UNI^2: UNIFIED TRAINING OF LATENT DIFFUSION OVER UNIFIED MODALITIES FOR DE NOVO BINDER DESIGN

Anonymous authors
Paper under double-blind review

ABSTRACT

Structure-based binder design asks a model to generate a peptide, antibody, or small molecule that binds a given target. One family of approaches handles this task with latent diffusion: an autoencoder compresses full-atom complexes into a learned latent space, and a diffusion model generates there. That latent space is usually obtained in two steps: an autoencoder is fit to reconstruct complexes, then frozen, and a diffusion model is trained on top of it. The autoencoder therefore meets the target only as data to be compressed, never as a constraint on what the generator will later have to produce, and a space that reconstructs cleanly can still hand the diffusion model a geometry that is difficult to occupy. We formulate autoencoding and latent diffusion as a unified optimization problem, jointly learning the latent representation and its generative dynamics. By allowing diffusion gradients to flow into the autoencoder, the diffusion objective directly shapes the latent manifold, aligning representation learning with the generative task while retaining the structural information required for reconstruction. On peptide, antibody, and small molecule benchmarks, our training strategy improves structural accuracy, physical plausibility, and binding energy together. These gains persist when the three modalities are trained together in one model. The change is visible in the latent space itself. Measured with a standardized sliced Wasserstein distance, the gap between posterior latents and generated latents narrows once training is unified. Finally, an ablation that treats $\mathbf{z}_x$ and $\mathbf{z}_h$ separately measures what unified training contributes through each channel: $\mathbf{z}_x$ accounts for the gain in geometry, $\mathbf{z}_h$ accounts for the gain in chemistry and interaction, and only when both receive gradients does the model become Pareto-optimal.

1 INTRODUCTION

Designing a molecule that binds a given protein target is the central problem of therapeutic discovery, and it recurs at several length scales: small molecules (tens of atoms), peptides (a handful of residues), and antibody complementarity embedded in a large framework. Recent work has shown that a *single* generative model can serve all three modalities by adopting a residue-level, full-atom representation and a shared latent diffusion backbone. The dominant paradigm is two-stage: (i) train a variational autoencoder (VAE) to compress full-atom blocks into a per-block latent pair $(\mathbf{z}_h, \mathbf{z}_x)$, consisting of an invariant chemistry code and an equivariant coordinate code, and (ii) freeze the encoder and train a diffusion model in that latent space.

This paradigm inherits a well-known but rarely addressed mismatch. Stage (i) minimizes reconstruction error subject to a weak KL penalty; nothing in this objective asks the latent distribution to be *easy to sample*. In geometric terms, the autoencoder is free to place its aggregate posterior on a latent manifold that supports reconstruction without necessarily being easy for the diffusion model to reproduce. The diffusion model in stage (ii) must spend its capacity learning the score of a manifold it had no part in shaping. Decoding then amplifies small latent errors into steric clashes and unfavourable interaction geometries, which are exactly the failure modes that energy-based metrics punish.

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.

Rethinking the latent space as the object of design. Latent representations are typically learned to preserve the information needed for reconstruction. Their suitability for subsequent diffusion modeling, however, may also depend on the geometry and organization of the latent distribution. This suggests that reconstruction quality alone may not fully characterize how well a representation supports generation, raising a natural question: *could the latent space benefit from feedback from the generative model while it is being learned?*

In conventional two-stage training, freezing the autoencoder makes the learned latent distribution a fixed target for diffusion modeling. Any mismatch must then be addressed by adapting the diffusion model alone, even when reshaping the representation could make that target more amenable to generation. The encoder receives no feedback about which aspects of its latent geometry facilitate or hinder this process.

We therefore introduce UNI^2, a framework that unifies both training and molecular modalities through a shared autoencoder and diffusion model for antibodies, peptides, and small molecules. Reconstruction and diffusion objectives guide representation learning together, with diffusion gradients propagated into the encoder rather than detached at the latent boundary. This feedback allows the latent manifold to adapt to generative modeling while remaining constrained by molecular reconstruction. The unified training formulation applies across all three modalities without adding model parameters or inference-time overhead.

Our contributions are:

- **Unified latent generative modeling for binder design.** We revisit the standard two-stage VAE + LDM pipeline from the perspective of latent manifold learning and introduce a unified training formulation. By propagating diffusion gradients into the encoder, this formulation allows the latent representation to adapt to the generative objective alongside reconstruction, improving generation quality across all three modalities (Sections 4.1–4.3).
- **Consistent gains under both single- and multi-modality training.** One-stage training helps whether the model is trained on one modality or on all three jointly, indicating the effect is a property of the training paradigm rather than of data scale.
- **A gradient-flow ablation that explains the mechanism.** By selectively detaching the diffusion gradient with respect to $\mathbf{z}_h$, to $\mathbf{z}_x$, or to both, we show that (a) even fully detached one-stage training helps, (b) $\mathbf{z}_x$ gradients drive geometric accuracy, (c) $\mathbf{z}_h$ gradients drive chemistry and interaction quality, and (d) only keeping both is Pareto-optimal (Section 5).
- **Latent-space evidence for the mechanism.** Posterior-to-generated distribution measurements and PCA with KDE contours show that one-stage training reduces the standardized Sliced Wasserstein distance between posterior and generated latents, most clearly in the content channel, while preserving VAE reconstruction quality (Section 6).

2 RELATED WORK

Structure-based binder design. Peptide design has been approached with backbone diffusion followed by inverse folding and relaxation (Watson et al., 2023; Dauparas et al., 2022; Alford et al., 2017), with multi-modal flow matching (Li et al., 2024), and with latent diffusion over full-atom residues (Kong et al., 2024). Antibody design has been tackled with equivariant refinement networks (Kong et al.; 2023) and with diffusion over amino-acid types, Cα coordinates and residue frames (Luo et al., 2022; Lin et al., 2024). Small molecule pocket-conditioned generation spans autoregressive (Peng et al., 2022; Liu et al.; Luo et al., 2021), diffusion (Guan et al., 2023; 2024; Lin et al., 2025a; Schneuing et al., 2024), fragment (Zhang et al., 2023; Lin et al., 2023), voxel (Masuda et al., 2010; Pinheiro et al., 2024), and Bayesian-flow (Qu et al., 2024) formulations. Among these, UNIMOMO is the most complete: one backbone, one latent space, and three modalities, covering peptides, antibody, and small molecules. It is therefore the natural testbed for a change in training procedure alone. We keep the backbone fixed and vary only how the latent space is learned, on each modality in isolation and on all three trained jointly.

Latent diffusion and the two-stage boundary. Latent diffusion decouples perceptual compression from generative modeling for efficiency. In the image domain the autoencoder is deliberately

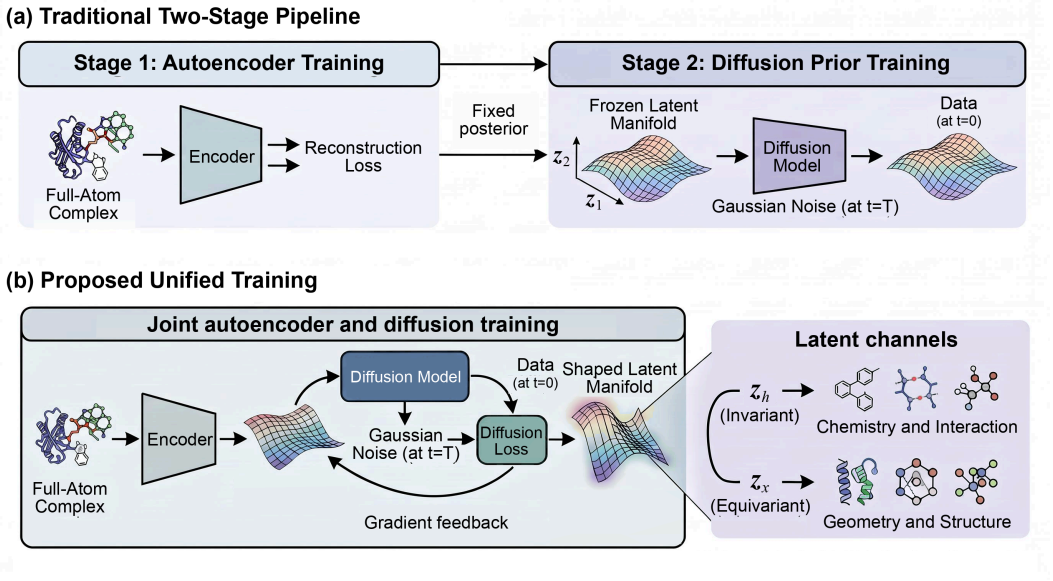


Figure 1: Overview of the training paradigms. **(a)** Conventional two-stage training first learns a reconstruction-oriented latent space and then freezes it while fitting the latent denoiser. Consequently, the encoder is unaware of whether its codes are reachable under the learned reverse process. **(b)** Unified Training Strategy removes this boundary and jointly adapts the encoder, decoder, and denoiser. Diffusion gradients through the invariant $\mathbf{z}_h$ and equivariant $\mathbf{z}_x$ channels shape a latent space that is simultaneously decodable, regularized, and reachable from noise.

kept weakly regularized and frozen. Several lines of work have examined how latent statistics affect the downstream prior through KL/VQ regularization strength, latent normalization, or "prior matching" losses, while retaining sequential optimization of the two objectives. End-to-end alternatives exist for VAE + flow and VAE + score models in low-dimensional settings; to our knowledge, joint one-stage training has not been shown to be the better default for full-atom, multi-modality biomolecular binder design, where the evaluation is not perceptual but physical (clashes, dihedral distributions, docking and Rosetta energies).

