decreases, showing that sufficient Gaussian corruption renders the two conformational distributions nearly indistinguishable. This motivates a shared noisy coordinate space in which a denoiser trained on holo complexes can generalize to apo protein inputs at inference time (Fig. 1). The derivation and underlying assumptions are provided in Apx. A.1.

Accordingly, given a reference protein conformation $\mathbf{X}_{\text{ref}}^{\mathcal{P}}$, we construct the perturbed coordinates as:

$$\mathbf{X}_t^{\mathcal{P}} = \mathbf{X}_{\text{ref}}^{\mathcal{P}} + \sigma(t)\boldsymbol{\epsilon}^{\mathcal{P}}, \quad \boldsymbol{\epsilon}^{\mathcal{P}} \sim \mathcal{N}(0, \mathbf{I}), \tag{3}$$

where $\sigma(t) = \sigma_{\min}\left(\sigma_{\max}/\sigma_{\min}\right)^{1-t}$ decreases from $\sigma_{\max}$ at $t = 0$ to $\sigma_{\min}$ at $t = 1$. During joint training on protein-ligand complexes, $\mathbf{X}_{\text{ref}}^{\mathcal{P}}$ corresponds to the holo conformation. In addition to the noisy coordinates $\mathbf{X}_t^{\mathcal{P}}$, the protein module is conditioned on time-invariant protein features. Specifically, given the protein sequence $\mathbf{r} = (r_1, \ldots, r_{n_R})$, we extract residue embeddings $\mathbf{H}^{\text{ESM}} = \text{ESMC}(\mathbf{r})$ from ESM-C (Candido et al., 2026) and combine them with the invariant protein attributes $F^{\mathcal{P}}$ and the binding-site mask $\mathbf{m}^{\mathcal{P}}$.

These features are then processed by a Pairformer module (Abramson et al., 2024):

$$(\mathbf{s}, \mathbf{z}) = \text{Pairformer}\left(\mathbf{H}^{\text{ESM}}, F^{\mathcal{P}}, \mathbf{m}^{\mathcal{P}}\right), \quad \mathbf{s} \in \mathbb{R}^{n_R \times d_s}, \quad \mathbf{z} \in \mathbb{R}^{n_R \times n_R \times d_z}, \tag{4}$$

where $\mathbf{s}$ and $\mathbf{z}$ denote the residue-wise single and pair representations, respectively, and d_s and d_z are their feature dimensions.

3.2.2 Ligand Modeling

We model ligand geometry and chemistry using a continuous coordinate path and categorical probability paths, respectively (Zhou et al., 2025; Cremer et al., 2026a). We define $\mathcal{M}_{\text{ref}} = (\mathbf{X}^{\mathcal{M}}_{\text{ref}}, \mathbf{a}_{\text{ref}}, \mathbf{q}_{\text{ref}}, \mathbf{b}_{\text{ref}})$ as the reference ligand associated with the observed protein–ligand complex.

For ligand coordinates, we construct a pocket-centered continuous path. Let $\mathbf{x}^{\mathcal{P}}_{\text{pocket}} \in \mathbb{R}^3$ denote the centroid of the selected pocket fusion atoms. We sample i.i.d. $\mathbf{g}^{(i)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and mean-center them to obtain offsets $\mathbf{z}^{\mathcal{M},(i)} = \mathbf{g}^{(i)} - n_{\mathcal{M}}^{-1} \sum_j \mathbf{g}^{(j)}$. The prior coordinates are $\mathbf{x}_0^{\mathcal{M},(i)} = \mathbf{x}^{\mathcal{P}}_{\text{pocket}} + \mathbf{z}^{\mathcal{M},(i)}$, collectively forming $\mathbf{X}_0^{\mathcal{M}}$. We construct the interpolated ligand coordinates as:

$$\mathbf{X}_t^{\mathcal{M}} = (1-t)\mathbf{X}_0^{\mathcal{M}} + t\mathbf{X}^{\mathcal{M}}_{\text{ref}} + \eta\boldsymbol{\epsilon}^{\mathcal{M}}, \quad \boldsymbol{\epsilon}^{\mathcal{M}} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{5}$$

where η is a small noise scale applied during training.

For ligand chemistry, we model atom types, formal charges, and bond states using categorical paths from uniform priors to their reference states (Zhou et al., 2025; Cremer et al., 2026a). For each categorical variable $\mathbf{c} \in \{\mathbf{a}, \mathbf{q}, \mathbf{b}\}$, we sample:

$$\mathbf{c}_t^{(\ell)} \sim \text{Cat}\left((1-t)\mathbf{u}_c + t\mathbf{c}^{(\ell)}_{\text{ref}}\right), \tag{6}$$

where $\mathbf{u}_c$ is uniform over admissible categories (excluding atom and charge padding) and $\mathbf{c}^{(\ell)}_{\text{ref}}$ is the corresponding one-hot reference state. Here $\ell = i$ for atom-level variables $\mathbf{a}$ and $\mathbf{q}$, and $\ell = (i, j)$ with $i < j$ for the bond variable $\mathbf{b}$. Bond states are mirrored across atom pairs to preserve symmetry, with diagonal entries fixed to the no-bond state.

3.2.3 Joint Protein-Ligand Network

With protein and ligand representations initialized as above, an $E(3)$-equivariant denoising network f_θ carries out the reverse diffusion process (Fig. 2 (b); equivariance proof in Apx. A.3). At each reverse step, the network receives the current protein and ligand states $\mathcal{M}_t$, together with the time step t, and predicts the clean protein and ligand coordinates $\widehat{\mathbf{X}}^{\mathcal{P}}_{\text{ref}}$ and $\widehat{\mathbf{X}}^{\mathcal{M}}_{\text{ref}}$, along with categorical distributions $\widehat{\mathbf{p}}^a$, $\widehat{\mathbf{p}}^q$, and $\widehat{\mathbf{p}}^b$ over atom types, formal charges, and bond states, respectively:

$$\left(\widehat{\mathbf{X}}^{\mathcal{P}}_{\text{ref}}, \widehat{\mathbf{X}}^{\mathcal{M}}_{\text{ref}}, \widehat{\mathbf{p}}^a, \widehat{\mathbf{p}}^q, \widehat{\mathbf{p}}^b\right) = f_\theta\left(\mathbf{X}_t^{\mathcal{P}}, \mathcal{M}_t, \mathbf{s}, \mathbf{z}, t\right). \tag{7}$$

The network consists of three components: a protein module responsible for conformational generation, a ligand module for *de novo* molecular generation, and an interaction module that exchanges geometric information at the protein-ligand interface.

The two modules first encode their inputs independently. On the protein side, an atom encoder extracts local geometric features from the noisy structure, pools them into residue-level features, and fuses them with the Pairformer context. On the ligand side, atomic coordinates, element types, formal charges, and bond orders are embedded into node and edge features. Both modules maintain scalar and equivariant vector channels conditioned on the time step. These encodings initialize the residue-level protein hidden state $\mathcal{H}_0^{\mathcal{P}} \in \mathbb{R}^{n_R \times d}$ and the atom-level ligand hidden state $\mathcal{H}_0^{\mathcal{M}} \in \mathbb{R}^{n_{\mathcal{M}} \times d}$.

The two modules are then interleaved over G interaction rounds, with the interaction module carrying information in both directions. Within each round, residue-level attention updates the protein state while cross-attending to the current ligand atoms. A pocket fusion adapter then gathers the updated protein state onto a fixed set of interface atoms, which the ligand atom-level attention in turn cross-attends to:

$$\mathcal{H}_{g+1}^{\mathcal{P}} = \text{ResAttn}_g\left(\mathcal{H}_g^{\mathcal{P}} \mid \mathcal{H}_g^{\mathcal{M}}\right), \quad \mathcal{H}_{g+1}^{\mathcal{M}} = \text{AtomAttn}_g\left(\mathcal{H}_g^{\mathcal{M}} \mid \text{PocketFusion}\left(\mathcal{H}_{g+1}^{\mathcal{P}}\right)\right), \tag{8}$$

where $\mathrm{ResAttn}$ and $\mathrm{AtomAttn}$ are the residue-level and atom-level attention blocks following the scalar-vector formulation of Irwin et al. (2024), and $\mathcal{H}_g^{\mathcal{P}}$, $\mathcal{H}_g^{\mathcal{M}}$ are their respective hidden states at round g.

After the final round, an atom decoder in the protein module maps residue-level features back to individual protein atoms, predicting a residue-level displacement and a per-atom refinement:

$$\widehat{\mathbf{x}}_{\mathrm{ref}}^{\mathcal{P},(i)} = \mathbf{x}_t^{\mathcal{P},(i)} + \Delta\mathbf{x}_{\pi(i)}^{\mathrm{res}} + \Delta\mathbf{x}_i^{\mathrm{atom}}, \tag{9}$$

where $\pi(i)$ is the parent residue of atom i. Separate ligand heads predict ligand atom coordinates and categorical probabilities for atom types, formal charges, and bond states.

To allow the model to iteratively refine its estimates, we adopt self-conditioning (Chen et al., 2022): the network is run twice at each step, with the preliminary pass supplying its detached predictions to a second pass that operates on the original noisy state. Module details appear in Apx. A.2.

3.2.4 Training Objective

The training objective combines coordinate reconstruction for both protein and ligand with categorical supervision for ligand chemistry:

$$\mathcal{L}_{\mathrm{pred}} = \lambda_{\mathcal{P}}\mathcal{L}_{\mathrm{coord}}^{\mathcal{P}} + \lambda_{\mathcal{M}}\mathcal{L}_{\mathrm{coord}}^{\mathcal{M}} + \lambda_a\mathcal{L}_a + \lambda_q\mathcal{L}_q + \lambda_b\mathcal{L}_b. \tag{10}$$

The coordinate losses use mean squared error, while $\mathcal{L}_a$, $\mathcal{L}_q$, and $\mathcal{L}_b$ use cross-entropy for atom types, formal charges, and bond states. We additionally regularize the geometry of the predicted complex:

$$\mathcal{L} = \mathcal{L}_{\mathrm{pred}} + \lambda_{\mathrm{lddt}}\mathcal{L}_{\mathrm{lddt}} + \lambda_{\mathrm{bond}}\mathcal{L}_{\mathrm{bond}} + \lambda_{\mathrm{angle}}\mathcal{L}_{\mathrm{angle}} + \lambda_{\mathrm{dist}}\mathcal{L}_{\mathrm{dist}}. \tag{11}$$

Here, $\mathcal{L}_{\mathrm{lddt}}$ is a smooth local Distance Difference Test (lDDT) loss (Mariani et al., 2013; Abramson et al., 2024) that measures the geometric accuracy of the predicted complex, while $\mathcal{L}_{\mathrm{bond}}$ and $\mathcal{L}_{\mathrm{angle}}$ regularize ligand bond geometry and bond angles, respectively. The distance term $\mathcal{L}_{\mathrm{dist}}$ supervises ligand-pocket distances relative to the reference complex. Detailed definitions are given in Apx. A.4.

3.2.5 Sampling

At inference, we perturb the input protein to the maximum protein noise level $\sigma_{\max}$ and initialize the ligand coordinates from a Gaussian prior centered at the specified pocket:

$$\mathbf{X}_0^{\mathcal{P}} = \mathbf{X}_{\mathrm{in}}^{\mathcal{P}} + \sigma_{\max}\boldsymbol{\epsilon}^{\mathcal{P}}, \quad \mathbf{X}_0^{\mathcal{M}} = \mathbf{x}_{\mathrm{pocket}}^{\mathcal{P}} + \mathbf{Z}^{\mathcal{M}}. \tag{12}$$

Here, $\mathbf{Z}^{\mathcal{M}}$ is the zero-centroid Gaussian offset matrix defined in Sec. 3.2.2. Ligand atom types, formal charges, and bond states are initialized from uniform categorical priors. At each step, the network predicts the clean state of both protein and ligand. Protein coordinates advance by deterministic integration of the denoiser-based vector field (Karras et al., 2022), while ligand coordinates follow Euler integration of the endpoint flow (Lipman et al., 2022; Cremer et al., 2026a). Atom types, formal charges, and bond states are sampled using discrete flow transitions parameterized by the predicted categorical probabilities (Gat et al., 2024). The coordinate updates, shared time schedule, and chemical decoding rules are given in Apx. A.5.

3.3 Data and Two-Stage Training

We train FlexiDD in two stages (Tab. 1). All training data in both stages predate a December 30, 2022 cutoff. In **Stage 1**, the protein module is pretrained on approximately 436k single-chain structures from PDB (Burley et al., 2017) and synthetic OpenFold (Ahdritz et al., 2024) data to learn a broad protein structural prior. In **Stage 2**, we initialize the protein module from Stage 1 and jointly train the full model on protein-ligand complexes filtered from GatorAffinity (Wei et al., 2025), PLINDER (Durairaj et al., 2024), and HiQBind (Wang et al., 2025), totaling 332k complexes for training. This stage teaches the model to jointly refine protein conformations and generate ligand molecules. Detailed data processing and training configurations are provided in Apx. A.8. Ablation experiments (Apx. B.6) confirm that this protein structural prior is critical for learning pocket flexibility.

Table 1: Training data statistics for each stage.

Stage	Dataset	Records
1	PDB (Burley et al., 2017)	∼302k
	OpenFold (Ahdritz et al., 2024)	∼133k
	Total	**∼436k**
2	GatorAffinity (Wei et al., 2025)	∼248k
	PLINDER (Durairaj et al., 2024)	∼64.6k
	HiQBind (Wang et al., 2025)	∼19.0k
	Total	**∼332k**

Table 2: Performance comparison on holo protein structures. The **first**, **second**, and **third** best results are shaded. The same convention applies to subsequent tables.

Method	Validity↑	PB-valid.↑	Strain↓	Vina↓	Vina Min↓	QED↑	SA↑	LogP	Lipinski↑	Div.2D↑	Uniq.2D↑
Reference	1.000	0.976	38.5	-6.99	-7.33	0.492	0.751	1.211	4.679	N/A	1.000
Pocket2Mol	0.976	0.649	155.4	-3.48	-5.24	0.620	0.789	1.225	4.908	0.854	0.988
TargetDiff	0.840	0.587	2539.9	-3.39	-5.15	0.489	0.632	1.240	4.711	0.894	1.000
DecompDiff	0.823	0.637	559.7	-5.02	-5.70	0.504	0.664	1.642	4.652	0.866	0.999
IPDiff	0.818	0.315	20026.5	-3.39	-5.07	0.490	0.609	3.342	4.766	0.891	1.000
PocketXMol	0.861	0.810	127.9	-4.63	-5.86	0.520	0.741	1.174	4.664	0.858	0.996
FLOWR.root	0.972	0.797	102.3	-5.14	-6.44	0.513	0.728	1.245	4.599	0.732	0.944
ODesign	0.990	0.423	1770.3	-3.35	-5.52	0.471	0.645	1.139	4.563	0.842	0.976
DrugFlow	0.840	0.742	525.0	-5.02	-5.85	0.568	0.741	1.863	4.747	0.855	0.998
FLEXIDD	0.979	0.916	76.8	-5.39	-6.67	0.550	0.658	1.185	4.761	0.871	0.980

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Benchmark: We construct a paired apo-holo benchmark from PLINDER (Durairaj et al., 2024), retaining complexes released after the December 30, 2022 training cutoff (Sec. 3.3). Each system pairs an experimental apo protein with its corresponding holo protein-ligand complex, with the apo and holo proteins sharing the same sequence. We further require each test protein to have less than 80% sequence similarity to any training protein. The resulting benchmark contains 330 paired apo-holo systems, with a mean maximum training-set sequence similarity of 46.5% and 25.8% of systems are below 30%. Further benchmark details are provided in Apx. A.10.

Baseline Methods: We compare FLEXIDD with both rigid-protein and flexible SBDD methods. Rigid-protein baselines include Pocket2Mol (Peng et al., 2022), TargetDiff (Guan et al., 2023), DecompDiff (Guan et al., 2024), IPDiff (Huang et al., 2024b), PocketXMol (Peng et al., 2026), FLOWR.root (Cremer et al., 2026b), ODesign (Zhang et al., 2025a), and DrugFlow (Schneuing et al., 2025). Flexible baselines include FlexFlow (Schneuing et al., 2025), Apo2Mol (Zheng et al., 2026), and YuelDesign (Wang et al., 2026). Rigid-protein methods are evaluated with both apo and holo inputs, while flexible baselines are compared under the apo setting. Implementation and configuration details of all baseline methods are provided in Apx. A.9.

Evaluation Metrics: We evaluate generated ligands in terms of validity, geometric plausibility, protein-ligand compatibility, molecular properties, and diversity. Validity measures the fraction of chemically valid and connected molecules, while PB-validity further requires passing all PoseBusters checks (Buttenschoen et al., 2024). Strain quantifies the MMFF94s energy drop upon local minimization of the generated conformation (Halgren, 1999). We report the AutoDock Vina score of the generated pose (Vina) and the score after local in-pocket minimization (Vina Min) (Eberhardt et al., 2021). QED measures drug-likeness (Bickerton et al., 2012), SA measures synthetic accessibility (Ertl & Schuffenhauer, 2009), LogP characterizes lipophilicity (Wildman & Crippen, 1999), and Lipinski summarizes compliance with common drug-likeness rules (Lipinski et al., 2001). Div.2D and Uniq.2D quantify 2D structural diversity and uniqueness, respectively. Detailed metric definitions and evaluation protocols are provided in Apx. B.1.