3 Method

3.1 Full-atom generation through a factorized latent interface

A shared representation across binder classes. A target–binder complex $\mathcal{C} = (\mathcal{P}, \mathcal{B})$ contains a fixed target pocket $\mathcal{P}$ and a binder $\mathcal{B}$ to be generated. We represent $\mathcal{B}$ as a set of *blocks*: an amino-acid residue for peptides and antibodies, and a chemical fragment for small molecules. This abstraction exposes the same learning problem across scales, namely predicting what each block is and where it lies in three-dimensional space, while retaining full-atom output geometry.

Factorized latent variables. An SE(3)-equivariant encoder E_ϕ compresses every binder block i into two complementary latent variables,

$$(\mathbf{z}_h^{(i)}, \mathbf{z}_x^{(i)}) = E_\phi(\mathcal{C}), \qquad \mathbf{z}_h^{(i)} \in \mathbb{R}^d, \quad \mathbf{z}_x^{(i)} \in \mathbb{R}^3. \tag{1}$$

Here $\mathbf{z}_h$ is invariant to global rigid transformations and represents block identity and local chemistry, whereas $\mathbf{z}_x$ is equivariant and represents spatial arrangement. The decoder $D_\psi(\mathbf{z}_h, \mathbf{z}_x, \mathcal{P})$ reconstructs block types and full-atom coordinates. Its objective combines type, coordinate, and local-distance reconstruction terms with channel-wise KL regularization, denoted compactly as $\mathcal{L}_{\text{recon}}$ and $\mathcal{L}_{\text{KL}}$.

Conditional latent diffusion. For diffusion time t, Gaussian perturbations produce $\mathbf{z}_{h,t} = \alpha_t \mathbf{z}_h + \sigma_t \epsilon_h$ and $\mathbf{z}_{x,t} = \alpha_t \mathbf{z}_x + \sigma_t \epsilon_x$. A pocket-conditioned denoiser $\epsilon_\theta(\mathbf{z}_{h,t}, \mathbf{z}_{x,t}, t, \mathcal{P})$ predicts the injected

noise in both channels:

$$\mathcal{L}^h_{\text{diff}} = \mathbb{E}_{t,\epsilon_h}\left[\|\epsilon_h - \epsilon_{\theta,h}(\mathbf{z}_{h,t}, \mathbf{z}_{x,t}, t, \mathcal{P})\|_2^2\right], \quad \mathcal{L}^x_{\text{diff}} = \mathbb{E}_{t,\epsilon_x}\left[\|\epsilon_x - \epsilon_{\theta,x}(\mathbf{z}_{h,t}, \mathbf{z}_{x,t}, t, \mathcal{P})\|_2^2\right]. \quad (2)$$

The invariant and equivariant predictions share contextual computation while preserving their transformation properties. At inference, reverse diffusion generates a latent pair conditioned on $\mathcal{P}$, which is then decoded into a full-atom binder.

Limitation of a frozen latent interface. Conventional two-stage training first optimizes (E_ϕ, D_ψ) for reconstruction, then freezes the encoder and trains the denoiser on the resulting latent distribution. This creates a fixed interface: the denoiser must adapt to the encoder posterior, but the encoder receives no feedback about whether its latent space is easy for the diffusion model to sample from. As shown later in our latent-space analysis, this can lead to a mismatch between posterior and generated latent manifolds, motivating unified training of the autoencoder and latent diffusion model.

3.2 Diffusion-aware latent co-adaptation

Removing the optimization boundary. UNI^2 replaces the one-way interface with a bidirectional one. We initialize the autoencoder and denoiser together and optimize Eq. 3 from scratch. The forward path is unchanged, but the diffusion loss is evaluated on latents produced by the *current* encoder and remains differentiable through them:

$$\nabla_\phi \mathcal{L} = \lambda_{\text{vae}} \nabla_\phi \mathcal{L}_{\text{recon}} + \nabla_\phi \mathcal{L}_{\text{KL}} + \underbrace{\nabla_\phi\left(\mathcal{L}^x_{\text{diff}} + \lambda_h \mathcal{L}^h_{\text{diff}}\right)}_{\text{diffusion-aware latent shaping; zero in two-stage training}}. \quad (3)$$

Consequently, the denoiser learns the encoder's distribution while the encoder simultaneously learns a representation the denoiser can traverse. We refer to this feedback loop as *latent co-adaptation*. It turns the latent space from a frozen compression artifact into a task-aware generative interface.

A rate–geometry–reachability trade-off. The joint objective has a useful interpretation. Reconstruction preserves the information needed to recover atom types and coordinates; KL regularization controls the latent rate; diffusion gradients penalize representations that are difficult to reach from noise under the actual conditional generator. Thus the learned code must be simultaneously *decodable*, *regularized*, and *reachable*. λ_{vae} controls the fidelity–reachability balance, while λ_h allocates diffusion pressure between invariant semantics and equivariant geometry. Importantly, this mechanism changes neither the architecture nor the sampling procedure: its entire effect is learned during training, so inference has the same parameter count and computational path as the two-stage model.

3.3 Dissecting the two-channel feedback path

The factorization in Eq. 1 gives a controlled way to test whether the gains come merely from simultaneous scheduling or from diffusion gradients reshaping particular information channels. We insert $\text{sg}[\cdot]$ (stop-gradient) at the latent boundary and define four regimes:

Regime	Latent fed to the diffusion model
w/o grad h-x	$(\text{sg}[\mathbf{z}_h],\ \text{sg}[\mathbf{z}_x])$
w/o grad h	$(\text{sg}[\mathbf{z}_h],\ \mathbf{z}_x)$
w/o grad x	$(\mathbf{z}_h,\ \text{sg}[\mathbf{z}_x])$
w grad h-x (default)	$(\mathbf{z}_h,\ \mathbf{z}_x)$

The fully detached variant remains one-stage in time: both modules are optimized throughout training and the encoder distribution continues to move, but the diffusion objective cannot alter that movement. It therefore isolates the benefit of a shared schedule from the benefit of latent shaping. The two channel-specific variants then localize the feedback: retaining only the $\mathbf{z}_x$ path asks whether diffusion-aware geometry is sufficient, while retaining only $\mathbf{z}_h$ probes semantic and chemical shaping. The default model keeps both paths and lets the two channels co-adapt under a common physical generation objective.

4 EXPERIMENTS

Datasets. We benchmark on three binder categories: peptides, antibodies, and small molecules. For peptides, we train on the PepBench and ProtFrag data of Kong et al. (2024), which provide 4,157 protein–peptide complexes and 70,498 synthetic examples, with 114 complexes held out for validation; evaluation uses the LNR benchmark (Kong et al., 2024; Tsaban et al., 2022), 93 protein–peptide complexes whose peptides span from 4 to 25 residues. For antibodies we follow the standard protocol of Luo et al. (2022) and Kong et al., in which SAbDab (Dunbar et al., 2014) entries deposited before September 24th are split into training and validation sets and 60 antigen–antibody complexes from RAbD (Adolf-Bryfogle et al., 2018) form the test set. After removing training entries with sequence identity above 40% to any test case (Kong et al.), 9,473 training and 400 validation complexes remain. For small molecules, we use CrossDocked2020 (Francoeur et al., 2020) with the established split of Peng et al. (2022), which provides 99,900 complexes for training, 100 for testing, and 100 additional complexes randomly drawn from the training pool for validation.

Setup and notation. We follow the evaluation protocols of the respective domains verbatim so that all numbers are directly comparable with published baselines. UNIMOMO denotes the two-stage baseline and UNI^2 our one-stage model; each appears in two variants: *(single)*, trained only on the domain-specific data, and *(all)*, trained jointly on all three datasets. In all tables the best result is **bold** and the second best is underlined; for small molecules we additionally underline the third best, following the CBGBench convention (Lin et al., 2025b).

4.1 PEPTIDE DESIGN

Setup. We report Amino Acid Recovery (AAR), the fraction of generated residues matching the reference peptide, computed with the standard bioinformatics alignment (Needleman & Wunsch, 1970; Henikoff & Henikoff, 1992). As AAR is known to be an unreliable proxy (Kong et al., 2024) we only include it for completeness. Complex RMSD (C-RMSD) is the Cα RMSD after aligning on the target protein, and Ligand RMSD (L-RMSD) after aligning on the peptide itself. ΔG is the pyRosetta binding energy (Alford et al., 2017) and IMP the percentage of targets where the generated peptide outperforms the native binder. Clash_{in} and $\text{Clash}_{\text{out}}$ are residue-level clash rates within the peptide and between peptide and target, respectively, defined by Cα distances below $3.6574\,\text{Å}$ (Ye et al., 2025). JSD_{bb} and JSD_{sc} are Jensen–Shannon divergences between generated and dataset dihedral distributions (backbone and sidechain, 10° bins) (Dunbrack Jr & Cohen, 1997). Sequence and struct diversity are ratios of unique clusters to total generations at 40% sequence identity and $2\,\text{Å}$ RMSD thresholds. For each target in the test set, we generate 100 candidate peptides for evaluation.

Results. Table 1 compares UNI^2 with the corresponding UNIMOMO model on peptide design. Under the single-task setting, UNI^2 substantially improves structural accuracy, reducing C-RMSD and L-RMSD from 2.48 Å to 1.92 Å, and from 1.48 Å to 0.85 Å, respectively. It also improves binding quality, with ΔG becoming 5.19 (kcal/mol) more favorable and IMP increasing from 29.03% to 38.71%. In terms of steric quality, UNI^2 reduces Clash_{in} from 1.53% to 0.22% and $\text{Clash}_{\text{out}}$ from 0.94% to 0.15%. The distributional similarity is also improved, with JSD_{bb} and JSD_{sc} reduced from 0.390 to 0.276 and from 0.365 to 0.219, respectively.

Under the all-task setting, UNI^2 shows consistent gains over UNIMOMO in most geometric and energetic metrics. C-RMSD and L-RMSD are reduced by 0.41 Å and 0.57 Å, respectively, while ΔG improves by 1.79 (kcal/mol) and IMP increases from 40.86% to 44.09%. The clash rates are further reduced, with Clash_{in} decreasing from 0.45% to 0.11% and $\text{Clash}_{\text{out}}$ from 0.93% to 0.07%. For distributional metrics, UNI^2 slightly improves side-chain geometry, reducing JSD_{sc} from 0.180 to 0.175, while JSD_{bb} increases from 0.205 to 0.246. Overall, these results show that UNI^2 improves the base model in structural accuracy, binding quality, and steric plausibility under both single-task and all-task settings.

4.2 ANTIBODY DESIGN

Setup. CDRs are defined with the Chothia numbering system (Chothia & Lesk, 1987), following Luo et al. (2022). Chothia is preferable to IMGT (Lefranc et al., 2003) here because IMGT is vulnerable to trivial unigram patterns that inflate AAR to 39% (Kong et al., 2023), whereas the

Table 1: Peptide design on the held-out test set. ↑/↓ indicate whether higher or lower is better. Best in each column in **bold**, second best underlined.

Model	AAR↑	C-RMSD↓	L-RMSD↓	ΔG↓	IMP↑	Clash_{in}↓	$\text{Clash}_{\text{out}}$↓	JSD_{bb}↓	JSD_{sc}↓	Seq.↑	Struct.↑
Reference	–	–	–	−37.25	–	0.31%	0.88%	–	–	–	–
RFDiffusion (Watson et al., 2023)	34.68%	4.69	1.88	−13.47	5.38%	**0.06%**	13.58%	0.273	0.798	0.155	0.616
PepFlow (Li et al., 2024)	35.47%	2.87	1.79	−21.71	15.22%	2.72%	4.62%	0.240	0.693	0.530	0.507
PepGLAD (Kong et al., 2024)	38.62%	2.74	1.60	−23.12	18.28%	1.82%	1.66%	0.474	0.398	**0.687**	**0.698**
UNIMOMO (single)	37.59%	2.48	1.48	−28.72	29.03%	1.53%	0.94%	0.390	0.365	0.626	0.629
UNI^2 (single)	38.91%	1.92	0.85	−33.91	38.71%	0.22%	0.15%	0.276	0.219	0.638	0.502
UNIMOMO (all)	39.45%	2.19	1.27	−34.35	40.86%	0.45%	0.93%	**0.205**	0.180	0.617	0.573
UNI^2 (all)	**39.57%**	**1.78**	**0.70**	**−36.14**	**44.09%**	0.11%	**0.07%**	0.246	**0.175**	0.623	0.460

Table 2: Antibody design. Metrics are reported over the best of $n \in \{1, 10, 100\}$ candidates per complex; deterministic baselines only admit $n = 1$. Best in **bold**, second best underlined.

	$n = 1$			$n = 10$			$n = 100$			Rationality			
Model	AAR↑	RMSD↓	IMP↑	AAR↑	RMSD↓	IMP↑	AAR↑	RMSD↓	IMP↑	Clash_{in}↓	$\text{Clash}_{\text{out}}$↓	JSD_{bb}↓	JSD_{sc}↓
Reference	–	–	–	–	–	–	–	–	–	0.08%	0.02%	–	–
MEAN (Kong et al.)	29.13%	1.87	6.67%	–	–	–	–	–	–	0.96%	0.16%	0.529	–
dyMEAN (Kong et al., 2023)	31.65%	8.21	11.86%	–	–	–	–	–	–	1.02%	2.98%	0.542	0.702
GeoAB-R (Lin et al., 2024)	**32.04%**	**1.67**	6.67%	–	–	–	–	–	–	0.59%	0.11%	0.529	–
DiffAb (Luo et al., 2022)	24.60%	2.77	10.34%	38.42%	2.08	34.48%	49.74%	1.46	60.34%	0.31%	0.25%	0.268	–
GeoAB-D (Lin et al., 2024)	29.74%	1.73	6.67%	38.20%	1.58	20.00%	45.96%	1.50	40.00%	0.75%	0.07%	0.430	–
UNIMOMO (single)	20.44%	2.71	15.00%	39.04%	1.90	35.00%	48.78%	1.39	63.33%	0.25%	0.06%	0.278	0.284
UNI^2 (single)	20.80%	1.99	18.33%	39.11%	1.38	38.33%	50.26%	1.08	65.00%	0.14%	0.05%	0.240	0.270
UNIMOMO (all)	21.44%	2.52	13.33%	**42.05%**	1.44	41.67%	**52.34%**	1.04	65.00%	0.18%	**0.03%**	0.224	0.221
UNI^2 (all)	19.32%	1.95	**23.33%**	40.11%	**1.18**	**43.33%**	50.32%	**0.85**	**68.33%**	**0.12%**	**0.03%**	**0.216**	**0.217**

same pattern reaches only 23% under Chothia. We report AAR, RMSD and IMP together with the CDR rationality metrics defined above, evaluated over $n \in \{1, 10, 100\}$ candidates per complex. Baselines include both deterministic predictors, namely MEAN (Kong et al.), dyMEAN (Kong et al., 2023), and GeoAB-R (Lin et al., 2024), as well as generative models, namely DiffAb (Luo et al., 2022) and GeoAB-D (Lin et al., 2024).

Results. Table 2 reports the antibody design results. Under the single-task setting, UNI^2 improves over UNIMOMO, reducing RMSD from 1.39 Å to 1.08 Å and increasing IMP from 63.33% to 65.00%. It also improves rationality metrics, with Clash_{in} decreasing from 0.25 % to 0.14 %, $\text{Clash}_{\text{out}}$ from 0.06 % to 0.05 %, JSD_{bb} from 0.278 to 0.240 , and JSD_{sc} from 0.284 to 0.270 . Under the all-task setting, UNI^2 also improves the base model, reducing RMSD from 1.04 Å to 0.85 Å and increasing IMP from 65.00% to 68.33%. In addition, UNI^2 achieves better rationality, reducing Clash_{in} from 0.18% to 0.12% and improving JSD_{bb} and JSD_{sc} from 0.224 to 0.216 and from 0.221 to 0.217 , respectively, while maintaining the same $\text{Clash}_{\text{out}}$ of 0.03%. Overall , these results show that UNI^2 improves structural accuracy and geometric rationality over UNIMOMO in antibody design.