Research Questions: Our experiments are designed to answer the following research questions: **(Q1)** How does joint protein-ligand generation compare with rigid-protein SBDD under holo inputs? **(Q2)** Can FLEXIDD generalize to apo inputs with only holo supervision? **(Q3)** Are the generated protein conformations structurally plausible while preserving overall protein consistency? **(Q4)** Can the generated pockets support structure-based ligand design by independent models without retraining?

4.2 HOLO-CONDITIONED LIGAND GENERATION

To answer **Q1**, we evaluate ligand generation under holo conditioning. All methods receive the crystal holo protein, while FLEXIDD additionally adapts the protein conformation during ligand generation. As shown in Tab. 2, FLEXIDD maintains a high validity of 97.9% and achieves the strongest performance on structure-sensitive metrics. Its PB-validity reaches 91.6%, exceeding the second-best PocketXMol by 10.6 percentage points, while attaining the lowest strain energy of 76.8.

Table 3: Performance comparison on apo protein structures. Holo Min denotes the Vina Min score of the generated ligand evaluated in the corresponding crystal holo protein.

Category	Method	Validity↑	PB-valid.↑	Strain↓	Vina↓	Vina Min↓	Holo Min↓	QED↑	SA↑	LogP	Lipinski↑	Div.2D↑	Uniq.2D↑
Reference		1.000	0.976	38.5	-6.99	-7.33	N/A	0.492	0.751	1.211	4.679	N/A	1.000
Rigid	Pocket2Mol	0.982	0.621	141.4	-2.50	-4.48	-3.90	0.609	0.793	1.033	4.912	0.857	0.985
	TargetDiff	0.756	0.462	2102.2	-0.15	-2.20	-3.60	0.466	0.634	1.105	4.632	0.892	0.999
	DecompDiff	0.733	0.492	16107.3	-1.55	-3.15	-4.26	0.483	0.655	1.373	4.610	0.875	1.000
	IPDiff	0.742	0.283	17096.0	-1.38	-3.32	-3.46	0.483	0.606	3.302	4.783	0.889	1.000
	PocketXMol	0.783	0.783	124.1	-3.78	-5.01	-4.44	0.518	0.725	0.836	4.676	0.867	0.996
	FLOWR.root	0.968	0.573	96.8	6.72	-2.18	-5.15	0.519	0.717	1.033	4.636	0.787	0.965
	ODesign	0.990	0.386	3078.6	-2.37	-4.83	-2.72	0.478	0.638	1.186	4.560	0.839	0.970
	DrugFlow	0.776	0.527	1704.0	4.39	-1.16	-4.26	0.561	0.730	1.838	4.775	0.866	0.998
Flexible	FlexFlow	0.765	0.461	1629.5	2.34	-2.82	-3.98	0.561	0.756	1.506	4.761	0.852	1.000
	Apo2Mol	0.915	0.342	697.7	-2.62	-4.09	-3.86	0.501	0.596	3.202	4.712	0.866	0.999
	YuelDesign	0.435	0.305	3704.0	-1.88	-3.98	-2.36	0.238	0.577	0.409	3.701	0.891	1.000
	FLEXIDD	0.976	0.916	59.4	-5.02	-6.42	-6.53	0.545	0.665	1.124	4.836	0.873	0.985

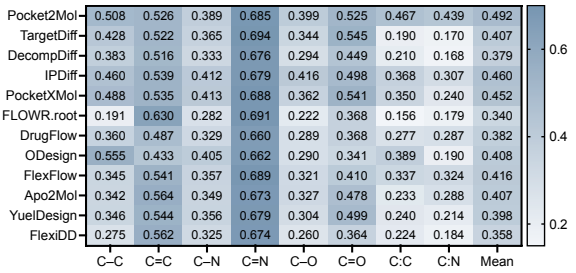

Figure 3: Bond-length distribution fidelity. Jensen-Shannon divergence (JSD) between bond-length distributions of generated and reference ligands. "-", "=", and ":" denote single, double, and aromatic bonds, respectively.

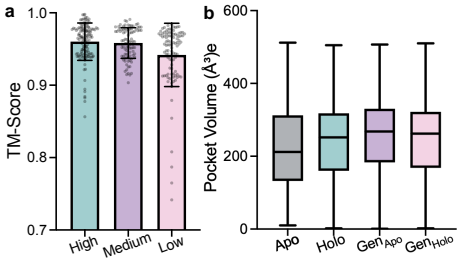

Figure 4: Structural analysis of generated proteins. **(a)** TM-score between generated and reference holo proteins. **(b)** Pocket-volume distributions of experimental apo and holo proteins and generated proteins.

FLEXIDD also achieves the most favorable Vina and Vina Min scores of -5.39 and -6.67, respectively. These consistent gains suggest that joint protein adaptation reduces geometric mismatch between the pocket and ligand, improving complex plausibility and binding quality without compromising chemical validity or molecular diversity.

We further examine whether these gains preserve the intrinsic geometry of the generated ligands. As shown in Fig. 3, FLEXIDD achieves a mean bond-length Jensen–Shannon divergence (JSD) of 0.358, ranking second overall after FLOWR.root at 0.340. Fig. 9 further compares all-atom pairwise distance distributions between generated and reference ligands. FLEXIDD obtains a JSD of 0.072, close to DrugFlow at 0.070 though above DecompDiff and FLOWR.root at 0.054. Together with the bond-length analysis, these results indicate that the complex-level gains of FLEXIDD do not come at the expense of ligand geometry, which remains competitive with the strongest baselines. Median results are reported in Apx. B.2.

4.3 GENERALIZATION TO APO PROTEIN INPUTS

To answer **Q2**, we evaluate all methods on apo protein inputs. As shown in Tab. 3, rigid SBDD methods struggle to generalize to apo inputs, as their fixed protein coordinates cannot accommodate the conformational mismatch; for example, FLOWR.root drops from 79.7% PB-validity under holo conditioning to 57.3% with apo inputs. Flexible baselines mitigate this limitation but remain constrained by restricted conformational degrees of freedom or pocket-only modeling, with the best PB-validity reaching only 46.1%. In contrast, FLEXIDD achieves 91.6% PB-validity, the lowest strain energy of 59.4, and the most favorable Vina and Vina Min scores of -5.02 and -6.42, while remaining close to its holo-conditioned performance. To disentangle ligand quality from protein conformation, we re-evaluate the generated ligands using Vina Min in the corresponding crystal holo proteins (denoted as Holo Min): FLEXIDD achieves the best Holo Min of -6.53, compared with -5.15 for the strongest baseline. These results confirm that combining noise perturbation with full-protein all-atom adaptation enables protein denoising learned from holo complexes to generalize effectively to apo inputs.

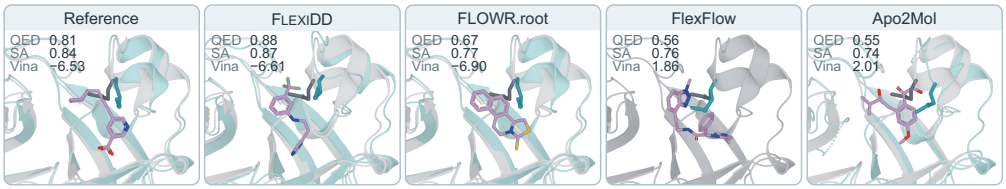

Figure 5: Qualitative comparison of ligand poses and local pocket conformations for PDB 7GRR. The apo structure (gray) is superimposed with the reference holo structure or the structure generated by each method (cyan). Ligands are shown as purple sticks. The Met49 side chains are highlighted as sticks to visualize conformational differences near the binding site. For FlexFlow, cyan largely overlaps with the gray apo structure since only pocket side-chain torsions are allowed to change while the protein backbone remains fixed.

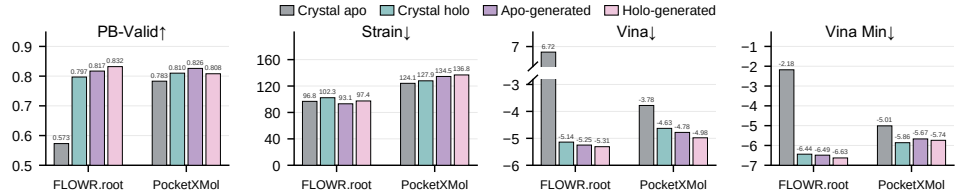

Figure 6: Transferable utility of FLEXIDD-generated pockets. Two independently trained ligand generators are conditioned on crystal and FLEXIDD-generated pockets without re-training.

4.4 STRUCTURAL CONSISTENCY OF GENERATED RECEPTORS

To answer **Q3**, we examine whether FLEXIDD preserves global protein structure while locally adapting pocket geometry. We further stratify the benchmark into high-, medium-, and low-similarity subsets by the maximum sequence similarity between each test protein and the training set (Apx. A.10.2). As shown in Fig. 4(a), the generated proteins remain highly similar to their matched holo structures, with mean TM-scores (Zhang & Skolnick, 2004) of 0.960, 0.958, and 0.942 on the three subsets, respectively. At the pocket level, we compute pocket volumes using POVME 3 (Wagner et al., 2017). Generated volumes shift toward the holo distribution: for apo-generated proteins, the JSD to holo structures is 0.089, compared with 0.164 to apo structures; for holo-generated proteins, the corresponding values are 0.064 and 0.120. Fig. 5 illustrates a representative case in PDB 7GRR, where FLEXIDD recovers the holo-like opening of the Met49 side chain from the apo protein alongside a compatible ligand pose. Overall, FLEXIDD maintains global structural integrity while driving local pocket geometry toward holo-like conformations.

4.5 TRANSFERABLE UTILITY OF GENERATED POCKETS

To answer **Q4**, we condition two independently trained ligand generators, FLOWR.root and PocketXMol, on FLEXIDD-generated pockets without modifying their parameters, and compare against their performance on crystal apo and crystal holo structures. As in Fig. 6 and Tab. 12, conditioning on crystal apo proteins degrades ligand quality: FLOWR.root PB-validity falls to 57.3% vs. 79.7% with crystal holo. In contrast, conditioning on FLEXIDD-generated pockets from apo or holo inputs largely recovers holo-level quality (FLOWR.root PB-validity 81.7% and 83.2%), with similar trends for PocketXMol. These results suggest that FLEXIDD-generated pockets retain enough binding-site geometry to remain compatible with the holo distribution that these generators were trained on.

5 CONCLUSION

We introduce FLEXIDD, a flexible SBDD framework that jointly generates ligands and refines full-protein conformations. Transferable protein coordinate denoising enables training on large-scale holo complexes without paired apo-holo supervision and generalization to apo inputs, while full-protein modeling supports globally consistent protein adaptation. FLEXIDD improves PB-validity and Vina scores over rigid-protein baselines under holo conditioning and substantially outperforms existing flexible SBDD methods under apo inputs. Generated pockets also transfer effectively to independently trained ligand generators, indicating broader applicability.

AI USE STATEMENT

In this work, we used generative AI tools to improve the organization, clarity, and readability of the manuscript, and to assist with qualitative data analysis and interpretation of results. We did not use generative AI tools to generate synthetic datasets, develop theoretical models, formulate or prove mathematical claims, propose or refine hypotheses, design experiments, implement the proposed method, or clean and reformat datasets. All AI-assisted content was critically reviewed and revised by the authors before submission. Experimental designs and interpretations were independently assessed by the authors, and all reported numerical results were obtained from our experimental pipeline and verified by the authors. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We have taken several steps to facilitate reproducibility. The formulation and overall design of FLEXIDD are described in Sec. 3, while detailed implementation and reproducibility information, including architectural specifications, loss weights, sampling configurations, data preparation and splitting procedures, and optimization settings, is provided in Apx. A.2, Apx. A.4, Apx. A.5, and Apx. A.8. The benchmark construction, evaluation protocol, and definitions of all reported metrics are described in Sec. 4 and Apx. B.1. The source code will be publicly released upon completion of the review process.

REFERENCES

Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, et al. Accurate structure prediction of biomolecular interactions with alphafold 3. Nature, 630(8016):493–500, 2024.

Rishal Aggarwal, Akash Gupta, and U Priyakumar. Apobind: a dataset of ligand unbound protein conformations for machine learning applications in de novo drug design. arXiv preprint arXiv:2108.09926, 2021.

Gustaf Ahdritz, Nazim Bouatta, Christina Floristean, Sachin Kadyan, Qinghui Xia, William Gerecke, Timothy J O'Donnell, Daniel Berenberg, Ian Fisk, Niccolò Zanichelli, et al. Openfold: retraining alphafold2 yields new insights into its learning mechanisms and capacity for generalization. Nature methods, 21(8):1514–1524, 2024.

Helen M Berman, John Westbrook, Zukang Feng, Gary Gilliland, Talapady N Bhat, Helge Weissig, Ilya N Shindyalov, and Philip E Bourne. The protein data bank. Nucleic acids research, 28(1): 235–242, 2000.

G Richard Bickerton, Gaia V Paolini, Jérémy Besnard, Sorel Muresan, and Andrew L Hopkins. Quantifying the chemical beauty of drugs. Nature chemistry, 4(2):90–98, 2012.

David D Boehr, Ruth Nussinov, and Peter E Wright. The role of dynamic conformational ensembles in biomolecular recognition. Nature chemical biology, 5(11):789–796, 2009.

Michal Brylinski and Jeffrey Skolnick. What is the relationship between the global structures of apo and holo proteins? Proteins: Structure, Function, and Bioinformatics, 70(2):363–377, 2008.

Stephen K Burley, Helen M Berman, Gerard J Kleywegt, John L Markley, Haruki Nakamura, and Sameer Velankar. Protein data bank (pdb): the single global macromolecular structure archive. Protein crystallography: methods and protocols, pp. 627–641, 2017.

Martin Buttenschoen, Garrett M Morris, and Charlotte M Deane. Posebusters: Ai-based docking methods fail to generate physically valid poses or generalise to novel sequences. Chemical science, 15(9):3130–3139, 2024.

Salvatore Candido, Thomas Hayes, Alexander Derry, Roshan Rao, Zeming Lin, Robert Verkuil, Bryan Z Wu, Jin Sub Lee, Elise S Bruguera, Jehan A Keval, et al. Language modeling materializes a world model of protein biology. bioRxiv, pp. 2026–06, 2026.

Ting Chen, Ruixiang Zhang, and Geoffrey Hinton. Analog bits: Generating discrete data using diffusion models with self-conditioning. arXiv preprint arXiv:2208.04202, 2022.

Jordan J Clark, Mark L Benson, Richard D Smith, and Heather A Carlson. Inherent versus induced protein flexibility: Comparisons within and between apo and holo structures. PLoS computational biology, 15(1):e1006705, 2019.

Gabriele Corso, Vignesh Ram Somnath, Noah Getz, Regina Barzilay, Tommi Jaakkola, and Andreas Krause. Composing unbalanced flows for flexible docking and relaxation. In International Conference on Learning Representations, volume 2025, pp. 27699–27722, 2025.

Julian Cremer, Tuan Le, Frank Noé, Djork-Arné Clevert, and Kristof T Schütt. Pilot: equivariant diffusion for pocket-conditioned de novo ligand generation with multi-objective guidance via importance sampling. Chemical Science, 15(36):14954–14967, 2024.

Julian Cremer, Ross Irwin, Alessandro Tibo, Jon Paul Janet, Simon Olsson, and Djork-Arné Clevert. Flowr: flow matching for structure-aware de novo, interaction-and fragment-based ligand generation. Nature Computational Science, pp. 1–10, 2026a.

Julian Cremer, Tuan Le, Mohammad M Ghahremanpour, Emilia Sługocka, Filipe Menezes, and Djork-Arné Clevert. Flowr. root–a flow matching-based foundation model for joint multi-purpose structure-aware 3d ligand generation and affinity prediction. Nature Communications, 17(1):5883, 2026b.

Giannis Daras, Jeffrey Ouyang-Zhang, Krithika Ravishankar, Constantinos Daskalakis, Adam Klivans, and Daniel Diaz. Ambient proteins-training diffusion models on noisy structures. Advances in Neural Information Processing Systems, 38:65530–65558, 2026.

Janani Durairaj, Yusuf Adeshina, Zhonglin Cao, Xuejin Zhang, Vladas Oleinikovas, Thomas Duignan, Zachary McClure, Xavier Robin, Gabriel Studer, Daniel Kovtun, et al. Plinder: The protein-ligand interactions dataset and evaluation resource. BioRxiv, pp. 2024–07, 2024.

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, 2021.