4.3 SMALL MOLECULE DESIGN

Setup. We evaluate with CBGBench (Lin et al., 2025b), which scores models in four categories. *Substructure* metrics are JSD/MAE between generated and reference distributions of atoms, rings and functional groups. *Chemical property* metrics are QED, SA, LogP and LPSK. *Geometry* metrics are JSD of bond lengths and bond angles plus atomic and molecular clash ratios. *Interaction* metrics are Vina energy in score, minimize and dock modes together with protein–ligand interaction type distributions. Models are ranked within each category and ranks are converted to scores as $(N - \text{rank})$ with N the number of models; the overall score is a weighted sum with weights $(0.2, 0.2, 0.4, 0.2)$ for substructure, chemistry, interaction and geometry.

Results. Table 3 summarizes the overall CBGBench performance for small molecule generation. Compared with the corresponding UNIMOMO model, UNI^2 improves the weighted score under both settings. In the single-task setting, UNI^2 improves the overall rank from 6th to 5th. In the all-task setting, UNI^2 achieves the best overall performance.

The per-category results in Appendix A provide a more detailed breakdown. For substructure distributions, UNI^2 improves over UNIMOMO in both settings, reducing the mean rank from 6th to 5th in the single-task setting and from 2nd to 1st in the all-task setting. For chemical properties, UNI^2 also achieves better ranks than UNIMOMO, improving from 5th to 4th in the single-task setting and from 2nd to 1st in the all-task setting. In geometry, UNI^2 substantially reduces clash ratios compared with UNIMOMO, although the bond-length and bond-angle JSDs are slightly higher, suggesting a trade-off between local distribution matching and steric plausibility. For binding and interaction metrics, UNI^2 improves the single-task model ranking from 9th to 7th, mainly through better docking-related energy and interaction scores, while the all-task setting remains comparable to UNIMOMO. Overall, UNI^2 provides consistent gains in the aggregate benchmark, with the most evident improvements in substructure distribution, chemical-property matching, and clash reduction.

Table 3: Small molecule: CBGBench overall comparison. Category scores are rank-derived; higher is better. Best in **bold**, second and third best underlined.

Model	Substruct.	Chem.	Interact.	Geom.	Wtd ↑	Rank
LIGAN	1.13	1.50	**4.87**	1.25	8.75	8
3DSBDD	1.13	1.80	2.50	0.80	6.23	11
GraphBP	0.17	1.70	0.40	0.10	2.37	16
Pocket2Mol	0.73	1.35	3.10	0.70	5.88	12
TargetDiff	1.77	1.60	4.03	1.90	9.30	7
DiffSBDD	0.77	1.95	1.27	1.00	4.98	14
DiffBP	0.27	1.20	2.33	1.45	5.25	13
FLAG	0.70	1.60	1.53	0.65	4.48	15
D3FG	1.47	2.45	2.00	0.70	6.62	10
DecompDiff	1.90	1.90	2.70	2.00	8.50	9
MolCRAFT	2.03	1.65	4.57	**2.40**	10.65	3
VoxBind	1.63	2.20	4.43	2.20	10.47	4
UNIMOMO (single)	2.30	2.35	2.83	2.10	9.58	6
UNI^2 (single)	2.67	2.45	3.20	2.00	10.32	5
UNIMOMO (all)	2.63	2.50	4.00	**2.40**	11.53	2
UNI^2 (all)	**2.70**	**2.55**	3.97	2.35	**11.57**	**1**

5 ABLATION : WHERE DOES THE GAIN COME FROM?

For each benchmark, we use the reported configuration and vary only the gradient-truncation regime. Tables 4, 5, report the full ablation results. Across peptide, antibody, and small molecule benchmarks, several consistent patterns emerge.

One-stage training already provides clear gains. Even when the latent variables are fully detached from the diffusion loss, *w/o grad* h-x improves over the two-stage UNIMOMO baseline on many structural and energetic metrics. This suggests that part of the improvement comes from replacing the separated encoder-denoiser training pipeline with a shared one-stage schedule. In this setting, the encoder and denoiser are exposed to a consistent training curriculum, reducing the mismatch between latent encoding and denoising.

The two latent channels control complementary factors. The ablations indicate that $\mathbf{z}_x$ and $\mathbf{z}_h$ contribute differently to generation quality. Removing the gradient through $\mathbf{z}_h$ tends to favor geometric quality, leading to strong RMSD , clash, and geometry-related scores. In contrast, removing the gradient through $\mathbf{z}_x$ can preserve or improve some chemistry- and interaction-related metrics, but often weakens geometric consistency. This supports the interpretation that $\mathbf{z}_x$ primarily captures spatial and geometric variation, whereas $\mathbf{z}_h$ carries information more closely related to component type, chemistry, and interaction patterns.

Optimizing only one channel leads to trade-offs. Although each truncated variant can perform well on a subset of metrics, neither dominates across all evaluation dimensions. Geometry -oriented variants may lose binding or interaction quality, while chemistry-oriented variants can introduce larger RMSD or clash rates. These trade-offs appear consistently across modalities, indicating that peptide, antibody, and small molecule design all require coordinated optimization of both structural and semantic latent representations.

Keeping both gradients gives the best overall balance. The full *w grad* h-x model is not always the best on every isolated metric, but it provides the most balanced performance across tasks. It achieves strong peptide structure and energy metrics, competitive antibody RMSD and rationality scores, and the best overall weighted score in the small molecule ablation. These results suggest that allowing the diffusion objective to shape both latent channels is important for obtaining a Pareto -balanced model rather than optimizing a single aspect of generation quality.

Table 4: Ablation, peptide, single-modality (top) and multi-modality (bottom) training.

Model	AAR	C-RMSD	L-RMSD	ΔG	IMP	Clash_{in}	$\text{Clash}_{\text{out}}$	JSD_{bb}	JSD_{sc}	Seq.	Struct.
single-modality											
UNIMOMO (single)	37.59%	2.48	1.48	−28.72	29.03%	1.53%	0.94%	0.390	0.365	0.626	**0.629**
UNI^2 w/o grad h-x	**39.47%**	2.10	0.90	−32.61	36.56%	0.30%	0.16%	0.294	0.225	**0.653**	0.586
UNI^2 w/o grad h	38.18%	**1.84**	**0.75**	−32.84	33.33%	**0.16%**	**0.13%**	0.280	**0.210**	0.518	0.544
UNI^2 w/o grad x	37.93%	1.99	0.94	−31.51	31.18%	0.33%	0.19%	0.303	0.235	0.645	0.584
UNI^2 w grad h-x	38.91%	1.92	0.85	**−33.91**	**38.71%**	0.22%	0.15%	**0.276**	0.219	0.638	0.502
multi-modality											
UNIMOMO (all)	39.45%	2.19	1.27	−34.35	40.86%	0.45%	0.93%	**0.205**	0.180	0.617	**0.573**
UNI^2 w/o grad h-x	38.80%	1.89	0.85	−35.47	43.01%	**0.10%**	0.08%	0.258	0.178	**0.646**	0.488
UNI^2 w/o grad h	38.04%	1.82	0.72	−35.49	43.16%	0.12%	**0.06%**	0.248	0.186	0.635	0.506
UNI^2 w/o grad x	**40.14%**	1.86	0.79	−36.07	**48.39%**	0.11%	0.07%	0.254	**0.175**	0.618	0.472
UNI^2 w grad h-x	39.57%	**1.78**	**0.70**	**−36.14**	44.09%	0.11%	0.07%	0.246	**0.175**	0.623	0.460

Table 5: Ablation on antibody (left metrics) and small molecule (right metrics), single-modality (top) and multi-modality (bottom) training. Antibody metrics are reported at $n = 100$; small molecule category scores are computed within each ablation group ($N = 5$).

	Antibody							*Small molecule*					
Model	AAR	RMSD	IMP	Clash_{in}	$\text{Clash}_{\text{out}}$	JSD_{bb}	JSD_{sc}	Substruct.	Chem.	Interact.	Geom.	Wtd ↑	Rank
single-modality													
UNIMOMO (single)	48.78%	1.39	63.33%	0.25%	0.06%	0.278	0.284	0.17	0.45	0.67	0.30	1.58	5
UNI^2 w/o grad h-x	51.20%	1.09	56.67%	0.19%	0.05%	0.256	0.279	**0.60**	0.25	0.60	**0.60**	2.05	4
UNI^2 w/o grad h	**53.12%**	**0.89**	61.67%	**0.11%**	**0.04%**	**0.230**	**0.258**	**0.60**	0.60	0.47	**0.60**	2.27	2.5
UNI^2 w/o grad x	49.42%	1.10	63.33%	0.15%	0.05%	0.242	0.269	0.17	**0.65**	**1.20**	0.25	2.27	2.5
UNI^2 w grad h-x	50.26%	1.08	**65.00%**	0.14%	0.05%	0.240	0.270	0.47	0.55	1.10	0.25	**2.37**	**1**
multi-modality													
UNIMOMO (all)	**52.34%**	1.04	65.00%	0.18%	**0.03%**	0.224	0.221	0.40	0.50	1.03	0.40	2.33	2
UNI^2 w/o grad h-x	50.46%	0.90	65.00%	0.18%	**0.03%**	0.232	0.224	0.10	0.45	0.67	0.25	1.47	5
UNI^2 w/o grad h	50.89%	**0.77**	65.00%	0.14%	**0.03%**	0.223	0.220	**0.53**	0.30	0.33	0.45	1.62	4
UNI^2 w/o grad x	50.12%	0.87	**73.33%**	0.19%	**0.03%**	0.236	0.223	0.47	**0.70**	0.63	0.40	2.20	3
UNI^2 w grad h-x	50.32%	0.85	68.33%	**0.12%**	**0.03%**	**0.216**	**0.217**	**0.53**	0.60	**1.33**	**0.55**	**3.02**	**1**

Per-category breakdowns for the small molecule ablations (substructure, chemistry, geometry, interaction) are deferred to Appendix B.

6 LATENT-SPACE ANALYSIS

Unified training aligns generated and posterior manifolds. We measure posterior-to-generated mismatch using standardized Sliced Wasserstein-1 (SW1) distance, where latent dimensions are normalized by posterior standard deviation before computing sliced Wasserstein distances. Lower SW1 indicates better agreement between generated and posterior latent distributions. We also visualize the two distributions by projecting standardized latents to two principal components and overlaying their density contours. As shown in Table 6, unified training consistently reduces the latent mismatch, especially for the content latent $\mathbf{z}_h$. Compared with the two-stage baseline, full unified training reduces $\mathbf{z}_h$ SW1 across all settings, from 0.0883 to 0.0206 on peptides, 0.0449 to 0.0373 on antibodies, 0.0886 to 0.0748 on small molecules, and 0.0194 to 0.0135 under multi-data training. The improvements for $\mathbf{z}_x$ are smaller, which is consistent with its already lower mismatch in the two-stage baseline. The gradient ablation further confirms this pattern: preserving the $\mathbf{z}_h$ gradient is important for aligning the generated and posterior $\mathbf{z}_h$ manifolds. Overall, the full model achieves the best SW1 in six of the eight benchmark-channel comparisons.

The visualization in Figure 2 supports the same conclusion. Under two-stage training, the generated and posterior contours are more separated, especially in $\mathbf{z}_h$. Unified training produces stronger overlap and more similar contour shapes, indicating that the LDM better reproduces the learned posterior manifold rather than merely producing a generic smooth latent distribution.

Reconstruction quality. The improved latent alignment does not come from a collapse in VAE reconstruction. As shown in Table 7, PB-valid reconstruction improves from 0.570 to 0.591 on peptides, while the decreases on antibodies and small molecules are modest, from 0.450 to 0.417 and

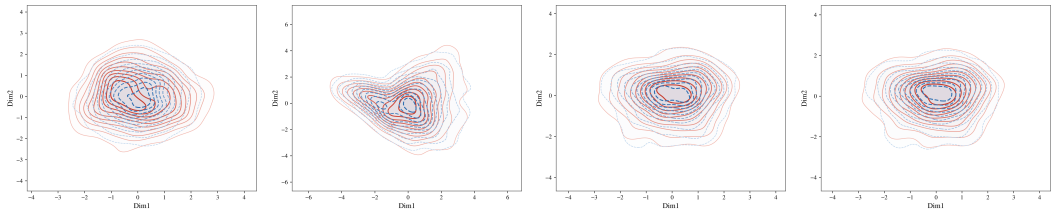


Figure 2: Latent-space visualization of the posterior and generated distributions under two-stage and unified training on the antibody dataset. From left to right, the panels show $\mathbf{z}_h$ (two-stage), $\mathbf{z}_h$ (unified), $\mathbf{z}_x$ (two-stage), and $\mathbf{z}_x$ (unified). Each panel overlays the standardized posterior and generated densities, with the standardized Sliced W1 distance reported below. Visualizations for peptide, small molecule, and multi-data settings are provided in Appendix C.

Table 6: Posterior-to-generated distribution mismatch measured by standardized SW1 on held-out complexes. Lower is better. Bold marks the lowest value in each benchmark and latent channel.

	Peptide		Antibody		Small molecule		Multi-data	
Training	$\mathbf{z}_h$	$\mathbf{z}_x$	$\mathbf{z}_h$	$\mathbf{z}_x$	$\mathbf{z}_h$	$\mathbf{z}_x$	$\mathbf{z}_h$	$\mathbf{z}_x$
UNIMOMO (single)	0.0883	0.0253	0.0449	0.0225	0.0886	0.0169	0.0194	0.0160
UNI^2 w/o grad h-x	0.0506	0.0217	0.0763	0.0218	0.1110	0.0163	0.0197	0.0155
UNI^2 w/o grad h	0.0405	**0.0206**	0.0732	0.0210	0.1106	0.0143	0.0181	0.0150
UNI^2 w/o grad x	0.0214	0.0210	**0.0359**	0.0214	0.0769	0.0154	0.0167	0.0152
UNI^2 w grad h-x	**0.0206**	0.0207	0.0373	**0.0209**	**0.0748**	**0.0137**	**0.0135**	**0.0149**

Table 7: VAE reconstruction quality measured by PB-valid fraction on the test dataset.

Training	Peptide ↑	Antibody ↑	Small molecule ↑
UNIMOMO	0.570	**0.450**	**0.570**
UNI^2	**0.591**	0.417	0.530

from 0.570 to 0.530, respectively. These results suggest that unified training improves the compatibility between the encoder posterior and the diffusion prior while largely preserving reconstruction quality.

7 CONCLUSION

Two-stage latent generative modeling is a common practice, but it is not a necessary choice for structure-based binder design. Our results show that training the autoencoder and the latent diffusion model jointly in a single stage allows the diffusion loss to directly shape the latent space that will later be sampled from . This leads to consistent improvements across peptide, antibody, and small molecule design tasks. The gains hold under both single-modality and multi-modality training settings, without introducing any additional inference cost.

The gradient- flow ablation further explains where these improvements come from . The equivariant latent channel mainly affects geometric modeling, whereas the invariant latent channel captures information more closely related to chemical properties and interactions. Keeping both channels in the gradient path enables the model to better balance structural plausibility and chemical validity. The latent-space analysis based on standardized SW1 supports the same conclusion: unified training makes the generated latent distribution better aligned with the encoder posterior manifold, especially in the content-related latent channel. At the same time, VAE reconstruction quality is largely preserved, indicating that this improved latent alignment is not achieved at the cost of reconstruction ability. Together, these findings suggest that diffusion feedback helps shape a latent space that supports both faithful reconstruction and generation, with complementary contributions from the geometric and chemical channels.

AI USE STATEMENT

We used generative AI tools to edit the manuscript for readability, grammar, and clarity. We have reviewed all AI-assisted text and take responsibility for the final content of this work.

REFERENCES

Jared Adolf-Bryfogle, Oleks Kalyuzhniy, Michael Kubitz, Brian D Weitzner, Xiaozhen Hu, Yumiko Adachi, William R Schief, and Roland L Dunbrack Jr. Rosettaantibodydesign (rabd): A general framework for computational antibody design. *PLoS computational biology*, 14(4):e1006112, 2018.

Rebecca F Alford, Andrew Leaver-Fay, Jeliazko R Jeliazkov, Matthew J O'Meara, Frank P DiMaio, Hahnbeom Park, Maxim V Shapovalov, P Douglas Renfrew, Vikram K Mulligan, Kalli Kappel, et al. The rosetta all-atom energy function for macromolecular modeling and design. *Journal of chemical theory and computation*, 13(6):3031, 2017.

Cyrus Chothia and Arthur M Lesk. Canonical structures for the hypervariable regions of immunoglobulins. *Journal of molecular biology*, 196(4):901–917, 1987.

Justas Dauparas, Ivan Anishchenko, Nathaniel Bennett, Hua Bai, Robert J Ragotte, Lukas F Milles, Basile IM Wicky, Alexis Courbet, Rob J de Haas, Neville Bethel, et al. Robust deep learning–based protein sequence design using proteinmpnn. *Science*, 378(6615):49–56, 2022.

James Dunbar, Konrad Krawczyk, Jinwoo Leem, Terry Baker, Angelika Fuchs, Guy Georges, Jiye Shi, and Charlotte M Deane. Sabdab: the structural antibody database. *Nucleic acids research*, 42(D1):D1140–D1146, 2014.

Roland L Dunbrack Jr and Fred E Cohen. Bayesian statistical analysis of protein side-chain rotamer preferences. *Protein science*, 6(8):1661–1681, 1997.