Peter Ertl and Ansgar Schuffenhauer. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. Journal of cheminformatics, 1(1):8, 2009.

Shikun Feng, Bicheng Lin, Yuanhuan Mo, Yuyan Ni, Wenyu Zhu, Bowen Gao, Wei-Ying Ma, Yanyan Lan, et al. Figrdock: Fast interaction-guided regression for flexible docking. Advances in Neural Information Processing Systems, 38:85757–85776, 2026.

Itai Gat, Tal Remez, Neta Shaul, Felix Kreuk, Ricky TQ Chen, Gabriel Synnaeve, Yossi Adi, and Yaron Lipman. Discrete flow matching. Advances in Neural Information Processing Systems, 37: 133345–133385, 2024.

Siyi Gu, Minkai Xu, Alexander Powers, Weili Nie, Tomas Geffner, Karsten Kreis, Jure Leskovec, Arash Vahdat, and Stefano Ermon. Aligning target-aware molecule diffusion models with exact energy optimization. Advances in neural information processing systems, 37:44040–44063, 2024.

Jiaqi Guan, Wesley Wei Qian, Xingang Peng, Yufeng Su, Jian Peng, and Jianzhu Ma. 3d equivariant diffusion for target-aware molecule generation and affinity prediction. arXiv preprint arXiv:2303.03543, 2023.

Jiaqi Guan, Xiangxin Zhou, Yuwei Yang, Yu Bao, Jian Peng, Jianzhu Ma, Qiang Liu, Liang Wang, and Quanquan Gu. Decompdiff: diffusion models with decomposed priors for structure-based drug design. arXiv preprint arXiv:2403.07902, 2024.

Thomas A Halgren. Mmff vi. mmff94s option for energy minimization studies. Journal of computational chemistry, 20(7):720–729, 1999.

Yufei Huang, Odin Zhang, Lirong Wu, Cheng Tan, Haitao Lin, Zhangyang Gao, Siyuan Li, Stan Li, et al. Re-dock: towards flexible and realistic molecular docking with diffusion bridge. arXiv preprint arXiv:2402.11459, 2024a.

Zhilin Huang, Ling Yang, Xiangxin Zhou, Zhilong Zhang, Wentao Zhang, Xiawu Zheng, Jie Chen, Yu Wang, Bin Cui, and Wenming Yang. Protein-ligand interaction prior for binding-aware 3d molecule diffusion models. In International Conference on Learning Representations, volume 2024, pp. 18453–18474, 2024b.

Ross Irwin, Alessandro Tibo, Jon Paul Janet, and Simon Olsson. Semlaflow–efficient 3d molecular generation with latent attention and equivariant flow matching. arXiv preprint arXiv:2406.07266, 2024.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. nature, 596(7873):583–589, 2021.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. Advances in neural information processing systems, 35:26565–26577, 2022.

Rohith Krishna, Jue Wang, Woody Ahern, Pascal Sturmfels, Preetham Venkatesh, Indrek Kalvet, Gyu Rie Lee, Felix S Morey-Burrows, Ivan Anishchenko, Ian R Humphreys, et al. Generalized biomolecular modeling and design with rosettafold all-atom. Science, 384(6693):eadl2528, 2024.

Christopher A Lipinski, Franco Lombardo, Beryl W Dominy, and Paul J Feeney. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Advanced drug delivery reviews, 46(1-3):3–26, 2001.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. arXiv preprint arXiv:2210.02747, 2022.

Meng Liu, Youzhi Luo, Kanji Uchino, Koji Maruhashi, and Shuiwang Ji. Generating 3d molecules for target protein binding. arXiv preprint arXiv:2204.09410, 2022.

Wei Lu, Jixian Zhang, Weifeng Huang, Ziqiao Zhang, Xiangyu Jia, Zhenyu Wang, Leilei Shi, Chengtao Li, Peter G Wolynes, and Shuangjia Zheng. Dynamicbind: predicting ligand-specific protein-ligand complex structure with a deep equivariant generative model. Nature Communications, 15 (1):1071, 2024.

Shitong Luo, Jiaqi Guan, Jianzhu Ma, and Jian Peng. A 3d generative model for structure-based drug design. Advances in neural information processing systems, 34:6229–6239, 2021.

Valerio Mariani, Marco Biasini, Alessandro Barbato, and Torsten Schwede. lddt: a local superposition-free score for comparing protein structures and models using distance difference tests. Bioinformatics, 29(21):2722–2728, 2013.

Alex Morehead and Jianlin Cheng. Flowdock: Geometric flow matching for generative protein–ligand docking and affinity prediction. Bioinformatics, 41(Supplement_1):i198–i206, 2025.

Xingang Peng, Shitong Luo, Jiaqi Guan, Qi Xie, Jian Peng, and Jianzhu Ma. Pocket2mol: Efficient molecular sampling based on 3d protein pockets. In International conference on machine learning, pp. 17644–17655. PMLR, 2022.

Xingang Peng, Ruihan Guo, Fenglin Guo, Ziyi Wang, Jiayu Sun, Jiaqi Guan, Yinjun Jia, Yan Xu, Yanwen Huang, Muhan Zhang, et al. Unified modeling of 3d molecular generation via atomic interactions with pocketxmol. Cell, 189(7):1904–1922, 2026.

Zhuoran Qiao, Weili Nie, Arash Vahdat, Thomas F Miller III, and Animashree Anandkumar. State-specific protein–ligand complex structure prediction with a multiscale deep generative model. Nature Machine Intelligence, 6(2):195–208, 2024.

Keyue Qiu, Yuxuan Song, Zhehuan Fan, Peidong Liu, Zhe Zhang, Mingyue Zheng, Hao Zhou, and Wei-Ying Ma. Piloting structure-based drug design via modality-specific optimal schedule. arXiv preprint arXiv:2505.07286, 2025.

Yanru Qu, Keyue Qiu, Yuxuan Song, Jingjing Gong, Jiawei Han, Mingyue Zheng, Hao Zhou, and Wei-Ying Ma. Molcraft: structure-based drug design in continuous parameter space. arXiv preprint arXiv:2404.12141, 2024.

Matthew Ragoza, Tomohide Masuda, and David Ryan Koes. Generating 3d molecules conditional on receptor binding sites with deep generative models. Chemical science, 13(9):2701–2713, 2022.

David Rogers and Mathew Hahn. Extended-connectivity fingerprints. Journal of chemical information and modeling, 50(5):742–754, 2010.

Arne Schneuing, Charles Harris, Yuanqi Du, Kieran Didi, Arian Jamasb, Ilia Igashov, Weitao Du, Carla Gomes, Tom L Blundell, Pietro Lio, et al. Structure-based drug design with equivariant diffusion models. Nature Computational Science, 4(12):899–909, 2024.

Arne Schneuing, Ilia Igashov, Adrian Dobbelstein, Thomas Castiglione, Michael Bronstein, and Bruno Correia. Multi-domain distribution learning for de novo drug design. In International Conference on Learning Representations, volume 2025, pp. 84997–85033, 2025.

Kristof Schütt, Pieter-Jan Kindermans, Huziel Enoc Sauceda Felix, Stefan Chmiela, Alexandre Tkatchenko, and Klaus-Robert Müller. Schnet: A continuous-filter convolutional neural network for modeling quantum interactions. Advances in neural information processing systems, 30, 2017.

Martin Steinegger and Johannes Söding. Mmseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nature biotechnology, 35(11):1026–1028, 2017.

Simon J Teague. Implications of protein flexibility for drug discovery. Nature reviews Drug discovery, 2(7):527–541, 2003.

Jeffrey R Wagner, Jesper Sørensen, Nathan Hensley, Celia Wong, Clare Zhu, Taylor Perison, and Rommie E Amaro. Povme 3.0: software for mapping binding pocket flexibility. Journal of chemical theory and computation, 13(9):4584, 2017.

Jian Wang, Dong Yan Zhang, Shreshty Budakoti, and Nikolay V Dokholyan. A diffusion-based framework for designing molecules in flexible protein pockets. Science Advances, 12(15):eaeb7045, 2026.

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(5):1209–1220, 2025.

Jinhang Wei, Yupu Zhang, Peter A Ramdhan, Zihang Huang, Gustavo Seabra, Zhe Jiang, Chenglong Li, and Yanjun Li. Gatoraffinity: Boosting protein-ligand binding affinity prediction with large-scale synthetic structural data. bioRxiv, 2025.

Scott A Wildman and Gordon M Crippen. Prediction of physicochemical parameters by atomic contributions. Journal of chemical information and computer sciences, 39(5):868–873, 1999.

Jeremy Wohlwend, Gabriele Corso, Saro Passaro, Noah Getz, Mateo Reveiz, Ken Leidal, Wojtek Swiderski, Liam Atkinson, Tally Portnoi, Itamar Chinn, et al. Boltz-1 democratizing biomolecular interaction modeling. BioRxiv, pp. 2024–11, 2025.

Odin Zhang, Xujun Zhang, Haitao Lin, Cheng Tan, Qinghan Wang, Yuanle Mo, Qiantai Feng, Gang Du, Yuntao Yu, Zichang Jin, et al. Odesign: A world model for biomolecular interaction design. arXiv preprint arXiv:2510.22304, 2025a.

Runze Zhang, Xinyu Jiang, Duanhua Cao, Zhaokun Wang, Jie Yu, Mingan Chen, Zhehuan Fan, Xiangtai Kong, Jiacheng Xiong, Zimei Zhang, et al. Flexible protein–ligand docking with diffusion-based side-chain packing. Proceedings of the National Academy of Sciences, 122(52):e2511925122, 2025b.

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.

Zaixi Zhang, Mengdi Wang, and Qi Liu. Flexsbdd: Structure-based drug design with flexible protein modeling. Advances in neural information processing systems, 37:53918–53944, 2024.

Zaixi Zhang, Jiaxian Yan, Yining Huang, Qi Liu, Enhong Chen, Mengdi Wang, and Marinka Zitnik. Structure-based drug design with geometric deep learning: A comprehensive survey. ACM Computing Surveys, 58(5):1–35, 2025c.

Zizhuo Zhang, Lijun Wu, Kaiyuan Gao, Jiangchao Yao, Tao Qin, and Bo Han. Fast and accurate blind flexible docking. In International Conference on Learning Representations, volume 2025, pp. 14542–14566, 2025d.

Xinzhe Zheng, Shiyu Jiang, Gustavo Seabra, Chenglong Li, and Yanjun Li. Apo2mol: 3d molecule generation via dynamic pocket-aware diffusion models. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 40, pp. 1614–1622, 2026.

Xiangxin Zhou, Yi Xiao, Haowei Lin, Xinheng He, Jiaqi Guan, Yang Wang, Qiang Liu, Feng Zhou, Liang Wang, and Jianzhu Ma. Integrating protein dynamics into structure-based drug design via full-atom stochastic flows. In International Conference on Learning Representations, volume 2025, pp. 18621–18649, 2025.

Jintao Zhu, Zhonghui Gu, Jianfeng Pei, and Luhua Lai. Diffbindfr: an se (3) equivariant network for flexible protein–ligand docking. Chemical Science, 15(21):7926–7942, 2024.

APPENDIX

A Extended Methods **16**

A.1 Gaussian Corruption and Conformational Distribution Shift 16

A.2 Network Architecture Details 17

A.3 Equivariance of the Deterministic Predictor 20

A.4 Training Objective Definitions 22

A.5 Noise Schedule and Sampling 23

A.6 Training Algorithm 25

A.7 Public Apo-Holo Paired Data 25

A.8 Training Data 25

A.9 Baseline Implementation Details 27

A.10 PLINDER Apo-Holo Test Benchmark 27

B Extended Results **29**

B.1 Evaluation Protocol and Metrics 29

B.2 Median Performance with Holo and Apo Inputs 30

B.3 All-Atom Pairwise-Distance Distributions 31

B.4 Detailed Pocket Transfer Results 33

B.5 Sampling Trajectories and Pocket Views 34

B.6 Stage 1 Pretraining Ablation 38

C Limitations **38**

A Extended Methods

A.1 Gaussian Corruption and Conformational Distribution Shift

This section derives the KL bound stated in Eq. 2 and makes its assumptions explicit, following the distribution-merging argument of Ambient Proteins (Daras et al., 2026).

A.1.1 Assumptions and Notation

Throughout this section, we consider a single protein with fixed identity, atom correspondences, and alignment frame. All other conditions, including site labels, fusion indices, and local reference features, are held fixed. We represent all-atom coordinates as vectors in $\mathbb{R}^d$ with $d = 3n_{\mathcal{P}}$. Let μ and ν be coordinate distributions with finite second moments, and $\gamma_\sigma = \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_d)$. Independent Gaussian noise gives $\mu_\sigma = \mu * \gamma_\sigma$ and $\nu_\sigma = \nu * \gamma_\sigma$, where $*$ denotes convolution. Define

$$W_2^2(\mu, \nu) = \inf_{\pi \in \Pi(\mu,\nu)} \int \|\mathbf{x} - \mathbf{y}\|_2^2 \mathrm{d}\pi(\mathbf{x}, \mathbf{y}), \tag{13}$$

where $\Pi(\mu, \nu)$ contains all joint probability measures with marginals μ and ν.

A.1.2 Contraction and a Quantitative Bound

Proposition 1 (contraction with increasing noise): For $0 < \sigma_1 \leq \sigma_2$,

$$D_{\mathrm{KL}}(\mu_{\sigma_2} \| \nu_{\sigma_2}) \leq D_{\mathrm{KL}}(\mu_{\sigma_1} \| \nu_{\sigma_1}). \tag{14}$$

Proof. For $\tau = (\sigma_2^2 - \sigma_1^2)^{1/2}$, $\gamma_{\sigma_2} = \gamma_{\sigma_1} * \gamma_\tau$. Applying this common noise channel to μ_{σ_1} and ν_{σ_1} contracts KL divergence by data processing; $\tau = 0$ gives equality. □

Proposition 2 (Wasserstein upper bound): For every $\sigma > 0$,

$$D_{\mathrm{KL}}(\mu_\sigma \| \nu_\sigma) \leq \frac{W_2^2(\mu, \nu)}{2\sigma^2}. \tag{15}$$

Proof. For any $\pi \in \Pi(\mu, \nu)$, both convolved measures are mixtures under the same joint measure:

$$\mu_\sigma = \int \mathcal{N}(\mathbf{x}, \sigma^2 \mathbf{I}_d) \mathrm{d}\pi(\mathbf{x}, \mathbf{y}), \quad \nu_\sigma = \int \mathcal{N}(\mathbf{y}, \sigma^2 \mathbf{I}_d) \mathrm{d}\pi(\mathbf{x}, \mathbf{y}). \tag{16}$$

Joint convexity of KL divergence and the equal-covariance Gaussian formula give

$$\begin{aligned} D_{\mathrm{KL}}(\mu_\sigma \| \nu_\sigma) &\leq \int D_{\mathrm{KL}} \left(\mathcal{N}(\mathbf{x}, \sigma^2 \mathbf{I}_d) \| \mathcal{N}(\mathbf{y}, \sigma^2 \mathbf{I}_d) \right) \mathrm{d}\pi(\mathbf{x}, \mathbf{y}) \\ &= \frac{1}{2\sigma^2} \int \|\mathbf{x} - \mathbf{y}\|_2^2 \mathrm{d}\pi(\mathbf{x}, \mathbf{y}). \end{aligned} \tag{17}$$

Taking the infimum over π proves Eq. 15. □

Setting $\mu = p^{\mathrm{apo}}$ and $\nu = p^{\mathrm{holo}}$ recovers the bound in Eq. 2, and interchanging the two distributions gives the reverse direction. Both KL divergences vanish as $\sigma \to \infty$, though a finite noise scale does not guarantee that the gap is negligibly small.

Two aligned structures: For fixed conformations, the corrupted laws are Gaussian and their divergence is

$$D_{\mathrm{KL}} \left(q_\sigma^{\mathrm{apo}} \,\|\, q_\sigma^{\mathrm{holo}} \right) = \frac{\left\| \mathbf{X}_{\mathrm{apo}}^{\mathcal{P}} - \mathbf{X}_{\mathrm{holo}}^{\mathcal{P}} \right\|_F^2}{2\sigma^2}. \tag{18}$$

For a fixed pair of conformations, the divergence therefore decreases as σ^{-2}. This provides a quantitative motivation for using Gaussian corruption to reduce the discrepancy between apo and holo inputs. The remaining discrepancy at a finite noise scale depends on the full coordinate displacement between the structures.