Paul G. Francoeur, Tomohide Masuda, Jocelyn Sunseri, Andrew Jia, Richard B. Iovanisci, Ian Snyder, and David Ryan Koes. 3d convolutional neural networks and a crossdocked dataset for structure-based drug design. *Journal of chemical information and modeling*, 60:4200 – 4215, 2020. URL `https://api.semanticscholar.org/CorpusID:221383064`.

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.

Steven Henikoff and Jorja G Henikoff. Amino acid substitution matrices from protein blocks. *Proceedings of the national academy of sciences*, 89(22):10915–10919, 1992.

Xiangzhe Kong, Wenbing Huang, and Yang Liu. Conditional antibody design as 3d equivariant graph translation (2023). In *URL https://openreview. net/forum*.

Xiangzhe Kong, Wenbing Huang, and Yang Liu. End-to-end full-atom antibody design. *arXiv preprint arXiv:2302.00203*, 2023.

Xiangzhe Kong, Yinjun Jia, Wenbing Huang, and Yang Liu. Full-atom peptide design with geometric latent diffusion. *Advances in Neural Information Processing Systems*, 37:74808–74839, 2024.

Marie-Paule Lefranc, Christelle Pommié, Manuel Ruiz, Véronique Giudicelli, Elodie Foulquier, Lisa Truong, Valérie Thouvenin-Contet, and Gérard Lefranc. Imgt unique numbering for immunoglobulin and t cell receptor variable domains and ig superfamily v-like domains. *Developmental & Comparative Immunology*, 27(1):55–77, 2003.

Jiahan Li, Chaoran Cheng, Zuofan Wu, Ruihan Guo, Shitong Luo, Zhizhou Ren, Jian Peng, and Jianzhu Ma. Full-atom peptide design based on multi-modal flow matching. *arXiv preprint arXiv:2406.00735*, 2024.

Haitao Lin, Yufei Huang, Odin Zhang, Yunfan Liu, Lirong Wu, Siyuan Li, Zhiyuan Chen, and Stan Z Li. Functional-group-based diffusion for pocket-specific molecule generation and elaboration. *Advances in Neural Information Processing Systems*, 36:34603–34626, 2023.

Haitao Lin, Lirong Wu, Yufei Huang, Yunfan Liu, Odin Zhang, Yuanqing Zhou, Rui Sun, and Stan Z Li. Geoab: towards realistic antibody design and reliable affinity maturation. *bioRxiv*, pp. 2024–05, 2024.

Haitao Lin, Yufei Huang, Odin Zhang, Siqi Ma, Meng Liu, Xuanjing Li, Lirong Wu, Jishui Wang, Tingjun Hou, and Stan Z Li. Diffbp: Generative diffusion of 3d molecules for target protein binding. *Chemical Science*, 16(3):1417–1431, 2025a.

Haitao Lin, Guojiang Zhao, Odin Zhang, Yufei Huang, Lirong Wu, Cheng Tan, Zicheng Liu, Zhifeng Gao, and Stan Z Li. Cbgbench: fill in the blank of protein-molecule complex binding graph. In *International Conference on Learning Representations*, volume 2025, pp. 19861–19887, 2025b.

M Liu, Y Luo, K Uchino, K Maruhashi, and S Ji. Generating 3d molecules for target protein binding, arxiv, 2022. *arXiv preprint arXiv:2204.09410*, 10.

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.

Shitong Luo, Yufeng Su, Xingang Peng, Sheng Wang, Jian Peng, and Jianzhu Ma. Antigen-specific antibody design and optimization with diffusion-based generative models for protein structures. *Advances in Neural Information Processing Systems*, 35:9754–9767, 2022.

T Masuda, M Ragoza, and DR Koes. Generating 3d molecular structures conditional on a receptor binding site with deep generative models, arxiv, 2020. *arXiv preprint arXiv:2010.14442*, 10, 2010.

Saul B Needleman and Christian D Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, 48(3):443–453, 1970.

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.

Pedro O Pinheiro, Arian Jamasb, Omar Mahmood, Vishnu Sresht, and Saeed Saremi. Structure-based drug design by denoising voxel grids. *arXiv preprint arXiv:2405.03961*, 2024.

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.

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.

Tomer Tsaban, Julia K Varga, Orly Avraham, Ziv Ben-Aharon, Alisa Khramushin, and Ora Schueler-Furman. Harnessing protein folding neural networks for peptide–protein docking. *Nature communications*, 13(1):176, 2022.

Joseph L Watson, David Juergens, Nathaniel R Bennett, Brian L Trippe, Jason Yim, Helen E Eisenach, Woody Ahern, Andrew J Borst, Robert J Ragotte, Lukas F Milles, et al. De novo design of protein structure and function with rfdiffusion. *Nature*, 620(7976):1089–1100, 2023.

Fei Ye, Zaixiang Zheng, Dongyu Xue, Yuning Shen, Lihao Wang, Yiming Ma, Yan Wang, Xinyou Wang, Xiangxin Zhou, and Quanquan Gu. Proteinbench: A holistic evaluation of protein foundation models. In *International Conference on Learning Representations*, volume 2025, pp. 29857–29891, 2025.

Zaixi Zhang, Yaosen Min, Shuxin Zheng, and Qi Liu. Molecule generation for target protein binding with structural motifs. In *The eleventh international conference on learning representations*, 2023.

A PER-CATEGORY SMALL MOLECULE RESULTS

Per-category CBGBench breakdowns for the small molecule benchmark (overall leaderboard in Table 3). Marking follows the main text: best in **bold**, second and third best underlined.

Table 8: Substructure analysis for de novo small molecule design.

Model	Atom JSD	Atom MAE	Ring JSD	Ring MAE	FG JSD	FG MAE	MeanRank	Pos
LIGAN	0.1167	0.8680	0.3163	0.2701	0.2468	0.0378	10.33	10
3DSBDD	0.0860	0.8444	0.3188	0.2457	0.2682	0.0494	10.33	11
GraphBP	0.1642	1.2266	0.5061	0.4382	0.6259	0.0705	15.17	16
Pocket2Mol	0.0916	1.0497	0.3550	0.3545	0.2961	0.0622	12.33	13
TargetDiff	0.0533	0.2399	0.2345	0.1559	0.2876	0.0441	7.17	7
DiffSBDD	0.0529	0.6316	0.3853	0.3437	0.5520	0.0710	12.17	12
DiffBP	0.2591	1.5491	0.4531	0.4068	0.5346	0.0670	14.67	15
FLAG	0.1032	1.7665	0.2432	0.3370	0.3634	0.0666	12.50	14
D3FG	0.0644	0.8154	0.1869	0.2204	0.2511	0.0516	8.67	9
DecompDiff	0.0431	0.3197	0.2431	0.2006	0.1916	0.0318	6.50	6
MolCRAFT	0.0490	0.3208	0.2469	0.0264	0.1196	0.0477	5.83	5
VoxBind	0.0942	0.3564	0.2401	0.0301	**0.1053**	0.0761	7.83	8
=UNIMOMO (single)	0.0352	0.2796	0.1252	0.0976	0.1562	0.0173	4.50	4
UNI^2 (single)	0.0336	0.2111	**0.0807**	**0.0216**	0.1349	0.0202	2.67	2
UNIMOMO (all)	**0.0280**	0.2108	0.1073	0.0726	0.1356	**0.0161**	2.83	3
UNI^2 (all)	0.0316	**0.1960**	0.0854	0.0388	0.1332	0.0190	**2.50**	**1**

Table 9: Results of chemical properties for de novo small molecule design. LogP between -0.4 and 5.6 are all reasonable (Lin et al., 2024d), without preference to higher or lower values.

Model	QED	LogP	SA	LPSK	MeanRank	Pos
LIGAN	0.46	0.56	0.66	4.39	8.50	14
3DSBDD	0.48	0.47	0.63	4.72	7.00	9
GraphBP	0.44	3.29	0.64	4.73	7.50	10
Pocket2Mol	0.39	2.39	0.65	4.58	9.25	15
TargetDiff	0.49	1.13	0.60	4.57	8.00	12
DiffSBDD	0.49	−0.15	0.34	4.89	6.25	7
DiffBP	0.47	5.27	0.59	4.47	10.00	16
FLAG	0.41	0.29	0.58	**4.93**	8.00	13
D3FG	0.49	1.56	0.66	4.84	3.75	3
DecompDiff	0.49	1.22	0.66	4.40	6.50	8
MolCRAFT	0.48	0.87	0.66	4.39	7.75	11
VoxBind	0.54	2.22	0.65	4.70	5.00	6
UNIMOMO (single)	0.53	1.36	0.68	4.69	4.25	5
UNI^2 (single)	0.54	2.06	0.71	4.65	3.75	4
UNIMOMO (all)	**0.55**	1.55	0.70	4.68	3.50	2
UNI^2 (all)	**0.55**	1.97	**0.72**	4.63	**3.25**	**1**

Table 10: Geometry analysis for de novo small molecule design.