A.1.3 Implication for a Shared Sampler

We now extend the bound to the full generation pipeline. For a fixed binding site and ligand size, let $p_0^{\mathcal{M}}$ be a ligand coordinate prior that is independent of the protein initialization, and assume the sampler accesses the protein coordinates only through the corrupted initialization. Its complete kernel K_θ includes deterministic coordinate steps and stochastic chemical decoding. Data processing gives

$$\begin{aligned} D_{\mathrm{KL}}\left((\mu_\sigma \otimes p_0^{\mathcal{M}})K_\theta \| (\nu_\sigma \otimes p_0^{\mathcal{M}})K_\theta\right) &\le D_{\mathrm{KL}}(\mu_\sigma \| \nu_\sigma) \\ &\le \frac{W_2^2(\mu,\nu)}{2\sigma^2}. \end{aligned} \tag{19}$$

Because the ligand prior is shared and independent of the protein coordinates, the tensor product does not increase the KL divergence. This result bounds the sensitivity of the generated distribution to the choice of input conformation, but does not guarantee that the output approximates the true bound-complex distribution or recovers any particular holo structure. The bound further assumes identical cropping, site annotations, and fusion selection across inputs; in practice, these may differ. Finally, unlike Ambient Proteins, which trains on noisy observations, FLEXIDD reconstructs reference structures (including synthetic ones) and uses a finite noise range, both of which are modeling choices validated empirically in Sec. 4.

A.2 Network Architecture Details

The denoising network f_θ in Eq. 7 receives the Pairformer context $(\mathbf{s}, \mathbf{z})$ together with noisy protein coordinates, ligand states, and the diffusion time. It consists of a protein module, a ligand module, and an interaction module that are interleaved over G rounds as described in Eq. 8. Tab. 4 summarizes module dimensions.

A.2.1 Inputs and Notation

Let $n_{\mathcal{P}}$, n_R, and $n_{\mathcal{M}}$ denote the numbers of protein atoms, residues, and ligand atom slots, respectively. Each atom i is assigned to its parent residue $\pi(i)$, and one representative atom defines each residue position $\mathbf{r}_j$. Time-invariant inputs comprise residue and atom identities, reference charges, bond connectivity, chain and residue indices, pocket-site labels, and chemical-component conformers expressed in local reference frames. The diffusion time is denoted as t.

A.2.2 Protein Module

The protein module converts noisy atom coordinates into residue-level scalar and vector features through four stages: (1) an atom encoder extracts local geometric descriptors, (2) mean-pooling lifts atom features to residue tokens, (3) time conditioning injects the diffusion step, and (4) residue self-attention layers refine the representation.

Atom encoder. For each protein atom i with noisy position $\mathbf{x}_i$, an invariant descriptor aggregates local geometric information:

$$\mathbf{d}_i = \left[\, \|\mathbf{x}_i - \mathbf{r}_{\pi(i)}\|,\ \mathrm{RBF}_8\big(\|\mathbf{x}_i - \mathbf{r}_{\pi(i)}\|\big),\ \textstyle\sum_{j\in\mathcal{W}_i} \mathrm{RBF}_{16}\big(\|\mathbf{x}_j - \mathbf{x}_i\|\big),\ n_i/|\mathcal{W}_i| \,\right], \tag{20}$$

where $\mathcal{W}_i$ is a local attention window, RBF_k denotes k equally spaced Gaussian radial basis functions (Schütt et al., 2017), and n_i counts valid atoms in the window. A normalized MLP maps $\mathbf{d}_i$ to scalar features, which are refined by atom-attention layers. The encoder retains atom scalar features $\mathbf{Q}^A$ for the output head and the pocket fusion adapter. Equivariant vector channels capture directional information from two geometric sources:

$$\mathbf{V}_i^A = W_{\mathrm{rep}} \frac{\mathbf{x}_i - \mathbf{r}_{\pi(i)}}{\|\mathbf{x}_i - \mathbf{r}_{\pi(i)}\| + \epsilon} + W_{\mathrm{nbr}}\, \boldsymbol{\mu}_i, \tag{21}$$

where $W_{\mathrm{rep}}, W_{\mathrm{nbr}}$ are bias-free linear maps, $\boldsymbol{\mu}_i$ averages radial-weighted displacement vectors from neighboring atoms, and each channel is normalized by its mean length over valid atoms. These vector channels are preserved for the coordinate prediction head.

Residue initialization. Atom scalar features are mean-pooled within each residue and concatenated with $\mathbf{s}$ to initialize residue scalars $\mathbf{h}^R$. Centered representative coordinates initialize vector channels $\mathbf{V}^R$ through a scaled, bias-free projection.

Time conditioning. A Fourier embedding of the diffusion time t conditions both the scalar and vector states:

$$\mathbf{h}^R \mathrel{+}= \mathrm{MLP}_s\big(\mathrm{Fourier}(t)\big), \quad \mathbf{V}^R \mathrel{*}= 1 + \tanh\big(\mathrm{MLP}_v\big(\mathrm{Fourier}(t)\big)\big). \tag{22}$$

The scalar branch receives an additive shift while the vector branch receives a multiplicative gate initialized near unity.

Residue layers. The residue layers use the scalar-vector self-attention of Irwin et al. (2024), in which each node maintains both scalar and equivariant vector channels that are jointly updated through attention. Residue edges are initialized from projected Pairformer pair features $\mathbf{z}$ and radial distance features. After every layer, a learned gate reinjects $\mathbf{z}$ into the edges, preventing the pair context from fading with depth. Because the initial vector channels derive from a single direction vector toward each residue center, providing only a rank-one directional signal, each layer also injects a neighbor vector that aggregates direction vectors to surrounding residues, weighted by radial basis functions:

$$\mathbf{V}_j^R \mathrel{+}= W_{\mathrm{inj}} \sum_{r=1}^{R} \frac{\sum_k w_r\big(\|\mathbf{r}_k - \mathbf{r}_j\|\big)\,(\mathbf{r}_k - \mathbf{r}_j)}{\sum_k w_r\big(\|\mathbf{r}_k - \mathbf{r}_j\|\big) + \epsilon}\,, \tag{23}$$

where w_r is the r-th Gaussian radial basis function with centers spanning 0 to 8 Å and W_{inj} is a bias-free projection. The denominator normalizes by total weight rather than neighbor count, which is critical because an unweighted sum over a roughly isotropic neighborhood collapses to the radial direction and carries no new directional information. After the final layer, the refined residue scalars $\mathbf{h}^R$ and vectors $\mathbf{V}^R$ are passed to the interaction module, which mediates information exchange with the ligand (see below).

A.2.3 Ligand Module

The ligand module shares the same scalar-vector attention architecture as the protein residue layers but operates on atom-level tokens. It follows two stages: (1) initialization encodes atom types, bonds, and coordinates into scalar, edge, and vector features with time conditioning, and (2) ligand self-attention layers refine the representation while periodically updating coordinates.

Initialization. Initial atom types, charges, and bonds are sampled uniformly from their respective vocabularies, excluding padding categories. Initial ligand coordinates for $n_{\mathcal{M}}$ real atoms are drawn from a centered Gaussian that removes the mean, so that the molecule starts near the origin:

$$\mathbf{Z}^{\mathcal{M}} = \left(\mathbf{I}_{n_{\mathcal{M}}} - \frac{\mathbf{1}\mathbf{1}^\top}{n_{\mathcal{M}}}\right)\mathbf{G}, \quad G_{ij} \overset{\mathrm{iid}}{\sim} \mathcal{N}(0,1). \tag{24}$$

From these raw inputs, ligand scalar features embed atom types, charges, diffusion time, atom count, and self-conditioning atom-type probabilities. Bond embeddings, self-conditioning bond probabilities, and radial distance features initialize edge features. Vector channels are initialized by bias-free mixing of current and self-conditioning coordinates, followed by the same multiplicative time gate as in Eq. 22.

Ligand layers. Each layer performs self-attention among ligand atoms and cross-attention to the pocket interface atoms provided by the interaction module (see below). Ligand coordinates are updated at evenly spaced intervals within the layer sequence, with distance features refreshed afterwards. After the final layer, the refined ligand scalars, vectors, and coordinates are passed to the output heads.

A.2.4 Interaction Module

The interaction module mediates information exchange between the protein and ligand modules across the G interleaving rounds of Eq. 8. It has three components: (1) a predetermined set of fusion atoms that serve as the protein-ligand interface, (2) a pocket fusion adapter that gathers updated protein state onto these atoms for the ligand layers to cross-attend to, and (3) a reverse fusion channel that injects ligand context into the Pairformer single representation before the interleaving rounds begin.

Fusion atom selection. Let $\mathcal{A}_S$ contain the atoms of labeled pocket residues, with $\boldsymbol{\mu}_S$ the centroid of their observed heavy atoms in the supplied structure. Fusion candidates are

$$\mathcal{F}_{\mathrm{cand}} = \mathcal{A}_S \cup \Big\{ i : \text{atom } i \text{ is observed and } \|\mathbf{x}^{\mathcal{P},(i)} - \boldsymbol{\mu}_S\|_2 \le 10\,\text{Å} \Big\}. \tag{25}$$

Pocket atoms precede other candidates, ordered respectively by atom index and distance to $\boldsymbol{\mu}_S$. The list is truncated to a set capacity and padded if needed; indices and masks are determined before corruption. All coordinates are internally centered on the fusion-atom centroid and restored before output.

Pocket fusion adapter. After each protein group, the pocket fusion adapter refreshes the interface representation that the subsequent ligand layers attend to. For each fusion atom i belonging to residue $\pi(i)$, the adapter concatenates layer-normalized atom-encoder features $\mathbf{Q}_i^A$ and current residue scalars $\mathbf{h}_{\pi(i)}^R$:

$$\mathbf{s}_i^{\mathcal{F}} = W_f \left[\mathrm{LN}(\mathbf{Q}_i^A) \; ; \; \mathrm{LN}(\mathbf{h}_{\pi(i)}^R) \right]. \tag{26}$$

Fusion equivariant channels combine a learned mixing of residue vectors $\mathbf{V}_{\pi(i)}^R$ with the atom-to-representative offset:

$$\mathbf{V}_i^{\mathcal{F}} = W_{\mathrm{mix}} \, \mathbf{V}_{\pi(i)}^R + W_{\mathrm{disp}} \, \frac{\mathbf{x}_i - \mathbf{r}_{\pi(i)}}{3} \, . \tag{27}$$

The residue coordinate head yields displacement $\Delta\mathbf{r}_{\pi(i)}$, placing each fusion atom at

$$\mathbf{x}_g^{\mathcal{F},(i)} = \mathbf{x}^{\mathcal{P},(i)} + \Delta\mathbf{r}_{\pi(i)}^{(g)}. \tag{28}$$

Cross-attention edges from ligand to fusion atoms are recomputed from these refreshed coordinates at each round. The subsequent ligand group performs self-attention, fusion cross-attention, and feed-forward updates; the next protein group therefore reads the updated ligand geometry.

Reverse fusion (ligand to protein). Before the interleaving rounds begin, a reverse fusion module injects noisy ligand context into the Pairformer single representation $\mathbf{s}$. For each residue j with representative position $\mathbf{r}_j$, the module aggregates type-and-distance features over ligand atoms within 12 Å:

$$\mathbf{c}_j = \frac{1}{|\mathcal{N}_j|} \sum_{k \in \mathcal{N}_j} \big(W_d \, \mathrm{RBF}_{32}\big(\|\mathbf{r}_j - \mathbf{x}_k^L\|\big) \big) \odot \mathbf{e}_{a_k}, \quad \mathbf{s}_j \mathrel{+}= \mathrm{MLP}_0(\mathbf{c}_j), \tag{29}$$

where W_d projects the 32-channel Gaussian RBF to a hidden dimension, $\mathbf{e}_{a_k}$ is a learned embedding of ligand atom type a_k, the element-wise product couples type identity with spatial proximity, and MLP_0 is zero-initialized. The zero initialization ensures the pathway contributes nothing at the start of training and grows in gradually. Together, the pocket fusion adapter and reverse fusion form a bidirectional bridge that is refreshed every round, allowing protein and ligand representations to co-evolve.

A.2.5 Output Heads

The output heads decode the refined residue and ligand representations into coordinate corrections and categorical predictions.

Protein coordinates. Normalized residue scalars are projected back to the atom-encoder dimension, gathered to atoms, and added to the saved atom-encoder features $\mathbf{Q}^A$. Local atom-decoder layers then produce per-atom features $\mathbf{q}_i$. A residue-level coordinate head projects normalized final residue vectors to a three-dimensional displacement $\Delta\mathbf{x}_j^{\mathrm{res}}$ for each residue j, which is shared across all atoms of that residue. With $\overline{\mathbf{V}}^R$ denoting the same normalized vectors, a scalar-weighted combination of three equivariant sources yields the atom-level coordinate correction:

$$\Delta\mathbf{x}_i^{\mathrm{atom}} = \sum_c w_c(\mathbf{q}_i) \, \overline{\mathbf{V}}_{\pi(i),:,c}^R + \sum_c u_c(\mathbf{q}_i) \, \mathbf{V}_{i,:,c}^A + g(\mathbf{q}_i) \, (\mathbf{x}_i - \mathbf{r}_{\pi(i)}). \tag{30}$$

Adding this correction and the final residue displacement to the noisy coordinates gives Eq. 9 for every nonpadding protein atom.

Ligand coordinates and discrete outputs. The ligand coordinate head projects normalized final ligand vectors to three dimensions and adds the result to the current ligand coordinates. Atom-type and charge logits are obtained by projecting normalized final ligand scalars. For bonds, a refinement layer concatenates pairwise distances and dot products from the predicted coordinates with node and edge features, passes them through an MLP to update the edge representation, symmetrizes the result, and projects to bond logits. Masked categorical softmax produces the probabilities in Eq. 7.

Table 4: Architecture settings of FLEXIDD.

Module	Layers	Scalar	Vector	Pair	Heads
Time-invariant atom encoder	3	128	–	16	4
Pairformer	12	384	–	128	16 and 4
Protein atom encoder	6	320	16	16	4
Protein residue layers	16	768	128	64	32
Ligand layers	12	512	128	128	32
Fusion adapter	–	256	128	–	–
Protein atom decoder	6	320	–	16	4

Component	Setting
Inputs	
Time-invariant atom input	389 ch (local coords, charge, padding, 128 element, four 64-ch name)
Residue input	384+23+1152+4=1563 ch (pooled atoms, identity, ESM, site)
Ligand embeddings	64 ch each (type, charge, time, size, bond)
Attention	
Pairformer heads	16 single, 4 pair; dropout 0.15
Atom-attention window	32 queries, 128 keys
Protein atom encoder	16 vector ch saved for output head; attention updates scalars only
Scalar message, hidden	64, 96
Radial basis features	
Atom offset	8 centers, 0 to 8 Å
Atom neighborhood	16 centers, 0 to 15 Å
Residue edges	32 centers, 0 to 32 Å
Residue vector injection	16 centers, 0 to 8 Å
Ligand edges	64 centers
Cross-module edges	
Residue-to-ligand	width 32, 16 radial features, 5 Å scale
Ligand-to-fusion	width 128, 32 initial and 16 refresh radial, 5 Å scale

Vector widths count three-dimensional channels. The time-invariant context is computed once and shared by two denoising passes. An em-dash denotes a nonapplicable entry.

A.2.6 SELF-CONDITIONING

Self-conditioning (Chen et al., 2022) lets the network refine its own predictions by running the denoising network f_θ twice with shared parameters. The first pass receives the ligand coordinate prior and zero atom and bond probabilities as conditioning inputs, and runs without gradient tracking. The second pass re-initializes all internal states from the original noisy inputs but replaces the zero conditioning with the detached predictions from the first pass, allowing the network to correct its initial estimate. During training, the two-pass procedure is activated with probability 0.5; otherwise the network receives only the zero-initialized conditioning. At inference, every sampling step uses the two-pass procedure. Protein-only examples omit ligand operations while retaining the protein atom encoder, residue layers, and atom decoder.

A.3 EQUIVARIANCE OF THE DETERMINISTIC PREDICTOR

We show that the deterministic predictor f_θ is E(3)-equivariant: rotating or translating every input coordinate produces identically transformed output coordinates and leaves discrete probabilities unchanged. The proof proceeds in three steps. We first factor out translation by centering both coordinate streams on the protein centroid, reducing the problem to $O(3)$. We then verify that every primitive operation the network uses preserves $O(3)$ symmetry. Finally, we trace the data flow from input through the three modules to the output heads, confirming that equivariance is maintained at each stage.

Statement: Given the time-invariant chemical context F and diffusion time t, let $\mathcal{C}$ contain the global protein, ligand, and ligand self-conditioning coordinates. For row vectors, define $g \cdot X = XR^\top + \mathbf{1}d^\top$ with $g = (R, d)$ and $R \in O(3)$, while F does not transform. The predictor operates in evaluation

mode with cross-product features disabled. Assuming a nonempty observed, nonpadding protein mask, its active outputs satisfy

$$\begin{aligned} \widehat{X}^s(g \cdot \mathcal{C}; F) &= \widehat{X}^s(\mathcal{C}; F) R^\top + \mathbf{1} d^\top, \quad && s \in \{\mathcal{P}, \mathcal{M}\}, \\ \widehat{p}^\ell(g \cdot \mathcal{C}; F) &= \widehat{p}^\ell(\mathcal{C}; F), && \ell \in \{a, b, q\}, \end{aligned} \tag{31}$$

where a, b, and q denote atom types, bond states, and formal charges.

Step 1. Shared centering: With observed, nonpadding protein mask m_i, subtract the protein centroid c from both streams and the ligand self-conditioning coordinates:

$$c = \frac{\sum_i m_i x_i^{\mathcal{P}}}{\sum_i m_i}, \quad c' = cR^\top + d^\top, \quad \widetilde{X} = X - \mathbf{1}c, \quad \widetilde{X}' = \widetilde{X} R^\top. \tag{32}$$

The affine action commutes with the masked mean, so translation is absorbed into the centroid. It therefore suffices to prove $O(3)$ equivariance in this centered frame and restore c at output.

Step 2. Building blocks: Before tracing the data flow, we verify that every primitive the network uses respects $O(3)$. Write each geometric channel as a row vector v_{ic}. For every orthogonal R, including reflections,

$$\|vR^\top\| = \|v\|, \quad \langle vR^\top, wR^\top \rangle = \langle v, w \rangle, \quad \sum_j \alpha_{ij}(v_j R^\top) = \left(\sum_j \alpha_{ij} v_j \right) R^\top \tag{33}$$

whenever α_{ij} is invariant. Distances, radial embeddings, norms, and dot products are therefore invariant, as are scalar networks and masked attention weights derived from them. Bias-free vector-channel mixing, invariant scalar gates and normalization, weighted sums, and residual additions preserve equivariance. Index gathers, broadcasts, and masks commute with R. Time enters vector channels only through invariant multiplicative gates.

Step 3. Data flow through the network: We now follow the data through each module, confirming that scalars remain invariant and vectors remain equivariant.

The Pairformer context $(\mathbf{s}, \mathbf{z})$ depends only on F and is therefore invariant; it conditions all three modules but introduces no geometric content.

In the protein module, global geometry enters the scalar atom encoder only through invariant radial descriptors. Vector channels originate from atom-representative and distance-weighted neighbor displacements and evolve through the building blocks of Step 2.

In the ligand module, centered input and self-conditioning coordinates enter through bias-free vector mixing with invariant time gates, while chemical embeddings remain invariant.

In the interaction module, the pocket fusion adapter gathers atoms at predetermined indices, broadcasts residue features, mixes vector channels without bias, and adds residue displacements to atom coordinates. The reverse fusion path injects distance-weighted ligand context into the protein stream. Cross-stream radial features and attention weights are invariant.

All layers use the operations cataloged in Step 2, including distance-weighted neighbor updates. Induction over the alternating protein and ligand layers with interleaved fusion therefore preserves invariant scalars and equivariant vectors throughout.

Output heads and conclusion: For atom i in residue $\pi(i)$ with representative coordinate $\mathbf{r}_{\pi(i)}$, the protein coordinate prediction is

$$\delta_i = \sum_c w_{ic} \, \overline{\mathbf{V}}^R_{\pi(i),c} + \sum_c u_{ic} \, \mathbf{V}^A_{i,c} + \gamma_i (\mathbf{x}_i - \mathbf{r}_{\pi(i)}), \quad \widehat{\mathbf{x}}_i^{\mathcal{P}} = \mathbf{x}_i + \Delta \mathbf{x}^{\mathrm{res}}_{\pi(i)} + \delta_i, \tag{34}$$

where invariant weights w_{ic}, u_{ic}, γ_i and equivariant residue displacement $\Delta \mathbf{x}^{\mathrm{res}}_{\pi(i)}$ make this equivariant. The ligand coordinate head uses bias-free vector mixing and a coordinate residual, preserving equivariance by the same argument. Atom and charge logits depend on invariant scalars; bond logits also use invariant distances and dot products before symmetrization. Their softmax probabilities are therefore invariant.

Restoring the centroid gives $(\widehat{\widetilde{X}}R^\top) + \mathbf{1}(cR^\top + d^\top) = (\widehat{\widetilde{X}} + \mathbf{1}c)R^\top + \mathbf{1}d^\top$, which proves Eq. 31 for one denoising pass.

Self-conditioning and scope: The second pass receives the first pass's equivariant ligand coordinates and invariant atom and bond probabilities in the same centered frame. The argument therefore applies identically; detachment leaves these values unchanged. External Cartesian self-conditioning seeds belong to $\mathcal{C}$ and must transform with it. The claim is restricted to this deterministic predictor with given prepared features. Changes to local reference conformers and stochastic sample paths with unchanged noise are outside its scope.

A.4 Training Objective Definitions

This section defines each term in the training objective (Eq. 10 and Eq. 11).

A.4.1 Coordinate and Chemical Reconstruction

For $U \in \{\mathcal{P}, \mathcal{M}\}$ and valid atom mask m_i^U, coordinate reconstruction is

$$\mathcal{L}_{\text{coord}}^U = \frac{\sum_i m_i^U \left\| \widehat{\mathbf{x}}_{\text{ref}}^{U,(i)} - \mathbf{x}_{\text{ref}}^{U,(i)} \right\|_2^2}{\max(1, \sum_i m_i^U)}. \tag{35}$$

This sums squared Cartesian errors before averaging over atoms.

For atom types and charges, $c \in \{a, q\}$, one-hot targets $c_{\text{ref},k}^{(i)}$ and predicted probabilities $\widehat{p}_k^{c,(i)}$ give

$$\mathcal{L}_c = -\frac{\sum_i m_i^{\mathcal{M}} \sum_{k=1}^{K_c} c_{\text{ref},k}^{(i)} \log \widehat{p}_k^{c,(i)}}{\max(1, \sum_i m_i^{\mathcal{M}})}. \tag{36}$$

Bond supervision averages the analogous cross entropy over all valid ordered atom pairs:

$$\mathcal{L}_b = -\frac{\sum_{i,j} m_i^{\mathcal{M}} m_j^{\mathcal{M}} \sum_{k=1}^{K_b} b_{\text{ref},k}^{(ij)} \log \widehat{p}_k^{b,(ij)}}{\max(1, \sum_{i,j} m_i^{\mathcal{M}} m_j^{\mathcal{M}})}. \tag{37}$$

The sum includes both orderings of each pair as well as nonbonded entries, with diagonal elements set to no-bond. Cross entropy is computed from logits for numerical stability.

A.4.2 Local Complex Geometry

For pair distances $\widehat{d}_{ij}$ and d_{ij}^{ref} in the concatenated predicted and reference complexes, define

$$e_{ij} = \left| \widehat{d}_{ij} - d_{ij}^{\text{ref}} \right|, \quad \ell_{ij} = \frac{1}{4} \sum_{\delta \in \{0.5,1,2,4\}} \text{sigmoid}(\delta - e_{ij}). \tag{38}$$

Let $\mathcal{S}$ contain valid distinct pairs within 15 Å in the reference complex. With distances in angstroms, smooth lDDT (Mariani et al., 2013; Abramson et al., 2024) is

$$\mathcal{L}_{\text{lddt}} = 1 - \frac{\sum_{(i,j)\in\mathcal{S}} w_{ij} \ell_{ij}}{\max(1, \sum_{(i,j)\in\mathcal{S}} w_{ij})}, \quad w_{ij} = 1 + \omega_{\mathcal{M}} \mathbb{I}[i \text{ or } j \text{ belongs to the ligand}]. \tag{39}$$

Here $\omega_{\mathcal{M}} = 5$ weights ligand-containing pairs independently of the overall loss coefficient, and gradients reach both structures. Even perfect distances give $\ell_{ij} < 1$, unlike discrete evaluation lDDT. When the pair set is empty the loss defaults to one, contributing no coordinate gradient.

A.4.3 Ligand Geometry

To regularize internal ligand geometry, we supervise predicted bond lengths and angles against reference values. From the reference graph, define valid bonded pairs $\mathcal{E}_{\text{ref}}$ with $i < j$ and angle triples $\mathcal{T}_{\text{ref}}$ with central atom j and neighbors $i < k$. For bond lengths d and angles α in radians,

$$\begin{aligned} \mathcal{L}_{\text{bond}} &= \frac{1}{\max(1, |\mathcal{E}_{\text{ref}}|)} \sum_{(i,j)\in\mathcal{E}_{\text{ref}}} (\widehat{d}_{ij} - d_{ij}^{\text{ref}})^2, \\ \mathcal{L}_{\text{angle}} &= \frac{1}{\max(1, |\mathcal{T}_{\text{ref}}|)} \sum_{(i,j,k)\in\mathcal{T}_{\text{ref}}} h_{0.5}(\widehat{\alpha}_{ijk} - \alpha_{ijk}^{\text{ref}}). \end{aligned} \tag{40}$$

Table 5: Weights of the terms in Eq. 10 and Eq. 11. Each multiplies a separately normalized loss.

Term	Stage 1	Stage 2
Protein coordinate reconstruction	1	1
Ligand coordinate reconstruction	0	3
Atom type classification	0	1
Charge classification	0	2
Bond classification	0	3
Smooth lDDT	0	3
Ligand bond length	0	5
Ligand bond angle	0	10
Ligand-pocket distance	0	10

Triplets with degenerate bond vectors are excluded. The Huber function is

$$h_\delta(e) = \begin{cases} \frac{1}{2}e^2, & |e| \le \delta, \\ \delta(|e| - \frac{1}{2}\delta), & |e| > \delta. \end{cases} \tag{41}$$

A.4.4 Ligand-Pocket Distances and Loss Weights

Stage 1 uses only $\mathcal{L}^{\mathcal{P}}_{\text{coord}}$; Stage 2 adds ligand and complex terms with the weights in Tab. 5.

To supervise relative ligand-pocket placement, let $\mathcal{I}$ contain valid ligand and interface-atom pairs with reference distance $d^{\text{ref}}_{ij} < 10\,\text{Å}$. Let $\bar{d}_{ij} = \left\| \widehat{\mathbf{x}}^{\mathcal{M},(i)}_{\text{ref}} - \text{sg}\left(\widehat{\mathbf{x}}^{\mathcal{P},(j)}_{\text{ref}}\right) \right\|_2$ denote the predicted distance between ligand atom i and protein atom j, with protein coordinates detached from the gradient. The loss is

$$\mathcal{L}_{\text{dist}} = \frac{\sum_{(i,j)\in\mathcal{I}} \exp(-d^{\text{ref}}_{ij}/5) h_2(\bar{d}_{ij} - d^{\text{ref}}_{ij})}{\max(1, |\mathcal{I}|)}, \tag{42}$$

with distances in angstroms. The denominator counts eligible pairs. This term penalizes distances both shorter and longer than the reference.

A.5 Noise Schedule and Sampling

This section specifies the noise schedule, integration rules, and categorical transition kernel used at inference.

A.5.1 Time Schedule and Coordinate Integration

We use the Karras noise ladder (Karras et al., 2022) with $\sigma_{\max} = 8\,\text{Å}$, $\sigma_{\min} = 0.05\,\text{Å}$, and $\rho = 7$. This is substantially smaller than the $\sigma_{\max} = 160\,\text{Å}$ used in Boltz-1 (Wohlwend et al., 2025) for protein structure prediction. For $u \in [0, 1]$, define

$$\begin{aligned} \sigma(u) &= \left[\sigma^{1/\rho}_{\max} + u\left(\sigma^{1/\rho}_{\min} - \sigma^{1/\rho}_{\max}\right)\right]^\rho, \\ t(u) &= 1 - \frac{\log(\sigma(u)/\sigma_{\min})}{\log(\sigma_{\max}/\sigma_{\min})}. \end{aligned} \tag{43}$$

Training samples u uniformly, inducing a nonuniform t shared by protein corruption and ligand interpolation. Sampling uses $K = 100$ intervals with $u_k = k/K$ and network evaluations at $k = 0, \ldots, K-1$, without stochastic corrections.

For protein coordinates, Euler integration in σ of the denoiser-based field $(\mathbf{X}^{\mathcal{P}} - \widehat{\mathbf{X}}^{\mathcal{P}}_{\text{ref}})/\sigma$ gives

$$\mathbf{X}^{\mathcal{P}}_{k+1} = \widehat{\mathbf{X}}^{\mathcal{P}}_{\text{ref},k} + \frac{\sigma_{k+1}}{\sigma_k}\left(\mathbf{X}^{\mathcal{P}}_k - \widehat{\mathbf{X}}^{\mathcal{P}}_{\text{ref},k}\right), \quad \sigma_k = \sigma(t_k). \tag{44}$$

The returned protein is the integrated state at $\sigma_{\min} > 0$.

The ligand endpoint velocity $(\widehat{\mathbf{X}}^{\mathcal{M}}_{\text{ref}} - \mathbf{X}^{\mathcal{M}})/(1-t)$ (Lipman et al., 2022; Irwin et al., 2024) gives the Euler step

$$\mathbf{X}^{\mathcal{M}}_{k+1} = \mathbf{X}^{\mathcal{M}}_k + \frac{t_{k+1} - t_k}{1 - t_k}\left(\widehat{\mathbf{X}}^{\mathcal{M}}_{\text{ref},k} - \mathbf{X}^{\mathcal{M}}_k\right), \quad t_k < 1. \tag{45}$$

Evaluations occur only at $t_k < 1$; the final step reaches the last predicted ligand endpoint. This ordinary endpoint velocity is not the exact probability flow of the training path in Eq. 5 when its perturbation $\eta > 0$.

A.5.2 Categorical Transitions and Decoding

For each categorical variable (Gat et al., 2024; Cremer et al., 2026a), let c be its current state and $\widehat{p}_j$ its endpoint probability over K_c channels. Define $\Delta t_k = t_{k+1} - t_k$, $\bar{t}_k = \mathrm{clip}(t_k, 10^{-5}, 1 - 10^{-5})$, and $\xi_k = 1$ except for $\xi_k = 0$ on the final step. The off-diagonal proposal masses are

$$r_j = \Delta t_k \left[\frac{1 + \xi_k + \xi_k (K_c - 1)\bar{t}_k}{1 - \bar{t}_k} \widehat{p}_j + \xi_k \widehat{p}_c \right], \quad j \neq c. \tag{46}$$

Clip each mass at one, assign the nonnegative remainder to c, normalize, and sample:

$$\widetilde{p}_j = \min(1, r_j) \quad (j \neq c), \quad \widetilde{p}_c = \max\left(0, 1 - \sum_{j \neq c} \widetilde{p}_j\right), \quad \mathbf{c}_{k+1} \sim \mathrm{Cat}\left(\frac{\widetilde{\mathbf{p}}}{\sum_j \widetilde{p}_j}\right). \tag{47}$$

The channel count K_c includes any padding channel. Atom or charge padding proposals retain the current state; bonds are symmetrized with no-bond diagonal entries.

Decoding rejects bond-order increases exceeding valence budgets and removes low-confidence bonds if excess valence remains at the final step. At the final step, discrete outputs are sampled from the predicted probabilities. These decoding constraints, separate from training losses, guarantee neither connectivity nor complete chemical validity.

The ligand heavy-atom count $n_{\mathcal{M}}$ is sampled before generation following prior work (Gat et al., 2024; Cremer et al., 2026a).

A.5.3 Sampling Algorithm

Alg. 1 uses the shared pocket-centered frame of Apx. A.2. In Alg. 1 and the training procedure (Apx. A.6), $\mathrm{Predict}_\theta$ returns $\widehat{\mathcal{Y}} = (\widehat{\mathbf{X}}^{\mathcal{P}}_{\mathrm{ref}}, \widehat{\mathbf{X}}^{\mathcal{M}}_{\mathrm{ref}}, \widehat{\mathbf{p}}^{a}, \widehat{\mathbf{p}}^{q}, \widehat{\mathbf{p}}^{b})$ from Eq. 7 using two self-conditioning passes. The first runs without gradient tracking; its detached ligand coordinates and atom and bond probabilities condition the second. Masks and prepared features are implicit.