Model	JSD_{BL}	JSD_{BA}	$Ratio_{c\text{-}c\text{-}a}$	$Ratio_{c\text{-}m}$	MeanRank	Pos
LIGAN	0.4645	0.5673	0.0096	0.0718	9.75	10
3DSBDD	0.5024	0.3904	0.2482	0.8683	12.00	12
GraphBP	0.5182	0.5645	0.8634	0.9974	15.50	16
Pocket2Mol	0.5433	0.4922	0.0576	0.4499	12.50	13
TargetDiff	0.2659	0.3769	0.0483	0.4920	6.50	8
DiffSBDD	0.3501	0.4588	0.1083	0.6578	11.00	11
DiffBP	0.3453	0.4621	0.0449	0.4077	8.75	9
FLAG	0.4215	0.4304	0.6777	0.9769	12.75	15
D3FG	0.3727	0.4700	0.2115	0.8571	12.50	14
DecompDiff	0.2576	0.3473	0.0462	0.5248	6.00	6
MolCRAFT	**0.2250**	**0.2683**	0.0264	0.2691	**4.00**	**1**
VoxBind	0.2701	0.3771	0.0103	0.1890	5.00	4
UNIMOMO (single)	0.3362	0.4166	0.0042	0.0715	5.50	5
UNI^2 (single)	0.3519	0.4334	0.0026	0.0482	6.00	7
UNIMOMO (all)	0.3223	0.3848	0.0040	0.0708	**4.00**	2
UNI^2 (all)	0.3482	0.4136	**0.0015**	**0.0279**	4.25	3

Table 11: Results of interaction analysis for de novo small molecule design.

Model	Vina Score	Vina Score IMP	Vina Min	Vina Min IMP	Vina Dock	Vina Dock IMP	MPBG	LBE	JSD_{OA}	MAE_{OA}	JSD_{PP}	MAE_{PP}	MeanRank	Pos
LIGAN	**−6.47**	**62.13**	**−7.14**	**70.18**	−7.70	**72.71**	4.22	0.3897	0.0346	0.0905	0.1451	**0.3416**	**3.83**	**1**
3DSBDD	–	3.99	−3.75	17.98	−6.45	31.46	9.18	0.3839	0.0392	0.0934	0.1733	0.4231	9.75	11
GraphBP	–	0.00	–	1.67	−4.57	10.86	−30.03	0.3200	0.0462	0.1625	0.2101	0.4835	15.00	16
Pocket2Mol	−5.23	31.06	−6.03	38.04	−7.05	48.07	−0.17	**0.4115**	0.0319	0.2455	0.1535	0.4152	8.25	8
TargetDiff	−5.71	38.21	−6.43	47.09	−7.41	51.99	5.38	0.3537	0.0198	0.0600	0.1757	0.4687	5.92	4
DiffSBDD	–	12.67	−2.15	22.24	−5.53	29.76	−23.51	0.2920	0.0333	0.1461	0.1777	0.5265	12.83	15
DiffBP	–	8.60	–	19.68	−7.34	49.24	6.23	0.3481	0.0249	0.1430	**0.1256**	0.5639	10.17	12
FLAG	–	0.04	–	3.44	−3.65	11.78	−47.64	0.3319	0.0170	0.0277	0.2762	0.3976	12.17	14
D3FG	–	3.70	−2.59	11.13	−6.78	28.90	−8.85	0.4009	0.0638	**0.0135**	0.1850	0.4641	11.00	13
DecompDiff	−5.18	19.66	−6.04	34.84	−7.10	48.31	−1.59	0.3460	0.0215	0.0769	0.1848	0.4369	9.25	10
MolCRAFT	−6.15	54.25	−6.99	56.43	**−7.79**	56.22	8.38	0.3638	0.0214	0.0780	0.1868	0.4574	4.58	2
VoxBind	−6.16	41.80	−6.82	50.02	−7.68	52.91	**9.89**	0.3588	0.0257	0.0533	0.1850	0.4606	4.92	3
UniMoMo (single)	−5.50	26.37	−5.96	38.21	−7.19	49.60	2.15	0.3521	0.0175	0.0447	0.2372	0.4706	8.92	9
Uni^2 (single)	−5.50	27.09	−6.08	38.45	−7.44	54.15	6.02	0.3438	0.0284	0.0649	0.2353	0.4404	8.00	7
UniMoMo (all)	−5.72	30.40	−6.08	39.23	−7.25	51.59	7.50	0.3473	**0.0135**	0.0173	0.2012	0.4277	6.00	5
Uni^2 (all)	−5.66	30.07	−6.18	40.86	−7.40	54.28	5.11	0.3455	0.0160	0.0177	0.2003	0.4270	6.08	6

B PER-CATEGORY SMALL MOLECULE ABLATIONS

Marking follows the main text: best in **bold**, second best underlined.

Table 12: Ablation study on substructure analysis for de novo small molecule design.

Model	Atom JSD	Atom MAE	Ring JSD	Ring MAE	FG JSD	FG MAE	MeanRank	Pos
single-modality								
UniMoMo (single)	0.0352	0.2796	0.1252	0.0976	0.1562	**0.0173**	4.17	4
Uni^2 w/o grad h-x	**0.0211**	**0.1036**	0.0858	0.0419	**0.1339**	0.0201	**2.00**	**1**
Uni^2 w/o grad h	0.0270	0.1726	**0.0755**	0.0341	0.1432	0.0198	**2.00**	2
Uni^2 w/o grad x	0.0400	0.1920	0.0890	0.0526	0.1546	0.0282	4.17	5
Uni^2 w grad h-x	0.0336	0.2111	0.0807	**0.0216**	0.1349	0.0202	2.67	3
multi-modality								
UniMoMo (all)	**0.0280**	0.2108	0.1073	0.0726	0.1356	**0.0161**	3.00	4
Uni^2 w/o grad h-x	0.0481	0.3092	0.0819	0.0479	0.1545	0.0255	4.50	5
Uni^2 w/o grad h	0.0385	0.2162	0.0807	**0.0273**	**0.1332**	0.0189	**2.33**	**1**
Uni^2 w/o grad x	0.0372	0.2032	**0.0786**	0.0293	0.1485	0.0202	2.67	3
Uni^2 w grad h-x	0.0316	**0.1960**	0.0854	0.0388	**0.1332**	0.0190	**2.33**	2

Table 13: Ablation results of chemical properties for de novo small molecule design.

Model	QED	LogP	SA	LPSK	MeanRank	Pos
single-modality						
UniMoMo (single)	0.53	1.36	0.68	**4.69**	2.75	4
Uni^2 w/o grad h-x	0.53	1.60	0.71	4.58	3.75	5
Uni^2 w/o grad h	**0.54**	1.69	0.72	4.65	2.00	2
Uni^2 w/o grad x	**0.54**	2.08	**0.73**	4.63	**1.75**	**1**
Uni^2 w grad h-x	**0.54**	2.06	0.71	4.65	2.25	3
multi-modality						
UniMoMo (all)	0.55	1.55	0.70	4.68	2.50	3
Uni^2 w/o grad h-x	0.54	2.26	**0.73**	4.61	2.75	4
Uni^2 w/o grad h	0.53	1.97	0.71	4.63	3.50	5
Uni^2 w/o grad x	**0.56**	1.99	0.72	**4.70**	**1.50**	**1**
Uni^2 w grad h-x	0.55	1.97	0.72	4.63	2.00	2

Table 14: Abaltion study on geometry analysis for de novo small molecule design.

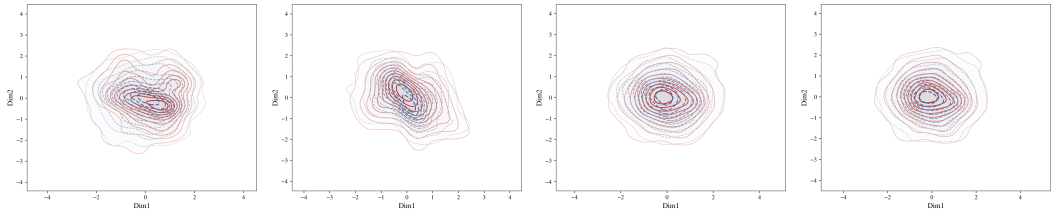
Model	JSD_{BL}	JSD_{BA}	$Ratio_{c\text{-}c\text{-}a}$	$Ratio_{c\text{-}m}$	MeanRank	Pos
single-modality						
UniMoMo (single)	**0.3362**	0.4166	0.0042	0.0715	3.50	3
Uni^2 w/o grad h-x	0.3388	**0.4112**	0.0025	0.0441	**2.00**	**1**
Uni^2 w/o grad h	0.3377	0.4331	**0.0023**	**0.0404**	**2.00**	2
Uni^2 w/o grad x	0.3559	0.4123	0.0032	0.0583	3.75	4
Uni^2 w grad h-x	0.3519	0.4334	0.0026	0.0482	3.75	5
multi-modality						
UniMoMo (all)	**0.3223**	**0.3848**	0.0040	0.0708	3.00	3
Uni^2 w/o grad h-x	0.3321	0.4375	0.0020	0.0375	3.75	5
Uni^2 w/o grad h	0.3414	0.4102	0.0017	0.0311	2.75	2
Uni^2 w/o grad x	0.3532	0.4163	**0.0015**	0.0282	3.00	4
Uni^2 w grad h-x	0.3482	0.4136	**0.0015**	**0.0279**	**2.25**	**1**

Table 15: Abaltion results of interaction analysis for de novo small molecule design.