Algorithm 1 Sampling: Joint Protein-Ligand Generation

Require: Supplied protein $\mathbf{X}^{\mathcal{P}}_{\mathrm{in}}$, prepared features $F^{\mathcal{P}}$, site labels $\mathcal{S}$, ligand size $n_{\mathcal{M}}$
1: Prepare the shared pocket-centered frame; store its translation and retain valid-atom masks
2: Compute and cache $(\mathbf{s}, \mathbf{z})$ using input embedding and Pairformer with 3 recycling iterations
3: Set $K = 100$; compute (σ_k, t_k) from Eq. 43 at $u_k = k/K$
4: $\mathbf{X}^{\mathcal{P}}_0 \leftarrow \mathbf{X}^{\mathcal{P}}_{\mathrm{in}} + \sigma_{\max} \boldsymbol{\epsilon}^{\mathcal{P}}$, $\boldsymbol{\epsilon}^{\mathcal{P}} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
5: Draw the centered Gaussian $\mathbf{Z}^{\mathcal{M}}$ from Eq. 24; set $\mathbf{X}^{\mathcal{M}}_0 \leftarrow \mathbf{Z}^{\mathcal{M}}$
6: Draw $\mathbf{a}_0, \mathbf{q}_0$ uniformly over nonpadding channels and $\mathbf{b}_0$ uniformly over bond channels
7: Symmetrize $\mathbf{b}_0$ and set its diagonal to no bond
8: **for** $k = 0, \ldots, K-1$ **do**
9: $\quad \widehat{\mathcal{Y}}_k \leftarrow \mathrm{Predict}_\theta(\mathbf{X}^{\mathcal{P}}_k, \mathcal{M}_k, \mathbf{s}, \mathbf{z}, t_k)$
10: $\quad \mathbf{X}^{\mathcal{P}}_{k+1} \leftarrow \widehat{\mathbf{X}}^{\mathcal{P}}_{\mathrm{ref}} + \frac{\sigma_{k+1}}{\sigma_k}(\mathbf{X}^{\mathcal{P}}_k - \widehat{\mathbf{X}}^{\mathcal{P}}_{\mathrm{ref}})$
11: $\quad \mathbf{X}^{\mathcal{M}}_{k+1} \leftarrow \mathbf{X}^{\mathcal{M}}_k + \frac{t_{k+1} - t_k}{1 - t_k}(\widehat{\mathbf{X}}^{\mathcal{M}}_{\mathrm{ref}} - \mathbf{X}^{\mathcal{M}}_k)$
12: $\quad$ Propose $\mathbf{a}_{k+1}, \mathbf{q}_{k+1}, \mathbf{b}_{k+1}$ using Eq. 46 and Eq. 47
13: $\quad$ Reject atom and charge padding proposals; enforce symmetric bonds with a no-bond diagonal
14: $\quad$ Reject bond-order increases that exceed valence budgets of the updated atom and charge states
15: **end for**
16: Reduce any remaining excess valence by removing low-confidence bonds
17: Restore the stored input-frame translation to both coordinate arrays
18: **return** $(\mathbf{X}^{\mathcal{P}}_K, \mathbf{X}^{\mathcal{M}}_K, \mathbf{a}_K, \mathbf{q}_K, \mathbf{b}_K)$

Table 6: Reported scale of public apo-holo paired data.

Method or resource	Reported pairs	Pair construction
APObind (Aggarwal et al., 2021)	10,599	Experimental apo-holo structures
Apo2Mol (Zheng et al., 2026)	24,601	Experimental apo-holo-ligand triplets
DynamicFlow (Zhou et al., 2025)	5,692	Apo pockets with MD-derived bound states
FlexSBDD (Zhang et al., 2024)	Not reported	Experimental or synthetic apo states

A.6 Training Algorithm

Alg. 2 summarizes the second-stage forward pass: for each reference complex, we construct $J = 16$ independently corrupted views, reuse the cached trunk representations $(\mathbf{s}, \mathbf{z})$ across them, and average the resulting losses. In the first training stage, the ligand stream is absent, and the protein stream and trunk are optimized solely with the protein coordinate loss $\mathcal{L}^{\mathcal{P}}_{\text{coord}}$.

Algorithm 2 Training: Forward Pass and Loss Computation (Stage 2)

Require: Reference complex $(\mathbf{X}^{\mathcal{P}}_{\text{ref}}, \mathcal{M}_{\text{ref}})$, prepared features $F^{\mathcal{P}}$, site labels $\mathcal{S}$
Prepare the complex in the shared pocket-centered frame; retain coordinate and padding masks
Sample recycling count $r \sim \text{Uniform}\{0, 1, 2, 3\}$
Compute and cache $(\mathbf{s}, \mathbf{z})$ using input embedding and Pairformer with r recycling iterations
for $j = 1, \ldots, J$ with $J = 16$ **do**
  Draw $u \sim \text{Uniform}[0, 1]$; compute $\sigma(u), t(u)$ from Eq. 43
  $\mathbf{X}^{\mathcal{P}}_t \leftarrow \mathbf{X}^{\mathcal{P}}_{\text{ref}} + \sigma(u)\boldsymbol{\epsilon}^{\mathcal{P}}$, $\boldsymbol{\epsilon}^{\mathcal{P}} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
  Draw $\mathbf{X}^{\mathcal{M}}_0$ from Eq. 24 and uniform categorical priors
  Match prior coordinates to the reference by minimum-cost assignment (Sec. 3.2.2)
  Apply this permutation to all prior atom variables and both axes of the bond tensor
  $\mathbf{X}^{\mathcal{M}}_t \leftarrow (1-t)\mathbf{X}^{\mathcal{M}}_0 + t\mathbf{X}^{\mathcal{M}}_{\text{ref}} + 0.3\boldsymbol{\epsilon}^{\mathcal{M}}$, $\boldsymbol{\epsilon}^{\mathcal{M}} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
  Draw $\mathbf{a}_t, \mathbf{q}_t, \mathbf{b}_t$ from the categorical mixture in Eq. 6
  Enforce symmetric bond states and a no-bond diagonal
  Apply one common random rotation to the noisy coordinates, coordinate priors, and reference targets
  $\widehat{\mathcal{Y}}_j \leftarrow \text{Predict}_\theta(\mathbf{X}^{\mathcal{P}}_t, \mathcal{M}_t, \mathbf{s}, \mathbf{z}, t)$
  Compute masked $\mathcal{L}_j$ from Eq. 11 with the weights in Tab. 5
end for
return $J^{-1} \sum_{j=1}^{J} \mathcal{L}_j$

A.7 Public Apo-Holo Paired Data

A key motivation in this paper (Sec. 3.2.1) is the limited availability of paired apo-holo data compared with abundant holo-ligand complex data. Tab. 6 surveys the public resources we are aware of. APObind (Aggarwal et al., 2021) and Apo2Mol (Zheng et al., 2026) curate experimentally determined apo and holo structures, DynamicFlow (Zhou et al., 2025) pairs apo pockets with molecular-dynamics-derived bound states, and FlexSBDD (Zhang et al., 2024) associates bound complexes with either experimental or synthetic apo conformations. The largest curated set contains fewer than 25k pairs, motivating our transferable denoising strategy that learns flexible generation from individual bound complexes via Gaussian corruption (Sec. 3.2.1).

A.8 Training Data

FlexiDD uses separate protein-only and protein-ligand complex corpora for its two training stages. Each protein-only record corresponds to one processed protein chain; each complex record corresponds to one protein-ligand pair.

A.8.1 Protein-Only Data

The first stage draws on 436,378 single-chain records from two components of the OpenFold training resources (Ahdritz et al., 2024): experimental PDB structures and AlphaFold2 self-distillation structures.

Experimental PDB subset: We processed the archived PDB mmCIF snapshot distributed with these resources (Berman et al., 2000; Ahdritz et al., 2024). Of 185,158 source entries, 170,671 passed our structure parser and chain filters. We extracted each valid protein chain as an individual record, retaining unresolved residues in the full polymer sequence and deduplicating identical sequences within each source entry. This produced 302,450 records.

Self-distillation subset: The OpenFold self-distillation set provides 269,455 AlphaFold2-predicted structures (Jumper et al., 2021; Ahdritz et al., 2024). We removed proteins with mean pLDDT below 80, retaining 136,857 candidates. Of these, 133,928 passed the common structural validity checks.

A.8.2 Protein-Ligand Complex Data

The second stage uses 331,738 filtered records from three sources: synthetic complexes from GatorAffinity (Wei et al., 2025) and experimental complexes from PLINDER (Durairaj et al., 2024) and HiQBind (Wang et al., 2025). Records whose ligand graph could not be constructed were discarded.

GatorAffinity: Among 463,867 indexed predictions (Wei et al., 2025), 279,210 had an available structure and a confidence score of at least 0.8. After single-chain parsing, sequence consistency checks, and structural validation, 248,126 records remained. Because GatorAffinity complexes were generated with Boltz-1 (Wohlwend et al., 2025), whose PDB training data predate December 30, 2022, we treat this corpus as satisfying our temporal cutoff.

PLINDER: We pooled the published source splits and applied a PDB release date cutoff of December 30, 2022, reducing 251,494 prepared pairs to 224,281 (Durairaj et al., 2024). We kept noncovalent druglike ligands that RDKit could parse and that contained at least one carbon atom. After structure conversion, we required exactly one valid protein chain, yielding 64,638 records.

HiQBind: From the published HiQBind pairs (Wang et al., 2025), we retained structures with exactly one protein chain and a release year no later than 2022, giving 19,358 candidates. After structure parsing and removing records with missing ligand chains or unsupported residue types, 18,974 records remained.

A.8.3 Splits and Sampling

For the first stage, we randomly sampled 100 of the 436,378 protein-only records for validation; the remaining 436,278 were used for training. For the second stage, we clustered proteins at 30% sequence identity across the three complex sources and held out entire clusters for validation, capping at 42 records per source (Tab. 7). The remaining records formed the training pool, sampled at the per-source probabilities shown in Tab. 7.

Table 7: Complex data splits and sampling weights.

Source	Filtered	Training	Validation	Sampling
GatorAffinity	248,126	248,084	42	0.5
PLINDER	64,638	64,596	42	0.3
HiQBind	18,974	18,932	42	0.2
Total	331,738	331,612	126	–

A.8.4 Cropping and Binding Site Conditioning

Both stages share the same per-example budget: 384 protein tokens, 3,456 protein atom slots, and 128 ligand slots, following prior work (Wohlwend et al., 2025). Records within these limits keep their full protein context; oversized records are cropped to fit. Every retained residue receives a categorical site label, POCKET for residues in the binding site and UNSELECTED for the rest.

In Stage 1, no ligand is available, so records are cropped to fit the budget and all residues are labeled UNSELECTED. In Stage 2, a residue is labeled POCKET if any of its heavy atoms lies within 6 Å of the ligand (Sec. 3.2.1), and the reference ligand is then removed from the protein input. Oversized Stage 2 complexes are either cropped around the binding site or, with probability 0.5, receive a

coverage crop without ligand supervision. It should be noted that the cropping strategy is only used for training; the validation and the inference processes apply no cropping.

A.8.5 Initialization and Optimization

Stage 1 trains the protein module from scratch. Stage 2 transfers these weights and randomly initializes the ligand and interaction modules.

Table 8: Optimization settings in the Stage 1 and Stage 2 scale configurations.

Setting	Stage 1	Stage 2
Optimizer	Adam	Adam
Base / new parameter learning rate	1.8×10^{-3}	1.8×10^{-3}
Transferred protein learning rate	-	4×10^{-4}
Ligand learning rate	-	3×10^{-4}
Warmup steps / subsequent schedule	500 / constant	500 / constant
Adam (β_1, β_2) / ϵ	$(0.9, 0.95)$ / 10^{-8}	$(0.9, 0.95)$ / 10^{-8}
Weight decay / gradient norm clipping	0 / 5	0 / 5
Records per device / gradient accumulation	1 / 4	1 / 4
Configured B200 GPUs / nodes	8 / 2	8 / 2
Corruption draws per prepared record	16	16
Prepared records per epoch	32,000	16,000
EMA decay	0.998	0.998
Seed	42	42

Training runs on 16 NVIDIA B200 GPUs with four gradient-accumulation steps, giving an effective batch of 64 prepared records per optimizer update. Each record contributes 16 independent corruption draws, so every update processes 1,024 noisy views in total. Stage 2 additionally applies a shared random rotation to protein and ligand coordinates and employs the self-conditioning procedure described in Apx. A.2. Both stages are capped at 200 epochs and two days of wall time.

A.9 Baseline Implementation Details

All baselines are retrained on the Stage 2 source complexes used by FlexiDD, following the released source code and default training configs. Rigid-protein methods keep their original pocket definitions for both training and inference. FlexSBDD (Zhang et al., 2024) and DynamicFlow (Zhou et al., 2025) are also closely related but are excluded because their source code is not publicly available.

Apo2Mol (Zheng et al., 2026) requires paired protein conformations, so we pair each training complex with its experimental apo structure when one exists, and otherwise pair the holo structure with itself. YuelDesign (Wang et al., 2026) and FlexFlow (Schneuing et al., 2025) train on the bound complexes directly without paired-conformation augmentation according to their default settings.

A.10 PLINDER Apo-Holo Test Benchmark

We construct the apo-holo benchmark from PLINDER (Durairaj et al., 2024), retaining complexes released after the December 30, 2022 training cutoff that have a matched experimental apo structure with the same protein sequence as the holo complex. Every test protein must share less than 80% sequence similarity with all training proteins. The resulting benchmark contains 330 paired apo-holo systems whose mean maximum sequence similarity to the training set is 46.5%; 25.8% of the systems fall below 30%.

Each system pairs an experimental apo protein with the corresponding holo protein-ligand complex. For known-site conditioning, a residue is included in the binding site when its minimum heavy-atom distance to the ligand is below 6 Å; hydrogen atoms are excluded. The apo structure supplies the input coordinates and the holo complex serves as the evaluation reference.

A.10.1 TRAIN-TEST SIMILARITY CALCULATION

Sequence similarity is computed with MMseqs2 (Steinegger & Söding, 2017) by searching each unique full-length test sequence against all unique training sequences. For an aligned query-target pair, we define

$$S_{\text{seq}} = \frac{N_{\text{ident}}}{L_{\text{test}}}, \tag{48}$$

where N_{ident} is the number of identical aligned residues and L_{test} is the full test-sequence length; the maximum score over all retained hits is reported for each test protein. Fig. 7 plots the resulting distribution across the 330 test systems, with a mean similarity of 0.465.

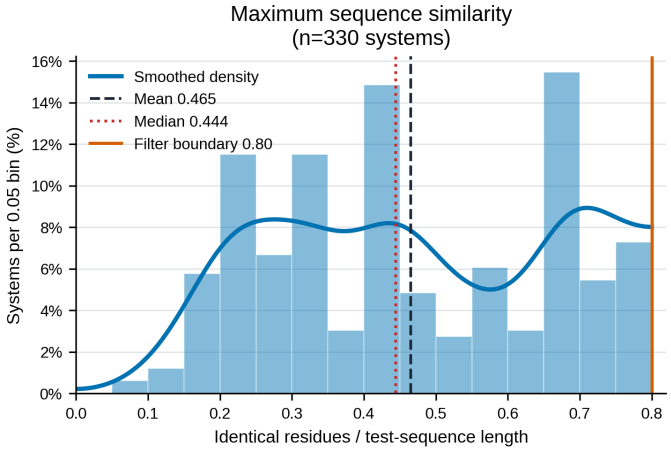

Figure 7: Maximum sequence similarity between each of the test systems and the training set.

A.10.2 SEQUENCE-SIMILARITY STRATIFICATION

To assess performance across different levels of train-test relatedness, we stratify the benchmark by each test protein's maximum sequence similarity to the training set. The low-, medium-, and high-similarity subsets are defined by $S_{\text{seq}} < 0.30$, $0.30 \leq S_{\text{seq}} < 0.50$, and $0.50 \leq S_{\text{seq}} < 0.80$, respectively, and contain 85, 113, and 132 systems. All systems sharing the same apo protein receive the same subset assignment.

A.10.3 APO-HOLO CONFORMATIONAL DIFFERENCES

We quantify apo-holo conformational differences by the Cα RMSD after a global Kabsch alignment over sequence-matched residues. As Fig. 8 shows, both whole-protein and binding-site RMSDs remain below 2 Å for over 90% of the test systems, with median values of 0.845 Å and 0.804 Å.

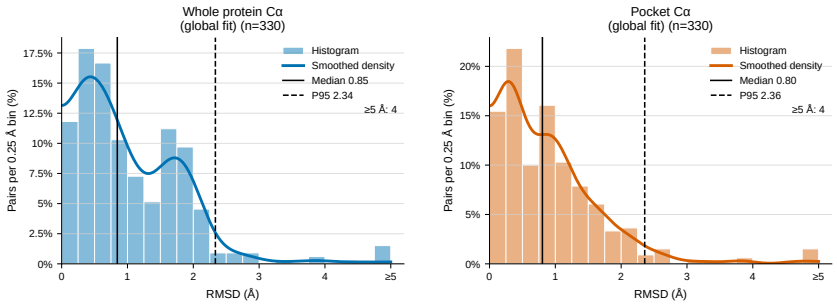

Figure 8: Apo-holo RMSD distributions of the test benchmark, computed over sequence-matched Cα atoms after a global Kabsch alignment. Left: whole-protein RMSD. Right: binding-site RMSD, where site residues are those with a heavy atom within 6 Å of the ligand.

B Extended Results

B.1 Evaluation Protocol and Metrics

B.1.1 Evaluation Inputs

The benchmark contains 330 PLINDER systems, each pairing a holo protein-ligand complex with its matched experimental apo structure (Durairaj et al., 2024). Each method uses its own default pocket configuration for generation. Rigid-protein methods use the experimental protein as input, and flexible SBDD methods use the generated protein for evaluation.

For each benchmark system, the ligand heavy-atom count $n_{\mathcal{M}}$ is sampled once from a prior distribution constructed from our training set following previous work (Cremer et al., 2026a). The same sampled atom count is then used across all baselines to ensure a fair comparison.

B.1.2 Metric Definitions

Table 9: Reported metrics and preferred direction.

Metric	Dir.	Definition
Validity	↑	RDKit-valid, single-fragment outputs per generation slot
PB-validity	↑	Fraction passing all PoseBusters `dock` checks
Strain	↓	MMFF94s energy decrease after local minimization (kcal/mol)
Vina	↓	AutoDock Vina score at the generated pose (kcal/mol)
Vina Min	↓	Vina score after local in-pocket minimization (kcal/mol)
QED	↑	Quantitative estimate of drug-likeness
SA	↑	Rescaled synthetic accessibility, $(10 - \mathrm{SA}_{\mathrm{raw}})/9$
LogP	-	Estimated octanol-water partition coefficient
Lipinski	↑	Number of satisfied drug-likeness rules
Uniq.2D	↑	Fraction of distinct canonical SMILES within a system
Div.2D	↑	Mean pairwise fingerprint Tanimoto distance within a system

Validity: Validity is the fraction of generated molecules passing RDKit sanitization as a single connected component.

PB-validity: PB-validity is the fraction passing all PoseBusters 0.3.1 `dock` checks (Buttenschoen et al., 2024), which assess internal chemistry, intramolecular geometry, and protein-ligand clashes.

Strain: We compute the MMFF94s energy of the generated ligand (with added hydrogens) and minimize locally for up to 1,000 steps with the MMFF94s force field (Halgren, 1999). Strain is $E_{\mathrm{pose}} - E_{\mathrm{local\ min}}$ in kcal/mol.

Vina and Vina Min: AutoDock Vina 1.2.7 (Eberhardt et al., 2021) scores each pose within a box centered on the reference ligand, with ligand preparation handled by Meeko 0.7.1. Vina reports the score at the generated pose; Vina Min reports the score after local optimization within the same box.

QED: RDKit's weighted drug-likeness estimate in $[0, 1]$ (Bickerton et al., 2012).

SA: The Ertl–Schuffenhauer synthetic accessibility score (Ertl & Schuffenhauer, 2009), rescaled as $(10 - \mathrm{SA}_{\mathrm{raw}})/9 \in [0, 1]$ so that higher values indicate easier synthesis.

LogP: RDKit's Crippen partition-coefficient estimate (Wildman & Crippen, 1999).

Lipinski: The number of satisfied Lipinski rules out of five (Lipinski et al., 2001; Cremer et al., 2026b).

Uniq.2D: The fraction of distinct canonical SMILES among the selected molecules within each system.

Div.2D: The mean pairwise Tanimoto distance $1 - T(f_a, f_b)$ over 2,048-bit radius-2 Morgan fingerprints (Rogers & Hahn, 2010) within each system. Systems with fewer than two fingerprintable molecules are omitted.

QED, SA, LogP, and Lipinski depend on the molecular graph and are computed with RDKit. Validity is evaluated over all 100 generated molecules per system, including failures. Because Vina scoring and PoseBusters checks are computationally expensive, all remaining metrics use only the first ten valid, single-fragment molecules in generation order. Each metric is first averaged within a system, then averaged across all systems that have at least one valid measurement. Uniqueness and diversity are computed within each system before averaging. Undefined values are omitted rather than zeroed, so the set of contributing systems can differ between metrics.

B.2 Median Performance with Holo and Apo Inputs

Tab. 10 and Tab. 11 complement the mean results in Tab. 2 and Tab. 3 with the same metrics and ordering. FlexiDD achieves median validity and PB-validity of 1.000 under both input settings, along with the lowest median Vina Min of -6.74 on holo inputs and -6.67 on apo inputs. FLOWR.root has lower median strain than FlexiDD under both settings, so the strain ranking reverses at the median.

Table 10: Median performance on holo protein structures. Format follows Tab. 2.

Method	Validity↑	PB-valid.↑	Strain↓	Vina↓	Vina Min↓	QED↑	SA↑	LogP	Lipinski↑	Div.2D↑	Uniq.2D↑
Reference	1.000	1.000	27.8	-7.06	-7.42	0.494	0.790	1.331	5.000	N/A	1.000
Pocket2Mol	0.988	0.600	140.6	-3.87	-5.32	0.640	0.794	1.285	5.000	0.865	1.000
TargetDiff	0.890	0.600	370.9	-5.30	-5.63	0.494	0.619	1.170	5.000	0.896	1.000
DecompDiff	0.860	0.600	222.3	-5.16	-5.77	0.507	0.662	1.660	5.000	0.873	1.000
IPDiff	0.930	0.200	2217.1	-5.66	-5.74	0.499	0.594	3.197	5.000	0.890	1.000
PocketXMol	0.900	0.800	87.6	-5.58	-6.17	0.523	0.751	1.443	5.000	0.872	1.000
FLOWR.root	0.990	1.000	32.7	-5.82	-6.66	0.525	0.727	1.571	5.000	0.819	1.000
ODesign	1.000	0.400	361.3	-3.28	-5.67	0.480	0.642	1.217	5.000	0.884	1.000
DrugFlow	0.860	0.800	160.8	-5.60	-6.09	0.582	0.747	1.973	5.000	0.864	1.000
FlexiDD	1.000	1.000	40.6	-5.82	-6.74	0.558	0.659	1.326	5.000	0.890	1.000

Table 11: Median performance with apo protein inputs. Format follows Tab. 3. Holo Min reports the locally minimized Vina score against the corresponding crystal holo protein.

Category	Method	Validity↑	PB-valid.↑	Strain↓	Vina↓	Vina Min↓	Holo Min↓	QED↑	SA↑	LogP	Lipinski↑	Div.2D↑	Uniq.2D↑
Reference		1.000	1.000	27.8	-7.06	-7.42	N/A	0.494	0.790	1.331	5.000	N/A	1.000
Static	Pocket2Mol	0.992	0.600	132.7	-2.98	-4.51	-4.24	0.618	0.798	1.083	5.000	0.868	1.000
	TargetDiff	0.870	0.400	397.2	-3.49	-4.31	-4.31	0.485	0.622	1.033	5.000	0.893	1.000
	DecompDiff	0.820	0.600	261.9	-3.12	-4.18	-4.50	0.486	0.644	1.385	5.000	0.879	1.000
	IPDiff	0.900	0.200	2568.5	-4.06	-4.42	-4.12	0.498	0.592	3.320	5.000	0.888	1.000
	PocketXMol	0.890	0.800	96.1	-4.79	-5.60	-4.87	0.529	0.732	0.989	5.000	0.874	1.000
	FLOWR.root	0.990	0.800	37.4	-2.62	-4.41	-5.78	0.516	0.722	1.360	5.000	0.852	1.000
	ODesign	0.990	0.400	360.5	-2.95	-5.02	-4.39	0.503	0.634	1.245	5.000	0.885	1.000
	DrugFlow	0.840	0.600	164.0	-2.67	-3.95	-4.46	0.584	0.734	1.883	5.000	0.873	1.000
Flexible	FlexFlow	0.780	0.400	139.1	-1.74	-3.82	-4.31	0.577	0.766	1.566	5.000	0.860	1.000
	Apo2Mol	0.970	0.200	476.0	-2.67	-4.29	-3.92	0.511	0.608	3.407	5.000	0.879	1.000
	YuelDesign	0.370	0.200	620.7	-2.08	-4.09	-2.10	0.214	0.572	0.284	4.000	0.893	1.000
	FlexiDD	1.000	1.000	39.1	-5.32	-6.67	-6.62	0.548	0.658	1.227	5.000	0.889	1.000

B.3 All-Atom Pairwise-Distance Distributions

Fig. 9 and Fig. 10 compare the intramolecular distance distributions of generated ligands with the reference ligands in the crystal structures. We collect all heavy-atom pair distances below 12 Å within each ligand. Each histogram is normalized to unit mass and agreement is quantified by the Jensen-Shannon distance (JSD) with natural logarithms.

On holo inputs, FlexiDD obtains a JSD of 0.072, compared with 0.054 for DecompDiff and FLOWR.root and 0.070 for DrugFlow. On apo inputs, FlexiDD achieves the lowest JSD among the four compared methods at 0.063. These comparisons supplement the bond-length analysis in Fig. 3 by assessing distances across the entire ligand.

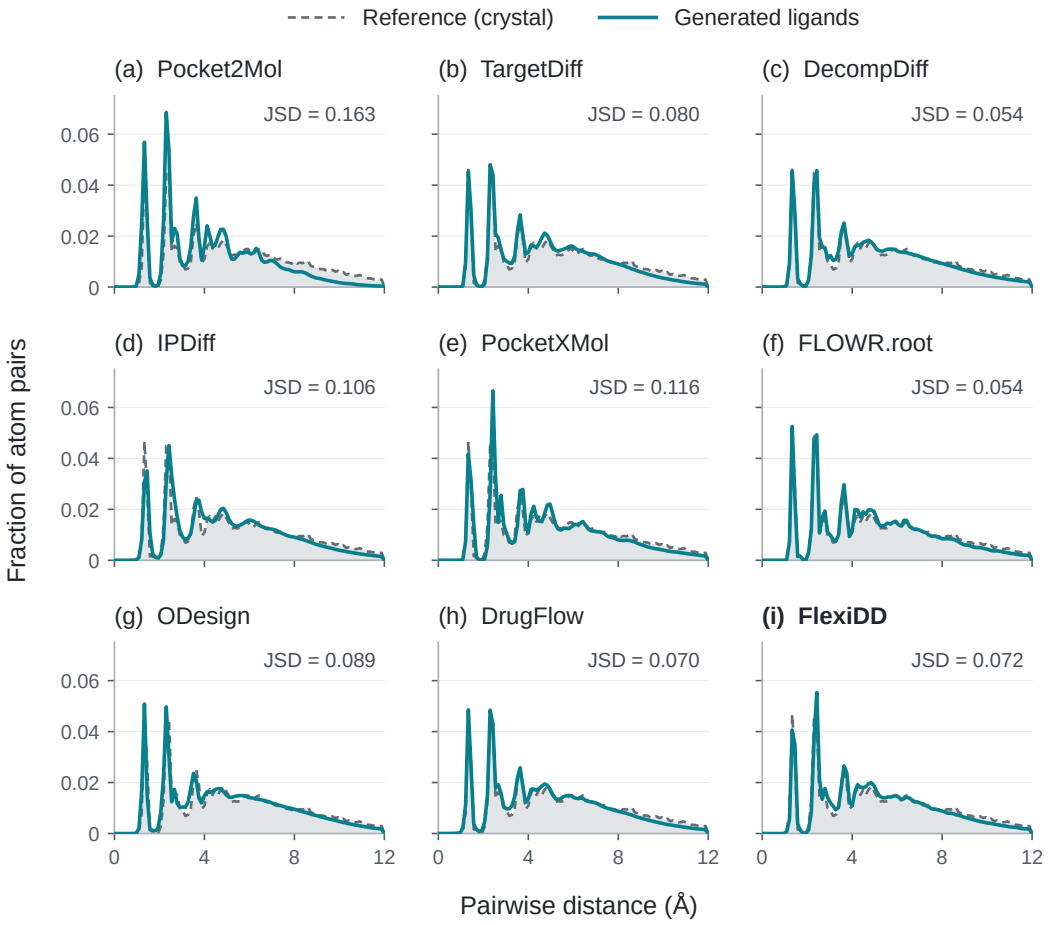

Figure 9: All-atom pairwise-distance distributions of ligands generated from holo inputs. Each panel overlays one method with the crystallographic Reference. Histograms are normalized to unit mass for heavy-atom pair distances below 12 Å. JSD denotes the Jensen-Shannon distance; lower values indicate closer agreement.

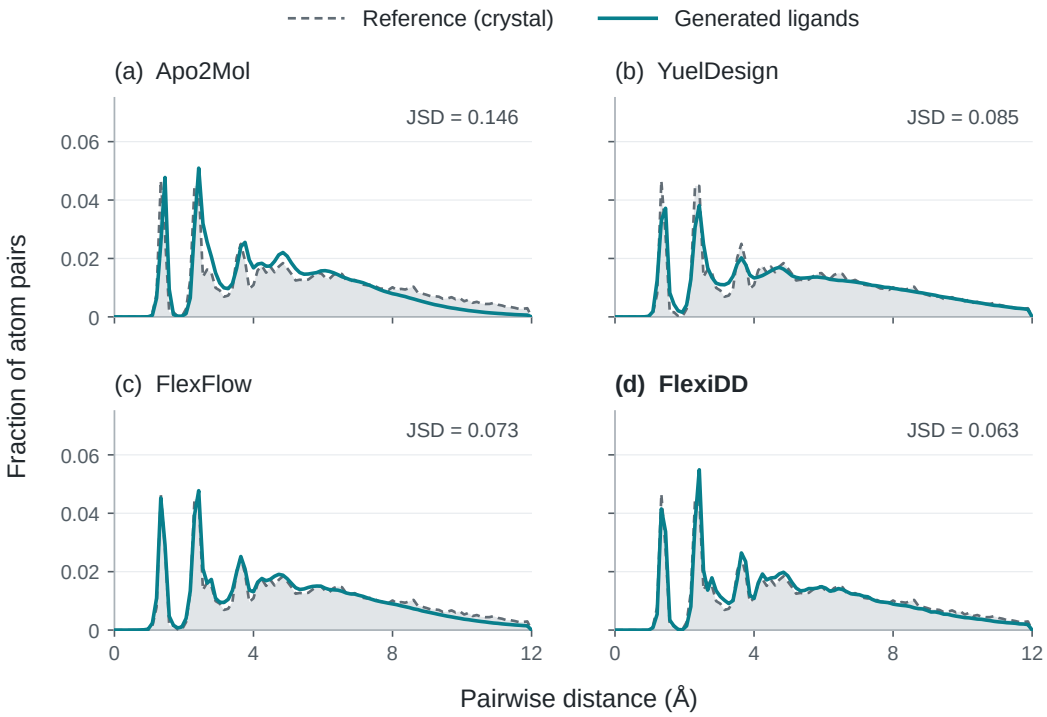

Figure 10: All-atom pairwise-distance distributions of ligands generated from apo inputs. Each panel overlays one method with the same crystallographic Reference used in Fig. 9. Histogram normalization, axis ranges, and JSD computation follow the same protocol. FLEXIDD has the lowest JSD among the four compared methods.

B.4 Detailed Pocket Transfer Results

Tab. 12 expands the pocket-transfer comparison in Sec. 4.5 to all reported metrics. When FLOWR.root receives FLEXIDD-generated pockets instead of crystal holo pockets, diversity and uniqueness increase while chemical validity decreases slightly. PocketXMol shows a similar but smaller diversity gain, with strain and Vina Min slightly less favorable on generated pockets.

Table 12: Detailed mean results for the pocket-transfer experiment in Sec. 4.5.

Model	Pocket	Validity↑	PB-valid.↑	Strain↓	Vina↓	Vina Min↓	QED↑	SA↑	LogP	Lipinski↑	Div.2D↑	Uniq.2D↑
FLOWR.root	Crystal apo	0.968	0.573	96.8	6.72	-2.18	0.519	0.717	1.033	4.636	0.787	0.965
	Crystal holo	0.972	0.797	102.3	-5.14	-6.44	0.513	0.728	1.245	4.599	0.732	0.944
	Apo-generated	0.936	0.817	93.1	-5.25	-6.49	0.531	0.716	1.158	4.604	0.825	0.991
	Holo-generated	0.936	0.832	97.4	-5.31	-6.63	0.532	0.716	1.142	4.634	0.832	0.995
PocketXMol	Crystal apo	0.783	0.783	124.1	-3.78	-5.01	0.518	0.725	0.836	4.676	0.867	0.996
	Crystal holo	0.861	0.810	127.9	-4.63	-5.86	0.520	0.741	1.174	4.664	0.858	0.996
	Apo-generated	0.856	0.826	134.5	-4.78	-5.67	0.525	0.728	1.024	4.685	0.872	0.998
	Holo-generated	0.866	0.808	136.8	-4.98	-5.74	0.521	0.724	0.998	4.604	0.870	0.997

B.5 Sampling Trajectories and Pocket Views

We visualize sampling for two PLINDER systems, 8C6N and 8SWE, under apo and holo conditioning. Each trajectory shows 20 snapshots saved every five sampling steps, followed by the final output. Protein cartoons follow Cα guide atoms and ligand sticks use the final molecule's connectivity throughout, so intermediate panels reflect coordinate and atom type changes.

The pocket views compare the crystal holo complex with the generated complexes under both input conditions. Annotations give per-example QED, SA, and Vina scores following the definitions in Apx. B.1.