Model	Vina Score	Vina Score IMP	Vina Min	Vina Min IMP	Vina Dock	Vina Dock IMP	MPBG	LBE	JSD_{OA}	MAE_{OA}	JSD_{PP}	MAE_{PP}	MeanRank	Pos
single-modality														
UniMoMo (single)	−5.50	26.37	−5.96	38.21	−7.19	49.60	2.15	**0.3521**	0.0175	0.0447	0.2372	0.4706	3.33	3
Uni^2 w/o grad h-x	−5.31	23.63	−5.93	34.89	−7.27	50.97	3.43	0.3359	**0.0147**	**0.0287**	0.2354	0.5024	3.50	4
Uni^2 w/o grad h	−5.40	23.84	−5.97	34.40	−7.28	50.33	3.06	0.3440	0.0352	0.0358	0.2585	0.4724	3.83	5
Uni^2 w/o grad x	**−5.54**	**29.00**	**−6.18**	**40.76**	**−7.45**	54.14	5.64	0.3486	0.0165	0.0318	0.2579	0.5025	**2.00**	**1**
Uni^2 w grad h-x	−5.50	27.09	−6.08	38.45	−7.44	**54.15**	**6.02**	0.3438	0.0284	0.0649	**0.2353**	**0.4404**	2.25	2
multi-modality														
UniMoMo (all)	**−5.72**	**30.40**	−6.08	39.23	−7.25	51.59	**7.50**	0.3473	**0.0135**	**0.0173**	0.2012	**0.4277**	2.42	2
Uni^2 w/o grad h-x	−5.59	29.49	−6.13	39.90	−7.38	54.06	4.27	0.3412	0.0402	0.1054	0.2210	0.4628	3.33	3
Uni^2 w/o grad h	−5.47	27.00	−6.11	38.46	−7.33	52.99	4.21	0.3415	0.0158	0.0549	0.2341	0.5330	4.17	5
Uni^2 w/o grad x	−5.55	28.13	−6.15	39.06	−7.36	52.54	4.39	**0.3499**	0.0379	0.0641	0.2333	0.4940	3.42	4
Uni^2 w grad h-x	−5.66	30.07	**−6.18**	**40.86**	**−7.40**	**54.28**	5.11	0.3455	0.0160	0.0177	**0.2003**	0.4270	**1.67**	**1**

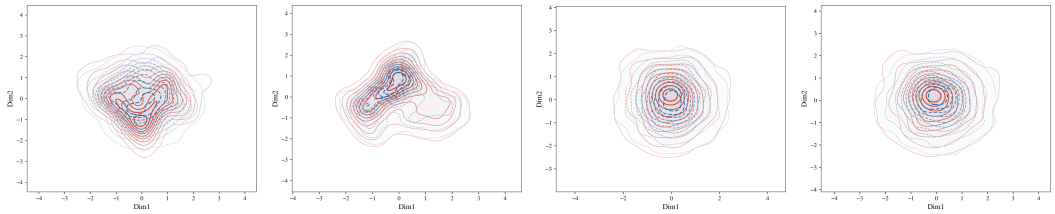
C ADDITIONAL LATENT-SPACE VISUALIZATIONS

This appendix provides additional latent-space visualizations for the peptide, small molecule, and multi-data settings. The visualization protocol is the same as in Figure 2: standardized posterior and generated latent distributions are projected to two principal components and overlaid as density contours. The standardized Sliced W1 distance is reported below each panel, with the corresponding quantitative results summarized in Table 6.



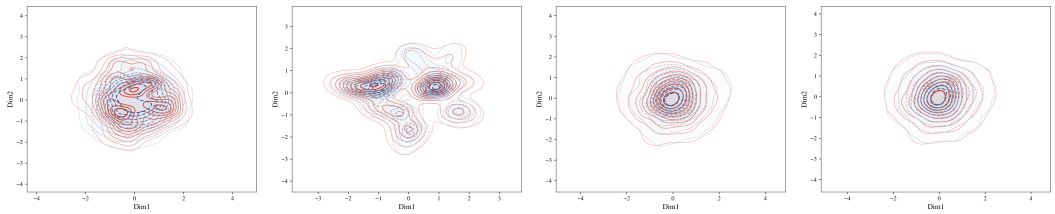
Sliced W1 = 0.0883 Sliced W1 = 0.0206 Sliced W1 = 0.0253 Sliced W1 = 0.0207

Figure 3: Latent-space visualization of the posterior and generated distributions under two-stage and unified training on the peptide dataset. From left to right: $\mathbf{z}_h$ (two-stage), $\mathbf{z}_h$ (unified), $\mathbf{z}_x$ (two-stage), and $\mathbf{z}_x$ (unified). The standardized Sliced W1 distance is reported below each panel.



Sliced W1 = 0.0886 Sliced W1 = 0.0748 Sliced W1 = 0.0169 Sliced W1 = 0.0137

Figure 4: Latent-space visualization of the posterior and generated distributions under two-stage and unified training on the small molecule dataset. From left to right: $\mathbf{z}_h$ (two-stage), $\mathbf{z}_h$ (unified), $\mathbf{z}_x$ (two-stage), and $\mathbf{z}_x$ (unified). The standardized Sliced W1 distance is reported below each panel.



Sliced W1 = 0.0194 Sliced W1 = 0.0135 Sliced W1 = 0.0160 Sliced W1 = 0.0149

Figure 5: Latent-space visualization of the posterior and generated distributions under two-stage and unified training in the multi-data setting. From left to right: $\mathbf{z}_h$ (two-stage), $\mathbf{z}_h$ (unified), $\mathbf{z}_x$ (two-stage), and $\mathbf{z}_x$ (unified). The standardized Sliced W1 distance is reported below each panel.

D IMPLEMENTATION DETAILS

We provide the hyperparameters for training our single-domain and multi-domain models in Table 16, and the loss weights used for each benchmark and training regime in Table 17. All models share the same backbone architecture; only the loss weights differ across benchmarks.

Table 16: Hyperparameters of UNI^2 with single-domain and multi-domain data.

	UNI^2 (single)	UNI^2 (all)	Description
Variational autoencoder			
Epoch	250	250	Number of epochs to train.
Warmup	2000	2000	Warmup steps for the KL (linear schedule).
Learning rate	1e-4	1e-4	Learning rate.
Embed size	512	512	Dimension of atom and block type embeddings.
Hidden size	512	512	Dimension of hidden states.
Latent size	8	8	Dimension of latent states.
Edge size	64	64	Dimension of edge type embeddings.
n_{rbf}	64	64	Number of RBF kernels for spatial distances.
Cutoff	10.0	10.0	Cutoff distance for RBF kernels.
n_{layers}	6	6	Number of layers in the encoder and the decoder.
n_{head}	8	8	Number of heads for multi-head attention.
λ_1	0.4 (mol.) / 0.6 (others)	0.6	KL divergence weight on the sequence.
λ_2	0.8	0.8	KL divergence weight on the structure.
λ_{dist}	0.5	0.5	Local distance loss weight.
Mask ratio	0.05	0.05	Ratio of binding-site residues for reconstruction.
N	10	10	Number of reconstruction iterations.
Latent diffusion model			
Hidden size	512	512	Dimension of hidden states.
T	100	100	Number of total steps for diffusion.
n_{rbf}	64	64	Number of RBF kernels for spatial distances.
Cutoff	3.0	3.0	RBF kernel cutoff in the normalized latent space.
n_{layers}	6	6	Number of layers in the denoising network.
n_{head}	8	8	Number of heads for multi-head attention.

Table 17: Loss weights of UNI^2 for each benchmark and training regime.

Benchmark	λ_{vae}	λ_{ldm}	$\lambda_{\text{ldm},h}$	$\lambda_{\text{ldm},x}$
Peptide (single)	0.3	1.0	0.375	1.0
Antibody (single)	0.3	1.0	0.5	1.0
Small molecule (single)	0.7	1.0	0.2	1.0
Multi-modality	1.0	1.0	0.5	1.0