B.5.1 8C6N

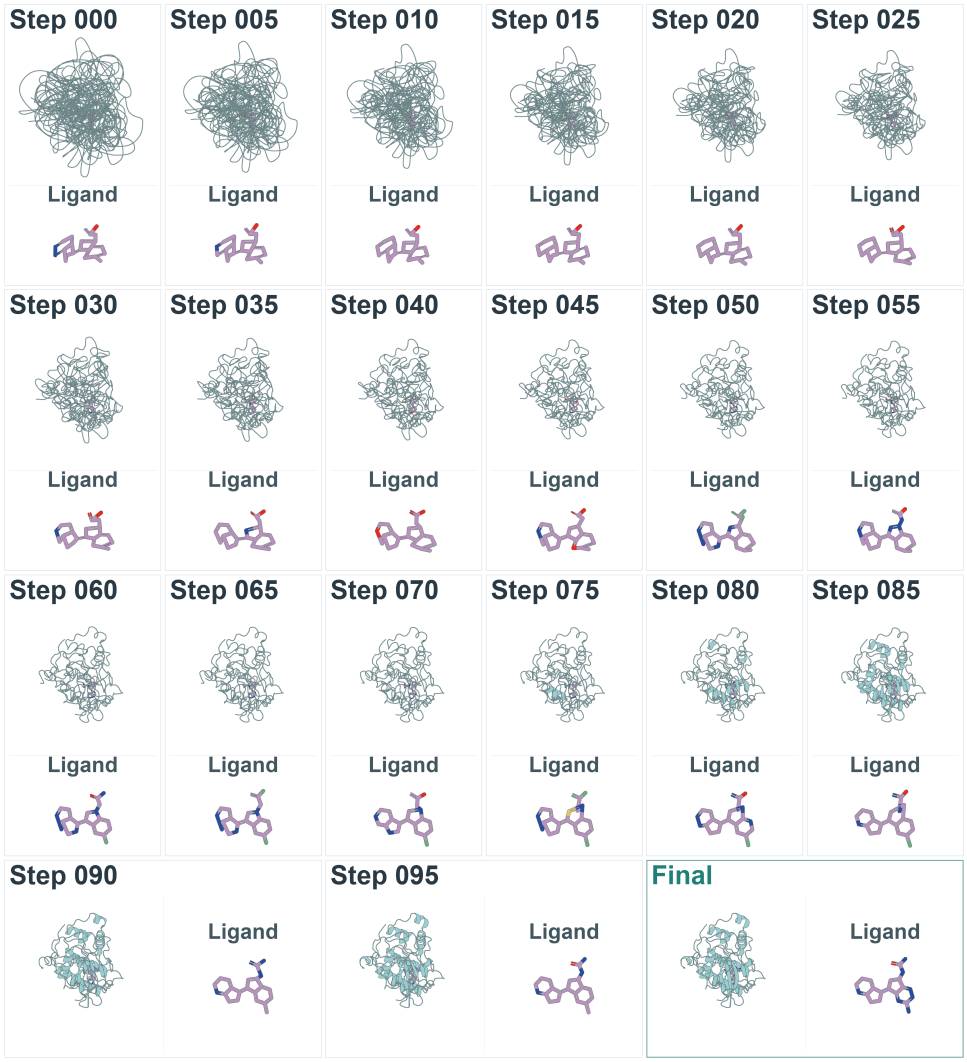

Figure 11: Sampling trajectory for 8C6N conditioned on an apo protein.

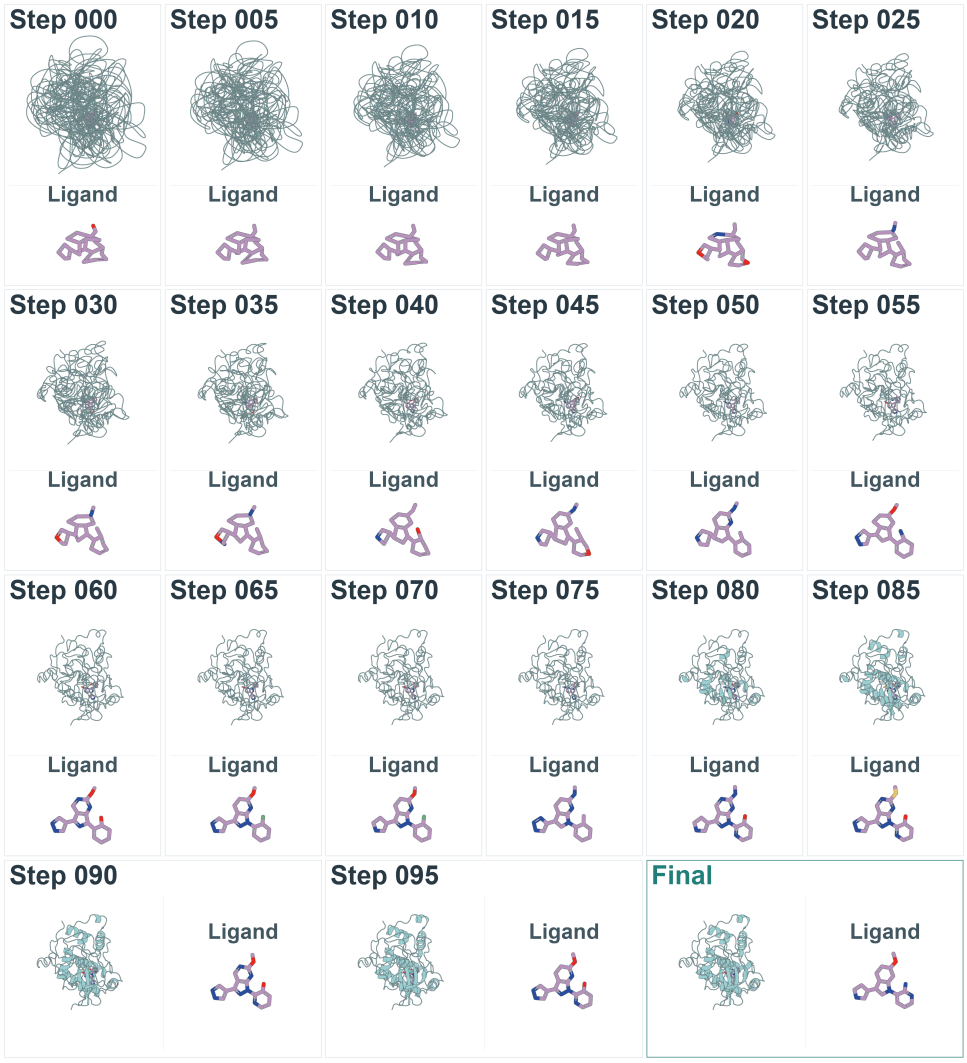

Figure 12: Sampling trajectory for 8C6N conditioned on a holo protein.

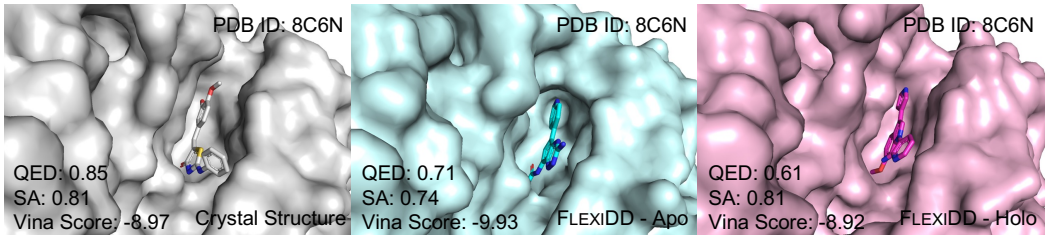

Figure 13: Pocket views for 8C6N comparing the crystal holo complex with FLEXIDD outputs under apo and holo conditioning. Generated ligands are shown within the corresponding generated protein surface.

B.5.2 8SWE

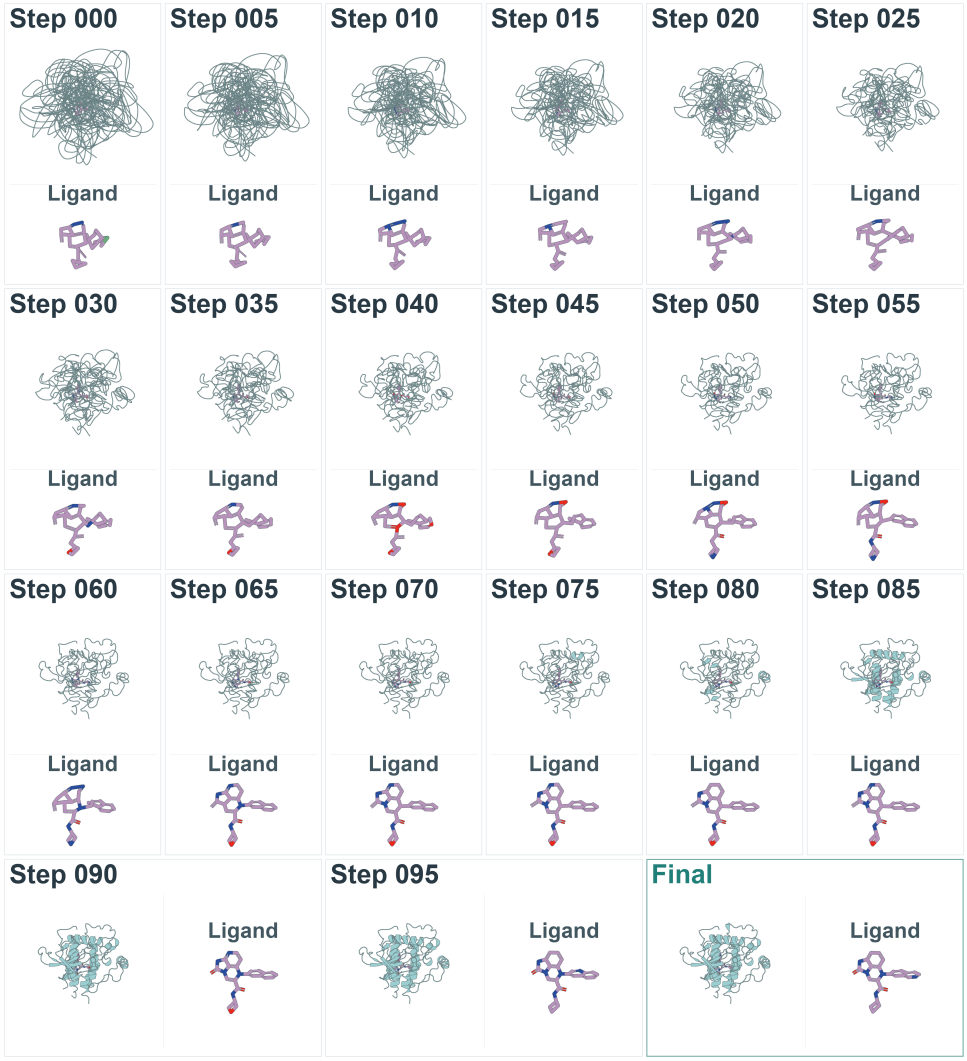

Figure 14: Sampling trajectory for 8SWE conditioned on an apo protein.

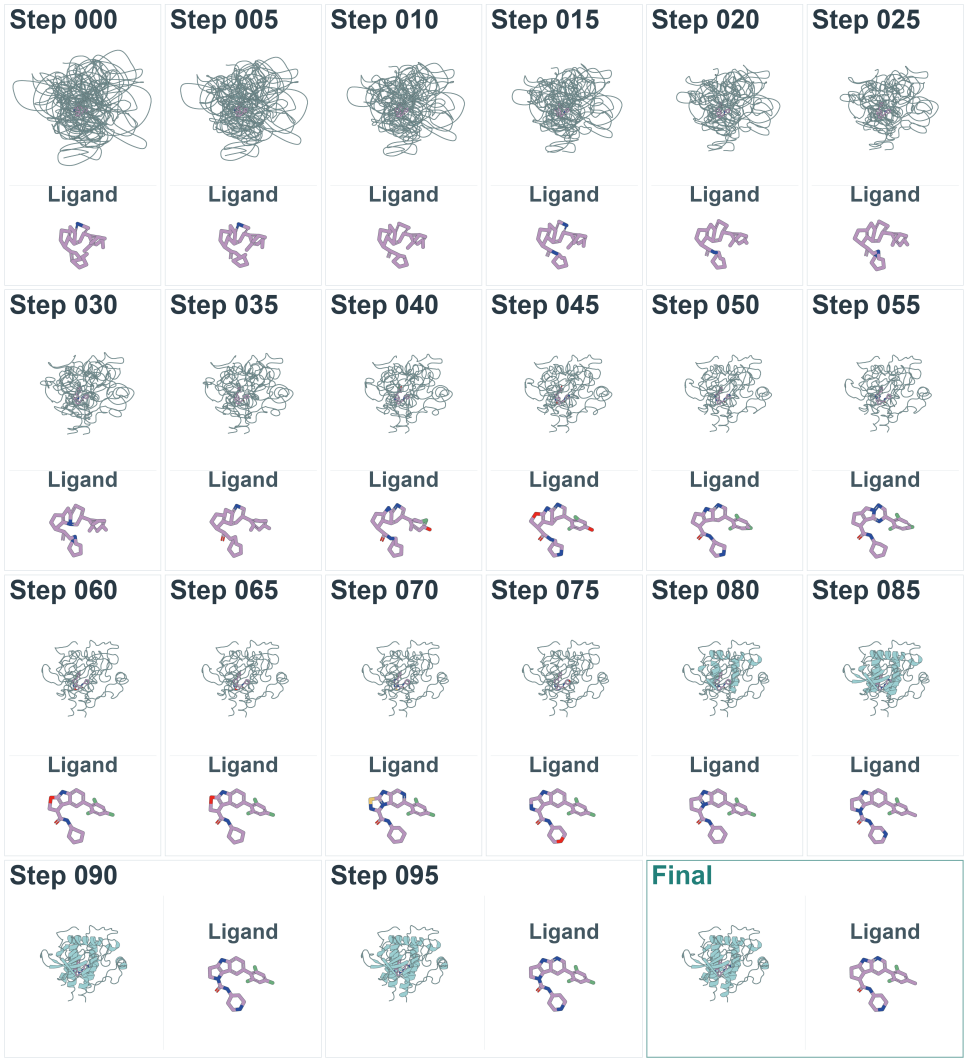

Figure 15: Sampling trajectory for 8SWE conditioned on a holo protein.

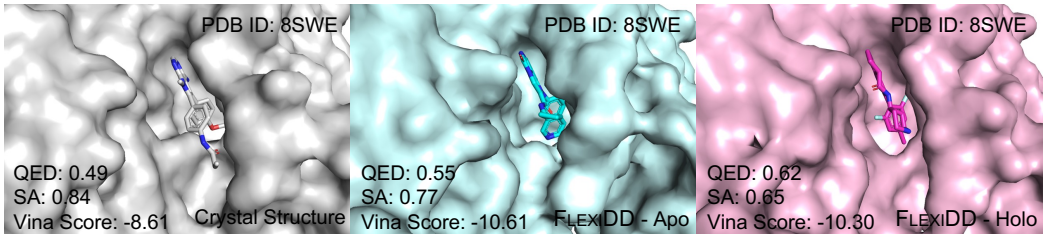

Figure 16: Pocket views for 8SWE comparing the crystal holo complex with FLEXIDD outputs under apo and holo conditioning. Generated ligands are shown within the corresponding generated protein surface.

B.6 Stage 1 Pretraining Ablation

We ablate the effect of Stage 1 protein pretraining by comparing FlexiDD trained with and without it. Tab. 13 reports the results on the PLINDER benchmark for both apo and holo inputs.

With Stage 1 pretraining, structure-sensitive metrics improve consistently under both input settings, including three to four percentage points higher PB-validity and stronger Vina scores, while graph-level properties such as validity and diversity remain comparable. The gap concentrates on pose and binding quality rather than chemical plausibility, indicating that the benefit stems from better pocket geometry. Learning a protein structure prior before joint training thus helps the model reconstruct more accurate pockets and generate ligands with stronger binding.

Table 13: Ablation of Stage 1 protein pretraining.

Input	Variant	Validity↑	PB-valid.↑	Strain↓	Vina↓	Vina Min↓	QED↑	SA↑	LogP	Lipinski↑	Div.2D↑	Uniq.2D↑
Apo	w/ Stage 1	0.976	**0.916**	**59.4**	**-5.02**	**-6.42**	**0.545**	**0.665**	1.124	**4.836**	**0.873**	0.985
	w/o Stage 1	**0.979**	0.878	61.2	-4.88	-6.21	0.529	0.652	1.288	4.778	0.864	**0.989**
Holo	w/ Stage 1	0.979	**0.916**	76.8	**-5.39**	**-6.67**	**0.550**	**0.658**	1.185	4.761	**0.871**	0.980
	w/o Stage 1	**0.982**	0.887	**60.5**	-5.11	-6.54	0.535	0.647	1.141	**4.796**	0.867	**0.997**

C Limitations

FlexiDD requires a pre-specified binding site as input and does not perform binding-site prediction. In most SBDD scenarios, this information is available from prior knowledge or computational tools, but integrating blind pocket prediction remains an open direction.

The current implementation processes a single protein chain. Binding sites that span protein-protein interfaces or involve multi-chain assemblies are therefore not directly supported. Extending the framework to multi-chain inputs is a natural next step.

All evaluations in this work are computational, relying on physics-based scoring functions and geometric validity checks. Experimental validation of generated molecules, such as binding assays or crystallographic confirmation, would further establish the practical utility of the framework but is beyond the scope of this study.